Plenum 04

Einführung in die Numerik
Sommersemester 2022

17.05.2022 und 19.05.2022

Zahldarstellung, Rechnerarithmetik

Was sind die Highlights der Woche?

- Fließkommadarstellung
- Beobachtung, bei Fließkommazahlen zur Basis $\beta = 2$ die führende Ziffer der normalisierten Mantisse nicht speichern zu müssen

Welche Fragen gibt es? I

- Wie sind die Darstellungen (5.2) und (5.3) zu verstehen?
- Wie ist die Darstellung des Exponenten $e = c - 1023$ im IEEE-Standard zu verstehen?
- Wie kommen im Beispiel 5.2 die darstellbaren Zahlen zustande?
- Rundungsregel *round to nearest, ties to even*
- Gesetzmäßigkeit der Verteilung der Zahlen im Fließkommagitter
- verschiedene Definitionen der Maschinengenauigkeit

Welche Fragen gibt es? II

- Rundungsfehlerabschätzung für Zahlen nahe der Null
- Quizfrage im Beweis von Lemma 5.6

Notwendigkeit der Rundung

Geben Sie Beispiele dafür an, dass die Grundoperationen $+$, $-$, $\cdot$ und $/$ selbst für Argumente in $\mathbb{F}$ im Allgemeinen Ergebnisse liefern, die keine Fließkommazahlen in $\mathbb{F}$ sind.

Verwenden Sie zum Beispiel das Fließkommasystem mit Basis $\beta = 10$ und Mantissenlänge $r = 3$.

Rundungsfehlerabschätzung

Lemma 5.6 besagt:

$$\frac{|x - \mathsf{rd}(x)|}{|x|} \leqslant \varepsilon_{\mathsf{mach}} := \frac{1}{2}\beta^{1-r}$$

für alle $x \in [-x_{\mathsf{posmax}}, -x_{\mathsf{posmin}}] \cup [x_{\mathsf{posmin}}, x_{\mathsf{posmax}}]$.

Welche Abschätzung gilt für $|x| < x_{\mathsf{posmin}}$?

Rechnen mit Rundung

Bestimmen Sie das Ergebnis der folgenden Aufgaben

- bei exakter Rechnung und
- bei Rechnung im Fließkommasystem mit Basis $\beta = 10$ und Mantissenlänge $r = 3$.

Bestimmen Sie jeweils auch den relativen Fehler.

1. $1.23 + 1.22$ und $1.23 \oplus 1.22$
2. $1.231 + 1.22$ und $\mathsf{rd}(1.231) \oplus 1.22$
3. $1.23 - 1.22$ und $1.23 \ominus 1.22$
4. $1.231 - 1.22$ und $\mathsf{rd}(1.231) \ominus 1.22$

Erklären Sie das Ergebnis.

IEEE Single Precision

Das IEEE-Format *single precision* verwendet die Basis $\beta = 2$, Mantissenlänge $r = 24$ und Exponenten $e \in [-126, 127]$. Der Speicherbedarf beträgt 32 Bit. Bestimmen Sie

- x_{posmin}
- x_{posmax}
- $\varepsilon_{\text{mach}}$

für dieses Fließkommagitter.

Lösung:

- $x_{\text{posmin}} = 2^{-126} \approx 1.2 \cdot 10^{-38}$
- $x_{\text{posmax}} = (2 - 2^{-23}) \cdot 2^{127} \approx 3.4 \cdot 10^{38}$
- $\varepsilon_{\text{mach}} = \frac{1}{2} 2^{1-24} = 2^{-24} \approx 6.0 \cdot 10^{-8}$