ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
XIV, 3 (1974)

M. M. SYSŁO (Wrocław)

REMARKS ON ADDITION PROCESSES OF POSITIVE FLOATING-POINT NUMBERS

1. This short note contains a proof of the following theorem:

The minimum of the upper bound of the absolute rounding error of a sum of positive floating-point numbers is attained for the algorithm of summation which in section 2 is called the algorithm ALN.

2. Paper [3] has dealt with the analysis of summation algorithms for the sum $A = a_1 + a_2 + \ldots + a_n$ of positive binary floating-point numbers to obtain an algorithm giving the sum A with a minimal error. There were considered the following summation sequences for the sum A:

the NS-sequence $\left(\ldots((a_1 + a_2) + a_3) + \ldots + a_n\right)$ and

the ALN-sequence obtained by the following operations:

Add the two least numbers and insert the result among the remaining ones. Repeat this step $n-1$ times.

The main theorem of paper [3] states that if all exponents of the numbers $a_1, a_2, \ldots, a_n$ are different, then the addition of these numbers according to the summation sequence ALN causes a minimal truncation error.

Next, it is assumed that there exists a constant d such that the rounding error $\alpha(a+b)$ of the sum $a+b$ is not greater than $d(a+b)$, i.e.

$$\alpha(a+b) \leqslant d(a+b).$$

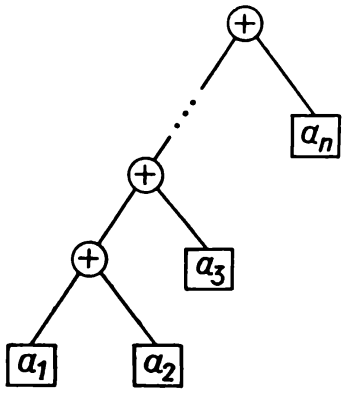
Hence, for the NS-sequence of summation we have

$$\alpha(A) \leqslant d\left[(n-1)a_1 + \sum_{i=2}^{n}(n-i+1)a_i\right]$$

and it is easy to see that this upper bound is minimal for a non-decreasing sequence $a_1, a_2, \ldots, a_n$.

Now we prove that the minimum of the upper bound of the absolute rounding error of the sum A is attained by the ALN-sequence of summation. Binary trees (see [1] and [2]) will be used in the proof.

Every way of computation of the sum A can be represented by a binary tree with n terminal nodes which correspond to the numbers $a_1, a_2, \ldots, a_n$ and with $n-1$ internal nodes which correspond to the summing operation. For example, the binary tree for the NS-sequence of summation is the following:



Let l_i $(i = 1, 2, \ldots, n)$ denote the length of the path (the number of tree arcs) from the terminal node i to the root of the tree. For the NS-sequence of summation we have $l_1 = n-1$, $l_i = n-i+1$ $(i = 2, 3, \ldots, n)$. It is easy to see that the maximal error for an arbitrary way of computation of the sum A is equal to

$$\sum_{i=1}^{n} l_i a_i. \tag{1}$$

Hence, the minimization of the maximal error of the sum A is equivalent to the minimization of the sum (1) over the sequences $l = \{l_1, l_2, \ldots, l_n\}$ which correspond to binary trees with n terminal nodes. (Paper [1] contains the conditions which must be fulfilled by the sequence l.) The solution of this problem is well known ([1] and [2]). The sum (1) is minimized by the sequence l^* which corresponds to Huffman's tree and, in turn, to the ALN-sequence of summation.

References

[1] T. C. Hu and A. C. Tucker, *Optimal computer search tree and variable-length alphabetical codes*, SIAM J. Appl. Math. 21 (1971), p. 514-532.

[2] D. E. Knuth, *The art of computer programming*, Vol. 1, Addison-Wesley, New York 1968, p. 399-415.

[3] A. Szurman, *On the minimum error in addition processes of positive floating-point numbers*, Zastosow. Matem. 13 (1973), p. 351-366.

DEPT. OF NUMERICAL METHODS
UNIVERSITY OF WROCŁAW
50-384 WROCŁAW

Received on 20. 7. 1973

M. M. SYSŁO (Wrocław)

UWAGI O SUMOWANIU CIĄGU ZMIENNOPOZYCYJNYCH LICZB DODATNICH

STRESZCZENIE

W notce przedstawiony jest bardzo prosty dowód twierdzenia, że minimum maksymalnego błędu zaokrąglenia sumy dodatnich liczb zmiennopozycyjnych jest osiągnięte dla kolejności ALN obliczania sumy.