4.4 Rundungsfehleranalyse der Gauß-Elimination
(LR-Zerlegung)

Absolutwertnotation

Für $A \in \mathbb{R}^{m \times n}$ ist

$$B = |A| \Rightarrow b_{ij} = |a_{ij}| \quad 1 \le i \le m,\ 1 \le j \le n$$

Die Abbildung $rd : \mathbb{R} \to \mathbb{F}$ erweitern wir entsprechend auf $\mathbb{R}^n$, $\mathbb{R}^{m \times n}$.

Dann gilt

$$rd(A) = A + A' \quad \text{mit} \quad \underbrace{|A'| \le |A| eps}_{\text{dies sind } m \cdot n \text{ Ungleichungen}}$$

Wdh. aus Analyse von rd:

$$rd(a_{ij}) = a_{ij}(1+\varepsilon_{ij}) = a_{ij} + \underbrace{a_{ij}\varepsilon_{ij}}_{a_{ij}'}$$

$$|a_{ij}'| \le |a_{ij}| eps$$

fl-Notation

→ ~~nächstes Mal früher einführen~~

Sei E eine Formel, dann bezeichne $fl(E)$ eine Berechnung der Formel E in Fließkommaarithmetik (dabei wird die Ausführungsreihenfolge meist offensichtlich sein, sonst ist sie anzugeben).
(links nach rechts, Klammern)

Somit ist zum Beispiel

$$fl(A+B) = \underbrace{(A+B)}_{\text{exakte Rechnung in } \mathbb{R}} + H \quad \text{mit} \quad |H| \le eps(|A+B|)$$

$A, B \in \mathbb{F}^{m \times n}$

Nochmal: $|.|$ ist keine Norm sondern eine Matrix.

(folgt aus

$$fl(a_{ij} + b_{ij}) = a_{ij} \oplus b_{ij}$$
$$= (a_{ij} + b_{ij})(1+\varepsilon_{ij}) \quad |\varepsilon_{ij}| \le eps.$$
$$= (a_{ij} + b_{ij}) + \varepsilon_{ij}(a_{ij} + b_{ij})$$

Man kann auch

$$fl(\sqrt{x})$$

oder

$$fl(\sin(x))$$

schreiben.

Rückwärtsanalyse

Bisher haben wir die „Vorwärtsanalyse" der Rundungsfehler betrieben.
Z.B. gilt für das Skalarprodukt $x^T y$:

$$|fl(x^T y) - x^T y| \leq n\,\text{eps}\,|x|^T|y| + O(\text{eps}^2) \quad \text{(absolut)}$$

oder

$$\frac{|fl(x^T y) - x^T y|}{|x^T y|} \leq n\,\text{eps}\,\frac{|x|^T|y|}{|x^T y|} + O(\text{eps}^2) \quad \text{(relativ)}.$$

Eine Alternative ist die sog. „Rückwärtsanalyse".
Dort versucht man das Fließkommaergebnis als exaktes Ergebnis eines modifizierten Ausdrucks zu schreiben.

Beispiel 4.8 Betrachte Lösung des LGS $Ax = b$. Die GEM berechnet die numerische Lösung $\hat{x}$.

Vorwärtsanalyse: $\|\hat{x} - x\| \leq F(\text{eps}, n, A, b)$

Rückwärtsanalyse: $(A+E)\hat{x} = b$ mit $|E| \leq F'(\text{eps}, n, A)$

Mit dem Störungssatz 3.18 und $\|E\|_\infty = \||E|\|_\infty$ folgt dann für d. Rundungsfehler:

$$\frac{\|\hat{x} - x\|_\infty}{\|x\|_\infty} \leq \frac{\text{cond}(A)}{1 - \text{cond}(A)\frac{\||E|\|_\infty}{\|A\|_\infty}} \underbrace{\frac{\||E|\|_\infty}{\|A\|_\infty}}$$

ist noch entsprechend abzuschätzen,
ideal wäre: $\||E|\|_\infty \leq n^p\,\text{eps}\,\|A\|_\infty$

Rundg.: $rd(A) = A + \delta A$
$|\delta A| \leq |A|\,\text{eps}$
Störungssatz:
$$\frac{\|\delta A\|_\infty}{\|A\|_\infty} = \frac{\||\delta A|\|_\infty}{\|A\|_\infty} \leq \frac{\|A\|_\infty\,\text{eps}}{\|A\|_\infty} = \text{eps}$$

Somit ist ein direkter Vergleich mit der Konditionsanalyse möglich.

Rückwärtsanalyse des Skalarproduktes.

Hilfssatz 4.9 Es gilt für $x, y \in \mathbb{F}^n$:

$$\hat{s} = fl(x^T y) = (x+f)^T y, \qquad |f| \le n \, eps \, |x| + O(eps^2)$$

Beweis: Induktion über n.

$\underline{n=1}$: $\hat{s}_1 = fl(x_1 y_1) = x_1 y_1 (1+\delta_1) = (x_1 + \underbrace{x_1\delta_1}_{=: f_1}) y_1$

also $|f_1| = |\delta_1||x_1| \le eps \, |x_1|$.

$\underline{n \ge 2}$: $\hat{s}_n = fl(x^T y) = fl(\tilde{x}^T \tilde{y} + x_n y_n)$

$x = \begin{pmatrix} \tilde{x} \\ x_n \end{pmatrix}, y = \begin{pmatrix} \tilde{y} \\ y_n \end{pmatrix}$

$$= \left(fl(\tilde{x}^T \tilde{y}) + fl(x_n y_n)\right)(1+\varepsilon_n)$$

$$= \left((\tilde{x}+\tilde{f})^T \tilde{y} + x_n y_n (1+\delta_n)\right)(1+\varepsilon_n)$$

$$= (\tilde{x}+\tilde{f})^T \tilde{y}(1+\varepsilon_n) + x_n y_n \left(1 + (\delta_n + \varepsilon_n) + \delta_n \varepsilon_n\right)$$

$$= (\tilde{x} + \underbrace{\tilde{f} + \varepsilon_n \tilde{x} + \varepsilon_n \tilde{f}})^T \tilde{y} + \left(x_n + (\delta_n + \varepsilon_n) x_n + \delta_n \varepsilon_n x_n\right) y_n$$

$$= \left[\underbrace{\begin{pmatrix} \tilde{x} \\ x_n \end{pmatrix}}_{=x} + \underbrace{\begin{pmatrix} \tilde{f} + \varepsilon_n \tilde{x} + \varepsilon_n \tilde{f} \\ (\delta_n + \varepsilon_n) x_n + \delta_n \varepsilon_n x_n \end{pmatrix}}_{=: f}\right] \underbrace{\begin{pmatrix} \tilde{y} \\ y_n \end{pmatrix}}_{= y}$$

mit $$|\tilde{f} + \varepsilon_n \tilde{x} + \varepsilon_n \tilde{f}| \le (n-1) eps \, |\tilde{x}| + eps \, |\tilde{x}| + O(eps^2)$$
$$\le n \, eps \, |\tilde{x}| + O(eps^2)$$

und $$|(\delta_n + \varepsilon_n) x_n + \delta_n \varepsilon_n x_n| \le 2 \, eps \, |x_n| + O(eps^2)$$

wg $n \ge 2$ gilt also $$|f| \le n \, eps \, |x| + O(eps^2)$$ ∎

Bemerkung 4.10 $n \, eps$ in ~~5.8~~ 4.9 ist der schlechteste Fall und sehr pessimistisch. Der Rundungsfehler in einer Operation hängt von den Argumenten ab und sie können sich auch weg heben (positiv / negativ!). Besser wäre eine statistische Betrachtung. ∎

Satz 4.11 (Lösen von Dreieckssystemen)

Es seien $\hat{x}$ bzw $\hat{y}$ die numerischen Lösungen des unteren bzw. oberen Dreieckssysteme $Lx = b$ und $Ry = c$. Dann gilt

$$(L + F)\hat{x} = b \qquad |F| \le n\,\mathrm{eps}\,|L| + O(\mathrm{eps}^2)$$

$$(R + G)\hat{y} = c \qquad |G| \le n\,\mathrm{eps}\,|R| + O(\mathrm{eps}^2)$$

Beweis: (evtl als Übung). Induktion über n,

$n = 1$: $\quad l_{11} x_1 = b_1 \to \quad \hat{x}_1 = fl(b_1 / l_{11}) = (b_1 / l_{11})(1 + \varepsilon_1)$

$$\Leftrightarrow \frac{1}{1+\varepsilon_1}\hat{x}_1 = \frac{b_1}{l_{11}} \Leftrightarrow \frac{l_{11}}{1+\varepsilon_1}\hat{x}_1 = b_1$$

$\frac{1}{1+\varepsilon_1} = 1 +$

$\partial_\varepsilon \frac{1}{1+\varepsilon_1} = -\frac{1}{(1+\varepsilon_1)^2}$

$$\overset{?}{\Longleftrightarrow} \left(1 - \varepsilon_1 + \tilde{R}(\varepsilon^2)\right) l_{11}\hat{x}_1 = b_1$$

$$\Leftrightarrow \Big(l_{11} \underbrace{-\varepsilon_1 l_{11} + l_{11}\tilde{R}(\varepsilon^2)}_{= f_1}\Big)\hat{x}_1 = b_1$$

mit $|f_1| \le \mathrm{eps}\,|l_{11}| + O(\mathrm{eps}^2)$

$n \ge 2$:

$$Lx = b \Leftrightarrow \begin{array}{c} \\ n-1 \\ \\ 1 \end{array}\!\!\begin{array}{c} \begin{array}{cc} n-1 & 1 \end{array} \\ \begin{pmatrix} L_1 & 0 \\ v^T & \alpha \end{pmatrix} \end{array} \begin{pmatrix} x_1 \\ w \end{pmatrix} = \begin{pmatrix} b_1 \\ \beta \end{pmatrix}$$

Berechne Lösg. $\hat{x}_1$ von $L_1 x_1 = b_1$ und rechne

$$\hat{w} = fl\left((\beta - v^T\hat{x}_1)/\alpha\right)$$

Fehler in $/\alpha$

$$= \left(fl(\beta - v^T\hat{x}_1)/\alpha\right)(1 + \varepsilon_n)$$

Skalarprodukt

$$= \left((\beta - fl(v^T\hat{x}_1))(1+\delta_n)/\alpha\right)(1+\varepsilon_n) = \left((\beta - fl(v^T\hat{x}_1))/\alpha\right)(1+\delta_n)(1+\varepsilon_n)$$

$$\overset{\text{HS. 5.8}}{=} \left((\beta - (v+f)^T\hat{x}_1)/\alpha\right)\left(1 + (\varepsilon_n + \delta_n) + \varepsilon_n\delta_n\right)$$

Satz 4.12 (Rückwärtsanalyse der **LR**-Zerlegung)

Sei $A \in \mathbb{F}^{n\times n}$. Es werde die LR-Zerlegung <u>ohne</u> Pivotierung berechnet. Dann gilt für die numerisch berechneten $\hat{L}, \hat{R}$:

$$\hat{L}\hat{R} = A + H \quad \text{mit} \quad |H| \leq 3(n-1)\,\text{eps}\left(|A| + |\hat{L}||\hat{R}|\right) + O(\text{eps}^2)$$

Beweis: [Golub/van Loan, THM 3.3.1] Induktion über n.

<u>$n=1$:</u> Es ist $\hat{l}_{11} = 1$ und $\hat{r}_{11} = a_{11}$ und damit $H = 0$.

<u>$n \geq 2$:</u> Schreibe

$$A = \begin{array}{cc} \begin{array}{cc} 1 & n-1 \end{array} & \\ \begin{bmatrix} \alpha & w^T \\ v & B \end{bmatrix} & \begin{array}{c} 1 \\ \\ n-1 \end{array} \end{array}$$

Dann macht die LR-Zerlegung:

a) $\hat{z} = fl(v/\alpha)$

b) $\hat{A}_1 = fl(B - \hat{z} w^T)$

c) Berechne LR-Zerlegung von $\hat{A}_1$

Fehler in $\hat{z}$:

$(\square_1)$ $\quad \hat{z} = \frac{v}{\alpha} + f \quad$ mit $\quad |f| \leq \text{eps} \frac{|v|}{|\alpha|}$

Fehler in $\hat{A}_1$:

$$\hat{A}_1 = fl(B - \hat{z} w^T)$$

exakt gerundet

$$= B - fl(\hat{z} w^T) + G \qquad |G| \leq \text{eps}\,|B - fl(\hat{z} w^T)|$$

Bistriff-

$$= B - (\hat{z} w^T + \underbrace{G') + G}_{=:\,F} \qquad |G'| \leq \text{eps}\,|\hat{z} w^T| \leq \text{eps}\,|\hat{z}||w|^T$$

exakt produkt! genau eine Operation pro Matrixeintrag.

$(\square_2)$ $\quad \hat{A}_1 = B - \hat{z} w^T + F$

$$\text{mit} \quad |F| = |G' + G| \leq |G'| + |G|$$

$$\leq \underbrace{\text{eps}\,|\hat{z}||w|^T}_{|G'|} + \text{eps}\,|B - \hat{z} w^T - G'|$$

$$\leq \text{eps}\,|\hat{z}||w|^T + \text{eps}\left(|B| + |\hat{z}||w|^T\right) + \text{eps}^2 |\hat{z}||w|^T$$

$$\text{also} \quad |F| \leq 2\,\text{eps}\left(|B| + |\hat{z}||w|^T\right) + O(\text{eps}^2)$$

Mit $\frac{1}{1+(\varepsilon_n+\delta_n)+\varepsilon_n\delta_n} = 1-(\varepsilon_n+\delta_n)+R(\varepsilon_n^2+\delta_n^2+\varepsilon_n\delta_n)$ gilt

$$\left(1-(\varepsilon_n+\delta_n)+R(..)\right)\hat{w} = \left(\beta-(v+f)^T\hat{x}_1\right)/\alpha$$

$$\Leftrightarrow \quad (v+f)^T\hat{x}_1 + \left(1-(\varepsilon_n+\delta_n)+R(\ldots)\right)\alpha\hat{w} = \beta.$$

Mit der Induktionsvoraussetzung $(L_1+F_1)\hat{x}_1 = b_1$ folgt also für $\hat{x}$.

$$\begin{pmatrix} L_1+F_1 & 0 \\ (v+f)^T & \left(1-(\varepsilon_n+\delta_n)+R(..)\right)\alpha \end{pmatrix}\begin{pmatrix}\hat{x}_1 \\ \hat{w}\end{pmatrix} = \begin{pmatrix} b_1 \\ \beta\end{pmatrix}$$

$$\Leftrightarrow \left[\underbrace{\begin{pmatrix} L_1 & 0 \\ v^T & \alpha\end{pmatrix}}_{=L} + \underbrace{\begin{pmatrix} F_1 & 0 \\ f^T & \left(-(\varepsilon_n+\delta_n)+R(\ldots)\right)\alpha\end{pmatrix}}_{=F}\right]\begin{pmatrix}\hat{x}_1 \\ \hat{w}\end{pmatrix} = \begin{pmatrix} b_1 \\ \beta\end{pmatrix}$$

Wegen
$$|F_1| \le (n-1)\,\mathrm{eps}\,|L| + O(\mathrm{eps}^2)$$
$$|f|^T \le (n-1)\,\mathrm{eps}\,|v|^T + O(\mathrm{eps}^2)$$
$$\left|\left(-(\varepsilon_n+\delta_n)+R(\ldots)\right)\alpha\right| \le 2\,\mathrm{eps}\,|\alpha| + O(\mathrm{eps}^2)$$

hier kommt das n-1 her!
nicht durch die Rekursion.
es geht also nicht besser

gilt für $n \geq 2$ dann

$$|F| \le n\,\mathrm{eps}\,|L| + O(\mathrm{eps}^2)$$

(möglich wäre auch $\max(n-1, 2)$, ist aber nicht wirklich besser).

Analog für $Ry=c$. ▨

Nun wird $\hat{A}_1$ LR-zerlegt und es gilt die Induktionsannahme

$$(U_3) \qquad \hat{L}_1 \hat{R}_1 = \hat{A}_1 + H_1 \quad \text{mit } |H_1| \le 3(n-2)\,eps\,(|\hat{A}_1| + |\hat{L}_1||\hat{R}_1|) + O(eps^2)$$

also ist die Blockform der LR-Zerlegung von A:

$$\hat{L}\hat{R} = \begin{bmatrix} 1 & 0 \\ \vec{z} & \hat{L}_1 \end{bmatrix} \begin{bmatrix} \alpha & w^T \\ 0 & \hat{R}_1 \end{bmatrix} = \begin{bmatrix} \alpha & w^T \\ \vec{z}\alpha & \vec{z}w^T + \underbrace{\hat{L}_1 \hat{R}_1}_{= \hat{A}_1 + H_1} \end{bmatrix}$$

Darstellung von oben einsetzen $(U_1), (U_2), (U_3)$

$$\overset{\downarrow}{=} \begin{bmatrix} \alpha & w^T \\ \left(\frac{v}{\alpha} + f\right)\alpha & \vec{z}w^T + \underbrace{(B - \vec{z}w^T + F)}_{= \hat{A}_1} + H_1 \end{bmatrix}$$

$$= \underbrace{\begin{bmatrix} \alpha & w^T \\ v & B \end{bmatrix}}_{= A} + \underbrace{\begin{bmatrix} 0 & 0 \\ \alpha f & H_1 + F \end{bmatrix}}_{=: H}$$

Damit hat man schon mal die Form $\hat{L}\hat{U} = A + H$. Bleibt noch H abzuschätzen.

In H steckt $H_1 + F$, in $|H_1|$ steckt $|\hat{A}_1|$ also:

$$|\hat{A}_1| = |B - \vec{z}w^T + F| \le |B| + |\vec{z}||w|^T + |F|$$

Abschätzung für $|F|$

$$\overset{\rightarrow}{\le} |B| + |\vec{z}||w|^T + 2eps\,(|B| + |\vec{z}||w|^T) + O(eps^2)$$

$$\le (1 + 2eps)(|B| + |\vec{z}||w|^T) + O(eps^2)$$

$$|H_1 + F| \le |H_1| + |F|$$

Induktionsannahme

$$\overset{\rightarrow}{\le} 3(n-2)\,eps\,(|\hat{A}_1| + |\hat{L}_1||\hat{R}_1|) + 2eps\,(|B| + |\vec{z}||w|^T) + O(eps^2)$$

Abschätzung für $|\hat{A}_1|$ einsetzen

$$\overset{\rightarrow}{\le} 3(n-2)\,eps\left[(1 + 2eps)(|B| + |\vec{z}||w|^T) + |\hat{L}_1||\hat{R}_1|\right] + 2eps\,(|B| + |\vec{z}||w|^T) + O(eps^2)$$

$$\le 3(n-1)\,eps\left[|B| + |\vec{z}||w|^T + |\hat{L}_1||\hat{R}_1|\right] + O(eps^2)$$

Und damit

$$|H| = \begin{bmatrix} 0 & 0 \\ |\alpha f| & |H_1 + F| \end{bmatrix} \overset{(a_1)}{\leq} \begin{bmatrix} 0 & 0 \\ \mathrm{eps}\,|v| & 3(n-1)\,\mathrm{eps}\left(|B| + |\vec{z}|\,|w|^T + |\vec{L}_1^{-1}|\,|\hat{R}_1|\right) \end{bmatrix} + O(\mathrm{eps}^2)$$

$$\leq 3(n-1)\,\mathrm{eps} \begin{bmatrix} 0 & 0 \\ |v| & |B| + |\vec{z}|\,|w|^T + |\vec{L}_1^{-1}|\,|\hat{R}_1| \end{bmatrix} + O(\mathrm{eps}^2)$$

$$\leq 3(n-1)\,\mathrm{eps}\left(\underbrace{\begin{bmatrix} |\alpha| & |w|^T \\ |v| & |B| \end{bmatrix}}_{|A|} + \underbrace{\underbrace{\begin{bmatrix} 1 & 0 \\ |\vec{z}| & |\vec{L}_1| \end{bmatrix}}_{|\vec{L}|} \underbrace{\begin{bmatrix} |\alpha| & |w|^T \\ 0 & |\hat{R}_1| \end{bmatrix}}_{|\hat{R}|}}_{\begin{bmatrix} |\alpha| & |w|^T \\ |\vec{z}|\,|\alpha| & |\vec{z}|\,|w|^T + |\vec{L}_1|\,|\hat{R}_1| \end{bmatrix}} \right)$$

0 Terme sind hinzugefügt. z.B.: (1,1)

Nun sind noch die Dreieckssysteme aufzulösen.

Folgerung

Nach Satz 4.13 ist $\hat{x}$ exakte Lösung des modifizierten Systems

$$(A+E)\hat{x} = b.$$

Mit dem Störungssatz gilt dann in $\|.\|_\infty$-Norm (Beachte: $\|B\|_\infty = \||B|\|_\infty$)

$$\frac{\|\hat{x}-x\|_\infty}{\|x\|_\infty} \leq \operatorname{cond}(A) \cdot \left\{ \underbrace{3n\,\mathrm{eps} \underbrace{\frac{\|A\|_\infty}{\|A\|_\infty}}_{=1}}_{\substack{\text{vergleichbar} \\ \text{mit Rundungs-} \\ \text{fehler } \|rd(A)\|/\|A\|}} + 5n\,\mathrm{eps} \underbrace{\frac{\|\hat{L}\|\,\|\hat{R}\|}{\|A\|_\infty}}_{\text{Problem!}} + O(\mathrm{eps}^2) \right\}$$

!

bis auf Faktor n.

O.K. letztes Mal $rd(A) = A + \delta A$ mit $|\delta A| \leq |A|\mathrm{eps}$ $\Rightarrow \frac{\|\delta A\|_\infty}{\|A\|_\infty} \leq \mathrm{eps}\frac{\|A\|_\infty}{\|A\|_\infty} = \mathrm{eps}$

Nehmen an, dass $\|E\|_\infty / \|A\|_\infty \ll 1$ (brauchen ohnehin $\|E\| < \frac{1}{\|A^{-1}\|}$)

$\rightsquigarrow$ Nenner im Vorfaktor ist ≈ 1

- Erster Term vergleichbar mit dem aus der Konditionsanalyse
- Zweiter Term ist möglicherweise problematisch.

$\hat{L}$ enthält Einträge der Form $\frac{\tilde{a}_{ij}}{\tilde{a}_{ii}}$, also $\frac{1}{\text{Pivotelement}}$.

|Pivotelement| klein $\Rightarrow$ $|\hat{L}|$ groß $\Rightarrow$ großer Rundungsfehler!

- Dies kann trotz guter Kondition von A passieren!

Beispiel: $A = \begin{bmatrix} \varepsilon & 1 \\ 1 & 0 \end{bmatrix}$

$\Rightarrow$ $\boxed{\text{Gauß-Elimination (LR-Zerlegung) ist in dieser Form } \underline{\text{nicht}} \text{ numerisch stabil!}}$

Satz 4.13

Seien $\hat{L}$ und $\hat{R}$ die numerisch berechnete LR-Zerlegung von $A \in \mathbb{F}^{n\times n}$ aus Satz 4.12. Sei weiter $\hat{y} \in \mathbb{F}^n$ die numerische Lösung von $\hat{L} y = b$ und schließlich $\hat{x} \in \mathbb{F}^n$ die numerische Lösung von $\hat{R} x = \hat{y}$. Dann gilt für $\hat{x}$ die Beziehung

$$(A+E)\hat{x} = b$$

mit

$$|E| \le n\,eps\,\big(3|A| + 5|\hat{L}||\hat{R}|\big) + O(eps^2).$$

Beweis:

Wg Satz 4.11 gilt

$$(\hat{L}+F)\hat{y} = b \qquad |F| \le n\,eps\,|\hat{L}| + O(eps^2)$$
$$(\hat{R}+G)\hat{x} = \hat{y} \qquad |G| \le n\,eps\,|\hat{R}| + O(eps^2)$$

Einsetzen liefert:

$$(\hat{L}+F)\hat{y} = (\hat{L}+F)(\hat{R}+G)\hat{x} = (\underbrace{\hat{L}\hat{R}}_{=A+H} + F\hat{R} + \hat{L}G + FG)\hat{x} = b$$

Wg Satz 4.12 gilt

$$\hat{L}\hat{R} = A + H \qquad |H| \le 3(n-1)\,eps\,\big(|A| + |\hat{L}||\hat{R}|\big) + O(eps^2)$$

also

$$(A+E)\hat{x} = b \quad \text{mit} \quad E = H + F\hat{R} + \hat{L}G + FG$$

und

$$|E| \le |H| + |F||\hat{R}| + |\hat{L}||G| + O(eps^2)$$

$$\le \underbrace{3(n-1)\,eps\,\big(|A| + |\hat{L}||\hat{R}|\big)}_{|H|} + \underbrace{n\,eps\,|\hat{L}|}_{|F|}|\hat{R}| + |\hat{L}|\underbrace{n\,eps\,|\hat{R}|}_{|G|} + O(eps^2)$$

$$\le 3n\,eps\,|A| + 5n\,eps\,|\hat{L}||\hat{R}| + O(eps^2). \qquad \blacksquare$$

$$\iff |e_{ij}| \le n\,eps\,\Big(3|a_{ij}| + 5\sum_{k=1}^{n} |l_{ik}||r_{kj}|\Big) + O(eps^2)$$

4.5 Pivotisierung

Die Rundungsfehleranalyse in Satz 4.13 führt auf den unvorteilhaften Term $|\hat{L}||\hat{R}|$.

Mit der Wahl von r im Gauß-Algorithmus so dass

$$|a_{rk}^{(k)}| \geq |a_{ik}^{(k)}| \quad \forall\; k \leq i \leq n$$

passt schon, aber $k < n$

gilt dann

$$|\hat{l}_{ij}| \leq 1 \quad \text{und damit} \quad \|L\|_\infty \leq n.$$

Diese Wahl nennt man „Spaltenpivotisierung" (oder maximales Spaltenpivot).

Beispiel 5.13 Aus [GvL] → Blatt

$$\begin{bmatrix} -10^{-5} & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

$$A^{-1} = \frac{-1}{2+10^{-5}} \begin{bmatrix} 1 & -1 \\ -2 & -10^{-5} \end{bmatrix}$$

In exakter Arithmetik führt das Gaußsche Verfahren nach Elimination von a_{21} auf

$$\begin{bmatrix} -10^{-5} & 1 \\ 0 & 1+2\cdot 10^5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2\cdot 10^5 \end{bmatrix}$$

mit der Lösung

$$x_1 = -0.4999975, \qquad x_2 = 0.999995 \quad .$$

Nun führen wir das Verfahren in $\mathbb{F}(10,4,1)$ durch. Beim Multiplikator (Normierung)

$$q_{21} = (0.2\cdot 10^1) \oslash (-0.1\cdot 10^{-4}) = -0.2\cdot 10^6$$

ergibt sich kein Rundungsfehler.

Für das neue a_{22} ergibt sich

$$\left[\begin{array}{cc|c} -0.1\cdot 10^{-4} & 1 & 1 \\ & 0{,}2\cdot 10^6 & 0{,}2\cdot 10^6 \end{array}\right]$$

$$a_{22}^{(1)} = \overbrace{0.1\cdot 10^1}^{a_{22}} \ominus (\overbrace{-0.2\cdot 10^6}^{q_{21}}) \odot (\overbrace{0.1\cdot 10^1}^{a_{12}})$$
$$= 0.1\cdot 10^1 \oplus 0.2\cdot 10^6 = \boxed{0.2\cdot 10^6}.$$

$0.2000 \cdot 10^6$
$0.000001 \cdot 10^6$

Hier wurde auf vier Stellen gerundet.

$0.1\cdot 10^{-5}$ $\underbrace{10^5\cdot 10^1}_{10^6}$

Damit ergibt sich (ohne Fehler)

$$b_2^{(1)} = 0 - (-0.2\cdot 10^6) \odot (0.1\cdot 10^1) = 0.2\cdot 10^6$$

und

$$x_2 = b_2^{(1)} \oslash a_{22}^{(1)} = 0.2\cdot 10^6 \oslash 0.2\cdot 10^6 = \boxed{1}, \quad \text{(statt 0.999995)}$$
$$x_1 = (0.1\cdot 10^1 \ominus 0.1\cdot 10^1 \odot 1) \oslash (-0.1\cdot 10^{-4}) = \boxed{0} \quad .$$

2
16.11.09

Es ist also *keine* Stelle im Ergebnis korrekt obwohl nur an einer *einzigen* Stelle (in der Berechnung von $a_{22}^{(1)}$) ein Rundungsfehler eingeführt wurde.

Darüberhinaus überprüfe man, dass für die Kondition von A gilt:

$$\kappa(A) = 3 \quad .$$

Demnach ist das System gut konditioniert! Der Algorithmus, so wie er ist, ist numerisch nicht stabil.

Das Problem ist offensichtlich der große Multiplikator q_{21} der aus dem sehr kleinen a_{11} resultiert und der dafür sorgt, dass das ursprüngliche a_{22} in $a_{22}^{(1)}$ vollkommen ignoriert wird.

Im Prinzip haben wir in Fließkommaarithmetik das System

$$\begin{bmatrix} -10^{-5} & 1 \\ 2 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

exakt gelöst, was eine völlig andere Lösung hat als das ursprüngliche (11.1) (Rückwärtsanalyse).

Der große Multiplikator kann ganz einfach vermieden werden indem man eine Zeilenvertauschung durchführt, d. h. wir lösen

$$\begin{bmatrix} 2 & 1 \\ -10^{-5} & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{11.2}$$

Nun erhält man

$$\begin{aligned}
q_{21} &= -0.1 \cdot 10^{-4} \oslash 0.2 \cdot 10^1 = -0.5 \cdot 10^{-5}, \\
a_{22}^{(1)} &= 0.1 \cdot 10^1 \ominus (-0.5 \cdot 10^{-5}) \odot 0.1 \cdot 10^1 = 0.1 \cdot 10^1 \oplus 0.5 \cdot 10^{-5} = 0.1 \cdot 10^1, \\
b_2^{(1)} &= 0.1 \cdot 10^1 \ominus 0.5 \cdot 10^{-5} \odot 0 = 0.1 \cdot 10^1, \\
x_2 &= 0.1 \cdot 10^1 \oslash 0.1 \cdot 10^1 = \boxed{1}, \\
x_1 &= (0 \ominus 0.1 \cdot 10^1 \odot 0.1 \cdot 10^1) \oslash 0.2 \cdot 10^1 = \boxed{-0.5},
\end{aligned}$$

was in $\mathbb{F}(10,4,1)$ völlig in Ordnung ist. □

Spaltenpivotisierung ist nicht ausreichend wie folgendes Beispiel zeigt.

Forts. von Beispiel S. 13

(ebenfalls aus [GO96]). Wir betrachten das 2×2 System

$$\begin{bmatrix} 10 & -10^6 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -10^6 \\ 0 \end{bmatrix}.$$

welches aus (11.1) durch Multiplikation der ersten Zeile mit -10^6 entsteht.

Die Spaltenpivotisierung erfordert keine Vertauschung. Allerdings entsteht für $a_{22}^{(1)} = 1 + 2 \cdot 10^5$ genau dasselbe Problem wie oben! □

Diese Probleme kann man durch eine Skalierung (Äquilibrierung) des Gleichungssystems vermindern:

$$Ax = b \rightarrow D^{-1}Ax = D^{-1}b \quad \text{mit } d_{ii} = \sum_{j=1}^{n} |a_{ij}|$$
$$\Leftrightarrow \tilde{A}x = \tilde{b} \quad (\text{also } \|\tilde{A}\|_\infty = 1)$$

Rundungsfehleranalyse bei Pivotisierung

ja

Analog zu Satz 4.13 zeigt man, dass für die Lösung $\hat{x}$ bei Spaltenpivotisierung gilt: (Referenz GvL)

$$(A+E)\hat{x} = b \quad \text{mit } |E| \leq n \, \text{eps} \left(3|A| + 5 P^T |\hat{L}||\hat{R}|\right) + O(\text{eps}^2).$$

Nach Konstruktion ist $\|L\|_\infty \leq n$ und mit der Definition

$$\rho = \max_{i,j,k} \frac{|\hat{a}_{ij}^{(k)}|}{\|A\|_\infty} \quad \text{„Wachstumsfaktor"}$$

zeigt man [GvL]

$$\|E\|_\infty \leq 8 n^3 \|A_\infty\| \rho \, \text{eps} + O(\text{eps}^2).$$

In der Praxis ist $\rho \approx 10$

schlechtester Fall bei Spaltenpivotisierung ist $\rho = 2^{n-1}$.

Mit totaler Pivotisierung erreicht man

$$|a_{ij}^{(k)}| \leq k^{1/2} \left(2 \cdot 3^{1/2} \cdots k^{1/k-1}\right)^{1/2} \max |a_{ij}|$$

also deutlich kleineres Wachstum.

Totale Pivotisierung

Wähle $r, s \in \{k, \dots, n\}$ so dass

$$|a_{rs}^{(k)}| \geq |a_{ij}^{(k)}| \quad \forall\ k \leq i,j \leq n$$

und erreiche durch <u>Zeilen und Spaltenvertauschung</u>, dass

$$\tilde{a}_{kk}^{(k)} = a_{rs}^{(k)}.$$

In Matrixform:

(sollten wir schon haben)

Rechtstransformation, Umbenennung der Variablen.

Schritt 1: $$G_1 P_{r_1} A P_{s_1} \underbrace{P_{s_1} x} = G_1 P_{r_1} b$$

Schritt 2: $$G_2 P_{r_2} G_1 P_{r_1} A P_{s_1} P_{s_2} P_{s_2} P_{s_1} x = G_2 P_{r_2} G_1 P_{r_1} b$$

$$\vdots$$

Schritt n und Umformulierung

$$\underbrace{\underbrace{G'_n \dots G'_1}\underbrace{P_{r_n} \dots P_{r_1}}_{P} A \underbrace{P_{s_1} \dots P_{s_n}}_{Q^T}}_{=R} \overbrace{\underbrace{P_{s_n} \dots P_{s_1} x}_{=:z}}^{Q} = G'_n \dots G'_1 P_{r_n} \dots P_{r_1} b$$

Und damit

$$PAQ^T z = Pb$$

$$\boxed{PAQ^T = LR} \qquad z = Qx$$

Lösen des LGS gelingt dann mit $\qquad LRz = Pb$

$$b' = Pb$$
$$Ly = b'$$
$$Rz = y$$
$$x = Q^T z \qquad (Q \text{ ist orthogonal})$$

Aufwand der Pivotisierung:

- $n^2/2$ Vergleiche bei Spaltenpivotisierung

- $n^3/3$ Vergleiche bei totaler Pivotisierung

Da Speicherzugriffe teurer als eigentliche Rechnung $\Rightarrow$ Verdopplung des Zeitbedarfs bei totaler Pivotisierung.

Praktische Erfahrung zeigt keine Vorteile für Rundungsfehler bei totaler Pivotisierung

$\Rightarrow$ Spaltenpivotisierung mit Zeilenskalierung ist effizient und numerisch stabil in der Praxis.

4.6 Spezielle Systeme

Symmetrisch positiv definite Matrizen

<u>Satz 4.15</u> Eine symmetrisch positiv definite Matrix $A \in \mathbb{R}^{n \times n}$ ist stets <u>ohne</u> Pivotisierung stabil LR-zerlegbar. Für die Diagonalelemente der im Eliminationsprozess auftretenden Matrizen gilt

$$a_{ii}^{(k)} \geq \lambda_{min}(A), \qquad k \leq i \leq n.$$

<u>Beweis:</u> Betrachte einen Schritt in der LR-Zerlegung

(a).

$$A = \begin{array}{cc|l} \left[\begin{array}{cc} \alpha & v^T \\ v & B \end{array}\right] & & \begin{array}{l} 1 \\ n-1 \end{array} \\ \begin{array}{cc} 1 & n-1 \end{array} & & \end{array}$$

Elimination der Spalte v liefert

$$\begin{bmatrix} 1 & 0 \\ -v/\alpha & I \end{bmatrix} \underbrace{\begin{bmatrix} \alpha & v^T \\ v & B \end{bmatrix}}_{A} = \begin{bmatrix} \alpha & v^T \\ 0 & B - \frac{1}{\alpha} v v^T \end{bmatrix}$$

$$= \begin{bmatrix} \alpha & 0 \\ 0 & B - \frac{1}{\alpha} v v^T \end{bmatrix} \begin{bmatrix} 1 & v^T/\alpha \\ 0 & I \end{bmatrix}$$

Mit $\begin{bmatrix} 1 & v^T/\alpha \\ 0 & I \end{bmatrix}^{-1} = \begin{bmatrix} 1 & -v^T/\alpha \\ 0 & I \end{bmatrix}$ folgt:

$$\underbrace{\begin{bmatrix} 1 & 0 \\ -v/\alpha & I \end{bmatrix}}_{X^T} \underbrace{\begin{bmatrix} \alpha & v^T \\ v & B \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} 1 & -v^T/\alpha \\ 0 & I \end{bmatrix}}_{X} = \underbrace{\begin{bmatrix} \alpha & 0 \\ 0 & B - \frac{1}{\alpha} v v^T \end{bmatrix}}_{\tilde{A}}$$

X hat vollen Rang, A ist s.p.d. nach Vor. $\Rightarrow X^T A X = \tilde{A}$ ist s.p.d.

$B - \frac{1}{\alpha} v v^T$ ist Hauptuntermatrix von $\tilde{A}$ und somit auch sym. pos. definit.

bitte 5.6.13

7
16.11.09

b) Es gilt (Raleigh-Quotient):

$$(Ax,x)_2 \geq \lambda_{min}(A)\,(x,x)_2$$

und damit für $x = e^{(i)}$ (kartesischer Einheitsvektor)

$$a_{ii} = (Ae^{(i)}, e^{(i)})_2 \geq \lambda_{min}(A)\,\underbrace{(e^{(i)}, e^{(i)})_2}_{=1} = \lambda_{min}(A).$$

Für die Diagonalelemente von $B - \frac{1}{\alpha}vv^T$ gilt mit $\tilde{e}^{(i)} = \begin{pmatrix} 0 \\ e^{(i)} \end{pmatrix}$

$$\left(B - \frac{1}{\alpha}vv^T\right)_{ii} = \left(\tilde{A}\tilde{e}^{(i)}, \tilde{e}^{(i)}\right)_2 = \left(X^TAX\tilde{e}^{(i)}, \tilde{e}^{(i)}\right)_2$$

$$= \left(AX\tilde{e}^{(i)}, X\tilde{e}^{(i)}\right)_2$$

$$\geq \lambda_{min}(A)\left(X\tilde{e}^{(i)}, X\tilde{e}^{(i)}\right)_2 = \lambda_{min}(A)\left(1 + \underbrace{\frac{v_i^2}{\alpha^2}}_{\geq 0}\right)$$

$$\underbrace{\begin{bmatrix} 1 & -v^T/\alpha \\ 0 & I \end{bmatrix}}_{X}\underbrace{\begin{bmatrix} 0 \\ e^{(i)} \end{bmatrix}}_{\tilde{e}^{(i)}} = \begin{bmatrix} -\frac{v_i}{\alpha} \\ e^{(i)} \end{bmatrix} \qquad \geq \lambda_{min}(A).$$

$\boxtimes$

Dies zeigt, dass die Pivotelemente bei der Gaußelimination echt nach unten beschränkt bleiben.

Ü) Rundungsfehleranalyse. Kann man $|\hat{C}|\,|\hat{R}|$ abschätzen?
⇝ Rannacher Lemma 4.1.

Cholesky-Zerlegung

P
16.11.09

Die Pivotelemente die bei der LR-Zerlegung auftreten sind nach 4.15 stets positiv.

Mit $D = \text{diag}(R_i)$ gilt

$$A = L D \underbrace{D^{-1} R}_{U} = L D U$$

U ist obere Dreiecksmatrix mit $u_{ii} = 1$. Wegen der Symmetrie gilt $U = L^T$, also

$$A = L D L^T.$$

Da $d_{ii} > 0$ ist die Matrix „$D^{1/2}$"

$$(D^{1/2})_{ii} = \sqrt{d_{ii}}$$

wohldefiniert und es gilt

$$A = \underbrace{L D^{1/2}}_{=:\tilde{L}} \underbrace{D^{1/2} L^T} = \tilde{L} \tilde{L}^T.$$

Dies ist die sog. Cholesky-Zerlegung einer symmetrisch positiv definiten Matrix.

Ausnutzung der Symmetrie erlaubt die Berechnung der Cholesky-Zerlegung in $n^3/3 + O(n^2)$ Operationen (Faktor 2 schneller).

Diagonaldominante Matrizen

Definition 4.16 Eine Matrix $A \in \mathbb{R}^{n\times n}$ heißt diagonaldominant falls

$$\sum_{j=1, j\neq i}^{n} |a_{ij}| \leq |a_{ii}| \qquad i = 1,\dots,n.$$

Satz 4.17 Diagonaldominante, reguläre Matrizen erlauben eine LR-Zerlegung ohne Pivotisierung.

Beweis. $A = \begin{bmatrix} \alpha & w^T \\ v & B \end{bmatrix}$, ein Schritt der GEM liefert

$$\begin{bmatrix} 1 & 0 \\ -v/\alpha & I \end{bmatrix} \begin{bmatrix} \alpha & w^T \\ v & B \end{bmatrix} = \begin{bmatrix} \alpha & w^T \\ 0 & B - \frac{1}{\alpha} v w^T \end{bmatrix}$$

Da $\alpha \neq 0$ wg. Diagonaldominanz ist dies wohldefiniert.

Aus der Diagonaldominanz von A ergibt sich:

(Zeile 1) $\displaystyle\sum_{j=1}^{n-1} |w_j| \leq |\alpha| \iff \sum_{j=1}^{n-1} \frac{|w_j|}{|\alpha|} \leq 1$

(Zeile 2…n) $\displaystyle |v_i| + \sum_{j=1, j\neq i}^{n-1} |b_{ij}| \leq |b_{ii}|$

Für $B - \frac{1}{\alpha} v w^T$ rechnen wir dann nach:

$$\sum_{j=1, j\neq i}^{n-1} \left| b_{ij} - \frac{1}{\alpha} v_i w_j \right| + \left| \frac{v_i w_i}{\alpha} \right| \leq \sum_{j=1, j\neq i}^{n-1} |b_{ij}| + |v_i| \sum_{j=1, j\neq i}^{n-1} \frac{|w_j|}{|\alpha|} + |v_i| \frac{|w_i|}{|\alpha|}$$

$$= \sum_{j=1, j\neq i}^{n-1} |b_{ij}| + |v_i| \underbrace{\sum_{j=1}^{n-1} \frac{|w_j|}{|\alpha|}}_{\leq 1} \leq |v_i| + \sum_{j=1, j\neq i}^{n-1} |b_{ij}| \leq |b_{ii}|$$

$$\iff \sum_{j=1, j\neq i}^{n-1} \left| b_{ij} - \frac{1}{\alpha} v_i w_j \right| \leq |b_{ii}| - \left| \frac{v_i w_i}{\alpha} \right| \leq \left| b_{ii} - \frac{v_i w_i}{\alpha} \right| = \left| \left(B - \frac{1}{\alpha} v w^T\right)_{ii} \right|$$

$|x| = |x-y+y| \leq |x-y| + |y| \iff |x| - |y| \leq |x-y|$ ■

Somit ist $B - \frac{1}{\alpha} v w^T$ wieder diagonaldominant.

Rangbestimmung

Sei $A \in \mathbb{K}^{n \times n}$. Gilt <u>nach</u> k-Schritten der Gauß-Elimination

$$A^{(k)} = \begin{array}{c|c} \overbrace{\qquad}^{k} & \overbrace{\qquad}^{n-k} & \\ \left[\begin{array}{c|c} U^{(k)} & * \\ \hline 0 & 0 \end{array}\right] & & \begin{array}{c} k \\ n-k \end{array} \end{array}$$

mit $u_{ii}^{(k)} \neq 0$ so ist $\mathrm{Rang}(A) = k$. ... Somit

- Kann die GE zu Ende geführt werden so gilt $k = n$, mithin $\mathrm{Rang}(A) = n$.
- Rangbestimmung erfordert totales Pivoting, da die erste Spalte der $n-k \times n-k$ Restmatrix 0 sein kann, was aber <u>nicht</u> $\mathrm{Rang}(A) = k$ bedeutet.

Inversenberechnung

Zur Berechnung der Inversen geht man so vor:

- Berechne LR-Zerlegung von A. Aufwand $\frac{2}{3} n^3 + O(n^2)$ bei Spaltenpivotisierung
- Für $i = 1, \dots, n$ löse $A x^{(i)} = e^{(i)}$.
  Aufwand $n \cdot 2 n^2$ für Lösen der Dreieckssysteme
- $A^{-1} = [x^{(1)}, \dots, x^{(n)}]$ besteht spaltenweise aus den $x^{(i)}$
- Gesamtaufwand ist $\frac{8}{3} n^3 + O(n^2)$.

11
16.11.09

Tridiagonalsysteme

Definition 4.18

$A \in \mathbb{R}^{n\times n}$ heißt Tridiagonalmatrix falls

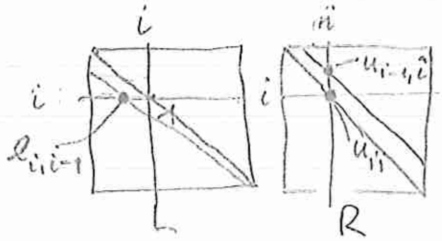
$$a_{ij} = 0 \quad \text{für } |i-j| > 1. \qquad 1 \le i \le n$$

Eine Tridiagonalmatrix ist Spezialfall einer Bandmatrix:

$$a_{ij} = 0 \quad \text{für } j < i - m_l \quad \text{oder} \quad j > i + m_r \qquad 1 \le i \le n$$

Die LR-Zerlegung einer Tridiagonalmatrix sei ohne Pivotisierung durchführbar. Dann gilt:

- L ist Tridiagonalmatrix
- R ist Tridiagonalmatrix



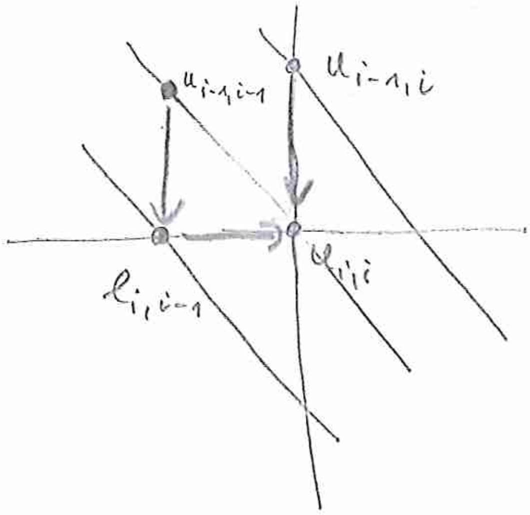
und damit:

Zeile i · Spalte i: $l_{i,i-1} \cdot r_{i-1,i} + r_{ii} = a_{ii} \Rightarrow r_{ii} = a_{ii} - l_{i,i-1} r_{i-1,i}$

$$l_{i,i-1} \cdot r_{i-1,i-1} = a_{i,i-1} \Rightarrow l_{i,i-1} = \frac{a_{i,i-1}}{u_{i-1,i-1}}$$

$$1 \cdot r_{i-1,i} = a_{i-1,i} \Rightarrow r_{i-1,i} = a_{i-1,i}$$

Nächstes Mal andere Reihenfolge!



$r_{11} = a_{11}$, ~~[illegible]~~

for ($i = 2$ to n) do

$r_{i-1,i} = a_{i-1,i}$

$l_{i,i-1} = \dfrac{a_{i,i-1}}{u_{i-1,i-1}}$

$r_{ii} = a_{ii} - l_{i,i-1} r_{i-1,i}$

Nichtreguläre Systeme

Es sei nun $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ sowie Rang(A) beliebig.
$Ax = b$ hat dann genau eine, unendlich viele oder gar keine Lösung.

Einige Grundbegriffe aus der linearen Algebra: $A : \mathbb{R}^n \to \mathbb{R}^m$ lineare Abb.

$$\mathrm{Bild}(A) = \{ y \in \mathbb{R}^m : y = Ax \text{ für ein } x \in \mathbb{R}^n \} \subseteq \mathbb{R}^m$$

$$\mathrm{Kern}(A) = \{ x \in \mathbb{R}^n : Ax = 0 \} \subseteq \mathbb{R}^n$$

$$\mathrm{Rang}(A) = \dim(\overset{\text{„Spalten"}}{\mathrm{Bild}(A)}) = \mathrm{Rang}(A^T) = \dim(\overset{\text{„Zeilenrang"}}{\mathrm{Bild}(A^T)})$$

(# l.u. Spalten = # l.u. Zeilen). nicht trivial → z.B. Beutelsp. S. 101

Orthogonales Komplement:

$$\mathrm{Bild}(A)^\perp = \{ y \in \mathbb{R}^m : (y, y')_2 = 0 \ \forall y' \in \mathrm{Bild}(A) \}$$

$$\begin{aligned}
& (y, y')_2 = 0 \ \forall y' \in \mathrm{Bild}(A) \\
\Longleftrightarrow\ & (y, Ax)_2 = 0 \ \forall x \in \mathbb{R}^n \\
\Longleftrightarrow\ & (A^T y)^T x = 0 \ \forall x \in \mathbb{R}^n \quad \text{geht für alle } x \text{ nur wenn } A^T y = 0 \\
\Longleftrightarrow\ & y \in \mathrm{Kern}(A^T)
\end{aligned}$$

und damit $\mathrm{Bild}(A)^\perp = \mathrm{Kern}(A^T)$.

Dies zeigt dann

$$\underbrace{\dim(\mathrm{Bild}(A))}_{= \mathrm{Rang}(A)} + \dim(\overbrace{\mathrm{Kern}(A^T)}^{\mathrm{Bild}(A)^\perp}) = m \qquad \text{\# Zeilen}$$

$$\overbrace{\dim(\mathrm{Bild}(A^T))} + \dim(\mathrm{Kern}(A)) = n$$

Den Lösungsbegriff für lineare Gleichungssysteme kann man auf die folgende Weise erweitern.

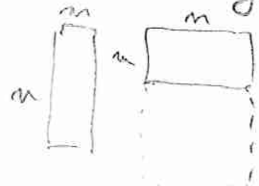
Satz 4.19 (Least Squares Lösung)

a) Es existiert ein $\bar{x} \in \mathbb{R}^n$ so dass

$$\|A\bar{x}-b\|_2 = \min_{x\in\mathbb{R}^n} \|Ax-b\|_2 .$$

b) Diese Bedingung ist äquivalent dazu dass $\bar{x}$ Lösung von

$$\underbrace{A^TA}_{n\times n}\bar{x} = A^Tb,$$



der sog. „Normalengleichung".

c) Falls $\text{Rang}(A)=n$ (dann ist zwingend $m\geq n$) ist $\bar{x}$ eindeutig bestimmt, sonst hat jede weitere Lösung die Form $\bar{x}+y$ mit $y\in \text{Kern}(A)$.

Beweis.

(b)$\Rightarrow$(a) $\bar{x}$ sei Lösung der Normalengleichung. Für ein beliebiges $x\in\mathbb{R}^n$ gilt:

$$\|Ax-b\|_2^2 = \|A(x-\bar{x}+\bar{x})-b\|_2^2 = \big(\underbrace{A\bar{x}-b}+\underbrace{A(x-\bar{x})}, \underbrace{A\bar{x}-b}+\underbrace{A(x-\bar{x})}\big)_2$$

$$= (A\bar{x}-b, A\bar{x}-b)_2 + 2(\underbrace{A\bar{x}-b}_{\in \text{Kern}(A^T) = \text{Bild}(A)^\perp}, \underbrace{A(x-\bar{x})}_{\in \text{Bild}(A)})_2 + (A(x-\bar{x}), A(x-\bar{x}))_2$$

da $A^T(A\bar{x}-b)=0$

$$= \|A\bar{x}-b\|_2^2 + \underbrace{\|A(x-\bar{x})\|_2^2}_{\geq 0}$$

$$\geq \|A\bar{x}-b\|_2^2$$

Damit erfüllt $\bar{x}$ die Minimalitätsbedingung.

(a)$\Rightarrow$(b) Setze $F:\mathbb{R}^n\to\mathbb{R}^1$, $F(x) = \|Ax-b\|_2^2$. Notwendige Bedingung für ein Minimum ist $\nabla F(\bar{x}) = 0 \Leftrightarrow \frac{\partial F}{\partial x_k}(\bar{x}) = 0 \quad \forall k=1,\dots,n$.

$$\frac{\partial F(\bar{x})}{\partial x_k} = \frac{\partial}{\partial x_k}(Ax-b, Ax-b)_2\Big|_{x=\bar{x}} = \frac{\partial}{\partial x_k}\left(\sum_{i=1}^m \Big[\underbrace{\sum_{j=1}^n a_{ij}x_j - b_i}_{i\text{-te Komponente}}\Big]^2\right)\Bigg|_{x=\bar{x}}$$

$$= \left(\sum_{i=1}^m 2\Big[\sum_{j=1}^n a_{ij}x_j - b_i\Big]\underbrace{a_{ik}}_{(j=k)}\right)\Bigg|_{x=\bar{x}} = 2\sum_{i=1}^m (A^T)_{ki}\Big[\underbrace{\sum_{j=1}^n a_{ij}\bar{x}_j - b_i}_{(A\bar{x}-b)_i}\Big]$$

Nachdifferenzieren.

$$= 2\big(A^T(A\bar{x}-b)\big)_k$$

Und damit $\nabla F(\bar{x}) = 2A^T(A\bar{x}-b) \overset{!}{=} 0 \Leftrightarrow A^TA\bar{x} = A^Tb$,

die Normalengleichung.

(c) Lösbarkeit der Normalengleichung.

$$\mathbb{R}^m = \text{Bild}(A) \oplus \text{Bild}(A)^\perp,$$

d.h. jedes $b \in \mathbb{R}^m$ besitzt eine <u>eindeutige</u> Zerlegung

$$b = s + r \quad \text{mit} \quad s \in \text{Bild}(A),\ r \in \text{Bild}(A)^\perp = \text{Kern}(A^T)$$

Da $s \in \text{Bild}(A)$ gibt es $\bar{x} \in \mathbb{R}^n$ mit $A\bar{x} = s$. Für dieses gilt dann

$$A^TA\bar{x} = A^Ts = A^Ts + \underbrace{A^Tr}_{\substack{=0 \text{ da} \\ r \in \text{Kern}(A^T)}} = A^Tb,$$

also löst dieses $\bar{x}$ auch die Normalengleichung.

<u>Sei Rang$(A) = n$.</u> und damit $m \geq n$ (wg. Rang$(A) \leq \min(m,n)$).

Betrachte $A^TAx = 0$ also Kern(A^TA).

Der alte, durchgestrichene Text bleibt gültig, der neue ist falsch ! (26.6.2013)

$(*)$ $A^T\underbrace{A x}_{y} = 0 \iff A^Ty = 0 \wedge y = Ax$

~~$y \in \text{Kern}(A^T) \iff y \in \text{Bild}(A)^\perp$. Da Rang$(A) = n$ ist Bild$(A)^\perp = \{0\}$~~

Nun ist ~~aber $\underbrace{\text{Kern}(A^T)}_{=\text{Bild}(A)^\perp} \perp \text{Bild}(A)$ d.h. Kern$(A^T) \cap \text{Bild}(A) = \{0\}$.~~

~~d.h. $(*)$ gilt nur für y~~ $= 0$. da Rang$(A) = n$ bedeutet dies $x = 0$ also A^TA regulär

(Alternative: A^TA ist symmetrisch und positiv definit wg. max. Rang).

<u>Sei Rang$(A) < n$.</u> Eine Lösung haben wir schon, nämlich $\bar{x}$. Sei x_1 eine weitere Lösung der Normalengleichung (also $A^TAx_1 = A^Tb$). Dann gilt

$$b = \underbrace{Ax_1}_{\in \text{Bild}(A)} + \underbrace{(b - Ax_1)}_{\in \text{Kern}(A^T),\text{ da } A^T(b-Ax_1) = A^Tb - A^TAx_1 \overset{x_1 \text{ Lsg der NG}}{=} 0}$$

Da die Zerlegung $\mathbb{R}^m = \text{Bild}(A) \oplus \text{Bild}(A)^\perp$ eindeutig ist muss

$\hookrightarrow$ d.h. $b = Ax_1 + (b - Ax_1) = A\bar{x} + (b - A\bar{x})$

$$Ax_1 = A\bar{x} \text{ sein}$$

Und das heißt $\overset{(\text{Ansatz})}{x_1 = \bar{x} + y}$ und

$$Ax_1 = A\bar{x} + Ay \overset{!}{=} A\bar{x} \iff \boxed{Ay = 0}. \quad \text{(was die Behauptung war)}$$

Bemerkung: A^TA ist symmetrisch und positiv semidefinit.

Lösung prinzipiell mit Cholesky-Zerlegung möglich.

QR-Zerlegung

Definition 4.20.

Sei $v \in \mathbb{R}^n$ gegeben, dann heißt die Matrix

$$Q_v = I - 2\frac{vv^T}{v^Tv} \quad \text{äußeres Produkt}$$

Householdermatrix oder Householderreflektion $\boxtimes$

Wir rechnen einige Eigenschaften dieser Matrix nach:

1) Q_v ist orthogonal, d.h. $Q_v^T Q_v = I$, $Q^T = Q$

$$Q^T = I - 2\frac{vv^T}{v^Tv} = I - \frac{2}{v^Tv}(vv^T)^T = I - \frac{2}{v^Tv}vv^T = Q$$

$$Q^TQ = Q^2 = \left(I - 2\frac{vv^T}{v^Tv}\right)\left(I - \frac{2vv^T}{v^Tv}\right) = I - 4\frac{vv^T}{v^Tv} + 4\frac{(vv^T)(vv^T)}{v^Tv\, v^Tv}$$

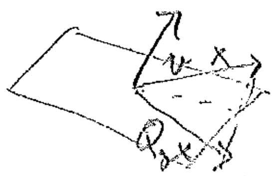
$$= I - 4\frac{vv^T}{v^Tv} + 4\frac{v\cancel{(v^Tv)}v^T}{(v^Tv)\cancel{(v^Tv)}} = I.$$

2) $$Q_v x = x - 2\frac{vv^T}{v^Tv}x = x - 2\frac{v^Tx}{v^Tv}v$$

$$\big((vv^T)x\big)_i = \sum_{j=1}^n v_i v_j x_j = (v^Tx)\,v_i$$

3) $x = \alpha v \Rightarrow Q_v x = -x$ „Reflektion an der Ebene mit Normale v"

$$Q_v x = \alpha v - 2\alpha\frac{\cancel{v^Tv}}{\cancel{v^Tv}}v = -\alpha v = -x$$



4) $(x,v)_2 = v^Tx = 0 \Rightarrow Q_v x = x$

$$Q_v x = x - 2\frac{\overbrace{v^Tx}^{=0}}{v^Tv}v = x$$

5) zusammen: $x = \alpha v + w$ mit $(w,v)_2 = 0 \Rightarrow Q_v x = -\alpha v + w$

Q_v kann in ähnlicher Weise wie die Frobenius-matrizen benutzt werden um eine Matrix auf obere Dreiecksgestalt zu transformieren.

Problem. Wähle v so, dass

$$Q_v A = \begin{pmatrix} \alpha & * & \cdots & * \\ 0 & | & & | \\ \vdots & & & \vdots \\ 0 & * & \cdots & * \end{pmatrix}$$

Sei x die erste Spalte von A so bestimme v so dass

$$Q_v x = \beta e_1$$

$$\Leftrightarrow x - 2\frac{v^T x}{v^T v} v = \beta e_1$$

Mit dem Ansatz $v = x + \alpha e_1$ erhalten wir

$$Q_v x = x - 2\frac{v^T x}{v^T v}(x + \alpha e_1) = \left(1 - 2\frac{v^T x}{v^T v}\right) x - 2\alpha \frac{v^T x}{v^T v} e_1$$

Um den ersten Term zu eliminieren:

$$1 - 2\frac{v^T x}{v^T v} = 1 - 2\frac{(x+\alpha e_1)^T x}{(x+\alpha e_1)^T (x+\alpha e_1)}$$

$$= 1 - 2\frac{x^T x + \alpha x_1}{x^T x + 2\alpha x^T e_1 + \alpha^2}$$

$$= \frac{x^T x + 2\alpha x_1 + \alpha^2 - 2x^T x - 2\alpha x_1}{x^T x + 2\alpha x_1 + \alpha^2} = \frac{-x^T x + \alpha^2}{x^T x + 2\alpha x_1 + \alpha^2} \overset{!}{=} 0$$

hier.

$$\Leftrightarrow \alpha^2 = x^T x = \|x\|_2^2 \Leftrightarrow \boxed{\alpha = \pm \|x\|_2}$$

d.h. für $v = x \pm \|x\|_2 e_1$ gilt $Q_v x = \mp 2\|x\|_2 e_1$

Welches Vorzeichen wählt man? Falls schon $x = \gamma e_1$ gilt, dann wäre $v = \gamma e_1 \pm |\gamma| e_1$, damit die Norm von v möglichst groß wird wählt man $+$ für $\gamma > 0$ und $-$ für $\gamma < 0$.

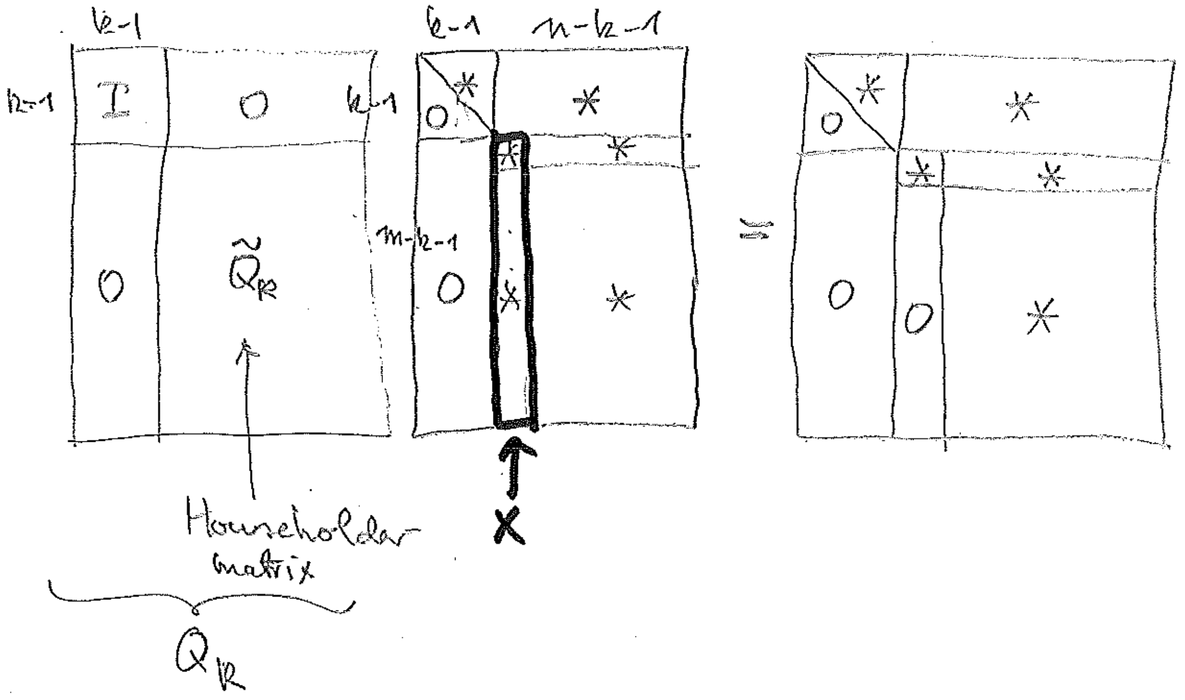
Satz 4.21 (QR-Zerlegung)

Zu jeder Matrix $A \in \mathbb{R}^{m \times n}$ mit $m \geq n$ und $\text{Rang}(A) = n$ existiert eine orthogonale Matrix $Q \in \mathbb{R}^{m \times m}$ und eine obere Dreiecksmatrix $R \in \mathbb{R}^{m \times n}$ so dass

$$A = QR.$$

Die ersten n Spalten von Q bilden eine orthonormale Basis von A.

__Beweis.__ Im Schritt $k = 1, \dots,$



- Q_k ist orthogonal, da $\tilde{Q}_k$ Householdermatrix. Somit gilt nach $n-1$ Schritten:

$$Q_{n-1} \dots Q_1 A = R$$

da $Q_k^{-1} = Q_k^T = Q_k$ gilt

$$A = \underbrace{Q_1 \dots Q_{n-1}}_{=: Q} R = QR$$

Q ist orthogonal, da $Q^T = (Q_1 \cdots Q_{n-1})^T = Q_{n-1}^T \cdots Q_1^T = Q_{n-1} \cdots Q_1$ und $Q^T Q = Q_{n-1} \cdots Q_1 Q_1 \cdots Q_{n-1} = I$.

$\boxtimes$

4
12.6.13

Anwendung zur Lösung von $A^T A x = A^T b$.

$A = QR \Rightarrow A^T A = (QR)^T QR = R^T \underbrace{Q^T Q}_{=I} R = R^T R$

also $R^T \underbrace{R x}_{=y} = A^T b$

Somit 1) $R^T y = A^T b$

2) $R x = y$

$\Rightarrow$ - $A^T A$ muss nicht explizit aufgestellt werden (n^3 Op, Fill in).

- Vorteile bei Rundungsfehlerfortpflanzung

- Anwendung bei Berechnung von Eigenwerten (QR-Iteration)

Anwendung: Gaußsche Ausgleichsrechnung.

Gegeben:

(i) n Funktionen $u_1, \dots, u_n : \mathbb{R} \to \mathbb{R}$, sowie

(ii) m Datenpunkte $(x_i, y_i) \in \mathbb{R}^2$, $1 \le i \le m \geq n$

← Bei = ungl. Eindeutig.

Gesucht: n Koeffizienten $c_1, \dots, c_n$ so dass

$$u(x) = \sum_{j=1}^{\cancel{m}\, n} c_j u_j(x)$$

und

$$(*) \qquad \sum_{i=1}^{m} (u(x_i) - y_i)^2 \longrightarrow \text{minimal}.$$

Dies lässt sich folgendermaßen formulieren:

$$c = (c_1, \dots, c_n)^T$$

$$y = (y_1, \dots, y_m)^T$$

$$a_{ij} = u_j(x_i)$$

(Finde $c \in \mathbb{R}^n$ so dass $\|Ac - y\|_2^2$ minimal.

Denn $$\sum_{i=1}^{m} (u(x_i) - y_i)^2 = \sum_{i=1}^{m} \Big(\underbrace{\sum_{j=1}^{n} c_j u_j(x_i) - y_i}_{(Ac-y)_i} \Big)^2$$

$$= \|Ac - y\|_2^2 .$$

Somit sind nach Satz ~~5.17~~ 4.19 die gesuchten c_i Lösung der Normalengleichung

$$A^T A c = A^T y.$$

Lösung: z.B. mit Cholesky-Zerl.

Ü: Zeige $\mathrm{cond}_2(A^T A) = \mathrm{cond}_2(A)^2$.