

Numerik Boxen

Florian Lubitz & Steffen Hecht

9. April 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

$$r \circ_{\mathbb{M}} s := rd(r \circ s)$$

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

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

Der absolute Rundungsfehler ist also

$$f_{rd} = (r +_{\mathbb{M}} s) - (r + s) = 2 - \frac{17}{8} = -\frac{1}{8}$$

21

und der relative Rundungsfehler ist

$$\begin{aligned}
|f_{rel}| &= \left| \frac{(r +_{\mathbb{M}} s) - (r + s)}{r + s} \right| \\
&= \frac{\frac{1}{8}}{\frac{17}{8}} \\
&= \frac{1}{17} \\
&= 0,0588... \approx 5,88\%
\end{aligned}$$

Das ist relativ gut, denn der maximale Rundungsfehler bei einer 3-stelligen Mantisse ist $\epsilon = 2^{-3} = 12,5\%$.

$$r \circ_{\mathbb{M}} s = rd(r \circ s) = (r \circ s) \cdot (1 + \epsilon_0)$$

23

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

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

24

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

25

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

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

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

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**

30-36

$$y + \delta_y = f(x + \delta_x) = f(x) + f'(x)\delta_x + O(\delta_x^2)$$

Unter Vernachlässigung der Terme höherer Ordnung gilt also:

$$\delta_y \doteq f'(x) \cdot \delta_x$$

37

Die Verstärkung des relativen Rundungsfehlers ϵ_y des Resultats y ergibt sich dann für $x, y \neq 0$:

$$\epsilon_y = f_{rel}(y) = \frac{\delta_y}{y} \doteq \frac{f'(x) \cdot \delta_x}{y} = \frac{x f'(x)}{y} \cdot \frac{\delta_x}{x} = \frac{x \cdot f'(x)}{y} \cdot \epsilon_x \quad (1.12)$$

aus dem relativen Rundungsfehler ϵ_x der Eingabe x .

... aus dem Verstärkungsfaktor

38

$$K_f(x) := \left| \frac{x \cdot f'(x)}{f(x)} \right|$$

zwischen den relativen Rundungsfehlern aus (1.12)

2 Lineare Gleichungssysteme



40

$$\begin{aligned}10x_1 + 7x_2 + 0x_3 &= 7 \\2,5x_2 + 5x_3 &= 2,5 \\6,2x_3 &= 6,2\end{aligned}$$

... oder in Matrixschreibweise:

$$\begin{pmatrix} 10 & -7 & 0 \\ 0 & 2,5 & 5 \\ 0 & 0 & 6,2 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 7 \\ 2,5 \\ 6,2 \end{pmatrix}$$

41

$$\begin{aligned}x_3 &= \frac{6,2}{6,2} = 1 \\x_2 &= \frac{2,5 - 5x_3}{2,5} = \frac{2,5 - 5 \cdot 1}{2,5} = -1 \\x_1 &= \frac{7 - 0x_2 + 7x_3}{10} = \frac{7 - 0 - 7}{10} = 0\end{aligned}$$

Somit resultiert der Lösungsvektor $\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}$

42

I) Berechnung $x_n = \frac{b_n}{a_{nn}}$

II) Schrittweise Berechnung :

Für $i = n - 1, n - 2, \dots, 3, 2, 1$ ist $x_i = \frac{b_i - \sum_{j=i+1}^n a_{ij} \cdot x_j}{a_{ii}}$

43

$$\begin{pmatrix} a_{11} & a_{12} & a_{1n} \\ a_{21} & a_{22} & a_{2n} \\ \vdots & & \\ a_{n1} & a_{n2} & a_{nn} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

44

$$\begin{pmatrix} 10 & -7 & 0 \\ -3 & 2 & 6 \\ 5 & -1 & 5 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 7 \\ 4 \\ 6 \end{pmatrix}$$

45

$$\begin{pmatrix} 10 & -7 & 0 \\ 0 & -0.1 & 6 \\ 5 & -1 & 5 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 7 \\ 6, 1 \\ 6 \end{pmatrix}$$

46

$$\begin{pmatrix} 10 & -7 & 0 \\ 0 & -0.1 & 6 \\ 0 & 2, 5 & 5 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 7 \\ 6, 1 \\ 2, 5 \end{pmatrix}$$

47

$$\begin{pmatrix} 10 & -7 & 0 \\ 0 & -0.1 & 6 \\ 0 & 0 & 155 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 7 \\ 6, 1 \\ 155 \end{pmatrix}$$