4.4 Maschinenzahlen

Definition 4.1. Es sei $b \in \mathbb{N}_{\geq 2}$. Für $\mathcal{E} \in \mathbb{Z}$ und $\sigma \in \{\pm 1\}$ ist eine Zahl $x \in \mathbb{R}$ der Form

$$x = \sigma \cdot \Big(\sum_{i=0}^{m-1} z_i \cdot b^{-i} \Big) \cdot b^{\mathcal{E}}$$

mit $z_i \in \{0, \ldots, b-1\}$ und $z_0 \neq 0$ eine m-stellige b-adische normalisierte Gleitkommazahl mit Mantisse $\sum_{i=0}^{m-1} z_i b^{-i}$ und Exponent $\mathcal{E}$.

Beispiel. $b = 10, m = 4$

$$x = 3,141 = (+1)(3 \cdot 10^0 + 1 \cdot 10^{-1} + 4 \cdot 10^{-2} + 1 \cdot 10^{-3}) \cdot 10^0$$
$$x = -87,3 = (-1)(8 \cdot 10^0 + 7 \cdot 10^{-1} + 3 \cdot 10^{-2}) \cdot 10^1$$

$x = \frac{1}{3}$ besitzt für $b \in \{2, 10\}$ und beliebiges m keine Darstellung als m-stellige normalisierte Gleitkommazahl.

Definition 4.2. Für $b \in \mathbb{N}_{\geq 2}, m \in \mathbb{N}_{\geq 1}$ bildet die Menge der b-adischen m-stelligen normalisierten Gleitkommazahlen mit Exponenten $\mathcal{E} \in \{\mathcal{E}_{min}, \ldots, \mathcal{E}_{max}\}$ zuzüglich der Zahl 0 den Maschinenzahlbereich $\mathcal{F}(b, m, \mathcal{E}_{min}, \mathcal{E}_{max})$

Beispiel. IEEE-Standard 754

- $b = 2$
- $m = 53$
- $\mathcal{E} \in \{-1022, -1021, \ldots, 1023\}$
- Exponentenwerte $E + 1023 \in \{0, 2047\}$ reserviert für ± 0 und $\pm\infty$.

4.5 Maschinengenauigkeit

Es sei $\mathcal{F} = F(b, m, \mathcal{E}_{min}, \mathcal{E}_{max})$ ein Maschinenzahlbereich. Eine Abbildung $rd : \mathbb{R} \to \mathcal{F}$ heißt Rundung zu F, wenn für alle $x \in \mathbb{R}$ gilt: $|x - rd(x)| = \min_{a \in \mathcal{F}} |x - a|$.

Beispiel.
- Kaufmännische Rundung
- IEEE 754: Runde im Zweifel so, dass letzte Stelle gerade wird.

Definition 4.3. Es sei $\tilde{x}$ eine Näherung von $x \in \mathbb{R}$.

(i) $|x - \tilde{x}|$ wird absoluter Fehler genannt.

(ii) $\frac{|x-\tilde{x}|}{x}$ wird relativer Fehler genannt.

Definition 4.4. Es sei $\mathcal{F}$ Maschinenzahlbereich mit Rundung rd. Die Maschinengenauigkeit von $\mathcal{F}$ ist

$$eps(F) := \sup \left\{ \left| \frac{x - rd(x)}{x} \right| \mid x \in \mathbb{R} \text{ und } |x| \in \text{ range}(\mathcal{F}) \right\},$$

mit range($\mathcal{F} := [\mathcal{F}_{min}, \mathcal{F}_{max}]$) und $\mathcal{F}_{min}(\mathcal{F}_{max})$ kleinste (größte) darstellbare positive Zahl in $\mathcal{F}$

Satz 4.5. *Für jeden Maschinenzahlbereich $\mathcal{F}(b, m, \mathcal{E}_{min}, \mathcal{E}_{max})$ mit $\mathcal{E}_{min} < \mathcal{E}_{max}$ gilt:*

$$eps(\mathcal{F}) = \frac{1}{1 + 2b^{m-1}}$$

Definition 4.6. Es sei $\mathcal{F}$ Maschinenzahlbereich und $s \in \mathbb{N}$. Dann hat $f \in \mathcal{F}$ (mindestens) s sogmofolamte Stellen in der b-adischen Gleitkommadarstellung, falls $f \neq 0$ und für jede Rundung rd und jede Zahl $x \in \mathbb{R}$ mit $rd(x) = f$ gilt:

$$|x - f| \leq \frac{1}{2} \cdot b^{\lfloor \log_b |f| \rfloor + 1 - s}$$

4.6 Maschinenbauarithmetik

Es sei $\mathcal{F}$ ein Maschinenzahlbereich und $\circ \in \{+, -, \cdot, /\}$ sei eine Operation. Problem:
Für $x, y \in \mathcal{F}$ gilt im Allgemeinen nicht $x \circ y \in \mathcal{F}$.
Pragmatische Lösung: Ersatzoperation $\circledcirc \in \{\oplus, \ominus, \odot, \oslash\}$ mit $x \circledcirc y = rd(x \circ y)$ für Rundung rd zu $\mathcal{F}$.

Beispiel. Es sei $\mathcal{F} = (10, 2, -5, 5), x = 4,5 \cdot 10^1 = 45$ und $y = 1,1 - 10^0 = 1,1$

$$x \oplus y = rd(x + y) = rd(46,1) = rd(4,61 \cdot 10^1) = 4,6 \cdot 10^1$$

Bemerkung.

- Zur Berechnung von $x \circledcirc y$ muss man $x \circ y$ nicht berechnen.
- Grundrechenarten für natürliche Zahlen reichen.
- Ist $|x \circ y| \in$ range$(\mathcal{F})$, so gilt für den relativen Fehler

$$\left| \frac{x \circ y - x \circledcirc y}{x \circ y} \right| = \left| \frac{x \circ y - rd(x \circ y)}{x \circ y} \right| \leq eps(\mathcal{F})$$

- Kommutativgesetz gilt für $\oplus, \odot$, nicht jedoch Assoziativ- und Distributivgesetz.