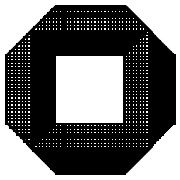

UNIVERSITÄT KARLSRUHE

Ein Kalkül für verläßliche absolute und relative Fehlerabschätzungen

A. Bantle und W. Krämer

Preprint Nr. 98/5

Institut für Wissenschaftliches Rechnen
und Mathematische Modellbildung

76128 Karlsruhe

Anschrift der Verfasser:

Armin Bantle

Walter Krämer
Institut für Wissenschaftliches Rechnen und
Mathematische Modellbildung (IWRMM)

Universität Karlsruhe
Postfach 6980
76128 Karlsruhe

Bundesrepublik Deutschland

Das Postscript-File dieses Preprints ist über FTP unter der Adresse

`iamk4515.mathematik.uni-karlsruhe.de`

im Verzeichnis

`/pub/iwrmm/preprints`

abrufbar.

Inhaltsverzeichnis

1 Zusammenfassung **2**

2 Voraussetzungen an die Arithmetik **3**

3 Fehlerschranken für Grundoperationen **4**

3.1 Absolute Fehlerschranken 4

3.1.1 Abschätzung des fortgepflanzten Datenfehlers 5

3.1.2 Abschätzung des Rundungsfehlers 7

3.1.3 Gesamtfehlerabschätzung 9

3.2 Relative Fehlerschranken 9

3.2.1 Abschätzung des fortgepflanzten Datenfehlers 10

3.2.2 Abschätzung des Rundungsfehlers 13

3.2.3 Gesamtfehlerabschätzung 15

3.3 Sonderfälle: Operanden, die keine Rundungsfehler verursachen 15

4 Fehlerschranken für Funktionen **20**

4.1 Allgemeine Fehlerabschätzung 20

4.2 Konkrete Fehlerschranken für einige mathematische Standardfunktionen 22

4.2.1 Die Funktion $\mathrm{sqrt}(x) = \sqrt{x}$ 22

4.2.2 Die Funktion $\exp(x) = e^x$ 24

4.2.3 Die Funktion $\ln(x)$ 25

5 Zusammenfassung **27**

6 Anhang: Notation und Bezeichnungen **29**

Literaturverzeichnis 31

1 Zusammenfassung

Die Arbeit befaßt sich mit der rigorosen Fehleranalyse numerischer Algorithmen, die das im IEEE-Standard 754-1985 [5], [14] festgelegte Datenformat doppelt genauer Gleitpunktzahlen (64 Bit) verwenden. Es werden Fehlerschranken sowohl für absolute als auch für relative Fehler hergeleitet. Dabei werden auch Operationen, die zu (Zwischen-)Ergebnissen im Unterlaufbereich führen, durch geeignete worst case Abschätzungen sicher erfaßt. Die Methodik kann in Verbindung mit einer Maschinenintervallrechnung und unter Verwendung eines Operatorkonzeptes verwendet werden, um auf elegante Weise Gesamtfehlerabschätzungen für Gleitkommaalgorithmen automatisch vom Rechner durchführen zu lassen. Auch für bereits existierende Programme können mit nur minimalen Quellcodeanpassungen (im wesentlichen durch Abändern von Datentypen) verläßliche Vorwärtsfehleranalysen durchgeführt werden. Eine so gefundene Gesamtfehlerschranke berücksichtigt alle Eingangs-, Konstanten- und Rundungsfehler. Sie ist eine abgesicherte worst case Schranke, die gleichmäßig über dem gesamten untersuchten Datenbereich (bzw. Parameterbereich) gilt.

Eine Umsetzung der hier beschriebenen Abschätzungen in eine Softwarebibliothek findet sich in [2]. Die notwendigen Routinen sind in der Programmiersprache C++ unter Verwendung der Klassenbibliothek C-XSC [16] realisiert. Das Überladen von Operatoren erlaubt die elegante Anwendung des Fehlerkalküls. Beispiele finden sich u. a. in [2], [11], [12], [13], [19]. Sie belegen die hohe numerische Güte der gefundenen Fehlerschranken.

Key Words: Rounding Errors, Reliable Error Estimates, Floating-point Calculations, Absolute Error Bounds, Relative Error Bounds, Automatic Error Analysis

MSC: 65G05, 65G10, 68M15

2 Voraussetzungen an die Arithmetik

Der in den Paragraphen 3 und 4 vorgestellte Fehlerkalkül ermöglicht sichere a priori Fehlerabschätzungen für mathematische Grundoperationen und Funktionen. Dabei sind die mit dem Kalkül berechneten Fehlerschranken gleichmäßige Schranken. Sie gelten jeweils für beliebige Argumente, sofern diese den in den Abschätzungen zugrundegelegten Datenbereich (Parameterbereich) nicht verlassen.

Es werden hier die bereits in den Arbeiten [12], [11], [17], [13], [19] verwendeten Bezeichnungen und Notationen verwendet (vgl. Anhang).

Im folgenden mögen als Generalvoraussetzungen gelten:

$$
\begin{aligned}
&\circ \in \{+,-,\cdot,/\}, \quad a \in A \in I\!I\!R, \quad b \in B \in I\!I\!R \quad \text{sowie} \\
&\tilde{a} = a + \Delta_a = a(1+\varepsilon_a) \quad \text{mit} \quad |\Delta_a| \le \Delta(a), |\varepsilon_a| \le \varepsilon(a) = \Delta(a)/|a| \text{ und} \\
&\tilde{b} = b + \Delta_b = b(1+\varepsilon_b) \quad \text{mit} \quad |\Delta_b| \le \Delta(b), |\varepsilon_b| \le \varepsilon(b) = \Delta(b)/|b|, \text{ also} \qquad \text{(GV)} \\
&\tilde{a} \in \tilde{A} := A + [-\Delta(a), \Delta(a)] = A \cdot [1-\varepsilon(a), 1+\varepsilon(a)], \\
&\tilde{b} \in \tilde{B} := B + [-\Delta(b), \Delta(b)] = B \cdot [1-\varepsilon(b), 1+\varepsilon(b)].
\end{aligned}
$$

Dabei seien die Intervalle $A = [\underline{a}, \overline{a}]$ und $B = [\underline{b}, \overline{b}]$ ($\underline{a}, \overline{a}, \underline{b}, \overline{b} \in I\!R$), die absoluten Fehlerschranken $\Delta(a)$ und $\Delta(b)$ und die relativen Fehlerschranken $\varepsilon(a)$ und $\varepsilon(b)$ gegeben (falls die jeweilige Größe in der betrachteten Formel oder Ungleichung auftritt).

Bei der Herleitung der Fehlerschranken wird vorausgesetzt, daß die Operationen des verwendeten Rechners dem IEEE-Standard genügen[1]. Insbesondere hat man also für die Grundoperation $\circ \in \{+,-,\cdot,/\}$ und deren Maschinenanalogon $\boxed{\circ} \in \{\boxplus, \boxminus, \boxdot, \boxslash\}$ die Abschätzung

$$
\left| \frac{(\tilde{a} \circ \tilde{b}) - (\tilde{a} \boxed{\circ} \tilde{b})}{\tilde{a} \circ \tilde{b}} \right| \le \overline{\varepsilon} \tag{1}
$$

für alle $\tilde{a}, \tilde{b} \in S$ mit $|\tilde{a} \circ \tilde{b}| \in [\texttt{MinReal}, \texttt{MaxReal}]$. $\overline{\varepsilon}$ bedeutet für gerichtet gerundete Operationen das Zweifache der Maschinengenauigkeit, bei Rundung zur nächsten Zahl die Maschinengenauigkeit ε^* selbst. Es wird außerdem vorausgesetzt, daß kein Überlauf auftritt[2], d. h. $a \circ b, \tilde{a} \boxed{\circ} \tilde{b} \in [-\texttt{MaxReal}, \texttt{MaxReal}]$ sei stets erfüllt.

Wenn $\tilde{a} \circ \tilde{b}$ im Unterlaufbereich $U := (-\texttt{MinReal}, \texttt{MinReal})$ liegt, gilt obige Abschätzung für den relativen Fehler einer Maschinenoperation i. allg. nicht mehr. Das

[1] Das hier vorgestellte methodische Vorgehen kann ohne Schwierigkeiten auf andere Gleitkommaraster übertragen werden.

[2] Durch die Routinen der Fehlerabschätzungsbibliothek wird der Fall eines Überlaufs automatisch erkannt. Es wird dann mit einer Fehlermeldung abgebrochen.

folgende Lemma liefert für diesen Fall zumindest eine Schranke für den absoluten Fehler.

Lemma 2.1 *Im Unterlaufbereich* $U := (-\texttt{MinReal}, \texttt{MinReal})$ *gilt die Abschätzung*

$$|(\tilde{a} \circ \tilde{b}) - (\tilde{a} \boxdot \tilde{b})| \leq \texttt{MinReal} \tag{2}$$

für alle $\circ \in \{+, -, \cdot, /\}$ *und* $\tilde{a}, \tilde{b} \in S$ *mit* $|\tilde{a} \circ \tilde{b}| < \texttt{MinReal}$.

Beweis: $\tilde{a} \circ \tilde{b}$ und $\tilde{a} \boxdot \tilde{b}$ haben das gleiche Vorzeichen, d. h. es gilt

$$\begin{aligned} &|(\tilde{a} \circ \tilde{b}) - (\tilde{a} \boxdot \tilde{b})| \in [0, \texttt{MinReal}] \\ &\quad \Longrightarrow |(\tilde{a} \circ \tilde{b}) - (\tilde{a} \boxdot \tilde{b})| \leq |[0, \texttt{MinReal}]| = \texttt{MinReal} \end{aligned}$$ ■

Bemerkung 2.1 Falls eine Arithmetik mit „gradual underflow“ verwendet wird und die Arithmetik auch im Unterlaufbereich maximal genaue Grundoperationen zur Verfügung stellt, gilt offensichtlich $|(\tilde{a} \circ \tilde{b}) - (\tilde{a} \boxdot \tilde{b})| \leq \texttt{dMinReal}$ für alle $\circ \in \{+, -, \cdot, /\}$ und $\tilde{a}, \tilde{b} \in S$ mit $|\tilde{a} \circ \tilde{b}| < \texttt{MinReal}$.

Über die $(1+\varepsilon)$–Eigenschaft (1) hinausgehend, wird in dieser Arbeit vorausgesetzt, daß die zugrundeliegende Rechnerarithmetik im Falle eines exakt darstellbaren Resultates diesen Wert als Maschinenergebnis liefert, d. h., daß

$$\tilde{a} \circ \tilde{b} \in S \Longrightarrow \tilde{a} \boxdot \tilde{b} = \tilde{a} \circ \tilde{b}. \tag{3}$$

Rechnerarithmetiken, die dem IEEE-Standard genügen, besitzen diese Eigenschaft.

3 Fehlerschranken für Grundoperationen

3.1 Absolute Fehlerschranken

Für das Ergebnis der Maschinenverknüpfung $\tilde{a} \boxdot \tilde{b}$ soll nun eine absolute Fehlerschranke $\Delta(\circ) \in I\!R^+$ berechnet werden, so daß

$$|(a \circ b) - (\tilde{a} \boxdot \tilde{b})| \leq \Delta(\circ)$$

für alle $a \in A$ mit $|a - \tilde{a}| \leq \Delta(a)$ bzw. $|a - \tilde{a}| \leq \varepsilon(a)|a|$ und für alle $b \in B$ mit $|b - \tilde{b}| \leq \Delta(b)$ bzw. $|b - \tilde{b}| \leq \varepsilon(b)|b|$ gilt.

3.1.1 Abschätzung des fortgepflanzten Datenfehlers

Satz 3.1 *Für die Fortpflanzung des absoluten Datenfehlers* $|(a \circ b) - (\tilde{a} \circ \tilde{b})|$ *gilt bei exakter Rechnung*

$$|(a+b)-(\tilde{a}+\tilde{b})| \leq \Delta(a)+\Delta(b) =: \Delta_{dat,+} \tag{4}$$

$$|(a-b)-(\tilde{a}-\tilde{b})| \leq \Delta(a)+\Delta(b) =: \Delta_{dat,-} \tag{5}$$

$$|(a\cdot b)-(\tilde{a}\cdot\tilde{b})| \leq \Delta(a)\Delta(b)+|a|\Delta(b)+|b|\Delta(a) =: \Delta_{dat,\cdot} \tag{6}$$

$$|(a/b)-(\tilde{a}/\tilde{b})| \leq \frac{\Delta(a)+|\frac{a}{b}|\Delta(b)}{|b|-\Delta(b)} =: \Delta_{dat,/} \quad \textit{für} \quad \Delta(b)<|b|\,. \tag{7}$$

Beweis:

$$\begin{aligned} \textit{Add.:}\quad |(a+b)-(\tilde{a}+\tilde{b})| &= |a+b-(a+\Delta_a+b+\Delta_b)| \\ &= |\Delta_a+\Delta_b| \\ &\leq \Delta(a)+\Delta(b)=\Delta_{dat,+} \end{aligned}$$

$$\begin{aligned} \textit{Sub.:}\quad |(a-b)-(\tilde{a}-\tilde{b})| &= |a-b-(a+\Delta_a-b-\Delta_b)| \\ &= |\Delta_a-\Delta_b| \\ &\leq \Delta(a)+\Delta(b)=\Delta_{dat,-} \end{aligned}$$

$$\begin{aligned} \textit{Mul.:}\quad |(a\cdot b)-(\tilde{a}\cdot\tilde{b})| &= |ab-(a+\Delta_a)(b+\Delta_b)| \\ &= |a\Delta_b+b\Delta_a+\Delta_a\Delta_b| \\ &\leq \Delta(a)\Delta(b)+|a|\Delta(b)+|b|\Delta(a)=\Delta_{dat,\cdot} \end{aligned}$$

$$\begin{aligned} \textit{Div.:}\quad |(a/b)-(\tilde{a}/\tilde{b})| &= \left|\frac{\frac{a}{b}(b+\Delta_b)-(a+\Delta_a)}{b+\Delta_b}\right| = \left|\frac{\frac{a}{b}\Delta_b-\Delta_a}{b+\Delta_b}\right| \\ &\leq \frac{\Delta(a)+|\frac{a}{b}|\Delta(b)}{|b|-\Delta(b)}=\Delta_{dat,/}\,,\ \text{falls } \Delta(b)<|b| \end{aligned}$$

■

Mit Hilfe von Satz 3.1 lassen sich nun gleichmäßige Schranken $\Delta_{dat}(+)$, $\Delta_{dat}(-)$, $\Delta_{dat}(\cdot)$ und $\Delta_{dat}(/)$ für die Fortpflanzung des absoluten Datenfehlers für alle $a \in A$ und $b \in B$ herleiten:

Fall I Für $|a-\tilde{a}|$ und $|b-\tilde{b}|$ sind die absoluten Schranken $\Delta(a)$ und $\Delta(b)$ gegeben.

$$\Delta_{dat,+} = \Delta(a)+\Delta(b) =: \Delta_{dat}(+)$$

$$
\begin{aligned}
\Delta_{dat,-} &= \Delta(a) + \Delta(b) =: \Delta_{dat}(-) \\
\Delta_{dat,\cdot} &= \Delta(a)\Delta(b) + |a|\Delta(b) + |b|\Delta(a) \\
&\leq \Delta(a)\Delta(b) + |A|\Delta(b) + |B|\Delta(a) =: \Delta_{dat}(\cdot) \\
\Delta_{dat,/} &= \frac{\Delta(a) + |\frac{a}{b}|\Delta(b)}{|b| - \Delta(b)} \leq \frac{\Delta(a) + \frac{|A|}{\langle B\rangle}\Delta(b)}{\langle B\rangle - \Delta(b)} =: \Delta_{dat}(/) \quad \text{für} \quad \Delta(b) < \langle B\rangle
\end{aligned}
$$

Fall II Für $|a-\tilde{a}|$ ist die absolute Schranke $\Delta(a)$ und für $|b-\tilde{b}|$ die relative Schranke $\varepsilon(b)$ gegeben.

$$
\begin{aligned}
\Delta_{dat,+} &= \Delta(a) + |b|\varepsilon(b) \leq \Delta(a) + |B|\varepsilon(b) =: \Delta_{dat}(+) \\
\Delta_{dat,-} &= \Delta(a) + |b|\varepsilon(b) \leq \Delta(a) + |B|\varepsilon(b) =: \Delta_{dat}(-) \\
\Delta_{dat,\cdot} &= \Delta(a) \cdot |b|\varepsilon(b) + |a| \cdot |b|\varepsilon(b) + |b|\Delta(a) \\
&\leq |B|\Big(\Delta(a)\varepsilon(b) + |A|\varepsilon(b) + \Delta(a)\Big) =: \Delta_{dat}(\cdot) \\
\Delta_{dat,/} &= \frac{\Delta(a) + |\frac{a}{b}| \cdot |b|\varepsilon(b)}{|b| - |b|\varepsilon(b)} \leq \frac{\Delta(a) + |A|\varepsilon(b)}{\langle B\rangle(1-\varepsilon(b))} =: \Delta_{dat}(/) \quad \text{für} \quad \varepsilon(b) < 1
\end{aligned}
$$

Fall III Für $|a-\tilde{a}|$ ist die relative Schranke $\varepsilon(a)$ und für $|b-\tilde{b}|$ die absolute Schranke $\Delta(b)$ gegeben.

$$
\begin{aligned}
\Delta_{dat,+} &= |a|\varepsilon(a) + \Delta(b) \leq |A|\varepsilon(a) + \Delta(b) =: \Delta_{dat}(+) \\
\Delta_{dat,-} &= |a|\varepsilon(a) + \Delta(b) \leq |A|\varepsilon(a) + \Delta(b) =: \Delta_{dat}(-) \\
\Delta_{dat,\cdot} &= |a|\varepsilon(a) \cdot \Delta(b) + |a|\Delta(b) + |b| \cdot |a|\varepsilon(a) \\
&\leq |A|\Big(\varepsilon(a)\Delta(b) + \Delta(b) + |B|\varepsilon(a)\Big) =: \Delta_{dat}(\cdot) \\
\Delta_{dat,/} &= \frac{|a|\varepsilon(a) + |\frac{a}{b}|\Delta(b)}{|b| - \Delta(b)} \leq |A|\frac{\varepsilon(a) + \frac{\Delta(b)}{\langle B\rangle}}{\langle B\rangle - \Delta(b)} =: \Delta_{dat}(/) \quad \text{für} \quad \Delta(b) < \langle B\rangle
\end{aligned}
$$

Fall IV Für $|a-\tilde{a}|$ und $|b-\tilde{b}|$ sind die relativen Schranken $\varepsilon(a)$ und $\varepsilon(b)$ gegeben.

$$
\begin{aligned}
\Delta_{dat,+} &= |a|\varepsilon(a) + |b|\varepsilon(b) \leq |A|\varepsilon(a) + |B|\varepsilon(b) =: \Delta_{dat}(+) \\
\Delta_{dat,-} &= |a|\varepsilon(a) + |b|\varepsilon(b) \leq |A|\varepsilon(a) + |B|\varepsilon(b) =: \Delta_{dat}(-) \\
\Delta_{dat,\cdot} &= |a|\varepsilon(a) \cdot |b|\varepsilon(b) + |a| \cdot |b|\varepsilon(b) + |b| \cdot |a|\varepsilon(a) \\
&\leq |A| \cdot |B| \cdot \Big(\varepsilon(a)\varepsilon(b) + \varepsilon(b) + \varepsilon(a)\Big) =: \Delta_{dat}(\cdot) \\
\Delta_{dat,/} &= \frac{|a|\varepsilon(a) + |\frac{a}{b}| \cdot |b|\varepsilon(b)}{|b| - |b|\varepsilon(b)} \leq \frac{|A|(\varepsilon(a) + \varepsilon(b))}{\langle B\rangle(1-\varepsilon(b))} =: \Delta_{dat}(/) \quad \text{für} \quad \varepsilon(b) < 1
\end{aligned}
$$

In Tabelle 1 sind die Ergebnisse noch einmal zusammengefaßt.

gegeben	$\Delta_{dat}(\pm)$	$\Delta_{dat}(\cdot)$	$\Delta_{dat}(/)$
$\Delta(a), \Delta(b)$	$\Delta(a) + \Delta(b)$	$\Delta(a)\Delta(b) + \lvert A\rvert\Delta(b) + \lvert B\rvert\Delta(a)$	$\frac{\Delta(a)+\frac{\lvert A\rvert}{\langle B\rangle}\Delta(b)}{\langle B\rangle - \Delta(b)}$
$\Delta(a), \varepsilon(b)$	$\Delta(a) + \lvert B\rvert\varepsilon(b)$	$\lvert B\rvert\Big(\Delta(a)\varepsilon(b) + \lvert A\rvert\varepsilon(b) + \Delta(a)\Big)$	$\frac{\Delta(a)+\lvert A\rvert\varepsilon(b)}{\langle B\rangle(1-\varepsilon(b))}$
$\varepsilon(a), \Delta(b)$	$\lvert A\rvert\varepsilon(a) + \Delta(b)$	$\lvert A\rvert\Big(\varepsilon(a)\Delta(b) + \Delta(b) + \lvert B\rvert\varepsilon(a)\Big)$	$\lvert A\rvert\frac{\varepsilon(a)+\frac{\Delta(b)}{\langle B\rangle}}{\langle B\rangle - \Delta(b)}$
$\varepsilon(a), \varepsilon(b)$	$\lvert A\rvert\varepsilon(a) + \lvert B\rvert\varepsilon(b)$	$\lvert A\rvert \cdot \lvert B\rvert \cdot \Big(\varepsilon(a)\varepsilon(b) + \varepsilon(b) + \varepsilon(a)\Big)$	$\frac{\lvert A\rvert(\varepsilon(a)+\varepsilon(b))}{\langle B\rangle(1-\varepsilon(b))}$

Tabelle 1: Schranken für die Fortpflanzung des absoluten Datenfehlers bei den Grundoperationen

3.1.2 Abschätzung des Rundungsfehlers

Satz 3.2 *Für den Rundungsfehler* $|(\tilde{a} \circ \tilde{b}) - (\tilde{a} \boxdot \tilde{b})|$ *einer Operation* $\circ \in \{+, -, \cdot, /\}$ *mit den Maschinenzahlen* $\tilde{a}, \tilde{b} \in S$ *gilt*

a) im Unterlaufbereich, d. h. $\tilde{a} \boxdot \tilde{b} \in U$ *und damit* $\tilde{a} \circ \tilde{b} \in U$*:*

$$|(\tilde{a} \circ \tilde{b}) - (\tilde{a} \boxdot \tilde{b})| \;\leq\; \texttt{MinReal} =: \Delta_{rnd,U}(\circ)\,. \tag{8}$$

Bei Verwendung einer Arithmetik mit maximal genauen Operationen im „gradual underflow" kann die Abschätzung noch verschärft werden zu:

$$|(\tilde{a} \circ \tilde{b}) - (\tilde{a} \boxdot \tilde{b})| \;\leq\; \left\{ \begin{array}{cc} 0 & \textit{für } \circ \in \{+, -\} \\ \texttt{dMinReal} & \textit{für } \circ \in \{\cdot, /\} \end{array} \right\} =: \Delta_{rnd,U}(\circ) \tag{9}$$

b) im normalisierten Bereich, d. h. $\tilde{a} \boxdot \tilde{b} \notin U$ *und damit* $\tilde{a} \circ \tilde{b} \notin U$*:*

$$|(\tilde{a} \circ \tilde{b}) - (\tilde{a} \boxdot \tilde{b})| \;\leq\; \overline{\varepsilon} \cdot (\Delta_{dat}(\circ) + |A \circ B|) \;=:\; \Delta_{rnd,N}(\circ) \tag{10}$$

für jedes $\circ \in \{+, -, \cdot, /\}$*.*

Beweis:

a) Ungleichung (8) folgt sofort aus Lemma 2.1.
Für $\circ \in \{\cdot, /\}$ ist Ungleichung (9) offensichtlich (siehe Bemerkung 2.1).
Seien also $\circ = +$ und $\tilde{a}, \tilde{b} \in S$ mit $\tilde{a} \boxplus \tilde{b} \in U$. Man kann ohne Beschränkung der Allgemeinheit $\tilde{a}, \tilde{b} \geq 0$ annehmen. (Für $\tilde{a}, \tilde{b} \leq 0$ betrachte $\tilde{a} \boxplus \tilde{b} =$

$-((-\tilde{a}) \boxplus (-\tilde{b}))$. Die beiden anderen Fälle kommen einer Subtraktion gleich, die weiter unten behandelt wird.) Es folgt

$$0 \leq \tilde{a}, \tilde{b} \leq \tilde{a} + \tilde{b} < \texttt{MinReal},$$

d. h. $\tilde{a}$, $\tilde{b}$ und $\tilde{a} + \tilde{b}$ haben denselben Exponenten: $e(\tilde{a}) = e(\tilde{b}) = e(\tilde{a} + \tilde{b}) = e_{min} - 1$. Die Summe $\tilde{a} \boxplus \tilde{b}$ ist daher exakt.
Seien nun $\circ = -$ und $\tilde{a}, \tilde{b} \in S$ mit $\tilde{a} \boxminus \tilde{b} \in U$. Wieder sei o. B. d. A. $\tilde{a}, \tilde{b} \geq 0$ und zusätzlich $\tilde{a} \geq \tilde{b}$. Es sind drei Fälle zu unterscheiden:

Fall I $e(\tilde{a}) = e(\tilde{b})$: klar!

Fall II(a) $e(\tilde{a}) = e(\tilde{b}) + 1$ und $\tilde{b} \in U$:
Es folgt $\tilde{a} = 1.a_1 \ldots a_{52} \cdot 2^{e_{min}}$ und $\tilde{b} = 0.b_1 \ldots b_{52} \cdot 2^{e_{min}}$, d. h.

$$\begin{array}{rr} & 1.a_1 \ldots a_{52} \cdot 2^{e_{min}} \\ - & 0.b_1 \ldots b_{52} \cdot 2^{e_{min}} \\ \hline = & c_0.c_1 \ldots c_{52} \cdot 2^{e_{min}} \end{array}$$

Da aber $\tilde{a} - \tilde{b} \in U$, ist $c_0 = 0$, und das exakte Ergebnis ist ohne Rundung im Gleitpunktsystem darstellbar.

Fall II(b) $e(\tilde{a}) = e(\tilde{b}) + 1$ und $\tilde{b} \notin U$:
Es folgt $\tilde{a} = a_0.a_1 \ldots a_{52} \cdot 2^{e(\tilde{a})}$ und $\tilde{b} = 0.b_0 b_1 \ldots b_{52} \cdot 2^{e(\tilde{a})}$ mit $e(\tilde{a}) = e(\tilde{b}) + 1 \geq e_{min} + 1$, d. h.

$$\begin{array}{rr} & a_0.a_1 \ldots a_{52} \quad \cdot 2^{e(\tilde{a})} \\ - & 0.b_0 \ldots b_{51} b_{52} \cdot 2^{e(\tilde{a})} \\ \hline = & c_0.c_1 \ldots c_{52} c_{53} \cdot 2^{e(\tilde{a})} \end{array}$$

Wegen $\tilde{a} - \tilde{b} \in U$ ist $e(\tilde{a} - \tilde{b}) = e_{min} - 1 \leq e(\tilde{a}) - 2$. Es muß also $c_0 = c_1 = 0$ gelten; das exakte Ergebnis ist ohne Rundung im Gleitpunktsystem darstellbar.

Fall III $e(\tilde{a}) \geq e(\tilde{b}) + 2$:
Dann gilt $\tilde{a} \notin U \Rightarrow \tilde{a} = 1.a_1 \ldots a_{52} \cdot 2^{e(\tilde{a})} \geq 2^{e(\tilde{a})} \geq 2^{e(\tilde{b})+2}$ und $\tilde{b} = b_0.b_1 \ldots b_{52} \cdot 2^{e(\tilde{b})} \leq 2^{e(\tilde{b})+1}$. Daher ist $\tilde{a} - \tilde{b} \geq 2^{e(\tilde{b})+1} \geq 2^{e_{min}} \notin U$, d. h. dieser Fall kann nicht eintreten, wenn die Differenz $\tilde{a} - \tilde{b}$ im Unterlauf liegen soll.

Es folgt in jedem Falle die Behauptung.

b) Nach Ungleichung (1) gilt

$$\begin{aligned} |(\tilde{a} \circ \tilde{b}) - (\tilde{a} \boxed{\circ} \tilde{b})| &\leq \overline{\varepsilon} \cdot |\tilde{a} \circ \tilde{b}| \\ &\leq \overline{\varepsilon} \cdot (|(a \circ b) - (\tilde{a} \circ \tilde{b})| + |a \circ b|) \\ &\leq \overline{\varepsilon} \cdot (\Delta_{dat}(\circ) + |A \circ B|) = \Delta_{rnd,N}(\circ)\,, \end{aligned}$$

womit die Behauptung gezeigt wäre ■

Beispiel 3.1 (zu Satz 3.2a) Gegeben sei das Gleitpunktsystem $S := S(2, 4, -6, 7)$ sowie die Maschinenzahlen $\tilde{a} := 1.011 \cdot 2^{-5} \in S$ und $\tilde{b} := 1.111 \cdot 2^{-6} \in S$. Die Differenz $\tilde{a} - \tilde{b} = 10.110 \cdot 2^{-6} - 1.111 \cdot 2^{-6} = 0.111 \cdot 2^{-6} \in U := (-2^{-6}, 2^{-6})$ ist ohne Rundung darstellbar und daher exakt.

3.1.3 Gesamtfehlerabschätzung

Mit der Dreiecksungleichung läßt sich der absolute Gesamtfehler folgendermaßen abschätzen:

$$|(a \circ b) - (\tilde{a} \boxdot \tilde{b})| \leq |(a \circ b) - (\tilde{a} \circ \tilde{b})| + |(\tilde{a} \circ \tilde{b}) - (\tilde{a} \boxdot \tilde{b})|\,,$$

d. h. man hat

- im Unterlaufbereich

$$|(a \circ b) - (\tilde{a} \boxdot \tilde{b})| \leq \Delta_{dat}(\circ) + \Delta_{rnd,U}(\circ) =: \Delta(\circ)\,,$$

- im normalisierten Bereich

$$|(a \circ b) - (\tilde{a} \boxdot \tilde{b})| \leq \Delta_{dat}(\circ) + \Delta_{rnd,N}(\circ) =: \Delta(\circ)$$

- und im gesamten Bereich

$$\begin{aligned} |(a \circ b) - (\tilde{a} \boxdot \tilde{b})| &\leq \Delta_{dat}(\circ) + \max\{\Delta_{rnd,U}(\circ), \Delta_{rnd,N}(\circ)\} \\ &\leq \Delta_{dat}(\circ) + \Delta_{rnd,U}(\circ) + \Delta_{rnd,N}(\circ) =: \Delta(\circ)\,. \end{aligned}$$

3.2 Relative Fehlerschranken

Für das Ergebnis der Maschinenverknüpfung $\tilde{a} \boxdot \tilde{b}$ soll nun eine relative Fehlerschranke $\varepsilon(\circ) \in \mathbb{R}^+$ berechnet werden, so daß

$$\left| \frac{(a \circ b) - (\tilde{a} \boxdot \tilde{b})}{a \circ b} \right| \leq \varepsilon(\circ)$$

für alle $a \in A$ mit $|a - \tilde{a}| \leq \Delta(a)$ bzw. $|a - \tilde{a}| \leq \varepsilon(a)|a|$ und für alle $b \in B$ mit $|b - \tilde{b}| \leq \Delta(b)$ bzw. $|b - \tilde{b}| \leq \varepsilon(b)|b|$ gilt.

3.2.1 Abschätzung des fortgepflanzten Datenfehlers

Satz 3.3 *Für die Fortpflanzung des relativen Datenfehlers*

$$\left|\frac{(a \circ b) - (\tilde{a} \circ \tilde{b})}{a \circ b}\right|$$

gilt bei exakter Rechnung

$$\left|\frac{(a+b)-(\tilde{a}+\tilde{b})}{a+b}\right| \leq \frac{\Delta(a)+\Delta(b)}{|a+b|} =: \varepsilon_{dat,+} \tag{11}$$

$$\left|\frac{(a-b)-(\tilde{a}-\tilde{b})}{a-b}\right| \leq \frac{\Delta(a)+\Delta(b)}{|a-b|} =: \varepsilon_{dat,-} \tag{12}$$

$$\left|\frac{(a\cdot b)-(\tilde{a}\cdot\tilde{b})}{a\cdot b}\right| \leq \frac{\Delta(a)}{|a|}\cdot\frac{\Delta(b)}{|b|}+\frac{\Delta(b)}{|b|}+\frac{\Delta(a)}{|a|} =: \varepsilon_{dat,\cdot} \tag{13}$$

$$\left|\frac{(a/b)-(\tilde{a}/\tilde{b})}{a/b}\right| \leq \frac{|\frac{b}{a}|\Delta(a)+\Delta(b)}{|b|-\Delta(b)} =: \varepsilon_{dat,/} \quad \textit{für} \quad \Delta(b) < |b|\,. \tag{14}$$

Beweis:

$$\textit{Add.}: \quad \left|\frac{(a+b)-(\tilde{a}+\tilde{b})}{a+b}\right| \overset{(4)}{\leq} \frac{\Delta(a)+\Delta(b)}{|a+b|} = \varepsilon_{dat,+}$$

$$\textit{Sub.}: \quad \left|\frac{(a-b)-(\tilde{a}-\tilde{b})}{a-b}\right| \overset{(5)}{\leq} \frac{\Delta(a)+\Delta(b)}{|a-b|} = \varepsilon_{dat,-}$$

$$\textit{Mul.}: \quad \left|\frac{(a\cdot b)-(\tilde{a}\cdot\tilde{b})}{a\cdot b}\right| \overset{(6)}{\leq} \frac{\Delta(a)\Delta(b)+|a|\Delta(b)+|b|\Delta(a)}{|a\cdot b|}$$

$$= \frac{\Delta(a)}{|a|}\cdot\frac{\Delta(b)}{|b|}+\frac{\Delta(b)}{|b|}+\frac{\Delta(a)}{|a|} = \varepsilon_{dat,\cdot}$$

$$\textit{Div.}: \quad \left|\frac{(a/b)-(\tilde{a}/\tilde{b})}{a/b}\right| \overset{(7)}{\leq} \frac{\Delta(a)+|\frac{a}{b}|\Delta(b)}{|b|-\Delta(b)}\cdot\frac{|b|}{|a|}$$

$$= \frac{|\frac{b}{a}|\Delta(a)+\Delta(b)}{|b|-\Delta(b)} = \varepsilon_{dat,/}\,, \text{ falls } \Delta(b) < |b| \qquad \blacksquare$$

Mit Hilfe von Satz 3.3 lassen sich nun gleichmäßige Schranken für die Fortpflanzung des relativen Datenfehlers für alle $a \in A$ und $b \in B$ herleiten. Für die Addition und Subtraktion werden dabei die beiden folgenden Lemmata verwendet.

Lemma 3.1 *Seien $x \in X = [\underline{x}, \overline{x}] \in I\!I\!R$, $y \in Y = [\underline{y}, \overline{y}] \in I\!I\!R$ beliebig und $c_1, c_2 \geq 0$ konstant. Dann gilt*

$$\max_{x\in X, y\in Y} \frac{c_1|x| + c_2}{|x \pm y|} = \max_{x\in\{\underline{x},\overline{x}\}, y\in\{\underline{y},\overline{y}\}} \frac{c_1|x| + c_2}{|x \pm y|} = \max_{x\in\{\underline{x},\overline{x}\}} \frac{c_1|x| + c_2}{\langle x \pm Y\rangle}\,. \tag{15}$$

Dabei wird vorausgesetzt, daß die auftretenden Nenner stets ungleich Null sind, d. h. $0 \notin X \pm Y$.

Beweis: Der Beweis wird für den Fall „+“ mit $\inf(X+Y) > 0$ gezeigt. Die restlichen drei Fälle („+“ mit $\sup(X + Y) < 0$, „−“ mit $\inf(X - Y) > 0$ und „−“ mit $\sup(X - Y) < 0$) werden ganz entsprechend behandelt; es ändern sich lediglich einige Vorzeichen.
Sei $D := X \times Y$, $f : D \to I\!R, (x, y) \mapsto f(x, y) := \frac{c_1|x|+c_2}{|x+y|} = \frac{c_1|x|+c_2}{x+y}$ und $c_1 > 0$ ($c_1 = 0 \Rightarrow \max_{x\in X, y\in Y} \frac{c_2}{x+y} = \max_{x\in\{\underline{x},\overline{x}\}} \frac{c_2}{\langle x+Y\rangle}$).
Zu $y_0 \in Y$, beliebig aber fest gewählt, wird $g_{y_0}(x) := f(x, y_0)$ definiert.
Für $x < 0$ ist

$$g'_{y_0}(x) = \frac{-c_1(x + y_0) - (-c_1x + c_2)}{(x + y_0)^2} = \frac{-c_1y_0 - c_2}{(x + y_0)^2} < 0\,,$$

da $y_0 > -x > 0$.

Für $x > 0$ ist

$$g'_{y_0}(x) = \frac{c_1(x + y_0) - (c_1x + c_2)}{(x + y_0)^2} = \frac{c_1y_0 - c_2}{(x + y_0)^2} \begin{cases} < 0 & \text{für} \quad y_0 < c_2/c_1\,, \\ \geq 0 & \text{für} \quad y_0 \geq c_2/c_1\,. \end{cases}$$

Somit ist $g_{y_0}(x)$ für $x < 0$ eine monoton fallende und für $x \geq 0$, in Abhängigkeit von der fest vorgegebenen Lage von y_0, eine monoton fallende bzw. monoton wachsende Funktion. Es folgt daher wegen der Stetigkeit von g_{y_0} auf X

$$\max_{x\in X} f(x, y_0) = \max_{x\in X} g_{y_0}(x) = \max_{x\in\{\underline{x},\overline{x}\}} g_{y_0}(x) \tag{16}$$

und weiter

$$\max_{x\in X, y\in Y} \frac{c_1|x| + c_2}{x + y} = \max_{x\in X} \max_{y\in Y} \frac{c_1|x| + c_2}{x + y} = \max_{x\in X} \max_{y\in\{\underline{y},\overline{y}\}} \frac{c_1|x| + c_2}{x + y}$$
$$\overset{(16)}{=} \max_{x\in\{\underline{x},\overline{x}\}} \max_{y\in\{\underline{y},\overline{y}\}} \frac{c_1|x| + c_2}{x + y} = \max_{x\in\{\underline{x},\overline{x}\}} \frac{c_1|x| + c_2}{\langle x + Y\rangle} \qquad \blacksquare$$

Lemma 3.2 *Seien $x \in X = [\underline{x}, \overline{x}] \in I\!I\!R$, $y \in Y = [\underline{y}, \overline{y}] \in I\!I\!R$ beliebig und $d_1, d_2 \geq 0$ konstant. Dann gilt*

$$\max_{x\in X, y\in Y} \frac{d_1|x| + d_2|y|}{|x \pm y|} = \max_{x\in\{\underline{x},\overline{x}\}, y\in\{\underline{y},\overline{y}\}} \frac{d_1|x| + d_2|y|}{|x \pm y|}\,. \tag{17}$$

Dabei wird vorausgesetzt, daß die auftretenden Nenner stets ungleich Null sind, d. h. $0 \notin X \pm Y$.

Beweis: Wieder wird nur der Beweis für den Fall „+“ mit $\inf(X+Y) > 0$ gezeigt. Die anderen Fälle werden ganz entsprechend gezeigt; es ändern sich lediglich einige Vorzeichen.
Seien o. B. d. A. $d_1, d_2 > 0$ ($d_1 = 0$ oder $d_2 = 0$ führt auf Gleichung (15) mit $c_2 = 0$ und evtl. Rollentausch von x und y). Dann folgt die Behauptung aus Lemma 3.1:

$$\max_{x\in X, y\in Y} \frac{d_1|x| + d_2|y|}{x+y} = \max_{x\in X} \max_{y\in Y} \frac{d_1|x| + d_2|y|}{x+y} =^{3} \max_{x\in X} \max_{y\in\{\underline{y},\overline{y}\}} \frac{d_1|x| + d_2|y|}{x+y}$$

$$= \max_{y\in\{\underline{y},\overline{y}\}} \max_{x\in X} \frac{d_1|x| + d_2|y|}{x+y} =^{4} \max_{y\in\{\underline{y},\overline{y}\}} \max_{x\in\{\underline{x},\overline{x}\}} \frac{d_1|x| + d_2|y|}{x+y} \qquad \blacksquare$$

Unter Verwendung von Lemma 3.1 und Lemma 3.2 können jetzt optimale Schranken $\varepsilon_{dat}(+), \varepsilon_{dat}(-), \varepsilon_{dat}(\cdot)$ und $\varepsilon_{dat}(/)$ für die oben genannte Fehlerfortpflanzung angegeben werden.

Fall I Für $|a-\tilde{a}|$ und $|b-\tilde{b}|$ sind die absoluten Schranken $\Delta(a)$ und $\Delta(b)$ gegeben.

$$\begin{aligned}
\varepsilon_{dat,+} &= \frac{\Delta(a)+\Delta(b)}{|a+b|} \leq \frac{\Delta(a)+\Delta(b)}{\langle A+B\rangle} =: \varepsilon_{dat}(+) \\
\varepsilon_{dat,-} &= \frac{\Delta(a)+\Delta(b)}{|a-b|} \leq \frac{\Delta(a)+\Delta(b)}{\langle A-B\rangle} =: \varepsilon_{dat}(-) \\
\varepsilon_{dat,\cdot} &= \frac{\Delta(a)}{|a|}\cdot\frac{\Delta(b)}{|b|} + \frac{\Delta(b)}{|b|} + \frac{\Delta(a)}{|a|} \\
&\leq \frac{\Delta(a)}{\langle A\rangle}\cdot\frac{\Delta(b)}{\langle B\rangle} + \frac{\Delta(b)}{\langle B\rangle} + \frac{\Delta(a)}{\langle A\rangle} =: \varepsilon_{dat}(\cdot) \\
\varepsilon_{dat,/} &= \frac{|\frac{b}{a}|\Delta(a)+\Delta(b)}{|b|-\Delta(b)} = \frac{\frac{\Delta(a)}{|a|}+\frac{\Delta(b)}{|b|}}{1-\frac{\Delta(b)}{|b|}} \\
&\leq \frac{\frac{\Delta(a)}{\langle A\rangle}+\frac{\Delta(b)}{\langle B\rangle}}{1-\frac{\Delta(b)}{\langle B\rangle}} =: \varepsilon_{dat}(/) \quad \text{für} \quad \Delta(b) < \langle B\rangle
\end{aligned}$$

Fall II Für $|a-\tilde{a}|$ ist die absolute Schranke $\Delta(a)$ und für $|b-\tilde{b}|$ die relative Schranke $\varepsilon(b)$ gegeben.

$$\begin{aligned}
\varepsilon_{dat,+} &= \frac{\Delta(a)+|b|\varepsilon(b)}{|a+b|} \overset{\text{Lemma 3.1}}{\leq} \max_{b\in\{\underline{b},\overline{b}\}} \frac{\Delta(a)+|b|\varepsilon(b)}{\langle A+b\rangle} =: \varepsilon_{dat}(+) \\
\varepsilon_{dat,-} &= \frac{\Delta(a)+|b|\varepsilon(b)}{|a-b|} \overset{\text{Lemma 3.1}}{\leq} \max_{b\in\{\underline{b},\overline{b}\}} \frac{\Delta(a)+|b|\varepsilon(b)}{\langle A-b\rangle} =: \varepsilon_{dat}(-)
\end{aligned}$$

[3] Lemma 3.1 mit $c_1 = d_2$, $c_2 = d_1|x|$
[4] Lemma 3.1 mit $c_1 = d_1$, $c_2 = d_2|y|$

$$\begin{aligned}
\varepsilon_{dat,\cdot} &= \frac{\Delta(a)}{|a|}\cdot\varepsilon(b)+\varepsilon(b)+\frac{\Delta(a)}{|a|} \;\le\; \frac{\Delta(a)}{\langle A\rangle}\cdot\varepsilon(b)+\varepsilon(b)+\frac{\Delta(a)}{\langle A\rangle} \;=:\; \varepsilon_{dat}(\cdot)\\
\varepsilon_{dat,/} &= \frac{|\frac{b}{a}|\Delta(a)+|b|\varepsilon(b)}{|b|-|b|\varepsilon(b)} \;=\; \frac{\frac{\Delta(a)}{|a|}+\varepsilon(b)}{1-\varepsilon(b)}\\
&\le \frac{\frac{\Delta(a)}{\langle A\rangle}+\varepsilon(b)}{1-\varepsilon(b)} \;=:\; \varepsilon_{dat}(/) \quad \text{für} \quad \varepsilon(b)<1
\end{aligned}$$

Fall III Für $|a-\tilde{a}|$ ist die relative Schranke $\varepsilon(a)$ und für $|b-\tilde{b}|$ die absolute Schranke $\Delta(b)$ gegeben.

$$\begin{aligned}
\varepsilon_{dat,+} &= \frac{|a|\varepsilon(a)+\Delta(b)}{|a+b|} \overset{\text{Lemma 3.1}}{\le} \max_{a\in\{\underline{a},\overline{a}\}} \frac{|a|\varepsilon(a)+\Delta(b)}{\langle a+B\rangle} \;=:\; \varepsilon_{dat}(+)\\
\varepsilon_{dat,-} &= \frac{|a|\varepsilon(a)+\Delta(b)}{|a-b|} \overset{\text{Lemma 3.1}}{\le} \max_{a\in\{\underline{a},\overline{a}\}} \frac{|a|\varepsilon(a)+\Delta(b)}{\langle a-B\rangle} \;=:\; \varepsilon_{dat}(-)\\
\varepsilon_{dat,\cdot} &= \varepsilon(a)\cdot\frac{\Delta(b)}{|b|}+\frac{\Delta(b)}{|b|}+\varepsilon(a) \;\le\; \varepsilon(a)\cdot\frac{\Delta(b)}{\langle B\rangle}+\frac{\Delta(b)}{\langle B\rangle}+\varepsilon(a) \;=:\; \varepsilon_{dat}(\cdot)\\
\varepsilon_{dat,/} &= \frac{|\frac{b}{a}|\cdot|a|\varepsilon(a)+\Delta(b)}{|b|-\Delta(b)} \;=\; \frac{\varepsilon(a)+\frac{\Delta(b)}{|b|}}{1-\frac{\Delta(b)}{|b|}}\\
&\le \frac{\varepsilon(a)+\frac{\Delta(b)}{\langle B\rangle}}{1-\frac{\Delta(b)}{\langle B\rangle}} \;=:\; \varepsilon_{dat}(/) \quad \text{für} \quad \Delta(b)<\langle B\rangle
\end{aligned}$$

Fall IV Für $|a-\tilde{a}|$ und $|b-\tilde{b}|$ sind die relativen Schranken $\varepsilon(a)$ und $\varepsilon(b)$ gegeben.

$$\begin{aligned}
\varepsilon_{dat,+} &= \frac{|a|\varepsilon(a)+|b|\varepsilon(b)}{|a+b|} \overset{\text{Lemma 3.2}}{\le} \max_{a\in\{\underline{a},\overline{a}\},b\in\{\underline{b},\overline{b}\}} \frac{|a|\varepsilon(a)+|b|\varepsilon(b)}{|a+b|} \;=:\; \varepsilon_{dat}(+)\\
\varepsilon_{dat,-} &= \frac{|a|\varepsilon(a)+|b|\varepsilon(b)}{|a-b|} \overset{\text{Lemma 3.2}}{\le} \max_{a\in\{\underline{a},\overline{a}\},b\in\{\underline{b},\overline{b}\}} \frac{|a|\varepsilon(a)+|b|\varepsilon(b)}{|a-b|} \;=:\; \varepsilon_{dat}(-)\\
\varepsilon_{dat,\cdot} &= \varepsilon(a)\varepsilon(b)+\varepsilon(b)+\varepsilon(a) \;=:\; \varepsilon_{dat}(\cdot)\\
\varepsilon_{dat,/} &= \frac{|\frac{b}{a}|\cdot|a|\varepsilon(a)+|b|\varepsilon(b)}{|b|-|b|\varepsilon(b)} \;=\; \frac{\varepsilon(a)+\varepsilon(b)}{1-\varepsilon(b)} \;=:\; \varepsilon_{dat}(/) \quad \text{für} \quad \varepsilon(b)<1
\end{aligned}$$

In Tabelle 2 sind die Ergebnisse noch einmal zusammengefaßt.

3.2.2 Abschätzung des Rundungsfehlers

Über den Rundungsfehler, der bei einer gleitpunktmäßigen Verknüpfung zweier Gleitpunktzahlen entsteht, macht der folgende Satz Aussagen.

gegeben	$\varepsilon_{dat}(\pm)$	$\varepsilon_{dat}(\cdot)$	$\varepsilon_{dat}(/)$
$\Delta(a), \Delta(b)$	$\frac{\Delta(a)+\Delta(b)}{\langle A \pm B\rangle}$	$\frac{\Delta(a)}{\langle A\rangle}\cdot\frac{\Delta(b)}{\langle B\rangle}+\frac{\Delta(b)}{\langle B\rangle}+\frac{\Delta(a)}{\langle A\rangle}$	$\frac{\frac{\Delta(a)}{\langle A\rangle}+\frac{\Delta(b)}{\langle B\rangle}}{1-\frac{\Delta(b)}{\langle B\rangle}}$
$\Delta(a), \varepsilon(b)$	$\max\limits_{b\in\{\underline{b},\overline{b}\}} \frac{\Delta(a)+\lvert b\rvert\varepsilon(b)}{\langle A\pm b\rangle}$	$\frac{\Delta(a)}{\langle A\rangle}\cdot\varepsilon(b)+\varepsilon(b)+\frac{\Delta(a)}{\langle A\rangle}$	$\frac{\frac{\Delta(a)}{\langle A\rangle}+\varepsilon(b)}{1-\varepsilon(b)}$
$\varepsilon(a), \Delta(b)$	$\max\limits_{a\in\{\underline{a},\overline{a}\}} \frac{\lvert a\rvert\varepsilon(a)+\Delta(b)}{\langle a\pm B\rangle}$	$\varepsilon(a)\cdot\frac{\Delta(b)}{\langle B\rangle}+\frac{\Delta(b)}{\langle B\rangle}+\varepsilon(a)$	$\frac{\varepsilon(a)+\frac{\Delta(b)}{\langle B\rangle}}{1-\frac{\Delta(b)}{\langle B\rangle}}$
$\varepsilon(a), \varepsilon(b)$	$\max\limits_{a\in\{\underline{a},\overline{a}\},b\in\{\underline{b},\overline{b}\}} \frac{\lvert a\rvert\varepsilon(a)+\lvert b\rvert\varepsilon(b)}{\lvert a\pm b\rvert}$	$\varepsilon(a)\varepsilon(b)+\varepsilon(b)+\varepsilon(a)$	$\frac{\varepsilon(a)+\varepsilon(b)}{1-\varepsilon(b)}$

Tabelle 2: Schranken für die Fortpflanzung des relativen Datenfehlers bei den Grundoperationen

Satz 3.4 *Für den relativen Rundungsfehler*

$$\left|\frac{(\tilde{a}\circ\tilde{b})-(\tilde{a}\boxed{\circ}\tilde{b})}{a\circ b}\right|$$

einer Operation $\circ\in\{+,-,\cdot,/\}$ *mit den Maschinenzahlen* $\tilde{a},\tilde{b}\in S$ *gilt*

a) im Unterlaufbereich, d. h. $\tilde{a}\boxed{\circ}\tilde{b}\in U$ *und damit* $\tilde{a}\circ\tilde{b}\in U$*:*

$$\left|\frac{(\tilde{a}\circ\tilde{b})-(\tilde{a}\boxed{\circ}\tilde{b})}{a\circ b}\right| \;\le\; \frac{\texttt{MinReal}}{\langle A\circ B\rangle} \;=:\; \varepsilon_{rnd,U}(\circ)\,. \tag{18}$$

Bei Verwendung einer Arithmetik mit maximal genauen Operationen im „gradual underflow“ kann die Abschätzung noch verschärft werden zu:

$$\left|\frac{(\tilde{a}\circ\tilde{b})-(\tilde{a}\boxed{\circ}\tilde{b})}{a\circ b}\right| \;\le\; \left\{\begin{array}{ll} 0 & \text{für } \circ\in\{+,-\} \\ \dfrac{\texttt{dMinReal}}{\langle A\circ B\rangle} & \text{für } \circ\in\{\cdot,/\} \end{array}\right\} \;=:\; \varepsilon_{rnd,U}(\circ) \tag{19}$$

b) im normalisierten Bereich, d. h. $\tilde{a}\boxed{\circ}\tilde{b}\notin U$ *und damit* $\tilde{a}\circ\tilde{b}\notin U$*:*

$$\left|\frac{(\tilde{a}\circ\tilde{b})-(\tilde{a}\boxed{\circ}\tilde{b})}{a\circ b}\right| \;\le\; \overline{\varepsilon}\cdot(\varepsilon_{dat}(\circ)+1) \;=:\; \varepsilon_{rnd,N}(\circ) \tag{20}$$

für jedes $\circ\in\{+,-,\cdot,/\}$*.*

Beweis:

a) Folgt sofort aus Satz 3.2a).

$$\text{b)} \quad \left|\frac{(\tilde{a} \circ \tilde{b}) - (\tilde{a} \boxed{\circ} \tilde{b})}{a \circ b}\right| \overset{(1)}{\leq} \overline{\varepsilon} \cdot \left|\frac{\tilde{a} \circ \tilde{b}}{a \circ b}\right|$$

$$\leq \overline{\varepsilon} \cdot \frac{|(a \circ b) - (\tilde{a} \circ \tilde{b})| + |a \circ b|}{|a \circ b|}$$

$$\leq \overline{\varepsilon} \cdot (\varepsilon_{dat}(\circ) + 1) = \varepsilon_{rnd,N}(\circ) \quad \blacksquare$$

3.2.3 Gesamtfehlerabschätzung

Mit der Dreiecksungleichung läßt sich der relative Gesamtfehler folgendermaßen abschätzen:

$$\left|\frac{(a \circ b) - (\tilde{a} \boxed{\circ} \tilde{b})}{a \circ b}\right| \leq \left|\frac{(a \circ b) - (\tilde{a} \circ \tilde{b})}{a \circ b}\right| + \left|\frac{(\tilde{a} \circ \tilde{b}) - (\tilde{a} \boxed{\circ} \tilde{b})}{a \circ b}\right| ,$$

d. h. man hat

- im Unterlaufbereich

$$\left|\frac{(a \circ b) - (\tilde{a} \boxed{\circ} \tilde{b})}{a \circ b}\right| \leq \varepsilon_{dat}(\circ) + \varepsilon_{rnd,U}(\circ) =: \varepsilon(\circ) ,$$

- im normalisierten Bereich

$$\left|\frac{(a \circ b) - (\tilde{a} \boxed{\circ} \tilde{b})}{a \circ b}\right| \leq \varepsilon_{dat}(\circ) + \varepsilon_{rnd,N}(\circ) =: \varepsilon(\circ)$$

- und im gesamten Bereich

$$\left|\frac{(a \circ b) - (\tilde{a} \boxed{\circ} \tilde{b})}{a \circ b}\right| \leq \varepsilon_{dat}(\circ) + \max\{\varepsilon_{rnd,U}(\circ), \varepsilon_{rnd,N}(\circ)\}$$

$$\leq \varepsilon_{dat}(\circ) + \varepsilon_{rnd,U}(\circ) + \varepsilon_{rnd,N}(\circ) =: \varepsilon(\circ) .$$

3.3 Sonderfälle: Operanden, die keine Rundungsfehler verursachen

In diesem Abschnitt kommt Forderung (3) aus Abschitt 2 zum Einsatz.

Falls eine Operation auf dem Rechner ohne Rundungsfehler durchführbar ist, so sollte für die mit dem Fehlerkalkül bestimmten Schranken $\Delta_{rnd}(\circ) = \varepsilon_{rnd}(\circ) = 0$ gelten. Die in den vorangegangenen Abschnitten hergeleiteten Rundungsfehlerschranken sind in diesen Fällen in der Regel jedoch größer als Null:

Beispiel 3.2 Seien $a = \underline{a} = \overline{a} = 10^{15}$ und $b = \underline{b} = \overline{b} = 10^{14}$ (A und B sind demnach Punktintervalle). Beide Zahlen sind im IEEE-`double`-Format exakt darstellbar, woraus folgt, daß $\Delta(a) = \varepsilon(a) = 0$ und somit $\Delta_{dat}(+) = \varepsilon_{dat}(+) = 0$. Man erhält schließlich nach (10) und (20)

$$\Delta_{rnd,N}(+) = \overline{\varepsilon} \cdot |10^{15} + 10^{14}| = 1.1 \cdot 10^{15} \cdot \overline{\varepsilon} > 0$$

und

$$\varepsilon_{rnd,N}(+) = \overline{\varepsilon} > 0\,,$$

obwohl das Ergebnis der Addition der beiden Zahlen, $1.1 \cdot 10^{15} \in I\!N \cap S$, auf dem Rechner exakt darstellbar ist.

Im folgenden werden hinreichende Bedingungen für rundungsfehlerfrei durchführbare (Maschinen-)Operationen genannt. Wenn nicht anders angegeben, wird dabei vorausgesetzt, daß das Ergebnis der Verknüpfung der Maschinenzahlen $\tilde{a}$ und $\tilde{b}$ im normalisierten Bereich liegt, d. h. $|\tilde{a} \circ \tilde{b}| \in [\texttt{MinReal}, \texttt{MaxReal}]$ für $\tilde{a}, \tilde{b} \in S$.
Für $x = (-1)^{s(x)} \cdot m(x) \cdot 2^{e(x)} \in S$ liefern die Funktionen z_{lead} und z_{trail} die Anzahl der führenden bzw. abschließenden Nullen in der Mantisse $m(x)$ von x ($z_{lead}(x) = 0$ für $x \notin U = (-\texttt{MinReal}, \texttt{MinReal})$). $e(x)$ bezeichne den Exponenten der IEEE-Darstellung von x.

Satz 3.5 (Subtraktion) *Das Ergebnis einer Subtraktion von zwei Maschinenzahlen ist exakt darstellbar, d. h.*

$$|(\tilde{a} - \tilde{b}) - (\tilde{a} \boxminus \tilde{b})| = \left| \frac{(\tilde{a} - \tilde{b}) - (\tilde{a} \boxminus \tilde{b})}{a - b} \right| = 0\,,$$

wenn eine der folgenden Bedingungen erfüllt ist:

$$\textit{(i)} \quad \frac{1}{2} \;\leq\; \frac{\tilde{a}}{\tilde{b}} \;\leq\; 2 \tag{21}$$

$$\textit{(ii)} \quad e(\tilde{a} - \tilde{b}) \;\leq\; \min\{\, e(\tilde{a}) + z_{trail}(\tilde{a}),\; e(\tilde{b}) + z_{trail}(\tilde{b}) \,\} \tag{22}$$

$$\textit{(iii)} \quad \tilde{a} = 0 \quad \textit{oder} \quad \tilde{b} = 0 \tag{23}$$

Beweis: Zu (i): siehe [26]. Zu (ii): siehe [8]. Zu (iii): trivial ∎

Satz 3.6 (Addition) *Das Ergebnis einer Addition von zwei Maschinenzahlen ist exakt darstellbar, d. h.*

$$|(\tilde{a} + \tilde{b}) - (\tilde{a} \boxplus \tilde{b})| = \left| \frac{(\tilde{a} + \tilde{b}) - (\tilde{a} \boxplus \tilde{b})}{a + b} \right| = 0\,,$$

wenn eine der folgenden Bedingungen erfüllt ist:

$$\textit{(i)} \quad -2 \;\leq\; \frac{\tilde{a}}{\tilde{b}} \;\leq\; -\frac{1}{2} \tag{24}$$

$$\textit{(ii)} \quad e(\tilde{a} + \tilde{b}) \;\leq\; \min\{\, e(\tilde{a}) + z_{trail}(\tilde{a}),\; e(\tilde{b}) + z_{trail}(\tilde{b}) \,\} \tag{25}$$

$$\textit{(iii)} \quad \tilde{a} = 0 \quad \textit{oder} \quad \tilde{b} = 0 \tag{26}$$

Beweis: Die Behauptung folgt wegen $\tilde{a} + \tilde{b} = \tilde{a} - (-\tilde{b})$ aus Satz 3.5 ■

Satz 3.7 (Multiplikation) *Das Ergebnis einer Multiplikation von zwei Maschinenzahlen ist exakt darstellbar, d. h.*

$$|(\tilde{a} \cdot \tilde{b}) - (\tilde{a} \boxdot \tilde{b})| = \left| \frac{(\tilde{a} \cdot \tilde{b}) - (\tilde{a} \boxdot \tilde{b})}{a \cdot b} \right| = 0\,,$$

wenn eine der folgenden Bedingungen erfüllt ist:

$$\textit{(i)} \quad \tilde{a} = 2^k,\ k \in \mathbb{Z}. \qquad (27)$$
Ist $\tilde{a} \cdot \tilde{b} \in U$, *muß* $k \geq e_{min} - e(\tilde{b}) - z_{trail}(\tilde{b})$ *sein.*

$$\textit{(ii)} \quad \tilde{b} = 2^k,\ k \in \mathbb{Z}. \qquad (28)$$
Ist $\tilde{a} \cdot \tilde{b} \in U$, *muß* $k \geq e_{min} - e(\tilde{a}) - z_{trail}(\tilde{a})$ *sein.*

$$\textit{(iii)} \quad z_{lead}(\tilde{a}) + z_{trail}(\tilde{a}) + z_{lead}(\tilde{b}) + z_{trail}(\tilde{b}) \geq t \qquad (29)$$

Beweis:

(i) Seien $\tilde{a} = 2^k \in S, k \in \mathbb{Z}$ und $\tilde{b} \in S$ mit der Darstellung $\tilde{b} = (-1)^{s(\tilde{b})} \cdot m(\tilde{b}) \cdot 2^{e(\tilde{b})}$. Dann ist $\tilde{a} \cdot \tilde{b} = 2^k \cdot (-1)^{s(\tilde{b})} \cdot m(\tilde{b}) \cdot 2^{e(\tilde{b})} = (-1)^{s(\tilde{b})} \cdot m(\tilde{b}) \cdot 2^{k+e(\tilde{b})} \in S$ und daher $\tilde{a} \boxdot \tilde{b} = \tilde{a} \cdot \tilde{b}$.
Für $\tilde{a} \cdot \tilde{b} \in U$ und $k \geq e_{min} - e(\tilde{b}) - z_{trail}(\tilde{b})$ ist $e(\tilde{a} \boxdot \tilde{b}) = e_{min}$, d. h. die Mantisse von $\tilde{b}$ muß bei der Multiplikation mit 2^k um $e_{min} - (k + e(\tilde{b}))$ Stellen nach rechts geschoben werden. Wegen $e_{min} - (k + e(\tilde{b})) \leq e_{min} - (e_{min} - e(\tilde{b}) - z_{trail}(\tilde{b}) + e(\tilde{b})) = z_{trail}(\tilde{b})$ fallen nur Nullen aus der Mantisse heraus, es gilt also auch hier $\tilde{a} \boxdot \tilde{b} = \tilde{a} \cdot \tilde{b}$.

(ii) Analog zu (i).

(iii) $\tilde{a}$ möge $t - (z_{lead}(\tilde{a}) + z_{trail}(\tilde{a}))$ und $\tilde{b}$ möge $t - (z_{lead}(\tilde{b}) + z_{trail}(\tilde{b}))$ signifikante Stellen in der Mantisse besitzen. Dann hat $\tilde{a} \cdot \tilde{b}$ höchstens $t - (z_{lead}(\tilde{a}) + z_{trail}(\tilde{a})) + t - (z_{lead}(\tilde{b}) + z_{trail}(\tilde{b})) = 2t - (z_{lead}(\tilde{a}) + z_{trail}(\tilde{a}) + z_{lead}(\tilde{b}) + z_{trail}(\tilde{b})) \leq 2t - t = t$ signifikante Stellen und damit ist $\tilde{a} \cdot \tilde{b} \in S$ ■

Satz 3.8 (Division) *Das Ergebnis einer Division von zwei Maschinenzahlen ist exakt darstellbar, d. h.*

$$|(\tilde{a}/\tilde{b}) - (\tilde{a} \boxplus \tilde{b})| = \left| \frac{(\tilde{a}/\tilde{b}) - (\tilde{a} \boxplus \tilde{b})}{a/b} \right| = 0\,,$$

wenn die folgende Bedingung erfüllt ist:

$$\tilde{b} = 2^k,\ k \in \mathbb{Z} \qquad (30)$$

Ist $\tilde{a} \cdot \tilde{b} \in U$, *muß zusätzlich* $k \leq e(\tilde{a}) + z_{trail}(\tilde{a}) - e_{min}$ *gefordert werden.*

Beweis: Die Behauptung folgt wegen $\tilde{a}/2^k = \tilde{a} \cdot 2^{-k}$ aus Satz 3.7 ∎

Für die spätere Implementierung werden Bedingungen für die einschließenden Intervalle $\tilde{A}$ und $\tilde{B}$ benötigt, die gewährleisten, daß sämtliche Operationen $\tilde{a} \circ \tilde{b}$ mit $\tilde{a} \in \tilde{A} \cap S$ und $\tilde{b} \in \tilde{B} \cap S$ rundungsfehlerfrei durchgeführt werden können, d. h. , daß

$$\tilde{a} \circ \tilde{b} = \tilde{a} \boxed{\circ} \tilde{b} \quad \forall \tilde{a} \in \tilde{A} \cap S \;\; \forall \tilde{b} \in \tilde{B} \cap S$$

gleichmäßig gilt. Solche Bedingungen werden in den folgenden Sätzen angegeben. Dabei sei die Funktion $e_{IS} : IS \mapsto I\!N$ gemäß

$$e_{IS}(X) := \begin{cases} e(\langle X \rangle), & \text{falls } X \text{ echtes Intervall} \\ e(x) + z_{trail}(x), & \text{falls } X = [x,x] \text{ Punktintervall} \end{cases}$$

definiert.

Satz 3.9 (Subtraktion) *Ist eine der Bedingungen*

$$\textit{(i)} \quad \inf(\tilde{A}/\tilde{B}) \geq \frac{1}{2} \quad \textit{und} \quad \sup(\tilde{A}/\tilde{B}) \leq 2 \tag{31}$$

$$\textit{(ii)} \quad e(|\tilde{A} - \tilde{B}|) \leq \min\{e_{IS}(\tilde{A}), e_{IS}(\tilde{B})\} \tag{32}$$

$$\textit{(iii)} \quad \tilde{A} = [0,0] \quad \textit{oder} \quad \tilde{B} = [0,0] \tag{33}$$

erfüllt, dann gilt

$$\tilde{a} - \tilde{b} = \tilde{a} \boxed{-} \tilde{b} \quad \forall \tilde{a} \in \tilde{A} \cap S \;\; \forall \tilde{b} \in \tilde{B} \cap S\,.$$

Beweis:

(i) $\inf(\tilde{A}/\tilde{B}) \geq \frac{1}{2} \;\wedge\; \sup(\tilde{A}/\tilde{B}) \leq 2$
$\Longrightarrow \; \frac{1}{2} \leq \frac{\tilde{a}}{\tilde{b}} \leq 2 \quad \forall \tilde{a} \in \tilde{A} \cap S \;\; \forall \tilde{b} \in \tilde{B} \cap S \overset{\text{Satz 3.5 (i)}}{\Longrightarrow}$ Beh.

(ii) $e(\tilde{a} - \tilde{b}) \leq e(|\tilde{A} - \tilde{B}|) \leq \min\{e_{IS}(\tilde{A}), e_{IS}(\tilde{B})\}$
$\leq \min\{\; e(\tilde{a}) + z_{trail}(\tilde{a}),\; e(\tilde{b}) + z_{trail}(\tilde{b}) \;\} \quad \forall \tilde{a} \in \tilde{A} \cap S \;\; \forall \tilde{b} \in \tilde{B} \cap S$
$\overset{\text{Satz 3.5 (ii)}}{\Longrightarrow}$ Beh.

(iii) Folgt aus Satz 3.5 (iii) ∎

Satz 3.10 (Addition) *Ist eine der Bedingungen*

$$\textit{(i)} \quad \inf(\tilde{A}/\tilde{B}) \geq -2 \quad \textit{und} \quad \sup(\tilde{A}/\tilde{B}) \leq -\frac{1}{2} \tag{34}$$

$$\textit{(ii)} \quad e(|\tilde{A} + \tilde{B}|) \leq \min\{e_{IS}(\tilde{A}), e_{IS}(\tilde{B})\} \tag{35}$$

$$\textit{(iii)} \quad \tilde{A} = [0,0] \quad \textit{oder} \quad \tilde{B} = [0,0] \tag{36}$$

erfüllt, dann gilt

$$\tilde{a} + \tilde{b} = \tilde{a} \boxed{+} \tilde{b} \quad \forall \tilde{a} \in \tilde{A} \cap S \;\; \forall \tilde{b} \in \tilde{B} \cap S\,.$$

Beweis:

(i) $\inf(\tilde{A}/\tilde{B}) \geq -2 \ \wedge \ \sup(\tilde{A}/\tilde{B}) \leq -\frac{1}{2}$
$\Longrightarrow \ -2 \leq \frac{\tilde{a}}{\tilde{b}} \leq -\frac{1}{2} \quad \forall \tilde{a} \in \tilde{A} \cap S \ \ \forall \tilde{b} \in \tilde{B} \cap S \overset{\text{Satz 3.6 (i)}}{\Longrightarrow}$ Beh.

(ii) $e(\tilde{a} + \tilde{b}) \leq e(|\tilde{A} + \tilde{B}|) \leq \min\{e_{IS}(\tilde{A}), e_{IS}(\tilde{B})\}$
$\leq \min\{\ e(\tilde{a}) + z_{trail}(\tilde{a}),\ e(\tilde{b}) + z_{trail}(\tilde{b})\ \} \quad \forall \tilde{a} \in \tilde{A} \cap S \ \ \forall \tilde{b} \in \tilde{B} \cap S$
$\overset{\text{Satz 3.6 (ii)}}{\Longrightarrow}$ Beh.

(iii) Folgt aus Satz 3.6 (iii) ■

Satz 3.11 (Multiplikation) *Ist eine der Bedingungen*

$$\textit{(i)} \quad \tilde{A} = [\tilde{a}, \tilde{a}],\ \tilde{a} = 2^k,\ k \in \mathbb{Z}; \qquad (37)$$
$$\textit{ist } (\tilde{A} \cdot \tilde{B}) \cap U \neq \emptyset,\ \textit{muß } k \geq e_{min} - e_{IS}(\tilde{B}) \textit{ sein}$$

$$\textit{(ii)} \quad \tilde{B} = [\tilde{b}, \tilde{b}],\ \tilde{b} = 2^k,\ k \in \mathbb{Z}; \qquad (38)$$
$$\textit{ist } (\tilde{A} \cdot \tilde{B}) \cap U \neq \emptyset,\ \textit{muß } k \geq e_{min} - e_{IS}(\tilde{A}) \textit{ sein}$$

$$\textit{(iii)} \quad \tilde{A} = [\tilde{a}, \tilde{a}],\ \tilde{B} = [\tilde{b}, \tilde{b}],\ z_{lead}(\tilde{a}) + z_{trail}(\tilde{a}) + z_{lead}(\tilde{b}) + z_{trail}(\tilde{b}) \geq t \qquad (39)$$

erfüllt, dann gilt

$$\tilde{a} \cdot \tilde{b} = \tilde{a} \boxdot \tilde{b} \quad \forall \tilde{a} \in \tilde{A} \cap S \ \ \forall \tilde{b} \in \tilde{B} \cap S\,.$$

Beweis: Die Behauptung folgt aus Satz 3.7 (i) bis (iii) ■

Satz 3.12 (Division) *Es gilt*

$$\tilde{a}/\tilde{b} = \tilde{a} \boxplus \tilde{b} \quad \forall \tilde{a} \in \tilde{A} \cap S \ \ \forall \tilde{b} \in \tilde{B} \cap S\,,$$

falls die folgende Bedingung erfüllt ist:

$$\tilde{B} = [\tilde{b}, \tilde{b}],\ \tilde{b} = 2^k,\ k \in \mathbb{Z}. \qquad (40)$$

Ist $(\tilde{A} \cdot \tilde{B}) \cap U \neq \emptyset$, muß zusätzlich $k \leq e_{IS}(\tilde{A}) - e_{min}$ gefordert werden.

Beweis: Die Behauptung folgt aus Satz 3.8 ■

Bemerkung 3.1 Die Aussagen über den Rundungsfehler im Unterlaufbereich sind natürlich nur zulässig, wenn die zugrundeliegende Arithmetik den „gradual underflow“ auch unterstützt!

Bemerkung 3.2 Ob für $\tilde{A} = [\tilde{a}, \tilde{a}]$, $\tilde{B} = [\tilde{b}, \tilde{b}]$ die Operation $\tilde{a} \circ \tilde{b}$ rundungsfehlerfrei durchführbar ist, läßt sich im Falle einer maximal genauen Maschinen-Intervallarithmetik auch mit der entsprechenden Maschinen-Intervalloperation prüfen. Bezeichnet $\diamondsuit$ eine Maschinen-Intervalloperation, so gilt

$$d(\tilde{A} \diamondsuit \tilde{B}) = 0 \implies \tilde{a} \circ \tilde{b} = \tilde{a} \boxdot \tilde{b}. \qquad (41)$$

Wegen (41) $\Rightarrow$ (39), kann in diesem Fall (39) übergangen werden.

4 Fehlerschranken für Funktionen

4.1 Allgemeine Fehlerabschätzung

Satz 4.1 *Sei $f : D \to \mathbb{R}$ differenzierbar auf dem Intervall*

$$D \supseteq \tilde{A} = A + [-\Delta(a), \Delta(a)] \subset \mathbb{R}\,.$$

Für die auf der Maschine implementierte Näherungsfunktion $\tilde{f} : D \cap S \to S$ an f sei $\varepsilon(f) \geq 0$ eine (bekannte) obere Schranke für den relativen Fehler im normalisierten Bereich und $\Delta(f) \geq 0$ eine (bekannte) obere Schranke für den absoluten Fehler im Unterlaufbereich, d. h. es gelte

$$|f(\tilde{x}) - \tilde{f}(\tilde{x})| \leq \varepsilon(f)|f(\tilde{x})| \tag{42}$$

für alle $\tilde{x} \in S$ mit $|f(\tilde{x})| \in [\texttt{MinReal}, \texttt{MaxReal}]$ und

$$|f(\tilde{x}) - \tilde{f}(\tilde{x})| \leq \Delta(f) \tag{43}$$

für alle $\tilde{x} \in S$ mit $|f(\tilde{x})| \in [0, \texttt{MinReal})$. Mit diesen Voraussetzungen erhält man die Fehlerabschätzung

$$|f(a) - \tilde{f}(\tilde{a})| \leq \Delta_{dat,f} + \varepsilon(f)\Big(\Delta_{dat,f} + |f(a)|\Big) + \Delta(f) \;\; \forall a \in A\,, \tag{44}$$

wobei

$$\begin{aligned} \Delta_{dat,f} &= \Delta(a) \cdot \Big|f'\Big(a + [-\Delta(a), \Delta(a)]\Big)\Big| \\ &= |a|\varepsilon(a) \cdot \Big|f'\Big(a \cdot [1 - \varepsilon(a), 1 + \varepsilon(a)]\Big)\Big|\,. \end{aligned} \tag{45}$$

Beweis: Seien $a \in A$ und $\tilde{a} \in \tilde{A} \cap S$ beliebig und $\Delta_a = \tilde{a} - a \in [-\Delta(a), \Delta(a)]$.

1. **Fortgepflanzter Datenfehler:**
   Aus dem Mittelwertsatz der Differentialrechnung folgt

   $$f(\tilde{a}) = f(a + \Delta_a) = f(a) + f'(a + \theta\Delta_a)\Delta_a \quad \text{mit} \quad \theta \in (0\,,1)\,,$$

   d. h. für die Fortpflanzung des Datenfehlers Δ_a gilt bei exakter Rechnung

   $$\begin{aligned} |f(a) - f(\tilde{a})| &= |\Delta_a f'(a + \theta\Delta_a)| \\ &\leq \Delta(a) \cdot \Big|f'\Big(a + [-\Delta(a), \Delta(a)]\Big)\Big| \\ &= |a|\varepsilon(a) \cdot \Big|f'\Big(a \cdot [1 - \varepsilon(a), 1 + \varepsilon(a)]\Big)\Big| = \Delta_{dat,f}\,. \end{aligned}$$

2. **Rundungsfehler:**
Nach (42) und (43) ist

$$|f(\tilde{a}) - \tilde{f}(\tilde{a})| \leq \begin{cases} \Delta(f)\,, & \text{falls } |f(\tilde{a})| \in [0, \texttt{MinReal})\,, \\ \varepsilon(f)|f(\tilde{a})|\,, & \text{falls } |f(\tilde{a})| \in [\texttt{MinReal}, \texttt{MaxReal}]\,. \end{cases}$$

Man hat also

$$\begin{aligned} |f(\tilde{a}) - \tilde{f}(\tilde{a})| &\leq \varepsilon(f)|f(\tilde{a})| + \Delta(f) \\ &\leq \varepsilon(f)\Big(|f(\tilde{a}) - f(a)| + |f(a)|\Big) + \Delta(f) \\ &= \varepsilon(f)\Big(\Delta_{dat,f} + |f(a)|\Big) + \Delta(f)\,. \end{aligned}$$

3. **Gesamtfehler:**
Da $a \in A$ und $\tilde{a} \in \tilde{A} \cap S$ beliebig waren, folgt die Behauptung jetzt aus der Dreiecksungleichung, angewandt auf den Gesamtfehler:

$$\begin{aligned} |f(a) - \tilde{f}(\tilde{a})| &\leq |f(a) - f(\tilde{a})| + |f(\tilde{a}) - \tilde{f}(\tilde{a})| \\ &\leq \Delta_{dat,f} + \varepsilon(f)\Big(\Delta_{dat,f} + |f(a)|\Big) + \Delta(f) \;\; \forall a \in A \end{aligned}$$ ∎

Korollar 4.1 *Für den relativen Gesamtfehler gilt nach Satz 4.1*

$$\left|\frac{f(a) - \tilde{f}(\tilde{a})}{f(a)}\right| \leq \varepsilon_{dat,f} + \varepsilon(f)\Big(\varepsilon_{dat,f} + 1\Big) + \frac{\Delta(f)}{|f(a)|} \;\; \forall a \in A\,, \tag{46}$$

wobei

$$\begin{aligned} \varepsilon_{dat,f} &= \Delta(a) \cdot \left|\frac{f'\Big(a + [-\Delta(a), \Delta(a)]\Big)}{f(a)}\right| \\ &= |a|\varepsilon(a) \cdot \left|\frac{f'\Big(a \cdot [1 - \varepsilon(a), 1 + \varepsilon(a)]\Big)}{f(a)}\right|\,. \end{aligned} \tag{47}$$

Korollar 4.2 *Gleichmäßige Fehlerabschätzungen über dem Intervall A erhält man, indem man (44) und (46) intervallmäßig für A auswertet, also*

$$|f(a) - \tilde{f}(\tilde{a})| \leq \Delta_{dat}(f) + \varepsilon(f)\Big(\Delta_{dat}(f) + |f(A)|\Big) + \Delta(f) \;\; \forall a \in A \tag{48}$$

$$\left|\frac{f(a) - \tilde{f}(\tilde{a})}{f(a)}\right| \leq \varepsilon_{dat}(f) + \varepsilon(f)\Big(\varepsilon_{dat}(f) + 1\Big) + \frac{\Delta(f)}{\langle f(A)\rangle} \;\; \forall a \in A\,, \tag{49}$$

wobei

$$\begin{aligned} \Delta_{dat}(f) &= \Delta(a) \cdot \left| f'\Big(A + [-\Delta(a), \Delta(a)]\Big)\right| \\ &= |A|\varepsilon(a) \cdot \left| f'\Big(A \cdot [1-\varepsilon(a), 1+\varepsilon(a)]\Big)\right| \end{aligned} \tag{50}$$

und

$$\begin{aligned} \varepsilon_{dat}(f) &= \Delta(a) \cdot \frac{\left| f'\Big(A + [-\Delta(a), \Delta(a)]\Big)\right|}{\langle f(A)\rangle} \\ &= |A|\varepsilon(a) \cdot \frac{\left| f'\Big(A \cdot [1-\varepsilon(a), 1+\varepsilon(a)]\Big)\right|}{\langle f(A)\rangle}\,. \end{aligned} \tag{51}$$

Bemerkung 4.1 Manchmal wird von der Approximation $\tilde{f}$ verlangt, daß sie dieselben Nullstellen wie f hat, also $f(\tilde{x}) = 0 \implies \tilde{f}(\tilde{x}) = 0$. Die oben aufgeführten Fehlerschranken behalten jedoch ihre Gültigkeit auch dann, wenn diese Eigenschaft nicht erfüllt ist.

Bemerkung 4.2 Die Ableitung von f in (50) bzw. (51) kann mit Hilfe von automatischer Differentiation auf einfache Weise berechnet werden. Treten in der intervallmäßigen Auswertung von (48) bzw. (49) zu große Überschätzungen auf, können die Fehlerschranken durch direkte Anwendung von Satz 4.1 und Korollar 4.1 eventuell verschärft werden[5].

Als Beispiel für die Anwendung von Satz 4.1 und der Korollare 4.1 und 4.2 sollen im folgenden Unterabschnitt Fehlerabschätzungen für einige mathematische Standardfunktionen hergeleitet werden.

4.2 Konkrete Fehlerschranken für einige mathematische Standardfunktionen

4.2.1 Die Funktion $\mathrm{sqrt}(x) = \sqrt{x}$

Lemma 4.1 *Für $c_1, c_2 \geq 0$ nimmt die Funktion $f : X \to I\!R$, $X = [\underline{x}, \overline{x}] \in I\!I\!R$, $\underline{x} > c_1$, die durch*

$$f(x) := \frac{c_1}{2\sqrt{x - c_1}} + c_2 \left(\frac{c_1}{2\sqrt{x - c_1}} + \sqrt{x} \right)$$

[5]Gelegentlich kann gezeigt werden, daß die Fehlerschranken ihr Maximum am Intervallrand annehmen.

definiert sei, ihr Maximum am Rand von X an, d. h.

$$\max_{x\in X} f(x) = \max_{x\in\{\underline{x},\overline{x}\}} f(x)\,.$$

Die Funktion $f(x) = \mathrm{sqrt}(x) = \sqrt{x}$ ist differenzierbar auf dem Intervall $\tilde{A} = A + [-\Delta(a), \Delta(a)] = A \cdot [1-\varepsilon(a), 1+\varepsilon(a)]$, wenn $\inf(\tilde{A}) > 0$ ist. Für das Maschinenanalogon $\tilde{f}$ zu f verlangt der IEEE-754-Standard

$$|f(\tilde{x}) - \tilde{f}(\tilde{x})| \le \overline{\varepsilon}|f(\tilde{x})|$$

für alle $\tilde{x} \in S$ mit $|f(\tilde{x})| \in [\texttt{MinReal}, \texttt{MaxReal}]$ und

$$|f(\tilde{x}) - \tilde{f}(\tilde{x})| = 0$$

für alle $\tilde{x} \in S$ mit $|f(\tilde{x})| \in [0, \texttt{MinReal})$[6]. Damit sind die Voraussetzungen von Satz 4.1 erfüllt, und es gilt mit $\varepsilon(\mathrm{sqrt}) := \overline{\varepsilon}$

$$|\sqrt{a} - \widetilde{\sqrt{\tilde{a}}}| \le \Delta_{dat,\mathrm{sqrt}} + \varepsilon(\mathrm{sqrt})\Big(\Delta_{dat,\mathrm{sqrt}} + |\sqrt{a}|\Big)\,,$$

wobei

$$\Delta_{dat,\mathrm{sqrt}} = \left|\frac{\Delta(a)}{2\sqrt{a+[-\Delta(a),\Delta(a)]}}\right| = \frac{\Delta(a)}{2\sqrt{a-\Delta(a)}} = \frac{\varepsilon(a)\sqrt{a}}{2\sqrt{1-\varepsilon(a)}}\,.$$

Die dabei auftretenden Radikanden sind wegen $\inf(\tilde{A}) > 0$ alle größer als Null. Für den absoluten Fehler erhält man demnach die folgenden Schranken:

$$\begin{aligned}
|\sqrt{a} - \widetilde{\sqrt{\tilde{a}}}| &\le \frac{\Delta(a)}{2\sqrt{a-\Delta(a)}} + \varepsilon(\mathrm{sqrt})\left(\frac{\Delta(a)}{2\sqrt{a-\Delta(a)}} + \sqrt{a}\right) \qquad (52)\\
&\overset{\text{Lemma 4.1}}{\le} \max_{a\in\{\underline{a},\overline{a}\}}\left(\frac{\Delta(a)}{2\sqrt{a-\Delta(a)}} + \varepsilon(\mathrm{sqrt})\left(\frac{\Delta(a)}{2\sqrt{a-\Delta(a)}} + \sqrt{a}\right)\right)\\
|\sqrt{a} - \widetilde{\sqrt{\tilde{a}}}| &\le \frac{\varepsilon(a)\sqrt{a}}{2\sqrt{1-\varepsilon(a)}} + \varepsilon(\mathrm{sqrt})\left(\frac{\varepsilon(a)\sqrt{a}}{2\sqrt{1-\varepsilon(a)}} + \sqrt{a}\right)\\
&\le |\sqrt{A}|\left(\frac{\varepsilon(a)}{2\sqrt{1-\varepsilon(a)}} + \varepsilon(\mathrm{sqrt})\left(\frac{\varepsilon(a)}{2\sqrt{1-\varepsilon(a)}} + 1\right)\right)\,.
\end{aligned}$$

Aus (52) ergeben sich die Schranken für den relativen Fehler entsprechend zu

$$\left|\frac{\sqrt{a} - \widetilde{\sqrt{\tilde{a}}}}{\sqrt{a}}\right| \le \frac{\frac{\Delta(a)}{a}}{2\sqrt{1-\frac{\Delta(a)}{a}}} + \varepsilon(\mathrm{sqrt})\left(\frac{\frac{\Delta(a)}{a}}{2\sqrt{1-\frac{\Delta(a)}{a}}} + 1\right)$$

[6] Dieser Fall tritt nur für $\tilde{x} = 0$ ein; nach IEEE-754 gilt $\widetilde{\sqrt{0}} = \sqrt{0} = 0$.

$$\leq \begin{cases} \dfrac{\frac{\Delta(a)}{\langle A\rangle}}{2\sqrt{1-\frac{\Delta(a)}{\langle A\rangle}}} + \varepsilon(\mathrm{sqrt})\left(\dfrac{\frac{\Delta(a)}{\langle A\rangle}}{2\sqrt{1-\frac{\Delta(a)}{\langle A\rangle}}}+1\right) \\ \dfrac{\varepsilon(a)}{2\sqrt{1-\varepsilon(a)}} + \varepsilon(\mathrm{sqrt})\left(\dfrac{\varepsilon(a)}{2\sqrt{1-\varepsilon(a)}}+1\right) . \end{cases}$$

Die Ergebnisse sind in Tabelle 3 noch einmal zusammengefaßt.

abzuschätzen	gegeben	Gesamtfehlerabschätzung
$\lvert\sqrt{a}-\widetilde{\sqrt{\tilde{a}}}\rvert$	$\Delta(a)$	$\max\limits_{a\in\{\underline{a},\overline{a}\}}\left(\frac{\Delta(a)}{2\sqrt{a-\Delta(a)}}+\varepsilon(\mathrm{sqrt})\left(\frac{\Delta(a)}{2\sqrt{a-\Delta(a)}}+\sqrt{a}\right)\right)$
	$\varepsilon(a)$	$\lvert\sqrt{A}\rvert\left(\frac{\varepsilon(a)}{2\sqrt{1-\varepsilon(a)}}+\varepsilon(\mathrm{sqrt})\left(\frac{\varepsilon(a)}{2\sqrt{1-\varepsilon(a)}}+1\right)\right)$
$\left\lvert\frac{\sqrt{a}-\widetilde{\sqrt{\tilde{a}}}}{\sqrt{a}}\right\rvert$	$\Delta(a)$	$\frac{\frac{\Delta(a)}{\langle A\rangle}}{2\sqrt{1-\frac{\Delta(a)}{\langle A\rangle}}}+\varepsilon(\mathrm{sqrt})\left(\frac{\frac{\Delta(a)}{\langle A\rangle}}{2\sqrt{1-\frac{\Delta(a)}{\langle A\rangle}}}+1\right)$
	$\varepsilon(a)$	$\frac{\varepsilon(a)}{2\sqrt{1-\varepsilon(a)}}+\varepsilon(\mathrm{sqrt})\left(\frac{\varepsilon(a)}{2\sqrt{1-\varepsilon(a)}}+1\right)$

Tabelle 3: Fehlerschranken für die Funktion $\mathrm{sqrt}(x)$

4.2.2 Die Funktion $\exp(x) = e^x$

Lemma 4.2 *Für $c_1, c_2, c_3 \geq 0$ nimmt die Funktion $f : X \to \mathbb{R}$, $X = [\underline{x}, \overline{x}] \in I\mathbb{R}$, die durch*

$$f(x) := c_1|x|e^{c_1|x|} + c_2\left(c_1|x|e^{c_1|x|}+1\right) + \frac{c_3}{e^x}$$

definiert sei, ihr Maximum am Rand von X an, d. h.

$$\max_{x\in X} f(x) = \max_{x\in\{\underline{x},\overline{x}\}} f(x) .$$

Beweis: Es ist

$$f'(x) = (-c_1 + c_1^2 x)e^{-c_1 x}(1+c_2) - \frac{c_3}{e^x} \leq 0 \quad \text{für} \quad x < 0$$

und

$$f''(x) = (2c_1^2 + c_1^3 x)e^{c_1 x}(1+c_2) + \frac{c_3}{e^x} \geq 0 \quad \text{für} \quad x > 0 ,$$

d. h. f fällt monoton für $x < 0$ und ist konvex für $x > 0$. Mit der Stetigkeit von f, insbesondere in $x = 0$, folgt die Behauptung ∎

Die Funktion $f(x) = \exp(x) = e^x$ ist auf ganz $\mathbb{R}$ differenzierbar, also auch auf dem Intervall $\tilde{A} = A + [-\Delta(a), \Delta(a)] = A \cdot [1 - \varepsilon(a), 1 + \varepsilon(a)]$. Falls zwei Zahlen $\varepsilon(\exp), \Delta(\exp) \geq 0$ mit

$$|f(\tilde{x}) - \tilde{f}(\tilde{x})| \leq \varepsilon(\exp)|f(\tilde{x})| \quad \forall \tilde{x} \in S \text{ mit } |f(\tilde{x})| \in [\texttt{MinReal}, \texttt{MaxReal}]$$

und

$$|f(\tilde{x}) - \tilde{f}(\tilde{x})| \leq \Delta(\exp) \quad \forall \tilde{x} \in S \text{ mit } |f(\tilde{x})| \in [0, \texttt{MinReal})$$

existieren und bekannt sind, sind die Voraussetzungen von Satz 4.1 erfüllt, und man hat die Abschätzung

$$|e^a - \widetilde{e^{\tilde{a}}}| \leq \Delta_{dat,\exp} + \varepsilon(\exp)\left(\Delta_{dat,\exp} + |e^a|\right) + \Delta(\exp)\,,$$

wobei

$$\Delta_{dat,\exp} = \Delta(a) \cdot \left|e^{a+[-\Delta(a),\Delta(a)]}\right| = \Delta(a)e^{\Delta(a)} \cdot e^a = |a|\varepsilon(a)e^{|a|\varepsilon(a)} \cdot e^a\,.$$

Weiter gilt für den absoluten Fehler

$$\begin{aligned}
|e^a - \widetilde{e^{\tilde{a}}}| &\leq \Delta(a)e^{\Delta(a)} \cdot e^a + \varepsilon(\exp)\left(\Delta(a)e^{\Delta(a)} \cdot e^a + e^a\right) + \Delta(\exp) \\
&\leq \begin{cases} e^{\overline{a}}\left(\Delta(a)e^{\Delta(a)} + \varepsilon(\exp)\left(\Delta(a)e^{\Delta(a)} + 1\right)\right) + \Delta(\exp) \\ e^{\overline{a}}\left(|A|\varepsilon(a)e^{|A|\varepsilon(a)} + \varepsilon(\exp)\left(|A|\varepsilon(a)e^{|A|\varepsilon(a)} + 1\right)\right) + \Delta(\exp) \end{cases}
\end{aligned}$$

und damit für den relativen Fehler

$$\begin{aligned}
\left|\frac{e^a - \widetilde{e^{\tilde{a}}}}{e^a}\right| &\leq \Delta(a)e^{\Delta(a)} + \varepsilon(\exp)\left(\Delta(a)e^{\Delta(a)} + 1\right) + \frac{\Delta(\exp)}{e^{\underline{a}}} \\
\left|\frac{e^a - \widetilde{e^{\tilde{a}}}}{e^a}\right| &\leq |a|\varepsilon(a)e^{|a|\varepsilon(a)} + \varepsilon(\exp)\left(|a|\varepsilon(a)e^{|a|\varepsilon(a)} + 1\right) + \frac{\Delta(\exp)}{e^a} \\
&\overset{\text{Lemma 4.2}}{\leq} \max_{a \in \{\underline{a}, \overline{a}\}} \left(|a|\varepsilon(a)e^{|a|\varepsilon(a)} + \varepsilon(\exp)\left(|a|\varepsilon(a)e^{|a|\varepsilon(a)} + 1\right) + \frac{\Delta(\exp)}{e^a}\right)\,.
\end{aligned}$$

Die Ergebnisse sind in Tabelle 4 noch einmal zusammengestellt.

4.2.3 Die Funktion $\ln(x)$

Für die Herleitung von Fehlerschranken für die Logarithmusfunktion wird das folgende Lemma verwendet.

abzuschätzen	gegeben	Gesamtfehlerabschätzung
$\lvert e^a - \widetilde{e^{\tilde{a}}}\rvert$	$\Delta(a)$	$e^{\overline{a}}\left(\Delta(a)e^{\Delta(a)} + \varepsilon(\exp)\left(\Delta(a)e^{\Delta(a)}+1\right)\right) + \Delta(\exp)$
	$\varepsilon(a)$	$e^{\overline{a}}\left(\lvert A\rvert\varepsilon(a)e^{\lvert A\rvert\varepsilon(a)} + \varepsilon(\exp)\left(\lvert A\rvert\varepsilon(a)e^{\lvert A\rvert\varepsilon(a)}+1\right)\right) + \Delta(\exp)$
$\left\lvert\frac{e^a - \widetilde{e^{\tilde{a}}}}{e^a}\right\rvert$	$\Delta(a)$	$\Delta(a)e^{\Delta(a)} + \varepsilon(\exp)\left(\Delta(a)e^{\Delta(a)}+1\right) + \frac{\Delta(\exp)}{e^{\underline{a}}}$
	$\varepsilon(a)$	$\max\limits_{a\in\{\underline{a},\overline{a}\}}\left(\lvert a\rvert\varepsilon(a)e^{\lvert a\rvert\varepsilon(a)} + \varepsilon(\exp)\left(\lvert a\rvert\varepsilon(a)e^{\lvert a\rvert\varepsilon(a)}+1\right) + \frac{\Delta(\exp)}{e^a}\right)$

Tabelle 4: Fehlerschranken für die Funktion $\exp(x)$

Lemma 4.3 *Für $c_1, c_2 \geq 0$ nimmt die Funktion $f : X \to I\!R$, $X = [\underline{x}, \overline{x}] \in I\!I\!R$, $\underline{x} > c_1$, die durch*

$$f(x) := \frac{c_1}{x - c_1} + c_2\left(\frac{c_1}{x - c_1} + |\ln(x)|\right)$$

definiert sei, ihr Maximum am Rand von X an, d. h.

$$\max_{x\in X} f(x) = \max_{x\in\{\underline{x},\overline{x}\}} f(x)\,.$$

Beweis: Es ist

$$f'(x) = -\frac{c_1(c_2+1)}{(x-c_1)^2} - \frac{c_2}{x} \leq 0 \quad \text{für} \quad c_1 < x < 1,$$

und für $x > 1$ gilt

$$\begin{aligned} f'(x) = -\frac{c_1(c_2+1)}{(x-c_1)^2} + \frac{c_2}{x} &= 0 \\ \iff x_{1/2} &= \frac{3c_1c_2 + c_1 \pm \sqrt{5c_1^2c_2^2 + 6c_1^2c_2 + c_1^2}}{2c_2}\,, \end{aligned}$$

wobei $x_2 \notin X$ wegen

$$x_2 = \frac{3c_1c_2 + c_1 - \sqrt{5c_1^2c_2^2 + 6c_1^2c_2 + c_1^2}}{2c_2} \leq \frac{3c_1c_2 + c_1 - \sqrt{(2c_1c_2 + c_1)^2}}{2c_2} = \frac{c_1}{2} \leq c_1\,.$$

Da $f''(x_1) > 0$ und f stetig ist, ist $f'(x) < 0$ für $x < \max\{1, x_1\}$ und $f'(x) > 0$ für $x > \max\{1, x_1\}$, woraus die Behauptung folgt ∎

Die Funktion $f(x) = \ln(x)$ ist differenzierbar auf dem Intervall $\tilde{A} = A + [-\Delta(a), \Delta(a)] = A \cdot [1 - \varepsilon(a), 1 + \varepsilon(a)]$, wenn $\inf(\tilde{A}) > 0$ ist. Falls zwei Zahlen $\varepsilon(\ln), \Delta(\ln) \geq 0$ mit

$$|f(\tilde{x}) - \tilde{f}(\tilde{x})| \leq \varepsilon(\ln)|f(\tilde{x})| \quad \forall \tilde{x} \in S \text{ mit } |f(\tilde{x})| \in [\texttt{MinReal}, \texttt{MaxReal}]$$

und

$$|f(\tilde{x}) - \tilde{f}(\tilde{x})| \leq \Delta(\ln) \quad \forall \tilde{x} \in S \text{ mit } |f(\tilde{x})| \in [0, \texttt{MinReal})$$

gegeben sind, sind die Voraussetzungen von Satz 4.1 erfüllt, und es ergibt sich die Abschätzung

$$|\ln(a) - \widetilde{\ln(\tilde{a})}| \leq \Delta_{dat,\ln} + \varepsilon(\ln)\Big(\Delta_{dat,\ln} + |\ln(a)|\Big) + \Delta(\ln)\,,$$

wobei

$$\Delta_{dat,\ln} = \Delta(a) \cdot \left| \frac{1}{a + [-\Delta(a), \Delta(a)]} \right| = \frac{\Delta(a)}{a - \Delta(a)} = \frac{\varepsilon(a)}{1 - \varepsilon(a)}\,.$$

Wegen $\inf(\tilde{A}) > 0$ sind die auftretenden Nenner größer als Null. Man hat also die folgenden Schranken für den absoluten Fehler:

$$\begin{aligned}
|\ln(a) - \widetilde{\ln(\tilde{a})}| &\leq \frac{\Delta(a)}{a - \Delta(a)} + \varepsilon(\ln)\left(\frac{\Delta(a)}{a - \Delta(a)} + |\ln(a)|\right) + \Delta(\ln) \\
&\overset{\text{Lemma 4.3}}{\leq} \max_{a \in \{\underline{a}, \overline{a}\}} \left(\frac{\Delta(a)}{a - \Delta(a)} + \varepsilon(\ln)\left(\frac{\Delta(a)}{a - \Delta(a)} + |\ln(a)|\right)\right) + \Delta(\ln) \\
|\ln(a) - \widetilde{\ln(\tilde{a})}| &\leq \frac{\varepsilon(a)}{1 - \varepsilon(a)} + \varepsilon(\ln)\left(\frac{\varepsilon(a)}{1 - \varepsilon(a)} + |\ln(A)|\right) + \Delta(\ln)\,.
\end{aligned}$$

Daraus lassen sich jetzt auch die Schranken für den relativen Fehler ableiten:

$$\begin{aligned}
\left|\frac{\ln(a) - \widetilde{\ln(\tilde{a})}}{\ln(a)}\right| &\leq \frac{\Delta(a)}{(a - \Delta(a))|\ln(a)|} + \varepsilon(\ln)\left(\frac{\Delta(a)}{(a - \Delta(a))|\ln(a)|} + 1\right) + \frac{\Delta(\ln)}{|\ln(a)|} \\
&\leq \begin{cases} \dfrac{\Delta(a)}{(\underline{a} - \Delta(a))\langle\ln(A)\rangle} + \varepsilon(\ln)\left(\dfrac{\Delta(a)}{(\underline{a} - \Delta(a))\langle\ln(A)\rangle} + 1\right) + \dfrac{\Delta(\ln)}{\langle\ln(A)\rangle} \\ \dfrac{\varepsilon(a)}{(1 - \varepsilon(a))\langle\ln(A)\rangle} + \varepsilon(\ln)\left(\dfrac{\varepsilon(a)}{(1 - \varepsilon(a))\langle\ln(A)\rangle} + 1\right) + \dfrac{\Delta(\ln)}{\langle\ln(A)\rangle}\,. \end{cases}
\end{aligned}$$

Die Ergebnisse sind in Tabelle 5 noch einmal zusammengefaßt.

Weitere häufig in Programmiersprachen vorgegebene mathematische Funktionen $(\sin, \cos, \arcsin, \cdots)$ können analog abgeschätzt werden.

5 Zusammenfassung

Für jede der in dieser Arbeit hergeleiteten Fehlerschranken kann unter Verwendung von Intervalloperationen und gerichtet gerundeten Operationen eine Gleitkommazahl automatisch bestimmt werden, die mit Sicherheit den theoretischen Wert der Schranke nicht unterschreitet. Durch Abschätzung jeder einzelnen Grundoperation $(+, -, \cdot, /, \sin, \cos, \ldots)$ des zu einem Gleitkommaalgorithmus äquivalenten

abzuschätzen	gegeben	Gesamtfehlerabschätzung
$\lvert \ln(a) - \widetilde{\ln(\tilde{a})} \rvert$	$\Delta(a)$	$\max\limits_{a \in \{\underline{a}, \overline{a}\}} \left(\frac{\Delta(a)}{a - \Delta(a)} + \varepsilon(\ln) \left(\frac{\Delta(a)}{a - \Delta(a)} + \lvert \ln(a) \rvert \right) \right) + \Delta(\ln)$
	$\varepsilon(a)$	$\frac{\varepsilon(a)}{1 - \varepsilon(a)} + \varepsilon(\ln) \left(\frac{\varepsilon(a)}{1 - \varepsilon(a)} + \lvert \ln(A) \rvert \right) + \Delta(\ln)$
$\left\lvert \frac{\ln(a) - \widetilde{\ln(\tilde{a})}}{\ln(a)} \right\rvert$	$\Delta(a)$	$\frac{\Delta(a)}{(\underline{a} - \Delta(a)) \langle \ln(A) \rangle} + \varepsilon(\ln) \left(\frac{\Delta(a)}{(\underline{a} - \Delta(a)) \langle \ln(A) \rangle} + 1 \right) + \frac{\Delta(\ln)}{\langle \ln(A) \rangle}$
	$\varepsilon(a)$	$\frac{\varepsilon(a)}{(1 - \varepsilon(a)) \langle \ln(A) \rangle} + \varepsilon(\ln) \left(\frac{\varepsilon(a)}{(1 - \varepsilon(a)) \langle \ln(A) \rangle} + 1 \right) + \frac{\Delta(\ln)}{\langle \ln(A) \rangle}$

Tabelle 5: Fehlerschranken für die Funktion $\ln(x)$

Ausdrucksbaumes (Code-Liste) kann damit eine Gesamtfehlerschranke automatisch mit dem Rechner bestimmt werden [19]. Das Ergebnis der a priori durchgeführten Vorwärtsfehleranalyse kann dann zur Laufzeit eines Programms durch Berücksichtigung der bekannten Gesamtfehlerschranke zur sicheren Einschließung des exakten Ergebnisses verwendet werden.

Die mit dem hergeleiteten Kalkül berechneten Fehlerschranken liefern **verläßliche quantitative Werte**. Sie sind damit z. B. den nur **qualitativen** Aussagen von Fehlerschätzungen erster Ordnung und Fehleraussagen in der Landau-Symbolik vorzuziehen.

Konkrete Anwendungsbeispiele des Kalküls für absolute Fehlerschranken (siehe z. B. [2], [12], [13], [19]) haben gezeigt, daß die gefundenen worst case Schranken realistisch sind. Erste Anwendungen des relativen Kalküls finden sich in [2]. Sie zeigen, daß gerade dann, wenn eine relative Gesamtfehlerschranke über einem größeren Datenbereich gesucht ist, im Vergleich zum absoluten Kalkül der Feinheitsgrad von Bereichsunterteilungen zum Teil wesentlich gröber ausfallen darf, ohne daß dadurch die Güte der Gesamtfehlerschranke leiden würde (Laufzeitgewinn).

Es muß hervorgehoben werden, daß die in der vorliegenden Arbeit durchgeführten Überlegungen ihre eigentliche Schlagkraft im Zusammenspiel mit einer Einbettung in eine geeignete Programmierumgebung erhalten. Bereits für kurze Programmstücke ist eine verläßliche Fehlerabschätzung per Handrechnung viel zu aufwendig und zu fehleranfällig. Erst die Möglichkeit der Funktionsüberladung, das Vorhandensein eines Operatorkonzeptes sowie die Durchführbarkeit von gerichtet gerundeten Operationen und Intervalloperationen erlaubt es, die Möglichkeiten des vorgestellten Fehlerkalküls voll auszuschöpfen.

Andere Arten der sicheren Fehlerabschätzung sind z. B. in [15], [9] (mittels Automatischer Differentiation), [23], [25] (mittels sogenannter Wiederberechnungsverfahren) vorgeschlagen. Erste numerische Vergleiche zu Abschätzungen, welche mittels der Rückwärtsmethode der Automatischen Differentiation gefunden wurden, sind in [2] durchgeführt. Die wenigen durchgerechneten Beispiele erlauben noch keine ab-

schließende Beurteilung. Die gewonnene Erfahrung zeigt jedoch bereits, daß der hier vorgestellte Kalkül wesentlich einfacher handhabbar ist.

6 Anhang: Notation und Bezeichnungen

$S = S(b, t, e_{min}, e_{max})$	Gleitpunktsystem mit Basis b, Mantissenlänge t und Exponent $e \in [e_{min}, e_{max}] \cap \mathbb{Z}$; in dieser Arbeit wird hauptsächlich das IEEE-`double`-Format $S = S(2, 53, -1022, +1023)$ verwendet
$\circ \in \{+, -, \cdot, /\}$	exakte reelle Operation
$\boxed{\circ}$, $\circ \in \{+, -, \cdot, /\}$	rundungsfehlerbehaftete Maschinenoperation
$\bigtriangledown$, $\circ \in \{+, -, \cdot, /\}$	Maschinenoperation mit Rundung nach unten
$\bigtriangleup$, $\circ \in \{+, -, \cdot, /\}$	Maschinenoperation mit Rundung nach oben
$\diamondsuit$, $\circ \in \{+, -, \cdot, /\}$	Maschinen-Intervalloperation
`MinReal`	kleinste positive normalisierte Gleitpunktzahl; für das IEEE-`double`-Format gilt: `MinReal`$:= 2^{-1022} = 2.22\ldots \cdot 10^{-308}$
`dMinReal`	kleinste positive denormalisierte Gleitpunktzahl; für das IEEE-`double`-Format gilt: `dMinReal`$:= 2^{-1074} = 4.94\ldots \cdot 10^{-324}$
`MaxReal`	größte normalisierte Gleitpunktzahl; für das IEEE-`double`-Format gilt `MaxReal`$:= 1.79\ldots \cdot 10^{308}$
U	Unterlaufbereich, $U := (-$`MinReal`$,$`MinReal`$)$
ε^*	relative Maschinengenauigkeit; für das IEEE-`double`-Format gilt $\varepsilon^* := \frac{1}{2} \cdot 2^{1-53} = 2^{-53}$
$\overline{\varepsilon}$	maximaler relativer Rundungsfehler; für das IEEE-`double`-Format gilt $\overline{\varepsilon} := 2^{1-53} = 2^{-52}$ für gerichtet gerundete Operationen, bei Rundung zur nächstgelegenen Gleitpunktzahl gilt $\overline{\varepsilon} := \varepsilon^*$
$\tilde{a}, \tilde{b}, \ldots$	auf der Maschine berechnete, i. allg. fehlerbehaftete Größen

$\tilde{A}$, $\tilde{B}$, ...	Einschließungen gestörter Größen
Δ_a, Δ_b, ...	exakte absolute Fehler; $\tilde{a} = a + \Delta_a$
ε_a, ε_b, ...	exakte relative Fehler; $\tilde{a} = a \cdot (1 + \varepsilon_a)$
$\Delta(a)$, $\Delta(b)$, ...	gleichmäßige Schranken für Δ_a, Δ_b, ...; $\lvert\Delta_a\rvert \leq \Delta(a)$
$\varepsilon(a)$, $\varepsilon(b)$, ...	gleichmäßige Schranken für ε_a, ε_b, ...; $\lvert\varepsilon_a\rvert \leq \varepsilon(a)$
Δ_{rnd}, ε_{rnd}	absoluter bzw. relativer Rundungsfehler einer Gleitkommaoperation
$\Delta_{rnd,N}$, $\varepsilon_{rnd,N}$	absoluter bzw. relativer Rundungsfehler einer Gleitkommaoperation mit Ergebnis im Bereich der normalisierten Zahlen
$\Delta_{rnd,U}$, $\varepsilon_{rnd,U}$	absoluter bzw. relativer Rundungsfehler einer Gleitkommaoperation mit Ergebnis im Unterlaufbereich
Δ_{dat}, ε_{dat}	absoluter bzw. relativer Datenfehler einer Gleitkommaoperation
$\Delta(\circ)$, $\varepsilon(\circ)$	gleichmäßige Schranke des absoluten bzw. relativen Gesamtfehlers einer Gleitkommaoperation
f	mathematische Standardfunktion
$\tilde{f}$	auf der Maschine implementierte Näherungsfunktion an f
$\Delta(f)$	Schranke für den absoluten Fehler der Funktion $\tilde{f} \approx f$ im Unterlaufbereich
$\varepsilon(f)$	Schranke für den relativen Fehler der Funktion $\tilde{f} \approx f$ im Bereich der normalisierten Zahlen
$s(x), m(x), e(x)$	Vorzeichen, Mantisse und Exponent von $x \in S$
z_{lead}, z_{trail}	Anzahl führender bzw. abschließender Nullen in der Mantisse einer Gleitkommazahl
$\mathbb{R}$	Menge der reellen Zahlen
$\mathbb{R}^+$	Menge der positiven reellen Zahlen
$I\mathbb{R}$	Menge der abgeschlossenen Intervalle über den reellen Zahlen

IS	$IS = \{[\underline{a}, \overline{a}] \mid \underline{a}, \overline{a} \in S,\ \underline{a} \leq \overline{a}\}$, Menge der Maschinenintervalle
$\lvert A\rvert$, $A \in I\!I\!R$	$\lvert A\rvert := \max_{a \in A} \lvert a\rvert$, Betragsmaximum eines Intervalls
$\langle A\rangle$, $A \in I\!I\!R$	$\langle A\rangle := \min_{a \in A} \lvert a\rvert$, Betragsminimum eines Intervalls
$d(A)$, $A \in I\!I\!R$	$d(A) := \sup(A) - \inf(A)$, Durchmesser eines Intervalls

Literatur

[1] Götz Alefeld, Jürgen Herzberger. *Einführung in die Intervallrechnung.* Bibliographisches Institut, Mannheim, 1974. xiii+398 S. ISBN 3-411-01466-0.

[2] Armin Bantle. *Ein Kalkül für verläßliche rechnergestützte Fehlerabschätzungen und dessen Anwendung.* Diplomarbeit am IWRMM, Universität Karlsruhe, 1998.

[3] F. L. Bauer. *Computational graphs and rounding error.* SIAM J. Numer. Anal., 11(1):87-96, 1974.

[4] Frithjof Blomquist, Walter Krämer. *Algorithmen mit garantierten Fehlerschranken für die Fehler- und die komplementäre Fehlerfunktion.* Preprint 97/3 des IWRMM, Universität Karlsruhe, 1997.

[5] Gerd Bohlender, Christian Ullrich. *Standards zur Computerarithmetik.* In: Jürgen Herzberger (Hrsg): *Wissenschaftliches Rechenen*, 9-46, Akademie Verlag, 1995.

[6] John W. Carr III. *Error analysis in floating point arithmetic.* Comm. ACM, 2(5):10-15, 1959.

[7] James W. Demmel. *Underflow and the reliability of numerical software.* SIAM J. Sci. Statist. Comput., 5(4):887-919, 1984.

[8] Warren E. Ferguson, Jr. *Exact computation of a sum or difference with applications to argument reduction.* In *Proc. 12th IEEE Symposium on Computer Arithmetic, Bath, England*, Simon Knowles and William H. McAllister, editors, IEEE Computer Society Press, Los Alamitos, CA, USA, 1995, pp. 216-221.

[9] Hans-Christoph Fischer. *Schnelle automatische Differentiation, Einschließungsmethoden und Anwendungen.* Dissertation, Universität Karlsruhe, 1990.

[10] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1996. xxviii+688 pp. ISBN 0-89871-355-2

[11] Werner Hofschuster, Walter Krämer. *Ein rechnergestützter Fehlerkalkül mit Anwendung auf ein genaues Tabellenverfahren.* Preprint 96/5 des IWRMM, Universität Karlsruhe, 1996.

[12] Werner Hofschuster, Walter Krämer. *Eine schnelle und portable Funktionsbibliothek für reelle Argumente und reelle Intervalle im IEEE-Double-Format.* Institut für Wissenschaftliches Rechnen und Mathematische Modellbildung, Universität Karlsruhe, 1997. (Diese Dokumentation kann im IWRMM eingesehen werden.)

[13] Werner Hofschuster, Walter Krämer. *A Computer Oriented Approach to Get Sharp Reliable Error Bounds*, *Reliable Computing 3*, pp. 239-248, 1997.

[14] *IEEE Standard for Binary Floating-Point Arithmetic, ANSI/IEEE Standard 754-1985.* Institute of Electrical and Electronics Engineers, New York, 1985. Reprinted in SIGPLAN Notices, 22(2):9-25, 1987.

[15] M. Iri. *Simultaneous Computation of Functions, Partial Derivatives and Error Estimates of Rounding Errors.* Japan J. Appl. Math. 1, pp 223-252, 1984.

[16] R. Klatte, U. Kulisch, C. Lawo, M. Rauch, A. Wiethoff. *C-XSC: A C++ Class Library for Extended Scientific Computing.* Springer-Verlag, Heidelberg, 1993. xvi+269 S. ISBN 3-540-56328-8.

[17] Walter Krämer. *Eine Fehlerfaktorarithmetik für zuverlässige a priori Fehlerabschätzungen.* Preprint 97/5 des IWRMM, Universität Karlsruhe, 1997.

[18] Walter Krämer. *A priori Worst Case Error Bounds for Floating-Point Computations.* Proceedings of the 13th IEEE Symp. on Computer Arithmetic, Asilomar, California, pp. 64-71, 1997.

[19] Walter Krämer. *Constructive Error Analysis.* Journal of Universal Computer Science (JUCS), Vol. 4, No. 2, pp. 147-163, 1998.

[20] Ulrich Kulisch. *Grundlagen des Numerischen Rechnens.* Bibliographisches Institut, Mannheim, 1976. 467 S. ISBN 3-411-01517-9.

[21] Webb Miller. *Software for roundoff analysis.* ACM Trans. Math. Software, 1(2):108-128, 1975.

[22] Douglas M. Priest. *On Properties of Floating Point Arithmetics: Numerical Stability and the Cost of Accurate Computations.* Ph.D. thesis, Mathematics Department, University of California, Berkeley, CA, USA, November

1992. 126 pp. URL = `ftp://ftp.icsi.berkeley.edu/pub/theory/priest-thesis.ps.Z`.

[23] Paul L. Richman. *Automatic error analysis for determining precision.* Comm. ACM, 15(9):813-817, 1972.

[24] R. Scherer, K. Zeller. *Shorthand Notation for Rounding Errors.* Computing Suppl. 2, 165-168, Springer Verlag, 1980.

[25] Günter Schumacher. *Genauigkeitsfragen bei algebraisch-numerischen Algorithmen auf Skalar- und Vektorrechnern.* Dissertation, Universität Karlsruhe, 1989.

[26] Pat H. Sterbenz. *Floating-Point Computation.* Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1974. xiv+316 pp. ISBN 0-13-322495-3.

[27] Josef Stoer. *Numerische Mathematik 1.* 7. Aufl., Springer-Verlag, Heidelberg, 1994. xi+367 S. ISBN 3-540-57823-4.

[28] F. Stummel. *Rounding Error Analysis of Elementary Numerical Algorithms.* Computing. Suppl. 2, 169-195, Springer Verlag, 1980.

Folgende Arbeiten sind bisher in der Preprintreihe des IWRMM erschienen:

Nr. 93/1: G. Aumann, K. Bentz: Geometrische Stetigkeit beliebiger Ordnung zwischen Tensor-Produkt-Bézier-Flächen

Nr. 93/2: G. Alefeld, G. Mayer: A Computer Aided Existence and Uniqueness Proof for an Inverse Matrix Eigenvalue Problem

Nr. 93/3: B. Weber: Symbolische Programmierung in der Mehrkörperdynamik

Nr. 93/4: R. Rihm: Über Einschließungsverfahren für gewöhnliche Anfangswertprobleme und ihre Anwendung auf Differentialgleichungen mit unstetiger rechter Seite

Nr. 93/5: J. Wittenburg: Explizite Lösungen für lineare Gleichungssysteme mit tridiagonalen Koeffizientenmatrizen. Anwendungen in der Mechanik

Nr. 93/6: N. Henze, B. Klar: Goodness–of–Fit Testing for a Space–Time Model for Daily Rainfall

Nr. 93/7: K. Schweizerhof, J. Riccius, M. Baumann: Verbesserung von Finite Element Berechnungen durch Adaptivität und Netzglättung am Beispiel ebener und gekrümmter Flächentragwerke

Nr. 93/8: G. Starke: Subspace Orthogonalization for Substructuring Preconditioners for Nonselfadjoint Elliptic Problems

Nr. 93/9: N. Henze, B. Klar: Empirical Distribution Function Tests for the Generalized Poisson Model

Nr. 94/1: G. Aumann: Geometric Continuity of Parametric Curves and Surfaces

Nr. 94/2: T. Dehn, M. Eiermann, K. Giebermann, V. Sperling: Structured Sparse Matrix-Vector Multiplication on Massively Parallel Architectures

Nr. 94/3: W. Krämer: Bericht über die Begutachtung des IWRMM im Dezember 1993

Nr. 95/1: L. Kobbelt: Interpolatory Refinement is Low Pass Filtering

Nr. 95/2: M. Paluszny, H. Prautzsch, M. Schäfer: Corner cutting and interpolatory refinement

Nr. 95/3: B. Klar: Analysis of and Goodness of Fit Testing for a Flexible Discrete Time Failure Model

Nr. 95/4: P. Vielsack: Regularisierung von Haftkräften bei Coulombscher Reibung

Nr. 95/5: P. Vielsack, M. Storz: Bifurcation of Motion in a Technical System with Stick-Slip and Impact

Nr. 95/6: M. Brühl: A Curve Tracing Algorithm for Computing the Pseudospectrum

Nr. 95/7: J. Riccius, K. Schweizerhof, M. Baumann: On the treatment of shell intersections in adaptive finite element analysis and combination with mesh smoothing

Nr. 96/1: M. Dormanns, H.-U. Heiß: Nutzung von Asynchronität bei iterativen Gleichungslösern auf Multirechnersystemen

Nr. 96/2: P. Vielsack, J. Kirillowa: Nichteindeutigkeit der Bewegungen eines Reibschwingers mit Selbsterregung

Nr. 96/3: L. Kobbelt, T. Hesse, H. Prautzsch, K. Schweizerhof: Diskrete Freiformflächenerzeugung für FEM-Anwendungen

Nr. 96/4: M. Brühl, M. Hanke, H. Wanzki: Ein Rekonstruktionsverfahren für die elektrische Impedanztomographie

Nr. 96/5 : W. Hofschuster, W. Krämer: Ein rechnergestützter Fehlerkalkül mit Anwendung auf ein genaues Tabellenverfahren

Nr. 96/6: W. Niethammer, W. Krämer (Herausgeber): Tagungsband zum Workshop „Wissenschaftliches Rechnen in den Ingenieurwissenschaften“

Nr. 96/7: G. Freimann: FAS-Verfahren zur Lösung strukturmechanischer Probleme

Nr. 97/1: P. Vielsack, A. Hartung: Orbitale Stabilität von Bewegungen mit Pausen bei Einwirkung permanenter Störungen

Nr. 97/2: J. G. Schmidt, G. Starke: Coarse Space Orthogonalization for Indefinite Linear Systems of Equations Arising in Geometrically Nonlinear Elasticity

Nr. 97/3: F. Blomquist, W. Krämer: Algorithmen mit garantierten Fehlerschranken für die Fehler- und die komplementäre Fehlerfunktion

Nr. 98/1: S. Doll, R. Hauptmann, K. Schweizerhof, C. Freischläger: Selective Reduced Integration and Volumetric Locking in Finite Deformation Elastoviscoplasticity

Nr. 98/2: P. Vielsack, H. Kammerer: Finite Element Formulierung nichtglatter Schwingungen eines Balkens mit Reibglied

Nr. 98/3: H. Prautzsch: How Smooth are Subdividable Surfaces at Extraordinary Points?

Nr. 98/4: W. Wu, W. Rodi, Th. Wenka: 3D Numerical Modeling of Flow and Sediment Transport in Open Channels

Nr. 98/5: A. Bantle, W. Krämer: Ein Kalkül für verläßliche absolute und relative Fehlerabschätzungen

Weitere Arbeiten sind in Vorbereitung.