ZAMM 54, 39 – 51 (1974)

A. NEUMAIER

Rundungsfehleranalyse einiger Verfahren zur Summation endlicher Summen

Der bei einer Summation auftretende Rundungsfehler kann als Maß für die Güte des verwendeten Verfahrens gelten. Im folgenden werden für mehrere Summierungsverfahren, unter anderem für das übliche und das Kahan-Babuška-Verfahren, a-priori-Schranken für diese Rundungsfehler angegeben und miteinander verglichen.

The rounding-error arising during summation can be interpreted as a measure for the quality of the procedure used. In the following, a-priori-bounds for this rounding-error are used to compare several summation procedures, e.g. the common procedure and the method of Kahan-Babuška.

Возникающие при суммировании ошибки скругления могут рассматриваться как мера качества для применяемого метода. В последующем приведены и сравнены между собой для многих методов суммирования, в том числе для "обычного" метода и для метода Каган-Бабуска, априорные пределы для этих ошибок скругления.

1. Aufgabenstellung

Zur Berechnung einer endlichen Summe

$$s := \sum_{m=1}^{n} a_m \tag{1}$$

können verschiedene Verfahren benutzt werden. Bei Durchführung der Rechnung im Körper der reellen Zahlen liefert jedes Verfahren das Ergebnis s; wird die Rechnung aber auf einer Rechenmaschine durchgeführt, so ergeben sich im allgemeinen verschiedene Ergebnisse, da die auftretenden Rundungsfehler vom Verfahren abhängen.

Die *Güte* einer Summationsvorschrift kann z. B. durch das Verhältnis von Genauigkeit des berechneten Wertes zum Arbeitsaufwand angegeben werden. Gesucht sind daher Schranken für die Abweichung $\tilde{s} - s$ des nach einem bestimmten Verfahren berechneten Näherungswertes $\tilde{s}$ für s.

2. Summationsvorschriften

Es werden die folgenden vier Summationsvorschriften untersucht[1]):

I. Normalverfahren

Man setzt

$$\begin{aligned} s_0 &:= 0 \\ s_m &:= s_{m-1} + a_m \qquad \text{für} \quad m = 1, \ldots, n \\ s &:= s_n \,, \end{aligned} \tag{I,1}$$

d. h., die Summe (1) wird von links nach rechts abgearbeitet, z. B.

$$s = \big((a_1 + a_2) + a_3\big) + a_4 \qquad (n = 4)$$

Der Rechenaufwand beträgt $n - 1$ Additionen (Addition von 0 wird vernachlässigt).

II. Verbessertes Normalverfahren

2^t sei die kleinste Potenz von 2 mit $2^t \geqq n$. Setzt man

$$\begin{aligned} s_{m\,0} &:= a_m \qquad \text{für} \quad m = 1, \ldots, n \\ s_{m\,0} &:= 0 \qquad \text{für} \quad m = n+1, \ldots, 2^t \,, \end{aligned}$$

so berechnet auch die Vorschrift

$$\begin{aligned} s_{m\,k} &:= s_{2\,m-1,\,k-1} + s_{2\,m,\,k-1} \qquad \text{für} \quad 1 \leqq k \leqq t \,, \quad 0 \leqq m \leqq 2^{t-k} - 1 \\ s &:= s_{1\,t} \end{aligned} \tag{II,1}$$

die Summe s, aber in anderer Klammerung, z. B.

$$s = \big((a_1 + a_2) + (a_3 + a_4)\big) + \big((a_5 + a_6) + (a_7 + a_8)\big) \qquad (t = 3,\ n = 8) \,.$$

Wieder sind $n - 1$ Additionen auszuführen.

---

[1]) BABUŠKA [1] behandelte die Verfahren I, II, III unter den Bezeichnungen $L_{1\,A}$, $L_{1\,B}$, $L_{1\,C}$.

III. KAHAN-BABUŠKA-Verfahren

Man setzt [1], [2], [3]

$$\begin{aligned} &s_0 := 0\,, && s_m := a_m + s_{m-1} && (m = 1, \ldots, n)\\ &w_0 := 0\,, && w_m := w_{m-1} + \big(a_m + (s_{m-1} - s_m)\big) && (m = 1, \ldots, n)\\ &s := s_n + w_n\,. \end{aligned} \tag{III,1}$$

Die w_m sind bei reeller Rechnung Null; wird auf einer Maschine gerechnet, so sammeln sich die Rundungsfehler in diesem Term an. Durch die Korrektur $+ w_n$ werden die Rundungsfehler etwa kompensiert. Es ist also eine größere Genauigkeit zu erwarten. Der Rechenaufwand beträgt hier $4n - 3$ Additionen/Subtraktionen, ist also etwa viermal so groß wie in Verfahren I und II.

IV. Verbesscrtes KAHAN-BABUŠKA-Verfahren

Man setzt

$$\begin{aligned} &s_0 := 0\,, && s_m := a_m + s_{m-1} && (m = 1, \ldots, n)\\ &w_0 := 0\,, && w_m := w_{m-1} + \big(a_m + (s_{m-1} - s_m)\big)\,, && \text{wenn } |a_m| \leqq |s_{m-1}|\\ &&& w_m := w_{m-1} + \big(s_{m-1} + (a_m - s_m)\big)\,, && \text{wenn } |a_m| > |s_{m-1}|\\ &s := s_n + w_n\,. \end{aligned} \tag{IV,1}$$

Das Verfahren unterscheidet sich von Verfahren III nur für $|a_m| > |s_{m-1}|$. (Z. B. ist IV = III, wenn (a_m) eine monoton fallende Folge positiver reeller Zahlen ist, da $|a_m| > |s_{m-1}|$ genau für $m = 1$.) Hier sind außer $4n - 3$ Additionen/Subtraktionen noch n Abfragen auszuführen. Rechnet man für eine Abfrage denselben Aufwand wie für eine Subtraktion, so ist der Aufwand für Verfahren IV etwa fünfmal so groß wie für Verfahren I oder II, und 1/4mal größer als in Verfahren III.

3. Rundungsfehler beim Normalverfahren (I und II)

Nach WILKINSON [4] genügt die Maschinennäherung $a \mathbin{\tilde{+}} b$ einer Summe $a + b$ für viele Rechenmaschinen der Formel

$$a \mathbin{\tilde{+}} b = (a + b)(1 + \delta)\,, \qquad |\delta| \leqq \varepsilon\,. \tag{2}$$

Wenn die Maschine mit L Gleitkomma-Dualziffern rechnet, ist speziell $\varepsilon = (1 + c) \cdot 2^{-L}$; dabei ist c eine nichtnegative Maschinenkonstante (bei optimaler Rundung ist $c = 0$).

Auf Grund dieser Formel erhält WILKINSON für Verfahren I folgendes Ergebnis ($\tilde{s}$ ist der berechnete Wert für s):

$$\tilde{s} = \sum_{m=1}^{n} a_m (1 + \eta_m) = s + \sum_{m=1}^{n} a_m \eta_m \tag{I,2}$$

$$1 + \eta_m = \prod_{i=m}^{n} (1 + \varepsilon_i)\,, \qquad |\varepsilon_i| \leqq \varepsilon \quad \text{für} \quad i > 1\,, \quad \varepsilon_1 = 0\,, \tag{I,3}$$

wobei die Größen η_m durch

$$\begin{aligned} &|\eta_{n-t+1}| \leqq (1 + \varepsilon)^t - 1 && (1 \leqq t \leqq n - 1)\\ &|\eta_1| \leqq (1 + \varepsilon)^{n-1} - 1 \end{aligned} \tag{I,4}$$

beschränkt sind; diese Schranken werden für $\varepsilon_i = \varepsilon$ (für $i > 1$) angenommen.

Für Verfahren II gilt: ist $\tilde{s}_{mk}$ der berechnete Wert für s_{mk}, so ist nach (2) und (II,1):

$$\begin{aligned} &\tilde{s}_{mk} = (\tilde{s}_{2m-1,k-1} + \tilde{s}_{2m,k-1})(1 + \varepsilon_{mk}) && \text{für} \quad 1 \leqq k \leqq t\,,\\ &0 \leqq m \leqq 2^{t-k} - 1\,, \qquad |\varepsilon_{mk}| \leqq \varepsilon\,. \end{aligned}$$

Durch vollständige Induktion erhält man daraus leicht das Ergebnis:

$$\tilde{s}_{mk} = \sum_{i=2^k(m-1)+1}^{2^k m} s_{i0}\,(1 + \delta_i^{(k)})\,, \qquad 1 + \delta_i^{(k)} = \prod_{L=1}^{k} \left(1 + \varepsilon_{1+\left[\frac{i-1}{2^L}\right], L}\right).$$

Für den berechneten Wert $\tilde{s} = \tilde{s}_{1t}$ der Summe s gilt also (unter Berücksichtigung von $s_{i0} = 0$ für $i > n$):

$$\tilde{s} = \sum_{m=1}^{n} a_m (1 + \eta_m) = s + \sum_{m=1}^{n} a_m \eta_m \tag{II,2}$$

$$1 + \eta_m = \prod_{i=1}^{t} \left(1 + \varepsilon_{1+\left[\frac{m-1}{2^i}\right], i}\right), \qquad |\varepsilon_{ik}| \leqq \varepsilon\,, \tag{II,3}$$

woraus man sofort die Schranke

$$|\eta_i| \leqq (1 + \varepsilon)^t - 1 \qquad (1 \leqq i \leqq n) \tag{II,4}$$

erhält, die für $\varepsilon_{ik} = \varepsilon$ angenommen wird.

4. Rundungsfehler beim Kahan-Babuška-Verfahren (III)

Die berechneten Näherungswerte für s_m, w_m, s seien bzw. $\tilde{s}_m, \tilde{w}_m, \tilde{s}$. Aus (III,1) folgt jetzt mit (2):

$$\begin{aligned} \tilde{s}_0 &= \tilde{w}_0 = 0 \\ \tilde{s}_m &= (\tilde{s}_{m-1} + a_m)(1+\varepsilon_m) && m = 1, \ldots, n \\ \tilde{w}_m &= (1+\alpha_m)\left(\tilde{w}_{m-1} + (1+\beta_m)\left(a_m + (1+\gamma_m)(\tilde{s}_{m-1} - \tilde{s}_m)\right)\right) && m = 1, \ldots, n \end{aligned} \tag{III,2}$$

mit $\alpha_m, \beta_m, \gamma_m, \varepsilon_m$, die betragsmäßig $\leqq \varepsilon$ sind. Da man annehmen kann, daß $z + 0 = z$ exakt berechnet wird, kann man voraussetzen:

$$\alpha_1 = \beta_1 = \gamma_1 = \varepsilon_1 = 0\,. \tag{III,2a}$$

Nun ist $q_m = \sum_{r=1}^{m} \left(\psi_r \prod_{i=r}^{m} \varphi_i\right)$ die Lösung der Differenzengleichung $q_m = \varphi_m (q_{m-1} + \psi_m)$ mit der Anfangsbedingung $q_0 = 0$, wie man durch Einsetzen feststellt.

Mit den Abkürzungen

$$\prod_{i=r}^{s}(1+\varepsilon_i) = b_{rs}\,, \qquad (1+\beta_r)\prod_{i=r}^{n}(1+\alpha_i) = c_r \tag{III,3}$$

ist daher

$$\tilde{s}_m = \sum_{r=1}^{m} a_r\, b_{rm} \tag{III,4}$$

$$\tilde{w}_n = \sum_{r=1}^{n} c_1 \left(a_r + (1+\gamma_r)(\tilde{s}_{r-1} - \tilde{s}_r)\right) \tag{III,5}$$

oder, unter Verwendung von (III,4):

$$\tilde{w}_n = \sum_{r=1}^{n} a_r \left\{ c_r \left(1 - (1+\gamma_r)(1+\varepsilon_r)\right) + \sum_{s=r+1}^{n} c_s (1+\gamma_s)(b_{r,s-1} - b_{rs}) \right\}. \tag{III,6}$$

Also ist

$$\tilde{s}_n + \tilde{w}_n = \sum_{r=1}^{n} a_r (1+\delta_r) \tag{III,7}$$

mit

$$\delta_r = 1 + b_{rn} - c_r\,\gamma_r\,(1+\varepsilon_r) - c_r\,\varepsilon_r + \sum_{s=r+1}^{n} c_s\,(b_{r,s-1} - b_{rs})\,(1+\gamma_s)\,. \tag{III,8}$$

Durch Umformen erhält man

$$\delta_r = -\varepsilon_r (c_r - 1) - c_r\,\gamma_r\,(1+\varepsilon_r) + \sum_{s=r+1}^{n} (b_{r,s-1} - b_{rs})\left(c_s(\gamma_s + 1) - 1\right)$$

(wegen $b_{rn} = b_{rr} - \sum_{s=r+1}^{n} (b_{r,s-1} - b_{rs})$ und $b_{rr} = 1 + \varepsilon_r$) und weiter

$$-\delta_r = \varepsilon_r (c_r - 1) + c_r\,\gamma_r\,(1+\varepsilon_r) + \sum_{s=r+1}^{n} \varepsilon_s\, b_{r,s-1}\,(c_s\,\gamma_s + c_s - 1)$$

$\left(\text{wegen } b_{rs} = (1+\varepsilon_s)\, b_{r,s-1}\right)$, also

$$-\delta_r = \sum_{s=r}^{n} \varepsilon_s\, b_{r,s-1}\,(c_s\,\gamma_s + c_s - 1) + \gamma_r\, c_r\,. \tag{III,9}$$

Aus (III,3) folgt für $r > 1$:

$$\begin{aligned} |b_{rs}| &\leqq (1+\varepsilon)^{s-r+1}\,, \qquad |c_s| \leqq (1+\varepsilon)^{n-s+2}\,, \\ |c_s - 1| &\leqq (1+\varepsilon)^{n-s+2} - 1 \end{aligned} \tag{III,10a}$$

und für $r = 1$ wegen (III,2a):

$$|b_{1s}| \leqq (1+\varepsilon)^{s-1} \tag{III,10b}$$

Man erhält damit aus (III,9) die Abschätzung

$$\begin{aligned} |\delta_r| &\leqq (1+\varepsilon)^{n-r+2}\,\varepsilon + \sum_{s=r}^{n} \varepsilon\,(1+\varepsilon)^{s-r}\left(\varepsilon\,(1+\varepsilon)^{n-s+2} + (1+\varepsilon)^{n-s+2} - 1\right) \\ &= \varepsilon\,(1+\varepsilon)^{n-r+2} + (n-r+1)\,\varepsilon\,(1+\varepsilon)^{n-r+3} - (1+\varepsilon)^{n-r+1} + 1 \end{aligned}$$

für $r > 1$, also

$$|\delta_{n-t+1}| \leqq \left(\varepsilon\,(1+\varepsilon)^2\,t + \varepsilon\,(1+\varepsilon) - 1\right)(1+\varepsilon)^t + 1 \qquad (1 \leqq t \leqq n-1) \tag{III,11a}$$

und

$$\begin{aligned} |\delta_1| &\leqq \sum_{s=2}^{n} \varepsilon\,(1+\varepsilon)^{s-2}\left(\varepsilon\,(1+\varepsilon)^{n-s+2} + (1+\varepsilon)^{n-s+2} - 1\right) \\ &= (n-1)\,\varepsilon\,(1+\varepsilon)^{n+1} - (1+\varepsilon)^{n-1} + 1\,, \end{aligned}$$

also

$$|\delta_1| \leqq \left(\varepsilon\,(1+\varepsilon)^2\,(n-1) - 1\right)(1+\varepsilon)^{n-1} + 1\,. \tag{III,11b}$$

Die Schranken in (III,11) werden für $\alpha_i = \beta_i = \gamma_i = \varepsilon_i = \varepsilon$ $(i > 1)$ erreicht, können also i. a. nicht mehr verbessert werden.

Nun ist

$$\tilde{s} = (\tilde{s}_n + \tilde{w}_n)\,(1+\delta) = \left(s + \sum_{i=1}^{n} \alpha_i\,\delta_i\right)(1+\delta)\,,$$

also

$$\tilde{s} = s + \delta s + \sum_{i=1}^{n} a_i\,\eta_i \tag{III,12}$$

mit

$$\eta_i = -(\delta+1)\left(\gamma_r\,c_r + \sum_{s=r}^{n} \varepsilon_s\,b_{r,s-1}\left(c_s(\gamma_s+1) - 1\right)\right) \tag{III,13}$$

$$\begin{aligned} &|\eta_{n-t+1}| \leqq \left(\varepsilon\,(1+\varepsilon)^2\,t + \varepsilon\,(1+\varepsilon) - 1\right)(1+\varepsilon)^{t+1} + 1 + \varepsilon \qquad (1 \leqq t \leqq n-1) \\ &|\eta_1| \leqq \left(\varepsilon\,(1+\varepsilon)^2\,(n-1) - 1\right)(1+\varepsilon)^n + 1 + \varepsilon \\ &|\delta| \leqq \varepsilon\,. \end{aligned} \tag{III,14}$$

5. Forderungen an eine Rechenmaschine. Ein Satz über Rundungsfehler

Zunächst werden einige Forderungen an die zu verwendende Rechenmaschine gestellt, die in der Praxis mehr oder weniger genau erfüllt sind. Zur Vorbereitung der Abschätzung für die Rundungsfehler beim Verfahren IV wird dann ein Satz über die Berechnung von $B - (A + B)$ für $|A| \leqq |B|$ in einer solchen Rechenmaschine bewiesen.

Über die Struktur der Maschinenzahlen wird folgende Annahme gemacht:

Voraussetzung 1: Maschinenzahlen sind genau die Zahl 0 und die Zahlen $x = \pm\, q \cdot Z^{-\alpha}$ mit $Z^{L-1} \leqq q < Z^L$, $q \in \mathbb{N}$, $\alpha \in \mathbb{G}$ *); dabei sind $Z > 1$ und L positive, ganzzahlige Maschinenkonstanten. In der Praxis können natürlich nur endlich viele α zugelassen werden. Alles Folgende bleibt jedoch richtig, wenn alle auftretenden Summen oder Differenzen eines dieser α als Exponent besitzen.

Hilfssatz 1: *Gilt Vor.* 1, *sind A und B Maschinenzahlen, gilt $A \geqq B \geqq 0$ und ist*

$$B = q\,Z^{-\alpha} \quad (q \in \mathbb{N}\,, \quad Z^{L-1} \leqq q < Z^L\,, \quad \alpha \in \mathbb{G})\,,$$

so ist $A = r\,Z^{-\alpha}$ mit einem $r \in \mathbb{N}$, $r \geqq q$.

Beweis: Es ist $A \geqq B \geqq Z^{L-\alpha-1}$. Da A eine Maschinenzahl ist und da $A > 0$, ist $A = r_0\,Z^{-\beta}$ mit $r_0 \in \mathbb{N}$, $Z^{L-1} \leqq r_0 < Z^L$. Also ist $Z^{L-\beta} > r_0\,Z^{-\beta} = A \geqq Z^{L-\alpha-1}$ oder $L - \beta > L - \alpha - 1$. Das bedeutet aber, daß $-\beta = \gamma - \alpha$ mit $\gamma > 0$. Mit $r = r_0\,Z^{\gamma} (\in \mathbb{N})$ ist dann $A = r\,Z^{-\alpha}$. Aus $A \geqq B$ folgt schließlich $r \geqq q$, w. z. z. w.

Hilfssatz 2: *Gilt Vor.* 1, *ist A Maschinenzahl, $Z^{L-\alpha} > A \geqq Z^{L-1-\alpha}$, $\alpha \in \mathbb{G}$, so ist*

$$A = r\,Z^{-\alpha}\,, \qquad r \in \mathbb{N}\,, \qquad Z^{L-1} \leqq r < Z^L\,.$$

Beweis: Wegen $Z^{L-1-\alpha} = q\,Z^{-\alpha}$ mit $q = Z^{L-1}$ folgt aus Hilfssatz 1: $A = r\,Z^{-\alpha}$ mit $r \in \mathbb{N}$, $Z^{L-1} \leqq r$. Wegen $A < Z^{L-\alpha}$ gilt dann $r < Z^L$, w. z. z. w.

Für die Maschinenoperationen $\tilde{+}$ (Näherung für $+$) und $\tilde{-}$ (Näherung für $-$) seien folgende Voraussetzungen erfüllt:

Voraussetzung 2: Sind A und B Maschinenzahlen, gilt

(V 0) $\qquad A \geqq B \geqq 0$

und existiert ein $\alpha \in \mathbb{G}$ mit

(V 1) $\qquad A\,Z^{\alpha} \in \mathbb{G}\,, \qquad B\,Z^{\alpha} \in \mathbb{G}$

*) $\mathbb{G}$: Menge der ganzen Zahlen.

(V 2) $B < Z^{L-\alpha}$

(V 3) $A - B \leqq Z^{L-\alpha}$

so ist $A \tilde{-} B = A - B$.

Voraussetzung 3: Sind A und B Maschinenzahlen, gilt

$$A \geqq B \geqq 0\,, \qquad B \leqq Z^{L-\alpha}\,,$$

so ist

(U 1) $A \leqq A \tilde{+} B \leqq A + Z^{L-\alpha}$

(U 2) $A - Z^{L-\alpha} \leqq A \tilde{-} B \leqq A$

(U 3) $0 \leqq A \tilde{-} B$.

Bemerkung: Bei optimaler Rundung[2]) sind unter Annahme von Vor. 1 die Voraussetzungen 2 und 3 erfüllt.

Satz 1: *Es mögen Vor.* 1—3 *erfüllt sein.* X, Y *seien Maschinenzahlen mit* $X \geqq Y > 0$. *Dann gilt*

a) $(X \tilde{+} Y) \tilde{-} X = (X \tilde{+} Y) - X$

b) $X \tilde{-} (X \tilde{-} Y) = X - (X \tilde{-} Y)$.

Beweis: Nach Voraussetzung 1 ist

$$X = q\,Z^{-\gamma}\,, \qquad Z^{L-1} \leqq q < Z^{L}\,, \qquad (q \in \mathbb{N},\ \gamma \in G)\,, \tag{3}$$

also gilt

$$Y < Z^{L-\gamma}\,, \qquad Z^{L-\gamma-1} \leqq X < Z^{L-\gamma}\,. \tag{4}$$

a) Setze $A := X \tilde{+} Y$, $B := X$, $\alpha := \gamma$.

Nach (U 1) ist $A \geqq X$. Daher ist nach Hilfssatz 1:

$$A = r\,Z^{-\alpha}\,, \qquad r \geqq q\,, \qquad (r \in \mathbb{N})$$

$A \geqq X = B \geqq 0$ liefert (V 0), $A\,Z^{\alpha} = r \in G$, $B\,Z^{\alpha} = q \in G$ liefert (V 1), $B = X < Z^{L-\alpha}$ liefert (V 2).

Wegen $Y < Z^{L-\alpha}$ ist nach Vor. 3 (U 1):

$$A - B = (X \tilde{+} Y) - X \leqq (X + Z^{L-\alpha}) - X = Z^{L-\alpha}\,.$$

Also gilt auch (V 3) und damit ist nach Vor. 2: $(X \tilde{+} Y) \tilde{-} X = (X \tilde{+} Y) - X$.

b) Setze $A := X$, $B := X \tilde{-} Y$.

Nach (U 2), (U 3) gilt

$$0 \leqq B \leqq A = X < Z^{L-\gamma} \tag{5}$$

(V 0) ist daher erfüllt.

Es werden folgende Fälle unterschieden:

Fall α: $Z^{L-1-\gamma} \leqq Y < Z^{L-\gamma}$

Fall β: $Z^{L-1-\gamma} \leqq B < Z^{L-\gamma}\,, \qquad Y < Z^{L-1-\gamma}$

Fall γ: $Z^{L-2-\gamma} \leqq B < Z^{L-1-\gamma}\,, \qquad Y < Z^{L-1-\gamma}$

Fall δ: $Z^{L-2-\gamma} \leqq Y < Z^{L-1-\gamma}\,, \qquad B < Z^{L-2-\gamma}\,.$

Die Möglichkeit $Y < Z^{L-2-\gamma}$, $B < Z^{L-2-\gamma}$ kann nicht eintreten, da daraus (im Widerspruch zu (4)) nach (U 2): $X - Z^{L-2-\gamma} \leqq B < Z^{L-2-\gamma}$, also $X < 2\,Z^{L-2-\gamma} \leqq Z^{L-1-\gamma}$ folgen würde.

Fall α: Setze $\alpha := \gamma$. Dann ist $Z^{L-1-\gamma} \leqq Y < Z^{L-\alpha}$, nach Hilfssatz 2 also

$$Y = r\,Z^{-\alpha}\,, \qquad Z^{L-1} \leqq r < Z^{L}\,, \qquad (r \in \mathbb{N})\,. \tag{6}$$

Nach Voraussetzung ist $X \geqq Y > 0$. Weiter ist $X\,Z^{\alpha} = q \in G$, $Y\,Z^{\alpha} = r \in \mathbb{N}$, $Y < Z^{L-\alpha}$, $X - Y \leqq X < Z^{L-\alpha}$. Vor. 2 läßt sich daher auf X und Y anwenden und liefert

$$B = X \tilde{-} Y = X - Y = (q - r)\,Z^{-\alpha}\,.$$

Nun folgt (V 1) aus $A\,Z^{\alpha} = q \in G$, $B\,Z^{\alpha} = q - r \in G$, (V 2) aus $B \leqq A = X < Z^{L-\alpha}$, (V 3) aus

$$A - B = q\,Z^{-\alpha} - (q - r)\,Z^{-\alpha} = r\,Z^{-\alpha} \leqq Z^{L-\alpha} \quad \text{nach (6)}\,.$$

Fall β: Setze $\alpha := \gamma$. Dann ist $Z^{L-1-\alpha} \leqq B < Z^{L-\alpha}$, nach Hilfssatz 2 also

$$B = r\,Z^{-\alpha}\,, \qquad Z^{L-1} \leqq r < Z^{L}\,, \qquad (r \in \mathbb{N})\,.$$

---

[2]) $x \tilde{\pm} y$ = nächste bei $x \pm y$ gelegene Maschinenzahl; liegt $x \pm y$ genau in der Mitte zwischen den beiden nächsten Maschinenzahlen, so ist entweder stets $x \tilde{\pm} y =$ die betragsmäßig größere der beiden Maschinenzahlen oder stets $x \tilde{\pm} y =$ die betragsmäßig kleinere der beiden Maschinenzahlen.

(V 1) folgt aus $A\,Z^\alpha = q \in \mathbb{G}$, $B\,Z^\alpha = r \in \mathbb{G}$, (V 2) aus $B \leqq A = X < Z^{L-\alpha}$, (V 3) aus $A - B \leqq A = X < Z^{L-\alpha}$.

Fall γ: Setze $\alpha := \gamma + 1$. Dann ist $Z^{L-1-\alpha} \leqq B < Z^{L-\alpha}$, nach Hilfssatz 2 also

$$B = r\,Z^{-\alpha}, \qquad Z^{L-1} \leqq r < Z^L, \qquad (r \in \mathbb{N}). \tag{7}$$

(V 1) folgt aus $A\,Z^\alpha = q\,Z \in \mathbb{G}$, $B\,Z^\alpha = r \in \mathbb{G}$, (V 2) aus (7), (V 3) aus

$$A - B = X - (X \tilde{-} Y) \leqq X - (X - Z^{L-\alpha}) = Z^{L-\alpha}$$

wegen (U 2) und $Y \leqq Z^{L-\alpha}$.

Fall δ: Setze $\alpha := \gamma + 1$. Dann ist $Z^{L-1-\alpha} \leqq Y \leqq Z^{L-\alpha}$, nach Hilfssatz 2 also

$$Y = r\,Z^{-\alpha}, \qquad Z^{L-1} \leqq r \leqq Z^L, \qquad (r \in \mathbb{N}) \tag{8}$$

Nach (U 2) gilt: $X - Z^{L-\alpha} \leqq B < Z^{L-\alpha-1}$, also

$$X - Y \leqq Z^{L-\alpha} + Z^{L-\alpha-1} - Z^{L-\alpha-1} = Z^{L-\alpha}.$$

Wegen $X\,Z^\alpha = q\,Z \in \mathbb{G}$, $Y\,Z^\alpha = r \in \mathbb{G}$, $Y < Z^{L-\alpha}$ läßt sich Vor. 2 anwenden und man erhält:

$$B = X \tilde{-} Y = X - Y = (Z\,q - r)\,Z^{-\alpha}. \tag{9}$$

(V 1) folgt jetzt aus $A\,Z^\alpha = Z\,q \in \mathbb{G}$, $B\,Z^\alpha = Z\,q - r \in \mathbb{G}$, (V 2) aus (8), (9), (V 3) aus $A - B = X - (X - Y) = Y < Z^{L-\alpha}$. Also sind in allen vier Fällen die Bedingungen (V 1), (V 2), (V 3) erfüllt. Wegen (5) ist daher Vor. 2 anwendbar und liefert $X \tilde{-} (X \tilde{-} Y) = X - (X \tilde{-} Y)$, w. z. z. w.

Um aus diesem Satz Aussagen über $B \tilde{-} (A \tilde{+} B)$ zu gewinnen, werden folgende weitere Voraussetzungen getroffen (die bei optimaler Rundung wieder erfüllt sind):

Voraussetzung 4: Für alle Maschinenzahlen A gilt:

(a) $A \tilde{+} 0 = 0 \tilde{+} A = A$

(b) $A \tilde{-} A = 0$

(c) $0 \tilde{-} A = -A$.

Voraussetzung 5: Für alle positiven Maschinenzahlen X, Y gilt:

(d) $Y \tilde{+} X = X \tilde{+} Y$

(e) $Y \tilde{-} X = -(X \tilde{-} Y)$

(f) $(-X) \tilde{+} (-Y) = -(Y \tilde{+} X)$

(g) $(-X) \tilde{-} (-Y) = -(X \tilde{-} Y)$

(h) $X \tilde{+} (-Y) = X \tilde{-} Y$.

Satz 2: *Unter Vor. 1—5 gilt:*

Sind A, B Maschinenzahlen mit $|A| \leqq |B|$, so ist

$$B \tilde{-} (A \tilde{+} B) = B - (A \tilde{+} B).$$

Beweis: Ist $A = 0$, so ist $B \tilde{-} (A \tilde{+} B) = B \tilde{-} B = 0 = B - B = B - (A \tilde{+} B)$ nach Vor. 4.
Ist $B = 0$, so ist $B \tilde{-} (A \tilde{+} B) = B \tilde{-} A = 0 \tilde{-} A = -A = 0 - A = B - A = B - (A \tilde{+} B)$ nach Vor. 4.
Ist $A > 0$, $B > 0$, so ist mit $A = Y$, $B = X$:

$$B \tilde{-} (A \tilde{+} B) \overset{(e)}{=} -((Y \tilde{+} X) - X) \overset{(d)}{=} -((X \tilde{+} Y) \tilde{-} X) \overset{\text{Satz 1}}{=} -((X \tilde{+} Y) - X)$$

$$= X - (X \tilde{+} Y) \overset{(d)}{=} B - (A \tilde{+} B).$$

Ist $A > 0$, $B < 0$, so ist mit $A = Y$, $B = -X$:

$$B \tilde{-} (A \tilde{+} B) \overset{(h)}{=} (-X) \tilde{-} (Y \tilde{-} X) \overset{(e)}{=} (-X) \tilde{-} (-(X \tilde{-} Y)) \overset{(g)}{=} -(X \tilde{-} (X \tilde{-} Y))$$

$$\overset{\text{Satz 1}}{=} -(X - (X \tilde{-} Y)) \overset{(e)}{=} (-X) - (Y \tilde{-} X) \overset{(h)}{=} (-X) - (Y \tilde{+} (-X)) = B - (A \tilde{+} B).$$

Ist $A < 0$, $B > 0$, so ist mit $A = -Y$, $B = X$:

$$B \tilde{-} (A \tilde{+} B) \overset{(d)}{=} X \tilde{-} (X \tilde{+} (-Y)) \overset{(h)}{=} X \tilde{-} (X \tilde{-} Y) \overset{\text{Satz 1}}{=} X - (X \tilde{-} Y) \overset{(h)}{=} X - (X \tilde{+} (-Y)) \overset{(d)}{=} B - (A \tilde{+} B).$$

Ist schließlich $A < 0$, $B < 0$, so ist mit $A = -Y$, $B = -X$:

$$B \tilde{-} (A \tilde{+} B) \overset{(f)}{=} (-X) \tilde{-} (-(X \tilde{+} Y)) \overset{(g)}{=} -(X \tilde{-} (X \tilde{+} Y)) \overset{(e)}{=} (X \tilde{+} Y) \tilde{-} X \overset{\text{Satz 1}}{=} (X \tilde{+} Y) - X$$

$$= -X - (-(X \tilde{+} Y)) \overset{(f)}{=} B - (A \tilde{+} B).$$

Damit ist Satz 2 bewiesen.

6. Das verbesserte Kahan-Babuška-Verfahren (IV)

Um einfache Abschätzungen zu gewinnen, sei angenommen, daß die verwendete Maschine die Voraussetzungen 1 bis 5 erfüllt (also z. B. optimale Rundung besitzt). Für andere Maschinen kann man die erhaltene Abschätzung als näherungsweise gültig ansehen. Die Bezeichnung der in der Maschine berechneten Werte sei wie in Nummer 4. Dann ist für $|a_m| \leqq |\tilde{s}_{m-1}|$ unter Verwendung von (2):

$$a_m \tilde{+} (\tilde{s}_{m-1} \tilde{-} \tilde{s}_m) = a_m \tilde{+} (\tilde{s}_{m-1} \tilde{-} (a_m \tilde{+} \tilde{s}_{m-1})) \overset{\text{Satz 2}}{=} a_m \tilde{+} (\tilde{s}_{m-1} - (a_m \tilde{+} \tilde{s}_{m-1}))$$
$$= a_m \tilde{+} (\tilde{s}_{m-1} - \tilde{s}_m) = (1 + \beta_m)(a_m + \tilde{s}_{m-1} - \tilde{s}_m)$$

und für $|a_m| > |\tilde{s}_{m-1}|$:

$$\tilde{s}_{m-1} \tilde{+} (a_m \tilde{-} \tilde{s}_m) = \tilde{s}_{m-1} \tilde{+} (a_m \tilde{-} (\tilde{s}_{m-1} \tilde{+} a_m)) \overset{\text{Satz 2}}{=} \tilde{s}_{m-1} \tilde{+} (a_m - (\tilde{s}_{m-1} \tilde{+} a_m))$$
$$= \tilde{s}_{m-1} \tilde{+} (a_m - \tilde{s}_m) = (1 + \beta_m)(\tilde{s}_{m-1} + a_m - \tilde{s}_m) .$$

Daher ist in beiden Fällen:

$$\tilde{s}_m = (1 + \varepsilon_m)(\tilde{s}_{m-1} + a_m) \tag{IV,2}$$
$$\tilde{w}_m = (1 + \alpha_m)(\tilde{w}_{m-1} + (1 + \beta_m)(a_m + \tilde{s}_{m-1} - \tilde{s}_m)) . \tag{IV,3}$$

Das sind aber gerade die Formeln (III,2) für $\gamma_m = 0$ $(m \geqq 1)$. (Man kann also eine Verbesserung gegenüber Verfahren III erwarten. Dies ist der Grund für die Einführung dieses Verfahrens.) Daher folgen aus (IV,2, 3) die folgenden Formeln ((III,7, 9, 10) für $\gamma_m = 0$):

$$s' := \tilde{s}_n + \tilde{w}_n = \sum_{r=1}^{n} a_r (1 + \delta_r) \tag{IV,4}$$

$$\delta_r = -\sum_{s=r}^{n} \varepsilon_s \, b_{r,s-1} (c_s - 1) \tag{IV,5}$$

$$|b_{rs}| \leqq (1 + \varepsilon)^{s-r+1} , \qquad |c_s - 1| \leqq (1 + \varepsilon)^{n-s+2} - 1 , \qquad |b_{1s}| \leqq (1 + \varepsilon)^{s-1} . \tag{IV,6}$$

Daraus erhält man für $r > 1$:

$$|\delta_r| \leqq \sum_{s=r}^{n} \varepsilon (1 + \varepsilon)^{s-r} \left((1 + \varepsilon)^{n+2-s} - 1\right) = \varepsilon (1 + \varepsilon)^{n-r+2}(n - r + 1) - (1 + \varepsilon)^{n-r+1} + 1 ,$$

also

$$|\delta_{n-t+1}| \leqq (\varepsilon (1 + \varepsilon)\, t - 1)(1 + \varepsilon)^t + 1 \qquad (1 \leqq t \leqq n - 1) \tag{IV,7a}$$

und für $r = 1$:

$$|\delta_1| \leqq \sum_{s=2}^{n} \varepsilon (1 + \varepsilon)^{s-2} \left((1 + \varepsilon)^{n-s+2} - 1\right) = \varepsilon (n - 1)(1 + \varepsilon)^n - (1 + \varepsilon)^{n-1} + 1 ,$$

also

$$|\delta_1| \leqq \left(\varepsilon (1 + \varepsilon)(n - 1) - 1\right)(1 + \varepsilon)^{n-1} + 1 . \tag{IV,7b}$$

Nun ist $\tilde{s} = (\tilde{s}_n + \tilde{w}_n)(1 + \delta) = s'(1 + \delta) = \left(s + \sum_{i=1}^{n} a_i \, \delta_i\right)(1 + \delta)$, also ist

$$\tilde{s} = s + \delta s + \sum_{i=1}^{n} a_i \, \eta_i \tag{IV,8}$$

$$\eta_r = -(1 + \delta) \sum_{s=r}^{n} \varepsilon_s \, b_{r,s-1} (c_s - 1) \tag{IV,9}$$

$$|\eta_{n-t+1}| \leqq (\varepsilon (1 + \varepsilon)\, t - 1)(1 + \varepsilon)^{t+1} + 1 + \varepsilon \qquad (1 \leqq t \leqq n - 1)$$
$$|\eta_1| \leqq \left(\varepsilon (1 + \varepsilon)(n - 1) - 1\right)(1 + \varepsilon)^n + 1 + \varepsilon \tag{IV,10}$$
$$|\delta| \leqq \varepsilon .$$

Wie man aus (IV,9) sieht, werden auch die Schranken in (IV,10) angenommen, und zwar für $\alpha_m = \beta_m = \varepsilon_m = \varepsilon$ $(m > 1)$.

7. Vereinfachung der Abschätzungen für kleine ε

Es ist

$$\varepsilon - \frac{\varepsilon^2}{2} \leqq r \leqq \varepsilon \qquad \text{für} \qquad r = \ln (1 + \varepsilon) . \tag{10}$$

Daher ist für Verfahren I nach (I,4):

$$|\eta_{n-t+2}| \leqq (1+\varepsilon)^{t-1} - 1 = \varepsilon\,(t-1) + r^2\,t^2\,\frac{\mathrm{e}^{rt} - r\,t - 1}{(r\,t)^2} - t\,(\varepsilon - r) - \varepsilon\left((1+\varepsilon)^{t-1} - 1\right)$$

$$\leqq \varepsilon\,(t-1) + \varepsilon^2\,t^2\,\frac{\mathrm{e}^k - k - 1}{k^2} \qquad \text{für} \qquad \varepsilon\,t \leqq k\,,$$

ebenso

$$|\eta_1| \leqq (1+\varepsilon)^{n-1} - 1 \leqq \varepsilon\,(n-1) + \varepsilon^2\,n^2\,\frac{\mathrm{e}^k - k - 1}{k^2} \qquad \text{für} \qquad \varepsilon\,n \leqq k\,,$$

also

$$|\eta_i| \leqq \varepsilon\,(t-1) + \varepsilon^2\,t^2\,\frac{\mathrm{e}^k - k - 1}{k^2} \qquad (\varepsilon\,n \leqq k) \tag{I,5}$$

$$t = n - i + 2 \quad \text{für} \quad i > 1\,, \qquad t = n \quad \text{für} \quad i = 1\,.$$

Für Verfahren II gilt (mit $u = t - 1$):

$$|\eta_i| \leqq (1+\varepsilon)^{u+1} - 1 = \varepsilon\,(u+1) + r^2\,u^2\,\frac{\mathrm{e}^{ru} - r\,u - 1}{(r\,u)^2} + \varepsilon\,r\,u\,\frac{\mathrm{e}^{ru} - 1}{r\,u} - u\,(\varepsilon - r)\,,$$

$$|\eta_i| \leqq \varepsilon\,t + \varepsilon^2\left((t-1)^2\,\frac{\mathrm{e}^k - k - 1}{k^2} + (t-1)\,\frac{\mathrm{e}^k - 1}{k}\right) \qquad (\varepsilon\,(t-1) \leqq k) \tag{II,5}$$

$$t \leqq \log_2 n + 1\,.$$

Für Verfahren III ist nach (III,14):

$$|\eta_{n-t+2}| \leqq \left(\varepsilon\,(1+\varepsilon)^2\,t - \varepsilon^2\,(1+\varepsilon) - 1\right)(1+\varepsilon)^t + 1 + \varepsilon$$

$$= \varepsilon + r^2\,t^2\,\frac{\mathrm{e}^{rt}\,(r\,t-1)+1}{(r\,t)^2} + \varepsilon\,r\,t\,\frac{5}{2}\,\mathrm{e}^{rt} + \varepsilon^2\,\mathrm{e}^{rt}\,(\ldots)$$

$$\leqq \varepsilon + \varepsilon^2\left(t^2\,\frac{\mathrm{e}^k\,(k-1)+1}{k^2} + t\,\frac{5}{2}\,\mathrm{e}^k\right) \qquad \left(\varepsilon\,t \leqq k \leqq \frac{4}{9}\,,\ \varepsilon \leqq \frac{1}{2}\right),$$

denn $\dfrac{\mathrm{e}^x\,(x-1)+1}{x^2} = \sum_{i=0}^{\infty} \dfrac{i+1}{(i+2)!}\,x^i$ ist für $x > 0$ monoton wachsend, und aus (10) erhält man durch eine einfache Abschätzung $(\ldots) \leqq 0$ für $\varepsilon\,t \leqq \dfrac{4}{9}$, $\varepsilon \leqq \dfrac{1}{2}$. Aus (III,14) sieht man, daß $|\eta_1|$ durch die Schranke von $|\eta_2|$ beschränkt wird. Daher gilt:

$$|\eta_i| \leqq \varepsilon + \varepsilon^2\left(t^2\,\frac{\mathrm{e}^k\,(k-1)+1}{k^2} + t\,\frac{5}{2}\,\mathrm{e}^k\right)\left(\varepsilon\,t \leqq k \leqq \frac{4}{9}\,,\ \varepsilon \leqq \frac{1}{2}\right) \tag{III,15}$$

$$t = n - i + 2 \quad \text{für} \quad i > 1\,, \qquad t = n \quad \text{für} \quad i = 1\,.$$

Für Verfahren IV erhält man aus (IV,10):

$$|\eta_{n-t+2}| \leqq \left(\varepsilon\,(1+\varepsilon)\,(t-1) - 1\right)(1+\varepsilon)^t + 1 + \varepsilon$$

$$= r^2\,t^2\,\frac{\mathrm{e}^{rt}\,(r\,t-1)+1}{(r\,t)^2} + \varepsilon\,r\,t\,\frac{\mathrm{e}^{rt}\,(3\,r\,t-2)+2}{2\,r\,t} + \varepsilon^2\,\mathrm{e}^{rt}\,(\ldots)$$

$$\leqq \varepsilon^2\left(t^2\,\frac{\mathrm{e}^k\,(k-1)+1}{k^2} + t\,\frac{\mathrm{e}^k\,(3\,k-2)+2}{2\,k}\right)\left(\varepsilon\,t \leqq k \leqq 1\,,\ \varepsilon \leqq \frac{1}{2}\right),$$

da $\dfrac{\mathrm{e}^x\,(3\,x-2)+2}{x^2} = \sum_{i=0}^{\infty} \dfrac{3\,i+1}{(i+1)!}\,x^i$ für $x > 0$ monoton wächst und da $(\ldots) \leqq 0$ für $\varepsilon\,t \leqq 1$, $\varepsilon \leqq \dfrac{1}{2}$.

Wieder wird $|\eta_1|$ durch die Schranke von $|\eta_2|$ beschränkt, also:

$$|\eta_i| \leqq \varepsilon^2\left(t^2\,\frac{\mathrm{e}^k\,(k-1)+1}{k^2} + t\,\frac{\mathrm{e}^k\,(3\,k-2)+2}{2\,k}\right)\left(\varepsilon\,t \leqq k \leqq 1\,,\ \varepsilon \leqq \frac{1}{2}\right), \tag{IV,11}$$

$$t = n - i + 2 \quad \text{für} \quad i > 1\,, \qquad t = n \quad \text{für} \quad i = 1\,.$$

Nun gilt für alle vier Verfahren eine Gleichung der Form

$$\tilde{s} = s + \delta s + \sum_{i=1}^{n} a_i\,\eta_i \tag{11}$$

($|\delta| \leqq \varepsilon$ für Verfahren III, IV; $\delta = 0$ für Verfahren I, II). Daraus lassen sich zwei einfache Formeln für den absoluten Fehler $s - \tilde{s}$ ableiten. Denn aus (11) folgt:

$$|s - \tilde{s}| \leqq |\delta|\,|s| + \sum_{i=1}^{n} |a_i|\,|\eta_i|;$$

also

$$|s - \tilde{s}| \leqq |\delta|\,|s| + \left(\operatorname*{Max}_{1 \leqq i \leqq n} |\eta_i|\right)\left(\sum_{i=1}^{n} |a_i|\right) \tag{12}$$

$$|s - \tilde{s}| \leqq |\delta|\,|s| + \left(\sum_{i=1}^{n} |\eta_i|\right)\left(\operatorname*{Max}_{1 \leqq i \leqq n} |a_i|\right). \tag{13}$$

Formel (12) ist nützlich, wenn alle a_i dasselbe Vorzeichen haben, Formel (13), wenn alle a_i von derselben Größenordnung sind. Auf diesen Formeln baut die Diskussion der Verfahren in diesem und dem nächsten Abschnitt auf.

Im folgenden wird $\varepsilon\, n \leqq k = 1/3$ vorausgesetzt, was in der Praxis wohl meist erfüllt ist. Dann gilt:

$$\frac{e^k - 1}{k} < \frac{6}{5},\quad \frac{e^k (k-1) + 1}{k^2} < \frac{3}{4},\quad \frac{e^k - k - 1}{k^2} < \frac{3}{5},\quad \frac{5}{2} e^k < \frac{7}{2},\quad \frac{e^k (3k-2) + 2}{2k} < 1. \tag{14}$$

Für die einzelnen Verfahren erhält man aus (12), (13), damit folgende Formeln:

Verfahren I: Aus (I,5) erhält man mit (14):

$$\operatorname*{Max}_{1 \leqq i \leqq n} |\eta_i| \leqq \varepsilon (n-1) + \frac{3}{5} \varepsilon^2 n^2$$

$$\sum_{i=1}^{n} |\eta_i| \leqq \varepsilon (n-1) + \frac{3}{5} \varepsilon^2 n^2 + \sum_{t=2}^{n} \left(\varepsilon (t-1) + \frac{3}{5} \varepsilon^2 t^2\right),$$

$$\sum_{i=1}^{n} |\eta_i| \leqq \varepsilon \frac{n^2 + n - 2}{2} + \varepsilon^2 \frac{2n^3 + 9n^2 + n - 6}{10},$$

$$\leqq \varepsilon \frac{n^2 + n - 2}{2} + \varepsilon^2 \frac{n^3 + 5n^2}{5}$$

also

$$|s - \tilde{s}| \leqq \left(\varepsilon (n-1) + \frac{2}{5} \varepsilon^2 n^2\right) \sum_{i=1}^{n} |a_i| \quad \left(\varepsilon n \leqq \frac{1}{3}\right) \tag{I,6}$$

$$|s - \tilde{s}| \leqq \left(\varepsilon \frac{n^2 + n - 2}{2} + \varepsilon^2 \frac{n^3 + 5n^2}{5}\right) \operatorname*{Max}_{1 \leqq i \leqq n} |a_i| \quad \left(\varepsilon n \leqq \frac{1}{3}\right) \tag{I,7}$$

Verfahren II: Aus (II,2), (II,5), (14) folgt unmittelbar:

$$|s - \tilde{s}| \leqq \left(\varepsilon (1 + \log_2 n) + \frac{3}{5} \varepsilon^2 \log_2 n\, (\log_2 n + 2)\right) \sum_{i=1}^{n} |a_i| \quad \left(\varepsilon n \leqq \frac{1}{3}\right), \tag{II,6}$$

$$|s - \tilde{s}| \leqq \left(\varepsilon (1 + \log_2 n) + \frac{3}{5} \varepsilon^2 \log_2 n\, (\log_2 n + 2)\right) n \operatorname*{Max}_{1 \leqq i \leqq n} |a_i| \quad \left(\varepsilon n \leqq \frac{1}{3}\right). \tag{II,7}$$

Verfahren III: Nach (III,15) ist

$$\operatorname*{Max}_{1 \leqq i \leqq n} |\eta_i| \leqq \varepsilon + \varepsilon^2 \left(\frac{3}{4} n^2 + \frac{7}{2} n\right)$$

$$\sum_{i=1}^{n} |\eta_i| \leqq \varepsilon + \varepsilon^2 \left(\frac{3}{4} n^2 + \frac{7}{2} n\right) + \sum_{t=2}^{n} \left(\varepsilon + \varepsilon^2 \left(\frac{3}{4} t^2 + \frac{7}{2} t\right)\right)$$

$$= \varepsilon n + \varepsilon^2 \left(\frac{1}{4} n^3 + \frac{23}{8} n^2 + \frac{43}{8} n - \frac{17}{4}\right) \leqq \varepsilon n + \varepsilon^2 \left(\frac{1}{4} n^3 + 3n^2 + 4n\right),$$

also

$$|s - \tilde{s}| \leqq \varepsilon |s| + \left(\varepsilon + \varepsilon^2 \left(\frac{3}{4} n^2 + \frac{7}{2} n\right)\right) \sum_{i=1}^{n} |a_i| \qquad \left(\varepsilon n \leqq \frac{1}{3}\right) \tag{III,16}$$

$$|s - \tilde{s}| \leqq \varepsilon |s| + \left(\varepsilon n + \varepsilon^2 \left(\frac{1}{4} n^3 + 3n^2 + 4n\right)\right) \operatorname*{Max}_{1 \leqq i \leqq n} |a_i| \qquad \left(\varepsilon n \leqq \frac{1}{3}\right). \tag{III,17}$$

Verfahren IV: Aus (IV,11) folgt

$$\operatorname*{Max}_{1\leqq i\leqq n}|\eta_i| \leqq \varepsilon^2\left(\frac{3}{4}n^2+n\right)$$

$$\sum_{i=1}^{n}|\eta_i| \leqq \varepsilon^2\left(\frac{3}{4}n^2+n\right)+\sum_{t=2}^{n}\varepsilon^2\left(\frac{3}{4}t^2+t\right)=\varepsilon^2\left(\frac{1}{4}n^3+\frac{17}{8}n^2+\frac{25}{8}n-\frac{11}{4}\right)$$

$$\leqq \varepsilon^2\left(\frac{1}{4}n^3+\frac{5}{2}n^2+n\right),$$

also

$$|s-\tilde{s}| \leqq \varepsilon\,|s|+\varepsilon^2\left(\frac{3}{4}n^2+n\right)\sum_{i=1}^{n}|a_i| \qquad \left(\varepsilon\, n\leqq\frac{1}{3}\right) \tag{IV,12}$$

$$|s-\tilde{s}| \leqq \varepsilon\,|s|+\varepsilon^2\left(\frac{1}{4}n^3+\frac{5}{2}n^2+n\right)\operatorname*{Max}_{1\leqq i\leqq n}|a_i| \qquad \left(\varepsilon\, n\leqq\frac{1}{3}\right). \tag{IV,13}$$

Die Ergebnisse ((I,6, 7), (II,6, 7), (III,16, 17), (IV,12, 13)) sind in Tabelle 1 zusammengefaßt.

Bemerkung: Der ε-Anteil der Verfahrensfehler von I, II, III wurde — in etwas anderer Form — schon von BABUŠKA [3] angegeben.

Tabelle 1. Fehlerkoeffizienten der Summierungsverfahren

$$\left\{\begin{aligned}&|s-\tilde{s}| \leqq A\,\varepsilon|s|+(B\,\varepsilon+C\,\varepsilon^2)\sum_{i=1}^{n}|a_i| \qquad (15)\\ &|s-\tilde{s}| \leqq A\,\varepsilon|s|+(D\,\varepsilon+E\,\varepsilon^2)\operatorname*{Max}_{1\leqq i\leqq n}|a_i| \qquad (16)\end{aligned}\right\} \text{ für } \varepsilon\, n\leqq\frac{1}{3}$$

Verf. Nr.	A	$B(n)$	$C(n)$	$D(n)$	$E(n)$
I	0	$n-1$	$0{,}6\,n^2$	$0{,}5\,n^2+0{,}5\,n-1$	$0{,}2\,n^3+n^2$
II	0	$1+\log_2 n$	$0{,}6\,(\log_2 n)^2+1{,}2\log_2 n$	$n\,(1+\log_2 n)$	$0{,}6\,n\log_2 n$
III	1	1	$0{,}75\,n^2+3{,}5\,n$	n	$0{,}25\,n^3+3\,n^2+4\,n$
IV	1	0	$0{,}75\,n^2+n$	0	$0{,}25\,n^3+2{,}5\,n^2+n$

8. Diskussion der Ergebnisse

8.1. Zunächst sollen alle a_i dasselbe Vorzeichen haben. Dann ist mit der Abkürzung

$$s_{\text{abs}} := \sum_{i=1}^{n}|a_i| \tag{17}$$

$s_{\text{abs}} = |s|$ und der relative Fehler ist nach (15)

$$\left|\frac{s-\tilde{s}}{s}\right| \leqq (A+B)\,\varepsilon+C\,\varepsilon^2\,. \tag{18}$$

Solange $n<\varepsilon^{-1/2}$ ist, spielt der Term $C\,\varepsilon^2$ in (18) gegenüber $(A+B)\,\varepsilon$ keine Rolle. Das erste Verfahren hat dann einen maximalen Fehler, der proportional zur Zahl n der Summanden wächst. Im zweiten Verfahren hat man trotz gleichen Rechenaufwands einen maximalen Fehler, der nur mit dem Logarithmus von n wächst (ist z. B. $n=10^6$, so ist der Verfahren-I-Fehler $10^6\,\varepsilon$, der Verfahren-II-Fehler nur $20\,\varepsilon$!!). Wenn daher die Zahl der Summanden von vornherein bekannt ist, empfiehlt es sich, anstelle von Verfahren I stets Verfahren II zu benutzen. (Bei der Summierung einer unendlichen Reihe ist das natürlich nicht möglich.)

Beim KAHAN-BABUŠKA-Verfahren (III) ist der maximale Fehler fast konstant. Mit etwa $2\,\varepsilon$ liegt er in derselben Größenordnung wie der Fehler bei einer Normalverfahrenrechnung mit doppelter Genauigkeit (ε^2 statt ε), bei der das Ergebnis auf einfache Genauigkeit gerundet wird. Da aber der Rechenaufwand für doppelte Genauigkeit nur doppelt so groß ist wie der für einfache, lohnt sich Verfahren II nur dann, wenn auf der Maschine keine Möglichkeit, mit doppelter Genauigkeit zu rechnen, besteht. Dann ist das Verfahren ein guter Ersatz dafür (das gilt nur für $s_{\text{abs}}\approx s$; vgl. unten!). Die Fehlerschranke von Verfahren IV hat ebenfalls die Größenordnung ε. Die geringfügige Verbesserung gegenüber III lohnt den höheren Aufwand zur Berechnung der Summe jedoch nicht.

Ist $\varepsilon^{-1}\leqq n^2\ll\varepsilon^{-1}\log_2(\varepsilon^{-1})$, so wachsen die Schranken in (18) zwar quadratisch in n bzw. $\log_2 n$, aber das Obengesagte bleibt qualitativ richtig. Für $n^2\approx\varepsilon^{-1}\log_2(\varepsilon^{-1})$ sind Verfahren II und III etwa gleich genau; für $n^2\geqq 2\,\varepsilon^{-1}\log_2(\varepsilon^{-1})$ ist schließlich Verfahren II das genaueste Verfahren (genauer als III und IV), da

$$0{,}6\,(\log_2 n)^2+1{,}2\log_2 n\leqq n \qquad \text{für alle } n\,,$$

$$\log_2 n\leqq\frac{3}{4}n^2\,\varepsilon \qquad \text{für} \qquad n^2\geqq 2\,\varepsilon^{-1}\log_2(\varepsilon^{-1})$$

$\left(\frac{\log_2 t}{t}$ ist für $t \geqq 3$ monoton fallend und $\leqq \frac{3}{2}\varepsilon$ für $t = 2\,\varepsilon^{-1}\log_2(\varepsilon^{-1})$, also ist $\log_2 n = \frac{1}{2}n^2\frac{\log_2 n^2}{n^2} \leqq \frac{3}{4}$ für $n \geqq 2\,\varepsilon^{-1}\log_2(\varepsilon^{-1})\right)$.

8.2. Es sei $s_{\text{abs}} \approx |s|$. Dann gilt (18) angenähert. Daher lassen sich alle obigen Aussagen näherungsweise übertragen. Insbesondere ist Verfahren III fast so gut wie Verfahren IV, für kleine n ist Verfahren III (IV) das beste, für große n ist Verfahren II am besten geeignet. Verfahren I ist wieder das schlechteste Verfahren von allen.

8.3. Es sei $s_{\text{abs}} \gg |s|$. Jetzt kann man das Glied $A\,\varepsilon\,|s|$ vernachlässigen und man erhält

$$|s - \tilde{s}| \leqq (B + C\,\varepsilon)\,(s_{\text{abs}}\,\varepsilon)\,. \tag{19}$$

8.3.1. Ist s_{abs} so groß, daß $s_{\text{abs}} \cdot \varepsilon$ mindestens die Größenordnung von $|s|$ hat (etwa wenn s fast Null ist), so ist der Näherungswert $\tilde{s}$ i. a. nur dann brauchbar, wenn $B = 0$ ist, also wenn s nach IV berechnet wird. In den anderen Fällen kann eine so starke Auslöschung auftreten, daß nur noch eine oder keine Ziffer von $\tilde{s}$ mehr richtig ist.

8.3.2. Auch wenn der Extremfall 8.3.1 nicht vorliegt, bestimmt die Größe von B zunehmend die Fehlerschranke. Für Verfahren I ist der maximale relative Fehler jetzt $\gg n\,\varepsilon$, für große n ist das Verfahren unbrauchbar. Verfahren II hat jetzt einen Fehler von ungefähr $\varepsilon_{\text{abs}}\log_2 n\; s_{\text{abs}}/|s|$, der in nicht zu ungünstigen Fällen etwa die Größenordnung $n\,\varepsilon$ haben kann. Dieses Verfahren ist also noch so gut wie Verfahren I im günstigsten Fall, also bedingt verwendbar. Für das KAHAN-BABUŠKA-Verfahren ist die Schranke ungefähr $\varepsilon\, s_{\text{abs}}/|s|$ (für kleine und mittlere n), wird also in der Mehrzahl der Fälle (z. B. Summation absolut konvergenter Reihen mit alternierenden Vorzeichen) noch gute Ergebisse liefern. Das verbesserte KAHAN-BABUŠKA-Verfahren hat jedoch eine Schranke von größenordnungsmäßig $n^2\,\varepsilon^2\,\frac{s_{\text{abs}}}{|s|}$ für den relativen, $n^2\,\varepsilon^2\,s_{\text{abs}}$ für den absoluten Fehler, liefert also für nicht allzu große n (z. B. für $n \ll \varepsilon^{-1/2}$) einen Maximalfehler, der in der Größenordnung von ε liegt! Hier wird also die Auslöschung fast vollständig kompensiert. In diesem Fall wird sich daher der gegenüber I fünffache Rechenaufwand praktisch immer lohnen (wenn man nicht mit doppelter Genauigkeit rechnen kann).

8.4. Wenn s_{abs} nicht sehr groß gegenüber $|s|$ ist, so ist wieder Verfahren II dem ersten Verfahren vorzuziehen. Ob man Verfahren III (oder IV) verwendet, wird von der benötigten Genauigkeit der Summe abhängen. Verfahren III und IV dürften dann etwa gleichwertig sein, der Vorteil größerer Genauigkeit bei IV wird durch den größeren Rechenaufwand aufgewogen.

Das hier Gesagte schließt natürlich nicht aus, daß in Einzelfällen Verfahren I bessere Ergebnisse liefert als Verfahren II oder III; oder das Verfahren I auch einmal für ein kleines n besser ist als Verfahren III; Beispiele dafür werden im nächsten Abschnitt gegeben.

9. Beispiele

Für die ersten drei Beispiele sei eine L-ziffrige Gleitkomma-Dualmaschine mit optimaler Rundung[3]) zugrunde gelegt; eine genaue 1 in der $(L+1)^{\text{ten}}$ Dualstelle soll abgerundet werden. Nach der Bemerkung in Abschnitt 3 ist daher $\varepsilon = 2^{-L}$.

a) Es sei $n = 7$,

$$\begin{aligned} a_1 &= (2^{-1} + 2^{1-L} + 2^{-L})\,2^k\,, \\ a_2 = a_3 = a_4 &= a_5 + (2^{-1})\,2^{k-1-L} \\ a_6 &= (2^{-1})\,2^{k+2} \\ a_7 &= -(2^{-1} + 2^{-3} + 2^{-L})\,2^{k+2}\,. \end{aligned}$$

Es ist $s = 0$. Tabelle 2 gibt die Zwischenwerte für die Summation nach Verfahren I, III und IV wieder, Tabelle 3 die für Verfahren II.

Tabelle 2. Zwischenwerte zu Beispiel a)

m	$\tilde{s}_m$	$(\tilde{w}_m$III)	$\tilde{w}_m$(IV)	Ergebnisse
1	$(2^{-1} + 2^{1-L} + 2^{-L})\,2^k$	0	0	$\tilde{s}_{\text{I}} = 0$
2	$(2^{-1} + 2^{1-L} + 2^{-L})\,2^k$	$(2^{-1})\,2^{k-1-L}$	$(2^{-1})\,2^{k-1-L}$	$\tilde{s}_{\text{III}} = (2^{-1})\,2^{k+1-L}$
3	$(2^{-1} + 2^{1-L} + 2^{-L})\,2^k$	$(2^{-1})\,2^{k-L}$	$(2^{-1})\,2^{k-L}$	$\tilde{s}_{\text{IV}} = 0$
4	$(2^{-1} + 2^{1-L} + 2^{-L})\,2^k$	$(2^{-1} + 2^{-2})\,2^{k-L}$	$(2^{-1} + 2^{-2})\,2^{k-L}$	
5	$(2^{-1} + 2^{1-L} + 2^{-L})\,2^k$	$(2^{-1})\,2^{k+1-L}$	$(2^{-1})\,2^{k+1-L}$	
6	$(2^{-1} + 2^{-3} + 2^{-L})\,2^{k+2}$	$(2^{-1})\,2^{k+1-L}$	0	
7	0	$(2^{-1})\,2^{k+1-L}$	0	

[3]) siehe Fußnote [2]).

Tabelle 3. Zwischenwerte zu Beispiel a)

m	$\tilde{s}_{m\,0}$	$\tilde{s}_{m\,1}$	$\tilde{s}_{m\,2}$	$\tilde{s}_{m\,3}$
1	$(2^{-1}+2^{1-L})\,2^k$	$(2^{-1}+2^{1-L}+2^{-L})\,2^k$	$(2^{-1}+2^{1-L}+2^{-L})\,2^k$	$-(2^{-1})\,2^{k+1}$
2	$(2^{-1})\,2^{k-1-L}$	$(2^{-1})\,2^{k-L}$	$-(2^{-1}+2^{2-L})\,2^k$	
3	$(2^{-1})\,2^{k-1-L}$	$(2^{-1})\,2^{k+2}$		
4	$(2^{-1})\,2^{k-1-L}$	$-(2^{-1}+2^{-3}+2^{-L})\,2^{k+2}$		
5	$(2^{-1})\,2^{k-1-L}$			
6	$(2^{-1})\,2^{k+2}$		Ergebnis: $\tilde{s}_{\mathrm{II}} = -(2^{-1})\,2^{k+1-L}$	
7	$-(2^{-1}+2^{-3}+2^{-L})\,2^{k+2}$			
8	0			

Für die absoluten Fehler erhält man

$$|s-\tilde{s}_{\mathrm{I}}| = |s-\tilde{s}_{\mathrm{IV}}| = 0\,, \qquad |s-\tilde{s}_{\mathrm{II}}| = |s-\tilde{s}_{\mathrm{III}}| = 2^{k-L} = \frac{1}{2}\,\varepsilon \max_{1\leqq i\leqq 7} |a_i|\,.$$

Die in Tabelle 1 angegebenen Schranken überschätzen den wirklichen Fehler bei Verfahren III um das Doppelte, bei Verfahren II um das Achtfache. Der nach I berechnete Wert ist sogar genau. Trotz der wesentlich schlechteren Schranke in Tab. 1 ist also Verfahren I hier genauer als Verfahren II oder III. Verfahren IV liefert das exakte Ergebnis.

b) Es sei $n = 4$, $a_1 = 1$, $a_2 = 2^{-2} + 2^{-1-L}$, $a_3 = -1$, $a_4 = -(2^{-2} + 2^{-1-L})$. Wie man leicht nachprüft, ist jetzt

$$s = 0\,, \qquad \tilde{s}_{\mathrm{II}} = \tilde{s}_{\mathrm{IV}} = 0\,, \qquad \tilde{s}_{\mathrm{I}} = \tilde{s}_{\mathrm{III}} = -2^{-L-1}\,,$$

also Verfahren II genauer als Verfahren I oder III. Verfahren IV gibt wieder das exakte Ergebnis.

c) Sei $n = 3$, $a_1 = 1$, $a_2 = 2^{L+1} + 2^L$, $a_3 = -(2^{L+1} + 2^L)$. Jetzt ist

$$s = 1\,, \qquad \tilde{s}_{\mathrm{I}} = \tilde{s}_{\mathrm{II}} = 0\,, \qquad \tilde{s}_{\mathrm{III}} = \tilde{s}_{\mathrm{IV}} = 1\,,$$

also Verfahren I und II ungenauer als Verfahren III oder IV.

Meistens werden von allen Verfahren ungenaue Ergebnisse geliefert, z. B. stets, wenn s keine Maschinenzahl ist. Ist aber s als Maschinenzahl darstellbar, so ist oft $\tilde{s}_{\mathrm{IV}} = s$; vgl. dazu Beispiel d), Tab. 4, 5. ($n = 7$; die a_m und die Zwischenwerte sind in den Tabellen angegeben.)

Tabelle 4. Zwischenwerte zu Beispiel d)

m	a_m	$\tilde{s}_m$	$\tilde{w}_m$(III)	$\tilde{w}_m$(IV)	Ergebnisse
1	0.5555_{10^0}	0.5555_{10^0}	0	0	$\tilde{s}_{\mathrm{I}} = 0.7780_{10^1}$
2	0.5555_{10^1}	0.6111_{10^1}	$-0.1000_{10^{-2}}$	$-0.5000_{10^{-3}}$	$\tilde{s}_{\mathrm{III}} = 0.7774_{10^1}$
3	0.5555_{10^0}	0.6667_{10^1}	$-0.1500_{10^{-2}}$	$-0.1000_{10^{-3}}$	$\tilde{s}_{\mathrm{IV}} = 0.7777_{10^1}$
4	0.5555_{10^1}	0.7223_{10^1}	$-0.2000_{10^{-3}}$	$-0.1500_{10^{-2}}$	
5	0.5555_{10^2}	0.6277_{10^2}	$-0.2000_{10^{-2}}$	$+0.1500_{10^{-2}}$	
6	0.5555_{10^0}	0.6333_{10^2}	$-0.6500_{10^{-2}}$	$-0.3000_{10^{-2}}$	
7	-0.5555_{10^2}	0.7780_{10^1}	$-0.6500_{10^{-3}}$	$-0.3000_{10^{-2}}$	

Tabelle 5. Zwischenwerte zu Beispiel d)

m	$\tilde{s}_{m\,0}$	$\tilde{s}_{m\,1}$	$\tilde{s}_{m\,2}$	$\tilde{s}_{m\,3}$
1	0.5555_{10^0}	0.6111_{10^1}	0.7222_{10^1}	0.7782_{10^1}
2	0.5555_{10^1}	0.1111_{10^1}	0.5600_{10^0}	
3	0.5555_{10^0}	0.5611_{10^2}		
4	0.5555_{10^1}	-0.5555_{10^2}		
5	0.5555_{10^3}			
6	0.5555_{10^0}		Ergebnis: $\tilde{s}_{\mathrm{II}} = 0.7782_{10^1}$	
7	-0.5555_{10^3}			
8	0			

Die Werte wurden unter Annahme einer Dezimalmaschine mit vier geltenden Ziffern und der üblichen Rundung berechnet. Das genaue Ergebnis ist $s = 0.7777_{10^1} = \tilde{s}_{\mathrm{IV}}$. *Wie zu erwarten war*, nimmt jetzt die Genauigkeit der Verfahren in der Reihenfolge I, II, III, IV zu.

An den Tabellen 2 und 4 sieht man gut, woran es liegt, daß Verfahren III i. a. schlechter ist als Verfahren IV. Immer wenn ein großes a_m zu einem kleinen s_{m-1} addiert wird, berechnet der Korrektor w_m des KAHAN-BABUŠKA-Verfahrens nicht mehr, wie geplant, den Rundungsfehler, sondern einen Wert, der mit dem Rundungsfehler nicht mehr viel zu tun hat. Die im verbesserten KAHAN-BABUŠKA-Verfahren eingebaute Abfrage vertauscht hier sozusagen die Rollen von a_m und s_{m-1} und berechnet dadurch — nach Satz 2 des fünften Abschnitts — den *richtigen* Fehler.

NICKEL [3] berechnete auf einer Maschine mit $\varepsilon = 2^{-30}$ die Summe $\sum_{i=1}^{2^t} \frac{1}{i}$ für $0 \leqq t \leqq 14$ nach Verfahren I und III (= IV in diesem Fall). Für $s = \sum_{i=1}^{2^{12}} \frac{1}{i}$ z. B. erhält er

$$|s - \tilde{s}_{\mathrm{I}}| \approx 25 \cdot 10^{-8}\,, \qquad |s - \tilde{s}_{\mathrm{IV}}| \leqq 0.5 \cdot 10^{-8}\,.$$

Unter Verwendung von $s < 9$ ergibt sich aus Tabelle 1:

$$|s - \tilde{s}_{\mathrm{I}}| \leqq 3500 \cdot 10^{-8}\,, \qquad |s - \tilde{s}_{\mathrm{IV}}| \leqq 1.7 \cdot 10^{-8}\,.$$

Der Fehler von Verfahren I wird also durch die Schranke in Tab. 1 um das 140fache, bei Verfahren IV mindestens um das dreifache überschätzt. Jedoch liegt die Tab. 1-Schranke für Verfahren IV immer noch niedriger als der maschinell berechnete Fehler von Verfahren I.

Das letzte Beispiel soll wieder auf einer L-ziffrigen Dualmaschine gerechnet werden. Diesmal soll die Rundung jedoch durch Abschneiden der $(L+i)$-ten Ziffern $(i \geqq 1)$ geschehen $(\varepsilon = 2^{1-L})$. Nun sei $n = 2^t$, $2 \leqq t \leqq L-1$,

$$a_1 = 1\,, \qquad a_2 = 2^{-L}\,, \ldots, a_{2^i+1} = \cdots = a_{2i+1} = 2^{-L} + 2^{i-2L} \qquad (1 \leqq i \leqq t-1)\,.$$

Man findet leicht:

$$\tilde{s} = 1 + 2^{t-L} - 2^{-L} + \frac{1}{3}\,2^{-2L}\,(2^{2t} - 4)\,,$$

$$\tilde{s}_{\mathrm{I}} = 1\,,$$

$$\tilde{s}_{\mathrm{II}} = 1 + 2^{t-L} - 2^{1-L} \qquad \left(t > \frac{L}{2} + 1\right)$$

$$\tilde{s}_{\mathrm{II}} = 1 + 2^{t-L} - 2^{1-L} + \frac{1}{3} \cdot 2^{-2L}(2^{2t} - 2^{L+r}) \left(t > \frac{L}{2} + 1\,;\, r = 2 \text{ für gerade } L,\ r = 1 \text{ für ungerade } L\right),$$

$$\tilde{s}_{\mathrm{III}} = 1 + 2^{t-L} - 2^{1-L}\,.$$

Man sieht unmittelbar, daß Verfahren II für $n^2 \geqq 8\,\varepsilon^{-1}$ genauer ist als Verfahren III (hier = Verfahren IV); das ist in guter Übereinstimmung mit der in Abschnitt 8 abgeleiteten Bedingung $n^2 \geqq \varepsilon^{-1} \log_2 \varepsilon^{-1}$. Daß $\tilde{s}_{\mathrm{II}}$ für kleine n hier $= \tilde{s}_{\mathrm{III}}$ ist, folgt daraus, daß wegen der speziellen Form der Summanden bei Verfahren II nur die $\tilde{s}_{1\,i}$ gerundet sind.

Für die Fehler erhält man

$$|\tilde{s}_{\mathrm{I}} - s| = 2^{t-L} + \frac{1}{3}(2^{t-L})^2 - 2^{-L} - \frac{1}{3}\,2^{2-2L} = \frac{1}{2}(n-1)\,\varepsilon + \left(\frac{1}{12}\,n^2 - \frac{1}{3}\right)\varepsilon^2\,,$$

$$|\tilde{s}_{\mathrm{II}} - s| = 2^{-L} + \frac{1}{3}\,2^{-2L}\,(2^{2t} - 4) = \frac{1}{2}\,\varepsilon + \left(\frac{1}{12}\,n^2 - \frac{1}{3}\right)\varepsilon^2 \qquad (\text{für } n^2 < 8\,\varepsilon^{-1})\,,$$

$$|\tilde{s}_{\mathrm{II}} - s| = \left(\frac{1}{3}\,2^r + 1\right)2^{-L} - \frac{1}{3}\,2^{2-2L} \leqq \frac{7}{6}\,\varepsilon - \frac{1}{3}\,\varepsilon^2 \qquad (\text{für } n^2 \geqq 8\,\varepsilon^{-1})\,,$$

$$|\tilde{s}_{\mathrm{III}} - s| = 2^{-L} + \frac{1}{3}\,2^{-2L}\,(2^{2t} - 4) = \frac{1}{2}\,\varepsilon + \left(\frac{1}{12}\,n^2 - \frac{1}{3}\right)\varepsilon^2\,.$$

Die Fehlerschranken aus Tab. 1 überschätzen hier den Fehler bei Verfahren I um den Faktor 2, bei Verfahren II um den Faktor $2 \log_2 \varepsilon^{-1}$, bei Verfahren III um den Faktor 4 (da $s \approx 1$). Der hohe Fehler (im Vergleich zur Schranke) bei Verfahren I und III und der kleine Fehler bei Verfahren II rühren von der speziellen Form der a_i her; im allgemeinen werden die Fehler bei Verfahren I und III kleiner, bei Verfahren II größer sein. Trotzdem gibt dieses Beispiel die Unterschiede der benutzten Summierungsverfahren gut wieder, wie sie in Abschnitt 8.1 diskutiert wurden.

Literatur

1 BABUŠKA, I., Numerical Stability in Mathematical Analysis, Information Processing 68, Amsterdam, 11–22 (1969).
2 KAHAN, W., Further Remarks on Reducing Truncation Errors, Comm. ACM, 8, 40 (1965).
3 NICKEL, K., Das Kahan-Babuškasche Summierungsverfahren in Triplex-ALGOL 60, ZAMM 50, 369–373 (1970).
4 WILKINSON, J. H., Rounding Errors in Algebraic Processes, London 1963; dt. Übersetzung: Rundungsfehler, Berlin-Heidelberg-New York 1969.

Eingereicht am 6. 3. 1973

Anschrift: ARNOLD NEUMAIER, 75 Karlsruhe 41, Kieselweg 1, BRD