# ON A MODEL FOR COMPUTING ROUND-OFF ERROR OF A SUM

BY

GEORGE B. DANTZIG

STAN-C S-70-156
MARCH 1970

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY

# ON A MODEL FOR COMPUTING ROUND-OFF ERROR OF A SUM

BY

GEORGE B. DANTZIG

March, 1970

Technical Report No. 156

Computer Science Department
Stanford University
Stanford,California

Reproduction and research of this report was partially supported by Office of Naval Research, Contract N00014-67-A0112-0011; U.S. Atomic Energy Commission, Contract AT[04-3] 326 PA #18; National Science Foundation, Grant GP 9329 and Grant GJ 320; U.S. Army Research Office, Contract DAHC 04-67-0028; and National Institutes of Health, Grant GM 14789-02.

Given real numbers $a_1, a_2, \ldots, a_n$ we are interested in the classic problem of the error in computing $S = \sum_1^n a_i$ when the sum is computed by $\tilde{S}_0 = \sum_1^n a_i^*$ where $a_i^*$ is the nearest integer to $a_i$. We shall first study this error as a function of a A shift, i.e., when all numbers $a_i$ are each shifted A and then rounded;

$$(1) \quad S - n\Delta = \sum_{i=1}^{n} (a_i - A)$$

$$(2) \quad \tilde{S}_\Delta - n\Delta = \sum_{i=1}^{n} (a_i - A)^*$$

We will then let A become a random variable that can take on uniformly any value in the interval $-\frac{1}{2} \leq A \leq +\frac{1}{2}$. Different choices of A give rise to different rounding errors $\tilde{S}_\Delta - s$ and the variance of the distribution of $\tilde{S}_\Delta - s$ can be used to measure the variability of the <u>rounding error due to the random selection of the origin of the real numbers</u> $a_i$ <u>with respect to that of the computer.</u>

The cumulative error from (1) and (2) is

$$(3) \quad \tilde{S}_\Delta - s = \sum_{i=1}^{n} [(a_i - \Delta)^* - (a_i - A)]$$

Let $f_i$ be the positive fractional part of $a_i$ and let $\alpha_i$ be the largest integer not exceeding $a_i$, i.e.,

$$(4) \quad a_1 = \alpha_i + f_i$$

Denoting by $r_i$ the error of the $i^{th}$ term, we have

$$(5)\quad ri = [(ai-A)^* - (a_i-A)] = \begin{cases} 1-(f_i-\Delta) & \text{if } -\frac{1}{2} \le A \le -\frac{1}{2} + f_i \\ -(f_i-\Delta) & \text{if } -\frac{1}{2} + f_i \le A \le +\frac{1}{2} \end{cases}$$

To prove the above, we note that $f_i-\Delta = (a_i-\Delta) + \alpha_i$. If $-\frac{1}{2} \le f_i - \Delta \le +\frac{1}{2}$ then $(a_i-\Delta)$ is rounded to $a_1$. Hence $a_i-\Delta$ is rounded down if $-\frac{1}{2} + f_i \le A$ otherwise rounded up.

Denoting expected value by E, we have by direct evaluation

$$(6)\qquad E(r_i) = \int_{-\frac{1}{2}}^{+\frac{1}{2}} r_i d\Delta = 0$$

Assume $f_i \le f_j$, then

$$E(r_ir_j) = \int_{-\frac{1}{2}}^{-\frac{1}{2}+f_i} r_ir_j dA + \int_{-\frac{1}{2}+f_i}^{-\frac{1}{2}-f_j} r_ir_j d\Delta + \int_{-\frac{1}{2}+f_j}^{+\frac{1}{2}} r_ir_j d\Delta$$

$$= \int_{-\frac{1}{2}}^{+\frac{1}{2}} (f_if_j - \Delta(f_i + f_j) + \Delta^2)\, d\Delta$$

$$+ \int_{-\frac{1}{2}}^{-\frac{1}{2}+f_i} [(1-f_i-f_j) + 2\Delta]\, d\Delta$$

$$+ \int_{-\frac{1}{2}+f_i}^{-\frac{1}{2}+f_j} [-f_1 + A)\, d\Delta$$

Performing indicated integration yields:

$$(7)\qquad E(r_ir_j) = \frac{1}{2}[\,|f_j-f_i|^2 - |f_j-f_i| + \frac{1}{6}]$$

which is one-half the $2^{nd}$ order Bernoulli Polynomial in $|f_j - f_i|$. For $f_j \leq f_i$ we also get (7). Note that the individual errors $r_i$ and $r_j$ are not independent of one another.

It now follows that

$$(9) \qquad E(\tilde{S}) = S$$

$$(10) \qquad E(\tilde{S}-S)^2 = E\left(\sum_{i=1}^{n}\sum_{j=1}^{n} r_i r_j\right) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[|f_i-f_j|^2 - |f_i-f_j| + \frac{1}{6}\right]$$

The usual value of variance, $E(\tilde{S}-S)^2 = n/12$, will result if we further assume $f_i$ are independently drawn from uniform distributions on $[0 \leq f_i \leq 1]$.

Theorem: If the fractional parts of all $a_i$ are equal to each other, then each term of (10) is maximum for $0 \leq f_i \leq 1$ and

$$(11) \qquad \text{Max } E(\tilde{S}-S)^2 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{1}{6}\right) = \frac{n^2}{12} .$$

From (10) we have an interesting inequality, namely for all $f_i$

$$(12) \qquad V(f) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{|f_i-f_j|^2 - |f_i-f_j| + \frac{1}{6}\right\} \geq 0$$

This function is not convex even for $n=2$, since $f^{(1)} = (\frac{1}{2}, 0)$ and $f^{(2)} = (-\frac{1}{2}, 0)$ yields $V(f^0) = V(f^1) = \frac{1}{12} + \frac{1}{12} - \frac{1}{12} = \frac{1}{12}$ but

$V\left(\frac{f^1 + f^2}{2}\right) = V(0) = \frac{3}{12}$. There appears to be no obvious direct way to establish that $V(f) \geq 0$ for all $0 \leq f_i \leq 1$. Our development shows $V(f)$ to be a variance and this, of course, constitutes an indirect proof. We can replace (12) by a convex realization: Assume $f_i \geq f_{i+1}$ for all $i$, then the problem of finding Min $V(f)$ can be rewritten:

$$\text{(13) Find Min } [V(f)] = \sum_{i<j}\sum (f_i - f_j)^2 + \frac{n^2}{12} - [(n-1)f_1 + (n-3)f_2 + (n-5)f_3$$

$$+ \ldots + (n-2k+1)\, f_k + \ldots - (n-1)f_n]$$

subject to

$$\text{(14)} \quad f_1 \geq f_2 \ \ldots \ \geq f_n$$

$$\text{(15)} \qquad 0 \leq f_i \leq 1$$

Formally (13), (14), (15), is a positive definite quadratic program. Fortunately, as we shall see this can be solved by classical calculus by ignoring inequalities (14) and (15).

Theorem: Equally spaced $f_i = (n - i)/n$, $(i = 1, \ldots, n)$ yields Min $V(f) = \frac{1}{12}$ independent of $n$, i.e., the variance of the sum in this case is minimum and is the same as the variance of the individual terms forming the sum.

Proof: Setting partials = 0 in (13) yields:

$$
(16)\quad \begin{cases}
2(n-1)f_1 - 2f_2 \; \cdots \; - 2f_n = (n-1) \\
-2f_1 + 2(n-1)f_2 \qquad - 2f_n = (n-3) \\
-2f_1 - 2f_2 \; \cdots \; 2(n-1)f_{n-1} - 2f_n = -(n-3) \\
-2f_1 - 2f_2 \qquad 2(n-1)f_n = -(n-1)
\end{cases}
$$

Adding shows the equations to be dependent. Hence we may drop the last equation as redundant. Moreover, we can always translate the $f_i$ so that the smallest $f_i$, namely $f_n = 0$

Re-adding yields:

$$2f_1 + 2f_2 + \ldots 2f_{n-1} + 0 = (n-1) \qquad , f_n = 0.$$

Adding this last equation to each of the others gives

$$2nf_i = (n - 2i + 1) + (n - 1) = 2(n - i)$$

$$(17) \qquad f_i = (n - i)/n$$

Evidently the conditions $0 \leq f_i \leq 1$ and $f_i \geq f_{i+1}$ are (by good luck) also satisfied so that (17) yields the minimum, namely

$$(18) \qquad \text{Min } V(f) = \frac{n^2}{12} - \frac{1}{2} \sum_{i=1}^{n} (n-2i+1) f_i = \frac{1}{12} \,,$$