

Numerik Boxen

Florian Lubitz & Steffen Hecht

28. März 2018

1 1

Die Gleitpunktzahldarstellung zerlegt jede reelle Zahl $r \in \mathbb{R}$ in drei Bestandteile:

- 1
- I) ein Vorzeichen $(-1)^v$ mit $v \in \{0, 1\}$
 - II) eine Mantisse $m \in \mathbb{R}$ und
 - III) einen Exponenten $e \in \mathbb{Z}$, sodass

$$r = (-1)^v \cdot m \cdot 2^e \quad (1.1)$$

2

$$r = (-1)^v \cdot \sum_{i=0}^{t-1} r_i \cdot 2^{-i} \text{ mit } r_0 = 1 \text{ (sodass } 1 \leq m \leq 2) \quad (1.2)$$

3

$$\begin{aligned} 13,6 &= 2^3 + 5,6 \\ &= 2^3 + 2^2 + 1,6 \\ &= \dots \\ &= 2^3 + 2^2 + 2^0 + 2^{-1} + 2^{-4} + 0,0375 \\ &= \dots \end{aligned}$$

Also können wir die Dezimalzahl $(13,6)_{10}$ als Binärzahl $(1101,1001\dots)_2$ darstellen.

4

$$\begin{aligned}(13, 6)_{10} &= (1101, 1001...)_{2} \\ &= (-1)^0 \cdot (1, 1011001...)_{2} \cdot 2^3\end{aligned}$$

also $v = 0$, $m = (1, 1011001...)_{2}$ und $e = 3$

5

Eine reelle Zahl $r \in \mathbb{R}$ mit der Darstellung

$$r = (-1)^v \cdot \sum_{i=0}^{t-1} r_i \cdot 2^{-i} \cdot 2^e \text{ mit } r_0 = 0 \quad (1.3)$$

gehört der Menge der nicht-normierten oder subnormalen Gleitpunktzahlen \mathbb{M}_s an.

6

Es gilt analog zu Definition 1.2: Das Vorzeichen $(-1)^v$ wird durch das Vorzeichen-Bit $v \in \{0, 1\}$, der Wert der Mantisse m durch die Ziffern $r_i \in \{0, 1\}$, $i = 1, \dots, t - 1$ und der Wert des Exponenten durch eine ganze Zahl $e \in \mathbb{Z}$ mit $e = e_{min}$ festgelegt.

7

$$0 = (-1)^v \cdot 0 \cdot 2^{e_{min}} \quad (1.4)$$

11

$$v + 134 = 16$$

12

$$f_{rd}(r) := \delta_r := rd(r) - r$$

13

$$f_{rel} := \epsilon_r := \frac{\delta_r}{r} = \frac{rd(r) - r}{r} \quad \text{für } r \neq 0$$

16

$$|\delta_r| \leq 2^{e-t} \quad \text{und} \quad |\epsilon_r| \leq 2^{-t}$$

17

$$\epsilon = 2^{-t}$$

18

...die entsprechenden Maschinenoperationen, d.h. die Anwendung der jeweiligen Operation auf Zahlen im normierten Gleitpunktzahlenformat nach Def. 1.2

19

- I) Verknüpfung der Maschinenzahlen mit höherer (ausreichend hoher) Genauigkeit
- II) Runden des Ergebnisses auf eine Maschinenzahl

20

$$r \circ_M s := rd(r \circ s)$$

21

$$\begin{aligned}
 r +_M s &= (1, 11)_2 \cdot 2^0 +_M (1, 10)_2 \cdot 2^2 = (111)_2 \cdot 2^{-2} + (1, 10)_2 \cdot 2^{-2} \\
 &= (1000, 10)_2 \cdot 2^{-2} \\
 &= (100010)_2 \cdot 2^1 \\
 &= (1, 00)_2 \cdot 2^1 = (2)_{10}
 \end{aligned}$$

wohingegen das exakte Ergebnis $r + s = \frac{7}{4} + \frac{3}{8} = \frac{17}{8}$ ist.

Der absolute Rundungsfehler

$$f_{rd} = (r +_M s) - (r + s)$$

und der relative

$$|f_{rel}| = \left| \frac{(r +_M s) - (r + s)}{r + s} \right| = \frac{\frac{1}{8}}{\frac{17}{8}} = \frac{1}{17}$$

Das ist relativ gut, denn der maximale Rundungsfehler bei einer 3-stelligen Mantisse ist

23

$$r \circ_M s = rd(r \circ s) = (r \circ s) \cdot (1 + \epsilon_0)$$

wobei der relative Fehler ϵ_0 der Gleitpunktoperation wegen Def. 1.10 stets durch die Maschinengenauigkeit beschränkt ist,

$$\epsilon_0 := \frac{(r \circ_M s) - (r \circ s)}{r \circ s} = \frac{rd(r \circ s) - (r \circ s)}{r \circ s} \leq \epsilon$$

24

$$\begin{aligned}
 \tilde{y} &= \tilde{x} +_M c \\
 &= (\tilde{x} + c) \cdot (1 + \epsilon_2) \\
 &= ((a +_M b) + c) \cdot (1 + \epsilon_2) \\
 &= ((a + b) \cdot (1 + \epsilon_1) + c) \cdot (1 + \epsilon_2) \\
 &= a + b + c + (a + b) \cdot \epsilon_1 + (a + b + c) \cdot \epsilon_2 + (a + b) \cdot \epsilon_1 \epsilon_2
 \end{aligned}$$

25

$$\tilde{y} \doteq a + b + c + (a + b) \cdot \epsilon_1 + (a + b + c) \cdot \epsilon_2$$

26

$$\begin{aligned} f_{rel}(y) &= \frac{\tilde{y} - y}{y} = \frac{(a + b) \cdot \epsilon_1 + (a + b + c) \cdot \epsilon_2}{a + b + c} \\ &= \frac{a + b}{a + b + c} \cdot \epsilon_1 + \epsilon_2 \end{aligned}$$

27

$$\begin{aligned} |f_{rel}(y)| &= \left| \frac{a + b}{a + b + c} \cdot \epsilon_1 + \epsilon_2 \right| \\ &\stackrel{(D.U.G)}{\leq} \left| \frac{a + b}{a + b + c} \right| \cdot |\epsilon_1| + |\epsilon_2| \end{aligned}$$

28

$$|f_{rel}(y)| \leq \left(1 + \left| \frac{a + b}{a + b + c} \right| \right) \cdot \epsilon = \left(a + \frac{1}{\left|1 + \frac{c}{a+b}\right|}\right) \cdot \epsilon$$

29

$$c \approx -(a + b)$$

ist. Denn für $c \rightarrow -(a + b)$ ist $\frac{|a+b|}{|a+b+c|} \rightarrow \infty$ und damit wird auch die obere Schranke von $|f_{rel}(y)|$ beliebig (unendlich) groß. In diesem Fall spricht man von **Auslöschung**