Computergestützte Mathematik zur Linearen Algebra

Pivotwahl bei Rundungsfehlern

Achim Schädle

Übungsleiter: Lennart Jansen

Tutoren: Marina Fischer, Kerstin Ignatzy, Narin Konar
Pascal Kuhn, Nils Sänger, Tran Dinh

8. Januar 2015

Instabilitäten bei Gauß-Elimination

Aufgabe: Berechne LR-Zerlegung von

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

LR-Zerlegung von A existiert nicht (Division durch 0)

$$A = \begin{bmatrix} 10^{-20} & 1 \\ 1 & 1 \end{bmatrix}$$

Pivotelement $10^{-20} \neq 0$, LR-Zerlegung von A existiert:

$$L = \begin{bmatrix} 1 & 0 \\ 10^{20} & 1 \end{bmatrix}, \qquad R = \begin{bmatrix} 10^{-20} & 1 \\ 0 & 1 - 10^{20} \end{bmatrix}$$

MATLAB liefert jedoch bei der Multiplikation $L \cdot R \neq A$

Gleitkommaarithmetik

Wie lässt sich dieser Fehler erklären?

- Wie werden Zahlen im Rechner dargestellt?
- Welche Fehler können bei Grundrechenarten passieren?
- Wie kann man Instabilitäten im Gauß-Algorithmus vermeiden?

Gleitpunktdarstellung reeller Zahlen

$x \in \mathbb{R}$, $x \neq 0$ kann eindeutig durch

$$x = \pm a \cdot 10^b$$

dargestellt werden

- a Mantisse, $0.1 \leq a < 1$
- b Exponent

Einschränkungen durch den Rechner

- nur ℓ Ziffern für die Mantisse
- nur ℓ' Ziffern für den Exponenten

Zur Vereinfachung: $\ell' = \infty$

Gleitpunktdarstellung reeller Zahlen

Sei $\overline{a}$ auf ℓ Ziffern gerundete Mantisse von $x = \pm a \cdot 10^b$

$$\mathrm{fl}(x) := \pm\overline{a} \cdot 10^b$$

Beispiel: $\ell = 8$, $x = \pi = 3.141592653\ldots$

$$\mathrm{fl}(\pi) = 0.31415927 \cdot 10^1$$

Maschinengenauigkeit: Kleinste positive Zahl eps, so dass

$$\mathrm{fl}(1 + \mathrm{eps}) > 1.$$

- Dezimalsystem $\quad \mathrm{eps} = 5 \cdot 10^{-\ell}$
- Binärsystem (Basis 2) $\quad \mathrm{eps} = 2^{-\ell}$

IEEE Standard 754 - 1985: Gleitkommazahlen

Standardisierung der Darstellung im Rechner durch IEEE

- single precision 32 bit (4 Byte)
- double precision 64 bit (8 Byte)

Genauigkeit	Vorzeichen s	Exponent e	Mantisse m
einfach = single	$1(31)$	$8(30-23)$	$23(22-0)$
doppelt = double	$1(63)$	$11(62-52)$	$52(51-0)$

Normalisierte Zahlendarstellung:

$$s \cdot 1.m \cdot 2^{e-bias}, \quad bias = 127 \text{ bzw. } 1023$$

Damit ergeben sich die Maschinengenauigkeiten

- (single) $\mathrm{eps} = 2^{-23} \approx 1.2 \cdot 10^{-7}$
- (double) $\mathrm{eps} = 2^{-52} \approx 2.2 \cdot 10^{-16}$

In **MATLAB** wird standardmäßig doppelt genau gerechnet

Relative Fehler

Satz: Für jedes $x \neq 0$ ist

$$|\operatorname{fl}(x) - x| \leq \mathtt{eps}|x|,$$

das heißt der relative Fehler ist beschränkt durch $\mathtt{eps}$.

Beweis: Sei $x = a \cdot 10^b$ und $\operatorname{fl}(x) = \overline{a} \cdot 10^b$. Bei ℓ signifikanten Stellen ist $|a - \overline{a}| \leq 5 \cdot 10^{-l-1}$, also

$$\frac{|\operatorname{fl}(x) - x|}{|x|} = \frac{|\overline{a} - a| \cdot 10^b}{|a| \cdot 10^b} \leq \frac{5 \cdot 10^{-l-1}}{10^{-1}} = \mathtt{eps},$$

da $\overline{a} \geq 10^{-1}$.

Schreibe daher $\operatorname{fl}(x) = x(1 + \varepsilon)$ mit $|\varepsilon| \leq \mathtt{eps}$

Kondition eines Problems

Ein Problem sei durch Auswertung einer Abbildung

$$F : \mathbb{R}^n \to \mathbb{R}$$

beschrieben, wobei F zum Beispiel

- ein Polynom,
- eine rationale Funktion,
- die Lösung von $Ax = b$ oder
- die Lösung eines Eigenwertproblems $Ax = \lambda x$ ist.

Frage: Wie wirken sich Störungen von $x = (x_1, \ldots, x_n)$ auf das Resultat $F(x)$ aus?

Kondition eines Problems

Definition: Die Kondition κ von F ist die kleinste Zahl, so dass

$$\frac{|\hat{x}_i - x_i|}{|x_i|} \leq \texttt{eps}, \quad \forall\, i \Longrightarrow \frac{|F(\hat{x}) - F(x)|}{|F(x)|} \leq \kappa \cdot \texttt{eps}.$$

Das Problem heißt gut konditioniert, falls κ nicht zu groß ist (ideal $\kappa = 1$) und andernfalls schlecht konditioniert.

Die Kondition kann näherungsweise durch Linearisierung mithilfe der Ableitung bestimmt werden.

Kondition der Grundrechenarten

Multiplikation zweier reeller Zahlen

Sei $F(x_1, x_2) = x_1 \cdot x_2$. Für die gestörten Werte

$$\hat{x}_1 = x_1(1+\varepsilon_1), \qquad \hat{x}_2 = x_2(1+\varepsilon_2), \qquad |\varepsilon_i| \leq \texttt{eps}$$

erhalten wir für $\dfrac{|F(\hat{x}) - F(x)|}{|F(x)|}$

$$\frac{\hat{x}_1\hat{x}_2 - x_1x_2}{x_1x_2} = (1+\varepsilon_1)(1+\varepsilon_2) - 1 = \varepsilon_1 + \varepsilon_2 + \varepsilon_1\varepsilon_2.$$

Da eps sehr klein, vernachlässige $\varepsilon_1\varepsilon_2$ und erhalte

$$\left|\frac{\hat{x}_1\hat{x}_2 - x_1x_2}{x_1x_2}\right| \leq 2\texttt{eps}.$$

Also $\kappa(F) = 2$, die Multiplikation ist gut konditioniert!

Kondition der Grundrechenarten

Subtraktion zweier reeller Zahlen

Für $F(x_1, x_2) = x_1 - x_2$ erhalten wir

$$\begin{aligned} \left| \frac{(\hat{x}_1 - \hat{x}_2) - (x_1 - x_2)}{x_1 - x_2} \right| &= \left| \frac{x_1 \varepsilon_1 - x_2 \varepsilon_2}{x_1 - x_2} \right| \\ &\leq \frac{|x_1| + |x_2|}{|x_1 - x_2|} \texttt{eps} =: \kappa \texttt{eps}. \end{aligned}$$

- Mit $\operatorname{sign} x_1 = -\operatorname{sign} x_2$ (Addition) ist $\kappa(F) = 1$.
  Die Addition ist gut konditioniert.
- Mit $\operatorname{sign} x_1 = \operatorname{sign} x_2$ (Subtraktion) und $x_1 \approx x_2$ ist $\kappa(F) >> 1$ sehr groß.
  Die Subtraktion zweier etwa gleich großer Zahlen ist sehr schlecht konditioniert (Auslöschung).

Gauss-Elimination

In double precision mit eps $= 10^{-16}$

$$A = \begin{bmatrix} 10^{-20} & 1 \\ 1 & 1 \end{bmatrix}$$

Pivotelement $10^{-20} \neq 0$, LR-Zerlegung von A existiert:

$$L = \begin{bmatrix} 1 & 0 \\ 10^{20} & 1 \end{bmatrix}, \qquad R = \begin{bmatrix} 10^{-20} & 1 \\ 0 & 1 - 10^{20} \end{bmatrix}$$

Jedoch wird das Ergebnis auf double precision gerundet!

$$\mathrm{fl}(L) = \tilde{L} = \begin{bmatrix} 1 & 0 \\ 10^{20} & 1 \end{bmatrix}, \qquad \mathrm{fl}(R) = \tilde{R} = \begin{bmatrix} 10^{-20} & 1 \\ 0 & -10^{20} \end{bmatrix}$$

und damit erhalten wir numerisch (vgl. **MATLAB**)

$$\tilde{L}\tilde{R} = \begin{bmatrix} 10^{-20} & 1 \\ 1 & 0 \end{bmatrix} \neq A = \begin{bmatrix} 10^{-20} & 1 \\ 1 & 1 \end{bmatrix}$$

Lösung eines Gleichungssystemes mit $\tilde{L}\tilde{R}$-Zerlegung

$$\mathrm{fl}(L) = \tilde{L} = \begin{bmatrix} 1 & 0 \\ 10^{20} & 1 \end{bmatrix}, \qquad \mathrm{fl}(R) = \tilde{R} = \begin{bmatrix} 10^{-20} & 1 \\ 0 & -10^{20} \end{bmatrix}$$

Löst man mit $\tilde{L}$ und $\tilde{R}$ das LGS $Ax = b$ mit $b = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, so ergibt sich aus $\tilde{L}\tilde{y} = b$ zunächst

$$\tilde{y} = \begin{bmatrix} 1 \\ -10^{20} \end{bmatrix} \qquad (\text{richtig})$$

und dann aus $\tilde{R}\tilde{x} = \tilde{y}$ die Lösung

$$\tilde{x} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \qquad (\text{statt richtig } x = \begin{bmatrix} -1 \\ 1 \end{bmatrix})$$

Spaltenpivotsuche

Problem: L-Faktor enthält großes Element.

Erinnerung: $k-1$ Schritte Gauß-Elimination liefern

$$A \to A_{k-1} = \begin{bmatrix} \times & \cdots & \times & \alpha_1 & \times & \cdots & \times \\ & \ddots & & \vdots & \vdots & & \vdots \\ & & \times & \alpha_{k-1} & \times & \cdots & \times \\ & & & \alpha_k & \times & \cdots & \times \\ & & & \vdots & \vdots & & \vdots \\ & & & \alpha_n & \times & \cdots & \times \end{bmatrix}$$

Spaltenpivotsuche: Wählt man $|\alpha_j| = \max\limits_{k \le i \le n} |\alpha_i|$ als Pivotelement, gilt $|\ell_{ik}| \le 1$

Beispiel für $n = 4$

$$A = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} A = \begin{bmatrix} 1 & & & \\ \frac{3}{4} & 1 & & \\ \frac{1}{2} & -\frac{2}{7} & 1 & \\ \frac{1}{4} & -\frac{3}{7} & \frac{1}{3} & 1 \end{bmatrix} \begin{bmatrix} 8 & 7 & 9 & 5 \\ & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \\ & & -\frac{6}{7} & -\frac{2}{7} \\ & & & \frac{2}{3} \end{bmatrix}$$

Gauß-Elimination mit Spaltenpivotsuche

Nach $n-1$ Schritten Gauß-Elimination erhält man

$$L_{n-1}P_{n-1}\cdots L_2P_2L_1P_1A = R$$

mit Permutationsmatrizen $P_1,\ldots,P_{n-1}$ oder

$$L_{n-1}P_{n-1}\cdots L_2P_2L_1P_1 = (L'_{n-1}\cdots L'_2L'_1)(P_{n-1}\cdots P_2P_1)$$

mit

$$L'_k = P_{n-1}\cdots P_{k+1}L_kP_{k+1}^{-1}\cdots P_{n-1}^{-1}$$

Da P_j nur Vertauschungen Zeilen j und m mit $m>j$ vertauscht, bleibt die Struktur von L'_k unverändert, lediglich die Elemente unterhalb der Diagonalen werden permutiert

LR-Zerlegung mit Spaltenpivotsuche

Satz: Jede nichtsinguläre Matrix $A \in \mathbb{K}^{n\times n}$ hat eine Zerlerung $PA = LR$ mit

- P Permutationsmatrix
- L untere Dreiecksmatrix mit $\ell_{i,i} = 1$ und $|\ell_{j,k}| \leq 1$
- R obere Dreiecksmatrix mit $r_{j,j} \neq 0$

Lösung linearer Gleichungssysteme $Ax = b$ in drei Schritten

1. Permutieren der rechten Seite $P \cdot b$
2. Lösen von $Ly = Pb$
3. Lösen von $Rx = y$

Beachte: Die Permutationsmatrix kann platzsparend mit Hilfe eines Vektors gespeichert werden