Numerische Analysis I
Vorlesung 2

Arnold Reusken

Hauke Saß, Stephanie Schwab

IGPM

Wintersemester 2022-2023

Übersicht

Themen: Dahmen & Reusken Kap 2.4-2.5

- Grundlagen: Taylor-Entwicklung (Wiederholung)
- Kondition eines Problems
- Rundungsfehler und Gleitpunktarithmetik

Was Sie mitnehmen sollten:

- Was ist die (relative) Kondition eines Problems?
- Wie kann (in bestimmten Fällen) die Kondition eines Problems berechnet werden?
- Wie sind die elementaren Rechenoperationen konditioniert?
- Wie ist die Menge der Maschinenzahlen definiert?
- Was ist Auslöschung?

Taylor-Entwicklung: Skalare Funktionen

Taylor-Polynom vom Grad $k-1$ in x_0

$$p_{k-1}(x) = f(x_0) + f'(x_0)\,(x-x_0) + \frac{f^{(2)}(x_0)}{2}(x-x_0)^2 + \ldots + \frac{f^{(k-1)}(x_0)}{(k-1)!}(x-x_0)^{k-1}.$$

- Für $k=1$ erhält man den Mittelwertsatz
$$\frac{f(x)-f(x_0)}{x-x_0} = f'(\xi),$$
wobei ξ eine Zahl zwischen x und x_0 ist.
- Oft verwendete Darstellung
$$f(x) = p_{k-1}(x) + \mathcal{O}(|x-x_0|^k) \qquad (x \to x_0)$$

Siehe auch Matlab-Demo 2.23.

Taylor-Entwicklung: Vektorwertige Funktionen

Kompakte Schreibweise

$$
\begin{aligned}
f(\tilde{x}) &= f(x) + (\nabla f(x))^T (\tilde{x} - x) \\
&\quad + \tfrac{1}{2}(\tilde{x} - x)^T f''(x)(\tilde{x} - x) + \mathcal{O}(\|\tilde{x} - x\|_2^3).
\end{aligned}
$$

oder

$$f(\tilde{x}) = f(x) + (\nabla f(x))^T (\tilde{x} - x) + \mathcal{O}(\|\tilde{x} - x\|_2^2).$$

Falls $\|\tilde{x} - x\| \ll 1$:

$$f(\tilde{x}) \doteq f(x) + (\nabla f(x))^T (\tilde{x} - x)$$

$\doteq$: Terme höherer Ordnung werden vernachlässigt.

Fehlerquellen

Fehler im Resultat auf Grund von

- Datenfehlern (oder Eingabefehlern)

  $\Rightarrow$ Kondition eines Problems

  – können häufig nicht vermieden werden

- Fehler(akkumulation) im Algorithmus (z.B. Rundungsfehler)

  $\Rightarrow$ Stabilität eines Algorithmus

  – kann man beeinflussen durch Anpassung des Verfahrens

Begriff der Kondition

Ungestörtes Problem

$$\underbrace{\boldsymbol{x}}_{\text{Eingabedaten}} \in \boldsymbol{X} \xrightarrow[\boldsymbol{f}]{\boxed{\text{Problem, Prozess}}} \underbrace{\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})}_{\text{Ausgabedaten}} \in \boldsymbol{Y}$$

Gestörtes Problem

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{\Delta x} \xrightarrow[\boldsymbol{f}]{\boxed{\text{Problem, Prozess}}} \tilde{\boldsymbol{y}} = \boldsymbol{f}(\tilde{\boldsymbol{x}})$$

mit Eingabefehler $\boldsymbol{\Delta x} = \tilde{\boldsymbol{x}} - \boldsymbol{x}$

Ausgabefehler $\boldsymbol{\Delta y} = \tilde{\boldsymbol{y}} - \boldsymbol{y} = \boldsymbol{f}(\tilde{\boldsymbol{x}}) - \boldsymbol{f}(\boldsymbol{x})$

Ziel: Verhältnis Ausgabefehler $\boldsymbol{\Delta y}$ zu Eingabefehler $\boldsymbol{\Delta x}$.

Kondition: $f : \mathbb{R} \to \mathbb{R}$

Taylor-Entwicklung 1. Ordnung von f um festes x

$$f(\tilde{x}) \doteq f(x) + f'(x)\,(\tilde{x} - x)$$

Daraus erhält man die Kondition für

- $f : \mathbb{R} \to \mathbb{R}$ (Eingabe: Skalar, Ausgabe: Skalar)

$$\left|\frac{f(\tilde{x}) - f(x)}{f(x)}\right| \doteq \kappa_{\mathrm{rel}}(x) \left|\frac{\tilde{x} - x}{x}\right|$$

mit

$$\kappa_{\mathrm{rel}}(x) := \left|f'(x)\frac{x}{f(x)}\right|$$

Kondition: $f : \mathbb{R}^n \longrightarrow \mathbb{R}$

Für $f : \mathbb{R}^n \to \mathbb{R}$ lautet die Taylor-Reihenentwicklung 1. Ordnung

$$f(\tilde{x}) \doteq f(x) + (\nabla f(x))^T \cdot (\tilde{x} - x)$$

mit

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n.$$

Hieraus folgt

$$\frac{f(\tilde{x}) - f(x)}{f(x)} \doteq \sum_{j=1}^{n} \left(\frac{\partial f(x)}{\partial x_j} \cdot \frac{x_j}{f(x)} \right) \cdot \frac{\tilde{x}_j - x_j}{x_j}$$

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

Mit den Verstärkungsfaktoren

$$\phi_j(x) = \frac{\partial f(x)}{\partial x_j} \cdot \frac{x_j}{f(x)}$$

erhält man

$$\underbrace{\frac{f(\tilde{x}) - f(x)}{f(x)}}_{\substack{\text{relativer Fehler}\\ \text{der Ausgabe}}} \doteq \sum_{j=1}^{n} \underbrace{\phi_j(x)}_{\substack{\text{Fehler-}\\ \text{verstärkung}}} \cdot \underbrace{\frac{\tilde{x}_j - x_j}{x_j}}_{\substack{\text{relativer Fehler}\\ \text{der Eingabe in } x_j}}$$

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

Damit erhält man die Kondition für

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

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

mit

$$\kappa_{\mathrm{rel}}(x) = \kappa_{\mathrm{rel}}^{\infty}(x) := \max_j |\phi_j(x)|$$

und den Verstärkungsfaktoren

$$\phi_j(x) = \frac{\partial f(x)}{\partial x_j} \cdot \frac{x_j}{f(x)}$$

wobei $\dot{\leq}$ entsprechend $\doteq$ zu verstehen ist.

Beispiel 2.26.

Gegeben sei

$$f : \mathbb{R} \to \mathbb{R}, \quad f(x) = e^{3x^2}.$$

Relative Konditionszahl:

$$\kappa_{\mathrm{rel}}(x) = \left| f'(x) \frac{x}{f(x)} \right| = 6\,x^2.$$

$\rightsquigarrow$ für $|x|$ klein/groß ist f gut/schlecht konditioniert.

Beispiel

- $x = 0.1,\ \tilde{x} = 0.10001$: $\kappa_{\mathrm{rel}}(0.1) = 6 \cdot 10^{-2}$

$$\left|\frac{\tilde{x}-x}{x}\right| = 10^{-4} \to \left|\frac{f(x)-f(\tilde{x})}{f(x)}\right| = 6.03 \cdot 10^{-6}$$

- $x = 4,\ \tilde{x} = 4.0004$: $\kappa_{\mathrm{rel}}(4) = 96$

$$\left|\frac{\tilde{x}-x}{x}\right| = 10^{-4} \to \left|\frac{f(x)-f(\tilde{x})}{f(x)}\right| = 9.65 \cdot 10^{-3}$$

Elementare Rechenoperationen

Kondition bei

- Multiplikation: $x = (x_1, x_2)^T$, $f(x) = x_1 \, x_2$

$$\kappa_{\text{rel}}(x) = 1 \text{ (von } x \text{ unabhängig!)}$$

  Multiplikation für alle Eingangsdaten gut konditioniert.
  Ein ähnliches Resultat gilt für die Division.

- Addition: $x = (x_1, x_2)^T$, $f(x) = x_1 + x_2$

$$\kappa_{\text{rel}}(x) = \max\left\{\left|\frac{x_1}{x_1 + x_2}\right|, \left|\frac{x_2}{x_1 + x_2}\right|\right\}$$

  Bei zwei Zahlen mit gleichem Vorzeichen: $\kappa_{\text{rel}} \leq 1$.

  ABER: $\kappa_{\text{rel}}(x) \gg 1$ wenn $x_1 \approx -x_2$.

Beispiel 2.29 (Nullstelle)

Bestimmung der kleineren Nullstelle y^* von $y^2 - 2x_1 y + x_2 = 0$:

$$x = (x_1, x_2)^T, \qquad y^* = f(x) = x_1 - \sqrt{x_1^2 - x_2}$$

- Partielle Ableitungen

$$\frac{\partial f(x)}{\partial x_1} = \frac{\sqrt{x_1^2 - x_2} - x_1}{\sqrt{x_1^2 - x_2}} = \frac{-y^*}{\sqrt{x_1^2 - x_2}}$$

$$\frac{\partial f(x)}{\partial x_2} = \frac{1}{2\sqrt{x_1^2 - x_2}}$$

- Verstärkungsfaktoren

$$\phi_j(x) = \frac{\partial f(x)}{\partial x_j} \cdot \frac{x_j}{f(x)}$$

Beispiel 2.29 (Nullstelle)

- Verstärkungsfaktoren

$$\phi_1(x) = \frac{-y^*}{\sqrt{x_1^2 - x_2}} \cdot \frac{x_1}{y^*} = \frac{-x_1}{\sqrt{x_1^2 - x_2}}$$

$$\phi_2(x) = \frac{1}{2\sqrt{x_1^2 - x_2}} \cdot \frac{x_2}{y^*} = \frac{1}{2} - \frac{1}{2}\phi_1(x)$$

- Kondition: $\kappa_{\mathrm{rel}}(x) = \max_j |\phi_j(x)|$

Kondition hängt stark von der Stelle (x_1, x_2) ab:

- Wenn $x_2 < 0$: $|\phi_1(x)| \leq 1$ und $\kappa_{\mathrm{rel}}(x) \leq 1$
- Wenn $x_2 \approx x_1^2$: $|\phi_1(x)| \gg 1$ und $\kappa_{\mathrm{rel}}(x) \gg 1$

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

Sei $f = B \in \mathbb{R}^{n \times n}$, mit $\det B \neq 0$, also

$$y = f(x) = Bx$$

bzw. für gestörte Daten

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

und damit

$$f(\tilde{x}) - f(x) = B\tilde{x} - Bx = B(\tilde{x} - x)$$
$$x = B^{-1}y$$

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

Wegen $\|x\| = \|B^{-1}y\| \le \|B^{-1}\|\,\|y\|$ gilt

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

wobei

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

die Konditionszahl der Matrix B ist.

Beachte:
$\kappa(B) = \kappa(B^{-1})$ hängt nur von der Matrix B (und der Norm $\|\cdot\|$) ab.

Beispiel 2.34.

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 $u = A^{-1}b$ mit

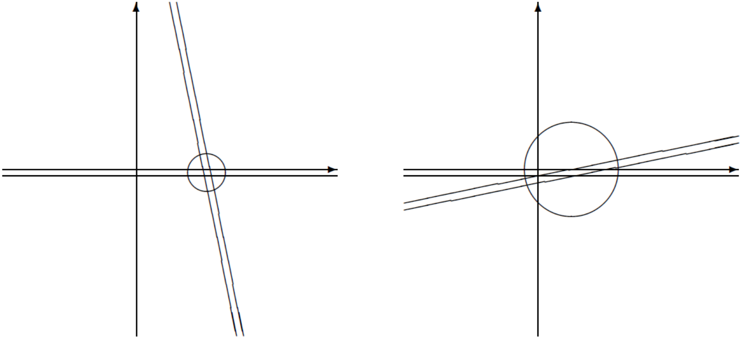
$$A = \begin{pmatrix} 3 & 1.001 \\ 6 & 1.997 \end{pmatrix}, \qquad b = \begin{pmatrix} 1.999 \\ 4.003 \end{pmatrix}.$$

Die Lösung ist

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

Also: $f : \mathbb{R}^2 \to \mathbb{R}^2$, $f(x) = A^{-1}x$.

Beispiel 2.34. Kondition bei Bestimmung eines Schnittpunktes



Beispiel 2.34.

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

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

Man erhält

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

Wir betrachten die Maximumnorm:

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

Beispiel 2.34.

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

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

Kondition einer Basis

Sei V ein linearer normierter Raum mit Basis $\Phi = \{\phi_1, \ldots, \phi_n\}$. Die Koordinaten-Abbildung ist durch

$$\mathcal{L} : \mathbb{R}^n \to V, \qquad \mathcal{L}(a) = \sum_{i=1}^{n} a_i \phi_i,$$

gegeben.

Kondition

Es gilt

$$\min\left\{ C/|c| \;\middle|\; |c|\,\|a\| \le \Big\| \sum_{j=1}^{n} a_j \phi_j \Big\|_V \le C\|a\| \;\; \forall\, a \in \mathbb{R}^n \right\}$$
$$= \kappa(\mathcal{L}) = \|\mathcal{L}\|_{\mathbb{R}^n \to V} \|\mathcal{L}^{-1}\|_{V \to \mathbb{R}^n}$$

Kondition einer Basis

Gram-Matrix

Annahmen: $\|\cdot\|_V$ entspricht $(\cdot,\cdot)_V$, und $\|\cdot\| := \|\cdot\|_2$ auf $\mathbb{R}^n$.

$$\text{Gram-Matrix:} \quad G_{i,j} := (\phi_i, \phi_j)_V, \quad 1 \leq i,j \leq n,$$

definiert. Es gilt $\kappa(\mathcal{L}) = \sqrt{\kappa_2(G)}$.

Beispiel

Sei $V = \Pi_m$, mit Dimension $n := m+1$, Skalarprodukt $(f,g)_V := \int_0^1 f(t)g(t)\,dt$ und die Basis $\phi_i(t) = t^{i-1}$, $i = 1, \ldots n$.

$$G_{i,j} = \int_0^1 \phi_i(t)\phi_j(t)\,dt = \int_0^1 t^{i+j-2}\,dt = \frac{1}{i+j-1}$$

Diese Hilbert-Matrix hat eine (sehr) große Konditionszahl.

Zahlendarstellungen: Beispiel 2.40.

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

- Dezimalsystem (Basis 10)

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

- Binärsystem (Basis 2)

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

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 $b = 2$)

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

- Dezimalsystem (Basis $b = 10$)

$$\frac{1}{3} \Rightarrow 0.33333\ldots.. \cdot 10^0$$

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

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

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 – Beispiel

Gleitpunktdarstellung: $b = 10$, $m = 6$

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: $b = 2$, $m = 10$

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}	$0.1010110000 * 2^{1111}$	$4.8 * 10^{-4}$
$\frac{1}{10}$	$0.1100110011 * 2^{-11}$	$2.4 * 10^{-4}$

Dahmen & Reusken

Maschinengenauigkeit

- Für den relativen Rundungsfehler erhält man
$$\left|\frac{\mathrm{fl}(x)-x}{x}\right| \le \frac{\frac{b^{-m}}{2}b^e}{b^{-1}b^e} = \frac{b^{1-m}}{2}.$$
- Die (relative) Maschinengenauigkeit
$$\text{eps} := \frac{b^{1-m}}{2}$$
charakterisiert das Auflösungsvermögen des Rechners, d.h.
$$\text{eps} = \inf\{\delta > 0 \mid \mathrm{fl}(1+\delta) > 1\}$$
- Der Rundungsfehler ε erfüllt $|\varepsilon| \le \text{eps}$ und es gilt
$$\mathrm{fl}(x) = x\,(1+\varepsilon).$$

Maschinengenauigkeit – Beispiel

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

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

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

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

Dahmen & Reusken

Gleitpunktarithmetik

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

Beispiel

$b = 10,\ m = 3$:

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

Ähnliches passiert bei Multiplikation und Division.

Exakte Arithmetik $\rightsquigarrow$ Gleitpunktarithmetik (Pseudoarithmetik),

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

Gleitpunktarithmetik

Forderung

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

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

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

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

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

Vorsicht bei 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

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

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

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

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.