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 arithmetic. 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

$$s = s^*(1+e) = (a_1 op a_2)(1+e)$$

(1)

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

The error analysis problem is to characterize $s$ which is

$$e = \frac{s - s^*}{s^*}.$$  

(2)

If $s = s^* F$ and $s^* = s^* F$ (the case where $s$ and $s^*$ have different exponents is discussed in [3]) where $E$ 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

$$e = \frac{x - x^*}{x^*}.$$  

(3)

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

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

$$r(x^*) = 1/(x^* \ln B) \text{ if } x^* \in [1/B,1).$$  

(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 $\epsilon$ is assumed to approximately follow a uniform distribution whose density function is

$$u(\epsilon) = \frac{1}{d-c} \text{ if } (c,d), c > 0$$  

(5)

where $c = 0$ and $d = B^{-t}$ when chopping is used and $c = -1/2 B^{-t}$ and $d = 1/2 B^{-t}$ 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 $\epsilon$. We do have enough information from our distributions for trailing digits to find approximate means and variances for $\epsilon$ 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 $c = ab$ with $a = xB^e$, $b = yB^d$, and $c = zB^u$, and $x, y, z \in [1/8, 1]$ with densities $f(x), g(y)$ and $h(z)$ respectively, then

$$ I_M(f,g)(z) = \frac{1}{B} \int_{1/8}^{z} f(x) g(z/Bx) dx \quad (6) $$

$$ + \frac{1}{z} \int_{1/8}^{1} f(x) g(z/x) dx $$

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

$$ I_D(f,g)(z) = \frac{1}{z^2} \int_{1/8}^{z} x f(x) g(x/z) dx \quad (7) $$

$$ + \frac{1}{z} \int_{1/8}^{1} x f(x) g(x/Bz) dx $$

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

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

then

$$ D(I_M(f,g)) \leq D\{g\} \quad (9a) $$

$$ D(I_D(f,g)) \leq D\{g\}. \quad (9b) $$

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 $[1/8,1]$. Since we also assume these density functions are bounded, the assumption $x \in [1/8,1]$ instead of $x \in [1/8,1]$ has no influence on the distributions associated with these densities.

Lemma 1. If $f$ and $g$ are bounded on $[1/8,1]$ 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 $z_M, z_D \in [1/8,1]$

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

and

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

respectively.

Theorem 1. Let $f$ and $g$ satisfy the hypothesis of Lemma 1; let $f(x) > 0$ a.e. on $[1/8,1]$; 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 $[1/8,1]$, 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 $[1/8,1]$ 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 $\{g_n\}_{n=1}^\infty$ and $\{h_n\}_{n=1}^\infty$ are sequences of continuous density functions on $[1/8,1]$ described by $g_1 = h_1 = f$ with $g_n = I_M(f, g_{n-1})$ and $h_n = I_D(f, h_{n-1})$ for $n = 1, 2, 3, \ldots$ then

* Left continuous at 1 and right continuous at 1/8.
\[ \lim_{n \to \infty} g_n(x) = \lim_{n \to \infty} h_n(x) = r(x) \]

The proof follows immediately from Theorem 1.

3. The Distribution 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 \( x^E \) 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

\[ |f(x) - f(y)| \leq K|x - y| \quad \forall x, y \in (1/B, 1). \]  

(10)

Let \( x^B \) be a number truncated to \( t \) digits. Define

\[ A = \left\{ (x - x^*)^B + k \right\} B^{-k} \in [x, x^* + 1] \]  

and let \( q_k^t(A) \) be the probability distribution of \( A \). Then

\[ \lim_{k \to \infty} q_k^t(A) = U(A) + 0(B^{-t}) \]  

(11a)

\[ \lim_{t \to \infty} q_k^t(A) = [AB^k] B^{-k} = U(A) + 0(B^{-t}) \]  

(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 subtraction \( k \) varies greatly and tends to be small more often than large [17]. Therefore we approximate \( q_k^t(A) \) by \( [AB^k] B^{-k} \) when dealing with error from these two operations.

4. Floating Point Arithmetic

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

For chopping

\[ h_c(\rho) = \begin{cases} 
(B-1)B^{-t} \lambda/nB & \text{if } \rho \in [0, B^{-t}] \\
0 & \text{if } \rho \in (B^{-t}, 1]
\end{cases} \]  

(12)

with first and second non-central moments

\[ E_c(\rho) = B^{-t}(B-1)/(2t/nB) \]  

(13a)

\[ E_c(\rho^2) = B^{-2t}(B-1)/6(nB) \]  

(13b)

For symmetric rounding

\[ h_R(\rho) = \begin{cases} 
(B-1)B^{-t} \lambda/nB & \text{if } \rho \in (0, 1/2, B^{-t}) \\
(1/2)B^{-t} \lambda/nB & \text{if } \rho \in (1/2, B^{-t}, 1/2, B^{-t}) \end{cases} \]  

(14)

with first and second moments

\[ E_R(\rho) = 0 \]  

(15a)

\[ E_R(\rho^2) = \text{Var}(\rho) = B^{-2t}(B^{-2t}-1)/24(nB). \]  

(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 \( E_1 \) and \( E_2 \) and \( E_1 \geq E_2 \). 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 \( \epsilon \).

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 \( \epsilon \) follows the distribution with density

\[ \rho(\epsilon) = \begin{cases} 
B^{-k} & \text{if } \epsilon = \frac{d}{B^{k+t}} \text{ and } 0, 1, \ldots, B-1 \\
1/2B^{-k} & \text{if } \epsilon = \frac{d}{B^{k+t}} \text{ and } 0, 1, \ldots, B-1/2, B^{-1}
\end{cases} \]  

(16)

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

\[ h(\rho|x^*, k) = \begin{cases} 
B^{-k} & \text{if } \rho = \frac{d}{x^*B^{k+t}} \text{ and } 0, 1, \ldots, 1/2B^{-1} \\
1/2B^{-k} & \text{if } \rho = \frac{d}{x^*B^{k+t}} \text{ and } 0, 1, \ldots, 1/2B^{-1}
\end{cases} \]  

(17)

By symmetry \( E(\rho|x^*, k) = 0 \) so \( E(\rho) = 0 \).
5. Extended Operations in Floating Point Arithmetic

Let

\[ s_n = \sum_{i=1}^{n} a_i \]  

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

\[ s_{n+1}^* (1 + a_{k+1}) = s_n^* + a_{k+1} \quad k=2, \ldots, n-1 \]  

where \( a_n^* \) is defined in (1) with \( op = + \) and \( s = a_2^* \).

It follows that

\[ \Delta s_{k+1} = \Delta s_{k+1}^* - s_{k+1}^* \]  

Solving the recursion relation we get

\[ \Delta s_n = \sum_{k=2}^{n} \Delta s_k \]  

Because the \( p_k \) are the result of independent machine operations they are independent and identically distributed which implies

\[ E(\Delta s_n) = \left( \sum_{k=2}^{n} \Delta s_k^* \right) E(p) \]  

\[ \text{Var}(\Delta s_n) = \left( \sum_{k=2}^{n} \text{Var}(a_k^*) \right) \text{Var}(p). \]  

By the Central Limit Theorem from probability theory

\[ \frac{\Delta s_n - E(\Delta s_n)}{\sigma(\Delta s_n)} \]

\[ \left\{ -15, 8.683 \times 10^{-15}, 1.302 \times 10^{-14} \right\} \]

The mean and variance for \( \Delta s_n \) when \( B \) is odd is derived in [3].
\[ s_n = \sum_{k=1}^{n} a_k b_i \]

and let \( c_i = a_i b_i \) then

\[ s_k^* = c_1 \]

\[ s_{k+1}^* (1 + \alpha_{k+1}) = s_k^* c_k + k = 1, 2, \ldots, n-1 \]

where

\[ \alpha_k (1 + \alpha_{k+1}) = c \]

\[ k = 1, 2, \ldots, n \]

and \( \alpha_k \) is the relative error from the floating point multiplication of \( a_k \) and \( b_k \).

The recursion relation for \( s_n \) is

\[ \Delta s_{k+1} = s_k^* a_k + s_{k+1}^* c_k + s_{k+1}^* k \]

\[ k = 1, 2, \ldots, n \]

The solution of (33) is

\[ \Delta s_n = \sum_{k=1}^{n} \alpha_k s_k^* + \sum_{k=1}^{n} \alpha_k s_k^* k \]

Therefore

\[ \text{E}(s_n) = (\sum_{k=1}^{n} \alpha_k) \text{E}(s_k^*) + (\sum_{k=1}^{n} \alpha_k) \text{E}(\alpha) \]

\[ \text{Var}(\Delta s_n) = \sum_{k=1}^{n} (\alpha_k)^2 \text{Var}(s_k^* ) + \sum_{k=1}^{n} (\alpha_k)^2 \text{Var}(\alpha) \]

where \( \alpha \) is the relative error from one multiplication and \( \rho \) is the relative error from one addition.

As with sums

\[ \Delta s_n = \text{E}(\Delta s_n) \]

\[ \sigma(\Delta s_n) \]

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

**References**


**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.