ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
XIII, 3 (1973)

A. SCHURMANN (Sopot)

ON THE MINIMUM ERROR IN ADDITION PROCESSES OF POSITIVE FLOATING-POINT NUMBERS

0. Summary. The paper considers the dependence of the error of addition of n floating-point binary numbers upon the used summation sequence. Let ALN denote the summation sequence of n numbers from M_n, described as follows:

1. Add the two least numbers a and b from M_n; normalize and round-off the result S_n.

2. Let M_{n-1} consist of all numbers (with exception of a and b) from M_n and the number S_n. Reduce the index n by 1. If $n \neq 1$, then return to 1.

There are given (Section 2) two examples, which show that the error of addition according to the ALN-summation sequence can be greater than the minimal one. In Section 3 is proved the main theorem of this paper, which states that if all exponents of the numbers from M_n are different, then the addition of these numbers according to the summation sequence ALN causes a minimal truncation error.

1. Introduction. Let M_n be a set of n positive normalized floating-point numbers. The approximate sum of these n numbers is calculated in the following way:

I. Find the two least numbers a and b of the set M_n.

II. Calculate and normalize the sum $a+b$, and then round-off the normalized sum.

III. Let M_{n-1} denote the set $M_n - \{a, b\} \cup \{(a+b) \pm \delta\}$, where δ is the absolute rounding error.

IV. Reduce the index n by 1.

V. If $n \neq 1$, then return to I. Otherwise stop; the number from M_1 is the approximate sum of the numbers from M_n.

We call this way of addition the *ALN-algorithm* (Addition according to the two Least Numbers). It is well known that the accuracy of a sum

of n numbers depends on the used summation sequence. The problem to be discussed here is:

Let ε denote the error of the sum of numbers from M_n, caused by the use of the summation sequence defined by the algorithm ALN. Is there a summation sequence of numbers from M_n which causes an error of the sum less than ε?

The examples given in Section 2 show that the accuracy of the sum obtained by the ALN-algorithm can be not maximal. The main theorem (Section 3) states that the accuracy of a sum of n numbers from M_n is maximal if the algorithm ALN is used and all numbers from M_n have different exponents. It is assumed that the numbers from M_n are expressed by the binary normalized floating-point representation.

2. Examples of non-minimal summation errors obtained by the ALN--algorithm. In this paper it is assumed that any number z is represented by an ordered pair of numbers x and c such that $z = x \cdot 2^c$, where x is a binary fractional number called *mantissa* and c is an integral called *exponent*. To abbreviate, instead of $x \cdot 2^c$ we shall write $x'c$. In what follows, we shall say that $x'c \neq 0$ is a *normalized t-figure number* if the mantissa x has the form $0.1x_2x_3 \ldots x_t$, where x_i is equal to 0 or to 1 ($i = 2, 3, \ldots, t$). Thus $1/2 \leqslant x \leqslant 1 - 2^{-t}$.

Let b denote a floating-point number. By $\tau(b)$ we denote the number obtained from b by the following operations:

1° normalize b, i. e. write b in the form $b = 0.1b_2b_3 \ldots b_k'c$;

2° if $k > t$, then round-off the normalized b to the t-figure number in such a way that to $0.1b_2 \ldots b_k'c$ add the number $0.1'(c-t)$ and then throw away the figures from the positions $t+1$ to k.

Now we give examples which show that the ALN-algorithm can cause non-minimal rounding errors.

Example 1. Let $t = 5$, $a_1 = 0.10000'c$, $a_2 = 0.10001'c$ and $a_3 = 0.10101'c$. According to the ALN-algorithm, the approximate sum of a_1, a_2 and a_3 equals

$$\tau\big(\tau(a_1 + a_2) + a_3\big) = \tau\big(0.10001'(c+1) + a_3\big) = 0.11100'(c+1).$$

The accurate sum equals

$$a_1 + a_2 + a_3 = 0.11011'(c+1).$$

Thus the absolute value of the rounding error is equal to $0.0001'c$.

Now we add the numbers in the following sequence: a_2 to a_3 and then to a_1. We have

$$\tau\big(\tau(a_2 + a_3) + a_1\big) = \tau\big(0.10011'(c+1) + a_1\big) = 0.11011'(c+1).$$

Thus, in this case, the rounding error is equal to 0.

In what follows, we understand by $\gamma(b)$ the number obtained from b by the following operations:

1° normalize b, i. e. write b in the form $b = 0.1b_2b_3 \dots b_k'c$;

2° if $k > t$, then throw away the figures from the positions $t+1$ to k.

Thus $\gamma(b) = 0.1b_2 \dots b_t'c$. Example 2 shows that the algorithm ALN can cause non-minimal errors also in the case of truncation by γ.

Example 2. Let a_1, a_2 and a_3 have the same values as in Example 1. Adding according the algorithm ALN, we obtain

$$\gamma\big(\gamma(a_1+a_2)+a_3\big) = \gamma\big(0.10000\ (c+1)+a_3\big) = 0.11010'(c+1).$$

The error is equal to $0.0001'c$. But if we add the numbers in the sequence $(a_2+a_3)+a_1$, then the result is accurate

$$\gamma\big(\gamma(a_2+a_3)+a_1\big) = \gamma\big(0.10011'(c+1)+a_3\big) = 0.11011'(c+1).$$

3. Minimal error of summation of numbers with different exponents. We shall assume that the result of addition of normalized t-figure numbers a and b is the number $\gamma(a+b)$. This kind of addition we call *γ-addition.* We shall now prove the main theorem of this paper.

THEOREM. *Let A_n denote a set of n positive normalized t-figure numbers represented by $a_i = m_i'c_i$ for $i = 1, 2, \dots, n$. It is assumed that $c_i < c_{i+1}$ for $i = 1, 2, \dots, n-1$. If the ALN-algorithm is used, then the error of γ-addition of n numbers from A_n is minimal.*

In order to prove the theorem, we shall show that lemmas 1 and 2 are true. The following notation will be useful:

We say that a summation sequence of n numbers from A_n is *natural* (in abbreviation, *NS-sequence*) if the sum of these numbers arises in such a way that every performed addition operation has as its argument at least one number from A_n, i. e. the first performed addition operation has two arguments from A_n, all other addition operations have only one argument. All other summation sequences we shall denote by PS.

Thus, n numbers from A_n are summed according to an NS-sequence if they are added according to the formula

$$\big(\dots\big((a_{v_1}+a_{v_2})+a_{v_3}\big)+\dots\big)+a_{v_n},$$

where $1 \leqslant v_i \leqslant n$ $(i = 1, \dots, n)$ and $v_i \neq v_j$ for $i \neq j$. If $c_i < c_{i+1}$, the ALN-algorithm applied to numbers from A_n defines an NS-sequence. We call it an *ALN-sequence.* An example of a PS-sequence of the numbers from the set $\{b_1, b_2, b_3, b_4\}$ is given by the formula $(b_1+b_2)+(b_3+b_4)$.

LEMMA 1. *Let the notation and assumptions of the theorem be valid. In the set of all γ-additions of n numbers from A_n according to NS-sequences, the γ-addition according to the ALN-sequence causes the minimal truncation error.*

Proof. The lemma will be proved by induction. It is easy to see that the lemma is true for $n = 2$. By the induction hypothesis, the lemma is true for $n = k$. We shall prove that it is also true for $n = k+1$.

Let us divide the set of all NS-sequences of numbers from set $A_{k+1} = \{a_1, a_2, \ldots, a_{k+1}\}$ into two sets Q and R. A set Q consists of all NS-sequences such that a_{k+1} is added as the last argument. Thus to Q belong all NS-sequences defined by the formula

$$\left(\ldots\left((a_{v_1} + a_{v_2}) + a_{v_3}\right) + \ldots + a_{v_k}\right) + a_{k+1},$$

where $1 \leqslant v_i \leqslant k$ $(i = 1, 2, \ldots, k)$ and $v_i \neq v_j$ for $i \neq j$. To a set R belong all NS-sequences which do not belong to Q, i. e. NS-sequences defined by the formula

$$\left(\ldots\left((a_{v_1} + a_{v_2}) + a_{v_3}\right) + \ldots\right) + a_{v_{k+1}},$$

where $v_{k+1} \neq k+1$, $1 \leqslant v_i \leqslant k+1$ $(i = 1, 2, \ldots, k+1)$ and $v_i \neq v_j$ for $i \neq j$ and $i, j \leqslant k+1$. In this way, γ-addition of numbers from set A_{k+1} according to an NS-sequence L is equivalent to γ-addition by the sequence: first the numbers from the set $A_k = \{a_1, a_2, \ldots, a_k\}$ according to an NS-sequence and then the number a_{k+1} (if an NS-sequence L belongs to Q); or by the sequence: first the numbers from the set $\bar{A}_k = \{a_1, a_2, \ldots, a_{j-1}, a_{j+1}, \ldots, a_{k+1}\}$ according to an NS-sequence and then a_j (if an NS-sequence L belongs to R).

By the induction hypothesis, the error of γ-addition of k numbers from A_k is minimal if these numbers are added according to the ALN-algorithm. From the above-mentioned, it follows that from among γ-additions of $k+1$ numbers from A_{k+1}, according to NS-sequences belonging to Q, the γ-addition according to the ALN-sequence causes the minimal error. This error is equal to

$$\varepsilon = \sum_{i=2}^{k+1} \varepsilon_i, \tag{1}$$

where ε_i denotes the error caused by γ-addition of a_i. More exactly, ε_i $(i = 2, 3, \ldots, k+1)$ are defined by partial sums S_i as follows:

$$\begin{aligned} S_1 &= a_1, \\ S_{i+1} &= \gamma(S_i + a_{i+1}) \qquad \text{for } i = 1, 2, \ldots, k, \\ \varepsilon_i &= (S_{i-1} + a_i) - S_i \qquad \text{for } i = 2, 3, \ldots, k+1. \end{aligned} \tag{2}$$

S_{k+1} is the result of γ-addition of $k+1$ numbers from A_{k+1} according to the ALN-algorithm.

By the induction hypothesis, the error of γ-addition of k numbers from $\bar{A}_k$ is minimal if these numbers are added according to the ALN-algorithm. Hence, it follows that the error of γ-addition of $k+1$ numbers from A_{k+1}, according to an NS-sequence belonging to R, is minimal if these numbers are added by the NS-sequence (from R) defined by the formula

$$(3) \qquad \Big(\dots\Big(\big(\dots\big((a_1+a_2)+a_3\big)+\dots+a_{j-1}\big)+a_{j+1}\Big)+\dots+a_{k+1}\Big)+a_j.$$

This minimal error is equal to

$$(4) \qquad \bar{\varepsilon} = \sum_{i=2}^{j-1}\varepsilon_i + \sum_{i=v+2}^{k+1}\bar{\varepsilon}_i+\bar{\varepsilon}_j,$$

where $v = \max(1, j-1)$ and $\bar{\varepsilon}_i$ (for $i = v+1, \dots, k+1$) denotes the error caused by γ-addition of a_i. More exactly, $\bar{\varepsilon}_i$ is defined by partial sums T_i as follows:

$$\begin{aligned}
T_v &= S_{j-1} && \text{if } j>1, \text{ otherwise } T_v = a_2,\\
T_i &= \gamma(T_{i-1}+a_{i+1}) && \text{for } i = v+1, v+2, \dots, k,\\
T_{k+1} &= \gamma(T_k+a_j),\\
\bar{\varepsilon}_i &= (T_{i-2}+a_i)-T_{i-1} && \text{for } i = v+2, v+3, \dots, k+1,\\
\bar{\varepsilon}_j &= (T_k+a_j)-T_{k+1}.
\end{aligned}$$

Thus, T_{k+1} is the result of γ-addition of numbers from A_{k+1} according to the NS-sequence defined by formula (3). From the above-mentioned it follows that in order to prove the lemma for $n = k+1$ we have to show that $\varepsilon \leqslant \bar{\varepsilon}$. By (1) and (4), this inequality is equivalent to the following inequality:

$$(5) \qquad \sum_{i=j}^{k+1}\varepsilon_i \leqslant \sum_{i=v+2}^{k+1}\bar{\varepsilon}_i+\bar{\varepsilon}_j.$$

Now we define an operation β. If $x = x_e x_{e+1} \dots x_0.x_1 \dots x_t\text{'}p$ (where $e \leqslant 0$), then by $\beta(x, i)$ we denote the number with the exponent p and the mantissa consisting of the i last bits of the mantissa of x. Formally, $\beta(x, i)$ $(i \leqslant t+e+1)$ is defined as follows:

$$\beta(x,i) = \begin{cases} 0.\underbrace{0\dots0}_{t-i}x_{t+1-i}\dots x_t\text{'}p & \text{if } i \leqslant t,\\ x_{t+1-i}\dots x_0.x_1\dots x_t\text{'}p & \text{if } i > t. \end{cases}$$

Using the operation β we can express the error ω of γ-addition of two normalized t-figure numbers $a = z_1'c$ and $b = z_2'o$ (where $o \geqslant c$) by the formula

$$\omega = \beta[\beta(a, d-c)+\beta(b, d-o), d-c] \quad (6)$$

where d denotes the exponent of the normalized number $\gamma(a+b)$.

In this paper, the symbol * denotes a right shift operation defined as follows:

if $y = y_e \dots y_0 . y_1 \dots y_v' c$ and m is a positive integer, then

$$y*m = \begin{cases} y_e \dots y_{e+m} . y_{e+m+1} \dots y_0 y_1 \dots y_{v-m}'(c+m) & \text{if } m \leqslant |e|, \\ 0.\underbrace{0 \dots 0}_{e+m-1} y_e \dots y_0 y_1 \dots y_{v-m}'(c+m) & \text{if } m > |e|. \end{cases}$$

Thus the number $y*m$ is obtained by the following operations:

1° shift right the mantissa of y m times and truncate the m less significant bits,

2° increase the exponent by m.

It is easy to show that

$$\beta(\gamma(a+b), l) = \beta[[\beta(a, d-c+l)+\beta(b, d-o+l)]*(d-c), l], \quad (7)$$

where a and b have the same meaning as in (6). In order to prove inequality (5), we show that

$$\beta(S_i, l) = \beta\left\{\left[\sum_{u=j}^{i}\beta(a_u, g_i-c_u+l)+\beta(S_{j-1}, g_i-g_{j-1}+l)\right]*(g_i-g_{j-1}), l\right\}, \quad (8)$$

where $l \leqslant t$ and g_e denotes the exponent of the normalized t-figure number S_e for $e = 1, 2, \dots, k$. Equality (8) will be proved by induction on i. Since $S_j = \gamma(S_{j-1}+a_j)$, we infer from (7) that equality (8) is true for $i = j$.

Now, let us assume that (8) is true for $i = e$. We shall prove that (8) is also true for $i = e+1$. From (2) and (7) we have

$$\beta(S_{e+1}, l) = \beta\{[\beta(a_{e+1}, g_{e+1}-c_{e+1}+l)+\beta(S_e, g_{e+1}-g_e+l)]*(g_{e+1}-g_e), l\}.$$

Hence, using the induction hypothesis,

$$\beta(S_{e+1}, l) = \beta\left\{\left\{\beta(a_{e+1}, g_{e+1}-c_{e+1}+l)+\beta\left[\left[\sum_{u=j}^{e}\beta(a_u, g_{e+1}-c_u+l)+\right.\right.\right.$$

$$\left.\left.\left.+\beta(S_{j-1}, g_{e+1}-g_{j-1}+l)\right]*(g_e-g_{j-1}), g_{e+1}-g_e+l\right]\right\}*(g_{e+1}-g_e), l\right\}.$$

---

(1) We assume that the addition + does not normalize the result. The exponent of ω is c. See the definition of addition in the appendix.

We use the following equality which can be verified:

$$(9)\qquad \beta\big[[\beta(b,\, d-o+l)+\beta(a,\, d-c+l)]*(d-c),\, l\big] = \beta\big[[(b,\, d-o+l)+a]*(d-c),\, l\big].$$

Substituting a_{e+1} for b, S_e for a and $g_{e+1}-g_e$ for $d-c$, we obtain, by formula (9), equality (8) for $i = e+1$, which completes the proof of (8).

Now we shall prove, by induction on k, the following formula:

$$(10)\qquad \sum_{i=j}^{k+1} \varepsilon_i = \beta\Big[\sum_{i=j}^{k+1} \beta(a_i,\, g_{k+1}-c_i)+\beta(S_{j-1},\, g_{k+1}-g_{j-1}),\, g_{k+1}-g_{j-1}\Big].$$

For $k = j-1$ formula (10) follows from (6).

Suppose (10) is true for $k = e-1$, where $e \geqslant j$. It will be proved that (10) is true also for $k = e$. Remembering that $S_{e+1} = \gamma(S_e + a_{e+1})$, using (6), we have

$$\varepsilon_{e+1} = \beta[\beta(S_e,\, g_{e+1}-g_e)+\beta(a_{e+1},\, g_{e+1}-c_{e+1}),\, g_{e+1}-g_e].$$

Applying formula (8) to $\beta(S_e,\, g_{e+1}-g_e)$ and the obvious equality

$$(11)\qquad \beta[\beta(a,\, d-c)+\beta(b,\, d-o),\, d-c] = \beta[a+\beta(b,\, d-o),\, d-c],$$

and substituting $b = a_{e+1}$, $d-c = g_{e+1}-g_e$ and

$$a = \Big[\sum_{u=j}^{e} \beta(a_u,\, g_{e+1}-g_u)+\beta(S_{j-1},\, g_{e+1}-g_{j-1})\Big]*(g_e-g_{j-1}),$$

we get

$$\varepsilon_{e+1} = \beta\Big\{\Big[\sum_{i=j}^{e} \beta(a_i,\, g_{e+1}-c_u)+\beta(S_{j-1},\, g_{e+1}-g_{j-1})\Big]*(g_e-g_{j-1})+ +\beta(a_{e+1},\, g_{e+1}-c_{e+1}),\, g_{e+1}-g_e\Big\}.$$

Applying the formula

$$(12)\qquad \beta\big[\beta(a,\, l-(d-c))+\beta(b,\, l-(d-o)),\, l\big] = \beta\big[\beta(a,\, l-(d-c)+w)+\beta(b,\, l-(d-o)+w),\, l\big] \qquad (w \geqslant 0)$$

to the right-hand side of the equality given by the induction hypothesis (substituting $w = g_{e+1}-g_e$), we obtain

$$\sum_{i=j}^{e} \varepsilon_i = \beta\Big[\sum_{i=j}^{e} \beta(a_i,\, g_{e+1}-c_i)+\beta(S_{j-1},\, g_{e+1}-g_{j-1}),\, g_e-g_{j-1}\Big].$$

One can verify the following property:

If $y = 0.y_1 \dots y_t \dots y_{t+m}'c$ and $b = 0.b_1 \dots b_l'o$ and $c \leqslant o \leqslant c+m+l$, then $\beta(y*m+b, l)+\beta(y, m) = \beta(y+b, m+l)$.

Substituting in this equality

$$b = a_{e+1}, \quad m = g_e - g_{j-1}, \quad l = g_{e+1} - g_e,$$

$$y = \sum_{i=j}^{e} \beta(a_i, g_{e+1} - c_i) + \beta(S_{j-1}, g_{e+1} - g_{j-1}),$$

and using the last formulas for ε_{e+1} and $\sum_{i=j}^{e} \varepsilon_i$, we get equality (10) for $k = e$, which completes the proof of (10).

Similarly as (10) we can show that

$$\sum_{i=j+1}^{k+1} \bar{\varepsilon}_i = \beta\Big[\sum_{i=j+1}^{k+1} \beta(a_i, q_k - c_i) + \beta(S_{j-1}, q_k - g_{j-1}), q_k - g_{j-1}\Big],$$

where q_i $(i = j, j+1, \dots, k+1)$ denotes the exponent of T_i. Applying (12) to the right-hand side of the above-mentioned equality (substituting $w = q_{k+1} - q_k$), we have

$$\sum_{i=j+1}^{k+1} \bar{\varepsilon}_i = \beta\Big[\sum_{i=j+1}^{k+1} \beta(a_i, q_{k+1} - c_i) + \beta(S_{j-1}, q_{k+1} - g_{j-1}), q_k - g_{j-1}\Big]. \tag{13}$$

Using formula (6) to $\bar{\varepsilon}_j$, we obtain

$$\bar{\varepsilon}_j = \beta[\beta(a_j, q_{k+1} - c_j) + \beta(T_k, q_{k+1} - q_k), q_{k+1} - c_j].$$

It is easy to verify the inequality

$$\beta(a, m) + \beta[\beta(a*m, l) + \beta(b, l-w), l-w] \geqslant \beta[\beta(a, l+m) + \beta(b, l-w), m+l],$$

where $w = q - c - m$. Substituting in this inequality for $\beta(a, m)$ the right-hand side of equality (13), $m = q_k - g_{j-1}$, $l = q_{k+1} - q_k$ and $b = a_j$, we obtain

$$\sum_{i=j+1}^{k+1} \bar{\varepsilon}_i + \bar{\varepsilon}_j \geqslant \beta\Big\{\beta\Big[\sum_{i=j+1}^{k+1} \beta(a_i, q_{k+1} - c_i) + \beta(S_{j-1}, q_{k+1} - g_{j-1}), q_{k+1} - g_{j-1}\Big] + \beta(a_j, q_{k+1} - c_j), q_{k+1} - g_{j-1}\Big\}$$

$$= \beta\Big[\sum_{i=j}^{k+1} \beta(a_i, q_{k+1} - c_i) + \beta(S_{j-1}, q_{k+1} - g_{j-1}), q_{k+1} - g_{j-1}\Big].$$

The last equality follows from (11). Hence, using (10), we infer that if $g_{k+1} \leqslant q_{k+1}$, then (5) is true for $n = k+1$. If $g_{k+1} > q_{k+1}$, i. e. if $g_{k+1} = c_{k+1}+1$ and $q_{k+1} = c_{k+1}$, then $S_{k+1} > T_{k+1}$. Hence, using the equality $\varepsilon + S_{k+1} = \bar{\varepsilon} + T_{k+1}$, we infer that $\varepsilon < \bar{\varepsilon}$. Thus, the lemma is proved in the case where $j > 1$. The proof of (5) is analogous in the case where $j = 1$ and $v = 1$. This completes the proof of Lemma 1.

LEMMA 2. *Let the notation and assumptions of the theorem be valid. The error caused by γ-addition of n numbers from A_n according to the ALN-sequence is not greater than the error of γ-addition of these numbers according to a PS-sequence.*

Proof. Let us observe that a γ-addition according to a PS-sequence can be defined recursively as follows:

Addition described by the formula

(14) $\gamma[(\gamma\text{-sum of numbers } a_{u_1}, \ldots, a_{u_k} \text{ according to an NS-sequence}) + (\gamma\text{-sum of numbers } a_{v_1}, \ldots, a_{v_s} \text{ according to an NS-sequence})]$, where $k > 1$ and $s > 1$,

is a γ-addition of numbers $a_{u_1}, \ldots, a_{u_k}, a_{v_1}, \ldots, a_{v_s}$ according to a PS-sequence.

Addition described by the formula

(15) $\gamma[(\gamma\text{-sum of numbers } a_{g_1}, \ldots, a_{g_r} \text{ according to a PS-sequence}) + (\gamma\text{-sum of numbers } a_{q_1}, \ldots, a_{q_e} \text{ according to a PS-sequence})]$

is a γ-addition of numbers $a_{g_1}, \ldots, a_{g_r}, a_{q_1}, \ldots a_{q_e}$ according to a PS-sequence.

We use the following notation:

Let T_u and T_v denote the results of γ-additions of numbers $a_{u_1}, \ldots, a_{u_k}$ and $a_{v_1}, \ldots, a_{v_s}$, respectively, according to NS-sequences. Then $\bar{\varepsilon}_u$ and $\bar{\varepsilon}_v$ defined by

$$\bar{\varepsilon}_u = \sum_{i=1}^{k} a_{u_i} - T_u \quad \text{and} \quad \bar{\varepsilon}_v = \sum_{i=1}^{s} a_{v_i} - T_v$$

denote the errors of γ-additions of numbers $a_{u_1}, \ldots, a_{u_k}$ and $a_{v_1}, \ldots, a_{v_s}$, respectively, according to NS-sequences. Let T denote the result of the sum given by (14), i. e. $T = \gamma(T_u + T_v)$, and let $\bar{\varepsilon}_1$ be the error of γ-addition of T_u to T_v. Thus $\bar{\varepsilon}_1 = (T_u - T_v) - T$. The entire error $\bar{\varepsilon}$ of the sum given by (14) is equal to $\bar{\varepsilon} = \bar{\varepsilon}_u + \bar{\varepsilon}_v + \bar{\varepsilon}_1$.

By S_u and S_v we denote the result of γ-additions of numbers $a_{u_1}, \ldots, a_{u_k}$ and $a_{v_1}, \ldots, a_{v_s}$, respectively, according to the ALN-algorithm. The errors arisen from these γ-additions we denote by ε_u and ε_v, respec-

tively. Thus

$$\varepsilon_u = \sum_{i=1}^{k} a_{u_i} - S_u \quad \text{and} \quad \varepsilon_v = \sum_{i=1}^{s} a_{v_i} - S_v.$$

Further, let $\bar{S}$ denote the sum $\gamma(S_u + S_v)$ and ε_1 the associated rounding error, i. e. $\varepsilon_1 = (S_u + S_v) - \bar{S}$.

By Lemma 1 we infer that $\varepsilon_u \leqslant \bar{\varepsilon}_u$ and $\varepsilon_v \leqslant \bar{\varepsilon}_v$. Hence, since $\varepsilon_u + S_u = \bar{\varepsilon}_u + T_u$ and $\varepsilon_v + S_v = \bar{\varepsilon}_v + T_v$, we infer that $S_u \geqslant T_u$ and $S_v \geqslant T_v$. Using these inequalities, we get $\gamma(S_u + S_v) \geqslant \gamma(T_u + T_v)$. Hence we have

$$\bar{S} \geqslant T. \tag{16}$$

Now we shall show that $\varepsilon \leqslant \varepsilon_u + \varepsilon_v + \varepsilon_1$, where ε denotes the error of γ-addition of numbers $a_{u_1}, \ldots, a_{u_k}, a_{v_1}, \ldots, a_{v_s}$ according to the ALN-algorithm. Suppose that $u_i < u_{i+1}$ $(i = 1, 2, \ldots, k-1)$ and $v_j < v_{j+1}$ $(j = 1, 2, \ldots, s-1)$. Applying formula (10), we obtain

$$\varepsilon_u = \beta\Big[\sum_{i=1}^{k} \beta(a_{u_i},\, g_u - c_{u_i}),\, g_u - c_{u_1}\Big]$$

and

$$\varepsilon_v = \beta\Big[\sum_{i=1}^{s} \beta(a_{v_i},\, g_v - c_{v_i}),\, g_v - c_{v_1}\Big].$$

Hence, using formula (12) (substituting $\bar{g} - g_u$ and $\bar{g} - g_v$ for w, $g_u - c_{u_1}$ and $g_v - c_{v_1}$ for l), we get

$$\varepsilon_u = \beta\Big[\sum_{i=1}^{k} \beta(a_{u_i},\, \bar{g} - c_{u_i}),\, g_u - c_{u_1}\Big] \tag{17}$$

and

$$\varepsilon_v = \beta\Big[\sum_{i=1}^{s} \beta(a_{v_i},\, \bar{g} - c_{v_i}),\, g_v - c_{v_1}\Big], \tag{18}$$

where $\bar{g}$ denotes the exponent of $\bar{S}$.

Let us suppose that $v_s > u_k$. Then, by the definition of ε_1, using formula (6) and then formula (8), we obtain

$$\begin{aligned}
\varepsilon_1 &= \beta[\beta(S_u,\, \bar{g} - g_u) + \beta(S_v,\, \bar{g} - g_v),\, \bar{g} - g_u] \\
&= \beta\Big\{\beta\Big\{\Big[\sum_{i=1}^{k} \beta(a_{u_i},\, \bar{g} - c_{u_i})\Big] * (g_u - c_{u_1}),\, \bar{g} - g_u\Big\} + \\
&\qquad + \beta\Big\{\Big[\sum_{i=1}^{s} \beta(a_{v_i},\, \bar{g} - c_{v_i})\Big] * (g_v - c_{v_1}),\, \bar{g} - g_v\Big\},\, \bar{g} - g_u\Big\}.
\end{aligned}$$

One can verify the inequality

$$\beta(a, d_1-c)+\beta(b, d_2-o)+\beta[\beta(a*(d_1-c), d-d_1)+\beta(b*(d_2-o), d-d_2), f] \geqslant \beta[\beta(a, d-c)+\beta(b, d-o), f] = \beta(a+b,f),$$

where d denotes the exponent of $\gamma(a+b)$, $d \geqslant d_1 \geqslant c$, $d \geqslant d_2 \geqslant 0$ and $f = \max(d-c, d-o)$. Substituting in the above-mentioned formula the right-hand side of equality (17) for $\beta(a, d_1-c)$, the right-hand side of equality (18) for $\beta(b, d_2-o)$, $\bar{g}$ for d and using the last formula for ε_1, we get

$$\varepsilon_u+\varepsilon_v+\varepsilon_1 \geqslant \beta\Big[\sum_{i=1}^{k}\beta(a_{u_i}, \bar{g}-c_{u_i})+\sum_{i=1}^{s}\beta(a_{v_i}, \bar{g}-c_{v_i}), f\Big] = \beta\Big[\sum_{i=1}^{k+s}\beta(a_{w_i}, \bar{g}-c_{w_i}), \bar{g}-c_{w_1}\Big],$$

where $f = \max(\bar{g}-c_{v_1}, \bar{g}-c_{u_1})$ and w: $\{1, 2, \ldots, k+s\} \to \{u_1, \ldots, u_k, v_1, \ldots, v_s\}$ is a one-to-one mapping (sequence) such that $w_i < w_{i+1}$ for $i = 1, 2, \ldots, s+k-1$. Let S denote the result of γ-addition of numbers $a_{u_1}, \ldots, a_{u_k}, a_{v_1}, \ldots, a_{v_s}$ according to the ALN-sequence and let q denote the exponent of the normalized number S. Then, applying (10), we get

$$\varepsilon = \beta\Big[\sum_{i=1}^{k+e}\beta(a_{w_i}, q-c_{w_i}), q-c_{w_1}\Big].$$

Hence and from the above-mentioned, if $q = c_{v_s}$, then we get $\varepsilon_u+\varepsilon_v+\varepsilon_1 \geqslant \varepsilon_u+\varrho \geqslant \varepsilon$, what was to be shown. Now, let us consider the case where $q > c_{v_s}$, i. e. where $q = c_{v_s}+1$. If $S \geqslant \bar{S}$, then, by the equality $\varepsilon_u+\varepsilon_v+\varepsilon_1+\bar{S} = \varepsilon+S$, we infer that $\varepsilon_u+\varepsilon_v+\varepsilon_1 \geqslant \varepsilon$. Finally, if $S < \bar{S}$, then $\bar{g} > q$, i. e. $g = c_{v_s}+1$ and $q = c_{v_s}$. But the last equality contradicts the assumption that $q > c_{v_s}$. Thus we proved that $\varepsilon_u+\varepsilon_v+\varepsilon_1 \geqslant \varepsilon$. Hence, using the equality $\bar{S}+\varepsilon_u+\varepsilon_v+\varepsilon_1 = T+\bar{\varepsilon}$ and inequality (16), we infer that $\bar{\varepsilon} \geqslant \varepsilon$ which completes the proof of Lemma 2 in the case of γ-additions described by formula (14).

Formulas (14) and (15) define inductively γ-addition according to a PS-sequence. In order to complete the proof of Lemma 2, we have to show that if the lemma is true in the case of γ-additions of numbers $a_{g_1}, \ldots, a_{g_r}$ and $a_{q_1}, \ldots, a_{q_e}$ according to a PS-sequence, then it is also true for γ-additions described by formula (15).

Let $\bar{\varepsilon}_g$ and $\bar{\varepsilon}_q$ denote the errors of γ-additions of numbers $a_{g_1}, \ldots, a_{g_r}$ and $a_{q_1}, \ldots, a_{q_e}$, respectively, according to a PS-sequence, and let ε_g and ε_q denote the errors of γ-additions of these numbers, respectively, by the ALN-algorithm. The induction hypothesis of the proof implies $\varepsilon_g \leqslant \bar{\varepsilon}_g$

and $\varepsilon_q \leqslant \bar{\varepsilon}_q$. Further, the proof is similarly to the above-described proof of the lemma in the case of γ-additions given by formula (14). Thus the lemma is true.

The proof of the theorem follows immediately from Lemmas 1 and 2.

APPENDIX

In the sequel we shall verify the formulas used in this paper without proof. First, we define the addition of two floating-point numbers.

Definition. Let

$$x = x_e x_{e+1} \dots x_0 . x_1 \dots x_t' c,$$

$$y = y_e y_{e+1} \dots y_0 . y_1 \dots y_t' q \quad \text{and} \quad n = q - c \geqslant 0.$$

Then

$$x + y = z_v z_{v-1} \dots z_0 . z_1 \dots z_t' c,$$

where $v = e - n - 1$ and z_i are defined recursively as follows:

(1.1)
$$z_i = x_i \quad \text{for } i = t, t-1, \dots, t-n+1,$$
$$p_{t-n} = 0;$$

$$z_i = (x_i + y_{i+n} + p_i) \pmod 2 \quad \text{for } i = t-n, t-n-1, \dots, e,$$

(1.2)
$$p_{i+1} = \begin{cases} 0 & \text{if } x_i + y_{i+n} + p_i \geqslant 2, \\ 1 & \text{otherwise;} \end{cases}$$

$$z_i = (y_{i+n} + p_i) \pmod 2 \quad \text{for } i = e-1, e-2, \dots, e-n,$$

(1.3)
$$p_{i+1} = \begin{cases} 0 & \text{if } y_{i+n} + p_i = 2, \\ 1 & \text{otherwise.} \end{cases}$$

(1.4)
$$z_v = p_v.$$

Now we shall show that the error ω of γ-addition of two normalized numbers $x = 0.1x_2 \dots x_t' c$ and $y = 0.1y_2 \dots y_t' q$ (where $q \geqslant c$) is given by the formula

(2)
$$\omega = \beta[\beta(x, n+j) + \beta(y, j), n+j],$$

where $n = q - c$, $j = d - q$, and d denotes the exponent of the normalized number $\gamma(x+y)$. Since $e = 1$, by the above-mentioned definition of addition, we have

$$x + y = z_{-n} z_{-n+1} \dots z_0 . z_1 \dots z_t' c,$$

where $z_i\,(i = -n, -n+1, \ldots, t)$ are defined in (1.1)-(1.4). Hence, after normalization,

$$x+y = 0.1z_{-n-j+2} \ldots z_t{}'d.$$

Thus

$$\omega = \underbrace{0.0 \ldots 0}_{t} z_{t-n-j+1} \ldots z_t{}'d$$

$$= \begin{cases} \underbrace{0.0 \ldots 0}_{t} x_{t-n+1} \ldots x_t{}'d & \text{if } j = 0, \\ \underbrace{0.0 \ldots 0}_{t} (x_{t-n}+y_t) (\mathrm{mod}\ 2) x_{t-n+1} \ldots x_t{}'d & \text{if } j = 1. \end{cases}$$

Using the definition of β, we obtain

$$\beta(x, n+j)+\beta(y, j)$$

$$= \begin{cases} \underbrace{0.0 \ldots 0}_{t-n} x_{t-n+1} \ldots x_t{}'c = \underbrace{0.0 \ldots 0}_{t} x_{t-n+1} \ldots x_t{}'d & \text{if } j = 0, \\ \underbrace{0.0 \ldots 0}_{t-n-2} p_{t-n-1} (x_{t-n}+y_t) (\mathrm{mod}\ 2) x_{t-n+1} \ldots x_t{}'c \end{cases}$$

$$= \underbrace{0.0 \ldots 0}_{t-1} p_{t-n-1} (x_{t-n}+y_t) (\mathrm{mod}\ 2)\, x_{t-n+1} \ldots x_t{}'d \qquad \text{if } j = 1.$$

Hence, using the last formula for ω, we get (2).

Let the notation of formula (2) be valid. We shall show that

$$\beta\big(\gamma(x+y), l\big) = \beta\big[[\beta(x, m)+\beta(y, j+l)]*(d-c), l\big], \tag{3}$$

where $m = n+j+l = d-c+l$. From the proof of (2) it follows that

$$\gamma(x+y) = 0.1z_{-n-j+2} \ldots z_{t-n-j}{}'d.$$

Hence

$$\beta\big(\gamma(x+y), l\big) = \underbrace{0.0 \ldots 0}_{t-l} z_{t-n-j+1-l} \ldots z_{t-n-j}{}'d. \tag{3.1}$$

Let $z = \beta(x, m)+\beta(y, j+l)$. Then

$$z = \underbrace{0.0 \ldots 0}_{t-m} x_{t-m+1} \ldots x_t{}'c + \underbrace{0.0 \ldots 0}_{t-j-l} y_{t-j-l+1} \ldots y_t{}'q$$

$$= \underbrace{0.0 \ldots 0}_{t-m-1} p_{t-m} z_{t+1-m} \ldots z_t{}'c.$$

Hence

$$z*(n+j) = \underbrace{0.0 \ldots 0}_{t-l-1} p_{t-m} z_{t+1-m} \ldots z_{t-n-j}{}'d. \tag{3.2}$$

Thus, using (3.1), we get (3).

Using the above-mentioned notation, we shall show that

$$\beta[[\beta(x, m)+\beta(y, j+l)]*(d-c), l] = \beta[[x+\beta(y, j+l)]*(d-c), l]. \tag{4}$$

We have

$$\begin{aligned}[x+\beta(y, j+l)]*(d-c) &= [0.1x_2 \dots x_t'c+0.\underbrace{0 \dots 0}_{t-j-l}y_{t-j-l+1}\dots y_t'q]*(n+j)\\ &= a_0.a_1 \dots a_{t-m}z_{t-m+1} \dots z_t'c*(n+j)\\ &= 0.\underbrace{0 \dots 0}_{n+j-1}a_0 \dots a_{t-m}z_{t-m+1} \dots z_{t-n-j}'d,\end{aligned}$$

where $a_i = (x_i+p_i) \pmod 2$ for $i = t-m, t-m-1, \dots, 1$; p_i are defined similarly as in (1.2) and (1.3); and $a_0 = p_0$. From the last equality and (3.2) we obtain formula (4).

Similarly as (4), we can prove the formula

$$\beta[\beta(x, d-c)+\beta(y, j), d-c] = \beta[x+\beta(y, j), d-c]. \tag{5}$$

Now we shall verify that

$$\begin{aligned}\beta[\beta(x, l-(d-c))+\beta(y, l-j), l]&\\ = \beta[\beta(x, l-(d-c)+w)+\beta(y, l-j+w), l],&\end{aligned} \tag{6}$$

where $w \geqslant 0$. This formula follows immediately from the following two equalities:

$$\beta(x, l-(d-c))+\beta(y, l-j) = 0.\underbrace{0 \dots 0}_{t-l+n+j-1}p_{t-l+n+j}z_{t+n-l+j+1} \dots z_t'c,$$

$$\beta(x, l-(d-c)+w)+\beta(y, l-j+w) = 0.\underbrace{0 \dots 0}_{t-l+n+j-w-1}p_{t-l+n+j-w}z_{t-l+n+j-w+1} \dots z_t'c.$$

Now we shall prove the following property:

if $x = 0.x_1 \dots x_t \dots x_{t+k}'c$, $y = 0.y_1 \dots y_t'q$ and $c \leqslant q \leqslant c+k+l$, then

$$\beta(x*k+y, l)+\beta(x, k) = \beta(x+y, k+l). \tag{7}$$

Let $q \geqslant c+k$ and $m = q-c-k = n-k$. Then

$$\begin{aligned}\beta(x*k+y, l)+\beta(x, k) &= \beta(\underbrace{0.0 \dots 0}_{k}x_1 \dots x_t'(c+k)+0.y_1 \dots y_t'q, l)+\beta(x, k)\\ &= \beta(z_{-n} \dots z_0.z_1 \dots z_t \dots z_{t+k}'(c+k), l)+\underbrace{0.0 \dots 0}_{t}x_{t+1} \dots x_{t+k}'c\\ &= 0.\underbrace{0 \dots 0}_{t+k-l}z_{t+k-l+1} \dots z_{t+k}x_{t+1} \dots x_{t+k}'c = \beta(x+y, k+l),\end{aligned}$$

where $z_i = x_{i-k}$ for $i = t+k, t+k-1, \dots, t-m+1$; $p_{t-m} = 0$; $z_i = (x_{i-k}+y_{i+m}+p_i) \pmod 2$ for $i = t-m, t-m-1, \dots, k+1$ and $z_i =$

$(y_{i+m}+p_i)(\mathrm{mod}\ 2)$ for $i = k, k-1, \ldots, 1-m$; $z_{-m} = p_{-m}$ and p_i are defined similarly as in (1.2) and (1.3).

Analogously we prove (7) if $c+k > q$.

Let m, l be positive integers, $k = q-c-m$ and $l \geqslant k$. Then

$$(8)\quad \beta(x, m)+\beta[\beta(y, l-k)+\beta(x*m, l), l-k] \geqslant \beta[\beta(y, l-k)+\beta(x, l+m), l+m].$$

Let E_1 and E_2 denote the left-hand and right-hand sides, respectively, of this inequality. We have

$$E_1 = x_{t-m+1} \ldots x_t'c+\beta[y_{t-l+k+1} \ldots y_t'q+a_{t-l+1} \ldots a_t'(c+m), l-k],$$

where $a_i = x_{i-m}$ for $i = t, t-1, \ldots, t-l+1$.

Hence, if $q \leqslant c+m$, then $k \leqslant 0$ and

$$E_1 = x_{t-m+1} \ldots x_t'c+\beta[b_{t-l+k} \ldots b_t'q, l-k]$$

$$= x_{t-m+1} \ldots x_t'c+b_{t-l+k+1} \ldots b_t'q = z_{t-l+k} \ldots z_t x_{t-(q-c)+1} \ldots x_t'c,$$

where $b_i = y_i$ for $i = t, t-1, \ldots, t+k+1$; $r_{t+k} = 0$; $b_i = (y_i+x_{i-m-k}+r_i)$ $(\mathrm{mod}\ 2)$ for $i = t+k, \ldots, t+k-l+1$; $b_{t-l+k} = r_{t-l+k}$; $p_t = 0$; $z_i = (y_i+x_{i-(q-c)}+p_i)\ (\mathrm{mod}\ 2)$ for $i = t, t-1, \ldots, t-l+k+1$; and $z_{t-l+k} = p_{t-l+k}$.

We have

$$E_2 = \beta[y_{t-l+k+1} \ldots y_t'q+x_{t-l-m+1} \ldots x_t'c, l+m]$$
$$= \beta[z_{t-l+k} \ldots z_t x_{t-(q-c)+1} \ldots x_t'c, l+m]$$
$$= z_{t-l+k+1} \ldots z_t x_{t-(q-c)+1} \ldots x_t'c.$$

Thus, $E_1 \geqslant E_2$. Analogously we can verify (8) if $q > c+m$.

Analogously as (8) we can prove that if $k = q-c+m$, then

$$(8.1)\quad \beta(y, m)+\beta[\beta(y*m, l)+\beta(x, l+k), l+k] \geqslant \beta[\beta(y, m+l)+\beta(x, l+k), l+k].$$

Using (8) and (8.1), we show that if d denotes the exponent of $\gamma(x+y)$ and $d \geqslant d_1 \geqslant c$, $d \geqslant d_2 \geqslant q$, then

$$(9)\quad \beta(x, d_1-c)+\beta(y, d_2-q)+ +\beta[\beta(x*(d_1-c), d-d_1)+\beta(y*(d_2-q), d-d_2), d-c] \geqslant \beta[\beta(x, d-c)+\beta(y, d-q), d-c] = \beta(x+y, d-c).$$

Let E_1 denote the left-hand side of this inequality. Substituting $m = d_1-c$, $y = y*(d_2-q)$, $l-k = d-d_2$, $l = d-d_1$, in (8), we obtain

$$E_1 \geqslant \beta(y, d_2-q)+\beta[\beta(x, d-c)+\beta(y*(d_2-q), d-d_2), d-c].$$

Applying (8.1) to the right-hand side of this inequality and substituting $m = d_2 - q$ and $l = d - d_2$, we get

$$E_1 \geqslant \beta[\beta(x, d-c)+\beta(y, d-q), d-c].$$

Thus, inequality (9) is proved. The second part of formula (9) follows from (5).

INSTITUTE OF MATHEMATICS
UNIVERSITY OF GDAŃSK

Received on 10. 3. 1971;
revised version on 11. 9. 1971

A. SCHURMANN (Sopot)

MINIMALNY BŁĄD SUMOWANIA DODATNICH LICZB ZMIENNOPOZYCYJNYCH

STRESZCZENIE

W pracy rozważa się zależność błędu sumowania n liczb zmiennopozycyjnych binarnych od kolejności dodawania tych liczb. Niech ALN oznacza dodawanie n liczb ze zbioru M_n według następującej kolejności:

1. Dodać dwie najmniejsze liczby a oraz b ze zbioru M_n, znormalizować i zaokrąglić wynik S_n.

2. Niech M_{n-1} składa się z wszystkich liczb z wyjątkiem liczb a i b ze zbioru M_n i liczby S_n. Zmniejszyć indeks n o 1. Jeżeli $n \neq 1$, to wrócić do 1.

Podano (§2) dwa przykłady, które pokazują, że błąd sumowania według kolejności ALN może być większy od minimalnego.

W §3 dowodzi się głównego twierdzenia tej pracy, które mówi, że jeżeli liczby ze zbioru M_n mają różne cechy, to dodawanie tych liczb według kolejności ALN daje maksymalną dokładność sumy. Przez *błąd dodawania* rozumie się w podanym twierdzeniu obcięcie najmniej znaczących bitów sumy.
