ON THE DISTRIBUTION OF ACCUMULATED ROUNDOFF ERROR IN FLOATING POINT ARITHMETIC*

Jesse Barlow
Department of Electrical Engineering and Computer Science
Northwestern University
Evanston, Illinois 60201

Abstract

This paper discusses longstanding problems in the probabilistic error analysis of numerical algorithms when they are performed in floating point arithmetic.

Local roundoff error in floating point addition is characterized and its mean and variance are approximated. We apply these results to finding distributions for the roundoff error accumulated in sums and long inner products.

We state theorems which resolve questions left open in Bustoz et al. [5] and Hamming [11]. These theorems are proven in [3].

1. Introduction

There are two significant purposes for the discussion of accumulated roundoff error in computer arithmetics. The first is to analyze the error performance of the arithmetic systems themselves, and the second is to analyze the error performance of numerical algorithms. In the case of floating point arithmetic, we describe methods to do both.

For example, if we let s^* denote the result of one floating point operation on two floating point numbers a_1 and a_2 and let s denote the exact result of the operation, then

$$s \equiv s^*(1+\rho) \equiv (a_1 \text{ op } a_2)(1+\rho) \qquad (1)$$

where op is the floating point approximation of one of the questions $+, -, *, \div$.

The error analysis problem is to characterize ρ which is

$$\rho = \frac{s - s^*}{s^*}. \qquad (2)$$

If $s = xB^E$ and $s^* = x^*B^E$ (the case where s and s^* have different exponents is discussed in [3]) where B is the base of the floating point number system and $x, x^* \in [1/B, 1)$ are the fractional parts of s and s^* respectively, then

$$\rho = \frac{x - x^*}{x^*}. \qquad (3)$$

A distribution for ρ can be found by assuming a distribution for x^* and a distribution for $\epsilon = x - x^*$ and performing a transformation.

We assume that the distribution for x^* is closely approximated by the reciprocal distribution which has the density function

$$r(x^*) = 1/(x^* \ell n B) \quad \text{if} \quad x^* \quad [1/B, 1). \qquad (4)$$

The use of this density for real fractions x is justified empirically in [4] and by its theoretical properties in [7], [15], [16]. Thus the reciprocal distribution is only an approximation of the distribution of floating point fractions x^*. The above justifications are summarized in [2], [13], and [14].

When the operation is multiplication or division ϵ is assumed to approximately follow a uniform distribution whose density function is

$$u(\epsilon) = 1/(d-c) \quad \text{if} \qquad [c,d) \quad d > c \qquad (5)$$

where $c = 0$ and $d = B^{-t}$ when chopping is used and $c = -1/2\ B^{-t}$ and $d = 1/2B^{-t}$ when symmetric rounding is used on a machine with t-digit fractions. We justify the use of this distribution and generalize results from Goodman and Feldstein [6], [8], [9] and Bustoz et al. [5].

If the operation is addition or subtraction the uniform distribution is not a good approximation of the distribution of ϵ. We do have enough information from our distributions for trailing digits to find approximate means and variances for ϵ under addition and subtraction.

The effect of repeated operations on the distribution of real fractions, and hence approximately the effect on floating point fractions is discussed by Adhikari and Sarkar [1] and

* The work in this paper was supported by the National Science Foundation under contract No. MCS-7920150.

Hamming [11]. Hamming left open questions about the effects of repeated multiplications and divisions on floating point and real fractions. We resolve those questions in this paper.

We apply our results to the problem of finding confidence intervals for the error from sums and long inner products.

2. Multiplication and Division Reinforce the Reciprocal Distribution

Hamming (1970) showed that if a and b are random real numbers and $c = a*b$ with $a = xB^E$, $b = yB^F$, and $c = zB^G$, and $x,y,z \in [1/B,1)$ with densities $f(x)$, $g(y)$ and $h(z)$ respectively, then $h = I_M(f,g)$, where

$$I_M(f,g)(z) = \frac{1}{B}\int_{1/B}^{z} \frac{f(x)}{x}\, g(z/Bx)dx \qquad (6)$$

$$+ \int_{z}^{1} \frac{f(x)}{x}\, g(z/x)dx$$

If $c = a/b$ then $h = I_D(f,g)$, where

$$I_D(f,g)(z) = \frac{1}{z^2}\int_{1/B}^{z} x\, f(x)g(x/z)dx \qquad (7)$$

$$+ \frac{1}{Bz^2}\int_{z}^{1} x\, f(x)g(x/Bz)dx$$

Hamming also showed:

(i) $I_M(f,r)(z) = I_D(f,r)(z) = r(z)$

regardless of what f is when r is the reciprocal density defined by (4).

(ii) If we define the distance functional

$$D\{f\} = \sup_{x \in [1/B,1)} \frac{|f(x) - r(x)|}{r(x)} \qquad (8)$$

then

$$D\{I_M(f,g)\} \le D\{g\} \qquad (9a)$$

$$D\{I_D(f,g)\} \le D\{g\}. \qquad (9b)$$

As new results we show that under minimal restrictions on f and g:

1) The inequalities (9) are strict.
2) $r(x)$ is the only density for floating point fractions that is preserved under multiplication or division.
3) Repeated multiplications and/or divisions force densities satisfying these restrictions to the reciprocal density.

The lemmas necessary for proof of these results are stated. The proof is in [3].

For simplicity we consider the domain for our probability density functions f and g to be $[1/B,1]$. Since we also assume these density functions are bounded, the assumption $x \in [1/B,1]$ instead of $x \in [1/B,1)$ has no influence on the distributions associated with these densities.

Lemma 1. If f and g are bounded on $[1/B,1]$ and g is continuous* on that interval, then $I_M(f,g)$ and $I_D(f,g)$ are bounded and continuous.*

Lemma 2. If f and g satisfy the hypothesis of Lemma 1 and if $h_M = I_M(f,g)$ and $h_D = I_D(f,g)$ then for some z_M, $z_D \in [1/B,1]$

$$\frac{|h_M(z_M) - r(z_M)|}{r(z_M)} = D\{h_M\}$$

and

$$\frac{|h_D(z_D) - r(z_D)|}{r(z_D)} = D\{h_D\}$$

respectively.

Theorem 1. Let f and g satisfy the hypothesis of Lemma 1; let $f(x) > 0$ a.e. on $[1/B,1]$; let I_M and I_D be defined by (6) and (7) respectively and let r be given by (4). If $g \neq r$ then

(a) $D\{I_M(f,g)\} < D\{g\}$

(b) $D\{I_D(f,g)\} < D\{g\}$

We conjecture that Theorem 1 is the strongest theorem with the weakest conditions we can derive. If $f(x) = 0$ over a measurable part of $[1/B,1]$, then the conclusions of Theorem 1 do not hold in general, as is shown in [3].

Corollary 1. If f and g satisfy the hypothesis of Theorem 1 and I_M and I_D are as described by (6) and (7) then r is the only continuous density on $[1/B,1]$ that is a fixed point of I_M and/or I_D.

Corollary 2. If f satisfies the hypothesis of Theorem 1; I_M and I_D are described by equations (6) and (7) respectively; and $\{g_n\}_{n=1}$ and $\{h_n\}_{n=1}$ are sequences of continuous density functions on $[1/B,1]$ described by $g_1 = h_1 = f$ with $g_{n+1} = I_M(f,g_n)$ and $h_{n+1} = I_D(f,h_n)$ for $n = 1,2,3,\ldots$ then

---

* Left continuous at 1 and right continuous at 1/B.

$$\lim_{n\to\infty} g_n(x) = \lim_{n\to\infty} h_n(x) = r(x)$$

The proof follows immediately from Theorem 1.

3. The Distributiton of the Intermediate and Trailing Digits of Floating Point Fractions

Our assumptions for the distribution of **discarded digits** in the four standard operations **is determined** by the following theorem.

Theorem 2. Let xB^E be a real number where $x \in [1/B,1)$ follows a probability distribution F with continuous density $f(x) = F'(x)$, satisfying the Lipschitz condition

$$|f(x) - f(y)| \leq K|x - y| \;\forall\; x,y \in [1/B,1). \qquad (10)$$

Let x^*B^E be xB^E truncated to t digits. Define $A = \lfloor (x - x^*)B^{t+k} \rfloor B^{-k} = [.x_{t+1}\ldots x_{t+k}]$ and let $Q_k^t(A)$ be the probability distribution of A. Then

$$\lim_{k\to\infty} Q_k^t(A) = U(A) + O(B^{-t}) \qquad (11a)$$

$$\lim_{t\to\infty} Q_k^t(A) = \lfloor AB^k \rfloor B^{-k} = U(A) + O(B^{-k}) \qquad (11b)$$

where
$$U(A) = \begin{cases} 0 & \text{if } A < 0 \\ A & \text{if } A \in [0,1) \\ 1 & \text{if } A > 1. \end{cases}$$

Here k is the number of discarded digits and t is the number of digits in the computer word. In multiplication $k = t$ and in division $k = \infty$ so we approximate $Q_k^t(A)$ by $U(A)$. In addition and subtaction k varies greatly and tends to be small more often then large [17]. Therefore we approximate $Q_k^t(A)$ by $\lfloor AB^k \rfloor B^{-k}$ when dealing with error from these two operations.

4. Floating Point Arithmetic

Using the assumptions of this paper, [12] and [18] derived density functions $h_c(\ )$ and $h_R(\ )$ for ρ of (2) and (3) when chopping and symmetric rounding respectively are used.

For chopping

$$h_c(\rho) = \begin{cases} (B-1)B^{t-1}/\ln B & \text{if } \rho \in [0,B^{-t}) \\ (\frac{1}{\rho} - B^{t-1})/\ln B & \text{if } \rho \in [B^{-t},B^{1-t}) \end{cases} \qquad (12)$$

with first and second non-central moments

$$E_c(\rho) = B^{-t}(B-1)/(2\ln B) \qquad (13a)$$

$$E_c(\rho^2) = B^{-2t}(B^2-1)/6\ln B). \qquad (13b)$$

For symmetric rounding

$$h_R(\rho) = \begin{cases} (B-1)B^{t-1}/\ln B & \text{if } |\rho| \in [0,1/2\,B^{-t}) \\ (\frac{1}{2|\rho|} - B^{t-1})/\ln B & \text{if } |\rho| \in [1/2\,B^{-t}, 1/2\,B^{-t}) \end{cases} \qquad (14)$$

with first and second moments

$$E_R(\rho) = 0 \qquad (15a)$$

$$E_R(\rho^2) = Var(\rho) = B^{-2t}(B^2-1)/24\ln B). \qquad (15b)$$

If the operation is addition or subtraction the assumption that $\epsilon = x - x^*$ follows a continuous uniform distribution is inaccurate. The reason for this is that the number of discarded digits varies greatly in addition and subtraction.

From Section 3 we may assume that the discarded digits approximately follow a discrete uniform distribution. Suppose we are adding $a_1 = x_1B^{E_1}$ and $a_2 = x_1B^{E_2}$ where $E_1 \geq E_2$. If $E_1 - E_2$ is large then the continuous uniform distribution is a good approximation to the distribution of $\epsilon = x - x^*$, but if $E_1 - E_2$ is small which is more often the case [17] then the continuous uniform distribution is an inappropriate model for the behavior of ϵ.

Unfortunately, there is no good assumption for the distribution of the exponents. For that reason we make no such assumption. We let k be the number of discarded digits, and assume that it is known and non-zero. There are two cases regarding B, which must be treated separately, namely when B is even and when B is odd.

When B is even it is assumed that the method of symmetric rounding which rounds to the nearest even fraction in case of a tie is employed. Then ϵ follows the distribution with density

$$\rho(\epsilon) = \begin{cases} B^{-k} & \text{if } \epsilon = \frac{i}{B^{k+t}}, i=0,\pm,\ldots,\pm B-1 \\ 1/2B^{-k} & \text{if } \epsilon = \frac{\pm 1}{2B^t}. \end{cases} \qquad (16)$$

Therefore, for a fixed x^*, the relative error $\rho = \epsilon/x^*$ follows the conditional density $h(\rho|x^*,k)$ given by

$$h(\rho|x^*,k) = \begin{cases} B^{-k} \text{if } \rho = \frac{i}{x^*B^{k+t}}, \; i=0,\pm 1,\ldots,\pm 1/2B^k - 1 \\ 1/2B^{-k} \text{ if } \rho = \frac{\pm 1}{2x^*B^t}. \end{cases} \qquad (17)$$

By symmetry $E(\rho|x^*,k) = 0$ so $E(\rho) = 0$.

$$\mathrm{Var}(\rho\,|\,x^*,k) = \sum_{(\rho)} \rho^2 \; h(\rho\,|\,x^*,k) \tag{18}$$

$$= \sum_{i=0}^{\frac{1}{2}B^k} 2\left(\frac{i}{x^*B^{k+t}}\right)^2 B^{-k} - \left(\frac{1}{2x^*B^t}\right)^2 B^k$$

$$= \frac{1}{(x^*)^2B^{2t}}\left[\frac{1}{12} + \frac{1}{6B^{2k}}\right]$$

$$\mathrm{Var}(\rho\,|k) = \int_{1/B}^{1} \mathrm{Var}(\rho\,|\,x^*,k)\,\frac{dx^*}{x^*\ell nB} \tag{19}$$

$$= \left[\frac{1}{12}+\frac{1}{4B^k}+\frac{1}{6B^{2k}}\right]B^{-2t}\int_{1/B}^{1}\frac{dx^*}{(x^*)^3\ell nB}$$

$$= \frac{B^{-2t}(B^2-1)}{\ell nB}\left[\frac{1}{24}+\frac{1}{6B^{2k}}\right].$$

If $k = 0$ then $\rho \equiv 0$. Thus $\mathrm{Var}(\rho|\,0) \equiv 0$. If $p_0 = \mathrm{Prob}(k = 0)$ then $\mathrm{Var}(\rho)$ can be bounded by

$$\inf_{k\neq 0}(1-p_0)\mathrm{Var}(\rho\,|k) \le \mathrm{Var}(\rho) \le \sup_{(k)}(1-p_0)\mathrm{Var}(\rho\,|k) \tag{20}$$

Since $\mathrm{Var}(\rho|k)$ is decreasing with increasing k,

$$\sup_{(k)} \mathrm{Var}(\rho\,|\,k) = \mathrm{Var}(\rho\,|\,1) = \frac{B^{-2t}(B^2-1)}{\ell nB}\left[\frac{1}{24}+\frac{1}{12B^2}\right]. \tag{21}$$

and

$$\inf_{k\neq 0} \mathrm{Var}(\rho\,|k) = \lim_{k\to\infty}\mathrm{Var}(\rho\,|k) = \frac{B^{-2t}(B^2-1)}{24\ell nB}. \tag{22}$$

Therefore

$$\frac{(1-p_0)B^{-2t}(B^2-1)}{24\ell nB} \le \mathrm{Var}(\rho) \le \frac{(1-p_0)B^{-2t}(B^2-1)}{\ell nB}\left[\frac{1}{24}+\frac{1}{12B^2}\right]. \tag{23}$$

p_0 can vary greatly depending upon the source of additions and subtractions. Note the $k = 0$ occurs only when adding or subtracting numbers with equal exponents and there is no overflow. In base two, this occurs only when subtracting numbers with equal exponents. Sweeney [17], p. 41 from a sample of 250,000 additions and subtractions found p_0 for base two to be approximately .153. Thus for a base two 22 bit floating point computer

$$\mathrm{Var}(\rho) \in [8.683 \times 10^{-15},\ 1.302 \times 10^{-14}].$$

The mean and variance for ρ when B is odd is derived in [3].

5. Extended Operations in Floating Point Arithmetic

Let

$$s_n = \sum_{i=1}^{n} a_i \tag{24}$$

where the a_i are floating point numbers. Let s_n^* be the machine computation of s_n and let $\Delta s_n = s_n - s_n^*$. The computational recursion equation is

$$s_{k+1}^*(1+\rho_{k+1}) \equiv s_k^*+a_{k+1} \quad k=2,\ldots,n-1 \tag{25}$$

where s_2^* is defined in (1) with op = + and $s = s_2$.

It follows that

$$\Delta s_{k+1} \equiv s_{k+1} - s_{k+1}^* \tag{26}$$

$$\equiv s_k - s_k^* + s_{k+1}^*\rho_{k+1} \equiv \Delta s_k + s_{k+1}^*\rho_{k+1}.$$

Solving the recursion relation we get

$$\Delta s_n \equiv \sum_{k=2}^{n} s_k^*\rho_k. \tag{27}$$

Because the ρ_k are the result of independent machine operations they are independent and identically distributed which implies

$$E(\Delta s_n) = \left(\sum_{k=2}^{n} s_k^*\right)E(\rho) \tag{28a}$$

$$\mathrm{Var}(\Delta s_n) = \left(\sum_{k=2}^{n} (s_k^*)\right)\mathrm{Var}(\rho). \tag{28b}$$

By the Central Limit Theorem from probability theory

$$\frac{\Delta s_n - E(\Delta s_n)}{\sigma(\Delta s_n)}$$

where $\sigma(\Delta s_n) = (\mathrm{Var}(\Delta s_n))^{1/2}$ is approximately normally distributed with mean zero and variance one for reasonably large n.

Thus an approximate $100\,\omega\%$ confidence interval for Δs_n is given by

$$[E(\Delta s_n) - q_{\omega_0}\sigma(\Delta s_n),\ E(\Delta s_n) + q_{\omega_0}\sigma(\Delta s_n)] \tag{29}$$

where

$$\Phi(q_{\omega_0}) = \omega_0 = \frac{1+\omega}{2},\ \omega \in (0,1)$$

and

$$\Phi(y) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y} e^{-t^2/2}dt.$$

If instead we let s_n denote the inner product

$$s_n = \sum_{k=1}^{n} a_i b_i$$

and let $c_i = a_i b_i$ then

$$s_1^* = c_1^* \qquad (30)$$

$$s_{k+1}^*(1+\rho_{k+1}) \equiv s_k^* + c_{k+1}^* \quad k=1,2,\ldots,n-1 \qquad (31)$$

where

$$c_k^*(1+\alpha_k) \equiv c \qquad k=1,2,\ldots,n \qquad (32)$$

and α_k is the relative error from the floating point multiplication of a_k and b_k.

The recursion relation for Δs_n is

$$\Delta s_{k+1} \; s_k - s_k^* + c_{k+1} - c_{k+1}^* + s_{k+1}^* c_{k+1} \qquad (33)$$

$$\equiv s_k + c_{k+1}^* \alpha_{k+1} + s_{k+1}^* \rho_{k+1}. \quad k=1,2,\ldots,n-1.$$

The solution of (33) is

$$\Delta s_n = \sum_{k=1}^{n} c_k^* \alpha_k + \sum_{k=1}^{n} s_k^* \rho_k.$$

Therefore

$$E(\Delta s_n) = (\sum_{k=1}^{n} c_k^*)E(\alpha) + (\sum_{k=1}^{n} s_k^*)E(\alpha) \qquad (34a)$$

$$\mathrm{Var}(\Delta s_n) = (\sum_{k=1}^{n} (c_k^*)^2)\mathrm{Var}(\rho) \qquad (34b)$$

$$+ (\sum_{k=1}^{n} (s_k^*)^2)\mathrm{Var}(\rho)$$

where α is the relative error from one multiplication and ρ is the relative error from one addition.

As with sums

$$\frac{\Delta s_n - E(\Delta s_n)}{\sigma(\Delta s_n)}$$

is approximately normally distributed with mean zero and variance one, and hence an approximate 100 ω% confidence interval is given by (29).

References

1. Adhikari, A. K. and B. F. Sarkar, "Distribution of Most Significant Digit in Certain Function whose Arguments are Random Variables," Indian J. of Statistics, Series B, 30, Part 1 & 2 (1968), 47-58.

2. Bareiss, E. H. and J. L. Barlow, "Probabilistic Error Analysis of Computer Arithmetics," Northwestern University DOE Report COO-2280-37, (December 1978).

3. Barlow, J. L., "Probabilistic Error Analysis of Floating Point and CRD Arithmetics," Ph.D. Thesis, Northwestern University, (June, 1981).

4. Benford, F., "The Law of Anomolous Numbers," Proc. Amer. Phil. Soc., 78 (1938), 551-572.

5. Bustoz, J., A. Feldstein, R. Goodman, and S. Linnainmaa, "Improved Trailing Digit Estimates Applied to Optimal Computer Arithmetic," JACM, Vol. 26, No. 4, (1979), 716-730.

6. Feldstein, A. and R. Goodman, "Convergence Estimates for the Distribution of Trailing Digits," JACM, 23 (1976), 287-297.

7. Flehinger, B. J., "On the Probability that a Random Integer has Initial Digit A," Amer. Math. Monthly, 73 (1966), 1056-1061.

8. Goodman, R. and A. Feldstein, "Effects of Guard Digits and Normalization Options on Floating Point Multiplication," Computing, Vol. 18 (1977), 93-106.

9. __________, "Roundoff Error in Products," Computing, Vol. 15 (1975), 263-273.

10. Grau, A. A. and E. H. Bareiss, "Statistical Aspects of Machine Rounding," Northwestern Univeristy ERDA Report COO-2280-34, (August 1977).

11. Hamming, R. W., "On the Distribution of Numbers," Bell System Technical Journal, Vol. 49, No. 8 (1970), 1609-1625.

12. Kaneko, T. and B. Liu, "On the Local Round-off Error in Floating Point Arithmetic," JACM, Vol. 20 (July 1973), 391-398.

13. Knuth, D. E., The Art of Computer Programming Vol. 2: Seminumerical Algorithms, Addison-Wesley, Reading, Mass., (1969).

14. Nathan, L. H., Probabilistic Distribution of the Most Significant Digit in Computer Represented Numbers and Its Behavior Under Iterated Fixed and Floating Point Operations, M.S. Thesis, Northwestern University, Evanston, Illinois, (August 1973).

15. Pinkham, R. S., "On the Distribution of First Significant Digits," Ann. Math. Stat., 32 (1961), 1223-1230.

16. Raimi, R. A., "On the Distribution of First Significant Digits," Amer. Math. Monthly, 74, No. 2 (1969), 342-348.

17. Sweeney, D., "An Analysis of Floating Point Addition," IBM Systems Journal, Vol. 4, No. 5, 31-42

18. Tsao, N., "On the Distribution of Significant Digits and Roundoff Errors," CACM, Vol. 17, (May 1974), 269-271.

Acknowledgment

The author would like to express his gratitude to his faculty adviser, Professor E. H. Bareiss, for his critique of an earlier version of this paper.