Numerische Mathematik für Maschinenbauer

Kondition, Rundungsfehler, Stabilität

A. Reusken

K.-H. Brakhage, Amira El Amouri, Thomas Jankuhn

Institut für Geometrie und Praktische Mathematik
RWTH Aachen

Sommersemester 2019

Heute in der Vorlesung

Themen: Dahmen & Reusken Kap 2.2-2.3

- Kondition eines linearen Problems
- Zahlendarstellung und Rundungsfehler
- Gleitpunktarithmetik
- Stabilität eines Algorithmus

Was Sie mitnehmen sollten:

- Welche Rolle spielt die Konditionszahl einer Matrix
- Wie werden Zahlen im Computer dargestellt
- Wichtige Eigenschaften der Gleitpunktarithmetik
- Stabilität vs. Kondition

Kondition: $f : \mathbb{R}^n \to \mathbb{R}^n$, f linear

Wir haben

$$y = f(x) = B\,x, \quad B \in \mathbb{R}^{n \times n}$$

bzw. für gestörte Daten

$$\tilde{y} = f(\tilde{x}) = B\,\tilde{x}$$

und damit

$$f(\tilde{x}) - f(x) = B\,\tilde{x} - B\,x = B\,(\tilde{x} - x)$$

Kondition: $f : \mathbb{R}^n \to \mathbb{R}^n$, f linear

Damit erhält man die Kondition für

- $f : \mathbb{R}^n \to \mathbb{R}^n$ (Eingabe: Vektor, Ausgabe: Vektor)

$$\frac{\|f(\tilde{x}) - f(x)\|}{\|f(x)\|} \leq \underbrace{\|B\| \cdot \|B^{-1}\|}_{\kappa(B)} \frac{\|\tilde{x} - x\|}{\|x\|}$$

wobei

$$\kappa(B) \equiv \|B\| \cdot \|B^{-1}\|$$

die Konditionszahl der Matrix B ist.

Beachte: $\kappa(B) = \kappa(B^{-1})$.

Beispiel 2.28.

Die Bestimmung des Schnittpunkts der Geraden

$$\begin{aligned} 3\,u_1 + 1.001\,u_2 &= 1.999 \\ 6\,u_1 + 1.997\,u_2 &= 4.003. \end{aligned}$$

(fast parallel!) ergibt das Problem $\boldsymbol{u} = \boldsymbol{A}^{-1}\boldsymbol{b}$ mit

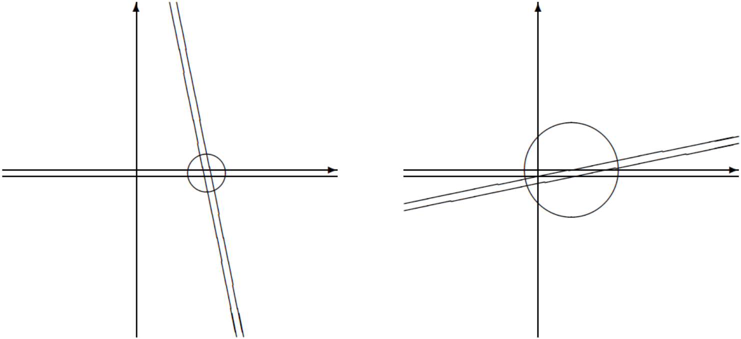
$$\boldsymbol{A} = \begin{pmatrix} 3 & 1.001 \\ 6 & 1.997 \end{pmatrix}, \qquad \boldsymbol{b} = \begin{pmatrix} 1.999 \\ 4.003 \end{pmatrix}.$$

Die Lösung ist

$$\boldsymbol{u} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Hier: $\boldsymbol{x} = \boldsymbol{b}$ (Eingabe), $\boldsymbol{y} = \boldsymbol{u}$ (Ausgabe), $\boldsymbol{B} = \boldsymbol{A}^{-1}$ (linear).

Beispiel 2.28. Kondition bei Bestimmung eines Schnittpunktes



Beispiel 2.28.

Effekt einer Störung in $\boldsymbol{b}$:

$$\tilde{\boldsymbol{b}} = \begin{pmatrix} \mathbf{2.002} \\ \mathbf{4} \end{pmatrix}, \qquad \tilde{\boldsymbol{u}} = \boldsymbol{A}^{-1}\,\tilde{\boldsymbol{b}}.$$

Man erhält

$$\boldsymbol{A}^{-1} = \frac{-\mathbf{1}}{\mathbf{0.015}} \begin{pmatrix} \mathbf{1.997} & -\mathbf{1.001} \\ -\mathbf{6} & \mathbf{3} \end{pmatrix}, \qquad \tilde{\boldsymbol{u}} = \begin{pmatrix} \mathbf{0.4004} \\ \mathbf{0.8} \end{pmatrix}$$

Wir betrachten die Maximumnorm:

$$\|\boldsymbol{x}\| = \|\boldsymbol{x}\|_\infty = \max_i |\boldsymbol{x}_i|.$$

Beispiel 2.28.

Es gilt

- Störung der Daten

$$\frac{\|\tilde{b}-b\|_\infty}{\|b\|_\infty} = \frac{3\cdot 10^{-3}}{4.003} \approx 7.5\cdot 10^{-4}$$

- Änderung des Resultats

$$\frac{\|\tilde{u}-u\|_\infty}{\|u\|_\infty} = \frac{1.8}{1} \approx 1.8$$

Schlechte Kondition wird quantifiziert durch

$$\|A\|_\infty\,\|A^{-1}\|_\infty = 4798.2.$$

Beispiel 2.31.

Wir betrachten als Beispiel die Zahl $\mathbf{123.75}$:

- Dezimalsystem (Basis 10)

$$\begin{aligned}&\mathbf{123.75}\\&= \mathbf{1 \cdot 10^2 + 2 \cdot 10^1 + 3 \cdot 10^0 + 7 \cdot 10^{-1} + 5 \cdot 10^{-2}}\\&= \mathbf{10^3\,(1 \cdot 10^{-1} + 2 \cdot 10^{-2} + 3 \cdot 10^{-3} + 7 \cdot 10^{-4} + 5 \cdot 10^{-5})}\end{aligned}$$

- Binärsystem (Basis 2)

$$\begin{aligned}&\mathbf{123.75}\\&= \mathbf{1 \cdot 2^6 + 1 \cdot 2^5 + 1 \cdot 2^4 + 1 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0}\\&\quad \mathbf{+ 1 \cdot 2^{-1} + 1 \cdot 2^{-2}}\\&= \mathbf{2^7\,(1 \cdot 2^{-1} + 1 \cdot 2^{-2} + 1 \cdot 2^{-3} + 1 \cdot 2^{-4} + 0 \cdot 2^{-5} + 1 \cdot 2^{-6}}\\&\quad \mathbf{+ 1 \cdot 2^{-7} + 1 \cdot 2^{-8} + 1 \cdot 2^{-9}) = 0.111101111 \cdot 2^{111}}\end{aligned}$$

Zahlendarstellung

Seien $b \in \mathbb{N}$, $b > 1$, fest gewählt. Jedes $x \in \mathbb{R}$, $x \neq 0$, lässt sich in der Form

$$x = \pm \left(\sum_{j=1}^{\infty} d_j \, b^{-j} \right) \cdot b^e$$

darstellen, mit $d_j \in \{0, 1, \ldots, b-1\}$, $d_1 \neq 0$,
und e eine ganze Zahl.

- Dezimalsystem (Basis $b = 10$)
  $$123.75 \Rightarrow 0.12375 \cdot 10^3$$
- Binärsystem (Basis $b = 2$)
  $$123.75 \Rightarrow 0.111101111 \cdot 2^{111}$$

Normalisierte Gleitpunktdarstellung

Gleitpunktdarstellung:

$$\begin{aligned} x &= \pm 0.d_1 d_2 \dots d_m \times b^e \\ &= \pm \left(\sum_{j=1}^{m} d_j \, b^{-j} \right) \times b^e \end{aligned}$$

wobei

- Basis $b \in \mathbb{N} \backslash \{1\}$
- Exponent $e \in \mathbb{Z}$ mit $r \leq e \leq R$
- Mantisse $f = \pm 0.d_1 d_2 \dots d_m, \; d_j \in \{0, 1, \dots, b-1\}$
- Mantissenlänge m
- Normalisierung: $d_1 \neq 0$ für $x \neq 0$

Maschinenzahlen

Nur endliche Anzahl von Zahlen darstellbar:

$$x = \pm \left(\sum_{j=1}^{m} d_j \, b^{-j} \right) \times b^{e}, \qquad r \leq e \leq R$$

$\Rightarrow$ Maschinenzahlen $\mathbb{M}(b, m, r, R)$,

$$\begin{aligned} x_{\mathrm{MIN}} &= b^{r-1}, \\ x_{\mathrm{MAX}} &= (1 - b^{-m}) \times b^{R}, \\ \mathbb{D} :&= [-x_{\mathrm{MAX}}, -x_{\mathrm{MIN}}] \cup [x_{\mathrm{MIN}}, x_{\mathrm{MAX}}] \end{aligned}$$

Maschinenzahlen

Nur endliche Anzahl von Zahlen darstellbar:

$$x = \pm \left(\sum_{j=1}^{m} d_j \, b^{-j} \right) \times b^e, \qquad r \leq e \leq R$$

$\Rightarrow$ Maschinenzahlen $\mathbb{M}(b, m, r, R)$,

Definition

Reduktionsabbildung $\mathrm{fl} : \mathbb{D} \to \mathbb{M}(b, m, r, R)$ definiert durch

$$\mathrm{fl}(x) := \pm \begin{cases} \left(\sum_{j=1}^{m} d_j \, b^{-j} \right) \times b^e & \text{falls } d_{m+1} < \frac{b}{2}, \\ \left(\sum_{j=1}^{m} d_j \, b^{-j} + b^{-m} \right) \times b^e & \text{falls } d_{m+1} \geq \frac{b}{2}, \end{cases}$$

Maschinengenauigkeit – Beispiel

Gleitpunktdarstellung: $\boldsymbol{b = 10}$, $\boldsymbol{m = 6}$

x	$\mathrm{fl}(x)$	$\left\lvert\frac{\mathrm{fl}(x)-x}{x}\right\rvert$
$\frac{1}{3} = 0.33333333\ldots$	$0.333333 * 10^0$	$1.0 * 10^{-6}$
$\sqrt{2} = 1.41421356\ldots$	$0.141421 * 10^1$	$2.5 * 10^{-6}$
$e^{-10} = 0.000045399927\ldots$	$0.453999 * 10^{-4}$	$6.6 * 10^{-7}$
$e^{10} = 22026.46579\ldots$	$0.220265 * 10^5$	$1.6 * 10^{-6}$
$\frac{1}{10} = 0.1$	$0.100000 * 10^0$	0.0

Gleitpunktdarstellung: $\boldsymbol{b = 2}$, $\boldsymbol{m = 10}$

x	$\mathrm{fl}(x)$	$\left\lvert\frac{\mathrm{fl}(x)-x}{x}\right\rvert$
$\frac{1}{3}$	$0.1010101011 * 2^{-1}$	$4.9 * 10^{-4}$
$\sqrt{2}$	$0.1011010100 * 2^{1}$	$1.1 * 10^{-4}$
e^{-10}	$0.1011111010 * 2^{-111}$	$3.3 * 10^{-4}$
e^{10}	$0.1010110000 * 2^{1111}$	$4.8 * 10^{-4}$
$\frac{1}{10}$	$0.1100110011 * 2^{-11}$	$2.4 * 10^{-4}$

Maschinengenauigkeit

- Für den relativen Rundungsfehler erhält man

$$\delta_x := \left|\frac{\mathrm{fl}(x) - x}{x}\right| \le \frac{\frac{b^{-m}}{2}\cdot b^e}{b^{-1}\cdot b^e} = \frac{b^{1-m}}{2}.$$

- Die (relative) Maschinengenauigkeit

$$\mathsf{eps} := \frac{b^{1-m}}{2}$$

charakterisiert das Auflösungsvermögen des Rechners, d.h.

$$\mathsf{eps} = \inf\{\delta > 0 \mid \mathrm{fl}(1+\delta) > 1\}$$

- Der Rundungsfehler ε erfüllt $|\varepsilon| \le \mathsf{eps}$ und es gilt

$$\mathrm{fl}(x) = x\,(1+\varepsilon).$$

Maschinengenauigkeit – Beispiel

Gleitpunktdarstellung: $b = 10$, $m = 6 \to$ eps $= \frac{1}{2} \cdot 10^{-5}$

x	$\mathrm{fl}(x)$	$\left\lvert\frac{\mathrm{fl}(x)-x}{x}\right\rvert$
$\frac{1}{3} = 0.33333333\ldots$	$0.333333 * 10^0$	$1.0 * 10^{-6}$
$\sqrt{2} = 1.41421356\ldots$	$0.141421 * 10^1$	$2.5 * 10^{-6}$
$e^{-10} = 0.000045399927\ldots$	$0.453999 * 10^{-4}$	$6.6 * 10^{-7}$
$e^{10} = 22026.46579\ldots$	$0.220265 * 10^5$	$1.6 * 10^{-6}$
$\frac{1}{10} = 0.1$	$0.100000 * 10^0$	0.0

Gleitpunktdarstellung: $b = 2$, $m = 10 \to$ eps $= 9.8 \cdot 10^{-4}$

x	$\mathrm{fl}(x)$	$\left\lvert\frac{\mathrm{fl}(x)-x}{x}\right\rvert$
$\frac{1}{3}$	$0.1010101011 * 2^{-1}$	$4.9 * 10^{-4}$
$\sqrt{2}$	$0.1011010100 * 2^{1}$	$1.1 * 10^{-4}$
e^{-10}	$0.1011111010 * 2^{-111}$	$3.3 * 10^{-4}$
e^{10}	$0.1010110000 * 2^{1111}$	$4.8 * 10^{-4}$
$\frac{1}{10}$	$0.1100110011 * 2^{-11}$	$2.4 * 10^{-4}$

Pseudoarithmetik

Exakte elementare arithmetische Operation von Maschinenzahlen
$\not\Rightarrow$ Maschinenzahl

Beispiel

$\boldsymbol{b = 10,\ m = 3}$:

$$\mathbf{0.346 \times 10^2 + 0.785 \times 10^2 = 0.1131 \times 10^3 \neq 0.113 \times 10^3}$$

Ähnliches passiert bei Multiplikation und Division.

Exakte Arithmetik $\rightsquigarrow$ Pseudoarithmetik (Gleitpunktarithmetik),

z.B.: $+ \rightsquigarrow \oplus$.

Pseudoarithmetik

Forderung

Für $\nabla \in \{+, -, \cdot, /\}$ gelte

$$x \circledtriangledown y = \mathrm{fl}(x \nabla y) \quad \text{für } x, y \in \mathbb{M}(b, m, r, R).$$

Da $\mathrm{fl}(x) = x\,(1+\varepsilon)$, folgt somit, dass für $\nabla \in \{+, -, \cdot, /\}$

$$x \circledtriangledown y = (x \nabla y)(1+\varepsilon) \quad \text{für } x, y \in \mathbb{M}(b, m, r, R)$$

und ein ε mit $|\varepsilon| \leq$ **eps** gilt.

Vorsicht bei Pseudoarithmetik:

- Grundlegende Regeln der Algebra, die bei exakter Arithmetik gelten, sind nicht mehr gültig.
- Reihenfolge der Verküpfung spielt eine Rolle (Assoziativität der Addition geht verloren).

Assoziativgesetz

Beispiel 2.36.

Zahlensystem mit $b = 10,\ m = 3$. Maschinenzahlen

$$\begin{array}{rcccl} x &=& 6590 &=& 0.659 \times 10^4 \\ y &=& 1 &=& 0.100 \times 10^1 \\ z &=& 4 &=& 0.400 \times 10^1 \end{array}$$

Exakte Rechnung:

$$(x + y) + z = (y + z) + x = 6595.$$

Pseudoarithmetik:

$$x \oplus y = 0.659 \times 10^4 \quad \text{und} \quad (x \oplus y) \oplus z = 0.659 \times 10^4,$$

aber

$$y \oplus z = 0.500 \times 10^1 \quad \text{und} \quad (y \oplus z) \oplus x = 0.660 \times 10^4.$$

Distributivgesetz

Beispiel 2.37.

Für $b = 10,\ m = 3, x = 0.156 \cdot 10^2$ und $y = 0.157 \cdot 10^2$

$$\begin{aligned}(x - y) \cdot (x - y) &= 0.01 \\ (x \ominus y) \odot (x \ominus y) &= 0.100 \times 10^{-1}\end{aligned}$$

aber

$$(x \odot x) \ominus (x \odot y) \ominus (y \odot x) \oplus (y \odot y) = -0.100 \times 10^1.$$

Auslöschung

Beispiel 2.38.

Betrachte

$$x = 0.73563, \quad y = 0.73441, \quad x - y = 0.00122.$$

Bei 3-stelliger Rechnung ($b = 10,\ m = 3,\ \text{eps} = \frac{1}{2} \cdot 10^{-2}$):

$$\begin{aligned} \tilde{x} &= \mathrm{fl}(x) = 0.736, & |\delta_x| &= 0.50 \cdot 10^{-3} \\ \tilde{y} &= \mathrm{fl}(y) = 0.734, & |\delta_y| &= 0.56 \cdot 10^{-3} \end{aligned}$$

Die relative Störung im Resultat:

$$\left| \frac{(\tilde{x} - \tilde{y}) - (x - y)}{x - y} \right| = \left| \frac{0.002 - 0.00122}{0.00122} \right| = 0.64$$

also sehr groß im Vergleich zu δ_x, δ_y.

Zusammenfassung Gleitpunktarithmetik

$$\boxed{\left|\frac{(x \bigtriangledown\!\!\!\!\bigcirc y) - (x \nabla y)}{(x \nabla y)}\right| \leq \textbf{eps}, \quad x, y \in \mathbb{M}, \quad \nabla \in \{+, -, \cdot, /\}}$$

Die relativen Rundungsfehler bei den elementaren Gleitpunktoperationen sind $\leq$ **eps**, wenn die Eingangsdaten x, y Maschinenzahlen sind.

Sei $f(x, y) = x \nabla y$, $x, y \in \mathbb{R}$, $\nabla \in \{+, -, \cdot, /\}$ und κ_{rel} die relative Konditionszahl von f. Es gilt

$$\boxed{\begin{array}{lll} \nabla \in \{\cdot, /\} : & \kappa_{\mathrm{rel}} \leq 1 & \text{für alle } x, y, \\ \nabla \in \{+, -\} : & \kappa_{\mathrm{rel}} \gg 1 & \text{wenn } |x \nabla y| \ll \max\{|x|, |y|\} \end{array}}$$

Sehr große Fehlerverstärkung bei $+, -$ möglich (Auslöschung).

Beispiel: Polynom 7. Grades

$$\begin{aligned} p(x) &= (x-1)^7 \\ &= x^7 - 7x^6 + 21x^5 - 35x^4 + 35x^3 - 21x^2 + 7x - 1 \end{aligned}$$

Matlab-Demo

Stabilität

Definition

Ein Algorithmus heißt gutartig oder stabil, wenn die durch ihn im Laufe der Rechnung erzeugten Fehler in der Größenordnung des durch die Kondition des Problems bedingten unvermeidbaren Fehlers bleiben.

- Kondition ist Eigenschaft des Problems
- Stabilität ist Eigenschaft des Verfahrens/Algorithmus

$\Rightarrow$ Wenn ein Problem schlecht konditioniert ist, kann man nicht erwarten, dass die Numerische Methode (ein stabiler Algorithmus) gute Ergebnisse liefert.

Ziel: Numerische Methode soll Fehlerverstärkung nicht noch weiter vergrößern

Beispiel 2.39.

Bestimmung der kleineren Lösung u^* von

$$y^2 - 2\,a_1\,y + a_2 = 0$$

für $a_1 = 6.000227$, $a_2 = 0.01$.

Algorithmus I

$$u^* = f(a_1, a_2) = -\frac{-2\,a_1}{2} - \sqrt{\left(\frac{-2\,a_1}{2}\right)^2 - a_2}$$
$$= a_1 - \sqrt{a_1^2 - a_2}.$$

$$\begin{array}{rrcl} & y_1 & = & a_1\,a_1 \\ \longrightarrow & y_2 & = & y_1 - a_2 \\ \longrightarrow & y_3 & = & \sqrt{y_2} \\ \longrightarrow & u^* & = & a_1 - y_3 \end{array}$$

Beispiel 2.39.

Algorithmus I

$$u^* = f(a_1, a_2) = a_1 - \sqrt{a_1^2 - a_2}.$$

In Gleitpunktarithmetik mit $b = 10$, $m = 5$ $\left(\text{eps} = \frac{1}{2} \times 10^{-4}\right)$:

$$\tilde{u}^* = 0.90000 \times 10^{-3}$$

Exakte Lösung:

$$u^* = 0.83336 \cdot 10^{-3}$$

- Problem für diese Eingangsdaten $a_1,\ a_2$ gut konditioniert.
- Durch Algorithmus erzeugte Fehler sind sehr viel größer als der unvermeidbare Fehler.

$\Rightarrow$ Algorithmus I ist nicht stabil

Beispiel 2.39.

Bestimmung der Lösung u^* von

$$y^2 - 2\,a_1\,y + a_2 = 0$$

für $a_1 = 6.000227$, $a_2 = 0.01$.

Algorithmus II (Alternative)

$$u^* = \frac{a_2}{a_1 + \sqrt{a_1^2 - a_2}}$$

$$\begin{array}{rrcl}
 & y_1 & = & a_1\,a_1 \\
\longrightarrow & y_2 & = & y_1 - a_2 \\
\longrightarrow & y_3 & = & \sqrt{y_2} \\
\longrightarrow & y_4 & = & a_1 + y_3 \\
\longrightarrow & u^* & = & a_2/y_4
\end{array}$$

Beispiel 2.39.

Algorithmus II

$$u^* = \frac{a_2}{a_1 + \sqrt{a_1^2 - a_2}}$$

In Gleitpunktarithmetik mit $b = 10$, $m = 5$ $\left(\text{eps} = \frac{1}{2} \times 10^{-4}\right)$:

$$\tilde{u}^* = 0.83333 \times 10^{-3}$$

Exakte Lösung:

$$u^* = 0.83336 \cdot 10^{-3}$$

- Gesamtfehler bleibt im Rahmen der Maschinengenauigkeit.
- Auslöschung tritt nicht auf.

$\Rightarrow$ Algorithmus II ist stabil

Rückwärtsstabilität

Ein Verfahren zur Berechnung von $f(x)$ liefert als Ergebnis $\tilde{f}(x)$.

Definition

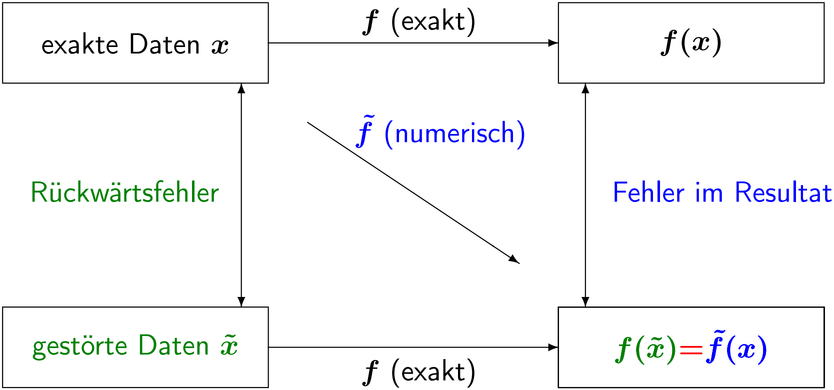
Das Verfahren heißt rückwärts stabil, wenn für alle $x \in X$,

$$\tilde{f}(x) = f(\tilde{x})$$

für ein $\tilde{x}$ mit $\frac{\|x-\tilde{x}\|}{\|x\|} = \mathcal{O}(\text{eps})$.

$\Rightarrow$ Ein rückwärts stabiler Algorithmus gibt die exakte Lösung des nahezu richtigen Problems (Daten, d.h. $x \to \tilde{x} = x + \Delta x$).

Rückwärtsanalyse



Rückwärtsstabilität

Satz

Wird ein rückwärts stabiler Algorithmus zur Lösung des Problems f mit Kondition $\kappa(x)$ angewendet, so gilt

$$\frac{\|\tilde{f}(x) - f(x)\|}{\|f(x)\|} = \mathcal{O}(\kappa(x)\,\mathrm{eps}).$$

Beweis:

$$\frac{\|\tilde{f}(x) - f(x)\|}{\|f(x)\|} = \frac{\|f(\tilde{x}) - f(x)\|}{\|f(x)\|} \lesssim \kappa(x)\,\underbrace{\frac{\|\tilde{x} - x\|}{\|x\|}}_{\mathcal{O}(\mathrm{eps})}.$$

Rückwärtsstabilität

Was haben wir gemacht?

Fehler im Algorithmus $\tilde{f}$ wurden

zurückgespiegelt auf Fehler in den Daten $\tilde{x}$.

$$\Downarrow$$

Vorteil: Auswertung von $f(\tilde{x})$ ist Frage nach Kondition von f.

Beispiel 2.40.: Summation ist rückwärts stabil

Geg.: Maschinenzahlen $x_1,\ x_2,\ x_3$, Maschinengenauigkeit **eps**.
Ges.: Summe $S = (x_1 + x_2) + x_3$.

Man erhält

$$\widetilde{S} = ((x_1 + x_2)(1 + \varepsilon_2) + x_3)(1 + \varepsilon_3)$$

mit $|\varepsilon_i| \leq \text{eps},\ i = 2, 3$.

Daraus folgt

$$\begin{aligned}\widetilde{S} &= x_1 (1 + \varepsilon_2)(1 + \varepsilon_3) + x_2 (1 + \varepsilon_2)(1 + \varepsilon_3) + x_3 (1 + \varepsilon_3) \\ &\doteq x_1 (1 + \varepsilon_2 + \varepsilon_3) + x_2 (1 + \varepsilon_2 + \varepsilon_3) + x_3 (1 + \varepsilon_3) \\ &= x_1 (1 + \delta_1) + x_2 (1 + \delta_2) + x_3 (1 + \delta_3)\end{aligned}$$

wobei

$$|\delta_1| = |\delta_2| = |\varepsilon_2 + \varepsilon_3| \leq 2\,\text{eps}, \quad |\delta_3| = |\varepsilon_3| \leq \text{eps}$$

Beispiel 2.40.: Summation ist rückwärts stabil

Es gilt

$$\begin{aligned}\widetilde{S} &= x_1\,(1+\delta_1) + x_2\,(1+\delta_2) + x_3\,(1+\delta_3) \\ &=: \tilde{x}_1 + \tilde{x}_2 + \tilde{x}_3,\end{aligned}$$

wobei

$$|\delta_1| = |\delta_2| = |\varepsilon_2 + \varepsilon_3| \leq 2\,\text{eps}, \quad |\delta_3| = |\varepsilon_3| \leq \text{eps}$$

$\Rightarrow$ Fehlerbehaftetes Resultat $\widetilde{S}$ als exaktes Ergebnis zu gestörten Eingabedaten $\tilde{x}_i = x_i(1+\delta_i)$.

$\Rightarrow$ Summenbildung ist rückwärts stabiles Verfahren.
Der durch Rechnung bedingte Fehler ist höchstens

$$\begin{aligned}\left|\frac{\tilde{S}-S}{S}\right| &\leq \kappa_{\text{rel}}(x) \cdot \sum_{j=1}^{3} \left|\frac{\tilde{x}_j - x_j}{x_j}\right| \\ &\leq \kappa_{\text{rel}}(x) \cdot \sum_{j=1}^{3} |\delta_j| \leq \kappa_{\text{rel}}(x) \cdot 5\,\text{eps}.\end{aligned}$$

Zusammenfassung

Was Sie mitnehmen sollten:

- $\kappa(B) = \|B\|\|B^{-1}\|$ beschreibt die relative Kondition des Problems $y = Bx$

Wie werden Zahlen im Computer dargestellt

- Maschinenzahlen $\mathbb{M}(b, m, r, R)$
  $\Rightarrow\ x_{\mathrm{MIN}},\ x_{\mathrm{MAX}},\ \mathsf{eps},\ |\varepsilon| \leq \mathsf{eps}$

Welche Probleme können dabei/deswegen auftreten?

- Assoziativ- und Distributivgesetz nicht mehr gültig
- Gefahr der Auslöschung bei $\nabla \in \{+, -\}$

Zusammenfassung

Stabilität vs. Kondition

- Bei einem stabilen Lösungsverfahren bleiben die im Laufe der Rechnung erzeugten Rundungsfehler in der Größenordnung der durch die Kondition des Problems bedingten unvermeidbaren Fehler.
- Kenntnisse über die Kondition eines Problems sind oft für die Interpretation oder Bewertung der Ergebnisse von entscheidender Bedeutung
  - "Schlechtes Ergebnis" bedeutet nicht unbedingt gleich "instabiler Algorithmus", sondern deutet evtl. auf eine schlechte Kondition des Problems hin.
- In einem Algorithmus sollen (wegen Stabilität) Auslöschungseffekte vermieden werden.

Verständnisfragen

Es seien $\boldsymbol{x}_{\mathrm{MIN}}$ bzw. $\boldsymbol{x}_{\mathrm{MAX}}$ die kleinste bzw. größte (strikt) positive Zahl sowie $\mathbf{eps}$ die relative Maschinengenauigkeit in der Menge $\mathbb{M}(\boldsymbol{b}, \boldsymbol{m}, \boldsymbol{r}, \boldsymbol{R})$ der Maschinenzahlen und $\mathbb{D} := [-\boldsymbol{x}_{\mathrm{MAX}}, -\boldsymbol{x}_{\mathrm{MIN}}] \cup [\boldsymbol{x}_{\mathrm{MIN}}, \boldsymbol{x}_{\mathrm{MAX}}]$. Ferner beschreibe $\ominus$ der Minusoperator in der Pseudoarithmetik. Alle Zahlen sind im Dezimalsystem angegeben.

Berechnen Sie $\boldsymbol{x}_{\mathrm{MAX}}$ für $\mathbb{M}(\mathbf{3}, \mathbf{2}, \mathbf{-1}, \mathbf{3})$ $\boxed{24}$

$\boxed{\text{w}}$ Es gilt $\left|\frac{(\boldsymbol{x}\ominus\boldsymbol{y})-(\boldsymbol{x}-\boldsymbol{y})}{(\boldsymbol{x}-\boldsymbol{y})}\right| \leq \mathbf{eps}$ für alle $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{M}(\boldsymbol{b}, \boldsymbol{m}, \boldsymbol{r}, \boldsymbol{R})$ mit $\boldsymbol{x} \neq \boldsymbol{y}$ und $\boldsymbol{x} - \boldsymbol{y} \in \mathbb{D}$.

$\boxed{\text{f}}$ Es gilt $\left|\frac{(\boldsymbol{x}\ominus\boldsymbol{y})-(\boldsymbol{x}-\boldsymbol{y})}{(\boldsymbol{x}-\boldsymbol{y})}\right| \leq \mathbf{eps}$ für alle $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{D}$ mit $\boldsymbol{x} \neq \boldsymbol{y}$ und $\boldsymbol{x} - \boldsymbol{y} \in \mathbb{D}$.

$\boxed{\text{f}}$ Bei einem stabilen Algorithmus ist der Ausgabefehler nicht viel größer als der Eingabefehler.