1.4 Stabilität der Gauß-Elimination

Bezeichne x die exakte Lösung von $Ax = b$ bzw. $\hat{x}$ die mit einem (zunächst beliebigen) Algorithmus berechnete Näherungslösung (inklusive aller Rundungsfehler).

Definition 4. *Der Algorithmus heißt numerisch stabil*
(i) im Sinne der Vorwärtsanalyse, falls

$$\frac{\|x - \hat{x}\|}{\|x\|} \leq C \cdot \mathrm{cond}(A) \cdot eps$$

mit nicht allzu großem C gilt, d.h. der Einfluss von Rundungsfehlern während der Rechnung ist nicht viel größer als der Einfluss von Rundungsfehlern (relative Abweichung der Größenordnung eps) in den Daten.

(ii) im Sinne der Rückwärtsanalyse, falls das numerische Ergebnis $\hat{x}$ als exakte Lösung einer Gleichung $\bar{A}\hat{x} = \bar{b}$ interpretiert werden kann mit

$$\frac{\|A - \bar{A}\|}{\|A\|} \leq C \cdot eps, \qquad \frac{\|b - \bar{b}\|}{\|b\|} \leq C \cdot eps$$

mit nicht allzu großem C.

Bemerkung 5.
(i) Mit der numerischen Stabilität im Sinne der Rückwärtsanalyse folgt die Stabilität der Vorwärtsanalyse aus Satz 4:

$$\frac{\|x - \hat{x}\|}{\|x\|} \leq 2C \cdot \mathrm{cond}(A) \cdot eps + \mathcal{O}(eps^2).$$

(ii) Für die Stabilität der Rückwärtsanalyse ist die Kenntnis der Konditionszahl von A nicht nötig.

(iii) (Deuflhard:) Die Idee der von J.H. Wilkinson eingeführten Rückwärtsanalyse besteht darin, die durch den Algorithmus verursachten Fehler auf die Eingabegröße zurückzuspielen und so als zusätzliche Eingabefehler zu interpretieren. Dazu fassen wir die fehlerbehafteten Resultate als exakte Ergebnisse zu gestörten Eingabegrößen auf.

Bezeichnungen: Im Folgenden interpretieren wir den Vergleich und den Betrag von Matrizen komponentenweise:

$$A \leq B :\Leftrightarrow \ a_{ij} \leq b_{ij} \ \ \forall_{ij}$$
$$|A| := (|a_{ij}|)_{i,j=1,\ldots,n}$$

Beispiel 9. *(Rückwärtsanalyse des Skalarprodukts)*
Das Skalarprodukt $< y, z >$, für $y, z \in \mathbb{R}^n$ lässt sich rekursiv berechnen durch

$$< y, z > = y_n z_n + < y^{n-1}, z^{n-1} >, \tag{1.13}$$

wobei $y^{n-1} := (y_1, \ldots, y_{n-1})^T$ und $z^{n-1} := (z_1, \ldots, z_{n-1})^T$.
Die Gleitpunktrealisierung des Skalarprodukts gemäß (1.13) berechnet für Gleitpunktzahlen y, z den Wert

$$< y, z >_{fl} = < \bar{y}, z >$$

für ein $\bar{y} \in \mathbb{R}^n$ mit

$$|y - \bar{y}| \leq n \cdot eps|y| + \mathcal{O}(eps^2).$$

Beweis durch Induktion: Für $n = 1$ erhalten wir

$$< y, z >_{fl} = y\hat{\cdot}z \;=\; y \cdot z(1+\delta),$$

wobei δ mit $|\delta| \leq eps$ den relativen Fehler der Multiplikation beschreibt. Setze $\bar{y} := y(1+\delta)$. Dann gilt offenbar

$$< y, z, >_{fl} =< \bar{y}, z >$$

und

$$|y - \bar{y}| = |y \cdot \delta| = |\delta||y| \leq eps|y|.$$

Sei $n > 1$ und die Behauptung für $n-1$ bereits bewiesen. Für die Gleitpunktrealisierung der Rekursion (1.13) *gilt:*

$$\begin{aligned} < y, z >_{fl} &= y_n\hat{\cdot}z_n\hat{+} < y^{n-1}, z^{n-1} >_{fl} \\ &= \big(y_n z_n(1+\delta) + < y^{n-1}, z^{n-1} >_{fl}\big)(1+\epsilon), \end{aligned}$$

wobei diesmal δ und ϵ mit $|\epsilon|, |\delta| \leq eps$ die relativen Fehler der Multiplikation bzw. der Addition bezeichnen. Nach Induktionsvoraussetzung gilt ferner

$$< y^{n-1}, z^{n-1} >_{fl} =< c, z^{n-1} >$$

für ein $c \in \mathbb{R}^{n-1}$ mit

$$|y^{n-1} - c| \leq (n-1)eps|y^{n-1}| + \mathcal{O}(eps^2).$$

Wir setzen $\bar{y}_n := y_n(1+\delta)(1+\epsilon)$ und $\bar{y}_k := c_k(1+\epsilon)$ für $k = 1, \ldots, n-1$. Damit folgt:

$$\begin{aligned} < y, z >_{fl} &= y_n z_n(1+\delta)(1+\epsilon) + < y^{n-1}, z^{n-1} >_{fl} (1+\epsilon) \\ &= \bar{y}_n z_n + < \underbrace{c \cdot (1+\epsilon)}_{=\bar{y}^{n-1}}, z^{n-1} > \\ &=< \bar{y}, z > \end{aligned}$$

und

$$\begin{aligned} |y_n - \bar{y}_n| &\leq 2eps|y_n| + eps^2|y_n| \\ |y_k - \bar{y}_k| &\leq |y_k - c_k| + |c_k - \bar{y}_k| \\ &\leq (n-1)eps|y_k| + eps|\bar{y}_k| + \mathcal{O}(eps^2) \\ &\leq n \cdot eps|y_k| + eps|y_k - \bar{y}_k| + \mathcal{O}(eps^2). \end{aligned}$$

Somit gilt auch

$$(1 - eps)|y_k - \bar{y}_k| \leq n \cdot eps|y_k| + \mathcal{O}(eps^2)$$

also

$$\begin{aligned} |y_k - \bar{y}_k| &\leq \frac{n}{1 - eps} \cdot eps|y_k| + \mathcal{O}(eps^2) \\ &= n \cdot eps|y_k| + \mathcal{O}(eps^2) \text{ für } k = 1, \ldots, n-1. \end{aligned}$$

Insgesamt folgt

$$|y - \bar{y}| \leq n \cdot eps|y| + \mathcal{O}(eps^2).$$

Insbesondere ist das Skalarprodukt im Sinne der Rückwärtsanalyse stabil mit $C = n$.

Satz 5. *(Rückwärtsanalyse der Vorwärtssubstitution)*
Sei $L \in \mathbb{R}^{n\times n}$ eine untere Dreiecksmatrix und $b \in \mathbb{R}^n$ ein Vektor jeweils aus Gleitpunktzahlen bestehend. Die Gleitpunktrealisierung der Vorwärtssubstitution zur Lösung eines gestaffelten Gleichungssystems $Lx = b$ berechnet eine Lösung $\hat{x}$, welche exakte Lösung eines Systems $\bar{L}x = b$ ist, mit $\bar{L}$ untere Dreiecksmatrix und

$$|L - \bar{L}| \leq n \cdot eps|L| + \mathcal{O}(eps^2),$$

d.h. die Vorwärtssubstitution ist stabil im Sinne der Rückwärtsanalyse mit $C = n$.

Beweis: Wir betrachten zunächst den einfachen Fall $n = 1$, d.h. die skalare Gleichung $lx = b$. Sei $\hat{x}$ die Lösung von

$$l\hat{\cdot}\hat{x} = b.$$

Es gilt $l\hat{\cdot}\hat{x} = lx(1 + \delta)$, wobei δ mit $|\delta| \leq eps$ den relativen Fehler der Multiplikation beschreibt. Mit $\bar{l} := l(1 + \delta)$ ist somit die Behauptung des Satzes erfüllt.

Im Fall $n > 1$ ist die k-te Komponente des Lösungsvektors $x = (x_1, \ldots, x_n)^T$ bestimmt durch

$$\begin{aligned} l_{kk}x_k &= b_k - (l_{k1}x_1 + \ldots + l_{k,k-1}x_{k-1}) \\ &= b_k - < l^{k-1}, x^{k-1} >, \ k = 1, \ldots, n, \end{aligned}$$

wobei wir wieder die abkürzenden Schreibweisen $l^{k-1} := (l_{k1}, \ldots, l_{k,k-1})^T$ und $x^{k-1} := (x_1, \ldots, x_{k-1})^T$ verwendet haben. Bezeichne $\hat{x}$ die Lösung der Realisierung in Gleitpunkt-Arithmetik

$$l_{kk}\hat{\cdot}\hat{x}_k = b_k \hat{-} < l^{k-1}, \hat{x}^{k-1} >_{fl} .$$

Offenbar gilt dann auch

$$l_{kk}\hat{x}_k(1 + \delta_k) = \big(b_k - < l^{k-1}, \hat{x}^{k-1} >_{fl}\big)(1 + \epsilon_k)$$

für $k = 1, \ldots, n$, wobei δ_k und ϵ_k die relativen Fehler der Multiplikation bzw. der Addition beschreiben mit $|\epsilon_k|, |\delta_k| \leq eps$.
Nach Beispiel 9 wissen wir bereits, dass

$$< l^{k-1}, \hat{x}^{k-1} >_{fl} = < \bar{l}^{k-1}, \hat{x}^{k-1} >$$

für einen Vektor $\bar{l}^{k-1} = (\bar{l}_{k1}, \ldots, \bar{l}_{k,k-1})^T$ mit

$$|l^{k-1} - \bar{l}^{k-1}| \leq (k - 1)eps|l^{k-1}| + \mathcal{O}(eps^2).$$

Setzen wir $\bar{l}_{kk} := l_{kk}(1 + \delta_k)/(1 + \epsilon_k)$, so ist $\bar{L}$ definiert und es gilt die Behauptung des Satzes.

□

Der folgende Satz liefert eine Aussage zur Stabilität der LR-Zerlegung im Sinne der Rückwärtsanalyse.

Satz 6. *(Rückwärtsanalyse der LR-Zerlegung durch Gauß-Elimination)*
Sei $A \in \mathbb{R}^{n\times n}$ eine Matrix von Gleitpunktzahlen, die eine LR-Zerlegung besitzt. Dann berechnet das durch Gleitpunkt-Arithmetik realisierte Gaußsche Eliminationsverfahren Matrizen $\hat{L}$ und $\hat{R}$ mit:

$$|A - \hat{L}\hat{R}| \leq (n + 3)eps|\hat{L}||\hat{R}| + \mathcal{O}(eps^2). \tag{1.14}$$

Beweis: Durch Induktion: $n = 1$ ist klar. Sei $n > 1$ und die Behauptung für $n-1$ bereits gezeigt.
Sei nun A eine $(n \times n)$-Gleitpunktmatrix. Wir schreiben

$$A = \begin{pmatrix} \alpha & w^T \\ v & C \end{pmatrix}$$

mit $\alpha \in \mathbb{R},\ v, w \in \mathbb{R}^{n-1}$ und $C \in \mathbb{R}^{(n-1)\times(n-1)}$.
Die Gauß-Elimination berechnet $z = \frac{v}{\alpha}$ und damit $C^{(1)} = C - zw^T$. Seien $\hat{z}$ und $\hat{C}^{(1)}$ in der entsprechenden Gleitpunktrealisierung berechnet, d.h.

$$\begin{aligned} \hat{z} &= v \hat{/} \alpha \\ \hat{C}^{(1)} &= C \hat{-} \hat{z} \hat{\cdot} w^T. \end{aligned}$$

Dann gilt

$$\begin{aligned} \hat{z}_i &= \frac{v_i}{\alpha}(1+\delta_i) \\ \hat{c}_{ij}^{(1)} &= (c_{ij} - \hat{z}_i w_j (1+\delta_{ij}))(1+\epsilon_{ij}) \end{aligned}$$

mit $|\delta_i|, |\delta_{ij}|, |\epsilon_{ij}| \leq eps$. Damit gilt:

$$|z - \hat{z}| \leq eps|z|.$$

Weiter folgt:

$$\begin{aligned} |\hat{c}_{ij}^{(1)} - c_{ij}^{(1)}| &= |\epsilon_{ij}||c_{ij}| + |\hat{z}_i w_j \underbrace{(1+\delta_{ij})(1+\epsilon_{ij})}_{1+\delta_{ij}+\epsilon_{ij}+\mathcal{O}(eps^2)} - z_i w_j| \\ &\leq eps|c_{ij}| + 2eps|z_i w_j| + |(\hat{z}_i - z_i)w_j| + \mathcal{O}(eps^2) \\ &\leq eps|c_{ij}| + 2eps|z_i w_j| + eps|z_i||w_j| + \mathcal{O}(eps^2) \\ &\leq \big(|c_{ij}| + 3|z_i||w_j|\big)eps + \mathcal{O}(eps^2) \end{aligned}$$

bzw.

$$\begin{aligned} |\hat{C}^{(1)} - C^{(1)}| &\leq eps\big(|\underbrace{C}| + 3|z||w|^T\big) + \mathcal{O}(eps^2) \\ &= C^{(1)} + zw^T \\ &\leq eps\big(|C^{(1)}| + 4|z||w|^T\big). \end{aligned}$$

Der Algorithmus berechnet nun die LR-Zerlegung von $\hat{C}^{(1)}$. Bezeichnen $\hat{L}^{(1)}$ und $\hat{R}^{(1)}$ die durch Gleitpunkt-Arithmetik erhaltenen Matrizen. Nach Induktionsvoraussetzung gilt:

$$|\hat{C}^{(1)} - \hat{L}^{(1)}\hat{R}^{(1)}| \leq (n+2)eps|\hat{L}^{(1)}||\hat{R}^{(1)}| + \mathcal{O}(eps^2).$$

Wir wissen

$$\begin{aligned} \hat{L}\hat{R} = \begin{pmatrix} 1 & 0 \\ \hat{z} & \hat{L}^{(1)} \end{pmatrix}\begin{pmatrix} \alpha & w^T \\ 0 & \hat{R}^{(1)} \end{pmatrix} &= \begin{pmatrix} \alpha & w^T \\ \alpha\hat{z} & \hat{z}w^T + \hat{L}^{(1)}\hat{R}^{(1)} \end{pmatrix} \\ A = LR = \begin{pmatrix} 1 & 0 \\ z & L^{(1)} \end{pmatrix}\begin{pmatrix} \alpha & w^T \\ 0 & R^{(1)} \end{pmatrix} &= \begin{pmatrix} \alpha & w^T \\ \alpha z & zw^T + L^{(1)}R^{(1)} \end{pmatrix}. \end{aligned} \tag{1.15}$$

Somit

$$A - \hat{L}\hat{R} = \begin{pmatrix} 0 & 0 \\ \alpha(z-\hat{z}) & (z-\hat{z})w^T + \underbrace{L^{(1)}R^{(1)}}_{=C^{(1)}} - \hat{L}^{(1)}\hat{R}^{(1)} \end{pmatrix}.$$

Wir schreiben $C^{(1)} = C^{(1)} - \hat{C}^{(1)} + \hat{C}^{(1)}$ und erhalten mit den obigen Abschätzungen

$$|A - \hat{L}\hat{R}| \leq eps \begin{pmatrix} 0 & 0 \\ |\alpha||z| & |z||w|^T + |C^{(1)}| + 4|z||w|^T + (n+2)|\hat{L}^{(1)}||\hat{R}^{(1)}| \end{pmatrix} + \mathcal{O}(eps^2).$$

Mit

$$\begin{aligned} |C^{(1)}| &= |C^{(1)} - \hat{C}^{(1)} + \hat{C}^{(1)} - \hat{L}^{(1)}\hat{R}^{(1)} + \hat{L}^{(1)}\hat{R}^{(1)}| \\ &\leq \underbrace{|C^{(1)} - \hat{C}^{(1)}|}_{=\mathcal{O}(eps)} + \underbrace{|\hat{C}^{(1)} - \hat{L}^{(1)}\hat{R}^{(1)}|}_{=\mathcal{O}(eps)} + |\hat{L}^{(1)}\hat{R}^{(1)}| \\ &= |\hat{L}^{(1)}\hat{R}^{(1)}| + \mathcal{O}(eps) \end{aligned}$$

finden wir

$$\begin{aligned} |A - \hat{L}\hat{R}| &\leq eps \begin{pmatrix} 0 & 0 \\ |\alpha||z| & 5|z||w|^T + (n+3)|\hat{L}^{(1)}||\hat{R}^{(1)}| \end{pmatrix} + \mathcal{O}(eps^2) \\ &\leq \underbrace{(n+3)}_{\geq 5} eps \begin{pmatrix} |\alpha| & |w|^T \\ |\alpha||z| & |z||w|^T + |\hat{L}^{(1)}||\hat{R}^{(1)}| \end{pmatrix} + \mathcal{O}(eps^2). \end{aligned}$$

Investieren wir nun abschließend $|z| = |\hat{z}| + \mathcal{O}(eps)$, so erhalten wir mit (1.15) die Behauptung

$$\begin{aligned} |A - \hat{L}\hat{R}| &\leq (n+3)eps \begin{pmatrix} |\alpha| & |w|^T \\ |\alpha||\hat{z}| & |\hat{z}||w|^T + |\hat{L}^{(1)}||\hat{R}^{(1)}| \end{pmatrix} + \mathcal{O}(eps^2) \\ &\leq (n+3)eps|\hat{L}||\hat{R}| + \mathcal{O}(eps^2). \end{aligned}$$

□

Bemerkung 6. *Wichtige Frage im Zusammenhang der Stabilität: Können $|\hat{L}|$ und $|\hat{R}|$ in Abschätzung (1.14) groß gegenüber den Einträgen in A werden?*
Bei Spaltenpivotsuche gilt:

$$|l_{ij}| \leq 1$$

für alle $i, j = 1, \ldots, n$. Für die Elemente der Matrix $\hat{R}$ sieht die Situation jedoch nicht so gut aus. Hier gilt im Allgemeinen:

$$\max_{i,j} |\hat{r}_{ij}| \leq 2^{n-1} \cdot \max_{i,j} |a_{ij}|.$$

Diese Abschätzung ist meist zu pessimistisch kann aber auftreten. Bei zufällig gewählten Matrizen A wird

$$\max_{i,j} |\hat{r}_{ij}| \approx n \cdot \max_{i,j} |a_{ij}|$$

beobachtet.

Satz 7. *(Rückwärtsanalyse der Gauß-Elimination ohne Pivotwahl)*
Seien $A \in \mathbb{R}^{n \times n}$ eine Matrix und $b \in \mathbb{R}^n$ ein Vektor von Gleitpunktzahlen. Des Weiteren besitze A eine LR-Zerlegung und es seien $\hat{L}, \hat{R}$ wie in Satz 6. Das in Gleitpunkt-Arithmetik erhaltene Ergebnis $\hat{x}$ von $\hat{L}\hat{c} = b,\ \hat{R}x = \hat{c}$ erfüllt

$$\bar{A}\hat{x} = b$$

für eine Matrix $\bar{A}$ mit

$$|A - \bar{A}| \leq 3(n+1)eps|\hat{L}||\hat{R}| + \mathcal{O}(eps^2).$$

Beweis: Ohne Rundungsfehler wäre

$$\left.\begin{array}{rcl} A & = & LR \\ Lc & = & b \\ Rx & = & c \end{array}\right\} \Rightarrow Ax \;=\; b.$$

Statt der exakten LR-Zerlegung haben wir $\hat{L}$ und $\hat{R}$. Nach Satz 5 erhalten wir in der Gleitpunkt-Arithmetik $\hat{x}$ als Lösung von

$$\bar{\hat{L}}\hat{c} = b$$
$$\bar{\hat{R}}x = \hat{c}$$

mit

$$|\hat{L} - \bar{\hat{L}}| \leq n \cdot eps|\hat{L}| + \mathcal{O}(eps^2)$$
$$|\hat{R} - \bar{\hat{R}}| \leq n \cdot eps|\hat{R}| + \mathcal{O}(eps^2).$$

Wir setzen $\bar{A} := \bar{\hat{L}}\bar{\hat{R}}$ und erhalten somit

$$\bar{A}\hat{x} = b$$

und

$$\begin{aligned} |A - \bar{A}| &= |A - \hat{L}\hat{R} + \hat{L}\hat{R} - \bar{\hat{L}}\hat{R} + \bar{\hat{L}}\hat{R} - \bar{\hat{L}}\bar{\hat{R}}| \\ &\leq \underbrace{|A - \hat{L}\hat{R}|}_{\leq (n+3)|\hat{L}||\hat{R}|eps + \mathcal{O}(eps^2)} + |\hat{L} - \bar{\hat{L}}||\hat{R}| + \underbrace{|\bar{\hat{L}}|}_{=|\hat{L}| + \mathcal{O}(eps)} |\hat{R} - \bar{\hat{R}}| \\ &\leq 3(n+1)eps|\hat{L}||\hat{R}| + \mathcal{O}(eps^2). \end{aligned}$$

□

Satz 8. *(Rückwärtsanalyse der Gauß-Elimination mit Spaltenpivotwahl)*
Seien $A \in \mathbb{R}^{n\times n}$ eine Matrix und $b \in \mathbb{R}^n$ ein Vektor von Gleitpunktzahlen. Des Weiteren sei die Gauß-Elimination mit Spaltenpivotwahl durchführbar, d.h. $PA = LR$ für eine Permutationsmatrix P und L, R der LR-Zerlegung. Die Gauß-Elimination mit Spaltenpivotwahl für das Gleichungssystem $Ax = b$ in der Gleitpunkt-Arithmetik berechnet ein $\hat{x}$, so dass

$$\bar{A}\hat{x} = b$$

für eine Matrix $\bar{A}$ mit

$$\frac{\|A - \bar{A}\|_\infty}{\|A\|_\infty} \leq 3(n+1)n^2 \frac{\alpha_{max}}{\max_{i,j}|a_{ij}|} eps + \mathcal{O}(eps^2), \tag{1.16}$$

wobei α_{max} der größte Betrag eines Elements ist, welches im Laufe des Verfahrens in den Matrizen $A^{(1)}$ bis $A^{(n-1)}$ auftritt.

Beweis: Das Verfahren liefert in der Gleitpunkt-Arithmetik $\hat{P}, \hat{L}, \hat{R}$ und $\hat{x}$. Dann besitzt $\hat{P}A$ eine LR-Zerlegung und $\hat{L}$ und $\hat{R}$ sind die in der Gleitpunkt-Arithmetik berechneten Dreiecksmatrizen. Nach Satz 7 existiert eine Matrix $\overline{PA}$ mit

$$\overline{PA}\hat{x} = \hat{P}b$$

und

$$|\hat{P}A - \overline{PA}| \leq 3(n+1)eps|\hat{L}||\hat{R}| + \mathcal{O}(eps^2).$$

Wir definieren $\bar{A} := \hat{P}^T\overline{PA}$ und finden mit der Identität $\hat{P}^T\hat{P} = I$ die Abschätzung

$$\begin{aligned}\|A - \bar{A}\|_\infty &= \|\hat{P}^T\hat{P}A - \hat{P}^T\overline{PA}\|_\infty \\ &\leq \underbrace{\|\hat{P}^T\|_\infty}_{=1}\|\hat{P}A - \overline{PA}\|_\infty \\ &\leq 3(n+1)eps\|\hat{L}\|_\infty\|\hat{R}\|_\infty + \mathcal{O}(eps^2).\end{aligned}$$

Die Spaltenpivotwahl sorgt dafür, dass alle Komopnenten von $\hat{L}$ vom Betrag kleiner oder gleich 1 sind, d.h.

$$\|\hat{L}\|_\infty \leq n.$$

Die Norm von $\hat{R}$ können wir abschätzen durch

$$\begin{aligned}\|\hat{R}\|_\infty &\leq n \cdot \max_{i,j}|\hat{r}_{ij}| \\ &\leq n \cdot \alpha_{max}.\end{aligned}$$

Insgesamt folgt also

$$\|A - \bar{A}\|_\infty \leq 3(n+1)n^2\alpha_{max}eps + \mathcal{O}(eps^2). \tag{1.17}$$

Die Behauptung folgt nun leicht aus (1.17) und aus $\max_{i,j}|a_{ij}| \leq \| A\|_\infty$.

□

Bemerkung 7.

(i) Tatsächlich gilt (1.16) *auch mit* $3(n+1)n^2$ *ersetzt durch* $2n^3$ *(siehe Deuflhard).*

(ii) Die Stabilität der Gauß-Elimination mit Spaltenpivotwahl im Sinne der Rückwärtsanalyse wird somit durch die Größe des Faktors

$$\rho_n(A) := \frac{\alpha_{max}}{\max_{ij}|a_{ij}|}$$

bestimmt. Allgemein gilt

$$\rho_n(A) \leq 2^{n-1},$$

wobei die Schranken (in pathologischen Fällen) tatsächlich angenommen wird. Die Gauß-Elimination mit Spaltenpivotwahl ist also über die ganze Menge der invertierbaren Matrizen nicht stabil. Doch für Matrizen mit bestimmten Strukturen ist $\rho_n(A)$ wesentlich kleiner und das Verfahren stabil. Für symmetrische positiv definite Matrizen gilt zum Beispiel $\rho_n(A) = 1$.
Denn nach Satz 7 gilt im Fall einer symmetrisch positiv definiten Matrix

$$|A - \bar{A}| \leq 3(n+1)eps|\hat{L}||\hat{L}^T| + \mathcal{O}(eps^2) \tag{1.18}$$

mit

$$|A - \hat{L}\hat{L}^T| \leq (n+3)eps|\hat{L}||\hat{L}^T| + \mathcal{O}(eps^2) \;=\; \mathcal{O}(eps),$$

also $\hat{L}\hat{L}^T = A + \mathcal{O}(eps)$. *Die Matrix* $|\hat{L}|$ *kann jedoch im Verhältnis zu* $a := \max_{ij} |a_{ij}|$ *nicht groß werden. Denn*

$$a_{ii} + \mathcal{O}(eps) = \sum_{k=1}^{i} \hat{l}_{ij}^2 \;\geq\; \hat{l}_{ij}^2$$

für alle j *und daher*

$$|\hat{l}_{ij}| \leq \sqrt{a} + \mathcal{O}(eps).$$

Mit der Abschätzung $\|\hat{L}\|_\infty \leq n\sqrt{a} + \mathcal{O}(eps)$, *welche so offenbar auch für die Transponierte von* $\hat{L}$ *gilt, folgt mit Ungleichung* (1.18) *die Abschätzung*

$$\frac{\|A - \bar{A}\|_\infty}{\|A\|_\infty} \leq 3(n+1)n^2 eps + \mathcal{O}(eps^2),$$

d.h. der Nachweis für $\rho_n(A) = 1$.
Für tridiagonale Matrizen

$$A = \begin{pmatrix} * & * & & \\ * & \ddots & \ddots & \\ & \ddots & \ddots & * \\ & & * & * \end{pmatrix}$$

gilt $\rho_n(A) \leq 2$ *und für obere Hessenberg-Matrizen*

$$A = \begin{pmatrix} * & \cdots & \cdots & * \\ * & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & * & * \end{pmatrix}$$

gilt $\rho_n(A) \leq n$ *(vgl. Ü).*