Einführung in Computational Engineering Grundlagen der Modellierung und Simulation

Prof. Jan Peters, C. Daniel, MSc. und H. van Hoof, MSc.

TECHNISCHE UNIVERSITÄT DARMSTADT

Wintersemester 2012/2013
Lösungsvorschlag der 5. Übung

Aufgabe 1 Arithmetik und Zahldarstellung nach IEEE 754 (10 Punkte)

(a) Welche charakteristische Problematik zeigt sich bei der Umwandlung von $\frac{1}{3}$ in eine Gleitpunktzahl?

(b) Wie groß ist der absolute Fehler bei der Darstellung der Zahl 1 000 000 009 mit einfacher Genauigkeit? Begründen Sie Ihre Antwort.

(c) Welche beiden Bedingungen müssen erfüllt sein, damit eine Zahl ohne Genauigkeitsverluste von double nach single konvertiert werden kann?

(d) Die Entfernung von Darmstadt und New Orleans ist ca. 8 100 km. Wenn man diese Zahl nach IEEE 754 als 32 Bit Zahl im Computer speichern will, tritt ein Rundungsfehler auf. Wie gross kann hier der absolute Rundungsfehler sein? *Hinweis: Betrachten Sie Werte der Größenordnung 8100 und treffen Sie eine allgemeine Aussage. Der Rundungsfehler entsteht durch die beschränkte Mantissenlänge.*

(e) Wie groß kann der maximale absolute Rundungsfehler werden, wenn man statt 8 100 km 8 100 000 m speichern will?

(f) Warum wird zur Darstellung der Zahl 0 als normalisierte, einfach genaue Gleitkommazahl nach IEEE 754 eine Ausnahmeregel benötigt?

Lösungsvorschlag

a)

$$\frac{1}{3} = 0.\overline{3}$$

$$s = 0, \qquad 0.\overline{3}_{10} = 0.\overline{01}_2 \qquad \text{periodisch!}$$

⇒ Runden an den letzten Stellen ist notwendig.

b)

$$1000000009 = 2^{29} + \ldots$$

⇒ Die Binärdarstellung benötigt hier 30 Stellen, abzüglich der 1. Stelle fallen die letzten 6 Stellen weg. Diese lauten:

$$\underbrace{\ldots 000}_{\text{Runden hier ist kein Problem}} \qquad 1001_2 \;\widehat{=}\; 9_{10}$$

⇒ Ungenauigkeit 9.

c) → Bei hoechstens 23 relevanten Mantissenbits,
→ Exponenten zwischen -126 und 127, oder Spezialfälle:
0, NaN, dann passt die Zahl in die Single-Darstellung.

d) Die Mantisse ist 23 Bit lang, weshalb das letzte Bit 2^{-23} ist. Werte, die kleiner als 2^{-23} sind, werden abgeschnitten. Der Rundungsfehler ist maximal 2^{-23}. Nun liegt 8000 zwischen 2^{12} und 2^{13}. Somit ist der Exponent der Binärdarstellung gleich 12. Beziehen diesen nun mit ein, so lautet der absolute Rundungsfehler $2^{12} \cdot 2^{-23} = 2^{-11}$km.

e) 8,000,000 liegt zwischen 2^{22} und 2^{23}. Analog zu a) gilt hier: $2^{22} \cdot 2^{-23} = 2^{-1}$m.

f) Nach $IEEE$ Darstellung

$$\pm \underbrace{(1+M)}_{\neq 0} \cdot \underbrace{2^{E-bias}}_{\neq 0}$$

⇒ Durch die Addition der 1 ist es nicht möglich für die Mantisse 0 zu erhalten.
⇒ Ausnahmeregel notwendig!

Aufgabe 2 Kondition und Stabilität (10 Punkte)

Wenn p die Wahrscheinlichkeit eines Ereignisses ist , dann ist die Wahrscheinlichkeit dass dieses Ereignis in n unabhängigen Versuchen zumindest ein mal auftritt

$$f(p) = 1 - (1-p)^n.$$

a) Berechnen Sie die Konditionszahlen von f. Hinweis: $0 \leq p \leq 1$ (p ist eine Wahrscheinlichkeit).

b) Betrachten Sie den Fall $n = 2$. Ist f gut konditioniert für $p > 0$ (Konditionszahl ≤ 1)?

c) Kann man die Konditionszahl durch die Verwendung eines anderen Algorithmus verbessern?

d) Was könnte ein Problem sein, wenn $p \approx 1$ ist?

Betrachten Sie die auf den reellen Zahlen $\mathbb{R}$ definierte Funktion

$$f(x) := -x^2 + x$$

e) Geben Sie eine Begründung an, warum sich die numerische Stabilität erhöht, wenn man die erste Funktion an der Stelle $x = 1$ mit $f(x) := x(1-x)$ auswertet. Ersetzen Sie dazu jeweils alle Gleitpunktimplementierungen $gl(x \Diamond y)$ durch $(x \Diamond y) \cdot (1 + \varepsilon_\Diamond)$, und betrachten Sie $x \to 1$.

Beispiel: $\sin(x) - 2$ wird nach ersetzen $(\sin(x)(1+\epsilon_1) - 2)(1+\epsilon_2)$.

Lösungsvorschlag

a) Die Konditionszahl (Verstärkungsfaktor)

$$\begin{aligned}
\text{cond}_{f(p)} &= \left|\frac{p}{f(p)}\frac{\partial f(p)}{\partial p}\right| \\
&= \left|\frac{p}{1-(1-p)^n}n(1-p)^{(n-1)}\right| \\
&= \frac{p}{1-(1-p)^n}n(1-p)^{(n-1)} \quad \text{Wenn } 0 \le p \le 1
\end{aligned}$$

(2 Punkte)

b) Die Funktion ist gut konditioniert wenn $\text{cond}_{f(p)} < 1$:

$$\begin{aligned}
\frac{p}{1-(1-p)^2}2(1-p)^1 &\le 1 \\
p &\le \frac{1}{2(1-p)} - \frac{(1-p)}{2} \quad (p \neq 0) \\
2p - 2p^2 &\le 2p - p^2 \\
-p^2 &\le 0 :
\end{aligned}$$

die Funktion ist gut konditioniert für $p \neq 0$. (2 Punkte)

c) Die Konditionszahl ist nur von der Funktion abhängig, und nicht von dem Algorithmus. Deswegen kann die Konditionszahl nicht durch ein anderes Verfahren verbessert werden. (Ein anderer Algorithmus könnte aber die numerische Stabilität erhöhen.) (1 Punkt)

d) Wenn $p \approx 1$ werden in f zwei nahezu gleich große Zahlen von einander abgezogen. Deswegen ist die Berechnung numerisch instabil. (1 Punkt)

e)

$$\begin{aligned}
i) \qquad & x - x^2 \\
\Rightarrow & (x - x^2(1+\epsilon_1))(1+\epsilon_2) \\
= & (x - x^2 - x^2\epsilon_1)(1+\epsilon_2) \\
= x - x^2 - x^2\epsilon_1 + (x-x^2)\epsilon_2 - x^2\epsilon_1\epsilon_2 & \longrightarrow_{(x\to 1)} \quad -\epsilon_1 - \epsilon_1\epsilon_2 \\
ii) \qquad & x(1-x) \\
\Rightarrow & x(1-x)(1+\epsilon_1)(1+\epsilon_2) \\
= & (x - x^2)(1 + \epsilon_1 + \epsilon_2 + \epsilon_1\epsilon_2) \\
= & f(x)(1 + \epsilon_1 + \epsilon_2 + \epsilon_1\epsilon_2) \longrightarrow_{(x\to 1)} 0
\end{aligned}$$

$\rightarrow$ bei $i)$ bleiben die Fehler erhalten, bei $ii)$ nicht. (4 Punkte)