§5 Zahlendarstellung auf Digitalrechnern und Fehleranalyse

Bemerkung 5.1, Zahlendarstellung:
Bei den Digitalrechnern werden reelle Zahlen durch endlich viele physikalische Zustände dargestellt. Weil nur endlich viele reelle Zahlen dargestellt werden können, müssen die anderen approximiert werden.
Sei $g \in \mathbb{N}$ fest mit $g \geq 2$. Zu $r \in \mathbb{R}$ gibt es ein eindeutig bestimmtes $j \in \{0, 1\}$, ein eindeutig festgelegtes $l \in \mathbb{Z}$ und zu $k \in \mathbb{Z}$ mit $k \leq l$ eindeutig bestimmte $a_k \in \mathbb{Z}$ mit:

$$r = (-1)^j \cdot \sum_{k=-\infty}^{l} a_k \cdot g^k \,, \tag{5.1}$$

$$a_l \neq 0 \quad \text{für } r \neq 0, \quad j = l = 0 \quad \text{für } r = 0 \,, \tag{5.1a}$$

$$a_k \in S_g := \{0, 1, \ldots, g-1\} \text{ für alle } k \leq l, \quad a_k < g-1 \text{ für unendlich viele } k < l. \tag{5.1b}$$

g hängt dabei von der Maschine ab;
in der Regel ist $g = 10$ (Dezimaldarstellung) oder $g = 2$ (Dualdarstellung)

Bemerkung 5.2, Die Festpunktdarstellung:
Bei der Festpunktdarstellung sind die Zahlen s und t der Stellen vor bzw. nach dem Komma durch die Maschine fixiert. Die darstellbaren

Zahlen r haben die Gestalt

(5.2) $$r = \pm \sum_{k=-t}^{s-1} b_k \cdot g^k \quad \text{mit } b_k \in S_g \text{ für } -t \le k < s.$$

Darstellung auf der Maschine

$\pm\ b_{s-1} \dots b_0$	$b_{-1} \dots b_{-t}$

Die darstellbaren Zahlen sind in dem offenen Intervall $(-g^s, g^s)$ äquidistant verteilt mit dem Abstand g^{-t}.

Die Festpunktdarstellung ist beispielsweise angemessen für kaufmännische Rechnungen.

Bemerkung 5.3, Die Gleitpunktdarstellung:
Zu $r \in \mathbb{R} \setminus \{0\}$ gibt es eindeutig bestimmte $a \in \mathbb{R}$ mit $\frac{1}{g} \le |a| < 1$ und $b \in \mathbb{Z}$ mit

(5.3) $$r = a \cdot g^b .$$

a heißt die Mantisse und b der Exponent von r.

Für gewisse, durch die Maschine fixierte, $s, t \in \mathbb{N}$ wird a bzw. b durch t bzw. s Stellen und das Vorzeichen dargestellt.
Ist r exakt darstellbar, so ist

$$-(g^s - 1) \le b \le g^s - 1$$

und folglich

$$-g^{(g^s-1)} < r < g^{(g^s-1)} .$$

Beispiele:

Für $g = 10, s = 2$ und $t = 8$ erhält man folgende Darstellungen:

5138	$\mapsto$	0.51380000	04
$-18,23197$	$\mapsto$	-0.18231970	02
$0,03164$	$\mapsto$	0.31640000	-01
$-0,03164$	$\mapsto$	-0.31640000	-01

Bemerkung 5.4, Rundung:

Es sei M die - endliche - Menge der in der Maschine darstellbaren Zahlen - auch Maschinenzahlen - genannt.

Eine Zahl $r \in \mathbb{R}$ zu runden heißt:

Suche ein $\tilde{r} \in M$ mit:

$$|r - \tilde{r}| \leq |r - g| \quad \text{für alle } g \in M.$$

Im Falle der Festpunktdarstellung setzen wir für r wie in (5.1) und $l \leq s - 1$:

$$(5.4) \qquad rd(r) := \begin{cases} (-1)^j \cdot \sum_{k=-t}^{l} a_k \cdot g^k & \text{falls } a_{-t-1} < \frac{1}{2} \cdot g \\ (-1)^j \cdot \left(\sum_{k=-t}^{l} a_k \cdot g^k + g^{-t} \right) & \text{falls } a_{-t-1} \geq \frac{1}{2} \cdot g. \end{cases}$$

Wir erhalten dann die - von r unabhängige - Abschätzung

$$(5.4a) \qquad |r - rd(r)| \leq \frac{1}{2} \cdot g^{-t}.$$

Im Falle der Gleitpunktdarstellung setzen wir

$$gl(0) := 0,$$

(5.4) $$gl(a \cdot g^b) := rd(a) \cdot g^b \quad \text{für } \frac{1}{g} \leq |a| < 1,\ b \in \mathbb{Z}.$$

Für den <u>absoluten Fehler</u> $|r - gl(r)|$ erhalten wir die - von $r = a \cdot g^b$ abhängige - Abschätzung

(5.4a) $$|r - gl(r)| = |a - rd(a)| \cdot g^b \leq \frac{1}{2} \cdot g^{-t} \cdot g^b.$$

Für den <u>relativen Fehler</u> $\frac{|r - gl(r)|}{|r|}$, der den absoluten Fehler in Relation zur betrachteten Zahl $r \neq 0$ setzt, erhalten wir die - von $r = a \cdot g^b$ unabhängige - Abschätzung

(5.4b) $$\frac{|r - gl(r)|}{|r|} \leq \frac{1}{2} \cdot g^{-t+b} \cdot \frac{g}{g^b} = \frac{1}{2} \cdot g^{-t+1} =: eps.$$

eps heißt die <u>Maschinengenauigkeit</u>.
Zu jedem $r \in \mathbb{R} \setminus \{0\}$ gibt es also ein $\varepsilon \in [-eps, eps]$ mit:

(5.4c) $$gl(r) = r \cdot (1 + \varepsilon).$$

<u>Beispiele für $g = 10$, $t = 8$, $s = 2$</u>

$$gl(246813795) = 24681380 = 0{,}2468138 \cdot 10^9,$$

$$gl\left(-\frac{1}{15}\right) = gl(-0{,}0\overline{6}) = -0{,}66666667 \cdot 10^{-1},$$

$$gl(e) = gl(2{,}71828183) = 0{,}27182818 \cdot 10^1.$$

Bemerkung 5.5. Rechenoperationen in der Gleitpunktdarstellung

Für $r_1, r_2 \in M$ setzen wir

(5.5 a) $\quad r_1 \oplus r_2 := gl(r_1 + r_2)$,

(5.5 b) $\quad r_1 \ominus r_2 := gl(r_1 - r_2)$,

(5.5 c) $\quad r_1 \odot r_2 := gl(r_1 \cdot r_2)$,

(5.5 d) $\quad r_1 \oslash r_2 := gl(r_1 : r_2)$ für $r_2 \neq 0$.

Das seien jeweils die Ergebnisse, die von der Maschine berechnet werden.

Bemerkung 5.6:

Die Gleitpunktoperationen genügen nicht den üblichen Gesetzen der arithmetischen Operationen.

Zum Beispiel gilt:

(5.6) $\quad eps = \frac{1}{2} \cdot g^{-t+1} = \min\{x \in M \mid 1 \oplus x > 1\}$.

Für $0 \leq x < eps$, $x \in M$, ist also $1 \oplus x = 1$.

Gegenbeispiel zum Assoziativgesetz:

Sei $g = 10$, $t = 4$, $s = 2$ sowie

$a = 0{,}1349 \cdot 10^{-1} = 0{,}01349$;

$b = 0{,}3368 \cdot 10^{2} = 33{,}68$;

$c = -0{,}3367 \cdot 10^{2} = -33{,}67$.

Dann folgt:

$a + b + c = 0{,}02349$;

$(a \oplus b) \oplus c = 33{,}69 \oplus (-33{,}67) = 0{,}02000$;

$a \oplus (b \oplus c) = 0{,}01349 \oplus 0{,}01 = 0{,}02349$.

Konvention 5.6:

Für $r \in \mathbb{R} \setminus \{0\}$ mit gerundetem $r' \in M$ bezeichne

(5.6) $$\varepsilon_r := \frac{|r - r'|}{|r|}$$

den relativen Fehler.

Satz 5.7, Fehlerfortpflanzung

- exakte Rechnung mit falschen Daten:

Es seien $r_1, r_2 \in \mathbb{R} \setminus \{0\}$; sind $r_1', r_2' \in M$ die gerundeten Werte, so gilt also:

(5.7) $$\varepsilon_{r_1} \cdot |r_1| = |r_1 - r_1'| \quad , \quad \varepsilon_{r_2} \cdot |r_2| = |r_2 - r_2'| .$$

Ferner nehmen wir an, dass $r_1' + r_2'$, $r_1' - r_2'$, $r_1' \cdot r_2'$ und $\frac{r_1'}{r_2'}$ exakt berechnet werden.

Dann gelten folgende Abschätzungen für die relativen Fehler $\varepsilon_{r_1 \circ r_2}$, wobei $\circ$ für die 4 Grundrechenarten steht:

(5.7a) $$\varepsilon_{r_1 + r_2} \leq \frac{1}{|r_1 + r_2|} \cdot (\varepsilon_{r_1} \cdot |r_1| + \varepsilon_{r_2} \cdot |r_2|) ,$$
falls $r_1 + r_2 \neq 0$;

(5.7b) $$\varepsilon_{r_1 - r_2} \leq \frac{1}{|r_1 - r_2|} \cdot (\varepsilon_{r_1} \cdot |r_1| + \varepsilon_{r_2} \cdot |r_2|) ,$$
falls $r_1 \neq r_2$;

(5.7c) $$\varepsilon_{r_1 \cdot r_2} \leq \varepsilon_{r_1} + \varepsilon_{r_2} + \varepsilon_{r_1} \cdot \varepsilon_{r_2} \approx \varepsilon_{r_1} + \varepsilon_{r_2} \quad ;$$

(5.7d) $$\varepsilon_{r_1 : r_2} \leq \varepsilon_{r_1} + (\varepsilon_{r_2} + \varepsilon_{r_1} \cdot \varepsilon_{r_2}) \cdot \frac{|r_2|}{|r_2'|} \approx \varepsilon_{r_1} + \varepsilon_{r_2} .$$

Beweis:

Verifikation von (5.7a):

Wir erhalten im Falle $r_1 + r_2 \neq 0$:

$$\varepsilon_{r_1+r_2} = \frac{|(r_1+r_2)-(r_1'+r_2')|}{|r_1+r_2|} \leq \frac{|r_1-r_1'|+|r_2-r_2'|}{|r_1+r_2|}$$

$$= \frac{1}{|r_1+r_2|} \cdot (\varepsilon_{r_1} \cdot |r_1| + \varepsilon_{r_2} \cdot |r_2|) \quad .$$

Verifikation von (5.7b):

Ist $r_1 \neq r_2$, so folgt:

$$\varepsilon_{r_1-r_2} = \frac{|(r_1-r_2)-(r_1'-r_2')|}{|r_1-r_2|} \leq \frac{|r_1-r_1'|+|r_2'-r_2|}{|r_1-r_2|}$$

$$= \frac{1}{|r_1-r_2|} \cdot (\varepsilon_{r_1} \cdot |r_1| + \varepsilon_{r_2} \cdot |r_2|) \quad .$$

Verifikation von (5.7c):

Wir erhalten:

$$\varepsilon_{r_1 \cdot r_2} = \frac{|r_1 \cdot r_2 - r_1' \cdot r_2'|}{|r_1 \cdot r_2|} = \frac{|r_1 \cdot (r_2-r_2') + (r_1-r_1') \cdot r_2'|}{|r_1 \cdot r_2|}$$

$$= \frac{|r_1 \cdot (r_2-r_2') + (r_1-r_1') \cdot r_2 - (r_1-r_1') \cdot (r_2-r_2')|}{|r_1 \cdot r_2|}$$

$$\leq \frac{|r_1| \cdot |r_2-r_2'|}{|r_1| \cdot |r_2|} + \frac{|r_1-r_1'| \cdot |r_2|}{|r_1| \cdot |r_2|} + \frac{|r_1-r_1'| \cdot |r_2-r_2'|}{|r_1| \cdot |r_2|}$$

$$= \varepsilon_{r_2} + \varepsilon_{r_1} + \varepsilon_{r_1} \cdot \varepsilon_{r_2} \quad .$$

Verifikation von (5.7 d):

Es folgt:

$$\varepsilon_{r_1 : r_2} \cdot \left| \frac{r_1}{r_2} \right| = \left| \frac{r_1}{r_2} - \frac{r_1'}{r_2'} \right| = \frac{|r_2' \cdot (r_1 - r_1') - r_1' \cdot (r_2 - r_2')|}{|r_2 \cdot r_2'|}$$

$$= \frac{|r_2' \cdot (r_1 - r_1') - r_1 \cdot (r_2 - r_2') + (r_1 - r_1') \cdot (r_2 - r_2')|}{|r_2 \cdot r_2'|}$$

$$\leq \frac{|r_2'| \cdot |r_1 - r_1'|}{|r_2| \cdot |r_2'|} + \frac{|r_1| \cdot |r_2 - r_2'|}{|r_2| \cdot |r_2'|} + \frac{|r_1 - r_1'| \cdot |r_2 - r_2'|}{|r_2| \cdot |r_2'|}$$

Multiplikation mit $\frac{|r_2|}{|r_1|}$ liefert:

$$\varepsilon_{r_1 : r_2} \leq \frac{|r_1 - r_1'|}{|r_1|} + \frac{|r_2 - r_2'|}{|r_2'|} + \frac{|r_1 - r_1'|}{|r_1|} \cdot \frac{|r_2 - r_2'|}{|r_2'|}$$

$$= \varepsilon_{r_1} + (\varepsilon_{r_2} + \varepsilon_{r_1} \cdot \varepsilon_{r_2}) \cdot \frac{|r_2|}{|r_2'|} .$$

□

Bemerkung 5.8, Analyse der Fehlerfortpflanzung - Auslöschung

Bei der Multiplikation und der Division verhält sich der relative Fehler unproblematisch. Bei Vernachlässigung sehr kleiner Produkte wie $\varepsilon_{r_1} \cdot \varepsilon_{r_2}$ addieren sich die relativen Fehler im ungünstigsten Fall.

Die Addition ist ebenfalls unproblematisch, wenn r_1 und r_2 gleiches Vorzeichen haben; gegebenenfalls ist $\varepsilon_{r_1 + r_2} \leq \varepsilon_{r_1} + \varepsilon_{r_2}$.

Auch die Subtraktion ist unproblematisch, wenn $r_1 - r_2$ nicht ungefähr 0 ist.

Gilt jedoch $v_1 \approx v_2$, so kann $\varepsilon_{v_1 - v_2}$ gemäß (S. 7b) groß werden!
Dieses Verhalten heißt <u>Fehlerverstärkung durch Auslöschung</u>.

<u>Beispiel S. 9:</u>
Wir betrachten das folgende lineare Gleichungssystem:

$$3x = 10 \quad \wedge \quad 5x - \frac{1}{1000} \cdot y = 16{,}66 .$$

Die exakte Lösung berechnet sich wie folgt:

$$x = \frac{10}{3} = 3{,}\overline{3} \; ;$$

$$y = 1000 \cdot (16{,}\overline{6} - 16{,}66) = 1000 \cdot 0{,}00\overline{6} = 6{,}\overline{6} = 6\frac{2}{3} .$$

Ein Rechner mit vierstelliger Gleitpunktdarstellung liefert jedoch folgende gerundete Werte:

$$x' = 3{,}333 \; ;$$
$$y' = 1000 \odot (5 \odot 3{,}333 \ominus 16{,}66) = 1000 \cdot (16{,}67 - 16{,}66) = 10 .$$

Während der Wert x' noch möglichst genau ist (bei Berücksichtigung von nur 4 Stellen!), ergibt sich aber für y' ein mehr als nur ungenauer Wert!

Beispiel S. 10, Der Horner-Algorithmus

Sei $m \in \mathbb{N}$, seien $a_0, \dots, a_m \in \mathbb{R}$, und definiere die Polynom-Funktion $f: \mathbb{R} \to \mathbb{R}$ durch

$$f(x) := \sum_{j=0}^{m} a_j \cdot x^j = a_m \cdot x^m + \dots + a_1 \cdot x + a_0$$
$$= (\dots(a_m \cdot x + a_{m-1}) \cdot x + \dots + a_1) \cdot x + a_0 .$$

Für $x \in \mathbb{R}$ ist der Horner-Algorithmus zur Berechnung von $f(x)$ gegeben durch:

$s_0 := a_m$
for $j = 1$ until m do
$s_j := s_{j-1} \cdot x + a_{m-j}$.

Für $m = 2$ werden - bei exakten Anfangsdaten - folgende Werte geliefert:

Exakte Werte	Werte im Rechner
$s_0 := a_2$	$s_0 := a_2$
$s_1 := s_0 \cdot x + a_1$	$\tilde{s}_1 := s_0 \odot x \oplus a_1$
$s_2 := s_1 \cdot x + a_0$	$\tilde{s}_2 := \tilde{s}_1 \odot x \oplus a_0$

Aus (5.4̃c) folgt - für passende $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4 \in [-eps, eps]$:

$s_0 = a_2$,
$\tilde{s}_1 = ((s_0 \cdot x)(1+\varepsilon_1) + a_1)(1+\varepsilon_2)$
$= a_2 \cdot (1+\varepsilon_1)(1+\varepsilon_2) \cdot x + a_1 \cdot (1+\varepsilon_2)$,
$\tilde{s}_2 = ((\tilde{s}_1 \cdot x)(1+\varepsilon_3) + a_0)(1+\varepsilon_4)$
$= a_2 \cdot (1+\varepsilon_1)(1+\varepsilon_2)(1+\varepsilon_3)(1+\varepsilon_4) \cdot x^2$
$+ a_1 \cdot (1+\varepsilon_2)(1+\varepsilon_3)(1+\varepsilon_4) \cdot x + a_0 \cdot (1+\varepsilon_4)$.