TU Berlin, Institut für Mathematik MA 4-5
Thomas Slawig, Uwe Prüfert

Zahlendarstellung auf dem Rechner (2)

Reelle Zahlen

Satz. Sei $b \in \mathbb{N}, b > 1$ und $z \in \mathbb{R}, |z| \leq b^n - 1, n \in \mathbb{N}$. Dann ist die Darstellung

$$z = (-1)^\nu \sum_{i=-\infty}^{n-1} z_i b^i \quad \text{mit} \quad \begin{cases} \nu \in \{0,1\}, \nu = 0 \text{ für } z = 0, \\ z_i \in \Sigma_b, i = 0, \ldots, n-1, \\ z_i < b - 1 \text{ für unendlich viele } i < n \end{cases}$$

eindeutig. Man schreibt $z = (\pm z_{n-1} \ldots z_0 . z_{-1} \ldots)_b$.

Darstellung reeller Zahlen auf dem Rechner

Gleitpunktdarstellung

$$z = (-1)^\nu m \cdot b^e, \quad \nu \in \{0,1\}$$

m : Mantisse, e: Exponent,
in dieser Form nicht eindeutig

Normalisierte Gleitpunktdarstellung

$$z = (-1)^\nu \underbrace{\sum_{i=0}^{k-1} m_i b^{-i}}_{=m} \cdot b^e, \quad \nu \in \{0,1\}, m_i \in \Sigma_b, m_0 \neq 0. \tag{1}$$

k : Stellenzahl der Mantisse. **Definition** $\mathbb{M}_N(b, k, e_{min}, e_{max})$ bezeichnet die Menge der durch (1) für $e_{min} \leq e \leq e_{max}$ darstellbaren normalisierten Gleitpunktzahlen.
Beispiel: $\mathbb{M}_N(2, 3, -1, 1)$. Darstellbar sind $\pm z$ mit z aus der folgenden Tabelle:

Mantisse/Exponent	2^{-1}	2^0	2^1
$(1.00)_2$	0.5	1	2
$(1.01)_2$	0.625	1.25	2.5
$(1.10)_2$	0.75	1.5	3
$(1.11)_2$	0.875	1.75	3.5

Definition: Die Zahl $z = 0$ wird durch $m_i = 0$ f.a. $i = 0, \ldots, k-1$ und $e = e_{min}$ definiert. Alle anderen Zahlen mit der Darstellung

$$z = (-1)^\nu \sum_{i=0}^{k-1} m_i b^{-i} \cdot b^{e_{min}}, \quad \nu \in \{0,1\}, m_i \in \Sigma_b, m_0 = 0$$

heißen subnormale Gleitpunktzahlen $\mathbb{M}_S(b, k, e_{min}, e_{max})$.
Die Menge $\mathbb{M}(b, k, e_{min}, e_{max}) := \mathbb{M}_N(b, k, e_{min}, e_{max}) \cup \{0\} \cup \mathbb{M}_S(b, k, e_{min}, e_{max})$ heißt die Menge der Gleitpunktzahlen oder Maschinenzahlen.
Beispiel: $\mathbb{M}(2, 3, -1, 1)$.

Mantisse	2^{-1}	2^0	2^1	2^{-1} (subnormal)
$(1.00)_2$	0.5	1	2	$(0.00)_2 = 0$
$(1.01)_2$	0.625	1.25	2.5	$(0.01)_2 = 0.125$
$(1.10)_2$	0.75	1.5	3	$(0.10)_2 = 0.25$
$(1.11)_2$	0.875	1.75	3.5	$(0.11)_2 = 0.375$

IEEE Standard

Auf dem Rechner gilt wegen $b = 2$ immer $m_0 = 1$. Daher wird auf den meisten Rechnern der sog. IEEE Standard verwendet. Er legt fest:

- Das führende $m_0 = 1$ wird nicht gespeichert, d.h. bei k-stelliger Mantisse:

$$z = (-1)^\nu \left(1 + \sum_{i=1}^{k} m_i 2^{-i}\right) 2^e, \quad \nu, m_i \in \{0, 1\},$$

  also $m = (1.z_1 \dots z_k)$.

- Der Exponent wird (wenn l Stellen benutzt werden) in der Form

$$e = \sum_{i=0}^{l-1} e_i 2^i - (2^{l-1} - 1), \quad e_i \in \{0, 1\}, i = 0, \dots, l-1$$

  dargestellt. Damit gilt

$$e \geq e_{min} := -(2^{l-1} - 1)$$

  und

$$e \leq e_{max} := \sum_{i=0}^{l-1} 2^i - (2^{l-1} - 1) = 2^l - 1 - (2^{l-1} - 1) = 2^l - 2^{l-1} = 2^{l-1}.$$

- Mit $e = e_{max}$ und $m_i = 0$ f.a. $i = 1, \dots, k$ werden die Werte $\pm\infty$ repräsentiert.
- Mit $e = e_{max}$ und $m_i \neq 0$ für mindestens ein $i \in \{1, \dots, k\}$ wird NaN = not a number dargestellt.
- Verteilung der zur Verfügung stehenden Stellen auf Mantisse und Exponenten für einfach und doppelt genaue Gleitpunktzahlen, s. Tabelle.

Beispiel: Mantissenlänge $k = 2$, Exponentenlänge $l = 2$, d.h.

$$e = \sum_{i=0}^{1} e_i 2^i - (2^1 - 1) = e_1 \cdot 2 + e_0 - 1$$

und daher $e_{min} = -1, e_{max} = 2$. Darstellbare Zahlen, jeweils + Vorzeichenbit:

	Exponent			
Mantisse	$(01) = 0$	$(10) = 1$	$(00) = -1$ (subnormal)	(11)
(00)	$(1.00)_2 \cdot 2^0 = 1 =: x_{min}$	$(1.00)_2 \cdot 2^1 = 2$	$(0.00)_2 = 0$	∞
(01)	$(1.01)_2 \cdot 2^0 = 1.25$	$(1.00)_2 \cdot 2^1 = 2.5$	$(0.01)_2 \cdot 2^{-1} = 0.125^*$	NaN
(10)	$(1.10)_2 \cdot 2^0 = 1.5$	$(1.00)_2 \cdot 2^1 = 3$	$(0.10)_2 \cdot 2^{-1} = 0.25$	NaN
(11)	$(1.11)_2 \cdot 2^0 = 1.75$	$(1.00)_2 \cdot 2^1 = 3.5 =: x_{max}$	$(0.11)_2 \cdot 2^{-1} = 0.375$	NaN

x_{min}: kleinste positive normalisierte Gleitpunktzahl,
x_{max}: größte darstellbare Zahl,
*: kleinste positive darstellbare Zahl (subnormale Gleitpunktzahl).
Das heißt: Im IEEE-Standard mit Mantissenlänge k und Exponentenlänge l gilt:

- $x_{min} = 1.0 \cdot 2^{e_{min}+1} = 2^{-2^{l-1}}$.
- $x_{max} = \sum_{i=0}^{k} 2^i 2^{e_{max}-1} = (2^{k+1} - 1) 2^{2^{l-1}-1}$
- kleinste darstellbare (subnormale) Zahl: $2^{-k} 2^{e_{min}} = 2^{-k} 2^{-2^{l-1}-1} = 2^{-2^{l-1}-k-1}$