**Skriptum zur Vorlesung**

# Mathematik für Data Science 2
# Mathematik für Physikstudierende B

**Sommersemester 2021**

**Dr. Daniel Tenbrinck**
**daniel.tenbrinck@fau.de**

**Department Mathematik**
**Lehrstuhl für Angewandte Mathematik (Modellierung und Numerik)**

Version 0.32 vom 19. Juli 2021

# Inhaltsverzeichnis

# Kapitel 1

# Einleitung

Das vorliegende Skript begleitet die Vorlesung „Mathematik für Data Science 2 / Physikstudierende B“und ist im Sommersemester 2021 an der FAU Erlangen-Nürnberg entstanden. Es soll den Studierenden zusätzlich zur virtuellen Vorlesung (in Form einer Videoreihe) als Nachschlagewerk dienen und ist ausführlicher und genauer gehalten als die Vorlesungsnotizen.

Das Skript orientiert sich teilweise an dem Vorlesungsskript *"Mathematik für Physikstudierende 2"* [Knauf2020] von Prof. Dr. Andreas Knauf (FAU) aus dem Sommersemester 2020. Der Inhalt der Vorlesung wird in diesem Skript in grob zwei Teile untergliedert, nämlich mathematischen Theorien der *Linearen Algebra* und der *Analysis*. Für die Inhalte der Linearen Algebra wird häufig auf das Standardlehrbuch *"Lineare Algebra"* [Fischer2005] von Prof. Dr. Gerd Fischer (TU München) verwiesen, das eine gute Balance zwischen Theorie und Anwendung bietet. Teile der Analysis sind dem Standardlehrbuch *Änalysis 2"* [Forster2017] von Prof. Dr. Otto Forster (LMU München) entnommen. Für die Optimierung wird zu großen Teilen der Notation von Jorge Nocedal und Stephen J. Wright [Nocedal2006] gefolgt. Mein besonderer Dank gilt Lea Föcke, Tim Roith und Philipp Werner für ihre Unterstützung bei der Erstellung dieses Skripts, sowie Doris Schneider für ihre gute Organisation und die Durchführung der Tafelübung.

Folgende Inhalte werden in diesem Skript insbesondere behandelt:

## Lineare Algebra:

- Eigenwerttheorie
- Diagonalisierbarkeit / Trigonalisierbarkeit von Endomorphismen
- Jordansche Normalform
- Euklidische und unitäre Vektorräume

- Bilinear- und Sesquilinearformen
- Orthogonalisierung / Orthonormalisierung
- Spezielle Endomorphismen

## Analysis:

- Topologie normierter Räume
- Dualräume
- Fixpunktsätze
- Integrationstechniken
- Differentialrechnung in mehreren Veränderlichen
- Gewöhnliche Differentialgleichungen
- Unrestringierte Optimierung
- Optimierung unter Nebenbedingungen

Sollten Ihnen beim Studium dieses Skripts inhaltliche oder sprachliche Fehler auffallen, so würde ich mich über eine kurze Email mit einem entsprechenden Hinweis an `daniel.tenbrinck@fau.de` freuen.

Ich wünsche Ihnen viel Erfolg und viel Spaß in der Vorlesung!

Dr. Daniel Tenbrinck Erlangen, 12.04.2021

# Teil I

# Lineare Algebra

# Kapitel 2

# Eigenwerttheorie

## 2.1 Motivation

Bevor wir uns mathematisch mit dem Begriff der Eigenwerte eines Endomorphismus widmen, wollen wir motivieren warum wir uns damit beschäftigen. Eigenwerte und ihre zugehörigen Eigenvektoren sind eine wesentliche Eigenschaft linearer Abbildungen, die der Charakterisierung dieser Operatoren dienen und beschreiben, wie sie sich mathematisch verhalten. Anschaulich ausgedrückt wird ein Vektor Eigenvektor genannt, wenn seine Richtung sich nicht durch Anwendung der linearen Abbildung ändert. Er wird also höchstens um einen Faktor skaliert, d.h., verlängert oder verkürzt, und diesen Faktor nennt man Eigenwert. Die Eigenwerte einer Matrix (als Darstellung eines Endomorphismus eines endlichdimensionalen Vektorraums) sind aus vielerlei Hinsicht interessant, wie die folgenden praktischen Beispiele illustrieren:

- Eigenwerte einer Matrix legen fest, ob ein zugehöriges **lineares Gleichungssystem** eindeutig lösbar ist.
- Eigenwerte sagen etwas über die Kondition eines linearen Operators aus und sind wichtig für die Stabilität von Lösungsverfahren bei **inversen Problemen**
- Lösungen von Eigenwertproblemen beschreiben die Hauptspannungen eines Körpers in der **Mechanik**
- Eigenwerte sind messbare Größen von Operatoren in der **Quantenmechanik**.
- Eigenwerte und Eigenvektoren einer Kovarianzmatrix von gegebenen Daten beschreiben die Geometrie und Varianz der Daten im Raum und erlauben eine effektive Datenreduktion oder -kompression mittels der sogenannten **Hauptkomponentenanalyse**.
- Eigenvektoren der sogenannten **Google-Matrix** eines Netzwerks bzw. Graphen sind wesentlich für die Berechnung des Google PageRanks, der die Wichtigkeit und Reihenfolge von Suchmaschinenergebnissen festlegt.

- Eigenvektoren einer Adjazenzmatrix werden beim **Spectral Clustering** für die Segmentierung von Bildern eingesetzt.

## 2.2 Mathematische Grundlagen

Zur Entwicklung der Eigenwerttheorie betrachten wir in diesem Kapitel spezielle lineare Abbildungen $F\colon V \to V$, den sogenannten *Endomorphismen*, die von einem Vektorraum $V$ wieder nach $V$ abbilden. Hierbei beschränken wir uns auf den endlich-dimensionalen Fall und nehmen im Folgenden immer an, dass $V$ ein $\mathbb{K}$-Vektorraum über einem Körper $\mathbb{K}$ ist und eine endliche Dimension $\dim(V) = n, n \in \mathbb{N}$, besitzt. Ein kanonisches Beispiel wäre der Vektorraum $V = \mathbb{R}^n$. Es sei angemerkt, dass Eigenwerte auch für den Fall von unendlich-dimensionalen Vektorräumen untersucht werden können, zum Beispiel bei der Spektraltheorie in der Funktionanalysis. Wir wollen uns in dieser Vorlesung jedoch auf den einfacheren, endlich-dimensionalen Fall beschränken.

Wir beginnen mit einer kurzen Auffrischung der Beziehung einer darstellenden Matrix und einer allgemeinen linearen Abbildung zwischen zwei $\mathbb{K}$-Vektorräumen. Wie wir wissen, besteht zwischen Matrizen aus $\mathbb{K}^{n\times m}$ und linearen Abbildungen $F : V \to W$ zwischen $n$– bzw. $m$–dimensionalen $\mathbb{K}$–Vektorräumen $V$ und $W$ ein enger Zusammenhang. Ist $B = (b_1, ..., b_n)$ eine Basis von $V$ und $C = (c_1, ..., c_m)$ eine Basis von $W$ , dann ist die darstellende Matrix $M_C^B(F) \in \mathbb{K}^{n\times m}$ diejenige Matrix, bezüglich derer der Vektor $\sum_{i=1}^n v_i b_i \in V$ unter $F$ auf den Vektor $\sum_{j=1}^m w_j c_j \in W$ mit

$$w_j \;=\; \sum_{i=1}^{n} (M_C^B(F))_{j,i} v_i$$

abgebildet wird. Die Matrix $M_C^B(F)$ drückt also aus, wie sich die lineare Abbildung auf Vektoren von $V$ bezüglich der gewählten Basen $B$ und $C$ verhält, so dass viele mathematische Sätze für lineare Abbildungen und Matrizen äquivalent formuliert werden können.

Besonders wichtig für die Eigenwerttheorie ist der folgende grundlegende Basiswechselsatz.

**Satz 2.1** (Basiswechselsatz)
*Seien $V$ ein $n$-dimensionaler $\mathbb{K}$-Vektorraum mit $n \in \mathbb{N}$ und seien*

$$B \;:=\; (b_1, \ldots, b_n), \quad C \;:=\; (c_1, \ldots, c_n)$$

*zwei Basen von $V$. Sei nun $v \in V$ ein Vektor, der in den beiden Basen folgende Darstellungen hat*

$$v \;=\; \sum_{i=1}^{n} x_i b_i \;=\; \sum_{i=1}^{n} y_i c_i.$$

*Die Koordinaten $(x_1, \ldots, x_n) \in \mathbb{K}^n$ und $(y_1, \ldots, y_n) \in \mathbb{K}^n$ sind eindeutig bestimmt und es existiert eine reguläre Matrix $T_C^B \in \mathrm{GL}(n; \mathbb{K})$, genannt* Transformationsmatrix*, die einen Basiswechsel von $B$ nach $C$ wie folgt realisiert:*

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = T_C^B \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

*Außerdem lassen sich die Koordinaten des Vektors $v$ bezüglich der Basis $B$ aus den Koordinaten bezüglich der Basis $C$ unter zu Hilfenahme der inversen Transformationsmatrix $T_B^C = (T_C^B)^{-1}$ berechnen und es gilt*

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = T_B^C \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

*Beweis.* Siehe [Fischer2005, Bemerkung 2.6.2 und 2.6.3]. □

Bevor wir uns dem eigentlich Studium von Eigenwerten und Eigenvektoren widmen, wollen wir noch einige grundlegende Begriffe aus der Linearen Algebra wiederholen, die aus [Burger2020] bereits bekannt sein sollten. Hierbei handelt es sich um nützliche Begriffe und Eigenschaften von quadratische Matrizen als Repräsentanten von Endomorphismen, die ein Spezialfall von linearen Abbildungen zwischen $\mathbb{K}$-Vektorräumen darstellen.

**Definition 2.2** (Spur einer Matrix)
*Sei $A \in \mathbb{K}^{n \times n}$ eine quadratische Matrix. Dann ist die* Spur $\mathrm{Spur}(A)$ *von $A$ definiert als die Summe der Diagonaleinträge von $A$, d.h.,*

$$\mathrm{Spur}(A) \coloneqq \sum_{i=1}^{n} A_{i,i}.$$

**Definition 2.3** (Kern einer Matrix)
*Sei $A \in \mathbb{K}^{n \times n}$ eine quadratische Matrix. Dann ist der* Kern $\mathrm{Kern}(A)$ *von $A$, auch* Nullraum *genannt, definiert als der Unterraum von $V$, dessen Vektoren von $A$ auf den Nullvektor $\vec{0} \in V$ abgebildet werden, d.h.,*

$$\mathrm{Kern}(A) \coloneqq \{v \in V \mid Av = 0\}.$$

*Der Kern von $A$ ist also gerade der Lösungsraum des homogenen, linearen Gleichungssystems $Av = 0$.*

**Definition 2.4** (Bild und Rang einer Matrix)
*Sei $A \in \mathbb{K}^{n \times n}$ eine quadratische Matrix.*

1. *Dann ist das mit* $\mathrm{Bild}(A)$ *bezeichnete* Bild *von* $A$*, auch* Bildraum *genannt, definiert als der Unterraum von* $V$*, auf den alle Vektoren* $v \in V$ *von* $A$ *abgebildet werden, d.h.,*

$$\mathrm{Bild}(A) \;:=\; \{w \in V \mid Av = w, v \in V\}.$$

2. *Weiterhin können wir den mit* $\mathrm{Rang}(A)$ *bezeichneten* Rang *von* $A$ *definieren als Dimension des Bildraums von* $A$*, d.h.,* $\mathrm{Rang}(A) := \dim \mathrm{Bild}(A)$.

Folgende Lemmata werden uns nützlich sein in Bezug auf die Determinante $\det(A)$ einer quadratischen Matrix $A$.

**Lemma 2.5**
*Eine quadratische Matrix* $A \in \mathbb{K}^{n\times n}$ *hat genau dann vollen Rang, wenn ihre Determinante ungleich Null ist, d.h.*

$$\mathrm{Rang}(A) = \dim V \;\Leftrightarrow\; \det(A) \neq 0.$$

*Beweis.* Sei $\mathrm{Rang}(A) = \dim \mathrm{Bild}(A) = \dim V$, dann wissen wir nach dem Satz über die Orthogonalität von Bild und Kern [Burger2020, Satz 3.28], dass der Kern von $A^T$ trivial sein muss, d.h., $\mathrm{Kern}(A^T) = \{\vec{0}\}$. Dies ist äquivalent dazu, dass die Matrix $A^T$ regulär ist. Dann folgt schon mit der äquivalenten Bedingung aus [Burger2020, Satz 3.41], dass die Determinante $\det(A^T) = \det(A) \neq 0$ ist. □

**Lemma 2.6** (Determinanten-Regel von Sarrus)
*Die Determinante* $\det(A)$ *einer quadratischen* $(3 \times 3)$*-Matrix* $A \in \mathbb{K}^{3\times 3}$ *mit*

$$A \;=\; \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

*lässt sich mit der* Regel von Sarrus *wie folgt berechnen.*

$$\det(A) \;=\; a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33}.$$

*Beweis.* Siehe [Fischer2005, Theorem 3.2.5 (Leibnizformel)]. □

Das folgende Lemma erlaubt es uns die Determinante für Blockmatrizen besonders leicht auszurechnen und das Problem in einfachere Unterprobleme zu zerlegen.

**Lemma 2.7** (Determinanten-Regel für Blockmatrizen)
*Sei* $n \in \mathbb{N}$ *mit* $n \geqslant 2$ *und sei* $A \in \mathbb{K}^{n\times n}$ *eine quadratische Blockmatrix von der Gestalt*

$$A \;=\; \begin{pmatrix} A_1 & C \\ 0 & A_2 \end{pmatrix},$$

*wobei* $A_1$ *und* $A_2$ *quadratische Blöcke sind. Dann gilt für die Determinante der Blockmatrix* $A$*:*

$$\det(A) \;=\; \det(A_1) \cdot \det(A_2).$$

*Beweis.* Siehe [Fischer2005, Satz 3.1.3, D9]. □

## 2.3 Eigenwerte und Eigenvektoren

In diesem Abschnitt definieren wir den Begriff der Eigenwerte und Eigenvektoren für Endomorphismen $F$ und deren darstellende Matrizen $A := M_B(F)$ bezüglich einer Basis $B$ von $V$.

**Definition 2.8** (Eigenwert, Eigenvektor und Spektrum)
*Sei $F\colon V \to V$ ein Endomorphismus von $V$. Ein Skalar $\lambda \in K$ heißt* Eigenwert *von $F$, wenn ein sogenannter* Eigenvektor *$v \in V, v \neq 0$ existiert, so dass*

$$F(v)\, v \;=\; \lambda\, v. \tag{2.1}$$

*Die Menge aller Eigenwerte des Endomorphismus $F$ nennt man das* Spektrum *von $F$.*

Die in (2.1) gegebene *Eigenwertgleichung* lässt sich ebenso für darstellende Matrizen $A \in \mathbb{K}^{n\times n}$ schreiben. In diesem Fall interessieren wir uns für Eigenwerte und Eigenvektoren, die folgende lineare Gleichung erfüllen

$$A\, v \;=\; \lambda\, v. \tag{2.2}$$

**Definition 2.9** (Eigenraum)
*Sei $V$ ein $\mathbb{K}$-Vektorraum und sei $F\colon V \to V$ ein Endomorphismus von $V$. Dann definieren wir den Eigenraum* $\mathrm{Eig}(F;\lambda)$ *zum Eigenwert $\lambda \in \mathbb{K}$ von $F$ als lineare Hülle der Eigenvektoren zum Eigenwert $\lambda$, d.h.*

$$\mathrm{Eig}(F;\lambda) \;=\; \mathrm{Kern}(F - \lambda\,\mathrm{id}_V) \;=\; \{v \in V \mid F(v) = \lambda v, v \neq \vec{0}\} \cup \{\vec{0}\} \subset V.$$

*Analog können wir den Eigenraum* $\mathrm{Eig}(A;\lambda)$ *zum Eigenwert $\lambda \in \mathbb{K}$ einer quadratischen Matrix $A \in \mathbb{K}^{n\times n}$ definieren als*

$$\mathrm{Eig}(A;\lambda) \;=\; \mathrm{Kern}(A - \lambda I_n) \;=\; \{v \in V \mid Av = \lambda v, v \neq \vec{0}\} \cup \{\vec{0}\} \subset V,$$

*wobei $I_n \in \mathbb{K}^{n\times n}$ die Einheitsmatrix bezeichnet.*

Wir können einen Eigenwert mittels der Dimension seines zugehörigen Eigenraums noch näher charakterisieren.

**Definition 2.10** (Geometrische Vielfachheit)
*Wir bezeichnen mit der* geometrischen Vielfachheit *eines Eigenwerts $\lambda \in \mathbb{K}$ die Dimension des zugehörigen Eigenraums* $\dim \mathrm{Eig}(F;\lambda)$ *bzw.* $\dim \mathrm{Eig}(A;\lambda)$.

**Bemerkung 2.11**
*Wir bemerken, dass der Eigenraum $\mathrm{Eig}(F;\lambda) \subset V$ zum Eigenwert $\lambda \in \mathbb{K}$ von $F$ ein $F$-invarianter Untervektorraum von $V$ ist, da ja wegen der Eigenwertgleichung* (2.1) *gelten muss*

$$F(v) = \lambda v \in \mathrm{Eig}(F;\lambda), \qquad \textit{für alle } v \in \mathrm{Eig}(F;\lambda).$$

*Das bedeutet, dass wenn wir $F$ auf $\mathrm{Eig}(F;\lambda)$ einschränken, so ist $F|_{\mathrm{Eig}(F;\lambda)}$ ein Endormorphismus von $\mathrm{Eig}(F;\lambda)$ mit*

$$F|_{\mathrm{Eig}(F;\lambda)}\colon\ \mathrm{Eig}(F;\lambda) \to \mathrm{Eig}(F;\lambda).$$

## 2.4 Das charakteristische Polynom

Nachdem wir in Abschnitt 2.3 die grundlegende Begriffe eingeführt haben, wollen wir uns nun damit beschäftigen, wie sich die Eigenwerte eines Endomorphismus bzw. einer darstellenden, quadratischen Matrix berechnen lassen und was wir an ihnen ablesen können.

**Definition 2.12** (Charakteristisches Polynom)
*Sei $A \in \mathbb{K}^{n\times n}$ eine quadratische Matrix und $I_n \in \mathbb{K}^{n\times n}$ die entsprechende Einheitsmatrix, so bezeichnen wir die Abbildung*

$$\begin{aligned} P_A\colon \mathbb{K} &\to \mathbb{K} \\ t &\mapsto \det(A - tI_n) \end{aligned} \tag{2.3}$$

*als charakteristisches Polynom von $A$.*

Der folgende Satz erlaubt es uns Eigenwerte als Nullstellen des charakteristischen Polynoms zu beschreiben und gibt uns gleichzeitig eine praktische Rechenvorschrift.

**Satz 2.13** (Satz zum charakteristischen Polynom)
*Sei $A \in \mathbb{K}^{n\times n}$ eine quadratische Matrix. Dann gilt folgende Äquivalenz:*

$$\lambda \in \mathbb{K} \text{ ist Eigenwert von } A \ \Leftrightarrow\ P_A(\lambda) = \det(A - \lambda I_n) \ = \ 0. \tag{2.4}$$

*Beweis.* Für einen festen Eigenwert $\lambda \in \mathbb{K}$ von $A$, sei $v \in V, v \neq 0$ der zugehörige Eigenvektor, so dass die Eigenwertgleichung (2.2) erfüllt ist. Dann gilt:

$$\begin{aligned}
& Av - \lambda v \ = \ 0 && \\
\Leftrightarrow\ & (A - \lambda I_n)v \ = \ 0 && \text{wegen Linearität} \\
\Leftrightarrow\ & \mathrm{Kern}(A - \lambda I_n) \ \neq \ \vec{0} && \text{wegen Definition 2.3} \\
\Leftrightarrow\ & \mathrm{Bild}(A - \lambda I_n) \ \neq \ V && \text{wegen Definition 2.4} \\
\Leftrightarrow\ & \mathrm{Rang}(A - \lambda I_n) \ \neq \ \dim(V) && \text{wegen Definition 2.4} \\
\Leftrightarrow\ & \det(A - \lambda I_n) \ = \ 0 && \text{wegen Lemma 2.5}
\end{aligned}$$

□

Die Dimension des $\mathbb{K}$-Vektorraums $V$ gibt den Grad des charakteristischen Polynoms vor, d.h., für $\dim(V) = n$ können wir erwarten, dass das charakteristische Polynom $P_A$ einer

Matrix $A \in \mathbb{K}^{n\times n}$ den Grad $n$ hat. Insgesamt ist nun also das geometrische Problem der Bestimmung von Eigenwerten eines Endomorphismus zurückgeführt auf das algebraische Problem der Bestimmung von Nullstellen eines Polynoms.

Wir können die Nullstellen des charakteristischen Polynoms noch näher charakterisieren durch folgende Definition.

**Definition 2.14** (Algebraische Vielfachheiten)
*Wir definieren die Häufigkeit mit der ein Eigenwert $\lambda \in \mathbb{K}$ einer Matrix $\mathbb{K}^{n\times n}$ als Nullstelle des charakteristischen Polynoms $P_A$ auftritt als die* algebraische Vielfachheit *des Eigenwerts $\lambda$.*

**Satz 2.15** (Vielfachheiten)
*Sei $F\colon V \to V$ ein Endomorphismus von $V$ und sei $\lambda \in \mathbb{K}$ ein Eigenwert von $F$ mit algebraischer Vielfachheit $r \in \mathbb{N}$ und geometrischer Vielfachheit $s \in \mathbb{N}$. Dann ist die algebraische Vielfachheit von $\lambda$ immer größer oder gleich seiner geometrischen Vielfachheit, d.h., es gilt*

$$n \geqslant r \geqslant s \geqslant 1.$$

*Beweis.* Sei $s \in \mathbb{N}$ die geometrische Vielfachheit des Eigenwerts $\lambda$ von $F$, so ist $\mathrm{Eig}(F;\lambda) = \mathrm{lin}(\{v_1,\dots,v_s\}) \subset V$ der zugehörige Eigenraum. Da $\mathrm{Eig}(F;\lambda)$ ein Untervektorraum von $V$ ist können wir $\{v_1,\dots,v_s\}$ als eine Basis des Eigenraums auffassen. Mit Hilfe des Basisergänzungssatzes [Burger2020, Lemma 3.24] können wir weitere Vektoren $v_{s+1},\dots,v_n \in V$ bestimmen, so dass $B \coloneqq \{v_1,\dots,v_s,v_{s+1},\dots v_n\}$ eine Basis von $V$ ist. Für die Vektoren $v_j \in V$ muss wegen der Eigenwertgleichung $F(v_j) = \lambda v_j$ gelten für $1 \leqslant j \leqslant s$. Gleichzeitig können wir den Endomorphismus $F$ durch die darstellende Matrix $A \coloneqq M_B(F)$ bezüglich der gewählten Basis $B$ ausdrücken. Die Einträge der Matrix $A$ werden hierbei durch ihre Wirkung auf die Basisvektoren von $B$ eindeutig festgelegt und wir wissen wegen der Eigenwertgleichung, dass gelten muss

$$F(v_j) \;=\; Av_j \;\overset{!}{=}\; \lambda v_j, \quad 1 \leqslant j \leqslant s.$$

Da die Basiselemente $v_j, 1 \leqslant j \leqslant n$ insbesondere linear unabhängig sind, muss zur Erfüllung der rechten Seite der Gleichung für $1 \leqslant j \leqslant s$ gelten

$$A_{i,j} \;=\; \begin{cases} 0, & \text{falls } i \neq j, \\ \lambda, & \text{falls } i = j. \end{cases}$$

Damit ergibt sich, dass die darstellende Matrix $A$ die folgende Gestalt haben muss

$$A \;=\; \begin{pmatrix} \lambda I_s & * \\ 0 & C \end{pmatrix},$$

wobei für den unteren, rechten Block $C \in \mathbb{K}^{(n-s)\times(n-s)}$ gilt. Für das charakteristische Polynom können wir also auf Grund der Gestalt von $A$ und der Determinanten-Produktregel

für Blockmatrizen in Lemma 2.7 folgern, dass gilt

$$P_F(t) \;=\; P_A(t) \;=\; (\lambda - t)^s \cdot \det(C - tI_{n-s}).$$

Das zeigt, dass der Eigenwert $\lambda$ von $F$ mindestens die algebraische Vielfachheit $r \geqslant s$ besitzt.

Abschließend lässt sich festhalten, dass $s \geqslant 1$ gelten muss, da mindestens ein Eigenvektor zum Eigenwert $\lambda$ von $F$ existieren muss. Gleichzeitig ist $r \leqslant n$, da das charakteristische Polynom $P_F$ vom Grad $n$ ist. □

Um ein wenig mehr Intuition für den Begriff der Eigenwerte zu bekommen wollen wir zwei konkrete Beispiele durchrechnen.

**Beispiel 2.16** (Bestimmung von Eigenwerten)
*Wir wollen im Folgenden die Bestimmung von Eigenwerten für $(2 \times 2)$-Matrizen veranschaulichen.*

1. *Sei $A \in \mathbb{R}^{2\times 2}$ eine Matrix deren Eigenwerte wir bestimmen wollen mit*

$$A \;:=\; \begin{pmatrix} -9 & -3 \\ 16 & 5 \end{pmatrix}$$

*Zur Berechnung der Eigenwerte stellen wir das charakteristische Polynom aus Definition 2.12 auf und setzen es gleich Null.*

$$P_A(t) \;=\; \det(A - tI_2) \;=\; \det\begin{pmatrix} -9-t & -3 \\ 16 & 5-t \end{pmatrix} \;\overset{!}{=}\; 0.$$

*Für das einfache Beispiel einer $(2 \times 2)$-Matrix können wir die Determinante der obigen Matrix in geschlossener Form angeben als*

$$\det\begin{pmatrix} -9-t & -3 \\ 16 & 5-t \end{pmatrix} \;=\; (-9-t)\cdot(5-t) - (-3)\cdot 16 \;=\; t^2 + 4t + 3.$$

*Da das charakteristische Polynom bereits in Normalform vorliegt, lassen sich die Nullstellen von $P_A$, die die Eigenwerte von $A$ darstellen, mittels der p-q-Formel für $p = 4$ und $q = 3$ angeben:*

$$\lambda_{1/2} \;=\; -\frac{p}{2} \pm \sqrt{\left(\frac{p}{2}\right)^2 - q} \;=\; -2 \pm \sqrt{4-3} \;=\; -2 \pm 1.$$

*Wir erhalten also die Eigenwerte $\lambda_1 = -3$ und $\lambda_2 = -1$ der Matrix $A$.*

2. *Für den einfachen Fall von $2 \times 2$-Matrizen lässt sich das charakteristische Polynom in allgemeiner Form angeben. Sei $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathbb{K}^{2\times 2}$, so lässt sich das charakteristische Polynom $P_A$ von $A$ schreiben als:*

$$\begin{aligned} P_A(t) &= \det\begin{pmatrix} a-t & b \\ c & d-t \end{pmatrix} = (a-t)\cdot(d-t) - b\cdot c \\ &= t^2 - (a+d)\cdot t + (ad-bc) = t^2 - \mathrm{Spur}(A)\cdot t + \det(A). \end{aligned}$$

*Durch Anwendung der p-q-Formel wie in Beispiel 2.16 lassen sich die Eigenwerte $\lambda_1, \lambda_2 \in \mathbb{K}$ als Nullstellen von $P_A$ auch in allgemeiner Form angeben als:*

$$\lambda_{1/2} = \frac{\mathrm{Spur}(A)}{2} \pm \sqrt{\left(\frac{\mathrm{Spur}(A)}{2}\right)^2 - \det(A)}. \tag{2.5}$$

*Aus Gleichung (2.5) können wir die folgenden interessanten mathematischen Zusammenhänge ablesen:*

$$\begin{aligned} \lambda_1 + \lambda_2 &= \frac{\mathrm{Spur}(A)}{2} + \frac{\mathrm{Spur}(A)}{2} = \mathrm{Spur}(A), \\ \lambda_1 \cdot \lambda_2 &= \left(\frac{\mathrm{Spur}(A)}{2}\right)^2 - \left(\frac{\mathrm{Spur}(A)}{2}\right)^2 + \det(A) = \det(A). \end{aligned} \tag{2.6}$$

**Bemerkung 2.17**
*Wir wollen folgende Beobachtungen zur Berechnung von Eigenwerten festhalten.*

*i) Wie man anhand der Berechnungsformel (2.5) sieht, kann das charakteristische Polynom $P_A$ komplexe Nullstellen besitzen, sogar wenn die Matrix nur reelle Einträge hat, d.h., $A \in \mathbb{R}^{2\times 2}$. Wir erinnern daran, dass wir eine Nullstelle des charakteristischen Polynoms nur dann Eigenwert nennen, wenn sie aus dem zu Grunde liegenden Körper $\mathbb{K}$ des Vektorraums $V$ stammt. Bettet man jedoch beispielsweise $\mathbb{R}$ in seinen algebraischen Abschluss $\mathbb{C}$ ein, so dass $V$ zu einem komplex-wertigen Vektorraum wird, dann lassen sich alle Nullstellen des charakteristischen Polynoms als Eigenwerte in $\mathbb{C}$ auffassen.*

*ii) Die Beobachtungen in Gleichung (2.6) lassen sich erstaunlicherweise auch auf beliebige Matrizen $A \in \mathbb{K}^{n\times n}$ verallgemeinern und man kann zeigen, dass gilt:*

$$\sum_{i=1}^{n} \lambda_i = \mathrm{Spur}(A), \qquad \prod_{i=1}^{n} \lambda_i = \det(A).$$

*iii) Für beliebige quadratische Matrizen $A \in \mathbb{K}^{n\times n}$ gibt es für $n > 3$ im Allgemeinen keine einfache Form zur Berechnung der Determinante. Hier nutzt man typischerweise das Gaußsche Eliminationsverfahren aus [Burger2020][Kapitel 3.2] um die Matrix*

*$A - \lambda I_n$ in eine obere, rechte Dreiecksmatrix zu überführen, deren Determinante man dann an der Hauptdiagonale ablesen kann. Es sei jedoch Vorsicht geboten: Die Determinante dieser oberen, rechten Dreiecksform ist im Allgemeinen verschieden von der ursprünglichen Determinante. Das folgende Beispiel erklärt welche Schritte man zusätzlich berücksichtigen muss, um das korrekte charakteristische Polynom zu erhalten.*

**Beispiel 2.18**
*Sei $A \in \mathbb{R}^{2\times 2}$ eine quadratische Matrix deren Determinante wir bestimmen wollen mit*

$$A \;=\; \begin{pmatrix} 6 & 1 \\ 3 & 2 \end{pmatrix}.$$

*Obwohl man in diesem einfachen Fall direkt die Determinante bestimmen kann als*

$$\det(A) \;=\; 6 \cdot 2 - 1 \cdot 3 \;=\; 9$$

*verwenden wir zunächst das Gaußsche Eliminationsverfahren, um die Matrix $A$ in einer obere, rechte Dreiecksform zu bringen.*

$$\begin{pmatrix} 6 & 1 \\ 3 & 2 \end{pmatrix} \overset{2\cdot II.}{\rightarrow} \begin{pmatrix} 6 & 1 \\ 6 & 4 \end{pmatrix} \overset{II.-I.}{\rightarrow} \begin{pmatrix} 6 & 1 \\ 0 & 3 \end{pmatrix} \;=:\; \tilde{A}.$$

*Man ist schnell versucht anzunehmen, dass die Determinante dieser oberen, rechten Dreiecksmatrix $\tilde{A}$ der Determinante von $A$ entspricht, jedoch zeigt sich, dass*

$$\det(A) \;=\; 6 \cdot 3 \;=\; 18 \;\neq\; 9 \;= \det(A).$$

*Um die korrekte Determinante von $A$ aus der Form von $\tilde{A}$ zu bestimmen, müssen wir die elementaren Zeilenoperationen des Gaußschen Eliminationsverfahrens mit Hilfe der entsprechenden Elementarmatrizen schreiben. Wir können schreiben:*

$$\tilde{A} \;=\; \begin{pmatrix} 6 & 1 \\ 0 & 3 \end{pmatrix} \;=\; \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \cdot \underbrace{\begin{pmatrix} 6 & 1 \\ 3 & 2 \end{pmatrix}}_{=A}$$

*Mit der Produktregel für Determinanten aus [Fischer2005, Satz 3.1.3, D11] können wir obige Gleichung schreiben als:*

$$\underbrace{\det \begin{pmatrix} 6 & 1 \\ 0 & 3 \end{pmatrix}}_{=18} \;=\; \underbrace{\det \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}}_{=1} \cdot \underbrace{\det \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}}_{=2} \cdot \det(A).$$

*Durch Umstellen erhalten wir dann die richtige Determinante mit* $\det(A) = \frac{18}{1\cdot 2} = 9$.

**Bemerkung 2.19**
*Bei einem genaueren Studium der Elementarmatrizen im Gaußschen Eliminationsverfahren fällt auf, dass*

1. ***skalare Multiplikation einer Zeile*** *mit Faktor $a \in \mathbb{K}$ immer durch eine Elementarmatrix mit Determinante $a$ ausgedrückt wird.*

2. ***Subtraktion zweier Zeilen*** *immer durch eine Elementarmatrix mit Determinante* 1 *ausgedrückt wird.*

3. ***Zeilenvertauschungen*** *immer durch eine Elementarmatrix mit Determinante* 1 *oder* $-1$ *ausgedrückt wird, in Abhängigkeit der ensprechenden Zeilenindizes.*

Da wir nun wissen, dass wir die Eigenwerte einer Matrix als Nullstellen des charakteristischen Polynoms berechnen können, wollen wir uns mit der Bestimmung der zugehörigen Eigenvektoren beschäftigen. Wir werden im folgenden Beispiel sehen, dass sich der Eigenraum zu einem Eigenwert einer Matrix als Lösungsraum des homogenen linearen Gleichungssystems $(A - \lambda I_n)v = 0$ bestimmen lässt.

**Beispiel 2.20** (Eigenraumbestimmung)
*Sei $A \in \mathbb{R}^{3\times 3}$ eine reellwertige, quadratische Matrix mit*

$$A \;=\; \begin{pmatrix} 1 & 2 & 0 \\ -1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}.$$

*Wie man mit der Determinantenregel von Sarrus aus Lemma 2.6 nachrechnen kann, sind die Nullstellen des charakteristischen Polynoms $P_A(t)$ gegeben durch $t_1 = 1$, $t_2 = i\sqrt{2}$ und $t_3 = -i\sqrt{2}$. Somit hat die Matrix $A$ als lineare Abbildung auf den $\mathbb{R}^3$ lediglich den reellen Eigenwert $\lambda_1 = 1$.*

*Zur Bestimmung eines Eigenvektors $v \in \mathbb{R}^3$ zum Eigenwert $\lambda_1 = 1$, der die Eigenwertgleichung* (2.1) *erfüllt, müssen wir das folgende lineare Gleichungssystem lösen:*

$$(A - \lambda_1 \cdot I_3)v \;=\; \begin{pmatrix} 1-1 & 2 & 0 \\ -1 & 0-1 & 1 \\ 1 & 0 & 0-1 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \;=\; \begin{pmatrix} 0 & 2 & 0 \\ -1 & -1 & 1 \\ 1 & 0 & -1 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \;=\; 0.$$

*Um den gesuchten Eigenvektor $v = (v_1, v_2, v_3)^T \in \mathbb{R}^3$ zu bestimmen, verwenden wir das Gaußsche-Eliminationsverfahren zum Lösen linearer Gleichungssysteme [Burger2020][Kapitel*

*3.2]. Hierzu betrachten wir die folgende Sequenz von Zeilenoperationen:*

$$(A-\lambda_1 I_3 \mid y) = \left[\begin{array}{ccc|c} 0 & 2 & 0 & 0 \\ -1 & -1 & 1 & 0 \\ 1 & 0 & -1 & 0 \end{array}\right] \to \left[\begin{array}{ccc|c} -1 & -1 & 1 & 0 \\ 0 & 2 & 0 & 0 \\ 1 & 0 & -1 & 0 \end{array}\right] \to \left[\begin{array}{ccc|c} -1 & -1 & 1 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{array}\right]$$
$$\to \left[\begin{array}{ccc|c} -1 & -1 & 1 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

*Durch Rückwärtseinsetzen erhalten wir, dass $v_2 = 0$ gelten muss. Aus der ersten Zeile folgert man nun, dass $-v_1 + v_3 = 0$ und somit $v_1 = v_3$ gelten muss.*

*Der zum Eigenwert $\lambda_1 = 1$ von $A$ zugehörige Eigenvektor $v \in \mathbb{R}^3$ ist also ein Vektor der Form $v = \alpha(1,0,1)^T$ für $\alpha \neq 0$. Der durch den Kern von $A - \lambda_1 I_3$ beschriebene Unterraum von $\mathbb{R}^3$ ist somit gegeben durch die lineare Hülle bzw. den Spann* $\mathrm{Kern}(A-\lambda_1 I_3) = \mathrm{lin}(\{v\})$.

**Satz 2.21**
*Sei $A \in \mathbb{R}^{n\times n}$ eine reellwertige Matrix aufgefasst als Abbildung auf $\mathbb{C}^n$. Wenn $\lambda \in \mathbb{C}$ Eigenwert von $A$ ist, so ist auch sein komplex konjugiertes Element $\overline{\lambda} \in \mathbb{C}$ Eigenwert von $A$. Außerdem sind die zugehörigen Eigenvektoren komplex konjugiert zueinander.*

*Beweis.* Aus den Rechenregeln für komplexe Zahlen in [Burger2020, Kapitel 2.2.7] können wir für zwei komplexe Zahlen $v, w \in \mathbb{C}$ mit $v := a + b\cdot i$ und $w := c + d\cdot i$ für $a,b,c,d \in \mathbb{R}$ folgern, dass $\overline{vw} = \overline{v}\,\overline{w}$ gilt, da

$$\begin{aligned} \overline{vw} &= \overline{(a+b\cdot i)(c+d\cdot i)} = \overline{(ac-bd)+(ad+bc)\cdot i} = (ac-bd)-(ad+bc)\cdot i \\ &= (ac-bd)+(-ad-bc)\cdot i = (a-b\cdot i)(c-d\cdot i) = \overline{v}\,\overline{w}. \end{aligned}$$

Außerdem kann man leicht für endliche Summen komplexer Zahlen $v_1, \ldots, v_n \in \mathbb{C}$ zeigen, dass gilt

$$\overline{\sum_{i=1}^{n} v_i} = \sum_{i=1}^{n} \overline{v_i}.$$

Mit den beiden hergeleiteten Eigenschaften kann man nun direkt folgern, dass ein ähnlicher Zusammenhang für die Matrix-Vektor Multiplikation gilt, nämlich $\overline{Av} = \overline{A}\overline{v}$ für eine quadratische Matrix $A \in \mathbb{C}^{n\times n}$ und einen Vektor $v \in \mathbb{C}^n$.

Sei nun $\lambda \in \mathbb{C}$ ein Eigenwert der reellen Matrix $A \in \mathbb{R}^{n\times n}$ zum Eigenvektor $v \in \mathbb{C}^n$. Dann gilt offensichtlich die Eigenwertgleichung $Av = \lambda v$. Durch komplexe Konjugation beider Seiten der Gleichung wissen wir dann, dass folgende Gleichung erfüllt ist

$$\overline{Av} = \overline{\lambda v}. \tag{2.7}$$

Damit können wir den Beweis nun direkt hinschreiben.

$$A\overline{v} \overset{A \text{ reell}}{=} \overline{A}\overline{v} = \overline{Av} \overset{(2.7)}{=} \overline{\lambda v} = \overline{\lambda}\,\overline{v}.$$

Es gilt also $A\overline{v} = \overline{\lambda}\,\overline{v}$ und wir haben damit gezeigt, dass $\overline{\lambda} \in \mathbb{C}$ auch Eigenwert von $A$ zum Eigenvektor $\overline{v} \in \mathbb{C}^n$ ist. □

Der folgende Satz macht eine entscheidende Beobachung, die es uns in Kapitel 2.5 erlauben wird, den Vektorraum $V$ als direkte Summe der Eigenräume einer quadratischen Matrix $A$ zu zerlegen, falls $A$ gewisse Eigenschaften erfüllt.

**Satz 2.22**
*Sei $A \in \mathbb{K}^{n\times n}$ eine Matrix und seien $\lambda_1, \ldots, \lambda_k \in \mathbb{K}, k \leqslant n$, Eigenwerte von $A$. Außerdem seien $v_1, \ldots, v_k \in V$ Eigenvektoren von $A$ zu den Eigenwerten $\lambda_1, \ldots, \lambda_k$. Sind die Eigenwerte von $A$ paarweise verschieden, d.h., $\lambda_i \neq \lambda_j$ für alle $i \neq j$, so sind die Eigenvektoren linear unabhängig.*

*Beweis.* Wir führen den Beweis dieser Aussage durch vollständige Induktion über $k \in \mathbb{N}$.

**Induktionsanfang:** $k = 1$
Die Aussage ist trivialerweise erfüllt, da für den Eigenvektor $v_1 \in V$ gelten muss $v_1 \neq \vec{0}$.

**Induktionsschritt:** $k - 1 \to k$
Die Induktionsannahme ist, dass die Aussage bereits für den Fall $k-1$ gezeigt wurde. Seien $\alpha_1, \ldots, \alpha_k \in \mathbb{K}$ so gewählt, dass gilt

$$\sum_{i=1}^{k} \alpha_i v_i \;=\; \vec{0}. \tag{2.8}$$

Sei nun $\lambda_k \in \mathbb{C}$ ein Eigenwert von $A$, der paarweise verschieden ist zu $\lambda_1, \ldots, \lambda_{k-1} \in \mathbb{C}$. Wir multiplizieren die Gleichung (2.8) mit dem Eigenwert $\lambda_k$ und erhalten

$$\sum_{i=1}^{k} \alpha_i \lambda_k v_i \;=\; \vec{0}. \tag{2.9}$$

Außerdem erhalten wir aus Gleichung (2.8) durch Multiplikation von links mit der Matrix $A$ aus der Linearität folgenden Zusammenhang:

$$A \cdot \sum_{i=1}^{k} \alpha_i v_i \;=\; \sum_{i=1}^{k} \alpha_i A v_i \;=\; \sum_{i=1}^{k} \alpha_i \lambda_i v_i \;=\; \vec{0}. \tag{2.10}$$

Wenn wir nun die rechte Seite von (2.10) von Gleichung (2.9) subtrahieren erhalten wir

$$\sum_{i=1}^{k} \alpha_i (\lambda_k - \lambda_i) v_i \;=\; \sum_{i=1}^{k-1} \alpha_i (\lambda_k - \lambda_i) v_i \;=\; \vec{0}.$$

Da wir $\lambda_k$ als paarweise verschieden zu $\lambda_1, \ldots, \lambda_{k-1}$ angenommen haben wissen wir, dass $\alpha_i = 0$ für $i = 1, \ldots, k-1$ gelten muss, damit obige Gleichung gilt. Setzen wir dies in Gleichung (2.8) ein, so folgt direkt, dass auch $\alpha_k = 0$ gelten muss, da für den Eigenvektor $v_k \neq \vec{0}$ gilt. Dies beweist also die lineare Unabhängigkeit der Eigenvektoren von $A$ für paarweise verschiedene Eigenwerte. □

## 2.5 Diagonalisierbarkeit

Wichtige Eigenschaften eines Endomorphismus lassen sich bereits am Rang und am Spektrum einer darstellenden Matrix ablesen. Leider lassen sich diese Charakteristika im Allgemeinen (bis auf Spezialfälle) nicht direkt an den Einträgen der Matrix ablesen. Die grundlegende Frage in diesem Abschnitt wird sein, wie wir durch eine geeignete Wahl von Basen eine besonders einfache Gestalt der darstellenden Matrix erreichen können, die die Eigenschaften des zu Grunde liegenden Endomorphismus erhält.

Aus dem Basiswechselsatz 2.1 wissen wir, dass ein Basiswechsel für Abbildungsmatrizen einer Multiplikation mit zwei regulären Basiswechselmatrizen von links und rechts entspricht. Die hierdurch beschriebene Relation der Abbildungsmatrizen motiviert die folgende Definition.

**Definition 2.23** (Äquivalenz und Ähnlichkeit von Matrizen)
*Wir definieren im folgenden zwei Begriffe, die eine spezielle Relation zweier Matrizen beschreibt.*

*i) Zwei Matrizen $A, B \in \mathbb{K}^{n\times m}$ heißen* äquivalent*, wenn es Matrizen $S \in \mathrm{GL}(n;\mathbb{K})$ und $T \in \mathrm{GL}(m;\mathbb{K})$ gibt mit*

$$B \;=\; SAT^{-1}. \tag{2.11}$$

*ii) Zwei Matrizen $A, B \in \mathbb{K}^{n\times n}$ heißen* ähnlich *oder* konjugiert*, wenn es eine Matrix $S \in \mathrm{GL}(n,\mathbb{K})$ gibt mit*

$$B \;=\; SAS^{-1}. \tag{2.12}$$

Man sieht sofort, dass ähnliche Matrizen ein Spezialfall von äquivalenten Matrizen sind für $m = n$ und $T^{-1} \coloneqq S^{-1}$. Während man für äquivalente Matrizen zeigen kann, dass der Rang der Matrizen unter den in Definition 2.23 beschriebenen Transformationen erhalten bleibt, so gilt für ähnliche Matrizen sogar die noch stärkere Invarianz des Spektrums, wie das folgende Lemma zeigt.

**Lemma 2.24**
*Seien $A, B \in \mathbb{K}^{n\times n}$ zwei quadratische Matrizen. Falls $A$ und $B$ ähnlich zueinander sind, so haben sie das gleiche charakteristische Polynom.*

*Beweis.* Seien $A, B \in \mathbb{K}^{n\times n}$ zwei ähnliche Matrizen, d.h., es existiert eine Matrix $S \in \mathrm{GL}(n;\mathbb{K})$, so dass $B = SAS^{-1}$. Ferner gilt wegen Linearität und der Kommutativität der Einheitsmatrix $I_n$ für jedes Skalar $t \in \mathbb{K}$

$$S \cdot t \cdot I_n \cdot S^{-1} = t \cdot I_n.$$

Wir können also schreiben:

$$B - t \cdot I_n \;=\; SAS^{-1} - t \cdot I_n \;=\; SAS^{-1} - S \cdot t \cdot I_n \cdot S^{-1} \;=\; S(A - t \cdot I_n)S^{-1}.$$

Wenden wir nun die Determinante an, so erhalten wir aus dem Produktsatz für Determinanten [Burger2020][Satz 3.40]

$$\det(B - t \cdot I_n) \;=\; \det(S(A - t \cdot I_n)S^{-1}) \;=\; \det(S) \cdot \det(A - t \cdot I_n) \cdot \underbrace{\det(S^{-1})}_{=\det(S)^{-1}} \;=\; \det(A - t \cdot I_n).$$

Die Ausdrücke auf der linken und rechten Seite der obigen Gleichung sind gerade die Definitionen der charakteristischen Polynome von $A$ bzw. $B$, was die Aussage dieses Lemmas beweist. □

Ein besonders interessanter Fall liegt vor, wenn eine Matrix $A \in \mathbb{K}^{n \times n}$ ähnlich zu einer Diagonalmatrix $D \in \mathbb{K}^{n \times n}$ ist. Dies wird in der folgenden Definition weiter präzisiert.

**Definition 2.25** (Diagonalisierbarkeit)
*Wir definieren den Begriff der* Diagonalisierbarkeit *im folgenden sowohl für Endomorphismen als auch für Matrizen.*

1. *Ein Endormorphismus $F \colon V \to V$ eines $\mathbb{K}$-Vektorraums $V$ heißt* diagonalisierbar*, wenn $V$ eine Basis aus Eigenvektoren von $F$ besitzt.*

2. *Eine Matrix $A \in \mathbb{K}^{n \times n}$ heißt* diagonalisierbar*, wenn sie ähnlich zu einer Diagonalmatrix ist.*

Auf Grund von Lemma 2.24 wird klar, dass eine diagonalisierbare Matrix $A$ ähnlich zu einer Diagonalmatrix $D$ sein muss, die die Eigenwerte von $A$ auf der Diagonalen enthält. Im Folgenden wollen wir verstehen, wie wir entscheiden können, ob ein Endomorphismus $F$ bzw. eine darstellende Matrix von $F$ diagonalisierbar ist.

**Satz 2.26** (Diagonalisierbarkeit)
*Sei $V$ ein endlichdimensionaler $\mathbb{K}$-Vektorraum und $F \colon V \to V$ ein Endomorphismus von $V$. Dann sind die folgenden Bedingungen äquivalent:*

*i) $F$ ist diagonalisierbar*

*ii) Das charakteristische Polynom $P_F$ zerfällt in Linearfaktoren über $\mathbb{K}$ und die algebraische Vielfachheit ist gleich der geometrischen Vielfachheit für alle Eigenwerte $\lambda \in \mathbb{K}$ von $F$.*

*iii) Sind $\lambda_1, \ldots, \lambda_k \in \mathbb{K}$ die paarweise verschiedenen Eigenwerte von $F$, so lässt sich $V$ als direkte Summe der korrespondierenden Eigenräume schreiben, d.h.*

$$V \;=\; \operatorname{Eig}(F; \lambda_1) \oplus \ldots \oplus \operatorname{Eig}(F; \lambda_k).$$

*Beweis.* Wir zeigen die Äquivalenz der drei Aussagen mittels eines Ringschlusses.

**i) → ii):**
Nehmen wir an, dass $F$ diagonalisierbar ist. Dann ist jede darstellende Matrix von $F$ ähnlich zu einer Diagonalmatrix $D \in \mathbb{K}^{n\times n}$ auf deren Hauptdiagonalen die Eigenwerte $\lambda_1, \ldots, \lambda_k \in \mathbb{K}$ von $F$ mit algebraischen Vielfachheiten $r_1, \ldots, r_k \in \mathbb{N}$ stehen. Das charakteristische Polynom $P_D$ von $D$ zerfällt offensichtlich in Linearfaktoren über $\mathbb{K}$ in der Form

$$P_D(t) \;=\; (\lambda_1 - t)^{r_1} \cdot \ldots \cdot (\lambda_k - t)^{r_k}$$

und die Summe der algebraischen Vielfachheiten entspricht dem Grad des Polynoms, d.h., $\sum_{i=1}^{k} r_i = n$. Aus Lemma 2.24 wissen wir aber schon, dass das charakteristische Polynom von $F$ und $D$ gleich sein müssen.

Da $F$ diagonalisierbar ist existiert eine Basis von $V$ aus Eigenvektoren von $F$. Die Eigenvektoren dieser Basis können wir anhand ihrer zugehörigen Eigenwerte sortieren, so dass sich Basen der jeweiligen Eigenräume mit geometrischen Vielfachheiten $s_1, \ldots, s_k$ ergeben, d.h., wir betrachten die Eigenvektoren $v_1^i, \ldots, v_{s_i}^i \in V$ von $F$ zum Eigenwert $\lambda_i \in \mathbb{K}$ mit geometrischer Vielfachheit $s_i \in \mathbb{N}$ als Basis des Eigenraums $\mathrm{Eig}(F; \lambda_i)$ für $1 \leqslant i \leqslant k$. Daraus ergibt sich, dass $\sum_{i=1}^{k} s_i = n$ gelten muss, da wir von einer Basis von $V$ ausgegangen sind. Gleichzeitig wissen wir aus dem Argument von oben, dass $\sum_{i=1}^{k} r_k = n$ gelten muss und nach Satz 2.15 die algebraischen Vielfachheiten größer oder gleich den geometrischen Vielfachheiten sind, d.h., es gilt $r_i \geqslant s_i$. Diese drei Bedingungen können jedoch nur dann erfüllt werden, wenn schon gilt $r_i = s_i$.

**ii) → iii):**
Seien $\lambda_1, \ldots, \lambda_k \in \mathbb{K}$ die paarweise verschiedenen Eigenwerte von $F$, deren algebraische Vielfachheit gleich der geometrischen Vielfachheit ist, d.h., $r_i = s_i$ für $1 \leqslant i \leqslant k$. Wir nehmen an, dass das charakteristische Polynom $P_F$ von $F$ in Linearfaktoren über $\mathbb{K}$ zerfällt und von der Form ist

$$P_F(t) \;=\; (t - \lambda_1)^{r_1} \cdot \ldots \cdot (t - \lambda_k)^{r_s}.$$

Wir betrachten die lineare Hülle der Eigenräume $\mathrm{Eig}(F; \lambda_i), 1 \leqslant i \leqslant k$, von $F$

$$W \;:=\; \mathrm{lin}(\mathrm{Eig}(F; \lambda_1) \cup \ldots \cup \mathrm{Eig}(F; \lambda_k)) \subset V.$$

Da die geometrische Vielfachheit $s_i = r_i$ für $i = 1, \ldots, k$ ist, wird $W$ durch $n$ Eigenvektoren aufgespannt. Aus Satz 2.22 wissen wir, dass Vektoren aus verschiedenen Eigenräumen paarweise linear unabhängig sind. Damit folgt aber schon, dass $W$ eine direkte Summe der Eigenräume sein muss mit

$$W \;=\; \mathrm{Eig}(F; \lambda_1) \oplus \ldots \oplus \mathrm{Eig}(F; \lambda_k) = V.$$

**iii) → i):**

Sei $B_i = (v_1^i, \ldots, v_{s_i}^i)$ eine Basis aus Eigenvektoren zum Eigenwert $\lambda_i \in \mathbb{K}$ vom Eigenraum

$\mathrm{Eig}(F;\lambda_i)$ für $1 \leqslant i \leqslant k$. Da die Eigenräume als direkte Summe den ganzen Vektorraum $V$ bilden, wissen wir, dass

$$B \;:=\; (v_1^1, \ldots, v_{s_1}^1, \ldots, v_1^k, \ldots, v_{s_k}^k)$$

eine Basis von $V$ ist. Da diese Basis aus Eigenvektoren von $F$ besteht ist $F$ schon diagonalisierbar per Definition. □

**Korollar 2.27**
*Aus Satz 2.26 wird direkt klar, dass der Endomorphismus $F$ diagonalisierbar ist, wenn er $n \in \mathbb{N}$ paarweise verschiedene Eigenwerte besitzt. Dies ist eine hinreichende aber keineswegs notwendige Bedingung, wie etwa das Beispiel $F = \mathrm{Id}_V$ zeigt.*

**Beispiel 2.28**
*Im Folgenden wollen wir zwei Beispiele untersuchen in denen eine reellwerte Matrix $A \in \mathbb{R}^{3\times 3}$ nicht diagonalisierbar ist und die Gründe hierfür genauer beleuchten.*

1. *Sei die Matrix $A \in \mathbb{R}^{3\times 3}$ gegeben durch:*

$$A \;=\; \begin{pmatrix} 1 & -\sqrt{3} & 0 \\ \sqrt{3} & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

*Wir bestimmen das charakteristische Polynom $P_A$ von $A$ mit der Produktregel für Determinanten von Blockmatrizen in Lemma 2.7 als*

$$\begin{aligned} P_A(t) \;&=\; \det(A - tI_3) \;=\; \begin{pmatrix} 1-t & -\sqrt{3} & 0 \\ \sqrt{3} & -1-t & 0 \\ 0 & 0 & 1-t \end{pmatrix} \\ &=\; (1-t)\big((1-t)(-1-t) + \sqrt{3}\cdot\sqrt{3}\big) \;=\; (1-t)(t^2 - 1 + 3) \\ &=\; (1-t)(t^2+2). \end{aligned}$$

*Da das quadratische Polynom $(t^2+2)$ keine Nullstellen in $\mathbb{R}$ besitzt, lässt sich das charakteristische Polynom $P_A$ nicht vollständig in Linearfaktoren über $\mathbb{R}$ zerlegen. Daraus folgt mit Satz 2.26, dass die Matrix $A$ nicht diagonalisierbar ist.*

2. *Sei die Matrix $A \in \mathbb{R}^{3\times 3}$ gegeben durch:*

$$A \;=\; \begin{pmatrix} 3 & 4 & -3 \\ 2 & 7 & -4 \\ 3 & 9 & -5 \end{pmatrix}.$$

*Wir bestimmen das charakteristische Polynom $P_A$ von $A$ durch die Regel von Sarrus in Lemma 2.6 oder Umformung mittels Gaußschen Eliminationsverfahren und erhalten:*

$$P_A(t) \;=\; \det(A - tI_3) \;=\; -(t-2)^2 \cdot (t-1).$$

*Das charakteristische Polynom zerfällt also in Linearfaktoren über $\mathbb{R}$ und wir können die beiden Eigenwerte $\lambda_1 = 2$ und $\lambda_2 = 1$ ablesen. Hierbei bemerken wir, dass der Eigenwert $\lambda_1$ die algebraische Vielfachheit 2 und der Eigenwert $\lambda_2$ die algebraische Vielfachheit 1 besitzt. Außerdem kann man die zugehörigen Eigenräume wie folgt bestimmen:*

$$\mathrm{Eig}(A;\lambda_1) \;=\; \left\{ \alpha \cdot \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} \mid \alpha \in \mathbb{R} \right\}, \quad \mathrm{Eig}(A;\lambda_2) \;=\; \left\{ \alpha \cdot \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \mid \alpha \in \mathbb{R} \right\}.$$

*Wir sehen also, dass die geometrischen Vielfachheiten der Eigenwerte $\lambda_1, \lambda_2 \in \mathbb{R}$ jeweils 1 betragen und somit die algebraische Vielfachheit von $\lambda_1$ nicht mit der geometrischen Vielfachheit übereinstimmt. Daraus folgt mit Satz 2.26, dass die Matrix $A$ nicht diagonalisierbar ist.*

Das folgende Beispiel untersucht wann eine allgemeine $(2 \times 2)$-Matrix $A \in \mathbb{R}^{2\times 2}$ nicht diagonalisierbar ist.

**Beispiel 2.29**
*Sei $A \in \mathbb{R}^{2\times 2}$ eine quadratische Matrix mit*

$$A \;:=\; \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

*für $a, b, c, d \in \mathbb{R}$. Um zu untersuchen wann $A$ nicht diagonalisierbar ist, betrachten wir das charakteristische Polynom $P_A$ von $A$ mit*

$$\begin{aligned} P_A(t) \;&=\; \det(A - tI_2) \;=\; \det \begin{pmatrix} a-t & b \\ c & d-t \end{pmatrix} \;=\; (a-t)(d-t) - bc \\ &=\; t^2 - (a+d)t + (ad - bc). \end{aligned}$$

*Um die Nullstellen des charakteristischen Polynoms zu bestimmen verwenden wir in diesem einfachen Fall die p-q-Formel mit $p := -(a+d)$ und $q := (ad - bc)$, so dass für die Eigenwerte von $A$ gilt*

$$\lambda_{1/2} \;=\; -\frac{p}{2} \pm \sqrt{\frac{p^2}{4} - q} \;=\; \frac{(a+d)}{2} \pm \sqrt{\frac{(a+d)^2}{4} - (ad - bc)}.$$

*Wir bemerken zuerst, dass der Radikand $R$ unter der Wurzel von der folgenden Form ist*

$$R \;=\; \frac{(a+d)^2}{4} - (ad - bc) \;=\; \frac{\mathrm{Spur}(A)^2}{4} - \det(A).$$

*Nun macht es Sinn eine Fallunterscheidung nach dem Vorzeichen von $R$ zu machen.*

1. *Falls $R > 0$ gilt, so existieren zwei Lösungen der quadratischen Gleichung und somit zwei verschiedene Eigenwerte $\lambda_1 \neq \lambda_2$ von $A$. Nach Korollar 2.27 wissen wir, dass $A$ dann schon diagonalisierbar ist.*

2. *Falls $R < 0$ gilt, so liegt die Wurzel nicht mehr im Körper $\mathbb{R}$ und somit gibt es keine reellen Eigenwerte von $A$. Für diesen Fall ist $A$ nicht diagonalisierbar.*

3. *Der spannende Fall tritt ein für $R = 0$. Hier besitzt die Matrix $A$ nur einen Eigenwert $\lambda = \frac{(a+d)}{2}$ mit algebraischer Vielfachheit 2. Nach Satz 2.26 ist $A$ genau dann diagonalisierbar, wenn die geometrische Vielfachheit des zugehörigen Eigenraums $\mathrm{Eig}(A;\lambda)$ auch 2 beträgt. Wir betrachten also den Eigenraum zum Eigenwert $\lambda$ im Folgenden.*

$$\mathrm{Eig}(A;\lambda) \;=\; \mathrm{Kern}(A-\lambda I_2) \;=\; \mathrm{Kern}(A-\frac{(a+d)}{2}I_2) \;=\; \mathrm{Kern}\begin{pmatrix} \frac{(a-d)}{2} & b \\ c & \frac{(d-a)}{2} \end{pmatrix}.$$

*Wir versuchen also folgendes lineares Gleichungssystem für einen unbekannten Vektor $x = (x_1, x_2)^T \in \mathbb{R}^2$ zu lösen:*

$$\begin{pmatrix} \frac{(a-d)}{2} & b \\ c & \frac{(d-a)}{2} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \;=\; \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

*Mittels Gaußschen-Eliminationsverfahren bringen wir die Matrix in eine obere, rechte Dreiecksgestalt und erhalten so:*

$$\begin{pmatrix} c\frac{(a-d)}{2} & bc \\ 0 & \frac{-(a-d)^2}{4} - bc \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \;=\; \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

*Wir erkennen, dass der Eintrag unten, rechts in der Matrix von folgender Gestalt ist:*

$$\begin{aligned} \frac{-(a-d)^2}{4} - bc \;&=\; \frac{-a^2+2ad-d^2}{4} - bc \;=\; \frac{-a^2-2ad-d^2}{4} + (ad-bc) \\ &=\; \frac{-(a+d)^2}{4} + (ad-bc) \;=\; -\frac{\mathrm{Spur}(A)^2}{4} + \det(A) \;=\; -R \;=\; 0. \end{aligned}$$

*Das bedeutet, dass wir zur Bestimmung des Kern Lösungen des folgenden Gleichungssystems bestimmen müssen.*

$$\begin{pmatrix} c\frac{(a-d)}{2} & bc \\ 0 & 0 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \;=\; \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{2.13}$$

*Auch hier gibt es zwei Möglichkeiten. Falls die Matrix $A$ bereits in Diagonalgestalt ist, so steht ihr Eigenwert $\lambda$ auf der Hauptdiagonalen und es gilt $a = d = \lambda$ und*

$b = c = 0$. *Wie man leicht einsieht ist die Matrix in* (2.13) *dann die Nullmatrix und jeder Vektor* $x \in \mathbb{R}^2$ *löst das lineare Gleichungssystem. Also ist der Kern bereits der gesamte Vektorraum* $V = \mathbb{R}^2$ *und die geometrische Vielfachheit des Eigenwerts* $\lambda$ *ist in der Tat* 2. *Damit ist die Matrix trivialerweise diagonalisierbar.*

*In allen anderen Fällen können wir den Eigenraum explizit angeben als*

$$\mathrm{Eig}(A;\lambda) \;=\; \mathrm{lin}(\{\begin{pmatrix} bc \\ -c\frac{(a-d)}{2} \end{pmatrix}\}).$$

*Der Eigenraum* $\mathrm{Eig}(A;\lambda)$ *hat also die Dimension* 1 *und somit stimmen geometrische Vielfachheit und algebraische Vielfachheit nicht überein. Nach Satz 2.26 ist die Matrix* $A$ *also nicht diagonalisierbar.*

*Um ein konkretes Beispiel für den dritten Fall der Fallunterscheidung oben anzugeben, müssen wir eine Matrix* $A \in \mathbb{R}^{2\times 2}$ *konstruieren, so dass* $R = 0$ *ist, bzw., so dass gilt* $\mathrm{Spur}(A)^2 = 4\det(A)$. *Hierzu betrachten wir die folgende Matrix*

$$A \;:=\; \begin{pmatrix} 3 & -1 \\ 1 & 1 \end{pmatrix}.$$

*Wir sehen sofort ein, dass gilt*

$$\mathrm{Spur}(A)^2 \;=\; (3+1)^2 \;=\; 16 \;=\; 4\cdot(3\cdot 1-(-1)\cdot 1) \;=\; 4\det(A).$$

*Mit unseren allgemeinen Vorüberlegungen oben, können wir den Eigenwert* $\lambda$ *von* $A$ *angeben als*

$$\lambda \;=\; -\frac{a+d}{2} \;=\; -\frac{3+1}{2} \;=\; -2.$$

*Und der Eigenraum* $\mathrm{Eig}(A;-2)$ *von* $A$ *wird aufgespannt durch den Vektor*

$$\begin{pmatrix} bc \\ -c\frac{(a-d)}{2} \end{pmatrix} \;=\; \begin{pmatrix} -1\cdot 1 \\ -1\cdot\frac{(3-1)}{2} \end{pmatrix} \;=\; \begin{pmatrix} -1 \\ -1 \end{pmatrix}.$$

*Die Matrix* $A$ *ist nach Satz 2.26 nicht diagonalisierbar, da geometrische und algebraische Vielfachheit des Eigenwerts* $\lambda = -2$ *nicht übereinstimmen.*

Sollte eine quadratische Matrix $A \in \mathbb{K}^{n\times n}$ diagonalisierbar sein, so hat die reguläre Matrix $S^{-1} \in \mathrm{GL}(n;\mathbb{K})$ in (2.12) eine besondere Gestalt, wie das folgende Lemma zeigt.

**Lemma 2.30**
*Sei* $A \in \mathbb{K}^{n\times n}$ *eine diagonalisierbare Matrix. In diesem Fall sind die Spaltenvektoren von* $S^{-1}$ *gerade die Eigenvektoren der zugehörigen Eigenwerte auf der Diagonalen von* $D$.

*Beweis.* Da $A$ diagonalisierbar ist existiert eine Diagonalmatrix $D \in \mathbb{K}^{n\times n}$ und eine reguläre Matrix $S \in \mathrm{GL}(n;\mathbb{K})$, so dass $SAS^{-1} = D$ gilt. Aus Lemma 2.24 wissen wir, dass die Eigenwerte $\lambda_1, \ldots, \lambda_n$ von $A$ durch die Einträge auf der Diagonalen von $D$, gegeben sind. Durch Multiplikation mit der Matrix $S^{-1}$ von links erhalten wir also:

$$\underbrace{S^{-1}S}_{=I_n} AS^{-1} \;=\; AS^{-1} \;=\; S^{-1}D.$$

Wir sehen also, dass $AS^{-1} = S^{-1}D$ gilt. Sei nun $v_k \in \mathbb{K}^n$ die $k$-te Spalte von $S^{-1}$ mit $1 \leqslant k \leqslant n$, dann sieht man ein, dass gilt

$$Av \;=\; \lambda_k v_k, \qquad \text{für alle } 1 \leqslant k \leqslant n.$$

Nach Definition 2.8 ist der Vektor $v_k$ also gerade der Eigenvektor zum Eigenwert $\lambda_k$ von $A$. □

Im folgenden Beispiel wollen wir diagonalisierbare $(2 \times 2)$-Matrizen untersuchen und die Beobachtung aus Lemma 2.30 verifizieren.

**Beispiel 2.31**
*Wir betrachten zwei Beispiele von $(2\times 2)$-Matrizen, für die wir eine ähnliche Diagonalmatrix $D$ berechnen wollen, auf deren Hauptdiagonalen die zugehörigen Eigenwerte stehen.*

1. *Für eine Matrix der Form*

$$A \;=\; \begin{pmatrix} -1 & 6 \\ -1 & 4 \end{pmatrix}$$

*bestimmen wir das charakteristische Polynom $P_A$ als*

$$P_A(t) \;=\; \det(A - tI_2) \;=\; (-1-t)(4-t)+6 \;=\; t^2 - 3t + 2 \;=\; (t-1)(t-2).$$

*Wir sehen also, dass das charakteristische Polynom $P_A$ in Linearfaktoren zerfällt und die Eigenwerte von $A$ gegeben sind durch $\lambda_1 = 1$ und $\lambda_2 = 2$. Es ist auf Grund von Satz 2.26 klar, dass $A$ diagonalisierbar ist, d.h., dass es eine reguläre Matrix $S \in \mathrm{GL}(2;\mathbb{K})$ gibt, so dass $SAS^{-1} = D$ gilt. Die Diagonalmatrix $D$ ist damit bis auf Permutation der Hauptdiagonale eindeutig bestimmt als*

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

*Aus Lemma 2.30 wissen wir, dass die Spalten von $S^{-1}$ gerade die Eigenvektoren von $A$ sind. Für die Bestimmung der Eigenräume $\mathrm{Eig}(A;\lambda_1)$ und $\mathrm{Eig}(A;\lambda_2)$ lösen wir die beiden homogenen Gleichungssysteme*

$$\begin{pmatrix} -2 & 6 \\ -1 & 3 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \;=\; (A - \lambda_1 I_2)\vec{v} = \vec{0},$$
$$\begin{pmatrix} -3 & 6 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \;=\; (A - \lambda_2 I_2)\vec{v} = \vec{0}.$$

*Wir sehen direkt, dass die jeweiligen Zeilen der beiden Matrizen linear abhängig sind und man den Eigenvektor zum Eigenwert $\lambda_1 = 1$ angeben kann als $v = (3,1)^T$ bzw. den Eigenvektor zum Eigenwert $\lambda_2 = 2$ als $v = (2,1)^T$. Schreiben wir die Eigenvektoren als Spalten der Matrix $S^{-1}$, so erhalten wir*

$$S^{-1} = \begin{pmatrix} 3 & 2 \\ 1 & 1 \end{pmatrix}$$

*Der Vollständigkeit halber bestimmen wir nun noch die Inverse $S$ zu $S^{-1}$ durch die Determinanten-Regel:*

$$S = (S^{-1})^{-1} = \begin{pmatrix} 3 & 2 \\ 1 & 1 \end{pmatrix}^{-1} = \frac{1}{3 \cdot 1 - 2 \cdot 1} \begin{pmatrix} 1 & -2 \\ -1 & 3 \end{pmatrix} = \begin{pmatrix} 1 & -2 \\ -1 & 3 \end{pmatrix}.$$

*Wir überprüfen unsere Rechnung abschließend durch das Diagonalisieren von $A$ als*

$$SAS^{-1} = \begin{pmatrix} 1 & -2 \\ -1 & 3 \end{pmatrix} \begin{pmatrix} -1 & 6 \\ -1 & 4 \end{pmatrix} \begin{pmatrix} 3 & 2 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} = D.$$

2. *Für eine Spiegelmatrix der Form*

$$A = \begin{pmatrix} \cos\alpha & \sin\alpha \\ \sin\alpha & -\cos\alpha \end{pmatrix}$$

*berechnen wir das charakteristische Polynom $P_A$ als*

$$\begin{aligned} P_A(t) &= \det(A - tI_2) = -(\cos\alpha - t)(\cos\alpha + t) - \sin^2\alpha = t^2 - \cos^2\alpha - \sin^2\alpha \\ &= t^2 - \underbrace{(\sin^2\alpha + \cos^2\alpha)}_{=1} = t^2 - 1 = (t-1)(t+1). \end{aligned}$$

*Wir sehen also, dass das charakteristische Polynom $P_A$ in Linearfaktoren zerfällt und die Eigenwerte dieser allgemeinen Spiegelmatrix unabhängig sind von der Wahl des Winkels $\alpha \in [0, 2\pi)$ immer $\lambda_1 = 1$ und $\lambda_2 = -1$. Es ist auf Grund von Satz 2.26 klar, dass $A$ diagonalisierbar ist, d.h., dass es eine reguläre Matrix $S \in \mathrm{GL}(2;\mathbb{K})$ gibt, so dass $SAS^{-1} = D$ gilt. Die Diagonalmatrix $D$ ist damit bis auf Permutation der Hauptdiagonale eindeutig bestimmt als*

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

*Aus Lemma 2.30 wissen wir, dass die Spalten von $S^{-1}$ gerade die Eigenvektoren von $A$ sind. Für die Bestimmung der Eigenräume $\mathrm{Eig}(A;\lambda_1)$ und $\mathrm{Eig}(A;\lambda_2)$ lösen wir die*

*beiden homogenen Gleichungssysteme*

$$\begin{pmatrix} \cos\alpha - 1 & \sin\alpha \\ \sin\alpha & -\cos\alpha - 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = (A - \lambda_1 I_2)\vec{v} = \vec{0},$$
$$\begin{pmatrix} \cos\alpha + 1 & \sin\alpha \\ \sin\alpha & -\cos\alpha + 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = (A - \lambda_2 I_2)\vec{v} = \vec{0}.$$

*Durch Multiplikation der unteren Zeile mit dem ersten Eintrag der ersten Zeile sieht man, dass beide Zeilen linear abhängig sind und man kann die Eigenvektoren als Lösungen der obigen Gleichungen ablesen. Für den Eigenwert* $\lambda_1 = 1$ *erhält man den zugehörigen Eigenvektor* $\vec{v}_1 = (\sin^2\alpha, \sin\alpha[1 - \cos\alpha])^T$ *und für den zweiten Eigenwert* $\lambda_2 = -1$ *erhält man entsprechend den zugehörigen Eigenvektor* $\vec{v}_2 = (\sin^2\alpha, -\sin\alpha[1 + \cos\alpha])^T$*. Damit ist die Transformationsmatrix* $S^{-1}$ *gegeben durch*

$$S^{-1} = \begin{pmatrix} \sin^2\alpha & \sin^2\alpha \\ \sin\alpha(1 - \cos\alpha) & -\sin\alpha(1 + \cos\alpha) \end{pmatrix}.$$

*Das Bestimmen der Transformationsmatrix* $S$ *und die Überprüfung der Diagonalisierung von* $A$ *überlassen wir an dieser Stelle der geneigten Leserin und machen dafür folgende Beobachtung. Die Vektoren* $\vec{v}_1, \vec{v}_2 \in \mathbb{R}^2$ *bilden eine Orthogonalbasis von* $\mathbb{R}^2$*, da gilt:*

$$\begin{aligned} \langle \vec{v}_1, \vec{v}_2 \rangle &= \langle (\sin^2\alpha, \sin\alpha[1 - \cos\alpha])^T, (\sin^2\alpha, -\sin\alpha[1 + \cos\alpha])^T \rangle \\ &= \sin^4\alpha - \sin\alpha(1 - \cos\alpha) \cdot \sin\alpha(1 + \cos\alpha) \\ &= \sin^4\alpha - \sin^2\alpha \underbrace{(1 - \cos^2\alpha)}_{=\sin^2\alpha} = \sin^4\alpha - \sin^4\alpha = 0. \end{aligned}$$

## 2.6 Trigonalisierbarkeit

Aus Kapitel 2.5 wissen wir nun, dass ein Endomorphismus genau dann diagonalisierbar ist, wenn das zugehörige charakteristische Polynom in Linearfaktoren über $\mathbb{K}$ zerfällt und die algebraischen und geometrischen Vielfachheiten der Eigenwerte übereinstimmen. Das Beispiel 2.28 hat uns jedoch gezeigt, dass es sehr wohl Matrizen gibt deren charakteristisches Polynom zerfällt, jedoch nicht diagonalisierbar sind, da die Vielfachheiten der Eigenwerte nicht übereinstimmen. Glücklicherweise werden wir im Folgenden jedoch feststellen, dass Endomorphismen, deren charakteristisches Polynom in Linearfaktoren zerfällt, zumindest ähnlich zu einer oberen, rechten Dreiecksmatrix sind.

**Definition 2.32** (Trigonalisierbarkeit)
*Wir definieren den Begriff der* Trigonalisierbarkeit *im Folgenden sowohl für Endomorphismen als auch für Matrizen.*

1. *Ein Endomorphismus $F$ von $V$ heißt* trigonalisierbar*, wenn es eine Basis $B$ von $V$ gibt, so dass die darstellende Matrix $M_B(F)$ von $F$ bezüglich $B$ eine obere, rechte Dreiecksmatrix ist.*

2. *Eine Matrix $A \in \mathbb{K}^{n\times n}$ heißt* trigonalisierbar*, wenn sie ähnlich zu einer oberen, rechten Dreiecksmatrix ist.*

Es ist klar, dass wenn eine Matrix $A$ trigonalisierbar ist, so stehen auf der oberen, rechten Dreiecksmatrix. zu der $A$ ähnlich ist, die Eigenwerte von $A$. Der folgende Satz gibt uns ein Kriterium mit dessen Hilfe wir direkt entscheiden können, ob ein Endomorphismus trigonalisierbar ist oder nicht.

**Satz 2.33** (Trigonalisierungssatz)
*Für einen Endormorphismus $F$ eines $n$-dimensionalen $\mathbb{K}$-Vektorraums $V$ sind folgende Bedingungen äquivalent:*

*i) $F$ ist trigonalisierbar*

*ii) Das charakteristische Polynom $P_F$ zerfällt in Linearfaktoren über $\mathbb{K}$, d.h.,*

$$P_F(t) \;=\; \pm(t-\lambda_1)\cdot\ldots\cdot(t-\lambda_n).$$

*Beweis.* Siehe [Fischer2005, Satz 4.4.3] □

**Korollar 2.34**
*Der Fundamentalsatz der Algebra [Fischer2005, Theorem 1.3.9] besagt, dass jedes Polynom über $\mathbb{C}$ in Linearfaktoren zerfällt. Hieraus folgt schon, dass jeder Endomorphismus von einem endlich-dimensionalen $\mathbb{C}$-Vektorraum trigonalisierbar ist.*

**Algorithmus 2.35** (Trigonalisierung einer Matrix)
*Sei $A \in \mathbb{K}^{n\times n}$ eine Matrix deren charakteristisches Polynom in Linearfaktoren zerfällt, d.h.,*

$$P_A(t) \;=\; (t-\lambda_1)\cdot\ldots\cdot(t-\lambda_n),$$

*wobei die Eigenwerte $\lambda_1,\ldots,\lambda_n \in \mathbb{K}$ von $A$ nicht paarweise verschieden sein müssen. Wir versuchen nun eine Matrix $S \in \mathrm{GL}(n;\mathbb{K})$ zu bestimmen, so dass*

$$D \;=\; SAS^{-1},$$

*wobei $D \in \mathbb{K}^{n\times n}$ eine obere, rechte Dreiecksmatrix ist. Für diese Trigonalisierung gehen wir folgt vor:*

***1. Schritt:***

*Wir betrachten zunächst den Vektorraum $W_1 := \mathbb{K}^n$ mit der kanonischen Einheitsbasis $B_1$ und den Endomorphismus $A_1 := A$. Zunächst berechnen wir den zugehörigen Eigenvektor $v_1 \in \mathbb{K}^n$ zum Eigenvektor $\lambda_1$ von $A$. Nach dem Basisaustauschlemma [Fischer2005, Lemma 1.5.4] können wir ein Element der kanonischen Einheitsbasis durch $v_1$ austauschen, so dass wir immer noch eine Basis erhalten. Das heißt, wir bestimmen einen Index $1 \leqslant j_1 \leqslant n$, so dass*

$$B_2 \;:=\; (v_1, e_1, \ldots, \widehat{e_{j_1}}, \ldots, e_n)$$

*wieder eine Basis von $\mathbb{K}^n$ ist. Hierbei bedeutet die Notation $\widehat{e_{j_1}}$, dass dieses Basiselement ausgetauscht wurde.*

*Wir schreiben nun die Basiselemente von $B_2$ als Spalten der Transformationsmatrix*

$$S_1^{-1} \;:=\; T_{B_1}^{B_2}.$$

*Dann wird klar, dass wir die erste Spalte von $A_1$ in obere rechte Dreiecksform bringen können durch:*

$$A_2 \;:=\; S_1 \cdot A \cdot S_1^{-1} \;=\; \begin{pmatrix} \lambda_1 & * & \ldots & * \\ 0 & & & \\ \vdots & & A_2' & \\ 0 & & & \end{pmatrix}.$$

***2. Schritt:***

*Nun betrachten wir $W_2 \subset \mathbb{K}^n$ mit der Basis*

$$B_2' \;:=\; (e_1, \ldots, \widehat{e_{j_1}}, \ldots, e_n)$$

*und dem Endomorphismus $A_2' \colon W_2 \to W_2$. Das charakteristische Polynom von $A_2'$ ist wegen der Determinantenregel für Blockmatrizen in Lemma 2.7 gegeben durch*

$$P_{A_2'}(t) \;=\; (t - \lambda_2) \cdot \ldots \cdot (t - \lambda_n).$$

*Wir berechnen zum Eigenwert $\lambda_2 \in \mathbb{K}$ wieder einen Eigenvektor $v_2 \in W_2$ und nutzen das Basisaustauschlemma, um einen Index $j_2 \neq j_1$ zu finden, für den*

$$B_3' \;:=\; (v_2, e_1, \ldots, \widehat{e_{j_1}}, \ldots, \widehat{e_{j_2}}, \ldots, e_n)$$

*wieder eine Basis von $W_2$ ist. Damit ist auch*

$$B_3 \;:=\; (v_1, v_2, e_1, \ldots, \widehat{e_{j_1}}, \ldots, \widehat{e_{j_2}}, \ldots, e_n)$$

*eine Basis von $W_1 = \mathbb{K}^n$. Schreiben wir wiederum die Basis $B_3$ als Spalten der Transformationsmatrix*

$$S_2^{-1} := \; T_{B_1}^{B_3},$$

*so erhalten wir*

$$A_3 \;:=\; S_2 \cdot A \cdot S_2^{-1} \;=\; \begin{pmatrix} \lambda_1 & * & \dots & \dots & * \\ 0 & \lambda_2 & * & \dots & * \\ \vdots & 0 & & & \\ \vdots & \vdots & & A_3' & \\ 0 & 0 & & & \end{pmatrix}.$$

*Bei der Berechnung der Matrix $S_2$ kann man ausnutzen, dass*

$$T_{B_1}^{B_3} \;=\; T_{B_1}^{B_2} \cdot T_{B_2}^{B_3}, \qquad \textit{wobei}\;\; T_{B_2}^{B_3} \;:=\; \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & T_{B_2'}^{B_3'} & \\ 0 & & & \end{pmatrix}$$

*Spätestens nach $(n-1)$ Schritten erhält man eine obere, rechte Dreiecksmatrix, da $A_n'$ eine $(1 \times 1)$-Matrix ist. Insgesamt erhalten wir also eine zu $A$ ähnliche obere, rechte Dreiecksmatrix $D$ durch*

$$D \;:=\; A_n \;=\; S_{n-1} \cdot A \cdot S_{n-1}^{-1}.$$

Das folgende Beispiel zeigt, wie Algorithmus 2.35 angewendet werden kann, um eine Matrix, deren charakteristisches Polynom in Linearfaktoren zerfällt, zu trigonalisieren.

**Beispiel 2.36**
*Sei $A \in \mathbb{R}^{3\times 3}$ eine quadratische Matrix, die wir trigonalisieren wollen, mit*

$$A \;:=\; \begin{pmatrix} 3 & 4 & 3 \\ -1 & 0 & -1 \\ 1 & 2 & 3 \end{pmatrix}.$$

*Wir bestimmen zuerst das charakteristische Polynom $P_A$ von $A$ als*

$$P_A(t) \;=\; -(t-2)^3,$$

*also ist $\lambda = 2$ der einzige Eigenwert von $A$ mit algebraischer Vielfachheit 3. Das charakteristische Polynom zerfällt offensichtlich in Linearfaktoren und nach Satz 2.33 wissen wir folglich, dass $A$ trigonalisierbar ist.*

*Wir betrachten den Eigenraum* $\mathrm{Eig}(A;2)$ *von $A$ durch Bestimmung von* $\mathrm{Kern}(A - 2\cdot I_3)$ *als*

$$\mathrm{Eig}(A;2) \;=\; \mathrm{Kern}(A - 2\cdot I_3) \;=\; \mathrm{Kern}\begin{pmatrix} 3-2 & 4 & 3 \\ -1 & 0-2 & -1 \\ 1 & 2 & 3-2 \end{pmatrix} \;=\; \mathrm{Kern}\begin{pmatrix} 1 & 4 & 3 \\ -1 & -2 & -1 \\ 1 & 2 & 1 \end{pmatrix}.$$

*Wir sehen direkt, dass die letzten beiden Zeilen der Matrix* $(A-2\cdot I_3)$ *linear abhängig sind und sie daher den Rang* 2 *hat. Daraus folgt mit der Dimensionsformel von Bild und Kern [Fischer2005, Satz 2.2.4], dass der Kern von* $A-2\cdot I_3$ *die Dimension* 1 *besitzt und daher der Eigenwert* $\lambda=2$ *die geometrische Vielfachheit* 1 *hat. Da geometrische und algebraische Vielfachheit des Eigenwerts nicht übereinstimmen wissen wir mit Satz 2.26, dass die Matrix* $A$ *nicht diagonalisierbar ist. Konkret können wir den Eigenvektor* $v_1\in\mathbb{R}^3$ *zum Eigenwert* $\lambda=2$ *angeben als* $v_1=(1,-1,1)^T$.

*Wir wenden nun den Algorithmus 2.35 an, um die Matrix* $A$ *zu trigonalisieren.*

***1. Schritt:***

*Wir wählen den Index* $j_1=1$, *d.h., wir tauschen den Einheitsvektor* $e_1$ *durch* $v_1$ *aus und erhalten* $B_2$ *als neue Basis von* $W_1=\mathbb{R}^3$ *mit*

$$B_2 \;:=\; (v_1,e_2,e_3).$$

*Daraus ergibt sich für die Transformationmatritzen*

$$S_1^{-1} \;:=\; T_{B_1}^{B_2} \;=\; \begin{pmatrix}1&0&0\\-1&1&0\\1&0&1\end{pmatrix}, \qquad S_1 \;=\; \begin{pmatrix}1&0&0\\1&1&0\\-1&0&1\end{pmatrix}.$$

*Damit erhalten wir für*

$$A_2 \;:= S_1\cdot A\cdot S_1^{-1} \;=\; \begin{pmatrix}2&4&3\\0&4&2\\0&-2&0\end{pmatrix}.$$

***2. Schritt:***

*Wir betrachten den verbleibenden Block* $A_2'$ *von* $A_2$, *der noch nicht in oberer, rechter Dreiecksgestalt ist mit*

$$A_2' \;:=\; \begin{pmatrix}4&2\\-2&0\end{pmatrix}.$$

*Wir bestimmen den Eigenvektor* $v_2'\in\mathbb{R}^2$ *von* $A_2'$ *zum Eigenwert* $\lambda_2=2$ *als* $v_2'=(1,-1)^T$. *Die Basis* $B_2'$ *ist gegeben durch* $B_2'=(e_1,e_2)$ *und wir tauschen das erste Basiselement durch den Eigenvektor* $v_2'$ *aus, so dass wir eine neue Basis* $B_3'=(v_2',e_2)$ *erhalten. Die entsprechende Transformationsmatrix ist dann gegeben durch*

$$T_{B_2'}^{B_3'} \;:=\; \begin{pmatrix}1&0\\-1&1\end{pmatrix}.$$

*Bezogen auf die Basen $B_2 = (v_1, e_2, e_3)$ von $W_1$ können wir das zweite Basiselement $e_2$ durch den um 0 erweiterten Eigenvektor $v_2'$ ersetzen und erhalten $B_3$ als neue Basis mit*

$$B_3 \;:=\; (v_1, v_2, e_3),$$

*wobei $v_2 := (0, 1, -1)^T$ ist. Wir betten $T_{B_2'}^{B_3'}$ als unteren, rechten Block in eine Transformationsmatrix von Basis $B_2$ zur neuen Basis $B_3$ ein und erhalten*

$$T_{B_2}^{B_3} \;:=\; \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}.$$

*Damit können wir nun die globale Transformationsmatrix $S_2^{-1}$ bestimmen durch*

$$S_2^{-1} \;=\; S_1^{-1} \cdot T_{B_2}^{B_3} \;=\; \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & -1 & 1 \end{pmatrix}, \qquad S_2 \;=\; \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}.$$

*Damit erhalten wir schlussendlich eine zu $A$ ähnliche Triagonalmatrix $D$ für*

$$A_3 \;:= S_2 \cdot A \cdot S_2^{-1} \;=\; \begin{pmatrix} 2 & 1 & 3 \\ 0 & 2 & 2 \\ 0 & 0 & 2 \end{pmatrix} \;=\; D.$$

Wie wir in Beispiel 2.36 gesehen haben, können wir eine trigonalisierbare Matrix zwar in eine obere, rechte Dreiecksform überführen, jedoch gibt es viele Möglichkeiten dies zu tun, abhängig von der Wahl der auszutauschenden Basiselemente in Algorithmus 2.35. Da Mathematiker im Allgemeinen allergisch auf Wahlfreiheit reagieren, wollen wir uns im Folgenden mit der Bestimmung eines kanonischen Repräsentanten für eine ähnliche obere, rechte Dreiecksmatrix beschäftigen, die besonders schöne Eigenschaft hat.

## 2.7 Die Jordansche Normalform

An einer Triagonalmatrix können wir bereits Rang und Eigenwerte ablesen, jedoch keine weiteren charakteristischen Eigenschaften des zu Grunde liegenden Endomorphismus, wie z.B. die Dimension der Eigenräume. Außerdem kann die obere, rechte Dreiecksmatrix voll besetzt in der oberen Hälfte sein, so dass sie für numerische Verfahren eher ungünstig ist. Das wirkt sich vor allem bei der Potenzierung von Endomorphismen in iterativen Verfahren aus, bei der diese Mehrfachanwendung zu unerwünschten Rechenoperationen führt. Ein weiteres Problem ist, das allgemeine Triagonalmatrizen ungeeignet sind um explizite Lösungen von Systemen linearer Differentialgleichungen anzugeben, da die Gleichungen nicht hinreichend entkoppeln in dieser Darstellung.

Es ist also ganz natürlich sich die Frage zu stellen, ob für eine gegebene trigonalisierbare Matrix $A$ eine kanonische Wahl einer oberen, rechten Dreiecksmatrix existiert, die einfach und interpretierbar aufgebaut ist und nur wenige Wahlmöglichkeiten bei der Bestimmung zulässt. Die Frage, wie durch geschickte Wahl einer Basis $B$ von des endlich-dimensionalen Vektorraums $V$ die darstellende Matrix $M_B(F)$ des Endomorphismus $F$ auf eine möglichst einfache und eindeutige Gestalt gebracht werden kann, ist allerdings deutlich schwieriger als die bereits bekannte Triagonalisierung einer Matrix. Diese Frage wird zentral in der Normalformentheorie von Endomorphismen behandelt. Wie wir im Folgenden sehen werden existiert glücklicherweise eine kanonische Darstellung einer trigonalisierbaren Matrix, welche *Jordansche Normalform* genannt wird, und die die gewünschten Eigenschaften hat.

Bevor wir uns jedoch mit dem Studium dieser Normalform beschäftigen führen die Definition eines nilpotenten Endomorphismus ein.

**Definition 2.37** (Nilpotenz)
*Wir definieren den Begriff der* Nilpotenz *im folgenden sowohl für Endomorphismen als auch für Matrizen.*

1. *Ein Endormorphismus $F\colon V \to V$ eines $\mathbb{K}$-Vektorraums $V$ heißt* nilpotent*, falls es einen Index $k \in \mathbb{N}$ gibt, so dass $F^k = \underbrace{F \circ \ldots \circ F}_{k\text{-fache Anwendung}} = 0$ ist.*
   *Der kleinste solche Index $k$ heißt dann* Nilpotenzindex*.*

2. *Eine Matrix $A \in \mathbb{K}^{n\times n}$ heißt* nilpotent*, falls es einen Index $k \in \mathbb{N}$ gibt, so dass $A^k = \underbrace{A \cdot \ldots \cdot A}_{k\text{-fache Anwendung}} = 0$ ist.*
   *Der kleinste solche Index $k$ heißt dann* Nilpotenzindex*.*

**Beispiel 2.38**
*Wir wollen im Folgenden den Nilpotenzindex zweier Matrizen durch ihre Potenzierung bestimmen.*

1. *Wir betrachten die Matrix $A \in \mathbb{R}^{3\times 3}$ mit*

$$A := \begin{pmatrix} 5 & -3 & 2 \\ 15 & -9 & 6 \\ 10 & -6 & 4 \end{pmatrix}.$$

   *Wir betrachten Potenzen von $A$ und erhalten:*

$$A^2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

   *Erstaunlicherweise ist der Nilpotenzindex der Matrix $A$ schon $k = 2$.*

*2. Wir betrachten die Matrix $A \in \mathbb{R}^{4\times 4}$ mit*

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

*Wir betrachten Potenzen von $A$ und erhalten:*

$$A^2 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad A^3 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad A^4 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

*Der Nilpotenzindex der Matrix $A$ ist also $k = 4$.*

Die Matrix $A$ aus dem zweiten Beispiel 2.38 ist in einer besonderen Form, welche wir näher betrachten werden.

**Definition 2.39** (Normalform für nilpotente Matrizen)
*Sei $k \in \mathbb{N}, k \geqslant 1$, dann definieren wir eine* nilpotente Matrix in Normalform $N_k \in \mathbb{K}^{k\times k}$ *durch:*

$$(N_k)_{i,j} := \delta_{i,j-1} = \begin{cases} 1, & \text{falls } i = j-1, \\ 0, & \text{sonst .} \end{cases}$$

*Hierbei bezeichnet $\delta_{i,j}$ das Kronecker-Delta. Das heißt $N_k$ ist eine Matrix, die nur auf der oberen, ersten Nebendiagonale Einsen besitzt und deren sonstige Einträge alle Null sind. Diese Normalform von nilpotenten Matrizen wird auch* Jordanmatrix *genannt.*

Wir wollen im Folgenden einige nützliche Eigenschaften von nilpotenten Matrizen angeben.

**Lemma 2.40**
*Sei $A \in \mathbb{K}^{n\times n}$ eine strikte obere, rechte Dreiecksmatrix, d.h., für die Diagonalelemente gilt $a_{ii} = 0, 1 \leqslant i \leqslant n$. Dann ist $A$ nilpotent ist und besitzt einen Nilpotenzindex von $k \leqslant n$.*

*Beweis.* Wir beweisen die Behauptung per vollständige Induktion über die Dimension $n$ von $A$:

**Induktionsanfang $n = 1$:**
Falls $n = 1$, so $A = (0)$ und $A$ ist offensichtlich nilpotent mit Index $1 \leqslant n$.

**Induktionsschritt $n - 1 \to n$, $n > 1$:**
Wenn $A \in \mathbb{K}^{n\times n}$ eine strikte obere Dreiecksmatrix ist, dann gibt es eine strikte obere

Dreiecksmatrix $A' \in \mathbb{K}^{(n-1)\times(n-1)}$, so dass

$$A = \begin{pmatrix} & & \star \\ & A' & \vdots \\ & & \star \\ 0_{n-1} & & 0 \end{pmatrix}.$$

Hierbei kennzeichnet $\star$ einen Eintrag, welcher nicht notwendigerweise Null ist. Wir sehen, dass

$$A^2 = \begin{pmatrix} (A')^2 & A'(\star,\ldots,\star)^T \\ 0_{n-1} & 0 \end{pmatrix}.$$

Per Induktionsvoraussetzung ist $A'$ nilpotent mit Nilpotenzindex $\ell \leqslant n-1$. Wir rechnen nun

$$A^\ell = \begin{pmatrix} (A')^\ell & (A')^{\ell-1}(\star,\ldots,\star)^T \\ 0_{n-1} & 0 \end{pmatrix} = \begin{pmatrix} \mathbf{0}_{(n-1)\times(n-1)} & (A')^{\ell-1}(\star,\ldots,\star)^T \\ 0_{n-1} & 0 \end{pmatrix}.$$

Eine weitere Multiplikation mit $A$ von links zeigt

$$A\begin{pmatrix} \mathbf{0}_{(n-1)\times(n-1)} & (A')^{\ell-1}(\star,\ldots,\star)^T \\ 0_{n-1} & 0 \end{pmatrix} = \begin{pmatrix} \mathbf{0}_{(n-1)\times(n-1)} & (A')^{\ell}(\star,\ldots,\star)^T \\ 0_{n-1} & 0 \end{pmatrix} = \mathbf{0}_{n\times n}$$

und der Nilpotenzindex von $A$ ist höchstens $\ell + 1 \leqslant n$. □

Wir wollen im folgenden Satz Kriterien herleiten, die aussagen wann eine obere, rechte Dreiecksmatrix nilpotent ist.

**Satz 2.41**
*Sei $A \in \mathbb{K}^{n\times n}$ eine obere, rechte Dreiecksmatrix. Dann gelten die folgenden Aussagen:*

*i) A ist genau dann nilpotent, wenn alle Diagonalelemente $a_{ii}, 1 \leqslant i \leqslant n$, gleich 0 sind.*

*ii) A ist genau dann nilpotent und diagonalisierbar, wenn A die Nullmatrix $\mathbb{0} \in \mathbb{K}^{n\times n}$ ist.*

*iii) Ist A nilpotent, so hat A nur den Eigenwert 0.*
*Beweis.*

i) Man kann leicht zeigen, dass für eine obere, rechte Dreiecksmatrix $A$ stets gilt:

$$A^j = \begin{pmatrix} a_{11}^j & & * \\ & \ddots & \\ 0 & & a_{nn}^j \end{pmatrix}.$$

Damit $A$ nilpotent ist, muss also für alle Diagonalelemente $a_{ii} = 0, 1 \leqslant i \leqslant n$, gelten.

Sei umgekehrt $A$ eine obere, rechte Dreiecksmatrix deren Diagonalelemente $a_{ii} = 0$ sind für $1 \leqslant i \leqslant n$. Dann folgt die Behauptung direkt mit Lemma 2.40.

ii) Falls $A = 0$ die Nullmatrix ist, so ist $A$ nilpotent vom Index 1 und trivialerweise diagonalisierbar.

Sei umgekehrt $A$ nilpotent und diagonalisierbar, dann existiert eine reguläre Matrix $S \in \mathrm{GL}(n;\mathbb{K})$, so dass

$$S \cdot A \cdot S^{-1} \;=\; \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \;=:\; D,$$

wobei $\lambda_i, 1 \leqslant i \leqslant n$, die Eigenwerte von $A$ sind. Sei $k$ der Nilpotenzindex von $A$. Wie man leicht einsieht gilt

$$D^k \;=\; (S \cdot A \cdot S^{-1})^k \;=\; S \cdot A^k \cdot S^{-1} \;=\; \begin{pmatrix} \lambda_1^k & & \\ & \ddots & \\ & & \lambda_n^k \end{pmatrix}.$$

Da $A^k = 0$ die Nullmatrix ist, folgt schon, dass $\lambda_i = 0, 1 \leqslant i \leqslant n$, gelten muss. Darum ist auch

$$A \;=\; S^{-1} \cdot D \cdot S \;=\; 0.$$

iii) Wir führen einen Beweis über Widerspruch. Nehmen wir an, dass die Behauptung nicht gelte, dann existiert ein Eigenwert $\lambda \neq 0$ und ein zugehörigen Eigenvektor $v \neq \vec{0}$, so dass $Av = \lambda v$.

Da $A$ nilpotent ist, existiert ein Index $k \in \mathbb{N}$, so dass $A^{k-1} \neq 0$, jedoch $A^k = 0$ gilt. Aus der Eigenwertgleichung können wir folgern, dass

$$0 \;=\; A^k v \;=\; A^{k-1}\lambda v \;=\; \lambda^k v.$$

Daraus folgt aber, dass $\lambda = 0$ oder $v = 0$ gilt, was zum Widerspruch führt.

□

Der folgende Satz sagt uns, dass wir für jeden nilpotenten Endomorphismus eine darstellende Matrix finden können, die eine strikte obere, rechte Dreiecksgestalt besitzt.

**Satz 2.42**
*Sei $V$ ein endlich-dimensionaler $\mathbb{K}$-Vektorraum und $F\colon V \to V$ ein nilpotenter Endomorphismus von $V$. Dann existiert eine Basis $B$ von $V$, so dass die darstellende Matrix $M_B(F)$ von $F$ bezüglich $B$ eine obere, rechte Dreiecksmatrix mit Nullen auf der Hauptdiagonale ist, d.h.*

$$M_B(F) \;=\; \begin{pmatrix} 0 & & * \\ & \ddots & \\ 0 & & 0 \end{pmatrix}$$

*und es gilt $P_F(t) = (-1)^n t^n$.*

*Beweis.* Wir führen den Beweis durch Induktion über $n = \dim V$.

**Induktionsanfang:** $n = 1$

Die Aussage ist trivialerweise erfüllt, da für einen nilpotenten Endomorphismus $F$ eines eindimensionalen Vektorraums $V$ gelten muss $F \equiv 0$. Dadurch ist die darstellende Matrix für eine beliebige Basis $B$ von $V$ gegeben durch $M_B(F) = 0$ und das charakteristische Polynom ist dementsprechend $P_F(t) = 0 - t = (-1)^1 \cdot t^1$.

**Induktionsschritt:** $n - 1 \to n$

Die Induktionsannahme ist, dass die Aussage bereits für den Fall $n - 1$ gezeigt wurde. Sei $F$ nun ein nilpotenter Endomorphismus von $V$ mit $F \not\equiv 0$ (da ansonsten die Situation vom Induktionsanfang vorliegt). Da nach Satz 2.41 Null der einzige Eigenwert von $F$ ist wissen wir, dass $\dim \text{Bild}(F(V)) < \dim V$ gilt und somit muss schon gelten, dass der Kern von $F$ nicht-trivial ist, d.h., $\text{Kern}\, F \neq \vec{0}$.

Sei nun $v_1 \in \text{Kern}(F), v \neq \vec{0}$. Wir ergänzen $v_1$ zu einer Basis $B' = (v_1, w_2, \ldots, w_n)$ von $V$. Mit Hilfe des Algorithmus 2.35 zur Trigonalisierung einer Matrix erhalten wir also:

$$M_{B'}(F) \;=\; \begin{pmatrix} 0 & a_{12} & \ldots & a_{1n} \\ \vdots & & & \\ \vdots & & B & \\ 0 & & & \end{pmatrix}$$

Da $W \coloneqq \text{lin}(\{w_2, \ldots, w_n\})$ im Allgemeinen nicht $F$-invariant ist, definieren wir die linearen Abbildungen

$$H(w_j) \;=\; a_{1j} v_1 \quad \text{und} \quad G(w_j) \;=\; a_{2j} w_2 + \ldots + a_{nj} w_n.$$

Dann können wir den Endomorphismus $F$ schreiben als: $F(w) = H(w) + G(w)$ für alle $w \in W$. Bezüglich der Basis $\tilde{B}' = (w_2, \ldots, w_n)$ gilt dann $B = M_{\tilde{B}'}(G)$. Außerdem gilt, dass $\text{Bild}(H) \subset \text{Kern}(F)$ und $G$ ist nilpotent, da auf Grund der Nilpotenz von $F$ für alle $w \in W$ gilt:

$$\begin{aligned} 0 \;=\; F^k(w) \;&=\; F^{k-1}(F(w)) \\ &=\; F^{k-1}(H(w) + G(w)) \;=\; F^{k-1}(\lambda v_1 + G(w)) \\ &=\; F^{k-1}(G(w)) \;=\; \ldots \;=\; F^{k-2}(G^2(w)) \;=\; \ldots \;=\; G^k(w). \end{aligned}$$

Da $\dim W = \dim V - 1$ gilt, können wir auf $G$ die Induktionsvoraussetzung anwenden, d.h., es gibt eine Basis $\tilde{B} = (v_2, \ldots, v_n)$ von $W$, so dass

$$M_{\tilde{B}}(G) \;=\; \begin{pmatrix} 0 & & * \\ & \ddots & \\ 0 & & 0 \end{pmatrix}$$

Damit folgt schon für die Basis $B = (v_1, \ldots, v_n)$ von $V$, dass

$$M_B(F) \;=\; \begin{pmatrix} 0 & & * \\ & \ddots & \\ 0 & & 0 \end{pmatrix}$$

und das charakteristische Polynom ist dementsprechend $P_F(t) \;=\; (-1)^n t^n$. □

Man kann sogar noch mehr zeigen als die Aussage der vorangegangenen Sätze und Lemmata, nämlich eine vollständige Charakterisierung von nilpotenten Endomorphismen, wie der folgende Satz zeigt.

**Satz 2.43**
*Sei $F\colon V \to V$ ein Endomorphismus von $V$. Dann sind folgende Aussagen äquivalent:*

*i) $F$ ist nilpotent.*

*ii) $F^k = 0$ für ein $k \in \mathbb{N}$.*

*iii) Das charakterstische Polynom $P_F$ von $F$ hat die Form $P_F(t) = (-1)^n t^n$.*

*iv) Es gibt eine Basis $B$ von $V$, so dass die darstellende Matrix $M_B(F)$ von $F$ die folgende Gestalt hat:*

$$M_B(F) \;=\; \begin{pmatrix} 0 & & * \\ & \ddots & \\ 0 & & 0 \end{pmatrix}$$

*Beweis.* Siehe [Fischer2005, Satz 4.5.7] □

**Bemerkung 2.44**
*Nilpotente Endomorphismen bzw. Matrizen besitzen nur den Eigenwert $\lambda = 0$, daher haben ihre darstellenden Matrizen keinen vollen Rang. Andersherum gibt es jedoch quadratische Matrizen, die nicht vollen Rang haben, jedoch nicht nilpotent sind, z.B. die Matrix*

$$A \;:=\; \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$$

*mit $A^k = A$ für alle $k \in \mathbb{K}$ und den Eigenwerten $\lambda_1 = 0$ und $\lambda_2 = 1$ von $A$.*

Eine wichtige Erkenntnis zur Konstruktion der Jordanschen Normalform ist, dass der Kern des Endomorphismus $G := (F - \lambda \operatorname{Id}_V)$ mit jeder Potenz von $G$ größer werden kann, wie folgendes Lemma zeigt.

**Lemma 2.45**
*Sei $F\colon V \to V$ ein Endomorphismus des endlich-dimensionalen $\mathbb{K}$-Vektorraums $V$ mit Eigenwert $\lambda \in \mathbb{K}$. Dann gilt für alle $k \in \mathbb{N}$:*

$$\operatorname{Kern}(F - \lambda \operatorname{Id}_V) \;\subset\; \operatorname{Kern}([F - \lambda \operatorname{Id}_V]^k).$$

*Beweis.* Sei $G \coloneqq (F - \lambda \operatorname{Id}_V)$. Wir müssen zeigen, dass für beliebiges $k \in \mathbb{N}$ gilt:

$$v \in \operatorname{Kern}(G) \;\Rightarrow\; v \in \operatorname{Kern}(G^k), \qquad \text{für alle } v \in V.$$

Sei also $v \in \operatorname{Kern}(G)$, dann gilt offensichtlich $Gv = 0$. Sei nun $k \in \mathbb{N}$ eine beliebige Potenz, dann betrachten wir

$$G^k v \;=\; G^{k-1} \underbrace{Gv}_{=0} \;=\; 0.$$

Daraus folgt also schon, dass $v \in \operatorname{Kern}(G^k)$ gilt. □

Nach Satz 2.26 wissen wir, dass sich der Vektorraum $V$ genau dann in eine direkte Summe von $F$-invarianten Eigenräumen $\operatorname{Eig}(F; \lambda_i), i = 1, \ldots, k$, zerlegen lässt, wenn die Dimension jedes Eigenraums der algebraischen Vielfachheit $r_i \in \mathbb{N}$ der Nullstellen des charakteristischen Polynoms entspricht, d.h.,

$$\dim \operatorname{Eig}(F; \lambda_i) \;=\; r_i, \quad \text{für } i = 1, \ldots, k.$$

Falls die Dimension eines Eigenraums $\operatorname{Eig}(F; \lambda_i)$ jedoch zu klein ist, so lässt sie sich durch Potenzieren mit $r_i$ passend vergrößern, denn nach Lemma 2.45 gilt:

$$\operatorname{Eig}(F; \lambda_i) \;=\; \operatorname{Kern}(F - \lambda_i \operatorname{Id}_V) \;\subset\; \operatorname{Kern}([F - \lambda_i \operatorname{Id}_V]^{r_i}). \tag{2.14}$$

Die Einbettung in (2.14) motiviert folgende Definition des Hauptraums.

**Definition 2.46** (Hauptraum und Hauptvektoren)
*Sei $F\colon V \to V$ ein Endomorphismus des $\mathbb{K}$-Vektorraums $V$. Sei außerdem $\lambda \in \mathbb{K}$ ein Eigenwert von $F$ der algebraischen Vielfachheit $r \geqslant 1$. Dann definieren wir den* Hauptraum *oder* verallgemeinerten Eigenraum $\operatorname{Haupt}(F; \lambda)$ *von $F$ zum Eigenwert $\lambda$ als Kern der $r$-fachen Anwendung von $(F - \lambda \operatorname{Id}_V)$, d.h.*

$$\operatorname{Haupt}(F; \lambda) \;\coloneqq\; \operatorname{Kern}([F - \lambda \operatorname{Id}_V]^r).$$

*Die Vektoren $v \in \operatorname{Haupt}(F; \lambda)$ werden* Hauptvektoren *der Stufe $d \geqslant 1$ genannt, wenn gilt*

$$[F - \lambda \operatorname{Id}_V]^d(v) \;=\; 0, \quad [F - \lambda \operatorname{Id}_V]^{d-1}(v) \;\neq\; 0.$$

*Damit ergibt sich, dass alle Eigenvektoren Hauptvektoren der Stufe $d = 1$ sind.*

Um zu verstehen, wie sich die Potenzierung der Endomorphismen auswirkt betrachten wir einen Eigenwert $\lambda \in \mathbb{K}$ des Endomorphismus $F\colon V \to V$ und Potenzen des Endomorphimus $G := F - \lambda \operatorname{Id}_V$. Wir stellen fest, dass wir folgende Inklusionsketten erhalten:

$$\{\vec{0}\} \subset \operatorname{Kern} G \subset \operatorname{Kern} G^2 \subset \ldots \subset \operatorname{Kern} G^l,$$
$$V \supset \operatorname{Bild} G \supset \operatorname{Bild} G^2 \supset \ldots \supset \operatorname{Bild} G^l.$$

Außerdem gilt nach dem Dimensionssatz [Fischer2005, Satz 2.2.4], dass $\dim \operatorname{Kern} G^l + \dim \operatorname{Bild} G^l = \dim V$ ist. Jedoch sind die Mengen im Allgemeinen nicht disjunkt wie bei einer direkten Summe, d.h., wir haben nicht

$$\operatorname{Kern} G^l \cap \operatorname{Bild} G^l \neq \{\vec{0}\}.$$

Da $V$ jedoch endlich-dimensional ist, können die beiden obigen Inklusionsketten nicht beliebig auf- bzw. absteigen.

Das folgende nützliche Lemma charakterisiert die Eigenschaften dieser Inklusionsketten noch genauer.

**Lemma 2.47** (Lemma von Fitting)
*Sei $G\colon V \to V$ ein Endomorphismus des $\mathbb{K}$-Vektorraums $V$. Sei außerdem $\lambda = 0$ ein Eigenwert von $G$ mit algebraischer Vielfachheit $r \in \mathbb{N}, r \geqslant 1$. Wir betrachten die kleinste Potenz $d \in \mathbb{N}$ für die der Kern von $G$ sich nicht mehr ändert, d.h.,*

$$d := \min\{l \in \mathbb{N} \mid \operatorname{Kern}(G^l) = \operatorname{Kern}(G^{l+1})\}, \tag{2.15}$$

*wobei $G^0 := \operatorname{Id}_V$ gilt. Dann gelten die folgenden Aussagen:*

1. $d = \min\{l \in \mathbb{N} \mid \operatorname{Bild}(G^l) = \operatorname{Bild}(G^{l+1})\}$,
2. $\operatorname{Kern}(G^{d+i}) = \operatorname{Kern}(G^d), \quad \operatorname{Bild}(G^{d+i}) = \operatorname{Bild}(G^d) \quad$ *für alle* $i \in \mathbb{N}$,
3. *Die Räume* $U := \operatorname{Kern}(G^d)$ *und* $W := \operatorname{Bild}(G^d)$ *sind $G$-invariant,*
4. $(G|_U)^d = 0$ *und* $G|_W\colon W \to W$ *ist ein Isomorphismus,*
5. $V = U \oplus W$.

*Beweis.* Wir nehmen an $d \in \mathbb{N}$ sei der kleinste Index mit der Eigenschaft aus (2.15). Dann können wir mit der Dimensionsformel [Fischer2005, Satz 2.2.4] folgern, dass gilt

$$\begin{aligned}\operatorname{Kern}(G^{d+1}) = \operatorname{Kern}(G^d) &\Leftrightarrow \dim(\operatorname{Kern}(G^{d+1})) = \dim(\operatorname{Kern}(G^d)) \\ &\Leftrightarrow \dim(\operatorname{Bild}(G^{d+1})) = \dim(\operatorname{Bild}(G^d)) \\ &\Leftrightarrow \operatorname{Bild}(G^{d+1}) = \operatorname{Bild}(G^d).\end{aligned}$$

Das bedeutet schon, dass die Abbildung $G|_W$ für $W := \text{Bild}(G^d)$ mit

$$G|_W \colon W \to \text{Bild}(G^{d+1}) = W$$

ein Isomorphismus ist. Aus dieser Beobachtung folgen schon die ersten drei Aussagen, sowie der zweite Teil der vierten Aussage. Die Nilpotenz der Abbildung $G|_U$ mit Nilpotenzindex $d$ ist auch klar, da für alle $v \in U = \text{Kern}(G^k)$ gilt, dass $G^d(v) = 0$ ist.

Sei nun $v \in U \cap W$, dann ist $G^d(v) = 0$ und es muss ein $w \in V$ geben, so dass $G^d(w) = v$ ist. Setzen wir die erste Beobachtung in die zweite Beobachtung ein erhalten wir, dass auch $G^{2d}(w) = 0$ sein muss und somit gilt nach der zweiten Aussage des Lemmas, dass

$$w \in \text{Kern}(G^{2d}) = \text{Kern}(G^d).$$

Damit folgt aber schon, dass

$$0 \;=\; G^d(w) \;=\; v,$$

und somit gilt $V = U \oplus W$. □

Durch die Betrachtung von Haupträumen anstatt Eigenräumen lässt sich eine mögliche Differenz zwischen algebraischen und geometrischen Vielfachheiten der Eigenwerte eines Endomorphismus ausgleichen. Wie wir im folgenden Satz sehen werden lässt sich der Vektorraum $V$ nun in eine innere direkte Summe der Haupträume zerlegen. Dies war bisher nur für diagonalisierbare Endomorphismen mit Hilfe der Eigenräume in Satz 2.26 möglich und bringt uns einen großen Schritt in Richtung der Jordanschen Normalform voran.

**Satz 2.48** (Hauptraumzerlegung)
*Sei $F \colon V \to V$ ein Endomorphismus des $\mathbb{K}$-Vektorraums $V$ und sei*

$$P_F(t) \;=\; \pm(t-\lambda_1)^{r_1} \cdot \ldots \cdot (t-\lambda_k)^{r_k}$$

*das charakteristische Polynom von $F$ mit paarweisen verschiedenen $\lambda_1, \ldots, \lambda_k \in \mathbb{K}$, die die Eigenwerte von $F$ darstellen und deren algebraischen Vielfachheiten $r_i \in \mathbb{N}, r_i \geqslant 1$ sind. Es sei außerdem $V_i := \text{Haupt}(F; \lambda_i) \subset V$ für jedes $\lambda_i$ der entsprechende Hauptraum. Dann gelten die folgenden Aussagen:*

1. $V \;=\; V_1 \oplus \ldots \oplus V_k$
2. $F(V_i) \subset V_i$ *und* $\dim V_i = r_i$ *für* $i = 1, \ldots, k$
3. $F$ *hat eine Zerlegung* $F = F_D + F_N$ *mit:*
   a) $F_D$ *ist diagonalisierbar*
   b) $F_N$ *ist nilpotent*
   c) $F_N$ *und* $F_D$ *kommutieren, d.h.,* $F_D \circ F_N = F_N \circ F_D$

*Beweis.* Wir führen den Beweis mittels vollständiger Induktion über die Zahl $k \geqslant 1$ der paarweise verschiedenen Eigenwerte von $F$.

**Induktionsanfang:** $k = 1$
Für $k = 1$ existiert nur ein Eigenwert $\lambda \in \mathbb{K}$ von $F$. Das bedeutet, dass das charakteristische Polynom $P_F$ von der Form ist

$$P_F(t) \;=\; \pm(t-\lambda)^n$$

und somit hat $\lambda$ die algebraische Vielfachheit $n = \dim(V)$. Damit gilt für $V_1 = \text{Haupt}(F;\lambda)$ schon

$$V_1 \;=\; \text{Kern}([F - \lambda\,\text{Id}_V]^n) \;=\; V,$$

da $F - \lambda\,\text{Id}_V$ nilpotent ist mit Nilpotenzindex $k \leqslant n$ und wir erhalten damit die triviale Zerlegung aus der ersten Behauptung.

Da $F$ Endomorphismus ist folgt trivialerweise, dass $F(V_1) \subset V_1 = V$ und $\dim(V_1) = \dim(V) = n$ gilt, was die zweite Behauptung zeigt.

Da das charakteristische Polynom $P_F$ von $F$ in Linearfaktoren zerfällt wissen wir mit Satz 2.33, dass eine Basis $B$ von $V$ existiert, so dass die darstellende Matrix $M_B(F)$ eine obere, rechte Dreiecksgestalt hat. Wir können die darstellende Matrix dann zerlegen in $M_B(F) = D + N$, wobei $D$ eine Diagonalmatrix der Form $D = \lambda E_n$ ist und $N$ eine strikte obere, rechte Dreiecksmatrix ist. Mit Satz 2.43 wissen wir, dass $N$ nilpotent sein muss. Eine einfache Rechnung zeigt, dass $D$ und $N$ kommutieren mit

$$D \cdot N \;=\; \lambda \cdot N \;=\; N \cdot D,$$

womit die dritte Behauptung gezeigt ist.

**Induktionsschritt:** $k - 1 \to k$
Die Induktionsannahme ist, dass die Aussage bereits für den Fall $k-1$ gezeigt wurde. Wir definieren uns also für den Eigenwert $\lambda_1 \in \mathbb{K}$ von $F$ mit algebraischer Vielfachheit $r_1 \in \mathbb{N}, r_1 \geqslant 1$ die Abbildung $G \;:=\; F - \lambda_1\,\text{Id}_V$. Seien $A$ eine darstellende Matrix von $G$ und $B$ eine darstellende Matrix von $F$ bezüglich einer beliebigen gemeinsamen Basis. Dann gilt offensichtlich $A = B - \lambda_1 I_n$. Damit sehen wir nun ein, dass gilt

$$\begin{aligned} P_G(t-\lambda_1) \;&=\; P_A(t-\lambda_1) \;=\; \det(A - (t-\lambda_1)I_n) \\ &=\; \det(B - \lambda_1 I_n - (t-\lambda_1)I_n) \;=\; \det(B - tI_n) \;=\; P_B(t) \;=\; P_F(t), \end{aligned}$$

womit schon folgt, dass 0 ein Eigenwert von $G$ mit algebraischer Vielfachheit $r_1$ ist, da

$$P_F(\lambda_1) \;=\; P_G(\lambda_1 - \lambda_1) \;=\; P_G(0).$$

Nach dem Lemma 2.47 von Fitting lässt sich $V$ als direkte Summe schreiben mit

$$V \;=\; \text{Haupt}(F;\lambda_1) \,\oplus\, \text{Bild}(G^{r_1}).$$

Für $v \in \text{Haupt}(F;\lambda_1)$ gilt, dass $[F-\lambda_1 \text{Id}_V]^{r_1}(v) = 0$. Außerdem sieht man durch die Kommutativität der Identität ein, dass

$$[F-\lambda_1 \text{Id}_V](F(v)) \;=\; (F^2-\lambda_1 F)(v) \;=\; F\circ(F-\lambda_1 \text{Id}_V)(v) \tag{2.16}$$

und somit durch sukzessive Anwendung von (2.16) auch

$$[F-\lambda_1 \text{Id}_V]^{r_1}(F(v)) \;=\; F\circ \underbrace{[F-\lambda_1 \text{Id}_V]^{r_1}(v)}_{=0} \;=\; 0.$$

Das zeigt, dass $F(v) \in \text{Haupt}(F;\lambda_1)$ für alle $v \in \text{Haupt}(F;\lambda_1)$, d.h, dass der Unterraum $\text{Haupt}(F;\lambda_1)$ $F$-invariant ist.

Für $v \in V$ gilt, dass $G^{r_1}(v) =: w \in \text{Bild}(G^{r_1})$ ist. Außerdem sehen wir ein, dass mit $F = (G+\lambda_1 \text{Id}_V)$ gilt:

$$F(w) \;=\; (G+\lambda_1 \text{Id}_V)(w) \;=\; G(w)+\lambda_1 w = \; G\circ G^{r_1}(v)+\lambda_1 w \;\in \text{Bild}(G^{r_1}),$$

da $\text{Bild}(G^{r_1+1}) \subset \text{Bild}(G^{r_1})$ ist. Das zeigt, dass $F(w) \in \text{Bild}([F-\lambda_1 \text{Id}_V]^{r_1})$ für alle $w \in \text{Bild}([F-\lambda_1 \text{Id}_V]^{r_1})$ ist, d.h., der Unterraum $\text{Bild}([F-\lambda_1 \text{Id}_V]^{r_1})$ ist auch $F$-invariant.

Betrachten wir die Einschränkung $F|_W$ von $F$ auf den Unterraum $W$ so stellen wir fest, dass das charakteristische Polynom $P_{F|_W}$ in Linearfaktoren zerfällt mit

$$P_{F|_W}(t) \;=\; \pm(t-\lambda_2)^{r_2}\cdot\ldots\cdot(t-\lambda_k)^{r_k}.$$

Da wir nun einen Endomorphismus betrachten, der $k-1$ verschiedene Eigenwerte besitzt und dessen charakteristischen Polynom in Linearfaktoren zerfällt, können wir die Induktionsvoraussetzung anwenden. Damit folgen direkt schon die ersten beiden Aussagen des Satzes.

Die Zerlegung aus der dritten Aussage des Satzes erhält man durch die folgenden darstellenden Matrizen in Blockdiagonalgestalt, die existieren, da der Endomorphismus $F$ trigonalisierbar ist nach Satz 2.33:

$$D \;:=\; \begin{pmatrix} \lambda_1 E_{r_1} & & \\ & \ddots & \\ & & \lambda_k E_{r_k} \end{pmatrix}, \quad N \;:=\; \begin{pmatrix} N_1 & & \\ & \ddots & \\ & & N_k \end{pmatrix}.$$

Man kann durch Nachrechnen leicht zeigen, dass gilt:

$$D\cdot N \;=\; \begin{pmatrix} \lambda_1 N_1 & & \\ & \ddots & \\ & & \lambda_k N_k \end{pmatrix} \;=\; N\cdot D.$$

Da $N$ und $D$ die darstellenden Matrizen der Endomorphismen $F_D$ und $F_N$ sind, folgt die Kommutativität jener. $\square$

Im Fall von Matrizen lässt sich die Aussage von Satz 2.48 wie folgt formulieren.

**Korollar 2.49**
*Sei $A \in \mathbb{K}^{n\times n}$ eine Matrix, für die das charakteristische Polynom $P_A$ in Linearfaktoren zerfällt, d.h.*

$$P_A(t) \;=\; \pm(t-\lambda_1)^{r_1}\cdot\ldots\cdot(t-\lambda_k)^{r_k}.$$

*Dann existiert eine invertierbare Matrix $S \in \mathrm{GL}(n;\mathbb{K})$, so dass*

$$SAS^{-1} \;=\; \begin{pmatrix} \boxed{\lambda_1 I_{r_1} + N_1} & & 0 \\ & \ddots & \\ 0 & & \boxed{\lambda_k I_{r_k} + N_k} \end{pmatrix} =:\; \tilde{A}.$$

*Jede Blockmatrix für $i = 1,\ldots,k$ hat hierbei die Gestalt einer rechten oberen Dreiecksmatrix, d.h.,*

$$\lambda_i I_{r_i} + N_i \;=\; \begin{pmatrix} \lambda_i & & * \\ & \ddots & \\ 0 & & \lambda_i \end{pmatrix} \in \mathbb{K}^{r_i\times r_i}, \quad i = 1,\ldots,k.$$

*Insbesondere lässt sich die Matrix $\tilde{A}$ zerlegen in $\tilde{A} = D + N$, wobei $D$ Diagonalmatrix und $N$ nilpotent ist. Schließlich gilt außerdem, dass $D$ und $N$ kommutieren, d.h.*

$$D\cdot N \;=\; N\cdot D.$$

**Bemerkung 2.50**
*Die in Satz 2.48 beschriebene Zerlegung $F = F_D + F_N$ ist die einzige Zerlegung in einen diagonalisierbaren und einen nilpotenten Endomorphismus, die kommutieren.*

Die Hauptraumzerlegung liefert uns zwar eine Blockdiagonalmatrix, die der Gestalt einer vollbesetzten oberen, rechten Dreiecksmatrix vorzuziehen ist, jedoch geben wir uns noch nicht zufrieden mit diesem Resultat. Bisher haben wir die nilpotenten Anteile des Endomorphismus als gegeben angesehen. Es stellt sich jedoch heraus, dass es möglich ist diese durch geschickte Basiswahl in die Normalform einer Jordanmatrix in Definition 2.39 zu überführen, wie der folgende Satz aussagt.

**Satz 2.51** (Normalisierung nilpotenter Endomorphismen)
*Es sei $G \in \mathrm{End}(V)$ nilpotent mit Nilpotenzindex $d \in \mathbb{N}$ über einem $\mathbb{K}$-Vektorraum $V$. Dann gilt*

$$\{0\} \subseteq \mathrm{Kern}\, G \subseteq \mathrm{Kern}\, G^2 \subseteq \cdots \subseteq \mathrm{Kern}\, G^d = V$$

*und es gibt Koeffizienten $s_i \in \mathbb{N}, 1 \leqslant i \leqslant d$, so dass eine Zahlpartition existiert mit*

$$s_1 \cdot 1 + s_2 \cdot 2 \cdots + s_d \cdot d = n := \dim V.$$

*Die Koeffizienten der Zahlpartition sind für den Endomorphismus $G$ eindeutig festgelegt durch die folgende Differenz:*

$$s_i \;=\; \Delta_i - \Delta_{i+1}, \quad 1 \leqslant i \leqslant d,$$

*wobei $\Delta_i := \dim \operatorname{Kern}(G_i) - \dim \operatorname{Kern}(G_{i-1})$ gerade die Anzahl der Hauptvektoren der Stufe $i$ sind.*

*Außerdem gibt es eine Basis $B$ von $V$, so dass die darstellende Matrix von $G$ bezüglich $B$ eine Blockdiagonalmatrix mit folgender Gestalt ist*

$$M_B(G) = \operatorname{diag}\left( \underbrace{J_d, \ldots, J_d}_{s_d-\text{mal}}, \underbrace{J_{d-1}, \ldots, J_{d-1}}_{s_{d-1}-\text{mal}}, \ldots \underbrace{J_1, \ldots, J_1}_{s_1-\text{mal}} \right) \tag{2.17}$$

*wobei, die Matrizen $J_k$, $1 \leqslant k \leqslant d$, $k$-dimensionale Jordanmatrizen aus Definition 2.39 sind mit*

$$J_k \;=\; \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \mathbf{0} \\ & & 0 & 1 & \\ & \mathbf{0} & & \ddots & \ddots \end{pmatrix} \in \mathbb{K}^{k \times k}.$$

*Beweis.* Siehe [Fischer2005, Theorem 4.6.5]. □

Wir verzichten an dieser Stelle bewusst auf einen konstruktiven Beweis dieses wichtigen Satzes, da wir für ein vollständiges Verständnis viel mehr Theorie benötigen, die jedoch nicht Bestandteil dieser Vorlesung sein kann. Diese unbefriedigende Lücke in der Normalformentheorie werden wir stattdessen mit der Diskussion eines Algorithmus zur Überführung der nilpotenten Anteile des Endomorphismus in die Normalform aus Satz 2.51 füllen.

**Algorithmus 2.52** (Normalisierung einer nilpotenten Matrix)
*Sei $B$ eine kanonische Basis des $\mathbb{K}$-Vektorraums $V$ und $A := M_B(G)$ darstellende Matrix eines nilpotenten Endomorphismus $G\colon V \to V$ mit Nilpotenzindex $d \in \mathbb{N}$.*

*Um eine Transformationsmatrix $S \in \mathrm{GL}(\mathbb{K}; n)$ zu erhalten, so dass gilt*

$$N \;=\; SAS^{-1},$$

*wobei $N$ eine Jordanmatrix ist, müssen wir geschickt Basisvektoren aus den verschiedenen Kernen der Potenzen von $G$ wählen.*

**Vorbereitung**

1. *Berechne Potenzen von $A$ als $A^i$ für $1 \leqslant i \leqslant d$*
2. *Bestimme Basen $K_i$ der jeweiligen Kerne $\operatorname{Kern}(A^i)$ für $1 \leqslant i \leqslant d$*
3. *Berechne die Differenzen der aufeinanderfolgenden Kerndimensionen:*
   $\Delta_1 = \dim \operatorname{Kern}(A) - \dim \operatorname{Kern}(E_n)$, $\ldots$, $\Delta_d = \dim \operatorname{Kern}(A^d) - \dim \operatorname{Kern}(A^{d-1})$

**0. Schritt: Hauptvektoren der Stufe $d$**

1. *Wähle $s_d := \Delta_d - \Delta_{d+1} = \Delta_d$ Hauptvektoren $v_1^{(d)}, \ldots, v_{s_d}^{(d)}$ der Stufe $d$ aus $K_d$*
2. *Notiere das Schema für den Aufbau von Jordanketten wie folgt*

| | | |
|---|---|---|
| $v_1^{(d)}$ | $\ldots$ | $v_{s_d}^{(d)}$ |

**1. Schritt: Hauptvektoren der Stufe $d-1$**

1. *Multipliziere alle Vektoren des vorigen Schritts (die unterste Zeile im Schema) mit $A$ und trage die resultierenden Vektoren $Av_1^{(d)}, \ldots, Av_{s_d}^{(d)}$ in eine neue Zeile **unter** das Schema ein.*
2. *Ergänze um $s_{d-1} = \Delta_{d-1} - \Delta_d$ Hauptvektoren $v_1^{(d-1)}, \ldots, v_{s_{d-1}}^{(d-1)}$ der Stufe $d-1$ aus $K_{d-1}$ und trage sie **rechts** neben die unterste Zeile des Schemas ein.*
3. *Das resultierende Schema für den Aufbau von Jordanketten sollte die folgende Gestalt haben:*

| | | | | | |
|---|---|---|---|---|---|
| $v_1^{(d)}$ | $\ldots$ | $v_{s_d}^{(d)}$ | | | |
| $Av_1^{(d)}$ | $\ldots$ | $Av_{s_d}^{(d)}$ | $v_1^{(d-1)}$ | $\ldots$ | $v_{s_{d-1}}^{(d-1)}$ |

**$i$. Schritt: Hauptvektoren der Stufe $d-i$**

1. *Multipliziere alle Vektoren des vorigen Schritts (die unterste Zeile im Schema) mit $A$ und trage die resultierenden Vektoren in eine neue Zeile **unter** das Schema ein.*
2. *Ergänze um $s_{d-i} = \Delta_{d-i} - \Delta_{d-i+1}$ Hauptvektoren $v_1^{(d-i)}, \ldots, v_{s_{d-i}}^{(d-i)}$ der Stufe $d-i$ aus $K_{d-i}$ und trage sie **rechts** neben die unterste Zeile des Schemas ein.*
3. *Das resultierende Schema für den Aufbau von Jordanketten sollte die folgende Gestalt haben:*

| | | | | | | |
|---|---|---|---|---|---|---|
| $\vdots$ | $\vdots$ | $\vdots$ | | | | |
| $A^{i-1}v_1^{(d)}$ | $\ldots$ | $A^{i-1}v_{s_d}^{(d)}$ | $\ldots$ | | | |
| $A^iv_1^{(d)}$ | $\ldots$ | $A^iv_{s_d}^{(d)}$ | $\ldots$ | $v_1^{(d-i)}$ | $\ldots$ | $v_{s_{d-i}}^{(d-i)}$ |

$d-1$**. Schritt: Hauptvektoren der Stufe** 1

1. *Multipliziere alle Vektoren des vorigen Schritts (die unterste Zeile im Schema) mit $A$ und trage die resultierenden Vektoren in eine neue Zeile* ***unter*** *das Schema ein.*

2. *Ergänze um $s_1 = \Delta_1 - \Delta_2$ Hauptvektoren $v_1^{(1)}, \ldots, v_{s_1}^{(1)}$ der Stufe 1 aus $K_1 = \mathrm{Eig}(G;\lambda)$, also* ***Eigenvektoren****, und trage sie* ***rechts*** *neben die unterste Zeile des Schemas ein.*

3. *Das resultierende Schema für den Aufbau von Jordanketten sollte die folgende Gestalt haben:*

| | | | | | | |
|---|---|---|---|---|---|---|
| $\vdots$ | $\vdots$ | $\vdots$ | | | | |
| $A^{d-2}v_1^{(d)}$ | $\ldots$ | $A^{d-2}v_{s_d}^{(d)}$ | $\ldots$ | | | |
| $A^{d-1}v_1^{(d)}$ | $\ldots$ | $A^{d-1}v_{s_d}^{(d)}$ | $\ldots$ | $v_1^{(1)}$ | $\ldots$ | $v_{s_1}^{(1)}$ |
| $\uparrow$ *Jordankette* | | $\uparrow$ *Jordankette* | | | | |

.

**Spaltenweises Eintragen des Schemas in $S^{-1}$:**

*Lesen wir die schließlich das fertige Schema zuerst von* ***unten nach oben*** *und dann von* ***links nach rechts*** *(entlang der Jordanketten) zellenweise ab und notieren die so gefundenen Vektoren* ***spaltenweise*** *von links nach rechts in die Transformationsmatrix $S^{-1}$, so liegt $N = SAS^{-1}$ in der Normalform nilpotenter Endomorphismen in Definition 2.39 vor.*

Durch geschickte Kombination der Hauptraumzerlegung aus Satz 2.48 und der Normalform niolpotenter Endormorphismen in Satz 2.51 lässt sich eine kanonische Normalform für Endomorphismen bestimmen, die schöne Eigenschaften hat. Diese *Jordansche Normalform* wird im folgenden Satz näher beschrieben.

**Satz 2.53** (Jordansche Normalform)
*Sei $A \in \mathbb{K}^{n\times n}$ eine Matrix, für die das charakteristische Polynom $P_A$ in Linearfaktoren zerfällt, d.h.*

$$P_A(t) \;=\; \pm(t-\lambda_1)^{r_1} \cdot \ldots \cdot (t-\lambda_k)^{r_k}.$$

*Dann existiert eine invertierbare Matrix $S \in \mathrm{GL}(n;\mathbb{K})$, so dass*

$$SAS^{-1} \;=\; \begin{pmatrix} \boxed{\lambda_1 I_{r_1} + N_1} & & 0 \\ & \ddots & \\ 0 & & \boxed{\lambda_k I_{r_k} + N_k} \end{pmatrix} \;=:\; J. \tag{2.18}$$

*Die nilpotenten Anteile $N_i$ für $i = 1, \ldots, k$ liegen hierbei (blockweise) in der Normalform aus Definition 2.39 vor. Die Blockmatrizen $\lambda_i I_{r_i} + N_i \in \mathbb{K}^{r_i \times r_i}$ in $J$ werden* Jordanblöcke *genannt und sind von der Gestalt*

$$\lambda_i I_{r_i} + N_i \;=\; \begin{pmatrix} \lambda_i & 1 & & & & & & & & & \\ & \ddots & \ddots & & & & & & & & \\ & & \ddots & 1 & & & & & & & \\ & & & \lambda_i & 0 & & & & & & \\ & & & & \lambda_i & 1 & & & & & \\ & & & & & \ddots & \ddots & & & & \\ & & & & & & \ddots & 1 & & & \\ & & & & & & & \lambda_i & 0 & & \\ & & & & & & & & \ddots & \ddots & \\ & & & & & & & & & \ddots & 0 \\ & & & & & & & & & & \lambda_i \end{pmatrix} \tag{2.19}$$

*Die* ***Anzahl der Kästchen*** *in einem Jordanblock der Form* (2.19) *ist gegeben durch die* ***geometrische Vielfachheit*** $\dim \mathrm{Eig}(F - \lambda_i I_n)$ *des Eigenwerts $\lambda_i \in \mathbb{K}$ von $F$. Insbesondere lässt sich die Jordansche Normalform $J$ von $A$ zerlegen in $J = D + N$, wobei $D$ Diagonalmatrix und $N$ nilpotent ist. Schließlich gilt außerdem, dass $D$ und $N$ kommutieren, d.h.*

$$D \cdot N \;=\; N \cdot D.$$

*Beweis.* Der Beweis der Jordanschen Normalform besteht im Prinzip nur aus Anwendung der Hauptraumzerlegung aus Satz 2.48 und dem Satz 2.51 über die Normalform für nilpotente Endomorphismen. Wir bezeichnen für $i = 1, \ldots, k$ die Haupträume von $F$ bezüglich der Eigenwerte $\lambda_i$ mit

$$V_i \;:=\; \mathrm{Haupt}(F;\lambda_i),$$

und wir betrachten die nilpotenten Endomorphismen

$$G_i \;:=\; (F - \lambda_i \operatorname{Id}_V)|_{V_i}.$$

Durch Anwendung des Satzes 2.51 können wir Basen $B_i$ der Haupträume $V_i$ finden, so dass die darstellenden Matrizen $M_{B_i}(G_i)$ der nilpotenten Endomorphismen in Normalform vorliegen. Diese Basen kann man dann wegen der Hauptraumzerlegung in Satz 2.48 zu einer Basis $B$ von $V$ zusammenführen, so dass die darstellende Matrix $M_B(F)$ in Jordanscher Normalform vorliegt. □

**Algorithmus 2.54** (Berechnung der Jordannormalform)
*Sei $B$ eine kanonische Basis des $\mathbb{K}$-Vektorraums $V$ und $A := M_B(F)$ darstellende Matrix eines Endomorphismus $F\colon V \to V$, dessen charakteristisches Polynom in Linearfaktoren zerfällt von der Gestalt ist*

$$P_F(t) \;=\; \pm(t-\lambda_1)^{r_1} \cdot \ldots \cdot (t-\lambda_k)^{r}_{k}$$

*für $k \in \mathbb{N}$ paarweise verschiedene Eigenwerte von $F$.*

*Das Ziel ist es eine Transformationsmatrix $S \in \mathrm{GL}(\mathbb{K}; n)$ zu konstruieren, so dass gilt*

$$J \;=\; SAS^{-1},$$

*wobei $J$ die Jordannormalform aus (2.18) ist.*

*Hierfür müssen wir nur die nilpotenten Einschränkungen von $F$ auf die Haupträume $V_i := \mathrm{Haupt}(F; \lambda_i)$ mit*

$$G_i \;:=\; (F - \lambda_i I_{r_i})|_{V_i}, \quad 1 \leqslant i \leqslant k$$

*betrachten und die nötigen Basen $B_i$ von $\mathrm{Haupt}(F; \lambda_i)$ mit dem Algorithmus 2.52 zur Normalisierung von nilpotenten Endomorphismen berechnen. Die Konkatenation der Basen $B_i, 1 \leqslant i \leqslant k$ ergibt wegen dem Satz zur Hauptraumzerlegung 2.48 eine Basis des Vektorraums $V$. Werden die Basisvektoren spaltenweise in die Transformationsmatrix $S^{-1}$ eingetragen, so erhält man unter dieser Ähnlichkeitstransformation die gewünschte Jordansche Normalform $J$ von $A$.*

*Diese Jordansche Normalform $J$ ist eindeutig bis auf Permutation der Jordanblöcke.*

Wir wollen ein abschließendes Beispiel zur Jordanschen Normalform für eine $(5 \times 5)$-Matrix durchrechnen.

**Beispiel 2.55** (Berechnung der Jordanschen Normalform mit Transformationsmatrix)
*Wir betrachten die Matrix*

$$A \;:=\; \begin{pmatrix} 5 & 0 & 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & -\frac{1}{2} & 0 \\ -1 & 0 & 3 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{3}{2} & 0 \\ 0 & 0 & 0 & 0 & 4 \end{pmatrix}$$

*Um die Matrix $A$ in eine Jordansche Normalform zu überführen verwenden wir Algorithmus 2.52 und 2.54.*

*Wir berechnen zuerst alle Eigenwerte von $A$ mit Hilfe des charakteristischen Polynoms:*

$$\begin{aligned}
P_A(\lambda) &= \det\begin{pmatrix} 5-\lambda & 0 & 1 & 0 & 0 \\ 0 & \frac{1}{2}-\lambda & 0 & -\frac{1}{2} & 0 \\ -1 & 0 & 3-\lambda & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{3}{2}-\lambda & 0 \\ 0 & 0 & 0 & 0 & 4-\lambda \end{pmatrix} \\
&= (-1)^{5+5}(4-\lambda)\det\begin{pmatrix} 5-\lambda & 0 & 1 & 0 \\ 0 & \frac{1}{2}-\lambda & 0 & -\frac{1}{2} \\ -1 & 0 & 3-\lambda & 0 \\ 0 & \frac{1}{2} & 0 & \frac{3}{2}-\lambda \end{pmatrix} \\
&= (4-\lambda)\left[(-1)^{4+2}\frac{1}{2}\det\begin{pmatrix} 5-\lambda & 1 & 0 \\ 0 & 0 & -\frac{1}{2} \\ -1 & 3-\lambda & 0 \end{pmatrix}\right. \\
&\qquad \left. + (-1)^{4+4}(\frac{3}{2}-\lambda)\det\begin{pmatrix} 5-\lambda & 0 & 1 \\ 0 & \frac{1}{2}-\lambda & 0 \\ -1 & 0 & 3-\lambda \end{pmatrix}\right] \\
&= (4-\lambda)\left[\frac{1}{2}(-1)^{2+3}(-\frac{1}{2})\det\begin{pmatrix} 5-\lambda & 1 \\ -1 & 3-\lambda \end{pmatrix}\right. \\
&\qquad \left. +(\frac{3}{2}-\lambda)(-1)^{2+2}(\frac{1}{2}-\lambda)\det\begin{pmatrix} 5-\lambda & 1 \\ -1 & 3-\lambda \end{pmatrix}\right] \\
&= (4-\lambda)\left[\frac{1}{4}\Big((5-\lambda)(3-\lambda)+1\Big) + (\frac{3}{2}-\lambda)(\frac{1}{2}-\lambda)\Big((5-\lambda)(3-\lambda)+1\Big)\right] \\
&= (4-\lambda)\Big((5-\lambda)(3-\lambda)+1\Big)\left(\frac{1}{4}+(\frac{3}{2}-\lambda)(\frac{1}{2}-\lambda)\right) \\
&= (4-\lambda)(16-8\lambda+\lambda^2)(1-2\lambda+\lambda^2) \\
&= (4-\lambda)^3(1-\lambda)^2.
\end{aligned}$$

*Es liegen somit die Eigenwerte $\lambda_1 = 1$ und $\lambda_2 = 4$ von $A$ mit den jeweiligen algebraischen Vielfachheiten $r_1 = 2$ und $r_2 = 3$ vor.*

*Für den **ersten Jordanblock** zum Eigenwert $\lambda_1 = 1$ von $A$ betrachten wir zunächst den Endomorphismus*

$$G_1 := A - 1 \cdot I_5 = \begin{pmatrix} 4 & 0 & 1 & 0 & 0 \\ 0 & -\frac{1}{2} & 0 & -\frac{1}{2} & 0 \\ -1 & 0 & 2 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 3 \end{pmatrix}$$

*und wenden Algorithmus 2.52 zur Bestimmung einer Normalform an.*

***Vorbereitung***

*Wir bestimmen eine Basis* $K_1$ *des Eigenraums* $\mathrm{Kern}(G_1)$ *mittels Gaußschen Eliminiationsverfahren:*

$$\begin{pmatrix} 4 & 0 & 1 & 0 & 0 \\ 0 & -\frac{1}{2} & 0 & -\frac{1}{2} & 0 \\ -1 & 0 & 2 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 3 \end{pmatrix} \overset{I+4\cdot III}{\mapsto} \begin{pmatrix} 0 & 0 & 9 & 0 & 0 \\ 0 & -\frac{1}{2} & 0 & -\frac{1}{2} & 0 \\ -1 & 0 & 2 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 3 \end{pmatrix} \overset{I\leftrightarrow III}{\mapsto}$$

$$\begin{pmatrix} -1 & 0 & 2 & 0 & 0 \\ 0 & -\frac{1}{2} & 0 & -\frac{1}{2} & 0 \\ 0 & 0 & 9 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 3 \end{pmatrix} \overset{IV+II}{\mapsto} \begin{pmatrix} -1 & 0 & 2 & 0 & 0 \\ 0 & -\frac{1}{2} & 0 & -\frac{1}{2} & 0 \\ 0 & 0 & 9 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{pmatrix} \overset{IV\leftrightarrow V}{\mapsto}$$

$$\begin{pmatrix} -1 & 0 & 2 & 0 & 0 \\ 0 & -\frac{1}{2} & 0 & -\frac{1}{2} & 0 \\ 0 & 0 & 9 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

*Wir erhalten also als mögliche Basis*

$$K_1 \;:=\; \left\{ \begin{pmatrix} 0 \\ 1 \\ 0 \\ -1 \\ 0 \end{pmatrix} \right\}.$$

*Da* $\dim K_1 = \dim \mathrm{Eig}(A;1) = 1$ *gilt, wissen wir nach Satz 2.53, dass es nur ein Jordankästchen innerhalb des Jordanblocks zum Eigenwert* $\lambda_1 = 1$ *von* $A$ *geben kann. Die Größe dieses Jordankästchens entspricht in diesem Fall der algebraischen Vielfachheit* $r_1 = 2$ *von* $\lambda_1$*.*

*Wir bestimmen nun eine Basis* $K_2$ *des Eigenraums* $\mathrm{Kern}(G_1^2)$ *mittels Gaußschen Eliminationsverfahren mit*

$$G_1^2 \;=\; \begin{pmatrix} 15 & 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -6 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 9 \end{pmatrix}$$

*und erhalten somit*

$$\begin{pmatrix} 15 & 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -6 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 9 \end{pmatrix} \overset{II\leftrightarrow V \& II\leftrightarrow III}{\mapsto} \begin{pmatrix} 15 & 0 & 6 & 0 & 0 \\ -6 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 9 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \overset{2\cdot I \& 5\cdot II}{\mapsto}$$

$$\begin{pmatrix} 30 & 0 & 12 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -30 & 0 & 15 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 9 \end{pmatrix} \overset{II+I}{\mapsto} \begin{pmatrix} 30 & 0 & 12 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 27 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 9 \end{pmatrix}.$$

*Wir erhalten somit als mögliche Basis von* $\mathrm{Kern}(G_1^2)$

$$K_2 \;:=\; \left\{ \begin{pmatrix} 0 \\ 1 \\ 0 \\ -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \\ 0 \\ -1 \\ 0 \end{pmatrix} \right\}.$$

*Wir haben den Nilpotenzindex von* $d = 2$ *von* $G_1|_{\mathrm{Haupt}(A;\lambda_1)}$ *erreicht. Das bedeutet, dass der Kern von* $G_1$ *sich nicht mehr ändert für jede weitere Potenz von* $G_1$*, da gilt*

$$2 \;=\; \dim \mathrm{Kern}(G_1^2) \;=\; \dim \mathrm{Haupt}(A;1) \;=\; r_1.$$

*Entsprechend brauchen wir keine weiteren Potenzen von* $G_1$ *mehr zu betrachten.*

*Wir berechnen abschließend zur Vorbereitung*

$$\Delta_2 \;:=\; \dim K_2 - \dim K_1 \;=\; 2 - 1 \;=\; 1, \quad \Delta_1 \;:=\; \dim K_1 \;=\; 1.$$

**1. Schritt: Hauptvektoren der Stufe** 2

*Wir wählen aus dem Kern* $K_2$ *einen (*$\Delta_2 = 1$*) Hauptvektor der Stufe* 2*, d.h., einen Vektor der linear unabhängig zu Vektoren aus* $K_1$ *ist, also beispielsweise* $(0,-1,0,-1,0)^T$*. Wir notieren diesen Vektor in ein Schema wie folgt:*

$$\boxed{\begin{pmatrix} 0 \\ -1 \\ 0 \\ -1 \\ 0 \end{pmatrix}}$$

**2. Schritt: Hauptvektoren der Stufe** 1

*Wir berechnen zunächst* $G_1 \cdot (0,-1,0,-1,0)^T = (0,1,0,-1,0)^T$ *und tragen diesen Vektor in einer neuen Zeile unten in das Schema ein. Wir berechnen*

$$s_1 \;=\; \Delta_1 - \Delta_2 \;=\; 1 - 1 \;=\; 0,$$

*also brauchen wir keine weiteren Vektoren hinzufügen. Dies ist konsistent zu der Beobachtung, dass das Schema bereits* $r_1 = 2$ *Vektoren enthält.*

*Das finale Schema für den* ***ersten Jordanblock*** *sieht entsprechend so aus:*

$$\begin{array}{|c|}\hline \begin{pmatrix} 0 \\ -1 \\ 0 \\ -1 \\ 0 \end{pmatrix} \\ \hline \begin{pmatrix} 0 \\ 1 \\ 0 \\ -1 \\ 0 \end{pmatrix} \\ \hline \end{array}$$

*Die Basis* $B_1$ *für den Hauptraum* $\mathrm{Haupt}(A;1)$ *von* $A$ *zum Eigenwert* $\lambda_1 = 1$ *ergibt sich entsprechend durch Ablesen des Schemas* „von unten nach oben, von links nach rechts“*:*

$$B_1 \;:=\; \begin{pmatrix} 0 & 0 \\ 1 & -1 \\ 0 & 0 \\ -1 & -1 \\ 0 & 0 \end{pmatrix}$$

*Für den* ***zweiten Jordanblock*** *zum Eigenwert* $\lambda_1 = 4$ *von* $A$ *betrachten wir zunächst den Endomorphismus*

$$G_2 \;:=\; A - 4 \cdot I_5 \;=\; \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & -\frac{7}{2} & 0 & -\frac{1}{2} & 0 \\ -1 & 0 & -1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & -\frac{5}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

*und wenden Algorithmus 2.52 zur Bestimmung einer Normalform an.*

***Vorbereitung***

*Wir bestimmen eine Basis* $K_1$ *des Eigenraums* $\operatorname{Kern}(G_2)$ *mittels Gaußschen Eliminiationsverfahren:*

$$\begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & -\frac{7}{2} & 0 & -\frac{1}{2} & 0 \\ -1 & 0 & -1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & -\frac{5}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \overset{III+I}{\mapsto} \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & -\frac{7}{2} & 0 & -\frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & -\frac{5}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \overset{III \leftrightarrow IV}{\mapsto}$$

$$\begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & -\frac{7}{2} & 0 & -\frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & -\frac{5}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \overset{II \leftrightarrow III}{\mapsto} \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & -\frac{5}{2} & 0 \\ 0 & -\frac{7}{2} & 0 & -\frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \overset{III+7\cdot II}{\mapsto}$$

$$\begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & -\frac{5}{2} & 0 \\ 0 & 0 & 0 & -18 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

*Wir erhalten also als mögliche Basis*

$$K_1 \;:=\; \left\{ \begin{pmatrix} 1 \\ 0 \\ -1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \right\}.$$

*Da* $\dim K_1 = \dim \operatorname{Eig}(A; 4) = 2$ *gilt, wissen wir nach Satz 2.53, dass es zwei Jordankästchen innerhalb des Jordanblocks zum Eigenwert* $\lambda_2 = 4$ *von* $A$ *gibt. Die Summe der Größen dieser Jordankästchen entspricht in diesem Fall der algebraischen Vielfachheit* $r_2 = 3$ *von* $\lambda_2$*.*

*Wir bestimmen nun eine Basis* $K_2$ *des Eigenraums* $\operatorname{Kern}(G_2^2)$ *mittels Gaußschen Eliminiationsverfahren mit*

$$G_2^2 \;=\; \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 12 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & -3 & 0 & 6 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

*und erhalten somit*

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 12 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & -3 & 0 & 6 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \overset{I \leftrightarrow IV}{\mapsto} \begin{pmatrix} 0 & -3 & 0 & 6 & 0 \\ 0 & 12 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \overset{II+4 \cdot I}{\mapsto} \begin{pmatrix} 0 & -3 & 0 & 6 & 0 \\ 0 & 0 & 0 & 27 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

*Wir erhalten somit als mögliche Basis von* $\operatorname{Kern}(G_1^2)$

$$K_2 \;:=\; \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \right\}.$$

*Wir haben den Nilpotenzindex von* $d = 3$ *von* $G_2|_{\operatorname{Haupt}(A;\lambda_2)}$ *erreicht. Das bedeutet, dass der Kern von* $G_2$ *sich nicht mehr ändert für jede weitere Potenz von* $G_2$*, da gilt*

$$3 \;=\; \dim \operatorname{Kern}(G_2^2) \;=\; \dim \operatorname{Haupt}(A;4) \;=\; r_2.$$

*Entsprechend brauchen wir keine weiteren Potenzen von* $G_2$ *zu betrachten.*

*Wir berechnen abschließend zur Vorbereitung*

$$\Delta_2 := \dim K_2 - \dim K_1 \;=\; 3 - 2 \;=\; 1, \quad \Delta_1 := \dim K_1 \;=\; 2.$$

**1. Schritt: Hauptvektoren der Stufe** 2

*Wir wählen aus dem Kern* $K_2$ *einen (*$\Delta_2 = 1$*) Hauptvektor der Stufe* 2*, d.h., einen Vektor der linear unabhängig zu Vektoren aus* $K_1$ *ist, also beispielsweise* $(1,0,0,0,0)^T$*. Wir notieren diesen Vektor in ein Schema wie folgt:*

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

**2. Schritt: Hauptvektoren der Stufe** 1

*Wir berechnen zunächst* $G_2 \cdot (1,0,0,0,0)^T = (1,0,-1,0,0)^T$ *und tragen diesen Vektor in einer neuen Zeile unten in das Schema ein. Wir berechnen*

$$s_1 \;=\; \Delta_1 - \Delta_2 \;=\; 2 - 1 \;=\; 1.$$

*Dies bedeutet, dass wir noch einen weiteren Hauptvektor der Stufe* 1 *aus* $K_1$ *zu unserem Schema hinzufügen müssen. Hierzu wählen wir den Vektor* $(0,0,0,0,1)^T$*. Dies ist konsistent zu der Beobachtung, dass das Schema nun* $r_2 = 3$ *Vektoren enthält.*

*Das finale Schema für den* ***zweiten Jordanblock*** *sieht entsprechend so aus:*

| | |
|---|---|
| $\begin{pmatrix}1\\0\\0\\0\\0\end{pmatrix}$ | |
| $\begin{pmatrix}1\\0\\-1\\0\\0\end{pmatrix}$ | $\begin{pmatrix}0\\0\\0\\0\\1\end{pmatrix}$ |

*Die Basis* $B_2$ *für den Hauptraum* $\mathrm{Haupt}(A;4)$ *von* $A$ *zum Eigenwert* $\lambda_2 = 4$ *ergibt sich entsprechend durch Ablesen des Schemas* „von unten nach oben, von links nach rechts“*:*

$$B_2 \;:=\; \begin{pmatrix}1&1&0\\0&0&0\\-1&0&0\\0&0&0\\0&0&1\end{pmatrix}.$$

*Wir fügen abschließend die beiden Basen* $B_1$ *und* $B_2$ *der zwei Haupträume zu einer Basis* $B$ *von* $V$ *zusammen und schreiben die Basisvektoren von* $B$ *als Spalten der Transformationsmatrix*

$$S^{-1} \;:=\; \begin{pmatrix}0&0&1&1&0\\1&-1&0&0&0\\0&0&-1&0&0\\-1&-1&0&0&0\\0&0&0&0&1\end{pmatrix}, \quad S \;:=\; \begin{pmatrix}0&\frac{1}{2}&0&-\frac{1}{2}&0\\0&-\frac{1}{2}&0&-\frac{1}{2}&0\\0&0&-1&0&0\\1&0&1&0&0\\0&0&0&0&1\end{pmatrix}.$$

*Entsprechend erhalten wir*

$$S \cdot A \cdot S^{-1} \;=\; \left(\begin{array}{ccccc} \boxed{1} & \boxed{1} & 0 & 0 & 0 \\ \boxed{0} & \boxed{1} & 0 & 0 & 0 \\ 0 & 0 & \boxed{4} & \boxed{1} & 0 \\ 0 & 0 & \boxed{0} & \boxed{4} & 0 \\ 0 & 0 & 0 & 0 & \boxed{4} \end{array}\right)$$

*Wir können uns während der Bestimmung einer Jordanschen Normalform auch mittels der Jordanketten die passende Jordannormalform schon überlegen. Zu* $\mathrm{Haupt}(A;1)$ *gehört nur ein Jordankästchen der Dimension* $2 \times 2$ *(eine Jordankette der Länge* $2$*). Zu* $\mathrm{Haupt}(A;4)$ *gehört ein Jordankästchen der Dimension* $2 \times 2$ *(eine Jordankette der Länge* $2$*) und eine Jordankästchen der Dimension* $1 \times 1$ *(eine Jordankette der Länge* $1$*).*

Glücklicherweise existieren Algorithmen, die die obigen Berechnungen automatisiert in einem Computer durchführen und dabei mögliche Rechenfehler vermeiden und uns somit viel Zeit sparen.

# Kapitel 3

# Euklidische und unitäre Vektorräume

Bisher haben wir immer einen endlich-dimensionalen $\mathbb{K}$-Vektorraum $V$ mit einem beliebigen Körper $\mathbb{K}$ betrachtet. Wenn wir uns jedoch konkret auf die beiden Körper $\mathbb{K} = \mathbb{R}$ und $\mathbb{K} = \mathbb{C}$ festlegen, erhalten wir zusätzliche Struktur durch das Vorhandenseins eines *Skalarprodukts*, dass ein essentielles Hilfsmittel zur Messung von Längen und Winkeln darstellt. Stattet man einen reellen Vektorraum mit einem Skalarprodukt aus, so nennt man diesen *euklidisch*, wohingegen ein komplexer Vektorraum mit einem Skalarprodukt *unitär* genannt wird.

Bevor wir die allgemeinen Begriffe von Bilinearformen und Sesquilinearformen einführen, wollen wir zur Motivation die kanonischen Beispiele für euklidische und unitäre Vektorräume diskutieren.

## 3.1 Das kanonische Skalarprodukt in $\mathbb{R}^n$

Sei im Folgenden der zu Grunde liegende Vektorraum gewählt als $V = \mathbb{R}^n$. Wir definieren zuerst das kanonische Skalarprodukt in $\mathbb{R}^n$.

**Definition 3.1** (Skalarprodukt in $\mathbb{R}^n$)
*Seien $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ und $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ zwei Vektoren. Wir bezeichnen die Abbildung*

$$\begin{aligned} \langle \cdot, \cdot \rangle : \ \mathbb{R}^n \times \mathbb{R}^n \ &\to \ \mathbb{R}, \\ (x, y) \ &\to \ \langle x, y \rangle \ := \ x_1 y_1 + \ldots + x_n y_n \ = \ \sum_{i=1}^{n} x_i y_i \end{aligned}$$

*als* kanonisches Skalarprodukt *von $x$ und $y$ in $\mathbb{R}^n$.*

Schreibt man $x, y \in \mathbb{R}^n$ als Spaltenvektoren, so lässt sich das kanonische Skalarprodukt auch schreiben als

$$\langle x, y \rangle \;=\; x^T \cdot y \;=\; (x_1, \ldots, x_n) \cdot \begin{pmatrix} y_1 \\ \vdots \\ v_n \end{pmatrix}.$$

Eine ebenfalls in der Literatur gängige Schreibweise ist ein Skalarprodukt mit runden Klammern, d.h., $(x, y) \coloneqq \langle x, y \rangle$.

Folgende Eigenschaften des Skalarprodukts lassen sich leicht nachrechnen.

**Lemma 3.2** (Eigenschaften des Skalarprodukts)
*Das kanonische Skalarprodukt von $\mathbb{R}^n$ in Definition 3.1 besitzt folgende Eigenschaften für Vektoren $x, x', y, y' \in \mathbb{R}^n$ und Skalare $\lambda \in \mathbb{R}$:*

*i) Bilinearität:*

$$\begin{aligned} &\langle x + x', y \rangle \;=\; \langle x, y \rangle + \langle x', y \rangle, \quad \langle x, y + y' \rangle \;=\; \langle x, y \rangle + \langle x, y' \rangle, \\ &\lambda \langle x, y \rangle \;=\; \langle \lambda x, y \rangle \;=\; \langle x, \lambda y \rangle. \end{aligned}$$

*ii) Symmetrie:*

$$\langle x, y \rangle \;=\; \langle y, x \rangle.$$

*iii) Positive Definitheit:*

$$\langle x, x \rangle \;=\; x_1^2 + \ldots + x_n^2 \;\geqslant\; 0, \quad \langle x, x \rangle \;=\; 0 \;\Leftrightarrow\; x = \vec{0}.$$

Die dritte Eigenschaft von Lemma 3.2 nutzt bereits eine Eigenschaft des Körpers $\mathbb{R}$ aus, nämlich, dass die Summe von quadratischen Termen nie negativ werden kann. Das bedeutet, wir können aus dem Skalarprodukt eines Vektors mit sich selbst immer die Wurzel ziehen ohne den Körper $\mathbb{R}$ zu verlassen. Diese wichtige Beobachtung induziert den Begriff der *Norm* in folgender Definition, welche wir zur Messung von Längen eines Vektor nutzen können.

**Definition 3.3** (Norm in $\mathbb{R}^n$)
*Sei $x \in \mathbb{R}^n$ ein Vektor. Dann definieren wir die Abbildung*

$$\begin{aligned} ||\cdot|| \;:\; \mathbb{R}^n \;&\to\; \mathbb{R}_0^+, \\ x \;&\to\; ||x|| \;\coloneqq\; \sqrt{\langle x, x \rangle} \;=\; \sqrt{x_1^2 + \ldots + x_n^2} \end{aligned}$$

*als* Euklidische Norm *von $x$ in $\mathbb{R}^n$.*

**Bemerkung 3.4**
*Im eindimensionalen Fall $V = \mathbb{R}$ reduziert sich die Norm eines Vektors $x \in \mathbb{R}$ auf die Betragsfunktion, d.h., $||x|| = |x|$ für $\dim(V) =: n = 1$.*

Die oben definierte Norm $||x||$ gibt uns den Abstand des Punktes $x \in \mathbb{R}^n$ zum Nullpunkt $\vec{0} \in \mathbb{R}^n$, also gerade die Länge des Vektors $(x - \vec{0}) \in \mathbb{R}^n$.

Folgender wichtiger Satz hilft uns dabei die Eigenschaften der Norm noch besser zu charakterisieren.

**Satz 3.5** (Cauchy-Schwarz Ungleichung)
*Seien $x, y \in \mathbb{R}^n$. Dann gilt*

$$|\langle x, y\rangle| \;\leqslant\; ||x|| \cdot ||y||. \tag{3.1}$$

*Und es gilt außerdem*

$$|\langle x, y\rangle| \;=\; ||x|| \cdot ||y|| \quad \Leftrightarrow \quad x \text{ und } y \text{ sind linear abhängig.}$$

*Beweis.* Da die Wurzelfunktion monoton ist, können wir beide Seiten quadrieren, ohne dass sich an der Aussage etwas ändert. Wir müssen also zeigen, dass gilt

$$||x||^2 \cdot ||y||^2 - |\langle x, y\rangle|^2 \;=\; \langle x, x\rangle \cdot \langle y, y\rangle - \langle x, y\rangle^2 \;\geqslant\; 0. \tag{3.2}$$

Wir bedienen uns des folgenden Tricks. Wir betrachten eine Matrix $A \in \mathbb{R}^{2\times n}$ mit

$$A \;:=\; \begin{pmatrix} x_1 & \dots & x_n \\ y_1 & \dots & y_n \end{pmatrix}.$$

Multiplizieren wir nun $A$ von rechts mit $A^T \in \mathbb{R}^{n\times 2}$, so erhalten wir

$$AA^T \;=\; \begin{pmatrix} \langle x, x\rangle & \langle x, y\rangle \\ \langle y, x\rangle & \langle y, y\rangle \end{pmatrix}.$$

Wegen der Symmetrie des Skalarprodukts können wir die linke Seite der Gleichung (3.2) ersetzen durch den Ausdruck $\det(AA^T)$ und müssen nur zeigen, dass $\det(AA^T) \geqslant 0$ gilt. Nach dem Determinanten-Multiplikationstheorem in [Fischer2005, Satz 3.3.7] lässt sich die Determinante von $AA^T$ über die Produkte aller $(2 \times 2)$-Minoren ausrechnen und im vorliegenden Fall gilt

$$\det(AA^T) \;=\; \sum_{1\leqslant k_1\leqslant k_2\leqslant n} \Big(\det(A^{k_1,k_2})\Big)^2 \;\geqslant\; 0.$$

Die Nichtnegativität der Determinante nutzt hier wieder eine Eigenschaft des Körpers $\mathbb{R}$, da die Summe von quadratischen Termen nie negativ werden kann.

Es bleibt nur noch die Gleichheit zu überprüfen.

$$\begin{aligned}
&|\langle x, y\rangle| \;=\; ||x|| \cdot ||y|| \\
\Leftrightarrow\; &\det(AA^T) \;=\; \sum_{1\leqslant k_1\leqslant k_2\leqslant n} \Big(\det(A^{k_1,k_2})\Big)^2 \;=\; 0 \\
\Leftrightarrow\; &\det(A^{k_1,k_2}) \;=\; 0, \quad \text{für alle } 1 \leqslant k_1 \leqslant k_2 \leqslant n \\
\Leftrightarrow\; &\mathrm{Rang}(A) < 2 \\
\Leftrightarrow\; &x \text{ und } y \text{ sind linear abhängig .}
\end{aligned}$$

□

Mit Hilfe der Cauchy-Schwarz Ungleichung können wir die wichtigsten Eigenschaften der Norm zeigen.

**Lemma 3.6** (Eigenschaften der Norm)
*Seien $x, y \in \mathbb{R}^n$ Vektoren und $\lambda \in \mathbb{R}$ ein Skalar, dann gelten für die durch das kanonische Skalarprodukt induzierte Norm folgende Eigenschaften.*

*i)* $||x|| \;=\; 0 \quad \Leftrightarrow \quad x \;=\; 0,$

*ii)* $||\lambda x|| \;=\; |\lambda| \cdot ||x||,$

*iii)* $||x + y|| \;\leqslant\; ||x|| + ||y||.$

*Beweis.* Die ersten beiden Eigenschaften *i)* und *ii)* lassen sich direkt durch Nachrechnen verifizieren. Die dritte Eigenschaft der Dreiecksungleichung können wir mit Hilfe der Cauchy-Schwarz Ungleichung aus Satz 3.5 zeigen. Es gilt

$$\begin{aligned} ||x + y||^2 \;&=\; \langle x + y, x + y\rangle \;=\; \langle x, x\rangle + 2\langle x, y\rangle + \langle y, y\rangle \\ &\leqslant\; \langle x, x\rangle + 2||x|| \cdot ||y|| + \langle y, y\rangle \;\leqslant\; (||x|| + ||y||)^2. \end{aligned}$$

Da die Wurzelfunktion monoton ist gilt obige Abschätzung auch ohne die Quadrate auf der linken und rechten Seite der Ungleichung. □

Dadurch, dass die Norm die Länge eines Vektors beschreibt, können wir sie nutzen um Abstände zu beschreiben. Um nun den Abstand zwischen zwei Punkten $x, y \in \mathbb{R}^n$ zu messen führt man den Begriff einer durch die Norm induzierten Metrik ein.

**Definition 3.7** (Metrik)
*Seien $x, y \in \mathbb{R}^n$ zwei Punkte. Dann induziert die Norm $||\cdot||$ des Vektorraums eine* Metrik *genannte Abbildung*

$$\begin{aligned} d\colon \mathbb{R}^n \times \mathbb{R}^n \;&\to\; \mathbb{R}_0^+, \\ (x, y) \;&\mapsto\; d(x, y) \;:=\; ||x - y|| \;=\; \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}, \end{aligned}$$

*mit der man den Abstand (bezüglich der Norm) zwischen $x$ und $y$ messen kann.*

**Bemerkung 3.8**
*Wie man im Laufe des Studiums noch lernt, wird je nach Wahl der Norm in Definition 3.3 eine andere Metrik induziert, die den Abstandsbegriff maßgeblich beeinflusst. Letzten*

*Endes bestimmt dies auch die sogenannte* Topologie *eines Vektorraums. Dies wird durch folgende Relationen dargestellt:*

***Skalarprodukt*** *(Euklidischer / Unitärer Vektorraum)*
↓ *induziert*
***Norm*** *(Normierter Vektorraum)*
↓ *induziert*
***Metrik*** *(Metrischer Vektorraum)*
↓ *induziert*
***Topologie*** *(Topologischer Vektorraum)*

Wir können nun wesentliche Eigenschaften einer Metrik festhalten.

**Lemma 3.9** (Eigenschaften einer Metrik)
*Seien $x, y \in \mathbb{R}^n$, dann gelten für die Metrik $d(x,y) = ||x-y||$ folgende Eigenschaften.*

*i)* $d(x,y) \;=\; 0 \quad\Leftrightarrow\quad x \;=\; y,$

*ii)* $d(x,y) \;=\; d(y,x),$

*iii)* $d(x,z) \;\leqslant\; d(x,y) + d(y,z).$

*Beweis.* Die ersten beiden Eigenschaften *i*) und *ii*) sind klar und lassen sich direkt durch Nachrechnen verifizieren. Die dritte Eigenschaft *iii*) folgt direkt aus der Dreiecksungleichung der Norm in Lemma 3.6, da

$$d(x,z) \;=\; ||x-z|| \;=\; ||x-y+y-z|| \;\leqslant\; ||x-y|| + ||y-z|| \;=\; d(x,y) + d(y,z).$$

□

**Beispiel 3.10** (Manhattan Metrik)
*Eine interessante Metrik, die nicht durch die Euklidische Norm induziert wird, ist die sogenannte* Manhattan Metrik*, die für zwei Vektoren $x, y \in \mathbb{R}^n$ wie folgt definiert ist:*

$$d(x,y) \;:=\; \sum_{i=1}^{n} |x_i - y_i| \;=:\; ||x-y||_1.$$

*Der Name lehnt sich an die orthogonalen Straßengitter Manhattans an, die dazu führen, dass man zum Erreichen eines Ziels nur durch Aneinanderreihung vertikaler und horizontaler Wegstücke zu überwinden. Hierbei weisen alle Wege, die einen näher zum Ziel bringen, immer die gleiche Entfernung auf.*

Die Cauchy-Schwarz Ungleichung in Satz 3.5 liefert uns nicht nur nützliche Aussagen bei der Längenmessung (wie die Dreiecksungleichung für die Norm), sondern liefert zudem eine Methode zur *Winkelmessung* im $\mathbb{R}^n$. Betrachtet man nämlich den Quotienten der linken und rechten Seite in der Cauchy-Schwarz Ungleichung (3.1), so erhält man für zwei Vektoren $x, y \in \mathbb{R}^n$ mit $x, y \neq \vec{0}$ folgende Abschätzung

$$-1 \leqslant \frac{\langle x, y\rangle}{||x|| \cdot ||y||} \leqslant 1. \tag{3.3}$$

Der Quotient ist also ähnlich beschränkt wie die Kosinusfunktion $\cos\colon [0, \pi] \to [-1, 1]$ und es wird klar, dass es für fixe Vektoren $x$ und $y$ genau ein Winkel $\alpha \in [0, \pi]$ existiert, so dass

$$\cos\alpha \;=\; \frac{\langle x, y\rangle}{||x|| \cdot ||y||}.$$

Dies macht auch geometrisch Sinn, da das Skalarprodukt $\langle x, y\rangle$ von $x$ und $y$ als Maß für den Anteil des einen Vektors, der in Richtung des anderen Vektors zeigt interpretiert werden kann. Je größer dieser Anteil ist, desto kleiner ist der Winkel $\alpha$ zwischen den beiden Vektoren $x$ und $y$.

Dies motiviert die folgende Definition zur Winkelmessung in $\mathbb{R}^n$.

**Definition 3.11** (Winkelmessung in $\mathbb{R}^n$)
*Seien $x, y \in \mathbb{R}^n$ zwei Vektoren mit $x, y \neq \vec{0}$. Dann definieren wir den Winkel $\measuredangle(x, y)$ zwischen $x$ und $y$ als den Arkuskosinus des Quotienten in* (3.3)*, d.h.*

$$\measuredangle(x, y) \;:=\; \arccos \frac{\langle x, y\rangle}{||x|| \cdot ||y||}. \tag{3.4}$$

**Bemerkung 3.12**
*Aus den Eigenschaften des Skalarprodukts in Lemma 3.2 folgen direkt folgende sinnvolle Beobachtungen für Vektoren $x, y \in \mathbb{R}^n$ mit $x, y \neq \vec{0}$ und positive Skalare $\lambda > 0$:*

*i)* $\measuredangle(x, y) \;=\; \measuredangle(y, x)$,

*ii)* $\measuredangle(\lambda x, y) \;=\; \measuredangle(x, \lambda y) \;=\; \measuredangle(x, y)$.

Zur Veranschaulichung dieser wichtigen geometrischen Erkenntnisse soll das folgende Beispiel dienen.

**Beispiel 3.13**
*Um zu verstehen, dass die Definition 3.11 für die Winkelmessung sich mit unserer Vorstellung eines unorientierten Winkels deckt, betrachten wir zwei Beispiele:*

1. *Es seien zwei Vektoren $x, y \in \mathbb{R}^3$ gegeben mit:*

$$x \;:=\; \begin{pmatrix}1\\2\\3\end{pmatrix}, \quad y \;:=\; \begin{pmatrix}4\\5\\6\end{pmatrix}.$$

*Um den Winkel $\alpha \in [0, \pi]$ zwischen $x$ und $y$ mittels der Formel (3.4) zu berechnen, sehen wir uns zunächst die einzelnen Terme an:*

$$\begin{aligned} \langle x, y \rangle &= \langle (1,2,3)^T, (4,5,6)^T \rangle = 4 + 10 + 18 = 32, \\ ||x|| &= \sqrt{1+4+9} = \sqrt{14}, \qquad ||y|| = \sqrt{16+25+36} = \sqrt{77}. \end{aligned}$$

*Damit können wir den Winkel $\alpha$ wie folgt bestimmen:*

$$\alpha \coloneqq \measuredangle(x, y) = \arccos \frac{\langle x, y \rangle}{||x|| \cdot ||y||} = \arccos \frac{32}{\sqrt{14 \cdot 77}} \approx 13^\circ.$$

2. *Für zwei allgemeine Vektoren $x, y \in \mathbb{R}^2$ führen wir normierte Varianten $x', y' \in \mathbb{R}^n$ mit*

$$x' \coloneqq \frac{x}{||x||}, \quad y' \coloneqq \frac{y}{||y||}.$$

*Aus der Normierung wird klar, dass $||x|| = ||y|| = 1$ gilt und auf Grund von Bemerkung 3.12 sehen wir ein, dass $\measuredangle(x, y) = \measuredangle(x', y')$.*

*Da $x'$ und $y'$ auf dem Einheitskreis $\mathbb{S}^1$ liegen, wissen wir aus der Analysis, dass eine Parametrisierung mit Winkeln $\alpha, \beta \in [0, 2\pi[$ gibt, so dass*

$$x' = (\cos\alpha, \sin\alpha), \quad y' = (\cos\beta, \sin\beta).$$

*Mit den Additionsregeln für den Kosinus in [Burger2020, Kapitel 5.5] können wir also schreiben*

$$\langle x', y' \rangle = \cos\alpha\cos\beta + \sin\alpha\sin\beta = \cos(\beta - \alpha),$$

*wobei wir $\beta - \alpha \in [0, \pi]$ annehmen können.*

*Insgesamt folgt also, dass*

$$\measuredangle(x, y) = \beta - \alpha.$$

Für zwei Vektoren $x, y \in \mathbb{R}^n$ mit $x, y \neq \vec{0}$ wird klar, dass der Winkel zwischen den beiden genau $\frac{\pi}{2}$ (oder $90^\circ$) entspricht, wenn das Skalarprodukt $\langle x, y \rangle = 0$ ist. Das motiviert die folgende Definition.

**Definition 3.14** (Orthogonalität von Vektoren)
*Zwei Vektoren $x, y \in \mathbb{R}^n$ heißen* senkrecht *oder* orthogonal*, wenn für sie gilt*

$$\langle x, y \rangle = 0.$$

*Diese Definition gilt auch, wenn einer der beiden Vektoren der Nullvektor ist. Sind die Vektoren $x$ und $y$ orthogonal, so schreiben wir auch häufig $x \perp y$.*

## 3.2 Das Vektorprodukt in $\mathbb{R}^3$

Im Folgenden beschäftigen wir uns mit einer ganz speziellen Abbildung in $\mathbb{R}^3$, die zum Einen aus Sicht der Geometrie sehr nützlich ist und zum Anderen in der Physik eine wichtige Rolle spielt. Diese Abbildung wird Vektorprodukt genannt und taucht beispielsweise bei der Berechnung der Lorentzkraft im Elektromagnetismus auf oder wird für die Berechnung des Drehmoments in der klassischen Mechanik verwendet.

Das Vektorprodukt zweier Vektoren ist wie folgt definiert.

**Definition 3.15** (Vektorprodukt in $\mathbb{R}^3$)
*Seien $x, y \in \mathbb{R}^3$ zwei Vektoren mit $x = (x_1, x_2, x_3)^T$ und $y = (y_1, y_2, y_3)^T$. Dann wird das* Vektorprodukt *von $x$ und $y$, auch häufig* Kreuzprodukt *oder* äußeres Produkt *genannt, als folgende Abbildung definiert*

$$\begin{aligned}(\cdot \times \cdot)\colon \mathbb{R}^3 \times \mathbb{R}^3 &\to \mathbb{R}^3 \\ (x, y) &\mapsto x \times y := \begin{pmatrix} x_2y_3 - x_3y_2 \\ x_3y_1 - x_1y_3 \\ x_1y_2 - x_2y_1 \end{pmatrix}.\end{aligned}$$

Wir wollen zuerst einige grundlegende Eigenschaften des Vektorprodukts in $\mathbb{R}^3$ festhalten.

**Lemma 3.16** (Eigenschaften des Vektorprodukts)
*Seien $x, x', y, y' \in \mathbb{R}^3$ Vektoren und $\lambda \in \mathbb{R}$ ein Skalar. Dann gelten folgende Eigenschaften für das Vektorprodukt:*

*i) Bilinearität:*

$$\begin{aligned}(x + x') \times y &= x \times y + x' \times y, \quad x \times (y + y') = x \times y + x \times y', \\ \lambda x \times y &= \lambda(x \times y) = (x \times \lambda y).\end{aligned}$$

*ii) Antikommutativität:*

$$x \times y = -y \times x.$$

*iii) Alternierende Abbildung:*

$$x \times y = \vec{0} \quad \Leftrightarrow \quad x \text{ und } y \text{ sind linear abhängig.}$$

*Beweis.* Man kann die Eigenschaften des Vektorprodukts durch Anwendung der Definition direkt nachrechnen. Daher verzichten wir an dieser Stelle auf einen Beweis. Es sei jedoch angemerkt, dass man sowohl aus Eigenschaft *ii*) als auch Eigenschaft *iii*) des Vektorprodukts folgern kann, dass $x \times x = \vec{0}$ gilt. □

Es gibt außerdem einen engen Zusammenhang zwischen dem Vektorprodukt in $\mathbb{R}^3$ und dem kanonischen Skalarprodukt aus Kapitel 3.1, wie die folgenden Rechenregeln zeigen.

**Lemma 3.17**
*Seien $x, y, z \in \mathbb{R}^3$ Vektoren, dann gelten die folgenden Rechenregeln für das Vektorprodukt in $\mathbb{R}^3$:*

$$\textit{i)}\ \langle x \times y, z\rangle \ = \ \det \begin{pmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ z_1 & z_2 & z_3 \end{pmatrix},$$

$$\textit{ii)}\ \langle x \times y, x\rangle \ = \ \langle x \times y, y\rangle \ = \ 0,$$

$$\textit{iii)}\ ||x \times y||^2 \ = \ ||x||^2 \cdot ||y||^2 - \langle x, y\rangle^2,$$

$$\textit{iv)}\ ||x \times y||^2 \ = \ ||x||^2 \cdot ||y||^2 \cdot (\sin \measuredangle(x, y))^2.$$

*Beweis.* In der Hausaufgabe zu zeigen. □

Die in Lemma 3.17 beschriebenen mathematischen Zusammenhänge lassen sich geometrisch sehr gut deuten. So sagt die zweite Eigenschaft *ii*), dass das Vektorprodukt immer orthogonal zur Ebene ausgerichtet ist, die von den beiden Vektoren $x$ und $y$ in $\mathbb{R}^3$ aufgespannt wird. Andererseits drückt die vierte Eigenschaft *iv*) aus, dass die Länge des Vektorprodukts gegeben ist durch die Fläche des Parallelograms, dass durch die Vektoren $x$ und $y$ beschrieben wird.

Abschließend wollen wir das Vektorprodukt für zwei Beispiele berechnen.

**Beispiel 3.18**
*Seien $x, y \in \mathbb{R}^3$ Vektoren für die wir im Folgenden das Kreuzprodukt berechnen wollen.*

1. *Seien die Vektoren $x$ und $y$ gegeben als*

$$x \ := \ \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \qquad y \ := \ \begin{pmatrix} -1 \\ 0 \\ 2 \end{pmatrix}.$$

*Wir berechnen das Vektorprodukt $x \times y$ von $x$ und $y$ als*

$$x \times y \ = \ \begin{pmatrix} x_2y_3 - x_3y_2 \\ x_3y_1 - x_1y_3 \\ x_1y_2 - x_2y_1 \end{pmatrix} \ = \ \begin{pmatrix} 2 \cdot 2 - 1 \cdot 0 \\ 1 \cdot (-1) - 1 \cdot 2 \\ 1 \cdot 0 - 2 \cdot (-1) \end{pmatrix} \ = \ \begin{pmatrix} 4 \\ -3 \\ 2 \end{pmatrix} \ =: \ z.$$

*Wir verifizieren, dass das Vektorprodukt $x \times y =: z$ orthogonal zu $x$ und $y$ steht.*

$$\langle x, z\rangle \;=\; (1,2,1)\cdot\begin{pmatrix}4\\-3\\2\end{pmatrix} \;=\; 1\cdot 4 + 2\cdot(-3) + 1\cdot 2 \;=\; 4-6+2 \;=\; 0,$$

$$\langle y, z\rangle \;=\; (-1,0,2)\cdot\begin{pmatrix}4\\-3\\2\end{pmatrix} \;=\; (-1)\cdot 4 + 0\cdot(-3) + 2\cdot 2 \;=\; -4+0+4 \;=\; 0.$$

2. *Seien die Vektoren $x$ und $y$ gegeben als*

$$x \;:=\; \begin{pmatrix}1\\2\\1\end{pmatrix}, \qquad y \;:=\; \begin{pmatrix}-2\\-4\\-2\end{pmatrix}.$$

*Wir berechnen das Vektorprodukt $x \times y$ von $x$ und $y$ als*

$$x \times y \;=\; \begin{pmatrix}x_2y_3 - x_3y_2\\x_3y_1 - x_1y_3\\x_1y_2 - x_2y_1\end{pmatrix} \;=\; \begin{pmatrix}2\cdot(-2) - 1\cdot(-4)\\1\cdot(-2) - 1\cdot(-2)\\1\cdot(-4) - 2\cdot(-2)\end{pmatrix} \;=\; \begin{pmatrix}0\\0\\0\end{pmatrix} \;=:\; z.$$

*Da die Vektoren $x$ und $y$ offensichtlich linear abhängig sind ist das Vektorprodukt $x \times y = \vec{0}$.*

## 3.3 Das kanonische Skalarprodukt in $\mathbb{C}^n$

Bevor wir allgemeine Bilinear- und Sesquilinearformen einführen, wollen wir noch einen wichtigen Spezialfall behandeln, nämlich das kanonische Skalarprodukt in $\mathbb{C}^n$.

Wie bereits in [Burger2020, Kapitel 2.2.7] beschrieben, betrachten wir komplexe Zahlen $z \in \mathbb{C}$, die für $(a, b) \in \mathbb{R}^2$ dargestellt werden können als $z = a + ib$. Das Produkt $z_1 \cdot z_2$ zweier komplexer Zahlen $z_1 = x_1 + iy_1$ und $z_2 = x_2 + iy_2$ ist definiert als

$$z_1 \cdot z_2 \;=\; (x_1x_2 - y_1y_2) + i(x_1y_2 + x_2y_1).$$

Auch bekannt ist, dass der *Betrag* $|z|$ von $z$ definiert ist als $|z| := \sqrt{a^2 + b^2}$, was mit der in Definition 3.3 eingeführten Norm in $\mathbb{R}^2$ übereinstimmt. Außerdem ist die komplexe Konjugation $\overline{z}$ von $z$ definiert als $\overline{z} := a - bi$, das heißt wir vertauschen das Vorzeichen von $b \in \mathbb{R}$.

Mit diesen Eigenschaften können wir das kanonische Skalarprodukt in $\mathbb{C}^n$ einführen.

**Definition 3.19** (Skalarprodukt in $\mathbb{C}^n$)
*Seien $z = (z_1, \ldots, z_n) \in \mathbb{C}^n$ und $w = (w_1, \ldots, w_n) \in \mathbb{C}^n$ zwei Vektoren. Wir bezeichnen die*

*Abbildung*

$$\begin{aligned}\langle\cdot,\cdot\rangle_c \colon \mathbb{C}^n \times \mathbb{C}^n &\to \mathbb{C},\\ (z,w) &\mapsto \langle z,w\rangle_c := z_1\overline{w_1} + \ldots + z_n\overline{w_n} = \sum_{i=1}^{n} z_i\overline{w_i}\end{aligned}$$

*als das* kanonische Skalarprodukt *in* $\mathbb{C}^n$*. Häufig wird das kanonische Skalarprodukt auch* komplexes Skalarprodukt *genannt.*

Der Index $c$ in der Notation des komplexen Skalarprodukts kann auch weggelassen werden, wenn aus dem Kontext ersichtlich ist, dass man es mit komplexen Zahlen rechnet. Wir werden daher auch im Folgenden die vereinfachte Notation ohne den Index $c$ verwenden.

Folgende Eigenschaften des komplexen Skalarprodukts lassen sich leicht nachrechnen.

**Lemma 3.20** (Eigenschaften des Skalarprodukts)
*Das kanonische Skalarprodukt von* $\mathbb{C}^n$ *in Definition 3.19 besitzt folgende Eigenschaften für Vektoren* $z, z', w, w' \in \mathbb{C}^n$ *und Skalare* $\lambda \in \mathbb{C}$*:*

*i) Sesquilinearität:*

$$\begin{aligned}\langle z+z', w\rangle &= \langle z,w\rangle + \langle z',w\rangle, \quad \langle z, w+w'\rangle = \langle z,w\rangle + \langle z,w'\rangle,\\ \langle \lambda z, w\rangle &= \lambda\langle z,w\rangle, \quad \langle z,\lambda w\rangle = \overline{\lambda}\langle z,w\rangle.\end{aligned}$$

*ii) Hermitezität:*

$$\langle z,w\rangle = \overline{\langle w,z\rangle}.$$

*iii) Positive Definitheit:*

$$\langle z,z\rangle = |z_1|^2 + \ldots + |z_n|^2 \in \mathbb{R}_0^+, \quad \langle z,z\rangle = 0 \Leftrightarrow z = \vec{0}.$$

Die dritte Eigenschaft von Lemma 3.20 erlaubt es uns den Begriff der Norm eines komplexwertigen Vektors in der folgenden Definition einzuführen.

**Definition 3.21** (Norm in $\mathbb{C}^n$)
*Sei* $z \in \mathbb{C}^n$ *ein Vektor. Dann definieren wir die Abbildung*

$$\begin{aligned}||\cdot|| : \mathbb{C}^n &\to \mathbb{R}_0^+,\\ z &\to ||z|| := \sqrt{\langle z,z\rangle} = \sqrt{|z_1|^2 + \ldots + |z_n|^2}\end{aligned}$$

*als* Norm *von* $z$ *in* $\mathbb{C}^n$*.*

Besonders interessant ist die Beziehung zum kanonischen Skalarprodukt in $\mathbb{R}^n$. Bezüglich der natürlichen Inklusion $\mathbb{R} \subset \mathbb{C}$ mit $x = x + i \cdot 0$ gibt es kein Problem. Daher können wir $\mathbb{R}^n \subset \mathbb{C}^n$ betrachten und es wird klar, dass $\langle \cdot, \cdot \rangle_c$ eine natürliche Fortsetzung des kanonischen Skalarprodukts $\langle \cdot, \cdot \rangle$ in $\mathbb{R}^n$ ist.

Betrachten wir jedoch die Darstellung

$$\begin{aligned} \mathbb{R}^{2n} &\to \mathbb{C}^n, \\ v = (x_1, y_1, \ldots, x_n, y_n) &\mapsto (x_1 + iy_1, \ldots, x_n + iy_n) = z, \end{aligned} \tag{3.5}$$

so können wir das komplexe Skalarprodukt in $\mathbb{C}^n$ in Relation zum kanonischen Skalarprodukt in $\mathbb{R}^{2n}$ wie folgt setzen. Seien $v, v' \in \mathbb{R}^{2n}$ zwei Vektoren, denen die komplexen Vektoren $z, z' \in \mathbb{C}^n$ mit Abbildung (3.5) entsprechen. Dann können wir schreiben:

$$\langle z, z' \rangle_c = \sum_{j=1}^{n} z_j \overline{z'_j} = \sum_{j=1}^{n} (x_j x'_j + y_j y'_j) - i \underbrace{\sum_{j=1}^{n} \det \begin{pmatrix} x_j & y_j \\ x'_j & y'_j \end{pmatrix}}_{=:\omega(v,v')} = \langle v, v' \rangle - i\,\omega(v, v').$$

Auf Grund der Determinante innerhalb der Summe ist klar, dass $\omega(v, v') = -\omega(v', v)$ und $\omega(v, v) = 0$ gilt.

Wir erkennen also, dass das kanonische Skalarprodukt in $\mathbb{R}^{2n}$ gerade der Realteil des komplexen Skalarprodukts in $\mathbb{C}^n$ ist, zu dem noch ein alternierender Imaginärteil hinzukommt. Da die Abbildung $\omega \colon \mathbb{R}^{2n} \times \mathbb{R}^{2n} \to \mathbb{R}$ für gleiche Vektoren verschwindet, stellen wir fest, dass das kanonische Skalarprodukt in $\mathbb{R}^{2n}$ und das komplexe Skalarprodukt in $\mathbb{C}^n$ die gleiche Norm induzieren.

## 3.4 Bilinear- und Sesquilinearformen

Die im vorigen Kapitel besprochenen Werkzeuge zur Messung von Längen, Abständen und Winkeln leiteten sich aus dem Begriff des Skalarprodukts ab. Wir werden im Folgenden sehen, dass ein Skalarprodukt nur ein Spezialfall einer ganzen Familie von allgemeineren Abbildungen ist, den sogenannten Bilinear- und Sesquilinearformen. Zur Untersuchung dieser abstrakteren mathematischen Begriffe sei im Folgenden $\mathbb{K}$ wieder ein beliebiger Körper.

**Definition 3.22** (Bilinearform)
*Seien $v, v', w, w' \in V$ und ein Skalar $\lambda \in \mathbb{K}$ gegeben. Wir nennen eine Abbildung*

$$\begin{aligned} s : V \times V &\to \mathbb{K}, \\ (v, w) &\to s(v, w), \end{aligned}$$

Bilinearform *auf $V$, wenn sie* bilinear *ist, d.h., linear in beiden Argumenten mit*

$$\begin{aligned} s(v + v', w) = s(v,w) + s(v', w), \quad s(v, w + w') &= s(v, w) + s(v, w'), \\ s(\lambda v, w) = \lambda s(v, w) &= s(v, \lambda w). \end{aligned}$$

*Die Abbildung $s$ heißt* symmetrisch*, falls gilt*

$$s(v,w) \;=\; s(w,v)$$

*und* alternierend *oder* schiefsymmetrisch*, falls gilt*

$$s(v,w) \;=\; -s(w,v).$$

**Beispiel 3.23**
*Sei $\mathbb{K} = \mathbb{R}$ und $I = [a,b] \subset \mathbb{R}$ ein Intervall. Wir betrachten den Vektorrraum der auf dem Intervall $I$ stetigen Funktionen $V = C(I;\mathbb{R})$ und definieren eine Abbildung*

$$\begin{aligned} s\colon V \times V \;&\to\; \mathbb{R}, \\ (f,g) \;&\mapsto\; s(f,g) \;:=\; \int_a^b f(x)g(x)\,\mathrm{d}x. \end{aligned}$$

*Es folgt aus den Rechenregeln für Integrale, dass die Abbildung $s$ eine symmetrische Bilinearform auf $V$ darstellt.*

Im obigen Beispiel war der Vektorraum $V$ nicht endlich-dimensional. Schränken wir uns jedoch auf endlich-dimensionale Vektorräume ein, so lässt sich erkennen, dass sich Bilinearformen durch Matrizen beschreiben lassen.

**Definition 3.24** (Darstellende Matrix einer Bilinearform)
*Sei $V$ ein endlich-dimensionaler $\mathbb{K}$-Vektorraum und sei $B := (v_1,\dots,v_n)$ eine Basis von $V$. Außerdem sei $s$ eine Bilinearform auf $V$. Dann nennen wir die Matrix $M_B(s) \in \mathbb{K}^{n\times n}$ mit*

$$M_B(s) \;:=\; (s(v_i,v_j))_{1\leqslant i,j\leqslant n}$$

*die* darstellende Matrix *von $s$ bezüglich $B$.*

**Satz 3.25**
*Sei $s$ eine Bilinearform auf $V$ mit Basis $B = (v_1,\dots,v_n)$. Sei außerdem $\Phi_B\colon \mathbb{K}^n \to V$ das zugehörige Koordinatensystem zur Basis $B$ und $A := M_B(s)$ die darstellende Matrix von $s$ bezüglich $B$. Wenn wir zwei Vektoren $v,w \in V$ betrachten mit den Koordinaten*

$$x \;=\; \Phi_B^{-1}(v), \qquad y \;=\; \Phi_B^{-1}(w),$$

*dann gilt*

$$s(v,w) \;=\; (x_1,\dots,x_n) \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \;=\; x^T A y.$$

*Beweis.* Wir können schreiben

$$s(v,w) \;=\; s(x_1v_1+\ldots+x_nv_n, y_1v_1+\ldots+y_nv_n) \;=\; \sum_{i=1}^{n}\sum_{j=1}^{n} x_i s(v_i,v_j) y_j.$$

Wir sehen also, dass durch Ausnutzung der Bilinearität von $s$ wir eine Summe aus $n^2$ Summanden erhalten. Da $A$ die darstellende Matrix von $s$ bezüglich der Basis $B$ ist gilt $s(v_i,v_j)=a_{ij}$, womit schon die Behauptung folgt. □

Wir wollen die Darstellung einer Bilinearform durch eine Matrix durch folgende zwei Beispiele besser verstehen.

**Beispiel 3.26**
*Sei $V=\mathbb{R}^2$ ein Euklidischer Vektorraum mit dem kanonischen Skalarprodukt als symmetrischer Bilinearform $s(v,w) := \langle v,w\rangle$ auf $V$. Seien außerdem $v,w \in \mathbb{R}^2$ mit*

$$v \;:=\; \begin{pmatrix}2\\-1\end{pmatrix}, \qquad w \;:=\; \begin{pmatrix}0\\5\end{pmatrix}.$$

*Wir betrachten die darstellenden Matrix des Skalarprodukts bezüglich zweier unterschiedlicher Basen von $V$ im Folgenden.*

*i) Wir wählen im ersten Beispiel die kanonische Standardbasis $B=(e_1,e_2)$ des $\mathbb{R}^2$. Wir berechnen die darstellende Matrix des Skalarprodukts bezüglich der Basis $B$ als*

$$A \;:=\; M_B(s) \;=\; \begin{pmatrix}\langle e_1,e_1\rangle & \langle e_1,e_2\rangle\\ \langle e_2,e_1\rangle & \langle e_2,e_2\rangle\end{pmatrix} \;=\; \begin{pmatrix}1&0\\0&1\end{pmatrix} \;=\; I_2.$$

*Bezüglich der Basis $B$ erhalten wir für die Vektoren $v$ und $w$ die folgenden Koordinaten:*

$$x \;=\; \Phi_B(v) \;=\; (2,-1)^T \;=\; v, \qquad y \;=\; \Phi_B(w) \;=\; (0,5)^T \;=\; w.$$

*Damit gilt dann insgesamt:*

$$-5 \;=\; \langle (2,-1)^T,(0,5)^T\rangle \;=\; \langle v,w\rangle \;=\; x^TAy \;=\; v^TI_2w \;=\; \langle v,w\rangle \;=\; -5.$$

*ii) Wir wählen im zweiten Beispiel die Basis $B=(v_1,v_2)$ des $\mathbb{R}^2$ mit*

$$v_1 \;:=\; \begin{pmatrix}1\\2\end{pmatrix}, \qquad v_2 \;:=\; \begin{pmatrix}2\\-1\end{pmatrix}$$

*Wir berechnen die darstellende Matrix des Skalarprodukts bezüglich der Basis $B$ als*

$$A \;:=\; M_B(s) \;=\; \begin{pmatrix}\langle v_1,v_1\rangle & \langle v_1,v_2\rangle\\ \langle v_2,v_1\rangle & \langle v_2,v_2\rangle\end{pmatrix} \;=\; \begin{pmatrix}5&0\\0&5\end{pmatrix}.$$

*Bezüglich der Basis $B$ erhalten wir für die Vektoren $v$ und $w$ die folgenden Koordinaten:*

$$x \;=\; \Phi_B(v) \;=\; (0,1)^T, \qquad y \;=\; \Phi_B(w) \;=\; (2,-1)^T.$$

*Damit gilt dann insgesamt:*

$$\begin{aligned} -5 \;&=\; \langle (2,-1)^T, (0,5)^T \rangle \;=\; \langle v, w \rangle \;=\; x^T A y \\ &=\; (0,1)^T \cdot \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ -1 \end{pmatrix} \;=\; (0,1)^T \begin{pmatrix} 10 \\ -5 \end{pmatrix} \;=\; -5. \end{aligned}$$

Zwischen Matrizen und Bilinearformen existiert ein starker Zusammenhang, wie wir im Folgenden feststellen.

**Bemerkung 3.27** (Bijektivität und Symmetrie von Bilinearformen)
*Sei $V$ ein endlich-dimensionaler $\mathbb{K}$-Vektorraum und $B$ eine zugehörige Basis von $V$. Dann ist die Abbildung*

$$s \;\mapsto\; M_B(s)$$

*von den Bilinearformen auf $V$ nach $\mathbb{K}^{n\times n}$* bijektiv*. Außerdem ist $s$ genau dann* symmetrisch*, wenn $M_B(s)$ symmetrisch ist.*

**Lemma 3.28**
*Seien $A, B \in \mathbb{K}^{n\times n}$ gegeben mit*

$$x^T A y \;=\; x^T B y,$$

*für alle $x, y \in \mathbb{K}^n$. Dann gilt schon $A = B$.*

*Beweis.* Da die Gleichung $x^T A y = x^T B y$ für alle Vektoren in $\mathbb{K}^n$ gilt, muss sie insbesondere auch für die Einheitsvektoren gelten und damit folgt für alle Indizes $1 \leqslant i, j \leqslant n$:

$$a_{i,j} \;=\; e_i^T A e_j \;=\; e_i^T B e_j \;=\; b_{i,j}.$$

Damit folgt also direkt, dass $A = B$ ist. □

**Satz 3.29** (Transformationsformel)
*Sei $V$ ein endlich-dimensionaler Vektorraum mit zwei Basen $A, B$. Sei außerdem $T_A^B$ die entsprechende Transformationsmatrix von Basis $B$ zu Basis $A$. Für jede Bilinearform $s$ auf $V$ gilt dann die folgende Transformationsformel:*

$$M_B(s) \;=\; (T_A^B)^T \cdot M_A(s) \cdot T_A^B. \tag{3.6}$$

*Beweis.* Seien $\Phi_A, \Phi_B \colon \mathbb{K}^n \to V$ Koordinatensysteme von $V$ bezüglich der Basen $A$ und $B$. Seien außerdem $v, v' \in V$ mit

$$v \;=\; \Phi_A(x) \;=\; \Phi_B(y), \qquad v' \;=\; \Phi_A(x') \;=\; \Phi_B(y').$$

Sei weiterhin $T := T_A^B$, so sehen wir ein, dass gilt

$$x = Ty, \qquad x' = Ty'.$$

Damit können wir folgern, dass gilt

$$y^T M_B(s) y' = s(v, v') = x^T M_A(s) x' = (Ty)^T M_A(s)(Ty') = y^T (T^T M_A(s) T) y'.$$

Mit Lemma 3.28 folgt dann schon, dass gelten muss:

$$M_B(s) = (T_A^B)^T \cdot M_A(s) \cdot T_A^B.$$

□

**Bemerkung 3.30**
*Das Transformationsverhalten bei Endomorphismen in Kapitel 2 unterscheidet sich von der Transformationsformel* (3.6)*, da für einen Endomorphismus $F\colon V \to V$ galt:*

$$M_B(F) = \underbrace{(T_A^B)^{-1}}_{=T_B^A} \cdot M_A(F) \cdot T_A^B.$$

**Definition 3.31** (Quadratische Form)
*Sei $s$ eine symmetrische Bilinearform auf $V$ und $v \in V$. Dann definieren wir die Abbildung*

$$\begin{aligned} q\colon V &\to \mathbb{K}, \\ v &\mapsto q(v) := s(v, v), \end{aligned}$$

*die zugehörige* quadratische Form *von $s$. Aus der Bilinearität von $s$ folgt für alle Skalare $\lambda \in \mathbb{K}$ direkt, dass gilt*

$$q(\lambda v) = \lambda^2 q(v).$$

Wir wollen die quadratische Form zu der wohl naheliegensten Bilinearform betrachten - dem kanonischen Skalarprodukt in $\mathbb{R}^n$.

**Beispiel 3.32**
*Sei $V = \mathbb{R}^n$ ein Euklidischer Vektorraum. Sei außerdem $v \in V$ und ein Skalar $\lambda \in \mathbb{R}$ gegeben. Wir betrachten das kanonische Skalarprodukt aus Definition 3.1 und bestimmen die zugehörige quadratische Form als*

$$q(v) = \langle v, v \rangle = v_1^2 + \ldots v_n^2 = ||v||^2.$$

*Wir sehen also ein, dass die quadratische Norm die quadratische Form des kanonischen Skalarprodukts darstellt und in der Tat gilt auch:*

$$q(\lambda v) = ||\lambda v||^2 = \lambda^2 ||v||^2 = \lambda^2 q(v).$$

**Definition 3.33** (Sesquilinearform)
*Sei $V$ ein komplexer Vektorraum und seien $v, v', w, w' \in V$ und ein Skalar $\lambda \in \mathbb{C}$ gegeben. Wir nennen eine Abbildung*

$$\begin{aligned} s \,:\, V \times V &\to \mathbb{C}, \\ (v, w) &\to s(v, w), \end{aligned}$$

Sesquilinearform *auf $V$, wenn sie linear im ersten Argument und semilinear im zweiten Argument ist, d.h.*

$$\begin{aligned} s(v + v', w) &= s(v, w) + s(v', w), \quad s(v, w + w') = s(v, w) + s(v, w'), \\ s(\lambda v, w) &= \lambda s(v, w), \quad s(v, \lambda w) = \overline{\lambda} s(v, w). \end{aligned}$$

*Die Abbildung $s$ heißt* hermitesch*, falls zusätzlich gilt*

$$s(v, w) = \overline{s(w, v)}.$$

*Zu einer Sesquilinearform $s$ auf $V$ definieren wir eine zugehörige* quadratische Form

$$\begin{aligned} q \,:\, V &\to \mathbb{C}, \\ v &\to q(v) := s(v, v). \end{aligned}$$

Neben dem komplexen Skalarprodukt $\langle \cdot, \cdot \rangle_c$ aus Definition 3.19 gibt es noch weitere Sesquilinearformen, wie das folgende Beispiel zeigt.

**Beispiel 3.34**
*Sei $\mathbb{K} = \mathbb{C}$ und $I = [a, b] \subset \mathbb{R}$ ein Intervall. Wir betrachten den Vektorrraum der auf dem Intervall $I$ stetigen, komplex-wertigen Funktionen $V = C(I; \mathbb{C})$ und definieren eine Abbildung*

$$\begin{aligned} s \colon V \times V &\to \mathbb{C}, \\ (f, g) &\mapsto s(f, g) := \int_a^b f(x) \overline{g}(x) \, \mathrm{d}x. \end{aligned}$$

*Es folgt aus den Rechenregeln für Integrale, dass die Abbildung $s$ eine Sesquilinearform auf $V$ darstellt.*

**Bemerkung 3.35**
*Ähnlich wie bei Bilinearformen können wir folgende Charakteristiken feststellen.*

*i) Eine Sesquilinearform kann auch durch eine Matrix beschrieben werden. Hierzu müssen wir allerdings die komplexe Konjugation berücksichtigen. Sei $B = (v_1, \ldots, v_n)$ eine*

*Basis eines komplexen Vektorraums $V$ und $s$ eine Sesquilinearform auf $V$. Die darstellende Matrix $M_B(s)$ von $s$ bezüglich der Basis $B$ ist gegeben durch:*

$$A \coloneqq (M_B(s))_{ij} = s(v_i, v_j), \qquad \text{für alle } 1 \leqslant i, j \leqslant n.$$

*Die Matrix $A$ ist genau dann hermitesch, d.h., $A^T = \overline{A}$, falls die Sesquilinearform $s$ hermitesch ist. Für zwei Vektoren $v, w \in V$ mit Koordinaten*

$$x = \Phi_B(v), \qquad y = \Phi_B(w),$$

*lässt sich die Sesquilinearform $s$ ausdrücken durch*

$$s(v,w) = (x_1, \ldots, x_n) \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} \overline{y_1} \\ \vdots \\ \overline{y_n} \end{pmatrix} = x^T A \overline{y}.$$

*ii) Sei $C$ eine weitere Basis von $V$ und sei $T \coloneqq T_B^C$ die Transformationsmatrix von Basis $C$ zu Basis $B$. Dann erhalten wir folgendes Transformationsverhalten für die darstellende Matrix $M_C(s)$ von $s$ bezüglich $C$:*

$$M_C(s) = T^T \cdot M_B(s) \cdot \overline{T}.$$

Es ist in der Tat möglich symmetrische Bilinearformen und Sesquilinearformen nur an Hand ihrer quadratischen Formen zu rekonstruieren. Diesen mathematischen Zusammenhang nennt man Polarisierung und wird durch folgenden Satz beschrieben.

**Satz 3.36** (Polarisierung)
*Sei $V$ ein $\mathbb{R}$-Vektorraum und $W$ ein $\mathbb{C}$-Vektorraum. Sei außerdem $b$ eine symmetrische Bilinearform auf $V$ mit zugehöriger quadratischer Form $q(v) = b(v,v)$ und $s$ eine Sesquilinearform auf $W$ mit zugehöriger quadratischer Form $p(v) = s(v,v)$. Dann gelten die folgenden,* Polarisierung *genannten, Zusammenhänge:*

*i)* $b(v,w) = \frac{1}{4}[q(v+w) - q(v-w)], \quad$ *für alle* $v, w \in V$,

*ii)* $s(v,w) = \frac{1}{4}[p(v+w) - p(v-w) + i \cdot p(v+iw) - i \cdot p(v-iw)], \quad$ *für alle* $v, w \in W$.

*Beweis.* In der Hausaufgabe zu zeigen. □

Im reellen Fall ist es entscheidend, dass man in Satz 3.36 eine symmetrische Bilinearform annimmt, da die Polarisierungsformel nicht für beliebige Bilinearformen gilt, wie das folgende Beispiel zeigt:

**Beispiel 3.37**
*Sei $V = \mathbb{R}^2$ und wir betrachten die folgende Bilinearform*

$$s(x,y) \;:=\; x^T \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} y \;=\; -x_1y_2 + x_2y_1.$$

*Es gilt offensichtlich für alle $x \in \mathbb{R}^2$*

$$s(x,x) \;=\; -x_1x_2 + x_1x_2 \;=\; 0.$$

*Andererseits gilt jedoch nicht die Polarisierungsformel, wie man im folgenden sieht*

$$1 \;=\; s((1,0)^T,(0,1)^T) \;\neq\; \frac{1}{4}[\underbrace{s((1,1)^T,(1,1)^T)}_{=0} - \underbrace{s((1,-1)^T,(1,-1)^T)}_{=0}] \;=\; 0.$$

**Definition 3.38** (Positive Definitheit)
*Wir wollen im folgenden den Begriff der positiven Definitheit für Matrizen und symmetrische Bilinearformen bzw. hermitesche Sesquilinearformen einführen. Sei $V$ ein $\mathbb{K}$-Vektorraum.*

*i) Eine symmetrische Bilinearform bzw. hermitesche Sesquilinearform $s\colon V \times V \to \mathbb{K}$ heißt* positiv definit *falls für alle $v \in V$ mit $v \neq 0$ gilt:*

$$s(v,v) \;>\; 0.$$

*ii) Eine symmetrische bzw. hermitesche Matrix $A \in \mathbb{K}^{n\times n}$ heißt* positiv definit *falls für alle $x \in \mathbb{K}^n$ mit $x \neq 0$ gilt:*

$$x^T A\overline{x} \;>\; 0.$$

Schlussendlich sind wir in der Lage den Begriff eines Euklidischen und unitären Vektorraums mit Hilfe von Bilinear- und Sesquilinearformen zu definieren.

**Definition 3.39** (Euklidischer und unitärer Vektorraum)
*Wir nennen eine positiv definite symmetrische Bilinearform bzw. eine positiv definite hermitesche Sesquilinearform ein* Skalarprodukt. *Einen reellen bzw. komplexen Vektorraum zusammen mit einem Skalarprodukt nennen wir* Euklidischen *bzw.* unitären Vektorraum.

## 3.5 Orthogonalisierung und Orthonormalisierung

In vielen Fällen ist es sinnvoll nicht eine beliebige Basis eines endlich-dimensionalen Vektorraums $V$ zu betrachten, sondern eine Familie von Vektoren, die *orthogonal* oder sogar *orthonormal* sind. Dies hat viele Vorteile für die Mathematik, da sich so manche Berechnung durch eine Orthonormalbasis deutlich vereinfachen lässt. Auch lassen sich durch orthonormale Vektoren längen- und winkelerhaltende Transformationen durchführen.

**Definition 3.40** (Orthogonalität und Orthonormalität)
*Sei $V$ ein Euklidischer bzw. unitärer Vektorraum. Dann können wir folgende Begriffe und Notation definieren:*

*i) Zwei Vektoren $u, v \in V$ heißen* orthogonal, *falls gilt:*

$$\langle v, w \rangle \;=\; 0.$$

*Wir notieren in diesem Fall häufig auch $v \perp w$.*

*ii) Zwei Untervektorräume $U, W \subset V$ heißen* orthogonal, *falls gilt:*

$$u \perp v \quad \text{für alle } u \in U \text{ und } w \in W.$$

*Wir notieren in diesem Fall häufig auch $U \perp W$.*

*iii) Ist $U \subset V$ ein Untervektorraum, so definieren wir sein* orthogonales Komplement *als*

$$U^{\perp} \;:=\; \{v \in V \mid v \perp u \text{ für alle } u \in U\}.$$

*Es ist klar, dass $U^{\perp}$ wieder ein Untervektorraum ist.*

*iv) Eine Familie von Vektoren $(v_1, \ldots, v_n)$ in $V$ heißt* orthogonal, *wenn gilt*

$$v_i \perp v_j \quad \text{für alle } i \neq j.$$

*Sie heißt* orthonormal, *falls zusätzlich gilt*

$$||v_i|| \;=\; 1 \quad \text{für alle } 1 \leqslant i \leqslant n.$$

*In diesem Fall gilt offenbar*

$$\langle v_i, v_j \rangle \;=\; \delta_{ij},$$

*wobei $\delta_{ij}$ das Kronecker-Delta bezeichnet (vgl. Definition 2.39).*

*v) Wir nennen eine Familie von orthonormalen Vektoren $(v_1, \ldots, v_n)$ in $V$ eine* Orthonormalbasis, *falls die Vektoren eine Basis von $V$ bilden.*

*vi) Ist $V = V_1 \oplus \ldots \oplus V_n$, so heißt die direkte Summe* orthogonal, *falls gilt*

$$V_i \perp V_j \quad \text{für alle } i \neq j.$$

*Wir notieren in diesem Fall häufig auch $V = V_1 \operatorname{\oplus} \ldots \operatorname{\oplus} V_n$.*

**Beispiel 3.41**
*Betrachten wir $\mathbb{R}^n$ oder $\mathbb{C}^n$ mit dem kanonischen bzw. komplexen Skalarprodukt, so ist die kanonische Basis $B = (e_1, \ldots, e_n)$ eine Orthonormalbasis.*

**Satz 3.42**
*Ist $(v_1, \ldots, v_n)$ eine orthogonale Familie von Vektoren in $V$ mit $v_i \neq 0$ für alle $1 \leqslant i \leqslant n$, so gelten die folgenden Aussagen.*

1. *Die Familie $(\alpha_1 v_1, \ldots, \alpha_n v_n)$ von Vektoren mit $\alpha_i := ||v_i||^{-1}$ ist orthonormal.*
2. *Die Familie $(v_1, \ldots, v_n)$ von Vektoren ist linear unabhängig.*

*Beweis.* Wir zeigen die beiden Behauptungen für ein allgemeines Skalarprodukt.

1. Da $v_i \perp v_j$ gilt für $i \neq j$ folgt schon, dass gilt

$$\langle \alpha_i v_i, \alpha_j v_j \rangle \;=\; \alpha_i \overline{\alpha_j} \langle v_i, v_j \rangle \;=\; 0, \qquad \text{für } i \neq j.$$

   Die Familie $(\alpha_1 v_1, \ldots, \alpha_n v_n)$ von Vektoren ist also orthogonal. Da $\alpha_i = ||v_i||^{-1} \in \mathbb{R}$ gilt sehen wir für den Fall $i = j$, dass gilt

$$\langle \alpha_i v_i, \alpha_i v_i \rangle \;=\; \alpha_i \overline{\alpha_i} \langle v_i, v_i \rangle \;=\; \frac{||v_i||^2}{||v_i||^2} \;=\; 1.$$

   Die Familie von Vektoren ist also orthonormal.

2. Wir müssen für die lineare Unabhängigkeit der Vektoren $v_1, \ldots, v_n \in V$ zeigen, dass aus der Gleichung

$$0 \;=\; \lambda_1 v_1 + \ldots + \lambda_n v_n \tag{3.7}$$

   bereits folgt, dass $\lambda_i = 0$ für $1 \leqslant i \leqslant n$ gelten muss. Multiplizieren wir also die Gleichung (3.7) von rechts mit $v_i^T$ so folgt:

$$0 \;=\; \langle 0, v_i \rangle \;=\; \langle \lambda_1 v_1 + \ldots + \lambda_n v_n, v_i \rangle \;=\; \sum_{j=1}^{n} \lambda_j \langle v_j, v_i \rangle \;=\; \lambda_i \langle v_i, v_i \rangle.$$

   Da das Skalarprodukt insbesondere positiv definit ist, muss also schon gelten, dass $\lambda_i = 0$ ist. Da dies unabhängig von der Wahl des Vektors $v_i$ gilt, müssen schon alle Koeffizienten $\lambda_i = 0$ für $1 \leqslant i \leqslant n$ gelten.

□

**Satz 3.43**
*Sei $(v_1, \ldots, v_n)$ eine Orthonormalbasis von $V$ und $v \in V$ ein beliebiger Vektor. Setzen wir $\lambda_i := \langle v_i, v \rangle$, so gilt:*

$$v \;=\; \lambda_1 v_1 + \ldots + \lambda_n v_n.$$

*Beweis.* Da $(v_1, \ldots, v_n)$ eine Orthonormalbasis von $V$ ist, existieren eindeutige Koeffizienten $\gamma_i, 1 \leqslant i \leqslant n$, so dass sich der beliebige Vektor $v \in V$ schreiben lässt als

$$v \;=\; \gamma_1 v_1 + \ldots + \gamma_n v_n.$$

Wir multiplizieren obige Gleichung von rechts mit dem Vektor $v_i^T$ und können die Koeffizienten damit eindeutig bestimmen als

$$\langle v, v_i \rangle \;=\; \langle \gamma_1 v_1 + \ldots + \gamma_n v_n, v_i \rangle \;=\; \gamma_i \langle v_i, v_i \rangle \;=\; \gamma_i.$$

Da dies für alle $1 \leqslant i \leqslant n$ gilt, können wir $\lambda_i \coloneqq \gamma_i = \langle v_i, v \rangle$ definieren und es gilt damit offensichtlich

$$v \;=\; \lambda_1 v_1 + \ldots + \lambda_n v_n.$$

□

In vielen Situationen ist es praktisch eine Orthonormalbasis zu betrachten, da sie viele Berechnungen vereinfacht. Lässt sich jedoch eine Orthonormalbasis für einen beliebigen Euklidischen bzw. unitären Vektorraum bestimmen? Darauf gibt glücklicherweise der folgende Satz eine zufriedenstellende Antwort.

**Satz 3.44** (Orthonormalisierungssatz)
*Sei $V$ ein endlich-dimensionaler Euklidischer bzw. unitärer Vektorraum und $W \subset V$ ein Untervektorraum mit Orthonormalbasis $(w_1, \ldots, w_m)$. Dann existiert eine Ergänzung aus Vektoren $w_{m+1}, \ldots, w_n \in V$, so dass*

$$(w_1, \ldots, w_m, w_{m+1}, \ldots, w_n)$$

*eine Orthonormalbasis von $V$ ergibt.*

*Beweis.* Da der Beweis des Satzes in konstruktiver Form erfolgt, formulieren wir diesen im Folgenden als einen konkreten Algorithmus. □

Da als Unterraum $W = \{0\}$ in Satz 3.44 erlaubt ist, erhalten wir direkt das folgende Korollar.

**Korollar 3.45**
*Jeder endlichdimensionale Euklidische bzw. unitäre Vektorraum besitzt eine Orthonormalbasis.*

Außerdem können wir folgendes Korollar aus dem Orthogonalisierungssatz 3.44 ableiten.

**Korollar 3.46**
*Ist $W \subset V$ Untervektorraum eines Euklidischen bzw. unitären Vektorraums $V$, so gilt*

$$V \;=\; W \oplus W^{\perp}, \quad \textit{und} \quad \dim V \;=\; \dim W + \dim W^{\perp}.$$

Die Konstruktion einer Orthonormalbasis geht auf die beiden Mathematiker *J. Gram* und *E. Schmidt* zurück und wird daher weitläufig auch als **Gram-Schmidtsches Orthogonalisierungsverfahren** bezeichnet. Bevor wir uns dem Verfahren widmen, wollen wir eine nützliche Abbildung einführen.

**Definition 3.47**
*Orthogonale Projektion Seien $V$ ein endlich-dimensionaler Euklidischer bzw. unitärer Vektorraum und $v, w \in V$ zwei linear unabhängige Vektoren. Dann bezeichnen wir die Abbildung*

$$\Pi_v(w) \;:=\; \frac{\langle w, v\rangle}{\langle v, v\rangle} \cdot v,$$

*als* orthogonale Projektion *von $w$ auf $v$. Die orthogonale Projektion berechnet den Anteil, den der Vektor $v$ an der Geometrie von $w$ hat.*

Die Kernidee des Gram-Schmidtschen Orthogonalisierungsverfahren ist es die Vektoren paarweise zueinander orthogonal auszurichten. Dabei spielt eine Korrektur mit Hilfe der orthogonalen Projektion eine zentrale Rolle, wie das folgende Lemma zeigt.

**Lemma 3.48**
*Seien $V$ ein endlich-dimensionaler Euklidischer bzw. unitärer Vektorraum und $v, w \in V$ zwei linear unabhängige Vektoren. Dann können wir einen Vektor $\hat{w} \in V$ berechnen mit*

$$\hat{w} \;:=\; w - \Pi_v(w) \;=\; w - \frac{\langle w, v\rangle}{\langle v, v\rangle} \cdot v,$$

*so dass $\hat{w} \perp v$ gilt.*

*Beweis.* Wir betrachten folgende Umformungen:

$$\begin{aligned}\langle v, \hat{w}\rangle \;&=\; \langle v, w - \frac{\langle w, v\rangle}{\langle v, v\rangle} \cdot v\rangle \;=\; \langle v, w\rangle - \overline{\left(\frac{\langle w, v\rangle}{\langle v, v\rangle}\right)} \cdot \langle v, v\rangle \\ &=\; \langle v, w\rangle - \overline{\langle w, v\rangle} \cdot \frac{1}{\overline{\langle v, v\rangle}} \cdot \langle v, v\rangle \;=\; \langle v, w\rangle - \langle v, w\rangle \cdot \underbrace{\frac{\langle v, v\rangle}{\langle v, v\rangle}}_{=1} \;=\; 0.\end{aligned}$$

Und somit gilt $\hat{w} \perp v$. Die obigen Umformungen wären einfacher zu zeigen gewesen, wenn man die Gleichung mit $\langle \hat{w}, v\rangle$ begonnen hätte. So konnte man jedoch sehen, dass die Aussage auch mit dem komplexen Standardskalarprodukt verträglich ist. □

Der folgende Algorithmus erklärt das Gram-Schmidt-Orthogonalisierungsverfahren zur Konstruktion einer Orthonormalbasis eines endlich-dimensionalen Euklidischen oder unitären Vektorraums.

**Algorithmus 3.49** (Gram-Schmidtsches Orthogonalisierungsverfahren)
*Sei $V$ ein endlich-dimensionaler Euklidischer bzw. unitärer Vektorraum mit* $\dim V = n$, *dann lässt sich eine Orthonormalbasis $(v_1, \ldots, v_n)$ aus einer Familie von linear unabhängig Vektoren $(w_1, \ldots, w_n)$ von $V$ wie folgt konstruieren.*

***1. Schritt:***

*Wähle den ersten Vektor $w_1 \in V$ und setze*

$$\tilde{v_1} \;:=\; w_1.$$

*Normiere den ersten Vektor wie folgt:*

$$v_1 \;:=\; \frac{\tilde{v_1}}{||\tilde{v_1}||}.$$

***2. Schritt:***

*Wähle den zweiten Vektor $w_2 \in V$ und berechne die orthogonale Projektion von $w_2$ auf den normierten Vektor $v_1$ als*

$$\Pi_{v_1}(w_2) \;:=\; \langle w_2, v_1 \rangle v_1.$$

*Ziehe die orthogonale Projektion $\Pi_{v_1}(w_2)$ vom ursprünglichen Vektor $w_2$ ab und erhalte damit*

$$\tilde{v_2} \;:=\; w_2 - \Pi_{v_1}(w_2) \;=\; w_2 - \langle w_2, v_1 \rangle v_1.$$

*Normiere den zweiten Vektor wie folgt:*

$$v_2 \;:=\; \frac{\tilde{v_2}}{||\tilde{v_2}||}.$$

***i. Schritt:***

*Wähle den i-ten Vektor $w_i \in V$ und berechne die orthogonale Projektion von $w_i$ auf die normierten Vektoren $v_1, \ldots, v_{i-1}$ und ziehe diese Projektionen vom ursprünglichen Vektor $w_i$ ab durch*

$$\tilde{v_i} \;:=\; w_i - \sum_{j=1}^{i-1} \langle w_i, v_j \rangle v_j.$$

*Normiere den i-ten Vektor wie folgt:*

$$v_i \;:=\; \frac{\tilde{v_i}}{||\tilde{v_i}||}.$$

***n. Schritt:***

*Wähle den n-ten Vektor $w_n \in V$ und berechne die orthogonale Projektion von $w_n$ auf die normierten Vektoren $v_1, \ldots, v_{n-1}$ und ziehe diese Projektionen vom ursprünglichen Vektor $w_n$ ab durch*

$$\tilde{v_n} \;:=\; w_n - \sum_{j=1}^{n-1} \langle w_n, v_j \rangle v_j.$$

*Normiere den n-ten Vektor wie folgt:*

$$v_n \;:=\; \frac{\tilde{v_n}}{||\tilde{v_n}||}.$$

*Die Familie $(v_1, \ldots, v_n)$ bilden nach Konstruktion nun eine Orthonormalbasis von $V$.*

Wir wollen das Gram-Schmidtsche Orthogonalisierungsverfahren mit einem kurzen Rechenbeispiel veranschaulichen.

**Beispiel 3.50**
*Sei $V = \mathbb{R}^2$ der Euklidische Vektorraum und wir betrachten zwei linear unabhängige Vektoren $w_1, w_2 \in V$, die wir orthonormalisieren wollen mit*

$$w_1 \;:=\; \begin{pmatrix} 3 \\ 1 \end{pmatrix}, \quad w_2 \;:=\; \begin{pmatrix} 2 \\ 2 \end{pmatrix}.$$

*Wir nutzen Algorithmus 3.49 zur Konstruktion einer Orthonormalbasis aus $w_1$ und $w_2$.*

***1. Schritt:***

*Wir setzen $\tilde{v_1} = w_1$ und normieren den Vektor wie folgt:*

$$v_1 \;=\; \frac{\tilde{v_1}}{||\tilde{v_1}||} \;=\; \frac{1}{\sqrt{10}} \cdot \begin{pmatrix} 3 \\ 1 \end{pmatrix}.$$

***2. Schritt:***

*Wir berechnen die orthogonale Projektion von $w_2$ auf $v_1$ wie folgt:*

$$\Pi_{v_1}(w_2) \;=\; \langle w_2, v_1 \rangle \cdot v_1 \;=\; \frac{1}{\sqrt{10}} \cdot \langle (2,2)^T, (3,1)^T \rangle \cdot \frac{(3,1)^T}{\sqrt{10}} \;=\; \frac{8}{10} \cdot \begin{pmatrix} 3 \\ 1 \end{pmatrix}.$$

*Damit können wir nun den Vektor $w_2$ orthogonalisieren mit*

$$\tilde{v_2} \;=\; w_2 - \Pi_{v_1}(w_2) \;=\; \begin{pmatrix}2\\2\end{pmatrix} - \frac{8}{10}\cdot\begin{pmatrix}3\\1\end{pmatrix} \;=\; \frac{1}{5}\cdot\begin{pmatrix}-2\\6\end{pmatrix}.$$

*Durch Normierung erhalten wir den zweiten Vektor der Orthonormalbasis:*

$$v_2 \;=\; \frac{\tilde{v_2}}{||\tilde{v_2}||} \;=\; \sqrt{\frac{25}{40}}\cdot\frac{1}{5}\cdot\begin{pmatrix}-2\\6\end{pmatrix} \;=\; \frac{1}{\sqrt{10}}\cdot\begin{pmatrix}-1\\3\end{pmatrix}.$$

*Damit haben wir eine Orthonormalbasis von $V$ konstruiert, da offensichtlich gilt:*

$$\begin{aligned}\langle v_1, v_2\rangle &= \frac{1}{10}\cdot\langle (3,1)^T, (-1,3)^T\rangle \;=\; 0,\\ \langle v_1, v_1\rangle &= \frac{1}{10}\cdot\langle (3,1)^T, (3,1)^T\rangle \;=\; 1,\\ \langle v_2, v_2\rangle &= \frac{1}{10}\cdot\langle (-1,3)^T, (-1,3)^T\rangle \;=\; 1.\end{aligned}$$

## 3.6 Orthogonale und unitäre Endomorphismen

Mit Hilfe des in den vorangegangenen Kapiteln eingeführten Skalarprodukts ist es uns möglich eine besondere Gruppe von Endomorphismen zu untersuchen - die orthogonalen und unitären Endomorphismen. Wir werden hierbei sehen, dass diese Abbildungen schöne mathematische Eigenschaften haben und insbesondere Längen- und Winkel-erhaltend sind. Daher sind sie besonders geeignet bei der Lösung von linearen Gleichungssystemen und Eigenwertproblemen.

Orthogonale bzw. unitäre Endomorphismen lassen sich durch ihr Verhalten bezüglich des Standardskalarprodukts charakterisieren, worauf die folgende Definition basiert.

**Definition 3.51** (Orthogonale und unitäre Endomorphismen)
*Sei $V$ ein Euklidischer bzw. unitärer Vektorraum und $F\colon V \to V$ ein Endomorphismus von $V$. Dann heißt $F$* orthogonal *bzw.* unitär*, wenn*

$$\langle F(v), F(w)\rangle \;=\; \langle v, w\rangle \qquad \textit{für alle } v, w \in V\,.$$

Wir können nur mit Hilfe von Definition 3.51 bereits viele interessante Eigenschaften für orthogonale bzw. unitäre Endomorphismen ableiten, wie der folgende Satz zeigt.

**Satz 3.52**
*Sei $V$ ein Euklidischer bzw. unitärer Vektorraum und $F\colon V \to V$ ein orthogonaler bzw. unitärer Endomorphismus von $V$. Seien außerdem $v, w \in V$ beliebige Vektoren. Dann besitzt $F$ folgende Eigenschaften:*

*i)* $\|F(v)\| \;=\; \|v\|$,

*ii)* $\measuredangle(F(v), F(w)) \;=\; \measuredangle(v, w)$,

*iii)* $v \perp w \;\Leftrightarrow\; F(v) \perp F(w)$,

*iv)* $F$ *ist Isomorphismus und* $F^{-1}$ *ist auch orthogonal bzw. unitär,*

*v) Für alle Eigenwerte* $\lambda \in \mathbb{K}$ *von* $F$ *gilt* $|\lambda| = 1$.

*Beweis.* Wir beweisen die verschiedenen Eigenschaften des Endomorphismus $F$ durch Ausnutzen der Definition von orthogonalen bzw. unitären Endomorphismen.

i) Sei $v \in V$, dann gilt

$$||F(v)|| \;=\; \sqrt{\langle F(v), F(v)\rangle} \;=\; \sqrt{\langle v, v\rangle} \;=\; ||v||.$$

ii) Seien $v, w \in V$, dann können wir unter Ausnutzung von i) zeigen, dass gilt:

$$\measuredangle(F(v), F(w)) \;=\; \arccos \frac{\langle F(v), F(w)\rangle}{||F(v)|| \cdot ||F(w)||} \;=\; \arccos \frac{\langle v, w\rangle}{||v|| \cdot ||w||} \;=\; \measuredangle(v, w).$$

iii) Seien $v, w \in V$ mit $v \perp w$, dann ist die Aussage natürlich ein Spezialfall von ii). Wir können aber auch direkt nachrechnen, dass gilt:

$$\langle F(v), F(w)\rangle \;=\; \langle v, w\rangle \;=\; 0.$$

iv) Wir müssen für die Aussage zeigen, dass $F\colon V \to V$ injektiv ist und der Kern von $F$ nur die Null enthält, denn dann gilt nach dem Dimensionssatz [Fischer2005, Satz 2.2.4] schon, dass $\text{Bild}(F) = V$ gelten muss und es sich somit um einen Isomorphismus handelt. Sei also $v \in \text{Kern}(F)$, dann gilt schon:

$$\langle v, v\rangle \;=\; \langle F(v), F(v)\rangle \;=\; \langle 0, 0\rangle \;=\; 0.$$

Wegen der positiven Definitheit des Skalarprodukt muss $v = 0$ gelten und damit haben wir gezeigt, dass $F$ injektiv und somit ein Isomorphismus ist.

Um zu zeigen, dass auch $F^{-1}$ orthogonal bzw. unitär ist betrachten wir für beliebiges $v \in V$ mit $F^{-1}(v) =: w$ folgende Gleichung:

$$\langle F^{-1}(v), F^{-1}(v)\rangle \;=\; \langle w, w\rangle \;=\; \langle F(w), F(w)\rangle \;=\; \langle F \circ F^{-1}(v), F \circ F^{-1}(v)\rangle \;=\; \langle v, v\rangle.$$

Daher ist $F^{-1}$ also auch orthogonal bzw. unitär.

v) Sei $\lambda \in \mathbb{K}$ ein Eigenwert von $F$ mit zugehörigem Eigenvektor $v \in V$, so gilt wegen i):

$$||v|| \;=\; ||F(v)|| \;=\; ||\lambda v|| \;=\; |\lambda| \cdot ||v||.$$

Diese Gleichung kann nur gelten, wenn $|\lambda| = 1$ ist.

□

Wir können einen Endomorphismus $F$ von $V$ schon als orthogonal bzw. unitär erkennen, sobald er Längen-erhaltend ist, wie folgendes Lemma zeigt. Solche Abbildungen werden *Isometrien* genannt. Damit gilt auch die Umkehrung von Aussage $i$) in Satz 3.52.

**Lemma 3.53**
*Sei $F\colon V \to V$ ein Endomorphismus mit*

$$\|F(v)\| \;=\; \|v\| \quad \textit{für alle } v \in V.$$

*Dann ist $F$ orthogonal bzw. unitär.*

*Beweis.* Aus der Invarianz der Norm $||F(v)|| = ||v||$ eines Vektors $v \in V$ folgt auch schon die Invarianz der quadratischen Norm $||v||^2$ von $v$. Dies ist jedoch gerade die quadratische Form des kanonischen Skalarprodukts in $V$. Mit Hilfe der Polarisierungsformel in Satz 3.36 folgt aus der Invarianz der quadratischen Form schon die Invarianz des Skalarprodukts selbst. Dies bedeutet nach Definition 3.51, dass $F$ orthogonal bzw. unitär sein muss. □

Nutzt man in $\mathbb{R}^n$ und $\mathbb{C}^n$ mit dem kanonischen Skalarprodukt die darstellende Matrix $A$ eines orthogonalen bzw. unitären Endomorphismus $F$ von $V$, so lässt sich beobachten, dass für alle $x, y \in \mathbb{K}^n$ gilt:

$$\langle x, y\rangle \;=\; \langle Ax, Ay\rangle \;=\; (Ax)^T\overline{Ay} \;=\; x^T(A^T\bar{A})\overline{y} \;=\; x^T(\bar{A}^TA)^T\overline{y}. \tag{3.8}$$

Die Matrix $\bar{A}^T$ in Gleichung (3.8) hat eine besondere Bedeutung wie die folgende Definition zeigt.

**Definition 3.54** (Adjungierte Matrix)
*Sei $A \in \mathbb{C}^{m\times n}$ eine komplexe Matrix mit*

$$A \;=\; \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix},$$

*dann definieren wir die* adjungierte Matrix $A^* \in \mathbb{C}^{n\times m}$*, auch* Adjungierte *genannt, als die komplexe Konjugation der transponierten Matrix von $A$, d.h.,*

$$A^* \;:=\; \bar{A}^T \;=\; \overline{A^T} \;=\; \begin{pmatrix} \bar{a}_{11} & \cdots & \bar{a}_{m1} \\ \vdots & & \vdots \\ \bar{a}_{1n} & \cdots & \bar{a}_{mn} \end{pmatrix}.$$

*Ist $A$ eine reelle Matrix, so ist die Adjungierte lediglich die transponierte Matrix von $A$, d.h., $A^* = A^T$.*

Für die darstellende Matrix eines unitären Endomorphismus muss also nach Gleichung (3.8) gelten, dass $A^* \cdot A = E_n$ gilt, was die folgende Definition motiviert.

**Definition 3.55** (Orthogonale und unitäre Matrix)
*Eine Matrix $A \in \mathrm{GL}(n;\mathbb{R})$ heißt* orthogonal*, falls gilt*

$$A^{-1} \;=\; A^T.$$

*Eine Matrix $A \in \mathrm{GL}(n;\mathbb{C})$ heißt* unitär*, falls gilt*

$$A^{-1} \;=\; A^*.$$

Aus der Gleichung $A^* \cdot A = E_n$ können wir folgende interessante Beobachtungen zur Gestalt und Determinante einer unitären Matrix machen.

**Lemma 3.56**
*Sei $A$ eine unitäre Matrix. Dann bilden sowohl die Spalten als auch die Zeilen von $A$ eine Orthonormalbasis des $\mathbb{K}^n$.*

*Beweis.* Da $A$ unitär ist gilt per Definition, dass $A^{-1} = A^*$ gilt. Daraus folgt einerseits

$$\delta_{i,j} \;=\; E_n \;=\; A \cdot A^{-1} \;=\; A \cdot A^*,$$

also müssen die Zeilen von $A$ eine Orthonormalbasis des $\mathbb{K}^n$ bilden. Gleichzeitig gilt aber auch

$$\delta_{i,j} \;=\; E_n \;=\; A^{-1} \cdot A \;=\; A^* \cdot A,$$

also müssen die Spalten von $A$ eine Orthonormalbasis des $\mathbb{K}^n$ bilden. □

**Lemma 3.57**
*Sei $A$ eine unitäre Matrix. Dann gilt entweder* $\det A = 1$ *oder* $\det A = -1$.

*Beweis.* Für eine komplexe Zahl $z \in \mathbb{C}$ mit $z := a + ib$ und $a, b \in \mathbb{R}$ gilt offensichtlich:

$$z \cdot \bar{z} \;=\; (a+ib)\cdot(a-ib) \;=\; a^2 - aib + aib - i^2b^2 \;=\; a^2 + b^2 \;=\; |z|^2.$$

Daher können wir für die Determinante von $A$ folgern, dass gilt:

$$|\det A|^2 \;=\; \det A \cdot \overline{\det A} \;=\; \det A \cdot \det A^* \;=\; \det(A \cdot A^*) \;=\; \det(E_n) \;=\; 1.$$

Damit wissen wir, dass für unitäre Matrizen gilt $\det A = \pm 1$. □

Das Vorzeichen der Determinante hat eine wichtige geometrische Bedeutung, denn falls $\det A = +1$ gilt, bleiben Orientierungen unter der Wirkung von $A$ erhalten. Solch orthogonale Matrizen nennt man daher auch *eigentlich orthogonal* und sie bilden eine abgeschlossene Gruppe, wie folgende Bemerkung feststellt.

**Bemerkung 3.58**
*Die Mengen*

$$\begin{aligned}
\mathrm{O}(n) &:= \left\{A \in \mathrm{GL}(n;\mathbb{R}) \mid A^{-1} = A^T\right\}, && \text{(orthogonale Gruppe)}\\
\mathrm{SO}(n) &:= \{A \in \mathrm{O}(n) \mid \det A = 1\}, && \text{(spezielle orthogonale Gruppe)}\\
\mathrm{U}(n) &:= \left\{A \in \mathrm{GL}(n;\mathbb{C}) \mid A^{-1} = \bar{A}^T\right\}, && \text{(unitäre Gruppe)}\\
\mathrm{SU}(n) &:= \{A \in \mathrm{U}(n) \mid \det A = 1\}, && \text{(spezielle unitäre Gruppe)}
\end{aligned}$$

*der* orthogonalen, speziellen orthogonalen, unitären *und* speziellen unitären *Matrizen sind Untergruppen von* $\mathrm{GL}(n;\mathbb{R})$ *bzw.* $\mathrm{GL}(n;\mathbb{C})$.

Wir diskutieren im Folgenden zwei Beispiele von orthogonalen bzw. unitären Matrizen.

**Beispiel 3.59**
*Sei* $A \in \mathbb{K}^{3\times 3}$, *dann betrachten wir zwei einfache Beispiele.*

1. *Die Einheitsmatrix* $A = E_3 \in \mathbb{R}^{3\times 3}$ *mit*

$$A := \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

   *ist eine orthogonale Matrix, da offensichtlich gilt* $A^T = A = A^{-1}$. *Es gilt sogar* $A \in \mathrm{SO}(3)$, *da* $\det A = +1$ *ist.*

2. *Die Matrix* $A \in \mathbb{C}^{3\times 3}$ *mit*

$$A := \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix}$$

   *ist eine unitäre Matrix, denn offensichtlich gilt* $A^* = A^{-1}$ *mit*

$$A \cdot A^* = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & \bar{i} \\ 0 & -\bar{i} & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

   *Es gilt sogar* $A \in \mathrm{SU}(3)$, *da* $\det A = +1$ *ist.*

Der folgende Satz stellt eine direkte Beziehung zwischen orthogonalen bzw. unitären Endomorphismen und den Eigenschaften ihrer darstellenden Matrizen bezüglich bestimmter Basen fest.

**Satz 3.60**
*Sei* $V$ *ein Euklidischer bzw. unitärer Vektorraum mit einer Orthonormalbasis* $B$ *und* $F$ *ein Endomorphismus von* $V$. *Dann gilt:*

$$F \text{ ist orthogonal (bzw. unitär)} \Leftrightarrow M_B(F) \text{ ist orthogonal (bzw. unitär)}.$$

*Beweis.* Wir zeigen die Aussage für den allgemeineren Fall eines unitären Vektorraums. Sei $A := M_B(F) \in \mathbb{K}^{n\times n}$ die darstellende Matrix von $F$ bezüglich der Orthonormalbasis $B$ und für $v, w \in V$ seien $x, y \in \mathbb{K}^n$ die zugehörigen Koordinaten, d.h.

$$v \;=\; \Phi_B(x), \quad w \;=\; \Phi_B(y).$$

Da $B$ eine Orthonormalbasis ist gilt offensichtlich

$$\langle v, w\rangle \;=\; \langle \Phi_B(x), \Phi_B(y)\rangle \;=\; \langle x, y\rangle.$$

Der Endomorphismus $F$ ist also genau dann unitär, falls gilt

$$\langle F(v), F(w)\rangle \;=\; \langle Ax, Ay\rangle \;=\; x^T A^T \bar{A}\bar{y} \;=\; x^T \bar{y} \;=\; \langle x, y\rangle \;=\; \langle v, w\rangle.$$

Also genau dann wenn $A^T \cdot \bar{A} = I_n$ gilt, was der Fall ist, wenn $A$ unitär ist. □

Wir wollen im folgenden untersuchen wie die Normalform eines orthogonalen bzw. unitären Endomorphismus aussieht. Hierbei werden wir die mathematischen Werkzeuge aus der Eigenwerttheorie in Kapitel 2 einsetzen können. Hierzu beginnen wir mit der (mathematisch schöneren) **Normalform von unitären Endomorphismen**, die im folgenden Satz beschrieben ist.

**Satz 3.61** (Diagonalisierungssatz)
*Jeder unitäre Endomorphismus $F$ eines unitären Vektorraums $V$ besitzt eine Orthonormalbasis aus Eigenvektoren von $F$. Insbesondere ist er diagonalisierbar.*

*Beweis.* Wir führen den Beweis mittels vollständiger Induktion über $n = \dim V$.

**Induktionsanfang:** $n = 1$
Da $F$ unitär ist muss gelten

$$\langle F(v), F(w)\rangle = \langle v, w\rangle, \qquad \text{für alle } v, w \in V.$$

Im eindimensionalen Fall, kann dies nur gelten, falls $F(v) = v$ oder $F(v) = -v$ für alle $v \in V$ gilt. Dies erfüllt aber schon die Eigenwertgleichung für den Eigenwert $\lambda_1 = 1$ oder $\lambda_1 = -1$. Für beide Fälle ist $v_1 = 1$ der zugehörige Eigenvektor und es ist klar, dass $v_1$ eine Orthonormalbasis von $V$ bildet. Damit ist $F$ nach Definition 2.25 diagonalisierbar.

**Induktionsschritt:** $n - 1 \to n$
Die Induktionsannahme ist, dass die Aussage bereits für den Fall $n - 1$ gezeigt wurde. Wir betrachten für $n \geqslant 1$ das charakteristische Polynom $P_F$ von $F$, welches nach dem Fundamentalssatz der Algebra in Linearfaktoren über dem Körper $\mathbb{C}$ zerfällt in

$$P_F(t) \;=\; \pm(t - \lambda_1) \cdot \ldots \cdot (t - \lambda_n) \qquad \text{mit} \quad \lambda_1, \ldots, \lambda_n \in \mathbb{C}.$$

Zum Eigenwert $\lambda_1 \in \mathbb{C}$ von $F$ wählen wir einen zugehörigen Eigenvektor $v_1$ mit $\|v_1\| = 1$. Wir betrachten das orthogonale Komplement zum Unterraum $\mathrm{lin}(\{v_1\})$, d.h.

$$W := \{w \in V \mid v_1 \perp w\} \; .$$

Wir müssen nun $F(W) = W$ zeigen, d.h., dass $W$ ein $F$-invarianter Unterraum ist. Da $F$ nach Satz 3.52 ein Isomorphismus ist, genügt es $F(W) \subset W$ zu beweisen. Für alle $w \in W$ folgt aus der Gleichung

$$\lambda_1 \langle v_1, F(w) \rangle \;=\; \langle \lambda_1 v_1, F(w) \rangle \;=\; \langle F(v_1), F(w) \rangle \;=\; \langle v_1, w \rangle \;=\; 0 \, ,$$

dass $\langle F(w), v_1 \rangle = 0$, da $|\lambda_1| = 1$ und damit insbesondere $\lambda_1 \neq 0$ gilt. Das zeigt jedoch schon, dass $F(W) \subset W$ gilt.

Nun betrachten wir den Endomorphismus $G := F|_W$ von $W$. Als Einschränkung von $F$ ist $G$ weiterhin unitär und wegen $\dim W = n-1$ können wir auf $G$ die Induktionsannahme anwenden. Danach gibt es eine Orthonormalbasis $(v_2, \ldots, v_n)$ von $W$ bestehend aus Eigenvektoren von $G$ und somit auch von $F$. Die um den Eigenvektor $v_1 \in V$ ergänzte Basis $(v_1, v_2, \ldots, v_n)$ ist orthonormal und besteht aus Eigenvektoren von $F$. □

Übertragen auf unitäre Matrizen liefert uns der Diagonalisierungssatz 3.61 folgende Ähnlichkeitstransformation.

**Korollar 3.62**
*Zu $A \in \mathrm{U}(n)$ gibt es ein $S \in \mathrm{U}(n)$ mit*

$$S \cdot A \cdot S^* \;=\; \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix},$$

*wobei $\lambda_i \in \mathbb{C}$ mit $|\lambda_i| = 1$ für $i = 1, \ldots, n$.*

*Beweis.* Als Spalten von $S^* = S^{-1}$ verwendet man eine orthonormale Basis des $\mathbb{C}^n$, die aus Eigenvektoren von $A$ besteht. □

Im folgenden Beispiel wollen wir die Diagonalisierbarkeit einer unitären Matrix nach Korollar 3.62 prüfen.

**Beispiel 3.63**
*Die Matrix $A \in \mathbb{R}^{2\times 2}$ mit*

$$A := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

*beschreibt eine Drehung um* $90^\circ$ *im* $\mathbb{R}^2$ *und ist offensichtlich orthogonal. Da das charakteristische Polynom von* $A$ *die Form* $P_A(t) = t^2 + 1$ *hat, sind die Eigenwerte* $\lambda_1 = i$ *und* $\lambda_2 = -i$ *nicht reell. Die zugehörigen Eigenvektoren in* $\mathbb{C}^2$ *sind:*

$$v_1 = \begin{pmatrix} i \\ 1 \end{pmatrix} \qquad v_2 = \begin{pmatrix} -i \\ 1 \end{pmatrix} \qquad \textit{mit} \quad \langle v_1, v_2 \rangle = v_1^T \bar{v_2} = 0 \quad \textit{und} \quad \|v_i\| = \sqrt{2}.$$

*Durch Normalisierung der Eigenvektoren mit dem Faktor* $\sqrt{2}$ *erhalten wir*

$$S^* \;=\; S^{-1} \;:=\; \frac{1}{\sqrt{2}} \begin{pmatrix} i & -i \\ 1 & 1 \end{pmatrix}, \quad S \;:=\; \frac{1}{\sqrt{2}} \begin{pmatrix} -i & 1 \\ i & 1 \end{pmatrix}$$

*und wir erhalten damit*

$$SAS^* \;=\; \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}.$$

Um den komplizierteren Fall von **orthogonalen Endomorphismen** besser zu verstehen beginnen wir mit einer Falldiskussion für kleine Dimensionen $n = 1, 2, 3$ des Vektorraums $V$. Sei also im Folgenden $F \colon V \to V$ ein orthogonaler Endomorphismus und $A := M_B(F)$ die darstellende Matrix von $F$ bezüglich einer Basis $B$. Aus Satz 3.60 wissen wir, dass $A$ dann auch orthogonal ist und wir können unsere Diskussion auf diese Matrix beschränken.

Im *eindimensionalen Fall* kann wegen $A^{-1} = A^T$ nur gelten $A = \pm 1$.

Im *zweidimensionalen Fall* erhalten wir im folgenden Lemma eine interessante geometrische Interpretation von orthogonalen Matrizen, die besagt, dass orthogonale $(2 \times 2)$-Matrizen spezielle geometrische Transformationen realisieren.

**Lemma 3.64**
*Ist* $A \in \mathrm{O}(2)$*, so gibt es ein* $\alpha \in [0, 2\pi[$*, so dass*

$$A = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \quad \textit{oder} \quad A = \begin{pmatrix} \cos\alpha & \sin\alpha \\ \sin\alpha & -\cos\alpha \end{pmatrix}.$$

*Beweis.* In der Hausaufgabe zu zeigen. □

Im *dreidimensionalen Fall* eines orthogonalen Endomorphismus $F \colon \mathbb{R}^3 \to \mathbb{R}^3$ betrachten wir das charakteristische Polynom $P_F$. Da $P_F$ den Grad 3 besitzt, existiert nach dem Zwischenwertsatz der Analysis mindestens eine reelle Nullstelle von $P_F$. Also hat $F$ mindestens einen Eigenwert $\lambda_1 \in \mathbb{R}$ und nach Satz 3.52 wissen wir, dass $\lambda_1 = \pm 1$ gilt. Sei $v_1 \in \mathbb{R}^3$ der zugehörige Eigenvektor, für den wir $\|v_1\| = 1$ annehmen können (durch Normalisierung). Dann können diesen Eigenvektor von $F$ nach Satz 3.44 zu einer Orthonormalbasis $B = (v_1, w_2, w_3)$ von $V$ ergänzen.

Wir bezeichnen mit $W \subset V = \mathbb{R}^3$ die von $w_2$ und $w_3$ aufgespannte zweidimensionale Ebene. Da $v_1$ ein Eigenvektor von $F$ ist, gilt natürlich $F(v_1) \subset \mathrm{lin}(\{v_1\})$. Da außerdem

$v_1 \perp W$ und $F$ nach Satz 3.52 ein Isomorphismus ist muss gelten, dass $F(W) = W$ gilt. Betrachten wir die darstellende Matrix $M_B(F)$ von $F$ bezüglich der Orthonormalbasis $B$ mit

$$M_B(F) \;=\; \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \boxed{A'} & \\ 0 & & \end{pmatrix} =:\; A,$$

so folgt aus Satz 3.60, dass $A' \in \mathrm{O}(2)$ orthogonal ist. Weiter gilt wegen der Determinantenregel für Blockmatrizen in Lemma 2.7, dass gilt $\det A = \lambda_1 \cdot \det A'$. Betrachten wir also die möglichen Fälle im Folgenden basierend auf unseren Erkenntnissen aus Lemma 3.64.

1. Sei $\det A = +1$. Ist $\lambda_1 = -1$, dann muss $\det A' = -1$ sein. Daher kann man $w_2$ und $w_3$ als Eigenvektoren zu den Eigenwerten $\lambda_2 = +1$ und $\lambda_3 = -1$ wählen, also

$$A = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

   Ist $\lambda_1 = +1$, dann muss auch $\det A' = +1$ sein, also gibt es ein $\alpha \in [0, 2\pi[$, so dass

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{pmatrix}.$$

2. Ist $\det A = -1$, dann gibt es bei geeigneter Wahl von $w_2$ und $w_3$ die Möglichkeiten

$$A \;=\; \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \qquad \text{und} \qquad A \;=\; \begin{pmatrix} -1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{pmatrix}.$$

Wir nutzen unsere Erkenntnisse aus der gerade durchgeführten Diskussion für die Fälle $\dim V \in \{1, 2, 3\}$ um eine *Normalform für unitäre und orthogonale Endomorphismen* im allgemeinen Fall anzugeben.

Im Gegensatz zu der gerade im komplexen Fall bewiesenen Diagonalisierbarkeit unitärer Endomorphismen gibt es im Reellen orthogonale Endomorphismen ohne Eigenwerte in $\mathbb{R}$. Das einfachste Beispiel sind Drehungen in der Ebene $\mathbb{R}^2$ wie in Beispiel 3.63 beschrieben.

Bevor wir uns der Normalform von orthogonalen Endomorphismen widmen benötigen wir das folgende hilfreiche Lemma.

**Lemma 3.65**
*Zu einem orthogonalen Endomorphismus $F\colon V \to V$ eines Euklidischen Vektorraums $V$ mit $\dim V \geqslant 1$ gibt es stets einen Untervektorraum $W \subset V$ mit*

$$F(W) \subset W \quad \textit{und} \quad 1 \leqslant \dim W \leqslant 2\,.$$

*Beweis.* Sei $A \in \mathbb{R}^{n\times n}$ die darstellende Matrix des orthogonalen Endomorphismus $F$ auf $V$. Wir führen den Beweis des Lemmas durch eine Symmetrisierung von $A$. Sei hierfür die Symmetrisierung $A^s$ von $A$ definiert als

$$A^s \;:=\; A + A^T \;=\; A + A^{-1}.$$

Offensichtlich ist $A^s$ symmetrisch, da für alle Indizes $1 \leqslant i, j \leqslant n$ gilt

$$A^s_{i,j} \;=\; A_{i,j} + A^T_{i,j} \;=\; A_{i,j} + A_{j,i} \;=\; A^T_{j,i} + A_{j,i} \;=\; A^s_{j,i}.$$

Wie wir im nächsten Kapitel sehen werden, bedeutet dies, dass ein reeller Eigenwert $\lambda \in \mathbb{R}$ und ein zugehöriger reeller Eigenvektor $v$ von $A^s$, so dass die Eigenwertgleichung $A^s v = \lambda v$ erfüllt ist.

Dieser Eigenvektor erzeugt den $F$-invarianten Unterraum $W$ mit

$$W := \operatorname{span}(v, Av).$$

Die Invarianz von $W$ unter Anwendung von $A$ folgt aus den folgenden Argumenten. Wenden wir $A$ auf $v$ an, so erhalten wir den Vektor $Av \in W$. Außerdem gilt, dass $A \cdot Av = A^2 v \in W$, da wir aus der Eigenwertgleichung sehen:

$$Av + A^{-1}v \;=\; A^s v \;=\; \lambda v \quad \Rightarrow \quad A^2 v \;=\; -v + \lambda Av \;\in W.$$

□

Wir zeigen nun, dass sich die Normalform eines orthogonalen Endomorphismus auf eine darstellende Matrix zurückführen lässt, die im Wesentlichen nur die Eigenwerte $\pm 1$ auf der Hauptdiagonalen und eben jene $(2 \times 2)$-Drehmatrizen besitzt.

**Satz 3.66**
*Ist $F$ ein orthogonaler Endomorphismus eines Euklidischen Vektorraums $V$, dann gibt es in $V$ eine Orthonormalbasis $B$ derart, dass die darstellende Matrix $M_B(F) \in \mathrm{GL}(n;\mathbb{R})$ folgende Gestalt einer* Normalform *hat*

$$M_B(F) = \begin{pmatrix} +1 & & & & & & & \\ & \ddots & & & & & & \\ & & +1 & & & 0 & & \\ & & & -1 & & & & \\ & & & & \ddots & & & \\ & & & & & -1 & & \\ & & 0 & & & & A_1 & \\ & & & & & & & \ddots \\ & & & & & & & & A_k \end{pmatrix},$$

*wobei für $j = 1, \ldots, k$ folgende Drehmatrizen existieren:*

$$A_j = \begin{pmatrix} \cos\alpha_j & -\sin\alpha_j \\ \sin\alpha_j & \cos\alpha_j \end{pmatrix} \in \mathrm{SO}(2) \qquad \textit{mit} \quad \alpha_j \in [0, 2\pi[, \ \textit{jedoch } \alpha_j \notin \{0, \pi\}\,.$$

*$F$ ist also charakterisiert durch die Anzahl $r \in \mathbb{N}$ der Eigenwerte $+1$ die Anzahl $s \in \mathbb{N}$ der Eigenwerte $-1$, sowie durch die Winkel $\alpha_1, \ldots, \alpha_k$, wobei gilt*

$$r + s + 2k \;=\; \dim V.$$

*Beweis.* Sei $A$ eine darstellende Matrix von $F$. Wir führen den Beweis durch vollständige Induktion über $n = \dim V$.

**Induktionsanfang:** $n = 1$ **und** $n = 2$
Der Induktionsanfang folgt direkt aus der Beobachtung, dass $A = \pm 1$ für $n = 1$ gilt und der Aussage von Lemma 3.64 für $n = 2$.

**Induktionsschritt:** $n - 1 \to n$ **für** $n \geqslant 3$
Die Induktionsannahme ist, dass die Aussage bereits für den Fall $n - 1$ und $n - 2$ gezeigt wurde.

Da $F$ orthogonal ist existiert nach Lemma 3.65 ein Untervektorraum $W \subset V$ mit

$$1 \leqslant \dim W \leqslant 2 \qquad \text{und} \qquad F(W) \subset W.$$

Aus Satz 3.52 wissen wir, dass $F$ injektiv und somit gilt schon, dass $F(W) = W$ sein muss. Außerdem existiert existiert ein Endomorphismus $F^{-1}$, der auch orthogonal ist und für den gilt $F^{-1}(W) = W$. Daher können wir für die Vektoren $w \in W$ und $v \in W^\perp$ folgern, dass gilt

$$\langle F(v), w\rangle \;=\; \langle F^{-1}(F(v)), F^{-1}(w)\rangle \;=\; \langle v, F^{-1}(w)\rangle \;=\; 0.$$

Daraus folgt schon, dass $F(W^\perp) \subset W^\perp$ gilt. Da $F$ nach Satz 3.52 insbesondere ein Isomorphismus ist, muss schon gelten, dass $F(W^\perp) = W^\perp$. Damit haben wir $F$ zerlegt in zwei orthogonale Abbildungen

$$G \;:=\; F|_W \colon W \to W \qquad \text{und} \qquad H \;:=\; F|_{W^\perp} \colon W^\perp \to W^\perp\,.$$

Da $n - 2 \leqslant \dim W^\perp < n$ ist können wir können wir auf $H$ die Induktionvorrausetzung anwenden und erhalten eine Basis $B'$ von $W^\perp$ der gewünschten Art.

Für den orthogonalen Endomorphismus $G$ müssen wir abschließend noch zwei Fälle in Abhängigkeit von $\dim W \in \{1, 2\}$ betrachten.

1. Ist $\dim W = 1$, so gibt es einen Eigenvektor $v \in W$ mit $\|v\| = 1$ zu einem Eigenwert $\pm 1$. Ergänzt man $B'$ an passender Stelle durch $v$ zu $B$, so hat diese Basis von $V$ die gewünschten Eigenschaften.

2. Im Fall $\dim W = 2$ gibt es eine Orthonormalbasis $(v_1, v_2)$ von $W$, bezüglich der $G$ nach Lemma 3.64 beschrieben wird durch eine Matrix der Form

$$\begin{pmatrix} \pm 1 & 0 \\ 0 & \pm 1 \end{pmatrix} \quad \text{oder} \quad \begin{pmatrix} \cos\alpha_j & -\sin\alpha_j \\ \sin\alpha_j & \cos\alpha_j \end{pmatrix} \qquad \text{mit} \quad \alpha_j \neq 0, \alpha_j \neq \pi.$$

Indem man $v_1$ und $v_2$ an den passenden Stellen in $B'$ einfügt, erhält man wieder einen gewünschte Basis $B$ von $V$.

Wie der Beweis zeigt lässt sich $V$ rekursiv in eine orthogonale direkte Summe von invarianten Unterräume der Dimension 1 oder 2 zerlegen. □

## 3.7 Selbstadjungierte Endomorphismen

Dieser Abschnitt widmet sich hauptsächlich den Eigenschaften und der Herleitung einer Normalform von symmetrischen bzw. hermetischen Matrizen. Wir werden insbesondere zeigen, dass diese Matrizen immer diagonalisierbar sind, was sie besonders attraktiv für numerische Verfahren macht. Da diese Familie von Matrizen auch in der Theorie sehr häufig auftaucht haben sie zahlreiche Anwendungen in Geometrie, Analysis und in der Physik.

Wir beginnen mit der Definition von adjungierten und selbstadjungierten Endomorphismen, die eine zentrale Rolle in diesem Abschnitt spielen werden.

**Definition 3.67** ((Selbst-)Adjungierter Endomorphismus)
*Wir nennen eine Abbildung $F^*\colon V \to V$ den* adjungierten Endomorphismus *von $F$, falls die folgende Beziehung bezüglich des Standardskalarprodukts gilt:*

$$\langle F(v), w\rangle \;=\; \langle v, F^*(w)\rangle \qquad \textit{für alle } v, w \in V. \tag{3.9}$$

*Ein Endomorphismus $F$ eines Euklidischen bzw. unitären Vektorraums $V$ heißt* selbstadjungiert*, wenn $F = F^*$, d.h.*

$$\langle F(v), w\rangle \;=\; \langle v, F(w)\rangle \qquad \textit{für alle } v, w \in V.$$

Für orthogonale bzw. unitäre Matrizen haben wir bereits in Kapitel 3.6 gesehen, dass $A^* = A^{-1}$ gilt. Diese Beobachtung überträgt sich mit folgendem Lemma auch auf adjungierte Endomorphismen.

**Lemma 3.68**
*Ist $F$ orthogonal bzw. unitär, so gilt $F^* = F^{-1}$.*

*Beweis.* Seien $u, v \in V$ mit $w = F(u)$, also $u = F^{-1}(w)$, so folgt

$$\langle F(v), w\rangle \;=\; \langle F(v), F(u)\rangle \;=\; \langle v, u\rangle \;=\; \langle v, F^{-1}(w)\rangle.$$

□

Folgender Satz garantiert uns die Existenz eines adjungierten Endomorphismus und stellt eine Beziehung zur Adjungierten der darstellenden Matrix her.

**Satz 3.69**
*Sei $V$ ein endlich-dimensionaler Euklidischer bzw. unitärer $\mathbb{K}$-Vektorraum und $F\colon V \to V$ ein Endomorphismus. Dann existiert genau ein adjungierter Endomorphismus $F^*$ von $F$. Ist $B$ eine Orthonormalbasis von $V$ und $A := M_B(F)$ die darstellende Matrix von $F$ bezüglich $B$, so gilt*

$$M_B(F^*) \;=\; A^*.$$

*Beweis.* Da $B$ orthonormal ist, gilt für $v = \Phi_B(x)$ und $w = \Phi_B(y)$ mit $x, y \in \mathbb{K}^n$, so dass

$$\langle v, w\rangle \;=\; x^T E_n \bar{y} \;=\; x^T \bar{y}.$$

Die Bedingung (3.9) bedeutet also

$$\langle F(v), w\rangle \;=\; (Ax)^T \bar{y} \;=\; x^T A^T \bar{y} \;\overset{!}{=}\; x^T \bar{C} \bar{y} \;=\; \langle v, F^*(w)\rangle \qquad \text{für alle } x, y \in \mathbb{K}^n\,.$$

Daraus folgt nun, dass $A^T = \bar{C}$ gelten muss. Also ist $F^*$ durch die darstellende Matrix $C \coloneqq \bar{A}^T$ eindeutig definiert. □

Aus dem obigen Satz lässt sich folgende interessante Beziehung zwischen selbstadjungierten Endomorphismen und symmetrischen bzw. hermiteschen Matrizen direkt ableiten.

**Korollar 3.70**
*Ist $F\colon V \to V$ ein Endomorphismus und $B$ eine Orthonormalbasis von $V$, so gilt*

$$F \text{ selbstadjungiert} \;\Leftrightarrow\; A \coloneqq M_B(F) \text{ ist symmetrisch bzw. hermetisch, d.h. } A = A^*.$$

*Beweis.* Mit den gleichen Argumenten wie im Beweis von Satz 3.69 können wir feststellen, dass $F$ selbstadjungiert ist genau dann wenn gilt:

$$\langle F(v), w\rangle \;=\; (Ax)^T \bar{y} \;=\; x^T A^T \bar{y} \;\overset{!}{=}\; x^T \bar{A} \bar{y} \;=\; \langle v, F(w)\rangle \qquad \text{für alle } x, y \in \mathbb{K}^n\,.$$

Das ist offensichtlich genau dann der Fall, wenn $A^T = \bar{A}$ gilt oder äquivalent, wenn gilt $A = \bar{A}^T = A^*$. □

Selbstadjungierte Endomorphismen haben die schöne Eigenschaft, dass sie nur reelle Eigenwerte besitzen, wie uns folgendes Lemma zeigt.

**Lemma 3.71**
*Sei $V$ ein Euklidischer bzw. unitärer $\mathbb{K}$-Vektorraum und $F\colon V \to V$ selbstadjungiert. Dann hat das charakteristische Polynom $P_F$ auch für $\mathbb{K} = \mathbb{C}$ nur reelle Nullstellen und zerfällt in Linearfaktoren über $\mathbb{R}$. Insbesondere sind alle Eigenwerte von $F$ reell.*

*Beweis.* Sei $\lambda \in \mathbb{K}$ ein Eigenwert von $F$ mit zugehörigem Eigenvektor $v \in V$, dann gilt $F(v) = \lambda v$. Wegen der Selbstadjungiertheit von $F$ können wir dann folgern:

$$\lambda\langle v, v\rangle \;=\; \langle \lambda v, v\rangle \;=\; \langle F(v), v\rangle \;=\; \langle v, F(v)\rangle \;=\; \langle v, \lambda v\rangle \;=\; \bar{\lambda}\langle v, v\rangle.$$

Da $v \neq 0$ ist als Eigenvektor folgt also schon, dass $\bar{\lambda} = \lambda$ gilt und somit sind alle Eigenwerte von $F$ reell.

Da das charakteristische Polynom $P_F$ von $F$ nach dem Fundamentalsatz der Algebra in Linearfaktoren über $\mathbb{C}$ zerfällt und die Nullstellen die Eigenwerte von $F$ darstellen, folgt die Aussage. □

Wenn wir die Aussagen aus Korollar 3.70 und Lemma 3.71 zusammen nehmen, können wir folgern, dass alle symmetrischen und hermiteschen Matrizen nur reelle Eigenwerte besitzen. Wir wollen diese Eigenschaft in folgendem Beispiel illustrieren.

**Beispiel 3.72**
*Wir untersuchen die Eigenwerte zweier hermitescher Matrizen $A \in \mathbb{K}^{2\times 2}$ im Folgenden.*

1. *Sei $A \in \mathbb{C}^{2\times 2}$ mit*

$$A \;:=\; \begin{pmatrix} 1 & i \\ -i & 2 \end{pmatrix}.$$

   *Dann ist das charakteristische Polynom gegeben durch $P_A(t) = t^2 - 3t + 1$ dessen Nullstellen gegeben sind durch:*

$$\lambda_{1/2} \;=\; \frac{1}{2}(3 \pm \sqrt{5}) \in \mathbb{R}.$$

2. *Sei $A \in \mathbb{R}^{2\times 2}$ mit*

$$A \;:=\; \begin{pmatrix} a & b \\ b & c \end{pmatrix}.$$

   *Dann ist das charakteristische Polynom gegeben durch*

$$P_A(t) \;=\; t^2 - (a+c)t + (ac - b^2).$$

   *Die Diskriminante von $P_A$ ist in diesem Fall $(a-c)^2 + 4b^2 \geqslant 0$, also sind die Eigenwerte reell.*

Die schönen mathematischen Eigenschaften von selbstadjungierten Endomorphismen führen dazu, dass ihre Normalform einer Diagonalmatrix entspricht wie uns folgender Satz erklärt.

**Satz 3.73** (Diagonalisierungssatz)
*Ist $F$ ein selbstadjungierter Endomorphismus auf einem Euklidischen bzw. unitären Vektorraum $V$, so gibt es eine Orthonormalbasis von $V$, die aus Eigenvektoren zu reellen Eigenwerten von $F$ besteht.*

*Beweis.* Wir führen den Beweis mittels vollständiger Induktion über $n = \dim V$.

**Induktionsanfang:** $n = 1$
Da $F$ selbstadjungiert ist, ist der einzige Eigenwert von $F$ reell. Sei $\lambda_1 \in \mathbb{R}$ der Eigenwert von $F$ und $v_1 \in V$ der zugehörige Eigenvektor mit $||v_1|| = 1$ (durch Normalisierung). Dann bildet $B = (v_1)$ eine Orthonormalbasis von $V$.

**Induktionsschritt:** $n - 1 \to n$
Die Induktionsannahme ist, dass die Aussage bereits für den Fall $n - 1$ gezeigt wurde. Nach Lemma 3.71 zerfällt das charakteristische Polynom $P_F$ von $F$ in Linearfaktoren über $\mathbb{R}$ mit:

$$P_F = \pm(t - \lambda_1) \cdot \ldots \cdot (t - \lambda_n) \qquad \text{mit} \qquad \lambda_1, \ldots, \lambda_n \in \mathbb{R}.$$

Wir wählen den Eigenwert $\lambda_1$ von $F$ und den zugehörigen Eigenvektor $v_1$ mit $\|v_1\| = 1$ (durch Normalisierung) und definieren das orthogonale Komplement zu $v$ als

$$W \;:=\; \{w \in V \mid v_1 \perp w\}\,.$$

Wir erkennen wieder, dass $W$ invariant ist unter $F$, d.h. $F(W) \subset W$, da für $w \in W$ wegen der Selbstadjungiertheit gilt:

$$\langle v_1, F(w)\rangle \;=\; \langle F(v_1), w\rangle \;=\; \langle \lambda_1 v_1, w\rangle \;=\; \lambda_1\langle v_1, w\rangle \;=\; 0\,.$$

Außerdem ist $F|_W \colon W \to W$ als Einschränkung von $F$ wieder selbstadjungiert. Da

$$P_{F|_W} \;=\; \pm(t - \lambda_2) \cdot \ldots \cdot \ldots \cdot (t - \lambda_n)$$

gilt können wir die Induktionsvorraussetzung anwenden, so dass eine Orthonormalbasis $B'$ von $W$ existiert aus Eigenvektoren von $F|_W$. Diese können wir nun zu der gewünschten Orthonormalbasis $B = (v_1, \ldots, v_n)$ von $V$ aus Eigenvektoren von $F$ ergänzen. □

Wegen der Diagonalisierbarkeit eines selbstadjungierten Endomorphismus lassen sich direkt zwei Beobachtungen ableiten.

**Korollar 3.74**
*Ist $A \in \mathbb{K}^{n\times n}$ eine symmetrische bzw. hermitesche Matrix, so gibt es eine orthogonale bzw. unitäre Matrix $S \in U(n)$, so dass*

$$S \cdot A \cdot S^* \;=\; \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} \qquad \textit{mit} \quad \lambda_1, \ldots, \lambda_n \in \mathbb{R}.$$

*Beweis.* Als Spalten von $S^* = S^{-1}$ verwendet man eine orthonormale Basis des $\mathbb{C}^n$, die aus Eigenvektoren von $A$ besteht. □

Außerdem lässt sich aus dem Satz 2.48 über die Hauptraumzerlegung folgende Zerlegung in orthogonale $F$-invariante Unterräume folgern.

**Korollar 3.75**
*Sind $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$ die paarweise verschiedenen Eigenwerte eines selbstadjungierten oder unitären Endomorphismus $F$ von $V$, so ist*

$$V \;=\; \mathrm{Eig}(F;\lambda_1) \oplus \cdots \oplus \mathrm{Eig}(F;\lambda_k),$$

*wobei $\oplus$ die orthogonale Summe bezeichnet.*

Für symmetrische bzw. hermitesche Matrizen können wir die positive Definitheit mittels des folgenden Satzes am Vorzeichen der Eigenwerte ablesen.

**Satz 3.76**
*Für eine symmetrische bzw. hermitesche Matrix $A \in \mathbb{K}^{n\times n}$ sind die folgenden Bedingungen äquivalent:*

*i) $A$ ist positiv definit.*

*ii) Alle Eigenwerte $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ von $A$ sind positiv.*

*Beweis.* Für führen den Beweis für den allgemeineren Fall von unitären Vektorräumen, da er den Euklidischen Fall mit abdeckt.

*i) $\Rightarrow$ ii)*: Sei $\lambda \in \mathbb{R}$ ein beliebiger Eigenwert von $A$ und $v \in \mathbb{C}^n$ der zugehörige Eigenvektor, so ist wegen der Eigenwertgleichung

$$Av \;=\; \lambda v \quad\Rightarrow\quad \bar{A}\bar{v} \;=\; \lambda\bar{v}\,.$$

Somit ist $\bar{v} \in \mathbb{C}^n$ offensichtlich ein Eigenvektor von $\bar{A}$ zum Eigenwert $\lambda$. Mit $w := \bar{v} \neq 0$ folgt wegen der Hermizität von $A$

$$0 \;<\; w^T A\bar{w} \;=\; (A^T w)^T\bar{w} \;=\; (\bar{A}w)^T\bar{w} \;=\; (\lambda w)^T\bar{w} \;=\; \lambda\cdot w^T\bar{w}\,.$$

und aus der positiven Definitheit des Skalarprodukts, d.h., $w^T\bar{w} > 0$, folgt $\lambda > 0$.

*ii) $\Rightarrow$ i)*: Wir wählen eine Orthonormalbasis $(w_1, \ldots, w_n)$ bestehend aus Eigenvektoren von $\bar{A}$, so dass $\bar{A}w_i = \lambda_i w_i$ für alle $1 \leqslant i \leqslant n$. Ähnlich wie im Beweis der ersten Folgerung erhält man

$$w_i^T A\bar{w}_j \;=\; \lambda_i\cdot w_i^T\bar{w}_j \;=\; \begin{cases}\lambda_i, & \text{für } i = j\,,\\ 0, & \text{für } i \neq j\,.\end{cases}$$

Jeder beliebige Vektor $v \in \mathbb{C}^n$ hat nun eine Darstellung aus Eigenvektoren mit $v = \sum_{i=1}^{n}\mu_i w_i$, also ist

$$v^T Av \;=\; \sum_{i=1}^{n}\lambda_i\mu_i\bar{\mu}_i \;>\; 0 \quad \text{für } v \neq 0\,.$$

□

Abschließend wollen wir die Behauptung aus Satz 3.76 an einem konkreten Beispiel nachvollziehen.

**Beispiel 3.77**
*Für $a, b \in \mathbb{R}$ und die hermetische Matrix*

$$A = \begin{pmatrix} a & \mathrm{i} \\ -\mathrm{i} & b \end{pmatrix} \in \mathbb{C}^{2\times 2}$$

*ist*

$$P_A(t) \;=\; t^2 - (a+b)t + (ab-1)\,.$$

*Also ist $A$ genau dann positiv definit, wenn $a+b>0$ und $ab>1$ ist. In diesem Fall kann man die Eigenwerte einfach ausrechnen:*

$$\lambda_{1,2} \;=\; \frac{1}{2}\left(a+b \pm \sqrt{(a-b)^2+4}\right)\,.$$

*Sie sind beide positiv, wenn $a+b>0$ und $ab>1$ ist, wie man leicht nachprüfen kann.*

## 3.8 Normale Endomorphismen

Zum Schluss betrachten wir noch eine weitere besondere Gruppe von Abbildungen. Die sogenannten *normalen Endomorphismen* zeichnen sich insbesondere dadurch aus, dass sie mit Ihrem adjungierten Endomorphismus kommutieren. Hierzu betrachten wir die folgende Definition.

**Definition 3.78** (Normaler Endomorphismus)
*Ist $V$ ein Euklidischer oder unitärer Vektorraum, so heißt ein Endomorphismus $F$ von $V$* normal*, wenn er mit seiner Adjungierten kommutiert, d.h.,*

$$F \circ F^* = F^* \circ F.$$

*Ist $A = \mathrm{M}_B(F)$ eine darstellende Matrix von $F$ bezüglich einer Orthonormalbasis $B$ von $V$, so bedeutet das*

$$A \cdot A^* \;=\; A^* \cdot A.$$

**Beispiel 3.79**
*Wir wollen im folgenden hinreichende Bedingungen für Normalität eines Endomorphismus angeben.*

*i) Jeder unitäre Endomorphismus $F$ ist normal, da wegen $F^* = F^{-1}$ gilt*

$$F \circ F^* \;=\; F \circ F^{-1} \;=\; \mathrm{Id}_V \;=\; F^{-1} \circ F \;=\; F^* \circ F.$$

*ii) Jeder selbstadjungierte Endomorphismus $F$ ist normal, da wegen $F^* = F$ gilt*

$$F \circ F^* \;=\; F \circ F \;=\; F^2 \;=\; F \circ F \;=\; F^* \circ F.$$

Für normale Endomorphismen stellt sich heraus, dass sowohl ihr Kern (sowie ihr Bild) mit denen der adjungierten Abbildung übereinstimmen, wie folgender Satz aussagt.

**Satz 3.80**
*Sei $V$ ein unitärer Vektorraum und $F\colon V \to V$ ein normaler Endomorphismus von $V$. Dann gilt*

$$\operatorname{Kern} F^* \;=\; \operatorname{Kern} F \quad \textit{und} \quad \operatorname{Bild} F^* \;=\; \operatorname{Bild} F..$$

*Beweis.* Sei $v \in \operatorname{Kern} F$, so können wir wegen der Normalität von $F$ folgern

$$\begin{aligned} 0 \;=\; \langle F(v), F(v)\rangle \;&=\; \langle v, F^* \circ F(v)\rangle \\ &=\; \langle v, F \circ F^*(v)\rangle \;=\; \overline{\langle F \circ F^*(v), v\rangle} \;=\; \overline{\langle F^*(v), F^*(v)\rangle}. \end{aligned}$$

Da das komplexe Skalarprodukt positiv definit ist muss also schon gelten, dass $F^*(v) = 0$ gilt und somit $v \in \operatorname{Kern} F^*$ ist. Damit haben wir gezeigt, dass $\operatorname{Kern} F^* = \operatorname{Kern} F$ gilt. Wegen der Dimensionsformel von Bild und Kern [Fischer2005, Satz 2.2.4] folgt dann auch schon direkt, dass $\operatorname{Bild} F^* = \operatorname{Bild} F$ ist. □

Um die Normalform von normalen Endomorphismen zu untersuchen, beweisen wir zunächst das folgende Lemma.

**Lemma 3.81**
*Sei $V$ ein Euklidischer bzw. unitärer Vektorraum und $F\colon V \to V$ ein Endomorphismus. Dann gelten die folgenden Aussagen:*

*i) $F$ ist genau dann normal, wenn*

$$\langle F^*(v), F^*(w)\rangle \;=\; \langle F(v), F(w)\rangle \quad \textit{für alle } v, w \in V. \tag{3.10}$$

*ii) Ist $F$ normal, so folgt für alle $v \in V$*

$$\|F^*(v)\| \;=\; \|F(v)\|.$$

*iii) Ist $F$ normal, so ist $G := F - \lambda \operatorname{Id}_V$ für alle $\lambda \in \mathbb{K}$ normal.*

*iv) Ist $F$ normal, so gilt für alle $v \in V$ und $\lambda \in \mathbb{K}$*

$$F(v) = \lambda v \quad \Leftrightarrow \quad F^*(v) = \bar{\lambda} v.$$

*Beweis.* Die einzelnen Aussagen folgen direkt aus der Definition 3.78 von normalen Endomorphismen.

i) Ist $F$ normal, so können wir wegen $(F^*)^* = F$ und der Definition 3.67 der Adjungierten folgern, dass gilt

$$\langle F(v), F(w)\rangle \;=\; \langle v, (F^* \circ F)(w)\rangle \;=\; \langle v, (F \circ F^*)(w)\rangle \;=\; \langle F^*(v), F^*(w)\rangle.$$

ii) Folgt sofort aus i), da $||v|| = \sqrt{\langle v, v\rangle}$ für alle $v \in V$ gilt.

iii) Wir leiten uns zunächst den adjungierten Endomorphismus $G^*$ von $G$ her. Seien $v, w \in V$ und $G := F - \lambda \operatorname{Id}_V$, dann gilt:

$$\begin{aligned}\langle G(v), w\rangle \;&=\; \langle (F(v) - \lambda \operatorname{Id}_V)(v), w\rangle \;=\; \langle F(v), w\rangle - \lambda\langle \operatorname{Id}_V(v), w\rangle \\ &=\; \langle v, F^*(w)\rangle - \lambda\langle v, \operatorname{Id}_V(w)\rangle \;=\; \langle v, (F^*(w) - \bar{\lambda} \operatorname{Id}_V)(w)\rangle \;=\; \langle v, G^*(w)\rangle.\end{aligned}$$

Es gilt also

$$G^* \;=\; F^* - \bar{\lambda} \operatorname{Id}_V .$$

Durch Einsetzen erhalten wir die Normalität von $G$ durch

$$\begin{aligned}G \circ G^* \;&=\; F \circ F^* - \lambda F^* - \bar{\lambda} F + \lambda\bar{\lambda} \\ &=\; F^* \circ F - \bar{\lambda} F - \lambda F^* + \lambda\bar{\lambda} \;=\; G^* \circ G.\end{aligned}$$

iv) Da $G = F - \lambda \operatorname{Id}_V$ nach iii) normal ist, können wir Satz 3.80 anwenden und erhalten damit:

$$\operatorname{Eig}(F; \lambda) \;=\; \operatorname{Kern} G \;=\; \operatorname{Kern} G^* \;=\; \operatorname{Eig}(F^*; \bar{\lambda}).$$

Dies zeigt insbesondere, dass $F$ und $F^*$ die gleichen Eigenvektoren besitzen.

□

Schließlich stellen wir mit dem folgenden Satz fest, dass die Normalform von normale Endomorphismen Diagonalgestalt besitzt.

**Satz 3.82** (Diagonalisierungssatz)
*Für einen Endomorphismus $F$ eines unitären Vektorraums $V$ sind die folgenden Eigenschaften äquivalent:*

*i) $F$ ist normal.*

*ii) Es gibt eine Orthonormalbasis $B$ von $V$ bestehend aus Eigenvektoren von $F$.*

*Beweis.*
*i) ⇒ ii)*: Wir führen den Beweis mittels vollständiger Induktion über $n = \dim V$.

**Induktionsanfang:** $n = 1$
Sei $\lambda_1 \in \mathbb{K}$ der Eigenwert von $F$ und $v_1 \in V$ der zugehörige Eigenvektor mit $||v_1|| = 1$ (durch Normalisierung). Dann bildet $B = (v_1)$ eine Orthonormalbasis von $V$.

**Induktionsschritt:** $n-1 \to n$
Die Induktionsannahme ist, dass die Aussage bereits für den Fall $n-1$ gezeigt wurde. Sei $\lambda_1 \in \mathbb{C}$ eine Nullstelle des charakteristischen Polynoms von $F$ und $v_1 \in V$ ein zugehöriger Eigenvektor mit $\|v_1\| = 1$ (durch Normalisierung). Wir definieren das orthogonale Komplement $W := \mathrm{lin}(\{v_1\})^\perp \subset V$ von $v_1$. Für jedes $w \in W$ gilt dann nach Lemma 3.81

$$\langle F(w), v_1 \rangle \;=\; \langle w, F^*(v_1) \rangle \;=\; \langle w, \bar{\lambda}_1 v_1 \rangle \;=\; \lambda_1 \langle w, v_1 \rangle \;=\; 0.$$

Daraus folgt $F(W) \subset W$. $F|_W \colon W \to W$ ist als Einschränkung von $F$ wieder normal und für die charakteristischen Polynome gilt

$$P_F \;=\; (t - \lambda_1) \cdot P_{F|_W}.$$

Nach Induktionsvorraussetzung wissen wir also, dass es eine Orthonormalbasis $B' = (v_2, \ldots, v_n)$ von $W$ aus Eigenvektoren von $F|_W$ gibt. Diese ergänzen wir um den Eigenvektor $v_1$ von $F$ zu einer Orthonormalbasis $B = (v_1, \ldots, v_n)$ von $V$ aus Eigenvektoren von $F$.

*ii)* $\Rightarrow$ *i)*: Die Matrix $D := \mathrm{M}_B(F)$ ist diagonal und es gilt

$$\mathrm{M}_{\mathcal{B}}(F^*) \;=\; D^* \;=\; \bar{D}.$$

Da die folgende Gleichung gilt

$$D \cdot D^* \;=\; D \cdot \bar{D} \;=\; \bar{D} \cdot D \;=\; D^* \cdot D$$

ist $F$ normal. $\square$

# Teil II

# Analysis

# Kapitel 4

# Normierte Vektorräume

Sie haben in [Burger2020, Kapitel 4 & 5] bereits die fundamentalen Konzepte von Folgen, Stetigkeit und Kompaktheit in *metrischen Räumen* kennengelernt. Wir werden diese grundlegenden Begriffe im Folgenden durch Verwendung der Norm aus Kapitel 3.1 wiederholen, weiter präzisieren und durch neue Erkenntnisse erweitern.

Die zugrunde liegende Struktur für dieses Kapitel ist ein Vektorraum auf dem eine Norm definiert ist.

**Definition 4.1** (Normierter Vektorraum)
*Sei $X$ ein $\mathbb{K}$-Vektorraum und sei*

$$||\cdot||_X \colon V \ \to \ \mathbb{R}_0^+$$

*eine Norm auf $X$. Dann nennen wir das Paar $(X, ||\cdot||_X)$ einen* normierten Vektorraum. *Wenn der mathematische Kontext eindeutig ist schreiben wir häufig nur $||\cdot||$ anstatt $||\cdot||_X$ und $X$ anstatt dem Paar $(X, ||\cdot||_X)$. In diesen Fällen nehmen wir immer die kanonische Norm des Vektorraums an, z.B., die Euklidische Norm für den $\mathbb{R}^n$.*

Zwischen Metriken und Normen gibt es einen direkten Zusammenhang, den wir kurz wiederholen wollen.

**Bemerkung 4.2**
*Es ist klar, dass jeder normierte Vektorraum auch ein metrischer Raum ist, da jede Norm $||\cdot||_X$ eine Metrik $d$ auf $X$ induziert durch:*

$$\begin{aligned} d \colon X \times X \ &\to \ \mathbb{R}_0^+, \\ (x,y) \ &\mapsto \ d(x,y) \ := \ \|x-y\|_X, \qquad \forall x, y \in X. \end{aligned}$$

*Metriken sind jedoch viel allgemeiner und nicht alle Metriken entstehen aus Normen. Beispielsweise ist die geodätische Distanz (der kürzeste Weg) auf der Erdoberfläche eine Metrik, die nicht durch eine Norm induziert wird.*

Sie haben bereits einige Normen in Ihrem Studium kennengelernt. Die wichtigsten Beispiele sind im Folgenden nochmal zusammengefasst.

**Beispiel 4.3**
*Wir können die folgenden Vektorräume mit einer Norm ausstatten.*

1. *Das Paar* $(\mathbb{R}^n, ||\cdot||_p)$ *mit der p-Norm*

$$\|x\|_p \;:=\; \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$$

   *ist für* $1 \leqslant p < \infty$ *ein normierter Vektorraum. Ein wichtiger Spezialfall für* $p = 2$ *ist der Euklidische Vektorraum.*

2. *Das Paar* $(\mathbb{R}^n, ||\cdot||_\infty)$ *mit der Maximumsnorm*

$$\|x\|_\infty \;:=\; \max_{i=1,\ldots,n} |x_i|$$

   *ist ein normierter Vektorraum.*

3. *Sei* $\ell^\infty$ *der Raum der beschränkten Folgen. Für unendliche Folgen gilt dann beispielsweise* $(x_n)_{n\in\mathbb{N}} := n \notin \ell^\infty$*, jedoch* $(x_n)_{n\in\mathbb{N}} := 1 - \frac{1}{n} \in \ell^\infty$*.*
   *Das Paar* $(\ell^\infty, ||\cdot||_\infty)$ *mit der Supremumsnorm*

$$\|x\|_\infty \;:=\; \sup_{i\in\mathbb{N}} |x_i|$$

   *ist ein normierter (unendlich-dimensionaler) Vektorraum.*

Jede Norm induziert eine Metrik, welche wiederum eine Topologie auf dem Vektorraum induziert. Hierdurch definieren sich insbesondere die fundamentalen Begriffe von offenen und abgeschlossenen Mengen in normierten Vektorräumen, wie wir im Folgenden sehen werden.

**Definition 4.4** (Umgebung)
*Sei* $X$ *ein normierter Raum und* $\epsilon \in \mathbb{R}_+$ *ein positives Skalar. Sei außerdem* $x \in X$ *ein beliebiger Punkt. Dann heisst für*

$$U_\epsilon(x) \;:=\; \{y \in X : ||x - y|| < \epsilon\}$$

*die* $\epsilon$-Umgebung *von* $x$*.*

**Definition 4.5** (Offene und abgeschlossene Mengen)
*Sei* $X$ *ein normierter Raum. Dann definieren wir folgende Begriffe für Teilmengen von* $X$*:*

1. *Eine Teilmenge $M \subset X$ heißt offen, wenn für alle Punkte $x \in M$ eine $\epsilon$-Umgebung $U_\epsilon(x) \subset M$ existiert.*

2. *Eine Teilmenge $M \subset X$ heißt abgeschlossen, wenn $X \backslash M$ offen ist.*

**Beispiel 4.6**
*Wir betrachten folgende typische Beispiele für offene und abgeschlossene Mengen in $\mathbb{R}$.*

*i) Sei $X = \mathbb{R}$ und $a, b \in \mathbb{R}$ mit $a < b$. Dann ist das Intervall $[a, b]$ abgeschlossen und das Intervall $(a, b)$ offen. Die Intervalle $[a, b)$ bzw. $(a, b]$ sind weder abgeschlossen noch offen.*

*ii) Die leere Menge $\varnothing$ sowie der gesamte normierte Raum $X$ sind die einzigen Mengen, die sowohl abgeschlossen als auch offen sind.*

*iii) In allgemeinen normierten Vektorräumen $X$ können wir für einen positiven Radius $r > 0$ abgeschlossene Kugeln $B_r^X$ um einen Punkt $x \in X$ betrachten mit*

$$B_r^X(x) \;:=\; \{y \in X : ||x - y||_X \leqslant r\}$$

*Wenn der mathematische Kontext eindeutig ist, schreiben wir häufig auch nur $B_r(x)$.*

## 4.1 Konvergenz von Folgen

Im Folgenden wollen wir uns noch einmal mit allgemeinen Konvergenzbegriffen in normierten Vektorräumen beschäftigen, welche sich direkt aus der Konvergenz in metrischen Räumen ableiten lassen indem man den in einer Metrik gemessenen Abstand durch die Norm der Differenz von Punkten ersetzt. Konvergente Folgen und Cauchy-Folgen können also in normierten Räumen analog wie in metrischen Räumen definiert werden. Sie zeichnen sich dadurch aus, dass der Abstand der Punkte gemessen in der Norm eine Nullfolge ist. Anders ausgedrückt, nennen wir eine Folge konvergent, wenn alle Folgeglieder ab einem bestimmten Index in einer $\epsilon$-Umgebung um den Grenzwert liegen.

**Definition 4.7** (Konvergente Folgen und Cauchy-Folgen)
*Sei $X$ ein normierter Raum und $(x_n)_{n\in\mathbb{N}}$ eine Folge in $X$.*

1. *Die Folge $(x_n)_{n\in\mathbb{N}}$ heißt* konvergent *mit Grenzwert $\overline{x} \in X$, falls folgende Bedingung erfüllt ist*
$$\forall \epsilon > 0 \; \exists n_0 \in \mathbb{N} \; \forall n \geqslant n_0 : ||x_n - \overline{x}|| < \epsilon.$$

2. *Die Folge $(x_n)_{n\in\mathbb{N}}$ heißt* Cauchy-Folge *mit Grenzwert $\overline{x} \in X$, falls folgende Bedingung erfüllt ist*
$$\forall \epsilon > 0 \; \exists n_0 \in \mathbb{N} \; \forall\, m, n \geqslant n_0 : ||x_n - x_m|| < \epsilon.$$

Genau wie im Fall von metrischen Räumen können wir auch leicht folgendes Resultat beweisen:

**Satz 4.8**
*Jede konvergente Folge in einem normierten Vektorraum ist eine Cauchy-Folge.*

*Beweis.* [Burger2020, Satz 2.30 & Satz 4.11] □

**Definition 4.9** (Vollständigkeit und Banachraum)
*Ein normierter Vektorraum $X$ heißt* vollständig*, wenn jede Cauchy-Folge in $X$ konvergiert. Ein vollständiger, normierter Vektorraum $X$ wird* Banachraum *genannt.*

**Beispiel 4.10**
*Wir untersuchen die folgenden Beispiele mit Blick auf Vollständigkeit.*

*i) Der normierte Vektorraum $(\mathbb{R}^n, ||\cdot||_2)$ mit der Euklidischen Norm ist vollständig wie bereits in [Burger2020, Satz 4.13] bewiesen wurde.*

*ii) Die Menge $C(\Omega;\mathbb{R}^n)$ der stetigen Funktionen auf einer kompakten Menge $\Omega \subset \mathbb{R}^n$ mit Werten in $\mathbb{R}^n$ zusammen mit der Supremumsnorm*

$$||f||_\infty := \sup_{x\in\Omega} ||f(x)||_2$$

*ist ein Banachraum.*

*iii) Die Menge der reellwertigen, stetigen Funktionen $C([a,b];\mathbb{R})$ auf dem abgeschlossenen Intervall $[a,b] \subset \mathbb{R}$ zusammen mit der $L^1$-Norm*

$$||f||_{L^1([a,b])} := \int_a^b |f(x)|\,\mathrm{d}x$$

*ist ebenfalls ein Banachraum.*

*iv) Der Vektorraum $\mathbb{Q}$ mit der Betragsfunktion aufgefasst als Norm ist nicht vollständig und somit kein Banachraum. Hierzu betrachten wir eine rekursiv definierte Folge*

$$x_{n+1} := \frac{1}{2}\left(x_n + \frac{2}{x_n}\right).$$

*Man kann zeigen, dass die Folge $(x_n)_{n\in\mathbb{N}}$ eine Cauchy-Folge ist bei der alle Folgenglieder offensichtlich rational sind. Anderseits konvergiert die Folge gegen den Grenzwert $\sqrt{2} \notin \mathbb{Q}$ (siehe [Burger2020, Satz 2.27]) und ist somit nicht konvergent in $(\mathbb{Q}, |\cdot|)$.*

*v) Die Menge der reellwertigen, stetig-differenzierbaren Funktionen $C^1(I;\mathbb{R})$ auf einem offenen Intervall $I \subseteq \mathbb{R}$ bildet zusammen mit der Supremumsnorm $||\cdot||_\infty$ als auch mit der sogenannten* Sobolev-Norm

$$||f||_{W^{1,\infty}(I)} \;:=\; ||f||_\infty + ||f'||_\infty$$

*einen normierten Vektorraum. Allerdings ist $(C^1(I;\mathbb{R}), ||\cdot||_\infty)$ nicht vollständig, sondern nur $(C^1(I;\mathbb{R}), ||\cdot||_{W^{1,\infty}(I)})$.*

*Intuitiv kann man sich klar machen, dass hier lediglich die Sobolev-Norm auch die Ableitung der Funktionen $u \in C^1(I;\mathbb{R})$ berücksichtigt und damit für konvergente Folgen sicherstellt, dass auch die erste Ableitung gleichmäßig konvergiert. Die Supremumsnorm hingegen „beachtet“ diese gar nicht.*

Wir wollen im Folgenden auch den Begriff einer konvergenten Teilfolge nochmals für normierte Vektorräume definieren und charakterisieren.

**Definition 4.11** (Teilfolge)
*Sei $(x_n)_{n\in\mathbb{N}}$ eine Folge in einem normierten Vektorraum $(X, ||\cdot||_X)$ und sei $K : \mathbb{N} \to \mathbb{N}$ eine streng monoton wachsende Abbildung, d.h. $K(i) > K(j)$ für $i > j$. Dann heißt $(x_{K(n)})_{n\in\mathbb{N}}$ Teilfolge von $(x_n)_{n\in\mathbb{N}}$. Wir werden im Folgenden für Teilfolgen auch kurz $(x_{k_n})_{n\in\mathbb{N}}$ schreiben.*

Um die Konvergenz von Teilfolgen festzustellen, führt man den Begriff eines Häufungspunktes einer Folge ein.

**Definition 4.12** (Häufungspunkt)
*Sei $(X, ||\cdot||_X)$ ein normierter Raum. Dann nennen wir einen Punkt $y \in X$* Häufungspunkt *der Folge $(x_n)_{n\in\mathbb{N}} \subset X$, wenn eine Teilfolge $(x_{n_k})_{k\in\mathbb{N}} \subset (x_n)_{n\in\mathbb{N}}$ existiert, die gegen $y$ konvergiert, d.h.*

$$\forall \epsilon > 0 \; \exists n_0 \in \mathbb{N} \; \forall k \geqslant n_0 : ||x_{n_k} - \overline{x}|| < \epsilon.$$

Das folgende Beispiel veranschaulicht die Beziehung zwischen Häufungspunkten und konvergenten Teilfolgen.

**Beispiel 4.13**
*Die reelle Folge $(x_n)_{n\in\mathbb{N}} \subset \mathbb{R}$ mit*

$$x_n \;:=\; (-1)^n + \frac{1}{n+1}$$

*besitzt zwei Häufungspunkte $\bar{x}_1 = +1$ und $\bar{x}_2 = -1$. Das bedeutet, dass die Folge $(x_n)_{n\in\mathbb{N}}$ selbst nicht konvergent ist, jedoch zwei konvergente Teilfolgen besitzt, die sich gerade aus den geraden und ungeraden Folgegliedern bilden.*

## 4.2 Stetigkeit

Aus [Burger2020, Kapitel 5] sind Sie bereits mit den verschiedenen Stetigkeitsbegriffen auf metrischen Räumen vertraut. Im Fall von normierten Vektorräumen lassen sich diese Konzepte durch Verwendung der Norm für den Abstandsbegriff analog definieren.

**Definition 4.14** (Stetigkeit)
*Wir wollen im Folgenden die wichtigsten Definitionen für Stetigkeit wiederholen. Hierbei betrachten wir eine Funktion* $f\colon X \to Y$ *zwischen zwei normierten Vektorräumen* $(X, ||\cdot||_X)$ *und* $(Y, ||\cdot||_Y)$.

1. *Die Funktion* $f$ *ist genau dann* stetig *in einem Punkt* $x_0 \in X$*, wenn es für alle* $\varepsilon > 0$ *ein* $\delta > 0$ *gibt, so dass für alle Punkte* $x \in X$ *mit*

   $$||x - x_0||_X \ < \ \delta$$

   *für den Abstand der Funktionswerte gilt*

   $$||f(x) - f(x_0)||_Y \ < \ \epsilon.$$

2. *Die Funktion* $f$ *ist genau dann* gleichmäßig stetig*, wenn es für alle* $\varepsilon > 0$ *ein* $\delta > 0$ *so gibt, dass für alle Punkte* $x, y \in X$ *mit*

   $$||x - y||_X \ < \ \delta$$

   *für den Abstand der Funktionswerte gilt*

   $$||f(x) - f(y)||_Y \ < \ \varepsilon.$$

3. *Die Funktion* $f$ *ist genau dann* Hölder-stetig *mit Exponent* $0 < \alpha \leqslant 1$*, wenn für alle* $x, y \in X$ *gilt:*

   $$||f(x) - f(y)||_Y \ \leqslant \ L \cdot ||x - y||_X^{\alpha}$$

   *für eine nicht-negative Konstante* $L \in \mathbb{R}_0^+$.

   *Im Spezialfall* $\alpha = 1$ *nennen wir* $f$ *auch* Lipschitz-stetig*. Gilt sogar* $0 \leqslant L < 1$*, so nennen wir die Funktion* $f$ *eine Kontraktion.*

Natürlich sind diese Stetigkeitsbegriffe unterschiedlich stark, wie folgende Bemerkung feststellt.

**Bemerkung 4.15**
*Für eine Funktion* $f\colon X \to Y$ *zwischen normierten Vektorräumen* $X$ *und* $Y$ *gelten folgende Implikationen:*

$$f \text{ ist Hölder-stetig} \ \Rightarrow \ f \text{ ist gleichmäßig stetig} \ \Rightarrow \ f \text{ ist stetig}\,.$$

Folgende Beispiele zeigen den Unterschied zwischen den Stetigkeitsbegriffen.

**Beispiel 4.16**
*Die Wurzel-Funktion*

$$\sqrt{\cdot} : \mathbb{R}_0^+ \ \rightarrow \ \mathbb{R}_0^+$$

*ist stetig. Sie ist jedoch nicht Lipschitz-stetig in $x = 0$.*

*Beweis.* In der Hausaufgabe zu zeigen. □

**Beispiel 4.17**
*Die Betragsfunktion*

$$|\cdot| : \mathbb{R} \ \rightarrow \ \mathbb{R}_0^+$$

*Lipschitz-stetig mit Lipschitz Konstante $L = 1$.*

Eine besonders interessante Klasse von stetigen Funktionen sind sogenannte *Homöomorphismen*. Diese Funktionen erlauben es zu untersuchen wann Teilmengen (topologischer) Vektorräume ineinander überführt werden können, z.B. durch Dehnen, Stauchen, Drehen oder Verzerren der Mengen. Die folgende Definition charakterisiert Homöomorphismen genauer

**Definition 4.18** (Homöomorphismus)
*Seien $X$ und $Y$ zwei normierte Vektorräume. Wir nennen eine Funktion $f\colon X \to Y$ einen* Homöomorphismus*, falls folgende Bedingungen erfüllt sind:*

- *$f$ ist stetig,*
- *$f$ ist bijektiv,*
- *die Umkehrfunktion $f^{-1}$ ist ebenfalls stetig.*

Der folgende berühmte Satz ist ein wichtiges Hilfsmittel zur Untersuchung von Fixpunktgleichungen, z.B. in der Numerik, und wird uns später bei der Behandlung von gewöhnlichen Differentialgleichungen noch hilfreich sein.

**Satz 4.19** (Fixpunktsatz von Banach)
*Es sei $X$ ein Banachraum und $F\colon X \to X$ eine Lipschitz-stetige Funktion mit Lipschitz-Konstante $L < 1$, d.h. es gilt*

$$||F(x) - F(y)|| \ \leqslant \ L \cdot ||x - y|| \quad \text{für alle } x, y \in X.$$

*Dann existiert ein genau ein Fixpunkt $\bar{x} \in X$, so dass*

$$F(\bar{x}) \ = \ \bar{x}.$$

*Beweis.* Es sei $x_0 \in X$ beliebiger Startwert der Folge $(x_k)_{k\in\mathbb{N}}$, die durch wiederholte Anwendung der Funktion $F$ definiert ist, d.h,

$$x_k \;:=\; F(x_{k-1}), \quad k \geqslant 1.$$

Wir werden im Folgenden zeigen, dass $(x_k)_{k\in\mathbb{N}}$ eine Cauchy-Folge in $X$ ist. Da $F$ Lipschitzstetig mit Lipschitz Konstante $L$ ist, gilt offensichtlich für alle $n \in \mathbb{N}$

$$||x_{n+1} - x_n|| \;=\; ||F(x_n) - F(x_{n-1})|| \;\leqslant L \cdot ||x_n - x_{n-1}||.$$

Damit folgt induktiv aber auch schon, dass

$$||x_{n+1} - x_n|| \;\leqslant L^n \cdot ||x_1 - x_0||.$$

Seien nun $n, m \in \mathbb{N}$ zwei beliebige Indizes mit $1 \leqslant n < m$, dann gilt wegen der Dreiecksungleichung der Norm und unter Ausnutzung der geometrischen Reihe:

$$\begin{aligned} ||x_m - x_n|| \;&\leqslant\; \sum_{k=n+1}^{m} ||x_k - x_{k-1}|| \;\leqslant\; ||x_1 - x_0|| \cdot \sum_{k=n+1}^{m} L^{k-1} \\ &=\; ||x_1 - x_0|| \cdot L^n \sum_{k=0}^{m-n-1} L^k \;\leqslant\; ||x_1 - x_0|| \cdot L^n \sum_{k=0}^{\infty} L^k \;=\; ||x_1 - x_0|| \cdot \frac{L^n}{1-L}. \end{aligned}$$

Da $L < 1$ nach Voraussetzung ist folgt, dass $L^n \to 0$ für $n \to \infty$ und somit gilt auch

$$\lim_{n,m\to\infty} ||x_m - x_n|| \;=\; 0$$

und daraus folgt, dass $(x_k)_{k\in\mathbb{N}}$ eine Cauchy-Folge in $X$ ist. Da $X$ vollständig ist, muss $(x_k)_{k\in\mathbb{N}}$ nach Definition 4.9 konvergieren gegen einen Grenzwert $\bar{x} \in X$. Das heißt wir haben gezeigt, dass die Funktion einen Fixpunkt $\bar{x} = F(\bar{x})$ besitzt, da gilt

$$\lim_{k\to\infty} F(x_k) \;=\; \lim_{k\to\infty} x_{k+1} \;=\; \bar{x}.$$

Sei $\hat{x} \in X$ nun ein weiterer Fixpunkt von $F$ mit $F(\hat{x}) = \hat{x}$, dann gilt

$$||\bar{x} - \hat{x}|| \;=\; ||F(\bar{x}) - F(\hat{x})|| \;\leqslant\; L \cdot ||\bar{x} - \hat{x}||.$$

Da $L < 1$ ist kann diese Ungleichung nur gelten wenn $||\bar{x} - \hat{x}|| = 0$ ist. Aus der positiven Definitheit der Norm folgt dann schon, dass $\bar{x} = \hat{x}$ sein muss, d.h., der Fixpunkt ist eindeutig. □

**Bemerkung 4.20**
*Der Banach'sche Fixpunktsatz kann noch allgemeiner für abgeschlossene Untermengen vollständiger metrischer Vektorräume formuliert werden, da man nur einen Abstandsbegriff (also eine Metrik) und die Existenz des Grenzwertes benötigt.*

## 4.3 Kompaktheit

Ein zentraler Begriff in der mathematischen Topologie ist Kompaktheit einer Menge, die für viele Existenzaussagen essentiell ist. Aus [Burger2020, Kapitel 5.1] kennen Sie bereits kompakte Mengen in metrischen Räumen, so dass die folgende Definition einer Wiederholung darstellt.

**Definition 4.21** (Kompaktheit)
*Eine Teilmenge $M \subseteq X$ eines normierten Raumes $X$ ist kompakt genau dann, wenn jede Folge $(x_n)_{n\in\mathbb{N}} \subseteq M$ eine konvergente Teilfolge enthält, deren Grenzwert in $M$ liegt.*

In endlichen Dimensionen haben Sie bereits ein wichtiges Resultat für konvergente Teilfolgen kennengelernt. Der Satz von Bolzano-Weierstrass, dass abgeschlossene Kugeln in $\mathbb{K}^n$ mit Körper $\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$ kompakt sind.

**Satz 4.22** (Bolzano-Weierstrass)
*Sei $\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$ ein Körper und $(x_k)_{k\in\mathbb{N}} \subset \mathbb{K}^n$ eine beschränkte Folge, d.h. es gibt eine Konstante $C \in \mathbb{R}^+$, so dass alle Folgenglieder $x_k, k \in \mathbb{N}$, in der abgeschlossenen Kugel $B_C(0) \subset \mathbb{K}^n$ liegen, d.h.*

$$\|x_k\| \leqslant C.$$

*Dann existiert eine konvergente Teilfolge $(x_{k_j})_{j\in\mathbb{N}}$ mit dem Grenzwert $\overline{x} \in X$ und es gilt $\|\overline{x}\| \leqslant C$.*

*Beweis.* Siehe [Burger2020, Satz 4.19] □

**Bemerkung 4.23**
*Wir können folgende Beobachtungen zum Satz von Bolzano-Weierstrass 4.22 festhalten.*

1. *In endlich-dimensionalen Vektorräumen $X$ lässt sich eine noch stärkere Aussage beweisen, die sich auf den Satz von Heine-Borel [Burger2020, Satz 5.18] zurückführen lässt. Anstatt nur einer hinreichenden Bedingung, kann man aussagen, dass eine Teilmenge $M \subset X$ genau dann kompakt ist, wenn Sie beschränkt and abgeschlossen ist.*

2. *Der Satz gilt nicht mehr, wenn der betrachtete normierte Raum unendlich-dimensional ist, beispielsweise in Funktionen- bzw. Folgenräumen wie $C([a,b];\mathbb{R})$ oder*

$$\ell^2(\mathbb{R}) \;=\; \{(x_n)_{n\in\mathbb{N}} : x_n \in \mathbb{R},\ \sum_{n=1}^{\infty} |x_n|^2 < \infty\}.$$

   *Betrachten wir beispielsweise die Folge $(x_n)_{n\in\mathbb{N}} \subset \ell^2(R)$ der unendlich-dimensionalen Einheitsvektoren $e_n \in \ell^2(\mathbb{R})$ mit*

$$x_n \;=\; e_n \;:=\; (0,\ldots,0,1,0,\ldots), \quad n \in \mathbb{N},$$

*d.h., das Folgenglied $x_n$ besitzt eine Eins an der n-ten Stelle und sonst nur Nullen.*

*Es wird klar, dass jedes Folgenglied auf der Einheitskugel des Vektorraums $\ell^2(\mathbb{R})$ liegt, da offensichtlich gilt $||x_n|| = 1$ für alle $n \in \mathbb{N}$. Das heißt, dass die Folge $(x_n)_{n\in\mathbb{N}}$ beschränkt ist und in der abgeschlossenen Einheitskugel liegt. Dennoch besitzt die Folge keine konvergente Teilfolge, da*

$$||x_m - x_n|| \;=\; ||e_m - e_n|| \;=\; \sqrt{2}, \qquad \text{für alle } m, n \in \mathbb{N}, m \neq n,$$

*d.h., der Abstand eines beliebigen Folgeglieds zu allen anderen Folgegliedern ist konstant $\sqrt{2}$. Somit kann keine konvergente Teilfolge von $(x_n)_{n\in\mathbb{N}}$ existieren und damit ist die Einheitskugel in $\ell^2(\mathbb{R})$ nicht kompakt.*

Wir können über den Begriff der Kompaktheit sogar erkennen, ob ein Vektorraum endlich erzeugt ist. Als Vorbereitung dieser Beobachtung benötigen wir jedoch zuerst das folgende Hilfslemma. Die Motivation für dieses Lemma ist, dass wir den Begriff eines senkrechten Vektors auf einen Unterraum aus Kapitel 3.5 nicht in beliebigen normierten Vektorräumen anwenden können (da ggfs. kein Skalarprodukt existiert). Dennoch können wir nach dem folgenden Lemma Einheitsvektoren finden, die außerhalb dieses Unterraums liegen und einen positiven Abstand zu diesem besitzen.

**Lemma 4.24** (Lemma von Riesz)
*Es sei $X$ ein normierter Vektorraum und $U \subsetneq X$ ein abgeschlossener, echter Unterraum von $X$. Sei außerdem ein $\delta \in \mathbb{R}$ mit $0 < \delta < 1$ gegeben. Dann existiert ein Element $y \in X$ der Einheitskugel mit $||y||_X = 1$, so dass*

$$d(y, U) \;:=\; \inf_{u\in U} ||y - u||_X \;\geqslant\; \delta.$$

*Beweis.* Es sei $\delta \in (0, 1)$ beliebig vorgegeben. Wir wählen ein beliebiges Element $z \in X\backslash U$ aus dem Komplement von $U$ und erkennen, dass der Abstand $d(z, U)$ von $z$ zu $U$ positiv ist. Da wir den Unterraum $U \subsetneq X$ als abgeschlossen angenommen haben, existiert ein Element $u_0 \in U$ für das gilt

$$d(z, U) \;:=\; \inf_{u\in U} ||z - u||_X \;\leqslant\; ||z - u_0||_X \;\leqslant\; \frac{d(z, U)}{\delta}.$$

Und damit können wir folgern, dass

$$\frac{\delta}{d(z, U)} \;\leqslant\; \frac{1}{||z - u_0||_X}. \tag{4.1}$$

Wählen wir nun den Einheitsvektor

$$y \;:=\; \frac{z - u_0}{||z - u_0||_X},$$

so ist offensichtlich $||y||_X = 1$.

Für ein beliebiges Element $v \in U$ können wir nun nachrechnen, dass gilt

$$||y - v||_X \;=\; \left\| \frac{z - u_0}{||z - u_0||_X} - v \cdot \frac{||z - u_0||_X}{||z - u_0||_X} \right\|_X \;=\; \frac{1}{||z - u_0||_X} \left\| z - u_0 - v \cdot ||z - u_0||_X \right\|_X .$$

Da der Vektor $(u_0 - v \cdot ||z - u_0||_X) \in U$ ist, können wir folgende Abschätzung treffen

$$\begin{aligned} \frac{1}{||z - u_0||_X} \left\| z - u_0 - v \cdot ||z - u_0||_X \right\|_X \;&\geqslant\; \frac{1}{||z - u_0||_X} \inf_{u \in U} ||z - u||_X \\ &\overset{(4.1)}{\geqslant}\; \frac{\delta}{d(z, U)} d(z, U) \;=\; \delta. \end{aligned}$$

□

Mit Hilfe des obigen Lemmas können wir den folgenden interessanten Kompaktheitssatz beweisen, der es uns erlaubt einen endlich-dimensionalen Vektorraum an der Kompaktheit seiner Einheitskugel zu erkennen.

**Satz 4.25** (Kompaktheitssatz von Riesz)
*Ein normierter Vektorraum $X$ ist genau dann endlich-dimensional, wenn seine entsprechende Einheitskugel*

$$B_1(0) \;:=\; \{x \in X : ||x||_X \leqslant 1\}$$

*kompakt ist.*

*Beweis.* Wir zeigen die beiden Implikationen der Äquivalenzaussage im Folgenden getrennt voneinander.

„⇒“: Aus der Linearen Algebra wissen wir, dass es für jeden endlich-dimensionalen $\mathbb{K}$-Vektorraum $X$ ein Homöomorphismus $\Phi\colon X \to \mathbb{K}^n$ existiert. Es sei nun $(x_k)_{k \in \mathbb{N}} \subseteq B_1^X(0)$ eine Folge in der Einheitskugel von $X$, so dass

$$(\Phi(x_k))_{k \in \mathbb{N}} \;\subseteq\; B_1^{\mathbb{K}^n}(0).$$

Da die Einheitskugel $B_1^{\mathbb{K}^n}(0)$ nach dem Satz 4.22 von Bolzano-Weierstraß kompakt ist, existiert eine konvergente Teilfolge

$$(y_k)_{k \in \mathbb{N}} \;\subseteq\; (\Phi(x_k))_{k \in \mathbb{N}}.$$

Weil $\Phi$ als Homöomorphismus insbesondere eine stetige Umkehrabbildung $\Phi^{-1}$ besitzt, ist dann aber auch

$$(z_k)_{k \in \mathbb{N}} \;:=\; (\Phi^{-1}(y_k))_{k \in \mathbb{N}}$$

eine konvergente Teilfolge in $B_1^X(0)$. Da die Folge $(x_k)_{k \in \mathbb{N}}$ beliebig gewählt war, folgt die Kompaktheit der Einheitskugel $B_1^X(0) \subset X$.

„⇐“: Um die Rückrichtung der Aussage zu beweisen, führen wir einen Beweis über die Kontraposition. Es sei $X$ nicht endlich-dimensional, d.h, unendlich-dimensional. Dann existiert eine aufsteigende Folge von echten, abgeschlossenen Unterräumen $(U_k)_{k\in\mathbb{N}} \subsetneq X$ von $X$ mit

$$U_1 \subsetneq U_2 \subsetneq \ldots \subsetneq U_k \subsetneq \ldots \ , \quad \dim U_k < \dim U_{k+1} \text{ für } k \in \mathbb{N}.$$

Für ein beliebiges $\delta \in (0,1)$ wählen wir mit dem Lemma 4.24 von Riesz einen Einheitsvektor $y_k \in U_k$, so dass

$$\inf_{u\in U_{k-1}} ||y-u||_X \ \geqslant \ \delta.$$

Für beliebige Wahl des ersten Einheitsvektors $y_1 \in X$ mit $||y_1|| = 1$ haben wir damit ein Folge $(y_k)_{k\in\mathbb{N}}$ auf der Einheitskugel $B_1^X(0)$ konstruiert, welche beschränkt ist, jedoch keine Cauchy-Folge sein kann, da der Abstand zwischen den Folgegliedern immer mindestens $\delta$ beträgt. Daraus folgt, dass die Einheitskugel in einem unendlich-dimensionalen Vektorraum $X$ nicht kompakt sein kann. □

Der Kompaktheitssatz 4.25 von Riesz besagt zwar, dass Kugeln in unendlich-dimensionalen Banachräumen nicht kompakt sind. Dennoch heißt das nicht, dass es in unendlich-dimensionalen Banachräumen keine kompakten Teilmengen gibt. Um dies zu verdeutlichen beschäftigen uns abschließend mit einem fundamentalen Satz, der Aussagen über die Kompaktheit bestimmter Teilmengen des Funktionenraumes $(C([a,b],\mathbb{R}^n), ||\cdot||_\infty)$ der stetigen Funktionen auf einem Intervall $[a,b] \subset \mathbb{R}$ zusammen mit der Supremumsnorm erlaubt. Dies ist nur ein Spezialfall des Satzes von Arzelà-Ascoli, der in der allgemeinen Formulierung abstraktere Räume untersucht.

**Satz 4.26** (Arzelà-Ascoli)
*Sei $[a,b] \subset \mathbb{R}$ ein Intervall und es sei $(f_n)_{n\in\mathbb{N}}$ eine Folge von Funktionen aus $C([a,b],\mathbb{R})$, welche punktweise beschränkt und gleichgradig stetig ist, d. h.*

*(i)* $\forall x \in [a,b]\ \exists K \in \mathbb{R}^+ \colon \sup_{n\in\mathbb{N}} |f_n(x)| \leqslant K$,

*(ii)* $\forall \varepsilon > 0\, \exists \delta > 0\, \forall n \in \mathbb{N}\ \forall x,y \in [a,b] \colon (|x-y| < \delta \Rightarrow |f_n(x) - f_n(y)| < \varepsilon)$.

*Dann besitzt die Folge $(f_n)_{n\in\mathbb{N}}$ eine gleichmäßig konvergente Teilfolge.*

*Beweis.* Der Beweis dieses Satzes beruht auf dem *Cantorschen Diagonalverfahren*, das rekursiv Teilfolgen konstruiert die partiell konvergieren und dann anschließend aus all diesen Teilfolgen eine überall konvergente Teilfolge zu konstruieren. Wir betrachten hierfür zunächst die abzählbar unendliche Teilmenge $\mathbb{Q}\cap[a,b]$ des Intervalls $[a,b]$ mit der Abzählung

$$\Phi \colon \mathbb{Q} \cap [a,b] \to \mathbb{N}.$$

Für $x \in \mathbb{Q} \cap [a,b]$ mit $\Phi(x) = i$ schreiben wir $x_i$. Weil alle Folgeglieder $(f_n(x_1))_{n\in\mathbb{N}}$ beschränkt sind nach Voraussetzung, existiert eine konvergente Teilfolge $(f_n^1(x_1))_{n\in\mathbb{N}}$ von

$(f_n(x_1))_{n\in\mathbb{N}}$ nach dem Satz 4.22 von Bolzano-Weierstraß. Aus dieser konvergenten Teilfolge können wir wiederum eine Teilfolge $(f_n^2(x_2))_{n\in\mathbb{N}}$ mit der gleichen Argumentation bestimmen. Dies können wir sukzessive weiterführen für alle Punkte $(x_n)_{n\in\mathbb{N}} \subset [a,b]$.

Wir definieren nun eine spezielle Funktionenfolge $(g_n)_{n\in\mathbb{N}} \subset (f_n)_{n\in\mathbb{N}}$ mit

$$\begin{aligned} g_n \colon [a,b] &\to \mathbb{R}, \\ x &\mapsto f_n^n(x) \end{aligned}$$

und beweisen, dass diese Folge gleichmäßig konvergiert in $C([a,b];\mathbb{R})$. Hiebei reicht es zu zeigen, dass $(g_n)_{n\in\mathbb{N}}$ eine Cauchy-Folge ist, da wir aus Beispiel 4.10 wissen, dass $C([a,b];\mathbb{R})$ vollständig bezüglich der Supemumsnorm ist.

Sei nun $\epsilon > 0$ beliebig gewählt. Nach Voraussetzung existiert nun ein entsprechendes $\delta > 0$ so dass für alle $x, y \in [a,b]$ folgt, dass $|f_n(x) - f_n(y)| < \epsilon$ falls $|x-y| < \delta$ gilt. Wir partitionieren das Intervall $[a,b]$ in kleine Teilintervalle $I_j, j = 1,\dots,m$ mit $\operatorname{diam}(I_j) < \delta$, so dass

$$[a,b] \;=\; \biguplus_{j=1}^{m} I_j.$$

Dann wählen wir Punkte aus jedem Teilintervall mit

$$x_1 \in I_1 \cap \mathbb{Q}, \quad x_2 \in I_2 \cap \mathbb{Q}, \;\dots, \quad x_m \in I_m \cap \mathbb{Q}.$$

Für einen beliebigen Index $j \in \{1,\dots,m\}$ seien $k(j) > \ell(j)$ so groß, dass die Teilfolge $(f_n^\ell(x_j))_{n\in\mathbb{N}}$ punktweise konvergiert. Als Teilfolge jener Folge konvergiert auch $(f_n^k(x_i))_{n\in\mathbb{N}}$ gegen den gleichen Grenzwert. Wir finden daher $\ell(j)$ so groß, dass

$$|g_k(x_i) - g_\ell(x_i)| \;=\; |f_k^k(x_i) - f_\ell^\ell(x_i)| \;<\; \frac{\varepsilon}{3}. \tag{4.2}$$

Nehmen wir nun einen globalen Index über alle Teilintervalle mit $\ell > \max\{\ell(1),\dots,\ell(m)\}$, so gilt die obige Abschätzung für alle $x_i$, $i \in \{1,\dots,m\}$.

Es sei nun $x \in [a,b]$ ein beliebiger Punkt. Dann existiert ein Index $j \in \{1,\dots,m\}$, so dass der Punkt $x$ im Teilintervall $I_j$ liegt. Für ein $k > \ell > \max\{\ell(1),\dots,\ell(m)\}$ gilt nun mit Abschätzung (4.2) und wegen der geradlinigen Stetigkeit der Folge:

$$|g_k(x) - g_\ell(x)| \;\leqslant\; |g_k(x) - g_k(x_j)| + |g_k(x_j) - g_\ell(x_j)| + |g_\ell(x_j) - g_\ell(x)| \;\leqslant\; \varepsilon$$

Da dies für beliebige Punkte $x \in [a,b]$ gilt ist $(g_k)_{k\in\mathbb{N}}$ eine Cauchy-Folge bezüglich der Supremumsnorm und konvergiert somit gleichmäßig in $C([a,b];\mathbb{R})$. □

## 4.4 Hilberträume

Im Folgenden führen wir den Begriff eines Hilbertraums als Spezialfall eines Banachraums ein. Hilberträume weisen besonders viel Struktur auf, da ihre Norm durch ein Skalarprodukt

induziert wird. Daher ist es nicht verwunderlich, dass sie in vielen Theorien der Physik eine wichtige Rolle spielen, wie zum Beispiel in der Quantenmechanik. Auch im Bereich des maschinellen Lernens sind Hilberträume ein gängiges Werkzeug, wie zum Beispiel bei den sogenannten *Kernel Methods* in der statistischen Lerntheorie.

**Definition 4.27** (Hilbertraum)
*Wir nennen einen Banachraum $(X, ||\cdot||_X)$, dessen Norm durch ein Skalarprodukt induziert ist einen* Hilbertraum.

Das folgende Beispiel diskutiert verschiedene Banachräume und erklärt ob es sich jeweils um einen Hilbertraum handelt oder nicht.

**Beispiel 4.28**
*Im Folgenden untersuchen wir Beispiele für endlich- und unendlich-dimensionale Banachräume, d.h., vollständige normierte Vektorräume, und entscheiden ob diese einen Hilbertraum darstellen.*

1. *Der Euklidische Vektorraum $\mathbb{R}^n$ mit $n \in \mathbb{N}$ ist ausgestattet mit der Euklidischen Norm*

$$||x||_2 \;:=\; \sqrt{\sum_{i=1}^{n} |x_i|^2} \;=\; \sqrt{\langle x, x\rangle}, \qquad \text{für alle } x \in \mathbb{R}^n$$

   *ein endlich-dimensionaler Hilbertraum.*

2. *Der Banachraum $(\mathbb{R}^n, ||\cdot||_\infty)$ ist kein Hilbertraum.*

3. *Der Matrizenraum $\mathbb{K}^{m\times n}$ der reellen oder komplexen Matrizen ausgestattet mit dem Frobenius-Skalarprodukt*

$$\langle A, B\rangle \;:=\; \sum_{i=1}^{m}\sum_{j=1}^{n} \bar{a}_{ij} b_{ij}, \qquad \text{für alle Matrizen } A, B \in \mathbb{K}^{m\times n}$$

   *ist ein endlich-dimensionaler Hilbertraum.*

4. *Der Banachraum $(\ell^2(\mathbb{R}), ||\cdot||_{\ell^2})$ der Folgen, deren Summe ihrer Quadrate endlich ist, zusammen mit der Norm*

$$||(x_n)_{n\in\mathbb{N}}||_{\ell^2} \;:=\; \sqrt{\sum_{n=1}^{\infty} |x_n|^2}$$

   *ist ein unendlich-dimensionaler Hilbertraum.*

5. *Der unendlich-dimensionale Banachraum* $(\ell^1(\mathbb{R}), ||\cdot||_{\ell^1})$ *der Folgen, deren Summe ihrer Beträge endlich ist, zusammen mit der Norm*

$$||(x_n)_{n\in\mathbb{N}}||_{\ell^1} \;:=\; \sum_{n=1}^{\infty} |x_n|$$

*ist kein Hilbertraum.*

6. *Der Raum der skalarwertigen, quadrat-integrierbaren Funktionen* $L^2$ *auf einem Gebiet* $\Omega \subset \mathbb{R}^n$ *ausgestattet mit der Norm*

$$||f||_{L^2} \;:=\; \left(\int_\Omega |f(x)|^2 \,\mathrm{d}x\right)^{\frac{1}{2}} \;=\; \left(\int_\Omega \overline{f(x)} f(x) \,\mathrm{d}x\right)^{\frac{1}{2}} \;=\; \sqrt{\langle f, f\rangle_{L^2}}$$

*ist ein unendlich-dimensionaler Hilbertraum.*

## 4.5 Dualräume

Zu einem normierten Vektorraum können wir seinen sogenannten topologischen Dualraum betrachten. Dieser Dualraum erlaubt es beispielsweise in der Differentialgeometrie eine Integration auf einer Mannigfaltigkeit zu definieren oder aber Optimierungsprobleme in eine (manchmal angenehmere) äquivalente, duale Form zu überführen. Wir definieren daher zunächst den Begriff des topologischen Dualraums im Folgenden.

**Definition 4.29** (Topologischer Dualraum und Funktional)
*Der* topologische Dualraum $X'$ *zu einem normierten* $\mathbb{K}$*-Vektorraum* $X$ *ist definiert als die Menge aller stetigen, linearen Funktionen von* $X$ *in den Körper* $\mathbb{K}$ *der reellen oder komplexen Zahlen. Oft spricht man nur von dem zugehörigen Dualraum, wenn der mathematische Kontext eindeutig ist.*

*Insbesondere wenn* $X$ *unendlich-dimensional ist, nennt man eine Funktion* $F \in X'$

$$F\colon X \to \mathbb{K}$$

*häufig ein* Funktional.

Folgendes Beispiel erinnert an Funktionale, die Sie bereits kennengelernt haben.

**Beispiel 4.30**
*Die Abbildung*

$$\begin{aligned} F\colon C([a,b];\mathbb{R}) \;&\to\; \mathbb{R}, \\ f \;&\mapsto\; F(f) \;:=\; \int_a^b f(x)\,\mathrm{d}x \end{aligned}$$

*ist linear und stetig für Funktionen $f$, die auf einem Intervall $I \subset \mathbb{R}$ integrierbar sind, d.h., es handelt sich um ein Funktional $F \in C([a,b];\mathbb{R})'$.*

*Ein physikalisch motiviertes Beispiel für die Anwendung von Funktionalen wäre die Bewegung eines Massepunkt entlang einer Kurve $\varphi\colon [0,1] \to \mathbb{R}^3$ in einem Potentialfeld $V\colon \mathbb{R}^3 \to [0,\infty)$. Hierbei berechnet man die verrichtete Arbeit durch das folgende Funktional*

$$F(V) \;:=\; \int_0^1 V(\varphi(s))\,\mathrm{d}s.$$

**Bemerkung 4.31**
*Folgende Anmerkungen zum Dualraum wollen wir festhalten.*

1. *Neben dem topologischen Dualraum $X'$ existiert in der Literatur der* algebraische Dualraum *$X^*$, aller linearen (aber nicht notwendigerweise stetigen) Funktionale von $X$ nach $\mathbb{K}$. Für endlich-dimensionale Vektorräume $X$ stimmen der algebraische und topologische Dualraum überein, da in diesem Fall alle linearen Operatoren auf $X$ automatisch stetig sind.*

2. *Da $X$ ein normierter Vektorraum ist, ist sein zugehöriger topologischer Dualraum $X'$ auch ein normierter Vektorraum ausgestattet mit der folgenden Operatornorm*

$$||F||_{X'} \;=\; \sup_{||x||_X \leqslant 1} |F(x)|.$$

   *Da die Funktionale $F \in X'$ per Definition in den vollständigen Körper $\mathbb{K}$ abbilden ist $X'$ selbst ein Banachraum, unabhängig davon ob $X$ ein Banachraum ist.*

3. *Falls $X$ sogar ein Hilbertraum ist, so ist sein zugehöriger topologischer Dualraum $X'$ nach dem Darstellungssatz von Fréchet-Riesz [] isometrisch isomorph zum Vektorraum $X$ selbst. Das bedeutet, dass ein isometrischer Isomorphismus $\Phi$ existiert, der jedem Element $x \in X$ ein eindeutiges Funktional $F \in X'$ zuordnet mit*

$$\begin{aligned} \Phi\colon X \;&\to\; X' \\ x \;&\mapsto\; \Phi(x) \;:=\; \langle x, \cdot\rangle_X \;=:\; F(x). \end{aligned}$$

Das folgende Beispiel illustriert noch einmal die Isomorphie eines Hilbertraums mit seinem Dualraum für den endlich-dimensionalen Fall.

**Beispiel 4.32**
*Wir betrachten den Hilbertraum $(\mathbb{R}^n, ||\cdot||_2)$ und versuchen zu verstehen, wie sein zugehöriger Dualraum $X'$ aussieht. Da alle Elemente $a \in X'$ mit $a\colon \mathbb{R}^n \to \mathbb{R}$ linear sein müssen, wird klar, dass diese Abbildungen von der Form $\langle a, \cdot\rangle$ sein müssen mit:*

$$a(x) \;:=\; \langle a, x\rangle \;=\; (a_1,\ldots,a_n)^T \cdot (x_1,\ldots,x_n) \;=\; \sum_{i=1}^n a_i x_i, \qquad \textit{für alle } x \in \mathbb{R}^n.$$

*Offensichtlich realisieren diese linearen Abbildungen Skalarprodukte mit n-dimensionalen Vektoren, welche auch häufig* kovariante Vektoren *oder* Kovektoren *genannt werden. Ein Isomorphismus zwischen $X$ und seinem Dualraum $X'$ lässt sich mittels darstellender Matrizen bezüglich der Standardbasen der beiden Räume bestimmen.*

# Kapitel 5

# Integralrechnung

Im letzten Semester haben Sie bereits die wichtige Klasse der Riemann-integrierbaren Funktionen [Burger2020, Kapitel 7.1] kennengelernt. Wir wollen damit beginnen die wichtigsten Konzepte und Beobachtungen nochmal kurz zu wiederholen und dann nützliche Formeln für die Integralrechnung für Riemann-integrierbare Funktionen in einer Variablen herleiten. Insbesondere werden wir die Substitutionsregel und die Partialbruchzerlegung für die Integration rationaler Funktionen einführen. Die Berechnung von Integralen für Funktionen in mehreren Variablen werden später im Studium behandelt und in diesem Zuge wird das Riemann-Integral durch einen allgemeineren Integralbegriff (dem Lebesgue-Integral) ersetzt.

Zunächst wollen wir wiederholen, was wir unter einem unbestimmten Integral und einer Stammfunktion verstehen.

**Definition 5.1** (Unbestimmtes Integral und Stammfunktion)
*Sei* $[a,b] \subset \mathbb{R}$ *ein Intervall und* $f : [a,b] \to \mathbb{R}$ *eine integrierbare Funktion, d.h., eine Funktion die auf jedem Teilintervall* $I \subset [a,b]$ *integrierbar ist. Wir nennen eine Funktion* $F : [a,b] \to \mathbb{R}$ unbestimmtes Integral *von* $f$*, wenn für alle* $y, z \in [a,b]$ *gilt*

$$F\Big|_y^z := F(z) - F(y) = \int_y^z f(x)\,\mathrm{d}x.$$

*Weiterhin nennen wir* $F : [a,b] \to \mathbb{R}$ *eine* Stammfunktion *von* $f$*, wenn* $F'(x) = f(x)$ *für alle* $x \in [a,b]$ *gilt.*

**Bemerkung 5.2**
*Wir beachten, dass wir aus Konsistenzgründen ein Integral mit vertauschten Integralgrenzen mit negativem Vorzeichen definieren, d.h.*

$$\int_z^y f(x)\ dx = -\int_y^z f(x)\,\mathrm{d}x.$$

Wegen seiner großen Bedeutung in der Analysis wiederholen wir im Folgenden den Hauptsatz der Differential- und Integralrechnung (auch Fundamentalsatz der Analysis genannt), der im Grunde aussagt, dass Integration und Differentiation die jeweilige Umkehrung voneinander sind. Außerdem erklärt er, dass die Begriffe Stammfunktion und unbestimmtes Integral übereinstimmen.

**Satz 5.3** (Fundamentalsatz der Analysis)
*Sei $f\colon [a,b] \to \mathbb{R}$ eine stetige Funktion. Dann ist*

$$\begin{aligned} F_a\colon [a,b] &\to \mathbb{R} \\ x &\mapsto F_a(x) := \int_a^x f(y)\,\mathrm{d}y \end{aligned}$$

*eine Stammfunktion von $f$. Eine Funktion $F\colon [a,b] \to \mathbb{R}$ ist genau dann Stammfunktion von $f$, wenn $F$ ein unbestimmtes Integral ist.*

*Beweis.* Siehe [Burger2020, Satz 7.10] □

**Bemerkung 5.4**
*Da unbestimmte Integrale und Stammfunktionen übereinstimmen schreibt man häufig etwas unpräzise für eine Stammfunktion $F(x)$ eines Integranden $f(x)$ folgenden Zusammenhang*

$$\int f(x)\,\mathrm{d}x = F(x) + C,$$

*wobei $C$ eine beliebige Integrationskonstante ist.*

Interpretieren wir die Integration als die Umkehrung der Differentiation, dann müssen wir die uns bekannten Ableitungsregeln noch einmal betrachten, da diese mit der Integration harmonieren sollen. Die wohl wichtigsten Regeln der Differentiation einer Funktion in einer Variablen, sind im Folgenden zusammengefasst.

**Satz 5.5** (Differentiationsregeln)
*Sei $[a,b] \subset \mathbb{R}$ ein Intervall und $f, g\colon [a,b] \to \mathbb{R}$ differenzierbare Funktionen und $c \in \mathbb{R}$ ein Skalar. Dann gelten die folgenden Regeln der Differentiationsrechnung für alle $x \in [a,b]$:*

$$\begin{aligned} (c \cdot f)'(x) &= c \cdot f'(x), && \text{(Faktorregel)} \\ (f+g)'(x) &= f(x) + g(x), && \text{(Summenregel)} \\ (f \cdot g)'(x) &= f'(x) \cdot g(x) + f(x) \cdot g'(x), && \text{(Produktregel)} \\ \left(\frac{f}{g}\right)'(x) &= \frac{f'(x) \cdot g(x) - f(x) \cdot g'(x)}{(g(x))^2}, && \text{(Quotientenregel)} \\ (f \circ g)'(x) &= (f' \circ g)(x) \cdot g'(x). && \text{(Kettenregel)} \end{aligned}$$

*Beweis.* Siehe [Burger2020, Satz 6.5] □

Im folgenden Beispiel werden die Rechenregeln der Differentiation aus Satz 5.5 angewendet.

**Beispiel 5.6**
*Es seien $f, g\colon \mathbb{R}^+ \to \mathbb{R}$ stetig differenzierbare Funktionen mit*

$$f(x) := \ln(x), \quad g(x) := x^2, \quad \text{für alle } x > 0.$$

*Wir wollen nun die Ableitung der folgenden Kombinationen von Funktionen mit den Regeln der Differentiation berechnen.*

*i) Wir wenden zuerst die* Produktregel *für die Differentiation an:*

$$\begin{aligned}(f \cdot g)'(x) &= (\ln(x) \cdot x^2)' = \ln'(x) \cdot x^2 + \ln(x) \cdot (x^2)' = \frac{1}{x} \cdot x^2 + \ln(x) \cdot 2x \\ &= x + 2x \cdot \ln(x) = x \cdot (1 + 2\ln(x)).\end{aligned}$$

*ii) Es folgt die* Quotientenregel *für die Differentiation:*

$$\left(\frac{f}{g}\right)'(x) = \left(\frac{\ln}{x^2}\right)'(x) = \frac{\ln'(x) \cdot x^2 - \ln(x)(x^2)'}{(x^2)^2} = \frac{\frac{x^2}{x} - \ln(x) \cdot 2x}{x^4} = \frac{1 - 2\ln(x)}{x^3}.$$

*iii) Zuletzt wenden wir die* Kettenregel *für die Differentiation an:*

$$(f \circ g)'(x) = (\ln \circ\, x^2)'(x) = \ln'(x^2) \cdot (x^2)' = \frac{1}{x^2} \cdot 2x = \frac{2}{x} = (2\ln(x))'.$$

## 5.1 Partielle Integration

Aus der Produktregel in Satz 5.5 können wir eine Regel für die *Partielle Integration* ableiten:

**Satz 5.7** (Partielle Integration)
*Seien $f, g\colon [a, b] \to \mathbb{R}$ stetig differenzierbare Funktionen. Dann gilt die folgende Rechenregel der partiellen Integration:*

$$\int_a^b f(x)g'(x)\,\mathrm{d}x = (f \cdot g)\Big|_a^b - \int_a^b f'(x)g(x)\,\mathrm{d}x\,. \tag{5.1}$$

*Beweis.* Siehe [Burger2020, Satz 7.13] □

**Bemerkung 5.8**
*Um die Stammfunktion $F$ einer Funktion $f$ mit $F'(x) = f(x)$ zu bestimmen, betrachten wir ein unbestimmtes Integral ohne Grenzen*

$$F(x) = \int f(x)\,\mathrm{d}x.$$

*Wir können dann mittels der Produktregel der Differentiation in Satz 5.5 eine Stammfunktion der Form $F(x) = (f \cdot g)(x)$ bestimmen als*

$$\begin{aligned} F(x) &= (f \cdot g)(x) = \int (f \cdot g' + f' \cdot g)(x)\,\mathrm{d}x \\ &= \int f(x) \cdot g'(x)\,\mathrm{d}x + \int f'(x) \cdot g(x)\,\mathrm{d}x. \end{aligned} \tag{5.2}$$

Das folgende Beispiel wendet die Regel der partiellen Integration an.

**Beispiel 5.9**
*Wir wenden die partielle Integration im Folgenden für zwei verschiedene Fälle an. Zuerst wollen wir eine Stammfunktion herleiten und danach ein Integral berechnen.*

*i) Wir wollen eine Stammfunktion des natürlichen Logarithmus $\ln\colon \mathbb{R}^+ \to \mathbb{R}$ herleiten. Wir können in diesem Fall zwei Funktionen $f$ und $g$ so definieren, dass wir die Rechenregel für partielle Integration in Satz 5.7 anwenden können. Sei also $f(x) := \ln(x)$ und $g(x) := x$ gewählt, dann gilt offensichtlich, dass $g'(x) \equiv 1$ konstant ist. Stellen wir die Formel (5.2) geeignet um und setzen diese Funktionen ein, so erhalten wir*

$$\begin{aligned} \int \ln(x) \cdot 1\,\mathrm{d}x &= \int f(x)g'(x)\,\mathrm{d}x = (f \cdot g)(x) - \int f'(x)g(x)\,\mathrm{d}x \\ &= \ln(x) \cdot x - \int \frac{1}{x} \cdot x\,\mathrm{d}x = \ln(x) \cdot x - x + C. \end{aligned}$$

*ii) Wir wollen das Integral der Arkussinus Funktion*

$$\arcsin\colon [-1, 1] \to [-\frac{\pi}{2}, \frac{\pi}{2}]$$

*in einem Intervall $[0, y]$ mit $0 < y \leqslant 1$ berechnen.*

*Wir können in diesem Fall zwei Funktionen $f$ und $g$ so definieren, dass wir die Rechenregel für partielle Integration in Satz 5.7 anwenden können. Sei also $f(x) := \arcsin(x)$ und $g(x) := x$ gewählt, dann gilt offensichtlich, dass $g'(x) \equiv 1$ konstant ist. Setzen wir dies in (5.1) ein, erhalten wir*

$$\begin{aligned} \int_0^y \arcsin(x) \cdot 1\,\mathrm{d}x &= \int_0^y f(x)g'(x)\,\mathrm{d}x = (f \cdot g)\Big|_0^y - \int_0^y f'(x)g(x)\,\mathrm{d}x \\ &= y \cdot \arcsin(y) - \int_0^y \frac{x}{\sqrt{1 - x^2}}\,\mathrm{d}x. \end{aligned}$$

*Das zweite Integral in dieser Gleichung, wollen wir zunächst so hinnehmen. Im folgenden Abschnitt werden wir eine praktische Integrationsregel, genannt* Substitutionsregel*, einführen mit der sich dieses Integral einfach berechnen lässt.*

## 5.2 Substitutionsregel

Aus der Kettenregel in Satz 5.5 können wir ein wichtiges Werkzeug für die Integralrechnung ableiten, die sogenannte *Substitutionsregel*. Die Idee ist es eine neue Integrationsvariable so geschickt einzuführen, dass ein Teil des Integranden ersetzt wird um das Integral zu vereinfachen oder auf eine bekannte Form zurückzuführen.

**Satz 5.10** (Substitutionsregel)
*Sei $I \subset \mathbb{R}$ eine Teilmenge und $[a,b] \subset \mathbb{R}$ ein Intervall. Sei außerdem $f\colon I \to \mathbb{R}$ eine integrierbare Funktion und $g\colon [a,b] \to I$ eine stetig differenzierbare Funktion. Dann gilt die folgende Substitutionsregel für die Integralrechnung:*

$$\int_a^b f(g(x)) \cdot g'(x)\,\mathrm{d}x \;=\; \int_{g(a)}^{g(b)} f(y)\,\mathrm{d}y. \tag{5.3}$$

*Beweis.* Sei $F$ eine Stammfunktion von $f$. Durch die Kettenregel für Differentiation in Satz 5.5 wissen wir, dass

$$(F \circ g)'(x) \;=\; F'(g(x)) \cdot g'(x) \;=\; f(g(x)) \cdot g'(x).$$

Damit können wir schreiben

$$\int_a^b f(g(x)) \cdot g'(x)\,\mathrm{d}x \;=\; \int_a^b (F \circ g)'(x)\,\mathrm{d}x \;=\; (F \circ g)\Big|_a^b \;=\; F\Big|_{g(a)}^{g(b)} \;=\; \int_{g(a)}^{g(b)} f(y)\,\mathrm{d}y.$$

□

**Bemerkung 5.11**
*Die Substitutionsregel für die Integralrechnung in Satz 5.10 ist besonders dann einfach, wenn die Ableitung der inneren Funktion $g(x)$ eine Konstante ist, d.h., $g'(x) \equiv c \neq 0$. Wegen der Linearität des Integrals kann man diese dann Konstante herausziehen und vor die rechte Seite von (5.3) schreiben und es ergibt sich damit folgender Zusammenhang:*

$$\int_a^b f(g(x))\,\mathrm{d}x \;=\; \frac{1}{c}\int_{g(a)}^{g(b)} f(y)\,\mathrm{d}y.$$

Das folgende Beispiel soll die Anwendung der Substitutionsregel für verschiedene Funktion mit unterschiedlichen Formen des Integrals illustrieren.

**Beispiel 5.12**
*Wir untersuchen drei Beispiele zur Anwendung der Substitutionsregel für die Integralrechnung mit aufsteigendem Schwierigkeitsgrad.*

*i) Sei $c \in \mathbb{R}, c \neq 0$, ein Skalierungsfaktor und $d \in \mathbb{R}$ eine Translationskonstante. Sei außerdem $f$ eine beliebige integrierbare Funktion auf dem Intervall $[ca+d, cb+d] \subset \mathbb{R}$. Wir wollen das folgende Integral vereinfachen:*

$$\int_a^b f(cx+d)\,\mathrm{d}x.$$

*Hierzu können wir eine einfache Substitution vornehmen:*

$$y \;:=\; g(x) \;=\; cx+d \qquad \Rightarrow \qquad g'(x) \;\equiv\; c.$$

*Wir erkennen, dass die einfache Situation aus Bemerkung 5.11 vorliegt, bei der die innere Ableitung eine Konstante ergibt.*

*Für die Substitution des Differentials* $\mathrm{d}x$ *berechnen wir:*

$$c \;=\; g'(x) \;=\; \frac{\mathrm{d}g}{\mathrm{d}x} \;=\; \frac{\mathrm{d}y}{\mathrm{d}x} \qquad \Rightarrow \qquad \mathrm{d}x \;=\; \frac{1}{g'(x)}\,\mathrm{d}y \;=\; \frac{1}{c}\,\mathrm{d}y.$$

*Damit ergibt sich für das Integral einer Funktion unter einer linearen Transformation die folgende Rechenvorschrift:*

$$\int_a^b f(cx+d)\,\mathrm{d}x \;=\; \int_a^b f(g(x))\,\mathrm{d}x \;=\; \int_{g(a)}^{g(b)} f(y)\frac{1}{c}\,\mathrm{d}y \;=\; \frac{1}{c}\int_{ca+d}^{cb+d} f(y)\,\mathrm{d}y.$$

*ii) Wir nutzen die Substitutionsregel für die Berechnung des Integrals*

$$\int_0^2 \cos(x^2+1)\cdot x\,\mathrm{d}x.$$

*Wir setzen $f(x) := \cos(x)$ und entscheiden uns für die folgende Substitution:*

$$y \;:=\; g(x) \;=\; x^2+1 \qquad \Rightarrow \qquad g'(x) \;=\; 2x.$$

*Die innere Ableitung $g'(x) = 2x$ ist leider keine Konstante, daher können wir sie nicht aus dem Integral herausziehen. Dennoch haben wir Glück, dass ein Vielfaches der Ableitung im Integranden vorkommt, nämlich der Faktor $x$, und somit können wir die Substitutionsregel für die Integralrechnung aus Satz 5.10 anwenden.*

*Für die Substitution des Differentials* $\mathrm{d}x$ *berechnen wir:*

$$2x \;=\; g'(x) \;=\; \frac{\mathrm{d}g}{\mathrm{d}x} \;=\; \frac{\mathrm{d}y}{\mathrm{d}x} \qquad \Rightarrow \qquad x\,\mathrm{d}x \;=\; \frac{1}{2}\,\mathrm{d}y.$$

*Damit ergibt sich für das Integral die folgende Rechenvorschrift:*

$$\begin{aligned}\int_0^2 \cos(x^2+1)\cdot x\,\mathrm{d}x \;&=\; \int_0^2 f(g(x))\cdot x\,\mathrm{d}x \;=\; \int_{g(0)}^{g(2)} f(y)\cdot\frac{1}{2}\,\mathrm{d}y \;=\; \frac{1}{2}\int_1^5 \cos(y)\,\mathrm{d}y \\ &=\; \frac{1}{2}\sin(y)\Big|_1^5 \;=\; \frac{1}{2}(\sin(5)-\sin(1)) \;\approx\; -0.900.\end{aligned}$$

*iii) Zuletzt wollen wir das verbliebene Integral aus Beispiel 5.9 mit Hilfe der Substitutionsregel berechnen. Wir hatten hierzu bereits hergeleitet, dass gilt*

$$\int_0^y \arcsin(x)\,\mathrm{d}x \;=\; y\cdot\arcsin(y) - \int_0^y \frac{x}{\sqrt{1-x^2}}\,\mathrm{d}x.$$

*Wir setzen* $f(x) := \frac{1}{\sqrt{1-x}}$ *und entscheiden uns für die folgende Substitution:*

$$z \;:=\; g(x) \;=\; x^2 \qquad \Rightarrow \qquad g'(x) \;=\; 2x.$$

*Die innere Ableitung* $g'(x) = 2x$ *ist leider keine Konstante, daher können wir sie nicht aus dem Integral herausziehen. Dennoch haben wir Glück, dass ein Vielfaches der Ableitung im Integranden vorkommt, nämlich der Faktor* $x$*, und somit können wir die Substitutionsregel für die Integralrechnung aus Satz 5.10 anwenden.*

*Für die Substitution des Differentials* $\mathrm{d}x$ *berechnen wir:*

$$2x \;=\; g'(x) \;=\; \frac{\mathrm{d}g}{\mathrm{d}x} \;=\; \frac{\mathrm{d}z}{\mathrm{d}x} \qquad \Rightarrow \qquad x\,\mathrm{d}x \;=\; \frac{1}{2}\,\mathrm{d}z.$$

*Damit ergibt sich für das verbliebene Integral die folgende Rechenvorschrift:*

$$\begin{aligned}\int_0^y \frac{x}{\sqrt{1-x^2}}\,\mathrm{d}x \;&=\; \int_0^y f(g(x))\cdot x\,\mathrm{d}x \;=\; \int_{g(0)}^{g(y)} f(z)\cdot\frac{1}{2}\,\mathrm{d}z \;=\; \frac{1}{2}\int_0^{y^2}\frac{1}{\sqrt{1-z}}\,\mathrm{d}z \\ &=\; \frac{1}{2}\cdot(-2\sqrt{1-z})\Big|_0^{y^2} \;=\; -(\sqrt{1-y^2}-\sqrt{1-0}) \;=\; 1-\sqrt{1-y^2}.\end{aligned}$$

*Insgesamt erhalten wir also für das Integral der Arkussinus Funktion in Beispiel 5.9:*

$$\int_0^y \arcsin(x)\,\mathrm{d}x \;=\; y\cdot\arcsin(y) - 1 + \sqrt{1-y^2}.$$

## 5.3 Integration rationaler Funktionen

Im Gegensatz zur Differentiation gibt es keine Aussage darüber, dass das Integral einer elementaren Funktion wieder eine elementare Funktion ergibt. Insbesondere gibt es keinen allgemein gültigen Algorithmus zur Bestimmung des Integrals, was die analytische Integration sehr schwierig macht. Im Spezialfall von rationalen Funktionen, die als Quotient zweier Polynome dargestellt werden können, kann man eine solche Rechenvorschrift erarbeiten.

### 5.3.1 Polynomfunktionen

Bevor wir uns mit der Integration rationaler Funktionen beschäftigen können, wollen wir im Folgenden die wichtigsten Begriffe und Erkenntnisse zu Polynomen wiederholen. Wir beginnen mit der Definition von Polynomfunktionen.

**Definition 5.13** (Polynomfunktion)
*Für den Körper $\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$ nennen wir jede Funktion $p\colon \mathbb{K} \to \mathbb{K}$ der Form*

$$p(x) \;=\; \sum_{k=0}^{n} a_k x^k \;=\; a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n \quad \textit{mit} \quad a_k \in \mathbb{K} \tag{5.4}$$

reelle *oder* komplexe Polynomfunktion *oder (ungenauerweise) häufig auch kurz* Polynom.

*Wir nennen den Koeffizienten $a_n \in \mathbb{K}$ mit dem höchsten Index den* Leitkoeffizienten *des Polynoms. Wenn der Leitkoeffizient ungleich Null ist, d.h., $a_n \neq 0$, dann sagen wir, dass $p$ ein Polynom vom* Grad $\deg(p) = n$ *ist. Ist der Leitkoeffizient gleich Eins, d.h, $a_n = 1$, so nennen wir das Polynom $p$* normiert.

*Der* Ring der Polynome $p$ *mit Koeffizienten aus dem Körper $\mathbb{K}$ wird häufig mit $\mathbb{K}[x]$ bezeichnet.*

**Bemerkung 5.14**
*Für Polynomfunktionen können wir folgende Beobachtungen festhalten:*

*i) Da die Monome der Form $1, x, x^2, \ldots$ eine Basis des Polynomrings $\mathbb{K}[x]$ (betrachtet als Vektorraum) bilden, sind Koeffizienten $a_k \in \mathbb{K}$ in (5.4) für jedes Polynom $p \in \mathbb{K}[x]$ eindeutig bestimmt.*

*ii) Allgemein soll gelten, dass die Summe zweier Polynome höchstens den maximalen Grad der beiden einzelnen Polynome hat, d.h., für zwei Polynome $f, g \in \mathbb{K}[x]$ gilt:*

$$\deg(f+g) \;\leqslant\; \max(\deg(f), \deg(g)).$$

*Der Grad von $(f+g) \in \mathbb{K}[x]$ kann sich jedoch verringern, z.B., wenn für die Polynome $\deg(f) = \deg(g) = n$ gilt und die jeweils höchsten Koeffizienten $a_n, b_n \in \mathbb{K}$ sich durch $a_n = -b_n$ aufheben.*

*Für die Multiplikation zweier Polynome stellen wir fest, dass sich der Grad des resultierenden Polynoms als Summe der Grade der einzelnen Polynome ergibt, d.h.,*

$$\deg fg \;=\; \deg(f) + \deg(f)\,, \tag{5.5}$$

*iii) Das konstante Polynom*

$$\mathbf{0}\colon \mathbb{K} \to \mathbb{K}$$

*mit $\mathbf{0}(x) = 0$ für alle $x \in \mathbb{K}$ heißt* Nullpolynom *und hat per Definition den Grad*

$$\deg(\mathbf{0}) \;:=\; -\infty.$$

*Die Festlegung des Grades des Nullpolynoms erscheint willkürlich, kann aber durch die Rechenregeln für den Grad von Polynomen in (5.5) gerechtfertigt werden, wenn man für beliebiges $k \in \mathbb{N}$ annimmt, dass $k + (-\infty) = -\infty$ gilt.*

Lassen wir als Argumente in (5.4) beliebige komplexe Zahlen zu, dann lässt sich ein reelles Polynom als komplexes Polynom auffassen mit der Einbettung $\mathbb{R}[x] \subset \mathbb{C}[x]$. Wir erinnern daran, dass reelle Polynome vom Grad $\deg(p) > 0$ keine Nullstellen besitzen müssen, im Gegensatz zu komplexen Polynomen. Der *Fundamentalsatzes der Algebra* [Fischer2005, Theorem 1.3.9] hat zur Folge, dass ein Polynom $p \in \mathbb{C}[x]$ vom Grad $\deg(p) = n$ sich in ein Produkt von Linearfaktoren, d.h., in Polynome von Grad Eins, zerlegen lässt, d.h.,

$$p(x) \;=\; \prod_{i=1}^{n}(x - x_i). \tag{5.6}$$

Die Nullstellen $x_i \in \mathbb{C}$ in der Linearfaktorzerlegung (5.6) müssen nicht zwingend paarweise verschieden sein. Tritt ein Linearfaktor $k$-mal auf so spricht man von einer $k$-fachen Nullstelle.

**Beispiel 5.15**
*Ein Beispiel für die Bedeutung der Einbettung zur Interpretation von Nullstellen ist das Polynom*

$$p(x) \;=\; x^2 + 1.$$

*Wird das Polynom $p$ über den reellen Zahlen aufgefasst, also $p \in \mathbb{R}[x]$, so besitzt $p$ keine Nullstellen. Interpretieren wir $p \in \mathbb{C}[x]$ jedoch als Polynom über den komplexen Zahlen so hat es die Nullstellen $x_1 = -i$ und $x_2 = +i$ und es gilt*

$$p(x) \;=\; x^2 + 1 \;=\; (x + i) \cdot (x - i).$$

Das folgende Lemma charakterisiert die Nullstellen eines reellen Polynoms noch genauer und besagt, dass die komplexe Konjugation einer Nullstelle selbst eine Nullstelle sein muss. Dies deckt sich mit der Aussage aus Satz 2.21, der besagt, dass zu jedem Eigenwert einer reellwertigen Matrix auch dessen komplexe Konjugation ein Eigenwert der Matrix sein muss.

**Lemma 5.16**
*Sei $p \in \mathbb{R}[x]$ ein reelles Polynom und sei $k \in \mathbb{N}$ mit $k \leqslant \deg(p)$. Ist $x_1 \in \mathbb{C}$ eine Nullstelle von $p$, dann ist auch $\overline{x}_1 \in \mathbb{C}$ eine Nullstelle von $p$.*

*Beweis.* Sei das Polynom $p \in \mathbb{R}[x] \subset C[x]$ aufgefasst über den komplexen Zahlen von der Form

$$p(x) \;=\; \sum_{k=0}^{n} a_k x^k.$$

Sei $x_1 \in \mathbb{C}$ nun eine Nullstelle von $p$, dann gilt

$$p(x_1) \;=\; \sum_{k=0}^{n} a_k x_1^k \;=\; 0.$$

Da die komplexe Konjugation verträglich mit der Addition und Multiplikation in $\mathbb{C}$ ist, können wir daraus folgern

$$\bar{0} \;=\; \overline{p(x_1)} \;=\; \overline{\sum_{k=0}^{n} a_k x_1^k} \;=\; \sum_{k=0}^{n} \overline{a_k x_1^k} \;=\; \sum_{k=0}^{n} \bar{a}_k \bar{x}_1^k$$

Da wir aber $p \in \mathbb{R}[x]$ als reelles Polynom angenommen haben gilt $\bar{a}_k = a_k$ für alle Koeffizienten. Damit folgt schon, dass

$$0 \;=\; \bar{0} \;=\; \overline{p(x_1)} \;=\; \sum_{k=0}^{n} \bar{a}_k \bar{x}_1^k \;=\; \sum_{k=0}^{n} a_k \bar{x}_1^k \;=\; p(\bar{x}_1).$$

Also haben wir gezeigt, dass $\bar{x}_1 \in \mathbb{C}$ auch eine Nullstelle von $p$ ist. □

Eine Zerlegung in Linearfaktoren wie in (5.6) können wir für reelle Polynome im Allgemeinen nicht erwarten. Dennoch lässt sich der folgende Satz für die Zerlegung reeller Polynom in eindeutige reelle Linearfaktoren und quadratische Polynome formulieren.

**Satz 5.17** (Faktorisierungssatz für reelle Polynome)
*Sei $p \in \mathbb{R}[x]$ ein reelles, normiertes Polynom mit Grad $\deg(p) = n$. Dann existieren Koeffizienten $a_1, \ldots, a_l, b_1, \ldots, b_l, c_1, \ldots, c_m \in \mathbb{R}$ mit*

$$2l + m \;=\; n, \qquad a_i^2 - b_i \;<\; 0,$$

*so dass sich das Polynom $p$ faktorisieren lässt in die folgende Form*

$$p(x) \;=\; \prod_{i=1}^{l} (x^2 - 2a_i x + b_i) \;\cdot\; \prod_{j=1}^{m} (x - c_j)\,. \tag{5.7}$$

*Beweis.* Nach dem Fundamentalsatz der Algebra können wir $p$ in Linearfaktoren zerlegen mit

$$p(x) \;=\; \prod_{i=1}^{n} (x - x_i) \;=\; \prod_{i=1}^{m} (x - x_i) \;\cdot\; \prod_{i=m+1}^{n} (x - x_i),$$

wobei $x_1, \ldots, x_m \in \mathbb{R}$ die reellen Nullstellen und $x_{m+1}, \ldots, x_n \in \mathbb{C}\backslash\mathbb{R}$ die echt komplexen Nullstellen von $p$ sind. Wir setzen einen neuen Index

$$l \;\coloneqq\; \frac{n-m}{2} \in \mathbb{N}_0$$

und wissen nach Lemma 5.16, dass für jede komplexe Nullstelle $x_i \in \mathbb{C}, i = m+1, \ldots, n$, von $p$ eine komplex-konjugierte Nullstelle $\bar{x}_i \in \mathbb{C}$ von $p$ existiert. Durch eine geeignete Umnummerierung der Indizes können wir also schreiben:

$$\prod_{i=m+1}^{n} (x - x_i) \;=\; \prod_{i=m+1}^{m+l} (x - x_i) \cdot (x - \bar{x}_i)\,.$$

Um die Form in (5.7) zu erhalten setzen wir für die Linearfaktoren

$$c_i \coloneqq x_i, \qquad i = 1, \ldots, m.$$

Für die quadratischen Polynome setzen wir schließlich

$$a_i \coloneqq \mathrm{Re}(x_{i+m}) \quad \text{und} \quad b_i \coloneqq x_{i+m} \cdot \bar{x}_{i+m} = |x_{i+m}|^2, \qquad i = 1 \ldots, l.$$

Es ist klar, dass $a_i^2 - b_i < 0$ für die quadratischen Polynome mit komplexen Nullstellen gelten muss, da dieser Term gerade den Radikand in der Formel zur Berechnung von Nullstellen $x_{1/2} \in \mathbb{C}$ normierter, quadratischer Polynome der Form $q(x) = x^2 - 2ax + b$ darstellt mit

$$x_{1/2} = a \pm \sqrt{a^2 - b}.$$

Falls $a_i^2 - b_i \geqslant 0$ wäre, so ließe sich das Polynom $q$ in zwei Linearfaktoren mit reellen Nullen faktorisieren. □

Wir wollen die Aussage aus Satz 5.17 im folgenden Beispiel mit konkreten Faktorisierungen reeller Polynome illustrieren.

**Beispiel 5.18**
*Wir interessieren uns für eine Faktorisierung verschiedener reeller Polynome $p \in \mathbb{R}[x]$ in lineare und quadratische Polynome.*

1. *Sei $p(x) = x^4 - 1$. Wir bezeichnen mit*

$$\zeta_k \coloneqq \exp\left(\frac{2\pi i k}{4}\right), \quad k = 0, \ldots, 3$$

*die 4. Einheitswurzeln. Dann lässt sich das Polynom $p$ wie folgt faktorisieren:*

$$p(x) = x^4 - 1 = \prod_{k=0}^{4} (x - \zeta_k) = (x-1)\cdot(x+1)\cdot(x-i)\cdot(x+i) = (x-1)\cdot(x+1)\cdot(x^2+1)$$

2. *Sei $p(x) = x^5 - 1$. Wir bezeichnen mit*

$$\zeta_k \coloneqq \exp\left(\frac{2\pi i k}{5}\right), \quad k = 0, \ldots, 4$$

*die 5. Einheitswurzeln. Dann lässt sich das Polynom $p$ wie folgt faktorisieren:*

$$p(x) = x^5 - 1 = \prod_{k=1}^{4} (x - \zeta_k) = (x-1) \cdot (x^2 - 2\,\mathrm{Re}(\zeta_1) + 1) \cdot (x^2 - 2\,\mathrm{Re}(\zeta_2) + 1)$$

3. *Sei $p(x) = x^4 - 6x^2 + 7x - 6$ und sei $c = \sqrt{-\frac{3}{4}} \in \mathbb{C}$. Dann lässt sich das Polynom $p$ wie folgt faktorisieren:*

$$\begin{aligned} p(x) &= x^4 - 6x^2 + 7x - 6 = (x-2) \cdot (x+3) \cdot (x + \frac{1}{2} + c) \cdot (x + \frac{1}{2} - c) \\ &= (x-2) \cdot (x+3) \cdot (x^2 - x + 1). \end{aligned}$$

### 5.3.2 Rationale Funktionen

Nachdem wir die wichtigsten Erkenntnisse und Begriffe für Polynomfunktionen wiederholt haben können wir uns im Folgenden mit den rationalen Funktionen beschäftigen, die als Quotient von Polynomen definiert sind.

**Definition 5.19** (Rationale Funktion)
*Seien $p, q \in \mathbb{K}[x]$ zwei Polynome mit $\deg(p) = n$ und $\deg(q) = m$ für $n, m \in \mathbb{N}$. Dann bezeichnen wir den Quotienten der beiden Polynome*

$$r(x) \;:=\; \frac{p(x)}{q(x)} \;=\; \frac{\sum_{k=0}^{n} a_k x^k}{\sum_{k=0}^{m} b_k x^k}$$

*als* rationale Funktion. *Hierbei ist es wichtig, dass wir den Definitionsbereich von $r \in \mathbb{K}[x]$ einschränken auf die $x \in \mathbb{K}$ für die $q(x) \neq 0$.*

*Enthält der Zähler von $r$ ein Polynom, dass einen echt kleineren Grad als der Nenner besitzt, d.h., $n < m$, so sprechen wir von einer* echt gebrochenrationalen *Funktion. Ist das Gegenteil der Fall, d.h., $n \geqslant m$, so sprechen wir von einer* unecht gebrochenrationalen *Funktion. Ist $r$ ein Polynom, d.h. es gilt $m = 0$, so sprechen wir von einer* ganzrationalen *Funktion.*

**Bemerkung 5.20**
*Wir wollen folgende Bemerkungen zu rationalen Funktionen festhalten.*

1. *Rationale Funktionen können als Elemente eines Funktionskörpers $\mathbb{K}(x)$ aufgefasst werden, der über einem Körper $\mathbb{K}$ definiert ist, da jedes Element $r = \frac{p}{q}$ nun auch ein multiplikativ-inverses Element $r^{-1} = \frac{q}{p}$ besitzt im Gegensatz zu Elementen des Polynomrings $\mathbb{K}[x]$. Dafür muss jedoch der Definitionsbereich in $D \subset \mathbb{K}$ so eingeschränkt werden, dass $q(x) \neq 0$ für alle $x \in D$. Insbesondere gilt $\mathbb{K}[x] \subset \mathbb{K}(x)$.*

2. *Rationale Funktionen sind nicht eindeutig, da sowohl Zähler als auch Nenner mit einem beliebigen Polynom $s \in \mathbb{K}[x], s \neq \mathbf{0}$, erweitert werden können. Jedoch lässt sich jede rationale Funktion $r \in \mathbb{K}(x)$ mittels* Polynomdivision *[Burger2020][Kapitel 1.4] in eine eindeutige Summe aus einer ganzrationalen Funktion $g \in \mathbb{K}[x]$ (also einem Polynom) und einer echt gebrochenrationalen Funktion $\tilde{r}$ schreiben, d.h.*

$$r \;=\; \frac{p}{q} \;=\; g + \frac{\tilde{p}}{q} \;=\; g + \tilde{r} \qquad \textit{mit} \qquad \deg(\tilde{p}) < \deg(q).$$

Wir wollen die letzte Bemerkung zusätzlich an Hand eines Beispiels mittels Polynomdivision nachvollziehen.

**Beispiel 5.21** (Polynomdivision)
*Sei $r \in \mathbb{K}(x)$ eine rationale Funktion mit*

$$r(x) \;=\; \frac{p(x)}{q(x)} \;=\; \frac{x^4}{x^3 + 2x^2 - 1}.$$

*Dann können wir $r$ schreiben als eine Summe eines Polynoms und einer echt gebrochenrationalen Funktion und es gilt:*

$$\frac{x^4}{x^3+2x^2-1} = x - 2 + \frac{4x^2+x-2}{x^3+2x^2-1},$$

*Hierzu führen wir eine Polynomdivision mit dem Euklidischen Algorithmus durch:*

$$\begin{array}{rr} \left(\quad x^4\right) : \left(x^3+2x^2-1\right) = x-2+\dfrac{4x^2+x-2}{x^3+2x^2-1} \\ \underline{-x^4-2x^3 \qquad +x} \\ \quad -2x^3 \qquad +x \\ \underline{\quad 2x^3+4x^2 \qquad -2} \\ \qquad\qquad 4x^2+x-2 \end{array}$$

### 5.3.3 Partialbruchzerlegung rationaler Funktionen

Wir wissen also nun, dass wir jede rationale Funktion $r \in \mathbb{K}(x)$ darstellen können als

$$r \;=\; g + \frac{\tilde{p}}{q} \qquad \text{mit} \qquad \deg(\tilde{p}) \;<\; \deg(q).$$

Die Polynome $g \in \mathbb{R}[x]$ können wir offensichtlich sehr einfach integrieren jedoch müssen wir uns noch um die Bestimmung des unbestimmten Integrals für echt gebrochenrationale Funktionen $\frac{\tilde{p}}{q} \in \mathbb{K}(x)$ kümmern. Ein sehr hilfreiches Werkzeug hierbei stellt die sogenannte *Partialbruchzerlegung* dar, die es ermöglicht rationale Funktionen in eine Summe rationaler Funktionen von einfacher Gestalt zu zerlegen. Wird die Ausgangsfunktion als komplexe rationale Funktion aus $\mathbb{C}(x)$ betrachtet, so besteht die Zerlegung sogar aus Summanden deren Nennerpolynom maximal eine Nullstelle besitzt.

Wir wollen zunächst den folgenden Satz für die Partialbruchzerlegung für den allgemeinen Fall $\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$ formulieren. In dieser Formulierung sagt der Satz aus, dass wir Nullstellen des Nennerpolynoms von echt gebrochenrationalen Funktionen aus dem Bruch herausziehen können, um somit den Grad des Nennerpolynoms zu verringern und dadurch einfachere Terme zu generieren.

**Satz 5.22** (Partialbruchzerlegung)
*Seien $p, q \in \mathbb{K}[x]$ beliebige Polynome mit $q \not\equiv \mathbf{0}$. Sei außerdem $x_1 \in \mathbb{K}$ eine $k$-fache Nullstelle von $q$ mit $k \in \mathbb{N}$. Falls die rationale Funktion $r \in \mathbb{K}(x)$ mit*

$$r(x) \;:=\; \frac{p(x)}{q(x)} \;=\; \frac{p(x)}{\tilde{q}(x)\cdot(x-x_1)^k}$$

*echt gebrochenrational ist, d.h., $\deg(p) < \deg(q)$, dann existiert ein eindeutig bestimmtes Polynom $\tilde{p} \in \mathbb{K}[x]$ und Skalare $A_1, \ldots, A_k \in \mathbb{K}$, so dass die folgende Partialbruchzerlegung*

*gilt*

$$r(x) \;=\; \frac{p(x)}{\tilde{q}(x)(x-x_1)^k} \;=\; \frac{\tilde{p}(x)}{\tilde{q}(x)} + \sum_{j=1}^{k} \frac{A_j}{(x-x_1)^j} \qquad \textit{mit} \qquad \deg(\tilde{p}) < \deg(\tilde{q}).$$

*Beweis.* Da $x_1 \in \mathbb{K}$ eine $k$-fache Nullstelle des Nennerpolynoms ist, kann man rekursiv die Vielfachheit dieser Nullstelle reduzieren, was wir im Folgenden für den ersten Schritt zeigen werden. Dazu müssen wir zeigen, dass

$$r(x) \;=\; \frac{p(x)}{q(x)} \;=\; \frac{p(x)}{\tilde{q}(x)(x-x_1)^k} \;\stackrel{!}{=}\; \frac{p_k(x)}{\tilde{q}(x)(x-x_1)^{k-1}} + \frac{A_k}{(x-x_1)^k}, \tag{5.8}$$

wobei $p_k \in \mathbb{K}[x]$ ein eindeutig bestimmtes Polynom vom Grad $\deg(p_k) < \deg(\tilde{q}) + k - 1$ ist und $A_k \in \mathbb{K}$ ein eindeutig bestimmtes Skalar.

Durch Multiplikation beider Seiten der Gleichung (5.8) mit $(x-x_1)^k$ erhalten wir

$$\frac{p(x)}{\tilde{q}(x)} \;\stackrel{!}{=}\; \frac{p_k(x)(x-x_1)}{\tilde{q}(x)} + A_k. \tag{5.9}$$

Setzen wir nun $x = x_1$ in diese Gleichung ein, so ergibt sich für das unbekannte Skalar

$$A_k \;=\; \frac{p(x_1)}{\tilde{q}(x_1)} \in \mathbb{K}.$$

Stellen wir jedoch (5.9) nach der unbekannten Funktion $p_k \in \mathbb{K}(x)$ um, so erhalten wir

$$p_k(x) \;=\; \frac{p(x) - A_k\tilde{q}(x)}{x-x_1} \;=\; \frac{p(x) \;-\; \frac{p(x_1)}{\tilde{q}(x_1)} \cdot \tilde{q}(x)}{x-x_1}.$$

Es wird klar, dass $p_k \in \mathbb{K}[x]$ ein Polynom ist, da das Zählerpolynom $(p(x) - A_k\tilde{q}(x)) \in \mathbb{K}[x]$ offensichtlich eine Nullstelle in $x = x_1$ besitzt und sich somit ein Linearfaktor $(x - x_1)$ abspalten lässt.

Abschließend lässt sich noch Folgendes feststellen. Da nach Voraussetzung

$$\deg(p) \;<\; \deg(q) \;=\; \deg(\tilde{q}) + k$$

galt, ist wegen Bemerkung 5.14 auch

$$\deg(p - A_k\tilde{q}) \;<\; \deg(\tilde{q}) + k.$$

Aus der Gestalt des Polynoms $p_k$ wissen wir daher, dass

$$\deg(p_k) < \deg(q) + k - 1$$

gilt. Wir haben also ein eindeutiges Skalar $A_k \in \mathbb{K}$ und ein eindeutiges Polynom $p_k \in \mathbb{K}[x]$ gefunden, so dass die Zerlegung in (5.8) gilt. Auf die übrigbleibende echt gebrochenrationale Funktion können wir also rekursiv den obigen Ansatz anwenden. □

Wenn der zu Grunde liegende Körper $\mathbb{K}$ explizit gewählt wird, so lässt sich die Aussage aus Satz 5.22 noch weiter präzisieren, wie folgende Bemerkung festhält.

**Bemerkung 5.23**
*Im Folgenden wollen wir für die spezielle Wahl des Körpers $\mathbb{K}$ erklären, wie sich eine echt gebrochenrationale Funktion in möglichst einfache Summanden durch mehrfache Anwendung der Partialbruchzerlegung zerlegen lässt.*

1. *Lässt sich das Nennerpolynom $q$ der echt gebrochenrationalen Funktion $r \in \mathbb{K}(x)$ in Satz 5.22 in Linearfaktoren zerlegen (was für alle Polynome $q \in \mathbb{C}[x]$ der Fall ist nach dem Fundamentalsatz der Algebra), d.h., es besitzt die Form*

$$q(x) \;=\; \prod_{i=1}^{m}(x - x_i)^{d_i},$$

*und sind die Nullstellen $x_i \in \mathbb{K}, i = 1, \ldots, m$ der Vielfachheit $k_i \in \mathbb{N}$ paarweise verschieden, dann führt mehrfache Anwendung der Partialbruchzerlegung in Satz 5.22 zu folgender Formel:*

$$\frac{p(x)}{q(x)} \;=\; \sum_{i=1}^{m}\sum_{j=1}^{k_i}\frac{A_j^{(i)}}{(x - x_i)^j}\,. \tag{5.10}$$

*Die Anzahl der zu bestimmenden Koeffizienten $A_j^{(i)}$ ist in diesem Fall $\sum\limits_{i=1}^{m} k_i = \deg(q)$.*

2. *Betrachtet man lediglich reelle Polynome $p, q \in \mathbb{R}[x]$, so wissen wir aus dem Faktorisierungssatz 5.17 für reelle Polynome, dass sich das Nennerpolynom in ein Produkt aus $\hat{k} \in \mathbb{N}$ Linearfaktoren und $\hat{l} \in \mathbb{N}$ quadratischen Polynomen mit je zwei echt komplexen Nullstellen faktorisieren lässt, d.h.,*

$$q(x) \;=\; \prod_{i=1}^{\hat{l}}(x^2 - 2a_i x + b_i) \;\cdot\; \prod_{j=1}^{\hat{k}}(x - c_j).$$

*Das bedeutet, dass $q \in \mathbb{R}[x]$ insgesamt $m = \hat{k} + 2\hat{l}$ Nullstellen besitzt. Wir nehmen an, dass $k \in \mathbb{N}, k \leqslant \hat{k}$ reelle Nullstellen paarweise verschieden sind und außerdem noch $l \in \mathbb{N}, l \leqslant \hat{l}$ Paare von paarweise verschiedenen, echt komplexen Nullstellen von $q$ existieren. Seien nun $s_1, \ldots, s_k \in \mathbb{N}$ die Vielfachheiten der reellen Nullstellen und $t_1, \ldots, t_l \in \mathbb{N}$ die Vielfachheiten der komplexen Nullstellen. Dann können wir das Polynom $q \in \mathbb{R}[x]$ schreiben als die folgende Faktorisierung*

$$q(x) \;=\; \prod_{i=1}^{l}(x^2 - 2a_i x + b_i)^{t_i} \;\cdot\; \prod_{j=1}^{k}(x - c_j)^{s_j}.$$

*Die mehrfache Anwendung der Partialbruchzerlegung in Satz 5.22 führt dann für reelle echt gebrochenrationale Funktionen zu folgender Formel:*

$$\frac{p(x)}{q(x)} \;=\; \sum_{i=1}^{k}\sum_{j=1}^{s_i}\frac{A_j^{(i)}}{(x-x_i)^j} \;+\; \sum_{i=1}^{l}\sum_{j=1}^{t_i}\frac{B_j^{(i)}x+C_j^{(i)}}{(x^2-2ax+b)^j}. \tag{5.11}$$

*Die Anzahl der zu bestimmenden Koeffizienten $A_j^{(i)}$ für die Linearfaktoren ist in diesem Fall $\sum_{i=1}^{k} s_i = \hat{k}$. Die Anzahl der zu bestimmenden Koeffizienten $B_j^{(i)}$ und $C_j^{(i)}$ für die quadratischen Polynome ist in diesem Fall jeweils $\sum_{i=1}^{l} t_i = \hat{l}$. Insgesamt müssen also $\hat{k}+2\hat{l} = m = \deg(q)$ unbekannte Koeffizienten bestimmt werden.*

3. *Nach Multiplikation beider Seiten von* (5.10) *und* (5.11) *mit dem Nennerpolynom $q$ lassen sich die unbekannten Koeffizienten durch Koeffizientenvergleich der Polynome auf beiden Seiten berechnen. Dies lässt sich am einfachsten durch die Lösung eines linearen Gleichungssystems realisieren.*

Basierend auf den Beobachtungen oben können wir nun informell einen Algorithmus für die Partialbruchzerlegung einer rationalen Funktion angeben.

**Algorithmus 5.24** (Partialbruchzerlegung)
*Sei zu Anfang eine rationale Funktion $r \in \mathbb{K}(x)$ gegeben mit*

$$r(x) \;:=\; \frac{p(x)}{q(x)}$$

*mit zwei Polynomen $p, q \in \mathbb{K}[x]$.*

***1. Schritt: Überführung von $r$ in eindeutige Darstellung***

*Ist die rationale Funktion $r$ bereits echt gebrochenrational, d.h. $\deg(q) > \deg(p)$ so ist in diesem Schritt nichts zu tun. Andernfalls führen wir eine Polynomdivision durch um $r$ in die eindeutige Darstellung*

$$r \;=\; g + \frac{\tilde{p}}{q}$$

*zu überführen, wobei $g \in \mathbb{K}[x]$ ein Polynom ist und $\deg(q) > \deg(\tilde{p})$ gilt.*

***2. Schritt: Bestimmung der Nullstellen des Nennerpolynoms***

*Wir bestimmen nun die Nullstellen des Nennerpolynoms $q \in \mathbb{K}[x]$ analytisch, d.h., wir versuchen das Polynom in die in* (5.6) *oder* (5.7) *beschriebene Form zu faktorisieren (in Abhängigkeit des Körpers $\mathbb{K}$). Ist eine analytische Bestimmung der Nullstellen von $q$ nicht möglich, können numerische Methoden angewendet werden zur Approximation der Nullstellen.*

***3. Schritt: Bestimmung der Partialbruchzerlegung***

*Basierend auf den im 2. Schritt bestimmten Nullstellen von $q$ kann eine Partialbruchzerlegung der Form* (5.10) *oder* (5.11) *durchgeführt werden (in Abhängigkeit des Körpers $\mathbb{K}$).*

***4. Schritt: Aufstellung des linearen Gleichungssystems***

*Durch Multiplikation beider Seiten von* (5.10) *oder* (5.11) *mit dem Nennerpolynom $q \in \mathbb{K}[x]$ kann ein Koeffizientenvergleich durchgeführt werden, der zu einem linearen Gleichungssystem mit einer Koeffizientenmatrix vom Rang* $\deg(q)$ *führt.*

***5. Schritt: Lösung des linearen Gleichungssystems***

*Die Lösung des entstehenden linearen Gleichungssystems kann entweder algebraisch oder numerisch erfolgen (basierend auf der Gestalt und Größe der Koeffizientenmatrix) und liefert die eindeutig bestimmten Koeffizienten $A_j^{(i)}, B_j^{(i)}, C_j^{(i)} \in \mathbb{K}$ der Partialbruchzerlegung.*

*Damit haben wir die rationale Funktion $r \in \mathbb{K}(x)$ in eine Summe einfacherer Terme durch Partialbruchzerlegung überführt.*

Wir wollen den Algorithmus 5.24 für die Partialbruchzerlegung an einem konkreten Beispiel anwenden und somit besser verdeutlichen.

**Beispiel 5.25** (Partialbruchzerlegung)
*Wir wollen eine Partialbruchzerlegung für die rationale Funktion $r \in \mathbb{R}(x)$ durchführen mit*

$$r(x) \;:=\; \frac{2x^3 + 4x}{x^4 - 2x^2 + 1}.$$

*Hierzu wenden wir Algorithmus 5.24 schrittweise an.*

***1. Schritt: Überführung von $r$ in eindeutige Darstellung***

*Wir bemerken, dass die rationale Funktion $r \in \mathbb{R}(x)$ bereits echt gebrochenrational ist und wir deshalb keine Polynomdivision durchführen müssen.*

***2. Schritt: Bestimmung der Nullstellen des Nennerpolynoms***

*Durch Ausprobieren stellen wir fest, dass das Nennerpolynom $q \in \mathbb{R}[x]$ Nullstellen in $x_1 = 1$ und $x_2 = -1$ besitzt. Faktorisieren wir diese Nullstellen heraus erhalten wir den Ausdruck*

$$q(x) \;=\; x^4 - 2x^2 + 1 \;=\; (x-1)\cdot(x+1)\cdot(x^2-1).$$

*Die Nullstellen des übrig gebliebenen quadratischen Polynoms bestimmen wir durch Ausklammern (oder mittels p-q-Formel) und erhalten die folgende Faktorisierung von $q$ in zwei Linearfaktoren der Vielfachheit zwei:*

$$q(x) \;=\; x^4 - 2x^2 + 1 \;=\; (x-1)\cdot(x+1)\cdot(x^2-1) \;=\; (x-1)^2\cdot(x+1)^2.$$

***3. Schritt: Bestimmung der Partialbruchzerlegung***

*Da wir Glück haben und sich das Nennerpolynom in Linearfaktoren über $\mathbb{R}$ faktorisieren lässt wie in* (5.6)*, können wir die allgemeine Darstellung der Partialbruchzerlegung in* (5.10) *anwenden:*

$$\frac{2x^3+4x}{(x+1)^2(x-1)^2} = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{A_j^{(i)}}{(x-x_i)^j} = \frac{A_1^{(1)}}{x+1} + \frac{A_2^{(1)}}{(x+1)^2} + \frac{A_1^{(2)}}{x-1} + \frac{A_2^{(2)}}{(x-1)^2}. \quad (5.12)$$

***4. Schritt: Aufstellung des linearen Gleichungssystems***

*Durch Multiplikation beider Seiten der Partialbruchzerlegung in* (5.12) *mit dem Nennerpolynom $q(x) = (x-1)^2(x+1)^2$ ergibt dann:*

$$\begin{aligned} 2x^3+4x &= A_1^{(1)}(x+1)(x-1)^2 + A_2^{(1)}(x-1)^2 + A_1^{(2)}(x+1)^2(x-1) + A_2^{(2)}(x+1)^2 \\ &= \left(A_1^{(1)} + A_1^{(2)}\right)x^3 + \left(-A_1^{(1)} + A_2^{(1)} + A_1^{(2)} + A_2^{(2)}\right)x^2 \\ &\quad + \left(-A_1^{(1)} - 2A_2^{(1)} - A_1^{(2)} + 2A_2^{(2)}\right)x + A_1^{(1)} + A_2^{(1)} + A_1^{(2)} + A_2^{(2)}\,. \end{aligned}$$

*Durch Koeffizientenvergleich erhalten wir das folgende lineare Gleichungssystem*

$$\begin{pmatrix} 1 & 0 & 1 & 0 \\ -1 & 1 & 1 & 1 \\ -1 & -2 & -1 & 2 \\ 1 & 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} A_1^{(1)} \\ A_2^{(1)} \\ A_1^{(2)} \\ A_2^{(2)} \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ 4 \\ 0 \end{pmatrix}.$$

***5. Schritt: Lösung des linearen Gleichungssystems***

*Mittels Gauß'schem Eliminationsverfahren können wir die eindeutige Lösung für die Koeffizienten bestimmen als*

$$A_1^{(1)} = 1, \quad A_2^{(1)} = -\frac{3}{2}, \quad A_1^{(2)} = 1, \quad A_2^{(2)} = \frac{3}{2}.$$

*Insgesamt erhalten wir also die folgende Partialbruchzerlegung*

$$\frac{2x^3+4x}{(x+1)^2(x-1)^2} = \frac{1}{x+1} + \frac{-\frac{3}{2}}{(x+1)^2} + \frac{1}{x-1} + \frac{\frac{3}{2}}{(x-1)^2}.$$

### 5.3.4 Stammfunktionen rationaler Funktionen

Mit Hilfe des Algorithmus 5.24 für die Partialbruchzerlegung können wir also beliebige rationale Funktionen in eine Summe aus wesentlich einfacheren Termen zerlegen. Nun müssen

wir nur noch festhalten, wie die Stammfunktionen dieser einfachen Summanden aussehen, damit wir rationale Funktionen integrieren können. Hierzu hält die folgende Bemerkung die relevanten Ergebnisse fest.

**Bemerkung 5.26**
*Wir erwarten als Resultat der Partialbruchzerlegung im Allgemeinen zwei verschiedene Arten von Summanden. Für diese zwei Arten werden wir im Folgenden Stammfunktionen angeben, so dass sich die Stammfunktion von beliebigen rationalen Funktionen bestimmen lässt. Auf eine explizite Herleitung der Stammfunktionen wird verzichtet, da sich die Resultate durch die Rechenregeln der Differentiation direkt verifizieren lassen.*

1. *Die erste Art von Summanden der Partialbruchzerlegung ist von der Gestalt*

$$s(x) \;=\; \frac{A}{(x-a)^j}, \quad j \in \mathbb{N}, A \in \mathbb{K}$$

*und entsteht durch das Ausklammern von Linearfaktoren aus dem Nennerpolynom $q \in \mathbb{K}[x]$ für eine Nullstelle $a \in \mathbb{K}$. Da das Skalar $A \in \mathbb{K}$ nicht von der Variable $x$ abhängt, lässt es sich linear aus dem Integral herausziehen und somit können wir die Stammfunktion direkt angeben als*

$$\int s(x)\,\mathrm{d}x \;=\; A\int \frac{1}{(x-a)^j}\,\mathrm{d}x \;=\; \begin{cases} A\cdot\ln|x-a| & \text{für } j=1\,, \\ \frac{-A}{(j-1)(x-a)^{j-1}} & \text{für } j>1\,. \end{cases}$$

2. *Die zweite Art von Summanden der Partialbruchzerlegung ist von der Gestalt*

$$s(x) \;=\; \frac{Bx+C}{(x^2-2ax+b)^j}, \quad j \in \mathbb{N},\ A, B \in \mathbb{K}$$

*und entsteht im Fall von reellen rationalen Funktionen durch das Ausklammern von quadratischen Polynomen aus dem Nennerpolynom $q \in \mathbb{R}[x]$, die je ein Paar von echt komplexe Nullstellen besitzen (da in diesem Fall $a^2 < b$ gilt). Um eine Stammfunktion für $s \in \mathbb{R}(x)$ zu bestimmen formen wir die Funktion zuerst geeignet um, so dass wir im Zähler ein Vielfaches des linearen Terms $(x-a)$ erhalten:*

$$\frac{Bx+C}{(x^2-2ax+b)^j} \;=\; \frac{B\cdot(x-a)}{(x^2-2ax+b)^j} + \frac{C+B\cdot a}{(x^2-2ax+b)^j}, \quad j \in \mathbb{N},\ B, C \in \mathbb{K} \quad (5.13)$$

*Für den linken Summanden in (5.13) können wir folgende Stammfunktion angeben:*

$$B\int \frac{x-a}{(x^2-2ax+b)^j}\,\mathrm{d}x \;=\; \begin{cases} \frac{B}{2}\ln(x^2-2ax+b)\,, & \text{für } j=1 \\ -\frac{B}{2(j-1)(x^2-2ax+b)^{j-1}}\,, & \text{für } j>1\,. \end{cases}$$

*Für den rechten Summanden in* (5.13) *können wir folgende Stammfunktion rekursiv angeben:*

$$
\begin{aligned}
I_j &:= \underbrace{(C + B \cdot a)}_{=:D} \int \frac{1}{(x^2 - 2ax + b)^j} \, \mathrm{d}x \\
&= \begin{cases} \frac{D}{\sqrt{b-a^2}} \cdot \arctan\left(\frac{x-a}{\sqrt{b-a^2}}\right), & \text{für } j = 1, \\ D \cdot \left(\frac{x-a}{2(b-a^2)(j-1)(x^2-2ax+b)^{j-1}} + \frac{2j-3}{2(b-a^2)(j-1)} I_{j-1}\right), & \text{für } j > 1. \end{cases}
\end{aligned}
$$

Glücklicherweise sind moderne Computerprogramme in der Lage die oben beschriebenen Stammfunktionen analytisch zu bestimmen. Abschließend wollen wir im folgenden Beispiel die Stammfunktion einer einfachen rationalen Funktion mit Hilfe der Aussagen aus Bemerkung 5.26 ausrechnen.

**Beispiel 5.27** (Integration einer rationalen Funktion)
*Sei $s \in \mathbb{R}(x)$ eine rationale Funktion, deren Nennerpolynom quadratisch ist mit*

$$s(x) := \frac{x+2}{x^2 - 4x + 7}.$$

*Zur Bestimmung einer Stammfunktion von $s$ sehen wir ein, dass die Koeffizienten $a = 2$ und $b = 7$ im Nennerpolynom sind. Außerdem ist $B = 1, C = 2$ und $j = 1$. Mit diesen Informationen formen wir den Zähler wie in* (5.13) *um und erhalten als Stammfunktion*

$$
\begin{aligned}
\int \frac{x+2}{x^2 - 4x + 7} \, \mathrm{d}x &= \int \frac{x-2}{x^2 - 4x + 7} \, \mathrm{d}x + 4 \int \frac{1}{x^2 - 4x + 7} \, \mathrm{d}x \\
&= \frac{1}{2} \ln(x^2 - 4x + 7) + \frac{4}{\sqrt{3}} \arctan\left(\frac{x-2}{\sqrt{3}}\right).
\end{aligned}
$$

# Kapitel 6

# Differentiation von Funktionen mehrerer Veränderlicher

In diesem Kapitel wollen wir uns der Frage widmen, wie sich das Konzept der Differentiation von Funktionen in einer Variablen auf Funktionen in mehreren Variablen übertragen lässt um zu verstehen wie sich Änderungen in den veränderlichen Eingabeparametern auf die Funktionswerte dieser Funktion auswirken. Insbesondere wollen wir verstehen welche verschiedenen Ableitungsbegriffe für mehrdimensionaler Funktionen $f : \mathbb{R}^n \supset U \to \mathbb{R}$ existieren und welche Eigenschaften sie besitzen. Darüber hinaus lernen wir wichtige Sätze der Differentialrechnung kennen. Die Erkenntnisse dieses Kapitels spielen eine entscheidende Rolle in der Numerik, der mathematischen Modellierung, oder der Optimierung.

## 6.1 Partielle Differenzierbarkeit

Um den Begriff der Ableitung auf Funktionen mehrerer Variablen zu übertragen, betrachtet man zunächst deren Änderungsverhalten entlang einzelner Koordinatenrichtungen. Wir fixieren also einen Punkt $x \in U$ und betrachten die Funktion

$$\tilde{f}(\xi) \;:=\; f(x_1, \ldots, x_{i-1}, \xi, x_{i+1}, \ldots, x_n).$$

Anschaulich gesprochen halten wir alle Variablen, bis auf die $i$-te fest und erhalten somit eine eindimensionale Funktion. Dies erlaubt es uns den Ableitungsbegriff auf die schon bekannte Definition für Funktionen in einer Veränderlichen herunterzubrechen, was zum Begriff der partiellen Differenzierbarkeit führt.

**Definition 6.1** (Partielle Differenzierbarkeit)
*Sei $U \subset \mathbb{R}^n$ eine offene Teilmenge und $f : U \to \mathbb{R}$ eine Funktion.*

*i) Wir nennen die Funktion $f$* (lokal) partiell differenzierbar *im Punkt $x \in U$ in die $i$-te*

*Koordinatenrichtung falls der Grenzwert*

$$\frac{\partial f}{\partial x_i}(x) \;:=\; \partial_i f(x) \;:=\; \lim_{h\to 0}\frac{f(x+h\cdot e_i)-f(x)}{h}$$

*existiert.*

*ii) Wir nennen die Funktion $f$* (global) partiell differenzierbar *falls sie für jedes $x \in U$ und jede Koordinatenrichtung $i \in \{1,\dots,n\}$ partiell differenzierbar ist.*

*iii) Ist die Funktion $\partial_i f : U \to \mathbb{R}$ partiell differenzierbar und zudem stetig für jedes $i \in \{1,\dots,n\}$, so nennen wir $f$* stetig partiell differenzierbar.

In der obigen Definition bezeichnet

$$e_i \;:=\; (0,\dots,0,\underbrace{1}_{i\text{-te Stelle}},0,\dots,0)\in\mathbb{R}^n$$

den $i$-ten Einheitsvektor des $\mathbb{R}^n$. Im Grenzwert für $h \to 0$ sind nur Nullfolgen $(h_k)_{k\in\mathbb{N}}$ erlaubt, welche die Bedingung

$$(x+h_k\cdot e_i)\in U \quad \text{mit} \quad h_k\neq 0,\ \forall k\in\mathbb{N}$$

erfüllen. Das Symbol $\partial$ ist hierbei eine stilisierte Version des Buchstaben d (manchmal auch „del“ ausgesprochen).

Im folgenden Beispiel berechnen wir die partielle Ableitung der Euklidischen Norm im $\mathbb{R}^n$.

**Beispiel 6.2** (Ableitung der Euklidischen Norm)
*Wir betrachten die Abbildung*

$$\begin{aligned} f:\mathbb{R}^n &\to \mathbb{R}_0^+ \\ x &\mapsto ||x||_2 := \sqrt{\sum_{i=1}^n x_i^2}, \end{aligned}$$

*welche die Euklidische Norm im $\mathbb{R}^n$ darstellt. Wir betrachten die* Niveaumengen $N_f \subset \mathbb{R}^n$ *von $f$ mit*

$$N_f(c) \;:=\; \{x\in\mathbb{R}^n : f(x)=c\} \;=\; \{f^{-1}(c)\}.$$

*In diesem Fall sind die Niveaumengen $N_f(c)$ Sphären mit Radius $c \in \mathbb{R}_0^+$. Weiterhin sieht man leicht, dass die Funktion $f$ auf der Menge $\mathbb{R}^n\backslash\{0\}$ partiell differenzierbar ist, da wir*

*für $x \neq 0$ mittels Kettenregel die folgende partielle Ableitung erhalten*

$$\begin{aligned}\partial_i ||x||_2 &= \partial_i \left( (\sum_{i=1}^{n} x_i)^{\frac{1}{2}} \right) = \partial_i \left( x_i^2 + \sum_{j \neq i} x_j^2 \right)^{1/2} \\ &= \frac{1}{2} \left( x_i^2 + \sum_{j \neq i} x_j^2 \right)^{-1/2} \cdot 2x_i = \frac{x_i}{||x||_2}.\end{aligned}$$

Der Begriff der partiellen Differenzierbarkeit in Definition 6.1 ist der schwächste Ableitungsbegriff für Funktionen in mehreren Veränderlichen. Anders als im eindimensionalen Fall gilt beispielsweise nicht, dass aus der partiellen Differenzierbarkeit einer Funktion schon folgt, dass diese stetig ist. Das soll uns das folgende Beispiel vor Augen führen.

**Beispiel 6.3**
*Wir betrachten im Folgenden eine Funktion auf $\mathbb{R}^n$ mit $n \geqslant 2$ die zeigt, dass aus partieller Differenzierbarkeit keine Stetigkeit folgt. Dazu definieren wir*

$$f(x) := \begin{cases} \frac{x_1 \cdot x_2 \cdot \ldots \cdot x_n}{||x||_2^n}, & \text{falls } x \neq 0, \\ 0, & \text{falls } x = 0. \end{cases}$$

*Man sieht sofort, dass die Funktion $f$ auf $\mathbb{R}^n \backslash \{0\}$ stetig und auch partiell differenzierbar ist.*

*Wir überprüfen nun die partielle Differenzierbarkeit im Punkt $x = 0$. Hierzu betrachten wir für $h \neq 0$*

$$f(0 + h \cdot e_i) - f(0) = \frac{0 \cdot \ldots \cdot 0 \cdot h \cdot 0 \cdot \ldots \cdot 0}{||h \cdot e_i||_2^n} - 0 = 0$$

*und somit folgt direkt*

$$\lim_{h \to 0} \frac{f(0 + h \cdot e_i) - f(0)}{h} = 0$$

*für alle Richtungen $i \in \{1, \ldots, n\}$. Daher wissen wir also, dass die Funktion $f$ auch im Punkt $x = 0$ partiell differenzierbar ist.*

*Tatsächlich ist die Funktion aber nicht stetig in der Null. Dafür betrachten wir eine Folge $(a_k)_{k \in \mathbb{N}} \subset \mathbb{R}^n \backslash \{0\}$, die diagonal in den Nullpunkt hineinläuft, d.h., konkret betrachten wir*

$$a_k := (\frac{1}{k}, \ldots, \frac{1}{k})^T, \quad k \in \mathbb{N}.$$

*Für jedes Folgeglied $a_k, k \in \mathbb{N}$, der Folge gilt dann $||a_k||_2 = \frac{\sqrt{n}}{k}$ und somit ist der Funktionswert von $f$ für diese Folgeglieder gegeben durch*

$$f(a_k) = \frac{(1/k)^n}{(\sqrt{n}/k)^n} \equiv \left( \frac{1}{\sqrt{n}} \right)^n = n^{-\frac{n}{2}}, \quad k \in \mathbb{N}.$$

*Daraus folgt aber schon, dass*

$$\lim_{k\to\infty} f(a_k) \;\equiv\; n^{-\frac{n}{2}} \;\neq\; 0 \;=\; f(0) \;=\; f(\lim_{k\to\infty} a_k).$$

*Das bedeutet also, dass die Funktion $f$ nicht stetig im Punkt $x = 0$ ist.*

*Man bemerke, dass diese Konstruktion nur für höherdimensionale Räume $\mathbb{R}^n$ mit $n \geqslant 2$ möglich ist, da die analog definierte Funktion im Eindimensionalen nicht differenzierbar ist.*

### 6.1.1 Differentialoperatoren erster Ordnung

In diesem Abschnitt wollen wir einige wichtige Operatoren einführen, welche direkt auf dem Konzept der partiellen Differenzierbarkeit von Funktionen basieren. Insbesondere werden wir den Gradienten, die Divergenz und die Rotation als Differentialoperatoren erster Ordnung kennenlernen, für deren Definition man nur erste partielle Ableitungen verwendet, und die eine zentrale Rolle in der Mathematik und Physik spielen.

Der erste Operator, den wir diskutieren wollen, ist der sogenannte *Gradient*, der alle partiellen Ableitungen einer Funktionen in einem Vektor sammelt.

**Definition 6.4** (Gradient)
*Sei $U \subset \mathbb{R}^n$ eine Teilmenge. Für eine partiell differenzierbare Funktion $f : U \to \mathbb{R}$ heißt der Spaltenvektor*

$$\nabla f(x) \;:=\; \left(\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_n}(x)\right)^T \;=\; (\partial_1 f(x), \ldots, \partial_n f(x))^T$$

*der* Gradient *von $f$ an der Stelle $x \in U$.*

*Das Symbol $\nabla$ bezeichnet man hierbei häufig als den* Nabla-Operator.

Wir wollen den Gradienten einer Funktion für ein bereits bekanntes Beispiel im Folgenden berechnen.

**Beispiel 6.5** (Gradient der Euklidischen Norm)
*Für die Euklidische Norm aus Beispiel 6.2 hat der Gradient für alle $x \neq 0$ die Form*

$$\nabla\, ||x||_2 \;=\; (\partial_1 ||x||_2, \ldots, \partial_n ||x||_2)^T \;=\; \left(\frac{x_1}{||x||_2}, \ldots, \frac{x_n}{||x||_2}\right)^T \;=\; \frac{x}{||x||_2}.$$

*Offensichtlich ist der Gradient der Norm selbst normiert, d.h., es gilt $\left\|\nabla ||x||_2\right\|_2 = 1$.*

Viele Rechenregeln der Differentiation für Funktionen mit nur einer Veränderlichen lassen sich problemlos auf Funktionen mit mehreren Veränderlichen generalisieren. Mit den bisherigen Definitionen können wir bereits die Produktregel für die Ableitung auf den Gradienten einer Funktionen verallgemeinern.

**Lemma 6.6** (Produktregel für den Gradienten)
*Sei $U \subset \mathbb{R}^n$ eine offene Teilmenge und seien $f, g : U \to \mathbb{R}$ zwei partiell differenzierbare Funktion. Dann gilt die folgende Produktregel für den Gradienten des Produkts der Funktionen*

$$\nabla(f \cdot g)(x) \;=\; [f \cdot (\nabla g) + (\nabla f) \cdot g](x), \qquad \textit{für alle } x \in U.$$

*Beweis.* Aus Gründen der Übersichtlichkeit verzichten wir bei diesem Beweis auf die Angabe des Funktionsarguments $x \in U$, wodurch die Notation zwar unpräzise, jedoch deutlich übersichtlicher wird. Wir schreiben den Gradienten des Produkts der beiden Funktionen zunächst mittels partiellen Ableitungen, so dass wir komponentenweise die Produktregel für eindimensionale Funktionen anwenden können, d.h.,

$$\begin{aligned}\nabla(f \cdot g) &= (\partial_1(f \cdot g), \ldots, \partial_n(f \cdot g))^T \\ &= (\partial_1 f \cdot g + f \cdot \partial_1 g, \ldots, \partial_n f \cdot g + f \cdot \partial_n g)^T\end{aligned}$$

Wir können den entstehenden Vektor nun linear auseinander ziehen und erhalten somit

$$\begin{aligned}\nabla(f \cdot g) &= (\partial_1 f \cdot g + f \cdot \partial_1 g, \ldots, \partial_n f \cdot g + f \cdot \partial_n g)^T \\ &= (\partial_1 f, \ldots, \partial_n f)^T \cdot g \;+\; f \cdot (\partial_1 g, \ldots, \partial_n g)^T \\ &= \nabla f \cdot g + f \cdot \nabla g.\end{aligned}$$

□

**Bemerkung 6.7**
*Sei $U \subset \mathbb{R}^n$ eine offene Teilmenge. Folgende Interpretationen des Gradienten-Operators sind ebenfalls geläufig.*

1. *Mit etwas missbräuchlicher Notation kann man den Gradienten auch als mehrdimensionale Abbildung interpretieren mit*

$$\begin{aligned}\nabla\colon U &\to \mathbb{R}^n \\ x &\mapsto \nabla(x).\end{aligned}$$

2. *Funktionen $v : U \to \mathbb{R}^n$ werden auch* Vektorfelder *genannt. In diesem Sinne ist kann der Gradient als Vektorfeld auf $U$ interpretiert werden..*

Für partiell differenzierbare Vektorfelder $v\colon U \to \mathbb{R}^n$ (d.h. jede Komponente $v_i$ von $v$ ist partiell differenzierbar) gibt es weitere Differentialoperatoren erster Ordnung. Im Folgenden führen wir den Begriff der Divergenz eines Vektorfelds $v$ ein.

**Definition 6.8** (Divergenz eines Vektorfelds)
*Sei $U \subset \mathbb{R}^n$ eine offene Teilmenge. Für ein partiell differenzierbares Vektorfeld $v : U \to \mathbb{R}^n$ nennt man*

$$\operatorname{div} v(x) \;:=\; \sum_{i=1}^{n} \frac{\partial v_i}{\partial x_i}(x) \;=\; \sum_{i=1}^{n} \partial_i v_i(x)$$

*die* Divergenz *von $v$ an der Stelle $x \in U$.*

*Oft wird die Divergenz auch als Skalarprodukt des Nabla-Operators $\nabla$ mit einen Vektorfeld $v$ notiert, d.h.,*

$$\operatorname{div} v \;=:\; \langle \nabla, v \rangle.$$

**Bemerkung 6.9**
*Die Divergenz eines Vektorfeldes $v\colon U \to \mathbb{R}^n$ besitzt eine einfache physikalische Interpretation. Es misst anschaulich den Fluss des Vektorfeldes im Punkt $x$. Man kann also an Hand der Divergenz entscheiden ob mehr Fluss in den Punkt $x \in U$ hinein- als hinausfließt oder anders herum.*

*i) Ist die Divergenz in einem Punkt $x \in U$* positiv *d.h., gilt* $\operatorname{div} v(x) > 0$ *so spricht man davon, dass das der Punkt eine* ***Quelle*** *des Vektorfelds darstellt. Anschaulich könnte man den Punkt $x$ mit einem $n$-dimensionalen Würfel umschließen und die Menge an Fluss, die aus dem Würfel austritt wäre größer als die Menge an Fluss die hineinfließt.*

*ii) Ist die Divergenz in einem Punkt $x \in U$* negativ *d.h., gilt* $\operatorname{div} v(x) < 0$ *so spricht man davon, dass das der Punkt eine* ***Senke*** *des Vektorfelds darstellt. Die Menge an Fluss, die in den umschließenden $n$-dimensionalen Würfel hineinfließt wäre größer als die Menge an Fluss die austritt.*

*iii) Ist die Divergenz in einem Punkt $x \in U$* Null*, d.h., gilt* $\operatorname{div} v(x) = 0$*, so nennen wir das Vektorfeld* divergenzfrei *im Punkt $x$. Die Menge an Fluss, die in den umschließenden $n$-dimensionalen Würfel hineinfließt entspricht in diesem Fall der Menge an Fluss die austritt.*

Im Folgenden wollen wir die Divergenz für ein dreidimensionales Vektorfeld konkret ausrechnen und bezüglich ihres Vorzeichens das Vektorfeld charakterisieren.

**Beispiel 6.10**
*Sei $v\colon \mathbb{R}^3 \to \mathbb{R}^3$ ein Vektorfeld mit*

$$v(x,y,z) \;:=\; \frac{1}{2}\begin{pmatrix} x^2 \\ -4xz \\ 6z \end{pmatrix}, \qquad \text{für alle } (x,y,z)^T \in \mathbb{R}^3.$$

*Wir berechnen die Divergenz des Vektorfelds für allgemeine Punkte $(x,y,z)^T \in \mathbb{R}^3$ als*

$$\begin{aligned} \operatorname{div} v(x,y,z) \;&=\; \partial_x v_1(x,y,z) + \partial_y v_2(x,y,z) + \partial_z v_3(x,y,z) \;=\; \frac{1}{2}\cdot(\partial_x x^2 - \partial_y 4xz + \partial_z 6z) \\ &=\; x - 0 + 3 \;=\; x + 3. \end{aligned}$$

*Wir erkennen also, dass die Divergenz des Vektorfelds $v$ maßgeblich von der $x$-Koordinate abhängt. Für $x > -3$ ist jeder Punkt $(x,y,z) \in \mathbb{R}^3$* Quelle *des Vektorfelds, für $x < -3$ ist jeder Punkt $(x,y,z) \in \mathbb{R}^3$* Senke *des Vektorfelds und nur bei $x = -3$ ist jeder Punkt $(x,y,z) \in \mathbb{R}^3$ des Vektorfelds* divergenzfrei.

Auch für die Divergenz lässt sich eine Produktregel für das Produkt einer reellwertigen Funktion und eines Vektorfeldes herleiten.

**Lemma 6.11** (Produktregel für die Divergenz)
*Sei $U \subset \mathbb{R}^n$ eine offene Teilmenge und es sei $v : U \to \mathbb{R}^n$ ein partiell differenzierbares Vektorfeld (d.h. jede Komponente $v_i$ von $v$ ist partiell differenzierbar) und $f : U \to \mathbb{R}$ eine partiell differenzierbare Funktion. Dann gilt die folgende Produktregel für die Divergenz:*

$$\operatorname{div}(f \cdot v)(x) \;=\; \langle \nabla f(x), v(x) \rangle + f(x) \cdot \operatorname{div} v(x), \qquad \text{für alle } x \in U.$$

*Beweis.* In der Hausaufgabe zu zeigen. □

Der letzte Differentialoperator erster Ordnung, den wir einführen wollen, ist a priori nur im Fall $U \subset \mathbb{R}^3$ definiert ist und spielt insbesondere in physikalischen Anwendungen eine wichtige Rolle. Der sogenannte Rotations-Operator hat seinen Namen dadurch erhalten, dass er bei Anwendung auf ein Strömungsfeld für jeden Ort das Doppelte der Winkelgeschwindigkeit angibt, mit der sich ein mitschwimmender Körper dreht.

**Definition 6.12** (Rotation)
*Sei $U \subset \mathbb{R}^3$ eine offene Teilmenge und $v : U \to \mathbb{R}^3$ ein partiell differenzierbares Vektorfeld (d.h. jede Komponente $v_i$ von $v$ ist partiell differenzierbar). Dann heißt der folgende Operator*

$$\operatorname{rot} v := \; (\partial_2 v_3 - \partial_3 v_2, \partial_3 v_1 - \partial_1 v_3, \partial_1 v_2 - \partial_2 v_1)$$

Rotation *von $v$. Er bildet also ein Vektorfeld wieder auf ein Vektorfeld ab.*

*Oft wird die Rotation als auch Vektorprodukt (vgl. Definition 3.15) des Nabla-Operators $\nabla$ mit einen Vektorfeld $v$ notiert, d.h.,*

$$\operatorname{rot} v \;=\; \nabla \times v.$$

Wir wollen die Rotation für das Vektorfeld aus Beispiel 6.10 nachrechnen.

**Beispiel 6.13**
*Sei $v\colon \mathbb{R}^3 \to \mathbb{R}^3$ ein Vektorfeld mit*

$$v(x,y,z) \;:=\; \frac{1}{2}\begin{pmatrix} x^2 \\ -4xz \\ 6z \end{pmatrix}, \qquad \text{für alle } (x,y,z)^T \in \mathbb{R}^3.$$

*Wir berechnen die Rotation des Vektorfelds für allgemeine Punkte $(x,y,z)^T \in \mathbb{R}^3$ als*

$$\begin{aligned} \operatorname{rot} v(x,y,z) \;&=\; (\partial_y v_3 - \partial_z v_2, \partial_z v_1 - \partial_x v_3, \partial_x v_2 - \partial_y v_1)\,(x,y,z) \\ &=\; \left(\partial_y 6z + \partial_z 4xz, \partial_z x^2 - \partial_x 6z, -\partial_x 4xz - \partial_y x^2\right) \\ &=\; (4x, 0, -4z). \end{aligned}$$

*Wir erkennen also, dass die Rotation des Vektorfelds nicht von der y-Koordinate abhängt.*

### 6.1.2 Differentialoperatoren höherer Ordnung

Analog zum eindimensionalen Fall kann man höhere Ableitungen definieren, d.h., wir wollen die Funktion $f : U \to \mathbb{R}$ mehrmals ableiten. Induktiv bezeichnet man $f$ als $(k+1)$-mal differenzierbar, falls $f$ $k$-mal differenzierbar ist und für eine beliebige Kombination von Richtungen $(i_1, \ldots, i_k)$ die Funktion $\partial_{i_1} \ldots \partial_{i_k} f$ wieder partiell differenzierbar ist. Hierbei ist a priori nicht klar, dass die partiellen Ableitungen vertauschbar sind. Wir stellen uns also die Frage ob für eine zweimal partiell differenzierbare Funktion $f$ die Gleichheit

$$\partial_i \partial_j f(x) \;=\; \partial_j \partial_i f(x) \tag{6.1}$$

für alle $i, j \in \{1, \ldots, n\}$ gilt.

In der Tat reicht die Eigenschaft zweimal partiell differenzierbar zu sein **nicht** für die Gleichheit in (6.1) aus, wie das folgende Beispiel zeigt.

**Beispiel 6.14** (Vertauschbarkeit partieller Ableitungen)
*Wir betrachten die Funktion*

$$f(x,y) \;:=\; xy \cdot \frac{x^2 - y^2}{x^2 + y^2}.$$

*Wie man nachrechnen kann ist die Funktion $f$ in allen Punkten $(x,y) \in \mathbb{R}^2$ zweimal partiell differenzierbar. Wir berechnen die ersten partiellen Ableitungen*

$$\begin{aligned}
\partial_x f(x,y) &= \frac{(3x^2y - y^3)(x^2+y^2) - (x^3y - xy^3)\cdot 2x}{(x^2+y^2)^2} = y\,\frac{x^4 + 4x^2y^2 - y^4}{(x^2+y^2)^2}, \\
\partial_y f(x,y) &= \frac{(x^3 - 3xy^2)(x^2+y^2) - (x^3y - xy^3)\cdot 2y}{(x^2+y^2)^2} = x\,\frac{x^4 - 4x^2y^2 - y^4}{(x^2+y^2)^2}.
\end{aligned}$$

*Wir betrachten nun die **zweiten** partiellen Ableitungen von $f$ im Punkt $(x,y) = (0,0)$ mit*

$$\begin{aligned}
\partial_y \partial_x f(0,0) &= \lim_{h\to 0} \frac{\partial_x f(0,h) - \partial_x f(0,0)}{h} = \lim_{h\to 0} \frac{h}{h}\,\frac{0 + 0 - h^4}{(0+h^2)^2} - 0 = -1, \\
\partial_x \partial_y f(0,0) &= \lim_{h\to 0} \frac{\partial_y f(h,0) - \partial_y f(0,0)}{h} = \lim_{h\to 0} \frac{h}{h}\,\frac{h^4 - 0 - 0}{(h^2+0)^2} - 0 = +1.
\end{aligned} \tag{6.2}$$

*Wir erkennen, dass die Funktion $f$ zwar zweimal partiell differenzierbar ist im Nullpunkt, jedoch ist die Reihenfolge der partiellen Ableitungen nicht vertauschbar.*

Um eine Bedingung zu erarbeiten, die die Vertauschbarkeit von höheren partiellen Ableitungen versichert, fragt man sich zunächst weshalb die Funktion aus Beispiel 6.14 diese Eigenschaft verletzt. Betrachtet man die zweiten partiellen Ableitungen $\partial_i \partial_j f$ in (6.2) so fällt auf, dass diese nicht stetig im Punkt $(0,0)$ sind. Die Vermutung liegt also nahe, dass gerade die fehlende Stetigkeit die Vertauschbarkeit verhindert. Diese Vermutung lässt sich tatsächlich auch beweisen, was gemeinhin als der folgende *Satz von Schwarz* bekannt ist.

**Satz 6.15** (Satz von Schwarz)
*Sei $U \subset \mathbb{R}^n$ eine offene Teilmenge und $f : U \to \mathbb{R}$ eine zweimal stetig partiell differenzierbare Funktion. Dann gilt*

$$\partial_i \partial_j f \;=\; \partial_j \partial_i f$$

*für alle Ableitungsrichtungen $i, j \in \{1, \ldots, n\}$.*

*Beweis.* Wir können o.B.d.A. annehmen, dass $U \subset \mathbb{R}^2$ eine offene Teilmenge ist. Wir betrachten einen beliebigen Punkt $(\bar{x}, \bar{y}) \in U$. Da $U$ offen ist existiert ein $\delta > 0$, so dass ein Quadrat $B_\delta^\infty$ mit Seitenlänge $\delta$ noch in $U$ liegt, d.h.,

$$B_\delta^\infty(\bar{x}, \bar{y}) \;:=\; \{(x, y) \in \mathbb{R}^2 : |x - \bar{x}| < \delta \wedge |y - \bar{y}| < \delta\} \;\subset\; U.$$

Für ein festes $y \in (\bar{y} - \delta, \bar{y} + \delta)$ betrachten wir nun die eindimensionale Funktion

$$\begin{aligned} R_y : [-\delta, \delta] \;&\to\; \mathbb{R} \\ x \;&\mapsto\; R_y(x) \;:=\; f(x + \bar{x}, y + \bar{y}) - f(x + \bar{x}, \bar{y}). \end{aligned}$$

Wir bemerken zunächst, dass für $y = 0$ gilt

$$R_0(x) \;=\; f(x + \bar{x}, \bar{y}) - f(x + \bar{x}, \bar{y}) \;=\; 0, \qquad \text{für alle } x \in [-\delta, \delta].$$

Da $R_y$ als Einschränkung von $f$ in eine Koordinatenrichtung stetig ist können wir für ein beliebiges $x \in [-\delta, \delta]$ mit Hilfe des Mittelwertsatzes [Burger2020, Kapitel 6.2] einen Punkt $\xi_x \in [-\delta, \delta]$ finden mit $|\xi_x| \leqslant |x|$, so dass

$$R_y'(\xi_x) \cdot x \;=\; R_y(x) - R_y(0).$$

Hierbei ist die Ableitung der Funktion $R_y$ gegeben durch

$$R_y'(\xi_x) \;=\; \partial_1 R_y(\xi_x) \;=\; \partial_1 f(\xi_x + \bar{x}, y + \bar{y}) \;-\; \partial_1 f(\xi_x + \bar{x}, \bar{y}).$$

Wir können nun basierend auf dem obigen Ausdruck erneut eine stetige eindimensionale Funktion $r : [-\delta, \delta] \to \mathbb{R}$ definieren mit $r(y) := R_y'(\xi_x)$. Auf diese Funktion wenden wir erneut den Mittelwertsatz an und finden analog für jedes beliebige $y \in [-\delta, \delta]$ einen Punkt $\eta_y \in [-\delta, \delta]$ mit $|\eta_y| \leqslant |y|$, so dass

$$r'(\eta_y) \cdot y \;=\; r(y) - r(0).$$

Hierbei ist die Ableitung der Funktion $r$ gegeben durch

$$r'(\eta_y) \;=\; \partial_2 \partial_1 f(\xi_x + \bar{x}, \eta_y + \bar{y}).$$

Insgesamt erhalten wir für alle $(x,y) \in B_\delta^\infty(0,0)$ die folgende Identität

$$\begin{aligned}\partial_2\partial_1 f(\xi_x+\bar{x},\eta_y+\bar{y}) &= r'(\eta_y) = \frac{r(y)-r(0)}{y} = \frac{R_y'(\xi_x)-R_0'(\xi_x)}{y}\\ &= \frac{\frac{R_y(x)-R_y(0)}{x}-\frac{R_0(x)-R_0(0)}{x}}{y} = \frac{R_y(x)-R_y(0)}{xy}\\ &= \frac{f(x+\bar{x},y+\bar{y})-f(x+\bar{x},\bar{y})-f(\bar{x},y+\bar{y})-f(\bar{x},\bar{y})}{xy}.\end{aligned}$$

Insgesamt gilt also für alle $(x,y) \in B_\delta^\infty(0,0)$ die folgende Gleichung

$$\partial_2\partial_1 f(\xi_x+\bar{x},\eta_y+\bar{y})\cdot xy = f(x+\bar{x},y+\bar{y})-f(x+\bar{x},\bar{y})-f(\bar{x},y+\bar{y})-f(\bar{x},\bar{y}). \quad (6.3)$$

Mit vertauschten Rollen wenden wir die Argumente für ein festes $x \in (\bar{x}-\delta,\bar{x}+\delta)$ analog auf die eindimensionale Funktion

$$\begin{aligned}T_x\colon [-\delta,\delta] &\to \mathbb{R}\\ y &\mapsto T_x(y) := f(x+\bar{x},y+\bar{y})-f(\bar{x},y+\bar{y})\end{aligned}$$

an. Für $(x,y) \in B_\delta^\infty(0,0)$ liefert das jeweils Skalare $\hat{\xi}_x, \hat{\eta}_y \in [-\delta,\delta]$ mit $|\hat{\xi}_x| \leqslant |x|, |\hat{\eta}_y| \leqslant |y|$, so dass

$$\begin{aligned}\partial_1\partial_2 f(\hat{\xi}_x+\bar{x},\hat{\eta}_y+\bar{y})\cdot yx &= T_x(y)-T_x(0)\\ &= f(x+\bar{x},y+\bar{y})-f(x+\bar{x},\bar{y})-f(\bar{x},y+\bar{y})-f(\bar{x},\bar{y}).\end{aligned}$$

Zusammen mit (6.3) können wir für $xy \neq 0$ schlussfolgern, dass

$$(\partial_2\partial_1 f)(\xi_x+\bar{x},\eta_y+\bar{y}) = (\partial_1\partial_2 f)(\hat{\xi}_x+\bar{x},\hat{\eta}_y+\bar{y}).$$

Es folgt nun der entscheidende Schritt. Wir betrachten im Folgenden eine Nullfolge $(x_n,y_n)_{n\in\mathbb{N}} \subset B_\delta^\infty(0,0)$ mit $(x_n,y_n)_{n\in\mathbb{N}} \to (0,0)$ und $x_n y_n \neq 0$ für alle $n \in \mathbb{N}$. Diese Folge induziert mit den obigen Argumenten zwei weitere Nullfolgen $(\xi_n,\eta_n)_{n\in\mathbb{N}}$ und $(\hat{\xi}_n,\hat{\eta}_n)_{n\in\mathbb{N}}$, so dass

$$(\partial_2\partial_1 f)(\xi_n+\bar{x},\eta_n+\bar{n}) = (\partial_1\partial_2 f)(\hat{\xi}_n+\bar{x},\hat{\eta}_n+\bar{y}).$$

Um zu sehen, dass diese beiden Folgen wieder Nullfolgen sind nutzen wir die Beschränktheit der Folgeglieder aus, d.h., dass $|\xi_x|, |\hat{\xi}_x| \leqslant |x|$ und $|\eta_y|, |\hat{\eta}_y| \leqslant |y|$ gilt.

Schließlich benutzen wir die Stetigkeit der zweiten partiellen Ableitungen und sehen damit ein, dass

$$\begin{aligned}\partial_2\partial_1 f(\bar{x},\bar{y}) &= \lim_{n\to\infty} \partial_2\partial_1 f(\xi_n+\bar{x},\eta_n+\bar{y})\\ &= \lim_{n\to\infty} \partial_1\partial_2 f(\hat{\xi}_n+\bar{x},\hat{\eta}_n+\bar{y}) = \partial_1\partial_2 f(\bar{x},\bar{y}).\end{aligned}$$

□

Das folgende Korollar sagt aus, dass die Vertauschbarkeit nicht nur für zwei Ableitungsrichtungen gilt, sondern für eine beliebige Permutation von partiellen Ableitungen gegeben ist.

**Korollar 6.16**
*Sei $U \subset \mathbb{R}^n$ eine offene Teilmenge und $f : U \to \mathbb{R}$ eine $k$-mal stetig partiell differenzierbare Funktion. Dann gilt für jede bijektive Abbildung $\pi : \{1, \dots, k\} \to \{1, \dots, k\}$ (auch* Permutation *genannt)*

$$\partial_{i_1} \dots \partial_{i_k} f \;=\; \partial_{i_{\pi(1)}} \dots \partial_{i_{\pi(k)}} f$$

*für beliebige Indizes $i_1, \dots, i_k \in \{1, \dots, n\}$.*

*Beweis.* In der Hausaufgabe zu zeigen. □

**Bemerkung 6.17** *1. Für mehrfache partielle Ableitung in die gleiche Koordinatenrichtung schreibt man häufig kurz*

$$\partial_i \partial_i f \;=:\; \partial_i^2 f.$$

*Dies lässt sich analog für höhere Ableitungen durch höhere Potenzen ausdrücken.*

*2. Insbesondere folgt aus dem Satz 6.15 von Schwarz, dass für eine zweimal stetig partiell differenzierbare Funktion $f$ die folgende Beobachtung für die Rotation gilt*

$$\operatorname{rot} \nabla f \;=\; 0.$$

*Dies folgt aus der Beobachtung, dass $x \times x = 0$ gilt für alle $x \in \mathbb{R}^3$ (vgl. Lemma 3.16) und der Notation*

$$\operatorname{rot} \nabla f \;=\; \nabla \times \nabla f \;=\; 0.$$

*Diese Erkenntnis wird im allgemeinen Sprachgebrauch häufig mit dem Ausspruch* „Gradientenfelder sind Rotationsfrei“ *gemeint.*

Analog zur zweiten Ableitung von Funktionen in einer Veränderlichen kann man für mehrdimensionale Funktionen einen Operator definieren, der alle zweiten Ableitungen sammelt - die sogenannte *Hesse-Matrix*.

**Definition 6.18** (Hesse-Matrix)
*Sei $U \subset \mathbb{R}^n$ eine offene Teilmenge und $f : U \to R$ eine zweimal stetig partiell differenzierbare Funktion. Dann ist die* Hesse-Matrix *von $f$ im Punkt $x \in U$ definiert als*

$$H_f(x) \;:=\; \left( \frac{\partial^2 f}{\partial x_i \partial x_j}(x) \right)_{1 \leqslant i,j \leqslant n} \;=\; (\partial_i \partial_j f(x))_{1 \leqslant i,j \leqslant n} \;=\; \begin{pmatrix} \partial_1 \partial_1 f(x) & \cdots & \partial_1 \partial_n f(x) \\ \vdots & \ddots & \vdots \\ \partial_n \partial_1 f(x) & \cdots & \partial_n \partial_n f(x) \end{pmatrix}.$$

**Bemerkung 6.19**
*Die Hesse-Matrix $H_f$ einer zweimal stetig partiell differenzierbaren Funktion $f$ ist symmetrisch, da die Reihenfolge der partiellen Ableitungen vertauscht werden kann und somit gilt $H_f = H_f^T$.*

Wir wollen im folgenden Beispiel die Hesse-Matrix einer zweidimensionalen Funktion berechnen.

**Beispiel 6.20**
*Sei $f\colon \mathbb{R}^2 \to \mathbb{R}$ eine zweimal stetig partiell differenzierbare Funktion mit*

$$f(x,y) \;:=\; x^2y - \sin(x).$$

*Um die Hesse-Matrix $H_f$ von $f$ zu berechnen geben wir zuerst die ersten partiellen Ableitungen von $f$ an mit*

$$\begin{aligned}
\partial_x f(x,y) &= \partial_x(x^2y - \sin(x)) = 2xy - \cos(x),\\
\partial_y f(x,y) &= \partial_y(x^2y - \sin(x)) = x^2.
\end{aligned}$$

*Nun betrachten wir die zweiten partiellen Ableitungen von $f$ mit*

$$\begin{aligned}
\partial_x\partial_x f(x,y) &= \partial_x(2xy - \cos(x)) = 2y + \sin(x),\\
\partial_y\partial_x f(x,y) &= \partial_y(2xy - \cos(x)) = 2x,\\
\partial_x\partial_y f(x,y) &= \partial_x(x^2) = 2x,\\
\partial_y\partial_y f(x,y) &= \partial_y(x^2) = 0,
\end{aligned}$$

*Daraus ergibt sich für die Hesse-Matrix von $f$ die folgende Gestalt:*

$$H_f(x) \;=\; \begin{pmatrix} \partial_x\partial_x f(x,y) & \partial_x\partial_y f(x,y) \\ \partial_y\partial_x f(x,y) & \partial_y\partial_y f(x,y) \end{pmatrix} \;=\; \begin{pmatrix} 2x + \sin(x) & 2x \\ 2x & 0 \end{pmatrix}.$$

Den letzten Operator den wir in diesem Abschnitt kennenlernen wollen ist der sogenannte Laplace-Operator.

**Definition 6.21** (Laplace-Operator)
*Sei $U \subset \mathbb{R}^n$ offen und $f : U \to \mathbb{R}$ eine zweimal stetig differenzierbare Funktion, dann heißt*

$$\Delta f(x) \;:=\; \operatorname{div} \nabla f(x) \;=\; \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2}(x) \;=\; \sum_{i=1}^{n} \partial_i^2 f(x).$$

*Laplace-Operator. Analog zu den vorherigen Überlegungen kann auch der Laplace-Operator als Differtialoperator geschrieben werden,*

$$\Delta f \;=:\; \langle \nabla, \nabla f \rangle \;=:\; \nabla^2 f.$$

Wir wollen den Laplace Operator für die Funktion aus Beispiel 6.20 im Folgenden berechnen.

**Beispiel 6.22**
*Sei $f\colon \mathbb{R}^2 \to \mathbb{R}$ eine zweimal stetig partiell differenzierbare Funktion mit*

$$f(x,y) \;:=\; x^2y - \sin(x).$$

*Um den Laplace-Operator $\Delta f$ von $f$ zu berechnen geben wir zunächst die relevanten zweiten partiellen Ableitungen von $f$ an mit*

$$\begin{aligned} \partial_x^2 f(x,y) &= \partial_x^2(x^2y - \sin(x)) = 2x + \sin(x), \\ \partial_y^2 f(x,y) &= \partial_y^2(x^2y - \sin(x)) = 0. \end{aligned}$$

*Nun können wir den Laplace Operator $\Delta f$ von $f$ angeben als*

$$\Delta f(x,y) \;:=\; \partial_x^2 f(x,y) + \partial_y^2 f(x,y) \;=\; 2x + \sin(x) + 0 \;=\; 2x + \sin(x).$$

**Bemerkung 6.23**
*Der Laplace Operator $\Delta f$ einer zweimal stetig partiell differenzierbaren Funktion $f$ lässt sich ebenfalls als Spur der Hesse-Matrix $H_f$ von $f$ darstellen durch:*

$$\Delta f(x) \;=\; \mathrm{Spur}(H_f(x)).$$

## 6.2 Totale Differenzierbarkeit

Im letzten Abschnitt haben wir gesehen, dass der Begriff der partiellen Ableitung die nächstliegendste Strategie ist um Ableitungen für Funktionen mehrerer Veränderlicher zu definieren. Allerdings wurde aus den obigen Beispielen auch klar, dass Definition über Einschränkung auf einzelne Koordinatenachsen, einerseits willkürlich ist, aber insbesondere auch keine befriedigende Verallgemeinerung des Ableitungsbegriffs darstellt. So gilt z.B. die aus dem Eindimensionalen bekannte Implikation für Funktionen $f\colon \mathbb{R} \to \mathbb{R}$

$$f \text{ ist differenzierbar } \;\Rightarrow\; f \text{ ist stetig}$$

nicht für den Begriff der partiellen Differenzierbarkeit.

Aus diesem Grund, wollen wir nun einen weiteren Ableitungsbegriff kennenlernen, welcher eine tatsächliche Verallgemeinerung dieser Beobachtung darstellt. Insbesondere erlaubt es uns dieser neue Begriff auch Ableitung von vektorwertigen Funktionen $f : U \to \mathbb{R}^m$ für eine offene Teilmenge $U \subset \mathbb{R}^n$ zu definieren.

**Definition 6.24** (Totale Differenzierbarkeit)
*Sei $U \subset \mathbb{R}^n$ eine offene Teilmenge. Dann heißt eine Funktion $f : U \to \mathbb{R}^m$* total differenzierbar *im Punkt $x \in U$, falls für einen beliebigen Vektor $\xi \in \mathbb{R}^n$ eine lineare Abbildung $L : \mathbb{R}^n \to \mathbb{R}^m$ existiert, so dass*

$$\lim_{\xi \to 0} \frac{||f(x+\xi) - f(x) - L\xi||}{||\xi||} \;=\; 0. \tag{6.4}$$

Die folgende Bemerkung beschreibt die Intuition hinter der Definition von totaler Differenzierbarkeit.

**Bemerkung 6.25**
*Zur totalen Differenzierbarkeit können wir folgende Beobachtungen festhalten.*

1. *In* (6.4) *betrachtet man das sogenannte* Fehlerfunktional

$$r(\xi) \;:=\; f(x+\xi) - f(x) - L\xi \tag{6.5}$$

*welches die Abweichung zwischen der Linearisierung und der eigentlichen Differenz misst. Bei der Definition von totaler Differenzierbarkeit fordern wir also, dass diese Diskrepanz schnell genug gegen Null konvergiert.*

*Zudem erkennen wir, dass Definition 6.24 konsistent mit dem herkömmlichen Begriff der Differenzierbarkeit einer Funktion $f$ im Eindimensionalen ($n = m = 1$) ist, da die Funktion $L$ in diesem Fall als*

$$L(h) \;:=\; f'(x) \cdot h$$

*gewählt werden kann.*

2. *Die lineare Abbildung $L$ wird typischerweise mit der darstellenden Matrix*

$$L \;=\; \begin{pmatrix} L_{11} & \dots & L_{1n} \\ \vdots & & \vdots \\ L_{m1} & \dots & L_{mn} \end{pmatrix} \in \mathbb{R}^{m,n}$$

*bezüglich der kanonischen Basen von $\mathbb{R}^n$ und $\mathbb{R}^m$ identifiziert. Das Fehlerfunktional, hat dann komponentenweise die Form*

$$r_i(\xi) \;=\; f_i(x+\xi) - f_i(x) - \sum_{j=1}^{n} L_{ij}\xi_j.$$

*Somit sehen wir, dass $f$ genau dann total differenzierbar ist, falls jede Komponente von $f$ im Bildraum total differenzierbar ist.*

**Beispiel 6.26**
*Sei $C \in \mathbb{R}^{n,n}$ eine symmetrische Funktion und die Funktion $f : \mathbb{R}^n \to \mathbb{R}$ als quadratische Form gegeben durch*

$$f(x) \;:=\; \langle x, Cx\rangle.$$

*Wir berechnen nun für einen beliebigen Punkt $x \in \mathbb{R}^n$ und einen Richtungsvektor $\xi \in \mathbb{R}^n$*

$$\begin{aligned} f(x+\xi) \;&=\; \langle x+\xi, C(x+\xi)\rangle \;=\; \langle x, Cx\rangle + \underbrace{\langle x, C\xi\rangle + \langle \xi, Cx\rangle}_{=2\langle Cx,\xi\rangle =: L\xi} + \langle \xi, C\xi\rangle \\ &= f(x) + L\xi + \langle \xi, C\xi\rangle. \end{aligned}$$

*Das Fehlerfunktional ist also gegeben durch*

$$r(\xi) \;:=\; f(x+\xi) - f(x) - L\xi \;=\; f(x) + L\xi + \langle \xi, C\xi\rangle - f(x) - L\xi \;=\; \langle \xi, C\xi\rangle.$$

*Mit Hilfe der Cauchy-Schwarz Ungleichung aus Satz 3.5 sehen wir, dass*

$$|r(\xi)| \;=\; \langle \xi, C\xi\rangle| \;\leqslant\; ||\xi|| \cdot ||C\xi||| \;\leqslant\; ||C|| \cdot ||\xi||^2.$$

*Damit können wir schließlich folgern*

$$\lim_{\xi\to 0} \frac{||r(\xi)||}{||\xi||} \;\leqslant\; \lim_{\xi\to 0} ||C|| \cdot ||\xi|| \;=\; 0.$$

*Wir sehen also, dass die Funktion $f(x) = \langle x, Cx\rangle$ total differenzierbar in allen Punkten $x \in \mathbb{R}^n$ ist.*

Der folgende Satz liefert uns nun die gewünschte Aussage, dass totale Differenzierbarkeit einer Funktion schon Stetigkeit impliziert. Zudem stellt er einen Bezug zum Begriff der partiellen Differenzierbarkeit her.

**Satz 6.27**
*Sei $U \subset \mathbb{R}^n$ eine offene Teilmenge und sei $f : U \to \mathbb{R}^m$ eine im Punkt $x \in U$ total differenzierbare Funktion, d.h., es existiert eine Matrix $L \in \mathbb{R}^{m\times n}$, so dass,*

$$r(\xi) \;=\; f(x+\xi) - f(x) - L\xi$$

*die Gleichung* (6.4) *erfüllt. Dann gilt*

1. *$f$ ist stetig im Punkt $x$,*
2. *jede Komponente von $f$ im Bildraum ist partiell differenzierbar in $x$ und die Einträge der Matrix $L$ sind gerade die partiellen Ableitungen von $f$, d.h.,*

$$\partial_j f_i \;=\; L_{ij}.$$

*Beweis.* Die Stetigkeit ist eine direkte Folgerung aus der Definition, denn mittels Dreiecksungleichung können wir zeigen, dass gilt

$$||f(x+\xi)-f(x)|| \;=\; ||f(x+\xi)-f(x)-L\xi+L\xi|| \;\leqslant\; ||f(x+\xi)-f(x)-L\xi||+||L\xi||.$$

Da auf Grund der Linearität von $L$ offensichtlich $\lim_{\xi\to 0}||L\xi||=0$ gilt und $f$ total differenzierbar ist, d.h.,

$$\lim_{\xi\to 0}||f(x+\xi)-f(x)-L\xi|| \;=\; 0$$

folgt somit schon

$$\lim_{\xi\to 0}||f(x+\xi)-f(x)|| \;\leqslant\; \lim_{\xi\to 0}||f(x+\xi)-f(x)-L\xi||+||L\xi|| \;=\; 0.$$

Da der obige Grenzwerte beliebige Nullfolgen betrachtet folgt die Stetigkeit von $f$.

Für den Zusammenhang mit der partiellen Differenzierbarkeit in der zweiten Aussage des Satzes betrachten wir eine Komponente $f_i$ von $f$ für $i\in\{1,\ldots,m\}$ und damit gilt nach Bemerkung 6.25

$$r_i(\xi) \;=\; f_i(x+\xi)-f_i(x)-\sum_{j=1}^{n}L_{ij}\xi_j.$$

Treffen wir nun die spezielle Wahl $\xi=h\cdot e_j$ für eine Koordinatenrichtung $e_j, 1\leqslant j\leqslant n$, so sehen wir

$$r_i(h\cdot e_j) \;=\; f_i(x+h\cdot e_j)-f_i(x)-L_{ij}\cdot h.$$

Setzen wir nun die Definition der totalen Differenzierbarkeit für die Komponente $r_i$ ein folgt

$$\begin{aligned}|\partial_j f_i(x)-L_{ij}| \;&=\; \lim_{h\to 0}\left|\frac{f_i(x+h\cdot e_j)-f_i(x)}{h}-L_{ij}\right| = \lim_{h\to 0}\frac{|f_i(x+h\cdot e_j)-f_i(x)-L_{ij}\cdot h|}{h}\\ &=\; \lim_{h\to 0}\frac{||r_i(h\cdot e_j)||}{||h\cdot e_j||} \;=\; \lim_{h\to 0}\frac{||r_i(\xi)||}{||\xi||} \;=\; 0.\end{aligned}$$

Damit haben wir gezeigt, dass die Einträge der Matrix $L$ mit den partiellen Ableitungen der Funktion $f$ übereinstimmen. □

Speziell die besondere Gestalt der Matrix $L$ in der zweiten Aussage des Satzes 6.27 motiviert die Definition der Jacobi-Matrix einer vektorwertigen, partiell differenzierbaren Funktion.

**Definition 6.28** (Jacobi-Matrix)
*Sei $U \subset \mathbb{R}^n$ eine offene Teilmenge und sei $f : U \to \mathbb{R}^m$ eine partiell differenzierbare Funktion (d.h. jede Komponente $f_i$ ist partiell differenzierbar), dann heißt die Matrix*

$$(Df)(x) \;:=\; J_f(x) \;:=\; \begin{pmatrix} \partial_1 f_1(x) & \dots & \partial_n f_1(x) \\ \vdots & & \vdots \\ \partial_1 f_m(x) & \dots & \partial_n f_m(x) \end{pmatrix}$$

Jacobi-Matrix *am Punkt $x \in U$.*

Im Folgenden wollen wir wichtige Beobachtungen zur Bedeutung der Jacobi-Matrix festhalten.

**Bemerkung 6.29**
*Sei $U \subset \mathbb{R}^n$ eine offene Teilmenge und sei $f \colon U \to \mathbb{R}^m$ eine partiell differenzierbare Funktion. Dann können wir folgendes feststellen.*

1. *Falls $f$ eine reellwertige Funktion ist, d.h., $m = 1$, so stimmt die Jacobi-Matrix von $f$ am Punkt $x \in U$ mit dem Gradienten von $f$ in $x$ überein, d.h.*

$$(Df)(x) \;=\; (\nabla f(x))^T.$$

2. *Aus Satz 6.27, folgt insbesondere, dass die lineare Abbildung $L$ in der Definition der totalen Differenzierbarkeit eindeutig bestimmt ist durch die Jacobi-Matrix. Das Fehlerfunktional* (6.5) *ist also gegeben durch*

$$r(\xi) = f(x+\xi) - f(x) - J_f(x)\xi.$$

Wir wissen nun, dass für Funktionen $f \colon U \to \mathbb{R}^m$ die Implikation

$$f \text{ ist total differenzierbar } \;\Rightarrow\; f \text{ ist partiell differenzierbar}$$

gilt. Die Umkehrung gilt offensichtlich nicht, wie wir in Beispiel 6.3 gesehen haben. Nehmen wir jedoch eine zusätzliche zusätzliche Stetigkeitsannahme hinzu, erhalten wir wieder totale Differenzierbarkeit, wie folgender Satz zeigt.

**Satz 6.30**
*Sei $U \subset \mathbb{R}^n$ eine offene Teilmenge und sei $f : U \to \mathbb{R}$ eine in $U$ partiell differenzierbare Funktion für die alle partiellen Ableitungen $\partial_i f$ stetig sind im Punkt $x \in U$.*
*Dann ist $f$ in $x \in U$ total differenzierbar.*

*Beweis.* Wir wählen $x \in U$ und ein $\delta > 0$, so dass $B_\delta(x) \subset U$. Wir betrachten nun einen beliebigen Vektor $\xi \in B_\delta(0)$ und definieren basierend auf $\xi$ einen Familie von Vektoren

$z_0, \dots, z_n \in B_\delta(x)$ mit

$$z_k \;:=\; x + \sum_{i=1}^{k} \xi_k \;=\; (x_1 + \xi_1, \dots, x_k + \xi_k, x_{k+1}, \dots, x_n)^T, \qquad \text{für} \quad k = 0, \dots, n.$$

Wir erkennen, dass $z_0 = x$ und $z_n = x + \xi$ gilt und die Vektoren $z_k$ dadurch entstehen, dass wir sukzessive weitere Komponenten von $\xi$ hinzunehmen. Das bedeutet, dass ich zwei aufeinanderfolgende Vektoren $z_k$ und $z_{k-1}$ nur in der $k$-ten Koordinatenrichtung unterscheiden.

Im Folgenden beschränken wir die Funktion $f$ auf die $k$-te Komponente, d.h., wir betrachten

$$\tilde{f}(\theta) \;:=\; f(z_{k-1} + \theta \cdot e_k) \;=\; f(x_1 + \xi_1, \dots, x_{k-1} + \xi_{k-1}, x_k + \theta, x_{k+1}, \dots, x_n)$$

Durch diese eindimensionale Beschränkung sehen wir, dass gilt

$$f(z_k) - f(z_{k-1}) \;=\; \tilde{f}(\xi_k) - \tilde{f}(0). \tag{6.6}$$

Wenden wir nun den Mittelwertsatz für Funktionen in einer Veränderlichen [Burger2020, Kapitel 6.2] an, so sehen wir, dass ein $\theta_k \in (0, \xi_k)$ existiert, so dass

$$\tilde{f}(\xi_k) - \tilde{f}(0) \;=\; \tilde{f}'(\theta_k) \cdot \xi_k.$$

Wegen der Identität (6.6) folgt damit schon, dass

$$f(z_k) - f(z_{k-1}) \;=\; \partial_k f(z_{k-1} + \theta_k e_k) \cdot \xi_k. \tag{6.7}$$

Da (6.7) für jedes $0 \leqslant k \leqslant n$ gilt, können wir über diese Indizes summieren und erhalten damit die folgende Teleskopsumme:

$$\begin{aligned} f(x + \xi) - f(x) \;&=\; f(z_n) - f(z_0) \;=\; \sum_{k=1}^{n} f(z_k) - f(z_{k-1}) \\ &=\; \sum_{k=1}^{n} \partial_k f(z_k + \theta_k e_k) \cdot \xi_k. \end{aligned}$$

Wir betrachten nun das folgende Fehlerfunktional für die spezielle Wahl der Linearform $L$ in Definition 6.24 als Jacobi-Matrix $J_f$. In diesem Fall entspricht die Jacobi-Matrix dem transponierten Gradienten von $f$ nach Bemerkung 6.29, d.h., $J_f(x) = (\nabla f(x))^T$. Man beachte, dass wir nur irgendeine Linearform finden müssen, die die Definition von totaler Differenzierbarkeit erfüllt.

$$\begin{aligned} r(\xi) = f(x + \xi) - f(x) - J_f(x)\xi \;&=\; \left( \sum_{k=1}^{n} \partial_k f(z_k + \theta_k e_k) \xi_k \right) - \langle \nabla F(x), \xi \rangle \\ &=\; \sum_{k=1}^{n} \left( \partial_k f(z_k + \theta_k e_k) - \partial_k f(x) \right) \xi_k. \end{aligned}$$

Mittels der Cauchy-Schwarz Ungleichung in Satz 3.5 können wir damit sofort folgern , dass

$$\begin{aligned}\frac{||r(\xi)||}{||\xi||} &\leqslant \frac{||\sum_{k=1}^{n} \partial_k f(z_k + \theta_k e_k) - \partial_k f(x)|| \cdot ||\xi||}{||\xi||} \\ &= \| \sum_{k=1}^{n} \partial_k f(z_k + \theta_k e_k) - \partial_k f(x) \|.\end{aligned}$$

Per Konstruktion gilt stets $\theta_k \leqslant \xi_k$ für $0 \leqslant k \leqslant n$ und somit

$$\lim_{\xi \to 0}(z_k + \theta_k e_k) \;=\; \lim_{\xi \to 0} (x_1 + \xi_1, \ldots, x_{k-1}\xi_{k-1}, x_k + \theta_k, x_{k+1}, \ldots, x_n) \;=\; x.$$

Benutzen wir nun die Stetigkeit der partiellen Ableitungen folgt damit schon die totale Differenzierbarkeit von $f$ in $x$ durch

$$\lim_{\xi \to 0} \frac{||r(\xi)||}{||\xi||} \;\leqslant\; \lim_{\xi \to 0} || \sum_{k=1}^{n} \partial_k f(z_k + \theta_k e_k) - \partial_k f(x)|| \;=\; 0.$$

□

Insgesamt haben wir nun eine Abstufung der Stärke der verschiedenen Begriffe von Differenzierbarkeit von Funktionen in mehreren Veränderlichen, wie folgende Bemerkung zusammenfasst.

**Bemerkung 6.31**
*Sei $U \subset \mathbb{R}^n$ eine offene Teilmenge und $f\colon U \to \mathbb{R}^m$ eine Funktion. Dann können wir folgende Beobachtungen festhalten:*

1. *Zusammengefasst habe wir folgende Implikationskette:*

$$\begin{array}{ccccc} & & f \text{ ist stetig partiell differenzierbar} & & \\ & & \Downarrow & & \\ f \text{ ist total differenzierbar} & \Rightarrow & & & f \text{ ist stetig} \\ & & \Downarrow & & \\ & & f \text{ ist partiell differenzierbar.} & & \end{array}$$

2. *Die Umkehrungen obiger Implikationen gelten im Allgemeinen nicht, wie wir in verschiedenen Beispielen zeigen konnten.*

### 6.2.1 Kettenregel

In diesem Abschnitt beweisen wir eine Verallgemeinerung der Kettenregel für Funktionen mehrerer Veränderlicher, welche im nächsten Satz beschrieben ist.

**Satz 6.32** (Kettenregel)
*Seien $U \subset \mathbb{R}^n, V \subset \mathbb{R}^k$ zwei offene Teilmengen. Außerdem sei $g : U \to V$ eine im Punkt $x \in U$ und $f : V \to \mathbb{R}^m$ eine im Punkt $g(x) \in V$ total differenzierbare Funktion. Dann ist die Funktion*

$$f \circ g\colon\ U \to \mathbb{R}^m$$

*im Punkt $x \in U$ total differenzierbar und für das Differential (d.h. für die Jacobi-Matrix) gilt*

$$D(f \circ g)(x) \ = \ Df(g(x)) \cdot Dg(x).$$

*Beweis.* Wir wählen einen beliebigen Vektor $\xi \in \mathbb{R}^n$, so dass $x + \xi \in U$ ist und betrachten zunächst das Fehlerfunktional $r_g$ für $g$ bezüglich $\xi$ im Punkt $x \in U$ mit

$$r_g(\xi) \ = \ g(x+\xi) - g(x) - Dg(x) \cdot \xi \quad \Leftrightarrow \quad g(x+\xi) \ = \ g(x) + \underbrace{r_g(\xi) + Dg(x) \cdot \xi}_{:=\eta}.$$

Damit gilt für die Konkatenation der Funktionen

$$(f \circ g)(x+\xi) \ = \ f(g(x+\xi)) \ = \ f(g(x)+\eta).$$

Analog sehen wir für das Fehlerfunktional $r_f$ für $f$ bezüglich $\eta$ im Punkt $g(x) \in V$,

$$\begin{aligned} r_f(\eta) \ &= \ f(g(x)+\eta) - f(g(x)) - Df(g(x)) \cdot \eta \\ \Leftrightarrow \quad f(g(x)+\eta) \ &= \ f(g(x)) + r_f(\eta) + Df(g(x)) \cdot \eta. \end{aligned}$$

Insgesamt erhalten wir also den folgenden Zusammenhang

$$\begin{aligned} (f \circ g)(x+\xi) \ &= \ f(g(x)+\eta) \ = \ f(g(x)) + r_f(\eta) + Df(g(x)) \cdot \eta \\ &= \ f(g(x)) + r_f(\eta) + Df(g(x)) \cdot (r_g(\xi) + Dg(x) \cdot \xi). \end{aligned} \tag{6.8}$$

Durch Umstellen von (6.8) erhalten wir folgende Identität:

$$\begin{aligned} r_{f\circ g}(\xi) \ &:= \ (f \circ g)(x+\xi) - (f \circ g)(x) - Df(g(x)) \cdot Dg(x) \cdot \xi \\ &= \ r_f(\eta) + Df(g(x)) \cdot r_g(\xi). \end{aligned}$$

Es ist klar, dass der Term $(Df(g(x)) \cdot Dg(x))$ ein linearer Operator ist. Um zu zeigen, dass es sich auch wirklich um das Differential von $(f \circ g)$ im Punkt $x \in U$ handelt müssen wir zeigen, dass das Fehlerfunktional $r_{f\circ g}(\xi)$ gegen Null konvergiert, wenn $\xi$ gegen Null geht und somit $(f \circ g)$ total differenzierbar in $x \in U$ ist.

Um zeigen, dass die Konkatenation $f \circ g$ total differenzierbar in $x \in U$ ist, wählen wir ein beliebiges $\varepsilon > 0$. Da $g$ total differenzierbar in $x \in U$ ist nach Voraussetzung, wissen wir, dass ein $\delta_1 > 0$ existiert, so dass für $||\xi|| \leqslant \delta_1$ gilt

$$||r_g(\xi)|| \ \leqslant \ ||\xi|| \ \leqslant \ \delta_1.$$

Somit gilt insbesondere durch Anwendung der Dreiecksungleichung und der Cauchy-Schwarz-Ungleichung

$$||\eta|| \;=\; ||r_g(\xi) + Dg(x)\xi|| \;\leqslant\; ||\xi|| + ||Dg(x)|| \cdot ||\xi|| \;\leqslant\; ||\xi|| \cdot (1 + ||Dg(x)||).$$

Da $f$ total differenzierbar im Punkt $g(x) \in V$ nach Voraussetzung ist, wissen wir, dass ein $\delta_2 \leqslant \delta_1$ existiert, so dass für beliebiges $\tilde{\eta} \in \mathbb{R}^k$, mit $||\tilde{\eta}|| < \delta_2$ gilt, dass

$$||r_f(\tilde{\eta})|| \;\leqslant\; \varepsilon||\tilde{\eta}||.$$

Wählen wir nun

$$\delta := \frac{\delta_2}{||1 + Dg(x)||},$$

so folgt, dass $||\eta|| \leqslant \delta_2$ und somit gilt schon für alle $\xi$ mit $||\xi|| \leqslant \delta$

$$||r_f(\eta)|| \;\leqslant\; \varepsilon||\eta|| \;\leqslant\; \varepsilon||\xi|| \cdot (1 + ||Dg(x)||) \quad \Leftrightarrow \quad \frac{||r^f(\eta)||}{||\xi||} \;\leqslant\; \varepsilon(1 + ||Dg(x)||).$$

Wir haben insgesamt also

$$\lim_{\xi \to 0} \frac{||r^f(\eta)||}{||\xi||} \;=\; 0$$

gezeigt und somit gilt wegen der totalen Differenzierbarkeit von $g$ in $x \in U$ für die Konkatenation

$$\lim_{\xi \to 0} \frac{||r_{f \circ g}(\xi)||}{||\xi||} \;\leqslant\; \lim_{\xi \to 0} \frac{||r_f(\eta)||}{||\xi||} + ||Df(g(x))|| \cdot \frac{||r_g(\xi)||}{||\xi||} \;=\; 0.$$

□

Im Folgenden wollen wir die Anwendung der mehrdimensionalen Kettenregel an Hand eines einfachen Beispiels illustrieren

**Beispiel 6.33**
*Wir betrachten zwei total differenzierbare Funktionen $f, g \colon \mathbb{R}^2 \to \mathbb{R}^2$ mit*

$$f(x,y) \;:=\; \begin{pmatrix} x - y \\ xy^2 \end{pmatrix}, \qquad g(x,y) \;:=\; \begin{pmatrix} \sin(x) \\ \cos(y) \end{pmatrix}.$$

*Wir betrachten die Konkatenation $h := f \circ g \colon \mathbb{R}^2 \to \mathbb{R}^2$ der beiden Funktionen mit*

$$h(x,y) \;:=\; (f \circ g)(x,y) \;=\; f(g(x,y)) \;=\; \begin{pmatrix} \sin(x) - \cos(y) \\ \sin(x) \cdot \cos^2(y) \end{pmatrix}.$$

*Wir können die Jacobi-Matrix $J_h$ direkt berechnen als*

$$J_h(x,y) \;=\; \begin{pmatrix} \partial_x h_1(x,y) & \partial_y h_1(x,y) \\ \partial_x h_2(x,y) & \partial_y h_2(x,y) \end{pmatrix} \;=\; \begin{pmatrix} \cos(x) & \sin(y) \\ \cos(x)\cdot\cos^2(y) & -2\sin(x)\cdot\sin(y)\cdot\cos(y) \end{pmatrix}.$$

*Andererseits können wir über die mehrdimensionale Kettenregel in Satz 6.32 das Differential berechnen als*

$$D(f\circ g)(x) \;=\; D(f(g(x))\cdot Dg(x).$$

*Wir berechnen also zunächst die Jacobi-Matrizen $Df = J_f$ von $f$ und $Dg = J_g$ von $g$:*

$$J_f(x,y) \;=\; \begin{pmatrix} 1 & -1 \\ y^2 & 2xy \end{pmatrix}, \qquad J_g(x,y) \;=\; \begin{pmatrix} \cos(x) & 0 \\ 0 & -\sin(y) \end{pmatrix}.$$

*Durch Einsetzen erhalten wir also insgesamt*

$$\begin{aligned} D(f\circ g)(x) \;&=\; J_f(g(x,y))\cdot J_g(x,y) \;=\; \begin{pmatrix} 1 & -1 \\ \cos^2(y) & 2\sin(x)\cos(y) \end{pmatrix}\cdot\begin{pmatrix} \cos(x) & 0 \\ 0 & -\sin(y) \end{pmatrix} \\ &=\; \begin{pmatrix} \cos(x) & \sin(y) \\ \cos(x)\cdot\cos^2(y) & -2\sin(x)\cdot\sin(y)\cdot\cos(y) \end{pmatrix} \;=\; J_h(x,y). \end{aligned}$$

*Die mehrdimensionale Kettenregel liefert also das gleiche Ergebnis für das Differential der Konkatenation von $f$ und $g$.*

### 6.2.2 Richtungsableitung

Wir führen nun zusätzlich noch das Konzept der Richtungsableitung ein, welches analog zur partiellen Ableitung in Kapitel 6.1 Differenzen entlang eindimensionaler Linien betrachtet, mit dem wichtigen Unterschied, dass wir nun beliebige Richtungen im $\mathbb{R}^n$ zulassen werden.

**Definition 6.34**
*Sei $U \subset \mathbb{R}^n$ eine offene Teilmenge und sei $f\colon U \to \mathbb{R}$ eine Funktion. Für einen Punkt $x \in U$ und einen normierten Richtungsvektor $v \in \mathbb{R}^n$ mit $||v|| = 1$ heißt der Grenzwert (sofern er existiert)*

$$D_v f(x) \;:=\; \frac{d}{dt} f(x+tv)\Big|_{t=0} \;=\; \lim_{t\to 0}\frac{f(x+tv)-f(x)}{t}$$

Richtungsableitung *von $f$ am Punkt $x$ in Richtung $v$.*

Für stetig partiell differenzierbare Funktionen lassen sich Richtungsableitungen leicht über den Gradienten darstellen, wie der folgende Satz aussagt.

**Satz 6.35**
*Sei $U \subset \mathbb{R}^n$ eine offene Teilmenge und sei $f : U \to \mathbb{R}$ eine stetig partiell differenzierbare Funktion. Dann gilt für jeden Richtungsvektor $v \in \mathbb{R}^n$ mit $||v|| = 1$, dass gilt*

$$D_v f(x) \;=\; \langle \nabla f(x), v \rangle$$

*für alle $x \in U$.*

*Beweis.* In der Hausaufgabe zu zeigen. □

Die folgende Bemerkung motiviert die Betrachtung der speziellen Richtung des stärksten Gradientenanstiegs bzw. -abstiegs in numerischen Methoden der Optimierung.

**Bemerkung 6.36**
*Sofern $\nabla f(x) \neq 0$ können wir den Winkel $\theta := \measuredangle(\nabla f(x), v)$ zwischen $\nabla f(x)$ und $v$ definieren. In diesem Fall gilt nach Definition 3.11 die Identität*

$$\langle \nabla f(x), v \rangle \;=\; ||v|| \cdot ||\nabla f(x)|| \cdot \cos(\theta) \;=\; ||\nabla f(x)|| \cdot \cos(\theta).$$

*Dieser Ausdruck wird maximal bzw. minimal wenn für den Winkel $\theta$ gilt*

$$\begin{aligned} \cos(\theta) \;=\; 1 \quad &\Leftrightarrow \quad \theta \;=\; 0 \quad \Leftrightarrow \quad v \;=\; \nabla f(x) \\ \cos(\theta) \;=\; -1 \quad &\Leftrightarrow \quad \theta \;=\; \pi \quad \Leftrightarrow \quad v \;=\; -\nabla f(x). \end{aligned}$$

*Anschaulich bedeutet diese Beobachtung, dass am Punkt $x$ der steilste Aufstieg bzw. Abstieg in Richtung des (negativen) Gradienten erfolgt. Diese Überlegung bildet die Grundlagen vieler numerischer Optimierungsverfahren, da diese Richtung offensichtlich die Funktionswerte am stärksten verändert.*

### 6.2.3 Mittelwertsatz

Für Funktionen mehrerer Veränderlicher haben wir bisher häufig alle Koordinatenrichtungen bis auf eine fixiert haben, so dass wir effektiv den Mittelwertsatz für Funktionen in einer Veränderlichen benutzen konnten. Analog können wir reellwertige Funktionen in mehreren Veränderlichen auch entlang beliebiger Richtungen betrachten was zu folgender Aussage führt.

**Satz 6.37** (Mittelwertsatz für reellwertige Funktionen)
*Sei $U \subset \mathbb{R}^n$ eine offene Teilmenge und $f : U \to \mathbb{R}$ eine stetig partiell differenzierbare Funktion. Für einen Punkt $x \in U$ und einem Richtungsvektor $\xi \in \mathbb{R}^n$ mit $(x + t\xi) \in U$ für alle $t \in [0, 1]$ existiert ein $\theta \in [0, 1]$, so dass*

$$f(x + \xi) - f(x) \;=\; \langle \nabla f(x + \theta\xi), \xi \rangle.$$

*Beweis.* Wir betrachten die eindimensionale Einschränkung $g(t) := f(x + t\xi)$ von $f$ und sehen mit Hilfe des Mittelwertsatzes für Funktionen in einer Veränderlichen [Burger2020, Kapitel 6.2], dass ein $\theta \in [0, 1]$ existiert, so dass wir mit der Kettenregel aus Satz 6.32 erhalten

$$\begin{aligned} f(x+\xi) - f(x) &= g(1) - g(0) = g'(\theta) \cdot 1 \\ &= \nabla f(g(\theta)) \cdot g'(\theta) = \langle \nabla f(x + \theta\xi), \xi \rangle. \end{aligned}$$

□

Die obige Darstellung des Mittelwertsatzes funktioniert leider nur für reellwertige Funktionen, wie die folgende Bemerkung feststellt.

**Bemerkung 6.38**
*Für vektorwertige Funktionen $f : U \to \mathbb{R}^m$ scheitert die Überlegung aus Satz 6.37 leider, da wir hier verschiedene unabhängige Komponenten im Bildbereich haben. Wir müssten also den Mittelwertsatz für reellwertige Funktionen auf jede Komponente $f_i$ von $f$ einzeln anwenden und erhalten in obiger Notation $m$ verschiedene Zwischenstellen $x + \theta_i\xi \in \mathbb{R}^n$. Da diese Zwischenstellen im Allgemeinen verschieden sind scheitert das Argument.*

*Das Konzept lässt sich allerdings durch folgende Überlegungen verallgemeinern. Für eine stetig differenzierbare Funktion $f : I \to \mathbb{R}$, wobei $I \subset \mathbb{R}$ eine offene Teilmenge sei, folgt aus dem Hauptsatz der Differential- und Integralrechnung [Burger2020, Kapitel 7.2], dass*

$$f(x+\xi) - f(x) = \left( \int_0^1 f'(x + t\xi) \, \mathrm{d}t \right) \cdot \xi.$$

*Diese Integralform lässt sich nun auch auf Funktionen mit mehreren Komponenten übertragen. Dazu bemerken wir kurz, dass das Integral einer matrixwertigen Funktion $A : \mathbb{R} \to \mathbb{R}^{n \times m}$ durch das Integral der einzelnen Matrix-Einträge gegeben ist, das heißt es gilt*

$$\left( \int_{t_0}^{t_1} A(t) \, \mathrm{d}t \right)_{i,j} = \int_{t_0}^{t_1} A_{i,j}(t) \, \mathrm{d}t$$

*für $1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m$.*

Basierend auf der Beobachtung in Bemerkung 6.38 können wir im Folgenden den Mittelwertsatz für vektorwertige Funktionen in mehreren Veränderlichen formulieren.

**Satz 6.39** (Mittelwertsatz)
*Sei $U \subset \mathbb{R}^n$ eine offene Teilmenge und sei $f : U \to \mathbb{R}^m$ eine stetig partiell differenzierbare, vektorwertige Funktion. Für einen beliebigen Punkt $x \in U$ und einen Richtungsvektor $\xi \in \mathbb{R}^n$ mit $x + t\xi \in U$ für alle $t \in [0, 1]$ gilt*

$$f(x+\xi) - f(x) = \left( \int_0^1 Df(x + t\xi) \, \mathrm{d}t \right) \cdot \xi.$$

*Beweis.* Für jede Komponente $f_i$ im Bild von $f$ betrachten wir eine eindimensionale Funktion $g_i : [0,1] \to \mathbb{R}$ mit

$$g_i(t) \;:=\; f_i(x + t\xi).$$

Wenden wir auf diese Funktionen den Hauptsatz der Differential- und Integralrechnung [Burger2020, Kapitel 7.2] an und benutzen die Darstellung der Richtungsableitung aus Satz 6.35, so sehen wir, dass für jede Komponente $i \in \{1, \ldots, m\}$ gilt

$$\begin{aligned} f_i(x+\xi) - f_i(x) \;&=\; g_i(1) - g_i(0) \;=\; \int_0^1 g_i'(t)\,\mathrm{d}t \\ &=\; \int_0^1 \langle \nabla f_i(x+t\xi), \xi\rangle\,\mathrm{d}t \;=\; \big\langle \int_0^1 \nabla f_i(x+t\xi)\,\mathrm{d}t, \xi\big\rangle. \end{aligned}$$

□

Der allgemeine Mittelwertsatz 6.39 erlaubt es uns zusätzlich eine sehr praktische Norm-Abschätzung herzuleiten. Dafür benötigen wir jedoch zunächst folgendes Hilfslemma.

**Lemma 6.40**
*Sei $f : [t_0, t_1] \to \mathbb{R}^m$ eine stetige Funktion, dann gilt*

$$\left\| \int_{t_0}^{t_1} f(t)\,\mathrm{d}t \right\| \;\leqslant\; \int_{t_0}^{t_1} ||f(t)||\,\mathrm{d}t.$$

*Beweis.* In der Hausaufgabe zu zeigen. □

Mit Hilfe des Lemmas 6.40 können wir nun ein nützliche Abschätzung für die Abstand zweier Funktionswerte in Abhängigkeit des Differentials zeigen.

**Satz 6.41**
*Sei $U \subset \mathbb{R}^n$ eine offene Teilmenge und sei $f : U \to \mathbb{R}^m$ eine stetig partiell differenzierbare, vektorwertige Funktion. Für einen Punkt $x \in U$ und einem Richtungsvektor $\xi \in \mathbb{R}^n$ mit $(x + t\xi) \in U$ für alle $t \in [0,1]$ sei außerdem*

$$M \;:=\; \sup_{t\in[0,1]} ||Df(x+t\xi)||.$$

*Dann gilt die folgende Abschätzung*

$$||f(x+\xi) - f(x)|| \;\leqslant\; M \cdot ||\xi||.$$

*Beweis.* In der Hausaufgabe zu zeigen. □

## 6.3 Taylor-Formel

Wie wir schon anhand der Definition des Fehlerfunktionals bei der totalen Differenzierbarkeit gesehen haben, liefert das Differential einer Funktion $f$ an einem Punkt $x$ eine affinlineare Approximation der Funktion an dieser Stelle. Mithilfe der Taylor-Formel wollen wir nun eine bessere Annäherung an die Funktion erhalten, in dem wir höhere Ableitungen hinzunehmen. Diese Formel ist ein sehr nützliches Werkzeug bei der Approximation von Funktionen durch einfachere Polynome und erfährt beispielsweise in der Physik oder in der Numerik regelmäßig Anwendung. Dadurch können beispielsweise nichtlineare Funktionen in einer lokalen Umgebung linear angenähert werden, was viele Berechnungen und Abschätzungen vereinfacht.

Bevor wir die Taylor-Formel für Funktionen in einer Veränderlichen betrachten, benötigen wir das folgende Hilfslemma.

**Lemma 6.42** (Satz von Rolle)
*Sei $[a,b] \subset \mathbb{R}$ ein Intervall mit $a < b$ und sei $f\colon [a,b] \to \mathbb{R}$ eine stetige Funktion, die auf dem offenen Intervall $(a,b)$ differenzierbar ist. Es sei außerdem $f(a) = f(b)$.*

*Dann existiert eine Punkt $x_0 \in (a,b)$, so dass $f'(x_0) = 0$.*

*Beweis.* Falls $f$ eine konstante Funktion ist mit $f(x) \equiv f(a) = f(b)$ für alle $x \in [a,b]$, so ist die Aussage des Lemmas trivialerweise erfüllt, da für konstante Funktionen $f'(x_0) = 0$ für alle $x_0 \in (a,b)$ gilt.

Sei also $f$ nicht konstant. Da $f$ auf dem kompakten Intervall $[a,b] \subset \mathbb{R}$ stetig ist nach Voraussetzung, nimmt die Funktion auf dem Intervall nach Satz dem Satz von Weierstraß [Burger2020, Satz 5.19] an einer Stelle $m \in [a,b]$ ein Minimum und an einer Stelle $M \in [a,b]$ ein Maximum an. Da $f$ nicht konstant ist muss wegen $f(a) = f(b)$ mindestens $m \in (a,b)$ oder $M \in (a,b)$ gelten. Wir bezeichnen diese Extremstelle als den Punkt $x_0 \in (a,b)$.

Falls es sich bei $x_0$ um ein Maximum handelt, so folgt aus der Differenzierbarkeit von $f$ in $(a,b)$, dass

$$f'(x_0) \;=\; \lim_{h \searrow 0} \frac{f(x_0+h) - f(x_0)}{h} \;\leqslant\; 0 \quad \text{und} \quad f'(x_0) \;=\; \lim_{h \nearrow 0} \frac{f(x_0+h) - f(x_0)}{h} \;\geqslant\; 0.$$

Damit folgt schon, dass $f'(x_0) = 0$ sein muss.

Falls $f$ in $x_0$ ein Minimum annimmt, so kann man analog argumentieren und erhält auch in diesem Fall $f'(x_0) = 0$. □

Mit Hilfe des obigen Lemmas können wir nun einen der wichtigsten Sätze zur Approximation von beliebigen Funktionen formulieren. Die Taylor-Formel erlaubt die Annäherung einer Funktion durch ein Polynom und erlaubt es den Fehler bei dieser Approximation abzuschätzen.

**Satz 6.43** (Eindimensionale Taylor-Formel)
*Es sei $I \subset \mathbb{R}$ ein offenes Intervall und es sei $f : I \to \mathbb{R}$ eine $(k+1)$-mal stetig differenzierbare Funktion für $k \in \mathbb{N}$. Dann existiert für ein beliebiges kompaktes Intervall $[t_0, t_1] \subset I$ ein $\theta \in [t_0, t_1]$, so dass die folgende* Taylor-Formel *gilt*

$$f(t_1) \;=\; \sum_{i=0}^{k} \frac{d^i}{dt^i} f(t_0) \cdot \frac{(t_1 - t_0)^i}{i!} + \frac{d^{k+1}}{dt^{k+1}} f(\theta) \cdot \frac{(t_1 - t_0)^{k+1}}{(k+1)!}$$

*Hierbei folgen wir der Konvention, dass $0! := 1$ ist.*

*Beweis.* Für ein beliebiges $a \in \mathbb{R}$ betrachten wir die Hilfsfunktion

$$g(t) \;:=\; f(t_1) - f(t) - \sum_{i=1}^{k} \frac{d^i}{dt^i} f(t) \cdot \frac{(t_1 - t)^i}{i!} - a \cdot \frac{(t_1 - t)^{k+1}}{(k+1)!}.$$

Offensichtlich ist $t_1 \in I$ eine Nullstelle von $g$ mit $g(t_1) = 0$. Wenn wir nun $a \in R$ wählen als

$$a \;:=\; \frac{(k+1)!}{(t_1 - t_0)^{k+1}} \left( f(t_1) - f(t_0) - \sum_{i=1}^{k} \frac{d^i}{dt^i} f(t_0) \cdot \frac{(t_1 - t_0)^i}{i!} \right) \tag{6.9}$$

so erhalten wir auch $g(t_0) = 0$.

Nach Voraussetzung ist $g$ stetig differenzierbar und somit insbesondere stetig. Damit folgt nach dem Satz 6.42 von Rolle, dass ein $\theta \in [t_0, t_1]$ existiert, so dass $g'(\theta) = 0$ gilt. Diese Stelle $\theta$ können wir nutzen um den Term $a$ noch näher zu charakterisieren. Wir erhalten mit der Produktregel der Differentialrechnung, dass

$$\begin{aligned} g'(\theta) \;&=\; -f'(\theta) - \sum_{i=1}^{k} \frac{d^{i+1}}{dt^{i+1}} f(\theta) \cdot \frac{(t_1 - \theta)^i}{i!} + \sum_{i=1}^{k} \frac{d^i}{dt^i} f(\theta) \cdot \frac{(t_1 - \theta)^{i-1}}{(i-1)!} + a \cdot \frac{(t_1 - \theta)^k}{k!} \\ &=\; -\frac{d^{k+1}}{dt^{k+1}} f(\theta) \cdot \frac{(t_1 - \theta)^k}{k!} + a \cdot \frac{(t_1 - \theta)^k}{k!} \;=\; \frac{(t_1 - \theta)^k}{k!} \cdot \left( a - \frac{d^{k+1}}{dt^{k+1}} f(\theta) \right) \;=\; 0. \end{aligned}$$

Wir sehen also, dass der Term $a$ in (6.9) gleichzeitig folgende Identität erfüllt:

$$a \;=\; \frac{d^{k+1}}{dt^{k+1}} f(\theta).$$

Setzen wir diese Darstellung in die Hilfsfunktion $g$ ein und werten diese in $t_0 \in I$ aus, so erhalten wir schließlich:

$$g(t_0) \;=\; f(t_1) - f(t_0) - \sum_{i=1}^{k} \frac{d^i}{dt^i} f(t) \cdot \frac{(t_1 - t_0)^i}{i!} - \frac{d^{k+1}}{dt^{k+1}} f(\theta) \cdot \frac{(t_1 - t_0)^{k+1}}{(k+1)!} \;=\; 0.$$

Durch Umstellen folgt nun direkt die Taylor-Formel. □

Im Folgenden wollen wir einige Beobachtungen zur Taylor-Formel festhalten.

**Bemerkung 6.44**
*Die zusätzlichen Bemerkungen helfen dabei die Taylor-Formel in Satz 6.43 vernünftig zu interpretieren.*

1. *Wir nennen*

$$f(t_1) \;=\; \underbrace{\sum_{i=0}^{k} \frac{d^i}{dt^i} f(t_0) \cdot \frac{(t_1 - t_0)^i}{i!}}_{=: T_k f(t_1; t_0)} + \underbrace{\frac{d^{k+1}}{dt^{k+1}} f(\theta) \cdot \frac{(t_1 - t_0)^{k+1}}{(k+1)!}}_{=: R_k f(t_1; t_0)}$$

   *die* Taylor-Entwicklung *$k$-ter Ordnung von $f(t_1)$ im Punkt $t_0 \in I$. Sie gliedert sich in eine Approximation der ursprünglichen Funktion $f$ im Punkt $t_1 \in I$ durch ein (Taylor-)Polynom $T_k f(t_1; t_0)$ vom Grad $k$ ausgewertet in $t_1$ und einen Fehlerterm $R_k f(t_1; t_0)$.*

2. *Der Term $R_k f(t_1; t_0)$ wird* Lagrangesches Restglied *der Taylor-Entwicklung $k$-ter Ordnung genannt. Er beschreibt den Fehler, den man macht, wenn man den Funktionswert $f(t_1)$ durch eine Auswertung des Taylorpolynoms von Grad $k$ im Punkt $t_0 \in I$ approximiert.*

   *Es wird klar, dass dieser Fehler maßgeblich durch den Abstand der beiden Punkte $t_0, t_1 \in I$ zueinander bestimmt wird. Je näher der Entwicklungspunkt $t_0$ an der untersuchten Stelle $t_1$ liegt, desto besser ist die Approximation.*

Das folgende Beispiel soll uns helfen zu verstehen, wie die Taylor-Entwicklung zur Approximation einer Funktion konkret eingesetzt werden kann.

**Beispiel 6.45**
*Wir betrachten eine Funktion $f\colon \mathbb{R} \to \mathbb{R}$ mit $f(x) := e^x$. Wir möchten den Funktionswert von $f$ in $t_1 = 1$ approximieren durch eine Taylor-Entwicklung im Punkt $t_0 = 0$.*

*Hierzu formulieren wir zunächst die Taylor-Entwicklung* ***erster Ordnung****:*

$$\begin{aligned} f(1) \;=\; T_1 f(1;0) + R_1 f(1;0) \;&=\; f(0) \cdot 1 + f'(0) \cdot \frac{(1-0)^1}{1!} + R_1 f(1;0) \\ &\approx\; f(0) + f'(0) \;=\; e^0 + e^0 \;=\; 2. \end{aligned}$$

*Durch den Wegfall des Restglieds $R_1 f(1;0)$ wird der Funktionswert nur linear approximiert. Vergleichen wir diese Approximation mit dem echten Funktionswert, so erhalten wir*

$$|f(1) - T_1 f(1;0)| \;=\; |e^1 - 2| \;\approx\; 0.72.$$

*Betrachten wir nun die Taylor-Entwicklung* ***zweiter Ordnung****:*

$$f(1) \;=\; T_2 f(1;0) + R_2 f(1;0) \;=\; f(0)\cdot 1 + f'(0)\cdot\frac{(1-0)^1}{1!} + f''(0)\cdot\frac{(1-0)^2}{2!} + R_2 f(1;0)$$
$$\approx\; f(0) + f'(0) + \frac{1}{2}f''(0) \;=\; e^0 + e^0 + \frac{1}{2}e^0 \;=\; 2.5.$$

*Durch den Wegfall des Restglieds $R_2 f(1;0)$ wird der Funktionswert nun quadratisch approximiert. Vergleichen wir diese Approximation mit dem echten Funktionswert, so erhalten wir*

$$|f(1) - T_2 f(1;0)| \;=\; |e^1 - 2.5| \;\approx\; 0.22.$$

*Wir erkennen also, dass die Approximation durch Erhöhung der Ordnung der Taylorentwicklung genauer wird.*

Die Taylor-Formel aus Satz 6.43 gilt auch für vektorwertige Funktionen in mehreren Veränderlichen. Für diese Verallgemeinerung müssen wir jedoch zunächst die folgende Notation einführen.

**Definition 6.46** (Multiindex)
*Für ein $n$-Tupel $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n$, genannt* Multiindex*, definieren wir die folgenden Operationen*

$$|\alpha| \;:=\; \sum_{i=1}^{n} \alpha_i \qquad \textit{und} \qquad \alpha! \;:=\; \prod_{i=1}^{n} (\alpha_i!).$$

*Für eine $|\alpha|$-mal stetig partiell differenzierbare Funktion $f$ definieren wir weiterhin*

$$\partial^\alpha f = \partial_1^{\alpha_1} \ldots \partial_n^{\alpha_n} f.$$

*Das heißt wir leiten die Funktion $f$ in jede Koordinatenrichtung $i \in \{1, \ldots n\}$ genau $\alpha_i$-mal ab. Analog setzen wir*

$$x^\alpha \;:=\; \prod_{i=1}^{n} x_i^{\alpha_i}.$$

Ähnlich zu den Überlegung zum Mittelwertsatz betrachten wir nun zunächst eine Funktion $g(t) := f(x + t\xi)$, das heißt eine eindimensionale Funktion welche entlang einer Richtung $\xi \in \mathbb{R}^n$ vom Punkt $x$ aus betrachtet wird. Das Konzept entlang von eindimensionalen Strahlen zu operieren ist elementar für die Taylor-Formel, weshalb wir zunächst folgendes Hilfsresultat zeigen.

**Lemma 6.47**
*Sei $U \subset \mathbb{R}^n$ eine offene Teilmenge und sei $f : U \to \mathbb{R}^m$ eine $k$-mal stetig partiell differenzierbare Abbildung. Für einen beliebigen Punkt $x \in U$ und einen Richtungsvektor$\xi \in \mathbb{R}^n$ mit $x + t\xi \in U$ für alle $t \in [0, 1]$, gilt für die $k$-te Ableitung der Funktion $g(t) := f(x + t\xi)$*

$$\frac{d^k}{dt^k} g(t) \;=\; \sum_{|\alpha|=k} \partial^\alpha f(x + t\xi) \cdot \frac{k!}{\alpha!} \cdot \xi^\alpha.$$

*Beweis.* Im ersten Schritt zeigen wir zunächst die Identität

$$\frac{d^k}{dt^k} g(t) \;=\; \sum_{i_1=1}^{n} \dots \sum_{i_k=1}^{n} \partial_{i_k} \dots \partial_{i_1} f(x + t\xi) \cdot \xi_{i_1} \cdot \ldots \cdot \xi_{i_k}$$

mittels vollständiger Induktion über $k \in \mathbb{N}$.

**Induktionsanfang $k = 1$:**
In diesem Fall betrachten wir die Funktion $\phi(t) := x + t\xi$ und erhalten so die Verkettung $g = f \circ \phi$. Somit können wir die Kettenregel aus Satz 6.32 benutzen und erhalten, dass

$$\frac{d}{dt} g(t) \;=\; \nabla f(\phi(t)) \cdot D\phi(t) \;=\; \sum_{i=1}^{n} \partial_i f(x + t\xi) \cdot \xi_i.$$

**Induktionsschritt $(k-1) \to k$:**
Die Induktionsannahme ist, dass die Aussage bereits für $k - 1$ gelte. Mit der gleichen Rechnung wie im Fall $k = 1$ sehen wir dann,

$$\begin{aligned}
\frac{d^k}{dt^k} g(t) \;&=\; \frac{d}{dt} \left( \frac{d^{k-1}}{dt^{k-1}} g \right)(t) \\
&=\; \frac{d}{dt} \left( \sum_{i_1,\dots,i_{k-1}=1}^{n} \partial_{i_{k-1}} \dots \partial_{i_1} f(x + t\xi) \cdot \xi_{i_1} \cdot \ldots \cdot \xi_{i_{k-1}} \right) \\
&= \sum_{j=1}^{n} \partial_j \left( \sum_{i_1,\dots,i_{k-1}=1}^{n} \partial_{i_{k-1}} \dots \partial_{i_1} f(x + t\xi) \cdot \xi_{i_1} \cdot \ldots \cdot \xi_{i_{k-1}} \right) \cdot \xi_j.
\end{aligned}$$

Um die eigentliche Behauptung zu zeigen müssen wir nun lediglich die Indizierung über das Tupel $(i_1, \dots, i_k)$ in die Multiindex Schreibweise umrechnen. Das funktioniert folgendermaßen: Für ein Tupel $\iota = (i_1, \dots, i_k)$ definieren wir den zugehörigen Multiindex $\alpha^\iota \in \mathbb{N}^n$ durch

$$\alpha^\iota_i \;:=\; \sum_{j=0}^{k} \delta_{i_j,i},$$

wobei

$$\delta_{i_j,i} \;=\; \begin{cases} 1 & \text{falls } i_j = i \\ 0 & \text{sonst.} \end{cases}$$

Anschaulich gesprochen zählt der Eintrag $\alpha^\iota_i$ wie oft der Index $i$ im Tupel $\iota$ vorkommt. Offensichtlich gilt für derartige Multiinidizes stets

$$|\alpha^\iota| \;=\; k$$

und insbesondere

$$\partial_{i_k} \ldots \partial_{i_1} f(x+t\xi) \cdot \xi_{i_1} \cdot \ldots \cdot \xi_{i_k} \;=\; \partial^{\alpha^\iota} f(x+t\xi) \cdot \xi^{\alpha^\iota}.$$

Der letzte Schritt um die Behauptung zu zeigen ist zu zählen wie viele Tupel $\iota = (i_1, \ldots, i_k)$ auf den gleichen Multiindex $\alpha^\iota$ führen durch obige Konstruktion. Aus der Kombinatorik wissen wir, dass es genau $\frac{k!}{\alpha!}$ verschiedene solcher Tupel gibt, was man folgendermaßen sieht:

- Zunächst wollen wir $\alpha_1$-mal den Index 1 auf $k$ Plätze verteilen, wofür es

  $$\binom{k}{\alpha_1} \;=\; \frac{k!}{\alpha_1! \cdot (k-\alpha_1)!}$$

  verschiedene Möglichkeiten gibt.

- Im zweiten Schritt verteilen wir $\alpha_2$-mal den Index 2 auf die verbliebenen $k-\alpha_1$ Plätze wofür es dann

  $$\binom{k-\alpha_1}{\alpha_2} \;=\; \frac{(k-\alpha_2)!}{\alpha_1! \cdot (k-\alpha_1-\alpha_2)!}$$

  verschiedene Möglichkeiten gibt.

- Führen wir diesen Prozess iterativ weiter und multiplizieren die jeweiligen Möglichkeiten auf, so erhalten wir insgesamt

  $$\frac{k! \cdot (k-\alpha_1) \cdot \ldots (k-\alpha_1-\ldots-\alpha_n)}{\alpha_1! \cdot (k-\alpha_1) \cdot \ldots \cdot (k-\alpha_1-\ldots-\alpha_n) \cdot \alpha_n! \cdot 0!} \;=\; \frac{k!}{\alpha!}.$$

  Dies vervollständigt den Beweis weil wir nun die folgende Identität gezeigt haben

$$\begin{aligned} \frac{d^k}{dt^k} g(t) \;&=\; \sum_{i_1,\ldots,i_k=1}^{n} \partial_{i_k} \ldots \partial_{i_1} f(x+t\xi) \cdot \xi_{i_1} \cdot \ldots \cdot \xi_{i_k} \\ &=\; \sum_{|\alpha|=k} \partial^\alpha f(x+t\xi) \cdot \frac{k!}{\alpha!} \cdot \xi^\alpha. \end{aligned}$$

□

Diese Hilfsaussage erlaubt es uns nun den eigentlichen Hauptsatz dieses Abschnitts zu zeigen, die mehrdimensionale Taylor-Formel.

**Satz 6.48** (Mehrdimensionale Taylor-Formel)
*Sei $U \subset \mathbb{R}^n$ eine offene Teilmenge und $f : U \to \mathbb{R}^m$ eine $(k+1)$-mal stetig partiell differenzierbare Abbildung. Für einen beliebigen Punkt $x \in U$ und einen Richtungsvektor $\xi \in \mathbb{R}^n$ mit $x + t\xi \in U$ für alle $t \in [0,1]$, existiert ein $\theta \in [0,1]$, so dass*

$$f(x+\xi) \;=\; \sum_{|\alpha| \leqslant k} \partial^\alpha f(x) \cdot \frac{1}{\alpha!} \cdot \xi^\alpha + \sum_{|\alpha| = k+1} \partial^\alpha f(x+\theta\xi) \cdot \frac{1}{\alpha!} \cdot \xi^\alpha.$$

*Beweis.* Wir betrachten erneut die eindimensionale Funktion

$$g(t) \;:=\; f(x+t\xi)$$

und da $g$ somit eine $(k+1)$-mal differenzierbare Funktion ist können wir die eindimensionale Taylor-Formel aus Satz 6.43 anwenden und erhalten somit ein $\theta \in [0,1]$, so dass

$$g(1) \;=\; \sum_{i=0}^{k} \frac{d^i}{dt^i} g(0) \cdot \frac{1}{i!} + \frac{d^{k+1}}{dt^{k+1}} g(\theta) \cdot \frac{1}{(k+1)!}.$$

In dieser Form können wir nun Lemma 6.47 auf jede der Ableitungen $\frac{d^i}{dt^i} g$ anwenden, was die Aussage beweist. □

Die folgende Bemerkung hält einige Beobachtungen zur Taylor-Formel fest.

**Bemerkung 6.49**
*Wir bemerken zur mehrdimensionalen Taylor-Formel in Satz 6.48 die folgenden Aussagen.*

1. *Wir sehen ein, dass der Mittelwertsatz 6.37 ein Spezialfall der Taylor-Formel für $k = 0$ ist, da er aussagt, dass*

$$f(x+\xi) - f(x) \;=\; \langle \nabla f(x+\theta\xi), \xi \rangle \quad \text{für ein} \quad \theta \in [0,1].$$

2. *Oftmals spricht man im Zusammenhang mit der Taylor-Form auch von den sogenannten* Taylorpolynomen*. In der typischen Schreibweise wird nun $x$ zur Variable, wir fixieren $a \in \mathbb{R}^n$ und nennen dann*

$$T_k(x;a) \;:=\; \sum_{|\alpha| \leqslant k} \partial^\alpha f(a) \cdot \frac{(x-a)^\alpha}{\alpha!}$$

*Taylorpolynom der Ordnung $k$ an der Entwicklungsstelle $a$. Hierbei ist zu beachten, dass $T_k(\cdot;a)$ ein Polynom ist für jedes* fixierte *$a \in \mathbb{R}^n$. Wir bemerken auch, dass in dieser Schreibweise $(x-a)$ nun die Rolle von $\xi$ in Satz 6.48 inne hat, während die Rolle von $x$ durch $a$ übernommen wurde.*

Im obigen Satz haben wir vorausgesetzt, dass die Funktion $(k+1)$-mal stetig partiell differenzierbar ist. Ist sie jedoch nur $k$-mal stetig partiell differenzierbar, so erhalten wir keinen expliziten Ausdruck für das Restglied, aber zumindest eine Fehlerabschätzung.

**Korollar 6.50**
*Sei $U \subset \mathbb{R}^n$ eine offene Teilmenge und sei $f : U \to \mathbb{R}$ eine $k$-mal stetig partiell differenzierbare Funktion. Dann gilt für jeden Punkt $x \in U$, dass die Funktion*

$$r(\xi) \;:=\; f(x+\xi) - \sum_{|\alpha| \leqslant k} \partial^\alpha f(x) \cdot \frac{1}{\alpha!} \cdot \xi^\alpha$$

*folgende Grenzwert-Eigenschaft hat:*

$$\lim_{\xi \to 0} \frac{|r(\xi)|}{||\xi||^k} \;=\; 0.$$

*Beweis.* Wir wählen $\delta > 0$, so dass $B_\delta(x) \subset U$. Wir wissen nach der Taylor-Formel in Satz 6.48, dass für $\xi \in B_\delta(x)$ ein $\theta \in [0,1]$ existiert, so dass

$$\begin{aligned} f(x+\xi) \;&=\; \sum_{|\alpha| \leqslant k-1} \partial^\alpha f(x) \cdot \frac{1}{\alpha!} \cdot \xi^\alpha + \sum_{|\alpha|=k} \partial^\alpha f(x+\theta\xi) \cdot \frac{1}{\alpha!} \cdot \xi^\alpha \\ &=\; \sum_{|\alpha| \leqslant k} \partial^\alpha f(x) \cdot \frac{1}{\alpha!} \cdot \xi^\alpha + \underbrace{\sum_{|\alpha|=k} (\partial^\alpha f(x+\theta\xi) - \partial^\alpha f(x)) \cdot \frac{1}{\alpha!} \cdot \xi^\alpha}_{=r(\xi)}. \end{aligned}$$

Weiterhin sehen wir, dass

$$|\xi^\alpha| \;\leqslant\; ||\xi||^{\alpha_1} \cdot \ldots \cdot ||\xi||^{\alpha_n} \;=\; ||\xi||^k.$$

Somit erhalten wir also

$$|r(\xi)| \;\leqslant\; ||\xi||^k \cdot \sum_{|\alpha|=k} |(\partial^\alpha f)(x+\theta\xi) - (\partial^\alpha f)(x)| \cdot \frac{1}{\alpha!}.$$

Wegen der Stetigkeit aller partiellen Ableitungen folgt, dass für jedes $\varepsilon > 0$ ein $\delta > 0$ existiert, so dass

$$|(\partial^\alpha f)(x+\theta\xi) - (\partial^\alpha f)(x)| \;\leqslant\; \varepsilon$$

für alle $\xi \in B_\delta(x)$ und alle $\alpha$ mit $|\alpha| \leqslant k$. Hierbei ist zu beachten, dass $\theta$ zwar jeweils von $\xi$ abhängt, aber da $\theta \in [0,1]$ gilt, wissen wir, dass

$$\xi \in B_\delta(x) \;\Rightarrow\; \theta\xi \in B_\delta(x).$$

Insgesamt haben wir damit

$$\lim_{\xi \to 0} \frac{|r(\xi)|}{||\xi||^k} \;=\; \lim_{\xi \to 0} \sum_{|\alpha|=k} |\partial^\alpha f(x+\theta\xi) - \partial^\alpha f(x)| \cdot \frac{1}{\alpha!} \;=\; 0.$$

□

**Bemerkung 6.51**
*Im Sinne des Korollars 6.50 gilt für den Fehler des Taylorpolynoms an einer Stelle $x$, dass der Approximationsfehler*

$$r(x) \;:=\; f(x) - T_k(x,a)$$

*den Grenzwert*

$$\lim_{x \to a} \frac{|r(x)|}{||x-a||^k} \;=\; 0$$

*erfüllt.*

*Speziell für den Fall $k = 1$ sehen wir, dass*

$$\sum_{|\alpha|=1} \partial^\alpha f(x) \cdot \frac{1}{\alpha!} \cdot \xi^\alpha \;=\; \sum_{i=1}^{n} \partial_i f(x) \cdot \xi_i \;=\; \langle \nabla f(x), \xi \rangle$$

*gilt. Für den Fall $k = 2$ erhalten wir*

$$\sum_{|\alpha|=2} \partial^\alpha f(x) \cdot \frac{1}{\alpha!} \cdot \xi^\alpha \;=\; \sum_{i,j=1}^{n} \partial_i \partial_j f(x) \cdot \frac{1}{2} \cdot \xi_i \cdot \xi_j \;=\; \frac{1}{2} \cdot \langle \xi, H_f(x)\xi \rangle,$$

*wobei $H_f(x)$ die Hesse-Matrix von $f$ an der Stelle $x$ bezeichnet.*

Da der Fall $k = 2$ in obiger Bemerkung relevant für die grundlegenden Aussagen der Optimierung im nächsten Kapitel sind, fassen wir die obigen Überlegungen in einem Korollar zusammen.

**Korollar 6.52**
*Sei $U \subset \mathbb{R}^n$ eine offene Teilmenge und sei $f : U \to \mathbb{R}$ eine zweimal stetig partiell differenzierbare Funktion. Für jeden Punkt $x \in U$ existiert dann ein $\delta > 0$ und eine Funktion $r \colon B_\delta(x) \to \mathbb{R}$, so dass*

$$f(x+\xi) \;=\; f(x) + \langle \nabla f(x), \xi \rangle + \frac{1}{2} \langle H_f(x)\xi, \xi \rangle + r(\xi),$$

*für alle $\xi \in B_\delta(0)$ gilt. Außerdem gilt*

$$\lim_{\xi \to 0} \frac{|r(\xi)|}{||\xi||^2} \;=\; 0.$$

# Kapitel 7

# Optimierung

In der Physik spielt Optimierung eine wesentliche Rolle bei der Modellierung von Energiezuständen auf unterschiedlichen Skalen: Moleküle formieren sich in einer Art, die die Gesamtenergie des Teilchensystems unter Berücksichtigung aller wechselseitigen Kräfte minimiert. Gleichzeitig strebt das Universum mit all seinen Planeten, Sternen und Galaxien nach einem Zustand von maximaler Verteilung, beschrieben durch die thermodynamische Größe der Entropie. Auch hier folgt die Zunahme der Entropie dem Prinzip der Energieminimierung des Gesamtsystems.

Menschen betreiben seit jeher Optimierung in den verschiedensten Anwendungen, oft mit unterschiedlichen Motivationen. Flugzeuge werden von Ingenieuren so entworfen und gebaut, dass sie möglichst stromlinienförmig aussehen, um damit den Reibungswiderstand in der Luft zu minimieren und gleichzeitig den nötigen Auftrieb für einen sicheren Flug zu erzeugen. Fondmanager streben danach Portfolios zu erstellen, deren Gewinn möglichst maximal ist und dennoch Spekulationsrisiken vermeiden.

Im Folgenden wollen wir die mathematischen Grundlagen zur Untersuchung von allgemeinen Optimierungsproblemen einführen. Wir beginnen mit der Definition des allgemeinen Optimierungsproblems, welches wir im weiteren Verlauf noch näher spezifizieren werden.

**Definition 7.1** (Allgemeines Optimierungsproblem)
*Sei $\Omega \subset \mathbb{R}^n$ eine offene, zusammenhängende Teilmenge und sei $F\colon \Omega \to \mathbb{R}$ eine reellwertige Funktion, welche wir* Zielfunktion *nennen. Unser Ziel ist es einen unbekannten Vektor $x \in \Omega$, auch* Parametervektor *genannt, zu finden, welcher das folgende* allgemeine Optimierungsproblem *löst.*

$$\min_{x\in\Omega} F(x) \quad \textit{mit} \quad \begin{cases} c_i(x) \;=\; 0, & i \in \mathcal{E}, \\ c_j(x) \;\geqslant\; 0, & j \in \mathcal{I}. \end{cases} \tag{7.1}$$

*Die reellwertigen Funktionen $c_i, c_j\colon \Omega \to \mathbb{R}$ bilden einen Vektor von* Nebenbedingungen, *welcher das Optimierungsproblem restringiert. Die Indexmengen $\mathcal{E}$ und $\mathcal{I}$ legen hierbei*

*fest, ob es ich bei der jeweiligen Nebenbedingung um eine Gleichung oder eine Ungleichung handelt.*

Zur Veranschaulichung betrachten wir ein zweidimensionales Beispiel für ein beschränktes, nichtlineares Optimierungsproblem.

**Beispiel 7.2**
*Wir interessieren uns für das folgende Optimierungsproblem*

$$\min_{x\in\mathbb{R}^2} (x_1-2)^2 + (x_2-1)^2$$

*unter den Nebenbedingungen* $x_1^2 - x_2 \leqslant 0$ *und* $x_1 + x_2 \leqslant 2$. *Wir können dieses Problem in die allgemeine Form des Optimierungsproblems* (7.1) *umschreiben als:*

$$\min_{x\in\mathbb{R}^2} F(x) \;=\; \min_{x\in\mathbb{R}^2} (x_1-2)^2 + (x_2-1)^2 \quad \textit{mit} \quad \begin{cases} c_1(x) \;=\; -x_1^2 + x_2 \;\geqslant\; 0, \\ c_2(x) \;=\; -x_1 - x_2 + 2 \;\geqslant\; 0. \end{cases}$$

*Hierbei gilt für die Indexmengen* $\mathcal{I} = \{1,2\}$ *und* $\mathcal{E} = \varnothing$. *Visualisiert man die Niveaulinien der Zielfunktion* $F$ *zusammen mit den Nebenbedingungen, so sieht man, dass das globale Minimum der quadratischen Funktion* $F$, *nämlich* $x = (x_1, x_2)^T = (2,1)^T$, *nicht in der erlaubten Menge der Parameter liegt, welche durch die Nebenbedingungen beschrieben ist. Trotzdem existiert ein eindeutiges globales Minimum des beschränkten Optimierungsproblems, nämlich* $x = (x_1, x_2)^T = (1,1)^T$.

Neben der Unterscheidung von Optimierungsproblemen in beschränkte und unbeschränkte Formulierungen, lassen sich noch weitere Kriterien zur Charakterisierung eines Optimierungsproblems heran ziehen:

- **Anzahl der unbekannten Parameter**, z.B. groß oder klein
- **Eigenschaften der Zielfunktion**, z.B. Linearität, Konvexität, Differenzierbarkeit
- **Charakteristik des Optimums**, z.B. Sattelpunkt, lokales oder globales Optimum
- **Modelleigenschaften**, z.B. stochastisch oder deterministisch
- **Wertebereich**, z.B. diskret oder kontinuierlich

## 7.1 Unrestringierte Optimierung

Im Folgenden wollen wir uns auf eine bestimmte Klasse von allgemeinen Optimierungsproblemen konzentrieren, den *unbeschränkten* oder *unrestringierten* Optimierungsproblemen.

**Definition 7.3** (Unrestringierte Optimierung)
*Liegt ein allgemeines Optimierungsproblem der Form* (7.1) *ohne Nebenbedingungen vor, d.h., für die Indexmengen gilt* $\mathcal{E} = \mathcal{I} = \varnothing$*, so sprechen wir von einem* unbeschränkten *oder* unrestringierten Optimierungsproblem.

**Bemerkung 7.4**
*Häufig lassen sich restringierte Optimierungsprobleme in unrestringierte Optimierungsprobleme überführen, indem man zusätzliche Strafterme zur Zielfunktion hinzufügt, die eine Verletzung der ursprünglichen Nebenbedingungen zwar mit Kosten belegt, diese jedoch grundsätzlich erlaubt. Hierbei spricht man auch von* relaxierten Optimierungsproblemen.

*Möchte man eine Zielfunktion* $F\colon \Omega \to \mathbb{R}$ *unter der Nebenbedingung* $c(x) = 0$ *für* $x \in \Omega$ *minimieren, so lässt sich beispielsweise folgendes relaxierte Optimierungsproblem gewinnen*

$$\min_{x\in\Omega} G_\lambda(x) \;:=\; \min_{x\in\Omega} F(x) + \lambda \cdot ||c(x)||^2.$$

*Hierbei ist* $\lambda > 0$ *ein fest gewählter Parameter, der es erlaubt den Einfluss der Nebenbedingung auf die Minimerung der Zielfunktion zu steuern.*

Zur Bestimmung von Optima müssen wir zunächst den Begriff eines stationären Punkts einer zu optimierenden Zielfunktion einführen.

**Definition 7.5** (Stationärer Punkt)
*Sei* $\Omega \subset \mathbb{R}^n$ *eine offene, zusammenhängende Teilmenge und sei* $F\colon \Omega \to \mathbb{R}$ *eine reellwertige Zielfunktion. Wir nennen einen Punkt* $x^* \in \Omega$ stationären Punkt *von* $F$*, falls er die Bedingung*

$$\nabla F(x^*) \;=\; 0$$

*erfüllt.*

Stationäre Punkte sind in der Optimierung interessante Kandidaten für Extremstellen. Für eine genaue Charakterisierung führen wir zunächst den Begriff eines lokalen bzw. globalen Minimums ein.

**Definition 7.6** (Lokales und globales Minimum)
*Sei* $\Omega \subset \mathbb{R}^n$ *eine offene, zusammenhängende Teilmenge und sei* $F\colon \Omega \to \mathbb{R}$ *eine reellwertige Zielfunktion. Wir nennen einen Punkt* $x^* \in \Omega$ *ein* lokales Minimum *der Funktion* $F$*, falls es eine lokale Umgebung* $U \subset \Omega$ *von* $x^* \in U$ *gibt, so dass für alle* $x \in U$ *gilt:*

$$F(x^*) \leqslant F(x), \qquad \text{für alle } x \in U. \tag{7.2}$$

*Wir nennen* $x^* \in \Omega$ *ein* globales Minimum *von* $F$*, falls die Ungleichung* (7.2) *für jede beliebige Umgebung* $U \subset \Omega$ *gilt und somit insbesondere für* $U = \Omega$*.*

**Bemerkung 7.7**
*In obiger Definition sprechen wir nur von Minima, jedoch ist klar, dass sich jedes Maximierungsproblem durch einen Vorzeichenwechsel leicht in ein Minimierungsproblem umschreiben lässt, d.h.,*

$$\max_{x\in\Omega} F(x) \;\Leftrightarrow\; \min_{x\in\Omega} -F(x) \;=:\; \min_{x_\in\Omega} G(x).$$

Da wir nun eine Bedingung für das Vorliegen eines lokalen Minimums haben, können wir mit folgenden Satz die notwendigen Bedingungen für solch ein lokales Minimum angeben.

**Satz 7.8** (Notwendige Optimalitätsbedingungen erster Ordnung)
*Sei $\Omega \subset \mathbb{R}^n$ eine offene, zusammenhängende Teilmenge und sei $F\colon \Omega \to \mathbb{R}$ eine reellwertige Zielfunktion. Sei $x^* \in \Omega$ ein lokales Minimum von $F$ in $\Omega$ und die Funktion $F$ sei stetig partiell differenzierbar in einer lokalen, offenen Umgebung $U \subset \Omega$ von $x^*$. Dann gilt*

$$\nabla F(x^*) \;=\; 0.$$

*Beweis:* Wir führen einen Beweis durch Widerspruch. Nehmen wir also an, dass $x^* \in \mathbb{R}$ ein lokales Minimum von $F$ sei, jedoch aber $\nabla F(x^*) \neq 0$ gelte. Wir wählen den Richtungsvektor $\vec{p} := -\nabla F(x^*) \neq 0$. Es ist somit klar, dass

$$\langle \vec{p}, \nabla F(x^*)\rangle \;=\; -\langle \nabla F(x^*), \nabla F(x^*)\rangle \;=\; -||\nabla F(x^*)||^2 \;<\; 0.$$

Da $\nabla F$ nach Voraussetzung stetig in einer lokalen Umgebung $U \subset \Omega$ von $x^*$ ist existiert ein $T > 0$, so dass auch gilt:

$$\langle \vec{p}, \nabla F(x^* + t\vec{p})\rangle < 0, \qquad \text{für alle } t \in [0, T].$$

Nach dem Satz 6.48 von Taylor gilt aber auch für jedes $t \in [0, T]$:

$$F(x^* + t\vec{p}) \;=\; F(x^*) + \underbrace{t\langle \vec{p}, \nabla F(x^* + \tilde{t}\vec{p})\rangle}_{<\,0}, \qquad \text{für ein } \tilde{t} \in (0, t).$$

Somit gilt also $F(x^* + t\vec{p}) < F(x^*)$ und wir haben offenbar eine Richtung $\vec{p} \in \mathbb{R}^n\backslash\{0\}$ gefunden in der die Funktionswerte von $F$ abnehmen. Also ist $x^* \in \Omega$ kein lokales Minimum von $F$. Das ist aber ein Widerspruch zur Annahme und somit ist die Behauptung bewiesen. □

Die Aussage des Satzes 7.8 lässt sich wie folgt zusammenfassen.

**Korollar 7.9**
*Jedes lokale Minimum $x^* \in \Omega$ einer Zielfunktion $F\colon \Omega \to \mathbb{R}$ ist ein stationärer Punkt.*

Die Umkehrung der Aussage in Satz 7.8 gilt im Allgemeinen nicht, wie uns das folgende Beispiel zeigt.

**Beispiel 7.10**
*Wir betrachten die Funktion*

$$F(x) := -x^3.$$

*Diese besitzt einen stationären Punkt in $x^* = 0$, d.h., es gilt $\nabla F(0) = 0$. Dennoch handelt es sich hierbei nicht um ein lokales Optimum, sondern lediglich um einen* Sattelpunkt*.*

Bei der Suche nach lokalen Minima einer Zielfunktion $F$ lässt sich ein weiteres Kriterium anwenden, welches die zweite Ableitung der Funktion verwendet.

**Satz 7.11** (Notwendige Optimalitätsbedingungen zweiter Ordnung)
*Sei $\Omega \subset \mathbb{R}^n$ ein offenes, zusammenhängendes Gebiet und sei $F\colon \Omega \to \mathbb{R}$ eine reellwertige Zielfunktion. Sei $x^* \in \Omega$ ein lokales Minimum von $F$ in $\Omega$ und die Hessematrix $H_F$ von $F$ sei stetig in einer offenen Umgebung $U \subset \Omega$ von $x^*$, d.h., die Funktion $F$ ist zweimal stetig partiell differenzierbar auf der Teilmenge $U$. Dann gilt $\nabla F(x^*) = 0$ und $H_F(x^*)$ ist positiv semidefinit, d.h., es gilt*

$$\langle \vec{p}, H_F(x^*)\vec{p} \rangle \;\geqslant\; 0 \quad \textit{für alle} \;\; \vec{p} \in \mathbb{R}^n.$$

*Beweis:* Der erste Teil der Behauptung folgt bereits aus Satz 7.8, so dass wir uns nur auf den Beweis für die zweite Behauptung konzentrieren müssen.

Wir führen wieder einen Beweis durch Widerspruch. Sei $x^* \in \Omega$ nach Voraussetzung ein lokaler Minimierer von $F$, das heißt nach Satz 7.8 gilt $\nabla F(x^*) = 0$. Wir nehmen an, dass $H_F F(x^*)$ nicht positiv semidefinit ist. Dann können wir einen Vektor $\vec{p} \in \mathbb{R}^n/\{0\}$ finden, so dass

$$\langle \vec{p}, H_F(x^*)\vec{p} \rangle \;<\; 0$$

gilt. Da $H_F$ nach Voraussetzung stetig ist in einer lokalen Umgebung $U \subset \Omega$ von $x^*$ existiert ein $T > 0$, so dass

$$\langle \vec{p}, H_F(x^* + t\vec{p})\vec{p} \rangle \;<\; 0, \qquad \text{für alle } t \in [0, T].$$

Nach dem Satz 6.48 von Taylor gilt jedoch für alle $t \in (0, T)$ die folgende Identität

$$F(x^* + t\vec{p}) \;=\; F(x^*) + \underbrace{t\langle \vec{p}, \nabla F(x^*)\rangle}_{=\,0} + \frac{1}{2}\,\underbrace{t^2 \langle \vec{p}, H_F(x^* + \tilde{t}\vec{p})\vec{p}\rangle}_{<\,0}, \qquad \text{für ein } \tilde{t} \in (0, t).$$

Durch Weglassen des Terms auf der rechten Seite der Gleichung folgt also, dass $F(x^* + \hat{t}\vec{p}) < F(x^*)$ gilt. Wir haben also eine Richtung $\vec{p} \in \mathbb{R}^n/\{0\}$ gefunden entlang der die Funktionswerte von $F$ abnehmen. Damit folgt, dass $x^*$ kein lokales Minimum von $F$ ist, was aber im Widerspruch zur Annahme steht. Das beweist die Aussage des Satzes. □

Schlussendlich wollen wir auch eine hinreichende Bedingung für das Vorliegen eines lokalen Minimums angeben.

**Satz 7.12** (Hinreichende Optimalitätsbedingungen zweiter Ordnung)
*Sei $\Omega \subset \mathbb{R}^n$ eine offene, zusammenhängende Teilmenge und sei $F\colon \Omega \to \mathbb{R}$ eine reellwertige Zielfunktion. Sei $x^* \in \Omega$ ein Punkt für den gelte*

*(i)* $\nabla F(x^*) \;=\; 0$,

*(ii)* $H_F(x^*)$ *ist positiv definit.*

*Außerdem sei die Hesse-Matrix $H_F$ von $F$ stetig in einer offenen Umgebung $U \subset \Omega$ von $x^* \in U$. Dann ist $x^* \in \Omega$ ein striktes lokales Minimum von $F$.*

*Beweis.* Da die Hesse-Matrix $H_F$ von $F$ stetig und positiv definit in $x^* \in \Omega$ ist nach Voraussetzung können wir einen Radius $r > 0$ finden, so dass $H_F(x)$ positiv definit ist für alle $x \in B_r(x^*)$. Für jeden Vektor $\vec{p} \in \mathbb{R}^n/\{0\}$ mit $||\vec{p}|| < r$ gilt dann nach dem Satz 6.48 von Taylor:

$$F(x^* + \vec{p}) \;=\; F(x^*) + \underbrace{\langle \vec{p}, \nabla F(x^*) \rangle}_{=\,0} + \frac{1}{2}\langle \vec{p}, H_F(x^* + t\vec{p})\vec{p} \rangle, \qquad \text{für ein } t \in (0,1).$$

Da $||t\vec{p}|| < r$ ist nach Konstruktion wissen wir, dass

$$\langle \vec{p}, H_F(x^* + t\vec{p})\vec{p} \rangle \;>\; 0$$

gilt und somit schon $F(x^* + \vec{p}) > F(x^*)$ gelten muss. Da $\vec{p} \in \mathbb{R}^n/\{0\}$ mit $||\vec{p}|| < r$ beliebig gewählt war handelt es sich bei $x^* \in \Omega$ um ein striktes lokales Minimum der Funktion $F$. □

**Bemerkung 7.13**
*An Hand der Definitheit der Hesse-Matrix können wir also in vielen Fällen die Art des stationären Punkts charakterisieren. Es stellt sich heraus, dass folgende Beobachtungen gelten:*

- *Ist die Hesse-Matrix* positiv definit*, handelt es sich um ein* ***lokales Minimum****.*
- *Falls die Hesse-Matrix* negativ definit *ist, so ist der stationäre Punkt ein* ***lokales Maximum****.*
- *Ist die Hesse-Matrix* indefinit*, so handelt es sich um einen* ***Sattelpunkt****.*

*Sollte die Hesse-Matrix in einem stationären Punkt* semidefinit *sein, so lässt sich keine eindeutige Aussage treffen.*

Im Folgenden wollen wir die notwendigen und hinreichenden Optimalitätsbedingungen für verschiedene Zielfunktionen prüfen.

**Beispiel 7.14**
*Wir untersuchen zwei unterschiedliche Zielfunktionen auf mögliche Extremstellen.*

1. *Zuerst diskutieren wir eine einfache eindimensionale Zielfunktion* $F\colon \mathbb{R} \to \mathbb{R}$ *mit*

$$F(x) = x^4.$$

*Wir prüfen die notwendigen Optimalitätsbedingungen erster Ordnung aus Satz 7.8 und bemerken, dass der einzige stationäre Punkt von* $F$ *in* $x^* = 0$ *vorliegt, da*

$$F'(x) \;=\; 4x^3 \;=\; 0 \quad \Leftrightarrow \; x \;=\; 0.$$

*Da* $F''(x) = 12x^2$ *ist gilt im stationären Punkt* $x^* = 0$ *nur* $F''(0) = 0$*. Daher können wir nicht die hinreichenden Optimalitätsbedingungen zweiter Ordnung aus Satz 7.12 anwenden, da die zweite Ableitung nicht positiv ist. Dennoch erkennen wir, dass* $F$ *ein striktes lokales Minimum in* $x = 0$ *besitzt mit* $\nabla F(0) = 0$*. Dies zeigt uns, dass die in Satz 7.12 genannten Bedingungen nur hinreichend sind, jedoch nicht notwendig für das Vorliegen eines strikten lokalen Minimums.*

2. *Nun diskutieren wir eine zweidimensionale Zielfunktion* $F\colon \mathbb{R}^2 \to \mathbb{R}$ *mit*

$$F(x,y) \;=\; 6x^2 - 2x - 4xy + y^2 + 1$$

*Zur Bestimmung von möglichen Extremstellen der Zielfunktion müssen wir zunächst den Gradienten berechnen. Hierzu bestimmen wir die ersten partiellen Ableitungen von* $F$ *als*

$$\begin{aligned} \partial_x F(x,y) &= 12x - 2 - 4y, \\ \partial_y F(x,y) &= -4x + 2y \end{aligned}$$

*Zur Bestimmung eines stationären Punkts setzen wir*

$$\nabla F(x,y) \;=\; (\partial_x F(x,y), \partial_y F(x,y))^T \;=\; (12x - 2 - 4y, -4x + 2y)^T \;\stackrel{!}{=}\; (0,0)^T.$$

*Um mögliche Kandidaten für ein lokales Extremum zu finden müssen wir also folgendes lineares Gleichungssystem lösen:*

$$Ax \;=\; \begin{pmatrix} 12 & -4 \\ -4 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \;=\; \begin{pmatrix} 2 \\ 0 \end{pmatrix} \;=\; f.$$

*Durch Lösen des linearen Gleichungssystems erhalten wir den einzigen stationären Punkt in*

$$x^* \;=\; (0.5, 1)^T.$$

*Wir berechnen als Nächstes die Hesse-Matrix $H_F$ von $F$ mit den zweiten Ableitungen:*

$$\begin{aligned}
\partial_x^2 F(x,y) &= \partial_x(12x-2-4y) &&= 12,\\
\partial_y\partial_x F(x,y) &= \partial_y(12x-2-4y) &&= -4,\\
\partial_x\partial_y F(x,y) &= \partial_x(-4x+2y) &&= -4,\\
\partial_y^2 F(x,y) &= \partial_y(-4x+2y) &&= 2.
\end{aligned}$$

*Insgesamt ergibt sich also für die Hesse-Matrix im stationären Punkt $x^* \in \mathbb{R}^2$*

$$H_F(x^*) = \begin{pmatrix} 12 & -4 \\ -4 & 2 \end{pmatrix}.$$

*Zur Bestimmung der Definitheit von $H_F$ stellen wir das charakteristische Polynom $P_{H_F}$ auf als*

$$P_{H_F}(t) = (12-t) * (2-t) - 16 = t^2 - 14t + 8.$$

*Wir bestimmen die Eigenwerte der Hesse-Matrix $H_F$ als Nullstellen des charakteristischen Polynoms $P_{H_F}$ mittels p-q-Formel und erhalten somit*

$$\begin{aligned}
\lambda_{1/2} &= -\frac{p}{2} \pm \sqrt{\frac{p^2}{4} - q} = \frac{14}{2} \pm \sqrt{\frac{196}{4} - 8}\\
&= 7 \pm \sqrt{41}
\end{aligned}$$

*Da beide Eigenwerte $\lambda_{1/2} = 7 \pm \sqrt{41}$ von $H_F$ positiv sind, ist die Hesse-Matrix positiv definit. Damit sind die hinreichenden Bedingungen aus Satz 7.12 für ein striktes lokales Minimum von $F$ im Punkt $x^* = (0.5, 1)$ erfüllt. Da die Hesse-Matrix auf ganz $\mathbb{R}^2$ positiv definit ist, handelt es sich sogar um ein globales Minimum.*

Eine äußerst wertvolle Eigenschaft bei der Optimierung ist die Konvexität einer Zielfunktion, da jedes lokale Optimum einer konvexen Funktion bereits ein globales Optimum ist.

**Definition 7.15** (Konvexität)
*Sei $\Omega \subset \mathbb{R}^n$ eine offene, zusammenhängende Teilmenge und sei $F\colon \Omega \to \mathbb{R}$ eine reellwertige Zielfunktion. Wir nennen $F$* konvex *wenn für beliebige Vektoren $x, y \in \Omega$ die folgende Ungleichung für alle $0 \leqslant \alpha \leqslant 1$ gilt:*

$$F(\alpha x + (1-\alpha)y) \leqslant \alpha F(x) + (1-\alpha)F(y). \tag{7.3}$$

*Wir nennen die Funktion $F$* strikt konvex, *falls* (7.3) *eine echte Ungleichung ist.*

*Anschaulich bedeutet Konvexität einer Funktion $F$, dass jede Verbindungsgerade zwischen zwei Punkten $x, y \in \Omega$ oberhalb des Graphen der Funktion $F$ durch die Punkte $x$ und $y$ verläuft.*

Folgende Bemerkung stellt die Bedeutung von Konvexität für die Optimierung fest.

**Bemerkung 7.16**
*Sei $\Omega \subset \mathbb{R}^n$ eine offene, zusammenhängende Teilmenge und $F\colon \Omega \to \mathbb{R}$ eine Zielfunktion. Man kann zeigen, dass die Hesse-Matrix $H_F(x)$ genau dann positiv definit ist für alle $x \in \Omega$, wenn $F$ eine strikt konvexe Funktion ist.*

*Dies hat zur Konsequenz, dass wenn man einen stationären Punkt $x^* \in \Omega$ findet mit $\nabla F(x^*) = 0$, so ist dieser Punkt das eindeutige, globale Minimum der Zielfunktion.*

## 7.2 Optimierung unter Nebenbedingungen

Bisher haben wir uns nur mit unrestringierten Optimierungsproblemen beschäftigt und für diese Optimalitätsbedingungen hergeleitet. Viele reale Aufgabenstellungen erfordern jedoch die Optimierung einer Zielfunktion unter Nebenbedingungen. Daher wollen wir uns im Folgenden mit solchen Problemen beschäftigen und eine Möglichkeit aufzeigen stationäre Punkte unter Nebenbedingungen zu identifizieren.

Wir beschränken uns hierbei auf Optimierungsprobleme, bei denen die Nebenbedingung eine Gleichheit erfüllen müssen, lassen jedoch keine Ungleichungen zu, d.h., für die Indexmenge $\mathcal{I}$ der Ungleichungen im allgemeinen Optimierungsproblem (7.1) gilt $\mathcal{I} = \varnothing$. Der Grund für diese Einschränkung ist, dass die zugehörige Theorie der sogenannten *Karush-Kuhn-Tucker (KKT)* Bedingungen für Optimierungsprobleme mit Ungleichungsnebenbedingungen den Rahmen dieser Vorlesung sprengen würde. Interessierte Leser*innen seien auf [Nocedal2006, Kapitel 12.3] verwiesen.

Wir beginnen direkt mit der wichtigen geometrischen Beobachtung, dass die Gradienten einer Zielfunktion und der zugehörigen Nebenbedingung in einem Minimierer parallel ausgerichtet sein müssen.

**Satz 7.17** (Lagrange-Multiplikatoren)
*Seien $F, c\colon \Omega \to \mathbb{R}$ zwei stetig partiell differenzierbare Funktionen und sei*

$$M \coloneqq \{x \in \Omega\colon c(x) = 0\} \subset \Omega$$

*eine Untermannigfaltigkeit (z.B. eine Kurve in $\Omega$), die alle Nullstellen von $c$ enthält.*

*Zu jeder Lösung $x^* \in M$ des Minimierungsproblems*

$$\min_{x \in \Omega} F(x) \qquad \text{mit} \qquad c(x) \;=\; 0$$

*und $\nabla c(x^*) \neq 0$ existiert ein* Lagrange-Multiplikator $\lambda^* \in \mathbb{R}$, *so dass*

$$\nabla F(x^*) + \lambda^* \nabla c(x^*) \;=\; 0.$$

*Das heißt die beiden Gradienten $\nabla F$ und $\nabla c$ sind parallel in $x^* \in \Omega$.*

*Beweis.* Sei $x^* \in M$ eine Lösung des Minimierungsproblems mit Nebenbedingung. Wir sehen zunächst ein, dass der Gradient $\nabla c(x^*)$ senkrecht zu allen Tangentialvektoren der Untermannigfaltigkeit $M$ steht, da sich $c$ entlang aller Tangentialrichtungen von $M$ nicht ändert. Wir schreiben nun den Gradienten der Zielfunktion $\nabla F$ mittels orthogonaler Projektion (vgl. Kapitel 3.5) als eindeutige Summe zweier Vektoren $v_\perp \in \mathbb{R}^n$ und $v_\parallel \in \mathbb{R}^n$, die jeweils orthogonal und parallel zu $\nabla c$ sind mit

$$\nabla F(x^*) \;=\; v_\perp + v_\parallel.$$

Wir führen nun den Beweis über einen Widerspruch. Nehmen wir an, dass $\nabla F$ und $\nabla c$ nicht parallel sind, dann folgt, dass $v_\perp \neq 0$ ist. Betrachten wir nun die Richtungsableitung von $F(x^*)$ in Richtung des Vektors $-v_\perp \in \mathbb{R}^n$, dann gilt

$$\begin{aligned} D_{-v_\perp}F(x^*) \;&=\; \langle \nabla F(x^*), -v_\perp \rangle \;=\; -\langle v_\perp + v_\parallel, v_\perp \rangle \\ &=\; -\langle v_\perp, v_\perp \rangle - \underbrace{\langle v_\parallel, v_\perp \rangle}_{=0} \;=\; -\langle v_\perp, v_\perp \rangle \;<\; 0. \end{aligned}$$

Das bedeutet, dass wir den Funktionswert der Zielfunktion $F$ noch weiter verkleinern können, indem wir entlang der Untermannigfaltigkeit $M$ in Richtung $-v_\perp \in \mathbb{R}^n$ gehen. Dies ist jedoch ein Widerspruch zur Optimalität des Punkts $x^* \in M$. Also müssen $\nabla F$ und $\nabla c$ parallel sein und es gilt:

$$\nabla F(x^*) \;=\; -\lambda \cdot \nabla c(x^*).$$

Der Lagrange-Multiplikator $\lambda \in \mathbb{R}$ taucht in der Formel auf, da die Gradienten parallel aber nicht gleich lang sein müssen. □

Die folgenden Bemerkungen erklären wie aus Satz 7.17 das sogenannte *Verfahren der Lagrange-Multiplikatoren* gewonnen werden kann, welches beispielsweise in der Physik bei Anwendungen der klassischen Mechanik eine Schlüsselrolle spielt.

**Bemerkung 7.18**
*Wir wollen folgende Beobachtungen festhalten.*

1. *Die notwendige Bedingung, dass die Gradienten der Zielfunktion $F$ und der Nebenbedingung $c$ in einem Minimierer $x^* \in M$ parallel ausgerichtet sein müssen, d.h.,*

   $$\nabla F(x^*) + \lambda^* \nabla c(x^*) \;=\; 0$$

   *lässt sich für das Verfahren der Lagrange-Multiplikatoren ausnutzen. Hierzu definieren wir zunächst eine neue Funktion $\Lambda\colon \Omega \times \mathbb{R} \to \mathbb{R}$, genannt* Lagrange-Funktion*, die eine zusätzliche Variable $\lambda \in \mathbb{R}$ für die Nebenbedingung besitzt, als*

   $$\Lambda(x, \lambda) \;:=\; F(x) + \lambda \cdot c(x).$$

*Betrachtet man nämlich nun stationäre Punkte $(x^*, \lambda^*) \in \Omega \times \mathbb{R}$ von $\Lambda$, d.h.,*

$$\nabla_x \Lambda(x^*, \lambda^*) \;=\; \nabla F(x^*) + \lambda^* \nabla c(x^*) \;\stackrel{!}{=}\; 0, \qquad \nabla_\lambda \Lambda(x^*, \lambda*) \;=\; c(x^*) \;\stackrel{!}{=}\; 0,$$

*so erfüllen diese Punkte automatisch die notwendigen Kriterien für einen Minimierer des ursprünglichen Optimierungsproblems unter Nebenbedingungen.*

2. *Das oben beschriebene Verfahren der Lagrange-Multiplikatoren lässt sich auch auf Optimierungsprobleme mit mehreren Nebenbedingungen $c_i \colon \Omega \to \mathbb{R}$ für $1 \leqslant i \leqslant m$ verallgemeinern. Hierzu wird die Lagrange-Funktion als eine Linearkombination der Zielfunktion und den $m \in \mathbb{N}$ Nebenbedingungen geschrieben als*

$$\Lambda(x, \lambda) \;=\; F(x) + \sum_{i=1}^{m} \lambda_i c_i(x).$$

*Zur Bestimmung stationärer Punkte geht man nun analog zum Fall mit nur einer Nebenbedingung vor.*

**Beispiel 7.19**
*Wir wollen im Folgenden eine Zielfunktion $F \colon \mathbb{R}^2 \to \mathbb{R}$ mit*

$$F(x, y) \;:=\; x + y$$

*minimieren unter der Nebenbedingung, dass der Lösungsvektor $x^* = (x, y)^T \in \mathbb{R}^2$ normiert sein soll, d.h., $x^2 + y^2 = 1$. Dies führt also zu einem Optimierungsproblem mit Nebenbedingung der folgenden Gestalt:*

$$\min_{(x,y) \in \mathbb{R}^2} F(x, y) \;=\; x + y \qquad \textit{mit} \qquad c(x, y) \;=\; x^2 + y^2 - 1 \;=\; 0. \tag{7.4}$$

*Wir identifizieren zunächst Punkte in denen der Gradient von $c(x, y)$ verschwindet, d.h.,*

$$\nabla c(x, y) \;=\; (2x, 2y)^T \;\stackrel{!}{=}\; 0.$$

*Dies kann nur im Ursprung $(0, 0)^T$ passieren. Da dieser Punkt nicht die Nebenbedingung erfüllt, müssen wir ihn also bei den folgenden Berechnungen nicht explizit beachten.*

*Um das Verfahren der Lagrange-Multiplikatoren anwenden zu können, stellen wir zunächst die Lagrange-Funktion auf:*

$$\Lambda(x, y, \lambda) \;:=\; F(x, y) + \lambda \cdot c(x, y) \;=\; x + y + \lambda \cdot (x^2 + y^2 - 1).$$

*Nach dem Satz 7.17 müssen wir für potentielle Lösungen des Optimierungsproblems (7.4) nur Extremstellen der Lagrange-Funktion berechnen, d.h., wir betrachten stationäre Punkte von $\Lambda$ durch*

$$\nabla \Lambda(x, y, \lambda) \;=\; \begin{pmatrix} \partial_x F(x, y) + \lambda \partial_x c(x, y) \\ \partial_y F(x, y) + \lambda \partial_y c(x, y) \\ c(x, y) \end{pmatrix} \;=\; \begin{pmatrix} 1 + \lambda \cdot 2x \\ 1 + \lambda \cdot 2y \\ x^2 + y^2 - 1 \end{pmatrix} \;\stackrel{!}{=}\; \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \in \mathbb{R}^3.$$

*Obwohl es sich nicht um lineares Gleichungssystem handelt können wir dennoch eine Lösung für die drei Gleichungen herleiten. Wenn wir annehmen, dass $\lambda \neq 0$ gilt, so können wir die ersten beiden Gleichungen jeweils nach $x$ und $y$ umstellen und erhalten so, dass*

$$x \;=\; -\frac{1}{2\lambda} \;=\; y.$$

*Setzen wir diese Identitäten in die dritte Gleichung können wir für den Lagrange-Multiplikator folgende Bedingung herleiten*

$$x^2 + y^2 - 1 \;=\; \frac{1}{4\lambda^2} + \frac{1}{4\lambda^2} - 1 \;=\; \frac{1}{2\lambda^2} - 1 \;=\; 0 \quad \Leftrightarrow \quad 2\lambda^2 \;=\; 1.$$

*Wir erhalten für $\lambda$ also die beiden möglichen Lösungen*

$$\lambda_1 \;=\; \frac{1}{\sqrt{2}}, \quad \lambda_2 \;=\; -\frac{1}{\sqrt{2}}.$$

*Setzen wir diese beiden Lagrange-Multiplikatoren wieder in die beiden ersten Optimalitätsbedingungen ein erhalten wir als stationäre Punkte der Lagrange-Funktion:*

$$x_1^* \;=\; (\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})^T, \quad x_2^* \;=\; (-\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2})^T.$$

*Durch Einsetzen der Punkte in die Zielfunktion $F$ sehen wir, dass es sich bei $x_1^*$ um ein Maximum des Optimierungsproblems handelt und bei $x_2^*$ um ein Minimum, da*

$$F(x_1^*) \;=\; \frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2} \;=\; \sqrt{2}, \qquad F(x_2^*) \;=\; -\frac{\sqrt{2}}{2} - \frac{\sqrt{2}}{2} \;=\; -\sqrt{2}.$$

# Kapitel 8

# Gewöhnliche Differentialgleichungen

Differentialgleichungen spielen seit je her eine fundamentale Rolle bei der physikalischen Modellierung von Naturgesetzen und erklären viele beobachtbare Phänomene unseres Alltags äußerst präzise. Eine Differentialgleichung beschreibt häufig das Änderungsverhalten von Größen, wie zum Beispiel die Fallgeschwindigkeit eines Steins, den man fallen lässt, oder die Vermehrung von Bakterien in einer Nährlösung. Während insbesondere in der Analysis die Existenz und Eindeutigkeit von Lösungen von Differentialgleichungen untersucht wird, widmet sich die Numerische Mathematik der Herleitung von Lösungsverfahren, die approximative Lösungen für Differentialgleichungen liefern, für die keine analytische Lösung bekannt ist.

Mathematisch gesehen ist eine Differentialgleichung eine Gleichung, in der eine unbekannte Funktion und ihre Ableitungen auftreten. Die größte vorkommende Ableitung bestimmt hierbei die *Ordnung der Differentialgleichung*. Je nachdem ob die es sich um eine Funktion in einer oder mehreren Veränderlichen handelt sprechen wir von einer *Gewöhnlichen Differentialgleichung* oder einer *Partiellen Differentialgleichung*. Werden gleich mehrere solche Funktionen durch mehrere Gleichungen beschrieben, so spricht man von einem *Differentialgleichungssystem*. Im Folgenden beschränken wir uns auf die Diskussion von gewöhnlichen Differentialgleichungen, d.h., wir untersuchen Gleichungen in denen eine unbekannte Funktion mit einer Variablen und deren Ableitungen vorkommen.

Das folgende Beispiel soll ein grundlegendes Verständnis von gewöhnlichen Differentialgleichungen und den zugehörigen mathematischen Fragestellungen liefern.

**Beispiel 8.1**
*Das einfachste Beispiel einer gewöhnlichen Differentialgleichung ist gegeben durch*

$$x(t) \;=\; x'(t), \qquad \text{für alle} \;\; t \in \mathbb{R}.$$

*Wir suchen also eine unbekannte Funktion $x\colon \mathbb{R} \to \mathbb{R}$, deren Ableitung gerade die Funktion*

*selbst ist. In diesem Fall können wir die Lösung der Gleichung quasi erraten, denn für die Exponentialfunktion gilt offensichtlich*

$$x(t) \;:=\; e^t \;=\; (e^t)' \;=\; x'(t).$$

*Jedoch ist dies nicht die einzige Lösung, denn die konstante Nullfunktion $x(t) \equiv 0$ für alle $t \in \mathbb{R}$ erfüllt ebenfalls die Differentialgleichung.*

In diesem Kapitel wollen wir uns damit beschäftigen, wann Lösungen einer Differentialgleichungen existieren und in welchen Fällen diese eindeutig sind. Darüber hinaus lernen wir praktische Werkzeuge zur analytischen Lösung von Differentialgleichungen kennen.

Wir beginnen mit der Definition einer grundlegenden Klasse von Differentialgleichungen.

**Definition 8.2** (Lineare Differentialgleichung)
*Sei $n \in \mathbb{N}$ und $I \subset \mathbb{R}$ ein offenes Intervall. Seien außerdem $a_i \colon I \to \mathbb{R}$ und $b \colon I \to \mathbb{R}$ stetige Funktionen für $0 \leqslant i \leqslant n$. Dann nennen wir eine gewöhnliche Differentialgleichung $n$-ter Ordnung mit einer unbekannten $n$-mal stetig differenzierbaren Funktion $x \colon I \to \mathbb{R}$* linear*, wenn sie in folgender Form vorliegt*

$$\sum_{i=0}^{n} a_i(t) x^{(i)}(t) \;=\; a_0(t) \cdot x(t) + a_1(t) \cdot x'(t) + \ldots + a_n(t) \cdot x^{(n)}(t) \;=\; b(t). \tag{8.1}$$

*Hierbei bezeichnen wir mit $x^{(k)}(t)$ die $k$-te Ableitung der Funktion $x$ im Punkt $t \in I$.*

*Ist eine Differentialgleichung nicht in der Form* (8.1)*, so bezeichnen wir sie als* nichtlinear*.*

*Eine lineare Differentialgleichung heißt* homogen*, falls $b(t) \equiv 0$ für alle $t \in \mathbb{R}$. Der Term $b$ wird auch häufig* Störfunktion *genannt.*

Im Folgenden wollen wir verschiedene gewöhnliche Differentialgleichungen diskutieren und gemäß der Definition 8.2 klassifizieren.

**Beispiel 8.3** (Radioaktiver Zerfall)
*Sei $c > 0$ eine Konstante. Dann beschreibt die folgende gewöhnliche Differentialgleichung*

$$\frac{d}{dt} x(t) \;=\; x'(t) \;=\; -c \cdot x(t)$$

*den radioaktiven Zerfall eines strahlenden Materials mit Stoffmenge $x$ zum Zeitpunkt $t$. In diesem Kontext beschreibt $c$ die charakteristische Zerfallskonstante des Materials. Es handelt sich hierbei um eine* ***lineare, gewöhnliche Differentialgleichung erster Ordnung****, die* ***homogen*** *ist (da die Störfunktion $b(t) \equiv 0$ ist).*

*Ist die anfängliche Stoffmenge $x_0 \in \mathbb{R}^+$ zum Zeitpunkt $t = 0$ mit $x(0) = x_0$ bekannt, so ist*

$$x(t) \;=\; x_0 \cdot e^{-ct}, \qquad \text{für} \;\; t \in \mathbb{R}_0^+$$

*die eindeutige Lösung der Differentialgleichung. Wir erhalten im allgemeinen Fall also eine einparametrige Schar von Lösungen, die linear vom Anfangswert $x_0$ abhängen.*

**Beispiel 8.4** (Steinwurf)
*In diesem Beispiel wollen wir die Bahn eines geworfenen Steins im konstanten Gravitationsfeld der Erde mit Erdbeschleunigung $g \approx 9,81 \frac{m}{s^2}$ berechnen. Wir vernachlässigen die Krafteinflüsse der Luftreibung und betrachten das folgende* ***inhomogene System von gewöhnlichen Differentialgleichungen zweiter Ordnung****:*

$$\frac{d^2}{dt^2} x_1(t) \;=\; 0, \qquad \frac{d^2}{dt^2} x_2(t) \;=\; -g.$$

*Wir bezeichnen hierbei die Horizontalkomponente der Position des Steins entlang der Wurfbahn zum Zeitpunkt $t \in \mathbb{R}_0^+$ mit $x_1 \colon \mathbb{R}_0^+ \to \mathbb{R}$ und die Vertikalkomponente mit $x_2 \colon \mathbb{R}_0^+ \to \mathbb{R}$. Die anfängliche Position zum Zeitpunkt $t_0 \in \mathbb{R}_0^+$ und die Anfangsgeschwindigkeit seien gegeben als*

$$x_0 \;=\; \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}, \qquad v_0 \;=\; \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}.$$

*Unter diesen Voraussetzungen können wir das obige Differentialgleichungssystem lösen und erhalten für die Wurfbahn des Steins, d.h., die Position des Steins in Abhängigkeit der Zeit $t \in \mathbb{R}_0^+$, folgende Lösung*

$$x_1(t) \;=\; z_1 + w_1 \cdot t, \qquad x_2(t) \;=\; z_2 + w_2 \cdot t - \frac{1}{2} \cdot g \cdot t^2.$$

*Hieraus lassen sich nun auch die Richtungskomponenten der Geschwindigkeit des Steins in Abhängigkeit der Zeit $t \in \mathbb{R}_0^+$ ermitteln, denn es gilt*

$$v_1(t) \;:=\; \frac{d}{dt} x_1(t) \;=\; x_1'(t) \;=\; w_1, \qquad v_2(t) \;:=\; \frac{d}{dt} x_2(t) \;=\; x_2'(t) \;=\; w_2 - g \cdot t.$$

**Beispiel 8.5** (Beschleunigung eines Raumschiffs durch Gravitation)
*Im letzten Beispiel wollen wir die Wirkung der Gravitation eines Himmelskörpers, wie zum Beispiel der Erde, auf ein wesentlich kleineres Objekt (wie ein Raumschiff) beschreiben. Sei also $r \colon \mathbb{R}_0^+ \to \mathbb{R}_0^+$ der Abstand des Raumschiffs zur Erde und $M \approx 5,972 \cdot 10^{24}$ kg die ungefähre Erdmasse.*

*Dann lässt sich die Beschleunigung, die das Raumschiff auf Grund der Gravitation der Erde erfährt, durch folgende* ***gewöhnliche Differentialgleichung zweiter Ordnung****, welche* ***homogen und nichtlinear*** *ist, beschreiben:*

$$\frac{d^2}{dt^2} r(t) \;=\; -\frac{M}{r^2(t)}.$$

*Das bedeutet, dass die Beschleunigung des Raumschiffs durch die Gravitation quadratisch mit dem Abstand zum Erdmittelpunkt zunimmt.*

## 8.1 Trennung der Variablen

Nachdem wir im letzten Abschnitt einige motivierende Beispiele für gewöhnliche Differentialgleichungen aus der Physik diskutiert haben, wollen wir im Folgenden eine erste Technik zur Bestimmung von Lösungen für bestimmte Spezialfälle herleiten. Die Methode der Trennung der Variablen, auch Separationsmethode genannt, erlaubt es separierbare Differentialgleichungen erster Ordnung zu lösen.

Hierzu führen wir zunächst den Begriff von separierbaren Differentialgleichungen ein.

**Definition 8.6** (Separierbare Differentialgleichung)
*Seien $I, J \subset \mathbb{R}$ zwei offene Intervalle und es seien $f\colon I \to \mathbb{R}$ und $g\colon J \to \mathbb{R}$ zwei stetige Funktionen. Wir setzen voraus, dass $g(y) \neq 0$ für alle $y \in J$. Eine gewöhnliche Differentialgleichung erster Ordnung der Form*

$$y'(x) \;=\; f(x) \cdot g(y(x)) \;=\; f(x) \cdot g(y)$$

*heißt* separierbare Differentialgleichung *und wird häufig auch* Differentialgleichung mit getrennten Variablen *genannt.*

Der folgende Satz sagt aus unter welchen Voraussetzungen eindeutige Lösungen für separierbare Differentialgleichungen existieren und welche Gestalt diese besitzen.

**Satz 8.7** (Trennung der Variablen)
*Seien $I, J \subset \mathbb{R}$ zwei offene Intervalle und $(x_0, y_0) \in I \times J$ ein beliebiger Punkt. Es seien nun $f\colon I \to \mathbb{R}$ und $g\colon J \to \mathbb{R}$ zwei stetige Funktionen, wobei $g(y) \neq 0$ für alle $y \in J$ gelte. Wir betrachten die separierbare Differentialgleichung*

$$y'(x) \;=\; f(x) \cdot g(y).$$

*Wir definieren außerdem Funktionen $F\colon I \to \mathbb{R}$ und $G\colon J \to \mathbb{R}$ durch*

$$F(x) \;:=\; \int_{x_0}^{x} f(t)\,\mathrm{d}t, \qquad G(y) \;:=\; \int_{y_0}^{y} \frac{1}{g(s)}\,\mathrm{d}s.$$

*Sei nun $I' \subset I$ ein Intervall mit $x_0 \in I'$ und $F(I') \subset G(J)$. Dann existiert eine eindeutige Lösung $\varphi\colon I' \to \mathbb{R}$ der separierbaren Differentialgleichung mit der Anfangsbedingung $\varphi(x_0) = y_0$. Die Lösung genügt außerdem der folgenden Beziehung*

$$G(\varphi(x)) \;=\; F(x) \qquad \text{für alle } \; x \in I'.$$

*Beweis.* Wir bemerken zunächst, dass wir für $g(y) \neq 0$ für alle $y \in J$ die separierbare Differentialgleichung umschreiben können in

$$\frac{y'(x)}{g(y)} \;=\; f(x).$$

Außerdem wissen wir, dass die stetige Funktion $g\colon J \to \mathbb{R}$ entweder positiv mit $g(y) > 0$ oder negativ mit $g(y) < 0$ für alle $y \in J$ sein muss. Denn ansonsten würde es nach dem Zwischenwertsatz [Burger2020, Satz 5.13] eine Stelle $y_0 \in J$ geben mit $g(y_0) = 0$, was nach Voraussetzung ausgeschlossen ist. Damit folgt aber schon, dass die Funktion $G\colon J \to \mathbb{R}$ mit

$$G(y) \;=\; \int_{y_0}^{y} \frac{1}{g(s)}\,\mathrm{d}s$$

streng monoton und somit insbesondere injektiv ist. Damit existiert also eine Umkehrabbildung $G^{-1}\colon G(J) \to J$. Diese Beobachtung können wir im Folgenden ausnutzen.

Zuerst wollen wir die Eindeutigkeit der Lösung zeigen. Sei also $\varphi\colon I' \to \mathbb{R}$ eine Lösung der separierbaren Differentialgleichung, dann erfüllt die Lösung offensichtlich

$$\frac{\varphi'(x)}{g(\varphi)} \;=\; f(x) \qquad \text{für alle } \; x \in I'.$$

Betrachten wir nun die Funktion $F\colon I \to \mathbb{R}$ so sehen wir unter Ausnutzung der Substitutionsregel der Integralrechnung aus Satz 5.10 ein, dass für alle $x \in I'$ gilt

$$F(x) \;=\; \int_{x_0}^{x} f(t)\,\mathrm{d}t \;=\; \int_{x_0}^{x} \frac{\varphi'(t)}{g(\varphi(t))}\,\mathrm{d}t \;=\; \int_{\varphi(x_0)}^{\varphi(x)} \frac{1}{g(s)}\,\mathrm{d}s \;=\; \int_{y_0}^{\varphi(x)} \frac{1}{g(s)}\,\mathrm{d}s \;=\; G(\varphi(x)).$$

Da $G$ eine Umkehrfunktion $G^{-1}$ besitzt können wir die Lösung der separierbaren Differentialgleichung also eindeutig charakterisieren durch

$$G^{-1}(F(x)) \;=\; \varphi(x) \qquad \text{für alle } x \in I'.$$

Nun müssen wir nur noch zeigen, dass die Funktion $\varphi := G^{-1} \circ F\colon I' \to \mathbb{R}$ eine Lösung der separierbaren Differentialgleichung darstellt. Durch Anwendung der Ketten- und Umkehrregel der Differentialrechnung und dem Hauptsatz der Differential- und Integralrechnung gilt für alle $x \in I'$:

$$\begin{aligned}\varphi'(x) \;&=\; (G^{-1} \circ F)'(x) \;=\; (G^{-1})'(F(x)) \cdot F'(x) \;=\; \frac{1}{G'(G^{-1}(F(x)))} \cdot F'(x) \\ &=\; \frac{1}{G'(\varphi(x))} \cdot F'(x) \;=\; g(\varphi(x)) \cdot f(x).\end{aligned}$$

Außerdem gilt $\varphi(x_0) = y_0$, wie man nachrechnen kann:

$$\varphi(x_0) \;=\; G^{-1}(F(x_0)) \;=\; G^{-1}\left(\int_{x_0}^{x_0} f(t)\,\mathrm{d}t\right) \;=\; G^{-1}(0) \;=\; y_0.$$

□

Das folgende Beispiel illustriert wie das Verfahren der Trennung der Variablen eingesetzt werden kann, um analytische Lösungen für separierbare Differentialgleichungen zu erhalten.

**Beispiel 8.8** (Trennung der Variablen)
*Betrachten wie die folgende gewöhnliche Differentialgleichung erster Ordnung*

$$y'(x) \;=\; -\frac{y(x)}{x},$$

*welche definiert ist für alle $x \in \mathbb{R}\backslash\{0\}$ und außerdem linear und homogen ist. Wir nehmen an, dass die Anfangsbedingung $y(1) = c$ für eine Konstante $c > 0$ gelte.*

*Es handelt sich hierbei um eine separierbare Differentialgleichung da wir sie mit getrennten Variablen wie folgt darstellen können:*

$$y'(x) \;=\; f(x) \cdot g(y) \qquad \textit{mit} \quad f(x) \coloneqq -\frac{1}{x} \quad \textit{und} \quad g(y) \coloneqq y\,.$$

*Nach Satz 8.7 berechnen wir nun die Funktionen $F$ und $G$ als*

$$F(x) \;=\; \int_1^x f(t)\,\mathrm{d}t \;=\; -\int_1^x \frac{1}{t}\,\mathrm{d}t \;=\; -\log\Big|_1^x \;=\; -\log x + \underbrace{\log 1}_{=0} \;=\; -\log x,$$

$$G(y) \;=\; \int_c^y \frac{1}{g(s)}\,\mathrm{d}s \;=\; \int_c^y \frac{1}{s}\,\mathrm{d}s \;=\; \log\Big|_c^y \;=\; \log(y) - \log(c) \;=\; \log\Big(\frac{y}{c}\Big)\,.$$

*Setzen wir den Definitionsbereich $J$ der Funktionen $g$ und $G$ auf $J = \mathbb{R}^+$, wissen wir, dass der Logarithmus $\mathbb{R}_+$ auf ganz $\mathbb{R}$ abbildet und daher können wir den Definitionsbereich $I$ der Funktionen $f$ und $F$ auf $I' = I = \mathbb{R}_+$ setzen. Offensichtlich gilt $\mathbb{R}^+ = F(I') \subset G(J) = \mathbb{R}$ und somit sind die Voraussetzungen von Satz 8.7 erfüllt.*

*Es existiert also eine auf $\mathbb{R}^+$ definierte eindeutige Lösung $\varphi\colon \mathbb{R}_+ \to \mathbb{R}$ mit*

$$\log\frac{\varphi(x)}{c} \;=\; G(\varphi(x)) \;=\; F(x) \;=\; -\log x \qquad \textit{für alle} \;\; x \in \mathbb{R}_+.$$

*Durch Anwendung der Exponentialfunktion auf beiden Seiten der Gleichung erhält man dann die explizite Lösung*

$$\varphi(x) \;=\; \frac{c}{x} \qquad \textit{für alle} \;\; x \in \mathbb{R}^+.$$

*Die Korrektheit der Lösung lässt sich leicht überprüfen, denn es gilt*

$$\varphi'(x) \;=\; \Big(\frac{c}{x}\Big)' \;=\; -\frac{c}{x^2} \;=\; -\frac{c}{x}\cdot\frac{1}{x} \;=\; -\frac{\varphi(x)}{x}.$$

Die Trennung der Variablen lässt sich von Beispiel 8.8 auf allgemeine lineare, homogene Differentialgleichungen erster Ordnung mit beliebiger Koeffizientenfunktion verallgemeinern, wie der folgende Satz zeigt.

**Satz 8.9** (Lineare und homogene gewöhnliche Differentialgleichung erster Ordnung)
*Sei $I \subset \mathbb{R}$ ein offenes Intervall. Betrachten wir eine lineare und homogene Differentialgleichung erster Ordnung, die im Allgemeinen von folgender Form ist:*

$$y'(x) \;=\; a(x) \cdot y(x), \qquad \text{für alle } x \in I.$$

*Für einen beliebigen Punkt $x_0 \in I$ und einen konstanten Anfangswert $c \in \mathbb{R}$ gibt es dann genau eine Lösung $\varphi\colon I \to \mathbb{R}$ der Differentialgleichung, die der Anfangsbedingung $\varphi(x_0) = c$ genügt. Diese ist für alle $x \in I$ von der Form*

$$\varphi(x) \;:=\; c \cdot \exp\left(\int_{x_0}^{x} a(t)\,\mathrm{d}t\right).$$

*Beweis.* Jede homogene lineare Differentialgleichung erster Ordnung ist ein Spezialfall einer separierbaren Differentialgleichung in Satz 8.7 für $f(x) := a(x)$ und $g(y) := y$.

Zuerst zeigen wir, dass $\varphi$ eine Lösung der Differentialgleichung ist. Man rechnet leicht nach, dass die Anfangsbedingung erfüllt ist, da

$$\varphi(x_0) \;=\; c \cdot \exp\left(\int_{x_0}^{x_0} a(t)\,\mathrm{d}t\right) \;=\; c \cdot \exp(0) \;=\; c.$$

Außerdem gilt für alle Punkte $x \in I$

$$\varphi'(x) \;=\; c \cdot \left(\exp\left(\int_{x_0}^{x} a(t)\,\mathrm{d}t\right)\right)' \;=\; a(x) \cdot c \cdot \exp\left(\int_{x_0}^{x} a(t)\,\mathrm{d}t\right) \;=\; a(x) \cdot \varphi(x).$$

Um die Eindeutigkeit der Lösung zu beweisen bemerken wir zunächst, dass

$$\varphi_0(x) \;:=\; \exp\left(-\int_{x_0}^{x} a(t)\,\mathrm{d}t\right)$$

die Differentialgleichung $\varphi_0'(x) = -a(x) \cdot \varphi_0(x)$ löst mit einem analogen Argument wie oben. Sei nun $\vartheta\colon I \to \mathbb{R}$ eine beliebige Lösung der Differentialgleichung $y'(x) = a(x) \cdot y(x)$ mit den Anfangsbedingungen $\vartheta(x_0) = c$. Wir bilden nun das Produkt $\vartheta_0(x) := \vartheta(x) \cdot \varphi_0(x)$ und differenzieren die Funktion $\vartheta_0$ mit der Produktregel:

$$\begin{aligned} \vartheta_0'(x) \;&=\; \vartheta'(x) \cdot \varphi_0(x) + \vartheta(x) \cdot \varphi_0'(x) \\ &=\; a(x) \cdot \vartheta(x) \cdot \varphi_0(x) - \vartheta(x) \cdot a(x) \cdot \varphi_0(x) \;=\; 0. \end{aligned}$$

Da dies für alle $x \in I$ gilt muss $\vartheta_0$ somit eine Konstante sein und es gilt

$$\vartheta_0(x) \;=\; \vartheta_0(x_0) \;=\; \vartheta(x_0) \cdot \varphi_0(x_0) \;=\; c \cdot \exp(0) \;=\; c.$$

Damit folgt nun schließlich aus $\vartheta_0(x) = \vartheta(x) \cdot \varphi_0(x)$ für alle $x \in I$:

$$\vartheta(x) \;=\; c \cdot \varphi_0(x)^{-1} \;=\; c \cdot \exp\left( \int_{x_0}^{x} a(t)\,\mathrm{d}t \right).$$

Hierbei können wir die Umkehrfunktion $\varphi_0^{-1}$ betrachten, da die Exponentialfunktion streng monoton und somit insbesondere injektiv ist. □

## 8.2 Variation der Konstanten

Bisher haben wir uns auf den Fall von linearen, homogenen Differentialgleichungen beschränkt. Das heißt wir haben nur den Spezialfall ohne Störfunktion, d.h., $b(x) \equiv 0$ diskutiert. Im Folgenden wollen wir nun inhomogene lineare Differentialgleichungen der Forn

$$y'(x) = a(x) \cdot y(x) + b(x),$$

betrachten, wobei $a, b\colon I \to \mathbb{R}$ stetige Funktionen auf dem offenen Intervall $I \subset \mathbb{R}$ sind. Für einen beliebigen Punkt $x_0 \in I$ und einen Anfangswert $c \in \mathbb{R}$ suchen wir eine Lösung $\varphi\colon I \to \mathbb{R}$ der Differentialgleichung die der Anfangsbedingung $\varphi(x_0) = c$ genügt.

Da eine Trennung der Variablen aus Kapitel 8.1 in diesem Fall nicht zum Ziel führt, wollen wir eine andere wichtige Technik zur analytischen Lösung von gewöhnlichen Differentialgleichungen untersuchen. Die sogenannte *Variation der Konstanten* ist ein Verfahren zur Bestimmung einer speziellen Lösung einer inhomogenen linearen Differentialgleichung erster Ordnung. Die einzige Zutat, die diese Technik voraussetzt, ist das Vorliegen einer Lösung der zugehörigen homogenen Differentialgleichung

$$y'(x) = a(x) \cdot y(x).$$

Aus Satz 8.9 wissen wir bereits, dass diese von der folgenden Form sind

$$\varphi_0(x) \;:=\; c \cdot \exp\left( \int_{x_0}^{x} a(t)\,\mathrm{d}t \right).$$

Der folgende Satz beschreibt analytische Lösungen der inhomogenen Differentialgleichung basierend auf der Lösung $\varphi_0$ der zugehörigen homogenen Differentialgleichung.

**Satz 8.10** (Variation der Konstanten)
*Sei $I \subset \mathbb{R}$ ein offenes Intervall und seien $a, b\colon I \to \mathbb{R}$ zwei stetige Funktionen. Dann gibt es zu einem beliebigen Punkt $x_0 \in I$ und einem Anfangswert $c \in \mathbb{R}$ eine eindeutige Lösung $\varphi\colon I \to \mathbb{R}$ der linearen, inhomogenen Differentialgleichung erster Ordnung*

$$y'(x) \;=\; a(x) \cdot y(x) + b(x) \quad \text{für alle} \quad x \in I$$

*unter der Anfangsbedingung $\varphi(x_0) = c$. Die Lösung hat die Form*

$$\varphi(x) \;=\; \varphi_0(x) \cdot \left( c + \int_{x_0}^{x} \varphi_0^{-1}(t) \cdot b(t)\,\mathrm{d}t \right),$$

*wobei*

$$\varphi_0(x) \;\coloneqq\; \exp\left( \int_{x_0}^{x} a(t)\,\mathrm{d}t \right).$$

*Beweis.* Sei $\varphi_0$ eine Lösung der zugehörigen homogenen Differentialgleichung und $\varphi$ die zu bestimmende Lösung der inhomogenen Differentialgleichung. Wir betrachten nun eine sogenannte Ansatzfunktion $u$ mit

$$u(x) \;\coloneqq\; \varphi(x) \cdot \varphi_0^{-1}(x) \;=\; \varphi(x) \cdot \exp\left( -\int_{x_0}^{x} a(t)\,\mathrm{d}t \right).$$

Durch Multiplikation mit $\varphi_0$ lässt sich der Ausdruck dieser Funktion nun umstellen zur unbekannten Lösung $\varphi$ mit

$$\varphi(x) \;=\; \varphi_0(x) \cdot u(x) \;=\; \exp\left( \int_{x_0}^{x} a(t)\,\mathrm{d}t \right) \cdot u(x). \tag{8.2}$$

Setzen wir diesen Ausdruck nun in die inhomogene Differentialgleichung ein, so erhalten wir für alle $x \in I$

$$(u \cdot \varphi_0)'(x) \;=\; a(x) \cdot u(x) \cdot \varphi_0(x) + b(x). \tag{8.3}$$

Durch Anwendung der Produkt- und Kettenregel der Differentiation und des Hauptsatzes der Differential- und Integralrechnung lässt sich die linke Seite der Gleichung noch vereinfachen zu

$$\begin{aligned}
(u \cdot \varphi_0)'(x) \;&=\; u'(x) \cdot \varphi_0(x) + u(x) \cdot \varphi_0'(x) \\
&=\; u'(x) \cdot \varphi_0(x) + u(x) \cdot \varphi_0(x) \cdot \left( \int_{x_0}^{x} a(t)\,\mathrm{d}t \right)' \\
&=\; u'(x) \cdot \varphi_0(x) + u(x) \cdot \varphi_0(x) \cdot a(x).
\end{aligned}$$

Wir erkennen, dass sich die Terme auf der linken und rechten Seite von (8.3) teilweise annullieren wodurch wir die folgende einfache Darstellung der Funktion $b$ erhalten:

$$\begin{aligned} & u'(x) \cdot \varphi_0(x) + u(x) \cdot \varphi_0(x) \cdot a(x) \;=\; a(x) \cdot u(x) \cdot \varphi_0(x) + b(x) \\ \Leftrightarrow \quad & u'(x) \cdot \varphi_0(x) \;=\; b(x). \end{aligned}$$

Multiplizieren wir diese Gleichung wiederum mit $\varphi_0^{-1}$, so erhalten wir die folgende Darstellung der Ableitung von $u$ mit

$$u'(x) \;=\; \varphi_0^{-1}(x) \cdot b(x) \qquad \text{für alle } x \in I.$$

Somit können wir die unbekannte Funktion $u$ charakterisieren als Stammfunktion durch

$$u(x) \;=\; \int_{x_0}^{x} \varphi_0^{-1}(t) \cdot b(t)\,\mathrm{d}t + C,$$

wobei $C \in \mathbb{R}$ eine Integrationskonstante ist. Setzen wir dies nun in die Gleichung (8.2) für die Lösung $\varphi$ der inhomogenen Differentialgleichung ein, so erhalten wir für alle $x \in I$:

$$\varphi(x) \;=\; \varphi_0(x) \cdot u(x) \;=\; \varphi_0(x) \cdot \left( \int_{x_0}^{x} \varphi_0^{-1}(t) \cdot b(t)\,\mathrm{d}t + C \right).$$

Zur Bestimmung der Integrationskonstante $C$ setzen wir $x = x_0$ und sehen durch Anwendung der Anfangswertbedingung

$$\varphi(x_0) \;=\; \underbrace{\varphi_0(x_0)}_{=1} \cdot \Big( \underbrace{\int_{x_0}^{x_0} \varphi_0^{-1}(t) \cdot b(t)\,\mathrm{d}t}_{=0} + C \Big) \;=\; C \;\stackrel{!}{=}\; c.$$

Damit ist nun die Aussage des Satzes vollständig bewiesen. □

Im folgenden Beispiel wollen wir das Verfahren der Variation der Konstanten nutzen, um die analytische Lösung einer inhomogenen Differentialgleichung herzuleiten.

**Beispiel 8.11** (Variation der Konstanten)
*Wir betrachten die folgende lineare, inhomogene Differentialgleichung erster Ordnung*

$$y'(x) \;=\; 2x \cdot y(x) + x^3, \tag{8.4}$$

*für die die Anfangsbedingung $y(0) = c$ mit $c \in \mathbb{R}$ erfüllt sein soll. Die zugehörige homogene Differentialgleichung $y'(x) = 2x \cdot y(x)$ besitzt nach Satz 8.9 Lösungen der Form*

$$\varphi_0(x) \;=\; \exp\left( \int_0^x 2t\,\mathrm{d}t \right) \;=\; e^{x^2}.$$

*Wir erhalten eine Lösung $\varphi\colon \mathbb{R} \to \mathbb{R}$ der inhomogenen Gleichung* (8.4) *unter der Anfangsbedingung $\varphi(0) = c$ durch*

$$\varphi(x) \;=\; e^{x^2} \cdot \left( c + \int_0^x t^3 e^{-t^2} \,\mathrm{d}t \right).$$

*Wir können glücklicherweise dieses Integral weiter berechnen indem wir die Substitution $s := t^2$ anwenden und darüber hinaus partielle Integration durchführen. Damit erhält man*

$$\int_0^x t^3 e^{-t^2} \,\mathrm{d}t \;=\; \frac{1}{2} - \frac{1}{2}\left(x^2 + 1\right) e^{-x^2},$$

*und somit insgesamt die Lösung der inhomogenen Differentialgleichung* (8.4) *mit*

$$\begin{aligned} \varphi(x) \;&=\; e^{x^2} \cdot \left( c + \int_0^x t^3 e^{-t^2} \,\mathrm{d}t \right) \;=\; e^{x^2} \cdot \left( c + \frac{1}{2} - \frac{1}{2}\left(x^2 + 1\right) e^{-x^2} \right) \\ &=\; \left( c + \frac{1}{2} \right) e^{x^2} - \frac{1}{2}\left(x^2 + 1\right). \end{aligned}$$

## 8.3 Differentialgleichungen höherer Ordnung

Obwohl wir gewöhnliche Differentialgleichungen höherer Ordnung bereits im vorigen Abschnitt erwähnt haben, haben wir diese noch nicht näher diskutiert. Wir wollen im Folgenden ein zusätzliches Kriterium zur Klassifikation von gewöhnlichen Differentialgleichungen einführen und beschreiben, wie sich eine gewöhnliche Differentialgleichung höherer Ordnung in ein System von gewöhnlichen Differentialgleichungen erster Ordnung umformulieren lässt.

Beginnen wir zunächst mit der Definition einer gewöhnlichen Differentialgleichung $n$-ter Ordnung. Eine Differentialgleichung $n$-ter Ordnung schreibt eine Bedingung für die $n$-te Ableitung der gesuchten Lösung $\varphi$ vor. Diese $n$-te Ableitung hängt sowohl von $x$ als auch von der Lösung $\varphi$ und ihren Ableitungen bis zur $(n-1)$-ten Ordnung im Punkt $x$ ab.

**Definition 8.12** (Differentialgleichung $n$-ter Ordnung)
*Sei $G \subset \mathbb{R} \times \mathbb{R}^n$ eine offene Teilmenge und*

$$f\colon G \to \mathbb{R}$$

*eine stetige Funktion.*

1. *Wir nennen eine Gleichung der Form*

$$y^{(n)}(x) \;=\; f(x, y(x), y'(x), \ldots, y^{(n-1)}(x))$$

*eine* explizite *gewöhnliche Differentialgleichung n-ter Ordnung. Ist die gewöhnliche Differentialgleichung nicht nach der höchsten vorkommenden Ableitung aufgelöst, d.h. wenn sie von der folgenden Form ist*

$$f(x, y(x), y'(x), \ldots, y^{(n-1)}(x), y^{(n)}(x)) \;=\; 0,$$

*so nennen wir sie eine* implizite *gewöhnliche Differentialgleichung n-ter Ordnung.*

2. *Wir nennen eine n-mal differenzierbare Funktion $\varphi\colon I \to \mathbb{R}$ auf dem offenen Intervall $I \subset \mathbb{R}$ eine* Lösung der Differentialgleichung, *wenn sie die folgenden Eigenschaften besitzt:*

   *i) Der von der unbekannten Lösung $\varphi$ abhängende Menge*

$$\Gamma_\varphi \;:=\; \Big\{(x, y_1, y_2, \ldots, y_n) \in I \times \mathbb{R}^n \;\Big|\; y_k = \varphi^{(k-1)}(x),\; 1 \leqslant k \leqslant n\Big\}$$

   *ist eine Teilmenge von G,*

   *ii) Für alle Punkte $x \in I$ gilt*

$$\varphi^{(n)}(x) \;=\; f(x, \varphi(x), \varphi'(x), \ldots, \varphi^{(n-1)}(x))\,.$$

Das folgende Beispiel illustriert den Begriff einer gewöhnlichen Differentialgleichung in expliziter und impliziter Form.

**Beispiel 8.13**
*Sei $I \subset \mathbb{R}$ ein offenes Intervall. Die Gleichung*

$$y'(x) \cdot y(x) + x \;=\; 0$$

*ist eine nichtlineare, implizite gewöhnliche Differentialgleichung erster Ordnung. Diese Differentialgleichung lässt sich in die folgende explizite Form umschreiben*

$$y'(x) \;=\; -\frac{x}{y(x)}$$

*wenn wir annehmen, dass $y(x) \neq 0$ für alle Punkte $x \in I$ gilt. Aus der Kreisgleichung $x^2 + y^2 = c > 0$ lässt sich eine explizite Lösung dieser gewöhnlichen Differentialgleichung angeben mit*

$$\varphi(x) \;=\; \pm\sqrt{c - x^2}, \qquad \text{für } |x| < \sqrt{c}.$$

Man kann eine gewöhnliche Differentialgleichung $n$-ter Ordnung in ein System von gewöhnlichen Differentialgleichungen erster Ordnung überführen wie folgende Bemerkung zeigt. Diese Beobachtung hat den großen Vorteil, dass man sich bei Lösungsmethoden von gewöhnlichen Differentialgleichungen, egal ob analytisch oder numerisch, nur auf das Lösen von Differentialgleichungssystemen erster Ordnung konzentrieren muss.

**Bemerkung 8.14**
*Sei $G \subset \mathbb{R} \times \mathbb{R}^n$ eine offene Teilmenge und $f\colon G \to \mathbb{R}$ eine stetige Funktion. Wir betrachten im Folgenden die explizite gewöhnliche Differentialgleichung $n$-ter Ordnung*

$$y^{(n)}(x) \;=\; f(x, y(x), y'(x), \ldots, y^{(n-1)}(x)). \tag{8.5}$$

1. *Wir können die Gleichung (8.5) zu einem Differentialgleichungssystem erster Ordnung umformulieren bestehend aus $n+1$ Gleichungen mit*

$$\begin{cases} y &= y_1, \\ y_1' &= y_2, \\ &\vdots \\ y_{n-1}' &= y_n, \\ y_n' &= f(x, y_1, \ldots, y_n). \end{cases} \tag{8.6}$$

   *Definieren wir nun zwei $n$-dimensionale Vektoren $Y, F \in \mathbb{R}^n$ mit*

$$Y(x) \;:=\; \begin{pmatrix} y_1(x) \\ \vdots \\ y_{n-1}(x) \\ y_n(x) \end{pmatrix} \qquad \textit{und} \qquad F(x,Y) \;:=\; \begin{pmatrix} y_2(x) \\ \vdots \\ y_n(x) \\ f(x, Y(x)) \end{pmatrix}$$

   *dann können wir das Differentialgleichungssystem (8.6) umschreiben zu*

$$Y'(x) \;=\; F(x, Y(x)).$$

   *Die Ableitung $Y'$ ist hierbei komponentenweise zu interpretieren.*

2. *Sei nun $\varphi\colon I \to \mathbb{R}$ eine Lösung der expliziten gewöhnlichen Differentialgleichung $n$-ter Ordnung in (8.5), d.h.,*

$$\varphi^{(n)}(x) \;=\; f(x, \varphi(x), \varphi'(x), \ldots, \varphi^{(n-1)}(x)).$$

   *Basierend auf $\varphi$ können wir uns nun eine vektorwertige Funktion $\Phi\colon I \to \mathbb{R}^n$ definieren mit*

$$\Phi(x) \;:=\; \begin{pmatrix} \varphi_1(x) \\ \varphi_2(x) \\ \vdots \\ \varphi_n(x) \end{pmatrix} = \begin{pmatrix} \varphi(x) \\ \varphi'(x) \\ \vdots \\ \varphi^{(n-1)}(x) \end{pmatrix}.$$

*Dann wird aus der obigen Definition klar, dass $\Phi$ eine Lösung des Differentialgleichungssystems* (8.6) *ist.*

*Ist nun umgekehrt vorausgesetzt, dass $\Phi\colon I \to \mathbb{R}^n$ eine Lösung des Differentialgleichungssystems erster Ordnung in* (8.6) *ist, dann ist*

$$\varphi \coloneqq \varphi_1\colon I \to \mathbb{R}$$

*eine Lösung der Differentialgleichung $n$-ter Ordnung* (8.5)*. Dieser Zusammenhang gilt, da aus den ersten $n$ Gleichungen des Systems* (8.6) *folgt*

$$\begin{aligned}
\varphi_1(x) &= \varphi(x)\\
\varphi_2(x) &= \varphi_1'(x) = \varphi'(x),\\
\varphi_3(x) &= \varphi_2'(x) = \varphi_1''(x) = \varphi''(x),\\
&\vdots\\
\varphi_n(x) &= \varphi_{n-1}'(x) = \ldots = \varphi^{(n-1)}(x).
\end{aligned}$$

*Da $\varphi_n$ einmal differenzierbar ist, folgt daraus, dass $\varphi$ $n$-mal differenzierbar sein muss. Die $(n+1)$-te Gleichung von* (8.6) *liefert dann*

$$\varphi_n'(x) = \varphi^{(n)}(x) = f(x, \varphi(x), \varphi'(x), \ldots, \varphi^{(n-1)}(x)).$$

*Wir sehen also, dass die Lösungen von* (8.5) *und* (8.6) *in einer eindeutigen Beziehung zueinander stehen.*

**Beispiel 8.15**
*Die Gleichung*

$$y''(x) + y(x) = 0 \tag{8.7}$$

*für $x \in \mathbb{R}$ ist eine implizite gewöhnliche Differentialgleichung zweiter Ordnung. Um diese Differentialgleichung in ein Differentialgleichungssystem erster Ordnung zu transformieren führen wir zwei neue Variablen wie folgt ein:*

$$y_1 \coloneqq y, \qquad y_2 \coloneqq y_1'.$$

*Mit diesen Variablen können wir die Differentialgleichung* (8.7) *umschreiben in die Gleichung*

$$y_2' + y_1 = 0.$$

*Zusammen mit der Bedingung $y_1' = y_2$ ergibt sich damit das folgende lineare, homogene Differentialgleichungssystem erster Ordnung:*

$$Y' \coloneqq \begin{pmatrix} y_1' \\ y_2' \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}.$$

*Die gewöhnliche Differentialgleichung* (8.7) *besitzt offensichtlich die auf ganz* $\mathbb{R}$ *definierten Lösungen*

$$\varphi(x) \;:=\; \cos x, \qquad \textit{und} \qquad \varphi(x) \;:=\; \sin x.$$

*Aufgrund der Linearität der Ableitung sind in diesem Fall auch sämtliche Linearkombinationen von Lösungen wiederum Lösungen. Das heißt wir können eine allgemeine Lösung der gewöhnlichen Differentialgleichung angeben für beliebige Konstanten* $c_0, c_1 \in \mathbb{R}$*:*

$$\varphi_{c_0,c_1}(x) \;:=\; c_0 \cdot \cos x + c_1 \cdot \sin x.$$

*Betrachten wir die beiden Anfangswertgleichungen*

$$\begin{aligned} \varphi_{c_0,c_1}(0) \;&=\; c_0 \cdot \cos(0) + c_1 \cdot \sin(0) \;=\; c_0 \\ \varphi'_{c_0,c_1}(0) \;&=\; -c_0 \cdot \sin(0) + c_1 \cdot \cos(0) \;=\; c_1\,, \end{aligned}$$

*so stellen wir fest, dass* $\varphi_{c_0,c_1} : \mathbb{R} \to \mathbb{R}$ *die eindeutig bestimmte Lösung* $\varphi$ *der Differentialgleichung* (8.7) *mit den Anfangswertbedingungen* $\varphi_{c_0,c_1}(0) = c_0$ *und* $\varphi'_{c_0,c_1}(0) = c_1$ *ist.*

## 8.4 Existenz und Eindeutigkeit von Lösungen

Bisher haben wir uns vornehmlich um analytische Lösungsverfahren für lineare gewöhnliche Differentialgleichungen erster Ordnung beschäftigt. Dabei haben wir theoretische Überlegungen zur Existenz und Eindeutigkeit von Lösungen allgemeiner gewöhnlicher Differentialgleichungen vernachlässigt. Daher wollen wir uns in diesem Abschnitt den notwendigen und hinreichenden Bedingungen widmen, die uns sagen wann eine Differentialgleichung lösbar ist und wann ihre Lösung sogar eindeutig ist.

Da wir in Kapitel 8.3 gesehen haben, dass sich beliebige gewöhnliche Differentialgleichungen höherer Ordnung in ein Differentialgleichungssystem erster Ordnung transformieren lassen, beschränken wir uns auf diese Differentialgleichungssysteme. Dies motiviert die folgende Definition.

**Definition 8.16** (Differentialgleichungssystem erster Ordnung)
*Sei* $G \subset \mathbb{R} \times \mathbb{R}^n$ *eine offene Teilmenge und*

$$\begin{aligned} f\colon G \;&\to\; \mathbb{R}^n, \\ (x, y) \;&\mapsto\; f(x, y) \end{aligned}$$

*eine vektorwertige stetige Funktion.*

*Dann nennt man die folgende Gleichung*

$$y'(x) \;=\; f(x, y(x))$$

*ein* System von $n$ Differentialgleichungen erster Ordnung. *Eine Lösung dieser Gleichung ist eine auf dem offenen Intervall* $I \subset \mathbb{R}$ *total differenzierbare Funktion* $\varphi\colon I \to \mathbb{R}^n$*, für die die folgenden Eigenschaften gelten:*

1. *Der Graph $\Gamma_\varphi$ von $\varphi$ liegt in der Teilmenge $G$, d.h,*

$$\Gamma_\varphi \;:=\; \{(x,y) \in I \times \mathbb{R}^n \mid y = \varphi(x)\} \subset G.$$

2. *Die Funktion $\varphi$ erfüllt die Gleichung*

$$\varphi'(x) \;=\; f(x,\varphi(x)) \qquad \textit{für alle } x \in I.$$

**Bemerkung 8.17**
*Im Gegensatz zu vorigen Abschnitten handelt es sich in Definition 8.16 nicht um eine Differentialgleichung einer skalarwertigen Funktion sondern um ein Differentialgleichungssystem einer vektorwertigen Funktion. Wir können die vektorwertigen Funktionen $y$ und $f$ also komponentenweise notieren als*

$$y(x) \;=\; \begin{pmatrix} y_1(x) \\ \vdots \\ y_n(x) \end{pmatrix} \qquad \textit{und} \qquad f(x,y) \;=\; \begin{pmatrix} f_1(x,y) \\ \vdots \\ f_n(x,y) \end{pmatrix},$$

*was zu folgendem Differentialgleichungssystem erster Ordnung führt:*

$$\begin{cases} y_1'(x) \;=\; f_1(x, y_1(x), \ldots, y_n(x)), \\ y_2'(x) \;=\; f_2(x, y_1(x), \ldots, y_n(x)), \\ \qquad\qquad \ldots \\ y_n'(x) \;=\; f_n(x, y_1(x), \ldots, y_n(x)). \end{cases}$$

Der folgende Satz charakterisiert die Lösung eines Differentialgleichungssystems mit Anfangswertbedingungen als Lösung einer zugehörigen Integralgleichung.

**Satz 8.18** (Integralgleichung)
*Sei $G \subset \mathbb{R} \times \mathbb{R}^n$ eine offene Teilmenge, $I \subset R$ ein offenes Intervall und sei $f\colon G \to \mathbb{R}^n$ eine stetige Abbildung. Weiter seien die Anfangwerte $(x_0, c) \in G$ mit $x_0 \in I$ gegeben. Sei nun $\varphi\colon I \to \mathbb{R}^n$ eine total differenzierbare Funktion, deren Graph in $G$ enthalten ist*

*Die Funktion $\varphi$ ist genau dann eine Lösung der Differentialgleichung*

$$y'(x) \;=\; f(x, y(x))$$

*unter der Anfangswertbedingung $\varphi(x_0) = c$, wenn sie die folgende Integralgleichung erfüllt:*

$$\varphi(x) \;=\; c + \int_{x_0}^{x} f(t, \varphi(t))\,\mathrm{d}t \qquad \textit{für alle } \; x \in I. \tag{8.8}$$

*Beweis.* Wir zeigen zunächst die Rückrichtung des Satzes. Sei also die Integralgleichung (8.8) erfüllt. Für die Anfangswertbedingung setzen wir $x = x_0$ und sehen somit, dass $\varphi(x_0) = c$ gilt. Da die Funktion $f$ nach Voraussetzung stetig ist, folgt aus dem Hauptsatz der Integral- und Differentialrechnung [Burger2020, Kapitel 7.2], dass

$$\frac{\mathrm{d}}{\mathrm{d}x} \int_{x_0}^{x} f(t, \varphi(t)) \, \mathrm{d}t \;=\; f(x, \varphi(x)).$$

Damit folgt aus der Definition der Funktion $\varphi$ in (8.8), dass $\varphi$ total differenzierbar ist und die gewöhnliche Differentialgleichung $\varphi'(x) = f(x, \varphi(x))$ erfüllt.

Für die Hinrichtung des Beweises nehmen wir an, dass die Funktion $\varphi$ eine Lösung der gewöhnlichen Differentialgleichung $\varphi'(x) = f(x, \varphi(x))$ ist, total differenzierbar sei und die Anfangswertbedingung $\varphi(x_0) = c$ erfülle. Dann können wir das folgende Integral betrachten und erhalten aus dem Hauptsatz der Integral- und Differentialrechnung:

$$\int_{x_0}^{x} f(t, \varphi(t)) \, \mathrm{d}t \;=\; \int_{x_0}^{x} \varphi'(t) \, \mathrm{d}t \;=\; \varphi(x) - \varphi(x_0) \;=\; \varphi(x) - c.$$

Durch Umstellen von $c$ auf die linke Seite erhält man die Identität der Integralgleichung in (8.8). □

Das folgende Lemma liefert uns ein hinreichendes Kriterium für die Lipschitz-Stetigkeit einer Funktion. Diese Eigenschaft wird für die Existenz und Eindeutigkeit von Lösungen gewöhnlicher Differentialgleichungen entscheidend sein.

**Lemma 8.19**
*Sei $G \subset \mathbb{R} \times \mathbb{R}^n$ eine offene Teilmenge und $f \colon G \to \mathbb{R}^n$ eine bezüglich der Variablen $y = (y_1, \ldots, y_n)$ stetig partiell differenzierbare Funktion.*

*Dann ist die Funktion $f$ lokal Lipschitz-stetig in $G$ bezüglich des $y$-Variable.*

*Beweis.* Sei $(a, b) \in G$ ein beliebiger Punkt. Dann existiert ein $r > 0$, so dass die kompakte Menge

$$B_r(a, b) \;:=\; \{(x, y) \in \mathbb{R} \times \mathbb{R}^n : \; |x - a| \leqslant r, \; ||y - b|| \leqslant r\}$$

ganz in $G$ liegt. Da nach Voraussetzung alle Komponenten der Jacobi-Matrix $J := (\frac{\partial f_i}{\partial y_j})_{1 \leqslant i,j \leqslant n}$ stetige Funktionen sind, können wir eine Konstante $L \in R_0^+$ finden mit

$$L \;:=\; \sup_{(x,y) \in B_r(a,b)} \|J(x, y)\| \;<\; +\infty.$$

Aus einer Folgerung des Mittelwertsatzes in 6.41 folgt für alle $(x, y), (x, \tilde{y}) \in B_r(a, b)$, dass gilt

$$||f(x, y) - f(x, \tilde{y})|| \;\leqslant\; L \cdot ||y - \tilde{y}||.$$

□

Nun sind wir in der Lage die hinreichenden Bedingungen für die Eindeutigkeit von Lösungen gewöhnlicher Differentialgleichungen zu formulieren.

**Satz 8.20** (Eindeutigkeitssatz)
*Sei $G \subset \mathbb{R} \times \mathbb{R}^n$ eine offene Teilmenge und sei $f\colon G \to \mathbb{R}^n$ eine stetige Funktion, die lokal Lipschitz-stetig in $G$ bezüglich der $y$-Variablen ist. Seien nun $\varphi, \vartheta\colon I \to \mathbb{R}^n$ zwei Lösungen der Differentialgleichung*

$$y'(x) \;=\; f(x, y(x)) \qquad \textit{für alle } x \in I.$$

*Stimmen die beiden Lösungen der Differentialgleichung in einem Punkt $x_0 \in I$ überein, d.h., es gilt $\varphi(x_0) = \vartheta(x_0)$, so folgt schon*

$$\varphi(x) \;=\; \vartheta(x) \qquad \textit{für alle } x \in I.$$

*Beweis.* Wir beginnen zunächst damit die Eindeutigkeit von Lösungen in einer lokalen $\epsilon$-Umgebung zu zeigen. Sei $x_0 \in I$ ein Punkt, so dass für die beiden Lösungen $\varphi, \vartheta\colon I \to \mathbb{R}^n$ der gewöhnlichen Differentialgleichung gilt $\varphi(x_0) = \vartheta(x_0)$. Auf Grund der Integraldarstellung in Satz 8.18 folgt nun

$$\varphi(x) - \vartheta(x) \;=\; \int_{x_0}^{x} f(t, \varphi(t)) - f(t, \vartheta(t))\,\mathrm{d}t. \tag{8.9}$$

Da $f$ nach Voraussetzung lokal Lipschitz-stetig bezüglich der $y$-Variablen ist, existieren Konstanten $L \geqslant 0$ und $\delta > 0$, so dass für alle $t \in I \cap B_\delta(x_0) \coloneqq \{t \in I\colon |t - x_0| < \delta\}$ gilt:

$$||f(t, \varphi(t)) - f(t, \vartheta(t))|| \;\leqslant\; L \cdot ||\varphi(t) - \vartheta(t)||.$$

Wir definieren nun die zwei Variablen

$$\epsilon \;\coloneqq\; \min(\delta, \frac{1}{2L}) \quad \text{ und } \quad M \;\coloneqq\; \sup\{||\varphi(x) - \vartheta(x)|| :\; x \in I \cap B_\epsilon(x_0)\}.$$

Aus Gleichung (8.9) und (8.4) können wir nun für alle $x \in I \cap B_\epsilon(x_0)$ folgern

$$||\varphi(x) - \vartheta(x)|| \;\leqslant\; L \cdot \left| \int_{x_0}^{x} ||\varphi(t) - \vartheta(t)||\,\mathrm{d}t \right| \;\leqslant\; L \cdot |x - x_0| \cdot M \;\leqslant\; \frac{M}{2}.$$

Da diese Abschätzung auch für den Punkt $x \in I \cap B_\epsilon(x_0)$ gilt, für den das Supremum $M$ angenommen wird, folgt also $M \leqslant \frac{M}{2}$. Dies ist jedoch nur möglich, wenn $M = 0$ gilt, d.h., wenn die Funktionen $\varphi$ und $\vartheta$ in $I \cap B_\epsilon(x_0)$ übereinstimmen.

Basierend auf obigem Resultat zeigen wir nun, dass $\varphi(x) = \vartheta(x)$ für alle $x \in I$ mit $x \geqslant x_0$ gilt. Der Beweis für den Fall $x \leqslant x_0$ funktioniert analog und wird daher nicht näher diskutiert. Wir definieren zunächst den am weitest rechts liegenden Punkt $\xi \in I$, so dass die Funktionen $\varphi$ und $\vartheta$ noch übereinstimmen durch

$$x_1 \;\coloneqq\; \sup\{\xi \in I :\; \varphi|_{[x_0,\xi]} \equiv \vartheta|_{[x_0,\xi]}\}.$$

Falls $x_1 = \infty$ oder gleich dem rechten Rand des Intervalls $I$ gilt, so ist die Aussage bereits bewiesen. Nehmen wir also an, dass ein Punkt $x_1 \in I$ existiert bis zu dem die Funktionen $\varphi$ und $\vartheta$ übereinstimmen und darüber hinaus ein $\delta > 0$ existiert, so dass $[x_1, x_1 + \delta] \subset I$ gilt. Nach Voraussetzung wissen wir nun, dass $\varphi(x_1) = \vartheta(x_1)$ gilt und die Funktionen stetig sind. Nun wissen wir aus dem ersten Teil des Beweises, dass ein $\epsilon > 0$ existiert, so dass gilt

$$\varphi(x) \;=\; \vartheta(x) \qquad \text{für alle } x \in I \cap B_\epsilon(x_1).$$

Das widerspricht jedoch offensichtlich der Definition von $x_1 \in I$. Daher gilt also $\varphi(x) = \vartheta(x)$ für alle $x \in I$ mit $x \geqslant x_0$. □

Um zu verstehen, wann die Lösung einer gewöhnlichen Differentialgleichung nicht eindeutig ist, diskutieren wir im Folgenden ein Gegenbeispiel.

**Beispiel 8.21**
*Wir betrachten die folgende nichtlineare gewöhnliche Differentialgleichung*

$$y'(x) \;=\; (y(x))^{\frac{2}{3}}. \tag{8.10}$$

*Wie man sofort einsieht ist eine spezielle Lösung der Differentialgleichung gegeben durch*

$$\varphi_0(x) \;\equiv\; 0, \qquad \textit{für alle } x \in \mathbb{R}.$$

*Da die Differentialgleichung separierbar ist, lassen sich die Lösungen unter der Annahme $y \neq 0$ mit Hilfe von Satz 8.7 bestimmen.*

*Man sieht aber auch leicht, dass für beliebiges $a \in \mathbb{R}$ die Funktion $\vartheta_a \colon \mathbb{R} \to \mathbb{R}$ mit*

$$\vartheta_a(x) \;:=\; \frac{1}{27}(x-a)^3$$

*die gewöhnliche Differentialgleichung (8.10) erfüllt, denn es gilt*

$$\vartheta_a'(x) \;=\; \frac{1}{9}(x-a)^2 \;=\; \left(\frac{1}{27}(x-a)^3\right)^{\frac{2}{3}} \;=\; \vartheta_a(x)^{\frac{2}{3}}.$$

*Obwohl die Lösung $\varphi$ und $\vartheta_0$ im Punkt $x_0 = a$ übereinstimmen mit*

$$\varphi_0(a) \;=\; \vartheta_a(a) \;=\; 0,$$

*sieht man ein, dass der der Eindeutigkeitssatz 8.20 nicht gilt. Der Grund hierfür ist eine Verletzung der Voraussetzungen des Satzes, denn die Funktion $f(x,y) := y^{\frac{2}{3}}$ ist nicht für alle Punkte $y \in \mathbb{R}$ bezüglich der y-Variable Lipschitz-stetig. Zwar ist $f$ für $y \neq 0$ stetig partiell differenzierbar bezüglich der Variablen $y$ mit*

$$\frac{\partial f}{\partial y}(x,y) \;=\; \frac{2}{3}y^{-\frac{1}{3}}.$$

*Nach Lemma 8.19 ist $f$ somit lokal Lipschitz-stetig in ganz $\mathbb{R} \times \mathbb{R}\backslash\{0\}$ in Bezug auf die y-Variable. Jedoch ist $f$ in keiner Umgebung eines Punktes $(a,0)$ Lipschitz-stetig bezüglich de y-Variablen, da dieser Punkt eine Singularität mit unendlicher Steigung darstellt.*

Der folgende wichtige Satz formuliert die hinreichenden Bedingungen für die Existenz von Lösungen einer gewöhnlichen Differentialgleichung.

**Satz 8.22** (Existenzsatz von Picard-Lindelöf)
*Sei $G \subset \mathbb{R} \times \mathbb{R}^n$ eine offene Teilmenge und sei $f\colon G \to \mathbb{R}^n$ eine stetige Funktion, die lokal Lipschitz-stetig auf $G$ bezüglich der $y$-Variablen ist. Dann existiert zu jedem Anfangswert $(x_0, c) \in G$ ein $\varepsilon > 0$, sowie eine Lösung*

$$\varphi\colon\ [x_0 - \varepsilon, x_0 + \varepsilon] \to \mathbb{R}^n$$

*der gewöhnlichen Differentialgleichung*

$$y'(x) \ = \ f(x, y(x))$$

*unter der Anfangsbedingung $\varphi(x_0) = c$.*

*Beweis.* Siehe [Forster2017, §12, Satz 4]. □

Selbst wenn die Funktion $f$ der Differentialgleichung $y'(x) = f(x, y(x))$ auf ganz $\mathbb{R} \times \mathbb{R}^n$ definiert ist und überall lokal Lipschitz-stetig bezüglich der $y$-Variablen ist, so kann eine Lösung, die die Anfangsbedingung $\varphi(x_0) = c$ erfüllt unter Umständen nur in einer sehr kleinen Umgebung von $x_0 \in \mathbb{R}$ definiert sein. Dies wird durch das folgende Beispiel klar.

**Beispiel 8.23**
*Wir betrachten die folgende nichtlineare gewöhnliche Differentialgleichung erster Ordnung*

$$y'(x) = 2x \cdot (y(x))^2 \,.$$

*Die Funktion $f(x, y) = 2xy^2$ ist offensichtlich auf ganz $\mathbb{R} \times \mathbb{R}$ definiert und stetig partiell differenzierbar bezüglich der Variable $y$ und somit nach Lemma 8.19 lokal Lipschitz-stetig bezüglich der $y$-Variablen.*

*Wir suchen eine Lösung $\phi\colon \mathbb{R} \to \mathbb{R}$ der Differentialgleichung, die die Anfangswertbedingung $\varphi(0) = c$ erfüllt. Für den Fall $c = 0$ ist dies offensichtlich die konstante Funktion $\varphi(x) \equiv 0$ für alle $x \in \mathbb{R}$. Für $c \neq 0$ können wir die Lösung durch Trennung der Variablen*

*(siehe Kapitel 8.1) wie folgt unter der Annahme $y(x) \neq 0$ berechnen:*

$$\begin{aligned} y' &= \frac{\mathrm{d}y}{\mathrm{d}x} &&= 2xy^2 \\ \Rightarrow \quad \frac{1}{y^2}\,\mathrm{d}y &= 2x\,\mathrm{d}x\,, \\ \Rightarrow \quad \int_c^y \frac{1}{\eta^2}\,\mathrm{d}\eta &= \int_0^x 2\xi\,\mathrm{d}\xi, \\ \Rightarrow \quad -\frac{1}{y} + \frac{1}{c} &= x^2\,, \\ \Rightarrow \quad \varphi(x) \;:=\; y &= \frac{1}{\gamma - x^2}, \qquad \textit{mit} \;\; \gamma := \frac{1}{c}. \end{aligned}$$

*Dies ist die Lösung der gewöhnlichen Differentialgleichung unter der Anfangswertbedingung $\varphi(0) = c$, was man direkt durch Einsetzen verifizieren kann.*

*Falls $c > 0$ gilt, so ist der maximale Definitionsbereich dieser Lösung gegeben durch*

$$I_c \;:=\; \left\{ x \in \mathbb{R} \,:\, |x| < \sqrt{\gamma} = c^{-\frac{1}{2}} \right\}.$$

*Nähert sich $x$ von innen dem Rand des Intervalls $I_c$, so strebt der Funktionswert der Lösung gegen $+\infty$. Für $c < 0$ ist die Lösung hingegen auf ganz $\mathbb{R}$ definiert.*

# Literaturverzeichnis

[Burger2020] Martin Burger: *Skript zur Vorlesung "Mathematik für Data Science 1 / Physikstudierende A"*. Wintersemester 2020/21, FAU Erlangen-Nürnberg, 2020.

[Fischer2005] Gerd Fischer: *Lineare Algebra*. vieweg studium, 15. Auflage, 2005.

[Forster2017] Otto Forster: *Analysis 2*. Springer, 11. Auflage, 2017.

[Knauf2020] Andreas Knauf: *Skript zur Vorlesung "Mathematik für Physikstudierende 2"*. Sommersemester 2020, FAU Erlangen-Nürnberg, 2020.

[Nocedal2006] Jorge Nocedal und Stephen J. Wright: *Numerical Optimization*. Springer Verlag, 2. Auflage, 2006.