# Wissenschaftliches Rechnen II

Vorlesungsskripte
Sommer-Semester 2006

Werner Römisch

Humboldt-Universität Berlin
Institut für Mathematik

**Inhalt:** Seite

# 0 Einleitung

Die Leistungsexplosionen von Rechnern und die damit mögliche Bearbeitung praxisnäherer Aufgaben war der Anlaß und die Motivation für eine enge Verknüpfung von Ingenieurwissenschaften, Informatik und Mathematik. Dabei entstand als Ergebnis das interdisziplinäre Gebiet des **Wissenschaftlichen Rechnens** (Scientific Computing).

Im zweiten Teil des Kurses *Wissenschaftliches Rechnen* sollen **Spezifika numerischer Algorithmen** behandelt werden. Dabei werden anhand von ausgewählten Aufgabenstellungen (lineare Gleichungssysteme, lineare Optimierung) typische Vorgehensweisen bei der Konstruktion von numerischen Algorithmen, deren mathematischer Analyse und deren Implementierung diskutiert. Zugleich soll in die Nutzung von Standardsoftware eingeführt und Grundlagen für die kritische Evaluierung numerischer Ergebnisse gelegt werden.

**Beispiel 0.1** *:*
*Wir berechnen die Ableitung der Wurzelfunktion* $f : \mathbb{R}_+ \to \mathbb{R}$, $f(x) := x^{\frac{1}{2}}$, $\forall x \in \mathbb{R}$, *approximativ durch Differenzenquotienten*

$$\frac{f(x+h) - f(x)}{h} = \frac{(x+h)^{\frac{1}{2}} - x^{\frac{1}{2}}}{h}$$

*in* $x = x_0$ *und verschiedene (kleine)* $h > 0$.
*Der exakte Wert ist* $f'(x_0) = \frac{1}{2} x_0^{-\frac{1}{2}}$ *und für* $x_0 = 1000$ *erhält man auf einem Taschenrechner mit* 10*-stelliger Rechnung* $f'(1000) = 0.0158113883$.
*Auf diesem Taschenrechner erhalten wir für DQ und* $DQ^*$ *aus der Identität*

$$DQ := \frac{(x_0+h)^{\frac{1}{2}} - x_0^{\frac{1}{2}}}{h} = \frac{1}{(x_0+h)^{\frac{1}{2}} + x_0^{\frac{1}{2}}} =: DQ^*$$

*die folgenden Ergebnisse*

| $h$ | $10^{-3}$ | $10^{-5}$ | $10^{-7}$ |
|---|---|---|---|
| $DQ$ | 0.0158114 | 0.01581 | 0 |
| $DQ^*$ | 0.01581138435 | 0.0158113882 | 0.0158113883 |

*Je nach Implementierung mathematisch völlig identischer Formeln können also große Fehler auftreten! Kritisch an der Formel für DQ ist das Auftreten einer Differenz nahezu gleichgroßer Zahlen. Man erkennt auch, wie man die Zahlenwerte modifizieren kann, um bei Rechnungen mit beliebiger Stellenzahl katastrophale Ergebnisse zu produzieren.*

"Bonmot": (K. Nickel, Karlsruhe)

- Der (naive) Anfänger glaubt an jede einzelne Ziffer eines Ergebnisses.
- Der (erfahrene) Programmierer vertraut auf die Hälfte der Stellen.
- Der (wissende) Pessimist mißtraut sogar noch dem Vorzeichen.

**Beispiel 0.2** *:*
*Es sei* $A = (a_{ij})_{i,j=1,\ldots,n}$ *eine quadratische Matrix. Gesucht sei die Determinante* $\det A$ *von* $A$*. Diese kann nach der induktiven Determinantendefinition wie folgt berechnet werden:*

$$\det A = \sum_{i=1}^{n} a_{1i}(-1)^{i+1} D_{1i}$$

*wobei* $D_{1i}$ *die Determinante der Matrix darstellt, die durch Streichung der ersten Zeile und i-ten Spalte aus* $A$ *entsteht.*
*Es bezeichne nun* $z_n$ *die Anzahl der Rechenoperationen zur Berechnung von* $\det A$*. Dann gilt* $z_1 = 0$ *und aus der obigen induktiven Formel folgt für* $n > 1$*:*

$$\begin{aligned} z_n &= n - 1 + n + n z_{n-1} \\ &= n(z_{n-1} + 2) - 1 \end{aligned}$$

*(*$n - 1$ *Additionen,* $n$ *Multiplikationen und* $n$ *Berechnungen von Determinanten von Matrizen mit* $(n - 1)$ *Zeilen und Spalten)*

*Folglich gilt:* $z_n > n!$*.*

*Für einen Rechner mit* $10^6$ *Operationen pro Sekunde bedeuten* $n!$ *Operationen:*

| $n$ | *Rechenzeit* |
|---|---|
| $n = 11$ | $\approx 40\ s$ |
| $n = 21$ | $1620083\ a$ *(Jahre!)* |
| $n = 24$ | $19.6 \cdot 10^9\ a$ |
| $n = 30$ | $8.4 \cdot 10^{18}\ a$ *(Trillionen Jahre!)* |

*Selbst bei kühnsten Gedankenexperimenten mit gigantischen Rechnern (Parallelrechner mit* $1km^3$ *Volumen und Rechenelementen in Atomradiusgröße ist die Anzahl der Rechenoperationen durch etwa* $3 \cdot 10^{57}$ *beschränkt. Gleichzeitig ist* $100! \approx 10^{159.9}$*.)*

Problem: Mathematisch gut nutzbare Formeln können sich für die konkrete Berechnung als völlig ungeeignet erweisen!

In der Vorlesung wird exemplarisch die numerische Lösung folgender *Grundaufgaben* anhand ausgewählter Algorithmen untersucht:

(i) Lineare Gleichungssysteme: $Ax = b$, wobei $A = (a_{ij})_{i,j=1,\ldots,n}$ regulär, $b \in \mathbb{R}^n$.

(ii) Lineare Optimierungsprobleme: $\min\{c^T x : Ax = b, x \geq 0\}$, wobei $A = (a_{ij})_{\substack{i=1,\ldots,m \\ j=1,\ldots,n}}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ und "$\geq$" komponentenweise zu verstehen ist.

Bei der numerischen Lösung von praktischen Problemen gibt es eine Reihe von unvermeidbaren *Fehlerquellen*:

- *Modellfehler.*
  Die Abbildung eines realen Sachverhalts auf ein mathematisches Modell ist nur näherungsweise möglich.

- *Datenfehler:*
  Parameter des realen Modells gehen nur *näherungsweise* in das mathematische Modell ein.

- *Fehler im numerischen Lösungsprozeß:*

  - *Diskretisierungsfehler:*
    Abbildung kontinuierlicher Prozesse auf diskrete Größen.
  - *Darstellungsfehler:*
    z. B. Darstellungen reeller Zahlen im Rechner.
  - *Abbruchfehler:*
    Aufgaben, für die kein endlicher Algorithmus existiert, müssen durch sukzessive, unendliche Prozesse gelöst werden, die geeignet „abgebrochen“ werden müssen.
  - *Rundungsfehler:*
    Rechenoperationen mit reellen Zahlen sind auf Rechnern grundsätzlich nicht exakt durchführbar.

Die theoretische Einsicht in benutzte Verfahren ist stets die Voraussetzung für die kritische Einschätzung der Ergebnisse!

**Literatur:**


* G. Hämmerlin und K.-H. Hoffmann: Numerische Mathematik, Springer-Verlag, Berlin 1994 (4. Auflage).

* A. Kiełbasiński und H. Schwetlick: Numerische lineare Algebra, Verlag der Wissenschaften, Berlin 1988.

P. Deuflhard und A. Hohmann: Numerische Mathematik I, Walter de Gruyter, Berlin 1993 (2. Auflage).

G. H. Golub und C. F. van Loan: Matrix Computations (Second Edition), John Hopkins University Press, Baltimore 1993.

# 1 Grundlagen der Fehleranalyse

## 1.1 Normen von Vektoren und Matrizen

Vektoren und Matrizen treten bei der Beschreibung der Grundaufgaben dieser Vorlesung (als Eingangsdaten bzw. Lösungen) als elementare Bestandteile auf. Um nun Fehler quantifizieren zu können („große“ bzw. „kleine“ Fehler), ist es notwendig, einen Abstandsbegriff für diese Größen zu besitzen.

**Definition 1.1** *Eine Abbildung $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ heißt Norm in $\mathbb{R}^n$, falls*

(1) $\|x\| \geq 0, \forall x \in \mathbb{R}^n$, *und* $\|x\| = 0$ *gdw.* $x = \Theta = (0, \ldots, 0) \in \mathbb{R}^n$;
(2) $\|\alpha x\| = |\alpha| \|x\|, \forall x \in \mathbb{R}^n, \forall \alpha \in \mathbb{R}$;
(3) $\|x + y\| \leq \|x\| + \|y\|, \forall x, y \in \mathbb{R}^n$ *(Dreiecksungleichung).*

**Eigenschaften 1.2** *Es sei $\|\cdot\|$ eine Norm in $\mathbb{R}^n$. Dann gilt:*

*(i)* $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ *mit* $d(x, y) := \|x - y\|, \forall x, y \in \mathbb{R}^n$, *ist eine Metrik in $\mathbb{R}^n$ (mit den üblichen Eigenschaften);*

*(ii)* $|\|x\| - \|y\|| \leq \|x - y\| \leq \|x\| + \|y\|, \ \forall x, y \in \mathbb{R}^n$.

Beweis:

(i) Die Metrik-Eigenschaften folgen aus (1)-(3) von Def. 1.1.

(ii) Die Ungleichung auf der linken Seite folgt aus

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

wobei die Rollen von $x$ bzw. $y$ vertauscht werden können,
und die auf der rechten Seite folgt aus (2) bzw. (3). □

**Beispiel 1.3**

*(i) Sei $p \in [1, \infty]$.*

$$\|x\|_p := \left( \sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}, \quad \forall x = (x_1, \ldots, x_n) \in \mathbb{R}^n, \forall p \in [1, \infty);$$

$$\|x\|_\infty := \max_{i=1,\ldots,n} |x_i|, \qquad \forall x = (x_1, \ldots, x_n) \in \mathbb{R}^n$$

*Dann ist $\|\cdot\|_p$ eine Norm auf $\mathbb{R}^n$.*
*(Dabei sind die Eigenschaften (1) und (2) einfach zu sehen, wie auch die Dreiecksungleichung für $\|\cdot\|_\infty$ und $\|\cdot\|_1$; für die Dreiecksungleichung von $\|\cdot\|_p$, $p > 1$, benötigt man die sog. Hölder-Ungleichung (vgl. Übungen).)*

*(ii) Für je zwei Elemente $x, y \in \mathbb{R}^n$ definiert der Ausdruck*

$$\langle x, y\rangle := x^T y = \sum_{i=1}^{n} x_i y_i$$

*ein sog. Skalarprodukt in $\mathbb{R}^n$. Es gilt $\|x\|_2 = \langle x, x\rangle^{\frac{1}{2}}$.*

**Definition 1.4** *Es sei $A$ eine $m \times n$-Matrix, $\|\cdot\|_X$ eine Norm in $\mathbb{R}^n$ und $\|\cdot\|_Y$ eine Norm in $\mathbb{R}^m$. Dann heißt*

$$\|A\| := \sup_{x \neq \Theta} \frac{\|Ax\|_Y}{\|x\|_X} = \sup_{\|x\|_X = 1} \|Ax\|_Y$$

*die (zu $\|\cdot\|_X$ und $\|\cdot\|_Y$) zugeordnete Matrixnorm.*

**Lemma 1.5** *Es seien $\|\cdot\|_X$ und $\|\cdot\|_Y$ Normen auf $\mathbb{R}^n$ bzw. $\mathbb{R}^m$, und $\mathbb{R}^{m\times n}$ bezeichne den linearen Raum aller $m \times n$-Matrizen.*
*Dann ist die zugeordnete Matrixnorm eine Norm auf $\mathbb{R}^{m\times n}$. Überdies gilt:*

*(i) Für alle $A \in \mathbb{R}^{m\times n}$ und $x \in \mathbb{R}^n$ gilt $\|Ax\|_Y \leq \|A\|\|x\|_X$ und $\|A\|$ ist die kleinste aller Zahlen $C > 0$ mit $\|Ax\|_Y \leq C\|x\|_X, \forall x \in \mathbb{R}^n$.*

*(ii) Sei zusätzlich $\|\cdot\|_Z$ eine Norm auf $\mathbb{R}^k$ und $B$ eine $k \times m$-Matrix.*
*Dann gilt für die entsprechenden Matrixnormen:*

$$\|BA\| \leq \|B\|\ \|A\|.$$

Beweis:
Wir diskutieren zunächst die Normeigenschaften der Matrixnorm. Es gilt natürlich $\|A\| \geq 0$ für alle $A \in \mathbb{R}^{m\times n}$ und $\|A\| = 0$ gdw. $\|Ax\|_Y = 0$ für alle $x \in \mathbb{R}^n$ gdw. $A = \Theta$ (Nullmatrix).
Ferner gilt für bel. $\alpha \in \mathbb{R}$ und $A, B \in \mathbb{R}^{m\times n}$:

$$\begin{aligned}
\|\alpha A\| &= \sup_{\|x\|_X=1} \|\alpha Ax\|_Y = |\alpha| \sup_{\|x\|_X=1} \|Ax\|_Y = |\alpha|\ \|A\| \\
\|A + B\| &= \sup_{\|x\|_X=1} \|(A+B)x\|_Y \leq \sup_{\|x\|_X=1} (\|Ax\|_Y + \|Bx\|_Y) \\
&\leq \sup_{\|x\|_X=1} \|Ax\|_Y + \sup_{\|x\|_X=1} \|Bx\|_Y = \|A\| + \|B\|.
\end{aligned}$$

Also ist $\|\cdot\|$ eine Norm auf dem linearen Raum $\mathbb{R}^{m\times n}$.

(i) Es sei $A \in \mathbb{R}^{m\times n}$ und $x \in \mathbb{R}^n$. Für $x = \Theta$ ist die behauptete Ungleichung richtig, für $x \neq \Theta$ gilt aber $\frac{\|Ax\|_Y}{\|x\|_X} \leq \|A\|$, d. h., die gewünschte Ungleichung.
Ist $C > 0$ eine Zahl mit $\|Ax\|_Y \leq C\|x\|_X$, so gilt für alle
$x \in \mathbb{R}^n$ mit $\|x\|_X = 1 : \|Ax\|_Y \leq C$, d. h., $\|A\| \leq C$.

(ii) Es sei $x \in \mathbb{R}^n, \|x\|_X = 1$. Dann gilt:
$\|BAx\|_Z \le \|B\| \, \|Ax\|_Y \le \|B\| \, \|A\| \, \|x\|_X = \|B\| \, \|A\|$ und damit
$\|BA\| = \sup\limits_{\|x\|_X=1} \|BAx\|_Z \le \|B\| \, \|A\|$. □

**Beispiel 1.6**

*a) Für $\|\cdot\|_X = \|\cdot\|_1$ auf $\mathbb{R}^n$ und $\|\cdot\|_Y = \|\cdot\|_1$ auf $\mathbb{R}^m$ gilt für $A = (a_{ij})$ :*
$\|A\|_1 := \sup\limits_{\|x\|_1=1} \|Ax\|_1 = \max\limits_{j=1,\ldots,n} \sum\limits_{i=1}^{m} |a_{ij}|$ *(Spaltensummennorm)*

*Beweis:*
*Für bel. $x \in \mathbb{R}^n$ mit $\|x\|_1 = \sum\limits_{j=1}^{n} |x_i| = 1$ gilt:*

$$\|Ax\|_1 = \sum_{i=1}^{m} \left| \sum_{j=1}^{n} a_{ij} x_j \right| \le \sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}| \, |x_j| \le \left( \max_{j=1,\ldots,n} \sum_{i=1}^{m} |a_{ij}| \right) \left( \sum_{j=1}^{n} |x_j| \right),$$

*also $\|A\|_1 \le \max\limits_{j=1,\ldots,n} \sum\limits_{i=1}^{m} |a_{ij}|$.*
*Um die Gleichheit zu zeigen, sei $j_0 \in \{1, \ldots, n\}$ der Index mit*
*$\sum\limits_{i=1}^{m} |a_{ij_0}| = \max\limits_{j=1,\ldots,n} \sum\limits_{i=1}^{m} |a_{ij}|$. Dann gilt für $\bar{x} := (0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^n$ :*
$\uparrow_{j_0}$
$\|\bar{x}\|_1 = 1$ *und* $\|A\bar{x}\|_1 = \sum\limits_{i=1}^{m} |a_{ij_0}| \le \|A\|_1$. □

*b) Für $\|\cdot\|_X = \|\cdot\|_\infty$ und $\|\cdot\|_Y = \|\cdot\|_\infty$ gilt*
$\|A\|_\infty = \sup\limits_{\|x\|_\infty=1} \|Ax\|_\infty = \max\limits_{i=1,\ldots,m} \sum\limits_{j=1}^{n} |a_{ij}|$ *(Zeilensummennorm)*
*(Übung)*

*c) Für $\|\cdot\|_X = \|\cdot\|_2$ und $\|\cdot\|_Y = \|\cdot\|_2$ gilt*
$\|A\|_2 = \sup\limits_{\|x\|_2=1} \|Ax\|_2 = \sqrt{\lambda_{\max}(A^T A)}$ *(Spektralnorm).*
*Hierbei ist $\lambda_{\max}(A^T A)$ der größte Eigenwert der symmetrischen und positiv semidefiniten Matrix $A^T A$.*
*Beweisskizze: Alle Eigenwerte von $A^T A$ sind reell und nichtnegativ. Überdies besitzt $A^T A$ zu den Eigenwerten $\lambda_i$, $i = 1, \ldots, n$, eine Orthonormalbasis von (reellen) Eigenvektoren $x_i$, $i = 1, \ldots, n$. Damit gilt:*

$$\begin{aligned} \|Ax\|_2^2 &= \langle Ax, Ax \rangle = \langle A^T Ax, x \rangle \\ &= \langle \sum_{i=1}^{n} \langle x, x_i \rangle A^T A x_i, \sum_{i=1}^{n} \langle x, x_i \rangle x_i \rangle = \sum_{i=1}^{n} \lambda_i \langle x, x_i \rangle^2 \\ &\le \lambda_{\max}(A^T A) \|x\|^2 \end{aligned}$$

*Also gilt: $\|A\|_2 \le \sqrt{\lambda_{\max}(A^T A)}$.*
*Es sei schließlich $x_{i_0}$ ein normierter Eigenvektor zu $\lambda_{\max}(A^T A)$. Dafür gilt*

$$\|x_{i_0}\|_2 = 1 \quad \textit{und} \quad \|Ax_{i_0}\|_2^2 = \lambda_{\max}(A^T A)$$

*und deshalb die gewünschte Aussage.*

*d) Ist $B \in \mathbb{R}^{n\times n}$ eine reguläre Matrix und $\|\cdot\|$ eine Norm auf $\mathbb{R}^n$, so ist $\|x\|_B := \|Bx\|$, $\forall x \in \mathbb{R}^n$, eine Norm auf $\mathbb{R}^n$ (Übung).*

**Bemerkung 1.7** *Die Spektralnorm $\|A\|_2$ für $A \in \mathbb{R}^{m\times n}$ ist schwer zu berechnen. Deshalb benutzt man häufig die folgenden oberen Schranken für $\|A\|_2$:*

$$\|A\|_F := \left(\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2\right)^{\frac{1}{2}} \quad (\textit{Frobenius-Norm})$$

$$\|A\|_{\max} := \sqrt{nm}\max_{\substack{i=1,\dots,m\\ j=1,\dots,n}} |a_{ij}|$$

*die beide auch Normen auf $\mathbb{R}^{m\times n}$ sind.*

**Definition 1.8** *Zwei Normen $\|\cdot\|_X$ und $\|\cdot\|_Y$ in $\mathbb{R}^n$ heißen äquivalent, falls es Konstanten $m > 0$, $M > 0$ gibt, so daß die Abschätzung*

$$m\|x\|_X \le \|x\|_Y \le M\|x\|_X$$

*für alle $x \in \mathbb{R}^n$ gilt.*

*Die Eigenschaft „äquivalent“ zu sein, definiert eine Äquivalenzrelation in der Menge aller Normen. Sie ist offenbar reflexiv, symmetrisch und transitiv. Auf $\mathbb{R}^n$ gilt nun die folgende, auf den ersten Blick überraschende, Aussage.*

**Satz 1.9** *Alle Normen in $\mathbb{R}^n$ sind äquivalent.*

Beweis: Wir zeigen, daß jede beliebige Norm $\|\cdot\|$ in $\mathbb{R}^n$ zur Euklidischen Norm $\|\cdot\|_2$ äquivalent ist. Dann sind auch je zwei beliebige Normen auf $\mathbb{R}^n$ äquivalent wegen der Transitivität der Norm-Äquivalenz.
Es sei $e^i = (0,\dots,0,\underset{\uparrow_i}{1},0,\dots,0)$ der $i$-te kanonische Einheitsvektor. Dann gilt für $x = (x_1,\dots,x_n) \in \mathbb{R}^n$, daß $x = \sum_{i=1}^{n} x_i e^i$ und

$$\|x\| = \|\sum_{i=1}^{n} x_i e^i\| \le \sum_{i=1}^{n} |x_i|\ \|e^i\| \le \|x\|_2 \left(\sum_{i=1}^{n}\|e^i\|^2\right)^{\frac{1}{2}}.$$

Wir setzen $M := \left(\sum_{i=1}^{n}\|e^i\|^2\right)^{\frac{1}{2}}$ und wissen nach Eigenschaft 1.2

$$|\,\|x\| - \|y\|\,| \le \|x-y\| \le M\|x-y\|_2, \forall x,y \in \mathbb{R}^n.$$

D. h. $\|\cdot\| : (\mathbb{R}^n, \|\cdot\|_2) \to \mathbb{R}$ ist (Lipschitz-) stetig. Da nun die Menge $S_n := \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$ beschränkt und abgeschlossen ist, nimmt $\|\cdot\|$ dort ein Minimum an, d. h. es existiert ein $\hat{x} \in S_1$ mit $m := \|\hat{x}\| = \inf_{x\in S_1}\|x\|$.
Wegen $\hat{x} \neq \Theta$ gilt $m > 0$ und $m \le \left\|\frac{x}{\|x\|_2}\right\|, \forall x \in \mathbb{R}^n, x \neq \Theta$.
Folglich gilt $m\|x\|_2 \le \|x\| \le M\|x\|_2, \forall x \in \mathbb{R}^n$. □

**Bemerkung 1.10** *Satz 1.9 gibt uns die Möglichkeit, bei Fehlerabschätzungen irgendeine Norm zu verwenden. Die Aussagen bleiben bis auf Konstanten qualitativ richtig. Allerdings sind oftmals ganz bestimmte Normen an ein gegebenes Problem besonders angepaßt und dann kann durch die Konstanten Informationsverlust auftreten.*

## 1.2 Kondition von Aufgaben

(Numerische) Lösung eines Problems:
Berechnung von Lösungen (Ergebnissen) aus gewissen Eingangsgrößen (Daten) nach wohlbestimmten Vorschriften. Wir bezeichnen die Daten mit $d$, die Ergebnisse mit $s$, die Vorschriften mit $P$ und schreiben:

$$s = P(d)$$

**Beispiel 1.11**

*a) Lösung eines linearen Gleichungssystems $Ax = b$ mit einer regulären Matrix $A$. Hierbei ist*

$$d = (A, b), \quad s = x, \quad P(d) = A^{-1}b$$

*b) Lösung eines linearen Optimierungsproblems* $\min\{\langle c, x\rangle : Ax = b, x \geq 0\}$. *Das Problem bedeutet: Minimiere die Funktion $f(x) := \langle c, x\rangle$ bzgl. aller $x \in \mathbb{R}^n$, die der "Restriktion" $x \in M := \{x \in \mathbb{R}^n : Ax = b, x \geq 0\}$, mit gegebenen $c \in \mathbb{R}^n, b \in \mathbb{R}^m$ und $A \in \mathbb{R}^{m\times n}$, genügen.*
*($M$ ist ein sog. Polyeder (vgl. Kap. 3).)*
*Bei diesem Problem ist $d = (A, b, c), \langle c, s\rangle = \min\limits_{x\in M}\langle c, x\rangle$ und $P(d)$ ist eine Vorschrift, die besagt, wie man eine Lösung $s$ des Problems aus $d = (A, b, c)$ berechnet. Eine einfache explizite Beschreibung von $P$ ist hier nicht mehr möglich.*

Die Eingangsgrößen für numerische Rechnungen sind (wie bereits in der Einleitung diskutiert) in der Regel fehlerbehaftet und (folglich) nur näherungsweise bekannt. Mit der Entscheidung darüber, was Daten sind, wird gleichzeitig impliziert, welche Größen als fehlerbehaftet anzusehen sind.
Deshalb ist es wichtig, das Verhalten des Problems $P$ bei "Störung" der Daten $d \in D$ zu kennen. Die nachfolgende Eigenschaft der *korrekten Gestelltheit (well posedness)* von $P$ ist dabei wünschenswert.

**Definition 1.12**

*(i) Das Problem $P$ heißt korrekt gestellt, falls eine Abbildung $L : D \times \mathbb{R}_+ \to \mathbb{R}_+$ existiert, so daß*

$$\|P(d) - P(\tilde{d})\| \leq L(d, \|d - \tilde{d}\|)\|d - \tilde{d}\| \tag{1.1}$$

*für alle $d, \tilde{d} \in D$ gilt und die Funktion $L(d, \cdot) : \mathbb{R}_+ \to \mathbb{R}_+$ stetig (in 0) ist.*
*Anderenfalls heißt $P$ inkorrekt oder schlecht gestellt.*

*(ii) Die Zahl* $K(d) := L(d,0)\frac{\|d\|}{\|P(d)\|}$ *heißt (relative)* Konditionszahl *des Problems* $P$ *in den Daten* $d$.

*(iii) Das Problem heißt* gut konditioniert *(in* $d$*), falls* $K(d)$ *nicht zu groß ist; andernfalls* schlecht konditioniert.

**Bemerkung 1.13** *Korrektheit des Problems ist gleichbedeutend mit der (lokal) Lipschitzstetigen Abhängigkeit der Lösung von den Daten (in einer Umgebung von* $d$*).*
*Die Bedingung (1.1) is äquivalent zur folgenden Eigenschaft von* $P$

$$\frac{\|P(d) - P(\tilde{d})\|}{\|P(d)\|} \leq K(d, \|d - \tilde{d}\|)\frac{\|d - \tilde{d}\|}{\|d\|}$$

*in die die relativen Fehler von Lösungen und Daten eingehen. Dabei ist* $K(d, \|d-\tilde{d}\|) := L(d, \|d - \tilde{d}\|)\frac{\|d\|}{\|P(d)\|}$.
*Ist* $P$ *nach* $d$ *differenzierbar (genauer: stetig differenzierbar; vgl. Mittelwertsätze im Grundkurs Analysis), so ist die Norm der Ableitung von* $P$*, d. h. der Jacobi-Matrix, in* $d$ *ein Orientierungspunkt für die Größe von* $L(d, \Delta_d)$ *mit kleinem* $\Delta_d$.

Da wir später gut konditionierte bzw. korrekt gestellte Probleme kennenlernen werden, beschränken wir uns bei den folgenden Beispielen auf schlecht konditionierte bzw. schlecht gestellt Probleme.

**Beispiel 1.14** *(Wilkinson)*
*Das Problem sei, die Nullstellen des Polynoms*

$$\begin{aligned} p(x) := \prod_{i=1}^{20}(x-i) &= x^{20} - \left(\sum_{i=1}^{20} i\right)x^{19} + \ldots + \prod_{i=1}^{20} i \\ &= x^{20} - 210x^{19} + \ldots + 20! \end{aligned}$$

*zu bestimmen. Die Nullstellen, d.h., das Ergebnis* $s$ *ist klar:* $s = (1, 2, \ldots, 20)$*. Die Daten* $d$ *bestehen aus den Koeffizienten des Polynoms* $p(\cdot)$*, die Vorschriften* $P$ *sind kompliziert, sollen aber hier auch nicht weiter interessieren.*
*Wir fragen uns nun, ob das Problem gut oder schlecht konditioniert ist. Dazu stören wir den Koeffizienten* $-210$ *von* $x^{19}$ *durch* $-(210 + \varepsilon)$ *mit* $\varepsilon = -2^{-23} = -1.192 \cdot 10^{-7}$*. Dies entspricht einer Störung im letzten Bit einer 32-Bit-Darstellung (vgl. Bem. 1.19). Die Nullstellen des gestörten Polynoms* $p_\varepsilon(\cdot)$ *sind dann*

$$\begin{aligned} \tilde{s}_1 &= 1.0000\,00000 \\ \tilde{s}_4 &= 4.0000\,00000 \\ \\ \tilde{s}_{10,11} &= 10.095266145 \pm 0.643500904\,i \\ \tilde{s}_{16,17} &= 16.730737466 \pm 2.812624894\,i \end{aligned}$$

*Die Fehler sind offenbar sehr groß und aus den einfachen, gut separierten Nullstellen sind teilweise sogar konjugiert komplexe Nullstellen geworden. Man kann für die zu* $s_{20}$ *bzw.* $s_{16}$ *gehörigen "absoluten Konditionszahlen" zeigen:*

$$L_{s_{20}} \approx 0.9 \cdot 10^{10}, \quad L_{s_{16}} \approx 3.7 \cdot 10^{14}$$

*Also ist die Aufgabe, diese Nullstellen zu berechnen, schlecht (!) konditioniert !*

**Beispiel 1.15** *Wir betrachten das lineare Optimierungsproblem*

$$\min\{-x_1 : x_1 \geq 0, x_2 \geq 0, x_1 \leq 1, x_2 \leq 1, \varepsilon x_1 + x_2 = 0\}$$

*oder äquivalent dazu*

$$\min\{-x_1 : x_i \geq 0, i = 1, \ldots, 4,\ x_1 + x_3 = 1,\ x_2 + x_4 = 1,\ \varepsilon x_1 + x_2 = 0\},$$

*wobei Äquivalenz bedeutet, daß die ersten beiden Komponenten der Lösung des letzteren Problems gerade die Lösung der Ausgangsaufgabe ist.*
*Damit hat das Problem jetzt die Form von Bsp. 1.11, wobei*

$$c = (-1, 0, 0, 0), \quad A = \begin{pmatrix} \varepsilon & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}.$$

*In den Daten $d = (A, b, c)$ wird nur $\varepsilon \geq 0$ gestört und man erhält die Lösungen:*

$$\begin{array}{ll} \varepsilon > 0 : \{(0,0)\} & (\textit{oder } \{(0,0,1,1)\}) \\ \varepsilon = 0 : \{(1,0)\} & (\textit{oder } \{(1,0,0,1)\}) \end{array}$$

*Also ist die Aufgabe <u>nicht</u> korrekt gestellt!*
*(Es sei angemerkt, daß unter gewissen Voraussetzungen ein lineares Optimierungsproblem durchaus korrekt gestellt bzgl. der Daten $d = (A, b, c)$ aus einer gewissen Menge $D$ sein kann. Diese Voraussetzungen sind jedoch hier nicht erfüllt.)*

## 1.3 Computerzahlen

Ausgangspunkt für die Darstellung von reellen Zahlen und von Computerzahlen sind Aussagen des folgenden Typs.

**Lemma 1.16** *Es sei $\beta \in \mathbb{N}, \beta \geq 2$ und $x \in \mathbb{R}, x \neq 0$.*
*Dann gibt es genau eine Darstellung der Gestalt*

$$x = \sigma \beta^N \sum_{i=1}^{\infty} x_{-i} \beta^{-i}$$

*mit $\sigma \in \{-1, 1\}, N \in \mathbb{Z}$ und $x_{-i} \in \{0, 1, \ldots, \beta - 1\}$, wenn man von den Zahlen $x_{-i}$ noch zusätzlich verlangt, daß $x_{-1} \neq 0$ und daß zu jedem $n \in \mathbb{N}$ ein Index $i \geq n$ existiert mit $x_{-i} \neq \beta - 1$.*

<u>Beweis:</u>
Grundvorlesung Analysis oder Hämmerlin/ Hoffmann, Kap. 1.1.

**Bemerkung 1.17** *Wichtige Spezialfälle von Lemma 1.16 sind $\beta = 2$ (Dualsystem), $\beta = 10$ (Dezimalsystem) und $\beta = 16$ (Hexadezimalsystem).*
*Für eine Zahl in der Basisdarstellung aus 1.16 (d. h. zur Basis $\beta$) wählt man nun eine spezielle Codierung (Darstellung):*

$$x = \sigma \cdot 0.x_{-1}x_{-2}x_{-3} \cdots \beta^N.$$

*Den Elementen von $\{0,\ldots,\beta-1\}$ ordnet man Zeichen zu, die Ziffern genannt werden (z. B. $\beta = 2$ : Ziffern 0,1; $\beta = 16$: Ziffern 0,1,...,9,A,...,F). Das Dualsystem ist für elektronische Rechenanlagen besonders geeignet, da es der Unterscheidung zweier Zustände entspricht. Bei der Benutzung eines anderen Zahlsystems müssen die Ziffern (wieder) binär codiert werden, was besonders einfach ist, wenn dessen Basis $\beta$ eine Zweierpotenz ist.*

Klar ist natürlich, daß auf einer elektronischen Rechenanlage die Menge der Computerzahlen ebenfalls endlich ist. Deswegen muß die Reihe bei einem $t \in \mathbb{N}$, der Mantissenlänge, abgebrochen und der Exponent $N$ beschränkt werden. Dies führt zu folgendem Begriff einer Menge von Computerzahlen.

**Definition 1.18** *Für gegebene $\beta, t, N_-, N_+ \in \mathbb{N}, \beta \geq 2$, nennen wir*

$$M(\beta,t,N_-,N_+) := \Big\{0, x = \sigma\, 0.x_{-1}x_{-2}\cdots x_{-t}\,\beta^N \,:\, \sigma \in \{-1,1\}, x_{-1} \neq 0, \\ x_{-i} \in \{0,1,\ldots,\beta-1\}, i = 1,2,\ldots,t, -N_- \leq N \leq N_+\Big\}$$

*Menge von Computerzahlen zur Basis $\beta$, mit Mantissenlänge $t$ und Exponentenbereich $[-N_-, N_+]$.*

**Bemerkung 1.19** *Alle Computerzahlen $x \neq 0$ liegen (also) im Bereich $\beta^{-N_- -1} \leq |x| < \beta^{N_+}$. Gilt $|x| < \beta^{-N_- -1}$, so wird $x$ durch Null ersetzt. Zahlen $x$ mit $|x| \geq \beta^{N_+}$ können nicht verarbeitet werden. Treten diese beiden Fälle auf, so spricht man von Exponentenüberlauf.*

Wegen des nötigen Übergangs von einer reellen Zahl $x$ zu einer Computerzahl, muß $x$ i. a. durch eine Näherung ersetzt werden. Dieser Prozeß wird als Runden bezeichnet. Dabei gehen wir davon aus, daß die folgende Rundungsvorschrift realisiert wird.

**Definition 1.20** *Es sei $\beta \in \mathbb{N}$ gerade, $t \in \mathbb{N}$ und $x \in \mathbb{R}\backslash\{0\}$ besitze die Darstellung $x = \sigma\beta^N \sum_{i=1}^{\infty} x_{-i}\beta^{-i}$ mit $N_- \leq N < N_+$. Dann definieren wir die folgende Rundungsvorschrift:*

$$rd_t(x) := \begin{cases} \sigma\beta^N \sum_{i=1}^{t} x_{-i}\beta^{-i} & , \text{ falls } x_{-t-1} < 0.5\ \beta \\ \sigma\beta^N \left(\sum_{i=1}^{t} x_{-i}\beta^{-i} + \beta^{-t}\right) & , \text{ falls } x_{-t-1} \geq 0.5\ \beta \end{cases}$$

*$rd_t(x)$ heißt der auf $t$ Stellen gerundete Wert von $x$.*

Im Fall des Dezimalsystems $\beta = 10$ entspricht die Vorschrift in Def. 1.20 der üblicherweise als „Runden“ bezeichneten Regel.

**Satz 1.21** *Es sei $\beta \in \mathbb{N}$ gerade, $t \in \mathbb{N}$ und $x \in \mathbb{R}\backslash\{0\}$ besitze die Darstellung wie in Definition 1.20. Dann gilt:*

*(i)* $rd_t(x)$ *gehört zu* $M(\beta, t, N_-, N_+)$*, d.h. hat eine Darstellung der Gestalt* $rd_t(x) = \sigma\beta^{\tilde{N}} \sum_{i=1}^{t} \tilde{x}_{-i}\beta^{-i}$*;*

*(ii) für den absoluten Fehler gilt:* $|rd_t(x) - x| \leq 0.5\,\beta^{N-t}$*;*

*(iii) für die relativen Fehler gelten die Abschätzungen:*

$$\left|\frac{rd_t(x) - x}{x}\right| \leq 0.5\ \beta^{-t+1}, \quad \left|\frac{rd_t(x) - x}{rd_t(x)}\right| \leq 0.5\ \beta^{-t+1}$$

*(iv)* $rd_t(x)$ *erlaubt die Darstellung* $rd_t(x) = x(1 + \varepsilon(x)) = \frac{x}{1-\eta(x)}$*, wobei* $\max\{|\varepsilon(x)|, |\eta(x)|\} \leq 0.5\,\beta^{-t+1}$*.*

*Deshalb heißt die Zahl* $\tau := 0.5\,\beta^{-t+1}$ *die relative Rechengenauigkeit der t-stelligen Gleitkommaarithmetik.*

Beweis:

(i) Es genügt, den Fall $x_{-t-1} \geq 0.5\beta$ zu betrachten. Wir unterscheiden 2 Fälle:

(1) $\exists i \in \{1, \ldots, t\} : x_{-i} < \beta - 1$.
Wir definieren $\ell := \max\{i \in \{1, \ldots, t\} : x_{-i} < \beta - 1\}$ und setzen
$\tilde{N} := N, \tilde{x}_{-i} := x_{-i}, i = 1, \ldots, \ell - 1, \tilde{x}_{-\ell} := x_{-\ell} + 1, \tilde{x}_{-i} := 0, i = \ell + 1, \ldots, t.$

(2) $x_{-i} = \beta - 1, \forall i = 1, \ldots, t.$
Hier setzen wir $\tilde{N} := N + 1, \tilde{x}_{-1} := 1, \tilde{x}_{-i} := 0, i = 2, \ldots, t.$

(ii) Für $x_{-t-1} < 0.5\beta$ gilt:

$$\begin{aligned} -\sigma(rd_t(x) - x) &= \beta^N \sum_{i=t+1}^{\infty} x_{-i}\beta^{-i} = \beta^{N-t-1}x_{-t-1} + \beta^N \sum_{i=t+2}^{\infty} x_{-i}\beta^{-i} \\ &\leq \beta^{N-t-1}(0.5\beta - 1) + \beta^{N-t-1} = 0.5\beta^{N-t} \end{aligned}$$

Für $x_{-t-1} \geq 0.5\beta$ ergibt sich:

$$\begin{aligned} \sigma(rd_t(x) - x) &= \beta^{N-t} - \beta^{N-t-1}x_{-t-1} - \beta^N \sum_{i=t+2}^{\infty} x_{-i}\beta^{-i} \\ &= \beta^{N-t-1}(\beta - x_{-t-1}) \leq 0.5\beta^{N-t} \end{aligned}$$

Außerdem folgt aus $\beta^{N-t-1}(\beta - x_{-t-1}) \geq \beta^{N-t-1} > \beta^N \sum_{i=t+2}^{\infty} x_{-i}\beta^{-i}$
auch $\sigma(rd_t(x) - x) > 0$.
In beiden Fällen folgt die Abschätzung für $|rd_t(x) - x|$.

(iii) Wegen $x_{-1} \geq 1$ folgt $|x| \geq \beta^{N-1}$ und gemeinsam mit (ii)

$$\left|\frac{rd_t(x) - x}{x}\right| \leq 0.5\beta^{N-t} \cdot \beta^{1-N} = 0.5\beta^{1-t}.$$

Die andere Abschätzung ergibt sich analog, da aus der Rundungsvorschrift $|rd_t(x)| \geq x_{-1}\beta^{N-1} \geq \beta^{N-1}$ folgt.

(iv) Hierbei handelt es sich nur um eine andere Schreibweise des Ergebnisses aus (iii). Man setzt einfach $\varepsilon(x) := \frac{rd_t(x)-x}{x}$ und $\eta(x) := \frac{rd_t(x)-x}{rd_t(x)}$. Dann gelten die angegebenen Darstellungen und Abschätzungen. □

Wir kommen nun zur Verknüpfung von Computerzahlen. Dazu bezeichne $\square$ eine der Rechenoperationen $+, -, *, /$. Wenn nun $x$ und $y$ zwei Computerzahlen mit Mantissenlänge $t$ sind, so ist $x \square y$ i. a. nicht mit $t$-stelliger Mantisse darstellbar. Es muß also i. a. gerundet werden. Deshalb werden die elementaren Rechenoperationen $\square$ auf Computern in 2 Schritten ausgeführt:

(i) Möglichst genaue Berechnung von $x \square y$;

(ii) Runden des Ergebnisses auf $t$ Stellen.

Das Ergebnis dieser Operation werde mit $fl_t(x \square y)$ bezeichnet.

**Postulat 1.22** *Die Computer-Arithmetik ist so organisiert, daß für zwei Computerzahlen $x$ und $y$ mit Mantissenlänge $t$ stets gilt:*

$$fl_t(x \square y) = rd_t(x \square y)$$

Wir nehmen im folgenden stets an, daß das Postulat erfüllt ist. Aus Satz 1.21(iv) ergibt sich dann

$$fl_t(x \square y) = (x \square y)(1 + \varepsilon) = \frac{(x \square y)}{1 - \eta}, \ |\varepsilon|, |\eta| \leq \tau.$$

Schreibt man $\tilde{x} = x(1 + \varepsilon), \tilde{y} = y(1 + \varepsilon)$, so bedeutet dies

$$fl_t(x \pm y) = \tilde{x} \pm \tilde{y}, \ fl_t(x * y) = \tilde{x} * y = x * \tilde{y}$$

usw., d. h. das Computerresultat einer arithmetischen Operation ist das exakte Resultat derselben Operation mit wenig gestörten Operanden.

**Bemerkung 1.23** *(Probleme der Computer-Arithmetik)*

| | |
|---|---|
| *Überlauf:* | *Falls der Betrag des Ergebnisses einer Rechenoperation grösser als $\beta^{N_+}$ ist, kann das Ergebnis nicht mehr dargestellt werden. Je nach Hardware wird dann entweder die Rechnung abgebrochen oder mit fiktiven Zahlen weitergerechnet.* |
| *Unterlauf:* | *Falls der Betrag des Ergebnisses einer Rechenoperation kleiner als $\beta^{-N_- - 1}$ ist, wird häufig mit 0 gerundet. Der relative Fehler des Resultats beträgt damit 100%!* |

*Auslöschung:* *Addition etwa gleichgroßer Zahlen mit entgegengesetztem Vorzeichen führt zu einer starken Verringerung der Zahl der gültigen Ziffern, so daß große Fehler die Folge sind (vgl. Beipsiele 0.1 und 1.24).*

*Rechenregeln:* *Selbst in den Fällen, wo weder Über- noch Unterlauf eintritt, gelten die Rechenregeln der reellen Zahlen nicht mehr. Zwar sind wegen des Postulats Addition und Multiplikation kommutativ, jedoch gelten Assoziativ- und Distributivgesetz nicht mehr.*

**Beispiel 1.24** *(Auslöschung)*
*Die Arithmetik der Addition im Dezimalsystem sei wie folgt organisiert. Es seien* $x = \sigma_1 m_1 \cdot 10^{N_1}$ *und* $y = \sigma_2 m_2 \cdot 10^{N_2}$ *mit* $0 \leq m_1, m_2 < 1$ *und* $N_2 \leq N_1$ *zwei Dezimalzahlen mit Mantissenlänge* $t$*. Beide Zahlen werden als Computerzahlen mit Mantissenlänge* $2t$ *und gleichen Exponenten dargestellt und dann addiert (Zwischenspeicherung mit doppelter Genauigkeit). Das Ergebnis wird anschließend normalisiert, so daß für Mantisse* $m$ *der Summe* $0 \leq m \leq 1$ *gilt. Danach folgt die Rundung auf* $t$ *Stellen. Für dieses Vorgehen ist das Postulat 1.22 erfüllt !*

*Bsp.:* $\beta := 10,\ t := 3,\ \tilde{x} := 0.9995 \cdot 10^0,\ \tilde{y} := -0.9984 \cdot 10^0$
*Ziel:* *Berechnung von* $\tilde{x} + \tilde{y}$ *in der obigen Computerarithmetik.*

| | |
|---|---|
| *0. Rundung:* | $x = rd_t(\tilde{x}) = 0.100 \cdot 10^1$ |
| | $y = rd_t(\tilde{y}) = -0.998 \cdot 10^0$ |
| *1. Darstellung* | *mit Mantissenlänge* $2t$*:* |
| | $x = \ 0.100000 \cdot 10^1$ |
| | $y = -0.099800 \cdot 10^1$ |
| *2. Addition:* | $0.000200 \cdot 10^1$ |
| *3. Normalisierung:* | $0.200000 \cdot 10^{-2}$ |
| *4. Rundung:* | $0.200 \cdot 10^{-2}$ |

*Also gilt:*

$$\begin{aligned} fl_t(x+y) &= fl_t(rd_t(\tilde{x}) + rd_t(\tilde{y})) = 0.200 \cdot 10^{-2} \\ \tilde{x} + \tilde{y} &= 0.0011 = 0.110 \cdot 10^{-2} \\ \rightsquigarrow fl_t(rd_t(\tilde{x}) + rd_t(\tilde{y})) &= (\tilde{x} + \tilde{y}) \cdot \tfrac{20}{11} = (\tilde{x} + \tilde{y})(1 + 0.818) \end{aligned}$$

*Der Betrag des relativen Fehlers beträgt also 81,8% !*
*Ursache: Ungünstige Fortpflanzung der Rundungsfehler !*

## 1.4 Rundungsfehlerfortpflanzung

Gegeben sei wieder eine Aufgabe bzw. Vorschrift $P$, die auf einer Datenmenge $D$ definiert ist und jetzt zusätzlich ein Algorithmus $P_A$ zur numerischen Lösung von $P$. Wir nehmen nun an, daß sich $P_A$ aus einer endlichen Folge $P_1, \ldots, P_N$ von elementaren Schritten zusammensetzt, d.h.

$$P_A = P_N \circ \cdots \circ P_2 \circ P_1.$$

Bisher kennen wir nur die Rundungsfehlerauswirkungen bei elementaren Schritten und fragen uns schließlich nach der Fortpflanzung des Rundungsfehlers im gesamten Algorithmus. Die folgenden Eigenschaften charakterisieren nun allgemein sein „freundliches“ Verhalten in dieser Hinsicht.

**Definition 1.25** *Es sei $\tau$ die relative Rechengenauigkeit der t-stelligen Gleitkomma-Arithmetik.*

*(i) Der Algorithmus $P_A$ heißt numerisch stabil für $P$ auf $D$, falls eine Konstante $F_s > 0$ existiert, so daß für alle $d \in D$ gilt*

$$\|P(d) - P_A(d)\| \leq F_s \|d\| \tau \,.$$

*(ii) Der Algorithmus $P_A$ heißt numerisch gutartig für $P$ auf $D$, falls eine Konstante $F_g > 0$ und für jedes $d \in D$ eine "Störung" $\delta_d$ mit $d + \delta_d \in D$ und $\|\delta_d\| \leq F_g \tau \|d\|$ existiert, so daß*

$$P_A(d) = P(d + \delta_d) \,.$$

Es wird sich zeigen, daß die numerische Gutartigkeit die bestmögliche Eigenschaft eines numerischen Algorithmus ist, weil die in den Eingangsdaten enthaltenen Informationen bestmöglich auf die Ergebnisse übertragen werden. Numerische Stabilität ist die Mindestanforderung an ein vernünftiges Verfahren zur Lösung eines korrekt gestellten Problems. Natürlich wird man sich in beiden Fällen wünschen, daß die Konstanten $F_s$ und $F_g$ nicht "zu groß" sind.

**Lemma 1.26** *Es sei $P$ korrekt gestellt auf $D$, d.h. es existiert eine Funktion $L : D \times \mathbb{R}_+ \to \mathbb{R}_+$ mit der Eigenschaft, daß für alle $d$, $\tilde{d} \in D$ gilt*

$$\|P(d) - P(\tilde{d})\| \leq L(d, \|d - \tilde{d}\|) \|d - \tilde{d}\| \,.$$

*Überdies sei $L(d, \cdot) : \mathbb{R}_+ \to \mathbb{R}_+$ monoton wachsend und es gelte $\sup_{d \in D} L(d, C\|d\|) < +\infty$ für jedes $C > 0$.*
*Ist ein Algorithmus $P_A$ numerisch gutartig für $P$ auf $D$, so ist er auch numerisch stabil mit $F_s := F_g \sup_{d \in D} L(d, F_g \tau \|d\|)$.*

Beweis: Es sei $d \in D$ beliebig gewählt. Wir wählen die Konstante $F_g$ und die Störung $\delta_d$ mit $d + \delta_d \in D$, $\|\delta_d\| \leq F_g \tau \|d\|$ und $P_A(d) = P(d + \delta_d)$.
Dann folgt aus der korrekten Gestelltheit von $P$, daß

$$\begin{aligned} \|P(d) - P_A(d)\| &= \|P(d) - P(d + \delta_d)\| \leq L(d, F_g \tau \|d\|) F_g \tau \|d\| \\ &\leq F_g \sup_{d \in D} \{L(d, F_g \tau \|d\|)\} \|d\| \tau \,, \end{aligned}$$

d.h. die gewünschte Aussage. □

Für die Erfülltheit der zusätzlichen Voraussetzung in Lemma 1.26 ist hinreichend, daß $P$ global Lipschitzstetig ist oder in vielen Fällen, daß die Menge $D$ beschränkt ist.
Ehe wir später in Kap. 2 die numerische Gutartigkeit des Gaußschen Algorithmus untersuchen, betrachten wir hier das folgende einfache, aber instruktive Beispiel. Dabei werden wir stets voraussetzen, daß weder Über- noch Unterlauf auftritt.

**Beispiel 1.27** *Gegeben: $n$ reelle Zahlen $x_1, \ldots, x_n$; Arithmetik mit Mantissenlänge $t$.*
*Gesucht: $S := P(x) := \sum_{i=1}^{n} x_i$, wobei $x \in \mathbb{R}^n$*
*Konzeptioneller Algorithmus: $S := 0$, $S := S + x_i$, $i = 1, \ldots, n$.*
*Realer Algorithmus: $S_0 := 0$, $s_i := fl_t(s_{i-1} + rd_t(x_i))$, $i = 1, \ldots, n$, d.h. $P_A(x) := s_n$.*

*Behauptung: $P$ ist korrekt gestellt und der Algorithmus $P_A$ ist numerisch stabil auf $D = \mathbb{R}^n$ mit $\|\cdot\|_1$ und*

$$L \equiv 1 \quad \textit{und} \quad F_s := \frac{n}{1-(n-1)\tau} \approx n\,,$$

*falls $(n-1)\tau < 1$ vorausgesetzt wird.*
*Er ist sogar numerisch gutartig auf $D = \mathbb{R}^n$ mit $\|\cdot\|_1$ und*

$$F_g = n + O(\tau) \approx F_s\,.$$

*Beweis: Es gilt $|P(x) - P(\tilde{x})| \le \|x - \tilde{x}\|_1$, d.h. $L \equiv 1$.*
*Nach Postulat 1.22 folgt für bel. $i \in \{1, \ldots, n\}$ :*

$$\begin{aligned}
s_i &= (s_{i-1} + rd_t(x_i))(1+\varepsilon_i) \text{ mit } |\varepsilon_i| \le \tau \\
&= s_{i-1}(1+\varepsilon_i) + rd_t(x_i)(1+\varepsilon_i) \\
&= \sum_{j=1}^{i} rd_t(x_j) \prod_{k=j}^{i}(1+\varepsilon_k) \\
\rightsquigarrow s_n &= \sum_{j=1}^{n} rd_t(x_j) \prod_{k=j}^{n}(1+\varepsilon_k) \\
&= \sum_{j=1}^{n} x_j \prod_{k=j}^{n}(1+\varepsilon_k) + \sum_{j=1}^{n}(rd_t(x_j) - x_j) \prod_{k=j}^{n}(1+\varepsilon_k) \\
\overset{1.21}{\rightsquigarrow} |S - s_n| &\le \sum_{j=1}^{n} |x_j| \left\{ \prod_{k=j}^{n}(1+|\varepsilon_k|) - 1 \right\} + \sum_{j=1}^{n} \tau |x_j| \prod_{k=j}^{n}(1+|\varepsilon_k|)
\end{aligned}$$

*Wegen $s_0 = 0$ gilt $\varepsilon_1 = 0$ und folglich*

$$\begin{aligned}
\prod_{k=j}^{n}(1+|\varepsilon_k|) &\le \prod_{k=2}^{n}(1+|\varepsilon_k|) = 1 + \sum_{k=2}^{n}|\varepsilon_k| + \cdots + |\varepsilon_2|\cdots|\varepsilon_n| \\
&\le 1 + (n-1)\tau + ((n-1)\tau)^2 + \cdots + ((n-1)\tau)^{n-1} \\
&= 1 + (n-1)\tau \sum_{i=0}^{n-2}((n-1)\tau)^i \le 1 + \tfrac{(n-1)\tau}{1-(n-1)\tau} \\
\rightsquigarrow |S - s_n| &\le \tfrac{(n-1)\tau}{1-(n-1)\tau} \sum_{j=1}^{n} |x_j| + \tfrac{\tau}{1-(n-1)\tau} \sum_{j=1}^{n} |x_j| \\
&= \left\{ \tfrac{n}{1-(n-1)\tau} \sum_{j=1}^{n} |x_j| \right\} \tau = \tfrac{n}{1-(n-1)\tau} \|x\|_1 \tau\,.
\end{aligned}$$

*Überdies kann die Summe $s_n = P_A(x)$ als die exakte Summe mit gestörten Eingangsdaten dargestellt werden kann. Es gilt nämlich*

$$P_A(x) = \sum_{i=1}^{n} rd_t(x_i) \prod_{k=i}^{n}(1+\varepsilon_k) = \sum_{i=1}^{n} rd_t(x_i)(1+\varepsilon_i^{(n)}),$$

*wobei* $|\varepsilon_i^{(n)}| = |\prod_{k=i}^{n}(1+\varepsilon_k) - 1| \leq \sum_{k=i}^{n} |\varepsilon_k| + O(\tau^2) \leq (n-i)\tau + O(\tau^2)$.

*Wir wählen nun die "Störung" $\delta_x$ von $x$ durch die Festlegung*

$$\delta_{x,i} := rd_t(x_i)(1+\varepsilon_i^{(n)}) - x_i = x_i(1+\hat{\varepsilon}_i)(1+\varepsilon_i^{(n)}) - x_i = x_i(\hat{\varepsilon}_i + \varepsilon_i^{(n)} + \hat{\varepsilon}_i\varepsilon_i^{(n)}),$$

*wobei* $|\hat{\varepsilon}_i| \leq \tau$, $i = 1, \ldots, n$. *Dann gilt*

$$P(x+\delta_x) = P_A(x) \quad \textit{und} \quad \|\delta_x\|_1 \leq \sum_{i=1}^{n} |x_i|(|\hat{\varepsilon}_i| + |\varepsilon_i^{(n)}| + O(\tau^2)) \leq \|x\|_1(n\tau + O(\tau^2)).$$

□

**Bemerkung 1.28** *Die Abschätzung in Beispiel 1.27 liefert*

$$\frac{|S - s_n|}{\sum_{i=1}^{n} |x_i|} \leq \frac{n}{1-(n-1)\tau}\tau \approx n\tau,$$

*da in der Regel $(n-1)\tau \ll 1$ gilt. Besitzen nun alle $x_i$, $i = 1, \ldots, n$, gleiches Vorzeichen, so gilt $|S| = \sum_{i=1}^{n} |x_i|$ und damit*

$$\frac{|S - s_n|}{|S|} \leq \frac{n}{1-(n-1)\tau}\tau \approx n\tau.$$

*D. h., in diesem Fall multipliziert sich der relative Fehler $\tau$ eines elementaren Additionsschrittes hin zum relativen Fehler bei der Berechnung der Summe mit $n$. Diese Abschätzung reflektiert den „schlechtesten Fall" (worst case), ist aber durchaus realistisch. Man kann zeigen, daß andere Summationsalgorithmen (z. B. wenn die $x_i$ in betragsmäßig wachsender Folge summiert werden oder im Fall $n = 2^k$ die sog. Kaskadensummation angewendet wird) zu besseren Schranken führen.*

Für andere Beispiele der Rundungsfehleranalyse einfacher Algorithmen, wie z.B. die Berechnung des Skalarproduktes, sei auf Kap. 2.3 in Kiełbasiński/Schwetlick verwiesen.

# 2 Numerische Lösung linearer Gleichungssysteme

Wir werden hier fast ausschließlich lineare Gleichungssysteme (GS) mit quadratischer, regulärer Koeffizientenmatrix betrachten, so daß diese Systeme für beliebige rechte Seiten stets genau eine Lösung besitzen. In den Anwendungen ist es zweckmäßig, Gleichungssysteme nach speziellen (analytisch, algebraisch) Eigenschaften, nach ihrer „Größe" sowie ihrer Struktur zu unterscheiden. Diese Unterscheidungen betreffen etwa

(i) normale, vollbesetzte Matrix,

(ii) symmetrische bzw. symmetrische und positiv definite Matrix,

(iii) sehr große Matrix, die viele Nullen enthält, mit den Spezialfällen:
Bandmatrix, sehr wenige irregulär verteilte Nicht-Nullelemente.

Die Lösungsverfahren unterscheiden sich je nach Situation (i), (ii) bzw. (iii). Wir werden uns im folgenden mit Verfahren beschäftigen, die allgemein anwendbar, aber im wesentlichen auf den Fall (i) orientiert sind.
Wir beginnen mit einer Fehleranalyse und der Kondition linearer Gleichungssysteme.

## 2.1 Kondition linearer Gleichungssysteme

Wir betrachten das lineare Gleichungssystem

$$Ax = b, \quad A \in \mathbb{R}^{n\times n}, \quad b \in \mathbb{R}^n$$

als gegeben und erinnern uns zunächst an einige Fakten aus der linearen Algebra.
Wir bezeichnen mit $\mathrm{rg}(A)$ den Rang der Matrix $A$, d.h., die Dimension des Wertebereichs $R(A) := \{Ax : x \in \mathbb{R}^n\}$ bzw. die maximale Anzahl linear unabhängiger Zeilen bzw. Spalten. Eine andere Formulierung dieser Definition von $\mathrm{rg}(A)$ ist, daß das homogene lineare Gleichungssystem $Ax = \theta$ gerade $n - \mathrm{rg}(A)$ linear unabhängige Lösungen besitzt. Man nennt $A$ eine reguläre Matrix, falls $rg(A) = n$. Ist $A$ regulär, so ist das lineare Gleichungssystem $Ax = b$ für jede rechte Seite $b$ eindeutig lösbar und es existiert eine Matrix $A^{-1} \in \mathbb{R}^{n\times n}$, so daß $A^{-1}b$ diese eindeutig bestimmte Lösung darstellt. $A^{-1}$ ist die zu $A$ inverse Matrix.
Zusätzlich betrachten wir das „gestörte“ lineare Gleichungssystem

$$(A + \Delta A)(x + \Delta x) = b + \Delta b,$$

wobei $\Delta A$ und $\Delta b$ Fehler in den Daten $A$ bzw. $b$ darstellen und $\Delta x$ den entstehenden Fehler der Lösung bezeichnet. Uns interessieren Abschätzungen für den absoluten bzw. relativen Fehler $\|\Delta x\|$ bzw. $\frac{\|\Delta x\|}{\|x\|}$ (mit einer Norm $\|\cdot\|$ im $\mathbb{R}^n$) im Sinne von Definition 1.12.
Als ersten Schritt einer theoretischen Analyse kümmern wir uns zunächst um die Regularität gestörter regulärer Matrizen.

**Lemma 2.1** *Es sei $A \in \mathbb{R}^{n\times n}$ und für eine zugeordnete Matrixnorm gelte $\|A\| < 1$. Dann ist die Matrix $I + A$ regulär und es gilt:*

$$\frac{1}{1 + \|A\|} \le \|(I + A)^{-1}\| \le \frac{1}{1 - \|A\|}.$$

*(Hierbei bezeichnet $I$ die Einheitsmatrix in $\mathbb{R}^{n\times n}$.)*

Beweis:
Für beliebiges $x \in \mathbb{R}^n$ gilt:

$$\|(I + A)x\| = \|x + Ax\| \ge \|x\| - \|Ax\| \ge (1 - \|A\|)\|x\| \ge 0$$

Also folgt aus $(I+A)x = \Theta$ sofort $x = \Theta$. Also gilt $rg(I+A) = n$ und es existiert $(I+A)^{-1} =: C \in I\!R^{n\times n}$. Für die Matrix $C$ gilt:

$$I = (I+A)C = C(I+A)$$

$$\rightsquigarrow \quad \|I\| = 1 \le \|(I+A)\|\|C\| \le (1+\|A\|)\|C\| \quad \text{(nach Lemma 1.5)}$$

$$\|I\| = 1 = \|C + AC\| \ge \|C\| - \|AC\| \ge (1-\|A\|)\|C\|.$$

Damit ist die behauptete Ungleichungskette bewiesen. □

**Satz 2.2** *(„Störungslemma")*
*Es seien $A, \tilde{A} \in I\!R^{n\times n}$ mit $A$ regulär und für die zugeordnete Matrixnorm gelte $\|A^{-1}\| \le \alpha, \|A - \tilde{A}\| \le \beta$ und $\alpha\beta < 1$. Dann ist auch $\tilde{A}$ regulär und es gilt*

$$\|\tilde{A}^{-1}\| \le \frac{\alpha}{1-\alpha\beta} \quad , \quad \|A^{-1} - \tilde{A}^{-1}\| \le \frac{\alpha^2\beta}{1-\alpha\beta}\,.$$

Beweis:
Nach Lemma 1.5 gilt: $\|A^{-1}(\tilde{A} - A)\| \le \|A^{-1}\|\|\tilde{A} - A\| \le \alpha\beta < 1$. Deshalb ist nach Lemma 2.1 die Matrix $I + A^{-1}(\tilde{A} - A) = A^{-1}\tilde{A}$ regulär, also auch $\tilde{A}$. Außerdem folgt aus Lemma 2.1:

$$\|\tilde{A}^{-1}\| \le \|\tilde{A}^{-1}A\|\,\|A^{-1}\| = \|(I + A^{-1}(\tilde{A} - A))^{-1}\|\,\|A^{-1}\| \le \frac{\|A^{-1}\|}{1-\alpha\beta}$$

$$\|A^{-1} - \tilde{A}^{-1}\| = \|A^{-1}(\tilde{A} - A)\tilde{A}^{-1}\| \le \|A^{-1}\|\,\|\tilde{A} - A\|\,\|\tilde{A}^{-1}\| \le \frac{\|A^{-1}\|^2\beta}{1-\alpha\beta}\,.$$

□

Als eine erste Schlußfolgerung aus dem Störungslemma untersuchen wir die Kondition der Aufgabenstellung aus einer regulären Matrix $A$ die inverse Matrix $A^{-1}$ zu berechnen.

**Folgerung 2.3** *Die Aufgabenstellung, aus einer regulären Matrix $A$ die inverse Matrix $A^{-1}$ zu berechnen, ist korrekt gestellt und für die absoluten bzw. relativen Konditionszahlen (in $A$) gilt:*

$$L(A, \Delta_A) = \frac{\|A^{-1}\|^2}{1 - \|A^{-1}\|\Delta_A}, \quad K(A, \Delta_A) = \frac{\|A^{-1}\|\ \|A\|}{1 - \|A^{-1}\|\Delta_A}.$$

*Überdies ist die Menge $\{A \in I\!R^{n\times n} : A$ ist regulär$\}$ offen im Raum $(I\!R^{n\times n}, \|\cdot\|)$.*

Beweis:
Mit $\Delta_A = \beta$ folgt die Aussage über die absolute Konditionszahl sofort aus Satz 2.2. Die Aussage über die relative Konditionszahl folgt sofort aus der Beziehung zwischen beiden Konditionszahlen (vgl. 1.13):

$$K(A, \Delta_A) = L(A, \Delta_A)\frac{\|A\|}{\|A^{-1}\|}\,.$$

Die letzte Aussage ist eine direkte Konsequenz aus Satz 2.2. □

Die Bedeutung des erstmals in Folgerung 2.3 auftauchenden Terms $\|A\|\ \|A^{-1}\|$ wird noch deutlicher im folgenden Störungsresultat für die Lösungen linearer Gleichungssysteme.

**Satz 2.4** *Es seien $A$ und $\Delta A$ Matrizen aus $\mathbb{R}^{n\times n}$, $A$ sei regulär, $b$ und $\Delta b$ seien aus $\mathbb{R}^n$ und $x = A^{-1}b$. Ferner gelte für die zu einer Norm $\|\cdot\|$ auf dem $\mathbb{R}^n$ zugeordnete Matrixnorm $\|A^{-1}\|\ \|\Delta A\| < 1$.*
*Dann existiert eine Lösung $x + \Delta x$ des linearen Gleichungssystems*

$$(A + \Delta A)(x + \Delta x) = b + \Delta b,$$

*die Aufgabe ist korrekt gestellt und es gelten die Abschätzungen*

$$\|\Delta x\| \le \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\ \|\Delta A\|}\|\Delta b\|,\ \textit{falls } b = \Theta,$$

$$\frac{\|\Delta x\|}{\|x\|} \le \frac{\|A^{-1}\|\ \|A\|}{1 - \|A^{-1}\|\ \|A\|\frac{\|\Delta A\|}{\|A\|}}\left(\frac{\|\Delta b\|}{\|b\|} + \frac{\|\Delta A\|}{\|A\|}\right),\ \textit{falls } b \neq \theta.$$

Beweis:
Die Voraussetzung $\|A^{-1}\|\ \|\Delta A\| < 1$ impliziert nach Satz 2.2, daß die Matrix $A + \Delta A$ regulär ist und daß

$$\|(A + \Delta A)^{-1}\| \le \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\ \|\Delta A\|}.$$

Aus der Gleichheit $(A + \Delta A)(x + \Delta x) = b + \Delta b$ folgt
$$(A + \Delta A)\Delta x = b + \Delta b - Ax - \Delta Ax = \Delta b - \Delta Ax$$

und deshalb

$$\begin{aligned} \Delta x &= (A + \Delta A)^{-1}(\Delta b - \Delta Ax) \\ \rightsquigarrow\ \|\Delta x\| &\le \tfrac{\|A^{-1}\|}{1-\|A^{-1}\|\ \|\Delta A\|}(\|\Delta b\| + \|\Delta A\|\ \|x\|). \end{aligned}$$

Im Fall $b = \Theta$ gilt auch $x = \Theta$ und alles ist gezeigt. Es sei $b \neq \Theta$.

$$\begin{aligned} \rightsquigarrow\ \tfrac{\|\Delta x\|}{\|x\|} &\le \tfrac{\|A^{-1}\|}{1-\|A^{-1}\|\ \|\Delta A\|}\left(\tfrac{\|\Delta b\|}{\|b\|}\tfrac{\|b\|}{\|x\|} + \tfrac{\|\Delta A\|}{\|A\|}\|A\|\right) \\ &\le \tfrac{\|A^{-1}\|\|A\|}{1-\|A^{-1}\|\ \|\Delta A\|}\left(\tfrac{\|\Delta b\|}{\|b\|} + \tfrac{\|\Delta A\|}{\|A\|}\right). \end{aligned}$$

□

**Definition 2.5** *Die Zahl $cond\,(A) := \|A\|\ \|A^{-1}\|$ heißt Konditionszahl der regulären Matrix $A$ (bzgl. der Matrixnorm $\|\cdot\|$).*

**Bemerkung 2.6** *Die Abschätzung für den relativen Fehler* $\frac{\|\Delta x\|}{\|x\|}$ *in Satz 2.4 kann i. a. nicht verbessert werden. Sie besagt, daß der relative Fehler der Lösungen abgeschätzt werden kann durch das Produkt eines Faktors mit der Summe der relativen Fehler der Eingangsdaten. Für eine „kleine Störung“* $\Delta A$ *ist dieser Faktor etwa gleichgroß mit* $cond\,(A)$.
*Die relativen Eingangsfehler multiplizieren sich also mit* $cond\,(A)$. *Analog ist die Situation auch in Folgerung 2.3.*
*Große Konditionszahlen führen also i. a. dazu, daß aus kleinen Eingabefehlern große Fehler bei den Lösungen resultieren! Offensichtlich hängt die Konditionszahl* $cond\,(A)$ *von der konkreten Wahl der Norm ab. Bei Verwendung von* $\|\cdot\|_p$ *schreiben wir* $cond_p(A)$.

**Beispiel 2.7** *Es sei* $H_n$ *die Hilbert-Matrix der Ordnung* $n$ *mit den Elementen*

$$a_{ij} = \frac{1}{i+j-1}, \quad i,j = 1,\ldots,n, n \in \mathbb{N}.$$

$H_n$ *ist symmetrisch und positiv definit, also auch regulär, mit der ganzzahligen Inversen* $H_n^{-1} = (h_{ij})$, $h_{ij} = \frac{(-1)^{i+j}}{i+j-1} r_i r_j$ *mit* $r_i := \frac{(n+i-1)!}{((i-1)!)^2 (n-i)!}$, $i,j = 1,\ldots,n$.
*Für* $n = 4$ *gilt zum Beispiel*

$$H_4 = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \end{pmatrix} \qquad \textit{und} \qquad H_4^{-1} = \begin{pmatrix} 16 & -120 & 240 & -140 \\ -120 & 1200 & -2700 & 1680 \\ 240 & -2700 & 6480 & -4200 \\ -140 & 1680 & -4200 & 2800 \end{pmatrix}$$

*Wird nun das lineare Gleichungssystem* $H_n x = b := H_n(1,1,\ldots,1)^T$ *für verschiedene* $n \in \mathbb{N}$ *mit dem Gaußschen Algorithmus gelöst, so ergeben sich für* $n = 8$ *bzw.* $n = 10$ *relative Fehler der Lösungen von etwa* 0.4 *bzw.* 3.4. *Dies Effekt nimmt mit wachsendem* $n$ *weiter zu (vgl. Hämmerlin/Hoffmann, Kap. 2.6).*
*Man erahnt, daß die Ursache für die aufgetretenen Fehler sich auch in den Konditionszahlen von* $H_n$ *manifestiert. Es gilt nämlich:*

| $n$ | 3 | 4 | 5 | 10 |
|---|---|---|---|---|
| $cond_2(H_n)$ | 520 | 16.000 | 480.000 | $1.6 \cdot 10^{13}$ |

*wobei* $cond_2(H_n) = \frac{\lambda_{\max}(H_n)}{\lambda_{\min}(H_n)}$ *(vgl. auch 1.6c)).*

## 2.2 Der Gaußsche Algorithmus

Die Aufgabe besteht in der Lösung des linearen Gleichungssystems $Ax = b$ mit regulärer Koeffizientenmatrix $A$. Der Gaußsche Algorithmus basiert auf der Beobachtung, daß die beiden folgenden Operationen die Lösung des linearen Gleichungssystems nicht verändern, wohl aber die Struktur von $A$:

- Multiplikation einer Zeile des linearen Gleichungssystems mit einem Faktor verschieden von 0,
- Addition einer Zeile des linearen Gleichungssystems zu einer anderen Zeile.

Ziel ist es, mit diesen Operationen aus $A$ eine reguläre Dreiecksmatrix $R$ zu erzeugen. Die Grundform des Gaußschen Algorithmus hat die folgende formale Beschreibung:

- $A^{(1)} := A,\ b^{(1)} := b$;
- für $k = 1, \ldots, n-1$:
  - finde einen Index $s(k) \in \{k, k+1, \ldots, n\}$ mit $a^{(k)}_{s(k),k} \neq 0$ und vertausche die Zeilen $k$ und $s(k)$ in $A^{(k)}$ und $b^{(k)}$, und bezeichne die Elemente wie vorher;
  - $a^{(k+1)}_{ij} := \begin{cases} 0 & , \quad i = k+1, \ldots, n, j = k \\ a^{(k)}_{ij} - \frac{a^{(k)}_{ik}}{a^{(k)}_{kk}} a^{(k)}_{kj} & , \quad i, j = k+1, \ldots, n \\ a^{(k)}_{ij} & , \quad \text{sonst} \end{cases}$
  - $b^{(k+1)}_{i} := \begin{cases} b^{(k)}_{i} - \frac{a^{(k)}_{ik}}{a^{(k)}_{kk}} b^{(k)}_{k} & , \quad i = k+1, \ldots, n \\ b^{(k)}_{i} & , \quad \text{sonst} \end{cases}$
- $R := A^{(n)},\ c := b^{(n)}$.

Unser nächstes Ziel besteht darin, diesen Algorithmus in Form von Matrixprodukten zu schreiben. Dadurch wird er für uns für eine Analyse zugänglicher. Dazu führen wir zunächst zwei Typen von (Transformations-) Matrizen ein:

$$P_{k,s(k)} := \begin{array}{l} (e^1, \ldots, e^{k-1}, \quad e^{s(k)}, e^{k+1}, \ldots, e^{s(k)-1}, \quad e^k, e^{s(k)+1}, \ldots, e^n) \\ \qquad\qquad\qquad\qquad\quad \uparrow \qquad\qquad\qquad\qquad\quad \uparrow \\ \qquad\qquad\qquad\qquad\quad \text{Spalte k} \qquad\qquad\qquad\quad \text{Spalte s(k)} \end{array}$$

wobei $e^i := (0, \ldots, 0, 1, 0, \ldots, 0)^T$ der $i$-te kanonische Einheitsvektor ist. $P_{k,s(k)}$ entsteht also aus der identischen Matrix I durch Vertauschung der $k$-ten mit der $s(k)$-ten Spalte. Sie bewirkt bei Multiplikation mit einer Matrix die Vertauschung der $k$-ten mit der $s(k)$-ten Zeile dieser Matrix. Man nennt sie auch *Vertauschungsmatrix*.
Der zweite Typ ist die folgende elementare Transformationsmatrix:

$$L_{ij}(\beta) := I + \beta \begin{pmatrix} 0 & \cdots & 0 & \cdots & 0 \\ \vdots & & & & \\ 0 & \cdots & 1 & \cdots & 0 \\ \vdots & & \vdots & & \\ 0 & \cdots & 0 & \cdots & 0 \end{pmatrix} = I + \beta e^i (e^j)^T \ (\beta \in I\!R, i \neq j)$$

Dabei besteht die Matrix aus Nullen mit Ausnahme einer 1 in der i-ten Zeile und j-ten Spalte. Eine Multiplikation von $L_{ij}(\beta)$ mit einer Matrix bedeutet Addition der mit $\beta$ multiplizierten $j$-ten Zeile zur $i$-ten Zeile der Matrix.

Hat nun im ersten Teilschritt des $k$-ten Schrittes des Gaußschen Algorithmus erhaltene Matrix $P_{k,s(k)}A^{(k)}$ die Gestalt

$$P_{k,s(k)}A^{(k)} = \begin{pmatrix} a_{11}^{(k)} & \cdots & a_{1,k-1}^{(k)} & a_{1,k}^{(k)} & \cdots & a_{1n}^{(k)} \\ 0 & \ddots & & & & \vdots \\ \vdots & & & & & \vdots \\ 0 & & a_{k-1,k-1}^{(k)} & & & a_{k-1,n}^{(k)} \\ 0 & & 0 & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ \vdots & & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{pmatrix}$$

so können die weiteren Teilschritte wie folgt kompakt geschrieben werden:

$$(A^{(k+1)}, b^{(k+1)}) := \underbrace{\prod_{i=k+1}^{n} L_{ik}\left(-\frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}\right)}_{=:L_k} P_{k,s(k)}(A^{(k)}, b^{(k)}).$$

Diese Vorschrift vereint die analogen Transformationen an $A^{(k)}$ bzw. $b^{(k)}$ im $k$-ten Schritt des Gaußschen Algorithmus durch Betrachtung der durch die jeweilige rechte Seite $b^{(k)}$ erweiterten Matrizen $(A^{(k)}, b^{(k)})$, $k = 1, \ldots, n$, aus $\mathbb{R}^{n\times(n+1)}$.
Der Gaußsche Algorithmus kann dann wie folgt in Form von Matrixprodukten geschrieben werden:

$$(R, c) = \prod_{k=1}^{n-1} L_k P_{k,s(k)}(A, b) = L_{n-1}P_{n-1,s(n-1)} \cdots L_1 P_{1,s(1)}(A, b)$$

Die weitere Analyse des Gaußschen Algorithmus beginnen wir nun mit einer Zusammenstellung der Eigenschaften der Transformationsmatrizen:

**Eigenschaften 2.8**

*a)* $P_{k,s(k)}$ *ist symmetrisch und regulär mit* $P_{k,s(k)}^2 = I$, $\det(P_{k,s(k)}) = -1$, *falls* $k \neq s(k)$ *und* $\mathrm{cond}_1(P_{k,s(k)}) = 1$.

*b)* $L_{ij}(\beta)$ *ist regulär für jedes* $\beta \in \mathbb{R}$ *mit* $(L_{ij}(\beta))^{-1} = L_{ij}(-\beta)$ *und es gilt* $\mathrm{cond}_1(L_{ij}(\beta)) = (1 + |\beta|)^2$.

Beweis:

a) ist klar nach Definition der Vertauschungsmatrizen und wegen $\mathrm{cond}_1(P_{k,s(k)}) = \|P_{k,s(k)}\|_1^2 = 1$, da $\|\cdot\|_1$ die Spaltensummennorm ist (vgl. 1.6 a)).

b) Wegen $i \neq j$ ist $L_{ij}(\beta)$ eine Dreiecksmatrix mit Einsen in der Hauptdiagonale, also regulär. Ferner gilt:

$$\begin{aligned} L_{ij}(\beta)L_{ij}(-\beta) &= (I + \beta e^i(e^j)^T)(I - \beta e^i(e^j)^T) \\ &= I - \beta^2 e^i(e^j)^T e^i(e^j)^T = I \text{ wegen } (e^j)^T e^i = 0, \\ \text{und damit } L_{ij}(-\beta) &= (L_{ij}(\beta))^{-1}. \\ \text{Außerdem gilt: } \mathrm{cond}_1(L_{ij}(\beta)) &= \|L_{ij}(\beta)\|_1 \|L_{ij}(-\beta)\|_1 = (1 + |\beta|)^2. \quad \square \end{aligned}$$

**Satz 2.9** *Für jede reguläre Matrix $A \in \mathbb{R}^{n\times n}$ existiert eine Matrix $P \in \mathbb{R}^{n\times n}$, die ein Produkt von Vertauschungsmatrizen darstellt, sowie Matrizen $L$ bzw. $R$ aus $\mathbb{R}^{n\times n}$ mit der Eigenschaft*

$$PA = LR = \begin{pmatrix} 1 & 0 & 0\cdots & 0 \\ \ell_{21} & 1 & 0 & \vdots \\ \vdots & & \ddots & 0 \\ \ell_{n1} & \cdots & \ell_{n,n-1} & 1 \end{pmatrix} \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ 0 & r_{22} & \cdots & r_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & r_{nn} \end{pmatrix}.$$

*$R$ ist dabei die Matrix, die der Gaußsche Algorithmus in exakter Arithmetik liefert, und es gilt $r_{ii} \neq 0, i = 1, \ldots, n$. Die Matrizen $L$ und $R$ sind durch $P$ und $A$ eindeutig festgelegt. $P$ heißt auch Permutationsmatrix.*

Beweis:
Nach unseren obigen Überlegungen kann der Gaußsche Algorithmus in der Form

$$R = A^{(n)} = L_{n-1}P_{n-1,s(n-1)} \cdots L_1 P_{1,s(1)} A \text{ mit } L_k = \prod_{i=k+1}^{n} L_{ik}\left(-\frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}\right)$$

geschrieben werden. Bezeichnet man $\ell_{ik} := \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$ so hat $L_k$ die Gestalt

$$\begin{aligned} L_k &= \prod_{i=k+1}^{n} (I - \ell_{ik} e^i (e^k)^T) = I - \sum_{i=k+1}^{n} \ell_{ik} e^i (e^k)^T \\ &= \begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & & & & \vdots \\ \vdots & & 1 & & & \vdots \\ \vdots & & -\ell_{k+1,k} & 1 & & \vdots \\ \vdots & & \vdots & & \ddots & 0 \\ 0 & & -\ell_{nk} & & & 1 \end{pmatrix}. \end{aligned}$$

Weiterhin kann man zeigen, daß $P_{k+1,s(k+1)}L_k = \hat{L}_k P_{k+1,s(k+1)}$, wobei $\hat{L}_k$ sich nur dadurch von $L_k$ unterscheidet, daß die Spalte $(0, \ldots, 0, 1, -\ell_{k+1,k}, \ldots, -\ell_{nk})^T$ ersetzt wird durch $P_{k+1,s(k+1)}(0, \ldots, 0, 1, -\ell_{k+1,k}, \ldots, -\ell_{nk})^T$ (vgl. Kielbasinski/ Schwetlick, Kap. 5.1). Macht man dies sukzessive, erhält man $R = \hat{L}_{n-1} \cdots \hat{L}_2 \hat{L}_1 PA$, wobei $P = \prod_{k=1}^{n-1} P_{k,s(k)}$ und die Matrizen $\hat{L}_k$ wie oben beschrieben aus den $L_k$ hervorgehen, aber die gleiche Struktur besitzen. Also folgt:

$$PA = LR \quad \text{mit} \quad L = \hat{L}_1^{-1}\hat{L}_2^{-1} \cdots \hat{L}_{n-1}^{-1}.$$

Klar ist nach Konstruktion, daß $R$ die behauptete obere Dreiecksgestalt besitzt und daß alle Hauptdiagonalelemente von $R$ verschieden von 0 sind. Wäre das nicht so, würde $R$ nicht regulär, damit $PA$ nicht regulär und damit $A$ nicht regulär sein.

Wir untersuchen nun die Gestalt von $L$. Bezeichnen $-\hat{\ell}_{ik}, i = k+1, \ldots, n$, die Elemente der $k$-ten Spalte von $\hat{L}_k$ unterhalb der Hauptdiagonale, so gilt nach 2.9b):

$$\hat{L}_k^{-1} = L_{k+1,k}^{-1}(-\hat{\ell}_{k+1,k}) \cdots L_{n,k}^{-1}(-\hat{\ell}_{n,k}) = L_{k+1,k}(\hat{\ell}_{k+1,k}) \cdots L_{n,k}(\hat{\ell}_{n,k})$$

$$= \begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & & & & \vdots \\ \vdots & & 1 & & & \vdots \\ \vdots & & \hat{\ell}_{k+1,k} & 1 & & \vdots \\ \vdots & & \vdots & & \ddots & 0 \\ 0 & & \hat{\ell}_{nk} & & & 1 \end{pmatrix}, \; k = 1, \ldots, n-1,$$

und folglich hat

$$L = \prod_{k=1}^{n-1} \hat{L}_k^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \hat{\ell}_{21} & 1 & & \vdots \\ \vdots & & \ddots & 0 \\ \hat{\ell}_{n1} & \cdots & \hat{\ell}_{n,n-1} & 1 \end{pmatrix}$$

die behauptete untere Dreiecksgestalt.
Es seien nun $P$ und $A$ gegeben und wir zeigen die Eindeutigkeit der Darstellung $PA = LR$, wobei $L$ und $R$ die Gestalt wie in der Behauptung besitzen. Es seien $\tilde{L}$ und $\tilde{R}$ zwei weitere Matrizen mit dieser Gestalt und der Eigenschaft $PA = LR = \tilde{L}\tilde{R}$. Dann gilt $\tilde{L}^{-1}L = \tilde{R}R^{-1}$ und wir wissen aus den obigen Überlegungen, daß $\tilde{L}^{-1}$ wieder eine untere und (analog) $R^{-1}$ wieder eine obere Dreiecksmatrix ist. Dies trifft dann auch auf die Produkte $\tilde{L}^{-1}L$ bzw. $\tilde{R}R^{-1}$ zu. Wegen der Gleichheit $\tilde{L}^{-1}L = \tilde{R}R^{-1}$ müssen beide gleich einer Diagonalmatrix $D$ sein: $\tilde{L}^{-1}L = \tilde{R}R^{-1} = D$. Wegen $L = D\tilde{L}$ muß dann aber $D = I$ und folglich $L = \tilde{L}$ und $R = \tilde{R}$ gelten. □

**Bemerkung 2.10** *Man nennt die Darstellung von $PA$ in der Form $PA = LR$ wie in Satz 2.9 die LR-Faktorisierung von $A$.*
*Der Gaußsche Algorithmus hat in Matrixschreibweise nun folgende Form:*
*- $PAx = Pb$ (Vertauschung von Zeilen)*
*- $(R, c) = L^{-1}(PA, Pb)$ (Dreieckszerlegung) und anschließend*
*- $x = R^{-1}c$ (Lösung eines linearen Gleichungssystems mit Dreiecksmatrix)*
*Ist umgekehrt eine LR-Faktorisierung von $A$ gegeben, so löst man das Gleichungssystem $Ax = b$ in den folgenden beiden Schritten:*
*- Berechne $c$ als Lösung von $Lc = Pb$ („Vorwärtselimination“ ),*
*- berechne $x$ als Lösung von $Rx = c$ („Rückwärtselimination“).*

**Bemerkung 2.11** *(Anzahl von Operationen)*
*Wir berechnen die Anzahl der Gleitkommaoperationen für den Gaußschen Algorithmus bzw. für die Lösung eines linearen Gleichungssystems. Dabei ist es üblich, mit* „opms“

*eine Gleitkommarechenoperation, bestehend aus einer Multiplikation und einer Addition/Subtraktion zugrunde zu legen. Dann erhält man für den Rechenaufwand zur*

$$\begin{aligned}
&\textit{Berechnung von } R: && \sum_{k=1}^{n-1}(n-k)^2 &&= \sum_{k=1}^{n-1} k^2 && = \tfrac{1}{6}(2n-1)n(n-1)\\
& && && && = \tfrac{1}{3}n^3 - \tfrac{1}{2}n^2 + \tfrac{1}{6}n \quad \text{opms},\\
&\textit{Berechnung von } c: && \sum_{k=1}^{n-1}(n-k) && = \tfrac{1}{2}n(n-1) && = \tfrac{1}{2}n^2 - \tfrac{1}{2}n \quad \text{opms},\\
&\textit{Lösung von } Rx{=}c: && \sum_{k=1}^{n-1}(n-k) && = \tfrac{1}{2}n^2 - \tfrac{1}{2}n \quad \text{opms}.
\end{aligned}$$

*Nicht gerechnet sind hier ingesamt 2n Divisionen durch Hauptdiagonalelemente. Zur Lösung eines linearen Gleichungssystems mit dem Gaußschen Algorithmus benötigt man also:*

$$\frac{1}{3}n^3 + \frac{1}{2}n^2 + O(n) \quad \text{opms},$$

*wobei der Term $O(n)$ ein Vielfaches von n bezeichnet.*

**Bemerkung 2.12** *(Pivotisierung)*
*Nicht eindeutig bestimmt ist bisher die Wahl der Permutationsmatrix P, d. h. die Wahl der Zeilenvertauschungen zur Bestimmung des Elementes $a^{(k)}_{s(k),k} \neq 0$ in der k-ten Spalte unterhalb des Hauptdiagonale.*
*Man nennt diesen Prozeß auch Pivotisierung und $a^{(k)}_{s(k),k}$ Pivotelement.*
*Wie sollte man nun pivotisieren ? Eine Antwort darauf gibt die Kondition der Transformationsmatrix $L_k$ im k-ten Schritt. Für diese gilt*

$$\begin{aligned}
\operatorname{cond}_1(L_k) &= \|L_k^{-1}\|_1\|L_k\|_1 = \|I + \sum_{i=k+1}^{n} \tfrac{a^{(k)}_{ik}}{a^{(k)}_{kk}} e^i (e^k)^T\|_1 \|I - \sum_{i=k+1}^{n} \tfrac{a^{(k)}_{ik}}{a^{(k)}_{kk}} e^i (e^k)^T\|_1\\
&= \left(1 + \sum_{i=k+1}^{n} \left|\tfrac{a^{(k)}_{ik}}{a^{(k)}_{kk}}\right|\right)^2 \qquad \textit{(vgl. 2.8b))}.
\end{aligned}$$

*Die Kondition von $L_k$ wird also möglichst klein, wenn $|a^{(k)}_{kk}|$ möglichst groß ist! Dies führt zur sogenannten Spaltenpivotisierung:*
*Bestimme*

$$s(k) \in \{k, k+1, \ldots, n\} \text{ so, daß } \left|a^{(k)}_{s(k),k}\right| = \max_{i=k,\ldots,n} \left|a^{(k)}_{ik}\right|.$$

*Analog zur Suche nach einem Pivotelement in einer Spalte in unserer Originalform des Gaußschen Algorithmus könnte diese auch in einer Zeile oder in der gesamten Restmatrix erfolgen. Dies führt zur sog. Zeilenpivotisierung bzw. vollständigen Pivotisierung oder Gesamt-Pivotisierung. Letztere ist i.a. zu aufwendig!*

**Folgerung 2.13** *(numerische Berechnung von Determinanten)*
*Es sei $A \in \mathbb{R}^{n\times n}$ eine reguläre Matrix mit gegebener LR-Faktorisierung gemäß Satz 2.9. Dann gilt für die Determinante von A*

$$\det(A) = (-1)^\mu \prod_{i=1}^{n} r_{ii}\ ,$$

*wobei $r_{ii}, i = 1, \ldots, n$, die Hauptdiagonalelemente von R bezeichnen und $\mu$ die Anzahl der $k \in \{1, \ldots, n-1\}$ mit $k < s(k)$ bezeichnet.*

Beweis:
Nach Satz 2.10 gilt $PA = LR$ und deshalb nach den Rechenregeln für Determinanten

$$\begin{aligned} \det(P)\det(A) &= \det(L)\det(R) \\ &= \det(R) = \prod_{i=1}^{n} r_{ii} \quad \text{wegen } \det(L) = 1. \end{aligned}$$

Aus den Eigenschaften 2.9 folgt ferner $\det(P) = \prod_{k=1}^{n-1} \det(P_{k,s(k)}) = (-1)^{\mu}$.
Insgesamt ergibt sich $(-1)^{\mu}\det(A) = \prod_{i=1}^{n} r_{ii}$. □

**Bemerkung 2.14** *(Cholesky-Faktorisierung)*
*Für spezielle Matrizen nimmt der Gaußsche Algorithmus spezielle Formen an. So kann man für symmetrische und positiv definite Matrizen $A \in \mathbb{R}^{n\times n}$ (d. h. $A = A^T$ und $\langle Ax, x\rangle > 0, \forall x \in \mathbb{R}^n$) zeigen, daß der Gaußsche Algorithmus durchführbar ist (d. h. die jeweiligen Hauptdiaonalelemente sind bei exakter Arithmetik stets positiv) und, daß die Dreieckszerlegung von A) die folgende Form annimmt:*

$$A = LD\hat{R} = LDL^T$$

*($\hat{R}$ ist dabei definiert als $D^{-1}R$ mit $D := \operatorname{diag}(r_{11}, \ldots, r_{nn})$ und besitzt damit Einsen in der Hauptdiagonale. Aus $A = A^T$ und der Eindeutigkeit der LR-Faktorisierung von A nach Satz 2.9 folgt dann $\hat{R} = L^T$.)*
*Definiert man nun noch $\hat{L} := LD^{\frac{1}{2}}$, wobei $D^{\frac{1}{2}} := \operatorname{diag}(r_{11}^{\frac{1}{2}}, \ldots, r_{nn}^{\frac{1}{2}})$, so entsteht die sog. Cholesky-Faktorisierung von A:*

$$A = \hat{L}\hat{L}^T$$

*Durch Ausnutzung der Symmetrie-Eigenschaften von A ist der Rechenaufwand des Gaußschen Algorithmus gegenüber Bem. 2.11 (etwa) halbiert.*
*(Literatur: Hämmerlin/Hoffmann, Kap. 2.2; Kielbasinski/Schwetlick, Kap. 6).*

Wir kommen nun zur Rundungsfehleranalyse des Gaußschen Algorithmus zur Lösung eines linearen Gleichungssystems. Das erste Resultat betrifft dabei die $LR$-Faktorisierung, das zweite die Rückwärtselimination.

**Satz 2.15** *Für $A \in \mathbb{R}^{n\times n}$ sei der Gaußsche Algorithmus durchführbar und L bzw. R seien die Matrizen der LR-Faktorisierung in einer t-stelligen Arithmetik. Dann existiert eine „Störung“ $\delta A \in \mathbb{R}^{n\times n}$ von A mit*

$$LR = A + \delta A \quad \text{und} \quad \|\delta A\|_p \le (\tau + O(\tau^2))F_p(A)\|A\|_p,$$

*wobei $F_p(A) := 1 + 3\sum_{k=2}^{n} \|M^{(k)}\|_p/\|A\|_p$, $p \in \{1, \infty\}$, $\tau$ die relative Rechengenauigkeit der t-stelligen Arithmetik, $M^{(k)} \in \mathbb{R}^{(n-k+1)\times(n-k+1)}$ und $A^{(k)} = \begin{pmatrix} \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \\ \hline 0 & \vert & M^{(k)} \\ & k-1 & \end{pmatrix}$.*

Beweis: Die bei der Pivotisierung vorgenommenen Zeilen- bzw. Spaltenvertauschungen entsprechen einer Umnumerierung der Gleichungen bzw. Unbekannten. Wir setzen deshalb o.B.d.A. voraus, daß diese Vertauschungen bereits vor Beginn des Gaußschen Algorithmus vorgenommen wurden und daß der Gaußsche Algorithmus mit $s(k) = k$ durchführbar ist. Mit dieser Vereinbarung nimmt der k-te Schritt für $k = 1, \ldots, n-1$ in exakter Arithmetik die folgende Form an:

$$\begin{aligned} A^{(k+1)} &= L_k A^{(k)} = \left(\begin{array}{c|c} I_{k-1} & 0 \\ \hline 0 & L^{(n-k+1)} \end{array}\right) \left(\begin{array}{c|c} 0 \quad \ddots & R^{(k)} \\ \hline 0 & M^{(k)} \end{array}\right) \\ &= \left(\begin{array}{c|c} 0 \quad \ddots & R^{(k)} \\ \hline 0 & L^{(n-k+1)} M^{(k)} \end{array}\right) \\ &= \left(\begin{array}{c|cc} 0 \quad \ddots & R^{(k)} & \\ \hline 0 & a_{kk}^{(k)} \quad a_{k,k+1}^{(k)} \quad \ldots & a_{kn}^{(k)} \\ \hline 0 & 0 & M^{(k+1)} \end{array}\right) \end{aligned}$$

Hierbei ist $L_k$ die entsprechende Transformationsmatrix (vgl. den Beweis von Satz 2.9), $R^{(k)}$ der "bereits fertige" Teil von $R$ (mit $k-1$ Zeilen), $I_{k-1} \in \mathbb{R}^{(k-1)\times(k-1)}$ die Einheitsmatrix und $M^{(k)}$ die "Restmatrix" von $A^{(k)}$.
Wir untersuchen zunächst die numerische Gutartigkeit der Transformation

$$M^{(k)} \mapsto L^{(n-k+1)} M^{(k)}$$

in $t$-stelliger Arithmetik. Die diese Transformation beschreibenden Operationen lassen sich dabei nach Postulat 1.22 in der Form

$$\begin{aligned} a_{ij}^{(k+1)} &:= [a_{ij}^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} a_{kj}^{(k)}(1+\varepsilon_{ij})](1+\theta_{ij}) \quad \text{bzw.} \\ (*) \quad a_{ij}^{(k)} &= a_{ij}^{(k+1)} \frac{1}{1+\theta_{ij}} + \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} a_{kj}^{(k)}(1+\varepsilon_{ij}) \end{aligned}$$

mit $|\varepsilon_{ij}| \le \tau$, $|\theta_{ij}| \le \tau$, $i, j = k+1, \ldots, n$, schreiben. Daraus folgt

$$\begin{aligned} (*^2) \quad a_{ij}^{(k+1)} &= a_{ij}^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} a_{kj}^{(k)} + \delta a_{ij}^{(k)}, \quad \text{wobei} \\ (*^3) \quad \delta a_{ij}^{(k)} &:= a_{ij}^{(k)} \theta_{ij} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} a_{kj}^{(k)} \eta_{ij} \quad \text{und} \quad \eta_{ij} := \varepsilon_{ij} + \theta_{ij} + \varepsilon_{ij}\theta_{ij} \end{aligned}$$

nebst $|\eta_{ij}| \le 2\tau + \tau^2$ für $i, j = k+1, \ldots, n$. Durch Einsetzen von $(*)$ in $(*^3)$ und Ausnutzung der Gleichung $(*^2)$ entsteht die Gleichung

$$\begin{aligned} \delta a_{ij}^{(k)} &= a_{ij}^{(k+1)} \frac{\theta_{ij}}{1+\theta_{ij}} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} a_{kj}^{(k)} \varepsilon_{ij} \\ &= a_{ij}^{(k+1)} \frac{\eta_{ij}}{(1+\theta_{ij})(1+\varepsilon_{ij})} - a_{ij}^{(k)} \frac{\varepsilon_{ij}}{1+\varepsilon_{ij}} \end{aligned}$$

für $i,j = k+1,\ldots,n$. Die Beträge von $\frac{1}{1+\theta_{ij}}$ und $\frac{1}{1+\varepsilon_{ij}}$ lassen sich nach Postulat 1.22 jeweils mit $1+\tau$ nach oben abschätzen. Daraus folgt für $i,j = k+1,\ldots,n$:

$$|\delta a_{ij}^{(k)}| \leq (\tau + O(\tau^2))(|a_{ij}^{(k)}| + 2|a_{ij}^{(k+1)}|)$$

Wir definieren eine Störungsmatrix $\delta M^{(k)} \in I\!R^{(n-k+1)\times(n-k+1)}$ so, daß sie in der ersten Zeile und Spalte Nullen hat und an der Stelle $ij$ das Element $\delta a_{ij}^{(k)}$ steht $(i,j = k+1,\ldots,n)$. Dann gilt

$$\left(\begin{array}{c|ccc} a_{kk}^{(k)} & a_{k,k+1}^{(k)} & \ldots & a_{kn}^{(k)} \\ \hline 0 & & M^{(k+1)} & \end{array}\right) = L^{(n-k+1)}(M^{(k)} + \delta M^{(k)}),$$

wobei die Abschätzung $\|\delta M^{(k)}\|_p \leq (\tau + O(\tau^2))(\|M^{(k)}\|_p + 2\|M^{(k+1)}\|_p)$ gültig ist. Mit

$$\delta A^{(k)} := \left(\begin{array}{c|c} 0 & 0 \\ \hline 0 & \delta M^{(k)} \end{array}\right)$$

gilt dann $A^{(k+1)} = L_k(A^{(k)} + \delta A^{(k)})$ und folglich

$$\begin{aligned} R = A^{(n)} &= L_{n-1}(A^{(n-1)} + \delta A^{(n-1)}) \\ &= L_{n-1}[L_{n-2}(A^{(n-2)} + \delta A^{(n-2)}) + \delta A^{(n-1)}] \\ &= L_{n-1}L_{n-2}[A^{(n-2)} + \delta A^{(n-2)} + \delta A^{(n-1)}], \end{aligned}$$

da wegen der speziellen Blockstruktur von $\delta A^{(k)}$ die Identität $L_{n-2}\delta A^{(n-1)} = \delta A^{(n-1)}$ gilt. Setzt man dies sukzessive fort, so folgt

$$R = L_{n-1}L_{n-2}\cdots L_2 L_1[A + \delta A^{(1)} + \cdots + \delta A^{(n-1)}].$$

Mit $L := L_1^{-1}\ldots L_{n-1}^{-1}$ und $\delta A := \sum_{k=1}^{n-1} \delta A^{(k)}$ ergibt sich deshalb die Darstellung

$$\begin{aligned} LR &= A + \delta A, \quad \text{wobei} \\ \|\delta A\|_p &\leq \sum_{k=1}^{n-1} \|\delta A^{(k)}\|_p = \sum_{k=1}^{n-1} \|\delta M^{(k)}\|_p \\ &\leq (\tau + O(\tau^2)) \sum_{k=1}^{n-1} (\|M^{(k)}\|_p + 2\|M^{(k+1)}\|_p) \\ &\leq (\tau + O(\tau^2))(\|A\|_p + 3\sum_{k=2}^{n} \|M^{(k)}\|_p) \end{aligned}$$

Damit ist die Aussage vollständig bewiesen. □

**Bemerkung 2.16** *Satz 2.15 besagt, daß in jeder Matrizenklasse $\mathcal{A} \subset I\!R^{n\times n}$, in der der Gaußsche Algorithmus durchführbar ist und*

$$\sup\{F_p(A) : A \in \mathcal{A}\} < \infty$$

*für eine der Normen $\|\cdot\|_p$ gilt, die LR-Faktorisierung numerisch gutartig ist. Klassen von solchen Matrizen werden in Folg. 2.17 und Kielbasinski/Schwetlick, Satz 5.3.2 angegeben.*

*Das folgende Beispiel zeigt aber, daß Satz 2.15 nicht die numerische Gutartigkeit des Gaußschen Algorithmus für alle regulären Matrizen impliziert. Es sei $n = 2$ und wir betrachten das lineare Gleichungssystem:*

$$\begin{array}{rcrcl} a_{11}x_1 & + & a_{12}x_2 & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & = & b_2 \end{array}$$

*wobei $a_{11}a_{22} - a_{12}a_{21} \neq 0$ (d.h. $A$ ist regulär) und $a_{11} \neq 0$ .*
*Wir erhalten dann ohne Zeilenvertauschungen:*

$$M^{(2)} = \left(a_{22} - \frac{a_{12}}{a_{11}}a_{21}\right).$$

*Also kann $\|M^{(2)}\|_p = \left|a_{22} - \frac{a_{12}}{a_{11}}a_{21}\right|$ beliebig groß werden, falls $a_{11}$ beliebig klein und $a_{12}a_{21} \neq 0$ fixiert ist.*
*Dieser Effekt tritt nicht auf, wenn eine Spaltenpivotisierung durchgeführt wird !*

**Folgerung 2.17** *Ist der Gaußsche Algorithmus mit Spaltenpivotisierung zur LR-Faktorisierung von $A \in \mathbb{R}^{n\times n}$ durchführbar, so ist er numerisch gutartig.*
*Ist $A$ regulär und gilt $(\tau + O(\tau^2))(1 + 3 \cdot 2^n)cond_\infty(A) < 1$, so ist der Gaußsche Algorithmus mit Spaltenpivotisierung durchführbar.*

Beweis:
Bei Verwendung von Spaltenpivotisierung (vgl. Bemerkung 2.12) gilt

$$\max_{i=k+1,\ldots,n} |\frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}| \leq 1 \quad \text{und folglich} \quad |a_{ij}^{(k+1)}| \leq |a_{ij}^{(k)}| + |a_{kj}^{(k)}|\,,\ \forall i,j = k+1,\ldots,n\,.$$

Deshalb folgt für die Zeilensummennorm $\|\cdot\|_\infty$ auf $\mathbb{R}^{n\times n}$:

$$\|M^{(k+1)}\|_\infty \leq 2\|M^{(k)}\|_\infty \quad \text{und folglich} \quad \|M^{(k)}\|_\infty \leq 2^{k-1}\|A\|_\infty\,.$$

(vgl. auch Kielbasinski/Schwetlick, Aussage 5.2.1).
Aus Satz 2.15 folgt dann, daß wegen $F_\infty(A) \leq 1 + 3\sum_{k=2}^{n} 2^{k-1} \leq 1 + 3 \cdot 2^n$ der Gaußsche Algorithmus mit Spaltenpivotisierung numerisch gutartig ist, falls er durchführbar ist. Für die Durchführbarkeit ist hinreichend, daß $LR$ regulär ist. Dies gilt im Fall $A \in \mathbb{R}^{n\times n}$ regulär, falls nach dem Störungslemma $\|A - LR\|_\infty\|A^{-1}\|_\infty < 1$ oder falls $(\tau + O(\tau^2))(1 + 3 \cdot 2^n)\text{cond}_\infty(A) < 1$. □

Dies bedeutet, daß der Gaußsche Algorithmus mit Spaltenpivotisierung auf der Menge aller regulären Matrizen $A$ mit nicht zu großer Kondition $\text{cond}_\infty(A)$ von $A$ durchführbar und numerisch gutartig ist. Die Abschätzung $F_\infty(A) \leq 1 + 3 \cdot 2^n$ in Folgerung 2.17 ist in vielen Anwendungsfällen zu pessimistisch. Praktisch geht man von einem linearen Wachstum von $F_\infty(A)$ mit $n$ aus.

Für die Rückwärtselimination ist die Situation einfacher.

**Satz 2.18** *Das lineare Gleichungssystem $Rx = c$ mit regulärer oberer Dreiecksmatrix $R$ werde durch den folgenden Algorithmus (Rückwärtselimination) gelöst:*

$$x_n := \frac{c_n}{r_{nn}}, \quad x_i := \left( c_i - \sum_{j=i+1}^{n} r_{ij} x_j \right) / r_{ii}, \quad i = n-1, \ldots, 1.$$

*Erfolgt dies in einer $t$-stelligen Gleitkommaarithmetik mit relativer Rechengenauigkeit $\tau$ wobei $\tau n < 1$, so genügt die berechnete Lösung $x$ der Gleichung $(R + \delta R)x = c$ mit einer oberen Dreiecksmatrix $\delta R$ mit der Eigenschaft $\|\delta R\|_\infty \leq \frac{\tau n}{1-\tau n} \|R\|_\infty$.*
*Insbesondere ist der Algorithmus numerisch gutartig.*

Beweis:
Bei Rechnung in einer $t$-stelligen Gleitkommaarithmetik gilt:

$$s_i = fl_t \left( \sum_{j=i+1}^{n} r_{ij} x_j \right) = \sum_{j=i+1}^{n} r_{ij} x_j (1 + \varepsilon_{ij}), \qquad \text{wobei}$$

$$|\varepsilon_{ij}| \leq \tfrac{(n-j+2)\tau}{1-n\tau}, \quad j = i+1, \ldots, n \quad \text{(vgl. Bsp. 1.27)}$$

$$x_n = \tfrac{c_n}{r_{nn}(1+\delta_n)} \qquad \text{mit } |\delta_n| \leq \tau \qquad \text{(vgl. Post. 1.22)}$$

$$x_i = \tfrac{c_i - s_i}{r_{ii}(1+\delta_i)} \qquad \text{mit } |\delta_i| \leq 2\tau \qquad \text{(vgl. Post. 1.22)}$$

Wir setzen nun

$$(\delta R)_{ij} := \begin{cases} r_{ij}\varepsilon_{ij} & , \quad i < j \\ r_{ii}\delta_i & , \quad i = j \\ 0 & , \quad \text{sonst} \end{cases}$$

und erhalten $(R + \delta R)x = c$. $\delta R$ ist eine obere Dreiecksmatrix und es genügt folglich, $\|\delta R\|_\infty$ abzuschätzen. Es gilt:

$$\begin{aligned} \|\delta R\|_\infty &= \max_{i=1,\ldots,n} \sum_{j=i}^{n} |(\delta R)_{ij}| = \max_{i=1,\ldots,n} \left\{ \sum_{j=i+1}^{n} |r_{ij}\varepsilon_{ij}| + |r_{ii}\delta_i| \right\} \\ &\leq \tfrac{n\tau}{1-n\tau} \max_{i=1,\ldots,n} \sum_{j=i}^{n} |r_{ij}| = \tfrac{n\tau}{1-n\tau} \|R\|_\infty \end{aligned}$$ □

**Bemerkung 2.19** *(Skalierung)*
*Die Lösungen des linearen Gleichungssystems $Ax = b$ verändern sich nicht, wenn $Ax = b$ zeilenweise mit geeigneten positiven Faktoren multipliziert wird. Diesem Vorgehen entspricht die Multiplikation von $A$ und $b$ mit einer Diagonalmatrix $D = \operatorname{diag}(d_1, \ldots, d_n)$ mit $d_i > 0$, $i = 1, \ldots, n$. Läßt sich durch geeignete Wahl von $D$ die Kondition von $A$ verkleinern ?*
*Es sei $A \in \mathbb{R}^{n \times n}$ regulär, $a^i$ bezeichne die $i$-te Zeile von $A$ und wir betrachten die Diagonalmatrix $D = \operatorname{diag}(d_1, \ldots, d_n)$, wobei*

$$(*) \quad d_i := \frac{\max\limits_{k=1,\ldots,n} \|a^k\|_1}{\|a^i\|_1} = \frac{\|A\|_\infty}{\|a^i\|_1} \quad (i = 1, \ldots, n).$$

*Dann gilt:*

$$\|DA\|_\infty = \|A\|_\infty \quad \textit{und} \quad \frac{\min\limits_{k=1,\ldots,n} \|a^k\|_1}{\max\limits_{k=1,\ldots,n} \|a^k\|_1} \mathrm{cond}_\infty(A) \leq \mathrm{cond}_\infty(DA) \leq \mathrm{cond}_\infty(A).$$

*Beweis: Es gilt zunächst* $\|DA\|_\infty = \max\limits_{i=1,\ldots,n} \sum\limits_{j=1}^{n} |d_i a_{ij}| = \max\limits_{i=1,\ldots,n} d_i \|a^i\|_1 = \|A\|_\infty$.
*Ferner gilt* $\|(DA)^{-1}\|_\infty \leq \|A^{-1}\|_\infty \|D^{-1}\|_\infty = \|A^{-1}\|_\infty \frac{1}{\min\limits_{i=1,\ldots,n} d_i} = \|A^{-1}\|_\infty$ *und*
$\|A^{-1}\|_\infty = \|(DA)^{-1}D\|_\infty \leq \|(DA)^{-1}\|_\infty \max\limits_{i=1,\ldots,n} d_i$. □

*Dies bedeutet: Wählt man D durch (∗), so verkleinert sich die Kondition* $\mathrm{cond}_\infty$ *von A bei Multiplikation mit D. Unter allen durch Zeilenskalierung aus A hervorgehenden Matrizen hat jede "zeilenäquilibrierte"( d.h., deren* $\|\cdot\|_1$ *der Zeilen alle gleich sind), die kleinste* $\mathrm{cond}_\infty$.

**Bemerkung 2.20** *(Nachiteration)*
*Es sei* $x \in \mathbb{R}^n$ *die durch LR-Faktorisierung und Rückwärtselimination berechnete Computer-Lösung des linearen Gleichungssystems* $Ax = b$ *und* $x_* \in \mathbb{R}^n$ *sei dessen exakte Lösung, d. h.,* $x_* = A^{-1}b$.
*Ferner sei* $h_* := x_* - x$ *und* $r_* := r_*(x) := b - Ax$ *das sog. Residuum von x. Dann gilt:*

$$h_* = x_* - x = A^{-1}b - x = A^{-1}(b - Ax) = A^{-1}r_* \textit{ oder } Ah_* = r_*.$$

*Also gilt:* $x_* = x + h_*$ *wobei* $Ah_* = b - Ax$.
*Man könnte also bei exakter Rechnung aus einer fehlerbehafteten Lösung durch Lösung eines weiteren Gleichungssystems (mit anderer rechter Seite) die exakte Lösung berechnen. Führt man diese Lösung wieder auf einem Computer unter Verwendung der LR-Faktorisierung durch Vorwärts- und Rückwärtselimination durch (vgl. Bemerkung 2.10), so ist auch diese Näherung fehlerbehaftet, aber i. a. besser. Dies führt zur Idee der iterativen Fortsetzung dieses Prozesses, d. h., zur sog. Nachiteration:*

- $x^0$ *Näherungslösung von* $Ax = b$ *(aus LR-Faktorisierung erhalten);*
- *für* $k = 0, \ldots, k_{\max}$ *bestimme* $h^k$ *aus dem linearen Gleichungssystem* $Ah = b - Ax^k$ *durch LR-Faktorisierung und setze*
  $x^{k+1} := x^k + h^k = x^k + (LR)^{-1}(b - Ax^k) = (LR)^{-1}(LR - A)x^k + (LR)^{-1}b$;
- *die letztere Iterierte ist eine "gute" Lösung von* $Ax = b$, *da die Folge* $(x^k)$ *wegen* $\|(LR)^{-1}(LR - A)\| < 1$ *(da* $\|LR - A\|$ *in der Regel klein ist) nach dem Banachschen Fixpunktsatz gegen* $x_*$ *konvergiert.*

*Achtung: Bei der Berechnung des Residuums können Auslöschungseffekte auftreten. Deshalb ist eine Berechnung mit höherer Genauigkeit, aber Abspeicherung mit einfacher Genauigkeit eine geeignete Vorgehensweise.*

**Bemerkung 2.21** *(Konditionsschätzung)*
*Das Ziel bestehe darin,* $cond_\infty(A)$ *für eine reguläre Matrix* $A \in \mathbb{R}^{n\times n}$ *näherungsweise zu berechnen. Die Berechnung von* $\|A\|_\infty$ *ist kein Problem, allerdings ist es ein Aufwandsproblem,* $A^{-1}$ *zu berechnen (*$O(n^3)$ *Operationen vgl. Bem. 2.11).*

*Es sei jetzt eine LR-Faktorisierung von $A$ gegeben und das Ziel sei, $\|A^{-1}\|_\infty$ näherungsweise zu berechnen. Es gilt:*

$$\|A^{-1}\|_\infty = \|(A^{-1})^T\|_1 = \|((LR)^T)^{-1}\|_1 \geq \frac{\|z\|_1}{\|x\|_1}$$

*für jedes $x \neq \theta$ und $(LR)^T z = x$.*
*Das Ziel ist nun, $x$ so zu wählen, daß $\frac{\|z\|_1}{\|x\|_1}$ möglichst groß wird.*

*Es sei $y$ so gewählt, daß $R^T y = x \rightsquigarrow L^T z = y$ und*

$$\frac{\|z\|_1}{\|x\|_1} = \frac{\|z\|_1}{\|y\|_1}\frac{\|y\|_1}{\|x\|_1} = \frac{\|(L^T)^{-1}y\|_1}{\|y\|_1}\frac{\|(R^T)^{-1}x\|_1}{\|x\|_1} \leq \|A^{-1}\|_\infty.$$

*Wir setzen nun voraus, daß die LR-Faktorisierung mit Spaltenpivotisierung erhalten wurde. Dann gilt:*

$$\ell_{ii} = 1,\ i = 1, \ldots, n,\ \text{und } |\ell_{ij}| \leq 1,\ 1 \leq j < i \leq n.$$

$$\rightsquigarrow \|L^T\|_1 \leq n \text{ und } \frac{1}{n} \leq \frac{1}{\|L^T\|_1} \leq \frac{\|(L^T)^{-1}y\|_1}{\|y\|_1} = \|(L^T)^{-1}\|_1 = \|L^{-1}\|_\infty.$$

*sowie $\|L^{-1}\|_\infty \leq 2^{n-1}$ (Übung).*

*Praktisch ist nun $\|L^{-1}\|_\infty$ meist wesentlich kleiner als $2^{n-1}$, deshalb variiert der Term $\frac{\|(L^T)^{-1}y\|_1}{\|y\|_1}$ oft nur wenig.*
*Daher versucht man, einen Vektor $x$ so zu konstruieren, daß der Term $\frac{\|(R^T)^{-1}x\|_1}{\|x\|_1}$ möglichst groß wird.*
*Ansatz: $x_1 := 1,\ x_i := \pm 1,\ i = 2, \ldots, n$. Für $y = (R^T)^{-1}x$ gilt dann $y_1 = \frac{x_1}{r_{11}}$,*
$y_i = -\sum_{j=1}^{i-1} \frac{r_{ji}}{r_{ii}} y_{j+} + \frac{x_i}{r_{ii}},\ i = 2, \ldots, n.$

*Da $x_1$ bekannt ist, ist auch $y_1$ bekannt. $y_2$ werde nun so bestimmt, daß $\|y\|_1 = \sum_{i=1}^{n} |y_i|$ möglichst groß wird. Näherungsweise bestimmt man $y_2$ aus $x_2 = \pm 1$ so, daß*

$$|y_1| + |y_2| + \sum_{i=3}^{n} \left| -\frac{r_{1i}}{r_{ii}} y_1 - \frac{r_{2i}}{r_{ii}} y_2 \right|$$

*möglichst groß wird. Diesen Prozeß setzt man dann mit $y_3$ analog fort (Details in Kielbasinski/ Schwetlick, Kap 5.4).*

## 2.3 Householder-Orthogonalisierung

Ziel: Transformation des linearen Gleichungssystems $Ax = b$ auf Dreiecksgestalt mit Hilfe von orthogonalen Matrizen, die die Kondition nicht „verschlechtern".

**Definition 2.22**
*Eine Matrix $Q \in \mathbb{R}^{n \times n}$ heißt orthogonal, falls $Q$ regulär mit $Q^{-1} = Q^T$.*
*Für jedes $u \in \mathbb{R}^n$ mit $\|u\|_2 = 1$ heißt die Matrix $H := I - 2uu^T$*
*Householder-Spiegelung(smatrix).*

**Lemma 2.23**

*a) Für $A \in \mathbb{R}^{n\times n}$ und jede orthogonale Matrix $Q \in \mathbb{R}^{n\times n}$ gilt:*

$$\|Qx\|_2 = \|x\|_2, \quad \forall x \in \mathbb{R}^n, \quad \|QA\|_2 = \|A\|_2.$$

*b) Jede Householder-Spiegelung ist symmetrisch und orthogonal.*

Beweis:

a) Für jedes $x \in \mathbb{R}^n$ gilt:
$\|Qx\|_2^2 = \langle Qx, Qx\rangle = \left\langle Q^TQx, x\right\rangle = \langle x, x\rangle = \|x\|_2^2$.
Es sei $\bar{y} \in \mathbb{R}^m$, $\|\bar{y}\|_2 = 1$ und $\|A\bar{y}\|_2 = \|A\|_2$
$\rightsquigarrow \|QA\bar{y}\|_2 = \|A\bar{y}\|_2 = \|A\|_2 \rightsquigarrow \|QA\|_2 \geq \|A\|_2$.
Außerdem gilt für jedes $y \in \mathbb{R}^n$ mit $\|y\|_2 = 1$.
$\rightsquigarrow \|QAy\|_2 = \|Ay\|_2 \leq \|A\| \rightsquigarrow \|QA\|_2 \leq \|A\|_2$.

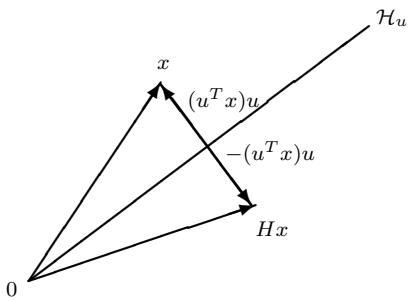
b) Eine Householder-Spiegelung ist nach Definition symmetrisch, d. h. es gilt $H = H^T$. Ferner gilt:

$$H^2 = HH = (I - 2uu^T)(I - 2uu^T) = I - 4uu^T + 4u(u^Tu)u^T = I$$

wegen $u^Tu = \langle u, u\rangle = \|u\|_2^2 = 1$. □

**Bemerkung 2.24** *Der Begriff „orthogonal“ rührt daher, daß die Zeilen und die Spalten einer orthogonalen Matrix als Vektoren in $\mathbb{R}^n$ orthogonal zueinander sind. Ist $Q$ orthogonal, so auch $Q^{-1} = Q^T$, und folglich gilt $cond_2(Q) = 1$. Der Begriff „Spiegelung“ hat seinen Ursprung darin, daß die Matrix $H = I - 2uu^T$ eine Spiegelung des $\mathbb{R}^n$ an der Hyperebene $\mathcal{H}_u := \{x \in \mathbb{R}^n : \langle x, u\rangle = 0\}$ bewirkt.*

*Schreibt man nämlich ein beliebiges $x \in \mathbb{R}^n$ in der Form $x = (u^Tx)u + (x - (u^Tx)u)$, so gilt $Hx = -(u^Tx)u + (x - (u^Tx)u)$. Obwohl von ähnlicher Art wie die Matrizen $L_{ij}(\beta)$ in Kap. 2.2, so sind Householder-Spiegelungen orthogonal und die $L_{ij}(\beta)$ nicht!*




**Lemma 2.25** *Es sei $e^1 := (1, 0, \ldots, 0)^T$, $x \in \mathbb{R}^n$ mit $x \neq \alpha e^1$, $\forall \alpha \in \mathbb{R}$.*
*Dann gilt $Hx = (I - 2uu^T)x = \sigma e^1$ und $\|u\|_2 = 1$ gdw. $u := \pm\frac{x-\sigma e^1}{\|x-\sigma e^1\|_2}$ und $\sigma := \pm\|x\|_2$.*

Beweis:
Aus der Gleichung $Hx = x - 2u^Txu = \sigma e^1$ und aus $\|u\|_2 = 1$ folgt, daß $u$ die Gestalt

$$u = \frac{x - \sigma e^1}{\|x - \sigma e^1\|_2}$$

haben muß. Nach Lemma 2.24 folgt $\|Hx\|_2 = \|x\|_2 = |\sigma|$, d.h. $\sigma = \pm\|x\|_2$.
Haben umgekehrt $u$ und $\sigma$ die angegebene Form, so gilt

$$\|x - \sigma e^1\|_2^2 = \|x\|^2 - 2\sigma\langle x, e^1\rangle + \sigma^2 = 2\|x\|^2 - 2\sigma\langle x, e^1\rangle.$$

Daraus folgt

$$Hx = x - 2u^T x\ u = x - 2\frac{(x - \sigma e^1)^T x}{\|x - \sigma e^1\|_2^2}(x - \sigma e^1) = \sigma e^1 \qquad \square$$

Das Lemma legt das folgende Orthogonalisierungsverfahren nach Householder zur Dreiecksfaktorisierung einer Matrix $A \in \mathbb{R}^{n\times n}$ nahe:

**Algorithmus 2.26** *(QR-Faktorisierung durch Householder-Orthogonalisierung)*
1. Schritt:
*$a_1$ sei die erste Spalte von $A$; hat $a_1$ die Form $\alpha e^1$, so ist der erste Schritt beendet; ansonsten bestimme $u_1 \in \mathbb{R}^n$ so, daß $H_1 a_1 = (I - 2u_1 u_1^T)a_1 = \|a_1\|_2 e^1$. Transformiere alle Spalten von $A$ mit $H_1$ und bezeichne die neue Matrix mit $A^{(1)}$.*
k-ter Schritt:
*$A^{(k-1)}$ habe die Gestalt*

$$A^{(k-1)} = \left(\begin{array}{ccc|ccc} * & \cdots & * & * & \cdots & * \\ & \ddots & \vdots & \vdots & & \vdots \\ & & * & * & \cdots & * \\ \hline & & & -- & --- & -- \\ & & & a_k^{(k-1)} & \cdots & a_n^{(k-1)} \\ & & & & & \end{array}\right) \begin{array}{l} \left.\vphantom{\begin{array}{c}a\\a\\a\end{array}}\right\} k-1 \\ \left.\vphantom{\begin{array}{c}a\\a\end{array}}\right\} n-k+1 \end{array},$$

*bestimme $u_k \in \mathbb{R}^{n-k+1}$ so, daß*

$$H_k a_k^{(k-1)} = (I - 2u_k u_k^T)a_k^{(k-1)} = \|a_k^{(k)}\|_2\ e^1 \in \mathbb{R}^{n-k+1}.$$

*Transformiere alle "Restspalten" $a_i^{(k-1)}$, $i = k, \ldots, n$, mit $H_k$ und bezeichne die neue Matrix mit $A^{(k)}$.*
n-ter Schritt:

$$\text{Setze } R = A^{(n-1)} = \left(\begin{array}{c} \\ \diagdown \\ \end{array}\right)$$

**Satz 2.27** *Zu jeder Matrix $A \in \mathbb{R}^{n\times n}$ existiert eine orthogonale Matrix $Q \in \mathbb{R}^{n\times n}$ und eine rechte obere Dreiecksmatrix $R \in \mathbb{R}^{n\times n}$, so daß*

$$A = QR \qquad (QR - Faktorisierung).$$

Beweis: Es seien $H_1, \ldots, H_{n-1}$ die in Algorithmus 2.26 definierten Householder-Spiegelungen und wir definieren die folgenden Matrizen in $\mathbb{R}^{n\times n}$:

$$Q_1 := H_1, \quad Q_k := \begin{pmatrix} I & 0 \\ 0 & H_k \end{pmatrix}, \quad k = 2, \ldots, n-1.$$

Dann sind alle $Q_k$ symmetrisch und orthogonal und $Q := Q_{n-1} \cdots Q_1$ orthogonal. Wegen $QA = A^{(n-1)} = R$ ist damit alles gezeigt. $\square$

**Bemerkung 2.28** *Anzahl der Rechenoperationen einer QR-Faktorisierung:*

$$\frac{2}{3}n^3 + 0(n^2) \text{ opms (vgl. Kielbasinski/Schwetlick, Kap. 10.2)}$$

*Die Anzahl der Rechenoperationen ist also etwa doppelt so groß wie beim Gaußschen Algorithmus. Allerdings gilt bei der Householder-Orthogonalisierung*

$$cond_2(A) = \ cond_2(R)$$

*im Unterschied zum Gaußschen Algorithmus. Deshalb ist die Householder-Orthogonalisierung besonders dann empfehlenswert, wenn $cond_2(A)$ „groß" ist. Die Householder-Orthogonalisierung ist ein numerisch gutartiges Verfahren (vgl. Kielbasinski/Schwetlick, Kap. 10.2).*

Eine wichtige weitere Anwendung von Householder-Orthogonalisierungen ist die Lösung von Ausgleichs- oder Quadratmittel-Problemen.

**Beispiel 2.29** *(lineare Regression)*
*Gegeben seien statistische Daten $(t_i, x_i) \in \mathbb{R} \times \mathbb{R}$, $i = 1, \ldots, m$, die z.B. gemessenen Werten an Zeitpunkten $t_i$ entsprechen, und reelle Funktionen $\varphi_j$, $j = 1, \ldots, n$.*
*Gesucht ist nun eine Linearkombination $\sum_{j=1}^n c_j\varphi_j$, so daß sie die gegebenen Daten bestmöglich im Quadratmittel-Sinn annimmt, d.h. die gesuchten Koeffizienten lösen das Problem*

$$\min_{c_1,\ldots,c_n} \sum_{i=1}^m (x_i - \sum_{j=1}^n c_i\varphi_j(t_i))^2 = \min_c \|Ac - x\|^2,$$

*wobei $A = (\varphi_j(t_i)) \in \mathbb{R}^{m\times n}$.*

Wir betrachten also ein Quadratmittel-Problem der Form
Gegeben: $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$.
Gesucht: $x \in \mathbb{R}^n$ mit $\frac{1}{2}\|Ax - b\|_2^2 = \min\limits_{y\in\mathbb{R}^n} \frac{1}{2}\|Ay - b\|_2^2$.

**Satz 2.30** *Es sei $A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$. Das Quadratmittel-Problem besitzt eine Lösung $x_*$. Alle solchen Quadratmittellösungen sind auch Lösungen der Normalgleichungen*

$$A^TAx = A^Tb$$

*und umgekehrt. Der affine Unterraum $\mathcal{L} = x_* + \ker(A)$, wobei $\ker(A)$ der Nullraum von $A$ ist, ist die Lösungsmenge des Quadratmittel-Problems und hat die Dimension $n - \mathrm{rg}(A)$. $\mathcal{L}$ ist einelementig, wenn $\mathrm{rg}(A) = n$, d.h., wenn $A$ spaltenregulär ist.*
*Es existiert genau ein $x^N \in \mathcal{L}$, so daß*

$$\|x^N\|_2 = \min_{x\in\mathcal{L}} \|x\|_2 \quad \textit{und} \quad x^N \perp \ker(A) \quad (\textit{Normallösung}).$$

Beweisskizze: Wir definieren $\Phi(x) := \frac{1}{2}\|Ax - b\|_2^2$ für alle $x \in \mathbb{R}^n$.
Es gilt: $\Phi(x) = \frac{1}{2}\langle Ax - b, Ax - b\rangle = \frac{1}{2}\left[\left\langle A^TAx, x\right\rangle - 2\left\langle A^Tb, x\right\rangle + \langle b, b\rangle\right]$
$\rightsquigarrow \Phi'(x) = A^TAx - A^Tb$ (Gradient), $\Phi''(x) = A^TA \in \mathbb{R}^{n\times n}$ (Hesse-Matrix).
Da die Hesse-Matrix symmetrisch und positiv semidefinit ist, ist $x$ eine Lösung des

Quadratmittel-Problems gdw. $\Phi'(x) = 0$. Die Normalgleichungen sind aber stets lösbar und besitzen gerade die angegebene Lösungsmenge $\mathcal{L}$. Überdies ist $A^TA$ regulär, falls $\mathrm{rg}(A) = n$. Ferner existiert ein Element $x^N$ in $\mathcal{L}$, so daß sein Euklidischer Abstand zum Nullelement in $\mathbb{R}^n$ minimal ist. Dieses Element steht senkrecht auf $\ker(A)$ und ist eindeutig bestimmt. □

**Definition 2.31** *Es sei $x^N$ die eindeutig bestimmte Normallösung des Quadratmittel-Problems. Dann heißt die Matrix $A^+ \in \mathbb{R}^{n\times m}$ mit der Eigenschaft $A^+b = x^N$ ($\forall b \in \mathbb{R}^m$) verallgemeinerte Inverse oder Pseudoinverse von $A$.*

**Bemerkung 2.32** *(Eigenschaften von Pseudoinversen)*
*Ist $A \in \mathbb{R}^{m\times n}$ eine Matrix mit Rang $r = \mathrm{rg}(A)$, so gilt:*

$$A^+ = \begin{cases} (A^TA)^{-1}A^T, & n = r \le m, \\ A^T(AA^T)^{-1}, & m = r \le n, \\ A^{-1}, & m = r = n. \end{cases}$$

*Erweiterung des Konditionsbegriffs für $A \in \mathbb{R}^{m\times n}$:* $\mathrm{cond}(A) := \|A^+\|\|A\|$.

Wir zeigen jetzt, wie man die $QR$-Faktorisierung zur Berechnung von Normallösungen benutzen kann. Dabei entsteht das geeignete Verfahren zur Lösung von Quadratmittel-Problemen, da bei solchen Aufgaben die Matrizen häufig sehr schlecht konditioniert sind! Es ist i.a. der Cholesky-Faktorisierung zur Lösung von $A^TAx = A^Tb$ vorzuziehen!

**Satz 2.33** *(Berechnung von Normallösungen)*
*Es sei $A \in \mathbb{R}^{m\times n}$ spaltenregulär, d.h. $\mathrm{rg}(A) = n \le m$, und $x^N$ sei die eindeutig bestimmte Normallösung. Dann existiert eine orthogonale Matrix $Q \in \mathbb{R}^{m\times m}$ und eine reguläre rechte obere Dreiecksmatrix $R \in \mathbb{R}^{n\times n}$, so daß*

$$Rx^N = r_1 \quad mit \quad Qb = \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} \begin{matrix} \}n \\ \}m-n \end{matrix}.$$

Beweis:
Analog zu Satz 2.27 existieren orthogonale Matrizen $Q_k$, $k = 1, \ldots, n-1$, so daß mit $Q = Q_{n-1}\cdots Q_1$ die Matrix $QA$ die Gestalt

$$QA = \underbrace{\begin{pmatrix} R \\ ----- \\ 0 \end{pmatrix}}_{n} \begin{matrix} \}n \\ \\ \}m-n \end{matrix}$$

mit einer regulären oberen Dreiecksmatrix $R$ besitzt. Dann gilt:

$$\begin{aligned} \Phi(x) &= \tfrac{1}{2}\|Ax-b\|_2^2 = \tfrac{1}{2}\|Q(Ax-b)\|_2^2 \qquad \text{(Lemma 2.23 !)} \\ &= \tfrac{1}{2}\left\| \begin{pmatrix} R \\ ----- \\ 0 \end{pmatrix} x - Qb \right\|_2^2 = \tfrac{1}{2}\|Rx-r_1\|_2^2 + \tfrac{1}{2}\|r_2\|_2^2. \end{aligned}$$

Also minimiert $x^N$ die Funktion $\Phi$ gdw. $Rx^N = r_1$ gilt. □

# 3 Numerik linearer Optimierungsprobleme

Wir betrachten das folgende lineare Optimierungsproblem:

$$\min\{\langle c, x\rangle : Ax = b,\, x \geq 0\} \tag{3.1}$$

wobei $c \in \mathbb{R}^m, A \in \mathbb{R}^{n\times m}, b \in \mathbb{R}^n$ gegeben sind. Die Aufgabenstellung ist so zu verstehen, daß die lineare Funktion $f(x) := \langle c, x\rangle, \forall x \in \mathbb{R}^m$ über der sogenannten Restriktionsmenge

$$M := \{x \in \mathbb{R}^m : Ax = b,\, x \geq 0\} \tag{3.2}$$

minimiert werden soll, d. h., ein Element $x_* \in M$ bestimmt werden soll, so daß $\langle c, x_*\rangle = \min\{\langle c, x\rangle : x \in M\}$. Dabei ist $\langle\cdot,\cdot\rangle$ das Skalarprodukt in $\mathbb{R}^m$ und die Relation „$\geq$“ (oder „$\leq$“) ist für Elemente aus dem $\mathbb{R}^m$ komponentenweise zu verstehen.
In komponentenweiser Form hat das Optimierungsproblem die folgende Gestalt:

$$\min\left\{\sum_{j=1}^{m} c_j x_j : \sum_{j=1}^{m} a_{ij}x_j = b_i,\;\; i = 1,\ldots,n,\;\; x_j \geq 0,\;\; j = 1,\ldots,m\right\}$$

Wir bevorzugen aber in der Regel die kompaktere Schreibweise in Matrixform.

**Bemerkung 3.1**
*Man nennt die obige Form eines linearen Optimierungsproblems die Standardform eines solchen. Andere Formen linearer Optimierungsprobleme lassen sich stets in ein Problem in Standardform umformulieren.*
*Wir diskutieren das im folgenden an einigen Beispielen.*

*(i) Nebenbedingungen der Form $Ax \leq b$ werden durch Vergrößerung des Vektors $x$ zu $\chi := (\hat{x}, \bar{x}, \tilde{x})$ und die folgenden Bedingungen in Standardform formuliert:*

$$(A, -A, I)\chi = b,\; \chi \geq 0$$

*Dabei setzt man $x = \hat{x} - \bar{x}$ mit $\hat{x} \geq 0$ und $\bar{x} \geq 0$. $\tilde{x}$ ist eine sog. Schlupfvariable, die aus der Ungleichung eine Gleichung erzeugt.*

*(ii) Nebenbedingungen der Form $Ax \geq b$ werden durch $(-A)x \leq (-b)$ auf den Fall (i) zurückgeführt. Haben die Restriktionen die Gestalt $x \geq 0$, $Ax \leq b$, so schreibt man diese in der äquivalenten Form $\begin{pmatrix} A \\ -I \end{pmatrix} x \leq \begin{pmatrix} b \\ 0 \end{pmatrix}$ und hat wieder (i).*

*(iii) Ist das ursprüngliche Problem von der Form $\max\{\langle c, x\rangle : x \in M\}$ so löst man das äquivalente lineare Optimierungsproblem*

$$\min\{\langle -c, x\rangle : x \in M\}.$$

## 3.1 Polyeder

Wir beschäftigen uns zunächst mit Eigenschaften der Restriktionsmenge des linearen Optimierungsproblems (3.1).

**Definition 3.2** *Eine Menge $H$ der Form $H = \{x \in \mathbb{R}^m : \langle a, x\rangle \leq b\}$ mit $a \in \mathbb{R}^m$ und $b \in \mathbb{R}$ heißt Halbraum. Eine Menge $P$ heißt Polyeder, wenn sie Durchschnitt endlich vieler Halbräume ist, d.h. die Form $P = \{x \in \mathbb{R}^m : Ax \leq b\}$ mit $A \in \mathbb{R}^{n\times m}$ und $b \in \mathbb{R}^n$ besitzt. Eine Menge $M \subseteq \mathbb{R}^m$ heißt konvex, falls für alle $x, y \in M$ und $\lambda \in [0, 1]$ gilt $\lambda x + (1 - \lambda)y \in M$.*

Nach Bemerkung 3.1 (i) kann jedes Polyeder evtl. durch Einführung neuer Variabler in die Form $M$ vgl. (3.2) transformiert werden. $M$ ist aber selbst ein Polyeder, da $M$ in der Form $\{x \in \mathbb{R}^m : (-I)x \leq 0, Ax \leq b, (-A)x \leq (-b)\}$ geschrieben werden kann.

**Folgerung 3.3** *Jedes Polyeder ist eine konvexe abgeschlossene Menge.*

Beweis: Es sei $P := \{x \in \mathbb{R}^m : Ax \leq b\}$ und es seien $x, y \in P$ und $\lambda \in [0, 1]$ beliebig gewählt. Dann gilt: $A(\lambda x + (1 - \lambda)y) = \lambda Ax + (1 - \lambda)Ay \leq \lambda b + (1 - \lambda)b = b$, also $\lambda x + (1 - \lambda)y \in P$. Ist $(x_n)$ eine Folge in $P$ mit $x_n \to x \in \mathbb{R}^m$, so folgt aus $Ax_n \leq b$ durch Grenzübergang $Ax \leq b$. □

**Definition 3.4** *Ein Element $x$ einer konvexen Menge $M$ heißt Extrempunkt von $M$, wenn aus der Gültigkeit der Beziehung $x = \lambda y + (1 - \lambda)z$ für $y, z \in M$ und $0 < \lambda < 1$ bereits $x = y = z$ folgt. $E_M$ bezeichne die Menge aller Extrempunkte von $M$. Ist $M$ ein Polyeder, so nennen wir einen Extrempunkt von $M$ eine Ecke.*

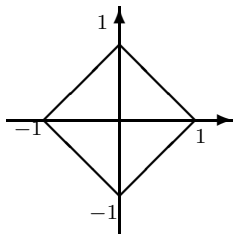
**Beispiel 3.5**

a) *Einheitskreisscheibe $\{x \in \mathbb{R}^2 : \|x\|_2 \leq 1\} = \{x \in \mathbb{R}^2 : x_1^2 + x_2^2 \leq 1\} = B_2$ im $\mathbb{R}^2$. Dann gilt: $E_{B_2} = \{x \in \mathbb{R}^2 : \|x\|_2 = 1\}$*
*Beweis:*
*Klar ist, daß Extrempunkte einer konvexen Menge keine inneren Punkte sein können. Ansonsten könnte man zwei Punkte aus einer Kugel um den Extrempunkt auswählen, auf deren Verbindungsstrecke der Extrempunkt liegt. Also gilt zunächst „$\subseteq$“.*
*Es sei $x \in \mathbb{R}^2$ mit $\|x\|_2 = 1$ und wir nehmen an, daß $y, z \in B_2, \lambda \in (0, 1)$ existieren mit $x = \lambda y+(1-\lambda)z$. Aus $1 = \|\lambda y+(1-\lambda)z\|_2^2 = \lambda^2\|y\|_2^2+(1-\lambda)^2\|z\|^2+ 2\lambda(1 - \lambda)\langle y, z\rangle$ folgt notwendig $\|y\|_2 = \|z\|_2 = 1$ und $< y, z >= 1 = \|y\|_2\|z\|_2$. Deshalb müssen $y, z$ linear abhängig sein und es folgt $x = y = z$.* □

b) *Einheitskugel bzgl. $\|\cdot\|_1$ in $\mathbb{R}^2$ : $B_1 := \{x \in \mathbb{R}^2 : \|x\|_1 \leq 1\}$. Es gilt:*

$$E_{B_1} = \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \end{pmatrix} \right\}.$$

*Bew.:* *Klar ist, daß jeder andere Punkt aus $B_1$, außer den angegebenen 4 Punkten, als Konvexkombination zweier (verschiedener) Punkte aus $B_1$ dargestellt werden kann. Für jeden der 4 Punkte ist dies aber unmöglich.* □




*$B_1$ ist ein Polyeder, es gilt nämlich*

$$B_1 = \{x \in \mathbb{R}^2 : x_2 \geq x_1 - 1, x_2 \leq x_1 + 1, x_2 \leq -x_1 + 1, x_2 \geq -x_1 - 1\}.$$

*c) Das Polyeder $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 \leq 0\}$ besitzt keine Ecke !*

**Satz 3.6** *Es seien $A = (a^1, \ldots, a^m) \in \mathbb{R}^{n \times m}$, d. h., $a^j \in \mathbb{R}^n, j = 1, \ldots, m$; $b \in \mathbb{R}^n$ und $x \in M := \{x \in \mathbb{R}^m : Ax = b, x \geq 0\}$.*
*Dann sind die Aussagen (i) und (ii) äquivalent:*
*(i) $x$ ist Ecke von $M$;*
*(ii) Die Elemente $a^j, j \in J(x) := \{j \in \{1, \ldots, m\} : x_j > 0\}$ sind linear unabhängig.*

Beweis:
(i) $\Rightarrow$ (ii): Es sei $x \in M$ eine Ecke von $M$. Wir nehmen o.B.d.A. an, daß die Komponenten von $x$ so numeriert sind, daß $J(x) = \{1, \ldots, r\}$ gilt. Für $r = 0$ ist die Aussage trivial, es sei also $r \geq 1$. Es gilt nun

$$\sum_{j=1}^{r} x_j a^j = \sum_{j=1}^{m} x_j a^j = b.$$

Annahme: $a^1, \ldots, a^r$ sind linear abhängig.
Dann existieren reelle Zahlen $\alpha_1, \ldots, \alpha_r$ mit $(\alpha_1, \ldots, \alpha_r) \neq 0$ und

$$\sum_{j=1}^{r} \alpha_j a^j = 0.$$

Wir wählen nun $\varepsilon > 0$ so klein, daß $x_j \pm \varepsilon \alpha_j > 0$, $\forall j \in J(x)$, und wir definieren Elemente $y_+$ und $y_-$ in $\mathbb{R}^m$ durch

$$y_\pm := (x_1 \pm \varepsilon \alpha_1, \ldots, x_r \pm \varepsilon \alpha_r, \ldots, 0)^T \in \mathbb{R}^m.$$

Dann gilt $y_+ \geq 0$ und $y_- \geq 0$ sowie

$$\sum_{j=1}^{m} (y_\pm)_j a^j = \sum_{j=1}^{r} (y_\pm)_j a^j = \sum_{j=1}^{r} x_j a^j \pm \varepsilon \sum_{j=1}^{r} \alpha_j a^j = b.$$

Also gilt $y_+, y_- \in M$ und $\frac{1}{2} y_+ + \frac{1}{2} y_- = (x_1, \ldots, x_r, 0, \ldots, 0) = x$, d.h., $x$ ist keine Ecke.
(ii) $\Rightarrow$ (i): Die Elemente $a^j, j \in J(x)$, seien linear unabhängig für $x \in M$. Es seien

$y, z \in M$ und $\lambda \in (0,1)$ mit $x = \lambda y + (1-\lambda)z$. Dann gilt natürlich $J(x) = J(y) \cup J(z)$. Wieder nehmen wir o.B.d.A. an, daß $J(x) = \{1, \ldots, r\}, r \geq 1$. Dann gilt

$$0 = \sum_{j=1}^{m} (y_j - z_j) a^j = \sum_{j=1}^{r} (y_j - z_j) a^j$$

und wegen der linearen Unabhängigkeit der $a^j, j \in J(x)$, auch $y_j = z_j, \ \forall j = 1, \ldots, r$, woraus $y_j = z_j, \ \forall j = 1, \ldots, m$, folgt, da die restlichen Komponenten gleich 0 sind. Deshalb gilt $x = y = z$ und $x$ ist eine Ecke von $M$. □

**Folgerung 3.7** *Die Menge $M$ sei wie in Satz 3.6 definiert und $a^j$, $j = 1, \ldots, m$ seien die Spalten von $A \in \mathbb{R}^{n \times m}$. Dann gilt:*
*$M$ besitzt höchstens endlich viele Ecken.*
*Ist* $\mathrm{rg}(A) = n$ *und $x$ eine Ecke von $M$, so existieren $j_i \in \{1, \ldots, m\}$, $i = 1, \ldots, n$, so daß $a^{j_1}, \ldots, a^{j_n}$ linear unabhängig sind und $J(x) \subseteq \{j_1, \ldots, j_n\}$ gilt.*

Beweis:
Mit den Bezeichnungen aus dem Beweis von Satz 3.6 gilt für jede Ecke $x \in M$, daß $|J(x)| \leq \mathrm{rg}(A) \leq \min\{n, m\}$. Überdies ist $x$ aus der Beziehung

$$\sum_{j \in J(x)} x_j a^j = b$$

eindeutig bestimmt. Also gibt es höchstens so viele Ecken, wie es Möglichkeiten gibt, aus einer endliche Menge von Elementen eine kleinere Anzahl von Elementen auszuwählen. Also ist $E_M$ eine endliche Menge.
Es sei nun $\mathrm{rg}(A) = n$ und $x$ eine Ecke von $M$. Dann sind die $a^j, j \in J(x)$, linear unabhängig nach (ii) und es gilt $r := |J(x)| \leq n$. Falls $r < n$, so ergänzen wir $a^j, j \in J(x) = \{j_1, \ldots, j_r\}$, durch $n-r$ weitere Spaltenvektoren $a^{j_{r+1}}, \ldots, a^{j_n}$ zu einem System linear unabhängiger Vektoren. Damit ist alles gezeigt. □

**Definition 3.8** *Es sei $A \in \mathbb{R}^{n \times m}$ mit $rg(A) = n$. Dann heißt $x \in M = \{x \in \mathbb{R}^m : Ax = b, x \geq 0\}$ Basispunkt von $M$, falls Indizes $j_i \in \{1, \ldots, m\}$, $i = 1, \ldots, n$, existieren, so daß die Matrix $A_B = (a^{j_1}, \ldots, a^{j_n})$ von Spalten von $A$ regulär ist, $x_j = 0$ für alle $j \notin \{j_1, \ldots, j_n\}$ und $Ax = \sum_{j=1}^{m} x_j a^j = \sum_{i=1}^{n} x_{j_i} a^{j_i} = b$ gilt.*
*Bezeichnung:* $J_B(x) := \{j_1, \ldots, j_n\}$, $J_N(x) := \{1, \ldots, m\} \setminus J_B(x)$.
*Wir nennen eine Komponente $x_j$ eines Basispunktes $x$ Basisvariable, falls $j \in J_B(x)$, und sonst Nichtbasisvariable.*
*Ein Basispunkt $x$ von $M$ heißt entartet, falls $|J(x)| < n$, anderenfalls nichtentartet.*

Nach 3.7 gilt für $A \in \mathbb{R}^{n \times m}$ mit $\mathrm{rg}(A) = n$: $\quad x \in E_M$ gdw. $x$ ist Basispunkt von $M$.
Als Vorbereitung für einen Darstellungssatz für Elemente von $M$ benötigen wir noch den Begriff einer Richtung.

**Definition 3.9** *Ein Element $d \in \mathbb{R}^m$, $d \neq 0$, heißt Richtung in einem Polyeder $P$, falls für jedes $x_0 \in P$ der Strahl $\{x_0 + \lambda d : \lambda \geq 0\}$ in $P$ liegt.*

$P$ besitzt eine Richtung gdw. $P$ unbeschränkt ist. Offenbar ist $d \neq 0$ eine Richtung in $M = \{x \in \mathbb{R}^m : Ax = b, x \geq 0\}$ gdw. $Ad = 0$ und $d \geq 0$.

**Satz 3.10** *(Darstellungssatz)*
*Jedes Element $x$ eines Polyeders $M = \{x \in \mathbb{R}^m : Ax = b, x \geq 0\}$ läßt sich in der Form*

$$x = \sum_{j=1}^{\ell} \lambda_j z^j + d$$

*darstellen, wobei $E_M = \{z^1, \ldots, z^\ell\}$ und $\lambda_j \geq 0$, $j = 1, \ldots, \ell$, mit $\sum_{j=1}^{\ell} \lambda_j = 1$ und $d$ entweder eine Richtung in $M$ ist oder $d = 0$ gilt.*

Beweis:
Es sei $x \in M$ beliebig mit $r := |J(x)|$. Wir führen den Beweis durch Induktion über $r$. Für $r = 0$ ist $x$ nach Satz 3.6 selbst eine Ecke und die Darstellung von $x$ ist trivial. Wir nehmen jetzt an, daß die Aussage für $0, 1, \ldots, r-1$ richtig ist. Es sei jetzt $r = |J(x)|$. Ist $x$ eine Ecke von $M$, so ist die Aussage trivial. Es sei also $x$ keine Ecke von $M$. Dann sind die Spalten $\{a^j\}_{j \in J(x)}$ linear abhängig und es existiert ein $w \neq 0$ mit $w_j = 0$ für $j \notin J(x)$ und $Aw = 0$. Wir unterscheiden die folgenden 3 Fälle:
Fall (a): $w$ hat Komponenten beiderlei Vorzeichens.
Wir betrachten die Gerade $x(\theta) = x + \theta w$, $\theta \in \mathbb{R}$, durch $x$. Für diese gilt $Ax(\theta) = b$. Es sei $\theta'$ der kleinste positive Wert von $\theta$, so dass $x(\theta)$ wenigstens eine weitere Nullkomponente als $x$ besitzt. Wegen $x \geq 0$ muß dann auch $x(\theta') \geq 0$ und damit $x(\theta') \in M$ gelten. Ähnlich wählen wir $\theta''$ als den größten negativen Wert von $\theta$, so daß $x(\theta)$ ebenfalls wenigstens eine weitere Nullkomponente als $x$ besitzt. Die Punkte $x' = x(\theta')$ und $x'' = x(\theta'')$ liegen beide in $M$ und es gilt $J(x') < r$ und $J(x'') < r$. Nach Induktionsvoraussetzung existieren für $x'$ und $x''$ Darstellungen der Form

$$x' = \sum_{j=1}^{\ell} \lambda_j' z^j + d' \quad \text{und} \quad x'' = \sum_{j=1}^{\ell} \lambda_j'' z^j + d'',$$

wobei $\lambda_j', \lambda_j'' \geq 0, j = 1, \ldots, \ell$, und $\sum_{j=1}^{\ell} \lambda_j' = 1 = \sum_{j=1}^{\ell} \lambda_j''$ sowie $d', d'' \geq 0, Ad' = Ad'' = 0$.
Da $x$ zwischen $x'$ und $x''$ liegt, existiert ein $\mu \in (0,1)$ mit

$$\begin{aligned} x &= \mu x' + (1-\mu)x'' = \mu \sum_{j=1}^{\ell} \lambda_j' z^j + d' + (1-\mu) \sum_{j=1}^{\ell} \lambda_j'' z^j + d'' \\ &= \sum_{j=1}^{\ell} (\mu \lambda_j' + (1-\mu)\lambda_j'') z^j + \mu d' + (1-\mu) d''. \end{aligned}$$

Außerdem gilt $d := \mu d' + (1-\mu)d'' \geq 0$ und $Ad = 0$ sowie $\lambda_j := \mu\lambda_j' + (1-\mu)\lambda_j'' \geq 0$, $j = 1, \ldots, \ell$, und $\sum_{j=1}^{\ell} \lambda_j = 1$. Also hat $x$ die gewünschte Form.
Fall (b): $w \leq 0$.
Wir definieren $x'$ wie im Fall (a). Dann kann $x$ in der Form

$$x = x' + \theta'(-w) \quad \text{mit} \quad \theta' > 0$$

geschrieben werden. Da $-w$ eine Richtung in $M$ ist und $x'$ wie in (a) die gewünschte Gestalt hat, gilt die Darstellung für $x$ mit $\lambda_j := \lambda_j'$, $j = 1, \ldots, \ell$, und $d := d' + \theta'(-w)$.

Fall (c): $w \geq 0$.
In diesem Fall ist $w$ eine Richtung in $M$ und $x$ kann in der Form $x = x'' + (-\theta'')w$ dargestellt werden, wobei $x''$ und $\theta'' < 0$ wie in Fall (a) gewählt werden. Schließlich setzt man $\lambda_j := \lambda_j''$, $j = 1, \ldots, \ell$, und $d := d'' + (-\theta'')w$. □

**Folgerung 3.11** *In jedem beschränkten Polyeder $M = \{x \in \mathbb{R}^m : Ax = b, x \geq 0\}$ ist jedes $x \in M$ eine Konvexkombination der Ecken von $M$.*

Beweis: folgt sofort mit $d = 0$ aus Satz 3.10. □

Mit Hilfe des Beweises von Satz 3.10 läßt sich nun auch die Existenz einer Ecke für ein Polyeder in Standardform einfach herleiten.

**Satz 3.12** *Jedes nichtleere Polyeder $M$ der Form $M = \{x \in \mathbb{R}^m : Ax = b, x \geq 0\}$ mit $A \in \mathbb{R}^{n \times m}$ und $b \in \mathbb{R}^n$ besitzt eine Ecke.*

Beweis: Es sei $\gamma := \min\{|J(x)| : x \in M\}$ und es sei $x \in M$ mit $\gamma := |J(x)|$. Wäre nun $x$ keine Ecke von $M$, so könnte man analog zum Beweis von Satz 3.10 ein Element $x'$ in $M$ konstruieren mit der Eigenschaft $|J(x')| < \gamma$. Da dies unmöglich ist, muß $x$ eine Ecke von $M$ sein. □

## 3.2 Existenz und Charakterisierung von Lösungen

**Satz 3.13** *(Existenzsatz)*
*Das Polyeder $M := \{x \in \mathbb{R}^m : Ax = b,\ x \geq 0\}$ sei nichtleer. Dann gilt entweder $\inf\{\langle c, x\rangle : x \in M\} = -\infty$ oder das Infimum wird in einer Ecke von $M$ angenommen.*

Beweis:
1. Fall: Es existiert eine Richtung $d \in \mathbb{R}^m$ mit $\langle c, d\rangle < 0$. In diesem Fall ist $M$ unbeschränkt und es gilt mit einem fixierten $x_0 \in M$, daß

$$\inf_{x \in M} \langle c, x\rangle \leq \inf_{\lambda \geq 0} \langle c, x_0 + \lambda d\rangle = -\infty$$

2. Fall: Es sei $x \in M$ beliebig. Nach Satz 3.10 hat $x$ eine Darstellung der Form

$$x = \sum_{j=1}^{\ell} \lambda_j z^j + d,$$

wobei $E_M = \{z^1, \ldots, z^\ell\}$ und $\lambda_j \geq 0$, $j = 1, \ldots, \ell$, mit $\sum\limits_{j=1}^{\ell} \lambda_j = 1$ und für die Richtung $d$ muss gelten $\langle c, d\rangle \geq 0$. Dann gilt

$$\langle c, x\rangle = \sum_{j=1}^{\ell} \lambda_j \langle c, z^j\rangle + \langle c, d\rangle \geq \sum_{j=1}^{\ell} \lambda_j \langle c, z^j\rangle \geq \min\{\langle c, z^j\rangle : j = 1, \ldots, \ell\}$$

und folglich

$$\min\{\langle c, x\rangle : x \in M\} = \min\{\langle c, z^j\rangle : j = 1, \ldots, \ell\}.$$

Damit ist die Aussage bewiesen. □

**Bemerkung 3.14** *Die Zielfunktion $f$ kann ihr Minimum auch in mehreren (genauer: unendlich vielen Punkten) annehmen. Wichtig ist hier nur, daß mindestens eine Ecke dazu gehört.*
*Beispiel: Wir betrachten das Polyeder*

$$M := \{(x_1, x_2) \in I\!R^2 \,:\, x_2 \leq -1,\, x_2 \leq x_1,\, 2x_2 \leq -x_1\}.$$

*Dann besitzt das lineare Optimierungsproblem* $\min\{\langle c, x\rangle : x \in M\}$ *für* $c := (0, 1)$ *keine Lösung (da* $d = (0, -1)$ *eine Richtung ist). Für* $c = (0, -1)$ *ist das Problem lösbar und es gilt*

$$\inf\{\langle c, x\rangle : x \in M\} = \inf\{-x_2 : x \in M\} = 1 = \{\langle c, (x_1, -1)\rangle : x_1 \in [-1, 2]\},$$

*d.h. die Strecke zwischen den Ecken* $(-1, -1)$ *und* $(2, -1)$ *von* $M$ *ist gerade die Lösungsmenge des Optimierungsproblems.*
*Im Prinzip zeigt Satz 3.13, wie man zur Lösung eines linearen Optimierungsproblems vorgehen wird: Elemente aus* $E_M$ *sind Kandidaten für optimale Lösungen. Satz 3.6 liefert dabei ein Kriterium zur Bestimmung der Ecken. Diese allgemeine Vorgehensweise muß aber noch "effektiviert" werden, da man Ausschau nach Richtungen* $d$ *in* $M$ *mit* $\langle c, d\rangle < 0$ *halten muß und ein "reines" Absuchen der Ecken (bei einer i.a. sehr großen Zahl von Ecken) zu lange dauert. Später werden wir sehen, wie man dieses Absuchen der Ecken ökonomischer gestaltet.*

Wir setzen im folgenden generell voraus, daß $M = \{x \in I\!R^m : Ax = b,\ x \geq 0\}$ nichtleer ist, daß $\operatorname{rg}(A) = n$ und das lineare Optimierungsproblem

$$\min\{\langle c, x\rangle : Ax = b,\ x \geq 0\}$$

lösbar ist. Es sei $\bar{x}$ eine Ecke von $M$ (vgl. Satz 3.12). Nach Folgerung 3.7 ist $\bar{x}$ ein Basispunkt. Für jedes $x \in I\!R^m$ bezeichnen wir nun

$$x_B := (x_{j_1}, \ldots, x_{j_n})^T \in I\!R^n\,,\ \text{falls } J_B(\bar{x}) = \{j_1, \ldots, j_n\}\,,\ \text{und}$$

$$x_N := (x_{j_{n+1}}, \ldots, x_{j_m})^T \in I\!R^{m-n}\,,\ \text{falls } J_N(\bar{x}) = \{j_{n+1}, \ldots, j_m\}\,,$$

sowie mit $A_B$ und $A_N$ die entsprechenden Teilmatrizen von $A$. Dann kann man das lineare Optimierungsproblem in der folgenden Form schreiben

$$\min\{\langle c_B, x_B\rangle + \langle c_N, x_N\rangle : A_B x_B + A_N x_N = b,\ x_B \geq 0, x_N \geq 0\}$$

oder wegen $x_B = A_B^{-1}(b - A_N x_N)$ in der Form

$$\min\{\langle c_N - (A_B^{-1}A_N)^T c_B, x_N\rangle + \langle c_B, A_B^{-1}b\rangle : A_B^{-1}(b - A_N x_N) \geq 0, x_N \geq 0\}. \quad (3.3)$$

Für die Ecke $\bar{x}$ gilt $\bar{x}_N = 0$ (Folgerung 3.7) und $\bar{x}_B = A_B^{-1}b$ $(\geq 0)$.

**Lemma 3.15** *Das lineare Optimierungsproblem (3.1) sei lösbar, es gelte* $rg(A) = n$ *und ausgehend von einer Ecke* $\bar{x}$ *von* $M$ *sei das Problem auf die Form (3.3) reduziert. Gilt* $c_N - (A_B^{-1}A_N)^T c_B \geq 0$*, so sind* $\bar{x}_N$ *bzw.* $\bar{x}$ *Lösungen von (3.3) bzw. (3.1).*

Beweis:
Aus den Bedingungen $c_N - (A_B^{-1}A_N)^T c_B \geq 0$ und $x_N \geq 0$ und der Gestalt der zu minimierenden Funktion in (3.3) folgt sofort, daß $x_N = 0$ eine Lösung von (3.3) ist. Also gilt dies für $\bar{x}_N$ und nach Konstruktion ist $\bar{x}$ eine Lösung von (3.1). □

## 3.3 Das Simplex-Verfahren

Nach Satz 3.13 und Lemma 3.15 sucht man zur Lösung von (3.1) also eine Ecke $\bar{x}$, so daß mit den Bezeichnungen aus dem letzten Kapitel gilt $c_N - (A_B^{-1}A_N)^T c_B \geq 0$. Wir beschreiben jetzt eine Methodik, wie man aus einer Ecke $\bar{x}$ von $M$, die keine Lösung von (3.1) ist, eine neue Ecke bestimmen kann, für die der Wert der Zielfunktion kleiner als in $\bar{x}$ ist ("Eckenaustausch").
Ist eine Ecke $\bar{x}$ von $M$ keine Lösung des linearen Optimierungsproblems, so muß nach Lemma 3.15 ein $k \in J_N(\bar{x})$ existieren, so daß

$$\langle c_N - (A_B^{-1}A_N)^T c_B,\ e^k\rangle < 0.$$

Hierbei ist $e^k \in I\!R^{m-n}$ der entsprechende kanonische Einheitsvektor. Wir definieren nun eine Richtung $\bar{d} \in I\!R^m$ durch die Festlegung $\bar{d}_B := -A_B^{-1}A_N e^k$ und $\bar{d}_N := e^k$, sowie eine „Schrittweite" $\bar{t}$ durch

$$\bar{t} := \sup\{t \geq 0 : \bar{x} + t\bar{d} \in M\}.$$

**Lemma 3.16**
*Es gelte $M := \{x \in I\!R^m : Ax = b, x \geq 0\} \neq \emptyset$, es sei $rg(A) = n$ und $\bar{x} \in E_M$.*
*Ist $\bar{x}$ keine Lösung von (3.1), so ist entweder $\bar{d}$ eine Richtung in $M$ mit $\langle c, \bar{d}\rangle < 0$ und folglich besitzt (3.1) keine Lösung oder $\bar{x} + \bar{t}\bar{d}$ ist eine Ecke von $M$ mit $\langle c, \bar{d}\rangle < 0$, wobei*

$$\begin{aligned}
\bar{d}_B &:= -A_B^{-1}A_N e^k, \quad \bar{d}_N := e^k, \quad J_B(\bar{x}) = \{j_1, \ldots, j_n\},\\
\bar{t} &:= \min\left\{\frac{[A_B^{-1}b]_{j_i}}{[A_B^{-1}A_N e^k]_{j_i}} : [A_B^{-1}A_N e^k]_{j_i} > 0, i = 1, \ldots, n\right\} \quad \textit{und}\\
k &\in J_N(\bar{x}) \quad \textit{mit} \quad \langle c_N - (A_B^{-1}A_N)^T c_B,\ e^k\rangle < 0.
\end{aligned}$$

*Ist $\bar{x}$ nichtentartet, so gilt $\langle c, \bar{x} + \bar{t}\bar{d}\rangle < \langle c, \bar{x}\rangle$.*
*Überdies gilt $J_B(\bar{x} + \bar{t}\bar{d}) = (J_B(\bar{x}) \setminus \{j_{i_0}\}) \cup \{k\}$ und $J_N(\bar{x} + \bar{t}\bar{d}) = (J_N(\bar{x}) \setminus \{k\}) \cup \{j_{i_0}\}$, wobei $i_0 \in \{1, \ldots, n\}$ ein Index ist, an dem das Minimum zur Bestimmung von $\bar{t}$ angenommen wird.*

Beweis:
Für den oben definierten Vektor $\bar{d}$ gilt nach Konstruktion

$$\begin{aligned}
A\bar{d} &= -A_B A_B^{-1}A_N e^k + A_N e^k = 0, \text{ also } A(\bar{x} + t\bar{d}) = A\bar{x} = b,\ \forall t \geq 0,\\
\langle c, \bar{d}\rangle &= \langle c_B, -A_B^{-1}A_N e^k\rangle + \langle c_N, e^k\rangle = \langle c_N - (A_B^{-1}A_N)^T c_B, e^k\rangle < 0\,.
\end{aligned}$$

Die erste Möglichkeit ist nun, daß $\bar{d} \geq 0$ und damit $\bar{x} + t\bar{d} \geq 0$ für alle $t \geq 0$ gilt. Dann ist $\bar{d}$ eine Richtung in $M$ im Sinne von Def. 3.9 und das lineare Optimierungsproblem (3.1) besitzt keine Lösung (vgl. Beweis von Satz 3.13).
Gilt nicht $\bar{d} \geq 0$, so erhalten wir für $x(t) := \bar{x} + t\bar{d}$, daß

$$\begin{aligned}
x_N(t) &= te^k \geq 0,\ \forall t \geq 0, \quad \text{und} \quad x_B(t) = A_B^{-1}(b - tA_N e^k)\,,\ \forall t \geq 0\,.\\
\rightsquigarrow x_{j_i}(t) &= [A_B^{-1}(b - tA_N e^k)]_{j_i} \geq 0, \quad \text{falls} \quad [A_B^{-1}A_N e^k]_{j_i} \leq 0\,.
\end{aligned}$$

Also gilt $x(t) \in M$ für alle $t \geq 0$ falls $x_{j_i}(t) \geq 0,\ \forall i = 1, \ldots, n$ mit $[A_B^{-1}A_N e^k]_{j_i} > 0$.
Letzteres gilt offenbar für alle $0 \leq t \leq \min\limits_i \frac{[A_B^{-1}b]_{j_i}}{[A_B^{-1}A_N e^k]_{j_i}}$, wobei $i = 1, \ldots, n$ mit

$[A_B^{-1}A_Ne^k]_{j_i} > 0$ und damit für $t \in [0, \bar{t}]$. Für mindestens ein $i \in \{1, \ldots, n\}$ muß $[A_B^{-1}A_Ne^k]_{j_i} > 0$ gelten, da anderenfalls $\bar{d} \geq 0$ gelten würde, was aber jetzt ausgeschlossen ist.
Ist $\bar{x}$ nichtentartet, so ist jede Komponente von $A_B^{-1}b$ positiv. Folglich gilt $\bar{t} > 0$ und damit $\langle c, \bar{x} + \bar{t}\bar{d}\rangle < \langle c, \bar{x}\rangle$.
Wegen $x_{j_{i_0}}(\bar{t}) = 0$ für jeden Index $i_0 \in \{1, \ldots, n\}$, an dem das Minimum angenommen wird, kann $j_{i_0}$ aus $J_B(\bar{x})$ in $J_N(\bar{x} + \bar{t}\bar{d})$ wechseln. Da außerdem $x_k(\bar{t}) = \bar{t}$ gilt, muß der Index $k$ aus $J_N(\bar{x})$ in $J_B(\bar{x} + \bar{t}\bar{d})$ wechseln. Gezeigt werden muß jedoch, daß die Spalten $a^{j_i}$, $a^k$, $i = 1, \ldots, n$, $i \neq i_0$, linear unabhängig sind. In diesem Fall würde nach Satz 3.6 folgen, daß $\bar{x} + \bar{t}\bar{d}$ eine Ecke von $M$ ist. Wir nehmen also an, daß die Spalten $a^{j_i}$, $a^k$, $i = 1, \ldots, n$, $i \neq i_0$, linear abhängig sind. Dies ist nur möglich, wenn $\alpha_i \in \mathbb{R}$, $i = 1, \ldots, n$, $i \neq i_0$, existieren, so daß $a^k = \sum\limits_{\substack{i=1 \\ i \neq i_0}}^{n} \alpha_i a^{j_i}$. Nun gilt aber nach Konstruktion $a^k = A_Ne^k = -A_B\bar{d}_B$, d.h. $a^k$ ist auch eine Linearkombination der $a^{j_i}$, $i = 1, \ldots, n$. Dies kann aber nur richtig sein, wenn in der letzteren Linearkombination der Faktor von $a^{j_{i_0}}$, also $\bar{d}_{j_{i_0}} = [A_B^{-1}A_Ne^k]_{j_{i_0}}$, gleich 0 ist. Dies ist aber nach Wahl von $i_0$ unmöglich und die gewünschte lineare Unabhängigkeit ist gezeigt. □

Lemma 3.16 offeriert also einen Eckenaustausch indem man einen Indexaustausch in den Mengen $J_B$ bzw. $J_N$ von Indizes für Basisvariable bzw. Nichtbasisvariable vornimmt. Dies führt zu folgendem Grundalgorithmus.

**Algorithmus 3.17** *(Simplex-Verfahren)*

*(i) Bestimme eine Ecke $x^0$ von $M := \{x \in \mathbb{R}^m : Ax = b, x \geq 0\}$, berechne die Indexmengen $J_B(x^0)$ und $J_N(x^0)$ und setze $\ell := 0$.*

*(ii) Prüfe $c_N - (A_B^{-1}A_N)^T c_B \geq 0$. Wenn ja, stop ($x^\ell$ ist Lösung)!*

*(iii) Wähle $k \in J_N(x^\ell)$ mit $\langle c_N - (A_B^{-1}A_N)^T c_B, e^k\rangle < 0$ und bestimme $\bar{d}$.
Falls $\bar{d} \geq 0$, stop (das Problem hat keine Lösung)!
Anderenfalls berechne $\bar{t}$ und die Indexmengen $J_B(\bar{x} + \bar{t}\bar{d})$ und $J_N(\bar{x} + \bar{t}\bar{d})$, wobei $\bar{x} = x^\ell$, durch Indexaustausch wie in Lemma 3.16.*

*(iv) Setze $x^{\ell+1} := x^\ell + \bar{t}\bar{d}$, $\ell := \ell + 1$ und gehe zu (ii).*

Das Verfahren besteht also aus zwei Phasen: In Phase I wird zunächst eine Ecke des zulässigen Bereichs $M$ bestimmt und in Phase II wird durch Ecken- bzw. Index-Austausch eine Ecke von $M$ bestimmt, die auch Lösung von (3.1) ist.

**Satz 3.18** *Das Polyeder $M := \{x \in \mathbb{R}^m : Ax = b, x \geq 0\}$ sei nichtleer, es gelte $\mathrm{rg}(A) = n$ und alle Ecken von $M$ seien nichtentartet.*
*Dann bricht das Simplex-Verfahren 3.17 entweder mit der Information der Unlösbakeit oder nach einer Anzahl von $\ell_*$ Schritten mit der Information, daß $x^{\ell_*}$ eine Lösung des linearen Optimierungsproblems ist, ab.*
*Es gilt $\langle c, x^{\ell+1}\rangle < \langle c, x^\ell\rangle, \forall \ell = 0, \ldots, \ell_* - 1$.*

Beweis:
Nach Lemma 3.16 sind alle $x^\ell, \ell = 0, 1, \ldots$ im Simplex-Verfahren Ecken von $M$. Beim Übergang von $x^\ell$ zu $x^{\ell+1}$ gilt $\langle c, x^{\ell+1}\rangle < \langle c, x^\ell\rangle$ nach Lemma 3.16, da alle Ecken von $M$ nach Voraussetzung nichtentartet sind. Da die Werte der Zielfunktion streng monoton fallend sind, kann keine Ecke mehrfach auftreten und die Aussage folgt aus der Tatsache, daß $M$ höchstens endlich viele Ecken besitzt (Folgerung 3.7). □

**Bemerkung 3.19** *(Phase I)*
*In den meisten Fällen läßt sich eine Startecke $x^0$ nicht auf einfache Art und Weise bestimmen. Jedoch läßt sich das Simplex-Verfahren, in Anwendung auf ein Hilfsproblem, selbst dazu verwenden, eine Startecke bzw. einen Basispunkt zu berechnen. Dazu betrachten wir das lineare Optimierungsproblem*

$$\min\left\{\sum_{i=1}^{n} y_i : Ax + y = b, x \geq 0, y \geq 0\right\}$$

*mit $m+n$ Variablen, wobei wir o.B.d.A. annehmen, daß $b \geq 0$ gilt. Offenbar ist $(x, y) = (0, b)$ eine Ecke des zulässigen Bereiches. Mit dieser Ecke starten wir das Simplex-Verfahren zur Lösung dieses Problems. Existiert eine zulässiger Punkt in (3.1) (d.h. gilt $M \neq \emptyset$), so endet das Verfahren in einer Ecke, die Lösung des Problems ist und folglich die Form $(\bar{x}, 0)$ hat. Damit ist $\bar{x}$ eine Ecke des Polyeders $M$ und kann als Startecke $x^0$ verwendet werden. Endet das Simplex-Verfahren mit einem Zielfunktionswert, der größer als 0 ist, so muß $M$ leer sein.*

**Beispiel 3.20** *(Klee/Minty 1972)*
*Wir betrachten das lineare Optimierungsproblem*

$$\min\{-\langle e^m, x\rangle = -x_m : 0 \leq x_1 \leq 1, \varepsilon x_{i-1} \leq x_i \leq 1 - \varepsilon x_{i-1}, i = 2, \ldots, m\},$$

*wobei $\varepsilon \in (0, 0.5)$, mit $m$ Variablen und $2m$ Ungleichungen. Offenbar sind $(0, \ldots, 0) \in \mathbb{R}^m$ und $(0, \ldots, 0, 1) \in \mathbb{R}^m$ Ecken des zulässigen Bereiches. Wird das Simplex-Verfahren mit $x_0 := (0, \ldots, 0)$ gestartet und wählt man $k$ stets so aus, daß $\langle c_N - (A_B^{-1}A_N)^T c_B, e^k\rangle < 0$ am kleinsten ist, so durchläuft das Verfahren alle Ecken und endet nach $2^m$ Schritten in $(0, \ldots, 0, 1)$. Dies ist das worst-case Verhalten des Algorithmus.*

**Bemerkung 3.21** *(mittlere Laufzeit des Simplex-Verfahrens)*
*Nach Einführung des Simplex-Verfahrens durch Dantzig 1947/48 und durchaus guten praktischen Erfahrungen, schienen Beispiele der Art wie 3.20 und damit eine mögliche exponentielle Anzahl von Schritten, das "Aus" für dieses Verfahren einzuläuten. Dies wurde noch dadurch verstärkt, daß Ende der 70er eine neue Klasse von Lösungsverfahren für lineare Optimierungsprobleme gefunden wurde, die sog. innere-Punkt Verfahren. Jedoch wurde Anfang der 80er das mittlere Laufzeitverhalten des Simplex-Verfahrens untersucht und gezeigt, daß die mittlere Anzahl der Schritte nur wie $\min\{m^2, n^2\}$ wächst. Gegenwärtig hält man es sogar noch für möglich, daß eine Index-Austausch-Regel existiert, die ein (nur) polynomiales Wachstum der Anzahl der Schritte mit $\min\{m, n\}$ für jedes lineare Optimierungsproblem garantiert (Todd 2002).*

**Bemerkung 3.22** *(Entartungen und Implementierung)*
*Treten entartete Ecken auf, so kann $\bar{t} = 0$ gelten. Dann bleibt das Simplex-Verfahren in derselben Ecke, berechnet aber eine neue Indexmenge $J_B$, d.h. eine neue Basis. Im nächsten Schritt könnte deshalb wieder $\bar{t} > 0$ gelten. Um zu verhindern, daß unendliche Zyklen beim Basisaustausch auftreten, werden sogenannte Pivot-Regeln verwendet, die festlegen, welche der Indizes ausgetauscht werden, wenn es mehrere Kandidaten dafür gibt. Z.B. kann man immer die mit dem kleinsten Index verwenden.*
*Schließlich sei angemerkt, daß auf keinen Fall im Laufe des Simplex-Verfahrens explizit inverse Matrizen der Form $A_B^{-1}$ berechnet werden müssen. In allen Fällen werden LR-Zerlegungen dieser Matrizen berechnet und die entsprechenden Gleichungssysteme damit gelöst. Man kann zeigen, daß bei Spalten-Austausch die neuen LR-Zerlegungen ökonomisch aus den alten berechnet werden können.*

Zusätzliche Literatur zu Kapitel 3:

D. Goldfarb, M.J. Todd: Linear Programming, in: Handbooks in Operations Research and Management Science, Vol. 1, Optimization (G.L. Nemhauser, A.H.G. Rinnooy Kan, M.J. Todd eds.), North-Holland, Amsterdam 1989, 73–170.
M.J. Todd: The many facets of linear programming, Mathematical Programming 91 (2002), 417–436.