Höhere Mathematik für Ingenieure

Teil I

gelesen von

Prof. Dr. Michael Junk
Universität des Saarlandes

im Wintersemester 2003/2004

Vorwort

Kurz gesagt, ist Mathematik ein Hilfsmittel zur **Beschreibung**, **Analyse** und **Vorhersage** von Zusammenhängen in unserer Umwelt, wie etwa

- Temperaturverlauf: Zeitpunkt → Temperaturwert
- Dynamo: Drehgeschwindigkeit → Spannungswert
- Studierende: Matrikelnummer → Person
- Steuer: Einkommen → Steuersatz
- Werkstücke: Punkt auf der Oberfläche → Härte
- Schumacher: Beschleunigungsverlauf → Rundenzeit
- Fußball: Tritt → Flugbahn

Allen diesen Beispielen ist gemeinsam, daß Elemente einer Menge D (Zeitpunkte, Studenten, etc.) bestimmte Elemente einer anderen Menge B (Temperaturwerte, Flugbahnen, etc.) zugeordnet werden.
Entscheidende Bestandteile der „Zuordnung“ bzw. „Funktion“

(1) Definitionsmenge D
(2) Zielmenge B
(3) eindeutige Zuordnungsvorschrift:
$x \in D$ wird $f(x) \in B$ zugeordnet

Diese Bestandteile fassen wir mit einer kompakten Notation zusammen

$$\begin{aligned} f \quad : \quad & D \to B \\ & x \to f(x) \end{aligned}$$

Im Folgenden werden wir typische Beispiele von Mengen und Funktionen kennenlernen. Das Ziel ist dabei, Ideen und Techniken zu lernen, die bei der **Analyse** von Funktionen hilfreich sind. Außerdem werden verschiedene Möglichkeiten der **Beschreibung** von Funktionen vorgestellt. Zum Beispiel kann die Berechnungsvorschrift durch endlich viele elementare Rechenoperationen gegeben sein, oder durch Grenzwerte unendlich vieler Operationen. Eine andere Möglichkeit ist, daß die Vorschrift nur indirekt durch eine Gleichung gegeben ist (z. B. eine Differentialgleichung, d. h. eine Beziehung zwischen Funktionswerten und deren Änderungsraten). In diesem Fall muß ein konkreter Funktionswert durch Lösen der Gleichung bestimmt werden, was einer klassischen

Vorhersage entspricht, wenn der Definitionsbereich aus Zeitpunkten besteht und der Funktionswert zu einem zukünftigen Zeitpunkt bestimmt werden soll.

KAPITEL 1

Elementare Mengen und Funktionen

1. Endliche Mengen

Wir beginnen mit dem Fall endlicher Mengen, die prinzipiell durch Auflistung aller ihrer Elemente beschrieben werden. So ist zum Beispiel die Menge der Monitorgrundfarben gegeben durch $M = \{\text{rot, grün, blau}\}$. Nummeriert man die Elemente der Menge (z. B. beim Abzählen), so erzeugt man in unserer Sichtweise einen „Zusammenhang“ $Z : M \to \{1, 2, 3\}$, etwa durch eine eindeutige Vorschrift der Form

$$Z(\text{rot}) = 1,\ Z(\text{blau}) = 2,\ Z(\text{grün}) = 3$$

Bei Funktionen mit endlichen Definitionsmengen bietet sich zur vollständigen Beschreibung eine Wertetabelle an

x	rot	blau	grün
$Z(x)$	1	2	3

Eine wichtige Klasse solcher Funktionen sind die logischen Operationen. Betrachten wir dazu die Menge $\mathcal{A}$ aller Aussagen, wobei eine Aussage ein grammatikalisch korrekter Satz sein soll, dem genau einer der beiden Wahrheitswerte „wahr“ oder „falsch“ zugeordnet werden kann. Damit existiert also ein Zusammenhang $E : \mathcal{A} \to \{0, 1\} = W$ gemäß der Vorschrift

$$E(A) = \begin{cases} 0 & A \text{ ist falsch} \\ 1 & A \text{ ist wahr} \end{cases}$$

Dem menschlichen Gefühl von Logik entsprechen nun bestimmte Einschränkungen an die Funktion E. Ist nämlich eine Aussage A falsch, z. B. $A =$ „alle Frauen sind blond“, so ist die verneinte Aussage

$$\neg A = \text{„mindestens eine Frau ist nicht blond“}$$

automatisch wahr (Vorsicht, „keine Frau ist blond“ ist *nicht* die Verneinung von „alle Frauen sind blond“).

Umgekehrt ist die Negation einer wahren Aussage falsch. Dieser logische Zusammenhang läßt sich also durch eine Funktion NOT : $W \to W$ mit der Vorschrift

x	0	1
NOT(x)	1	0

beschreiben. Die Einschränkung an E lautet dann

$$E(\neg A) = \mathrm{NOT}(E(A))$$

Entsprechend fordert man weitere Eigenschaften für Verknüpfungen von Aussagen wie etwa der und-Verknüpfung $A \wedge B =$ „A und B“. Betrachten wir zunächst das Beispiel

$$A = \text{„Lampe 1 brennt“}, \; B = \text{„Lampe 2 brennt“}$$

Dann sollte $A \wedge B =$ „Lampe 1 brennt und Lampe 2 brennt“ nur dann wahr sein, wenn beide Lampen brennen, d. h. wenn sowohl A als auch B wahr sind. Bezüglich und-Verknüpfungen verlangen wir also die Einschränkung

$$E(A \wedge B) = \mathrm{AND}((E(A), A(B))),$$

wobei AND : $W \times W \to W$ auf der Menge der Paare

$$W \times W = \{(0,0), (0,1), (1,0), (1,1)\}$$

definiert ist und der Vorschrift

(x, y)	(0,0)	(0,1)	(1,0)	(1,1)
AND((x, y))	0	0	0	1

genügt. (Bemerkung: zur Abkürzung wird die Paar-Klammer oft unterdrückt, d. h. man schreibt AND(x, y) statt AND$((x, y))$.)
Eine weitere wichtige Verknüpfung ist die oder-Verknüpfung $A \vee B$, die in der Logik nicht als entweder-oder verstanden wird, sondern als und-oder. Im obigen Beispiel ist also $A \vee B$ sowohl wahr, wenn genau eine der beiden Lampen brennt, als auch wenn beide Lampen brennen. Anders herum „Lampe 1 brennt oder Lampe 2 brennt“ ist nur dann falsch, wenn beide Lampen aus sind. Die oder-Verknüpfung schränkt

damit unsere „Wahrheitsfunktion“ ein, durch

$$E(A \vee B) = \mathrm{OR}(E(A), E(B))$$

wobei $\mathrm{OR} : W \times W \to W$ gegeben ist durch

(x, y)	(0,0)	(0,1)	(1,0)	(1,1)
$\mathrm{OR}(x, y)$	0	1	1	1

Die umgangssprachliche oder-Verknüpfung im Sinne entweder/oder wird dagegen durch folgenden Zusammenhang ausgedrückt

$\mathrm{XOR} : W \times W \to W$

(x, y)	(0,0)	(0,1)	(1,0)	(1,1)
XOR(x,y)	0	1	1	0

Die Funktion XOR kann auch durch die elementaren Funktionen NOT, AND und OR dargestellt werden.

$$\mathrm{XOR}(x, y) = \mathrm{OR}(\mathrm{AND}(x, \mathrm{NOT}(y)), \mathrm{AND}(\mathrm{NOT}(x), y))$$

Um solche Formeln lesbarer zu machen, benutzt man gerne die gleichen Symbole wie bei den Verknüpfungen der Aussagen

$$\begin{aligned} \neg x &:= \mathrm{NOT}(x) \\ x \wedge y &:= \mathrm{AND}(x, y) \\ x \vee y &:= \mathrm{OR}(x, y) \end{aligned}$$

Zu beachten ist aber, daß wir nun die gleichen Symbole für verschiedene Zusammenhänge benutzen. So ist $\neg : \mathcal{A} \to \mathcal{A}$ die Funktion die jeder Aussage ihre negierte Aussage zuordnet, aber $\neg : W \to W$ ist die Funktion für die gilt $\neg 0 = 1$ und $\neg 1 = 0$. Eine Verwechslung kann hier nicht auftreten, da man an den Argumenten stets erkennt, welcher Zusammenhang gemeint ist. Vorteil der Schreibweise ist die Eleganz. So ergibt sich im Fall der entweder/oder Funktion

$$\mathrm{XOR}(x, y) = (x \wedge (\neg y)) \vee ((\neg x) \wedge y)$$

Außerdem gelten die Zusammenhänge

$$E(\neg A) = \neg E(A),\ E(A \wedge B) = E(A) \wedge E(B),\ E(A \vee B) = E(A) \vee E(B)$$

aufgrund derer man sogar die Aussage $\mathcal{A}$ mit ihrem Wahrheitsgehalt $E(A)$ identifizieren könnte (was wir aber nicht tun, da es zu Verständnisproblemen führen kann).

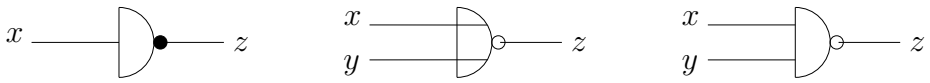
Zum Nachweis der Gleichheit zweier Funktionen mit endlicher Definitionsmenge vergleicht man übrigens einfach alle Einträge in der Wertetabelle.
So gilt zum Beispiel

$$\neg(x \vee y) = (\neg x) \wedge (\neg y)$$

denn

(x, y)	(0,0)	(0,1)	(1,0)	(1,1)
$\neg(x \vee y)$	1	0	0	0
$(\neg x) \wedge (\neg y)$	1	0	0	0

In der Digitaltechnik spielen die Logikfunktionen eine fundamentale Rolle und werden durch spezielle Symbole repräsentiert, z. B.



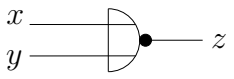
(a) NOT Gatter (b) OR Gatter (c) AND Gatter

Abbildung 1: Elementare Schaltbilder der Digitaltechnik

Häufig werden allgemeine Logikfunktionen aus wenigen Grundfunktionen aufgebaut. So kann z. B. aus der Funktion $\mathrm{NOR}(x, y) = \neg(x \vee y)$ jede andere Logikfunktion aufgebaut werden. Beispielsweise gilt der Zusammenhang $\mathrm{NOT}(x) = \mathrm{NOR}(x, x)$ und damit dann

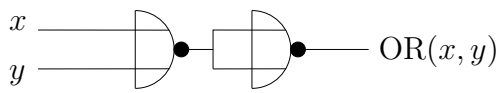
$$\mathrm{OR}(x, y) = \mathrm{NOT}(\mathrm{NOR}(x, y)) = \mathrm{NOR}(\mathrm{NOR}(x, y), \mathrm{NOR}(x, y))$$

Benutzt man das Symbol der NOR Funktion

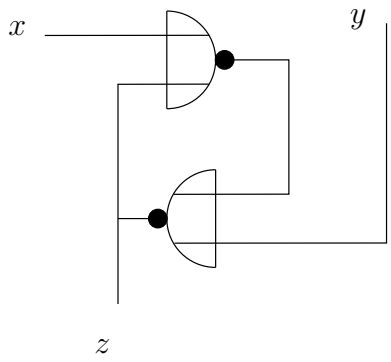


(x,y)	(0,0)	(0,1)	(1,0)	(1,1)
NOR(x,y)	1	0	0	0

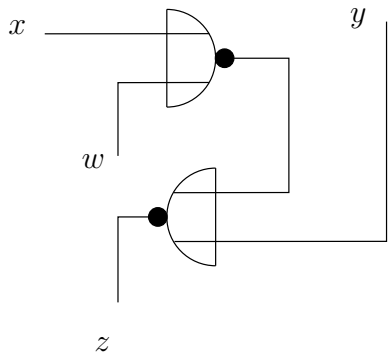
so kann man die OR-Funktion also durch folgende Schaltung darstellen



Zum Abschluß betrachten wir den interessanten Fall, daß der Ausgang einer Schaltung gleichzeitig als Eingang in das System zurückgeführt wird, etwa



Welchen Wert kann man abhängig von den Eingängen x, y am Ausgang z erwarten? Um die mathematische Struktur dieses Problems deutlich zu machen, definieren wir zunächst die Hilfsfunktion $f(x, y, w) = \mathrm{NOR}(\mathrm{NOR}(x, w), y)$ mit dem Schaltbild



Die ursprüngliche Schaltung entspricht dann offensichtlich der *Gleichung*

$$f(x, y, z) = z.$$

Die Lösungen dieser Gleichung bestimmt man einfach dadurch, daß man alle Funktionswerte von f ermittelt und dann jeweils die Bedingung $f(x, y, z) = z$ überprüft. Die (gedrehte) Wertetabelle lautet

x	y	w	$f(x,y,w)$
0	**0**	**0**	0
0	**0**	**1**	1
0	**1**	**0**	0
0	1	1	0
1	0	0	1
1	**0**	**1**	1
1	**1**	**0**	0
1	1	1	0

Offensichtlich entsprechen die durch Fettdruck hervorgehobenen Tripel genau den Lösungen der Gleichung $f(x,y,z) = z$. Was aber bedeuten die Lösungen der Gleichung für die elektrische Schaltung?
Nehmen wir zunächst an, daß sich das System in einem Zustand befindet, der die Gleichung *nicht* löst, also etwa $(x,y,z) = (0,1,1)$. Der Funktionswert $f(0,1,1) = 0$ sagt aus, daß das System dann im nächsten Moment bereits den Ausgang auf $z = 0$ ändert, d. h. der Zustand $(0,1,1)$ kann nicht bestehen bleiben, er ist *instabil*. Wir werden also einen Zustandswechsel nach $(0,1,0)$ beobachten und da $f(0,1,0) = 0$ gilt, tritt nun *kein* weiterer Zustandswechsel mehr auf; d. h. das System ist *stabil*.
Lösungen der Gleichung entsprechen also genau den stabilen Zuständen des Systems. Unsere Schaltung ist übrigens ein Flip-Flop. Aus der Wertetabelle kann man das Verhalten ablesen:Wir starten aus dem Zustand $(0,0,0)$; geht der Eingang x irgendwann einmal auf 1, so wird der Ausgang ebenfalls $z = 1$. Eine weitere Änderung von x auf 0 bzw. zurück auf 1 ändert den Ausgang dann nicht mehr. Erst wenn der Eingang y einmal auf 1 geht, wird der Ausgang gelöscht und bleibt 0, falls $x = 0$ ist, egal wie sich y danach ändert.
Das Flip-Flop kann also eine kurze 1-Meldung auf x speichern und durch „Betätigung“ von y wieder gelöscht werden. Sehen Sie, daß Mathematik hilft, Zusammenhänge vorherzusagen?

2. Abzählbar unendliche Mengen

Der Begriff „unendlich“ ist letztlich eine Verallgemeinerung des Prinzips „ein Bier geht immer noch“ (natürlich ist dieses Prinzip praktisch nicht durchführbar, aber doch hinreichend lange in menschlichen Maßstäben, so daß es unserem Verstand plausibel erscheint). Das Konzept der natürlichen Zahlen beruht auf dem gleichen Prinzip

- 1 ist eine natürliche Zahl

- zu jeder natürlichen Zahl gibt es eine nachfolgende natürliche Zahl.

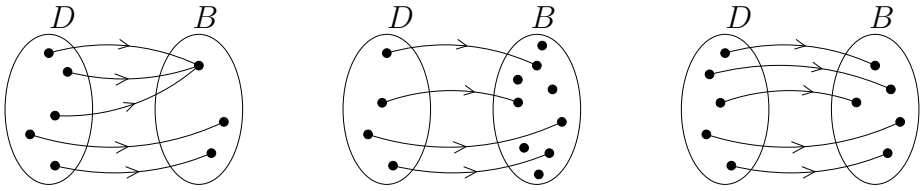
Selbstverständlich existieren diese Zahlen nicht in dem Sinne, daß man sie alle aufschreiben oder sonstwie kodieren könnte. Mit endlich vielen Elementarteilchen im Universum stößt man dabei irgendwann auf eine natürliche Grenze. Da die menschliche Abstraktion aber offensichtlich die Vernachlässigung solcher „nebensächlichen Details“ unterstützt, erscheint uns das Konzept der natürlichen Zahlen trotzdem sinnvoll – und tatsächlich ist es an vielen Stellen sehr hilfreich.
Wir bezeichnen die Menge aller natürlichen Zahlen mit $\mathbb{N}$. Jede Menge M, deren Durchnummerierung alle natürlichen Zahlen benötigt, bezeichnen wir als *abzählbar unendlich*. Durchnummerierung bedeutet dabei, daß es eine Funktion $Z : \mathbb{N} \to M$ gibt mit der Eigenschaft, daß jedes Element von M eine Nummer bekommt, also

$$(S) \text{ für jedes } m \in M \text{ gibt es } n \in \mathbb{N} \text{ mit } m = Z(n)$$

und kein Element von M mehrere Nummern hat, also

$$(I)\ n_1 \neq n_2 \text{ impliziert } Z(n_1) \neq Z(n_2).$$

Die Eigenschaft (S) nennt man Surjektivität der Funktion Z und die Eigenschaft (I) Injektivität.
Eine Funktion, die beide Eigenschaften erfüllt, nennt man auch bijektiv.



(a) Surjektivität (b) Injektivität (c) Bijektivität

Abbildung 2: Abbildungseigenschaften

Beispiele: $\mathbb{N}_0 = \{0\} \cup \mathbb{N}$ ist abzählbar unendlich

$$\begin{aligned} Z &: \mathbb{N} \to \mathbb{N}_0 \\ &\quad n \to n-1 \end{aligned}$$

Die ganzen Zahlen $\mathbb{Z} = \mathbb{N}_0 \cup \{-n | n \in \mathbb{N}\}$ sind abzählbar unendlich. Hier ist eine Abbildung aber schon komplizierter, z. B.

$$0, -1, +1, -2, +2, -3, +3, \ldots \qquad Z(n) = \begin{cases} \frac{n-1}{2} & n \text{ ungerade} \\ -\frac{n}{2} & n \text{ gerade} \end{cases}$$

Die Bruchzahlen $\mathbb{Q} = \{\frac{p}{q} | p \in \mathbb{Z}, q \in \mathbb{N}\}$ sind ebenfalls abzählbar unendlich. Die Idee ist hierbei, die Bruchzahlen folgendermaßen zu durchlaufen

$$\begin{array}{ccccccccccccccc}
\ldots & & -3 & & -2 & \leftarrow & -1 & & 0 & \rightarrow & 1 & & 2 & \rightarrow & 3 & \ldots \\
 & & & & \downarrow & & \uparrow & & & & \downarrow & & \uparrow & & \downarrow & \\
\ldots & & -\frac{3}{2} & & -\frac{\not 2}{2} & & -\frac{1}{2} & \leftarrow & \frac{\not 0}{2} & \leftarrow & \frac{1}{2} & & \frac{\not 2}{2} & & \frac{3}{2} & \ldots \\
 & & & & \downarrow & & & & & & & & \uparrow & & \downarrow & \\
\ldots & & -\frac{3}{3} & & -\frac{2}{3} & \rightarrow & -\frac{1}{3} & \rightarrow & \frac{\not 0}{3} & \rightarrow & \frac{1}{3} & \rightarrow & \frac{2}{3} & & \frac{3}{3} & \ldots \\
 & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots &
\end{array}$$

und nur den Brüchen eine Nummer zuzuweisen, die noch nicht (in gekürzter Form) aufgetreten sind (z. B. erhält $-2/3$ die Nummer 7, obwohl es die neunte Zahl auf dem Weg ist, denn $0/2 = 0$ und $-2/2 = -1$ wurden auf dem Weg schon einmal erfaßt).
Funktionen, deren Definitionsmenge abzählbar unendlich ist, nennen wir auch **Folgen**. Hier ein Beispiel:
Sie haben ein Sparkonto eröffnet mit dem Startkapitel S und dem jährlichen Zinssatz p z. B. $p = 3\% = 0,03$. Wenn Sie wissen wollen, welches Kapital sich nach n Jahren auf dem Konto befindet, interessieren Sie sich für die Folge

$$\mathcal{K} : \mathbb{N}_0 \to M$$

wobei M die Menge aller möglichen Kontostände ist, die der Beziehung

$$\mathcal{K}(0) = S, \quad \mathcal{K}(n) = \mathcal{K}(n-1) + p \cdot \mathcal{K}(n-1), \quad n \geq 1$$

genügt (streng genommen, muß das Ergebnis noch kaufmännisch gerundet werden, aber das lassen wir hier mal außer Acht).
Bei Folgen schreibt man auch oft $\mathcal{K}_n$ statt $\mathcal{K}(n)$, nur um Klammern zu sparen.
Die Kontostandfolge ist übrigens eine *rekursive Folge*: um den Wert an der Stelle n ausrechnen zu können, benötigen Sie die Werte der Folge an vorherigen Stellen $n-1, n-2$ etc. Mit anderen Worten, die Vorhersagen der Werte von rekursiven Folgen ist nicht ganz einfach!

$$\begin{array}{lrcl}
\text{Beispiel:} & S = 1 \text{ EURO} & , & \text{p} = 3\% \\
 & \mathcal{K}(10) & = & \text{1,34 EURO} \\
 & \mathcal{K}(100) & = & \text{19,21 EURO} \\
 & \mathcal{K}(500) & \approx & \text{2,6 Mio. EURO}
\end{array}$$

Das rasante Anwachsen des Kapitals nennt man *exponentielles* Wachstum. Die exponentielle Abhängigkeit sieht man gut in der folgenden Darstellung der Kapitalfunktion

$$\mathcal{K}(n) = S(1+p)^n, \quad n \in \mathbb{N}_0;$$

die wir jetzt nachweisen wollen. Offensichtlich wird hier die Gleichheit zweier Funktionen postuliert, nämlich $\mathcal{K} : \mathbb{N}_0 \to M$ und $E : \mathbb{N}_0 \to M, E(n) = S(1+p)^n$.
Im Fall von Funktionen mit endlicher Definitionsmenge haben wir die Gleichheit durch Vergleich der Wertetabellen überprüft. Das geht im Fall von abzählbar unendlichen Definitionsmengen natürlich nicht mehr. Hier hilft statt dessen der „Dominoeffekt“: Man zeigt, daß die Gleichheit der Funktionswerte für ein beliebiges Element der Definitionsmenge gilt, falls sie für den (bzw. die) Vorgänger gilt (das entspricht dem Aufstellen von Dominosteinen in einem Abstand, daß der Vorgänger umfällt).
Kann man nun die Gleichheit für ein bestimmtes Element der Definitionsmenge nachweisen, so gilt sie automatisch für *alle* Nachfolger (dieser Schritt entspricht dem Anstoßen eines Dominostein, der alle nachfolgenden Stein umfallen läßt.)
Der Dominoeffekt wird in der Mathematik *Induktionsprinzip* genannt, das Überprüfen eines speziellen Elements heißt *Induktionsanfang* und der Nachweis, daß eine Aussage gilt, wenn die vorherige Aussage gilt, heißt *Induktionsschritt*.
Führen wir diesen Prozeß einmal an unserem Beispiel durch. Zunächst wissen wir, daß $\mathcal{K}(0) = S$ gilt und außerdem ist $E(0) = S(1+p)^0 = S$, d. h. der Induktionsanfang ist schon erledigt für das spezielle Element $n = 0$ der Definitionsmenge. Im Induktionsschritt müssen wir zeigen, daß $\mathcal{K}(n) = E(n)$ gilt, falls $\mathcal{K}(n-1) = E(n-1)$ richtig ist. Durch Ausnutzung der Definition der Kapitalfunktionen wissen wir

$$\mathcal{K}(n) = \mathcal{K}(n-1) + p\mathcal{K}(n-1) = (1+p)\mathcal{K}(n-1)$$

Nun nutzen wir die Annahme, daß der Vorgänger „umfällt“, d. h. daß $\mathcal{K}(n-1) = E(n-1)$ gilt. Dies liefert zusammen mit der Definition von E

$$\mathcal{K}(n) = (1+p)E(n-1) = (1+p)S(1+p)^{n-1} = S(1+p)^n = E(n)$$

also genau das, was wir zeigen wollten. Damit gilt die Gleichheit der Funktionswerte für *alle* Elemente von $\mathbb{N}_0$. Die Berechnung der n-ten Potenz $(1+p)^n$ erfordert übrigens *nicht* $(n-1)$ Multiplikationen, es geht auch wesentlich schneller! Der Trick besteht darin, eine schnellere Rekursion zu benutzen, nämlich

$$x^n = \begin{cases} x^{n/2} \cdot x^{n/2} & n \text{ gerade} \\ x^{\frac{n-1}{2}} \cdot x^{\frac{n-1}{2}} \cdot x & n \text{ ungerade} \end{cases}$$

Damit berechnet sich $(1+p)^{1024}$ mit nur 10 statt 1023 Multiplikationen und bei größeren Exponenten wird der Unterschied noch deutlicher.
Als weiteres Beispiel für das Induktionsprinzip betrachten wir folgende Aussage

$$1 + 2 + \ldots + n = \frac{n(n+1)}{2}, \quad n \in \mathbb{N} \tag{1}$$

Auch hier wird wieder die Gleichheit von zwei Funktionen postuliert, nämlich der Funktion $G : \mathbb{N} \to \mathbb{N}, G(n) = n(n+1)/2$ und der Funktion $F : \mathbb{N} \to \mathbb{N}$

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

Das Summensymbol $\sum$ hilft hierbei, die Punkte zu vermeiden und erlaubt damit eine kompaktere und klarere Schreibweise. Allgemein bedeutet für $m \leq n,$ mit $m, n \in \mathbb{N}$

$$\sum_{k=m}^{n} a(k) = a(m) + a(m+1) + \ldots + a(n)$$

wobei a irgendeine Funktion auf $\{m, \ldots, n\}$ sein kann (in unserem Fall ist $a : \mathbb{N} \to \mathbb{N}, a(k) = k$).
Im Fall $n < m$ ist die Summe „leer“, d. h. wir definieren

$$\sum_{k=m}^{n} a(k) = 0 \qquad n < m.$$

Nun aber zurück zu dem Nachweis von (1). Der Induktionsanfang ist wieder einfach, denn

$$F(1) = \sum_{k=1}^{1} k = 1, \qquad G(1) = \frac{1 \cdot 2}{2} = 1$$

Im Induktionsschritt nehmen wir an, daß $F(n-1) = G(n-1)$ gilt und müssen zeigen, daß dann auch $F(n) = G(n)$ stimmt.

$$F(n) = \sum_{k=1}^{n} k = \left(\sum_{k=1}^{n-1} k \right) + n = F(n-1) + n$$

Nutzen wir jetzt $F(n-1) = G(n-1) = (n-1)n/2$

$$F(n) = \frac{(n-1)n}{2} + n = \frac{(n-1)n + 2n}{2} = \frac{(n+1)n}{2} = G(n).$$

und damit gilt $F(n) = G(n)$ für alle $n \in \mathbb{N}$. Das Induktionsprinzip läßt sich auch auf andere Aussagen anwenden, z. B. auf Ungleichungen zwischen Funktionen wie die Bernoulli-Ungleichung (für $p \geq -1$)

$$(1+p)^n \geq 1 + np, \quad n \in \mathbb{N}.$$

Überprüfen wir die Aussage zunächst für $n = 1$:

$$(1+p)^1 = 1 + p = 1 + 1 \cdot p$$

Hier tritt sogar Gleichheit der beiden Ausdrücke auf und da Gleichheit in $\geq$ enthalten ist, stimmt die Aussage für $n = 1$. Im Induktionsschritt nehmen wir nun wieder an, daß $(1+p)^{n-1} \geq 1 + (n-1)p$ stimmt für ein beliebiges $n \leq 2$. Es muß nun gezeigt werden, daß dann auch $(1+p)^n \geq 1 + np$ gilt (d. h. wenn der Stein $(n-1)$ fällt, dann auch der Stein n). Zunächst formen wir so um, daß ein Ausdruck $(1+p)^{n-1}$ erscheint, über den wir ja was wissen

$$(1+p)^n = (1+p)^{n-1}(1+p)$$

Da nun $p \geq -1$ gilt, ist $(1+p) \geq 0$. Deshalb dreht sich das $\geq$ Zeichen im Zusammenhang $(1+p)^{n-1} \geq 1 + (n-1)p$ nicht um, wenn wir mit $(1+p)$ multiplizieren. Es gilt also

$$(1+p)^n \geq (1 + (n-1)p)(1+p) = 1 + np + (n-1)p^2$$

Da $(n-1)p^2$ nicht kleiner als Null sein kann, können wir weiter abschätzen

$$(1+p)^n \geq 1 + np$$

und damit ist der Induktionsschritt erfolgreich beendet.

3. Kontinuierliche Mengen - eindimensional

Betrachten wir als Beispiel die Menge aller Zeitpunkte. Wählen wir als Zeiteinheit 1 Tag, so helfen uns die ganzen Zahlen, Zeitabstände in Vielfachen dieser Einheit anzugeben (etwa in 5 Tagen wird durch 5 repräsentiert und vor 3 Tagen durch -3). Kleinere Zeiträume kann man durch rationale Zahlen (Bruchzahlen) charakterisieren (also $\frac{1}{2}$ steht für $\frac{1}{2}$ Tag oder $\frac{1}{24}$ für $\frac{1}{24}$ Tag, was einer Stunde entspricht etc.). Theoretisch kann man mit $\mathbb{Q}$ beliebig kleine Zeitschnipsel beschreiben, z. B. 1/100000 Tag.
Wir könnten also, nach Festlegung einer Zeiteinheit und einer Referenzzeit die Menge der Zeitpunkte mit $\mathbb{Q}$ identifizieren gemäß:

$$q \in \mathbb{Q} \longleftrightarrow q$$

Zeiteinheiten entfernt von der Referenzzeit und es ließen sich so beliebig lange und auch beliebig kurze Zeitabschnitte beschreiben. Ist die Menge Q also ein gutes Modell zur Beschreibung des Kontinuums Zeit?
Überraschenderweise lautet die Antwort: Nein! Stellen wir uns vor, wir führen das gleiche Fallexperiment durch wie Galilei im 16. Jahrhundert in Pisa. Wie erstellen wir (mit etwas Idealismus beim Auswerten der Messungen) fest: Das Verhältnis der zurückgelegten Strecken entspricht dem Quadrat der dabei verstrichenen Zeiten. Beschreibt $s(t)$ die zurückgelegte Strecke zur Zeit $t > 0$, so gilt also

$$\frac{s(t_2)}{s(t_1)} = \left(\frac{t_2}{t_1}\right)^2.$$

Stellen wir jetzt die natürliche Frage: Wann ist der Fallkörper (z. B. eine Kugel) doppelt so weit vom Ausgangspunkt entfernt wie zum Zeitpunkt t_1?
Zur Beantwortung dieser Frage müssen wir also t_2 suchen, so daß $s(t_2) = 2s(t_1)$ gilt. Nehmen wir an, daß Zeitpunkte durch rationale Zahlen beschrieben werden, so ergibt sich das Problem, eine rationale Zahl $r = t_2/t_1$ zu finden, für die $r^2 = 2$ gilt. Bekanntlich existiert eine solche Zahl nicht, d. h. der Fallkörper erreicht die doppelte Entfernung *nie*! Die rationalen Zahlen genügen also *nicht* unserer Vorstellung von einer kontinuierlich ablaufenden Zeit.
Erst wenn man die offensichtlichen Lücken zwischen den rationalen Zahlen stopft, kommt man zum Modell eines Kontinuums – den *reellen* Zahlen $\mathbb{R}$.
In diesem Modell existiert, per Konstruktion, die Lösung der Gleichung $r^2 = 2$. Wir bezeichnen sie mit $r = \sqrt{2}$.
Übrigens, für die, die noch nicht wissen, daß die Gleichung $r^2 = 2$ in $\mathbb{Q}$ nicht lösbar ist, hier der kurze Beweis. Es handelt sich dabei um einen Widerspruchsbeweis, d. h. wir nehmen an, daß die Gleichung lösbar sei und folgern daraus eine unsinnige Aussage. Da eine richtige Annahme zusammen mit korrekten Folgerungen nur auf wahre Aussagen führen kann, zeigt sich damit, daß unsere Annahme, $r^2 = 2$ sei lösbar in $\mathbb{Q}$, falsch sein muß. Soviel zur Strategie. Nehmen wir also an, es gibt ein $r \in \mathbb{Q}$ mit der Eigenschaft $r^2 = 2$. Da jede Zahl in $\mathbb{Q}$ als Bruch geschrieben werden kann, haben wir $r = p/q$ ist und dieser Bruch soll *vollständig gekürzt* sein, d. h. p und q haben keine gemeinsamen Primfaktoren. Dann gilt zunächst $p^2 = 2q^2$, woraus wir erkennen, daß p^2 gerade ist.

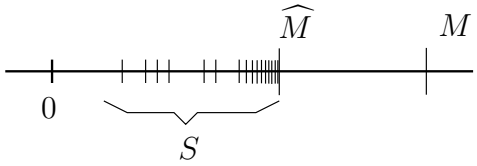
Da das Quadrat einer ungeraden Zahl immer ungerade ist

$$(2m+1)^2 = 4m^2 + 4m + 1 \qquad m \in \mathbb{Z}$$

kann somit p nicht ungerade sein, d. h. es gibt ein $k \in \mathbb{N}$ mit $p = 2k$. Dann ist aber $4k^2 = 2q^2$ bzw. $q^2 = 2k^2$ und somit ist q ebenfalls gerade, d. h. $q^2 = 2j$ mit $j \in \mathbb{N}$. Dies ist aber jetzt wirklich Unsinn, denn es kann ja nicht gleichzeitig $p = 2k$ und $q = 2j$ gelten *und* p und q keine gemeinsamen Faktoren haben. Irgendetwas ist also faul hier. Da alle unsere Schlußfolgerungen korrekt waren, kann die faule Stelle nur in der (ungesicherten) Annahme liegen, die damit als falsch entlarvt ist.
Jetzt aber zurück zum Modell der reellen Zahlen. Was sind eigentlich reelle Zahlen? Nun, reelle Zahlen sind Objekte, die bestimmten Regeln und Gesetzen gehorchen und da diese Regeln die (aus $\mathbb{Q}$) bekannten Rechenregeln umfassen, bezeichnet man die Objekte eben als Zahlen. Im einzelnen bestehen die Eigenschaften der reellen Zahlen aus

(1) Rechenregeln für Grundrechenarten „Addition“ und „Multiplikation“ (d. h. Kommutativ-, Assoziativ- und Distributivgesetze etc. genau wie in $\mathbb{Q}$),
(2) Existenz einer $\leq$ Relation und deren Zusammenspiel mit Addition und Multiplikation (genau wie in $\mathbb{Q}$)
(3) Kontinuum-Eigenschaft (Unterschied zu $\mathbb{Q}$!)

Die Kontinuum-Eigenschaft, die uns hier besonders interessiert, besagt dabei, daß eine beschränkte Teilmenge von $\mathbb{R}$ stets von reellen Zahlen (und nicht von Lücken!) begrenzt ist.
Etwas genauer formuliert man die Eigenschaft der reellen Zahlen so: Ist $S \neq \emptyset$ eine Teilmenge von $\mathbb{R}$ und hat S eine obere Schranke, d. h. gibt es ein $M \in R$, so daß $s \leq M$ für alle $s \in S$, so gibt es auch eine *kleinste* obere Schranke. $\widehat{M} \in \mathbb{R}$ von S.



Diese kleinste obere Schranke ist sozusagen der Grenzstein der Menge S, denn es gilt ja $s \leq \widehat{M}$ für alle $s \in S$, da $\widehat{M}$ eine obere Schranke ist und außerdem gibt es keine kleinere Zahl für die diese Eigenschaft noch gilt – daher Grenzstein. Der Grenzstein selbst muß übrigens nicht zu der Menge S dazugehören.
Zum Beispiel ist der obere Grenzstein der Menge $S = \{x \in \mathbb{R} | 0 < x < 1\}$ offensichtlich die Zahl 1 aber $1 \notin S$. Als Name für den oberen

Grenzstein einer Menge S wird das Symbol $\sup S$ benutzt (sup steht dabei für Supremum). Analog gibt es einen unteren Grenzstein, wenn S nach unten beschränkt ist. Diese Zahl wird mit $\inf S$ (für Infimum) bezeichnet.
Nochmal zurück zu Galilei: Was hat $\inf, \sup$ jetzt mit der fallenden Kugel zu tun?
Betrachten wir die Menge aller Zeitpunkte für die die Kugel die Marke $2s(t_1)$ noch nicht erreicht hat

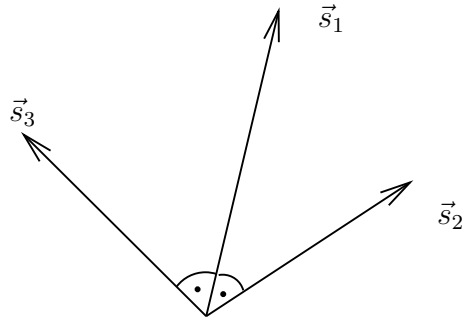
$$T = \{t \in \mathbb{R} | s(t) < 2s(t_1)\}.$$

Diese ist offensichtlich nicht leer, denn zum Beispiel gilt $t_1 \in T$. Um zu zeigen, daß T nach oben beschränkt ist, müssen wir (entsprechend der Definition) eine Zahl M finden, so daß $t \leq M$ für alle $t \in T$ gilt.
Im vorliegenden Fall gibt es sicherlich ein solches M, denn irgendwann einmal wird die Kugel an der Marke $2s(t_1)$ vorbeigeflogen sein (hier können Sie sehr großzügig sein, z. B. „nach 10.0000 Jahren“) und da sie nicht von alleine zurückkehrt, gilt mit diesem großen M sicherlich $t \leq M$ für alle $t \in T$.
Dann gibt es aber nach der Kontinuum-Eigenschaft auch einen Grenzstein von T, nämlich die reelle Zahl $\sup T$. Als Grenzstein hat der Zeitpunkt $\sup T$ die Eigenschaft, daß $t \leq \sup T$ für alle Zeitpunkte, an denen die Kugel die Marke $2s(t_1)$ noch nicht erreicht hat (d. h. für alle $t \in T$) und $\sup T$ ist der früheste Zeitpunkt mit dieser Eigenschaft.
Mit anderen Worten, $\sup T$ entspricht unserer Vorstellung nach genau dem Zeitpunkt, an dem die Kugel die doppelte Entfernung $2s(t_1)$ erreicht. Während die Kontinuums-Eigenschaft der reellen Zahlen garantiert, daß der Grenzstein $\sup T$ der Menge T existiert, gilt dies eben für die rationalen Zahlen nicht. In den rationalen Zahlen gibt es beschränkte Mengen, die keinen rationalen Grenzstein besitzen.
Zusammenfassend stellen wir fest, daß die reellen Zahlen ein sinnvolles Modell für Kontinua darstellen, also Längen, Zeiten, Temperaturen, Drücke, Spannungen, Ströme, etc.

4. Kontinuierliche Mengen – mehrdimensional

Das Modell der reellen Zahlen beinhaltet eine $\leq$ Relation, d. h. für zwei Zahlen $x, y \in \mathbb{R}$ gilt stets $x \leq y$ oder $y \leq x$. Damit eignet sich $\mathbb{R}$ für die Beschreibung der Zeit ($\leq$ bedeutet hier früher als) oder zur Beschreibung von Punktposition auf einer Geraden ($\leq$ bedeutet hier links von). Für die Beschreibung von Punktpositionen im Raum ist die Menge $\mathbb{R}$ alleine aber ungeeignet.

Jeder Schatzsucher weiß, daß hier mehrere Zahlen zur genauen Ortsbeschreibung notwendig sind: starte bei der Palme am Strand, gehe 100 Schritte nach Norden, dann 25 Schritte nach Westen und dann findest Du in 5 Fuß Tiefe eine Kiste.
Wir werden also Positionen im uns umgebenden Raum mit drei Zahlen beschreiben, und da uns der Raum kontinuierlich erscheint, werden wir drei reelle Zahlen benutzen. Die Zahlen müssen dabei als Entfernung in vorgegebenen festen Richtungen von einem bestimmten Punkt aus interpretiert werden.
In unserem obigen Beispiel waren zwei der Richtungen durch Kompaßmarkierungen gegeben und die dritte Richtung durch die Gravitationskraft. Streng genommen sind diese Richtungen zwar nicht vollständig ortsunabhängig, aber auf einer kleinen Insel fällt diese Ungenauigkeit nicht ins Gewicht.
Da wir unsere Raumbeschreibung unabhängig von der Erde (Magnet- und Gravitationsfeld) durchführen wollen, nehmen wir an, wir hätten spezielle Zeiger endlicher Länge, die an einem Ende spitz sind und die die Eigenschaft haben, daß sie im Raum immer in dieselbe Richtung zeigen, also nur verschoben, aber nicht gedreht werden können (genau wie die Kompaßnadel oder die Achsen von schnell rotierenden Kreiseln, die ja auch im sogenannten Kreiselkompaß benutzt werden). Wir wählen dann drei solcher Zeiger gleicher Länge aus (unsere Längeneinheit), die paarweise senkrecht aufeinander stehen und nennen sie z. B. $\vec{s}_1, \vec{s}_2, \vec{s}_3$ (der Pfeil deutet die Orientierung der Zeiger vom stumpfen zum spitzen Ende an). Dabei ist die Reihenfolge der Nummerierung der Zeiger so gewählt, daß die Anordnung der Konstellation Daumen (1), Zeigefinger (2) und Mittelfinger (3) einer rechten Hand entspricht, die gerade eine Pistole simuliert.

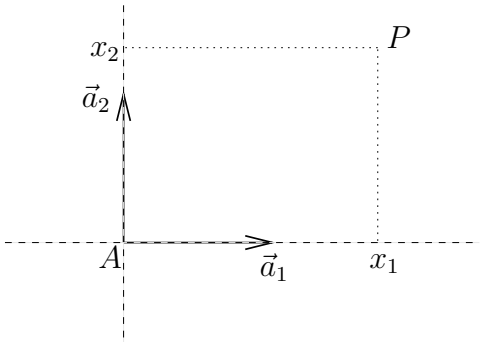


Jetzt brauchen wir noch einen Punkt A des Raumes zu markieren und los geht's: Starte in A und nimm' die Zeiger mit; gehe die Entfernung

$|x_1|$ in die Richtung von Zeiger $\vec{s}_1$ (bzw. in die entgegengesetzte Richtung, falls $x_1 < 0$); gehe von diesem Punkt dann $|x_2|$ in Richtung von $\vec{s}_2$ (bzw. entgegengesetzt, wenn $x_2 < 0$) und wiederhole die Prozedur vom jetzt erreichten Zwischenpunkt mit der Entfernung $|x_3|$ und der Richtung $\vec{s}_3$.
Der Endpunkt P ist in diesem Sinne eindeutig durch die drei Zahlen $x_1, x_2, x_3 \in \mathbb{R}$ charakterisiert, den sogenannten Koordinaten des Punktes P im (kartesischen) Koordinatensystem $(A, \vec{s}_1, \vec{s}_2, \vec{s}_3)$.
Mit unseren drei verallgemeinerten Kompaßnadeln $\vec{s}_1, \vec{s}_2, \vec{s}_3$ und dem Bezugspunkt A können wir also jedem Tripel von Zahlen (x_1, x_2, x_3) einen Punkt im Raum zuordnen. Umgekehrt können wir jeden Punkt im Raum bezüglich seiner Lage zu A vermessen und ihm drei Maßzahlen zuordnen.
Als Modell des physikalischen Raumes erscheint uns also die Menge $\mathbb{R} \times \mathbb{R} \times \mathbb{R}$ aller reellen Zahlentripel geeignet. Wir schreiben für diese Menge auch kurz $\mathbb{R}^3 = \mathbb{R} \times \mathbb{R} \times \mathbb{R}$.
Allgemeiner betrachten wir die n-dimensionalen Räume

$$\mathbb{R}^n = \underbrace{\mathbb{R} \times \mathbb{R} \times \ldots \times \mathbb{R}}_{n-\text{mal}}$$

Insbesondere ist $\mathbb{R}^1 = \mathbb{R}$ die Menge der reellen Zahlen und $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ kann dazu benutzt werden, alle Punkte einer Ebene zu beschreiben. Wir wählen dazu wieder einen Punkt A im Raum, aber jetzt eben nur zwei Zeiger $\vec{a}_1, \vec{a}_2$ gleicher Länge, die senkrecht aufeinander stehen. Dem Zahlenpaar $(x_1, x_2) \in \mathbb{R}^2$ ordnen wir dann den Punkt P des Raumes zu, den man dadurch erreicht, daß man von A die Entfernung $|x_1|$ dem Zeiger $\vec{a}_1$ folgt und dann die Entfernung $|x_2|$ entlang $\vec{a}_2$ zurücklegt (wobei die Richtung von den Vorzeichen von x_1, x_2 bestimmt wird).

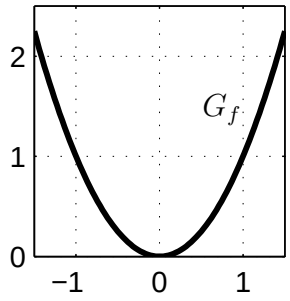


Den Fall $\mathbb{R}^3$ als Modell des Raumes haben wir ja schon besprochen, aber wofür braucht man $\mathbb{R}^4, \mathbb{R}^5, \mathbb{R}^6$? Die geometrische Interpretation

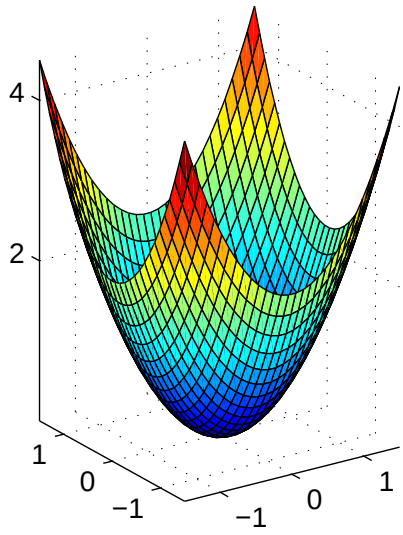
übersteigt hierbei die menschlichen Fähigkeiten. Die Räume $\mathbb{R}^n$ mit $n \geq 4$ werden daher meist als *Zustandsräume* benutzt. Stellen Sie sich vor, sie wollen den Zustand der Luft, die Sie gerade umgibt, charakterisieren. Wieviele Zahlen brauchen Sie dazu? Je nach dem wie genau Sie die Luft beschreiben wollen, fällt die Antwort sicherlich unterschiedlich aus. Beispielsweise können sie die Größen Dichte, Druck und Geschwindigkeit benutzen.
Das sind schon fünf Zahlen, denn die Geschwindigkeit im Raum beschreibt man dadurch, daß man angibt, wie schnell die Bewegung in $\vec{s}_1$-Richtung, in $\vec{s}_2$-Richtung und $\vec{s}_3$-Richtung abläuft, also mit drei Zahlen. Der Raum $\mathbb{R}^5$ kann also dazu benutzt werden, den Zustand der Luft zu beschreiben.
Da die Werte für Dichte, Druck und Geschwindigkeit in jedem Punkt des Raumes prinzipiell verschieden sein können, würde man den Luftzustand insgesamt durch eine Funktion $f : \mathbb{R}^3 \to \mathbb{R}^5$ angeben, wobei z. B. $(f(x_1, x_2, x_3))_1$ die Dichte des Gases am zu (x_1, x_2, x_3) gehörenden Punkt beschreibt, $(f(x_1, x_2, x_3))_2$ den Druck, und f_3, f_4, f_5 die Komponenten der Geschwindigkeit (beachten Sie, daß $f(x_1, x_2, x_3)$ ein Element von $\mathbb{R}^5$ ist und damit fünf Komponenten hat). Nun besteht Luft aber aus einem Gemisch von Gasen wie Stickstoff, Sauerstoff, Kohlendioxid etc. und wenn Sie eine genauere Zustandsbeschreibung wünschen, so brauchen Sie entsprechend mehr Zahlen. Statt einem Wert für die Dichte der Luft kann man mehrere Partialdichten einführen, also die Dichte des Stickstoffs, des Sauerstoffs, des Kohlendioxids und so weiter. Beschränken wir uns auf die drei Gase, so steigt die Dimension des Zustandsraums auf sieben, d. h. das Gas wird durch eine Funktion $g : \mathbb{R}^3 \to \mathbb{R}^7$ beschrieben. Bleiben wir aber zunächst in den niedrig dimensionalen Räumen.
Zur Veranschaulichung einer Funktion $f : \mathbb{R}^n \to \mathbb{R}^m$ kann man, zumindest im Fall $n + m \leq 3$, den Graph G_f der Funktion verwenden. G_f ist eine Teilmenge des $\mathbb{R}^{m+n}$, definiert durch

$$G_f = \{(x_1, \ldots, x_n, y_1, \ldots, y_m) | \\ (y_1, \ldots, y_m) = f(x_1, \ldots, x_n), (x_1, \ldots, x_n) \in \mathbb{R}^n\}$$

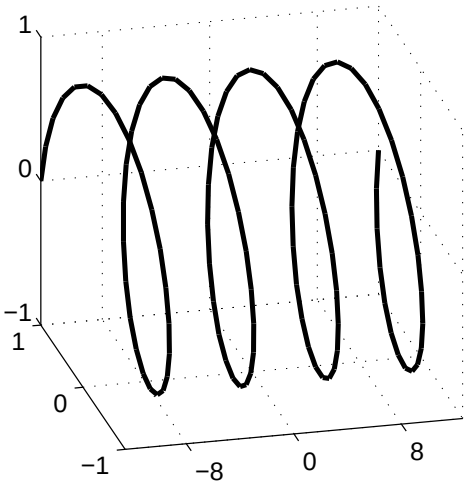
Im Fall $f : \mathbb{R} \to \mathbb{R}$ ist $G_f \subset \mathbb{R}^2$. Identifiziert man $\mathbb{R}^2$ durch Wahl eines Koordinatensystems mit der Zeichenebene, so entsprechen die Elemente von G_f Punkten auf einer Kurve in der Zeichenebene. Der Graph der Funktion $f(x) = x^2, x \in \mathbb{R}$ ist zum Beispiel eine Parabel



Entsprechend kann man den Graph einer Funktion $f : \mathbb{R}^2 \to \mathbb{R}$ als Fläche im Raum veranschaulichen, d. h. markiert man alle Raumpunkte, die zu Tripeln der Form $(x, y, f(x, y))$ gehören, so ist das Gesamtgebilde eine Fläche. Als Beispiel betrachten wir die Funktion $f(x, y) = x^2+y^2$. Der zugehörige Graph hat die Form eines Rotationsparaboloids.



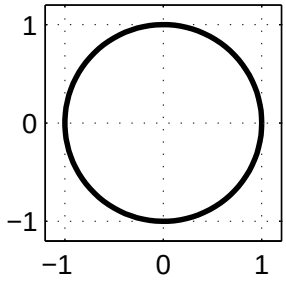
Die dritte Kombination von Definitions- und Bildmenge für die $n+m \leq 3$ gilt, ist durch $f : \mathbb{R} \to \mathbb{R}^2$ gegeben. Hier entspricht dem Graph eine Kurve im Raum. Für den Fall $f(t) = (\cos(2\pi t), \sin(2\pi t))$ ist diese Kurve z. B. eine Schraubenlinie.



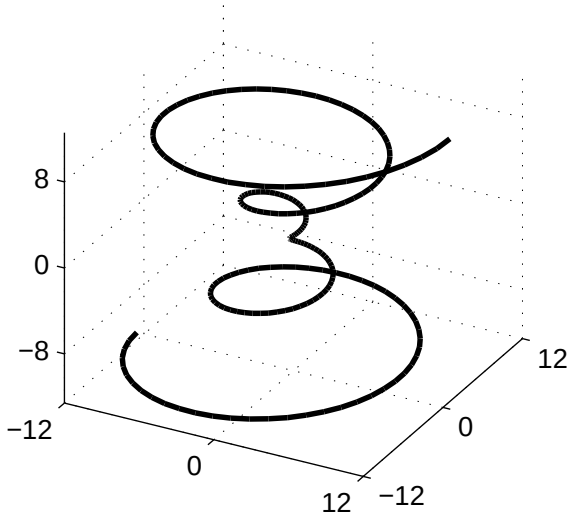
Zeichnet man hingegen nur den *Wertebereich*

$$f(\mathbb{R}) = \{f(x)|x \in \mathbb{R}\} \subset \mathbb{R}^2$$

der Funktion, so ergibt sich eine Kreiskurve in der Zeichenebene



Diese Darstellung charakterisiert die Funktion jedoch nicht vollständig. So hat die Funktion $g : \mathbb{R} \to \mathbb{R}^2, g(t) = (\cos(20\pi t), \sin(20\pi t))$ den gleichen Wertebereich wie f, obwohl es sich offensichtlich um eine andere Funktion handelt. Die Funktionsgraphen G_f, G_g sind dagegen unterschiedlich. Beide Graphen sind Schraubenlinien, aber G_g hat nur ein Zehntel der Ganghöhe des Graphen G_f (d. h. G_g windet sich zehn Mal, wenn sich G_f einmal windet). Trotzdem ist die Darstellung des Wertebereichs hilfreich für das Verständnis der Funktion.
Im Fall von Funktionen $f : \mathbb{R} \to \mathbb{R}^3$ ist die Darstellung des Graphen G_f nicht mehr möglich (Teilmenge von $\mathbb{R}^4$), aber der Wertebereich kann noch als Kurve im Raum dargestellt werden. Als Beispiel nennen wir hier $f(t) = (t \sin t, t \cos t, t)$,



Genauso wie Funktionsgraphen lassen sich auch Lösungsmengen von Gleichungen und Ungleichungen graphisch veranschaulichen. So bildet z. B. die Lösungsmenge einer Gleichung mit zwei Unbekannten typischerweise eine Kurve in der Zeichenebene. Allgemein kann man eine Gleichung mit n Unbekannten immer in der Form $\phi(x_1, \ldots, x_n) = 0$ darstellen, wobei $\phi : \mathbb{R}^n \to \mathbb{R}$ eine gegebene Funktion ist. Die Lösungsmenge ist dann

$$L_0 = \{(x_1, \ldots, x_n) \in \mathbb{R}^n | \phi(x_1, \ldots, x_n) = 0\}$$

Als Beispiel betrachten wir zunächst die Funktion $\phi : \mathbb{R}^2 \to \mathbb{R}$

$$\phi(x, y) = \left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 - 1,$$

wobei $a, b > 0$ feste Parameter sind. Um herauszufinden, wie die Lösungsmenge der Gleichung $\phi(x, y) = 0$ aussieht, können wir jetzt beliebige Paare $(x, y) \in \mathbb{R}^2$ in die Funktion ϕ einsetzen und nachrechnen, ob der Funktionswert Null ist. Ist dies der Fall, so markieren wir den entsprechenden Punkt in der Ebene, wenn nicht, so wählen wir einen anderen Punkt. Mit dieser Vorgehensweise würden wir aber bald enttäuscht aufgeben, da man so fast nie einen Punkt der Lösungsmenge erwischt. Ein besserer Ansatz besteht darin, die *implizite* Gleichung $\phi(x, y) = 0$ zunächst in eine *explizite* Gleichung der Form $y = f(x)$ bzw. $x = g(y)$ umzuformen. Schauen wir uns unser Beispiel an. Ist (x, y) ein Element

der Lösungsmenge, so gilt $\phi(x,y)=0$ und damit

$$y^2 = b^2\Big(1-\Big(\frac{x}{a}\Big)^2\Big) \tag{2}$$

Da immer $y^2 \geq 0$, sehen wir sofort, daß für Elemente der Lösungsmenge $x^2 \leq a^2$ gelten muß. Ziehen wir die Wurzel und beachten $\sqrt{x^2}=|x|$, so folgt zunächst $|x| \leq |a|$. Da $a>0$ vorausgesetzt war, erhalten wir also insgesamt $|x| \leq a$. Ziehen wir nun die Wurzel aus dem Zusammenhang (2), so folgt für jedes Element (x,y) der Lösungsmenge der Zusammenhang

$$|y| = b\sqrt{1-\Big(\frac{x}{a}\Big)^2}, \quad |x| \leq a.$$

Um diese Darstellung in die Form $y=f(x)$ zu bringen, müssen wir nur noch die Betragsstriche loswerden. Dazu erinnern wir zunächst an die Definition des Betrags

$$|y| = \begin{cases} y & y \geq 0 \\ -y & y<0 \end{cases} \tag{3}$$

Um Beträge aufzulösen, muß man daher immer Fallunterscheidungen durchführen! Betrachten wir den ersten Fall $y \geq 0$. Hier gilt einfach $|y|=y$, und damit haben wir ausgerechnet, daß, wenn (x,y) ein Element der Lösungsmenge ist, mit $y \geq 0$, die Beziehung

$$y = b\sqrt{1-\Big(\frac{x}{a}\Big)^2}, \quad |x| \leq a$$

gelten muß.
Entsprechend gilt für (x,y) in der Lösungsmenge mit $y<0$, daß

$$-y = b\sqrt{1-\Big(\frac{x}{a}\Big)^2}, \quad |x| < a$$

erfüllt ist. Definieren wir

$$f_+(x) = b\sqrt{1-\Big(\frac{x}{a}\Big)^2}, \quad f_-(x) = -b\sqrt{1-\Big(\frac{x}{a}\Big)^2}, \quad |x| \leq a$$

so haben wir bisher gezeigt, daß

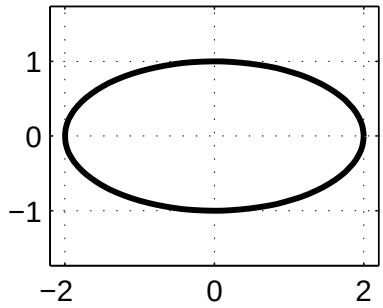
$$L_0 \subset \{(x,y)|y=f_+(x), |x| \leq a\} \cup \{(x,y)|y=f_-(x), |x|<a\}$$

Um die Gleichheit der beiden Mengen nachzuweisen, bleibt noch zu zeigen, daß jedes Element $(x,f_+(x))$ bzw. $(x,f_-(x))$ in der expliziten

Darstellung auch eine Lösung der Gleichung $\phi(x, y)$ ist, d. h. es fehlt noch die umgekehrte Inklusion

$$\{(x, y)|y = f_+(x), |x| \leq a\} \cup \{(x, y)|y = f_-(x), |x| < a\} \subseteq L_0$$

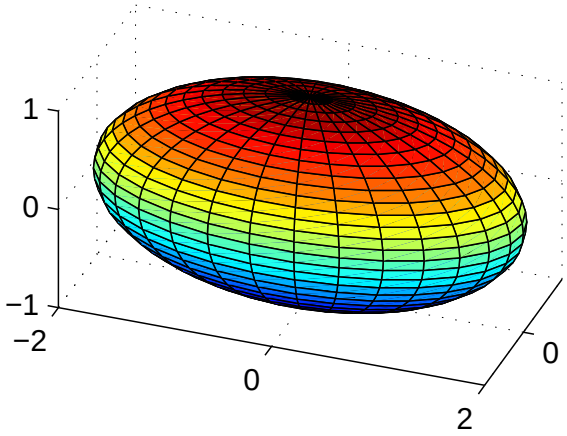
Dies ist aber einfach, man muß nur noch einsetzen und nachrechnen, daß $\phi(x, f_+(x)) = 0$ und $\phi(x, f_-(x)) = 0$ gilt.
Zum Darstellen der Lösungsmenge L_0 können wir also die Graphen der Funktionen f_+, f_- zeichnen. Dazu fertigen wir eine Wertetabelle an, berechnen endlich viele Funktionswerte, tragen die Paare $(x, f_+(x))$ und $(x, f_-(x))$ in der Zeichenebene ein und verbinden danach benachbarte Punkte. Wenn wir das mit hinreichend vielen Punkten x aus der Definitionsmenge $\{x||x| \leq a\}$ machen, so entsprechen die resultierenden Polygonzüge ziemlich genau den Funktionsgraphen (da die Funktionen f_+, f_- stetig sind). Im vorliegenden Fall findet man eine Ellipse mit den Halbachsen a, b.



Beinhaltet die Gleichung drei Unbekannte, wird also durch eine Funktion $\Psi : \mathbb{R}^3 \to \mathbb{R}$ beschrieben, so bildet die veranschaulichte Lösungsmenge von $\Psi(x, y, z) = 0$ typischerweise eine Fläche im Raum. Ist etwa

$$\Psi(x, y, z) = \left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 + \left(\frac{z}{c}\right)^2 - 1$$

mit $a, b, c > 0$, so entspricht die Lösungsmenge der Oberfläche eines Ellipsoids.



Mit den gleichen Schritten wie oben formt man auch hier die implizite Gleichung zunächst in explizit Gleichungen um. Es gilt

$$\begin{aligned}\{(x,y,z)|\Psi(x,y,z)=0\} &= \{(x,y,g_+(x,y))|(x,y)\in E\}\\ &\cup \{(x,y,g_-(x,y))|(x,y)\in E\}\end{aligned}$$

wobei E eine Teilmenge der Form

$$E=\left\{(x,y)\in\mathbb{R}^2|\left(\frac{x}{a}\right)^2+\left(\frac{y}{b}\right)^2-1\leq 0\right\} \tag{4}$$

ist und $g_+ : E\to\mathbb{R}, g_- : E\to\mathbb{R}$ die Form

$$g_+(x,y)=c\sqrt{1-\left(\frac{x}{a}\right)^2-\left(\frac{y}{b}\right)^2},\quad g_-(x,y)=-g_+(x,y)$$

haben. Beachten Sie, daß E eine Lösungsmenge einer Ungleichung ist, d. h. alle Paare (x,y) enthält, die der Ungleichung $\phi(x,y)\leq 0$ gehorchen, wobei ϕ die „Ellipsenfunktion“ ist. Wie löst man nun so eine Ungleichung? Hier hilft eine Eigenschaft der Funktion, die wir später noch genauer untersuchen werden: die *Stetigkeit*. Diese Eigenschaft besagt anschaulich, daß die Funktionswerte der Funktion ϕ sich nicht abrupt ändern, wenn sich die Argumente kontinuierlich ändern. Bei stetigen Funktion ϕ geht man zur Bestimmung der Lösungsmenge einer Ungleichung $\phi(x,y)\leq 0$ folgendermaßen vor. Zunächst berechnet man die Kurve, die zur Lösungsmenge der Gleichung $\phi(x,y)=0$ gehört, was ja in unserem Fall eine Ellipse ergibt. Danach überprüft man in den durch die Kurve separierten Gebieten jeweils in einem Punkt das Vorzeichen der Funktion ϕ. So findet man im Innern der Ellipse z. B. $\phi(0,0)=-1<0$ und außerhalb $\phi(2a,0)=4-1=3>0$. Dann muß

aber auch *überall* im Innern der Ellipse $\phi(x,y) < 0$ gelten und *überall* außerhalb $\phi(x,y) > 0$. Wäre nämlich $\phi(\overline{x},\overline{y}) > 0$ für einen Punkt im Innern der Ellipse, dann würde die Funktion ϕ auf der Verbindungslinie zwischen $(0,0)$ und $(\overline{x},\overline{y})$ mit einem negativen Wert starten und mit einem positiven Wert enden, ohne zwischendurch die Null zu durchlaufen (die Punkte mit $\phi(x,y) = 0$ liegen ja auf der Ellipse und die schneidet die Verbindung zwischen $(0,0)$ und $(\overline{x},\overline{y})$ nicht). Das kann aber nur dann möglich sein, wenn der Wert von ϕ auf der Verbindung plötzlich von negativen auf positive Werte springt, einem Fall, der der Stetigkeit widerspricht. Genauso zeigt man, daß außerhalb der Ellipse $\phi(x,y) > 0$ gilt, wobei man hier nicht immer die direkte Verbindungslinie für das Argument benutzen kann (denken Sie z. B. an die beiden Punkte $(2a,0), (-2a,0)$, deren Verbindungslinie die Ellipse schneidet). Statt dessen wird man eine gekrümmte, aber kontinuierliche (stetige) Verbindung wählen, die die Ellipse nicht berührt und die beiden gewünschten Punkte verbindet. Damit haben wir gezeigt, daß die Menge E in (4) eine elliptische Scheibe in der Zeichenebene beschreibt. Trägt man über bzw. unter jedem Punkt (x,y) dieser Scheibe gerade die beiden Werte $g_+(x,y), g_-(x,y)$ im Raum ein, so ergibt das Gesamtgebilde eben die Ellipsoidfläche.

Zum Abschluß dieses Kapitels betrachten wir noch Gleichungen bzw. Ungleichungen, die Betragsfunktionen beinhalten. Traditionell bereitet die Auflösung solcher Gleichungen Schwierigkeiten und diesen Schwierigkeiten kann man am besten durch *sorgfältiges* Arbeiten begegnen. Diese Sorgfalt ist deshalb nötig, weil die Definition (3) der Betragsfunktion zwei Fälle beinhaltet. Beim Auflösen muß man deshalb für jedes Auftreten der Betragsfunktion zwei Fälle unterscheiden. Beinhaltet eine Gleichung also drei Betragsfunktionen, so müssen bereits $2^3 = 8$ Fälle unterschieden werden, und wenn man hier nicht buchhalterisch (also sorgfältig) vorgeht, so ist einer der Fälle schnell einmal vergessen. Bevor wir uns einige Beispiele anschauen, zunächst noch eine Bemerkung zu Gleichungen bzw. Ungleichungen, die auf der Funktion $\lambda(x,y) = ax + by + c$ beruhen.

Ist $b \neq 0$, so sieht man schnell, daß die veranschaulichte Lösungsmenge von $\lambda(x,y) = 0$ die Form einer Geraden hat, denn jedes Paar (x,y) der Lösungsmenge ist eindeutig durch die explizite Geradengleichung

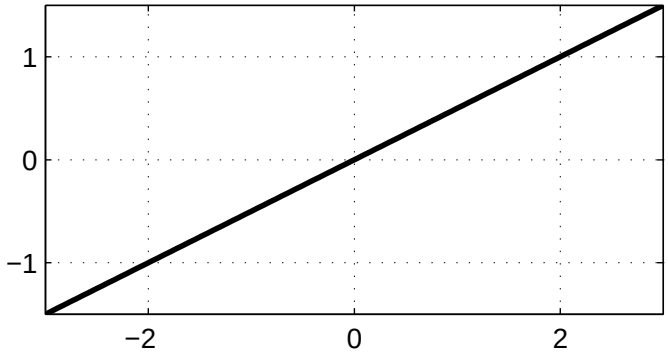
$$y = -\frac{a}{b}x - \frac{c}{b}, \quad x \in \mathbb{R}$$

charakterisiert. Ist dagegen $b = 0$, aber $a \neq 0$, so beschreibt die Lösungsmenge von $\lambda(x,y) = 0$ eine Gerade parallel zur y-Achse. Die Einschränkung an Paare (x,y) der Lösungsmenge ist nämlich nur eine

Einschränkung an die x-Komponente

$$x = -\frac{c}{a}, \quad y \in \mathbb{R}$$

und y kann frei in $\mathbb{R}$ variieren.
Der Fall $a = b = 0$ ist natürlich langweilig. Hier ist die Lösungsmenge der Gleichung $\lambda(x, y) = 0$ entweder leer (wenn $c \neq 0$), oder der ganze Raum $\mathbb{R}^2$ (wenn $c = 0$). Sie sehen aber, daß die Lösungsmenge einer Gleichung mit zwei Unbekannten nicht unbedingt eine Kurve sein muß. Im interessanten Fall $|a| + |b| \neq 0$ ist die Lösungsmenge aber durch eine Gerade in der Zeichenebene gegeben und entsprechend sind die Lösungsmengen der Ungleichungen $\lambda(x, y) < 0, \lambda(x, y) > 0$ Halbräume. Da λ stetig ist, genügt es dazu, das Vorzeichen von λ an einem Punkt ober- bzw. unterhalb der Geraden auszuwerten. Als Beispiel betrachten wir $\lambda(x, y) = x - 2y$. Die Gleichung $\lambda(x, y) = 0$ führt auf die Geradengleichung $y = \frac{1}{2}x$. Die Lösungsmenge entspricht also der um 1 nach oben verschobenen Geraden mit Steigung 1/2.



Auswertung von λ am Punkt $(2, 0)$ liefert $\lambda(2, 0) = 2 > 0$, d. h. die Ungleichung $\lambda(x, y) > 0$ liefert als Lösungsmenge den Halbraum unterhalb der Geraden. Entsprechend findet man den oberen Halbraum als Lösungsmenge der Ungleichung $\lambda(x, y) < 0$. Schauen wir uns jetzt aber einmal eine Gleichung $\alpha(x, y) = 0$ mit einer Betragsfunktion an.

$$\alpha(x, y) = |x - y| - 1, \quad x, y \in \mathbb{R}.$$

Eine Betragsfunktion bedeutet, daß zwei Fälle zu unterscheiden sind, nämlich hier $x - y \geq 0$ (Fall 1) und $x - y < 0$ (Fall 2). Beachten Sie, daß die beiden Fälle geometrisch zwei Halbräumen in der Zeichenebene entsprechen: im Fall 1 suchen wir nach Punkten der Lösungsmenge von $\alpha(x, y) = 0$ unterhalb der Winkelhalbierenden ($y \leq x$) und im Fall 2 nach Lösungspunkten oberhalb der Winkelhalbierenden ($y > x$).

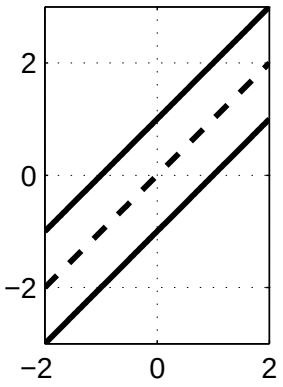
Schauen wir uns aber zunächst Fall 1 an. Nach Definition der Betragsfunktion entspricht $|x-y|$ gerade $x-y$, wenn $x-y \geq 0$ ist. In diesem Fall lautet die Gleichung also einfach

$$x - y - 1 = 0, \quad x - y \geq 0$$

und die Lösungsmenge ist, wie wir wissen, eine Gerade. Diese zeichnen wir unterhalb der Winkelhalbierenden. Im Fall 2 besagt die Definition der Betragsfunktion, daß $|x-y| = -(x-y)$ ist (jetzt ist nämlich $(x-y)$ negativ und daher $-(x-y)$ der positive Betragswert!) Nicht wundern, nur einsetzen:

$$-(x-y) - 1 = 0, \quad x - y < 0$$

Auch hier erhalten wir wieder eine Gerade als Lösungsmenge, die oberhalb der Winkelhalbierenden gezeichnet wird. Insgesamt erhält man die graphische Darstellung der Lösungsmenge



Die Ungleichung $\alpha(x,y) < 0$ entspricht übrigens dem Streifen zwischen den beiden Geraden, da $\alpha(0,0) = -1 < 0$ und $\alpha(2,0) = \alpha(-2,0) = 1 > 0$. Im Fall einer Gleichung mit mehreren Betragsfunktionen ist der Lösungsprozeß übrigens nicht schwieriger, sondern nur aufwendiger. Betrachten wir z. B.

$$\beta(x,y) = \Big|x + |x-2y|\Big| - 1, \quad x,y \in \mathbb{R}$$

und die zugehörige Gleichung $\beta(x,y) = 0$.
Wir arbeiten uns von innen nach außen vor und unterscheiden zunächst $x - 2y \geq 0$ (Fall 1) und $(x-2y) < 0$ (Fall 2). Im Fall 1 stellt sich die Gleichung als

$$|x + (x-2y)| - 1 = 0, \quad x - 2y \geq 0$$

dar, was erneut auf zwei Fälle führt, nämlich $2x - 2y \geq 0$ (Fall 1.1) und $2x - 2y < 0$. Wir stellen fest, daß die Halbebenen $y \leq x/2$, auf die

wir uns im Fall 1 beschränken, durch die zusätzliche Unterscheidung wiederum in zwei Hälften zerlegt wird und zwar durch den Schnitt mit der Halbebene unterhalb der Winkelhalbierenden ($y \leq x$) im Fall 1.1 und oberhalb der Winkelhalbierenden ($y > x$) im Fall 1.2. Im Fall 1.1. reduziert sich die Gleichung also auf

$$2x - 2y - 1 = 0 \quad y \leq 1/2x, \quad y \leq x$$

und liefert eine Gerade, die im Durchschnitt der Halbräume $y \leq x/2$ und $y \leq x$ gezeichnet werden darf. Entsprechend liefert Fall 1.2.

$$-(2x - 2y) - 1 = 0 \quad y \leq x/2, \quad y > x$$

Im Fall 2 lautet die Gleichung wegen $|x - 2y| = -(x - 2y)$ nun

$$0 = |x - (x - 2y)| - 1 = |2y| - 1, \quad x - 2y < 0$$

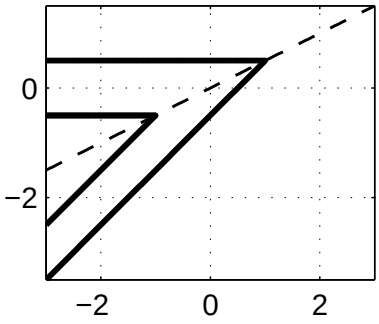
Auch hier ergeben sich zwei Unterfälle, nämlich $y \geq 0$ (Fall 2.1) und $y < 0$ (Fall 2.2.). In jedem der beiden Fälle ist die Lösungsmenge eine Gerade parallel zur x-Achse und zwar im Fall 2.1 gegeben durch

$$2y - 1 = 0 \quad y > x/2, \quad y \leq 0$$

und im Fall 2.2 durch

$$-2y - 1 = 0 \quad y > x/2, \quad y < 0$$

Insgesamt hat die Lösungsmenge die Struktur



wobei wieder im Inneren des abknickenden Streifens $\beta(x, y) < 0$ gilt, denn $\beta(0, 0) = -1 < 0$.
Als Schlußbemerkung sei erwähnt, daß die Lösungsmenge eines Systems von Gleichungen bzw. Ungleichungen der Durchschnitt der Lösungsmengen der einzelnen Gleichungen bzw. Ungleichungen ist.

KAPITEL 2

Lineare Funktionen

1. Vektorräume

Erinnern wir uns noch einmal an die Kompaßnadel-Zeiger aus Abschnitt (4), die beliebig im Raum verschoben werden können, die aber ihre Richtung unverändert beibehalten. Stellen wir uns vor, daß wir zu je zwei Punkten A und B im Raum einen solchen Zeiger $\vec{a}$ „schnitzen" können, dessen Spitze auf B zeigt, wenn sich das andere Ende im Punkt A befindet. Ist $x \geq 0$ eine reelle Zahl, so bezeichne $x\vec{a}$ den neugeschnitzen Zeiger, der in die gleiche Richtung wie $\vec{a}$ zeigt, aber x-mal so lang ist. Also $3\vec{a}$ ist dreimal so lang wie $\vec{a}$ und $1/2\vec{a}$ einhalb mal. Der „Zeiger" $0\vec{a}$ ist besonders einfach zu schnitzen - hier ist gar nichts zu tun, da der Null-Zeiger Länge Null hat. Ein Zeiger, der in die entgegengesetzte Richtung wie $\vec{a}$ zeigt, aber genauso lang ist, nennen wir $-\vec{a}$ und für eine Zahl $x < 0$ soll $x\vec{a}$ bedeuten, daß wir einen Zeiger schnitzen, mit der $|x|$-fachen Länge von $\vec{a}$, der in die entgengesetzte Richtung von $\vec{a}$ zeigt. Das Neuschnitzen definiert somit eine Abbildung: jedem Paar $(x, \vec{a})$ bestehend aus einer reellen Zahl x und einem Zeiger $\vec{a}$ wird ein neuer Zeiger $x\vec{a}$ zugeordnet. Bezeichnet S Zeiger, so haben wir also

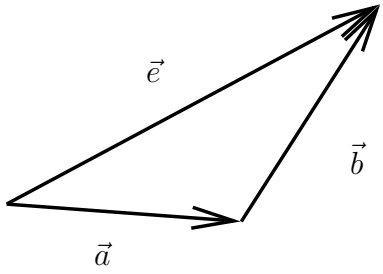
$$\begin{aligned} \mathbb{R} \times S &\longrightarrow S \\ (x, \vec{a}) &\longrightarrow x\vec{a} \end{aligned}$$

Im Folgenden interessieren wir uns nur für die *Zeigewirkung* unserer Zeiger, d. h. wir werden zwei Zeiger als gleich(-wertig) betrachten, wenn die Spitzen auf denselben Endpunkt zeigen, sofern die stumpfen Enden am gleichen Punkt sind. Das bedeutet z. B., daß wir schreiben können

$$y(x\vec{a}) = (yx)\vec{a}, \quad x(-\vec{a}) = (-x)\vec{a}$$

wobei das Gleichheitszeichen zwischen zwei Zeigern eben besagt, daß sie die gleiche Zeigewirkung haben. Zum Zeigen von einem Ausgangspunkt auf einen Endpunkt muß man übrigens nicht unbedingt einzelne Zeiger benutzen. Stellen wir uns vor, wir können das spitze Ende eines Zeigers fest mit dem stumpfen Ende eines anderen Zeigers $\vec{b}$ verbinden. Dann hat die Gesamtkonstruktion immer noch ein stumpfes und ein spitzes Ende und ist somit in der Zeigewirkung gleich(-wertig) mit

einem einzelnen Zeiger $\vec{e}$, der geradlinig das stumpfe mit dem spitzen Ende der Konstruktion verbindet



Beachten Sie auch, daß der aus $\vec{a}$ und $\vec{b}$ bestehende abgeknickte Zeigestock genau wie $\vec{e}$ frei im Raum verschoben werden kann, und daß eine Drehung nicht möglich ist, da sich die Bestandteile $\vec{a}, \vec{b}$ der Richtungsänderung widersetzen.
In diesem Sinne können wir das Verkleben von Zeigern als eine Abbildung verstehen, die jedem Zeigerpaar $(\vec{a}, \vec{b})$ den gleichwertigen Zeiger $\vec{e}$ zuordnet, der das stumpfe Ende von $\vec{a}$ mit dem spitzen Ende des an $\vec{a}$ befestigten Zeigers $\vec{b}$ verbindet. Ist S die Menge aller Zeiger, so bezeichnen wir die Abbildung mit

$$\begin{array}{rcl} + : S \times S & \longrightarrow & S \\ (\vec{a}, \vec{b}) & \longrightarrow & \vec{a} + \vec{b} \end{array}$$

Das Pluszeichen für die Verklebeoperation ist hier übrigens nicht zufällig gewählt. Tatsächlich erfüllt die Verklebeoperation unserer Erfahrung nach alle Eigenschaften, die wir vom „üblichen“+ kennen. So ändert z. B. das Ankleben des Null-Zeigers die Zeigewirkung nicht, d. h.

$$\vec{a} + \vec{0} = \vec{a}$$

Auch ist das Verkleben assoziativ

$$(\vec{a} + \vec{b}) + \vec{c} = \vec{a} + (\vec{b} + \vec{c}),$$

denn es ergibt sich die gleiche Zeigewirkung, wenn man zuerst $\vec{b}$ an $\vec{a}$ klebt und dann $\vec{c}$ an $\vec{b}$ befestigt, oder zuerst $\vec{b}$ und $\vec{c}$ verbindet (immer die Spitze des ersten Zeigers am stumpfen Ende des zweiten Zeigers) und dann diese Kombination an der Spitze von $\vec{a}$ befestigt. Auch zeigt unsere Erfahrung, daß $\vec{a} + \vec{b} = \vec{b} + \vec{a}$ gilt (ob ich zuerst fünf Schritte nach Norden und dann zwei Schritte nach Westen gehe, oder zuerst zwei Schritte nach Westen und dann fünf nach Norden, in jedem Fall komme ich am selben Punkt an. Beachten Sie aber, daß diese Aussage in einem

gekrümmten Raum nicht unbedingt gilt und damit möglicherweise eine Approximation der Realität darstellt).
Durch „Experimente“ können wir uns davon überzeugen, daß die Zeiger zusammen mit Schnitz- und Verklebeoperation folgende Eigenschaften erfüllen (zur Erinnerung: die Gleichheit zweier Konstruktionen bezieht sich auf die Zeigewirkung)

$$\begin{aligned}
\vec{a}+\vec{b} &= \vec{b}+\vec{a} \\
(\vec{a}+\vec{b})+\vec{c} &= \vec{a}+(\vec{b}+\vec{c}) \\
\vec{a}+\vec{0} &= \vec{a} \\
\vec{a}+(-\vec{a}) &= \vec{0} \\
1\vec{a} &= \vec{a} \\
x(y\vec{a}) &= (xy)\vec{a} \\
x(\vec{a}+\vec{b}) &= x\vec{a}+x\vec{b} \\
(x+y)\vec{a} &= x\vec{a}+y\vec{a}
\end{aligned}$$

wobei $\vec{a},\vec{b},\vec{c}$ Elemente der Zeigermenge S sind und $x,y,\in\mathbb{R}$.
Allgemeiner nennt man eine Menge V mit zwei Operationen

$$\begin{array}{ccc} \mathbb{R}\times S & \longrightarrow & S \\ (x,\vec{a}) & \longrightarrow & x\vec{a} \end{array} \qquad \begin{array}{ccc} S\times S & \longrightarrow & S \\ (\vec{a},\vec{b}) & \longrightarrow & \vec{a}+\vec{b} \end{array}$$

einem sogenannten *Nullelement* $\vec{0}$ und *inversen* Elementen $(-\vec{a})$ zu allen $\vec{a}\in V$, die zusammen den obigen Regeln gehorchen, einen *Vektorraum* über $\mathbb{R}$. Die Elemente eines solchen Raumes werden als *Vektoren* bezeichnet und die Operationen heißen Addition bzw. skalare Multiplikation. Die inversen Elemente und das neutrale Element müssen übrigens nicht lange gesucht werden. Man kann nämlich zeigen, daß in jedem Vektorraum die Zusammenhänge

$$0\vec{a}=\vec{0} \qquad (-1)\vec{a}=-\vec{a},\ \vec{a}\in V$$

gelten, wozu man ausschließlich die acht Rechenregeln benutzt. Hat man sich also zu einer Menge V eine Operation $\mathbb{R}\times V\to V$ überlegt, mit dem Ziel, einen Vektorraum zu konstruieren, dann gibt es nur noch eine Wahl für den Nullvektor und die inversen Elemente, mit denen das Ziel erreicht werden kann, nämlich $\vec{0}=0\vec{a}$ für ein beliebiges Element $\vec{a}$ von V und $(-\vec{a})=(-1)\vec{a}$ für alle $\vec{a}\in V$. In der Mathematik findet man häufig Vektorräume von Funktionen. Betrachten wir z. B. alle Funktionen auf einer Menge D mit Werten in der Menge $\mathbb{R}$

$$V=\{f:D\to\mathbb{R}\}.$$

Zur Definition der Skalarmultiplikation müssen wir uns überlegen, welche Funktion αf sein soll, wenn $f \in V$ und $\alpha \in \mathbb{R}$ ist. Die natürliche Definition für das α-fache einer Funktion ist sicherlich die Funktion mit den α-fachen Funktionswerten

$$(\alpha f)(x) = \alpha f(x), \quad x \in D.$$

Entsprechend ist es naheliegend, die Funktion $f + g$ durch die Summe der Funktionswerte von f und g zu definieren

$$(f + g)(x) = f(x) + g(x), \quad x \in D.$$

Mit dieser Vereinbarung für die skalare Multiplikation und die Addition ist der Funktionenraum V tatsächlich ein Vektorraum über $\mathbb{R}$, was man durch Nachprüfen aller Rechengesetze nachweisen kann. Beispielsweise gilt $f + g = g + f$, denn

$$(f + g)(x) = f(x) + g(x) = g(x) + f(x) = (g + f)(x),$$

wobei die Kommutativität der reellen Zahlen ausgenutzt wurde. In dieser Weise können alle Bedingungen leicht nachgeprüft werden. Beachten Sie, daß der Nullvektor in V der Nullfunktion entspricht, die jedem Argument den Wert $0 \in \mathbb{R}$ zuordnet.
Ein anderer Funktionenvektorraum, den Sie gewiß kennen, ist der Raum aller Polynome auf $\mathbb{R}$. Ein Polynom ist eine Funktion der Form

$$p(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n, \quad x \in \mathbb{R}$$

mit beliebigen reellen Koeffizienten a_i. Den höchsten Exponenten, der mit einem nichtverschwindenden Koeffizienten auftritt, nennt man den *Grad* des Polynoms, oder kurz deg p. Die Menge aller reellen Polynome ist also

$$\mathcal{P} = \{p : \mathbb{R} \to \mathbb{R} \,|\, p \text{ Polynom } \}$$

Offensichtlich ist diese Menge im Vektorraum aller reellen Funktionen auf $\mathbb{R}$ enthalten (obiges Beispiel mit $D = \mathbb{R}$). Mit den gleichen Definitionen für skalare Multiplikation und Addition ist aber $\mathcal{P}$ selbst ein Vektorraum für sich, da das Vielfache eines Polynoms wieder ein Polynom ist

$$(\alpha p)(x) = \alpha a_0 + \alpha a_1 x + \ldots + \alpha a_n x^n$$

und auch die Summe zweier Polynome wieder ein Polynom ergibt. Die Rechengesetze müssen jetzt gar nicht mehr überprüft werden, da sie ja für *alle* reelle Funktionen im obigen Beispiel bereits bestätigt wurden. Dieser Fall tritt allgemein für jede Teilmenge U eines Vektorraums V auf, bei der sowohl die skalare Multiplikation als auch die Addition keine Ergebnisse außerhalb von U liefern, genauer, falls für alle $\vec{u}, \vec{w} \in U$

und alle $\alpha \in \mathbb{R}$ $\vec{u} + \vec{w} \in U$ und $\alpha\vec{u} \in U$ gilt. Solch eine Teilmenge nennen wir *Untervektorraum* von V. Die Polynome bilden in dieser Sprechweise also einen Untervektorraum des Vektorraums aller reellen Funktionen auf $\mathbb{R}$. Wir finden sogar eine unendliche Kette von Untervektorräumen $\mathcal{P}_0 \subset \mathcal{P}_1 \subset \mathcal{P}_2 \subset \ldots \subset \mathcal{P}$, wenn wir die Räume $\mathcal{P}_n =$ aller Polynome vom Grad $\leq n$ betrachten

$$\mathcal{P}_n = \{p : \mathbb{R} \to \mathbb{R} | \text{Polynom}, \ \deg p \leq n\}$$

Wir werden noch viele andere Untervektorräume der reellen Funktionen kennen lernen, z. B. die stetigen Funktionen, k-mal differenzierbare Funktionen, integrierbare Funktionen, quadrat-integrierbare Funktionen etc.
Zum Schluß betrachten wir noch einen Spezialfall des ersten Beispiels, den Fall einer endlichen Definitionsmenge $D = \{1, \ldots, n\}$. Aus Abschnitt (1) wissen wir, daß eine reelle Funktion $f : D \to \mathbb{R}$ durch ihre Wertetabelle vollständig beschrieben ist, d. h. die n reellen Zahlen

$$(f(1), \ldots, f(n)) \in \mathbb{R}^n$$

haben den gleichen Informationsgehalt wie f. Die oben eingeführten Operationen Addition und skalare Multiplikation für reelle Funktionen übertragen sich damit auf das Rechnen mit Wertetabellen (n-Tupeln) gemäß

$$\alpha(f(1), \ldots, f(n)) = (\alpha f(1), \ldots, \alpha f(n)), \alpha \in \mathbb{R}$$

und

$$(f(1), \ldots, f(n)) + (g(1), \ldots, g(n)) = (f(1) + g(1), \ldots, f(n) + g(n))$$

Mit diesen Definitionen für Addition und skalare Multiplikation bildet der $\mathbb{R}^n$ folglich einen Vektorraum.

2. Basis und Dimension

Wir entwickeln den Begriff Basis zunächst für unseren anschaulichen Zeiger-Vektorraum. Dazu erinnern wir uns an das kartesische Koordinatensystem $(A, \vec{s}_1, \vec{s}_2, \vec{s}_3)$, das aus einem Referenzpunkt A und drei zueinander senkrecht angeordneten Zeigern $\vec{s}_1, \vec{s}_2, \vec{s}_3 \in S$ besteht. Ist $\vec{a} \in S$ ein beliebiger Zeiger, so zeigt er auf einen bestimmten Punkt P, wenn sein stumpfes Ende in A befestigt wird. Jeder Punkt P im Raum läßt sich aber mit Hilfe des Koordinatensystems durch drei Zahlen (x_1, x_2, x_3) beschreiben, wobei x_i angibt, wie weit man von A aus in Richtung $\vec{s}_i$ gehen muß, um nach P zu gelangen. Anders ausgedrückt, schnitzen wir drei Zeiger $x_1\vec{s}_1, x_2\vec{s}_2$ und $x_3\vec{s}_3$ und befestigen das stumpfe Ende von $x_1\vec{s}_1$ in A, das stumpfe Ende von $x_2\vec{s}_2$ an der Spitze von

$x_1\vec{s}_1$ und schließlich das stumpfe Ende von $x_3\vec{s}_3$ an der Spitze von $x_2\vec{s}_2$, so zeigt die Spitze von $x_3\vec{s}_3$ genau auf den Punkt P, d. h. es gilt

$$x_1\vec{s}_1 + x_2\vec{s}_2 + x_3\vec{s}_3 = \vec{a}$$

Daraus können wir zwei wichtige Schlußfolgerungen ziehen. Erstens kann man die Menge aller Raumpunkte mit der Menge aller Zeiger identifizieren, wenn man einen Referenzpunkt A im Raum auszeichnet. Zu einem beliebigen Punkt P gehört dann der Zeiger, der von A auf P zeigt und einen beliebigen Zeiger kann man umgekehrt als Stellvertreter für den Punkt betrachten, auf den der Zeiger zeigt, wenn sein stumpfes Ende sich im Referenzpunkt A befindet. In dieser Interpretation bezeichnet man $\vec{a} \in S$ auch als *Ortsvektor*.
Die zweite Schlußfolgerung ist, daß jeder Zeiger als Kombination der drei Zeiger $\vec{s}_1, \vec{s}_2, \vec{s}_3$ erzeugt werden kann, wobei diese Eigenschaft verloren geht, wenn wir auf nur einen der *Basis-Zeiger* $\vec{s}_1, \vec{s}_2, \vec{s}_3$ verzichten. Eine ähnliche Situation finden wir im Vektrorraum $\mathbb{R}^2$ mit den beiden Vektoren $(1, 1)$ und $(1, 0)$. Können mit diesen Vektoren alle anderen Zahlenpaare dargestellt werden? Ist $\vec{a} = (a_1, a_2)$ ein beliebiges Paar, so fragen wir uns, ob es Kombinationskoeffizienten $x_1, x_2 \in \mathbb{R}$ gibt, so daß

$$(a_1, a_2) = x_1(1, 1) + x_2(1, 0) = (x_1 + x_2, x_1)$$

gilt. Die resultierenden Bedingungen $a_1 = x_1 + x_2$ und $a_2 = x_1$ an die Koeffizienten lassen sich auflösen

$$x_1 = a_2, \qquad x_2 = a_1 - x_1 = a_1 - a_2$$

und wir sehen, daß es *genau eine* Möglichkeit der Darstellung mit $(1, 1)$ und $(1, 0)$ gibt, nämlich

$$(a_1, a_2) = a_1(1, 1) + (a_1 - a_2)(1, 0)$$

Bei der Generierung beliebiger Paare $\vec{a} = (a_1, a_2)$ sind offensichtlich beide Vektoren unverzichtbar. Würden wir $(1, 0)$ weglassen, so ließe sich nicht mehr jeder Vektor durch das übrig gebliebene Zahlenpaar $(1, 1)$ darstellen. Zum Beispiel ist $(1, 0)$ selbst durch $(1, 1)$ nicht darstellbar, da $x(1, 1) \neq (1, 0)$ für alle $x \in \mathbb{R}$. Genauso läßt sich der Vektor $(1, 1)$ nicht durch $(1, 0)$ darstellen, so daß auch auf $(1, 1)$ nicht verzichtet werden kann. Zusammenfassend können wir also sagen, daß die Vektoren $\{(1, 1), (1, 0)\}$ den ganzen $\mathbb{R}^2$ erzeugen können, wobei diese Eigenschaft verloren geht, wenn wir auf einen der Vektoren verzichten. In diesem Sinne kann man $\{(1, 1), (1, 0)\}$ als minimales Erzeugendensystem des $\mathbb{R}^2$ bezeichnen.

Als weiteres Beispiel betrachten wir den Raum $\mathcal{P}_1$ aller Polynome vom Grad ≤ 1 und fragen uns, ob $\{q_0, q_1, q_2\}$ mit

$$q_0(x) = 1, \quad q_1(x) = x - 1, \quad q_2(x) = x + 1$$

ein Erzeugendensystem von $\mathcal{P}_1$ ist. Dazu müssen wir nur zeigen, daß jedes Element von $\mathcal{P}_1$ als Linearkombination der q_i darstellbar ist. Sei $p(x) = a_0 + a_1 x$ solch ein allgemeines Polynom. Dann gilt offensichtlich

$$p = a_0 q_0 + a_1(q_1 + q_0) = (a_0 + a_1)q_0 + a_1 q_1$$

d. h. $\{q_0, q_1, q_2\}$ ist ein Erzeugendensystem. Da der Vektor q_2 bei der Darstellung gar nicht benötigt wurde, könnte er auch aus dem Erzeugendensystem entfernt werden, ohne daß die Darstellbarkeit beliebiger Polynome darunter leidet. Das Erzeugendensystem ist also nicht minimal. Außerdem können wir beobachten, daß mit solch einem nichtminimalen Erzeugendensystem die Darstellung der Vektoren mehrdeutig ist. So hat das Polynom $p(x) = 2x$ mindestens zwei Darstellungen

$$p = 2q_0 + 2q_1, \quad p = q_1 + q_2$$

Als abschließendes Beispiel betrachten wir noch den Raum $\mathcal{P}$ aller Polynome auf $\mathbb{R}$. Durch die Polynome $q_n(x) = x^n$ mit $n \in \mathbb{N}_0$ wird offensichtlich ein minimalen Erzeugendensystem definiert. Wir sehen mit diesem Beispiel, daß auch Erzeugendensysteme mit unendlich vielen Elementen auftreten können.

Zur Verallgemeinerung unserer Beobachtungen wenden wir uns nun einem allgemeinen Vektorraum V über $\mathbb{R}$ zu, von dem wir nur die grundlegenden Vektorraumeigenschaften kennen. Eine nichtleere Menge $E \subset V$ von Vektoren bezeichnen wir als *Erzeugendensystem* des Vektorraums, wenn jeder Vektor $\vec{v} \in V$ als endliche Kombination (sogenannte *Linearkombination*)

$$\vec{v} = x_1\vec{e}_1 + x_2\vec{e}_2 + \ldots + x_m\vec{e}_m \quad x_i \in \mathbb{R}, \vec{e}_i \in E, m \in \mathbb{N}$$

geschrieben werden kann (beachten Sie, daß E nicht endlich sein muß und daß die Anzahl der Summanden je nach Vektor $\vec{v}$ variieren kann). Ist das Erzeugendensystem *minimal* in dem Sinne, daß beim Weglassen auch nur eines Vektors die Erzeugendeneigenschaft von E verloren geht, so bezeichnen wir E auch als *Basis* und die Elemente als Basisvektoren. Eine Basis ist also ein minimales Erzeugendensystem.

Der Vollständigkeit halber sei erwähnt, daß man die leere Menge $E = \emptyset$ als Erzeugendensystem des Nullvektorraums $V = \{\vec{0}\}$ definiert. Der Nullvektorraum ist offensichtlich ein sehr langweiliges Objekt, da er nur ein einziges Element, den Nullvektor, enthält. Trotzdem erfüllt diese Menge zusammen mit den durch $\vec{0} + \vec{0} = \vec{0}$ und $x \cdot \vec{0} = \vec{0}, x \in \mathbb{R}$ definierten Operationen alle Vektrorraumeigenschaften und ist somit

ein vollwertiger reeller Vektorraum. Durch den Trick, die leere Menge als Erzeugendensystem von $\{\vec{0}\}$ zu definieren, verhindert man, daß der Nullvektor in einem minimalen Erzeugendensystem auftauchen kann. In „vernünftigen“ Vektorräumen $V \neq \{\vec{0}\}$ ist das sowieso der Fall, da man bei der Darstellung von Vektoren sicherlich aufs Vielfache des Nullvektors verzichten kann. Im Fall $V = \{\vec{0}\}$ dagegen kann man nur dann auch auf den Nullvektor im Erzeugendensystem verzichten, wenn eben $\{\vec{0}\} \setminus \{\vec{0}\} = \emptyset$ immer noch ein Erzeugendensystem ist.
Wenden wir uns nun der Frage zu, wann ein Vektor $\vec{v} \in V$ eindeutig durch ein Erzeugendensystem darstellbar ist. Dabei betrachten wir zunächst die Darstellbarkeit des Nullvektors. Dieser spezielle Vektor läßt sich in trivialer Weise darstellen. Es gilt nämlich $\vec{0} = x_1\vec{e}_1$ mit $x_1 = 0$ und $\vec{e}_1 \in E$, oder auch $\vec{0} = x_1\vec{e}_1 + x_2\vec{e}_2$ mit $x_1 = x_2 = 0$ und $\vec{e}_1, \vec{e}_2 \in E$ oder $\vec{0} = 1 \cdot \vec{e}_1 + (-1)\vec{e}_1$ usw. Läßt sich der Nullvektor auch weniger langweilig darstellen? Nehmen wir einmal an, daß

$$\vec{0} = x_1\vec{e}_1 + \ldots + x_n\vec{e}_n, \quad \vec{e}_i \in E$$

wobei die Vektoren $\vec{e}_i$ paarweise verschieden sind und *nicht* alle $x_i = 0$ seien; also z. B. $x_1 \neq 0$. Dann gilt aber

$$\vec{e}_1 = -\frac{x_2}{x_1}\vec{e}_2 - \ldots - \frac{x_n}{x_1}\vec{e}_n$$

d. h. der Vektor $\vec{e}_1$ im Erzeugendensystem ist überflüssig, da er selbst ja bereits durch die Vektoren $\vec{e}_2, \ldots, \vec{e}_n$ dargestellt werden kann. Wir können $\vec{e}_1$ also aus E entfernen, ohne daß die Erzeugendeneigenschaft verloren geht, d. h. E kann *keine* Basis sein, wenn sich die Null nichtlangweilig darstellen läßt. Anders herum gesagt, gibt es mit einer Basis nur eine einzige Möglichkeit, den Nullvektor durch paarweise verschiedene Vektoren darzustellen, indem man alle Koeffizienten gleich Null wählt. Diese Besonderheit bezeichnen wir als *lineare Unabhängigkeit*, d. h. kann mit Vektoren $\vec{a}_1, \ldots, \vec{a}_n$ der Nullvektor nur dadurch dargestellt werden, daß alle Koeffizienten gleich Null gewählt werden, so heißen $\vec{a}_1, \ldots, \vec{a}_n$ *linear unabhängig*. Ist dagegen eine nicht-triviale Darstellung möglich, so heißen die Vektoren *linear abhängig*.
Unsere Beobachtung, daß der Nullvektor mit paarweise verschiedenen Basisvektoren nur trivial darstellbar ist (d. h. nur mit Null- Koeffizienten) kann also auch so formuliert werden, daß paarweise verschiedene Basisvektoren $\vec{e}_1, \ldots, \vec{e}_m$ immer linear unabhängig sind. In diesem Sinne ist also eine Basis ein linear unabhängiges Erzeugendensystem. Umgekehrt ist aber auch jedes linear unabhängige Erzeugendensystem eine Basis. Um dies zu zeigen, nehmen wir einmal an, wir hätten ein linear

unabhängiges Erzeugendensystem E, das keine Basis, also nicht minimal ist. Insbesondere gibt es dann einen Vektor $\vec{e} \in E$, den man aus E herausnehmen kann, ohne daß die Erzeugendeneigenschaft darunter leidet. Stellen wir den Vektor $\vec{e}$ durch paarweise verschiedene Vektoren $\vec{b}_1, \ldots, \vec{b}_m$ aus dem verkleinerten Erzeugendensystem $E \setminus \{\vec{e}\}$ dar, so folgt

$$x_1\vec{b}_1 + \ldots + x_m\vec{b}_m - \vec{e} = \vec{0}.$$

Das ist aber ein Widerspruch zur linearen Unabhängigkeit von E und somit ist die Annahme falsch, daß man einen Vektor aus E entfernen kann, ohne die Erzeugendeneigenschaft zu verlieren. Wir haben damit gezeigt

Satz 1. *Die Basen eines Vektorraums sind genau die linear unabhängigen Erzeugendensysteme .*

Wie sieht es aber jetzt mit der eindeutigen Darstellung anderer Vektoren als dem Nullvektor aus? Nehmen wir zunächst an, daß ein Vektor $\vec{v}$ zwei Darstellungen der Form

$$\vec{v} = x_1\vec{e}_1 + \ldots + x_n\vec{e}_n$$

und

$$\vec{v} = y_1\vec{e}_1 + \ldots + y_n\vec{e}_n$$

mit den paarweise verschiedenen Vektoren $\vec{e}_1, \ldots, \vec{e}_n$ der Basis E hat. Durch Gleichsetzen und Vorklammern ergibt sich sofort

$$\vec{0} = (x_1 - y_1)\vec{e}_1 + \ldots + (x_n - y_n)\vec{e}_n$$

und da es in einer Basis nur die langweilige Darstellung der Null gibt, folgt, daß die Koeffizienten x_i, y_i in der Darstellung von $\vec{v}$ gleich sein müssen. Ein aufmerksamer Beobachter wird bemerkt haben, daß diese Aussage noch nicht ganz ausreicht, um zu schließen, daß die Darstellung von Vektoren in einer Basis eindeutig ist. Es könnte ja sein, daß der gleiche Vektor $\vec{v}$ die beiden Darstellungen mit paarweise verschiedenen Basisvektoren

$$\vec{v} = x_1\vec{e}_1 + \ldots + x_n\vec{e}_n + x_{k+1}\vec{a}_1 + \ldots + x_{k+n}\vec{a}_n$$

und

$$\vec{v} = y_1\vec{e}_1 + \ldots + y_n\vec{e}_n + y_{k+1}\vec{b}_1 + \ldots + y_{k+m}\vec{b}_m$$

hat, wobei ein Teil der Basisvektoren in beiden Darstellungen auftreten $(\vec{e}_1, \ldots, \vec{e}_n)$, ein anderer Teil der Basisvektoren $(\vec{a}_1, \ldots, \vec{a}_n)$ bzw.

$(\vec{b}_1, \ldots, \vec{b}_m)$ aber nur in jeweils einer der beiden Darstellungen. Tatsächlich ist dieser Fall aber einfach abzuhandeln, denn es gilt ja

$$\begin{aligned}\vec{v} &= x_1\vec{e}_1 + \ldots + x_k\vec{e_k} + x_{k+n}\vec{a}_1 + \ldots + x_{k+n}\vec{a}_n + 0\vec{b}_1 + \ldots + 0\vec{b}_m \\ &= y_1\vec{e}_1 + \ldots + y_k\vec{e_k} + 0\vec{a}_1 + \ldots + 0\vec{a}_n + y_{k+1}\vec{b}_1 + \ldots + y_{k+m}\vec{b}_m\end{aligned}$$

wobei jetzt in beiden Darstellungen die gleichen paarweise verschiedenen Basisvektoren auftauchen. Für diesen Fall wissen wir aber schon, daß die Koeffizienten dann übereinstimmen, d. h. es gilt $x_i = y_i$ für $i = 1, \ldots, k$ und $x_i = y_i = 0$ für $i > k$. Wir fassen unser Ergebnis in einem Satz zusammen

Satz 2. *Sei V ein Vektorraum über $\mathbb{R}$ und sei E eine Basis von V. Dann läßt sich der Nullvektor nur durch Linearkombinationen mit Null-Koeffizienten darstellen und für jeden Vektor $\vec{0} \neq \vec{v} \in V$ gibt es genau eine Darstellungsmöglichkeit der Form*

$$\vec{v} = x_1\vec{e}_1 + \ldots + x_n\vec{e}_n$$

mit paarweise verschiedenen Vektoren $\vec{e}_i \in E$ und $0 \neq x_i \in \mathbb{R}$.

Basen sind also nützlich zur eindeutigen Charakterisierung jedes Vektors $\vec{v} \in V$ und es stellt sich die Frage, ob solche Erzeugendensysteme in jedem Vektrorraum zur Verfügung stehen. Beschränken wir uns auf *endlich erzeugte* Vektorräume, also solche, für die ein endliches Erzeugendensystem existiert, dann läßt sich die Frage mit Ja beantworten. Eine Möglichkeit, eine Basis zu finden, besteht offensichtlich darin, solange überflüssige Vektoren aus dem endlichen Erzeugendensystem zu entfernen, bis das nicht weiter möglich ist, ohne die Erzeugendeneigenschaft zu verlieren. Das Ergebnis ist ein minimales Erzeugendensystem, also eine Basis. Eine weitere Möglichkeit beruht auf der Feststellung, daß jedes Erzeugendensystem aus linear unabhängigen Vektoren eine Basis bildet. Wir können also eine Basis aufbauen, indem wir so viele linear unabhängige Vektoren sammeln, bis sie ein Erzeugendensystem bilden. Als Sammelgrundlage nehmen wir dazu ein endliches Erzeugendensystem E von $V \neq \{\vec{0}\}$, da dann sichergestellt ist, daß wir irgendwann ein Erzeugendensystem zusammen haben. Wir beginnen mit der Wahl eines Vektors $\vec{u}_1 \neq \vec{0}$ aus E. Der Vektor ist automatisch linear unabhängig, da mit ihm der Nullvektor nur trivial dargestellt werden kann, d. h. $0 \cdot \vec{u}_1 = \vec{0}$. Wir sind damit in folgender Situation (mit $r = 1$)

$$E = \{\vec{u}_1, \ldots, \vec{u}_r, \vec{v}_1, \ldots, \vec{v}_s\} \quad \vec{u}_1, \ldots, \vec{u}_r \quad \text{linear unabhängig.}$$

Ist $\vec{u}_1, \ldots, \vec{u}_r$ noch kein Erzeugendensystem, dann gibt es ein $\vec{v}_i$, so daß $\vec{u}_1, \ldots, \vec{u}_r, \vec{v}_i$ linear unabhängig sind. Das liegt daran, daß mindestens ein $\vec{v}_i$ sich noch nicht durch $\vec{u}_1, \ldots, \vec{u}_r$ darstellen läßt (sonst könnten

die Vektoren $\vec{u}_i$ ja bereits jedes Element des Erzeugendensystems und damit auch jeden Vektor in V darstellen). Die Überprüfung der linearen Unabhängigkeit ist einfach. Nehmen wir an, daß

$$\lambda\vec{v}_i + \lambda_1\vec{u}_1 + \lambda_2\vec{u}_2 + \ldots + \lambda_r\vec{u}_r = \vec{0}.$$

Dann muß aber $\lambda = 0$ sein, denn sonst wäre $\vec{v}_i$ ja durch $\vec{u}_1, \ldots, \vec{u}_r$ darstellbar. Da $\vec{u}_1, \ldots, \vec{u}_r$ linear unabhängig sind, folgt schließlich auch $\lambda_1 = \ldots = \lambda_r = 0$, da nur die triviale Darstellung möglich ist. Insgesamt folgt $\lambda = \lambda_1 = \ldots = \lambda_r = 0$ und die Vektoren sind damit linear unabhängig. Nennen wir den Vektor $\vec{v}_i$ nun $\vec{u}_{r+1}$, so können wir die Prozedur solange wiederholen, bis entweder $\vec{u}_1, \ldots, \vec{u}_m$ ein Erzeugendensystem geworden ist, oder bis keine Vektoren $\vec{v}_i$ mehr übrig sind. In diesem Fall umfassen die Vektoren $\vec{u}_i$ alle Elemente von E und sind damit wiederum ein Erzeugendensystem.
In beiden Situationen bilden die linear unabhängigen Vektoren $\vec{u}_i$ ein Erzeugendensystem und damit eine Basis. Eine interessante Eigenschaft der Basen von endlich erzeugten Vektorräumen ist deren konstante Länge: Jede Basis hat gleich viele Elemente und diese Anzahl nennt man die *Dimension* des Vektorraums V, oder kurz $\dim V$. Um das einzusehen, nehmen wir an, wir hätten zwei Basen E, B unterschiedlicher Länge, wobei $E = \{\vec{e}_1, \ldots, \vec{e}_n\}$ endlich sei und B mehr Elemente hätte als E. Entfernen wir den Vektor $\vec{e}_1$ aus E, so bilden die übrigen Vektoren $\vec{e}_i$ kein Erzeugendensystem mehr, sind aber noch linear unabhängig. Wenn wir nun $F_1 = \{\vec{e}_2, \ldots, \vec{e}_n\} \cup B$ betrachten, sind wir in einer Situation wie oben: F_1 ist ein Erzeugendensystem und die Vektoren $\vec{e}_2, \ldots, \vec{e}_n \in F_1$ sind linear unabhängig, bilden aber für sich genommen noch kein Erzeugendensystem. Wir haben gesehen, daß wir dann aus den restlichen Vektoren B einen Vektor $\vec{b}_1$ auswählen können, so daß $\vec{b}_1, \vec{e}_2, \ldots, \vec{e}_n$ linear unabhängig ist. Diese Vektoren bilden sogar ein Erzeugendensystem und damit eine Basis, denn der weggelassene Vektor $\vec{e}_1$ läßt sich mit ihnen darstellen. Wir wissen nämlich, daß sich $\vec{b}_1$ durch $\vec{e}_1, \ldots, \vec{e}_n$ darstellen läßt, da E eine Basis ist, also

$$\vec{b}_1 = x_1\vec{e}_1 + x_2\vec{e}_2 + \ldots + x_n\vec{e}_n$$

Dabei ist $x_1 \neq 0$, denn sonst wären $\vec{b}_1, \vec{e}_2, \ldots, \vec{e}_n$ linear abhängig und folglich

$$\vec{e}_1 = \frac{1}{x_1}\vec{b}_1 - \frac{x_2}{x_1}\vec{e}_2 - \ldots - \frac{x_n}{x_1}\vec{e}_n$$

Betrachten wir nun $F_2 = \{\vec{b}_1, \vec{e}_3, \ldots, \vec{e}_n\} \cup B$, so folgt mit dem gleichen Argument, daß es einen Vektor $\vec{b}_2 \in B$ gibt, so daß $\vec{b}_1, \vec{b}_2, \vec{e}_3, \ldots, \vec{e}_n$ eine Basis ist. Nach n Schritten haben wir schließlich alle Vektoren $\vec{e}_i$ durch

Vektoren $\vec{b}_i$ der Basis B ausgetauscht, wobei $\vec{b}_1, \ldots, \vec{b}_n$ eine Basis darstellt. Das kann aber eigentlich nicht sein, da B ja mehr als n Elemente hat und ein minimales Erzeugendensystem ist. Mit einer echten Teilmenge von B dürfte also kein Erzeugendensystem zustande kommen, was der Basiseigenschaft von $\vec{b}_1, \ldots, \vec{b}_n$ widerspricht. Folglich kann die Annahme, daß B mehr Elemente enthält als E, nicht richtig sein und damit folgt, daß je zwei Basen eines endlich erzeugten Vektorraums die gleiche Anzahl von Elementen besitzen.
Betrachten wir nun einige Beispiele. Zunächst bestimmen wir die Dimension des Vektorraums $\mathbb{R}^n$. Dazu genügt es, *irgendeine* Basis zu finden und deren Elemente zu zählen. Eine besonders einfache Basis, die sogenannte kanonische Basis des $\mathbb{R}^n$, ist durch die Vektoren $\vec{e}_i$ gegeben, wobei das n-Tupel $\vec{e}_i \in \mathbb{R}^n$ in allen Einträgen eine Null hat bis auf den i-ten Eintrag, wo eine Eins steht, also z. B. in $\mathbb{R}^3$ die Vektoren

$$\vec{e}_1 = (1, 0, 0), \quad \vec{e}_2 = (0, 1, 0), \quad \vec{e}_3 = (0, 0, 1).$$

Ein beliebiges n-Tupel $(a_1, \ldots, a_n)$ läßt sich damit offensichtlich darstellen, denn

$$(a_1, \ldots, a_n) = \sum_{i=1}^{n} a_i \vec{e}_i = a_1 \vec{e}_1 + \ldots + a_n \vec{e}_n.$$

Die Menge $\{\vec{e}_1, \ldots, \vec{e}_n\}$ ist damit ein Erzeugendensystem des $\mathbb{R}^n$ und außerdem ist sie linear unabhängig. Nehmen wir nämlich an, daß

$$x_1 \vec{e}_1 + x_2 \vec{e}_2 + \ldots + x_n \vec{e}_n = \vec{0}$$

ist, so folgt in der Tupel-Schreibweise

$$(x_1, x_2, \ldots, x_n) = (0, 0, \ldots, 0)$$

und damit gibt es nur die triviale Darstellung des Nullvektors. Wir haben also eine Basis des $\mathbb{R}^n$ gefunden und da sie n Elemente enthält, ist die Dimension des Raumes $\mathbb{R}^n$ gerade n. Ähnlich zeigt man, daß die Räume $\mathcal{P}_n$ die Dimension $n+1$ besitzen, wobei hier z. B. die Polynome $q_0(x) = 1, \ldots, q_n(x) = x^n$ als Basis dienen. Die Kenntnis der Dimension eines Vektorraumes hilft übrigens in einigen Fällen sehr schnell zu entscheiden, ob eine gegebene Menge von Vektoren $\{\vec{a}_1, \ldots, \vec{a}_m\}$ ein Erzeugendensystem, eine Basis, oder linear unabhängig sein kann. ist nämlich $m > \dim V$, so können $\vec{a}_1, \ldots, \vec{a}_m$ *nicht* linear unabhängig sein. Es wäre nämlich sonst möglich, aus $\vec{a}_1, \ldots, \vec{a}_m$ eine Basis zu konstruieren (genauso wie oben beschrieben), die mehr als $\dim V$ Vektoren enthält, also sind mehr als $\dim V$ Vektoren stets linear abhängig. Umgekehrt können die Vektoren $\vec{a}_1, \ldots, \vec{a}_m$ im Fall $m < \dim V$ kein

Erzeugendensystem bilden, da es sonst ein minimales Erzeugendensystem (also eine Basis) von V gäbe, das weniger als $\dim V$ Elemente hat. Die Möglichkeit, daß $\vec{a}_1, \ldots, \vec{a}_m$ eine Basis bildet, kommt schließlich nur dann in Betracht, wenn $m = \dim V$ gilt. Allerdings bleibt in diesem Fall entweder die Erzeugendeneigenschaft *oder* die lineare Unabhängigkeit zu überprüfen.
Sie können also jetzt ohne nachzurechnen sagen, daß die Vektoren

$$(1,1), (2,-1), (0,1)$$

linear abhängig sind und daß

$$(1,1,5), \quad (0,1,2)$$

kein Erzeugendensystem $\mathbb{R}^3$ sein kann. Ob dagegen

$$(1,1,5), \quad (0,1,2), \quad (1,0,3)$$

eine Basis des $\mathbb{R}^3$ ist, entscheidet sich erst durch genaueres Hinsehen. Nehmen wir an,

$$x_1(1,1,5) + x_2(0,1,2) + x_3(1,0,3) = (0,0,0),$$

dann gilt

$$x_1 + x_3 = 0, \quad x_1 + x_2 = 0, \quad 5x_1 + 2x_2 + 3x_3 = 0$$

was sich in

$$x_3 = -x_1, \quad x_2 = -x_1, \quad 0 = 0$$

umformen läßt. Eine spezielle Lösung ist also z. B. durch $x_1 = 1, x_2 = -1, x_3 = -1$ gegeben, d. h. es gibt eine nicht-triviale Darstellung des Nullvektors und damit bildet obige Menge aus drei Vektoren *keine* Basis des $\mathbb{R}^3$, da die Vektoren linear abhängig sind.

3. Lineare Abbildungen

Betrachten wir zunächst den Vektorraum S der Zeiger im physikalischen Raum. Wie wir gesehen haben, läßt sich jeder Zeiger $\vec{s}$ durch die drei speziellen Zeiger $\vec{s}_1, \vec{s}_2, \vec{s}_3$, mit denen wir ein kartesisches Koordinatensystem aufgebaut hatten, darstellen

$$\vec{s} = x_1\vec{s}_1 + x_2\vec{s}_2 + x_3\vec{s}_3.$$

Versucht man den Nullzeiger zu konstruieren, so merkt man schnell, daß nur die triviale Darstellung in Frage kommt. Wäre z. B. $x_1 \neq 0$, dann könnten wir durch Ankleben beliebig langer oder kurzer Zeiger die nach links bzw. rechts oder nach oben bzw. unten zeigen, nie mehr mit der Spitze der Konstruktion an das stumpfe Ende zurückkehren. Mit anderen Worten, die Zeiger $\vec{s}_1, \vec{s}_2, \vec{s}_3$ bilden ein linear unabhängiges Erzeugendensystem, also eine Basis. Wir schließen daraus $\dim S = 3$,

was sich mit unserer umgangssprachlichen Nutzung der Bezeichnung dreidimensionaler Raum deckt (denken Sie daran, daß wir ja die Zeiger durch Auszeichnung eines Referenzpunktes als Ortsvektoren betrachten und mit Punkten des Raumes identifizieren können).
Insbesondere wissen wir jetzt, daß jeder Zeiger durch die drei Koeffizienten x_i in der Basisdarstellung, den sogenannten *Koordinaten* bezüglich der Basis $\vec{s}_1, \vec{s}_2, \vec{s}_3$ eindeutig charakterisiert ist. Jeder Zeiger gibt Anlaß zu einem Koordinatentripel und zu jedem Tripel gehört genau ein Zeiger. Mit anderen Worten haben wir es hier mit einer Abbildung zwischen zwei Vektorräumen zu tun, der sogenannten *Koordinatenabbildung*, die jedem $\vec{s} \in S$ ein Tripel $(x_1, x_2, x_3) \in \mathbb{R}^3$ zuordnet, eben die Koordinaten von $\vec{s}$. Schauen wir uns diese Abbildung, die wir hier einmal $\Phi : S \to \mathbb{R}^3$ nennen wollen, etwas genauer an. Wenn $\Phi(\vec{s}) = (x_1, x_2, x_3) = \vec{x}$ ist, was ist dann der Funktionswert von dem verlängerten Zeiger $\Phi(\alpha\vec{s})$? Dazu müssen wir uns nur der definierenden Eigenschaft des Zusammenhangs Φ erinnern

$$\vec{s} = x_1\vec{s}_1 + x_2\vec{s}_2 + x_3\vec{s}_3, \quad \vec{x} = \Phi(\vec{s})$$

Offensichtlich ist unter Ausnutzung der Vektorraum Rechenregeln

$$\alpha\vec{s} = (\alpha x_1)\vec{s}_1 + (\alpha x_2)\vec{s}_2 + (\alpha x_3)\vec{s}_3$$

so daß wir die Koordinaten von $\alpha\vec{s}$, also $\Phi(\alpha\vec{s})$, direkt ablesen können. Es gilt

$$\Phi(\alpha\vec{s}) = \alpha\vec{x} = \alpha\Phi(\vec{s}), \qquad \alpha \in \mathbb{R}.$$

Ähnlich verhält es sich mit dem Verkleben von zwei Zeigern $\vec{s}, \vec{v}$ mit Koordinaten $\Phi(\vec{s}) = \vec{x}$ und $\Phi(\vec{v}) = \vec{y}$. Für die Summe der beiden Zeiger gilt offensichtlich

$$\begin{aligned} \vec{s} + \vec{v} &= (x_1\vec{s}_1 + x_2\vec{s}_2 + x_3\vec{s}_3) + (y_1\vec{s}_1 + y_2\vec{s}_2 + y_3\vec{s}_3) \\ &= (x_1 + y_1)\vec{s}_1 + (x_2 + y_2)\vec{s}_2 + (x_3 + y_3)\vec{s}_3 \end{aligned}$$

d. h. die Koordinaten der Summe sind durch die Summe der Korrdineten gegeben,

$$\Phi(\vec{s} + \vec{v}) = \vec{x} + \vec{y} = \Phi(\vec{s}) + \Phi(\vec{v})$$

Beachten Sie, daß die beiden Pluszeichen zwar identisch aussehen, aber doch etwas ganz Verschiedenes bedeuten. Es kommt eben darauf an, was addiert wird. Eine Addition $\Phi(\vec{s}+\vec{v})$ im Argument von Φ bedeutet Verkleben von Zeigern. Die Addition der Funktionswerte $\Phi(\vec{s}) + \Phi(\vec{v})$ bezieht sich dagegen auf komponentenweises Addieren in Zahlentripeln. Entsprechendes gilt für die beiden skalaren Multiplikationen $\Phi(\alpha\vec{s})$

und $\alpha\phi(\vec{s})$. Jetzt ahnen Sie vielleicht den großen Vorteil von Abbildungen, die wie Φ die Operationen auf zwei Vektorräumen ineinander übersetzen (sogenannten *linearen Abbildungen*).
Statt vor der Berechnung des Funktionswerte aufwendige Operationen in einem Vektorraum durchführen zu müssen, wie z. B. mehrmaliges Neuschnitzen und Verkleben von Zeigern

$$\Phi(\lambda_1\vec{a}_1 + \lambda_2\vec{a}_2 + \ldots + \lambda_n\vec{a}_n) =?$$

kann man bei linearen Abbildungen auch zunächst die Funktionswerte $\Phi(\vec{a}_i)$ der beteiligten Vektoren bestimmen und dann die Operationen im Zielvektorraum durchführen, was im vorliegenden Fall wesentlich weniger aufwendig ist, da man nur ein paar Zahlen zusammenrechnen muß

$$\lambda_1\Phi(\vec{a}_1) + \lambda_2\Phi(\vec{a}_2) + \ldots + \lambda_n\Phi(\vec{a}_n)$$

Daß die Werte tatsächlich übereinstimmen, folgt aus den beiden Vektorraum-Rechenregeln. Nennen wir z. B. $\lambda_1\vec{a}_1 = \vec{s}$ und $\lambda_2\vec{a}_2 + \ldots + \lambda_n\vec{a}_n = \vec{v}$, so gilt $\Phi(\vec{s}) = \Phi(\lambda_1\vec{a}_1) = \lambda_1\Phi(\vec{a}_1)$ und

$$\begin{aligned}\Phi(\lambda_1\vec{a}_1 + \ldots + \lambda_n\vec{a}_n) &= \Phi(\vec{s}+\vec{v}) = \Phi(\vec{s}) + \Phi(\vec{v}) \\ &= \lambda_1\Phi(\vec{a}_1) + \Phi(\lambda_2\vec{a}_2 + \ldots + \lambda_n\vec{a}_n)\end{aligned}$$

Durch wiederholtes Anwenden dieser Regeln sieht man, daß lineare Abbildungen beliebige Linearkombinationen im Argument in entsprechende Linearkombinationen der Funktionswerte verwandelt.
Allgemein nennen wir eine Abbildung $L : V \to W$ zwischen zwei reellen Vektorräumen V und W eine lineare Abbildung, falls die Bedingungen

$$L(x\vec{u}) = xL(\vec{u}), \quad x \in \mathbb{R}, \vec{u} \in V \qquad L(\vec{u}+\vec{v}) = L(\vec{u}) + L(\vec{v}) \quad \vec{u}, \vec{v} \in V$$

erfüllt sind. Wie wir gesehen haben, ist die Koordinatenabbildung von einem endlich dimensionalen Vektorraum V mit $\dim V = n$ in den Raum $\mathbb{R}^n$ aller n-Tupel linear. Beachten Sie, daß Koordinatenabbildungen immer von der Wahl der Basis abhängen. So hat beispielweise das Polynom $p(x) = 1 - x^2$ bezüglich der Basis $\{q_0, q_1, q_2\}$ des $\mathcal{P}_2$ mit $q_0(x) = 1, q_1(x) = x, q_2(x) = x^2$ die Koordinaten $1, 0, -1$, aber in der Basis $\{\widehat{q}_0, \widehat{q}_1, \widehat{q}_2\}$ mit $\widehat{q}_0(x) = 1 - x^2, \widehat{q}_1(x) = x, \widehat{q}_2(x) = 1$ die Koordinaten $1, 0, 0$. Traditionell schreibt man übrigens Koordinatentupel nicht zeilenförmig, sondern man ordnet die Koordinaten in einer Spalte an. Das hilft besonders dann, wenn man mit Koordinaten von n-Tupeln arbeitet, da man dann zwischen dem Vektor $\vec{a} = (a_1, \ldots, a_n)$ und seinen Koordinaten in einer bestimmten Tupelbasis optisch unterscheiden kann. So hat z. B. das Zahlenpaar $(1, 2)$ in der kanonischen Basis $(1, 0), (0, 1)$ die Koordinaten $\binom{1}{2}$ aber in der Basis $(1, 1), (1, 0)$ die

Koordinaten $\binom{2}{-1}$. Wenn Sie also einen Spaltenvektor sehen, fragen Sie immer, bezüglich welcher Basis sind diese Koordinaten zu benutzen?
Schauen wir uns aber jetzt noch einige lineare Abbildungen an. So ist etwa die Translationsoperation $T_h : \mathcal{P}_n \to \mathcal{P}_n$ für jedes $n \in \mathbb{R}$ linear, wobei das Polynom $T_h(p)$ gegeben ist durch $[T_h(p)](x) = p(x-h)$, also z. B. für $p(x) = 1 - x^2$

$$[T_h(p)](x) = 1 - (x-h)^2 = 1 - x^2 + 2hx - h^2$$

Die Operation heißt Translation, weil sie den Funktionsgraphen G_p von p um die Distanz h nach rechts verschiebt. (Stellen Sie sich vor, p habe ein Maximum in $x = 0$, dann hat $T_h(p)$ das Maximum an der Stelle $x - h = 0$, also bei $x = h$. Das Maximum, und auch jeder andere Punkt des Graphen haben sich also um h verschoben.)
Um die Linearität nachzuweisen, müssen wir die beiden Bedingungen überprüfen. Zunächst gilt

$$[T_h(\alpha p)](x) = (\alpha p)(x-h) = \alpha p(x-h) = \alpha[T_h(p)](x) = [\alpha T_h(p)](x)$$

und da dies für jedes $x \in \mathbb{R}$ richtig ist, folgt die Gleichheit der Funktionen, $T_h(\alpha p) = \alpha T_h(p)$. Genauso folgt $T_h(p+q) = T_h(p) + T_h(q)$, denn

$$\begin{aligned}(T_h(p+q))(x) &= (p+q)(x-h) \\ &= p(x-h) + q(x-h) = [T_h(p)](x) + [T_h(q)](x).\end{aligned}$$

Sie haben vielleicht gemerkt, daß wir nirgends benutzt haben, daß die Argumente von T_h Polynome sind. Es gilt also allgemein für den Vektorraum V aller reellen Funktionen auf $\mathbb{R}$, daß $T_h : V \to V$ linear ist. Daß auch die Einschränkung von T_h auf $\mathcal{P}_n$ eine lineare Abbildung nach $\mathcal{P}_n$ ist, liegt daran, daß T_h aus einem Polynom wieder ein Polynom macht. Eine andere lineare Abbildung auf $\mathcal{P}_n$ bzw. V ist durch die Punktauswertung von Funktionen gegeben. Und zwar ist $\delta_{\overline{x}}(f) = f(\overline{x})$ offensichtlich eine lineare Abbildung $\delta_{\overline{x}} : V \to \mathbb{R}$, denn

$$\delta_{\overline{x}}(\alpha f) = (\alpha f)(\overline{x}) = \alpha \delta_{\overline{x}}(f)$$

$$\delta_{\overline{x}}(f+g) = (f+g)(\overline{x}) = f(\overline{x}) + g(\overline{x}) = \delta_{\overline{x}}(f) + \delta_{\overline{x}}(g).$$

Mit der Zielmenge $\mathbb{R}$ verlassen wir übrigens nicht das Konzept, daß lineare Abbildungen Vektorräume in Vektorräume abbilden. Die Menge $\mathbb{R}$ kann nämlich auch als reeller Vektorraum aufgefaßt werden, wenn wir die normale Addition und die normale Multiplikation als Vektorraum-Addition bzw. skalare Multiplikation nehmen. Prüfen Sie einfach die

Rechenregeln nach. (Sie können $\mathbb{R}$ auch als Spezialfall von $\mathbb{R}^n$ mit $n = 1$ betrachten.)
Schauen wir uns jetzt einmal eine Abbildung $L : \mathbb{R}^2 \to \mathbb{R}$ an.

$$L(x_1, x_2) = \sqrt{x_1^2 + x_2^2}$$

Ist L linear? Wieder sind nur die beiden Bedingungen zu überprüfen.

$$\begin{aligned} L(\alpha(x_1, x_2)) &= L(\alpha x_1, \alpha x_2) = \sqrt{(\alpha x_1)^2 + (\alpha x_2)^2} \\ &= \sqrt{\alpha^2(x_1^2 + x_2^2)} = |\alpha|\sqrt{x_1^2 + x_2^2} = |\alpha| L(x_1, x_2) \end{aligned}$$

Wir können auf die Überprüfung der zweiten Bedingung also verzichten, da schon die erste nicht erfüllt ist. Geben Sie einfach ein Gegenbeispiel an, um die Sache klar zu machen, etwa

$$L((-1) \cdot (1, 0)) = 1 \neq -1 = (-1)L(1, 0)$$

Das gleiche Problem tritt bei allen Funktionen auf, aus denen man einen skalaren Faktor α nicht einfach „herausziehen“ kann, etwa

$$\begin{aligned} L(x_1, x_2) &= x_1 \cdot x_2, \quad L(x_1, x_2) = \ln(|x_1| + 1), \\ L(x_1, x_2) &= e^{x_1 + x_2}, \quad L(x_1, x_2) = \sin x_1 + \cos x_2. \end{aligned}$$

Diese Funktionen sind also alle *nichtlinear*. Wie sehen denn aber nun lineare Abbildungen von $\mathbb{R}^2$ nach $\mathbb{R}$, oder allgemeiner, von $\mathbb{R}^n$ nach $\mathbb{R}$, aus?
Eine lineare Abbildung von $\mathbb{R}^n \to \mathbb{R}$ ist offensichtlich durch

$$L(\vec{x}) = a_1 x_1 + a_2 x_2 + \ldots + a_n x_n \qquad \vec{x} = (x_1, \ldots, x_n)$$

gegeben, wobei $a_1, \ldots, a_n$ beliebige reelle Zahlen sind. Es gilt nämlich

$$\begin{aligned} L(\alpha\vec{x}) &= a_1(\alpha x_1) + \ldots + a_n(\alpha x_n) \\ &= \alpha(a_1 x_1 + \ldots + a_n x_n) = \alpha L(\vec{x}) \end{aligned}$$

und

$$\begin{aligned} L(\vec{x} + \vec{y}) &= a_1(x_1 + y_1) + \ldots + a_n(x_n + y_n) \\ &= (a_1 x_1 + \ldots + a_n x_n) + (a_1 y_1 + \ldots + a_n y_n) = L(\vec{x}) + L(\vec{y}). \end{aligned}$$

Entsprechend erhält man eine lineare Abbildung $L : \mathbb{R}^n \to \mathbb{R}^m$, indem man obige Konstruktion m-mal wiederholt. Der Funktionswert $\vec{y} = L(\vec{x})$ ist dann gegeben durch

$$\begin{array}{ccccccccc} y_1 & = & a_{11}x_1 & + & a_{12}x_2 & + & \ldots & + & a_{1n}x_n \\ y_2 & = & a_{21}x_1 & + & a_{22}x_2 & + & \ldots & + & a_{2n}x_n \\ \vdots & & \vdots & & \vdots & & & & \vdots \\ y_m & = & a_{m1}x_1 & + & a_{m2}x_2 & + & \ldots & + & a_{mn}x_n \end{array}$$

also letztlich durch die $m \cdot n$ Koeffizienten $a_{ij} \in \mathbb{R}$. Interessanterweise ist diese Konstruktion bereits die allgemeinste Form der linearen Abbildung zwischen endlich dimensionalen Vektorräumen. Immer dann, wenn wir uns die lineare Abbildung in Koordinaten anschauen, sieht sie so aus. Nehmen wir also an, $L : V \to W$ sei eine lineare Abbildung zwischen dem n-dimensionalen Vektorraum V und dem m-dimensionalen Vektorraum W. Um mit Koordinaten arbeiten zu können, müssen wir zunächst Basen in V und W einführen, also etwa $\{\vec{v}_1, \ldots, \vec{v}_n\}$ für V und $\{\vec{w}_1, \ldots, \vec{w}_m\}$ für W. Ein beliebiger Vektor $\vec{v} \in V$ mit Koordinaten $x_1, \ldots, x_n$, also

$$\vec{v} = x_1\vec{v}_1 + x_2\vec{v}_2 + \ldots + x_n\vec{v}_n$$

wird durch die lineare Abbildung auf den Vektor

$$L(\vec{v}) = x_1L(\vec{v}_1) + x_2L(\vec{v}_2) + \ldots + x_nL(\vec{v}_n)$$

abgebildet. Um die Koordinaten von $L(\vec{v})$ zu ermitteln, wenden wir die Koordinatenabbildung $\Phi : W \to \mathbb{R}^m$ auf $L(\vec{v})$ an, wobei wieder wegen der Linearität folgt

$$\Phi(L(\vec{v})) = x_1\Phi(L(\vec{v}_1)) + x_2\Phi(L(\vec{v}_2)) + \ldots + x_n\Phi(L(\vec{v}_n)) \tag{5}$$

Wir sehen jetzt ganz deutlich, wie die Koordinaten des Bildes $L(\vec{v})$ von den Koordinaten des Urbildes $\vec{v}$ abhängen. Es werden einfach die Koordinaten der Bilder der Basisvektoren $\Phi(L(\vec{v}_i))$ mit den Koordinaten x_i des Vektors $\vec{v}$ multipliziert und dann addiert. Insbesondere ist die Wirkung der Abbildung L vollkommen klar, wenn die n-Koordinatentupel $\Phi(L(\vec{v}_i))$ berechnet wurden. Die Auswertung von L mit verschiedenen Vektoren $\vec{v}$ entspricht nur einer Linearkombination dieser n Koordinatentupel. Kurz gesagt, kennt man die Koordinaten der Bilder der Basisvektoren, so kennt man die ganze lineare Abbildung! Rechnen wir die Koordinaten der Bilder der Basisvektoren aus, so erhalten wir n mal m reelle Zahlen

$$\Phi(L(\vec{v}_1)) = \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{pmatrix}, \ldots, \Phi(L(\vec{v}_n)) = \begin{pmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{pmatrix}$$

Das Bild eines Vektors mit Koordinaten $x_1, \ldots, x_n$ hat dann gemäß (5) die Koordinaten $y_1, \ldots, y_m$ mit

$$\begin{array}{ccccccc} y_1 & = & a_{11}x_1 & + & a_{12}x_2 & + \ldots + & a_{1n}x_n \\ y_2 & = & a_{21}x_1 & + & a_{22}x_2 & + \ldots + & a_{2n}x_n \\ \vdots & & \vdots & & \vdots & & \vdots \\ y_m & = & a_{m1}x_1 & + & a_{m2}x_2 & + \ldots + & a_{mn}x_n \end{array}$$

also genau die Form, die wir uns schon als Beispiel einer linearen Abbildung von $\mathbb{R}^n$ nach $\mathbb{R}^m$ angesehen haben.
Da die Zahlen a_{ij} die lineare Abbildung L vollständig beschreiben (sofern die zugehörigen Basen bekannt sind), faßt man sie auch in einem einzigen Objekt, einer sogenannten *Matrix* optisch zusammen

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

Merken Sie sich dabei, daß in der j-ten Spalte die Koordinaten des Bildes des j-ten Basisvektors stehen, oder kurz, in den Spalten stehen die Bilder der Basisvektoren. Schauen wir uns jetzt einmal in einem konkreten Fall an, wie man die Matrix zu einer linearen Abbildung findet, nachdem Basen in Ziel- und Startraum gewählt wurden.
Als lineare Abbildung wählen wir $L : \mathcal{P}_2 \to \mathcal{P}_3$, die dadurch gegeben ist, daß jedes Polynom in $\mathcal{P}_2$ mit einem fest vorgegebenen Polynom $Q(x) = 1 + 2x$ multipliziert wird. Offensichtlich wird dabei aus einem Polynom vom Grad ≤ 2 höchstens ein Polynom vom Grad ≤ 3, so daß $\mathcal{P}_3$ eine sinnvolle Wahl für die Zielmenge ist. Die Linearität der Abbildung weist man wie üblich nach. So gilt

$$\begin{array}{rclcrcl} [L(\alpha p)](x) & = & Q(x)(\alpha p)(x) & = & Q(x)\alpha p(x) & = & \alpha Q(x)p(x) \\ & = & \alpha[L(p)](x) & = & [\alpha L(p)](x) & & \end{array}$$

und entsprechend zeigt man $L(p + q) = L(p) + L(q)$.
Um die Abbildung durch eine Matrix repräsentieren zu können, müssen wir zunächst Basen in $\mathcal{P}_2$ und $\mathcal{P}_3$ einführen, also z. B. $\{q_0, q_1, q_2\}$ für $\mathcal{P}_2$ und $\{q_0, q_1, q_2, q_3\}$ für $\mathcal{P}_3$, wobei $q_n(x) = x^n, n \in \mathbb{N}_0$ die sogenannten Monome sind. Als nächstes müssen wir die Bilder der Basisvektoren von $\mathcal{P}_2$ ausrechnen

$$\begin{array}{rclcrcl} [L(q_0)](x) & = & Q(x)q_0(x) & = & Q(x)1 & = & 1 + 2x \\ [L(q_1)](x) & = & Q(x)q_1(x) & = & Q(x)x & = & x + 2x^2 \\ [L(q_2)](x) & = & Q(x)q_2(x) & = & Q(x)x^2 & = & x^2 + 2x^3 \end{array}$$

Von diesen Bildern werden jetzt die Koordinaten in der gewählten Basis von $\mathcal{P}_3$ benötigt, d. h. wir müssen sie als Linearkombination der Basis $\{q_0, q_1, q_2, q_3\}$ darstellen

$$\begin{array}{rclcrcrcl} L(q_0) & = & 1q_0 & + & 2q_1 & + & 0q_2 & + & 0q_3 \\ L(q_1) & = & 0q_0 & + & 1q_1 & + & 2q_2 & + & 0q_3 \\ L(q_2) & = & 0q_0 & + & 0q_1 & + & 1q_2 & + & 2q_3 \end{array}$$

Die gefundenen Koordinaten werden dann einfach in die Spalten einer Matrix eingetragen

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix}$$

Um nun das Bild des Polynoms $p(x) = 3 - 5x + 2x^2$ auszurechnen, brauchen wir nur noch die Koordinaten $3, -5, 2$ von p in unserer $\mathcal{P}_2$-Basis, um damit die Spalten der Matrix linear zu kombinieren. Das Ergebnis ist

$$3 \begin{pmatrix} 1 \\ 2 \\ 0 \\ 0 \end{pmatrix} - 5 \begin{pmatrix} 0 \\ 1 \\ 2 \\ 0 \end{pmatrix} + 2 \begin{pmatrix} 0 \\ 0 \\ 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \\ -8 \\ 4 \end{pmatrix}$$

Dies sind nun die Koordinaten des Bildes in der von uns gewählten $\mathcal{P}_3$-Basis, d. h. das resultierende Polynom ist

$$L(p) = 3q_0 + 1q_1 - 8q_2 + 4q_3$$

Der Wert dieses Polynoms an einer Stelle $x \in \mathbb{R}$ ist also

$$[L(p)](x) = 3 + x - 8x^2 + 4x^3$$

Beachten Sie, daß die Matrix von der Wahl der Basen abhängt. Hätten wir im Raum $\mathcal{P}_3$ z. B. die „maßgeschneiderte“ Basis $\{\widehat{q}_0, \widehat{q}_1, \widehat{q}_2, \widehat{q}_3\}$ mit

$$\widehat{q}_0(x) = 1,\ \widehat{q}_1(x) = 1 + 2x,\ \widehat{q}_2(x) = x + 2x^2,\ \widehat{q}_3(x) = x^2 + 2x^3$$

gewählt, so ergeben sich andere Koordinaten für die Bilder der Basisvektoren

$$\begin{array}{rcccccccc} L(q_0) & = & 0\widehat{q}_0 & + & 1\widehat{q}_1 & + & 0\widehat{q}_2 & + & 0\widehat{q}_3 \\ L(q_1) & = & 0\widehat{q}_0 & + & 0\widehat{q}_1 & + & 1\widehat{q}_2 & + & 0\widehat{q}_3 \\ L(q_2) & = & 0\widehat{q}_0 & + & 0\widehat{q}_1 & + & 0\widehat{q}_2 & + & 1\widehat{q}_3 \end{array}$$

und damit eine andere Matrix

$$\begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Trotzdem beschreibt diese Matrix immer noch die Abbildung L. Rechnen wir dazu das gleiche Beispiel $L(p)$ mit $p(x) = 3 - 5x + 2x^2$ noch

einmal aus. Wieder müssen wir die Spalten der Matrix mit $3, -5, 2$ linear kombinieren

$$3\begin{pmatrix}0\\1\\0\\0\end{pmatrix} - 5\begin{pmatrix}0\\0\\1\\0\end{pmatrix} + 2\begin{pmatrix}0\\0\\0\\1\end{pmatrix} = \begin{pmatrix}0\\3\\-5\\2\end{pmatrix}$$

Die Ergebniskoordinaten sind jetzt natürlich anders als vorher, da sich ja die Matrix geändert hat. Allerdings beziehen sie sich auch auf eine andere Basis. Das Ergebnispolynom ist dagegen das gleiche, denn

$$\begin{aligned}3\widehat{q}_1(x) - 5\widehat{q}_2(x) + 2\widehat{q}_3(x) &= 3 + 6x - 5x - 10x^2 + 2x^2 + 4x^3 \\ &= 3 + x - 8x^2 + 4x^3 = [L(p)](x)\end{aligned}$$

die Matrix beschreibt also immer noch dieselbe Abbildung, sie bezieht sich nur auf eine andere Basiswahl. Das Gleiche passiert übrigens, wenn wir die Basis des Ausgangsraums $\mathcal{P}_2$ ändern. Wählen wir z. B. $\{\widetilde{q}_0, \widetilde{q}_1, \widetilde{q}_2\}$ mit

$$\widetilde{q}_0(x) = 1 - x, \quad \widetilde{q}_1(x) = 1 + x, \quad \widetilde{q}_2(x) = x^2.$$

Wie prüft man eigentlich nach, daß $E = \{\widetilde{q}_0, \widetilde{q}_1, \widetilde{q}_2\}$ eine Basis ist? Da wir schon wissen, daß $\dim \mathcal{P}_2 = 3$ ist, genügt es zu zeigen, daß E ein Erzeugendensystem darstellt. Ist dies der Fall, dann muß E auch minimal sein, denn sonst könnte man durch Weglassen von Elementen eine Basis konstruieren, die weniger als drei Elemente hat und das geht ja in einem Raum der Dimension drei nicht. Zum Nachweis, daß E ein Erzeugendensystem ist, wählen wir ein beliebiges Polynom $q(x) = a_0 + a_1 x + a_2 x^2$ in $\mathcal{P}_2$. Offensichtlich gilt

$$\begin{aligned}q &= a_0\frac{\widetilde{q}_0 + \widetilde{q}_1}{2} + a_1\frac{\widetilde{q}_1 - \widetilde{q}_0}{2} + a_2\widetilde{q}_2 \\ &= \frac{a_0 - a_1}{2}\widetilde{q}_0 + \frac{a_0 + a_1}{2}\widetilde{q}_1 + a_2\widetilde{q}_2\end{aligned}$$

Als Basis für den Zielraum $\mathcal{P}_3$ wählen wir die Monome $\{q_0, q_1, q_2, q_3\}$. Zur Bestimmung der Matrix benötigen wir zunächst die Bilder der Basisvektoren

$$\begin{aligned}[L(\widetilde{q}_0)](x) &= Q(x)\widetilde{q}_0(x) = (1 + 2x)(1 - x) = 1 + x - 2x^2 \\ [L(\widetilde{q}_1)](x) &= Q(x)\widetilde{q}_1(x) = (1 + 2x)(1 + x) = 1 + 3x + 2x^2 \\ [L(\widetilde{q}_2)](x) &= Q(x)\widetilde{q}_2(x) = (1 + 2x)x^2 = x^2 + 2x^3\end{aligned}$$

Da man die Koordinaten in der Monombasis aus dieser Darstellung sofort ablesen kann, erhalten wir die Matrixdarstellung

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 3 & 0 \\ -2 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix}$$

Auch hier beschreibt die Matrix natürlich wieder die Abbildung L, was wir an unserem Beispiel mit $p(x) = 3 - 5x + 2x^2$ noch einmal nachvollziehen sollen. Zunächst brauchen wir dazu die Koordinaten des Polynoms p in unserer Basis. $\{\widetilde{q}_0, \widetilde{q}_1, \widetilde{q}_2\}$. Nach unserer obigen Rechnung folgt mit $a_0 = 3, a_1 = -5, a_2 = 2$, daß die Koordinaten durch

$$\frac{a_0 - a_1}{2} = 4, \quad \frac{a_0 + a_1}{2} = -1, \quad a_2 = 2$$

gegeben sind. Eine entsprechende Linearkombination der Matrixspalten liefert

$$4 \begin{pmatrix} 1 \\ 1 \\ -2 \\ 0 \end{pmatrix} + (-1) \begin{pmatrix} 1 \\ 3 \\ 2 \\ 0 \end{pmatrix} + 2 \begin{pmatrix} 0 \\ 0 \\ 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \\ -8 \\ 4 \end{pmatrix}$$

wobei die Ergebniskoordinaten bezüglich der Monombasis zu verstehen sind, also zum Polynom

$$3q_0 + q_1 - 8q_2 + 4q_3 = L(p)$$

was dem bekannten Ergebnis entspricht.
Sie sehen also, daß, wenn Ihnen jemand eine Matrix als Vertreter einer linearen Abbildung zwischen zwei endlich dimensionalen Vektorräumen anbieten will, so ist die Information unbrauchbar, falls Sie nicht auch die zugehörigen Basen genannt bekommen. Während sie also bei einem Koordinatenvektor stets fragen sollten, auf welche Basis er sich bezieht, müssen Sie bei einer Matrix fragen, auf welche *Basen* sie sich bezieht. Vielleicht werden Sie jetzt einwenden, daß man doch eine Basis in jedem Vektorraum ein für allemal auszeichnen könnte, z. B. die Basis $\{(1,0),(0,1)\}$ des $\mathbb{R}^2$, die Basis $\{(1,0,0),(0,1,0),(0,0,1)\}$ des $\mathbb{R}^3$, oder die Monombasis $\{q_0, q_1, q_2\}$ des $\mathcal{P}_2$usw. Ist das denn nicht die einfachste Wahl? Wer will denn überhaupt die Basis $\{\widetilde{q}_0, \widetilde{q}_1, \widetilde{q}_2\}$ des $\mathcal{P}_2$ oder $\{\widehat{q}_0, \widehat{q}_1, \widehat{q}_2, \widehat{q}_3\}$ des $\mathcal{P}_3$ benutzen? Nun, wir werden sehen, daß durch die Wahl einer günstigen Basis Berechnungen sehr stark vereinfacht werden

können. Was dabei günstig oder ungünstig ist, hängt vom betrachteten Problem ab und kann eben nicht unabhängig davon entschieden werden.

4. Matrizenrechnung

Im vorangegangenen Abschnitt haben wir gesehen, daß das Arbeiten mit linearen Abbildungen zwischen endlich dimensionalen Vektorräumen auf das Rechnen mit Matrizen führt, sobald wir Basen in den beteiligten Vektorräumen einführen.
In vielen Anwendungen werden Abbildungen natürlicher Weise miteinander verknüpft, z. B. durch Addition oder Hintereinanderausführung, wobei die resultierende Abbildung wieder linear ist. Entsprechend ergeben sich Verknüpfungen von Matrizen, die wir in diesem Abschnitt genauer untersuchen wollen. Schauen wir uns aber zunächst ein praktisches Beispiel an, bei dem Verknüpfungen linearer Abbildungen auftreten. Eine Bäckerei bietet eine bestimmte Produktpalette an: Brötchen, Weißbrot, Roggenbrot, Sauerteigbrot, Kekse, ... Zur Herstellung jedes Produkts werden bestimmte Mengen von Zutaten benötigt, wobei die Zutatenliste aus Weizenmehl, Roggenmehl, Hefe, Sauerteig, Backfett, Milch, Wasser, Zucker, Salz etc. besteht. Nehmen wir an, daß n Produkte angeboten werden, wobei insgesamt m Zutaten zum Einsatz kommen. Eine Bestellung kann man dann durch $\vec{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ beschreiben. Die Komponente x_i gibt dabei an, wieviel kg des Produkts i gewünscht sind. Die zur Fertigstellung der Bestellung $\vec{x}$ benötigte Zutatenmenge bezeichnen wir mit $Z(\vec{x}) = \vec{y} \in \mathbb{R}^m$. Hier gibt y_j an, wieviel kg von der Zutat j benötigt werden. Ist diese Zutatenabbildung Z linear? Tatsächlich ist das der Fall, da eine α-fache Bestellung auch die α-fache Zutatenmenge benötigt und wenn zwei Bestellungen $\vec{x}$ und $\vec{u}$ hereinkommen, dann erfordert die Fertigstellung der gesamten Anforderung $\vec{x} + \vec{u}$ auch die Summe der Zutaten $Z(\vec{x}) + Z(\vec{u})$ der Einzelbestellungen. Es gilt also

$$Z(\alpha\vec{x}) = \alpha Z(\vec{x}), \qquad Z(\vec{u} + \vec{x}) = Z(\vec{u}) + Z(\vec{x}).$$

Obwohl das alles sehr überzeugend klingt, wäre es doch falsch, an dieser Stelle zu behaupten, daß $Z : \mathbb{R}^n \to \mathbb{R}^m$ eine lineare Abbildung ist. Tatsächlich haben wir Z noch gar nicht für alle $\vec{x} \in \mathbb{R}^n$ definiert. Was ist denn z. B. $Z(-2, 0, 0 \ldots)$? Will hier der Kunde 2 kg Brötchen zurückgeben und soll das Ergebnis die Menge der darin verbackenen Zutaten sein? Unsere Abbildung Z hat nämlich zunächst nur eine Bedeutung für $\vec{x} = (x_1, \ldots, x_1)$ mit $x_i \geq 0$ und entsprechend ist $Z(\alpha\vec{x})$ nur sinnvoll, für $\alpha \geq 0$. Um dennoch in den „Genuß“ der Linearität zu kommen, *definieren* wir Z einfach so, daß die Abbildung linear ist. Dabei

hilft uns die Tatsache, daß $\mathbb{R}^n$ eine Basis aus sinnvollen „Bestellungs-Vektoren“ besitzt. Nehmen wir z. B. die kanonische Basis $(\vec{e}_1, \ldots, \vec{e}_n)$. Der Basisvektor $\vec{e}_i$ steht dann für die Bestellung von 1 kg des Produktes i und 0 kg aller anderen Produkte. Das Bild dieses Basisvektors $Z(\vec{e}_i)$ kann man also als Rezept für das Produkt i betrachten. Für eine beliebige Bestellung $\vec{x} = (x_1, \ldots, x_n)$ gilt dann

$$Z(\vec{x}) = x_1 Z(\vec{e}_1) + \ldots + x_n Z(\vec{e}_n)$$

und nun können wir diesen Zusammenhang einfach für *alle* $\vec{x} \in \mathbb{R}^n$ definieren. Damit ist dann $Z : \mathbb{R}^n \to \mathbb{R}^m$ linear. Natürlich werden wir in der Anwendung Z nie für unsinnige Bestellungen ausrechnen.
Wenn wir die kanonischen Basen von $\mathbb{R}^n$ und $\mathbb{R}^m$ wählen, können wir Z durch eine Matrix darstellen, wobei in den Spalten die Bilder der Basisvektoren (also die Rezepte) stehen.
Wird für 1 kg des Produkts i $\mathcal{Z}_{ij}$ kg der Zutat j benötigt, so ist die Matrix gerade durch

$$\mathcal{Z} = \begin{pmatrix} \mathcal{Z}_{11} & \ldots & \mathcal{Z}_{1n} \\ \vdots & & \vdots \\ \mathcal{Z}_{m1} & \ldots & \mathcal{Z}_{mn} \end{pmatrix} \in \mathbb{R}^{m \times n}$$

gegeben.
Entsprechend läßt sich eine Kostenfunktion $K : \mathbb{R}^m \to \mathbb{R}$ einführen, wobei $K(\vec{y})$ die Kosten sind, die durch die Zutaten $\vec{y}$ entstehen. Die Matrixdarstellung von K führt auf

$$\mathcal{K} = (\mathcal{K}_1 \ldots \mathcal{K}_m) \in \mathbb{R}^{1 \times m}$$

wobei $\mathcal{K}_j$ der Preis für 1 kg der Zutat j ist. Für den Gesamtpreis einer Bestellung ist aber auch der Arbeitsaufwand und die gewünschte Gewinnspanne wichtig.
Nehmen wir einfachheitshalber an, daß die Herstellung von 2 kg Brötchen doppelt so lange dauert wie die Herstellung von 1 kg (dabei lassen wir mögliche Zeiten außer Acht, die unabhängig von der Menge anfallen, wie z. B. das Reinigen der Geräte etc.). Unter dieser Annahme ist auch die Arbeitskosten-Abbildung $A : \mathbb{R}^n \to \mathbb{R}$ linear mit zugehöriger Matrix

$$\mathcal{A} = (\mathcal{A}_1 \ldots \mathcal{A}_n)$$

wobei $\mathcal{A}_i$ die anfallenden Lohnkosten für die Herstellung von 1 kg des Produkts i darstellt. Insgesamt belaufen sich die Fixkosten für eine Bestellung $\vec{x}$ damit auf

$$K(Z(\vec{x})) + A(\vec{x})$$

Will der Produzent nun p Prozent Umsatzrendite erwirtschaften, so muß er für die Bestellung den Gesamtpreis

$$U(\vec{x}) = \frac{1}{1-p}(K(Z(\vec{x})) + A(\vec{x}))$$

verlangen. Der Gewinn $G(\vec{x})$ nach Abzug der Kosten ist nämlich

$$\begin{aligned} G(\vec{x}) = U(\vec{x}) - (K(Z(\vec{x}))) + A(\vec{x}) &= \left(\frac{1}{1-p} - 1\right)\left(K(Z(\vec{x})) + A(\vec{x})\right) \\ &= p\frac{1}{1-p}(K(Z(\vec{x})) + A(\vec{x})) = pU(\vec{x}) \end{aligned}$$

Sie sehen, daß bei diesen Berechnungen die konkrete Bestellung $\vec{x}$ selbst völlig unerheblich ist. Vielmehr sind die Beziehungen für alle Bestellungen gültig – wir rechnen eben mit Kosten-Abbildungen und nicht mit einer bestimmten Bestellung. In diesem Sinne unterdrücken wir einfach die Angabe des Arguments $\vec{x}$ und schreiben etwa

$$U = \frac{1}{1-p}(K \circ Z + A)$$

wobei $K \circ Z$ die Hintereinanderausführung der Abbildungen K und Z andeutet. Die Summe zweier Abbildungen $K \circ Z + A$ ist im obigen Sinne durch

$$(K \circ Z + A)(\vec{x}) = K \circ Z(\vec{x}) + A(\vec{x})$$

gegeben, d. h. das Bild einer Summe von Abbildungen ist die Summe der Bilder. Genauso definiert man das Vielfache einer Abbildung durch das entsprechende Vielfache des Funktionswerts, etwa

$$\left(\frac{1}{1-p}A\right)(\vec{x}) = \frac{1}{1-p}A(\vec{x}).$$

In diesem Sinne ist $U = \frac{1}{1-p}(K \circ Z + A)$ zu verstehen. Wie wir sehen werden, ist die Verknüpfung von linearen Abbildungen wieder eine lineare Abbildung und folglich läßt sich $U : \mathbb{R}^n \to \mathbb{R}$ auch direkt durch eine Matrix $\mathcal{U} = (\mathcal{U}_1 \ldots \mathcal{U}_n)$ beschreiben. Genau diese Matrixeinträge finden Sie übrigens in der Bäckerei neben der ausgestellten Ware, denn $\mathcal{U}_i$ ist ja gerade der Endpreis für das Produkt i (pro kg). Wie sich die Matrix $\mathcal{U}$ aus einer Verknüpfung der Rezeptmatrix $\mathcal{Z}$, der Zutaten-Preismatrix $\mathcal{R}$, der Lohnmatrix $\mathcal{A}$ und der Umsatzrendite p berechnen läßt, wollen wir uns nun im allgemeinen Fall linearer Abbildungen genauer ansehen. Seien dazu L, M beliebige lineare Abbildungen vom Vektorraum V in den Vektorraum W. Definieren wir die Summe der beiden Abbildungen $L + M : V \to W$ durch

$$(L + M)(\vec{v}) = L(\vec{v}) + M(\vec{v}), \quad \vec{v} \in V$$

und das α Vielfache $\alpha L : V \to W$ durch

$$(\alpha L)(\vec{v}) = \alpha L(\vec{v})$$

so bildet die Menge

$$\text{Hom}\,(V,W) = \{L : V \to W | L \text{ linear}\}$$

wieder einen Vektorraum.
Dazu müssen wir zunächst überprüfen, ob $L+M$ und αL wieder lineare Abbildungen sind, denn von den Operationen Addition und skalare Multiplikation wird ja bei einem allgemeinen Vektorraum Z verlangt, daß die Zielmenge jeweils Z ist. Es ist

$$(\alpha L)(\beta\vec{v}) = \alpha L(\beta\vec{v}) = \alpha\beta L(\vec{v}) = \beta\alpha L(\vec{v}) = \beta(\alpha L)(\vec{v})$$

und

$$\begin{aligned}(\alpha L)(\vec{v}+\vec{u}) &= \alpha L(\vec{v}+\vec{u}) = \alpha(L(\vec{v}) + L(\vec{u})) \\ &= \alpha L(\vec{v}) + \alpha L(\vec{u}) = (\alpha L)(\vec{v}) + (\alpha L)(\vec{u})\end{aligned}$$

und damit ist αL tatsächlich linear. Genauso zeigt man, daß $L + M$ eine lineare Abbildung ist. Zum Nachweis, daß Hom (V, W) tatsächlich ein Vektorraum ist müssen jetzt nur noch die Vektorraumrechenregeln nachgeprüft werden. Die Bezeichnung Hom stammt übrigens daher, daß lineare Abbildungen auch als *Homomorphismen* bezeichnet werden.
Neben der Summe und dem Vielfachen von linearen Abbildungen ist auch die Verkettung von linearen Abbildungen wieder linear. Sind U, V und W Vektorräume und $K \in \text{Hom}\,(U, V), L \in \text{Hom}\,(V, W)$, so kann man L und K „hintereinanderschalten“, d. h. eine Abbildung $L \circ K$ von U nach W konstruieren gemäß der Vorschrift

$$(L \circ K)(\vec{u}) = L(K(\vec{u})) \qquad \vec{u} \in U.$$

Diese Verkettung ist wieder linear, denn

$$\begin{aligned}(L \circ K)(\alpha\vec{u}) &= L(K(\alpha\vec{u})) \;=\; L(\alpha[K(\vec{u})]) \;=\; \alpha L(K(\vec{u})) \\ &= \alpha(L \circ K)(\vec{u})\end{aligned}$$

und

$$\begin{aligned}(L \circ K)(\vec{u}_1 + \vec{u}_2) &= L(K(\vec{u}_1 + \vec{u}_2)) = L(K(\vec{u}_1) + K(u_2)) \\ &= L(K(\vec{u}_1)) + L(K(\vec{u}_2)) = (L \circ K)(\vec{u}_1) + (L \circ K)(\vec{u}_2)\end{aligned}$$

Es gilt also $L \circ K \in \text{Hom}\,(U, W)$. Wenn wir nun in allen beteiligten Vektorräumen Basen einführen, so lassen sich die Operationen $\alpha L, L + M$ und $L \circ K$ in entsprechende Matrixoperationen übersetzen.
Seien dazu $\{\vec{u}_1, \ldots, \vec{u}_r\}, \{\vec{v}_1, \ldots, \vec{v}_n\}$ und $\{\vec{w}_1, \ldots, \vec{w}_m\}$ Basen von U, V und W. Die Abbildung K läßt sich dann durch eine $n \times r$ Matrix repräsentieren (n Zeilen, r Spalten)

$$C = \begin{pmatrix} c_{11} & \dots & c_{1r} \\ \vdots & & \vdots \\ c_{n1} & \dots & c_{nr} \end{pmatrix}$$

wobei in den Spalten die Koordinaten von $K(\vec{u}_1), \dots, K(\vec{u}_r)$ stehen. Entsprechend gibt es $m \times n$ Matrizen A, B zu $L, M \in \text{Hom}\,(V, W)$

$$A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix}, \qquad B = \begin{pmatrix} b_{11} & \dots & b_{1n} \\ \vdots & & \vdots \\ b_{m1} & \dots & b_{mn} \end{pmatrix}$$

Die Matrix zur linearen Abbildung αL erhält man nun dadurch, daß man die Koordinaten von $(\alpha L)(\vec{v}_i)$ in einer Matrix anordnet. Da aber $(\alpha L)(\vec{v}_i) = \alpha L(\vec{v}_i)$ ist, bedeutet dies gerade daß man alle Koordinaten von $L(\vec{v}_i)$ mit α multiplizieren muß. Da die Koordinaten von $L(\vec{v}_i)$ in der Matrix A ablesbar sind, läßt sich die Matrix zu αL leicht aus der Matrix zu L konstruieren. Wir definieren dazu allgemein

$$\alpha \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} = \begin{pmatrix} \alpha a_{11} & \dots & \alpha a_{1n} \\ \vdots & & \vdots \\ \alpha a_{m1} & \dots & \alpha a_{mn} \end{pmatrix}$$

Damit ist die Matrix zu αL gerade αA, wenn A die Matrix zu L ist. Entsprechend gewinnt man die Matrix zur linearen Abbildung $L + M$ aus den einzelnen Matrizen A und B. Die Spalteneinträge sind die Koordinaten von $(L+M)(\vec{v}_i) = L(\vec{v}_i) + M(\vec{v}_i)$, also gerade die Vektorsummen entsprechender Spalten von A und B. Definieren wir also

$$\begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} + \begin{pmatrix} b_{11} & \dots & b_{1n} \\ \vdots & & \vdots \\ b_{m1} & \dots & b_{mn} \end{pmatrix} = \begin{pmatrix} a_{11} + b_{11} & \dots & a_{1n} + b_{1n} \\ \vdots & & \vdots \\ a_{m1} + b_{m1} & \dots & a_{mn} + b_{mn} \end{pmatrix}$$

so ist die Matrix zu $L + M$ gerade $A + B$, wenn A, B die Matrizen zu L, M sind. Wie man leicht nachprüft, haben wir damit schon wieder eine Vektorraumstruktur gefunden, diesmal auf der Menge der $m \times n$ Matrizen. Wir bezeichnen die Menge als $\mathbb{R}^{m \times n}$, wobei der erste Index m die Anzahl der Zeilen und der zweite Index n die Anzahl der Spalten angibt. Welche Dimension hat der Vektorraum $\mathbb{R}^{m \times n}$? Um diese Frage zu beantworten, müssen wir eine Basis von $\mathbb{R}^{m \times n}$ finden.
Die erste Wahl fällt hier wohl auf die kanonische Basis

$$\{E_{ij} | i = 1, \dots, m, j = 1, \dots, n\}$$

wobei E_{ij} eine Matrix ist, die in allen Einträgen eine Null hat bis auf die j-te Spalte der i-ten Zeile, wo sie eine Eins trägt, also etwa in $\mathbb{R}^{2\times 2}$

$$E_{11} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, E_{12} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, E_{21} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, E_{22} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

Wie man leicht nachprüft, bilden diese Matrizen ein Erzeugendensystem und sind linear unabhängig, da es nur die triviale Darsellung der Nullmatrix (alle Einträge = Null) gibt. Da die Basis $m \cdot n$ Elemente hat, gilt also dim $\mathbb{R}^{m\times n} = m \cdot n$.
Genauso kann man übrigens eine Basis des Raumes Hom (V, W) konstruieren. Wir wählen einfach zu E_{ij} gehörig lineare Abbildungen $b_{ij} \in$ Hom (V, W). Nehmen wir dazu an, daß dim $V = n$ und dim $W = m$, d. h. es gibt Basen $(\vec{v}_1, \ldots, \vec{v}_n)$ und $(\vec{w}_1, \ldots, \vec{w}_m)$ von V und W. Wie muß eine Abbildung $b_{ij} : V \to W$ aussehen, damit sie in den gewählten Basen die Matrixdarstellung E_{ij} hat? Die Matrix nennt uns doch die W-Koordinaten der Bilder der V-Basisvektoren. In der k-ten Spalte von E_{ij} finden wir also die Koordinaten von $b_{ij}(\vec{v}_k)$ in der $\vec{w}_l$ Basis. Insbesondere bildet $b_{ij}(\vec{v}_k)$ auf $\vec{0}$ ab für $k \neq j$ und genau auf $\vec{w}_i$ im Fall $k = j$. Wir kennen damit alle Bilder der Basisvektoren und da b_{ij} linear sein soll, folgt für einen beliebigen Vektor $\vec{v} = x_1\vec{v}_1 + \ldots + x_n\vec{v}_n$

$$b_{ij}(\vec{v}) = x_1 b(\vec{v}_1) + \ldots + x_n b(\vec{v}_n) = x_j \vec{w}_i$$

Für die so definierten Abbildungen b_{ij} kann man nun die Basiseigenschaft zeigen. Es folgt damit

$$\text{dim Hom } (V, W) = \text{dim } V \cdot \text{dim } W$$

Die dritte wichtige Matrixoperation entsteht durch das Übersetzen der Hintereinanderausführung linearer Abbildungen in Matrixschreibweise. Die Frage ist, wie kann die Matrix der Verkettung $L \circ K$ durch die Matrizen A, C der beteiligten Abbildungen L und K berechnet werden. Auf jeden Fall wissen wir, daß in den Spalten der Ergebnismatrix die Koordinaten von $(L \circ K)(\vec{u}_j) = L(K(\vec{u}_j))$ stehen müssen. Da aber $K(\vec{u}_j)$ in der Basis $\{\vec{v}_1, \ldots, \vec{v}_n\}$ gerade durch

$$K(\vec{u}_j) = c_{1j}\vec{v}_1 + c_{2j}\vec{v}_2 + \ldots + c_{nj}\vec{v}_n$$

gegeben ist, erhalten wir insgesamt

$$L(K(\vec{u}_j)) = c_{1j}L(\vec{v}_1) + c_{2j}L(\vec{v}_2) + \ldots + c_{nj}L(\vec{v}_n)$$

Da die Koordinaten von $L(\vec{v}_k)$ aber gerade die Spalten von A bilden, bedeutet dies, daß wir die Spalten von A mit den Einträgen der j-ten Spalte von C linear kombinieren müssen, um die j-te Spalte der Matrix zu $L \circ K$ zu bekommen.

Da an der i-ten Stelle in den Spalten $L(\vec{v}_k)$ die Werte a_{ik} stehen, ergibt sich somit für den Eintrag an der Stelle (i,j) der Ergebnismatrix

$$c_{1j}a_{i1} + c_{2j}a_{i2} + \ldots + c_{nj}a_{in} = \sum_{k=1}^{n} a_{ik}c_{kj}$$

Um den Ursprung der Operation in der Hintereinanderausführung von Abbildungen anzudeuten, nennt man die Ergebnismatrix AC und da die Operation das Produkt von Matrixelementen beinhaltet, spricht man vom *Matrixprodukt*. Die Berechnungsvorschrift lautet im Detail

$$\begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \ldots & a_{mn} \end{pmatrix} \begin{pmatrix} c_{11} & \ldots & c_{1r} \\ \vdots & & \vdots \\ c_{n1} & \ldots & c_{nr} \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{n} a_{1k}c_{k1} & \ldots & \sum_{k=1}^{n} a_{1k}c_{kr} \\ \vdots & & \vdots \\ \sum_{k=1}^{n} a_{mk}c_{k1} & \ldots & \sum_{k=1}^{n} a_{mk}c_{kr} \end{pmatrix}$$

Aus dem Produkt einer $m \times n$ Matrix mit einer $n \times r$ Matrix entsteht also eine $m \times r$ Matrix, was zu erwarten war, da AC ja die Matrix zur Abbildung $L \circ K$ von einem r-dimensionalen Raum in einen m-dimensionalen Raum ist.
An dieser Stelle wird übrigens ein weiterer Vorteil der Darstellung von Koordinatentupeln als Spaltenvektoren deutlich. Ein Spaltenvektor ist in unserer neuen Sichtweise ein Element von $\mathbb{R}^{n\times 1}$ (n Zeilen, 1 Spalte). Die Multiplikation einer solchen Spalte mit einer $m \times n$ Matrix bewirkt, wie wir gesehen haben, gerade eine Linearkombination der Spalten

$$\begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \ldots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = x_1 \begin{pmatrix} a_{11} \\ \vdots \\ a_{m1} \end{pmatrix} + \ldots + x_n \begin{pmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{pmatrix}$$

und diese Operation wird ja benötigt, wenn wir die Wirkung der zur Matrix gehörenden linearen Abbildung auf den Vektor mit den Koordinaten $x_1, \ldots, x_n$ berechnen wollen (dieses spezielle Matrixprodukt heißt auch Matrix-Vektor-Produkt). Anders ausgedrückt, ist $\vec{x}$ der Koordinatenvektor zu einem Element $\vec{v} \in V$ bezüglich der Basis $\{\vec{v}_1, \ldots, \vec{v}_n\}$ von V und ist A die Matrix zu $L \in \text{Hom}\,(V,W)$ und den Basen $\{\vec{v}_1, \ldots, \vec{v}_n\}, \{\vec{w}_1, \ldots, \vec{w}_m\}$, so ist $\vec{y} = A\vec{x}$ der Koordinatenvektor von $L(\vec{v})$ in der Basis $\{\vec{w}_1, \ldots, \vec{w}_m\}$ von W. Sobald also die Basen für ein gegebenes Problem fest gewählt sind, kann man statt mit Vektoren $\vec{v} \in V, \vec{w} \in W$ einfach mit den zugehörigen Koordinatenvektoren in $\mathbb{R}^{n\times 1}$ bzw. $\mathbb{R}^{m\times 1}$ rechnen, wobei lineare Abbildungen durch Matrizen ersetzt werden, die Anwendung von linearen Abbildungen durch

Matrix-Vektor-Produkte und die Verkettung von Abbildungen durch die Matrixprodukte. Beachten Sie, welch starke Vereinheitlichung dies für Berechnungen in Vektorräumen bedeutet: Egal, ob der Vektorraum aus Keiselkompaß-Zeigern, Polynomen, Drehstreckungen in der Ebene, n-Tupeln, Bestellungen, Zutatenlisten oder Matrizen besteht, die Durchführung der Vektorraumoperationen und die Anwendung von linearen Abbildungen läßt sich immer mit Koordinatenvektoren und Matrizen bewerkstelligen. Schauen wir uns dazu zum Abschluß einige Beispiele an. Wir beginnen mit einer linearen Abbildung von $\mathbb{R}^2$ nach $\mathbb{R}^3$, die durch

$$L(u_1, u_2) = (u_1 + 2u_2, -\pi u_1, 3u_1 - u_2)(u_1, u_2) \in \mathbb{R}^2$$

gegeben ist. Wählt man nun die kanonischen Basen für $\mathbb{R}^2$ und $\mathbb{R}^3$, so ist die zugehörige Matrix A sehr einfach zu bestimmen. Man braucht nur die Bilder von $(1, 0)$ und $(0, 1)$ auszurechnen und das Ergebnis in die Spalten einzutragen (die Vektoren sind hier identisch mit ihren Koordinatenvektoren bis auf die optische Anordnung).

$$A = \begin{pmatrix} 1 & 2 \\ -\pi & 0 \\ 3 & -1 \end{pmatrix}$$

Möchte man $L(\sqrt{2}, 1/\sqrt{2})$ ausrechnen, kann man dies entweder direkt, oder aber einheitlich als Matrixvektorprodukt durchführen

$$\begin{pmatrix} 1 & 2 \\ -\pi & 0 \\ 3 & -1 \end{pmatrix} \begin{pmatrix} \sqrt{2} \\ \frac{1}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} \sqrt{2} + \frac{2}{\sqrt{2}} \\ -\pi\sqrt{2} \\ 3\sqrt{2} - \frac{1}{\sqrt{2}} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 4 \\ -2\pi \\ 5 \end{pmatrix}.$$

Statt nach der Berechnung den Faktor $\frac{1}{\sqrt{2}}$ vorzuziehen, kann man das bei linearen Abbildungen auch vorher tun, denn skalare Faktoren dürfen ja aus dem Argument nach vorne gezogen werden.

$$\begin{pmatrix} 1 & 2 \\ -\pi & 0 \\ 3 & -1 \end{pmatrix} \begin{pmatrix} \sqrt{2} \\ \frac{1}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ -\pi & 0 \\ 3 & -1 \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 2 \\ -\pi & 0 \\ 3 & -1 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$
$$= \frac{1}{\sqrt{2}} \begin{pmatrix} 4 \\ -2\pi \\ 5 \end{pmatrix}$$

Es lohnt sich also, so früh wie möglich zu vereinfachen. Genauso gilt für die lineare Abbildung $M : \mathbb{R}^2 \to \mathbb{R}^3$

$$M(u_1, u_2) = (4u_1 + 4u_2, 8u_2, -16u_1 + 4u_2)$$

mit der Matrix (bezüglich kanonischen Basen)

$$B = \begin{pmatrix} 4 & 4 \\ 0 & 8 \\ -16 & 4 \end{pmatrix}$$

daß man vor Berechnungen den gemeinsamen Faktor 4 vorziehen kann. Dabei nutzen wir übrigens die Tatsache, daß $\mathbb{R}^{3\times 2}$ als reeller Vektorraum betrachtet werden kann, wobei das Skalarprodukt komponentenweise zu verstehen ist, also

$$\begin{pmatrix} 4 & 4 \\ 0 & 8 \\ -16 & 4 \end{pmatrix} = 4 \begin{pmatrix} 1 & 1 \\ 0 & 2 \\ -4 & 1 \end{pmatrix}$$

Die Anwendung von L auf einen Vektor $(-1, 1)$ kann man damit durch

$$4 \begin{pmatrix} 1 & 1 \\ 0 & 2 \\ -4 & 1 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = 4 \begin{pmatrix} 0 \\ 2 \\ 5 \end{pmatrix} = \begin{pmatrix} 0 \\ 8 \\ 20 \end{pmatrix}$$

berechnen. Die Matrix zu einer kombinierten Abbildung, z. B. $-2L + \frac{1}{2}M$, berechnet sich schließlich durch Linearkombination der Matrizen

$$-2A + \frac{1}{2}B = \begin{pmatrix} -2 & -4 \\ 2\pi & 0 \\ -6 & 2 \end{pmatrix} + \begin{pmatrix} 2 & 2 \\ 0 & 4 \\ -8 & 2 \end{pmatrix} = \begin{pmatrix} 0 & -2 \\ 2\pi & 4 \\ -13 & 4 \end{pmatrix}$$

Um das Matrixprodukt in einem konkreten Fall zu sehen, benutzen wir die 2×2 Matrix

$$C = \begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix}$$

die z. B. zur linearen Abbildung $K : \mathbb{R}^2 \to \mathbb{R}^2$

$$K(x_1, x_2) = (x_1 + 2x_2, -2x_1 + x_2)$$

gehört, falls wir jeweils die kanonische Basis zugrunde legen. Das Produkt

$$AC = \begin{pmatrix} 1 & 2 \\ -\pi & 0 \\ 3 & -1 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix}$$

kann man z. B. spaltenweise bestimmen, wobei in der *ersten* Spalte von AC das Matrix-Vektor-Produkt von A mit der *ersten* Spalte von C erscheint

$$\begin{pmatrix} 1 & 2 \\ -\pi & 0 \\ 3 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ -2 \end{pmatrix} = \begin{pmatrix} -3 \\ -\pi \\ 5 \end{pmatrix}$$

und in der *zweiten* Spalte das Produkt von A mit der *zweiten* Spalte von C

$$\begin{pmatrix} 1 & 2 \\ -\pi & 0 \\ 3 & -1 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 4 \\ -2\pi \\ 5 \end{pmatrix}$$

Das Ergebnis ist also

$$AC = \begin{pmatrix} -3 & 4 \\ -\pi & -2\pi \\ 5 & 5 \end{pmatrix}.$$

Übrigens, wenn Sie beim Matrix-Vektor-Produkt zum Verrechnen neigen, schreiben Sie einfach mehr Zwischenschritte auf, also etwa

$$\begin{aligned} \begin{pmatrix} 1 & 2 \\ -\pi & 0 \\ 3 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ -2 \end{pmatrix} &= 1 \cdot \begin{pmatrix} 1 \\ -\pi \\ 3 \end{pmatrix} + (-2) \begin{pmatrix} 2 \\ 0 \\ -1 \end{pmatrix} \\ &= \begin{pmatrix} 1 \\ -\pi \\ 3 \end{pmatrix} + \begin{pmatrix} -4 \\ 0 \\ 2 \end{pmatrix} = \begin{pmatrix} -3 \\ -\pi \\ 5 \end{pmatrix} \end{aligned}$$

Ist $D \in \mathbb{R}^{2\times 2}$ eine weitere Matrix

$$D = \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}$$

so gibt es übrigens zwei Möglichkeiten für das Matrixprodukt von D und C, nämlich

$$CD = \begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 3 \\ -4 & -1 \end{pmatrix}$$

und

$$DC = \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 5 \\ -2 & 1 \end{pmatrix}$$

Sie sehen also $CD \neq DC$, d. h. das Matrixprodukt ist *nicht* kommutativ. Natürlich heißt das nicht, daß nicht hin und wieder doch mal die Matrizen in einem Produkt vertauscht werden können. Betrachten Sie z. B. die spezielle Matrix

$$E = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \in \mathbb{R}^{2\times 2}$$

Hier gilt sogar für *jede* Matrix

$$F = \begin{pmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{pmatrix} \in \mathbb{R}^{2\times 2}$$

daß

$$EF = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{pmatrix} = \begin{pmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{pmatrix} = F$$

und

$$FE = \begin{pmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{pmatrix} = F$$

also $EF = F = FE$. Diese spezielle Matrix kommutiert also im Matrixprodukt mit jeder 2×2 Matrix und läßt dabei die beteiligte Matrix sogar unverändert. Die Matrix E hat bezüglich des Matrixprodukts somit eine neutrale Wirkung, ist das *Neutralelement* der Matrixmultiplikation. Überlegen wir uns einmal, zu welcher linearen Abbildung E gehört, wenn wir kanonische Basen voraussetzen. Dazu berechnen wir

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

d. h. das Bild des Vektors

$$x_1(1,0) + x_2(0,1) = (x_1, x_2)$$

ist wieder

$$x_1(1,0) + x_2(0,1) = (x_1, x_2)$$

und damit ist die zugehörige lineare Abbildung die Identitätsabbildung auf $\mathbb{R}^2$. Entsprechende Schlußfolgerungen gelten natürlich in $\mathbb{R}^n$ mit den Einheitsmatrizen

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \text{etc.}$$

5. Geraden, Ebenen und Gaußalgorithmus

In diesem Abschnitt wollen wir kurz darauf eingehen, wie man geometrische Objekte (Geraden, Ebenen) im Zeigervektorraum S darstellen kann. Betrachten wir zunächst den Fall der Geraden. Eine Gerade ist durch zwei verschiedene Punkte A und P im Raum eindeutig charakterisiert. Dabei gibt der Zeiger $\overrightarrow{AP}$ (oder $\overrightarrow{PA}$) die Richtung an, die man vom Punkt A (oder P) aus einschlagen muß, um der Geraden zu folgen. Anders ausgedrückt, erhält man alle Punkte der Geraden durch die Zeigerspitzen von $t \cdot \overrightarrow{AP}, t \in \mathbb{R}$, wenn wir A als Referenzpunkt für die stumpfen Enden der Zeiger nehmen. Die Menge

$$G = \{t\overrightarrow{AP} | t \in \mathbb{R}\}$$

bildet übrigens einen Untervektorraum von S, wie man sich leicht überzeugt. Das α-Vielfache eines Elements $t\overrightarrow{AP} \in G$ ist nämlich von der Form $(\alpha t)\overrightarrow{AP}$ und damit ein Element von G. Ebenso ist die Summe zweier Elemente $t\overrightarrow{AP}, s\overrightarrow{AP}$ wieder von der Form $u\overrightarrow{AP}$ mit der reellen Zahl $u = t + s$ und damit in G enthalten. Dies sind genau die beiden Bedingungen, die wir überprüfen müssen: Addition und skalare Multiplikation von Elementen aus G liefert wieder Elemente aus G. Als Untervektorraum ist G auch ein eigenständiger Vektorraum und wir können nach der Dimension fragen. Dazu benötigen wir eine Basis von G, wobei die erste Wahl sicherlich auf $\{\overrightarrow{AP}\}$ fällt. Offensichtlich werden alle Elemente von $\overrightarrow{AP}$ erzeugt und das Erzeugendensystem ist minimal. Damit gilt also dim $G = 1$ und wir können Geraden mit eindimensionalen Untervektorräumen von S identifizieren (denn umgekehrt ist jeder eindimensionale Untervektorraum ja von der Form $\{t\vec{b} | t \in \mathbb{R}\}$ wobei $\vec{b}$ ein Basisvektor ist).

Bisher haben wir den Fall betrachtet, daß unser Referenzpunkt A auf der Gerade liegt. Für die Beschreibung einer einzelnen Gerade ist das sicherlich eine günstige Wahl. Wollen wir allerdings zwei oder mehr Geraden gleichzeitig betrachten und z. B. nach Schnittpunkten oder Parallelität fragen, so kann der Referenzpunkt A höchstens auf einer der beteiligten Geraden liegen. Eine Gerade mit dem *Richtungsvektor* $\vec{b} \neq \vec{0}$, die nicht durch den Punkt A läuft, kann man dadurch beschreiben, daß man einen beliebigen Punkt B der Geraden nimmt und dann halt von B aus der Richtung $\vec{b}$ folgt. Die Punkte der Gerade können also mit

$$\overrightarrow{AB} + t\vec{b}, \quad t \in \mathbb{R}$$

angesprochen werden. Hätten wir einen anderen Startpunkt auf der Gerade gewählt, z. B. den Punkt C mit dem Ortsvektor

$$\overrightarrow{AC} = \overrightarrow{AB} + t_0\vec{b}$$

für ein bestimmtes $t_0 \in \mathbb{R}$, so ergibt sich damit die gleiche Gerade

$$\overrightarrow{AB} + (t + t_0)\vec{b}, \quad t \in \mathbb{R},$$

denn $(t + t_0)\vec{b}$ liefert ja dieselben Punkte wie $t\vec{b}$, wenn t alle reellen Zahlen durchläuft.
Geometrisch beschreibt $A(\vec{x})$ übrigens den Zeiger, der vom Ebenenpunkt $E(x_2, x_3)$ zum Geradenpunkt $G(x_1)$ zeigt. Ist dieser Abstandszeiger gleich $\vec{0}$, so fallen die beiden Punkte zusammen, d. h. Nullstellen von A entsprechen Schnittpunkten.
Genau wie Geraden, kann man auch Ebenen im Zeigervektorraum beschreiben. Der einzige Unterschied ist, daß man sich in einer Ebene in zwei *unabhängige* Richtungen bewegen kann. Eine Ebene durch den Referenzpunkt A ist daher durch zwei Richtungsvektoren $\vec{a}, \vec{b}$ definiert, wobei $\vec{a}$ kein Vielfaches von $\vec{b}$ sein soll (d. h. $\vec{a}, \vec{b}$ sind linear unabhängig). Die Punkte der Ebene sind dann durch die Ortsvektoren

$$s\vec{a} + t\vec{b}, \qquad s, t \in \mathbb{R}$$

gegeben und bilden offensichtlich einen zweidimensionalen Unterraum von S. Wie bei den Geraden wird man im allgemeinen auch Ebenen betrachten, die den Referenzpunkt A nicht enthalten. In diesem Fall wird die Ebene durch einen *Aufpunkt* P bzw. den zugehörigen Ortsvektor $\overrightarrow{AP}$ und zwei unabhängige Richtungen $\vec{a}$ und $\vec{b}$ beschrieben

$$\overrightarrow{AP} + s\vec{a} + t\vec{b}, \qquad s, t \in \mathbb{R}$$

Allgemein nennt man Teilmengen eines Vektorraums V der Form

$$\vec{v} + U = \{\vec{v} + \vec{u} | \vec{u} \in U\}$$

affine Teilräume, wenn U ein Untervektorraum (oder synonym, ein Teilraum) von V ist. Affine Teilräume sind also Teilräume, die um einen festen Vektor $\vec{v}$ verschoben sind. Beachten Sie, daß jeder Untervektorraum auch ein affiner Teilraum ist. Der Verschiebungsvektor ist hier $\vec{v} = \vec{0}$.
Geraden und Ebenen kann man übrigens auch als Bildmengen von Abbildungen betrachten. So ist das Bild von $G : \mathbb{R} \to S$ mit $G(t) = \vec{c} + t\vec{a}$ im Fall $\vec{a} \neq \vec{0}$ die Gerade durch $\vec{c}$ mit Richtung $\vec{a}$ und $E : \mathbb{R}^2 \to S$ mit $E(s, t) = \vec{c} + s\vec{a} + t\vec{b}$ liefert eine Ebene als Bildmenge, falls $\vec{a}, \vec{b}$ linear unabhängig sind.

Allgemein nennt man eine Abbildung $A : V \to W$ zwischen zwei Vektorräumen V und W *affin linear*, wenn sie von der Form $A(\vec{v}) = L(\vec{v})+\vec{c}$ ist, mit einer linearen Abbildung $L : V \to W$. Der konstante Vektor $\vec{c}$ ist übrigens durch $A(\vec{0}) = L(\vec{0}) + \vec{c} = \vec{c}$ gegeben, denn lineare Abbildungen liefern im Punkt $\vec{0}$ immer den Wert $\vec{0}$, da

$$L(\vec{0}) = L(0 \cdot \vec{0}) = 0 \cdot L(\vec{0}) = \vec{0}.$$

Damit ist also $A : V \to W$ affin linear, wenn $L(\vec{v}) = A(\vec{v}) - A(\vec{0})$ linear ist. Jetzt wollen wie aber endlich mit Geraden und Ebenen arbeiten. Als erste Frage interessiert uns, ob sich eine gegebene Gerade mit einer bestimmten Ebene schneidet. Wir nehmen an, daß Gerade und Ebene als Bilder von zwei affin linearen Abbildungen $G : \mathbb{R} \to S$ und $E : \mathbb{R}^2 \to S$ vorliegen. Die Frage nach einem Schnittpunkt entspricht damit der Frage nach einem gemeinsamen Bildpunkt der Abbildungen, d. h. wir suchen $t, s_1, s_2 \in \mathbb{R}$ mit der Eigenschaft $G(t) = E(s_1, s_2)$. Die Fragestellung läßt sich noch knapper formulieren, wenn wir die Abbildung $A(t, s_1, s_2) = G(t) - E(s_1, s_2)$ einführen, die $\mathbb{R}^3$ nach S abbildet. Offensichtlich ist A affin linear, denn

$$A(\vec{x}) - A(\vec{0}) = G(x_1) - G(0) + E(x_2, x_3) - E(0, 0)$$

ist, wie man sich leicht überzeugt, linear. Das Schnittproblem führt damit auf die Gleichung $A(\vec{x}) = \vec{0}$ mit einer affin linearen Funktion $A : \mathbb{R}^3 \to S$.
In genau der gleichen Weise lassen sich alle anderen Schnittprobleme (Gerade - Gerade, Ebene - Ebene, Gerade - Ebene ...) auf die Nullstellenbestimmung von affin linearen Abbildungen zurückführen. Als Beispiel erwähnen wir noch den Fall des Geradeschnitts. Seien zwei Geraden durch die Bildmengen von $G_1, G_2 : \mathbb{R} \to S$ gegeben.
Die Frage nach gemeinsamen Punkten stellt sich dann folgendermaßen: Finde $s, t \in \mathbb{R}$, so daß $G_1(s) = G_2(t)$ gilt. Mit der affinlinearen Abbildung $B(x_1, x_2) = G_1(x_1) - G_2(x_2)$ von $\mathbb{R}^2$ nach S kann man das Problem auch so formulieren: Finde $\vec{x} \in \mathbb{R}^2$, so daß $B(\vec{x}) = \vec{0}$ gilt. Wir können also allgemeine Schnittprobleme lösen, falls wir in der Lage sind, Nullstellen von allgemeinen affinen Abbildungen zu finden, und diesem Thema wollen wir uns jetzt widmen.
Sei $A : V \to W$ dazu eine affin lineare Abbildung zwischen zwei Vektorräumen. Wir wissen dann, daß $L(\vec{v}) = A(\vec{v}) - A(\vec{0})$ eine lineare Abbildung ist. Die Nullstellensuche von A kann man also auch als Gleichung mit L formulieren, denn $\vec{v}$ ist genau dann eine Nullstelle von A, wenn $L(\vec{v}) = -A(\vec{0})$ gilt. Eine solche Gleichung nennt man auch *lineare Gleichung*.

Mit der Umformulierung des Nullstellenproblems in eine lineare Gleichung sind wir der Lösung zwar noch nicht viel näher gekommen, aber wir haben jetzt die Möglichkeit, die Berechnung in Koordinatenräume $\mathbb{R}^{n\times 1}, \mathbb{R}^{m\times 1}$ zu verlagern, wenn V, W endlich dimensional sind.
Dazu führen wir Basen $(\vec{v}_1, \ldots, \vec{v}_n)$ von V und $(\vec{w}_1, \ldots, \vec{w}_m)$ von W ein und ermitteln die Matrix M zur Abbildung L sowie die Koordinaten $\vec{y} \in \mathbb{R}^{m\times 1}$ des Vektors $-A(\vec{0})$. Das Problem: Finde $\vec{v} \in V$, so daß $L(\vec{v}) = -A(\vec{0})$ gilt, stellt sich dann so dar: Finde die Koordinaten $\vec{x} \in \mathbb{R}^{n\times 1}$, so daß $M\vec{x} = \vec{y}$ gilt. Hat man dieses Problem gelöst, so liefert jeder Lösungsvektor $\vec{x}$ eine Lösung $\vec{v}$ des ursprünglichen Problems mit $\vec{v} = x_1\vec{v}_1 + \ldots + x_n\vec{v}_n$.
Schreiben wir die lineare Gleich $M\vec{x} = \vec{y}$ ausführlich, so ergibt sich das Gleichungssystem

$$\begin{array}{ccccccc} M_{11}x_1 & + & M_{12}x_2 & + \ldots + & M_{1n}x_n & = & y_1 \\ M_{21}x_1 & + & M_{22}x_2 & + \ldots + & M_{2n}x_n & = & y_2 \\ \vdots & & \vdots & & \vdots & & \vdots \\ M_{m1}x_1 & + & M_{m2}x_2 & + \ldots + & M_{mn}x_n & = & y_m \end{array}$$

aus m skalaren linearen Gleichungen mit n Unbekannten $x_1, \ldots, x_n$. Wie würden Sie so ein Gleichungssystem nach den Unbekannten x_i auflösen? Eine Möglichkeit ist, die Unbekannten nacheinander aus dem Gleichungssystem zu entfernen. Nehmen wir an, in der zweiten Gleichung ist $M_{21} \neq 0$. Dann kann man diese Gleichung ja nach x_1 auflösen und erhält x_1 in Abhängigkeit von $x_2, \ldots, x_n$ und y_2. Setzt man diesen Ausdruck in die übrigen Gleichungen ein, so ergeben sich $m-1$ Gleichungen für die $n-1$ Unbekannten $x_2, \ldots, x_n$. Diese Prozedur kann man nun wiederholen und man erhält somit immer weniger Gleichungen für immer weniger Unbekannte. Endet dieser Prozeß so, daß die letzte Variable eindeutig bestimmt ist, so kann man mit diesem Wert nacheinander die Werte der vorher eliminierten Variablen ausrechnen. Die detailierte Ausführung dieser Idee führt auf den Gaußschen Eliminationsalgorithmus, den wir nun entwickeln wollen.
Zunächst stellen wir fest, daß die Anordnung der Gleichungen sicherlich keine Auswirkung auf den Lösungsprozeß hat, d. h. es ist egal, welche der Gleichungen die oberste, die zweite, die dritte usw. ist. Wir dürfen also *Zeilen vertauschen* ohne daß dadurch die Lösungsmenge beeinflußt wird. Diese Eigenschaft benutzen wir dazu, unser Vorgehen zu automatisieren. Wir suchen uns dazu eine Zeile, in der der Koeffizient vor x_1 nicht gleich Null ist und bringen sie in die oberste Position (sind alle Koeffizienten gleich Null, so wird x_1 durch das Gleichungssystem offensichtlich nicht eingeschränkt, kann also beliebig gewählt werden).

Die erste Gleichung hat dann die Form

$$a_1x_1 + \ldots a_nx_n = A$$

mit $a_1 \neq 0$ und kann aufgelöst werden

$$x_1 = \frac{1}{a_1}(A - a_2x_2 - \ldots - a_nx_n)$$

Dieser Zusammenhang, eingesetzt in eine beliebige andere Gleichung

$$b_1x_1 + \ldots + b_nx_n = B$$

liefert

$$\frac{b_1}{a_1}(A - a_2x_2 - \ldots - a_nx_n) + b_2x_2 + \ldots + b_nx_n = B$$

oder nach geeigneter Sortierung

$$\left(b_2 - \frac{b_1}{a_1}a_2\right)x_2 + \ldots + \left(b_n - \frac{b_1}{a_1}a_n\right)x_n = B - \frac{b_1}{a_1}A$$

Wir sehen, daß x_1 auf diese Weise in allen anderen Gleichungen eliminiert werden kann. Auch wird deutlich, wie sich die Koeffizienten ändern. Schreiben wir nur die Koeffizienten untereinander, so haben wir vor der Umformung

$$\begin{array}{ccccc|c} a_1 & a_2 & a_3 & \ldots & a_n & A \\ b_1 & b_2 & b_3 & \ldots & b_n & B \end{array}$$

und danach

$$\begin{array}{ccccc|c} 1 & \frac{1}{a_1}a_2 & \frac{1}{a_1}a_3 & \ldots & \frac{1}{a_1}a_n & \frac{1}{a_1}A \\ 1 & b_2 - \frac{b_1}{a_1}a_2 & b_3 - \frac{b_1}{a_1}a_3 & \ldots & b_n - \frac{b_1}{a_1}a_n & B - \frac{b_1}{a_1}A \end{array}$$

Auf die Koeffizienten im Gleichungssystem hat das Auflösen und Einsetzen also folgende Konsequenz

- die oberste Zeile wird mit einer von Null verschiedenen Zahl multipliziert (komponentenweise)
- zu den übrigen Zeilen wird ein Vielfaches der ersten Zeile addiert (komponentenweise)

Um Schreibarbeit zu vermeiden, verzichten wir ab jetzt auf die Angabe von x_i und $+, =$ in den Gleichungen und schreiben nur die Koeffizienten auf, wobei die Faktoren von x_1 in der ersten Spalte stehen, die von x_2 in der zweiten usw. Beachten Sie, daß eine Null als Koeffizient in der j-ten Spalte einzutragen ist, wenn eine Variable x_j in einer Gleichung gar nicht auftritt. Das Vertauschen von Gleichungen, das Auflösen und Einsetzen äußert sich dann nur noch in den entsprechenden Zeilen-Operationen.

Die Elimination von x_1 aus dem Gleichungssystem liefert also folgende Situation der Koeffizientenmatrix

$$\left(\begin{array}{ccccc|c} * & * & * & \cdots & * & * \\ 0 & * & * & \cdots & * & * \\ 0 & \vdots & & & \vdots & \vdots \\ \vdots & \vdots & & & \vdots & \vdots \\ 0 & * & * & \cdots & * & * \end{array}\right)$$

wobei $*$ einen beliebigen Wert andeutet. Beachten Sie, daß wir das erste Element der ersten Zeile nicht als 1 schreiben. Damit behandeln wir gleichzeitig den einfachen Spezialfall, daß kein von Null verschiedener Koeffizient von x_1 existiert. In diesem Fall ist der obere linke Stern nämlich eine 0. Ist die Variable x_1 eliminiert, so bleiben $m-1$ Gleichungen für die $n-1$ unbekannten $x_2, \ldots, x_n$. Die Koeffizienten dieses neuen Systems finden sich in dem Koeffizientenblock unterhalb der ersten Zeile und rechts von der ersten Spalte. Auch hier können wir durch Zeilentausch und anschließende Elimination (d. h. Addition von Vielfachen der zweiten Zeile zu den Zeilen drei bis n) folgende Situation herstellen.

$$\left(\begin{array}{ccccc|c} * & * & * & \cdots & * & * \\ 0 & * & * & \cdots & * & * \\ 0 & 0 & * & \cdots & * & * \\ 0 & 0 & * & \cdots & * & \vdots \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & * & & * & * \end{array}\right)$$

Entsprechend eliminiert man $x_3, x_4, \ldots$, solange noch Gleichungen zur Verfügung stehen. Insgesamt erhält man also mit den Operationen Vertauschen, Multiplizieren mit nicht-Null Faktoren und addieren von Zeilenvielfachen ein äquivalentes Gleichungssystem der dreieckigen Form

$$\begin{array}{ccccccccccc} \widetilde{M}_{11}x_1 & + & \widetilde{M}_{12}x_2 & + & \widetilde{M}_{13}x_3 & + \ldots + & \widetilde{M}_{1n}x_n & = & \widetilde{y}_1 \\ & & \widetilde{M}_{22}x_2 & + & \widetilde{M}_{23}x_3 & + \ldots + & \widetilde{M}_{2n}x_n & = & \widetilde{y}_2 \\ & & & & \widetilde{M}_{23}x_3 & + \ldots + & \widetilde{M}_{3n}x_n & = & \widetilde{y}_3 \\ & & & & & & \vdots & & \vdots \end{array}$$

Je nach Größenverhältnis von n und m ist das Dreieck dabei natürlich mehr oder weniger verunstaltet. Im Spezialfall $n = m$ und $\widetilde{M}_{ii} \neq 0$ für alle i sehen wir aber sehr schnell ein, daß das Gleichungssystem

genau eine Lösung hat. Die letzte Gleichung führt nämlich auf $x_n = \widetilde{y}_n/\widetilde{M}_{nn}$. Dieser Wert liefert dann in der vorletzten Gleichung $x_{n-1} = (\widetilde{y}_{n-1} - \widetilde{M}_{n-1n}x_n)/\widetilde{M}_{n-1n-1}$ und entsprechend lassen sich alle Werte x_i bestimmen. Andere Fälle ($n > m, n < m$, nicht alle $\widetilde{M}_{ii} \neq 0$) werden wir uns jetzt an Beispielen genauer ansehen.
Betrachten wir zunächst den Schnitt von zwei Geraden, die durch affine Funktionen $G_1 : \mathbb{R} \to S$ gegeben sind. Um das Schnittproblem in ein lineares Gleichungssystem vom Typ $M\vec{x} = \vec{y}$ zu verwandeln, müssen wir zunächst eine Basis von S einführen und alle beteiligten Elemente von S in dieser Basis ausdrücken. Ist $\Phi : S \to \mathbb{R}^{3\times 1}$ die Koordinatenabbildung, so betrachten wir die Geraden in Koordinatendarstellung, die als Bilder der verketteten Abbildungen $g_i = \Phi \circ G_i$ gegeben sind. Als konkretes Beispiel nehmen wir

$$g_1(t) = \begin{pmatrix}1\\0\\1\end{pmatrix} + t\begin{pmatrix}2\\2\\1\end{pmatrix}, \quad g_2(s) = \begin{pmatrix}1\\-1\\-1\end{pmatrix} + s\begin{pmatrix}-1\\3\\-1\end{pmatrix}$$

Die Abbildung $a(t,s) = g_1(t) - g_2(s)$ ist dann

$$a(t,s) = \begin{pmatrix}2 & 1\\2 & -3\\1 & 1\end{pmatrix}\begin{pmatrix}t\\s\end{pmatrix} + \begin{pmatrix}0\\1\\2\end{pmatrix}$$

und das Schnittproblem: Finde $(t,s) \in \mathbb{R}^2$, so daß $a(t,s) = \vec{0}$ gilt, führt auf das lineare Gleichungssystem mit den Koeffizienten

$$\left(\begin{array}{cc|c}2 & 1 & 0\\2 & -3 & -1\\1 & 1 & -2\end{array}\right)$$

Dieses System wollen wir nun so manipulieren, daß es (so weit wie möglich) Dreicksform annimmt. Zur Vereinfachung der Rechnung tauschen wir zunächst die erste und die letzte Zeile

$$\left(\begin{array}{cc|c}1 & 1 & -2\\2 & -3 & -1\\2 & 1 & 0\end{array}\right)$$

und subtrahieren dann das Zweifache der ersten Zeile von der zweiten Zeile

$$\left(\begin{array}{cc|c}1 & 1 & -2\\0 & -5 & -5\\2 & 1 & 0\end{array}\right)$$

und das Zweifache der ersten von der letzten Zeile

$$\left(\begin{array}{cc|c} 1 & 1 & -2 \\ 0 & -5 & -5 \\ 0 & -1 & 2 \end{array}\right)$$

Hätten wir am Anfang die Zeilen nicht getauscht, so hätten wir entweder die Hälfte der ersten Zeile von der letzten abziehen müssen, was zu Brüchen in den Einträgen führt, oder aber die letzte Zeile zunächst mit zwei durchmultiplizieren müssen, um danach die erste Zeile abzuziehen. Das geht natürlich auch. Schauen wir uns das Endsystem einmal genauer an. Die letzte Zeile sagt uns, daß $-s = 2$ gelten muß, wenn eine Lösung vorliegt. Allerdings besagt die vorletzte Zeile, daß gleichzeitig $-5s = -5$ zutreffen muß, aber s kann ja nicht gleichzeitig -2 und 1 sein! Das Gleichungssystem führt also zu widersprüchlichen Anforderungen an t, s und wird deshalb durch *kein* Paar t, s erfüllt. Wir sagen, daß die Lösungsmenge dieses Gleichungssystems leer ist. Wenn wir uns zurückbesinnen auf die geometrische Bedeutung des Problems, so ist das Ergebnis nicht überraschend. Zwei beliebige Geraden im Raum werden sich eben tyischerweise *nicht* schneiden. Damit ein Schnittpunkt auftreten kann, müssen Aufpunkte und Richtungsvektoren sorgfältig gewählt werden, z. B.

$$g_1(t) = \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix} + t \begin{pmatrix} 2 \\ 2 \\ -2 \end{pmatrix}, \quad g_2(s) = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} + s \begin{pmatrix} -2 \\ -4 \\ 3 \end{pmatrix}$$

Das zugehörige Gleichungssystem ist jetzt durch die Koeffizienten

$$\left(\begin{array}{cc|c} 2 & 2 & 2 \\ 2 & 4 & 0 \\ -2 & -3 & -1 \end{array}\right)$$

gegeben. Subtrahieren bzw. Addieren der ersten Zeile zur zweiten bzw. dritten Zeile liefert

$$\left(\begin{array}{cc|c} 2 & 2 & 2 \\ 0 & 2 & -2 \\ 0 & -1 & 1 \end{array}\right)$$

Wieder ergeben sich zwei Bedingungen an s, die aber jetzt kompatibel sind. Die beiden letzten Gleichungen verlangen nämlich $s = -1$ und die erste Gleichung ergibt dann $2t = 2 - 2s = 4$, also $t = 2$. In diesem Fall gibt es also genau eine Lösung des Gleichungssystems, die genau einem Geradenschnittpunkt entspricht. Die Koordinaten dieses Schnittpunkts sind übrigens $g_1(2)$ bzw. $g_2(-1)$. Zur Probe, ob wir uns nicht verrechnet

haben, empfiehlt es sich, beide Koordinatentripel zu berechnen. Wir erhalten

$$g_1(2) = \begin{pmatrix} 3 \\ 3 \\ -3 \end{pmatrix}, \quad g_2(-1) = \begin{pmatrix} 3 \\ 3 \\ -3 \end{pmatrix}$$

Eine noch speziellere Situation ergibt sich in folgendem Fall

$$g_1(t) = \begin{pmatrix} 1 \\ -1 \\ 3 \end{pmatrix} + t \begin{pmatrix} -2 \\ 2 \\ 1 \end{pmatrix}, \quad g_2(s) = \begin{pmatrix} 0 \\ 0 \\ \frac{7}{2} \end{pmatrix} + s \begin{pmatrix} 1 \\ -1 \\ -\frac{1}{2} \end{pmatrix}$$

mit

$$\left(\begin{array}{cc|c} -2 & -1 & -1 \\ 2 & 1 & 1 \\ 1 & \frac{1}{2} & \frac{1}{2} \end{array}\right)$$

was durch Zeilenumformung auf

$$\left(\begin{array}{cc|c} -2 & -1 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}\right)$$

führt. Sie sehen, daß in diesem Fall an s gar keine Bedingung gestellt wird, denn die Gleichung $0 \cdot s = 0$ ist ja für alle $s \in \mathbb{R}$ automatisch erfüllt. Für t ergibt sich die Bedingung $-2t - s = -1$ bzw. $t = (1-s)/2$. Wir erhalten hier also eine richtig große Lösungsmenge

$$\{ \binom{t}{s} \mid t = (1-s)/2, \quad s \in \mathbb{R} \}$$

Geometrisch gesprochen bedeutet dies, daß die Geraden unendlich viele Schnittpunkte haben, also aufeinander liegen. Es gilt ja $g_1((1-s)/2) = g_2(s)$, d. h. jedem Punkt $g_2(s)$ auf der einen Gerade entspricht der Punkt $g_1((1-s)/2)$ auf der anderen.
Obwohl wir bisher nur drei spezielle Beispiele betrachtet haben, ist es doch einfach nachvollziehbar, was in einem allgemeinen Fall von m Gleichungen mit n Unbekannten ($m > n$) im Lösungsprozeß passieren wird. Zunächst läßt sich die Koeffizientenmatrix auf die Form

$$\left(\begin{array}{ccccc|c} * & * & \dots & * & * & * \\ & * & \dots & * & * & * \\ & & \ddots & & \vdots & \vdots \\ & & & * & * & * \\ & & & & * & * \\ & & & & * & * \\ & & & & \vdots & \vdots \\ & & & & * & * \end{array}\right)$$

transformieren. Im allgemeinen werden dabei die letzten Gleichungen widersprüchlich sein und das Gleichungssystem damit keine Lösung haben (der Fall $m > n$ heißt auch *überbestimmtes* Gleichungssystem, da mehr Bedingungen als Unbekannte vorhanden sind). In ganz besonderen Fällen sind die Anforderungen an die letzte Variable kompatibel, so daß mindestens eine Lösung existiert. Ist aber ein Diagonalelement des Dreiecks gleich Null, so läßt sich für die entsprechende Variable keine Bedingung ableiten. Sie kann damit frei gewählt werden und es ergibt sich eine ganze Lösungsschar.
Als Beispiel für den Fall *unterbestimmter* Gleichungssysteme ($m < n$) betrachten wir nun den Schnitt von zwei Ebenen, die wir auch wieder direkt durch ihre Koordinatendarstellung beschreiben

$$\begin{aligned} e_1(t_1,t_2) &= t_1 \begin{pmatrix} 1 \\ 2 \\ 5 \end{pmatrix} + t_2 \begin{pmatrix} -2 \\ 1 \\ 1 \end{pmatrix}, \\ e_2(s_1,s_2) &= \begin{pmatrix} -1 \\ 3 \\ 1 \end{pmatrix} + s_1 \begin{pmatrix} 0 \\ 1 \\ 3 \end{pmatrix} + s_2 \begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix} \end{aligned}$$

Das Schnittproblem: Finde $(s_1,s_2),(t_1,t_2) \in \mathbb{R}^2$ so daß $e_1(t_1,t_2) = e_2(s_1,s_2)$ gilt, führt auf ein Nullstellenproblem für die affine Funktion

$$a(t_1,t_2,s_1,s_2) = \begin{pmatrix} 1 & -2 & 0 & -1 \\ 2 & 1 & -1 & 0 \\ 5 & 1 & -3 & 2 \end{pmatrix} \begin{pmatrix} t_1 \\ t_2 \\ s_1 \\ s_2 \end{pmatrix} + \begin{pmatrix} 1 \\ -3 \\ -1 \end{pmatrix}$$

und damit auf das Gleichungssystem mit Koeffizienten

$$\left(\begin{array}{cccc|c} 1 & -2 & 0 & -1 & -1 \\ 2 & 1 & -1 & 0 & 3 \\ 5 & 1 & -3 & 2 & 1 \end{array}\right)$$

Durch elementare Zeilenumformungen erhält man zunächst

$$\left(\begin{array}{cccc|c} 1 & -2 & 0 & -1 & -1 \\ 0 & 5 & -1 & 2 & 5 \\ 0 & 11 & -3 & 7 & 6 \end{array}\right)$$

und dann durch Multiplikation der zweiten Zeile mit 11 und der letzten Zeile mit 5

$$\left(\begin{array}{cccc|c} 1 & -2 & 0 & -1 & -1 \\ 0 & 55 & -11 & 22 & 55 \\ 0 & 55 & -15 & 35 & 30 \end{array}\right)$$

Schließlich ergibt sich durch Subtraktion der zweiten Zeile von der letzten

$$\left(\begin{array}{cccc|c} 1 & -2 & 0 & -1 & -1 \\ 0 & 55 & -11 & 22 & 55 \\ 0 & 0 & -4 & 13 & -25 \end{array}\right)$$

und man kann den Faktor 11 aus der zweiten Zeile wieder herausdividieren

$$\left(\begin{array}{cccc|c} 1 & -2 & 0 & -1 & -1 \\ 0 & 5 & -1 & 2 & 5 \\ 0 & 0 & -4 & 13 & -25 \end{array}\right)$$

Offensichtlich liefert der Lösungsprozeß im unterbestimmten Fall immer eine Koeffizientenstruktur der Form

$$\left(\begin{array}{cccccc|c} * & * & \dots & * & \dots & * & * \\ & * & \dots & * & \dots & * & \vdots \\ & & \ddots & & & & \vdots \\ & & & * & \dots & * & * \end{array}\right)$$

da einfach nicht genügend Gleichungen (d. h. einschränkende Bedingungen) für die Unbekannten vorhanden sind. Einige der gesuchten Größen werden also typischerweise nicht durch das System festgelegt und können somit frei gewählt werden. In unserem Beispiel kann z. B. die Variable s_2 beliebige Werte annehmen, wenn nur die anderen Variablen dazu passen, d. h. es muß gelten

$$\begin{aligned} t_1 - 2t_2 + 0s_1 &= -1 + s_2 \\ 5t_2 - s_1 &= 5 - 2s_2 \\ -4s_1 &= -25 - 13s_2 \end{aligned}$$

Dieses Gleichungssystem hat jetzt Dreiecksgestalt und kann daher sukzessive von unten nach den Unbekannten s_1, t_2, t_1 aufgelöst werden. Das erfordert offensichtlich fortgesetztes Einsetzen, was, wie wir wissen, auch durch Zeilenoperationen dargestellt werden kann. Betrachten wir das System

$$\left(\begin{array}{ccc|c} 1 & -2 & 0 & s_2 - 1 \\ & 5 & -1 & 5 - 2s_2 \\ & & -4 & -25 - 13s_2 \end{array}\right)$$

Multiplikation der zweiten Zeile mit (-4) und Addition der letzten Zeile liefert

$$\left(\begin{array}{ccc|c} 1 & -2 & 0 & s_2 - 1 \\ 0 & -20 & 0 & -45 - 5s_2 \\ & & -4 & -25 - 13s_2 \end{array}\right)$$

und eine Multiplikation der ersten Zeile mit (-10) und Addition der zweiten Zeile ergibt

$$\left(\begin{array}{ccc|c} -10 & 0 & 0 & -35 - 15s_2 \\ & -20 & 0 & -45 - 5s_2 \\ & & -4 & -25 - 13s_2 \end{array}\right)$$

Durch passende Zeilenmultiplikation kann man hier noch aufräumen

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & \frac{7}{2}+\frac{3}{2}s_2 \\ & 1 & 0 & \frac{9}{4}+\frac{1}{4}s_2 \\ & & 1 & \frac{25}{4}+\frac{13}{4}s_2 \end{array}\right)$$

Die Lösungsvektoren des Gleichungssystems haben also die Form

$$\begin{pmatrix} t_1 \\ t_2 \\ s_1 \\ s_2 \end{pmatrix} = \begin{pmatrix} \frac{7}{2}+\frac{3}{2}\alpha \\ \frac{9}{4}+\frac{1}{4}\alpha \\ \frac{25}{4}+\frac{13}{4}\alpha \\ \alpha \end{pmatrix}, \quad \alpha \in \mathbb{R}$$

Was heißt das für das Schnittgebilde der beiden Ebenen? Setzen wir ein

$$e_2\left(\frac{25}{4}+\frac{13}{4}\alpha, \alpha\right) = \begin{pmatrix} -1 \\ \frac{37}{4} \\ \frac{79}{4} \end{pmatrix} + \alpha \begin{pmatrix} 1 \\ \frac{13}{4} \\ \frac{31}{4} \end{pmatrix}, \quad \alpha \in \mathbb{R}$$

so sehen wir, daß die Schnittpunkte eine Gerade beschreiben. Natürlich finden wir die gleiche Gerade, wenn wir $e_1(\frac{7}{2}+\frac{3}{2}\alpha, \frac{9}{4}+\frac{1}{4}\alpha)$ ausrechnen. In Spezialfällen kann natürlich auch ein unterbestimmtes Gleichungssystem keine Lösung haben. Denken Sie z. B. an den Schnitt von zwei parallelen Ebenen. In diesem Fall sind halt einige $*$ Symbole in der abgeschnittenen Dreiecksform gerade gleich Null, was dann zu Widersprüchen führt. Beispielsweise kann folgende Situation auftreten,

$$\left(\begin{array}{cccccc|c} * & \dots & * & * & \dots & * & * \\ & \ddots & \vdots & \vdots & & \vdots & \vdots \\ & & * & * & \dots & * & * \\ & & & 0 & \dots & 0 & \neq 0 \end{array}\right)$$

was auf die widersprüchliche Bedingung $0 \neq 0$ führt.
Gleichungssysteme, die weder über- noch unterbestimmt sind, haben genausoviele Gleichungen wie Unbekannte ($n = m$). Dieser Fall tritt zum Beispiel beim Schnitt von einer Gerade mit einer Ebene auf. Aus der Anschauung heraus können Sie damit die Struktur der Lösungsmenge schon erraten. Eine beliebig gewählte Gerade schneidet eine beliebig gewählte Ebene typischerweise in genau einem Punkt, d. h. die Lösungsmenge enthält nur einen Vektor, bzw. das Gleichungssystem

ist eindeutig lösbar. Wenn Sie sich bei der Auswahl der Richtungsvektoren sehr anstrengen, kann aber auch der Fall eintreten, daß keine Lösung existiert (Gerade parallel zur Ebene), oder daß unendlich viele Schnittpunkte auftreten (Gerade liegt in der Ebene). Dafür muß aber in jedem Fall der Richtungsvektor der Gerade durch die Richtungsvektoren der Ebene darstellbar sein, d. h. die beteiligten Richtungsvektoren sind linear abhängig. Da uns die Spezialfälle prinzipiell in den vorherigen Beispielen schon begegnet sind, beschränken wir uns hier auf den allgemeinen Fall des Schnitts von Gerade und Ebene. Sei dazu

$$g(t) = \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix} + t \begin{pmatrix} 4 \\ 1 \\ -2 \end{pmatrix}, \quad e(s_1, s_2) = s_1 \begin{pmatrix} 3 \\ 3 \\ 1 \end{pmatrix} + s_2 \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$$

Das zugehörige Schnittproblem führt auf das Gleichungssystem

$$\left(\begin{array}{ccc|c} 4 & -3 & -1 & -1 \\ 1 & -3 & 0 & 2 \\ -2 & -1 & 1 & 1 \end{array}\right)$$

Vertauscht man erste und zweite Zeile, so läßt sich die erste Variable leicht eliminieren

$$\left(\begin{array}{ccc|c} 1 & -3 & 0 & 2 \\ 0 & 9 & -2 & -9 \\ 0 & -7 & 1 & 5 \end{array}\right)$$

Will man Bruchzeilen vermeiden, so bietet es sich an, die zweite und letzte Zeile zunächst zu erweitern

$$\left(\begin{array}{ccc|c} 1 & -3 & 0 & 2 \\ 0 & 63 & -7 & -63 \\ 0 & -63 & 9 & 45 \end{array}\right)$$

wobei die Erweiterung der zweiten Zeile nach Durchführung der Summe mit der dritten auch wieder rückgängig gemacht werden kann

$$\left(\begin{array}{ccc|c} 1 & -3 & 0 & 2 \\ 0 & 9 & -1 & -9 \\ 0 & 0 & 2 & -18 \end{array}\right)$$

Wie versprochen, erhalten wir also ein „echtes“ Dreieck, d. h. die Diagonalelemente sind alle von Null verschieden.
Das Ausrechnen der x_i kann man wieder durch Zeilenoperationen durchführen, aber jetzt von unten nach oben. Zunächst erhalten wir

$$\left(\begin{array}{ccc|c} 1 & -3 & 0 & 2 \\ 0 & 9 & 0 & -18 \\ 0 & 0 & 1 & -9 \end{array}\right)$$

und schließlich

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & -4 \\ 0 & 1 & 0 & -2 \\ 0 & 0 & 1 & -9 \end{array}\right)$$

Den eigentlichen Schnittpunkt erhält man wie immer dadurch, daß man $t = -4, s_1 = -2, s_2 = -9$ in die Geradenfunktion und zur Kontrolle auch in die Ebenenfunktion einsetzt. In beiden Fällen erhalten wir

$$g(-4) = \begin{pmatrix} -15 \\ -6 \\ 7 \end{pmatrix} = e(-2, -9)$$

Zusammenfassend halten wir fest, daß man mit Schnittproblemen praktisch alle Situationen antreffen kann, die beim Lösen linearer Gleichungssysteme auftreten können. Bei unterbestimmten Gleichungssystemen enthält die Lösungsmenge typischerweise unendlich viele Elemente, da mindestens eine Unbekannte unbestimmt bleibt. Bei überbestimmten Systemen besteht dagegen typischerweise die Gefahr, daß widersprüchliche Bedingungen auftreten und die Lösungsmenge damit leer ist. Enthält das Gleichungssystem dagegen genauso viele Unbekannte wie Gleichungen, so besteht die Lösungsmenge typischerweise aus genau einem Element. In allen Fällen ist es aber auch möglich, daß die Lösungsmengenstruktur vom typischen Fall abweicht. Diese Spezialsituationen treten dann auf, wenn die Spalten (für $n \leq m$) bzw. die Zeilen (für $m \leq n$) der Koeffizientenmatrix linear abhängig sind (entspricht Parallelität bei Schnittproblemen).

6. Invertierung linearer Abbildungen

Während die geometrischen Schnittprobleme gut als Gedankenstütze für mögliche Situationen beim Lösen von linearen Gleichungssystemen dienen können, so stellen sie natürlich nicht die einzige Quelle von solchen Systemen dar. Lineare Systeme treten tatsächlich in sehr vielen Anwendungsfällen auf.
Wenn in einem solchen Fall das Gleichungssystem nur einmal gelöst werden soll, so kann man diese Lösung z. B. mit dem Gaußalgorithmus konstruieren. Soll die Lösung dagegen für viele verschiedene rechte Seiten aber die *gleiche* Matrix bestimmt werden, dann bietet sich

eine andere Sichtweise an. In diesem Fall hilft es, den Lösungsprozeß als Abbildung zu betrachten, die jeder rechten Seite die Lösung des Gleichungssystems zuordnet. Natürlich geht das nur, wenn das System eindeutige Lösungen liefert, da sonst nicht klar ist, welche Lösung der rechten Seite zugeordnet werden soll. Wenn das Gleichungssystem aus der Koordinatendarstellung der linearen Gleichung $L(\vec{v}) = \vec{w}$ mit $L : V \to W$ folgt, so interessieren wir uns letztlich für die Frage, wann die Abbildung L invertierbar ist, denn bei Invertierbarkeit gibt es ja gerade zu jedem $\vec{w} \in W$ ein eindeutiges $\vec{v} = L^{-1}(\vec{u})$ mit $L(\vec{v}) = \vec{w}$. Nehmen wir für einen Moment an, daß L invertierbar ist. Dann sieht man leicht, daß L^{-1} auch linear ist, denn für $\vec{w} = L(\vec{v})$ und $\alpha \in \mathbb{R}$ gilt $\alpha\vec{w} = L(\alpha\vec{v})$, so daß

$$L^{-1}(\alpha\vec{w}) = \alpha\vec{v} = \alpha L^{-1}(\vec{w})$$

Entsprechend zeigt man, daß auch $L^{-1}(\vec{w}+\vec{u}) = L^{-1}(\vec{w})+L^{-1}(\vec{w})$ gilt. Um eine lineare Abbildung vollständig zu verstehen, genügt es aber, die Bilder der Vektoren einer beliebigen Basis $(\vec{w}_1, \ldots, \vec{w}_m)$ zu kennen, da alle anderen Bilder als Linearkombination der $L^{-1}(\vec{w}_j)$ dargestellt werden können. Die Berechnung von $L^{-1}(\vec{w}_j)$ entspricht nun der Lösung der Gleichung $L(\vec{v}) = \vec{w}_j$, d. h. zur Bestimmung von L^{-1} müssen m Gleichungen gelöst werden (die V-Koordinaten dieser Lösungen stehen dann in den Spalten der Matrix zu L^{-1}). Die Konstruktion der inversen Abbildung ist also sicher erst dann rentabel, wenn das System mehr als m-mal gelöst werden soll, wobei m die Dimensionalität der rechten Seite ist.

Um Kriterien für die Invertierbarkeit linearer Abbildungen zu finden, erinnern wir uns, daß Invertierbarkeit sowohl Injektivität als auch Surjektivität voraussetzt. Dabei ist L injektiv, wenn zu jedem $\vec{w} \in W$ *höchstens ein* Urbild $\vec{v}$ korrespondiert, also falls $\vec{v}_1 \neq \vec{v}_2$ auch $L(\vec{v}_1) \neq L(\vec{v}_2)$ impliziert. Anders ausgedrückt, lautet diese Bedingung, wenn $\vec{u} = \vec{v}_1 - \vec{v}_2 \neq \vec{0}$, gilt, dann soll $L(\vec{u}) = L(\vec{v}_1) - L(\vec{v}_2) \neq \vec{0}$ gelten, d. h. L ist injektiv, wenn die einzige Nullstelle bei $\vec{0}$ ist. Die Menge aller Nullstellen einer linearen Abbildung L bezeichnet man als

$$\text{Kern } L = \{\vec{v} \in V | L(\vec{v}) = \vec{0}\}$$

Mit dieser Notation ist L also injektiv genau dann, wenn Kern $L = \{\vec{0}\}$. Im Gegensatz zur Injektivität bedeutet Surjektivität, daß zu jedem $\vec{w} \in W$ *mindestens ein* $\vec{v} \in V$ existiert mit $L(\vec{v}) = \vec{w}$, also daß der Wertebereich von L dem ganzen Raum W entspricht. Den Wertebereich einer linearen Abbildung L bezeichnet man auch als

$$\text{Bild } L = \{L(\vec{v}) | \vec{v} \in V\}$$

und damit ist Surjektivität gleichbedeutend mit der Bedingung Bild $L = W$. Bei linearen Abbildungen zwischen endlich dimensionalen Vektorräumen gibt es einen nützlichen Zusammenhang, die sogenannte Dimensionsformel, mit dem man den Test auf Surjektivität durch einen Test auf Injektivität ersetzen kann und umgekehrt. Die Grundlage dazu ist zunächst die Beobachtung, daß sowohl Kern L als auch Bild L Untervektorräume von V bzw. W sind. Nehmen wir an, daß $\vec{v}_1, \ldots, \vec{v}_s$ eine Basis von Kern L ist, d. h. dim Kern $L = s$. Ergänzen wir diese Basis zu einer Basis $\vec{v}_1, \ldots, \vec{v}_s, \vec{v}_{s+1}, \ldots, \vec{v}_n$ von V, so können wir zeigen, daß $L(\vec{v}_{s+1}), \ldots, L(\vec{v}_n)$ linear unabhängig in W sind. Nehmen wir dazu an, daß

$$\lambda_{s+1} L(\vec{v}_{s+1}) + \ldots + \lambda_n L(\vec{v}_n) = \vec{0}$$

gilt, mit dem Ziel, $\lambda_{s+1} = \ldots = \lambda_n = 0$ zu zeigen. Da L linear ist, folgt zunächst

$$L(\lambda_{s+1}\vec{v}_{s+1} + \ldots + \lambda_n \vec{v}_n) = \vec{0}$$

d. h. die Linearkombination ist eine Nullstelle von L und damit in Kern L enthalten. Da $\vec{v}_1, \ldots, \vec{v}_s$ eine Basis des Kerns ist, gibt es eindeutige Koeffizienten $\lambda_1, \ldots, \lambda_s$, so daß

$$\lambda_1\vec{v}_1 + \ldots + \lambda_s\vec{v}_s = \lambda_{s+1}\vec{v}_{s+1} + \ldots + \lambda_n\vec{v}_n$$

Dabei müssen nun aber alle λ_i identisch Null sein, denn $\vec{v}_1, \ldots, \vec{v}_n$ ist eine Basis und erlaubt daher nur die triviale Darstellung der Null. Insbesondere gilt also $\lambda_{s+1} = \ldots = \lambda_n = 0$ und die lineare Unabhängigkeit von $L(\vec{v}_{s+1}), \ldots, L(\vec{v}_n)$ ist gezeigt. Weiterhin sehen wir schnell, daß diese Vektoren den Raum Bild L erzeugen, denn das Bild eines beliebigen Vektors

$$\vec{v} = \mu_1\vec{v}_1 + \ldots + \mu_s\vec{v}_s + \mu_{s+1}\vec{v}_{s+1} + \ldots + \mu_n\vec{v}_n$$

ist ja gerade

$$L(\vec{v}) = \mu_{s+1} L(\vec{v}_{s+1}) + \ldots + \mu_n L(\vec{v}_n)$$

Die Vektoren stellen also ein linear unabhängiges Erzeugendensystem von Bild L dar und sind damit eine Basis. Für die Dimension folgt dim Bild $L = n - s$ und da dim Kern $L = s$,

$$\text{dim Bild } L + \text{dim Kern } L = \text{dim } V$$

Diese Dimensionsformel erleichtert nun z. B. das Nachprüfen von Surjektivität. Die Abbildung L ist surjektiv, wenn Bild $L = W$ gilt, d. h. wenn

$$\text{dim } V - \text{dim } W = \text{dim Kern } L$$

Die Frage nach der Surjektivität reduziert sich damit auf eine Dimensionsuntersuchung von Kern L. Da dim Kern $L \geq 0$, kann L offensichtlich nur dann surjektiv sein, wenn der Ausgangsraum eine höhere Dimension hat als der Zielraum. Stellen wir durch Dimensionsuntersuchung fest, daß dim Kern L = dim V − dim W gilt, so sagt uns die Dimensionsformel, daß dim Bild L = dim W ist und damit folgt dann tatsächlich Bild $L = W$, da Bild L ja durch eine maximale linear unabhängige Menge von Vektoren aus W erzeugt wird. Umgekehrt läßt sich die Injektivität (dim Kern $L = 0$) mit Hilfe der Dimensionsformel auf eine Dimensionsuntersuchung von Bild L zurückführen. Soll die Abbildung L sowohl injektiv als auch surjektiv sein, so muß dim Bild L = dim W gelten und dim Kern $L = 0$. Als notwendiges Kriterium ergibt die Dimensionsformel deshalb dim V = dim W. Nur wenn Dimension von Start und Zielraum übereinstimmen, kann die lineare Abbildung invertierbar sein. Ist dies der Fall und haben wir z. B. die Injektivitätsbedingung Kern $L = \{\vec{0}\}$ überprüft, so zeigt die Dimensionsformel, daß auch dim Bild L = dim V = dim W gilt. Die Invertierbarkeit einer linearen Abbildung läßt sich also einfach durch den Vergleich zweier Zahlen (den Dimensionen von Start- und Zielraum) und die Bestimmung von Kern L überprüfen. Wie bestimmt man denn eigentlich Kern L? Nun, dazu muß man alle $\vec{v} \in V$ finden, für die $L(\vec{v}) = \vec{0}$ gilt. Mit anderen Worten, wir müssen eine lineare Gleichung mit spezieller rechter Seite lösen und das können wir ja (z. B. mit dem Gaußalgorithmus). Zunächst führen wir Basen ein und stellen wie gewohnt das Gleichungssystem auf.
Da die rechte Seite identisch Null ist, werden Zeilenoperationen hier keine Veränderung hervorrufen. Nach Durchführung des Gaußalgorithmus wird daher nie ein Widerspruch auftreten, wie wir ihn bei den Schnittproblemen gesehen haben. Das ist aber auch nicht verwunderlich, da $\vec{0}$ immer eine Lösung des Problems darstellt und die Lösungsmenge des Systems nicht leer sein kann. Es verbleiben damit zwei Möglichkeiten. Im ersten Fall besteht die Lösungsmenge aus einem einzigen Element, was zwangsläufig der Nullvektor ist. Dies bedeutet Kern $L = \{\vec{0}\}$ und L ist damit injektiv.
Enthält die Lösungsmenge dagegen unendlich viele Elemente, so ist die Abbildung nicht injektiv und die Anzahl der frei wählbaren Parameter entspricht der Dimension von Kern L.
Ist z. B.

$$A = \begin{pmatrix} 1 & 2 & -1 \\ -1 & 1 & -1 \\ 1 & 2 & 1 \end{pmatrix}$$

die Matrix zu einer linearen Abbildung L, so sehen wir an der *quadratischen* Form, daß $\dim V = \dim W = 3$ gilt. Um die Invertierbarkeit zu überprüfen, lösen wir das System

$$\left(\begin{array}{ccc|c} 1 & 2 & -1 & 0 \\ -1 & 1 & -1 & 0 \\ 1 & 2 & 1 & 0 \end{array}\right)$$

Elimination liefert

$$\left(\begin{array}{ccc|c} 1 & 2 & -1 & 0 \\ 0 & 3 & -2 & 0 \\ 0 & 0 & 2 & 0 \end{array}\right)$$

und da alle Diagonalelemente von Null verschieden sind, können wir das System weiter vereinfachen zu

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{array}\right)$$

Hier können wir leicht ablesen, daß es nur die Null-Lösung gibt, d. h. Kern $L = \{\vec{0}\}$ und L ist somit invertierbar. Sie sehen auch, daß das Mitführen der rechten Seite nur unnötig Schreibarbeit erfordert und deshalb unterdrückt werden kann.
Wir haben ja bereits diskutiert, daß die Matrix zur inversen Abbildung L^{-1} durch das Lösen von $\dim V = \dim W$ Gleichungssystemen ermittelt werden kann. Es müssen nämlich die Koordinaten von $\vec{u}_j = L^{-1}(\vec{w}_j)$ gefunden werden, d. h. die Gleichungen $L(\vec{u}_j) = \vec{w}_j$ müssen gelöst werden. Bezeichnen wir die unbekannten Koordinaten mit $\vec{x}_j \in \mathbb{R}^{3\times 1}$ und beachten wir, daß die Koordinaten von $\vec{w}_1, \vec{w}_2, \vec{w}_3$ gerade die kanonischen Basisvektoren in $\mathbb{R}^3$ sind, so ergeben sich die Gleichungssysteme

$$A\vec{x}_j = \vec{e}_j \qquad j = 1, 2, 3$$

Wir müssen also den Gaußalgorithmus mit drei verschiedenen rechten Seiten durchführen. Da die benötigten Umformungen aber durch die Matrix vorgegeben werden und bei jeder rechten Seite die gleichen sind, können wir die Lösungen simultan ermitteln. Wir tragen dazu zunächst rechts vom Strich alle drei rechten Seiten ein

$$\left(\begin{array}{ccc|ccc} 1 & 2 & -1 & 1 & 0 & 0 \\ -1 & 1 & -1 & 0 & 1 & 0 \\ 1 & 2 & 1 & 0 & 0 & 1 \end{array}\right)$$

und führen dann die Elimination durch, indem wir zunächst die erste Zeile zur zweiten addieren und von der dritten abziehen

$$\left(\begin{array}{ccc|ccc} 1 & 2 & -1 & 1 & 0 & 0 \\ 0 & 3 & -2 & 1 & 1 & 0 \\ 0 & 0 & 2 & -1 & 0 & 1 \end{array}\right)$$

jetzt multiplizieren wir die erste Zeile mit zwei und addieren dann die letzte Zeile zu den anderen

$$\left(\begin{array}{ccc|ccc} 2 & 4 & 0 & 1 & 0 & 1 \\ 0 & 3 & 0 & 0 & 1 & 1 \\ 0 & 0 & 2 & -1 & 0 & 1 \end{array}\right)$$

Addition des (-4)-fachen der zweiten Zeile zum Dreifachen der ersten ergibt

$$\left(\begin{array}{ccc|ccc} 6 & 0 & 0 & 3 & -4 & -1 \\ 0 & 3 & 0 & 0 & 1 & 1 \\ 0 & 0 & 2 & -1 & 0 & 1 \end{array}\right)$$

Aufräumen durch passende Zeilenmultiplikation führt denn zum Endergebnis

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & \frac{1}{2} & -\frac{2}{3} & -\frac{1}{6} \\ 0 & 1 & 0 & 0 & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 1 & -\frac{1}{2} & 0 & \frac{1}{2} \end{array}\right)$$

In den Spalten rechts vom Strich können wir die drei Lösungen jetzt direkt ablesen, die ja in die Spalten der Matrix zu L^{-1} eingetragen werden. Diese Matrix bezeichnen wir mit A^{-1}

$$A^{-1} = \begin{pmatrix} \frac{1}{2} & -\frac{2}{3} & -\frac{1}{6} \\ 0 & \frac{1}{3} & \frac{1}{3} \\ -\frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}$$

Zur Probe können wir die Spalten von A^{-1} mit A multiplizieren, was ja die kanonischen Basisvektoren ergeben sollte. Auch diese Operationen kann man simultan durchführen indem wir das Matrixprodukt AA^{-1} berechnen

$$AA^{-1} = \begin{pmatrix} 1 & 2 & -1 \\ -1 & 1 & -1 \\ 1 & 2 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{2} & -\frac{2}{3} & -\frac{1}{6} \\ 0 & \frac{1}{3} & \frac{1}{3} \\ -\frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Diese Relation entspricht übrigens der Verkettung $L \circ L^{-1} = J_W$, wobei J_W die Identitätsabbildung auf W ist. Genauso gilt natürlich $L^{-1} \circ L = J_V$, was dem Matrixprodukt $A^{-1}A = E$ entspricht, mit der Einheitsmatrix $E \in \mathbb{R}^{3\times 3}$.
Wenn zwischen zwei Vektorräumen V, W eine invertierbare lineare Abbildung existiert L, so sind die beiden Räume in gewisser Weise sehr ähnlich (man sagt auch *isomorph*). So sind z. B. $\vec{v}_1, \ldots \vec{v}_n$ linear unabhängig in V genau dann, wenn $L(\vec{v}_1), \ldots, L(\vec{v}_n)$ linear unabhängig in W sind. Genauso ist ein System von Vektoren $\vec{w}_1, \ldots, \vec{w}_r$ ein Erzeugendensystem in W, genau dann, wenn $L^{-1}(\vec{w}_1), \ldots, L^{-1}(\vec{w}_r)$ den Raum V erzeugt. Außerdem werden Untervektorräume von V in gleichdimensionale Untervektorräume von W abgebildet und umgekehrt. Wir haben diese *Isomorphie* übrigens beim Rechnen mit Koordinaten bereits nutzbringend eingesetzt. Hier ist die invertierbare lineare Abbildung die Koordinatenabbildung, die einen n-dimensionalen Vektorraum V in den Raum $\mathbb{R}^{n\times 1}$ abbildet. Sie erlaubt uns, mit reellen Zahlen und den zugehörigen Grundoperationen zu arbeiten anstelle der jeweiligen Operationen Addition und skalare Multiplikation in V. Da V zu $\mathbb{R}^{n\times 1}$ isomorph ist, kann man also alle Eigenschaften wie lineare Unabhängigkeit, Erzeugendeneigenschaft, Untervektorraumeigenschaft, Linearität von Abbildung an den Koordinaten überprüfen. Die Antworten gelten wegen Isomorphie dann auch für die zugeordneten Vektoren im Vektorraum V.
Obwohl man in diesem Sinne V und $\mathbb{R}^{n\times 1}$ identifizieren kann, ist es doch ratsam, die beiden Räume konzeptionell zu trennen. Oft ist es so, daß man im Ausgangsraum V Strukturen klarer erkennt als in Koordinaten, die ja von der willkürlichen Wahl einer Basis abhängen. Als Beispiel betrachten wir die Abbildung $T_h : \mathcal{P}_n \to \mathcal{P}_n$, wobei $T_h(p)$ das um $h \in \mathbb{R}$ „verschobene“ Polynom $[T_h(p)](x) = p(x - h)$ darstellt. Um nachzuweisen, daß T_h invertierbar ist, müssen wir wegen der Dimensionsformel nur nachprüfen, ob Kern $T_h = \{N\}$ gilt, wobei N das Nullpolynom ist. Wir wissen, daß man die lineare Gleichung $T_h(p) = N$ in ein Gleichungssystem verwandeln kann, wenn man eine Basis einführt und zu Koordinatenvektoren übergeht, also die Isomorphie zu $R^{(n+1)\times 1}$ ausnutzt. Es ist jedoch ratsam, die Koordination hier *nicht* zu benutzen. Die Gleichung $T_h(p) = N$ impliziert ja $p(x - h) = 0$ für alle $x \in \mathbb{R}$

und damit auch $p(y) = 0$ für alle $y \in \mathbb{R}$ (man muß nur $y + h$ für x einsetzen).
Es folgt also sofort Kern $T_h = \{N\}$, ohne auch nur einen Koordinatenvektor auszurechnen oder über eine geeignete Basis nachdenken zu müssen. Versuchen Sie also so lange wie möglich, Koordinaten zu vermeiden!
Als abschließendes praktisches Beispiel für eine Isomorphie betrachten wir die Menge $\mathcal{F}$ aller Punktkräfte. Eine Punktkraft ist dabei eine (idealisierte) Kraft, die in einem Punkt eines materiellen Objekts im Raum angreifen kann. Denken Sie z. B. an Ihre Muskelkraft, die Sie mit einer feinen Spitze an einem Körper wirken lassen können, oder an die von einem Gewicht erzeugte Gewichtskraft, die über ein Seil wirkt, das an einem feinen Haken in einem Punkt eines Körpers befestigt ist. Unabhängig vom jeweiligen Ursprung der Punktkraft (Gewichtskraft, Federkraft, Muskelkraft, ...) betrachten wir zwei Punktkräfte als identisch, wenn sie im gleichen Punkt die gleiche Kraftwirkung (Verformung, Beschleunigung) hervorrufen.
Tatsächlich können wir Kräfte überhaupt nur an ihren Wirkungen erkennen. Um diese Wirkungen zu quantifizieren, benötigen wir ein Meßinstrument. Eine Möglichkeit ist dabei, Kräfte mit Hilfe einer (idealisierten) dünnen, beliebig dehnbaren Feder in *Zeiger* zu verwandeln. Wir halten dazu die Feder in ihrem Anfangspunkt A fest so daß sie sich frei um diesen Punkt drehen kann. Dann lassen wir eine Punktkraft am Endpunkt der Feder wirken, so daß sich die Feder *dehnt* (versucht die Kraft, die Feder zusammenzudrücken, so entspricht dies einer Zugkraft, wenn man die Feder in die entgegengesetzte Richtung zeigen läßt). Durch die Kraftwirkung wird das Ende der Feder einen Punkt E im Raum annehmen und sich die Länge von L auf L' verlängern. Einer Punktkraft $F \in \mathcal{F}$ läßt sich so also ein Zeiger

$$M(F) = \frac{L' - L}{L'} \overrightarrow{AE}$$

zuordnen, d. h. wir haben eine Abbildung $M : \mathcal{F} \to S$ konstruiert. Da wir die Feder als beliebig dehnbar angenommen haben, ist die Abbildung M surjektiv. Durch Einwirkung beliebig kleiner und großer Kräfte können beliebig lange und kurze Zeiger gemessen werden und im Prinzip kann jedes Element von S auftreten. (Für eine reale nicht idealisierte Feder gilt dies natürlich nicht.) Die Abbildung M ist auch injektiv wegen unserer Definition der Gleichheit von Punktkräften. Gilt nämlich $M(F_1) = M(F_2)$, d. h. haben F_1 und F_2 die gleiche Wirkung, so ist $F_1 = F_2$, bzw. umgekehrt impliziert $F_1 \neq F_2$, daß $M(F_1) \neq M(F_2)$ gilt.

Auf der Menge $\mathcal{F}$ sind in natürlicher Weise zwei Operationen gegeben. Die erste Operation ordnet zwei Kräften $F, G \in \mathcal{F}$ die *resultierende* Kraft zu, die entsteht, wenn F und G im selben Punkt angreifen. Experimentell stellt sich heraus, daß diese Operation Additionsregeln gehorcht. Deshalb wird die resultierende Kraft auch als $F+G$ bezeichnet. Das Ergebnis von Experimenten ist dann (streng genommen mit gewisser Idealisierung)

$$M(F+G) = M(F) + M(G)$$

wobei sich das linke + Zeichen auf Kraftüberlagerung und das rechte auf Verkleben von Zeigern bezieht.
Die zweite Operation ordnet einer reellen Zahl α und einer Kraft $F \in \mathcal{F}$ die Kraft zu, die die α-fache Wirkung hat. Diese Kraft wird mit αF bezeichnet. Es gilt also *per Definition*

$$M(\alpha F) = \alpha M(F)$$

Damit ist $M : \mathcal{F} \to S$ eine bijektive Abbildung, bei der reelle Faktoren aus dem Argument herausgezogen werde dürfen und die der resultierenden Kraft aus der Überlagerung von F und G die Summe der Bilder $M(F)$ und $M(G)$ zuordnet. Beachten Sie, daß wir M streng genommen noch nicht als lineare Abbildung bezeichnen dürfen, da wir uns noch nicht überzeugt haben, ob denn $\mathcal{F}$ mit den beiden Operationen einen Vektorraum bildet. Um diese Überprüfung durchzuführen, müssen wir einfach die Vektorraum-Rechenregeln nachprüfen. Als Beispiel betrachten wir das Kommutativgesetz der Addition. Es gilt

$$M(F+G) = M(F) + M(G) = M(G) + M(F) = M(G+F)$$

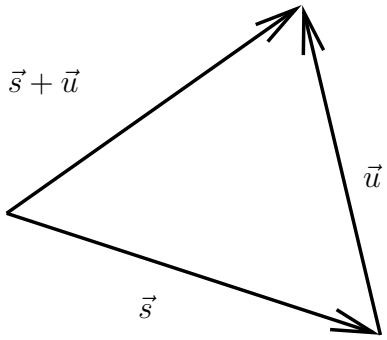
und da M injektiv ist, folgt $F+G = G+F$. Die anderen Rechenregeln folgen mit einer ähnlichen Schlußweise. Beachten Sie, daß neutrale und inverse Elemente durch die Definition des Produkts αF praktisch festgelegt sind, da nur $0F$ für das neutrale Element und $(-1)F$ für inverse Elemente möglich ist.
Nach dieser Überprüfung können wir also festhalten, daß der Raum $\mathcal{F}$ der (idealisierten) Punktkräfte einen reellen Vektorraum bildet und *isomorph* zum Raum aller Zeiger ist. Wir können daher statt mit Kräften zu operieren, auch rein geometrisch mit Zeigern argumentieren, wenn es um Vektrorraumfragestellungen geht, also z. B. um Superpositionen, Zerlegungen (bzw. allgemein um Linearkombinationen), lineare Unabhängigkeit oder Abhängigkeit, Erzeugendeneigenschaft und so weiter.

Diese Tatsache lernt man schon sehr früh in der Schule und durch häufige Wiederholung kann man kaum noch anders über Kräfte denken als in Form von Zeigern.
Beachten Sie aber, daß Zeiger und Punktkräfte *verschiedene* Objekte sind und daß Addition von Zeigern etwas anderes bedeutet als Addition von Kräften. Isomorphie bedeutet eben nicht vollständige Gleichheit, sondern nur gleiches Verhalten bei linearen Operationen.

7. Längen und Winkel

In diesem Abschnitt geht es um geometrische Konzepte in Vektorräumen. Zunächst scheinen dabei Längen und Winkel, aber auch Flächeninhalte und Volumina Konstruktionen zu sein, die *nur* im Zeigervektorraum eine Bedeutung haben, da hier die Vektoren ja mit elementargeometrischen Strecken in Verbindung gebracht werden können. Es stellt sich aber heraus, daß die wesentlichen Eigenschften z. B. von Längen bzw. Abständen auch auf andere Vektorräume übertragen werden können und dort sehr nützlich sind. Ein Abstand zwischen zwei Vektoren in einem Funktionen-Vektorraum kann etwa die Verschiedenheit der beiden Funktionen quantifizieren und dann bei Approximationsproblemen benutzt werden: Finde eine Funktion P in einer Teilmenge des Funktionenraums (z. B. in der Menge $\mathcal{P}$ der Polynome), die möglichst *nahe* an einer gegebenen Funktion f ist. Hat man ein solches P gefunden, so kann man, natürlich mit gewissen Abstrichen an die Genauigkeit, mit P anstelle von f weiterarbeiten (z. B. zur schnelleren Berechnung, zur Integration, zur Differentiation etc.). Im Vektrorraum der Bestellungen (Abschnitt 4) kann ein Abstandsbegriff dazu benutzt werden, zu einer neuen Bestellung eine ähnliche, schon abgewickelte Bestellung zu finden, um z. B. die Preisgestaltung ähnlich durchzuführen. Die Länge einer Punktkraft wird man, wegen der Identifikation zwischen Zeigern und Kräften, natürlich mit der Stärke der Kraft in Verbindung bringen. Die Länge einer Matrix A bzw. einer linearen Abbildung L könnte dagegen quantifizieren, wie stark sich die Länge eines Vektors $\vec{v}$ von der Länge seines Bildes $L(\vec{v})$ unterscheidet, d. h. die Länge einer Matrix könnte ihr maximaler „Streckfaktor“ sein.
Um alle diese Längenbegriffe unter einen Hut zu bringen, wollen wir nun einige charakteristische Eigenschaften der elementargeometrischen Länge im Zeigervektorraum S herausarbeiten, und diese dann als Kriterien für ein allgemeines Längenkonzept fordern.
Die Funktion, die jedem Zeiger $\vec{s} \in S$ seine elementargeometrische Länge zuordnet, bezeichnen wir mit $||\cdot||$. Der Wert $||\vec{s}||$ ist also die Länge von $\vec{s}$ und als solche immer nicht-negativ. Wir stellen damit

als erste Eigenschaft fest, daß $||\cdot|| : S \to [0,\infty)$ gilt. Eine weitere offensichtliche Tatsache ist, daß, abgesehen vom Nullzeiger, jeder Zeiger zwei verschiedene Punkte im Raum verbindet und damit eine positive Länge hat. Nur der Nullzeiger hat Länge Null. Wenn wir uns erinnern wie die skalare Multiplikation in S definiert war, so sehen wir, daß die Länge von $\alpha\vec{s}$ gerade das $|\alpha|$-fache der Länge von $\vec{s}$ ist, denn $\alpha\vec{s}$ ist ja der um den Faktor $|\alpha|$ in der Länge geänderte Zeiger $\vec{s}$, bei dem noch die Richtung umgekehrt wird, falls α negativ ist. Es gilt also $||\alpha\vec{s}|| = |\alpha|\ ||\vec{s}||$ für die Längenfunktion, egal wie $\alpha \in \mathbb{R}$ und $\vec{s} \in S$ gewählt sind. Die letzte offensichtliche Eigenschaft von Zeigerlängen bezieht sich auf das Zusammenspiel mit der Verklebeoperation. Sind $\vec{s}$ und $\vec{u}$ zwei Zeiger, so ist $\vec{s}+\vec{u}$ der Zeiger, der das stumpfe Ende von $\vec{s}$ mit dem spitzen Ende von $\vec{u}$ verbindet, wobei das stumpfe Ende von $\vec{u}$ an der Spitze von $\vec{s}$ befestigt wird. Die drei Zeiger $\vec{s}, \vec{u}$ und $\vec{s}+\vec{u}$ bilden somit geometrisch ein Dreieck.



Offensichtlich ist die Länge von $\vec{s}+\vec{u}$ nie größer als die der Knick-Konstruktion, bestehend aus $\vec{s}$ und $\vec{u}$, d. h. es gilt die sogenannte *Dreiecksungleichung* $||\vec{s}+\vec{u}|| \leq ||\vec{s}|| + ||\vec{u}||$. Letztlich liegt diese Eigenschaft an einer Modellannahme über unseren physikalischen Raum: Die kürzeste Verbindung zwischen zwei Punkten ist die gerade Verbindung. Die Gleichheit in der Dreiecksungleichung kann natürlich auch auftreten. Denken Sie nur an den Fall

$$||\vec{s}+\vec{s}|| = ||2\vec{s}|| = 2||\vec{s}|| = ||\vec{s}|| + ||\vec{s}||$$

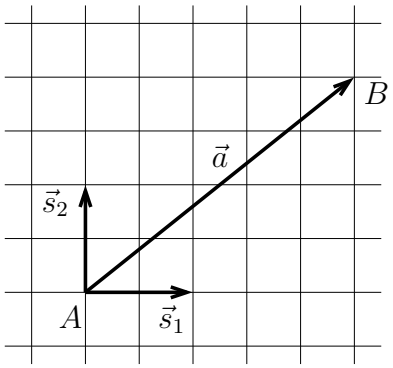
Wenn wir die bisher beschriebenen Eigenschaften etwas abstrakter betrachten, so sehen wir, daß sie das Zusammenspiel des Längenbegriffs mit den Vektorraumoperationen beschreiben. Insbesondere lassen sich die Eigenschaften in jedem reellen Vektorraum formulieren.

Definition 1. *Sei V ein reeller Vektorraum. Eine Funktion $||\cdot|| : V \longrightarrow [0,\infty)$ heißt Norm auf V, wenn folgende Bedingungen erfüllt sind*

(i) $||\vec{v}|| = 0$ *nur für* $\vec{v} = \vec{0}$
(ii) $||\alpha\vec{v}|| = |\alpha|\ ||\vec{v}||$ *für alle* $\alpha \in \mathbb{R}, \vec{v} \in V$
(iii) $||\vec{v} + \vec{u}|| \leq ||\vec{v}|| + ||\vec{u}||$ *für alle* $\vec{u}, \vec{v} \in V$

Ein Vektorraum mit Norm nennt man auch normierter Vektorraum.

Im Fall des Zeigervektorraums S stellt sich nun die Frage, ob bereits alle charakterisierenden Eigenschaften der elementargeometrischen Länge duch die Bedingung $(i), (ii)$ und (iii) erfaßt sind. Daß dies *nicht* der Fall ist, zeigt das Beispiel der New-York-Taxifahrer-Norm. Stellen wir uns den Stadtplan von New York vor mit gleichmäßig parallel laufenden Straßen in Nord-Süd- und in Ost-West-Richtung.



Durch $(A, \vec{s}_1, \vec{s}_2)$ sei ein kartesisches Koordinatensystem in der Kartenebene gegeben.
Stellen wir uns weiter vor, daß ein Taxi vom Punkt A zum Punkt B fahren soll, wobei B durch den Ortsvektektor $\vec{a} = a_1\vec{s}_1 + a_2\vec{s}_2$ gegeben ist. Welche Strecke muß das Taxi dafür zurücklegen? Da das Taxi *nicht* die direkte Verbindung nehmen kann, sondern den Straßen folgen muß, ergibt sich als Distanz $|a_1| + |a_2|$. Offensichtlich ist dies ein vernünftiges Längenmaß mit einer praktischen Bedeutung (Ihr Geldbeutel wird proportional zu dieser Distanz belastet und nicht abhängig von der elementargeometrischen Distanz!). Wir definieren also für beliebige Zeiger in der Ebene $V = \{a_1\vec{s}_1 + a_2\vec{s}_2 | a_1, a_2 \in \mathbb{R}\}$

$$||a_1\vec{s}_1 + a_2\vec{s}_2||_1 = |a_1| + |a_2| \qquad a_1, a_2 \in \mathbb{R}$$

wobei die Normnotation eigentlich etwas verfrüht ist, da wir die Normeigenschaften noch nicht nachgeprüft haben. Zunächst gilt offensichtlich $||\vec{s}||_1 \geq 0$ für alle Zeiger $\vec{s} \in V$ und wenn $||a_1\vec{s}_1 + a_2\vec{s}_2||_1 = 0$ ist, dann sind mit $|a_1| + |a_2| = 0$ sowohl $|a_1|$ als auch $|a_2|$ gleich Null, was schließlich $a_1 = a_2 = 0$ impliziert. Mit anderen Worten, $||\vec{s}||_1 = 0$ ist nur im Fall $\vec{s} = \vec{0}$ möglich, womit die erste Bedingung an die Norm bereits überprüft ist. Die zweite Bedingung folgt mit

$$||\alpha(a_1\vec{s}_1+a_2\vec{s}_2)||_1 = |\alpha a_1|+|\alpha a_2| = |\alpha|\ |a_1|+|\alpha|\ |a_2| = |\alpha|\ ||a_1\vec{s}_1+a_2\vec{s}_2||_1$$

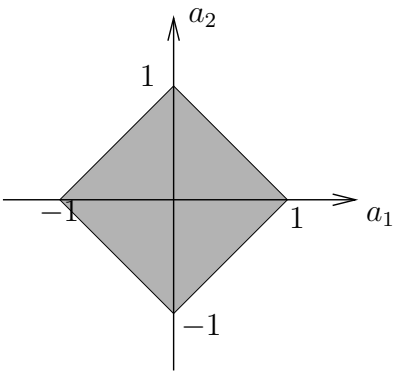
Die Dreiecksungleichung ergibt sich anschaulich folgendermaßen: Die Taxientfernung zu einem Punkt $\vec{s} + \vec{u}$ kann nämlich nicht dadurch verkürzt werden, daß man zunächst zu einem anderen Punkt $\vec{s}$ fährt und von dort weiter nach $\vec{s} + \vec{u}$. Bestenfalls liegt $\vec{s}$ auf einer direkten Taxiverbindung nach $\vec{s}+\vec{u}$ und die Entfernungen sind identisch. Wenn $\vec{s}$ aber ungünstig liegt, so ist die Fahrt via $\vec{s}$ länger. In mathematischer Sprache ergibt sich diese Tatsache aus der Dreiecksungleichung für den Betrag von reellen Zahlen. Tatsächlich ist der Raum $\mathbb{R}$ mit dem Betrag $|\cdot|$ ein normierter reeller Vektorraum! Die Vektoreigenschaften haben wir ja schon früher nachgeprüft ($\mathbb{R}$ als Spezialfall von $\mathbb{R}^n$ mit $n = 1$). Aus der Definition des Betrags folgt außerdem, daß $|x| \geq 0$ für alle $x \in \mathbb{R}$ und $|x| = 0$ kann nur im Fall $x = 0$ eintreten (ist nämlich $x > 0$, so gilt $|x| = x > 0$ und entsprechend impliziert $x < 0$, daß $|x| = -x > 0$ ist).
Außerdem folgt sofort, daß $|\alpha x| = |\alpha|\ |x|$ ist. Betrachten Sie einfach die vier möglichen Vorzeichenkombinationen und benutzen Sie die Definition des Betrags. Die Dreiecksungleichung $|x + y| \leq |x| + |y|$ kann man ebenfalls durch Fallunterscheidung der Vorzeichen von x und y und durch Benutzung der Rechenregeln für Ungleichungen nachweisen. Wenn Sie eine anschauliche Argumentation bevorzugen, können Sie $\mathbb{R}$ auch als Koordinatenraum einer Gerade in S betrachten (Zahlengerade). Ist $\vec{e}$ der Richtungsvektor der Gerade mit elementargeometrischer Länge Eins, so hat der Zeiger $x\vec{e}$ die Länge $|x|$. Da unsere Betrachtungen zur elementargeometrischen Länge für alle Zeiger gelten, so treffen sie auch auf die Elemente der Gerade zu, wo die Länge durch den Betrag der Koordinate gegeben ist. Insbesondere gelten also auch $(i), (ii)$ und (iii) für die Betragsfunktion auf den reellen Zahlen $\mathbb{R}$.
Kommen wir nun aber zurück zur Dreiecksungleichung der Taxifahrer-Norm. Es gilt für zwei Vektoren $\vec{s} = a_1\vec{s}_1 + a_2\vec{s}_2 \in V$ und $\vec{u} = b_1\vec{s}_1 + b_2\vec{s}_2 \in V$, daß

$$||\vec{s}+\vec{u}||_1 = |a_1+b_1|+|a_2+b_2| \leq (|a_1|+|b_1|)+(|a_2|+|b_2|) = ||\vec{s}||_1+||\vec{u}||_1$$

wobei wir zweimal die Dreiecksungleichung in $\mathbb{R}$ ausgenutzt haben.
Damit haben wir alle Normeigenschaften nachgewiesen und gesehen, daß es andere sinnvolle Längenbegriffe geben kann als den der Elementargeometrie.
Um das Verhalten der $||\cdot||_1$-Norm zu veranschaulichen, kann man alle Punkte der Ebene markieren, für die die Länge des zugehörigen Ortsvektors kleiner oder gleich einer festen Zahl (z. B. Eins) ist, d. h. alle Punkte, deren Abstand zum Referenzpunkt höchstens ein vorgegebener Wert ist. Für den Taxifahrer entspricht dies dem erreichbaren Gebiet mit einer gegebenen Tankfüllung. Die Menge zum Wert Eins wird allgemein *Einheitskugel* genannt, obwohl sie geometrisch nicht unbedingt kugel- bzw. kreisförmig sein muß. Im Fall der Taxifahrer-Norm müssen wir dazu

$$E = \{a_1\vec{s}_1 + a_2\vec{s}_2|\ |a_1| + |a_2| \leq 1\}$$

skizzieren. Wie man Lösungsmengen von Ungleichungen der Form $|a_1|+|a_2|-1 \leq 0$ findet, haben wir ja bereits in Kapitel 1 gelernt. Es ergibt sich ein rautenförmiges Gebiet



Die Mengen $E_c = \{\vec{v}|\ ||\vec{v}v||_1 \leq c\}$ für $c > 0$ sehen übrigens wegen der Eigenschaft (ii) genauso aus, nur die Größe ändert sich proportional zu c; ist $\vec{v} \in E$, dann ist $||c\vec{v}||_1 = |c|\ ||\vec{v}||_1 \leq c$ also $c\vec{v} \in E_c$.
Als weiteres Beispiel, wo eine nicht elementargeometrische Länge auftritt, betrachten wir die Auswertung einer Temperaturmeßreihe. Stellen wir uns vor, an jedem Tag im Juni wird um Punkt 12 Uhr die Außentemperatur gemessen und vermerkt. Diese Meßreihe liefert also

30 Temperaturwerte, d. h. wir können die Meßreihe durch einen Vektor $\vec{x}$ aus $\mathbb{R}^{30}$ beschreiben.
Die mittlere Mittagstemperatur ist dann durch

$$M(\vec{x}) = \frac{1}{30}\sum_{i=1}^{30} x_i$$

gegeben. Beachten Sie, daß $M : \mathbb{R}^{30} \to \mathbb{R}$ eine lineare Abbildung ist. Fragen wir nun nach der größten Abweichung vom Mittelwert, so müssen wir

$$\max\{|x_i - M(\vec{x})| \; | i = 1, \ldots, 30\}$$

ausrechnen. Mit dem Vektor $\vec{m} = (M(\vec{x}), \ldots, M(\vec{x}))$ können wir diese maximale Abweichnung auch als $||\vec{x} - \vec{m}||_\infty$ schreiben, wenn wir

$$||(v_1, \ldots, v_{30})||_\infty = \max\{|v_i| \; | i = 1, \ldots, 30\}$$

als Abkürzung einführen. Auch hier prüfen wir nach, ob die Norm-Notation gerechtfertigt ist. Da das Maximum über die Beträge der Komponenten von $\vec{v}$ genommen wird, gilt sicherlich $||\vec{v}||_\infty \geq 0$.
Ist $||\vec{v}||_\infty = 0$, so ist die betragsgrößte Komponente gleich Null und damit müssen alle Komponenten Null sein, d. h. $\vec{v} = \vec{0}$. Wegen

$$||\alpha(v_1, \ldots, v_n)||_\infty = \max_{i=1,\ldots,30} |\alpha v_i| = \max_{i=1,\ldots,30} |\alpha| \; |v_i| = |\alpha| \max_{i=1,\ldots,30} |v_i|$$

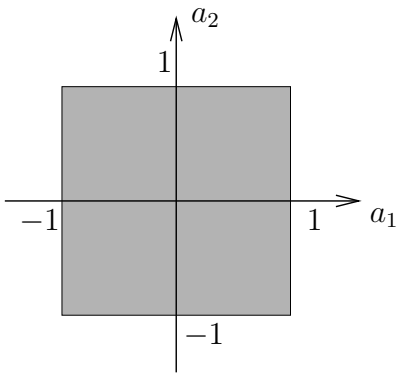
folgt auch Bedingung (ii). Die Dreiecksungleichung ergibt sich schließlich aus der Tatsache, daß $|v_i| \leq ||\vec{v}||_\infty$ für alle $i = 1, \ldots, 30$ gilt. Ist $\vec{u} = (u_1, \ldots, u_{30})$ ein weiterer Vektor, so folgt mit der Dreiecksungleichung in $\mathbb{R}$

$$\begin{aligned} ||\vec{v} + \vec{u}||_\infty &= \max_{i=1,\ldots,30} |v_i + u_i| \leq \max_{i=1,\ldots,30} (|v_i| + |u_i|) \\ &\leq \max_{i=1,\ldots,30} |v_i| + \max_{i=1,\ldots,30} |u_i|_\infty + ||\vec{u}||_\infty) = ||\vec{v}||_\infty + ||\vec{u}||_\infty \end{aligned}$$

Die sogenannte Maximum-Norm $||\cdot||_\infty$ liefert also die betragsmäßig größte Komponente eines Vektors und tritt deshalb überall da auf, wo es um maximale Abweichungen geht. Im $\mathbb{R}^n$ mit allgemeinem $n \in \mathbb{N}$ ist die Definition natürlich entsprechend zu modifizieren. Für den Fall $n = 2$ können wir zur Veranschaulichung wieder die Einheitskugel

$$E = \{(a_1, a_2) | \max\{|a_1|, |a_2|\} \leq 1\}$$

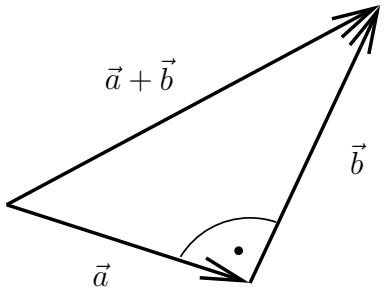
zeichnen. Die „Kugel“ ist hier das Quadrat $[-1, 1]^2$, da jede Koordinate unabhängig voneinander in $[-1, 1]$ variieren kann.



Kommen wir nun aber wieder zurück zur elementargeometrischen Länge im Zeigerraum. Da unsere bisher gefundenen Normbedingungen auch andere Längenbegriffe zulassen, muß die elementargeometrische Länge weitere Eigenschaften haben, die wir bisher noch nicht isoliert haben. Ein solches Charakteristikum ist der Zusammenhang der Länge mit *Winkeln*. Denken Sie nur an den Satz des Pythagoras über die Seitenlängen in einem *rechtwinkligen* Dreieck.
Sind $\vec{a}, \vec{b}$ zwei Zeiger, die im rechten Winkel zueinander stehen, so gilt für die Länge von $\vec{a} + \vec{b}$ eben

$$||\vec{a} + \vec{b}||^2 = ||\vec{a}||^2 + ||b||^2$$

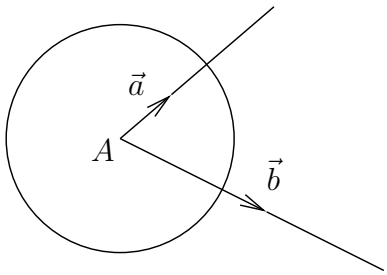
da $\vec{a}, \vec{b}$ und $\vec{a} + \vec{b}$ ein rechtwinkliges Dreieck bilden, wenn das stumpfe Ende von $\vec{b}$ an die Spitze von $\vec{a}$ gebracht wird und $\vec{a} + \vec{b}$ im stumpfen Ende von $\vec{a}$ beginnt



Diese Eigenschaft der elementargeometrischen Länge haben andere Normen im allgemeinen *nicht*. Schauen wir uns dazu einmal die Norm $||\cdot||_1$ an. Sind $\vec{s}_1, \vec{s}_2$ die Zeiger eines kartesischen Koordinatensystems (also senkrecht zueinander), so gilt

$$||\vec{s}_1 + \vec{s}_2||_1^2 = (|1| + |1|)^2 = 4 \neq 2 = |1|^2 + |1|^2 = ||\vec{s}_1||_1^2 + ||\vec{s}_2||_1^2$$

Der Zusammenhang zwischen elementargeometrischer Länge und Winkeln wird deutlicher, wenn wir den Begriff des Winkels etwas konkretisieren. Wie lassen sich Winkel überhaupt messen? Sind $\vec{a}$ und $\vec{b}$ zwei Vektoren mit positiver Länge, so messen wir den von ihnen eingeschlossenen Winkel $\sphericalangle(\vec{a},\vec{b})$ folgendermaßen: Wir bringen zunächst die beiden stumpfen Enden in einen gemeinsamen Punkt A. Dann schlagen wir in einer Ebene, die $\vec{a}$ und $\vec{b}$ enthält, einen Kreis mit Radius Eins um diesen Punkt.
Die beiden durch $\vec{a}$ und $\vec{b}$ gegebenen Halbgeraden $\{t\vec{a}|t \geq 0\}$ und $\{t\vec{b}|t \geq 0\}$ zerschneiden den Kreis nun in zwei Kreisbögen



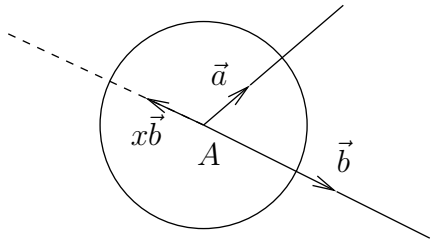
Die Länge des *kürzeren* der beiden Bögen können wir dann als Maß für den eingeschlossenen Winkel nehmen. Dies ist das sogenannte *Bogenmaß*. Sehen Sie, wie eng an dieser Stelle Länge und Winkel miteinander verbunden sind!
Traditionell nimmt man übrigens statt des Einheitskreises oft einen Kreis mit Umfang 360, also mit Radius $360/2\pi$. Das hat den Vorteil, daß man mit ganzzahligen Winkelwerten schon relativ feine Winkelabstufungen beschreiben kann. Die Umrechnung dieses Winkelgrad-Systems in unser Bogenmaßsystem ist aber sehr einfach. Ist φ das Bogenmaß eines Winkels, so ist $\varphi \cdot 360/2\pi$ der Bogen auf dem größeren Kreis, also die sogenannte Gradzahl des Winkels. Umgekehrt kommt man von einer Grundzahl Φ mit $\Phi 2\pi/360$ zum Bogenmaß des Winkels. Das Bogenmaß $\frac{\pi}{2}$ entspricht also $90°$, $\frac{\pi}{3}$ entspricht $60°$, $\frac{\pi}{4}$ entspricht $45°$ usw.
Durch die Meßvorschrift des Winkels zwischen zwei Vektoren $\vec{a},\vec{b}$ erhalten wir letztlich eine Abbildung

$$\sphericalangle(\cdot,\cdot) : S \setminus \{\vec{0}\} \times S \setminus \{\vec{0}\} \to [0,\pi]$$

Die Zielmenge können wir auf das Intervall $[0,\pi]$ einschränken, da der kürzere der beiden Kreisbögen maximal die Hälfte des Kreisumfangs 2π annehmen kann. Diese Winkelfunktion hat einige offensichtliche Eigenschaften

(i) $\sphericalangle(\vec{a},\vec{b}) = \sphericalangle(\vec{b},\vec{a})$
(ii) $\sphericalangle(\vec{a},x\vec{b}) = \sphericalangle(\vec{a},\vec{b})$ für $x > 0$
(iii) $\sphericalangle(\vec{a},x\vec{b}) = \pi - \sphericalangle(\vec{a},\vec{b})$ für $x < 0$
(iv) $\sphericalangle(\vec{a},\vec{a}) = 0$

Beachten Sie zum Verständnis der Bedingung (*iii*), daß die beiden Halbgeraden mit $\vec{b}$ und $x\vec{b}$ im Fall $x < 0$ zusammen eine Gerade bilden, die den Einheitskreis in zwei Hälften zerschneidet.

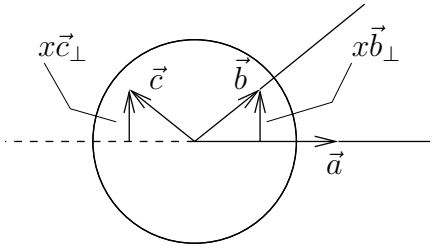


Einen dieser beiden Bögen der Länge π zerschneidet die Halbgerade entlang $\vec{a}$ dann in zwei Winkel, d. h.

$$\sphericalangle(\vec{a},\vec{b}) + \sphericalangle(\vec{a},x\vec{b}) = \pi$$

woraus Bedingung (*iii*) folgt. Zur Bedingung (*iv*) sei angemerkt, daß hier die Auswahl der Ebene, in der der Kreis gebildet wird, nicht eindeutig festgelegt ist. Unabhängig davon, wie wir diese Ebene wählen, ist der ausgeschnittene Kreisbogen aber immer von der Länge 0.
Mit der eleganten Definition von $\sphericalangle(\vec{a},\vec{b})$ haben wir die Winkelmessung durch eine Längenmessung ersetzt. Die ganze Sache hat nur einen Haken, der Ihnen vielleicht schon aufgefallen ist: Wie kann man denn die Länge eines *Bogens* messen? Sie kennen wahrscheinlich folgenden Trick: Man nimmt eine Schnur, legt sie entlang des Kreisbogens und markiert die beiden Stellen, wo die Halbgeraden schneiden. Danach zieht man die Schnur gerade und kann dann in gewohnter Weise (mit einer Vergleichsstrecke, dem Lineal) die Länge ablesen. Die Genauigkeit dieser Prozedur hängt aber unter anderem davon ab, wie genau die Schnur entlang des Kreisbogens läuft und der schwierige Teil ist ja gerade das genaue Anbringen der Schnur. Um die ganze Sache zu vereinfachen, könnten wir uns doch ein anderes Maß für den Winkel zwischen $\vec{a}$ und $\vec{b}$ überlegen, bei dem nur *gerade* Strecken gemessen werden müssen. Eine Möglichkeit ist dabei, einen der beiden Vektoren, z. B. den Vektor $\vec{b}$, in eine Komponente $\vec{b}_{||} = \lambda(\vec{b})\vec{a}/||\vec{a}||$ parallel zu $\vec{a}$ und die dazu *senkrechte* Komponente $\vec{b}_{\perp} = \vec{b} - \vec{b}_{||}$ zerlegen. Die Funktion $\lambda(\vec{b})$ gibt

dabei den Anteil von $\vec{b}$ in Richtung $\vec{a}/||\vec{a}||$ an und ist somit die vorzeichenbehaftete Länge von $\vec{b}_{||}$.
Zur Winkelquantifizierung eignet sich dann die Länge von $\vec{b}_{||}$, oder, genauer gesagt, die vorzeichenbehaftete Länge $\lambda(\vec{b})$. Die Länge von $b_\perp$ kommt nicht in Frage, da sie den eingeschlossenen Winkel nicht eindeutig wiedergibt. Es kann nämlich vorkommen, daß $||\vec{c}_\perp|| = ||\vec{b}_\perp||$ aber $\sphericalangle(\vec{a},\vec{c}) \neq \sphericalangle(\vec{a},\vec{b})$



Betrachten wir deshalb $\lambda(\vec{b})$ als Rohmaterial zur Winkelquantifizierung. Den Wert $\lambda(\vec{b})$ müssen wir aber noch modifizieren, denn wenn $\vec{b}$ durch $x\vec{b}$ ersetzt wird mit $x > 0$, so gilt zunächst

$$x\vec{b} = x(\vec{b}_{||} + \vec{b}_\perp) = x\vec{b}_{||} + x\vec{b}_\perp$$

und da wegen Bedingung (i) und (ii) unserer Winkelfunktion

$$\sphericalangle(x\vec{b}_{||}, x\vec{b}_\perp) = \sphericalangle(\vec{b}_{||}, \vec{b}_\perp) = \frac{\pi}{2}$$

und $\sphericalangle(\vec{a}, x\vec{b}_{||}) = \sphericalangle(\vec{a}, \vec{b}_{||}) = 0$ gilt, haben wir $x\vec{b}_{||}$ als $(x\vec{b})_{||}$ erkannt. Damit ist aber auch $\lambda(x\vec{b}) = x\lambda(\vec{b})$. Die Länge der Parallelkomponente ändert sich also mit der Länge von $\vec{b}$, obwohl der Winkel derselbe bleibt (Bedingung (ii)). Da sich aber $\lambda(x\vec{b})$ und $x\vec{b}$ mit dem gleichen Faktor x ändern, kann man diese Längenabhängigkeit herauskürzen und

$$c(\vec{b}) = \frac{\lambda(\vec{b})}{||\vec{b}||}$$

als Winkelmaß betrachten. Wenn $\sphericalangle(\vec{a},\vec{b})$ zwischen 0 und π variiert, wird der Wert $c(\vec{b})$ zwischen $+1$ und -1 variieren, wobei zu jedem Winkel genau ein Wert c gehört und umgekehrt. Den funktionalen Zusammenhang zwischen Winkel und der normierten und vorzeichenbehafteten Projektionslänge $c(\vec{b})$ nennt man $\cos : [0,\pi] \to [-1,1]$, also

$$\cos \sphericalangle(\vec{a}, \vec{b}) = \frac{\lambda(\vec{b})}{||\vec{b}||}$$

Der Vorteil des Winkelmaßes $\cos \sphericalangle(\vec{a}, \vec{b})$ ist der, daß zu seiner Bestimmung die Länge von zwei geraden Strecken $\lambda(\vec{b})$ und $||\vec{b}||$ gemessen werden muß und nicht die eines Kreisbogens wie bei $\sphericalangle(\vec{a}, \vec{b})$. Allerdings ist die Situation bei der Wahl dieses Winkelmaßes etwas verwickelt: Um die benötigte Länge $\lambda(\vec{b})$ zu ermitteln, muß man bereits einen Winkel benutzen, nämlich den für die Zerlegung des Vektors $\vec{b}$ erforderlichen rechten Winkel. Diese Komplikation kann man durch folgende Überlegung umgehen. Die Zerlegung $\vec{b} = \vec{b}_{||} + \vec{b}_{\perp}$ ist nämlich dadurch ausgezeichnet, daß die Komponente $\vec{b}_{\perp}$ so kurz wie möglich ist. Damit haben wir ein neues Winkelmeßverfahren: Sind $\vec{a}, \vec{b}$ zwei Zeiger, so betrachtet man alle Zerlegungen

$$\vec{b} = \mu\vec{a}/||\vec{a}|| + (\vec{b} - \mu\vec{a}/||\vec{a}||)$$

und wählt μ so, daß $\vec{b} - \mu\vec{a}/||\vec{a}||$ so kurz wie möglich ist. Der gefundene Wert ist dann gerade $\lambda(\vec{b})$ und durch Normierung mit $||\vec{b}||$ erhalten wir einen Wert, der den eingeschlossenen Winkel eindeutig charakterisiert, wobei im Meßprozeß nur gerade Strecken gemessen werden müssen.
Die normierte Länge des senkrechten Anteils

$$s(\vec{b}) = \frac{||\vec{b} - \lambda(\vec{b})\vec{a}/||\vec{a}|| \; ||}{||\vec{b}||}$$

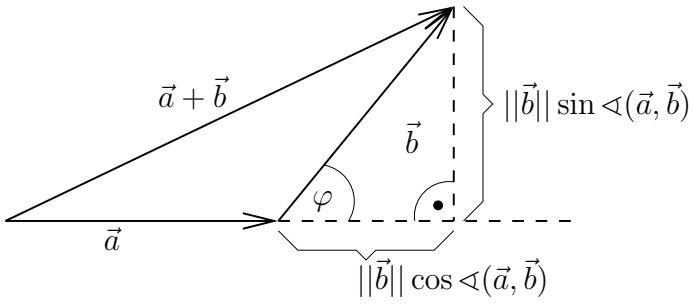
bezeichnet man übrigens als $\sin \sphericalangle(\vec{a}, \vec{b})$. Da nach dem Satz des Pythagoras

$$||\vec{b}||^2 = ||\vec{b}_{||} \; ||^2 + ||\vec{b}_{\perp}||^2$$

gilt, erhalten wir den Zusammenhang

$$1 = \frac{||\vec{b}_{||}||^2}{||\vec{b}||^2} + \frac{||\vec{b}_{\perp}||^2}{||\vec{b}||^2} = (\cos \sphericalangle(\vec{a}, \vec{b}))^2 + (\sin \sphericalangle(\vec{a}, \vec{b}))^2$$

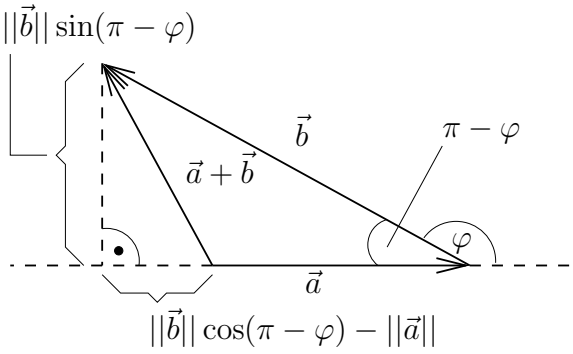
Unabhängig davon, wie wir den Winkel charakterisieren, ob durch $\sphericalangle(\vec{a}, \vec{b})$ oder $\cos \sphericalangle(\vec{a}, \vec{b})$, in jedem Fall beruht die Messung auf einer Längenmessung, was die enge Verzahnung der elementargeometrischen Länge mit Winkeln verdeutlicht. Dies wird auch deutlich, wenn wir ein allgemeines, von den Zeigern $\vec{a}, \vec{b}$ und $\vec{a} + \vec{b}$ aufgespanntes Dreieck, betrachten.



Im Fall $\varphi = \sphericalangle(\vec{a},\vec{b}) \leq \frac{\pi}{2}$ folgt mit dem Satz von Pythagoras

$$\begin{aligned} ||\vec{a}+\vec{b}||^2 &= (||\vec{a}|| + ||\vec{b}||\cos\varphi)^2 + ||\vec{b}||^2(\sin\varphi)^2 \\ &= ||\vec{a}||^2 + ||\vec{b}||^2 + 2||\vec{a}||\ ||\vec{b}||\cos\varphi \end{aligned}$$

wobei wir die Relation $\cos^2\varphi + \sin^2\varphi = 1$ ausgenutzt haben. Im Fall $\varphi = \sphericalangle(\vec{a},\vec{b}) > \frac{\pi}{2}$ ist die Länge der Grundseite durch $\big|\ ||\vec{a}|| - ||\vec{b}||\cos(\pi - \varphi)\big|$ gegeben, wie die folgende Skizze verdeutlicht



Deshalb gilt

$$\begin{aligned} ||\vec{a}+\vec{b}||^2 &= (||\vec{a}|| - ||\vec{b}||\cos(\pi-\varphi))^2 + ||\vec{b}||^2\sin^2(\pi-\varphi) \\ &= ||\vec{a}||^2 + ||\vec{b}||^2 - 2||\vec{a}||\ ||\vec{b}||\cos(\pi-\varphi) \end{aligned}$$

Wegen Eigenschaft (iii) der Winkelfunktion gilt mit $\varphi = \sphericalangle(\vec{a},\vec{b})$, daß $\pi - \varphi = \sphericalangle(\vec{a},-\vec{b})$ und

$$\cos\sphericalangle(\vec{a},-\vec{b}) = \frac{\lambda(-\vec{b})}{||\vec{b}||} = -\frac{\lambda(\vec{b})}{||\vec{b}||} = -\cos\sphericalangle(\vec{a},\vec{b})$$

Damit gilt sowohl im Fall $\varphi \leq \frac{\pi}{2}$ als auch für $\varphi > \frac{\pi}{2}$ der Zusammenhang

$$||\vec{a}+\vec{b}||^2 = ||\vec{a}||^2 + ||\vec{b}||^2 + 2||\vec{a}||\,||\vec{b}||\cos\sphericalangle(\vec{a},\vec{b}) \tag{6}$$

Die Seitenlängen im Dreieck sind also durch den Winkel $\sphericalangle(\vec{a},\vec{b})$ miteinander verknüpft. Für den winkelabhängigen Term in (6) führen wir folgende abkürzende Schreibweise ein.

Definition 2. *Die durch*

$$\langle \vec{a},\vec{b}\rangle = \begin{cases} ||\vec{a}||\,||\vec{b}||\cos\sphericalangle(\vec{a},\vec{b}) & \vec{a},\vec{b}\neq\vec{0} \\ 0 & \textit{sonst} \end{cases}$$

auf $S\times S$ gegebene Funktion heißt Skalarprodukt auf S.

Mit dieser Notation läßt sich (6) kurz als

$$||\vec{a}+\vec{b}||^2 = ||\vec{a}||^2 + ||\vec{b}||^2 + 2\langle\vec{a},\vec{b}\rangle \tag{7}$$

schreiben und wegen

$$\begin{aligned} \langle\vec{a},-\vec{b}\rangle &= ||\vec{a}||\,||-\vec{b}||\cos\sphericalangle(\vec{a},-\vec{b}) \\ &= ||\vec{a}||\,||\vec{b}||(-\cos\sphericalangle(\vec{a},\vec{b})) = -\langle\vec{a},\vec{b}\rangle \end{aligned} \tag{8}$$

für den nichttrivialen Fall $\vec{a},\vec{b}\neq\vec{0}$ (ist ein Zeiger der Nullzeiger, so ist die Beziehung trivial erfüllt), folgt ebenso

$$||\vec{a}-\vec{b}||^2 = ||\vec{a}||^2 + ||\vec{b}||^2 - 2\langle\vec{a},\vec{b}\rangle \tag{9}$$

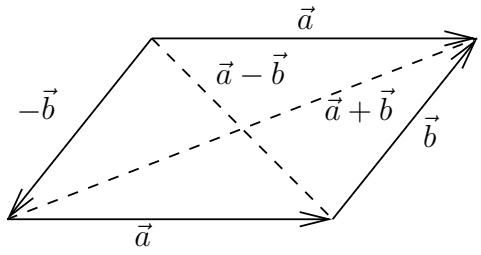
Subtraktion von (7) und (9) liefert

$$\langle\vec{a},\vec{b}\rangle = \frac{1}{4}(||\vec{a}+\vec{b}||^2 - ||\vec{a}-\vec{b}||^2), \tag{10}$$

also erneut einen Zusammenhang, der es erlaubt, $\cos\sphericalangle(\vec{a},\vec{b})$ nur durch Messung von vier Zeigern $\vec{a},\vec{b},\vec{a}+\vec{b}$ und $\vec{a}-\vec{b}$ zu ermitteln. Durch Addition von (7) und (9) fällt das Skalarprodukt dagegen heraus und wir finden das sogenannte *Parallelogramm Gesetz*

$$||\vec{a}+\vec{b}||^2 + ||\vec{a}-\vec{b}||^2 = 2||\vec{a}||^2 + 2||\vec{b}||^2 \tag{11}$$

was die Längenquadrate von den Seiten und den Diagonalen eines Parallelogramms in Relation setzt



Mit den fundamentalen Beziehungen (10) und (11) läßt sich eine wichtige Eigenschaft des Skalarprodukts zeigen, die *Additivität* $\langle \vec{a}+\vec{b},\vec{c}\rangle = \langle \vec{a},\vec{c}\rangle + \langle \vec{b},\vec{c}\rangle$. Wir beginnen mit der rechten Seite, die wir, um Brüche zu vermeiden, mit 8 multiplizieren.
Unter Ausnutzung von (10) ergibt sich zunächst

$$8\langle \vec{a},\vec{c}\rangle + 8\langle \vec{b},\vec{c}\rangle = 2(||\vec{a}+\vec{c}||^2 - ||\vec{a}-\vec{c}||^2) + 2(||\vec{b}+\vec{c}||^2 - ||\vec{b}-\vec{c}||^2)$$

Zu der ersten Klammer addieren wir nun Null in der Form $0 = 2||\vec{b}||^2 - 2||\vec{b}||^2$ und zur zweiten als $0 = 2||\vec{a}||^2 - 2||\vec{a}||^2$. Viermaliges Benutzen des Parallelogramm-Gesetzes

$$\begin{aligned}
2||\vec{a}+\vec{c}||^2 + 2||\vec{b}||^2 &= ||\vec{a}+\vec{c}+\vec{b}||^2 + ||\vec{a}+\vec{c}-\vec{b}||^2 \\
2||\vec{b}||^2 + 2||\vec{a}-\vec{c}||^2 &= ||\vec{b}+\vec{a}-\vec{c}||^2 + ||\vec{b}-\vec{a}+\vec{c}||^2 \\
2||\vec{b}+\vec{c}||^2 + 2||\vec{a}||^2 &= ||\vec{b}+\vec{c}+\vec{a}||^2 + ||\vec{b}+\vec{c}-\vec{a}||^2 \\
2||\vec{a}||^2 + 2||\vec{b}-\vec{c}||^2 &= ||\vec{a}+\vec{b}-\vec{c}||^2 + ||\vec{a}-\vec{b}+\vec{c}||^2
\end{aligned}$$

ergibt dann schließlich unter Berücksichtigung der Vorzeichen

$$\begin{aligned}
8\langle \vec{a},\vec{c}\rangle + 8\langle \vec{b},\vec{c}\rangle &= 2||\vec{a}+\vec{b}+\vec{c}||^2 - 2||\vec{a}+\vec{b}-\vec{c}||^2 \\
&= 8\langle \vec{a}+\vec{b},\vec{c}\rangle
\end{aligned}$$

Die Additivität im zweiten Argument des Skalarprodukts kann man mit einer ähnlichen Rechnung nachweisen. Es ist allerdings einfacher, die *Symmetrie* $\langle \vec{a},\vec{b}\rangle = \langle \vec{b},\vec{a}\rangle$ auszunutzen, die sofort aus der Definition des Skalarprodukts folgt. Wir haben dann

$$\langle \vec{a},\vec{b}+\vec{c}\rangle = \langle \vec{b}+\vec{c},\vec{a}\rangle = \langle \vec{b},\vec{a}\rangle + \langle \vec{c},\vec{a}\rangle = \langle \vec{a},\vec{b}\rangle + \langle \vec{a},\vec{c}\rangle$$

Das Zusammenspiel des Skalarprodukts mit der skalaren Multiplikation ergibt sich ebenfalls direkt aus der Definition. Für einen Faktor $x > 0$

gilt

$$\begin{aligned}\langle x\vec{a},\vec{b}\rangle &= ||x\vec{a}||\,||\vec{b}||\cos\sphericalangle(x\vec{a},\vec{b})\\ &= x||\vec{a}||\,||\vec{b}||\cos\sphericalangle(\vec{a},\vec{b}) = x\langle\vec{a},\vec{b}\rangle\end{aligned}$$

(die entsprechende Relation für den Fall $\vec{a}=\vec{0}$ oder $\vec{b}=\vec{0}$ ist wieder trivial erfüllt). Ist der Faktor $x<0$, so ist zu beachten, daß $x=-|x|$ gilt und

$$\cos\sphericalangle(x\vec{a},\vec{b}) = \cos(\pi-\sphericalangle(\vec{a},\vec{b})) = -\cos\sphericalangle(\vec{a},\vec{b})$$

so daß

$$\langle x\vec{a},\vec{b}\rangle = |x|\,||\vec{a}||\,||\vec{b}||(-\cos\sphericalangle(\vec{a},\vec{b})) = x\langle\vec{a},\vec{b}\rangle.$$

Wie oben, zeigt man unter Ausnutzung der Symmetrie, daß skalare Faktoren auch aus dem zweiten Argument nach vorne gezogen werden können. Zusammen mit der Additivität ist das Skalarprodukt also linear sowohl im ersten als auch im zweiten Argument und damit eine sogenannte *bilineare* Funktion auf S.
Als letzte Eigenschaft erwähnen wir noch die Beziehung

$$\langle\vec{a},\vec{a}\rangle = ||\vec{a}||^2\cos 0 = ||\vec{a}||^2$$

aus der sowohl $\langle\vec{a},\vec{a}\rangle\geq 0$ für alle $\vec{a}\in S$ folgt, als auch die Tatsache, daß $\langle\vec{a},\vec{a}\rangle$ nur im Fall $\vec{a}=\vec{0}$ Null sein kann.
Beim Arbeiten mit der elementargeometrischen Länge stellt sich heraus, daß die gefundenen Eigenschaften immer wieder auftreten. Außerdem sind sie unabhängig von speziellen Eigenschaften des Zeigervektorraums definiert und lassen sich somit auf andere Vektorräume übertragen. Dieser Schritt hat enorme Vorteile, da wir aus dem Zeigerraum motivierte geometrische Zusammenhänge auch auf andere Räume übertragen können und so unsere Anschauung z. B. in abstrakten Funktionenräumen einsetzen können. Als Anwendung werden wir später die Methode der kleinsten Quadrate in der Approximationstheorie oder die Fourieranalyse in der Signalverarbeitung kennenlernen. Wir verallgemeinern dazu die Situation im Zeigervektorraum mit folgender

Definition 3. *Ist V ein reeller Vektorraum und $\langle\cdot,\cdot\rangle : V\times V\to\mathbb{R}$ eine Funktion, die für alle $\vec{u},\vec{v},\vec{w}\in V$ und $\alpha\in\mathbb{R}$ folgende Bedingungen erfüllt*

(i) $\langle\vec{u},\vec{v}\rangle = \langle\vec{v},\vec{u}\rangle$
(ii) $\langle\vec{u}+\vec{v},\vec{w}\rangle = \langle\vec{u},\vec{w}\rangle + \langle\vec{v},\vec{w}\rangle$
(iii) $\langle\alpha\vec{u},\vec{v}\rangle = \alpha\langle\vec{u},\vec{v}\rangle$
(iv) $\langle\vec{v},\vec{v}\rangle\geq 0$
(v) $\langle\vec{v},\vec{v}\rangle = 0$ *nur für* $\vec{v}=\vec{0}$

so heißt $\langle \cdot, \cdot \rangle$ inneres Produkt bzw. Skalarprodukt auf V. In Analogie zum Zeigerraum nennt man $\vec{u}, \vec{v} \in V$ senkrecht zueinander bzw. orthogonal bezüglich des Skalarprodukt $\langle \cdot, \cdot \rangle$, falls $\langle \vec{u}, \vec{v} \rangle = 0$ gilt (in Zeichen $\vec{u} \perp \vec{v}$).

Als erstes Beispiel für den Umgang mit allgemeinen Skalarprodukten wollen wir die orthogonale Zerlegung des Vektors $\vec{v}$ bezüglich eines anderes Vektors $\vec{u} \neq \vec{0}$ betrachten. Wir suchen also $\lambda \in \mathbb{R}$, so daß

$$\vec{v} = \lambda\vec{u} + (\vec{v} - \lambda\vec{u}) \qquad \text{wobei} \qquad \vec{v} - \lambda\vec{u} \perp \vec{u}$$

Die Orthogonalitätsbedingung liefert dabei die Bestimmungsgleichung für λ, denn

$$0 = \langle \vec{v} - \lambda\vec{u}, \vec{u} \rangle = \langle \vec{v}, \vec{u} \rangle - \lambda \langle \vec{u}, \vec{u} \rangle$$

liefert unter Ausnutzung der Eigenschaft (v) des Skalarprodukts

$$\lambda = \frac{\langle \vec{v}, \vec{u} \rangle}{\langle \vec{u}, \vec{u} \rangle}$$

und damit die Zerlegung

$$\vec{v} = \vec{v}_{||} + \vec{v}_{\perp}, \quad \vec{v}_{||} = \frac{\langle \vec{v}, \vec{u} \rangle}{\langle \vec{u}, \vec{u} \rangle}\vec{u}, \quad v_{\perp} \perp \vec{u}$$

Wegen der Orthogonalität von $\vec{v}_{||}$ und $\vec{v}_{\perp}$ gilt für den quadratischen Ausdruck

$$\begin{aligned} \langle \vec{v}, \vec{v} \rangle &= \langle \vec{v}_{||} + \vec{v}_{\perp}, \vec{v}_{||} + \vec{v}_{\perp} \rangle = \langle \vec{v}_{||}, \vec{v}_{||} + \vec{v}_{\perp} \rangle + \langle \vec{v}_{\perp}, \vec{v}_{||} + \vec{v}_{\perp} \rangle \\ &= \langle \vec{v}_{||}, \vec{v}_{||} \rangle + \langle \vec{v}_{\perp}, \vec{v}_{\perp} \rangle \end{aligned}$$

und mit Eigenschaft (iv) sowie $\vec{v}_{||} = \lambda\vec{u}$

$$\langle \vec{v}, \vec{v} \rangle \geq \langle \vec{v}_{||}, \vec{v}_{||} \rangle = \lambda^2 \langle \vec{u}, \vec{u} \rangle = \frac{\langle \vec{v}, \vec{u} \rangle^2}{\langle \vec{u}, \vec{u} \rangle}$$

Durch Multiplikation mit $\langle \vec{u}, \vec{u} \rangle$ und Ziehen der Wurzel ergibt sich schließlich die *Cauchy-Schwarzsche Ungleichung*

$$|\langle \vec{u}, \vec{v} \rangle| \leq \sqrt{\langle \vec{u}, \vec{u} \rangle \langle \vec{v}, \vec{v} \rangle} \qquad \vec{u}, \vec{v} \in V \tag{12}$$

Als unmittelbare Konsequenz dieser Ungleichung können wir zeigen, daß ein Skalarprodukt $\langle \cdot, \cdot \rangle$ immer eine Norm induziert.

Satz 3. *Sei V ein reeller Vektorraum mit Skalarprodukt $\langle \cdot, \cdot \rangle$. Dann ist $||\vec{v}|| = \sqrt{\langle \vec{v}, \vec{v} \rangle}$ eine Norm auf V.*

Zum Nachweis dieser Aussage müssen wir nur die drei Normeigenschaften nachweisen. Zunächst ist $||\vec{v}|| = 0$ genau dann, wenn $\langle \vec{v}, \vec{v} \rangle = 0$, was ja nur im Fall $\vec{v} = \vec{0}$ eintreten kann. Für das Zusammenspiel mit der skalaren Multiplikation benötigen wir Eigenschaft (*iii*) des Skalarprodukts

$$||\alpha\vec{v}|| = \sqrt{\langle \alpha\vec{v}, \alpha\vec{v} \rangle} = \sqrt{\alpha^2 \langle \vec{v}, \vec{v} \rangle} = |\alpha| \sqrt{\langle \vec{v}, \vec{v} \rangle} = |\alpha| \; ||\vec{v}||$$

Schließlich kommt beim Nachweis der Dreiecksungleichung die Cauchy-Schwarzsche Ungleichung zum Einsatz, die man mit $|| \cdot ||$ auch in der Form $|\langle \vec{u}, \vec{v} \rangle| \leq ||\vec{u}|| \; ||\vec{v}||$ schreiben kann.

$$\begin{aligned} ||\vec{u} + \vec{v}||^2 &= \langle \vec{u} + \vec{v}, \vec{u} + \vec{v} \rangle = \langle \vec{u}, \vec{u} \rangle + \langle \vec{u}, \vec{v} \rangle + \langle \vec{v}, \vec{u} \rangle + \langle \vec{v}, \vec{v} \rangle \\ &= ||\vec{u}||^2 + ||\vec{v}||^2 + 2\langle \vec{u}, \vec{v} \rangle \\ &\leq ||u||^2 + ||\vec{v}||^2 + 2||\vec{u}|| \; ||\vec{v}|| = (||\vec{u}|| + ||\vec{v}||)^2 \end{aligned}$$

Im Fall des Zeigervektorraums, wo wir das Skalarprodukt ja als $\langle \vec{a}, \vec{b} \rangle = ||\vec{a}|| \; ||\vec{b}|| \cos \sphericalangle(\vec{a}, \vec{b})$ definiert hatten, mit der elementargeometrischen Länge $|| \cdot ||$, gilt $\sqrt{\langle \vec{a}, \vec{a} \rangle} = ||\vec{a}||$, d. h. die mit dem Skalarprodukt definierte Länge ist gerade die elementargeometrische. Den Zusammenhang $\cos \sphericalangle(\vec{a}, \vec{b}) = \langle \vec{a}, \vec{b} \rangle / (||\vec{a}|| \; ||\vec{b}||)$ nimmt man nun in allgemeinen Vektorräumen als Grundlage für die Definition eines Winkels zwischen zwei Vektoren. Dabei wird die inverse Abbildung arccos zu $\cos : [0, \pi] \to [-1, 1]$ benutzt.

Definition 4. *Sei V ein reeller Vektorraum mit Skalarprodukt $\langle \cdot, \cdot \rangle$. Seien $\vec{u}, \vec{v} \neq \vec{0}$ in V. Dann heißt*

$$\sphericalangle(\vec{u}, \vec{v}) = \arccos \frac{\langle \vec{u}, \vec{v} \rangle}{||\vec{u}|| \; ||\vec{v}||} \in [0, \pi]$$

der Winkel zwischen $\vec{u}$ und $\vec{v}$ bezüglich $\langle \cdot, \cdot \rangle$.

Beachten Sie, daß die Cauchy-Schwarzsche Ungleichung erneut benötigt wird, um zu zeigen, daß der Quotient $\langle \vec{u}, \vec{v} \rangle / (||\vec{u}|| \; ||\vec{v}||)$ tatsächlich in der Definitionsmenge $[-1, 1]$ der arccos-Funktion liegt.
Nach diesen allgemeinen Betrachtungen wollen wir uns nun der Frage zuwenden, wie man mit Skalarprodukten konkret arbeitet, genauer gesagt, wie man ein Skalarprodukt bzw. die zugehörige Norm ausrechnet. Sei dazu $(\vec{v}_1, \ldots, \vec{v}_n)$ die Basis eines Vektorraums V mit Skalarprodukt. Zur Berechnung des Skalarprodukts $\langle \vec{a}, \vec{b} \rangle$ stellen wir die beteiligten Vektoren zunächst in der Basis dar

$$\vec{a} = \alpha_1 \vec{v}_1 + \ldots + \alpha_n \vec{v}_n, \qquad \vec{b} = \beta_1 \vec{v}_1 + \ldots + \beta_n \vec{v}_n.$$

Nutzen wir nun die Linearität des Skalarprodukts in beiden Komponenten, so folgt

$$\langle \vec{a}, \vec{b} \rangle = \sum_{i=1}^{n} \alpha_i \langle \vec{v}_i, \sum_{j=1}^{n} \beta_j v_j \rangle = \sum_{i=1}^{n} \alpha_i \sum_{j=1}^{n} \langle \vec{v}_i, \vec{v}_j \rangle \beta_j$$

Neben den Koordinatenvektoren von $\vec{a}$ und $\vec{b}$ benötigt man zur Berechnung offensichtlich die $n \cdot (n+1)/2$ Skalarproduktkombinationen $\langle \vec{v}_i, \vec{v}_j \rangle$ zwischen den Basisvektoren. Zur kompakteren Schreibweise fassen wir diese Zahlen in einer Matrix $M \in \mathbb{R}^{n \times n}$ zusammen, d. h. $M_{ij} = \langle \vec{v}_i, \vec{v}_j \rangle$. Dann ist aber

$$\sum_{j=1}^{n} \langle \vec{v}_i, \vec{v}_j \rangle \beta_j = \sum_{j=1}^{n} M_{ij} \beta_j = (M\vec{\beta})_i$$

d. h. zur Berechnung von $\langle \vec{a}, \vec{b} \rangle$ bildet man zunächst das Matrixvektorprodukt $M\vec{\beta}$ zwischen der Matrix M und dem Koordinatenvektor von $\vec{b}$ und multipliziert den resultierenden Vektor anschließend Komponente für Komponente mit den Koordinaten von $\vec{a}$ und summiert über alle Werte. Zur Beschreibung des zweiten Teils dieser Prozedur definieren wir ein spezielles Skalarprodukt auf dem Koordinatenvektorraum.

Definition 5. *Sei $n \in \mathbb{N}$. Dann heißt*

$$\langle \vec{x}, \vec{y} \rangle = \sum_{i=1}^{n} x_i y_i \qquad \vec{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \vec{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Standardskalarprodukt auf $\mathbb{R}^{n \times 1}$.

Die Skalarprodukteigenschaften rechnet man sehr leicht nach. Beachten Sie, daß wir das gleiche Symbol $\langle \cdot, \cdot \rangle$ benutzen wie im allgemeinen Vektorraum V und im Zeigervektorraum S. Das kann nicht zur Verwechslung führen, da man an den eingesetzten Vektoren immer erkennen kann, welches Skalarprodukt gemeint ist. Nur wenn auf dem gleichen Raum mehrere Skalarprodukte benutzt werden, müssen wir uns verschiedene Schreibweisen überlegen.
Mit dem Standardskalarprodukt können wir nun die Berechnung eines allgemeinen Skalarprodukts auf V knapp formulieren

$$\langle \vec{a}, \vec{b} \rangle = \langle \vec{\alpha}, M\vec{\beta} \rangle$$

Beachten Sie, daß links das V-Skalarprodukt steht und rechts das Standardskalarprodukt auf $\mathbb{R}^{n \times 1}$. Besonders einfach wird diese Berechnung offensichtlich, wenn die Matrix M die Einheitsmatrix ist. Das ist dann

der Fall, wenn $M_{ij} = \langle \vec{v}_i, \vec{v}_j \rangle = 0$ ist für $i \neq j$, d. h. wenn die Basisvektoren paarweise senkrecht zueinander sind. Außerdem bedeutet $\langle \vec{v}_i, \vec{v}_i \rangle = 1$, daß alle Basisvektoren die Länge Eins haben. Eine solche Basis nennt man Orthonormalbasis.

Definition 6. *Sei V ein n-dimensionaler Vektorrraum mit Skalarprodukt $\langle \cdot, \cdot \rangle$. Die Vektoren $(\vec{v}_1, \ldots, \vec{v}_n)$ heißen Orthonormalbasis von V bezüglich $\langle \cdot, \cdot \rangle$, falls*

$$\langle \vec{v}_i, \vec{v}_j \rangle = \delta_{ij} = \begin{cases} 1 & i = j \\ \\ 0 & i \neq j \end{cases} \qquad i, j = 1, \ldots, n$$

Das Symbol δ_{ij} heißt auch *Kronecker Delta*. Es gibt gerade die Komponenten der $n \times n$ Einheitsmatrix an. Haben Sie gemerkt, daß die Definition eine versteckte Aussage enthält? Der Begriff Orthonormalbasis suggeriert natürlich, daß es sich dabei um eine Basis handelt, allerdings wurde dies nicht vorausgesetzt. Man kann es aber bereits aus den Othonormalitätsbedingungen ableiten. Nehmen wir dazu an, wir hätten eine beliebige Darstellung des Nullvektors durch die Vektoren $\vec{v}_1, \ldots, \vec{v}_n$

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

Multiplizieren wir diese Gleichung mit $\vec{v}_i$ und beachten

$$\langle \vec{v}_i, \vec{0} \rangle = \langle \vec{v}_i, 0 \cdot \vec{0} \rangle = 0 \langle \vec{v}_i, \vec{0} \rangle = 0$$

so erhalten wir

$$0 = \langle \vec{v}_i, \vec{0} \rangle = \langle \vec{v}_i, \lambda_1 \vec{v}_1 + \ldots + \lambda_n \vec{v}_n \rangle = \lambda_1 \langle \vec{v}_i, \vec{v}_1 \rangle + \ldots + \lambda_n \langle \vec{v}_i, \vec{v}_n \rangle$$

Da alle Skalarprodukte außer $\langle \vec{v}_i, \vec{v}_i \rangle = 1$ verschwinden, folgt somit $0 = \lambda_i$ und damit sind die Vektoren $\vec{v}_1, \ldots, \vec{v}_n$ ein maximales, linear unabhängiges System von Vektoren, d. h. eine Basis von V. Die Koordinaten c_i eines Vektors

$$\vec{u} = c_1 \vec{v}_1 + \ldots + c_n \vec{v}_n$$

lassen sich übrigens sehr einfach bestimmen. Multipliziert man $\vec{u}$ mit $\vec{v}_i$, so fischt man genau den Faktor c_i vor $\vec{v}_i$ heraus, da alle anderen Produkte $\langle \vec{v}_i, \vec{v}_j \rangle$ mit $i \neq j$ verschwinden. Es gilt also $c_i = \langle \vec{u}, \vec{v}_i \rangle$ bzw.

$$\vec{u} = \langle \vec{u}, \vec{v}_1 \rangle \vec{v}_1 + \langle \vec{u}, \vec{v}_2 \rangle \vec{v}_2 + \ldots + \langle \vec{u}, \vec{v}_n \rangle \vec{v}_n \quad \vec{u} \in V$$

Außerdem ist die Berechnung des Skalarprodukts zweier Vektoren $\vec{a}, \vec{b} \in V$ mit Koordinatenvektoren $\vec{\alpha}$ und $\vec{\beta}$ bezüglich einer Orthonormalbasis sehr einfach. Wie wir gesehen haben, gilt

$$\langle \vec{a}, \vec{b} \rangle = \langle \vec{\alpha}, \vec{\beta} \rangle = \alpha_1 \beta_1 + \ldots + \alpha_n \beta_n$$

und damit

$$||\vec{a}|| = ||\vec{\alpha}|| = \sqrt{\alpha_1^2 + \alpha_2^2 + \ldots + \alpha_n^2}$$

wobei die zweite Norm die des Standardskalarprodukts in $\mathbb{R}^{n\times 1}$ ist.
Im Fall des Zeigervektorraums können wir damit Winkel einfach ausrechnen. Welchen Winkel bildet z. B. der Vektor $\vec{a} = \vec{s}_1 + \vec{s}_2 + \vec{s}_3$ in Richtung der Raumdiagonalen des ersten Oktanden mit den Koordinatenvektoren $\vec{s}_i$? Dazu finden wir zunächst die Koordinaten $\vec{\alpha}$ von $\vec{a}$ und $\vec{\beta} = \vec{s}_1$ bezüglich der Orthonormalbasis $(\vec{s}_1, \vec{s}_2, \vec{s}_3)$

$$\vec{\alpha} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \vec{\beta} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

Offensichtlich gilt $||\vec{\alpha}|| = \sqrt{3}$ und $||\vec{\beta}|| = 1$ und nach der Winkelformel folgt

$$\cos \sphericalangle(\vec{a}, \vec{b}) = \frac{\langle \vec{a}, \vec{b} \rangle}{||\vec{a}||\, ||\vec{b}||} = \frac{1}{\sqrt{3}}(1 \cdot 1 + 1 \cdot 0 + 1 \cdot 0) = \frac{1}{\sqrt{3}}$$

Anwendung der Umkehrfunktion arccos liefert ungefähr $\sphericalangle(\vec{a}, \vec{b}) \approx 0,96$, was einer Gradzahl von etwa $54,7°$ entspricht. Die Entfernung zwischen zwei Raumpunkten berechnet man mit der Länge der Differenz der Ortsvektoren. So beträgt der Abstand zwischen den Punkten auf die $\vec{a}$ und $\vec{b}$ zeigen, wenn die stumpfen Enden in einem Referenzpunkt liegen

$$||\vec{a} - \vec{b}|| = \sqrt{(\alpha_1 - \beta_1)^2 + (\alpha_2 - \beta_2)^2 + (\alpha_3 - \beta_3)^2} = \sqrt{2}$$

Alle Punkte, die vom Ursprung des Koordinatensystems höchstens den Abstand Eins haben, bilden die Einheitskugel der Norm

$$E = \{\vec{v} \in S| \;\; ||\vec{v}|| \leq 1\}$$

bzw. in Koordinaten

$$E = \{v_1\vec{s}_1 + v_2\vec{s}_2 + v_3\vec{s}_3 \mid v_1^2 + v_2^2 + v_3^2 \leq 1\}$$

Geometrisch beschreibt dieses Objekt eine Kugel vom Radius Eins, so daß hier der Name Einheitskugel voll und ganz gerechtfertigt ist.
Während man im Zeigervektorraum im allgemeinen immer eine Orthonormalbasis zugrunde legt, können in anderen Vektorräumen die natürlichen Basen durchaus nicht orthonormal sein. Beispielsweise ist die Monombasis im Raum $\mathcal{P}_n$ nicht orthonormal bezüglich dem Skalarprodukt

$$\langle P, Q \rangle = \int_{-1}^{1} P(x)Q(x)dx$$

Ein anderes Beispiel sind Ebenen im Zeigervektorraum, bei denen die Richtungsvektoren auf, für die konkrete Anwendung wichtigen, Richtungen basieren, die nicht orthogonal sind. Beispielsweise kann eine Ebene durch drei Punkte bei einer Landvermessung charakterisiert werden, wobei die Meßpunkte relativ willkürlich im Raum liegen. Um dennoch in den Genuß einer Orthonormalbasis zu kommen, muß man in solchen Fällen zunächst Vektoren $\vec{v}_i$ mit der Eigenschaft $\langle \vec{v}_i, \vec{v}_j \rangle = \delta_{ij}$ konstruieren. Die Idee des Schmidtschen Orthonormalisierungsverfahrens ist, eine gegebene Basis $(\vec{w}_1, \ldots, \vec{w}_1)$ so zurechtzustutzen, daß sie orthonormal ist. Der wesentliche Punkt ist dabei die Herstellung der Orthogonalität. Das Normieren der orthogonalen Vektoren ist ja sehr einfach. Dividiert man einen gegebenen Vektor durch seine Norm, so hat der resultierende Vektor Länge Eins, ist also normiert. Um dagegen einen Vektor bezüglich einer Gruppe anderer Vektoren zu orthogonalisieren, muß man sich etwas mehr anstrengen, aber schauen wir uns den Prozeß Schritt für Schritt an. Zunächst nimmt man den ersten Vektor $\vec{w}_1$ der Basis und normiert ihn

$$\vec{v}_1 = \frac{\vec{w}_1}{||\vec{w}_1||}$$

Der zweite Vektor $\vec{w}_2$ der Basis wird typischerweise nicht senkrecht zu $\vec{v}_1$ sein, d. h. er wird etwas in $\vec{v}_1$ Richtung (oder entgegengesetzt) gekippt sein. Diesen Anteil in $\vec{v}_1$-Richtung muß man nun subtrahieren, um einen zu $\vec{v}_1$ senkrechten Vektor zu erhalten. Die Berechnung der orthogonalen Zerlegung eines Vektors hatten wir ja schon einmal durchgeführt. Es ist

$$\vec{w}_2 = \langle \vec{w}_2, \vec{v}_1 \rangle \vec{v}_1 + (\vec{w}_2 - \langle \vec{w}_2, \vec{v}_1 \rangle \vec{v}_1)$$

wobei wir $\langle \vec{v}_1, \vec{v}_1 \rangle = 1$ benutzt haben. Der senkrechte Anteil ist also

$$\vec{u}_2 = \vec{w}_2 - \langle \vec{w}_2, \vec{v}_1 \rangle \vec{v}_1$$

und Normierung liefert einen Vektor

$$\vec{v}_2 = \frac{\vec{u}_2}{||\vec{u}_2||}$$

der Länge Eins hat und zu $\vec{v}_1$ orthogonal ist. Beachten Sie, daß beliebige Linearkombinationen der Vektoren $\vec{v}_1$ und $\vec{v}_2$ denselben Unterraum erzeugen wie die Linearkombinationen von $\vec{w}_1$ und $\vec{w}_2$. Wir haben ja nur die Richtung von $\vec{w}_2$ geändert, indem wir den $\vec{w}_1$ Anteil herausgenommen haben. Das stört aber bei Linearkombinationen nicht, da die Richtung $\vec{w}_1$ ja immer noch durch $\vec{v}_1$ zur Verfügung steht und damit wieder dazukombiniert werden kann. Bei der Herstellung des dritten Vektors

der Orthonormalbasis verfahren wir prinzipiell wie bei der Konstruktion von $\vec{v}_2$. Wir nehmen den Vektor $\vec{w}_3$ und subtrahieren zunächst den Anteil $\langle \vec{w}_3, \vec{v}_1 \rangle \vec{v}_1$ in $\vec{v}_1$-Richtung und dann den Anteil $\langle \vec{w}_3, \vec{v}_2 \rangle \vec{v}_2$ in $\vec{v}_2$-Richtung. Somit erhalten wir einen Vektor $\vec{u}_3$, der noch normiert werden muß

$$\vec{u}_3 = \vec{w}_3 - \langle \vec{w}_3, \vec{v}_1 \rangle \vec{v}_1 - \langle \vec{w}_3, \vec{v}_2 \rangle \vec{v}_2, \qquad \vec{v}_3 = \frac{\vec{u}_3}{||\vec{u}_3||}$$

Man rechnet leicht nach, daß nun $\langle \vec{v}_3, \vec{v}_1 \rangle = \langle \vec{v}_3, \vec{v}_2 \rangle = 0$ gilt. Außerdem liefern erneut Linearkombinationen von $\vec{v}_1, \vec{v}_2, \vec{v}_3$ den gleichen Teilraum wie Linearkombinationen von $\vec{w}_1, \vec{w}_2, \vec{w}_3$, da die von $\vec{w}_3$ abgezogenen Anteile ja durch $\vec{v}_1$ und $\vec{v}_2$ noch zur Verfügung stehen. Den allgemeinen Schritt der Konstruktion von $\vec{v}_k$ bei bereits konstruierten Vektoren $\vec{v}_1, \ldots, \vec{v}_{k-1}$ können Sie nun sicherlich erraten

$$\vec{u}_k = \vec{w}_k - \sum_{i=1}^{k-1} \langle \vec{w}_k, \vec{v}_i \rangle \vec{v}_i, \qquad \vec{v}_k = \frac{\vec{u}_k}{||\vec{u}_k||} \tag{13}$$

wobei die Orthogonalitätsbedingungen sehr einfach nachzurechnen sind. Hoffentlich haben Sie sich während der Konstruktion gefragt, ob denn die Norm des Hilfsvektors $\vec{u}_k$ immer von Null verschieden ist, da wir doch durch diese Norm dividieren. Beim Dividieren muß man nämlich höllisch aufpassen, daß man nicht durch Null teilt, da diese Operation nicht definiert ist, d. h. niemand kann Ihnen sagen, was das Ergebnis sein sollte. Denken Sie also *immer* darüber nach, ob das, was Sie gerne in den Nenner schreiben wollen, Null ist. Im vorliegenden Fall kann das nicht passieren. Stellen Sie sich vor, $\vec{u}_k$ habe Länge Null, d. h. $\vec{u}_k$ sei der Nullvektor. Dann steht aber in (13), daß $\vec{w}_k$ als Linearkombination von $\vec{v}_1, \ldots, \vec{v}_{k-1}$ geschrieben werden kann und nach unseren obigen Überlegungen auch als Linearkombination der $\vec{w}_1, \ldots, \vec{w}_{k-1}$. Dies widerspricht aber der Annahme, daß $(\vec{w}_1, \ldots, \vec{w}_n)$ eine Basis ist und damit aus linear unabhängigen Vektoren besteht. Der Vektor $\vec{w}_n$ zeigt also im übertragenen Sinne in eine Richtung, die von allen Richtungen, die durch $\vec{v}_1, \ldots, \vec{v}_{n-1}$ erzeugt werden können, verschieden ist. Beim Entfernen der $\vec{v}_i$-Komponenten bleibt also auf jeden Fall etwas übrig, und das ist gerade die senkrechte Komponente $\vec{u}_k$.
Schauen wir uns jetzt einmal ein konkretes Beispiel im Zeigervektorraum S an. Sei

$$E = \{ s\vec{a} + t\vec{b} | (s,t) \in \mathbb{R}^2 \}$$

eine Ebene, wobei die Koordinaten von $\vec{a}$ und $\vec{b}$ bezüglich einer Orthonormalbasis $(\vec{s}_1, \vec{s}_2, \vec{s}_2)$ von S

$$\vec{\alpha} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \quad , \qquad \vec{\beta} = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}$$

gegeben sind. Um eine Orthonormalbasis von E zu konstruieren, benutzen wir das Schmidtsche Orthonormalisierungsverfahren, beruhend auf der Basis $(\vec{a}, \vec{b})$ von E. Zunächst normieren wir $\vec{u}_1 = \vec{a}$.

$$||\vec{u}_1|| = \sqrt{1^2 + 1^2 + 1^2} = \sqrt{3} \ , \qquad \vec{v}_1 = \frac{1}{\sqrt{3}} \, \vec{a}$$

Danach ziehen wir die $\vec{v}_1$-Komponente von $\vec{b}$ ab, wozu wir das Skalarprodukt $\langle \vec{b}, \vec{v}_1 \rangle$ benötigen

$$\langle \vec{b}, \vec{v}_1 \rangle = \frac{1}{\sqrt{3}} \langle \vec{b}, \vec{a} \rangle = \frac{1}{\sqrt{3}} \langle \vec{\beta}, \vec{\alpha} \rangle = \frac{1}{\sqrt{3}} (2 \cdot 1 + 1 \cdot 1 + 2 \cdot 1) = \frac{5}{\sqrt{3}}$$

Damit ergibt sich

$$\vec{u}_2 = \vec{b} - \langle \vec{b}, \vec{v}_1 \rangle \vec{v}_1 = \vec{b} - \frac{5}{3} \vec{a}$$

bzw. für die Koordinaten $\vec{\mu}_2$ von $\vec{u}_2$

$$\vec{\mu}_2 = \vec{\beta} - \frac{5}{3} \vec{\alpha} = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix} - \frac{5}{3} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \frac{1}{3} \begin{pmatrix} 1 \\ -2 \\ 2 \end{pmatrix}$$

Zur Normierung brauchen wir noch die Länge von $\vec{u}_2$

$$||\vec{u}_2|| = ||\vec{\mu}_2|| = \frac{1}{3} \sqrt{1^2 + (-2)^2 + 1^2} = \frac{\sqrt{6}}{3}$$

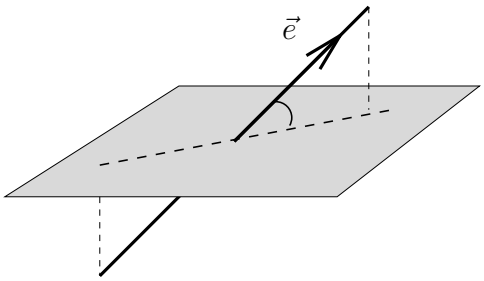
so daß schließlich $\vec{v}_2 = \frac{3}{\sqrt{6}} \, \vec{u}_2$. Die Koordinaten der Orthonormalbasisvektoren $\vec{v}_1$ und $\vec{v}_2$ sind also

$$\vec{\gamma}_1 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \ , \qquad \vec{\gamma}_2 = \frac{1}{\sqrt{6}} \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} .$$

Mit der so gewonnenen Orthonormalbasis kann man nun leicht interessante Fragen beantworten. Wie weit ist ein gegebener Punkt mit Ortsvektor $\vec{c}$ von der Ebene entfernt? Unter welchem Winkel schneidet eine Gerade die Ebene? Welcher Ebenenvektor liegt am dichtesten an einem gegebenen Vektor?

Beginnen wir mit der Frage nach dem Schnittwinkel. Zunächst ist dabei zu klären, wie man diesen Winkel überhaupt definiert. Ist $\vec{e}$ der

Richtungsvektor der Geraden und $\vec{v}$ ein beliebiger Vektor der Ebene, so kann man mit Hilfe des Skalarprodukts $\langle \vec{e}, \vec{v} \rangle$ den Winkel $\sphericalangle(\vec{e}, \vec{v})$ bestimmen. Allerdings wird je nach Wahl von $\vec{v}$ ein anderer Wert resultieren. Welchen soll man denn dann als Winkel zwischen Gerade und Ebene nehmen? Ausgezeichnet unter allen möglichen Winkeln sind offensichtlich der kleinste und der größte auftretende Winkel und der kleinste ist wohl der, den man intuitiv als Schnittwinkel bezeichnen würde



Der zugehörige Vektor $\vec{v}$ der Ebene, zu dem der kleinste Winkel auftritt, hat anschaulich die Eigenschaft, daß er genau unterhalb von $\vec{e}$ liegt; er bildet sozusagen den Schatten von $\vec{e}$, wenn die Ebene senkrecht von oben mit Licht bestrahlt wird. Mathematisch gesehen, ist $\vec{v}$ also durch die *Projektion* von $\vec{e}$ in die Ebene gegeben. Die Berechnung dieser Projektion führt auf folgende Aufgabenstellung: Zerlege den Vektor $\vec{e}$ in eine Ebenenkomponente und in eine Komponente senkrecht zur Ebene. Die Ebenenkomponente ist dann gerade die (orthogonale) Projektion von $\vec{e}$. Symbolisch übersetzt sich die Frage in die Beziehungen

$$\vec{e} = \vec{e}_{||} + \vec{e}_{\perp} \ , \qquad \vec{e}_{||} = \lambda_1 \vec{v}_1 + \lambda_2 \vec{v}_2 \ ,$$

wobei $\vec{v} \perp \vec{e}_{\perp}$ für alle $\vec{v} \in E$. Aus der Orthogonalitätsbeziehung folgern wir zunächst für die beiden orthogonalen Basisvektoren $\vec{v}_i \in E$

$$0 = \langle \vec{v}_i, \vec{e}_{\perp} \rangle = \langle \vec{v}_i, \vec{e} - \vec{e}_{||} \rangle = \langle \vec{v}_i, \vec{e} - \lambda_1 \vec{v}_1 - \lambda_2 \vec{v}_2 \rangle = \langle \vec{v}_i, \vec{e} \rangle - \lambda_i$$

d. h. die Koordinaten der Projektion $\vec{e}_{||}$ sind $\lambda_i = \langle \vec{v}_i, \vec{e} \rangle$. Tatsächlich ist $\vec{e}_{\perp}$ dann auch senkrecht zu *allen* Ebenenvektoren, da für einen beliebigen Vektor $\vec{v} = \mu_1 \vec{v}_1 + \mu_2 \vec{v}_2$ folgt

$$\langle \vec{v}, \vec{e}_{\perp} \rangle = \mu_1 \langle \vec{v}_1, \vec{e}_{\perp} \rangle + \mu_2 \langle \vec{v}_2, \vec{e}_{\perp} \rangle = 0.$$

Die gefundene Projektion $\vec{e}_{||}$ von $\vec{e}$ auf E bezeichnen wir auch als

$$P_E(\vec{e}) = \langle \vec{v}_1 \rangle \vec{v}_1 + \langle \vec{e}, \vec{v}_2 \rangle \vec{v}_2$$

und den Überrest der Projektion $\vec{e}_{||}$ als

$$Q_E(\vec{e}) = \vec{e} - P_E(\vec{e})$$

Den Winkel zwischen dem Richtungsvektor $\vec{e}$ und der Ebene E ergibt sich somit als $\sphericalangle(\vec{e}, P_E(\vec{e}))$. Aber Vorsicht, es könnte ja sein, daß $P_E(\vec{e}) = \vec{0}$ und der Winkel somit gar nicht definiert ist! In diesem Fall ist $\vec{e} = \vec{e}_\perp = Q_E(\vec{e})$ und eigentlich sollte der Winkel $\frac{\pi}{2}$ resultieren. Als Ausweg berechnen wir zunächst den Winkel zwischen $\vec{e}$ und einem normierten Vektor $\vec{n}$, der senkrecht auf der Ebene steht (ein sogenannter *Normalenvektor*). Der Schnittwinkel ergibt sich dann als Differenz von $\frac{\pi}{2}$ und $\sphericalangle(\vec{e}, \vec{n})$. Die Konstruktion des Normalenvektors ist einfach. Wir nehmen einen beliebigen Vektor $\vec{c}$, der nicht in der Ebene liegt, berechnen $Q_E(\vec{c})$ und normieren diesen Vektor.
In unserem Beispiel führen wir diese Konstruktion mit dem Vektor $\vec{c} = \vec{s}_1$ durch, der den Koordinatenvektor $\vec{e}_1$ hat. Zunächst benötigen wir die Skalarprodukte von $\vec{c}$ mit den orthonormalen Basisvektoren $\vec{v}_1, \vec{v}_2$ der Ebene, deren Koordinaten ja $\vec{\gamma}_1$ und $\vec{\gamma}_2$ sind.

$$\langle \vec{c}, \vec{v}_1 \rangle = \langle \vec{e}_1, \vec{\gamma}_1 \rangle = \frac{1}{\sqrt{3}},$$
$$\langle \vec{c}, \vec{v}_2 \rangle = \langle \vec{e}_1, \vec{\gamma}_2 \rangle = \frac{1}{\sqrt{6}}.$$

Die Koordinaten von

$$Q_E(\vec{c}) = \vec{c} - \frac{1}{\sqrt{3}}\vec{v}_1 - \frac{1}{\sqrt{6}}\vec{v}_2$$

sind dann

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} - \frac{1}{3}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - \frac{1}{6}\begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$$

und nach Normierung ergibt sich der Koordinatenvektor von $\vec{n}$

$$\vec{\eta} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$$

Zur Probe können Sie die Orthogonalität $\langle\vec{\eta},\vec{\gamma}_1\rangle = \langle\vec{\eta},\vec{\gamma}_2\rangle = 0$ noch einmal schnell nachrechnen. Beachten Sie, daß bei der Wahl des Normalenvektors das Vorzeichen willkürlich ist. Wir hätten auch $-\vec{\eta}$ als Koordinatenvektor nehmen können. Der entsprechende Zeiger zeigt dann halt auf die andere Seite der Ebene.
Will man nun den Schnittwinkel der Ebene mit einer Gerade mit Richtungsvektor $\vec{e}$ ermitteln, so wählt man den Normalenvektor, der auf die gleiche Seite wie $\vec{e}$ zeigt, also einen Winkel kleiner oder gleich $\frac{\pi}{2}$ mit $\vec{e}$ bildet. Das Skalarprodukt der beiden Vektoren hat also den Wert $\langle\vec{e},\vec{n}\rangle$ falls $\sphericalangle(\vec{e},\vec{n}) \leq \frac{\pi}{2}$ und $\langle\vec{e},-\vec{n}\rangle$ falls $\sphericalangle(\vec{e},\vec{n}) > \frac{\pi}{2}$, d. h. insgesamt $|\langle\vec{e},\vec{n}\rangle|$. Der Schnittwinkel ist damit

$$\varphi = \frac{\pi}{2} - \arccos\frac{|\langle\vec{e},\vec{n}\rangle|}{||\vec{e}||}.$$

Für den Schnittwinkel von zwei Ebenen ergibt sich mit ähnlichen Überlegungen

$$\varphi = \arccos|\langle\vec{n}_1,\vec{n}_2\rangle|$$

wobei $\vec{n}_1$ und $\vec{n}_2$ Normalenvektoren der beteiligten Ebenen sind.

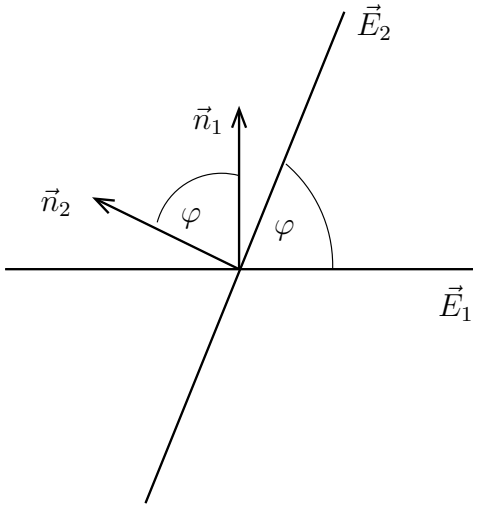


Abbildung 1: Seitenansicht des Schnitts zweier Ebenen

Der Normalenvektor $\vec{n}$ einer Ebene ist übrigens nützlich zur Beschreibung der Ebene selbst. Wir können knapp schreiben

$$E = \{\vec{v} \in S | \langle\vec{v},\vec{n}\rangle = 0\}$$

Allgemein führen wir folgende Bezeichnung ein.

Definition 7. *Sei V ein reeller Vektorraum mit Skalarprodukt und $\emptyset \neq M \subseteq V$. Dann heißt $M^\perp = \{\vec{v} \in V | \langle \vec{v}, \vec{m} \rangle = 0$ für alle $\vec{m} \in M\}$ das orthogonale Komplement von M.*

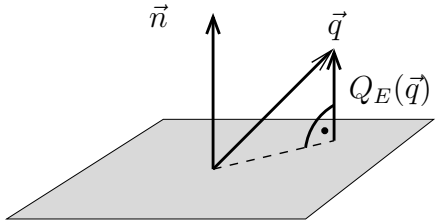
In unserem Fall ist also $E = \{\vec{n}\}^\perp$ das orthogonale Komplement des Vektors $\vec{n}$ und die Gerade $\{\alpha\vec{n} | \alpha \in \mathbb{R}\} = E^\perp$ das orthogonale Komplement der Ebene E. Man überzeugt sich leicht davon, daß das orthogonale Komplement einer Menge stets einen Untervektorraum bildet. So ist das orthogonale Komplement von zwei Zeigern $\{\vec{n}_1, \vec{n}_2\}$ typischerweise eine Gerade, nämlich der Schnitt der beiden Ebenen mit Normalenvektoren $\vec{n}_1$ und $\vec{n}_2$. Die Elemente von $\{\vec{n}_1, \vec{n}_2\}^\perp$ müssen ja senkrecht zu $\vec{n}_1$ *und* $\vec{n}_2$ sein, also in beiden Ebenen liegen.
Will man eine Ebene, die nicht durch den Referenzpunkt des Koordinatensystems läuft, durch ihren Normalenvektor $\vec{n}$ beschreiben, so kann man dies in folgender Form tun

$$\widetilde{E} = \{\vec{v} \in S | \langle \vec{v} - \vec{g}, \vec{n} \rangle = 0\}$$

wobei $\vec{g}$ ein beliebiger Punkt der Ebene ist. Umgekehrt beschreibt die Menge

$$\{\vec{v} \in S | \langle \vec{v}, \vec{n} \rangle = \alpha\}$$

eine affine Ebene. Wählt man nämlich $\vec{g} = \alpha\vec{n}$, so hat diese Menge genau die Form von $\widetilde{E}$, da $\langle -\vec{g}, \vec{n} \rangle = -\alpha$ ist.
Als weitere Anwendung des Normalenvektors berechnen wir zum Schluß noch den Abstand eines Punktes zu einer Ebene. Wir erinnern uns, daß der Ortsvektor $\vec{q}$ eines beliebigen Punktes in eine Komponente $P_E(\vec{q})$ in der Ebene und eine senkrechte Komponente $Q_E(\vec{q})$ zerlegt werden kann. Die Länge der senkrechten Komponente liefert dabei gerade den Abstand der Spitze von $\vec{q}$ zu E.



Da $Q_E(\vec{q}) \in E^\perp$ ist, gilt $Q_E(\vec{q}) = \alpha\vec{n}$, wobei

$$\alpha = \langle Q_E(\vec{q}), \vec{n} \rangle = \langle Q_E(\vec{q}) + P_E(\vec{q}), \vec{n} \rangle = \langle \vec{q}, \vec{n} \rangle.$$

Bis auf das Vorzeichen ist α die Länge von $Q_E(\vec{q})$, da $||\alpha\vec{n}|| = |\alpha|\ ||\vec{n}|| = |\alpha|$. Berechnen wir in unserem Beispiel den Abstand des Punktes $\vec{q}$ mit den Koordinaten

$$\vec{\lambda} = \begin{pmatrix} 2 \\ 3 \\ 5 \end{pmatrix}$$

Es ergibt sich $\langle\vec{\lambda}, \vec{\eta}\rangle = -3/\sqrt{2}$, d. h. der Punkt befindet sich auf der Seite, wo $\vec{n}$ nicht hinzeigt und hat den Abstand $3/\sqrt{2}$.
Das hilfreiche Konzept der orthogonalen Projektion läßt sich vom Fall der Ebene im Zeigervektorraum leicht verallgemeinern.

Definition 8. *Sei V ein reeller Vektorraum mit Skalarprodukt $\langle\cdot,\cdot\rangle$ und U ein m-dimensionaler Untervektorraum von V mit Orthonormalbasis $(\vec{v}_1, \ldots, \vec{v}_m)$. Dann heißt die lineare Abbildung*

$$P_U(\vec{v}) = \sum_{i=1}^{m} \langle\vec{v}, \vec{v}_i\rangle v_i \qquad \vec{v} \in V$$

(orthogonale) Projektion auf U.

Für den Rest $Q_U = I - P_U$ gilt die Orthogonalitätsbeziehung

$$Q_U(\vec{v}_k) = \vec{v}_k - \sum_{i=1}^{m} \langle\vec{v}_k, \vec{v}_i\rangle\vec{v}_i = \vec{v}_k - \vec{v}_k = \vec{0} \qquad k = 1, \ldots, m$$

und damit ist $Q_U(\vec{v}) \in U^\perp$ für alle $\vec{v} \in V$. (Man kann zeigen, daß Q_U eine orthogonale Projektion auf $U^\perp$ ist).
Wegen der Orthogonalität der Zerlegungsvektoren $P_u(\vec{v})$ und $Q_u(\vec{v})$ gilt der Satz des Pythagoras im verallgemeinerten Sinne. Es ist

$$||\vec{v}||^2 = ||P_U(\vec{v}) + Q_U(\vec{v})||^2 = ||P_U(\vec{v})||^2 + ||Q_U(\vec{v})||^2 + 2\langle P_U(\vec{v}), Q_U(\vec{v})\rangle$$

und da $P_U(\vec{v}) \perp Q_U(\vec{v})$

$$||\vec{v}||^2 = ||P_U(\vec{v})||^2 + ||Q_U(\vec{v})||^2$$

Mit dieser Beziehung kann man einen Zusammenhang zwischen der Projektion und der Lösung des folgenden Optimierungsproblems herstellen: Für einen gegebenen Vektor $\vec{v} \in V$, finde den Vektor $\vec{u} \in U$, der am dichtesten an $\vec{v}$ liegt, d. h. der den Abstand $||\vec{v} - \vec{u}||$ so klein wie möglich macht.
Für den Abstand bzw. das Quadrat des Abstandes, gilt nämlich wegen $\langle Q_U(\vec{v}), \vec{u}\rangle = 0$

$$\begin{aligned} ||\vec{v}-\vec{u}||^2 &= ||\vec{u}||^2+||\vec{v}||^2-2\langle\vec{v},\vec{u}\rangle \\ &= ||\vec{u}||^2+||P_U(\vec{v})||^2+||Q_U(\vec{v})||^2-2\langle P_U(\vec{v})+Q_U(\vec{v}),\vec{u}\rangle \\ &= ||\vec{u}-P_U(\vec{v})||^2+||Q_U(\vec{v})||^2 \end{aligned}$$

Der Anteil $||Q_U(\vec{v})||^2$ ist dabei nur von $\vec{v}$ abhängig und bezüglich einer Variation von $\vec{u} \in U$ konstant. Der Abstand wird also alleine durch $||\vec{u}-P_U(\vec{v})|| \geq 0$ bestimmt, wobei der Minimalwert offensichtlich im Punkt

$$\vec{u}_{\text{opt}} = P_U(\vec{v})$$

angenommen wird. Die Projektion $P_U(\vec{v})$ liefert also die *Bestapproximation* an $\vec{v}$ im Unterraum U und diese Eigenschaft verleiht der Projektion eine große Bedeutung. Exemplarisch sei hier eine Anwendung erwähnt, die eigentlich ein wenig aus dem hier betrachteten Rahmen fällt. Sie bezieht sich auf die bilineare Funktion

$$\langle f,g\rangle = \sum_{i=1}^{N} f(x_i)g(x_i) \qquad f,g \in \mathcal{F}$$

auf dem Vektorraum $\mathcal{F}$ der reellwertigen Funktionen auf $\mathbb{R}$, wobei $x_1,\ldots,x_N \in \mathbb{R}$ paarweise verschieden sind. Es ist $\langle\cdot,\cdot\rangle$ zwar kein Skalarprodukt auf $\mathcal{F}$ (welche Bedingungen sind nicht erfüllt?), wohl aber auf den Polynom-Unterräumen $\mathcal{P}_0,\ldots,\mathcal{P}_N$. Trotzdem lassen sich alle unsere Überlegungen bezüglich Projektionen auf diese Unterräume übertragen, insbesondere die zur Bestapproximation. Wir können also die Frage, welches Polynom in $\mathcal{P}_n$ am dichtesten an $f \in \mathcal{F}$ liegt, genau wie oben beantworten: Es ist die $\langle\cdot,\cdot\rangle$ Projektion von f auf $\mathcal{P}_n$. Dieses Polynom $p = P_{\mathcal{P}_n}(f)$ minimiert also die Summe der quadratischen Abweichungen an den Stellen x_i

$$||p-f||^2 = \sum_{i=1}^{N}(p(x_i)-f(x_i))^2$$

(Bei dieser Notation ist Vorsicht geboten, da $||\cdot||$ nicht alle Normbedingungen erfüllt; welche nicht?)
Angewendet wird diese *Methode der kleinsten Quadrate* häufig dann, wenn man einen Zusammenhang f zwischen Parameterwerten x_i und Meßgrößen $y_i = f(x_i)$ durch eine einfache Funktion beschreiben will. Sucht man beispielsweise eine Funktion $g(x) = ax + b$, so daß die Bedingung $g(x_i) \approx y_i$ möglichst gut erfüllt ist, so ergibt sich g gerade als $g = P_{\mathcal{P}_1}(f)$. Zur Berechnung der Projektion kann man die Monombasis von $\mathcal{P}_1$ zunächst mit dem Schmidtschen Orthonormalisierungsverfahren

in eine Orthonormalbasis umwandeln. Die Berechnung von g erfordert dann nur noch die Berechnung zweier Skalarprodukte zwischen f und den beiden Basisvektoren. Genauso kann man die am besten passende Parabel durch Projektion auf $\mathcal{P}_2$ bestimmen und Entsprechendes gilt für Polynome höheren Grades.
Von herausragender Bedeutung ist die Bestapproximation auch beim Lösen partieller Differentialgleichungen aus dem Bereich der mathematischen Physik. Hier ist die Lösung der Differentialgleichung oft durch ein Energieprinzip chrakterisiert, d. h. die Lösung minimiert eine bestimmte Energiefunktion. Diese Energiefunktion spielt beim genaueren Hinsehen die Rolle einer Norm in einem geeigneten Vektorraum von Funktionen und diese Norm ist durch ein Skalarprodukt gegeben. Da man an die Lösung der Gleichung typischerweise Nebenbedingungen stellt (z. B. das Randverhalten der Lösung) und die Menge aller Funktionen, die die Nebenbedingungen erfüllen, oft einen Untervektorraum bilden, ergibt sich die Lösung des Problems als orthogonale Projektion einer Funktion auf den Untervektorraum. Man kann also Differentialgleichungen durch Ausnutzung geometrischer Konzepte in Vektorräumen lösen und diese Beobachtung hat zu einer stürmischen Entwicklung in der Theorie partieller Differentialgleichungen Anlaß gegeben.

8. Rechnen in $\mathbb{R}^2$: komplexe Zahlen

Bisher sind wir in der Lage, Elemente von $\mathbb{R}^n$ mit den üblichen Rechenregeln zu addieren und mit reellen Zahlen zu multiplizieren. Auch haben wir gesehen, daß sich zwei Elemente von $\mathbb{R}^n$ mit einem Skalarprodukt multiplizieren lassen, wobei das Ergebnis allerdings eine reelle Zahl ist. Hierbei gelten zumindest die gewohnten Vertauschungs- und Klammerungsregeln sowie das Distributivgesetz. Die Multiplikation zweier vom Nullvektor verschiedener Tupel kann allerdings Null ergeben (wenn die Tupel bzgl. des Skalarprodukts orthogonal sind) und inverse Elemente kann man nicht sinnvoll definieren, da das Skalarprodukt keine Tupel, sondern Zahlen liefert. Es stellt sich heraus, daß ein Produkt mit *allen* Eigenschaften des Produkts in $\mathbb{R}$ nur für $\mathbb{R}^2$ definiert werden kann (abgesehen natürlich von $\mathbb{R}^1$). Diese sogenannte *komplexe Multiplikation* ist folgendermaßen definiert

$$(x_1, x_2) \cdot (y_1, y_2) = (x_1y_1 - x_2y_2, x_1y_2 + x_2y_1)$$

Der Raum $\mathbb{R}^2$ mit komponentenweiser Addition und komplexer Multiplikation wird *Körper der komplexen Zahlen* genannt und mit dem Symbol $\mathbb{C}$ bezeichnet. Der Grund, warum man plötzlich Zahlenpaare

als komplexe *Zahlen* bezeichnet, ist dabei ganz einfach: Mit den genannten Rechenregeln lassen sich die Paare *genauso* verarbeiten, wie ganze, rationale und reelle Zahlen. Wenn man die Paare mit einzelnen Symbolen abkürzt, also z. B. $z = (x_1, x_2), w = (y_1, y_2)$ u. s. w., dann merkt man beim Rechnen gar nicht, daß z, w *komplexe* Zahlen sind. So gelten z. B. die binomischen Formeln

$$\begin{aligned}(z+w)^2 &= z^2 + 2zw + w^2,\\ (z-w)^2 &= z^2 - 2zw + w^2,\\ (z-w)(z+w) &= z^2 - q^2\end{aligned}$$

für komplexe Zahlen in gewohnter Weise und genauso die Bruchrechenregeln

$$\frac{a}{b} + \frac{c}{d} = \frac{da}{bd} + \frac{bc}{bd} = \frac{ad+bc}{bd}$$

denn alle diese Regeln folgen direkt aus den grundsätzlichen Eigenschaften der Addition und Multiplikation, die ja für komplexe Zahlen genau wie für reelle Zahlen gelten.
Zunächst ist dies aber noch eine Behauptung. Die Grundrechenregeln der Multiplikation

(i) $a \cdot b = b \cdot a$ für alle a, b
(ii) $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ für alle a, b, c
(iii) Es gibt ein Einselement e, so daß $e \cdot a = a$ für alle a
(iv) Zu jedem $a \neq 0$ (wobei 0 das Neutralelement der Addition ist) gibt es ein inverses Element a^{-1}, so daß $a^{-1}a = e$
(v) $a(b+c) = ab + ac$ für alle a, b, c

müssen wir zunächst noch für die komplexe Multiplikation überprüfen. Die Regeln $(i), (ii)$ und (v) folgen unmittelbar durch Vergleich von linker und rechter Seite, z. B. für (v)

$$\begin{aligned}(x_1, x_2)\big((y_1, y_2) + (z_1, z_2)\big) &= (x_1, x_2)(y_1 + z_1, y_2 + z_2)\\ &= \big(x_1(y_1+z_1) - x_2(y_2+z_2), x_1(y_2+z_2) + x_2(y_1+z_1)\big)\\ &= (x_1y_1 - x_2y_2, x_1y_2 + x_2y_1) + (x_1z_1 - x_2z_2, x_1z_2 + x_2z_1)\\ &= (x_1, x_2)(y_1, y_2) + (x_1, x_2)(z_1, z_2)\end{aligned}$$

Interessanter ist da schon die Suche nach dem Einselement $e = (e_1, e_2)$, für das ja gelten muß

$$(e_1x_1 - e_2x_2, e_1x_2 + e_2x_1) = (x_1, x_2)$$

für *alle* $(x_1, x_2) \in \mathbb{R}^2$.
Schon mit der Wahl $(x_1, x_2) = (1, 0)$ ergibt sich $(e_1, e_2) = (1, 0)$ und tatsächlich gilt für beliebige (x_1, x_2)

$$(1,0)(x_1, x_2) = (1 \cdot x_1 - 0 \cdot x_2, 1 \cdot x_2 + 0 \cdot x_1) = (x_1, x_2)$$

Mit dem Einselement $(1, 0)$ können wir uns nun auf die Suche nach inversen Elementen machen. Sei dazu $(x_1, x_2) \neq (0, 0)$. Gesucht ist (y_1, y_2), so daß

$$(y_1, y_2)(x_1, x_2) = (1, 0)$$

gilt, bzw. ausgeschrieben

$$\begin{aligned} y_1 x_1 - y_2 x_2 &= 1, \\ y_1 x_2 + y_2 x_1 &= 0 \end{aligned}$$

Dieses lineare Gleichungssystem läßt sich folgendermaßen lösen. Wir multiplizieren die erste Gleichung mit x_1 und die zweite mit x_2 und addieren die resultierenden Gleichungen, was

$$y_1(x_1^2 + x_2^2) = x_1$$

ergibt. Multiplikation der beiden Gleichungen mit x_2 bzw. x_1 und anschließender Subtraktion liefert

$$y_2(x_1^2 + x_2^2) = -x_2$$

Da $x_1^2 + x_2^2 > 0$ nach Annahme, kann man weiter auflösen

$$(y_1, y_2) = \left(\frac{x_1}{x_1^2 + x_2^2}, \; -\frac{x_2}{x_1^2 + x_2^2} \right)$$

So sehen also die inversen Elemente aus. Der im Nenner auftretende Wert $x_1^2 + x_2^2$ ist übrigens das Quadrat der Norm zum Standardskalarprodukt. Bezüglich der Rechenregeln in $\mathbb{C}$ verhält sich diese Norm genauso, wie der Betrag in den reellen Zahlen. Man definiert deshalb

$$|z| = \sqrt{x_1^2 + x_2^2} \qquad z = (x_1, x_2)$$

als Betrag der komplexen Zahlt z.
Die beiden Komponenten x_1, x_2 einer komplexen Zahl $z = (x_1, x_2)$ werden traditionell als Realteil und Imaginärteil bezeichnet

$$Re(x_1, x_2) = x_1, \qquad Im(x_1, x_2) = x_2$$

Beim Arbeiten mit komplexen Zahlen folgt man wie bei den reellen Zahlen dem Prinzip, daß das Einselement e bei Multiplikationen und das Nullelement bei Additionen schreibtechnisch unterdrückt wird. Führt

man für die komplexe Zahl $(0,1)$ das spezielle Symbol i ein (die sogenannte imaginäre Einheit), dann gilt

$$z = (x_1, x_2) = x_1(1,0) + x_2(0,1) = x_1 e + x_2 i$$

bzw. nach Unterdrückung von e in der skalaren Multiplikation $x_1 e$

$$z = x_1 + x_2 i$$

Als Spezialfall dieser Notation ergibt sich

$$(\alpha, 0) = \alpha + 0 \cdot i = \alpha, \qquad \alpha \in \mathbb{R}$$

wobei hier die Null in der Addition unterdrückt wurde. In diesem Sinne sind die reellen Zahlen $\alpha \in \mathbb{R}$ also als $(\alpha, 0) \in \mathbb{C}$ in den komplexen Zahlen enthalten. Insbesondere gilt $e = (1,0) = 1$. Diese Interpretation paßt auch wunderbar mit der skalaren Multiplikation zusammen, denn

$$(\alpha, 0)(x_1, x_2) = (\alpha x_1, \alpha x_2) = \alpha(x_1, x_2), \quad \alpha \in \mathbb{R}$$

d. h. die skalare Multiplikation entspricht der komplexen Multiplikation im Spezialfall eines reellen Multiplikators. Insbesondere gilt wegen der Kommutativität

$$x_2 i = (x_2, 0)(0,1) = (0,1)(x_2, 0) = i x_2$$

Als Zusammenfassung dieser kosmetischen Betrachtungen folgt die sogenannten *Normalform* einer komplexen Zahl

$$z = Re(z) + iIm(z)$$

Natürlich ist das Spannende an den komplexen Zahlen der imaginäre Anteil, da wir mit dem reellen Anteil ja schon aus vielen Schuljahren bestens vertraut sind. Eine sehr wichtige Beobachtung ist dabei die folgende:

$$i^2 = ii = (0,1)(0,1) = (0 \cdot 0 - 1 \cdot 1, 0 \cdot 1 + 1 \cdot 0) = (-1, 0) = -1$$

also kurz $i^2 = -1$. Das gibt's in den reellen Zahlen nicht! Die komplexen Zahlen leisten also etwas mehr als die reellen Zahlen. Mit ihnen lassen sich Probleme lösen, die in den reellen Zahlen nicht lösbar sind, wie z. B. das Problem, eine Zahl z zu finden mit der Eigenschaft $z^2 = -1$. Das kann *keine* reelle Zahl, da für reelle Zahlen z immer gilt $z^2 \geq 0$. Trotzdem hat das Problem Lösungen, wenn wir komplexe Zahlen zulassen, nämlich $z = i$ oder $z = -i$, wie wir oben gesehen haben. Erinnern Sie sich an den Grund für die Einführung der reellen Zahlen? Die reellen Zahlen wurden eingeführt, weil in den rationalen Zahlen bestimmte Gleichungen wie z. B. $x^2 = 2$ nicht lösbar sind, was im Widerspruch zu unserer Vorstellung eines Kontinuums steht. Genauso kann man die komplexen Zahlen als Erweiterung der reellen Zahlen betrachten, wobei man nun auch Gleichungen wie $z^2 = -1$ lösen will. Als Beispiel,

wozu dies nützlich sein kann, sei hier die kardanische Formel erwähnt. Sie liefert alle reellen Nullstellen des kubischen Polynoms $x^3 + 3px + q$ mit $p, q \in \mathbb{R}$, wobei jedes Polynom dritten Grades durch eine einfache Substitution in diese Form gebracht werden kann. Die Aussage ist folgende: Die reellen Lösungen von $x^3 + 3px + q = 0$ sind gegeben durch $x = u + v \in \mathbb{R}$ wobei u^3, v^3 die beiden Lösungen der quadratischen Gleichung $y^2 + qy - p^3 = 0$ sind. Wie wir später sehen werden, kann es passieren, daß in der Rechnung komplexe Zahlen auftreten, obwohl das Endergebnis drei *reelle* Nullstellen liefert. Die komplexen Zahlen sind dann offensichtlich ein gutes Hilfsmittel, um an reelle Lösungen heranzukommen.

Der Punkt, an dem die komplexen Zahlen ins Spiel kommen, ist die Lösung der quadratischen Gleichung. In $\mathbb{C}$ hat eine quadratische Gleichung immer zwei Lösungen, oder genauer gesagt, ein Polynom $x^2 + ax + b$ mit $a, b \in \mathbb{R}$ läßt sich immer faktorisieren als

$$x^2 + ax + b = (x - \lambda_1)(x - \lambda_2) \qquad \lambda_1, \lambda_2 \in \mathbb{C}$$

Zur Bestimmung von λ_1, λ_2 benutzen wir wie aus $\mathbb{R}$ gewohnt, den Prozeß der quadratischen Ergänzung. Zunächst gilt

$$x^2 + ax + b = \left(x + \frac{a}{2}\right)^2 - \frac{a^2}{4} + b$$

so daß mit $w = x + \frac{a}{2}$ das Nullstellenproblem auf die Gleichung

$$w^2 = \frac{a^2}{4} - b \tag{14}$$

führt. Diese Gleichung hat nur dann reelle Lösungen, wenn $\frac{a^2}{4} - b \geq 0$ ist. Wie sieht es aber mit komplexen Lösungen aus? Nehmen wir an, daß $w = w_1 + iw_2$ eine komplexe Lösung von (14) ist. Zunächst gilt mit der ersten binomischen Formel

$$w^2 = (w_1 + iw_2)^2 = w_1^2 + 2w_1 iw_2 + (iw_2)^2 = w_1^2 - w_2^2 + 2w_1 w_2 i$$

Eingesetzt in (14) erhalten wir die Gleichung

$$w_1^2 - w_2^2 + 2iw_1 w_2 = \frac{q^2}{4} - b$$

für die beiden Unbekannten w_1 und w_2. Genügt eine Gleichung überhaupt, um zwei Unbekannte zu finden? Natürlich nicht, aber beachten Sie, daß eine Gleichung mit komplexen Zahlen aus *zwei* Bedingungen besteht. Zwei komplexe Zahlen sind nämlich genau dann gleich, wenn beide Komponenten, der Real- und der Imaginärteil, gleich sind. Dabei entspricht die reelle Zahl auf der rechten Seite der Gleichung gemäß unserer kosmetischen Konvention dem Zahlenpaar

$$\frac{a^2}{4} - b = \left(\frac{a^2}{4} - b\right) e + i0 = \left(\frac{a^2}{4} - b, 0\right)$$

Wir finden also durch Vergleich der Real- und Imaginärteile die beiden Bedingungen

$$w_1^2 - w_2^2 = \frac{a^2}{4} - b, \qquad 2w_1 w_2 = 0$$

Die zweite Bedingung erzwingt, daß w_1 oder w_2 gleich Null ist. Zusammen mit der ersten Bedingung ergibt dies die Lösungen

$$w_1 \in \left\{ \sqrt{\frac{a^2}{4} - b}, -\sqrt{\frac{a^2}{4} - b} \right\}, \quad w_2 = 0, \qquad \frac{a^2}{4} - b \geq 0$$

beziehungsweise

$$w_2 \in \left\{ \sqrt{b - \frac{a^2}{4}}, -\sqrt{b - \frac{a^2}{4}} \right\}, \quad w_1 = 0, \qquad \frac{a^2}{4} - b < 0$$

Der erste Fall entspricht dabei den üblichen reellen Lösungen und im zweiten Fall sind die Lösungen rein imaginär. Da wir die Abkürzung $w = x + \frac{q}{2}$ eingeführt hatten, ergeben sich die Lösungen des Ausgangsproblems

$$x \in \left\{ -\frac{q}{2} \pm \sqrt{\frac{a^2}{4} - b} \right\} \qquad \frac{a^2}{4} - b \geq 0$$

$$x \in \left\{ -\frac{q}{2} \pm i\sqrt{b - \frac{a^2}{4}} \right\} \qquad \frac{a^2}{4} - b < 0$$

Beachten Sie, daß die beiden Lösungen im zweiten Fall sich nur im Vorzeichen des Imaginärteils unterscheiden. Darauf werden wir später noch einmal zurückkommen.

Die gerade gezeigte Tatsache, daß sich jedes reelle Polynom zweiten Grades in Terme $x - \lambda_i$ mit $\lambda_i \in \mathbb{C}$ faktorisieren läßt, hat folgende wichtige Verallgemeinerung: Jedes *komplexe Polynom* vom Grad n

$$z^n + a_{n-1} z^{n-1} + a_{n-2} z^{n-2} + \ldots + a_1 z + a_o \qquad a_i \in \mathbb{C}$$

läßt sich in n Faktoren $(x - \lambda_i)$ mit geeigneten $\lambda_i \in \mathbb{C}$ zerlegen, d. h.

$$z^n + a_{n-1} z^{n-1} + \ldots + a_1 z + a_o = (z - \lambda_1)(z - \lambda_2) \ldots (z - \lambda_n)$$

Anders ausgedrückt bedeutet dieser sogenannte *Fundamentalsatz der Algebra*, daß jedes komplexe Polynom n-ten Grades n Nullstellen besitzt (wenn wir die mögliche Vielfachheit von Nullstellen berücksichtigen). Den Spezialfall $n = 2$ werden wir später explizit behandeln. Dazu

benötigen wir unter anderem den Begriff der komplexen Wurzel, wozu noch einige Vorbereitungen nötig sind.
Zunächst sei aber noch eine andere Konsequenz der Gleichung $i^2 = -1$ erwähnt. Sie kann nämlich als einfache Gedankenstütze für die komplexe Multiplikation genutzt werden. Sie müssen sich nur zusätzlich merken, daß man mit komplexen Zahlen genau wie mit reellen Zahlen rechnet. Dann ergibt sich sofort

$$\begin{aligned}(x_1, x_2) \cdot (y_1, y_2) \\ &= (x_1 + ix_2)(y_1 + iy_2) \;=\; x_1y_1 + i^2x_2y_2 + ix_1y_2 + ix_2y_1 \\ &= x_1y_1 - x_2y_2 + i(x_1y_2 + x_2y_1) \\ &= (x_1y_1 - x_2y_2,\; x_1y_2 + x_2y_1)\end{aligned}$$

Eine zweite Bemerkung bezieht sich auf das Arbeiten mit inversen Elementen. Zunächst führt man wie bei den reellen Zahlen die Bruchschreibweise ein, d. h.

$$\frac{w}{z} = z^{-1}w, \qquad z \neq 0$$

Insbesondere ist also

$$\frac{1}{z} = z^{-1}, \qquad z \neq 0.$$

Will man die komplexe Zahl $\frac{w}{z}$ in Normalform bringen, z. B. um den Real- und Imaginärteil leicht ablesen zu können, bedient man sich des folgenden Tricks. Zunächst führt man die zu einer komplexen Zahl *konjugiert komplexe* Zahl $\overline{z}$ ein gemäß

$$\overline{z} = Re(z) - iIm(z)$$

Mit der dritten binomischen Formel rechnet man dann leicht aus, daß für $z = x + iy$ mit $x, y \in \mathbb{R}$ folgender Zusammenhang gilt

$$z \cdot \overline{z} = (x + iy)(x - iy) = x^2 - (iy)^2 = x^2 + y^2 = |z|^2$$

Das Produkt $z \cdot \overline{z}$ ist also immer eine reelle Zahl, und diese Eigenschaft kann man nutzen, um den Nenner in einem komplexen Bruch reell zu machen. Man erweitert dazu w/z einfach mit $\overline{z}$. Sind die Normalformen von w und $z \neq 0$ gegeben durch $w_1 + iw_2$ und $z_1 + iz_2$, so erhalten wir

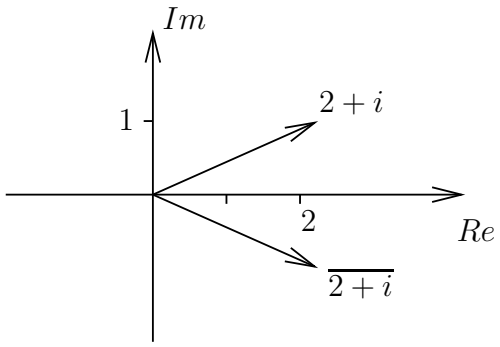
$$\frac{w}{z} = \frac{\overline{z}w}{\overline{z}z} = \frac{\overline{z}w}{|z|^2} = \frac{z_1w_1 + z_2w_2}{z_1^2 + z_2^2} + i\frac{z_1w_2 - z_2w_1}{z_1^2 + z_2^2}$$

Das ist der einzige Trick, den man beim komplexen Bruchrechnen beachten muß.

Beim Arbeiten mit reellen Zahlen ist es oft hilfreich, daß man die Zahlen mit Punkten auf einer Geraden identifiziert, was geometrische Anschauung liefert und damit die Intuition fördert. In ähnlicher Weise kann man auch die komplexen Zahlen mit geometrischen Objekten in Verbindung bringen.
Wir betrachten dazu $Re(z)$ und $Im(z)$ als Punktkoordinaten bezüglich eines kartesischen Koordinatensystems $(A, \vec{s}_1, \vec{s}_2)$ in einer Ebene. Jede komplexe Zahl z entspricht damit genau einem Punkt in dieser sogenannten komplexen Zahlenebene. Da Punkte wiederum mit Ortsvektoren bezüglich A identifiziert werden können, läßt sich die Zahl z auch als Zeiger

$$Re(z)\vec{s}_1 + Im(z)\vec{s}_2$$

veranschaulichen.



Der Addition in $\mathbb{C}$ entspricht dann unsere gewohnte Zeigeraddition und die komplexe Konjugation bedeutet geometrisch eine Spiegelung an der Re-Achse. Offensichtlich ist die Spiegelung $(x_1, x_2) \to (x_1, -x_2)$ eine lineare Abbildung, d. h.

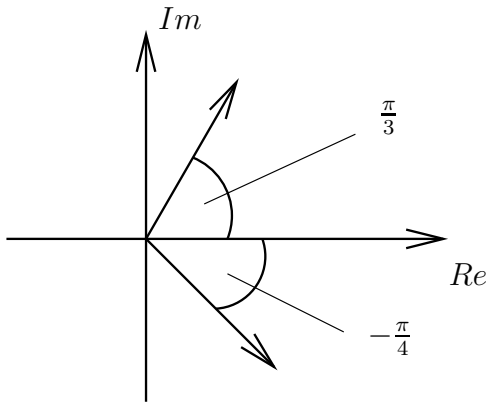
$$\overline{z + w} = \overline{z} + \overline{w}, \quad \overline{\alpha z} = \alpha \overline{z}, \qquad \alpha \in \mathbb{R}.$$

Interessanter ist aber sicherlich die geometrische Bedeutung der neuen Operation, d. h. der komplexen Multiplikation. Um die Multiplikation geometrisch zu verstehen, führt man zunächst die *Polardarstellung* komplexer Zahlen ein. Dazu definiert man einen vorzeichenbehafteten Winkel zwischen $z \neq 0$ und $e = (1, 0)$

$$\arg(z) = \begin{cases} \arccos \frac{Re(z)}{|z|} & Im(z) \geq 0 \\ -\arccos \frac{Re(z)}{|z|} & Im(z) < 0 \end{cases}$$

das sogenannte Argument von z.

Geometrisch ist dies gerade der Winkel zwischen dem zu z gehörenden Zeiger und der reellen Achse, wobei der Winkel negativ gerechnet wird, wenn der Zeiger nach unten zeigt.



Für eine beliebige komplexe Zahl $z \neq 0$ gilt dann

$$z = |z| \, (\cos \ \arg(z) + i \ \sin \ \arg(z))$$

Seien nun

$$z = r(\cos\varphi + i \ \sin\varphi), \qquad w = s(\cos\psi + i \ \sin\psi)$$

zwei komplexe Zahlen in der Polardarstellung. Für das Produkt wz gilt dann

$$\begin{aligned} wz &= rs(\cos\varphi \ \cos\psi - \sin\varphi \ \sin\psi + i(\cos\varphi \ \sin\psi + \sin\varphi \ \cos\psi)) \\ &= rs(\cos(\varphi+\psi) + i \ \sin(\varphi+\psi)) \end{aligned}$$

wobei im letzten Schritt die Additionstheoreme benutzt wurden. Geometrisch gesehen, multiplizieren sich also die Längen der Vektoren und die Winkel addieren sich. Die Abbildung $z \mapsto w \cdot z$ mit einer festen komplexen Zahl $w \neq 0$ „streckt“ also alle komplexen Zahlen um den Faktor $s = |w|$ und „dreht“ die Zahlen um den Winkel $\psi = \arg(w)$.
Die Multiplikation mit einer komplexen Zahl entspricht also geometrisch einer Drehstreckung in der Zahlenebene.
Die Addition der Argumente bei Multiplikation der komplexen Zahlen wird besonders deutlich mit der Eulerformel, die die Polardarstellung mit der Exponentialfunktion in Verbindung bringt.

$$\exp(i\varphi) = e^{i\varphi} = \cos\varphi + i \ \sin\varphi$$

Dabei berechnet sich die Exponentialfunktion für komplexe Argumente genau wie für reelle Argumente durch die Potenzreihe

$$\exp(z) = \sum_{n=0}^{\infty} \frac{z^n}{n!}, \qquad z \in \mathbb{C}$$

Wie diese Rechenvorschrift genau zu verstehen ist, werden wir später noch detailliert betrachten. Im Prinzip müssen unendlich viele Terme zusammenaddiert werden

$$\exp(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24} + \dots$$

Für das Arbeiten mit komplexen Zahlen spielt diese Rechenvorschrift nur insofern eine Rolle, als daß sie die Exponentialeigenschaft

$$\exp(z+w) = \exp(z)\exp(w) \qquad z, w \in \mathbb{C}$$

impliziert und eben die Eulerformel liefert.
Mit der Eulerformel läßt sich also die Polardarstellung von z und w knapp als

$$z = |z|e^{i\ \arg(z)}, \quad w = |w|e^{i\ \arg(w)}$$

schreiben und das Produkt ergibt sich auch durch die exp-Eigenschaft

$$wz = |w|\ |z|e^{i\ \arg(w)}e^{i\ \arg(z)} = |w|\ |z|e^{i(\arg(w)+\arg(z))}$$

Als unmittelbare Konsequenz der Eulerformel sei erwähnt, daß $|e^{i\varphi}| = 1$ gilt, d. h. die Punkte $e^{i\varphi}$ liefern den Einheitskreis in der Zahlenebene, wenn φ das Intervall $-\pi$ bis π durchläuft. Außerdem folgt mit der Eulerformel aus $e^{i\varphi} = 1$, daß $\cos\varphi = 1$ und $\sin\varphi = 0$ ist, was wiederum auf $\varphi = 2\pi k$ mit irgendeinem $k \in \mathbb{Z}$ führt. Diese Mehrdeutigkeit der Lösungen von $e^{i\varphi} = 1$ darf man nicht vernachlässigen, wie wir nun am Beispiel der komplexen Wurzel sehen werden.
Sei dazu $0 \neq a \in \mathbb{C}$ eine vorgegebene Zahl. Jede Zahl $w \in \mathbb{C}$ für die $w^n = a$ gilt, bezeichnen wir als n-te Wurzel von a. Beachten Sie, daß in diesem Sinne $a = 1$ zwei reelle Wurzeln hat, denn sowohl $1^2 = 1$ als auch $(-1)^2 = 1$. Der Begriff zweite Wurzel unterscheidet sich also von der Wurzelfunktion, die per Definition immer nur die positive Wurzel liefert. Wir wollen nun allgemein die n-ten Wurzeln von a bestimmen. Da es beim Potenzieren um n-fache Multiplikation geht, bietet es sich an, mit der Polardarstellung zu arbeiten, da in der Polardarstellung die Multiplikation eine einfach nachzuvollziehende Operation ist. Sei also $a = re^{i\varphi}$ und die gesuchte Zahl $w = se^{i\psi}$. Wir wollen nun s und ψ so wählen, daß

$$a = re^{i\varphi} = w^n = (se^{i\psi})^n = s^n e^{in\psi}$$

gilt. Offensichtlich erreichen wir dies, wenn $s = \sqrt[n]{r}$ ist und $e^{i\varphi} = e^{in\psi}$. Die Exponentialrelation formen wir noch um, indem wir durch $e^{in\psi}$ dividieren. Beachten Sie, daß wegen der Exponential-Eigenschaft

$$e^{-in\psi}e^{in\psi} = e^0 = 1$$

gilt, d. h. allgemein ist $|z|^{-1}e^{-i\arg(z)}$ das inverse Element zu z. Wir erhalten damit

$$e^{i(\varphi - n\psi)} = 1$$

woraus wegen der Mehrdeutigkeit

$$\varphi - n\psi = 2\pi k, \qquad k \in \mathbb{Z}$$

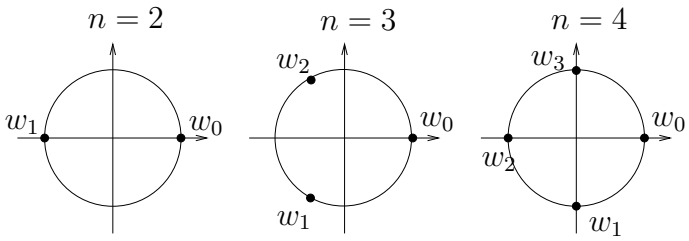
folgt. Aufgelöst nach ψ sehen wir, daß sich mehr als eine n-te Wurzel von a ergibt

$$w_k = \sqrt[n]{r}e^{i(\frac{\varphi}{n} - \frac{2\pi}{n}k)}, \qquad k \in \mathbb{Z}, \quad a = re^{i\varphi}.$$

Schauen wir uns diese Lösungen einmal genauer an. Sie bestehen aus der Grundlösung w_0 multipliziert mit $\exp(-i2\pi k/n)$

$$w_k = w_0 e^{-i\frac{2\pi}{n}k}, \qquad w_0 = \sqrt[n]{r}e^{i\frac{\varphi}{n}}$$

Dabei ist sofort einsichtig, daß $w_0^n = re^{i\varphi}$ ist. Die anderen Lösungen w_k erhält man nun, indem man w_0 um das Vielfache des Winkels $\frac{-2\pi}{n}$ dreht, denn Multiplikation mit $\exp(-i2\pi k/n)$ hat ja gerade diese Wirkung. Dabei ist klar, daß nach n Drehungen der Ausgangspunkt wieder erreicht wird und für größeres k die gleichen Zahlen durchlaufen werden. Es gibt also genau n *verschiedene* komplexe Wurzeln $w_0, \ldots, w_{n-1}$. Als Beispiel schauen wir uns einige *Einheitswurzeln* an, d. h. Wurzeln von $a = 1$. Zunächst brauchen wir dazu die Polardarstellung von a, also $1 = 1 \cdot e^{i0}$. Die Grundlösung ist damit immer $w_0 = 1$. Im Fall $n = 2$ ist die andere Lösung durch Drehung um $-2\pi/2 = -\pi$ gegeben, also $w_1 = -1$. Dies sind natürlich genau die bekannten Wurzeln von 1. Im Fall der dritten Wurzeln wird die Grundlösung w_0 um $-2\pi/3$ gedreht (120°) und bei den vierten Wurzeln um $-2\pi/4$.



Als weiteres Beispiel berechnen wir die zweiten Wurzeln von $a = i$. Wieder benötigen wir dazu die Polardarstellung $i = 1e^{i\frac{\pi}{2}}$ was zur Grundlösung $w_0 = e^{i\frac{\pi}{4}}$ führt. Diese wird um 180° gedreht, um die andere zweite Wurzel zu finden.

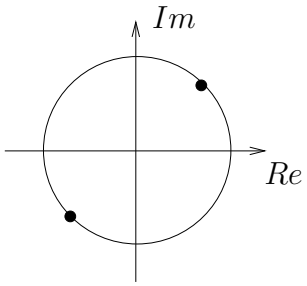


Abbildung 2: Zweite Wurzeln von $a = i$

Da bei der Grundlösung w_0 das Argument $\arg(a)$ von a durch n dividiert wird, rutscht w_0 im Vergleich zu a also näher an die reelle Achse und zwar umso mehr, je größer n ist. Da $\overline{a}$ spiegelbildlich zu a bezüglich der reellen Achse liegt, verhält sich die n-te Wurzel-Grundlösung zu $\overline{a}$ ebenfalls spiegelbildlich. Wir finden also $\overline{w}_0$ als Grundlösung zu $\overline{a}$. Tatsächlich gilt wegen der Euler-Formel allgemein

$$re^{-i\varphi} = r(\cos(-\varphi) + i\sin(-\varphi))$$

und da $\cos(-\varphi) = \cos\varphi,\ \sin(-\varphi) = -\sin\varphi$ ergibt sich also

$$re^{-i\varphi} = r(\cos\varphi - i\sin\varphi) = \overline{re^{i\varphi}}$$

Konjugation äußert sich also als Vorzeichenänderung des Arguments, was auch sofort klar ist, wenn man sich die Definition von $\arg(z)$ anschaut, wo ja das Vorzeichen von φ nur vom Vorzeichen von $Im(z)$ abhängt. Als direkte Konsequenz aus dieser Beobachtung können wir für das Produkt von $z = re^{i\varphi}$ und $w = se^{i\psi}$ nachrechnen

$$\overline{w}\,\overline{z} = re^{i\varphi}se^{-i\psi} = rs\,e^{-i(\varphi+\psi)} = \overline{wz}$$

Insbesondere ergibt sich durch mehrmaliges Anwenden von $\overline{z}\,\overline{z} = \overline{zz}$,

$$\overline{z^n} = \overline{z}^n \qquad z \in \mathbb{C}.$$

Damit sehen wir den Zusammenhang zwischen den Wurzeln von a und denen von $\overline{a}$. Sind $w_0, \ldots, w_{n-1}$ die n-ten Wurzeln von a, so sind $\overline{w}_0, \ldots, \overline{w}_{n-1}$ die von $\overline{a}$, denn

$$a = w^n \Longleftrightarrow \overline{a} = \overline{w^n} = \overline{w}^n$$

Da wir nun in der Lage sind, Quadratwurzeln von beliebigen komplexen Zahlen zu finden, können wir auch allgemeine quadratische Gleichungen lösen. Für gegebene $a, b \in \mathbb{C}$ seien die Lösungen von

$$z^2 + az + b = 0$$

gesucht. Mit Hilfe der quadratischen Ergänzung läßt sich diese Gleichung umschreiben als

$$\left(z + \frac{a}{2}\right)^2 = \frac{a^2}{4} - b$$

und wir sehen, daß die Lösungen die Form

$$z = -\frac{a}{2} + w, \qquad w^2 = \frac{a^2}{4} - b$$

haben. Zur Bestimmung von w müssen Sie nur $\frac{a^2}{4} - b$ in die Polardarstellung bringen, das Argument halbieren und die positive Wurzel des Betrags ermitteln (die andere Quadratwurzel erhält man durch Drehung um 180°, d. h. Änderung des Vorzeichens).
Als Anwendung der kubischen komplexen Wurzeln betrachten wir noch einmal die kardanische Formel. Mit ihr lassen sich reelle Lösungen der Gleichung

$$x^3 + 3px + q = 0, \qquad p, q \in \mathbb{R} \tag{15}$$

finden. Wir unterteilen den Lösungsprozeß in drei Schritte. Im ersten Schritt ermittelt man a, b, so daß

$$y^2 + qy - p^3 = (y - a)(y - b) = y^2 - (a + b)y + ab \tag{16}$$

gilt, d. h. wir suchen die Nullstellen a, b vom Polynom $y^2 + qy - p^3$. Im zweiten Schritt werden Kubikwurzeln u von a und v von b ermittelt, so daß u, v entweder reell oder konjugiert komplex sind. Im letzten Schritt ergibt sich für jedes zulässige Paar u, v eine reelle Lösung von

(15) gemäß $x = u + v$. Bevor wir ein konkretes Beispiel durchrechnen, soll noch kurz erklärt werden, warum dieser Prozeß funktioniert. Zunächst gilt für Schritt 1, daß die Nullstellen a, b entweder beide reell oder konjugiert komplex sind. Für den Fall einer quadratischen Gleichung mit reellen Koeffizienten haben wir dies ja bereits explizit nachgerechnet. Die Aussage gilt aber sogar allgemeiner. Ist $z \in \mathbb{C}$ Nullstelle eines *reellen* Polynoms, d. h. gilt

$$a_n z^n + a_n + z^{n-1} + \ldots + a_1 z + a_0 = 0, \quad a_i \in \mathbb{R}$$

dann ist $\overline{z}$ ebenfalls eine Nullstelle. Wie man leicht nachrechnet, ist nämlich

$$0 = \overline{0} = \overline{a}_n \overline{z^n} + \ldots + \overline{a}_1 \overline{z} + \overline{a}_0 = a_n \overline{z}^n + \ldots + a_1 \overline{z} + a_0$$

da für reelle Zahlen a_i die Beziehung $\overline{a}_i = a_i$ gilt. Kurz gesagt, liegen Nullstellen von reellen Polynomen immer spiegelbildlich zur reellen Achse. Sind die beiden Nullstellen a und b reell und verschieden, so sind u und v eindeutig durch $u = \sqrt[3]{a}$ bzw. $v = \sqrt[3]{b}$ gegeben. Ist dagegen $b = \overline{a}$, so hat a die Kubikwurzeln w_0, w_1, w_2 und $\overline{a}$ die konjugiert komplexen $\overline{w}_0, \overline{w}_1, \overline{w}_2$, so daß das Bilden von konjugiert komplexen Paaren $u = w_k, v = \overline{w}_k$ möglich ist.
Aus einem Koeffizientenvergleich in (16) sehen wir, daß

$$-p^3 = a \cdot b = (uv)^3, \qquad -q = a + b = u^3 + v^3$$

gilt. Da $u \cdot v$ in jedem Fall reell ist, finden wir insbesondere $uv = -p$. Für $x = u + v$ gilt dann

$$\begin{aligned} x^3 &= (u+v)^3 = u^3 + 3u^2 v + 3uv^2 + v^3 \\ &= -q + 3uv(u+v) = -q - 3px \end{aligned}$$

d. h. x löst (15). Im Fall von nicht-reellen $u, v = \overline{u}$ ist x dabei trotzdem reell, denn für beliebige komplexe Zahlen gilt

$$z + \overline{z} = 2Re(z), \qquad z - \overline{z} = 2iIm(z)$$

Schauen wir uns zum Abschluß ein Beispiel an, wo wirklich der Weg über die komplexen Zahlen genommen wird, um drei reelle Lösungen von (15) zu ermitteln

$$x^3 - 3x - \sqrt{2} = 0, \qquad p = -1, q = -\sqrt{2}$$

Zunächst lösen wir dazu

$$y^2 - \sqrt{2}y + 1 = 0$$

was nach quadratischer Ergänzung auf

$$y = \frac{1}{\sqrt{2}} + q, \qquad w^2 = \frac{1}{2} - 1 = -\frac{1}{2}$$

führt. Die Quadratwurzeln von $-\frac{1}{2}$ sind $\frac{i}{\sqrt{2}}$ und $-\frac{i}{\sqrt{2}}$, so daß

$$a = \frac{1}{\sqrt{2}}(1+i), \qquad b = \frac{1}{\sqrt{2}}(1-i)$$

Zur Bestimmung der Kubikwurzeln von a benötigen wir zunächst die Polardarstellung. Die Zahl a hat offensichtlich Länge 1 und zeigt in Diagonalrichtung, also $a = e^{i\frac{\pi}{4}}$. Damit ergibt sich die Grundlösung $w_0 = e^{i\frac{\pi}{12}}$, die dann zweimal um den Winkel $\frac{2\pi}{3}$ gedreht wird, um die anderen Wurzeln zu erhalten.

$$w_0 = e^{i\frac{\pi}{12}}, \quad w_1 = e^{i\left(\frac{\pi}{12}-\frac{2\pi}{3}\right)}, \quad w_2 = e^{i\left(\frac{\pi}{12}-\frac{4\pi}{3}\right)}$$

Addition mit den entsprechenden konjugiert komplexen Kubikwurzeln von $b = \overline{a}$, liefert schließlich die drei reellen Lösungen

$$x_0 = 2\cos\frac{\pi}{12}, \quad x_1 = 2\cos\left(\frac{7}{12}\pi\right), \quad x_2 = 2\cos\left(\frac{5}{4}\pi\right)$$

Neben dem Lösen von algebraischen Gleichungen eignen sich komplexe Zahlen besonders zur Vereinfachung von Rechnungen mit den trigonometrischen Funktionen sin und cos. Dies kommt z. B. in der Elektrotechnik zum Tragen, wenn mit Wechselstromkreisen gearbeitet wird. Durch den Herstellungsprozeß hat dieser Strom annähernd den zeitlichen Verlauf

$$I(t) = J_0\cos(wt+\alpha), \quad t \in \mathbb{R}.$$

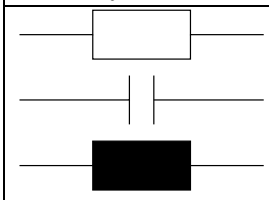
Will man das Verhältnis von Strom und Spannung an elektrischen Grundelementen wie Ohmsche Widerstände, Kondensatoren und Spulen untersuchen, so ist es vorteilhaft, U und I als Realteile von einer komplexen Spannung U^c bzw. eines komplexen Stroms I^c zu betrachten

$$I(t) = Re\left(J_0 e^{i(wt+\alpha)}\right), \qquad U(t) = Re\left(U_0 e^{i(wt+\beta)}\right)$$

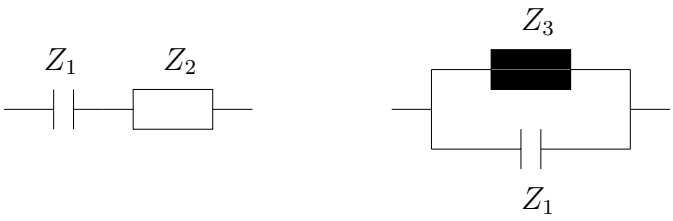
Mit diesem Trick gehorchen die am Bauelement anliegenden Größen $I^c(t) = J_0\exp(i(wt+\alpha))$ und $U(t) = U_0\exp(i(wt+\beta))$ nämlich *einheitlich* dem Ohmschen Gesetz

$$U^c(t) = ZI^c(t)$$

wobei $Z \in \mathbb{C}$ der sogenannte komplexe Widerstand des Bauelements ist. Für die genannten Bauelemente gelten folgende Zusammenhänge mit den reellen Kenngrößen

Symbol	Kenngröße	komplexer Widerstand
	Widerstand R	$Z = R$
	Kapazität C	$Z = \frac{1}{i\omega C}$
	Induktivität L	$Z = i\omega L$

Außerdem gelten für die Grundverknüpfungen der Bauelemente (Reihen- und Parallelschaltung) die üblichen, von Ohmschen Widerständen bekannten Regeln für den Gesamtwiderstand



$$Z = Z_1 + Z_2 \qquad \frac{1}{Z} = \frac{1}{Z_1} + \frac{1}{Z_3}$$

Beachten Sie aber, daß Addition und Division nun komplexe Operationen sind.
Da sich kompliziertere Netzwerke sukzessiv auf diese Grundverknüpfungen zurückführen lassen, ermöglicht die komplexe Darstellung von Strom, Spannung und Widerständen ein einfaches Berechnen von Gesamtwiderständen allgemeiner Netzwerke.
Ein anderes Beispiel für die vereinfachende Wirkung der komplexen Schreibweise werden wir im Zusammenhang mit der Fourieranalyse kennenlernen. Die Grundaussage dieser Theorie ist, daß sich jede (vernünftige) periodische Funktion beliebig gut allein durch geschickte Überlagerung von sin- und cos-Funktionen darstellen läßt. Ist $f : \mathbb{R} \to \mathbb{R}$ differenzierbar und periodisch mit Länge 2π, d. h. $f(t + 2\pi) = f(t)$, so gilt

$$f(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos(kt) + b_k \sin(kt))$$

wobei $a_k, b_k \in \mathbb{R}$ geeignet gewählt werden müssen. Praktisch bedeutet dies, daß f durch die Amplituden a_k, b_k zu den Frequenzen $k =$

$1, 2, 3 \ldots$ eindeutig charakterisiert ist, d. h. die Funktion f wird durch abzählbar viele Zahlen beschrieben (und approximativ durch endlich viele, wenn man die Reihe bei einer hohen Frequenz abbricht). Ist f z. B. der von einem Mikrofon gelieferte Spannungsverlauf bei Aufnahme eines Geräusches, so bedeuten große Werte a_k, b_k für hohe k, daß das Geräusch starke hochfrequente Komponenten hat. Sind dagegen alle a_k, b_k ab einem kleinen k praktisch vernachlässigbar, so handelt es sich um ein niederfrequentes Geräusch. Die Fourierdarstellung erlaubt auch, das Signal zu modifizieren. Bestimmte Frequenzen können z. B. abgeschnitten werden (d. h. die entsprechenden Amplituden werden auf Null gesetzt), was zur Konstruktion von Hochpaß-, Tiefpaß- oder Bandpaßfiltern benutzt werden kann. Die notwendigen Arbeiten sind dabei typischerweise einfacher, wenn die Fourierreihe in komplexer Schreibweise benutzt wird. Man betrachtet dabei $f(t)$ einfach als Realteil einer komplexen Reihe

$$f(t) = Re\left(\sum_{k=-\infty}^{+\infty} c_k e^{ikt}\right)$$

wobei die komplexen Amplituden c_k jetzt die Rolle der beiden reellen Zahlen a_k, b_k übernimmt und e^{ikt} für $\sin(kt)$ und $\cos(kt)$ sorgt.
Zum Abschluß dieser keineswegs vollständigen Liste der Anwendungen komplexer Zahlen sei noch erwähnt, daß die Theorie der komplexwertigen Funktionen auf $\mathbb{C}$ zum Lösen von partiellen Differentialgleichungen in der Ebene benutzt werden kann. Der Grund ist, daß Real- und Imaginärteil von komplex-differenzierbaren Funktionen automatisch Lösungen der Potentialgleichung liefern. Da es einfach ist, komplex differenzierbare Funktionen zu konstruieren, führt dies zu einem großen Vorrat an expliziten Lösungen der Potentialgleichung, die dann beispielsweise in der Elektro- oder Magnetostatik und der Theorie der Potentialströmungen in der Strömungsdynamik benutzt werden können.

KAPITEL 3

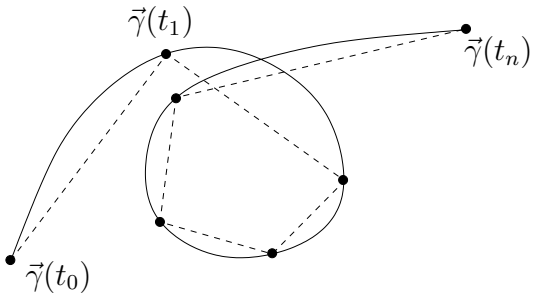
Approximation mit linearen Funktionen

Nachdem wir uns ausführlich mit der Theorie der linearen Funktionen (d. h. den einfachsten proportionalen Zusammenhängen) beschäftigt haben, wollen wir nun allgemeinere Abbildungen untersuchen. Um die nun über uns hereinbrechende Vielfalt zu verstehen und zu ordnen, ist folgender Hinweis wichtig: Im täglichen Leben haben wir es häufig mit Zusammenhängen zu tun, die sich zeitlich bzw. räumlich *vorhersagbar* ändern. Wenn z. B. jetzt die Sonne am Himmel steht, wird sie nicht direkt im nächsten Moment verschwinden, sondern nur ein klein wenig ihre Position ändern. Allgemeiner wird sich jedes bewegte Objekt in einer sehr kurzen Zeit nur beliebig wenig von der Position entfernen, die es zu Beginn des kurzen Zeitintervalls hatte. Auch ändern sich andere physikalische Eigenschaften, wie z. B. die Raumtemperatur vorhersagbar: Bei einer Bewegung um einen Millimeter nach rechts, wird man sich nicht plötzlich den Arm verbrennen, wenn es in der momentanen Position angenehm ist.
Auch kann man beim Tauchen davon ausgehen, daß der Umgebungsdruck bei einer kleinen Tiefenänderung nur wenig wenig zunimmt und nicht plötzlich so stark ansteigt, daß man wie eine Flunder zerquetscht wird. Diese Beispiele zeigen, daß Vorhersagbarkeit geradezu lebensnotwenig ist - zumindest für Leben wir wir es kennen. Es gibt aber auch Bereiche, in denen Unvorhersagbarkeit sehr wichtig ist. Wenn Sie z. B. diesen Text lesen, so nutzen Sie ständig aus, daß das Reflexionsvermögen auf dem Papier sprunghaft wechselt. Stellen Sie sich einen Meßpunkt vor, der über das Blatt wandert und an einer weißen Stelle beginnt. Dort findet er ein hohen Reflexionsvermögen, das über eine gewisse Distanz konstant bleibt. Doch plötzlich, bei einer extrem kleinen Bewegung nach rechts, sinkt der Reflexionskoeffizient rapide ab - der Meßpunkt ist auf einen Buchstaben gestoßen. Dieser Sprung war aus den bisherigen Meßdaten nicht zu erwarten, also nicht *vorhersagber*. Unvorhersagbarkeit spielt also eine wichtige Rolle beim Abgrenzen, Unterscheiden, Einteilen etc. Die Buchstaben können Sie umso besser erkennen, je stärker der Sprung im Reflexionsvermögen ist.

Wäre der Übergang von Weiß nach Schwarz vorhersagbar, so wären die Buchstaben nur schlecht vom Blatt zu trennen und sie würden ineinander verlaufen. Andererseits ist für das Erkennen des Buchstabens E auch wichtig, daß der Reflexionsverlauf in gewisse Richtungen wiederum vorhersagbar ist (horizontal, vertikal).
Die gleichen Überlegungen gelten für andere Informationsübertragungen (Sprache, elektrische Impulse in Nervenbahnen etc.).
Wenn wir uns in Erinnerung rufen, daß die Mathematik zur Beschreibung und Analyse von Zusammenhängen in unserer Umwelt dient, so ist der Begriff der Vorhersagbarkeit also offensichtlich ganz zentral. Eine Funktion, über die wir wissen, daß der Funktionswert sich nur wenig ändert, wenn das Argument wenig variiert (deren Verhalten also vorhersagbar ist), nennen wir *stetig*. Wissen wir sogar, daß das Verhalten bei kleinen Änderungen im Argument fast *linear* ist, so nennen wir die Funktion *differenzierbar*. Das Wort „fast“ bedeutet hierbei, daß man das Funktionsverhalten lokal durch eine lineare Funktion annähern (approximieren) kann, und der Fehler dabei klein ist.
Diese Approximationsidee wird im Folgenden eine sehr wichtige Rolle spielen. Der Begriff der Approximation ist übrigens eng verknüpft mit dem Konzept der *Folge*. Wenn Sie einen schwachen Radiosender optimal einstellen wollen, werden Sie eine Folge von immer feineren Senderknopfbewegungen durchführen. Genauso werden Sie, um Ihre ideale Geschmacksvorstellung von einer Suppe zu approximieren, eine Folge von immer feineren Salzzugaben tätigen.
Ein weiteres Beispiel ergibt sich, wenn Sie die Länge einer gekrümmten Linie messen wollen. Hier ist die naheliegende Idee, die Kurve durch eine stückweise gerade Linie anzunähern, bei der die Längenbestimmung sehr einfach ist. Um das Ergebnis zu verbessern, kann man die geraden Stücke immer weiter verkürzen und erhält so eine Folge von stückweise geraden Kurven, die die gegebene Kurve immer besser approximieren. Mit diesem Beispiel wollen wir uns jetzt genauer beschäftigen.

1. Folgen und Grenzwerte

Durch eine Funktion $\vec{\gamma} : [0, 1] \to \mathbb{R}^2$ sei eine Kurve in der Ebene gegeben, deren Länge bestimmt werden soll. Da wir mit einem Lineal nur gerade Strecken abmessen können, versuchen wir einen Näherungswert für die Länge zu gewinnen, indem wir die Kurve durch stückweise gerade Abschnitte annähern. Dazu wählen wir eine Zerlegung Z^{n+1} von $[0, 1]$, d. h. ein Zahlentupel $Z = (t_0, \ldots, t_n) \in \mathbb{R}^{n+1}$ mit der Eigenschaft $0 = t_0 < t_1 < \ldots < t_n = 1$ und verbinden die Punkte $\vec{\gamma}(t_i)$ durch gerade Linien.



Das Ausmessen dieser Hilfskurve ergibt dann die approximative Länge

$$L(\vec{\gamma}, Z) = \sum_{i=1}^{n} ||\vec{\gamma}(t_i) - \vec{\gamma}(t_{i-1})||_2$$

Um einen genaueren Wert zu vermitteln, muß man dann nur die Zerlegung verfeinern, d. h. mehr Stützpunkte entlang der Kurve hinzufügen. Genauer nennen wir eine Zerlegung $Z' = (s_0, \ldots, s_m)$ feiner als $Z = (t_0, \ldots, t_n)$, falls

$$\{t_0, \ldots, t_n\} \subset \{s_0, \ldots, s_m\}.$$

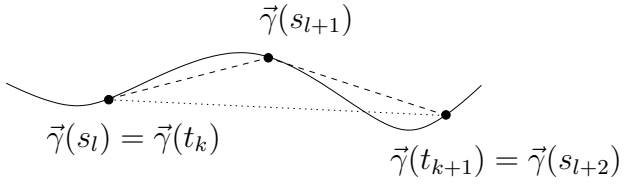
Als Beispiel sei die Folge

$$Z_n = \frac{1}{2^n}(0, 1, 2, \ldots, 2^n) \qquad n \in \mathbb{N}_0$$

erwähnt, die aus sukzessive feineren Zerlegungen besteht.

$$Z_0 = (0, 1), \quad Z_1 = (0, \frac{1}{2}, 1), \quad Z_2 = (0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1), \ldots$$

Allgemein beobachtet man beim Übergang von einer Zerlegung Z zu einer Verfeinerung Z', daß die approximative Länge nie kleiner werden kann. Liegt etwa zwischen den Stützpunkten $s_l = t_k$ und $s_{l+2} = t_{k+1}$ ein weiterer Punkt s_{l+1} der feineren Zerlegung Z', so wird durch den zusätzlichen Zwischenstopp die ursprüngliche Verbindung $\vec{\gamma}(t_{k+1}) - \vec{\gamma}(t_k)$ durch zwei Streckenabschnitte $\vec{\gamma}(s_{l+1}) - \vec{\gamma}(s_l)$ und $\vec{\gamma}(s_{l+2}) - \vec{\gamma}(s_{l+1})$ ersetzt.



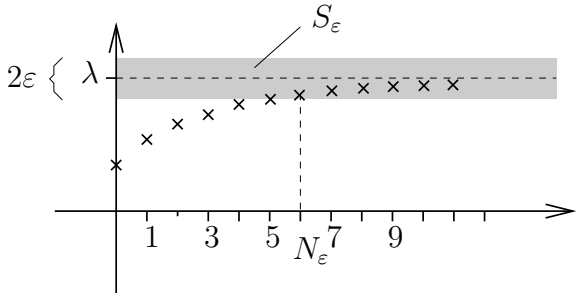
Die Dreiecksungleichung ergibt dann

$$\begin{aligned} ||\vec{\gamma}(t_{k+1}) - \vec{\gamma}(t_k)||_2 &= ||\vec{\gamma}(s_{l+2}) - \vec{\gamma}(s_{l+1}) + \vec{\gamma}(s_{l+1}) - \vec{\gamma}(s_l)||_2 \\ &\leq ||\vec{\gamma}(s_{l+2}) - \vec{\gamma}(s_{l+1})||_2 + ||\vec{\gamma}(s_{l+1}) - \vec{\gamma}(s_l)||_2 \end{aligned}$$

und somit

$$L(\vec{\gamma}, Z) \leq L(\vec{\gamma}, Z').$$

Benutzen wir die Verfeinerungsfolge Z_n, so ergibt sich also eine Folge $l : N_0 \to \mathbb{R}, l(n) = L(\vec{\gamma}, Z_n)$. Zur Erinnerung: Eine Folge ist eine Funktion auf einer abzählbar unendlichen Definitionsmenge. Die Funktionswerte bezeichnet man auch als $l_n = l(n)$ und die gesamte Funktion auch als $(l_n)_{n\in\mathbb{N}_0}$. Aufgrund unserer Konstruktion approximieren die Werte l_n die wahre Länge der Kurve um so besser, je größer wir den Index n wählen. Aber was ist eigentlich die wahre Länge der Kurve? Intuitiv ist es die Zahl, der wir bei feiner werdenden Messungen immer näher kommen, d. h. wir werden eine Zahl $\lambda \in \mathbb{R}$ nur dann als Längenwert zulassen, wenn für jeden noch so kleinen Toleranzwert $\varepsilon > 0$, die feinsten Meßwerte l_n mit sehr großen n *alle* im Toleranzbereich $(\lambda - \varepsilon, \lambda + \varepsilon)$ liegen. Anschaulich kann man diese Forderung an dem Graph der Folge l verdeutlichen. Da der Definitionsbereich von l durch die diskrete Menge $\mathbb{N}_0$ gegeben ist, besteht der Graph $G_l = \{(n, l(n)) : n \in \mathbb{N}_0\}$ aus isolierten Punkten. Der Toleranzbereich ist in dieser Darstellung durch den Streifen $S_\varepsilon(\lambda) = \{(x, y)|\lambda - \varepsilon < y < \lambda + \varepsilon, x \in \mathbb{R}\}$ der Breite 2ε um $y = \lambda$ gekennzeichnet.



Der intuitive Längenwert λ ist also die Zahl, für die jeder Streifen $S_\varepsilon(\lambda), \varepsilon > 0$ alle bis auf endlich viele Punkte des Graphen enthält. Anders ausgedrückt, muß es zu jedem noch so schmalen Streifen $S_\varepsilon(\lambda), \varepsilon > 0$ immer eine Nummer N_ε geben, ab der die Punkte $(n, l(n)), n \geq N_\varepsilon$ des Graphen komplett im Streifen liegen. Die Überpfügung mit *beliebig kleinen* Streifenbreiten $\varepsilon > 0$ ist hierbei wichtig. Würden wir uns

auf eine bestimmte Streifenbreite $\varepsilon = \varepsilon_0 > 0$ bzw. auf $\varepsilon \geq \varepsilon_0 > 0$ einschränken, so könnten wir natürlich aus der Feststellung, daß irgendwann (d. h. ab einem bestimmten N_ε) der Graph im ε-Streifen bleibt nur schließen, daß λ von den Meßwerten höchstens um $\varepsilon_0 > 0$ abweicht, was ja nicht ausschließen würde, daß z. B. $\lambda + \frac{\varepsilon_0}{2}$ ein besserer Kandidat für die Länge wäre als λ. Überprüfen wir den Graphen aber auch mit dem Streifen $S_{\frac{\varepsilon_0}{4}}(\lambda)$, so können wir dagegen in dieser Frage Klarheit schaffen: Liegt der Graph irgendwann ganz in diesem Streifen, so ist $\lambda + \frac{\varepsilon_0}{2}$ sicherlich als Längenwert ausgeschieden. Aber jetzt könnte doch $\lambda + \frac{\varepsilon_0}{8}$ vielleicht der bessere Längenwert sein? Na, dann überprüfen wir das halt mit dem Streifen $S_{\frac{\varepsilon_0}{16}}(\lambda)$ und so weiter und so weiter. Sie sehen, um Klarheit zu schaffen, ob λ der geeignete Längenwert ist, müssen wir für *beliebig* schmale Streifen nachprüfen, ob irgendwann der hintere Teil des Graphen ganz im Streifen bleibt. Nach diesen Vorüberlegungen verstehen Sie jetzt sicherlich folgende

Definition 9. *Sei $a : \mathbb{N} \to \mathbb{R}$ eine Folge. Dann heißt $A \in \mathbb{R}$ Grenzwert der Folge (bzw. man sagt a_n konvergiert gegen A, oder strebt gegen A und schreibt $\lim\limits_{n\to\infty} a_n = A$ bzw. $a_n \xrightarrow[n\to\infty]{} A$), wenn für jeden Toleranzwert $\varepsilon > 0$ ein $N_\varepsilon \in \mathbb{N}$ existiert, so daß der N_ε-Rest der Folge $a_{N_\varepsilon}, a_{N_\varepsilon+1}, a_{N_\varepsilon+2} \ldots$ im Toleranzbereich $(\lambda - \varepsilon, \lambda + \varepsilon)$ liegt, also falls*

$$|A - a_n| < \varepsilon \qquad \textit{für alle} \qquad n \geq N_\varepsilon$$

Der Übersichtlichkeit halber ist die Definition für Folgen mit Definitionsbereich $\mathbb{N}$ angegeben. In analoger Weise definieren wir den Grenzwert aber auch für andere abzählbare Definitionsmengen wie z. B. $\mathbb{N}_0$ oder $\{n \in \mathbb{Z} | n \geq n_0\}$.
Nach dieser begrifflichen Klärung wollen wir jetzt für unser Beispiel nachrechnen, ob die Längenfolge l einen Grenzwert besitzt. Da die Folge monoton wächst, ist der größte Wert sicher ein heißer Kandidat für den Grenzwert. Allerdings müssen wir hier etwas vorsichtig sein. Die Kurve könnte ja unbeschränkt lang sein (eine seltsame Spirale zum Beispiel), d. h. wir werden wohl zunächst fordern müssen, daß die Längen nicht zu groß werden. Die Wertemenge $l(\mathbb{N}_0)$ sollte also zumindest beschränkt sein. Außerdem kann es auch wachsende beschränkte Folgen geben, die niemals ihren größten Wert annehmen (denken Sie etwa an die Folge $a_n = 1 - \frac{1}{n}$, die ja dem Wert 1 immer näher kommt, ihn aber nie erreicht). Als Ersatz für das fehlende Maximum können wir aber die kleinste obere Schranke $\sup l(\mathbb{N}_0)$ der Menge nehmen. Das ist ja ein Wert, der größer als alle l_n ist, aber trotzdem so dicht an $l(\mathbb{N}_0)$ liegt, daß kein kleinerer Wert existiert, an dem die l_n nicht irgendwann (für

großes n) vorbeilaufen. Nennen wir also $\lambda = \sup l(\mathbb{N}_0)$ und überprüfen nun, ob λ im obigen Sinne der Grenzwert der Folge ist. Dazu müssen wir zu beliebigem vorgegebenem $\varepsilon > 0$ zeigen, daß ab einer bestimmten Nummer alle l_n im Intervall $(\lambda - \varepsilon, \lambda + \varepsilon)$ liegen. Nehmen wir dazu einmal an, *kein* l_n würde das Intervall je betreten. Dann wäre aber auch $l_n \leq \lambda - \varepsilon$ für alle n, denn λ ist ja eine obere Schranke, d. h. alle l_n müssen links von λ liegen und wenn $(\lambda - \varepsilon, \lambda)$ keine Werte enthält, dann sind sie sogar links von $\lambda - \varepsilon$. Da $\varepsilon > 0$ ist, hätten wir aber eine kleinere obere Schranke als die kleinste obere Schranke λ gefunden – ein offensichtlicher Widerspruch. Die Annahme, daß sich nie ein l_n in das Intervall $(\lambda - \varepsilon, \lambda + \varepsilon)$ verirrt, ist also falsch, d. h. es gibt ein N_ε, so daß $l_{N_\varepsilon} \in (\lambda - \varepsilon, \lambda + \varepsilon)$. Nun wissen wir aber, daß l_n monoton wächst und durch λ beschränkt ist, also $\lambda - \varepsilon \leq l_n \leq \lambda$ für alle $n \geq N_\varepsilon$.
Bei der obigen Argumentation haben Sie wohl bemerkt, daß wir nur zwei Eigenschaften der Folge ausgenutzt haben. Erstens haben wir angenommen, daß die Folge beschränkt ist und zweitens haben wir genutzt, daß die Folge monoton ist. Letztlich haben wir damit folgende Aussage bewiesen.

Satz 4. *Jede monoton wachsende (fallende) und nach oben (unten) beschränkte reellwertige Folge hat einen Grenzwert. Der Grenzwert ist das Supremum (Infimum) der Folgenwerte.*

Ohne die Kontinuumeigenschaft der reellen Zahlen hätten wir an dieser Stelle schlechte Karten. Wir könnten nämlich die Länge der Kurve unter Umständen nicht durch eine Zahl beschreiben, wenn die Meßwerte gegen eine „Lücke“ streben würden. Mit der Existenz des Infimums und Supremums von beschränkten Mengen in $\mathbb{R}$ ist aber für geeignete Grenzwerte von monotonen Folgen gesorgt.
Einige weitere Beispiele von Folgen, über deren Grenzwert wir mit dem obigen Satz eine Aussage machen können, sind

$$a_n = \frac{1}{n}, \quad b_n = \frac{1}{2^n}, \quad c_n = \frac{1}{\sqrt{n}}$$

die jeweils monoton fallend sind und Null als größte untere Schranke und damit auch als Grenzwert besitzen.
Mit unserer Meßmethode sind wir nun in der Lage, Kurven einen Längenwert zuzuordnen. Bei glatten, kontinuierlichen Kurven ist dieser Längenwert wohl genau die Zahl, auf die sich die meisten Menschen einigen würden. Bei Kurven, die springen, muß das aber nicht mehr so sein. Betrachten wir dazu folgende Funktion, die bei $t = \frac{1}{2}$ „springt“

$$\vec{\gamma}(t) = \begin{cases} (t, t^2) & t \in \left[0, \frac{1}{2}\right] \\ \left(t - \frac{1}{4}, \frac{1}{4}\right) & t \in \left(\frac{1}{2}, 1\right] \end{cases}$$

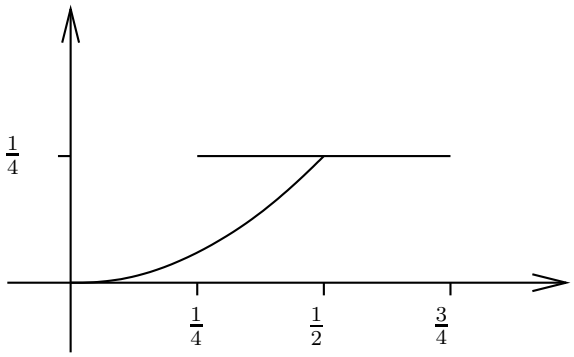


Abbildung 1: Wertebereich der Funktion $\vec{\gamma}$

Wenden wir auf diese Kurve unser Meßverfahren an, so wird schnell klar, daß $L(\vec{\gamma}, Z_n)$ die Sprungweite von $\left(\frac{1}{2}, \frac{1}{2}\right)$ nach $\left(\frac{1}{4}, \frac{1}{2}\right)$ mitmißt. Ob der so ermittelte Längenwert sinnvoll ist, hängt nun stark von der Anwendung ab, um die es geht. Für eine Baufirma, die verschiedene Straßenabschnitte teeren soll, ist es offensichtlich nicht sinnvoll, den Abstand zwischen zwei Straßen (d. h. die Sprungweite) für den Teerbedarf mit einzuplanen. Geht es bei der Anwendung aber um die Auswertung der zurückgelegten Entfernung eines Insekts in einem gewissen Zeitintervall (biologisches Experiment), bei dem der durch eine Kamera aufgenommene Weg unterbrochen ist (Insekt krabbelt unter ein Blatt etc.), so ist es sinnvoll, *mindestens* die direkte Verbindung zwischen den Wegabschnitten hinzuzurechnen. An diesen Beispielen sehen Sie, daß bei Sprüngen im Kurvenverlauf auf jeden Fall Interpretationsbedarf bezüglich der Zuordnung einer Länge besteht. Außerdem erkennen Sie auch im Hinblick auf die in der Einleitung angegebenen Beispiele, daß sprunghaftes Verhalten einer Funktion ein wichtiges Merkmal ist. Diesen Aspekt wollen wir uns im nächsten Abschnitt genauer ansehen.

2. Stetige Funktionen

Als Beispiel einer „springenden“ Funktion haben wir im letzten Abschnitt die Kurve $\vec{\gamma}$ betrachtet. Am Wertebereich (Abbildung 1) kann man den Sprung von $\vec{\gamma}$ bei $t = \frac{1}{2}$ allerdings gar nicht richtig erkennen, da die beiden Äste der Kurve sich im Punkt $\left(\frac{1}{2}, \frac{1}{4}\right)$ berühren. Deutlicher

wird der Sprung dagegen am Graph der Funktion, der als Raumkurve gezeichnet werden kann.

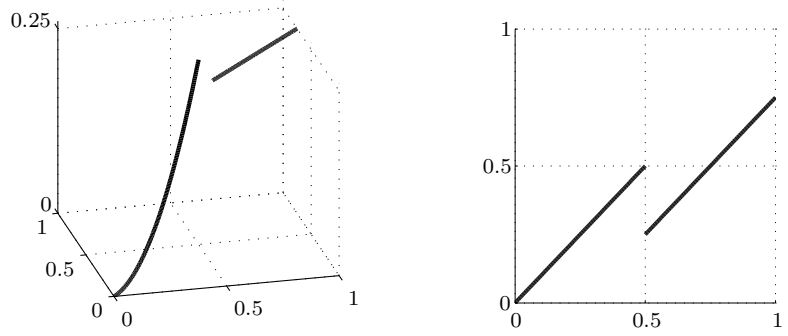


Abbildung 2: Graph der Funktion $\vec{\gamma}$. Die t-Achse läuft jeweils von links nach rechts. Das untere Bild zeigt den Blick auf den Graph von oben

In der Draufsicht sieht das Auge eine „Lücke“ bzw. eine Lückenweite $\varepsilon_0 > 0$, um die sich Funktionswerte beliebig nahe bei $t_0 = \frac{1}{2}$ vom Wert in t_0 unterscheiden. Anders ausgedrückt: Variiert man die Argumente in einem winzigen Intervall $(t_0-\delta, t_0+\delta)$ um t_0, so kann der Funktionswert trotzdem vergleichsweise riesige Änderungen der Größe ε_0 ausführen. In mathematischer Sprechweise ist ein Sprung an der Stelle t_0 also dadurch charakterisiert, daß ein $\varepsilon_0 > 0$ existiert, so daß für *jedes* noch so kleine $\delta > 0$ ein $t \in (t_0 - \delta, t_0 + \delta)$ existiert mit der Eigenschaft $||\vec{\gamma}(t) - \vec{\gamma}(t_0)|| \geq \varepsilon_0$.
Umgekehrt liegt an der Stelle t_0 kein Sprung vor, wenn wir eben kein solches ε_0 finden können. Hier ist nun etwas Scharfsinn gefragt, um die Sprungbedingung zu negieren. Die Idee ist die: Nehmen Sie einmal ein $\varepsilon_0 > 0$ her. Wenn kein Sprung vorliegt, können wir eben *nicht* für jedes noch so kleine $\delta > 0$ einen Punkt $t \in (t_0 - \delta, t_0 + \delta)$ finden, für den der Funktionswert $\vec{\gamma}(t)$ um mehr als ε_0 in der Norm von $\vec{\gamma}(t_0)$ abweicht. Das bedeutet aber doch, daß ab einem bestimmten $\delta_0 > 0$ die Funktion in $(t_0 - \delta, t_0 + \delta)$ mit $\delta \leq \delta_0$ nicht um mehr als ε_0 schwanken kann bezüglich ihrem Wert im Zentrum t_0 des Intervalls. Schreiben wir diese Bedingung auf, so erhalten wir die sogenannte $\varepsilon - \delta$-Definition der Stetigkeit. Wir formulieren die Definition direkt für Funktionen zwischen allgemeinen Vektorräumen.

Definition 10. *Seien V, W normierte Vektorräume und $D \subset V$. Eine Funktion $f : D \to W$ heißt stetig in $\vec{v}_0 \in D$, wenn für jedes $\varepsilon > 0$ ein passendes $\delta > 0$ existiert, so daß stets $||f(\vec{v}) - f(\vec{v}_0)||_w < \varepsilon$ gilt, falls $||\vec{v} - \vec{v}_0||_v < \delta$ und $\vec{v} \in D$ ist.*
Die Funktion f heißt stetig, wenn f für alle $\vec{v}_0 \in D$ stetig ist.

Wenn Sie mit dieser Definition die Stetigkeit einer Funktion nachprüfen wollen, müssen Sie so vorgehen: Der Wert ε wird Ihnen vorgegeben. Sie wissen von ihm nur, daß er strikt positiv ist. Ihre Aufgabe besteht darin, einen Wert $\delta > 0$ zu finden, so daß für jeden Punkt $\vec{v}$ der höchstens den Abstand δ von $\vec{v}_0$ hat, der zugehörige Funktionswert $f(\vec{v})$ höchstens um ε vom Wert $f(\vec{v}_0)$ in der W-Norm abweicht. Die von Ihnen zu findende Zahl δ wird dabei natürlich von ε abhängen: Typischerweise muß δ um so kleiner gewählt werden, je kleiner die Vorgabe $\varepsilon > 0$ ausfällt. Außerdem wird δ von der „Steilheit“ der Funktion an der Stelle $\vec{v}_0$ beeinflußt. Ändert sich die Funktion bei $\vec{v}_0$ stark, dann heißt das ja gerade, daß kleine Variationen im Argument große Variationen im Funktionswert zur Folge haben. In solch einem Fall wird δ deutlich kleiner sein müssen als ε. Umgekehrt kann δ auch größer als ε ausfallen, wenn die Funktion bei $\vec{v}_0$ sehr flach ist. Dann schwankt ja der Funktionswert sowieso nicht stark, selbst wenn wir deutlich am Argument wackeln. Beachten Sie aber, daß die Definition nicht die Kenntnis des maximalen Radius $\delta > 0$ verlangt, für die die Funktionswerte gerade noch um weniger als ε variieren. Jeder kleinere Radius erfüllt natürlich auch den Zweck! Sie können also großzügig beim Abschätzen sein. Wenn Sie durch das von Ihnen gefundene $\delta > 0$ sicherstellen, daß der Funktionswert sogar um weniger als $\frac{\varepsilon}{1000}$ variiert, haben Sie Ihre Aufgabe auch erfüllt, und wenn die maximale Variation nur $\frac{\varepsilon}{100000}$ ist, so ist das auch ok. Prinzipiell laufen Stetigkeitsuntersuchungen immer nach demselben Muster ab. Das Ziel ist die Variation der Funktionswerte $||f(\vec{v}) - f(\vec{v}_0)||_W$ durch die Variation der Argumente $||\vec{v} - \vec{v}_0||_V$ abzuschätzen, also etwa in der Form

$$||f(\vec{v}) - f(\vec{v}_0)||_W \leq C_{\vec{v}_0} ||\vec{v} - \vec{v}_0||_V \tag{17}$$

wobei die Konstante $C_{\vec{v}_0} > 0$ nur von $\vec{v}_0$, nicht aber von $\vec{v}$ abhängt. Hat man solch eine Abschätzung gefunden, so ist die Stetigkeit gezeigt. Eine vorgegebene Variationsweite $\varepsilon > 0$ ist dann nämlich sichergestellt, wenn $||\vec{v} - \vec{v}_0||_V$ kleiner als $\delta = \varepsilon / C_{\vec{v}_0}$ ist. Die Form der Abschätzung (17) ist dabei nicht zwingend. Wenn man zum Beispiel nur einen Zusammenhang

$$||f(\vec{v}) - f(\vec{v}_0)||_W \leq \sqrt{C_{\vec{v}_0}||\vec{v} - \vec{v}_0||_V} \tag{18}$$

herleiten kann, so muß die Schwankungsbreite des Arguments auf $\delta = \varepsilon^2/C_{\vec{v}_0}$ eingeschränkt werden, damit der Funktionswert um höchstens ε variiert. Für sehr kleine ε ist in diesem Fall δ deutlich keiner als ε. Das ist nötig, weil die Wurzelfunktion $x \to \sqrt{x}$ bei $x = 0$ sehr steil ist. Mit einer ähnlichen Argumentation sieht man, daß alle Abschätzungen der Form

$$||f(\vec{v}) - f(\vec{v}_0)||_W \leq (C_{\vec{v}_0}||\vec{v} - \vec{v}_0||_V)^\alpha \tag{19}$$

mit $0 < \alpha \leq 1$ benutzt werden können, um die Stetigkeit von f nachzuweisen. Der Spezialfall $\alpha = 1$ ist dabei gerade (17). Funktionen f, die sich so abschätzen lassen, nennt man *Lipschitz-stetig* in $\vec{v}_0$. Im Fall $\alpha < 1$ spricht man von Hölder-stetigen Funktionen (beachten Sie, daß $\alpha = \frac{1}{2}$ gerade (18) entspricht).
In seltenen Fällen trifft man auch Funktionen, die stetig sind, aber für kein $\alpha > 0$ einer Abschätzung der Form (19) genügen. Dann muß man sich typischerweise etwas mehr anstrengen, um die Stetigkeit nachzuweisen.
Schauen wir uns jetzt aber einmal einige stetige Funktionen an. Das erste Beispiel handelt von einem sehr langweiligen Zusammenhang, der konstanten Funktion. Sind V, W normierte Vektorräume und $\vec{w} \in W$, so ist durch $f(\vec{v}) = \vec{w}$ für alle $\vec{v} \in V$ die Funktion mit dem konstanten Wert $\vec{w}$ gegeben (das krasse Gegenteil eines unvorhersagbaren Zusammenhangs). Die Variation der Funktionswerte ist hier gleich Null und damit sehr einfach durch die Variation der Argumente abzuschätzen

$$||f(\vec{v}) - f(\vec{v}_0)||_W = ||\vec{w} - \vec{w}||_W = 0 \leq ||\vec{v} - \vec{v}_0||_V$$

f ist also Lipschitz-stetig (hier mit Konstante $C_{\vec{v}_0} = 1$ aber jede andere positive Konstante wäre auch möglich gewesen) und damit stetig. Nochmal: Die Variation bleibt kleiner als jedes vorgegebene $\varepsilon > 0$, wenn die Argumente höchstens um $\delta = \varepsilon/C_{\vec{v}_0}$ von $\vec{v}_0$ abweichen. Konstante Funktionen sind also, wie zu erwarten war, stetig. Ein weiteres einfaches Beispiel ist die Identität $f(\vec{v}) = (\vec{v})$ auf V. Hier gilt

$$||f(\vec{v}) - f(\vec{v}_0)||_V = ||\vec{v} - \vec{v}_0||_V$$

d. h. die Identität ist Lipschitz-stetig mit Konstante Eins. Das nächste Beispiel ist die Normfunktion $||\cdot||_V : V \to \mathbb{R}$. Der Zielraum W ist hier der reelle Vektorraum $\mathbb{R}$ mit dem Betrag als Norm $||\cdot||_W = |\cdot|$. Vergleichen wir zwei Funktionswerte, so folgt mit der Dreiecksungleichung

$$\begin{aligned} ||\vec{v}||_V - ||\vec{v}_0||_V &= ||\vec{v} - \vec{v}_0 + \vec{v}_0||_V - ||\vec{v}_0||_V \\ &\leq ||\vec{v} - \vec{v}_0||_V + ||\vec{v}_0||_V - ||\vec{v}_0||_V = ||\vec{v} - \vec{v}_0||_V \end{aligned}$$

Die gleiche Argumentation mit vertauschten Rollen liefert $||\vec{v}_0|| - ||\vec{v}||_V \leq ||\vec{v}_0 - \vec{v}||_V$ und da $||\vec{v} - \vec{v}_0||_V = ||\vec{v}_0 - \vec{v}||_V$ ergibt sich

$$-||\vec{v} - \vec{v}_0||_V \leq ||\vec{v}||_V - ||\vec{v}_0||_V \leq ||\vec{v} - \vec{v}_0||_V$$

oder, mit anderen Worten

$$|| \, ||\vec{v}||_V - ||\vec{v}_0||_V \, ||_W = | \, ||\vec{v}||_V - ||\vec{v}_0||_V \, | \leq ||\vec{v} - \vec{v}_0||_V$$

Also auch hier finden wir Lipschitz-Stetigkeit, d. h. $|| \cdot ||_V$ ist stetig auf V.

Beachten Sie, daß unsere allgemeine Betrachtungsweise den Fall $V = \mathbb{R}$ mit Norm $|| \cdot ||_V = | \cdot |$ beinhaltet. Wir haben also auch gezeigt, daß die Betragsfunktion $| \cdot | : \mathbb{R} \to \mathbb{R}$ eine stetige Funktion ist. Mit dem gleichen Beweis haben wir aber auch festgestellt, daß sich die Länge eines Zeigers nicht sprunghaft ändert, wenn Sie, bei festgehaltenem Endpunkt, die Spitze ein wenig verschieben. Dieses Ergebnis sollte intuitiv klar sein. Umso bedeutender ist deshalb unsere Beobachtung, daß die „Längenfunktion“ $|| \cdot ||$ in *jedem* normierten Vektorraum nicht sprunghaft ist, denn wir können unsere Intuition im geometrischen dreidimensionalen Anschauungsraum in dieser Hinsicht auf *alle Vektorräume* übertragen, ohne dabei Fehler zu machen.

Im nächsten Schritt schauen wir uns die Vektorraumoperationen Addition und skalare Multiplikation an. Die Addition können wir dabei als Funktion auf $V \times V$ auffassen, die jedem Vektorpaar $(\vec{x}, \vec{y})$ die Summe $f(\vec{x}, \vec{y}) = \vec{x} + \vec{y}$ zuordnet. Um im Rahmen unserer Definition von Stetigkeit zu bleiben, müssen wir dazu die Menge $V \times V$ mit einer Vektorraumstruktur versehen. Die naheliegende Variante für eine Addition in $V \times V$ ist dabei

$$(\vec{x}_1, \vec{y}_1) + (\vec{x}_2, \vec{y}_2) = (\vec{x}_1 + \vec{x}_2, \vec{y}_1 + \vec{y}_2)$$

und für die skalare Multiplikation

$$\alpha(\vec{x}, \vec{y}) = (\alpha\vec{x}, \alpha\vec{y}) \quad \alpha \in \mathbb{R}$$

Als Norm auf $V \times V$ wählen wir

$$||(\vec{x}, \vec{y})||_{V \times V} = \max\{||\vec{x}||_V, ||\vec{y}||_V\}$$

Der Nachweis der Vektorraum- und Normeigenschaften ist dabei eine gute Übung, die Sie sich nicht entgehen lassen sollten. Mit entsprechenden Operationen kann man übrigens auch allgemeinere Produktmengen $V \times W$ aus zwei normierten Vektorräumen V, W zu einem Vektorraum machen. Schauen wir uns aber nun die Additionsfunktion $f : V \times V \to V$ an einer Stelle $(\vec{x}_0, \vec{y}_0) \in V \times V$ genauer an.

$$\begin{aligned} ||f(\vec{x},\vec{y}) - f(\vec{x}_0,\vec{y}_0)||_V &= ||\vec{x}+\vec{y}-(\vec{x}_0+\vec{y}_0)||_V \\ &\leq ||\vec{x}-\vec{x}_0||_V + ||\vec{y}-\vec{y}_0||_V \end{aligned}$$

Um die (Lipschitz)-Stetigkeit zu zeigen, müssen wir die rechte Seite nur noch durch die Argumentdifferenz

$$||(\vec{x},\vec{y}) - (\vec{x}_0,\vec{y}_0)||_{V\times V} = \max\{||\vec{x}-\vec{x}_0||_V, ||\vec{y}-\vec{y}_0||_V\}$$

abschätzen. Offensichtlich ist jeder einzelne der beiden Terme durch die Norm der Argumentdifferenz beschränkt, so daß

$$||f(\vec{x},\vec{y}) - f(\vec{x}_0,\vec{y}_0)||_V \leq 2\, ||(\vec{x},\vec{y}) - (\vec{x}_0,\vec{y}_0)||_{V\times V}$$

d. h. die Addition ist Lipschitz-stetig und damit auch stetig. Im Fall der skalaren Multiplikation haben wir die Funktion $f : \mathbb{R} \times V \to V$ zu untersuchen, mit $f(\alpha, \vec{x}) = \alpha\vec{x}$. Auch hier ist die Definitionsmenge wieder ein Produkt-Vektorraum, wobei wir Addition, skalare Multiplikation und Norm wie oben definieren. Untersuchen wir die Variation der Funktion

$$\begin{aligned} ||f(\alpha,\vec{x}) - f(\alpha_0,\vec{x}_0)||_V &= ||\alpha\vec{x} - \alpha_0\vec{x}_0||_V \\ &= ||(\alpha-\alpha_0)\vec{x} + \alpha_0(\vec{x}-\vec{x}_0)||_V \end{aligned}$$

Hier haben wir *den* Standardtrick für die Abschätzung von Produktdifferenzen benutzt. Durch das Zwischenschieben des Terms $\alpha_0\vec{x} - \alpha_0\vec{x}$, ist es möglich, von der Differenz der Produkte auf Produkte mit Differenzen zu kommen. Mit der Dreiecksungleichung schätzen wir weiter ab

$$\begin{aligned} ||f(\alpha,\vec{x}) - f(\alpha_0,\vec{x}_0)||_V &\leq |\alpha-\alpha_0|\, ||\vec{x}||_V + |\alpha_0|\, ||\vec{x}-\vec{x}_0||_V \\ &\leq (||\vec{x}||_V + |\alpha_0|)||(\alpha,\vec{x}) - (\alpha_0,\vec{x}_0)||_{\mathbb{R}\times V} \end{aligned}$$

wobei sowohl $|\alpha - \alpha_0|$ als auch $||\vec{x} - \vec{x}_0||_V$ durch die Norm auf $\mathbb{R} \times V$

$$||(\alpha,\vec{x}) - (\alpha_0,\vec{x}_0)||_{\mathbb{R}\times V} = \max\{|\alpha-\alpha_0|, ||\vec{x}-\vec{x}_0||_V\}$$

abgeschätzt werden können. An dieser Stelle ist nun Vorsicht geboten. Wir können *nicht* schließen, daß f Lipschitz stetig ist, da der Vorfaktor vor der Norm der Argumentdifferenz keine Konstante ist. Der Faktor hängt noch von $\vec{x}$ ab und wird für große $\vec{x}$ beliebig groß! Was nun? Besinnen wir uns auf die Definition der Stetigkeit zurück. Bei der Stetigkeitsüberprüfung wird uns eine maximale Variation $\varepsilon > 0$ der Funktionswerte vorgegeben und wir müssen ein passendes $\delta > 0$ finden, so daß die Funktionswerte um weniger als ε variieren, sofern die Argumente um weniger als δ schwanken. Es liegt damit in unserem Ermessen, die Variation des Arguments einzuschränken. Wenn wir z. B. nur Abweichungen

$$||(\alpha, \vec{x}) - (\alpha_0, \vec{x}_0)||_{\mathbb{R}\times V} < 1 \tag{20}$$

zulassen, so kann $||\vec{x}-\vec{x}_0||_V$ nicht größer als Eins und folglich $||\vec{x}||_V$ nie viel größer als $||\vec{x}_0||_V$ werden. Details regelt wie immer die Dreiecksungleichung

$$||\vec{x}||_V = ||\vec{x} - \vec{x}_0 + \vec{x}_0||_V \leq ||\vec{x} - \vec{x}_0||_V + ||\vec{x}_0||_V \leq 1||\vec{x}_0||_V + 1$$

wobei wir 20 benutzt haben. Insgesamt gilt also unter dieser Prämisse

$$||f(\alpha, \vec{x}) - f(\alpha_0, \vec{x}_0)||_V \leq (|\alpha_0| + ||\vec{x}_0||_V + 1)||(\alpha, \vec{x}) - (\alpha_0, \vec{x})||_{\mathbb{R}\times V}$$

Wenn wir also

$$||(\alpha, \vec{x}) - (\alpha_0, \vec{x}_0)||_{\mathbb{R}\times V} < \delta = \min\left\{1, \frac{\varepsilon}{|\alpha_0| + ||\vec{x}_0||_V + 1}\right\}$$

wählen, so unterbieten wir mit der maximalen Funktionswertschwankung den vorgegebenen Wert ε, und die skalare Multiplikation hat sich somit als stetige Funktion geoutet.
Beachten Sie, daß wir auch wieder etwas über den Spezialfall $V = \mathbb{R}$ dazugelernt haben. In diesem Fall entspricht das Skalarprodukt gerade der gewöhnlichen Multiplikation $(x, y) \to xy$, bei der zwei reellen Zahlen x, y ihr Produkt xy zu geordnet wird. Durch Anwendung des Multiplikationstricks zur Verwandlung von Produktdifferenzen in Differenzpunkte kann man auch zeigen, daß jedes Skalarprodukt $\langle\cdot,\cdot\rangle : V : V \to \mathbb{R}$ bezüglich der Norm

$$||(\vec{v}, \vec{y})||_{V\times V} = \max\left\{\sqrt{\langle\vec{x}, \vec{x}\rangle}, \sqrt{\langle\vec{y}, \vec{y}\rangle}\right\}$$

auf $V \times V$ und $|\cdot|$ auf $\mathbb{R}$ stetig ist. Als Zutat muß man hier die Cauchy-Schwarz-Ungleichung benutzen.

Damit haben wir gesehen, daß alle Vektrorraumoperationen auf normierten Räumen stetig sind. Neben diesen Grundoperationen sind fast alle „Taschenrechnerfunktionen“ stetig. Dazu zählen wir Funktionen $f : D \to \mathbb{R}$ aus folgender Liste

D	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R} \setminus \{0\}$	$[0, \infty)$	$\mathbb{R}$	$\mathbb{R}$	$(0, \infty)$	$\mathbb{R}$	$\mathbb{R}$
$f(x)$	1	x	$\lvert x \rvert$	$\frac{1}{x}$	$\sqrt[2n]{x}$	$\sqrt[2n+1]{x}$	exp	ln	sin	cos

Die Stetigkeit der ersten drei Einträge haben wir bereits im allgemeinen Rahmen untersucht. Für $f(x) = \frac{1}{x}$ schätzen wir die Variation der Funktionswerte folgendermaßen ab

$$|f(x) - f(x_0)| = \left| \frac{1}{x} - \frac{1}{x_0} \right| = \left| \frac{x_0 - x}{x_0 x} \right| = \frac{1}{|x_0 x|} |x_0 - x|$$

Beachten Sie, daß hieraus noch *nicht* folgt, daß f Lipschitz-stetig ist, denn der Faktor vor $|x - x_0|$ hängt noch von x ab und kann beliebig groß werden, wenn x nahe an Null kommt. Da Stetigkeit in $x_0 \neq 0$ sich aber auf das Verhalten der Funktionswerte nahe bei x_0 bezieht, können wir diese Gefahr leicht ausschließen. Wir verlangen einfach, daß x von x_0 nicht mehr als $|x_0|/2$ abweichen darf. Dann kann x offensichtlich nur noch bis auf $|x_0|/2$ an die Null herankommen. Wieder konkretisiert die Dreiecksungleichung diese Überlegung, denn

$$|x_0| \leq |x_0 - x| + |x| \leq \frac{|x_0|}{2} + |x|$$

impliziert $|x| \geq |x_0|/2$ und dann

$$\left| \frac{1}{x} - \frac{1}{x_0} \right| \leq \frac{2}{|x_0|^2} |x - x_0|$$

Folglich ist die Schwankung des Funktionswerts kleiner als jedes vorgegebene $\varepsilon > 0$, wenn nur

$$|x - x_0| < \delta = \min \left\{ \frac{|x_0|}{2}, \frac{\varepsilon |x_0|^2}{2} \right\}$$

gilt. Als weiteres Beispiel aus der Taschenrechnerliste betrachten wir die Wurzelfunktion $f(x) = \sqrt{x}$. An einer Stelle $x_0 \neq 0$ hilft uns die dritte binomische Formel, um von der Funktionswertedifferenz auf die Argumentdifferenz zu kommen

$$|\sqrt{x}-\sqrt{x_0}| = \frac{\sqrt{x}+\sqrt{x_0}}{\sqrt{x}+\sqrt{x_0}}|\sqrt{x}-\sqrt{x_0}| = \frac{|x-x_0|}{\sqrt{x}+\sqrt{x_0}} \leq \frac{|x-x_0|}{\sqrt{x_0}}$$

Im letzten Schritt haben wir benutzt, daß $\sqrt{x} \geq 0$ ist. Offensichtlich ist die Wurzelfunktion Lipschitz-stetig in $x_0 > 0$ und damit stetig. An der Stelle $x_0 = 0$ versagt der angegebene Trick, aber hier ist die Situation eigentlich einfacher.

$$|\sqrt{x}-\sqrt{0}| = \sqrt{x} = \sqrt{|x-0|}$$

Also ist die Wurzelfunktion Hölder-stetig am Punkt $x_0 = 0$ zum Exponent $\alpha = \frac{1}{2}$ und damit stetig. Das vorgegebene $\varepsilon > 0$ wird offentlichtlich unterboten, wenn $|x-0| < \delta = \varepsilon^2$ gewählt wird.
Die Stetigkeit der anderen Funktionen unserer der Taschenrechnerliste können wir im Moment noch nicht im Detail nachprüfen, da wir die Berechnungsvorschrift dieser Funktionen noch nicht genau angegeben haben. Wir kommen später darauf zurück und begnügen uns im Moment damit, daß man die Stetigkeit nachweisen kann. Wichtiger ist an dieser Stelle die Beobachtung, daß mit der Stetigkeit dieser wenigen Funktionen bereits *automatisch* die Stetigkeit von daraus zusammengesetzten Funktionen folgt. Zum Beispiel kann die Funktion

$$f(x) = \sqrt[3]{\sin(5x-3)} \qquad x \in \mathbb{R}$$

als Verkettung von stetigen Funktionen geschrieben werden. Sehen Sie alle beteiligten Funktionen? Natürlich treten $f_1(x) = \sqrt[3]{x}$ und $f_2(x) = \sin(x)$ auf. Außerdem sehen wir noch $f_3(x) = x$ und die konstanten Funktionen $f_4(x) = -3$ und $f_5(x) = 5$. Etwas schwieriger ist dann schon die Produktfunktion $f_6(y_1, y_2) = y_1 y_2$ zu erkennen, die mit der Tupelfunktion $f_7(x) = (f_5(x), f_3(x))$ verknüpft ist, denn

$$f_6(f_7(x)) = f_6(f_5(x), f_4(x)) = f_5(x) f_4(x) = 5x$$

Genauso tritt die Additionsfunktion $f_8(y_1, y_2) = y_1 + y_2$ auf, die mit der Tupelfunktion $f_9(x) = (f_6(f_7(x)), f_4(x))$ verknüpft ist

$$f_8(f_9(x)) = f_6(f_7(x)) + f_4(x) = 5x + (-3)$$

Insgesamt ist also

$$f(x) = f_1(f_2(f_8(f_9(x))))$$

Die auftretende Tupelbildung ist übrigens leicht als stetige Funktion identifizierbar. Sind $h_i : D \to W_i, i = 1, \ldots, n$ stetige Funktionen auf $D \subset V$, so ist die Tupelfunktion

$$h(\vec{v}) = (h_1(\vec{v}), \ldots, h_n(\vec{v})) \qquad \vec{v} \in D$$

ebenfalls eine stetige Funktion mit dem Zielraum $W = W_1 \times \ldots \times W_n$ bezüglich der Norm

$$||(\vec{w}_1, \ldots, \vec{w}_n)||_W = \max\{||\vec{w}_1||_{w_1}, \ldots, ||\vec{w}_n||_{W_n}\}$$

Kurz gesagt, sind alle Komponenten eines Tupels stetig, so ist die Tupelfunktion auch stetig.
Ein weiteres Beispiel für eine Verkettung stetiger Funktionen ist

$$f(x_1, x_2) = \cos(x_1) \qquad (x_1, x_2) \in \mathbb{R}^2$$

Erkennen Sie hier die beteiligten Funktionen? Klar ist $f_1(x) = \cos(x)$ dabei. Dazu kommt aber noch die (langweilige) Funktion

$$f_2(x_1, x_2) = x_1$$

die $\mathbb{R}^2$ auf $\mathbb{R}$ abbildet (übrigens eine lineare Funktion). Es ist also

$$f(x_1, x_2) = f_1(f_2(x_1, x_2))$$

oder kurz $f = f_1 \circ f_2$. Die Stetigkeit von f_2 ist leicht in allgemeinem Rahmen beweisbar. So ist die Funktion $\Pi_k : W_1 \times \ldots \times W_n \to W_k$, die einem Tupel $(\vec{w}_1, \ldots \vec{w}_n)$ die k-te Komponente $\Pi_k(\vec{w}_1, \ldots, \vec{w}_n) = \vec{w}_k$ zuordnet, offensichtlich Lipschitz-stetig, wenn man die oben angegebene Norm zugrunde legt (Übungsaufgabe).
Nach diesen Beispielen können Sie wohl auch folgende Funktion

$$g(x_1, x_2) = (\cos(x_2 - x_1), x_1^{\frac{1}{3}}, \tan(x_1^2 + x_2^2))$$

zerlegen, die für $\{(x_1, x_2) | \cos(x_1^2 + x_2^2) \neq 0\}$ definiert ist. Diese Einschränkung kommt von der Tangens-Funktion

$$\tan(x) = \frac{\sin(x)}{\cos(x)} \qquad \cos(x) \neq 0$$

die ja für sich genommen bereits eine Verknüpfung aus $\sin, \cos, x \to \frac{1}{x}$ sowie der Produktfunktion $(y_1, y_2) \to y_1 y_2$ ist. Warum eine Verkettung aus stetigen Funktionen wieder stetig ist, sagt folgender

Satz 5. *Seien V, W, X normierte Vektorräume und $D \subset V, E \subset W$. Die Funktionen $f : D \to E$ und $g : E \to X$ seien stetig. Dann ist auch $g \circ f : D \to X$ stetig.*

Der Beweis dieser Aussage geht so: Vorgegeben wird ein $\varepsilon > 0$ und ein Punkt $\vec{x}_0 \in D$. Da g in $f(\vec{x}_0)$ stetig ist, gibt es ein $\eta > 0$, so daß

$$||g(\vec{y}) - g(f(\vec{x}_0))||_X < \varepsilon \qquad \text{falls} \qquad ||\vec{y} - f(\vec{x}_0)||_W < \eta$$

Zu diesem η gibt es wiederum ein $\delta > 0$, jetzt wegen der Stetigkeit von f in $\vec{x}_0$, so daß

$$||f(\vec{x}) - f(\vec{x}_0)||_W < \eta \qquad \text{falls} \qquad ||\vec{x} - \vec{x}_0||_V < \delta$$

Zusammengesetzt gilt also im Falle $||\vec{x} - \vec{x}_0||_V < \delta$

$$||g(f(\vec{x})) - g(f(\vec{x}_0))||_X < \varepsilon$$

d. h. $g \circ f$ ist stetig in $\vec{x}_0$. Da $\vec{x}_0 \in D$ beliebig war, ist $g \circ f$ stetig auf der gesamten Definitionsmenge.
In den obigen Beispielen haben wir gesehen, daß Summen von stetigen Funktionen wieder stetig sind, denn $f + g$ kann als Verkettung der stetigen Addition mit einer stetigen Tupelfunktion $h = (f, g)$ betrachtet werden. Da die Addition in jedem normierten Vektorraum stetig ist, kann man diese Beobachtung verallgemeinern. Seien dazu V, W normierte Vektorräume und $D \subset V$. Wie wir wissen, bildet

$$\mathcal{F}(D, W) = \{f : D \to W\}$$

zusammen mit der punktweisen Addition und skalaren Multiplikation einen Vektorraum. Die Teilmenge aller stetigen Funktionen

$$C^0(D, W) = \{f \in \mathcal{F}(D, W) | f \text{ stetig } \}$$

bildet darin einen Untervektorraum. Um dies nachzuprüfen, müssen wir zeigen, daß mit $f, g \in C^0(D, W)$ und $\alpha \in \mathbb{R}$ auch $f + g \in C^0(D, W)$ und $\alpha g \in C^0(D, W)$ gilt. In Worten lauten diese Bedingungen: Summen und Vielfache stetiger Funktionen sind wieder stetig.
Betrachten wir zunächst die Addition. Wir definitieren die Funktion $h(\vec{v}) = (f(\vec{v}), g(\vec{v}))$ die als Tupel zweier stetigen Funktionen ebenfalls stetig ist als Funktion von D nach $W \times W$. Außerdem ist die Addition $A : W \times W \to W$ stetig und damit auch die Verkettung $A \circ h : D \to W$, wobei diese Verkettung gerade die Summe der Funktionen f und g ist

$$(A \circ h)(\vec{v}) = A(h(\vec{v})) = A(f(\vec{v}), g(\vec{v})) = f(\vec{w}) + g(\vec{v}) = (f + g)(\vec{v}).$$

In ähnlicher Weise zeigt man, daß das α-Vielfache von g stetig ist. Dazu definieren wir die Konstante (also stetige) Funktion $f : D \to \mathbb{R}$ mit $f(\vec{v}) = \alpha$ und bilden wieder ein stetiges Paar $h = (f, g) : D \to \mathbb{R} \times W$.

Verknüpft mit der stetigen skalaren Multiplikation $M : \mathbb{R} \times W \to W$ ergibt sich die Stetigkeit von αg, denn

$$(M \circ h)(\vec{v}) = M(h(\vec{v})) = M(f(\vec{v}), g(\vec{v})) = \alpha g(\vec{v}) = (\alpha g)(\vec{v})$$

Beachten Sie, daß die *beiden* Aussagen über Stetigkeit von Summen und Vielfachen von stetigen Funktionen in der *einen* Aussage enthalten ist, daß $C^0(D, W)$ einen Vektorraum bildet.
Wir können also mit stetigen Funktionen intuitiv so arbeiten, wie mit Zeigern oder anderen möglichen Ausprägungen von Vektoren. So ist z. B. eine beliebige Linearkombination

$$\lambda_0 f_0 + \lambda_1 f_1 + \ldots + \lambda_n f_n$$

mit $\lambda_i \in \mathbb{R}$ und $f_i \in C^0(D, W)$ wieder eine stetige Funktion auf D. Wenden wir diese Beobachtung auf den Fall $f_i = q_i \in C^0(\mathbb{R}, \mathbb{R})$ an mit $q_i(x) = x^i$, so sehen wir, daß *alle* Polynome stetige Funktionen sind.
Die Stetigkeit der Monome q_i folgt dabei induktiv aus der Tatsache, daß $q_i = q_1 \cdot q_{i-1}$ gilt und daß q_0 und q_1 stetig sind.
Genauso wie Summen und Vielfache von stetigen Funktionen wieder stetig sind, gilt dies nämlich auch für Produkte und Quotienten von *reellwertigen* Funktionen. Die Produktbildung von f und g betrachtet man dazu einfach als Verkettung der (skalaren) Multiplikation $M : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ mit dem stetigen Paar $h = (f, g)$. Ein Quotient f/g ist zunächst eine Multiplikation $f \cdot \frac{1}{g}$, wobei $\frac{1}{g}$ stetig ist als Verkettung von $x \to \frac{1}{x}$ mit g. Beachten Sie hier besonders, daß die Verkettung nur dort definiert ist, wo $g(\vec{v}) \neq 0$ ist, d. h. der Quotient f/g zweier stetiger Funktionen ist überall dort stetig, wo er definiert ist.
Als direkte Konsequenz können wir nun alle *rationalen* Funktionen, also Funktionen der Form

$$r(x) = \frac{P(x)}{Q(x)} = \frac{a_0 + a_1 x + \ldots + a_n x^n}{b_0 + b_1 x + \ldots + b_m x^m} \qquad x \in \mathbb{R}, Q(x) \neq 0$$

als stetig verbuchen. Die Klasse der rationalen Funktionen spielt in der Mathematik eine wichtige Rolle, da die Funktionswerte mit endlich vielen elementaren Operationen $(+, -, \cdot, /)$ berechnet werden können. Kurz gesagt, der Computer kennt nur rationale Funktionen. Jeder andere Zusammenhang muß letztlich (approximativ) auf rationale Funktionen zurückgeführt werden.

3. Das Rechnen mit konvergenten Folgen

Am Anfang unserer Betrachtung von stetigen (d. h. vorhersagbaren) Zusammenhängen stand ein Kriterium für sprunghaftes Verhalten einer

Funktion. Wenn eine Funktion $f : D \to W$ bei $\vec{v}_0 \in D$ springt, bedeutet dies, daß es eine zugehörige Mindestsprungweite $\varepsilon_0 > 0$ gibt, so daß beliebig nahe bei $\vec{v}_0$ Funktionswertvariationen von mindestens ε_0 auftreten. Es gibt dann also für *jedes* $\delta > 0$ ein $\vec{v} \in D$ mit $||\vec{v}-\vec{v}_0|| < \delta$ aber $||f(\vec{v})-f(\vec{v}_0)|| > \varepsilon_0$. Wählen wir nacheinander für δ die Werte $\frac{1}{n}, n \in \mathbb{N}$ so gibt es folglich zu jeder Wahl einen Punkt $\vec{v}_n$ mit $||\vec{v}_n-\vec{v}_0|| < \frac{1}{n}$ aber $||f(\vec{v}_n)-f(\vec{v}_0)|| > \varepsilon_0$. Die Punkte $\vec{v}_n$ kommen dem Punkt $\vec{v}_0$ also immer näher, aber die Funktionswerte nähern sich *nicht* dem Funktionswert bei $\vec{v}_0$ an.
Wir können also den Sprung von f bei $\vec{v}_0$ durch das Beobachten von Funktionswerten $f(\vec{v}_n)$ entlang von Folgen $(\vec{v}_n)_{n\in\mathbb{N}}$ in D detektieren, wenn $\vec{v}_n$ gegen $\vec{v}_0$ strebt. Und zwar gibt es bei $\vec{v}_0 \in D$ einen Sprung in f, wenn man sich so an $\vec{v}_0$ „heranpirschen“ kann, daß die zugehörigen Funktionswerte *nicht* gegen $f(\vec{v}_0)$ streben. Heranpirschen bedeutet hier, daß man Punkte $\vec{v}_n \in D$ wählt, die immer näher an $\vec{v}_0$ liegen, je größer n ist, d. h. für die $\lim ||\vec{v}_n - \vec{v}_0|| = 0$ gilt.
Als Beispiel, wie man sich in $V = \mathbb{R}^{2n\to\infty}$ an den Punkt $(1, 2)$ heranpirschen kann, betrachten wir

$$\vec{a}_n = \left(1 + \frac{2}{n}, 2 + \frac{1}{n}\right), \quad \vec{b}_n = \left(1 + \frac{1}{n}, 2 + \frac{8}{n^2}\right)$$

und

$$\vec{c}_n = \left(1 - \frac{1}{n}, 2 + \frac{1}{n}\right), \quad \vec{d}_n = \left(1, 2 - \frac{1}{\sqrt{n}}\right)$$

Wenn Sie die Folgenglieder als Koordinaten von Punkten in der Ebene auffassen, dann nähert sich $(\vec{a}_n)$ von oben rechts entlang einer Geraden an, $(\vec{b}_n)$ von oben rechts auf einer Parabel, $(\vec{c}_n)$ von oben links auf einer Geraden und $(\vec{d}_n)$ genau von unten

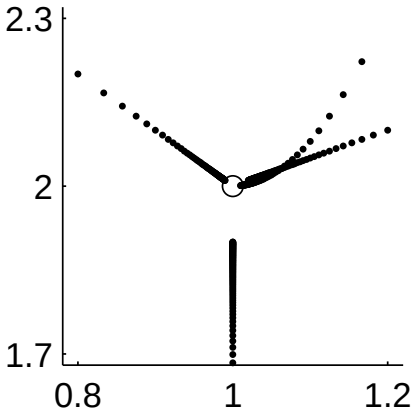


Abbildung 3: Wertebereich der Funktion $\vec{\gamma}$

Wenn eine Funktion in $\vec{v}_0 \in D$ einen Sprung hat, bedeutet das natürlich nicht, daß Sie dies entlang jeder Folge bemerken.
Umgekehrt verlangt es etwas Geschick, eine Folge zu finden, bei der der Sprung offensichtlich wird. Betrachten Sie dazu folgende Skizze des Funktionsgraphen einer Funktion $f : D \to \mathbb{R}$ mit $D \subset \mathbb{R}^2$, (der Zielvektorraum W ist hier der Vektorraum der reellen Zahlen)

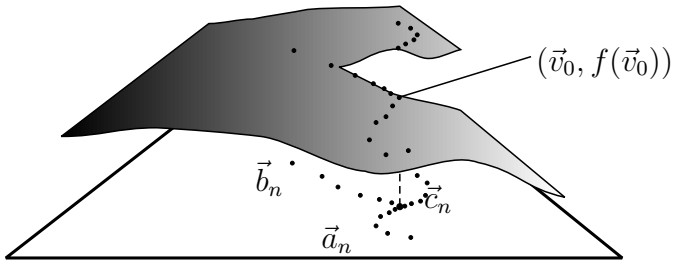


Abbildung 4: Wertebereich der Funktion $\vec{\gamma}$

Entlang der Folge $(\vec{a}_n)_{n\in\mathbb{N}}$ oder $(\vec{b}_n)_{n\in\mathbb{N}}$ wird der Sprung offensichtlich nicht bemerkt, da die Funktionswerte gegen $f(\vec{v}_0$ streben, d. h. die Punkte auf dem Graphen $(\vec{a}_n, f(\vec{a}_n))$ kommen immer näher an $(\vec{v}_0, f(\vec{v}_0))$. Nur bei Folgen, die wie $(\vec{c}_n)_{n\in\mathbb{N}}$ sich aus der richtigen Richtung annähern, merkt man, daß die Funktionswerte nicht gegen $f(\vec{v}_0)$ streben (d. h. auf dem Graph behalten die Punkte $(\vec{c}_n, f(\vec{c}_n))$ einen bestimmten Mindestabstand von $(\vec{v}_0, f(\vec{v}_0)))$.
Natürlich sagt Ihnen in diesem Fall Ihr Auge, wie die Annäherungsfolge zu wählen ist, bei der der Sprung offensichtlich wird. In allgemeinen

Fällen läßt sich der Graph allerdings nicht mehr ohne weiteres darstellen und Sie müssen in gewissem Sinne blind den richtigen Pfad finden – und das kann schon deutlich schwieriger sein.
Insgesamt können wir aber festhalten, daß man Sprünge durch die Beobachtung der Funktionswerte entlang von Folgen detektieren kann: Wenn es einen Sprung gibt, so gibt es auch eine Folge, die diesen Sprung anzeigt. Umgekehrt hat die Funktion bei $\vec{v}_0$ *keinen* Sprung, wenn entlang *jeder* Folge die sich an $\vec{v}_0$ heranpirscht, die Funktionswerte brav auf $f(\vec{v}_0)$ zulaufen. Diesen Zusammenhang zwischen Folgen und stetigen Funktionen wollen wir präzise festhalten. Zunächst brauchen wir dazu folgende

Definition 11. *Sei $\vec{a} : \mathbb{N} \to V$ eine Folge mit Werten in einem normierten Vektorraum V. Dann heißt $\vec{a} \in V$ Grenzwert der Folge (bzw. $\vec{a}_n$ konvergiert gegen $\vec{a}$, Symbole $\lim\limits_{n\to\infty} \vec{a}_n = \vec{A}$ oder $\vec{a}_n \xrightarrow[n\to\infty]{} \vec{A}$), falls $\lim\limits_{n\to\infty} ||\vec{a}_n = \vec{a}|| = 0$ gilt.*

Mit dieser Schreibweise lautet unsere Beobachtung

Satz 6. *Seien V, W normierte Vektorräume und $D \subset V$. Dann ist die Funktion f stetig in $\vec{v}_0 \in D$, genau dann, wenn für alle Folgen $\vec{v} : \mathbb{N} > D$ mit $\lim\limits_{n\to\infty} \vec{v}_n = \vec{v}_0$ auch $\lim\limits_{n\to\infty} f(\vec{v}_n) = f(\vec{v}_0)$ gilt.*

Um die etwas sperrige Formulierung: Für alle Folgen $\vec{v} : \mathbb{N} \to D$ mit $\lim \vec{v}_n = \vec{v}_0$ weiter zu verkürzen, einigen wir uns noch auf folgende

Definition 12. *Seien V, W normierte Vektorräume, $D \subset V$ und $f : D \to W$. Weiter sei $\vec{x}_0 \in V$ von D aus approximierbar, d. h. es gibt mindestens eine Folge $\vec{v} : \mathbb{N} \to D$ mit $\lim\limits_{n\to\infty} \vec{v}_1 = \vec{x}_0$. Man schreibt*

$$\lim_{\vec{x}\to\vec{x}_0} f(\vec{x}) = \vec{w}$$

wenn $\lim\limits_{n\to\infty} f(\vec{v}_n) = \vec{w}$ für jede Folge $\vec{v} : \mathbb{N} \to D$ mit $\lim\limits_{n\to\infty} \vec{v}_n = \vec{x}_0$.

Damit können wir jetzt ganz knapp schreiben

$$f \text{ stetig in } \vec{x}_0 \qquad \Longleftrightarrow \qquad \lim_{\vec{x}\to\vec{x}_0} f(\vec{x}) = f(\vec{x}_0)$$

Da wir wissen, daß die Identitätsfunktion $\vec{x} \mapsto \vec{x}$ auf jedem normierten Vektorraum stetig ist, so folgt $\lim\limits_{\vec{x}\to\vec{x}_0} \vec{x} = \vec{x}_0$ und die rechte Seite der obigen Äquivalenz läßt sich auch folgendermaßen formulierten

$$\lim_{\vec{x}\to\vec{x}_0} f(\vec{x}) = f\left(\lim_{\vec{x}\to\vec{x}_0} \vec{x}\right)$$

Stetige Funktionen f sind also genau die Funktionen, bei denen man f und lim *vertauschen* darf bzw. bei denen man lim ins Argument ziehen kann.
Diese Tatsache ist enorm hilfreich beim Berechnen von Grenzwerten, da Sie mit der Kenntnis sehr weniger konvergenter Grundfolgen bereits die Konvergenz komplizierter Folgen nachweisen können. Nehmen wir einmal an, die einzige konvergente Folge, die wir kennen, sei $a_n = \frac{1}{n}, n \in \mathbb{N}$ mit $\lim\limits_{n\to\infty} a_n = 0$. Wenn wir jetzt über die Konvergenz einer komplizierten Folge gefragt werden. z. B. $b_n = \frac{1}{n^5}, n \in \mathbb{N}$, so müssen wir nur versuchen, diesen Ausdruck als $f(a_n)$ zu schreiben, mit einer stetigen Funktion f. Im vorliegenden Beispiel erreicht man dies offensichtlich mit $f(x) = x^5$ und somit gilt

$$\lim_{n\to\infty} b_n = \lim_{n\to\infty} f(a_n) = f\left(\lim_{n\to\infty} a_n\right) = f(0) = 0$$

Mit der gleichen Argumentation sehen wir, daß jede Folge vom Typ $\frac{1}{n^k}, n \in \mathbb{N}$ gegen Null konvergiert. Auch andere Folgen, wie $\frac{1}{\sqrt{n}}$ oder $\frac{1}{\sqrt[k]{n}}$ werden als Nullfolgen entlarvt, da sie von der Form $g_k(a_n)$ sind mit den im Punkt Null stetigen Funktionen $g_k(x) = \sqrt[k]{x}$.
Bei einem komplizierteren Ausdruck, wie

$$c_n = \frac{3n + 5n^2 - 1}{n^2 - 2n + 2}, \qquad n \in \mathbb{N}$$

muß man erst ein wenig Vorarbeit leisten, um die bekannte Folge $a_n = \frac{1}{n}$ zu entdecken. Durch Erweitern mit dem Kehrwert $\frac{1}{n^2}$ der höchsten auftretenden n-Potenz ergibt sich

$$c_n = \frac{\frac{3}{n} + 5 - \frac{1}{n^2}}{1 - \frac{2}{n} + \frac{2}{n^2}} = f\left(\frac{1}{n}\right)$$

wobei

$$f(x) = \frac{3x + 5 - x^2}{1 - 2x + 2x^2}$$

Um die Argumentation erfolgreich zu beenden, muß nur noch die Stetigkeit von f im Punkt Null nachgewiesen werden, wobei nur eine Nullstelle im Nenner Schwierigkeiten machen könnte. Da der Nenner auch als

$$1 - 2x + 2x^2 = 2\left(x - \frac{1}{2}\right)^2 + \frac{1}{2} > 0$$

geschrieben werden kann, ist f aber sogar auf ganz $\mathbb{R}$ stetig, also insbesondere im Punkt Null. Es gilt

$$\lim_{n\to\infty} c_n = \lim_{n\to\infty} f(a_n) = f\left(\lim_{n\to\infty} a_n\right) = f(0) = 5$$

Wendet man den Satz über die Vertauschbarkeit von Grenzwertbildung und Anwendung einer stetigen Funktion auf die Operationen Addition und skalare Multiplikation an, so erhält man die nützlichen Regeln, daß Summen und Vielfache konvergenter Folgen wieder konvergent sind. Dabei ist der Grenzwert der Summenfolge gleich der Summe der Grenzwerte und eine entsprechende Regel gilt für die Multiplikation mit Skalaren. Für diese Aussage benötigt man ein Resultat über Tupel von Folgen: Sind $\vec{v}^{(i)} : \mathbb{N} \to V^{(i)}, i = 1, \ldots k$ Folgen in normierten Vektorräumen $V^{(i)}$, so ist das Tupel

$$(\vec{v}^{(1)}, \ldots, \vec{v}^{(k)}) : \mathbb{N} \to V^{(1)} \times \ldots \times V^{(k)} = W$$

genau dann konvergent bezüglich der Norm

$$||(\vec{w}_1, \ldots, \vec{w}_k)||_W = \max_{i=1\ldots k} ||\vec{w}_i||_{V^{(i)}}$$

wenn jede der Folgen $\vec{v}^{(i)}$ konvergent in $V^{(i)}$ ist. Kurz gesagt, eine Tupelfolge konvergiert genau dann, wenn jede Komponente konvergiert. Zum Beweis nehmen wir zunächst an, daß $\lim_{n\to\infty} \vec{v}_n^{(i)} = \vec{w}^{(i)}$ für $i = 1, \ldots, k$. Damit gibt es zu einem beliebig vorgegebenen $\varepsilon > 0$ ein $N_i \in \mathbb{N}$, so daß $||\vec{v}_n^{(i)} - \vec{w}^{(i)}||_{V^{(i)}} < \varepsilon$ für alle $n \geq N_i$. Wählen wir nun N als Maximum aller $N_i, i = 1, \ldots k$, so gilt für jede Folge, daß

$$||\vec{v}_n^{(i)} - \vec{w}^{(i)}||_{V^{(i)}} < \varepsilon \quad \text{für} \quad n \geq N$$

und damit auch für das Maximum

$$||(\vec{v}_n^{(1)}, \ldots, \vec{v}_n^{(k)}) - (\vec{w}^{(1)}, \ldots, \vec{w}^{(k)})||_W = \max_{i=1,\ldots,k} ||\vec{v}_n^{(i)} - \vec{w}^{(i)}||_{V^{(i)}} < \varepsilon$$

falls $n \geq N$. Nach Definition konvergiert also die Tupelfolge gegen das Tupel aller Grenzwerte. Umgekehrt sieht man genauso, daß, wenn die Tupelfolge gegen ein Tupel $(\vec{w}^{(1)}, \ldots, \vec{w}^{(k)})$ konvergiert, dann gilt für jedes $i = 1, \ldots, k$

$$0 \leq ||\vec{v}_n^{(i)} - \vec{w}^{(i)}||_{V^{(i)}} \leq \max_{j=1,\ldots,k} ||\vec{v}_n^{(j)} - \vec{w}^{(j)}||_{V^{(i)}} \xrightarrow[n\to\infty]{} 0$$

und somit $\lim_{n\to\infty} \vec{v}_n^{(i)} = \vec{w}^{(i)}$.

Nach dieser sorgfältigen Betrachtung konvergenter Tupel können wir uns nun der Summe von Folgen zuwenden. Sind $\vec{a}_n, \vec{c}_n : \mathbb{N} \to V$ konvergente Folgen im normierten Vektorraum V, so kann man die Summenfolge $\vec{a}+\vec{c}_n$ schreiben als $A(\vec{g}_n)$, wobei $\vec{g}_n = (\vec{a}_n, \vec{c}_n)$ eine Tupelfolge ist und $A : V \times V \to V$ die stetige Vektorraumaddition darstellt. Gilt $\lim\limits_{n\to\infty} \vec{a}_n = \vec{a}_\infty$ und $\lim\limits_{n\to\infty} \vec{c}_n = \vec{c}_\infty$, so ist nach dem oben Gesagten

$$\lim_{n\to\infty} \vec{g}_n = (\vec{a}_\infty, \vec{c}_\infty).$$

Die Stetigkeit der Addition A impliziert dann, daß auch $\vec{a}_n+\vec{c}_n = A(\vec{g}_n)$ konvergiert und

$$\lim_{n\to\infty} (\vec{a}_n + \vec{c}_n) = \lim_{n\to\infty} A(\vec{g}_n) = A(\lim_{n\to\infty} \vec{g}_n) = A(\vec{a}_\infty, \vec{c}_\infty) = \vec{a}_\infty + \vec{c}_\infty.$$

Genauso zeigt man, daß, wenn $a_n : \mathbb{N} \to \mathbb{R}$ und $\vec{c}_n : \mathbb{N} \to V$ konvergent sind, auch $a_n\vec{c}_n : \mathbb{N} \to V$ konvergiert. Hier benutzt man die Stetigkeit der skalaren Multiplikation $M : \mathbb{R} \times V \to V$ und

$$\lim_{n\to\infty} (a_n, \vec{c}_n) = (\lim_{n\to\infty} a_n, \lim_{n\to\infty} \vec{c}_n)$$

was als Spezialfall $V^{(1)} = \mathbb{R}, V^{(2)} = V$ unserer Betrachtung von Tupelfolgen gesehen werden kann.
Im Fall $V = \mathbb{R}$ ergibt sich eine Regel zum Produkt von reellen Folgen

$$\lim_{n\to\infty} a_n c_n = \lim_{n\to\infty} a_n \lim_{n\to\infty} c_n$$

Wählt man die skalare Folge konstant, also $a_n = \alpha$ für alle $n \in \mathbb{N}$, so folgt insbesondere

$$\lim_{n\to\infty} \alpha\vec{c}_n = \alpha \lim_{n\to\infty} \vec{c}_n$$

Das Verhalten von konvergenten Folgen bezüglich Addition und Vielfachenbildung können wir auch mit der Feststellung zusammenfassen, daß die konvergenten Folgen einen Untervektorraum aller Folgen bilden. Nennen wir die konvergenten Folgen

$$\mathcal{K}(\mathbb{N}, V) = \{(\vec{v}_n)_{n\in\mathbb{N}} \in \mathcal{F}(\mathbb{N}, V) | (\vec{v}_n)_{n\in\mathbb{N}} \text{ konvergent}\}$$

so besagt die Unterraumeigenschaft ja gerade, daß Summen von konvergenten Folgen wieder kongergent sind ($\vec{a} + \vec{c} \in \mathcal{K}(\mathbb{N}, V)$ falls $\vec{a}, \vec{c} \in \mathcal{K}(\mathbb{N}, V)$) und daß das jedes α-Vielfache einer konvergenten Folge auch konvergent ist ($\alpha\vec{c} \in \mathcal{K}(\mathbb{N}, V)$ falls $\vec{c} \in \mathcal{K}(\mathbb{N}, V)$)

Das Verhalten der Grenzwerte läßt sich dann dadurch ausdrücken, daß $\lim\limits_{n\to\infty} : \mathcal{K}(\mathbb{N}, V) \to V$ eine lineare Funktion ist (sie ordnet jeder konvergenten Folge ihren Grenzwert zu).
Beachten Sie, daß alle diese Aussagen über das Verhalten von konvergenten Folgen stets Konsequenzen des einen Sachverhalts sind, das lim mit stetigen Funktionen vertauscht. Mit dieser Grundregel sollte es Ihnen jetzt nicht schwer fallen, weiter „Rechenregeln“ für konvergente Folgen herzuleiten. Ist z. B. die Folge $a : \mathbb{N} \to R \setminus \{0\}$ konvergent und $\lim\limits_{n\to\infty} a_n \neq 0$, so gilt auch

$$\lim_{n\to\infty} \frac{1}{a_n} = \frac{1}{\lim\limits_{n\to\infty} a_n}$$

Hier steckt natürlich die stetige Funktion $f(x) = \frac{1}{x}$ dahinter. Die Bedingungen an die Folgenglieder und den Grenzwert stammen dabei offensichtlich daher, daß man die Stetigkeit von f nur an Punkten des Definitionsgebietes $\mathbb{R} \setminus \{0\}$ ausnutzen kann und daß die Folgenglieder a_n alle im Definitionsgebiet liegen müssen. Außerdem können Sie jetzt zeigen, daß Summen von konvergenten Folgen komplexerer Zahlen wieder konvergent sind. Erinnern Sie sich einfach daran, daß $\mathbb{C}$ bezüglich Addition gerade dem $\mathbb{R}^2$ entspricht. Das Gleiche gilt auch im Zusammenhang mit der komplexen Multiplikation.

$$(x_1, x_2)(y_1, y_2) = (x_1y_1 - x_2y_2, x_1y_2 + x_2y_1)$$

die ja offensichtlich eine stetige Funktion von $\mathbb{R}^2 \times \mathbb{R}^2$ nach $\mathbb{R}^2$ darstellt. (Sehen Sie, welche stetigen Grundfunktionen hier verkettet werden?) Sind nun $(z_n)_{n\in\mathbb{N}}$ und $(w_n)_{n\in\mathbb{N}}$ konvergente Folgen komplexer Zahlen, und $M : \mathbb{C} \times \mathbb{C} \to \mathbb{C}$ die stetige komplexe Multiplikation, so ergibt sich die Rechenregel

$$\lim_{n\to\infty} z_n w_n = \lim_{n\to\infty} M(z_n, w_n) = M\left(\lim_{n\to\infty}(z_n, w_n)\right) = \lim_{n\to\infty} z_n \lim_{n\to\infty} w_n.$$

Am Ende dieses Abschnitts soll nicht verschwiegen werden, daß es auch Folgen gibt, bei denen die Konvergenz nicht einfach daraus resultiert, daß sie als Verknüpfungen stetiger Funktionen mit bekannter Folge geschrieben werden können. Ein Beispiel ist

$$a_n = \left(1 + \frac{1}{n}\right)^n \qquad n \in \mathbb{N}$$

Versucht man, diesen Ausdruck als Funktion von $\frac{1}{n}$ zu schreiben, so findet man z. B. $a_n = f\left(\frac{1}{n}\right)$ mit

$$f(x) = (1+x)^{\frac{1}{x}} = \exp\left(\frac{1}{x}\ln(1+x)\right), \quad x > 0$$

Dies ist zwar eine stetige Funktion für $x > 0$, aber genau in $x = 0$, wo wir die Stetigkeit bräuchten, um $\lim\limits_{n\to\infty} f\left(\frac{1}{n}\right) = f\left(\lim\limits_{n\to\infty}\frac{1}{n}\right)$ schreiben zu können, ist f nicht definiert (und damit auch nicht stetig). Die Argumentfolge läuft im Grenzwert gerade aus dem Definitionsgebiet heraus und dabei kann natürlich alles Mögliche passieren. Denken Sie z. B. einmal an die Funktion $g_1(x) = \frac{1}{x}, x > 0$. Hier wachsen die Funktionswerte für $x \to 0$ unbeschränkt an und folglich kann man keinen vernünftigen Wert bei $x = 0$ erwarten. Die Funktion $g_2(x) = \frac{\frac{2}{x}+1}{\frac{1}{x}+1}$, die auch für $x = 0$ nicht definiert ist, hat dagegen zumindest ein vernünftiges Grenzverhalten. Nähert sich x dem Wert Null, so nähert sich $g_2(x)$ immer mehr dem Wert Zwei an und zwar unabhängig davon, wie man sich an $x = 0$ heranpirscht. (Erweitern Sie einfach den Bruch mit x, um dies nachzurechnen.) Mit unserer Definition des Grenzwerts können wir also schreiben

$$\lim_{x\to 0} g_2(x) = 2.$$

Genauso können Sie Funktionen konstruieren, die für $x = 0$ nicht definiert sind, aber dort jeden beliebigen Wert $\alpha \in \mathbb{R}$ immer näher kommen, etwa

$$g_\alpha(x) = \frac{\frac{\alpha}{x}+5}{\frac{1}{x}+1}, \quad x > 0$$

Neben Konvergenz und unbeschränktem Wachstum, können stetige Funktionen am Rand ihres Definitionsgebiets aber auch ein total ausgeflipptes Verhalten zeigen. Schauen wir uns z. B. die Funktion

$$f(x) = \sin\frac{1}{x}, \quad x > 0$$

an. Wenn x sich auf Null zubewegt, durchlaufen die Funktionswerte immer schneller die Perioden der Sinusfunktion. Schließlich ist das Verhalten so rasant, daß man mit dem Zeichnen des Graphen nicht mehr nachkommt.

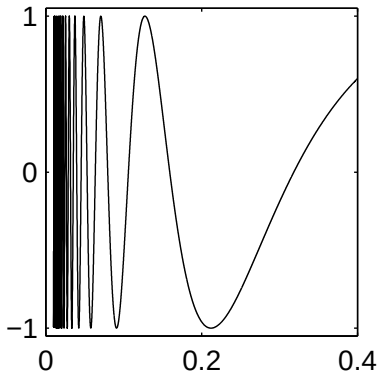


Abbildung 5: Graph der Funktion $x \mapsto \sin \frac{1}{x}$

Obwohl die Funktionswerte hier offensichtlich beschränkt sind, ist anschaulich auch klar, daß die Funktionswerte in diesem Beispiel nicht auf einen bestimmten Wert zulaufen, d. h. $\lim_{x\to 0} f(x)$ existiert *nicht*. Um dies nachzuweisen, genügt es zu zeigen, daß man unterschiedliches Grenzverhalten der Funktionswerte beobachtet, abhängig davon, wie man sich an $x = 0$ heranpirscht. Nehmen Sie einfach die beiden Nullfolgen

$$a_n = \frac{1}{2\pi n}, \qquad b_n = \frac{1}{2\pi n + \frac{\pi}{2}}$$

In einem Fall ist $f(a_n) = 0$ für alle n, im anderen Fall gilt stets $f(b_n) = 1$. Diese Beispiele zeigen, daß wir nicht ohne weiteres beurteilen können, wie sich die Funktion

$$(1+x)^{\frac{1}{x}} = \exp\left(\frac{1}{x}\ln(x+1)\right), \quad x > 0$$

bei $x = 0$ verhalten wird. Während $\frac{1}{x}$ unbeschränkt wächst für $x \to 0$, nähert sich $\ln(1 + x)$ immer mehr dem Wert 0 an. Wer gewinnt im Produkt? Geht $\frac{1}{x}$ schneller gegen Unendlich als $\ln(1 + x)$ gegen Null? Dann würde das Produkt immer größer werden. Ist $\ln(1 + x)$ schneller klein als $\frac{1}{x}$ groß? Dann würde das Produkt gegen Null gehen. Oder halten sich beide Faktoren irgendwie die Waage? Dann könnte jeder Wert zwischen Null und Unendlich in Frage kommen. Oder schwankt das Verhalten ständig wie bei $\sin \frac{1}{x}$? Im Zusammenhang mit der Regel von l'Hospital werden wir später ausrechnen können, daß der Grenzwert

$$\lim_{x\to 0} \frac{1}{x}\ln(1+x) = 1$$

tatsächlich existiert. Wegen der Stetigkeit der Exponentialfunktion folgt daher

$$\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n = \exp(1) = e.$$

Ein weiteres Beispiel, bei dem das Verhalten einer stetigen Funktion am Rand dieses Definitionsgebietes entscheidend ist, ist

$$c_n = \sqrt[n]{n}, \qquad n \in \mathbb{N}.$$

Es gilt $c_n = f\left(\frac{1}{n}\right)$ mit

$$f(x) = \left(\frac{1}{x}\right)^x = \exp\left(x \ln \frac{1}{x}\right) = \exp(-x \ln x) \quad x > 0$$

Hier geht im Produkt $x \ln x$ der Faktor $\ln x$ für $x \to 0$ gegen $-\infty$ während der Faktor x auf Null zuläuft. In diesem Fall ist aber x schneller bei Null als $\ln x$ bei $-\infty$ und es ergibt sich (wieder mit der Regel von l'Hospital)

$$\lim_{n\to\infty} \sqrt[n]{n} = \exp(0) = 1$$

Schauen wir uns noch eine Folge vom spannenden Typ an

$$az_n \frac{(-1)^n}{n}, \qquad n \in \mathbb{N}$$

Dieser Ausdruck sieht harmlos aus und Sie sagen sicher sofort, daß der Grenzwert Null sein wird. Können Sie diesen Ausdruck aber als $f\left(\frac{1}{n}\right)$ schreiben mit $f : (0,1] \to \mathbb{R}$? $f(x) = (-1)^x \cdot x$ macht keinen Sinn, da wir Potenzen nur für positive Basen erklärt haben. Zur Beschreibung des oszillierenden Verhaltens kann man aber die wilde Funktion $\sin \frac{1}{x}$ benutzen. Setzen wir

$$f(x) = x \cdot \sin\left(\frac{\pi}{x} + \frac{\pi}{2}\right), \qquad x > 0$$

so gilt tatsächlich $f\left(\frac{1}{n}\right) = a_n$. Wieder benötigen wir den Grenzwert der Funktion f am Rand der Definitionsmenge $x = 0$ und in diesem Fall existiert er auch, denn die erweiterte Funktion $g : \mathbb{R} \to \mathbb{R}$ mit

$$g(x) = \begin{cases} x \sin\left(\frac{\pi}{x} + \frac{\pi}{2}\right) & x \neq 0 \\ 0 & x = 0 \end{cases}$$

ist stetig. An Punkten $x_0 \neq 0$ folgt dies sofort mit dem Verknüpfungssatz aus der Stetigkeit der beteiligten elementaren Funktionen. An der Stelle $x_0 = 0$ mit $g(x_0) = 0$ zeigen wir für $x \neq 0$

$$|g(x) - g(x_0)| = |g(x)| = \left| x \sin\left(\frac{\pi}{x} + \frac{\pi}{2}\right)\right| \leq |x| = |x - x_0|$$

so daß $|g(x) - g(x_0)| < \varepsilon$ falls $|x - x_0| < \delta = \varepsilon$. Damit gilt also

$$\lim_{n\to\infty} \frac{(-1)^n}{n} = \lim_{n\to\infty} g\left(\frac{1}{n}\right) = g\left(\lim_{n\to\infty} \frac{1}{n}\right) = g(0) = 0$$

Natürlich ist es bei diesem Beispiel nicht unbedingt notwendig, eine passende stetige Funktion g zu finden. Man kann auch direkt mit der Grenzwertdefinition argumentieren. Zu vorgegebenem $\varepsilon > 0$ wählen wir nämlich einfach N_ε als eine natürliche Zahl, die größer als $\frac{1}{\varepsilon}$ ist und dann gilt für $n \geq N_\varepsilon$

$$|a_n - 0| = \left|\frac{(-1)^n}{n}\right| = \frac{1}{n} < \varepsilon$$

d. h. $\lim\limits_{n\to\infty} a_n = 0$. Der kleine Umweg über die Funktion g hat allerdings den Vorteil, daß zusätzliche Information herausspringt, obwohl die Abschätzung praktisch die gleiche ist.
Eine weitere Klasse von Folgen, bei denen Konvergenzaussagen etwas schwieriger sind, sind die sogenannten *rekursiven Folgen*. Bei rekursiven Folgen ist das n-te Folgenglied nicht explizit als Funktion von n gegeben, sondern nur seine Abhängigkeit zu Vorgängerwerten.
Als Beispiel betrachten wir

$$a_{n+1} = \frac{1}{1 + a_n}, \qquad a_0 = 0$$

Schreiben Sie einfach einmal ein paar Folgenglieder hin

$$a_0 = 0,\ a_1 = 1,\ a_2 = \frac{1}{1+1},\ a_3 = \frac{1}{1+\frac{1}{1+1}},\ a_4 = \frac{1}{1+\frac{1}{1+\frac{1}{1+1}}}$$

Die Folge wird (aus offensichtlichem Grund) auch als Kettenbruch bezeichnet. Die Frage nach der Konvergenz ist schon etwas kniffelig, da wir nicht mal ohne weiteres den Wert von a_{100} hinschreiben können. *Wenn* die Folge aber konvergiert, dann ist die Menge der möglichen Grenzwerte stark eingeschränkt! Nehmen wir an, daß $\lim\limits_{n\to\infty} a_n = a_\infty$. Da dann auch a_{n+1} gegen a_∞ konvergiert, liefert die Rekursionsvorschrift

$$a_\infty = \lim_{n\to\infty} \ a_{n+1} = \lim_{n\to\infty} \ \frac{1}{1 + a_n} \ \frac{1}{1 + \lim\limits_{n\to\infty} a_n} = \frac{1}{1 + a_\infty}$$

d. h. a_∞ muß eine Lösung der Gleichung

$$a_\infty^2 + a_\infty - 1 = 0$$

sein, was auf die beiden Werte

$$a_\infty \in \left\{ \frac{-1+\sqrt{5}}{2}, \quad \frac{-1-\sqrt{5}}{2} \right\}$$

führt. Die Grenzwertkandidatenliste ist also sehr klein. Trotzdem wissen wir noch nicht *ob* $(a_n)_{n\in\mathbb{N}}$ gegen einen dieser Werte konvergiert. Hilfreich wäre es, wenn wir zeigen könnten, daß die Folge z. B. monoton und beschränkt ist, denn für solche Folgen haben wir ja eine Konvergenzaussage. Um an diese Information zu gelangen, muß übrigens meist ein Induktionsargument benutzt werden, da ja bereits die Konstruktionsvorschrift induktiv gegeben ist. In unserem Fall sehen wir z. B. durch einen einfachen Induktionsbeweis, daß alle Folgenglieder nicht negativ sind und damit folgt sofort $0 \le a_n \le 1$ für alle n. Die Beschränktheit ist also gesichert und der negative Kandidat $(-1-\sqrt{5})/2$ ist auch schon aus dem Rennen. Aber wie sieht es mit der Monotonie aus? Schauen wir einmal auf einige Folgenglieder

$$a_0 = 0,\ a_1 = 1,\ a_2 = \frac{1}{2},\ a_3 = \frac{2}{3}, \ldots$$

Bestenfalls ist hier zu hoffen, daß die beiden Teilfolgen $b_n = a_{2m}$ und $c_m = a_{2m+1}$ monoton wachsend beziehungsweise fallen sind. Tatsächlich kann man dies rekursiv nachweisen, wozu in einem weiteren Rekursionsbeweis gezeigt werden muß, daß stets

$$b_n \le \frac{-1+\sqrt{5}}{2} \qquad \text{und} \qquad c_n \ge \frac{-1+\sqrt{5}}{2}$$

gilt. Alle diese Eigenschaften erhält man aus den Rekursionsformeln für b_m, c_m

$$b_{m+1} = \frac{1}{1+\frac{1}{1+b_m}} = \frac{1+b_m}{2+b_m}, \qquad b_0 = 0$$

und

$$c_{m+1} = \frac{1+c_m}{2+c_m}, \qquad c_0 = 1$$

Da beide Teilfolgen als beschränkte, monotone Folgen konvergieren und auch für sie gilt, daß $(-1+\sqrt{5})/2$ der einzig mögliche Grenzwert ist, folgt schließlich

$$\lim_{n\to\infty} a_n = (-1+\sqrt{5})/2.$$

4. Offene, abgeschlossene und kompakte Mengen

Am Ende des letzten Abschnitts haben wir gesehen, daß das Verhalten von stetigen Funktionen am *Rand* ihres Definitionsbereichs sehr interessant sein kann.
Denken Sie nur an die Funktionen

$$f(x) = \frac{1}{x}, \quad g(x) = \left(\frac{1}{x} - 1\right) \Big/ \left(\frac{2}{x} + 5\right), \quad h(x) = \sin\frac{1}{x}$$

die alle auf dem Intervall $(0, 1)$ definiert sind, aber bei Annäherung an den Randpunkt $x_0 = 0$ ganz verschiedenes Verhalten zeigen (unbeschränktes Wachstum, Konvergenz, rasante Oszillation). Der Begriff *Rand* für den Punkt x_0 ist dabei sinnvoll gewählt: x_0 selbst gehört zwar nicht zur Definitionsmenge $(0, 1)$, ist aber so dicht dran, daß ein *beliebig* kleiner Schritt nach rechts genügt, um in der Menge zu landen.
Im folgenden wollen wir uns Ränder von mehrdimensionalen Mengen anschauen. Sei dazu V ein Vektorraum mit Norm $||\cdot||$. Die Menge

$$B_r(\vec{v}) = \{\vec{v} \in V | \; ||\vec{v} - \vec{v}_0|| < r\}, \qquad r > 0$$

bezeichnen wir dann als Kugel mit dem Radius r um den Punkt $\vec{v}$. Eine wirkliche Kugel ist dies natürlich nur für den Fall $V = S$ (Zeigervektorraum) mit der elementargeometrischen Länge als Norm. Da uns dieser Spezialfall als Intuitionsgeber aber sowieso im Kopf herumspukt, übernehmen wir die Bezeichnung Kugel direkt in alle anderen normierten Vektorräume. Mit Hilfe von Kugeln $B_r(\vec{x}_0)$ wollen wir nun untersuchen, ob ein gegebener Punkt $\vec{x}_0 \in V$ ein Randpunkt einer vorgegebenen Menge $X \subset V$ ist. Wie im Eingangsbeispiel müssen wir dazu überprüfen, ob *beliebig* nahe bei $\vec{x}_0$ sowohl ein Punkt von X liegt als auch ein Element der Komplementärmenge X^c. Wie überprüft man aber die Eigenschaft *beliebig nahe*? Naja, für *jeden* noch so geringen Radius $r > 0$ muß in der Kugel $B_r(\vec{x}_0)$ halt ein Punkt der entsprechenden Menge zu finden sein.
Wir definieren die Menge ∂X aller Randpunkte von X (kurz, den Rand von X) also durch

$$\partial X = \{\vec{x}_0 \in V | B_r(\vec{x}_0) \cap X \neq \emptyset \text{ und } B_r(\vec{x}_0) \cap X^c \neq \emptyset \text{ für jedes } r > 0\}$$

Da ein Punkt $\vec{x}_0 \in V$ immer entweder in X oder in X^c liegen muß, ist eine der beiden Bedingungen einfach nachzuprüfen. Ist z. B. $\vec{x}_0 \notin X$, dann folgt $\vec{x}_0 \in X^c$ und da $\vec{x}_0$ in jeder Kugel $B_r(\vec{x}_0)$ mit $r > 0$ als Zentrum natürlich enthalten ist, folgt sofort

$$\vec{x}_0 \in B_r(\vec{x}_0) \cap X^c \neq \emptyset \text{ für jedes } r > 0.$$

In diesem Fall kann man sich bei der Untersuchung der Randpunkteigenschaft also darauf beschränken, die zweite Bedingung zu untersuchen. Dazu denkt man sich einen (sehr kleinen) Radius $r > 0$. Die Aufgabe besteht nun darin, in der zugehörigen kleinen Kugel $B_r(\vec{x}_0)$ mindestens einen Punkt $\vec{x}$ auszumachen, der zur Menge X dazugehört. Damit die Sache klar ist, sollte ein Punkt konkret angegeben werden (der von $\vec{x}_0, r$ und der Menge X abhängen wird). Wenn die Angabe des Punktes $\vec{x}$ für jedes $r > 0$ funktioniert, so ist auch die zweite Bedingung erfüllt und $\vec{x}_0$ ist damit ein Randpunkt. Findet man dagegen ein $\overline{r} > 0$, so daß in der Kugel $B_{\overline{r}}(\vec{x}_0)$ *kein* Punkt von X liegt, so scheidet $\vec{x}_0$ als Randpunkt aus. Der Grund, warum es genügt, nur an sehr kleine Kugeln zu denken, liegt daran, daß, wenn in einer kleinen Kugel um $\vec{x}_0$ ein Punkt von X gefunden wird, dann liegt dieser Punkt natürlich auch in jeder größeren Kugel um $\vec{x}_0$. Mit anderen Worten, die Randpunkteigenschaft ist eine *lokale* Eigenschaft, bei der es nur auf allernächste Nachbarschaftsverhältnisse ankommt.
Um unsere Definition auszuprobieren, schauen wir uns zunächst noch einmal unser Eingangsbeispiel $X = (0, 1)$ an. (Der Vektorraum ist hier $V = \mathbb{R}$ und $||\cdot||$ ist der Betrag.) Wie lautet die Menge ∂X in diesem Fall? Unsere Vermutung ist natürlich $\partial X = \{0, 1\}$. Prüfen wir dies einfach nach, wobei wir den für Mengenvergleiche üblichen Zugang wählen: Wir zeigen, daß $\{0, 1\} \subset \partial X$ und auch $\partial X \subset \{0, 1\}$ gilt, woraus die Gleichheit der beiden Mengen folgt. Beginnen wir mit $\{0, 1\} \subset \partial X$. Wieso ist $0 \in \partial X$? Zunächst stellen wir fest, daß $0 \notin X$ gilt und damit ist $B_r(0) \cap X^c \neq \emptyset$, da ja der Punkt 0 in der Schnittmenge für alle $r > 0$ enthalten ist (die Kugeln $B_r(0)$ sind übrigens in diesem Fall Intervalle $(-r, r)$, denn $|v - 0| < r$ ist genau für $v \in (-r, r)$ erfüllt). Umgekehrt können wir zu jedem $r > 0$ einen Punkt $x \in X$ finden, der in der Kugel $B_r(0)$ liegt. Hier ist jetzt eine Berechnungsformel für dieses x gefragt! Versuchen wir zunächst einfach $x = \frac{r}{2}$. Offensichtlich ist $\frac{r}{2} < r$ für $r > 0$ und damit $x \in B_r(0)$. Wenn r sehr klein ist, gilt natürlich auch $\frac{r}{2} \in (0, 1) = X$ und damit hätten wir einen Punkt in $B_r(0) \cap X$ gefunden. Damit die Formel wirklich für alle r stimmt, müssen wir nur die (langweiligen) großen Kugeln im Auge behalten. Ist nämlich $r > 2$, so fällt der Punkt $\frac{r}{2}$ wieder aus X heraus. Wie schon gesagt, interessieren uns große Kugeln aber gar nicht, wenn wir in kleinen Kugeln schon Punkte von X finden können. Wir modifizieren daher einfach die Formel für den Punkt x, z. B.

$$x = \frac{\min(r,1)}{2}, \qquad r > 0.$$

Sehen Sie, wie wir die lästigen großen Kugeln losgeworden sind? Wir ersetzen einfach das kleine r durch einen Ausdruck $\min\{r, r_0\}$. Dann bleibt die Formel für die wichtigen kleinen Kugeln ($r \leq r_0$) genau dieselbe, aber für große Kugeln ($r > r_0$) nehmen wir immer den gleichen Punkt. Sie sehen auch, daß die Auswahl des Punktes x etwas willkürlich ist. Natürlich hätten wir auch

$$\widetilde{x} = \frac{\min\{r,2\}}{8}, \quad \widehat{x} = \frac{\min\{r^2, \frac{1}{2}\}}{\pi}, \ldots$$

nehmen können. Wichtig ist nur, daß der Punkt sowohl in der Kugel $B_r(0)$ und der Menge X liegt.
Genauso finden wir heraus, daß $1 \in \partial X$ gilt. Hier nimmt man z. B. den Punkt

$$x = 1 - \frac{\min\{r,1\}}{2} \in B_r(1) \cap X$$

Damit gilt also $\{0,1\} \subset \partial X$. Für die umgekehrte Inklusion benutzen wir einen Widerspruchsbeweis. Angenommen, $\partial X \subset \{0,1\}$ wäre falsch. Dann gäbe es einen Punkt $x_0 \in \partial X$, der *nicht* 0 oder 1 ist. Es bleiben somit drei Fälle übrig: $x_0 \in (-\infty, 0)$, $x_0 \in (0,1)$ oder $x_0 \in (1,\infty)$. Im ersten Fall wäre x_0 strikt kleiner als Null, d. h. zwischen Null und x_0 ist genug Platz! Entfernt man sich nur um $\frac{|x_0|}{2}$ von x_0, so findet man keinen positiven Wert. Den größten Wert, den man entdeckt, ist nämlich wegen $|x_0| = -x_0$ für $x_0 < 0$ gerade $x_0 + \frac{|x_0|}{2} = x_0 - \frac{x_0}{2} = \frac{x_0}{2} < 0$. Mit anderen Worten, die Kugel $B_{\frac{|x_0|}{2}}(x_0)$ enthält nur Punkte aus X^c und folglich kann x_0 kein Randpunkt von X sein. Genauso sieht man, daß $x_0 \in (1,\infty)$ nicht möglich ist, da $B_{(x_0-1)/2}(x_0)$ ganz in X^c liegt. Schließlich ist auch der Fall $x_0 \in (0,1)$ ausgeschlossen, da wir hier eine Kugel um x_0 finden können, die ganz in X liegt. Ein passender Radius ist z. B.

$$r = \min\left\{\frac{x_0}{2}, \frac{1-x_0}{2}\right\}.$$

Wir sehen also, daß in keinem Fall $x_0 \in \partial X$ möglich ist, was im Widerspruch zur Annahme $x_0 \in \partial X$ steht. Damit ist also $\partial X \not\subset \{0,1\}$ ausgeschlossen und wir haben $\partial X \subset \{0,1\}$ gezeigt und insgesamt hat sich unsere Vermutung $\partial X = \{0,1\}$ bestätigt. Nach dieser ausführlichen Beschreibung sind Sie jetzt sicherlich in der Lage nachzurechnen, daß

$$\partial[1,5] = \{1,5\}, \quad \partial(-2,3] = \{-2,3\}, \quad \partial\mathbb{R} \setminus \{0\} = \{0\}$$

gilt. Eine unmittelbare Konsequenz unserer Randdefinition ist übrigens, daß die gesamte Menge und die leere Menge keine Randpunkte besitzen, also im vorliegenden Fall $\partial\mathbb{R} = \emptyset, \partial\emptyset = \emptyset$. Es gilt nämlich $B_r(x_0) \cap \emptyset = \emptyset$ für jedes x_0 und jedes $r > 0$ und auch $B_r(x_0) \cap \mathbb{R}^c = B_r(x_0) \cap \emptyset = \emptyset$. Um den Spezialfall der Teilmengen von $\mathbb{R}$ abzuschließen, stellen wir noch fest, daß unsere Randdefinition eine sehr natürliche Eigenschaft garantiert. Und zwar besitzt jede beschränkte Menge in $\mathbb{R}$ (also jede Menge, die an „beiden Seiten irgendwo aufhört“) einen Rand. Erstaunlicherweise ist dieser intuitiv sofort einleuchtende Sachverhalt gar nicht so offensichtlich. Um ihn nachzuweisen, müssen wir auf die Kontinuum-Eigenschaft von $\mathbb{R}$ zurückgreifen. Sei $M \subset \mathbb{R}$ also eine nicht leere, beschränkte Menge. Dann ist M insbesondere nach unten beschränkt und es existiert damit die größte untere Schranke, genannt $\inf M$. Untere Schranke bedeutet aber, daß $\inf M \leq m$ ist für alle $m \in M$. Links von $\inf M$ liegen also keine Punkte von M mehr und damit kristallisiert sich $\inf M$ auch schon als möglicher Randpunkt heraus. Die Frage ist nur noch, ob wir beliebig nahe rechts von $\inf M$ immer einen Punkt von M finden können. Das wiederum folgt aus der Tatsache, daß $\inf M$ die *größte* untere Schranke ist, d. h. jede Zahl $\inf M + s$ mit $s > 0$ ist *keine* untere Schranke mehr. Es gibt also zu jedem $S > 0$ ein $m \in M$, so daß $\inf M \leq m < \inf M + S$. Damit ist also für Elemente beliebig nahe rechts von $\inf M$ gesorgt. Formal funktioniert der Beweis jetzt so. Ein $r > 0$ ist vorgegeben. Dann ist

$$\inf M - \frac{r}{2} \in B_r(\inf M) \cap M^c$$

und es gibt ein $m \in M$ mit $\inf M \leq m < \inf M + r$, also

$$B_r(\inf M) \cap M \neq \emptyset$$

so daß

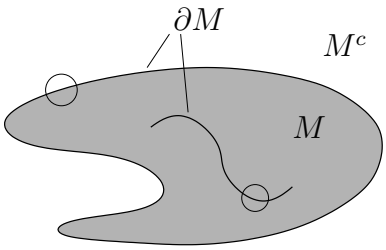
$$\inf M \in \partial M.$$

Genauso zeigt man, daß die kleinste obere Schranke zum Rand der Menge gehört und damit folgt für jede nicht leere, beschränkte Menge $M \subset \mathbb{R}$

$$\inf M, \sup M \in \partial M.$$

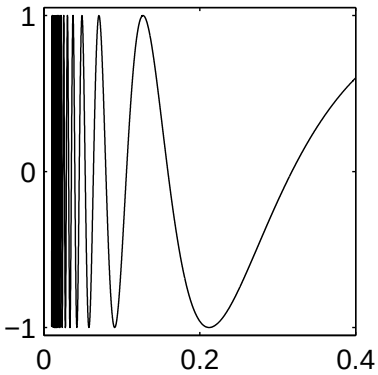
Es bleibt zu bemerken, daß der Rand einer beschränkten Teilmenge von $\mathbb{R}$ natürlich weitere Punkte neben dem Infimum und dem Supremum

beinhalten kann. So ist der Rand von $M = (-1, 0] \cup (1, 2]$ z. B. durch die Punkte $\partial M = \{-1, 0, 1, 2\}$ gegeben.
Beim Rand von mehrdimensionalen Mengen denkt man typischerweise an folgende Situation



Ein Punkt $\vec{x}$ gehört zum Rand von M, wenn jede Kugel um $\vec{x}$ sowohl Punkte aus M als auch aus M^c enthält. Gehören die Punkte auf der Kurve in M nicht zur Menge, so sind sie ebenfalls Randpunkte von M (sogenannter innerer Rand). Eine interessante zweidimensionale Menge ist durch den Graph der Funktion $x \to \sin \frac{1}{x}$ gegeben

$$M = \left\{ (x, \sin \frac{1}{x}) | x \in (0, 1) \right\} \subset \mathbb{R}^2$$



Die gesamte Menge M besteht hier aus Randpunkten! Der Rand ist sogar größer als die Menge selbst, denn der Punkt $(1, \sin 1)$ gehört dazu genauso wie die Punkte auf der y-Achse, deren Betrag kleiner oder gleich Eins ist

$$\partial M = M \cup \{(0, y) | y \in [-1, 1]\} \cup \{(1, \sin 1)\}$$

Als Beispiel in einem abstrakten Raum betrachten wir den Rand einer Kugel $B_R(\vec{0}), R > 0$ in einem normierten Vektorraum V. Da $B_R(\vec{0})$ alle Vektoren mit Länge *kleiner* als R enthält, liegt die Vermutung nahe, daß der Rand dieser Menge durch die Vektoren gegeben ist, die *exakt* die Länge R haben

$$S_R = \{\vec{v} \in V | \ ||\vec{v}|| = R\}$$

Für $\vec{v} \in S_R$ gilt tatsächlich, daß jede Kugel $B_r(\vec{v})$ Elemente aus $B_R(\vec{0})$ und $B_R(\vec{0})^c$ enthält. Konkrete Beispiele sind etwa

$$(1-\delta)\vec{v} \in B_R(\vec{0}) \cap B_r(\vec{v}), \qquad (1+\delta)\vec{v} \in B_R(\vec{v})^c \cap B_r(\vec{v})$$

wobei $\delta = \min\{1, \frac{r}{2R}\}$ vom Radius r der Testkugel abhängt. Damit gilt also $S_R \subset \partial B_R(\vec{0})$. Die umgekehrte Inklusion kann man durch die Beziehung $S_R^c \subset (\partial B_R(\vec{0}))^c$ nachweisen. Es gilt nämlich allgemein für zwei Mengen A, B, daß $A = B$ genau dann richtig ist, wenn $A \subset B$ und $A^c \subset B^c$ richtig ist. Sei also $\vec{v} \in S_R^c$, d. h. entweder $||\vec{v}|| < R$ oder $||\vec{v}|| > R$. Dann können wir leicht eine Kugel $B_\varepsilon(\vec{v})$ finden, die ganz zu $B_R(\vec{0})$ bzw. zum Komplement gehört, so daß $\vec{v}$ kein Randpunkt ist und damit $S_R^c \subset (\partial B_R(\vec{0}))^c$ gilt. Als Radius ε wählen wir einfach $\varepsilon = |\ ||\vec{v}|| - R|/2$.
Je nachdem, ob der Rand einer Menge gar nicht oder ganz zu der Menge dazugehört, sprechen wir von offenen bzw. abgeschlossenen Mengen.

Definition 13. *Sei V ein normierter Vektorraum und $X \subset V$. Dann heißt X offen, wenn $\partial X \cap X = \emptyset$. Die Menge X heißt abgeschlossen, falls $\partial X \cap X = \partial X$ gilt.*

Der Grund für die Namensgebung liegt am möglichen Grenzverhalten von Folgen in der Menge. Bei einer konvergenten Folge $(\vec{x}_n)_{n\in\mathbb{N}}$ in einer offenen Menge X (also $\vec{x}_n \in X$ für alle $n \in \mathbb{N}$) kann es passieren, daß der Grenzwert $\vec{x}_\infty = \lim_{n\to\infty} \vec{x}_n$ aus der Menge herausfällt. Die Menge X ist also offen bezüglich Grenzwertbildung. Natürlich kann sich der Grenzwert $\vec{x}_\infty$ nicht weit von den Folgengliedern $\vec{x}_n$ mit sehr großem n entfernen. Beliebig nahe bei $\vec{x}_\infty$ wird immer noch ein $\vec{x}_n$ zu finden sein, d. h. $\vec{x}_\infty$ ist nicht in X enthalten aber beliebig dicht an Elementen von X, d. h. $\vec{x}_\infty \in \partial X$. Wenn also der Grenzwert einer Folge in X die Menge verläßt, so wird er es höchstens bis auf den Rand schaffen. Diese Überlegung zeigt uns auch, wie wir eine Folge in einer offenen Menge basteln können, die konvergent ist und deren Grenzwert die Menge verläßt. Sei M dazu offen und nicht leer mit einem nichtleeren Rand. Wir wählen zunächst unseren gewünschten Grenzpunkt $\vec{x}_\infty \in \partial X$. Zu jedem Radius $R_n = \frac{1}{n}$ gibt es dann nach Definition von ∂X mindestens

ein Element im Durchschnitt $X \cap B_{r_n}(\vec{x}_\infty)$. Wir wählen ein Element aus und nennen es $\vec{x}_n$. So erhalten wir eine Folge von Punkten in X mit der Eigenschaft

$$||\vec{x}_n - \vec{x}_\infty|| < r_n = \frac{1}{n} \qquad n \in \mathbb{N}.$$

Nach Definition der Konvergenz gilt also tatsächlich $\lim_{n\to\infty} \vec{x}_n = \vec{x}_\infty$. Ist die Menge X dagegen abgeschlossen und $(\vec{x}_n)_{n\in\mathbb{N}}$ eine Folge in X die konvergiert, so sehen wir mit einer ähnlichen Argumentation, daß der Grenzwert $\vec{x}_\infty$ nun in X liegen muß. Läge der Grenzwert nämlich außerhalb von X, so wäre er insbesondere nicht auf dem Rand, da der Rand bei abgeschlossenen Mengen ja zur Menge dazugehört. Dann gibt es aber eine Kugel, um $\vec{x}_\infty$, die *nur* Elemente aus X^c enthält, was klar der Grenzwerteigenschaft widerspricht. Finden sich in einer kleinen Kugel um $\vec{x}_\infty \in X^c$ nur Punkte aus X^c, so kann $\vec{x}_\infty$ nicht beliebig gut mit Punkten aus X approximiert werden, da die Kugel immer einen bestimmten Mindestabstand erzwingt. Die Bezeichnung *abgeschlossen* bezieht sich also auf die Tatsache, daß die Menge bezüglich Granzwertbildung abgeschlossen ist — der Grenzwert kann nicht aus der Menge herausfallen.

Beachten Sie, daß die Klassifizierung in offene und abgeschlossene Mengen nicht vollständig ist. Neben der Eigenschaft, daß der Rand ganz oder gar nicht zur Menge gehört, gibt es nämlich noch viele andere Möglichkeiten. So könnten einige Randpunkte oder Randabschnitte zur Menge dazugehören und andere nicht, wie etwa bei dem Intervall $[0, 1)$, das weder offen noch abgeschlossen ist. Offene und abgeschlossene Mengen bilden also nur die beiden Extremfälle der Randzugehörigkeit. Dabei spielen abgeschlossene Mengen deshalb eine wichtige Rolle, weil, wie wir gesehen haben, Grenzwerte von konvergenten Folgen aus der Menge nicht herauskönnen. Offene Mengen sind dagegen wichtig, weil jeder Punkt in einer offenen Menge *berührungslos approximierbar* ist. Um diese Eigenschaft zu erklären, betrachten wir einen beliebigen Punkt $\vec{x}$ einer offenen Menge X. Offensichtlich gilt $\vec{x} \notin \partial X$, denn $X \cap \partial X = \emptyset$ für offene Mengen. Damit gibt es aber ein $\overline{r} > 0$, so daß $B_{\overline{r}}(\vec{x}) \cap X^c = \emptyset$, denn wäre für alle $r > 0$ der Schnitt der Kugel $B_r(\vec{x})$ mit dem Komplement X^c nicht leer, so würde jede Kugel um $\vec{x}$ Elemente von X^c und X enthalten ($\vec{x}$ gehört ja auch zur Kugel) und damit wäre dann doch $\vec{x} \in \partial X$, was wir ja oben ausgeschlossen haben. Diese Charakterisierung von offenen Mengen wollen wir festhalten.

Satz 7. *Sei V ein normierter Vektorraum und $X \subset V$. Dann ist X genau dann offen, wenn zu jedem $\vec{x} \in X$ ein Radius $r > 0$ existiert, so daß die gesamte Kugel $B_r(\vec{x})$ zu X gehört.*

In offenen Mengen hat also *jeder* Punkt eine umgebende Kugel, die ganz aus Punkten von X besteht (wie eine Schutzhülle gegen die Außenwelt X^c). Damit hat jeder Punkt einer offenen Menge aber auch in jeder Richtung sehr viele, sehr nahe Nachbarpunkte der Menge. Insbesondere kann jeder Punkt $\vec{x}$ einer offenen Menge X berührungslos approximiert werden. d. h. wir können eine Folge von Punkten $\vec{x}_n \in X$ finden, für die $\vec{x}_n \neq \vec{x}$ für alle $n \in \mathbb{N}$ gilt (kein Punkt $\vec{x}_n$ berührt $\vec{x}$) und die dem Punkt $\vec{x}$ trotzdem beliebig nahe kommen, d. h. $\lim_{n\to\infty} \vec{x}_n = \vec{x}$. Beispiele sind hier schnell gefunden. Wenn $B_r(\vec{x})$ eine Kugel mit positivem Radius $r > 0$ ist, die ganz zu X dazugehört und $\vec{e}$ ein beliebiger Vektor mit $||\vec{e}|| = 1$ ist (ein sogenannter Richtungsvektor), so ist $\vec{x}_n = \frac{1}{2n} r\vec{e} + \vec{x}$ eine Folge von Punkten, die ganz in $B_r(\vec{x})$ und damit in X liegt, die $\vec{x}$ nie berührt, die aber $\vec{x}$ beliebig nahe kommt (aus der Richtung $\vec{e}$ von $\vec{x}$ aus gesehen).
Warum die berührungslose Approximierbarkeit wichtig ist, soll hier kurz an einem eindimensionalen Beispiel erläutert werden. Sei dazu f eine reellwertige Funktion auf dem offenen Intervall $(0, 1)$ und x_0 ein beliebiger Punkt der Definitionsmenge. Dann gilt

$$f(x) = f(x_0) + \frac{f(x) - f(x_0)}{x - x_0} \cdot (x - x_0) \qquad x \neq x_0$$

Dieser simple Zusammenhang wird sehr nützlich, wenn wir den Differenzenquotienten $(f(x) - f(x_0))/(x - x_0)$ approximativ durch einen x-unabhängigen Wert ersetzen können, was möglich ist, wenn sich der Differenzenquotient für alle x sehr nahe bei x_0 praktisch nicht mehr ändert, d. h. falls der Grenzwert $c = \lim_{x\to x_0}(f(x) - f(x_0))/(x - x_0)$ existiert. In diesem Fall gilt dann

$$f(x) \approx f(x_0) + c(x - x_0)$$

wobei die Genauigkeit umso großer ist, je näher x an x_0 und damit $(f(x) - f(x_0))/(x - x_0)$ an c ist. Wir können also das Verhalten der Funktion f in einer ganzen Umgebung von x_0 gut vorhersagen und brauchen dazu nur eine Zahl, den Grenzwert des Differnzenquotienten. Zur Berechnung dieses Grenzwerts werden offensichtlich Folgen benötigt, die beliebig nahe an x_0 herankommen, die aber nie den Wert x_0 selbst annehmen (für $x = x_0$ ist der Differenzenquotient nämlich nicht definiert). Mit anderen Worten, der Punkt x_0 muß *berührungslos approximierbar* sein, um den Grenzwert des Differenzenquotienten zu ermitteln. Die Situation wird noch komplizierter, wenn die Funktion f von mehreren Variablen abhängt. Im dreidimensionalen Fall möchte man z. B. ganz analog eine approximative Beschreibung der Form

$$f(x_1, x_2, x_3) = f(\overline{x}_1, \overline{x}_2, \overline{x}_3) + a(x_1 - \overline{x}_1) + b(x_2 - \overline{x}_2) + c(x_3 - \overline{x}_3)$$

haben und braucht dazu die berührungslose Approximierbarkeit von $(\overline{x}_1, \overline{x}_2, \overline{x}_3)$ aus verschiedenen Richtungen, um die Grenzwerte a, b, c von gewissen Differenzenquotienten ausrechnen zu können. Diese Anforderung an die Definitionsmenge von f, daß jeder Punkt berührungslos aus allen Richtungen mit Elementen der Defintionsmenge approximierbar ist, ist sicherlich dann gegeben, wenn die Menge offen ist, und das ist letztlich unsere Motivation zur Betrachtung dieses Mengentyps.
Es bleibt zu erwähnen, daß offene und abgeschlossene Mengen in gewisser Weise komplementär sind. Das liegt daran, daß der Rand einer Menge X auch immer der Rand des Komplements X^c ist. Wenn also $\partial X \cap X = \emptyset$ gilt, also wenn X offen ist, dann ist offensichtlich $\partial X \subset X^c$ also $\partial X \cap X^c = \partial X$. Wegen $\partial X = \partial(X^c)$ zeigt dies aber, daß X^c abgeschlossen ist. Umgekehrt sieht man, daß das Komplement einer abgeschlossenen Menge immer offen ist.

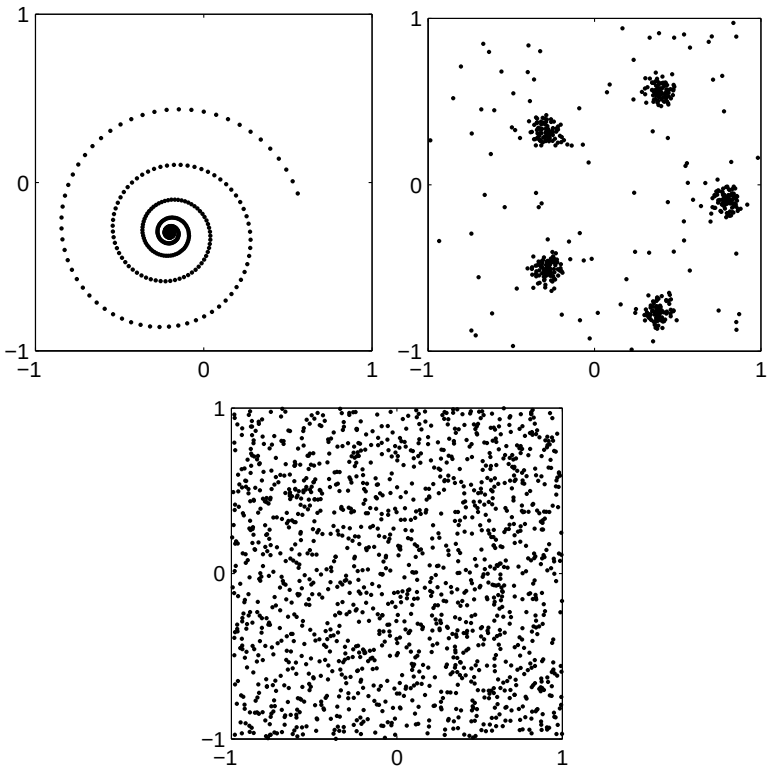
Satz 8. *Sei V ein normierter Vektorraum und $X \subset V$. Dann ist X abgeschlossen genau dann, wenn X^c offen ist.*

Neben offenen und abgeschlossenen Mengen spielen auch *beschränkte* Mengen eine wichtige Rolle. Dabei ist eine Menge M beschränkt, wenn sie in eine Kugel mit genügend großem Radius paßt.

Definition 14. *Sei V ein normierter Vektorraum. $X \subset V$ heißt beschränkt, wenn ein Radius $R > 0$ existiert, so daß X ganz in der Kugel $B_R(\vec{0})$ enthalten ist.*

Wie bei offenen und abgeschlossenen Mengen, betrachten wir das Verhalten von Folgen. Natürlich können sich die Folgenglieder in einer beschränkten Menge nicht sehr weit voneinander entfernen und außerdem müssen alle unendlich viele Folgenglieder in der beschränkten Menge Platz finden. Stellen Sie sich einmal vor, Sie müssen im Einheitsquadrat (eine beschränkte Menge in $\mathbb{R}^n$) unendlich viele Punkte einzeichnen. Irgendwann *müssen* die Punkte anfangen, sich an mindestens einer Stelle immer näher zu kommen. Dabei gibt es natürlich mehrere Szenarien, die unten in der Zeichnung angedeutet sind: (1) die Folgenglieder nähern sich immer mehr einem einzelnen Punkt an (d. h. die Folge konvergiert), (2) die Folgenglieder verdichten sich an mehreren Punkten (das klassische eindimensionale Beispiel für diese Situation ist die Folge $(-1)^n + \frac{1}{n}$, bei der sich alle Punkte mit gerader Nummer dem Punkt 1 nähern und alle Folgenglieder mit ungerader

Nummer dem Punkt -1); (3) es ist gar keine Konzentration der Folgenglieder zu erkennen, d. h. die Folge füllt die Menge gleichmäßig aus (d. h. aber, daß *jedem* Punkt der Menge gewisse Folgenglieder immer näher kommen).



Nach dieser Überlegung ist folgender Satz nicht mehr überraschend.

Satz 9. *Sei V ein endlich dimensionaler normierter Vektorraum und $(\vec{v}_n)_{n\in\mathbb{N}}$ eine beschränkte Folge (d. h. die Menge aller Folgenglieder ist beschränkt). Dann hat $(\vec{v}_n)$ mindestens eine konvergente Teilfolge, d. h. es gibt eine streng monotone funktion $M : \mathbb{N} \to \mathbb{N}$, so daß $(\vec{v}_{M(k)})_{k\in\mathbb{N}}$ konvergiert*

Für unsere Beispielfolge $v_n = (-1)^n + \frac{1}{n}$, die ja offensichtlich beschränkt ist

$$|v_n| \leq |(-1)^n| + |\frac{1}{n}| = 1 + \frac{1}{n} \leq 1 + 1 = 2,$$

wären zwei verschiedene konvergente Teilfolgen durch die „Herauspick-funktionen“ $M_g(k) = 2k$ und $M_u(k) = 2k + 1$ gegeben. Und zwar ist

$$v_{M_g(k)} = (-1)^{2k} + \frac{1}{2k} = 1 + \frac{1}{2k} \xrightarrow[k\to\infty]{} 1$$

und

$$v_{M_u(k)} = (-1)^{2k+1} + \frac{1}{2k+1} = -1 + \frac{1}{2k+1} \xrightarrow[k\to\infty]{} -1.$$

Bei unserer Überlegung zum obigen Satz haben wir benutzt, daß in einer beschränkten Menge nur endlich viel Platz ist. Dieses anschauliche Argument ist sicherlich richtig im uns umgebenden dreidimensionalen Raum, der unsere Intuition leitet. Es bleibt auch richtig in allen endlich dimensionalen Räumen. Wenn wir allerdings zu unendlich dimensionalen Vektorräumen übergehen, dann gibt es sozusagen unendlich viele unabhängige Raumrichtungen und dadurch kann selbst in beschränkten Mengen unendlich viel Platz sein. Man kann dann unendlich viele Folgenglieder in die unendlich vielen Raumrichtungen verteilen, *ohne* daß sich die Punkte irgendwo verdichten. Die Voraussetzung der endlichen Dimension ist also wesentlich für die Richtigkeit der Aussage.
Um dies durch ein Beispiel zu belegen, betrachten wir den unendlich dimensionalen Vektorraum aller beschränkten reellen Zahlenfolgen

$$V = \mathcal{F}_b(\mathbb{N}, \mathbb{R}) = \{a : \mathbb{N} \to \mathbb{R} | a(\mathbb{N}) \text{ beschränkt } \}$$

Als Norm einer beschränkten Folge a nehmen wir die halbe Breite r des kleinsten Intervalls $[-r, r]$, in dem alle Folgenglieder a_n enthalten sind, also

$$||a|| = \sup_{n\in\mathbb{N}} |a_n|$$

Eine Folge im Vektorraum V ist nun ein etwas verzwicktes Objekt, denn jedes Folgenglied ist als Element von V wieder eine komplette Folge mit Werten in $\mathbb{R}$. Betrachten wir als Beispiel die Folge, bei der das k-te Folgenglied durch

$$v_k(n) = \begin{cases} 1 & n = k \\ \\ 0 & n \neq k \end{cases} \qquad\qquad n \in \mathbb{N}$$

gegeben ist, also

$$\begin{array}{lcl} v_1 & : & 1,0,0,0,0, \ldots \\ v_2 & : & 0,1,0,0,0, \ldots \\ v_3 & : & 0,0,1,0,0 \ldots \end{array}$$

Die Folge $(v_k)_{k\in\mathbb{N}}$ ist beschränkt, denn $||v_k|| = 1$ für jedes $k \in \mathbb{N}$. Nehmen wir an, daß der obige Satz für diese Folge richtig wäre. Dann gäbe es eine in V konvergente Teilfolge $(v_{M(k)})_{n\in\mathbb{N}}$, die gegen ein $v_\infty \in V$ konvergieren würde. Insbesondere gilt mit der Dreiecksungleichung

$$||v_{M(k)} - v_{M(k+1)}|| \leq ||v_{M(k)} - v_\infty|| + ||v_\infty - v_{M(k+1)}|| \xrightarrow[k\to\infty]{} 0$$

Andererseits hat die Differenzfolge $v_{M(k)} - v_{M(k+1)}$ folgende Struktur

$$0, 0, \ldots, 0, 1, 0, \ldots, 0, -1, 0, 0, \ldots$$

und damit gilt

$$||M_{(k)} - v_{M(k+1)}|| = 1 \qquad \text{für alle } k \in \mathbb{N}$$

was im Widerspruch zu $||v_{M(k)} - v_{M(k+1)}|| \longrightarrow 0$ steht. Damit kann also $(v_k)_{k\in\mathbb{N}}$ keine konvergente Teilfolge besitzen und wir sehen, daß die Voraussetzung der endlichen Dimension für den obigen Satz von zentraler Bedeutung ist.
Dennoch möchte man auch in unendlich dimensionalen Räumen mit Mengen arbeiten, in denen jede Folge mindestens eine konvergente Teilfolge hat. Wie wir gleich sehen werden, nehmen stetige reellwertige Funktionen auf solchen Mengen ihren Maximal- und Minimalwert an. Diese Eigenschaft ist offensichtlich wichtig für die Lösung von Optimierungsaufgaben. Da die Beschränktheit einer Menge im allgemeinen nicht ausreicht, um die Existenz von konvergenten Teilfolgen zu garantieren und auch sonst keine elementare Eigenschaft dies garantiert, führen wir einen neuen Namen für solche Mengen ein.

Definition 15. *Sei V ein normierter Vektorraum. Eine Teilmenge $X \subset V$ heißt kompakt, wenn jede Folge in X eine in X konvergente Teilfolge hat.*

Die Kompaktheit beinhaltet übrigens die Beschränktheit der Menge. Wäre nämlich eine kompakte Menge $\mathcal{K}$ unbeschränkt, so könnte man eine Folge $(\vec{x}_n)_{n\in\mathbb{N}}$ in $\mathcal{K}$ konstruieren, für die $||\vec{x}_n|| \geq n$ gilt. Eine solche Folge kann aber keine konvergente Teilfolge haben, da für jede Teilauswahl $(\vec{x}_{M(k)})_{n\in\mathbb{N}}$ die Folgenglieder $\vec{x}_{M(k)}$ betragsmäßig immer größer werden und folglich nicht gegen einen festen Vektor $\vec{x}_\infty \in \mathcal{K}$ konvergieren können.
Jede kompakte Menge ist also zwangsläufig beschränkt. Außerdem sind kompakte Mengen auch abgeschlossen. Um dies zu sehen, wählen wir einen beliebigen Punkt $\vec{x}_\infty \in \partial\mathcal{K}$ des Randes einer kompakten Menge $\mathcal{K}$ aus. Dann konstruieren wir eine Folge $(\vec{x}_n)_{n\in\mathbb{N}}$ von Punkten aus $\mathcal{K}$,

die gegen $\vec{x}_\infty$ konvergieren (genauso wie wir das im Fall offener Mengen getan haben). Da die Menge $\mathcal{K}$ kompakt ist, gibt es eine Teilfolge $(\vec{x}_{M(k)})_{k\in\mathbb{N}}$, die gegen einen Punkt $\vec{y} \in \mathcal{K}$ konvergiert. Nun konvergieren aber alle Teilfolgen einer konvergenten Folge gegen den gleichen Grenzwert und folglich ist $\vec{x}_\infty = \vec{y} \in \mathcal{K}$. Der Rand $\partial\mathcal{K}$ gehört damit zu $\mathcal{K}$, d. h. $\mathcal{K}$ ist abgeschlossen.
Umgekehrt gilt in endlich dimensionalen Vektorräumen, daß beschränkte und abgeschlossene Mengen bereits kompakt sind. Ist $(\vec{x}_n)_{n\in\mathbb{N}}$ eine Folge in einer beschränkten Menge, so wissen wir, daß $(\vec{x}_n)_{n\in\mathbb{N}}$ eine konvergente Teilfolge besitzt (das gilt ja in endlich dimensionalen Räumen). Ist die Menge zusätzlich abgeschlossen, so kann der Grenzwert der Teilfolge nicht aus der Menge herausfallen, d. h. die Teilfolge konvergiert in der Menge, die damit kompakt ist.

Satz 10. *Sei V ein endlich dimensionaler normierter Vektorraum und $\mathcal{K} \subset V$. Dann ist $\mathcal{K}$ kompakt genau dann, wenn $\mathcal{K}$ beschränkt und abgeschlossen ist.*

Insbesondere ist die abgeschlossene Einheitskugel

$$\{\vec{x} \in V |\ ||\vec{x}|| \leq 1\}$$

in jedem endlich dimensionalen Vektorraum kompakt. In unendlich dimensionalen Räumen ist das dagegen nicht der Fall (wie wir z. B. im Folgenraum gesehen haben). Kriterien für Kompaktheit sind dort wesentlich schwieriger zu finden und in jedem Einzelfall neu zu suchen.
Schauen wir uns jetzt einmal an, wie stetige Funktionen mit kompakten Mengen zusammenarbeiten.
Für stetige Funktionen $f : V \to W$ zwischen zwei normierten Vektorräumen gilt nun, daß die Bilder von kompakten Mengen wieder kompakt sind.
Um dies zu zeigen, nehmen wir an, daß $\mathcal{K} \subset V$ kompakt ist. Für jede Folge $(\vec{y}_n)_{n\in\mathbb{N}}$ im Bild $f(\mathcal{K})$ ist dann nachzuweisen, daß eine in $f(\mathcal{K})$ konvergente Teilfolge existiert. Zunächst wissen wir, daß zu jedem $\vec{y}_n \in f(\mathcal{K})$ ein Urbild $\vec{v}_n \in \mathcal{K}$ existiert, so daß $\vec{y}_n = f(\vec{x}_n)$ gilt. Da $\mathcal{K}$ kompakt ist, gibt es zu $(\vec{x}_n)_{n\in\mathbb{N}}$ eine konvergente Teilfolge $(\vec{x}_{M(k)})_{k\in\mathbb{N}}$ mit $\vec{x}_{M(k)} \longrightarrow \vec{x}_\infty \in \mathcal{K}$ für $k \to \infty$. Jetzt ist aber f stetig und vertauscht deshalb mit Grenzübergängen, so daß

$$\vec{y}_{M(k)} = f(\vec{x}_{M(k)}) \xrightarrow[k\to\infty]{} f(\vec{x}_\infty) \in f(\mathcal{K})$$

Damit haben wir eine Teilfolge $(\vec{y}_{M(k)})_{k\in\mathbb{N}}$ gefunden, die in der Menge $f(\mathcal{K})$ konvergiert und folglich ist $f(\mathcal{K})$ eine kompakte Menge.
Für den Spezialfall $W = \mathbb{R}$ erhalten wir sofort die oben bereits erwähnte Eigenschaft.

Satz 11. *Sei V ein normierter Vektorraum und $f : V \to \mathbb{R}$ stetig. Ist $\mathcal{K} \subset V$ kompakt, so nimmt f auf $\mathcal{K}$ sein Maximum und sein Minimum an. Es gibt also Punkte $\vec{x}_{\min}, \vec{x}_{\max} \in \mathcal{K}$, so daß*

$$f(\vec{x}_{\min}) = \inf_{\vec{x}\in\mathcal{K}} f(\vec{x}), \qquad f(\vec{x}_{\max}) = \sup_{\vec{x}\in\mathcal{K}} f(\vec{x}).$$

Der Nachweis dieser Aussage beruht auf einer Kombination mehrerer Vorüberlegungen. Zunächst ist das Bild $f(\mathcal{K})$ eine kompakte Teilmenge von $\mathbb{R}$ und damit beschränkt und abgeschlossen. Für beschränkte Mengen in $\mathbb{R}$ haben wir aber gesehen, daß sowohl Infimum als auch Supremum zum Rand der Menge gehören. also

$$\inf f(\mathcal{K}), \sup f(\mathcal{K}) \in \partial f(\mathcal{K})$$

Andererseits besagt die Abgeschlossenheit, daß der Rand von $f(\mathcal{K})$ in der Menge $f(\mathcal{K})$ enthalten ist, also insbesondere

$$\inf f(\mathcal{K}), \sup f(\mathcal{K}) \in f(\mathcal{K})$$

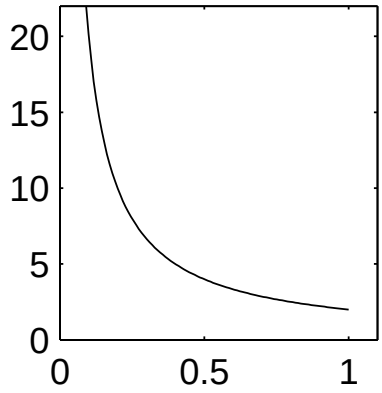
Andererseits besagt die Abgeschlossenheit, daß der Rand von $f(\mathcal{K})$ in der Menge $f(\mathcal{K})$ enthalten ist, also insbesondere

$$\inf f(\mathcal{K}), \sup f(\mathcal{K}) \in f(\mathcal{K})$$

Damit gibt es dann aber auch mindestens zwei Punkte $\vec{x}_{\min}, \vec{x}_{\max} \in \mathcal{K}$ mit der gewünschten Eigenschaft.
Daß stetige Funktionen auf nicht kompakten Mengen weder Maximum noch Minimum annehmen müssen, sieht man leicht an einem Beispiel

$$f(x) = \frac{2}{x}, \quad x \in (0,1)\cdot$$

Die Menge $(0,1)$ ist nicht abgeschlossen und damit nicht kompakt. Das Infimum des Bildes ist offensichtlich $\inf_{x\in(0,1)} f(x) = 2$ und wird für kein $x \in (0,1)$ als Bild angenommen. Das Supremum existiert nicht einmal, da die Funktion für $x \to 0$ unbeschränkt wächst.



Auf der kompakten Menge $[1/10, 1]$ greift dagegen unser Satz und es gilt

$$f(1) = \inf f(\mathcal{K}) = 2, \quad f\left(\frac{1}{10}\right) = \sup f(\mathcal{K}) = 20.$$

Als Anwendung des Satzes über Maximum und Minimum wollen wir uns die Äquivalenz von Normen in endlich dimensionalen Vektorräumen anschauen. Vielleicht haben Sie sich ja schon gefragt, ob es passieren kann, daß eine Folge $(\vec{x}_m)_{m\in\mathbb{N}}$ von n-Tupeln in einer Norm konvergiert, aber in einer anderen Norm vielleicht nicht. Normen auf $\mathbb{R}^n$ gibt es ja viele, z. B.

$$||\vec{x}||_\infty = \max_{i=1,\ldots,n} |x_i|, \quad ||\vec{x}||_1 = \sum_{i=1}^{n} |x_i|, \quad ||\vec{x}||_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

und vielleicht macht es ja bei Konvergenz- und Stetigkeitsuntersuchungen einen Unterschied, ob die Norm $||\cdot||_\infty$ benutzt wird oder eine andere Norm $||\cdot||_1, ||\cdot||_2$, etc. Daß dies nicht der Fall ist, hat viel mit Kompaktheit zu tun. Nehmen wir an, $||\cdot||$ ist irgendeine Norm auf $\mathbb{R}^n$. Dann gilt für einen Vektor $\vec{x} = (x_1, \ldots, x_n)$

$$\begin{aligned} ||\vec{x}|| = ||x_1\vec{e}_1 + \ldots + x_n\vec{e}_n|| &\leq |x_1|\ ||\vec{e}_1|| + \ldots + |x_2|\ ||\vec{e}_n|| \\ &\leq \max_{i=1,\ldots,n} |x_i| \sum_{i=1}^{n} ||\vec{e}_i|| = C||\vec{x}||_\infty \end{aligned}$$

wobei $\vec{e}_i$ die kanonischen Basisvektoren sind und

$$C = \sum_{i=1}^{n} ||\vec{e}_i||$$

eine Konstante, die nur von der Norm $||\cdot||$ abhängt.
Die gewonnene Abschätzung $||\vec{x}|| \leq C||\vec{x}||_\infty$ bedeutet aber, daß die Funktion $||\cdot|| : \mathbb{R}^n \to \mathbb{R}$ Lipschitz-stetig bezüglich der Norm $||\cdot||_\infty$ auf $\mathbb{R}^n$ ist mit Lipschitz-Konstante C. Es gilt nämlich

$$| \, ||\vec{x}|| - ||\vec{y}|| \, | \leq ||\vec{x} - \vec{y}|| \leq C||\vec{x} - \vec{y}||_\infty$$

wobei die erste Abschätzung für jede Norm gilt, wie wir bereits bei den Stetigkeitsuntersuchungen gesehen haben. Außerdem wissen wir, daß der Rand der Einheitskugel

$$\mathcal{K} = \{\vec{x} \in \mathbb{R}^n | \; ||\vec{x}||_\infty = 1\}$$

beschränkt und abgeschlossen und damit kompakt ist. Insbesondere gibt es $\vec{x}_{\min} \in \mathcal{K}$ mit der Eigenschaft

$$||\vec{x}_{\min}|| = \inf_{\vec{x} \in \mathcal{K}} ||\vec{x}||$$

denn $||\cdot||$ ist eine stetige Funktion, wie wir oben gesehen haben. Da der Nullvektor nicht in der Menge $\mathcal{K}$ enthalten ist, muß folglich $\vec{x}_{\min} \neq \vec{0}$ gelten und damit $||\vec{x}_{\min}|| = c > 0$. Schließlich liegt für einen beliebigen Vektor $\vec{0} \neq \vec{x} \in \mathbb{R}^n$ der normale Vektor $\vec{x}/||\vec{x}||_\infty$ in der Menge $\mathcal{K}$ und es gilt

$$C = ||\vec{x}_{\min}|| \leq \left|\left| \frac{\vec{x}}{||\vec{x}||_\infty} \right|\right| = \frac{||\vec{x}||}{||\vec{x}||_\infty}$$

Zusammen mit der obigen Abschätzung erhalten wir damit

$$c||\vec{x}||_\infty \leq ||\vec{x}|| \leq C||\vec{x}||_\infty$$

für zwei positive Konstanten c, C. Geometrisch bedeutet diese Abschätzung, daß jede $||\cdot||$-Kugel in eine genügend große $||\cdot||_\infty$-Kugel paßt, aber auch eine hinreichend kleine $||\cdot||_\infty$ Kugel umfaßt. Wir sagen in diesem Fall, die Normen $||\cdot||$ und $||\cdot||_\infty$ sind *äquivalent*.

Definition 16. *Sei V ein Vektorraum. Zwei Normen $||\cdot||$ und $||\cdot||_\sim$ auf V heißen äquivalent, wenn positive Konstanten c, C existieren, so daß*

$$c||\vec{v}||_\sim \leq ||\vec{v}|| \leq C||\vec{v}||_\sim \qquad \textit{für alle } \vec{v} \in V.$$

Der Begriff Äquivalenz für diesen Fall erklärt sich dadurch, daß Konvergenzaussagen in beiden Normen die gleichen sind (und damit auch Stetigkeitsaussagen, die sich ja mit konvergenten Folgen beschreiben lassen).
Konvergiert z. B. eine Folge $(\vec{x}_n)_{n\in\mathbb{N}}$ in einem Raum V bezüglich der Norm $||\cdot||$ gegen den Punkt $\vec{x}_\infty$, d. h. zu jedem $\varepsilon > 0$ gibt es ein $N \in \mathbb{N}$, so daß

$$||\vec{x}_n - \vec{x}_\infty|| < \varepsilon \quad \text{für alle } n \geq N_\varepsilon,$$

so konvergiert die Folge auch in jeder äquivalenten Norm. Gilt nämlich $||\cdot||_\sim \leq \frac{1}{c}||\cdot||$, so wählt man zu gegebenem $\varepsilon > 0$ den Index $\widetilde{N}_\varepsilon = N_{c\cdot\varepsilon}$ und hat dann für alle $n \geq \widetilde{N}_\varepsilon$

$$||\vec{x}_n - \vec{x}_\infty||_\sim \leq \frac{1}{c}||\vec{x}_n - \vec{x}_\infty|| < \frac{c\cdot\varepsilon}{c} = \varepsilon,$$

also Konvergenz bezüglich der $||\cdot||_\sim$ Norm. Genauso folgt mit der Abschätzung $||\cdot|| \leq C||\cdot||_\sim$, daß jede $||\cdot||_\sim$-konvergente Folge auch $||\cdot||$-konvergent ist.
Im Vektorraum $\mathbb{R}^n$ (und allgemeiner in allen endlich dimensionalen Vektorräumen) gilt nun, daß *alle* Normen zueinander äquivalent sind. Wir müssen also bei Konvergenz und Stetigkeit nicht genau nachfragen, auf welche Norm sich die jeweilige Eigenschaft bezieht. Den Nachweis dieser Eigenschaft haben wir eigentlich schon erbracht. Sind $||\cdot||$ und $||\cdot||_\sim$ zwei Normen auf $\mathbb{R}^n$, so gilt wegen der Äquivalenz zur $||\cdot||_\infty$-Norm

$$c||\cdot||_\infty \leq ||\cdot|| \leq C||\cdot||_\infty$$
$$\widetilde{c}||\cdot||_\infty \leq ||\cdot||_\sim \leq \widetilde{C}||\cdot||_\infty$$

und damit auch

$$\frac{c}{\widetilde{C}}||\vec{x}||_\sim \leq c||\vec{x}||_\sim \leq ||\vec{x}|| \leq C||\vec{x}||_\infty \leq \frac{C}{\widetilde{c}}||\vec{x}||_\sim$$

Mit den Konstanten $d = c/\widetilde{C}$ und $D = C/\widetilde{c}$ erhalten wir also die Äquivalenz zwischen $||\cdot||$ und $||\cdot||_\sim$

$$d||\vec{x}||_\sim \leq ||\vec{x}|| \leq D||\vec{x}||_\sim$$

Mit exakt der gleichen Argumentation kann man auch zeigen, daß alle Normen auf $\mathbb{R}^{n\times 1}$ äquivalent sind und da $\mathbb{R}^{n\times 1}$ als Koordinatenräume beliebiger endlich dimensionaler Vektorräume auftreten, kann man die Aussage auch auf allgemeine Räume übertragen (Übungsaufgabe).

Satz 12. *Seien $||\cdot||, ||\cdot||_\sim$ zwei Normen auf einem endlich dimensionalen Vektorraum. Dann sind die beiden Normen äquivalent.*

Als Folgerung aus diesem Satz können wir nachrechnen, daß *alle* linearen Abbildungen auf endlich dimensionalen Vektorräumen Lipschitzstetig sind. Sei dazu $L : V \to W$ gegeben und $\vec{x} \neq \vec{y}$. Es gilt

$$\begin{aligned} ||L\vec{x} - L\vec{y}||_W &= ||L(\vec{x} - \vec{y})||_W = \frac{||L(\vec{x} - \vec{y})||_W}{||\vec{x} - \vec{y}||_V} ||\vec{x} - \vec{y}||_V \\ &= \left|\left| L\left(\frac{\vec{x} - \vec{y}}{||\vec{x} - \vec{y}||_V}\right)\right|\right|_W ||\vec{x} - \vec{y}||_V \end{aligned}$$

Stellen wir den Einheitsvektor $\vec{e} = (\vec{x} - \vec{y})/||\vec{x} - \vec{y}||_V$ in einer Basis $(\vec{b}_1, \ldots, \vec{b}_n)$ von V dar, also

$$\vec{e} = y_1\vec{b}_1 + \ldots + y_n\vec{b}_n$$

so erhalten wir

$$||L(\vec{e})||_W = ||\sum_{i=1}^{n} y_i L(\vec{b}_i)||_W \leq \max_{i=1,\ldots,n} |y_i| \sum_{i=1}^{n} ||L(\vec{b}_i)||_W$$

Da

$$||x_1\vec{b}_1 + \ldots + x_1\vec{b}_n||_{\infty,V} = \max_{i=1,\ldots,n} |x_i|$$

eine Norm auf V ist, liefert die Normäquivalenz eine Konstate C mit

$$||\vec{e}||_{\infty,V} \leq C||\vec{e}||_V = C$$

und somit

$$||L(\vec{e})||_W \leq C\sum_{i=1}^{n} ||L(\vec{b}_i)||_W = \widetilde{C} \quad \text{falls } ||\vec{e}||_V = 1$$

Für unsere Stetigkeitsabschatzung der linearen Abbildung folgt damit

$$||L(\vec{x} - \vec{y})||_W \leq \widetilde{C}||\vec{x} - \vec{y}||_V$$

d. h. L ist Lipschitz-stetig mit Konstante $\widetilde{C}$.
Wenn Sie den Beweis sorgfältig anschauen, dann sehen Sie, daß die Annahme der endlichen Dimension nur bei der Abschätzung des *maximalen Verzerrungsfaktors*

$$\alpha = \sup\{||L(\vec{e})||_W \mid ||\vec{e}||_V = 1\}$$

benutzt wurde. Der Faktor α beschreibt, wie stark die Länge eines Bildvektors $L(\vec{x})$ maximal von der Länge des Urbildes $\vec{x}$ abweicht. Im Fall $\vec{x} \neq \vec{0}$ gilt nämlich

$$\frac{||L(\vec{x})||_W}{||\vec{x}||_V} = \left|\left| L\left(\frac{\vec{x}}{||\vec{x}||_V}\right)\right|\right|_W \leq \alpha.$$

Wir haben also gesehen, daß linear Abbildungen auf endlich dimensionalen Vektorräumen stets einen endlichen Verzerrungsfaktor haben.
In unendlich dimensionalen Räumen muß das nicht mehr der Fall sein. Dort ist die Beschränktheit des Verzerrungsfaktors eine eigenständige Qualität der linearen Abbildung und man bezeichnet Abbildungen, die diese Eigenschaft haben, kurz als *beschränkt*. Auf der Menge aller beschränkten lineare Abbildungen zwischen zwei Vektorräumen V und W bildet der maximale Verzerrungsfaktor sogar eine Norm

$$||L|| = \sup_{||\vec{e}||_V=1} ||L(\vec{e})||_W = \sup_{\vec{v}\neq\vec{0}} \frac{||L(\vec{v})||_W}{||\vec{v}||_V}$$

Mit der Bezeichnung $||L||$ für den maximalen Verzerrungsfaktor von L gilt dann

$$||L(\vec{x} - \vec{y})||_W \leq ||L||\ ||\vec{x} - \vec{y}||_V.$$

Wir fassen unsere Betrachtungen zur Stetigkeit von linearen Abbildungen in einem abschließenden Satz zusammen.

Satz 13. *Seien V, W normierte Vektorräume und sei $L : V \to W$ eine lineare Abbildung. Dann gilt*

(i) *Ist L beschränkt, so ist L stetig.*
(ii) *Ist V endlich dimensional, so ist L beschränkt.*

5. Differenzierbare Funktionen

Zu Beginn dieses Kapitels haben wir uns überlegt, daß die Vorhersagbarkeit von gewissen Zusammenhängen in unserer Umwelt sehr wichtig und teilweise sogar überlebensnotwendig ist. Betrachten wir noch ein weiteres Beispiel. Stellen Sie sich vor, Sie wollen eine Straße überqueren und ein Fahrzeug nähert sich von rechts. Für ein sicheres Überqueren ist offensichtlich der Zusammenhang zwischen der zurückgelegten Entfernung des Fahrzeugs (bezüglich Ihrer Position) und der dafür benötigten Zeit von zentraler Bedeutung. Nennen wir diese Funktion $t \mapsto f(t)$, wobei das Argument t für die Zeit steht. In dem Moment t_0, wo Sie das Fahrzeug erblicken, erhalten Sie Information über die Entfernung $f(t_0)$.

Gleichzeitig wissen Sie aber auch, daß der vorliegende Zusammenhang vorhersagbar ist. Wäre die Funktion f unstetig so wäre es nämlich vollkommen nutzlos, überhaupt nach einem Fahrzeug zu schauen! Die Information $f(t_0)$ sagt nämlich bei unstetigen Funktionen *gar nichts* über benachbarte Funktionswerte $f(t_0 + w)$ aus. Würde sich das Fahrzeug unstetig bewegen, könnte es im nächsten Moment schon bei Ihnen sein und Sie überrollen, oder aber plötzlich links von Ihnen sein ohne daß es jemals direkt vor Ihnen vorbeigefahren ist. Die Straße überqueren, wäre dann ein Lotteriespiel und Kindern müßte man nicht jahrelang predigen, erst links, dann rechts und dann wieder links zu schauen, da dies ja sowieso keinen Sinn hätte.
Aus unserer täglichen Erfahrung wissen wir aber, daß Fahrzeuge sich vorhersagbar, also stetig bewegen. Der tatsächliche Prozeß des Straßeüberquerens nutzt aber noch eine tieferliegende Eigenschaft des Fahrzeugverhaltens aus. Wüßten wir nur, daß die Entfernungsfunktion stetig ist, so würde der Blick zum Fahrzeug uns nur sagen, daß die Entfernung $f(t_0 + w)$ zu einem späteren Zeitpunkt nicht sehr stark von der Entfernung $f(t_0)$ zum Zeitpunkt t_0 abweichen wird, also etwa: In einer Millisekunde wird sich das Fahrzeug nicht wesentlich annähern. Das ist zwar beruhigender als von springenden Fahrzeugen ausgehen zu müssen, aber immer noch nicht wirklich nützlich, da das eigene Überqueren der Straße eine etwas längere Zeit in Anspruch nimmt, während der sich das Fahrzeug deutlich bewegen wird.
Tatsächlich nutzen wir für den alltäglichen Vorgang des Strasseüberquerens die *Differenzierbarkeit* der Fahrzeugbewegung aus. Wenn wir das Fahrzeug anpeilen, nehmen wir dabei nicht nur die Entfernung $f(t_0)$ wahr, sondern beobachten auch noch die Entfernung eine kurze Zeit später, also $f(t_0 + w)$ mit $w > 0$, und entwickeln dabei ein Gefühl für die *Änderungsrate* der Entfernung $(f(t_0+w)-f(t_0))/w$, die aussagt, wie stark sich die Entfernung des Fahrzeugs pro Zeiteinheit ändert. (Wir ermitteln also ungefähr die Geschwindigkeit des Fahrzeugs.) Dann schlagen wir noch einen Sicherheitsbonus auf und erhalten als konservative Abschätzung für die Änderungsrate einen Wert V, der hier negativ ist, weil die Entfernung mit der Zeit abnimmt. Wie weit das Fahrzeug dann nach der Zeit T entfernt sein wird, ist also ungefähr durch $f(t_0) + V \cdot T$ gegeben, und wir können nun unsere eigene Geschwindigkeit für das Überqueren so wählen, daß in der dafür benötigten Zeit T die Entfernung $f(t_0)+TV$ nicht zu klein wird. Dieser Trick funktioniert offensichtlich nur dann, wenn sich die Entfernungsfunktion $f(t)$ vernünftig durch die affin lineare Funktion

$$g(t) = f(t_0) + V(t - t_0)$$

annähern läßt, denn als Berechnungsgrundlage benutzen wir ja g anstelle der eigentlichen Funktion f. Dabei genügt es natürlich, wenn die Näherung nur für Zeitpunkte t in der Nähe von t_0 tauglich ist. Welche Entfernung das Fahrzeug eine Stunde später hat, wird sich wohl nicht durch die Funktion g, d. h. durch die abgeschätzte Geschwindigkeit V beschreiben lassen, aber das wollen wir ja auch gar nicht.
Allgemein nennen wir Funktionen, die sich (hinreichend nahe) bei einem Punkt ihres Definitionsgebiets gut durch eine affin lineare Funktion approximieren lassen *differenzierbar* in diesem Punkt. Hängt die Funktion von mehr als einer Variablen ab (z. B. ein ortsabhängiges Temperaturfeld), dann wird die affin lineare Funktion wie gewohnt durch eine lineare Abbildung und einen konstanten Vektor beschrieben. In unserem eindimensionalen Beispiel ist die lineare Funktion dabei $L(t) = V \cdot t$ und der konstante Wert $f(t_0) - Vt_0$. Die Differenzierbarkeit der Entfernungsfunktion können wir etwas salopp mit

$$f(t_0 + w) \approx f(t_0) + L(w), \qquad w \text{ klein}$$

beschreiben. In der Verallgemeinerung auf vektorabhängige und vektorwertige Funktionen $f : X \to Y$ bedeutet Differenzierbarkeit in einem Punkt $\vec{x}_0 \in X$, daß entsprechend

$$f(\vec{x}_0 + \vec{w}) \approx f(\vec{x}_0) + L(\vec{w}), \qquad ||\vec{w}||_x \text{ klein}$$

gilt, wobei $L : X \to Y$ eine lineare Abbildung zwischen den Vektorräumen X und Y ist. Als Beispiel sei $f : \mathbb{R}^{3\times1} \to \mathbb{R}^{3\times1}$ eine Funktion, die ein elektrisches Feld im Raum beschreibt. Die Argumente $\vec{x}$ von f sind dabei Koordinaten von Raumpunkten und der Funktionswert $f(\vec{x})$ beschreibt das elektrische Feld am Ort der zum Vektor $\vec{x}$ gehört. Dabei gibt $||f(\vec{x})||_2$ die Stärke des Feldes an und $f(\vec{x})/||f(\vec{x})||_2$ seine räumliche Richtung (falls $f(\vec{x}) \neq \vec{0}$ ist).
Welche Bedeutung hat in diesem Fall eine affin lineare Approximation von f?
Stellen wir uns vor, das Feld f ist im Punkt $\vec{x}_0$ bekannt. Wie stark $f(\vec{x}_0 + \vec{w})$ von $f(\vec{x}_0)$ abweichen wird, wenn wir uns um den kleinen Differenzvektor $\vec{w}$ vom Ort $\vec{x}_0$ entfernen, kann man annäherungsweise durch eine lineare Funktion L beschreiben. Im vorliegenden Fall ist $L : \mathbb{R}^{3\times1} \to \mathbb{R}^{3+1}$ direkt durch eine Matrix gegeben

$$L = \begin{pmatrix} L_{11} & L_{12} & L_{13} \\ L_{21} & L_{22} & L_{23} \\ L_{31} & L_{33} & L_{33} \end{pmatrix}$$

Der Eintrag L_{ij} besagt dabei, wie stark sich die Komponente i des Feldes ändert pro zurückgelegter Entfernung in Richtung j (die Einheit ist also Feldstärke/Länge). Bewegt man sich z. B. ausschließlich in Richtung des ersten räumlichen Basisvektors, so hat der Vektor $\vec{w}$ die Form $\vec{w} = w_1\vec{e}_1$ und damit ist

$$f(\vec{x}_0) + L(\vec{w}) = f(\vec{x}_0) + \begin{pmatrix} L_{11}w_1 \\ L_{21}w_1 \\ L_{31}w_1 \end{pmatrix}$$

der ungefähre Wert des elektrischen Feldes am Punkt $\vec{x}_0 + w_1\vec{e}_1$.
Ein entsprechender Zusammenhang gilt für Bewegungen $\vec{x}_0 + w_j\vec{e}_j$ in die beiden anderen Raumrichtungen und bei einer Gesamtbewegung $\vec{x}_0 + \vec{w} = \vec{x}_0 + w_1\vec{e}_1 + w_2\vec{e}_2 + w_3\vec{e}_3$ überlagern sich die entsprechenden Änderungen des elektrischen Feldes annähernd linear

$$f(\vec{x}_0 + \vec{w}) \approx f(\vec{x}_0) + w_1 \begin{pmatrix} L_{11} \\ L_{21} \\ L_{31} \end{pmatrix} + w_2 \begin{pmatrix} L_{12} \\ L_{22} \\ L_{32} \end{pmatrix} + w_3 \begin{pmatrix} L_{13} \\ L_{23} \\ L_{33} \end{pmatrix}$$

Ist also das elektrische Feld f differenzierbar, so läßt sich die lokale Änderung durch neun Zahlen L_{ij} beschreiben, die die Änderungsraten der Komponenten i bei jeweiliger Bewegung in den Raumrichtungen j angeben. Zusammen mit den drei Zahlen in $f(\vec{x}_0)$ können wir also (zumindest approximativ) das Feld in der Nähe von $\vec{x}_0$ allein durch zwölf Zahlen beschreiben. Natürlich muß die lineare Abbildung L passend zum Feld gewählt werden und wie das funktioniert, wollen wir uns jetzt anschauen.
Prinzipiell gibt es mehrere Möglichkeiten, zu einer gegebenen Funktion eine gut approximierende lineare Abbildung zu finden. Die Wahl wird allerdings eindeutig (und damit verschwinde die *Qual* der Wahl), wenn wir fordern, daß der Approximationsfehler im Abstand $||\vec{w}||$ nicht nur proportional zu $||\vec{w}||$, sondern deutlich kleiner als die Entfernung $||\vec{w}||$ sein soll. Zur Konkretisierung sei $f : X \to Y$ eine Abbildung zwischen zwei Vektorräumen X, Y. In der Nähe eines Punktes $\vec{x}_0$ sei durch die lineare Funktion L eine Approximation gegeben. Der Approximationsfehler ist dann

$$E(\vec{w}) = f(\vec{x}_0 + \vec{w}) - [f(\vec{x}_0) + L(\vec{w})]$$

und die Größe des Fehlers ist $e(\vec{w}) = ||E(\vec{w})||$. Die Forderung, daß $e(\vec{w})$ deutlich kleiner als $||\vec{w}||$ sein soll, kann man auch so formulieren, daß

$$e(\vec{w}) \leq C(\vec{w})||\vec{w}||$$

gilt, wobei die Proportionalitätskonstante für kleiner werdendes $\vec{w}$ ebenfalls immer kleiner werden soll, also

$$\lim_{\vec{w}\to\vec{0}} C(\vec{w}) = \lim_{\vec{w}\to\vec{0}} \frac{e(\vec{w})}{||\vec{w}||} = 0.$$

Diese Bedingung an den Approximationsfehler läßt tatsächlich höchstens eine lineare Abbildung L zur Approximation zu. Um dies zu sehen, nehmen wir an, daß L_1 und L_2 zwei lineare Abbildungen seien, die zu zwei Approximationsfehlern E_1, E_2 Anlaß geben

$$f(\vec{x}_0 + \vec{w}) = f(\vec{x}_0) + L_1(\vec{w}) + E_1(\vec{w})$$
$$f(\vec{x}_0 + \vec{w}) = f(\vec{x}_0) + L_2(\vec{w}) + E_2(\vec{w})$$

für die jeweils gilt

$$\lim_{\vec{w}\to\vec{0}} \frac{||E_1(\vec{w})||}{||\vec{w}||} = 0, \qquad \lim_{\vec{w}\to\vec{0}} \frac{||E_2(\vec{w})||}{||\vec{w}||} = 0$$

Durch Bildung der Differenz erhält man

$$(L_1 - L_2)(\vec{w}) = E_2(\vec{w}) - E_1(\vec{w})$$

Wählen wir $\vec{w} = t\vec{u}$ für einen beliebigen Vektor $\vec{u}$ der Länge $||\vec{u}|| = 1$, so ergibt sich wegen $(L_1 - L_2)(\vec{u}) = \frac{1}{t}(L_1 - L_2)(t\vec{u})$ und $||t\vec{u}|| = t$

$$(L_1 - L_2)(\vec{u}) = \lim_{t\to 0} \frac{1}{t}(L_1 - L_2)(t\vec{u}) = \lim_{t\to 0} \frac{E_2(t\vec{u})}{||tu||} - \lim_{t\to 0} \frac{E_e(t\vec{u})}{||t\vec{u}||} = \vec{0}$$

Insgesamt gilt also für beliebige Vektoren $\vec{w} \neq \vec{0}$

$$(L_1 - L_2)(\vec{w}) = ||\vec{w}||(L_1 - L_2)\left(\frac{\vec{w}}{||\vec{w}||}\right) = ||\vec{w}|| \cdot \vec{0} = \vec{0}$$

und natürlich auch $(L_1 - L_2)(\vec{0}) = \vec{0}$, d. h. die beiden Abbildungen sind zwangsläufig identisch. Folglich kann es nicht zwei verschiedene lineare Abbildungen geben, die beide der Approximationsbedingung

$$\lim_{\vec{w}\to\vec{0}} \frac{||f(\vec{x}_0 + \vec{w}) - f(\vec{x}_0) - L(\vec{w})||}{||\vec{w}||} = 0$$

genügen. Wenn wir also überhaupt solch eine Abbildung L finden, dann ist sie eindeutig. Funktionen, für die eine affin lineare Approximation

im obigen Sinne gefunden werden kann, stecken wir in eine Tüte mit der Aufschrift „differenzierbare Funktion“.

Definition 17. *Seien X, Y reelle normierte Vektorräume und sei $\emptyset \neq D \subset X$ offen. Eine Funktion $f : D \to Y$ heißt differenzierbar in $\vec{x}_0 \in D$, falls eine beschränkte lineare Abbildung $L : X \to Y$ existiert, so daß*

$$\lim_{\vec{w} \to 0} \frac{||f(\vec{x}_0 + \vec{w}) - f(\vec{x}_0) - L(\vec{w})||_Y}{||\vec{w}||_X} = 0.$$

f heißt differenzierbar, wenn f für alle $\vec{x}_0 \in D$ differenzierbar ist. Die lineare Abbildung L nennt man auch das (totale) Differential $df_{\vec{x}_0}$ von f an der Stelle $\vec{x}_0$

Bevor wir uns verschiedene Anwendungen der Funktionsapproximation mit dem totalen Differential anschauen, wollen wir zunächst verstehen, wie man das Differential einer differenzierbaren Funktion berechnet. Da das Differential $df_{\vec{x}_0}$ selbst eine lineare Funktion ist, können wir dabei unsere Kenntnisse über lineare Funktionen einsetzen. Ein großer Vorteil von linearen Funktionen ist z. B., daß man *alle* Funktionswerte leicht angeben kann, wenn man nur die Bilder der Vektoren einer Basis kennt. Nehmen wir nun an, daß der Vektorraum X endlich dimensional ist und daß $(\vec{a}_1, \ldots, \vec{a}_n)$ eine Basis von X ist. Kennen wir die endlich vielen Vektoren $df_{\vec{x}_0}(\vec{a}_1), \ldots, df_{\vec{x}_0}(\vec{a}_n)$, so können wir auch das Differential für den allgemeinen Vektor $\vec{w} = w_1\vec{a}_1 + \ldots + w_n\vec{a}_n$ berechnen. Wegen der Linearität gilt nämlich

$$df_{\vec{x}_0}(\vec{w}) = w_1 df_{\vec{x}_0}(\vec{a}_1) + \ldots + w_n df_{\vec{x}_0}(\vec{a}_n).$$

Die Berechnung der Abbildung $df_{\vec{x}_0}$ reduziert sich damit auf die Aufgabe, n Vektoren $df_{\vec{x}_0}(\vec{a}_1), \ldots, df_{\vec{x}_0}(\vec{a}_n)$ zu finden, also n mal ein einzelnes Bild zu bestimmen. Erinnern wir uns dazu an die Interpretation des Differentials. Für „kurze“ Vektoren $\vec{w}$ sollte ja $df_{\vec{x}_0}(\vec{w})$ die Änderung des Funktionswerte $f(\vec{x}_0+\vec{w})-f(\vec{x}_0)$ so gut beschreiben, daß der Fehler klein ist relativ zur Länge $||\vec{w}||$.
Wie kann man dann aber den Wert des Differentials für Vektoren „handelsüblicher“ Länge bestimmen, wenn wir nur etwas über $df_{\vec{x}_0}(\vec{w})$ für kurze Vektoren $\vec{w}$ wissen? Auch hier hilft wieder die Linearität, denn wegen der Eigenschaft $t df_{\vec{x}_0}(\vec{a}) = df_{\vec{x}_0}(t\vec{a})$ für $t \in \mathbb{R}$ sehen wir, daß lineare Abbildungen sich für kleine Vektoren ($t\vec{a}$ mit $|t|$ sehr klein) genauso verhalten wir für große Vektoren $\vec{a}$. Es ändert sich einfach nur die Länge des Bildes mit dem gleichen Faktor t. Wenn wir also $df_{\vec{x}_0}(\vec{a})$ berechnen wollen für einen festen Vektor $\vec{a}$, so schreiben wir einfach für $t \neq 0$

$$df_{\vec{x}_0}(\vec{a}) = \frac{1}{t} df_{\vec{x}_0}(t\vec{a}) \approx \frac{1}{t}(f(\vec{x}_0 + t\vec{a}) - f(\vec{x}_0))$$

wobei die Approximation um so genauer wird, je kleiner t ist. Gleichheit wird nach Definition im Grenzwert $t \to 0$ erreicht, wo der relative Fehler ja verschwindet. Wir sehen also

$$df_{\vec{x}_0}(\vec{a}) = \lim_{t \to 0} \frac{f(\vec{x}_0 + t\vec{a}) - f(\vec{x}_0)}{t}$$

Für jeden Funktionswert $df_{\vec{x}_0}(\vec{a}_i)$ müssen wir also einen Grenzwert berechnen, d. h. durch Berechnen von $\dim X$ Grenzwerten kennen wir das Differential vollständig. Im eindimensionalen Fall ist dies nur ein einziger Grenzwert. Betrachten wir dazu die Situation $X = \mathbb{R}$. Als Basis von $\mathbb{R}$ wählen wir den Vektor $1 \in \mathbb{R}$. Ist nun $D \subset \mathbb{R}$ offen (z. B. ein offenes Intervall) und $f : D \to \mathbb{R}$ differenzierbar in $x_0 \in D$, dann berechnet sich das Differential $df_{\vec{x}_0}$ durch

$$df_{\vec{x}_0}(w) = w df_{\vec{x}_0}(1) \qquad w \in \mathbb{R}$$

mit

$$df_{\vec{x}_0}(1) = \lim_{t \to 0} \frac{f(x_0 + t) - f(x_0)}{t}$$

Den Ausdruck auf der rechten Seite kennen Sie sicherlich aus Ihrer Schulzeit als Grenzwert des Differenzenquotienten der Funktion f an der Stelle x_0, oder kurz als *Ableitung* $f'(x_0)$. Der Begriff des Differentials ist also eng verwandt mit der Ableitung und zwar gilt

$$df_{\vec{x}_0}(w) = w f'(x_0), \qquad w \in \mathbb{R}$$

Die Ableitung $f'(x_0)$ ist also die Steigung der linearen Funktion $w \to df_{\vec{x}_0}(w) = w f'(x_0)$ und da $df_{\vec{x}_0}(w)$ die Änderung der Funktion f in der Nähe von $f(x_0)$ beschreibt, gibt $f'(x_0)$ damit auch einen sinnvollen Wert für die Steigung von f in x_0 an. Auch das werden Sie sicher in der Schule gelernt haben, und Sie sehen nun den Zusammenhang mit unserer allgemeineren Herangehensweise.
Als konkretes Beispiel betrachten wir die Funktion $f(x) = \frac{1}{x}$ auf $D = \mathbb{R} \setminus \{0\}$. Durch Umformung erhalten wir

$$\frac{f(x_0 + t) - f(x_0)}{t} = \frac{1}{t}\left(\frac{1}{x_0 + t} - \frac{1}{x_0}\right) = \frac{x_0 - (x_0 + t)}{t(x_0 + t)x_0} = -\frac{1}{(x_0 + t)x_0}$$

und wegen der Stetigkeit der Funktion

$$t \to -\frac{1}{(x_0+t)x_0} \qquad t \neq -x_0$$

bei $t = 0$ folgt

$$f'(x_0) = \lim_{t\to 0} \frac{f(x_0+t)-f(x_0)}{t} = -\frac{1}{x_0^2}$$

Das Differential ist also

$$df_{\vec{x}_0}(w) = w \cdot f'(x_0) = -\frac{w}{x_0^2}, \qquad w \in \mathbb{R}.$$

Eine Liste der Taschenrechnerfunktionen und ihrer Ableitungen sollte man sich merken. Die Herleitungen der Ableitungsausdrücke werden wir später nachholen. Zusammengesetzte Ausdrücke aus diesen elementaren Funktionen können nach bestimmten Regeln differenziert werden, auf die wir weiter unten eingehen.

$f(x)$	D	$f'(x_0)$	$df_{\vec{x}_0}(w)$
$ax+b$	$\mathbb{R}$	a	wa
$\frac{1}{x}$	$\mathbb{R}\setminus\{0\}$	$-\frac{1}{x_0^2}$	$-\frac{w}{x_0^2}$
$\exp(x)$	$\mathbb{R}$	$\exp(x_0)$	$w\exp(x_0)$
$\ln(x)$	$(0,\infty)$	$\frac{1}{x_0}$	$\frac{w}{x_0}$
$\sin(x)$	$\mathbb{R}$	$\cos(x_0)$	$w\cos(x_0)$
$\cos(x)$	$\mathbb{R}$	$-\sin(x_0)$	$-w\sin(x_0)$

Das erste Beispiel $f(x) = ax + b$ der Liste ist eine allgemeine affin lineare Funktion. Hier kann man das Differential sofort angeben, ohne einen Grenzwert berechnen zu müssen. Dazu müssen wir uns nur an die Interpretation des Differentials erinnern, nach der $w \mapsto f(x_0)+df_{\vec{x}_0}(w)$ eine affin lineare Approximation an die Funktion f in der Nähe von x_0 ist. Natürlich ist die beste affin lineare Approximation an eine affin lineare Funktion die Funktion selbst! Wegen

$$f(x_0+w) = a(x_0+w)+b = aw+ax_0+b = f(x_0)+aw$$

lesen wir das Differential $df_{\vec{x}_0}(w) = aw$ einfach ab und damit auch die Ableitung $f'(x_0) = df_{\vec{x}_0}(1) = a$. Diese Aussage gilt offenbar im allgemeinen Fall einer affin linearen Funktion $f(\vec{x}) = A(\vec{x})+\vec{b}$ zwischen

zwei Vektorräumen X und Y. Genauso wie im eindimensionalen Fall schreibt man

$$f(\vec{x}_0 + \vec{w}) = A(\vec{x}_0 + \vec{w}) + \vec{b} = A(\vec{x}_0) + \vec{b} + A(\vec{w}) = f(\vec{x}_0) + A(\vec{w})$$

woraus sofort $df_{\vec{x}_0}(\vec{w}) = A(\vec{w})$ ablesbar ist. Das Differential einer linearen Funktion $A : X \to Y$ ist also die lineare Funktion A selbst, also $dA_{\vec{x}_0} = A$ und zwar für jeden Punkt $\vec{x}_0 \in X$. Genauso ist das Differential einer affin linearen Funktion an jedem Punkt $\vec{x}_0$ der lineare Anteil.
Nachdem wir im Fall der rellwertigen Funktionen auf $D \to \mathbb{R}$ den Begriff des Differentials mit der Ableitung zusammengebracht haben, schauen wir uns im nächsten Schritt vektorwertige Funktionen mit einer Variablen an, also z. B.

$$f(x) = \begin{pmatrix} \cos x \\ \sin x \end{pmatrix}, \qquad x \in \mathbb{R}$$

Für die Komponentenfunktionen $f^{(1)}(x) = \cos x$ und $f^{(2)}(x) = \sin x$ kennen wir dabei affin lineare Approximationen in der Nähe von $x_0 \in \mathbb{R}$

$$\begin{aligned} f^{(1)}(x_0 + w) = \cos(x_0 + w) &\approx f^{(1)}(x_0) + df^{(1)}x_0(w) \\ &= \cos(x_0) + (-\sin(x_0))w \\ f^{(2)}(x_0 + q) = \sin(x_0 + w) &\approx f^{(2)}(x_0) + df^{(2)}_{x_0}(w) \\ &= \sin(x_0) + \cos(x_0)w \end{aligned}$$

Fassen wir die beiden approximativen Gleichungen zu einer Vektorgleichung zusammen, so erhalten wir

$$f(x_0 + w) = \begin{pmatrix} \cos(x_0 + w) \\ \sin(x_0 + w) \end{pmatrix} \approx \begin{pmatrix} \cos x_0 \\ \sin x_0 \end{pmatrix} + w \begin{pmatrix} -\sin x_0 \\ \cos x_0 \end{pmatrix}$$

Der erste Vektor auf der rechten Seite ist dabei gerade $f(x_0)$, so daß der zweite das Differential angibt

$$df_{x_0}(w) = \begin{pmatrix} df^{(1)}_{x_0}(w) \\ df^{(2)}_{x_0}(w) \end{pmatrix} = w \begin{pmatrix} -\sin x_0 \\ \cos x_0 \end{pmatrix}$$

Analog zum skalaren Fall definieren wir

$$f'(x_0) = df_{x_0}(1)$$

also in unserem Beispiel

$$f'(x_0) = \begin{pmatrix} -\sin x_0 \\ \cos x_0 \end{pmatrix}$$

Das Differential df_{x_0} ist dann wieder allgemein von der Form

$$df_{x_0}(w) = wf'(x_0), \qquad w \in \mathbb{R}.$$

Bei der Bildung des Differentials wird also komponentenweise vorgegangen, d. h. ist

$$f(x) = \begin{pmatrix} f^{(1)}(x) \\ \vdots \\ f^{(n)}(x) \end{pmatrix}, \qquad x \in D \subset \mathbb{R}$$

so ist das Differential einfach der Vektor der Differentiale, d. h.

$$df_{x_0}(w) = \begin{pmatrix} df_{x_0}^{(1)}(w) \\ \vdots \\ df_{x_0}^{(n)}(w) \end{pmatrix}, \qquad w \in \mathbb{R}.$$

Diese Regel gilt sogar in dem Fall, wo die Komponentenfunktionen allgemeinere Funktionen zwischen Vektorräumen sind. Nehmen wir dazu an, $D \neq \emptyset$ sei eine offene Teilmenge eines normierten Vektorraums X und $f^{(i)} : D \to Y_i, i = 1, \ldots, n$. Seien differenzierbare Funktionen auf D mit Werten in normierten Vektorräumen Y_i. Die Tupelfunktion

$$f(\vec{x}) = (f^{(1)}(\vec{x}), \ldots, f^{(n)}(\vec{x})) \quad \vec{x} \in D$$

hat dann Werte im Produktraum $Y = Y_1 \times \ldots \times Y_n$, der mit komponentenweiser Addition und skalarer Multiplikation sowie

$$||(y^{(1)}, \ldots, y^{(n)})||_y = \max_{i=1,\ldots,n} ||y^{(i)}||Y_i, \quad y^{(i)} \in Y_i$$

ebenfalls ein normierter Vektorraum ist. Anhand der Definition des Differentials können wir uns leicht davon überzeugen, daß für $\vec{x}_0 \in D$ die lineare Funktion

$$L(\vec{w}) = \left(df_{\vec{x}_0}^{(1)}(\vec{w}), \ldots, df_{\vec{x}_0}^{(n)}(\vec{w}) \right), \qquad \vec{w} \in X$$

das Differential der Funktion f an der Stelle $\vec{x}_0$ ist. Für den Approximationsfehler berechnen wir

$$\begin{aligned} &f(\vec{x}_0 + \vec{w}) - (f(\vec{x}_0) + L(\vec{w})) \\ &= (f^{(1)}(\vec{x}_0+\vec{w}) - (f^{(1)}(\vec{x}_0) + df_{\vec{x}_0}^{(1)}(\vec{w})), \ldots, f^{(n)}(\vec{x}_0+\vec{w}) - (f^{(1)}(\vec{x}_0) + df_{\vec{x}_0}^{(n)}(\vec{w}))) \end{aligned}$$

und damit

$$\frac{||f(\vec{x}+\vec{w})-f(\vec{x}_0)-L(\vec{w})||_Y}{||\vec{w}||_X} = \max_{i=1,\ldots,n} \frac{||f^{(i)}(\vec{x}_0+\vec{w})-f^{(i)}(\vec{x}_0)-df^{(i)}_{\vec{x}_0}(\vec{w})||_{Y_i}}{||\vec{w}||_X}$$

Da die rechte Seite wegen der Differenzierbarkeit der Komponenten im Grenzwert $\vec{w} \to 0$ verschwindet, sehen wir, daß die Tupelfunktion f differenzierbar ist und daß das Differential von f durch L, also durch das Tupel aller Differentiale, gegeben ist. Wir können also die Regel „komponentenweise differenzieren“ ganz allgemein anwenden.
Kommen wir nun aber zurück zu konkreten und elementaren Beispielen. Da wir den Fall von vektorwertigen Funktionen einer Variablen abgehandelt haben, wenden wir uns nun Funktionen mit mehreren Veränderlichen zu. Als Beispiel betrachten wir die Funktion

$$f(x_1,x_2) = x_2 \sin(x_1) \qquad x_1,x_2 \in \mathbb{R}$$

Wieder stellen wir die Frage nach dem Differential von f d. h. wir wollen die Funktionswertvariation $f(\vec{x}+\vec{w})-f(\vec{x})$ in der Nähe eines Punktes $\vec{x} \in \mathbb{R}^2$ durch eine lineare Funktion $L(\vec{w})$ approximativ beschreiben. Dabei können wir bereits für ganz bestimmte Vektoren $\vec{w}$ eine Aussage über die Variation der Funktionswerte machen. Ist nämlich $\vec{w}$ von der Form $(w_1,0)$, so wird ja mit $\vec{x}+\vec{w} = (x_1+w_1,x_2)$ nur das erste Argument variiert, d. h. es geht in diesem Fall nur um die Variation $h(x_1+w_1)-h(x_1)$ der Funktion $h(t) = f(t,x_2)$ von *einer* Variablen $t \in \mathbb{R}$. Um ganz klar zu machen, daß x_2 in diesem Prozeß konstant ist, wählen wir kurzfristig den Namen c für x_2, da c „konstanter“ aussieht und wir damit weniger durch die Anwesenheit des Wertes x_2 irritiert werden. Die Hilfsfunktion für Variationen im ersten Argument ist also

$$h(t) = f(t,c) = c\sin(t_1),$$

und deren Variation $h(x_1,w_1)-h(x_1)$ wird ja, wie wir wissen, durch das Differential

$$dh_{x_1}(w_1) = w_1 h'(x_1) = w_1 c\cos(x_1) = w_1 x_2 \cos(x_1)$$

beschrieben. Eine sinnvolle Approximation $L(\vec{w})$ für $f(\vec{x}+\vec{w})-f(\vec{x})$ ist also im Fall $\vec{w} = (w_1,0)$ durch $L(w_1,0) = w_1 x_2 \cos(x_1)$ gegeben.
Genauso können wir die speziellen Vektoren $\vec{w} = (0,w_2)$ betrachten, die auf Variationen der Hilfsfunktion

$$g(t) = f(x_1, t) = \sin(x_1)t \qquad t \in \mathbb{R}$$

führen, also auf

$$dg_{x_2}(w_2) = w_2 g'(x_2) = w_2 \sin(x_1).$$

Hier ist also $L(0, w_2) = w_2 \sin(x_1)$ ein sinnvoller Wert.
Was können wir aber über „schräge“ Variationen $\vec{w} = (w_1, w_2)$ mit allgemeinen w_1, w_2 aussagen? Hier haben wir zwei Möglichkeiten. Zunächst können wir auch diesen Fall durch eine Hilfsfunktion mit einer Variablen beschreiben. z. B.

$$k(t) = f(x_1 + tw_1, x_2 + tw_2) \qquad t \in \mathbb{R}$$

Dann gilt

$$f(\vec{x} + \vec{w}) - f(\vec{x}) = k(1) - k(0) \approx dk_0(1) = k'(0)$$

Schreiben wir k explizit

$$k(t) = (x_2 + tw_2)\sin(x_1 + tw_1)$$

und beachten wir, daß x_1, x_2, w_1, w_2 als Konstante zu behandeln sind, so können wir mit genügend Vorwissen über Ableitungsregeln (Ketten- und Produktregel) $k'(0)$ ausrechnen

$$k'(0) = x_2 \sin(x_1)w_1 + w_2 \sin(x_1)$$

Ein sinnvoller Wert für $L(w_1, w_2)$ ist also

$$L(w_1, w_2) = w_1 x_2 \sin(x_1) + w_2 \sin(x_1)$$

Vergleichen wir dieses Ergebnis mit unseren vorherigen Berechnungen, so sehen wir, daß die aufwendige Ableitungsbestimmung unnötig war. Wir hätten einfach von der Linearität des Differentials L Gebrauch machen können — das ist die zweite Möglichkeit der Berechnung

$$L(w_1, w_2) = L(w_1, 0) + L(0, w_2) = w_1 x_2 \cos(x_1) + w_2 \sin(x_1)$$

Fassen wir zusammen: Wenn Sie das Differential einer Funktion f von n Variablen $x_1, \ldots, x_n$ an einer Stelle $\vec{y} = (y_1, \ldots, y_n)$ berechnen wollen, dann sind zuerst n Hilfsfunktionen einer Variablen zu differenzieren. Die erste Hilfsfunktion ist $h(x_1) = f(x_1, y_2, \ldots, y_n)$ deren Ableitung an der Stelle $x_1 = y_1$ benötigt wird. Sie müssen also f nach der ersten Variablen ableiten. Diese Ableitung nennen wir auch *partielle Ableitung* und bezeichnen sie mit

$$\partial_1 f(y_1, \ldots, y_n) = h'(y_1)$$

Danach werden alle Variablen außer der zweiten eingefroren, d. h. wir betrachten $g(x_2) = f(y_1, x_2, y_3, \ldots, y_n)$ zur Bestimmung der Variation von f bezüglich Schwankungen im zweiten Argument. Die entsprechende Ableitung wird analog mit

$$\partial_2 f(y_1, \ldots, y_n) = g'(y_2)$$

bezeichnet. Entsprechend ermitteln wir die partiellen Ableitungen $\partial_3 f(\vec{y}), \ldots, \partial_n f(\vec{y})$ an der Stelle $\vec{y}$. Für das Differential der Funktion f an der Stelle $\vec{y}$ gilt dann

$$df_{\vec{y}}\,(\vec{e}_k) = \partial_k f(\vec{y}) \qquad k = 1, \ldots, n$$

wobei $\vec{e}_k$ die kanonischen Basisvaktoren des $\mathbb{R}^n$ sind. Wegen der Linearität von $df_{\vec{y}}$ folgt schließlich für einen allgemeinen Vektor $\vec{w} = w_1\vec{e}_1 + \ldots + w_n\vec{e}_n$

$$df_{\vec{y}}\,(\vec{w}) = w_1\partial_1 f(\vec{y}) + \ldots + w_n\partial_n f(\vec{y})$$

Bei dieser Argumentation spielt es übrigens gar keine Rolle, ob f skalar oder vektorwertig ist. Im Fall einer vektorwertigen Funktion, wie etwa

$$f(x_1, x_2, x_3) = \begin{pmatrix} \cos(x_1)\exp(x_3) \\ 2(x_1 + x_2 + 3x_3) \end{pmatrix}$$

ist der einzige Unterschied, daß die partiellen Ableitungen Vektoren statt Skalare ergeben. Für die erste partielle Ableitung müssen wir $h(x_1) = \binom{\cos(x_1)\exp(y_2)}{2x_1+2(y_2+3y_3)}$ an der Stelle y_1 differenzieren, was

$$\partial_1 f(\vec{y}) = h'(y_1) = \begin{pmatrix} -\sin(y_1)\exp(y_3) \\ 2 \end{pmatrix}$$

ergibt. Entsprechend ist

$$\partial_2 f(\vec{y}) = \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \qquad \partial_3 f(\vec{y}) = \begin{pmatrix} \cos(y_1)\exp(y_3) \\ 6 \end{pmatrix}$$

so daß

$$df_{\vec{y}}(\vec{w}) = w_1 \begin{pmatrix} -\sin(y_1)\exp(y_3) \\ 2 \end{pmatrix} + w_2 \begin{pmatrix} 0 \\ 2 \end{pmatrix} + w_3 \begin{pmatrix} \cos(y_1)\exp(y_3) \\ 6 \end{pmatrix}$$

Die Berechnung des Differentials einer Funktion f von n Variablen, die Vektoren mit m Komponenten liefert, ist offensichtlich mit Schulmathematik zu bewältigen. Es müssen n partielle Ableitungen $\partial_k f$ berechnet, d. h. alle m Komponenten nach den einzelnen Variablen x_k differenziert

werden. Insgesamt fallen also $n \cdot m$ Ableitungen an zur Berechnung des Differentials. Das kann zwar aufwendig werden, wenn $n \cdot m$ groß ist, aber nicht schwieriger als das Ableiten von reellwertigen Funktionen mit einer Variablen.
Um Differentiale auch von komplizierten zusammengesetzten Funktionsausdrücken virtuos berechnen zu können, fehlt uns noch eine Zutat, aus der dann alle weiteren Differentiationsregeln folgen: die Kettenregel.
Nehmen wir an, h sei eine zusammengesetzte Funktion $h(\vec{x}) = f(g(\vec{x}))$ und wir wollen $dh_{\vec{x}_0}$ berechnen. Das Differential $dh_{\vec{x}_0}(\vec{w})$ beschreibt approximativ die Differenz

$$h(\vec{x}_o + \vec{w}) - h(\vec{x}_0) = f(g(\vec{x}_0 + \vec{w})) - f(g(\vec{x}_0))$$

für kurze Vektoren $\vec{w}$. Wenn g in $\vec{x}_0$ differenzierbar ist, wissen wir aber, daß

$$g(\vec{x}_0 + \vec{w}) \approx g(\vec{x}_0) + dg_{\vec{x}_0}(\vec{w})$$

gilt, wobei die beschränkte lineare Abbildung $dg_{\vec{x}_0}$ den kurzen Vektor $\vec{w}$ nicht beliebig stark verlängern kann. Also ist auch $dg_{\vec{x}_0}(\vec{w})$ ein kurzer Vektor, wenn nur $\vec{w}$ kurz genug ist. Damit ist $g(\vec{x}_0)+dg_{\vec{x}_0}(\vec{w})$ eine kleine Abweichung vom Punkt $g(\vec{x}_0)$ und wenn f in $g(\vec{x}_0)$ differenzierbar ist, folgt

$$f(g(\vec{x}_0 + \vec{w}) \approx f(g(\vec{x}_0) + dg_{\vec{x}_0}(\vec{w})) \approx f(g(\vec{x}_0)) + df_{g(\vec{x}_0)}(dg_{\vec{x}_0}(\vec{w}))$$

Fassen wir die Teilschritte zusammen, so folgt

$$h(\vec{x}_0 + \vec{w}) - h(\vec{x}_0) \approx df_{g(\vec{x}_0)}(dg_{\vec{x}_0}(\vec{w}))$$

und wir haben damit das Differential von h im Punkt $\vec{x}_0$ gefunden

$$dh_{\vec{x}_0}(\vec{w}) = df_{g(\vec{x}_0)}(dg_{\vec{x}_0}(\vec{w}))$$

Die Merkregel zur Berechnung lautet also: Das Differential einer Verkettung ist die Verkettung der Differentiale. Dieses wichtige Ergebnis läßt sich mit Hilfe der Definition des Differentials beweisen (die Idee ist dabei wie oben in der $\approx$ Motivation) und wir fassen es als Satz zusammen.

Satz 14. *Seien X, Y, Z normierte Vektorräume und $\emptyset \neq D \subset X, \emptyset \neq E \subset Y$ offen. Die Funktionen $g : D \to E$ und $f : E \to Z$ seien differenzierbar in $\vec{x}_0 \in D$ bzw. $g(\vec{x}_0) \in E$. Dann ist die Verkettung $f \circ g$ ebenfalls in $\vec{x}_0$ differenzierbar und es gilt $d(f \circ g)_{\vec{x}_0} = df_{g(\vec{x}_0)} \circ dg_{\vec{x}_0}$.*

Im eindimensionalen Fall ist die Kettenregel für Differentiale übrigens genau die Kettenregel der Ableitung. Sind f, g reellwertige Funktionen auf $\mathbb{R}$, so ist das Differential ja

$$dg_{x_0}(w) = wg'(x_0), \quad df_{g(x_0)}(v) = vf'(g(x_0))$$

und die Verkettung liefert das Differential

$$d(f \circ g)_{x_0}(w) = df_{g(x_0)}(dg_{x_0}(w)) = (wg'(x_0))f'(g(x_0)).$$

Da auch für die Verkettung $d(f \circ g)_{x_0}(w) = w(f \circ g)'(x_0)$ gilt, folgt durch Gleichsetzen mit $w = 1$

$$(f \circ g)'(x_0) = f'(g(x_0))g'(x_0)$$

die Kettenregel der Ableitung. Zur Benutzung der Kettenregel müssen Sie nur ihr Auge trainieren, daß es in einem gegebenen Ausdruck

$$h(x) = \frac{1}{\cos x}$$

Verkettungen erkennt. Die Funktionen f, g werden nämlich nicht mitgeliefert. Sie zu definieren, bleibt Ihrer Kreativität überlassen. Natürlich hilft uns die Kettenregel bei der Berechnung des Differentials oder der Ableitung nur dann weiter, wenn wir die Funktionen f, g so definieren, daß $f \circ g = h$ gilt *und* wir die Differentiale von f und g kennen (bzw. selbst wiederum mit Hilfe der Kettenregel ausrechnen können). Dieser Prozeß wird also umso einfacher, je mehr Funktionen Sie zusammen mit ihren Ableitungen kennen. Im Moment stehen uns nur die elementaren Taschenrechnerfunktionen aus unserer Liste zur Verfügung. Mit diesem Vorrat können wir h als Verkettung von

$$f(y) = \frac{1}{y}, \qquad g(x) = \cos x$$

schreiben, denn

$$f(g(x)) = \frac{1}{g(x)} = \frac{1}{\cos(x)} = h(x).$$

Die Ableitung von h ergibt sich nun indirekt aus

$$f'(y) = -\frac{1}{y^2}, \qquad g'(x) = -\sin(x)$$

zu

$$h'(x_0) = -\frac{1}{(g(x_0))^2}(-\sin(x_0)) = \frac{\sin x_0}{\cos^2 x_0}$$

Oft muß man die Kettenregel auch mehrfach anwenden, um zum Ziel zu gelangen. Betrachten wir

$$h(x) = \exp(\sin(3x+2))$$

Wollen wir diese Funktion auf unsere Taschenrechnerfunktionen zurückführen, so funktioniert dies z. B. mit

$$f(y) = \exp(y), \qquad g(x) = \sin(3x+2)$$

wobei g selbst wieder zerlegt wird in

$$g(y) = \sin(y), \qquad g_2(x) = 3x+2$$

Jetzt brauchen wir

$$g_1'(y) = \cos(y), \qquad g_2'(x) = 3$$

zur Berechnung von

$$g'(x) = g_1'(g_2(x))g_2'(x) = 3\cos(3x+2)$$

was zusammen mit $f'(y) = \exp(y)$ benötigt wird, um

$$h'(x) = f'(g(x))g'(x) = 3\exp(\sin(3x+2))\cos(3x+2)$$

zu berechnen. Die Verkettung $h(x) = f(ax+b)$ mit einer affin linearen Funktion $g(x) = ax+b$ tritt übrigens recht häufig auf. Hier kann man sich merken, daß wegen $g'(x) = a$ der Faktor vor der Variablen als Vorfaktor vor f' erscheint, also $h'(x) = af'(ax+b)$.
Manchmal ist die Verkettung in einem Ausdruck gar nicht so leicht zu erkennen. Nehmen wir an, f und g sind zwei reellwertige Funktionen auf einer offenen Menge $D \subset X$, die im Punkt $\vec{x}_0 \in D$ differenzierbar sind. Was können wir dann über das Produkt $h(\vec{x}) = f(\vec{x})g(\vec{x})$ aussagen? Haben Sie die Verkettung entdeckt? Hier wird die Produktfunktion

$$P(y_1, y_2) = y_1 y_2 \qquad (y_1, y_2) \in \mathbb{R}^2$$

verknüpft mit der Tupelfunktion

$$T(\vec{x}) = (f(\vec{x}), g(\vec{x}))$$

Über Tupel wissen wir ja bereits, daß Differentiale komponentenweise berechnet werden, also

$$dT_{\vec{x}_0}(\vec{w}) = (df_{\vec{x}_0}(\vec{w}), dg_{\vec{x}_0}(\vec{w}))$$

Fehlt noch das Differential von P an der Stelle $T(\vec{x}_0)$. Dazu berechnen wir die beiden partiellen Ableitungen von P

$$\partial_1 P(y_1, y_2) = y_2, \qquad \partial_1 P(Y_1, Y_2) = y_1$$

(für festes y_2 ist $y_1 \mapsto y_1 y_2$ eine lineare Funktion mit der Ableitung y_2 — Entsprechendes gilt für $y_2 \mapsto y_1 y_2$ bei festem y_1).
Das Differential $dP_{\vec{y}}(\vec{v})$ ist damit

$$dP_{\vec{y}}(\vec{v}) = v_1 \partial_1 P(\vec{y}) + v_2 \partial_2 P(\vec{y}) = v_1 y_2 + v_2 y_1$$

Schließlich ergibt die Kettenregel für das Differential des Produkts

$$\begin{aligned} d(fg)_{\vec{x}_0}(\vec{w}) &= dP_{T(\vec{x}_0)}(dT_{\vec{x}_0}(\vec{w})) = dP_{T(\vec{x}_0)}(df_{\vec{x}_0}(\vec{w}), dg_{\vec{x}_0}(\vec{w})) \\ &= df_{\vec{x}_0}(\vec{w}) g(\vec{x}_0) + dg_{\vec{x}_0}(\vec{w}) f(\vec{x}_0) \end{aligned}$$

Diese sogenannte Produktregel

$$d(fg)_{\vec{x}_0} = g(\vec{x}_0) df_{\vec{x}_0} + f(\vec{x}_0) dg_{\vec{x}_0}$$

führt auf die bekannte Produktregel der Ableitung, wenn f, g nur von einer Variablen abhängen und die Beziehung $df_{x_0}(1) = f'(x_0)$ für f, g und fg benutzt wird. Wir finden

$$\begin{aligned} (fg)'(x_0) &= d(fg)_{x_0}(1) = g(x_0) df_{x_0}(1) + f(x_0) dg_{x_0}(1) \\ &= g(x_0) f'(x_0) + f(x_0) g'(x_0) \end{aligned}$$

Im Spezialfall, daß $g(\vec{x}) = a$ eine konstante Funktion ist, ergibt die Produktregel wegen $dg_{\vec{x}_0} = 0$ (Nullabbildung)

$$d(af)_{\vec{x}_0} = a df_{\vec{x}_0}$$

Konstanten können also bei der Bildung des Differentials nach vorne gezogen werden. Für die Summe $h = f + g$ zweier differenzierbarer Funktionen gilt

$$d(f+g)_{\vec{x}_0} = df_{\vec{x}_0} + dg_{\vec{x}_0}$$

d. h. das Differential einer Summe ist die Summe der Differentiale. Der Grund für diese Summenregel ist übrigens wieder die Kettenregel. Hier wird die lineare Summenfunktion

$$S(Y_1, Y_2) = Y_1 + Y_2 \qquad (Y_1, Y_2) \in \mathbb{R}^2$$

verknüpft mit der Tupelfunktion $T(\vec{x}) = (f(\vec{x}), g(\vec{x}))$. Wegen der Linearität von S gilt

$$dS_{\vec{y}}(\vec{v}) = S(\vec{v})$$

und damit liefert die Kettenregel

$$\begin{aligned} d(f+g)_{\vec{x}_0}(\vec{w}) &= dS_{T(\vec{x}_0)}(df_{\vec{x}_0}(\vec{w}), dg_{\vec{x}_0}(\vec{w})) \\ &= S(df_{\vec{x}_0}(\vec{w}), dg_{\vec{x}_0}(\vec{w})) \\ &= df_{\vec{x}_0}(\vec{w}) + dg_{\vec{x}_0}(\vec{w}) \end{aligned}$$

Zusammengefaßt gilt somit die Regel

$$d(af+bg)_{\vec{x}_0} = adf_{\vec{x}_0} + bdg_{\vec{x}_0}$$

wenn $a, b \in \mathbb{R}$ Konstanten sind. Die entsprechende Ableitungsregel im eindimensionalen reellwertigen Fall ist

$$(af+bg)'(x_0) = af'(x_0) + bg'(x_0)$$

Am Beispiel $h(x) = 2\sin(x) + \cos(x)$ identifiziert man $a = 2, b = 1, f(x) = \sin(x)$ und $g(x) = \cos(x)$, so daß

$$h'(x) = 2\cos(x) - \sin(x)$$

Als Beispiel für die Produktregel betrachten wir

$$\tan(x) = \frac{\sin(x)}{\cos(x)}$$

Hier erkennen wir das Produkt der Funktionen $f(x) = \sin(x)$ und $g(x) = 1/\cos(x)$. Von g haben wir die Ableitung mit Hilfe der Kettenregel bereits ausgerechnet: $g'(x) = \sin(x)/\cos^2(x)$. Die Produktregel liefert damit

$$\begin{aligned} \tan'(x) &= \frac{1}{\cos(x)}\cos(x) + \sin(x)\frac{\sin(x)}{\cos^2(x)} \\ &= 1 + \frac{\sin^2(x)}{\cos^2(x)} = 1 + \tan^2(x) \end{aligned}$$

Eine nützliche Regel, die sich aus mehrfacher Anwendung der Produktregel ergibt, bezieht sich auf Potenzfunktionen $f_n(x) = x^n$ mit $n \in \mathbb{N}_0$. Für $n = 0$ und $n = 1$ wissen wir schon $f_0'(x) = 0, f_1'(x) = 1$. Im Falle $f_2(x) = x^2 = f_1(x)f_1(x)$ ergibt die Produktregel

$$f_2'(x) = f_1(x)f_1'(x) + f_1(x)f_1'(x) = x + x = 2x$$

Die Funktion f_3 schreiben wir ebenfalls als Produkt zweier Faktoren $f_3 = f_1 f_2$ und erhalten

$$f_3'(x) = f_2(x)f_1'(x) + f_1(x)f_2'(x) = x^2 + x2x = 3x^2$$

Damit zeichnet sich die Regel

$$f_n'(x) = nf_{n-1}(x) \qquad n \in \mathbb{N}$$

ab, die sich per Induktion beweisen läßt. Da wir den Induktionsanfang schon gemacht haben (für $n = 1$), fehlt nur noch der Induktionsschritt. Nehmen wir an, $f_m' = mf_{m-1}$ für $m \leq n$. Dann ist zu zeigen, daß die Beziehung auch für $n + 1$ gilt. Nun ist $f_{n+1} = f_1 f_n$, da $x^{n+1} = xx^n$ ist. Die Produktregel liefert tatsächlich

$$f'_{n+1}(x) = f_n(x)f'_1(x) + f_1(x)f'_n(x) = x^n + xnx^{n-1} = (n+1)x^n$$

Mit den Hilfsmitteln, die wir nun zur Verfügung haben, sollten Sie in der Lage sein, Ableitungen (und damit auch Differentiale) von Kombinationen elementarer differenzierbarer Funktionen berechnen zu können. Zum Training schauen wir uns noch eine vektorwertige Funktion mit mehreren Variablen an $f : \mathbb{R}^{2\times 1} \to \mathbb{R}^{3\times 1}$

$$f\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \sin(x_1x_2) \\ x_1^2 + x_2^2 \\ x_1 - \exp(x_2) \end{pmatrix}$$

Um das Differential an einer Stelle $\left(\begin{smallmatrix} y_1 \\ y_2 \end{smallmatrix}\right)$ zu ermitteln, benötigen wir letztlich die Ableitungen aller Koeffizientenfunktionen $f_1(\vec{x}) = \sin(x_1x_2)$, $f_2(\vec{x}) = x_1^2 + x_2^2$ und $f_3(\vec{x}) = x_1 - \exp(x_2)$ nach x_1 bei festgehaltenem $x_2 = y_2$ und nach x_2 bei festgehaltenem $x_1 = y_1$. Bilden wir zunächst die partielle Ableitung nach der ersten Variablen x_1.

$$\partial_1 f(\vec{y}) = \begin{pmatrix} y_2\cos(y_1y_2) \\ 2y_1 \\ 1 \end{pmatrix}$$

(Nicht vergessen: Die Funktion wird komponentenweise nach x_1 abgeleitet und die zweite Variable wird dabei als Konstante y_2 behandelt. Wenn Ihnen y_2 nicht konstant genug aussieht, nennen Sie den Wert während dieser Rechnung ruhig c. Am Ende können Sie c wieder durch y_2 ersetzen.) Entsprechend ist die partielle Ableitung nach der zweiten Variable

$$\partial_1 f(\vec{y}) = \begin{pmatrix} y_2\cos(y_1y_2) \\ 2y_1 \\ 1 \end{pmatrix} + w_2 \begin{pmatrix} y_1\cos(y_1y_2) \\ 2y_2 \\ -\exp(y_2) \end{pmatrix}$$

Das Differential ergibt sich schließlich durch Linearkombination der partiellen Ableitungen

$$\begin{aligned} df_{\vec{y}}(\vec{w}) &= w_1\partial_1 f(\vec{y}) + w_2\partial_2 f(\vec{y}) \\ &= w_1 \begin{pmatrix} y_2\cos(y_1y_2) \\ 2y_1 \\ 1 \end{pmatrix} + w_2 \begin{pmatrix} y_1\cos(y_1y_2) \\ 2y_2 \\ -\exp(y_2) \end{pmatrix} \end{aligned}$$

Da $\vec{w}$ stets zum gleichen Vektorraum gehört wie die Argumente von f (sonst würde $f(\vec{y}+\vec{w})$ ja keinen Sinn machen) ist $\vec{w}$ im vorliegenden Fall ein Spaltenvektor $\vec{w} = \binom{w_1}{w_2} \in \mathbb{R}^{2\times 1}$. Im Ausdruck für das Differential $df_{\vec{y}}(\vec{w})$ erkennen wir damit aber ein Matrix-Vektor-Produkt

$$df_{\vec{y}}(\vec{w}) = \begin{pmatrix} y_2\cos(y_1y_2) & y_1\cos(y_1y_2) \\ 2y_1 & 2y_2 \\ 1 & -\exp(y_2) \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$

wobei in den Spalten gerade die partiellen Ableitungen von f an der Stelle $\vec{y}$ stehen (erste Spalte $\partial_1 f(\vec{y})$, zweite Spalte $\partial_2 f(\vec{y})$). Diese Matrix enthält damit in kompakter Schreibweise alle partiellen Ableitungen aller Komponentenfunktionen. Sie wird als *Jacobimatrix* von f in $\vec{y}$ bezeichnet

$$J_f(\vec{y}) = \begin{pmatrix} y_2\cos(y_1y_2) & y_1\cos(y_1y_2) \\ 2y_1 & 2y_2 \\ 1 & -\exp(y_2) \end{pmatrix}$$

gemäß der Beziehung

$$df_{\vec{y}}(\vec{w}) = J_f(\vec{y})\vec{w}$$

ist die Kenntnis von $J_f(\vec{y})$ gleichbedeutend mit der Kenntnis des Differentials von f. Im allgemeinen Fall $f : \mathbb{R}^{n\times 1} \to \mathbb{R}^{m\times 1}$ ist die Jacobimatrix $J_f(\vec{y})$ eine $m \times n$ Matrix. In der ersten Spalte steht die partielle Ableitung $\partial_1 f(\vec{y})$, in der zweiten $\partial_2 f(\vec{y})$ und in der letzten $\partial_n f(\vec{y})$

$$J_f(\vec{y}) = \begin{pmatrix} \partial_1 f_1(\vec{y}) & \dots & \partial_n f_1(\vec{y}) \\ \vdots & & \vdots \\ \partial_1 f_m(\vec{y} & & \partial_n f_m(\vec{y}) \end{pmatrix}$$

Auch in diesem Fall kann man wegen

$$df_{\vec{y}}(\vec{w}) = w_1\partial_1 f(\vec{y}) + w_2\partial_2 f(\vec{y}) + \ldots + w_n\partial_n f(\vec{y})$$

das Differential mit Hilfe der Jacobimatrix schreiben

$$df_{\vec{y}}(\vec{w}) = J_f(\vec{y})\vec{w}, \qquad \vec{w} = \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix}$$

Der Spezialfall von Funktionen, die von einer Teilmenge des $\mathbb{R}^{n\times 1}$ in den Raum $\mathbb{R}^{m\times 1}$ abbilden, ist deshalb sehr wichtig, da $\mathbb{R}^{n\times 1}, \mathbb{R}^{m\times 1}$ als *Koordinatenräume* beliebiger n bzw. m dimensionaler Vektorräume auftreten. Ist eine differenzierbare Funktion $g : D \to Y$ gegeben, wobei D eine offene Teilmenge eines n dimensional Raumes X ist und $\dim Y = m$, so kann man nach Auswahl von Basen $A = (\vec{a}_1, \ldots, \vec{a}_n)$ von X und $B = (\vec{b}_1, \ldots, \vec{b}_m)$ von Y, die Funktion schreiben als

$$g(x_1\vec{a}_1 + \ldots + x_n\vec{a}_n) = g_1(x_1\vec{a}_1 + \ldots + x_n\vec{a}_n)\vec{b}_1 + \ldots +, \\ g_m(x_1\vec{a}_1 + \ldots + x_n\vec{a}_n)\vec{b}_m$$

Als zulässige Werte für die Variablen $x_1, \ldots x_n$ sind dabei alle Kombinationen erlaubt, die auf Vektoren $x_1\vec{a}_1 + \ldots + x_n\vec{a}_n \in D$ führen. Sie sehen, daß hier zwei Koordinatenabbildungen eine Rolle spielen und zwar $\Phi : \mathbb{R}^{n\times 1} \to X, \Psi : \mathbb{R}^{m\times 1} \to Y$

$$\Phi \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = x_1\vec{a}_1 + \ldots + x_n\vec{a}_n, \qquad \Psi \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = y_1\vec{b}_1 + \ldots + y_m\vec{b}_m$$

Die Komponentenfunktionen g_i sind dann gerade $\Psi_i^{-1}(g)$ und das Argument der Funktionen ist jeweils $\Phi(\vec{x})$. Führen wir zur Abkürzung

$$f = \Psi^{-1} \circ g \circ \Phi$$

als Funktion von $\Phi^{-1}(D) \subset \mathbb{R}^{n\times 1}$ nach $\mathbb{R}^{m\times 1}$ ein, so ist

$$g(x_1\vec{a}_1 + \ldots + x_n\vec{a}_n) = f_1(\vec{x})\vec{b}_1 + \ldots + f_m(\vec{x})\vec{b}_m$$

Die Funktion f ist sozusagen die Übersetzung der Funktion g in Koordinaten bezüglich der Basen A, B von X, Y. Insbesondere erhalten wir eine andere Funktion f, wenn wir andere Basen als Ausgangspunkt wählen. Diese Wahlfreiheit nutzt man normalerweise aus, um eine möglichst einfache Koordinatendarstellung zu erhalten.
Sind die Basen gewählt und damit f definiert, so können wir in gewohnter Weise die Jacobimatrix $J_f(\vec{x}_0)$ an einer Stelle $\vec{x}_0 \in \Phi^{-1}(D)$ berechnen und damit kennen wir auch das Differential $df_{\vec{x}_0}$.
Da wegen $f = \Psi^{-1} \circ g \circ \Phi$ umgekehrt die Beziehung

$$g = \Psi \circ f \circ \Phi^{-1}$$

gilt, ergibt die Kettenregel

$$dg_{\Phi(\vec{x}_0)} = \Psi \circ df_{\vec{x}_0} \circ \Phi^{-1}$$

Hierbei haben wir bereits ausgenutzt, daß die Differentiale der Koordinatenabbildungen $d\Phi^{-1}, d\Psi$ aufgrund der Linearität wieder die Abbildungen selbst sind. Damit ist das Differential der Funktion $g : D \to Y$ durch die Jacobimatrix einer Koordinatendarstellung f beschreibbar. Die Jacobimatrix spielt hier übrigens die Rolle der Matrixdarstellung der linearen Abbildung $L = dg_{\Phi(\vec{x}_0)}$ bezüglich der Basen A, B. Wir erinnern uns: In den Spalten der zu L und A, B gehörenden Matrix stehen in den Spalten die B-Koordinaten der Bilder der A-Basisvektoren. Wegen der Beziehung $L = \Psi \circ df_{\vec{x}_0} \circ \Phi^{-1}$ und $\Phi(\vec{e}_k) = \vec{a}_n$ für die kanonischen Basisvektoren $\vec{e}_n \in \mathbb{R}^{n\times 1}$ folgt

$$L(\vec{a}_k) = \Psi(df_{\vec{x}_0}(\vec{e}_k)) = \Psi(\partial_k f(\vec{x}_0))$$

Die Koordinaten des Bildes $L(\vec{a}_k)$ sind in unserer Schreibweise durch $\Psi^{-1}(L(\vec{a}_k))$ gegeben, so daß $\partial_k f(\vec{x}_0)$ in der k-ten Spalte der zu L gehörigen Matrix steht. Da dies aber auch der k-te Eintrag in der Jacobimatrix von f ist, sehen wir, daß die Matixdarstellung von $L = dg_{\Phi(\vec{x}_0)}$ die Jacobimatrix $J_f(\vec{x}_0)$ der Koordinatendarstellung f der Funktion gegeben ist (kurz: die Koordinatendarstellung des Differentials ist das Differential der Koordinatendarstellung).
Den Unterschied zwischen einer Funktion und ihrer Koordinatendarstellung sollte man nicht aus den Augen verlieren, da sonst Verwirrung entstehen kann. Dies gilt besonders für Vektorräume, bei denen eine Basiswahl besonders natürlich erscheint und damit die Funktion mit ihrer Koordinatendarstellung bezüglich diesen natürlichen Basen gleichgesetzt wird. Das beste Beispiel hierfür sind die Vektorräume $\mathbb{R}^n, \mathbb{R}^m$. Natürliche (kanonische) Basisdarstellungen sind durch

$$\Phi \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = (x_1, \ldots, x_n), \qquad \Psi \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = (y_1, \ldots, y_m)$$

gegeben. Die Koordinatendarstellung einer Funktion

$$g(x_1, \ldots, x_n) = (g_1(x_1, \ldots, x_n), \Phi \ldots, g_m(x_1, \ldots, x_n))$$

ist bezüglich dieser Basen durch

$$f\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} g_1(x_1,\ldots,x_n) \\ \vdots \\ g_m(x_1,\ldots,x_n) \end{pmatrix}$$

gegeben, d. h. die Komponentenfunktionen sind identisch. Nur die Zeilenschreibweise wird zu Spaltenschreibweise. Die Jacobimatrix ist f

$$J_f(\vec{y}) = \begin{pmatrix} \partial_1 g_1(y_1,\ldots,y_n) & \ldots & \partial_n g_1(y_1,\ldots,y_n) \\ \vdots & & \vdots \\ \partial_1 g_m(y_1,\ldots,y_n) & \ldots & \partial_n g_m(y_1,\ldots,y_n) \end{pmatrix}$$

kann daher unmittelbar aus der Funktion g berechnet werden.
Daß diese Situation nicht den Allgemeinfall darstellt, wollen wir an folgendem Beispiel untersuchen. Wir betrachten die Menge aller harmonischen Schwingungen mit Frequenz 1, d. h. die Menge aller Funktionen $S_{\lambda,\alpha}(t) = \lambda \sin(t+\alpha)$. Der Parameter $\lambda \geq 0$ ist dabei die Amplitude der Schwingung und $\alpha \in [0, 2\pi)$ die Phase

$$X = \{S_{\lambda,\alpha} | \lambda \geq 0, \alpha \in [0, 2\pi)\}$$

Die Menge X kann man zum Beispiel als Modell für die möglichen zeitlichen Spannungs- und Sromstärkeverläufe eines Wechselstroms mit konstanter Frequenz benutzen (die Frequenz 1 ist dabei keine Einschränkung, da man die Zeitskala, d. h. die Einheit für die Zeit t, frei wählen kann). Als Funktion auf X betrachten wir die Abbildung

$$Q(f) = f^2 \qquad f \in X$$

Beachten Sie, daß das Argument von Q eine Funktion ist (eine harmonische Schwingung aus X) und daß das Bild $Q(f) = f^2$ ebenfalls eine Funktion ist. Es gilt

$$[Q(S_{\lambda,\alpha})](t) = [\lambda \sin(t+\alpha)]^2 = \lambda^2 \sin^2(t+\alpha)$$

Nutzen wir die Beziehungen

$$\cos(2\varphi) = \cos^2\varphi - \sin^2\varphi$$
$$\cos^2(\varphi) + \sin^2\varphi = 1$$

so finden wir

$$[Q(S_{\lambda,\alpha})](t) = \frac{\lambda^2}{2}(1 - \cos(2(t+\alpha)))$$

Als Zielmenge kann man also den um Konstante erweiterten Raum aller Schwingungen der Frequenz 2 betrachten

$$Y = \{t \to c + \lambda \sin(2t + \alpha) | c \in \mathbb{R}, \lambda \geq 0, \alpha \in [0, 2\pi)\}$$

Wir wollen nun die Frage untersuchen, wie sich die Funktion $Q : X \to Y$ in der Nähe von $f \in X$ verhält, also wie stark sich $Q(f+w)$ von $Q(f)$ unterscheidet, wenn $f + w$ eine kleine Abweichung von f ist. Um diese Frage mit Hilfe des Differentials von Q zu beantworten, sind zunächst noch einige Details zu klären: Zunächst ist zu überprüfen, ob die Mengen X und Y normierte Vektorräume sind. Dann müssen Basen in X und Y gewählt, die entsprechende Koordinatendarstellung von Q ermittelt und deren Jacobimatrix berechnet werden. Die Rückübersetzung der Jacobimatrix mit den Koordinatenabbildungen ergibt dann das Differential von Q.
Schauen wir uns zunächst an, ob X ein reeller Vektorraum ist.
Offensichtlich sind alle Elemente $S_{\lambda,\alpha} \in X$ durch zwei reelle Parameter λ, α beschreibbar und das riecht nach einer zweidimensionalen Menge. Die Parameter λ, α sind aber offensichtlich keine Koordinaten, da die Koordinaten eines zweidimensionalen Vektorraums durch die Menge $\mathbb{R}^{2\times 1}$ gegeben sind und nicht durch

$$\{(\lambda, \alpha) | \lambda \geq 0, \alpha \in [0, 2\pi)\}$$

wie im vorliegenden Fall. Also ist X nun ein Vektorraum oder nicht? Auf jeden Fall ist X eine Teilmenge von $\mathcal{F}(\mathbb{R}, \mathbb{R})$, dem Vektorraum aller reellwertigen Funktionen auf $\mathbb{R}$. Wenn wir zeigen können, daß Summen und Vielfache von Elementen aus X wieder in X liegen, dann wäre X ein Untervektorraum von $\mathcal{F}(\mathbb{R}, \mathbb{R})$ und damit selbst ein Vektorraum. Wie ist das z. B. mit Vielfachen? Ist $aS_{\lambda,\alpha}$ für $a \in \mathbb{R}$ wieder eine harmonische Schwingung? Es gilt

$$(aS_{\lambda,\alpha})(t) = a\lambda \sin(t + \alpha) = S_{a\lambda,\alpha}(t)$$

so daß $aS_{\lambda,\alpha} = S_{a\lambda,\alpha}$ sicherlich für $a \geq 0$ wieder in X enthalten ist (die Amplitude ist $a\lambda \geq 0$ und die Phase bleibt unverändert). Aber wie steht es mit negativen Vielfachen? Kann man $-S_{\lambda,\alpha}$ wieder in der Form $S_{\mu,\beta}$ schreiben? Wenn man sich den Funktionsgraph von $S_{\lambda,\alpha}$ anschaut, dann ist $-S_{\lambda,\alpha}$ durch den an der x-Achse gespiegelte Graph gegeben. Die Amplitude dieser gespiegelten Schwingung ist offensichtlich immer noch λ, aber die Schwingung ist um die Phase π gegenüber $S_{\lambda,\alpha}$ verschoben. Wir haben also $-S_{\lambda,\alpha} = S_{\mu,\beta}$ mit $\mu = \lambda \geq 0$ und

$$\beta = \begin{cases} \alpha + \pi & \alpha < \pi \\ \alpha - \pi & \alpha \geq \pi \end{cases}$$

Beliebige negative Vielfache $aS_{\lambda,\alpha}$ schreibt man als $aS_{\lambda,\alpha} = -|a|S_{\lambda,\alpha} = -S_{|a|\lambda,\alpha}$ und kann sie so auf den vorherigen Fall zurückführen. Insgesamt sehen wir, daß Vielfache der Elemente von X wieder in X liegen – das ist die halbe Miete zum Untervektorraum. Die Rechnung wird übrigens eleganter, wenn man die Darstellung der harmonischen Schwingungen mit der komplexen Exponentialfunktion benutzt.
Der grundlegende Zusammenhang ist

$$S_{\lambda,\alpha}(t) = Re(\lambda e^{i(t+\alpha)})$$

Für beliebige Vielfache gilt

$$(aS_{\lambda,\alpha})(t) = Re(a\lambda e^{i(t+\alpha)})$$

Nun läßt sich jede komplexe Zahl und damit auch $a\lambda \exp(i\alpha)$ in Polardarstellung $\mu \exp(i\beta)$ schreiben mit $\mu \geq 0$ und $\beta \in (0, 2\pi)$, so daß

$$a\lambda e^{i(t+\alpha)} = e^{it} a\lambda e^{i\alpha} = e^{it}\mu e^{i\beta} = \mu e^{i(t+\beta)}$$

und folglich

$$(aS_{\lambda,\alpha})(t) = Re(\mu e^{i(t+\beta)}) = S_{\mu,\beta}$$

Mit diesem Trick läßt sich auch zeigen, daß Summen von harmonischen Schwingungen aus X wieder in X liegen. Es ist

$$\begin{aligned}(S_{\lambda_1,\alpha_1} + S_{\lambda_2,\alpha_2})(t) &= Re(\lambda_1 e^{i(t+\alpha_1)} + \lambda_2 e^{i(t+\alpha_2)}) \\ &= Re(e^{it}(\lambda_1 e^{i\alpha_1} + \lambda_2 e^{i\alpha_2})\end{aligned}$$

und die komplexe Zahl $\lambda_1 \exp(i\alpha_1) + \lambda_2 \exp(i\alpha_2)$ besitzt wieder eine Polardarstellung $\lambda_{12} \exp(i\alpha_{12})$, mit $\lambda_{12} \geq 0, \alpha_{12} \in [0, 2\pi)$

$$(S_{\lambda_1,\alpha_1} + S_{\lambda_2,\alpha_2})(t) = Re(\lambda_{12} e^{i(t+\alpha_{12})}) = S_{\lambda_{12},\alpha_{12}(t)}$$

also $S_{\lambda_1,\alpha_1} + S_{\lambda_2,\alpha_2} \in X$. Mit einer sehr ähnlichen Argumentation kann man zeigen, daß auch Y ein Untervektorraum von $\mathcal{F}(\mathbb{R}, \mathbb{R})$ ist.
Im nächsten Schritt suchen wir eine Basis von X, d. h. ein Erzeugendensystem aus linear unabhängigen Vektoren. Die Frage lautet also, durch welche endliche Auswahl von Schwingungen aus X lassen sich *alle* Schwingungen in X per Linearkombination darstellen. Hier hilft wieder die Darstellung mit komplexen Zahlen.

$$\begin{aligned}S_{\lambda,\alpha}(t) &= Re(e^{it}\lambda e^{i\alpha}) \\ &= Re((\cos(t) + i\sin(t))\lambda(\cos(\alpha) + i\sin(\alpha)) \\ &= \lambda\cos(\alpha)\cos(t) - \lambda\sin(\alpha)\sin(t)\end{aligned}$$

Sie zeigt uns, daß eine beliebige Schwingung $S_{\lambda,\alpha}$ immer durch Linearkombination der beiden Schwingungen

$$\cos(t) = S_{1,0}(t), \qquad \sin(t) = S_{1,\frac{3}{2}\pi}(t)$$

dargestellt werden können. Damit ist $\{\cos, \sin\} \subset X$ ein Erzeugendensystem von X. Es fehlt noch, die lineare Unabhängigkeit der beiden Schwingungen nachzuweisen. Nehmen wir dazu an, $A\cos + B\sin = N$ sei das Nullelement in X, also die Nullfunktion $N(t) = 0$ für alle $t \in \mathbb{R}$. Dann gilt insbesondere

$$0 = N(0) = A\cos(0) + B\sin(0) = A$$
$$0 = N\left(\frac{\pi}{2}\right) = A\cos\left(\frac{\pi}{2}\right) + B\sin(\frac{\pi}{2}) = B$$

d. h. es kann nur die triviale Darstellung der Nullfunktion mit sin und cos geben. Die Elemente des Erzeugendensystems sind daher linear unabhängig und das Erzeugendensystem ist eine Basis. Insbesondere gilt $\dim X = 2$. Entsprechend weist man nach, daß

$$E_1(t) = 1, \quad E_2(t) = \cos(2t), \quad E_3(t) = \sin(2t)$$

ein Erzeugendensystem von Y bildet. Die lineare Unabhängigkeit zeigt man dann ähnlich wie im Fall des Raumes X. Gilt nämlich $AE_1 + BE_2 + CE_3 = N$, so erhalten wir

$$0 = N(0) = A + B\cos(0) + C\sin(0) = A + B$$
$$0 = N\left(\frac{2\pi}{4}\right) = A + B\cos\left(\frac{3\pi}{2}\right) + C\sin\left(\frac{3\pi}{2}\right) = A - C$$
$$0 = N\left(\frac{\pi}{8}\right) = A + B\cos\left(\frac{\pi}{4}\right) + C\sin\left(\frac{\pi}{\varphi}\right) = A + \frac{B+C}{\sqrt{2}}$$

Aus der Differenz der ersten beiden Gleichungen ergibt sich dann $B + C = 0$, was in der dritten Gleichung auf $A = 0$ führt. Damit liest man auch $B = C = 0$ ab, was die lineare Unabhängigkeit belegt und $\dim Y = 3$ liefert.
Da sowohl X und Y endlich dimensional sind, brauchen wir uns über die Wahl der Norm nicht viele Gedanken machen, denn in endlich dimensionalen Vektorräumen sind alle Normen äquivalent. Eine Möglichkeit, die eine Norm in beiden Räumen liefert, ist

$$||f|| = \sup_{t\in[0,2\pi)} |f(t)|$$

(Beim Nachweis der Normbedingung $||f|| = 0 \Rightarrow F = N$ muß die Periodizität der Elemente von X und Y ausgenutzt werden). Nachdem wir nun Basen in X und Y eingeführt haben, können wir auch die Koordinatendarstellung der Funktion Q angeben.

$$Q(A\cos + B\sin) = A^2\cos^2 + 2AB\cos\sin + B^2\sin^2$$

Wie wir bereits bemerkt haben, gilt

$$\sin^2(t) = \frac{1}{2}(1 - \cos(2t)) = \frac{1}{2}E_1(t) - \frac{1}{2}E_2(t)$$

und $\cos^2(t) = 1 - \sin^2(t)$ so daß

$$A^2\cos^2 + B^2\sin^2 = \left(A^2 + \frac{1}{2}B^2\right)E_1 - \left(A^2 + \frac{1}{2}B^2\right)E_2$$

Benutzen wir außerdem

$$2\sin(t)\cos(t) = \sin(2t) = E_3(t)$$

so ergibt sich

$$Q(A\cos + B\sin) = \left(A^2 + \frac{1}{2}B^2\right)E_1 - \left(A^2 + \frac{1}{2}B^2\right)E_2 + ABE_3$$

Die Koordinatendarstellung ist somit

$$f\begin{pmatrix}A\\B\end{pmatrix} = \begin{pmatrix}A^2 + \frac{1}{2}B^2\\ -A^2 - \frac{1}{2}B^2\\ AB\end{pmatrix}$$

mit Jacobimatrix

$$J_f\begin{pmatrix}A\\B\end{pmatrix} = \begin{pmatrix}2A & B\\ -2A & -B\\ B & A\end{pmatrix}$$

Das Differential $dQ_{A\cos + B\sin}$ von Q an der Stelle $A\cos + B\sin$ ist hier eine lineare Abbildung, die jedem Element $w_1\cos + w_2\sin$ von X ein Element aus Y zuordnet und zwar

$$\begin{aligned} dQ_{A\cos + B\sin}(w_1\cos + w_2\sin) &= (2Aw_1 + Bw_2)E_1 \\ &- (2Aw_1 + Bw_2)E_2 + (Bw_1 + Aw_2)E_3 \end{aligned}$$

Als Beispiel betrachten wir die Variation $Q(\sin + 0.01\cos) - Q(\sin)$, also den Fall $A = 0, B = 1, w_1 = 0.01, w_2 = 0$.
Das Differential liefert

$$[dQ_{\sin}(0.01\cos)](t) = 0.01\sin(2t)$$

als Approximation für den exakten Wert

$$\begin{aligned}(\sin(t)+0.01\cos(t))^2-\sin^2(t) &= 0.01\ 2\sin(t)\cos(t)+10^{-4}\cos^2(t)\\ &= 0.01\sin(2t)+10^{-4}\cos^2(t)\end{aligned}$$

Mit dem Beispiel $Q : X \to Y$ sind wir auf der allgemeinsten Stufe der Differentialberechnung angekommen. Sie haben gesehen, wie dieser Fall durch Einführung von Koordinaten auf die Differentialberechnung einer Abbildung von $\mathbb{R}^{n\times 1}$ nach $\mathbb{R}^{m\times 1}$ zurückgeführt wird. Diese wiederum bewerkstelligen wir mit dem Berechnen von n partiellen Ableitungen der m Koeffizientenfunktionen, also letztlich durch Ableiten von reellwertigen Funktionen mit einer Veränderlichen.
Mit etwas Übung sollten Sie damit in der Lage sein, Differentiale von allgemeinen differenzierbaren Abbildungen zwischen endlich dimensionalen Vektorräumen zu berechnen. Voraussetzung für die Berechnung ist natürlich, daß die Funktion differenzierbar ist. Zum Schluß dieses Abschnitts wollen wir uns deshalb mit der Frage beschäftigen, wie man differenzierbare Funktionen erkennt. Eine wichtige Beobachtung, die uns zeigt, an welchen Punkten eine Funktion *nicht* differenzierbar ist, faßt der folgende Satz zusammen.

Satz 15. *Seien X, Y normierte Vektorräume und $D \subset X$ offen. Ist die Funktion $f : D \to Y$ differenzierbar in $\vec{x}_0$, so ist sie dort auch stetig. Insbesondere kann f in einem Punkt nicht differenzierbar sein, wenn f dort nicht stetig ist.*

Der Nachweis dieser Aussage funktioniert so: Wenn f in $\vec{x}_0$ differenzierbar ist, dann geht der Approximationsfehler

$$e(\vec{w}) = ||f(\vec{x}_0+\vec{w}) - f(\vec{x}_0) - df_{\vec{x}_0}(\vec{w})||$$

schneller gegen Null als $||\vec{w}||$. Ist $(\vec{v}_n)_{n\in\mathbb{N}}$ eine Folge, die in D gegen $\vec{x}_0$ konvergiert, so ist die Differenz $\vec{w}_n = \vec{y}_n - \vec{x}_0$ eine Nullfolge und

$$\begin{aligned}||f(\vec{y})_n - f(\vec{x}_0)|| &= ||f(\vec{x}_0+\vec{w}_n) - f(\vec{x}_0)||\\ &= ||f(\vec{x}_0+\vec{w}_n) - f(\vec{x}_0) - df_{\vec{x}_0}(\vec{w}_n) + df_{\vec{x}_0}(\vec{w}_n)||\\ &\le e(\vec{w}_n) + ||df_{\vec{x}_0}(\vec{w}_n)|| \le e(\vec{w}_n) + ||df_{\vec{x}_0}||\ ||\vec{w}_n||\end{aligned}$$

Hierbei ist $||df_{\vec{x}_0}||$ der maximale Verzerrungsfaktor der beschränkten linearen Abbildung $df_{\vec{x}_0}$. Da $\vec{w}_n$ gegen Null kovergiert, streben sowohl $e(\vec{w}_n)$ als auch $||df_{\vec{x}_0}||\ ||\vec{w}_n||$ gegen Null, woraus wir

$$\lim_{n\to\infty} f(\vec{v}_n) = f(\vec{x}_0)$$

schließen können, für jede Folge $(\vec{v}_n)_{n\in\mathbb{N}}$, die in D gegen $\vec{x}_0$ konvergiert. Die Funktion f ist also stetig in $\vec{x}_0$.
Wie schon erwähnt, wird die Aussage eher in ihrer Negation benutzt. Zu Beispiel ist die Funktion

$$H(x) = \begin{cases} 1 & x \geq 0 \\ \\ 0 & x < 0 \end{cases} \qquad x \in \mathbb{R}$$

an der Stelle $x = 0$ *nicht* differenzierbar. Wäre sie nämlich differenzierbar, so müßte sie dort auch stetig sein, was aber nicht der Fall ist. Das Gleiche gilt für

$$f(x) = \begin{cases} \frac{1}{x} & x \neq 0 \\ 0 & x = 0 \end{cases}, \qquad g(x) = \begin{cases} \sin\frac{1}{x} & x \neq 0 \\ 0 & x = 0 \end{cases}$$

an der Stelle $x = 0$. An Sprungstellen, Polstellen und Punkten mit unendlich starker Oszillation kann eine Funktion also nicht differenzierbar sein, da sie dort nicht einmal stetig ist. Diese Beispiele für die Aussage *nicht stetig* $\Rightarrow$ *nicht differenzierbar* sollten nicht zu dem falschen Umkehrschluß verleiten. Es gibt nämlich durchaus stetige Funktionen, die *nicht* differenzierbar sind. Als einfaches Beispiel betrachten wir die Betragsfunktion $f(x) = |x|$ auf der Menge der reellen Zahlen. Wir wissen bereits, daß diese Funktion in jedem Punkt ihrer Definitionsmenge stetig ist. Dagegen kann man leicht nachrechnen, daß f in $x_0 = 0$ *nicht* differenzierbar sein kann. Wäre f nämlich dort differenzierbar mit Ableitung $a \in \mathbb{R}$, so würde der Fehler

$$\left| f\left(\frac{1}{n}\right) - f(0) - a \cdot \frac{1}{n} \right| = \left| \frac{1}{n} - a\frac{1}{n} \right|$$

per Definition selbst bei Division durch $\frac{1}{n}$ für $n \to \infty$ verschwinden. Dies erzwingt offensichtlich $|1 - a| = 0$ also $a = 1$. Genauso müßte aber auch

$$n\left| f\left(-\frac{1}{n}\right) - f(0) - a\left(-\frac{1}{n}\right) \right| = n\left| \frac{1}{n} + a\frac{1}{n} \right| = \left| 1 + a \right|$$

verschwinden, was wiederum $a = -1$ erzwingt. Da keine reelle Zahl gleichzeitig plus und minus Eins sein kann, ist es also unmöglich, daß f in $x_0 = 0$ differenzierbar ist.

Bezeichnen wir mit $\mathcal{D}(D,Y)$ die Menge aller differenzierbaren Funktionen auf der offenen Menge $D \subset X$, so stellen wir damit fest, daß alle Inklusionen in der Kette

$$\mathcal{D}(D,Y) \subset C^0(D,Y) \subset \mathcal{F}(D,Y)$$

strikt sind, d. h. die linke Menge enthält jeweils echt weniger Elemente als die rechte. Daß die stetigen Funktionen $C^0(D,Y)$ einen Untervektorraum des Vektorraums $\mathcal{F}(D,Y)$ aller Funktionen von D nach Y bilden, haben wir schon überprüft. Das liegt einfach daran, daß Summen und Vielfache von stetigen Funktionen wieder stetig sind. Genauso haben wir gesehen, daß Summen und Vielfache differenzierbarer Funktionen wieder differenzierbar sind. Damit ist aber auch $\mathcal{D}(D,Y)$ ein Untervektorraum sowohl von $C^0(D,Y)$ als auch von $\mathcal{F}(D,Y)$.
Ob eine gegebene Funktion $f \in \mathcal{F}(D,Y)$ nun differenzierbar ist oder nicht, kann letztlich nur durch Nachprüfen des Kriteriums der Differenzierbarkeit ermittelt werden (siehe Definition Differenzierbarkeit). Dieser Rückgriff auf die Grenzwertbetrachtung des Approximationsfehlers $||f(\vec{x}_0+\vec{w}) - f(\vec{x}_0) - L(\vec{w})||$ ist aber meist mühsam und man versucht daher, einfachere Kriterien für die Differenzierbarkeit zu finden. Die *stetige* Differenzierbarkeit ist ein solches Kriterium, das sich auf die partiellen Ableitungen der Funktion bzw. einer ihrer Koordinatendarstellungen bezieht, die ja bei der Bestimmung des Differentials zum Einsatz kommen.

Definition 18. *Seien $n, m \in \mathbb{N}$ und $E \subset \mathbb{R}^{n\times 1}$ eine offene Menge. Eine Funktion $f : E \to \mathbb{R}^{m\times 1}$ heißt stetig differenzierbar auf E, wenn alle partiellen Ableitungen $\partial_1 f_\phi, \ldots, \partial_n f$ von f in jedem Punkt von E existieren und die partiellen Ableitungen selbst wieder stetige Funktionen auf E sind. Allgemein heißt eine Abbildung $g : D \to Y$ von einer offenen Teilmenge $D \subset X$ eines endlich dimensionalen normierten Vektorraums X in einem endlich dimensionalen normierten Vektorraum Y stetig differenzierbar, wenn g eine stetig differenzierbare Koordinatendarstellung hat. Die Menge aller stetig differenzierbaren Funktionen von D nach Y wird mit $C^1(D,Y)$ bezeichnet.*

Mit dieser neuen Menge, die ebenfalls einen Untervektorraum von $\mathcal{F}(D,Y)$ bildet, haben wir folgende Inklusionskette

$$C^1(D,Y) \subset \mathcal{D}(D,Y) \subset C^0(D,Y) \subset \mathcal{F}(D,Y)$$

Auch hier werden wir sehen, daß die Inklusion $C^1 \subset \mathcal{D}$ strikt ist, allerdings sind alle praktisch relevanten differenzierbaren Funktionen meist in C^1 enthalten. Der Vorteil von C^1 im Gegensatz zu $\mathcal{D}$ ist, daß der Zugehörigkeitstest deutlich einfacher ist. Es müssen zwei Dinge überprüft

werden: (1) Existieren partielle Ableitungen nach allen Variablen? (2) Sind diese Ableitungen selbst wieder stetige Funktionen? Bei der Beantwortung der Frage (1) geht es aber, wie wir gesehen haben, nur um die Differenzierbarkeit von Funktionen *einer* Variablen (die anderen sind beim partiellen Ableiten ja eingefroren). Aussagen wie Summen, vielfache Produkte und Verkettungen von differenzierbaren Funktionen sind wider differenzierbar, helfen dann, über die Differenzierbarkeit zu entscheiden, wobei man aber nur von vergleichsweise wenigen elementaren Funktionen *einer* Unbekannten über die Differenzierbarkeit Kenntnis haben muß. Die Stetigkeit der resultierenden partiellen Ableitung kann man dann ebenso auf die Stetigkeit von wenigen elementaren Funktionen zurückzuführen, wobei man nun ausnutzt, daß Summen, Vielfache, Produkte und Verkettungen stetiger Funktionen stetig sind.
Daß die Inklusion $C^1 \subset \mathcal{D}$ dennoch strikt ist, zeigt das folgende Beispiel.

$$f(x) = \begin{cases} x^2 \cos \frac{1}{x} & x \neq 0 \\ 0 & x = 0 \end{cases}$$

Mit Hilfe der Kettenregel und der Produktregel sieht man sofort, daß f an Punkten $x \neq 0$ differenzierbar ist mit der Ableitung

$$f'(x) = 2x \cos \frac{1}{x} + \sin \frac{1}{x}$$

Durch direkten Rückgriff auf die Definition zeigt man, daß auch an der Stelle $x = 0$ Differenzierbarkeit vorliegt mit der Ableitung $f'(0) = 0$. Tatsächlich gilt für den relativen Fehler

$$\frac{1}{n}\left| f(0+h) - f(0) - 0 \cdot h \right| = h \left| \cos \frac{1}{h} \right| \xrightarrow[h \to 0]{} 0$$

Die Ableitung

$$f'(x) = \begin{cases} 2x \cos \frac{1}{x} + \sin \frac{1}{x} & x \neq 0 \\ 0 & x = 0 \end{cases}$$

ist allerdings nicht stetig, so daß f zwar in $\mathcal{D}(\mathbb{R}, \mathbb{R})$, nicht aber in $C^1(\mathbb{R}, \mathbb{R})$ enthalten ist.
Zum Abschluß soll noch ein warnendes Beispiel diskutiert werden, das verdeutlicht, daß die Existenz der partiellen Ableitungen alleine noch *nicht* die Differenzierbarkeit der Funktion impliziert. Mit anderen Worten, die Forderung der *Stetigkeit* der partiellen Ableitungen bei der Definition der Menge C^1 ist nicht etwa unnötiges Beiwerk, sondern wesentlich für die Inklusion $C^1 \subset \mathcal{D}$. Betrachten wir dazu die Funktion

$$f(x,y) = \begin{cases} \frac{x^3-3xy^2}{x^2+y^2} & (x,y) \neq (0,0) \\ 0 & (x,y) = (0,0) \end{cases}$$

auf $\mathbb{R}^2$. Der interessante Punkt ist hier der Urspung, wo beide partiellen Ableitungen existieren. Wegen $f(x,0) = x$ sehen wir sofort, daß $\partial_1 f(0,0) = 1$ ist. Genauso liefert $f(0,y) = 0$ das Ergebnis $\partial_2 f(0,0) = 0$. Wenn f an der Stelle $(0,0)$ differenzierbar wäre mit Differential L, dann wäre zwangsläufig

$$L(w_1,w_2) = w_1\partial_1 f(0,0) + w_2\partial_2 f(0,0) = w_1$$

Nun stellt sich aber heraus, daß L *nicht* das Differential von f an der Stelle $(0,0)$ sein kann, und damit ist f nicht differenzierbar, obwohl die partiellen Ableitungen existieren.
Um zu sehen, daß L nicht das Differential von f sein kann, nehmen wir das Gegenteil an. Dann müßte für die Vektoren $t(1,1)$, mit kleinem $t \in \mathbb{R}$ gelten, daß $L(t,t) = t$ eine Approximation an die Variation $f(t,t) - f(0,0) = f(t,t)$ darstellt, die schneller als $||(t,t)||$ verschwindet. Wählen wir z. B. die Maximumnorm, so ist $||(t,t)|| = |t|$ und daher müßte

$$\frac{|f(t,t) - f(0,0) - L(t,t)|}{|t|} = \left| \frac{f(t,t)}{t} - 1 \right|$$

für $t \to 0$ verschwinden. Nun ist aber

$$\frac{f(t,t)}{t} = \frac{-2t^3}{2t^3} = -1$$

was auf die Bedingung $2 = 0$ führt, was offensichtlich Unsinn ist, d. h. die Annahme, daß L das Differential von f ist, ist falsch.
Tatsächlich sind die partiellen Ableitungen im vorliegenden Fall auch nicht stetig im Ursprung. Berechnen wir z. B.

$$\partial_1 f(x,y) = \begin{cases} \frac{3x^2-3y^2}{x^2+y^2} - \frac{(x^3-3xy^2)2x}{(x^2+y^2)^2} & (x,y) \neq 0 \\ 1 & (x,y) = 0 \end{cases}$$

so sehen wir, wenn wir uns entlang der Kurve $t \mapsto (0,t), t > 0$ an den Ursprung heranpirschen, daß

$$\lim_{t\to 0} \partial_1 f(t,t) = \lim_{t\to 0} \frac{-3t^2}{t^2} = -3 \neq \partial_1 f(0,0)$$

gilt und $\partial_1 f$ daher an der Stelle $(0,0)$ nicht stetig sein kann.