Prof. Dr. Stefan Funken
M.Sc. Andreas Bantle
Dipl.-Math. oec. Klaus Stolle

Universität Ulm
Institut für Numerische Mathematik
Wintersemester 2014/2015

Numerische Lineare Algebra - Theorie-Blatt 4
Lösung

(Abgabe am 10.12.2014 **vor** der Übung!)

Hinweise

Die Hinweise zur Abgabe der Übungsblätter finden Sie auf dem ersten Übungsblatt!

Aufgabe 9 *(Number representation, LaTeX)* (8+8+6 Punkte)

(i) Let $n \in \mathbb{N}$ and $b \in \mathbb{N}_{\geq 2}$. Proof that the mapping

$$\varphi : \{0, 1, \ldots, b-1\}^n \to \{0, 1, \ldots, b^n - 1\} \quad \text{mit} \quad (a_0, \ldots, a_{n-1}) \mapsto \sum_{k=0}^{n-1} a_k b^k$$

is bijective.

(ii) Show that every real number has a unique b-adic representation if "$a_k < b-1$ for infinitely many $k \leq n$" (see Remark 2.1.5).

(iii) In Praxis, the number representation with respect to the bases $b = 10$ (decimal numbers), $b = 2$ (binary numbers) and $b = 16$ (hexadecimal numbers) are of main importance.
In the representation of hexadecimal numbers we have digits in the range of 0 to 15. In order to obtain a monadic notation we use for the values 10-15 the characters A-F.
Fill out the following table (indicate all significant calculations):

decimal	binary	hexadecimal
30.125		
	110001.0101	
		B9.9

Lösung:

(a) Um die Bijektivität zu zeigen, müssen die Injektivität und die Surjektivität gezeigt werden.

- Injektivität:

$$\varphi(l) = \varphi(m) \Rightarrow \sum_{k=0}^{n-1} (l_k - m_k) b^k = 0$$

$$\Rightarrow |l_{n-1} - m_{n-1}| b^{n-1} \leq \sum_{k=0}^{n-2} |l_k - m_k| b^k \leq \sum_{k=0}^{n-2} (b-1) b^k = b^{n-1} - 1$$

$$\Rightarrow |l_{n-1} - m_{n-1}| < 1 \Rightarrow l_{n-1} = m_{n-1} \Rightarrow \sum_{k=0}^{n-2} (l_k - m_k) b^k = 0$$

analog $l_{n-2} = m_{n-2}$, usw.
Hiermit ist die Injektivität gezeigt.

- Surjektivität:
  $\{0,\cdots,b-1\}^n$ und $\{0,\cdots,b^n-1\}$haben die Mächtigkeit b^n. Surjektivität folgt demnach mit Schubfachprinzip aus Injektivität.

(b) Die Existenz einer b-adischen Entwicklung wurde in Satz 2.1.4 gezeigt. Es bleibt noch die Eindeutigkeit zu zeigen. Sei $x \in \mathbb{R}$, o.B.d.A. $0 \leq x < 1$, $x = \sum_1^\infty a_k b^{-k} = \sum_1^\infty \tilde{a}_k b^{-k}$.
Annahme: Es gibt ein $p \in \mathbb{N}_0$ mit $a_k = \tilde{a}_k$, $k = 1, ..p-1$, $a_p \neq \tilde{a}_p$. Sei o.B.d.A. $a_p < \tilde{a}_p$. Dann gilt

$$0 = \sum_{k=1}^{\infty} a_k b^{-k} - \sum_{k=1}^{\infty} \tilde{a}_k b^{-k} = \sum_{k=1}^{\infty} (a_k - \tilde{a}_k) b^{-k} = (a_p - \tilde{a}_p) b^{-p} + \sum_{k=p+1}^{\infty} (a_k - \tilde{a}_k) b^{-k}$$

Wegen $(a_p - \tilde{a}_p) \leq -1$ und $(a_k - \tilde{a}_k) \leq b-1$ für $k > p$ gilt weiter

$$0 = (a_p - \tilde{a}_p) b^{-p} + \sum_{k=p+1}^{\infty} (a_k - \tilde{a}_k) b^{-k} \overset{*}{\leq} -b^{-p} + \sum_{k=p+1}^{\infty} (b-1) b^{-k} = -b^{-p} + b^{-p-1} \cdot b = 0$$

Bei $*$ gilt nur "=", falls, $a_k - \tilde{a}_k = b-1$ für $k \geq p+1$, was äquivalent ist zu $a_k = b-1$, $\tilde{a}_k = 0$ für $k \geq p+1$, was bedeutet, dass unendlich viele a_k gleich $b-1$ sind. Ein Widerspruch.

(c) Folgende Tabelle zeigt die Ergebnisse der Umwandlungen:

Dezimal	Dual	Hexadezimal
30.125	11110.001	1E.2
49.3125	110001.0101	31.5
185.5625	10111001.1001	B9.9

Aufgabe 10 *(Umwandlung in und Operationen auf Gleitpunkt-Darstellungen)* (6+6+6 Punkte)
Gegeben seien $a = \frac{7}{8}, b = -\frac{6}{8}, c = \frac{3}{16} \in \mathbb{M}(2,3,2)$. Es gilt:

$$a = (0.111)_2 \cdot 2^{(00)_2} = (0.111)_2 \cdot 2^0 \qquad b = -(0.110)_2 \cdot 2^{(00)_2} = -(0.110)_2 \cdot 2^0$$
$$c = (0.110)_2 \cdot 2^{-(10)_2} = (0.110)_2 \cdot 2^{-2}.$$

(i) Zeigen Sie für die Operationen $\oplus$, $\ominus$ auf $\mathbb{M}(2,3,2)$, dass $(a \oplus b) \oplus c = (0.101)_2 \cdot 2^{-1} = \frac{5}{16}$ und dass $a \oplus (b \oplus c) = (0.110)_2 \cdot 2^{-1} = \frac{3}{8}$. Verwenden Sie dazu die Standardrundung. Bestimmen Sie den relativen Fehler beider Ergebnisse.

(ii) Bestimmen Sie für $a = \frac{3}{5}$ und $b = \frac{4}{7}$ die Darstellungen von a und b in $\mathbb{M}(2,5,3)$ und $\mathbb{M}(2,3,3)$ (da die Zahlen nicht exakt darstellbar sind, muss hier mit Standardrundung gerundet werden).

(iii) Berechnen Sie auf beiden Gleitpunkt-Darstellungen aus Teil (iii) $a \ominus b$. Wie groß ist der jeweilige Fehler und wie heißt das beobachtete Phänomen?

Lösung:

(i) Wir halten zunächst fest, dass wenn wir exakt rechnen, $a + b + c = \frac{5}{16}$. Nun prüfen wir, was bei der Rechnung auf Gleitkomma-Darstellungen rauskommt.

$$\begin{aligned}(a \oplus b) \oplus c &= \big((0.111) \cdot 2^0 \ominus (0.110) \cdot 2^0\big) \oplus (0.110)_2 \cdot 2^{-2} = (0.001)_2 \cdot 2^0 \oplus (0.110)_2 \cdot 2^{-2} \\ &= (0.100)_2 \cdot 2^{-2} \oplus (0.110)_2 \cdot 2^{-2} = (1.010)_2 \cdot 2^{-2} = (0.101)_2 \cdot 2^{-1} = \frac{5}{16}.\end{aligned}$$

$$\begin{aligned}a \oplus (b \oplus c) &= (0.111) \cdot 2^0 \oplus \big(-(0.110) \cdot 2^0 \oplus (0.110)_2 \cdot 2^{-2}\big) \\ &= (0.111) \cdot 2^0 \oplus \big(-(0.110)_2 \cdot 2^0 + (0.010)_2 \cdot 2^0\big) = (0.111)_2 \cdot 2^0 \ominus (0.100)_2 \cdot 2^0 \\ &= (1.010)_2 \cdot 2^{-2} = (0.011)_2 \cdot 2^0 = (0.110)_2 \cdot 2^{-1} = \frac{3}{8}.\end{aligned}$$

Das erste Ergebnis ist exakt. Für das zweite Ergebnis ergibt sich ein relativer Fehler von $|\frac{3}{8} - \frac{5}{16}| / \frac{5}{16} = \frac{1}{5} = 20\%$.

(ii) Es gilt $a = (0.1\overline{0011})_2 \cdot 2^0$ und $b = (0.\overline{100})_2 \cdot 2^0$. Also erhalten wir in $\mathbb{M}(2,5,3)$ $a = (0.10011)_2 \cdot 2^0$ und $b = (0.10010)_2 \cdot 2^0$ und in $\mathbb{M}(2,3,3)$ $a = (0.101)_2 \cdot 2^0$ und $b = (0.101)_2 \cdot 2^0$.

(iii) In $\mathbb{M}(2,5,3)$ erhalten wir

$$a \ominus b = (0.10011)_2 \cdot 2^0 \ominus (0.10010)_2 \cdot 2^0 = (0.00001)_2 \cdot 2^0 = (0.10000) \cdot 2^{-4} = \frac{1}{32},$$

und einen relativen Fehler von $|\frac{1}{32} - \frac{1}{35}|/|\frac{1}{35}| \approx 8.6\%$. Für $\mathbb{M}(2,3,3)$ hingegen gilt:

$$a \ominus b = (0.101)_2 \cdot 2^0 \ominus (0.101)_2 \cdot 2^0 = (0.000)_2 \cdot 2^0 = 0,$$

und somit ein relativer Fehler von 100% (Auslöschung).