Einführung in Computational Engineering
Grundlagen der Modellierung und Simulation

Dr. Arne Nägel

TECHNISCHE UNIVERSITÄT DARMSTADT

Wintersemester 2012/2013
Lösungsvorschlag der 5. Übung

Aufgabe 1 Gleittkommadarstellung (8 Punkte)

a) Es sei $\mathbb{M}$ die Menge aller positiven Maschinenzahlen in der normalisierten Gleitkommadarstellung zur Basis 2 mit Mantissenlänge $t = 4$ und 2-stelligen Exponenten $-1 \leq E \leq 2$. Bestimmen Sie die Zahlen dieser Menge in Dezimaldarstellung. Stellen Sie die Zahlen zudem auf dem Zahlenstrahl im Intervall $[0, 10] \subset \mathbb{R}$ dar.

Betrachten Sie im Folgenden den IEEE 754-Standard:

b) Wie lautet die Dezimaldarstellung der Zahl $0|10000101|10010111000000000000000$? Diese sei mit einfacher Genauigkeit (`bias=127`) im Format $S|E|M$ gespeichert.

c) Wandeln Sie die Dezimalzahl -83.625 in eine Gleittkommazahl mit einfacher Genauigkeit um.

d) Existiert eine Darstellung der Zahl 0 als normalisierte Gleittkommazahl? Geben Sie diese an bzw. begründen Sie, warum eine solche nicht existiert.

Lösungsvorschlag

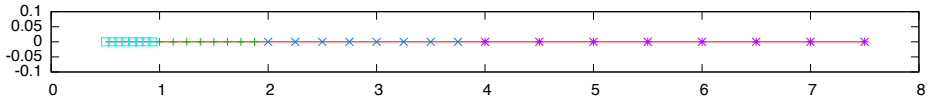
a) (3 Punkte) Verwendet wird die Darstellung $\mathsf{v} = M \cdot 2^E$. Die normalisierte Mantisse besitzt die Binärdarstellungen

$$M \in \{1.000, 1.001, 1.010, 1.011, 1.100, 1.101, 1.110, 1.111\}.$$

Die Exponenten sind $E \in \{-1, 0, 1, 2\}$. In Dezimaldarstellung gilt:

$$M \in \{\frac{8}{8}, \frac{9}{8}, \frac{10}{8}, \frac{11}{8}, \frac{12}{8}, \frac{13}{8}, \frac{14}{8}, \frac{15}{8}\}$$

sowie $2^E \in \{\frac{1}{2}, 1, 2, 4\}$.



b) (2 Punkte) Die gegebene Zahl wird im Format $S|E|M$ gespeichert, wobei die Dezimalzahl sich aus der Formel

$$\mathsf{v} = (-1)^S \cdot (1 + M) \cdot 2^{E-127}$$

ergibt. Man liest ab:

$$\begin{aligned} S &= 0, \\ E &= 10000101_2 = (128 + 4 + 1)_{10} = 133_{10}, \\ M &= 10010111000000000000000 \end{aligned}$$

Einsetzen liefert:

$$v = (1.10010111)_2 \cdot 2^6 = (1100101.11)_2 = (2^6 + 2^5 + 2^2 + 2^0 + 2^{-1} + 2^{-2})_{10} = 101.75_{10}$$

c) (2 Punkte) Gesucht ist eine Darstellung mit $-83.625 = (-1)^S \cdot (1+M) \cdot 2^{E-127}$. Betrachte zunächst die Binärdarstellung:

$$(-83.625)_{10} = (-1010011.101)_2 = (-1.010011101)_2 \cdot 2^6.$$

Nun müssen S, E, und M bestimmt werden. Es ist $S = 1$. Für E gilt $6 = E - 127$, also $E = 133_{10} = 10000101_2$. Gemäß IEEE 754 werden Zahlen mit einfacher Genauigkeit mit 32 bits gespeichert, wobei 1 Bit für S, 8 Bit für E und 23 Bit für M verwendet werden damit:

$$1|10000101|01001110100000000000000$$

d) (1 Punkt) Nein, die Zahl Null hat keine Darstellung, da die Normalisierung erfordert, dass mindestens ein bit gesetzt ist.

Aufgabe 2 Rundungsfehler (4 Punkte)

a) Die Maschinenzahlen $a = 7/4 = (1.11)_2 \cdot 2^0$ und $b = 3/8 = (1.10)_2 \cdot 2^{-2}$ sollen addiert werden. Berechnen Sie das Ergebnis $a +_{\mathbb{M}} b$ und verwenden Sie für die Mantisse zwei Nachkommastellen. Wie groß ist der relative Fehler?

b) Es seien drei Maschinenzahlen $m_1, m_2, m_3 > 0$ in Gleitkommadarstellung gegeben mit $m_1 \gg m_2, m_3$. Welcher der beiden folgenden Algorithmen zur Berechnung von $m_1 + m_2 + m_3$ ist numerisch besser geeignet und warum?

1. $(m_1 +_{\mathbb{M}} m_2) +_{\mathbb{M}} m_3$
2. $m_1 +_{\mathbb{M}} (m_2 +_{\mathbb{M}} m_3)$

Geben sie ein Beispiel an, das Ihre Antwort bestätigt.

Lösungsvorschlag

a) (2 Punkte) Die Rechnung lautet:

$$\begin{aligned}
(1.11)_2 \cdot 2^0 + (1.10)_2 \cdot 2^{-2} &= (111)_2 \cdot 2^{-2} + (1.10)_2 \cdot 2^{-2} \\
&= (1000.10)_2 \cdot 2^{-2} \\
&= (1.00010)_2 \cdot 2^1 \\
&\to (1.00)_2 \cdot 2^1 \\
&= 2
\end{aligned}$$

Das exakte Ergebnis ist $7/4 + 3/8 = 17/8$, der relative Fehler beträgt $\frac{|17/8-2|}{17/8} = 1/17 \approx 5.88\%$.

b) (2 Punkte) Es ist tendenziell besser, zunächst die kleineren Zahlen zu addieren (Variante 2). Das Ergebnis der Addition fällt dann größer aus, so dass tendenziell weniger Stellen bei der Rundung verloren gehen. Ein Beispiel ist $m_1 = 3/2, m_2 = 1/8, m_3 = 1/8$. Addiert man erst m_1 und m_2, so folgt:

$$(1.10)_2 \cdot 2^0 + (1.00)_2 \cdot 2^{-3} = (1100.0)_2 \cdot 2^{-3} + (1.00)_2 \cdot 2^{-3} = (1101.0)_2 \cdot 2^{-3} = (1.10|1)_2 \cdot 2^0 \to (1.10)_2 \cdot 2^0.$$

somit $(m_1 +_{\mathbb{M}} m_2) +_{\mathbb{M}} m_3 = (m_1 +_{\mathbb{M}} m_3) = m_1 \neq 7/4$.

Addiert man jedoch zunächst m_2 und m_3 so gilt:

$$(1.00)_2 \cdot 2^{-3} + (1.00)_2 \cdot 2^{-3} = (10.00)_2 \cdot 2^{-3} = (1.00)_2 \cdot 2^{-2} = 1/4$$

sowie weiter durch Addition von m_1:

$$(1.10)_2 \cdot 2^0 + (1.00)_2 \cdot 2^{-2} = (110.00)_2 \cdot 2^{-2} + (1.00)_2 \cdot 2^{-2} = (111.00)_2 \cdot 2^{-2} = (1.11)_2 \cdot 2^0$$

also das exakte Ergebnis $m_1 +_{\mathbb{M}} (m_2 +_{\mathbb{M}} m_3) = 7/4 = 1\frac{3}{4}$.

Aufgabe 3 Kondition und Stabilität (8 Punkte)

Die Lösungen der quadratischen Gleichung $y^2 - 2py + q = 0$ (mit $p, q \in \mathbb{R}$ und $p^2 > q$) sind durch die pq-Formel gegeben:

$$y_1 = p + \sqrt{p^2 - q} \quad (1)$$
$$y_2 = p - \sqrt{p^2 - q} \quad (2)$$

a) Untersuchen Sie die Kondition des o.g. Problems. Für welche Werte von p, q ist es gut bzw. schlecht konditioniert?

b) Untersuchen Sie den Algorithmus für $p^2 \gg q$ auf Stabilität. Welches Problem tritt dann bei der Berechnung von y_2 auf? Leiten Sie aus der Beziehung $(y - y_1)(y - y_2) = y^2 - 2py + q$ ein numerisch stabiles Berechungsverfahren für y_2 her.

Lösungsvorschlag

a) Definiere $f_1(p,q) := p + \sqrt{p^2 - q}$ und $f_2(p,q) := p - \sqrt{p^2 - q}$. Für die relativen Fehler $\varepsilon_1, \varepsilon_2, \varepsilon_p, \varepsilon_q$ gelten folgende Beziehungen (2 Punkte):

$$\varepsilon_1 \approx \frac{p}{y_1}\frac{\partial f_1}{\partial p}\varepsilon_p + \frac{q}{y_1}\frac{\partial f_1}{\partial q}\varepsilon_q$$
$$\varepsilon_2 \approx \frac{p}{y_2}\frac{\partial f_2}{\partial p}\varepsilon_p + \frac{q}{y_2}\frac{\partial f_2}{\partial q}\varepsilon_q$$

Untersucht man die Konditionszahlen (Verstärkungsfaktoren) folgt (3 Punkte):

$$\begin{aligned}\frac{p}{y_1}\frac{\partial f_1}{\partial p} &= \frac{p}{p + \sqrt{p^2 - q}}\left(1 + \frac{p}{\sqrt{p^2 - q}}\right) \\ &= \frac{p}{p + \sqrt{p^2 - q}}\frac{\sqrt{p^2 - q} + p}{\sqrt{p^2 - q}} \\ &= \frac{p}{\sqrt{p^2 - q}}\end{aligned}$$

und

$$\begin{aligned}\frac{q}{y_1}\frac{\partial f_1}{\partial q} &= \frac{q}{p+\sqrt{p^2-q}}\frac{-1}{2\sqrt{p^2-q}} \\ &= \frac{q}{p+\sqrt{p^2-q}}\frac{p-\sqrt{p^2-q}}{p-\sqrt{p^2-q}}\frac{-1}{2\sqrt{p^2-q}} \\ &= \frac{q(p-\sqrt{p^2-q})}{q}\frac{-1}{2\sqrt{p^2-q}} \\ &= -\frac{p-\sqrt{p^2-q}}{2\sqrt{p^2-q}}\end{aligned}$$

sowie analog

$$\frac{p}{y_2}\frac{\partial f_2}{\partial p} = -\frac{p}{\sqrt{p^2-q}},\ \frac{q}{y_2}\frac{\partial f_2}{\partial q} = \frac{p+\sqrt{p^2-q}}{2\sqrt{p^2-q}}.$$

Das Problem ist somit für $p^2 \approx q$ schlecht konditioniert (1 Punkt). Umgekehrt ist es für $q < 0$ wegen der Abschätzungen

$$\left|\frac{p}{y_2}\frac{\partial f_2}{\partial p}\right| = \left|\frac{p}{\sqrt{p^2-q}}\right| \leq 1, \left|\frac{q}{y_2}\frac{\partial f_2}{\partial q}\right| = \left|\frac{p\pm\sqrt{p^2-q}}{2\sqrt{p^2-q}}\right| \leq 1.$$

gut konditioniert.

b) Für $p^2 \gg q$ ist das Problem zwar gut konditioniert, wegen der potentiellen Auslöschung jedoch numerisch instabil. Betrachtet man den Fall $p > 0$, so ist die Berechnung von y_2 problematisch. Aus $(y-y_1)(y-y_2) = y^2 - 2py + q$ folgt $q = y_1 y_2$ und daher

$$y_2 = \frac{q}{y_1}$$

Bei Verwendung dieser Formel wird Auslöschung vermieden (2 Punkte).

(Umgekehrt ist für $p < 0$ die Berechnung von y_2 numerisch stabil; die Berechnung von y_1 jedoch instabil. In diesem Fall verwende man $y_1 = \frac{q}{y_2}$.)