浮動小数点演算における絶対丸め誤差のモーメントの存在とその推定†

小沢 一文††

この文献では，まず浮動小数点演算における絶対丸め誤差の確率分布を記述する式を与え，そしてそれをもとに，誤差のモーメントが存在するための十分条件および存在しないための十分条件を与えている．さらに，この2つの条件を満たす2つの例が提示されている．

また，誤差の平均，2乗平均を推定するための公式も提案されている．この公式によれば，真の値の平均，2乗平均さえ得られれば誤差の平均，2乗平均の推定が可能になる．すなわち，誤差を推定するにあたり，指数の値あるいは相対誤差の値についての情報は全く必要としない．

実験結果より，ここで得られた公式は従来のそれよりすぐれた結果を与えていることが判明した．

1. まえがき

浮動小数点演算の丸め誤差の解析については，これまでに多くの研究が行われてきた[1)~8)]．それらの研究では，絶対丸め誤差の平均，2乗平均を推定するための方法がいくつか提案されている[1)~6)]．

その一つ[1)~4)]は，相対誤差は真値と互いに独立でしかも一様に分布していると仮定することによって導かれているものであり，従来からよく用いられてきた．

しかし，相対誤差は絶対誤差を真値で割ったものであるから，これが真値に対して独立となることはあり得ず，また相対誤差分布が一様分布であるという仮定も必ずしも妥当ではない[7),8)]．しかも，この方法によって得られた誤差の推定値は実験結果とあまり一致していない[2)]．

その他，真値の指数の値が一定かつ既知である場合のみ適用できる方法も提案されているが[5),6)]，我々が数値計算を行うとき，そこで扱うデータはその平均，2乗平均のみ既知で指数については何も情報が得られないものが多い．したがって，この方法の適用範囲は非常に限られたものになる．

本論文では，絶対丸め誤差の確率分布を求め，それをもとにモーメントの存在について考察する．そして，2次までのモーメントが存在する場合を対象として，平均，2乗平均を推定する公式を導く．この公式は，従来のものとは異なり，相対誤差分布，指数については全く情報を必要としない．最後に，この公式によって得られる推定値と実験結果との比較も行う．

† Existence and Estimation of the Moments of Absolute Round off Error in Floating-point Arithmetic by KAZUFUMI OZAWA (Sendai Radio Technical College).

†† 仙台電波工業高等専門学校

2. 絶対丸め誤差の分布

実数Xの浮動小数点表現は次式である．

$$X=sgn(X)\cdot b^{e}\cdot\alpha \tag{1}$$

ここで，bは2以上の整数とし$sgn(X)$はXの符号を表わすものとする．

$$sgn(X)=\begin{cases}1, & X>0\\ 0, & X=0\\ -1, & X<0\end{cases} \tag{2}$$

また指数eは整数で，仮数αは正規化されているとする．すなわち，

$$b^{-1}\leq\alpha<1 \tag{3}$$

を満たすものとする．

仮数αをb進法を用いて表現すると，

$$\alpha=m_1b^{-1}+m_2b^{-2}+\cdots+m_ib^{-i}+\cdots \tag{4}$$

となる．ここで各m_iは，

$$1\leq m_1\leq b-1,\ 0\leq m_i\leq b-1,\ i=2,3\cdots \tag{5}$$

を満たす整数である．

実際，電子計算機においてXを記憶するときは，$sgn(X)$, e, $m_i(i=1,2,\cdots)$をコード化して記憶するわけであるが，いかなる計算機においてもこれらすべてを正確に記憶できるわけではない．

ここでは，指数e，符号$sgn(X)$は常に正確に記憶されているとし，また仮数αについてはその上位t (≥ 2) 桁しか記憶できず，$t+1$桁以下は切り捨てられるものとする．このとき実数Xとそのt桁近似との差，すなわち絶対丸め誤差の解析を行う．

実数Xのt桁近似を$fl(X)$とすれば，

$$fl(X)=sgn(X)\cdot b^{e}\cdot\gamma \tag{6}$$

となる．ここで，γ は

$$\gamma=m_1b^{-1}+m_2b^{-2}+\cdots+m_tb^{-t} \quad (7)$$

とする．このとき，絶対丸め誤差 R は

$$R=fl(X)-X=-sgn(X)\cdot b^{e}\cdot\beta \quad (8)$$

で与えられる．ここで β は

$$\beta=m_{t+1}b^{-(t+1)}+m_{t+2}b^{-(t+2)}+\cdots \quad (9)$$

とする．これは式（5）より

$$0\leq\beta<d,\ d=b^{-t} \quad (10)$$

なる不等式を満たす．

次に丸め誤差 R の分布関数について考察する．まず下に示す量を定義する．

定義 1

$$F_i^{+}(r)=P_r\{R<r, e=i, sgn(X)=1\} \quad (11)$$

$$F_i^{-}(r)=P_r\{R<r, e=i, sgn(X)=-1\} \quad (12)$$

ここで，X の分布関数を $G(x)(=P_r\{X<x\})$ とし，$G(x)$ は $-\infty<x<\infty$ において連続であると仮定する．

なお，この仮定を設けたことにより，X として離散的な値をとる確率変数は除外される．しかし，我々が数値計算で取り扱うデータは，ほとんど連続的な確率変数とみなして差し支えないものと思われる．

定理 1

$$F_i^{+}(r)=\begin{cases}0 & ;\ r\leq-b^id\\ \sum_{l=1}^{L}(G(b^i\gamma_{l+1})-G(b^i\gamma_l-r)) & \\ & -b^id<r\leq0\\ G(b^i)-G(b^{i-1}) & ;\ 0<r\end{cases} \quad (13)$$

ここで，$d=b^{-t}, \gamma_l=b^{-1}+(l-1)d, l=1,2,\cdots,L, L=(b-1)b^{t-1}$ である．

[証明] 式(8)において $e=i, sgn(X)=1$ とすると，式(10)より

$$-b^id<R=-b^i\beta\leq0 \quad (14)$$

が得られる．したがって，$r\leq-b^id$ のとき $F_i^{+}(r)=0$ が成立する．

つぎに，$-b^id<r\leq0$ のとき式(14)より

$$\begin{aligned}F_i^{+}(r)&=P_r\{-b^id<R<r, e=i, sgn(X)=1\}\\&=P_r\{-b^{-i}r<\beta<d, e=i, sgn(X)=1\}\end{aligned} \quad (15)$$

となる．ここで，α の上位 t 桁によって表わされる値が γ であるとするならば，条件 $-b^{-i}r<\beta<d$ は

$$\gamma-b^{-i}r<\alpha=\gamma+\beta<\gamma+d \quad (16)$$

と等価である．γ としては全部で $L(=(b-1)b^{t-1})$ 個の値をとり得るから，

$$F_i^{+}(r)=\sum_{l=1}^{L}P_r\{\gamma_l-b^{-i}r<\alpha<\gamma_{l+1},\ e=i, sgn(X)=1\} \quad (17)$$

となる．これを，X の分布関数 $G(x)$ を用いて表わすと，$X=b^i\alpha$ であるから，

$$\begin{aligned}F_i^{+}(r)&=\sum_{l=1}^{L}P_r\{b^i(\gamma_l-b^{-i}r)<X<b^i\gamma_{l+1}\}\\&=\sum_{l=1}^{L}(G(b^i\gamma_{l+1})-G(b^i\gamma_l-r+0))\end{aligned}$$

となるが，$G(x)$ は連続であると仮定されているから結局

$$F_i^{+}(r)=\sum_{l=1}^{L}(G(b^i\gamma_{l+1})-G(b^i\gamma_l-r)) \quad (18)$$

を得る．

つぎに，$r>0$ なる場合を考察する．$sgn(X)=1$ のときは，つねに $R\leq0$ となるから，$F_i^{+}(r)$ は

$$\begin{aligned}F_i^{+}(r)&=P_r\{R\leq0, e=i, sgn(X)=1\}\\&=P_r\{e=i, sgn(X)=1\}\end{aligned} \quad (19)$$

となる．これを $G(x)$ を用いて表わすと，

$$\begin{aligned}F_i^{+}(r)&=P_r\{b^{i-1}\leq X<b^i\}\\&=G(b^i)-G(b^{i-1})\end{aligned}$$

となり，これですべての場合について定理が証明された．

$F_i^{-}(r)$ については次の定理が成り立つ．

定理 2

$$F_i^{-}(r)=\begin{cases}0 & ;\ r\leq0\\ \sum_{l=1}^{L}(G(-b^i\gamma_l)-G(-b^i\gamma_l-r)) & \\ & 0<r\leq b^id\\ G(-b^{i-1})-G(-b^i) & ;\ b^id<r\end{cases} \quad (20)$$

[証明] 省略

ここで，指数 $e=i$ でしかも $R<r$ となる確率および R の分布関数を定義する．

定義 2

$$F_i(r)=F_i^{+}(r)+F_i^{-}(r)=P_r\{R<r, e=i\} \quad (21)$$

$$F(r)=P_r\{R<r\} \quad (22)$$

このとき，$F(r)$ は

$$F(r)=\sum_{i=-\infty}^{\infty}F_i(r)+P_r\{R<r, sgn(X)=0\}$$

として表わされる．ところが，$G(x)$ は連続であると仮定されているため，$X=0$ となる確率は 0 になり

$$F(r)=\sum_{i=-\infty}^{\infty}F_i(r) \quad (23)$$

を得る．

当然 $F(r)$ は単調非減少で $F(+\infty)=1$, $F(-\infty)=0$

を満たす.

ところで，実際の計算機では指数の値には限界があり，式(23)において i が $-\infty$ から ∞ まで加えられているのは現実的には問題があるように思われる.

しかし，定理1, 2および分布関数の性質より

$$F_i(r)<1-G(b^{i-1})+G(-b^{i-1})\to 0,\ (i\to\infty)$$

$$\begin{aligned}F_i(r)&<(G(b^i)-G(-b^i))+(G(-b^{i-1})-G(b^{i-1}))\\&\to(G(+0)-G(-0))+(G(-0)-G(+0))\\&=0,\ (i\to-\infty)\end{aligned}$$

となり，また指数の限界を $-E_l\le e\le E_u (E_l>0, E_u>0)$ としたとき E_l, E_u は十分大きいのが常であるから

$$F(r)\simeq\sum_{i=-E_l}^{E_u}F_i(r)$$

が成り立ち，指数の限界を越えた所において X が分布していても差し支えない.

また，逆に指数の限界を越えた所には X が分布していないとすれば，そのような i については $F_i(r)\equiv 0$ となるから，この場合は式(23)は全く問題ない.

次に，式 (23) 右辺における収束性について考察する．まず，$F_i(r)$ の部分和を定義する.

定義 3 自然数 n に対して，

$$S_n(r)=\sum_{i=-n}^{n}F_i(r)\tag{24}$$

と置く.

定理 3 $r\in(-\infty,\infty)$ において一様に

$$\lim_{n\to\infty}S_n(r)=F(r)\tag{25}$$

が成り立つ.

[証明] 任意の整数 N に対して，$n>m>N$ なる n, m を選べば，定理1, 2より

$$\begin{aligned}0&<\sup_r|S_n(r)-S_m(r)|\\&=\sum_{i=m+1}^{n}\sup_r F_i(r)\\&=\sum_{i=m+1}^{n}(G(b^i)-G(b^{i-1})+G(-b^{i-1})-G(-b^i))\\&=G(b^n)-G(b^m)+G(-b^m)-G(-b^n)\\&<1-G(b^N)+G(-b^N)\end{aligned}\tag{26}$$

となる．上式右辺は分布関数の性質より $N\to\infty$ のとき0に収束する．したがって，$S_n(r)$ は一様に収束する．|

系 1 $F(r)$ は $r\in(-\infty,\infty)$ において連続である.

[証明] $G(x)$ が連続であるから $S_n(r)$ も連続になる．したがって，$S_n(r)$ の一様収束極限 $F(r)$ も連続になる．|

これにより丸め誤差 R は連続な確率変数となる.

3. 絶対丸め誤差のモーメント

ここでは，絶対丸め誤差 R のモーメントについて考察する．まず，モーメントの存在条件を考察する．X のモーメントについては，次の結果が得られている[11].

ある定数 $k>0$ に対して，

$$\begin{aligned}&G(x)=O(|x|^{-k}),\ (x\to-\infty)\\&1-G(x)=O(x^{-k}),\ (x\to+\infty)\end{aligned}\tag{27}$$

が同時に成り立つとする.

このとき，整数 $\nu(0<\nu<k)$ に対して X の ν 次のモーメントが存在する．またこのとき，式(8)より

$$|R|=b^e\beta\le bd|X|\tag{28}$$

が成り立っているので，R の ν 次のモーメントも存在することになる．式 (27) の条件が満足されているとき，R の ν 次モーメント μ_ν については次の関係が成り立つ（付録参照）.

$$\mu_\nu=\int_{-\infty}^{\infty}r^\nu dF(r)=\lim_{n\to\infty}\int_{-\infty}^{\infty}r^\nu dS_n(r)\tag{29}$$

これは，μ_ν の計算において積分と極限の順序が交換可能であることを示している．この性質は，後に考察するモーメントの計算において大変役立つ.

次に，いかなる場合に R のモーメントは存在しないかを考察する.

定理 4 整数 i について，条件

$$b^{i-1}\le x\le y\le b^i\tag{30}$$

を満たす定数 x, y を任意に選んだとき，

$$G(y)-G(x)\ge K(y-x)b^{-ki}\tag{31}$$

となる i に無関係な定数 $K>0, k>0$ が常に存在するとする.

このとき，$\nu\ge k-1$ なる整数 ν に対して丸め誤差 R の ν 次モーメントは存在しない.

また，

$$-b^i\le x\le y\le -b^{i-1}\tag{32}$$

なる x, y について式(31)が成立している場合も同様に R の ν 次モーメントは存在しない.

[証明] まず，定理1, 2より

$$\begin{aligned}&F(-b^{i-1}d)-F(-b^id)\\&=\sum_{j=-\infty}^{\infty}(F_j{}^+(-b^{i-1}d)-F_j{}^+(-b^id))\\&\ge F_i{}^+(-b^{i-1}d)-F_i{}^+(-b^id)\\&=\sum_{l=1}^{L}(G(b^i\gamma_l+b^id)-G(b^i\gamma_l+b^{i-1}d))\end{aligned}$$

となる．そして，ここで

$$b^i\ge b^i\gamma_l+b^id>b^i\gamma_l+b^{i-1}d>b^{i-1},$$

$$l=1,2,\cdots,L$$

を満たすから，式(31)より

$$F(-b^{i-1}d)-F(-b^id)\geq\sum_{l=1}^{L}Kb^{i-1}(b-1)db^{-ki}$$
$$>Kd(b-1)b^{-1}b^{-(k-1)i}$$

となる．これより，正整数νに対して

$$\left|\int_{-b^id}^{-b^{i-1}d}r^\nu dF(r)\right|=\int_{-b^id}^{-b^{i-1}d}|r^\nu|dF(r)$$
$$>(b^{i-1}d)^\nu(F(-b^{i-1}d)-F(-b^id))$$
$$>Kd^{\nu+1}(b-1)b^{-(\nu+1)}b^{(\nu-k+1)i}$$

を得る．したがって，$\nu-k+1\geq 0$ ならば任意の n, m $(n>m)$ について，

$$\left|\sum_{i=m}^{n}\int_{-b^id}^{-b^{i-1}d}r^\nu dF(r)\right|=\sum_{i=m}^{n}\int_{-b^id}^{-b^{i-1}d}|r^\nu|dF(r)$$
$$>Kd^{\nu+1}(b-1)b^{-(\nu+1)}\sum_{i=m}^{n}b^{(\nu-k+1)i}\to\infty,\ (n\to\infty)$$
(33)

となり，$\int_{-\infty}^{\infty}r^\nu dF(r)$ は存在しない．

同様に，式(32)を満たす x, y について式(31)が成り立つとすれば，

$$F(b^id)-F(b^{i-1}d)>Kd(b-1)b^{-1}b^{-(k-1)i}$$

となり，やはり $\nu-k+1\geq 0$ を満たす整数νに対して $\int_{-\infty}^{\infty}r^\nu dF(r)$ は存在しない． |

系 2 定理4の仮定が成り立つとき，$\nu\geq k-1$ を満たす整数νに対して，Xのν次モーメントは存在しない．

［証明］ 式(28)より明らかである． |

例 1 t-分布を考える．このとき，$G(x)$ は

$$G(x)=\int_{-\infty}^{x}\frac{q}{(1+px^2)^\tau}dx,\ \tau>0,\ p>0,\ q>0$$
(34)

なる形で表わされるので，$b^{i-1}\leq x\leq y\leq b^i$ なる x, y について，つねに

$$G(y)-G(x)=\int_x^y\frac{q}{(1+px^2)^\tau}dx>\frac{q(y-x)}{(p+1)^\tau}b^{-\tau i}$$
(35)

となる．また

$$G(x)=O(|x|^{-(2\tau-1)}),\ x\to-\infty$$
$$1-G(x)=O(x^{-(2\tau-1)}),\ x\to\infty$$
(36)

が成立する．

したがって，$\nu\geq 2\tau-1$ なる整数νに対してはRのν次モーメントは存在せず，$\nu<2\tau-1$ なる整数νに対しては存在する．

この結果より，Cauchy 分布では $\tau=1$ であるから，Rの一次モーメント（平均）およびより高次のモーメントは存在しないことがわかる．

4. 平均，2乗平均を与える公式

Henrici[1] も指摘しているように，微分方程式の数値解法等アルゴリズムが差分方程式で記述される場合，演算の各ステップで生ずる局所的な誤差の累積によって生ずる累積丸め誤差の分布は正規分布に十分近いものになる．したがって，累積丸め誤差の平均，2乗平均は特に重要であり，これは局所的な誤差の平均，2乗平均から容易に求まる[1]．

実数Xを各ステップにおける演算結果の真の値と考えれば，Rは局所的な丸め誤差に相当するからRの平均，2乗平均を知ることは重要である．

ここでは，式 (27) の仮定が $k=3$ としたとき成り立っているような場合を対象とした平均，2乗平均を与える公式を導く．

4.1 平　均

丸め誤差Rの平均 μ_1 は式 (29) より

$$\mu_1=\lim_{n\to\infty}\int_{-\infty}^{\infty}rdS_n(r)$$
$$=\sum_{i=-\infty}^{\infty}\left(\int_0^{b^id}rdF_i^-(r)+\int_{-b^id}^{0}rdF_i^+(r)\right)$$

となる．ここで，Xの密度関数を $g(x)$ としRの平均 μ_1 を $g(x)$ を用いて表現すれば

$$\mu_1=\sum_{i=-\infty}^{\infty}\sum_{l=1}^{L}\left(\int_0^{b^id}rg(-r-b^i\gamma_l)dr\right.$$
$$\left.+\int_{-b^id}^{0}rg(-r+b^i\gamma_l)dr\right)$$
(37)

となる．上式において，$s=-r-b^i\gamma_l$，および $s=-r+b^i\gamma_l$ なる変数変換を行うと

$$\mu_1=-\sum_{i=-\infty}^{\infty}\sum_{l=1}^{L}\left(\int_{b^i\gamma_l}^{b^i\gamma_{l+1}}sg(s)ds+\int_{-b^i\gamma_{l+1}}^{-b^i\gamma_l}sg(s)ds\right)$$
$$+\sum_{i=-\infty}^{\infty}\sum_{l=1}^{L}b^i\gamma_l\left(\int_{b^i\gamma_l}^{b^i\gamma_{l+1}}g(s)ds\right.$$
$$\left.-\int_{-b^i\gamma_{l+1}}^{-b^i\gamma_l}g(s)ds\right)$$
$$=-\eta_1+\sum_{i=-\infty}^{\infty}\sum_{l=1}^{L}b^i\gamma_l\left(\int_{b^i\gamma_l}^{b^i\gamma_{l+1}}g(s)ds\right.$$
$$\left.-\int_{-b^i\gamma_{l+1}}^{-b^i\gamma_l}g(s)ds\right)$$

となる．ここで，η_1 は X の平均である．すなわち，

$$\eta_1=\int_{-\infty}^{\infty}sg(s)ds$$

である．上式の右辺第二項は，$s=b^i\sigma$ および $s=-b^i\sigma$

なる変数変換を行い，積分と和の順序を交換すると

$$第二項=\sum_{l=1}^{L}\gamma_l\int_{\gamma_l}^{\gamma_{l+1}}\left\{\sum_{i=-\infty}^{\infty}b^{2i}(g(b^i\sigma)-g(-b^i\sigma))\right\}d\sigma \tag{38}$$

となる．この式の被積分関数は，積分

$$\int_{-\infty}^{\infty}b^{2\theta}(g(b^\theta\sigma)-g(-b^\theta\sigma))d\theta=\frac{\eta_1}{\sigma^2 lnb} \tag{39}$$

をキザミ巾1の台形公式によって近似したものにほかならない．そこで，第二項の台形和を式 (39) の右辺に置き換えることを試みる．このとき

$$第二項=\simeq\sum_{l=1}^{L}\gamma_l\left(\int_{\gamma_l}^{\gamma_{l+1}}\frac{\eta_1}{\sigma^2 lnb}d\sigma\right)$$

$$=\frac{\eta_1 d}{lnb}\sum_{l=1}^{L}\frac{1}{\gamma_{l+1}} \tag{40}$$

を得る．式 (40) の和は，Eulen-Maclaurin の公式[12]より

$$d\sum_{l=1}^{L}\frac{1}{\gamma_{l+1}}=lnb-\frac{d}{2}(b-1)+O(d^2)$$

と表わされるから，

$$第二項\simeq\frac{\eta_1}{lnb}(lnb-\frac{d}{2}(b-1))+O(d^2)$$

となる．これより，

$$\mu_1\simeq-\frac{\eta_1}{2\,lnd}(b-1)d+O(d^2) \tag{41}$$

を得る．また従来から用いられてきた公式は，

$$\mu_1=-bd\eta_1/2$$

である[1)~4)]．

4.2 2乗平均

2乗平均 μ_2 は

$$\mu_2=\lim_{n\to\infty}\int_{-\infty}^{\infty}r^2ds_n(r)$$

$$=\sum_{i=-\infty}^{\infty}\sum_{l=1}^{L}\left(\int_{-b^id}^{0}r^2g(b^i\gamma_l-r)dr+\int_{0}^{b^id}r^2g(-b^i\gamma_l-r)dr\right)$$

$$=\eta_2+\sum_{l=1}^{L}\int_{\gamma_l}^{\gamma_{l+1}}(\gamma_l^2-2\gamma_l\sigma)\left\{\sum_{i=-\infty}^{\infty}b^{3i}(g(b^i\sigma)+g(-b^i\sigma))\right\}d\sigma \tag{42}$$

となる．ここで，η_2 は X の2乗平均である．

式 (42) の台形和は，積分

$$\int_{-\infty}^{\infty}b^{3\theta}(g(b^\theta\sigma)+g(-b^\theta\sigma))d\theta=\frac{\eta_2}{\sigma^3 lnb} \tag{43}$$

の近似であるから，式 (42) の台形和を式 (43) の右辺

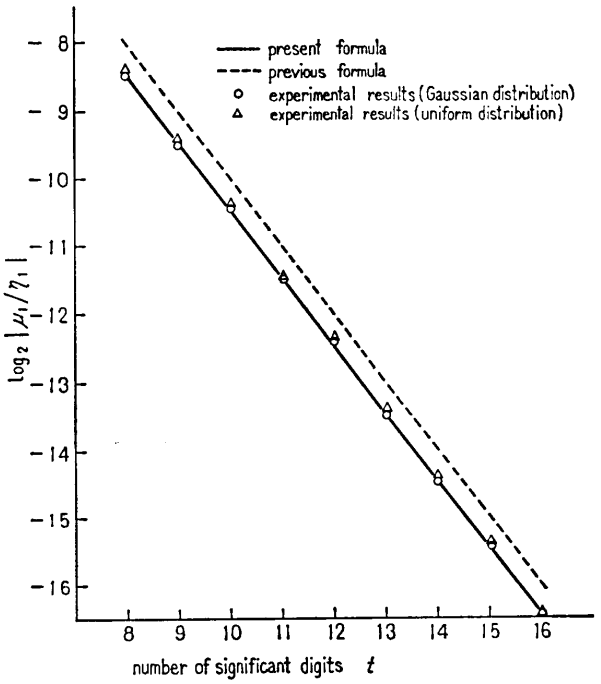

図 1 絶対丸め誤差の平均 ($b=2$)

Fig. 1 Mean of absolute round-off error ($b=2$).

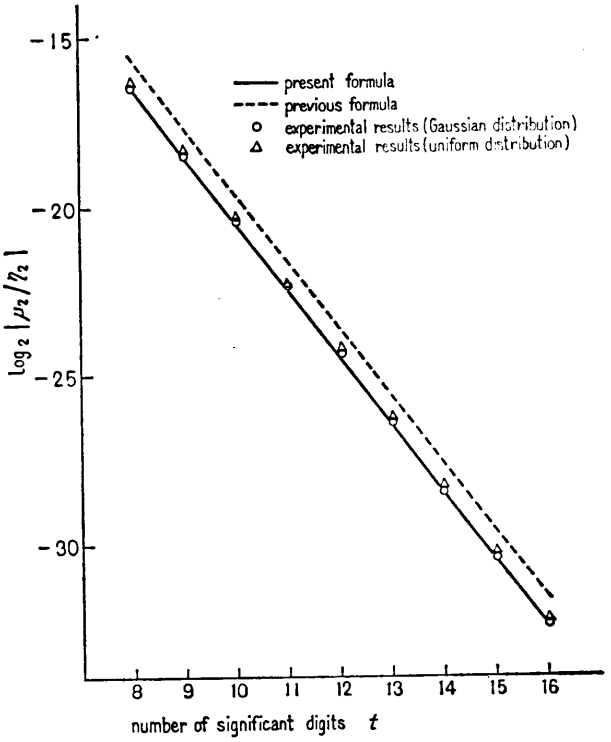

図 2 絶対丸め誤差の2乗平均 ($b=2$)

Fig. 2 Mean square of absolute round-off error ($b=2$).

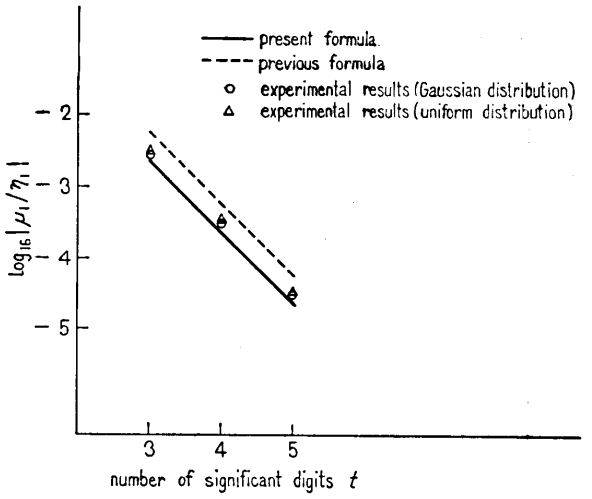

図 3 絶対丸め誤差の平均 ($b=16$)
Fig. 3 Mean of absolute round-off error ($b=16$)

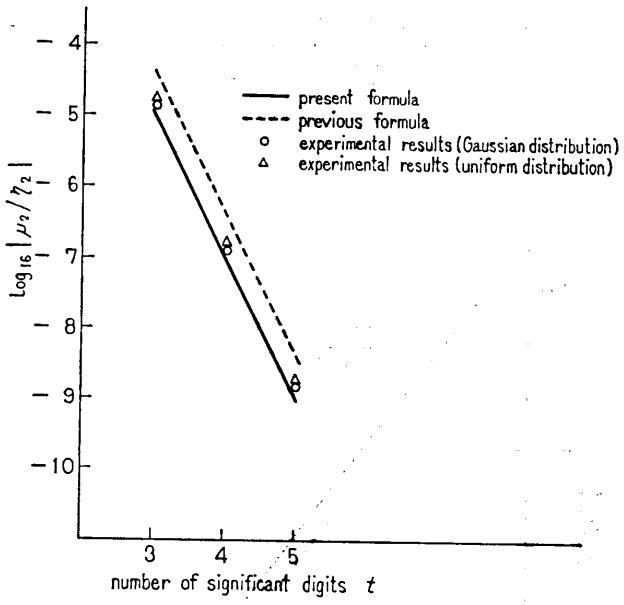

図 4 絶対丸め誤差の2乗平均 ($b=16$)
Fig. 4 Mean square of absolute round-off error ($b=16$).

に置き換えると,

$$\mu_2 \simeq \eta_2\left\{1-\frac{1}{lnb}\sum_{i=1}^{L}\left(\frac{d}{\gamma_{i+1}}+\frac{d^2}{2\gamma^2_{i+1}}\right)\right\} \tag{44}$$

を得る. これは Euler-Maclaurin の公式[12]より

$$\mu^2 \simeq \frac{\eta_2}{6lnb}(b^2-1)d^2+O(d^3) \tag{45}$$

と表わせる. 従来から用いられてきた公式は,

$$\mu_2=\frac{b^2d^2}{3}\eta_2$$

である[1)~4)].

ところで, 丸めの方式としてここで取り扱っている切り捨て方式 (chopping) のほかに, 四捨五入方式 (rounding) がある. これは十進法の四捨五入に相当するもので, b 進法においては下位桁の値 β が $d/2$ 以上ならば t 桁の値 m_t に1を加え, $d/2$ 未満ならばそのままで $t+1$ 桁以下を切り捨てる方式である. この方式の丸め誤差の分布, モーメントの存在性等については類似の結果が成り立つ. しかし, ここでは省略し平均, 2乗平均の公式[13]のみを下に示す.

$$\mu_1 \simeq -\frac{(b^2-1)\eta_1}{24\,lnb}d^2+O(d^4) \tag{46}$$

$$\mu_2 \simeq \frac{(b^2-1)\eta_2}{24\,lnb}d^2+O(d^4) \tag{47}$$

5. 実験および考察

ここでは, 真の値 X の平均, 2乗平均のみ与えられているとし, 丸め誤差 R の平均, 2乗平均の推定を行う. X の分布としては, 正規分布, 一様分布というもっとも代表的なものを扱う.

例 2 X の分布として, 平均 m, 2乗平均 $m^2+\sigma^2$ の正規分布を考える. すなわち,

$$G(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{x}e^{-(s-m)^2/2\sigma^2}ds \tag{48}$$

とする. ここで, $y=(x-m)/\sigma$ とおくと

$$G(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y}e^{-s^2/2}ds$$

と表わされ, これを $H(y)$ とおく. $s<-2$ のとき $e^{-s^2/2}<e^{s}$ が成り立つので, $y>2$ とすれば

$$1-H(y)=H(-y)$$
$$=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{-y}e^{-s^2/2}ds<\frac{1}{\sqrt{2\pi}}$$
$$\int_{-\infty}^{-y}e^{s}ds=\frac{1}{\sqrt{2\pi}}e^{-y} \tag{49}$$

を得る. したがって, $G(x)$ は $k=3$ として式(27)の仮定を満たすから公式 (41), (45) が使用できる.

ここでは, まず $b=2$ として2進27桁の仮数をもつ計算機にて有効桁 t を変化させて実験を行う.

なお, ここでは平均=1, 2乗平均=2の正規乱数を10,000回発生させ, 標本平均, 標本2乗平均を真の母集団平均, 2乗平均として採用することにする.

次に, $b=16$ として16進6桁の仮数をもつ計算機にて同じ実験を行う.

最後に, これらの結果を相対誤差が一様に分布して

表 1 関数表における丸め誤差

Table 1 Round-off error in function table.

Abscissas		Ordinates		
		Exact	Rounded	Round-off error
i	x_i	$y_i=\log(x_i)$	$\tilde{y}_i$	$r_i=\tilde{y}_i-y_i$
1	0.02	-.3912023E 1	-.391E 1	.202E-2
2	0.04	-.3218876E 1	-.321E 1	.887E-2
3	0.06	-.2813411E 1	-.281E 1	.341E-2
4	0.08	-.2525729E 1	-.252E 1	.572E-2
5	0.10	-.2302585E 1	-.230E 1	.258E-2
6	0.12	-.2120264E 1	-.212E 1	.263E-3
7	0.14	-.1966113E 1	-.196E 1	.611E-2
8	0.16	-.1832581E 1	-.183E 1	.258E-2
9	0.18	-.1714798E 1	-.171E 1	.479E-2
10	0.20	-.1609438E 1	-.160E 1	.943E-2
11	0.22	-.1514128E 1	-.151E 1	.412E-2
12	0.24	-.1427116E 1	-.142E 1	.711E-2
13	0.26	-.1347074E 1	-.134E 1	.707E-2
14	0.28	-.1272966E 1	-.127E 1	.296E-2
15	0.30	-.1203973E 1	-.120E 1	.397E-2
16	0.32	-.1139434E 1	-.113E 1	.943E-2
17	0.34	-.1078810E 1	-.107E 1	.880E-2
18	0.36	-.1021651E 1	-.102E 1	.165E-2
19	0.38	-.9675840E 0	-.967E 0	.584E-3
20	0.40	-.9162907E 0	-.916E 0	.290E-3
21	0.42	-.8675006E 0	-.867E 0	.500E-3
22	0.44	-.8209806E 0	-.820E 0	.980E-3
23	0.46	-.7765288E 0	-.776E 0	.528E-3
24	0.48	-.7339692E 0	-.733E 0	.969E-3
25	0.50	-.6931472E 0	-.693E 0	.147E-3
26	0.52	-.6539265E 0	-.653E 0	.926E-3
27	0.54	-.6161861E 0	-.616E 0	.186E-3
28	0.56	-.5798185E 0	-.579E 0	.818E-3
29	0.58	-.5447272E 0	-.544E 0	.727E-3
30	0.60	-.5108256E 0	-.510E 0	.825E-3
31	0.62	-.4780358E 0	-.478E 0	.358E-4
32	0.64	-.4462871E 0	-.446E 0	.287E-3
33	0.66	-.4155154E 0	-.415E 0	.515E-3
34	0.68	-.3856625E 0	-.385E 0	.662E-3
35	0.70	-.3566749E 0	-.356E 0	.674E-3
36	0.72	-.3285041E 0	-.328E 0	.504E-3
37	0.74	-.3011051E 0	-.301E 0	.105E-3
38	0.76	-.2744368E 0	-.274E 0	.436E-3
39	0.78	-.2484614E 0	-.248E 0	.461E-3
40	0.80	-.2231436E 0	-.223E 0	.143E-3
41	0.82	-.1984509E 0	-.198E 0	.450E-3
42	0.84	-.1743534E 0	-.174E 0	.353E-3
43	0.86	-.1508229E 0	-.150E 0	.822E-3
44	0.88	-.1278334E 0	-.127E 0	.833E-3
45	0.90	-.1053605E 0	-.105E 0	.360E-3
46	0.92	-.8338161E-1	-.833E-1	.816E-4
47	0.94	-.6187540E-1	-.618E-1	.754E-4
48	0.96	-.4082199E-1	-.408E-1	.219E-4
49	0.98	-.2020271E-1	-.202E-1	.270E-5
50	1.00	.0000000E 0	.000E 0	.000E 0

$$\eta_1=\frac{1}{50}\sum_{i=1}^{50} y_i=-.942\,E\,0,\quad \mu_1=\frac{1}{50}\sum_{i=1}^{50} r_i=.211\,E\text{-}2,$$

$$\eta_2=\frac{1}{50}\sum_{i=1}^{50} y_i{}^2=.166\,E\,1,\quad \mu_2=\frac{1}{50}\sum_{i=1}^{50} r_i{}^2=.122\,E\text{-}4,$$

Experimental $\begin{bmatrix}\mu_1/\eta_1=-.223\,E\text{-}2\\ \mu_2/\eta_2=\ .735\,E\text{-}5\end{bmatrix}$

Theoretical (present formula) $\begin{bmatrix}\mu_1/\eta_1=-.195\,E\text{-}2\\ \mu_2/\eta_2=\ .717\,E\text{-}5\end{bmatrix}$

Theoretical (previous formula) $\begin{bmatrix}\mu_1/\eta_1=.500\,E\text{-}2\\ \mu_2/\eta_2=.333\,E\text{-}4\end{bmatrix}$

いるとして導出された従来の公式[1)~4)]と比較する．結果を図 1~4 に示す．

例 3 Xの分布が有限区間における一様分布であるとする．このとき，$G(x)$ は $k=3$ として明らかに式 (27) の条件を満たすので，公式 (41), (45) が使用できる．

例 2 と同じ平均，2 乗平均をもつ一様乱数を 10,000 回発生させ，同じ要領で実験を行う．

例 4 次に，10 進法で表わされた関数表において，上位 3 桁を残し下位桁を切り捨てた場合の誤差について考察する．表 1 に関数の値 y と丸め誤差の値 r を示す．

μ_1, μ_2 を誤差の標本平均，二乗平均より求め，η_1, η_2 は y の標本平均，二乗平均より求める．そして，μ_1/η_1, μ_2/η_2 を計算し理論値 (41), (45) と比較する．

さて，これらの結果より，我々の推定値は従来の公式より高い精度をもっていることが判明した．また，$b=16$ のとき多少精度が落ちるようである．これは，積分 (39), (43) において b を大きくするということは，対応する台形和のキザミ巾を大きくすることに相当するから，b が大きくなるに従って積分に対する台形和による近似が不十分になるためと思われる．しかし，実用上は問題ない．

ここでは，分布関数 $G(x)$ が連続であると仮定されているため，Xが離散的な分布をもつ場合はこの公式は適用できない．

なお，これらの実験では一様乱数は乗算合同法 $x_{n+1}\equiv 65539\ x_n (\mathrm{Mod}\ 2^{31})$ を区間 (0, 1) に正規化したものを用い，正規乱数はこの一様乱数列を 12 回加え平均 6 を引いたものに* 標準偏差を掛け，平均を加えることによって合成した．

また，用いた計算機は 2 進の実験は ACOS 77-900 で 16 進の場合は FACOM 230-28 である．最後の関数表の例は YHP-25 を用いた．

6. あとがき

本論文では，浮動小数点演算における絶対丸め誤差の分布を求め，それをもとに絶対丸め誤差のモーメン

* 平均 0，分散 1 の正規乱数としては，この方法によるものが大変優れていることが証明されている[4)]．

トが存在する条件および存在しない条件を明らかにした.

また,2次までのモーメントが存在する場合を対象に,平均,2乗平均の推定値を与える公式を導いた.この公式は,これまでに得られた同様の公式と異なり,真の値の平均,2乗平均さえ得られれば使用できるものである.

本論文の結果は,数値計算の各種アルゴリズムにおける局所的な丸め誤差の平均,2乗平均の推定に役立つものと思われる.

最後に,日頃ご指導頂く東北大学工学部竹田研究室の諸氏に謝意を表す.

参考文献

1) Henrici, P.: Discrete Variable Methods in Ordinary Differential Equations, pp. 108-184, Wiley, New York (1962).
2) Liu, B. and Kaneko, T.: Error Analysis of Digital Filters Realyzed with Floating-Point Arithmetic, Proc. IEEE, Vol. 57, No. 10, pp. 1735-1747 (1969).
3) Kan, E. P. and Aggarwal, J. K.: Error Analysis of Digital Filter Employing Floating-Point Arithmetic, IEEE, Vol. CT-18, No. 6, pp. 678-686 (1971).
4) Thong, T. and Liu, B.: Accumulation of Roundoff Errors in Floating-Point FFT, IEEE, Vol. CAS-24, No. 3, pp. 132-143 (1977).
5) Linnainma, S.: Towards Accurate Statistical Estimation of Rounding Errors in Floating-Point Computation, BIT, Vol. 15, pp. 165-173 (1975).
6) Goodman, R. and Feldstein, A.: Round-off Errors in Products, Computing, Vol. 15, pp. 263-275 (1975).
7) Kaneko, T. and Liu, B.: On Local Roundoff Errors in Floating-Point Arithmetic, J. ACM, Vol. 20, No. 3, pp. 391-398 (1973).
8) 小沢一文: 浮動小数点演算における相対丸め誤差の平均と分散について,情報処理学会論文誌, Vol. 21, No. 2, pp. 100-107 (1980).
9) Laha, R. G. and Rohatgi, V. K.: Probability Theory 4 pp. Wiley, New York (1979).
10) ibid.: p. 138.
11) Cramer, H.: Mathematical Methods of Statistics, 71 pp. Princeton Univ. Press, Princeton (1946).
12) 森口繁一他: 数学公式集II, 167 pp., 岩波書店, 東京 (1957).
13) 小沢一文: 浮動小数点方式の丸め誤差の新しい評価法I, II, 仙台電波高専研究紀要8号, pp. 25-44 (1978).
14) 清水良一: 中心極限定理, 232 pp., 教育出版, 東京 (1976).

付 録 式(29)の証明

定理3において,r の各点で $S_n(r)\to F(r)$, $(n\to\infty)$ となることが証明されているから,式(29)の後半の等式が成り立つためには,r^ν が $S_n(r)$ に関して一様可積分であることを証明すれば十分である[10]. 以下これを証明する.

定理1, 2より, $c>b^id$ なるときは,

$$\int_c^\infty |r^\nu|\,dF_i^-(r)=0,\quad \int_{-\infty}^{-c}|r^\nu|\,dF_i^+(r)=0$$

となる. したがって, $c>b^nd$ なる n に対して

$$\int_{|r|\geq 2c}|r^\nu|\,dS_n(r)=\sum_{i=-n}^{n}\left(\int_c^\infty |r^\nu|\,dF_i^-(r)+\int_{-\infty}^{-c}|r^\nu|\,dF_i^+(r)\right)=0 \tag{50}$$

が成り立つ.

また, $c\leq b^id$ なるときは,

$$\int_c^\infty |r^\nu|\,dF_i^-(r)=\int_c^{b^id}|r^\nu|\,dF_i^-(r)\leq (b^id)^\nu(F_i^-(b^id)-F_i^-(c))$$

$$\int_{-\infty}^{-c}|r^\nu|\,dF_i^+(r)=\int_{-b^id}^{-c}|r^\nu|\,dF_i^+(r)\leq (b^id)^\nu(F_i^+(-c)-F_i^+(-b^id))$$

が成り立つから, $b^nd>c$ なる n に対して

$$b^nd>b^{n-1}d>\cdots>b^{n_c}d\geq c>b^{n_c-1}d \tag{51}$$

を仮定すると,

$$\begin{aligned}\int_{|r|\geq c}|r^\nu|\,dS^n(r)&=\sum_{i=-n}^{n_c-1}\int_{|r|\geq c}|r^\nu|\,dF_i(r)+\sum_{i=n_c}^{n}\int_{|r|>c}|r^\nu|\,dF_i(r)\\&=\sum_{i=n_c}^{n}\int_{|r|\geq c}|r^\nu|\,dF_i(r)\\&\leq\sum_{i=n_c}^{n}(b^id)^\nu(F_i^-(b^id)-F_i^-(c)+F_i^+(-c)-F_i^+(-b^id))\\&\leq\sum_{i=n_c}^{n}(b^id)^\nu(F_i^-(b^id)+F_i^+(0))\end{aligned}$$

となる. ここで,定理1, 2の結果を用いると,

$$\int_{|r|\geq c}|r^\nu|\,dS_n(r)\leq\sum_{i=n_c}^{n}(b^id)^\nu(G(-b^{i-1})-G(-b^i)+G(b^i)-G(b^{i-1}))$$

$$\leq \sum_{i=n_c}^{n} (b^i d)^\nu (G(-b^{i-1})+1-G(b^{i-1})) \quad (52)$$

となる．仮定より，ある定数 $U>0$ に対して

$$G(-b^{i-1})<Ub^{-ki},\ 1-G(b^{i-1})<Ub^{-ki}$$

が成り立つから，$k-\nu>0$ なるとき

$$\int |r^\nu| dS_n(r) \leq 2 \sum_{i=n_c}^{n} (b^i d)^\nu U b^{-ki}$$

$$=2Ud^\nu \sum_{i=n_c}^{n} b^{-(k-\nu)i} < 2Ud^\nu b^{-(k-\nu)n_c}/(1-b^{-(k-\nu)}) \quad (53)$$

を得る．式 (51) より $b^{-n_c} \leq d/c$ が成立しているから，式 (53) は

$$\int_{|r|\geq c} |r^\nu| dS_n(r) < 2Ud^\nu (d/c)^{k-\nu}/(1-b^{-(k-\nu)}) \quad (54)$$

となる．式 (50)，(54) より任意の $\varepsilon>0$ に対して

$$\int_{|r|\geq c} |r^\nu| dS_n(r) < \varepsilon$$

を満たす c が n に無関係に存在する．すなわち，r^ν は $S_n(r)$ に関して一様に可積分である．

（昭和55年2月14日受付）
（昭和55年10月23日採録）