Algorithmische Mathematik I

Wintersemester 2011 / 2012
Prof. Dr. Sven Beuchler
Peter Zaspel

universitätbonn

Übungsblatt 2.

Abgabe am **02.11.2011**.

Aufgabe 1. (Festkommazahlen)

Ein – zugegeben etwas primitiver – Rechner stellt reelle Zahlen im Festkommaformat mit einem Byte dar. Dabei werden ein Vorzeichen–Bit, vier Bits vor dem Komma und drei Bits hinter dem Komma verwendet. Somit haben Zahlen im Rechner die Form

$$z = (-1)^s \sum_{i=1}^{7} d_i \cdot 2^{i-4}$$

a) Welche Darstellung haben die Zahlen 7.25 und -5.625?

b) Wie viele verschiedene Zahlen können in obigem Format dargestellt werden?

c) Geben Sie die maximal und minimal darstellbaren Zahlen $z_{\max}$ und $z_{\min}$ sowie die betragsmäßig kleinste Zahl an.

d) Nicht darstellbare Zahlen x im Bereich $[z_{\min}, z_{\max}]$ werden auf die nächste darstellbare Zahl z gerundet. Dabei tritt ein (absoluter) Rundungsfehler $e_{\text{abs}} := |x-z|$ auf. Relativ zu $|x|$ (d.h. prozentual) ist dieser Fehler definiert als $e_{\text{rel}} := \frac{|x-z|}{|x|}$. Geben Sie den absoluten und den relativen Rundungsfehler bei der Darstellung der Zahl $x = \frac{1}{3}$ an.

e) Bestimmen Sie den maximalen absoluten und relativen Rundungsfehler für reelle Zahlen im Bereich $[z_{\min}, z_{\max}]$.

Beachten Sie, dass eine Rundung von $|x| > 0$ auf $z = 0$ als Underflow gewertet wird und daher nicht in die Rundungsfehleranalyse eingeht.

(5 Punkte)

Aufgabe 2. (Gleitkommazahlen)

a) Wir betrachten nun einen Rechner mit Gleitkommaarithmetik, der einen deutlich geringeren maximalen Rundungsfehler aufweisen soll. Dazu setzen wir für die Basis $b = 2$, verwenden $t = 3$ Ziffern für die Matissenlänge und $p = 1$ für die Exponentenlänge. Damit ist unsere *nomalisierte Gleitkommadarstellung* im Rechner gegeben durch:

$$\pm 1.d_1 d_2 \cdot 2^e \text{ wobei } d_1, d_2 \in \{0, 1\} \text{ und } e \in \{-1, 0, 1\}.$$

Die Null wird gesondert dargestellt mit $d_1 = d_2 = 0$, $e = -1$, also:

$$\pm 0 := \pm 1.00 \cdot 2^{-1}$$

Geben Sie alle darstellbaren, nichtnegativen normalisierten Gleitkommazahlen an (das Vorzeichen vernachlässigen wir) und markieren Sie diese auf einem Zahlenstrahl. Zur Darstellung einer beliebigen reellen Zahl x verwenden wir wieder die nächstgelegene Gleitkommazahl. Geben Sie den absoluten und den relativen Rundungsfehler bei der Darstellung der Zahlen $x = \frac{16}{5}$ und $x = \frac{3}{8}$ an.

b) Zeigen Sie durch Induktion, dass $x = \frac{1}{4}$ zur Basis $b = 5$ die Darstellung

$$x = \sum_{i=1}^{\infty} 1 \cdot 5^{-i}$$

besitzt.

(7 Punkte)

Aufgabe 3. (Rundung)

a) Zeigen Sie, daß die folgenden Ausdrücke mathematisch äquivalent sind:

- $((a+b)(a-b))^2$
- $(a^2+b^2)^2 - 4(ab)^2$
- $(a^2-b^2)^2$

b) Seien nun $a = 10^6 + 1$ und $b = 10^6 - 2$. Multiplizieren Sie damit obige Ausdrücke aus. *Jedes* Zwischenergebnis, das nicht mit 10 Dezimalstellen (also einer Gleitkommazahl mit einer Mantisse der Länge 10) dargestellt werden kann, soll auf 10 Stellen gerundet werden.

c) Berechnen Sie jeweils den relativen Fehler der Resultate (2 gültige Ziffern genügen). Was ist der Grund für dieses Verhalten?

(5 Punkte)

Aufgabe 4. (Kondition und Stabilität)

Die Funktion $s(x) = \dfrac{\sin x}{x}$ soll bei $x \approx 0$, $x \neq 0$ ausgewertet werden.

a) Erklären sie die Begriffe „Kondition“ und „Stabilität“.

b) Geben Sie die absolute und relative Kondition von s an der Stelle x an und ermitteln Sie so ob die Auswertung von s gut konditioniert ist. Verwenden Sie hierbei die Definition der Kondition über die Ableitung (siehe Bemerkung 1 im Abschnitt 1.3.2)

(3 Punkte)

Programmieraufgabe 1. (B-adische Zahlendarstellung)

Implementieren Sie Algorithmus 1.3 zur Berechnung der B-adischen Zahlendarstellung aus der Vorlesung. Schreiben Sie dazu analog zur Vorlesung die Methode `BasisN(int a, int B, int* z, int* L)`. Beachten Sie, dass das zurückgegebene Array `z` von der Methode in der entsprechenden Größe dynamisch angelegt wird und zum schluss auch korrekt wieder frei gegeben wird. Berechnen Sie nun die Darstellung der Dezimalzahl $a = 1234$ für $B = 2, \ldots, 10$.

(5 Punkte)

Programmieraufgabe 2. (Wurzel-Berechnung)

Implementieren Sie Algorithmus 1.6 aus der Vorlesung, mit Hilfe dessen Sie $\sqrt{r}$ annähern können. Implementieren Sie dazu die Funktion `double Wurzel(double r, double epsilon)` und bestimmen Sie die Wurzelnäherungen für $r = 2, 3, 4, 5$ und $\epsilon = 10^{-2}, 10^{-5}, 10^{-8}, 10^{-12}, 10^{-20}$. Geben Sie jeweils auch den absoluten und relativen Fehler gegenüber der durch die Funktion `sqrt()` (aus der `math.h`) berechneten Lösung an.

(5 Punkte)

Die Abgabe der Programmieraufgaben erfolgt in den CIP-Pools in der Woche vom 31.10. bis 04.11.2011. Die Listen für die Anmeldung zu den Abgabe- Terminen hängt in der Woche vom 24.10. bis 28.10.2011 aus. Bitte beachten Sie, dass Sie lauffähige HRZ-Login-Accounts benötigen.

Präsenzaufgaben

Aufgabe 5. (Gleitkommazahlen)

Unser Ein-Byte-Rechner soll nun mit Gleitkomma-Arithmetik ausgestattet werden. Bei der (normalisierten) Zahldarstellung werden ein Bit für das Vorzeichen s, vier Bits für die Mantisse m ($1 \leq m < 2$) und drei Bits für den Exponenten e bei einem Bias von 4 verwendet. Die führende Eins in der Mantisse wird nicht abgespeichert. Somit haben Zahlen im Rechner die Form

$$z = (-1)^s \cdot m \cdot 2^{e-4} \text{ mit } m = 1 + \sum_{i=1}^{4} m_i \cdot 2^{-i} \text{ und } e = \sum_{j=1}^{3} e_j \cdot 2^{j-1}$$

a) Welche Darstellung haben die Zahlen 6.5 und -0.375?

b) Wie viele verschiedene Zahlen können in diesem Gleitkomma-Format dargestellt werden?

c) Geben Sie die maximal und minimal darstellbaren Zahlen $z_{\max}$ und $z_{\min}$ sowie die betragsmässig kleinste darstellbare Zahl $z_{|\min|}$ an.

d) Skizzieren (bzw. plotten) Sie alle darstellbaren Zahlen auf einer Zahlengeraden.

e) Auch hier werden nicht darstellbare Zahlen auf die nächste darstellbare Zahl gerundet. Bestimmen Sie den maximalen absoluten und relativen Rundungsfehler für reelle Zahlen im Bereich $[z_{|\min|}, z_{\max}]$.

Aufgabe 6. (Fehlerfortpflanzung)

Wir wollen nun die Fortpflanzung von Fehlern bei der Durchführung der vier arithmetischen Grundoperationen $(+, -, \cdot, /)$ betrachten. Die Zahlen x und y seien mit Fehlern Δx und Δy behaftet, wobei $|\frac{\Delta x}{x}|$ und $|\frac{\Delta y}{y}|$ weit kleiner als 1 seien.[1] Zeigen Sie, daß selbst bei exakter Rechnung (also ohne weitere Rundungsfehler) für die relativen Fehler der Ergebnisse die folgenden Aussagen gelten:

a) $$\frac{(x+\Delta x)+(y+\Delta y)-(x+y)}{x+y} = \frac{x}{x+y}\cdot\frac{\Delta x}{x} + \frac{y}{x+y}\cdot\frac{\Delta y}{y}$$

b) $$\frac{(x+\Delta x)-(y+\Delta y)-(x-y)}{x-y} = \frac{x}{x-y}\cdot\frac{\Delta x}{x} - \frac{y}{x-y}\cdot\frac{\Delta y}{y}$$

c) $$\frac{(x+\Delta x)\cdot(y+\Delta y)-(x\cdot y)}{x\cdot y} \approx \frac{\Delta x}{x} + \frac{\Delta y}{y}$$

d) $$\frac{(x+\Delta x)/(y+\Delta y)-(x/y)}{x/y} \approx \frac{\Delta x}{x} - \frac{\Delta y}{y}$$

e) Falls der relative Fehler des Ergebnisses sehr viel größer ist, als der relative Fehler der Eingabedaten, so spricht man von Auslöschung. In welchen der obigen Fälle kann Auslöschung auftreten?

Hinweis: Das „$\approx$“ heißt hier, dass die sehr kleinen Produkte $\Delta x \cdot \Delta y$ weggelassen werden können.

Aufgabe 7. (Zahldarstellung)

Ein 16-Bit Rechner stellt reelle Zahlen im Festkommaformat mit einem Vorzeichen-Bit, 8 Bits vor dem Komma und 7 Bits hinter dem Komma dar.

[1] Man schreibt dies üblicherweise als $|\frac{\Delta x}{x}| \ll 1$

a) Geben Sie die Formeldarstellung einer solchen Festkommazahl sowie die grösste und die kleinste positive im Rechner darstellbare Festkommazahl an.

b) Definieren Sie den Begriff „Rundungsfehler“ und bestimmen Sie den maximalen absoluten und relativen Rundungsfehler für alle reellen Zahlen im Darstellungsbereich aus Aufgabe (a).