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

WiSe 2016 / 2017

Numerisches Programmieren, Übungen

Musterlösung 1. Übungsblatt: Zahlendarstellung, Rundungsfehler

1) Ganzzahlen, Fest- und Gleitkommazahlen

Nr.	Frage		$A = I$	$A = F$	$A = G$
1	Darstellungsbeispiele	1 0000000	-0	-0,0	$-2^{+0} \cdot 2^0$
2		0 0000000	+0	+0,0	$2^{+0} \cdot 2^0$
3		0 1000000	64	8,0	$2^{-0} \cdot 2^0$
4		0 1000100	68	8,5	$2^{-1} \cdot 2^0$
5		0 1000110	70	8,75	$0,75 = 2^{-1} \cdot (2^0 + 2^{-1})$
6		0 1000111	71	8,875	$0,875 = 2^{-1} \cdot (2^0 + 2^{-1} + 2^{-2})$
7	Allg. Eigenschaften	$\lvert A \rvert$	256	256	256
8		$\max_{a \in A} a$	127	15,875	$57344 = 2^{15} \cdot (2^0 + 2^{-1} + 2^{-2})$
9		$\min_{a \in A} a$	-127	-15,875	-57344
10		$\min_{a \in A, a>0} a$	1	0,125	$2^{-15} \cdot 2^0$
11		Ist 0 in A?	ja	ja	nein
12	Anzahl Zahlen von A im	$[2^{-15}, 2^{-14})$	0	0	4
13		$[1, 2)$	1	8	4
14		$[2, 4)$	2	16	4
15		$[4, 8)$	4	32	4
16		$[8, 16)$	8	64	4
17		$[16, 32)$	16	0	4
18		$[32, 64)$	32	0	4
19		$[64, 128)$	64	0	4
20		$[2^{15}, 2^{16})$	0	0	4
21		$[0, 1)$	1	8	$60 = (2^4 - 1) \cdot 2^2$

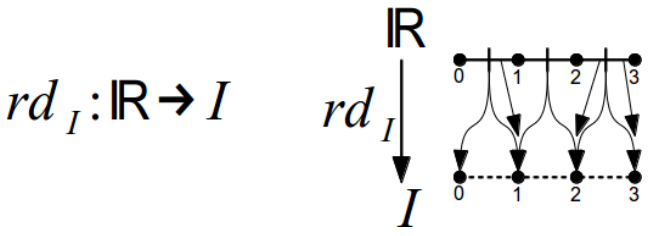

Abb. 1: Veranschaulichung von rd_I.

Nr.	Frage		$A = I$	$A = F$	$A = G$
22	Rundungen: $\mathrm{rd}_A(x)$	3^{-5}	0	0	2^{-8}
23		2,1	2	2,125	2
24		3,1	3	3,125	3
25		9	9	9	8 (oder 10)
26		18	18	15,875	16 (oder 18)
27		1023	127	15,875	1024
28	Absoluter Rundungsfehler: $\lvert x - \mathrm{rd}_A(x)\rvert$	3^{-5}	3^{-5}	3^{-5}	$2{,}09{\cdot}10^{-4}$
29		2,1	0,1	0,025	0,1
30		3,1	0,1	0,025	0,1
31		9	0	0	1
32		18	0	2,125	2
33		1023	...	...	1
34	Relativer Rundungsfehler: $\left\lvert\frac{x-\mathrm{rd}_A(x)}{x}\right\rvert$	3^{-5}	1	1	0,051
35		$2,1$	0,048	0,012	0,048
36		$3,1$	0,032	0,008	0,032
37		9	0	0	0,111
38		18	0	0,118	0,111
39		1023	...	...	0,001

Beobachtungen

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

- Bei vorzeichenbehafteten Ganzzahlen ist 0 nicht eindeutig definiert.

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

- Ohne bestimmte Rundungsregeln gibt es Zahlen, die man nicht eindeutig runden kann.

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?

- Es kann vorkommen, dass Rundungsfehlern auftreten, wenn man Objekte einfach zählen möchte.

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

- $2^{24} + 1 \approx 17$ Millionen. Vergleichswerte: 1 Gigabyte $= 2^{30}$ Bytes,
  1 Mililiter Wasser enthält $\approx 2^{74}$ Moleküle.

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

- Man weiß nicht, ob 01111111_F genau 15,875 entspricht, oder etwas größerem.

f) Für eine Mantissendarstellung mit 2 Bits ist die Maschinengenauigkeit $\varepsilon_{Ma} = 2^{-3}$. Ver-

gleichen Sie ihre Antworten zu Fragen Nr. 34-39 mit der Maschinengenauigkeit. Ist

$$\left|\frac{x - \mathrm{rd}_G(x)}{x}\right| \leq \varepsilon_{Ma} \tag{1}$$

immer erfüllt?

- Die Rundungsfehler, die durch rd_G entstehen sind immer kleiner oder gleich ε_{Ma} (solange die Zahl im Range ist, natürlich). Man kann Gleichung (1) beweisen (betrachten Sie dazu Ihre Antworten zu Fragen Nr. 12-20.).

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

- Nein. Die kleinste darstellbare positive Zahl ist nur von der Anzahl an Bits in den Exponenten abhängig. Die Maschinengenauigkeit ist nur von der Anzahl an Bits in der Mantisse abhängig.

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

- Fast ein Viertel für beide. Bei G: $\frac{60}{256}$ =0,23.

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

- 2 (Vorzeichen) $\cdot$ 2 (Exponentenkombinationen) $\cdot$ 2^{23} (Mantissenmöglichkeiten) = 2^{25} = 33554432 Zahlendarstellungen.

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?

- Es gibt kaum Unterschiede. Man kann F durch I und einen Skalierungsfaktor implementieren.

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

- Logarithmisch.

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

- Die Anzahl an darstellbare Zahlen im $[2^k, 2^{k+1})$ und die Maschinengenauigkeit ε_{Ma}.

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

- Die Zahlen aus Fragen Nr. 8-10.

2) Gleitkomma-Zahlen im IEEE-Standard

Bit-Verwendungszweck	gesetztes Bit für $-1.0\overline{0011} \cdot 2^0$
Vorzeichen (1 Bit) ('-' wird zu '1')	1
Exponent (8 Bits) (0 + 127 = 127)	0 1 1 1 1 1 1 1
Mantisse (23 Bits)	0 0 0 1 1 0 0 1 1 0 0 1 1 0 0 1 1 0 0 1 1 0 1

Das letzte Bit der Mantisse ergibt sich nach der Rundungsregel.

3) Assoziativgesetz

dezimal	binär
-8	-1000_2
11	1011_2
0.75	0.11_2

Damit gilt:

$$\begin{aligned}(-8+11)+0.75 &= (-1000_2+1011_2)+0.11_2 = (11_2)+0.11_2\\ &= 11.11_2 = \boxed{3.75} \text{ (kein Runden nötig)}\\ -8+(11+0.75) &= -1000_2+(1011_2+0.11_2) = -1000_2+(1011.11_2) \overset{Runden!}{=} -1000_2+(1100_2)\\ &= 100_2 = \boxed{4}\end{aligned}$$

Das Ergebnis ist offensichtlich abhängig von der Reihenfolge der abgearbeiteten Operationen. Diese Rundungsfehler können auch zu sehr großen Fehlern führen. Dabei sei auf die Aufgabe zur Stabiliätsanalyse auf dem nächsten Blatt verwiesen.