

# VORLESUNGSSKRIPT EINFÜHRUNG IN DIE NUMERIK

SOMMERSEMESTER 2022

Roland Herzog\*

2022-05-12

\*Interdisciplinary Center for Scientific Computing, Heidelberg University, 69120 Heidelberg, Germany  
([roland.herzog@iwr.uni-heidelberg.de](mailto:roland.herzog@iwr.uni-heidelberg.de), <https://scoop.iwr.uni-heidelberg.de/team/roland-herzog>).

Material für 13 Wochen.

# Inhaltsverzeichnis

o	Einführung	5
§ 1	Was ist Numerische Mathematik?	5
§ 2	Wiederholung und Notation	9
§ 2.1	Vektor- und Matrix-Normen	10
§ 2.2	Differenzierbarkeit und der Satz von Taylor	15
§ 2.3	Landau-Notation	19
1	Sensitivitätsanalyse und Kondition von Aufgaben	22
§ 3	Sensitivitätsanalyse und Kondition	22
§ 3.1	Absolute Betrachtungsweise	23
§ 3.2	Relative Betrachtungsweise	27
§ 3.3	Kondition linearer Gleichungssysteme	32
§ 3.4	Kondition zusammengesetzter Funktionen	34
§ 4	Singulärwertzerlegung von Matrizen	36
§ 4.1	Spektralzerlegung	36
§ 4.2	Singulärwertzerlegung	37
2	Zahldarstellung und Rechnerarithmetik	45
§ 5	Zahldarstellung	45
§ 6	Rechnerarithmetik	50



# Kapitel 0 Einführung

## § 1 WAS IST NUMERISCHE MATHEMATIK?

Die **numerische Mathematik** (kurz: **Numerik**, englisch: *numerical analysis*) beschäftigt sich mit der Konstruktion und Analyse von Rechenverfahren (Algorithmen) zur Lösung von Aufgaben der „kontinuierlichen Mathematik“ (im Gegensatz zur diskreten Mathematik) auf Computern. Das englische Originalzitat „Numerical analysis is the study of algorithms for the problems of continuous mathematics.“ hinter dieser Definition stammt von [Trefethen, 1992](#). Ein typisches Beispiel einer Grundaufgabe der Numerik ist etwa das Lösen linearer Gleichungssysteme  $Ax = b$  mit Hilfe des Gauß-Algorithmus.

Die Numerik ist damit ein Teilgebiet des **wissenschaftlichen Rechnens** (englisch: *scientific computing, computational science*), bei dem es darum geht, eine konkrete Fragestellung (etwa aus der Physik) in ein Modell und anschließend in ein numerisches Verfahren zu übersetzen, dieses auf geeigneter Hardware effizient zu implementieren und die Simulationsergebnisse mit realen Beobachtungen oder Experimenten abzugleichen, um damit die ursprüngliche Fragestellung beantworten zu können.

Nicholas Higham hat 2016 eine Liste der „Top 10“-Algorithmen in der angewandten Mathematik veröffentlicht<sup>1</sup>, indem er die Verweise im Index des über 1000-seitigen Buches *The Princeton Companion to Applied Mathematics* ([Higham u. a., 2016](#)) gezählt hat. Herausgekommen ist dabei folgende Liste:

- (1) Newton and quasi-Newton methods
- (2) Matrix factorizations (LU, Cholesky, QR)
- (3) Singular value decomposition, QR and QZ algorithms
- (4) Monte-Carlo methods
- (5) Fast Fourier transform
- (6) Krylov subspace methods (conjugate gradients, Lanczos, GMRES, MINRES)
- (7) JPEG
- (8) PAGERANK
- (9) Simplex algorithm
- (10) Kalman filter

Wir behandeln im Verlauf dieser Lehrveranstaltung von dieser Liste voraussichtlich die **Punkte (1) bis (3), (6) und (8)**. Für das Simplex-Verfahren (**Punkt (9)**) verweisen wir auf die Lehrveranstaltung *Grundlagen der Optimierung*.

Zur Einführung gehen wir hier schon einmal auf die PAGERANK-Aufgabe ein. Dabei handelt es sich um

---

<sup>1</sup><https://press.princeton.edu/ideas/nicholas-higham-on-the-top-10-algorithms-in-applied-mathematics>

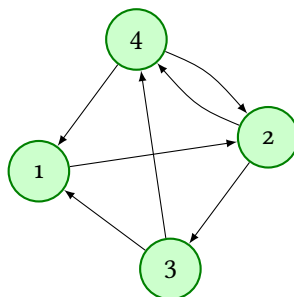


Abbildung 1.1: Ein Miniatur-Web zur Demonstration des PAGERANK-Verfahrens.

die von Sergey Brin und Lawrence Page in [Brin, Page, 1998](#) vorgeschlagenen Ranking-Methode für Webseiten, die seit dem Gründungsjahr 1998 von Google bestimmt, in welcher Reihenfolge Suchergebnisse der Google-Suche erscheinen. Unsere Beschreibung folgt [Higham u. a., 2016](#), Chapter VI.9.

Die in das Ranking einzubeziehenden Webseiten werden als sogenannter einfacher gerichteter Graph dargestellt, siehe [Abbildung 1.1](#). Dabei sind die Webseiten die **Knoten** (englisch: *vertex*) des Graphen, und ein Pfeil (eine **Kante**, englisch: *edge, arc*) bedeutet, dass eine Webseite per Link auf eine andere verweist.

Der Graph wird mittels einer **Adjazenzmatrix** (englisch: *adjacency matrix*) codiert:

$$A = \begin{matrix} & \textcircled{1} & \textcircled{2} & \textcircled{3} & \textcircled{4} \\ \begin{matrix} \textcircled{1} \\ \textcircled{2} \\ \textcircled{3} \\ \textcircled{4} \end{matrix} & \begin{bmatrix} 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix} & \begin{matrix} \textcircled{1} \\ \textcircled{2} \\ \textcircled{3} \\ \textcircled{4} \end{matrix} \end{matrix} \quad (1.1)$$

Dabei sind in der ersten Spalte diejenigen Kanten aufgeführt, die in Knoten  $\textcircled{1}$  starten usw.<sup>2</sup> In der ersten Zeile finden wir diejenigen Kanten, die im Knoten  $\textcircled{1}$  enden.

**Quizfrage:** Welche Bedeutung haben die Spaltensummen von  $A$ ? Welche Bedeutung haben die Zeilensummen von  $A$ ?

Wir stellen uns nun eine Person vor, die sich aktuell auf einer dieser Webseiten  $\textcircled{7}$  aufhält. Ihr Verhalten wird wie folgt modelliert. Mit einer Wahrscheinlichkeit  $\alpha$  wird die Person einem (von der aktuellen Webseite wegführenden) Link folgen. Jedem zur Verfügung stehenden Link wird dabei mit derselben Wahrscheinlichkeit gefolgt. Andernfalls, also mit Wahrscheinlichkeit  $1 - \alpha$ , begibt sich die Person zu einer zufälligen Webseite aus  $\textcircled{1} - \textcircled{4}$ .

Befindet sich die Person beispielsweise zu einem Zeitpunkt auf Webseite  $\textcircled{2}$ , so wird sie sich zum

<sup>2</sup>**Beachte:** Oft wird die Adjazenzmatrix gerichteter Graphen gerade als die Transponierte unserer Matrix  $A$  definiert. Dann stehen die Zeilen für die Anfangsknoten und die Spalten für die Endknoten. Für uns ist aber die Konvention wie in (1.1) günstig.

nächsten Zeitpunkt

$$\begin{aligned}
 &\text{auf Webseite ① mit Wahrscheinlichkeit } \alpha \cdot \frac{0}{2} + (1 - \alpha) \cdot \frac{1}{4} \\
 &\text{auf Webseite ② mit Wahrscheinlichkeit } \alpha \cdot \frac{0}{2} + (1 - \alpha) \cdot \frac{1}{4} \\
 &\text{auf Webseite ③ mit Wahrscheinlichkeit } \alpha \cdot \frac{1}{2} + (1 - \alpha) \cdot \frac{1}{4} \\
 &\text{auf Webseite ④ mit Wahrscheinlichkeit } \alpha \cdot \frac{1}{2} + (1 - \alpha) \cdot \frac{1}{4}
 \end{aligned}$$

befinden. (**Quizfrage:** Klar?) Der Faktor  $\frac{1}{2}$  kommt daher, dass die Webseite ② gerade zwei wegführende Links hat.

Insgesamt kann man diese Überlegungen in der Matrix  $\Pi$  der Übergangswahrscheinlichkeiten zusammenfassen, die in unserem Beispiel wie folgt aussieht:

$$\Pi = \alpha \begin{bmatrix} 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix} + (1 - \alpha) \frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}. \quad (1.2)$$

Die Einträge  $\pi_{ij}$  dieser Matrix beschreiben, mit welcher Wahrscheinlichkeit die Person von Webseite ① (Spalte) im nächsten Zeitpunkt auf die Webseite ① (Zeile) gelangt.

Besteht das Netzwerk nun allgemeiner aus  $n$  Webseiten, ist weiterhin  $A$  die Adjazenzmatrix und  $D$  die Diagonalmatrix mit den Kehrwerten der Anzahlen der wegführenden Links<sup>3</sup> sowie  $e = (1, 1, \dots, 1)^T \in \mathbb{R}^n$ , dann hat die Matrix der Übergangswahrscheinlichkeiten die Gestalt

$$\Pi = \alpha A D + (1 - \alpha) \frac{1}{n} e e^T. \quad (1.3)$$

**Quizfrage:** Welche Eigenschaften fallen Ihnen zu dieser Matrix ein? Was sind insbesondere die Spaltensummen  $\sum_{i=1}^n \pi_{ij}$ ?

Welcher Zusammenhang besteht nun zwischen der Matrix  $\Pi$  und dem Ranking der Webseiten in unserem Netzwerk? Dazu betrachten wir folgendes Gedankenexperiment. Nehmen wir an, die Person startet zum Zeitpunkt 0 auf Webseite ②, also mit einer Wahrscheinlichkeitsverteilung von  $x^{(0)} = (0, 1, 0, 0)^T$ . Zum Zeitpunkt 1 ist sie dann mit den Wahrscheinlichkeiten auf den Seiten ①–④ angekommen, die sich aus den Einträgen von

$$x^{(1)} = \Pi x^{(0)}$$

ergeben. Zum Zeitpunkt 2 haben wir dann die Verteilung  $x^{(2)} = \Pi x^{(1)} = \Pi^2 x^{(0)}$  usw. Man kann im Fall von  $\alpha \in (0, 1)$  zeigen, dass diese Folge  $(x^{(n)})$  der Wahrscheinlichkeiten gegen eine stationäre Verteilung konvergiert, und zwar unabhängig davon, mit welcher Anfangsverteilung  $x^{(0)}$  wir beginnen.<sup>4</sup> Die Eigenschaft, dass  $x^{(n)}$  eine Wahrscheinlichkeitsverteilung beschreibt (also  $x^{(n)} \geq 0$  und  $e^T x^{(n)} = 1$ ) überträgt sich von der Anfangsverteilung  $x^{(0)}$  auf alle Nachfolger. **Quizfrage:** Warum?

<sup>3</sup>Wir gehen hier der Einfachheit davon aus, dass jede Webseite im betrachteten Netzwerk wenigstens einen von der Seite wegführenden Link hat.

<sup>4</sup>Wir haben es hier mit einer stationären Markow-Kette mit endlichem Zustandsraum zu tun.

Die Stationarität bedeutet, dass sich die Wahrscheinlichkeiten nicht mehr ändern, dass also gilt:

$$\Pi x = x. \quad (1.4)$$

Diese Gleichung (1.4) kann man auf verschiedene Arten interpretieren. Eine davon besagt, dass  $x$  ein Eigenvektor der Matrix  $\Pi$  zum Eigenwert  $\lambda = 1$  des Eigenwertproblems

$$\Pi x = \lambda x \quad (1.5)$$

ist. Man kann unter geeigneten Annahmen zeigen (Stichwort: **Perron-Frobenius-Theorem**),

- dass tatsächlich  $\lambda = 1$  immer ein Eigenwert der Übergangsmatrix  $\Pi$  ist und daher eine Lösung  $x \neq 0$  von (1.5) existiert,
- dass alle anderen Eigenwerte von  $\Pi$  vom Betrag her  $\leq 1$  sind,
- dass  $\lambda = 1$  der einzige strikt positive, reelle Eigenwert von  $\Pi$  ist – die anderen Eigenwerte sind entweder negativ reell oder komplexwertig,
- dass die algebraische, also auch die geometrische Vielfachheit des Eigenwerts  $\lambda = 1$  gleich eins ist,
- dass der zugehörige Eigenvektor  $x$  so gewählt werden kann, dass er nur Einträge  $\geq 0$  hat und für die Gesamtsumme gilt:  $e^\top x = 1$ .

Die Größe der Einträge des gesuchten Eigenvektors  $x$  gibt gerade die Wichtigkeit der zugehörigen Webseite ①–④ im Sinne der PAGERANK-Idee an. Die PAGERANK-Aufgabe ist also letztendlich eine **Eigenwertaufgabe** (englisch: *eigenvalue problem*)!

Angesichts der zur Zeit (Stand: 2022) geschätzten Anzahl von etwa  $n \approx 2 \cdot 10^9$  Webseiten und der entsprechenden Dimensionen der Matrizen  $A$  und  $\Pi$  ist neben der Frage, wie/ob man diese Matrizen speichert, auch die Frage relevant, wie man den gesuchten Eigenvektor effizient berechnet.

**Quizfrage:** Die Adjazenzmatrix  $A$  des heutigen Internet ist extrem dünn besetzt. Trifft das auch für die Matrix  $\Pi$  zu? Was hat das für Konsequenzen?

Wir nehmen hier vorweg, dass glücklicherweise für die Bestimmung von Eigenpaaren für betragsgrößte Eigenvektoren einer Matrix mit der **Potenzmethode** (auch **Vektoriteration**, **Von-Mises-Iteration**, englisch: *power iteration*) eine matrixfreie Methode zur Verfügung steht. Wir geben sie im Folgenden an und kommen darauf später in der Lehrveranstaltung zurück. Für die Matrix  $\Pi$  aus (1.2) kann man die Potenzmethode wie in **Algorithmus 1.1** gezeigt aufschreiben.

**Algorithmus 1.1** (PAGERANK-Algorithmus mittels Potenzmethode).

**Eingabe:** Adjazenzmatrix  $A \in \{0, 1\}^{n \times n}$

**Eingabe:** Link-Folge-Wahrscheinlichkeit  $\alpha \in (0, 1)$

**Ausgabe:** ungefähre Eigenvektor  $x \geq 0$  der Matrix  $\Pi$  aus (1.3) zum Eigenwert  $\lambda = 1$

- 1: Initialisiere  $x^{(0)} := e/n$  //  $e = (1, 1, \dots, 1)^\top \in \mathbb{R}^n$
- 2: Setze  $k := 0$
- 3: **while** noch nicht konvergiert **do**
- 4:     Setze  $x^{(k+1)} := \alpha A D x^{(k)} + (1 - \alpha) \frac{1}{n} e$  //  $x^{(k+1)} = \Pi x^{(k)}$
- 5:     Setze  $x^{(k+1)} := x^{(k+1)} / e^\top x^{(k+1)}$  // Normalisierung



6:     Setze  $k := k + 1$   
 7: **end while**

Als Konvergenzkriterium in Zeile 3 kann man etwa heranziehen, ob sich die Einträge von  $x^{(k+1)}$  nur noch sehr wenig von denen von  $x^{(k)}$  unterscheiden. Die Initialisierung in Zeile 1 kann auch anders erfolgen, etwa indem man  $x^{(0)}$  mit Werten in  $[0, 1]$  zufällig wählt und dann normalisiert, sodass  $e^T x^{(0)} = 1$  gilt, oder durch die Wahl eines Einheitsvektors (ein Eintrag eins, Rest Nullen).

**Quizfrage:** Wie kommt man auf die Anweisung in Zeile 4? Was ist mit der Matrix in (1.3) passiert, die als lauter Einsen besteht?

**Quizfrage:** Wozu könnte der Schritt in Zeile 5 dienen? Sollte dieser nicht eigentlich überflüssig sein?

Für unser obiges Beispiel aus Abbildung 1.1 ergibt sich bei  $\alpha = 0.85$  und Startvektor  $x^{(0)} = (0, 1, 0, 0)^T$  auf vier Stellen gerundet die Folge

$$\begin{aligned} x^{(0)} &= (0.0000, 1.0000, 0.0000, 0.0000)^T \\ x^{(1)} &= (0.0375, 0.0375, 0.4625, 0.4625)^T \\ x^{(2)} &= (0.4306, 0.2659, 0.0534, 0.2500)^T \\ &\vdots \\ x^{(23)} &= (0.2240, 0.3374, 0.1809, 0.2578)^T \\ x^{(24)} &= (0.2239, 0.3374, 0.1809, 0.2578)^T \\ x^{(25)} &= (0.2239, 0.3374, 0.1809, 0.2578)^T \\ &\vdots \end{aligned}$$

Die Abhängigkeit der PAGERANK:s von der Link-Folge-Wahrscheinlichkeit  $\alpha$  für das Beispiel-Netzwerk aus Abbildung 1.1 wird in Abbildung 1.2 gezeigt. (**Quizfrage:** Wie kann man Abbildung 1.2, vor allem für  $\alpha$  in der Nähe von 0 und 1 erklären?) Im Original-Paper Brin, Page, 1998 wird eine Link-Folge-Wahrscheinlichkeit von  $\alpha = 0.85$  angenommen.

## § 2 WIEDERHOLUNG UND NOTATION

Wir gehen davon aus, dass der Leser mit Grundkonzepten der mehrdimensionalen Analysis und der linearen Algebra vertraut ist, darunter insbesondere Stetigkeit und Differenzierbarkeit von Funktionen sowie Matrix- und Vektorrechnung im Vektorraum  $\mathbb{R}^n$ .<sup>5</sup>

Wir bezeichnen die nicht-negativen bzw. die positiven reellen Zahlen mit

$$\mathbb{R}_{\geq 0} := \{x \in \mathbb{R} \mid x \geq 0\} \quad \text{und} \quad \mathbb{R}_{> 0} := \{x \in \mathbb{R} \mid x > 0\}.$$

Die natürlichen Zahlen sind  $\mathbb{N} = \{1, 2, \dots\}$ , und wir setzen  $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$ .

<sup>5</sup>Vieles kann drekt auch auf den komplexen Vektorraum  $\mathbb{C}^n$  übertragen werden.

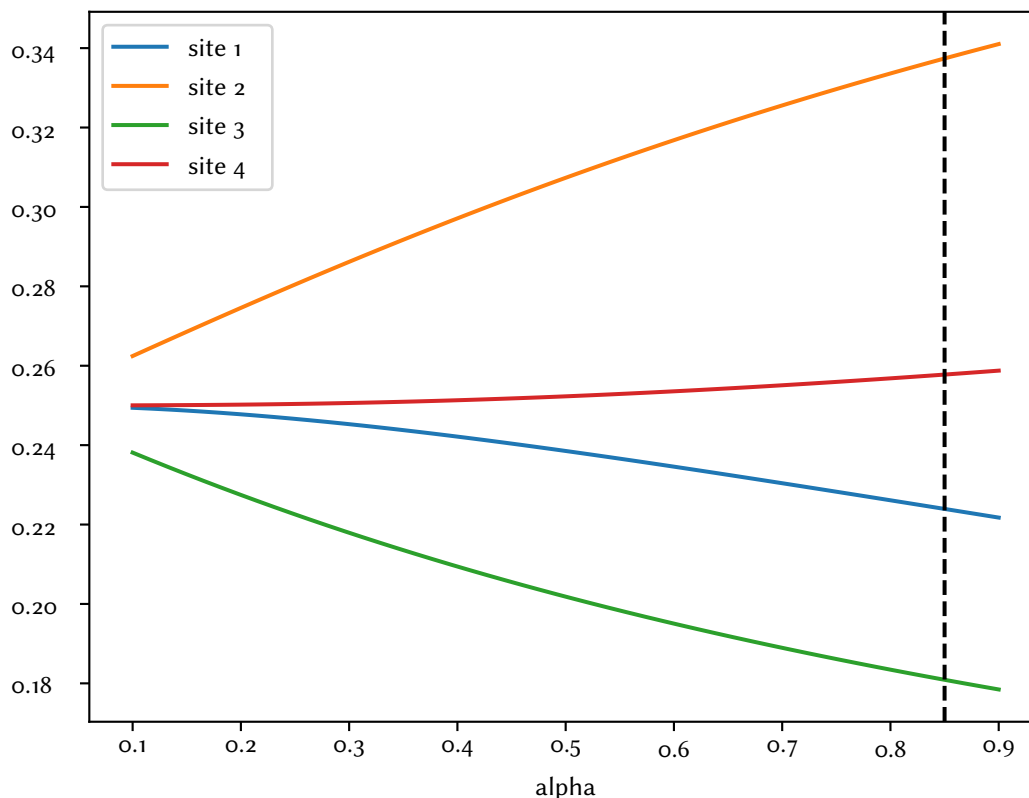


Abbildung 1.2: Gewichte (Einträge im Eigenvektor zum Eigenwert  $\lambda = 1$ ) der Webseiten für das Beispiel-Netzwerk aus [Abbildung 1.1](#) in Abhängigkeit der Link-Folge-Wahrscheinlichkeit  $\alpha$ .

## § 2.1 VEKTOR- UND MATRIX-NORMEN

**Definition 2.1** (Norm). *Es sei  $V$  irgendein Vektorraum über  $\mathbb{R}$ . Eine Abbildung  $\|\cdot\|: V \rightarrow \mathbb{R}$  heißt eine **Norm** (englisch: **norm**) auf  $V$ , wenn für beliebige  $x, y \in V$  und  $\alpha \in \mathbb{R}$  folgende Eigenschaften gelten:*

(i)  $\|x\| \geq 0$  und  $\|x\| = 0 \Rightarrow x = 0$ .

(**Definitheit**)

(ii)  $\|\alpha x\| = |\alpha| \|x\|$ .

(**positive Homogenität**)

(iii)  $\|x + y\| \leq \|x\| + \|y\|$ .

(**Subadditivität / Dreiecksungleichung**)

## VEKTORNORMEN

Wir arbeiten in dieser Lehrveranstaltung im Vektorraum  $\mathbb{R}^n$  der Spaltenvektoren, soweit nichts anderes gesagt wird, mit dem **Euklidischen Innenprodukt** (englisch: *Euclidean inner product*)

$$x^T y = \sum_{i=1}^n x_i y_i \quad \text{für } x, y, \in \mathbb{R}^n,$$

der daraus erzeugten **Euklidischen Norm** (auch: **2-Norm**)

$$\|x\|_2 := \sqrt{x^T x} \quad \text{für } x \in \mathbb{R}^n$$

und dem wiederum daraus erzeugten Abstand (der **Metrik**, (englisch: *metric*))

$$\|x - y\|_2 \quad \text{für } x, y, \in \mathbb{R}^n.$$

Wie für jedes Innenprodukt gilt die **Cauchy-Schwarz-Ungleichung** (englisch: *Cauchy-Schwarz inequality*)

$$x^T y \leq |x^T y| \leq \|x\|_2 \|y\|_2 \quad \text{für alle } x, y, \in \mathbb{R}^n,$$

wobei die Gleichheit  $|x^T y| = \|x\|_2 \|y\|_2$  genau dann gilt, wenn  $x$  und  $y$  linear abhängig sind.

**Quizfrage:** Wie sehen das Euklidische Innenprodukt und die daraus erzeugte Euklidische Norm für **Zeilenvektoren** (englisch: *row vectors*) im  $\mathbb{R}_n$  aus?

Im Folgenden gelten alle Aussagen für  $\mathbb{R}^n$  sinngemäß auch für  $\mathbb{R}_n$ .

Wir benötigen außerdem die **1-Norm** und die  **$\infty$ -Norm** von Vektoren im  $\mathbb{R}^n$ , definiert durch

$$\|x\|_1 := \sum_{i=1}^n |x_i| \quad \text{und} \quad \|x\|_\infty := \max_{i=1, \dots, n} |x_i|.$$

**Beachte:** In den endlich-dimensionalen Vektorräumen  $\mathbb{R}^n$  sind alle Normen äquivalent. Es ist also unerheblich, in welcher Norm wie z. B. Konvergenz oder Fehler messen. Die Wahl erfolgt häufig danach, wie es am einfachsten geht. Für die 1-, 2- und  $\infty$ -Normen im  $\mathbb{R}^n$  gelten die folgenden Abschätzungen, die die Äquivalenz dieser Normen untereinander bestätigen:

$$\|x\|_2 \leq \|x\|_1 \leq \sqrt{n} \|x\|_2, \tag{2.1a}$$

$$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n} \|x\|_\infty, \tag{2.1b}$$

$$\|x\|_\infty \leq \|x\|_1 \leq n \|x\|_\infty. \tag{2.1c}$$

**Quizfrage:** Für welche Vektoren im  $\mathbb{R}^n$  sind die Abschätzungen im (2.1) jeweils scharf?

## MATRIXNORMEN

Matrizen kennzeichnen wir i. d. R. mit Großbuchstaben und ihre Einträge mit kleinen Buchstaben, beispielsweise hat eine Matrix  $A \in \mathbb{R}^{m \times n}$  die Einträge  $a_{ij}$  mit  $i = 1, \dots, m$  und  $j = 1, \dots, n$ . Manchmal

schreiben wir statt  $A$  dann auch  $(a_{ij})$ , um die gesamte Matrix zu adressieren. Die **Einheitsmatrix** bezeichnen wir mit  $\text{Id}$ . Wenn wir die Dimension betonen wollen, schreiben wir auch  $\text{Id}_n$  für die Einheitsmatrix der Größe  $n \times n$ .

Auch im Vektorraum der Matrizen  $A \in \mathbb{R}^{m \times n}$  bzw. der durch sie repräsentierten linearen Abbildung  $\mathbb{R}^n \ni x \mapsto Ax \in \mathbb{R}^m$  lassen sich Normen definieren. Wir betrachten hier vorwiegend die durch die Vektornormen  $\|\cdot\|_2$ ,  $\|\cdot\|_1$  und  $\|\cdot\|_\infty$  induzierten Operatornormen:

**Definition 2.2** (Operatornorm einer Matrix). *Es sei  $A \in \mathbb{R}^{m \times n}$ . Weiter seien  $\|\cdot\|_X$  bzw.  $\|\cdot\|_Y$  die in den Vektorräumen  $\mathbb{R}^n$  und  $\mathbb{R}^m$  verwendeten Vektornormen. Die Größe*

$$\|A\|_{X \rightarrow Y} := \sup_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|Ax\|_Y}{\|x\|_X} \quad (2.2)$$

heißt die (durch die Normen  $\|\cdot\|_X$  im  $\mathbb{R}^n$  und  $\|\cdot\|_Y$  im  $\mathbb{R}^m$  induzierte) **Operatornorm** (englisch: operator norm) von  $A$ .

**Lemma 2.3** (Eigenschaften der Operatornorm). *Es seien  $A \in \mathbb{R}^{m \times n}$  und  $B \in \mathbb{R}^{\ell \times m}$ . In den Räumen  $\mathbb{R}^n$ ,  $\mathbb{R}^m$  und  $\mathbb{R}^\ell$  verwenden wir die Normen  $\|\cdot\|_X$ ,  $\|\cdot\|_Y$  bzw.  $\|\cdot\|_Z$ . Für die Operatornorm (2.2) gelten folgende Eigenschaften:*

(i) *Die Operatornorm ist eine Norm auf  $\mathbb{R}^{m \times n}$ .*

(ii)  $\|Ax\|_Y \leq \|A\|_{X \rightarrow Y} \|x\|_X$  für alle  $x \in \mathbb{R}^n$ .

(iii) *Es gilt*

$$\|A\|_{X \rightarrow Y} = \sup_{\substack{x \in \mathbb{R}^n \\ \|x\|_X=1}} \|Ax\|_Y = \max_{\substack{x \in \mathbb{R}^n \\ \|x\|_X=1}} \|Ax\|_Y.$$

(iv) *Es gilt*

$$\begin{aligned} \|A\|_{X \rightarrow Y} &= \inf\{c \geq 0 \mid \|Ax\|_Y \leq c \|x\|_X \text{ für alle } x \in \mathbb{R}^n\} \\ &= \min\{c \geq 0 \mid \|Ax\|_Y \leq c \|x\|_X \text{ für alle } x \in \mathbb{R}^n\}. \end{aligned}$$

(v)  $\|BA\|_{X \rightarrow Z} \leq \|B\|_{Y \rightarrow Z} \|A\|_{X \rightarrow Y}$  (**Submultiplikativität**).

**Beachte:** Aussage (iv) bedeutet: Kennt man eine Zahl  $c \geq 0$ , mit der  $\|Ax\|_Y \leq c \|x\|_X$  für alle  $x \in \mathbb{R}^n$  gilt, so besteht die Abschätzung  $\|A\|_{X \rightarrow Y} \leq c$ .

*Beweis.* Findet sich als Übungsaufgabe 1 auf Übungsblatt 1. □

**Lemma 2.4** (Operatornormen). *Für Matrizen  $A \in \mathbb{R}^{m \times n}$  und ihre Operatornormen, induziert durch die Verwendung jeweils der 1-Norm, der 2-Norm bzw. der  $\infty$ -Norm im Definitionsraum  $\mathbb{R}^n$  wie auch im*

Zielraum  $\mathbb{R}^m$  gelten die folgenden Aussagen:

$$\|A\|_{1 \rightarrow 1} = \max_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|Ax\|_1}{\|x\|_1} = \max_{j=1, \dots, n} \sum_{i=1}^m |a_{ij}| \quad (\text{Spaltensummennorm}), \quad (2.3a)$$

$$\|A\|_{2 \rightarrow 2} = \max_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|Ax\|_2}{\|x\|_2} = \text{Wurzel des maximalen Eigenwerts von } A^T A \quad (\text{Spektralnrm}), \quad (2.3b)$$

$$\|A\|_{\infty \rightarrow \infty} = \max_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|Ax\|_\infty}{\|x\|_\infty} = \max_{i=1, \dots, m} \sum_{j=1}^n |a_{ij}| \quad (\text{Zeilensummennorm}). \quad (2.3c)$$

*Beweis.* Zum Beweis von (2.3a) schätzen wir ab:

$$\begin{aligned} \|Ax\|_1 &= \sum_{i=1}^m \left| \sum_{j=1}^n a_{ij} x_j \right| \\ &\leq \sum_{i=1}^m \sum_{j=1}^n |a_{ij}| |x_j| && \text{(Dreiecksungleichung)} \\ &= \sum_{j=1}^n |x_j| \sum_{i=1}^m |a_{ij}| \\ &\leq \sum_{j=1}^n |x_j| \max_{k=1, \dots, n} \sum_{i=1}^m |a_{ik}| \\ &= \|x\|_1 \max_{j=1, \dots, n} \sum_{i=1}^m |a_{ij}| \quad \text{für alle } x \in \mathbb{R}^n. \end{aligned}$$

Es gilt also (siehe Text nach Lemma 2.3)  $\|A\|_{1 \rightarrow 1} \leq \max_{j=1, \dots, n} \sum_{i=1}^m |a_{ij}|$ . Um die Gleichheit zu zeigen, sei  $j^*$  einer derjenigen Indizes, für die

$$\sum_{i=1}^m |a_{ij^*}| = \max_{j=1, \dots, n} \sum_{i=1}^m |a_{ij}|$$

gilt. Wir wählen  $x^* := e_{j^*}$  als  $j^*$ -ten Einheitsvektor in  $\mathbb{R}^n$ . Dann gilt  $\|x^*\|_1 = 1$  und

$$\|Ax^*\|_1 = \|Ae_{j^*}\|_1 = \left\| \begin{pmatrix} a_{1j^*} \\ \vdots \\ a_{mj^*} \end{pmatrix} \right\|_1 = \sum_{i=1}^m |a_{ij^*}| = \max_{j=1, \dots, n} \sum_{i=1}^m |a_{ij}| = \|x^*\|_1 \max_{j=1, \dots, n} \sum_{i=1}^m |a_{ij}|,$$

also  $\|A\|_{1 \rightarrow 1} \geq \max_{j=1, \dots, n} \sum_{i=1}^m |a_{ij}|$ . Zusammen mit der obigen umgekehrten Abschätzung ist (2.3a) gezeigt.

Zum Beweis von (2.3c) schätzen wir wie folgt ab:

$$\begin{aligned}
 \|Ax\|_\infty &= \max_{i=1,\dots,m} \left| \sum_{j=1}^n a_{ij} x_j \right| \\
 &\leq \max_{i=1,\dots,m} \sum_{j=1}^n |a_{ij}| |x_j| && \text{(Dreiecksungleichung)} \\
 &\leq \max_{j=1,\dots,n} |x_j| \max_{i=1,\dots,m} \sum_{j=1}^n |a_{ij}| \\
 &= \|x\|_\infty \max_{i=1,\dots,m} \sum_{j=1}^n |a_{ij}| \quad \text{für alle } x \in \mathbb{R}^n.
 \end{aligned}$$

Es gilt also  $\|A\|_{\infty \rightarrow \infty} \leq \max_{i=1,\dots,m} \sum_{j=1}^n |a_{ij}|$ . Um die Gleichheit zu zeigen, sei  $i^*$  einer derjenigen Indizes, für die

$$\sum_{j=1}^n |a_{i^*j}| = \max_{i=1,\dots,m} \sum_{j=1}^n |a_{ij}|$$

gilt. Wir wählen  $x^* := (\operatorname{sgn} a_{i^*1}, \dots, \operatorname{sgn} a_{i^*n})^\top$ .<sup>6</sup> Dann gilt  $\|x^*\|_\infty = 1$  und

$$\|Ax^*\|_\infty = \max_{i=1,\dots,m} \left| \sum_{j=1}^n a_{ij} x_j^* \right| \geq \left| \sum_{j=1}^n a_{i^*j} x_j^* \right| = \left| \sum_{j=1}^n |a_{i^*j}| \right| = \sum_{j=1}^n |a_{i^*j}| = \max_{i=1,\dots,m} \sum_{j=1}^n |a_{ij}|,$$

also  $\|A\|_{\infty \rightarrow \infty} \geq \max_{i=1,\dots,m} \sum_{j=1}^n |a_{ij}|$ . Zusammen mit der obigen umgekehrten Abschätzung ist (2.3c) gezeigt.

Den Beweis von (2.3b) liefern wir in § 4 nach. □

Für die Operatornormen von Spaltenvektoren  $y \in \mathbb{R}^m \cong \mathbb{R}^{m \times 1}$  und Zeilenvektoren  $y \in \mathbb{R}_n \cong \mathbb{R}^{1 \times n}$  gilt Folgendes:

**Folgerung 2.5** (Operatornormen von Zeilen- und Spaltenvektoren).

(i) Für *Spaltenvektoren*  $y \in \mathbb{R}^m$ , die wir auch als Matrix  $y \in \mathbb{R}^{m \times 1}$  interpretieren, gilt:

$$\|y\|_{1 \rightarrow 1} = \|y\|_1, \quad \|y\|_{2 \rightarrow 2} = \|y\|_2, \quad \|y\|_{\infty \rightarrow \infty} = \|y\|_\infty \quad (2.4)$$

(ii) Für *Zeilenvektoren*  $y \in \mathbb{R}_n$ , die wir auch als Matrix  $y \in \mathbb{R}^{1 \times n}$  interpretieren, gilt:

$$\|y\|_{1 \rightarrow 1} = \|y\|_\infty, \quad \|y\|_{2 \rightarrow 2} = \|y\|_2, \quad \|y\|_{\infty \rightarrow \infty} = \|y\|_1 \quad (2.5)$$

*Beweis.* Aussage (i): Für  $y \in \mathbb{R}^m \cong \mathbb{R}^{m \times 1}$  gilt nach (2.3a) bzw. (2.3c)

$$\|y\|_{1 \rightarrow 1} = \sum_{i=1}^m |y_{i1}| = \|y\|_1 \quad \text{und} \quad \|y\|_{\infty \rightarrow \infty} = \max_{i=1,\dots,m} |y_{i1}| = \|y\|_\infty.$$

<sup>6</sup>Das **Signum** einer reellen Zahl  $x$  ist definiert als  $\operatorname{sgn} x = 1$  für  $x > 0$ ,  $\operatorname{sgn} x = -1$  für  $x < 0$  und  $\operatorname{sgn} x = 0$  für  $x = 0$ .

Da wir (2.3b) noch nicht bewiesen haben, untersuchen wir die 2-Norm direkt:

$$\|y\|_{2 \rightarrow 2} = \max_{\substack{\alpha \in \mathbb{R} \\ |\alpha|=1}} \|\alpha y\|_2 = \|y\|_2.$$

**Aussage (ii):** Für  $y \in \mathbb{R}^n \cong \mathbb{R}^{1 \times n}$  gilt nach (2.3a) bzw. (2.3c)

$$\|y\|_{1 \rightarrow 1} = \max_{j=1, \dots, n} |y_{1j}| = \|y\|_\infty \quad \text{und} \quad \|y\|_{\infty \rightarrow \infty} = \sum_{j=1}^n |y_{1j}| = \|y\|_1.$$

Die 2-Norm untersuchen wir wieder direkt:

$$\|y\|_{2 \rightarrow 2} = \max_{\substack{x \in \mathbb{R}^n \\ \|x\|_2=1}} |y x| \leq \max_{\substack{x \in \mathbb{R}^n \\ \|x\|_2=1}} \|y\|_2 \|x\|_2.$$

Im Fall  $y = 0$  sind ohnehin beide Seiten gleich Null. Andernfalls erreichen wir durch die Wahl von  $x = y/\|y\|_2$  die Gleichheit (Cauchy-Schwarz-Ungleichung).  $\square$

## § 2.2 DIFFERENZIERBARKEIT UND DER SATZ VON TAYLOR

Zwei wesentliche Hilfsmittel zur Analyse von Aufgaben und Algorithmen sind die Kettenregel und der Satz von Taylor, die wir hier kurz wiederholen:

**Definition 2.6** (Differenzierbarkeit, siehe z. B. Heuser, 2002, Definition 164). *Es sei  $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$  eine Funktion.  $F$  heißt **differenzierbar** (englisch: **differentiable**) an der Stelle  $x_0 \in \mathbb{R}^n$ , wenn es eine Matrix  $A \in \mathbb{R}^{m \times n}$  gibt, sodass für die durch*

$$r(x_0; \Delta x) := F(x_0 + \Delta x) - F(x_0) - A \Delta x$$

definierte **Restgliedfunktion** (englisch: **remainder function**) gilt:<sup>7</sup>

$$\lim_{\Delta x \rightarrow 0} \frac{r(x_0; \Delta x)}{\|\Delta x\|} = 0. \quad (2.6)$$

In diesem Fall heißt die durch die Definition eindeutig bestimmte Matrix  $A$  die **Ableitung** (englisch: **derivative**) von  $F$  an der Stelle  $x_0$  und wird mit  $F'(x_0)$  bezeichnet. Sie stimmt überein mit der **Jacobi-matrix** (der Matrix aus allen partiellen Ableitungen erster Ordnung, englisch: **Jacobian**) von  $F$  an der Stelle  $x_0$ :

$$J_F(x_0) := \begin{bmatrix} \frac{\partial F_1(x_0)}{\partial x_1} & \dots & \frac{\partial F_1(x_0)}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial F_m(x_0)}{\partial x_1} & \dots & \frac{\partial F_m(x_0)}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}. \quad (2.7)$$

<sup>7</sup>Da diese Aussage wegen der Äquivalenz der Normen auf  $\mathbb{R}^n$  für jede Vektornorm  $\|\cdot\|$  gilt, wird die Norm hier nicht näher spezifiziert.

Partielle Ableitungen zweiter Ordnung werden, sofern sie existieren, nach dem Muster

$$\frac{\partial^2 F_j(x_0)}{\partial x_2 \partial x_1} := \frac{\partial}{\partial x_2} \frac{\partial F_j}{\partial x_1}(x_0)$$

gebildet. Analog gibt es auch Ableitungen höherer Ordnung. **Quizfrage:** Wieviele verschiedene partielle Ableitungen der Ordnung  $k \in \mathbb{N}_0$  gibt es?

**Definition 2.7** ( $C^k$ -Funktion). Es sei  $k \in \mathbb{N}_0$  und  $U \subseteq \mathbb{R}^n$  offen. Eine Funktion  $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$  heißt  **$k$ -mal stetig partiell differenzierbar** (englisch: *k times continuously partially differentiable*) auf  $U$  oder eine  **$C^k$ -Funktion** auf  $U$  (kurz:  $C^k(U, \mathbb{R}^m)$ ), wenn  $F$  sowie alle partiellen Ableitungen bis einschließlich zur Ordnung  $k$  existieren und auf  $U$  stetig sind.

**Beachte:** Eine Funktion  $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ , die auf der offenen Menge  $U \subseteq \mathbb{R}^n$  eine  $C^1$ -Funktion ist, ist in jedem Punkt von  $U$  auch differenzierbar, erfüllt also die Restgliedbedingung (2.6) für jedes feste  $x_0 \in U$ . Kurz: Eine einmal stetig partiell differenzierbare Funktion ist differenzierbar.

**Satz 2.8** (Kettenregel).

Es seien  $F: \mathbb{R}^m \rightarrow \mathbb{R}^p$  und  $G: \mathbb{R}^n \rightarrow \mathbb{R}^m$  zwei Funktionen, und es sei  $G$  an der Stelle  $x_0 \in \mathbb{R}^n$  differenzierbar sowie  $F$  an der Stelle  $G(x_0)$  differenzierbar. Dann ist auch die Komposition  $F \circ G$  an der Stelle  $x_0$  differenzierbar, und es gilt

$$(F \circ G)'(x_0) = F'(G(x_0)) G'(x_0)$$

oder ausgeschrieben

$$\begin{bmatrix} \frac{\partial(F \circ G)_1(x_0)}{\partial x_1} & \dots & \frac{\partial(F \circ G)_1(x_0)}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial(F \circ G)_p(x_0)}{\partial x_1} & \dots & \frac{\partial(F \circ G)_p(x_0)}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial F_1(G(x_0))}{\partial x_1} & \dots & \frac{\partial F_1(G(x_0))}{\partial x_m} \\ \vdots & & \vdots \\ \frac{\partial F_p(G(x_0))}{\partial x_1} & \dots & \frac{\partial F_p(G(x_0))}{\partial x_m} \end{bmatrix} \begin{bmatrix} \frac{\partial G_1(x_0)}{\partial x_1} & \dots & \frac{\partial G_1(x_0)}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial G_m(x_0)}{\partial x_1} & \dots & \frac{\partial G_m(x_0)}{\partial x_n} \end{bmatrix}.$$

Die Ableitung (Jacobimatrix) der zusammengesetzten Funktion  $F \circ G$  ergibt sich also aus dem Produkt der Jacobimatrizen  $F'(G(x_0))$  und  $G'(x_0)$ .

Wir geben nun zwei Versionen des Satzes von Taylor für *reellwertige* Funktionen an, die sich in der Darstellung des Restglieds unterscheiden. Analog zur Schreibweise für Intervalle

$$[a, b] := \{x \in \mathbb{R} \mid a \leq x \text{ und } x \leq b\}$$

ist es günstig, eine ähnliche Schreibweise auch für die **Verbindungsstrecke** (englisch: *segment*) zwischen Punkten  $a, b$  im  $\mathbb{R}^n$  einzuführen, also

$$\langle a, b \rangle := \{(1-t)a + tb \mid t \in [0, 1]\}$$

für  $a, b \in \mathbb{R}^n$ . **Quizfrage:** Gilt für reelle Zahlen  $a$  und  $b$  immer  $[a, b] = \langle a, b \rangle$ ?

**Definition 2.9** (Konvexe Menge). Eine Menge  $A \subseteq \mathbb{R}^n$  heißt **konvex** (englisch: *convex*), wenn für  $x, y \in A$  immer auch die gesamte Verbindungsstrecke in  $A$  liegt, also  $\langle x, y \rangle \subseteq A$ .



**Satz 2.10** (Taylor mit Restglied in Zwischenwertform, siehe z. B. Heuser, 2002, Satz 168.1).

Es sei  $U \subseteq \mathbb{R}^n$  offen,  $k \in \{1, 2\}$  und  $f: U \rightarrow \mathbb{R}$  eine  $C^k$ -Funktion auf  $U$ , kurz:  $C^k(U, \mathbb{R})$ . Die Punkte  $x_0$  und  $x_0 + d$  seien so gewählt, dass  $\langle x_0, x_0 + d \rangle$  in  $U$  liegt. Dann existiert eine Zahl  $\xi \in [0, 1]$ , sodass gilt:

$$\text{im Fall } k = 1: \quad f(x_0 + d) = f(x_0) + f'(x_0 + \xi d) d,$$

$$\text{im Fall } k = 2: \quad f(x_0 + d) = f(x_0) + f'(x_0) d + \frac{1}{2} d^T f''(x_0 + \xi d) d.$$

Die erste Aussage ist gerade der **Mittelwertsatz** (englisch: *mean value theorem*). Weiter sind  $f'(x_0)$  und  $f''(x_0)$  die Ableitungen erster bzw. zweiter Ordnung von  $f$  an der Stelle  $x_0$ , die folgende Gestalt besitzen:

$$f'(x_0) = \begin{bmatrix} \frac{\partial f(x_0)}{\partial x_1} & \dots & \frac{\partial f(x_0)}{\partial x_n} \end{bmatrix} \quad (\text{Jacobimatrix})$$

$$f''(x_0) = \begin{bmatrix} \frac{\partial^2 f(x_0)}{\partial x_1 \partial x_1} & \dots & \frac{\partial^2 f(x_0)}{\partial x_1 \partial x_n} \\ \vdots & & \vdots \\ \frac{\partial^2 f(x_0)}{\partial x_n \partial x_1} & \dots & \frac{\partial^2 f(x_0)}{\partial x_n \partial x_n} \end{bmatrix} \quad (\text{Hessematrix, englisch: Hessian}).$$

Als Resultat des Satzes von Schwarz (Heuser, 2002, Satz 162.2) ist die Hessematrix  $f''(x_0)$  dabei symmetrisch, also die Reihenfolge der partiellen Differentiationen unerheblich.

**Quizfrage:** Gilt der **Satz von Taylor 2.10** auch für vektorwertige  $C^k$ -Funktionen  $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ ?

**Satz 2.11** (Taylor mit integralem Restglied, siehe z. B. Heuser, 2002, Satz 168.1).

Unter denselben Voraussetzungen wie in Satz 2.10 gilt

$$\text{im Fall } k = 1: \quad f(x_0 + d) = f(x_0) + \left[ \int_0^1 f'(x_0 + t d) dt \right] d$$

$$= f(x_0) + \int_0^1 f'(x_0 + t d) d dt,$$

$$\text{im Fall } k = 2: \quad f(x_0 + d) = f(x_0) + f'(x_0) d + \frac{1}{2} d^T \left[ \int_0^1 (1-t) f''(x_0 + t d) dt \right] d$$

$$= f(x_0) + f'(x_0) d + \frac{1}{2} \int_0^1 (1-t) d^T f''(x_0 + t d) d dt.$$

Die erste Aussage folgt direkt aus dem **Hauptsatz der Differential- und Integralrechnung**.<sup>8</sup>

**Beachte:** Unter dem jeweils ersten Integral steht eine vektorwertige bzw. matrixwertige Funktion, die einfach komponentenweise integriert wird. Im jeweils zweiten Integral wurden die konstanten (von  $t$  unabhängigen) Inkremente  $d$  in das Integral hineingezogen, wodurch der Integrand reellwertig wird.

**Quizfrage:** In dieser Form gilt **Satz 2.11** auch für vektorwertige  $C^k$ -Funktionen  $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ . Wie genau sollte der Satz dann formuliert werden?

<sup>8</sup> „Eine differenzierbare Funktion ist das Integral ihrer Ableitung.“

**Definition 2.12** (Lipschitz-Stetigkeit). Eine Funktion  $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$  heißt **Lipschitz-stetig** (englisch: Lipschitz continuous) auf der Menge  $A \subseteq \mathbb{R}^n$ , falls es eine Zahl  $L \geq 0$  (genannt: **Lipschitz-Konstante**, englisch: Lipschitz constant) gibt mit der Eigenschaft<sup>9</sup>

$$\|F(x) - F(y)\|_2 \leq L \|x - y\|_2 \quad \text{für alle } x, y \in A. \quad (2.8)$$

In diesem Fall heißt  $F$  dann auch  **$L$ -Lipschitz-stetig** auf  $A$ .

Eine einfache aber wichtige Konsequenz aus dem **Satz von Taylor 2.10** ist die Tatsache, dass  $C^1$ -Funktionen genau dann Lipschitz-stetig sind, wenn ihre Ableitung beschränkt ist, genauer:

**Folgerung 2.13** (Lipschitz-Stetigkeit von  $C^1$ -Funktionen). Es sei  $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$  eine  $C^1$ -Funktion auf der offenen, konvexen Menge  $U \subseteq \mathbb{R}^n$ . Dann sind folgende Aussagen äquivalent:

- (i)  $F$  ist Lipschitz-stetig auf  $U$ .
- (ii) Die Ableitung  $F'$  ist auf  $U$  beschränkt, also  $\sup\{\|F'(x)\|_{2 \rightarrow 2} \mid x \in U\} < \infty$ .

Falls die Aussagen wahr sind, dann ist  $L^* := \sup\{\|F'(x)\|_{2 \rightarrow 2} \mid x \in U\} < \infty$  die kleinstmögliche Lipschitz-Konstante in (2.8).

**Beachte:**  $\|F'(x)\|_{2 \rightarrow 2}$  ist die durch die 2-Norm induzierte Operatornorm der Jacobimatrix von  $F$  an der Stelle  $x$ , vgl. auch (2.3b).

**Beweis.** Wir nehmen zunächst **Aussage (i)** an. Es seien  $x \in U$  und  $d \in \mathbb{R}^n$  fest gewählt sowie  $\alpha > 0$ . Aus der Dreiecksungleichung folgt

$$\begin{aligned} \|F'(x)(\alpha d)\|_2 &= \|F'(x)(\alpha d) \pm [F(x + \alpha d) - F(x)]\|_2 \\ &\leq \|F(x + \alpha d) - F(x) - F'(x)(\alpha d)\|_2 + \|F(x + \alpha d) - F(x)\|_2. \end{aligned}$$

Es gilt also auch

$$\|F'(x)d\|_2 = \frac{\|F'(x)(\alpha d)\|_2}{\alpha} \leq \frac{\|F(x + \alpha d) - F(x) - F'(x)(\alpha d)\|_2}{\alpha} + \frac{\|F(x + \alpha d) - F(x)\|_2}{\alpha}.$$

Für hinreichend kleine  $\alpha > 0$  liegt  $x + \alpha d \in U$ , und wir können für den ersten Term auf der rechten Seite die Voraussetzung der Differenzierbarkeit und für den zweiten Term die Voraussetzung der  $L$ -Lipschitz-Stetigkeit von  $F$  auf  $U$  mit einer Lipschitz-Konstanten  $L$  nutzen. Im Grenzübergang  $\alpha \searrow 0$  folgt:

$$\|F'(x)d\|_2 \leq 0 + L \|d\|_2.$$

Da  $d \in \mathbb{R}^n$  beliebig war, folgt aus **Lemma 2.3** jetzt  $\|F'(x)\|_2 \leq L$ .

<sup>9</sup>Auf der linken und rechten Seite von (2.8) dürften wegen der Äquivalenz der Normen auch andere Vektornormen im  $\mathbb{R}^n$  bzw.  $\mathbb{R}^m$  stehen. Die konkrete Wahl beeinflusst nicht die Frage, ob eine Funktion Lipschitz-stetig ist, sondern nur, welche Lipschitz-Konstanten gültig sind. Die Wahl der 2-Norm ist hier günstig wegen der Formulierung der **Folgerung 2.13**.

Wir nehmen nun **Aussage (ii)** an. Es seien  $x, y \in U$ . Aufgrund der Konvexität von  $U$  liegt auch  $\langle x, y \rangle$  in  $U$ . Aus dem **Satz von Taylor 2.11** folgt

$$F(y) - F(x) = \int_0^1 F'(x + t(y - x)) dt (y - x),$$

also (**Quizfrage:** Warum gelten jeweils die Abschätzungen?)

$$\begin{aligned} \|F(y) - F(x)\|_2 &= \left\| \int_0^1 F'(x + t(y - x)) (y - x) dt \right\| \\ &\leq \int_0^1 \|F'(x + t(y - x)) (y - x)\| dt \\ &\leq \int_0^1 \|F'(x + t(y - x))\|_2 \|y - x\|_2 dt. \end{aligned}$$

Der Integrand  $\|F'(x + t(y - x))\|_2$  ist aber durch  $\sup\{\|F'(x)\|_2 \mid x \in U\}$  beschränkt, da  $x + t(y - x)$  auf der Verbindungsstrecke  $\langle x, y \rangle$  und damit in  $U$  liegt. Damit ist  $L^* := \sup\{\|F'(x)\|_2 \mid x \in U\} < \infty$  eine gültige Lipschitz-Konstante für  $F$  auf  $U$ , also gilt **Aussage (i)**.

Wir müssen noch zeigen, dass  $L^*$  die kleinstmögliche Lipschitz-Konstante ist. Aus dem Beweisteil **Aussage (i)  $\Rightarrow$  Aussage (ii)** erhalten wir aber

$$F \text{ ist } L\text{-Lipschitz} \quad \Rightarrow \quad \|F'(x)\|_2 \leq L \text{ für alle } x \in U \quad \Rightarrow \quad L^* := \sup\{\|F'(x)\|_2 \mid x \in U\} \leq L.$$

□

### § 2.3 LANDAU-NOTATION

Wir geben jetzt noch die (Teile der) Definition der **Landau-Notation** (englisch: *Landau notation*) an, mit deren Hilfe man asymptotisches Verhalten, z. B. von Fehlern, charakterisieren kann.

**Definition 2.14** (Landau-Notation für Funktionen).

Es sei  $g: \mathbb{R}^n \setminus \{0\} \rightarrow \mathbb{R}_{>0}$  eine Funktion.

(i) Es sei  $m \in \mathbb{N}$ . Wir bezeichnen die Menge von Funktionen

$$O(g) := \left\{ F: \mathbb{R}^n \setminus \{0\} \rightarrow \mathbb{R}^m \mid \begin{array}{l} \text{es existiert ein } C \geq 0 \text{ und ein } \varepsilon > 0, \text{ sodass gilt:} \\ \|F(x)\| \leq C g(x) \text{ für alle } x \text{ mit } 0 < \|x\| < \varepsilon \end{array} \right\}$$

als „**Groß-O von g bei Null**“ (englisch: *big-O of g near zero*).

Für  $F \in O(g)$  sagt man auch,  $g$  sei eine **asymptotisch obere Schranke** für  $F$  bei Null.

(ii) Es sei  $m \in \mathbb{N}$ . Wir bezeichnen die Menge von Funktionen

$$o(g) := \left\{ F: \mathbb{R}^n \setminus \{0\} \rightarrow \mathbb{R}^m \mid \begin{array}{l} \text{für alle } c > 0 \text{ existiert ein } \varepsilon > 0, \text{ sodass gilt:} \\ \|F(x)\| \leq c g(x) \text{ für alle } x \text{ mit } 0 < \|x\| < \varepsilon \end{array} \right\}$$

als „**Klein-o von  $g$  bei Null**“ (englisch: little-o of  $g$  near zero).

Für  $F \in o(g)$  sagt man auch,  $F$  sei gegenüber  $g$  **asymptotisch vernachlässigbar** bei Null.

**Beachte:** Wir lassen in der obigen Definition offen, welche Norm im  $\mathbb{R}^m$  wir verwenden, um  $\|F(x)\|$  zu messen, weil es wegen der Äquivalenz der Normen unerheblich ist.

**Quizfrage:** Wie kann man die Restgliedbedingung (2.6) in der Landau-Notation ausdrücken?

**Quizfrage:** Welche Beziehung besteht zwischen den Funktionenklassen  $o(\|\Delta x\|)$  und  $O(\|\Delta x\|^2)$ ?

**Beispiel 2.15** (Beispiele für Landau-Notation für Funktionen).

(i) Die Funktion  $f: \mathbb{R} \rightarrow \mathbb{R}$ , definiert durch  $f(x) = \sin(x)$ , gehört zu  $O(g)$  für  $g(x) = |x|$ .  
Kurz schreibt man auch:  $\sin(x) \in O(|x|)$ .

(ii) Die Funktion  $f: \mathbb{R} \rightarrow \mathbb{R}$ , definiert durch  $f(x) = x^2$ , gehört zu  $o(g)$  für  $g(x) = |x|$ .  
Kurz schreibt man auch:  $x^2 \in o(|x|)$ .

(iii) Die Funktion  $F: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ , definiert durch  $F(x) = \begin{pmatrix} x_1 - x_2 \\ x_1^2 + x_2 \end{pmatrix}$ , gehört zu  $O(g)$  für  $g(x) = \|x\|$ .  
Kurz schreibt man auch:  $F(x) \in O(\|x\|)$ .

Auch für die Beschreibung der Asymptotik von Folgen ist die Landau-Notation hilfreich. Dann ist nicht der Fall „bei Null“, sondern der Fall „bei Unendlich“ interessant.

**Definition 2.16** (Landau-Notation für Folgen).

Es sei  $(y^{(n)})$  eine  $\mathbb{R}_{\geq 0}$ -wertige Folge.

(i) Es sei  $m \in \mathbb{N}$ . Wir bezeichnen die Menge von Folgen

$$O(y^{(n)}) := \left\{ (x^{(n)}) : \mathbb{N} \rightarrow \mathbb{R}^m \left| \begin{array}{l} \text{es existiert ein } C \geq 0 \text{ und ein } n_0 \in \mathbb{N}, \text{ sodass gilt:} \\ \|x^{(n)}\| \leq C y^{(n)} \text{ für alle } n \geq n_0 \end{array} \right. \right\}$$

als „**Groß-O von  $(y^{(n)})$  bei Unendlich**“.

(ii) Es sei  $m \in \mathbb{N}$ . Wir bezeichnen die Menge von Folgen

$$o(y^{(n)}) := \left\{ (x^{(n)}) : \mathbb{N} \rightarrow \mathbb{R}^m \left| \begin{array}{l} \text{für alle } c > 0 \text{ existiert ein } n_0 \in \mathbb{N}, \text{ sodass gilt:} \\ \|x^{(n)}\| \leq c y^{(n)} \text{ für alle } n \geq n_0 \end{array} \right. \right\}$$

als „**Klein-o von  $(y^{(n)})$  bei Unendlich**“.

**Beachte:** Auch in dieser Definition ist wieder irrelevant, welche Norm im  $\mathbb{R}^m$  wir verwenden, um  $\|x^{(n)}\|$  zu messen.

**Beispiel 2.17** (Beispiele für Landau-Notation für Folgen).

- (i) Die reellwertige Folge  $x^{(n)} = 1/n$  gehört zu  $\mathcal{O}(y^{(n)})$  für  $y^{(n)} = 1$ .  
Kurz schreibt man auch:  $1/n \in \mathcal{O}(1)$ .
- (ii) Die reellwertige Folge  $x^{(n)} = 1/n$  gehört zu  $\mathcal{o}(y^{(n)})$  für  $y^{(n)} = 1$ .  
Kurz schreibt man auch:  $1/n \in \mathcal{o}(1)$ .
- (iii) Die  $\mathbb{R}^2$ -wertige Folge  $x^{(n)} = \begin{pmatrix} 1/n \\ 1/n^2 \end{pmatrix}$  gehört zu  $\mathcal{O}(y^{(n)})$  für  $y^{(n)} = 1/n$ .  
Kurz schreibt man auch:  $x^{(n)} \in \mathcal{O}(1/n)$ .

**Quizfrage:** Was bedeuten diese Aussagen?

Ende der Woche 1

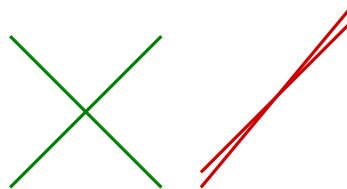
# Kapitel 1 Sensitivitätsanalyse und Kondition von Aufgaben

In diesem Kapitel beschäftigen wir uns zunächst mit Eigenschaften von Aufgaben, insbesondere mit dem Konzept der **Kondition** eines Problems. Der Grund dafür ist, dass jedes numerische Verfahren höchstens so gut sein kann, wie die Kondition der Aufgabe es zulässt.

## § 3 SENSITIVITÄTSANALYSE UND KONDITION

Zur Einstimmung zunächst eine anschauliche Situation.

**Beispiel 3.1** (Schnitt zweier Geraden). *Wir bestimmen grafisch den Schnittpunkt zweier Geraden im  $\mathbb{R}^2$ . Im ersten Fall (links im Bild) ist das Ergebnis, also die Lage des Schnittpunkts, vergleichsweise unempfindlich gegenüber Störungen in den Daten (der Eingabe) der Aufgabe.<sup>1</sup> Wir sprechen von einer **gut konditionierten Aufgabe**. Im zweiten Fall (rechts im Bild) ist dagegen das Ergebnis viel empfindlicher gegenüber Veränderungen in der Eingabe. Wir sprechen dann von einer **schlecht konditionierten Aufgabe**.*



**Quizfrage:** Welche weiteren Beispiele für Aufgaben, für die Sie eine gute oder schlechte Kondition vermuten, fallen Ihnen ein?

Die **Sensitivitätsanalyse** (englisch: *sensitivity analysis*) beschäftigt sich mit der Frage, wie stark das Ergebnis  $y \in \mathbb{R}^m$  einer mathematischen Funktion

$$y = F(x)$$

von Änderungen bzw. Störungen in der Eingabe (den Daten)  $x \in \mathbb{R}^n$  abhängt. Die Darstellung  $y = F(x)$  bedeutet nicht, dass  $y$  notwendigerweise durch eine endliche Abfolge von Operationen explizit

<sup>1</sup>Die Störungen können etwa durch die begrenzte Zeichengenauigkeit auftreten.

berechenbar ist. Vielmehr kann  $y = F(x)$  zum Beispiel auch dafür stehen, dass  $y$  die Lösung eines nichtlinearen Gleichungssystems ist, in das  $x$  als Datum eingeht. Wie man aber bereits an einfachen Beispielen wie  $y = x^2$  sieht, wird die Antwort *lokal* sein, also abhängig davon, welchen Wert die Eingabe  $x$  selbst hat; vgl. [Abbildung 3.1](#).

Wir setzen voraus, dass  $F$  am interessierenden Punkt  $x$  differenzierbar ist. Dann erhalten wir mit Hilfe der Definition der Differenzierbarkeit, dass

$$F(x + \Delta x) \in F(x) + F'(x) \Delta x + o(\|\Delta x\|) \quad (3.1)$$

gilt.<sup>2</sup> Bezeichnen wir mit

$$\Delta y := F(x + \Delta x) - F(x)$$

die durch die Änderung  $\Delta x$  der Eingangsdaten hervorgerufenen Änderungen im Resultat an der Stelle  $x$ , so können wir (3.1) schreiben als

$$\Delta y \in F'(x) \Delta x + o(\|\Delta x\|). \quad (3.2)$$

Das heißt also, dass die durch die Ableitung  $F'(x)$  definierte lineare Abbildung  $\Delta x \mapsto F'(x) \Delta x$  eine derart gute Näherung für  $\Delta y$  ist, dass der dabei gemachte Fehler asymptotisch vernachlässigbar gegenüber  $\|\Delta x\|$  ist. Man schreibt statt (3.2) auch:

$$\Delta y \doteq F'(x) \Delta x \quad (3.3)$$

und sagt,  $\Delta y$  sei gleich  $F'(x) \Delta x$  „bis auf Terme, die gegenüber  $\|\Delta x\|$  vernachlässigbar sind“ oder „bis auf Terme höherer Ordnung in  $\Delta x$ “. Letztere Sprechweise ist dadurch begründet, dass unter der Voraussetzung, dass  $F$  eine  $C^2$ -Funktion in einer Umgebung von  $x$  ist, aus dem [Satz von Taylor 2.11](#) folgt, dass statt (3.2) sogar gilt:  $\Delta y \in F'(x) \Delta x + O(\|\Delta x\|^2)$ . Der letzte Term ist dabei „von höherer Ordnung“ (nämlich quadratisch) in  $\|\Delta x\|$ .

In der **differentiellen Sensitivitätsanalyse** (englisch: *differential sensitivity analysis*) vernachlässigt man tatsächlich die Restgliedterme und stützt sich allein auf die Ableitung  $F'(x)$ . Daraus wollen wir nun im Folgenden verschiedene Informationen gewinnen.

### § 3.1 ABSOLUTE BETRACHTUNGSWEISE

Zunächst lautet (3.3) in Komponenten ausgeschrieben für eine Funktion  $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ :

$$\Delta y = \begin{pmatrix} \Delta y_1 \\ \vdots \\ \Delta y_m \end{pmatrix} \doteq F'(x) \Delta x = \begin{bmatrix} \frac{\partial F_1(x)}{\partial x_1} & \cdots & \frac{\partial F_1(x)}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial F_m(x)}{\partial x_1} & \cdots & \frac{\partial F_m(x)}{\partial x_n} \end{bmatrix} \begin{pmatrix} \Delta x_1 \\ \vdots \\ \Delta x_n \end{pmatrix}. \quad (3.4)$$

<sup>2</sup>**Beachte:** Auf der rechten Seite von (3.1) steht die Summe aus einer Funktion, nämlich  $F(x) + F'(x) \Delta x$ , und einer Menge von Funktionen  $o(\|\Delta x\|)$ . Das ist im Sinne der Minkowski-Summe zu verstehen, d. h., auf der rechten Seite steht die Menge aller Funktionen der Bauart  $\Delta x \mapsto F(x) + F'(x) \Delta x + G(\Delta x)$ , wobei  $G(\Delta x)$  zu  $o(\|\Delta x\|)$  gehört. Alternativ können wir (3.1) auch lesen als:  $F(x + \Delta x) - F(x) - F'(x) \Delta x \in o(\|\Delta x\|)$ .

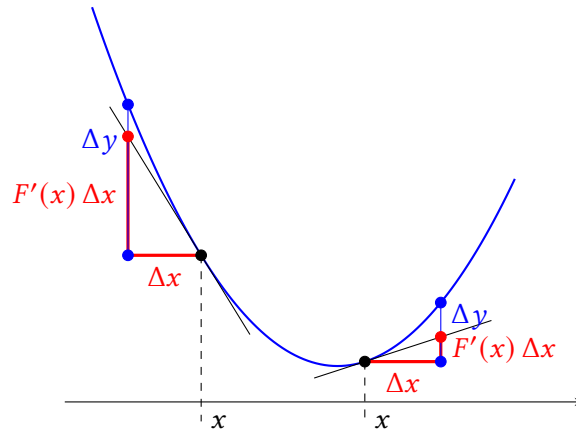


Abbildung 3.1: Wie stark das Ergebnis  $y = F(x) := x^2$  von Änderungen  $\Delta x$  in der Eingabe  $x$  abhängt, ist i. d. R. abhängig vom Wert von  $x$  abhängig.

Der Eintrag

$$K_{ij}(x) := \frac{\partial F_i(x)}{\partial x_j} \quad (3.5)$$

in Zeile  $i$  und Spalte  $j$  der Jacobimatrix gibt also den Faktor an, mit dem sich Änderungen  $\Delta x_j$  in der  $j$ -ten Komponente der Eingabe  $x$  auswirken auf die  $i$ -te Komponente des Ergebnisses.

**Definition 3.2** (Absolute partielle Konditionszahlen). Die Zahlen  $K_{ij}(x)$  heißen die **absoluten partiellen Konditionszahlen** (englisch: absolute partial condition numbers) der Funktion  $F$  an der Stelle  $x$ .

Wollen wir Störungen in **allen Komponenten der Eingabe** gemeinsam betrachten, aber uns im **Ergebnis** auf die  **$i$ -te Komponente** beschränken, also in (3.4) nur die  $i$ -Zeile

$$\Delta y_i \doteq \underbrace{\begin{bmatrix} \frac{\partial F_i(x)}{\partial x_1} & \dots & \frac{\partial F_i(x)}{\partial x_n} \end{bmatrix}}_{F'_i(x)} \begin{pmatrix} \Delta x_1 \\ \vdots \\ \Delta x_n \end{pmatrix} = \sum_{j=1}^n \frac{\partial F_i(x)}{\partial x_j} \Delta x_j \quad (3.6)$$

herausgreifen, so können wir (3.6) mit Hilfe der Cauchy-Schwarz-Ungleichung abschätzen durch<sup>3</sup>

$$|\Delta y_i| \lesssim \|F'_i(x)\|_2 \|\Delta x\|_2. \quad (3.7)$$

Dabei ist  $\|F'_i(x)\|_2$  die 2-Norm des Zeilenvektors bestehend aus den partiellen Ableitungen der  $i$ -ten Komponentenfunktion  $F_i$  von  $F$ , also die Norm der  $i$ -ten Zeile der Jacobimatrix  $F'(x)$ . Diese Abschätzung gilt gleichmäßig, unabhängig von der Richtung des Störungsvektors  $\Delta x$ , es handelt sich also um eine *Worst-Case*-Abschätzung. Aus der Cauchy-Schwarz-Ungleichung wissen wir auch, dass die Abschätzung genau dann mit Gleichheit angenommen wird, also  $|\Delta y_i| \doteq \|F'_i(x)\|_2 \|\Delta x\|_2$  gilt, wenn die Störungsrichtung  $\Delta x$  und die Ableitung  $F'_i(x)^T$  linear abhängig sind, also einer dieser Vektoren ein Vielfaches des anderen ist.

<sup>3</sup>Auch hier bedeutet der Punkt auf dem Symbol  $\lesssim$ , dass die Abschätzung gilt bis auf Terme, die gegenüber  $\|\Delta x\|$  vernachlässigbar sind.



Möchte man umgekehrt die Auswirkung der Störung in nur **einer Komponente  $x_j$  der Eingabe**, aber gleichzeitig auf **alle Komponenten des Ergebnisses** betrachten, so setzen wir in (3.4) alle  $\Delta x_k = 0$  bis auf den Index  $k = j$  und erhalten

$$\Delta y \doteq \underbrace{\begin{bmatrix} \frac{\partial F_1(x)}{\partial x_j} \\ \vdots \\ \frac{\partial F_m(x)}{\partial x_j} \end{bmatrix}}_{\frac{\partial F(x)}{\partial x_j}} \Delta x_j \quad (3.8)$$

und daraus die Abschätzung

$$\|\Delta y\|_2 \leq \left\| \frac{\partial F(x)}{\partial x_j} \right\|_2 |\Delta x_j|. \quad (3.9)$$

Wollen wir nun abschließend gleichzeitig Störungen **aller Komponenten in der Eingabe** und die Auswirkung auf **alle Komponenten im Ergebnis** betrachten, so können wir (3.4) mit Hilfe von Lemma 2.3 abschätzen als

$$\|\Delta y\|_2 \leq \|F'(x)\|_{2 \rightarrow 2} \|\Delta x\|_2. \quad (3.10)$$

Dabei ist  $\|F'(x)\|_{2 \rightarrow 2}$  die durch die 2-Norm induzierte Operatornorm (Definition 2.2) der Jacobimatrix  $F'(x)$ . Die Abschätzung (3.10) besagt also, dass die Größe (Norm)  $\|\Delta y\|_2$  der Störung  $\Delta y$  im Ergebnis nicht größer sein kann als die Operatornorm der Jacobimatrix, multipliziert mit der Größe (Norm)  $\|\Delta x\|_2$  der Störung  $\Delta x$  in der Eingabe. Diese Abschätzung gilt wieder gleichmäßig, unabhängig von der Richtung des Störungsvektors  $\Delta x$ , es handelt sich also um eine *Worst-Case*-Abschätzung. In § 4 werden wir sehen, für welche Richtungen  $\Delta x$  diese Abschätzung scharf ist.

**Definition 3.3** (Absolute Konditionszahl). *Es sei  $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$  an der Stelle  $x$  differenzierbar.*

(i) *Dann heißt die Operatornorm*

$$K(x) := \|F'(x)\|_{2 \rightarrow 2} = \|(K_{ij}(x))\|_{2 \rightarrow 2} \quad (3.11)$$

*die **absolute Konditionszahl** (englisch: absolute condition number) von  $F$  an der Stelle  $x$ . Man spricht auch von der absoluten Konditionszahl der Aufgabe,  $y = F(x)$  auszuwerten.*

(ii) *Die Funktion  $F$  heißt an der Stelle  $x$  **schlecht konditioniert im absoluten Sinne**, wenn  $\|F'(x)\|_{2 \rightarrow 2}$  „sehr groß“ ist, ansonsten **gut konditioniert im absoluten Sinne**.*

**Beachte:** Was „sehr groß“ im Einzelfall bedeutet, wird von den Anforderungen der jeweiligen Aufgabe bestimmt und kann nicht allgemeingültig festgelegt werden. Man beachte dabei auch, dass sich beim Umskalieren der Eingabe oder des Resultats – z. B. durch Wechsel der physikalischen Einheit von m zu mm – die absolute Konditionszahl ändert.

**Beispiel 3.4** (Absolute Konditionierung). *Bei einer gleichmäßig beschleunigten Bewegung mit der Beschleunigung  $a$  und Anfangsgeschwindigkeit  $v_0$  legt ein Körper innerhalb der Zeit  $t$  die Strecke*

$$s(t, v_0, a) = v_0 t + \frac{1}{2} a t^2$$

zurück. Wir betrachten die Strecke  $s$  als Funktion der Eingangsgrößen  $x = (t, v_0, a)$ . Bei einer Beschleunigung von  $a = 5 \text{ m s}^{-2}$ , einer Anfangsgeschwindigkeit von  $v_0 = 0 \text{ m s}^{-1}$  und einer Zeit von  $t = 10 \text{ s}$  beispielsweise ist  $s = 250 \text{ m}$ . Die Jacobimatrix an dieser Stelle ist

$$\left[ \frac{\partial s}{\partial t}(t, v_0, a) \quad \frac{\partial s}{\partial v_0}(t, v_0, a) \quad \frac{\partial s}{\partial a}(t, v_0, a) \right] = \left[ v_0 + a t \quad t \quad \frac{1}{2} t^2 \right] = \left[ 50 \text{ m s}^{-1} \quad 10 \text{ s} \quad 50 \text{ s}^2 \right].$$

Die Daten  $x = (t, v_0, a)$  stammen vielleicht aus einer Messung und sind Unsicherheiten unterworfen. Änderungen einer einzelnen der drei Größen der Größenordnung  $\Delta t = 0.1 \text{ s}$ ,  $\Delta v_0 = 0.5 \text{ m s}^{-1}$  bzw.  $\Delta a = 0.2 \text{ m s}^{-2}$  rufen beispielsweise folgende Änderungen im Ergebnis hervor:

$$\begin{aligned} \Delta s &\doteq \frac{\partial s}{\partial t}(t, v_0, a) \Delta t = 50 \text{ m s}^{-1} \cdot 0.1 \text{ s} = 5 \text{ m}, \\ \Delta s &\doteq \frac{\partial s}{\partial v_0}(t, v_0, a) \Delta v_0 = 10 \text{ s} \cdot 0.5 \text{ m s}^{-1} = 5 \text{ m}, \\ \Delta s &\doteq \frac{\partial s}{\partial a}(t, v_0, a) \Delta a = 50 \text{ s}^2 \cdot 0.2 \text{ m s}^{-2} = 10 \text{ m}. \end{aligned}$$

**Quizfrage:** Wie ändern sich die Zahlenwerte der absoluten partiellen Konditionszahlen, wenn man die zurückgelegte Strecke  $s$  in mm misst statt wie oben in m? **Quizfrage:** Und was ändert sich, wenn man die Anfangsgeschwindigkeit  $v_0$  in  $\text{km h}^{-1}$  misst statt wie oben in  $\text{m s}^{-1}$ ? **Quizfrage:** Ist die Frage nach der absoluten Konditionszahl (3.11) überhaupt physikalisch sinnvoll?

**Beachte:** Im Fall von  $F: \mathbb{R} \rightarrow \mathbb{R}$  gibt es nur eine einzige absolute partielle Konditionszahl  $K_{11} = F'(x)$ , siehe (3.5). Diese ist vorzeichenbehaftet, während die absolute Konditionszahl nach Definition 3.3 immer nicht-negativ ist. Im betrachteten Fall gilt  $\|F'(x)\|_{2 \rightarrow 2} = |F'(x)| = |K_{11}|$ .

**Bemerkung 3.5** (Alternative Definition der absoluten Konditionszahl). In der Literatur findet man auch die folgende Definition<sup>4</sup> der absoluten Konditionszahl einer Funktion  $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ :

$$\limsup_{\Delta x \rightarrow 0} \frac{\|F(x + \Delta x) - F(x)\|_2}{\|\Delta x\|_2} = \lim_{\substack{\varepsilon \searrow 0 \\ \Delta x \neq 0}} \sup_{\|\Delta x\|_2 < \varepsilon} \frac{\|F(x + \Delta x) - F(x)\|_2}{\|\Delta x\|_2}. \quad (3.12)$$

Bzgl. welcher Norm die Größe  $\Delta x$  im Supremum kleiner als  $\varepsilon$  genommen wird, ist hier unerheblich. Die Gleichheit der beiden Ausdrücke in (3.12) gilt nach Definition des Limes superior. Man betrachtet also beliebige Störungen  $\Delta x$  in immer kleiner werdenden Kugeln und bestimmt die größtmögliche Änderung, die solche Störungen im Ergebnis – bezogen auf ihre Größe  $\|\Delta x\|_2$  – hervorrufen können.

Man kann zeigen, dass (3.12) genau dann endlich ist, wenn  $F$  an der Stelle  $x$  differenzierbar ist. In diesem Fall stimmt (3.12) mit (3.11) aus Definition 3.3, also mit  $\|F'(x)\|_{2 \rightarrow 2}$ , überein. Wenn  $F$  an der Stelle  $x$  nicht differenzierbar ist, dann ist (3.12) gleich  $\infty$ .

**Beachte:** Die Wahl der  $2 \rightarrow 2$ -Norm (also der Wahl der 2-Norm in Zähler und Nenner von (3.12)) ist willkürlich. Stattdessen hätten wir z. B. auch die  $\infty \rightarrow \infty$ -Norm nehmen können, wie es im Anschluss bei der relativen Betrachtungsweise gleich der Fall sein wird. Aufgrund der Äquivalenz der Normen ändert sich dadurch die absolute Konditionszahl nur um einen (allerdings dimensionsabhängigen) konstanten Faktor.

<sup>4</sup>vgl. etwa Bornemann, 2018, Kapitel 11.2, S.41

### § 3.2 RELATIVE BETRACHTUNGSWEISE

Neben den bisherigen absoluten Betrachtungsweise kann man die Sensitivitäten auch relativ betrachten. Das bedeutet, dass wir die Störung in der Eingabe  $\Delta x_j$  auf  $x_j$  beziehen und die Änderung in der Ausgabe  $\Delta y_i$  auf  $y_i$ . Aus (3.4) wird also

$$\begin{pmatrix} \frac{\Delta y_1}{y_1} \\ \vdots \\ \frac{\Delta y_m}{y_m} \end{pmatrix} \doteq \begin{bmatrix} \frac{\partial F_1(x)}{\partial x_1} \cdot \frac{x_1}{y_1} & \dots & \frac{\partial F_1(x)}{\partial x_n} \cdot \frac{x_n}{y_1} \\ \vdots & & \vdots \\ \frac{\partial F_m(x)}{\partial x_1} \cdot \frac{x_1}{y_m} & \dots & \frac{\partial F_m(x)}{\partial x_n} \cdot \frac{x_n}{y_m} \end{bmatrix} \begin{pmatrix} \frac{\Delta x_1}{x_1} \\ \vdots \\ \frac{\Delta x_n}{x_n} \end{pmatrix}. \tag{3.13}$$

Die Größe

$$k_{ij}(x) := \frac{\partial F_i(x)}{\partial x_j} \cdot \frac{x_j}{y_i} = \frac{\partial F_i(x)}{\partial x_j} \cdot \frac{x_j}{F_i(x)} = K_{ij}(x) \cdot \frac{x_j}{F_i(x)} \tag{3.14}$$

gibt den Faktor an, mit dem sich relative Änderungen  $\Delta x_j/x_j$  in der  $j$ -ten Komponente der Eingabe relativ auswirken auf die  $i$ -te Komponente des Ergebnisses. Die relative Betrachtungsweise setzt natürlich voraus, dass die Komponenten  $y_i$  des Ergebnisses  $y = F(x)$  ungleich null sind und außerdem die Komponenten  $x_j$  der Eingabe ungleich null sind.

**Definition 3.6** (Relative partielle Konditionszahlen). *Es sei  $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$  an der Stelle  $x$  differenzierbar. Die Zahlen  $k_{ij}(x)$  heißen die **relativen partiellen Konditionszahlen**<sup>5</sup> (englisch: **relative partial condition numbers**) der Funktion  $F$  an der Stelle  $x$ .*

In Analogie zu Definition 3.3 und Bemerkung 3.5 wollen wir noch eine **relative Konditionszahl**  $k(x)$  für die Auswertung von  $F$  an einer Stelle  $x$  angeben. Es stellt sich dabei – um die relative Konditionszahl aus den relativen partiellen Konditionszahlen  $k_{ij}(x)$  berechenbar zu machen – als praktisch heraus, bzgl.  $x$  komponentenweise vorzugehen.

Wir betrachten dazu zunächst eine reellwertige Funktion  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  und untersuchen die größtmögliche relative Änderung, die Störungen  $\Delta x$  im Ergebnis  $f(x)$  hervorrufen können, wenn sie klein werden. Die Größe der Störung  $\Delta x$  messen wir dabei wie gesagt erstens komponentenweise und zweitens relativ zur Größe von  $x$ . Es sei dazu  $D$  die Diagonalmatrix  $D = \text{diag}(x_1, \dots, x_n)$ . Dann hat  $D^{-1}\Delta x$  die Gestalt

$$D^{-1}\Delta x = \begin{pmatrix} \frac{\Delta x_1}{x_1} \\ \vdots \\ \frac{\Delta x_n}{x_n} \end{pmatrix}.$$

Aufgrund der komponentenweise Betrachtung von  $\Delta x$  messen wir die Größe dieses Ausdrucks als

$$\|D^{-1}\Delta x\|_\infty = \max \left\{ \left| \frac{\Delta x_i}{x_i} \right| \mid i = 1, \dots, n \right\}.$$

Die obige Frage nach der größtmöglichen relativen Änderung in  $f(x)$  durch Störungen  $\Delta x$  führt uns nun zu dem Ausdruck

$$\limsup_{\Delta x \rightarrow 0} \frac{|f(x + \Delta x) - f(x)|}{|f(x)| \|D^{-1}\Delta x\|_\infty},$$

<sup>5</sup>In den Wirtschaftswissenschaften werden die relativen partiellen Konditionszahlen  $k_{ij}(x)$  auch als **Elastizitäten** bezeichnet.

den wir jetzt untersuchen. Es gilt

$$\begin{aligned} & \limsup_{\Delta x \rightarrow 0} \frac{|f(x + \Delta x) - f(x)|}{|f(x)| \|D^{-1} \Delta x\|_\infty} \\ &= \lim_{\varepsilon \searrow 0} \sup_{\substack{\|\Delta x\| < \varepsilon \\ \Delta x \neq 0}} \frac{|f(x + \Delta x) - f(x)|}{|f(x)| \|D^{-1} \Delta x\|_\infty} \end{aligned}$$

Bzgl. welcher Norm  $\Delta x$  im Supremum kleiner als  $\varepsilon$  genommen wird, ist hier unerheblich. Weiter:

$$\begin{aligned} &= \lim_{\varepsilon \searrow 0} \sup_{\substack{\|\Delta x\| < \varepsilon \\ \Delta x \neq 0}} \frac{|f'(x) \Delta x|}{|f(x)| \|D^{-1} \Delta x\|_\infty} \\ &= \sup_{\Delta x \neq 0} \frac{|f'(x) \Delta x|}{|f(x)| \|D^{-1} \Delta x\|_\infty} \\ &= \sup_{\Delta y \neq 0} \frac{|f'(x) D \Delta y|}{|f(x)| \|\Delta y\|_\infty} \quad (\text{setze } \Delta y = D^{-1} \Delta x) \\ &= \frac{\|D f'(x)^\top\|_1}{|f(x)|} \quad (\text{wegen (2.5) mit der } \|\cdot\|_{\infty \rightarrow \infty}\text{-Norm des Zeilenvektors } f'(x) D) \\ &= \sum_{j=1}^n \left| \frac{\partial f(x)}{\partial x_j} \right| \frac{|x_j|}{|f(x)|} = \sum_{j=1}^n |k_{1j}|. \end{aligned}$$

Ist nun  $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$  vektorwertig, so betrachten wir die obige Rechnung komponentenweise in  $F_i$ . Wir erhalten ganz analog wie oben

$$\max_{i=1, \dots, m} \limsup_{\Delta x \rightarrow 0} \frac{|F_i(x + \Delta x) - F_i(x)|}{|F_i(x)| \|D^{-1} \Delta x\|_\infty} = \max_{i=1, \dots, m} \sum_{j=1}^n \left| \frac{\partial f(x)}{\partial x_j} \right| \frac{|x_j|}{|f(x)|} = \max_{i=1, \dots, m} \sum_{j=1}^n |k_{ij}(x)| \quad (3.15)$$

als Maß für die relative Änderung im Ergebnis. Das ist gerade die Zeilensummennorm der Matrix, die aus den Einträgen  $k_{ij}(x)$  besteht.

**Quizfrage:** Wie würde dieser Ausdruck aussehen, wenn wir statt der  $\infty$ -Normen überall die 2-Normen verwenden würden?

**Definition 3.7** (Relative Konditionszahl). *Es sei  $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$  an der Stelle  $x$  differenzierbar. Weiter sei  $x \neq 0$  und  $F(x) \neq 0$ .*

(i) Dann heißt der Ausdruck

$$k(x) := \|(k_{ij}(x))\|_{\infty \rightarrow \infty} = \max_{i=1, \dots, m} \sum_{j=1}^n |k_{ij}(x)| \quad (3.16)$$

die **relative Konditionszahl** (englisch: relative condition number) von  $F$  an der Stelle  $x$ .

(ii) Die Funktion  $F$  heißt an der Stelle  $x$  **schlecht konditioniert im relativen Sinne**, wenn (3.16)  $\gg 1$  ist, ansonsten **gut konditioniert im relativen Sinne**.

**Beispiel 3.8** (Relative Konditionierung). Wir greifen das *Beispiel 3.4* nochmals auf, dieses Mal aber in relativer Betrachtungsweise. Die Matrix der relativen partiellen Konditionszahlen der Funktion  $s(t, v_0, a)$  an der Stelle  $x = (t, v_0, a) = (10 \text{ s}, 0 \text{ m s}^{-1}, 5 \text{ m s}^{-2})$  ist gleich

$$\begin{bmatrix} 50 \text{ m s}^{-1} \cdot \frac{10 \text{ s}}{250 \text{ m}} & 10 \text{ s} \cdot \frac{0 \text{ m s}^{-1}}{250 \text{ m}} & 50 \text{ s}^2 \cdot \frac{5 \text{ m s}^{-2}}{250 \text{ m}} \end{bmatrix} = \begin{bmatrix} 2 & 0 & 1 \end{bmatrix}.$$

Diese Zahlen sind wie erwartet einheitenlos.

Eine Änderung von  $\Delta t = 0.1 \text{ s}$  wie in *Beispiel 3.4* entspricht in relativer Betrachtungsweise gegenüber dem ungestörten Wert von  $t = 10 \text{ s}$  einer Änderung von  $\frac{0.1 \text{ s}}{10 \text{ s}} = 0.01$ . Diese erzeugt (bis auf Terme höherer Ordnung) eine relative Änderung von  $2 \cdot 0.01 = 0.02$  im Ergebnis, also absolut gesehen  $0.02 \cdot 250 \text{ m} = 5 \text{ m}$ . Das stimmt (wie erwartet) mit dem aus *Beispiel 3.4* bekannten Ergebnis überein. Analog ergibt die Änderung von  $\Delta a = 0.2 \text{ m s}^{-2}$ , also in relativer Betrachtung von  $\frac{0.2 \text{ m s}^{-2}}{5 \text{ m s}^{-2}} = 0.04$ , eine relative Änderung von  $1 \cdot 0.04 = 0.04$  im Ergebnis, also absolut gesehen  $0.04 \cdot 250 \text{ m} = 10 \text{ m}$ , was wiederum mit dem aus *Beispiel 3.4* bekannten Ergebnis übereinstimmt.

Die relative Konditionszahl bzgl. der Anfangsgeschwindigkeit  $v_0$  ist an der betrachteten Stelle  $x$  ohne Aussagekraft, da  $v_0 = 0 \text{ m s}^{-1}$  ist. Änderungen  $\Delta v_0$  in  $v_0$  lassen sich also nicht relativ ausdrücken, vgl. (3.13).

In relativer Betrachtungsweise lassen sich in jedem Fall die partiellen relativen Konditionszahlen  $k_{ij}(x)$  in physikalisch sinnvoller Weise zur relativen Konditionszahl gemäß (3.16) zusammenfassen, da die  $k_{ij}(x)$  immer einheitenlos sind. **Quizfrage:** Wie ändern sich die Zahlenwerte der relativen partiellen Konditionszahlen, wenn man die zurückgelegte Strecke  $s$  in mm misst statt wie oben in m? **Quizfrage:** Und was ändert sich, wenn man die Anfangsgeschwindigkeit  $v_0$  in  $\text{km h}^{-1}$  misst statt wie oben in  $\text{m s}^{-1}$ ? **Quizfrage:** Wie groß ist die relative Konditionszahl (3.16) in diesem Beispiel?

Im weiteren Verlauf der Vorlesung werden vor allem relative Konditionszahlen betrachtet. Das liegt (neben der Tatsache, dass sie vor allem bei Einbeziehung von Einheiten sinnvoller ist) auch daran, dass später wesentliche Fehlereinflüsse durch die Zahldarstellung im Computer eingebracht werden, die gut im relativen Sinn gemessen werden können, siehe ??.

Wir geben nun die absoluten und relativen partiellen Konditionszahlen der Grundrechenarten und einiger weiterer elementarer Operationen an. Es handelt sich dabei jeweils um Funktionen  $F: \mathbb{R}^2 \rightarrow \mathbb{R}$  bzw.  $F: \mathbb{R} \rightarrow \mathbb{R}$ .

**Lemma 3.9** (Absolute und relative partielle Konditionszahlen für einige Grundoperationen). Die absoluten und relativen Konditionszahlen (im Fall von  $x \neq 0$ ) der folgenden Grundoperationen sind (auf ihren jeweiligen Definitionsbereichen) gegeben durch

Operation	$K_{11}(x)$	$K_{12}(x)$	$k_{11}(x)$	$k_{12}(x)$	$k(x)$
$x_1 + x_2$	1	1	$\frac{x_1}{x_1+x_2}$	$\frac{x_2}{x_1+x_2}$	$\frac{ x_1 + x_2 }{ x_1+x_2 }$
$x_1 - x_2$	1	-1	$\frac{x_1}{x_1-x_2}$	$\frac{-x_2}{x_1-x_2}$	$\frac{ x_1 + x_2 }{ x_1-x_2 }$
$x_1 \cdot x_2$	$x_2$	$x_1$	1	1	2
$\frac{x_1}{x_2}$	$\frac{1}{x_2}$	$-\frac{x_1}{x_2^2}$	1	-1	2
$\frac{1}{x_1}$	$-\frac{1}{x_1^2}$		-1		1
$\sqrt{x_1}$	$\frac{1}{2\sqrt{x_1}}$		$\frac{1}{2}$		$\frac{1}{2}$

Beweis. Nachrechnen □

**Bemerkung 3.10** (Bedeutung von Lemma 3.9).

- (i) Die Ergebnisse von Lemma 3.9 besagen, dass die Multiplikation, Division, Kehrwerte und Quadratwurzeln auf ihrem gesamten Definitionsbereich im relativen Sinne überall gut konditioniert sind.
- (ii) Für die Addition ist die relative Konditionszahl  $k(x) = 1$ , sofern  $x_1$  und  $x_2$  dasselbe Vorzeichen haben. Man bezeichnet diesen Fall auch als **echte Addition**. Dort wo  $x_1$  und  $x_2$  jedoch verschiedene Vorzeichen haben (also  $x_1 + x_2$  einer **echten Subtraktion** zweier positiver Zahlen entspricht), ist die relative Konditionszahl  $k(x) > 1$ . Für  $x_1 \approx -x_2$  kann die relative Konditionszahl beliebig groß werden, siehe Abbildung 3.2. Beispielsweise für festes  $x_1 > 0$  und  $x(\varepsilon) = (x_1, \varepsilon - x_2)^T$  mit  $\varepsilon > 0$  ist  $k(x(\varepsilon)) \in O(\varepsilon^{-1})$ . Ein analoges Resultat gilt natürlich auch für die Subtraktion zweier Zahlen, wobei dann der Fall kritisch ist, bei dem beide dasselbe Vorzeichen haben.

**Beachte:** Die echte Subtraktion zweier fast gleicher Zahlen ist schlecht konditioniert!

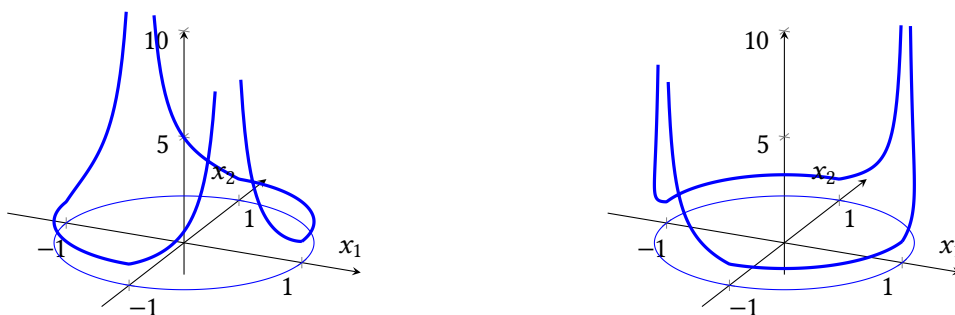


Abbildung 3.2: Relative Konditionszahl  $k(x)$  der Addition (links) und der Subtraktion (rechts) zweier Zahlen  $x_1, x_2$  auf dem Einheitskreis.

Das folgende Beispiel wurde nachträglich eingefügt.

**Beispiel 3.11** (Auslöschung). Die Subtraktion von

$$\begin{aligned}x_1 &= 1.000\,001 \\x_2 &= 1.000\,000\end{aligned}$$

ergibt das Ergebnis  $y = x_1 - x_2 = 10^{-6}$ . Die relative partielle Konditionszahl  $k_{11}(x)$  ist

$$k_{11}(x) = \frac{x_1}{x_1 - x_2} \approx 1.0 \cdot 10^6.$$

Änderungen etwa von  $\Delta x_1 = 10^{-6}$  ergeben eine Änderung von  $\Delta y = 10^{-6}$  im Ergebnis. Relativ gesehen sind diese Änderungen  $\Delta x_1/x_1 \approx 10^{-6}$  und  $\Delta y/y \approx 10^{-6}/10^{-6} = 1$ . Wie erwartet verstärkt sich die relative Störung in der Eingabe um den Faktor  $k_{11}(x) \approx 10^6$ .

Nochmal konkret: Statt  $y = 1 \cdot 10^{-6}$  ergibt sich bei Änderung von  $x_1$  zu  $x_1 + \Delta x_1$  nun das Ergebnis  $y + \Delta y = 2 \cdot 10^{-6}$ . Während sich also bei  $x_1$  die siebte Dezimalstelle ändert, führt dies im Ergebnis  $y$  zu einer Änderung bereits der ersten Dezimalstelle. Wir verlieren damit sechs Dezimalstellen an Genauigkeit! Diesen Effekt nennt man **Auslöschung** (englisch: *cancellation*).

**Beachte:** Gegenüber den Daten verliert man (im schlechtesten Fall) etwa  $\log_{10} k(x)$  Dezimalziffern an Genauigkeit, wobei  $k(x)$  die relative Konditionszahl der betrachteten Operation darstellt.

Der Klarheit halber liefern wir noch eine genaue Definition der Begriffe **echte Addition** und **echte Subtraktion** nach:

**Definition 3.12** (Echte Addition, echte Subtraktion).

- (i) Wir bezeichnen die Operation  $x_1 + x_2$  als **echte Addition** der Zahlen  $x_1$  und  $x_2$ , wenn  $|x_1 + x_2| = |x_1| + |x_2|$  gilt. Andernfalls heißt sie **echte Subtraktion**.
- (ii) Wir bezeichnen die Operation  $x_1 - x_2$  als **echte Addition** der Zahlen  $x_1$  und  $x_2$ , wenn  $|x_1 - x_2| = |x_1| + |x_2|$  gilt. Andernfalls heißt sie **echte Subtraktion**.

Das folgende Beispiel wurde nachträglich eingefügt.

**Beispiel 3.13** (Echte Addition, echte Subtraktion).

- (i) Die folgenden Operationen sind echte Additionen:

$$3 + 5, \quad -3 - 5$$

- (ii) Die folgenden Operationen sind echte Subtraktionen:

$$-3 + 5, \quad 5 - 3$$

### § 3.3 KONDITION LINEARER GLEICHUNGSSYSTEME

Lineare Gleichungssysteme sind eine Grundaufgabe in der angewandten und insbesondere in der numerischen Mathematik. Wir betrachten hier Systeme

$$Ax = b \quad \Leftrightarrow \quad x = A^{-1}b$$

mit invertierbaren Matrizen  $A \in \mathbb{R}^{n \times n}$  und rechten Seiten  $b \in \mathbb{R}^n$ . Zugunsten der für Gleichungssysteme üblichen Notation müssen wir in unserem allgemeinen Setting  $y = F(x)$  also entsprechende Umbenennungen vornehmen.

Wir betrachten zunächst Störungen  $\Delta b$  der rechten Seite  $b$ . Dann übernimmt  $b$  die Rolle der Eingabe  $x$ , und die Lösung  $x$  übernimmt die Rolle des Ergebnisses  $y$ . Statt  $y = F(x)$  haben wir also  $x = A^{-1}b$ . Diese Funktion ist linear in der Eingabe  $b$ , sodass wir an Stelle der Abschätzung (3.3) sogar die Gleichheit

$$\Delta x = A^{-1}\Delta b \tag{3.17}$$

bekommen. Hier könnten wir jetzt komponentenweise wie in § 3.2 vorgehen. Es zeigt sich aber, dass wir das einfachere Resultat erhalten, wenn wir an Stelle der Norm der Vektoren der relativen Änderungen  $(\Delta b_i/b_i)$  und  $(\Delta x_i/x_i)$  nur die relativen Änderungen in den Normen  $\|\Delta b\|/\|b\|$  und  $\|\Delta x\|/\|x\|$  betrachten und dabei jeweils die 2-Norm verwenden. Die zugehörigen Abschätzungen

$$\begin{aligned} \|\Delta x\|_2 &= \|A^{-1}\Delta b\|_2 \leq \|A^{-1}\|_{2 \rightarrow 2} \|\Delta b\|_2, \\ \|b\|_2 &= \|Ax\|_2 \leq \|A\|_{2 \rightarrow 2} \|x\|_2 \end{aligned}$$

lassen sich kombinieren zu

$$\frac{\|\Delta x\|_2}{\|x\|_2} \leq \|A\|_{2 \rightarrow 2} \|A^{-1}\|_{2 \rightarrow 2} \frac{\|\Delta b\|_2}{\|b\|_2}. \tag{3.18}$$

Wir betrachten nun Störungen  $\Delta A$  der Matrix  $A$ . Jetzt übernimmt also  $A$  die Rolle der Eingabe  $x$ . Die differentielle Sensitivitätsanalyse, siehe (3.3), verlangt nun die Bestimmung der Ableitung der Funktion  $A \mapsto F(A) := A^{-1}b$ , also insbesondere die Bestimmung der Ableitung der Matrixinversion.

**Lemma 3.14** (Ableitung der Matrixinversion). *Es sei  $\Phi(A) := A^{-1}$  die Matrixinversions-Funktion. Diese ist für jede reguläre (englisch: **non-singular**) Matrix  $A \in \mathbb{R}^{n \times n}$  in einer ganzen Umgebung von  $A$  definiert und dort stetig differenzierbar.<sup>6</sup> Es gilt die Ableitungsformel*

$$\Phi'(A) \Delta A = -A^{-1} \Delta A A^{-1} \quad \text{für alle regulären } A \in \mathbb{R}^{n \times n}. \tag{3.19}$$

*Beweis.* Wir wollen den Satz über implizite Funktionen anwenden auf die implizite Gleichung

$$A \Phi(A) = \text{Id}$$

für  $\Phi(A)$ . Die Differentiation dieser Beziehung in Richtung einer beliebigen Matrix  $\Delta A \in \mathbb{R}^{n \times n}$  ergibt nach Produktregel

$$\Delta A \Phi(A) + A (\Phi'(A) \Delta A) = 0_{n \times n}. \tag{*}$$

<sup>6</sup> $\Phi$  ist sogar eine  $C^\infty$ -Funktion, was wir hier aber nicht benötigen.



Durch Umstellen erhalten wir zunächst

$$A (\Phi'(A) \Delta A) = -\Delta A \Phi(A)$$

und dann durch Multiplikation von links mit  $A^{-1}$

$$\Phi'(A) \Delta A = -A^{-1} \Delta A \Phi(A) = -A^{-1} \Delta A A^{-1}.$$

Da also die Auflösung der Gleichung (\*) nach der gesuchten Richtungsableitung  $\Phi'(A) \Delta A$  eindeutig möglich war, können wir aus dem Satz über implizite Funktionen schließen, dass die Funktion  $\Phi$  in einer Umgebung von  $A$  tatsächlich definiert und stetig differenzierbar ist und dass die Richtungsableitungen  $\Phi'(A) \Delta A$  durch (3.19) gegeben sind.  $\square$

Mit dem Wissen aus Lemma 3.14 können wir nun die Gleichung (3.3) für unsere Aufgabe mit  $F(A) = A^{-1}b$  auswerten, also

$$\Delta x \doteq F'(A) \Delta A = [\Phi'(A) \Delta A] b = -A^{-1} \Delta A A^{-1} b = -A^{-1} \Delta A x. \quad (3.20)$$

Wir schätzen ähnlich wie oben ab:

$$\|\Delta x\|_2 = \|A^{-1} \Delta A x\|_2 \leq \|A^{-1}\|_{2 \rightarrow 2} \|\Delta A\|_{2 \rightarrow 2} \|x\|_2$$

und erhalten schließlich

$$\frac{\|\Delta x\|_2}{\|x\|_2} \leq \|A^{-1}\|_{2 \rightarrow 2} \|A\|_{2 \rightarrow 2} \frac{\|\Delta A\|_{2 \rightarrow 2}}{\|A\|_{2 \rightarrow 2}}. \quad (3.21)$$

**Definition 3.15** (Konditionszahl einer Matrix). *Es sei  $A \in \mathbb{R}^{n \times n}$ . Wenn  $A$  regulär ist, dann bezeichnen wir die Zahl*

$$\kappa(A) := \|A\|_{2 \rightarrow 2} \|A^{-1}\|_{2 \rightarrow 2} \quad (3.22)$$

als die **Konditionszahl der Matrix**  $A$ . Wenn  $A$  singularär ist, dann setzen wir die Konditionszahl auf  $\kappa(A) := \infty$ .

Nach (3.18) und (3.21) gibt die Konditionszahl eine obere Schranke für den Sensitivitätsfaktor an, mit dem sich eine relative Änderung der 2-Norm der rechten Seite bzw. der  $2 \rightarrow 2$ -Norm der Matrix überträgt auf die relative Änderung der 2-Norm der Lösung. In § 4 kommen wir nochmal darauf und auch auf Möglichkeiten zur Berechnung von  $\kappa(A)$  zurück.

Die Konditionszahl einer Matrix hat aber noch eine andere Interpretation:

**Satz 3.16** (Kahan, 1966, vgl. Bornemann, 2018, Kapitel 11.9). *Es sei  $A \in \mathbb{R}^{n \times n}$  regulär. Dann gilt*

$$\frac{1}{\kappa(A)} = \min \left\{ \frac{\|E\|_{2 \rightarrow 2}}{\|A\|_{2 \rightarrow 2}} \mid E \in \mathbb{R}^{n \times n}, A + E \text{ ist singularär} \right\}. \quad (3.23)$$

Der Kehrwert der Konditionszahl einer Matrix ist also ihr relativer Abstand zur Menge der singularären Matrizen in der  $2 \rightarrow 2$ -Norm.

*Beweis.* Wenn  $A + E$  eine singuläre Matrix ist, dann gibt es ein  $x \neq 0$  mit der Eigenschaft  $(A + E)x = 0$ , also auch  $A^{-1}(A + E)x = x + A^{-1}Ex = 0$ . Es gilt also

$$0 < \|x\|_2 = \|A^{-1}Ex\|_2 \leq \|A^{-1}\|_{2 \rightarrow 2} \|E\|_{2 \rightarrow 2} \|x\|_2 = \kappa(A) \frac{\|E\|_{2 \rightarrow 2}}{\|A\|_{2 \rightarrow 2}} \|x\|_2.$$

Daraus folgt

$$\frac{1}{\kappa(A)} \leq \frac{\|E\|_{2 \rightarrow 2}}{\|A\|_{2 \rightarrow 2}} \frac{\|x\|_2}{\|x\|_2}.$$

Um den Beweis abzuschließen, geben wir jetzt eine spezielle Matrix  $E$  an, für die hier die Gleichheit gilt, also  $\kappa(A) = \frac{\|A\|_{2 \rightarrow 2}}{\|E\|_{2 \rightarrow 2}}$  oder äquivalent  $\|E\|_{2 \rightarrow 2} = 1/\|A^{-1}\|_{2 \rightarrow 2}$ . Dazu sei  $y \in \mathbb{R}^n$  ein Vektor mit  $\|y\|_2 = 1$ , der das Maximum in der Definition der  $2 \rightarrow 2$ -Operatornorm für  $A^{-1}$  realisiert, also

$$\|A^{-1}\|_{2 \rightarrow 2} = \max_{\substack{z \in \mathbb{R}^n \\ \|z\|_2=1}} \|A^{-1}z\|_2 = \|A^{-1}y\|_2,$$

vgl. Lemma 2.3 (iii). Außerdem sei  $x := A^{-1}y \neq 0$ . Wir wählen  $E := -\frac{yx^T}{\|x\|_2^2}$ . Es gilt

$$\|E\|_{2 \rightarrow 2} = \frac{\|x\|_2 \|y\|_2}{\|x\|_2^2} = \frac{1}{\|x\|_2} = \frac{1}{\|A^{-1}y\|_2} = \frac{1}{\|A^{-1}\|_{2 \rightarrow 2}}.$$

Für den Beweis der ersten Gleichheit (Spektralnorm für Rang-1-Matrizen) betrachten wir

$$\begin{aligned} \|E\|_{2 \rightarrow 2} &= \max_{\substack{z \in \mathbb{R}^n \\ \|z\|_2=1}} \|Ez\|_2 \quad (\text{Lemma 2.3}) \\ &= \max_{\substack{z \in \mathbb{R}^n \\ \|z\|_2=1}} \frac{\|yx^Tz\|_2}{\|x\|_2^2} \quad (\text{Definition von } E) \\ &= \max_{\substack{z \in \mathbb{R}^n \\ \|z\|_2=1}} \frac{\|y\|_2}{\|x\|_2^2} |x^Tz| \quad (\text{positive Homogenität der Norm}) \\ &= \frac{\|y\|_2}{\|x\|_2^2} \frac{x^Tx}{\|x\|_2} \quad (z = x/\|x\|_2 \text{ maximiert den Ausdruck, Cauchy-Schwarz-Ungleichung}) \\ &= \frac{\|x\|_2 \|y\|_2}{\|x\|_2^2}. \end{aligned}$$

□

### § 3.4 KONDITION ZUSAMMENGESETZTER FUNKTIONEN

In diesem Abschnitt geht es darum, dass wir eine Funktionsauswertung  $y = F(x)$  für  $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$  in kleinere Einheiten zerlegen, im einfachsten Fall

$$y = F(x) = (H \circ G)(x) = H(G(x))$$

mit Funktionen

$$G: \mathbb{R}^n \rightarrow \mathbb{R}^\ell \quad \text{und} \quad H: \mathbb{R}^\ell \rightarrow \mathbb{R}^m.$$

Die Funktionen  $G$  und  $H$  seien an der Stelle  $x$  bzw. an der Stelle  $G(x)$  differenzierbar. Nach Kettenregel gilt für die differentielle Störungsanalyse, vgl. (3.3),

$$\Delta y \doteq (H \circ G)'(x) \Delta x = H'(G(x)) G'(x) \Delta x. \tag{3.24}$$

Wir können also die absolute Konditionszahl von  $F$  an der Stelle  $x$  mit Hilfe von Lemma 2.3 (v) (Submultiplikativität der Matrixnormen) abschätzen durch

$$K(x) = \|F'(x)\|_{2 \rightarrow 2} \leq \|H'(G(x))\|_{2 \rightarrow 2} \|G'(x)\|_{2 \rightarrow 2}. \tag{3.25}$$

Eine solche Abschätzung wird i. A. nicht scharf sein (**Quizfrage:** Warum?) Dennoch ist die Betrachtung später hilfreich für die Stabilitätsanalyse von Algorithmen.

Wir können eine zu (3.24) analoge Beziehung auch für die relative Sensitivitätsanalyse finden. Wir schreiben dazu (3.24) einmal komponentenweise auf:

$$\begin{pmatrix} \Delta y_1 \\ \vdots \\ \Delta y_m \end{pmatrix} \doteq \begin{bmatrix} \frac{\partial H_1(G(x))}{\partial z_1} & \dots & \frac{\partial H_1(G(x))}{\partial z_\ell} \\ \vdots & & \vdots \\ \frac{\partial H_m(G(x))}{\partial z_1} & \dots & \frac{\partial H_m(G(x))}{\partial z_\ell} \end{bmatrix} \begin{bmatrix} \frac{\partial G_1(x)}{\partial x_1} & \dots & \frac{\partial G_1(x)}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial G_\ell(x)}{\partial x_1} & \dots & \frac{\partial G_\ell(x)}{\partial x_n} \end{bmatrix} \begin{pmatrix} \Delta x_1 \\ \vdots \\ \Delta x_n \end{pmatrix}$$

und ergänzen wie in (3.13) die Terme, die zu einer relativen Betrachtungsweise führen:

$$\begin{pmatrix} \frac{\Delta y_1}{y_1} \\ \vdots \\ \frac{\Delta y_m}{y_m} \end{pmatrix} \doteq \underbrace{\begin{bmatrix} \frac{\partial H_1(G(x))}{\partial z_1} \cdot \frac{z_1}{y_1} & \dots & \frac{\partial H_1(G(x))}{\partial z_\ell} \cdot \frac{z_\ell}{y_1} \\ \vdots & & \vdots \\ \frac{\partial H_m(G(x))}{\partial z_1} \cdot \frac{z_1}{y_m} & \dots & \frac{\partial H_m(G(x))}{\partial z_\ell} \cdot \frac{z_\ell}{y_m} \end{bmatrix}}_{=: (k_{ik}^H(G(x)))} \underbrace{\begin{bmatrix} \frac{\partial G_1(x)}{\partial x_1} \cdot \frac{x_1}{z_1} & \dots & \frac{\partial G_1(x)}{\partial x_n} \cdot \frac{x_n}{z_\ell} \\ \vdots & & \vdots \\ \frac{\partial G_\ell(x)}{\partial x_1} \cdot \frac{x_1}{z_1} & \dots & \frac{\partial G_\ell(x)}{\partial x_n} \cdot \frac{x_n}{z_\ell} \end{bmatrix}}_{=: (k_{kj}^G(x))} \begin{pmatrix} \frac{\Delta x_1}{x_1} \\ \vdots \\ \frac{\Delta x_n}{x_n} \end{pmatrix} \tag{3.26}$$

Auch hier können wir also wieder die Submultiplikativität der Matrixnormen (Lemma 2.3 (v)) verwenden, um abzuschätzen:

$$\left\| \begin{pmatrix} \frac{\Delta y_1}{y_1} \\ \vdots \\ \frac{\Delta y_m}{y_m} \end{pmatrix} \right\| \leq \| (k_{ik}^H(G(x))) \|_{\infty \rightarrow \infty} \| (k_{kj}^G(x)) \|_{\infty \rightarrow \infty} \left\| \begin{pmatrix} \frac{\Delta x_1}{x_1} \\ \vdots \\ \frac{\Delta x_n}{x_n} \end{pmatrix} \right\|. \tag{3.27}$$

Auch diese Abschätzung ist i. A. nicht scharf, aber eine ähnliche Betrachtung wird später bei der Stabilitätsanalyse von Algorithmen noch nützlich sein.

**Bemerkung 3.17.** Wir betonen nochmal zur Verdeutlichung, dass die Kondition eine Eigenschaft der betrachteten Aufgabe ist. Sie hat nichts damit zu tun, mit welchem Verfahren wir einmal diese Aufgabe versuchen werden zu lösen.

Bei einer schlecht konditionierten Aufgabe muss man sich die Frage stellen, ob die Aufgabe überhaupt sachgerecht formuliert ist.

## § 4 SINGULÄRWERTZERLEGUNG VON MATRIZEN

### § 4.1 SPEKTRALZERLEGUNG

Eine Zahl  $\lambda \in \mathbb{C}$  heißt **Eigenwert** (englisch: *eigenvalue*) und ein Vektor  $v \in \mathbb{C}^n \setminus \{0\}$  zugehöriger **Eigenvektor** (englisch: *eigenvector*) einer Matrix  $A \in \mathbb{R}^{n \times n}$  (oder  $A \in \mathbb{C}^{n \times n}$ ), wenn gilt:

$$Av = \lambda v, \quad (4.1)$$

d. h., dass  $v$  unter der Abbildung  $v \mapsto Av$  auf ein Vielfaches von sich selbst abgebildet wird. Zusammen heißt  $(\lambda, v)$  ein **Eigenpaar** von  $A$ . Im Allgemeinen muss man damit rechnen, dass Eigenwerte und Eigenvektoren komplex sind, selbst wenn  $A$  eine reelle Matrix ist.

In der linearen Algebra untersucht man u. a. die Frage, unter welchen Umständen eine Basis aus lauter Eigenvektoren von  $A$  existiert und wann diese als Orthonormalbasis gewählt werden kann.

**Definition 4.1** (Orthogonale Matrix). *Eine Matrix  $V \in \mathbb{R}^{n \times n}$  heißt **orthogonal**, wenn  $V^T V = \text{Id}$  ist.*

Diese Bedingung bedeutet, dass die Spaltenvektoren  $v_i$  von  $V$  paarweise senkrecht (bzgl. des Euklidischen Innenprodukts) aufeinander stehen und außerdem  $\|v_i\|_2 = 1$  ist, denn es gilt

$$(V^T V)_{ij} = v_i^T v_j = \begin{cases} \|v_i\|_2^2 = 1 & \text{im Fall } i = j \\ 0 & \text{im Fall } i \neq j. \end{cases}$$

Eine Matrix  $V \in \mathbb{R}^{n \times n}$  ist also genau dann orthogonal, wenn ihre Spaltenvektoren eine (reelle) **Orthonormalbasis** von  $\mathbb{R}^n$  bilden, also eine Menge linear unabhängiger Vektoren, die paarweise senkrecht stehen und deren 2-Norm gleich eins ist. Für orthogonale Matrizen gilt offenbar  $V^{-1} = V^T$  und daher auch  $VV^T = \text{Id}$ .

Außerdem ist das Produkt orthogonaler Matrizen wieder eine orthogonale Matrix.<sup>7</sup>

**Beachte:** Der Begriff „orthogonale Matrix“ ist für reelle Matrizen reserviert. Das komplexe Analogon heißt **unitäre Matrix**.

Wir interessieren uns hier nur für die Frage, wann eine Matrix  $A \in \mathbb{R}^{n \times n}$  die Eigenschaft besitzt, eine Orthonormalbasis von  $\mathbb{R}^n$  aus lauter reellen Eigenvektoren zusammenstellen zu können. Die Antwort gibt folgender Satz.

**Satz 4.2** (Spektralsatz). *Es sei  $A \in \mathbb{R}^{n \times n}$ . Der Raum  $\mathbb{R}^n$  besitzt genau dann eine reelle Orthonormalbasis aus lauter Eigenvektoren von  $A$ , wenn  $A = A^T$  ist. In diesem Fall besitzt  $A$  die folgende **Spektralzerlegung** (englisch: *spectral decomposition*):*

$$A = V \Lambda V^T, \quad (4.2)$$

<sup>7</sup>D. h., die orthogonalen Matrizen bilden die algebraische Struktur einer Gruppe, die sog. **orthogonale Gruppe** (englisch: *orthogonal group*). Diese ist eine Untergruppe der **allgemeinen linearen Gruppe** (englisch: *general linear group*), die aus den invertierbaren  $n \times n$ -Matrizen besteht.

wobei  $\Lambda \in \mathbb{R}^{n \times n}$  die Diagonalmatrix der Eigenwerte ist und die orthogonale Matrix  $V \in \mathbb{R}^{n \times n}$  aus Eigenvektoren von  $A$  besteht.

Die Gleichung (4.2) können wir auch lesen als

$$A \begin{bmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{bmatrix} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \begin{bmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{bmatrix},$$

also spaltenweise:  $A v_i = \lambda_i v_i$ . (**Quizfrage:** Warum?)

### § 4.2 SINGULÄRWERTZERLEGUNG

Die Spektralzerlegung mit Hilfe der Eigenpaare ist nicht immer die geeignete Darstellung einer Matrix. Insbesondere für rechteckige Matrizen  $A \in \mathbb{R}^{m \times n}$  existieren überhaupt keine Eigenpaare. Daher führen wir jetzt die Singulärwertzerlegung ein, die immer existiert.

**Definition 4.3** (Singulärwertzerlegung). Unter einer **Singulärwertzerlegung** (englisch: **singular value decomposition**, kurz: **SVD**) einer Matrix  $A \in \mathbb{R}^{m \times n}$  versteht man eine Darstellung der Gestalt

$$A = U \Sigma V^T, \tag{4.3}$$

wobei

- $U \in \mathbb{R}^{m \times m}$  und  $V \in \mathbb{R}^{n \times n}$  orthogonale Matrizen sind und
- $\Sigma \in \mathbb{R}^{m \times n}$  eine (rechteckige) Diagonalmatrix ist, deren Diagonaleinträge  $\sigma_1, \dots, \sigma_{\min\{m,n\}} \geq 0$  sind.

Die Spaltenvektoren  $u_i$  von  $U$  heißen **Linkssingulärvektoren** (englisch: **left singular vectors**) von  $A$ . Die Spaltenvektoren  $v_i$  von  $V$  heißen **Rechtssingulärvektoren** (englisch: **right singular vectors**) von  $A$ . Die Zahlen  $\sigma_i$  heißen **Singulärwerte** (englisch: **singular values**) von  $A$ .

Im Falle von  $n > m$  hat die Singulärwertzerlegung beispielsweise die folgende Gestalt:

$$A = \begin{bmatrix} | & & | \\ u_1 & \cdots & u_m \\ | & & | \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_m \end{bmatrix} \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \text{---} & v_1^T & \text{---} \\ \vdots & \vdots & \\ \text{---} & v_n^T & \text{---} \end{bmatrix} \tag{4.4a}$$

und im Falle von  $m > n$

$$A = \begin{bmatrix} | & & & | \\ u_1 & \cdots & \cdots & u_m \\ | & & & | \end{bmatrix} \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_n & \\ \hline 0 & \cdots & 0 & \\ \vdots & & \vdots & \\ 0 & \cdots & 0 & \end{bmatrix} \begin{bmatrix} \text{---} & v_1^\top & \text{---} \\ \vdots & \vdots & \vdots \\ \text{---} & v_n^\top & \text{---} \end{bmatrix}. \quad (4.4b)$$

Jede Matrix  $A \in \mathbb{R}^{m \times n}$  besitzt eine Singulärwertzerlegung. Um das zu begründen (**Quizfrage:** Können Sie die Details ausarbeiten?), kann man Spektralzerlegungen

$$A^\top A = V \Lambda V^\top \quad \text{und} \quad A A^\top = U \tilde{\Lambda} U^\top \quad (4.5)$$

der reellen, symmetrischen, positiv semi-definiten Matrizen  $A^\top A \in \mathbb{R}^{n \times n}$  und  $A A^\top \in \mathbb{R}^{m \times m}$  gemäß [Satz 4.2](#) betrachten. Die Singulärwerte sind gerade die Wurzeln der Diagonaleinträge der kleineren der beiden Diagonalmatrizen  $\Lambda \in \mathbb{R}^{n \times n}$  und  $\tilde{\Lambda} \in \mathbb{R}^{m \times m}$ .

**Bemerkung 4.4** (Eindeutigkeit und Anordnung der Singulärwerte). *Die Singulärwerte sind eindeutig und werden i. d. R. absteigend angeordnet:*

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0 = \sigma_{r+1} = \cdots = \sigma_{\min\{m,n\}}. \quad (4.6)$$

Wir gehen ab jetzt immer von absteigend sortierten Singulärwerten aus.

Die Singulärwertzerlegung enthält sehr viel Information über die Matrix  $A$  und die zu ihr gehörende lineare Abbildung  $x \mapsto Ax$ . Beispielsweise ist die Anzahl  $0 \leq r \leq \min\{m, n\}$  der echt positiven Singulärwerte gerade gleich  $\text{Rang}(A)$ , also gleich der Dimension des Bildraumes  $\text{Bild } A$ . Die „späten“ Rechtssingulärvektoren  $v_{r+1}, \dots, v_n$  bilden eine orthonormale Basis von  $\ker A$ , und die „frühen“ Rechtssingulärvektoren  $v_1, \dots, v_r$  bilden eine orthonormale Basis des orthogonalen Komplements  $(\ker A)^\perp$ . Weiterhin bilden die „frühen“ Linkssingulärvektoren  $u_1, \dots, u_r$  eine Basis von  $\text{Bild } A$ , und die „späten“ Linkssingulärvektoren  $u_{r+1}, \dots, u_m$  bilden eine Basis von  $(\text{Bild } A)^\perp$ . **Quizfrage:** Klar?

**Quizfrage:** Was vermuten Sie: Ist die Singulärwertzerlegung einer Matrix eindeutig?

**Bemerkung 4.5** (Dünne Singulärwertzerlegung). *Im Falle von  $n > m$  sehen wir aus der Darstellung (4.4a), dass die „späten“ Zeilen von  $V^\top$  (Spalten von  $V$ ) mit den Indizes  $m+1, \dots, n$  zur Darstellung von  $A$  nicht benötigt werden. Es gilt also auch*

$$A = \begin{bmatrix} | & & & | \\ u_1 & \cdots & \cdots & u_m \\ | & & & | \end{bmatrix} \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_m & \\ & & & \end{bmatrix} \begin{bmatrix} \text{---} & v_1^\top & \text{---} \\ \vdots & \vdots & \vdots \\ \text{---} & v_m^\top & \text{---} \end{bmatrix}. \quad (4.7a)$$

Im Falle von  $m > n$  werden hingegen die „späten“ Spalten von  $U$  mit den Indizes  $n+1, \dots, m$  nicht benötigt. Es gilt also auch

$$A = \begin{bmatrix} | & & | \\ u_1 & \cdots & u_n \\ | & & | \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix} \begin{bmatrix} - & v_1^T & - \\ & \vdots & \\ - & v_n^T & - \end{bmatrix}. \tag{4.7b}$$

Man spricht dann von einer **dünnen Singulärwertzerlegung** (englisch: *thin SVD*, *economy-size SVD*).

**Beispiel 4.6** (Singulärwertzerlegung). Die Matrix

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 6 & 8 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

besitzt die Singulärwertzerlegung (aus Darstellungsgründen auf zwei Nachkommastellen gerundet)

$$A = \begin{bmatrix} -0.45 & 0.02 & -0.89 \\ -0.89 & 0.04 & 0.45 \\ -0.05 & -1.00 & 0.00 \end{bmatrix} \begin{bmatrix} 12.26 & 0 & 0 & 0 \\ 0 & 1.30 & 0 & 0 \\ 0 & 0 & 0.00 & 0 \end{bmatrix} \begin{bmatrix} -0.19 & -0.37 & -0.55 & -0.73 \\ -0.69 & -0.61 & 0.23 & 0.31 \\ -0.69 & 0.69 & 0.07 & -0.22 \\ -0.13 & 0.13 & -0.80 & 0.57 \end{bmatrix}^T$$

Bild A
(Bild A)<sup>⊥</sup>
(ker A)<sup>⊥</sup>
ker A

Der Rang der Matrix ist also gleich 2. Mit Hilfe von `NUMPY` kann diese Singulärwertzerlegung wie folgt berechnet werden:

```
import numpy as np
A = np.array([[1, 2, 3, 4], [2, 4, 6, 8], [1, 1, 0, 0]])
U, Sigma, V = np.linalg.svd(A)
```

Die Darstellung (4.3) erlaubt es uns, die Matrix  $A$  als eine Summe von  $r$  Rang-1-Matrizen darzustellen:

$$A = U\Sigma V^T = \sum_{i=1}^r \sigma_i \underbrace{u_i v_i^T}_{\text{Rang-1-Matrix}}. \tag{4.8}$$

Damit können wir auch Matrix-Vektor-Produkte genau analysieren:

$$Ax = U\Sigma V^T x = \sum_{i=1}^r \sigma_i u_i (v_i^T x). \tag{4.9}$$

**(Quizfrage:** Wie können wir diese Darstellung interpretieren? Was also „macht“  $v_i^T x$ ? Was bedeutet die Multiplikation mit  $\sigma_i$ ? Und was die Multiplikation mit  $u_i$ ?) Insbesondere folgt aus (4.9) sofort

$$Av_i = \begin{cases} \sigma_i u_i & \text{für } i = 1, \dots, r, \\ 0 & \text{für } i = r + 1, \dots, n. \end{cases}$$

Mit Hilfe der Darstellung (4.9) können wir nun die Frage auflösen, warum die Matrixnorm  $\|A\|_{2 \rightarrow 2}$  durch (2.3b) gegeben ist. Die (quadrierte) Norm des Bildes  $Ax$  lässt sich nun unter Ausnutzung der Orthonormalität der Vektoren  $u_i$  wie folgt bestimmen:

$$\begin{aligned}
 \|Ax\|_2^2 &= \left\| \sum_{i=1}^r \sigma_i u_i (v_i^\top x) \right\|_2^2 = \left( \sum_{i=1}^r \sigma_i u_i (v_i^\top x), \sum_{j=1}^r u_j \sigma_j (v_j^\top x) \right) \quad (\text{wegen (4.9)}) \\
 &= \sum_{i=1}^r \sum_{j=1}^r \sigma_i \sigma_j u_i^\top u_j (v_i^\top x) (v_j^\top x) \\
 &= \sum_{i=1}^r \sigma_i^2 \|u_i\|_2^2 (v_i^\top x)^2 \quad (\text{wegen der Orthogonalität von } u_i \text{ und } u_j \text{ für } i \neq j) \\
 &= \sum_{i=1}^r \sigma_i^2 (v_i^\top x)^2 \quad (\text{wegen } \|u_i\|_2 = 1) \\
 &\leq \sigma_1^2 \sum_{i=1}^r (v_i^\top x)^2 \\
 &= \sigma_1^2 \|x\|_2^2.
 \end{aligned}$$

Die obige Abschätzung ist mit Gleichheit erfüllt, wenn  $x$  ein Vielfaches von  $v_1$  ist. In dem Fall gilt  $\|Ax\|_2 = \sigma_1 \|x\|_2$ . Da der größte Singulärwert  $\sigma_1$  gleichzeitig die Wurzel des größten Eigenwertes von  $A^\top A$  ist, haben wir damit (2.3b) gezeigt und außerdem:

**Satz 4.7** (Die  $\|\cdot\|_{2 \rightarrow 2}$ -Norm einer Matrix ist ihr größter Singulärwert).

(i) Für jede Matrix  $A \in \mathbb{R}^{m \times n}$  gilt:

$$\|A\|_{2 \rightarrow 2} = \sigma_1, \quad (4.10)$$

also ist die Spektralnorm gerade der größte Singulärwert der Matrix.

(ii) Für *symmetrische* Matrizen  $A \in \mathbb{R}^{n \times n}$  gilt gleichzeitig:

$$\|A\|_{2 \rightarrow 2} = \max\{|\lambda| \mid \lambda \text{ ist ein Eigenwert von } A\}, \quad (4.11)$$

also ist die Spektralnorm gerade der Betrag des betragsgrößten Eigenwerts der Matrix.

*Beweis.* Aussage (i) haben wir oben bereits gezeigt. Es gilt nach Konstruktion der Singulärwertzerlegung

$$\sigma_1^2 = \max\{\lambda \mid \lambda \text{ ist ein Eigenwert von } A^\top A\}.$$

Wenn  $A$  symmetrisch ist, dann existiert die Spektralzerlegung  $A = V\Lambda V^\top$ , siehe (4.2) mit einer orthogonalen Matrix  $V$ . Es gilt also  $A^\top A = V\Lambda V^\top V\Lambda V^\top = V\Lambda^2 V^\top$ . Die Eigenwerte von  $A^\top A$  sind demnach – wenn  $A$  symmetrisch ist – gerade die Quadrate der Eigenwerte von  $A$ . Wir haben also

$$\begin{aligned}
 \sigma_1^2 &= \max\{\lambda \mid \lambda \text{ ist ein Eigenwert von } A^\top A\} \\
 &= \max\{\lambda^2 \mid \lambda \text{ ist ein Eigenwert von } A\}.
 \end{aligned}$$

Durch Wurzelziehen folgt Aussage (ii). □



Im Beweis von [Satz 3.16](#) haben wir die Norm  $\|y x^T\|_{2 \rightarrow 2}$  der Rang-1-Matrix  $y x^T$  benötigt. **Quizfrage:** Können Sie diese Norm mit dem jetzigen Wissen über die Singulärwertzerlegung bestimmen?

Falls  $A \in \mathbb{R}^{n \times n}$  invertierbar ist (**Quizfrage:** Wie drückt man das mit Hilfe der Singulärwerte von  $A$  aus?), dann erlaubt es uns die Singulärwertzerlegung  $A = U \Sigma V^T$ , auch direkt eine Singulärwertzerlegung von  $A^{-1}$  anzugeben:

$$A^{-1} = V \Sigma^{-1} U^T. \quad (4.12)$$

Hieraus sieht man sofort, dass  $1/\sigma_n$  der maximale Singulärwert von  $A^{-1}$  ist.

Weiter ergibt sich aus diesen Überlegungen, dass die Konditionszahl (3.22) einer regulären Matrix  $A \in \mathbb{R}^{n \times n}$  die Beziehung

$$\kappa(A) = \|A\|_{2 \rightarrow 2} \|A^{-1}\|_{2 \rightarrow 2} = \frac{\sigma_1}{\sigma_n} \quad (4.13)$$

erfüllt.

**Quizfrage:** Was ist die Konditionszahl einer orthogonalen Matrix? **Quizfrage:** Welche quadratischen (notwendigerweise invertierbaren) Matrizen haben Konditionszahl gleich eins? **Quizfrage:** Wie sieht die Singulärwertzerlegung einer Diagonalmatrix aus?

Auf Basis der SVD kann man sehr schön sehen, warum die Abschätzung

$$\|BA\|_{2 \rightarrow 2} \leq \|B\|_{2 \rightarrow 2} \|A\|_{2 \rightarrow 2}$$

aus [Lemma 2.3 \(v\)](#) (Submultiplikativität) insbesondere für die Spektralnorm i. A. nicht scharf ist. Wir betrachten dazu Singulärwertzerlegungen beider Matrizen:

$$B = U^{(2)} \Sigma^{(2)} V^{(2)T} \quad \text{und} \quad A = U^{(1)} \Sigma^{(1)} V^{(1)T}.$$

Es gilt also  $\|A\|_{2 \rightarrow 2} = \sigma_1^{(1)}$  und  $\|B\|_{2 \rightarrow 2} = \sigma_1^{(2)}$ . Es sei nun  $x$  ein Rechtssingulärvektor von  $A$  zum größten Singulärwert  $\sigma_1^{(1)}$  mit  $\|x\|_2 = 1$ . Dann gilt  $Ax = \sigma_1^{(1)} u_1^{(1)}$  mit  $\|Ax\|_2 = \sigma_1^{(1)}$ . Falls nun  $u^{(1)}$  ein Rechtssingulärvektor von  $B$  zum größten Singulärwert  $\sigma_1^{(2)}$  ist, so gilt weiter

$$BAx = \sigma_1^{(1)} B u^{(1)} = \sigma_1^{(1)} \sigma_1^{(2)} u^{(2)} \quad \text{mit} \quad \|BAx\|_2 = \sigma_1^{(1)} \sigma_1^{(2)}.$$

In diesem Fall haben wir also tatsächlich einen Vektor gefunden, für den  $\|BAx\|_2 = \|B\|_{2 \rightarrow 2} \|A\|_{2 \rightarrow 2} \|x\|_2$  gilt und damit  $\|BA\|_{2 \rightarrow 2} = \|B\|_{2 \rightarrow 2} \|A\|_{2 \rightarrow 2}$ . Im anderen Fall besitzt  $u^{(1)}$  auch Anteile von Rechtssingulärvektoren der Matrix  $B$ , die nicht zum größten Singulärwert von  $B$  gehören. Das impliziert  $\|B u^{(1)}\|_2 < \sigma_1^{(2)} \|u^{(1)}\|_2$ .

Mit diesen Überlegungen kann man folgenden Satz beweisen:

**Satz 4.8** (Submultiplikativität der  $2 \rightarrow 2$ -Norm). *Es seien  $A \in \mathbb{R}^{m \times n}$  und  $B \in \mathbb{R}^{\ell \times m}$ . Dann sind die folgenden Aussagen äquivalent:*

- (i) *Es gilt  $\|BA\|_{2 \rightarrow 2} = \|B\|_{2 \rightarrow 2} \|A\|_{2 \rightarrow 2}$ .*

- (ii) Es gibt einen Vektor  $x \in \mathbb{R}^n$ ,  $x \neq 0$ , mit der Eigenschaft  $\|Ax\|_2 = \sigma_1^{(1)} \|x\|_2$  und  $\|BAx\|_2 = \sigma_1^{(2)} \|Ax\|_2$ .
- (iii) Der Unterraum von  $\mathbb{R}^m$ , der durch die Linkssingulärvektoren von  $A$  aufgespannt wird, die zum größten Singulärwert  $\sigma_1^{(1)}$  gehören, schneidet den Unterraum von  $\mathbb{R}^m$ , der durch die Rechtssingulärvektoren von  $B$  aufgespannt wird, die zum größten Singulärwert  $\sigma_1^{(2)}$  gehören, in mehr als nur dem Nullvektor.

**Folgerung 4.9** (Submultiplikativität der  $2 \rightarrow 2$ -Norm). Insbesondere in folgenden Fällen sind die Aussagen (i) bis (iii) wahr:

- Die „mittlere“ Dimension ist  $m = 1$ .
- Mindestens eine der Matrizen  $A$  oder  $B$  ist die Nullmatrix.
- Mindestens eine der Matrizen  $A$  oder  $B$  ist eine orthogonale Matrix.

**Quizfrage:** Beweis?

Eine weitere wichtige Folgerung daraus ist:

**Folgerung 4.10** (Multiplikation mit einer orthogonalen Matrix ändert die Konditionszahl nicht). Es sei  $A \in \mathbb{R}^{n \times n}$  beliebig und  $V \in \mathbb{R}^{n \times n}$  orthogonal. Dann gilt  $\|AV\|_{2 \rightarrow 2} = \|VA\|_{2 \rightarrow 2} = \|A\|_{2 \rightarrow 2}$ .

Beweis. □

Schließlich geben wir noch eine Eigenschaft der Singulärwertzerlegung an.

**Satz 4.11** (Die SVD löst die Aufgabe der Best-Approximation mit konstantem Rang, vgl. Golub, van Loan, 1983, Theorem 2.5.3). Es sei  $A \in \mathbb{R}^{m \times n}$  eine Matrix mit Singulärwertzerlegung  $A = U\Sigma V^T$ . Weiter sei  $r = \text{Rang}(A)$  und  $0 \leq k \leq r$ . Dann löst die Matrix

$$A_k = \sum_{i=1}^k \sigma_i u_i v_i^T = \begin{bmatrix} | & & | \\ u_1 & \cdots & u_k \\ | & & | \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{bmatrix} \begin{bmatrix} - & v_1^T & - \\ & \vdots & \\ - & v_k^T & - \end{bmatrix} \in \mathbb{R}^{m \times n} \quad (4.14)$$

die folgende Optimierungsaufgabe:

$$\begin{aligned} & \text{Minimiere} && \|A - B\|_{2 \rightarrow 2}, && B \in \mathbb{R}^{m \times n} \\ & \text{unter der Bedingung} && \text{Rang}(B) = k. \end{aligned} \quad (4.15)$$

Der Optimalwert ist  $\|A - A_k\|_2 = \sigma_{k+1}$ .

*Beweis.* Wir überzeugen uns zunächst davon, dass die Matrix  $A_k$  den Rang  $k$  hat und dadurch als Kandidat in der Optimierungsaufgabe (4.15) zur Verfügung steht. Unter Ausnutzung der Orthogonalität von  $U$  und  $V$  ergibt sich aus (4.14)

$$U^T A_k V = \left[ \begin{array}{ccc|ccc} \sigma_1 & & & 0 & \cdots & 0 \\ & \ddots & & \vdots & & \vdots \\ & & \sigma_k & 0 & \cdots & 0 \\ \hline 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{array} \right] \in \mathbb{R}^{m \times n}.$$

Da die Multiplikation von  $A_k$  mit invertierbaren Matrizen den Rang nicht verändert und  $\sigma_1 \geq \cdots \geq \sigma_k > 0$  gilt, gilt tatsächlich  $\text{Rang}(A_k) = k$ . **Quizfrage:** Warum lässt die Multiplikation einer Matrix mit einer invertierbaren Matrix den Rang unverändert? **Quizfrage:** Warum ist  $\sigma_k > 0$ ?

Weiter gilt (Illustration für den Fall  $m > n$ )

$$U^T (A - A_k) V = \left[ \begin{array}{ccc|ccc} 0 & & & 0 & \cdots & 0 \\ & \ddots & & \vdots & & \vdots \\ & & 0 & 0 & \cdots & 0 \\ \hline 0 & \cdots & 0 & \sigma_{k+1} & & \\ \vdots & & \vdots & & \ddots & \\ 0 & \cdots & 0 & & & \sigma_p \\ \hline 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{array} \right] \in \mathbb{R}^{m \times n}$$

mit  $p = \min\{m, n\}$ . Aus **Folgerung 4.10** folgt  $\|A - A_k\|_{2 \rightarrow 2} = \|U^T (A - A_k) V\|_{2 \rightarrow 2}$ , und letztere Matrix hat offenbar  $\sigma_{k+1}$  als größten Singulärwert, also gilt  $\|A - A_k\|_{2 \rightarrow 2} = \sigma_{k+1}$ .

Wir zeigen jetzt noch, dass für jede beliebige Matrix  $B \in \mathbb{R}^{m \times n}$  mit  $\text{Rang}(B) = k$  gilt:  $\|A - B\|_{2 \rightarrow 2} \geq \sigma_{k+1}$ . Damit ist dann die behauptete Optimalität von  $A_k$  gezeigt und der Beweis abgeschlossen. Die Dimensionsformel  $k = \dim \text{Bild } B = n - \dim \ker B$  zeigt, dass der Kern von  $B$  die Dimension  $n - k$  besitzt. Wir können also eine Menge orthonormaler Vektoren  $x_1, \dots, x_{n-k}$  finden, sodass

$$\ker B = \text{span}\{x_1, \dots, x_{n-k}\}$$

gilt. Aus Dimensionsgründen gilt

$$\text{span}\{x_1, \dots, x_{n-k}\} \cap \text{span}\{v_1, \dots, v_{k+1}\} \neq \{0\}.$$

Es sei nun  $z$  ein Vektor in diesem Durchschnitt mit  $\|z\|_2 = 1$ . Wegen  $Bz = 0$  und

$$Az = \sum_{i=1}^r \sigma_i u_i (v_i^T z) = \sum_{i=1}^{k+1} \sigma_i u_i (v_i^T z)$$

erhalten wir

$$\begin{aligned}
 \|A - B\|_{2 \rightarrow 2}^2 &\geq \|(A - B)z\|_{2 \rightarrow 2}^2 \quad \text{siehe Lemma 2.3} \\
 &= \|Az\|_{2 \rightarrow 2}^2 \quad \text{wegen } Bz = 0 \\
 &= \left\| \sum_{i=1}^{k+1} \sigma_i u_i (v_i^\top z) \right\|_2^2 \\
 &= \sum_{i=1}^{k+1} \sigma_i^2 (v_i^\top z)^2 \quad \text{wegen der Orthonormalität der Vektoren } u_i \\
 &\geq \sigma_{k+1}^2 \sum_{i=1}^{k+1} (v_i^\top z)^2 \quad \text{wegen } \sigma_1 \geq \dots \geq \sigma_{k+1} \\
 &= \sigma_{k+1}^2 \quad \text{wegen } 1 = \|z\|_2^2 = \left\| \sum_{i=1}^{k+1} (v_i^\top z) v_i \right\|_2^2 = \sum_{i=1}^{k+1} (v_i^\top z)^2.
 \end{aligned}$$

□

Dieses Resultat zeigt, dass die führenden Terme in einer Singulärwertzerlegung die beste Approximation der gegebenen Matrix im Sinne der  $2 \rightarrow 2$ -Norm unter allen Matrizen mit gegebenem **Rang** <sup>GM</sup> darstellen. Der dabei gemachte Approximationsfehler ist gerade der erste Singulärwert, der nicht berücksichtigt wird.

**Folgerung 4.12** (Abstand zu Matrizen mit Rangdefekt). *Es sei  $A \in \mathbb{R}^{m \times n}$  eine Matrix und  $p = \min\{m, n\}$ . Dann gilt*

$$\min\{\|A - B\|_{2 \rightarrow 2} \mid B \in \mathbb{R}^{m \times n}, \text{Rang}(B) < p\} = \sigma_p > 0. \quad (4.16)$$

*Beweis.*

□

Der kleinste Singulärwert gibt also den Abstand einer Matrix  $A$  zur Menge der Matrizen gleicher Dimensionen an, die nicht vollen Rang haben.

Ende der Woche 3

# Kapitel 2 Zahldarstellung und Rechnerarithmetik

## § 5 ZAHLDARSTELLUNG

Computer rechnen mit bestimmten Zahldarstellungen. Man unterscheidet zwischen **Ganzzahlen** (englisch: *integer numbers*), **Festkommazahlen** (englisch: *fixed-point numbers*) und **Fließkommazahlen** (englisch: *floating-point numbers*). Da sich die numerische Mathematik mit Aufgaben der „kontinuierlichen Mathematik“ beschäftigt, sind Ganzzahlen i. d. R. nur als Zähler usw. interessant. Wir behandeln deshalb jetzt hier vorrangig Fließkommazahlen.

**Definition 5.1** (Fließkommazahl, vgl. Bornemann, 2018, Kapitel 12.1). Eine **Fließkommazahl** zur **Basis** (englisch: *base, radix*)  $\beta \in \{2, 3, \dots\} \subset \mathbb{N}$  ist eine Zahl der Form

$$m \beta^e. \quad (5.1)$$

Dabei heißt

$$m = \pm(m_0 \beta^0 + m_{-1} \beta^{-1} + \dots + m_{1-r} \beta^{1-r}) \quad \text{mit Ziffern } m_0, \dots, m_{1-r} \in \{0, 1, \dots, \beta - 1\} \quad (5.2)$$

die (rationale) **Mantisse** (englisch: *significand, mantissa*) der Zahl und

$$e = \pm(e_{s-1} \beta^{s-1} + e_{s-2} \beta^{s-2} + \dots + e_1 \beta^1 + e_0 \beta^0) \quad \text{mit Ziffern } e_{s-1}, \dots, e_0 \in \{0, 1, \dots, \beta - 1\} \quad (5.3)$$

ihr (ganzzahliger) **Exponent** (englisch: *exponent*). Es gilt die Konvention, dass die führende Stelle der Mantisse  $m_0 \neq 0$  ist.<sup>1</sup> Die einzige Ausnahme ist die Zahl 0, die durch die Mantisse  $m = 0$  und beliebigen Exponenten  $e$  dargestellt wird.

Wir bezeichnen die Menge der darstellbaren Fließkommazahlen (zu gegebener Basis, Mantissenlänge und Wertebereich für den Exponenten) mit  $\mathbb{F}$ . Man spricht manchmal auch von einem **Fließkommagitter** (englisch: *floating point grid*).

**Beachte:** Die Fließkommazahlen  $\mathbb{F}$  sind eine Teilmenge der rationalen Zahlen  $\mathbb{Q}$ . (**Quizfrage:** Klar?)

Mit Hilfe von Fließkommazahlen lassen sich Zahlen sehr unterschiedlicher Größenordnung darstellen, etwa (zur Basis  $\beta = 10$ )

- die Ruhemasse eines Elektrons, etwa  $9.109\,383 \cdot 10^{-31}$  kg,

<sup>1</sup>Zur Verdeutlichung kann man dann auch von einer **normalisierten Fließkommazahl** sprechen.

- die Lichtgeschwindigkeit im Vakuum, etwa  $2.997\,924\,58 \cdot 10^8 \text{ m s}^{-1}$ ,
- die Avogadro-Konstante, etwa  $6.022\,140\,76 \cdot 10^{23} \text{ mol}$ .

Wird eine Zahl nicht zur Basis  $\beta = 10$  notiert, so deutet man die Basis häufig durch einen tiefgestellten Index an. Beispielsweise ist

$$1001_2 = 9 \quad \text{und} \quad 1001.11_2 = 9 + \frac{1}{2} + \frac{1}{4} = 9.75.$$

Die Anzahl  $r$  der Stellen der Mantisse sowie die Anzahl  $s$  der Stellen des Exponenten in einem System von Fließkommazahlen sind in der Regel fest. Die Menge der damit darstellbaren Zahlen ist somit endlich.

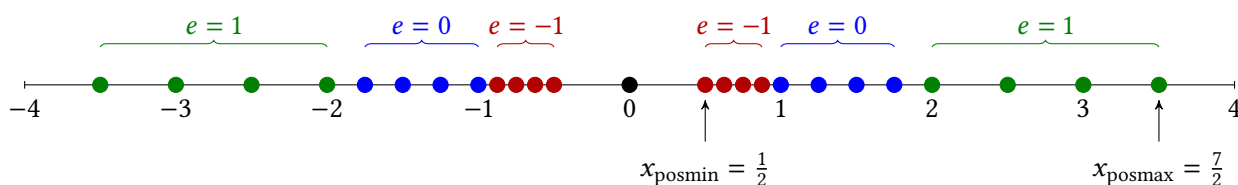
**Beispiel 5.2** (Darstellbare Fließkommazahlen). *Im Fall der Basis  $\beta = 2$ , Mantissenlänge  $r = 3$  und Exponentenlänge  $s = 1$  sind die folgenden positiven Mantissen möglich:*

$$\begin{aligned} m = 1.00_2 &= \frac{4}{4}, \\ m = 1.01_2 &= \frac{5}{4}, \\ m = 1.10_2 &= \frac{6}{4}, \\ m = 1.11_2 &= \frac{7}{4}. \end{aligned}$$

Durch Kombination mit den zur Verfügung stehenden Exponenten  $e \in \{-1, 0, 1\}$ , also den Faktoren  $2^{-1}$ ,  $2^0$  und  $2^1$ , sind also die positiven Zahlen

$$\left\{ \frac{4}{8}, \frac{5}{8}, \frac{6}{8}, \frac{7}{8}, \frac{4}{4}, \frac{5}{4}, \frac{6}{4}, \frac{7}{4}, \frac{4}{2}, \frac{5}{2}, \frac{6}{2}, \frac{7}{2} \right\}$$

darstellbar, dazu noch ihre negativen Gegenstücke sowie schließlich die Zahl 0. (**Quizfrage:** Ist die Darstellung einer Fließkommazahl immer eindeutig?)



Wie das **Beispiel 5.2** zeigt, sind Fließkommazahlen nicht gleichmäßig verteilt. Weiterhin gibt es eine kleinste und eine größte positive Zahl in  $\mathbb{F}$ . Die kleinste positive Zahl  $x_{\text{posmin}}$  wird dargestellt durch die Mantisse  $m = (1.0 \cdots 0)_\beta = 1$  und den Exponenten

$$e = -((\beta - 1)(\beta - 1) \cdots (\beta - 1))_\beta = -(\beta - 1) \sum_{k=0}^{s-1} \beta^k = -(\beta - 1) \frac{\beta^s - 1}{\beta - 1} = 1 - \beta^s.$$

Es ergibt sich also

$$x_{\text{posmin}} = \beta^{1-\beta^s} \quad (5.4)$$

Die größte positive Zahl  $x_{\text{posmax}}$  dagegen wird dargestellt durch die Mantisse

$$m = ((\beta-1).(\beta-1) \cdots (\beta-1))_\beta = (\beta-1) \sum_{k=1-r}^0 \beta^k = (\beta-1)\beta^{1-r} \sum_{k=0}^{r-1} \beta^k = (\beta-1)\beta^{1-r} \frac{\beta^r - 1}{\beta - 1} = \beta^{1-r}(\beta^r - 1)$$

und den Exponenten

$$e = ((\beta-1)(\beta-1) \cdots (\beta-1))_\beta = (\beta-1) \sum_{k=0}^{s-1} \beta^k = (\beta-1) \frac{\beta^s - 1}{\beta - 1} = \beta^s - 1.$$

Daraus ergibt sich

$$x_{\text{posmax}} = \beta^{1-r}(\beta^r - 1)\beta^{\beta^s - 1} = (\beta - \beta^{1-r})\beta^{\beta^s - 1}. \quad (5.5)$$

Im [Beispiel 5.2](#) etwa mit  $\beta = 2$ ,  $r = 3$  und  $s = 1$  gilt

$$\begin{aligned} x_{\text{posmin}} &= \beta^{1-\beta^s} = 2^{1-2} = \frac{1}{2}, \\ x_{\text{posmax}} &= (\beta - \beta^{1-r})\beta^{\beta^s - 1} = (2 - 2^{1-3})2^{2^1 - 1} = \frac{7}{4} \cdot 2^1 = \frac{7}{2}. \end{aligned}$$

Das stimmt mit unseren Beobachtungen in [Beispiel 5.2](#) überein.

**Definition 5.3** (Zulässiger Bereich). Die Menge

$$\mathbb{D} := [-x_{\text{posmax}}, -x_{\text{posmin}}] \cup \{0\} \cup [x_{\text{posmin}}, x_{\text{posmax}}] \subset \mathbb{R} \quad (5.6)$$

heißt der **zulässige Bereich** eines Fließkommagitters  $\mathbb{F}$ .

**Quizfrage:** Was ist die Bedeutung des zulässigen Bereichs?

**Definition 5.4** (IEEE<sup>2</sup>-Standard 754). Der 1985 eingeführte *IEEE-Standard 754* ist auch heute noch die Grundlage der Darstellung von Fließkommazahlen auf sehr vielen Computerarchitekturen. In diesem Standard spielt das Format **double precision** (auch: **binary64**) die wichtigste Rolle. Es verwendet die Basis  $\beta = 2$ , Mantissenlänge  $r = 53$  und ganzzahlige Exponenten  $e \in [-1022, 1023]$ . Daraus können wir ähnlich wie oben die kleinste und größte positive darstellbare (normalisierte) Fließkommazahl herleiten:

$$\begin{aligned} x_{\text{posmin}} &= 1 \cdot 2^{-1022} = 2^{-1022} \approx 2.2 \cdot 10^{-308}, \\ x_{\text{posmax}} &= 2^{1-53}(2^{53} - 1) \cdot 2^{1023} = (2 - 2^{-52}) \cdot 2^{1023} \approx 1.8 \cdot 10^{308}. \end{aligned}$$

An Stelle eines vorzeichenbehafteten Exponenten  $e$  wird dieser als  $e = c - 1023$  interpretiert, wobei die Zahl  $c \in [0, 2047]$  mit 11 Bits gespeichert wird. Da die Extremfälle  $c = 0$  und  $c = 2047$  besondere Bedeutung haben, ergeben sich die nutzbaren Werte  $e \in [-1022, 1023]$  für den Exponenten. Da die Mantisse normalisiert ist, ist die führende Ziffer für Fließkommazahlen  $x \neq 0$  immer  $m_0 = 1$  und muss nicht gespeichert werden. Zusammen mit dem Vorzeichen der Mantisse benötigt eine Zahl nach dem double precision-Standard daher  $1 + 52 + 11 = 64$  Bits oder 8 Bytes an Speicher.

Zusätzlich definiert der IEEE-Standard die erweiterten Fließkommazahlen  $+\infty$ ,  $-\infty$  und NaN („not a number“), die mit Hilfe von  $c = 2047$  kodiert werden.

<sup>2</sup>Institute of Electrical and Electronics Engineers

**Definition 5.5** (Rundung). Es sei  $\mathbb{F}$  eine Menge von Fließkommazahlen. Eine Funktion  $\text{rd}: \mathbb{R} \cup \{\pm\infty\} \rightarrow \mathbb{F} \cup \{\pm\infty\}$  heißt **Rundung**, wenn sie folgende Eigenschaften erfüllt:

(i) Sie ist monoton, d. h.,

$$\text{rd}(x) \leq \text{rd}(y) \quad \text{für } x, y \in \mathbb{R} \cup \{\pm\infty\} \text{ mit } x \leq y.$$

(ii) Sie ist idempotent, d. h.,

$$\text{rd}(x) = x \quad \text{für } x \in \mathbb{F} \cup \{\pm\infty\}.$$

**Quizfrage:** Wofür könnten die beiden Eigenschaften der Rundung nützlich sein?

Wir betrachten hier nur die gängigste Strategie des IEEE-Standards, **die Rundung zur nächstgelegenen Fließkommazahl** (englisch: *round to nearest*). Diese Funktion bildet Zahlen im Bereich  $[-x_{\text{posmax}}, x_{\text{posmax}}]$  auf die jeweils nächstgelegene Zahl in  $\mathbb{F}$  ab, die bis auf endlich viele Ausnahmewerte für  $x$  eindeutig ist. Die Ausnahmewerte sind gerade diejenigen Zahlen  $x$ , die genau in der Mitte zwischen zwei benachbarten Zahlen in  $\mathbb{F}$  liegen. Diese Grenzfälle werden bei der Unterstrategie *round to nearest, ties to even* so gerundet, dass das niederwertigste Bit der Mantisse = 0 wird. Weiterhin werden Zahlen  $|x| > x_{\text{posmax}} + c$  auf  $+\infty$  bzw.  $-\infty$  gerundet, wobei  $c > 0$  eine bestimmte Konstante ist.

Wenn wir in Zukunft von Rundung sprechen, so ist immer diese Strategie gemeint.

**Lemma 5.6** (Relativer Fehler bei der Rundung). Der relative Fehler bei der Rundung in einem System von Fließkommazahlen  $\mathbb{F}$  mit Mantissenlänge  $r$  erfüllt die Abschätzung

$$\frac{|x - \text{rd}(x)|}{|x|} \leq \epsilon_{\text{mach}} := \frac{1}{2}\beta^{1-r} \quad (5.7)$$

für alle Zahlen  $x \neq 0$ , die im zulässigen Bereich  $\mathbb{D}$  liegen. Dabei ist  $\epsilon_{\text{mach}}$  die sogenannte **Maschinengenauigkeit** (englisch: *machine precision, unit roundoff*).

*Beweis.* Wir müssen nur den Fall  $x > 0$  betrachten, da der Beweis im Fall  $x < 0$  analog läuft. Aufgrund der Annahme an  $x$  existieren zwei benachbarte Zahlen  $\underline{x}$  und  $\bar{x}$  in  $\mathbb{F}$  mit der Eigenschaft  $0 < \underline{x} \leq x < \bar{x}$ . Nach Definition der Rundung gilt

$$|x - \text{rd}(x)| \leq \frac{1}{2}(\bar{x} - \underline{x}).$$

Aus der Darstellung (4.2)–(5.3) sieht man, dass sich  $\underline{x}$  und  $\bar{x}$  um die Zahl

$$(0.00 \dots 01)_\beta \cdot \beta^e = \beta^{1-r} \cdot \beta^e$$

unterscheiden, wobei  $e$  der Exponent von  $\underline{x}$  ist. (**Quizfrage:** Warum stimmt das auch dann, wenn  $\underline{x} = (\beta.\beta\beta \dots \beta\beta)_\beta \cdot \beta^e$  ist?)

Es gilt also

$$|x - \text{rd}(x)| \leq \frac{1}{2}\beta^{1-r} \cdot \beta^e$$



und

$$x \geq \underline{x} \geq (1.00 \cdots 00)_\beta \cdot \beta^e,$$

also

$$\frac{|x - \text{rd}(x)|}{|x|} \leq \frac{1}{2} \beta^{1-r} \cdot \beta^e \cdot \frac{1}{\beta^e} = \frac{1}{2} \beta^{1-r}$$

wie behauptet. □

**Quizfrage:** Gilt die Abschätzung (5.7) auch für Zahlen  $x$  mit  $|x| < x_{\text{posmin}}$ ? Falls nicht, wie groß kann der relative Fehler bei der Rundung für diese Zahlen höchstens werden?

**Folgerung 5.7** (zur Maschinengenauigkeit). Für alle  $x \in \mathbb{D}$  (einschließlich  $x = 0$ ) gilt

$$\text{rd}(x) = x(1 + \varepsilon) \tag{5.8}$$

mit einer Zahl  $\varepsilon \in \mathbb{R}$ , die  $|\varepsilon| \leq \varepsilon_{\text{mach}}$  erfüllt.

**Beispiel 5.8** (Maschinengenauigkeit).

(i) Für das Fließkommagitter aus *Beispiel 5.2* gilt

$$\varepsilon_{\text{mach}} = \frac{1}{2} 2^{1-3} = 2^{-3} = \frac{1}{8}.$$

(ii) Im Falle des double precision-Standards haben wir

$$\varepsilon_{\text{mach}} = \frac{1}{2} 2^{1-53} = 2^{-53} \approx 1.11 \cdot 10^{-16}.$$

(iii) Im Zehnersystem ( $\beta = 10$ ) mit zwei Nachkommastellen, also für Zahlen der Bauart 1.23 mit Mantissenlänge  $r = 3$ , ergibt sich die „Maschinen“genauigkeit  $\varepsilon_{\text{mach}} = \frac{1}{2} 10^{1-3} = 0.005$ .

**Bemerkung 5.9** (Maschinengenauigkeit).

- (a) Die Maschinengenauigkeit ist gerade die Hälfte des Stellenwertes der letzten Ziffer der Mantisse; vgl. (5.2). Auf die darstellbaren Exponenten kommt es dabei nicht an.
- (b) Die doppelte Fließkommagenauigkeit  $2\varepsilon_{\text{mach}}$  ist gerade der Abstand zwischen der Fließkommazahl 1 und der nächst größeren Fließkommazahl.

**Bemerkung 5.10** (Alternative Definition der Maschinengenauigkeit). Manche Autoren definieren die Maschinengenauigkeit als das Doppelte von unserer Definition. Das liegt darin begründet, dass (5.7) dann auch für andere Rundungsarten als round to nearest gilt, etwa round toward 0. Siehe auch die Diskussion auf [https://en.wikipedia.org/wiki/Machine\\_epsilon](https://en.wikipedia.org/wiki/Machine_epsilon).

## § 6 RECHNERARITHMETIK

Die Grundoperationen  $+$ ,  $-$ ,  $\cdot$  und  $/$  liefern, selbst für Argumente in  $\mathbb{F}$ , im Allgemeinen Ergebnisse, die keine Fließkommazahlen in  $\mathbb{F}$  sind. (**Quizfrage:** Können Sie Beispiele dafür angeben, wann das der Fall ist, etwa für das Fließkommasystem  $\mathbb{F}$  aus [Beispiel 5.2](#)?) Die Grundoperationen lassen sich also nicht fehlerfrei in  $\mathbb{F}$  ausführen. Stattdessen werden Ersatzoperationen, genannt **Maschinenoperationen**  $\oplus$ ,  $\ominus$ ,  $\odot$  und  $\oslash$  implementiert (in Hardware), die Ergebnisse in  $\mathbb{F}$  liefern. Der IEEE-Standard sieht vor, dass diese Grundoperationen mit korrekter Rundung ausgeführt werden. Das heißt konkret, dass folgende Beziehungen gelten:

$$x \oplus y = \text{rd}(x + y), \quad (6.1a)$$

$$x \ominus y = \text{rd}(x - y), \quad (6.1b)$$

$$x \odot y = \text{rd}(x \cdot y), \quad (6.1c)$$

$$x \oslash y = \text{rd}(x/y). \quad (6.1d)$$

Dabei sind jeweils  $x, y \in \mathbb{F}$ , und das Ergebnis liegt in  $\mathbb{F} \cup \{\pm\infty\}$  bzw. ergibt NaN im Falle von  $0 \oslash 0$ . Aus [Lemma 5.6](#) bzw. [Folgerung 5.7](#) ergibt sich direkt folgendes Resultat:

**Satz 6.1** (Relativer Fehler der Maschinenoperationen  $\oplus$ ,  $\ominus$ ,  $\odot$  und  $\oslash$ ). *Es seien  $x, y \in \mathbb{F}$ .*

(i) Falls  $x + y \in \mathbb{D}$  liegt, so gilt mit einer Zahl  $\varepsilon \in \mathbb{R}$ ,  $|\varepsilon| \leq \varepsilon_{\text{mach}}$ :

$$x \oplus y = (x + y)(1 + \varepsilon).$$

(ii) Falls  $x - y \in \mathbb{D}$  liegt, so gilt mit einer Zahl  $\varepsilon \in \mathbb{R}$ ,  $|\varepsilon| \leq \varepsilon_{\text{mach}}$ :

$$x \ominus y = (x - y)(1 + \varepsilon).$$

(iii) Falls  $x \cdot y \in \mathbb{D}$  liegt, so gilt mit einer Zahl  $\varepsilon \in \mathbb{R}$ ,  $|\varepsilon| \leq \varepsilon_{\text{mach}}$ :

$$x \odot y = (x \cdot y)(1 + \varepsilon).$$

(iv) Falls  $y \neq 0$  ist und  $x/y \in \mathbb{D}$  liegt, so gilt mit einer Zahl  $\varepsilon \in \mathbb{R}$ ,  $|\varepsilon| \leq \varepsilon_{\text{mach}}$ :

$$x \oslash y = (x/y)(1 + \varepsilon).$$

**Beachte:** Die Voraussetzung  $x + y \in \mathbb{D}$  besagt, dass das korrekte Resultat im Bereich der von den zur Verfügung stehenden Fließkommazahlen überdeckten Intervalle [\(5.6\)](#) liegt.

Wie passt die Aussage von [Satz 6.1](#), dass  $x \ominus y$  nur um einen kleinen relativen Faktor vom echten Ergebnis  $x - y$  abweicht, mit der in [§ 3](#) gemachten Beobachtung zusammen, dass die Aufgabe,  $x - y$  zu berechnen, eine beliebig große relative Konditionszahl haben kann?

Das sind zwei verschiedene Paar Schuhe. Die Aussage von [Satz 6.1](#) bezieht sich auf die Berechnung von  $x \ominus y$ , wobei  $x$  und  $y$  bereits Fließkommazahlen sind und keiner Rundung unterliegen. (Eine Rundung erfolgt erst **nach** der Berechnung von  $x - y$ .) Die Aussage zur Kondition bezieht sich darauf, dass das Ergebnis von  $x - y$  stark abweichen kann, wenn die Daten  $x$  und  $y$  nicht exakt sind, etwa durch Rundung **vor** der Berechnung, um  $x$  und  $y$  in Fließkommazahlen zu verwandeln.

**Bemerkung 6.2** (Assoziativ- und Distributivgesetz).

(a) Das Assoziativgesetz und das Distributivgesetz gelten für die Maschinenoperationen nicht mehr! Es ist also i. A.

$$\begin{aligned}(x \oplus y) \oplus z &\neq x \oplus (y \oplus z), \\ (x \oplus y) \odot z &\neq (x \odot z) \oplus (y \odot z)\end{aligned}$$

für  $x, y, z \in \mathbb{F}$ .

(b) Das Kommutativgesetz der Addition und der Multiplikation gelten für  $\oplus$  und  $\odot$  aber weiter, d. h.,

$$\begin{aligned}x \oplus y &= y \oplus x, \\ x \odot y &= y \odot x\end{aligned}$$

für  $x, y \in \mathbb{F}$ .

(c) Diese Beobachtung führt dazu, dass man Algorithmen, bevor man den Einfluss von Rundungsfehlern untersucht, zweifelsfrei notieren muss, weil bereits die Reihenfolge einen Einfluss auf das Ergebnis haben kann. Dabei helfen Klammersetzung und Einführung von Zwischenresultaten. Beispielsweise sind zur Auswertung von  $y = x_1^2 - x_2^2 = (x_1 + x_2)(x_1 - x_2)$  folgende Realisierungen in Algorithmen denkbar:

$u_1 := x_1 \odot x_1$	$u_1 := x_1 \oplus x_2$	$u_1 := x_1 \ominus x_2$
$u_2 := x_2 \odot x_2$	$u_2 := x_1 \ominus x_2$	$u_2 := x_1 \oplus x_2$
$y := u_1 \ominus u_2$	$y := u_1 \odot u_2$	$y := u_1 \odot u_2$

Mit den Operationen  $+$ ,  $-$ ,  $\cdot$  in  $\mathbb{R}$  führen diese Algorithmen zu identischen Ergebnissen, in Maschinearithmetik mit  $\oplus$ ,  $\ominus$ ,  $\odot$  sind die zweite und dritte Variante identisch (**Quizfrage: Warum?**), aber verschieden von der ersten.

Ende der Woche 4

# Index

- 1-Norm, 11
- 2-Norm, 11
- $C^k$ -Funktion, 16
- $O$ , *siehe* Groß- $O$
- $o$ , *siehe* Klein- $o$
- $\infty$ -Norm, 11
- $k$ -mal stetig partiell differenzierbar, 16
  
- Ableitung, 15
- absolute Konditionszahl, 25
- absolute partielle Konditionszahlen, 24
- Adjazenzmatrix, 6
- allgemeine linearen Gruppe, 36
- asymptotisch obere Schranke, 19
- asymptotisch vernachlässigbar, 20
- Auslöschung, 31
  
- Basis einer Fließkommazahl, 45
- binary64, *siehe* double precision
  
- Cauchy-Schwarz-Ungleichung, 11
  
- Definitheit einer Norm, 10
- die Rundung zur nächstgelegenen Fließkommazahl, 48
- differentielle Sensitivitätsanalyse, 23
- differenzierbare Funktion, 15
- double precision, 47
- Dreiecksungleichung, 10
- dünne Singulärwertzerlegung, 39
  
- echte Addition, 31
- echte Subtraktion, 31
- Eigenpaar, 36
- Eigenvektor, 36
- Eigenwert, 36
- Eigenwertaufgabe, 8
- Einheitsmatrix, 12
- Elastizitäten, 27
- Euklidisches Innenprodukt, 11
  
- Euklidische Norm, 11
- Exponent, 45
  
- Festkommazahlen, 45
- Fließkommagitter, 45
- Fließkommazahl, 45
- Fließkommazahlen, 45
  
- Ganzzahlen, 45
- Groß- $O$ , 19, 20
- gut konditioniert im absoluten Sinne, 25
- gut konditioniert im relativen Sinne, 28
  
- Hessematrix, 17
  
- Jacobimatrix, 15
  
- Kante eines Graphen, 6
- Klein- $o$ , 20
- Knoten eines Graphen, 6
- Konditionszahl einer Matrix, 33
- konvexe Menge, 16
  
- Landau-Notation, 19
- Linkssingulärvektoren, 37
- Lipschitz-Konstante, 18
- Lipschitz-stetig, 18
  
- Mantisse, 45
- Maschinengenauigkeit, 48
- Maschinenoperation, 50
- Metrik, 11
- Mittelwertsatz, 17
  
- Norm, 10
- normalisierte Fließkommazahl, 45
- Numerik, 5
- numerische Mathematik, 5
  
- Operatornorm, 12
- orthogonale Gruppe, 36

orthogonale Matrix, 36  
Orthonormalbasis, 36

positive Homogenität einer Norm, 10  
Potenzmethode, 8

Rechtssingulärvektoren, 37  
relative Konditionszahl, 28  
relative partielle Konditionszahlen, 27  
Restglied, 15  
Rundung, 48

schlecht konditioniert im absoluten Sinne, 25  
schlecht konditioniert im relativen Sinne, 28  
Sensitivitätsanalyse, 22  
Signum, 14  
Singularwerte, 37  
Singularwertzerlegung, 37  
Spaltensummennorm, 13  
Spektralnorm, 13  
Spektralzerlegung einer reellen, symmetrischen  
Matrix, 36  
Subadditivität einer Norm, 10  
SVD, *siehe* Singularwertzerlegung

Vektoriteration, 8  
Verbindungsstrecke, 16  
Von-Mises-Iteration, 8

wissenschaftliches Rechnen, 5

Zeilensummennorm, 13  
Zeilenvektoren, 11  
zulässige Bereich, 47

## Literatur

- Bornemann, F. (2018). *Numerische lineare Algebra*. 2. Aufl. Springer Fachmedien Wiesbaden. DOI: [10.1007/978-3-658-24431-6](https://doi.org/10.1007/978-3-658-24431-6).
- Brin, S.; L. Page (1998). „The anatomy of a large-scale hypertextual Web search engine“. *Computer Networks and ISDN Systems* 30.1-7, S. 107–117. DOI: [10.1016/s0169-7552\(98\)00110-x](https://doi.org/10.1016/s0169-7552(98)00110-x).
- Golub, G.; C. van Loan (1983). *Matrix Computations*. Baltimore: Johns Hopkins University Press.
- Heuser, H. (2002). *Lehrbuch der Analysis. Teil 2*. 12. Aufl. Stuttgart: B.G.Teubner. DOI: [10.1007/978-3-322-96826-5](https://doi.org/10.1007/978-3-322-96826-5).
- Higham, N. J.; M. R. Dennis; P. Glendinning; P. A. Martin; F. Santosa; J. Tanner, Hrsg. (2016). *The Princeton Companion to Applied Mathematics*. Princeton University Press. DOI: [10.1515/9781400874477](https://doi.org/10.1515/9781400874477).
- Kahan, W. (1966). „Numerical linear algebra“. *Canadian Mathematical Bulletin* 9.05, S. 757–801. DOI: [10.4153/cmb-1966-083-2](https://doi.org/10.4153/cmb-1966-083-2).
- Trefethen, L. N. (1992). „The definition of numerical analysis“. *SIAM News*. URL: [https://people.maths.ox.ac.uk/trefethen/publication/PDF/1992\\_55.pdf](https://people.maths.ox.ac.uk/trefethen/publication/PDF/1992_55.pdf).