Aufgabe 1

(a) Es ist $(0,5731 \times 10^5)_8 = 5 \cdot 8^4 + 7 \cdot 8^3 + 3 \cdot 8^2 + 1 \cdot 8 = (0,24264 \times 10^5)_{10}$

(b) Es ist $0.3 = 1228.8 \cdot 2^{-12} \approx \left(0.10011001101 \cdot 2^{-1}\right)_2 = 1229 \cdot 2^{-12} = 0.300048828 \cdots \cdot 10^0$. Dieses Ergebnis ist genau dann gleich 0.3, wenn $r \leq 4$ ist.

(c) Sei $x_2 \in \mathbb{F}(4,6,2)$ und $x_3 \in \mathbb{F}(3,7,1)$. Es ist $e_{\max}(x_2) = \sum_{j=0}^{1} 3 \cdot 4^j = 15$. Die größte Zahl in 4er System hat überall die Zahl 3 stehen:

$$\max |x_2| = (0,333333 \times 10^{15})_4$$

Es ist $e_{\max}(x_3) = \sum_{j=0}^{0} 2 \cdot 3^j = 2$. Die größte Zahl im 3er System hat an jeder Mantissestelle ein 2 stehen:

$$\max |x_3| = (0,2222222 \times 10^2)_3$$

Umgerechnet zur Basis 10 ist

$$\max |x_2| = 3 \cdot 4^{14} + 3 \cdot 4^{13} + 3 \cdot 4^{12} + 3 \cdot 4^{11} + 3 \cdot 4^{10} + 3 \cdot 4^9 = 1.073.479.680$$

und

$$\max |x_3| = 2 \cdot 3 + 2 \cdot 3^{-1} + 2 \cdot 3^{-2} + 2 \cdot 3^{-3} + 2 \cdot 3^{-4} + 2 \cdot 3^{-5} + 2 \cdot 3^{-6} = \frac{2186}{243} \approx 8,995884774.$$

Da wir eine der beiden Zahlen mit -1 multiplizieren können, um den größten Abstand zwischen beiden Zahlen zu erhalten, gilt:

$$\max_{x_2,x_3} |x_2 - x_3| = |x_2| + |x_3| \approx 1.073.479.680 + 8,995884774 = 1.073.479.688,995884774$$

(d) Ein solches Gegenbeispiel wurde bereits in der Vorlesung gegeben: In $\mathbb{F}(2,2,1)$ ist für $x_4 = \frac{1}{4}$: $x_5 = 0, x_6 = \frac{3}{8}$ und daher $|x_4 - x_5| = \frac{1}{4} \neq \frac{1}{8} = |x_6 - x_4|$

Aufgabe 2

Seien $x, y \in \mathbb{F}(10,3,1)$ mit $x = 2,46$ und $y = -0,755$.

- (natürliches Runden): Es ist $x_0 = 0,246 \times 10^1$.

$$\begin{aligned} x_1 &= (0,246 \times 10^1 \oplus 0,755 \times 10^0) \ominus 0,755 \times 10^0 \\ &= \mathrm{rd}(0,246 \times 10^1 + 0,0755 \times 10^1) \ominus 0,755 \times 10^0 \\ &= \mathrm{rd}(0,3215 \times 10^1) \ominus 0,755 \times 10^0 \end{aligned}$$

In diesem Schritt wird um 0,005 aufgerundet

$$\begin{aligned} &= \mathrm{rd}(0,322 \times 10^1 - 0,0755 \times 10^1) \\ &= \mathrm{rd}(0,2465 \times 10^1) \end{aligned}$$

In diesem Schritt wird noch einmal um 0,005 aufgerundet

$$= 0,247 \times 10^1.$$

Es wird in jedem Iterationsschritt um $0,01$ erhöht. Somit ist $x_{10} = 2,56$.

- (gerades Runden): Es gilt

$$\begin{aligned} x_1 &= rd(rd(0,246 \times 10^1 + 0,755 \times 10^1) - 0,755 \times 10^1) \\ &= rd(rd(0,3215 \times 10^1) - 0,755 \times 10^1) \\ &= rd(0,322 \times 10^1 - 0,755 \times 10^1) \\ &= 0,246 \end{aligned}$$

  Somit gilt offensichtlich $x_{10} = 2,46$.

Aufgabe 3

(a) Listing 1: Programm zum Testen der Präzision von double bzw. float

```
#include <iomanip>
#include <iostream>

using namespace std;

int main()
{
    //float x;
    double x;
    cin >> x;
    x = x + 1;
    cout << setprecision(50) << x;
}
```

Dieses Programm liefert für den Datentyp float bei Eingabe 0.00000006 $= 6 \cdot 10^8$ bzw. 0.00000005 $= 5 \cdot 10^8$ die Ausgabe 1.00000011920928955078125 bzw. 1.

Für den Datentyp double erhält man bei Eingabe 0.0000000000000002 $= 2 \cdot 10^{16}$ die Ausgabe 1.0000000000000002220446049250313080847263336181641, während 0.0000000000000001 $= 10^{16}$ einfach die Ausgabe 1 liefert.

(b) Lässt man im Programm das Addieren der 1 weg, so kann man wesentlich kleinere Zahlen eingeben, bevor einfach nur 0 zurückgegeben wird. Folglich müssen double und float noch kleinere Zahlen darstellen können.