

AD708691

AN ALGORITHM FOR FLOATING-POINT ACCUMULATION
OF SUMS WITH SMALL RELATIVE ERROR

BY

MICHAEL MALCOLM

STAN-CS-70-163
JUNE 1970

Reproduced by the
CLEARINGHOUSE
for Federal Scientific & Technical
Information Springfield Va. 22151

This document has been approved
for public release and sale; its
distribution is unlimited.

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY



AN ALGORITHM FOR FLOATING-POINT ACCUMULATION OF SUMS
WITH SMALL RELATIVE ERROR

by
Michael Malcolm

Reproduction in whole or in part is permitted
for any purpose of the United States Government.

The preparation of this report was sponsored by the
Office of Naval Research under grant number N0013-67-A-
0112-0029, the National Science Foundation under grant
number NSF GJ 408 and the Atomic Energy Commission under
grant number AT (04-3) 326,PA 30.

I. Introduction

Many algorithms require the calculation of a sum

$$s = \sum_{i=1}^n x_i, \quad n \geq 3,$$

where $x_1, x_2, \dots, x_n$ are numbers represented in floating point. In practice, an approximate sum $\hat{s}$ is computed with rounding errors.

Wilkinson [1] shows that if the sum is accumulated in a single-precision accumulator (using floating binary arithmetic with t bits of precision and proper rounding), then

$$\hat{s} - s = \sum_{i=1}^n x_i \eta_i,$$

where

$$(1 - 2^{-t})^{n+1-r} \leq 1 + \eta_r \leq (1 + 2^{-t})^{n+1-r} \quad (r = 1, \dots, n).$$

Thus the error bound is dependent on the order of summation. This result has led to the well-known rule of thumb that it is usually best to add a list of numbers in order of increasing magnitude. If one has a priori knowledge of the x_i (e.g., $\sum |x_i| < 1$) or if the accumulation is performed with more precision (say double precision), then much smaller error bounds can be found. However, as Wilkinson points out, "It should be emphasized that we still cannot guarantee that an accumulated sum ... has a low relative error."

Large relative error in an accumulated sum is often the result of a phenomenon which Professor D. H. Lehmer calls catastrophic cancellation. This occurs when an intermediate partial sum is much larger in magnitude than the final sum. Then one or more additions result in a loss of

significant digits. The post-normalization step of a subsequent addition thus introduces zeros in place of significant digits. However, as Professor William Kahan has observed, this large cancellation is not the cause of the error -- it merely reveals the error. That is, the real villain here is not the cancellation, but rather the large intermediate sums within a floating-point system of given precision. Perhaps "catastrophic loss of precision" would be a more appropriate name. Catastrophic cancellation is fairly common with poorly designed algorithms; most good algorithms have built-in precautions which avoid (or usually avoid) this phenomenon.

Large relative errors can occur without catastrophic cancellation. This happens in large summations ($n \gg 3$) where the intermediate sums become much larger in magnitude than the individual addends, but not larger than the final sum. This sort of error can occur in numerical integration using a large number of intervals. Wolfe [2] proposed a technique for avoiding this type of error. It is described in the following section.

In the remainder of this report, a modification of Wolfe's algorithm is presented, followed by a detailed error analysis. This algorithm has the advantage that the final sum is guaranteed to have a very small relative error.

II. Extended Summation With Cascading Accumulators

Wolfe [2] suggests a technique which is easily programmed and requires only a small number of additional storage locations. These extra locations, called cascading accumulators, are denoted by $s_1, s_2, \dots$. The separate accumulators hold sums that are in various intervals; for example,

$$1.000 \leq c(s_1) \leq 9.999$$

$$10.00 \leq c(s_2) \leq 99.99$$

$$100.0 \leq c(s_3) \leq 999.9$$

$$\vdots \quad \quad \quad \vdots$$

where $c(s_i)$ denotes the contents of s_i . The summing is done at the lowest level accumulator (s_1) until it is about to overflow. At that point it is added to the next accumulator (s_2) and reset to zero. Similarly, if s_2 is about to overflow, it is added to s_3 and reset to zero, and so on.

By this technique the intermediate sums never become much larger than the addends. However, catastrophic cancellation can occur just as before. Wolfe does not discuss how to go about summing the accumulators at the end; in an example he uses the order of increasing magnitude. For certain problems, this is a useful technique; however, there is no guarantee that the final result has a small relative error.

III. A Modification of Wolfe's Algorithm

The following algorithm requires little if any more execution time than the algorithm of the last section, and nearly full-precision accuracy is achieved, provided exponent underflow or overflow do not occur. Such exceptional conditions are normally brought to the attention of the user by the system software and, if so, inaccurate results cannot go unnoticed. As in Wolfe's algorithm, additional intermediate accumulators are used -- typically fewer than 50.

The following discussion assumes the algorithm is implemented on a machine using a floating-point number system F of base β (usually β is 2, 8, 10 or 16) with a t -digit mantissa. The exponent e is assumed to lie in the range

$$-m \leq e \leq M .$$

Thus each nonzero $x \in F$ has the normalized representation

$$x = \pm .d_1 d_2 \dots d_t \cdot \beta^e , \quad (1)$$

where $d_1, \dots, d_t$ are integers satisfying

$$1 \leq d_1 \leq \beta - 1 ,$$

$$0 \leq d_i \leq \beta - 1 \quad (i = 2, \dots, t) .$$

The number 0 belongs to F , and has the structure

$$0 = .00 \dots 0 \cdot \beta^{-m} .$$

All floating-point addition is assumed to be normalized. The machine may do either proper rounding or truncation (chopping).

To facilitate discussion, the function lev (similar to that used by Møller [5]) is defined as follows: If $x \in F$ then $lev(x) = e + m$. lev is the biased exponent having the mnemonic "level". Note that lev is a function of the representation of a number and not the number itself. For example, suppose x is to be added to y , where $|y| > |x|$, and that x must be unnormalized during operand alignment. Suppose also that no nonzero digits are lost from the mantissa of x while it is being

unnormalized. If we denote the unnormalized representation of x by $\hat{x}$, then x and $\hat{x}$ both represent the same real number exactly, but $\text{lev}(\hat{x}) > \text{lev}(x)$.

The algorithm for computing $\sum_{i=1}^n x_i$ can now be described as follows. There are two positive parameters, l and η :

Assume there are $\eta+1$ accumulators, the contents of which are denoted by $\alpha_0, \alpha_1, \dots, \alpha_\eta$.

1. Set each of the accumulators to zero.
2. For each x_i form $a_{i1}, a_{i2}, \dots, a_{iq}$ ($q \geq 1$), where $a_{i1} + a_{i2} + \dots + a_{iq} = x_i$ and each a_{ij} has the property that the last l digits are 0 (i.e., $d_{t-l+1} = d_{t-l} = \dots = d_t = 0$).
3. Each a_{ij} is added to the k -th accumulator, where k is determined by

$$vk \leq \text{lev}(a_{ij}) \leq vk + v - 1, \quad (2)$$

$$v = \lceil (M+m+1)/(\eta+1) \rceil$$

where $\lceil \xi \rceil$ denotes the smallest integer not less than ξ .

(Thus

$$k = \text{lev}(a_{ij}) \div v, \quad (3)$$

in the sense of Algol 60.)

4. The accumulators are summed in decreasing order (i.e., $\eta, \eta-1, \dots, 0$).

The second step appears, at first sight, to be quite complicated. However, in practice it is easily done, especially on a machine with double-precision arithmetic. An illustration of this in Fortran for the IBM System/360 is contained in Section VI.

The parameters l and η are chosen so that the addition in step 3 retains all the significant digits involved. That is, until step 4, there are no rounding errors. More insight into choosing l and η will be given in the following section. Also, an important restriction on the magnitude of the product q_n will be revealed.

Step number 4 is certainly the most interesting step of the algorithm. If, instead, the accumulators are summed in increasing order (as one is tempted to do after reading Wilkinson [1]), catastrophic cancellation can occur. When this algorithm is incorporated in an innerproduct routine, it often happens that

$$\begin{aligned}\alpha_{\eta} &= 0 \\ &\vdots \\ \alpha_{k+1} &= 0 \\ \alpha_k &= -B \\ \alpha_{k-1} &= B \\ \alpha_{k-2} &= 0 \\ \alpha_{k-3} &= 0 \\ \alpha_{k-4} &= \epsilon_1 \\ \alpha_{k-5} &= \epsilon_2 \\ &\vdots\end{aligned}$$

where $\text{lev}(B) - \text{lev}(\varepsilon_1) > t$. Summing in increasing order will yield 0. However, as D. Jordan pointed out in [4], summing the accumulators in decreasing order $(\eta, \dots, 0)$ precludes the chance of this type of error. The remaining question is: Does summing the accumulators in decreasing order lead to some other case where a large relative roundoff error can occur? The answer to this question is no. Proof of this assertion and a sharp bound on the roundoff error are given in the next section.

IV. Error Analysis

Another convenient function is defined as follows: Let $x \in F$ be an approximation of some real number x^* . If $x = x^*$ and $x^* \neq 0$, then $\text{pad}(x, x^*)$ is defined to be the number of digits by which the mantissa of x can be shifted to the right before a significant digit is lost (i.e., before a non-zero digit is shifted out of the low-order position). If $x \neq x^*$, then $\text{pad}(x, x^*)$ is negative, and defined as follows: suppose x has the representation (1); if there exists a ξ such that ξ can be represented as

$$\xi = \pm .\hat{a}_1\hat{a}_2\dots\hat{a}_T \cdot \beta^{e-t} \quad \text{with } \hat{a}_T \neq 0, \hat{a}_1 \neq 0$$

and $x + \xi = x^*$ and T is finite, then $\text{pad}(x, x^*)$ is defined to be $-T$.

Otherwise, $\text{pad}(x, x^*)$ is defined as $-\infty$. For completeness,

$\text{pad}(0, 0) = \infty$. For example, if $\beta = 2$ and $t = 6$, $\text{pad}(-.101000 \cdot 2^3, -5_{10}) = 3$.

If $x = +.111111 \cdot 2^0$ and $y = +.111111 \cdot 2^1$, and $\oplus$ represents floating-point addition, and $\text{pad}(x \oplus y, x + y) = -2$ since two digits are lost during the floating-point addition. When $\text{pad}(x, x^*)$ is positive, the mantissa of x has a "padding" of zero digits at the end.

It follows that

$$\text{pad}(x, x^*) \geq 0 \Leftrightarrow x = x^*,$$

$$\text{pad}(x, x^*) < 0 \Leftrightarrow x \neq x^*,$$

$$\text{pad}(x, x^*) \geq t \Rightarrow \text{pad}(x, x^*) = \infty \Leftrightarrow x = x^* = 0.$$

In step 2 of the algorithm, it is required that $\text{pad}(a_{ij}, a_{ij}^*) \geq l > 0$, for all i, j .

It is also expedient to define

$$\rho(x, x^*) = \text{lev}(x) + \text{pad}(x, x^*) \quad (4)$$

$\rho(x, x^*)$ is invariant with respect to operand alignment (un-normalization) and post-normalization of x , provided no exponent underflow or overflow occurs.

Lemma 1: If x and y are two floating-point numbers and $\oplus$ represents floating-point addition, then

$$\rho(x \oplus y, x^* + y^*) \geq \min\{\rho(x, x^*), \rho(y, y^*)\},$$

provided no exponent underflow or overflow occurs.

Proof: Assume $\text{lev}(x) \geq \text{lev}(y)$. Let z denote an accumulator, a floating-point number with a $t+2$ digit mantissa and an overflow digit. Set $z \leftarrow y$ and, if necessary, unnormalize z so that $\text{lev}(z) = \text{lev}(x)$. The accumulator z can be treated as a floating-point number if one ignores the overflow digit and

considers only the first t digits of the mantissa. In this way, pad is defined for z . Let w denote another accumulator with the same structure as z . Set $w \leftarrow z+x$. Prior to the post-normalization step in forming w ,

$$\text{lev}(w) = \text{lev}(z) = \text{lev}(x) \quad \text{and}$$

$$\text{pad}(w, x^* + y^*) \geq \min\{\text{pad}(z, y^*), \text{pad}(x, x^*)\} ,$$

and equality occurs whenever the low-order digits of x and y don't cancel. From Equation (4) and the fact that $\rho(w, x^* + y^*)$ remains unchanged during the post-normalization step, it follows that

$$\rho(x \oplus y, x^* + y^*) = \rho(w, x^* + y^*) \geq \min\{\rho(z, y^*), \rho(x, x^*)\} .$$

Since $\rho(z, y^*) = \rho(y, y^*)$,

$$\rho(x \oplus y, x^* + y^*) \geq \min\{\rho(x, x^*), \rho(y, y^*)\} .$$

Lemma 2: In the notation of Section III,

$$\rho(\alpha_k, \alpha_k^*) \geq vk + 1.$$

Proof: Any term (y) that is added to the k -th accumulator satisfies

$$\text{lev}(y) \geq vk \quad \text{and} \quad \text{pad}(y, y) \geq 1.$$

By Lemma 1, Equation (4) and the fact that $\rho(0,0) = \infty$, the lemma follows by induction.

Lemma 3: If $\alpha_k \neq 0$, and N is the number of a_{ij} added to the accumulators, then

$$\begin{aligned} vk \leq \text{lev}(\alpha_k) &\leq v(k+1) - 1, & \text{if } N = 1 \\ vk - t + l + 1 \leq \text{lev}(\alpha_k) &\leq v(k+1) + \lfloor \log_{\beta}(N-1) \rfloor, & \text{if } N > 1 \end{aligned}$$

$$(k = 0, 1, \dots, \eta)$$

where $\lfloor \xi \rfloor$ denotes the largest integer not greater than ξ .

Proof: The inequalities for $N=1$ are obviously the same as those satisfied by a single term added to the k -th accumulator. The upper bound for $N > 1$ is found by considering the largest number (ξ) which can be added to the k -th accumulator, i.e.,

$$\xi = +.zz\dots z00\dots 0 \cdot \beta^{v(k+1)-1-m}, \quad (z = \beta-1)$$

and observing the $\text{lev}(\alpha_k)$ during repeated additions of ξ to α_k .

If the lower bound for $N > 1$ were not true, i.e., if

$$\text{lev}(\alpha_k) < vk - t + l + 1 ,$$

then, by Lemma 2,

$$\text{pad}(\alpha_k, \alpha_k^*) \geq t \Rightarrow \alpha_k = 0 ,$$

which is a contradiction.

Lemma 4: If $N \leq \beta^{l-v+1}$, then each of the α_k ($k = 0, 1, \dots, \eta$) is exact.

Proof: By Lemma 3,

$$\text{lev}(\alpha_k) < vk + l + 1 .$$

Combining this with Equation (4) and Lemma 2 gives

$$\text{pad}(\alpha_k, \alpha_k^*) \geq 0 ,$$

which, by the definition of pad , implies α_k is exact.

Loss of precision in an extended summation can result from either

1. repeated truncations (roundings) of the sum, or
2. post-normalization left shift of the approximate sum.

The post-normalization error can be formalized as follows: Let the accumulation of the floating-point sum

$$\psi_n = \sum_{i=1}^n x_i ,$$

where the x_i ($i = 1, 2, \dots, n$) are exact, be defined as

$$\psi_0 = 0$$

$$\psi_i = \psi_{i-1} \oplus x_i, \quad (i = 1, 2, \dots, n),$$

The function Δ of two floating-point variables is defined as

$$\Delta(x, y) = \max\{\text{lev}(x), \text{lev}(y)\} - \text{lev}(x \oplus y).$$

Thus, during post-normalization of the floating-point sum $x \oplus y$, the mantissa undergoes a left shift of $\Delta(x, y)$ digits. Clearly, if a carry occurs, $\Delta(x, y) = -1$. Also, $\Delta(x, y) > 1$ only if $|\text{lev}(x) - \text{lev}(y)| \leq 1$.

During the formation of $\psi_i = \psi_{i-1} \oplus x_i$, any truncation error already present in the low order digits of ψ_{i-1} is multiplied by $\beta^{\Delta(\psi_{i-1}, x_i)}$.

The accumulators are summed in decreasing order. Thus, the sum S_0 can be defined by

$$S_{\eta+1} = 0$$

$$S_k = S_{k+1} \oplus \alpha_k \quad (k = \eta, \eta-1, \dots, 0). \quad (5)$$

Lemma 5: If S_{k+1} is exact and $\text{lev}(S_{k+1}) \leq \text{lev}(\alpha_k)$ and if S_{k+1} is un-normalized so that $\text{lev}(S_{k+1}) = \text{lev}(\alpha_k)$, then $\text{pad}(S_{k+1}, S_{k+1}^*) > 0$, provided $N \leq \beta^{l-v+1}$.

Proof: Lemma 1 and Equation (5) yield

$$\rho(S_{k+1}, S_{k+1}) \geq \min\{\rho(\alpha_{k+1}, \alpha_{k+1}), \rho(\alpha_{k+2}, \alpha_{k+2}), \dots, \rho(\alpha_\eta, \alpha_\eta)\},$$

which, with Lemma 2 and the definition of ρ , gives

$$\text{lev}(S_{k+1}) + \text{pad}(S_{k+1}, S_{k+1}) \geq v(k+1) + l.$$

Substituting $\text{lev}(\alpha_k)$ for $\text{lev}(S_{k+1})$ and using Lemma 3, we obtain the desired result:

$$\text{pad}(S_{k+1}, S_{k+1}^*) \geq l - \lfloor \log_B(N-1) \rfloor.$$

Suppose that, in the process of summing the accumulators as described by Equation (5), $k = j$ is the first k such that $\text{pad}(S_k, S_k^*) < 0$. As a result of Lemma 5, the truncation (rounding) error in adding S_{j+1} and α_j must be caused in one of three ways:

- i) When the operands are aligned, $\text{pad}(\alpha_j, \alpha_j^*) = 0$ and a carry occurs.
- ii) When the operands are aligned, $\text{pad}(S_{j+1}, S_{j+1}^*) = 0$ and a carry occurs.
- iii) $\text{lev}(S_{j+1}) > \text{lev}(\alpha_j)$ and, when α_j is aligned so that $\text{lev}(\alpha_j) = \text{lev}(S_{j+1})$, $\text{pad}(\alpha_j, \alpha_j^*) < -\Delta(S_{j+1}, \alpha_j) \leq 0$.

Lemma 6:

$$\text{lev}(S_j) \geq vj + l + 1.$$

Proof: Case i) In the aligned position, using Lemma 2, we find that

$$\text{lev}(\alpha_j) = \rho(\alpha_j, \alpha_j^*) \geq vj + l.$$

$$\text{Thus } \text{lev}(S_j) = \text{lev}(\alpha_j) + 1 > vj + l + 1.$$

$$\begin{aligned} \text{Case ii) } \text{lev}(S_{j+1}) = \rho(S_{j+1}, S_{j+1}^*) &\geq \min\{\rho(\alpha_{j+1}, \alpha_{j+1}^*), \dots, \rho(\alpha_\eta, \alpha_\eta^*)\} \\ &\geq vj + v + l. \end{aligned}$$

Therefore $\text{lev}(S_j) > vj + v + l \geq vj + l + 1$.

Case iii) When α_j is aligned, $\text{lev}(S_{j+1}) = \text{lev}(\alpha_j) > \rho(\alpha_j, \alpha_j^*) \geq vj + l$.

Now,

$$\begin{aligned} \text{lev}(S_j) &= \text{lev}(S_{j+1}) + \Delta(S_{j+1}, \alpha_j) \\ &\geq \text{lev}(S_{j+1}) > \text{lev}(\alpha_j) = vj + l. \end{aligned}$$

Thus,

$$\text{lev}(S_j) \geq vj + l + 1.$$

Since $\text{lev}(\alpha_{j-1}) \leq vj + \lfloor \log_\beta(N-1) \rfloor$,

$$\text{lev}(S_j) - \text{lev}(\alpha_{j-1}) \geq l + 1 - \lfloor \log_\beta(N-1) \rfloor.$$

The assumption in Lemma 4 (i.e., $N \leq \beta^{l-v+1}$) is sufficient to guarantee that $\text{lev}(S_j) - \text{lev}(\alpha_{j-1}) > 1$, from which it follows that $\Delta(S_j, \alpha_{j-1}) \leq 1$. Similarly, each of the subsequent additions can undergo a post-normalization left shift of at most one digit. In fact, at most one of the additions

$$S_k = S_{k+1} \oplus \alpha_k \quad (k = j-1, j-2, \dots, 0)$$

will undergo a post-normalization left shift of one digit.

Lemma 7: If $N \leq \beta^{l-v+1}$, the mantissa of each of the accumulators

$\alpha_{j-\lambda}, \alpha_{j-\lambda-1}, \dots, \alpha_0$ is shifted at least t digits during operand alignment, where

$$\lambda = \lceil (t+1)/v \rceil. \quad (6)$$

Proof: By Lemma 6,

$$\text{lev}(S_{j-i}) \geq vj + l, \quad (i = 0, 1, \dots, j).$$

By Lemma 3,

$$\text{lev}(\alpha_{j-i}) \leq v(j-i) + l + 1, \quad (i = 0, 1, \dots, j),$$

$$\text{and } \text{lev}(S_{j-i+1}) - \text{lev}(\alpha_{j-i}) \geq vi - 1 \geq t, \quad (i = \lambda, \lambda+1, \dots, j).$$

Thus the mantissa of each of the accumulators $\alpha_{j-\lambda}, \alpha_{j-\lambda-1}, \dots, \alpha_0$ is shifted at least t digits during operand alignment.

Theorem 1: If $N \leq \beta^{l-v+1}$, if the accumulator used in accumulating S_0 has at least $t+1$ digits, and if no underflow or overflow occurs, then the absolute error in S_0 is bounded by

$$|s - S_0| \leq \lambda \delta \beta^{\text{lev}(S_0) - m - t + 1},$$

where

$$\delta = \begin{cases} 1 & \text{for chopped arithmetic} \\ \frac{1}{2} & \text{for rounded arithmetic} \end{cases}$$

and λ is given by Equation (6).

Proof: Since a post-normalization left shift of at most one digit can occur only once while the accumulators are summed, the worst case occurs when it is caused by the addition of $\alpha_{j-\lambda}$ (see Lemma 7). Subsequent additions of $\alpha_{j-\lambda-1}, \alpha_{j-\lambda-2}, \dots$ cannot affect the computed value of S_0 (see Knuth's [6] discussion of problem 5, page 498). Prior to the addition of $\alpha_{j-\lambda}$, a maximum of λ truncations (roundings) can occur, each resulting in an error of $\beta^{\text{lev}(S_0) - m - t}$ or less.

Q.E.D.

For machines which use t -digit accumulators and chopped arithmetic, the error bound is $(\lambda-1)\beta^{\text{lev}(S_0)-m-t+1}$ (a stronger result!). Note that, although the above theorem gives a bound on the absolute error, it also provides a bound on the relative error. Specifically, if the true value of the sum s is zero, then $S_0 = 0$.

Theorem 2: If $N \leq \beta^{l-v+1}$ and no underflow or overflow occurs, then

$$s = S_0(1+\epsilon) ,$$

where

$$|\epsilon| \leq \lambda \delta \beta^{2-t} .$$

Proof: i) If $S_0 = 0$, then, since total cancellation of significant digits cannot occur in summing the accumulators, $s = 0$.

ii) If $S_0 \neq 0$, then assume $s - S_0 = S_0 \epsilon$. By Theorem 1,

$$|s - S_0| = |S_0| |\epsilon| \leq \lambda \delta \beta^{\text{lev}(S_0)-m-t+1} .$$

Since $|S_0| \geq \beta^{\text{lev}(S_0)-m-1}$,

$$|\epsilon| \leq \lambda \delta \beta^{2-t} .$$

These theoretical results are substantiated by an experiment reported by D. Jordan [4]. Jordan used this technique for accumulating innerproducts on an IBM 360 ($\beta = 16$, $t = 14$). He chose $\eta = 32$, $l = 6$ and $q = 2$, and states:

"Empirical tests were run to determine the small amount of roundoff that might be expected from the procedure. The tests used 1000 dot products of 15-component vectors where the components were randomly generated in the range $(-10^{30}, 10^{30})$. The results of this routine were checked against results obtained using 256 hex-digit arithmetic through the multiple precision arithmetic package written by J. R. Ehrman of SLAC. Of the thousand cases, 467 were in exact agreement, 537 had an erroneous last bit and 3 had an erroneous penultimate bit." [4, p. 3].

If the last sentence in Jordan's statement were changed to read "... 467 were in exact agreement, 537 had an erroneous last (hexadecimal) digit and 3 had an erroneous penultimate digit.", then Jordan's results are consistent with those of Michael Saunders at Stanford University. Saunders performed several experiments on an IBM 360 using $\beta = 16$, $t = 14$, $\eta = 43$, $l = 6$ and $q = 2$. He found examples where the 13-th hexadecimal digit of the result was in error, but none with errors in the 12-th digit. Errors of this size are consistent with Theorem 2.

V. Additional Modifications to the Algorithm

If one desires the final floating-point result to be correct in all digits, the following procedure can be used immediately after calculating S_0 :

1. Form $a_1, a_2, \dots, a_q$ ($q \geq 1$), where $a_1 + a_2 + \dots + a_q = S_0$ and each a_i has the property that the last l digits of its mantissa are 0.
2. Add $-a_1, -a_2, \dots, -a_q$ to the accumulators.
3. Sum the accumulators in decreasing order. Call the result Δ .
4. $S_0 + \Delta$ is the full precision result.

In problems where N may get arbitrarily large (e.g., numerical integration), all is not lost. One merely increments an integer every time a term is added to the accumulators and when the integer is equal to $\beta^{l-v+1}-m(\eta+1)$, the following procedure is executed:

1. reset the integer to zero.
2. for $i := 0$ step 1 until η do
begin
 $a := \alpha_i; \alpha_i := 0;$
 $\text{addtoaccumulators}(a)$
end

where the procedure addtoaccumulators forms the $a_{i1}, \dots, a_{iq}$ variables, adds q to the integer and adds the a_{ij} ($j = 1, \dots, q$) to the accumulators.

3. Resume the original algorithm.

VI. Conclusion

The generality of the preceding discussion tends to obscure the simplicity of the algorithm. For this reason, a simple illustrative example of this technique programmed in Fortran for the IBM 360 is included:

```
      REAL FUNCTION SUM(X,N)
      EQUIVALENCE (IEQ,W)
      DIMENSION X(1)
      REAL*8 R(43),S8,DBLE
      DO 10 I=1,43
10    R(I)=0.0D0
      DO 20 I=1,N
      W=ABS(X(I))
      IEXP=IEQ/50331648 + 1
C 50331648 IS 3*(2**24) WHICH SHIFTS RIGHT 24 BITS AND
C DIVIDES BY 3. ABS GETS RID OF THE SIGN BIT.
20    R(IEXP)=R(IEXP) + DBLE(X(I))
      S8=0.0D0
      DO 30 I=1,43
30    S8=S8 + R(44-I)
      SUM=SNGL(S8)
      RETURN
      END
```

This subroutine finds the sum of a vector of short-precision ($t = 6$) numbers. Since the 360 used has long-precision ($t = 14$) floating-point hardware, it is convenient to use 14 digit accumulators and append 8 zero digits at the end of each x_i . Thus, $q = 1$ and $l = 8$. The value $\eta = 43$ was chosen to give $v = 3$. Thus, since $\beta = 16$, the short-precision result is guaranteed to have full-precision (chopped) accuracy provided $N \leq 16^6 = 16,777,216$.

This example is typical in that q is usually small (1 or 2), l is usually chosen for convenience, and η is usually chosen so the accumulators can be quickly indexed and so v is sufficiently small.

The algorithm is currently used in several innerproduct routines (see Malcolm [3] for descriptions of these routines, including running times). Since efficiency is satisfactory, it may well be feasible to implement this technique through a microprogram so that the programmer can specify by a certain operation code that a summation is to be performed with a set of these accumulators rather than with a single accumulator.

VII. Acknowledgment

The author is indebted to Professor George E. Forsythe for his helpful comments and criticisms of the manuscript. The author would also like to thank Michael Saunders for several enlightening discussions during the evolution of this algorithm and for his wealth of numerical counterexamples.

BIBLIOGRAPHY

- [1] Wilkinson, J. H., Rounding Errors in Algebraic Processes, Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1963.
- [2] Wolfe, J. M., "Reducing Truncation Errors by Programming," CACM, Vol. 7, No. 6, June 1964, 355-356.
- [3] Malcolm, M. A., "A Description and Comparison of Subroutines for Computing Euclidean Inner Products of Vectors," Technical Report (to appear), Computer Science Department, Stanford University, 1970.
- [4] Jordan, D. F., "ANL F154S - DOTP, Extra-Precision Accumulating Inner Product," Argonne National Laboratory, Applied Mathematics Division, System/360 Library Subroutine, Argonne, Illinois, November, 1967.
- [5] Møller, Ole, "Quasi Double-precision in Floating Point Addition," BIT, 5 (1965), 37-50.
- [6] Knuth, D. E., "Seminumerical Algorithms," The Art of Computer Programming, Vol. 2, Reading, Mass.: Addison-Wesley Publishing Co., 1969.

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Computer Science Department Stanford University Stanford, California 94305		2a. REPORT SECURITY CLASSIFICATION Unclassified	
		2b. GROUP - - -	
3. REPORT TITLE AN ALGORITHM FOR FLOATING-POINT ACCUMULATION OF SUMS WITH SMALL RELATIVE ERROR			
4. DESCRIPTIVE NOTES (Type of report and inclusive dates) Manuscript for Publication (Technical Report)			
5. AUTHOR(S) (First name, middle initial, last name) Michael Malcolm			
6. REPORT DATE June 1970		7a. TOTAL NO. OF PAGES 22	7b. NO. OF REFS 6
8a. CONTRACT OR GRANT NO N00014-67-A-0112-0029		9a. ORIGINATOR'S REPORT NUMBER(S) STAN-CS-70-163	
b. PROJECT NO NR 044-211		9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) none	
c.			
d.			
10. DISTRIBUTION STATEMENT Releasable without limitations on dissemination.			
11. SUPPLEMENTARY NOTES - - -		12. SPONSORING MILITARY ACTIVITY Office of Naval Research	
13. ABSTRACT A practical algorithm for floating-point accumulation is presented. Through the use of multiple accumulators, errors due to cancellation are avoided. An example in Fortran is included. An error analysis providing a sharp bound on the <u>relative</u> error is also given.			

DD FORM 1 NOV 66 1473

(PAGE 1)

S/N 0101-807-6801

Unclassified
Security Classification

Unclassified

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Floating point arithmetic error analysis						