Numerische lineare Algebra

Vorlesungsskripte
Sommer-Semester 2012

Werner Römisch

Humboldt-Universität Berlin
Institut für Mathematik

Inhalt: Seite

0 Einleitung

Gegestand dieser Vorlesung ist die numerische Lösung linearer Gleichungssysteme. Diese stellen eine Grundaufgabe der Numerischen Mathematik dar. Die Lösung komplizierterer Gleichungen (nichtlineare, Differential- oder Integralgleichungen) wird meist auf die Lösung linearer Gleichungssysteme zurückgeführt.

Neben dem wohlbekannten Gaußschen Algorithmus, der in der Vorlesung genauer untersucht wird, werden auch andere Methoden, u.a. iterative Verfahren, behandelt, die insbesondere für (sehr) große lineare Gleichungssysteme geeigneter sind als der Gaußsche Algorithmus.

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.

* C. Kanzow: Numerik linearer Gleichungssysteme, Springer, Berlin, 2005.

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 und Kondition

1.1 Normen von Vektoren und Matrizen

Vektoren und Matrizen sind die elementaren "Bauteile" linearer Gleichungssysteme. Um 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.* $x = 0 = (0, \ldots, 0) \in \mathbb{R}^n$;
(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* $\|x\|_2 = \langle x, x \rangle^{\frac{1}{2}}$.

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

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*

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

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) $\|A\| = \max_{\|x\|_X=1} \|Ax\|_Y = \sup_{\|x\|_X\le 1} \|Ax\|_Y = \sup_{x\neq 0} \frac{\|Ax\|_Y}{\|x\|_X}$ *für jede Matrix $A \in \mathbb{R}^{m\times n}$.*

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

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

$$\|BA\| \le \|B\|\ \|A\|.$$

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

$$\|\alpha A\| = \sup_{\|x\|_X=1} \|\alpha Ax\|_Y = |\alpha| \sup_{\|x\|_X=1} \|Ax\|_Y = |\alpha|\ \|A\|$$

$$\|A+B\| = \sup_{\|x\|_X=1} \|(A+B)x\|_Y \le \sup_{\|x\|_X=1} (\|Ax\|_Y + \|Bx\|_Y)$$

$$\le \sup_{\|x\|_X=1} \|Ax\|_Y + \sup_{\|x\|_X=1} \|Bx\|_Y = \|A\| + \|B\|.$$

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

(i) Das Supremum der stetigen Funktion $x \to \|Ax\|_Y$ wird auf der abgeschlossenen und beschränkten (also kompakten) Menge $\{x \in \mathbb{R}^n : \|x\|_X = 1\}$ angenommen. Die restlichen beiden Identitäten sind eine Übungsaufgabe.

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

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

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:

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

also $\|A\|_1 \leq \max\limits_{j=1,\ldots,n} \sum\limits_{i=1}^{m} |a_{ij}|$.
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* $\|A\bar{x}\|_1 = \sum\limits_{i=1}^{m} |a_{ij_0}| \leq \|A\|_1$. □

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 λ_i, $i = 1, \ldots, n$, eine Orthonormalbasis von (reellen) Eigenvektoren x_i, $i = 1, \ldots, n$. Damit gilt:

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

Also gilt: $\|A\|_2 \leq \sqrt{\lambda_{\max}(A^T A)}$.
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 gilt* $\|A\|_B := \max_{\|x\|_B=1} \|Ax\|_B = \|BAB^{-1}\|$, *wobei die Norm* $\|\cdot\|_B$ *auf* $\mathbb{R}^n$ *in Beispiel 1.3(iii) eingeführt wurde (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* $\|A\|_2$*:*

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

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

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*

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

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 Aussage.

Satz 1.9 *Alle Normen auf* $\mathbb{R}^n$ *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 $e^i = (0,\ldots,0,\underset{\uparrow_i}{1},0,\ldots,0)$ der i-te kanonische Einheitsvektor. Dann gilt für $x = (x_1,\ldots,x_n) \in \mathbb{R}^n$, daß $x = \sum_{i=1}^{n} x_i e^i$ und

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

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

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

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 $\bar{x} \in S_n$ mit $m := \|\bar{x}\| = \inf_{x \in S_n} \|x\|$.
Wegen $\bar{x} \neq 0$ gilt $m > 0$ und $m \leq \left\| \frac{x}{\|x\|_2} \right\|, \forall x \in \mathbb{R}^n, x \neq 0$.
Folglich gilt $m\|x\|_2 \leq \|x\| \leq M\|x\|_2, \forall x \in \mathbb{R}^n$. □

Satz 1.9 gibt uns die Möglichkeit, bei Fehlerabschätzungen irgendeine Norm zu verwenden. Die Aussagen bleiben bis auf Konstanten für jede Norm qualitativ richtig. Allerdings sind oftmals ganz bestimmte Normen an ein gegebenes Problem besonders angepaßt. In unendlichdimensionalen Räumen gilt die Aussage von Satz 1.9 nicht mehr (Ursache: Die Einheitskugel ist nicht mehr kompakt).

1.2 Fehler in linearen Gleichungssystemen und Kondition von Matrizen

Wir betrachten das lineare Gleichungssystem (mit quadratischer Matrix)

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

Ist A regulär, so ist es für jede rechte Seite $b \in \mathbb{R}^m$ eindeutig lösbar.
Da in den Anwendungen die Eingangsdaten A und b häufig fehlerbehaftet sind, betrachten wir das „gestörte“ lineare Gleichungssystem

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

wobei ΔA und Δb Fehler in den Daten A bzw. b darstellen und Δx den entstehenden Fehler der Lösung bezeichnet. Uns interessieren Abschätzungen für den relativen Fehler $\frac{\|\Delta x\|}{\|x\|}$ (mit einer Norm $\|\cdot\|$ im $\mathbb{R}^m$).
Als ersten Schritt einer theoretischen Analyse kümmern wir uns zunächst um die Invertierbarkeit gestörter invertierbarer Matrizen.

Lemma 1.10 *Es sei $B \in \mathbb{R}^{m \times m}$ und für eine zugeordnete Matrixnorm gelte $\|B\| < 1$. Dann ist die Matrix $I + B$ invertierbar und es gilt:*

$$\frac{1}{1 + \|B\|} \leq \|(I + B)^{-1}\| \leq \frac{1}{1 - \|B\|}.$$

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

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

$$\|(I + B)x\| = \|x + Bx\| \geq \|x\| - \|Bx\| \geq (1 - \|B\|)\|x\| \geq 0$$

Also folgt aus $(I+B)x = 0$ sofort $x = 0$. Also gilt $rg(I+B) = m$ und es existiert $(I+B)^{-1} =: C \in \mathbb{R}^{m\times m}$. Für die Matrix C gilt:

$$I = (I+B)C = C(I+B)$$

$$\rightsquigarrow \quad \|I\| = 1 \leq \|(I+B)\|\|C\| \leq (1+\|B\|)\|C\|$$

$$\|I\| = 1 = \|C+BC\| \geq \|C\| - \|BC\| \geq (1-\|B\|)\|C\|.$$

Damit ist die behauptete Ungleichungskette bewiesen. □

Satz 1.11 *(„Störungslemma")*
Es seien $A, \tilde{A} \in \mathbb{R}^{m\times m}$ mit A invertierbar und für eine zugeordnete Matrixnorm gelte $\|A^{-1}\| \leq \beta, \|A-\tilde{A}\| \leq \alpha$ und $\alpha\beta < 1$. Dann ist auch $\tilde{A}$ invertierbar und es gilt

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

Beweis:
Wir betrachten $B := A^{-1}(\tilde{A}-A)$. Dann gilt: $\|B\| \leq \|A^{-1}\|\|\tilde{A}-A\| \leq \alpha\beta < 1$. Deshalb ist nach Lemma 1.1 die Matrix $I + A^{-1}(\tilde{A}-A) = A^{-1}\tilde{A}$ invertierbar, also auch $\tilde{A}$. Außerdem folgt aus Lemma 1.1:

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

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

□

Folgerung 1.12
Die Menge $\{A \in \mathbb{R}^{m\times m} : A$ ist invertierbar$\}$ offen im Raum $(\mathbb{R}^{m\times m}, \|\cdot\|)$ (mit jedem invertierbaren A gehört auch die Kugel $\{\tilde{A} \in \mathbb{R}^{m\times m} : \|\tilde{A}-A\| < \frac{1}{\|A^{-1}\|}\}$ zur Menge).

Beweis:
Die Aussage folgt sofort aus Satz 1.11. □

Im folgenden Störungsresultat für Lösungen linearer Gleichungssysteme taucht nun erstmals der Term $\|A\|\,\|A^{-1}\|$ auf der rechten Seite der Abschätzung auf.

Satz 1.13 *Es seien A und ΔA Matrizen aus $\mathbb{R}^{m\times m}$, A sei invertierbar, b und Δb seien aus $\mathbb{R}^m$ und $x = A^{-1}b$. Ferner gelte für die zu einer Norm $\|\cdot\|$ auf dem $\mathbb{R}^m$ zugeordnete Matrixnorm $\|A^{-1}\|\,\|\Delta A\| < 1$.*
Dann existiert eine Lösung $x + \Delta x$ des linearen Gleichungssystems

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

und es gelten die Abschätzungen

$$\|\Delta x\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\ \|\Delta A\|}\|\Delta b\|, \ \textit{falls } b = 0,$$

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\|A^{-1}\|\,\|A\|}{1 - \|A^{-1}\|\,\|A\|\frac{\|\Delta A\|}{\|A\|}}\left(\frac{\|\Delta b\|}{\|b\|} + \frac{\|\Delta A\|}{\|A\|}\right), \ \textit{falls } b \neq 0.$$

Beweis:
Die Voraussetzung $\|A^{-1}\|\ \|\Delta A\| < 1$ impliziert nach Satz 1.11, daß die Matrix $A + \Delta A$ invertierbar ist und daß

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

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

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

Im Fall $b = 0$ gilt auch $x = 0$ und alles ist gezeigt. Es sei $b \neq 0$ ($\rightsquigarrow\ x \neq 0$).

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

wobei im letzten Schritt die Ungleichung $\|A\| \geq \frac{\|b\|}{\|x\|}$ verwendet wurde. □

Definition 1.14 *Die Zahl* $\textit{cond}\,(A) := \|A\|\ \|A^{-1}\|$ *heißt* Konditionszahl *der invertierbaren Matrix* $A \in \mathbb{R}^{m\times m}$ *(bzgl. der Matrixnorm* $\|\cdot\|$*).*

Bemerkung 1.15 *Die Abschätzung für den relativen Fehler* $\frac{\|\Delta x\|}{\|x\|}$ *in Satz 1.13 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“* ΔA *ist dieser Faktor etwa gleichgroß mit* $\mathrm{cond}(A)$*.*
Große Konditionszahlen führen also i.a. dazu, daß aus kleinen Eingabefehlern große Fehler bei den Lösungen resultieren! Offensichtlich hängt die Konditionszahl $\mathrm{cond}(A)$ *von der konkreten Wahl der Norm ab.*
Es gilt aber stets $\mathrm{cond}(A) \geq 1$ *wegen* $\|A^{-1}\|\,\|A\| \geq \|I\| = 1$*. Bei Verwendung von* $\|\cdot\|_p$ *als Matrixnorm verwenden wir die Bezeichnung* $\mathrm{cond}_p(A)$*.*

Beispiel 1.16 *Es sei H_m die Hilbert-Matrix der Ordnung m mit den Elementen*

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

H_m ist symmetrisch und positiv definit, also auch invertierbar, mit der ganzzahligen Inversen $H_m^{-1} = (h_{ij})$, $h_{ij} = \frac{(-1)^{i+j}}{i+j-1} r_i r_j$ mit $r_i := \frac{(m+i-1)!}{((i-1)!)^2(m-i)!}$, $i,j = 1,\ldots,m$.
Für $m = 4$ 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 $H_m x = b := H_m(1,1,\ldots,1)^T$ für verschiedene $m \in \mathbb{N}$ mit dem Gaußschen Algorithmus gelöst, so ergeben sich für $m = 8$ bzw. $m = 10$ relative Fehler der Lösungen von etwa 0.4 bzw. 3.4. Dies Effekt nimmt mit wachsendem m weiter zu (vgl. Hämmerlin/Hoffmann, Kap. 2.6).
Man erahnt, daß die Ursache für die aufgetretenen Fehler sich auch in den Konditionszahlen von H_m manifestiert. Es gilt nämlich:

m	*3*	*4*	*5*	*10*
$cond_2(H_m)$	*520*	*16.000*	*480.000*	$1.6 \cdot 10^{13}$

wobei $\mathrm{cond}_2(H_m) = \|H_m^{-1}\|_2 \|H_m\|_2 = \frac{\lambda_{\max}(H_m)}{\lambda_{\min}(H_m)}$.

1.3 Rundungsfehlerfortpflanzung

Definition 1.17 *Für gegebene $\beta, t, N_-, N_+ \in \mathbb{N}, \beta \geq 2$, nennen wir*

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

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

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.
Der Übergang von einer reellen Zahl $x \in \mathbb{R}$ zu einer Computerzahl wird als Runden bezeichnet. Dabei gehen wir davon aus, daß die folgende Rundungsvorschrift realisiert wird.

Definition 1.18 *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. 1.18 der üblicherweise als „Runden“ bezeichneten Regel.

Satz 1.19 *Es sei $\beta \in \mathbb{N}$ gerade, $t \in \mathbb{N}$ und $x \in \mathbb{R}\backslash\{0\}$ besitze die Darstellung wie in Definition 1.18. 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:

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

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 $\mathrm{fl}_t(x\square y)$ bezeichnet.

Postulat 1.20 *Die Computer-Arithmetik ist so organisiert, daß für zwei Computerzahlen x und y mit Mantissenlänge t stets gilt:*

$$\mathrm{fl}_t(x\square y) = rd_t(x\square y)$$

Wir nehmen im folgenden stets an, dass das Postulat erfüllt ist. Aus Satz 1.19(iv) ergibt sich dann

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

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

$$\mathrm{fl}_t(x \pm y) = \tilde{x} \pm \tilde{y},\ \mathrm{fl}_t(x * y) = \tilde{x} * y = x * \tilde{y}$$

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

Bemerkung 1.21 *(Probleme der Computer-Arithmetik)*

Überlauf:	*Falls der Betrag des Ergebnisses einer Rechenoperation grösser als β^{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.*
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.*

Definition 1.22 *Unter einem numerischen Algorithmus P_A zur Lösung eines Problems P wollen wir im folgenden stets eine Nacheinanderausführung von N elementaren Rechenoperationen O_i, $i = 1, \ldots, N$, verstehen, d.h.*

$$P_A = O_N \circ O_{N-1} \circ \cdots \circ O_2 \circ O_1 \,.$$

Durch Postulat 1.20 kennen wir die Rundungsfehlerauswirkungen bei elementaren Rechenoperationen. Wir fragen uns nun nach der Fortpflanzung der Rundungsfehler im einem Algorithmus, wobei wir mit einer (sehr) großen Zahl N von elementaren Operationen zu rechnen haben.
Um geeignete Eigenschaften von Algorithmen zu formulieren, beschreiben wir ein Problem P als eine Abbildung von einer Datenmenge D in eine Lösungsmenge und bezeichnen mit $P(d)$ das Ergebnis des Problems für die konkreten Daten $d \in D$. Der Algorithmus P_A überführt diese Daten d ebenfalls in ein Ergebnis $P_A(d)$.

Definition 1.23 *Es sei τ die relative Rechengenauigkeit der t-stelligen Gleitkomma-Arithmetik und es seien Normen im Raum der Daten und der Lösungen gewählt.*

(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" δ_d mit $d + \delta_d \in D$ und $\|\delta_d\| \leq F_g \|d\| \tau$ existiert, so dass

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

Die numerische Gutartigkeit ist die bestmögliche Eigenschaft eines numerischen Algorithmus; die numerische Stabilität ist eine Mindestanforderung. Natürlich wird man sich in beiden Fällen wünschen, dass die Konstanten F_s und F_g nicht "zu groß" sind. Da alle elementaren Operationen nach Postulat 1.20 im wesentlichen den gleichen Fehler wie die Rundung (vgl. Satz 1.19) erzeugen, nämlich proportional zur relativen Rechengenauigkeit τ, erhofft man sich, dass die Fortpflanzung der Fehler bis auf eine Konstante den Rundungsfehlern der Daten entspricht.

Beispiel 1.24 *Die Datenmenge D ist eine Menge von Paaren $d = (A, b)$ mit einer regulären Matrix $A \in \mathbb{R}^{m\times m}$ und $b \in \mathbb{R}^m$. Es gilt $P(d) = A^{-1}b$ für $d = (A, b)$. Der Algorithmus P_A wird für alle $d = (A, b) \in D$ eine Näherung von $P(d)$ berechnen. Numerische Stabiltät von P_A bedeutet, dass der relative Fehler*

$$\frac{\|A^{-1}b - P_A(A, b)\|}{\|(A, b)\|}$$

höchstens proportional zur relativen Rechengenauigkeit τ ist, wobei die Proportionalitätskonstante für alle Daten gleichmäßig fix (obwohl evtl. sehr groß) ist. Dabei setzt man $\|d\| := \|A\| + \|b\|$, wobei die Matrixnorm der gewählten Norm auf $\mathbb{R}^m$ zugeordnet ist. Die Menge D ist nach dem Störungslemma (Satz 1.11) sinnvollerweise als

$$D_\alpha = \{(\tilde{A}, b) \in \mathbb{R}^{m\times m} \times \mathbb{R}^m : \|A^{-1}\|\|\tilde{A} - A\| \leq \alpha\}$$

mit $\alpha < 1$ zu wählen.

Satz 1.25 *Es seien $A \in \mathbb{R}^{m\times m}$ eine reguläre Matrix und $b \in \mathbb{R}^m$. Die Datenmenge D_α und die Norm sei wie in Beispiel 1.24 definiert.*
Ist ein Algorithmus P_A zur Lösung des linearen Gleichungssystems $Ax = b$ numerisch gutartig aud D, so ist er auch numerisch stabil mit $F_s := F_g \frac{\|A^{-1}\|}{1-\alpha} \max\{1, \|A^{-1}\|\|b\|\}$.

Beweis: Aus dem Beweis von Satz 1.13 wissen wir, dass mit $(A + \delta_A, b + \delta_b) \in D_\alpha$ gilt

$$\|A^{-1}b - (A + \delta_A)^{-1}(b + \delta_b)\| \leq \frac{\|A^{-1}\|}{1 - \alpha}(\|\delta_b\| + \|\delta_A\| \, \|A^{-1}b\|).$$

Wir wählen nun $F_g > 0$ und $\delta_d = (\delta_A, \delta_b)$ so, dass $\|\delta_d\| \leq F_g\|d\|\tau$ und $P_A(d) = P(d + \delta_d) = (A + \delta_A)^{-1}(b + \delta_b)$.
Dann gilt

$$\begin{aligned}
\|P(d) - P_A(d)\| &= \|A^{-1}b - (A + \delta_A)^{-1}(b + \delta_b)\| \\
&\leq \frac{\|A^{-1}\|}{1 - \alpha} \max\{1, \|A^{-1}\|\|b\|\}(\|\delta_b\| + \|\delta_A\|) \\
&\leq \frac{\|A^{-1}\|}{1 - \alpha} \max\{1, \|A^{-1}\|\|b\|\}\|\delta_d\| \\
&\leq \frac{\|A^{-1}\|}{1 - \alpha} \max\{1, \|A^{-1}\|\|b\|\}F_g\tau\|d\|\tau
\end{aligned}$$

und die Aussage ist bewiesen. □

Beispiel 1.26 *(Übung)*
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: $S := x_1$, $S := S + x_i$, $i = 2, \ldots, n$.
Realer Algorithmus: $s := \mathrm{rd}_t(x_1)$, $s := \mathrm{fl}_t(s + \mathrm{rd}_t(x_i))$, $i = 2, \ldots, n$, d.h. $P_A(x) := s$.

Dann ist der Algorithmus P_A numerisch stabil auf $D = \mathbb{R}^n$ mit $\|\cdot\|_1$ und

$$F_s := \frac{n}{1-(n-1)\tau} \approx n\,,$$

falls $(n-1)\tau < 1$ vorausgesetzt wird. Er ist sogar numerisch gutartig mit

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

2 Direkte Verfahren für lineare Gleichungssysteme

Wir werden hier fast ausschließlich lineare Gleichungssysteme

$$Ax = b \quad (A \in \mathbb{R}^{m\times m},\, b \in \mathbb{R}^m)$$

mit quadratischer, invertierbarer Koeffizientenmatrix betrachten, so dass diese Systeme für beliebige rechte Seiten stets genau eine Lösung besitzen. In den Anwendungen ist es zweckmäßig, Gleichungssysteme nach speziellen (analytischen, algebraischen) Eigenschaften, nach ihrer „Größe“ sowie ihrer Struktur zu unterscheiden. Diese Unterscheidungen betreffen eine

(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). In diesem Kapitel behandeln wir nur Verfahren, die für (i) und (ii) relevant sind.

2.1 Der Gaußsche Algorithmus

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

- Vertauschung von zeilen,
- 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 invertierbare Dreiecksmatrix R zu erzeugen. Die Grundform des Gaußschen Algorithmus hat die folgende formale Beschreibung:

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

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:

$$\begin{array}{lll} P_{k,s(k)} := E & & (k = s(k)) \\ P_{k,s(k)} := (e^1, \ldots, e^{k-1}, & e^{s(k)}, e^{k+1}, \ldots, e^{s(k)-1}, & e^k, e^{s(k)+1}, \ldots, e^n) \ (k \neq s(k)) \\ & \uparrow & \uparrow \\ & \text{Spalte k} & \text{Spalte s(k)} \end{array}$$

wobei $e^i := (0, \ldots, 0, 1, 0, \ldots, 0)^T$ der i-te kanonische Einheitsvektor in $\mathbb{R}^m$ ist. $P_{k,s(k)}$ entsteht also aus der Einheitsmatrix $E \in \mathbb{R}^{m \times m}$ 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*:

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

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 β 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

$$P_{k,s(k)}A^{(k)} = \begin{pmatrix} a_{11}^{(k)} & \cdots & a_{1,k-1}^{(k)} & a_{1,k}^{(k)} & \cdots & a_{1m}^{(k)} \\ 0 & \ddots & & & & \vdots \\ \vdots & & & & & \vdots \\ 0 & & a_{k-1,k-1}^{(k)} & & & a_{k-1,m}^{(k)} \\ 0 & & 0 & a_{kk}^{(k)} & \cdots & a_{km}^{(k)} \\ \vdots & & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & a_{mk}^{(k)} & \cdots & a_{mm}^{(k)} \end{pmatrix}$$

so können die weiteren Teilschritte wie folgt kompakt geschrieben werden:

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

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 $(A^{(k)}, b^{(k)})$, $k = 1, \ldots, m$, aus $\mathbb{R}^{m\times(m+1)}$.
Der Gaußsche Algorithmus kann dann wie folgt in Form von Matrixprodukten geschrieben werden:

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

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

Eigenschaften 2.1

a) $P_{k,s(k)}$ *ist symmetrisch und invertierbar 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 invertierbar 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.

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

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

Satz 2.2 *Für jede invertierbare Matrix $A \in \mathbb{R}^{m\times m}$ existiert eine Matrix $P \in \mathbb{R}^{m\times m}$, die ein Produkt von Vertauschungsmatrizen darstellt, sowie Matrizen L bzw. R aus $\mathbb{R}^{m\times m}$ mit der Eigenschaft*

$$PA = LR = \begin{pmatrix} 1 & 0 & 0\cdots & 0 \\ \ell_{21} & 1 & 0 & \vdots \\ \vdots & & \ddots & 0 \\ \ell_{m1} & \cdots & \ell_{m,m-1} & 1 \end{pmatrix} \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ 0 & r_{22} & \cdots & r_{2m} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & r_{mm} \end{pmatrix}.$$

R ist dabei die Matrix, die der Gaußsche Algorithmus in exakter Arithmetik liefert, und es gilt $r_{ii} \neq 0, i = 1, \ldots, m$. 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

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

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

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

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 $(0, \ldots, 0, 1, -\ell_{k+1,k}, \ldots, -\ell_{mk})^\top$ ersetzt wird durch $P_{k+1,s(k+1)}(0, \ldots, 0, 1, -\ell_{k+1,k}, \ldots, -\ell_{mk})^\top$ (Kielbasinski-Schwetlick, Kap. 5.1). Macht man dies sukzessive, erhält man $R = \hat{L}_{m-1} \cdots \hat{L}_2 \hat{L}_1 PA$, wobei $P = \prod_{k=1}^{m-1} P_{k,s(k)}$ und die Matrizen $\hat{L}_k$ wie oben beschrieben aus den L_k hervorgehen, aber die gleiche Struktur besitzen. Also folgt:

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

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 invertierbar, damit PA nicht invertierbar und damit A nicht invertierbar sein.

Wir untersuchen nun die Gestalt von L. Bezeichnen $-\hat{\ell}_{ik}, i = k+1, \ldots, m$, die Elemente der k-ten Spalte von $\hat{L}_k$ unterhalb der Hauptdiagonale, so gilt nach 2.1b):

$$\begin{aligned}\hat{L}_k^{-1} &= L_{k+1,k}^{-1}(-\hat{\ell}_{k+1,k}) \cdots L_{m,k}^{-1}(-\hat{\ell}_{m,k}) = L_{k+1,k}(\hat{\ell}_{k+1,k}) \cdots L_{m,k}(\hat{\ell}_{m,k}) \\ &= \begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & & & & \vdots \\ \vdots & & 1 & & & \vdots \\ \vdots & & \hat{\ell}_{k+1,k} & 1 & & \vdots \\ \vdots & & \vdots & & \ddots & 0 \\ 0 & & \hat{\ell}_{mk} & & & 1 \end{pmatrix}, \ k = 1, \ldots, m-1,\end{aligned}$$

und folglich hat

$$L = \prod_{k=1}^{m-1} \hat{L}_k^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \hat{\ell}_{21} & 1 & & \vdots \\ \vdots & & \ddots & 0 \\ \hat{\ell}_{m1} & \cdots & \hat{\ell}_{m,m-1} & 1 \end{pmatrix}$$

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 $D = E$ und folglich $L = \tilde{L}$ und $R = \tilde{R}$ gelten. □

Bemerkung 2.3 *Man nennt die Darstellung von PA in der Form $PA = LR$ wie in Satz 2.2 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 2.4 *(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

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

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

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

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

Bemerkung 2.5 *(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 der 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

$$\begin{aligned}
\operatorname{cond}_1(L_k) &= \|L_k^{-1}\|_1\|L_k\|_1 = \|I + \sum_{i=k+1}^{m} \tfrac{a^{(k)}_{ik}}{a^{(k)}_{kk}} e^i (e^k)^\top\|_1 \|I - \sum_{i=k+1}^{m} \tfrac{a^{(k)}_{ik}}{a^{(k)}_{kk}} e^i (e^k)^\top\|_1\\
&= \left(1 + \sum_{i=k+1}^{m} \left|\tfrac{a^{(k)}_{ik}}{a^{(k)}_{kk}}\right|\right)^2 \qquad \textit{(vgl. 2.1b))}.
\end{aligned}$$

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

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

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 Variante ist i.a. zu aufwändig!

Folgerung 2.6 *(numerische Berechnung von Determinanten)*
Es sei $A \in \mathbb{R}^{m\times m}$ eine invertierbare Matrix mit gegebener LR-Faktorisierung gemäß Satz 2.2. Dann gilt für die Determinante von A

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

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

Beweis:
Nach Satz 2.2 gilt $PA = LR$ und deshalb nach den Rechenregeln für Determinanten

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

Aus den Eigenschaften 2.1a) folgt ferner $\det(P) = \prod_{k=1}^{m-1} \det(P_{k,s(k)}) = (-1)^{\mu}$.

Insgesamt ergibt sich $(-1)^{\mu}\det(A) = \prod_{i=1}^{m} r_{ii}$. □

Bemerkung 2.7 *(Cholesky-Faktorisierung)*
Für spezielle Matrizen kann der Gaußsche Algorithmus spezielle Formen annehmen. Ist z.B. A eine symmetrische und positiv definite Matrix in $\mathbb{R}^{m\times m}$ (d.h. $A = A^{\top}$ und $\langle Ax, x\rangle > 0, \forall x \in \mathbb{R}^m$), so ist der Gaußsche Algorithmus ohne Zeilenvertauschungen durchführbar (die jeweiligen Hauptdiaonalelemente sind bei exakter Arithmetik stets positiv) und die Dreieckszerlegung von A) hat die Form:

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

Dabei ist $\hat{R}$ definiert als $D^{-1}R$ mit $D := \operatorname{diag}(r_{11}, \ldots, r_{mm})$ und besitzt deshalb Einsen in der Hauptdiagonalen. Aus $A = A^{\top}$ und der Eindeutigkeit der LR-Faktorisierung von A nach Satz 2.2 folgt dann $\hat{R} = L^{T}$. Anders interpretiert, kann der Gaußsche Algorithmus dazu verwendet werden, um die positive Definitheit von A zu testen. Das Kriterium ist $r_{ii} > 0$, $i = 1, \ldots, m$.

Definiert man nun noch $\hat{L} := LD^{\frac{1}{2}}$, wobei $D^{\frac{1}{2}} := \operatorname{diag}(r_{11}^{\frac{1}{2}}, \ldots, r_{mm}^{\frac{1}{2}})$, 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. 2.4 (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. Die ersten Resultate (Satz 2.8 und Folg. 2.10) betreffen dabei die numerische Gutartigkeit der LR-Faktorisierung, das letzte (Satz 2.11) die Rückwärtselimination.

Satz 2.8 *Für $A \in \mathbb{R}^{m\times m}$ sei der Gaußsche Algorithmus durchführbar und L bzw. R seien die Matrizen der LR-Faktorisierung in einer t-stelligen Arithmetik. Dann existiert eine „Störung“ $\delta A \in \mathbb{R}^{m\times m}$ 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 $F_p(A) := 1+3\sum_{k=2}^{m} \|M^{(k)}\|_p/\|A\|_p, p \in \{1,\infty\}$, τ *die relative Rechengenauigkeit der* t*-stelligen Arithmetik,* $M^{(k)} \in \mathbb{R}^{(m-k+1)\times(m-k+1)}$ *und* $A^{(k)} = \left(\begin{array}{c|c} \begin{matrix}\cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots\end{matrix} & \\ \hline \begin{matrix}0\\ k-1\end{matrix} & M^{(k)} \end{array}\right)$.

Beweis: Wir setzen o.B.d.A. voraus, dass der Gaußsche Algorithmus mit $s(k) = k$ durchführbar ist. Mit dieser Vereinbarung nimmt der k-te Schritt für $k = 1, \ldots, m-1$ in exakter Arithmetik die folgende Form an:

$$\begin{aligned} A^{(k+1)} &= L_k A^{(k)} = \left(\begin{array}{c|c} I_{k-1} & 0 \\ \hline 0 & L^{(m-k+1)} \end{array}\right) \left(\begin{array}{c|c} \begin{matrix}0 & \ddots\\ 0 & \end{matrix} & \begin{matrix}R^{(k)}\\ M^{(k)}\end{matrix} \end{array}\right) \\ &= \left(\begin{array}{c|c} \begin{matrix}0 & \ddots\\ 0 & \end{matrix} & \begin{matrix}R^{(k)}\\ L^{(m-k+1)}M^{(k)}\end{matrix} \end{array}\right) \\ &= \left(\begin{array}{c|ccc} 0 & \ddots & & R^{(k)} \\ \hline 0 & a_{kk}^{(k)} & a_{k,k+1}^{(k)} & \cdots \quad a_{km}^{(k)} \\ \hline 0 & 0 & \multicolumn{2}{c}{M^{(k+1)}} \end{array}\right) \end{aligned}$$

Hierbei ist L_k die entsprechende Transformationsmatrix (vgl. den Beweis von Satz 2.2), $R^{(k)}$ der "bereits fertige" Teil von R (mit $k-1$ Zeilen), $I_{k-1} \in \mathbb{R}^{(k-1)\times(k-1)}$ die Einheitsmatrix der Dimension $k-1$ 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 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 $|\varepsilon_{ij}| \le \tau$, $|\theta_{ij}| \le \tau$, $i,j = k+1, \ldots, m$, 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 $i,j = k+1, \ldots, m$. 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 $i,j = k+1,\ldots,m$. Die Beträge von $\frac{1}{1+\theta_{ij}}$ und $\frac{1}{1+\varepsilon_{ij}}$ lassen sich jeweils mit $1+\tau$ nach oben abschätzen. Daraus folgt für $i,j = k+1,\ldots,m$:

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

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

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

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

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

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

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

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

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

Damit ist die Aussage vollständig bewiesen. □

Bemerkung 2.9 *Satz 2.8 besagt, daß in jeder Matrizenklasse $\mathcal{A} \subset \mathbb{R}^{m\times m}$, in der der Gaußsche Algorithmus durchführbar ist und für eine der Normen $\|\cdot\|_p$*

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

gilt, die LR-Faktorisierung numerisch gutartig ist. Klassen solcher Matrizen werden in Folg. 2.10 und in Kielbasinski/Schwetlick, Satz 5.3.2 angegeben.
Das folgende Beispiel zeigt aber, dass Satz 2.8 nicht die numerische Gutartigkeit des Gaußschen Algorithmus für alle invertierbaren Matrizen impliziert. Es sei $m = 2$ 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 invertierbar) 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 sowie die Norm $\|A\|_p$ nicht wächst.

Der Effekt in Bemerkung 2.9 tritt nicht auf, wenn eine Spaltenpivotisierung durchgeführt wird !

Folgerung 2.10 *Ist der Gaußsche Algorithmus mit Spaltenpivotisierung zur LR-Faktorisierung von $A \in \mathbb{R}^{m\times m}$ durchführbar, so ist er numerisch gutartig.*
Ist A invertierbar, so ist der Gaußsche Algorithmus mit Spaltenpivotisierung durchführbar, falls $\text{cond}_\infty(A)$ nicht zu groß ist.

Beweis:
Bei Verwendung von Spaltenpivotisierung (vgl. Bemerkung 2.5) gilt

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

wegen $a_{ij}^{(k+1)} = a_{ij}^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} a_{kj}^{(k)}$, $i,j = k+1,\ldots,m$. Für die Zeilensummennorm $\|\cdot\|_\infty$ auf $\mathbb{R}^{m\times m}$ gilt deshalb

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

(vgl. auch Kielbasinski/Schwetlick, Aussage 5.2.1).
Aus Satz 2.8 ergibt sich dann, daß wegen $F_\infty(A) \leq 1 + 3\sum_{k=2}^{m} 2^{k-1} \leq 1 + 3\cdot 2^m$ der Gaußsche Algorithmus mit Spaltenpivotisierung numerisch gutartig ist, falls er durchführbar ist. Für die Durchführbarkeit ist hinreichend, daß LR invertierbar ist. Dies gilt im Fall $A \in \mathbb{R}^{m\times m}$ invertierbar, falls nach dem Störungslemma $\|A - LR\|_\infty \|A^{-1}\|_\infty < 1$ oder $(\tau + O(\tau^2))(1 + 3\cdot 2^m)\text{cond}_\infty(A) < 1$. □

Dies bedeutet, daß der Gaußsche Algorithmus mit Spaltenpivotisierung auf der Menge aller invertierbaren Matrizen A mit nicht zu großer Kondition $\text{cond}_\infty(A)$ von A

durchführbar und numerisch gutartig ist. Die Abschätzung $F_\infty(A) \leq 1 + 3 \cdot 2^m$ in Folgerung 2.10 ist in vielen Anwendungsfällen zu pessimistisch. Praktisch geht man von einem linearen Wachstum von $F_\infty(A)$ mit m aus ("Faustregel": $\tau\, m\, \mathrm{cond}_\infty(A) < 1$).

Für die Rückwärtselimination ist die Situation einfacher.

Satz 2.11 *Das lineare Gleichungssystem $Rx = c$ mit invertierbarer oberer Dreiecksmatrix R werde durch den folgenden Algorithmus (Rückwärtselimination) gelöst:*

$$x_m := \frac{c_m}{r_{mm}}, \quad x_i := \frac{1}{r_{ii}} \left(c_i - \sum_{j=i+1}^{m} r_{ij} x_j \right), \quad i = m-1, \ldots, 1.$$

Erfolgt dies in einer t-stelligen Gleitkommaarithmetik mit relativer Rechengenauigkeit τ wobei $\tau m < 1$, so genügt die berechnete Lösung x der Gleichung $(R + \delta R)x = c$ mit einer oberen Dreiecksmatrix δR mit der Eigenschaft $\|\delta R\|_\infty \leq (m\tau + O(\tau^2))\|R\|_\infty$. Insbesondere ist der Algorithmus numerisch gutartig.

Beweis:
Bei Rechnung in einer t-stelligen Gleitkommaarithmetik gilt für $i = 1, \ldots, m-1$:

$$s_i = \mathrm{fl}_t \left(\sum_{j=i+1}^{m} r_{ij} x_j \right) = \sum_{j=i+1}^{m} r_{ij} x_j (1 + \eta_{ij}) \prod_{k=j}^{m} (1 + \varepsilon_{ik}) = \sum_{j=i+1}^{m} r_{ij} x_j (1 + \varepsilon_{ij}),$$

wobei $|\varepsilon_{ik}|, |\eta_{ij}| \leq \tau$, $k = j, \ldots, m$, $j = i+1, \ldots, m$, und

$$\begin{aligned} |\varepsilon_{ij}| &= \left| (1 + \eta_{ij}) \prod_{k=j}^{m} (1 + \varepsilon_{ik}) - 1 \right| \leq (m - j + 1)\tau + O(\tau^2) \\ x_m &= \frac{c_m}{r_{mm}(1 + \delta_m)} \quad \text{mit} \quad |\delta_m| \leq \tau \\ x_i &= \frac{c_i - s_i}{r_{ii}(1 + \delta_i)} \quad \text{mit} \quad |\delta_i| \leq 2\tau + O(\tau^2). \end{aligned}$$

Wir setzen nun

$$\delta R_{ij} := \begin{cases} r_{ij}\varepsilon_{ij} & , \ i < j \\ r_{ii}\delta_i & , \ i = j \\ 0 & , \ \text{sonst} \end{cases}$$

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

$$\begin{aligned} \|\delta R\|_\infty &= \max_{i=1,\ldots,m} \sum_{j=i}^{m} |\delta R_{ij}| = \max_{i=1,\ldots,m} \left\{ \sum_{j=i+1}^{m} |r_{ij}\varepsilon_{ij}| + |r_{ii}\delta_i| \right\} \\ &\leq (m\tau + O(\tau^2)) \max_{i=1,\ldots,m} \sum_{j=i}^{m} |r_{ij}| = (m\tau + O(\tau^2))\|R\|_\infty \end{aligned}$$ □

Bemerkung 2.12 *(Skalierung)*
Die Lösungen des linearen Gleichungssystems $Ax = b$ verändern sich nicht, wenn $Ax = b$ zeilenweise mit geeigneten positiven Faktoren multipliziert wird. Dies entspricht der Multiplikation von A und b mit einer Diagonalmatrix $D = \operatorname{diag}(d_1, \ldots, d_m)$ mit $d_i > 0$, $i = 1, \ldots, m$. Läßt sich durch geeignete Wahl von D die Kondition von A verkleinern ?
Es sei $A \in \mathbb{R}^{m\times m}$ invertierbar, a^i bezeichne die i-te Zeile von A und wir betrachten die Diagonalmatrix $D = \operatorname{diag}(d_1, \ldots, d_m)$, wobei

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

Dann gilt:

$$\|DA\|_\infty = \|A\|_\infty\,,\ \|(DA)^{-1}\|_\infty \le \|A^{-1}\|_\infty \quad \textit{und}$$

$$\frac{\min\limits_{k=1,\ldots,m} \|a^k\|_1}{\max\limits_{k=1,\ldots,m} \|a^k\|_1} \operatorname{cond}_\infty(A) \le \operatorname{cond}_\infty(DA) \le \operatorname{cond}_\infty(A).$$

<u>Beweis:</u> *Es gilt zunächst* $\|DA\|_\infty = \max\limits_{i=1,\ldots,m} \sum\limits_{j=1}^{m} |d_i a_{ij}| = \max\limits_{i=1,\ldots,m} d_i \|a^i\|_1 = \|A\|_\infty$.
Ferner gilt $\|(DA)^{-1}\|_\infty \le \|A^{-1}\|_\infty \|D^{-1}\|_\infty = \|A^{-1}\|_\infty \frac{1}{\min\limits_{i=1,\ldots,m} d_i} = \|A^{-1}\|_\infty$ *und*
$\|A^{-1}\|_\infty = \|(DA)^{-1}D\|_\infty \le \|(DA)^{-1}\|_\infty \max\limits_{i=1,\ldots,m} d_i$. □
Dies bedeutet: Wählt man D durch $(*)$, so verkleinert sich die Kondition $\operatorname{cond}_\infty$ 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 $\operatorname{cond}_\infty$.

Bemerkung 2.13 *(Nachiteration)*
Es sei $x \in \mathbb{R}^m$ die durch LR-Faktorisierung und Rückwärtselimination berechnete Computer-Lösung des linearen Gleichungssystems $Ax = b$ und $x_* \in \mathbb{R}^m$ 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 2.3), so ist auch diese Näherung fehlerbehaftet, aber i. a. besser. Dies führt zur Idee der iterativen Fortsetzung dieses Prozesses, d. h., zur sog. <u>Nachiteration</u>:

- *x^0 Näherungslösung von $Ax = b$ (aus LR-Faktorisierung erhalten);*

- *für $n = 0, \ldots, n_{\max}$ bestimme h^n aus dem linearen Gleichungssystem $Ah = b - Ax^n$ durch LR-Faktorisierung und setze*
  $x^{n+1} := x^n + h^n = x^n + (LR)^{-1}(b - Ax^n) = (LR)^{-1}(LR - A)x^n + (LR)^{-1}b$;

- *die letztere Iterierte ist eine "gute" Lösung von $Ax = b$, da die Folge (x^n) 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: (Konditionsschätzung)
Das Ziel bestehe darin, $\text{cond}_\infty(A)$ für eine invertierbare Matrix $A \in \mathbb{R}^{m\times m}$ näherungsweise zu berechnen. Die Berechnung von $\|A\|_\infty$ ist kein Problem, allerdings ist es ein Aufwandsproblem, A^{-1} zu berechnen ($O(m^3)$ Operationen vgl. Bem. 2.4).
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:

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

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

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

Praktisch ist nun $\|L^{-1}\|_\infty$ meist wesentlich kleiner als 2^{m-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: $x_1 := 1,\ x_i := \pm 1,\ i = 2, \ldots, m$. Für $y = (R^T)^{-1}x$ gilt dann $y_1 = \frac{x_1}{r_{11}}$,
$y_i = -\sum_{j=1}^{i-1} \frac{r_{ji}}{r_{ii}} y_{j+} + \frac{x_i}{r_{ii}},\ i = 2, \ldots, m.$

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

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

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

2.2 Householder-Orthogonalisierung und Quadratmittel-Probleme

Ziel: Transformation einer Matrix auf Dreiecksgestalt mit Hilfe von orthogonalen Matrizen, die die Kondition nicht „verschlechtern". Anwendung auf Quadratmittel-Probleme der Form $\min_{x \in \mathbb{R}^m} \|Ax - b\|_2^2$ für $A \in \mathbb{R}^{n \times m}$, $b \in \mathbb{R}^n$ mit $n \geq m$.

Definition 2.14
Eine Matrix $Q \in \mathbb{R}^{m \times m}$ heißt orthogonal, falls Q invertierbar mit $Q^{-1} = Q^\top$.
Für jedes $u \in \mathbb{R}^m$ mit $\|u\|_2 = 1$ heißt die Matrix $H := I - 2uu^\top$ Householder-Spiegelung(smatrix).

Lemma 2.15

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

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

Insbesondere gilt $\|Q\|_2 = 1$, $cond_2(Q) = 1$ und $cond_2(QA) = cond_2(A)$ (falls A invertierbar ist).

b) Jede Householder-Spiegelungsmatrix ist symmetrisch und orthogonal.

Beweis:

a) Für jedes $x \in \mathbb{R}^m$ gilt:
$\|Qx\|_2^2 = \langle Qx, Qx \rangle = \langle Q^\top Qx, x \rangle = \langle x, x \rangle = \|x\|_2^2 \rightsquigarrow \|Q\|_2 = 1$.
Mit Q ist auch $Q^\top$ orthogonal wegen $(Q^\top)^{-1} = (Q^{-1})^\top = (Q^\top)^\top = Q$.
$\rightsquigarrow \|Q^\top\|_2 = 1$ und $\mathrm{cond}_2(Q) = \|Q^\top\|_2 \|Q\|_2 = 1$.
Ist A invertierbar, so gilt
$\mathrm{cond}_2(QA) = \|(QA)^{-1}\|_2 \|QA\|_2 = \|A^{-1}Q^\top\|_2 \|A\|_2 = \|A^{-1}\|_2 \|A\|_2 = \mathrm{cond}_2(A)$.
Es sei nun $\bar{y} \in \mathbb{R}^m$, $\|\bar{y}\|_2 = 1$ 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: $\|QA\|_2 \leq \|Q\|_2 \|A\|_2 = \|A\|_2$.

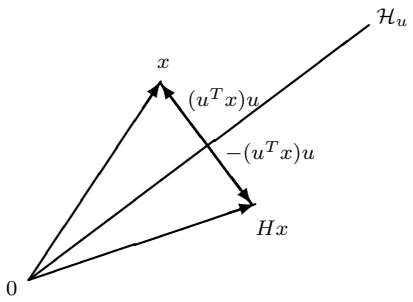
b) Eine Householder-Spiegelung ist nach Definition symmetrisch, d. h. es gilt $H = H^\top$. Ferner gilt:

$$H^2 = HH = (I - 2uu^\top)(I - 2uu^\top) = I - 4uu^\top + 4u(u^\top u)u^\top = I$$

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

Bemerkung 2.16 *Der Begriff „orthogonal“ rührt daher, daß die Zeilen und die Spalten einer orthogonalen Matrix als Vektoren in $\mathbb{R}^m$ orthogonal zueinander sind. Der Begriff „Spiegelung“ hat seinen Ursprung darin, daß die Matrix $H = I - 2uu^\top$ eine Spiegelung des $\mathbb{R}^m$ an der Hyperebene $\mathcal{H}_u := \{x \in \mathbb{R}^n : u^\top x = 0\}$ bewirkt.*

Schreibt man nämlich ein beliebiges $x \in \mathbb{R}^n$ in der Form $x = (u^\top x)u + (x - (u^\top x)u)$, so gilt $Hx = -(u^\top x)u + (x - (u^\top x)u)$. Obwohl von ähnlicher Art wie die Matrizen $L_{ij}(\beta)$ in Kap. 2.1, so sind Householder-Spiegelungen orthogonal und die $L_{ij}(\beta)$ nicht!



Lemma 2.17
Seien H eine Householder-Matrix, $e^1 := (1, 0, \ldots, 0)^\top$, $x \in \mathbb{R}^m$ mit $x \neq \alpha e^1$, $\forall \alpha \in \mathbb{R}$. Dann gilt $Hx = (I - 2uu^\top)x = \sigma e^1$ 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 $Hx = x - 2u^\top x u = \sigma e^1$ 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 2.15 folgt $\|Hx\|_2 = \|x\|_2 = |\sigma|$, d.h. $\sigma = \pm\|x\|_2$.
Haben umgekehrt u und σ die angegebene Form, so gilt

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

Daraus folgt

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

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

Algorithmus 2.18 *(QR-Faktorisierung durch Householder-Orthogonalisierung)*
1. Schritt:
a_1 sei die erste Spalte von A; hat a_1 die Form αe^1, so ist der erste Schritt beendet; ansonsten bestimme $u_1 \in \mathbb{R}^m$ so, daß $H_1 a_1 = (I - 2u_1 u_1^\top)a_1 = \|a_1\|_2 e^1$ und $\|u_1\|_2 = 1$. Setze $A^{(1)} = H_1 A$.
k-ter Schritt:
$A^{(k-1)}$ habe die Gestalt

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

mit den "Restspalten" $a_i^{(k-1)}$, $i = k, \ldots, m$. Bestimme $u_k \in \mathbb{R}^{m-k+1}$ so, daß

$$H_k a_k^{(k-1)} = (I - 2u_k u_k^\top) a_k^{(k-1)} = \|a_k^{(k-1)}\|_2 \, e^1 \in \mathbb{R}^{m-k+1} \text{ und } \|u_k\|_2 = 1.$$

Transformiere alle "Restspalten" $a_i^{(k-1)}$, $i = k, \ldots, m$, mit H_k und bezeichne die neue Matrix mit $A^{(k)}$.
m-ter Schritt:
Setze $R = A^{(m-1)} = \begin{pmatrix} \diagdown \end{pmatrix}$

Satz 2.19 *Zu jeder Matrix $A \in \mathbb{R}^{m\times m}$ existiert eine orthogonale Matrix $Q \in \mathbb{R}^{m\times m}$ und eine rechte obere Dreiecksmatrix $R \in \mathbb{R}^{m\times m}$, so daß*

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

Ist A invertierbar, so auch R und es gilt $cond_2(A) = cond_2(R)$.

Beweis: Es seien $H_1, \ldots, H_{m-1}$ die in Algorithmus 2.19 definierten Householder-Spiegelungen und wir definieren die folgenden Matrizen in $\mathbb{R}^{m\times m}$:

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

Dann sind alle Q_k symmetrisch und orthogonal. Überdies ist auch $Q := Q_{m-1} \cdots Q_1$ orthogonal. Wegen $QA = A^{(m-1)} = R$ ist damit der erste Teil gezeigt. Der zweite Teil folgt aber unmittelbar aus Lemma 2.16. □

Bemerkung 2.20 *Anzahl der Rechenoperationen einer QR-Faktorisierung:*

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

Die Anzahl der Rechenoperationen ist also etwa doppelt so groß wie beim Gaußschen Algorithmus. Die Householder-Orthogonalisierung ist besonders dann empfehlenswert, wenn $cond_2(A)$ „groß" ist. Die Householder-Orthogonalisierung ist ein numerisch gutartiges Verfahren (vgl. Kielbasinski/Schwetlick, Kap. 10.2). Soll die Matrix Q bestimmt werden, so sind $\frac{4}{3}m^3 + O(m^2)$ opms erforderlich.

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

Beispiel 2.21 *(lineare Regression)*
Gegeben seien statistische Daten $(t_i, x_i) \in \mathbb{R} \times \mathbb{R}$, $i = 1, \ldots, n$, *die z.B. gemessenen Werten an Zeitpunkten* t_i *entsprechen, und reelle Funktionen* φ_j, $j = 1, \ldots, m$.
Gesucht ist nun eine Linearkombination $\sum_{j=1}^{m} c_j \varphi_j$, *so daß sie die gegebenen Daten bestmöglich im Quadratmittel-Sinn annimmt, d.h. die gesuchten Koeffizienten lösen das Problem*

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

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

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

Satz 2.22 *Es sei* $A \in \mathbb{R}^{n \times m}, b \in \mathbb{R}^n$, $m \leq n$. *Das Quadratmittel-Problem besitzt eine Lösung* x_*. *Alle solchen Quadratmittellösungen sind auch Lösungen der Normalgleichungen*

$$A^\top A x = A^\top b$$

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* $m - \mathrm{rg}(A)$. $\mathcal{L}$ *ist einelementig, wenn* $\mathrm{rg}(A) = m$. *Es existiert genau ein* $x^N \in \mathcal{L}$, *so daß*

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

Beweisskizze: Wir definieren $\Phi(x) := \frac{1}{2}\|Ax - b\|_2^2$ für alle $x \in \mathbb{R}^m$.
Es gilt: $\Phi(x) = \frac{1}{2}\langle Ax - b, Ax - b\rangle = \frac{1}{2}\left[\langle A^\top A x, x\rangle - 2\langle A^\top b, x\rangle + \langle b, b\rangle\right]$
$\rightsquigarrow \Phi'(x) = A^\top A x - A^\top b$ (Gradient), $\Phi''(x) = A^\top A \in \mathbb{R}^{n \times n}$ (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 $A^\top A$ invertierbar, falls $\mathrm{rg}(A) = m$. Ferner existiert ein Element x^N in $\mathcal{L}$, so daß sein Euklidischer Abstand zum Nullelement in $\mathbb{R}^m$ minimal ist. Dieses Element steht senkrecht auf $\ker(A)$ und ist eindeutig bestimmt. □

Definition 2.23 *Es sei* x^N *die eindeutig bestimmte Normallösung des Quadratmittel-Problems. Dann heißt die Matrix* $A^+ \in \mathbb{R}^{m \times n}$ *mit der Eigenschaft* $A^+ b = x^N$ $(\forall b \in \mathbb{R}^n)$ *verallgemeinerte Inverse, Moore-Penrose-Inverse oder Pseudoinverse von* $A \in \mathbb{R}^{n \times m}$.

Bemerkung 2.24 *(Pseudoinverse und Singulärwertzerlegung)*
Die Pseudoinverse A^+ *von A ist die einzige Lösung des folgenden Systems von Matrizengleichungen (Penrose 1955):*

$$\begin{aligned} AXA &= A & (AX)^\top &= AX \\ XAX &= X & (XA)^\top &= XA. \end{aligned}$$

Singulärwertzerlegung: Für $A \in \mathbb{R}^{n\times m}$ mit $\mathrm{rg}(A) = r \leq \min\{n,m\}$ existieren Zahlen $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$ ("Singulärwerte") und orthogonale Matrizen $U \in \mathbb{R}^{n\times n}$ und $V \in \mathbb{R}^{m\times m}$, so dass

$$\Sigma = U^\top AV\,, \ \text{wobei } \Sigma = (\sigma_i\delta_{ij})_{\substack{i=1,\ldots,n\\ j=1,\ldots,m}} \in \mathbb{R}^{n\times m}$$

mit $\sigma_{r+1} = \cdots = \sigma_n = 0$, $\sigma_i = \sqrt{\lambda_i}$, $i = 1,\ldots,r$, und λ_i ist Eigenwert von $A^\top A$.

Die Pseudoinverse von $A \in \mathbb{R}^{n\times m}$ besitzt die Darstellung

$$A^+ = V\Sigma^+U^\top\,, \ \text{wobei } \Sigma^+ = (\tau_j\delta_{ji})_{\substack{j=1,\ldots,m\\ i=1,\ldots,n}} \in \mathbb{R}^{m\times n}$$

mit $\tau_j = \sigma_j^{-1}$, $j = 1,\ldots,r$, $\tau_j = 0$, $j = r+1,\ldots,\min\{n,m\}$.
(Lit.: Hämmerlin-Hoffmann, Kap. 2.6)

Ist $A \in \mathbb{R}^{n\times m}$ eine Matrix mit Rang $r = \mathrm{rg}(A) \leq \min\{n,m\}$, so gilt:

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

Erweiterung des Konditionsbegriffs für $A \in \mathbb{R}^{n\times m}$: $\mathrm{cond}(A) := \|A^+\|\|A\|$.
Mit Hilfe der Singulärwertzerlegung von $A \in \mathbb{R}^{n\times m}$ gilt:

$$\mathrm{cond}_2(A) = \|A^+\|_2\|A\|_2 = \|\Sigma^+\|_2\|\Sigma\|_2 = \frac{\sigma_1}{\sigma_r} = \sqrt{\frac{\lambda_1}{\lambda_r}}$$

wobei $r = \mathrm{rg}(A)$ und $\lambda_1 \geq \cdots \geq \lambda_r > 0$ die positiven Eigenwerte von $A^\top A$ sind.
(Literatur: A. Ben-Israel, T. Greville: Generalized Inverses (Second Edition), Springer, New York, 2003)

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 $A^\top Ax = A^\top b$ vorzuziehen!

Satz 2.25 *(Berechnung von Normallösungen)*
Es sei $A \in \mathbb{R}^{n\times m}$ spalteninvertierbar, d.h. $\mathrm{rg}(A) = m \leq n$, und x^N sei die eindeutig bestimmte Normallösung. Dann existiert eine orthogonale Matrix $Q \in \mathbb{R}^{n\times n}$ und eine invertierbare rechte obere Dreiecksmatrix $R \in \mathbb{R}^{m\times m}$, so daß

$$Rx^N = r_1 \quad \textit{mit} \quad QA = \begin{pmatrix} R \\ - - - - - \\ 0 \end{pmatrix} \quad \textit{und} \quad Qb = \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} \begin{matrix} \}m \\ \}n-m \end{matrix}.$$

Beweis:
Analog zu Satz 2.19 existieren orthogonale Matrizen Q_k, $k = 1,\ldots,m$, so daß mit $Q = Q_mQ_{m-1}\cdots Q_1$ die Matrix QA die Gestalt

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

mit einer invertierbaren oberen Dreiecksmatrix R besitzt. Dann gilt:

$$\begin{aligned}\Phi(x) &= \tfrac{1}{2}\|Ax-b\|_2^2 = \tfrac{1}{2}\|Q(Ax-b)\|_2^2 \qquad \text{(Lemma 2.15 !)}\\ &= \tfrac{1}{2}\left\|\left(\begin{array}{c} R \\ ----- \\ 0 \end{array}\right)x - Qb\right\|_2^2 = \tfrac{1}{2}\|Rx-r_1\|_2^2 + \tfrac{1}{2}\|r_2\|_2^2.\end{aligned}$$

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

3 Iterative Verfahren für große lineare Gleichungssysteme

Gegeben: $A = (a_{ij}) \in \mathbb{R}^{m\times m}$ invertierbar, $b \in \mathbb{R}^m$, m groß (d.h. $10^3 < m \leq 10^8$).
Gesucht: Lösung $x \in \mathbb{R}^m$ von $Ax = b$.

Spezielle Situation: A ist "schwach besetzt" (engl.: sparse), d.h. A besitzt relativ wenige von Null verschiedene Elemente. Insbesondere treten 2 Fälle auf:
(i) Matrizen mit spezieller regelmäßiger Struktur, z.B. Bandstruktur.
(ii) Matrizen, die unregelmäßig schwach besetzt sind.

Bemerkung 3.1 *Bei großen schwach besetzten Matrizen führen die auf Dreieckszerlegung basierenden Verfahren (Gaußscher Algorithmus, Householder-Orthogonalisierung) zu großen Rechenzeiten (obwohl meist nur Nullen multipliziert oder addiert werden) und evtl. auch zu Speicherplatzproblemen (da die entstehenden Dreiecksmatrizen häufig "voll besetzt" sind). Zum Beispiel benötigt der Gaußsche Algorithmus* $\frac{m^3}{3} + O(m^2)$ *Operationen. Damit sind für ein großes lineares Gleichungssystem mit* $m = 10^5$ *etwa* $0.33 \cdot 10^{15}$ *Operationen erforderlich. Steht nun ein Computer mit* 10^8 *Operationen pro Sekunde zur Verfügung, so benötigt er etwa* $0.33 \cdot 10^7$ *Sekunden. Im Vergleich entspricht ein Jahr etwa* $0.82 \cdot 10^7$ *Sekunden.*

Ausweg: Iterative Verfahren, die pro Schritt nur eine Multiplikation von Matrix mal Vektor erfordern (d.h. etwa m^2 *Operationen).*

3.1 Splitting-Methoden

Grundidee von Splitting-Methoden:
Mit einer invertierbaren Matrix $C \in \mathbb{R}^{m\times m}$ wird $Ax = b$ äquivalent umgeformt in eine Fixpunktgleichung:

$$Ax = b \Longleftrightarrow Cx = (C-A)x + b \Longleftrightarrow x = (E - C^{-1}A)x + C^{-1}b$$

Die Matrix A wird aufgesplittet in die Matrizen C und $A - C$.
Ausgehend von der Fixpunktgleichung wird wie im Banachschen Fixpunktsatz das Iterationsverfahren

$$x_n = (E - C^{-1}A)x_{n-1} + C^{-1}b\,,\ n = 1, 2, \ldots,\ x_0 \in \mathbb{R}^m,$$

angesetzt. Die Verfahren unterscheiden sich durch die Wahl der Matrix C.

Problemstellung: Wann konvergieren Iterationsverfahren vom allgemeinen Typ

$$\text{(IV)} \quad x_n = Bx_{n-1} + d\,,\ n = 1, 2, \ldots,\ x_0 \in \mathbb{R}^m,$$

wobei $B \in \mathbb{R}^{m\times m}$, $d \in \mathbb{R}^m$? Da auf $\mathbb{R}^m$ alle Normen äquivalent sind, interessiert uns ein norm-unabhängiges Konvergenzkriterium.

Im folgenden bezeichnet $\mathbb{C}$ die Menge der komplexen Zahlen und $\mathbb{C}^m$ den linearen Raum aller Elemente der Form $(x_1, \ldots, x_m)$ mit $x_i \in \mathbb{C}$, $i = 1, \ldots, m$. Eine Reihe von Normen auf dem $\mathbb{R}^m$ lassen sich sofort zu Normen auf $\mathbb{C}^m$ erweitern (z.B. $\|\cdot\|_p$ mit $p \in [1, +\infty]$). Für $B \in \mathbb{R}^{m\times m}$ bezeichnet $\rho(B) := \max\{|\lambda| : \lambda \in \mathbb{C},\ \det(B - \lambda E) = 0\}$ den Spektralradius von B.

Lemma 3.2 *Es sei $B \in \mathbb{R}^{m\times m}$. Dann gilt $\rho(B) < 1$ gdw. eine Norm $\|\cdot\|_*$ auf $\mathbb{C}^m$ existiert, so dass für die zugehörige Matrixnorm gilt $\|B\|_* < 1$.*

Beweis:

($\Leftarrow$) Es sei $\|\cdot\|_*$ eine Norm auf $\mathbb{C}^m$, so daß $\|B\|_* < 1$. Ferner sei $\lambda \in \mathbb{C}$ ein beliebiger Eigenwert von B. Dann existiert ein $0 \neq z \in \mathbb{C}^m$ (Eigenvektor) mit $Bz = \lambda z$. Folglich gilt

$$\|Bz\|_* = |\lambda|\|z\|_* \quad \rightsquigarrow \quad \left\| B\frac{z}{\|z\|_*} \right\|_* = |\lambda| \quad \rightsquigarrow \quad \|B\|_* \geq |\lambda|$$

und damit $\rho(B) \leq \|B\|_* < 1$.

($\Rightarrow$) Es gelte $\rho(B) < 1$ und es sei $\varepsilon > 0$ so gewählt, dass $\rho(B) + \varepsilon < 1$. Es sei nun $J \in \mathbb{C}^{m\times m}$ die Jordan-Normalform zu B, d.h. es existiert eine invertierbare Matrix $T \in \mathbb{C}^{m\times m}$, so dass

$$J = T^{-1}BT = \begin{pmatrix} \lambda_{i_1} & * & 0 & \cdots & 0 & 0 \\ 0 & \lambda_{i_2} & * & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_{i_{m-1}} & * \\ 0 & 0 & 0 & \cdots & 0 & \lambda_{i_m} \end{pmatrix}$$

wobei λ_{i_j}, $j = 1, \ldots, m$, die Eigenwerte von B sind und $*$ für 0 oder 1 steht. Es bezeichne nun D_ε die Diagonalmatrix

$$D_\varepsilon = \mathrm{diag}(1, \varepsilon, \ldots, \varepsilon^{m-1})$$

und wir betrachten die Matrix

$$J_\varepsilon := D_\varepsilon^{-1} J D_\varepsilon = (TD_\varepsilon)^{-1} B (TD_\varepsilon).$$

Dann hat J_ε die Form

$$J_\varepsilon = \begin{pmatrix} \lambda_{i_1} & *_\varepsilon & 0 & \cdots & 0 & 0 \\ 0 & \lambda_{i_2} & *_\varepsilon & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_{i_{m-1}} & *_\varepsilon \\ 0 & 0 & 0 & \cdots & 0 & \lambda_{i_m} \end{pmatrix}$$

wobei $*_\varepsilon$ für 0 oder ε steht.
Wir definieren nun die Norm $\|\cdot\|_*$ auf $\mathbb{C}^m$ durch

$$\|x\|_* := \|(TD_\varepsilon)^{-1}x\|_1 \quad (\forall x \in \mathbb{C}^m),$$

wobei $\|y\|_1 := \sum_{i=1}^m |y_i|$ für jedes $y = (y_1, \ldots, y_m) \in \mathbb{C}^m$. Dann ergibt sich

$$\begin{aligned} \|Bx\|_* &= \|(TD_\varepsilon)^{-1}B(TD_\varepsilon)(TD_\varepsilon)^{-1}x\|_1 = \|J_\varepsilon(TD_\varepsilon)^{-1}x\|_1 \le \|J_\varepsilon\|_1\|x\|_* \\ \|B\|_* &\le \|J_\varepsilon\|_1 \le \max\{|\lambda_i| + \varepsilon : i = 1, \ldots, m\} = \rho(B) + \varepsilon < 1. \end{aligned}$$

□

Mit Hilfe des Lemmas beweisen wir jetzt ein notwendiges und hinreichendes Konvergenzkriterium für Itegrationsverfahren.

Satz 3.3 *(Konvergenz von (IV))*
Sei $B \in \mathbb{R}^{m\times m}$. Wir betrachten das Iterationsverfahren

(IV) $\quad x_n := Bx_{n-1} + d\,,\ n = 1, 2, \ldots,$

Das Verfahren (IV) ist für alle $x_0 \in \mathbb{R}^m$ und $d \in \mathbb{R}^m$ konvergent gegen $(I-B)^{-1}d$ gdw. $\rho(B) < 1$.

Beweis:

($\Rightarrow$) Sei λ betragsgrößter Eigenwert von B, d.h. $|\lambda| = \rho(B)$ und $z \in \mathbb{C}^m$, $z \neq 0$ ein Eigenvektor zu λ. Wir betrachten das Iterationsverfahren (IV) für $x_0 = z$ und $d = 0$, d.h.

$$x_n = Bx_{n-1} = B^n z = \lambda^n z\,.$$

Da nach Voraussetzung die Folge (x_n) gegen 0 konvergiert, muss $\rho(B) = |\lambda| < 1$ gelten.

($\Leftarrow$) Es gelte $\rho(B) < 1$. Dann ist $\lambda = 1$ kein Eigenwert von B. Folglich ist $I - B$ invertierbar. Seien $x_0 \in \mathbb{R}^m$ und $d \in \mathbb{R}^m$. Für die von (IV) erzeugte Folge (x_n) gilt

$$\begin{aligned} x_n &= Bx_{n-1} + d = B(Bx_{n-2} + d) + d = B^n x_0 + \sum_{j=0}^{n-1} B^j d \\ &= B^n x_0 + (I-B)^{-1}(I-B)\sum_{j=0}^{n-1} B^j d = B^n x_0 + (I-B)^{-1}(I-B^n)d\,. \end{aligned}$$

Gemäß Lemma 3.2 wählen wir eine Norm $\|\cdot\|_*$ auf $\mathbb{C}^m$, so daß $\|B\|_* < 1$. Dann gilt $\|B^n x_0\|_* \le \|B^n\|_* \|x_0\|_* \le \|B\|_*^n \|x_0\|_*$ und

$$\begin{aligned} \|x_n - (I-B)^{-1}d\|_* &= \|B^n x_0 - (I-B)^{-1}B^n d\|_* \\ &\le \|B^n x_0\|_* + \|(I-B)^{-1}B^n d\|_* \\ &\le \|B\|_*^n \|x_0\|_* + \|(I-B)^{-1}\|_* \|B\|_*^n \|d\|_* \end{aligned}$$

Da $(\|B\|_*^n)$ eine Nullfolge ist, konvergiert die Folge (x_n) gegen $(I-B)^{-1}d$. □

Folgerung 3.4 *Seien $A, C \in \mathbb{R}^{m\times m}$ invertierbar und $b \in \mathbb{R}^m$. Das Splitting-Verfahren*

$$x_n = (I - C^{-1}A)x_{n-1} + C^{-1}b,\ n = 1, 2, \ldots,\ x_0 \in \mathbb{R}^m,$$

konvergiert gegen $A^{-1}b$ gdw. $\rho(I - C^{-1}A) < 1$.

Beweis: Die Aussage folgt mit $B = I - C^{-1}A$ und $d = C^{-1}b$ aus Satz 3.3. Es gilt $(I-B)^{-1}d = (C^{-1}A)^{-1}C^{-1}b = A^{-1}CC^{-1}b = A^{-1}b$. □

Beispiel 3.5 *(Splittingverfahren)*
Für die Diagonalelemente von A gelte $a_{ii} \neq 0$, $i = 1, \ldots, m$ (für reguläres A läßt sich diese Voraussetzung immer erfüllen).

a) Gesamtschrittverfahren oder Jacobi-Verfahren:
Wir wählen $C := D := diag\,(a_{11}, \ldots, a_{mm})$ und erhalten

$$B := I - D^{-1}A = \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} & \cdots & -\frac{a_{1m}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & \cdots & -\frac{a_{2m}}{a_{22}} \\ \vdots & \vdots & \vdots & \vdots \\ -\frac{a_{m1}}{a_{mm}} & -\frac{a_{m2}}{a_{mm}} & \cdots & 0 \end{pmatrix}$$

Das Gesamtschrittverfahren nimmt dann mit $x_n = (x_{n,1}, \ldots, x_{n,m})$ die Form an

$$x_n := (I - D^{-1}A)x_{n-1} + D^{-1}b\,,\ n = 1, 2, \ldots\,,\ \ \textit{mit } x_0 \in \mathbb{R}^m \textit{ beliebig.}$$

In Komponenten-Schreibweise hat es die Form

$$x_{n,i} := \frac{1}{a_{ii}} \left(-\sum_{\substack{j=1 \\ j\neq i}}^{m} a_{ij} x_{n-1,j} + b_i \right)\,,\ i = 1, \ldots, m\,,\ n = 0, 1, 2, \ldots\,,\ x_0 \in \mathbb{R}^m \textit{ bel.}$$

Konvergenz gilt gdw. $\rho(I - D^{-1}A) = \rho(-D^{-1}(L+R)) < 1$.

b) Einzelschrittverfahren oder Gauß-Seidel Verfahren:
Wir zerlegen A in eine untere Dreiecksmatrix L, Diagonalmatrix D und eine obere Dreiecksmatrix R, d.h. $A = L + D + R$ mit $D := diag(a_{11}, \ldots, a_{mm})$, und

definieren $C := L + D$ *und* $B := I - C^{-1}A = -(L+D)^{-1}R$.
Dann nimmt das Einzelschrittverfahren die Form an

$$x_n := -(L+D)^{-1}Rx_{n-1} + (L+D)^{-1}b \quad \textit{oder} \quad (L+D)x_n := -Rx_{n-1} + b$$

für $n = 0, 1, 2, \ldots$ *bzw. in Komponenten-Schreibweise*

$$x_{n,i} = \frac{1}{a_{ii}} \left(-\sum_{j=1}^{i-1} a_{ij} x_{n,j} - \sum_{j=i+1}^{m} a_{ij} x_{n-1,j} + b_i \right), \; i = 1, \ldots, m \,, \; n = 1, 2, \ldots,$$

jeweils für beliebig gewähltes $x_0 \in \mathbb{R}^m$. *Man berechnet also für* $i = 1$ *zunächst* $x_{n,1}$, *setzt dies für* $i = 2$ *in die rechte Seite ein und berechnet* $x_{n,2}$, *setzt dies in die rechte Seite ein usw.*
Konvergenz gilt gdw. $\rho(-(L+D)^{-1}R) < 1$ *?*

c) *Relaxationsverfahren:*
Unter Beachtung der Zerlegung $A = L + \frac{1}{\omega}D + (1 - \frac{1}{\omega})D + R$ *wählen wir* $C := L + \frac{1}{\omega}D$ *und* $B(\omega) := -(L+\frac{1}{\omega}D)^{-1}((1-\frac{1}{\omega})D + R) = (D+\omega L)^{-1}((1-\omega)D - \omega R)$ *mit* $\omega \in \mathbb{R} \setminus \{0\}$. *Analog zum Einzelschrittverfahren hat das Relaxationsverfahren in Komponenten-Schreibweise die Gestalt*

$$x_{n,i} = \frac{1}{a_{ii}} \left(-\omega \sum_{j=1}^{i-1} a_{ij} x_{n,j} - (\omega - 1) a_{ii} x_{n-1,i} - \omega \sum_{j=i+1}^{m} a_{ij} x_{n-1,j} + \omega b_i \right)$$

für jedes $i = 1, \ldots, m$ *und* $n = 1, 2, \ldots$.
Frage: Wann und für welche $\omega \in \mathbb{R} \setminus \{0\}$ *gilt* $\rho(-(D+\omega L)^{-1}((\omega-1)D + \omega R)) < 1$*?*

Satz 3.6
Es sei $A \in \mathbb{R}^{m \times m}$ *strikt diagonaldominant, d.h.* $\sum\limits_{\substack{j=1 \\ j \neq i}}^{m} |a_{ij}| < |a_{ii}|$ *für jedes* $i = 1, \ldots, m$.

Dann sind das Gesamtschritt- und das Einzelschrittverfahren für jeden Startwert x_0 *in* $\mathbb{R}^m$ *gegen* $A^{-1}b$ *konvergent. Das Einzelschrittverfahren konvergiert i.a. schneller als das Gesamtschrittverfahren.*

Beweis: Wir definieren $\gamma := \max\limits_{i=1,\ldots,m} \frac{1}{|a_{ii}|} \sum\limits_{\substack{j=1 \\ j \neq i}}^{m} |a_{ij}|$ und wissen, dass nach Voraussetzung $\gamma < 1$ gilt. Ferner gilt $\| - D^{-1}(L+R)\|_\infty = \gamma < 1$. Daraus folgt nach Lemma 1.10, dass $I + D^{-1}(L+R) = D^{-1}A$ invertierbar ist. Deshalb ist A invertierbar.

(i) Gesamtschrittverfahren: Es gilt

$$\|B\|_\infty = \|E - D^{-1}A\|_\infty = \| - D^{-1}(L+R)\|_\infty = \gamma < 1$$

Aus Lemma 3.2 folgt deshalb $\rho(B) \leq \|B\|_\infty < 1$ und aus Satz 3.3 die Konvergenz des Gesamtschrittverfahrens gegen $(E-B)^{-1}d = (D^{-1}A)^{-1}D^{-1}b = A^{-1}b$.

(ii) Einzelschrittverfahren: Wir zeigen, dass $\|(L+D)^{-1}R\|_\infty \leq \gamma < 1$.
Dazu wählen wir ein beliebiges $x \in \mathbb{R}^m$ mit $\|x\|_\infty = 1$ und setzen $y := Bx$.

$$\rightsquigarrow (D+L)y = -Rx \quad \rightsquigarrow \sum_{j=1}^{i} a_{ij}y_j = -\sum_{j=i+1}^{m} a_{ij}x_j$$

$$\rightsquigarrow y_i = -\frac{1}{a_{ii}}\left(\sum_{j=1}^{i-1} a_{ij}y_j - \sum_{j=i+1}^{m} a_{ij}x_j\right)$$

$$\rightsquigarrow |y_i| \leq \frac{1}{|a_{ii}|}\left(\sum_{j=1}^{i-1}|a_{ij}||y_j| + \sum_{j=i+1}^{m}|a_{ij}||x_j|\right) \leq \frac{1}{|a_{ii}|}\left(\sum_{j=1}^{i-1}|a_{ij}||y_j| + \sum_{j=i+1}^{m}|a_{ij}|\right)$$

für jedes $i = 1, \ldots, m$. Wir definieren nun

$$\gamma_i := \frac{1}{|a_{ii}|}\left(\sum_{j=1}^{i-1}|a_{ij}|\gamma_j + \sum_{j=i+1}^{m}|a_{ij}|\right) \quad (\forall i = 1, \ldots, m)$$

und zeigen induktiv, dass $|y_i| \leq \gamma_i \leq \gamma$, $i = 1, \ldots, m$, gilt.
Sei zunächst $i = 1$. Dann gilt nach oben:

$$|y_1| \leq \frac{1}{|a_{11}|}\sum_{j=2}^{m}|a_{1j}| = \gamma_1 \leq \gamma\,.$$

Es gelte nun bereits $|y_j| \leq \gamma_j \leq \gamma$, $j = 1, \ldots, i-1$. Es folgt wieder

$$\begin{aligned}
|y_i| &\leq \frac{1}{|a_{ii}|}\left(\sum_{j=1}^{i-1}|a_{ij}|\gamma_j + \sum_{j=i+1}^{m}|a_{ij}|\right) = \gamma_i \\
&\leq \frac{1}{|a_{ii}|}\left(\sum_{j=1}^{i-1}|a_{ij}|\gamma + \sum_{j=i+1}^{m}|a_{ij}|\right) \\
&\leq \frac{1}{|a_{ii}|}\left(\sum_{j=1}^{i-1}|a_{ij}| + \sum_{j=i+1}^{m}|a_{ij}|\right) \leq \gamma\,.
\end{aligned}$$

Deshalb gilt $\|y\|_\infty = \|Bx\|_\infty \leq \max\{\gamma_i : i = 1, \ldots, m\}$ und folglich $\|B\|_\infty \leq \max\{\gamma_i : i = 1, \ldots, m\} \leq \gamma < 1$. □

Als nächstes beschäftigen wir uns mit dem Relaxationsverfahren und der Wahl von $\omega \in \mathbb{R} \setminus \{0\}$.

Satz 3.7 *Für jede Matrix $A \in \mathbb{R}^{m\times m}$ mit $a_{ii} \neq 0$, $i = 1, \ldots, m$, und jedes $\omega \in \mathbb{R} \setminus \{0\}$ gilt $\rho(B(\omega)) \geq |\omega - 1|$. Insbesondere ist für die Bedingung $\rho(B(\omega)) < 1$ notwendig, dass $\omega \in (0, 2)$ gilt.*

Beweis: Sei $\omega \in \mathbb{R} \setminus \{0\}$ und es seien λ_i, $i = 1, \ldots, m$, die Eigenwerte von $B(\omega)$. Dann gilt für das charakteristische Polynom $\varphi(\lambda) := \det(B(\omega) - \lambda E)$ nach dem Satz von Vieta, dass

$$\begin{aligned}\left|\prod_{j=1}^{m} \lambda_j\right| &= |\varphi(0)| = |\det(B(\omega))| \\ &= |\det((D+\omega L)^{-1})||\det((\omega-1)D + \omega R)| \\ &= \prod_{j=1}^{m} \frac{1}{|a_{jj}|} \prod_{i=1}^{m} |(\omega-1)a_{ii}| = |\omega - 1|^m.\end{aligned}$$

Daraus folgt

$$\rho(B(\omega)) = \max_{j=1,\ldots,m} |\lambda_j| \geq \left|\prod_{j=1}^{m} \lambda_j\right|^{\frac{1}{m}} = |\omega - 1|. \qquad \square$$

Satz 3.8 *Ist $A \in \mathbb{R}^{m\times m}$ symmetrisch und positiv definit, so gilt $\rho(B(\omega)) < 1$ für alle $\omega \in (0,2)$. Folglich konvergiert das Relaxationsverfahren für $\omega \in (0,2)$ gegen $A^{-1}b$.*

Beweis: Nach Voraussetzung erlaubt A eine Zerlegung der Form

$$A = L + D + L^\top, \text{ wobei } D = \operatorname{diag}(a_{11}, \ldots, a_{mm})$$

und $a_{ii} = \langle Ae_i, e_i\rangle > 0$, $i = 1, \ldots, m$, mit den kanonischen Einheitsvektoren $e_i \in \mathbb{R}^m$, $i = 1, \ldots, m$, gilt. Die Matrix $B(\omega)$ hat die Form

$$B(\omega) = \Big(\frac{1}{\omega}D + L\Big)^{-1}\Big(\frac{1-\omega}{\omega}D - L^\top\Big) = \Big(\frac{1}{\omega}D + L\Big)^{-1}\Big(\Big(\frac{1}{\omega}D + L\Big) - A\Big)$$

da $R = L^\top$ gilt wegen der Symmetrie von A. Wir setzen nun

$$B = \frac{1}{\omega}D + L \quad \text{and} \quad C = B - A.$$

Dann gilt also $B(\omega) = B^{-1}C$.
1. Schritt: Wir betrachten die Matrix $B + B^\top - A$. Es gilt

$$\begin{aligned}B + B^\top - A &= \frac{1}{\omega}D + L + \frac{1}{\omega}D + L^\top - (A + D + L^\top) \\ &= \Big(\frac{2}{\omega} - 1\Big)D = \frac{2-\omega}{\omega}D.\end{aligned}$$

Also ist $B + B^\top - A$ symmetrisch und positiv definit, falls $\omega \in (0,2)$.
2.Schritt: Wir betrachten die Matrix $Q = A - (B^{-1}C)^\top A(B^{-1}C)$. Es gilt

$$\begin{aligned}Q &= A - (B^{-1}C)^\top A(B^{-1}C) = A - (I - B^{-1}A)^\top A(I - B^{-1}A) \\ &= A - (A - (B^{-1}A)^\top A)(I - B^{-1}A) \\ &= (B^{-1}A)^\top A - AB^{-1}A + (B^{-1}A)^\top AB^{-1}A \\ &= (B^{-1}A)^\top (B + B^\top - A)(B^{-1}A).\end{aligned}$$

Deshalb ist auch Q symmetrisch und positiv definit, falls $\omega \in (0,2)$.
3.Schritt: Es sei $\omega \in (0,2)$ und $\lambda \in \mathbb{C}$ ein Eigenwert von $B(\omega) = B^{-1}C$ mit Eigenvektor $v \in \mathbb{C}^m$, $v \neq 0$. Dann gilt

$$\begin{aligned}\langle Av, v\rangle &= \langle Qv, v\rangle + \left\langle (B^{-1}C)^\top A(B^{-1}C)\right\rangle \\ &\geq \langle A(B^{-1}C)v, (B^{-1}C)v\rangle = \langle A\lambda v, \lambda v\rangle \\ &= \lambda\bar{\lambda}\langle Av, v\rangle = |\lambda|^2\langle Av, v\rangle.\end{aligned}$$

Also folgt $|\lambda| < 1$, da $\langle Av, v\rangle$ und $\langle Qv, v\rangle$ auch bzgl. des (hermiteschen) Skalarproduktes auf $\mathbb{C}^m$ positiv sind. Deshalb gilt $\rho(B(\omega)) < 1$ und die Aussage folgt aus Satz 3.3 bzw. Folgerung 3.4. □

Bemerkung 3.9 *Satz 3.7 liefert für $\omega = 1$ ein Konvergenzresultat für das Einzelschrittverfahren unter substantiell anderen Voraussetzungen als Satz 3.5. Die Idee der Relaxationsverfahren besteht darin, durch eine geeignete Wahl des Relaxationsparameters $\omega \in (0,2)$ eine Konvergenzbeschleunigung zu erreichen. $\omega \in (0,2)$ sollte so gewählt werden, dass $\rho(B(\omega))$ (näherungsweise) minimal wird. Für eine große Klasse von anwendungsrelevanten Matrizen A ist bekannt, dass für die Iterationsmatrizen B_G, B_E und $B(\omega)$ des Gesamtschritt-, Einzelschritt- und Relaxationsverfahrens gilt:*

$$\rho(B_E) = (\rho(B_G))^2 \qquad \min_{\omega\in(0,2)} \rho(B(\omega)) = \rho(B(\omega_*)) = \left(\frac{\rho(B_G)}{1+\sqrt{1-\rho(B_G)^2}}\right)^2,$$

falls die Eigenwerte von B_G reell sind und $\rho(B_G) < 1$ gilt. Überdies gilt:

$$\omega_* = \frac{2}{1+\sqrt{1-\rho(B_G)^2}} \in [1,2)$$

(Literatur: Stoer-Bulirsch: Numerische Mathematik, Band 2, Springer, Berlin 1990; Kapitel 8.3).

Bemerkung 3.10 *Für jede quadratische Matrix $A \in \mathbb{R}^{m\times m}$ gilt*

$$\rho(A) = \lim_{n\to\infty} \|A^n\|^{\frac{1}{n}} = \inf_{n\in\mathbb{N}} \|A^n\|^{\frac{1}{n}}.$$

Diese Formel bleibt auch für alle linearen stetigen Abbildungen eines linearen normierten Raumes X in sich richtig. Damit läßt sich ein alternativer Beweis von Lemma 3.2 zur Konstruktion einer Norm $\|\cdot\|_*$ auf X mit der Eigenschaft

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

für jedes $\varepsilon > 0$ ohne Verwendung der Jordan-Normalform von A führen. Dieser Beweis ist auch für den unendlichdimensionalen Fall verwendbar. Die Norm $\|\cdot\|_*$ wird dabei wie folgt definiert: Für $\varepsilon > 0$ sei $n \in \mathbb{N}$ mit $\|A^n\| \leq (\rho(A)+\varepsilon)^n$ gewählt.

$$\|x\|_* := \sum_{j=0}^{n-1} (\rho(A)+\varepsilon)^{n-1-j}\|A^j x\| \quad (\forall x \in X)$$

Wegen

$$(\rho(A)+\varepsilon)^{n-1}\|x\| \le \|x\|_* \le \Big(\sum_{j=0}^{n-1}(\rho(A)+\varepsilon)^{n-1-j}\|A^j\|\Big)\|x\| \quad (\forall x \in X)$$

gilt die Äquivalenz der Normen $\|\cdot\|$ bzw. $\|\cdot\|_*$ und aus der Definition folgt auch (*).

3.2 Konjugierte Gradienten-Methoden

Konjugierte Gradienten-Methoden (kurz: CG-Methoden) dienen zur iterativen Lösung von *großen* linearen Gleichungssystemen

$$Ax = b \quad (A \in \mathbb{R}^{m\times m}, b \in \mathbb{R}^m)$$

mit *symmetrischer* und *positiv definiter* Koeffizientenmatrix A. Es wird sich zeigen, dass sie in höchstens m Schritten mit der exakten Lösung enden. Prinzipiell handelt es sich hierbei um ein direktes Verfahren, tatsächlich liefert es oft schon nach sehr viel weniger Schritten eine brauchbare Näherung, was bei großem m wesentlich ist. Deshalb wird die CG-Methode üblicherweise als iteratives Verfahren angesehen. Es ist i.a. deutlich effizienter als das Einzelschritt-Verfahren und ist heute die wohl wichtigste Methode zur Lösung großer linearer Gleichungssysteme mit symmetrischer, positiv definiter Matrix.

Lemma 3.11
Seien $A \in \mathbb{R}^{m\times m}$ symmetrisch und positiv definit, und $b \in \mathbb{R}^m$. Die Funktion

$$\Phi(x) := \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$$

nimmt ihr einziges Minimum auf $\mathbb{R}^m$ in $x^* = A^{-1}b$ an.
(Dabei ist $\langle\cdot,\cdot\rangle$ das Euklidische Skalarprodukt und $\|\cdot\|$ die Euklidische Norm.)

Beweis: Für beliebiges x und h in $\mathbb{R}^m$ zeigt man

$$\begin{aligned}\Phi(x+th) &= \frac{1}{2}\langle A(x+th), x+th\rangle - \langle b, x+th\rangle \\ &= \frac{1}{2}[\langle Ax, x\rangle + t(\langle Ax, h\rangle + \langle x, Ah\rangle] + t^2\|h\|^2] - \langle b, x\rangle - t\langle b, h\rangle \\ &= \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle + t\langle Ax - b, h\rangle + t^2\|h\|^2\end{aligned}$$

Die Funktion Φ ist eine quadratische Funktion, die natürlich zweimal stetig differenzierbar. Aus den obigen Überlegungen folgt für die Richtungsableitung von Φ in x in Richtung h:

$$\frac{\mathrm{d}}{\mathrm{dt}}\,\Phi(x+th)\big|_{t=0} = \langle \Phi'(x), h\rangle = \langle Ax - b, h\rangle$$

bzw. für den Gradienten $\Phi'(x) = Ax - b$ und die Hesse-Matrix $\Phi''(x) = A$. Da A symmetrisch und positiv definit ist, ist die einzige Lösung $x^* = A^{-1}b$ von $\Phi'(x) = Ax - b = 0$

das einzige Minimalelement von Φ auf $\mathbb{R}^m$. $\square$

Idee: Löse anstelle des linearen Gleichungssystems $Ax = b$ das Minimumproblem

$$\min_{x\in\mathbb{R}^m} \Phi(x)\,.$$

Lemma 3.12 *Seien $A \in \mathbb{R}^{m\times m}$ symmetrisch und positiv definit, $b \in \mathbb{R}^m$ und es sei $\Phi(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$, $\forall x \in \mathbb{R}^m$. Es seien $0 \neq d^j \in \mathbb{R}^m$ gewählt mit*

$$(*) \qquad \langle Ad^i, d^j\rangle = 0 \quad (i \neq j, \forall i, j = 0, \ldots, m-1).$$

Dann liefert das iterative Verfahren

$$x_{k+1} = x_k + t_k d^k\,, \quad k = 0, 1, \ldots, m-1,$$

mit t_k aus

$$\Phi(x_k + t_k d^k) = \min_{t\in\mathbb{R}} \Phi(x_k + td^k)$$

nach (höchstens) m Schritten mit x_m das Minimum x^* von Φ.
Weiterhin gelten für $k = 0, 1, \ldots, m-1$ mit den Gradienten $g_k = Ax_k - b$ die Beziehungen

$$t_k = -\frac{\langle g_k, d^k\rangle}{\langle Ad^k, d^k\rangle} \quad \textit{und} \quad \langle g_{k+1}, d^j\rangle = 0\,, \quad j = 0, 1, \ldots, k.$$

Beweis: Es sei $\varphi(t) := \Phi(x_k + td^k)$ $(t \in \mathbb{R})$. Wie im Beweis von Lemma 3.9 folgt als notwendige und hinreichende Bedingung an t_k

$$0 = \varphi'(t_k) = \langle \Phi'(x_k + t_k d^k), d^k\rangle = \langle A(x_k + t_k d^k) - b, d^k\rangle = \langle g_k, d^k\rangle + t_k\langle Ad^k, d^k\rangle$$

und damit $t_k = -\frac{\langle g_k, d^k\rangle}{\langle Ad^k, d^k\rangle}$. Wegen der positiven Definitheit von A ist der Nenner stets positiv. Ferner ergibt sich aus der Formel $x_{k+1} = x_k + t_k d^k$ für das iterative Verfahren

$$\langle g_{k+1}, d^k\rangle = \langle Ax_{k+1} - b, d^k\rangle = \langle Ax_k - b + t_k Ad^k, d^k\rangle = \langle g_k, d^k\rangle + t_k\langle Ad^k, d^k\rangle = 0$$

für $k = 0, \ldots, m-1$. Unter Verwendung der an die d^j gestellten Bedingungen (*) erhält man ferner für $i \neq j$

$$\langle g_{i+1} - g_i, d^j\rangle = \langle Ax^{i+1} - Ax^i, d^j\rangle = t_i\langle Ad^j, d^i\rangle = 0\,.$$

Deshalb folgt für $j = 0, \ldots, k$

$$\langle g_{k+1}, d^j\rangle = \langle g_{j+1}, d^j\rangle + \sum_{i=j+1}^{k} \langle g_{i+1} - g_i, d^j\rangle = 0\,.$$

Da die $d^0, \ldots, d^{m-1}$ wegen (*) paarweise orthogonal bzgl. des Skalarprodukts

$$\langle u, v\rangle_A := \langle Au, v\rangle$$

sind, sind sie linear unabhängig. Deshalb ist g_m orthogonal zu $d^0, \ldots, d^{m-1}$ und es muss $g_m = 0$ und damit $x_m = x^* = A^{-1}b$ nach Lemma 3.9 gelten. □

Vektoren $d^0, \ldots, d^{m-1}$ mit der Eigenschaft (*) heißen auch A-konjugiert.
Um sich solche Vektoren zu verschaffen, kann man das Gram-Schmidtsche Orthogonalisierungsverfahren bzgl. des Skalarproduktes $\langle \cdot, \cdot \rangle_A$ auf eine beliebige Basis des $\mathbb{R}^m$ anwenden. Da evtl. gar nicht alle d^j benötigt werden, sollen $d^0, d^1, \ldots$ im Laufe des Verfahrens sukzessive erzeugt werden. Dies wird so eingerichtet, dass der neu berechnete Vektor d^k eine *Abstiegsrichtung* für Φ in x_k ist. Da nach dem Satz von Taylor gilt

$$\Phi(x_{k+1}) = \Phi(x_k) + t_k \langle \Phi'(x_k), d^k \rangle + \frac{t_k^2}{2} \langle \Phi''(x_k) d^k, d^k \rangle = \Phi(x_k) + t_k \langle g_k, d^k \rangle + \frac{t_k^2}{2} \langle A d^k, d^k \rangle,$$

ist d^k eine Abstiegsrichtung (d.h. $\Phi(x_{k+1}) \leq \Phi(x_k)$), falls

$$\langle \Phi'(x_k), d^k \rangle = \langle g_k, d^k \rangle < 0 \quad \text{und} \quad t_k > 0$$

gilt. Man startet deshalb mit

$$d^0 := -\Phi'(x_0) = -g_0 = b - Ax_0 \,.$$

Es seien jetzt bereits $d^0, \ldots, d^{k-1}$ für ein $k \in \{0, 1, \ldots, m-1\}$ konstruiert, so dass

$$\langle A d^i, d^j \rangle = 0, \quad \forall i, j \in \{0, \ldots, k-1\},\ i \neq j.$$

Wegen $\langle g_k, d^i \rangle = 0$, $\forall i = 0, 1, \ldots, k-1$, nach Lemma 3.10 und im Fall $g^k \neq 0$ ist es sinnvoll, den Ansatz

$$d^k = -g_k + \sum_{i=0}^{k-1} \beta_i^k d^i$$

zu machen. Dann gilt $\langle d^k, A d^j \rangle = 0$, $j = 0, \ldots, k-1$, gdw.

$$0 \;=\; -\langle g_k, A d^j \rangle + \sum_{i=0}^{k-1} \beta_i^k \langle d^i, A d^j \rangle = -\langle g_k, A d^j \rangle + \beta_j^k \langle d^j, A d^j \rangle \quad \text{d.h.}$$

$$\beta_j^k \;=\; \frac{\langle g_k, A d^j \rangle}{\langle d^j, A d^j \rangle}$$

für jedes $j = 0, 1, \ldots, k-1$. Wir berechnen also die Abstiegsrichtung d^k durch

$$(**) \qquad d^k = -g_k + \sum_{i=0}^{k-1} \frac{\langle g_k, A d^i \rangle}{\langle d^i, A d^i \rangle} d^i \,.$$

Lemma 3.13
Berechnet man die Abstiegsrichtungen d^k, $k \in \{0, \ldots, m-1\}$ aus (**), so gilt

$$\langle g_k, d^k \rangle = -\|g_k\|^2 < 0\,, \quad d^k = -g_k + \frac{\|g_k\|^2}{\|g_{k-1}\|^2} d^{k-1} \quad \textit{und} \quad t_k = \frac{\|g_k\|^2}{\langle A d^k, d^k \rangle} > 0$$

für jedes $k \in \{1, \ldots, m-1\}$.

Beweis: Multipliziert man die Gleichung (**) mit g_k, so gilt nach Lemma 3.10

$$\langle g_k, d^k\rangle = -\|g_k\|^2 + \sum_{i=0}^{k-1} \frac{\langle g_k, Ad^i\rangle}{\langle d^i, Ad^i\rangle}\langle g_k, d^i\rangle = -\|g_k\|^2 .$$

Überdies folgt aus (**) für j (anstelle von k) ebenfalls nach Lemma 3.10

$$\langle g_k, g_j\rangle = -\langle g_k, d^j\rangle + \sum_{i=0}^{j-1} \frac{\langle g_j, Ad^i\rangle}{\langle d^i, Ad^i\rangle}\langle g_k, d^i\rangle = 0 \quad (j = 0, 1, \ldots, k-1).$$

Wegen $g_{j+1} - g_j = Ax_{j+1} - Ax_j = t_j Ad^j$ für $j = 0, 1, \ldots, k-1$ gilt

$$\langle g_k, Ad^j\rangle = \frac{1}{t_j}\langle g_k, g^{j+1} - g^j\rangle = 0 \quad (j = 0, 1, \ldots, k-2)$$

und damit $\beta_j^k = 0$, $j = 0, 1, \ldots, k-2$, sowie nach Lemma 3.10

$$\beta_{k-1}^k = \frac{1}{t_{k-1}} \frac{\langle g_k, g_k\rangle}{\langle Ad^{k-1}, d^{k-1}\rangle} = \frac{\|g_k\|^2}{\|g_{k-1}\|^2} =: \beta_k .$$

Also gilt: $d^k = -g_k + \beta_k d^{k-1}$. □

Algorithmus 3.14 *(CG-Verfahren)*
Wähle $x_0 \in \mathbb{R}^m$ und $\varepsilon \in (0, 1)$, setze $k = 0$, $g_0 = Ax_0 - b$ und $d^0 = -g_0$.
Schritt k: Falls $\|b - Ax_k\| \le \varepsilon\|b - Ax^0\|$, so STOP.

$$\begin{aligned} t_k &:= \frac{\|g_k\|^2}{\langle Ad^k, d^k\rangle} \qquad x_{k+1} := x_k + t_k d^k \\ g_{k+1} &:= g_k + t_k Ad^k \qquad \beta_k := \frac{\|g_{k+1}\|^2}{\|g_k\|^2} \\ d^{k+1} &:= -g_{k+1} + \beta_k d^k. \end{aligned}$$

(Die Vorschrift zur Berechnung von g_{k+1} erfordert gegenüber $g_{k+1} = Ax_{k+1} - b$ eine deutlich geringere Anzahl von Operationen, da Ad^k und t_k sowieso berechnet werden.)

Satz 3.15 *(Konvergenz)*
Der Algorithmus 3.12 liefert nach höchstens m Schritten die eindeutige Lösung $x^* \in \mathbb{R}^m$ von $Ax = b$ mit symmetrischer und positiv definiter Matrix $A \in \mathbb{R}^{m\times m}$ und $b \in \mathbb{R}^m$. Ist $n \le m$ die kleinste natürliche Zahl mit $x_n = x^*$, so gelten die Eigenschaften

$$\begin{aligned} \langle Ad^j, d^k\rangle &= 0 \qquad \langle d^j, g_k\rangle = 0 \\ \langle g_j, g_k\rangle &= 0 \qquad \langle g_k, d^k\rangle = -\|g_k\|^2 \end{aligned}$$

für $j = 0, 1, \ldots, k-1$, $k = 1, \ldots, n$.

Beweis: Ist $g_n = 0$ für ein $n < m$, so gilt $Ax_n = b$ und damit $x_n = x^*$. Anderenfalls sind die von 0 verschiedenen Richtungen $d^0, d^1, \ldots, d^{m-1}$ A-konjugiert und $x_m = x^*$ folgt aus Lemma 3.10. Die genannten Eigenschaften folgen aus 3.10 und 3.11. □

Bemerkung 3.16 *(Charakterisierung von CG-Verfahren)*
Mit den Residuen $r_i := b - Ax_i$, $i = 0, 1, \ldots, k$, *gilt*

$$\operatorname{span}\{r_0, r_1, \ldots, r_{k-1}\} = \operatorname{span}\{r_0, Ar_0, \ldots, A^{k-1}r_0\}$$

(Beweis mittels Induktion über k*). Der letztere Teilraum des* $\mathbb{R}^m$ *heißt auch der vom Anfangsresiduum* r_0 *und der Matrix* A *aufgespannte* Krylov-Raum $\mathcal{K}_k(r_0, A)$, *d.h.*

$$\mathcal{K}_k(r_0, A) = \operatorname{span}\{r_0, Ar_0, \ldots, A^{k-1}r_0\}\,.$$

Es gilt: $\mathcal{K}_k(r_0, A) = \operatorname{span}\{d^0, d^1, \ldots, d^k\}$ *(nach Konstruktion).*
Für die k*-te Iterierte* x_k *des CG-Verfahrens in Algorithmus 3.12 gilt:*

(a) x_k *minimiert die Funktion* Φ *in Lemma 3.9 auf* $x_0 + \mathcal{K}_k(r_0, A)$.

(b) x_k *minimiert das Residuum* $\|b - Ax\|^2_{A^{-\frac{1}{2}}}$ *auf* $x_0 + \mathcal{K}_k(r_0, A)$.

Dabei ist $A^{\frac{1}{2}}$ *die eindeutig bestimmte symmetrische und positiv definite Matrix* $B \in \mathbb{R}^{m\times m}$ *mit* $B^2 = A$.
Wegen (b) ist das CG-Verfahren ein spezielles Krylov-Raum-Verfahren *zur Lösung von* $Ax = b$. *Solche Krylov-Raum-Verfahren, bei denen schrittweise* $\|b - Ax\|^2$ *für eine gewisse Norm* $\|\cdot\|$ *auf* $\mathbb{R}^m$ *über dem Krylov-Raum* $x_0 + \mathcal{K}_k(r_0, A)$ *minimiert wird, existieren und konvergieren auch für allgemeine reguläre Matrizen (vgl. Kap. 6 in Kanzow).*

Satz 3.17 *(Konvergenzgeschwindigkeit)*
Seien $A \in \mathbb{R}^{m\times m}$ *symmetrisch und positiv definit,* $b \in \mathbb{R}^m$, $x^* \in \mathbb{R}^m$ *die eindeutig bestimmte Lösung des linearen Gleichungssystems* $Ax = b$ *und* (x_n) *die durch Algorithmus 3.12 erzeugte Folge. Dann gilt*

$$\|x_n - x^*\|_{A^{\frac{1}{2}}} \le 2\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^n \|x_0 - x^*\|_{A^{\frac{1}{2}}},$$

wobei $\kappa = \operatorname{cond}_2(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$.

Beweis: vgl. Satz 5.11 in C. Kanzow: Numerik linearer Gleichungssysteme, 2005.

Bemerkung 3.18 *(Präkonditionierer für CG-Verfahren)*
Satz 3.15 legt nahe, dass das CG-Verfahren vermutlich schneller konvergiert, je kleiner die Konditionszahl $\operatorname{cond}_2(A)$ *ist. Aus diesem Grunde ist es naheliegend, die symmetrische, positiv definite Matrix* A *und damit das lineare Gleichungssystem* $Ax = b$ *zu transformieren, indem man mit einer regulären Matrix* $C \in \mathbb{R}^m$ *das folgende transformierte lineare Gleichungssystem*

$$\begin{aligned} C^{-1}A(C^{-1})^\top C^\top x &= C^{-1}b \qquad \textit{oder} \\ \bar{A}\bar{x} &= \bar{b} \quad \textit{mit} \quad \bar{A} = C^{-1}A(C^{-1})^\top,\ \bar{x} = C^\top x,\ \bar{b} = C^{-1}b \end{aligned}$$

betrachtet. Die Matrix $\bar{A}$ *ist in jedem Fall wieder symmetrisch und positiv definit.*
Man kann dann zeigen, dass sich Algorithmus 3.12 nur dadurch verändert, dass zur Bestimmung von d^k *in jedem Iterationsschritt das lineare Gleichungssystem*

$$Pd = r^k = b - Ax_k \quad \textit{mit} \quad P = CC^\top$$

gelöst werden muss. Dieses sollte natürlich relativ leicht lösbar sein und überdies der sog. Präkonditionierer P *eine gute Approximation für* A *sein. Möglichkeiten dafür sind*
(i) $P = D = \operatorname{diag}(a_{11}, \ldots, a_{mm})$ *aus dem Gesamtschrittverfahren,*
(ii) die Durchführung einiger Schritte des Gesamtschrittverfahrens, oder
(iii) die sog. unvollständige Cholesky-Zerlegung von A*.*
(vgl. Kap. 5.5 und 5.6 aus Kanzow 2005.)