2. Großübung

1 Normalisierte Gleitpunktdarstellung

1.1 Darstellung:

$$x = f \cdot b^e, \tag{1}$$

wobei x die Maschinenzahl, f die Mantisse, b die Basis und e der Exponent ist. Mit

$$f = \pm 0.\underset{\neq 0}{d_1} d_2 ... d_m, \; 0 \leq d_j \leq b-1, \; r \leq e \leq R \tag{2}$$

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

Beispiel:

Sei $\mathbb{M}(10, 3, -4, 4)$

$$'123' = 0.123 \cdot 10^3 \tag{3}$$

betragsmäßig kleinste Zahl: $0.100 \cdot 10^{-4} =' 0.00001'$
betragsmäßig größte Zahl: $0.999 \cdot 10^4 =' 9990'$

1.2 Menge der Maschinenzahlen

betragsmäßig kleinste Zahl:

$$0.10...0 \cdot b^r = b^{r-1}$$

betragsmäßig größte Zahl:

$$0.\underbrace{(b-1)}_{d_1}\underbrace{(b-1)}_{d_2}...\underbrace{(b-1)}_{d_m} \cdot b^R = (1 - b^{-m}) \cdot b^R$$

Wie viele verschiedene $x \in \mathbb{M}(b, m, r, R)$ gibt es?

- pro Mantissenstelle d_j: b Möglichkeiten $(0, 1, ..., b-1)$
- Ausnahme d_1: nur $(b-1)$ Möglichkeiten $(1, ..., b-1)$, da $d_1 \neq 0$
- $R - r + 1$ verschiedene Exponenten $(r, r+1, ..., R-1, R)$

- 2 Vorzeichen
- die Zahl Null (entspricht $d_1 = 0$)

Daraus folgt insgesamt:

$$\underbrace{1}_{x=0} + \underbrace{2}_{\pm} \cdot \underbrace{(R-r+1)}_{r \le e \le R} \cdot \underbrace{(b-1)}_{d_1 \neq 0} \cdot \underbrace{b^{m-1}}_{d_2 \dots d_m}$$

Beispiel:

Sei $\mathbb{M}(2, 53, -1022, 1023)$. Dies entspricht dem Datentyp 'double' nach IEEE 754 Standard.

- betragsmäßig kleinste Zahl $\approx 1.11 \cdot 10^{-308}$
- betragsmäßig größte Zahl $\approx 8.99 \cdot 10^{307}$
- Anzahl Maschinenzahlen $\approx 1.84 \cdot 10^{19}$

1.3 Standardrundung:

Reduktionsabbildung einer reellen Zahl in eine Maschinenzahl
Letzte Stelle d_m der Mantisse um 1 erhöhen/beibehalten, falls die nächste $[(m+1)$-ste$]$ Stelle $d_{m+1} \geq \frac{b}{2}$ bzw. $< \frac{b}{2}$

Beispiel:

$m = 2, b = 10$

$$\text{5 mal aufrunden} \begin{cases} \geq 5 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \end{cases} \quad \text{4 mal abrunden} \begin{cases} < 5 \\ (0) \\ 1 \\ 2 \\ 3 \\ 4 \end{cases}$$

Beispiel:

$m = 2, b = 3$

$$\text{1 mal aufrunden} \begin{cases} \geq' 1.5' \\ 2 \end{cases} \quad \text{1 mal abrunden} \begin{cases} <' 1.5' \\ (0) \\ 1 \end{cases}$$

1.3.1 Absoluter Rundungsfehler:

$$|fl(x) - x| \leq \frac{b^{-m}}{2} \cdot b^e = \frac{b^{e-m}}{2}$$

1.3.2 Relativer Rundungsfehler

$$\left|\frac{fl(x) - x}{x}\right| \leq \frac{\frac{b^{-m}}{2} \cdot b^e}{|x|} \leq \frac{\frac{b^{-m}}{2} \cdot b^e}{b^{-1} \cdot b^e} = \frac{b^{1-m}}{2} =: eps \text{ (relative Maschinengenauigkeit)}$$

Beispiel:

Sei $\mathbb{M}(2, 8, -1024, 1024)$ und $x = 0.2_{10}$. Dann gilt:

$$\begin{aligned}
0.2_{10} &= 0.\overline{0011}_2 \\
fl(0.2_{10}) &= fl(0.\overset{\frac{1}{2}}{0}\overset{\frac{1}{4}}{0}\overset{\frac{1}{8}}{1}\,\overset{\frac{1}{16}}{1}\,\dot{0}\dot{0}\dot{1}\dot{1}\dot{0}\dot{0}\dot{1}\dot{1}...) = 0.11001101 \cdot \underbrace{2^{-2}}_{10_2^{-10_2}} \\
|fl(x) - x| &= |(\frac{1}{8} + \frac{1}{16} + \frac{1}{128} + \frac{1}{256} + \frac{1}{1024}) - 0.2| = 1.953125 \cdot 10^{-4} \\
&\leq \frac{b^{e-m}}{2} = \frac{2^{-2-8}}{2} = 4.88... \cdot 10^{-4} \\
\left|\frac{fl(x) - x}{x}\right| &= 9.765... \cdot 10^{-4} \leq eps = \frac{b^{1-m}}{2} = \frac{2^{1-8}}{2} = 2^{-8} = 3.906.. \cdot 10^{-3}
\end{aligned}$$

2 Umrechnung Dezimalsystem$\leftrightarrow$Binärsystem

2.1 Vorkommastellen:

Beispiel:

$x = 25_{10}$

$$\begin{aligned} 25 &= 2 \cdot 12 + \underline{1} \\ 12 &= 2 \cdot 6 + \underline{0} \\ 6 &= 2 \cdot 3 + \underline{0} \\ 3 &= 2 \cdot 1 + \underline{1} \\ 1 &= 2 \cdot 0 + \underline{1} \end{aligned}$$

Die unterstrichenen Restbeträge ergeben nun von unten nach oben gelesen die gesuchte Zahl:

$$25_{10} = 11001_2 = 0.11001 \cdot 2_{10}^{5_{10}} = 0.11001 \cdot (10)_2^{101_2}$$

2.2 Nachkommastellen:

Beispiel:

$x = 0.8125_{10}$

$$\begin{aligned} 2 \cdot 0.8125 &= \underline{1}.625 \\ 2 \cdot 0.625 &= \underline{1}.25 \\ 2 \cdot 0.25 &= \underline{0}.5 \\ 2 \cdot 0.5 &= \underline{1}.0 \end{aligned}$$

Die unterstrichenen Übertrage ergeben nun von oben nach unten gelesen die gesuchte Zahl:

$$0.8125_{10} = 0.1101_2$$

2.3 Allgemeine Darstellung Binärsystem:

$$\begin{matrix} & 16 & 8 & 4 & 2 & 1 & & \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{16} & \\ ::: & * & * & * & * & * & . & * & * & * & * & ::: \end{matrix}$$

<u>Hinweis:</u> Im Minitest häufig 'einfache' Zahlen im Kopf umrechnen.

2.4 Beispiel periodische Zahl:

$x = 0.2_{10}$

$$\begin{aligned} 2 \cdot 0.2 &= \underline{0}.4 \\ 2 \cdot 0.4 &= \underline{0}.8 \\ 2 \cdot 0.8 &= \underline{1}.6 \\ 2 \cdot 0.6 &= \underline{1}.2 \\ 2 \cdot 0.2 &= \underline{0}.4 \\ 2 \cdot 0.4 &= \underline{0}.8 \\ 2 \cdot 0.8 &= \underline{1}.6 \\ 2 \cdot 0.6 &= \underline{1}.2 \end{aligned}$$

Die Zahl wird also periodisch im Binärsystem $0.2_{10} = 0.\overline{0011}_2$. Exakte Zahlen im Dezimalsystem sind bei endlicher Mantissenlänge nciht zwangsläufig auch im Binärsystem exakt darstellbar!

3 Kondition $\leftrightarrow$ Stabilität

3.1 Kondition:

Die Kondition beschreibt die max. Verstärkung des Eingabefehlers bei <u>exakter</u> Rechnung.

3.2 Pseudoarithmetik:

Siehe Folie 2.42.

$$\rightarrow x \nabla y = (x \nabla y)(1+\epsilon) \qquad\qquad |\epsilon| \leq eps$$

3.3 Stabilität:

Siehe Folie 2.47.

stabil $\rightarrow$: Fehler in der Größenordnung des unvermeidbaren Fehlers.

3.4 Rückwärtsanalyse:

Siehe Folie 2.50.

Rückwärtsanalyse ermöglicht eine Abschätzung des durch den Algorithmus bedingten Fehlers im direkten Vergleich mit dem unvermeidbaren Fehler aufgrund gestörter Eingabedaten.

<u>Vorgehensweise:</u>

- Durch Pseudoarithmetik entstehen bei jeder Rechenoperation rel. Fehler in der Größe ϵ (mit $|\epsilon| \leq eps$). Setze diese in den Algorithmus ein.
- Forme den Algorithmus so um, dass die Fehler ϵ als Störung

  $$\delta_x = \frac{\|\triangle x\|}{\|x\|}$$

  der Eingangsdaten interpretiert werden können.
- Schätze den durch die gestörten Eingangsdaten entstehenden Fehler mit Hilfe der Konditionszahl des Problems ab.

3.4.1 Beispiel:

$$f(x) = 2x_1 + x_2 \qquad\qquad x_1, x_2 \in \mathbb{M}$$

Algorithmus:

$$\begin{aligned}
y_1 &= (2x_1)(1+\epsilon_1)\\
y_2 &= (y_1 + x_2)(1+\epsilon_2)\\
\Rightarrow f(x) &= (2x_1(1+\epsilon_1) + x_2)(1+\epsilon_2)\\
&= 2x_1(1+\epsilon_1)(1+\epsilon_2) + x_2(1+\epsilon_2)\\
&= 2x_1 \underbrace{(1+\epsilon_1+\epsilon_2+\overset{\approx 0}{\epsilon_1\epsilon_2})}_{1+\delta_{x_1}} + x_2 \underbrace{(1+\epsilon_2)}_{1+\delta_{x_2}}
\end{aligned}$$

$$\begin{aligned}
|\delta_{x_2}| &= |\epsilon_2| \leq eps\\
|\delta_{x_1}| &= |\epsilon_1 + \epsilon_2| \leq 2eps
\end{aligned}$$

Jetzt Störung der Eingangsdaten.

Kondition:

$$\begin{aligned}
\kappa_{rel} &= \max_j \left| \frac{\partial f(x)}{\partial x_j} \frac{x_j}{f(x)} \right|\\
\frac{\partial f(x)}{\partial x_1} &= 2 \rightarrow \left| \frac{\partial f(x)}{\partial x_1} \frac{x_1}{f(x)} \right| = \left| \frac{2x_1}{2x_1 + x_2} \right|\\
\frac{\partial f(x)}{\partial x_2} &= 1 \rightarrow \left| \frac{\partial f(x)}{\partial x_2} \frac{x_2}{f(x)} \right| = \left| \frac{x_2}{2x_1 + x_2} \right|
\end{aligned}$$

Vergleich:

x_1, x_2 geg.

$$\begin{aligned}
\underset{Algorithmus}{F(x)} &= \left| \frac{f(\hat{x}) - f(x)}{f(x)} \right| \leq \kappa_{rel}(x) \sum_{j=1}^{2} \left| \frac{\hat{x}_j - x_j}{x_j} \right|\\
&\leq \kappa_{rel}(x) \sum_{j=1}^{2} |\delta_{x_j}| \leq \kappa_{rel}(x) \cdot 3eps
\end{aligned}$$

Annahme: Eingangsdaten höchstens mit Maschinengenauigkeitgestört: $\tilde{x}_j = x_j(1+\epsilon)$, $|\epsilon| \leq eps$

$$F(x)_{Daten} = \left| \frac{f(\tilde{x}) - f(x)}{f(x)} \right| \leq \kappa_{rel}(x) \sum_{j=1}^{2} \left| \frac{\tilde{x}_j - x_j}{x_j} \right| \leq \kappa_{rel}(x) \cdot 2eps$$

Daraus folgt der Algorithmus ist stabil, da $F(x)_{Algorithmus}$ in der Größenordnung des unvermeidbaren Datenfehlers $F(x)_{Daten}$ ist.