8. Fehleranalyse und Gleitpunktarithmetik

Lernziele:

- Sie kennen Fehlerquellen bei der numerischen Lösung eines Problems und sind in der Lage, diese zu analysieren.
- Sie berücksichtigen bei der numerischen Lösung eines Problems Effekte der Gleitpunktarithmetik.
- Sie unterscheiden zwischen der Kondition und Stabilität eines Algorithmus und beziehen die Stabilität als Kriterium für die Brauchbarkeit eines Algorithmus ein.

Literatur:

- Huckle T., Schneider S.: Numerische Methoden, Kap. 4-7
- Chapra S. C.: Applied Numerical Methods with Matlab, Chap. 4

8.1 Fehlerquellen

Beispiel: Berechnung der Erdoberfläche mit der Formel

$$O = 4\pi R^2$$

Fehlerquellen

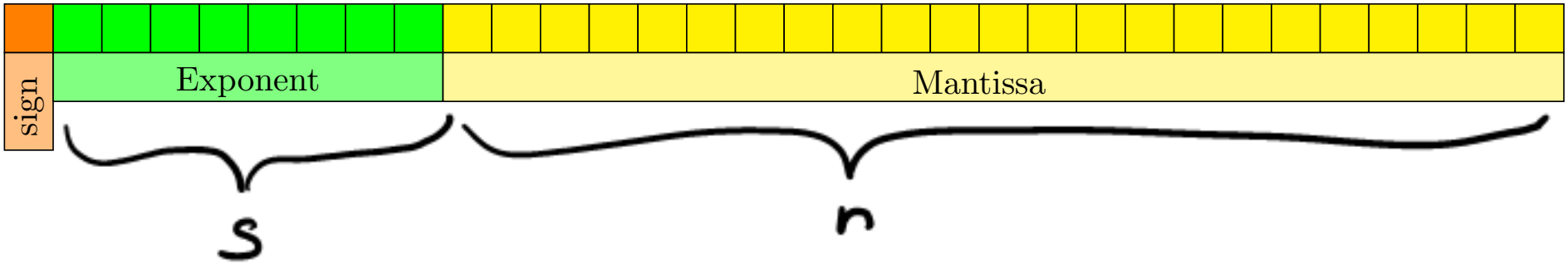
- Modellierungsfehler
- Fehler in den Eingangsdaten
- Diskretisierungsfehler
- Rundungsfehler bei Darstellung als Maschinenzahl
- Fehler aufgrund von Gleitpunktarithmetik

8.2 Maschinenzahlen

Darstellung: $M = \pm m \cdot b^E$

- Basis $b = 2, 10, 16$
- Exponent $E \in [L, U] \subset \mathbb{Z}$ der Länge s
- Mantisse $m = m_0.m_1 m_2 m_3 \ldots m_n$ mit $m_i \in \{0, 1, \ldots, b-1\}$
  $m_0 \neq 0$ in normalisierter Darstellung

Beispiel: Single Precision



IEEE Standard

- Untergrenze L für den Exponent
- Obergrenze U für den Exponent
- Länge s des Exponenten
- Anzahl n der signifikanten Ziffern

Precision	s	n	Bytes	L	U	Min	Max
Single	8	23	4	-126	127	2^{-126}	$2^{128}-1$
Double	11	52	8	-1022	1023	2^{-1022}	$2^{1024}-1$

Beispiel 8.2.1: Binärzahldarstellung von 10.1

Ganzzahliger Anteil:

$$10 : 2 = 5 \,\mathrm{R}\, \mathbf{0}$$
$$5 : 2 = 2 \,\mathrm{R}\, \mathbf{1}$$
$$2 : 2 = 1 \,\mathrm{R}\, \mathbf{0}$$
$$1 : 2 = 0 \,\mathrm{R}\, \mathbf{1}$$

Gebrochenrationaler Anteil:

$$0.1 * 2 = 0.2 + \mathbf{0}$$
$$0.2 * 2 = 0.4 + \mathbf{0}$$
$$0.4 * 2 = 0.8 + \mathbf{0}$$
$$0.8 * 2 = 0.6 + \mathbf{1}$$
$$0.6 * 2 = 0.2 + \mathbf{1}$$

$$(10)_{10} = 8 + 2 = (1010)_2 \qquad (0.1)_{10} = (.0\overline{0011})_2$$

$$(10.1)_{10} = (1010.0\overline{0011})_2 = 1.0100\overline{0011}_2 \cdot 2^3$$

Nicht exakt darstellbar $\Longrightarrow$ Rundungsfehler

Beispiel 8.2.2: Binärzahldarstellung von 0.3

$$(0.3)_{10} = (0.0\overline{1001})_2 = 1.\overline{0011} \cdot 2^{-2}$$

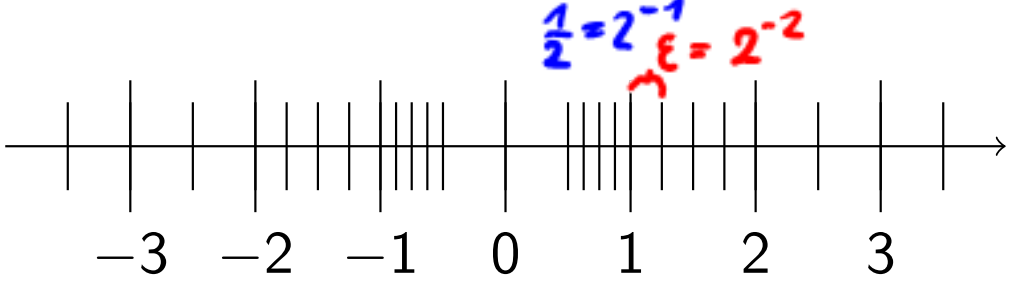
normalisierte Darstellung

$M = \pm\, m \cdot 2^E$

E: Exponent

Beispiel 8.2.3:

Geben Sie alle in einem Gleitpunktsystem mit $b = 2, n = 2, L = -1, U = 1$ darstellbaren Maschinenzahlen an.



Welche Dezimalzahlen sind exakt darstellbar?

m	$E=-1$	$E=0$	$E=1$
$(1.00)_2$	0.50	1.00	2.00
$(1.01)_2$	0.625	1.25	2.50
$(1.10)_2$	0.75	1.50	3.00
$(1.11)_2$	0.875	1.75	3.50

Eigenschaften von Gleitpunktsystemen

- Sie sind endlich und diskret.
  $\Longrightarrow$ Rundungsfehler, da nicht alle Zahlen exakt darstellbar und Ergebnisse von Gleitpunktoperationen evtl. wieder keine Maschinenzahlen sind
- Sie sind nach oben und unten beschränkt.
  $\Longrightarrow$ Probleme mit Overflow bzw. Underflow (s. Chapra 4.2.1)
- Der Abstand zwischen zwei aufeinanderfolgenden Maschinenzahlen wächst mit dem Exponenten E (d. h. der Größenordnung).
- Maschinenzahlen haben nur zwischen zwei aufeinanderfolgenden Exponenten b^E und b^{E+1} gleichen Abstand, nämlich b^{E-n}.

Maschinengenauigkeit $\varepsilon = 2^{-n}$

- ist der Abstand von $b^0 = 1$ zur nächstgrößeren Maschinenzahl, d. h. die kleinste Zahl x mit $rd(1 + x) \neq 1$.

  Bestimmen Sie ε in einem binären Gleitpunktsystem mit $n = 2$ bzw. geben Sie eine allgemein Formel für ε in einem binären Gleitpunktsystem mit Mantissenlänge n an.

- ist die obere Schranke für den relativen Rundungsfehler.

  Absoluter Rundungsfehler: $|rd(x) - x| \leq |x| \cdot \varepsilon$ abhängig von Größenord. $|x|$

  Relativer Rundungsfehler: $\frac{|rd(x)-x|}{|x|} \leq \varepsilon$

  größenordnungsbereinigt

Beispiel 8.2.4:
Testen Sie

$$0.1 + 0.1 + 0.1 = 0.3$$

in R auf Gleichheit

8.3 Gleitpunktarithmetik

Bei Gleitpunktoperationen gelten nicht die gewohnten Rechengesetze.

$n = 6$

Beispiel 8.3.1: $a = 1.23456 \cdot 10^{-3}$, $b = 1.00000 \cdot 10^{0}$, $c = -b$

$a + (b + c) =$ $rd(a+0) = 1.23456 \cdot 10^{-3}$ (exaktes Ergebnis)

$(a + b) + c =$ $rd(rd(a+b)+c) = 0.00123 = 1.23000 \cdot 10^{-3}$

Gleitpunktaddition:

(1) Exponentenabgleich durch Verschiebung der Mantisse (zum größeren Exponenten)

$a = 0.00123456 \cdot 10^{0}$, $b = 1.00000 \cdot 10^{0}$

(2) Addition in höherer Genauigkeit

$a+b = 1.00123456$ (exaktes Ergebnis)

(3) Runden des Ergebnisses auf Maschinenzahl

$rd(a+b) = 1.00123$

Beachte: Die Rechengesetze

- Assoziativgesetz
- Kommutativgesetz

gelten bei der Gleitpunktaddition nicht.

Die Reihenfolge der Berechnung kann das Ergebnis beeinflussen.

Idealerweise sollte man in aufsteigender Reihenfolge addieren, da signifikante Stellen verloren gehen, wenn man zu einer großen Zahl eine kleine addiert.

Beispiel 8.3.2: Ist folgende Berechnung numerisch geschickt?
Falls nein: Wie lässt sich die Summe alternativ berechnen?

$$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \cdots + \frac{1}{10^6}$$

Die Subtraktion von zwei Gleitpunktzahlen, die sich nur in ihren weniger signifikanten Stellen unterscheiden, führt zu einer Erhöhung der Signifikanz dieser Stellen im Ergebnis und damit zu einer Verstärkung des Rundungsfehlers. Dieser Effekt heißt **Auslöschung**.

Beispiel 8.3.3: $b = 10, n = 5$

$$3.97403 \cdot 10^2 - 3.97276 \cdot 10^2 = 1.27000 \cdot 10^{-1}$$

höhere Signifikanz der gerundeten Stelle

Beispiel: Lösungen einer quadratischen Gleichung $ax^2 + bx + c = 0$

Welches numerische Problem kann bei der Lösung der Gleichung

$$x^2 + 200x - 0.000015 = 0$$

in Gleitpunktarithmetik auftreten?
Welche Formel ist zur numerischen Lösung geeignet?

$$(1) \quad x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \qquad \text{oder} \qquad (2) \quad x_{1,2} = \frac{2c}{-b \mp \sqrt{b^2 - 4ac}}$$

8.4 Kondition und Stabilität

Fortpflanzung von Fehlern in den Daten vs. Verfahrensfehlern

- Numerische Lösung eines Problems
  - x: exakte Eingangsdaten
  - f: analytische Lösung
  - $\hat{x}$: fehlerbehaftete Eingangsdaten
  - $\hat{f}$: numerisches Lösungsverfahren
- Gesamtfehler

$$f(x) - \hat{f}(\hat{x}) = \underbrace{f(x) - f(\hat{x})}_{\text{Fehlerfortpflanzung}} + \underbrace{f(\hat{x}) - \hat{f}(\hat{x})}_{\text{Verfahrensfehler}}$$

→ Kondition → stabiler Algorithmus

- Die Fehlerfortpflanzung hängt nur von f ab, nicht von dem numerischen Lösungsverfahren.

Kondition

Ein Problem heißt gut konditioniert, wenn kleine relative Fehler in den Eingangsdaten einen kleinen relativen Fehler im Ergebnis verursachen.

Ein Problem heißt schlecht konditioniert, wenn der relative Fehler im Ergebnis deutlicher größer ist als die Fehler in den Eingangsdaten.

Konditionszahl:

$$\begin{aligned} \text{cond}(x) &= \left| \frac{\text{relativer Fehler im Ergebnis}}{\text{relativer Fehler in den Eingabedaten}} \right| \\ &= \left| \frac{\frac{f(\tilde{x})-f(x)}{f(x)}}{\frac{\tilde{x}-x}{x}} \right| \end{aligned}$$

Δf

Δx

Ein Problem ist schlecht konditioniert, wenn cond $\gg 1$

Herleitung für Konditionszahl

Ziel ist: $\frac{\Delta f}{f}$ verglichen mit $\frac{\Delta x}{x}$

$\frac{\Delta f}{f}$: relativer Fehler im Ergebnis

$\frac{\Delta x}{x}$: relativer Fehler in den Eingangsdaten

$$\frac{\Delta f}{f} = \kappa \cdot \frac{\Delta x}{x}$$

Man sagt, dass ein Problem gut konditioniert ist, wenn der Faktor κ erfüllt: $0 < |\kappa| < 100$

Es gilt: $\frac{\Delta f}{\Delta x} \approx f'(x)$, falls Δx klein

$$\left(\lim_{\Delta x \to 0} \frac{\Delta f}{\Delta x} = f'(x) \right)$$

$$\Rightarrow \quad \Delta f \approx f'(x) \cdot \Delta x \qquad \Big| \cdot \frac{1}{f}$$

$$\frac{\Delta f}{f(x)} \approx \underbrace{\frac{f'(x)}{f(x)} \cdot x}_{\kappa(x) = \text{cond}(x)} \frac{\Delta x}{x}$$

Fehlerfortpflanzung

$$\begin{aligned} z &= f(x) \\ \Delta f &= f(\tilde{x}) - f(x) \approx f'(x)\Delta x \end{aligned}$$

Konditionszahl:

$$\begin{aligned} \text{cond} &= \left| \frac{\frac{f(\tilde{x})-f(x)}{f(x)}}{\frac{\tilde{x}-x}{x}} \right| \\ &= \left| \frac{f(\tilde{x}) - f(x)}{\Delta x} \cdot \frac{x}{f(x)} \right| \approx \left| \frac{f'(x)}{f(x)} x \right| \end{aligned}$$

Bemerkungen:

(1) Die Kondition eines Problems ist abhängig von f und x.

(2) Falls ein Problem schlecht konditioniert ist, dann gibt es keinen numerisch günstigen Algorithmus zur Lösung des Problems.

Kondition von elementaren Funktionen:

(1) Lineare Funktion: $f(x) = ax + b$

$$\text{cond}(x) = \frac{f'(x)}{f(x)} \cdot x = \frac{a \cdot x}{ax+b} = \frac{ax+b}{ax+b} - \frac{b}{ax+b}$$

$$= 1 - \frac{b}{ax+b}$$

$\to 0$ für $x \to \pm\infty$

gut konditioniert !

(2) Gebrochen rationale Fkt. $f(x) = \frac{a}{x}$

$$\text{cond}(x) = \frac{f'(x)}{f(x)} \cdot x = \frac{-\frac{a}{x^2} \cdot x}{\frac{a}{x}} = -1$$

$\Rightarrow$ gut konditioniert

(3) Exponentialfkt. $f(x) = e^x$

$$\text{cond}(x) = \frac{e^x}{e^x} \cdot x = x$$

für große x schlecht konditioniert

(4) Logarithmus $f(x) = \ln x$

$$\text{cond}(x) = \frac{\frac{1}{x}}{\ln x} \cdot x = \frac{1}{\ln x}$$

schlecht konditioniert in der Nähe von $x = 1$

Beispiel 8.4.1:

Bestimmung der Konditionszahl von

$$f(x) = \ln(x - \sqrt{x^2 - 1}), \quad x = 30$$

$$f'(x) = \frac{1}{x - \sqrt{x^2-1}} \cdot \left(1 - \frac{2x}{2\sqrt{x^2-1}}\right) = \frac{1}{(x - \sqrt{x^2-1})} \cdot \frac{(\sqrt{x^2-1} - x)}{\sqrt{x^2-1}}$$

$$= -\frac{1}{\sqrt{x^2-1}}$$

$$|\text{cond}(x)| = \left| \frac{x}{\sqrt{x^2-1} \cdot \ln(x - \sqrt{x^2-1})} \right| \Rightarrow |\text{cond}(30)| \approx 0.244$$

für $x = 30$ gut konditioniert

Beispiel 8.4.2:

Bestimmung der Konditionszahl von

$$f(x) = 1 - \sqrt{1 - x^2}$$

Auslöschung: Wenn man kleine Zahlen als Ergebnis einer Differenz von großen Zahlen berechnet. Rundungsfehler werden beim weiterrechnen verstärkt!

$$f'(x) = -\frac{-2x}{2\sqrt{1-x^2}}$$

$$\text{cond}(x) = \frac{x^2}{\sqrt{1-x^2}\left(1-\sqrt{1-x^2}\right)} = \frac{x^2}{\sqrt{1-x^2}-(1-x^2)} = \frac{x^2}{\underbrace{\left(\sqrt{1-x^2}-1\right)}_{\approx 0 \text{ für } |x| \text{ klein}}+x^2}$$

$\text{cond}(x) \approx 1$ für $|x|$ klein

Was ergibt sich für Werte von x, die nahe bei 0 liegen?

Dort ist das Problem gut konditioniert

Aber: Es tritt der Effekt der "Auslöschung" auf.

Man versucht eine andere Formulierung zu finden, die numerisch günstiger ist.

Trick: Erweitern unter Verwendung der 3. binom. Formel

$$1-\sqrt{1-x^2} = \frac{(1-\sqrt{1-x^2})\cdot(1+\sqrt{1-x^2})}{1+\sqrt{1-x^2}} =$$

$$= \frac{1-(1-x^2)}{1+\sqrt{1-x^2}} = \frac{x^2}{1+\sqrt{1-x^2}}$$

numerisch stabil !

Stabilität

Ein Algorithmus heißt stabil, wenn das Ergebnis nicht empfindlich ist gegenüber numerischen Berechnungsfehlern (z. B. Rundungsfehlern).

Ein stabiler Algorithmus liefert ein Ergebnis, das nur wenig vom exakten Ergebnis abweicht.

Instabile Algorithmen sind nicht geeignet für numerische Berechnungen.

Beispiel 8.4.3:

Finden Sie einen äquivalenten, numerisch stabilen Ausdruck für

$$f(x) = 1 - \sqrt{1 - x^2}$$

s. oben