# Kapitel 40

# Orthogonale Matrizen

**Bemerkung 40.1 Motivation.** Im euklidischen Raum $\mathbb{R}^n$ haben wir gesehen, dass Orthonormalbasen zu besonders einfachen und schönen Beschreibungen führen. Nun soll das Konzept der Orthonormalität auf Matrizen erweitert werden. Dies führt auf die wichtige Klasse der orthogonalen Matrizen, die eine Reihe von schönen Eigenschaften aufweisen. Mit ihnen lassen sich unter anderem Drehungen und Spiegelungen beschreiben. □

**Definition 40.2 Orthogonale Matrix, $\mathbb{O}(n)$.** Besitzt eine Matrix $Q \in \mathbb{R}^{n\times n}$ orthonormale Spaltenvektoren $\mathbf{q}_{*1}, \ldots, \mathbf{q}_{*n}$, so wird sie orthogonale Matrix genannt (orthonormale Matrix wäre präziser, ist aber unüblich). Die Menge aller Orthogonalmatrizen des $\mathbb{R}^{n\times n}$ wird mit

$$\mathbb{O}(n) := \{Q \in \mathbb{R}^{n\times n} \mid Q \text{ ist orthogonal}\}$$

bezeichnet. □

**Satz 40.3 Eigenschaften orthogonaler Matrizen.** *Ist $Q \in \mathbb{O}(n)$, so gelten*

*i) $Q$ ist invertierbar und es gilt $Q^{-1} = Q^T$.*

*ii) Multiplikation mit $Q$ erhält das euklidische Produkt zweier Vektoren*

$$(Q\mathbf{u}) \cdot (Q\mathbf{v}) = \mathbf{u} \cdot \mathbf{v} \quad \forall\ \mathbf{u}, \mathbf{v} \in \mathbb{R}^n.$$

*iii) Multiplikation mit $Q$ erhält die euklidische Norm*

$$\|Q\mathbf{v}\|_2 = \|\mathbf{v}\|_2 \quad \forall\ bv \in \mathbb{R}^n.$$

*Man nennt $Q$ daher auch Isometrie.*

**Beweis:** *i):* Sei $A = (a_{ij}) = Q^TQ$. Dann gilt

$$a_{ij} = \sum_{k=1}^{n} q_{ki}q_{kj} = \mathbf{q}_{*i} \cdot \mathbf{q}_{*j} = \delta_{ij}.$$

Also ist $Q^TQ = I$. Analog zeigt man $QQ^T = I$. Somit ist $Q$ invertierbar mit $Q^T = Q^{-1}$.

*ii):* Die Aussage folgt aus

$$(Q\mathbf{u}) \cdot (Q\mathbf{v}) = (Q\mathbf{u})^T(Q\mathbf{v}) = \mathbf{u}^T \underbrace{Q^TQ}_{I}\mathbf{v} = \mathbf{u}^T\mathbf{v} = \mathbf{u} \cdot \mathbf{v}.$$

*iii):* Folgt aus ii) mit $\mathbf{u} = \mathbf{v}$. ■

**Bemerkung 40.4** Es gilt sogar dass eine Matrix $Q$ genau dann orthogonal ist, falls $Q^T = Q^{-1}$ gilt. □

**Beispiel 40.5**

1. *Rotation.* Orthogonale Matrizen können beispielsweise Rotationen beschreiben. Solche Matrizen besitzen die Gestalt

$$Q = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix},$$

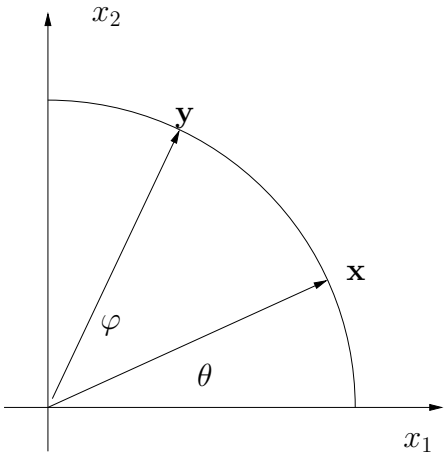
wobei $\varphi$ der Drehwinkel im mathematisch positiven Drehsinn (entgegen der Uhrzeigerrichtung) ist. Sei ein Vektor $\mathbf{x} \in \mathbb{R}^2$ gegeben

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} r\cos\theta \\ r\sin\theta \end{pmatrix}$$

mit $r = \sqrt{x_1^2 + x^2}$ und $\theta$ dem Winkel von $\mathbf{x}$ mit der positiven $x$–Achse. Mit Additionstheoremen für Winkelfunktionen, siehe zum Beispiel (20.2), erhält man

$$\mathbf{y} = Q\mathbf{x} = r \begin{pmatrix} \cos\varphi\cos\theta - \sin\varphi\sin\theta \\ \sin\varphi\cos\theta + \cos\varphi\sin\theta \end{pmatrix} = r \begin{pmatrix} \cos(\varphi+\theta) \\ \sin(\varphi+\theta) \end{pmatrix}.$$

Der Vektor $\mathbf{x}$ wurde also um den Winkel $\varphi$ gedreht.




Es ist offensichtlich, dass $Q$ orthogonal ist, da die beiden Spaltenvektoren orthogonal sind. Demzufolge gilt

$$Q^{-1} = Q^T = \begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix}.$$

Diese Matrix beschreibt eine Drehung um den Winkel $-\theta$.
Es gilt $\det Q = \cos^2\varphi + \sin^2\varphi = 1$.

2. *Spiegelung.* Orthogonale Matrizen können auch Spiegelungen an Geraden beschreiben. Zum Beispiel beschreibt die Matrix

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

die Spiegelung an der Gerade $y = x$. Diese Spiegelung vertauscht die $x_1$– und $x_2$–Komponente eines Vektors

$$Q\mathbf{x} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_2 \\ x_1 \end{pmatrix}.$$

Es gilt $\det Q = -1$.

□

**Satz 40.6 Determinante orthogonaler Matrizen.** *Ist $Q \in \mathbb{O}(n)$, so gilt $|\det Q| = 1$.*

**Beweis:** Aus $QQ^T$ und den Rechenregeln für Determinanten folgt

$$1 = \det(I) = \det\left(QQ^T\right) = \det Q \det Q^T = (\det Q\,)^2.$$

■

**Definition 40.7** $S\mathbb{O}(n)$**.** Man bezeichnet die Menge der orthogonale Matrizen $Q \in \mathbb{O}(n)$ mit $\det Q = 1$ mit

$$S\mathbb{O}(n) := \mathbb{O}(n)^+ := \{Q \in \mathbb{O}(n) \;\mid\; \det Q = 1\}.$$

□

**Satz 40.8 Gruppeneigenschaft von $\mathbb{O}(n)$ und $S\mathbb{O}(n)$.** *Die Mengen $\mathbb{O}(n)$ und $S\mathbb{O}(n)$ sind Untergruppen der allgemeinen linearen Gruppe $GL(n, \mathbb{R})$.*

**Beweis:** Übungsaufgabe. ■

**Bemerkung 40.9 Basiswechsel zwischen Orthonormalbasen.** Sei die Menge $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ eine Orthonormalbasis des euklidischen Raums $\mathbb{R}^n$. Dann existieren zu jedem Vektor $\mathbf{u} \in \mathbb{R}^n$ eindeutig bestimmte Koeffizienten $a_1, \ldots, a_n \in \mathbb{R}$ mit

$$\mathbf{u} = \sum_{k=1}^{n} a_k \mathbf{v}_k, \quad \text{das heißt} \quad \mathbf{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$$

ist der Koordinatenvektor von $\mathbf{u}$ bezüglich der Orthonormalbasis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$.

Sei nun $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$ eine weitere Orthonormalbases des $\mathbb{R}^n$ und $\mathbf{u}$ habe den Koordinatenvektor

$$\mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

bezüglich $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$.

Gesucht ist jetzt eine lineare Abbildung $Q$, die den einen Koordinatenvektor in den andern überführt, also etwa $\mathbf{b} = Q\mathbf{a}$.

Insbesondere lassen sich die Vektoren der einen Basis durch die andere Basis ausdrücken, etwa

$$\mathbf{v}_k = \sum_{i=1}^{n} (\mathbf{v}_k^T \mathbf{w}_i)\mathbf{w}_i,$$

siehe Satz 38.8. Es folgt

$$\begin{aligned} \mathbf{u} &= \sum_{k=1}^{n} a_k \mathbf{v}_k = \sum_{k=1}^{n} a_k \left( \sum_{i=1}^{n} (\mathbf{v}_k^T \mathbf{w}_i)\mathbf{w}_i \right) = \sum_{i=1}^{n} \left( \sum_{k=1}^{n} a_k (\mathbf{v}_k^T \mathbf{w}_i) \right) \mathbf{w}_i \\ &= \sum_{i=1}^{n} b_i \mathbf{w}_i \quad \text{mit} \quad b_i = \sum_{k=1}^{n} \underbrace{(\mathbf{v}_k^T \mathbf{w}_i)}_{=:q_{ik}} a_k. \end{aligned}$$

Für die gesuchte Matrix $Q = (q_{ik})$ gilt also

$$q_{ik} = \mathbf{v}_k^T \mathbf{w}_i = \mathbf{w}_i^T \mathbf{v}_k \quad \Longrightarrow \quad Q = \begin{pmatrix} \mathbf{w}_1^T \\ \vdots \\ \mathbf{w}_n^T \end{pmatrix} (\mathbf{v}_1, \ldots, \mathbf{v}_n) \,.$$

Wegen der Gruppeneigenschaft von $\mathbb{O}(n)$ ist $Q$ als Produkt zweier orthogonaler Matrizen wieder orthogonal. □

# Kapitel 41

# Eigenwerte und Eigenvektoren

**Bemerkung 41.1 Motivation.** Seien $\mathbf{v} \in \mathbb{R}^n$ und $A \in \mathbb{R}^{n \times n}$. Dann sind $\mathbf{v}$ und $A\mathbf{v}$ normalerweise nicht parallel. Man kann nun die Frage stellen, ob es ausgezeichnete Richtungen $\mathbf{v}$ gibt, so dass $A\mathbf{v}$ ein skalares Vielfaches von $\mathbf{v}$ ist, also $A\mathbf{v} = \lambda\mathbf{v}$.

Es wird sich herausstellen, dass diese ausgezeichneten Richtungen und vor allem die ausgezeichneten Skalare $\lambda$ sehr wichtige Informationen über die Eigenschaften der Matrix $A$ liefern. □

**Definition 41.2 Eigenvektor, Eigenwert.** Sei $A \in \mathbb{R}^{n \times n}$. Ein von $\mathbf{0}$ verschiedener Vektor $\mathbf{v} \in \mathbb{R}^n$ heißt Eigenvektor von $A$, wenn es ein $\lambda \in \mathbb{C}$ gibt mit

$$A\mathbf{v} = \lambda\mathbf{v}.$$

Der Skalar $\lambda$ heißt dann Eigenwert von $A$. □

**Bemerkung 41.3 Bedeutung von Eigenvektoren und Eigenwerten.** Eigenvektor– und Eigenwertprobleme sind wichtig in der Mathematik, Statik, Elektrotechnik, Maschinenbau, Biologie, Informatik und Wirtschaftswissenschaften. Oft beschreiben sie besondere Zustände von Systemen.

Beispiel: 1850 haben Soldaten die Hängebrücke von Angers zum Einsturz gebracht, indem sie mit einer Frequenz marschiert sind, die einen Eigenwert des Brückensystems getroffen hat. Es kam zur Resonanzkatastrophe, 226 Soldaten von 730 fanden den Tod. Seitdem geht man nicht mehr im Gleichschritt über Brücken. □

**Beispiel 41.4** Der Vektor

$$\mathbf{v} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

ist ein Eigenvektor von

$$A = \begin{pmatrix} 3 & 0 \\ 8 & -1 \end{pmatrix},$$

denn

$$A\mathbf{v} = \begin{pmatrix} 3 & 0 \\ 8 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 3 \\ 6 \end{pmatrix} = 3\mathbf{v}.$$

Der zugehörige Eigenwert ist $\lambda = 3$. □

**Bemerkung 41.5 Bestimmung von Eigenwerten.** Aus $A\mathbf{v} = \lambda\mathbf{v}$ folgt

$$(A - \lambda I)\,\mathbf{v} = 0.$$

Zur Bestimmung der Eigenwerte $\lambda$ muss man nun nichttriviale Lösungen dieses homogenen lineaeren Gleichungssystems suchen. Sie existieren nur falls $\operatorname{rg}(A-\lambda I) < n$, das heißt falls

$$\det(A - \lambda I) = 0,$$

siehe Definition 36.2. Für $A \in \mathbb{R}^{n\times n}$ ist dies ein Polynom $n$–ten Grades in $\lambda$. Dieses Polynom nennt man charakteristisches Polynom von $A$. Seine Nullstellen sind die gesuchten Eigenwerte. □

**Beispiel 41.6** Für

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

erhält man

$$\begin{aligned} 0 &= \det(A - \lambda I) = \begin{vmatrix} 2-\lambda & 1 \\ 6 & 1-\lambda \end{vmatrix} = (2-\lambda)(1-\lambda) - 6 \\ &= 2 - 2\lambda - \lambda + \lambda^2 - 6 = \lambda^2 - 3\lambda - 4. \end{aligned}$$

Die Nullstellen dieses Polynoms sind

$$\lambda_{1,2} = \frac{3 \pm \sqrt{9+16}}{2} = \frac{3 \pm 5}{2} \quad \Longrightarrow \quad \lambda_1 = 4,\ \lambda_2 = -1.$$

Die Matrix besitzt also die beiden Eigenwerte $\lambda_1 = 4$ und $\lambda_2 = -1$. □

**Bemerkung 41.7**

1. Selbst wenn $A$ nur reelle Einträge hat, kann das charakteristische Polynom komplexe Nullstellen besitzen. Komplexe Eigenwerte sind also nicht ungewöhnlich.
2. Sucht man Eigenwerte einer $n \times n$–Matrix $A$ als Nullstellen des charakteristischen Polynoms, kann dies für $n \geq 3$ schwierig werden. Für $n \geq 5$ ist dies im allgemeinen nicht mehr analytisch möglich. Dann werden numerische Approximationen benötigt. Diese sind ebenfalls nicht einfach.
3. Man kann zeigen, dass $\det A$ das Produkt der Eigenwerte ist und dass $A$ genau dann invertierbar ist, wenn der Eigenwert Null nicht auftritt.

□

Bei bestimmten Matrizen kann man die Eigenwerte jedoch sehr einfach bestimmen.

**Satz 41.8 Eigenwerte von Dreiecksmatrizen.** *Ist $A \in \mathbb{R}^{n\times n}$ eine obere oder untere Dreiecksmatrix, so sind die Eigenwerte durch die Diagonaleinträge gegeben.*

**Beweis:** Die Determinante einer Dreiecksmatrix ist das Produkt der Diagonaleinträge. Für $A = (a_{ij})$ folgt aus

$$0 = \det(A - \lambda I) = (a_{11} - \lambda)\dots(a_{nn} - \lambda),$$

dass die Eigenwerte durch $\lambda_1 = a_{11}, \dots, \lambda_n = a_{nn}$ gegeben sind. ■

**Beispiel 41.9** Die Matrix

$$A = \begin{pmatrix} 3 & 0 \\ 8 & -1 \end{pmatrix},$$

hat die Eigenwerte $\lambda_1 = 3$ und $\lambda_2 = -1$, vergleiche Beispiel 41.4. □

**Bemerkung 41.10 Bestimmung der Eigenvektoren.** Sei $\lambda$ ein bekannter Eigenwert der Matrix $A$. Dann sind die zugehörigen Eigenvektoren nichttriviale Lösungen von

$$(A - \lambda I)\mathbf{v} = 0. \tag{41.1}$$

Die Eigenvektoren sind nicht eindeutig bestimmt. Mit $\mathbf{v}$ ist auch $\alpha\mathbf{v}$ mit $\alpha \in \mathbb{R} \setminus \{0\}$ Eigenvektor.

Der Lösungsraum von (41.1) heißt Eigenraum von $A$ zum Eigenwert $\lambda$. Man sucht daher nach Basisvektoren im Eigenraum und gibt diese als Eigenvektoren an. □

**Beispiel 41.11** Gesucht sind die Basen der Eigenräume von

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

Zuerst bestimmt man die Eigenwerte aus $0 = \det(A - \lambda I)$. Man erhält

$$0 = \det(A - \lambda I) = (\lambda - 1)(\lambda - 2)^2.$$

Demzufolge ist $\lambda_1 = 1$ ein einfacher Eigenwert und $\lambda_2 = 2$ ein doppelter Eigenwert.

Der Eigenraum zu $\lambda_1 = 1$ ist der Lösungsraum von

$$\begin{pmatrix} -1 & 0 & -2 \\ 1 & 1 & 1 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Die erste und dritte Gleichung sind linear abhängig. Addiert man die erste zur zweiten Gleichung, erhält man

$$x_2 - x_3 = 0 \quad \Longrightarrow \quad x_2 := s \quad \Longrightarrow \quad x_3 = s.$$

Aus der dritten Gleichung folgt $x_1 = -2x_3 = -2s$. Man erhält damit den eindimensionalen Eigenraum

$$\left\{ \begin{pmatrix} -2s \\ s \\ s \end{pmatrix} \;\middle|\; s \in \mathbb{R} \right\},$$

der zum Beispiel vom Basisvektor

$$\begin{pmatrix} -2 \\ 1 \\ 1 \end{pmatrix}$$

aufgespannt wird.

Der Eigenraum zu $\lambda_2 = 2$ ist Lösungsraum von

$$\begin{pmatrix} -2 & 0 & -2 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Das sind drei linear abhängige Gleichungen mit der Lösungsmenge

$$\left\{ \begin{pmatrix} s \\ t \\ -s \end{pmatrix} \;\middle|\; s, t \in \mathbb{R} \right\}.$$

Eine Basis dieses zweidimensionalen Eigenraums ist zum Beispiel

$$\left\{ \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right\}.$$

□

**Satz 41.12 Eigenwerte von Potenzen einer Matrix.** *Seien $A \in \mathbb{R}^{n\times n}, k \in \mathbb{N}$ und $\lambda$ Eigenwert von $A$ mit Eigenvektor $\mathbf{v}$. Dann ist $\lambda^k$ Eigenwert von $A^k$ mit dem zugehörigen Eigenvektor $\mathbf{v}$.*

**Beweis:** Es gilt

$$\begin{aligned} A^k\mathbf{v} &= A^{k-1}(A\mathbf{v}) = A^{k-1}(\lambda\mathbf{v}) = \lambda A^{k-1}\mathbf{v} \\ &= \lambda A^{k-2}(A\mathbf{v}) = \lambda^2 A^{k-2}\mathbf{v} = \ldots = \lambda^k\mathbf{v}. \end{aligned}$$

■

**Beispiel 41.13** Betrachte die Matrix

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

siehe Beispiel 41.11. Die Eigenwerte von $A^8$ sind $\lambda_1 = 1^8 = 1$ und $\lambda_2 = 2^8 = 256$. □

# Kapitel 42

# Eigenwerte und Eigenvektoren symmetrischer Matrizen

Symmetrische Matrizen, das heißt es gilt $A = A^T$, kommen in der Praxis recht häufig vor. Die Eigenwerte und Eigenvektoren dieser Matrizen besitzen besondere Eigenschaften.

**Satz 42.1 Eigenwerte und Eigenvektoren symmetrischer Matrizen.** *Für eine symmetrische Matrix $A \in \mathbb{R}^{n\times n}$ gelten:*

*i) A hat nur reelle Eigenwerte.*

*ii) Eigenvektoren zu verschiedenen Eigenwerten sind orthogonal.*

**Beweis:** *i):* Für Vektoren und Matrizen definiert man die komplexe Konjugation komponentenweise. Für eine reelle Matrix gilt $\overline{A} = A$.

Sei nun $\lambda$ Eigenwert von $A$ zum Eigenvektor $\mathbf{v}$. Aus $\overline{A} = A$ und der Symmetrie von $A$ folgt

$$\begin{aligned}\overline{\lambda}\overline{\mathbf{v}}^T\mathbf{v} &= (\overline{\lambda\mathbf{v}})^T\mathbf{v} = (\overline{A\mathbf{v}})^T\mathbf{v} = \overline{\mathbf{v}}^T\overline{A}^T\mathbf{v} = \overline{\mathbf{v}}^T A^T\mathbf{v} = \overline{\mathbf{v}}^T A\mathbf{v}\\ &= \overline{\mathbf{v}}^T(\lambda\mathbf{v}) = \lambda\overline{\mathbf{v}}^T\mathbf{v}.\end{aligned}$$

Analog zu komplexen Zahlen und zu reellwertigen Vektoren ist für komplexwertige Vektoren $\overline{\mathbf{v}}^T\mathbf{v} = \|\mathbf{v}\|_2^2$. Da $\mathbf{v} \neq \mathbf{0}$ ist, gilt $\overline{\mathbf{v}}^T\mathbf{v} \in \mathbb{R}\setminus\{0\}$. Es folgt $\overline{\lambda} = \lambda$, das bedeutet $\lambda \in \mathbb{R}$.

*ii):* Seien $\mathbf{v}_1, \mathbf{v}_2$ Eigenvektoren von $A$ zu verschiedenen Eigenvektoren $\lambda_1, \lambda_2$. Dann gilt

$$\lambda_1\mathbf{v}_1^T\mathbf{v}_2 = (A\mathbf{v}_1)^T\mathbf{v}_2 = \mathbf{v}_1^T A^T\mathbf{v}_2 = \mathbf{v}_1^T A\mathbf{v}_2 = \mathbf{v}_1^T(A\mathbf{v}_2) = \lambda_2\mathbf{v}_1^T\mathbf{v}_2.$$

Daraus folgt

$$(\lambda_1 - \lambda_2)\mathbf{v}_1^T\mathbf{v}_2 = 0.$$

Da der erste Faktor nach Voraussetzung ungleich Null ist, muss der zweite Faktor verschwinden. Also sind $\mathbf{v}_1$ und $\mathbf{v}_2$ orthogonal. ■

**Beispiel 42.2** Die Matrix

$$A = \begin{pmatrix} 1 & 2\\ 2 & 4\end{pmatrix}$$

ist symmetrisch. Zur Bestimmung ihrer Eigenwerte löst man

$$\begin{aligned}0 &= \det(A-\lambda I) = \begin{vmatrix} 1-\lambda & 2\\ 2 & 4-\lambda\end{vmatrix} = (1-\lambda)(4-\lambda) - 4\\ &= 4 - \lambda - 4\lambda + \lambda^2 - 4 = \lambda^2 - 5\lambda = \lambda(\lambda - 5).\end{aligned}$$

Damit findet man die beiden reellen Eigenwerte $\lambda_1 = 0$ und $\lambda_2 = 5$.

Bestimme nun Eigenvektoren zu $\lambda_1 = 0$. Es ist

$$(A - \lambda_1 I)\mathbf{v} = \mathbf{0} \quad \Longrightarrow \quad \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Das sind zwei linear abhängige Gleichungen. Wählt man als freien Parameter $x_2 = s$, so folgt $x_1 = -2s$. Damit folgt für den Eigenraum und einen Eigenvektor

$$\left\{ \begin{pmatrix} -2s \\ s \end{pmatrix} \;\middle|\; s \in \mathbb{R} \right\}, \quad \mathbf{v}_1 = \begin{pmatrix} -2 \\ 1 \end{pmatrix}.$$

Den Eigenraum zum Eigenwert $\lambda_2 = 5$ bestimmt man analog

$$(A - \lambda_2 I)\mathbf{v} = \mathbf{0} \quad \Longrightarrow \quad \begin{pmatrix} -4 & 2 \\ 2 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Das sind auch zwei linear abhängige Gleichungen. Wählt man $x_1 = s$ als freien Parameter, so ist $x_2 = 2s$. Der Eigenraum und ein Eigenvektor sind

$$\left\{ \begin{pmatrix} s \\ 2s \end{pmatrix} \;\middle|\; s \in \mathbb{R} \right\}, \quad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

Man sieht sofort, dass $\mathbf{v}_1$ und $\mathbf{v}_2$ orthogonal sind. □

Symmetrische Matrizen lassen sich mit Hilfe ihrer Eigenwerte und Eigenvektoren elegant zerlegen.

**Satz 42.3 Hauptachsentransformation, Spektraldarstellung.** *Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch. Sei $\mathbf{v}_1, \ldots, \mathbf{v}_n$ das nach Satz 42.1 exisitierende Orthonormalsystem von Eigenvektoren mit zugehörigen, nicht notwendigerweise verschiedenen, Eigenwerten $\lambda_1, \ldots, \lambda_n$. Dann gilt*

$$A = Q\Lambda Q^T$$

*mit der Orthogonalmatrix $Q = (\mathbf{v}_1, \ldots, \mathbf{v}_n) \in \mathbb{R}^{n\times n}$ und der Diagonalmatrix*

$$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{pmatrix}.$$

**Beweis:** Nach Satz 42.1 sind die Eigenräume zu verschiedenen Eigenwerten orthogonal. Man kann zeigen, dass für symmetrische Matrizen ein $k$–facher Eigenwert einen Eigenraum der Dimension $k$ besitzt.

Verwendet man innerhalb jedes Eigenraums das Gram–Schmidt–Verfahren und normiert, entsteht ein Orthonormalsystem $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ von Eigenvektoren von $A$. Somit ist $Q := (\mathbf{v}_1, \ldots, \mathbf{v}_n)$ eine orthogonale Matrix. Die $k$–te Spalte von $Q^T AQ$ lautet

$$Q^T A\mathbf{v}_k = \lambda_k Q^T \mathbf{v}_k = \lambda_k \mathbf{e}_k.$$

Somit ist $Q^T AQ = \Lambda$. Mit der Orthogonalität von $Q$ folgt die Behauptung. ■

**Bemerkung 42.4**

1. Die Aussage von Satz 42.3 bedeutet, dass $A$ auf Diagonalgestalt transformiert werden kann
$$\Lambda = Q^T AQ.$$
Durch den Übergang in das durch $Q = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$ definierte Koordinatensystem hat $A$ eine besonders einfache Gestalt.

2. Multipliziert man das Produkt $Q\Lambda Q^T$ aus, so erhält man die Darstellung

$$A = \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T + \ldots + \lambda_n \mathbf{v}_n \mathbf{v}_n^T.$$

An dieser Darstellung erkennt man sofort, dass $\lambda_1, \ldots, \lambda_n$ Eigenwerte und $\mathbf{v}_1, \ldots, \mathbf{v}_n$ Eigenvektoren von $A$ sind. Es gilt nämlich

$$A\mathbf{v}_k = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^T \mathbf{v}_k = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \delta_{ik} = \lambda_k \mathbf{v}_k.$$

□

**Beispiel 42.5** Die Matrix

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$$

soll auf Diagonalgestalt transformiert werden.

Nach Beispiel 42.2 hat $A$ die Eigenwerte $\lambda_1 = 0$ und $\lambda_2 = 5$ mit zugehörigen Eigenvektoren

$$\mathbf{w}_1 = \begin{pmatrix} -2 \\ 1 \end{pmatrix}, \quad \mathbf{w}_2 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

Normierung der Eigenvektoren ergibt

$$\mathbf{v}_1 = \frac{1}{\sqrt{5}} \begin{pmatrix} -2 \\ 1 \end{pmatrix}, \quad \mathbf{v}_2 = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

Mit der orthogonalen Matrix

$$Q = (\mathbf{v}_1, \mathbf{v}_2) = \frac{1}{\sqrt{5}} \begin{pmatrix} -2 & 1 \\ 1 & 2 \end{pmatrix}$$

ergibt sich

$$\begin{aligned} Q^T A Q &= \frac{1}{5} \begin{pmatrix} -2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} -2 & 1 \\ 1 & 2 \end{pmatrix} = \frac{1}{5} \begin{pmatrix} -2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 0 & 5 \\ 0 & 10 \end{pmatrix} \\ &= \frac{1}{5} \begin{pmatrix} 0 & 0 \\ 0 & 25 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 5 \end{pmatrix} = \operatorname{diag}(\lambda_1, \lambda_2). \end{aligned}$$

□

# Kapitel 43

# Quadratische Formen und positiv definite Matrizen

**Bemerkung 43.1 Motivation.** Positiv definite symmetrische Matrizen sind eine wichtige Klasse von symmetrischen Matrizen. Sie treten bei Anwendungen in der Physik, der Computergraphik und auf anderen Gebieten auf.

In einem engen Zusammenhang mit symmetrisch positiv definiten Matrizen sehen quadratische Funktionen in mehreren Variablen. Spezielle Funktionen dieser Gestalt beschreiben wichtige Kurven oder Flächen. □

**Definition 43.2 Quadratische Form, quadratisches Polynom, Quadrik.** *Seien* $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$ *und* $A \in \mathbb{R}^{n \times n}$ *symmetrisch. Dann heißt*

$$\mathbf{x}^T A \mathbf{x} = \sum_{i,j=1}^{n} a_{ij} x_i x_j$$

*quadratische Form. Ferner seien* $\mathbf{b} \in \mathbb{R}^n$ *und* $c \in \mathbb{R}$*. Dann nennt man*

$$q(\mathbf{x}) = \mathbf{x}^T A \mathbf{x} + b^T \mathbf{x} + c$$

*quadratisches Polynom in* $x_1, \ldots, x_n$*.*

*Die Menge aller Punkte, welche die quadratische Gleichung*

$$q(\mathbf{x}) = \mathbf{x}^T A \mathbf{x} + b^T \mathbf{x} + c = 0$$

*erfüllt, heißt Quadrik.* □

**Beispiel 43.3**

1. Eine quadratische Form ist durch

$$\begin{aligned} & 7x_1^2 + 6x_2^2 + 5x_3^2 - 4x_1x_2 + 2x_2x_3 \\ = \; & 7x_1^2 + 6x_2^2 + 5x_3^2 - 2x_1x_2 - 2x_2x_1 + x_2x_3 + x_3x_2 \\ = \; & (x_1, x_2, x_3) \begin{pmatrix} 7 & -2 & 0 \\ -2 & 6 & 1 \\ 0 & 1 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \end{aligned}$$

   gegeben.

2. Ein quadratisches Polynom ist durch

$$q(\mathbf{x}) = 5x_1^2 - 3x_2^2 + 4x_1x_2 - 7x_2 + 3$$

   gegeben.

3. Die Ellipsengleichung
$$\frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} - 1 = 0$$
beschreibt eine Quadrik mit $n = 2$.
Quadriken mit $n = 2$ können generell als Kegelschnitte (Ellipsen, Parabeln, Hyperbeln, ...) interpretiert werden. Für $n = 3$ ergeben sich Ellipsoide, Paraboloide, Hyperboloide.

□

**Definition 43.4 Definite, semidefinite, indefinite Matrizen.** Seien $A \in \mathbb{R}^{n\times n}$ eine symmetrische Matrix und $\lambda_1, \ldots, \lambda_n$ die Eigenwerte von $A$. Dann heißt $A$

1.) positiv definit falls $\lambda_i > 0$ für alle $i = 1, \ldots, n$,
2.) positiv semidefinit falls $\lambda_i \geq 0$ für alle $i = 1, \ldots, n$,
3.) negativ definit falls $\lambda_i < 0$ für alle $i = 1, \ldots, n$,
4.) negativ semidefinit falls $\lambda_i \leq 0$ für alle $i = 1, \ldots, n$,
5.) indefinit, falls $\lambda_i, \lambda_j$ existieren mit $\lambda_i\lambda_j < 0$.

□

Es besteht ein enger Zusammenhang zwischen positiver Definitheit, quadratischen Formen und Determinanten von Untermatrizen.

**Satz 43.5 Positiv definite Matrizen und quadratische Formen.** *Es sei $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ symmetrisch. Dann ist $A$ positiv definit genau dann, wenn*
$$\mathbf{x}^T A\mathbf{x} > 0 \quad \forall\, \mathbf{x} \in \mathbb{R}^n,\ \mathbf{x} \neq \mathbf{0}.$$

**Beweis:** $\Longrightarrow$: Es sei $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ eine Orthonormalbasis von Eigenvektoren von $A$ mit zugehörigen Eigenwerten $\lambda_1, \ldots, \lambda_n$. Dann lässt sich jeder Vektor $\mathbf{x} \in \mathbb{R}^n$ als Linearkombination der Eigenvektoren darstellen $\mathbf{x} = \sum_{i=1}^n x_i\mathbf{v}_i$. Für $\mathbf{x} \neq \mathbf{0}$ gilt
$$\begin{aligned}
\mathbf{x}^T A\mathbf{x} &= \left(\sum_{i=1}^n x_i\mathbf{v}_i\right)^T A \left(\sum_{j=1}^n x_j\mathbf{v}_j\right) \\
&= \left(\sum_{i=1}^n x_i\mathbf{v}_i\right)^T \left(\sum_{j=1}^n x_j \underbrace{A\mathbf{v}_j}_{\lambda_j\mathbf{v}_j}\right) \\
&= \sum_{i,j=1}^n \lambda_j x_i x_j \underbrace{\mathbf{v}_i^T\mathbf{v}_j}_{=\delta_{ij}} = \sum_{i=1}^n \lambda_i x_i^2 > 0.
\end{aligned}$$

$\Longleftarrow$: Ist umgekehrt $A$ nicht positiv definit, so existiert ein Eigenwert $\lambda \leqq 0$ von $A$ mit zugehörigem Eigenvektor $\mathbf{v} \neq \mathbf{0}$. Damit ist
$$\mathbf{v}^T A\mathbf{v} = \mathbf{v}^T\lambda\mathbf{v} = \underbrace{\lambda}_{\leq 0}\underbrace{\mathbf{v}^T\mathbf{v}}_{>0} \leqq 0$$
im Widerspruch zu $\mathbf{x}^T A\mathbf{x} > 0$ für alle $\mathbf{x} \neq \mathbf{0}$. ■

Der Zusammenhang zwischen positiver Definitheit und Determinanten von Untermatrizen ist Gegenstand des folgenden Satzes. Er stellt ein wichtiges Kriterium zum Überprüfen der positiven Definitheit ohne Eigenwertberechnung dar.

**Satz 43.6 Hauptminorenkriterium.** *Eine symmetrische Matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ ist genau dann positiv definit, wenn ihre Hauptminoren*
$$\begin{vmatrix} a_{11} & \ldots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \ldots & a_{kk} \end{vmatrix}$$
*für $k = 1, \ldots, n$ positiv sind.*

**Beweis:** Siehe Literatur. ■

**Bemerkung 43.7** Ein ähnliches Kriterium für Semidefinitheit anzugeben ist nicht so einfach, man muss dann alle quadratischen Untermatrizen einbeziehen und nicht nur die Hauptminoren. □

**Beispiel 43.8** Betrachte die Matrix

$$A = \begin{pmatrix} 2 & -1 & -3 \\ -1 & 2 & 4 \\ -3 & 4 & 9 \end{pmatrix}.$$

Es gilt für die Hauptminoren von $A$

$$\begin{aligned} \det(2) &= 2 > 0, \\ \begin{vmatrix} 2 & -1 \\ -1 & 2 \end{vmatrix} &= 4 - 1 = 3 > 0, \\ \begin{vmatrix} 2 & -1 & -3 \\ -1 & 2 & 4 \\ -3 & 4 & 9 \end{vmatrix} &\overset{\text{Sarrus}}{=} 36 + 12 + 12 - 18 - 32 - 9 = 1 > 0. \end{aligned}$$

Nach Satz 43.6 folgt, dass $A$ positiv definit ist. □

Analog zu Quadratwurzeln aus nichtnegativen Zahlen lassen sich auch Wurzeln aus einer positiv semidefiniten Matrix definieren.

**Satz 43.9 Wurzel einer positiv semidefiniten Matrix.** *Es sei $A \in \mathbb{R}^{n\times n}$ symmetrisch positiv semidefinit. Dann existiert eine symmetrisch positiv semidefinite Matrix $B \in \mathbb{R}^{n\times n}$ mit $B^2 = A$.*

**Beweis:** Es seien $\lambda_1, \ldots, \lambda_n$ die Eigenwerte von $A$ und $\mathbf{v}_1, \ldots, \mathbf{v}_n$ die zugehörigen normierten Eigenvektoren. Dann gilt mit $Q = (\mathbf{v}_1 \ldots \mathbf{v}_n)$ und $\Lambda = \operatorname{diag}(\lambda_i)$, dass $A = Q\Lambda Q^T$, Satz 42.3.

Setze nun $\Lambda^{1/2} := \operatorname{diag}(\sqrt{\lambda_i})$ und $B := Q\Lambda^{1/2}Q^T$. Dann folgt

$$B^2 = Q\Lambda^{1/2}\underbrace{Q^TQ}_{I}\Lambda^{1/2}Q^T = Q\Lambda^{1/2}\Lambda^{1/2}Q^T = Q\Lambda Q^T = A.$$

■

Positiv und negativ definite Matrizen spielen eine wichtige Rolle beim Nachweis von Minima und Maxima von Funktionen mehrerer Variablen, siehe Kapitel 51.

Die nächste Fragestellung betrifft die Existenz oberer und unterer Schranken für Werte quadratischer Formen.

**Definition 43.10 Rayleigh[1]–Quotient.** Es seien $A \in \mathbb{R}^{n\times n}$ symmetrisch und $\mathbf{x} \in \mathbb{R}^n$ mit $\mathbf{x} \neq \mathbf{0}$. Dann nennt man

$$R_A(\mathbf{x}) := R(\mathbf{x}) := \frac{\mathbf{x}^TA\mathbf{x}}{\mathbf{x}^T\mathbf{x}}$$

den Rayleigh–Quotienten. □

Der Rayleigh–Quotient lässt sich durch die Eigenwerte von $A$ abschätzen.

**Satz 43.11 Rayleigh–Prinzip.** *Es sei $A \in \mathbb{R}^{n\times n}$ symmetrisch mit den Eigenwerten $\lambda_1 \geq \ldots \geq \lambda_n$ und zugehörigen orthonormierten Eigenvektoren $\mathbf{v}_1, \ldots, \mathbf{v}_n$. Dann gelten:*

[1] Rayleigh ??

*i)* $\lambda_n \leq R(\mathbf{x}) \leq \lambda_1$,
*ii) Diese Grenzen werden tatsächlich angenommen*

$$\lambda_1 = \max_{\mathbf{x}\in\mathbb{R}^n, \mathbf{x}\neq\mathbf{0}} R(\mathbf{x}), \quad \lambda_n = \min_{\mathbf{x}\in\mathbb{R}^n, \mathbf{x}\neq\mathbf{0}} R(\mathbf{x}).$$

**Beweis:** *i):* Sei $\mathbf{x} \in \mathbb{R}^n$ beliebig. Dann lässt sich $\mathbf{x}$ als Linearkombination der Eigenvektoren von $A$ darstellen $\mathbf{x} = \sum_{i=1}^n x_i \mathbf{v}_i$. Daraus folgt $\mathbf{x}^T\mathbf{x} = \sum_{i=1}^n x_i^2$. Analog zum Beweis von Satz 43.5 ist außerdem $\mathbf{x}^T A\mathbf{x} = \sum_{i=1}^n \lambda_i x_i^2$. Damit lässt sich der Rayleigh–Quotient wie folgt darstellen.

$$R(\mathbf{x}) = \frac{\sum_{i=1}^n \lambda_i x_i^2}{\sum_{i=1}^n x_i^2}.$$

Man erhält unmittelbar

$$R(\mathbf{x}) \leq \frac{\sum_{i=1}^n \lambda_1 x_i^2}{\sum_{i=1}^n x_i^2} = \lambda_1, \quad R(\mathbf{x}) \geq \frac{\sum_{i=1}^n \lambda_n x_i^2}{\sum_{i=1}^n x_i^2} = \lambda_n.$$

*ii):* Setzt man $\mathbf{x} = \mathbf{v}_k$, so folgt

$$R(\mathbf{v}_k) = \frac{\mathbf{v}_k^T A\mathbf{v}_k}{\mathbf{v}_k^T\mathbf{v}_k} = \frac{\mathbf{v}_k^T \lambda_k \mathbf{v}_k}{\mathbf{v}_k^T\mathbf{v}_k} = \lambda_k.$$

Insbesondere sind $R(\mathbf{v}_1) = \lambda_1$ und $R(\mathbf{v}_n) = \lambda_n$. ■

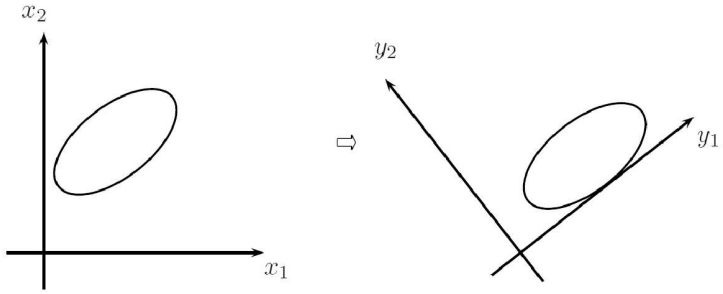
# Kapitel 44

# Quadriken

**Bemerkung 44.1 Motivation.** Quadriken, siehe Definition 43.2, stellen eine wichtige Klasse geometrischer Objekte dar. Diese besitzen unter anderem Anwendungen in der Computergraphik und der Physik. Oft kann man aber aus der angegebenen Form einer Quadrik deren Gestalt nicht unmittelbar erkennen. Ziel ist es, eine gegebene Quadrik auf einfache Form transformieren, sodass sich ihre geometrische Gestalt unmittelbar ablesen lässt. □

**Bemerkung 44.2 Grundlegende Verfahrensweise.** Gegeben sei eine Quadrik $q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x} + \mathbf{b}^T\mathbf{x} + c = 0$ mit $A \in \mathbb{R}^{n\times n}$ symmetrisch, $\mathbf{x}, \mathbf{b} \in \mathbb{R}^n,\ c \in \mathbb{R}$.

Schritt 1: *Elimination der gemischten quadratischen Terme.* Hierzu wird das Koordinatensystem so gedreht, dass $A$ in eine Diagonalmatrix übergeht.




Dazu berechnet man die Eigenwerte $\lambda_i$ von $A$ und eine Orthonormalbasis $\{\mathbf{v}_1, \dots \mathbf{v}_n\}$ aus Eigenvektoren mit $\det(\mathbf{v}_1, \dots, \mathbf{v}_n) = 1$. Ist $\det(\mathbf{v}_1, \dots, \mathbf{v}_n) = -1$, ersetzt man etwa $\mathbf{v}_1$ durch $-\mathbf{v}_1$. Mit $Q = (\mathbf{v}_1 \dots \mathbf{v}_n)$ gilt dann

$$\Lambda = \operatorname{diag}(\lambda_i) = Q^T A Q.$$

Aus $\mathbf{x}^T A\mathbf{x} + \mathbf{b}^T\mathbf{x} + c = 0$ folgt

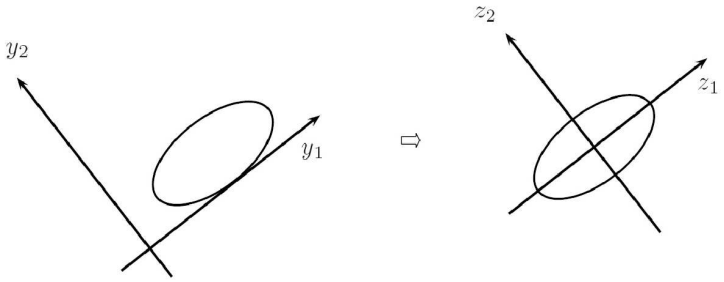
$$\mathbf{x}^T Q\Lambda Q^T\mathbf{x} + \mathbf{b}^T \underbrace{QQ^T}_{=I}\mathbf{x} + c = 0.$$

Mit $\mathbf{y} := Q^T\mathbf{x}$, $\tilde{\mathbf{b}} := Q^T\mathbf{b}$ ergibt sich daher $\mathbf{y}^T\Lambda\mathbf{y} + \tilde{\mathbf{b}}^T\mathbf{y} + c = 0$ beziehungsweise ausgeschrieben

$$\lambda_1 y_1^2 + \ldots + \lambda_n y_n^2 + \tilde{b}_1 y_1 + \ldots + \tilde{b}_n y_n + c = 0.$$

Die gemischten quadratischen Terme sind eliminiert.

Schritt 2: *Elimination linearer Terme (soweit möglich).* Durch Translation des Koordinatensystems kann erreicht werden, dass $\lambda_k y_k^2$ und $\tilde{b}_k y_k$ nicht zugleich vorkommen für jedes $k = 1, \dots, n$.




Es sei dazu o.B.d.A. $\lambda_i \neq 0$ für $i = 1, \ldots, r$ sowie $\lambda_{r+1} = \ldots = \lambda_n = 0$. Für $i = 1, \ldots, r$ wird der lineare Term $\tilde{b}_i y_i$ durch die quadratische Ergänzung eliminiert

$$z_i := y_i + \frac{\tilde{b}_i}{2\lambda_i} \; i = 1, \ldots, r, \quad z_i := y_i \; i = r+1, \ldots, n.$$

Damit erhält man

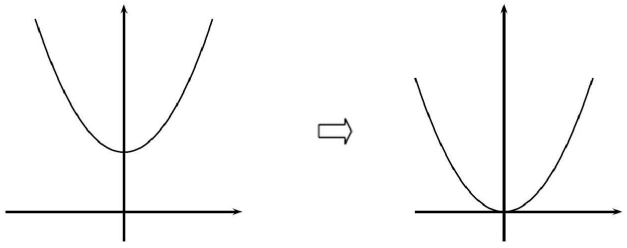
$$\lambda_1 z_1^2 + \ldots + \lambda_r z_r^2 + \tilde{b}_{r+1} z_{r+1} + \ldots + \tilde{b}_n z_n + \tilde{c} = 0$$

mit $\tilde{c} = c - \sum_{i=1}^{r} \tilde{b}_i^2/(4\lambda_i)$ und $r = \mathrm{rg}(A)$.

Schritt 3: *Elimination der Konstanten (falls möglich).* Ist einer der Koeffizienten $\tilde{b}_{r+1}, \ldots, \tilde{b}_n$ ungleich Null, o.B.d.A. sei dies $\tilde{b}_n$, so kann $\tilde{c}$ eliminiert werden durch

$$z_n \mapsto z_n - \frac{\tilde{c}}{\tilde{b}_n}.$$

Das ist ebenfalls eine Translation des Koordinatensystems, zum Beispiel



Resultat: *Normalformen der Quadrik.* Das ist eine Darstellung in Koordinatensystem, in dem möglichst viele Koeffizienten verschwinden. Für $\mathrm{rg}(A) = n = r$ erhält man

$$\lambda_1 z_1^2 + \ldots + \lambda_n z_n^2 + d = 0.$$

Ist $r < n$ ergibt sich eine der beiden folgenden Normalformen

$$\begin{aligned} \lambda_1 z_1^2 + \ldots + \lambda_r z_r^2 + e_{r+1} z_{r+1} + \ldots + e_n z_n &= 0, \\ \lambda_1 z_1^2 + \ldots + \lambda_r z_r^2 + d &= 0. \end{aligned}$$

□

**Beispiel 44.3** Die Quadrik

$$q(\mathbf{x}) = 5x_1^2 - 4x_1x_2 + 8x_2^2 + \frac{20}{\sqrt{5}}x_1 - \frac{80}{\sqrt{5}}x_2 + 4 = 0$$

soll auf Normalform gebracht werden. Es ist

$$q(\mathbf{x}) = \mathbf{x}^T A \mathbf{x} + b^T \mathbf{x} + c = 0 \quad \text{mit} \quad A = \begin{pmatrix} 5 & -2 \\ -2 & 8 \end{pmatrix}, \; \mathbf{b} = \frac{20}{\sqrt{5}} \begin{pmatrix} 1 \\ -4 \end{pmatrix}, \; c = 4.$$

*Hauptachsentransformation von $A$.* Man berechnet zuerst die Eigenwerte und Eigenvektoren von $A$. Für die Eigenwerte gilt

$$0 = \det(A - \lambda I) = (5-\lambda)(8-\lambda) - 4 = \lambda^2 - 13\lambda + 36 \quad \Longrightarrow \quad \lambda_1 = 9,\ \lambda_2 = 4.$$

Dann berechnet man Eigenvektoren durch die Lösung der singulären homogenen Systeme. Man erhält

$$Q = \frac{1}{\sqrt{5}} \begin{pmatrix} -1 & -2 \\ 2 & -1 \end{pmatrix}, \quad \det Q = 1.$$

Mit $\Lambda = Q^T A Q = \operatorname{diag}(9,4)$ und $\tilde{\mathbf{b}} = Q^T \mathbf{b} = (-36, 8)^T$ ergibt sich für $\mathbf{y} = Q^T \mathbf{x}$

$$9y_1^2 + 4y_2^2 - 36y_1 + 8y_2 + 4 = 0.$$

*Elimination linearer Terme.* Quadratische Ergänzung ergibt

$$9\left(y_1^2 - 4y_1 + 4\right) + 4\left(y_2^2 + 2y_2 + 1\right) = -4 + 36 + 4.$$

Mit $z_1 = y_1 - 2$ und $z_2 = y_2 + 1$ erhält man

$$9z_1^2 + 4z_2^2 = 36 \quad \Longrightarrow \quad \frac{z_1^2}{4} + \frac{z_2^2}{9} = 1.$$

Das ist eine Ellipse mit den Halbachsen 2 und 3, siehe auch Abbildung 44.1

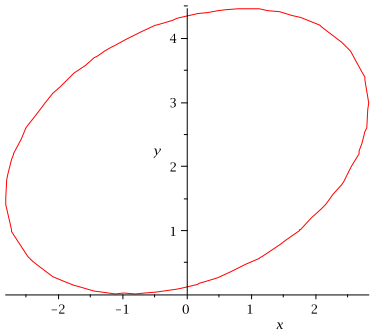



Abbildung 44.1: Originallage der Quadrik von Beispiel 44.3.

□

**Bemerkung 44.4 Normalformen der Quadriken im $\mathbb{R}^2$, Kegelschnitte.**
1.Fall $\operatorname{rg}(A) = 2$, das heißt alle Eigenwerte von $A$ sind ungleich Null.

i) Ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - 1 = 0,$$

ii) Hyperbel

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} - 1 = 0,$$

iii) leere Menge

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + 1 = 0,$$

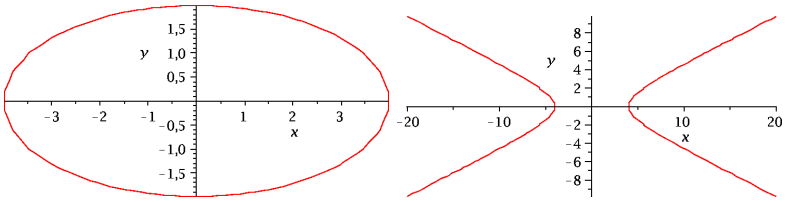



Abbildung 44.2: Ellipse (links) und Hyperbel (rechts).

iv) Punkt $(0,0)^T$

$$x^2 + a^2 y^2 = 0,\ a \neq 0,$$

v) Geradenpaar

$$x^2 - a^2 y^2 = 0,\ a \neq 0 \quad \Longrightarrow \quad y = \pm \frac{1}{a} x.$$

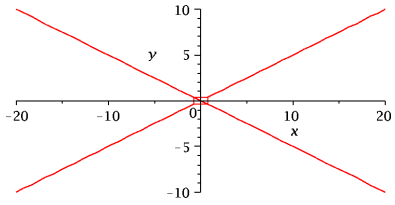



Abbildung 44.3: Geradenpaar.

2.Fall $\operatorname{rg}(A) = 1$, das heißt, ein Eigenwert von $A$ ist Null.

i) Parabel

$$x^2 - 2py = 0,$$

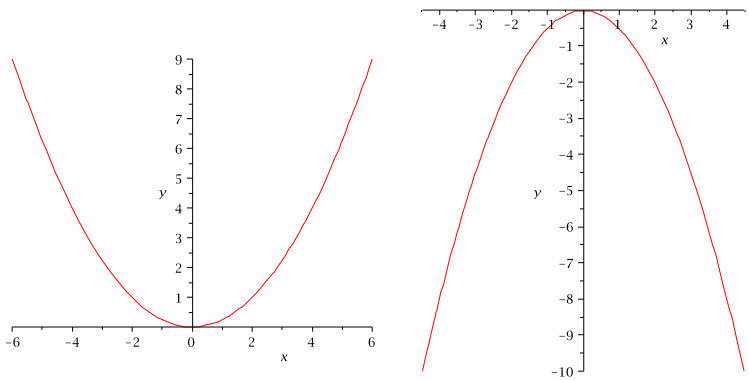



Abbildung 44.4: Parabeln.

ii) parallele Geraden

$$x^2 - a^2 = 0,\ a \neq 0 \quad \Longrightarrow \quad x = \pm a,$$

iii) leere Menge

$$x^2 + a^2 = 0,\ a \neq 0,$$

iv) Doppelgerade $x = 0$ ($y$–Achse)

$$x^2 = 0.$$

3.Fall $\operatorname{rg}(A) = 0$, das heißt beide Eigenwerte von $A$ sind Null.

i) Gerade

$$b_1 x + b_2 y + c = 0.$$

□

**Bemerkung 44.5 Normalformen der Quadriken im $\mathbb{R}^3$.**

1.Fall $\operatorname{rg}(A) = 3$, das heißt alle Eigenwerte von $A$ sind ungleich Null.

i) Ellipsoid

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} - 1 = 0,$$

ii) leere Menge

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} + 1 = 0,$$

iii) einschaliges Hyperboloid

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} - 1 = 0.$$

Betrachtet man den Schnitt dieses Körpers mit Ebenen, so erhält man eine Ellipse in einer Ebene ($x$–$y$–Ebene) und Hyperbeln in zwei Ebenen ($x$–$z$–, $y$–$z$–Ebene).

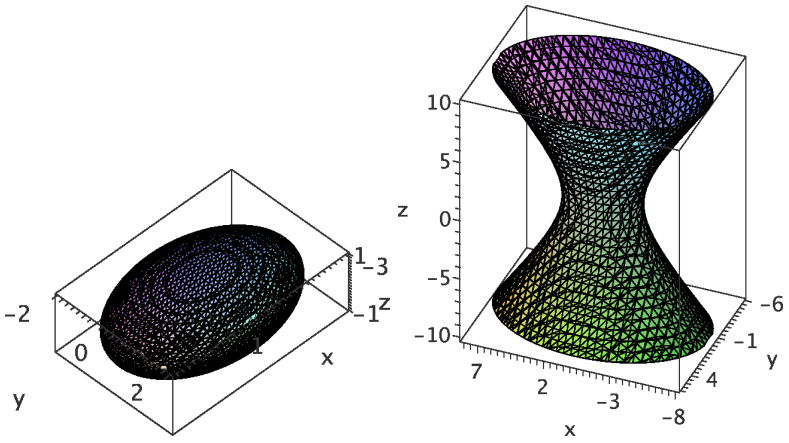



Abbildung 44.5: Ellipsoid (links) und einschaliges Hyperboloid (rechts).

iv) zweischaliges Hyperboloid

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} + 1 = 0.$$

Auch hier erhält man beim Schneiden mit Ebenen eine Ellipse in einer Ebene ($x$–$y$–Ebene) und Hyperbeln mit zwei Ebenen ($x$–$z$–, $y$–$z$–Ebene).

v) Punkt $(0, 0, 0)^T$

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 0,$$

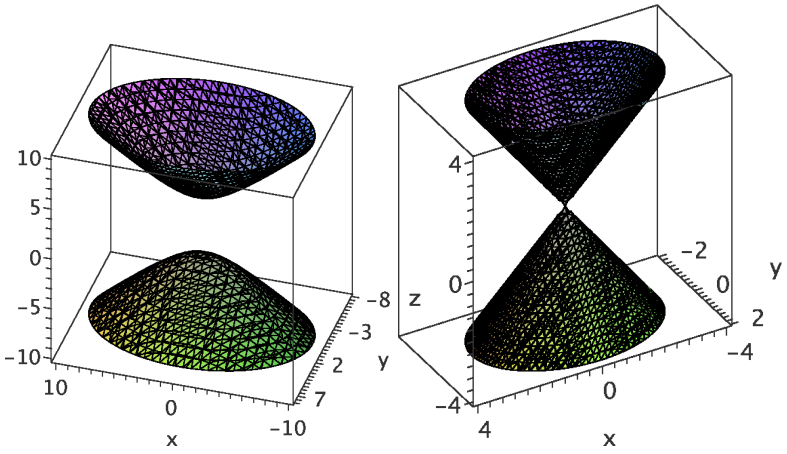



Abbildung 44.6: zweischaliges Hyperboloid (links) und elliptischer Kegel (rechts).

vi) elliptischer Kegel

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 0.$$

2.Fall $\operatorname{rg}(A) = 2$, das bedeutet ein Eigenwert ist gleich Null.

i) elliptisches Paraboloid

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - 2pz = 0.$$

Der Schnitt mit einer Ebene ist eine Ellipse ($x$–$y$–Ebene) und der Schnitt mit zwei Ebenen sind Parabeln ($x$–$z$–, $y$–$z$–Ebene).

ii) hyperbolisches Paraboloid, Sattelfläche

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} - 2pz = 0.$$

Der Schnitt mit einer Ebene ist eine Hyperbel ($x$–$y$–Ebene) und der Schnitt mit zwei Ebenen sind Parabeln ($x$–$z$–, $y$–$z$–Ebene).

iii) leere Menge

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + 1 = 0,$$

iv) elliptischer Zylinder

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - 1 = 0,$$

v) hyperbolischer Zylinder

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} + 1 = 0,$$

vi) Gerade ($z$–Achse)

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 0,$$

vii) Ebenenpaar mit Schnittgerade ($z$–Achse)

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 0,$$

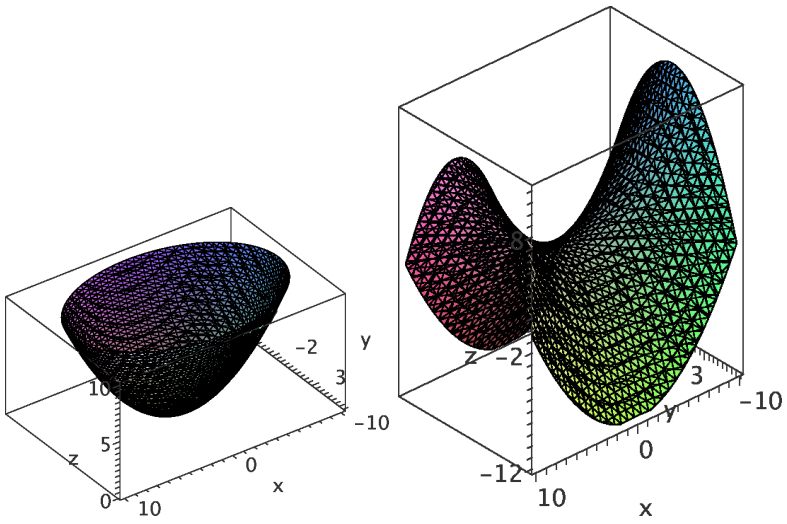



Abbildung 44.7: elliptisches Paraboloid (links) und hyperbolisches Paraboloid (rechts).

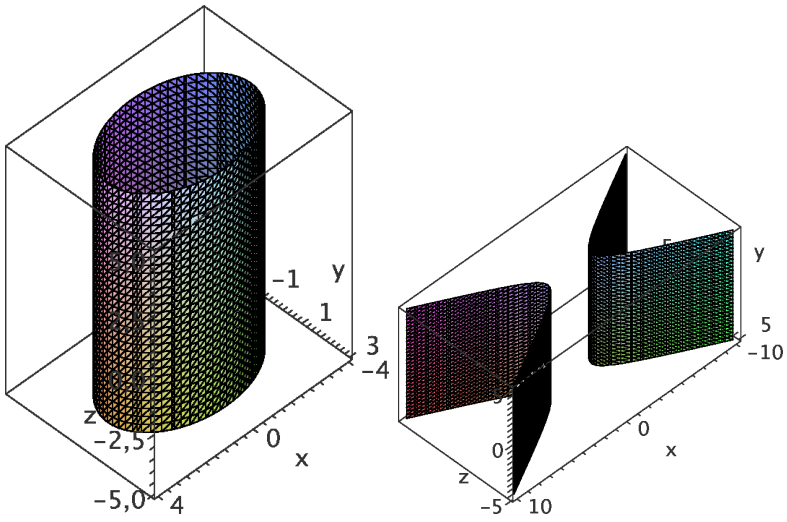



Abbildung 44.8: elliptischer Zylinder (links) und hyperbolischer Zylinder (rechts).

3.Fall $\operatorname{rg}(A) = 1$, also zwei Eigenwerte von $A$ sind Null.

i) parabolischer Zylinder
$$x^2 - 2pz = 0,$$

ii) paralleles Ebenenpaar
$$x^2 - a^2 = 0, \quad a \neq 0,$$

iii) leere Menge
$$x^2 + a^2 = 0, \quad a \neq 0,$$

iv) Ebene ($y$–$z$–Ebene)
$$x^2 = 0.$$

4.Fall $\operatorname{rg}(A) = 0$, das bedeutet, $A$ ist die Nullmatrix. Dann erhält man die allgemeine Ebenengleichung
$$b_1x + b_2y + b_3z + c = 0.$$

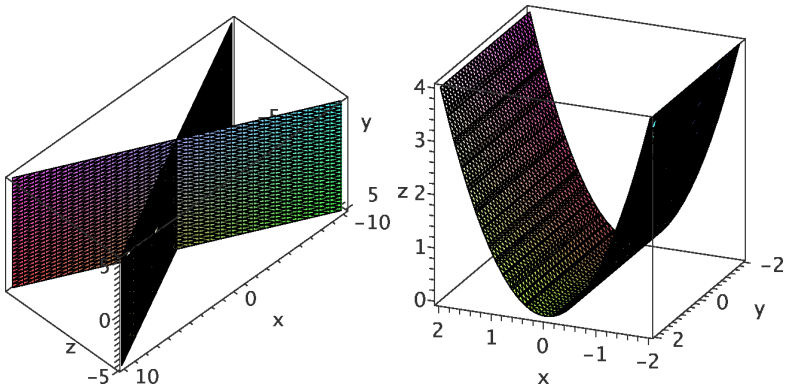



Abbildung 44.9: Ebenenpaar mit Schnittgerade (links) und parabolischer Zylinder (rechts).

□

# Kapitel 45

# Matrixnormen und Eigenwertabschätzungen

**Bemerkung 45.1 Motivation.** Matrixnormen sind ein wichtiges Hilfsmittel zur Beschreibung von Eigenschaften von Matrizen, die zum Beispiel in numerischen Verfahren wichtig sind. Insbesondere kann man mit Hilfe von Matrixnormen die Eigenwerte einer Matrix mit geringem Aufwand abschätzen. □

**Definition 45.2 Matrixnorm.** Unter einer Matrixnorm versteht man eine Funktion $\|\cdot\| : \mathbb{R}^{m\times n} \to \mathbb{R}$ mit folgenden Eigenschaften:

1.) $\|A\| \geq 0$ für alle $A \in \mathbb{R}^{m\times n}$ und $\|A\| = 0$ genau dann wenn $A = 0$,
2.) $\|\lambda A\| = |\lambda| \, \|A\|$ für alle $A \in \mathbb{R}^{m\times n}$ und $\lambda \in \mathbb{R}$,
3.) Dreiecksungleichung: $\|A + B\| \leq \|A\| + \|B\|$ für alle $A, B \in \mathbb{R}^{m\times n}$,
4.) Submultiplikativität: $\|AB\| \leq \|A\| \, \|B\|$ für alle $A, B \in \mathbb{R}^{n\times n}$.

□

**Beispiel 45.3** Sei $A \in \mathbb{R}^{m\times n}$.

1. Gesamtnorm: nur definiert falls $m = n$ ist,
$$\|A\|_G := n \max_{i,j=1,\ldots,n} |a_{ij}| \,,$$
2. Zeilensummennorm:
$$\|A\|_\infty := \max_{i=1,\ldots,m} \sum_{j=1}^{n} |a_{ij}| \,,$$
3. Spaltensummennorm:
$$\|A\|_1 := \max_{j=1,\ldots,n} \sum_{i=1}^{m} |a_{ij}| \,,$$
4. Frobenius–Norm[1]
$$\|A\|_F := \left( \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2 \right)^{1/2},$$
5. Spektralnorm
$$\|A\|_2 := \sqrt{\lambda_{\max}\left(A^T A\right)},$$
wobei $\lambda_{\max}\left(A^T A\right)$ der größte Eigenwert von $A^T A$ ist,

[1] Ferdinand Georg Frobenius (1849 – 1917)

6. Maximumsnorm:
$$\|A\|_M := \max_{j=1,\ldots,m;j=1,\ldots,n} |a_{ij}|\,,$$
ist eine Norm, erfüllt also die Eigenschaften 1.) – 3.), ist aber nicht submultiplikativ.

□

Es gibt Matrixnormen, die direkt etwas mit bekannten Vektornormen, siehe Beispiel 16.4 zu tun haben.

**Definition 45.4 Induzierte Matrixnorm.** Sei $A \in \mathbb{R}^{m\times n}$. Die durch die $l^p$–Vektornorm induzierte Matrixnorm ist gegeben durch
$$\|A\|_p := \max_{\mathbf{x}\in\mathbb{R}^n,\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|_p}{\|\mathbf{x}\|_p} = \max_{\|\mathbf{x}\|_p=1} \|A\mathbf{x}\|_p\,.$$

□

**Beispiel 45.5** Es gelten:
1. Die Betragssummen–Vektornorm induziert die Spaltensummennorm.
2. Die euklidische Vektornorm induziert die Spektralnorm.
3. Die Maximums–Vektornorm induziert die Zeilensummennorm.

□

Oft muss man die Norm von Matrix–Vektor–Produkten abschätzen. Dafür braucht man die folgende Eigenschaft.

**Definition 45.6 Kompatibilität, Verträglichkeit von Normen.** Eine Matrixnorm $\|\cdot\|_m$ heißt kompatibel (verträglich) mit einer Vektornorm $\|\cdot\|_v$, falls gilt
$$\|A\mathbf{x}\|_v \leq \|A\|_m \|\mathbf{x}\|_v$$
für alle $A \in \mathbb{R}^{m\times n}$ und für alle $\mathbf{x} \in \mathbb{R}^n$. □

**Beispiel 45.7** Zu den $l^p$–Vektornormen
$$\|\mathbf{x}\|_p := \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}, \quad p \in [1,\infty); \qquad \|\mathbf{x}\|_\infty := \max_{i=1,\ldots,n} |x_i|\,,$$
bestehen folgende Verträglichkeiten:
1. $\|A\|_G\,, \|A\|_1$ sind kompatibel zur Betragssummennorm $\|\mathbf{x}\|_1$,
2. $\|A\|_G\,, \|A\|_F\,, \|A\|_2$ sind kompatibel zur euklidischen Norm $\|\mathbf{x}\|_2$,
3. $\|A\|_G\,, \|A\|_\infty$ sind kompatibel zur Maximumsnorm $\|\mathbf{x}\|_\infty$.

Die Beweis der Verträglichkeit einer Vektornorm mit ihrer induzierten Matrixnorm ist Übungsaufgabe. Man kann zeigen, dass die induzierte Matrixnorm die kleinste aller Matrixnormen ist, die zu einer gegebenen Vektornorm verträglich ist.

Hier wird nur die Kompatibilität von $\|A\|_G$ und $\|\mathbf{x}\|_\infty$ gezeigt. Sei $A \in \mathbb{R}^{n\times n}$, dann gilt
$$\begin{aligned}
\|A\mathbf{x}\|_\infty &= \max_{i=1,\ldots,n} \left\{\left|\sum_{k=1}^n a_{ik}x_k\right|\right\} \leq \max_{i=1,\ldots,n} \left\{\sum_{k=1}^n |a_{ik}x_k|\right\} \\
&\leq \max_{i=1,\ldots,n} \left\{n \max_{s=1,\ldots,n} |a_{is}| \max_{l=1,\ldots,n} |x_l|\right\} \\
&= n \max_{i,s=1,\ldots,n} |a_{is}| \max_{l=1,\ldots,n} |x_l| = \|A\|_G \|\mathbf{x}\|_\infty\,.
\end{aligned}$$

□

Matrixnormen sind nützlich zur Abschätzung von Eigenwerten.

**Satz 45.8 Eigenwertabschätzung mit Matrixnormen.** *Sind $\lambda$ ein Eigenwert von $A \in \mathbb{R}^{n\times n}$ und $\|A\|$ eine beliebige, zu einer Vektornorm kompatible Matrixnorm, so gilt*

$$|\lambda| \leq \|A\| .$$

**Beweis:** Sei $\mathbf{v}$ ein Eigenvektor zu $\lambda$. Dann gilt

$$|\lambda| \, \|\mathbf{v}\| = \|\lambda \mathbf{v}\| = \|A\mathbf{v}\| \leq \|A\| \, \|\mathbf{v}\| .$$

Da $\mathbf{v} \neq \mathbf{0}$, gilt $\|\mathbf{v}\| > 0$. Damit folgt $|\lambda| \leq \|A\|$. ■

**Beispiel 45.9** Betrachte

$$A = \begin{pmatrix} 1 & 0.1 & -0.1 \\ 0 & 2 & 0.4 \\ -0.2 & 0 & 3 \end{pmatrix} .$$

Dann sind

$$\begin{aligned} \|A\|_G &= 3 \max_{i,j=1,2,3} |a_{ij}| = 3 \cdot 3 = 9, \\ \|A\|_\infty &= \max\{1.2, 2.4, 3.2\} = 3.2, \\ \|A\|_1 &= \max\{1.2, 2.1, 3.5\} = 3.5, \\ \|A\|_F &= \sqrt{1^2 + 0.1^2 + (-0.1)^2 + 2^2 + 0.4^2 + (-0.2)^2 + 3^2} = \sqrt{14.22} \approx 3.77. \end{aligned}$$

$\|A\|_\infty$ liefert die schärfste Abschätzung: $|\lambda| \leq 3.2$. Tatsächlich gilt

$$\lambda_1 \approx 3.0060, \quad \lambda_2 \approx 2.0078, \quad \lambda_3 \approx 0.9862.$$

□

Offenbar erlaubt Satz 45.8 nur die Abschätzung des betragsgrößten Eigenwertes. Es gibt auch Abschätzungen für alle Eigenwerte.

**Satz 45.10 Satz von Gerschgorin[2].** *Sei $A = (a_{ij}) \in \mathbb{R}^{n\times n}$.*

*i) Die Vereinigung aller Kreisscheiben*

$$K_i := \left\{ \mu \in \mathbb{C} \;\middle|\; |\mu - a_{ii}| \leq \sum_{k=1, k\neq i}^{n} |a_{ik}| \right\}$$

*enthält alle Eigenwerte von $A$.*

*ii) Jede Zusammenhangskomponente aus $m$ solchen Kreisen enthält genau $m$ Eigenwerte, wobei die Vielfachheit mitgezählt wird.*

**Beweis:** Siehe Literatur. ■

**Beispiel 45.11** Betrachte wie im Beispiel 45.9

$$A = \begin{pmatrix} 1 & 0.1 & -0.1 \\ 0 & 2 & 0.4 \\ -0.2 & 0 & 3 \end{pmatrix} .$$

[2]Semjon Aranowitsch Gerschgorin (1901 – 1933)

Dann sind die Gerschgorin–Kreise durch

$$\begin{aligned} K_1 &= \{\mu \in \mathbb{C} \mid |\mu - 1| \leq 0.2\}, \\ K_2 &= \{\mu \in \mathbb{C} \mid |\mu - 2| \leq 0.4\}, \\ K_3 &= \{\mu \in \mathbb{C} \mid |\mu - 3| \leq 0.2\} \end{aligned}$$

gegeben. Sämtliche Eigenwerte liegen in $K_1 \cup K_2 \cup K_3$. Da sich $K_1, K_2, K_3$ nicht überlappen, liegt nach Satz 45.10, ii) in jeder der Kreisscheiben genau ein Eigenwert. Aus den Gerschgorin–Kreisen folgt, dass $A$ invertierbar ist, da $0 \notin K_1 \cup K_2 \cup K_3$. Demzufolge ist Null kein Eigenwert. □

**Folgerung 45.12 Invertierbarkeit strikt diagonal–dominanter Matrizen.** *Sei $A \in \mathbb{R}^{n \times n}$ strikt diagonal–dominant, das heißt*

$$|a_{ii}| > \sum_{k=1, k \neq i}^{n} |a_{ik}|$$

*für alle $i = 1, \ldots, n$, so ist $A$ invertierbar.*

**Beweis:** Nach dem Satz von Gerschgorin liegt Null außerhalb der Gerschgorin–Kreise, kann also kein Eigenwert sein. ■

# Kapitel 46

# Klassische Iterationsverfahren zur Lösung linearer Gleichungssysteme

**Bemerkung 46.1 Motivation.** Direkte Lösungsverfahren für ein lineares System

$$A\mathbf{x} = \mathbf{b}, \quad A \in \mathbb{R}^{n\times n}, \mathbf{x}, \mathbf{b} \in \mathbb{R}^n, \tag{46.1}$$

sind für große $n$ recht teuer. So kostet das in Kapitel 35 vorgestellte Gaußsche Verfahren $\mathcal{O}(n^3)$ Rechenoperationen. In Anwendungen ist $n$ oft in der Größenordnung von $10^5$ – $10^8$. Darüber hinaus besitzen die Matrizen in Anwendungen oft viele Nullen, die man nicht abspeichern muss. Diese Matrizen nennt man schwach besetzt. Das Besetztheitsmuster solcher Matrizen wird bei direkten Verfahren im allgemeinen zerstört, das heißt, es werden viele neue Nichtnullelemente erzeugt. Eine Alternative zu direkten Verfahren sind iterative Verfahren zur Lösung von (46.1). In diesem Abschnitt werden die einfachsten dieser Verfahren vorgestellt. □

## 46.1 Allgemeine Theorie

**Bemerkung 46.2 Fixpunktiteration.** Die Konstruktion klassischer Iterationsverfahren zur Lösung von (46.1) startet mit der Zerlegung der Systemmatrix

$$A = M - N, \quad M, N \in \mathbb{R}^{n\times n}, \quad M \text{ ist regulär.} \tag{46.2}$$

Mit Hilfe dieser Zerlegung, kann man (46.1) in eine Fixpunktgleichung umformen

$$M\mathbf{x} = \mathbf{b} + N\mathbf{x} \quad \Longleftrightarrow \quad \mathbf{x} = M^{-1}\left(\mathbf{b} + N\mathbf{x}\right). \tag{46.3}$$

Ist eine Anfangsiterierte $\mathbf{x}^{(0)} \in \mathbb{R}^n$ gegeben, kann man nun (46.3) mit Hilfe der Fixpunktiteration

$$\mathbf{x}^{(k+1)} = M^{-1}\left(\mathbf{b} + N\mathbf{x}^{(k)}\right), \quad k = 0, 1, 2, \ldots \tag{46.4}$$

zu lösen. Der Banachsche Fixpunktsatz liefert eine Aussage über die Konvergenz dieser Iteration. □

Der Banachsche Fixpunktsatz in $\mathbb{R}$ ist in Satz 24.13 formuliert. Hier wird eine allgemeine Variante in vollständigen metrischen Räumen angegeben.

**Satz 46.3** ***Banachscher Fixpunktsatz für vollständige metrische Räume.*** *Seien $(\mathcal{X}, d)$ ein vollständiger metrischer Raum und $f : \mathcal{X} \to \mathcal{X}$ eine kontraktive Abbildung, das heißt $f$ ist Lipschitz–stetig mit einer Lipschitz–Konstanten $L < 1$. Dann besitzt die Gleichung $x = f(x)$ eine eindeutige Lösung $\bar{x} \in \mathcal{X}$, einen sogenannten Fixpunkt. Das Iterationsverfahren*

$$x^{(k+1)} = f\left(x^{(k)}\right), \quad k = 0, 1, 2, \ldots$$

*konvergiert für jede Anfangsinterierte $x^{(0)}$ gegen $\bar{x}$.*

**Beweis:** Ähnlich wie Satz 24.13, siehe Literatur. ■

Diese Aussage soll jetzt auf (46.4) angewandt werden. Dazu wird der Begriff des Spektralradius gebraucht.

**Definition 46.4 Spektralradius.** Sei $A \in \mathbb{R}^{n\times n}$. Der Spektralradius von $A$ ist definiert durch

$$\rho(A) = \max\{|\lambda| \ : \ \lambda \text{ ist Eigenwert von } A\}.$$

□

**Lemma 46.5** *Sei $A \in \mathbb{R}^{n\times n}$. Dann gilt für jede induzierte Matrixnorm $\rho(A) \leq \|A\|$.*

**Beweis:** Die Aussage folgt direkt aus Satz 45.8. ■

**Lemma 46.6** *Seien $A \in \mathbb{R}^{n\times n}$ und ein beliebiges $\varepsilon > 0$ gegeben. Dann gibt es eine Vektornorm $\|\cdot\|_*$ so, dass für die induzierte Matrixnorm $\|\cdot\|_*$*

$$\rho(A) \leq \|A\|_* \leq \rho(A) + \varepsilon$$

*gilt.*

**Beweis:** Siehe Literatur. ■

**Satz 46.7 Notwendige und hinreichende Bedingung für Konvergenz der Fixpunktiteration.** *Das Iterationsverfahren (46.4) konvergiert gegen die Lösung $\mathbf{x}$ von (46.1) für jede Anfangsiterierte $\mathbf{x}^{(0)}$ genau dann, wenn der Spektralradius der Iterationsmatrix $G = M^{-1}N$ kleiner als 1 ist: $\rho(G) < 1$.*

**Beweis:** i) Die Iteration (46.4) ist eine Fixpunktiteration mit

$$f \ : \ \mathbb{R}^n \to \mathbb{R}^n, \quad \mathbf{x} \mapsto M^{-1}N\mathbf{x} + M^{-1}\mathbf{b}.$$

Die Matrizenmultiplikation (mit $G = M^{-1}N$) und Vektoraddition (mit $M^{-1}\mathbf{b}$) sind stetig und sogar differenzierbar. Damit ist die Abbildung $f(\mathbf{x})$ Lipschitz–stetig. Sei $\|G\|_*$ zunächst irgendeine Matrixnorm. Auch im Mehrdimensionalen ist die Ableitung eines linearen Polynoms der Term vor dem linearen Faktor, siehe Kapitel 49. In Analogie zu Satz 24.11 erhält man die Lipschitz–Konstante bezüglich $\|\cdot\|_*$ durch

$$L_* = \sup_{\mathbf{x}\in\mathbb{R}^n} \|G\|_* = \|G\|_*.$$

ii) Sei $\rho(G) < 1$. Dann kann man $\varepsilon > 0$ so wählen, dass $\rho(G) + \varepsilon < 1$. Außerdem gibt es eine Matrixnorm $\|\cdot\|_*$ gemäß Lemma 46.6, so dass $\|G\|_* \leq \rho(G) + \varepsilon < 1$ gilt. Somit folgt $L_* < 1$ und $f(\mathbf{x})$ ist eine kontraktive Abbildung.

iii) Man konstruiert eine Startiterierte, für die man die Divergenz zeigt. Das ist etwas aufwändig, siehe Literatur. ■

## 46.2 Beispiele für klassische Iterationsverfahren

Klassische Iterationsverfahren nutzen eine Zerlegung der Matrix $A$ der Gestalt

$$A = D + L + U,$$

wobe $D$ die Diagonle von $A$ ist, $L$ der strikte unter Dreiecksteil und $U$ der strikte obere Dreiecksteil.

**Beispiel 46.8 Das Jacobi–Verfahren.** Man erhält das Jacobi–Verfahren, indem man

$$M = D,\ N = -(L + U)$$

setzt. Das ist natürlich nur möglich, wenn auf der Hauptdiagonalen von $A$ keine Nullen stehen. Eine direkte Rechung ergibt, dass in diesem Fall die Fixpunktgleichung (46.3) die Gestalt

$$\mathbf{x} = D^{-1}(\mathbf{b} - (L + U)\mathbf{x}) = \mathbf{x} + D^{-1}(\mathbf{b} - A\mathbf{x})$$

besitzt. Man erhält das folgende Iterationsverfahren, das sogenannte Jacobi–Verfahren,

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + D^{-1}\left(\mathbf{b} - A\mathbf{x}^{(k)}\right), \quad k = 0, 1, 2, \ldots.$$

Die Iterationsmatrix ist

$$G_{\text{Jac}} = -D^{-1}(L + U) = I - D^{-1}A.$$

□

**Beispiel 46.9 Gedämpftes Jacobi–Verfahren.** Sei $\omega \in \mathbb{R}$, $\omega > 0$. Die Matrizen, welche die Fixpunktgleichung für das gedämpfte Jacobi–Verfahren definieren, sind durch

$$M = \omega^{-1}D,\ N = \omega^{-1}D - A$$

gegeben. Das gedämpfte Jacobi–Verfahren besitzt die Gestalt

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \omega D^{-1}\left(\mathbf{b} - A\mathbf{x}^{(k)}\right), \quad k = 0, 1, 2, \ldots$$

und seine Iterationsmatrix ist

$$G_{\text{dJac}} = I - \omega D^{-1}A.$$

□

**Beispiel 46.10 Gauß–Seidel–Verfahren.** Im Gauß–Seidel[1]–Verfahren ist die invertierbare Matrix $M$ eine Dreiecksmatrix

$$M = D + L,\ N = -U.$$

Auch hier muss man voraussetzen, dass die Hauptdiagonale von $A$ keine Nullen besitzt. Es folgt

$$\mathbf{x}^{(k+1)} = (D + L)^{-1}\left(\mathbf{b} - U\mathbf{x}^{(k)}\right), \quad k = 0, 1, 2, \ldots,$$

so dass die Iterationsmatrix die Gestalt

$$G_{\text{GS}} = -(D + L)^{-1}U = I - (D + L)^{-1}A$$

[1]Philipp Ludwig von Seidel (1821 – 1896)

besitzt. Multipliziert man die Gleichung des Gauß–Seidel–Verfahrens mit $(D+L)$ und ordnet die Terme entsprechend an, so erhält man eine gebräuchlichere Form dieses Verfahrens

$$\begin{aligned}\mathbf{x}^{(k+1)} &= D^{-1}\left(\mathbf{b} - L\mathbf{x}^{(k+1)} - U\mathbf{x}^{(k)}\right)\\ &= \mathbf{x}^{(k)} + D^{-1}\left(\mathbf{b} - L\mathbf{x}^{(k+1)} - (D+U)\,\mathbf{x}^{(k)}\right), \quad k = 0,1,2,\ldots.\end{aligned}$$

Schreibt man diese Iteration in Komponentenschreibweise, so sieht man, dass man die rechte Seite auswerten kann obwohl dort bereits die neue Iterierte auftaucht, da nur die bereits berechneten Werte der neuen Iterierten für die Auswertung benötigt werden

$$x_i^{(k+1)} = x_i^{(k)} + \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i}^{n} a_{ij}x_j^{(k)}\right),\ i = 1,\ldots,n,\ k = 0,1,2,\ldots.$$

□

**Beispiel 46.11 SOR–Verfahren.** Die Matrizen, die das (Vorwärts–) SOR–Verfahren (successive over relaxation, nacheinander folgende Entspannung = Gegenteil von Dämpfung) definieren, sind durch

$$M = \omega^{-1}D + L,\ N = \omega^{-1}D - (D+U),$$

gegeben, wobei $\omega \in \mathbb{R}, \omega > 0$, ist. Dieses Verfahren kann in der Gestalt

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \omega D^{-1}\left(\mathbf{b} - L\mathbf{x}^{(k+1)} - (D+U)\,\mathbf{x}^{(k)}\right), \quad k = 0,1,2,\ldots$$

geschrieben werden. Für $\omega = 1$ erhält man das Gauß–Seidel–Verfahren. Die Iterationsmatrix des SOR–Verfahrens besitzt die Gestalt

$$\begin{aligned}G_{\mathrm{SOR}}(\omega) &= \left(\omega^{-1}D + L\right)^{-1}\left(\omega^{-1}D - (D+U)\right)\\ &= \omega\left(D+\omega L\right)^{-1}\left(\omega^{-1}D - (D+U)\right)\\ &= \left(D+\omega L\right)^{-1}\left((1-\omega)\,D - \omega U\right)\\ &= I - \left(D+\omega L\right)^{-1}\left(\omega U - (1-\omega)D + D + \omega L\right)\\ &= I - \omega(D+\omega L)^{-1}A = I - (\omega^{-1}D + L)^{-1}A.\end{aligned}$$

□

**Beispiel 46.12 SSOR–Verfahren.** Im SOR–Verfahren kann man die Rollen von $L$ und $U$ vertauschen und erhält eine Rückwärts–SOR–Verfahren

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \omega D^{-1}\left(\mathbf{b} - U\mathbf{x}^{(k+1)} - (D+L)\,\mathbf{x}^{(k)}\right), \quad k = 0,1,2,\ldots.$$

Dieses Verfahren berechnet die Komponenten der neuen Iterierten in umgekehrter Reihenfolge wie das SOR–Verfahren. Das Vorwärts– und das Rückwärts–SOR–Verfahren verhalten sich im allgemeinen unterschiedlich. Es gibt Beispiele, in denen das eine effizienter als das andere ist. Man weiß aber im allgemeinen vorher nicht, welche Variante die bessere ist. Das SSOR–Verfahren (symmetric SOR) kombiniert beide Methoden. Ein Iterationsschritt des SSOR–Verfahrens besteht aus zwei Teilschritten, einem Vorwärts–SOR–Schritt und einem Rückwärts–SOR–Schritt

$$\begin{aligned}\mathbf{x}^{(k+1/2)} &= \mathbf{x}^{(k)} + \omega D^{-1}\left(\mathbf{b} - L\mathbf{x}^{(k+1/2)} - (D+U)\,\mathbf{x}^{(k)}\right)\\ \mathbf{x}^{(k+1)} &= \mathbf{x}^{(k+1/2)} + \omega D^{-1}\left(\mathbf{b} - U\mathbf{x}^{(k+1)} - (D+L)\,\mathbf{x}^{(k+1/2)}\right),\end{aligned}$$

$k = 0,1,2,\ldots$. □

## 46.3 Einige Konvergenzaussagen

**Satz 46.13 Konvergenz für stark diagonal–dominante Matrizen.** *Sei $A \in \mathbb{R}^{n\times n}$ eine stark diagonal–dominante Matrix, das heißt*

$$|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}|, \quad \text{für alle } i = 1, \ldots, n.$$

*Dann konvergieren das Jacobi–Verfahren und das Gauß–Seidel–Verfahren für jede Anfangsiterierte $\mathbf{x}^{(0)} \in \mathbb{R}^n$.*

**Beweis:** Gemäß Satz 46.7 hat man zu zeigen, dass der Spektralradius der jeweiligen Iterationsmatrix kleiner als Eins ist.

*Jacobi–Verfahren.* Sei $\mathbf{z} \in \mathbb{R}^n, \mathbf{z} \neq \mathbf{0}$. Dann gilt

$$\begin{aligned} \left|(G_{\text{Jac}}\mathbf{z})_i\right| &= \left|\left(-D^{-1}(L+U)\,\mathbf{z}\right)_i\right| = \left|\frac{1}{a_{ii}} \sum_{j=1, j\neq i}^{n} a_{ij} z_j\right| \leq \frac{1}{|a_{ii}|} \sum_{j=1, j\neq i}^{n} |a_{ij}|\,|z_j| \\ &\leq \frac{1}{|a_{ii}|} \underbrace{\sum_{j=1, j\neq i}^{n} |a_{ij}|}_{<|a_{ii}|} \|\mathbf{z}\|_\infty < \|\mathbf{z}\|_\infty . \end{aligned}$$

Es folgt

$$\|G_{\text{Jac}}\|_\infty = \max_{i=1,\ldots,n} \left|(G_{\text{Jac}}\mathbf{z})_i\right| < \|\mathbf{z}\|_\infty .$$

Nun erhält man mit Lemma 46.5

$$\rho(G_{\text{Jac}}) \leq \|G_{\text{Jac}}\|_\infty = \max_{\mathbf{z}\in\mathbb{R}^n, \mathbf{z}\neq\mathbf{0}} \frac{\|G\mathbf{z}\|_\infty}{\|\mathbf{z}\|_\infty} < 1.$$

*Gauß–Seidel–Verfahren.* Siehe Literatur. ■

Als nächstes wird das gedämpfte Jacobi–Verfahren betrachtet.

**Lemma 46.14 Jacobi– und gedämpftes Jacobi–Verfahren.** *Sei $\omega > 0$. Dann ist $\lambda \in \mathbb{C}$ ein Eigenwert von $G_{\text{Jac}}$ genau dann, wenn $\mu = 1 - \omega + \omega\lambda$ ein Eigenwert von $G_{\text{dJac}}$ ist.*

**Beweis:** Es ist

$$G_{\text{dJac}} = I - \omega D^{-1}A = I - \omega D^{-1}D - \omega \underbrace{D^{-1}(L+U)}_{-G_{\text{Jac}}} = (1-\omega)\,I + \omega G_{\text{Jac}}.$$

Damit folgt für den Eigenwert $\lambda$ von $G_{\text{Jac}}$ und eine zugehörigen Eigenvektor $\mathbf{v}$

$$\begin{aligned} G_{\text{dJac}}\mathbf{v} &= \left[(1-\omega)\,I + \omega G_{\text{Jac}}\right]\mathbf{v} = (1-\omega)\,\mathbf{v} + \omega G_{\text{Jac}}\mathbf{v} = (1-\omega)\,\mathbf{v} + \omega\lambda\mathbf{v} \\ &= (1-\omega+\omega\lambda)\,\mathbf{v}. \end{aligned}$$

■

**Bemerkung 46.15** Das Lemma besagt, dass bei einer geeignete Wahl des Dämpfungsfaktors $\omega$ unter Umständen die Möglichkeit besteht, dass das gedämpfte Jacobi–Verfahren für jede Anfangsiterierte konvergiert währenddessen dies für das Jacobi–Verfahren nicht der Fall ist, Übungsaufgabe. □

Nun betrachten wir das SOR–Verfahren.

**Lemma 46.16 *Lemma von Kahan*[2].** *Falls das SOR–Verfahren für jede Anfangsiterierte $\mathbf{x}^{(0)} \in \mathbb{R}^n$ konvergiert, so ist $\omega \in (0, 2)$.*

[2] William M. Kahan, geboren 1933

**Beweis:** Seien $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ die Eigenwerte von $G_{\mathrm{SOR}}(\omega)$. Es ist

$$\begin{aligned}\prod_{i=1}^{n} \lambda_i &\overset{\text{Bem. 41.7}}{=} \det\left(G_{\mathrm{SOR}}(\omega)\right) = \det\left((D+\omega L)^{-1}\left((1-\omega)D - \omega U\right)\right) \\ &= \det\left(\underbrace{(D+\omega L)^{-1}}_{\text{untere Dreiecksmatrix}}\right) \det\left(\underbrace{\left((1-\omega)D - \omega U\right)}_{\text{obere Dreiecksmatrix}}\right) \\ &= \det\left(D^{-1}\right)(1-\omega)^n \det(D) = (1-\omega)^n .\end{aligned}$$

Somit gilt

$$\prod_{i=1}^{n} |\lambda_i| = |1-\omega|^n .$$

Es gibt also wenigstens einen Eigenwert $\lambda_i$ mit $|\lambda_i| \geq |1-\omega|$, woraus $\rho\left(G_{\mathrm{SOR}}(\omega)\right) \geq |1-\omega|$ folgt. Die Anwendung von Satz 46.7 liefert nun, dass das SOR–Verfahren nicht für alle Anfangsiterierten konvergieren kann falls $\omega \notin (0,2)$. ■

Die Umkehrung des Lemmas von Kahan gilt im Falle, dass $A$ eine symmetrisch, positiv definite Matrix ist.

**Satz 46.17** *Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit. Dann konvergiert das SOR–Verfahren für alle Anfangsiterierten $\mathbf{x}^{(0)} \in \mathbb{R}^n$ falls $\omega \in (0,2)$.*

**Beweis:** Etwas länglich, siehe Literatur. ■

**Bemerkung 46.18** Zur Wahl eines optimalen Parameters $\omega$ im SOR–Verfahren benötigt man Informationen über die Eigenwerte von $A$. Wie man diese erhält, wird im Kapitel 47 behandelt. □

**Bemerkung 46.19** Bei klassischen Iterationsverfahren hat man im wesentlichen Matrix–Vektor–Produkte zu berechnen. Man wird sie vor allem anwenden, wenn diese Produkte billig sind, das heißt wenn viele Matrixeinträge Null sind (schwach besetzte Matrizen). Man stellt allerdings fest, dass die Anzahl der Iterationen sich proportional zur Konditionszahl

$$\kappa_*(A) = \|A\|_* \left\|A^{-1}\right\|_*$$

verhält. In vielen Anwendungen erhält man aber Matrizen mit hoher Konditionszahl. Wenn man versucht, in diesen Anwendungen genauere Ergebnisse zu berechnen, erhöht man die Anzahl der Unbekannten. Man stellt dann fest, dass sich die Konditionszahl der System quadratisch mit der Anzahl der Unbekannten erhöht. Das bedeutet, die Anzahl der benötigten Iterationen erhöht sich auch quadratisch. Man erhält sehr ineffiziente Verfahren. Deshalb ist der Einsatzbereich klassischer Iterationsverfahren ziemlich beschränkt. Es gibt jedoch wesentlich bessere iterative Verfahren zur Lösung linearer Gleichungssysteme (Spezialvorlesung). □

# Kapitel 47

# Numerische Berechnung von Eigenwerten und Eigenvektoren

**Bemerkung 47.1 Eigenwertberechnung über das charakteristische Polynom.** Die Verwendung des charakteristischen Polynoms zur Berechnung der Eigenwerte von $A \in \mathbb{R}^{n\times n}$ besitzt in der Praxis entscheidende Nachteile. Zuerst müssen die Koeffizienten des Polynoms berechnet werden. Das ist für große $n$ aufwändig. Desweiteren hängen die Nullstellen oft sensibel von den Koeffizienten des charakteristischen Polynoms ab. Das Problem ist also schlecht konditioniert, das heißt kleine Fehler in den Daten führen zu großen Fehlern im Ergebnis. Kleine Fehler in den Daten hat man immer, da reelle Zahlen gerundet werden müssen, bevor sie vom Computer dargestellt werden können. Insgesamt ist das charakteristische Polynom zur numerischen Berechnung von Eigenwerten einer Matrix nicht brauchbar. □

## 47.1 Die Potenzmethode

**Bemerkung 47.2 Potenzmethode, Vektoriteration.** Die Potenzmethode oder Vektoriteration ist ein Verfahren zur Berechnung des betragsgrößten Eigenwertes und eines zugehörigen Eigenvektors einer Matrix $A$. Dieses Verfahren geht auf von Mises[1] zurück. Es liefert das Grundkonzept für die Entwicklung weiterer Verfahren zur Berechnung von Eigenwerten und –vektoren. □

**Bemerkung 47.3 Grundidee.** Der Einfachheit halber werden für die Konvergenzanalyse einige Annahmen gemacht. Die Matrix $A \in \mathbb{R}^{n\times n}$ sei diagonalisierbar, das heißt es existiert eine Basis aus Eigenvektoren $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{C}^n$ mit $\|\mathbf{v}_i\|_2 = 1$, $i = 1, \ldots, n$. Weiter gelte für die Eigenwerte

$$|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|,$$

das heißt, der betragsgrößte Eigenwert soll einfach sein. Da $\lambda_1$ damit eine einfache Nullstelle des charakteristischen Polynoms ist, welches nur reelle Koeffizienten besitzt, gilt damit $\lambda_1 \in \mathbb{R}$. Damit besitzt die Eigenwertgleichung eine nichttriviale reelle Lösung, also $\mathbf{v}_1 \in \mathbb{R}^n$.

[1] Richard von Mises (1883 – 1953)

Für die Potenzmethode benötigt man einen Startvektor $\mathbf{x}^{(0)} \in \mathbb{R}^n$. Dieser lässt sich theoretisch als Linearkombination der Eigenvektoren darstellen

$$\mathbf{x}^{(0)} = \sum_{j=1}^{n} c_j \mathbf{v}_j.$$

Sei $\mathbf{x}^{(0)}$ so gewählt, dass $c_1 \neq 0$ gilt. Multipliziert man die Darstellung von $\mathbf{x}^{(0)}$ mit der $k$–ten Potenz $A^k$ von, so erhält man mit Satz 41.12

$$A^k \mathbf{x}^{(0)} = \sum_{j=1}^{n} c_j A^k \mathbf{v}_j = \sum_{j=1}^{n} c_j \lambda_j^k \mathbf{v}_j.$$

Damit gilt

$$\mathbf{x}^{(k)} := A^k \mathbf{x}^{(0)} = \lambda_1^k \left( c_1 \mathbf{v}_1 + \sum_{j=2}^{n} c_j \left( \frac{\lambda_j}{\lambda_1} \right)^k \mathbf{v}_j \right) =: \lambda_1^k \left( c_1 \mathbf{v}_1 + \mathbf{r}^{(k)} \right). \tag{47.1}$$

Wegen $|\lambda_j / \lambda_1| < 1$ folgt

$$\lim_{k \to \infty} \mathbf{r}^{(k)} = \lim_{k \to \infty} \sum_{j=2}^{n} c_j \left( \frac{\lambda_j}{\lambda_1} \right)^k \mathbf{v}_j = \mathbf{0}.$$

Das bedeutet, für große $k$ dominiert in (47.1) der Beitrag vom ersten Eigenwert und Eigenvektor. □

**Satz 47.4** *Sei $A \in \mathbb{R}^{n \times n}$ und erfülle $A$ die Voraussetzungen aus Bemerkung 47.3. Sei $\mathbf{x}^{(k)} \in \mathbb{R}^n$ die $k$–te Iterierte der Potenzmethode und sei*

$$\lambda^{(k)} = \frac{\left(\mathbf{x}^{(k)}\right)^T A \mathbf{x}^{(k)}}{\left\|\mathbf{x}^{(k)}\right\|_2^2} = \frac{\left(\mathbf{x}^{(k)}\right)^T \mathbf{x}^{(k+1)}}{\left\|\mathbf{x}^{(k)}\right\|_2^2}.$$

*Dann gilt*

$$\left| \lambda_1 - \lambda^{(k)} \right| = \mathcal{O}\left( \left| \frac{\lambda_2}{\lambda_1} \right|^k \right).$$

**Beweis:** Man betrachtet den Abstand zwischen dem Unterraum $S^{(k)} := \{\alpha \mathbf{x}^{(k)} : \alpha \in \mathbb{R}\}$ und dem Eigenvektor $\mathbf{v}_1$

$$d\left(S^{(k)}, \mathbf{v}_1\right) := \min_{\mathbf{x} \in S^{(k)}} \|\mathbf{x} - \mathbf{v}_1\|_2 = \min_{\alpha \in \mathbb{R}} \left\| \alpha \mathbf{x}^{(k)} - \mathbf{v}_1 \right\|_2.$$

Nun formt man (47.1) äquivalent um

$$\alpha_k \mathbf{x}^{(k)} := \left( \lambda_1^k c_1 \right)^{-1} \mathbf{x}^{(k)} = \mathbf{v}_1 + c_1^{-1} \mathbf{r}^{(k)}. \tag{47.2}$$

Wählt man in $d\left(S^{(k)}, \mathbf{v}_1\right)$ den Wert $\alpha = \alpha_k$, so erhält man damit und der Definition von $\mathbf{r}^{(k)}$

$$d\left(S^{(k)}, \mathbf{v}_1\right) \leq \left\| \alpha_k \mathbf{x}^{(k)} - \mathbf{v}_1 \right\|_2 = \left| c_1^{-1} \right| \left\| \mathbf{r}^{(k)} \right\|_2 = \mathcal{O}\left( \left| \frac{\lambda_2}{\lambda_1} \right|^k \right). \tag{47.3}$$

Damit ist $\alpha_k \mathbf{x}^{(k)}$ eine Approximation an $\mathbf{v}_1$, also

$$A \alpha_k \mathbf{x}^{(k)} \approx \lambda_1 \alpha_k \mathbf{x}^{(k)} \quad \Longrightarrow \quad A \mathbf{x}^{(k)} \approx \lambda_1 \mathbf{x}^{(k)}.$$

Durch Multiplikation von links mit $\left(\mathbf{x}^{(k)}\right)^T$ folgt, dass

$$\lambda^{(k)} = \frac{\left(\mathbf{x}^{(k)}\right)^T A\mathbf{x}^{(k)}}{\|\mathbf{x}^{(k)}\|_2^2} = \frac{\left(\mathbf{x}^{(k)}\right)^T \mathbf{x}^{(k+1)}}{\|\mathbf{x}^{(k)}\|_2^2}$$

eine Approximation an $\lambda_1$ ist. Analog ist $\alpha_{k+1}\mathbf{x}^{(k+1)}$ mit $\alpha_{k+1} = \left(\lambda_1^{k+1}c_1\right)^{-1} = \alpha_k/\lambda_1$ eine Approximation von $\mathbf{v}_1$. Man erhält mit (47.2), (47.3) und $\|\mathbf{v}_1\|_2 = 1$

$$\begin{aligned}
\lambda^{(k)} &= \frac{\left(\alpha_k\mathbf{x}^{(k)}\right)^T \left(\alpha_k\mathbf{x}^{(k+1)}\right)}{\|\alpha_k\mathbf{x}^{(k)}\|_2^2} = \lambda_1 \frac{\left(\alpha_k\mathbf{x}^{(k)}\right)^T \left(\alpha_{k+1}\mathbf{x}^{(k+1)}\right)}{\|\alpha_k\mathbf{x}^{(k)}\|_2^2} \\
&= \lambda_1 \frac{\left(\mathbf{v}_1 + c_1^{-1}\mathbf{r}^{(k)}\right)^T \left(\mathbf{v}_1 + c_1^{-1}\mathbf{r}^{(k+1)}\right)}{\left\|\mathbf{v}_1 + c_1^{-1}\mathbf{r}^{(k)}\right\|_2^2} \\
&= \lambda_1 \frac{\|\mathbf{v}_1\|_2^2 + c_1^{-1}\left(\mathbf{r}^{(k)}\right)^T \mathbf{v}_1 + c_1^{-1}\mathbf{v}_1^T\mathbf{r}^{(k+1)} + c_1^{-2}\left(\mathbf{r}^{(k)}\right)^T \mathbf{r}^{(k+1)}}{\|\mathbf{v}_1\|_2^2 + 2c_1^{-1}\left(\mathbf{r}^{(k)}\right)^T \mathbf{v}_1 + c_1^{-2}\|\mathbf{r}^{(k)}\|_2^2} \\
&= \lambda_1 \frac{1 + \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)}{1 + \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)} = \lambda_1\left(1 + \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)\right).
\end{aligned}$$

Die Gültigkeit des letzten Schrittes sieht man aus

$$\begin{aligned}
\left(1 + \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)\right)\left(1 + \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)\right) &= 1 + \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right) + \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{2k}\right) \\
&= 1 + \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right),
\end{aligned}$$

vergleiche Definition des Landau–Symbols im Kapitel 15. Durch Umstellen folgt

$$\left|\lambda_1 - \lambda^{(k)}\right| = \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right).$$

■

**Bemerkung 47.5 Symmetrische Matrizen.** Falls $A$ eine symmetrische Matrix ist, kann man sogar

$$\left|\lambda_1 - \lambda^{(k)}\right| = \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{2k}\right).$$

zeigen. □

**Bemerkung 47.6 Skalierung der Iterierten.** Wendet man das bisherige Verfahren an, so gelten

$$\begin{aligned}
\left\|\mathbf{x}^{(k)}\right\|_2 &\to \infty \quad \text{falls} \quad |\lambda_1| > 1, \\
\left\|\mathbf{x}^{(k)}\right\|_2 &\to 0 \quad \text{falls} \quad |\lambda_1| < 1.
\end{aligned}$$

Aus diesen Gründen ist es zweckmäßig, die Iterierten zu skalieren. Damit werden starke Änderungen in der Größenordnung vermieden. Die Konvergenzaussagen ändern sich durch Skalierung auch nicht, da weder der Unterraum $S^{(k)}$ noch die Iterierte $\lambda^{(k)}$ von einer Skalierung von $\mathbf{x}^{(k)}$ abhängen. □

**Algorithmus 47.7 Potenzmethode, Vektoriteration.** Seien $A \in \mathbb{R}^{n\times n}$ und $\mathbf{y}^{(0)} \neq \mathbf{0}$ mit $\left\|\mathbf{y}^{(0)}\right\|_2 = 1$ gegeben. Für $k = 0, 1, \ldots$ berechne

$$\begin{aligned} \tilde{\mathbf{y}}^{(k+1)} &= A\mathbf{y}^{(k)} \\ \lambda^{(k)} &= \left(\tilde{\mathbf{y}}^{(k+1)}\right)^T \mathbf{y}^{(k)} \\ \mathbf{y}^{(k+1)} &= \frac{\tilde{\mathbf{y}}^{(k+1)}}{\left\|\tilde{\mathbf{y}}^{(k+1)}\right\|_2}. \end{aligned}$$

□

**Bemerkung 47.8** Wählt man $\mathbf{y}^{(0)} = \mathbf{x}^{(0)}$, so weist man mit vollständiger Induktion nach, dass

$$\mathbf{y}^{(k)} = \frac{\mathbf{x}^{(k)}}{\left\|\mathbf{x}^{(k)}\right\|_2} = \frac{A^k\mathbf{x}^{(0)}}{\left\|A^k\mathbf{x}^{(0)}\right\|_2}.$$

Also liefert Algorithmus 47.7 bis auf Skalierung in $\mathbf{x}^{(k)}$ die oben analysierten Folgen $\{\mathbf{x}^{(k)}\}$ und $\{\lambda^{(k)}\}$.

Die Konvergenzgeschwindigkeit der Potenzmethode hängt wesentlich vom Verhältnis von $|\lambda_1|$ und $|\lambda_2|$ ab. □

## 47.2 Das Jacobi–Verfahren

Das Jacobi–Verfahren ist ein einfaches und robustes Verfahren zur Bestimmung aller Eigenwerte und Eigenvektoren einer symmetrischen Matrix. Es basiert auf folgendem Lemma.

**Lemma 47.9** *Seien $A \in \mathbb{R}^{n\times n}$ und $Q \in \mathbb{R}^{n\times n}$ eine Orthogonalmatrix, dann besitzen $A$ und $Q^TAQ$ dieselben Eigenwerte.*

**Beweis:** Es wird gezeigt, dass $A$ und $Q^TAQ$ dasselbe charakteristische Polynom besitzen:

$$\begin{aligned} \det(Q^TAQ - \lambda I) &= \det(Q^TAQ - \lambda Q^TQ)) \\ &= \det(Q^T(A - \lambda I)Q) \\ &= \det(Q^T)\det(A - \lambda I)\det(Q) \\ &= \det(\underbrace{Q^TQ}_{I})\det(A - \lambda I) \\ &= \det(A - \lambda I). \end{aligned}$$

Also haben $A$ und $Q^TAQ$ diesselben Eigenwerte. ■

**Bemerkung 47.10 Grundidee des Jacobi–Verfahrens.** Mit Hilfe einer Sequenz $\{Q_k\}_{k=1,2\ldots}$ von orthogonalen Matrizen transformiert man $A$ auf Diagonalgestalt. Die Diagonalelemente geben die Eigenwerte an, und aus $\{Q_k\}_{k=1,2\ldots}$ berechnet man die Eigenvektoren. Da alle Matrizen reelle Einträge besitzen, kann das nur funktionieren, wenn alle Eigenwerte reell sind, da man bei den Matrizenmultiplikationen immer Matrizen mit reellen Einträgen als Ergebnis erhält. Diese Eigenschaft ist für symmetrische Matrizen erfüllt, siehe Satz 42.1.

Als orthogonale Transformationen verwendet man sogenannte Givens[2]–Matrizen. Das sind Rotationsmatrizen vom Typ

$$Q_k = \begin{pmatrix} 1 & & & & & & & & & 0 & \\ & \ddots & & & & & & & & & \\ & & 1 & & & & & & & & \\ & & & \cos\varphi_k & \dots & \dots & \dots & \sin\varphi_k & & & \leftarrow i \\ & & & & 1 & & & & & & \\ & & & \vdots & & \ddots & & \vdots & & & \\ & & & & & & 1 & & & & \\ & & & -\sin\varphi_k & \dots & \dots & \dots & \cos\varphi_k & & & \leftarrow j \\ & & & & & & & & 1 & & \\ & & & & & & & & & \ddots & \\ 0 & & & & & & & & & 1 & \end{pmatrix}$$

Man rechnet leicht nach, dass $Q_k^T A Q_k$ nur die $i$–ten und $j$–ten Zeilen und Spalten von $A$ verändert.

Den Rotationswinkel $\varphi_k$ wählt man so, dass $a_{ij}$ und $a_{ji}$ zum Verschwinden gebracht werden. Führt man dies iterativ für alle Nichtdiagonalelemente von $A$ durch, kann man zeigen, dass das Verfahren gegen eine Diagonalmatrix konvergiert

$$Q_k^T \dots Q_2^T Q_1^T A Q_1 Q_2 \dots Q_k \to \operatorname{diag}(\lambda_1, \dots, \lambda_n) \quad \text{für} \quad k \to \infty.$$

Die $m$–te Spalte von $P := Q_1 Q_2 \dots Q_k$ enthält somit eine Approximation an den Eigenvektor $\mathbf{v}_m$ zum Eigenwert $\lambda_m$. Da $P$ orthogonal ist, erhält man insbesondere auch im Fall von mehrfachen Eigenwerten ein vollständiges Orthonormalsystem von Eigenvektoren. Eine Gram–Schmidt–Orthonormierung ist daher nicht erforderlich.

In einem Zyklus des Jacobi–Verfahrens wird jedes Nichtdiagonalelement einmal auf Null transformiert. Dabei wird es im allgemeinen passieren, dass bereits vorhandene Nullen wieder verschwinden. Typischerweise erhält man nach 6–8 Zyklen brauchbare Approximationen für die Eigenwerte und Eigenvektoren. Die Kosten pro Zyklus betragen etwa $32n^3$ Multiplikationen. Das ist recht teuer.

Mehr Details zum Jacobi–Verfahren findet man in der Literatur. □

**Bemerkung 47.11 Fazit.** Die Lösung von Eigenwertproblemen ist numerisch aufwändig. In der Praxis werden wesentlich kompliziertere numerische Verfahren verwendet, als die hier vorgestellten. □

[2] James Wallace Givens, Jr. (1910 – 1993)

# Teil V

# Mehrdimensionale Analysis

In diesem Teil wird die Differential– und Integralrechnung von Funktionen mehrerer Veränderlicher betrachtet. Dazu müssen die Konzepte aus Teil III erweitert und verallgemeinert werden. Funktionen mehrerer Veränderlicher sind die natürlichen Funktionen für die Beschreibung von Naturprozessen und industriellen Prozessen.

# Kapitel 48

# Stetigkeit vektorwertiger Funktionen

**Bemerkung 48.1 Motivation.** Vektorwertige Funktionen mehrerer Variabler treten in der Informatik zum Beispiel bei der Verarbeitung von Farbbildern auf. Kontinuierliche Farbbilder lassen sich als Funktionen $\mathbf{f} : D \to \mathbb{R}^3$ auffassen. Dabei ist $D$ ein rechteckiger Bildbereich, und der Wertebereich beschreibt die drei Kanäle rot, grün, blau.

Einer vektorwertige Funktion $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$ entsprechen genau $m$ skalare Funktionen $f_i : D \subset \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m$,

$$\mathbf{y} = \mathbf{f}(\mathbf{x}) = \mathbf{f}(x_1, \ldots, x_n) \quad \Longleftrightarrow \quad \begin{array}{ccccc} y_1 & = & f_1(\mathbf{x}) & = & f_1(x_1, \ldots, x_n) \\ \vdots & & \vdots & \vdots & \\ y_m & = & f_m(\mathbf{x}) & = & f_m(x_1, \ldots, x_n) \end{array} .$$

□

**Beispiel 48.2** Eine vektorwertige Funktion von $\mathbb{R}^2$ nach $\mathbb{R}^3$ ist beispielsweise gegeben durch $\mathbf{f} : \mathbb{R}^2 \to \mathbb{R}^3$ mit

$$\mathbf{f} : \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} x^2 y \sin(xy) \\ x + \cos y \\ e^{xy^2} \end{pmatrix}.$$

□

**Definition 48.3 Stetigkeit vektorwertiger Funktionen.** Seien $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$ und $\boldsymbol{\xi} \in D$, $D$ offen. Dann heißt $\mathbf{f}(\mathbf{x})$ im Punkte $\boldsymbol{\xi}$ stetig, wenn zu jedem $\varepsilon > 0$ ein $\delta(\varepsilon) > 0$ existiert mit

$$\max_{i \in \{1,\ldots,n\}} |x_i - \xi_i| < \delta \quad \Longrightarrow \quad \max_{j \in \{1,\ldots,m\}} |f_j(\mathbf{x}) - f_j(\boldsymbol{\xi})| < \varepsilon.$$

□

**Bemerkung 48.4 Grenzwertdefinition.** Genauso wie für skalare Funktionen einer reellen Veränderlichen läßt sich die Stetigkeit auch mit Hilfe von Grenzwerten definieren. Eine Funktion $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$, $D$ offen, ist für $\boldsymbol{\xi} \in D$ stetig, wenn für jede Folge $\{\mathbf{x}_n\}_{n \in \mathbb{N}}$, $\mathbf{x}_n \in D$ für alle $n \in \mathbb{N}$, mit

$$\lim_{n \to \infty} \mathbf{x}_n = \boldsymbol{\xi} \quad \Longleftrightarrow \quad \lim_{n \to \infty} \max_{i \in \{1,\ldots,n\}} |x_{n,i} - \xi_i| = 0$$

gilt

$$\lim_{n\to\infty} \mathbf{f}(\mathbf{x}_n) = \mathbf{f}(\boldsymbol{\xi}) \quad \Longleftrightarrow \quad \lim_{n\to\infty} \max_{j\in\{1,\ldots,m\}} |f_j(\mathbf{x}_n) - f_j(\boldsymbol{\xi})| = 0.$$

Man kann genauso gut die Euklidische Norm nehmen. Im Unterschied zum Fall $n = 1$, bei welchem sich die Argumente $x_n$ nur auf einer Geraden dem Punkt $\xi$ nähern, nämlich auf der $x$–Achse, können sich im Fall $n > 1$ die Punkte irgendwie nähern. (Bild) □

Die Stetigkeit vektorwertiger Funktionen lässt sich direkt durch die Stetigkeit der einzelnen Komponenten untersuchen.

**Satz 48.5** *Die vektorwertige Funktion* $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ *ist im Punkt* $\boldsymbol{\xi} \in \mathbb{R}^n$ *genau dann stetig, wenn jede Komponente* $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m$, *in* $\boldsymbol{\xi}$ *stetig ist.*

**Beweis:** Sei zunächst $\mathbf{f}(\mathbf{x})$ in $\boldsymbol{\xi}$ stetig. Dann gilt für $\max_{i\in\{1,\ldots,n\}} |x_i - \xi_i| < \delta$

$$\varepsilon > \max_{j\in\{1,\ldots,m\}} |f_j(\mathbf{x}) - f_j(\boldsymbol{\xi})| > |f_k(\mathbf{x}) - f_k(\boldsymbol{\xi})|$$

für jede Komponente $f_k(\mathbf{x})$. Damit ist jede Komponente stetig.

Sei umgekehrt jede Komponente $f_k(\mathbf{x})$, $k = 1, \ldots, m$, eine stetige Funktion. Das heißt, zu jedem $\varepsilon > 0$ gibt es ein $\delta_k(\varepsilon) > 0$, so dass

$$\max_{i\in\{1,\ldots,n\}} |x_i - \xi_i| < \delta_k(\varepsilon) \quad \Longrightarrow \quad |f_k(\mathbf{x}) - f_k(\boldsymbol{\xi})| < \varepsilon.$$

Setzt man $\delta(\varepsilon) := \min_{k\in\{1,\ldots,m\}} \delta_k(\varepsilon)$, so gilt

$$\max_{i\in\{1,\ldots,n\}} |x_i - \xi_i| < \delta(\varepsilon) \quad \Longrightarrow \quad \max_{k\in\{1,\ldots,m\}} |f_k(\mathbf{x}) - f_k(\boldsymbol{\xi})| < \varepsilon.$$

Das ist die Stetigkeit von $\mathbf{f}(\mathbf{x})$ in $\boldsymbol{\xi}$. ■

**Beispiel 48.6** Die Funktion aus Beispiel 48.2 ist für alle $(x, y) \in \mathbb{R}^2$ stetig, weil alle drei Komponenten stetige Funktionen sind. □

**Bemerkung 48.7 Rechenregeln für Grenzwerte.** Viele Rechenregeln für Grenzwerte übertragen sich von skalaren auf vektorwertige Funktionen. Seien zum Beispiel $\mathbf{f}, \mathbf{g} : D \subset \mathbb{R}^n \to \mathbb{R}^m$, $\alpha, \beta \in \mathbb{R}$, $\boldsymbol{\xi} \in D$, dann gilt

$$\lim_{\mathbf{x}\to\boldsymbol{\xi}, \mathbf{x}\in D} (\alpha\mathbf{f}(\mathbf{x}) + \beta\mathbf{g}(\mathbf{x})) = \alpha \lim_{\mathbf{x}\to\boldsymbol{\xi}, \mathbf{x}\in D} \mathbf{f}(\mathbf{x}) + \beta \lim_{\mathbf{x}\to\boldsymbol{\xi}, \mathbf{x}\in D} \mathbf{g}(\mathbf{x}).$$

□

# Kapitel 49

# Differenzierbarkeit

## 49.1 Definition und grundlegende Eigenschaften

**Bemerkung 49.1 Motivation.** Auch die Differenzierbarkeit von Funktionen mehrerer Veränderlicher lässt sich aus einer Eigenschaft differenzierbarer reellwertiger Funktionen einer reellen Veränderlichen motivieren. Wenn eine reellwertige Funktion einer reellen Veränderlichen im Punkt $\xi$ ihres Definitionsbereiches differenzierbar ist, kann man in diesem Punkt eine Tangente anlegen. Die Tangente ist eine Gerade, also kann man die Funktion in $\xi$ durch ein lineares Polynom (Taylor–Polynom 1. Grades, Satz 22.2) approximieren. Auch bei der Definition der Differentiation von vektorwertigen Funktionen mehrerer Veränderlicher besteht die Grundidee darin, dass man die Funktion in einer Umgebung eines Punktes durch ein lineares Polynom approximieren kann. □

**Definition 49.2 Differenzierbarkeit vektorwertiger Funktionen.** Seien $D \subset \mathbb{R}^n$ offen und $\mathbf{f} : D \to \mathbb{R}^m$. Die Funktion $\mathbf{f}$ heißt im Punkt $\boldsymbol{\xi} \in D$ differenzierbar, wenn es eine lineare Abbildung $M : \mathbb{R}^n \to \mathbb{R}^m$ und eine in $\boldsymbol{\xi}$ stetige Funktion $\mathbf{r} : D \to \mathbb{R}^m$ gibt, so dass

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(\boldsymbol{\xi}) + M(\mathbf{x} - \boldsymbol{\xi}) + \mathbf{r}(\mathbf{x}) \|\mathbf{x} - \boldsymbol{\xi}\|_2, \quad \mathbf{x} \in D, \tag{49.1}$$

und $\mathbf{r}(\boldsymbol{\xi}) = \mathbf{0}$ gelten. Man nennt $M$ die erste Ableitung von $\mathbf{f}(\mathbf{x})$ in $\boldsymbol{\xi}$, im Zeichen $\mathbf{f}'(\boldsymbol{\xi}) := M$. □

**Bemerkung 49.3 Komponentenschreibweise.** Aus der linearen Algebra ist bekannt, dass sich lineare Abbildungen $M : \mathbb{R}^n \to \mathbb{R}^m$ durch $m \times n$–Matrizen und die Wirkung von $M$ auf den Vektor $\mathbf{x} - \boldsymbol{\xi}$ als Matrix–Vektor–Produkt schreiben lassen. In Komponentenschreibweise kann (49.1) daher mit $M = (m_{ij})_{i=1,\ldots,m;j=1,\ldots,n}$ in der Form

$$f_i(\mathbf{x}) = f_i(\boldsymbol{\xi}) + \sum_{j=1}^{n} m_{ij}(x_j - \xi_j) + r_i(\mathbf{x}) \|\mathbf{x} - \boldsymbol{\xi}\|_2, \ \mathbf{x} \in D, \ i = 1, \ldots, m, \tag{49.2}$$

geschrieben werden. □

**Bemerkung 49.4 Eindeutigkeit von $M$ und $\mathbf{r}(\mathbf{x})$.** Es muss geklärt werden, ob $M$ und $\mathbf{r}$ in (49.1) eindeutig bestimmt sind. Seien $\mathbf{e}_j$ der $j$–te kartesische Einheitsvektor und $t \in \mathbb{R}$ aus einer hinreichend kleinen Umgebung von Null, so dass $\mathbf{x} = \boldsymbol{\xi} + t\mathbf{e}_j$ in $D$ liegt. Dann folgt aus (49.2)

$$m_{ij} = \frac{f_i(\mathbf{x}) - f_i(\boldsymbol{\xi}) - r_i(\mathbf{x}) \|\mathbf{x} - \boldsymbol{\xi}\|_2}{t}.$$

Wegen

$$\|\mathbf{x}-\boldsymbol{\xi}\|_2 = t\,\|\mathbf{e}_j\|_2 = |t|\,,$$

der Stetigkeit von $\mathbf{r}(\mathbf{x})$ in $\boldsymbol{\xi}$ und $\mathbf{r}(\boldsymbol{\xi})=\mathbf{0}$ gilt

$$\begin{aligned} m_{ij} &= \lim_{t\to 0}\frac{f_i(\mathbf{x})-f_i(\boldsymbol{\xi})-r_i(\mathbf{x})\,\|\mathbf{x}-\boldsymbol{\xi}\|_2}{t} = \lim_{t\to 0}\frac{f_i(\mathbf{x})-f_i(\boldsymbol{\xi})}{t} - \lim_{t\to 0}\frac{|t|\,r_i(\mathbf{x})}{t} \\ &= \lim_{t\to 0}\frac{f_i(\boldsymbol{\xi}+t\mathbf{e}_j)-f_i(\boldsymbol{\xi})}{t} =: \frac{\partial f_i}{\partial x_j}(\boldsymbol{\xi}). \end{aligned}$$

Die Abbildung $M$ ist also eindeutig bestimmt. Sie wird Jacobi–Matrix genannt, Schreibweise:

$$M = \mathbf{f}'(\boldsymbol{\xi}) = J\mathbf{f}(\boldsymbol{\xi}).$$

Aus (49.1) folgt somit auch, dass $\mathbf{r}(\mathbf{x})$ in eindeutiger Weise gegeben ist

$$\mathbf{r}(\mathbf{x}) = \begin{cases} \dfrac{\mathbf{f}(\mathbf{x})-\mathbf{f}(\boldsymbol{\xi})-M\,(\mathbf{x}-\boldsymbol{\xi})}{\|\mathbf{x}-\boldsymbol{\xi}\|_2} & \mathbf{x}\neq\boldsymbol{\xi} \\ \mathbf{0} & \mathbf{x}=\boldsymbol{\xi}. \end{cases}$$

Die Differenzierbarkeitsbedingung (49.1) kann nun komponentenweise wie folgt geschrieben werden

$$\begin{pmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} f_1(\boldsymbol{\xi}) \\ \vdots \\ f_m(\boldsymbol{\xi}) \end{pmatrix} + \begin{pmatrix} \dfrac{\partial f_1}{\partial x_1}(\boldsymbol{\xi}) & \cdots & \dfrac{\partial f_1}{\partial x_n}(\boldsymbol{\xi}) \\ \vdots & & \vdots \\ \dfrac{\partial f_m}{\partial x_1}(\boldsymbol{\xi}) & \cdots & \dfrac{\partial f_m}{\partial x_n}(\boldsymbol{\xi}) \end{pmatrix} \begin{pmatrix} x_1-\xi_1 \\ \vdots \\ x_n-\xi_n \end{pmatrix} + \begin{pmatrix} r_1(\mathbf{x}) \\ \vdots \\ r_m(\mathbf{x}) \end{pmatrix} \|\mathbf{x}-\boldsymbol{\xi}\|_2\,.$$

□

**Satz 49.5** *Ist die Funktion* $\mathbf{f} : D\subset\mathbb{R}^n\to\mathbb{R}^m$ *in* $\boldsymbol{\xi}\in D$ *differenzierbar, so ist* $\mathbf{f}(\mathbf{x})$ *in* $\boldsymbol{\xi}$ *stetig.*

**Beweis:** Aus der Differenzierbarkeit von $\mathbf{f}(\mathbf{x})$ in $\boldsymbol{\xi}$ folgt mit der Dreiecksungleichung

$$\|\mathbf{f}(\mathbf{x})-\mathbf{f}(\boldsymbol{\xi})\|_2 \le \|M\,(\mathbf{x}-\boldsymbol{\xi})\|_2 + \|\mathbf{r}(\mathbf{x})\|_2\,\|\mathbf{x}-\boldsymbol{\xi}\|_2\,.$$

Mit der Cauchy–Schwarzschen Ungleichung (37.1) erhält man

$$\begin{aligned} \|M\,(\mathbf{x}-\boldsymbol{\xi})\|_2^2 &= \sum_{i=1}^m\left(\sum_{j=1}^n m_{ij}(x_j-\xi_j)\right)^2 \\ &\le \sum_{i=1}^m\left[\left(\sum_{j=1}^n m_{ij}^2\right)^{1/2}\left(\sum_{j=1}^n (x_j-\xi_j)^2\right)^{1/2}\right]^2 \\ &= \left(\sum_{i=1}^m\sum_{j=1}^n m_{ij}^2\right)\left(\sum_{j=1}^n (x_j-\xi_j)^2\right) = \|M\|_F^2\,\|\mathbf{x}-\boldsymbol{\xi}\|_2^2\,. \end{aligned}$$

Somit gibt es eine Konstante $K>0$, so dass

$$\|M\,(\mathbf{x}-\boldsymbol{\xi})\|_2 \le K\,\|\mathbf{x}-\boldsymbol{\xi}\|_2$$

gilt. Nun ist $\mathbf{r}(\mathbf{x})$ stetig in $\boldsymbol{\xi}$ mit $\mathbf{r}(\boldsymbol{\xi})=\mathbf{0}$. Man findet also zu jedem $\varepsilon>0$ ein $\delta(\varepsilon)>0$ derart, dass aus $\|\mathbf{x}-\boldsymbol{\xi}\|_2<\delta(\varepsilon)$ die Ungleichung $\|\mathbf{r}(\mathbf{x})\|_2<\varepsilon$ folgt. Setzt man

$$\delta^*(\varepsilon) := \min\left\{\delta(\varepsilon), \frac{\varepsilon}{K+\varepsilon}\right\},$$

so gilt für alle $\mathbf{x}$ mit $\|\mathbf{x}-\boldsymbol{\xi}\|_2 < \delta^*(\varepsilon)$ die Beziehung

$$\|\mathbf{f}(\mathbf{x})-\mathbf{f}(\boldsymbol{\xi})\|_2 < (K+\varepsilon)\,\|\mathbf{x}-\boldsymbol{\xi}\|_2 < (K+\varepsilon)\,\delta^*(\varepsilon) \le \varepsilon.$$

Das heißt, $\mathbf{f}(\mathbf{x})$ ist im Punkt $\boldsymbol{\xi}$ stetig. ■

Wir haben in Satz 48.5 gesehen, dass die Stetigkeit vektorwertiger Funktionen $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$ im Punkte $\boldsymbol{\xi} \in D$ äquivalent zur Stetigkeit aller Komponenten $f_i : D \subset \mathbb{R}^n \to \mathbb{R}$ im Punkte $\boldsymbol{\xi} \in D$ ist. Eine ähnliche Aussage gilt auch für die Differenzierbarkeit.

**Satz 49.6** *Die Funktion $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$ ist im Punkte $\boldsymbol{\xi} \in D$ genau dann differenzierbar, wenn jede ihrer Komponenten $f_i : D \subset \mathbb{R}^n \to \mathbb{R}$ im Punkte $\boldsymbol{\xi} \in D$ differenzierbar ist.*

**Beweis:** Der Beweis erfolgt direkt mit Hilfe der Definition der Differenzierbarkeit, aus Zeitgründen siehe Literatur. ■

**Beispiel 49.7** Die Funktion

$$\mathbf{f} : \mathbb{R}^2 \to \mathbb{R}^3, \quad \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} x_1^2+x_2^2 \\ x_1 \\ x_2 \end{pmatrix}$$

ist für alle $\mathbf{x} \in \mathbb{R}^2$ differenzierbar. Es gilt

$$J\mathbf{f}(\mathbf{x}) = \begin{pmatrix} 2x_1 & 2x_2 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Zum Beweis dieser Behauptung betrachtet man die Definitionsgleichung (49.1) mit beliebigem $\mathbf{x}, \boldsymbol{\xi} \in \mathbb{R}^2$. Ziel ist es, eine Funktion $\mathbf{r}(\mathbf{x})$ zu finden, die die Bedingungen der Definition 49.2 erfüllt. Es ist

$$\begin{aligned}
\mathbf{r}(\mathbf{x})\,\|\mathbf{x}-\boldsymbol{\xi}\|_2 &= \begin{pmatrix} x_1^2+x_2^2 \\ x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} \xi_1^2+\xi_2^2 \\ \xi_1 \\ \xi_2 \end{pmatrix} - \begin{pmatrix} 2\xi_1 & 2\xi_2 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1-\xi_1 \\ x_2-\xi_2 \end{pmatrix} \\
&= \begin{pmatrix} x_1^2+x_2^2-\xi_1^2-\xi_2^2-2\xi_1x_1+2\xi_1^2-2\xi_2x_2+2\xi_2^2 \\ 0 \\ 0 \end{pmatrix} \\
&= \begin{pmatrix} (x_1-\xi_1)^2+(x_2-\xi_2)^2 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \|\mathbf{x}-\boldsymbol{\xi}\|_2^2 \\ 0 \\ 0 \end{pmatrix}.
\end{aligned}$$

Daraus folgt

$$\mathbf{r}(\mathbf{x}) = \begin{pmatrix} \|\mathbf{x}-\boldsymbol{\xi}\|_2 \\ 0 \\ 0 \end{pmatrix}.$$

Diese Funktion ist für alle $\mathbf{x} \in \mathbb{R}^2$ definiert, sie ist in $\mathbb{R}^2$ stetig und für $\mathbf{x} \to \boldsymbol{\xi}$ folgt $\mathbf{r}(\boldsymbol{\xi}) = \mathbf{0}$. Damit sind alle Bedingungen von Definition 49.2 erfüllt. □

## 49.2 Richtungsableitung und Gradient

**Bemerkung 49.8 Motivation.** In Definition 21.2 wurde die Ableitung einer skalaren Funktion einer Variablen in einem inneren Punkt $\xi$ ihres Definitionsbereiches eingeführt

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

Dieser Ableitungsbegriff ist auch anwendbar bei vektorwertigen Funktionen einer Variablen. Für

$$\mathbf{f}(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{pmatrix}$$

definiert man die Ableitung in einem inneren Punkt $\xi$ der Definitionsbereiche aller Komponenten

$$\mathbf{f}'(\xi) := \begin{pmatrix} f_1'(\xi) \\ \vdots \\ f_m'(\xi) \end{pmatrix}.$$

In diesem Abschnitt wird das Konzept der Ableitung in eine Dimension (eine Richtung) auf skalare Funktionen mehrerer Variablen $f : D \subset \mathbb{R}^n \to \mathbb{R}$ verallgemeinert. □

**Definition 49.9 Richtungsableitung, partielle Ableitungen.** Seien $D \subset \mathbb{R}^n$ offen, $f : D \to \mathbb{R}$ und $\boldsymbol{\xi} \in D$. Für einen Vektor $\mathbf{v} \in \mathbb{R}^n$ mit $\|\mathbf{v}\|_2 = 1$ heißt

$$\frac{\partial f}{\partial \mathbf{v}}(\boldsymbol{\xi}) = \lim_{h \to 0} \frac{f(\boldsymbol{\xi} + h\mathbf{v}) - f(\boldsymbol{\xi})}{h}$$

die Richtungsableitung (Gateaux[1]–Ableitung) von $f(\mathbf{x})$ in $\boldsymbol{\xi}$ in Richtung $\mathbf{v}$.

Die Ableitungen in Richtung der kartesischen Einheitsvektoren $\mathbf{e}_1, \ldots, \mathbf{e}_n$ nennt man partielle Ableitungen von $f(\mathbf{x})$ nach $x_i$ im Punkt $\boldsymbol{\xi}$. Man schreibt

$$f_{x_i}(\boldsymbol{\xi}) := \frac{\partial f}{\partial x_i}(\boldsymbol{\xi}) := \lim_{h \to 0} \frac{f(\boldsymbol{\xi} + h\mathbf{e}_i) - f(\boldsymbol{\xi})}{h}.$$

□

**Beispiel 49.10** Betrachte

$$f : \mathbb{R}^2 \to \mathbb{R} \quad \text{mit} \quad f(x_1, x_2) = x_1^3 \cos x_2.$$

Es lässt sich mit Hilfe der Definition der partiellen Ableitungen leicht nachrechnen, dass

$$\frac{\partial f}{\partial x_1}(x_1, x_2) = 3x_1^2 \cos x_2, \quad \frac{\partial f}{\partial x_2}(x_1, x_2) = -x_1^3 \sin x_2$$

gelten. Das bedeutet, man erhält die partielle Ableitung nach $x_i$ indem man wie gewohnt nach $x_i$ differenziert und alle anderen Variablen $x_j$, $j \neq i$, dabei als Konstante betrachtet. □

**Bemerkung 49.11 Zur Jacobi–Matrix.** Die $i$–te Zeile der Jacobi–Matrix $J\mathbf{f}(\boldsymbol{\xi})$ enthält gerade die partiellen Ableitungen der $i$–ten Komponente $f_i(\mathbf{x})$ im Punkt $\boldsymbol{\xi}$. □

**Satz 49.12 Existenz und Darstellung der Richtungsableitungen.** *Sei $f : D \subset \mathbb{R}^n \to \mathbb{R}$ in $\boldsymbol{\xi} \in D$ differenzierbar. Dann existiert in $\boldsymbol{\xi}$ die Richtungsableitung in jede Richtung $\mathbf{v}$ und es gilt*

$$\frac{\partial f}{\partial \mathbf{v}}(\boldsymbol{\xi}) = Jf(\boldsymbol{\xi}) \cdot \mathbf{v} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\boldsymbol{\xi}) v_i.$$

[1] René Eugène Gateaux (1889 – 1914)

**Beweis:** Man setzt in die Definition der Differenzierbarkeit $\mathbf{x} = \boldsymbol{\xi} + h\mathbf{v}$. Durch Umstellen erhält man unter Nutzung der Linearität von $Jf(\boldsymbol{\xi})$

$$\frac{f(\boldsymbol{\xi} + h\mathbf{v}) - f(\boldsymbol{\xi})}{h} = Jf(\boldsymbol{\xi}) \cdot \mathbf{v} + r(\boldsymbol{\xi} + h\mathbf{v}) \underbrace{\|\mathbf{v}\|_2}_{=1}.$$

Für $h \to 0$ verschwindet der letzte Term, da $r(\mathbf{x})$ stetig ist und $r(\boldsymbol{\xi}) = 0$ ist. ■

**Beispiel 49.13** Gesucht ist die Richtungsableitung von

$$f(x, y) = e^{-x} \sin y$$

in Richtung

$$\mathbf{v} = \begin{pmatrix} 3/5 \\ 4/5 \end{pmatrix}$$

im Punkt $(0, \pi/6)$.

Es gelten $\|\mathbf{v}\|_2 = 1$,

$$\begin{aligned} f_x(x, y) &= -e^{-x} \sin y \quad \Longrightarrow \quad f_x\left(0, \frac{\pi}{6}\right) = (-1)\frac{1}{2} = -\frac{1}{2}, \\ f_y(x, y) &= e^{-x} \cos y \quad \Longrightarrow \quad f_y\left(0, \frac{\pi}{6}\right) = (1)\frac{1}{2}\sqrt{3} = \frac{1}{2}\sqrt{3}. \end{aligned}$$

Daraus folgt

$$\frac{\partial f}{\partial \mathbf{v}}\left(0, \frac{\pi}{6}\right) = \mathbf{v}^T \nabla f\left(0, \frac{\pi}{6}\right) = \frac{3}{5}\left(-\frac{1}{2}\right) + \frac{4}{5}\frac{1}{2}\sqrt{3} = -\frac{3}{10} + \frac{2}{5}\sqrt{3}.$$

□

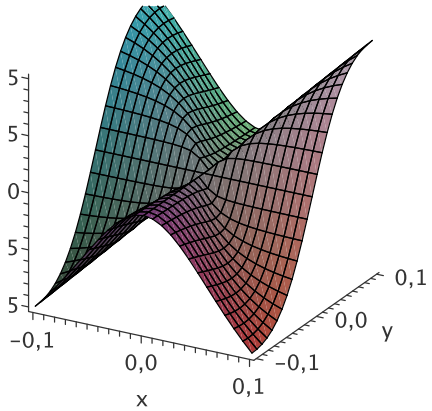
**Beispiel 49.14 Existenz aller Richtungsableitungen reicht nicht für Differenzierbarkeit.** Die Umkehrung des eben bewiesenen Satzes 49.12 gilt nicht. Betrachte dafür die Funktion $f : \mathbb{R}^2 \to \mathbb{R}$ mit

$$f(x, y) = \begin{cases} \dfrac{x^2 y}{x^2 + y^2} & \text{für } (x, y) \neq (0, 0), \\ 0 & \text{für } (x, y) = (0, 0). \end{cases}$$

Betrachtet man den Koordinatenursprung, dann hat ein beliebiger Einheitsvektor dort die Gestalt $\mathbf{v} = (v_1, v_2)$ mit $v_1^2 + v_2^2 = 1$. Man erhält

$$\frac{\partial f}{\partial \mathbf{v}}(0, 0) = \lim_{h \to 0} \frac{h^3 v_1^2 v_2 - 0}{h(h^2(v_1^2 + v_2^2))} = \lim_{h \to 0} \frac{h^3 v_1^2 v_2}{h^3} = v_1^2 v_2.$$

Diese Richtungsableitung ist keine lineare Funktion von $\mathbf{v}$, daher kann $f(x, y)$ in $(x, y) = (0, 0)$ nicht differenzierbar sein.




□

**Definition 49.15 Gradient, Nabla–Operator.** Seien $D \subset \mathbb{R}^n$ offen und $f : D \to \mathbb{R}$ in $\boldsymbol{\xi} \in D$ nach allen $x_i$ partiell differenzierbar. Dann nennt man den Vektor der ersten partiellen Ableitungen

$$\operatorname{grad} f(\boldsymbol{\xi}) := \nabla f(\boldsymbol{\xi}) := \begin{pmatrix} \dfrac{\partial f}{\partial x_1}(\boldsymbol{\xi}) \\ \vdots \\ \dfrac{\partial f}{\partial x_n}(\boldsymbol{\xi}) \end{pmatrix}$$

den Gradienten von $f(\mathbf{x})$ in $\boldsymbol{\xi}$. Manche Bücher definieren $\operatorname{grad} f(\mathbf{x})$ als Zeilenvektor und $\nabla f(\mathbf{x})$ als Spaltenvektor.

Den vektorwertigen Operator

$$\nabla := \begin{pmatrix} \dfrac{\partial}{\partial x_1} \\ \vdots \\ \dfrac{\partial}{\partial x_n} \end{pmatrix}$$

nennt man Nabla–Operator (nach der Form eines ägyptischen Musikinstruments). □

**Folgerung 49.16 Existenz des Gradienten und Darstellung der Richtungsableitungen für skalare Funktionen.** *Sei $f : D \subset \mathbb{R}^n \to \mathbb{R}$ in $\boldsymbol{\xi} \in D$ differenzierbar. Dann existiert der Gradient von $f(\mathbf{x})$ in $\boldsymbol{\xi}$ und für die Richtungsableitung von $f(\mathbf{x})$ in $\boldsymbol{\xi}$ in Richtung $\mathbf{v}$ gilt*

$$\frac{\partial f}{\partial \mathbf{v}}(\boldsymbol{\xi}) = \nabla f(\boldsymbol{\xi}) \cdot \mathbf{v}.$$

**Beispiel 49.17 Aus Existenz des Gradienten folgt nicht Existenz aller Richtungsableitungen.** Aus der Existenz des Gradienten folgt nicht notwendig die Existenz der Richtungsableitung in einer von den Koordinatenrichtungen verschiedenen Richtung. Betrachte dazu die Funktion

$$f(x, y) = \begin{cases} 0 & \text{für } xy = 0 \\ 1 & \text{sonst.} \end{cases}$$

Für diese Funktion existiert der Gradient im Punkt $(0, 0)$. In allen anderen Richtungen als den Koordinatenrichtungen ist die Funktion im Punkt $(0, 0)$ jedoch unstetig, also existiert in diese Richtungen keine Richtungsableitung. □

**Bemerkung 49.18 Bedeutung des Gradienten für skalare Funktionen.** Sei $f : D \subset \mathbb{R}^n \to \mathbb{R}$ eine skalare Funktion.

1. Für $\nabla f(\boldsymbol{\xi}) \neq \mathbf{0}$ nimmt die Richtungsableitung ihren größten Wert für

$$\mathbf{v} = \frac{\nabla f(\boldsymbol{\xi})}{\|\nabla f(\boldsymbol{\xi})\|_2}$$

an. Das folgt aus der Beziehung

$$\mathbf{v}^T \nabla f(\boldsymbol{\xi}) = \cos(\mathbf{v}, \nabla f(\boldsymbol{\xi})) \|\mathbf{v}\|_2 \|\nabla f(\boldsymbol{\xi})\|_2 = \cos(\mathbf{v}, \nabla f(\boldsymbol{\xi})) \|\nabla f(\boldsymbol{\xi})\|_2,$$

da der Kosinus am größten wird, wenn $\mathbf{v}$ in Richtung $\nabla f(\boldsymbol{\xi})$ zeigt. Dann gilt

$$\frac{\partial f}{\partial \mathbf{v}}(\boldsymbol{\xi}) = \mathbf{v}^T \nabla f(\boldsymbol{\xi}) = \frac{(\nabla f(\boldsymbol{\xi}))^T}{\|\nabla f(\boldsymbol{\xi})\|_2} \nabla f(\boldsymbol{\xi}) = \frac{\|\nabla f(\boldsymbol{\xi})\|_2^2}{\|\nabla f(\boldsymbol{\xi})\|_2} = \|\nabla f(\boldsymbol{\xi})\|_2.$$

Der Gradient zeigt in Richtung des steilsten Anstiegs.

2. In der Gegenrichtung $\mathbf{v} = -\dfrac{\nabla f(\boldsymbol{\xi})}{\|\nabla f(\boldsymbol{\xi})\|_2}$ gilt

$$\frac{\partial f}{\partial \mathbf{v}}(\boldsymbol{\xi}) = -\|\nabla f(\boldsymbol{\xi})\|_2 .$$

3. Wählt man eine Richtung $\mathbf{v}$ die orthogonal zu $\nabla f(\boldsymbol{\xi})$ ist, erhält man

$$\frac{\partial f}{\partial \mathbf{v}}(\boldsymbol{\xi}) = \frac{(\nabla f^{\perp}(\boldsymbol{\xi}))^T}{|\nabla f^{\perp}(\boldsymbol{\xi})|}\nabla f(\boldsymbol{\xi}) = 0.$$

Somit ist die Steigung Null, wenn man eine Richtung orthogonal zum Gradienten betrachtet. Damit steht der Gradient $\nabla f(\boldsymbol{\xi})$ in $\boldsymbol{\xi}$ senkrecht zu den sogenannten Höhenlinien $\{\mathbf{x} \in D \mid f(\mathbf{x}) = f(\boldsymbol{\xi}) = const\}$.

□

**Satz 49.19 Hinreichendes Kriterium für die Differenzierbarkeit.** *Existieren für $f : D \subset \mathbb{R}^n \to \mathbb{R}$ in einer Umgebung $B(\boldsymbol{\xi}, \rho)$ (Kugel mit Mittelpunkt $\boldsymbol{\xi}$ und Radius $\rho$) von $\boldsymbol{\xi}$ alle partiellen Ableitungen*

$$\frac{\partial f}{\partial x_i} \; : \; B(\boldsymbol{\xi}, \rho) \subset \mathbb{R}^n \to \mathbb{R},$$

*und seien diese in $B(\boldsymbol{\xi}, \rho)$ stetig. Dann ist $f(\mathbf{x})$ in $\boldsymbol{\xi}$ differenzierbar.*

**Beweis:** Ohne Beschränkung der Allgemeinheit beschränken wir uns auf den Fall $n = 2$. Mit Hilfe des ersten Mittelwertsatzes der Differentialrechnung, Satz 21.22, gilt

$$\begin{aligned} f(x_1, x_2) - f(\xi_1, \xi_2) &= f(x_1, x_2) - f(\xi_1, x_2) + f(\xi_1, x_2) - f(\xi_1, \xi_2) \\ &= \frac{\partial f}{\partial x_1}(\xi_1 + \theta_1(x_1 - \xi_1), x_2)\,(x_1 - \xi_1) \\ &\quad + \frac{\partial f}{\partial x_2}(\xi_1, \xi_2 + \theta_2(x_2 - \xi_2))\,(x_2 - \xi_2) \\ &= \sum_{i=1}^{2} \frac{\partial f}{\partial x_i}(\xi_1, \xi_2)(x_i - \xi_i) + R, \end{aligned}$$

mit $\theta_1, \theta_2 \in (0, 1)$ und

$$\begin{aligned} R &= \left(\frac{\partial f}{\partial x_1}(\xi_1 + \theta_1(x_1 - \xi_1), x_2) - \frac{\partial f}{\partial x_1}(\xi_1, \xi_2)\right)(x_1 - \xi_1) \\ &\quad + \left(\frac{\partial f}{\partial x_2}(\xi_1, \xi_2 + \theta_2(x_2 - \xi_2)) - \frac{\partial f}{\partial x_2}(\xi_1, \xi_2)\right)(x_2 - \xi_2). \end{aligned}$$

Da die partiellen Ableitungen in $(\xi_1, \xi_2)$ stetig sind, gibt es zu jedem $\varepsilon > 0$ ein $\delta(\varepsilon)$, so dass für $\|\mathbf{x} - \boldsymbol{\xi}\|_2 < \delta(\varepsilon)$

$$|R| < \varepsilon \|\mathbf{x} - \boldsymbol{\xi}\|_2$$

gilt. Dies zeigt, dass $R$ in der Form $R = r(\mathbf{x}) \|\mathbf{x} - \boldsymbol{\xi}\|_2$ geschrieben werden kann und $r(\mathbf{x})$ in $B(\boldsymbol{\xi}, \rho)$ stetig ist mit $r(\boldsymbol{\xi}) = 0$. ■

**Bemerkung 49.20**

- Die Funktion aus Beispiel 49.17 besitzt keine stetigen partiellen Ableitungen auf den Koordinatenachsen, mit Ausnahme des Koordinatenursprungs.
- Das Kriterium von Satz 49.19 ist nicht notwendig. Es gibt Funktionen in mehreren Dimensionen, die an einer Stelle differenzierbar sind, deren partielle Ableitungen aber in der Umgebung dieser Stelle nicht sind sind, siehe Heuser, Analysis II, S. 266, Aufg. 7.

□

## 49.3 Eigenschaften differenzierbarer Funktionen

Die vertrauten Differentiationsregeln übertragen sich ohne Änderung, wenn auch äußerlich etwas umständlicher, auf vektorwertige Funktionen von mehreren Veränderlichen.

**Satz 49.21 Linearität der Differentiation.** *Seien* $\mathbf{f}, \mathbf{g} : D \subset \mathbb{R}^n \to \mathbb{R}^m$ *zwei vektorwertige Abbildungen, die im Punkt* $\boldsymbol{\xi} \in D$ *differenzierbar sind. Dann ist mit beliebigem* $\alpha, \beta \in \mathbb{R}$ *die Linearkombination* $(\alpha\mathbf{f} + \beta\mathbf{g})(\mathbf{x})$ *in* $\boldsymbol{\xi}$ *differenzierbar und es gilt*

$$(\alpha\mathbf{f} + \beta\mathbf{g})'(\boldsymbol{\xi}) = \alpha\mathbf{f}'(\boldsymbol{\xi}) + \beta\mathbf{g}'(\boldsymbol{\xi}).$$

**Beweis:** Der Beweis erfolg direkt mit der Definition der Differenzierbarkeit, Übungsaufgabe. ■

**Bemerkung 49.22 Linearität der partiellen Ableitungen.** Da die Jacobi–Matrix $\mathbf{f}'(\boldsymbol{\xi})$ die Matrix aller partiellen Ableitungen erster Ordnung ist, sind die partiellen Ableitungen auch linear

$$\frac{\partial}{\partial x_i}(\alpha f_j(\mathbf{x}) + \beta g_j(\mathbf{x})) = \alpha\frac{\partial f_j}{\partial x_i}(\mathbf{x}) + \beta\frac{\partial g_j}{\partial x_i}(\mathbf{x}) \quad \forall\ \alpha, \beta \in \mathbb{R}.$$

□

**Satz 49.23 Kettenregel.** *Seien* $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{f} : \mathbb{R}^m \to \mathbb{R}^p$ *und* $\boldsymbol{\phi}(\mathbf{x}) = (\mathbf{f} \circ \mathbf{g})(\mathbf{x}) = \mathbf{f}(\mathbf{g}(\mathbf{x}))$ *die Verkettung. Sind* $\mathbf{g}(\mathbf{x})$ *in* $\boldsymbol{\xi} \in \mathbb{R}^n$ *und* $\mathbf{f}(\mathbf{y})$ *in* $\boldsymbol{\eta} = \mathbf{g}(\boldsymbol{\xi})$ *differenzierbar, so ist auch die Verkettung* $\boldsymbol{\phi}(\mathbf{x})$ *in* $\boldsymbol{\xi}$ *differenzierbar und es gilt*

$$\boldsymbol{\phi}'(\boldsymbol{\xi}) = \mathbf{f}'(\mathbf{g}(\boldsymbol{\xi})) \circ \mathbf{g}'(\boldsymbol{\xi})$$

*beziehungsweise*

$$\frac{\partial \phi_i}{\partial x_j}(\boldsymbol{\xi}) = \frac{\partial f_i\left(g_1(\xi_1, \ldots, \xi_n), \ldots, g_m(\xi_1, \ldots, \xi_n)\right)}{\partial x_j}(\boldsymbol{\xi}) = \sum_{k=1}^{m} \frac{\partial f_i}{\partial y_k}\left(\mathbf{g}(\boldsymbol{\xi})\right)\frac{\partial g_k}{\partial x_j}(\boldsymbol{\xi}).$$

**Beweis:** Der Beweis erfolgt analog zum Fall $m = n = p = 1$, siehe Satz 21.7. ■

**Bemerkung 49.24 Kettenregel in Matrixschreibweise.** Die Verkettung linearer Abbildungen kann als Matrizenprodukt geschrieben werden, siehe Bemerkung 33.10. Interpretiert man die Ableitungen als entsprechende rechteckige Matrizen, so entspricht der Verkettung von $\mathbf{f}'(\mathbf{g}(\boldsymbol{\xi}))$ mit $\mathbf{g}'(\boldsymbol{\xi})$ also dem Produkt dieser Matrizen. In Komponentenschreibweise gilt daher

$$\begin{pmatrix} \frac{\partial \phi_1}{\partial x_1}(\boldsymbol{\xi}) & \cdots & \frac{\partial \phi_1}{\partial x_n}(\boldsymbol{\xi}) \\ \vdots & & \vdots \\ \frac{\partial \phi_p}{\partial x_1}(\boldsymbol{\xi}) & \cdots & \frac{\partial \phi_p}{\partial x_n}(\boldsymbol{\xi}) \end{pmatrix}$$
$$= \begin{pmatrix} \frac{\partial f_1}{\partial y_1}(\mathbf{g}(\boldsymbol{\xi})) & \cdots & \frac{\partial f_1}{\partial y_m}(\mathbf{g}(\boldsymbol{\xi})) \\ \vdots & & \vdots \\ \frac{\partial f_p}{\partial y_1}(\mathbf{g}(\boldsymbol{\xi})) & \cdots & \frac{\partial f_p}{\partial y_m}(\mathbf{g}(\boldsymbol{\xi})) \end{pmatrix} \begin{pmatrix} \frac{\partial g_1}{\partial x_1}(\boldsymbol{\xi}) & \cdots & \frac{\partial g_1}{\partial x_n}(\boldsymbol{\xi}) \\ \vdots & & \vdots \\ \frac{\partial g_m}{\partial x_1}(\boldsymbol{\xi}) & \cdots & \frac{\partial g_m}{\partial x_n}(\boldsymbol{\xi}) \end{pmatrix}.$$

Hat man insbesondere eine Funktion $\phi(s,t) = f(x(s,t), y(s,t)) \;:\; D \subset \mathbb{R}^2 \to \mathbb{R}$ gegeben, so erhält man

$$\begin{pmatrix} \frac{\partial \phi}{\partial s}(s_0,t_0) & \frac{\partial \phi}{\partial t}(s_0,t_0) \end{pmatrix} = \begin{pmatrix} \frac{\partial f}{\partial x}(x_0,y_0) & \frac{\partial f}{\partial y}(x_0,y_0) \end{pmatrix} \begin{pmatrix} \frac{\partial x}{\partial s}(s_0,t_0) & \frac{\partial x}{\partial t}(s_0,t_0) \\ \frac{\partial y}{\partial s}(s_0,t_0) & \frac{\partial y}{\partial t}(s_0,t_0) \end{pmatrix}$$

oder ausgeschrieben

$$\begin{aligned} \frac{\partial \phi}{\partial s}(s_0,t_0) &= \frac{\partial f}{\partial x}(x_0,y_0)\frac{\partial x}{\partial s}(s_0,t_0) + \frac{\partial f}{\partial y}(x_0,y_0)\frac{\partial y}{\partial s}(s_0,t_0), \\ \frac{\partial \phi}{\partial t}(s_0,t_0) &= \frac{\partial f}{\partial x}(x_0,y_0)\frac{\partial x}{\partial t}(s_0,t_0) + \frac{\partial f}{\partial y}(x_0,y_0)\frac{\partial y}{\partial t}(s_0,t_0) \end{aligned} \tag{49.3}$$

mit $(x_0,y_0) = (x(s_0,t_0), y(s_0,t_0))$. □

**Beispiel 49.25** Betrachte die Abbildungen

$$\begin{aligned} \mathbf{g} &: \mathbb{R}^2 \to \mathbb{R}^2 \quad \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} x_1^5 - 3x_1^2x_2^5 + x_2^2 - 7 \\ x_1 + x_2^2 \end{pmatrix} \\ f &: \mathbb{R}^2 \to \mathbb{R} \quad \begin{pmatrix} g_1 \\ g_2 \end{pmatrix} \mapsto g_1^{1700} + g_2^{28} \end{aligned}$$

sowie ihre Komposition

$$\varphi(x_1,x_2) = f(\mathbf{g}(x_1,x_2)) = f\left(g_1(x_1,x_2), g_2(x_1,x_2)\right).$$

Für die partielle Ableitung nach $x_1$ erhält man nach Kettenregel

$$\begin{aligned} \frac{\partial \varphi}{\partial x_1}(x_1,x_2) &= \left(\frac{\partial f}{\partial g_1}\frac{\partial g_1}{\partial x_1} + \frac{\partial f}{\partial g_2}\frac{\partial g_2}{\partial x_1}\right)(x_1,x_2) \\ &= \left(1700 g_1^{1699}\frac{\partial g_1}{\partial x_1} + 28 g_2^{27}\frac{\partial g_2}{\partial x_1}\right)(x_1,x_2) \\ &= 1700 g_1^{1699}\left(5x_1^4 - 6x_1x_2^5\right) + 28 g_2^{27}\cdot 1 \\ &= 1700\left(x_1^5 - 3x_1^2x_2^5 + x_2^2 - 7\right)^{1699}\left(5x_1^4 - 6x_1x_2^5\right) + 28\left(x_1 + x_2^2\right)^{27}. \end{aligned}$$

□

## 49.4 Der Mittelwertsatz

**Beispiel 49.26 Einfache Übertragung des Mittelwertsatzes aus einer Dimension gilt nicht.** Der Mittelwertsatz in einer Dimension, Satz 21.22, besagt, dass man für eine in einem Intervall $[a,b]$ differenzierbare Funktion $f \;:\; [a,b] \to \mathbb{R}$ einen Wert $\xi \in (a,b)$ findet, so dass

$$f(b) - f(a) = f'(\xi)(b-a)$$

gilt. Solche eine Aussage gilt für vektorwertige Funktionen nicht. Wir betrachten dazu die Funktion

$$\mathbf{f} \;:\; [0,2\pi] \quad x \mapsto \begin{pmatrix} \cos x \\ \sin x \end{pmatrix}.$$

Dann gilt für alle $\xi \in (0,2\pi)$

$$\mathbf{f}(b) - \mathbf{f}(a) = \begin{pmatrix} 1 \\ 0 \end{pmatrix} - \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \neq 2\pi \begin{pmatrix} -\sin\xi \\ \cos\xi \end{pmatrix} = \mathbf{f}'(\xi)(b-a).$$

□

Man kann aber für skalare Funktionen mehrerer Veränderlicher einen Mittelwertsatz formulieren.

**Definition 49.27 Konvexe Menge.** Eine Menge $D \subset \mathbb{R}^n$ heißt konvex, wenn mit zwei beliebigen Punkten $\mathbf{a}, \mathbf{b} \in D$ immer die Verbindungsstrecke zwischen diesen Punkten, das heißt die Menge

$$\{\boldsymbol{\xi} \mid \boldsymbol{\xi} = \mathbf{a} + \theta(\mathbf{b} - \mathbf{a}), \theta \in (0,1)\}$$

in $D$ liegt. □

**Satz 49.28 Mittelwertsatz für skalare Funktionen mehrerer Veränderlicher.** *Seien $\mathbf{a}, \mathbf{b} \in D$ zwei Punkte einer konvexen, offenen Menge $D$ und $f : D \subset \mathbb{R}^n \to \mathbb{R}$ auf $D$ differenzierbar. Dann gibt es auf dem $\mathbf{a}$ und $\mathbf{b}$ verbindenden Segment einen Punkt $\boldsymbol{\xi} = \mathbf{a} + \theta(\mathbf{b} - \mathbf{a})$, $\theta \in (0,1)$, so dass*

$$f(\mathbf{b}) - f(\mathbf{a}) = f'(\boldsymbol{\xi}) \cdot (\mathbf{b} - \mathbf{a})$$

*gilt.*

**Beweis:** Der Beweis erfolgt durch Anwendung der Kettenregel. Man betrachtet $f(\mathbf{x})$ auf dem Segment von $\mathbf{a}$ nach $\mathbf{b}$. Die Kurve über diesem Segment kann mit einer skalarwertigen Funktion einer reellen Veränderlichen identifiziert werden

$$\phi : [0,1] \to \mathbb{R} \quad t \mapsto f(\mathbf{a} + t(\mathbf{b} - \mathbf{a})).$$

Diese Funktion genügt auf $[0,1]$ den Voraussetzungen des Mittelwertsatzes 21.22. Damit gibt es ein $\theta \in (0,1)$ mit

$$f(\mathbf{b}) - f(\mathbf{a}) = \phi(1) - \phi(0) = \phi'(\theta) \cdot 1. \tag{49.4}$$

Andererseits kann man auf $\phi(t)$ die Kettenregel (49.3) anwenden, womit man

$$\phi'(t) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{a}+t(\mathbf{b}-\mathbf{a}))\frac{\partial x_i}{\partial t} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{a}+t(\mathbf{b}-\mathbf{a}))(b_i-a_i) = f'(\mathbf{a}+t(\mathbf{b}-\mathbf{a}))\cdot(\mathbf{b}-\mathbf{a}).$$

erhält. Für $t = \theta$ erhält man in (49.4) gerade die Behauptung des Satzes. ■

**Beispiel 49.29** Sei

$$f(x,y) = \cos x + \sin y, \quad \mathbf{a} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \ \mathbf{b} = \frac{1}{2}\begin{pmatrix} \pi \\ \pi \end{pmatrix}.$$

Es gilt

$$f(0,0) = 1 = f\left(\frac{\pi}{2}, \frac{\pi}{2}\right).$$

Nach dem Mittelwertsatz existiert also ein $\theta \in (0,1)$ mit

$$\begin{aligned} 0 &= f\left(\frac{\pi}{2}, \frac{\pi}{2}\right) - f(0,0) = f'\left(\frac{\theta}{2}\begin{pmatrix} \pi \\ \pi \end{pmatrix}\right) \cdot \frac{1}{2}\begin{pmatrix} \pi \\ \pi \end{pmatrix} \\ &= \frac{1}{2}\left(-\sin\left(\frac{\theta\pi}{2}\right), \cos\left(\frac{\theta\pi}{2}\right)\right)\begin{pmatrix} \pi \\ \pi \end{pmatrix} \\ &= \frac{\pi}{2}\left(\cos\left(\frac{\theta\pi}{2}\right) - \sin\left(\frac{\theta\pi}{2}\right)\right). \end{aligned}$$

In der Tat ist diese Gleichung für $\theta = 1/2$ erfüllt. □

## 49.5 Höhere partielle Ableitungen

Nun wird der Begriff der partiellen Ableitung auf partielle Ableitungen höherer Ordnung verallgemeinert.

**Definition 49.30 Partielle Ableitungen höherer Ordnung.** Seien $D \subset \mathbb{R}^n$ offen und $f : D \to \mathbb{R}$. Ist $f(\mathbf{x})$ partiell differenzierbar auf $D$ und sind die partiellen Ableitungen selbst wieder differenzierbar, erhält man als partielle Ableitungen zweiter Ordnung

$$\frac{\partial^2 f}{\partial x_j \partial x_i}(\boldsymbol{\xi}) := \frac{\partial}{\partial x_j}\left(\frac{\partial f}{\partial x_i}(\boldsymbol{\xi})\right).$$

Induktiv definiert man die partiellen Ableitungen $k$–ter Ordnung durch

$$\frac{\partial^k f}{\partial x_{i_k} \partial x_{i_{k-1}} \dots \partial x_{i_1}}(\boldsymbol{\xi}) := \frac{\partial}{\partial x_{i_k}}\left(\frac{\partial^{k-1} f}{\partial x_{i_{k-1}} \dots \partial x_{i_1}}(\boldsymbol{\xi})\right) \quad \text{für } k \geq 2,$$

$i_1, \dots, i_k \in \{1, \dots n\}$. □

Man kann zeigen, dass unter gewissen Voraussetzungen die Reihenfolge, in welcher man partielle Ableitungen höherer Ordnung berechnet, keine Rolle spielt.

**Satz 49.31 Vertauschbarkeitssatz von Schwarz.** *Seien $D \subset \mathbb{R}^n$ offen, $\boldsymbol{\xi} \in D$ und $f : D \to \mathbb{R}$. Falls die partiellen Ableitungen von $f(\mathbf{x})$ bis zur zweiten Ordnung in einer Umgebung $B(\boldsymbol{\xi}, \rho) \subset D$ stetig sind, so sind sie vertauschbar*

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{\xi}) = \frac{\partial^2 f}{\partial x_j \partial x_i}(\boldsymbol{\xi}) \quad \forall i, j \in \{1, \dots, n\}.$$

**Beweis:** Der Beweis erfolgt durch Anwendung des Mittelwertsatzes, für Details siehe Literatur. ■

**Folgerung 49.32** *Sind die partiellen Ableitungen von $f(\mathbf{x})$ bis zur Ordnung $k$ stetig, so kann man die Reihenfolge der partiellen Ableitungen von $f(\mathbf{x})$ bis zur Ordnung $k$ beliebig vertauschen.*

**Bemerkung 49.33 Multiindexschreibweise.** Hat man eine Funktion, bei welcher die Reihenfolge der partiellen Ableitungen beliebig gewählt werden kann, verwendet man häufig eine abkürzende Schreibweise unter Verwendung von Multiindizes. Ein Multiindex ist ein $n$–Tupel $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n)$ nichtnegativer ganzer Zahlen $\alpha_i \geq 0$ mit $|\boldsymbol{\alpha}| = \sum_{i=1}^n \alpha_i$. Man schreibt dann

$$D^{\boldsymbol{\alpha}} f(\mathbf{x}) := \frac{\partial^{|\boldsymbol{\alpha}|} f}{\partial x_1^{\alpha_1} \dots \partial x_n^{\alpha_n}}(\mathbf{x}) = f_{\alpha_1 \alpha_2, \dots, \alpha_n}(\mathbf{x}).$$

□

**Beispiel 49.34** Für die Funktion

$$f(x, y, z) = z^2 \sin(x^3) + (\cos y \sin x - e^{-x}) z^2$$

soll $f_{111}(x, y, z) =: f_{xyz}(x, y, z)$ berechnet werden. Es ist zunächst klar, dass von dieser Funktion die partiellen Ableitungen beliebiger Ordnung existieren und stetig sind. Hier bietet es sich an, die Funktion zuerst nach $y$ partiell zu integrieren, weil dann einige Terme sofort verschwinden

$$\begin{aligned} f_{xyz}(x, y, z) &= \frac{\partial^2}{\partial x \partial z}(f_y) = \frac{\partial^2}{\partial x \partial z}\left(-(\sin y \sin x) z^2\right) = \frac{\partial}{\partial x}(-2z \sin y \sin x) \\ &= -2z \sin y \cos x. \end{aligned}$$

□

**Definition 49.35 Hesse[2]–Matrix.** Seien $D \subset \mathbb{R}^n$ offen und $f : D \to \mathbb{R}$ sei im Punkt $\boldsymbol{\xi} \in D$ zweimal partiell differenzierbar. Dann nennt man die Matrix der partiellen zweiten Ableitungen

$$Hf(\boldsymbol{\xi}) := \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(\boldsymbol{\xi}) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(\boldsymbol{\xi}) & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\boldsymbol{\xi}) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(\boldsymbol{\xi}) & \frac{\partial^2 f}{\partial x_2^2}(\boldsymbol{\xi}) & \dots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(\boldsymbol{\xi}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\boldsymbol{\xi}) & \frac{\partial^2 f}{\partial x_n \partial x_2}(\boldsymbol{\xi}) & \dots & \frac{\partial^2 f}{\partial x_n^2}(\boldsymbol{\xi}) \end{pmatrix}$$

die Hesse–Matrix von $f(\mathbf{x})$ in $\boldsymbol{\xi}$. □

**Bemerkung 49.36**

1. Für eine Funktion mit stetigen partiellen Ableitungen zweiter Ordnung ist die Hesse–Matrix nach dem Vertauschbarkeitssatz von Schwarz symmetrisch.
2. Die Hesse–Matrix spielt eine große Rolle bei der Klassifikation von Extrema einer Funktion mehrerer Variablen, siehe Kapitel 51.
3. Die Euklidische Norm eines Vektors ist invariant unter Rotationen, siehe Satz 40.3. Ähnlich wie bei $\nabla f(\mathbf{x})$ die wichtige rotationsinvariante Größe $\|\nabla f(\mathbf{x})\|_2$ abgeleitet werden kann, kann man auch aus $Hf(\mathbf{x})$ eine rotationsinvariante skalare Größe gewinnen. Hierzu betrachtet man die Summe der Diagonalelemente, die sogenannte Spur, von $Hf(\mathbf{x})$.

□

**Definition 49.37 Laplace–Operator.** Seien $D \subset \mathbb{R}^n$ offen und $f : D \to \mathbb{R}$ in $\boldsymbol{\xi} \in D$ zweimal partiell differenzierbar. Dann nennt man

$$\Delta f(\boldsymbol{\xi}) := \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2}(\boldsymbol{\xi}) = \left(f_{x_1 x_1} + \ldots + f_{x_n x_n}\right)(\boldsymbol{\xi})$$

den Laplace–Operator von $f(\mathbf{x})$ in $\boldsymbol{\xi}$. □

**Bemerkung 49.38 Anwendungen des Laplace–Operators, partielle Differentialgleichungen.** Viele Prozesse in der Physik und den Ingenieurswissenschaften lassen sich durch Gleichungen beschreiben, welche die Beziehungen zwischen einer gesuchten Funktion, zum Beispiel Druck, Temperatur, Konzentration, ..., und ihren partiellen Ableitungen der Ordnung kleiner oder gleich Zwei beinhalten. Dabei spielt der Laplace–Operator oft eine große Rolle. Man nennt diese Gleichungen partielle Differentialgleichungen 2. Ordnung, wegen der Ableitungen zweiter Ordnung im Laplace–Operator. Im folgenden werden einige Beispiele angegeben.

1. Die Laplace–Gleichung oder Potentialgleichung hat die Gestalt

$$\Delta u = 0 \quad \text{in } D.$$

Sie spielt zum Beispiel bei statischen Problemen eine Rolle, beispielsweise in der Elastizitätstheorie. Funktionen, die die Laplace–Gleichung erfüllen, nennt man harmonische Funktionen.

[2]Ludwig Otto Hesse (1811 – 1874)

2. Die Wellengleichung besitzt die Form

$$u_{tt} = \Delta u \quad \text{in } (0,T) \times D.$$

Diese Gleichung beschreibt Schwingungsprobleme, zum Beispiel die Ausbreitung von Schallwellen. Hierbei ist $t$ die Zeit und $\Delta u = u_{xx} + u_{yy} + u_{zz}$ misst die örtliche Veränderung.

3. Die Diffusionsgleichung oder Wärmeleitungsgleichung hat die Form

$$u_t = \Delta u \quad \text{in } (0,T) \times D.$$

Sie beschreibt die Dynamik von Ausgleichsprozessen wie Diffusion und Wärmeleitung. Hierbei ist $u(t, \mathbf{x})$ die Konzentration oder die Temperatur, und $t$ ist die Zeit.
In der Bildverarbeitung verwendet man Diffusionsprozesse zum Entrauschen (Diffusionsfilter). Dabei beschreibt $u(t, \mathbf{x})$ den Grauwert, und die Diffusionszeit $t$ ist ein Maß für die Glättung.

□

## 49.6 Die Taylorsche Formel

**Bemerkung 49.39 Motivation.** Ziel ist die Verallgemeinerung der Taylorschen Formel aus Satz 22.2 auf den Fall skalarer Funktionen $f : D \subset \mathbb{R}^n \to \mathbb{R}$. □

**Satz 49.40 Taylorsche Formel mit dem Restglied von Lagrange.** *Seien $D \subset \mathbb{R}^n$ offen und konvex und sei $f(\mathbf{x})$ $(m+1)$–mal stetig differenzierbar in $D$. Dann gilt für $\boldsymbol{\xi}, \boldsymbol{\xi} + \mathbf{h} \in D$*

$$\begin{aligned} f(\boldsymbol{\xi} + \mathbf{h}) &= \sum_{i=0}^{m} \frac{1}{i!} \left( (\mathbf{h} \cdot \nabla)^i f \right) (\boldsymbol{\xi}) + \frac{1}{(m+1)!} \left( (\mathbf{h} \cdot \nabla)^{m+1} f \right) (\boldsymbol{\xi} + \theta \mathbf{h}) \\ &=: T_m(\mathbf{h}, \boldsymbol{\xi}) + R_m(\mathbf{h}, \boldsymbol{\xi}), \end{aligned}$$

*mit $\theta \in (0,1)$. Dabei werden $T_m(\mathbf{h}, \boldsymbol{\xi})$ Taylor–Polynom vom Grad $m$ und $R_m(\mathbf{h}, \boldsymbol{\xi})$ Restglied nach Lagrange genannt.*

**Beweis:** Die formale Schreibweise der Taylorschen Formel wird innerhalb des Beweises erklärt.

Wie im Beweis des Mittelwertsatzes wird eine Funktion einer reellen Veränderlichen betrachtet, nämlich

$$\phi(t) = f(\boldsymbol{\xi} + t\mathbf{h}), \quad t \in \mathbb{R},\ |t| < \delta.$$

Diese Funktion erfüllt die Voraussetzungen des Satzes von Taylor, Satz 22.2. Um diesen Satz anwenden zu können, muss man $\phi(t)$ differenzieren. Mit Hilfe der Kettenregel, Bemerkung 49.24, folgt

$$\begin{aligned} \phi'(t) &= \left( \frac{\partial f}{\partial x_1}(\boldsymbol{\xi} + t\mathbf{h}), \ldots, \frac{\partial f}{\partial x_n}(\boldsymbol{\xi} + t\mathbf{h}) \right) \cdot (h_1, \ldots, h_n) \\ &= \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\boldsymbol{\xi} + t\mathbf{h}) h_i =: (\mathbf{h} \cdot \nabla f)(\boldsymbol{\xi} + t\mathbf{h}) \end{aligned}$$

mit der formalen Schreibweise

$$(\mathbf{h} \cdot \nabla) := h_1 \frac{\partial}{\partial x_i} + \ldots + h_n \frac{\partial}{\partial x_n}.$$

Für die zweite Ableitung erhält man analog (betrachte am einfachsten die Summendarstellung der ersten Ableitung)

$$\phi''(t) = \sum_{i_1,i_2=1}^{n} \frac{\partial^2 f}{\partial x_{i_1}\partial x_{i_2}}(\boldsymbol{\xi}+t\mathbf{h})h_{i_1}h_{i_2} =: \left((\mathbf{h}\cdot\nabla)^2 f\right)(\boldsymbol{\xi}+t\mathbf{h})$$

mit

$$\begin{aligned}(\mathbf{h}\cdot\nabla)^2 &:= \left(h_1\frac{\partial}{\partial x_1}+\ldots+h_n\frac{\partial}{\partial x_n}\right)\cdot\left(h_1\frac{\partial}{\partial x_1}+\ldots+h_n\frac{\partial}{\partial x_n}\right)\\ &= h_1^2\frac{\partial^2}{\partial x_1^2}+h_1h_2\frac{\partial^2}{\partial x_1\partial x_2}+\ldots+h_2h_1\frac{\partial^2}{\partial x_2\partial x_1}+\ldots+h_n^2\frac{\partial^2}{\partial x_n^2}\\ &= \sum_{i_1,i_2=1}^{n}\frac{\partial^2}{\partial x_{i_1}\partial x_{i_2}}h_{i_1}h_{i_2}.\end{aligned}$$

Für die $k$–te Ableitung erhält man schließlich

$$\phi^{(k)}(t) = \sum_{i_1,\ldots,i_k=1}^{n} \frac{\partial^k f}{\partial x_{i_1}\ldots\partial x_{i_k}}(\boldsymbol{\xi}+t\mathbf{h})h_{i_1}\ldots h_{i_k} =: \left((\mathbf{h}\cdot\nabla)^k f\right)(\boldsymbol{\xi}+t\mathbf{h}).$$

Der Satz von Taylor für die Funktion $\phi(t)$ ergibt die Darstellung

$$\phi(1) = \sum_{i=0}^{m}\frac{\phi^{(i)}(0)}{i!}+\frac{\phi^{(m+1)}(\theta)}{(m+1)!}$$

mit $\theta\in(0,1)$. Setzt man die berechneten Ableitungen von $\phi(t)$ ein, ergibt sich gerade die Aussage des Satzes. ■

**Bemerkung 49.41 Zur Schreibweise der Taylor–Formel.** In der Literatur findet man die äquivalente Darstellung

$$(\mathbf{h}\cdot\nabla)^i = \left(\mathbf{h}^T\nabla\right)^i.$$

Wie bereits im Beweis angedeutet, wird dieser Ausdruck formal ausmultipliziert und dann nach Ableitungen sortiert. Die ersten drei Summanden lassen sich übersichtlich wie folgt schreiben:

$$f(\boldsymbol{\xi}+\mathbf{h}) = f(\boldsymbol{\xi})+\mathbf{h}^T\nabla f(\boldsymbol{\xi})+\frac{1}{2}\mathbf{h}^T Hf(\boldsymbol{\xi})\mathbf{h}+\ldots$$

mit der Hesse–Matrix $Hf(\boldsymbol{\xi})$.

Oft wird auch $\mathbf{x}=\boldsymbol{\xi}+\mathbf{h}$ gesetzt. Dann hat die Taylor–Formel die Gestalt

$$f(\mathbf{x}) = \sum_{i=0}^{m}\frac{1}{i!}\left(((\mathbf{x}-\boldsymbol{\xi})\cdot\nabla)^i f\right)(\boldsymbol{\xi})+\frac{1}{(m+1)!}\left(((\mathbf{x}-\boldsymbol{\xi})\cdot\nabla)^{m+1} f\right)(\boldsymbol{\xi}+\theta(\mathbf{x}-\boldsymbol{\xi}))$$

mit $\theta\in(0,1)$. Dann wird das Taylor–Polynom mit $T_m(\mathbf{x},\boldsymbol{\xi})$ bezeichnet und das Restglied mit $R_m(\mathbf{x},\boldsymbol{\xi})$. □

**Beispiel 49.42** Gesucht ist das Taylor–Polynom zweiten Grades von

$$f(x,y,z) = xy^2\sin z$$

im Entwicklungspunkt $\boldsymbol{\xi}=(1,2,0)^T$. Zur Lösung dieser Aufgabe benötigt man die Funktionswerte aller partiellen Ableitungen bis zur zweiten Ordnung im Entwicklungspunkt:

| Ableitungen | Auswertung in $(1,2,0)^T$ |
|---|---|
| $f = xy^2 \sin z$ | 0 |
| $f_x = y^2 \sin z$ | 0 |
| $f_y = 2xy \sin z$ | 0 |
| $f_z = xy^2 \cos z$ | 4 |
| $f_{xx} = 0$ | 0 |
| $f_{yy} = 2x \sin z$ | 0 |
| $f_{zz} = -xy^2 \sin z$ | 0 |
| $f_{xy} = 2y \sin z$ | 0 |
| $f_{xz} = y^2 \cos z$ | 4 |
| $f_{yz} = 2xy \cos z$ | 4 |

Damit erhält man

$$f(\boldsymbol{\xi}) = 0, \quad \nabla f(\boldsymbol{\xi}) = \begin{pmatrix} 0 \\ 0 \\ 4 \end{pmatrix}, \quad Hf(\boldsymbol{\xi}) = \begin{pmatrix} 0 & 0 & 4 \\ 0 & 0 & 4 \\ 4 & 4 & 0 \end{pmatrix}$$

und es folgt für das Taylor–Polynom 2. Grades mit $\mathbf{h} = \mathbf{x} - \boldsymbol{\xi}$

$$\begin{aligned} & T_2(\mathbf{x}, \boldsymbol{\xi}) \\ = \; & 0 + \begin{pmatrix} x_1 - 1 \\ x_2 - 2 \\ x_3 \end{pmatrix}^T \begin{pmatrix} 0 \\ 0 \\ 4 \end{pmatrix} + \frac{1}{2} \left( \begin{pmatrix} x_1 - 1 \\ x_2 - 2 \\ x_3 \end{pmatrix}^T \begin{pmatrix} 0 & 0 & 4 \\ 0 & 0 & 4 \\ 4 & 4 & 0 \end{pmatrix} \begin{pmatrix} x_1 - 1 \\ x_2 - 2 \\ x_3 \end{pmatrix} \right) \\ = \; & 4x_3 + \frac{1}{2} \left( 8(x_1 - 1)x_3 + 8(x_2 - 2)x_3 \right) = -8x_3 + 4x_1x_3 + 4x_2x_3. \end{aligned}$$

□

**Bemerkung 49.43**

- Für $m = 0$ liefert der Satz von Taylor

$$f(\mathbf{x}) \quad = \quad f(\boldsymbol{\xi}) + (\mathbf{x} - \boldsymbol{\xi})^T \nabla f(\boldsymbol{\xi} + \theta(\mathbf{x} - \boldsymbol{\xi})), \quad \theta \in (0, 1).$$

  Dies ist äquivalent zum Mittelwertsatz 49.28 angewendet auf $\mathbf{a} = \boldsymbol{\xi}$, $\mathbf{b} = \mathbf{x}$. Der Mittelwertsatz ist also ein Spezialfall des Satzes von Taylor.
- Für beschränkte partielle Ableitungen der Ordnung $m+1$ hat das Restglied die Fehlerordnung $\mathcal{O}(\|\mathbf{x} - \boldsymbol{\xi}\|_2^{m+1})$. Somit folgt für die Approximationsgüte des Taylorpolynoms

$$f(\mathbf{x}) = T_m(\mathbf{x}, \boldsymbol{\xi}) + \mathcal{O}\left( \|\mathbf{x} - \boldsymbol{\xi}\|_2^{m+1} \right).$$

□

# Kapitel 50

# Ableitungsoperatoren

**Bemerkung 50.1 Motivation.** Als Ableitungsoperatoren wurden bisher der Gradient und der Laplace–Operator eingeführt. In diesem Abschnitt werden weitere Ableitungsoperatoren mit Hilfe des Nabla–Operators definiert. □

**Definition 50.2 Vektorfeld.** Sei $\mathbf{f} : D \to \mathbb{R}^n$ mit $D \subset \mathbb{R}^n$, dann wird $\mathbf{f}(\mathbf{x})$ ein Vektorfeld auf $D$ genannt. Ist jede Komponente $f_i(\mathbf{x})$ $k$–mal stetig differenzierbar auf $D$, so heißt $\mathbf{f}(\mathbf{x})$ ein $C^k(D)$–Vektorfeld. □

**Beispiel 50.3** Die Geschwindigkeit strömender Gase oder Flüssigkeiten ordnet zu einem festen Zeitpunkt jedem Raumpunkt aus $D \subset \mathbb{R}^3$ einen dreidimensionalen Vektor zu, der in Richtung der Strömung zeigt und dessen Länge der Betrag der Geschwindigkeit ist. Demzufolge ist die Geschwindigkeit eine Abbildung aus einer Teilmenge des $\mathbb{R}^3$ nach $\mathbb{R}^3$. Sie ist ein Vektorfeld. □

**Definition 50.4 Divergenz.** Sei $D \subset \mathbb{R}^n$ offen. Für ein differenzierbares Vektorfeld $\mathbf{f} : D \to \mathbb{R}^n$ definiert man die Divergenz durch die Summe der partiellen Ableitungen erster Ordnung

$$\mathrm{div}\mathbf{f}(\boldsymbol{\xi}) := \sum_{i=1}^{n} \frac{\partial f_i}{\partial x_i}(\boldsymbol{\xi}).$$

Man schreibt die Divergenz auch formal als Skalarprodukt des Nabla–Operators und des Vektorfeldes

$$\mathrm{div}\mathbf{f}(\boldsymbol{\xi}) = (\nabla \cdot \mathbf{f})(\boldsymbol{\xi}).$$

□

**Satz 50.5 Rechenregeln für die Divergenz.** *Sei $D \subset \mathbb{R}^n$ offen. Für differenzierbare Vektorfelder $\mathbf{f}, \mathbf{g} : D \to \mathbb{R}^n$ und eine skalarwertige differenzierbare Funktion $\varphi : D \to \mathbb{R}$ gelten:*

*i) Linearität:*

$$\mathrm{div}\,(\alpha\mathbf{f} + \beta\mathbf{g})(\boldsymbol{\xi}) = \alpha\mathrm{div}\mathbf{f}(\boldsymbol{\xi}) + \beta\mathrm{div}\mathbf{g}(\boldsymbol{\xi}) \quad \forall\ \alpha, \beta \in \mathbb{R},$$

*ii) Produktregel:*

$$\mathrm{div}\,(\varphi\mathbf{f})(\boldsymbol{\xi}) = (\nabla\varphi(\boldsymbol{\xi}))^T\mathbf{f}(\boldsymbol{\xi}) + \varphi(\boldsymbol{\xi})\mathrm{div}\mathbf{f}(\boldsymbol{\xi}).$$

**Beweis:** Direktes Nachrechnen mit Hilfe der Linearität und der Produktregel für reellwertige Funktionen einer reellen Veränderlichen, Übungsaufgabe. ■

**Bemerkung 50.6 Zur Divergenz.**

- Obwohl der Gradient und die Divergenz ähnlich aussehen, unterscheiden sich ihre Wirkung sehr stark:
  - Der Gradient erzeugt aus einer skalarwertigen Funktion eine vektorwertige Funktion.
  - Die Divergenz macht aus einem Vektorfeld eine skalarwertige Funktion.
- Mit Hilfe der Divergenz kann man den Laplace-Operator umschreiben. Ist $f : D \to \mathbb{R}$ eine skalarwertige, zweimal stetig differenzierbare Funktion mehrerer Variabler, so findet man durch direktes Nachrechnen

$$\Delta f = \operatorname{div}(\nabla f).$$

  Somit kann man zum Beispiel die Diffusionsgleichung $u_t = \Delta u$ als

$$u_t = \operatorname{div}(\nabla u) = \nabla \cdot \nabla u$$

  schreiben.
- Die Modellierung von Strömungen erfolgt mittels physikalischer Gesetze, nämlich der Impulserhaltung und der Massenerhaltung. Man erhält die Grundgleichungen der Strömungsmechanik, die sogenannten Navier[1]–Stokes[2]–Gleichungen. Sei $\mathbf{u}(\mathbf{x})$ das Geschwindigkeitsfeld zu einem festen Zeitpunkt. Dann hat die Massenerhaltung für inkompressible Fluide, zum Beispiel Wasser, die Gestalt $\nabla \cdot \mathbf{u}(\mathbf{x}) = 0$ für alle Punkte $\mathbf{x}$ aus dem Strömungsgebiet.

□

Neben der Divergenz gibt es noch einen weiteren wichtigen Differentialausdruck für Vektorfelder.

**Definition 50.7 Rotation.** Sei $D \subset \mathbb{R}^3$ offen. Für ein differenzierbares Vektorfeld $\mathbf{f} : D \to \mathbb{R}^3$ definiert man die Rotation in einem Punkt $\boldsymbol{\xi} \in D$ durch

$$\operatorname{rot}\mathbf{f}(\boldsymbol{\xi}) := \begin{pmatrix} \dfrac{\partial f_3}{\partial x_2}(\boldsymbol{\xi}) - \dfrac{\partial f_2}{\partial x_3}(\boldsymbol{\xi}) \\ \dfrac{\partial f_1}{\partial x_3}(\boldsymbol{\xi}) - \dfrac{\partial f_3}{\partial x_1}(\boldsymbol{\xi}) \\ \dfrac{\partial f_2}{\partial x_1}(\boldsymbol{\xi}) - \dfrac{\partial f_1}{\partial x_2}(\boldsymbol{\xi}) \end{pmatrix}.$$

Die formale Schreibweise mit Hilfe des Nabla–Operators ist

$$\operatorname{rot}\mathbf{f}(\boldsymbol{\xi}) =: \nabla \times \mathbf{f}(\boldsymbol{\xi}).$$

□

**Bemerkung 50.8 Zur Rotation, das Kreuzprodukt.**

1. Die Rotation ist nur für Vektorfelder im $\mathbb{R}^3$ definiert. Vektorfelder im $\mathbb{R}^2$ kann man formal um eine dritte Komponente erweitern und auch für sie die Rotation berechnen. Im Ergebnis sind die ersten beiden Komponenten der Rotation in diesem Falle Null.
2. Für zwei Vektoren

$$\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix}$$

[1] Claude Louis Marie Henri Navier (1785 – 1836)
[2] George Gabriel Stokes (1819 – 1903)

kann man das Kreuzprodukt $\mathbf{v} \times \mathbf{w}$ definieren durch

$$\mathbf{v} \times \mathbf{w} := \begin{pmatrix} v_2 w_3 - v_3 w_2 \\ v_3 w_1 - v_1 w_3 \\ v_1 w_2 - v_2 w_1 \end{pmatrix} = \det \begin{pmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{pmatrix}$$

mit den kartesischen Einheitsvektoren. Zu Eigenschaften des Kreuzproduktes wird es Übungsaufgaben geben.

3. Sei $\mathbf{u}(\mathbf{x})$ die Geschwindigkeit einer Strömung zu einem festen Zeitpunkt. Ist $\mathrm{rot}(\mathbf{x}) = \mathbf{0}$ für alle Punkte $\mathbf{x}$ des Strömungsgebietes, so ist die Strömung rotationsfrei oder wirbelfrei.

□

**Beispiel 50.9** Betrachte das Vektorfeld

$$\mathbf{f} \; : \; \mathbb{R}^3 \to \mathbb{R}^3, \quad \begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto \begin{pmatrix} 3x^2 y - z \\ z^2 + x^2 \\ y + 2x^2 \end{pmatrix}.$$

Dann ist

$$\begin{aligned} \mathrm{rot}\mathbf{f}(x,y,z) &= \nabla \times \mathbf{f} = \det \begin{pmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_2 \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ 3x^2 y - z & z^2 + x^2 & y + 2x^2 \end{pmatrix} \\ &= \left( \frac{\partial (y + 2x^2)}{\partial y} - \frac{\partial (z^2 + x^2)}{\partial z} \right) \mathbf{e}_1 \\ &\quad + \left( \frac{\partial (3x^2 y - z)}{\partial z} - \frac{\partial (y + 2x^2)}{\partial x} \right) \mathbf{e}_2 \\ &\quad + \left( \frac{\partial (z^2 + x^2)}{\partial x} - \frac{\partial (3x^2 y - z)}{\partial y} \right) \mathbf{e}_3 \\ &= (1 - 2z)\mathbf{e}_1 + (-1 - 4x)\mathbf{e}_2 + (2x - 3x^2)\mathbf{e}_3 \\ &= \begin{pmatrix} 1 - 2z \\ -1 - 4x \\ 2x - 3x^2 \end{pmatrix}. \end{aligned}$$

□

Man kann ähnliche Rechenregeln wie beim Divergenzoperator zeigen.

**Satz 50.10 Rechenregeln für die Rotation.** *Sei $D \subset \mathbb{R}^3$ offen. Für differenzierbare Vektorfelder $\mathbf{f}, \mathbf{g} \; : \; D \to \mathbb{R}^3$ und eine differenzierbare skalarwertige Funktion $\varphi \; : \; D \to \mathbb{R}$ gelten:*

*i) Linearität:*

$$\mathrm{rot}(\alpha \mathbf{f} + \beta \mathbf{g})(\boldsymbol{\xi}) = \alpha \mathrm{rot}\mathbf{f}(\boldsymbol{\xi}) + \beta \mathrm{rot}\mathbf{g}(\boldsymbol{\xi}) \quad \forall \; \alpha, \beta \in \mathbb{R},$$

*ii) Produktregel:*

$$\mathrm{rot}(\varphi \mathbf{f})(\boldsymbol{\xi}) = (\nabla \varphi(\boldsymbol{\xi})) \times \mathbf{f}(\boldsymbol{\xi}) + \varphi(\boldsymbol{\xi}) \mathrm{rot}\mathbf{f}(\boldsymbol{\xi}).$$

**Beweis:** Der Beweis erfolgt durch direktes Nachrechnen. ■

# Kapitel 51

# Lokale Extrema von Funktionen mehrerer Variabler

**Bemerkung 51.1 Motivation.** Bei skalarwertigen Funktionen einer Variablen gibt es notwendige und hinreichende Bedingungen für das Vorliegen von lokalen Extrema:

- Sei $f : (a,b) \to \mathbb{R}$ in $C^1(a,b)$. Besitzt $f(x)$ in $\xi \in (a,b)$ ein lokales Extremum, so gilt $f'(\xi) = 0$, Satz von Fermat, Satz 21.17.
- Seien $f : (a,b) \to \mathbb{R}$ in $C^2(a,b)$, $\xi \in (a,b)$ und $f'(\xi) = 0$. Ist $\xi$ ein lokales Minimum (beziehungsweise lokales Maximum), so gilt $f''(\xi) \geq 0$ (beziehungsweise $f(\xi) \leq 0$), siehe Satz 23.12. Ist umgekehrt $f''(\xi) > 0$ (beziehungsweise $f''(\xi) < 0$), so ist $\xi$ ein striktes lokales Minimum (striktes lokales Maximum), siehe auch Satz 23.12.

In diesem Abschnitt werden ähnliche Aussagen im Fall skalarwertiger Funktionen mehrerer Variabler hergeleitet. □

**Definition 51.2 (Strikte) Lokale Minima und Maxima.** Seien $D \subset \mathbb{R}^n$ offen und $f : D \to \mathbb{R}$. Ein Punkt $\boldsymbol{\xi} \in D$ heißt lokales Minimum (beziehungsweise lokales Maximum), falls eine Umgebung $U \subset D$ existiert mit $f(\boldsymbol{\xi}) \leq f(\mathbf{y})$ (beziehungsweise $f(\boldsymbol{\xi}) \geq f(\mathbf{y})$) für alle $\mathbf{y} \in U$. Tritt Gleichheit nur für $\mathbf{y} = \boldsymbol{\xi}$ auf, heißt $\boldsymbol{\xi}$ striktes lokales Minimum (beziehungsweise striktes lokales Maximum). □

Für mehrdimensionale Definitionsbereiche gibt es ein analoges notwendiges Kriterium für die Existenz eines lokalen Extremums.

**Satz 51.3 Notwendige Bedingung für lokale Extrema.** *Seien $D \subset \mathbb{R}^n$ offen und $f : D \to \mathbb{R}$ stetig differenzierbar in $D$. Hat $f(\mathbf{x})$ in $\boldsymbol{\xi} \in D$ ein lokales Extremum (Minimum oder Maximum), so gilt $\nabla f(\boldsymbol{\xi}) = \mathbf{0}$.*

**Beweis:** Für beliebiges $\mathbf{v} \in \mathbb{R}^n$ mit $\mathbf{v} \neq \mathbf{0}$ ist

$$\varphi(t) := f(\boldsymbol{\xi} + t\mathbf{v})$$

in einer Umgebung von $t = 0$ erklärt und dort stetig differenzierbar. Ferner besitzt $\varphi(t)$ in $t = 0$ ein lokales Extremum. Mit dem Satz von Fermat, Satz 21.17, und der Kettenregel, Satz 49.23, gilt mit $\mathbf{x} = \boldsymbol{\xi} + t\mathbf{v}$

$$0 = \varphi'(0) = \sum_{i=0}^{n} \frac{\partial f}{\partial x_i}(\boldsymbol{\xi}) \frac{\partial x_i}{\partial t} = \nabla f(\boldsymbol{\xi}) \cdot \mathbf{v}.$$

Da dies für beliebige $\mathbf{v} \neq \mathbf{0}$ gilt, ist $\nabla f(\boldsymbol{\xi}) = \mathbf{0}$. ■

**Bemerkung 51.4**

- Ein Punkt $\boldsymbol{\xi} \in D$ mit $\nabla f(\boldsymbol{\xi}) = \mathbf{0}$ heißt auch stationärer Punkt von $f(\mathbf{x})$.
- Bei der Suche nach stationären Punkte muss man im Allgemeinen ein nichtlineares System von $n$ Gleichungen mit $n$ Unbekannten lösen.
  Betracht zum Beispiel die Funktion

$$f(x, y) = e^{xy} + x^2 \sin y.$$

  Das zu lösende System zum Finden stationärer Punkte lautet

$$\mathbf{0} \stackrel{!}{=} \nabla f(x, y) = \begin{pmatrix} ye^{xy} + 2x \sin y \\ xe^{xy} + x^2 \cos y \end{pmatrix}.$$

  Das sind zwei Gleichungen mit zwei Unbekannten. Ähnlich wie bei Funktionen einer Variablen kann man hierbei Fixpunkt–Interationsverfahren, beispielsweise das Newton–Verfahren, einsetzen.
- Analog wie in einer Dimension ist nicht jeder stationäre Punkt ein lokales Extremum.

□

**Beispiel 51.5** Betrachte die Funktion

$$f(x, y) = x^2 - y^2.$$

Die stationären Punkte erfüllen die Gleichung

$$\mathbf{0} = \nabla f = \begin{pmatrix} 2x \\ -2y \end{pmatrix}.$$

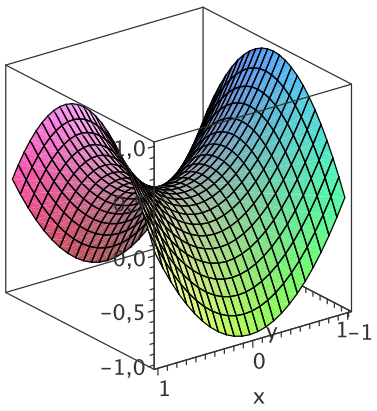
Damit ist

$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

der einzige stationäre Punkt. Allerdings existieren in jeder Umgebung um diesen stationären Punkt Punkte $(x_1, y_1)$ und $(x_2, y_2)$ mit

$$f(x_1, y_1) < f(x_0, y_0) < f(x_2, y_2).$$

Solche stationären Punkte nennt man Sattelpunkte.




□

Zur Klassifikation stationärer Punkte gibt es auch ein Analogon zum skalaren Fall.

**Satz 51.6 Klassifikation stationärer Punkte.** *Seien $D \subset \mathbb{R}^n$ offen, $f : D \to \mathbb{R}$ zweimal stetig differenzierbar, $\boldsymbol{\xi} \in D$ sowie $\nabla f(\boldsymbol{\xi}) = \mathbf{0}$. Dann gelten:*

*i) Notwendige Bedingungen. Besitzt $f(\mathbf{x})$ in $\boldsymbol{\xi}$ ein lokales Minimum (beziehungsweise lokales Maximum), so ist die Hesse–Matrix $Hf(\boldsymbol{\xi})$ positiv semidefinit (beziehungsweise negativ semidefinit).*

*ii) Hinreichende Bedingungen.*

*a) Ist $Hf(\boldsymbol{\xi})$ positiv definit (beziehungsweise negativ definit), so besitzt $f(\mathbf{x})$ in $\boldsymbol{\xi}$ ein striktes lokales Minimum (beziehungsweise striktes lokales Maximum).*

*b) Ist $Hf(\boldsymbol{\xi})$ indefinit, so ist $\boldsymbol{\xi}$ ein Sattelpunkt, das heißt in jeder Umgebung $U \subset D$ existieren $\mathbf{y}, \mathbf{z} \in U$ mit $f(\mathbf{y}) < f(\boldsymbol{\xi}) < f(\mathbf{z})$.*

**Beweis:** *i).* Sei $\boldsymbol{\xi}$ ohne Beschränkung der Allgemeinheit ein lokales Minimum. Für $\mathbf{v} \neq \mathbf{0}$ und hinreichend kleines $\varepsilon > 0$ gilt nach dem Satz von Taylor, Satz 49.40,

$$0 \leq f(\boldsymbol{\xi} + \varepsilon\mathbf{v}) - f(\boldsymbol{\xi}) = \varepsilon\mathbf{v}^T \underbrace{\nabla f(\boldsymbol{\xi})}_{\mathbf{0} \text{ nach Vor.}} + \frac{1}{2}\varepsilon\mathbf{v}^T Hf(\boldsymbol{\xi} + \theta\varepsilon\mathbf{v})\varepsilon\mathbf{v}$$

mit $\theta \in (0,1)$. Also ist $\mathbf{v}^T Hf(\boldsymbol{\xi} + \theta\varepsilon\mathbf{v})\mathbf{v} \geq 0$. Wegen der Stetigkeit der zweiten Ableitung folgt

$$0 \leq \lim_{\varepsilon \to 0} \mathbf{v}^T Hf(\boldsymbol{\xi} + \theta\varepsilon\mathbf{v})\mathbf{v} = \mathbf{v}^T Hf(\boldsymbol{\xi})\mathbf{v}.$$

Da $\mathbf{v}$ beliebig war, ist $Hf(\boldsymbol{\xi})$ somit positiv semidefinit, was man analog zu Satz 43.5 zeigen kann.

*ii, a).* Sei $Hf(\boldsymbol{\xi})$ positiv definit. Da $f(\mathbf{x})$ zweimal stetig differenzierbar ist, ist $Hf(\mathbf{x})$ auch in einer hinreichend kleinen Kugelumgebung $B(\boldsymbol{\xi}, \varepsilon)$ mit Radius $\varepsilon$ und Mittelpunkt $\boldsymbol{\xi}$ positiv definit. Für $\mathbf{x} \in B(\boldsymbol{\xi}, \varepsilon) \setminus \{\boldsymbol{\xi}\}$ gilt also nach dem Satz von Taylor

$$f(\mathbf{x}) - f(\boldsymbol{\xi}) = \underbrace{(\nabla f(\boldsymbol{\xi}))^T}_{=0}(\mathbf{x} - \boldsymbol{\xi}) + \frac{1}{2}\underbrace{(\mathbf{x} - \boldsymbol{\xi})^T Hf(\boldsymbol{\xi} + \theta(\mathbf{x} - \boldsymbol{\xi}))(\mathbf{x} - \boldsymbol{\xi})}_{>0} > 0,$$

mit $\theta \in (0,1)$. Da $\mathbf{x} \in B(\boldsymbol{\xi}, \varepsilon) \setminus \{\boldsymbol{\xi}\}$ beliebig war, folgt aus $f(\mathbf{x}) - f(\boldsymbol{\xi}) > 0$, dass $\boldsymbol{\xi}$ ein striktes lokales Minimum ist.

*ii), b).* Sei nun $Hf(\boldsymbol{\xi})$ indefinit. Dann existieren Eigenwerte $\lambda_1, \lambda_2$ von $Hf(\boldsymbol{\xi})$ mit $\lambda_1 > 0$, $\lambda_2 < 0$. Für die zugehörigen Eigenvektoren $\mathbf{v}, \mathbf{w}$ gelten also

$$\mathbf{v}^T Hf(\boldsymbol{\xi})\mathbf{v} > 0, \quad \mathbf{w}^T Hf(\boldsymbol{\xi})\mathbf{w} < 0.$$

Wie eben zeigt man, dass es $\varepsilon_1, \varepsilon_2 > 0$ gibt mit

$$\begin{aligned} f(\boldsymbol{\xi} + \varepsilon\mathbf{v}) - f(\boldsymbol{\xi}) &= \frac{1}{2}(\varepsilon\mathbf{v})^T Hf(\boldsymbol{\xi} + \theta_1\varepsilon\mathbf{v})\varepsilon\mathbf{v} > 0, \\ f(\boldsymbol{\xi} + \varepsilon\mathbf{w}) - f(\boldsymbol{\xi}) &= \frac{1}{2}(\varepsilon\mathbf{w})^T Hf(\boldsymbol{\xi} + \theta_2\varepsilon\mathbf{w})\varepsilon\mathbf{w} < 0 \end{aligned}$$

für alle $\varepsilon \in (0, \min\{\varepsilon_1, \varepsilon_2\})$. Somit ist $\boldsymbol{\xi}$ Sattelpunkt von $f(\mathbf{x})$. ■

**Bemerkung 51.7 Implikationen.** Nach Satz 51.6 gelten folgende Implikationen:

$$\begin{array}{ccc} \boldsymbol{\xi} \text{ lokales Minimum} & \Longleftarrow\!\!= & \boldsymbol{\xi} \text{ striktes lokales Minimum} \\ \Downarrow & & \Uparrow \\ Hf(\boldsymbol{\xi}) \text{ positiv semidefinit} & \Longleftarrow\!\!= & Hf(\boldsymbol{\xi}) \text{ positiv definit} \end{array}$$

Keine Implikation ist umkehrbar! □

**Beispiel 51.8** Betrachte die Funktion

$$f(x,y) = y^2(x-1) + x^2(x+1) \quad \Longrightarrow \quad \nabla f(x,y) = \begin{pmatrix} y^2 + 3x^2 + 2x \\ 2y(x-1) \end{pmatrix}.$$

Zur Berechnung der stationären Punkte setzt man den Gradienten zu Null. Insbesondere ist die zweite Komponente Null. Diese kann nur Null sein, wenn einer der Faktoren Null ist.

*1. Fall.* $x - 1 = 0$. Damit folgt $x = 1$. Einsetzen in die erste Komponente liefert $y^2 + 5 = 0$. Diese Gleichung besitzt keine reelle Lösung.

*2. Fall.* $y = 0$. Einsetzen in die erste Komponente ergibt $x(3x + 2) = 0$. Diese Gleichung besitzt die Lösungen $x_1 = 0$ und $x_2 = -2/3$.

Es gibt also 2 stationäre Punkte

$$\boldsymbol{\xi} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{\eta} = \begin{pmatrix} -2/3 \\ 0 \end{pmatrix}.$$

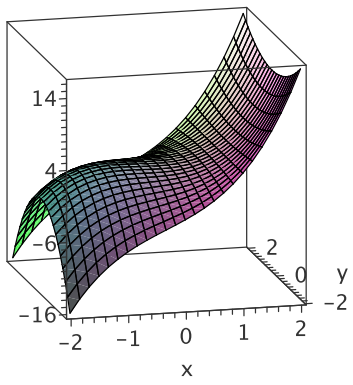
Zu ihrer Klassifikation wird noch die Hessematrix

$$Hf(x,y) = \begin{pmatrix} 6x+2 & 2y \\ 2y & 2(x-1) \end{pmatrix}$$

ausgewertet

$$Hf(0,0) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}, \quad Hf(-2/3,0) = \begin{pmatrix} -2 & 0 \\ 0 & -10/3 \end{pmatrix}.$$

Die erste Matrix ist indefinit, also ist $\boldsymbol{\xi}$ ein Sattelpunkt. Die zweite Matrix ist negativ definit, also ist $\boldsymbol{\eta}$ ein striktes lokales Maximum.




□

# Kapitel 52

# Extrema mit Nebenbedingungen

**Bemerkung 52.1 Motivation.** Im Kapitel 51 wurden notwendige und hinreichende Kriterien vorgestellt, um lokale Extrema einer skalaren Funktion mehrerer Variabler zu bestimmen. Oft sucht man jedoch die Extrema einer Funktion unter der Einschränkung, dass bestimmte Nebenbedingungen erfüllt sein müssen. □

**Beispiel 52.2 Verpackungsminimierung im $\mathbb{R}^2$.** Gesucht ist ein Rechteck maximalen Inhalts bei einem vorgegebenen Umfang $u$. Das bedeutet, man muss

$$f(x,y) = xy, \quad x, y - \text{Seitenlängen},$$

unter der Nebenbedingung

$$g(x,y) = 2x + 2y - u = 0.$$

maximieren.

*Lösungsmöglichkeit 1 (spezieller Weg):* Man löst die Nebenbedingung nach einer Variablen auf

$$2y = u - 2x \quad \Longrightarrow \quad y = \frac{u}{2} - x. \tag{52.1}$$

Dies setzt man in die zu maximierende Funktion ein. Damit ist das neue Optimierungsproblem, die Funktion

$$\tilde{f}(x) = x\left(\frac{u}{2} - x\right)$$

zu maximieren. Das bedeutet, die Nebenbedingung reduziert die Freiheitsgrade von 2 auf 1. Das neue Optimierungsproblem kann man mit Hilfe von Mitteln der skalaren Analysis lösen. Die notwendige Bedingung für ein Extremum lautet

$$0 \stackrel{!}{=} \tilde{f}'(x) = \frac{u}{2} - 2x \quad \Longrightarrow \quad x = \frac{u}{4}.$$

Einsetzen in (52.1) ergibt

$$y = \frac{u}{2} - x = \frac{u}{2} - \frac{u}{4} = \frac{u}{4}.$$

Da

$$\tilde{f}''\left(\frac{u}{2}\right) = -2 < 0$$

ist, handelt es sich um ein Maximum. Das optimale Rechteck ist ein Quadrat mit Seitenlänge $x = y = u/4$.

Diese Lösungsmöglichkeit ist speziell, weil die Nebenbedingung nach einer der Variablen aufgelöst wurde. Das ist nicht immer möglich. Falls dies nicht möglich ist oder umständlich erscheint, muss man einen anderen Weg wählen.

*Lösungsmöglichkeit 2 (allgemeiner Weg):* Man formt die Aufgabenstellung so um, dass man mit Mitteln der Extremwerttheorie ohne Nebenbedingungen in $\mathbb{R}^n$ arbeiten kann. Dazu erweitert man die Zielfunktion, indem man die Nebenbedingung einarbeitet

$$F(x, y, \lambda) := f(x, y) + \lambda g(x, y) = xy + \lambda(2x + 2y - u).$$

Die zusätzliche Variable $\lambda$ nennt man Lagrange–Multiplikator. Nun kann man für $F(x, y, \lambda)$ die im vorigen Kapitel vorgestellte Extremwerttheorie anwenden. Notwendige Bedingung für ein Maximum ist

$$\mathbf{0} \stackrel{!}{=} \begin{pmatrix} \dfrac{\partial F}{\partial x} \\ \dfrac{\partial F}{\partial y} \\ \dfrac{\partial F}{\partial \lambda} \end{pmatrix} = \begin{pmatrix} y + 2\lambda \\ x + 2\lambda \\ 2x + 2y - u \end{pmatrix}.$$

Man sieht, dass die dritte Gleichung die Nebenbedingung $g(x, y) = 0$ repräsentiert. Auflösen der 1. und 2. Gleichung nach $\lambda$ ergibt

$$y = -2\lambda, \quad x = -2\lambda \quad \Longrightarrow \quad x = y.$$

Durch einsetzen in die 3. Gleichung erhält man

$$4x - u = 0 \quad \Longrightarrow \quad x = \frac{u}{4} = y \quad \Longrightarrow \quad \lambda = -\frac{u}{8}.$$

Man kann nachweisen, dass dies tatsächlich ein Maximum ist. □

**Bemerkung 52.3 Allgemeine Vorgehensweise.** Seien $D \subset \mathbb{R}^n$ offen und $f : D \to \mathbb{R}$. Gesucht sind Extrema der Funktion $f(x_1, \ldots, x_n) = 0$ unter den $m$ Nebenbedingungen

$$\begin{array}{rcl} g_1(x_1, \ldots, x_n) & = & 0 \\ & \vdots & \\ g_m(x_1, \ldots, x_n) & = & 0 \end{array} \qquad \Longleftrightarrow \quad \mathbf{g}(\mathbf{x}) = \mathbf{0}$$

mit $m < n$.

Statt der $m$ Nebenbedingungen $\mathbf{g}(\mathbf{x}) = \mathbf{0}$ führt man $m$ zusätzliche Variablen $(\lambda_1, \ldots, \lambda_m)^T =: \boldsymbol{\lambda}$ ein, die so genannten Lagrange–Multiplikatoren. Dann maximiert/minimiert man statt $f(\mathbf{x})$ die Lagrange–Funktion

$$F(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}, \boldsymbol{\lambda}) + \boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x}).$$

Damit hat man ein Extremalproblem mit $n$ Variablen und $m$ Nebenbedingungen in ein Extremalproblem mit $n + m$ Variablen ohne explizite Nebenbedingungen umgewandelt. Für dieses Problem kann man das Vorgehen aus Kapitel 51 anwenden.

Die Anwendung der Hessematrix als hinreichende Bedingung ist jedoch nicht so einfach wie im Fall der Extremwertbestimmung ohne Nebenbedingungen. Das liegt daran, dass das Verhalten des Lagrange–Multiplikators an der Extremstelle für den Typ der Extremstelle unwichtig ist und die Auswahl der Vektoren eingeschränkt ist, da diese die Nebenbedingungen erfüllen müssen, siehe Literatur für Details. Um festzustellen, von welcher Art das Extremum ist, ist es oft sinnvoll sich Argumentationen zu überlegen, die auf das jeweilige Beispiel zugeschnitten sind. □

Man kann zeigen, dass der Ansatz aus Bemerkung 52.3 funktioniert.

**Satz 52.4** *Seien $D \subset \mathbb{R}^n$ offen und seien $f, g_1, \ldots, g_m : D \to \mathbb{R}$ stetig differenzierbare Funktionen. Ferner sei $\boldsymbol{\xi} \in D$ ein lokales Extremum von $f(\mathbf{x})$ unter der Nebenbedingung $\mathbf{g}(\mathbf{x}) = \mathbf{0}$ und es gelte die Regularitätsbedingung*

$$rang\,(Jg(\boldsymbol{\xi})) = m.$$

*Dann existieren Lagrange–Multiplikatoren $\lambda_1, \ldots, \lambda_m$, so dass die Lagrange–Funktion*

$$F(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x})$$

*die notwendige Bedingung $\nabla F(\boldsymbol{\xi}, \lambda) = \mathbf{0}$ erfüllt.*

**Beweis:** Siehe Literatur. ■

**Beispiel 52.5** Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch. Man bestimme die Extrema der quadratischen Form $\mathbf{x}^T A\mathbf{x}$ unter der Nebenbedingung $\|\mathbf{x}\|_2 = 1$.

Die Lagrange–Funktion ist

$$F(\mathbf{x}, \lambda) = \mathbf{x}^T A\mathbf{x} + \lambda(1 - \mathbf{x}^T\mathbf{x}) = \sum_{i,j=1}^{n} x_i x_j a_{ij} + \lambda\left(1 - \sum_{i=1}^{n} x_i^2\right).$$

Die notwendige Bedingung für ein Extremum ist

$$\mathbf{0} = \left(\frac{\partial F}{\partial x_1}, \ldots, \frac{\partial F}{\partial x_n}, \frac{\partial F}{\partial \lambda}\right)^T.$$

Mit

$$\begin{aligned}
\frac{\partial}{\partial x_k}\sum_{i,j=1}^{n} x_i x_j a_{ij} &= \frac{\partial}{\partial x_k}\sum_{j=1,j\neq k}^{n} x_k x_j a_{kj} + \frac{\partial}{\partial x_k}\sum_{i=1,i\neq k}^{n} x_i x_k a_{ik} + \frac{\partial}{\partial x_k} a_{kk} x_k^2 \\
&= \sum_{j=1,j\neq k}^{n} x_j a_{kj} + \sum_{i=1,i\neq k}^{n} x_i \underbrace{a_{ik}}_{=a_{ki}} + 2a_{kk}x_k \\
&= 2\sum_{j=1}^{n} a_{kj}x_j, \quad k = 1, \ldots, n,
\end{aligned}$$

folgt

$$\mathbf{0} = \begin{pmatrix} \frac{\partial F}{\partial x_1} \\ \vdots \\ \frac{\partial F}{\partial x_n} \\ \frac{\partial F}{\partial \lambda} \end{pmatrix} = \begin{pmatrix} 2\sum_{j=1}^{n} a_{1j}x_j - 2\lambda x_1 \\ \vdots \\ 2\sum_{j=1}^{n} a_{nj}x_j - 2\lambda x_n \\ 1 - \sum_{i=1}^{n} x_i^2 \end{pmatrix}.$$

Gleichungen 1 bis $n$ besagen, dass $\mathbf{x}$ Eigenvektor zu einem Eigenwert $\lambda$ von $A$ ist

$$\sum_{j=1}^{n} a_{kj}x_j = \lambda x_k, \quad k = 1, \ldots n \quad \Longrightarrow \quad A\mathbf{x} = \lambda\mathbf{x}.$$

Gleichung $n+1$ fordert, dass $\|\mathbf{x}\|_2$ auf 1 normiert ist. Für einen Eigenvektor $\mathbf{x}$ mit $\|\mathbf{x}\|_2 = 1$ gilt

$$f(\mathbf{x}) = \mathbf{x}^T A\mathbf{x} = \mathbf{x}^T \lambda \mathbf{x} = \lambda.$$

Somit wird $f(\mathbf{x})$ durch die Eigenvektoren zum größten (kleinsten) Eigenwert maximiert (minimiert). Das wurde bereits im Rayleigh–Prinzip, Satz 43.11, gezeigt.

In der Informatik benötigt man dieses Resultat zum Beispiel bei der Bewegungsanalyse in digitalen Bildfolgen. □

**Beispiel 52.6** Man bestimme eine notwendige Bedingung für das Vorliegen von Extrema von

$$f(x, y, z) = 5x + y - 3z$$

auf dem Schnitt der Ebene

$$x + y + z = 0$$

mit der Kugeloberfläche

$$x^2 + y^2 + z^2 = 1.$$

Das ist ein Problem mit zwei Nebenbedingungen, man benötigt also zwei Langrange–Multiplikatoren in der Lagrange–Funktion

$$F(x, y, z, \lambda, \mu) = 5x + y - 3z + \lambda(x + y + z) + \mu(x^2 + y^2 + z^2 - 1).$$

Die notwendige Bedingung lautet

$$\mathbf{0} = \begin{pmatrix} F_x \\ F_y \\ F_z \\ F_\lambda \\ F_\mu \end{pmatrix} = \begin{pmatrix} 5 + \lambda + 2\mu x \\ 1 + \lambda + 2\mu y \\ -3 + \lambda + 2\mu z \\ x + y + z \\ x^2 + y^2 + z^2 - 1 \end{pmatrix}.$$

Nach einigen Umformungen sieht man, dass

$$\begin{pmatrix} x \\ y \\ z \\ \lambda \\ \mu \end{pmatrix} = \begin{pmatrix} 1/\sqrt{2} \\ 0 \\ -1/\sqrt{2} \\ -1 \\ -2\sqrt{2} \end{pmatrix} \quad \text{und} \quad \begin{pmatrix} x \\ y \\ z \\ \lambda \\ \mu \end{pmatrix} = \begin{pmatrix} -1/\sqrt{2} \\ 0 \\ 1/\sqrt{2} \\ 1 \\ 2\sqrt{2} \end{pmatrix}$$

dieses Gleichungssystem lösen. □

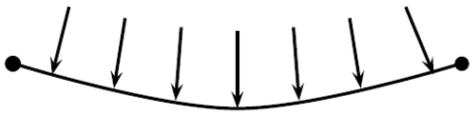
# Kapitel 53

# Variationsrechnung

**Bemerkung 53.1 Motivation.** Bisher wurden Optimierungsprobleme betrachtet, bei denen ein einzelner optimaler Wert (Maximum oder Minimum einer Funktion, eventuell mit Nebenbedingungen) gesucht war. In diesem Abschnitt befasst sich mit Problemen, bei denen eine gesamte Funktion gesucht wird, die ein Optimalitätskriterium erfüllt. Solche Probleme treten in der Praxis oft auf. □

**Beispiel 53.2 Typische Variationsprobleme.**

1. Welche Form nimmt ein Bücherbrett an, wenn es gleichmäßig belastet wird?



Die gesuchte Funktion minimiert ein Integral, in dem eine Biegeenergie auftritt. Ein verwandtes Problem ist die Crashtestsimulation.

2. Ein verrauschtes Signal $f(x)$, $x \in [0,1]$, soll so gefiltert werden, dass das resultierende Signal $u(x)$ zwar noch nahe an $f(x)$ liegt, aber nur noch geringe Schwankungen aufweist. Man sucht zum Beispiel $u(x)$ als Minimierer des sogenannten Energiefunktionals (ein Funktional ist eine Abbildung von einer Funktion nach $\mathbb{R}$)

$$E(u) = \frac{1}{2}\int_0^1 \left[\underbrace{(u(x)-f(x))^2}_{\text{Ähnlichkeit}} + \alpha \underbrace{(u'(x))^2}_{\text{Glattheit}}\right] dx.$$

Der freie Parameter $\alpha > 0$ heißt Regularisierungsparameter. Er steuert die Glattheit der Lösung. Je größer $\alpha$ ist, desto stärker werden Abweichungen von der Glattheit bestraft, das heißt desto glatter ist die Lösung.

□

**Beispiel 53.3 Diskretes Variationsproblem.** Das Signal aus Beispiel 53.2, 2. soll diskret vorliegen. Das bedeutet, es liegt nicht als Funktion $f(x)$ vor sondern es liegen nur die Funktionswerte in einer endlichen Anzahl von Punkten des (normierten) Definitionsbereiches $[0,1]$ vor, zum Beispiel auf einem äquidistanten Gitter,

$$f_0, f_1, \ldots, f_N \quad \text{mit} \quad f_i = f(ih),\ h = \frac{1}{N}, i = 0, \ldots, N.$$

Gesucht ist ein gefiltertes diskretes Signal $u_0, \ldots, u_N$, das die skalare Funktion mehrerer Variabler

$$E(u_0, \ldots, u_N) = \frac{1}{2}\sum_{i=0}^{N}(u_i - f_i)^2 + \frac{\alpha}{2}\sum_{i=0}^{N-1}\left(\frac{u_{i+1}-u_i}{h}\right)^2$$

minimiert. Die notwendige Bedingung für ein Minimum ist

$$\mathbf{0} \stackrel{!}{=} \left(\frac{\partial E}{\partial u_0}, \ldots, \frac{\partial E}{\partial u_N}\right)^T.$$

Diese Bedingung liest sich komponentenweise wie folgt

$$\begin{aligned}
0 &= \frac{\partial E}{\partial u_0} = (u_0 - f_0) - \alpha\frac{u_1 - u_0}{h^2}, \\
0 &= \frac{\partial E}{\partial u_k} = (u_k - f_k) - \alpha\left(\frac{u_{k+1}-u_k}{h^2} - \frac{u_k - u_{k-1}}{h^2}\right), \quad k = 1, \ldots, N-1, \\
0 &= \frac{\partial E}{\partial u_N} = (u_N - f_N) - \alpha\left(-\frac{u_N - u_{N-1}}{h^2}\right).
\end{aligned} \tag{53.1}$$

Das bedeutet, die unbekannten Werte $u_0, \ldots, u_N$ müssen folgendes lineares Gleichungssystem erfüllen

$$\begin{pmatrix}
1+\frac{\alpha}{h^2} & -\frac{\alpha}{h^2} & & & & 0 \\
-\frac{\alpha}{h^2} & 1+2\frac{\alpha}{h^2} & -\frac{\alpha}{h^2} & & & \\
& \ddots & \ddots & \ddots & & \\
& & \ddots & \ddots & \ddots & \\
& & & -\frac{\alpha}{h^2} & 1+2\frac{\alpha}{h^2} & -\frac{\alpha}{h^2} \\
0 & & & & -\frac{\alpha}{h^2} & 1+\frac{\alpha}{h^2}
\end{pmatrix}
\begin{pmatrix} u_0 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ u_N \end{pmatrix}
=
\begin{pmatrix} f_0 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ f_N \end{pmatrix}$$

Es gilt

$$\frac{u_{i+1}-u_i}{h^2} - \frac{u_i - u_{i-1}}{h^2} = \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2}.$$

Man kann mit Taylor–Entwicklung zeigen, dass dieser Differenzenquotient eine Approximation an die zweite Ableitung ist (Übungsaufgabe)

$$\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} = u''(ih) + \mathcal{O}(h^2), \quad u_i = u(ih).$$

Damit sieht (53.1) aus wie eine diskrete Form der Differentialgleichung

$$0 = u(x) - f(x) - \alpha u''(x) \quad \Longleftrightarrow \quad -\alpha u''(x) + u(x) = f(x).$$

Man kann zeigen, dass diese Differentialgleichung direkt aus dem kontinuierlichen Minimierungsproblem für

$$E(u) = \frac{1}{2}\int_0^1 \left[(u-f)^2(x) + \alpha\left(u'(x)\right)^2\right] dx$$

folgt. Das geschieht mit dem folgenden Satz. □

**Satz 53.4 Euler–Lagrange–Gleichung.** *Betrachte folgendes Variationsproblem: Finde eine differenzierbare Funktion $u : [a,b] \to \mathbb{R}$, die ein Funktional*

$$E(u) := \int_a^b F(x,u,u_x)\ dx$$

*minimiert und die Randbedingungen*

$$u(a) = \alpha, \quad u(b) = \beta$$

*erfüllt. Jede Lösung dieses Variationsproblems ist notwendigerweise Lösung der sogenannten Euler–Lagrange–Gleichung*

$$F_u - \frac{\mathrm{d}}{\mathrm{d}x} F_{u_x} = 0$$

*mit den Randbedingungen*

$$u(a) = \alpha, \quad u(b) = \beta.$$

**Beweis:** Man nimmt an, dass $u_0(x)$ eine differenzierbare Lösung des Variationsproblems ist und bettet $u_0(x)$ in eine Schar von Konkurrenzfunktionen ein

$$u(x,\varepsilon) := u_0(x) + \varepsilon h(x).$$

Hierbei ist $\varepsilon \in \mathbb{R}$ und die Funktionen $h(x)$ sollen stetig differenzierbar sein und am Rand verschwinden, das heißt $h(a) = h(b) = 0$. Da $u_0(x)$ das Integral $E(u)$ minimiert, besitzt die Funktion

$$g(\varepsilon) := E(u_0 + \varepsilon h)$$

in $\varepsilon = 0$ ein Minimum. Daher muss gelten

$$\begin{aligned} 0 &= g'(0) = \frac{\mathrm{d}}{\mathrm{d}\varepsilon} E(u_0 + \varepsilon h)\Big|_{\varepsilon=0} \\ &= \frac{\mathrm{d}}{\mathrm{d}\varepsilon} \int_a^b F(x, u_0 + \varepsilon h, u_{0x} + \varepsilon h_x)\ dx\Big|_{\varepsilon=0} \\ &\overset{\text{Kettenregel}}{=} \int_a^b \left[F_u(x,u_0,u_{0x})\ h(x) + F_{u_x}(x,u_0,u_{0x})\ h_x(x)\right] dx. \end{aligned}$$

Mit partieller Integration

$$\int_a^b F_{u_x}(x,u_0,u_{0x})\ h_x(x)\ dx = \underbrace{F_{u_x}(x,u_0,u_{0x})\ h(x)\big|_a^b}_{=0 \text{ da } h(a)=h(b)=0} - \int_a^b \frac{\mathrm{d}}{\mathrm{d}x}(F_{u_x}(x,u_0,u_{0x}))\ h(x)\ dx$$

folgt

$$0 = \int_a^b \left[F_u(x,u_0,u_{0x}) - \frac{\mathrm{d}}{\mathrm{d}x} F_{u_x}(x,u_0,u_{0x})\right] h(x)\ dx$$

für beliebige differenzierbare Funktionen $h(x)$. Das kann nur gelten, wenn

$$0 = F_u(x,u_0,u_{0x}) - \frac{\mathrm{d}}{\mathrm{d}x} F_{u_x}(x,u_0,u_{0x})$$

ist. ∎

**Bemerkung 53.5 Erweiterungen von Satz 53.4.** Man kann die folgenden Aussagen zeigen.

- *Keine Randvorgaben.* Hat das Variationsproblem keine Randbedingungen, so besitzen die Euler–Lagrange–Gleichungen die sogenannten natürlichen Randbedingungen
$$u'(a) = 0, \quad u'(b) = 0.$$

- *Funktionen mehrerer Variabler.* Die Variationsgleichung

$$E(u) = \int_{a_1}^{b_1} \dots \int_{a_n}^{b_n} F\big(x_1, \dots, x_n, u(x_1, \dots, x_n), u_{x_1}, \dots u_{x_n}\big)\ dx_n \dots dx_1$$

führt auf die Euler–Lagrange–Gleichung

$$\begin{aligned} 0 &= F_u(\mathbf{x}, u, u_{x_1}, \dots u_{x_n}) - \frac{\partial}{\partial x_1} F_{u_{x_1}}(\mathbf{x}, u, u_{x_1}, \dots u_{x_n}) - \dots \\ &\quad - \frac{\partial}{\partial x_n} F_{u_{x_n}}(\mathbf{x}, u, u_{x_1}, \dots u_{x_n}). \end{aligned}$$

- *Höhere Ableitungen.* Treten Ableitungen höherer Ordnung im zu minimierenden Funktional auf

$$E(u) = \int_a^b F(x, u, u_x, u_{xx}, \dots)\ dx,$$

so erhält man die Euler–Lagrange–Gleichungen

$$\begin{aligned} 0 &= F_u(x, u, u_x, u_{xx}, \dots) - \frac{\mathrm{d}}{\mathrm{d}x} F_{u_x}(x, u, u_x, u_{xx}, \dots) \\ &\quad + \frac{\mathrm{d}^2}{\mathrm{d}x^2} F_{u_{xx}}(x, u, u_x, u_{xx}, \dots) - + \dots\ . \end{aligned}$$

Das wechselnde Operationszeichen kommt durch die unterschiedliche Anzahl von partiellen Integrationen die man zur Herleitung dieser Terme anwendet.

- *Vektorwertige Funktionen.* Die Minimierung des Funktionals

$$E(u_1, \dots, u_m) = \int_a^b F(x, u_1, \dots, u_m, u_{1x}, \dots, u_{mx})\ dx$$

führt auf ein System von Euler–Lagrange–Gleichungen

$$\begin{aligned} 0 &= F_{u_1}(x, u_1, \dots, u_m, u_{1x}, \dots, u_{mx}) - \frac{\mathrm{d}}{\mathrm{d}x} F_{u_{1x}}(x, u_1, \dots, u_m, u_{1x}, \dots, u_{mx}) \\ &\vdots \\ 0 &= F_{u_m}(x, u_1, \dots, u_m, u_{1x}, \dots, u_{mx}) - \frac{\mathrm{d}}{\mathrm{d}x} F_{u_{mx}}(x, u_1, \dots, u_m, u_{1x}, \dots, u_{mx}). \end{aligned}$$

□

**Beispiel 53.6 Zusammenhang zwischen den Beispielen 53.2 und 53.3.** Betrachte Beispiel 53.2, 2, das heißt es ist

$$F(x, u, u_x) = \frac{1}{2}\left(u(x) - f(x)\right)^2 + \frac{\alpha}{2}\left(u_x(x)\right)^2.$$

Durch partielle Differentiation erhält man

$$F_u(x, u, u_x) = u(x) - f(x), \quad F_{u_x}(x, u, u_x) = \alpha u_x(x).$$

und somit nach Satz 53.4 und Bemerkung 53.5 die Euler–Lagrange–Gleichung

$$0 = F_u(x, u, u_x) - \frac{\mathrm{d}}{\mathrm{d}x} F_{u_x}(x, u, u_x) = u(x) - f(x) - \alpha u_{xx}(x) = u(x) - f(x) - \alpha u''(x)$$

mit den Randbedingungen

$$u'(0) = u'(1) = 0.$$

Das diskrete Variationsbeispiel 53.3 stellt ein Diskretisierungsverfahren für diese Differentialgleichung dar. □

**Beispiel 53.7 Zweidimensionales Variationsproblem.** Das Entrauschen von Bildern führt auf zweidimensionale Variationsprobleme, da man dort das Energiefunktional

$$E(u) = \frac{1}{2} \int_{a_1}^{a_2} \int_{b_1}^{b_2} \big[ (u-f)^2 + \alpha \underbrace{(u_x^2 + u_y^2)}_{\|\nabla u\|_2^2} \big](x,y) \; dx \; dy$$

minimiert. Man erhält die Euler–Lagrange–Gleichung

$$0 = (u-f) - \alpha \underbrace{(u_{xx} + u_{yy})}_{\Delta u} \quad \Longleftrightarrow \quad -\alpha \Delta u + u = f$$

Nummeriert man im diskreten Fall die Punkte zeilen- oder spaltenweise durch, entsteht ein großes lineares Gleichungssystem, das schwach besetzt ist (maximal 5 Einträge pro Zeile/Spalte)

$$\begin{pmatrix} \ddots & \ddots & & & \ddots & & 0 \\ \ddots & \ddots & \ddots & & & \ddots & \\ & \ddots & \ddots & \ddots & 0 & & \ddots \\ & & \ddots & \ddots & \ddots & & \\ \ddots & & 0 & \ddots & \ddots & \ddots & \\ & \ddots & & & \ddots & \ddots & \ddots \\ 0 & & \ddots & & & \ddots & \ddots \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ u_N \end{pmatrix} = \begin{pmatrix} f_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ f_N \end{pmatrix}$$

Zur Lösung dieses Systems bieten sich iterative Verfahren, zum Beispiel die aus Abschnitt 46.2, an. □

# Kapitel 54

# Zur Integration in mehreren Dimensionen

**Bemerkung 54.1 Motivation.** Bisher wurden viele Konzepte von Funktionen einer Variablen auf Funktionen mehrerer Variablen verallgemeinert: Differenzierbarkeit, Mittelwertsatz, Satz von Taylor, Extrema von Funktionen. Dieser Abschnitt wird sich mit der Integration in mehreren Dimensionen und dort speziell mit der Verallgemeinerung der Substitutionsregel befassen. Diese Verallgemeinerung wird Transformationsregel genannt zu ihrer Herleitung muss auch der Begriff der Umkehrbarkeit einer Funktion verallgemeinert werden. □

**Bemerkung 54.2 Erinnerung: Umkehrbarkeit von Funktionen einer Variablen.** Die Umkehrbarkeit einer Funktion $f : (a,b) \to \mathbb{R}$ mit $f \in C^1(a,b)$ wurde im Satz 21.7 behandelt. Sei $\xi \in (a,b)$ mit $f'(\xi) \neq 0$. Dann ist $f(x)$ streng monoton in einer Umgebung von $\xi$ und $f(x)$ lässt sich in dieser Umgebung invertieren. Dann existiert in einer Umgebung von $\eta = f(\xi)$ eine stetig differenzierbare Umkehrfunktion $g(y)$ mit

$$g'(\eta) = \frac{1}{f'(\xi)}.$$

□

Für vektorwertige Funktionen mehrerer Variabler kann man das folgende Resultat zeigen.

**Satz 54.3 Umkehrsatz.** *Seien $D \subset \mathbb{R}^n$ offen und $\mathbf{f} : D \to \mathbb{R}^n$ eine stetig differenzierbare Funktion. Ferner sei $\boldsymbol{\xi} \in D$. Ist die Jacobi–Matrix $J\mathbf{f}(\boldsymbol{\xi})$ invertierbar, so ist $\mathbf{f}(\mathbf{x})$ lokal umkehrbar. Das bedeutet, es gibt offene Umgebungen $U_1$ von $\boldsymbol{\xi}$ und $U_2$ von $\boldsymbol{\eta} = \mathbf{f}(\boldsymbol{\xi})$, so dass $\mathbf{f}(\mathbf{x})$ die Menge $U_1$ bijektiv auf $U_2$ abbildet. Die Umkehrfunktion $\mathbf{g} = \mathbf{f}^{-1} : U_2 \to U_1$ ist stetig differenzierbar uns es gilt*

$$J\mathbf{g}(\boldsymbol{\eta}) = (J\mathbf{f}(\boldsymbol{\xi}))^{-1}.$$

**Beweis:** Beweisidee. Das Ableitungsformel folgt aus $\mathbf{g} \circ \mathbf{f} = \mathrm{id}_{U_1}$ (identische Abbildung auf $U_1$) mit Hilfe der Kettenregel, Satz 49.23

$$J\mathbf{g}(\boldsymbol{\eta}) J\mathbf{f}(\boldsymbol{\xi}) = I.$$

Die lokale Existenz einer stetig differenzierbaren Umkehrfunktion $\mathbf{g}(\mathbf{y})$ erfordert ein aufwändiges Resultat, das den Rahmen der Vorlesung sprengt, den Satz über implizite Funktionen. ■

**Bemerkung 54.4**

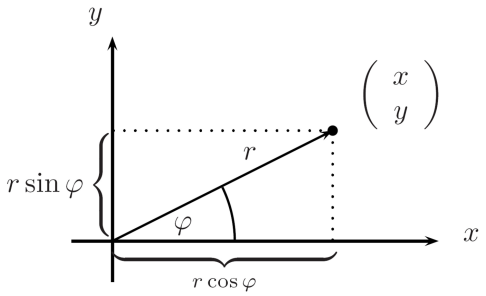
- Wegen der Umkehrbarkeit müssen sowohl Definitionsbereich als auch Wertebereich im $\mathbb{R}^n$ liegen.
- Bijektive stetig differenzierbare Funktion, deren Umkehrung ebenfalls stetig differenzierbar ist, heißen auch $C^1$–Diffeomorphismen.
- Der Umkehrsatz besagt, dass stetig differenzierbare Funktionen mit regulärer Jacobi–Matrix lokal umkehrbar sind. Solche Abbildungen kommen in den Anwendungen oft bei Koordinatentransformationen vor.

□

**Beispiel 54.5 Polarkoordinaten.** Einen Vektor

$$\begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2, \quad \begin{pmatrix} x \\ y \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

kann man durch seinen Abstand $r$ zum Ursprung und seinen Winkel $\varphi$ zur positiven $x$–Achse eindeutig charakterisiert werden.




Die Funktion $\mathbf{f} : (0,\infty) \times \mathbb{R} \to \mathbb{R}^2$

$$\begin{pmatrix} r \\ \varphi \end{pmatrix} \mapsto \begin{pmatrix} f_1(r,\varphi) \\ f_2(r,\varphi) \end{pmatrix} = \begin{pmatrix} r\cos\varphi \\ r\cos\varphi \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix}$$

besitzt die Jacobi–Matrix

$$J\mathbf{f}(r,\varphi) = \begin{pmatrix} \dfrac{\partial f_1}{\partial r} & \dfrac{\partial f_1}{\partial \varphi} \\ \dfrac{\partial f_2}{\partial r} & \dfrac{\partial f_2}{\partial \varphi} \end{pmatrix} = \begin{pmatrix} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{pmatrix}.$$

Da

$$\det(J\mathbf{f}(r,\varphi)) = r\cos^2\varphi + r\sin^2\varphi = r > 0,$$

ist $\mathbf{f}(r,\varphi)$ auf $(0,\infty)\times\mathbb{R}$ lokal invertierbar. Für $r \in (0,\infty)$ und $\varphi \in (-\pi/2, \pi/2)$ lautet die Umkehrfunktion

$$\mathbf{g} : \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} \sqrt{x^2+y^2} \\ \arctan\left(\dfrac{y}{x}\right) \end{pmatrix} = \begin{pmatrix} r \\ \varphi \end{pmatrix}.$$

Sie hat die Jacobi–Matrix

$$\begin{aligned} J\mathbf{g}(x,y) &= \begin{pmatrix} \dfrac{\partial g_1}{\partial x} & \dfrac{\partial g_1}{\partial y} \\ \dfrac{\partial g_2}{\partial x} & \dfrac{\partial g_2}{\partial y} \end{pmatrix} = \begin{pmatrix} \dfrac{x}{\sqrt{x^2+y^2}} & \dfrac{y}{\sqrt{x^2+y^2}} \\ \dfrac{-y}{x^2+y^2} & \dfrac{x}{x^2+y^2} \end{pmatrix} \\ &= \begin{pmatrix} \dfrac{r\cos\varphi}{r} & \dfrac{r\sin\varphi}{r} \\ \dfrac{-r\sin\varphi}{r^2} & \dfrac{r\cos\varphi}{r^2} \end{pmatrix} = \begin{pmatrix} \cos\varphi & \sin\varphi \\ \dfrac{-\sin\varphi}{r} & \dfrac{\cos\varphi}{r} \end{pmatrix}. \end{aligned}$$

Damit gilt

$$\begin{aligned} J\mathbf{g}(x,y)J\mathbf{f}(r,\varphi) &= \begin{pmatrix} \cos^2\varphi+\sin^2\varphi & -r\cos\varphi\sin\varphi+r\cos\varphi\sin\varphi \\ \dfrac{\sin\varphi\cos\varphi}{r}+\dfrac{\sin\varphi\cos\varphi}{r} & \sin^2\varphi+\cos^2\varphi \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \end{aligned}$$

das heißt $J\mathbf{g}(x,y) = (J\mathbf{f})^{-1}(r,\varphi)$.

Man beachte, dass $\mathbf{f}(r,\varphi)$ den Bereich $(0,\infty)\times\mathbb{R}$ auf $\mathbb{R}^2 \setminus \{(0,0)^T\}$ abbildet. Wenn man den Wertebereich dementsprechend einschränkt, hat man eine surjektive Abbildung. Die Abbildung ist aber nicht injektiv, da

$$\mathbf{f}(r,\varphi) = \mathbf{f}(r,\varphi+2k\pi) \quad \text{für alle } k\in\mathbb{Z}.$$

Damit ist die Funktion nicht bijektiv und nicht global umkehrbar. □

Die Jacobi–Matrix von Koordinatentransformationen wird unter anderem beim Berechnen von Mehrfachintegralen benötigt. Hier verwendet man oft die Transformationsregel, eine Verallgemeinerung der Substitutionsregel.

**Bemerkung 54.6 Erinnerung: Substitutionsregel.** Die Substitutionsregel für die Berechnung bestimmter Integrale

$$\int_c^d f(x)\ dx$$

wurde in Satz 26.8 bewiesen. Bei dieser Regel setzt man $x = g(t)$ mit einer streng monotonen Funktion $g\in C^1([a,b])$ mit $g(a)=c$, $g(b)=d$ und formal $\mathrm{d}x = g'(t)\mathrm{d}t$. Man führt damit eine Koordinatentransformation vom Intervall durch. Es gilt

$$\int_{x=c}^{x=d} f(x)\ dx = \int_{t=a}^{t=b} f(g(t))g'(t)\ dt.$$

Ist $c<d$, dann ist $g(t)$ wegen $c=g(a)<g(b)=d$ streng monoton wachsend. Dann kann man im linken Integral $g'(t)$ durch $|g'(t)|$ ersetzen. Im Fall $c>d$ ist $g(t)$ streng monoton fallend und man hat $g'(t)<0$. Es folgt

$$\begin{aligned} \int_{x=c}^{x=d} f(x)\ dx &= -\int_{t=b}^{t=a} f(g(t))g'(t)\ dt = \int_{t=b}^{t=a} f(g(t))(-g'(t))\ dt \\ &= \int_{t=b}^{t=a} f(g(t))\,|g'(t)|\ dt, \end{aligned}$$

womit nach der Transformation die natürliche Situation ist, dass die untere Integrationsgrenze kleiner als die obere ist. □

**Bemerkung 54.7 Transformationsregel.** Die Verallgemeinerung der Substitutionsregel auf Integrale über mehrdimensionale Gebiete nennt man Transformationsregel. Seien $D\subset\mathbb{R}^n$ und $f : D\to\mathbb{R}$. Zu berechnen ist das Integral

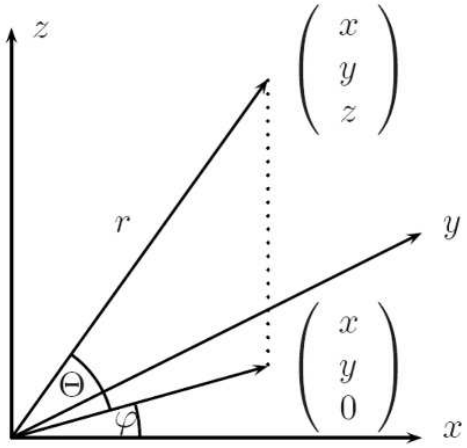
$$\int_D f(\mathbf{x})\ d\mathbf{x}.$$

Es existiere eine Menge $E\subset\mathbb{R}^n$ und ein $C^1$–Diffeomorphismus $\mathbf{g} : E\to D$. Man setzt $\mathbf{x}=\mathbf{g}(\mathbf{t})$. Unter geeigneten technischen Voraussetzungen, auf die hier nicht näher eingegangen wird, gilt dann

$$\int_{\mathbf{x}\in D} f(\mathbf{x})\ d\mathbf{x} = \int_{\mathbf{t}\in\mathbf{g}^{-1}(D)} f(\mathbf{g}(\mathbf{t}))\,|\det(J\mathbf{g}(\mathbf{t}))|\ d\mathbf{t}.$$

In einer Dimension ist gerade $|\det(J\mathbf{g}(\mathbf{t}))| = |g'(t)|$. □

**Beispiel 54.8 Integration in Kugelkoordinaten.** Kugelkoordinaten sind die dreidimensionale Verallgemeinerung von Polarkoordinaten.




Der Zusammenhang zwischen kartesischen und Kugelkoordinaten ist

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} r\cos\varphi\cos\theta \\ r\sin\varphi\cos\theta \\ r\sin\theta \end{pmatrix} = \mathbf{g}(r,\varphi,\theta).$$

Damit ist die Jacobi–Matrix

$$\begin{aligned} J\mathbf{g}(r,\varphi,\theta) &= \begin{pmatrix} \frac{\partial g_1}{\partial r} & \frac{\partial g_1}{\partial \varphi} & \frac{\partial g_1}{\partial \theta} \\ \frac{\partial g_2}{\partial r} & \frac{\partial g_2}{\partial \varphi} & \frac{\partial g_2}{\partial \theta} \\ \frac{\partial g_3}{\partial r} & \frac{\partial g_3}{\partial \varphi} & \frac{\partial g_3}{\partial \theta} \end{pmatrix} \\ &= \begin{pmatrix} \cos\varphi\cos\theta & -r\sin\varphi\cos\theta & -r\cos\varphi\sin\theta \\ \sin\varphi\cos\theta & r\cos\varphi\cos\theta & -r\sin\varphi\sin\theta \\ \sin\theta & 0 & r\cos\theta \end{pmatrix}. \end{aligned}$$

Für die Determinante erhält man

$$\begin{aligned} \det(J\mathbf{g}) &= \sin\theta \begin{vmatrix} -r\sin\varphi\cos\theta & -r\cos\varphi\sin\theta \\ r\cos\varphi\cos\theta & -r\sin\varphi\sin\theta \end{vmatrix} \\ &\quad + r\cos\theta \begin{vmatrix} \cos\varphi\cos\theta & -r\sin\varphi\cos\theta \\ \sin\varphi\cos\theta & r\cos\varphi\cos\theta \end{vmatrix} \\ &= \sin\theta\left(r^2\sin^2\varphi\sin\theta\cos\theta + r^2\cos^2\varphi\sin\theta\cos\theta\right) \\ &\quad + r\cos\theta\left(r\cos^2\varphi\cos^2\theta + r\sin^2\varphi\cos^2\theta\right) \\ &= \sin\theta r^2\sin\theta\cos\theta + r\cos\theta r\cos^2\theta \\ &= r^2\cos\theta. \end{aligned}$$

Diese Informationen werden jetzt dazu benutzt, das Volumen eines Kugeloktanden mit Radius $R$ zu berechnen

$$K = \left\{ \begin{pmatrix} x \\ y \\ z \end{pmatrix} \in \mathbb{R}^3 \;\middle|\; x^2+y^2+z^2 \le R^2 \text{ und } x,y,z \ge 0 \right\}.$$

Es ist

$$\begin{aligned}
\int_{(x,y,z)\in K} dx\ dy\ dz &= \int_{(r,\varphi,\theta)\in \mathbf{g}^{-1}(K)} |\det(J\mathbf{g})|\ dr\ d\varphi\ d\theta \\
&= \int_{\theta=0}^{\pi/2} \int_{\varphi=0}^{\pi/2} \int_{r=0}^{R} r^2 \cos\theta\ dr\ d\varphi\ d\theta \\
&= \int_{\theta=0}^{\pi/2} \int_{\varphi=0}^{\pi/2} \frac{R^3}{3} \cos\theta\ d\varphi\ d\theta \\
&= \int_{\theta=0}^{\pi/2} \frac{R^3}{3}\frac{\pi}{2} \cos\theta\ d\theta \\
&= \frac{1}{6}\pi R^3 \sin\theta\Big|_0^{\pi/2} \\
&= \frac{1}{6}\pi R^3.
\end{aligned}$$

Da das Kugelvolumen bekanntermaßen $4\pi R^3/3$ ist, stimmt das Ergebnis. □

# Teil VI

# Stochastik

Stochastik (aus dem Griechischen: vermuten, erwarten) ist die Mathematik des Zufalls. Sie beinhaltet die Lehre von Wahrscheinlichkeiten und die Statistik.

# Kapitel 55

# Grundbegriffe

**Bemerkung 55.1 Motivation.** Die Stochastik ist von großer Bedeutung in der Informatik:

- Analyse der Auslastung von Datennetzen,
- Modellierung von Antwortzeiten im Rechner,
- Zuverlässigkeitsanalyse von Hardware,
- Raytracing in der Computergrafik (Monte–Carlo–Methoden),
- stochastische Optimierungsalgorithmen (genetische Algorithmen, simulated annealing),
- Analyse der mittleren Laufzeit von Algorithmen,
- Kombinatorische Probleme in der Bioinformatik.

□

**Bemerkung 55.2 Gebiete der Stochastik.** Die Stochastik gliedert sich in zwei Gebiete.

- *Wahrscheinlichkeitstheorie.* Ausgehend von einem stochastischen Modell werden Wahrscheinlichkeiten berechnet.
  Beispiel: Die Modellannahme ist, dass beim Wurf eines Würfels jede der Augenzahlen $1, \ldots, 6$ die Wahrscheinlichkeit $1/6$ hat. Eine Folgerung ist, dass die Wahrscheinlichkeit für Augenzahl 1 oder 3 geich $1/6 + 1/6 = 1/3$ ist.
- *Statistik.* Ausgehend von realen Daten oder Messungen zieht man Schlussfolgerungen.
  Beispiele:
  - Man möchte eine möglichst gute Approximationskurve durch fehlerbehaftete Messwerte legen.
  - Man untersucht Hypothesen (Hypothesentest), zum Beispiel die Hypothese, dass ein neues Medikament wirksam ist.

□

**Definition 55.3 Wahrscheinlichkeit.** Die Wahrscheinlichkeit eines Ereignisses beschreibt die zu erwartende relative Häufigkeit, mit der dieses Ereignis eintritt, wenn man den zu Grunde liegenden Prozess immer wieder unter den gleichen Bedingungen wiederholt. □

**Beispiel 55.4**

1. Bei einer fairen Münze beträgt die Wahrscheinlichkeit, Zahl zu werfen, 1/2.
2. Bei einem fairen Würfel beträgt die Wahrscheinlichkeit, eine 6 zu würfeln, 1/6.

3. Man kann aber auch unendlich viele mögliche Ausgänge des Experiments haben. Zum Beispiel kann man zufällig eine Zahl aus $[0, 1]$ wählen und dann fragen, wie hoch die Wahrscheinlichkeit ist, dass diese Zahl rational ist.
   Die Wahrscheinlichkeitsrechnung wird als korrekte Antwort geben, dass die Wahrscheinlichkeit gleich Null ist. Das hängt mit der Mächtigkeit von abzählbaren und überabzählbaren Mengen zusammen, siehe Kapitel 5.2. Das Beispiel zeigt aber auch, dass bei unendlich vielen Ausgängen die Wahrscheinlichkeit zwar Null sein kann, aber das Ereignis durchaus eintreten kann.

□

**Definition 55.5 Laplace–Experiment.** Ein Experiment heißt Laplace–Experiment, wenn es endlich viele, einander ausschließende Ausgänge hat, die alle gleich wahrscheinlich sind. □

**Beispiel 55.6**

1. Der Wurf eines Würfels hat 6 gleich berechtigte Ausgänge. Dies ist ein Laplace–Experiment.
2. Fällt ein Marmeladenbrot zu Boden, gibt es zwei Ausgänge, die nach Murphy nicht gleich berechtigt sind. Falls dies stimmt, ist es kein Laplace–Experiment.

□

Zur Beschreibung von Zufallsexperimenten benötig man noch einige Begriffe.

**Definition 55.7 Ereignismenge, Ereignis, Wahrscheinlichkeitsverteilung.** Gegeben sei ein Zufallsexperiment mit endlich vielen Ausgängen.

1. Die Ereignismenge $\Omega$ (Stichprobenraum, Grundraum) ist eine endliche, nicht leere Menge, deren Elemente $\omega_i$ die Versuchsausgänge beschreiben.
2. Ein Ereignis ist eine Teilmenge von $\Omega$.
3. Die Wahrscheinlichkeitsverteilung oder das Wahrscheinlichkeitsmaß ist eine Abbildung $P$ von der Potenzmenge $\mathcal{P}(\Omega)$ (Definition 1.4) nach $\mathbb{R}$ mit folgenden Eigenschaften:
   1. Normiertheit: $P(\Omega) = 1$,
   2. Nichtnegativität: $P(A) \geq 0$ für alle $A \in \mathcal{P}(\Omega)$,
   3. Additivität: $P(A \cup B) = P(A) + P(B)$ für alle disjunkten $A, B \in \mathcal{P}(\Omega)$.

□

**Beispiel 55.8** Beim Würfelexperiment ist $\Omega = \{1, 2, \ldots, 6\}$. Ein Ereignis $A$ sind beispielsweise Würfe, bei denen 2 oder 5 gewürfelt wird: $A = \{2, 5\}$. □

**Folgerung 55.9** *Der Wertebereich von $P$ liegt in* $[0, 1]$.

**Beweis:** Sei $A \in \mathcal{P}(\Omega)$, dann gilt

$$1 = P(\Omega) = P(\Omega \setminus A) + P(A).$$

Da beide Summanden nichtnegativ sind, beweist dies die Behauptung. ■

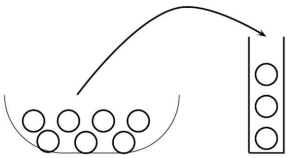
# Kapitel 56

# Kombinatorik

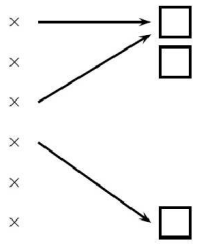
**Bemerkung 56.1 Motivation.**Die Kombinatorik liefert wichtige Modelle zum Berechnen von Wahrscheinlichkeiten bei Laplace–Experimenten. Sie spielt eine grundlegende Rolle in der Informatik. □

**Bemerkung 56.2 Zwei äquivalente Modelle.**

- *Stichprobensprechweise, Urnenmodell.* Aus einer Urne mit $n$ unterscheidbaren Kugeln werden $k$ Kugeln gezogen. Dabei kann das Ziehen mit oder ohne Zurücklegen erfolgen, und die Reihenfolge eine oder keine Rolle spielen.



- *Zuordnungssprechweise, Schubladenmodell.* Bei diesem Modell werden $k$ Objekte auf $n$ Schubladen verteilt. Dabei sind die Objekte entweder unterscheidbar oder nicht unterscheidbar, und die Schubladen dürfen einfach oder mehrfach besetzt werden.



Urnen– und Schubladenmodell sind äquivalent:

| Urnenmodell | Schubladenmodell |
|---|---|
| mit / ohne Zurücklegen | mit / ohne Mehrfachbesetzung |
| mit / ohne Reihenfolge | unterscheidbare / ununterscheidbare Objekte |

□

**Bemerkung 56.3 Produktregel der Kombinatorik.** Bei einem $k$–stufigen Experiment habe der Ausgang einer Stufe keinen Einfluss auf die Anzahl der möglichen Ausgänge bei späteren Stufen. Haben die einzelnen Stufen $n_1, \ldots, n_k$ Ausgänge, so hat das Gesamtexperiment $n_1 \ldots n_k$ Ausgänge. Die Produktregel ist wichtig bei der Beschreibung der vier kombinatorischen Grundsituationen. □

**Bemerkung 56.4 Die vier kombinatorischen Grundsituationen.**

1. *Geordnete Stichprobe mit Wiederholung.* Bei $k$ Ziehungen mit Zurücklegen aus einer Urne mit $n$ Objekten, wobei die Reihenfolge eine Rolle spielt, gibt es
$$\underbrace{n}_{\text{1. Ziehung}} \quad \underbrace{n}_{\text{2. Ziehung}} \quad \dots \quad \underbrace{n}_{\text{k–te Ziehung}} \quad = n^k$$
Möglichkeiten.
2. *Geordnete Stichprobe ohne Wiederholung.* Betrachte das Urnenmodell: $k$ Ziehungen aus $n$ Objekten ohne Zurücklegen, aber mit Berücksichtigung der Reihenfolge. Die Anzahl der Möglichkeiten ist
$$\underset{\substack{\uparrow\\ \text{1. Ziehung}}}{n}(n-1)(n-2)\dots \quad \underset{\substack{\uparrow\\ k\text{–te Ziehung}}}{(n-k+1)} = \frac{n!}{(n-k)!}.$$
Im Spezialfall $n = k$ hat man $n!$ Möglichkeiten.
Für große $n$ kann man $n!$ mit der Stirling[1]–Formel approximieren
$$n! \approx \sqrt{2\pi n}\, n^n\, e^{-n}.$$
3. *Ungeordnete Stichprobe ohne Wiederholung.* Die Situation ist ähnlich wie im Fall 2, jedoch müssen die $k!$ Permutationen für $k$ gezogene Elemente miteinander identifiziert werden. Daher gibt es nur
$$\frac{n!}{(n-k)!k!} = \binom{n}{k}$$
Möglichkeiten.
4. *Ungeordnete Stichprobe mit Wiederholung.* Hier ist das Schubladenmodell hilfreich. Es sollen $k$ nicht unterscheidbare Objekte in $n$ Schubladen verstaut werden, wobei Mehrfachbesetzung möglich ist:
$$\underbrace{\circ\,\circ\,\circ}_{\text{1. Schublade}} \Big| \underbrace{\circ}_{\text{2. Schublade}} \Big| \dots \Big| \underbrace{\circ\,\circ}_{n\text{–te Schublade}}$$
Der Gesamtzustand wird beschrieben durch die möglichen Aneinanderreihungen von $k$ Punkten und $n-1$ Trennungsstrichen, das sind $(k+n-1)!$ Möglichkeiten. Bei jeder Verteilung von Objekten und Trennungsstrichen kann man die Objekte permutieren ($k!$) und die Trennungsstrichen ($(n-1)!$). Diese Permutationen muss man noch identifizieren, womit man letztlich
$$\frac{(k+n-1)!}{k!(n-1)!} = \binom{k+n-1}{k} = \binom{k+n-1}{n-1}$$
Möglichkeiten erhält.

□

**Beispiel 56.5**

zu 1. Ein Vater will seinen ungezogenen Sohn mit 10 Ohrfeigen bestrafen. Auf wie viele Arten kann er das tun, wenn er bei jedem Schlag zwei Möglichkeiten hat (rechts oder links)? Es gibt $2^{10} = 1024$ Möglichkeiten.

zu 2. Ein Vater will seine 5 Kinder in einer Reihe anordnen für eine Gruppenaufnahme. Es gibt $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$ Möglichkeiten.

[1] James Stirling (1692 – 1770)

zu 3. Beim Lottoschein gibt es

$$\binom{49}{6} = \frac{49 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6} = 13\ 983\ 816$$

Möglichkeiten, 6 der 49 Zahlen anzukreuzen. Die Wahrscheinlichkeit, 6 Richtige zu tippen, ist daher $\frac{1}{13\ 983\ 816} \approx 7 \cdot 10^{-8}$.

zu 4. Auf wie viele Arten können 60 Parlamentssitze auf 3 Parteien verteilt werden? In diesem Beispiel sind $k = 60$ und $n = 3$. Man erhält

$$\binom{62}{60} = \binom{62}{2} = \frac{62 \cdot 61}{2 \cdot 1} = 1891$$

Möglichkeiten.

□

# Kapitel 57

# Erzeugende Funktionen

**Bemerkung 57.1 Motivation.** Erzeugende Funktionen wirken auf den ersten Blick etwas abstrakt, aber sie sind ein wichtiges Werkzeug, um kombinatorische Probleme systematischer und eleganter zu lösen. Sie sind zudem in verschiedenen anderen Gebieten der Stochastik nützlich. □

**Definition 57.2 Permutation, Kombination.**Betrachte eine $k$–elementige Stichproben einer $n$–elementigen Menge. Ist die Stichprobe geordnet, so nennt man sie $k$–Permutation, ist sie ungeordnet, wird sie $k$–Kombination genannt. □

**Definition 57.3 Erzeugende Funktion.** Eine Funktion $f : \mathbb{R} \to \mathbb{R}$ mit

$$f(x) = \sum_{k=0}^{n} a_k x^k$$

heißt erzeugende Funktion für die Koeffizienten $a_k$. Lassen sich die Zahlen $a_k$ mit einem kombinatorischen Problem identifizieren, so nennt man $f(x)$ die erzeugende Funktion für dieses Problem. □

**Beispiel 57.4 Erzeugende Funktion.** Nach dem Binomialsatz, Satz 18.7, gilt

$$(x+1)^n = \sum_{k=0}^{n} \binom{n}{k} x^k 1^{n-k} = \sum_{k=0}^{n} \binom{n}{k} x^k.$$

Der Koeffizient $\binom{n}{k}$ von $x^k$ beschreibt die Anzahl der $k$–Kombinationen einer $n$–elementigen Menge ohne Wiederholung, vergleich Beispiel 56.5, 3. In den Koeffizienten der Funktion

$$f(x) = (x+1)^n = \binom{n}{0} x^0 + \binom{n}{1} x^1 + \ldots + \binom{n}{n} x^n$$

stecken somit alle Informationen über dieses kombinatorische Problem.

Betrachte die Schubladeninterpretation. Jeder der $n$ Faktoren $(1+x)$ wird als Schublade aufgefasst, in die null oder ein Element passt. Wählt man beim Ausmultiplizieren von $(1+x)$ den Faktor 1, bleibt die Schublade leer. Wählt man $x$, wird sie mit einem Element besetzt. Beispielsweise beschreibt

$$f(x) = (1+x)^3 = 1 + 3x + 3x^2 + x^3$$

alle Kombinationen um $0, 1, 2$ oder 3 Elemente auf 3 Schubladen zu verteilen. Dabei bedeuten die Summanden:

$$
\begin{array}{lcll}
1 & = & 1 \cdot 1 \cdot 1 & \text{1 Möglichkeit bei 0 Elementen,} \\
3x & = & x \cdot 1 \cdot 1 + 1 \cdot x \cdot 1 + 1 \cdot 1 \cdot x & \text{3 Möglichkeiten bei 1 Element,} \\
3x^2 & = & x \cdot x \cdot 1 + x \cdot 1 \cdot x + 1 \cdot x \cdot x & \text{3 Möglichkeiten bei 2 Elementen,} \\
x^3 & = & x \cdot x \cdot x & \text{1 Möglichkeit bei 3 Elementen.}
\end{array}
$$

□

**Beispiel 57.5 Erzeugende Funktion: Verallgemeinerung auf mehrere Objekte.** Lässt man pro Schublade bis zu zwei Objekte zu, lautet der Faktor $(1+x+x^2)$ statt $(1+x)$. Zum Beispiel kann man die Anzahl der Kombinationen mit Wiederholung aus einer 2–elementigen Menge bestimmen, wenn jedes Element 0–, 1– oder 2–mal ausgewählt werden kann:

$$
\begin{aligned}
f(x) &= (1+x+x^2)^2 = (1+x+x^2)(1+x+x^2) \\
&= 1 \cdot 1 + 1 \cdot x + 1 \cdot x^2 + x \cdot 1 + x \cdot x + x \cdot x^2 + x^2 \cdot 1 + x^2 \cdot x + x^2 \cdot x^2 \\
&= 1 + 2x + 3x^2 + 2x^3 + x^4.
\end{aligned}
$$

Damit folgt für den Fall, dass man zwei Schubladen hat und in jede Schublade höchstens zwei Objekte hinein dürfen:

| | |
|---|---|
| Es gibt hier 1 Möglichkeit, | 0 Objekte zu verteilen. |
| Es gibt hier 2 Möglichkeiten, | 1 Objekt zu verteilen. |
| Es gibt hier 3 Möglichkeiten, | 2 Objekte zu verteilen. |
| Es gibt hier 2 Möglichkeiten, | 3 Objekte zu verteilen. |
| Es gibt hier 1 Möglichkeit, | 4 Objekte zu verteilen. |

□

**Beispiel 57.6 Erzeugende Funktion: unterschiedliche Beschränkungen für die Elemente.** Man bestimme die Kombinationen einer 4–elementigen Menge $\{x_1, \ldots, x_4\}$ mit den folgenden Beschränkungen:

| Element | Beschränkung | Polynom |
|---|---|---|
| $x_1$ | 0–, 1– oder 3–mal | $1 + x + x^3$, |
| $x_2$ | 1– oder 2–mal | $x + x^2$, |
| $x_3$ | 1–mal | $x$, |
| $x_4$ | 0– oder 4–mal | $1 + x^4$. |

Die erzeugende Funktion ist

$$
\begin{aligned}
f(x) &= (1+x+x^3)(x+x^2)x(1+x^4) \\
&= x^2 + 2x^3 + x^4 + x^5 + 2x^6 + 2x^7 + x^8 + x^9 + x^{10}.
\end{aligned}
$$

Damit gibt es beispielsweise zwei Kombinationen mit 6 Elementen: $(3, 2, 1, 0)$ und $(0, 1, 1, 4)$ als die Anzahl für $\{x_1, \ldots, x_4\}$. □

**Satz 57.7 Kombinationen mit vorgegebenen Wiederholungen.** *Die erzeugende Funktion der Kombinationen mit Wiederholung aus einer $n$–elementigen Menge $\{x_1, \ldots, x_n\}$, in der $x_i$ in den Anzahlen $v_1^{(i)}, \ldots, v_{k_i}^{(i)}$ auftreten darf, ist gegeben durch*

$$
f(x) = \prod_{i=1}^{n} \left( x^{v_1^{(i)}} + x^{v_2^{(i)}} + \ldots + x^{v_{k_i}^{(i)}} \right).
$$

**Beweis:** Siehe Literatur. ■

**Satz 57.8 Kombinationen mit beliebigen Wiederholungen.** *Die erzeugende Funktion der Kombinationen mit beliebigen Wiederholungen von n Elementen lautet*

$$f(x) = \frac{1}{(1-x)^n} = \left(\sum_{i=0}^{\infty} x^i\right)^n.$$

**Beweis:** Siehe Literatur. Die obige Formel nutzt die Summe der geometrischen Reihe

$$\sum_{i=0}^{\infty} x^i = \frac{1}{1-x} \quad \text{für } |x| < 1,$$

siehe Beispiel 16.8. ■

**Bemerkung 57.9**

- Die gesuchten Koeffizienten vor $x^k$, $k \in \mathbb{N}_0$, ergeben sich durch eine formale Potenzreihenentwicklung. Dabei kann man Konvergenzbetrachtungen ignorieren, da man nur an den Koeffizienten interessiert ist und nicht einen Wert für $x$ einsetzen will. Man kann auch ohne Beschränkung der Allgemeinheit $|x| < 1$ annehmen.
- Muss jedes Element mindestens $p$–mal auftreten, ergibt sich wegen

  $$x^p + x^{p+1} + \ldots = x^p \sum_{k=0}^{\infty} x^k = \frac{x^p}{1-x}$$

  die erzeugende Funktion

  $$f(x) = \frac{x^{np}}{(1-x)^n}.$$

□

Um das obige Vorgehen für Kombinationen auf Permutationen zu übertragen, muss man den Begriff der erzeugenden Funktion geeignet modifizieren.

**Definition 57.10 Exponentiell erzeugende Funktion.** Eine Funktion $f : \mathbb{R} \to \mathbb{R}$ mit

$$f(x) = \sum_{k=0}^{n} a_k \frac{x^k}{k!}$$

heißt exponentiell erzeugende Funktion für die Koeffizienten $a_k$. □

**Bemerkung 57.11**

- Der Name wird motiviert durch die Exponentialreihe

  $$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

- Nach Bemerkung 56.4, 2. gibt es bei einer $n$–elementigen Menge

  $$n(n-1)\ldots(n-k+1) = \frac{n!}{(n-k)!}$$

  $k$–Permutationen ohne Wiederholung, $k = 0, \ldots, n$. Wegen

  $$(x+1)^n = \sum_{k=0}^{n} \binom{n}{k} x^k = \sum_{k=0}^{n} \frac{n!}{(n-k)!k!} x^k = \sum_{k=0}^{n} \frac{n!}{(n-k)!} \frac{x^k}{k!}$$

  sind dies die Koeffizienten der exponentiell erzeugenden Funktion

  $$f(x) = (1+x)^n.$$

  Die exponentiell erzeugende Funktion spielt also bei Permutationen dieselbe Rolle wie die erzeugende Funktion bei Kombinationen.

□

Das Analogon zu Satz 57.7 lautet:

**Satz 57.12 Permutationen mit vorgegebenen Wiederholungen.** *Die exponentiell erzeugende Funktion der Permutationen mit Wiederholung aus einer n–elementigen Menge $\{x_1, \dots x_n\}$, in der $x_i$ in den Anzahlen $v_1^{(i)}, \dots, v_{k_i}^{(i)}$ auftreten darf, lautet*

$$f(x) = \prod_{i=1}^{n} \left( \frac{x^{v_1^{(i)}}}{v_1^{(i)}!} + \frac{x^{v_2^{(i)}}}{v_2^{(i)}!} + \dots + \frac{x^{v_{k_i}^{(i)}}}{v_{k_i}^{(i)}!} \right)$$

**Beweis:** Siehe Literatur. ■

**Satz 57.13 Permutationen mit beliebigen Wiederholungen.** *Die exponentiell erzeugende Funktion der Permutationen mit beliebigen Wiederholungen von n Elementen lautet*

$$f(x) = e^{nx} = (e^x)^n = \left( 1 + \frac{x}{1!} + \frac{x^2}{2!} + \dots \right)^n .$$

**Beweis:** Siehe Literatur, siehe auch Beispiel 20.15 für die Potenzreihendarstellung der Exponentialfunktion. ■

# Kapitel 58

# Bedingte Wahrscheinlichkeiten

**Bemerkung 58.1 Motivation.** In Kapitel 55 wurde der Begriff der Wahrscheinlichkeit eingeführt. Oft hat man Zusatzinformationen, mit denen sich präzisere Wahrscheinlichkeitsaussagen machen lassen. Dies führt zu so genannten bedingten Wahrscheinlichkeiten. □

**Beispiel 58.2** Betrachte die Aufteilung der 1500 Angehörigen eines Betriebs nach Geschlecht und Rauchergewohnheiten:

| | Frauen $B$ | Männer $\overline{B}$ |
|---|---|---|
| Raucher $A$ | 600 | 200 |
| Nichtraucher $\overline{A}$ | 300 | 400 |

Diese Informationen kann man folgendermaßen mit Mengen beschreiben:

| | | |
|---|---|---|
| $\Omega$: | Menge der Betriebsangehörigen, | |
| $A$: | Menge der Raucher | $\overline{A} = \Omega \setminus A$ : Nichtraucher, |
| $B$: | Menge der Frauen | $\overline{B} = \Omega \setminus B$ : Männer. |

Mit $|A|$ wird die Mächtigkeit einer Menge $A$ bezeichnet.

Man lost eine Person zufällig aus. Dabei treffen folgende Wahrscheinlichkeit zu:

$$\begin{aligned} P(A) &= \frac{|A|}{|\Omega|} = \frac{800}{1500} = \frac{8}{15} && \text{Raucheranteil,} \\ P(B) &= \frac{|B|}{|\Omega|} = \frac{900}{1500} = \frac{3}{5} && \text{Frauenanteil,} \\ P(A \cap B) &= \frac{|A \cap B|}{|\Omega|} = \frac{600}{1500} = \frac{2}{5} && \text{Anteil der rauchenden Frauen.} \end{aligned}$$

Die Frage ist nun, wie groß die Wahrscheinlichkeit ist, dass eine Person raucht, falls es sich um eine Frau handelt. Die Antwort erhält man wie folgt

$$P(A \mid B) = \frac{|A \cap B|}{|B|} = \frac{\frac{|A \cap B|}{|\Omega|}}{\frac{|B|}{|\Omega|}} = \frac{P(A \cap B)}{P(B)} = \frac{2/5}{3/5} = \frac{2}{3}.$$

Das sieht man auch sofort aus der ersten Tabelle. □

**Definition 58.3 Bedingte Wahrscheinlichkeit, unabhängige Ereignisse.** Man nennt $P(A \mid B)$ die bedingte Wahrscheinlichkeit von $A$ unter der Bedingung (Hypothese) $B$. Es gilt stets

$$P(A \cap B) = P(A \mid B)P(B).$$

Zwei Ereignisse $A$, $B$ heißen unabhängig, wenn gilt

$$P(A \cap B) = P(A)P(B).$$

□

**Satz 58.4 Verallgemeinerung auf $n$ Ereignisse.** *Seien $A_1, \ldots, A_n \subset \Omega$ und $P(A_1 \cap ... \cap A_{n-1}) > 0$. Dann gilt*

$$P(A_1 \cap \ldots \cap A_n) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \ldots P(A_n \mid A_1 \cap \ldots \cap A_{n-1}).$$

**Beweis:** Die rechte Seite lässt sich unter Nutzung der Formel aus Definition 58.3 schreiben als

$$P(A_1)\frac{P(A_1 \cap A_2)}{P(A_1)}\frac{P(A_1 \cap A_2 \cap A_3)}{P(A_1 \cap A_2)} \cdots \frac{P(A_1 \cap A_2 \cap \ldots \cap A_n)}{P(A_1 \cap \ldots \cap A_{n-1})} = P(A_1 \cap \ldots \cap A_n).$$

■

**Beispiel 58.5 Würfeln.** Untersuche die Wahrscheinlichkeit, mit 6 Würfeln 6 verschiedene Zahlen zu würfeln. Dazu kann man systematisch wie folgt vorgehen:

$A_1$ : irgendein Ergebnis für 1. Würfel,
$A_2$ : ein vom 1. Ergebnis verschiedenes Ergebnis für Würfel 2,
$\vdots$
$A_6$ : ein von $A_1, \ldots, A_5$ verschiedenes Ergebnis für Würfel 6.

Damit folgt

$$\begin{aligned} P(A_1 \cap \ldots \cap A_6) &= P(A_1)P(A_2 \mid A_1) \ldots P(A_6 \mid A_1 \cap \ldots \cap A_5) \\ &= 1 \cdot \frac{5}{6} \cdot \ldots \cdot \frac{1}{6} = \frac{6!}{6^6} \approx 0.015. \end{aligned}$$

□

**Beispiel 58.6 Unsicherer Krebstest.** Ein Krebstest sei mit 96%–iger Sicherheit positiv, falls der Patient Krebs hat, mit 94%–iger Sicherheit negativ, falls er keinen Krebs hat. Bei einem Patienten, in dessen Altersgruppe 0.5% aller Personen Krebs haben, verläuft der Test positiv. Wie groß ist die Wahrscheinlichkeit, dass er tatsächlich Krebs hat?

Wir bezeichnen das Ereignis, dass der Patient Krebs hat, mit $K$ und das Ereignis, dass der Test positiv ist, mit $T$. Dann folgt

$$\begin{aligned} & P(K \mid T) \\ &= \frac{P(K \cap T)}{P(T)} \\ &= \frac{P(\text{Patient Krebs \& Test positiv})}{P(\text{Patient Krebs \& Test positiv}) + P(\text{Patient nicht Krebs \& Test positiv})} \\ &= \frac{0.005 \cdot 0.96}{0.005 \cdot 0.96 + 0.995 \cdot 0.06} \approx 0.074. \end{aligned}$$

Die Wahrscheinlichkeit, dass der Patient tatsächlich Krebs hat, ist etwa 7.4 %. Diese kleine Wahrscheinlichkeit kommt dadurch zustande, dass der prozentuale Anteil der

Nichtkrebspatienten sehr hoch ist und die absolute Anzahl der fehlerhaften positiven Tests groß ist im Vergleich zu den korrekten positiven Tests.

Fazit: Um eine seltene Krankheit zuverlässig zu erkennen, darf ein Test nur sehr wenige „false positives“ haben. □

**Satz 58.7 Satz von der totalen Wahrscheinlichkeit.** *Sei*

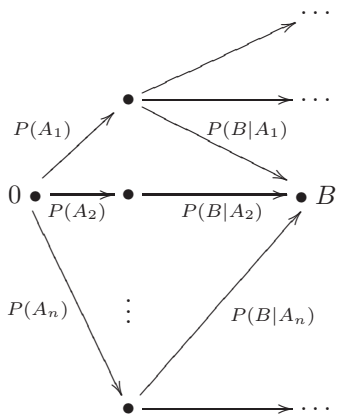
$$\Omega = \bigcup_{i=1}^{n} A_i$$

*eine Partition von $\Omega$ in mögliche Ereignisse $A_i, i = 1, \ldots, n$, siehe Definition 4.7. Ferner sei $P(A_i) > 0$ für alle $i$. Dann gilt für jedes Ereignis $B$*

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B \mid A_i).$$

**Beweis:** Siehe Literatur. ■

**Bemerkung 58.8 Veranschaulichung.**




Ein Fahrer startet bei 0 und fährt mit Wahrscheinlichkeit $P(A_1), \ldots, P(A_n)$ zu $A_1, \ldots, A_n$. Die Wahrscheinlichkeit von dort nach $B$ zu fahren, beträgt $P(B \mid A_1), \ldots, P(B \mid A_n)$. Die Gesamtwahrscheinlichkeit, dass der Fahrer nach $B$ gelangt, ist die Summe der Wahrscheinlichkeiten aller Pfade von 0 nach $B_i$

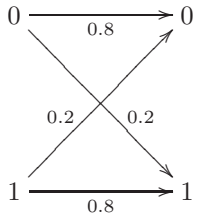
$$P(B) = \sum_{i=1}^{n} P(A_i)P(B \mid A_i).$$

Im Beispiel 58.6 wurde im Nenner bereits den Satz von der totalen Wahrscheinlichkeit verwendet. □

**Beispiel 58.9** Um einen binären Nachrichtenkanal robuster gegenüber Störungen zu machen, sendet man die Bitfolge 0000000 statt 0 und 1111111 statt 1. Störungen treten in 20% aller Fälle auf. Die Wahrscheinlichkeit dafür, dass 0000000 gesendet wird, sei 0.1. Demzufolge ist die Wahrscheinlichkeit, dass 1111111 gesendet wird, neunmal so hoch.

Es werde 0100110 empfangen. Wie groß ist die Wahrscheinlichkeit, dass 0000000 gesendet wurde? Die Wahrscheinlichkeiten für alle auftretende Fälle der Sendung

eines Bits sind wie folgt:




Damit erhält man

$$\begin{aligned} P(0000000 \mid 0100110) &= \frac{P(0000000 \text{ gesendet})}{P(0000000 \text{ gesendet}) + P(1111111 \text{ gesendet})} \\ &= \frac{0.1 \cdot 0.2^3 0.8^4}{0.1 \cdot 0.2^3 0.8^4 + 0.9 \cdot 0.2^4 0.8^3} \approx 0.308. \end{aligned}$$

Man wird den Block also als 1 lesen, obwohl die Mehrzahl der Bits Nullen sind! □

Der folgende Satz ist eng verwandt mit dem Satz von der totalen Wahrscheinlichkeit.

**Satz 58.10 Formel von Bayes**[1]**.** *Sei $P(B) > 0$ und seien die Voraussetzungen von Satz 58.7 erfüllt. Dann gilt*

$$P(A_k \mid B) = \frac{P(A_k)P(B \mid A_k)}{\sum_{i=1}^{n} P(A_i)P(B \mid A_i)}.$$

**Beweis:** Die Aussage folgt aus

$$P(A_k \mid B) = \frac{P(A_k \cap B)}{P(B)},$$

indem man

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B \mid A_i), \quad P(A_k \cap B) = P(A_k)P(B|A_k)$$

einsetzt. ■

**Bemerkung 58.11 Inverse Probleme.** In Satz 58.10 wird $P(A_k \mid B)$ aus $P(B \mid A_i)$, $i = 1, \ldots, n$, berechnet, das heißt Ursache und Wirkung kehren sich um. Eine typische Anwendung besteht darin, dass man eine Wirkung misst und nach der wahrscheinlichsten Ursache fragt. Solche Fragestellungen nennt man inverse Probleme. Diese sind besonders schwierig zu lösen. □

**Beispiel 58.12 Inverse Probleme in den Anwendungen.**

- Ein Arzt beobachtet bei einem Patienten ein Symptom $B$. Es kann von $n$ verschiedenen Krankheiten $A_k$, $k = 1, \ldots, n$, herrühren. Um die wahrscheinlichste Ursache zu finden, muss man also $P(A_k \mid B)$ abschätzen.
- Aus einem verrauschten Bild will man das wahrscheinlichste unverrauschte Bild rekonstruieren, vergleiche auch Beispiel 58.9.
- In der Computertomographie schickt man Röntgenstrahlung in verschiedenen Richtungen durch den Patienten und misst die durchgedrungene Intensität. Aus diesen Auswirkungen versucht man, Rückschlüsse auf die Ursache (Gewebe, Knochen, Tumor, ...) zu ziehen.

□

[1] Thomas Bayes (1702 – 1761)

# Kapitel 59

# Zufallsvariablen, Erwartungswert, Varianz

**Bemerkung 59.1 Motivation.** Oft möchte man dem Resultat eines Zufallsexperiments reelle Zahlen zuordnen. Der Gewinn bei einem Glücksspiel ist ein Beispiel hierfür. In diesem Fall interessiert man sich auch für den zu erwartenden Gewinn und für ein Maß für die statistischen Schwankungen. Dies führt auf Begriffe wie Zufallsvariable, Erwartungswert und Varianz. In der Informatik werden sie unter anderem bei der Zuverlässigkeitsanalyse von Systemen benötigt. □

**Definition 59.2 Zufallsvariable.** Sei $\Omega$ ein Stichprobenraum. Eine Funktion $X$, die jedem Ergebnis $\omega \in \Omega$ eine reelle Zahl $X(\omega)$ zuordnet, heißt Zufallsvariable. □

**Bemerkung 59.3** Eine Zufallsvariable ist also weder zufällig noch eine Variable, sondern eine Funktion. Man kann sie stets als Gewinn bei einem Glücksspiel interpretieren. □

**Beispiel 59.4** Eine faire Münze mit Seiten 0 und 1 werde drei Mal geworfen. Die Anzahl der Einsen sei der Gewinn. Man kann $\Omega = \{000, 001, \ldots, 111\}$ als Stichprobenraum und den Gewinn als Zufallsvariable $X(\omega)$ auffassen.

| Ergebnis $\omega$ | 000 | 001 | 010 | 011 | 100 | 101 | 110 | 111 |
|---|---|---|---|---|---|---|---|---|
| Gewinn $X(\omega)$ | 0 | 1 | 1 | 2 | 1 | 2 | 2 | 3 |
| Ws. $P(\omega)$ | 1/8 | 1/8 | 1/8 | 1/8 | 1/8 | 1/8 | 1/8 | 1/8 |

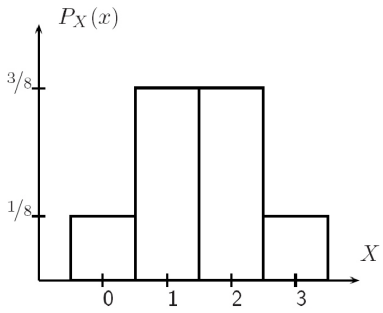
Verschiedene Ereignisse, beispielsweise 001 und 100, können zum selben Gewinn führen. □

**Definition 59.5 Verteilung.** Die Verteilung $P_X$ einer Zufallsvariablen $X$ ordnet jedem Wert $x \in X(\Omega)$ eine Wahrscheinlichkeit $P_X(x)$ zu. □

**Beispiel 59.6** Im Beispiel 59.4 kann man dem Gewinn folgende Wahrscheinlichkeiten zuordnen:

| Gewinn $x \in X(\Omega)$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Ws. $P_X(x)$ | 1/8 | 3/8 | 3/8 | 1/8 |

Oft veranschaulicht man die Verteilung einer Zufallsvariablen durch ein Histogramm:




Betrachtet man diskrete Zufallsvariablen, dann sind die Verteilungen ebenfalls diskret. □

Interessiert man sich für den Durchschnittsgewinn je Versuchswiederholung, gelangt man zum Begriff des Erwartungswertes.

**Definition 59.7 Erwartungswert.** Unter dem Erwartungswert $E(X)$ einer (diskreten) Zufallsvariablen $X$ versteht man das gewichtete Mittel der Funktionswerte $X(\omega)$ über $\Omega$, wobei jeder Wert mit seiner Wahrscheinlichkeit gewichtet wird

$$E(X) = \sum_{\omega \in \Omega} X(\omega) P(\omega).$$

□

**Beispiel 59.8** Für die Zufallsvariable $X$ aus Beispiel 59.4 erhält man als Erwartungswert

$$E(X) = 0\frac{1}{8} + 1\frac{1}{8} + 1\frac{1}{8} + 2\frac{1}{8} + 1\frac{1}{8} + 2\frac{1}{8} + 2\frac{1}{8} + 3\frac{1}{8} = \frac{12}{8} = \frac{3}{2}.$$

Man hätte den Erwartungswert mit Hilfe der Verteilung $P_X$ berechnet können

$$E(X) = 0\frac{1}{8} + 1\frac{3}{8} + 2\frac{3}{8} + 3\frac{1}{8} = \frac{3}{2}.$$

Es gilt also auch

$$E(X) = \sum_{x \in X(\Omega)} x P_X(x).$$

□

**Bemerkung 59.9 Kontinuierliche Zufallsvariablen.** Bei kontinuierlichen Zufallsvariablen verwendet man die übliche Verallgemeinerung von Summen, nämlich Integrale. □

**Satz 59.10 Linearität des Erwartungswerts.** *Seien $X, Y$ Zufallsvariablen und $\alpha, \beta \in \mathbb{R}$. Dann gilt*

$$E(\alpha X + \beta Y) = \alpha E(X) + \beta E(Y).$$

**Beweis:** Die Aussage des Satzes folgt mit Hilfe der Linearität der Summation

$$\begin{aligned} E(\alpha X + \beta Y) &= \sum_{\omega \in \Omega} \Big( \alpha X(\omega) + \beta Y(\omega) \Big) P(\omega) \\ &= \alpha \sum_{\omega \in \Omega} X(\omega) P(\omega) + \beta \sum_{\omega \in \Omega} Y(\omega) P(\omega) \\ &= \alpha E(X) + \beta E(Y). \end{aligned}$$

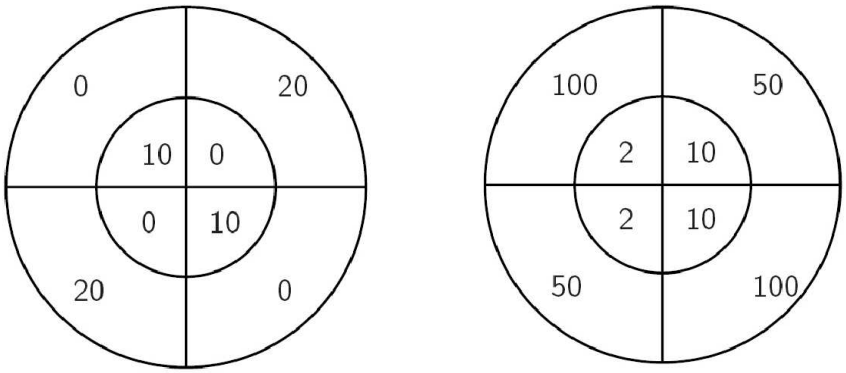
■

**Bemerkung 59.11 Multiplikativität von Erwartungswerten.** Man kann zeigen, dass $E(XY) = E(X)E(Y)$ gilt, falls $X$ und $Y$ unabhängig sind, das heißt

$$P\big((X = a) \cap (Y = b)\big) = P(X = a)P(Y = b) \quad \forall\, a, b \in \Omega,$$

vergleiche Definition 58.3. □

**Beispiel 59.12** Betrachte Glücksräder mit vier Ergebnissen, auf denen die Zufallsvariable $X$ (äußerer Ring) und $Y$ (innerer Ring) definiert sind.




Im linken Rad gelten

$$\left.\begin{array}{rcl} E(X) & = & \frac{1}{4}20 + \frac{1}{4}0 + \frac{1}{4}20 + \frac{1}{4}0 = 10 \\ E(Y) & = & \frac{1}{4}0 + \frac{1}{4}10 + \frac{1}{4}0 + \frac{1}{4}10 = 5 \end{array}\right\} E(X)E(Y) = 50$$

und

$$E(XY) = \frac{1}{4}\cdot 20\cdot 0 + \frac{1}{4}\cdot 0\cdot 10 + \frac{1}{4}\cdot 20\cdot 0 + \frac{1}{4}\cdot 0\cdot 10 = 0 \neq E(X)E(Y).$$

Damit sind $X$ und $Y$ nicht unabhängig. Das Ereignis $Y = 0$ hat die Wahrscheinlichkeit $\frac{1}{2}$. Weiß man jedoch, dass $X = 20$ eingetreten ist, dann hat $Y = 0$ die Wahrscheinlichkeit 1.

Im rechten Glücksrad gelten

$$\left.\begin{array}{rcl} E(X) & = & \frac{1}{4}50 + \frac{1}{4}100 + \frac{1}{4}50 + \frac{1}{4}100 = 75 \\ E(Y) & = & \frac{1}{4}10 + \frac{1}{4}10 + \frac{1}{4}2 + \frac{1}{4}2 = 6 \end{array}\right\} \Rightarrow E(X)E(Y) = 450$$

und

$$E(XY) = \frac{1}{4}\cdot 50\cdot 10 + \frac{1}{4}\cdot 100\cdot 10 + \frac{1}{4}\cdot 50\cdot 2 + \frac{1}{4}\cdot 100\cdot 2 = 450 = E(X)E(Y).$$

Die Zufallsvariablen $X$ und $Y$ sind unabhängig. $Y = 2$ und $Y = 10$ sind gleich wahrscheinlich. Ist das Ergebnis von $X$ bekannt, dann sind $Y = 2$ und $Y = 10$ immer noch gleich wahrscheinlich. □

Oft ist man nicht nur am Erwartungswert interessiert. Man möchte auch wissen, wie stark die Verteilung um den Erwartungswert streut.

**Definition 59.13 Varianz, Standardabweichung, Streuung.** Sei $X$ eine Zufallsvariable mit Erwartungswert $\mu = E(X)$. Dann versteht man unter der Varianz $V(X) = \sigma^2$ den Erwartungswert von $(X - \mu)^2$

$$\sigma^2 = V(X) := E\left((X - \mu)^2\right).$$

Die Größe $\sigma := \sqrt{V(X)}$ nennt man die Standardabweichung oder Streuung von $X$. □

**Bemerkung 59.14 Berechnung der Varianz.** Wegen der Linearität des Erwartungswerts und wegen $E(c) = c$ für $c \in \mathbb{R}$ gilt

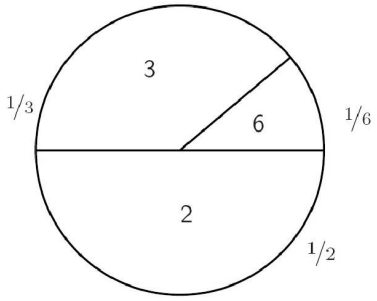
$$\begin{aligned} E((X-\mu)^2) &= E(X^2 - 2\mu X + \mu^2) = E(X^2) - 2\mu \underbrace{E(X)}_{\mu} + \mu^2 \\ &= E(X^2) - \mu^2. \end{aligned}$$

Man erhält somit eine wichtige Formel zur Berechnung der Varianz, den sogenannten Verschiebesatz

$$\sigma^2 = E(X^2) - \mu^2.$$

Wegen $\sigma^2 \geq 0$ folgt $E(X^2) \geq (E(X))^2 = \mu^2$. □

**Beispiel 59.15** Sei $X$ der Gewinn auf dem Glücksrad




Dann gelten:

$$\begin{aligned} \text{mittlerer Gewinn:} \quad \mu = E(X) &= \frac{1}{2}2 + \frac{1}{3}3 + \frac{1}{6}\,6 = 3, \\ E(X^2) &= \frac{1}{2}4 + \frac{1}{3}9 + \frac{1}{6}36 = 11, \\ \text{Varianz:} \quad \sigma^2 &= E(X^2) - \mu^2 = 11 - 3^2 = 2, \\ \text{Standardabweichung:} \quad \sigma &= \sqrt{2}. \end{aligned}$$

□

**Satz 59.16 Eigenschaften der Varianz.** *Sei $X$ eine Zufallsvariable und seien $\alpha, \beta \in \mathbb{R}$. Dann gelten*

*i)* $V(\alpha X) = \alpha^2 V(X)$,
*ii)* $V(X + \beta) = V(X)$.

**Beweis:** Der Beweis erfolgt mit Hilfe der Homogenität des Erwartungswertes

$$V(\alpha X) = E((\alpha X - \alpha\mu)^2) = E(\alpha^2 (X-\mu)^2) = \alpha^2 V(X)$$

und

$$V(X+\beta) = E\left((X + \beta - (\mu + \beta))^2\right) = V(X).$$

■

**Definition 59.17 Standardisierte Zufallsvariable.** Ist $X$ eine Zufallsvariable mit Erwartungswert $\mu$ und Varianz $\sigma^2$, so nennt man

$$X^* := \frac{X - \mu}{\sigma}$$

die Standardisierte von $X$. □

**Bemerkung 59.18** Eine solche Standardisierung ist nützlich, wenn man die Verteilung einer Zufallsvariablen mit einer tabellierten Verteilung vergleichen möchte, da letzterer oft nur in standardisierter Form vorliegt. Die standardisierte Zufallsvariable besitzt folgende wichtige Eigenschaften:

$$\begin{aligned} E(X^*) &= \frac{1}{\sigma}\underbrace{E(X)}_{\mu} - \frac{\mu}{\sigma} = \frac{\mu}{\sigma} - \frac{\mu}{\sigma} = 0, \\ V(X^*) &\overset{\text{Satz 59.16}}{=} \frac{1}{\sigma^2}\underbrace{V(X)}_{\sigma^2} = \frac{\sigma^2}{\sigma^2} = 1. \end{aligned}$$

□

**Satz 59.19 Ungleichung von Jensen**[1]**.** *Seien $r : \mathbb{R} \to \mathbb{R}$ eine konvexe Funktion und $X$ eine Zufallsvariable. Dann gilt*

$$E(r(X)) \geq r(E(X)).$$

**Beweis:** Sei $X$ zunächst eine diskrete Zufallsvariable. Dann wird

$$E(X) = \sum_{\omega\in\Omega} X(\omega)P(\omega)$$

als Konvexkombination der $X(\omega)$, $\omega \in \Omega$, mit Gewichten $P(\omega)$ aufgefasst, das heißt es gelten

$$P(\omega) \geq 0 \ \ \forall \ \ \omega \in \Omega \quad \text{und} \quad \sum_{\omega\in\Omega} P(\omega) = 1.$$

Aus der Konvexität von $r(X)$ folgt mit Definition 23.3

$$E(r(X)) = \sum_{\omega\in\Omega} r(X(\omega))P(\omega) \geq r\left(\sum_{\omega\in\Omega} X(\omega)P(\omega)\right) = r(E(X)).$$

Ist $X$ eine kontinuierliche Zufallsvariable, ersetzt man Summen durch Integrale. ■

**Beispiel 59.20**

- Die Funktion $r(t) = t^2$ ist konvex, da $r''(t) = 2 > 0$. Daher gilt

  $$E\left(X^2\right) \geq (E(X))^2,$$

  siehe Bemerkung 59.14.
- Sei $\theta > 0$. Dann ist $r(t) = e^{\theta t}$ konvex, und es gilt

  $$E\left(e^{\theta X}\right) \geq e^{\theta E(X)}.$$

  Die Funktion $E\left(e^{\theta X}\right)$ wird später eine wichtige Rolle spielen.

□

**Bemerkung 59.21 Stoppzeit.** Seien $X_1, X_2, \ldots$ unabhängige Zufallsvariablen. Manchmal interessiert man sich für die erste Zeit, die sogenannte Stoppzeit, $N = n$, in der die Summe $X_1 + \ldots + X_N$ einen vorgegebenen Wert $y$ übersteigt, das heißt

$$\sum_{i=1}^{n-1} X_i \leq y \quad \text{und} \quad \sum_{i=1}^{n} X_i > y.$$

Dann ist $N$ selbst wieder eine Zufallsvariable. □

[1] Johan Ludwig William Valdemar Jensen (1859 – 1925)

Für ihren Erwartungswert kann man den folgenden Satz zeigen.

**Satz 59.22 Gleichung von Wald[2].** *Seien $X_1, X_2, \ldots$ unabhängige, gleich verteilte Zufallsvariablen, das heißt $P(\omega)$ ist gleich für alle $\omega \in \Omega$, mit endlichem Erwartungswert, und sei $N$ eine Stoppzeit für $X_1, X_2, \ldots$. Ferner sei $S_N = \sum_{i=1}^{N} X_i$. Dann gilt*

$$E(S_N) = E(X)E(N).$$

**Beweis:** Siehe Literatur. ■

**Beispiel 59.23** Sei $X_i = 1$, falls beim $i$–ten Münzwurf Kopf eintritt, ansonsten $X_i = 0$. Gesucht ist der Erwartungswert $E(X)$ für die Stoppzeit $N := \min\{n \mid X_1 + ... + X_n \geq 10\}$.

Es gelten

$$E(X) = \frac{1}{2} \quad \text{und} \quad E(S_N) = 10.$$

Daraus folgt $E(N) = 20$. Nach 20 Würfen kann man erwarten, dass man zehnmal Kopf gewürfelt hat. □

Eng verwandt mit dem Begriff der Varianz ist die Kovarianz. Sie ist ein Maß für die Unabhängigkeit zweier Zufallsvariablen.

**Definition 59.24 Kovarianz, Korrelationskoeffizient.** Seien $X, Y$ Zufallsvariablen mit Erwartungswert $\mu_X, \mu_Y$ und Varianz $\sigma_X^2, \sigma_Y^2 > 0$. Dann heißen

$$\mathrm{Cov}(X,Y) := \sigma_{XY}^2 := E\big((X - \mu_X)(Y - \mu_Y)\big)$$

die Kovarianz von $X$ und $Y$, und

$$\varrho_{XY} := \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

der Korrelationskoeffizient von $X$ und $Y$.

Ist $\mathrm{Cov}(X,Y) = 0$, so heißen $X$ und $Y$ unkorreliert. □

**Bemerkung 59.25**

- Aus der Definition folgt direkt, dass $V(X) = \mathrm{Cov}(X,X)$.
- Sind $X - \mu_X$ und $Y - \mu_Y$ Funktionen auf einem endlichen Stichprobenraum $\Omega = \{\omega_1, \ldots, \omega_n\}$, so kann man sie auch als Vektoren im $\mathbb{R}^n$ mit der $i$–ten Komponente $X(\omega_i) - \mu_X$ beziehungsweise $Y(\omega_i) - \mu_Y$ ansehen. Dann bezeichnen:
  - die Standardabweichungen die induzierten euklidischen Normen,
  - die Varianzen die quadrierten euklidischen Normen,
  - die Kovarianz das euklidische Skalarprodukt,
  - der Korrelationskoeffizient den Kosinus des Winkels zwischen beiden Vektoren, insbesondere ist $-1 \leq \varrho_{XY} \leq 1$,
  - die Unkorreliertheit die Orthogonalität,

  falls das gewichtete euklidische Skalarprodukt

$$(u,v) := \sum_{i=1}^{n} p_i u_i v_i$$

  zu Grunde gelegt wird. Dabei ist $p_i := P(\omega_i)$.

[2] Abraham Wald (1902 – 1950)

□

**Satz 59.26 Rechenregeln für die Korrelation.** *Seien $X, Y$ Zufallsvariablen mit Erwartungswert $\mu_X, \mu_Y$. Dann gelten*

*i) $Cov(X,Y) = E(XY) - \mu_X \mu_Y$, vergleiche Bemerkung 59.14,*
*ii) $Cov(\alpha X + \beta, \gamma Y + \delta) = \alpha\gamma\ Cov(X,Y)$, vergleich Satz 59.16,*
*iii) $Cov(X,Y) = Cov(Y,X)$,*
*iv) für $m$ Zufallsvariablen $X_1, \ldots, X_m$ gilt*

$$V(X_1 + \ldots + X_m) = \sum_{i=1}^{m} V(X_i) + \sum_{i \neq j} Cov(X_i, X_j),$$

*v) sind $X, Y$ unabhängig, so sind sie auch unkorreliert,*
*vi) für unabhängige $X_1, \ldots, X_n$ gilt*

$$V(X_1 + \ldots + X_n) = \sum_{i=1}^{m} V(X_i).$$

**Beweis:** Die Aussagen überprüft man mit bekannten Eigenschaften des Erwartungswertes, Übungsaufgaben. ■

**Bemerkung 59.27**

- Die Umkehrung der Aussage v) von Satz 59.26 gilt nicht.
  Betrachte dazu folgendes Beispiel

| Ergebnis $\omega$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Zufallsvar. $X$ | 1 | -1 | 2 | -2 |
| Zufallsvar. $Y$ | -1 | 1 | 2 | -2 |
| Wahrscheinlichkeit $P(\omega)$ | 2/5 | 2/5 | 1/10 | 1/10 |

  Dann sind $E(X) = 0 = E(Y)$ und

$$\mathrm{Cov}(X,Y) = -1\frac{2}{5} - 1\frac{2}{5} + 4\frac{1}{10} + 4\frac{1}{10} = 0,$$

  aber $X$ und $Y$ sind nicht unabhängig, denn $X(\omega)$ bestimmt $\omega$ und $Y(\omega)$ eindeutig.
- In der Informatik tauchen Erwartungswerte und Varianzen zum Beispiel bei der Zuverlässigkeitsanalyse von Systemen oder bei der Abschätzung von Wartezeiten bei Internetanfragen auf. Kovarianzen sind unter anderem wichtig im Bereich des maschinellen Lernens.

□

**Beispiel 59.28 Zuverlässigkeitsanalyse von Systemen.** Ein System bestehe aus $n$ Komponenten. Der Zustand der $k$–ten Komponente wird durch die Zufallsvariable (Indikator)

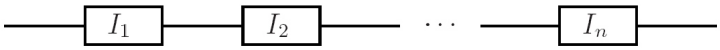
$$I_k := \begin{cases} 1 & k\text{–te Komponente funktioniert,} \\ 0 & k\text{–te Komponente funktioniert nicht,} \end{cases}$$

beschrieben. Ihr Erwartungswert beschreibt die Zuverlässigkeit der Komponente $k$. Man setzt

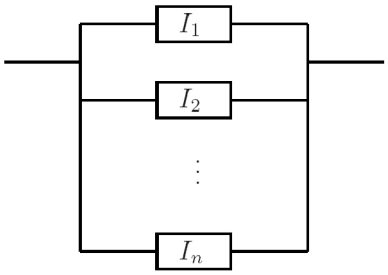
$$p_k := E(I_k), \quad q_k := 1 - p_k,$$

wobei $q_k$ Ausfallwahrscheinlichkeit genannt wird. Interessiert man sich für die Zuverlässigkeit $p$ des Gesamtsystems, muss man verschiedene Fälle unterscheiden.

1. *Reihensysteme.*



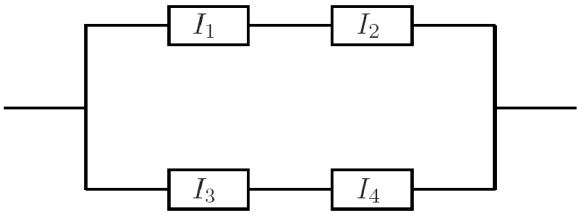

   Ein Reihensystem arbeitet, wenn alle Komponenten arbeiten: $p = p_1 \dots p_n$.
2. *Parallelsysteme.*




   Ein Parallelsystem fällt aus, wenn alle Komponenten ausfallen

$$q = q_1 \dots q_n \quad \Longrightarrow \quad p = 1 - q = 1 - (1 - p_1) \dots (1 - p_n).$$

3. *Gemischte Systeme.* Diese Systeme werden hierarchisch in Reihensysteme und Parallelsystem zerlegt.




   Die Zuverlässigkeit dieses Systems berechnet sich aus der Zuverlässigkeit der Teilsysteme:
   - oberes Reihensystem: $p_1 p_2$,
   - unteres Reihensystem: $p_3 p_4$,
   - äußeres Parallelsystem: $p = 1 - (1 - p_1 p_2)(1 - p_3 p_4)$.

□

# Kapitel 60

# Abschätzungen für Abweichungen vom Erwartungswert

**Bemerkung 60.1 Motivation.** Mit der Varianz beziehungsweise Standardabweichungen ist bereits ein Maß für die Fluktuation einer Zufallsvariablen um ihren Erwartungswert eingeführt. Dieser Abschnitt stellt weitere nützliche Maße vor. Außerdem werden Abschätzungen vorgestellt, die wahrscheinlichkeitstheoretische Aussagen darüber liefern, wie häufig eine Zufallsvariable außerhalb eines Intervalls um den Erwartungswert liegt. □

**Definition 60.2** $k$**–tes (zentrales) Moment.** Sei $X$ eine Zufallsvariable mit Erwartungswert $\mu = E(X)$. Dann bezeichnet man mit $E(X^k)$, $k \in \mathbb{N}$, das $k$–te Moment von $X$. Ferner definiert $E((X-\mu)^k)$ das $k$–te zentrale Moment von $X$. □

**Bemerkung 60.3**

- Der Erwartungswert $\mu = E(X)$ ist das erste Moment.
- Die Varianz $\sigma^2 = E((X-\mu)^2)$ ist das zweite zentrale Moment.
- Mit den 3. und 4. zentralen Momenten lassen sich Aussagen über die Schiefe beziehungsweise Flachheit der Verteilung einer Zufallsvariablen gewinnen.
- Höhere Momente sind anschaulich schwieriger zu interpretieren, liefern jedoch ebenfalls wichtige Aussagen.
- Momente haben eine große Bedeutung in der Mustererkennung, bei der Analyse von Texturen und der Erkennung von Objekten unter Rotationen und Skalierungen.

□

Ähnlich wie man die Menge aller $k$–Permutationen und $k$–Kombinationen einer $n$–elementigen Menge durch den Begriff der erzeugenden Funktion kompakt beschreiben kann, gibt es eine Funktion, die sämtliche Momente einer Zufallsvariablen beinhaltet.

**Definition 60.4 Momenten–erzeugende Funktion.** Sei $X$ eine Zufallsvariable. Falls $M_X(\Theta) := E(e^{\Theta X})$ existiert, nennt man $M_X(\Theta)$ die Momenten–erzeugende Funktion von $X$. □

**Satz 60.5 Eigenschaften Momenten–erzeugender Funktionen.** *Die Momenten–erzeugende Funktion $M_X(\Theta) = E(e^{\Theta X})$ einer Zufallsvariablen $X$ hat folgende Eigenschaften:*

*i) Die n–te Ableitung in $\Theta = 0$ liefert das n–te Moment*

$$M_X^{(n)}(0) = E\left(X^n\right).$$

*ii) Skalierungsverhalten. Sei $Y = aX + b$. Dann ist*

$$M_Y(\Theta) = e^{b\Theta} M_X(a\Theta).$$

*iii) Falls $X$ und $Y$ unabhängige Zufallsvariablen sind gilt*

$$M_{X+Y}(\Theta) = M_X(\Theta) M_Y(\Theta).$$

**Beweis:** *i).* Mit der Potenzreihenentwicklung von $\exp(x)$ und der Linearität des Erwarungswerts gilt

$$M_X(\Theta) = E\left(e^{\Theta X}\right) = E\left(\sum_{k=0}^{\infty} \frac{\Theta^k x^k}{k!}\right) = \sum_{k=0}^{\infty} \frac{\Theta^k}{k!} E\left(X^k\right).$$

Gliedweise Differentiation liefert

$$\begin{aligned} M_X^{'}(\Theta) &= \sum_{k=1}^{\infty} \frac{\Theta^{k-1}}{(k-1)!} E\left(X^k\right) = \sum_{k=0}^{\infty} \frac{\Theta^k}{k!} E\left(X^{k+1}\right) \\ &\vdots \\ M_X^{(n)}(\Theta) &= \sum_{k=0}^{\infty} \frac{\Theta^k}{k!} E\left(X^{k+n}\right) \end{aligned}$$

und somit

$$M_X^{(n)}(0) = E(X^n).$$

*ii), iii).* Siehe Literatur. ∎

Nun werden Abschätzungen für Fluktuationen jenseits eines vorgegebenen Abstands um den Erwartungswert einer Zufallsvariablen gezeigt. Die Grundlage ist der folgende Satz.

**Satz 60.6 Markow[1]sche Ungleichung.** *Seien $X$ eine Zufallsvariable und $h : \mathbb{R} \to \mathbb{R}$ eine nichtnegative und nichtfallende Funktion mit $h(t) > 0$. Dann gilt*

$$P(X \geq t) \leq \frac{E(h(X))}{h(t)}.$$

**Beweis:** Der Beweis wird nur für eine diskrete Zufallsvariable angegeben. Nach Beispiel 59.8 gilt

$$E(h(X)) = \sum_{z \in X(\Omega)} h(z) P_X(z).$$

Wegen

$$\sum_{z \in X(\Omega)} \underbrace{h(z)}_{>0} \underbrace{P_X(z)}_{\geq 0} \geq \sum_{\substack{z \in X(\Omega) \\ z \geq t}} h(z) P_X(z) \overset{\text{nicht-fallend}}{\geq} h(t) \underbrace{\sum_{\substack{z \in X(\Omega) \\ z \geq t}} P_X(z)}_{P(X \geq t)}$$

folgt

$$P(X \geq t) \leq \frac{1}{h(t)} E(h(X)).$$

∎

[1] Andrei Andrejewitsch Markow (1856) – (1922)

**Bemerkung 60.7**

1. Setzt man
$$h(x) = \begin{cases} x & \text{für } x > 0, \\ 0 & \text{sonst,} \end{cases}$$
folgt die sogenannte einfache Markow–Ungleichung
$$P(X \geq t) \leq \frac{E(X)}{t} \quad \text{für} \quad t > 0.$$

2. Mit $Y := (X - E(X))^2$ und $h(x) = x$ für $x > 0$ kann man die Tschebyschew[2]–Ungleichung für den Erwartungswert $\mu$ und die Standardabweichung $\sigma$ von $X$ beweisen
$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$
Eine Alternativschreibweise ist
$$P(|X - \mu| \geq c\sigma) \leq \frac{1}{c^2}.$$

3. Mit $h(X) = e^{\Theta X}$ für $\Theta \geq 0$ ergibt sich
$$P(X \geq t) \leq e^{-\Theta t} M_X(\Theta).$$
Dies führt zur Chernoff[3]–Schranke
$$P(X \geq t) \leq \inf_{\Theta \geq 0} e^{-\Theta t} M_X(\Theta).$$

4. Für die einfache Markow–Ungleichung benötigt man das erste Moment (Erwartungswert $\mu$), für die Tschebyschew–Ungleichung die ersten beiden Momente $\mu, \sigma^2$, und für die Chernoff–Ungleichung alle Momente, das heißt die Momenten–erzeugende Funktion. Je mehr Momente man kennt, desto mehr weiß man über die Verteilung von $X$ und desto schärfere Abschätzungen kann man erwarten.

□

**Beispiel 60.8** Eine faire Münze werde $n$ Mal geworfen. Tritt beim $k$–ten Wurf Kopf auf, setzt man $Y_k := 1$, sonst $Y_k := 0$. Wir interessieren uns für die „Kopfhäufigkeit“ nach $n$ Würfen
$$X_n := Y_1 + \ldots + Y_n.$$
$Y_1, \ldots, Y_n$ sind unabhängige Zufallsvariablen. Für $X_n$ gilt
$$\begin{aligned} \mu &= E(X_n) = \frac{n}{2}, \\ \sigma^2 &= E((X_n - \mu)^2) = \sum_{k=1}^{n} \left[ \left(0 - \frac{1}{2}\right)^2 \frac{1}{2} + \left(1 - \frac{1}{2}\right)^2 \frac{1}{2} \right] = \frac{n}{4}. \end{aligned}$$
Wegen
$$M_{Y_k}(\Theta) = E\left(e^{\Theta Y_k}\right) = e^{\Theta \cdot 0} \frac{1}{2} + e^{\Theta \cdot 1} \frac{1}{2} = \frac{1 + e^{\Theta}}{2}$$
folgt mit Satz 60.5, iii),
$$M_{X_n}(\Theta) = M_{Y_1 + \ldots + Y_n} = \left(\frac{1 + e^{\Theta}}{2}\right)^n.$$

Sei $\alpha = 0.8$. Gesucht ist die Wahrscheinlichkeit, nach $n = 100$ Würfen $X_n \geq \alpha n = 80$ zu erhalten. Vergleicht man dazu die drei Ungleichungen aus Bemerkung 60.7, erhält man:

[2] Pafnuti Lwowitsch Tschebyschew (1821 – 1894)
[3] Herman Chernoff geb. 1923

- *Einfache Markow–Ungleichung.* Mit $\mu = 50$ und $t = 80$ ergibt sich

$$P(X_{100} \geq 80) \leq \frac{\mu}{t} = \frac{50}{80} = 0.625.$$

- *Tschebyschew–Ungleichung.* Mit $\mu = 50, t = 30$ und $\sigma^2 = 25$ ergibt sich

$$P(X_{100} \geq 80) \leq P(|X_{100} - 50| \geq 30) \leq \frac{25}{30^2} \approx 0.028.$$

  Obwohl die Tschebyschew–Ungleichung Abweichungen nach beiden Seiten berücksichtigt, ist die Abschätzung schärfer als bei der einfachen Markow–Ungleichung.
- *Chernoff–Schranke.* Setze

$$P(X_{100} \geq 80) \leq \inf_{\Theta \geq 0} \underbrace{e^{-80\Theta} \left( \frac{1 + e^{\Theta}}{2} \right)^{100}}_{=: f(\Theta)}$$

  Durch Differentiation zeigt man, dass $f(\Theta)$ minimiert wird für $\Theta = \ln 4$. Damit folgt

$$P(X_{100} \geq 80) \leq 4^{-80} \left( \frac{1+4}{2} \right)^{100} \approx 4.26 \cdot 10^{-9}.$$

  Erwarungsgemäß ist dies eine wesentlich schärfere Schranke als bei den beiden anderen Ungleichungen.

□

**Bemerkung 60.9 Schwaches Gesetz der großen Zahlen.** Mit Hilfe der Tschebyschew–Ungleichung kann man das sogenannte schwache Gesetz der großen Zahlen beweisen. Es werde ein Versuch $n$ Mal wiederholt, bei dem das Ereignis $A$ mit Wahrscheinlichkeit $p$ eintritt. Dann strebt die Wahrscheinlichkeit, dass sich die relative Häufigkeit $h_n(A)$ um weniger als $\varepsilon$ von $p$ unterscheidet, gegen 1 für $n \to \infty$

$$\lim_{n \to \infty} P(|h_n(A) - p| < \varepsilon) = 1.$$

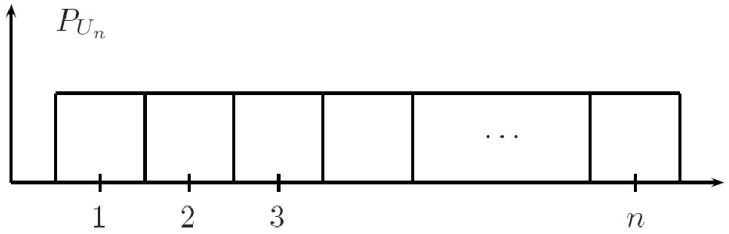
Dabei ist $\varepsilon$ eine beliebig kleine positive Zahl. □

# Kapitel 61

# Wichtige diskrete Verteilungen

**Bemerkung 61.1 Motivation.** Einige diskrete Verteilungen treten sehr häufig auf und tragen einen eigenen Namen. In diesem Abschnitt werden vier dieser Verteilungen genauer betrachten: Gleichverteilung, Binomialverteilung, Poisson–Verteilung und geometrische Verteilung. □

**Bemerkung 61.2 Gleichverteilung.** Sei $\Omega = \{\omega_1, ..., \omega_n\}$ ein Stichprobenraum mit $P(\omega_k) = 1/n$, $k = 1, \ldots, n$, siehe das Laplace–Experiment aus Definition 55.5. Ferner sei $U_n$ eine Zufallsvariable mit $U_n(\omega_k) = k$. Dann ist $P_{U_n}(k) = 1/n$ für $k = 1, \ldots, n$. Eine solche Verteilung heißt (diskrete) Gleichverteilung auf der Menge $\{1, \ldots, n\}$.




Es gelten:

$$\begin{aligned} \mu = E(U_n) &= \sum_{k=1}^{n} k P_{U_n}(k) = \sum_{k=1}^{n} k \frac{1}{n} = \frac{1}{n} \sum_{k=1}^{n} k, \quad \text{arithmetisches Mittel,} \\ \sigma^2 = V(U_n) &= \sum_{k=1}^{n} k^2 P_{U_n}(k) - \mu^2 = \frac{1}{n} \sum_{k=1}^{n} k^2 - \frac{1}{n^2} \left( \sum_{k=1}^{n} k \right)^2. \end{aligned}$$

□

**Beispiel 61.3 Gleichverteilung, Würfeln.** Beim Würfeln liegt eine diskrete Gleichverteilung auf der Menge $\{1, \ldots, 6\}$ vor mit

$$\begin{aligned} \mu &= \frac{1}{6} \sum_{k=1}^{6} k = \frac{21}{6} = 3.5, \\ \sigma^2 &= \frac{1}{6} \left( 1^2 + 2^2 + \ldots + 6^2 \right) - 3.5^2 \approx 2.917 \quad \Longrightarrow \quad \sigma \approx 1.708. \end{aligned}$$

□

**Bemerkung 61.4 Binomialverteilung, Bernoulli[1]–Experiment.** Betrachte ein Experiment, das aus einer $n$–fachen Wiederholung eines Einzelexperiments besteht, bei dem ein Ereignis $A$ jeweils mit Wahrscheinlichkeit $p$ auftritt. Ein solches Experiment heißt Bernoulli–Experiment. Man setzt

$$X_i(\omega) = \begin{cases} 1 & \text{falls } A \text{ beim } i\text{–ten Versuch eintritt,} \\ 0 & \text{sonst.} \end{cases}$$

$A_k$ bezeichnet das Ereignis, dass im Gesamtexperiment gilt

$$X(\omega) := \sum_{i=1}^{n} X_i(\omega) = k.$$

Nach Bemerkung 56.4 (ungeordnete Stichprobe ohne Wiederholung) hat $A_k$ insgesamt $\binom{n}{k}$ Elemente. Jedes solche Element tritt mit Wahrscheinlichkeit $p^k(1-p)^{n-k}$ auf. Die entsprechende Verteilung

$$b_{n,p}(k) := \binom{n}{k} p^k (1-p)^{n-k}$$

heißt Binomialverteilung mit den Parametern $n$ und $p$.

Für den Erwartungswert von $X_i$ gilt

$$E(X_i) = 1 \cdot p + 0 \cdot (1-p) = p \quad \Longrightarrow \quad E(X) = \sum_{i=1}^{n} E(X_i) = np.$$

Die zweite Gleichung folgt aus der Linearität des Erwartungswertes. Da die Einzelereignisse unabhängig sind, folgt

$$V(X) = \sum_{i=1}^{n} V(X_i) = np(1-p),$$

da

$$V(X_i) = E(X_i^2) - (E(X_i))^2 = 1^2 p + 0^2(1-p) - p^2 = p(1-p).$$

□

**Beispiel 61.5 Binomialverteilung, Zufallsabhängigkeit sportlicher Resultate.** A und B tragen ein Tischtennisturnier mit $n = 1, 3, 5, \ldots, 2m+1$ Spielen aus. Wer die meisten Einzelspiele gewinnt, ist Sieger. A gewinnt ein Einzelspiel mit Wahrscheinlichkeit $p = 0.6$. Wie groß sind die Siegeschancen für den schlechteren Spieler B?

B siegt, wenn A $S_n \leq m$ Erfolge erzielt. Die Wahrscheinlichkeit hierfür beträgt

$$P(S_n \leq m) = b_{n,p}(0) + b_{n,p}(1) + \ldots + b_{n,p}(m) = \sum_{k=0}^{m} \binom{n}{k} p^k (1-p)^{n-k}.$$

Man erhält für unterschiedliche Anzahlen von Spielen im Turnier:

$$\begin{aligned}
n = 1: \quad P(S_1 \leq 0) &= \binom{1}{0} 0.6^0 0.4^1 = 10.4 = 0.4, \\
n = 3: \quad P(S_3 \leq 1) &= \binom{3}{0} 0.6^0 0.4^3 + \binom{3}{1} 0.6^1 0.4^2 = 0.4^3 + 3.6\ 0.4^2 \approx 0.352, \\
n = 5: \quad P(S_5 \leq 2) &= \binom{5}{0} 0.6^0 0.4^5 + \binom{5}{1} 0.6^1 0.4^4 + \binom{5}{2} 0.6^2 0.4^3 \approx 0.317, \\
n = 7: \quad P(S_7 \leq 3) &= \ldots \approx 0.290, \\
n = 9: \quad P(S_9 \leq 4) &= \ldots \approx 0.267.
\end{aligned}$$

[1] Jakob Bernoulli (1655 – 1705)

Mit zunehmender Anzahl von Spielen sinkt also die Wahrscheinlichkeit, dass der schlechtere Spieler das Turnier gewinnt. □

**Bemerkung 61.6 Poisson[2]–Verteilung.** Für große $n$ wird das Arbeiten mit der Binomialverteilung unhandlich. Ist $p$ klein ($0 \leq p \leq 0.1$), gibt es eine gute Approximation, die Poisson–Verteilung zum Parameter $\lambda$

$$p(k) := \frac{\lambda^k}{k!} e^{-\lambda}, \quad \lambda > 0, \quad k = 0, 1, 2, \ldots.$$

Man kann zeigen, dass $p(k)$ für $\lambda = np$ die Binomialverteilung $b_{n,p}(k)$ approximiert. Ferner hat eine Poisson–verteilte Zufallsvariable den Erwartungswert $\lambda$ und die Varianz $\lambda$. Durch Umbenennung von „Erfolg“ und „Fehlschlag“ ist die Poisson–Verteilung auch für $0.9 \leq p \leq 1$ eine gute Approximation an die Binomialverteilung. Generell beschreibt die Poisson–Verteilung Ereignisse, die im zeitlichen Verlauf zufällig und unabhängig voneinander auftreten, zum Beispiel

- atomarer Zerfall,
- das Eintreffen von Bedienwünschen an einem Server,
- Anrufe in einem Call–Center,
- das Auftreten von Softwarefehlern in einem Programmsystem.

□

**Beispiel 61.7 Poisson–Verteilung, der große Jubiläumstag.** Genau in einem Jahr feiert ein großer Betrieb seinen 100. Geburtstag. Die Direktion beschließt, allen Kindern von Betriebsangehörigen, die an diesem Tag geboren werden, ein Sparkonto von 3000 Euro anzulegen. Da rund 730 Kinder pro Jahr geboren werden, erwartet man Auslagen von 6000 Euro. Um Zufallsschwankungen vorzubeugen, plant man 15.000 Euro ein. Wie groß ist die Wahrscheinlichkeit, dass das Geld nicht reicht?

Hierbei sind $n = 730$ Kinder/Jahr und $p = 1/365$ Wahrscheinlichkeit, dass der Geburtstag auf den Jubiläumstag fällt. Damit ist der Parameter der Poisson–Verteilung $\lambda = pn = 2$.

Das Geld reicht nicht, falls $k \geq 6$ Kinder geboren werden. Die Wahrscheinlichkeit dafür ist

$$\begin{aligned} p(k \geq 6) &= 1 - p(k \leq 5) = 1 - p(0) - p(1) - \ldots - p(5) \\ &= 1 - \frac{2^0}{0!} e^{-2} - \frac{2^1}{1!} e^{-2} - \ldots - \frac{2^5}{5!} e^{-2} \approx 0.0168. \end{aligned}$$

Die Wahrscheinlichkeit einer unangenehmen Zufallsüberraschung ist also gering. □

**Bemerkung 61.8 Geometrische Verteilung.** Eine diskrete Zufallsvariable $X$, die in einem Bernoulli–Experiment angibt, bei welchem Versuch ein Ereignis $A$ mit Wahrscheinlichkeit $p(A) = p$ zum ersten Mal eintritt, heißt geometrisch verteilt mit Parameter $p$. Die entsprechende Verteilung lautet

$$P_X(k) = p(1-p)^{k-1}.$$

Man kann zeigen:

$$E(X) = \frac{1}{p}, \quad V(X) = \frac{1-p}{p^2}.$$

Die geometrische Verteilung spielt in der Informatik eine große Rolle bei der Modellierung von Wartezeiten. □

[2]Siméon-Denis Poisson (1781 – 1840)

**Beispiel 61.9** Wie lange muss man beim Würfeln warten, bis die Augenzahl 4 erstmalig auftritt?

Mit $p = 1/6$ betragen der Erwartungswert

$$E(X) = \frac{1}{p} = 6$$

und die Varianz

$$V(X) = \frac{1-p}{p^2} = \frac{5/6}{1/36} = 30.$$

□

# Kapitel 62

# Wichtige kontinuierliche Verteilungen

**Bemerkung 62.1 Motivation.** Zufallsvariablen sind nicht immer diskret, sie können oft auch jede beliebige reelle Zahl in einem Intervall $[a, b]$ einnehmen. Beispiele für solche kontinuierlichen Zufallsvariablen sind Größe, Gewicht oder Zeit. In diesen Fällen macht es wenig Sinn, die Wahrscheinlichkeit anzugeben, dass die Zufallsvariable einen bestimmten Wert annimmt (diese Wahrscheinlichkeit ist 0). Man muss Wahrscheinlichkeiten für Intervalle betrachten. Hierzu sind Begriffe wie Dichten notwendig. □

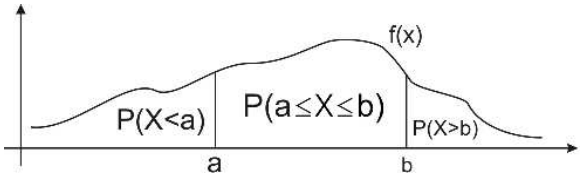
**Definition 62.2 Dichte, Erwartungswert, Varianz.** Sei $X$ eine kontinuierliche Zufallsvariable. Existiert eine Funktion $f : \mathbb{R} \to \mathbb{R}$ mit

1.) $f(x) \geq 0$ für alle $x \in \mathbb{R}$,

2.)
$$\int_{-\infty}^{\infty} f(x)\, dx = 1,$$

3.)
$$P(a \leq X \leq b) = \int_a^b f(x)\, dx,$$

so nennt man $f(x)$ die Dichte von $X$.




Erwartungswert (1. Moment) und Varianz (zentrales 2. Moment) von $X$ sind definiert durch

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)\, dx, \quad \sigma^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx.$$

□

**Definition 62.3 Verteilungsfunktion.** Sei $X$ eine kontinuierliche Zufallsvariable mit Dichte $f(x)$. Dann nennt man ihre Stammfunktion

$$F(X) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt$$

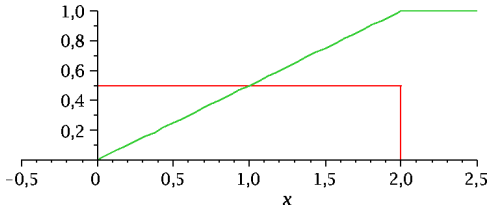
die Verteilungsfunktion von $X$. □

**Beispiel 62.4 Kontinuierliche Gleichverteilung.** Eine kontinuierliche Zufallsvariable, die auf $[a, b]$ gleichverteilt ist, hat die Dichte

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{für } a \le x \le b, \\ 0 & \text{sonst} \end{cases}$$

und die Verteilungsfunktion

$$F(x) = \begin{cases} 0 & \text{für } x < a, \\ \dfrac{x-a}{b-a} & \text{für } a \le x \le b, \\ 1 & \text{für } x > b. \end{cases}$$




□

**Beispiel 62.5 Standardnormalverteilung.** Das ist die wichtigste kontinuierliche Verteilung. Ist $X$ eine kontinuierliche Zufallsvariable mit der Dichte

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

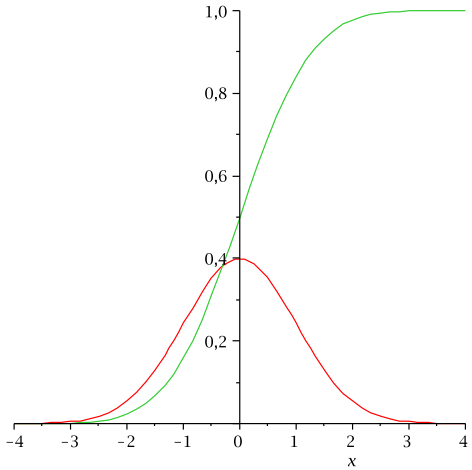
so kann man $P(a < X \le b)$ mit der Stammfunktion

$$\Phi(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt$$

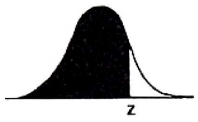
berechnen

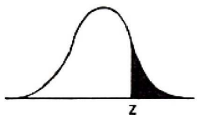
$$P(a < X \le b) = \Phi(b) - \Phi(a).$$

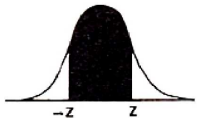
Die Dichte $\varphi(x)$ nennt man normale Dichte, und $\Phi(x)$ ist die Standardnormalverteilung oder $N(0, 1)$–Verteilung.

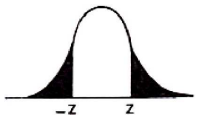



Die Standardnormalverteilung $\Phi(x)$ ist nicht analytisch auswertbar, liegt aber tabelliert vor, siehe Tabelle 62.1. Eine standardnormalverteilte Zufallsvariable hat Erwartungswert 0 und Varianz 1. Daher heißt sie $N(0, 1)$–verteilt. □









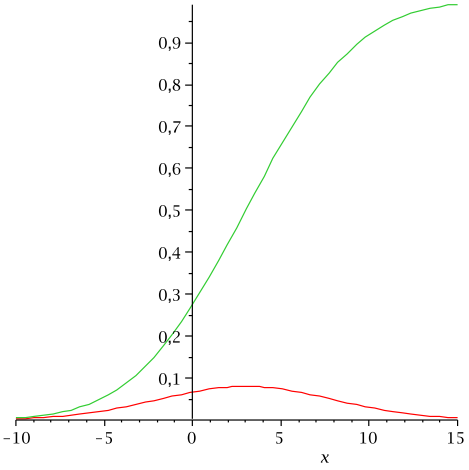

| $z$ | $\Phi(z)$ | $1-\phi(z)$ | $2\Phi(z)-1$ | $2-2\Phi(z)$ |
|---|---|---|---|---|
| 0.0 | .500 | .500 | .0000 | 1.0000 |
| 0.1 | .540 | .460 | .0797 | .9203 |
| 0.2 | .579 | .421 | .159 | .841 |
| 0.3 | .618 | .382 | .236 | .764 |
| 0.4 | .655 | .345 | .311 | .689 |
| 0.5 | .691 | .309 | .383 | .617 |
| 0.6 | .726 | .274 | .451 | .549 |
| 0.7 | .758 | .242 | .516 | .484 |
| 0.8 | .788 | .212 | .576 | .424 |
| 0.9 | .816 | .184 | .632 | .368 |
| 1.0 | .841 | .159 | .683 | .317 |
| 1.1 | .864 | .136 | .729 | .271 |
| 1.2 | .885 | .115 | .770 | .230 |
| 1.3 | .9032 | .0968 | .806 | .194 |
| 1.4 | .9192 | .0808 | .838 | .162 |
| 1.5 | .9332 | .0668 | .866 | .134 |
| 1.6 | .9452 | .0548 | .890 | .110 |
| 1.7 | .9554 | .0446 | .9109 | .0891 |
| 1.8 | .9641 | .0359 | .9281 | .0719 |
| 1.9 | .9713 | .0287 | .9425 | .0575 |
| 2.0 | .9772 | .0228 | .9545 | .0455 |
| 2.1 | .9821 | .0179 | .9643 | .0357 |
| 2.2 | .9861 | .0139 | .9722 | .0278 |
| 2.3 | .9893 | .0107 | .9786 | .0217 |
| 2.4 | .99180 | .00820 | .9836 | .0164 |
| 2.5 | .99379 | .00621 | .9876 | .0124 |
| 2.6 | .99534 | .00466 | .99068 | .00932 |
| 2.7 | .99653 | .00347 | .99307 | .00693 |
| 2.8 | .99744 | .00256 | .99489 | .00511 |
| 2.9 | .99813 | .00187 | .99627 | .00373 |
| 3.0 | .99865 | .00135 | .99730 | .00270 |
| 3.1 | .999032 | .000968 | .99806 | .00194 |
| 3.2 | .999313 | .000687 | .99863 | .00137 |
| 3.3 | .999517 | .000483 | .999033 | .000967 |
| 3.4 | .999663 | .000337 | .999326 | .000674 |
| 3.5 | .999767 | .000233 | .999535 | .000465 |
| 3.6 | .999841 | .000159 | .999682 | .000318 |
| 3.7 | .999892 | .000108 | .999784 | .000216 |
| 3.8 | .9999277 | .0000723 | .999855 | .000145 |
| 3.9 | .9999519 | .0000481 | .9999038 | .0000962 |
| 4.0 | .9999683 | .0000317 | .9999367 | .0000633 |

Tabelle 62.1: Werte der standardisierten Normalverteilung $N(0,1)$.

**Beispiel 62.6 Allgemeine Normalverteilung.** Eine kontinuierliche Zufallsvariable $X$ genügt einer allgemeinen Normalverteilung oder $N(\mu,\sigma^2)$–Verteilung oder Gauß–Verteilung mit Erwartungswert $\mu$ und Varianz $\sigma^2$, wenn ihre Dichte gegeben ist durch

$$\varphi(x)=\frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

siehe Abbildung für $\sigma = 5, \mu = 3$.




Ist $X$ $N(\mu, \sigma^2)$-verteilt, so ist $Y := \dfrac{X - \mu}{\sigma}$ $N(0, 1)$–verteilt. Somit ist die Tabelle der Standardnormalverteilung ausreichend.

Aus $X = \sigma Y + \mu$ folgt beispielsweise

$$\begin{aligned} P(\mu - \sigma \le X \le \mu + \sigma) &= P(\mu - \sigma \le \sigma Y + \mu \le \mu + \sigma) \\ &= P(-\sigma \le \sigma Y \le \sigma) \\ &= P(-1 \le Y \le 1) \approx 0.683 = 68.3\%, \end{aligned}$$

wobei man den Wert für die $N(0, 1)$–Verteilung aus der Tafel abliest. □

**Bemerkung 62.7 Approximation der Binomialverteilung durch die Gauß–Verteilung.** Eine Binomialverteilung mit $n$ Einzelexperimenten mit Wahrscheinlichkeit $p$ kann man durch eine allgemeine Normalverteilung mit Erwartungswert $np$ und Standardabweichung $\sigma = \sqrt{np(1-p)}$ approximieren

$$b_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k} \approx \Phi\left(\frac{k + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right). \qquad (62.1)$$

Diese Approximation ist gut für $np > 5$ und $n(1-p) > 5$, das heißt insbesondere für große $n$ oder $p \approx 0.5$. □

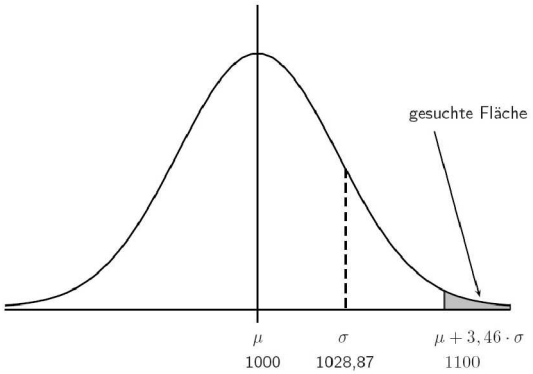
**Beispiel 62.8** Gesucht ist die Wahrscheinlichkeit, dass in 6000 Würfen eines fairen Würfels die Sechs mindestens 1100 Mal auftritt. Man hat $n = 6000$ und $p = 1/6$. Wegen $np = 1000 > 5$ und $n(1-p) = 5000 > 5$ ist die Approximation durch die Gaußverteilung sinnvoll. Man verwendet für die Approximation

$$\mu = np = 1000, \quad \sigma = \sqrt{np(1-p)} = \sqrt{6000\ \frac{1}{6}\ \frac{5}{6}} \approx 28.87.$$

Für die Anzahl der betrachteten Ereignisse $k \ge 1100$ muss man die Beiträge für alle

$k \geq 1100$ addieren. Das ist eine Teleskopsumme und man erhält mit (62.1)

$$\begin{aligned}
b_{6000,\frac{1}{6}}(k \geq 1100) &= \sum_{k=1100}^{6000} b_{6000,\frac{1}{6}}(k) \\
&\approx \sum_{k=1100}^{6000} \left[ \Phi\left( \frac{k+\frac{1}{2}-np}{\sqrt{np(1-p)}} \right) - \Phi\left( \frac{k-\frac{1}{2}-np}{\sqrt{np(1-p)}} \right) \right] \\
&= \sum_{k=1100}^{6000} \left[ \Phi\left( \frac{k+\frac{1}{2}-1000}{28.87} \right) - \Phi\left( \frac{k-\frac{1}{2}-1000}{28.87} \right) \right] \\
&= \Phi\left( \frac{6000+\frac{1}{2}-1000}{28.87} \right) - \Phi\left( \frac{1100-\frac{1}{2}-1000}{28.87} \right) \\
&\approx 1-(3.446) \approx 1-0.999716 = 0.000284.
\end{aligned}$$




Die Wahrscheinlichkeit, mehr als 1100 Sechsen zu würfeln, beträgt demnach ungefähr 0.0284%. □

Der wichtigste Grund für die Bedeutung der Gauß–Verteilung ist der folgende Satz.

**Satz 62.9 Zentraler Grenzwertsatz.** *Seien $X_1, X_2, \ldots$ unabhängige Zufallsvariablen, die alle die gleiche Verteilung und somit auch denselben Erwartungswert $\mu$ und dieselbe Varianz $\sigma^2$ besitzen. Ferner seien $Y_n := X_1 + \ldots + X_n$ und*

$$Z_n := \frac{Y_n - n\mu}{\sigma\sqrt{n}}$$

*bezeichne die Standardisierte von $Y_n$, vergleiche Definition 59.17. Dann konvergiert die Verteilungsfunktion $F_n(x)$ von $Z_n$ für $n \to \infty$ gegen die Standardnormalverteilung $\Phi(x)$.*

**Beweis:** Der Beweis ist aufwändig, siehe zum Beispiel R. Nelson: Probability, Stochastic Processes and Queueing Theory, Springer, New York, 1995, Abschnitt 5.5.6. ■

**Bemerkung 62.10**

- Man beachte, dass die einzelnen Zufallsvariablen nicht normalverteilt sein müssen. Ihre Verteilung kann beliebig sein.
- Die Normalverteilung ist also eine sinnvolle Approximation in allen Fällen, in denen sich eine Zufallsvariable aus vielen gleichartigen Einzeleinflüssen zusammensetzt.
  Beispiel: Eine Messung wird oft wiederholt. Dann approximieren die Ergebnisse eine Gauß–Verteilung.

□

# Kapitel 63

# Multivariate Verteilungen und Summen von Zufallsvariablen

**Bemerkung 63.1 Motivation.** Manchmal möchte man das Zusammenwirken mehrerer Zufallsvariabler $X_1, \ldots, X_n$ studieren. Man sucht in diesem sogenannten multivariaten Fall Aussagen über die gemeinsame Verteilung. Desweiteren sucht man Aussagen über die Verteilung von Summen von Zufallsvariablen. □

**Bemerkung 63.2 Erweiterung der Grundbegriffe auf multivariate Zufallsvariablen.** Mehrere Zufallsvariablen $X_1, \ldots, X_n$ fasst man zu einem Vektor $\mathbf{X} = (X_1, \ldots, X_n)^T$ zusammen. Im kontinuierlichen Fall ist die resultierende Dichte eine Funktion mehrerer Variabler. Für diese gemeinsame Dichte gilt

$$\begin{aligned} f(x_1, \ldots, x_n) &\geq 0 \quad \forall x_1, \ldots, x_n \in \mathbb{R}, \\ \int_{\mathbb{R}^n} f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n &= 1, \\ P(a_1 \leq X_1 \leq b_1, \ldots, a_n \leq X_n \leq b_n) &= \int_{a_n}^{b_n} \ldots \int_{a_1}^{b_1} f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n. \end{aligned}$$

Ferner betrachtet man die multivariate Verteilungsfunktion

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_n} \ldots \int_{-\infty}^{x_1} f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n.$$

Statt eines einzelnen Erwartungswerts hat man einen Erwartungswertvektor

$$\boldsymbol{\mu} = (E(X_1), \ldots, E(X_n))^T.$$

Varianzen und Kovarianzen fasst man zu einer symmetrischen und positiv definiten Kovarianzmatrix zusammen

$$\Sigma = \begin{pmatrix} \mathrm{V}(X_1) & \mathrm{Cov}(X_1, X_2) & \ldots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{V}(X_2) & \ldots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \ldots & \mathrm{V}(X_n) \end{pmatrix}.$$

□

**Beispiel 63.3 Multivariate Normalverteilung.** Die wichtigste multivariate Verteilung ist die multivariate Normalverteilung oder $N_n(\boldsymbol{\mu}, \Sigma)$–Verteilung. Sei $\mathbf{X} = (X_1, \ldots, X_n)^T$ ein Vektor von normalverteilten Zufallsvariablen mit Erwartungswert $\boldsymbol{\mu} = (E(X_1), \ldots, E(X_n))^T$ und Kovarianzmatrix $\Sigma$, dann besitzt die multivariate Normalverteilung die Dichte

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})} \quad \text{mit} \quad \mathbf{x}, \boldsymbol{\mu} \in \mathbb{R}^n.$$

□

**Beispiel 63.4** Eine Apfelbaumplantage mit gleich alten Bäumen werde durch drei normalverteilte Zufallsvariablen beschrieben:

| | | |
|---|---|---|
| $X_1$: | Höhe eines Baumes [m] | $N(4,1)$–verteilt, |
| $X_2$: | Ertrag [kg] | $N(20,100)$–verteilt, |
| $X_3$: | Zahl der Blätter [1000 Stück] | $N(20,225)$–verteilt. |

Diese Zufallsvariablen seien korreliert mit

$$\mathrm{Cov}(X_1, X_2) = 9, \quad \mathrm{Cov}(X_1, X_3) = 12.75, \quad \mathrm{Cov}(X_2, X_3) = 120.$$

Dann liegt eine $N_3(\boldsymbol{\mu}, \Sigma)$–Verteilung vor mit

$$\boldsymbol{\mu} = \begin{pmatrix} 4 \\ 20 \\ 20 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & 9 & 12.75 \\ 9 & 100 & 120 \\ 12.75 & 120 & 225 \end{pmatrix}.$$

□

Es stellt sich die Frage, ob man unter geeigneten Voraussetzungen die gemeinsame Dichte $f(x_1, \ldots, x_n)$ aus den einzelnen Dichten $f_1(x_1), \ldots, f_n(x_n)$ berechnen kann.

**Satz 63.5 Gemeinsame Dichte unabhängiger Zufallsvariablen.** *Seien $X_1, \ldots, X_n$ unabhängige Zufallsvariablen mit Dichten $f_1(x_1), \ldots, f_n(x_n)$, so hat $\mathbf{X} = (X_1, \ldots, X_n)^T$ die gemeinsame Dichte*

$$f(x_1, \ldots, x_n) = f_1(x_1) \cdots f_n(x_n). \tag{63.1}$$

*Hat umgekehrt $\mathbf{X} = (X_1, \ldots, X_n)^T$ eine gemeinsame Dichte in der Produktdarstellung (63.1), so sind $X_1, \ldots, X_n$ unabhängig.*

**Beweis:** Siehe Literatur. ■

**Beispiel 63.6** Zwei unabhängige Zufallsvariablen $X_1, X_2$ seien $N(\mu_1, \sigma_1^2)$– beziehungsweise $N(\mu_2, \sigma_2^2)$–verteilt. Da Unabhängigkeit Unkorreliertheit impliziert, vergleiche Satz 59.26,v), hat die Korellationsmatrix Diagonalgestalt

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}.$$

Mit $\det \Sigma = \sigma_1^2 \sigma_2^2$ und

$$\Sigma^{-1} = \begin{pmatrix} 1/\sigma_1^2 & 0 \\ 0 & 1/\sigma_2^2 \end{pmatrix},$$

hat $X = (X_1, X_2)^T$ nach Beispiel 63.3 eine multivariate Normalverteilung mit Dichte

$$\begin{aligned} f(x_1, x_2) &= \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}} e^{-\frac{1}{2}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]} \\ &= \frac{1}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}} \frac{1}{\sqrt{2\pi}\,\sigma_2} e^{-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}}. \end{aligned}$$

Dies ist gerade das Produkt der zwei einzelnen Dichten $f_1(x_1)$ und $f_2(x_2)$. □

Nun wird die Summe zweier Zufallsvariablen betrachtet.

**Definition 63.7 Faltung.** Falls für die Funktionen $f, g : \mathbb{R} \to \mathbb{R}$ das Integral

$$(f * g)(x) := \int_{-\infty}^{\infty} f(x-y)g(y)\ dy$$

existiert, so wird $(f * g)(x)$ die Faltung (englisch convolution) von $f(x)$ und $g(x)$ genannt. □

**Satz 63.8 Summe unabhängiger kontinuierlicher Zufallsvariabler.** *Seien $X_1, X_2$ unabhängige kontinuierliche Zufallsvariable mit den Dichten $f_1(x), f_2(x)$, so hat $X_1 + X_2$ die Dichte $(f_1 * f_2)(x)$.*

**Beweis:** Mit $B := \{(x_1, x_2) \mid x_1 + x_2 \leq s\}$ ergibt sich für die Verteilung von $X_1 + X_2$

$$P(X_1 + X_2 \leq s) = \iint_B \underbrace{f_1(x_1)f_2(x_2)}_{\text{unabhängig}}\ dx_1 dx_2.$$

Mit der Substitution $u := x_1 + x_2$ folgt:

$$P(X_1 + X_2 \leq s) = \int_{-\infty}^{s} \underbrace{\left( \int_{-\infty}^{\infty} f_1(u - x_2)f_2(x_2)\ dx_2 \right)}_{(f_1 * f_2)(u)} du.$$

■

**Folgerung 63.9 Summe unabhängiger normalverteilter Zufallsvariabler.** *Seien $X_1, X_2$ unabhängige kontinuierliche Zufallsvariablen mit $N(\mu_1, \sigma_1^2)$– beziehungsweise $N(\mu_2, \sigma_2^2)$–Verteilung. Dann ist $X = X_1 + X_2$ ebenfalls $N(\mu, \sigma^2)$–verteilt, und es gelten*

$$\mu = \mu_1 + \mu_2, \quad \sigma^2 = \sigma_1^2 + \sigma_2^2.$$

**Beweis:** Die Aussage folgt aus Satz 63.8, siehe U. Krengel (2003), Satz 11.9. ■

Auch im Fall diskreter Zufallsvariablen gibt es vergleichbare Aussagen.

**Definition 63.10 Diskrete Faltung.** Für $\mathbf{f} = (f_i)_{i\in\mathbb{Z}}, \mathbf{g} = (g_i)_{i\in\mathbb{Z}}$ definiert man die diskrete Faltung von $f$ und $g$ durch

$$(\mathbf{f} * \mathbf{g})_i := \sum_{j\in\mathbb{Z}} f_{i-j}g_j.$$

□

**Satz 63.11 Summe unabhängiger diskreter Zufallsvariabler.** *Seien $X_1, X_2$ unabhängige diskrete Zufallsvariablen mit Verteilungen $P_{X_1}, P_{X_2}$. Dann hat $P_{X_1+X_2}$ die Verteilung $P_{X_1} * P_{X_2}$, wobei $*$ die diskrete Faltung bezeichnet.*

**Beweis:** Siehe Literatur. ■

**Folgerung 63.12 Summe unabhängiger Poisson–verteilter Zufallsvariabler.** *Seien $X_1, X_2$ unabhängige diskrete Zufallsvariablen, die einer Poisson–Verteilung mit Parameter $\lambda_1$ beziehungsweise $\lambda_2$ genügen, kurz $P(\lambda_1)$-, $P(\lambda_2)$–verteilt. Dann ist $X_1 + X_2$ $P(\lambda_1 + \lambda_2)$–verteilt.*

**Beweis:** Die Aussage folgt aus Satz 63.11, Übungsaufgabe. ■

**Beispiel 63.13** Beim radioaktiven Zerfall einer Substanz werden ionisierende Teilchen frei. Mit einem Geiger[1]–Müller[2]–Zählrohr zählt man die innerhalb einer Minute eintreffenden Teilchen. Sie sind Poisson–verteilt. Hat man zwei radioaktive Substanzen mit Poisson–Verteilung $P(\lambda_1)$ und $P(\lambda_2)$, so genügt die Gesamtheit der pro Zeitintervall produzierten Teilchen einer $P(\lambda_1 + \lambda_2)$–Verteilung. □

[1] Johannes Wilhelm Geiger (1882 – 1945)
[2] Walther Müller (1905 – 1979)

# Kapitel 64

# Parameterschätzung und Konfidenzintervalle

**Bemerkung 64.1 Motivation.** Bisher sind wir stets von theoretischen Modellen, zum Beispiel einem fairen Würfel, ausgegangen, die erlauben, Parameter wie Erwartungswert oder Varianz einer Verteilung exakt zu berechnen. In vielen realen Situationen kennt man jedoch nur den Verteilungstyp und muss auf Grund von Stichproben die Parameter schätzen. Dieses Kapitel stellt Vorgehensweisen dazu vor.

Die geschätzten Parameter werden im allgemeinen fehlerhaft sein, weil zum Beispiel nur wenige Stichproben vorhanden sind. Eine wichtige Frage besteht darin, ob sich ein Vertrauensintervall angeben lässt, innerhalb dessen die Parameter mit vorgegebener Sicherheit liegen. Diese Fragestellung wird ebenfalls in diesem Kapitel betrachtet. □

**Definition 64.2 Umfang einer Stichprobe, Stichprobenwerte.** Gegeben seien $n$ Beobachtungswerte $x_1, \ldots, x_n$ eines Zufallsexperiments. Dann nennt man $(x_1, \ldots, x_n)^T$ Stichprobe vom Umfang $n$. Die einzelnen $x_i$ heißen Stichprobenwerte. □

**Beispiel 64.3** In einer Kiste befinden sich 10.000 Schrauben. Ein Teil davon ist fehlerhaft. Für eine Stichprobe werden 100 Schrauben entnommen. Die Zufallsvariable $X_i$ beschreibt den Zustand der $i$–ten entnommenen Schraube

$$X_i(\omega) = \begin{cases} 0 & \text{falls } i\text{–te Schraube in Ordnung,} \\ 1 & \text{falls } i\text{–te Schraube defekt.} \end{cases}$$

Eine konkrete Realisierung des Zufallsvektors $(X_1, \ldots, X_{100})^T$ liefert beispielsweise die Stichprobe

$$(x_1, \ldots, x_{100})^T = (0, 1, 0, 0, 1, 0, \ldots, 0, 1)^T.$$

□

So wie man bei Zufallsvariablen Parameter wie Erwartungswert oder Varianz zugeordnet, kann man auch für Stichproben Kenngrößen definieren.

**Definition 64.4 Mittelwert, Varianz, Standardabweichung.** Für eine Stichprobe $(x_1, \ldots, x_n)^T$ definiert man:

- den Mittelwert durch

$$\overline{x} := \frac{1}{n}(x_1 + \ldots + x_n),$$

- die Varianz durch
$$s^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$
- die Standardabweichung durch $s := \sqrt{s^2}$.

□

**Bemerkung 64.5**

- Man kann zeigen, dass der Mittelwert $\overline{x}$ und die Varianz $s^2$ geeignete Approximationen an den Erwartungswert $\mu$ und die Varianz $\sigma^2$ einer Zufallsvariablen sind.
- Die Tatsache, dass im Nenner von $s^2$ die Größe $n-1$ statt $n$ steht, hat tiefere theoretische Hintergründe, auf die hier nicht eingangen wird, siehe zum Beispiel Hartmann, Satz 21.9.
- Ähnlich zum Verschiebungssatz, Satz 59.14, gibt eine häufig benutzte Formel zum Berechnen der Varianz einer Stichprobe
$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right).$$
Das folgt aus
$$\sum_{i=1}^{n}(x_i-\overline{x})^2 = \sum_{i=1}^{n}\left(x_i^2 - 2x_i\overline{x} + \overline{x}^2\right) = \sum_{i=1}^{n} x_i^2 - 2\overline{x}\underbrace{\sum_{i=1}^{n} x_i}_{n\overline{x}} + n\overline{x}^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2.$$

□

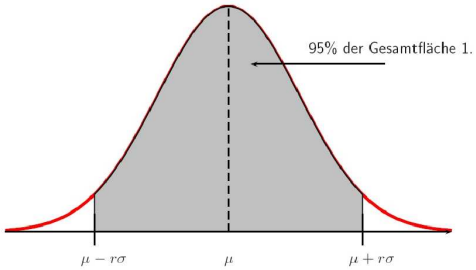
**Beispiel 64.6 Wahlumfrage.** Bei einer Wahlumfrage geben 400 von 1000 Personen an, die Partei $A$ wählen zu wollen. Das Umfrageinstitut prognostiziert auf Grund dieser Stichprobe einen Wahlausgang mit 40% aller Stimmen für Partei $A$. □

**Bemerkung 64.7 Konfidenzintervalle.** In Beispiel 64.6 werden verschiedene Stichproben zu leicht unterschiedlichen Resultaten führen, die wiederum im Allgemeinen alle vom tatsächlichen Wahlausgang abweichen. In der Praxis ist es nötig, statt eines einzelnen Werts $p = 0.4$ ein Vertrauensintervall (Konfidenzintervall) $[p_u, p_o]$ anzugeben, innerhalb dessen das Endresultat mit einer vorgegebenen Wahrscheinlichkeit (Konfidenzniveau) von beispielsweise 95% liegt. □

**Beispiel 64.8 Wahlumfrage aus Beispiel 64.6.** Man geht von einer Binomialverteilung aus und schätzt $p$ durch $p = 400/1000 = 0.4$ ab. Seien
$$X_i = \begin{cases} 1 & \text{Befragte/r } i \text{ wählt } A, \\ 0 & \text{Befragte/r } i \text{ wählt } A \text{ nicht.} \end{cases}$$
und $X := \sum_{i=1}^{1000} X_i$. Dann gelten nach Bemerkung 61.4 mit $p = 0.4$ und $n = 1000$
$$\begin{aligned} \mu &= E(X) = np = 400, \\ \sigma^2 &= V(X) = np(1-p) = 240 \quad \Longrightarrow \quad \sigma \approx 15.49. \end{aligned}$$
Wegen $np = 400 > 5$ und $n(1-p) = 600 > 5$ kann man für $X$ auch eine Normalverteilung mit $\mu = 400$ und $\sigma = 15.49$ annehmen. Man sucht ein Intervall

$[\mu - r\sigma, \mu + r\sigma]$, innerhalb dessen das Integral über die Dichtefunktion den Wert 0.95 annimmt

$$\frac{1}{\sqrt{2\pi}\sigma} \int_{\mu - r\sigma}^{\mu + r\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx = 0.95.$$

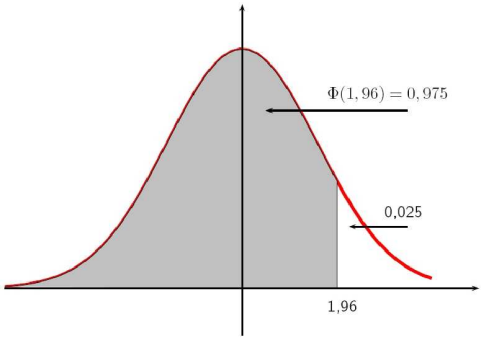



Tabelliert ist die Standardnormalverteilung $\Phi(x)$. Mit der Substitution $y = (x - \mu)/\sigma$ erhält man

$$\frac{1}{\sqrt{2\pi}} \int_{-r}^{r} e^{-\frac{y^2}{2}} \, dy = 0.95.$$

Aus Symmetriegründen ist dies gleichbedeutend mit

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{r} e^{-\frac{y^2}{2}} \, dy = 0.975.$$




Man findet $\Phi(1.96) \approx 0.975$. Somit ist $r = 1.96$ und man erhält das Konfidenzintervall

$$[\mu - r\sigma, \mu + r\sigma] = [400 - 1.96 \cdot 15.49, 400 + 1.96 \cdot 15.49] \approx [369.6, 430.4].$$

Bei einem Konfidenzniveau von 95% erzielt Partei $A$ also zwischen 36.96% und 43.04% der Stimmen.

Möchte man ein kleineres Konfidenzintervall, muss man mehr Personen befragen. Der Erwartungswert $\mu$ skaliert sich linear zur Anzahl der Befragten, die Standardabweichung aber nur mit $\sqrt{n}$. Damit wird das Verhältnis der Abweichung $r\sigma$ zum Erwartungswert $\mu$ geringer und das Konfidenzintervall kleiner. □

**Beispiel 64.9 Überbuchung eines Flugzeugs.** Ein Flugzeug hat 200 Sitze. Wie viele Reservierungen dürfen angenommen werden, wenn erfahrungsgemäß 5% aller Passagiere nicht erscheinen? Die Fluggesellschaft ist bereit, in 1 von 50 Fällen in Verlegenheit zu geraten.

Die Aufgabenstellung kann als das Vorhandensein einer Stichprobe aufgefasst werden, in welcher 5% der Passagiere nicht erschienen sind. Sei $n$ die Anzahl der Reservierungen und $X$ die Anzahl der tatsächlich erscheinenden Passagiere. Legt man eine Binomialverteilung zu Grunde mit $p = 0.95$, so gelten

$$\begin{aligned} \mu &= E(X) = np = 0.95n, \\ \sigma^2 &= np(1-p) = 0.0475n \quad \Longrightarrow \quad \sigma = 0.2179\sqrt{n}. \end{aligned}$$

Es ist klar, dass die Anzahl der Reservierungen $n \geq 200$ ist. Damit sind

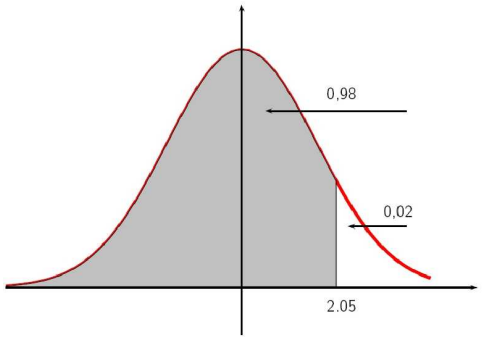
$$0.95n \geq 0.95 \cdot 200 = 190 > 5, \quad np(1-p) \geq 200 \cdot 0.95 \cdot 0.05 = 9.5 > 5.$$

Man darf also die Normalverteilung als Approximation nutzen. Die Wahrscheinlichkeit, dass höchstens 200 Passagiere kommen, soll mindestens 0.98 sein

$$\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{200} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx \geq 0.98 = 1 - \frac{1}{50}.$$

Transformation auf Normalverteilung, analog zum Beispiel 64.8, ergibt

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{200-\mu}{\sigma}} e^{\frac{-y^2}{2}}\, dy \geq 0.98.$$




Der Tabelle der Standardnormalverteilung entnimmt man

$$\Phi(2.05) \approx 0.98.$$

Damit folgt

$$\frac{200-\mu}{\sigma} \geq 2.05 \quad \Longleftrightarrow \quad 200 \geq 2.05 \cdot 0.2179\sqrt{n} + 0.95n.$$

Das ist eine quadratische Gleichung in $\sqrt{n}$ mit der einzigen positiven Lösung

$$\sqrt{n} = 14.276279 \quad \Longleftrightarrow \quad n = 203.81215.$$

Da $n$ eine natürliche Zahl sein muss, dürfen somit 203 Reservierungen angenommen werden. □

# Kapitel 65

# Hypothesentests

**Bemerkung 65.1 Motivation.** Bei Hypothesentests will man eine gewisse Annahme über eine Zufallsvariable daraufhin überprüfen, ob sie korrekt ist, beispielsweise:

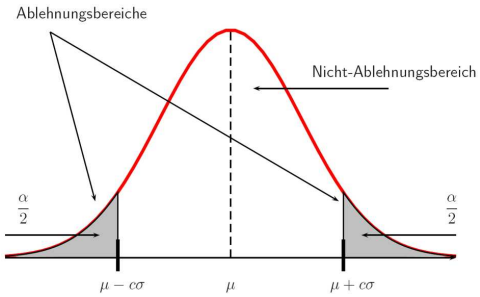
- Ist eine Münze fair, das heißt $p = 1/2$?
- Sind die Rechner von Hersteller $A$ zuverlässiger als von Hersteller $B$?

Ein statistisches Experiment soll dabei eine Entscheidung mit einer vorgegebenen Irrtumswahrscheinlichkeit (Signifikanzniveau) ermöglichen. Gegenüber den Verfahren aus Kapitel 64 kann man in den Rechnungen die Hypothese mit verwenden, hat also mehr in der Hand. □

**Beispiel 65.2 Parametertest am Beispiel eines Münzexperiments.** Man beobachtet das Ereignis $A =$ „Münze zeigt Kopf“ und will die Hypothese $p_0 = p(A) = 1/2$ überprüfen, indem man 200 Münzwürfe durchführt

$$X_i = \begin{cases} 1 & \text{Münze zeigt Kopf beim } i\text{–ten Wurf,} \\ 0 & \text{Münze zeigt Zahl beim } i\text{–ten Wurf.} \end{cases}$$

Wie weit darf $S_{200} := \sum_{i=1}^{200} X_i$ sich vom Erwartungswert 100 unterscheiden, damit man mit einer Irrtumswahrscheinlichkeit $\alpha = 0.05$ (Signifikanzniveau von $1 - \alpha = 0.95$) die Hypothese $p_0 = 1/2$ nicht ablehnt?




Man legt eine Binomialverteilung mit $n = 200$, $p = 0.5$ zu Grunde, die man wiederum durch eine Normalverteilung mit

$$\mu = np = 100, \quad \sigma = \sqrt{np(1-p)} = \sqrt{50} \approx 7.07$$

approximieren kann. Die Rechnung ist analog zum Beispiel 64.8. Wegen $\Phi(1.96) \approx 0.975$ ist $c = 1.96$ für $\alpha = 0.05$. Tritt bei 200 Würfen eine Kopfzahl $S_n$ außerhalb von

$$[\mu - c\sigma, \mu + c\sigma] \approx [86.14, 113.86]$$

auf, wird man die Hypothese $p_0 = 1/2$ auf einem Signifikanzniveau 0.95 ablehnen. Andernfalls wird man sie nicht ablehnen. □

**Bemerkung 65.3**

- Eine Hypothese an einen Parameter, etwa $p_0 = 1/2$, nennt man auch Nullhypothese $H_0$, die Gegenannahme, zum Beispiel $p \neq 1/2$, ist die Gegenhypothese $H_1$.
- Bei Hypothesentests können 2 Arten von Fehlern auftreten:
  - Fehler 1. Art: Die Hypothese wird abgelehnt, obwohl sie richtig ist. Das wird durch die Irrtumswahrscheinlichkeit $\alpha$ beschrieben.
  - Fehler 2. Art: Die Hypothese wird angenommen, obwohl sie falsch ist. Dieser Fehler kann insbesondere für kleines $\alpha$ sehr groß sein.

□

**Bemerkung 65.4 Der $\chi^2$–Test („Chi–Quadrat–Test").** Der $\chi^2$–Test ist einer der wichtigsten Tests. Er wird bei folgendem Problem angewandt. Ein Versuch habe $m$ mögliche Ausgänge. Man testet die Hypothese $H_0$, dass die Resultate mit vorgegebenen Wahrscheinlichkeiten $p_1, \ldots, p_m$ auftreten. Trifft $H_0$ zu, erwartet man bei $n$ Versuchen als Häufigkeit für die einzelnen Ausgänge: $np_1, \ldots, np_m$. In Wirklichkeit werden die Häufigkeiten $X_1, \ldots, X_m$ beobachtet. Als Maß für die Abweichung von $X_i$ und $np_i$ verwendet man

$$\chi^2 := \sum_{i=1}^{m} \frac{(X_i - np_i)^2}{np_i}$$

Ist $\chi^2$ „zu groß", wird man $H_0$ ablehnen.

Das Maß für die Abweichung $\chi^2$ ist als Summe von Zufallsvariablen selbst eine Zufallsvariable. Um zu beurteilen, was „zu groß" bedeutet, ist es sinnvoll den Erwartungswert von $\chi^2$ zu kennen. Ist jedes $X_i$ $b_{n,p_i}$–verteilt, gilt

$$V(X_i) = E((X_i - np_i)^2) = np_i(1 - p_i)$$

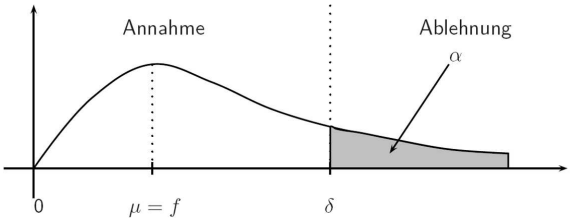
und aus der Linearität des Erwartungswerts folgt

$$\begin{aligned} E(\chi^2) &= \sum_{i=1}^{m} \frac{1}{np_i} E((X_i - np_i)^2) = \sum_{i=1}^{m} \frac{1}{np_i} np_i(1 - p_i) \\ &= \sum_{i=1}^{m} 1 - \sum_{i=1}^{m} p_i = m - 1. \end{aligned}$$

$f := m - 1$ bezeichnet die Freiheitsgrade der $\chi^2$–Verteilung, das heißt $m - 1$ der $p_i$, $i = 1, \ldots, m$, sind frei wählbar. Es gilt also:

$$\mu = E(\chi^2) = f.$$

Der typische Verlauf einer $\chi^2$-Verteilung sieht folgendermaßen aus:




Für einen gegebenen Freiheitsgrad $f$ und eine Irrtumswahrscheinlichkeit $\alpha$ ist die $\chi^2_f$–Verteilung tabelliert. Man nimmt $H_0$ an, falls der berechnete $\chi^2$–Wert kleiner oder gleich $\delta$ ist. Anderenfalls lehnt man $H_0$ ab. □

**Beispiel 65.5** Man will mit 120 Würfen nachprüfen, ob ein Würfel „fair“ ist, das heißt ob alle Augenzahlen gleich wahrscheinlich sind

$$p_1 = \ldots = p_6 = \frac{1}{6}.$$

Als Ergebnis erhält man

| Augenzahl $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| beobachtete Häufigkeit $X_i$ | 15 | 21 | 25 | 19 | 14 | 26 |
| erwartete Häufigkeit $np_i$ | 20 | 20 | 20 | 20 | 20 | 20 |

Daraus folgt

$$\chi^2 = \frac{(15-20)^2}{20} + \frac{(21-20)^2}{20} + \ldots + \frac{(26-20)^2}{20} \approx 6.2.$$

Man hat $f = 6 - 1 = 5$ Freiheitsgrade. Gibt man eine Irrtumswahrscheinlichkeit von $\alpha = 0.1$ vor, so findet man in einer Tabelle

$$p\big(\chi^2 \leq \underbrace{9.24}_{\delta}\big) = \underbrace{0.9}_{1-\alpha}.$$

Wegen $\chi^2 = 6.2 \leq 9.24 = \delta$ akzeptiert man die Hypothese $H_0$, dass alle Augenzahlen gleich wahrscheinlich sind. □

**Bemerkung 65.6** Möchte man eine $N(\mu, \sigma^2)$–Verteilung mit dem $\chi^2$–Test für gegebenes $\mu, \sigma^2$ verifizieren, teilt man die Verteilung in $m$, beispielsweise gleich wahrscheinliche, Klassen ein, siehe Abbildung 65.1.

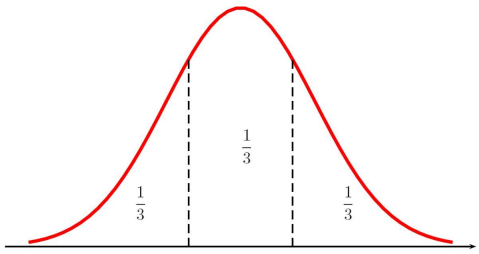



Abbildung 65.1: Einteilung einer Normalverteilung in 3 gleich wahrscheinliche Klassen.

Dann überprüft man, ob die experimentell ermittelten Häufigkeiten in jeder Klasse das $\chi^2$–Kriterium zu einer vorgegebenen Irrtumswahrscheinlichkeit $\alpha$ bei $f = m - 1$ Freiheitsgraden erfüllen. □

| $f \setminus \alpha$ $(P)$ | 0.99 | 0.975 | 0.95 | 0.90 | 0.10 | 0.05 | 0.025 | 0.01 |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.00016 | 0.00098 | 0.00393 | 0.01579 | 2.70554 | 3.84146 | 5.02389 | 6.63490 |
| 2 | 0.00201 | 0.00506 | 0.10259 | 0.21072 | 4.60517 | 5.99147 | 7.37776 | 9.21034 |
| 3 | 0.11483 | 0.21580 | 0.35185 | 0.58438 | 6.25139 | 7.81473 | 9.34840 | 11.3449 |
| 4 | 0.29711 | 0.48442 | 0.71072 | 1.06362 | 7.77944 | 9.48773 | 11.1433 | 13.2767 |
| 5 | 0.55430 | 0.83121 | 1.14548 | 1.61031 | 9.23635 | 11.0705 | 12.8325 | 15.0863 |
| 6 | 0.87209 | 1.23735 | 1.63539 | 2.20413 | 10.6446 | 12.5916 | 14.4494 | 16.8119 |
| 7 | 1.23904 | 1.68987 | 2.16735 | 2.83311 | 12.0170 | 14.0671 | 16.0128 | 18.4753 |
| 8 | 1.64648 | 2.17973 | 2.73264 | 3.48954 | 13.3616 | 15.5073 | 17.5346 | 20.0902 |
| 9 | 2.08781 | 2.70039 | 3.32511 | 4.16816 | 14.6837 | 16.9190 | 19.0228 | 21.6660 |
| 10 | 2.55821 | 3.24697 | 3.94030 | 4.86518 | 15.9871 | 18.3070 | 20.4831 | 23.2093 |
| 11 | 3.0535 | 3.8158 | 4.5748 | 5.5778 | 17.275 | 19.675 | 21.920 | 24.725 |
| 12 | 3.5706 | 4.4038 | 5.2260 | 6.3038 | 18.549 | 21.026 | 23.337 | 26.217 |
| 13 | 4.1069 | 5.0087 | 5.8919 | 7.0415 | 19.812 | 22.362 | 24.736 | 27.688 |
| 14 | 4.6604 | 5.6287 | 6.5706 | 7.7895 | 21.064 | 23.685 | 26.119 | 29.143 |
| 15 | 5.2294 | 6.2621 | 7.2604 | 8.5468 | 22.307 | 24.996 | 27.488 | 30.578 |
| 16 | 5.812 | 6.908 | 7.962 | 9.312 | 23.54 | 26.30 | 28.85 | 32.00 |
| 17 | 6.408 | 7.564 | 8.672 | 10.09 | 24.77 | 27.59 | 30.19 | 33.41 |
| 18 | 7.015 | 8.231 | 9.390 | 10.86 | 25.99 | 28.87 | 31.53 | 34.81 |
| 19 | 7.633 | 8.907 | 10.12 | 11.65 | 27.20 | 30.14 | 32.85 | 36.19 |
| 20 | 8.260 | 9.591 | 10.85 | 12.44 | 28.41 | 31.41 | 34.17 | 37.57 |
| 21 | 8.897 | 10.28 | 11.59 | 13.24 | 29.62 | 32.67 | 35.48 | 38.93 |
| 22 | 9.542 | 10.98 | 12.34 | 14.04 | 30.81 | 33.92 | 36.78 | 40.29 |
| 23 | 10.20 | 11.69 | 13.09 | 14.85 | 32.00 | 35.17 | 38.08 | 41.64 |
| 24 | 10.86 | 12.40 | 13.85 | 15.66 | 33.20 | 36.42 | 39.36 | 42.98 |
| 25 | 11.52 | 13.12 | 14.61 | 16.47 | 34.38 | 37.65 | 40.65 | 44.31 |
| 26 | 12.20 | 13.84 | 15.38 | 17.29 | 35.56 | 38.89 | 41.92 | 45.64 |
| 27 | 12.88 | 14.57 | 16.15 | 18.11 | 36.74 | 40.11 | 43.19 | 46.96 |
| 28 | 13.56 | 15.31 | 16.93 | 18.94 | 37.92 | 41.34 | 44.46 | 48.28 |
| 29 | 14.26 | 16.05 | 17.71 | 19.77 | 39.09 | 42.56 | 45.72 | 49.59 |
| 30 | 14.95 | 16.79 | 18.49 | 20.60 | 40.26 | 43.77 | 46.98 | 50.89 |

Tabelle 65.1: Schranken für $\chi^2$ bei $f$ Freiheitsgraden.
Bei 5 Freiheitsgraden sind beispielsweise

$$\begin{aligned} P(\chi^2 \geqq 0.5543) &= 0.99, \\ P(\chi^2 \geqq 0.83121) &= 0.975, \\ &\vdots \\ P(\chi^2 \geqq 15.0863) &= 0.01. \end{aligned}$$

# Kapitel 66

# Die Methode der kleinsten Quadrate

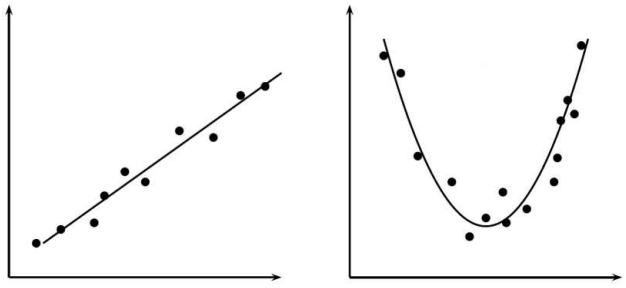
**Bemerkung 66.1 Motivation, Problemstellung.** In einem Experiment interessiert man sich für die Beziehung zwischen zwei Variablen $x$ und $y$. Hierzu hat man viele Wertepaare

$$(x_1, y_1), \ldots, (x_n, y_n)$$

gemessen. Die Messungen können Fehler in Form von statistischen Fluktuationen enthalten. Man möchte nun die Beziehung zwischen $x$ und $y$ durch ein Polynom

$$y = f(x) = \sum_{k=0}^{m} a_k x^k$$

approximieren, beispielsweise durch eine Gerade $m = 1$ oder eine Parabel $m = 2$, und sucht im gewissen Sinne optimale Koeffizienten $a_0, \ldots, a_m$. Man muss festlegen, was optimal bedeuten soll.



□

**Bemerkung 66.2 Die Methode der kleinsten Quadrate.** Jede der $n$ Messungen $(x_i, y_i)$ beschreibt eine lineare Gleichung für die Unbekannten $a_0, \ldots, a_m$

$$\begin{aligned} a_0 + a_1 x_1 + \ldots + a_m x_1^m &= y_1 \\ &\vdots \\ a_0 + a_1 x_n + \ldots + a_m x_n^m &= y_n. \end{aligned}$$

Im Allgemeinen hat man sehr viel mehr Gleichungen als Unbekannte, $n \gg m$, und das lineare Gleichungssystem

$$\underbrace{\begin{pmatrix} 1 & x_1 & \dots & x_1^m \\ \vdots & & & \vdots \\ 1 & x_n & \dots & x_n^m \end{pmatrix}}_{M \in \mathbb{R}^{n\times(m+1)}} \underbrace{\begin{pmatrix} a_0 \\ \vdots \\ a_m \end{pmatrix}}_{\mathbf{a}\in\mathbb{R}^{m+1}} = \underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}_{\mathbf{y}\in\mathbb{R}^n}$$

ist inkonsistent. Beispielsweise kann man nicht erwarten, dass 50 Messwerte exakt auf einer Geraden liegen, $n = 50$, $m = 1$.

Da $M\mathbf{a} = \mathbf{y}$ nicht lösbar ist, sucht man statt dessen eine „Lösung“ $\mathbf{a}^*$, die das Quadrat der Euklidischen Norm des Fehlers, den sogenannten quadratischen Fehler,

$$\|M\mathbf{a} - \mathbf{y}\|_2^2 = \left(\sum_{k=0}^m a_k x_1^k - y_1\right)^2 + \ldots + \left(\sum_{k=0}^m a_k x_n^k - y_n\right)^2$$

minimiert. Dieses Vorgehen nennt man Regression und das damit berechnete Polynom Ausgleichskurve oder Regressionskurve. □

**Bemerkung 66.3 Minimierung des quadratischen Fehlers.** Der quadratische Fehler kann als Funktion mt $m+1$ Unbekannten aufgefasst werden

$$\begin{aligned} f(a_0,\ldots,a_m) &= (M\mathbf{a}-\mathbf{y})^T(M\mathbf{a}-\mathbf{y}) = \mathbf{a}^T M^T M\mathbf{a} - \mathbf{a}^T M^T\mathbf{y} - \underbrace{\mathbf{y}^T M\mathbf{a}}_{\mathbf{a}^T M^T\mathbf{y}} + \mathbf{y}^T\mathbf{y} \\ &= \mathbf{a}^T M^T M\mathbf{a} - 2\mathbf{a}^T M^T\mathbf{y} + \mathbf{y}^T\mathbf{y}. \end{aligned}$$

Die notwendige Bedingung für ein Extremum ist, siehe Satz 51.3,

$$\mathbf{0} \stackrel{!}{=} \nabla f := \begin{pmatrix} \dfrac{\partial f}{\partial a_0} \\ \vdots \\ \dfrac{\partial f}{\partial a_m} \end{pmatrix} = 2M^T M\mathbf{a} - 2M^T\mathbf{y},$$

da mit $B = M^T M = B^T$ für den ersten Term gilt

$$\begin{aligned} \frac{\partial}{\partial a_k}\sum_{i,j=0}^m a_i a_j b_{ij} &= \frac{\partial}{\partial a_k}\sum_{j=0,j\neq k}^m a_k a_j b_{kj} + \frac{\partial}{\partial a_k}\sum_{i=0,i\neq k}^m a_i a_k b_{ik} + \frac{\partial}{\partial a_k} b_{kk} a_k^2 \\ &= \sum_{j=0,j\neq k}^m a_j b_{kj} + \sum_{i=0,i\neq k}^n a_i \underbrace{b_{ik}}_{=b_{ki}} + 2b_{kk}a_k \\ &= 2\sum_{j=0}^m b_{kj}a_j, \quad k = 0,\ldots,n. \end{aligned}$$

Der gesuchte Vektor $\mathbf{a}^*$ löst also die sogenannte Normalgleichung

$$M^T M\mathbf{a}^* = M^T\mathbf{y}.$$

Dies ist ein System aus $m+1$ Gleichungen mit $m+1$ Unbekannten $a_0,\ldots,a_m$. Ist $M^T M$ invertierbar, dann gilt

$$\mathbf{a}^* = (M^T M)^{-1} M^T\mathbf{y}.$$

Man nennt

$$M^+ := (M^T M)^{-1} M^T$$

die Pseudoinverse oder Moore[1]–Penrose[2]–Inverse der $n \times (m+1)$–Matrix $M$. □

[1] Eliakim Hastings Moore (1862 – 1932)
[2] Roger Penrose, geb. 1931

**Bemerkung 66.4**

- Die Hesse–Matrix $Hf(\mathbf{a}) = 2M^TM$ ist positiv semidefinit, da

$$\mathbf{a}^T 2M^T M\mathbf{a} = 2(M\mathbf{a})^T(M\mathbf{a}) = 2\|M\mathbf{a}\|_2^2 \geq 0 \quad \forall\, \mathbf{a} \in \mathbb{R}^{m+1}.$$

  Da $M^TM$ invertierbar sein soll, ist $Hf(\mathbf{a})$ sogar positiv definit. Nach Satz 51.6 folgt, dass $\mathbf{a}^*$ ein Minimum ist.
- Man kann zeigen, dass $M^TM$ invertierbar ist, falls $\text{rg}(M) = m+1$, das heißt $m+1$ der $n$ Gleichungen des Sytems $M\mathbf{a} = \mathbf{y}$ sind linear unabhängig oder $M$ besitzt vollen Zeilenrang.
- Das hergeleitete Verfahren kann allgemein verwendet werden, um ein überbestimmtes und im Allgemeinen inkonsistentes Gleichungssystem zu „lösen“.
- Pseudoinversen sind eine Verallgemeinerung von Inversen. Sei $M \in \mathbb{R}^{n\times n}$ invertierbar, dann gilt

$$M^+ := (M^TM)^{-1}M^T = M^{-1}M^{-T}M^T = M^{-1}.$$

- Man kann Pseudoinversen für alle Matrizen definieren, Vollrang der Matrix ist nicht erforderlich. Diese Definition fällt im Falle eines vollen Zeilenranges mit der obigen Definition zusammen.

□

**Satz 66.5 Pseudolösung überbestimmter Gleichungssysteme.** *Seien $n > m$, $A \in \mathbb{R}^{n\times m}$, $rg(A) = m$ und $\mathbf{b} \in \mathbb{R}^n$. Das überbestimmte lineare Gleichungssystem $A\mathbf{x} = \mathbf{b}$ besitzt eine eindeutige Pseudolösung $\mathbf{x}^*$, welche den quadratischen Fehler $\|A\mathbf{x} - \mathbf{b}\|_2^2$ minimiert*

$$\mathbf{x}^* = A^+\mathbf{b}$$

*mit der Pseudoinversen $A^+ = (A^TA)^{-1}A^T$.*

**Beweis:** Der Beweis folgt aus den Bemerkungen 66.3 und 66.4. ■

**Bemerkung 66.6**

- Pseudolösungen spielen auch in der Informatik eine wichtige Rolle. Überbestimmte Gleichungssyteme treten beispielsweise bei der Suche in Internetdatenbaken auf.
- In der numerischen Praxis löst man nicht die Normalgleichungen

$$A^TA\mathbf{x} = A^T\mathbf{b},$$

  da $A^TA$ wesentlich schlechtere Stabilitätseigenschaften als $A$ hat. Das bedeutet, bei direkten Verfahren werden die Rundungsfehler wesentlich größer. Stattdessen verwendet man Verfahren, die nur $A$ benutzen, deren Konstruktion aber recht kompliziert ist.

□

**Beispiel 66.7** Man soll mit der Methode der kleinsten Quadrate die Regressionsgerade durch die Punkte

$$(0,1),\ (1,3),\ (2,4),\ (3,4)$$

bestimmen. Die allgemeine Form der Regressionsgeraden ist $y = a_0 + a_1x$. Einsetzen der Punkte liefert vier Bestimmungsgleichungen

$$M\mathbf{a} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 4 \\ 4 \end{pmatrix} = \mathbf{y}.$$

Damit folgt

$$M^T M = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix} = \begin{pmatrix} 4 & 6 \\ 6 & 14 \end{pmatrix}.$$

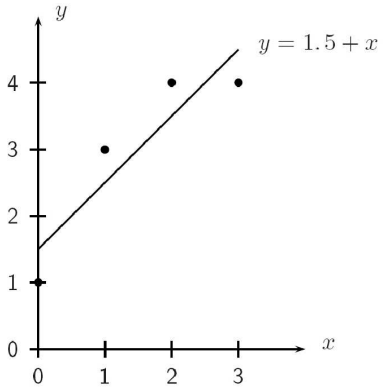
Diese Matrix ist invertierbar, da $\det(M^T M) = 4 \cdot 14 - 6 \cdot 6 = 20 \neq 0$. Man erhält

$$(M^T M)^{-1} = \frac{1}{10} \begin{pmatrix} 7 & -3 \\ -3 & 2 \end{pmatrix}$$

Damit lautet die Pseudolösung des obigen Systems

$$\begin{aligned} \mathbf{a}^* = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} &= (M^T M)^{-1} M^T \mathbf{y} \\ &= \frac{1}{10} \begin{pmatrix} 7 & -3 \\ -3 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \\ 4 \\ 4 \end{pmatrix} \\ &= \frac{1}{10} \begin{pmatrix} 7 & -3 \\ -3 & 2 \end{pmatrix} \begin{pmatrix} 12 \\ 23 \end{pmatrix} = \frac{1}{10} \begin{pmatrix} 15 \\ 10 \end{pmatrix} = \begin{pmatrix} 1.5 \\ 1 \end{pmatrix}. \end{aligned}$$

Die Ausgleichgerade lautet somit $y = 1.5 + x$.




□

# Kapitel 67

# Robuste Statistik

**Bemerkung 67.1 Motivation.** Will man aus realen Daten statistische Parameter schätzen, zum Beispiel nimmt man das arithmetisches Mittel als Schätzer für den Erwartungswert oder schätzt man die Parameter einer Regressionskurve, kann es sein, dass das Ergebnis auf Grund von Ausreißern stark verfälscht wird.

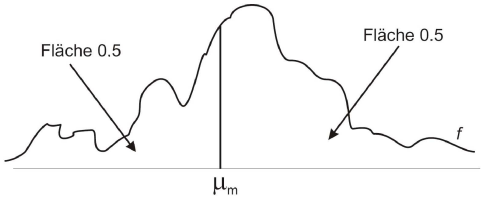
Zum Beispiel: 9 Studierende benötigen 10 Semester für ihr Studium, 1 benötigt 40 Semester. Das arithmetische Mittel ergibt eine mittlere Studiendauer von 13 Semestern. Das ist jedoch nicht repräsentativ für die Mehrzahl der Studierenden.

Die Frage ist, ob es statistische Verfahren gibt, die robuster gegenüber Ausreißern sind. □

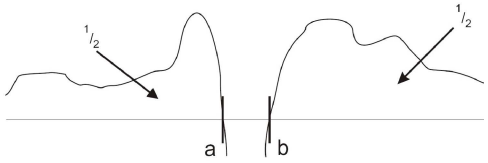
**Definition 67.2 Median.** Sei $X$ eine Zufallsvariable. Dann nennt man jede Zahl $\mu_m$ mit $P(X \geq \mu_m) \geq 0.5$ und $P(X \leq \mu_m) \geq 0.5$ einen Median von $X$. □

**Bemerkung 67.3**

- Veranschaulichung für kontinuierliche Zufallsvariable mit Dichte $f(x)$:




  Für die Verteilungsfunktion $F(x)$ gilt: $F(\mu_m) = 0.5$.
- Nicht immer gibt es einen eindeutigen Median. Gibt es ein Intervall $[a, b]$ mit $P(X \leq a) = 0.5$ und $P(X \geq b) = 0.5$, so ist jede Zahl aus $[a, b]$ ein Median.




- Im Allgemeinen stimmen Erwartungswert und Median nicht überein.

□

**Definition 67.4 Empirischer Median.** Hat man $2k+1$ der Größe nach geordnete Werte

$$x_1 \leq x_2 \leq \ldots \leq x_{2k+1},$$

dann nennt man $\hat{\mu}_m := x_{k+1}$ den empirischen Median dieser Daten. Es sind 50% der Daten größer oder gleich $\hat{\mu}_m$ und 50% sind kleiner oder gleich $\hat{\mu}_m$. Bei einer geraden Anzahl von Werten

$$x_1 \leq x_2 \leq \ldots \leq x_{2k}$$

definiert man

$$\hat{\mu}_m := \frac{1}{2}(x_{k+1} + x_k)$$

als empirischen Median. □

**Beispiel 67.5**

- Der empirische Median der Studiendauer in Bemerkung 67.1 beträgt 10 Semester. Der Ausreißer mit 40 Semestern hat somit keinen Einfluss auf den empirischen Median.
- In der Bildverarbeitung ersetzt der Medianfilter einen Grauwert durch seinen empirischen Median innerhalb eines $(2k+1) \times (2k+1)$–Fensters:

| | | |
|---|---|---|
| 32 | 17 | 24 |
| 35 | 251 | 21 |
| 12 | 24 | 25 |

Ordnen der Grauwerte:

$$12 \leq 17 \leq 21 \leq 24 \leq \underset{\substack{\uparrow \\ \text{Median}}}{24} \leq 25 \leq 32 \leq 35 \leq 251$$

Der Grauwert 251 (Ausreißer) wird durch den empirischen Median 24 ersetzt. Medianfilter sind robust gegenüber Impulsrauschen (Ausreißer nach oben oder unten) und erhalten Kanten.

□

**Definition 67.6 M–Schätzer.** Seien $x_1, \ldots, x_n$ Werte und $\Psi : [0, \infty) \to \mathbb{R}_0^+$ eine monoton wachsende Straffunktion (englisch penaliser). Dann nennt man dasjenige $\mu$, welches

$$\mu = \operatorname{argmin} \sum_{i=1}^{n} \Psi(|x - x_i|)$$

minimiert, den M–Schätzer von $x_1, \ldots, x_n$. □

**Beispiel 67.7** Beliebt ist die Familie $\Psi(s) = s^p$ mit $p \geq 0$. Man kann zeigen:

- $p = 2$ liefert das arithmetische Mittel $\overline{x}$. Der quadratische Abstand wird minimiert

$$\overline{x} = \operatorname{argmin} \sum_{i=1}^{n} (x_i - x)^2.$$

- $p = 1$ liefert den empirischen Median $\hat{\mu}$. Die Abstandssumme wird minimiert

$$\hat{\mu} = \operatorname{argmin} \sum_{i=1}^{n} |x_i - x|.$$

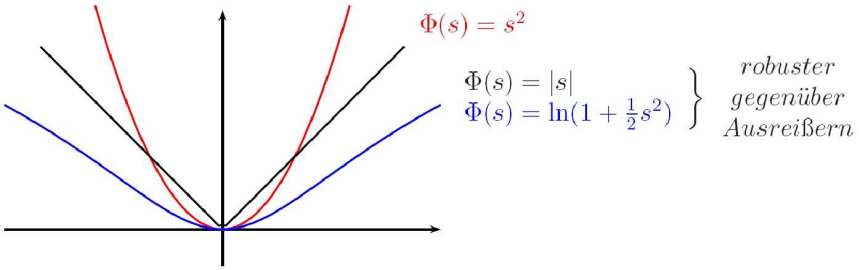
- $p \to 0$ liefert als Minimierer die sogenannten Modalwerte, das sind die Maxima des Histogramms, das heißt, die Werte mit der höchsten Wahrscheinlichkeit.
- $p \to \infty$ ergibt den sogenannte Midrange

$$\frac{\max_{i=1,\ldots,n}\{x_i\} + \min_{i=1,\ldots,n}\{x_i\}}{2}.$$

Kleinere Werte für $p$ liefern robustere M–Schätzer, da sie Ausreißer $x_i$, für die $\Psi(|x_i - x|) = |x_i - x|^p$ groß wird, weniger stark berücksichtigen. Der Midrange hängt dagegen nur von den Ausreißern ab. □

**Beispiel 67.8** Eine Straffunktion, die robuster als die übliche quadratische Straffunktion $\Psi(s) = s^2$ ist, ist zum Beispiel die Lorentz[1]–Straffunktion

$$\Psi(s) = \ln\left(1 + \frac{1}{2}s^2\right).$$




□

[1] Lorentz

# Kapitel 68

# Markowketten

**Bemerkung 68.1 Motivation.** Der Zustand eines Systems zur Zeit $n \in \mathbb{N}$ werde durch eine Zufallsvariable $X_n$ beschrieben und soll nur von $X_{n-1}$ abhängen, nicht jedoch von früheren Zuständen $X_{n-2}, X_{n-3}, \ldots$. Man möchten das zeitliche Verhalten dieses Systems studieren, insbesondere das Langzeitverhalten für $n \to \infty$.

Prozesse dieser Art sind in der Informatik zum Beispiel bei der Untersuchung der Auslastung von Servern wichtig (Warteschlangenmodelle). □

**Definition 68.2 Markowkette, Stochastischer Prozess.** Ein stochastischer Prozess ist eine Familie $(X_t)$ von Zufallsvariablen mit $t \in \mathbb{R}$ oder $t \in \mathbb{N}$. Man denkt dabei an $t$ als Zeitparameter, der kontinuierlich oder diskret ist. Ein diskreter stochastischer Prozess $(X_n)$, $n \in \mathbb{N}$, heißt Markowkette, wenn die Verteilung von $X_n$ bei gegebenen $X_{n-1}$ nicht von der früheren Verteilungen $X_k$, $k < n-1$, abhängt

$$P(X_n = i_n \mid X_{n-1} = i_{n-1},\ X_{n-2} = i_{n-2},\ \ldots) = P(X_n = i_n \mid X_{n-1} = i_{n-1}).$$

□

**Bemerkung 68.3** Wichtig sind insbesondere Markowketten, die nur endlich viele Zustände annehmen

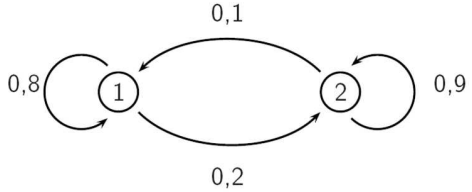
$$P(X_n \in \{1, \ldots, k\}) = 1.$$

□

**Beispiel 68.4** Jedes Jahr ziehen 10% der Bevölkerung außerhalb Kaliforniens nach Kalifornien, und 20% der Bevölkerung Kaliforniens zieht aus. Eine Person befindet sich im Jahr $n-1$ in einem von zwei Zuständen

$$X_{n-1} := \begin{cases} 1 & \text{Person wohnt in Kalifornien,} \\ 2 & \text{Person wohnt nicht in Kalifornien.} \end{cases}$$

Der Zustand im Jahr $n$ läßt sich dann durch ein graphisches Modell mit Übergangswahrscheinlichkeiten beschreiben



Sei $p_{ij}^n$ die Wahrscheinlichkeit, dass ein Zustand $j$ zur Zeit $n-1$ in den Zustand $i$ übergeht, zum Beispiel $p_{12}^n = 0.1$,

$$p_{ij}^n = P\left(X_n = i \mid X_{n-1} = j\right).$$

Man definiert die Matrix der Übergangswahrscheinlichkeiten oder Übergangsmatrix durch

$$M_n := \left(p_{ij}^n\right) \in \mathbb{R}^{k\times k},$$

wobei man $k$ Zustände hat. Das ist im Beispiel

$$M_n = \begin{pmatrix} 0.8 & 0.1 \\ 0.2 & 0.9 \end{pmatrix}.$$

Die Verteilung von $X_n$ auf die Zustände $i = 1, \ldots, k$, werde durch einen Vektor $\mathbf{u}_n \in \mathbb{R}^k$ beschrieben

$$\mathbf{u}_n = \begin{pmatrix} u_{n1} \\ \vdots \\ u_{nk} \end{pmatrix} \in \mathbb{R}^k.$$

Dann berechnet sich $\mathbf{u}_n$ aus $\mathbf{u}_{n-1}$ durch

$$\mathbf{u}_n = M_n \mathbf{u}_{n-1}.$$

Sind im Beispiel im Jahr $(n-1)$ 60% der Bevölkerung außerhalb Kaliforniens, so gilt

$$\mathbf{u}_{n-1} = \begin{pmatrix} 0.4 \\ 0.6 \end{pmatrix}.$$

Es folgt für das Jahr $n$

$$\mathbf{u}_n = \begin{pmatrix} 0.8 & 0.1 \\ 0.2 & 0.9 \end{pmatrix} \begin{pmatrix} 0.4 \\ 0.6 \end{pmatrix} = \begin{pmatrix} 0.38 \\ 0.62 \end{pmatrix}.$$

Im Jahr $n$ sind somit 62% der Bevölkerung außerhalb Kaliforniens. □

**Definition 68.5 Kette mit stationären Übergangswahrscheinlichkeiten.** Eine Markowkette $(X_n)$ heißt homogen oder Kette mit stationären Übergangswahrscheinlichkeiten, wenn die Übergangsmatrix $M_n$ unabhängig von der Zeit $n$ ist

$$M_n = M \quad \forall n \in \mathbb{N}.$$

□

**Bemerkung 68.6**

- Beispiel 68.4 beschreibt eine homogene Markowkette.
- Man nennt eine Marix $A = (a_{ij}) \in \mathbb{R}^{k\times k}$ eine stochastische Matrix, wenn alle Einträge nichtnegativ sind und die Spaltensummen Eins sind

$$\begin{aligned} a_{ij} &\geq 0 \quad \forall i, j \in \{1, \ldots, n\}, \\ \sum_{i=1}^{k} a_{ij} &= 1 \quad \forall j \in \{1, \ldots, n\}. \end{aligned}$$

Übergangsmatrizen sind Beispiele für stochastische Matrizen.

- In der Stochastikliteratur werden oft Zeilenvektoren $\mathbf{u}_n$ betrachtet, und man definiert $p_{ij}^n$ als die Übergangswahrscheinlichkeit von Zustand $i$ nach Zustand $j$. Dann ist

$$\mathbf{u}_n = \mathbf{u}_{n-1} M_n$$

und stochastische Matrizen haben Zeilensumme Eins.

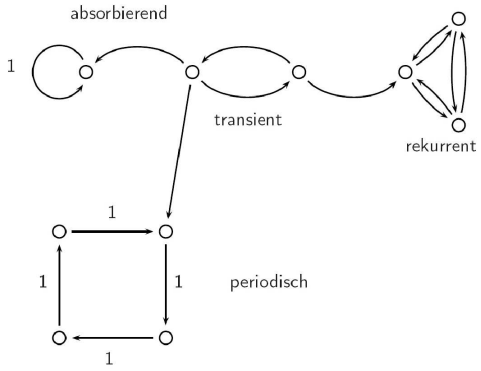
□

**Definition 68.7 Zustandsbeschreibung endlicher homogener Markowketten (transient, rekurrent, periodisch, absorbierend).** Sei $(X_n)$ eine endliche homogene Markowkette, die im Zustand $i$ ist. Der Zustand $i$ heißt:

1. rekurrent (wiederkehrend), wenn $p_i = 1$, wobei $p_i$ die Wahrscheinlichkeit ist, dass der Zustand $i$ irgendwann wieder erreicht wird. Für $n \to \infty$ wird ein rekurrenter Zustand mit Wahrscheinlichkeit 1 unendlich mal oft angenommen.
2. transient (vorübergehend), wenn $p_i < 1$ ist,
3. periodisch mit Periode $l$, wenn

$$P(X_{n+l} = i \mid X_n = i) = 1.$$

Eine Menge $I$ heißt absorbierend, wenn $P(X_{n+1} \in I \mid X_n \in I) = 1$.




Die Pfeile stellen Übergänge mit positiver Übergangswahrscheinlichkeit dar. □

Um das Zeitverhalten von Markowketten zu verstehen, muss man stochastische Matrizen näher untersuchen.

**Satz 68.8 Eigenwerte stochastischer Matrizen.** *Sei $M = (p_{ij}) \in \mathbb{R}^{k \times k}$ eine stochastische Matrix. Dann gelten:*

*i) $\lambda = 1$ ist Eigenwert von $M^T$.*

*ii) $|\lambda| \leq 1$ für alle Eigenwerte $\lambda$ von $M$ und $M^T$.*

*iii) $\lambda = 1$ ist einziger Eigenwert von $M^T$ mit $|\lambda| = 1$, falls $\min_{j=1,\ldots,k} p_{jj} > 0$.*

**Beweis:** *i).* Ist $M$ stochastisch, so hat $A = M^T$ Zeilensumme Eins. Damit gilt

$$A \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{k} a_{1j} \\ \vdots \\ \sum_{j=1}^{k} a_{kj} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Somit ist

$$\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

Eigenvektor von $A$ zum Eigenwert $\lambda = 1$.

*ii).* Betrachte die Spaltensummennorm

$$\|M\|_1 = \max_{j=1,\dots,k} \left( \sum_{i=1}^{k} |p_{ik}| \right) = 1.$$

Dann gilt nach Satz 45.8 für jeden Eigenwert $\lambda$ von $M$

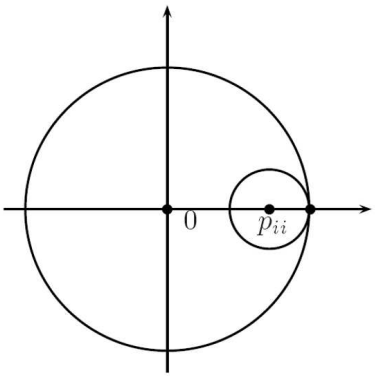
$$|\lambda| \leq \|M\|_1 = 1.$$

Für $M^T$ betrachtet man die Zeilensummennorm.

*iii).* Nach dem Satz von Gerschgorin, Satz 45.10, gilt für jeden Eigenwert $\lambda$ von $M^T = (a_{ij})$

$$|\lambda - p_{ii}| \leq \sum_{\substack{j=1 \\ j \neq i}}^{k} |a_{ij}| = 1 - p_{ii}.$$

Also liegt $\lambda$ in dem Kreis mit Mittelpunkt $(p_{ii}, 0)$ und Radius $1 - p_{ii}$. Er berührt den Einheitskreis von innen in $(1, 0)$




Aus $|\lambda| = 1$ folgt somit $\lambda = 1$. ■

**Bemerkung 68.9**

- Die Aussage von Satz 68.8, iii) gilt auch für $M$ statt $M^T$.
- *Bedeutung des Eigenwerts* $\lambda = 1$. Man interessiert sich für das Verhalten einer endlichen homogenen Markowkette $(X_n)$ für $n \to \infty$. Für einen Anfangszustand $\mathbf{u}_0$ und eine Übergangsmatrix $M = (p_{ij})$ gilt

$$\begin{aligned} \mathbf{u}_1 &= M\mathbf{u}_0, \\ \mathbf{u}_2 &= M\mathbf{u}_1 = M^2\mathbf{u}_0, \\ &\vdots \\ \mathbf{u}_n &= M^n\mathbf{u}_0. \end{aligned}$$

  Ist $\mathbf{u}_0$ Eigenvektor von $M$ zum Eigenwert $\lambda = 1$, dann folgt

$$\mathbf{u}_n = M^n\mathbf{u}_0 = \lambda^n\mathbf{u}_0 = \mathbf{u}_0 \quad \forall n \in \mathbb{N},$$

  das heißt, der Zustand $\mathbf{u}_0$ ist stabil.

□

Darüberhinaus kann man Folgendes zeigen.

**Satz 68.10 Potenzen stochastischer Matrizen.** *Sei $M$ eine stochastische Matrix. Der Grenzwert $\lim_{n\to\infty} M^n$ existiert genau dann, wenn 1 der einzige Eigenwert von $M$ mit Betrag 1 ist.*

**Beweis:** Es wird nur eine Richtung bewiesen: Sei $\lambda$ ein Eigenwert von $M$ mit $|\lambda| = 1$. Falls $\lim_{n\to\infty} M^n$ existiert, so ist 1 einziger Eigenwert von $M$ vom Betrag 1.

Sei $\mathbf{v} \neq \mathbf{0}$ ein Eigenvektor von $M$ zum Eigenwert $\lambda$ mit $|\lambda| = 1$

$$M\mathbf{v} = \lambda\mathbf{v} \quad \Longrightarrow \quad M^n\mathbf{v} = \lambda^n\mathbf{v} \quad \forall\, n \in \mathbb{N}.$$

Existiert $\lim_{n\to\infty} M^n$, dann folgt

$$\left(\lim_{n\to\infty} M^n\right)\mathbf{v} = \lim_{n\to\infty}(M^n\mathbf{v}) = \lim_{n\to\infty}(\lambda^n\mathbf{v}) = \left(\lim_{n\to\infty}\lambda^n\right)\mathbf{v}.$$

Also existiert $\mu = \lim_{n\to\infty}\lambda^n$. Dann ist auch

$$\mu = \lim_{n\to\infty}\lambda^{n+1} = \lambda\lim_{n\to\infty}\lambda^n = \lambda\mu.$$

Wegen $|\lambda| = 1$ ist auch $|\lambda^n| = 1$ und somit $|\mu| = 1 \neq 0$. Aus $\mu = \lambda\mu$ folgt dann $\lambda = 1$ nach Division durch $\mu$. ■

**Beispiel 68.11 Beispiel für die andere Beweisrichtung.** Betrachte eine stochastische Matrix, die einen Eigenwert $\lambda$ mit $|\lambda| = 1$, aber $\lambda \neq 1$ besitzt

$$M = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 1 & & & 0 \end{pmatrix} \in \mathbb{R}^{k\times k}.$$

Für den $i$–ten Einheitsvektor $\mathbf{e}_i$ gilt

$$M\mathbf{e}_i = \begin{cases} \mathbf{e}_{i-1} & \text{falls } 2 \leq i \leq k \\ \mathbf{e}_k & \text{falls } i = 1. \end{cases}$$

Das bedeutet, jede Multiplikation mit $M$ verschiebt in jeder Spalte die Einsen um eine Stelle nach oben beziehungsweise eine Eins in der ersten Zeile wird in die letzte Zeile verschoben. Damit beschreibt die Matrix $M$ eine zyklische Vertauschung der Einsen, die Potenz $M^n$ konvergiert nicht.

Sei

$$\alpha := e^{2\pi i/k} = \cos\frac{2\pi}{k} + i\sin\frac{2\pi}{k} \quad \Longrightarrow \quad \alpha^k = e^{2\pi i} = 1.$$

Setzt man

$$\mathbf{v}_j := \begin{pmatrix} 1 \\ \alpha^j \\ \alpha^{2j} \\ \vdots \\ \alpha^{(k-1)j} \end{pmatrix}, \quad 0 \leq j \leq k-1,$$

so folgt wegen $\alpha^k = 1$

$$M\mathbf{v}_j = \begin{pmatrix} \alpha^j \\ \alpha^{2j} \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} \alpha^j \\ \alpha^{2j} \\ \vdots \\ \alpha^{kj} \end{pmatrix} = \alpha^j \begin{pmatrix} 1 \\ \alpha^j \\ \alpha^{2j} \\ \vdots \\ \alpha^{(k-1)j} \end{pmatrix} = \alpha^j\mathbf{v}_j.$$

Also sind $1, \alpha^1, \alpha^2, \ldots, \alpha^{k-1} \in \mathbb{C}$ sämtliche Eigenwerte von $M$. Alle haben Betrag 1. □

**Definition 68.12 Markowketten im Gleichgewicht.** Eine endliche homogene Markowkette $(X_n)$ ist im Gleichgewicht, falls zu ihrer Übergangsmatrix $M = (p_{ij})$ ein Zustand $\mathbf{u}$ existiert mit

$$M\mathbf{u} = \mathbf{u},$$

das heißt $\mathbf{u}$ ist Eigenvektor von $M$ zum Eigenwert 1, und es gilt

$$\sum_{i=1}^{k} u_i = 1, \text{ mit } u_i \geq 0 \text{ für } i = 1, \ldots, k.$$

□

**Beispiel 68.13** Im Beispiel 68.4 war die Übergangsmatrix gegeben durch

$$M = \begin{pmatrix} 0.8 & 0.1 \\ 0.2 & 0.9 \end{pmatrix}$$

Für die Eigenwerte dieser Matrix erhält man

$$\begin{aligned} 0 &\stackrel{!}{=} \begin{vmatrix} 0.8-\lambda & 0.1 \\ 0.2 & 0.9-\lambda \end{vmatrix} = (0.8-\lambda)(0.9-\lambda) - 0.02 \\ &= 0.72 - 0.8\lambda - 0.9\lambda + \lambda^2 - 0.02 \\ &= \lambda^2 - 1.7\lambda + 0.7 \quad \Longrightarrow \\ \lambda_{1/2} &= \frac{1.7 \pm \sqrt{1.7^2 - 4 \cdot 0.7}}{2} = \frac{1.7 \pm \sqrt{2,89 - 2,8}}{2} \\ &= \frac{1.7 \pm 0.3}{2} \quad \Longrightarrow \\ \lambda_1 &= 1, \quad \lambda_2 = 0.7. \end{aligned}$$

Für den Eigenvektor zum Eigenwert $\lambda_1 = 1$ ergibt sich

$$\begin{aligned} \begin{pmatrix} -0.2 & 0.1 \\ 0.2 & -0.1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \Longrightarrow \quad -0.2x + 0.1y = 0 \\ & \qquad\qquad\qquad\qquad\quad 2x = y \\ \mathbf{v} &= \begin{pmatrix} \alpha \\ 2\alpha \end{pmatrix}, \quad \alpha \neq 0. \end{aligned}$$

Da $\mathbf{v}$ eine Verteilung auf die Zustände beschreiben soll, muss gelten

$$v_i \geq 0, \quad \sum_{i=1}^{n} v_i = 1 \quad \Longrightarrow \quad \mathbf{v} = \begin{pmatrix} 1/3 \\ 2/3 \end{pmatrix}.$$

Dies beschreibt den Gleichgewichtszustand: 1/3 der Personen wohnt in Kalifornien, 2/3 außerhalb. □

Als nächstes wird die Frage untersucht, unter welchen Bedingungen es möglich ist, bei einer Markowkette von einem Zustand in einen beliebigen anderen zu gelangen.

**Definition 68.14 Transitive oder irreduzible Matrix.** Eine stochastische Matrix $M = (p_{ij}) \in \mathbb{R}^{k \times k}$ heißt transitiv (irrduzibel), wenn man von jedem Zustand $n$ zu jedem anderen Zustand $m$ mit positiver Wahrscheinlichkeit in endlich vielen Schritten gelangen kann. Das heißt, es gibt ein $r \in \mathbb{N}$, so dass für $M^r = B = (b_{ij})$ gilt

$$b_{mn} > 0.$$

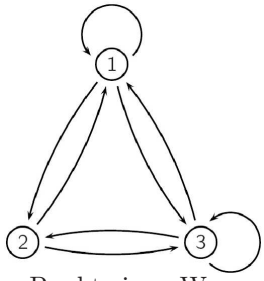
□

**Bemerkung 68.15 Praktisches Kriterium für Irreduzibilität.** Man zeichnet zur Übergangsmatrix $M = (p_{ij})$ einen gewichteten Graphen. Ist $p_{ij} > 0$, zeichnet man einen Pfeil von $j$ nach $i$. Falls man von jedem Zustand längs solcher Pfeile zu jedem anderen Zustand gelangt, ist $M$ transitiv, andernfalls nicht. □

**Beispiel 68.16**

- Sei

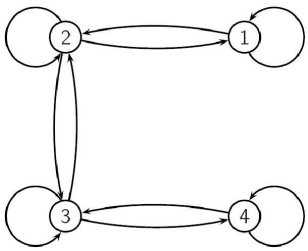
$$M = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & 0 & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$




  Irreduzibel bedeutet, es gibt von jedem Punkt einen Weg zu jedem anderen Punkt. Man beachte, dass der Weg mehrere Pfeile umfassen darf. Daher führt auch ein Weg von ② nach ②.

- Betrachte

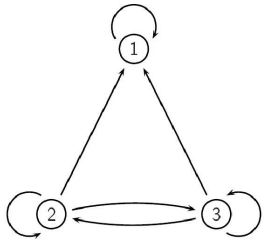
$$M = \begin{pmatrix} \frac{3}{4} & \frac{1}{4} & 0 & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 \\ 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & 0 & \frac{1}{4} & \frac{3}{4} \end{pmatrix}.$$




  Diese Matrix ist irreduzibel.

- Betrachte

$$M = \begin{pmatrix} 1 & \frac{2}{5} & \frac{1}{4} \\ 0 & \frac{1}{5} & \frac{1}{4} \\ 0 & \frac{2}{5} & \frac{1}{2} \end{pmatrix}.$$




  Diese Matrix ist reduzibel, da zum Beispiel kein Weg von ① nach ② führt. □

# Kapitel 69

# Pseudozufallszahlen und Monte–Carlo–Simulation

**Bemerkung 69.1 Motivation.** Es gibt Anwendungsgebiete (zum Beispiel Ray Tracing in der Computergrafik, Strömungssimulation im Bereich Computational Fluid Dynamics, Berechnung hochdimensionaler Integrale), bei denen stochastische Simulationen einfache oder effiziente Alternativen zu deterministischen Algorithmen sind. Solche Simulationen nennt man auch Monte–Carlo–Verfahren. Sie benötigen die Erzeugung von Zufallszahlen auf dem Rechner. Da funktionierende Computer jedoch deterministisch arbeiten, verwendet man statt dessen Algorithmen, die Zahlen liefern, welche echten Zufallszahlen ähneln. Solche Zahlen heißen Pseudozufallszahlen. □

**Bemerkung 69.2 Erzeugung von gleichverteilten Pseudozufallszahlen.**

- *Ziel:* Erzeugung einer Sequenz $Z_n$ von gleichverteilten Pseudozufallszahlen aus $[0,1]$. Eine beliebte Vorgehensweise sind lineare Kongruenzmethoden, die in vielen Compilern verwendet werden.
- *gegeben:*

| | |
|---|---|
| $m \in \mathbb{N}$ | Modus, |
| $a \in \{1, \ldots, m-1\}$ | Multiplikator, |
| $b \in \{0, \ldots, m-1\}$ | Inkrement, |
| $x_0 \in \{0, \ldots, m-1\}$ | Startwert. |

- *Verfahren:*

$$\begin{aligned} x_n &:= (ax_{n-1} + b) \text{ (modulo m)}, \\ Z_n &:= \frac{x_n}{m}. \end{aligned}$$

  Dann approximiert die Sequenz $\{Z_n\}$ eine Folge von stochastisch unabhängigen Zufallsvariablen, die auf $[0,1]$ gleichverteilt sind. Die Approximationsgüte hängt von der Parameterwahl ab.
- *Eigenschaften:* Nach spätestens $m$ Schritten wiederholt sich die Folge, da $x_n$ nur die Werte $0, \ldots, m-1$ annimmt. Die Zahl 1 tritt nie auf, da $x_n \neq m$.
- *Parameterwahl:* Häufig verwendet, aber schlecht:

$$m = 2^{16} = 65536, \quad a = 25173, \quad b = 13849.$$

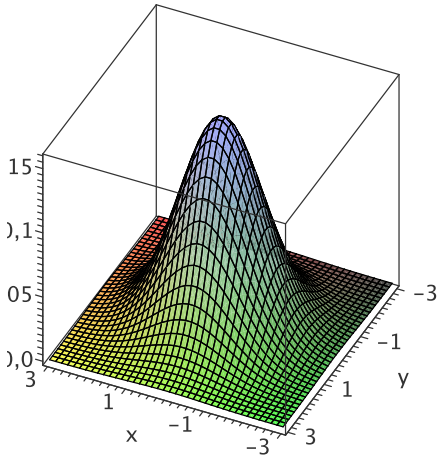
  Eine bessere Wahl ist (der sogenannte Minimalstandard):

$$m = 2^{31} - 1 = 2147483647, \quad a = 7^5 = 16807, \quad b = 0.$$

□

**Bemerkung 69.3 Erzeugung von $N(0,1)$–verteilten Pseudozufallszahlen.** Die Standardnormalverteilung im $\mathbb{R}^2$ hat die Dichte

$$\varphi(x,y) = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}.$$




Sie ist das Produkt zweier unabhängiger eindimensionaler $N(0,1)$–verteilten Zufallsgrößen $X$ und $Y$, siehe Beispiel 63.6. Für eine Kreisscheibe $B(\mathbf{0},t)$ um $\mathbf{0}$ mit Radius $t$ und diese Zufallsvariablen gilt

$$P(X^2+Y^2 \leq t^2) = \int_{B(\mathbf{0},t)} \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}\, dxdy.$$

Dieses Integral transformiert man auf Polarkoordinaten und wendet dann die Transformationsregel, Bemerkung 54.7, an. Man hat

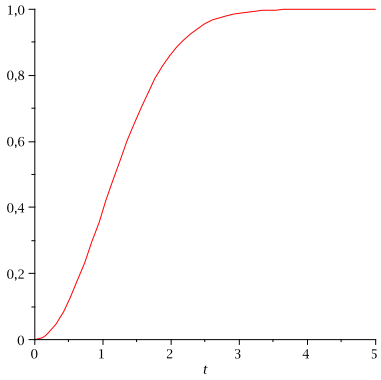
$$\begin{aligned}
x &= r\cos\varphi =: f_1(r,\varphi),\\
y &= r\sin\varphi =: f_2(r,\varphi),\\
Jf(r,\varphi) &= \begin{pmatrix} \frac{\partial f_1}{\partial r} & \frac{\partial f_1}{\partial \varphi} \\ \frac{\partial f_2}{\partial r} & \frac{\partial f_2}{\partial \varphi} \end{pmatrix} = \begin{pmatrix} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{pmatrix},\\
\det(Jf(r,\varphi)) &= r\cos^2\varphi + r\sin^2\varphi = r.
\end{aligned}$$

Mit der Transformationsregel erhält man nun

$$\begin{aligned}
P(X^2+Y^2 \leq t^2) &= \int_{\varphi=0}^{2\pi}\int_{r=0}^{t} \frac{1}{2\pi} e^{-\frac{r^2}{2}} r\; drd\varphi\\
&= 2\pi \int_{r=0}^{t} \frac{1}{2\pi} e^{-\frac{r^2}{2}} r\; dr = -e^{-\frac{r^2}{2}}\Big|_0^t = 1 - e^{-\frac{t^2}{2}}.
\end{aligned}$$

Der Radius $R$ des Zufallszahlenvektors $(X,Y)$ erfüllt also

$$P(R \leq t) = 1 - e^{-\frac{t^2}{2}} =: g(t).$$




Der Wertebereich von $g$ ist $[0,1)$. Die Umkehrfunktion zu $g(t)$ ist

$$t = g^{-1}(z) = \sqrt{-2\ln(1-z)}.$$

Hat man eine in $[0,1]$ gleichverteilte Zufallsvariable $Z_1$, ergibt sich damit als Zufallsvariable $R$ für den Radius der 2D–Standardnormalverteilung:

$$R = \sqrt{-2\ln(1-Z_1)}.$$

Der Winkel der 2D–Standardnormalverteilung ist gleichverteilt. Aus einer zweiten, auf $[0,1]$ gleichverteilten Zufallsvariablen $Z_2$, erhält man als Zufallsvariable für einen auf $[0,2\pi]$ gleichverteilten Winkel $T$

$$T = 2\pi Z_2.$$

Dies motiviert folgenden Algorithmus (Box[1]–Muller[2]–Verfahren, 1958) zur Erzeugung zweier $N(0,1)$–verteilter Pseudozufallszahlen $X, Y$ aus zwei auf $[0,1]$ gleichverteilten Pseudozufallszahlen $Z_1, Z_2$:

$$\begin{aligned} R &:= \sqrt{-2\ln(1-Z_1)}, \\ T &:= 2\pi Z_2, \\ X &:= R\cos T, \\ Y &:= R\sin T. \end{aligned}$$

□

**Bemerkung 69.4** Benötigt man $N(\mu,\sigma^2)$-verteilte Zufallszahlen $\tilde{X}, \tilde{Y}$, setzt man, vergleiche Definition 59.17:

$$\begin{aligned} \tilde{X} &= \mu + \sigma X \\ \tilde{Y} &= \mu + \sigma Y \end{aligned}$$

□

(Pseudo–)Zufallszahlen benötigt man zum Beispiel auch bei probabilistischen Algorithmen.

**Beispiel 69.5 Quicksort.** Sortieren einer Liste mit $n$ Elementen mit Quicksort:

- Suche ein zufälliges Element $z_1$ der Liste, mit gleichverteilten Pseudozufallszahlen aus Bemerkung 69.2.
- Sortiere Teilliste mit Elementen kleiner oder gleich $z_1$ mit Quicksort.

[1] George Edward Pelham Box
[2] Mervin Edgar Muller

- Sortiere Teilliste mit Elementen größer als $z_1$ mit Quicksort.

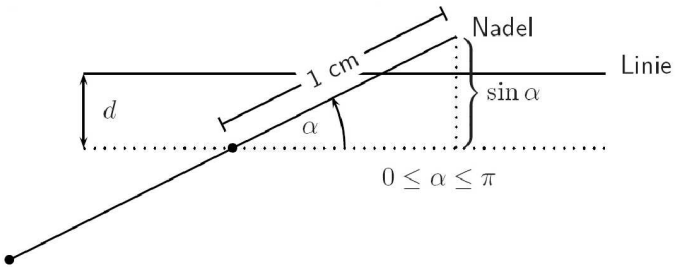
Das Verfahren wird auf diese Art und Weise fortgesetzt. Quicksort hat eine mittlere Laufzeit von $\mathcal{O}(n \log n)$, ist also nicht schlecht. □

**Beispiel 69.6 Buffon[3]'sches Nadelexperiment.** Dabei handelt es sich um einen probabilistischen Algorithmus zur Approximation von $\pi$.

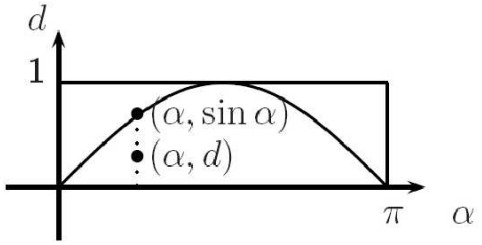
- Zeichne auf einem Blatt Papier parallele Linien im Abstand einer Stecknadellänge.
- Lasse die Stecknadel auf das Papier fallen und überprüfe, ob sie eine der Linien trifft.
- Zähle die Zahl $N$ der Versuche und die Zahl $T$ der Treffer. Dann gilt

$$\frac{N}{T} \to \frac{\pi}{2} \quad \text{für } N \to \infty.$$

Das ergibt sich aus den folgenden Überlegungen. Seien ohne Beschränkung der Allgemeinheit Nadellänge und Linienabstand gleich 2 [cm]. Die Nadel kann nur die nächstgelegene Linie schneiden und nur dann, wenn ihr Abstand $d$ zwischen Nadelmitte und Linie $d < 1$ erfüllt:




Es folgt, dass die Nadel die Linie schneidet, genau dann wenn $d < \sin\alpha$ gilt. Zu jedem Wurf gehört ein Wertepaar $(\alpha, d) \in [0, \pi] \times [0, 1]$. Die Bedingung $d < \sin\alpha$ ist genau dann erfüllt, wenn $(\alpha, d)$ unter Graphen von $\sin x$ liegt.




Im Zufallsexperiment sind die Punkte $(\alpha, d)$ gleichverteilt auf $[0, \pi] \times [0, 1]$, das bedeutet es gilt

$$\frac{T}{N} \to \frac{\text{Fläche unter Kurve}}{\text{Fläche des Rechtecks}} = \frac{\int_0^\pi \sin x \; dx}{\pi \cdot 1} = \frac{-\cos x|_0^\pi}{\pi} = \frac{2}{\pi}.$$

Damit stellt sich dieses Experiment als ein stochastisches Integrationsverfahren für

$$\int_0^\pi \sin x \; dx$$

heraus.

Im 1D–Fall sind solche Verfahren ineffizient, aber bei hochdimensionalen Integrationsproblemen haben sie eine bessere Komplexität als deterministische numerische Verfahren. □

[3]Georges–Louis Leclerc, Comte de Buffon (1707 – 1788)