MASTER

NOTICE
This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Department of Energy, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

C00-2280-37

PROBABILISTIC ERROR ANALYSIS
OF COMPUTER ARITHMETICS

Erwin H. Bareiss and Jesse L. Barlow
Northwestern University

December 1978

PREPARED FOR THE U.S. DEPARTMENT OF ENERGY
UNDER CONTRACT NO. EY-76-S-02-2280

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED *EDB*

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

PROBABILISTIC ERROR ANALYSIS OF COMPUTER ARITHMETICS

Erwin H. Bareiss and Jesse L. Barlow
Northwestern University

ABSTRACT

The problem of continuous and discrete error distribution for real computer arithmetics is discussed. The existing literature is surveyed. Several new and important theorems are proven. Results are illustrated with figures and tables.

CONTENTS

Introduction	1
I. Deterministic Error Bounds	2
1. Fixed Point Arithmetic	3
2. Floating Point Computation	4
3. Common Floating Point Operations	5
4. Floating Point Accumulation of Sums and Inner Products	8
5. Parallel Computation	10
References	14
II. The Distribution of the Most and Least Significant Digits	15
1. The Benford Hypothesis	15
2. Multiplication and Division Reinforce the Reciprocal Distribution	19
3. The Distribution of the Intermediate and Least Significant Digits	27
References	36
III. Floating Point Arithmetic	37
1. The Ideals: Symmetric Rounding and Chopping	38
2. Guard Digits and Choice of Base	45
3. ROM Rounding	61
4. Non-Symmetric Rounding Intervals	63
References	65
IV. Fixed Point Arithmetic	66
1. The Distribution of a Constant Times a Discrete Random Variable	68
2. Fixed Point Multiplication of Two Random Numbers	74
3. Inner Product with Fixed Point Operations	82
References	87

V.	The CRD Computer	88
1.	Definition of the Number System	88
2.	Geometric Error	89
3.	CRD Arithmetic	91
	References	95
	Conclusions	96
	Appendix	97

List of Tables

1.1	Fixed and Floating Point Operations	13
2.1	Benford's Sample of Numbers	16
2.2	The Distance from the Reciprocal Distribution of an Extended Product as Function of the Number of Factors Selected from a uniform distribution	26
2.3	Values of $P_t(a, B)$ for $B = 2, 4, 8, 10$ and 16 Reproduced from Feldstein and Goodman (1976)	31
3.1	Reproduced from Kuki and Cody (1973)	
	(a) Binary Significance Figures for Summation Tests 500 Sums of 1024 Numbers Each	48
	(b) Binary Significance Figures for 500 Products of 20 Factors	48
	(c) Binary Significance Figures for Inner Produce Tests 500 Inner Products with 50 Operands	49
	(d) Binary Significance Figure of Summation Tests Comparison with R^* -rounding	49
3.2	Reproduced from Cody (1973). Comparison of Analytical and Experiment Results	51
3.3	Values of K and M if $u = B^{-t}$ is the Machine Unit Reproduced from Bustoz, Feldstein, and Goodman (1978)	57
3.4	Means and Variances of r_1 and r_2 for Various Rounding Methods Reproduced from Bustoz, Feldstein, and Goodman (1978)	60
3.5	Bias of Rounding Methods for Base B , a Power of Two. Reproduced from Kuck, Parker, and Sameh (1977)	62
3.6	Multiplication Roundoff Error Performance Reproduced from Kuck, Parker, and Sameh (1977). The Base B is a Power of Two.	62

List of Tables (continued)

3.7	Addition Roundoff Error Performance Reproduced from Kuck, Parker, and Sameh (1977). The Base B is a Power of Two.	63
4.1	Roundoff Error for Multiplication. Base B Prime First Factor must be Positive. Means and Variances. Reproduced from Goodman (1976)	67
4.2	Roundoff Error in Multiplication. Means and Variance. Reproduced from Goodman (1976). Both Factors May be Zero. Base B prime.	67
5.1	Error η_s in exponents from 2000 samples on 20 CRD additions with $\delta = 2^{2^{-t}}$	94

List of Figures

2.1	The Region of Integration for Multiplication, Equation (7). Adapted from Hamming (1970)	21
2.2	The Region of Integration for Division, Equation (9). Adapted from Hamming (1970)	22
3.1	Domain of Integration for Cumulative Floating Point Error Distribution Under Multiplication with Chopping (Equation (3))	40
3.2	Error Probability Density $h_c(\epsilon)$, Equation (6), for Floating Point Multiplication with Chopping on a Binary Computer ($B = 2$)	40
3.3	Error Density Function $h_R(\epsilon)$, Equation (9), for Multiplication when Symmetric Rounding is used on a Binary Computer ($B = 2$). Adapted from [Tsao, 1974].	41
3.4	Histogram of Multiplication Error. Reproduced from Kaneko and Liu (1973)	44
3.5	Histograms of Addition Error for $p_0 = 0.62$ and $p_0 = 0.75$. Reproduced from Kaneko and Liu (1973)	45
3.6	Trailing Digits and $R_2(\cdot)$. Reproduced from Bustoz, Feldstein, and Goodman (1978)	56
4.1	Graphs of the Densities of $\epsilon^{(1)}$, $\epsilon^{(2)}$ and $\epsilon^{(3)}$	86

INTRODUCTION

We consider the problem of the probabilistic analysis of errors in real arithmetic on digital computing devices. The material surveyed has implication both in numerical analysis and computer hardware design.

This paper is divided into five sections. The first section summarizes important results from deterministic error analysis. The second section discusses the theory of the distribution of "naturally occurring" real numbers. The last three sections analyze floating point arithmetic, fixed point arithmetic, and arithmetic on a logarithmic computer.

In our discussion of the distribution of numbers we present simplifications of proofs concerning the distribution of intermediate and trailing digits.

Our discussion of floating point arithmetic assesses the accuracy of various rounding methods employed in digital computers.

The section on fixed point arithmetic includes a new theorem concerning the multiplication of a random fixed point number by a constant, and three new theorems solve the problem of the discrete error densities if both factors are random, and under different rules for rounding.

The final section discusses the problem of addition on a logarithmic computer. We present empirical results from a simulation of CRD addition showing that its accuracy is significantly better than the deterministic error bounds for the operation.

I. DETERMINISTIC ERROR BOUNDS

There are two common methods of representing real numbers on modern digital computers and calculators, the fixed point number system and the floating point number system.

The fixed point system represents real numbers as sequences of digits in some number base, usually 10 or some power of two, and a sign at the beginning. The number of digits determine the size of the computer word.

The floating point system usually represents a real number x as $\pm m \cdot B^E$ where $m \in [1/B, 1)$ is the mantissa, $E \in \{-k, -k+1, \dots, 0, 1, \dots, n\}$, for some positive integers k and n , is the exponent and B is the base of the system. On most machines B is 10 or some power of two.

Both the fixed point system and the floating point system can represent only a finite set of numbers. Since the real numbers are an uncountably infinite set, most real numbers cannot be represented exactly (such a statement holds not only for the fixed and floating point systems but for any conceivable number system). First, this introduces the problem of appropriate rounding and representation of error. Second, except for fixed point addition and subtraction, the four standard arithmetic operations are often not exact because of the restriction on word size. The first problem is not directly within the scope of this paper and is treated in Grau and Bareiss (1977). The second problem is attacked deterministically in this section, and probabilistically in Sections III and IV.

1. Fixed Point Arithmetic

Without loss of generality we may restrict our discussion of fixed point numbers to the interval $(-1,1)$. Scaling factors may be introduced if necessary. We assume the numbers have t digits.

An equation of the form

$$(1) \quad z = fi(x_1 \text{ op } x_2)$$

where op is one of the operations $+, -, \times, \div$, implies the x_1, x_2 , and z are standard fixed point numbers and that z is obtained by performing the appropriate fixed point operation on x_1 and x_2 . In the case of addition, subtraction and multiplication we assume the computed z lies in the permitted range.

There are no rounding errors in addition and subtraction so that we have

$$(2) \quad z \equiv fi(x_1 \pm x_2) \equiv x_1 \pm x_2$$

Throughout our discussion unless otherwise specified we assume that the rounding procedure is such that

$$(3) \quad z \equiv fi(\overset{x}{x_1} \overset{x}{\div} x_2) \equiv \overset{x}{x_1} \div x_2 + \epsilon \text{ where } |\epsilon| \leq \frac{1}{2} B^{-t}$$

and, if division is performed $x_2 \neq 0$.

Accumulation of Inner Products

Let $S = \sum_{i=1}^n a_i b_i$, the standard Euclidean inner product. If each

product is rounded separately then the corresponding computational equation is

$$(4) \quad \bar{S} \equiv fi \left(\sum_{i=1}^n a_i b_i \right) \equiv S + \epsilon \text{ where } |\epsilon| \leq \frac{1}{2} n B^{-t}.$$

The exact representation of S requires $2t$ digits. If we restrict $n < B^t$, and $\bar{S}_2$ is the number obtained by the use of a double precision accumulator and then rounding to the standard single precision result, we write

$$(5) \quad \bar{S}_2 \equiv fi_2 \left(\sum_{i=1}^n a_i b_i \right) \equiv S + \epsilon \text{ where } |\epsilon| \leq \frac{1}{2} B^{-t}.$$

A computer designed with the above facility frequently also has a facility for dividing a $2t$ digit number by a t digit number to obtain a t digit number. Thus

$$(6) \quad \bar{W}_2 \equiv fi_2 \left(\left[\sum_{i=1}^n a_i b_i \right] / C \right) \equiv \left(\left[\sum_{i=1}^n a_i b_i \right] / C \right) + \epsilon$$

where $|\epsilon| \leq \frac{1}{2} B^{-t}$ provided $n < B^t$

$$(7) \quad \left| \bar{W}_2 C - \sum_{i=1}^n a_i b_i \right| \equiv |C\epsilon| \leq \frac{1}{2} |C| B^{-t}.$$

2. Floating Point Computation

Addition and Subtraction

Let $x_1 = a_1 B^{b_1}$, $x_2 = a_2 B^{b_2}$, assume $b_1 \geq b_2$ and let t be the number of digits in a_1 and a_2 . Addition is defined as follows:

Case I: $b_1 - b_2 > t$. Here x_2 is too small to effect the first t significant digits of the sum.

Hence

$$(8) \quad fl(x_1 + x_2) \equiv x_1.$$

Case II: $b_1 - b_2 \leq t$. In this case a_2 is divided by $B^{b_1 - b_2}$ by shift-

ing it $b_1 - b_2$ places to the right and the sum $a_1 + a_2 B^{b_1 - b_2}$ is calcula-

ted exactly. The sum is then multiplied by an appropriate power of B so that the mantissa lies in the interval $[\frac{1}{B}, 1)$ and the result is rounded.

Since $a_1 + a_2 B^{b_1 - b_2} \leq 2$ the exponent of the sum is at most $b_1 + 1$. Thus the computational equation is

$$(9) \quad fl(x_1 + x_2) \equiv (x_1 + x_2)(1 + \epsilon) \text{ where } |\epsilon| \leq \frac{1}{2} B^{1-t}.$$

Since no assumption is made about the sign of the numbers being added, the argument holds for subtraction as well.

Multiplication and Division

Multiplication of two floating point numbers is performed by simply multiplying the mantissas and adding the exponents. The resulting $2t$ -digit product is rounded to the t -digit mantissa of the computed product. Division is similar. Thus

$$(10) \quad fl(x_1 \overset{x}{\div} x_2) \equiv (x_1 \overset{x}{\div} x_2) (1 + \epsilon) \text{ where } |\epsilon| \leq \frac{1}{2} B^{1-t} \text{ and if division is performed } x_2 \neq 0.$$

3. Common Floating Point Operations

$$(11) \quad \text{Let } p_n \equiv fl(\prod_{i=1}^n x_i)$$

Using the recursive relation

$$(12) \quad p_1 \equiv x_1, p_r \equiv fl(p_{r-1} x_r) \equiv p_{r-1} x_r (1 + \epsilon), |\epsilon| \leq \frac{1}{2} B^{1-t}$$

then

$$(13) \quad p_n \equiv \left(\prod_{i=1}^n x_i \right) \prod_{i=2}^n (1 + \epsilon_i) \text{ where } |\epsilon_i| \leq \frac{1}{2} B^{1-t} \text{ for } i=2, \dots, n$$

and

$$(14) \quad p_n \equiv \left(\prod_{i=1}^n x_i \right) (1 + E) \text{ where } (1 - \frac{1}{2} B^{1-t})^{n-1} \leq 1 + E \leq (1 + \frac{1}{2} B^{1-t})^{n-1}$$

In a similar manner

$$(15) \quad fl(x_1 x_2 \dots x_m / y_1 y_2 \dots y_n) \equiv (x_1 x_2 \dots x_m / y_1 y_2 \dots y_n) \cdot (1 + E)$$

$$\text{where } (1 - \frac{1}{2} B^{1-t})^{m+n-1} \leq 1 + E \leq (1 + \frac{1}{2} B^{1-t})^{m+n-1}$$

For extended sequences of additions it is simpler to use backward error bounds. If we let

$$(16) \quad S_n \equiv fl\left(\sum_{i=1}^n x_i\right)$$

and use the recursion relations

$$(17) \quad S_1 \equiv x_1$$

$$S_r \equiv fl(S_{r-1} + x_r) \equiv (S_{r-1} + x_r) (1 + \epsilon_r) \text{ where } |\epsilon_r| \leq \frac{1}{2} B^{1-t}$$

we obtain

$$(18) \quad S_n \equiv \sum_{i=1}^n x_i (1 + N_i) \quad (N_i = \prod_{r=i}^n (1 + \epsilon_r); N_1 = N_2)$$

$$\text{where } (1 - \frac{1}{2} B^{1-t})^{n-1} \leq 1 + N_1 \leq (1 + \frac{1}{2} B^{1-t})^{n-1}$$

$$(1 - \frac{1}{2} B^{1-t})^{n-r+1} \leq 1 + N_r \leq (1 + \frac{1}{2} B^{1-t})^{n-r+1} \quad r = 2, \dots, n$$

The upper bound on the forward error is dependent on the order of summation, it is smallest when $x_1 \leq x_2 \leq \dots \leq x_n$.

Floating Point Euclidean Inner Product

$$\text{Let } s_n \equiv fl\left(\sum_{i=1}^n a_i b_i\right), \quad t_r \equiv fl(a_r b_r).$$

We use the recursion relationship

$$s_1 \equiv t_1, \quad s_r \equiv fl(s_{r-1} + t_r)$$

hence

$$(19) \quad t_r \equiv a_r b_r (1 + \xi_r) \quad (|\xi_r| \leq \frac{1}{2} B^{1-t})$$

$$s_r \equiv (s_{r-1} + t_r)(1 + \eta_r) \quad (|\eta_r| \leq \frac{1}{2} B^{1-t}).$$

Then

$$(20) \quad s_n \equiv \sum_{i=1}^n a_i b_i (1 + \epsilon_i)$$

$$\text{where } 1 + \epsilon_1 = (1 + \xi_1)(1 + \eta_2) \dots (1 + \eta_n)$$

$$\text{and } 1 + \epsilon_i = (1 + \xi_i)(1 + \eta_i) \dots (1 + \eta_n) \quad (i = 2, \dots, n)$$

hence

$$(21) \quad (1 - \frac{1}{2} B^{1-t})^n \leq 1 + \epsilon_1 \leq (1 + \frac{1}{2} B^{1-t})^n$$

$$(1 - \frac{1}{2} B^{1-t})^{n-r+2} \leq 1 + \epsilon_r \leq (1 + \frac{1}{2} B^{1-t})^{n-r+2} \quad (r = 2, \dots, n)$$

For $B = 2$ Wilkinson (1963) simplifies these expressions

for $n2^{-t} < 0.1$. In most cases this is a mild restriction on n .

Let

$$(22) \quad (1 + 2^{-t})^n < 1 + (1.06) n2^{-t} \quad \text{and} \quad (1 - 2^{-t})^n > 1 - (1.06) n2^{-t}$$

and define

$$(23) \quad t_1 = t - \log_2(1.06) = t - .08406$$

which implies $2^{-t_1} = (1.06) 2^{-t}$.

Hence inequalities of the form

$$(1 - 2^{-t})^r \leq 1 + \epsilon \leq (1 + 2^{-t})^r$$

may be replaced by

$$(24) \quad |\epsilon| \leq r 2^{-t_1}.$$

4. Floating Point Accumulation of Sums and Inner Products

Let $fl_2\left(\sum_{i=1}^n a_i\right)$ denote the double precision sum of the a_i 's. Then

$$(25) \quad fl_2\left(\sum_{i=1}^n a_i\right) \equiv \left[\sum_{i=1}^n a_i(1 + \epsilon_i)\right](1 + \epsilon)$$

where $1 + \epsilon$ is the factor from rounding from double precision to single precision.

The factors have the bounds

$$(26) \quad |\epsilon| \leq \frac{1}{2}B^{1-t}$$

$$(1 - \frac{1}{2}(B^{-2t} + B^{-2t-1}))^{n-1} \leq 1 + \epsilon_1 \leq (1 + \frac{1}{2}(B^{-2t} + B^{-2t-1}))^{n-1}$$

$$(1 - \frac{1}{2}(B^{-2t} + B^{-2t-1}))^{n-r+1} \leq 1 + \epsilon_r \leq (1 + \frac{1}{2}(B^{-2t} + B^{-2t-1}))^{n-r+1}$$

$$(r = 2, \dots, n)$$

For $B = 2$ if we restrict (25) according to the inequality
 $n \cdot 3/2 \cdot 2^{-2t} < 0.1$, then for $r \leq n$,

$$(1 + 3/2 \cdot 2^{-2t})^r < 1 + 3/2 \cdot r \cdot (1.06) \cdot 2^{-2t}$$

$$(1 - 3/2 \cdot 2^{-2t})^r > 1 - 3/2 \cdot r \cdot (1.06) \cdot 2^{-2t}$$

Define t_2 by

$$2t_2 = 2t - \log_2 (1.06)$$

which implies

$$(1.06) \cdot 2^{-2t} = 2^{-2t_2}$$

so that (26) may be replaced by

$$(27) \quad \begin{aligned} |\epsilon_r| &\leq 2^{-t} \\ |\epsilon_r| &< 3/2 \cdot (n-1) \cdot 2^{2t_2}, \quad (r = 2, \dots, n). \end{aligned}$$

By a similar analysis we obtain

$$fl_2\left(\sum_{i=1}^n a_i b_i\right) \equiv \left[\sum_{i=1}^n a_i b_i (1 + \epsilon_i)\right] (1 + \epsilon)$$

where

$$(28) \quad \begin{aligned} |\epsilon| &\leq 2^{-t} \\ |\epsilon_1| &< 3/2n \cdot 2^{-2t_2} \\ |\epsilon_r| &< 3/2(n+2 \cdot r) \cdot 2^{-2t_2} \quad r = 2, \dots, n \end{aligned}$$

5. Parallel Computation

We evaluate the sum $S_n = \sum_{i=1}^n a_i$, but instead of using the recursion relation (17) we define

$$S_{i,k} \equiv fl_p \left(\sum_{j=1}^k x_j \right)$$

where $fl_p(\cdot)$ denotes "floating point parallel" and use the recursion relationship

$$S_{1,r} \equiv fl_p (S_{1,\lceil r/2 \rceil} + S_{\lceil r/2 \rceil+1,r})$$

where $\lceil x \rceil$ denotes the smallest integer greater than or equal to x .

Such an algorithm can be implemented on a standard sequential computer, or on a parallel computer with p processors in $\lceil n/p \rceil \lceil \log_2 (\min(n, p)) \rceil$ time units such as the ILLIAC IV or the CDC STAR 100 [Stone, 1973]. Similar algorithms can be devised for repeated multiplications, divisions and for inner products.

Following Brent (1973) and Wilkinson (1963), we propose the following error bounds:

$$(29) \quad S_{1,n} \equiv fl_p \left(\sum_{i=1}^n x_i \right) \equiv \left(\sum_{i=1}^n x_i \right) (1 + \epsilon), \quad n \geq 2$$

where

$$(1 - 1/2B^{1-t})^{\lceil \log_2 n \rceil} \leq 1 + \epsilon \leq (1 + 1/2B^{1-t})^{\lceil \log_2 n \rceil}$$

Proof: We use induction. For $n = 2$, we obtain

$$S_{1,2} \equiv fl_p (S_{1,1} + S_{2,2}) \equiv (x_1 + x_2) (1 + \epsilon)$$

where

$$|\epsilon| \leq 1/2 B^{1-t}.$$

Assume the equation (29) holds for 2, ..., n-1, then

$$\begin{aligned} S_{1,n} &\equiv f\ell_p (S_{1, \lceil n/2 \rceil} + S_{\lceil n/2 \rceil + 1, n}) \\ &\equiv S_{1, \lceil n/2 \rceil} + S_{\lceil n/2 \rceil + 1, n} (1 + \epsilon^*) \quad (\epsilon^* \leq \frac{1}{2} B^{1-t}) \end{aligned}$$

By the induction hypothesis

$$S_{1, \lceil n/2 \rceil} \equiv \left[\sum_{i=1}^{\lceil n/2 \rceil} x_i \right] (1 + \epsilon_1)$$

and

$$S_{\lceil n/2 \rceil + 1, n} \equiv \left[\sum_{i=\lceil n/2 \rceil + 1}^n x_i \right] (1 + \epsilon_2)$$

where

$$(1 - \frac{1}{2} B^{1-t})^{\lceil \log_2 \lceil n/2 \rceil \rceil} \leq 1 + \epsilon_i \leq (1 + \frac{1}{2} B^{1-t})^{\lceil \log_2 \lceil n/2 \rceil \rceil} \quad i = 1, 2$$

It is simple to show that

$$\lceil \log_2 \lceil n/2 \rceil \rceil = \lceil \log_2 n/2 \rceil = \lceil \log_2 n \rceil - 1.$$

Hence

$$\begin{aligned} S_{1,n} &\equiv \left[\left(\sum_{i=1}^{\lceil n/2 \rceil} x_i \right) (1 + \epsilon_1) + \left(\sum_{i=\lceil n/2 \rceil + 1}^n x_i \right) (1 + \epsilon_2) \right] (1 + \epsilon^*) \\ &\equiv \left(\sum_{i=1}^n x_i \right) (1 + E) \end{aligned}$$

where

$$\begin{aligned} (1 - \frac{1}{2} B^{1-t})^{\lceil \log_2 n \rceil - 1} (1 - \frac{1}{2} B^{1-t}) &\leq 1 + E \\ 1 + E &\leq (1 + \frac{1}{2} B^{1-t})^{\lceil \log_2 n \rceil - 1} (1 + \frac{1}{2} B^{1-t}) \end{aligned}$$

Thus

$$(1 - \frac{1}{2} B^{1-t})^{\lceil \log_2 n \rceil} \leq 1 + E \leq (1 + \frac{1}{2} B^{1-t})^{\lceil \log_2 n \rceil}$$

q.e.d.

This proof can be repeated if the iterated operations are $-$, $\times$, or $\div$. Hence the same error bound holds for n of any of these four operations. If we have the inner product as $S = \sum_{i=1}^n a_i b_i$, then

$$fl_p \left(\sum_{i=1}^n a_i b_i \right) \equiv \left(\sum_{i=1}^n a_i b_i \right) (1 + \epsilon)$$

where

$$(1 - \frac{1}{2}B^{1-t})^{\lceil \log_2 n \rceil + 1} \leq 1 + \epsilon \leq (1 + 1/2B^{1-t})^{\lceil \log_2 n \rceil + 1}$$

by a similar argument.

Table 1.1 Fixed and Floating Point Operations

(a) Fixed Point Operations		
Operation	Computational Result	Error Bounds
1. $x_1 \pm x_2$	$x_1 \pm x_2$	
2. $x_1 \div x_2$	$x_1 \div x_2 + \epsilon$	$ \epsilon \leq 1/2 B^{-t}$
3. $\sum_{i=1}^n x_i$	$\sum_{i=1}^n x_i$	
4. $\sum_{i=1}^n a_i b_i$	$\sum_{i=1}^n a_i b_i + \epsilon$	$ \epsilon \leq 1/2n B^{-t}$
(b) Floating Point Operations		
Operation	Computational Result	Error Bounds
1. $x_1 \pm x_2$	$(x_1 \pm x_2)(1+\epsilon)$	$ \epsilon \leq 1/2 B^{1-t}$
2. $x_1 \div x_2$	$x_1 \div x_2 + \epsilon$	$ \epsilon \leq 1/2 B^{1-t}$
3. $\sum_{i=1}^n x_i$	$\sum_{i=1}^n x_i (1+\epsilon_i)$	$(1-1/2B^{1-t})^{n-1} \leq 1+\epsilon_1 \leq (1+1/2B^{1-t})^{n-1}$ $(1-1/2B^{1-t})^{n-r+1} \leq 1+\epsilon_r \leq (1+1/2B^{1-t})^{n-r+1}$ $(r = 2, \dots, n)$
4. $\sum_{i=1}^n a_i b_i$	$\sum_{i=1}^n a_i b_i (1+\epsilon_i)$	$(1-1/2B^{1-t})^n \leq 1+\epsilon_1 \leq (1+1/2B^{1-t})^n$ $(1-1/2B^{1-t})^{n-r+2} \leq 1+\epsilon_r \leq (1+1/2B^{1-t})^{n-r+2}$ $(r = 2, \dots, n)$
5. $\prod_{i=1}^n x_i$	$\left(\prod_{i=1}^n x_i\right)(1+E)$	$(1-1/2B^{1-t})^{n-1} \leq 1+E \leq (1+1/2B^{1-t})^{n-1}$

References

- Brent, Richard P., "The Parallel Evaluation of Arithmetic Expressions in Logarithmic Time", Complexity of Sequential and Parallel Algorithms, Academic Press, New York, 1973.
- Grau, Albert A. and Erwin H. Bareiss, Statistical Aspects of Machine Rounding, ERDA Report COO - 2280 - 34, Northwestern University, 1977.
- Stone, Harold S., "Problems of Parallel Computation", Complexity of Sequential and Parallel Algorithms, Academic Press, New York, 1973.
- Sturbenz, Pat H., Floating Point Computation, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1974.
- Wilkinson, J. H., The Algebraic Eigenvalue Problem, Clarendon Press, Oxford, 1965.
- _____, Rounding Errors in Algebraic Processes, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1963.

II. THE DISTRIBUTION OF THE MOST AND LEAST SIGNIFICANT DIGITS

Let $m \cdot B^E$ be the exact result of some floating point computation, and let $m^* \cdot B^E$ be the computed machine result. Then the error in representation is $(m-m^*) \cdot B^E$, and the relative error is $\frac{(m-m^*) \cdot B^E}{m \cdot B^E} = \frac{m-m^*}{m}$.

(Not all cases of floating point rounding fall into this category. Exceptions are discussed in Section III.) To characterize the distribution of this error we must characterize the distribution of m and the distribution of $m-m^*$.

The distribution of m is closely approximated by the distribution of the most significant digit, and the distribution of $m-m^*$ is closely approximated by the distribution of the least significant digit. We show that the most significant digits have the density $r(x) = 1/x \ln B$ on $[1, B)$ and the least significant digits have approximately the density $f(x) = 1/B$ on $[0, B)$.

1. The Benford Hypothesis

The first known work on the most significant digit was done by Benford (1938). He observed that a common logarithm table showed more wear in the beginning pages for those numbers beginning with "1" or "2" than the last pages, for those numbers beginning with "8" or "9". On the surface, at least, this appeared to be a counter intuitive result.

So Benford took a sample of 20,229 numbers from newspapers, magazines, scientific tables, addresses, areas of rivers, population table entries, etc. He disregarded decimal points and leading zeroes, that is, he

checked only to see whether the first significant digit was 1,2,3,4,5, 6,7,8, or 9. His results are reproduced in Table 2.1.

Benford, from the assumption that "nature" counts logarithmically, proved that for numbers of length r the probability of the first significant digit being A was given by the formula

$$(1) \quad F_A^r = \begin{cases} \left[\ln [10 \cdot (2 \cdot 10^{r-1} - 1) / (10^r - 1)] + 8/10^r \right] / \ln 10 & \text{for } A = 1 \\ \left[\ln [(A + 1) (10^{r-1} - 1) / (A \cdot 10^{r-1} - A)] \right] / \ln 10 & \text{for } A \neq 1 \end{cases}$$

One may then show that

$$(2) \quad \lim_{r \rightarrow \infty} F_A^r = \log_{10} [(A + 1) / A] = \int_A^{A+1} \frac{dx}{x \ln 10}$$

meaning that the longer the number the more closely it conforms to the logarithmic frequency.

Table 2.1

Benford's sample of numbers

Natural Number	Number Interval	Observed Frequency*	Logarithmic Frequency**	Difference Observed Computed
1	[1,2)	0.306	0.301	0.005
2	[2,3)	0.185	0.176	0.009
3	[3,4)	0.124	0.125	-0.001
4	[4,5)	0.094	0.097	-0.003
5	[5,6)	0.080	0.079	0.001
6	[6,7)	0.064	0.064	-0.003
7	[7,8)	0.051	0.058	-0.007
8	[8,9)	0.049	0.051	-0.002
9	[9,10)	0.047	0.046	0.001

*Observed Frequency from 20,229 number sampled.

**Logarithmic frequency = $\log_{10} [(A + 1) / A]$ where A is a natural number.

Obviously we cannot prove mathematically that a given sample space conforms to a certain probabilistic law. We must make some assumptions about the sample space involved. However, suppose that the areas of the rivers Benford used in his sample had been measured in square kilometers instead of square miles. The distribution of the first significant digit should be the same. A reasonable assumption then is that the distribution of the first significant digit is invariant under scaling.

Proofs of Benford's hypothesis have since been proposed by Pinkham (1961), Raimi (1969) and Flehinger (1966). We follow Raimi's proof.

Theorem 2.1 The only continuous density function on $[1, B)$ that is invariant under scaling is the density

$$(3) \quad f(x) = 1/x \ln B.$$

Proof. Let R be the set of real numbers, let $(R, +)$ be the additive group of real numbers. There exists a large class J of real-valued set-functions j , defined for all subsets of R , which have the following properties.

For all $j \in J$

$$(a) \quad j(A+S) = j(A) \text{ for every } S \in R, A \subseteq R$$

$$(b) \quad j(A \cup B) = j(A) + j(B) \text{ if } A \cap B = \emptyset$$

$$(c) \quad j(A) \geq 0 \text{ for every } A \subseteq R$$

$$(d) \quad j(R) = 1$$

where $A + S = \{a+S: a \in A\}$. Note that the j 's are finitely but not countably additive.

For certain subsets $A \subset R$ if $j_1 \neq j_2$, then in general $j_1(A) \neq j_2(A)$.

A is said to be J -measurable or Banach measurable if $\{j(A): j \in J\}$ is a

single number, called its Banach measure. There exists at least one family of sets which is Banach measurable namely:

$$(4) \quad K_a = \bigcup_{n=-\infty}^{\infty} [n, n+a)$$

where $a \in (0, 1]$. The Banach measure of K_a is a^* .

The mapping $\log_B: R^+ \rightarrow R$, where $(R^+, *)$ is the multiplicative group of positive real numbers, is an isomorphism of groups from $(R^+, *)$ onto $(R, +)$. We define a set T of scale-invariant finitely additive, positive, and normalized set functions over R^+ , such that, if $t \in T$ then

$$(a) \quad t(AS) = t(A) \quad \text{for all } S \in R^+, \quad A \subseteq R^+$$

$$(b) \quad t(A \cup B) = t(A) + t(B) \quad \text{if } A \cap B = \emptyset$$

$$(c) \quad t(A) \geq 0 \quad \text{for all } A \subseteq R^+$$

$$(d) \quad t(R^+) = 1$$

where $AS = \{a*S: a \in A\}$.

If $j \in J$, its image $t \in T$ is defined by $t(A) = j(\log_B A)$. This is a bijection of J onto T . A subset $A \subseteq R^+$ is scale measurable if $\{t(A): t \in T\}$ is a single number. Hence A is scale measurable if and only if $\log_B A$ is Banach measurable.

Let $D_a \subseteq R^+$ be the set of all positive reals of the form $m*B^E$, $m \in [1, B)$ such that $m < a$, $a \in (1, B]$. Thus

$$(5) \quad D_a = \bigcup_{n=-\infty}^{\infty} [B^n, aB^n)$$

We define a measure of D_a as a subset of R^+ by computing $t \in T$. Since

* There is also a family of sets which is not Banach measurable, i.e., for any $x \in R$, the interval $(-\infty, x)$ is not Banach measurable.

$t(D_a) = j(\log_B D_a)$ where j is the Banach measure set function corresponding to t , and

$$(6) \log_B D_a = \bigcup_{n=-\infty}^{\infty} [n, n + \log_B a) = K_{\log_B a}$$

it follows that $t(D_a) = j(\log_B D_a) = \log_B a$, regardless of which j is used.

Hence every scaled measure on R^+ gives measure $\log_B a$ to the set D_a .

$$\text{Since } \log_B a = \int_1^a \frac{dx}{x \ln B} \quad a \in (1, B)$$

we have proven the theorem.

The set of all real positive numbers with first digit less than or equal to a is D_{a+1} . Hence Benford's hypothesis for the first significant digit holds. By a simple change of variable we can show that $(\log_B A) \pmod{1}$, $A \in R^+$, has a uniform distribution. If $A = m \cdot B^E$, $E \in \{\text{integers}\}$ the $(\log_B A) \pmod{1} = \log_B m$. Since m has a reciprocal distribution then $x = \log_B m$ has the density

$$f(x) = \frac{1}{m \ln B} \cdot \frac{dm}{dx} = \frac{1}{B^x \cdot \ln B} (B^x \cdot \ln B) = 1 \quad \text{on } [0, 1).$$

2. Multiplication and Division Reinforce the Reciprocal Distribution

Adhikari and Sarkar (1968) and Hamming (1970) showed that repeated multiplications and divisions move any "reasonable" underlying distribution toward the reciprocal distribution and once the reciprocal distribution is obtained, it persists.

Let $a_1 = x \cdot B^{E_1}$ and $a_2 = y \cdot B^{E_2}$ where B is the base of the floating point number system. If we consider a_1 and a_2 to be normalized then $x, y \in [1/B, 1)$. Let $a_3 = a_1 \cdot a_2 = z \cdot B^{E_3}$ be normalized, and let $f(x)$, $g(y)$, and $h(z)$ be the probability density functions for x, y , and z respectively, and $F(x), G(y)$ and $H(z)$ the corresponding cumulative distributions. We determine $h(x)$ from $f(x)$ and $g(x)$ as follows.

For both multiplication and division the exponents are directly combined and thus do not enter into the distribution of the result.

For multiplication, $H(z)$ is given by

$$(7) \quad H(z) = \int_{1/B}^z \int_{1/B}^{z/Bx} f(x)g(y)dydx + \int_{1/B}^z \int_{1/Bx}^1 f(x)g(y)dydx \\ + \int_z^1 \int_{1/Bx}^{z/x} f(x)g(y)dydx$$

where the region of integration is illustrated in Figure 2.1. Hence

$$H(z) = \int_{1/B}^z f(x) [G(z/Bx) - G(1/B) + G(1) - G(1/Bx)] dx \\ + \int_z^1 f(x) [G(z/x) - G(1/Bx)] dx$$

Differentiating with respect to z yields the density

$$(8) \quad h(z) = f(z) [G(1/B) - G(1/B) + G(1) - G(1/Bz) - G(1) + G(1/Bz)]$$

$$\begin{aligned}
 (8) \quad & + \int_{\frac{1}{B}}^z \frac{f(x)}{Bx} g(z/Bx) dx + \int_z^1 \frac{f(x)}{x} g(z/x) dx \\
 & = \frac{1}{B} \int_{\frac{1}{B}}^z \frac{f(x)}{x} g(z/Bx) dx + \int_z^1 \frac{f(x)}{x} g(z/x) dx
 \end{aligned}$$

Figure 2.1

The region of integration for multiplication, equation (7).
Adapted from Hamming (1970).

For division, the argument is similar. If we have $a_3 = a_1/a_2$ where $a_1 = x \cdot B^{E_1}$, $a_2 = y \cdot B^{E_2}$ and $a_3 = z \cdot B^{E_3}$, the cumulative distribution function of z is

$$(9) \quad H(z) = \int_{\frac{1}{B}}^z \int_{\frac{1}{B}}^x f(x)g(y)dydx + \int_{\frac{1}{B}}^z \int_{\frac{x}{z}}^1 f(x)g(y)dydx + \int_z^1 \int_{\frac{x}{Bz}}^x f(x)g(y)dydx$$

where the region of integration is illustrated in Figure 2.2.

Figure 2.2

The region of integration for division, equation (9)
Adapted from Hamming (1970).

Once again differentiating with respect to z yields the probability density function

$$h(z) = f(z) [G(z) - G(\frac{1}{B}) + G(1) - G(1) + G(z) + G(\frac{1}{B})]$$

$$(10) + \int_{\frac{1}{B}}^x f(x) [-g(\frac{x}{z}) (\frac{-x}{z^2})] dx + \int_z^1 f(x) [-g(\frac{x}{Bz}) (\frac{-x}{Bz^2})] dx$$

$$= \frac{1}{z^2} \int_{\frac{1}{B}}^z xf(x) g(\frac{x}{z}) dx + \frac{1}{Bz^2} \int_z^1 xf(x) g(\frac{x}{Bz}) dx$$

Persistence of the Reciprocal Distribution

For the case of the product set $g(y) = \frac{1}{y \ln B}$. Then for any density

$f(\cdot)$ we get

$$(11) \quad h(z) = \frac{1}{B} \int_{\frac{1}{B}}^z \frac{f(x)}{x} \frac{Bx}{z \ln B} dx + \int_z^1 \frac{f(x)}{x} \frac{x}{z \ln B} dx = \frac{1}{z \ln B}$$

Making the same assumption for the quotient

$$(12) \quad h(z) = \frac{1}{z^2} \int_{\frac{1}{B}}^z xf(x) \frac{z}{x \ln B} dx + \frac{1}{Bz^2} \int_z^1 xf(x) \frac{Bz}{x \ln B} dx = \frac{1}{z \ln B}$$

In the special case of $f(x)$ being the "spike distribution" with all its probability at $x = 1$ we see that the reciprocal of a random variable having the reciprocal distribution has the reciprocal distribution. Hence, the case of the quotient x/y , where x has the reciprocal distribution is covered by the product form.

If in a long sequence of multiplications and divisions at least one factor has the reciprocal distribution, then regardless of the distributions of the other factors, the reciprocal distribution is the one which persists.

Convergence to the reciprocal distribution

Following Hamming (1970) let $r(z) = 1/z \ln B$. We define the measure of distance for a general distribution $h(\cdot)$ by

$$\max_{z \in [1/B, 1]} \left| \frac{h(z) - r(z)}{r(z)} \right| = D\{h\}$$

and show that starting from any continuous distribution $h(z)$, repeated multiplications and divisions give us the reciprocal distribution. We begin with repeated multiplication.

With $r(z) = 1/z \ln B$, equation (11) is

$$r(z) = \frac{1}{B} \int_{\frac{1}{B}}^z \frac{f(x)}{x} r\left(\frac{z}{Bx}\right) dx + \int_z^1 \frac{f(x)}{x} r\left(\frac{z}{x}\right) dx$$

Let equation (8) generate $h(z)$ from the continuous distributions $f(x)$ and $g(y)$.

Then

$$\begin{aligned} \frac{h(z) - r(z)}{r(z)} &= \frac{1}{B} \int_{\frac{1}{B}}^z \frac{f(x)}{x} \left[\frac{g(z/Bx) - r(z/Bx)}{r(z)} \right] dx \\ &+ \int_z^1 \frac{f(x)}{x} \left[\frac{g(z/x) - r(z/x)}{r(z)} \right] dx \end{aligned}$$

Since $Bxr(z) = Bx/z \ln B = r(z/Bx)$

and $xr(z) = x/z \ln B = r(z/x)$

we have

$$\begin{aligned} \frac{h(z) - r(z)}{r(z)} &= \int_{\frac{1}{B}}^z f(x) \left[\frac{g(z/Bx) - r(z/Bx)}{r(z/Bx)} \right] dx \\ &+ \int_z^1 f(x) \left[\frac{g(z/x) - r(z/x)}{r(z/x)} \right] dx \end{aligned}$$

which means

$$\left| \frac{h(z) - r(z)}{r(z)} \right| \leq \int_{\frac{1}{B}}^z f(x) D\{g\} dx + \int_z^1 f(x) D\{g\} dx = D\{g\}$$

and therefore

$$D\{h\} \leq D\{g\}.$$

We remark that equality occurs when $f(x) = \delta(1-x)$, the "spike distribution" at $x = 1$.

Hamming (1970) examined the uniform distribution. Let

$$(13) \quad f(x) = g(y) = 1/(1-1/B) = B/(B-1)$$

Then the distribution for $z = xy$ becomes

$$\begin{aligned} h(z) &= \frac{1}{B} \left(\frac{B}{B-1} \right)^2 \int_{\frac{1}{B}}^z \frac{dx}{x} + \left(\frac{B}{B-1} \right)^2 \int_z^1 \frac{dx}{x} \\ &= \frac{B}{(B-1)^2} \{ \ln B - (B-1) \ln z \} \end{aligned}$$

If $B = 10$ then

$$(14) \quad h(z) = \frac{10}{81} \{ \ln 10 - 9 \ln z \}$$

The distance between the uniform distribution (13) for $B = 10$, and the reciprocal distributions (3) is

$$\max_{1/10 \leq z \leq 1} \left| z \frac{10 \ln 10}{9} - 1 \right| = 1.558$$

while the distance for equation (14) is 0.3454. Table 2.2 gives these results and those for the distribution of three and four factors.

Table 2.2 The Distance from the reciprocal distribution of an extended product as function of the number of factors selected from a uniform distribution.

Number of Factors	Distance
1	1.558
2	0.3454
3	0.0980
4	0.0289

Using a similar argument for division:

$$\frac{h(z) - r(z)}{r(z)} = \frac{1}{z^2} \int_{\frac{1}{B}}^z xf(x) \left[\frac{g(\frac{x}{z}) - r(\frac{x}{z})}{r(z)} \right] dx$$

$$+ \frac{1}{Bz^2} \int_z^1 xf(x) \left[\frac{g(\frac{x}{Bz}) - r(\frac{x}{Bz})}{r(z)} \right] dx$$

Since

$$\frac{z^2 r(z)}{x} = r\left(\frac{x}{z}\right) \text{ and } \frac{Bz^2 r(z)}{x} = r\left(\frac{x}{Bz}\right)$$

we have

$$\left| \frac{h(z) - r(z)}{r(z)} \right| \leq D\{g\} \left[\int_{\frac{1}{B}}^z f(x) dx + \int_z^1 f(x) dx \right] = D\{g\}$$

which means $D\{h\} \leq D\{g\}$.

When $g(y) = f(x) = B/(B-1)$ are uniform

$$h(z) = \frac{1}{2(B-1)} \left[B + \frac{1}{z^2} \right]$$

which for base 10 is

$$h(z) = \frac{1}{18} \left[10 + \frac{1}{z^2} \right]$$

and has a distance of 0.4071.

3. The Distribution of the Intermediate and Least Significant Digits

Benford (1938) extended his hypothesis to the t^{th} position for base 10. First we prove the hypothesis for a general base B , following a proof by Feldstein and Goodman (1976). From Theorem 2.1 we know that for a real number A taken as a random variable, the distribution of $(\log_B A) \bmod 1$ is uniform. This is the fundamental assumption underlying

Theorem 2.2. The probability that the first k digits of a number A are exactly $a_1, a_2, \dots, a_k$ and in that order is

$$(15) \quad P_1(G, B^k) = \log_B \frac{G+1}{G}$$

where $G = \sum_{i=1}^k a_i B^{k-i}$, and B is the radix of the number system.

Proof: Let $A = mB^E > 0$ where E is a signed integer, and $m \in [1/B, 1)$. Let

$$\psi(s) = \sum_{i=1}^k a_i s^{k-i}$$

We find the probability that the following sharp inequality holds,

$$(16) \quad 0 \leq mB^k - \psi(B) < 1$$

Since $mB = B^{(\log_B A) \bmod 1}$, this inequality is equivalent to

$$(17) \quad \log_B [\psi(B)/B^{k-1}] \leq (\log_B A) \bmod 1 < \log_B [(\psi(B) + 1)/B^{k-1}]$$

By Theorem 2.1, $(\log A) \bmod 1$ is uniformly distributed on $[0,1]$. Since the inequality is sharp, the probability that the inequality (16) is true is equal to the length of the interval indicated by (17), i.e., $\log_B [(\psi(B) + 1)/\psi(B)]$. Since $G = \psi(B)$, the theorem is proven.

Theorem 2.3. Given the integer $a \in [0, B-1]$, and let A be a number in the system over radix B . Then the probability that the t^{th} digit of A is a , is

$$(18) \quad P_t(a, B) = \begin{cases} \log_B \frac{a+1}{a} & \text{if } a \geq 1 \text{ and } t = 1 \\ \sum_{\ell=B^{t-2}}^{B^{t-1}-1} \log_B \frac{\ell B + a + 1}{\ell B + a} & \text{if } a \geq 0 \text{ and } t > 1 \end{cases}$$

Proof. The first equation in (18) follows from Theorem 2.1. The second equation follows from Theorem 2.2 by summing up the probabilities over all values for a_i ($i = 1, 2, \dots, t-1$) such that $0 \leq a_i \leq B-1$ with the exception that $a_1 > 0$. Hence, the numbers G in (15) have the form

$$G = \ell B + a$$

where $k = t$, $a_t = a$ is fixed, and $\ell = \sum_{i=1}^{t-1} a_i B^{t-1-i}$. Clearly, the range

of ℓ is determined by the $t-1$ tuple $(a_1, a_2, \dots, a_{t-1})$ with $(1, 0, \dots, 0)$

corresponding to $\ell = B^{t-2}$, the lowest value of ℓ , and $(B-1, B-1, \dots, B-1)$ corresponding to $\ell = B^{t-1} - 1$, the highest value of ℓ . Summing over all values of ℓ yields (18).

Theorem 2.4. Let $G = \sum_{i=0}^{k-1} a_{t+i} B^{k-1-i}$. Then the probability that the digits in positions $t, t+1, \dots, t+k-1$ of a number A in the radix system B are the given digits $a_t, a_{t+1}, \dots, a_{t+k-1}$ respectively is

$$(19) \quad P_t(G, B^k) = \sum_{\ell=B^{t-2}}^{B^{t-1}-1} \log_B \frac{\ell B^k + G + 1}{\ell B^k + G}$$

where $0 \leq a_i \leq B-1$, and $t > 1$.

Proof. [Feldstein and Goodman, 1976]: We combine the methods that yielded the proofs for theorems 2.2 and 2.3, and sum over all $(t-1)$ -tuples $(a_1, a_2, \dots, a_{t-1})$, $0 < a_1 < B$, $0 \leq a_i < B$ ($i=2, \dots, t-1$), which map into all possible combinations of the $t-1$ leading digits or equivalently into the numbers

$$\ell = \sum_{i=1}^{t-1} a_i B^{t-1-i}$$

For fixed ℓ , the probability that the digits in position $t, t+1, \dots, t+k-1$ are the given $a_t, a_{t+1}, \dots, a_{t+k-1}$ is

$$\log_B \frac{\ell B^k + G + 1}{\ell B^k + G}$$

Summing over the range of ℓ yields (19).

The expressions for the probabilities (15), (18), (19) have the common form (Definition):

$$(20) \quad P_t(G, B^k) = \sum_{\ell=\lfloor B^{t-2} \rfloor}^{B^{t-1}-1} \log_B \frac{\ell B^k + G + 1}{\ell B^k + G} = \frac{1}{\ell \ln B} \times \sum_{\ell=B}^{B^{t-1}-1} \ell \ln \frac{1 + \frac{G+1}{B^k \ell}}{1 + \frac{G}{B^k \ell}} \quad \begin{matrix} (t=1) \\ (t>1) \end{matrix}$$

$$\text{where } G = \sum_{i=0}^{k-1} a_{t+i} B^{k-1-i}$$

The $t-1$ tuple $(a_t, \dots, a_{t+k-1})_B$ indicates the position and value for

the leading, intermediate, or trailing digits of a number $A = m \cdot B^E$,

$$m = \sum_{i=1}^{\infty} a_i B^{-i}. \quad \text{The range of } G \text{ is } [1, B^k-1] \text{ when } t=1, \text{ and } [0, B^k-1]$$

when $t > 1$. Several cases for (20) are tabulated in Table 2.3.

Theorem 2.5. The probability for intermediate and trailing digits $a_t, a_{t+1}, \dots, a_{t+k-1}$ of a random number A over the radix system B is

$$(21) \quad \lim_{t \rightarrow \infty} P_t(G, B^k) = B^{-k}$$

for all G . Thus, intermediate and trailing digits approach a uniform distribution for portions $t \gg 1$.

To establish this theorem we prove

Lemma 2.1 For $t > 1$, equation (20) has the expansion

$$(22) \quad P_t(G, B^k) = \frac{1}{B^k} + \frac{(B^k - G) - (G+1)}{2B^{2k} \ln B} \cdot \frac{B(B-1)}{B^t} + O(B^{-2t})$$

Proof: [Bustoz, Feldstein, and Goodman, 1978]

Table 2.3

Values of $P_t(a, B)$ for $B = 2, 4, 8, 10$ and 16
 Reproduced from Feldstein and Goodman (1976)

$B = 2$

t	a = 0	1
1		1.0000
2	.5849	.4150
3	.5443	.4557
4	.5224	.4766
5	.5113	.4887
6	.5056	.4944
7	.5028	.4972
8	.5014	.4986
9	.5007	.4993
10	.5004	.4996
11	.5002	.4998
12	.5001	.4999
13	.5000	.5000
	.	.
	.	.

$B = 4$

t	a = 0	1	2	3
1		.5000	.2925	.2075
2	.3037	.2610	.2297	.2056
3	.2630	.2539	.2455	.2377
4	.2532	.2510	.2489	.2469
5	.2508	.2503	.2497	.2492
6	.2502	.2501	.2499	.2498
7	.2501	.2500	.2500	.2499
8	.2500	.2500	.2500	.2500
	.	.	.	.
	.	.	.	.
	.	.	.	.

B = 8

t	a =	0	1	2	3	4	5	6	7
1			.3333	.1950	.1383	.1073	.0877	.0741	.0642
2		.1505	.1410	.1330	.1260	.1199	.1145	.1097	.1053
3		.1279	.1271	.1262	.1254	.1246	.1237	.1230	.1222
4		.1254	.1253	.1252	.1251	.1249	.1248	.1247	.1246
5		.1250	.1250	.1250	.1250	.1250	.1250	.1250	.1250
		.	.	.	.	.	.	.	.

B = 10

t	a =	0	1	2	3	4	5	6	7	8	9
1			.3010	.1761	.1249	.0969	.0792	.0669	.0580	.0512	.0458
2		.1197	.1139	.1088	.1043	.1003	.0967	.0943	.0904	.0876	.0850
3		.1018	.1014	.1010	.1006	.1002	.0998	.0994	.0990	.0986	.0983
4		.1002	.1001	.1001	.1001	.1000	.1000	.0999	.0999	.0999	.0998
5		.1000	.1000	.1000	.1000	.1000	.1000	.1000	.1000	.1000	.1000
		.	.	.	.	.	.	.	.	.	.

B = 16

t	a=0	1	2	3	4	5	6	7	8	9	A(10)	B(11)	C(12)	D(13)	E(14)	F(15)
1		.2500	.1462	.1038	.0805	.0658	.0556	.0482	.0425	.0380	.0344	.0314	.0289	.0267	.0249	.0233
2		.0737	.0717	.0698	.0681	.0665	.0650	.0636	.0623	.0611	.0599	.0588	.0578	.0568	.0558	.0541
3		.0631	.0630	.0630	.0629	.0628	.0627	.0626	.0625	.0625	.0624	.0623	.0622	.0621	.0620	.0619
4		.0625	.0625	.0625	.0625	.0625	.0625	.0625	.0625	.0625	.0625	.0625	.0625	.0625	.0625	.0625
		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

If this Lemma is true, Theorem 2.5 is proven. To prove Lemma 2.1 we use the sums

$$(23) \quad S_j = \sum_{\ell=B^{t-2}}^{B^{t-1}-1} \frac{1}{\ell^j} = \int_{B^{t-2}}^{B^{t-1}} \frac{d\tau}{\tau^j} + \frac{1}{2} g(B^j) + \frac{1}{12} g(B^{j+1}) + R_j$$

where

$$(24) \quad g(x) = \frac{x(x-1)}{x^t}$$

$$(25) \quad R_j = -\frac{1}{30} \binom{j+3}{4} \frac{B^{2(j+3)}(B-1)}{B^{(j+3)t} [1+\theta(B-1)]^{j+4}} = O(B^{(j+3)t}) \quad (0 < \theta < 1)$$

The right side of (23) is obtained from the Euler sum formula with remainder term. In particular

$$(26.1) \quad S_1 = \ln B + \frac{1}{2} g(B) + \frac{1}{12} g(B^2) + O(B^{-4t})$$

$$(26.2) \quad S_2 = g(B) + \frac{1}{2} g(B^2) + \frac{1}{6} g(B^3) + O(B^{-5t})$$

$$(26.3) \quad S_3 = \frac{1}{2} g(B^2) + \frac{1}{2} g(B^3) + O(B^{-4t})$$

$$(26.4) \quad S_4 = \frac{1}{3} g(B^3) + O(B^{-4t})$$

Next we expand $(\ln B)P_t(G, B^k)$ for $t = 1$ into a Taylor series with respect to B^{-k} (i.e., we expand $\ln(1+B(x+1)) - \ln(1+Bx)$ with respect to $x = B^{-k}$).

Let

$$(27) \quad \alpha_j = \frac{(G+1)^j - G^j}{jB^{kj}}$$

Then the expansion, with the 4th term the remainder, is

$$(28) \quad \ln B P_t(G, B^k) = \alpha_1 S_1 - \alpha_2 S_2 + \alpha_3 S_3 + \alpha_4 S_4 / \theta^{4k} (0 < \theta < 1)$$

$$= \alpha_1 \ln B + \frac{1}{2} (\alpha_1 - \alpha_2) g(B) + \left(\frac{1}{12} \alpha_1 - \frac{1}{2} \alpha_2 + \frac{1}{2} \alpha_3 \right) g(B^2) + O(B^{-4t})$$

We evaluate this expression, using (24) and (27), simplify and retain the first three terms. This yields (22). Hence Lemma 2.1 is proven.

For later reference, an additional term has been carried along in equations (26) and (28).

Corollary 2.1 The probability that the t^{th} significant figure in a number A is a , is

$$(29) \quad P_t(a, B) = \frac{1}{B} + \frac{B-2a-1}{2B \ln B} \cdot \frac{B-1}{B^t} + O(B^{-2t})$$

when $t > 1$. If $t = 1$, it is $P_1(a, B) = \log_B(1+a^{-1})$.

Proof: Equation (29) is the special case $k = 1$ of Lemma 2.1

The case $t = 1$ is given by (15).

Corollary 2.2 Define the maximum difference between the probability distribution of the t^{th} significant figure a of a number A and the uniform distribution $1/B$ for $t > 1$ by

$$(30) \quad \Delta(P_t, B) = \max_{0 \leq a < B} |P_t(a, B) - B^{-1}|$$

For $t \gg 1$ follows $\Delta(P_t, B) \rightarrow 0$, and in particular

$$(31) \quad \lim_{t \rightarrow \infty} B^t \Delta(P_t, B) = \frac{(B-1)^2}{2B \ln B}$$

Proof: From Corollary 2.1 follows, by (29),

$$\Delta(P_t, B) = \frac{B-1}{2B \ln B} \max_{(a)} |B-1-2a| + O(B^{-t})$$

The maximum is reached for $a = 0$ and $a = B-1$. It is interesting to note that a minimum difference is attained by $a = \lfloor (B-1)/2 \rfloor$.

References

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

Benford, Frank, "The Law of Anamolous Numbers", Proc. American Phil. Soc., 78(1938), 551-72.

Bustoz, Joaquin, Alan Feldstein and Richard Goodman, "Improved Trailing Digits Estimates Applied to Optimal Computer Arithmetic", To appear, JACM, 1978.

Feldstein, Alan and Richard Goodman, "Convergence Estimates for the Distribution of Trailing Digits", JACM, 23(1976), 287-97.

Flehinger, B. J., "On the Probability that a Random Integer has Initial Digit A", Amer. Math. Monthly, 73(1966), 1056-61.

Hamming, R. W., "On the Distribution of Numbers", Bell System Technical Journal, Vol. 49, No. 8 (1970), 1609-25.

Nathan, Lew H., "Probabilistic Distribution of the Most Significant Digit in Computer Represented Numbers and Its Behavior Under Iterated Fixed and Floating Point Operations", MS Thesis, Northwestern University, Evanston, Illinois, August 1973.

Pinkham, Roger S., "On the Distribution of First Significant Digits", Ann. Math. Stat., 32(1961), 1223-30.

Raimi, R. A., "On the Distribution of First Significant Digits", Amer. Math. Monthly, 74 No. 2 (1966), 342-8.

III. FLOATING POINT ARITHMETIC

There are several rounding options in use for floating point arithmetic units on digital computers. What is commonly called rounding is a method which rounds to the nearest floating point number and rounds up in the case of a tie. We call this method symmetric rounding.

Another rounding method is chopping or truncating which simply disregards the remaining digits. An alternative version of chopping is jamming (also called the "Von Neumann round") which unconditionally forces the last binary digit to a "1".

R* - rounding is identical to rounding except that it breaks ties by jamming. However, most rounding does not fit the ideals of rounding or chopping. Symmetric rounding is usually implemented through the use of guard digits.

Kaneko and Liu (1973) and Tsao (1974) analyzed symmetric rounding and chopping. They used the assumptions proven in Section II to establish distributions of the round-off error for these two methods. The distributions they derive are continuous.

Kuki and Cody (1973) compared a six-digit hexadecimal floating point system with a 22-bit binary system. They sampled thousands of operations and determined the accuracy of several rounding methods with these two systems. Cody (1973) gave a theoretical comparison of these two systems with a system using a 23-bit binary mantissa and quaternary exponents.

Goodman and Feldstein (1975, 1977) with Bustoz (1978) established

discrete distributions for rounding errors from floating point multiplication considering the parameters of guard digits and normalization options.

Kuck, Parker, and Sameh (1977) tested the bias of the rounding methods of jamming, symmetric rounding, R*-rounding, chopping and a new kind of rounding called ROM rounding. We also point to the exceptions which are always present in standard symmetric rounding intervals as considered by Grau and Bareiss (1977). These intervals occur around the mantissa $1/B$.

1. The Ideals: Symmetric Rounding and Chopping

Symmetric rounding takes a real number in the interval $[(m^* - \frac{1}{2}B^{-t}) \cdot B^E, (m^* + \frac{1}{2}B^{-t}) \cdot B^E]$ and rounds it to m^*B^E . By Theorem 2.5 we may consider the numbers in this interval to have a uniform distribution and the number m to have a reciprocal distribution. If we let $z = m - m^*$ where mB^E is the number to be rounded, then $e = z/m$ is the relative rounding error.

For chopping we make the same assumptions about the distributions of z and m . In particular we assume that z has constant distribution

$$(1) \quad f(z) = B^t \quad \text{for } z \in [0, B^{-t})$$

and that the fractional part or mantissa has the inverse distribution

$$(2) \quad g(m) = 1/m \ln B, \quad m \in [1/B, 1).$$

For floating point multiplication we know that these assumptions are reasonable (Section II). From (1) and (2) we conclude that the

range of a given relative error ϵ_o is bounded by

$$0 \leq \epsilon_o = z/m < B^{-t}/B^{-1} = B^{1-t},$$

and we can calculate the cumulative distribution

$$H_C(\epsilon_o) = \text{Prob}(\epsilon < \epsilon_o)$$

for ϵ_o if chopping is used. We distinguish between the cases (a) when $\epsilon_o < B^{-t}$ and (b) when $B^{-t} \leq \epsilon_o < B^{1-t}$, observing the bounds of z given by (1). Thus

$$(3) \quad H_C(\epsilon_o) = \int_0^{\epsilon_o/B} dz \int_{1/B}^1 dm f(z)g(m) + \int_{\epsilon_o/B}^{d(\epsilon_o)} dz \int_{z/\epsilon_o}^1 dm f(z)g(m)$$

where

$$(4) \quad d(\epsilon_o) = \begin{cases} \epsilon_o & \text{if } 0 \leq \epsilon_o < B^{-t} \\ B^{-t} & \text{if } B^{-t} \leq \epsilon_o < B^{1-t} \end{cases} \quad \begin{matrix} (a) \\ (b) \end{matrix}$$

The domains of integration are illustrated in Figure 3.1. Substituting (1) and (2) in (3) yields

$$(5) \quad H_C(\epsilon_o) = B^t / \ln B \left\{ \epsilon_o \ln B / B + \int_{\epsilon_o/B}^{d(\epsilon_o)} dz (\ln \epsilon_o - \ln z) \right\}.$$

To obtain the density $h_C(\epsilon_o)$, we differentiate (5) with respect to ϵ_o . Hence

$$(6) \quad h_C(\epsilon_o) = \begin{cases} (B-1)B^{t-1} / \ln B & \text{if } 0 \leq \epsilon_o \leq B^{-t} \\ (1/\epsilon_o - B^{t-1}) / \ln B & \text{if } B^{-t} \leq \epsilon_o < B^{1-t} \end{cases}$$

Figure 3.2 illustrates the density $h_C(\epsilon_o)$ for the base $B = 2$.

By a similar procedure we derive the density h_R for symmetric rounding. We note that in this case

$$(7) \quad -\frac{1}{2}B^{-t} \leq z = m - m^* < \frac{1}{2}B^{-t},$$

Figure 3.1

Domain of integration for cumulative floating point error distribution under multiplication with chopping. (Equation(3)).

Figure 3.2

Error probability density $h_C(\epsilon)$, Equation (6), for floating point multiplication with chopping on a binary computer ($B = 2$).

and hence the domain for the relative error ϵ_0 is given by

$$(8) \quad -\frac{1}{2}B^{1-t} \leq \epsilon_0 < \frac{1}{2}B^{1-t}$$

The result, using again $f(z) = B^t$ as in (1) and $g(m)$ as in (2), is

$$(9) \quad h_R(\epsilon_0) = \begin{cases} (B-1)B^{t-1}/\ln B & \text{if } |\epsilon_0| \leq \frac{1}{2}B^{-t} \\ (1/2|\epsilon_0| - B^{t-1})/\ln B & \text{if } \frac{1}{2}B^{-t} \leq |\epsilon_0| \leq \frac{1}{2}B^{1-t} \end{cases}$$

Figure 3.3 illustrates the density $h_R(\epsilon_0)$ for the base $B = 2$.

Figure 3.3

Error density function $h_R(\epsilon)$, Equation (9), for multiplication when symmetric rounding is used on a binary computer ($B = 2$). Adapted from [Tsao, 1974].

The mean and mean squared values for (6) and (9) are for chopping

$$(10) \quad E_C(\epsilon_o) = B^{-t}(B - 1)/2\ln B$$

$$(11) \quad E_C(\epsilon_o^2) = B^{-2t}(B^2 - 1)/6\ln B$$

and for symmetric rounding

$$(12) \quad E_R(\epsilon_o) = 0$$

$$(13) \quad E_R(\epsilon_o^2) = B^{-2t}(B^2 - 1)/24\ln B.$$

Division is considered to be similar to multiplication, but the addition error is not so simple to characterize. First, there is positive probability that an addition error is exactly zero. Second, to make a good assumption about the distribution of the mantissa of the result, we must make general assumptions about the distribution of numbers on the entire range of the computing device.

Hamming (1970) assumes that if we let $x + y = z$ where $x = m_x \cdot B^{E_x}$, $y = m_y \cdot B^{E_y}$, $z = m_z \cdot B^{E_z}$ and $m_x \in [1/B, 1)$, $m_y \in [u, m_x]$ for given

$u = \min_{s \in [0, \infty)} \{s: fl(m_x + s) \neq m_x\}$, both m_x and m_y have reciprocal

distributions. From this assumption, he proves that m_z has a reciprocal distribution over a very wide range. However, the assumption that, e.g., m_y has a reciprocal distribution over a wide range is not supported. It implies that all exponents in the wide range $[-t, 0]$ are equally likely which as Brent (1973) points out, holds only locally, and which empirical tests have shown is not true globally [Sweeney, 1965], [Kuck, Parker, and Sameh, 1977]. Nonetheless, we assume that the mantissa of the

result z has a reciprocal distribution and compare our result with empirical tests.

Let p_0 be the positive probability that the roundoff error ϵ_0 is zero. Kaneko and Liu (1973) found that p_0 ranges from approximately 0.3 to 0.8. By a simple application of the laws of conditional probability, we obtain the error distributions for addition,

$$(14) \quad h_C(\epsilon_0) = \begin{cases} p_0 \delta(\epsilon_0) + [(1 - p_0)(B - 1)/B^{1-t} \ln B] & \text{if } 0 \leq \epsilon_0 < B^{-t} \\ (1 - p_0)(1/\epsilon_0 - B^{t-1})/\ln B & \text{if } B^{-t} \leq \epsilon_0 < B^{1-t} \\ 0 & \text{otherwise} \end{cases}$$

for chopping, and

$$(15) \quad h_R(\epsilon_0) = \begin{cases} p_0 \delta(\epsilon_0) + [(1 - p_0)(B - 1)/B^{1-t} \ln B] & \text{if } |\epsilon_0| < B^{-t}/2 \\ (1 - p_0)(1^{1-t}/2|\epsilon_0| - B^{t-1}/\ln B) & \text{if } B^{-t}/2 \leq |\epsilon_0| < B^{1-t}/2 \\ 0 & \text{otherwise} \end{cases}$$

for symmetric rounding where $\delta(\cdot)$ is Dirac's delta function.

The mean and mean squared values are,

for chopping

$$(16) \quad E_C(\epsilon_0) = B^{-t}(1 - p_0)(B - 1)/2 \ln B$$

$$(17) \quad E_C(\epsilon_0^2) = B^{-2t}(1 - p_0)(B^2 - 1)/6 \ln B$$

and for symmetric rounding

$$(18) \quad E_R(\epsilon_0) = 0$$

$$(19) \quad E_R(\epsilon_0^2) = B^{-2t}(1 - p_0)(B^2 - 1)/24 \ln B.$$

Kaneko and Liu tested equations (6) and (14) on calculations that simulated a second order digital filter. They compared the chopped six digit hexadecimal arithmetic to arithmetic done on double precision

accumulators. The input sequence to the filter was taken from a zero mean normal distribution and the calculation of each output sample involved two multiplications and three additions. Local roundoff errors in these arithmetic operations were monitored. Figure 3.4 shows a typical histogram of the multiplication errors and the theoretical density function. The sample mean and variance are -2.5×10^{-6} and $(2.4 \times 10^{-6})^2$ respectively compared to theoretical values from equations (10) and (11) of -2.7×10^{-6} and $(2.8 \times 10^{-6})^2$. The agreement of the tests with the model seems reasonable.

Figure 3.4. Histogram of multiplication error.
Reproduced from Kaneko and Liu (1973).

The addition sample also showed good agreement. A histogram is plotted in Figure 3.5 along with the theoretical distribution. Two sets of data were examined, one with $p_0 = 0.62$, the other with $p_0 = 0.75$. The sample means for the two sets of data were -1.2×10^{-6} and -0.7×10^{-6} , and the sample variances were $(2.2 \times 10^{-6})^2$ and $(1.8 \times 10^{-6})^2$ respectively. The theoretical mean and variance for $p_0 = 0.62$ are -1.08×10^{-6} and

$(2.18 \times 10^{-6})^2$ respectively and for $p_0 = 0.75$, they are -0.68×10^{-6} and $(1.80 \times 10^{-6})^2$ respectively.

Figure 3.5. Histograms of addition error for $p_0 = 0.62$ and $p_0 = 0.75$. Reproduced from Kaneko and Liu (1973).

2. Guard Digits and Choice of Base

Symmetric rounding as considered by Kaneko and Liu (1973) is rarely implemented. Most units employ guard digits, and some bases B give better rounding results than others. A discrete consideration of roundoff error is necessary in order to understand the roundoff on real computers.

Kuki and Cody (1973) tested several kinds of rounding using a 22-digit, base two (binary) system and a six digit hexadecimal system. (These two systems represent almost the same range of numbers in a 32 bit word). Their tests consisted of repeated additions, multiplications,

and long inner products.

The operands were generated in double precision as follows: Let N be the size of a sample, and for $j = 1, 2, \dots, N$, let $\{X_{ij} : i = 1, 2, \dots, M\}$ be the j^{th} operand set for which we evaluate an arithmetic expression where M is the number of operands in the expression. X_{ij} is generated by $X_{ij} = R_j * S_{ij}$ where R_j are random numbers from the reciprocal distribution in the interval $(1/16, 16)$ and the S_{ij} are random numbers from the uniform distribution on $(-1, 1)$ for the case of signed operands and on $(0, 1)$ for strictly positive cases.

The scale factor R_j is varied from one operand set to another to neutralize the effect of the fluctuation in the number of significant digits in hexadecimal data. The need for this is illustrated by the example of an inner product. If $M = 50$ and we have $\sum X_i * Y_i$ where X_i and Y_i are uniform on $(0, 1)$ then the distribution of $X_i \times Y_i$ is $p(z) = -\ln z$, and the expectation of the result is $-\sum \int_0^1 z \ln z_i dz = 12.5$ with a standard deviation of $[\sum (-\int_0^1 z^2 \ln z dz - \bar{z}^2)]^{1/2} = 1.56$. This means for most cases the result in the six digit hexadecimal representation has a mantissa $m = (1 \text{ xxx, xxxx, } \dots)_{16}$ and hence a nominal significance of 24 digits as opposed to the 22-digit significance of the binary system. This favors the hexadecimal form. However, if $M = 50$ and the operands are selected from $(0, \sqrt{2})$ or if $M = 100$ and the operands are selected from $(0, 1)$, the expected result is 25, and most of the results in the hexadecimal form have a mantissa starting with three leading zeros, $m = (0001, \text{ xxxx, } \dots)_{16}$ and thus have a nominal significance of only 21 bits, passing the significance advantage to the binary form of representation.

The double precision sample operands are chopped or rounded to six hexadecimal digits or 22 binary digits before arithmetic is performed. Thus the rounding method refers to data preparation as well as operational

rounding. The arithmetic expressions are computed in the various arithmetic systems and in double precision. The negative of the base two logarithm of the difference between the final double precision result and the result from the various arithmetic systems is then computed. This result is called the binary significance. The binary significance for each set of computations is listed in Table 3.1, giving the minimum and the average value.

If s is the binary significance and r is the relative error, then $s = -\log_2 |r|$ and $r = \pm 2^{-s}$. If we assume s is normally distributed, and $E(s) = \mu$, $\text{Var}(s) = \sigma^2$, then since

$$E(e^{ts}) = \exp(\mu t + \sigma^2 t^2 / 2) \quad [\text{Hogg, Craig, 1970}],$$

it follows $E(|r|) = E(e^{-s \ln 2}) = 2^{-\mu} * 2^{\sigma^2 \ln 2 / 2}$ and thus,

$$-\log_2 (E(|r|)) = \mu - (\ln 2 / 2) \sigma^2. \quad \text{From Section I, Eq. (18) we know}$$

that if $S = \sum_{i=1}^n a_i$ and $\bar{S} = f(S)$, $\bar{S}_i \equiv (\bar{S}_{i-1} + a_i) \rho_i \equiv \hat{S}_i \rho_i$, then

(20)

$$(\bar{S} - S)/S = (\rho_n - 1)(a_n/S) + (\rho_n \rho_{n-1} - 1)(a_{n-1}/S) + \dots + (\rho_n \rho_{n-1} \dots \rho_1 - 1)(a_1/S)$$

where $\rho_i = 1 + \epsilon_i$, $\epsilon_1 = 0$.

For chopping on the 22 digit binary system, $\epsilon_i = (\bar{S}_i - \hat{S}_i)/\hat{S}_i$. Thus

$$E(\epsilon_i) = - \int_0^{2^{-22}} \hat{E}_i dP_{\hat{E}_i} \int_{\frac{1}{2}}^1 (1/\hat{S}_i) dP_{\hat{S}} \quad \text{where } \hat{E}_i \text{ has a uniform distribution}$$

and $\hat{S}$ has a reciprocal distribution. Then

$$E(\epsilon_i) = - 2^{-23} \int_{\frac{1}{2}}^1 d\hat{S}/\hat{S}^2 \ln 2 = -2^{-22.471}.$$

Denote this quantity simply by $E(\epsilon)$ since it is independent of i . We

Table 3.1
Reproduced from Kuki and Cody (1973)

(a) Binary Significance Figures for Summation Tests
500 sums of 1024 numbers each

Rounding Method	Guard Digits	Sample*	Binary Significance Six Digit Hexadecimal			Binary Significance 22 Digit Binary		
			Min.	Average	†	Min.	Average	†
Chop	≥ 0	P	12.85	13.71	± 0.69	13.34	13.47	± 0.13
Round	≥ 1	P	16.57	20.01	± 1.65	13.37	19.88	± 1.57
Chop	0	S	8.96	18.23	± 2.04	7.48	18.10	± 2.05
Chop	1	S	5.86	13.53	± 1.93	5.68	14.17	± 1.84
Chop	∞	S	5.85	13.52	± 1.94	4.80	13.23	± 1.83
Round	1	S	9.35	12.53	± 2.06	6.11	14.20	± 1.77
Round	2	S	10.26	18.08	± 2.00	6.69	15.16	± 1.88
Round	∞	S	10.26	18.09	± 2.01	8.81	17.78	± 2.10

(b) Binary Significance Figures for 500 products of
20 factors

Rounding Method	Guard Digits	Binary Significance Six Digit Hexadecimal			Binary Significance 22 Digit Binary		
		Min.	Average	†	Min.	Average	†
Chop	0	14.11	15.63	± 0.62	16.49	16.88	± 0.20
Chop	≥ 1	16.47	17.33	± 0.34	16.73	17.19	± 0.20
Round	≥ 2	18.73	21.36	± 1.66	19.11	21.51	± 1.55

* P = a sample where all operands are positive

S = a sample with signed operands

† Average figures followed by standard deviation

Table 3.1 continued

(c) Binary Significance Figures for Inner Product Tests
500 Inner Products with 50 operands

Rounding Method	Guard Digits	Sample*	Binary Significance Six Digit Hexadecimal			Binary Significance 22 Digit Binary		
			Min.	Average	†	Min.	Average	†
Chop	≥ 0	P	16.42	17.84	± 0.69	17.27	17.69	± 0.22
Chop	≥ 1	P	16.57	17.91	± 0.67	17.27	17.69	± 0.23
Round	≥ 2	P	18.53	22.26	± 2.18	19.59	22.12	± 1.76
Chop	0	S	8.56	19.11	± 2.13	12.52	19.78	± 1.65
Chop	1	S	10.94	18.01	± 1.80	12.33	18.36	± 1.49
Chop	∞	S	10.94	17.99	± 1.80	10.96	17.69	± 1.54
Round	1	S	13.19	21.01	± 2.09	13.40	18.92	± 1.76
Round	2	S	13.14	21.13	± 2.17	13.82	19.55	± 1.81
Round	∞	S	13.14	21.12	± 2.18	13.82	20.37	± 1.92

(d) Binary Significance Figure of Summation Tests
Comparison with R*-rounding

Rounding Method	Guard Digits	Number of operands††	Binary Significance Six Digit Hexadecimal			Binary Significance 22 Digit Binary		
			Min.	Average	†	Min.	Average	†
Chop	0	8	12.61	21.58	± 2.13	13.51	21.36	± 1.90
Round	∞	8	12.58	22.52	± 2.46	13.49	21.99	± 2.15
R*-round	∞	8	12.58	22.51	± 2.45	13.49	22.30	± 2.27
Chop	0	32	10.01	20.67	± 2.13	9.48	20.56	± 2.05
Round	∞	32	8.27	21.66	± 2.48	8.54	20.67	± 2.48
R*-round	∞	32	8.58	21.74	± 2.50	8.54	21.28	± 2.50
Chop	0	256	10.57	19.29	± 1.99	11.09	29.19	± 1.85
Round	∞	256	11.15	19.73	± 2.11	8.30	18.86	± 2.02
R*-round	∞	256	12.49	20.14	± 2.17	9.56	20.03	± 2.00
Chop	0	1024	8.96	18.23	± 2.04	7.48	18.10	± 2.05
Round	∞	1024	10.26	18.09	± 2.01	8.81	17.78	± 2.10
R*-round	∞	1024	11.31	19.03	± 2.08	9.69	19.07	± 2.16

†† 500 sums of 1024 numbers with the indicated number of signed operands.

have $E(o) = E(o_i) = 1 + E(e)$. Assuming that the o_i 's and a_i/S 's are independently distributed in equation (20) implies

$$E\left(\frac{\bar{S}-S}{S}\right) = \frac{E(o)-1}{n} + \frac{E(o)^2-1}{n} + \dots + \frac{E(o)^{n-1}}{n} + \frac{E(o)^{n-1}}{n} = \frac{n+1}{2} - \frac{1}{n} E(e) + \dots$$

Hence $E(|r|) = 2^{-13.470}$ for $n = 1024$ additions. Experimental results for this case yielded $E(|r|) = 2^{-13.46}$ and are recorded in Table 3.2.

Cody (1973) figured theoretical values for chopping and R^* -rounding for the bases two, four and sixteen (see Table 3.2). The base four system has a 23 digit, binary mantissa and a base four exponent. (One may verify that this system uses the same word size and has approximately the same number range as the binary and hexadecimal systems considered).

For all of the operations surveyed in Table 3.1, R^* -rounding does better than chopping. Symmetric rounding beats chopping for all but one case, the case of the sum of all signed operands. Kuki and Cody (1973) attribute this result to the biases of chopping cancelling one another whereas the bias of symmetric rounding is the same no matter what the sign of the operand is.

We also clearly see that whereas rounding with two guard digits gives better results than rounding with only one guard digit, we gain little in the use of more than two guard digits.

Base four gives the best results in all of the above cases examined. Base two and base sixteen are roughly equivalent. The difference in the ranges of the three systems is insignificant. If we consider the maximum relative rounding error (MRRE) which we found for symmetric rounding in

Table 3.2
Reproduced from Cody (1973)

Comparison of Analytical and Experiment Results

Experiment*	Mode	Guard Digits	Base	Statistic**	
				analytic	experimental
I	Chop	0	2	13.470	13.46
			4	13.885	-----
			16	13.563	13.55
	R*-round	≥ 1	2	19.943	19.88
			4	20.282	-----
			16	19.738	20.01
II	Chop	0	2	16.904	16.87
			4	16.953	-----
			16	15.624	15.50
		1	2	17.186	17.18
			4	17.601	-----
			16	17.279	17.28
	R*-round	∞	2	21.509	21.51
			4	21.848	-----
			16	21.305	21.36

* Experiment I is 500 sums of 1024 positive numbers.
Experiment II is 500 products of 20 factors each.

** The statistic for chopping is $-\log_2(E(|r|))$.
That for R*-rounding is $E(s)$.

Section I,

$$(21) \quad \text{MRRE}(t, B) = \frac{1}{2}B^{1-t},$$

we find that the base four and base two systems examined have MRRE of $.5 \times 2^{-21}$ whereas the six digit hexadecimal system has an MRRE of 2^{-21} . Another important parameter is $E_R(|\epsilon_o|)$ where ϵ_o has the density described by equation (6). This parameter is called the average relative rounding error (ARRE).

$$(22) \quad \text{ARRE}(t, B) = E_R(|\epsilon_o|) = B^{-t}(B - 1)/4 \ln B$$

If we rate the three bases by this measure, base four has an ARRE of $.14 \times 2^{-21}$ compared to $.17 \times 2^{-21}$ for base sixteen and $.18 \times 2^{-21}$ for base two. It would seem then that the best binary system would have base four exponents.

We cannot explain the effect of guard digits using the continuous distribution of Kaneko and Liu (1973) and of Tsao (1974). We must instead consider roundoff error in terms of the discrete distributions derived in Section II.

Let $x_i = m_i B^{E_i}$ for $i = 1, 2, 3$ denote three positive floating point numbers where $x_3 = x_1 * x_2$. Division is similar to multiplication. We do not as yet have discrete distributions for addition errors.

Define the quantities r_i ($i = 0, 1, 2, 3$) by

$$(23a) \quad \text{Actual error:} \quad r_0 = fl(x_1 x_2) - x_1 x_2$$

$$(23b) \quad \text{Standardized error:} \quad r_1 = r_0 / B^{E_1 + E_2}$$

$$(23c) \quad \text{Fraction error:} \quad r_2 = r_0 / B^{E_3}$$

$$(23d) \quad \text{Relative error:} \quad r_3 = r_0 / x_3 \quad \equiv \quad \epsilon_o$$

Example: $x_1 = \pi = 10^1 \times (.314\dots)$ $E_1 = 1, m_1 = \pi/10$
 $x_2 = e = 10^1 \times (.271\dots)$ $E_2 = 1, m_2 = e/10$
 $x_3 = \pi e = 10^1 \times (.853\dots)$ $E_3 = 1, m_3 = m_1 m_2 \times 10$

Use a 3-digit computer and chopping, then normalization (CHN) to obtain

$$\begin{aligned} r_0 &= \text{CHN}(.314)(.271) \times 10^{1+1} - \pi e = -.0397\dots \\ r_1 &= -.0397 \times 10^{-2} = -.397 \times 10^{-3} \\ r_2 &= -.0397 \times 10^{-1} = -.397 \times 10^{-2} \\ r_3 &= -.0397\dots/\pi e = -.00464\dots \end{aligned}$$

For brevity we introduce the following notational conventions. We abbreviate rounding by chopping, symmetric rounding with an even base, symmetric rounding with an odd base by the symbols CH, SE, and SO respectively. Let N denote "normalize" so that CHN means normalize after chopping (which we call Option A). NCH means normalize before chopping (Option B). NSE, SEN, NSO and SON have a similar meaning. The number of guard digits is written in parentheses; for example, NSE (≥ 1) means normalize before symmetric rounding with an even base and at least one guard digit. Let $u = B^{-t}$, the machine unit.

We now establish distributions for roundoff errors in floating point multiplication using the measure of standardized error, r_1 , and fraction error, r_2 . We begin by stating a sequence of lemmas by Bustoz, Feldstein and Goodman (1978).

Lemma 3.1 Let x be rounded (by CH, SE, or SO) to $\tilde{x}$, a number whose fractional part has M digits. Let each of the K trailing digits a_i of

x range over its B possible values, $0 \leq a_i \leq B - 1$, where
 $M + 1 \leq i \leq M + K$. Let

$$(24) \quad G = \sum_{i=1}^K a_{M+i} B^{K-i}$$

Define the integer q by

$$(25) \quad q = \begin{cases} B^K - G & \text{for rounding up} \\ -G & \text{for rounding down} \end{cases}$$

Then q takes on exactly once each of the B^K integer values in the appropriate part of (26) below, and conversely if q is an integer satisfying the appropriate part of (26), then q satisfies (25).

$$(26a) \quad -B^K + 1 \leq q \leq 0 \quad \text{for CH}$$

$$(26b) \quad -\frac{1}{2}B^K + 1 \leq q \leq \frac{1}{2}B^K \quad \text{for SE}$$

$$(26c) \quad -\frac{1}{2}B^K + \frac{1}{2} \leq q \leq \frac{1}{2}B^K - \frac{1}{2} \quad \text{for SO}$$

Proof: We verify the lemma for SE. When $G = 0, 1, 2, \dots, \frac{1}{2}B^K - 1$,

then $0 \leq GB^{-K-M} < \frac{1}{2}B^{-M}$ so that rounding down occurs, and (25) implies

that $q = 0, -1, -2, \dots, -\frac{1}{2}B^K + 1$. On the other hand, when $G = \frac{1}{2}B^K$,

$\frac{1}{2}B^K + 1, \dots, B^K - 1$, then $GB^{-K-M} \geq \frac{1}{2}B^{-M}$ and rounding up occurs and

(25) implies that $q = \frac{1}{2}B^K, \frac{1}{2}B^K - 1, \dots, 1$. This verifies (26b). The

verification of the converse and the other rounding rules is similar.

Lemma 3.2 Let x as in Lemma 3.1 be rounded to $\tilde{x}$ (by CH, SE, or SO),

a number whose mantissa has M digits. Let G and q be given by equations

(24) and (25). The K trailing digits of m where $x = mB^E$ are

$(a_{M+1}, a_{M+2}, \dots, a_{M+K})$ if and only if

$$(27) \quad r_2 = qB^{-K-M}$$

where r_2 is given by $(\tilde{X} - X)/B^E$ corresponding to (23c).

Proof: Refer to Figure 3.6. Note that the trailing digits are

$(a_{M+1}, a_{M+2}, \dots, a_{M+K})$ if and only if

$$(28) \quad r_2 = \begin{cases} B^{-M} - GB^{-K-M} & \text{for rounding up} \\ -GB^{-K-M} & \text{for rounding down} \end{cases}$$

Eq. (25) implies that the right side of (27) is qB^{-K-M} .

Consider the product $x_1 * x_2$ where $x_1 = m_1 B^{E_1}$ as before. We know that $m_1 m_2 \in [1/B^2, 1)$ and we partition this interval into two separate cases:

Case 0: $1/B \leq m_1 m_2 < 1$ (no normalization needed)

Case 1: $1/B^2 \leq m_1 m_2 < 1/B$ (1 digit left shift needed)

These two cases each occur with probability $\frac{1}{2}$ (if $p(m) = 1/m \ln B$).

Let $E^* = E_3 - (E_1 + E_2)$. Then by (23b, c),

$$(29) \quad r_1 = r_2 B^{E^*}$$

Suppose x_3 has an $M + K$ digit fraction, and for Case 0, $m_1 m_2$ has $2t$ digits and m_3 has t digits. Hence, $K = M = t$ and $B^{-K} = B^{-M} = B^{-t}$. The choice of K and M in Case 1 depends on the rounding process, the normalization process, and the number of guard digits. For Case 1, rounding with no guard digits is identical to chopping with no guard digits. Suppose there is at least one guard digit. When Option A is used, the result is the same as though normalization were done first,

FIGURE 3.6

TRAILING DIGITS AND $R_2(\cdot)$

Reproduced from Bustoz, Feldstein, and Goodman (1978)

ROUND UP

	Digits 1 to M			Digits M+1 to M+K		
m rounded =	...	1 + a_M		0		
m	a_1	...	a_M	a_{M+1}	...	a_{M+K}
r_2	0	...0	1	-G		

ROUND DOWN

	Digits 1 to M			Digits M+1 to M+K		
	...	a_M		0		
	a_1	...	a_M	a_{M+1}	...	a_{M+K}
	0	...0	0	-G		

then the $2t - 1$ digit result were rounded to the $t - 1$ digit result, and the final t^{th} digit were set to zero. Thus $K = t$ and $M = t - 1$ and $B^{-K} = u$ and $B^{-M} = Bu$, i.e., $K = t$ and $M = t - 1$. This also holds in the special situation when symmetric rounding produces a carry which changes the leading digit from zero to one. That can occur only when rounding is up and the normalized result has $m_3 = 1/B$. On the other hand, when normalization occurs first and Case 1 occurs, then a $2t - 1$ digit number is rounded to t digits. Thus $K = t - 1$ and $M = t$ and $B^{-K} = Bu$ and $B^{-M} = u$. Notice that $\text{NSE}(1)$ for Case 1 causes normalization to occur first with no guard digit remaining with which to round symmetrically. Hence, in essence, chopping occurs.

Table 3.3

Values of K and M if $u = B^{-t}$ is the machine unit
Reproduced from Bustoz, Feldstein, and Goodman (1978)

<u>Rounding Rule</u>	<u>Case 0</u>		<u>Case 1</u>	
	<u>K</u>	<u>M</u>	<u>K</u>	<u>M</u>
$\text{NCH}(0), \text{CHN}(\geq 0), \text{SEN}(\geq 1), \text{SON}(N)$	t	t	t	$t-1$
$\text{NCH}(\geq 1), \text{NSE}(1)^\dagger, \text{NSE}(\geq 2), \text{NSO}(N)$	t	t	$(t-1)^\dagger$	$t^\dagger$

† Chopping is done for $\text{NSE}(1)$ in Case 1, so CH formula must be used.

We partition the sample space of the random variables r_1 and r_2 into two subspaces corresponding to Cases 0 and 1. Define

$$(30) \quad \text{Pr}^{(j)}\{\text{Event}\} \equiv \text{Prob}\{\text{Event} | \text{Case } j \text{ occurs}\} \quad (j = 0, 1)$$

In particular

$$(31) \quad P^{(j)}(q) = \text{Pr}^{(j)}\{r_2 = qB^{-M-K}\}$$

For integers q satisfying (25), define

$$(32) \quad \delta_q = \begin{cases} -\frac{1}{2} & \text{for rounding up} \\ \frac{1}{2} & \text{for rounding down} \end{cases}$$

By (25) of Lemma 3.1 rounding up occurs if and only if $1 \leq q \leq \lfloor B^K/2 \rfloor$

where $\lfloor x \rfloor = (\text{greatest integer} \leq x)$. The δ_q may be written as

$$(33) \quad \delta_q = \begin{cases} -\frac{1}{2} & \text{if SE or SO is used and } 1 \leq q \leq \lfloor \frac{1}{2} B^K \rfloor \\ \frac{1}{2} & \text{otherwise} \end{cases}$$

Lemma 3.3 Let $x_i = m_i B^{E_i}$ for $i = 1, 2$ be two t digit normalized, floating point numbers selected at random from the reciprocal distribution.

Let $x_1 * x_2$ have the normalized floating point representation $x_3 = m_3 B^{E_3}$.

Then

$$(34) \quad P^{(j)}(q) = B^{-K} + B^{-M-K} (qB^{-K} + \delta_q)(B-1)/\ln B + O(u^2)$$

Proof: Since m_1 and m_2 are reciprocally distributed, then so is m_3 .

Refer to Theorem 2.4, apply Lemma 3.2 with $x = x_3$; then (26) and (30)

imply that

$$P^{(j)}(q) = P_{M+1}^{(j)}(G, B^K)$$

where $P_t^{(j)}(\cdot)$ is defined in Theorem 2.2. Use (33) to rewrite (25) as

$$q = -G + (\frac{1}{2} - \delta_q) B^K$$

Since $B^{-K} = O(u)$ and $B^{-M} = O(u)$, then $P_{M+1}^{(j)}(G, B^K) = B^{-K} + T_1 B^{-M-1} + O(u^2)$

where

$$(\ln B) T_1 = B(B-1) \{ \frac{1}{2} B^{-K} - G B^{-2K} \} + O(u^2)$$

Since $G = -q + (\frac{1}{2} - \delta_q) B^K$, we have

$$(\ln B) T_1 = B(B-1) B^{-K} (\delta_q + q B^{-K}) + O(u^2).$$

Hence,

$$\begin{aligned} P^{(j)}(q) &= P_{M+1}(G, B^K) = B^{-K} + B^{-M}[(B-1)B^{1-K}(\delta_q + qB^{-K}) + O(u^2)]/B \ln B \\ &\quad + O(u^2) \\ &= B^{-K} + B^{-M-K}(B-1)(qB^{-K} + \delta_q)/\ln B + O(u^2) \end{aligned}$$

Theorem 3.1

$$(35) \quad \Pr(r_2 = qB^{-M-K}) = \frac{1}{2}P^{(0)}(q) + \frac{1}{2}P^{(1)}(q)$$

where the $P^{(j)}(\cdot)$ are given by (34) and

$$(36) \quad \Pr(r_1 = qB^{E^*-M-K}) = \frac{1}{2}P^{(E^*)}(q)$$

Proof: Eq. (35) follows from the simple principle of conditional probability,

$$\Pr(r_2 = qB^{-M-K}) = P^{(0)}(q) \cdot \Pr(\text{Case 0 occurs}) + P^{(1)}(q) \cdot \Pr(\text{Case 1 occurs})$$

$$\begin{aligned} \Pr(r_1 = qB^{-M-K+E^*}) &= \Pr(r_2 = qB^{-M-K} \text{ and case } E^* \text{ occurs}) \\ &= P^{(E^*)}(q) \cdot \Pr(\text{case } E^* \text{ occurs}) = \frac{1}{2}P^{(E^*)}(q) \quad (E^* = E_3 - E_1 - E_2) \end{aligned}$$

Table 3.4 lists the means and variances for r_1 and r_2 with the various rounding options. The proofs for these parameters have been omitted. A proof of the means is in Bustoz, Feldstein, and Goodman (1978), a proof of the variances is in Goodman and Feldstein (1977).

There are three important points which stand out in Table 3.4. Normalization after rounding produces less bias and a smaller variance for the case of symmetric rounding with an even base and one guard digit. However, Option A has other disadvantages. For instance, we cannot multiply by one and be guaranteed that we will get the same

Table 3.4

Means and Variances of r_1 and r_2 for various rounding methods

Reproduced from Bustoz, Feldstein, and Goodman (1978)

<u>Rounding Method</u>	<u>1</u>	<u>$E(r_1) + 0(u^3)$</u>	<u>$Var(r_1) + 0(u^3 \text{ or } u^4)$</u>
NCH (≥ 1)	1	$-u(1+B^{-1})/4 + u^2\{12+(B-B^{-1})/\ell n B\}/24$	$u^2(5-6B^{-1}+5B^{-2})/48$
	2	$-u/2 + u^2\{3+3B+(B-1)/\ell n B\}/12$	$u^2/12$
CHN (≥ 0), NCH (0)	1	$-u/2 + u^2\{12+(B^2-1)/\ell n B\}/24$	$u^2/12$
	2	$-u(1+B)/4 + u^2\{6+6B+(B-1)(B^2+1)/\ell n B\}/24$	$u^2(5B^2-6B+5)/48$
NSE (1)	1	$-u/4B + u^2\{24-(B-1)(B-2)/B\ell n B\}/48$	$u^2(2+5B^{-2})/48$
	2	$-u/4 + u^2\{12+12B+(B-1)/\ell n B\}/48$	$7u^2/48$
NSE (≥ 2)	1	$u^2\{24-(B-B^{-1})/\ell n B\}/48$	$u^2(1+B^{-2})/24$
	2	$u^2\{6+6B-(B-1)/\ell n B\}/24$	$u^2/12$
SEN (≥ 1)	1	$u^2\{24-(B^2-1)/\ell n B\}/48$	$u^2/12$
	2	$u^2\{12+12B-(B-1)(B^2+1)/\ell n B\}/48$	$u^2(1+B^2)/24$
NSO (≥ 1)	1	$-u^2(B-B^{-1})/(48\ell n B)$	$u^2(1+B^{-2})/24$
	2	$-u^2(B-1)/(24\ell n B)$	$u^2/12$
SON (≥ 1)	1	$-u^2(B^2-1)/(48\ell n B)$	$u^2/12$
	2	$-u^2(B-1)(B^2+1)/(48\ell n B)$	$u^2(B^2+1)/24$

number back. Two guard digits in symmetric even rounding with Option B perform dramatically better than one guard digit which confirms the results obtained by Kuki and Cody (1973). Symmetric rounding does consistently better with odd base than with even base.*

3. ROM Rounding

ROM rounding takes the last ℓ significant digits of the significant part of the result mantissa and the first digit to be rounded as the $\ell + 1$ digit input to a ROM. The output of the ROM is the ℓ digit rounded result except when all ℓ low digits of the significant part of the mantissa are $(B - 1, B - 1, \dots, B - 1)$ in which case chopping is performed.

Kuck, Parker, and Sameh (1977) compared the average roundoff errors for addition and multiplication and the average bias of the rounding methods chopping, symmetric rounding, jamming, R^* -rounding, and ROM rounding.

If we define average bias for base two by

$$\text{Bias}(t, s) = \sum_{x \in M(t+s)} (\ell(x) - x) / \|M(t+s)\|$$

where $M(t)$ is the set of all t digit-base-two-mantissas and s is the number of guard digits. $\ell(x)$ as before rounds x by the appropriate rule to t digits and $\|X\|$ denotes the number of elements in the set X . Biases are listed for various rounding methods in Table 3.5.

*Note however the difficulty in hardware implementation.

Table 3.5
Bias of Rounding Methods for Base B, a Power of Two.
Reproduced from Kuck, Parker, and Sameh (1977)

Rounding Scheme	Average Bias	
Chopping	$\frac{1}{2}2^{-t}(2^{-s}-1)/\log_2 B$	$s \geq 0$
Jamming	$\frac{1}{2}2^{-t}(2^{-s})/\log_2 B$	$s \geq 0$
Rounding	$\frac{1}{2}2^{-t}(2^{-s})/\log_2 B$	$s > 0$
R*	0	$s \geq 0$
ROM(ℓ)*	$\frac{1}{2}2^{-t}(2^{-s}-2^{1-\ell})/\log_2 B$	$s > 0$

* ROM(ℓ) means an ROM which uses digits $t-\ell+1, t-\ell+2, \dots, t, t+1$ as input to the ROM.

The average roundoff errors for multiplication in each of these schemes is listed in Table 3.6.

Table 3.6
Multiplication Roundoff Error Performance
Reproduced from Kuck, Parker, and Sameh (1977)
The base B is a power of two.

Rank	Scheme	Average Roundoff
1	R*	0
2	Rounding, jamming	$B^{-2t}((B-1)/\log_2 B)^2/2\ln 2$
3	ROM(ℓ)	$-B^{-t}2^{-\ell}(B-1)/\ln B$
4	Chopping	$-B^{-t}(B-1)/2\ln B$

Addition roundoff error depends greatly on the problem. There is positive probability in floating point that addition will give exact results. Let P_0 be the probability that the addition is not zero; let $P_s = \Pr(x + y \in M(t + s))$ where x and y are floating point numbers and let $P_\tau = \sum_{s=1}^{\infty} B^{-s} P_s$. Then the average roundoff errors for addition are listed in Table 3.7.

Table 3.7
Addition Roundoff Error Performance
Reproduced from Kuck, Parker, and Sameh (1977)
The base B is a power of two.

Rank	Scheme	Average Roundoff
1	R^*	0
2	ROM(ℓ)	$B^{-1}(B-1)((B-1)P_\tau/2\log_2 B - 2^{-\ell}P_s)/\ln B$
3	Rounding	$(B^{-t}/2\ln 2)((B-1)/\log_2 B)^2 P_\tau$
4	Jamming	$B^{-t}((B-1)/\log_2 B)P_\tau/2\ln 2$
5	Chopping	$B^{-t}(B-1)((B-1)P_\tau/2\log_2 B - P_s/2)/\ln B$

4. Non-Symmetric Rounding Intervals [Grau, Bareiss, 1977]

Throughout our discussion of floating point arithmetic we have assumed that standard rounding considers an interval of the form

$[(m - \frac{1}{2}B^{-t})B^E, (m + \frac{1}{2}B^{-t})B^E]$ as the rounding interval associated with mB^E . However, Grau and Bareiss (1977) point out that mantissas in the

interval $[(1/B - \frac{1}{2}B^{-t-1})B^E, (1/B + \frac{1}{2}B^{-t})B^E]$ are rounded to $1/B \cdot B^E$.

Notice that the parts of the interval to the right of $1/B$ is B times as large as the interval to the left of $1/B$. Since we consider mantissas as being in $[1/B, 1)$ and not $(1/B, 1]$, the assumptions made in this section assume this interval is $[1/B - \frac{1}{2}B^{-t}, 1/B + \frac{1}{2}B^{-t})$ which is larger than the above asymmetric interval. Thus with respect to these non-symmetric rounding intervals, the distribution we derive are conservative. In other words, the cumulative distribution over all possible errors is > 1 . This indicates that the available results need refinement.

REFERENCES

Brent, Richard P., "On the Precision Attainable with Various Floating Point Number Systems," IEEE Transactions on Computers, vol. C-22, (June 1973), 601-7.

Bustoz, Joaquin, Alan Feldstein, and Richard Goodman, "Improved Trailing Digit Estimates Applied to Optimal Computer Arithmetic," to appear, JACM, 1978.

Cody, William J., Jr., "Static and Dynamic Numerical Characteristics of Floating-Point Arithmetic," IEEE Transaction on Computers, vol. C-22, (June 1973), 598-601.

Goodman, Richard and Alan Feldstein, "Effect of Guard Digits and Normalization Options on Floating Point Multiplication," Computing, vol. 18 (1977), 93-106.

Goodman, Richard and Alan Feldstein, "Roundoff Error in Products," Computing, vol. 15 (1975), 263-73.

Grau, Albert A. and Erwin H. Bareiss, "Statistical Aspects of Machine Rounding," ERDA Report COO-2280-34, Northwestern University, 1977.

Hamming, R. W., "On the Distribution of Numbers," Bell System Technical Journal, vol. 49, no. 8 (1970), 1609-25.

Hogg, Robert V. and Allen T. Craig, Introduction to Mathematical Statistics, Wiley, New York, 1970.

Kaneko, T. and B. Liu, "On the Local Roundoff Error in Floating Point Arithmetic," JACM, vol. 20 (July 1973), 391-8.

Kuck, D. J., D. S. Parker, and A. H. Sameh, "Analysis of Rounding Methods in Floating Point Arithmetic," IEEE Transactions on Computers, vol. C-26, no. 7 (July 1977), 643-50.

Kuki, H. and W. J. Cody, Jr., "A Statistical Study of the Accuracy of Floating Point Number Systems," CACM, vol. 16 (April 1973), 223-30.

Sterbenz, Pat H., Floating Point Computation, Prentice-Hall, Englewood Cliffs, N.J., 1974.

Sweeney, D., "An Analysis of Floating Point Addition," IBM System Journal, vol. 4 (1965), 31-42.

Tsao, N., "On the Distribution of Significant Digits and Roundoff Errors," CACM, vol. 17 (May 1974), 269-71.

IV. FIXED POINT ARITHMETIC

In our analysis of fixed point arithmetic we need only consider the operations of multiplication and division. Addition and subtraction are exact operations. We consider division to be similar to multiplication. Hence our discussion considers only multiplication.

A statistical analysis of fixed point multiplication is discussed in two works on numerical analysis ([Von Neumann, Goldstine, 1947],[Henrici, 1962]). Von Neumann and Goldstine assume that if we consider $fi(xy) = xy + \epsilon$, then ϵ has a uniform distribution on $[\frac{1}{2}B^{-t}, \frac{1}{2}B^{-t}]$. Hence if $u = B^{-t}$ is the machine unit then $E(\epsilon) = 0$; $Var(\epsilon) = u^2/12$.

Henrici, in the study of ordinary differential equations, considered the distribution of ϵ when $x > 0$ is constant and $y \geq 0$ is a uniform random variable on $[0,1]$. Henrici's (1962) distribution is very close to the uniform distribution. The distribution Henrici derives for $B=2$ has a slight negative bias. Timlake (1962), following a suggestion from Henrici, derived a distribution for the case when both x and y are random variables.

Urabe (1968) studied the problem for base two when both x and y are random variables. He obtained a distribution which is slightly different from Timlake's because Timlake required that the first factor be positive.

Goodman (1976) extended the distributions of Timlake and Urabe to a general base B and with Feldstein (1975) generalized Henrici's distribution to a general base. However, they considered only the case where the last non-zero digit of the first factor is relatively prime to B . The results are summarized in Tables 4.1 and 4.2. We derive the distributions for the unrestricted cases, which contain the Goodman and Feldstein results as special cases.

Table 4.1
Roundoff error for multiplication. Base B prime.
First factor must be positive. Means and variances.*
Reproduced from Goodman (1976)

Rounding Method	Mean	Variance
Chopping	$-u/2 + u^2(1-B^{-1})t(1-u)/2$	$u^2 + u^3/6B$
Symmetric Rounding Base two	$u^2 t(1-u)^{-1}/4$	$u^2/12 + u^3/12 - t^2 u^4(1-u)^{-2}/16$
Symmetric Rounding B odd	0	$u^2(1-uB^{-1})/12$

Table 4.2
Roundoff error in multiplication
Means and variance. Reproduced from Goodman (1976)
Both factors may be zero. Base B prime.*

Rounding Method	Mean	Variance
Chopping	$(-u+u^2(1-B^{-1})t)(1-u)^{-1}/2$	$u^2/12 + u^3/6 + u^3(1-u)/6B$
Symmetric Rounding Base two	$\frac{1}{2}u^2 t$	$(u^2 - u^4)/12 - u^4 t^2/16$
Symmetric Rounding B odd	0	$u^2/12 - u^3/12 - u^3(1-u)/12B$

* B is base of number system; t is number of digits per word; $u = B^{-t}$.

1. The Distribution of a Constant times a discrete Random Variable

The calculation of $x \cdot y$ where x is a positive constant occurs frequently in the numerical solution of ordinary differential equations. Therefore, we consider the behavior of the error from this calculation separately from the case where both x and y are random variables.

Theorem 4.1 Let $x = x_1 B^{-1} + x_2 B^{-2} + \dots + x_k B^{-k}$ be a constant where each x_i is an integer in $[0, B-1]$ and at least the last k digits of y , i.e. $y_t, y_{t-1}, \dots, y_{t-l+1}$, ($t \geq l \geq k$), are drawn from a uniform random sample. Then, if g is the greatest common divisor of $G = x \cdot B^k$ and B^k , i.e. $g = (G, B^k)$, then the error $\epsilon = xy - fi(xy)$ in the fixed point product xy , rounded to t digits, takes exactly B^k/g distinct values with equal probability g/B^k . These values are, when $i = 0, 1, 2, \dots, B^k/g - 1$,

$$(1) \quad \epsilon = \begin{cases} igB^{-k-t} & \text{for chopping} \\ -\frac{1}{2}B^{-t} + (i)gB^{-k-t} & \text{for symmetric rounding, } B \text{ even} \\ -\frac{1}{2}B^{-t} + (i + \frac{1}{2})gB^{-k-t} & \text{for symmetric rounding, } B \text{ odd.} \end{cases}$$

To prove this theorem, we transform the problem to an equivalent problem over the integral domain of real numbers. The proof makes use of well known results in number theory which we state as

Lemma 4.1 Let a, b, q ($0 \leq q \leq b$) be given, and let $g = (a, b)$ be the greatest common divisor of a and b . Then the congruence

$$(2) \quad ay \equiv q \pmod{b}$$

has (a) exactly g principal roots $y_{q,j}$ ($j = 0, 1, \dots, g-1$; $0 \leq y_{q,j} < b$) if g divides q ; and (b) has no solution if g is not a factor of q . Hence a solution exists only if $q = ig$ ($i = 0, 1, 2, \dots, b/g - 1$).

If a solution exists, and $y_{q,0}$ is the smallest one, then all solutions are given by

$$(3) \quad y_{q,j} = y_{q,0} + j b/g \quad (j = 0, 1, \dots, g-1)$$

Conversely, instead of q , let the b distinct numbers $y_k = k$ ($k = 0, 1, \dots, b-1$) be given. Then the congruence (2) admits exactly b/g distinct residues q_k ($0 \leq q_k < b$). Hence

$$(4) \quad ak \equiv q_k \pmod{b} \Rightarrow a(k + \ell b/g) \equiv q_k \pmod{b}$$

where the g different integers, ℓ satisfy

$$-kg/b \leq \ell < (b - k)g/b$$

Proof of Theorem 4.1: Let $G = B^k x$, $H = B^t y$. Then

$$(5) \quad GH = B^k q_1 + q_2 \Rightarrow GH \equiv q_2 \pmod{B^k}$$

where

$$q_1 = \sum_{i=1}^t s_i B_i^{t-i}, \quad q_2 = \sum_{i=1}^k s_{t+i} B^{k-i} \quad (0 \leq s_j < B)$$

For given y , the number q_2 is uniquely determined and related to the round-off error. We must determine which values q_2 can assume, and then determine their distribution. Now,

$$(6) \quad H = B^k h_1 + h_2 \Rightarrow H \equiv h_2 \pmod{B^k}$$

where

$$h_1 = \sum_{i=1}^{t-k} y_i B^{t-k-i}, \quad h_2 = \sum_{i=1}^k y_{t-k+i} B^{k-i} < B^k$$

Hence

$$GH = B^k Gh_1 + Gh_2 = B^k q_1 + q_2$$

which implies

$$(7) \quad GH \equiv Gh_2 \equiv q_2 \pmod{B^k}$$

By assumption, $g = (G, B^k)$. For every H , there exists an h_2 by (6) and a q_2 by (5). By Lemma 4.1,

$$(8) \quad q_2 \in \{ig : i=0, \dots, B^k/g - 1\}$$

and therefore q_2 can assume only one value from the set given in (8).

But conversely, for any given $q_2 = ig$ there exist g different solutions $h_{2,\ell}$ ($\ell = 0, 1, \dots, g-1$; $0 \leq h_{2,\ell} < B^k$), which are related to $h_2 = h_{2,0}$ by

$$h_{2,\ell} = h_2 + \ell B^k/g \quad (-h_2 g/B_k \leq \ell < g - h_2 g/B^k)$$

The B^k/g sets of congruent solutions $\{h_{2,\ell} | q_2\}$ are mutually disjoint,

$$\{h_{2,\ell} | ig\} \cap \{h_{2,\ell} | jg\} = \emptyset, \quad (\forall i \neq j)$$

The cardinality of the union of these sets

$$\text{Card} \bigcup_{i=0}^{B^k/g-1} \{h_{2,\ell} | ig\} = B^k$$

Since each k -tuple corresponding to h_2 , $(y_{t-k+1}, \dots, y_t)$ occurs with equal probability, each corresponding q_2 , represented by the k -tuple $(s_{t+1}, \dots, s_{t+k})$ occurs either with probability 0 or probability gB^{-k} .

This reasoning extends to $H = B^t y$.

Finally, we divide q_2 by B^{k+t} and round to t digits. This yields the

different possibilities for the remainder

$$(9) \quad \epsilon = xy - fi(xy)$$

listed in (1) of the theorem.

It is of interest to note that only the distribution of the trailing k digits of y determine the distribution of the error ϵ , while the distribution of the $t-k$ leading digits of y is immaterial.

Example. Let $x = .3750$, $\{y\} = \{10^{-6} \leq y < 1\}$, $B = 10$. Hence $k = 3$ (since $x_4 = 0$) and $t = 6$. The greatest common divisor is

$$g = (G, B^k) = (375, 1000) = 125$$

Therefore, the rounding error ϵ of xy can assume only

$$B^k/g = 10^3/125 = 8$$

different values. These values are obtained from 8 mutually incongruent numbers modulo 10^3 , $y_k \times 10^6 \in [1, \dots, 10^6], (k=0, 1, \dots, 7)$. For example, choose $y_k \times 10^6 = k$. Then, xy_k yields for $k=0, 1, \dots, 7$ respectively

$$xy_0 = 0 \times 10^{-6}$$

$$xy_1 = .375 \times 10^{-6}$$

$$xy_2 = .750 \times 10^{-6}$$

$$xy_3 = 1.125 \times 10^{-6}$$

$$xy_4 = 1.500 \times 10^{-6}$$

$$xy_5 = 1.875 \times 10^{-6}$$

$$xy_6 = 2.250 \times 10^{-6}$$

$$xy_7 = 2.625 \times 10^{-6}$$

Note that $xy_8 = 3.000 \times 10^{-6}$ starts a new cycle. According to (1), the

possible values of the rounding errors for $t=6$ decimal place calculations are for chopping

$$\{\epsilon\} = \{igB^{-k-t}\} = \{125 \times 10^{-9} \times i : i = 0, 1, \dots, 7\}$$

and for symmetric rounding, with $B=10$ even,

$$\begin{aligned} \{\epsilon\} &= \{-\frac{1}{2}B^{-t} + igB^{-k-t}\} = \{[-500 + 125i] \times 10^{-9}\} \\ &= \{-.500, -.375, -.250, -.125, 0, .125, .250, .375 \text{ in units of } 10^{-6}\} \end{aligned}$$

This example illustrates also that the restriction $x_k \neq 0$ in Theorem 4.1 is theoretically irrelevant. The same results are obtained, using $x = .3750$ and $k=4$ instead of $k=3$. From a practical point of view, a small k is more desirable, since the distribution of the k trailing digits of y has been shown to be much closer to a uniform distribution if $k \ll t$ than if $k \sim t$.

Theorem 4.2 For the distribution of ϵ in Theorem 4.1 the expectation is

$$(10) \quad E(\epsilon) = \begin{cases} \frac{1}{2}(1 - g/B^k)B^{-t} & \text{for chopping} \\ -\frac{1}{2}gB^{-k-t} & \text{for symmetric rounding, } B \text{ even} \\ 0 & \text{for symmetric rounding, } B \text{ odd} \end{cases}$$

and the variance is for chopping

$$(11) \quad \text{Var}(\epsilon) = (1 - g^2/B^{2k})/12B^{2t}$$

Proof: For chopping, by (1)

$$E(\epsilon) = \sum_{(e)} \epsilon \text{ Prob}(\epsilon) = \sum_{i=0}^{Bk/g-1} igB^{-k-t} gB^{-k}$$

which yields the desired result immediately.

For symmetric rounding, B odd, all ϵ values are symmetric and equally probable. Hence $E(\epsilon) = 0$.

For symmetric rounding, B even, by (1)

$$E(\epsilon) = \sum_{i=0}^{Bk/g-1} (-\frac{1}{2} + igB^{-k}) B^{-t} gB^{-k} = \frac{1}{2} gB^{-t-k}$$

Because the three expressions for ϵ in (1) differ only from shifting the corresponding variances are close. We calculate for chopping, using (1) and (10),

$$\text{Var}(\epsilon) = E(\epsilon^2) - E^2(\epsilon)$$

$$(12) \quad E(\epsilon^2) = \sum_{i=0}^{Bk/g-1} i^2 g^2 B^{-2k-2t} g/B^k = (1 - g/B^k)(2 - g/B^k)/6B^{2t}$$

$$E^2(\epsilon) = (1 - g/B^k)^2/4B^{2t}$$

The difference between the last two expressions yields (11).

2. Fixed Point Multiplication of Two Random Numbers.

We consider the two cases where (a) $x \in (0,1)$ and (b) $x \in [0,1)$. In both cases x and $y \in [0,1)$ are random variables. Let

$$(13) \quad \begin{aligned} x &= x_1 B^{-1} + x_2 B^{-2} + \dots + x_t B^{-t} & (0 \leq x_i \leq B-1) \\ y &= y_1 B^{-1} + y_2 B^{-2} + \dots + y_t B^{-t} & (0 \leq y_i \leq B-1) \end{aligned}$$

be fixed point numbers which assume all possible values with equal probability in their respective domains.

Definition 4.1 Let $\phi(k)$ be the cardinality of the set of integers less than k which are relatively prime to k , including one.

Examples: $\phi(10) = \text{Card} \{1,3,7,9\} = 4$

$\phi(5) = \text{Card} \{1,2,3,4\} = 4$

$\phi(2) = 1, \phi(1) = 1$

The function $\phi(k)$ is called the totient function or Euler's ϕ -function.

Theorem 4.3 Let $x \in (0,1)$ and $y \in [0,1)$ be random t -digit fixed point numbers of the form (13), and define the rounding error by

$$\epsilon = xy - fi(xy)$$

For chopping, $\epsilon \in \{vB^{-2t}, v=0,1,\dots, B^t-1\}$. Let

$$(14) \quad p_v = \text{Prob}(\epsilon = vB^{-2t}) \quad v=0,1,\dots, B^t-1$$

and

$$(15) \quad S_v = \{g: g|(v, B^t), 1 \leq g < B^t\}$$

Then

$$(16) \quad p_v = \frac{1}{B^t(B^t-1)} \sum_{g \in S_v} g \phi(B^t/g)$$

The expected error of ϵ is

$$(17) \quad E(\epsilon) = \frac{1}{2B^t} \left[1 - \frac{C}{B^t(B^t-1)} \right]$$

and the variance of ϵ is

$$(18) \quad \text{Var}(\epsilon) = \frac{1}{12B^{2t}} \left[1 - \frac{1}{B^{2t}(B^t-1)^2} (3C^2 - 2(B^t-1)D) \right]$$

where

$$C = \sum_{G=1}^{B^t-1} (G, B^t), \quad D = \sum_{G=1}^{B^t-1} (G, B^t)^2$$

For symmetric rounding, see Corollary 4.1 below.

Proof: First we prove the expression (16) for p_v . Let $G = B^t x$ and note that $0 < G < B^t$. If we partition the event,

$$(19) \quad p_v = \sum_{x=B^{-t}}^{1-B^{-t}} p_{v|x} \cdot \text{Prob}(x) = \frac{1}{B^t-1} \sum_{G=1}^{B^t-1} p_{v|x}$$

If $g = (G, B^t)$ denotes the greatest common divisor of G and B^t , and $g|v$ indicates that g divides v , then by Theorem 4.1,

$$(20) \quad p_{v|x} = \begin{cases} 0 & \text{if } g|v \text{ false} \\ g/B^t & \text{if } g|v \text{ true.} \end{cases}$$

Next we determine the cardinality of the set

$$\{G\} = \{G: G B^t y \equiv v \pmod{B^t}\}$$

This set determines then

$$(21) \quad \{x\}_v = \{x: p_{v|x} \neq 0\}$$

which has the same cardinality as $\{G\}$. We observe that

$$g|v, \quad g|B^t = g|(v, B^t) \quad (1 \leq g < B^t)$$

Hence

$$\begin{aligned} \{G\} &= \{G: (G/g, B^t/g) = 1; \quad g|(v, B^t)\} \\ &= \bigcup_{g \in S_v} \{G: (G/g, B^t/g) = 1\} \end{aligned}$$

where S_v is given by (15). Since $0 < G < B^t$, it follows that the cardinality of $\{G\}$ is

$$(22) \quad \text{Card}\{G\} = \sum_{g \in S_v} \phi(B^t/g) = \text{Card}\{x\}_v$$

Since each $p_{v|x}$ with $x \in \{x\}_v$ has the same probability, g/B^t , (16) follows by substitution of (20), (21), and (22) into (19).

Example*: Let $B=10$, $t=1$. By direct calculation it is easily verified that for

$$\begin{aligned} v &= 0, \quad 1, \quad 2, \quad 3, \quad 4, \quad 5, \quad 6, \quad 7, \quad 8, \quad 9 \\ p_v &= \frac{17}{90}, \quad \frac{4}{90}, \quad \frac{12}{90}, \quad \frac{4}{90}, \quad \frac{12}{90}, \quad \frac{9}{90}, \quad \frac{12}{90}, \quad \frac{4}{90}, \quad \frac{12}{90}, \quad \frac{4}{90} \end{aligned}$$

respectively. Using (16) we obtain for

$$\begin{aligned} \underline{v=0}: \quad (v, B^t) &= (0, 10) = 10 \\ S_0 &= \{g: g|10\} = \{1, 2, 5\} \\ \phi(10) &= 4, \quad \phi(10/2) = 4, \quad \phi(10/5) = 1 \\ p_0 &= \frac{1}{(10)(9)} [(1)(4) + (2)(4) + (5)(1)] = 17/90 \end{aligned}$$

*For the example $B = 10$, $t = 2$, see Appendix.

$$\underline{v = 4}: (4,10) = 2; \quad S_4 = \{1,2\}, \quad \bar{\Phi}(10) = 4, \quad \bar{\Phi}(5) = 4,$$

$$p_4 = [(1)(4) + (2)(4)]/90 = 12/90$$

$$\underline{v = 5}: (5,10) = 5, \quad S_5 = \{1,5\}, \quad \bar{\Phi}(10) = 4, \quad \bar{\Phi}(2) = 1$$

$$p_5 = [(1)(4) + (5)(1)]/90 = 9/90$$

$$\underline{v = 7}: (7,10) = 1, \quad S_7 = \{1\}, \quad \bar{\Phi}(10) = 4$$

$$p_7 = (1)(4)/90 = 4/90$$

To prove (17), we partition the mean.

$$E(\epsilon) = \sum_{G=1}^{B^t-1} E(\epsilon|x)/(B^t-1)$$

where $E(\epsilon|x)$ is given by (10) for chopping with $g = (G, B^t)$, and $k=t$. Hence

$$E(\epsilon) = \frac{1}{2B^t(B^t-1)} \sum_{G=1}^{B^t-1} [1 - (G, B^t)/B^t]$$

Example: Let $B = 10$, $t = 1$ as in the previous example. Using (17), we obtain for $C = 1+2+1+2+5+2+1+2+1 = 17$, and

$$E(\epsilon) = (1 - 17/90)/20 = .0405555$$

To derive (18), we first determine

$$E(\epsilon^2) = \sum_{v=0}^{B^t-1} \frac{v^2}{B^{2t}} \sum_{G=1}^{B^t-1} p_{v|x} \text{Prob}(x) = \frac{1}{B^t-1} \sum_{G=1}^{B^t-1} E(\epsilon^2|x)$$

Recalling (12),

$$E(\epsilon^2|x) = (1 - (G, B^t)/B^t)(2 - (G, B^t)/B^t)/6B^{2t}$$

since $g = (G, B^t)$ and $k = t$. Thus

$$E(\epsilon^2) = \left[2 - \frac{3}{B^t(B^t-1)} \sum_{G=1}^{B^t-1} (G, B^t) + \frac{1}{B^{2t}(B^t-1)} \sum_{G=1}^{B^t-1} (G, B^t)^2 \right] / 6B^{2t}$$

and (18) is obtained from $\text{Var}(\epsilon) = E(\epsilon^2) - E^2(\epsilon)$.

Example: Let $B = 10$, $t = 1$ as in the two previous examples. Using (18), $C = 17$, $D = 1^2 + 2^2 + 1^2 + 2^2 + 5^2 + 2^2 + 1^2 + 2^2 + 1^2 = 45$ yields

$$\text{Var}(\epsilon) [1 - (3 \times 17 - 18 \times 45) / 8100] / 1200 = 8.275 \times 10^{-4}$$

If B is a prime number, the formulas for the means and variances are given in Table 4.1 and somewhat simpler.

Corollary 4.1 Using the assumptions of Theorem 4.3, but replacing chopping by symmetric rounding, the rounding errors $\epsilon = xy - fi(xy)$ take the values $r(\mu)$, $\mu = 0, 1, \dots, B^t - 1$, where

$$(23) \quad r(\mu) = \begin{cases} -\frac{1}{2}B^{-t} + \mu B^{-2t} & (B \text{ even}) \\ -\frac{1}{2}B^{-t} + (\mu + \frac{1}{2})B^{-2t} & (B \text{ odd}) \end{cases}$$

The probabilities that ϵ assumes a given value $r(\mu)$ are related to p_v of (16) for chopping, by

$$(24) \quad \text{Prob}(\epsilon = r(\mu)) = p_v$$

where

$$(25) \quad v = \begin{cases} (\mu + \frac{1}{2}B^t) \bmod B^t & (B \text{ even}) \\ (\mu + \frac{1}{2}(1 + B^t)) \bmod B^t & (B \text{ odd}) \end{cases}$$

The corresponding mean values are for symmetric rounding,

$$(26) \quad E(\epsilon) = \begin{cases} -\frac{1}{2}B^{-t} P_{\frac{1}{2}Bt} & (B \text{ even}) \\ 0 & (B \text{ odd}) \end{cases}$$

Proof: We observe from (23) that if B is odd, $r(\mu) = -r(B^t - 1 - \mu)$.

The two associated probabilities are equal, except for $\mu = \frac{1}{2}(B^t - 1)$, in which case $r(\mu) = 0$. Hence $E(\epsilon) = 0$.

If B is even, then (23) shows that $r(\mu) = -r(B^t - \mu)$ for $\mu \neq 0$, and for $\mu = B^t/2$, $r(\mu) = 0$. Hence, in calculating $E(\epsilon)$, only the term $r(0) \text{Prob}(\epsilon = r(0)) = -\frac{1}{2}B^{-t} P_{Bt/2}$ remains. Note that the symmetric relationship $p_v = p_{Bt-v}$ ($v \neq 0$) follows from the uniform distribution of x and y . Since $r(B^t/2) = 0$, the associated probability p_0 , determined by (24) and (25) does not enter the calculation.

Example: Let $B = 10$, $t = 1$ as in the previous examples of this section.

Then, using (26), $p_5 = 1/10$ as previously determined, yields $E(\epsilon) = -(\frac{1}{2})(10^{-1})(1/10) = -5 \times 10^{-3}$.

Theorem 4.4 Let $x \in [0,1)$ and $y \in [0,1)$ be random t -digit fixed point numbers of the form (13). Define the rounding error by $\epsilon = xy - fi(xy)$ where ϵ can assume the B^t values

$$(27) \quad r(v) = \begin{cases} vB^{-2t} & \text{for chopping} \\ -\frac{1}{2}B^{-t} + vB^{-2t} & \text{for symmetric rounding, } B \text{ even} \\ -\frac{1}{2}B^{-t} + (v + \frac{1}{2})B^{-2t} & \text{for symmetric rounding, } B \text{ odd} \end{cases}$$

when $v = 0, 1, 2, \dots, B^{t-1}$. Let

$$(28) \quad \mathcal{J}_v = \{g: g | (B^{2t} r(v), B^t), 1 \leq g < B^t\}$$

and define by P_v the probability that ϵ assumes a predetermined value $r(v)$. Then

$$(29) \quad P_v = \text{Prob}(\epsilon = r(v)) = B^{-2t} \sum_{g \in \mathcal{J}_v} g \delta(B^t/g) + B^{-t} \delta_{0, r(v)} B^{2t}$$

where $\delta_{i,k}$ is the Kronecker symbol. The corresponding expectations are

$$(30) \quad E(\epsilon) = \begin{cases} \frac{1}{2} B^{-2t} \sum_{G=1}^{B^t-1} [1 - (G, B^t)/B^t] & \text{for chopping} \\ -\frac{1}{2} B^{-t} P_0 & \text{for symmetric rounding, } B \text{ even} \\ 0 & \text{for symmetric rounding, } B \text{ odd} \end{cases}$$

Proof: We partition the problem so that we can use the results of Theorems 4.1 to 4.3 and Corollary 4.1. We note that the probability for a given value of x is now B^{-t} and not $(B^t-1)^{-1}$ as previously. Thus, for chopping, using (16), we have for $\epsilon \neq 0$,

$$(31) \quad P_v = (B^t - 1) B^{-t} P_v \quad (v \neq 0)$$

and for $\epsilon = 0$,

$$(32) \quad \begin{aligned} P_0 &= \text{Pr}(\epsilon = 0 | x = 0) + \text{Pr}(\epsilon = 0 | x \neq 0) \\ &= B^{-t} + (B^t - 1) B^{-t} P_0 \quad (v = 0) \end{aligned}$$

This proves (29) if $r(v) = v B^{-2t}$. For symmetric rounding, we proceed as in Corollary 4.1. But instead of adjusting v to a parameter μ as was

done in equations (23) to (25), we choose to adjust the set S_v of (15) to the set $\mathcal{S}_v$ of (28) which has the same effect as the shifting operations (25), but in addition permits a more compact representation. Thus, if B is even, $r(\frac{1}{2}B^t) = 0$, and if B is odd, $r(v) = 0$ for $v = \frac{1}{2}(B^t - 1)$. For these two cases, $\epsilon = 0$, and the probability for $\epsilon = 0$ is that of (32). This proves the term $B^{-t} \delta_{0, r(v)B^{2t}}$ in (29).

The expectation (30) is derived as for (17) and (26). Since the terms $r(v)P_v$ where $r(v) = 0$ do not enter the calculation, (30) is obtained from (17) and (26) by multiplying these two equations by $(B^t - 1)/B^t$ to account for the different distribution for x in $(0,1)$ and $[0,1)$.

Note that the variances for the different rounding procedures are different, but related to the cases $x \in (0,1)$ by

$$(33) \quad \text{Var}(\epsilon) = E(\epsilon^2) - E^2(\epsilon) = \frac{B^t - 1}{B^t} E(\epsilon^2 | x=0) - \left[\frac{B^t - 1}{B^t} E(\epsilon | x=0) \right]^2$$

This follows from the partitioning of

$$E(\epsilon^2) = E(\epsilon^2 | x \neq 0) \cdot \text{Prob}(x \neq 0) + E(\epsilon^2 | x = 0) \text{Prob}(x = 0)$$

where $\text{Prob}(x \neq 0) = (B^t - 1)B^{-t}$ and $E(\epsilon^2 | x = 0) = 0$. A similar argument applies to $E(\epsilon)$ as shown above, and hence for $E^2(\epsilon)$ also.

Examples: Let $B = 10$, $t = 1$. By direct calculation it is easily verified that for chopping, and

$$v = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9$$

$$P_v = .27, .04, .12, .04, .12, .09, .12, .04, .12, .04$$

respectively. Using (27) to (33) we obtain

$$\mathcal{J}_0 = \{1, 2, 5\}, P_0 = \frac{1}{100} \{1 \times 4 + 2 \times 4 + 5 \times 1\} + 10^{-1} = .27$$

$$E(\epsilon) = \frac{1}{200} \left[9 - \frac{1}{10} [1 + 2 + 1 + 2 + 5 + 2 + 1 + 2 + 1] \right] = .0365$$

$$\text{Var}(\epsilon) = \frac{9}{10} \left[8.275 \times 10^{-4} + (.0405556)^2 \right] - (.0365)^2 = 8.928 \times 10^{-4}$$

The term in brackets [] = $\text{Var}(\epsilon|x=0) + E^2(\epsilon^2|x=0)$, thus using previously calculated values.

For symmetric rounding we obtain with $B=10$, $t=1$, and

$$\begin{aligned} v &= 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 \\ 100r(v) &= -5, -4, -3, -2, -1, 0, 1, 2, 3, 4 \\ P_v &= .09, .12, .04, .12, .04, .27, .04, .12, .04, .12 \end{aligned}$$

respectively. Using (27) to (33) we obtain

$$\mathcal{J}_0 = \{1, 5\}, P_0 = \frac{1}{100} \{1 \times 4 + 5 \times 1\} = .09$$

$$E(\epsilon) = -\frac{.09}{20} = -.0045, E(\epsilon^2) = 7.85 \times 10^{-4}$$

$$\text{Var}(\epsilon) = [7.85 - (.45)^2] \times 10^{-4} = 7.6475 \times 10^{-4}$$

The formulas for the means and variances simplify if B is prime. They are listed in Table 4.2.

As in floating point arithmetic, symmetric rounding when B is odd, produces no bias in the the truncation error under multiplication.

3. Inner Product with Fixed Point Operations

We consider the computation of

$$(34) \quad S_n = \sum_{i=1}^n a_i b_i$$

by t-digit fixed point arithmetic . If a_i and b_i are t-digit fixed point machine numbers, then each term introduces an error $\epsilon_i = r_i$, $r_i = a_i b_i - \text{fi}(a_i b_i)$. If a_i and b_i are points from the continuous number line, then we first round to a t-digit number such that $a_i = \bar{a}_i + \alpha_i$, $b_i = \bar{b}_i + \beta_i$, and hence

$$(35) \quad a_i b_i = f_i(\bar{a}_i \bar{b}_i) + \epsilon_i$$

$$(36) \quad \epsilon_i = r_i + \bar{a}_i \beta_i + \bar{b}_i \alpha_i + \alpha_i \beta_i$$

In both cases, because of strict fixed point arithmetic, the total error is given by

$$(37) \quad \epsilon^{(n)} = S_n - \text{fi}(\sum a_i b_i) = \sum_{i=1}^n \epsilon_i$$

It is clear, that the probability distributions of ϵ_i depend on the sample space from which a_i, b_i are taken. In many cases, all numbers a_i, b_i ($i=1, 2, \dots, n$) are independently drawn from the same sample space. We make this assumption. Then we can draw some general conclusions, using the Central Limit Theorem which is stated below.

Theorem 4.5 (Central Limit): Let $\{\xi_k\}_{k=1}^{\infty}$ be a sequence of independent identically distributed random variables such that $E(\xi_k) = \mu$ and $\text{Var}(\xi_k) = \sigma^2 < \infty$ for each k. Let $s_n = \sum_{i=1}^n \xi_k$, $z_n = (s_n - n\mu)/\sqrt{n}\sigma$, and let $f_n(z)$ be the probability density function for z_n . Then

$$(38) \quad \lim_{n \rightarrow \infty} f_n(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad z \in (-\infty, \infty).$$

For a proof see any standard text on probability theory, for example, Parzen (1960).

The density function (38) is a special case of the normal density functions given by

$$(39) \quad \psi(z; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-(z - \mu)^2 / 2\sigma^2) \quad \mu, z \in (-\infty, \infty), \sigma^2 > 0.$$

One important property of this family of distributions is that if x_1 and x_2 have the densities $\psi(x_1; \mu_1, \sigma_1^2)$ and $\psi(x_2; \mu_2, \sigma_2^2)$ respectively, then $y = c_1 x_1 + c_2 x_2$ has the density $\psi(y; c_1 \mu_1 + c_2 \mu_2, c_1^2 \sigma_1^2 + c_2^2 \sigma_2^2)$ for any real constants c_1 and c_2 such that c_1 and c_2 are not both zero. Hence, if $E(\epsilon_i) = \mu$, $\text{Var}(\epsilon_i) = \sigma^2$ for (36), then it follows from (37) that $\frac{(\epsilon^{(n)} - n\mu)}{\sqrt{n}\sigma}$ has approximately the density $\psi(z; 0, 1)$. Using $c_1 = \sqrt{n}\sigma$, $c_2 = 0$, we see that $\epsilon^{(n)} - n\mu$ has approximately the density $\psi(\epsilon^{(n)}; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi n}\sigma} \exp - \frac{(\epsilon^{(n)} - n\mu)^2}{2n\sigma^2}$ if n is large. All of the distributions discussed in this section fit the hypotheses of Theorem 4.5.

Example. Assume the distribution of ϵ_i is uniform, i.e. that its density is

$$(40) \quad f(\epsilon_i) = \begin{cases} B^{-t} & -\frac{1}{2}B^{-t} \leq \epsilon < \frac{1}{2}B^{-t} \\ 0 & \text{elsewhere} \end{cases}$$

for $i = 1, 2, \dots$. Then $E(\epsilon_i) = 0$, $\text{Var}(\epsilon_i) = B^{-2t}/12$,

Let $h_n(\epsilon^{(n)})$ be the density of $\epsilon^{(n)}$. From probability theory we know that

$$(41) \quad h_{n+1}(\epsilon^{(n)}) = \int_{-\infty}^{\infty} f(\epsilon_{n+1}) h_n(\epsilon^{(n+1)} - \epsilon_{n+1}) d\epsilon_{n+1}$$

Using this convolution formula we find that

$$(42) \quad h_2(\epsilon^{(2)}) = \begin{cases} B^t(1 - |\epsilon^{(2)}|) & -B^{-t} \leq \epsilon^{(2)} < B^{-t} \\ 0 & \text{elsewhere} \end{cases}$$

and

$$(43) \quad h_3(\epsilon^{(3)}) = \begin{cases} B^t(\frac{1}{2}(\epsilon^{(3)})^2 - 3/2 |\epsilon^{(3)}| + 9/8) & \frac{1}{2}B^{-t} < |\epsilon^{(3)}| < 3/2B^{-t} \\ B^t(3/4 - (\epsilon^{(3)})^2) & 0 < |\epsilon^{(3)}| \leq \frac{1}{2}B^{-t} \\ 0 & \text{elsewhere} \end{cases}$$

Graphs of equations (41) to (43) are present in Figure 4.1 and the normal density is plotted with equation (43).

The density $h_n(\cdot)$ of $\epsilon^{(n)}$ is a piecewise polynomial of degree $n-1$ with n "pieces" of equal length in the interval $(-nB^{-t}/2, nB^{-t}/2)$. For large n , rather than computing these n polynomials, we employ then Central Limit Theorem to prove that the distribution approximates the normal.

For the fixed point computation $S_n = \sum_{i=1}^n a_i b_i$ we can establish the error confidence intervals:

$(-1.96 \sqrt{nu}/\sqrt{12}, 1.96 \sqrt{nu}/\sqrt{12})$ with probability .95

$(-2.05 \sqrt{nu}/\sqrt{12}, 2.05 \sqrt{nu}/\sqrt{12})$ with probability .96

$(-2.326 \sqrt{nu}/\sqrt{12}, 2.326 \sqrt{nu}/\sqrt{12})$ with probability .98

$(-2.55 \sqrt{nu}/\sqrt{12}, 2.55 \sqrt{nu}/\sqrt{12})$ with probability .99

where $u = B^{-t}$. For the distributions derived in Section 4.2 we can obtain similar confidence interval. Thus, the .95 probability interval becomes

$$(-1.96 \sqrt{n\sigma} + n\mu, 1.96 \sqrt{n\sigma} + n\mu)$$

where $\mu = E(\epsilon)$, $\sigma^2 = \text{Var}(\epsilon)$.

Densities of $\epsilon^{(1)}$ and $\epsilon^{(2)}$ normalized to the intervals $[-\frac{1}{2}, \frac{1}{2}]$ and $[-1, 1]$.

Graph of the density of $\epsilon^{(3)}$ normalized to the interval $[-3/2, 3/2]$ and the corresponding normal distribution. Reproduced from Lippmann (1971).

Figure 4.1
Graphs of the densities of $\epsilon^{(1)}$, $\epsilon^{(2)}$ and $\epsilon^{(3)}$

REFERENCES

Dickson, L.E., Introduction to the Theory of Numbers, University of Chicago Press, 1929.

Lippman, Steven A., Elements of Probability and Statistics, Holt, Rinehart, and Winston, Inc., New York, 1971.

Goodman, Richard, "On Roundoff Error in Fixed Point Multiplication," BIT 16 (1976), 263-73.

_____ and Alan Feldstein, "Roundoff Error in Products," Computing, vol. 15, no. 3 (1975), 263-73.

Henrici, Peter, Discrete Variable Methods in Ordinary Differential Equations, Wiley, New York, 1962.

Hogg, Robert and Allen T. Craig, Introduction to Mathematical Statistics, Third Edition, MacMillan, New York, 1976.

Parzen, Emanuel, Modern Probability Theory and its Application, Wiley, New York, 1960.

Ralston, Anthony, A First Course in Numerical Analysis, McGraw-Hill, New York, 1965.

Timlake, W.P., Problems in Discrete Variable Methods in Ordinary Differential Equations, (Official set of solutions to problem in Henrici (1962)).

Urabe, Minoru, "Roundoff Error Distribution in Fixed Point Arithmetic and a Remark About the Rounding Rule", SIAM J. Numerical Analysis, vol. 5, no. 2 (1968), 202-10.

Von Neumann, John and H.H. Goldstine, "Numerical Inversion of Matrices of High Order", Bulletin of the AMS, vol. 53, no. 11 (1947), 1021-99.

V. THE CRD COMPUTER

1. Definition of the Number System

We consider the number system

(1) $\{\pm \delta^i: i = -k, -k+1, \dots, 0, 1, 2, \dots, k\} \cup \{0\}$ where $\delta > 1$ and k is some positive integer. δ is a parameter which can be changed by the programmer and is chosen very close to 1 to get good precision. The numbers are represented in a computing device by the exponent i and we can let zero have some special designation. Without loss of generality and for most of our considerations, we can refer to the set of unsigned numbers $\{\delta^i: i = -k, -k+1, \dots, 0, 1, \dots, m\}$.

The relative difference between two consecutive numbers δ^i and δ^{i+1} is

$$(2) \quad \frac{\delta^{i+1} - \delta^i}{\delta^i} = \delta - 1 = \xi$$

which depends only upon δ . Hence, the number system is called the constant relative difference (CRD) system.

On a floating point computer the relative difference between two numbers is

$$(3) \quad \frac{(m + B^{-t})B^E - mB^E}{mB^E} = \frac{B^{-t}}{m}$$

Since $m \in [1/B, 1)$ the relative difference between two floating point numbers ranges from B^{-t} to B^{1-t} . If we let ξ in (2) be B^{-t} then

$$\xi = \delta - 1 = B^{-t}$$

and thus,

$\delta = 1 + B^{-t}$ gives the precision at least as good as a t digit floating point system with base B .

However, if we choose

$$(4) \quad \delta = B^{B^{-t}}$$

we can represent B^t numbers between B^j and B^{j+1} for any j where B^{j+1} is in the range of the system. There are as many numbers as can be represented by a $t+1$ digit floating point system with base B .

2. Geometric Error

[Grau, Bareiss, 1977]

The CRD computer represents a real number x as an integer k_x where

$$(5) \quad x = \epsilon \delta^{k_x} \quad \text{with} \quad \epsilon \in [\delta^{-\frac{1}{2}}, \delta^{\frac{1}{2}})$$

This implies

$$(6) \quad \log_{\delta} x = \log_{\delta} \epsilon + k_x \quad \text{where} \quad \log_{\delta} \epsilon \in [-\frac{1}{2}, \frac{1}{2})$$

Let $\eta = \log_{\delta} \epsilon$. From Theorem 2.1 we can assume that η has a uniform density,

$$(7) \quad f(\eta) = 1 \quad \eta \in [-\frac{1}{2}, \frac{1}{2})$$

therefore ϵ has the density

$$(8) \quad g(\epsilon) = f(\log_{\delta} \epsilon) \cdot \frac{d\eta}{d\epsilon} = (1) \cdot (1/\epsilon \ln \delta) \quad \epsilon \in [\delta^{-\frac{1}{2}}, \delta^{\frac{1}{2}}).$$

η and ϵ have the following means and variances

$$(9) \quad (a) \quad E(\eta) = 0$$

$$(b) \quad \text{Var}(\eta) = 1/12$$

$$(c) \quad E(\epsilon) = E(e^{(-\ln \delta)\eta}) = (\delta^{\frac{1}{2}} - \delta^{-\frac{1}{2}}) / \ln \delta$$

$$\begin{aligned}
 (d) \quad \text{Var}(\epsilon) &= E(\epsilon^2) - (E(\epsilon))^2 \\
 &= E(e^{2(\ln \delta)\eta}) - (E(e^{(\ln \delta)\eta}))^2 \\
 &= (\delta - \delta^{-1}) / 2 \ln \delta - (\delta^{\frac{1}{2}} - \delta^{-\frac{1}{2}})^2 / (\ln \delta)^2.
 \end{aligned}$$

[Hogg, Craig, 1970].

Intuitively we would like to have $E(\epsilon) = 1$. But one can easily see that the above expectation for ϵ is not one for any δ . However, if we consider a rounding interval of form

$$\delta^\alpha \leq \epsilon < \delta^{\alpha+1}$$

where $\eta = \log_\delta \epsilon$ has a uniform distribution on $[\alpha, \alpha + 1)$ then

$$E(\epsilon) = \int_{\alpha}^{\alpha+1} \delta^\eta d\eta = (\delta^{\alpha+1} - \delta^\alpha) / \ln \delta = \delta^\alpha (\delta - 1) / \ln \delta$$

If we set $E(\epsilon) = 1$ then

$$\delta^\alpha (\delta - 1) / \ln \delta = 1$$

and hence

$$\alpha = \ln \ln \delta - \ln (\delta - 1)$$

For δ 's small enough for most practical computations α is very close to $-\frac{1}{2}$.

For example

$$\delta = e^{.01} \quad \alpha = -.500417$$

$$\delta = 1 + 10^{-5} \quad \alpha = -.5 - 4 \cdot 10^{-7}$$

$$\delta = 1 + 10^{-10}$$

$$\alpha = -.5 - 4.2 \cdot 10^{-12}$$

The differences of these α 's from $-\frac{1}{2}$ scarcely warrant any change in the original rounding interval. It does, however, seem to indicate that

$$\delta^{-\frac{1}{2}} \leq \epsilon < \delta^{\frac{1}{2}} \text{ is slightly less biased than}$$

$$\delta^{-\frac{1}{2}} < \epsilon \leq \delta^{\frac{1}{2}}.$$

3. CRD Arithmetic

[Bareiss, Grau, 1977]

Multiplication and division on the CRD computer are simply the addition and subtraction of integer exponents. Thus like fixed point addition and subtraction there is no rounding necessary.

For addition let x_1 and x_2 be two numbers to be added on the CRD computer where $x_1 = \delta^{k_1}$ and $x_2 = \delta^{k_2}$. Then

$$x_1 + x_2 = x_1(1 + x_2/x_1)$$

Hence

$$\begin{aligned} \log_{\delta}(x_1 + x_2) &= \log_{\delta} x_1 + \log_{\delta}(1 + x_2/x_1) \\ &= k_1 + \log_{\delta}(1 + \delta^{k_2 - k_1}) \end{aligned}$$

Since the CRD computer deals only in integer exponents we let $\delta^{k_3^+}$ be the machine result, thus

$$(10) \quad k_3^+ = k_1 + \text{round}(\log_{\delta}(1 + \delta^{k_2 - k_1}))$$

By a similar reduction if we wish to subtract two numbers x_1 and x_2

then if $\delta^{k_3^-}$ is the machine result of the subtraction then

$$(11) \quad k_3^- = k_1 + \text{round} (\log_{\delta}(1 - \delta^{k_2 - k_1}))$$

The operation error comes from the second terms of equations (10) and (11). We know that

$$(12) \quad \log_{\delta}(x_1 \pm x_2) = k_3^{\pm} + \eta \quad \text{where } \eta \in [-\frac{1}{2}, \frac{1}{2}).$$

by our definition of the rounding operation.

Consider the sum

$$(13) \quad S_n = \sum_{i=1}^n x_i$$

and define $\text{CRD}(x)$ to be the machine representation of x . Hence, if

$\text{CRD}(x_i) = k_i$ where δ^{k_i} is the machine approximation for x then

$$\text{CRD}(S_n) = \text{CRD}(S_{n-1}) + \text{round}(1 + x_n / \text{CRD}(S_{n-1})).$$

By induction

$$(14) \quad \text{CRD}(S_n) = \log_{\delta}(S_n) + \eta_S \text{ where } \eta_S \in [^{(1-n)}/2, ^{(n-1)}/2).$$

Experiments taken of 2000 CRD sums of 20 numbers taken from a reciprocal distribution on $[\frac{1}{2}, 2)$ show that the bounds of η_S are very conservative. The operations were simulated on a CDC 6600. The bases δ

were 2^{-t} with $t = 8, 14, \text{ and } 22$. The results are listed in Table 5.1.

One may observe that in all case η_s does not go outside the range $[-5,4)$ and only two errors out of all 6000 samples were outside the interval $[-4,4)$. Research to find stronger error bounds than these for the computation of (13) is warranted.

Since the variable η in (12) has positive probability on a bounded interval $E(\eta)$ and $\text{Var}(\eta)$ exist. Since $\eta = \sum_{i=1}^n \eta_i$ where $\{\eta_i\}_{i=1}^n$ satisfy

the criteria of the Central Limit Theorem and for large n , η must have a distribution which is approximately normal.

Table 5.1

Error η_s in exponents from 2000 samples on 20 CRD additions with $\delta = 2^{2^{-t}}$.

<u>t</u>	<u>Interval</u>	<u>Number of Errors in Interval</u>	<u>Percent</u>
8	[-5, -4)	2	.1
	[-4, -3)	22	1.1
	[-3, -2)	116	5.8
	[-2, -1)	423	21.15
	[-1, 0)	649	32.45
	[0, 1)	517	25.85
	[1, 2)	212	10.6
	[2, 3)	58	2.9
	[3, 4)	1	.05
14	[-4, -3)	10	.5
	[-3, -2)	73	3.65
	[-2, -1)	304	15.2
	[-1, 0)	604	30.2
	[0, 1)	622	31.1
	[1, 2)	302	15.1
	[2, 3)	76	3.8
	[3, 4)	9	.45
22	[-4, -3)	10	.5
	[-3, -2)	91	4.55
	[-2, -1)	275	13.75
	[-1, 0)	621	31.05
	[0, 1)	605	30.25
	[1, 2)	317	15.85
	[2, 3)	75	3.75
	[3, 4)	6	.3

In all other intervals no errors of that size occurred.

References

Bareiss, Erwin H. and Albert A. Grau, "Basics of the CRD Computer, Northwestern University ERDA Report, COO-2280-35, (August 1977).

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

Hogg, Robert V. and Allen T. Craig, Introduction to Mathematical Statistics Third Edition, MacMillan, New York, 1970.

Marasa, John D., and David W. Matula, "A Simulated Study of Correlated Error in Various Finite Precision Arithmetics", IEEE Symposium on Computer Arithmetic, College Park, Maryland, 1972.

CONCLUSIONS

The results so far obtain for floating point multiplication seem to indicate that an odd base floating point unit would perform significantly better than an even base floating point unit. Base four seems to perform better than any other power of two. The results also indicate that the employment of two guard digits works much better than using only one, but that more than two digits give the user very little extra precision.

Odd base computers also work better for fixed point arithmetic. R^* -rounding eliminates the bias of even base fixed point rounding but at the cost of a larger variance of the error.

CRD addition seems to be as accurate as floating point addition, and the deterministic error bounds we obtain for it are unreasonably high for most purposes.

The probabilistic error distributions for addition and subtraction on both the CRD and floating point computers are areas where future research is certainly needed. We also need to establish probabilistic error bounds for extended operations similar to the deterministic error bounds in Section I.

APPENDIX

Example for the distribution of the error obtained by chopping the product of two random fixed point numbers for $B = 10$ and $t = 2$.

TABLE A1
Distribution p_v of chopping error e in units of 10^{-4}

v	00	01	02	03	04	05	06	07	08	09		
00	520	40	80	40	160	80	80	40	160	40	160	90
10	160	40	160	40	80	80	160	40	80	40	320	80
20	320	40	80	40	160	130	80	40	160	40	160	70
30	160	40	160	40	80	80	160	40	80	40	320	60
40	320	40	80	40	160	80	80	40	160	40	260	50
		09	08	07	06	05	04	03	02	01	00	v

Table A1 gives the probability that the error assumes a specified value. For example the error 75×10^{-4} occurs with probability

$$p_{75} = 130 \times 10^{-4} = .013.$$

The probability for the chopping error to have a specified leading digit is given below.

Leading digit:	0	1	2	3	4	5	6	7	8	9
Probability in units of 10^{-4} :	1240	880	1090	880	1040	980	1040	930	1040	880

For example, an error e , $.70 \leq 100e \leq .79$ occurs with probability $930 \times 10^{-4} = .093$. For further explanation, see page 74 of the text.