

Numerische Mathematik für Maschinenbauer

Gleitpunktdarstellung, Stabilität

A. Reusken

K.-H. Brakhage, I. Voulis, H. Saß

Institut für Geometrie und Praktische Mathematik
RWTH Aachen

Sommersemester 2017

Heute in der Vorlesung

Themen: Dahmen & Reusken Kap. 2.2/2.3

- ▶ Zahlendarstellung und Rundungsfehler
- ▶ Gleitpunktarithmetik
- ▶ Stabilität eines Algorithmus

Was Sie mitnehmen sollten:

- ▶ Wie werden Zahlen im Computer dargestellt
- ▶ Wichtige Eigenschaften der Gleitpunktarithmetik
- ▶ Stabilität vs. Kondition

Beispiel 2.31

Wir betrachten als Beispiel die Zahl **123.75**:

- ▶ Dezimalsystem (Basis 10)

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

- ▶ Binärsystem (Basis 2)

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

Zahlendarstellung

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

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

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

- ▶ Dezimalsystem (Basis $b = 10$)

$$123.75 \Rightarrow 0.12375 \times 10^3$$

- ▶ Binärsystem (Basis $b = 2$)

$$123.75 \Rightarrow 0.111101111 \times 2^{111}$$

Normalisierte Gleitpunktdarstellung

Floating Point Representation:

$$\begin{aligned}x &= \pm 0.d_1d_2\dots d_m \times b^e \\ &= \pm \left(\sum_{j=1}^m d_j b^{-j} \right) \times b^e\end{aligned}$$

wobei

- ▶ Basis $b \in \mathbb{N} \setminus \{1\}$
- ▶ Exponent $e \in \mathbb{Z}$ mit $r \leq e \leq R$
- ▶ Mantisse $f = \pm 0.d_1d_2\dots d_m$, $d_j \in \{0, 1, \dots, b-1\}$
- ▶ Mantissenlänge m
- ▶ Normalisierung: $d_1 \neq 0$ für $x \neq 0$

Maschinenzahlen

Nur endliche Anzahl von Zahlen darstellbar (m, r, R vs. ∞):

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

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

$$x_{\text{MIN}} = b^{r-1}, \quad x_{\text{MAX}} = (1 - b^{-m})b^R.$$

$$\mathbb{D} := [-x_{\text{MAX}}, -x_{\text{MIN}}] \cup [x_{\text{MIN}}, x_{\text{MAX}}].$$

Reduktionsabbildung

$\text{fl} : \mathbb{D} \rightarrow \mathbb{M}(b, m, r, R)$ definiert durch

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

Bildbereich und Genauigkeit

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

- ▶ Es gibt einen begrenzten Bereich von Zahlen, die dargestellt werden können

Die Endlichkeit von e beschränkt den **Bildbereich**.

- ▶ Es gibt nur eine endliche Anzahl von Zahlen, die innerhalb des Bildbereichs dargestellt werden können

Die Endlichkeit von f beschränkt die **Genauigkeit**.

Bildbereich

Die Endlichkeit von e beschränkt den **Bildbereich**:

- ▶ Betragsmäßig kleinste ($\neq 0$) Zahl:

$$x_{\text{MIN}} = b^{r-1}$$

- ▶ Betragsmäßig größte Zahl:

$$x_{\text{MAX}} = (1 - b^{-m}) b^R$$

Achtung:

- ▶ Unterlauf, wenn $0 \neq |x| < |x_{\text{MIN}}|$;
- ▶ Überlauf, wenn $|x| > |x_{\text{MAX}}|$.

Maschinengenauigkeit – Beispiel

Gleitpunktdarstellung mit $b = 10$ und $m = 6$

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

Gleitpunktdarstellung mit $b = 2$ und $m = 10$

x	$\text{fl}(x)$	$\left \frac{\text{fl}(x) - x}{x} \right $
$\frac{1}{3}$	$0.1010101011 * 2^{-1}$	$4.9 * 10^{-4}$
$\sqrt{2}$	$0.1011010100 * 2^1$	$1.1 * 10^{-4}$
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

$$\left| \frac{\text{fl}(x) - x}{x} \right| \leq \frac{\frac{b^{-m}}{2} b^e}{b^{-1} b^e} = \frac{b^{1-m}}{2}.$$

- ▶ Die (relative) **Maschinengenauigkeit**

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

charakterisiert das Auflösungsvermögen des Rechners, d.h.

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

- ▶ Der Rundungsfehler ε erfüllt $|\varepsilon| \leq \text{eps}$ und es gilt

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

Maschinengenauigkeit – Beispiel

Gleitpunktdarstellung: $b = 10, m = 6 \rightarrow \text{eps} = \frac{1}{2} \times 10^{-5}$

x	$\text{fl}(x)$	$\frac{ \text{fl}(x)-x }{x}$
$\frac{1}{3} = 0.33333333 \dots$	$0.333333 \cdot 10^0$	$1.0 \cdot 10^{-6}$
$\sqrt{2} = 1.41421356 \dots$	$0.141421 \cdot 10^1$	$2.5 \cdot 10^{-6}$
$e^{-10} = 0.000045399927 \dots$	$0.453999 \cdot 10^{-4}$	$6.6 \cdot 10^{-7}$
$e^{10} = 22026.46579 \dots$	$0.220265 \cdot 10^5$	$1.6 \cdot 10^{-6}$
$\frac{1}{10} = 0.1$	$0.100000 \cdot 10^0$	0.0

Gleitpunktdarstellung: $b = 2, m = 10 \rightarrow \text{eps} = 9.8 \times 10^{-4}$

x	$\text{fl}(x)$	$\frac{ \text{fl}(x)-x }{x}$
$\frac{1}{3}$	$0.1010101011 \cdot 2^{-1}$	$4.9 \cdot 10^{-4}$
$\sqrt{2}$	$0.1011010100 \cdot 2^1$	$1.1 \cdot 10^{-4}$
e^{-10}	$0.1011111010 \cdot 2^{-111}$	$3.3 \cdot 10^{-4}$
e^{10}	$0.1010110000 \cdot 2^{1111}$	$4.8 \cdot 10^{-4}$
$\frac{1}{10}$	$0.1100110011 \cdot 2^{-11}$	$2.4 \cdot 10^{-4}$

Pseudoarithmetik

Exakte elementare arithmetische Operation von Maschinenzahlen
 \Rightarrow Maschinenzahl

Beispiel

$b = 10, m = 3:$

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

Ähnliches passiert bei Multiplikation und Division.

Exakte Arithmetik \rightsquigarrow Pseudoarithmetik (Gleitpunktarithmetik).

Z.B. $+$ \rightsquigarrow \oplus .

Pseudoarithmetik

Forderung

Für $\nabla \in \{+, -, \times, \div\}$ gelte

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

Da $\text{fl}(x) = x(1 + \varepsilon)$, folgt somit, dass für $\nabla \in \{+, -, \times, \div\}$

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

und ein ε mit $|\varepsilon| \leq \text{eps}$ 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

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

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

$$(x - y) \times (x - y) = 0.01$$

$$(x \ominus y) \otimes (x \ominus y) = 0.100 \times 10^{-1}$$

aber

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

Auslöschung

Beispiel 2.38:

Betrachte

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

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

$$\tilde{x} = \text{fl}(x) = 0.736, \quad |\delta_x| = 0.50 \times 10^{-3}$$

$$\tilde{y} = \text{fl}(y) = 0.734, \quad |\delta_y| = 0.56 \times 10^{-3}$$

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

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

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

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

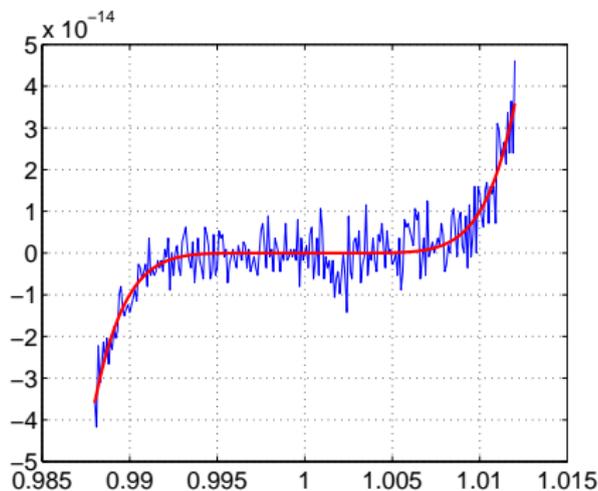
$$\nabla \in \{\times, \div\} : \kappa_{\text{rel}} \leq 1 \quad \text{für alle } x, y,$$

$$\nabla \in \{+, -\} : \kappa_{\text{rel}} \gg 1 \quad \text{wenn } |x \nabla y| \ll \max\{|x|, |y|\}$$

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

Beispiel: Polynom 7. Grades

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



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

⇒ 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: $y^2 - 2a_1y + a_2 = 0$

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

Algorithmus I

$$y_1 = a_1 a_1$$

$$y_2 = y_1 - a_2$$

$$y_3 = \sqrt{y_2}$$

$$u^* = a_1 - y_3$$

Für $a_1 = 6.000227$, $a_2 = 0.01$.

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

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

Exakte Lösung: $u^* = 0.83336 \times 10^{-3}$.

- ▶ Problem für diese Eingangsdaten a_1 , a_2 gut konditioniert.
- ▶ Durch Algorithmus erzeugte Fehler sehr viel größer als unvermeidbarer Fehler.

⇒ Algorithmus I ist nicht stabil

Beispiel 2.39: $y^2 - 2a_1y + a_2 = 0$

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

Algorithmus II

$$y_1 = a_1 a_1$$

$$y_2 = y_1 - a_2$$

$$y_3 = \sqrt{y_2}$$

$$y_4 = a_1 + y_3$$

$$u^* = \frac{a_2}{y_4}$$

In Gleitpunktarithmetik mit $b = 10$, $m = 5$:

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

Exakte Lösung: $u^* = 0.83336 \times 10^{-3}$

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

⇒ Algorithmus II ist **stabil**

Rückwärtsstabilität

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

Definition

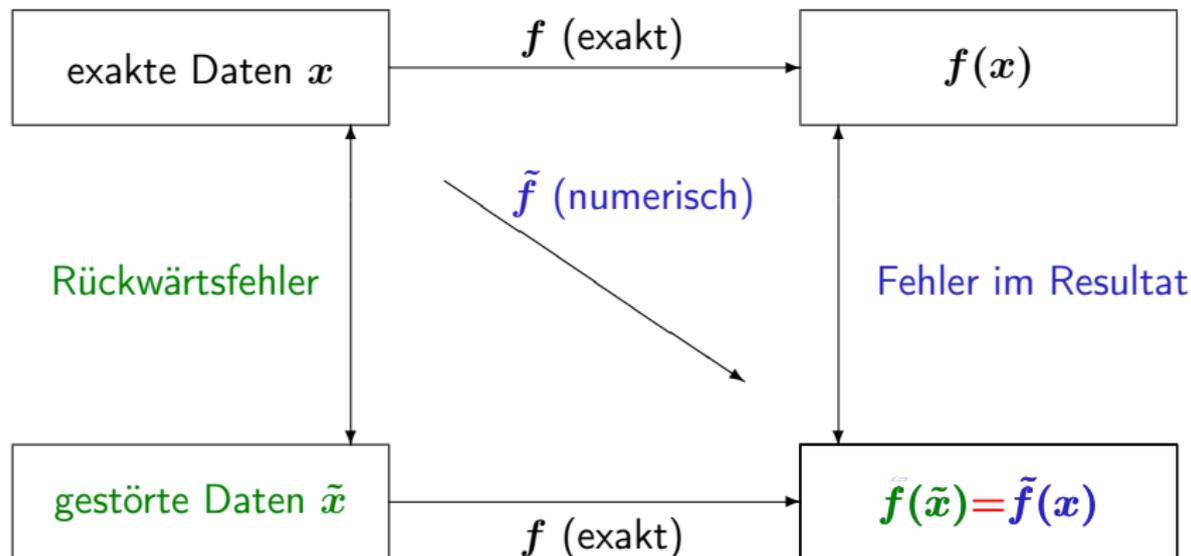
Das Verfahren heißt **rückwärts stabil**, wenn für alle $x \in X$,

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

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

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

Rückwärtsanalyse



Rückwärtsstabilität

Satz

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

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

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

Was haben wir gemacht?

- ▶ Fehler im Algorithmus wurden

zurückgespiegelt auf Fehler in den Daten.

⇒ 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

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

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

$$\begin{aligned} \tilde{S} &= x_1(1 + \varepsilon_2)(1 + \varepsilon_3) + x_2(1 + \varepsilon_2)(1 + \varepsilon_3) + x_3(1 + \varepsilon_3) \\ &\doteq x_1(1 + \varepsilon_2 + \varepsilon_3) + x_2(1 + \varepsilon_2 + \varepsilon_3) + x_3(1 + \varepsilon_3) \\ &= x_1(1 + \delta_1) + x_2(1 + \delta_2) + x_3(1 + \delta_3) \\ &=: \hat{x}_1 + \hat{x}_2 + \hat{x}_3, \end{aligned}$$

wobei

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

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

Beispiel 2.40

Der durch Rechnung bedingte Fehler ist höchstens

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

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

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

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

Größenordnung der Fehler identisch \Rightarrow Berechnung von S ist ein stabiler Algorithmus.

Zusammenfassung

Was Sie mitnehmen sollten:

Wie werden Zahlen im Computer dargestellt

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

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 x_{MIN} bzw. x_{MAX} die kleinste bzw. größte (strikt) positive Zahl sowie eps die relative Maschinengenauigkeit in der Menge $\mathbb{M}(\mathbf{b}, \mathbf{m}, \mathbf{r}, \mathbf{R})$ der Maschinenzahlen und $\mathbb{D} := [-x_{\text{MAX}}, -x_{\text{MIN}}] \cup [x_{\text{MIN}}, x_{\text{MAX}}]$. Ferner beschreibe $\mathbf{fl} : \mathbb{D} \rightarrow \mathbb{M}(\mathbf{b}, \mathbf{m}, \mathbf{r}, \mathbf{R})$ die Standardrundung. Alle Zahlen sind im Dezimalsystem angegeben.

Berechnen Sie x_{MAX} für $\mathbb{M}(\mathbf{3}, \mathbf{2}, -\mathbf{1}, \mathbf{3})$

- Es gilt $\left| \frac{\mathbf{fl}(x-y) - (x-y)}{x-y} \right| \leq \text{eps}$ für alle $x, y \in \mathbb{M}(\mathbf{b}, \mathbf{m}, \mathbf{r}, \mathbf{R})$ mit $x \neq y$.
- Es gilt $\left| \frac{\mathbf{fl}(x-y) - (x-y)}{x-y} \right| \leq \text{eps}$ für alle $x, y \in \mathbb{D}$ mit $x \neq y$.
- Bei einem stabilen Algorithmus ist der Ausgabefehler nicht viel größer als der Eingabefehler.