BIT 15 (1975), 165–173

TOWARDS ACCURATE STATISTICAL ESTIMATION OF ROUNDING ERRORS IN FLOATING-POINT COMPUTATIONS

SEPPO LINNAINMAA

Abstract.

A new method of estimating *a posteriori* the statistical characteristics of the rounding errors of an arbitrary algorithm is presented. This method is based on a discrete model of the distribution of rounding errors which makes more accurate estimates possible. The analysis is given for both rounding and truncating arithmetic. Finally, some experimental results are reported.

1. Introduction.

The accumulated rounding error R_N of the resulting value u_N of a numerical algorithm can generally be expressed quite accurately by a Taylor expansion of first degree with respect to the local errors r_{u_i} of the initial and intermediate values u_i of the algorithm [10,5]. Thus

$$R_N \approx \sum_i c_i r_{u_i} \tag{1}$$

for some coefficients c_i. In statistical analysis the local errors r_{u_i} are usually treated as mutually independent random variables, and thus the expected value ER_N and variance D^2R_N of R_N can be estimated using formulae [1]

$$ER_N \approx \sum_i c_i Er_{u_i}\,, \tag{2a}$$

$$D^2R_N \approx \sum_i c_i^2 D^2 r_{u_i}\,. \tag{2b}$$

Estimation of the behavior of R_N may be attempted when the computing process has either arbitrary or fixed initial values. The former case involves the use of *global analysis*. In this type of analysis, an attempt is made to estimate *a priori* the general behavior of the rounding errors in a certain computing algorithm.

Local analysis involves the use of fixed initial data. Of course, with given initial values and a fixed arithmetic, R_N is unique, but in this case the analysis is concerned with clarification of the behavior of R_N when

Received Feb. 11, 1975. Revised April 10, 1975.

the computing precision is varied or when the initial values are slightly perturbed.

This article describes a new discrete model of the distribution of r_{u_i}, for the purposes of local analysis. Used together with an existing effective *a posteriori* algorithm for determining the coefficients c_i in (1) with fixed initial data [5, 11], more accurate estimates can be made for the characteristics of R_N than was previously possible. The analysis is given for both rounding and truncating arithmetic.

A more thorough treatment of the results of this article is given in the report [7], together with other existing methods for the statistical estimation of rounding errors on the basis of the Taylor expansion (1). In that report the analysis is also given for a computer-oriented unbiased arithmetic, called parity arithmetic [6].

2. The idea of discretization.

The following considerations are based on the generally used *normalized signed-magnitude representation* of a real number u to base b

$$u = s_u \times (0.u_1u_2u_3\ldots)_b \times b^{e_u} = s_u f_u b^{e_u},\ u_1 \neq 0 \text{ if } u \neq 0\,, \tag{3}$$

where $s_u \in \{-1, 1\}$ is the *sign*, f_u the *fraction* and e_u the *exponent* of u. In a computer u is *rounded to t digits*. Thus a *floating-point number* of *precision t*, $\mathrm{fl}(u,t)$, is obtained. The number

$$r_u = fl(u,t) - u \tag{4}$$

is known as the (*absolute*) *rounding error* of the number $fl(u,t)$.

It has been shown both experimentally and theoretically that the fraction f_u of an arbitrary non-zero real number u, used in computations, obeys approximately the *logarithm law* [4]. This means that $\log_b f_u$ is uniformly distributed in the interval $[-1, 0)$. On this basis it is possible to determine the probability p_{kj} that the kth digit u_k of an arbitrary non-zero real number u is equal to j. We have [7]

$$p_{kj} = \begin{cases} \log_b(1+1/j), & k = 1,\ \ j = 1,2,\ldots,b-1, \\ 1/b + ((b-1)/(2b\ln b))\,(b-2j-1)b^{-k} + O(b^{-2k}), & \\ & k = 2,3,\ldots,\ \ j = 0,1,\ldots,b-1\,. \end{cases} \tag{5}$$

On the basis of (5) it is natural to expect r_u to be practically uniformly distributed for all practical values of the precision t, since b^{-k} can be assumed to be negligible with respect to $1/b$ when $k \geqq t$.

Thus the usual assumption that the local absolute rounding error is uniformly distributed between its extreme values [e.g. 1, 2, 3] is quite

well justified, and is surely accurate enough for the purposes of global analysis. However, it is obvious that when two floating-point numbers, containing at most t non-zero digits, are added, subtracted or multiplied the resulting value is a terminating real number. Thus it is possible to obtain more accurate estimates for the distribution of local rounding errors in local analysis by treating local rounding errors as discrete rather than continuous random variables. The most significant discrepancy between discrete and continuous models occurs in additions and subtractions.

3. Some practical units.

As we are performing a statistical analysis, a number u to be rounded should not be thought as a fixed number but as one which may attain all the values which are produced when the operands producing u are perturbed slightly without increasing the number of their significant digits. Thus the digits u_i, $i = 1, 2, \ldots$ in representation (3) are, except for the first few, not unique in local analysis. If it is known that $u_1, u_2, \ldots, u_l$ may differ from zero, but that $u_i = 0$, $i > l$, then u may be represented as

$$u = s_u \times (0.u_1 u_2 \ldots u_l)_b \times b^{e_u},\ u_1 \neq 0 \quad \text{if} \quad u \neq 0\,. \tag{6}$$

The concept of the *number unit* n_u (to the base b) of u is now introduced and defined to be the largest possible number of the form b^k, where k is an integer, such that all possible values of u are reached by multiplying this number by an integer. Since the number zero is reached by multiplying any finite number by zero, it is natural to assume its number unit to be ∞. Correspondingly, it is natural to assume the number unit of u to be zero, if it has no such terminating representation as (6). Thus

$$n_u = \begin{cases} \infty, & \text{if } u = 0, \\ 0\,, & \text{if } u \text{ has no terminating representation,} \\ b^{e_u - l} = (0.0_1 0_2 \ldots 0_{l-1} 1_l)_b \times b^{e_u}, & \text{otherwise}\,, \end{cases} \tag{7}$$

where l is as defined in (6).

When the number u is rounded to $\mathrm{fl}(u,t)$, the unit of the least significant digit which still can be expressed, called the *machine unit* of the number u, is defined as[1]

$$m_u = \begin{cases} s_u b^{e_u - t} = s_u \times (0.0_1 0_2 \ldots 0_{t-1} 1_t)_b \times b^{e_u}, & u \neq 0\,, \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad\quad , & u = 0\,. \end{cases} \tag{8}$$

[1] The case of renormalization is not noted since it occurs very rarely, in statistical sense, and this article deals with statistical analysis.

Further, the *rounding unit* of u is defined as

$$d_u = n_u/|m_u| \,. \tag{9}$$

Obviously, $d_u \geqq 1$ implies $\mathrm{fl}(u,t) = u$. If $d < 1$, then the number of digits to be rounded is equal to $l - t = -\log_b d_u$. If u is a product of two floating-point numbers, then $l-t$ normally equals t or $t-1$. In the case of division $l-t$ is normally ∞, and if u is a sum or difference then any positive integer value of $l-t$ is possible, although small values occur more frequently than large ones [9].

4. A method for statistical analysis of rounding errors.

As previously mentioned, it is natural to expect that if $0 < d_u < 1$ then the number $(0.u_{t+1}u_{t+2}\ldots u_l)_b$ attains all its possible values with equal probability, i.e.

$$P\{(0.u_{t+1}u_{t+2}\ldots u_l)_b = id_u\} = d_u, \quad i = 0,1,\ldots,d_u^{-1}-1 \,, \tag{10}$$

where $P\{A\}$ denotes the probability of event A.

There exists an obvious and quite troublesome exception to distribution (10). It occurs when u is the sum (or difference) of two floating-point numbers, say v and w, having quite different exponents. If the perturbation which causes the statistical distribution is quite small, then only the last few digits of $(0.u_{t+1}u_{t+2}\ldots u_l)_b$ vary (see Fig. 1) and (10)

$$\begin{array}{rl} & v\ v\ v\ v\ v\ v\ \underline{v}\ \underline{v} \\ + & \qquad\qquad w\ w\ w\ w\ w\ w\ \underline{w}\ \underline{w} \\ \hline & u\ u\ u\ u\ u\ u\ \underline{u}\ \underline{u}\ u\ u\ \underline{u}\ \underline{u} \end{array}$$

Figure 1. Addition of two floating-point numbers of different magnitude such that the perturbed digits (underlined) do not essentially affect the rounding error of the addition.

obviously does not hold. This effect is clearest when $e_v - e_w > t$. In rounding arithmetic, for example, the rounding error is then always equal to $-w$. This failure in formula (10) is not easily corrected, but nevertheless the theory does not break down since it is concerned with statistical considerations and large values of $|e_v - e_w|$ appear quite seldom, as the article by Sweeney [9] demonstrates. In addition, (10) always essentially holds when "perturbation" means computation by varying precision. In such cases t varies and thus the positions of the digits of u chosen in (10) also vary.

In view of the rounding rules of rounding[1] and truncating arithmetic, (10) implies, *provided that* $d_u < 1$,

(11a) $Er_u = \frac{1}{2}d_u m_u$, $D^2 r_u = (1/12)(1-d_u^2)m_u^2$, $r_u \in [-(\frac{1}{2}-d_u)m_u, \frac{1}{2}m_u]$

for rounding arithmetic with even b,

(11b) $Er_u = 0$, $D^2 r_u = (1/12)(1-d_u^2)m_u^2$, $r_u \in [-\frac{1}{2}(1-d_u)m_u, \frac{1}{2}(1-d_u)m_u]$

for rounding arithmetic with odd b, and

(11c) $Er_u = -\frac{1}{2}(1-d_u)m_u$, $D^2 r_u = (1/12)(1-d_u^2)m_u^2$, $r_u \in [-(1-d_u)m_u, 0]$

for truncating arithmetic. As noted after (9), $d \geqq 1$ implies $Er_u = D^2 r_u = 0$.

When formulae (11) are employed in (2), our *accurate local method* for estimating the distribution of rounding errors is obtained.

The utilization of (11) requires the value of the rounding unit d_u. Equation (9) can be used to determine this value once the value of the number unit n_u is known. Determining the value of the number unit is not a trivial problem since, when a number is computed with a computer, only its rounded and not its accurate value is known. However, there exist quite obvious formulae for this purpose, utilizing the number units of the operands producing u [7].

In binary-based computers (i.e. $b = 2^k$ for some integer k) our accurate local method can be further refined, in some situations, when binary number units are used, instead of the number units to the base b. E.g. the special properties of the number 2 as a multiplier and a divisor will be automatically observed if $n_2 = (10)_2 = 2$.

It is interesting to note that the well-known [8] but until now unmeasurable bias of rounding arithmetic can be measured using (11a). It has statistical importance only when the number of digits to be rounded is small. As Sweeney [9] has found experimentally, this situation is quite common in addition and subtraction. Thus, it is not surprising that the bias of the rounding arithmetic can clearly be observed in the experimental results reported below. It is natural for the bias to become stronger with smaller base numbers, since the maximum value of Er_u is $\frac{1}{2}m_u b^{-1}$.

5. Application to matrix inversion.

In the experiment described below, in which a matrix is inverted using the Gauss-Jordan method, the present method of estimation is

[1] In rounding arithmetic, $m_u/2$ is added to u, before truncation.

tested with the algorithm and the initial data remaining fixed but the precision of the arithmetic being varied. This is a typical local problem.

In order to perform this experiment, a 5×5 matrix was constructed whose elements were random numbers obeying the logarithm law, with random signs and the absolute values chosen from the interval $(\frac{1}{2}, 2)$. The column vectors of this matrix were orthonormalized using the Gram-Schmidt method. The matrix U thus obtained was employed as the eigenvector matrix of the final test matrix $A = U\Lambda U^T$. The eigenvalues of A, i.e. the diagonal elements of the diagonal matrix Λ, were positive random numbers chosen from the interval $(2^{-6}, 2^6)$. Thus the following random matrix was obtained:

$$A \approx \begin{pmatrix} 8.293 & -2.210 & 7.697 & -1.977 & 10.14 \\ -2.210 & 22.05 & -5.222 & -11.39 & 2.308 \\ 7.697 & -5.222 & 9.265 & -1.837 & 8.848 \\ -1.977 & -11.39 & -1.837 & 9.671 & -4.946 \\ 10.14 & 2.308 & 8.848 & -4.946 & 13.99 \end{pmatrix}$$

whose inverse is

$$A^{-1} \approx \begin{pmatrix} 7.021 & 4.365 & 4.185 & 3.719 & -7.142 \\ 4.365 & 4.991 & 5.867 & 4.821 & -5.993 \\ 4.185 & 5.867 & 7.464 & 5.767 & -6.684 \\ 3.719 & 4.821 & 5.767 & 4.868 & -5.417 \\ -7.142 & -5.993 & -6.684 & -5.417 & 8.550 \end{pmatrix}.$$

The largest and smallest eigenvalues af A are 30.642 and 0.035014.

An arithmetic simulator, programmed for a Burroughs B6700 computer, was utilized in these experiments, thus enabling the use of different bases, rounding rules and precisions. The "exact" inverse of A was first computed employing a precision of 300 bits and the fact that $A^{-1} = U\Lambda^{-1}U^T$. Element (2,3) of A^{-1}, having a value of 5.867, was then arbitrarily chosen for further investigation.

Of primary consideration was the Gauss-Jordan method for computing A^{-1}. Pivoting was not utilized during the inversion in order to ensure that no change in the computing order occurred when the precision was varied.

Employing this method, matrix A was first inverted 250 times using base 2 and varying the precision from 21 to 270 bits, then 100 times using base 16 and varying the precision from 6 to 105 hexadecimal digits. Thus the experimental statistical characteristics of the accumulated rounding error of element (2,3) of A^{-1} were obtained for bases 2

and 16. The experiment was performed for both rounding and truncating arithmetic.

The theoretical statistical characteristics of element (2,3) were computed using our *accurate local method*. For comparison, the experiment was repeated for the so-called *basic local* and *global* methods [7]. The former assumes that the rounding error is distributed (continuously) uniformly between its extreme values. The latter assumes, in addition, that the fraction is distributed according to the logarithm law.

Although some obvious dependencies between the local errors of the Gauss-Jordan method can be pointed out [7], excellent results were achieved in the theoretical estimations, as shown in Table 1.

Table 1. *Estimates of the statistical characteristics of the accumulated rounding error of the inverse element* (2,3) *when the precision is varied. All estimates are given in machine units of the accurate result.*

Arithm.	Method	Base 2, Sample size 250				Base 16, Sample size 100			
		Expected v.		Stand. dev.		Expected v.		Stand. dev.	
		Estim.	t	Estim.	t	Estim.	t	Estim.	t
Round.	Experim.	16.4		121		21.1		239	
	Acc. local	18.0	−0.22	116	0.91	−1.41	1.04	217	1.43
	Basic local	0	2.21	117	0.71	0	0.96	220	1.24
	Global	0	2.26	114	1.31	0	0.80	263	−1.31
Trunc.	Experim.	−151		111		12.0		241	
	Acc. local	−151	0.01	116	−1.07	19.0	−0.32	217	1.58
	Basic local	−151	−0.06	117	−1.26	0.31	0.53	220	1.39
	Global	−156	0.74	121	−1.85	−293	8.02	381	−5.18

The experimental expected value (m) and standard deviation (s) of the accumulated rounding error were tested for consistency with the corresponding theorical values (ER and DR) for each theoretical method. Student's t-test was utilized, the t-value being computed from the formula $t=(m-ER)\times\sqrt{n}/DR$ for the expected value and from the formula $t=(s-DR)\times\sqrt{2n}/DR$ for the standard deviation. The sample size is denoted by n. Since the percentile value $t_{0.975}$ for Student's t-distribution is 1.96 for large samples, t-values of magnitude greater than 1.96, at the 5% significance level, can be rejected.

As expected, the global method generally produced the worst estimates, since the problem is typically local. The estimates for standard

deviation are particularly poor with base 16. This is a logical consequence of the fact that the range within which the value of the fraction may vary is much larger with hexadecimal than with binary numbers. Greater error is therefore possible when the standard deviation of local errors is estimated.

The bias of the rounding arithmetic is apparent from the experimental results for binary arithmetic. The hypothesis that the expected value of the accumulated rounding error equals zero can be rejected at the 5% significance level. Only the accurate local method was able to indicate this bias. As expected, the biasity was not equally obvious in hexadecimal arithmetic, due to the larger base number.

The preceding experiment was repeated so that perturbation of the initial data was used instead of varying precision, to obtain the experimental statistical characteristics of the accumulated rounding error. The results of this experiment, presented in [7], are in accordance with the results given in Table 1.

6. Conclusions.

Several minor sources of inaccuracy still remain in our *a posteriori* method for predicting the statistical behavior of accumulated rounding errors: the local rounding errors are assumed to be independent, the exponents of the local numbers are treated as though they remain unchanged over the whole range of variation considered, and the distribution of the rounding errors is assumed to be caused by varying precision rather than perturbation. However, as the experimental results demonstrate, all these weaknesses are quite insignificant. In fact, our method, based upon new statistical considerations, has shown its value as a method giving more trustworthy predictions of the statistical behavior of rounding errors than has previously been possible.

Acknowledgement.

I am very obliged to Professor Martti Tienari for his stimulating guidance and criticism in the preparation of this article.

REFERENCES

1. P. Henrici, *Elements of numerical analysis*, Ch. 16, Wiley, New York, 1964.
2. T. E. Hull and J. R. Swenson, *Tests of probabilistic models for propagation of roundoff errors*, Comm. ACM 9 (1966), 108–113.
3. T. Kaneko and B. Liu, *On local round-off errors in floating-point aritmetic*, J. ACM 20 (1973), 391–398.

4. D. E. Knuth, *Seminumerical algorithms*, The Art of Computer Programming, Vol.2, Sec. 4.2, Addison-Wesley, Reading, Massachusetts, 1969.
5. S. Linnainmaa, *The representation of the cumulative rounding error of an algorithm as a Taylor expansion of local rounding errors* (*In Finnish*), Master's Thesis, Department of Computer Science, University of Helsinki, Helsinki, Finland, 1972.
6. S. Linnainmaa, *Analysis of some known methods of improving the accuracy of floating-point sums*, BIT 14 (1974), 167–202.
7. S. Linnainmaa, *Statistical estimation methods of rounding errors in floating-point computations*, Report A-1975-1, Department of Computer Science, University of Helsinki, Helsinki, Finland.
8. J. B. Scarborough, *Numerical mathematical analysis*, 5th ed., Ch. I, Johns Hopkins Press, Baltimore, 1962.
9. D. W. Sweeney, *An analysis of floating-point addition*, IBM Systems Journal 4 (1965), 31–42.
10. M. Tienari, *A statistical model of roundoff errors for varying length floating-point arithmetic*, BIT 10 (1970), 355–365.
11. M. Tienari, *On some topological properties of numerical algorithms*, BIT 12 (1972), 409–433.

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF HELSINKI
TÖÖLÖNKATU 11
SF-00100 HELSINKI 10
FINLAND