Technische Universität München
Institut für Informatik
Prof. Dr. Daniel Cremers
Dr. Frank Schmidt
Nikola Tchipev
Michael Rippl

WiSe 2016 / 2017

Numerisches Programmieren, Übungen

1. Übungsblatt: Zahlendarstellung, Rundungsfehler

Auf diesem Übungsblatt schauen wir uns die Vorteile von Gleitkommazahlen gegenüber Festkommazahlen und den IEEE-Standard an. Binär Gleitkommazahlen haben das folgende format:

$$(-1)^{Vorzeichen} \cdot 2^{Exponent} * Mantisse.$$

1) Ganzzahlen, Fest- und Gleitkommazahlen

Wir vergleichen jetzt drei Arten in den wir Zahlen mit 8 Bits speichern können:

- I: **Ganzzahlen** - 1 Bit für Vorzeichen (0 für "+", 1 für "-") und 7 Bits für den Ganzteil (in dieser Reihenfolge),
- F: **Festkommazahlen** - 1 Bit für Vorzeichen, 4 Bits für den Ganzteil und 3 Bits für die Nachkommastellen (in dieser Reihenfolge),
- G: **Gleitkommazahlen** - 1 Bit für Vorzeichen, 5 Bits für den Exponenten und 2 Bits für die Mantisse (in dieser Reihenfolge).
  In diese Aufgabe betrachten wir den Exponent als vorzeichenbehaftete Ganzzahl.
  Bei der Mantisse wird von einer Normalisierung mit führender Eins ausgegangen, die nicht gespeichert wird.

Aufgabe: Füllen Sie die fehlende Einträge der Tabelle aus:

Nr.	Frage		$A = I$	$A = F$	$A = G$
1	Darstellungs-beispiele	1 0000000	-0	-0,0	$-1 = -2^{+0} \cdot 2^0$
2		0 0000000	+0		
3		0 1000000	64		
4		0 1000100	68		
5		0 1000110	70	8,75	
6		0 1000111	71	8,875	$0,875 = 2^{-1} \cdot (2^0 + 2^{-1} + 2^{-2})$

In Aufgaben Nr. 22-29 wird die Rundungsfunktion $\mathrm{rd}_A : \mathbb{R} \to A$ gebraucht. Für jedes $A \subset \mathbb{R}$, ist sie definiert wie folgt:

$$\mathrm{rd}_A(x) = a \in A, \text{ so dass } |x - a| \leq |x - b| \ \forall b \in A. \tag{1}$$

Zum Beispiel $\mathrm{rd}_F(\pi) = 3,125$.

Nr.	Frage		$A = I$	$A = F$	$A = G$
7	Allg. Eigenschaften	$\lvert A\rvert$	256		
8		$\max_{a\in A} a$	127		
9		$\min_{a\in A} a$	-127		
10		$\min_{a\in A, a>0} a$	1		
11		Ist 0 in A?	ja		
12	Anzahl Zahlen von A in	$[2^{-15}, 2^{-14})$	0		
13		$[1, 2)$	1		
14		$[2, 4)$	2		
15		$[4, 8)$	4		
16		$[8, 16)$	8		
17		$[16, 32)$	16		
18		$[32, 64)$	32		
19		$[64, 128)$	64		
20		$[2^{15}, 2^{16})$	0		
21		$[0, 1)$	1		
22	Rundungen: $\mathrm{rd}_A(x)$	3^{-5}	0		2^{-8}
23		2,1	2		
24		3,1	3		
25		9	9		
26		18	18	15,875	
27		1023	127	15,875	
28	Absoluter Rundungsfehler: $\lvert x - \mathrm{rd}_A(x)\rvert$	3^{-5}	3^{-5}		
29		2,1	0,1		
30		3,1	0,1		
31		9	0		
32		18	0		
33		1023	...	...	
34	Relativer Rundungsfehler: $\left\lvert\frac{x-\mathrm{rd}_A(x)}{x}\right\rvert$	3^{-5}	1		
35		2, 1	0,048		
36		3, 1	0,032		
37		9	0		
38		18	0		
39		1023	...	...	

Beobachtungen

a) Betrachten Sie Ihre Antworten zu Fragen Nr. 2 und 3 für $A = G$. Was ist zu beobachten?

b) Betrachten Sie Ihre Antworten zu Fragen Nr. 25 und 26 für $A = G$. Was ist zu beobachten?

c) Betrachten Sie Ihre Antwort zu Frage Nr. 25 für $A = G$. In manche Programmiersprachen (z.B. JavaScript) gibt es kein Format für Ganzzahlen. Welche Konsequenzen kann das haben?

d) Was ist die kleinste Ganzzahl, die bei 32-Bit IEEE Gleitkommazahlen nicht exakt dar-

stellbar ist?

e) Betrachten Sie die Antworten zu den Fragen 26.-27. für $A = F$. Welche Konsequenzen kann die Abwesenheit einer Inf-Darstellung haben?

f) Für eine Mantissendarstellung mit 2 Bits ist die Maschinengenauigkeit $\varepsilon_{Ma} = 2^{-3}$. Vergleichen Sie ihre Antworten zu Fragen Nr. 34-39 mit der Maschinengenauigkeit. Ist

$$\left| \frac{x - \mathrm{rd}_G(x)}{x} \right| \leq \varepsilon_{Ma} \tag{2}$$

immer erfüllt?

g) Gibt es eine Relation zwischen der Maschinengenauigkeit und der Zahl aus Frage Nr. 10?

h) Was ist der Anteil an Zahlen im $[0, 1)$ beim Format G (ungefähr)? Bei 32-Bit IEEE?

i) Wie viele Zahlendarstellungen sind für spezielle Werte (Exponentenkombinationen $00\ldots0$ und $11\ldots1$) bei IEEE reserviert?

j) Gibt es einen Unterschied zwischen I und F? Betrachten Sie Ihre Antworten zu Fragen 12-21 für $A = I$ und $A = F$. Kann man F anhand I implementieren?

k) Die darstellbare Zahlen bei I und F sind *linear* verteilt. Wie sind die Zahlen G verteilt?

l) Was hängt von der Anzahl an Bits in der Mantisse ab?

m) Was hängt von der Anzahl an Bits in dem Exponent ab?

Der einfachen Format G ist leider an mehrere Stellen mehrdeutig. Alle Mehrdeutigkeiten müssen in einem Standard wie den IEEE-Standard gelöst werden um Konsistenz und Reproduzierbarkeit zu garantieren.

2) Gleitkomma-Zahlen im IEEE-Standard

Wir betrachten zunächst den 32-Bit IEEE-Standard:

- Das erste Bit bestimmt das **Vorzeichen**: Eine Null bedeutet positive Zahl, eine Eins bedeutet negative Zahl.
- Die nächsten 8 Bits sind für den **Exponenten** reserviert. Die gespeicherte Binärzahl entspricht dem Exponenten plus 127. Die Bit-Kombinationen 00000000 (entspräche einem Exponenten von -127) und 11111111 (entspräche $+128$) sind allerdings für spezielle Werte reserviert (0, Inf, NaN).
- Die letzten 23 Bits dienen der Speicherung der **Nachkommastellen**. Dabei wird von einer Normalisierung mit führender Eins ausgegangen (die nicht gespeichert werden muss). Genügen 23 Bits Genauigkeit für die Mantisse nicht, so wird zum nächstgelegenen darstellbaren Wert gerundet. Beispiele:
  - $x = 1,0|\mathbf{101} \rightarrow$ Aufrunden $(1,1)$
  - $x = 1,0|\mathbf{0}11 \rightarrow$ Abrunden $(1,0)$
  - Uneindeutiger Fall $|x_t x_{t+1} x_{t+2} \ldots = 100\ldots$
    - $x = 1,\mathbf{1}|100 \rightarrow$ Aufrunden $(10,0)$
    - $x = 1,\mathbf{0}|100 \rightarrow$ Abrunden $(1,0)$

Aufgabe:
Wandeln Sie die Zahl $-\frac{11}{10} = -1,0\overline{0011}_2 \cdot 2^0$ in 32-Bit IEEE Gleitkommazahl um!

3) Assoziativgesetz

Eine Eigenschaft, die bei Gleitkommazahlen leider verloren wird, ist das Assoziativgesetz. Betrachten Sie eine binäre Gleitkomma-Darstellung mit 4 signifikanten Stellen (Bsp: -0.01011_2, 10.10_2).

Berechnen Sie im gegebenen Format die Werte

- $(-8 + 11) + 0.75$ und
- $-8 + (11 + 0.75)$.

Runden Sie dabei nach jedem Rechenschritt. Was ist zu beobachten?