On the Error Estimation in Floating-point Arithmetic

Shin-ichiro Yamashita*

Abstract

A new method to estimate calculation errors in floating-point arithmetic is shown. The new error bound formula in addition or subtraction gives more precise estimates of error bound than Wilkinson's formula.

1. Introduction

The recent development of computers has remarkably increased calculation ability and made possible enormous calculation, and the numerical values used are mainly floating-points. Floating-points have been designed to keep significant digits constant, and are convenient to simultaneously process a variety of small and large numerical values. However, the results of calculations do not always hold errorless significant digits in them. The cause is that the accumulation of calculation errors (the term "error" hereafter means calculation error) generated at the respective stages of calculation may become too large to be disregarded for the final results, and that the significant digits cannot be well kept because of digit cancelling, etc. Therefore, it becomes necessary to estimate to what extent the final results are precise, i.e, how large the error is. However, it is difficult because in the floating-point arithmetic an associative law and a distributive law are not always established, etc.

These difficulties have been overcome by J.H. Wilkinson[5) 6)] by introducing equations (7) to (10). For the error estimations, there is a problem of time and labour spent for it, besides the problem of theoretical difficulties. Time and labour mainly depend upon the number of digits in calculations. The error estimations can be precisely calculated as the number of digits increases, namely, as time and labour to be spent is large. Increasing the number of digits excessively is, however, undesirable, and the

This paper first appeared in Japanese in Joho-Shori (Journal of the Information Processing Society of Japan), Vol. 15, No. 12 (1974), pp. 935~939.

* Department of software, Fujitsu Limited

normally allowable number of digits is almost the same as the number of digits in calculations at the most. Under these restrictions, it is impossible to calculate precisely the errors, but possible error bounds instead. However, error bounds are often overestimated compared to the real errors. Estimation formulae by J. H. Wilkinson cannot overcome this defect, either. Conversely, even in the most fundamental accumulated sums, they sometimes give overestimations. The most prominent reasons why overestimations are made are that there is a defect in the fundamental estimation formulae, and that error vanishing is not taken into account. The object of this paper is to improve the overestimations of errors by complementing the disadvantage of J.H. Wilkinson in almost the same time and labour as of his estimation formulae of error bounds.

2. Fundamental Nature of Floating-point Numbers

Let us assume that the numerical values of floating-point mentioned hereafter are expressed by base - M - system L-digit. Obtaining a floating-point number A from an arbitrary real number A* is expressed as

(1) $A = FL(A^*)$,

where A is an approximately value of A*. The relation between A and A* are, when $A^* \neq 0$. and $A \neq 0$, as follows.

Assuming

(2) $M^e \leqq |A^*| < M^{e+1}, \qquad M^e \leqq |A| \leqq M^{e+1}$,

where e is an appropriate integer, we get

(3) $|A - A^*| \leqq \gamma \times M^{e+1-L}$,

where $0.5 \leqq \gamma < 1.0$. The term γ determines the nature of FL (rounding-down, rounding-up, count fractions more than 0.5 as one and less than 0.5 as zero, etc.), and can be referred to as "rounding-off coefficient". Next, u is defined in the form

(4) $u = \gamma \times M^{1-L}$,

Using such u, the right side of equation (3) can be expressed by A* and A.

(5) $|A - A^*| \leqq |A^*|u, |A - A^*| \leqq |A|u.$

Furthermore, we obtain the following result from this.

(6) $A = A^*(1 + \alpha) = A^*(1 + \beta)^{-1}, \qquad |\alpha|, |\beta| \leqq u.$

The term u in equation (4) is called to as "unit roundoff"[9], and is basic unit for floating-point numbers.

3. Fundamental Operations of Floating-point Numbers

For four arithmetic operations as to floating-point numbers X and Y, J. H. Wilkinson has proposed the following relations.

(7) $FL(X \pm Y) = X(1+\alpha) \pm Y(1+\beta);\ |\alpha|\ ,|\beta| \leqq u$

(8) $FL(X \overset{\times}{\div} Y) = (X \overset{\times}{\div} Y)(1+\gamma);\ |\gamma| \leqq u$

(9) $|FL(X \pm Y) - (X \pm Y)| \leqq (|X| + |Y|)u$

(10) $|FL(X \overset{\times}{\div} Y) - (X \overset{\times}{\div} Y)| \leqq |X \overset{\times}{\div} Y|\ u$

On the basis of these, where Y ≠ 0 is assumed in division, he has made error estimations. However, when these Wilkinson's formulae and those in the reference 1) are applied for various problems, it turns out that the formers give somewhat overestimations. Then the latters are improved here using FL, and the following relations are obtained.

(11) $FL(X \pm Y) = (X \pm Y) + \max\ \{|X|,|Y|,|X \pm Y|\}\varepsilon, |\varepsilon| \leqq u$

(12) $FL(X \overset{\times}{\div} Y) = (X \overset{\times}{\div} Y)(1+\gamma)^{\pm 1},\ |\gamma| \leqq u$

(13) $|FL(X \pm Y) - (X \pm Y)| \leqq \max\ \{|X|,|Y|,|X \pm Y|\}u$

$\leqq \max\{|X|,|Y|,|FL(X \pm Y)|\}(1+u)u$

(14) $|FL(X \overset{\times}{\div} Y) - (X \overset{\times}{\div} Y)| \leqq |X \overset{\times}{\div} Y|\ u;\ \ |FL(X \overset{\times}{\div} Y) - (X \overset{\times}{\div} Y)| \leqq |FL(X \overset{\times}{\div} Y)|\ u$

The results by J. H. Wilkinson and this paper are essentially different in addition and subtraction. As an example, we shall explain formula (9) by Wilkinson and formula (13) by this paper. When the results of addition and subtraction of floating-point numbers X and Y are approximated by floating-point numbers, the upper bound of errors, in formula (9), can be regarded as the sum of |X| and |Y| multiplied by u, respectively; whereas, in formula (13), it can be regarded as the largest one among |X|,|Y| and |X ± Y| multiplied by u.

The author's formula can give an explanation to the errors separated the error in digit adjustment and the rounding error[7]. Comparing formulae (9) and (13), for arbitrary X and Y, we obtain

(15) $\max\{|X|,|Y|,|X \pm Y|\}u \leqq (|X| + |Y|\ )u.$

From this equation, it can be concluded that the error bound derived from author's result cannot get larger than that of J. H. Wilkinson. This ratio can be written by

(16) $1/2 \leqq \dfrac{\max\ \{|X|,|Y|,|X \pm Y|\}}{|X| + |Y|} \leqq 1,$

where $|X| + |Y| \neq 0.$

The author's result gives the same at the most or a half error bound in comparison with Wilkinson's result. In short, for addition, these two methods are the same if X and Y have the same signs, and the author's method is better if opposite sign. It can be concluded that J. H. Wilkinson does not take into account the signs of X and Y, but the author does.

4. Error Estimations of Accumulated sums

Since it has described that the error estimations by this paper can be made a half of the results by J. H. Wilkinson at its best in the sum of two terms, it can be easily presumed that in the accumulated sums of n terms, they will be able to decrease 1/n. There are several methods to obtain the solution of accumulated sum $\sum_{k=1}^{n} X_k$. Here, we shall consider the following successive addition system normally used.

(17) $Y_1 = X_1$; $Y_k = FL(Y_{k-1} + X_k)$, K=2,3,...,n; $Y \equiv Y_n$.

Using formula (7), J. H. Wilkinson has transformer equation (17) in the form

(18) $Y_k = (1+\alpha_k)Y_{k-1} + (1+\beta_k)X_k$, K=2,3,...n; where $Y_1 = X_1$; $|\alpha_k|, |\beta_k| \leqq u$.

From this, the following equations are derived.

(19) $Y = \sum_{k=1}^{n} X_k + |X_1|(n-1)\theta_1 + \sum_{r=2}^{n} |X_r|(n+1-r)\theta_r$,

(20) $|Y - \sum_{k=1}^{n} X_k| \leqq \{|X_1|(n-1) + \sum_{r=2}^{n} |X_r|(n+1-r)\}(1+\delta)u$,

where $|\theta_k| = (1+\delta)u$, k=1,2,...,n; $(n-1)u \leqq \delta \leqq 1$.

Equations (19) and (20) are the result of accumulated sum by J. H. Wilkinson.

In contrast to it, the author transforms equation (17) on the basis of equation (11).

(21) $Y_k = (Y_{k-1} + X_k) + \max\{|Y_{k-1}|, |X_k|, |Y_k|\}\theta_k$

where $Y_1 = X_1$; $|\theta_k| \leqq (1+u)u$; k=2,3,....,n.

If the sums are taken for k which is 2 to n on both sides and if $\sum_{k=1}^{n-1} Y_k$ are subtracted from both sides,

(22) $Y = \sum_{k=1}^{n} X_k + \sum_{r=2}^{n} \max\{|Y_{r-1}|, |X_r|, |Y_r|\}\theta_r$, $|\theta_r| \leqq (1+u)u$

is obtained. From this,

(23) $|Y - \sum_{k=1}^{n} X_k| \leqq \sum_{r=2}^{n} \max\{|Y_{r-1}|, |X_r|, |Y_r|\}(1+u)u$

is moreover obtained. This is the reuslt for accumulated sum by the author.

5. Examples of Error Estimations of Accumulated Sums

Examples for comparison of accumulated sum by J. H. Wilkinson and the author are shown below.

Example 1: The case of M=2, L=t, $u=2^{-t}$; and $X_1=1$; $X_2=1-u$; $X_3 \sim X_4=1-2u$; $X_5 \sim X_8=1-2^2u$; $X_9 \sim X_{16}=1-2^3u$;...; $X_{2m-1} \sim X_{2m}=1-2^{m-1}u$; $2^m=n$.

Solution: This example has been shown by J. H. Wilkinson.

$$\text{(Correct solution)} = \sum_{k=1}^{n} X_k = 2^m - \frac{1}{3}(4^m - 1)u = n - \frac{1}{3}(n^2 - 1)u, Y \equiv Y_n = 2^m = n$$

$$\text{(Absolute value of the error)} \equiv E = |Y - \sum_{k=1}^{n} X_k| = \frac{1}{3}(n^2 - 1)u$$

$$\text{(Right side of equation (20))} \equiv E_1 = \frac{(n + 2)(n - 1)}{2}(1 + \delta)u$$

$$\text{(Right side of equation (23))} \equiv E_2 = \frac{(n + 2)(n - 1)}{2}(1 + u)u$$

The ratio of E, E_1, E_2 is $E:E_1:E_2 \doteqdot 2:3:3$ and both estimation formulae by J. H. Wilkinson and the author give almost the same estimation bounds as the real error. This example is suitable for Wilkinson's formula because of little difference between the real error and the estimated error, but the next example brings a bad estimation when his formula is used.

Example 2: The case of $X_k = (-1)^k X$

Solution: $$\text{(Right side of equation (20))} \equiv E_1 = \frac{(n + 2)(n - 1)}{2}|X|(1 + \delta)u$$

$$\text{(Right side of equation (23))} \equiv E_2 = (n - 1)|X|(1 + u)u$$

Accordingly

$$E_1 : E_2 \doteqdot n/2 : 1$$

This shows that the result of J. H. Wilkinson makes an overestimation of nearly $\frac{n}{2}$ times in comparison with the author's one.

6. Conclusion

The improved estimation formulae have been obtained. When these formulae are used, the error bound of accumulated sums of n terms, which plays an important role in error estimation of floating-point arithmetic, becomes, in the best, one-n th as large as the result of J. H. Wilkinson, and, even in the worst, gives almost the same result as his.

An important subject of error estimations is to obtain the estimation value of error bounds as closer to real error as possible, but often the excessive error estimations are obtained. By the result of this paper, the overestimation can be overcome to some extent. However, many problems still remain, and error bounds are still liable to overestimations. This mainly results from error vanishing, which will be mentioned at another time. The author would like to express his sincere thanks to Prof. T. Uno for his valuable suggestions and continuing discussions. And the author would like to express his appreciation to Prof. S. Satake for valuable suggestion for the reasoning.

References

1) S. Yamashita: Accuracy test of $\sum_{i=1}^{n} X_i$ and its applications (in Japanese), Japan Electronic Association, documents of numerical analysis executive committee, Jun. '64.

2) S. Yamashita: On calculation errors and calculation limit in a definite digits calculation, (in Japanese), Research Institute for Mathematical Science, Kyoto Univ., No. 153, pp.152-175, Feb. '72.

3) S. Yamashita, S. Satake: On the calculation limit of roots of algebraic equations (in Japanese), Joho shori, Vol. 7, No. 4 pp.197-201, '66.

4) S. Yamashita, S. Satake: On the Calculation Limit of Roots of Algebraic Equation, Information Processing in Japan, Vol. 7, pp.18~23, '67.

5) J.H. Wilkinson: Error Analysis of Floating-Point Computation, Numer. Math., Vol. 2, pp.319~340, '60

6) J.H. Wilkinson: Rounding Error in Algebraic Processes, Her Majesty's Stationary office, '63.

7) S. Yamashita: On the Error Estimation in Floating-point Arithmetic, thesis, '73. 10.

8) S. Yamashita: On the Error Estimation and Error Vanishing in Floating-point Arithmetic (in Japanese), Research Institute for Mathematical Science, Kyoto Univ., No. 215, pp.143-189, Feb. '74.

9) G.E. Forsythe, C.B. Moler: Computer Solution of Linear Algebraic Systems, Prentice-Hall, '67.