Wissenschaftliches Rechnen II

Vorlesungsskripte
Sommer-Semester 2009

Werner Römisch

Humboldt-Universität Berlin
Institut für Mathematik

Inhalt: Seite

0 Einleitung

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

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

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

(i) Lineare Gleichungssysteme: $Ax = b$, wobei $A = (a_{ij})_{i,j=1,\dots,m} \in \mathbb{R}^{m\times m}$ invertierbar ist und $b \in \mathbb{R}^m$.

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

Literatur:

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

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

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

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

1 Numerische Lösung linearer Gleichungssysteme

Wir werden hier fast ausschließlich lineare Gleichungssysteme (GS) 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). Wir beginnen mit einer Fehleranalyse und der Kondition linearer Gleichungssysteme.

1.1 Kondition linearer Gleichungssysteme

Wir betrachten das lineare Gleichungssystem

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

als gegeben und erinnern uns zunächst an einige Fakten aus der linearen Algebra.
Wir bezeichnen mit $\mathrm{rg}(A)$ den Rang der Matrix A, d.h., die Dimension des Wertebereichs $R(A) := \{Ax : x \in \mathbb{R}^m\}$ bzw. die maximale Anzahl linear unabhängiger Zeilen bzw. Spalten. Eine andere Formulierung dieser Definition von $\mathrm{rg}(A)$ ist, daß das homogene lineare Gleichungssystem $Ax = 0$ gerade $m - \mathrm{rg}(A)$ linear unabhängige Lösungen besitzt. Man nennt A eine invertierbare Matrix, falls $\mathrm{rg}(A) = m$. Ist A invertierbar, so ist das lineare Gleichungssystem $Ax = b$ für jede rechte Seite b eindeutig lösbar und es existiert eine Matrix $A^{-1} \in \mathbb{R}^{m\times m}$, so daß $A^{-1}b$ diese eindeutig bestimmte Lösung darstellt. A^{-1} ist die zu A inverse Matrix.
Zusätzlich betrachten wir das „gestörte“ lineare Gleichungssystem

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

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

Lemma 1.1 *Es sei $B \in \mathbb{R}^{m\times m}$ und für eine zugeordnete Matrixnorm gelte $\|B\| < 1$. Dann ist die Matrix $E + B$ invertierbar und es gilt:*

$$\frac{1}{1 + \|B\|} \le \|(E + B)^{-1}\| \le \frac{1}{1 - \|B\|}.$$

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

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

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

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

$$E = (E+B)C = C(E+B)$$

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

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

Damit ist die behauptete Ungleichungskette bewiesen. □

Satz 1.2 *(„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 $E + A^{-1}(\tilde{A} - A) = A^{-1}\tilde{A}$ invertierbar, also auch $\tilde{A}$. Außerdem folgt aus Lemma 1.1:

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

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

□

Folgerung 1.3 *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 1.2. □

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

Satz 1.4 *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

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

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

Beweis:
Die Voraussetzung $\|A^{-1}\|\,\|\Delta A\| < 1$ impliziert nach Satz 1.2, daß die Matrix $A + \Delta A$ invertierbar ist und daß

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

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

$$(A + \Delta A)\Delta x = b + \Delta b - Ax - \Delta Ax = \Delta b - \Delta Ax$$

und deshalb

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

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

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

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

Definition 1.5 *Die Zahl $cond\,(A) := \|A\|\,\|A^{-1}\|$ heißt Konditionszahl der invertierbaren Matrix $A \in \mathbb{R}^{m\times m}$ (bzgl. der Matrixnorm $\|\cdot\|$).*

Bemerkung 1.6 *Die Abschätzung für den relativen Fehler $\frac{\|\Delta x\|}{\|x\|}$ in Satz 11.4 kann i. a. nicht verbessert werden. Sie besagt, daß der relative Fehler der Lösungen abgeschätzt werden kann durch das Produkt eines Faktors mit der Summe der relativen Fehler der Eingangsdaten. Für eine „kleine Störung“ 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* $\mathrm{cond}(A) \ge 1$ *wegen $\|A^{-1}\|\,\|A\| \ge \|E\| = 1$. Bei Verwendung von $\|\cdot\|_p$ als Matrixnorm verwenden wir die Bezeichnung* $\mathrm{cond}_p(A)$.

Beispiel 1.7 *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 $\mathrm{cond}_2(H_m) = \|H_m^{-1}\|_2 \|H_m\|_2 = \frac{\lambda_{\max}(H_m)}{\lambda_{\min}(H_m)}$.

1.2 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*:

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

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 1.8

a) $P_{k,s(k)}$ *ist symmetrisch und invertierbar mit* $P_{k,s(k)}^2 = E$, $\det(P_{k,s(k)}) = -1$, *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:

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

Satz 1.9 *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

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

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 $P_{k+1,s(k+1)}(0,\ldots,0,1,-\ell_{k+1,k},\ldots,-\ell_{mk})^\top$ (vgl. 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 1.8b):

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

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 1.10 *Man nennt die Darstellung von PA in der Form $PA = LR$ wie in Satz 1.9 die LR-Faktorisierung von A.*
Der Gaußsche Algorithmus hat in Matrixschreibweise nun folgende Form:
- $PAx = Pb$ (Vertauschung von Zeilen)
- $(R, c) = L^{-1}(PA, Pb)$ (Dreieckszerlegung) und anschließend
- $x = R^{-1}c$ (Lösung eines linearen Gleichungssystems mit Dreiecksmatrix)
Ist umgekehrt eine LR-Faktorisierung von A gegeben, so löst man das Gleichungssystem $Ax = b$ in den folgenden beiden Schritten:
- Berechne c als Lösung von $Lc = Pb$ („Vorwärtselimination“),
- berechne x als Lösung von $Rx = c$ („Rückwärtselimination“).

Bemerkung 1.11 *(Anzahl von Operationen)*
*Wir berechnen die Anzahl der Gleitkommaoperationen für den Gaußschen Algorithmus bzw. für die Lösung eines linearen Gleichungssystems. Dabei ist es üblich, mit „*opms*“ eine Gleitkommarechenoperation, bestehend aus einer Multiplikation und einer Addition/Subtraktion zugrunde zu legen. Dann erhält man für den Rechenaufwand zur*

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

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 1.12 *(Pivotisierung)*
Nicht eindeutig bestimmt ist bisher die Wahl der Permutationsmatrix P, d. h. die Wahl der Zeilenvertauschungen zur Bestimmung des Elementes $a^{(k)}_{s(k),k} \neq 0$ in der k-ten Spalte unterhalb 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

$$\begin{aligned} \mathrm{cond}_1(L_k) &= \|L_k^{-1}\|_1 \|L_k\|_1 = \|E + \sum_{i=k+1}^{m} \tfrac{a^{(k)}_{ik}}{a^{(k)}_{kk}} e^i (e^k)^\top\|_1 \|E - \sum_{i=k+1}^{m} \tfrac{a^{(k)}_{ik}}{a^{(k)}_{kk}} e^i (e^k)^\top\|_1 \\ &= \left(1 + \sum_{i=k+1}^{m} \left|\tfrac{a^{(k)}_{ik}}{a^{(k)}_{kk}}\right|\right)^2 \qquad \textit{(vgl. 1.8b))}. \end{aligned}$$

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

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

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 1.13 *(numerische Berechnung von Determinanten)*
Es sei $A \in \mathbb{R}^{m\times m}$ eine invertierbare Matrix mit gegebener LR-Faktorisierung gemäß Satz 1.9. 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 1.9 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 1.8 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 1.14 *(Cholesky-Faktorisierung)*
Für spezielle Matrizen nimmt der Gaußsche Algorithmus spezielle Formen an. Ist 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 1.9 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. 1.11 (etwa) halbiert.
(Literatur: Hämmerlin/Hoffmann, Kap. 2.2; Kielbasinski/Schwetlick, Kap. 6).

Wir kommen nun zur Rundungsfehleranalyse des Gaußschen Algorithmus zur Lösung eines linearen Gleichungssystems. Die ersten Resultate (Satz 1.15 und Folg. 1.17) betreffen dabei die numerische Gutartigkeit der *LR*-Faktorisierung, das letzte (Satz 1.18) die Rückwärtselimination. Vorher sei an die Fehlerfortpflanzung bei Gleitkomma-Operationen erinnert (vgl. WR I):

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

wobei $|\varepsilon|, |\eta| \le \tau$ gilt und $\tau := 0.5\beta^{-t+1}$ die relative Rechengenauigkeit bei t-stelliger Arithmetik mit Basis $\beta \in \mathbb{N}$, $\beta \ge 2$ ist.

Satz 1.15 *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 $A^{(k)} = \left(\begin{array}{ccc} \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots \\ \hline 0 & \vert & M^{(k)} \\ & k-1 & \end{array}\right)$.

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

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

Hierbei ist L_k die entsprechende Transformationsmatrix (vgl. den Beweis von Satz 1.9), $R^{(k)}$ der "bereits fertige" Teil von R (mit $k-1$ Zeilen), $E_{k-1} \in \mathbb{R}^{(k-1)\times(k-1)}$ die Einheitsmatrix und $M^{(k)}$ die "Restmatrix" von $A^{(k)}$.
Wir untersuchen zunächst die numerische Gutartigkeit der Transformation

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

in t-stelliger Arithmetik. Die diese Transformation beschreibenden Operationen lassen sich dabei 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 1.16 *Satz 1.15 besagt, daß in jeder Matrizenklasse $\mathcal{A} \subset \mathbb{R}^{m\times m}$, in der der Gaußsche Algorithmus durchführbar ist und*

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

für eine der Normen $\|\cdot\|_p$ gilt, die LR-Faktorisierung numerisch gutartig ist. Klassen von solchen Matrizen werden in Folg. 1.17 und Kielbasinski/Schwetlick, Satz 5.3.2 angegeben.
Das folgende Beispiel zeigt aber, daß Satz 1.15 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 1.16 tritt nicht auf, wenn eine Spaltenpivotisierung durchgeführt wird !

Folgerung 1.17 *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 1.12) gilt

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

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

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

(vgl. auch Kielbasinski/Schwetlick, Aussage 5.2.1).
Aus Satz 1.15 folgt 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 1.17 ist in vielen Anwendungsfällen zu pessimistisch. Praktisch geht man von einem linearen Wachstum von $F_\infty(A)$ mit m aus.

Für die Rückwärtselimination ist die Situation einfacher.

Satz 1.18 *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 1.19 *(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 $D = \mathrm{diag}(d_1, \ldots, d_m)$ 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 $D = \mathrm{diag}(d_1, \ldots, d_m)$, 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:

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

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

<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 cond_∞ 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 cond_∞.

Bemerkung 1.20 *(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 1.10), 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 1.21 *(Konditionsschätzung)*
Das Ziel bestehe darin, $\mathrm{cond}_\infty(A)$ *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. 1.11).*
Es sei jetzt eine LR-Faktorisierung von A gegeben und das Ziel sei, $\|A^{-1}\|_\infty$ näherungsweise zu berechnen. Es gilt:

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

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

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

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

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

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

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

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

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).

1.3 Householder-Orthogonalisierung

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

Definition 1.22
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 $H := E - 2uu^\top$
Householder-Spiegelung(smatrix).

Lemma 1.23

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: $\|QA\|_2 \leq \|Q\|_2\|A\|_2 = \|A\|_2$.

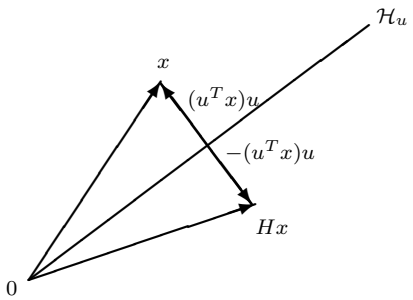
b) Eine Householder-Spiegelung ist nach Definition symmetrisch, d. h. es gilt $H = H^\top$. Ferner gilt:

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

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

Bemerkung 1.24 *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 $H = E - 2uu^\top$ 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. 1.2, so sind Householder-Spiegelungen orthogonal und die $L_{ij}(\beta)$ nicht!



Lemma 1.25 *Es sei $e^1 := (1, 0, \ldots, 0)^\top$, $x \in \mathbb{R}^m$ mit $x \neq \alpha e^1$, $\forall \alpha \in \mathbb{R}$.*
Dann gilt $Hx = (E - 2uu^\top)x = \sigma e^1$ 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 1.23 folgt $\|Hx\|_2 = \|x\|_2 = |\sigma|$, d.h. $\sigma = \pm\|x\|_2$.
Haben umgekehrt u und σ die angegebene Form, so gilt

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

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:

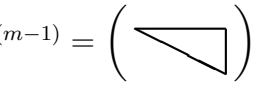
Algorithmus 1.26 *(QR-Faktorisierung durch Householder-Orthogonalisierung)*
1. Schritt:
a_1 sei die erste Spalte von A; hat a_1 die Form αe^1, so ist der erste Schritt beendet; ansonsten bestimme $u_1 \in \mathbb{R}^m$ so, daß $H_1 a_1 = (E - 2u_1 u_1^\top)a_1 = \|a_1\|_2 e^1$ und $\|u_1\|_2 = 1$. Setze $A^{(1)} = H_1 A$.
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*

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

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 $R = A^{(m-1)} =$ 

Satz 1.27 *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 1.26 definierten Householder-Spiegelungen und wir definieren die folgenden Matrizen in $\mathbb{R}^{m\times m}$:

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

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 1.23. □

Bemerkung 1.28 *Anzahl der Rechenoperationen einer QR-Faktorisierung:*

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

Die Anzahl der Rechenoperationen ist also etwa doppelt so groß wie beim Gaußschen Algorithmus. Die Householder-Orthogonalisierung besonders dann empfehlenswert, wenn $cond_2(A)$ *„groß" ist. Die Householder-Orthogonalisierung ist ein numerisch gutartiges Verfahren (vgl. Kielbasinski/Schwetlick, Kap. 10.2).*

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

Beispiel 1.29 *(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 $\frac{1}{2}\|Ax-b\|_2^2 = \min\limits_{y\in\mathbb{R}^m} \frac{1}{2}\|Ay-b\|_2^2$.

Satz 1.30 *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*

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

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: $\Phi(x) = \frac{1}{2}\langle Ax-b, Ax-b\rangle = \frac{1}{2}\left[\left\langle A^\top Ax, x\right\rangle - 2\left\langle A^\top b, x\right\rangle + \langle b, b\rangle\right]$
$\rightsquigarrow \Phi'(x) = A^\top Ax - A^\top b$ (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 1.31 *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 1.32 *(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

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

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

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

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:*

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

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 <u>das</u> 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 1.33 *(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*

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

<u>Beweis</u>:
Analog zu Satz 1.27 existieren orthogonale Matrizen Q_k, $k = 1, \ldots, m$, so daß mit $Q = Q_m Q_{m-1} \cdots Q_1$ die Matrix QA die Gestalt

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

mit einer invertierbaren oberen Dreiecksmatrix R besitzt. Dann gilt:

$$\begin{aligned} \Phi(x) &= \tfrac{1}{2}\|Ax-b\|_2^2 = \tfrac{1}{2}\|Q(Ax-b)\|_2^2 \qquad \text{(Lemma 1.23 !)} \\ &= \tfrac{1}{2}\left\| \begin{pmatrix} R \\ \hline 0 \end{pmatrix} x - Qb \right\|_2^2 = \tfrac{1}{2}\|Rx - r_1\|_2^2 + \tfrac{1}{2}\|r_2\|_2^2. \end{aligned}$$

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

1.4 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 1.34 *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. Steht nun ein Computer mit* 10^{7+x} *Operationen pro Sekunde zu Verfügung, so führt dieser etwa* $0.315 \cdot 10^{15+x}$ *Operationen im Jahr aus. Es ergeben sich folgende Rechenzeiten bei variierendem* m*:*

m	10^3	10^5	10^6	10^8
Rechenzeit (sec)	$\frac{1}{3}10^{2-x}$	$\frac{1}{3}10^{8-x}$	$\frac{1}{3}10^{11-x}$	$\frac{1}{3}10^{17-x}$

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

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

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

Ausgehend von der Fixpunktgleichung wird wie im Banachschen Fixpunktsatz das Iterationsverfahren

$$x_{n+1} = (E - C^{-1}A)x_n + C^{-1}b\,,\ n = 0, 1, 2, \ldots,\ x_0 \in \mathbb{R}^m,$$

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

Problemstellung: Wann konvergieren Iterationsverfahren vom allgemeinen Typ

$$\text{(IV)} \quad x_{n+1} = Bx_n + d\,,\ n = 0, 1, 2, \ldots,\ x_0 \in \mathbb{R}^m,$$

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 1.35 *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

$$J = T^{-1}BT = \begin{pmatrix} \lambda_1 & * & 0 & \cdots & 0 & 0 \\ 0 & \lambda_2 & * & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_{m-1} & * \\ 0 & 0 & 0 & \cdots & 0 & \lambda_m \end{pmatrix}$$

wobei λ_i, $i = 1, \ldots, m$, 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

$$J_\varepsilon = \begin{pmatrix} \lambda_1 & *_\varepsilon & 0 & \cdots & 0 & 0 \\ 0 & \lambda_2 & *_\varepsilon & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_{m-1} & *_\varepsilon \\ 0 & 0 & 0 & \cdots & 0 & \lambda_m \end{pmatrix}$$

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

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

□

Satz 1.36 *(Konvergenz von (IV))*
Sei $B \in \mathbb{R}^{m\times m}$. Dann ist das Iterationsverfahren

$$(IV) \qquad x_{n+1} := Bx_n + d\,,\ n = 0, 1, 2, \ldots,$$

für jedes $x_0 \in \mathbb{R}^m$ und jedes $d \in \mathbb{R}^m$ konvergent gegen $(E-B)^{-1}d$ gdw. $\rho(B) < 1$.

Beweis:

($\Leftarrow$) Sei λ betragsgrößter Eigenwert von B, d.h. $|\lambda| = \rho(B)$ und z 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.

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

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

Gemäß Lemma 1.35 wählen wir eine Norm $\|\cdot\|_*$, so daß $\|B\|_* < 1$. Dann gilt $\|B^n x_0\|_* \le \|B^n\|_*\|x_0\|_* \le \|B\|_*^n\|x_0\|_*$ und

$$\left\| \sum_{j=0}^{n-1} B^j - (E-B)^{-1} \right\|_* = \left\|(I-B)^{-1}B^n\right\|_* \le \left\|(I-B)^{-1}\right\|_* \|B\|_*^n\,.$$

Da $(\|B\|_*^n)$ eine Nullfolge ist, konvergiert die Folge $(\sum_{j=0}^{n-1} B^j d)$ gegen $(E-B)^{-1}d$ und damit auch (x_n) gegen $(E-B)^{-1}d$ in $\mathbb{R}^m$. □

Beispiel 1.37

a) *Gesamtschrittverfahren: Vor.: $a_{ii} \neq 0$, $i = 1, \ldots, m$.*
Wir wählen $C := D := diag(a_{11}, \ldots, a_{mm})$ und erhalten

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

Das Gesamtschrittverfahren nimmt dann die Form an

$$x_{n+1} := (E - D^{-1}A)x_n + D^{-1}b\,,\ n = 0, 1, 2, \ldots\,,\ \ \textit{mit } x_0 \in \mathbb{R}^m \textit{ beliebig.}$$

In Komponenten-Schreibweise hat es die Form

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

Frage: Wann gilt $\rho(E - D^{-1}A) = \rho(-D^{-1}(L+R)) < 1$ *?*

b) *Einzelschrittverfahren bzw. Gauß-Seidel Verfahren: Vor.:* $a_{ii} \neq 0$, $i = 1, \ldots, m$. *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* $B := E - C^{-1}A = -(L+D)^{-1}R$.
Dann nimmt das Einzelschrittverfahren die Form an

$$x_{n+1} := -(L+D)^{-1}Rx_n + (L+D)^{-1}b \quad oder \quad (L+D)x_{n+1} := -Rx_n + b$$

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

$$x_{n+1}^i := \frac{1}{a_{ii}} \left(-\sum_{j=1}^{i-1} a_{ij} x_{n+1}^j - \sum_{j=i+1}^{m} a_{ij} x_n^j + b_i \right) \text{ , } i = 1, \ldots, m \text{ , } n = 0, 1, 2, \ldots \text{ ,}$$

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

c) *Relaxationsverfahren: Vor.:* $a_{ii} \neq 0$, $i = 1, \ldots, m$. *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

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

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

Satz 1.38
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.*

Beweis: Wir definieren $\gamma := \max\limits_{i=1,\ldots,m} \frac{1}{|a_{ii}|} \sum\limits_{\substack{j=1 \\ j\neq i}} |a_{ij}|$ und wissen, dass nach Voraussetzung $\gamma < 1$ gilt. Ferner gilt $\| - D^{-1}(L+R)\|_\infty = \gamma < 1$. Daraus folgt nach Lemma 1.1, dass $E + D^{-1}(L+R) = D^{-1}A$ 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 1.35 folgt deshalb $\rho(B) \le \|B\|_\infty < 1$ und aus Satz 1.36 die Konvergenz des Gesamtschrittverfahrens gegen $(E-B)^{-1}d = (D^{-1}A)^{-1}D^{-1}b = A^{-1}b$.

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

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

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

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

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

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

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

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

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

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

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

Satz 1.39 *Für jede Matrix $A \in \mathbb{R}^{m\times m}$ mit $a_{ii} \neq 0$, $i = 1, \dots, m$, und jedes $\omega \in \mathbb{R}\setminus\{0\}$ gilt $\rho(B(\omega)) \ge |\omega - 1|$. 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 1.40 *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. Es sei $\omega \in (0,2)$ und $\lambda \in \mathbb{C}$ ein Eigenwert von $B(\omega)$. Es sei $d_0 \neq 0$ Eigenvektor von $B(\omega)$ zum Eigenwert λ. Wir definieren $d_1 := B(\omega)d_0$ und erhalten

$$(*) \quad (D+\omega L)d_1 = ((1-\omega)D - \omega L^\top)d_0\,.$$

Daraus ergibt sich

$$\begin{aligned}(D+\omega L)(d_1 - d_0) &= -\omega(L + D + L^\top)d_0 = -\omega A d_0 \\ ((1-\omega)D - \omega L^\top)(d_0 - d_1) &= \omega A d_1\,.\end{aligned}$$

Wir multiplizieren die beiden letzteren Gleichungen skalar in $\mathbb{C}^m$ ("von links") mit d_0 bzw. d_1, subtrahieren die entstandenen Gleichungen voneinander und erhalten

$$\begin{aligned}\omega(\langle d_0, Ad_0\rangle - \langle d_1, Ad_1\rangle) &= \langle d_0, (D+\omega L)(d_0-d_1)\rangle - \langle d_1, ((1-\omega)D - \omega L^\top)(d_0-d_1)\rangle \\ &= \langle d_0, D(d_0-d_1)\rangle - (1-\omega)\langle d_1, D(d_0-d_1)\rangle \\ &\quad +\omega(\langle d_0, L(d_0-d_1)\rangle + \langle d_1, L^\top(d_0-d_1)\rangle) \\ &= \langle Dd_0, d_0-d_1\rangle - (1-\omega)\langle Dd_1, d_0-d_1\rangle \\ &\quad +\omega\langle L^\top d_0 + Ld_1, d_0-d_1\rangle \\ &= \langle Dd_0, d_0-d_1\rangle - (1-\omega)\langle Dd_1, d_0-d_1\rangle \\ &\quad +\langle (1-\omega)Dd_0 - Dd_1, d_0-d_1\rangle \qquad \text{(wegen (*))} \\ &= (2-\omega)\langle D(d_0-d_1), d_0-d_1\rangle\,.\end{aligned}$$

Da λ Eigenwert mit Eigenvektor d_0 von $B(\omega)$ ist, folgt $d_1 = B(\omega)d_0 = \lambda d_0$ und damit

$$(2-\omega)\langle D(d_0-d_1), (d_0-d_1)\rangle = \omega(1-|\lambda|^2)\langle Ad_0, d_0\rangle\,.$$

Da A und D positiv definit sind und $\omega \in (0,2)$ gilt, folgt $1 - |\lambda|^2 > 0$, d.h. $|\lambda| < 1$. Deshalb gilt $\rho(B(\omega)) < 1$ und die Aussage folgt aus Satz 1.36. $\square$

Bemerkung 1.41 *Satz 1.40 liefert für $\omega = 1$ ein Konvergenzresultat für das Einzelschrittverfahren unter substantiell anderen Voraussetzungen als Satz 1.38. 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). Für lineare Gleichungssysteme, die bei Diskretisierung von (partiellen) Differentialgleichungen entstehen, existieren noch effizientere Iterationsverfahren auf der Basis von Matrizen, die verschiedenen Diskretierungsniveaus entsprechen: Mehrgitterverfahren *(engl.: multigrid methods).*

2 Numerik linearer Optimierungsprobleme

Wir betrachten das folgende lineare Optimierungsproblem:

$$\min\{\langle c, x\rangle : Ax = b,\ x \geq 0\} \tag{2.1}$$

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

$$M := \{x \in \mathbb{R}^m : Ax = b,\ x \geq 0\} \tag{2.2}$$

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

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

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

Bemerkung 2.1
Man nennt die obige Form eines linearen Optimierungsproblems die Standardform eines solchen. Andere Formen linearer Optimierungsprobleme lassen sich stets in ein Problem in Standardform umformulieren.
Wir diskutieren das im folgenden an einigen Beispielen.

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

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

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

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

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

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

Beispiel 2.2 *(Produktionsplanung)*
In einer Firma wird mit m Maschinen ein Produkt produziert, für das der Bedarf d (in einen gewissen Zeitraum) existiert. Die i-te Maschine produziere zu den Kosten c_i und mit maximaler Produktmenge b_i (in diesem Zeitraum), $i = 1, \ldots, m$. Gesucht ist der kostenminimale Einsatz x_i, $i = 1, \ldots, m$, aller Maschinen, um den Bedarf zu decken. Dies führt zu folgendem linearen Optimierungsproblem

$$\min\left\{\sum_{i=1}^{m} c_i x_i : \sum_{i=1}^{m} x_i = d,\ 0 \leq x_i \leq b_i, i = 1, \ldots, m\right\},$$

wobei $\sum_{i=1}^{m} b_i \geq d$ angenommen wird.
Dieses Problem erlaubt eine einfache Lösung. Sind nämlich die Kosten c_i so geordnet, daß $c_1 \leq c_2 \leq \cdots \leq c_m$, so ist

$$x_i := b_i,\ i = 1, \ldots, k-1,\ x_k := d - \sum_{i=1}^{k-1} b_i,\ x_i = 0,\ i = k+1, \ldots, m,$$

eine optimale Lösung, falls $k \in \{1, \ldots, m\}$ so gewählt ist, dass $0 \leq d - \sum_{i=1}^{k-1} b_i \leq b_k$. Kommen aber weitere Restriktionen, wie z.B. die Forderung, daß es eine obere Schranke für den Produktionsausstoß einer Teilgruppe von Maschinen gibt, hinzu oder sind die Kosten stückweise linear konvex, so läßt sich das lineare Optimierungsproblem nicht mehr einfach lösen.

2.1 Polyeder

Wir beschäftigen uns zunächst mit Eigenschaften der Restriktionsmenge des linearen Optimierungsproblems (2.1).

Definition 2.3 *Eine Menge H der Form $H = \{x \in \mathbb{R}^m : \langle a, x\rangle \leq b\}$ mit $a \in \mathbb{R}^m$ und $b \in \mathbb{R}$ heißt Halbraum. Eine Menge P heißt Polyeder, wenn sie Durchschnitt endlich vieler Halbräume ist, d.h. die Form $P = \{x \in \mathbb{R}^m : Ax \leq b\}$ mit $A \in \mathbb{R}^{n\times m}$ und $b \in \mathbb{R}^n$ besitzt. Eine Menge $M \subseteq \mathbb{R}^m$ heißt konvex, falls für alle $x, y \in M$ und $\lambda \in [0, 1]$ gilt $\lambda x + (1 - \lambda)y \in M$.*

Nach Bemerkung 2.1 (i) kann jedes Polyeder evtl. durch Einführung neuer Variabler in die Form M vgl. (2.2) transformiert werden. M ist aber selbst ein Polyeder, da M in der Form $\{x \in \mathbb{R}^m : (-E_m)x \leq 0, Ax \leq b, (-A)x \leq (-b)\}$ geschrieben werden kann.

Folgerung 2.4 *Jedes Polyeder ist eine konvexe abgeschlossene Menge.*

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

Definition 2.5 *Ein Element x einer konvexen Menge M heißt Extrempunkt von M, wenn aus der Gültigkeit der Beziehung $x = \lambda y + (1 - \lambda)z$ für $y, z \in M$ und $0 < \lambda < 1$ bereits $x = y = z$ folgt. E_M bezeichne die Menge aller Extrempunkte von M. Ist M ein Polyeder, so nennen wir einen Extrempunkt von M eine Ecke.*

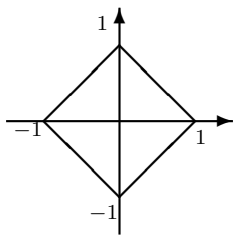
Beispiel 2.6

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

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

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

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



B_1 ist ein Polyeder, es gilt nämlich

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

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

Satz 2.7 *Es seien* $A = (a^1, \ldots, a^m) \in \mathbb{R}^{n \times m}$, *d. h.,* $a^j \in \mathbb{R}^n, j = 1, \ldots, m$; $b \in \mathbb{R}^n$ *und* $x \in M := \{x \in \mathbb{R}^m : Ax = b, x \geq 0\}$.
Dann sind die Aussagen (i) und (ii) äquivalent:
(i) x *ist Ecke von* M*;*
(ii) Die Elemente $a^j, j \in J(x) := \{j \in \{1, \ldots, m\} : x_j > 0\}$ *sind linear unabhängig.*

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

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

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

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

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

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

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

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

Also gilt $y_+, y_- \in M$ und $\frac{1}{2}y_+ + \frac{1}{2}y_- = (x_1, \ldots, x_r, 0, \ldots, 0) = x$, d.h., x ist keine Ecke.
(ii) $\Rightarrow$ (i): Die Elemente $a^j, j \in J(x)$, seien linear unabhängig für $x \in M$. Es seien $y, z \in M$ und $\lambda \in (0, 1)$ mit $x = \lambda y + (1 - \lambda)z$. Dann gilt natürlich $J(x) = J(y) \cup J(z)$. Wieder nehmen wir o.B.d.A. an, daß $J(x) = \{1, \ldots, r\}, r \geq 1$. Dann gilt

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

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

Folgerung 2.8 *Die Menge* M *sei wie in Satz 2.7 definiert und* a^j, $j = 1, \ldots, m$ *seien die Spalten von* $A \in \mathbb{R}^{n \times m}$. *Dann gilt:*
M *besitzt höchstens endlich viele Ecken.*
Ist $\operatorname{rg}(A) = n$ *und* x *eine Ecke von* M, *so existieren* $j_i \in \{1, \ldots, m\}$, $i = 1, \ldots, n$, *so daß* $a^{j_1}, \ldots, a^{j_n}$ *linear unabhängig sind und* $J(x) \subseteq \{j_1, \ldots, j_n\}$ *gilt.*

Beweis:
Mit den Bezeichnungen aus dem Beweis von Satz 2.7 gilt für jede Ecke $x \in M$, daß $|J(x)| \leq \mathrm{rg}(A) \leq \min\{n, m\}$. Überdies ist x aus der Beziehung

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

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

Definition 2.9 *Es sei $A \in \mathbb{R}^{n \times m}$ mit $rg(A) = n$. Dann heißt $x \in M = \{x \in \mathbb{R}^m : Ax = b, x \geq 0\}$ Basispunkt von M, falls Indizes $j_i \in \{1, \ldots, m\}$, $i = 1, \ldots, n$, existieren, so daß die Matrix $A_B = (a^{j_1}, \ldots, a^{j_n})$ von Spalten von A invertierbar ist, $x_j = 0$ für alle $j \notin \{j_1, \ldots, j_n\}$ und $Ax = \sum_{j=1}^{m} x_j a^j = \sum_{i=1}^{n} x_{j_i} a^{j_i} = b$ gilt.*
Bezeichnung: $J_B(x) := \{j_1, \ldots, j_n\}$, $J_N(x) := \{1, \ldots, m\} \setminus J_B(x)$.
Wir nennen eine Komponente x_j eines Basispunktes x Basisvariable, falls $j \in J_B(x)$, und sonst Nichtbasisvariable.
Ein Basispunkt x von M heißt entartet, falls $|J(x)| < n$, anderenfalls nichtentartet.

Nach 2.8 gilt für $A \in \mathbb{R}^{n \times m}$ mit $\mathrm{rg}(A) = n$: $\quad x \in E_M$ gdw. x ist Basispunkt von M.
Als Vorbereitung für einen Darstellungssatz für Elemente von M benötigen wir noch den Begriff einer Richtung.

Definition 2.10 *Ein Element $d \in \mathbb{R}^m$, $d \neq 0$, heißt Richtung in einem Polyeder P, falls für jedes $x_0 \in P$ der Strahl $\{x_0 + \lambda d : \lambda \geq 0\}$ in P liegt.*

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

Satz 2.11 *(Darstellungssatz)*
Es seien $M = \{x \in \mathbb{R}^m : Ax = b, x \geq 0\}$ mit $A \in \mathbb{R}^{n \times m}$, $b \in \mathbb{R}^n$ und $E_M = \{z^1, \ldots, z^\ell\}$. Für jedes $x \in M$ existieren $\lambda_j \geq 0$, $j = 1, \ldots, \ell$, mit $\sum_{j=1}^{\ell} \lambda_j = 1$ und $d \in \mathbb{R}^m$ mit $Ad = 0$, $d \geq 0$, so dass

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

Beweis:
Es sei $x \in M$ beliebig mit $r := |J(x)|$. Wir führen den Beweis durch Induktion über r. Für $r = 0$ ist x nach Satz 2.7 selbst eine Ecke und die Darstellung von x ist trivial. Wir nehmen jetzt an, daß die Aussage für $0, 1, \ldots, r-1$ richtig ist. Es sei jetzt $r = |J(x)|$.

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

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

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

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

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

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

geschrieben werden. Da $-w$ eine Richtung in M ist und x' wie in (a) die gewünschte Gestalt hat, gilt die Darstellung für x mit $\lambda_j := \lambda'_j$, $j = 1, \ldots, \ell$, und $d := d' + \theta'(-w)$.
Fall (c): $w \geq 0$.
In diesem Fall ist w eine Richtung in M und x kann in der Form $x = x'' + (-\theta'')w$ dargestellt werden, wobei x'' und $\theta'' < 0$ wie in Fall (a) gewählt werden. Schließlich setzt man $\lambda_j := \lambda'_j$, $j = 1, \ldots, \ell$, und $d := d'' + (-\theta'')w$. □

Folgerung 2.12 *Jedes Polyeder $M = \{x \in \mathbb{R}^m : Ax = b, x \geq 0\}$ läßt sich in der Form*

$$M = B + K$$

darstellen, wobei B ein beschränktes Polyder und K ein polyedrischer Kegel (d.h. es gilt $\lambda x \in K$, falls $x \in K$ und $\lambda \geq 0$) sind. Überdies gilt $B = \operatorname{conv} E_M$ (konvexe Hülle) *und $K = \{d \in \mathbb{R}^m : Ad = 0, d \geq 0\}$.*

Beweis: folgt sofort aus Satz 2.11. □

Mit Hilfe des Beweises von Satz 2.11 läßt sich nun auch die Existenz einer Ecke für ein Polyeder in Standardform einfach herleiten.

Satz 2.13 *Jedes nichtleere Polyeder M der Form $M = \{x \in \mathbb{R}^m : Ax = b, x \geq 0\}$ mit $A \in \mathbb{R}^{n\times m}$ und $b \in \mathbb{R}^n$ besitzt eine Ecke.*

Beweis: Es sei $\gamma := \min\{|J(x)| : x \in M\}$ und es sei $x \in M$ mit $\gamma := |J(x)|$. Wäre nun x keine Ecke von M, so könnte man analog zum Beweis von Satz 2.11 ein Element $x^{'}$ in M konstruieren mit der Eigenschaft $|J(x^{'})| < \gamma$. Da dies unmöglich ist, muß x eine Ecke von M sein. □

2.2 Existenz und Charakterisierung von Lösungen

Satz 2.14 *(Existenzsatz)*
Das Polyeder $M := \{x \in \mathbb{R}^m : Ax = b,\ x \geq 0\}$ sei nichtleer. Dann gilt entweder $\inf\{\langle c, x\rangle : x \in M\} = -\infty$ oder das Infimum wird in einer Ecke von M angenommen.

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

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

2. Fall: Es sei $x \in M$ beliebig. Nach Satz 2.11 hat x eine Darstellung der Form

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

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

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

und folglich

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

Damit ist die Aussage bewiesen. □

Bemerkung 2.15 *Die Zielfunktion f kann ihr Minimum auch in mehreren (genauer: unendlich vielen Punkten) annehmen. Wichtig ist hier nur, daß mindestens eine Ecke dazu gehört.*

Beispiel: *Wir betrachten das Polyeder*

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

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

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

d.h. die Strecke zwischen den Ecken $(-1, -1)$ *und* $(2, -1)$ *von* M *ist gerade die Lösungsmenge des Optimierungsproblems.*
Im Prinzip zeigt Satz 2.14, wie man zur Lösung eines linearen Optimierungsproblems vorgehen wird: Elemente aus E_M *sind Kandidaten für optimale Lösungen. Satz 2.7 liefert dabei ein Kriterium zur Bestimmung der Ecken. Diese allgemeine Vorgehensweise muß aber noch "effektiviert" werden, da man Ausschau nach Richtungen* d *in* M *mit* $\langle c, d\rangle < 0$ *halten muß und ein "reines" Absuchen der Ecken (bei einer i.a. sehr großen Zahl von Ecken) zu lange dauert. Später werden wir sehen, wie man dieses Absuchen der Ecken ökonomischer gestaltet.*

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

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

lösbar ist. Es sei $\bar{x}$ eine Ecke von M (vgl. Satz 2.13). Nach Folgerung 2.8 ist $\bar{x}$ ein Basispunkt. Für jedes $x \in \mathbb{R}^m$ bezeichnen wir nun

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

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

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

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

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

$$\min\{\langle c_N - (A_B^{-1}A_N)^T c_B, x_N\rangle + \langle c_B, A_B^{-1}b\rangle : A_B^{-1}(b - A_N x_N) \geq 0, x_N \geq 0\}. \quad (2.3)$$

Für die Ecke $\bar{x}$ gilt $\bar{x}_N = 0$ (Folgerung 2.8) und $\bar{x}_B = A_B^{-1}b$ (≥ 0).

Lemma 2.16 *Das lineare Optimierungsproblem (2.1) sei lösbar, es gelte* $rg(A) = n$ *und ausgehend von einer Ecke* $\bar{x}$ *von* M *sei das Problem auf die Form (2.3) reduziert. Gilt* $c_N - (A_B^{-1}A_N)^T c_B \geq 0$*, so sind* $\bar{x}_N$ *bzw.* $\bar{x}$ *Lösungen von (2.3) bzw. (2.1).*

Beweis:
Aus den Bedingungen $c_N - (A_B^{-1}A_N)^T c_B \geq 0$ und $x_N \geq 0$ und der Gestalt der zu minimierenden Funktion in (2.3) folgt sofort, daß $x_N = 0$ eine Lösung von (2.3) ist. Also gilt dies für $\bar{x}_N$ und nach Konstruktion ist $\bar{x}$ eine Lösung von (2.1). □

2.3 Das Simplex-Verfahren

Nach Satz 2.14 und Lemma 2.16 sucht man zur Lösung von (2.1) also eine Ecke $\bar{x}$, so daß mit den Bezeichnungen aus dem letzten Kapitel gilt $c_N - (A_B^{-1}A_N)^T c_B \geq 0$. Wir beschreiben jetzt eine Methodik, wie man aus einer Ecke $\bar{x}$ von M, die keine Lösung von (2.1) ist, eine neue Ecke bestimmen kann, für die der Wert der Zielfunktion kleiner als in $\bar{x}$ ist ("Eckenaustausch").
Ist eine Ecke $\bar{x}$ von M keine Lösung des linearen Optimierungsproblems, so muß nach Lemma 2.16 ein $k \in J_N(\bar{x})$ existieren, so daß

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

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

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

Lemma 2.17
Es gelte $M := \{x \in \mathbb{R}^m : Ax = b, x \geq 0\} \neq \emptyset$, es sei $rg(A) = n$ und $\bar{x} \in E_M$.
Ist $\bar{x}$ keine Lösung von (2.1), so ist entweder $\bar{d}$ eine Richtung in M mit $\langle c, \bar{d}\rangle < 0$ und folglich besitzt (2.1) keine Lösung oder $\bar{x} + \bar{t}\bar{d}$ ist eine Ecke von M mit $\langle c, \bar{d}\rangle < 0$, wobei

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

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

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

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

Die erste Möglichkeit ist nun, daß $\bar{d} \geq 0$ und damit $\bar{x} + t\bar{d} \geq 0$ für alle $t \geq 0$ gilt. Dann ist $\bar{d}$ eine Richtung in M im Sinne von Def. 2.10 und das lineare Optimierungsproblem (2.1) besitzt keine Lösung (vgl. Beweis von Satz 2.14).
Gilt nicht $\bar{d} \geq 0$, so erhalten wir für $x(t) := \bar{x} + t\bar{d}$, daß

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

Also gilt $x(t) \in M$ für alle $t \geq 0$ falls $x_{j_i}(t) \geq 0,\ \forall i = 1, \ldots, n$ mit $[A_B^{-1} A_N e^k]_{j_i} > 0$. Letzteres gilt offenbar für alle $0 \leq t \leq \min\limits_i \frac{[A_B^{-1} b]_{j_i}}{[A_B^{-1} A_N e^k]_{j_i}}$, wobei $i = 1, \ldots, n$ mit $[A_B^{-1} A_N e^k]_{j_i} > 0$ und damit für $t \in [0, \bar{t}]$. Für mindestens ein $i \in \{1, \ldots, n\}$ muß $[A_B^{-1} A_N e^k]_{j_i} > 0$ gelten, da anderenfalls $\bar{d} \geq 0$ gelten würde, was aber jetzt ausgeschlossen ist.
Ist $\bar{x}$ nichtentartet, so ist jede Komponente von $A_B^{-1} b$ positiv. Folglich gilt $\bar{t} > 0$ und damit $\langle c, \bar{x} + \bar{t}\bar{d} \rangle < \langle c, \bar{x} \rangle$.
Wegen $x_{j_{i_0}}(\bar{t}) = 0$ für jeden Index $i_0 \in \{1, \ldots, n\}$, an dem das Minimum angenommen wird, kann j_{i_0} aus $J_B(\bar{x})$ in $J_N(\bar{x} + \bar{t}\bar{d})$ wechseln. Da außerdem $x_k(\bar{t}) = \bar{t}$ gilt, muß der Index k aus $J_N(\bar{x})$ in $J_B(\bar{x} + \bar{t}\bar{d})$ wechseln. Gezeigt werden muß jedoch, daß die Spalten a^{j_i}, a^k, $i = 1, \ldots, n$, $i \neq i_0$, linear unabhängig sind. In diesem Fall würde nach Satz 2.7 folgen, daß $\bar{x} + \bar{t}\bar{d}$ eine Ecke von M ist. Wir nehmen also an, daß die Spalten a^{j_i}, a^k, $i = 1, \ldots, n$, $i \neq i_0$, linear abhängig sind. Dies ist nur möglich, wenn $\alpha_i \in \mathbb{R}$, $i = 1, \ldots, n$, $i \neq i_0$, existieren, so daß $a^k = \sum\limits_{\substack{i=1 \\ i \neq i_0}}^{n} \alpha_i a^{j_i}$. Nun gilt aber nach Konstruktion $a^k = A_N e^k = -A_B \bar{d}_B$, d.h. a^k ist auch eine Linearkombination der a^{j_i}, $i = 1, \ldots, n$. Dies kann aber nur richtig sein, wenn in der letzteren Linearkombination der Faktor von $a^{j_{i_0}}$, also $\bar{d}_{j_{i_0}} = [A_B^{-1} A_N e^k]_{j_{i_0}}$, gleich 0 ist. Dies ist aber nach Wahl von i_0 unmöglich und die gewünschte lineare Unabhängigkeit ist gezeigt. □

Lemma 2.17 offeriert also einen Eckenaustausch indem man einen Indexaustausch in den Mengen J_B bzw. J_N von Indizes für Basisvariable bzw. Nichtbasisvariable vornimmt. Dies führt zu folgendem Grundalgorithmus.

Algorithmus 2.18 *(Simplex-Verfahren)*

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

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

*(iii) Wähle $k \in J_N(x^\ell)$ mit $\langle c_N - (A_B^{-1} A_N)^T c_B, e^k \rangle < 0$ und bestimme $\bar{d}$.
Falls $\bar{d} \geq 0$, stop (das Problem hat keine Lösung)!
Anderenfalls berechne $\bar{t}$ und die Indexmengen $J_B(\bar{x} + \bar{t}\bar{d})$ und $J_N(\bar{x} + \bar{t}\bar{d})$, wobei $\bar{x} = x^\ell$, durch Indexaustausch wie in Lemma 2.17.*

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

Das Verfahren besteht also aus zwei Phasen: In Phase I wird zunächst eine Ecke des zulässigen Bereichs M bestimmt und in Phase II wird durch Ecken- bzw. Index-Austausch eine Ecke von M bestimmt, die auch Lösung von (2.1) ist.

Satz 2.19 *Das Polyeder $M := \{x \in \mathbb{R}^m : Ax = b,\ x \geq 0\}$ sei nichtleer, es gelte $\mathrm{rg}(A) = n$ und alle Ecken von M seien nichtentartet.*
Dann bricht das Simplex-Verfahren 2.18 entweder mit der Information der Unlösbakeit oder nach einer Anzahl von ℓ_* Schritten mit der Information, daß x^{ℓ_*} eine Lösung des linearen Optimierungsproblems ist, ab.
Es gilt $\langle c, x^{\ell+1} \rangle < \langle c, x^\ell \rangle, \forall \ell = 0, \ldots, \ell_* - 1$.

Beweis:
Nach Lemma 2.17 sind alle $x^\ell, \ell = 0, 1, \ldots$ im Simplex-Verfahren Ecken von M. Beim Übergang von x^ℓ zu $x^{\ell+1}$ gilt $\langle c, x^{\ell+1}\rangle < \langle c, x^\ell\rangle$ nach Lemma 2.17, da alle Ecken von M nach Voraussetzung nichtentartet sind. Da die Werte der Zielfunktion streng monoton fallend sind, kann keine Ecke mehrfach auftreten und die Aussage folgt aus der Tatsache, daß M höchstens endlich viele Ecken besitzt (Folgerung 2.8). □

Bemerkung 2.20 *(Phase I)*
In den meisten Fällen läßt sich eine Startecke x^0 nicht auf einfache Art und Weise bestimmen. Jedoch läßt sich das Simplex-Verfahren, in Anwendung auf ein Hilfsproblem, selbst dazu verwenden, eine Startecke bzw. einen Basispunkt zu berechnen. Dazu betrachten wir das lineare Optimierungsproblem

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

mit $m+n$ Variablen, wobei wir o.B.d.A. annehmen, daß $b \geq 0$ gilt. Offenbar ist $(x, y) = (0, b)$ eine Ecke des zulässigen Bereiches. Mit dieser Ecke starten wir das Simplex-Verfahren zur Lösung dieses Problems. Existiert ein zulässiger Punkt in (2.1) (d.h. gilt $M \neq \emptyset$), so endet das Verfahren in einer Ecke, die Lösung des Problems ist und folglich die Form $(\bar{x}, 0)$ hat. Damit ist $\bar{x}$ eine Ecke des Polyeders M und kann als Startecke x^0 verwendet werden. Endet das Simplex-Verfahren mit einem Zielfunktionswert, der größer als 0 ist, so muß M leer sein.

Beispiel 2.21 *(Klee-Minty 1972)*
Wir betrachten das lineare Optimierungsproblem

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

wobei $\varepsilon \in (0, 0.5)$, mit m Variablen und $2m$ Ungleichungen. Offenbar sind $(0, \ldots, 0) \in \mathbb{R}^m$ und $(0, \ldots, 0, 1) \in \mathbb{R}^m$ Ecken des zulässigen Bereiches. Wird das Simplex-Verfahren mit $x_0 := (0, \ldots, 0)$ gestartet und wählt man k stets so aus, dass

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

so durchläuft das Verfahren alle Ecken und endet nach 2^m Schritten in $(0, \ldots, 0, 1)$. Dies ist das worst-case Verhalten des Algorithmus.

Bemerkung 2.22 *(mittlere Laufzeit des Simplex-Verfahrens)*
Nach Einführung des Simplex-Verfahrens durch G.B. Dantzig 1947/48 und durchaus guten praktischen Erfahrungen, schienen Beispiele der Art wie 2.21 und damit eine mögliche exponentielle Anzahl von Schritten, das "Aus" für dieses Verfahren einzuläuten. Dies wurde noch dadurch verstärkt, daß Ende der 70er eine neue Klasse von Lösungsverfahren für lineare Optimierungsprobleme gefunden wurde, die sog. innere-Punkt Verfahren. Jedoch wurde Borgwardt (Augsburg) und Smale Anfang der 80er das mittlere Laufzeitverhalten des Simplex-Verfahrens untersucht und gezeigt, daß die mittlere Anzahl der Schritte nur wie $\min\{m^2, n^2\}$ wächst. Gegenwärtig hält man es sogar noch für möglich, daß eine Index-Austausch-Regel existiert, die ein (nur) polynomiales Wachstum der Anzahl der Schritte mit $\min\{m, n\}$ für jedes lineare Optimierungsproblem garantiert (Todd 2002).

Bemerkung 2.23 *(Entartungen und Implementierung)*
Treten entartete Ecken auf, so kann $\bar{t} = 0$ gelten. Dann bleibt das Simplex-Verfahren in derselben Ecke, berechnet aber eine neue Indexmenge J_B, d.h. eine neue Basis. Im nächsten Schritt könnte deshalb wieder $\bar{t} > 0$ gelten. Um zu verhindern, daß unendliche Zyklen beim Basisaustausch auftreten, werden sogenannte Pivot-Regeln verwendet, die festlegen, welche der Indizes ausgetauscht werden, wenn es mehrere Kandidaten dafür gibt. Z.B. kann man immer die mit dem kleinsten Index verwenden.
Schließlich sei angemerkt, daß auf keinen Fall im Laufe des Simplex-Verfahrens explizit inverse Matrizen der Form A_B^{-1} berechnet werden müssen. In allen Fällen werden LR-Zerlegungen dieser Matrizen berechnet und die entsprechenden Gleichungssysteme damit gelöst. Man kann zeigen, daß bei Spalten-Austausch die neuen LR-Zerlegungen ökonomisch aus den alten berechnet werden können.

Zusätzliche Literatur zu Kapitel 2:
D. Goldfarb, M.J. Todd: Linear Programming, in: Handbooks in Operations Research and Management Science, Vol. 1, Optimization (G.L. Nemhauser, A.H.G. Rinnooy Kan, M.J. Todd eds.), North-Holland, Amsterdam 1989, 73–170.
M.J. Todd: The many facets of linear programming, Mathematical Programming 91 (2002), 417–436.