Mathematisches Institut
Prof. Dr. Christiane Helzel
Yanick Kiechle
Dr. Christoph Matern
Erik Chudzik
11. April 2024

Heinrich Heine Universität Düsseldorf

Numerik I – 1. Übungsblatt

Aufgabe 1: (4 Punkte)

Geben Sie für die folgenden Dezimalzahlen die binäre Gleitkommadarstellung an. Die Mantisse $m = 1 + f$ und der Exponent $c - B$ sollen mit je 4 Bits gespeichert werden. Als Bias soll $B = 7$ verwendet werden. Geben Sie den relativen Fehler an, falls gerundet werden muss.

(a) 44

(b) $0,3$

(c) $-0,8$

Aufgabe 2: (8 Punkte)

Ähnlich wie das in der Vorlesung eingeführte 64 Bit Format existiert auch ein 32 Bit Format. Hier werden üblicherweise 1 Bit für das Vorzeichen, 8 Bit für den Exponenten sowie 23 Bit für die Mantisse gespeichert. Die Darstellung der Zahlen sieht dann wie folgt aus:

$$x = \pm(1 + f)2^{c-127}$$

(a) Bestimmen Sie alle möglichen Zahlen c und somit alle möglichen Exponenten. Der höchste und niedrigste Exponent sind wieder reserviert für subnormale Zahlen bzw. `inf` und `nan`. Bestimmen Sie dann die größte und kleinste Gleitkommazahl $x_{\max}$, $x_{\min}$. Bestimmen Sie außerdem die kleinste positive sowie die größte negative Zahl x_{posmin}, x_{negmax}, die unter Vernachlässigung der subnormalen Zahlen dargestellt werden können.

(b) Bestimmen Sie die Maschinengenauigkeit `eps`.

(c) Bestimmen Sie aus den positiven, subnormalen Zahlen die kleinste sowie die größte Zahl x_{submin}, x_{submax}.

(d) Wie vielen Zahlen wird der Wert `nan` zugeordnet?

Aufgabe 3: (6 Punkte)

Für $p, q \in \mathbb{R}$ sei

$$f(x) = x^2 + 2px + q.$$

(a) Berechnen Sie für $p = q = 10^3$ die Nullstellen von f nach der bekannten Lösungsformel

$$x_{1,2} = -p \pm \sqrt{p^2 - q}$$

in der dezimalen Gleitkommaarithmetik mit Mantissenlänge 2.

Welcher der beiden Näherungen würden Sie vertrauen?

Hinweis: Denken Sie daran, nach jeder Rechenoperation mittels der bekannten Rundungsoperation

$$\mathrm{rd}(x) = \mathrm{sign}(x) \cdot \begin{cases} (f_1 \cdot 10^{-1} + f_2 \cdot 10^{-2}) \cdot 10^e, & f_3 \in \{0, \ldots, 4\} \\ (f_1 \cdot 10^{-1} + f_2 \cdot 10^{-2} + 10^{-2}) \cdot 10^e, & f_3 \in \{5, \ldots, 9\} \end{cases}$$

zu runden.

(b) Seien x_1, x_2 wie im Teil (a) gegeben. Zeigen Sie, dass gilt

$$x_1 x_2 = q.$$

Leiten Sie damit eine bessere Näherung für Teil (a) her.

(c) Skizzieren Sie einen Pseudo-Code für eine Funktion `pqsolve(p,q)`, welche die beiden Nullstellen von f berechnet und dabei die bessere Näherung aus Teil (b) verwendet.

Aufgabe 4: **Programmieraufgabe**

(a) Seien A das numerische Gleitkommagitter und $x, y, z \in A$. Verifizieren Sie anhand geeigneter Beispiele

(i) $(x \oplus y) \oplus z \neq x \oplus (y \oplus z)$

(ii) $(x \oplus y) \odot z \neq (x \odot z) \oplus (y \odot z)$

(b) Seien $m_i = 1 + f_1 2^{-1} + f_2 2^{-2} + \ldots + f_i 2^{-i}$, $i \in \mathbb{N}_0$ mit $f_k \in \{0, 1\}$, $k = 1, \ldots, i$ und $e \in \{0, 1, 2\}$. Sei ferner A_i die Menge aller in der Form $x = \pm m_i 2^{\pm e}$ darstellbaren reellen Zahlen.
Stellen Sie $A_i \cup \{0\}$, $i = 1, \ldots, 3$ sowie die zugehörigen Maschinengenauigkeiten in einem Plot grafisch dar.

Abgabe am 18. April 2024 bis 10:30 Uhr im Ilias.