1.5 Ersatzarithmetik

Durch Runden oder Abschneiden gelingt es, reelle Zahlen x mit $x_{min} \le |x| \le x_{max}$ in ein gegebenes Gleitpunktzahlensystem $F(b,t,e_{min},e_{max})$ abzubilden. Deshalb werden die Grundoperationen $\circ \in \{+,-,*,/\}$ oft durch

$$x \tilde{\circ} y = rd(x \circ y) \quad \text{für } x,y \in F,\ x_{min} \le |x \circ y| \le x_{max} \tag{1.15}$$

oder

$$x \tilde{\circ} y = tc(x \circ y) \quad \text{für } x,y \in F,\ x_{min} \le |x \circ y| \le x_{max} \tag{1.16}$$

auf Rechner realisiert (bei der Division soll $y \neq 0$ sein).

Theorem 1.7

Bezüglich der durch (1.15) bzw. (1.16) definierten Ersatzoperationen $\tilde{+}, \tilde{-}, \tilde{*}, \tilde{/}$ in F abgeschlossen, d.h. im Ergebnis dieser Operationen erhält man Elemente aus F.

Außerdem gilt die Beziehung bzw. Darstellung

$$x \tilde{\circ} y = (x \circ y)(1+\varepsilon) \quad \text{mit } |\varepsilon| \le k\,eps, \tag{1.17}$$

wobei k im Fall von (1.15) gleich 1 und im Fall von (1.15) gleich 2 ist (Ehrlich: Darstellungsfehler)

Beweis

Die Abgeschlossenheit von F bezügl. $\tilde{\circ}$ folgt aus Th. 1.3

Die Darstellung (1.16) ergibt sich im Falle von (1.15) aus

$$\frac{|rd(x \circ y) - (x \circ y)|}{|x \circ y|} \le eps,$$

also aus (1.14)

1.6 Fehlerakkumulation

Wir betrachten Zahlen $x, y \in \mathbb{R}$. Durch eine evtl. Rundung erhalten wir mit

$$rd(x) = x + \Delta x \in F \qquad \frac{|\Delta x|}{|x|} \le \varepsilon$$
$$rd(y) = y + \Delta y \in F \quad \text{mit} \quad \frac{|\Delta y|}{|y|} \le \varepsilon$$

Zahlen aus einem Gleitpunkt-Zahlensystem F.
$\tilde{\circ}, \circ$ sei nun Multiplikation oder Division.
Mit (1.15) und (1.17) erhält man

$$\begin{aligned}(x+\Delta x)\,\tilde{\circ}\,(y+\Delta y) &= \big(x(1+\tau_x)\big)\,\tilde{\circ}\,\big(y(1+\tau_y)\big), \quad |\tau_x|, |\tau_y| \le \varepsilon \\ &= (x \circ y)\big((1+\tau_x)\circ(1+\tau_y)\big)(1+\alpha), \quad |\alpha| \le \varepsilon \\ &= (x \circ y)(1+\beta),\end{aligned}$$

wobei man benutzt, dass

$$(1+\tau_x)\circ(1+\tau_y)(1+\alpha) = 1+\beta$$

mit einem β mit der Eigenschaft $|\beta| \le \frac{3\varepsilon}{1-3\varepsilon}$ gilt (Beweis: Platz). Damit ergibt sich für die Multiplikation/Division das

Theorem 1.8
In einem Gleitpunkt-Zahlensystem $F(b, t, e_{min}, e_{max})$ seien die Zahlen $x, y \in \mathbb{R}$ und $\Delta x, \Delta y \in \mathbb{R}$ gegeben mit

$$x + \Delta x \in F, \quad y + \Delta y \in F, \quad \frac{|\Delta x|}{|x|} \le \varepsilon, \quad \frac{|\Delta y|}{|y|} \le \varepsilon \text{ mit } \varepsilon \le \tfrac{1}{4}$$

$\circ$ stehe für die Grundoperation $\times$ bzw. $/$ und für $x \circ y$ soll $x_{min} \le |x \circ y| \le x_{max}$ gelten.
Dann gilt die Fehlerdarstellung

$$(x+\Delta x)\,\tilde{\circ}\,(y+\Delta y) = x \circ y + \eta \quad \text{mit} \quad \frac{|\eta|}{|x \circ y|} \le \frac{3\varepsilon}{1-4\varepsilon} \tag{1.18}$$

Die Darstellung (1.18) zeigt, dass die Multiplikation bzw. Division verhältnismäßig gutartig mit einem kleinen rel. Fehler ist.

Im Folgenden soll die Fehlerverstärkung bei der Hintereinanderausführung von Additionen in einer gegebenen Gleitpunkt-Zahlenmenge F betrachtet werden. $\tilde{+}$ und $\sum^{\sim}$ sollen für die Addition in F und die Summation von links nach rechts in F stehen.

Theorem 1.9

Es $F(b,t,e_{min},e_{max})$ seien $x_1,\dots,x_n \in \mathbb{R}$ und $\Delta x_1,\dots,\Delta x_n \in \mathbb{R}$ Zahlen mit

$$x_k + \Delta x_k \in F, \quad \frac{|\Delta x_k|}{|x_k|} \le \varepsilon \quad \text{für } k=1,\dots,n$$

und es bezeichne

$$\tilde{S}_k := \sum_{j=1}^{k}{}^{\sim} (x_j + \Delta x_j), \quad S_k := \sum_{j=1}^{k} x_j, \quad k=1,\dots,n$$

die entsprechenden Partialsummen (Summation von links nach rechts). Dann gilt

$$|\tilde{S}_k - S_k| \le \underbrace{\left(\sum_{j=1}^{k} (1+\varepsilon)^{k-j} (2|x_j| + |S_j|)\right)}_{=:M_k} \varepsilon \quad \text{für } k=1,\dots,n \tag{1.19}$$

falls die Partialsummen (Notation $M_0 = 0$) innerhalb gewisser Schranken liegen:

$$x_{min} + (M_{k-1} + |x_k|)\varepsilon \le |S_k| \le x_{max} - (M_{k-1} + |x_k|)\varepsilon$$
$$k=1,\dots,n$$

Beweis:

Vollst. Induktion über k

(1.19) gilt für $k=1$, denn $|\tilde{S}_1 - S_1| = |\Delta x_1| \le 3|x_1|\varepsilon$

Wir nehmen jetzt an, dass (1.19) für ein $k \ge 1$ richtig ist

Mit der Notation

$$\Delta S_j := \tilde{S}_j - S_j \quad \text{für } j \geq 1, \quad \Delta S_0 = 0$$

berechnet man mit einer gewissen Zahl $\tau_k \in \mathbb{R}, |\tau_k| \leq \varepsilon$

$$\begin{aligned}\Delta S_k &= \tilde{S}_k - S_k = \tilde{S}_{k-1} \tilde{+} (x_k + \Delta x_k) - S_k \\ &= (S_{k-1} + \Delta S_{k-1}) \tilde{+} (x_k + \Delta x_k) - S_k \\ &= (S_k + \Delta S_{k-1} + \Delta x_k)(1+\tau_k) - S_k \\ &= (1+\tau_k)\Delta S_{k-1} + \tau_k S_k + (1+\tau_k)\Delta x_k\end{aligned}$$

und damit

$$|\Delta S_k| \leq (1+\varepsilon)|\Delta S_{k-1}| + \varepsilon(|S_k| + 2|x_k|) \qquad (1.20)$$

Aus (1.20) und der Induktionsannahme folgt die Behauptung (1.19) des Theorems

Bemerkung 1.10

Der Faktor $(1+\varepsilon)^{n-j}$ in der Abschätzung (1.19) ist umso größer, je kleiner j ist. Daher ist es vorteilhaft beim Aufsummieren mit den betragsmäßig kleinen Zahlen zu beginnen. Dies gewährleistet zudem, dass die Partialsumme S_k betragsmäßig nicht unnötig anwachsen

Theorem 1.9 liefert mit (1.19) nur eine Abschätzung für den absoluten Fehler. Der relative Fehler $\frac{|\tilde{S}_n - S_n|}{|S_n|}$ kann jedoch groß ausfallen, falls $|S_n|$ klein gegenüber $\sum_{j=1}^{n-1}(|x_j| + |S_j|) + |x_n|$ ist !