Compensation for Truncation Errors in Accumulative Addition Using Random Methods

The Problem

In simple numerical integration or other forms of accumulative addition, sufficient accuracy frequently requires retention of only the most significant digits of the final result. Yet, to avoid truncation errors, all digits of the intermediate sums must usually be retained, a requirement which often means using multiple precision storage of these intermediate sums. This correspondence presents a method for performing accumulation while eliminating the multiple precision storage. The method, which is new to the computer literature, is based on a statistically favorable use of random numbers and is optimum under some circumstances.

Manuscript received March 24, 1966; revised April 18, 1967.

The Method

Basically, the method may be described as follows. Whenever a summation is performed, the result is truncated to the desired accuracy. When the result is to be used in the next summation, the truncated information is simulated by entering in its place a suitably chosen random number.

In the more explicit discussion to follow, only digits to the left of the decimal point are desired in the final answer. Letting n_j represent the jth number to be accumulated, then

$$S_i = \sum_{j=1}^{i} n_j$$

represents the usual accumulated total of the first i numbers. This can also be expressed as

$$S_i = L_i + R_i$$
$$= S_{i-1} + n_i$$
$$= L_{i-1} + R_{i-1} + n_i \quad \text{(in alternative form)}$$

where L_i and R_i are the left-hand and right-hand parts, respectively, of the sum.

An analogous random method is now defined in which S_i, L_i, and R_i are replaced respectively by Σ_i, Λ_i, and P_i. This alternate summation is

$$\Sigma_i = \Lambda_i + P_i$$
$$= (\Lambda_{i-1} + \rho_i) + n_i \quad \text{(in alternative form)}$$

where ρ_i is a random number of the same sign as Λ_{i-1} (or Σ_i or ρ_{i-1}) and whose magnitude is uniformly distributed over the range from zero to one. If Λ_0 is set to zero, the resulting Λ_i has an expected value equal to S_i.

The method does not eliminate the multiple precision computation, but only multiple precision storage. The real savings accrue when many separate accumulations are being performed simultaneously. The additional complexity and computation time of random number generation can be offset by using the same random number in each of the accumulations. In a program on a PDP-5 computer, the method required only six additional words of program storage (including a pseudo-random-number generator), over the storage needed for conventional double precision accumulation. This means that a net saving of storage is realized if more than six simultaneous multiple precision accumulative additions or integrations are to be carried out.

List of Proved Results

The author has proved the following results.

1) There is no general deterministic method of accumulative addition which can code stored intermediate results more efficiently than conventional addition. There is no method of round-off or recoding which will give accurate accumulative addition that does not have storage requirements equal to or exceeding that of conventional full precision arithmetic.

2) The probabilistic method works. The expected value of Λ_i is equal to the sum of the first i values of n_j.

3) The variance of Λ_i:

$$Var(\Lambda_i) = \sum_{j=1}^{i} r_j(1 - r_j)$$

where r_j is the right-hand part of n_j.

4) This implies that if n_j are all positive (or all negative),

$$Var(\Lambda_i) = \sum_{j=1}^{i} r_j(1 - r_j) \leq \sum_{j=1}^{i} |r_j| \leq \sum_{j=1}^{i} |n_j|$$

Then,

$$StandardDeviation(\Lambda_i) \leq \sqrt{\sum_{j=1}^{i} |n_j|}$$

Using Standard Deviation as the measure of accuracy, this implies that for n_j all positive, even under worst conditions where $n_i \ll 1$, only the least significant half of the digits stored will be inaccurate. This inaccuracy may be an arbitrarily small percentage of the sum, depending on the number of digits stored.

5) The scheme is optimum in the minimax sense in that no round-off scheme which permits a given amount of storage and a given maximum range will give a smaller worst-case change in variance per addition.

Comment

The method is practical and can give good results in many circumstances. It represents a practical application of the theory of probabilistic finite automata as developed by M. Rabin.[1] Because the results can be related to the deterministic methods by the above theorems, the method defines one set of bounds to the applicability of probabilistic methods.

After this work was completed, the author was informed of unpublished work on this method by E. Fredkin of Information International, Inc. The method, which was being considered for use in the floating point package for the PDP-1 computer, appears to date well back in the statistical literature. An analogous method has also been proposed for compensating the truncation which is due to analog-to-digital conversion.[2]

Edgar H. Bristol
The Foxboro Company
Foxboro, Mass. 02035

1. M. O. Rabin, "Probabilistic automata," *Information and Control, vol. 5, pp. 230-245,* September 1963.

2. G. A. Korn, *Random-Process Simulation and Measurement.* New York: McGraw-Hill, 1966, pp. 136-139.