Technische Universität Chemnitz
Fakultät für Mathematik
Dr. Ralf Hielscher
Dr. Roman Unger

SoSe 2018

Numerische Methoden für Ingenieure

http://www.tu-chemnitz.de/∼ rahi

Übungsblatt 2 - Maschinenzahlen

Aufgabe 1: Machen Sie sich mit folgenden Zahldarstellungen vertraut. Beispielhaft verwenden wir die Basis $b = 2$ und die Bitlänge $N = 8$.
nicht negative ganze Zahlen: $x = [x_7x_6x_5x_4x_3x_2x_1x_0]_2$,

$$x = x_7 2^7 + x_6 2^6 + x_5 2^5 + x_4 2^4 + x_3 2^3 + x_2 2^2 + x_1 2^1 + x_0 2^0.$$

ganze Zahlen: $x = [x_7x_6x_5x_4x_3x_2x_1x_0]_{\overline{2}}$,

$$x = -x_7 2^7 + x_6 2^6 + x_5 2^5 + x_4 2^4 + x_3 2^3 + x_2 2^2 + x_1 2^1 + x_0 2^0.$$

Fixpunktzahlen: $x = [x_7.x_6x_5x_4x_3x_2x_1x_0]_{\overline{2}}$,

$$x = -x_7 2^0 + x_6 2^{-1} + x_5 2^{-2} + x_4 2^{-3} + x_3 2^{-4} + x_2 2^{-5} + x_1 2^{-6} + x_0 2^{-7}.$$

Gleitpunktzahlen: $x = [\underbrace{\pm x_3.x_2x_1x_0}_{\text{Mantisse}} \mid \underbrace{e_2e_1e_0}_{\text{Exponent}}]_{\overline{2}}$, (nicht normalisiert)

$$x = \pm[x_3.x_2x_1x_0]_2 \cdot 2^{[e_2e_1e_0]_{\overline{2}}}.$$

Aufgabe 2: Finden Sie für die folgenden Zahlen alle möglichen Darstellungen.

a) 1 b) -1 c) 99 d) -35 e) $\frac{15}{8}$ f) $\frac{3}{32}$

Aufgabe 3: Überzeugen Sie sich, dass man in der Ganzzahlendarstellung mit gewöhlicher schriftlicher Addition addieren kann. Überprüfen Sie dafür

a) 1 + (-1) = 0, b) 13 + (-5) = 8, c) 5 + (-13) = -8.

Aufgabe 4: Bestimmen Sie die doppelte Maschinengenauigkeit indem Sie ε in einer Schleife so lange verringern bis $1 \oplus \varepsilon = 1$.

Aufgabe 5: Bestimmen Sie auf ähnliche Art und Weise die betragsgrößte darstellbare Zahl.

Aufgabe 6: Berechnen Sie in einfacher Genauigkeit die Summen

$$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \dots$$

und

$$1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \dots$$

Wie kann man die Summen genauer berechnen? Welcher Wert ergibt sich tatsächlich bei diesen unendlichen Summen?

Aufgabe 7: Bestimmen Sie die Nullstellen x_1, x_2 des Polynoms $P(x) = (x+1/N)(x+N)$,

a) direkt, ohne Nutzung der Lösungsformel,

b) mit Hilfe der Lösungsformel,

c) über den Satz von Vieta, d.h. mittels $x_1 = 1/x_2$.

Veranschaulichen Sie sich die relativen Fehler für die betragskleinere Nullstelle für große N und interpretieren Sie diese.