≪論　文≫

# 浮動小数点演算における誤差評価について*

山　下　真一郎**

**Abstract**

An new method to estimate calculation errors in floating-point arithmetic is shown.
In addition or subtraction, the error bound is given by a formula-
$|FL(x\pm y)-(x\pm y)|\leqq\max\{|x|, |y|, |x\pm y|\}u$.
It includes Wilkinson's formula and gives more precise estimates of error bound than that.

## 1. 序　文

近年，電子計算機の発達によって，計算能力が飛躍的に増大し，尨大な計算が可能になったが，そこで取扱われる数値は主に浮動小数点方式である．浮動小数点方式は有効数字を一定に保持するために工夫されたものであり，大小様々な大きさの数値を同時に取扱うのに便利である．しかし，計算結果がいつも誤差***のない有効数字を保持しているとは限らない．それは，計算途中の各段階で発生した誤差の集積が最後の結果にとって無視しえない大きさになったり，桁落ちなどのために有効数字を保持しえなくなるからである．そこで，最後の結果がどの程度正しいか，すなわち，誤差がどの程度の大きさであるかを評価する必要がある．しかし，それは困難である．その理由は，浮動小数点演算では，結合律，分配律が必ずしも成立しないことなどによる．

J. H. Wilkinson 5) 6) によって，(9)～(12)式が導入され，これらの困難が回避された．誤差評価は理論的に困難であるという外に，手間が問題である．手間は計算桁数に集約でき，桁数を増せば，いくらでも精密に評価できる．しかし，桁数をむやみに増やすことは望ましくなく，普通に許容される桁数は高々計算桁数と同程度である．このような制約の下では誤差を精密に確定することはできず，誤差限界を求めることになる．ところが，誤差限界はしばしば実際の誤差よりも，非常に過大に評価しがちである．

Wilkinson の評価式もこの欠点を免れえない．しかも，もっとも基本的な累和においてさえ，過大な評価をもたらすことがある．この最大の理由は，基本的な評価式に欠点があるためと誤差消失を考慮に入れないことがあげられる．

この論文の目的は Wilkinson の誤差限界の評価式と同程度の手間で彼の欠点を補い，誤差の過大評価を解消することにある．

## 2. 浮動小数点の数値の基本性質

以下に述べる浮動小数点の数値は $M$ 進法 $L$ 桁で表わされているとする．

任意の実数 $A^*$ から浮動小数点の数値 $A$ を得ることを

(1)　$A=FL(A^*)$

と書く．$A$ は $A^*$ の近似値であって，$A$ と $A^*$ の間の関係は $A^*\neq 0$, $A\neq 0$ のとき，以下に述べる通りである．

まず，

(2)　$M^e\leqq|A^*|<M^{e+1}$，但し $e$ は適当な整数と仮定する．この仮定は，実数 $A^*$ が $M$ 進法の正規化された浮動小数点で表わされたときの指数が $e$ になるような仮定である（負数を補数表示する計算機では，$A^*$ が負で，ちょうど $M$ の幅乗のときに少し事情を異にするがそれは無視する）．

数直線上に浮動小数点の数値を並べると，$A^*$ の仮定の下では，$A$ は

(3)　$M^e\leqq|A|\leqq M^{e+1}$

の範囲にあり，適当な整数を選ぶことによって，

(4)　$|A|=b_i$

但し $b_i=M^e+i\times M^{e+1-L}$, $0\leqq i\leqq M^L-M^{L-1}$

のように表わされる．$A$ がこのような $b_i$ となるため

* On the Error Estimation in Floating-point Arithmetic, by Shin-ichiro YAMASHITA (Fujitsu Limited)

** 富士通（株）

*** ここで述べる「誤差」は計算誤差の意味である．

には，$A^*$ は，切捨て法によると，$b_i \leqq |A^*| < b_{i+1}$ のときであり，切上げ法によると，$b_{i-1} < |A^*| \leqq b_i$ のときである．いわゆる四捨五入法：(一般には，$M/2-1$ 捨 $M/2$ 入法) によると $b_i - 0.5\times M^{e+1-L} \leqq |A^*| < b_i + 0.5\times M^{e+1-L}$ のときである．その外，いろいろな方法でも，

(5) $$|A-A^*| \leqq \gamma \times M^{e+1-L},$$
但し $0.5 \leqq \gamma < 1.0$

となる．$\gamma$ は $FL$ の性質（切捨て，切上げ，四捨五入等）を規定する値で，“丸め係数”とでも呼ぶべきである．

次のように $u$ を定義する．

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

このような $u$ を用いて，(5)式の右辺を $A^*$ 及び $A$ で表わせば，(2)，(3)式を勘案して，

(7) $$|A-A^*| \leqq |A^*|u, \quad |A-A^*| \leqq |A|u$$

を得る．これから，また，次式を得る．

(8) $$A = A^*(1+\alpha) = A^*(1+\beta)^{-1},$$
但し $|\alpha|, |\beta| \leqq u$

(6)式の $u$ は文献9）によると“丸め誤差の単位”(unit roundoff) と呼び，浮動小数点の数値にとって，基本的な単位である．

## 3. 浮動小数点の基本演算

Wilkinson は浮動小数点の数値 $x, y$ に対する四則演算について，

(9) $$FL(x \pm y) = x(1+\alpha) \pm y(1+\beta);$$
$$|\alpha|, |\beta| \leqq u$$

(10) $$FL(x \overset{\times}{\div} y) = (x \overset{\times}{\div} y)(1+\gamma); \quad |\gamma| \leqq u$$

(11) $$|FL(x \pm y) - (x \pm y)| \leqq (|x| + |y|)u$$

(12) $$|FL(x \overset{\times}{\div} y) - (x \overset{\times}{\div} y)| \leqq |x \overset{\times}{\div} y|u$$

が成立することを述べ，これを基礎に誤差評価を行なっている．但し除算では $y \neq 0$ とする．

しかし，いろいろな問題について，これによる結果と参考文献1）の結果と比較すると，幾分過大な結果が得られる．著者の結果は稚拙であったのでこれを $FL$ を使って改め，次のような結果を得た．

(13) $$FL(x \pm y) = (x \pm y) + \max\{|x|, |y|, |x \pm y|\}\varepsilon, \quad |\varepsilon| \leqq u$$

(14) $$FL(x \overset{\times}{\div} y) = (x \overset{\times}{\div} y)(1+\gamma)^{\pm 1}, \quad |\gamma| \leqq u$$

(15) $$|FL(x \pm y) - (x \pm y)| \leqq \max\{|x|, |y|, |x \pm y|\}u \leqq \max\{|x|, |y|, |FL(x \pm y)|\}(1+u)u$$

(16) $$|FL(x \overset{\times}{\div} y) - (x \overset{\times}{\div} y)| \leqq |x \overset{\times}{\div} y|u;$$
$$|FL(x \overset{\times}{\div} y) - (x \overset{\times}{\div} y)| \leqq |FL(x \overset{\times}{\div} y)|u$$

Wilkinson と著者の結果が本質的に違うのは，加減算であるが，例えば，Wilkinson の結果を (11) 式で説明すると，浮動小数点の数値 $x, y$ の加減算の結果を浮動小数点の数値で近似したときの誤差の上限は，$|x|, |y|$ の下位 $u$ 単位が別々に伝播して加わったと見ることができる．これに対し，著者の結果の(15)式は，$|x|, |y|$ または，$|x \pm y|$ の大きい方の下位 $u$ 単位が伝播したものと見ることができる．

著者の結果は，加減算のアルゴリズムをよく検討すれば理解されるが，次章で説明する．

(11)式と(15)式の右辺を比較すると，任意の $x, y$ に対して，

(17) $$\max\{|x|, |y|, |x \pm y|\}u \leqq (|x| + |y|)u$$

となって，著者の結果から導かれる加減算の誤差限界は Wilkinson のそれよりも大きくなることがない．この比は

(18) $$1/2 \leqq \frac{\max\{|x|, |y|, |x \pm y|\}}{|x| + |y|} \leqq 1,$$
但し $|x| + |y| \neq 0$

となって，Wilkinson の結果に比べて，著者の結果はたかだか同じか半分になる．

Wilkinson と著者の結果は，端的に言えば，加算に対して，$x$ と $y$ が同符号ならば同じになり，異符号ならば，著者の方がよいことになる．

Wilkinson は $x$ と $y$ の符号について考慮を払わなかったが，著者はその点にも考慮を払ったと言うことができる．

## 4. 加減算の誤差

$x, y$ を $M$ 進法 $L$ 桁の浮動小数点の数値とする．加減算の結果は，桁上がりがあるか，桁上がりがないかのどちらかであり，誤差は演算の始めに小数点の位置を揃える，いわゆる桁揃えの段階に発生するものと最後の結果を丸める段階で発生するものとに分けて考えることができる．

桁揃えは，$x$ と $y$ の指数差だけ $\min(|x|, |y|)$ の上位に0を補い，常に，不要な下位の桁を切捨てる．それは，切捨てておけば，最後の結果を丸めるとき，どのような方法で丸めても，そのような丸めが1回行なわれるときと同じ結果になるが，切上げたり，いわゆる四捨五入すると，切上げが2回行なわれることが

あり得るからである.

加減算器の桁数は，切捨て，切上げ方式の演算では少なくとも，$L$桁あると考えてよいから，桁揃えの誤差は，$\max(|x|,|y|)$ の末尾1の単位，すなわち，$\max(|x|,|y|)u$ を越えない．いわゆる四捨五入方式の演算では，加減算器の桁数は，$L$桁よりも多くなければならないから，やはり，桁揃えの誤差は $\max(|x|,|y|)u$ を越えない.

(13)，(15)式について，桁上がりがある場合と桁上りがない場合に分けて説明する.

桁上りがある場合には，桁揃えの誤差が発生しても桁揃えを行なったのちの加減算の結果の指数と $(x\pm y)$ の指数の大きさは変らず，加減算の結果の誤差は，実数 $(x\pm y)$ を浮動小数点の数値で近似したときの誤差の上限を越えない．したがって，桁上りする場合には，

(19)　$|FL(x\pm y)-(x\pm y)|\leqq|x\pm y|u$

である.

桁上りがない場合には，加減算の結果の誤差は，桁落ちがあれば，桁揃えの誤差を越えず，桁落ちがなければ，桁揃えの誤差があったとしても，指数が $\max(|x|,|y|)$ と同じ指数をもつ浮動小数点の数値で近似したときの誤差限界を越えない．したがって，いずれも，

(20)　$|FL(x\pm y)-(x\pm y)|\leqq\max(|x|,|y|)u$

である.

桁上がりがある場合は，$|x\pm y|>\max(|x|,|y|)$ であり，(19)と(20)は排他的であるから，結局，加減算においては

(21)　$|FL(x\pm y)-(x\pm y)|\leqq\max\{|x|,|y|,|x\pm y|\}u$

が成立する．これから (13)，(15) 式が得られる.

## 5. 累和の誤差評価

加減算に対する誤差限界が Wilkinson の結果よりも都合がよければ半分に出来ることを示したから，その結果として，$n$ 個の累和では，$1/n$ に出来るだろことは簡単に推測できる．累和 $\sum_{k=1}^{n}x_k$ を求める方法にはいろいろあるが[7]，普通用いられる次のような逐次加算方式について考察する.

(22)　$y_1=x_1;\ y_k=FL(y_{k-1}+x_k),\ k=2,3,\cdots,n;\ y\equiv y_n$

Wilkinson は(9)式に基づき，(22)式を

(23)　$y_k=(1+\alpha_k)y_{k-1}+(1+\beta_k)x_k,\ k=2,3,\cdots n;$

但し $y_1=x_1;\ |\alpha_k|,|\beta_k|\leqq u$

とおいて，これから次式を導いた.

(24)　$y=\sum_{k=1}^{n}(1+\eta_k)x_k$

但し $1+\eta_1=\prod_{i=2}^{n}(1+\alpha_i)$

$1+\eta_r=(1+\beta_r)\prod_{i=r+1}^{n}(1+\alpha_i),\quad r=2,3,\cdots,n$

$\prod_{i=n+1}^{n}(1+\alpha_i)=1.$

これを評価するには次の補題が役立つ.

**補題**　$m=1,2,3,\cdots$に対して，$|\varepsilon_k|\leqq u,\ k=1,2,\cdots,m$ かつ $0\leqq mu\leqq\delta\leqq1$ ならば，次式が成立する.

(25)　$1-mu\leqq\prod_{k=1}^{m}(1+\varepsilon_k)\leqq1+(1+\delta)mu.$

これから，(24)式は $0\leqq(n-1)u\leqq\delta\leqq1$ と仮定して

$$\begin{cases}1-(n-1)u\leqq1+\eta_1\leqq1+(1+\delta)(n-1)u\\ 1-(n+1-r)u\leqq1+\eta_r\leqq1+(1+\delta)(n+1-r)u\end{cases}$$

となるから，

(26)　$y=\sum_{k=1}^{n}x_k+\{|x_1|(n-1)\theta_1+\sum_{r=2}^{n}|x_r|(n+1-r)\theta_r\}$

但し $|\theta_k|\leqq(1+\delta)u,\ k=1,2,\cdots,n$

(27)　$|y-\sum_{k=1}^{n}x_k|\leqq\{|x_1|(n-1)+\sum_{r=2}^{n}|x_r|(n+1-r)\}(1+\delta)u$

が得られる．これが Wilkinson の累和に対する結果である.

これに対し，著者は(13)式に基づき，(22)式を

(28)　$y_k=(y_{k-1}+x_k)+\max\{|y_{k-1}|,|x_k|,|y_k|\}\theta_k$

但し $y_1=x_1;\ |\theta_k|\leqq(1+u)u;\ k=2,3,\cdots,n$

とおいて，この両辺の $k=2$ から $k=n$ までの和を求め，その両辺から $\sum_{k=1}^{n-1}y_k$ を辺々引いて

(29)　$y=\sum_{k=1}^{n}x_k+\sum_{r=2}^{n}\max\{|y_{r-1}|,|x_r|,|y_r|\}\theta_r,\ |\theta_r|\leqq(1+u)u$

を得た．これからまた

(30)　$|y-\sum_{k=1}^{n}x_k|\leqq\sum_{r=2}^{n}\max\{|y_{r-1}|,|x_r|,|y_r|\}(1+u)u$

**Table. 1** Result of $\sum_{k=1}^{n} x_k$, $x_k=(-1)^k \sin(k\theta)$, $n=1000$

| $\theta$ | $y_n$ | $y_n^*$ | $y_n-y_n^*$ | $A_n$ | $B_n$ |
|---|---|---|---|---|---|
| 0.01 | −0.276608214 | −0.276606245 | $3.079\times10^{-6}$ | $4.741\times10^{-3}$ | $9.185\times10^{-6}$ |
| 0.10 | −0.256627545 | −0.256627773 | $2.285\times10^{-7}$ | $4.753\times10^{-3}$ | $9.532\times10^{-6}$ |
| 0.90 | 0.273375422 | 0.273374510 | $9.119\times10^{-7}$ | $4.753\times10^{-3}$ | $1.052\times10^{-5}$ |
| 1.00 | 0.293903798 | 0.293902961 | $8.366\times10^{-7}$ | $4.753\times10^{-3}$ | $1.097\times10^{-5}$ |
| 1.30 | −0.360859275 | −0.360860019 | $7.443\times10^{-7}$ | $4.750\times10^{-3}$ | $1.183\times10^{-5}$ |
| 1.57 | −0.507606521 | −0.507607939 | $1.418\times10^{-6}$ | $4.500\times10^{-3}$ | $1.322\times10^{-5}$ |
| 2.00 | −0.599824384 | −0.599826427 | $2.044\times10^{-6}$ | $4.753\times10^{-3}$ | $1.689\times10^{-5}$ |
| 3.00 | −13.8203630 | −13.8203672 | $4.193\times10^{-6}$ | $4.753\times10^{-3}$ | $1.101\times10^{-4}$ |

を得る．これが著者の累和に対する結果である．

## 6. 累和の計算例

Wilkinsonと著者との累和に対する結果を比較する例題を以下に示す。

**例題 1** $M=2$, $L=t$, $u=2^{-t}$ で次の場合，

$x_1=1$; $x_2=1-u$; $x_3\sim x_4=1-2u$; $x_5\sim x_8=1-2^2u$; $x_9\sim x_{16}=1-2^3u$; …; $x_{2^{m-1}+1}\sim x_{2^m}=1-2^{m-1}u$; $2^m\equiv m$

**解** この例題は Wilkinson 2), p.19) が示したものである．

$$\begin{cases} (\text{真値})\equiv\sum_{k=1}^{n} x_k=2^m-\frac{1}{3}(4^m-1)u \\ \qquad\qquad =n-\frac{1}{3}(n^2-1)u \\ y\equiv y_n=2^m=n \\ (\text{誤差の絶対値})\equiv E=|y-\sum_{k=1}^{n} x_k| \\ \qquad\qquad =\frac{1}{3}(n^2-1)u \\ ((27)\text{式の右辺})\equiv E_1=\frac{(n+2)(n-1)}{2}(1+\delta)u \\ ((30)\text{式の右辺})\equiv E_2=\frac{(n+2)(n-1)}{2}(1+u)u \end{cases}$$

となるから

$$E : E_1 : E_2 \fallingdotseq 2 : 3 : 3$$

となって，Wilkinsonの評価式も著者の評価式も，ほとんど実際の誤差に近い限界を与える．この例題は真の誤差と評価式による誤差とがあまり差がなくWilkinsonにとって都合がよいものであるが，次の例題は都合が悪い．

**例題 2** $x_k=(-1)^k x$ の場合

**解**
$$\begin{cases} ((27)\text{式の右辺})\equiv E_1 \\ \qquad\qquad =\frac{(n+2)(n-1)}{2}|x|(1+\delta)u \\ ((30)\text{式の右辺})\equiv E_2=(n-1)|x|(1+u)u \end{cases}$$

となるから，

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

となって，Wilkinsonの評価式は著者のそれに比べて，$n$（のオーダ）倍過大な限界を与える．

**例題 3** $x_k=(-1)^k \sin(k\theta)$ の場合：

**解** これは例題1と例題2が両極端の例なので，その中間的な例題として示すものである．いくつかの $\theta$ について，計算結果を **Table. 1** に示す．表中の値は次式で求めたものである．

$$y_k=FL(y_{k-1}+x_k);\ k=1,2,\cdots,n;\ y_0=0$$

$$y_n^*=\sum_{k=1}^{n} x_k$$

$$A_n=\sum_{k=1}^{n} |x_k|(n+1-k)(1+nu)u$$

$$B_n=\sum_{n=1}^{n} \max\{|y_{k-1}|, |x_k|, |y_k|\}(1+u)u$$

$x_k$ は $(-1)^k \sin(k\theta)$ を倍精度で計算し，それを単精度に丸めたものである．それを倍精度で累和したのが $y_n^*$ である．$A_n$ は Wilkinson の結果の近似値であり，$B_n$ は著者の結果による評価値である．いずれも単精度で求めた．なお，計算機は FACOM 230-60 によった．すなわち，$M=2$, $L=26$, $u=2^{-26}$ で計算した．

$y_n-y_n^*$ はほぼ誤差を表わし，著者の結果である $B_n$ はほぼ $\sqrt{n}$ 倍過大であり，Wilkinson の結果である $A_n$ は $B_n$ のほぼ $n/2$ 倍過大である．

## 7. 結 論

浮動小数点演算の誤差評価は累和が重要な役割をはたすが，$n$ 項の累和の誤差限界を Wilkinson の評価よりも，最良の場合，ほぼ $n$ 倍改善し，最悪の場合でも同等の結果をもたらす評価式を得た．

誤差評価の重要な課題は実際の誤差に近い誤差限界の評価値を得ることであるが，過大な誤価値が得られがちである．この論文の結果によって，その過大評価

を幾分緩和することができた．しかし，これですべてが解決したわけではない．誤差限界はなお過大に見積られがちである．その最大の原因は誤差同士が打消し合うためである．これについては，別の機会に述べたい．

終りに，本論文は宇野利雄日本大学理工学部顧問教授の示唆によるものである．先生の日頃の御指導に対し深謝の意を表します．

## 参考文献

1) 山下：$\sum_{i=1}^{n} x_i$ の精度判定とその応用，電子協，NA 委員会資料，'64.
2) 一：有限桁計算における計算誤差と計算限界について，京大数研，講究録 153, p. 152～175, '72.
3) 山下，佐竹：高次代数方程式の根の計算限界について，情報処理，Vol. 7, No. 4, p.197-201, '66.
4) 一，一：On the Calculation Limit of Roots of Algebraic Equation, I. P. J, Vol. 7, p.18～23, '67.
5) J. H. Wilkinson, Error Analysis of Floating-Point Computation, Numer. Math., Vol. 2, p.319～340, '60.
6) 一：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) 山下：浮動小数点演算の誤差評価と誤差消失について，京大数研，「計算の手間と能率化」，研究集会予稿集，'74. 2.
9) G. E. Forsythe, C. B. Moler, Compter Solution of Linear Algebraic Systems, Prentice-Hall, '67.（渋谷政昭，田辺国士共訳，線形計算の基礎，培風館，'69.）

（昭和 49 年 3 月 6 日受付）
（昭和 49 年 8 月 24 日再受付）