Numerische Analysis I
Vorlesung 3

Arnold Reusken

Hauke Saß, Stephanie Schwab

IGPM

Wintersemester 2022-2023

Übersicht

Themen: Dahmen & Reusken Kap 2.5-2.6

- Maschinenzahlen (Wiederholung)
- Rundungsfehler und Gleitpunktarithmetik (Wiederholung)
- Stabilität eines Algorithmus - Rückwärtsstabilität

Was Sie mitnehmen sollten:

- Wie ist die Menge der Maschinenzahlen definiert?
- Was ist Auslöschung?
- Wie hängt Stabilität mit Kondition zusammen?
- Wie ist Rückwärtsstabilität defininiert?

Normalisierte Gleitpunktdarstellung

Floating Point Representation:

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

wobei

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

Maschinenzahlen

Nur endliche Anzahl von Zahlen darstellbar:

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

$\Rightarrow$ Maschinenzahlen $\mathbb{M}(b, m, r, R)$.
Betragsmäßig kleinste bzw. größte Zahl in $\mathbb{M}(b, m, r, R)$: x_{MIN}, x_{MAX}.

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

Für $x \in \mathbb{D} := [-x_{\mathrm{MAX}}, -x_{\mathrm{MIN}}] \cup [x_{\mathrm{MIN}}, x_{\mathrm{MAX}}]$

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

d.h. die letzte Stelle der Mantisse wird um eins erhöht bzw. beibehalten, falls die Ziffer in der nächsten Stelle $\geq \frac{b}{2}$ bzw. $< \frac{b}{2}$

Maschinengenauigkeit

- Für den relativen Rundungsfehler erhält man
$$\left|\frac{\mathrm{fl}(x)-x}{x}\right| \leq \frac{\frac{b^{-m}}{2}b^e}{b^{-1}b^e} = \frac{b^{1-m}}{2}.$$
- 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 $|\varepsilon| \leq \mathsf{eps}$ und es gilt
$$\mathrm{fl}(x) = x\,(1+\varepsilon).$$

Gleitpunktarithmetik

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

Beispiel

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

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

Ähnliches passiert bei Multiplikation und Division.

Exakte Arithmetik $\rightsquigarrow$ Gleitpunktarithmetik (Pseudoarithmetik),

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

Gleitpunktarithmetik

Forderung

Für $\nabla \in \{+, -, \cdot, \div\}$ 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 $\nabla \in \{+, -, \cdot, \div\}$

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

und ein ε mit $|\varepsilon| \leq \mathbf{eps}$ gilt.

Vorsicht bei Gleitpunktarithmetik:

- 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 2.45

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

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

Exakte Rechnung:

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

Pseudoarithmetik:

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

aber

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

Distributivgesetz

Beispiel 2.46

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

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

aber

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

Auslöschung

Beispiel 2.47

Betrachte

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

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

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

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

$$\left|\frac{(x \circledtriangledown y) - (x \nabla y)}{(x \nabla y)}\right| \leq \textbf{eps}, \quad x, y \in \mathbb{M}, \quad \nabla \in \{+, -, \cdot, \div\}$$

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

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

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

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

Beispiele

In Matlab:

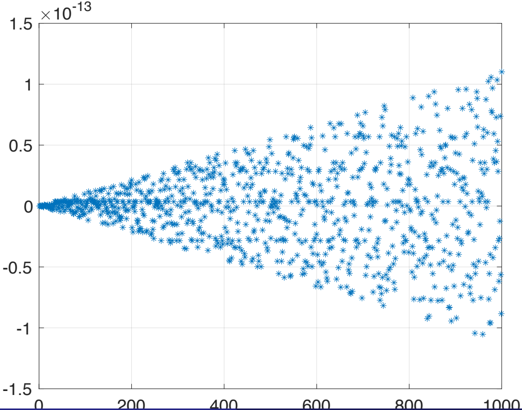
- $\mathtt{u} = 0.3/0.1$
  - Das Ergebnis ist nicht gleich 3, da Zähler etwas kleiner als 0.3 und Nenner etwas größer als 0.1.
- $\mathtt{a} = 2^{100};\ \mathtt{b} = \mathtt{a} + 2^{47};\ \mathtt{b} - \mathtt{a} = 0$
  - Die relative Differenz zwischen a und b is kleiner als **eps**
  - Es gibt keine Maschinenzahl zwischen $\mathbf{2^{100}}$ und $\mathbf{2^{100} + 2^{48}}$
- eps/3+eps/3+1-1= 2.220446049250313e-16
  eps/3+1+eps/3-1= 0
  - Assoziativgesetz gilt nicht

Beispiele

- Auswerten der Funktion $f(x) = 1 - x\left(\frac{x+1}{x} - 1\right)$

  Exakt: $f(x) = 1 - x\,\frac{x+1-x}{x} = 0$ für alle $x > 0$

  Auswertung in Matlab:



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 eine numerische Methode (ein stabiler Algorithmus) gute Ergebnisse liefert.

Ziel: Numerische Methode soll Fehlerverstärkung nicht signifikant weiter vergrößern

Beispiel 2.50

Bestimmung der Lösung u^* von

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

für $a_1 = 6.000227$, $a_2 = 0.01$.

Algorithmus I

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

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

Beispiel 2.50

Algorithmus I

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

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

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

Exakte Lösung:

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

- Problem ist 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
Ursache: Auslöschung

Beispiel 2.50

Bestimmung der Lösung u^* von

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

für $a_1 = 6.000227$, $a_2 = 0.01$.

Algorithmus II (Alternative)

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

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

Beispiel 2.50

Algorithmus II

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

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

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

Exakte Lösung:

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

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

$\Rightarrow$ Algorithmus II ist stabil

Rückwärtsstabilität

Wunsch: Auswertung von $f : X \to Y$

Wirklichkeit: berechnetes Ergebnis $\tilde{f} : X \to Y$

wobei $f \neq \tilde{f}$ aufgrund von

- Rundungsfehlern (Maschinengenauigkeit),
- Gleitpunktarithmetik.

Das Ziel

$$\frac{\|\tilde{f}(x) - f(x)\|}{\|f(x)\|} = \mathcal{O}(\text{eps})$$

ist zu ehrgeizig.

Grund: Wenn Problem f schlecht konditioniert ist, werden Datenstörungen um Kondition $\kappa \gg 1$ des Problems verstärkt.

Rückwärtsstabilität

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

Definition

Das Verfahren heißt rückwärts stabil, wenn

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

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

$\Rightarrow$ Ein rückwärts stabiler Algorithmus gibt die exakte Lösung des Problems mit nahezu richtigen Eingabedaten (d.h. $x \to \tilde{x} = x(1+\epsilon)$, $|\epsilon| \leq \text{eps}$).

Rückwärtsstabilität

Satz

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

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

Beweis:

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

Rückwärtsanalyse

$$
\begin{tikzcd}[column sep=huge]
\boxed{\text{exakte Daten } x} \arrow[rr, "{f\,[\text{exakt}]}"] \arrow[dd] \arrow[ddrr, "{\tilde f = f\,[\text{numerisch}]}"'] & & \boxed{f(x)} \arrow[d, dotted] \\
& & \text{Fehler } f(x) - f(\tilde x)\text{: Kondition des Problems} \\
\boxed{\text{gestörte Daten } \tilde x} \arrow[rr, "{f\,[\text{exakt}]}"] & & \boxed{f(\tilde x) = \tilde f(x)} \arrow[u, dotted]
\end{tikzcd}
$$

Beispiel 2.54: 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 2.54: 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

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

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

Der durch Rechnung bedingte Fehler ist höchstens

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

Beispiel 2.54

Der für die Summation $f(x) = f(x_1, x_2, x_3) = x_1 + x_2 + x_3$ unvermeidbare Fehler ist

$$\left| \frac{f(\tilde{x}) - f(x)}{f(x)} \right| \leq \kappa_{\mathrm{rel}}(x) \sum_{j=1}^{3} \left| \frac{\tilde{x}_j - x_j}{x_j} \right| \leq \kappa_{\mathrm{rel}}(x) 3\,\mathrm{eps},$$

wenn Daten höchstens mit Maschinengenauigkeit gestört werden ($\tilde{x}_i = x_i(1+\varepsilon)$, $|\varepsilon| \leq \mathrm{eps}$).

$\Rightarrow$ Berechnung von S ist ein stabiler Algorithmus

Zusammenfassung

- Maschinenzahlen:

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

- Reduktionsabbildung $\mathrm{fl} : \mathbb{R} \to \mathbb{M}$,
  Maschinengenauigkeit: $\mathrm{eps} = \frac{b^{1-m}}{2}$
- Pseudoarithmetik $\bigtriangledown$: siehe (2.67).
- Bei $\bigtriangledown$ gelten Assoziativität und Distributivität im Allg. nicht.
- Auslöschung ist eine Konsequenz der schlechten Kondition der Funktion $f(x_1, x_2) = x_1 + x_2$ wenn $x_1 \approx -x_2$.

Zusammenfassung

Stabilität ?

- Stabilität ist eine Eigenschaft des Algorithmus.
- Stabilität kann man beeinflussen durch Anpassung des Algorithmus.
- Konzept der Rückwärtsstabilität: Interpretiere sämtliche im Laufe der Rechnung auftretenden Fehler als Ergebnis *exakter* Rechnung zu geeignet *gestörten Daten*.
- Summenbildung ist Rückwärtsstabil.