

実数値の浮動小数点表現における仮数の分布と下位桁の漸近的性質†

小 沢 一 文††

一般に、浮動小数点演算においては仮数の分布は必ずしも逆数分布であるとは限らず様々な分布を呈する。したがって、丸め誤差も様々な分布を呈する。

この論文では、任意の分布をもった実数値の b 進浮動小数点表現において仮数の第 n 桁の値が a (a は整数で、 $0 \leq a \leq b-1$) である確率を求め、その確率分布と離散的な一様分布との差を導出している。そして、仮数が微分可能な確率密度関数を持ち、その導関数が有界であるならば、下位桁の分布はほとんど一様分布に近いものになることを証明している。したがって、ほとんどの場合仮数部の丸め誤差は一様に分布しているとして良い。

この論文の結果は、A. Feldstein, R. Goodman 等が仮数分布が一様分布である場合に導出した結果の一般化でもあり、また浮動小数点演算における相対誤差分布の導出に役立つであろう。

1. はじめに

実数値を、整数 b ($b \geq 2$) を底とした浮動小数点方式によって正確に表現すれば、仮数部の b 進表示は無限桁になる。

しかし、実際に電子計算機により実数値を記憶するときは、この無限桁表現の下位の桁に適切な処理を施し、有限桁に切り詰めて記憶しなければならない。

有限桁に切り詰めることによって生ずる誤差の分布、すなわち丸め誤差のそれを知ることは、数値解析の分野では重要かつ困難な課題である。

丸め誤差分布を導出する際には、仮数の切り捨てられた下位の桁の分布が必要になるが、これについては A. Feldstein 等¹⁾は R. S. Pinkham²⁾, D. E. Knuth³⁾ 等の研究に基づいて、仮数分布が逆数分布になるとき、仮数は下位の桁へ行く程一様分布に近づくことを明らかにした*。さらに、仮数部における丸め誤差はほぼ一様分布になることも証明した**。

これは、これまでの逆数分布の丸め誤差に関する研究²⁾⁻⁵⁾の正当性をうらづけたものでもある。

しかし、実際に数値計算で扱うデータは様々な分布を呈する。それにもかかわらず、逆数分布以外の分布については、いま述べたようなことは何も明らかにされていない。

本論文では、逆数分布を含む広い範囲の分布について A. Feldstein 等¹⁾と同様の考察を行い、仮数における丸め誤差分布を求める。

2. 浮動小数点方式による実数の表現

実数 x の浮動小数点表現について述べる。実数 x は b 進浮動小数点方式によって

$$x = \text{sgn}(x) \cdot b^e \cdot \alpha \quad (1)$$

と表わされる。ここで、 $\text{sgn}(x)$ は符号を表わすものとする。すなわち

$$\text{sgn}(x) = \begin{cases} x/|x|, & x \neq 0 \\ 0, & x = 0. \end{cases} \quad (2)$$

また、指数 e は整数で仮数 α が区間 $[b^{-1}, 1)$ の中に入るような値を選ぶものとする。

ここで、仮数 α を b 進法で表現すると、

$$\alpha = d_1 b^{-1} + d_2 b^{-2} + \dots + d_n b^{-n} + \dots \quad (3)$$

となる。これを、

$$\alpha = (0. d_1 d_2 \dots d_n \dots)_b \quad (4)$$

と略記する。ここで、各 d_i は条件

$$\begin{aligned} 1 \leq d_1 \leq b-1, \\ 0 \leq d_i \leq b-1, \quad i=2, 3, \dots \end{aligned} \quad (5)$$

を満たす整数である。

つぎに、実数 x は確率変数とし、その仮数 α は区間 $[b^{-1}, 1)$ で有界でしかも連続微分可能な密度関数 $P_\alpha(\alpha)$ をもつものとする。

3. 各 d_i の分布

まず、(4)における各桁 d_i の同時確率について考察する。仮数 α の上位 k 桁の値がそれぞれ d_1, d_2, \dots, d_k である確率を $P(d_1, d_2, \dots, d_k)$ で表わす。この

† The Mantissa Distribution and the Asymptotic Property of Tail End Digits in Floating-point Representation of Real Numbers by KAZUFUMI OZAWA (Sendai Radio Technical College).

†† 仙台電波工業高等専門学校

* 文献 1) の Theorem 4. 参照.

** 文献 1) の Theorem 7. 参照.

とき,

$$P(d_1, d_2, \dots, d_k) = \int_{\rho}^{\rho+b^{-k}} P_a(\alpha) d\alpha \quad (6)$$

を得る。ここで,

$$\rho = d_1 b^{-1} + d_2 b^{-2} + \dots + d_k b^{-k} \quad (7)$$

とする。これは以下の通り証明される。

仮数 α の上位 k 桁が d_1, d_2, \dots, d_k であるならば, α は区間 $[(0. d_1 d_2 \dots d_k)_b, (0. d_1 d_2 \dots (d_k+1))_b)$ の中に存在するはずである。すなわち,

$$\rho \leq \alpha < \rho + b^{-k} \quad (8)$$

が成り立つ。この確率は(6)で表わされる。

この同時確率と密接な関係にある各 d_i の分布を求める。まず, 第 n 桁の値 d_n が a である確率を $P_{n,b}(a)$ で表わす。このとき, $P_{n,b}(a)$ は $P(d_1, d_2, \dots, d_n)$ の周辺分布として求められるが, 次のように表現しておくこともできる。

$$P_{1,b}(a) = \int_{ab^{-1}}^{(a+1)b^{-1}} P_a(\alpha) d\alpha, \quad (9)$$

$$a=1, 2, \dots, b-1.$$

$$P_{n,b}(a) = \sum_{i=1}^{I(n)} \int_{\varphi_i(a)}^{\varphi_i(a+1)} P_a(\alpha) d\alpha, \quad n \geq 2, \quad (10)$$

$$a=0, 1, \dots, b-1.$$

ここで,

$$I(n) = b^{n-2}(b-1) \quad (11)$$

$$\varphi_i(a) = b^{-1} + ((i-1)b+a)b^{-n}, \quad (12)$$

$$i=1, 2, \dots, I(n), \quad a=0, 1, 2, \dots, b-1.$$

とする。以下これを証明する。

$d_1 = a$ ならば, α は区間 $[(0. a)_b, (0. (a+1))_b)$ の中に存在するはずである。したがって(9)が成立する。

つぎに, $n \geq 2$ なる場合を考察する。まず仮数 α の上位 $n-1$ 桁で表現される値を γ と置く。すなわち,

$$\gamma = (0. d_1 d_2 \dots d_{n-1})_b \quad (13)$$

とする。条件(5)より, γ として全部で $(b-1) \times b^{n-2} (= I(n))$ 通りの値が考えられる。そこで, それらに小さい順に番号をつける。

$$\gamma_i = b^{-1} + (i-1)b^{-(n-1)}, \quad i=1, 2, \dots, I(n). \quad (14)$$

$\gamma = \gamma_i$ でしかも $d_n = a$ であるとき, α は区間 $[\gamma_i + ab^{-n}, \gamma_i + (a+1)b^{-n})$ の中に存在するはずである。

したがって, この確率 $P_r\{d_n = a, \gamma = \gamma_i\}$ は

$$\begin{aligned} P_r\{d_n = a, \gamma = \gamma_i\} &= \int_{\gamma_i + ab^{-n}}^{\gamma_i + (a+1)b^{-n}} P_a(\alpha) d\alpha \\ &= \int_{\varphi_i(a)}^{\varphi_i(a+1)} P_a(\alpha) d\alpha \end{aligned} \quad (15)$$

で表わされる。これを i について1から $I(n)$ まで加えたものが $P_{n,b}(a)$ であるから(10)が成り立つ。

そこで, 各 d_i が統計的に独立であると仮定すると,

$$P(d_1, d_2, \dots, d_k) = \prod_{n=1}^k P_{n,b}(d_n) \quad (16)$$

が成り立つことに注意しておく。

つぎに, いかなる条件の下で $P_{n,b}(a)$ が一様分布になるかを考察する。その前にいくつかの定義と定理を与える。

定義 1

$P_a(\alpha)$ と一様分布との偏差を $\Delta(\alpha)$ で表わす。すなわち,

$$\Delta(\alpha) \triangleq P_a(\alpha) - b/(b-1) \quad (17)$$

とする。

また, この上限を D で表わす。

$$D \triangleq \sup_{b^{-1} \leq \alpha < 1} |\Delta(\alpha)|. \quad (18)$$

つぎに, $P_{n,b}(a)$ の一様分布との偏差を定義する。

定義 2

$P_{n,b}(a)$ と離散的な一様分布との偏差を $\delta_{n,b}(a)$ で表わす。

$$\delta_{n,b}(a) \triangleq \begin{cases} P_{n,b}(a) - 1/(b-1), & n=1, \\ P_{n,b}(a) - 1/b, & n \geq 2. \end{cases} \quad (19)$$

以上の定義にもとづいて, いくつかの定理を証明する。

定理 1

$\delta_{n,b}(a)$ は,

$$\delta_{1,b}(a) = \int_{ab^{-1}}^{(a+1)b^{-1}} \Delta(\alpha) d\alpha, \quad (20)$$

$$\delta_{n,b}(a) = \sum_{i=1}^{I(n)} \int_{\varphi_i(a)}^{\varphi_i(a+1)} \Delta(\alpha) d\alpha, \quad n \geq 2. \quad (21)$$

で与えられる。さらに, $\delta_{n,b}(a)$ の絶対値について次式が成り立つ。

$$|\delta_{n,b}(a)| \leq D/b, \quad n=1, 2, \dots. \quad (22)$$

[証明]

まず, $n=1$ のときは(9)より

$$\begin{aligned} \delta_{1,b}(a) &= P_{1,b}(a) - 1/(b-1) \\ &= \int_{ab^{-1}}^{(a+1)b^{-1}} \{P_a(\alpha) - b/(b-1)\} d\alpha \\ &= \int_{ab^{-1}}^{(a+1)b^{-1}} \Delta(\alpha) d\alpha. \end{aligned}$$

したがって,

$$|\delta_{1,b}(a)| \leq \int_{ab^{-1}}^{(a+1)b^{-1}} |\Delta(\alpha)| d\alpha$$

$$\leq D/b.$$

つぎに, $n \geq 2$ のときは (10) より

$$\begin{aligned} \delta_{n,b}(a) &= P_{n,b}(a) - 1/b \\ &= \sum_{i=1}^{I(n)} \int_{\varphi_i(a)}^{\varphi_i(a+1)} P_a(\alpha) d\alpha - 1/b \\ &= \sum_{i=1}^{I(n)} \left\{ \int_{\varphi_i(a)}^{\varphi_i(a+1)} P_a(\alpha) d\alpha - b^{-n} \cdot b/(b-1) \right\} \\ &= \sum_{i=1}^{I(n)} \int_{\varphi_i(a)}^{\varphi_i(a+1)} \Delta(\alpha) d\alpha. \end{aligned}$$

したがって,

$$\begin{aligned} |\delta_{n,b}(b)| &\leq \sum_{i=1}^{I(n)} \int_{\varphi_i(a)}^{\varphi_i(a+1)} |\Delta(\alpha)| d\alpha \\ &\leq D \sum_{i=1}^{I(n)} b^{-n} \\ &= D(b-1)/b^2 \\ &< D/b. \end{aligned}$$

定理 2

すべての n について, d_n の分布 $P_{n,b}(a)$ が一様分布でしかも d_n が確率変数として互いに独立となるための必要十分条件は, $P_a(\alpha)$ が一様分布になることである. すなわち,

$$P_a(\alpha) = b/(b-1) \tag{23}$$

である.

[証明]

十分性; 式(23)が成り立てば $D=0$ である. したがって, 定理1より

$$\delta_{n,b}(a) = 0, \quad n=1, 2, \dots \tag{24}$$

すなわち, $P_{n,b}(a)$ は一様分布になる. このとき, $P(d_1, d_2, \dots, d_k)$ を計算すれば, 式(6)より

$$\begin{aligned} P(d_1, d_2, \dots, d_k) &= \int_{\rho}^{\rho+b^{-k}} b/(b-1) d\alpha \\ &= b^{-(k-1)}/(b-1) \end{aligned} \tag{25}$$

となる. また, 式(24)より

$$\begin{aligned} \prod_{n=1}^k P_{n,b}(d_n) &= (b-1)^{-1} \overbrace{b^{-1} b^{-1} \dots b^{-1}}^{k-1} \\ &= b^{-(k-1)}/(b-1) \end{aligned} \tag{26}$$

となる. したがって, 式(16)が成立する. すなわち, 各 $P_{n,b}(a)$ は互いに独立になる.

必要性; 仮数 α の分布が一様分布であるということは, α の分布関数を $F_a(\alpha)$ としたとき, 任意の実数 $y(b^{-1} \leq y < 1)$ に対して

$$F_a(y) = (b/(b-1))(y-1/b) \tag{27}$$

となることである. 以下, $P_{n,b}(a)$ が一様分布で互いに独立であると仮定し(27)が成り立つことを証明す

る.

まず, y を

$$\begin{aligned} y &= (0. d_1 d_2 \dots)_b \\ &= \sum_{n=1}^{\infty} d_n b^{-n} \end{aligned} \tag{28}$$

と表わす. この y に対して, $b^{-1} \leq \alpha < y$ を満たす実数 α の集合 $\{\alpha | b^{-1} \leq \alpha < y\}$ を, 以下に示す通りに分解する.

$$\begin{aligned} \{\alpha | (0. 1)_b \leq \alpha < y\} &= \{\alpha | (0. 1)_b \leq \alpha < (0. d_1)_b\} \\ &\cup \{\alpha | (0. d_1)_b \leq \alpha < (0. d_1 d_2)_b\} \\ &\cup \{\alpha | (0. d_1 d_2)_b \leq \alpha < (0. d_1 d_2 d_3)_b\} \\ &\vdots \\ &\cup \{\alpha | (0. d_1 d_2 \dots d_{n-1})_b \\ &\leq \alpha < (0. d_1 d_2 \dots d_{n-1} d_n)_b\} \\ &\vdots \\ &\cup \dots \end{aligned} \tag{29}$$

上に示した各事象は, 互いに排反であるので

$$\begin{aligned} P_r\{b^{-1} \leq \alpha < y\} &= P_r\{(0. 1)_b \leq \alpha < (0. d_1)_b\} \\ &+ P_r\{(0. d_1)_b \leq \alpha < (0. d_1 d_2)_b\} + \dots \\ &+ P_r\{(0. d_1 d_2 \dots d_{n-1})_b \\ &\leq \alpha < (0. d_1 d_2 \dots d_{n-1} d_n)_b\} \\ &+ \dots \end{aligned} \tag{30}$$

が成り立つ.

さらに, 式(30)の右辺第一項を分解すると,

$$\begin{aligned} P_r\{(0. 1)_b \leq \alpha < (0. d_1)_b\} &= P_r\{(0. 1)_b \leq \alpha < (0. 2)_b\} \\ &+ P_r\{(0. 2)_b \leq \alpha < (0. 3)_b\} \\ &+ P_r\{(0. (d_1-1))_b \leq \alpha < (0. d_1)_b\} \\ &= P_{1,b}(1) + P_{1,b}(2) + \dots + P_{1,b}(d_1-1) \end{aligned} \tag{31}$$

となり, 仮定より

$$\begin{aligned} P_r\{(0. 1)_b \leq \alpha < (0. d_1)_b\} &= \sum_{i=1}^{d_1-1} 1/(b-1) \\ &= (b/(b-1))(d_1-1)b^{-1}. \end{aligned} \tag{32}$$

つぎに, 式(30)第 $n(n \geq 2)$ 項は仮定より式(16)が成立するから,

$$\begin{aligned} P_r\{(0. d_1 d_2 \dots d_{n-1})_b \leq \alpha < (0. d_1 d_2 \dots d_{n-1} d_n)_b\} &= \sum_{i=0}^{d_n-1} P(d_1, d_2, \dots, d_{n-1}, i) \\ &= \prod_{j=1}^{n-1} P_{j,b}(d_j) \sum_{i=0}^{d_n-1} P_{n,b}(i) \\ &= (b-1)^{-1} \cdot b^{-1} \cdot b^{-1} \dots d_n b^{-1} \\ &= (b/(b-1)) d_n b^{-n}. \end{aligned} \tag{33}$$

ここで, (32), (33)を(30)に代入すると

$$\begin{aligned}
 P_r\{b^{-1} \leq \alpha < y\} \\
 &= (b/(b-1)(d_1-1)b^{-1} + (b/(b-1)) \sum_{n=2}^{\infty} d_n b^{-n}) \\
 &= (b/(b-1)) \left(\sum_{n=1}^{\infty} d_n b^{-n} - 1/b \right) \\
 &= (b/(b-1))(y-1/b). \quad (34)
 \end{aligned}$$

上式左辺は、 $F_\alpha(y)$ であるから (27) が得られた。 ■

3. $P_{n,s}(\alpha)$ の漸近的性質

つぎに、 $n \rightarrow \infty$ するときの $P_{n,s}(\alpha)$ の極限とその収束のオーダーについて考察する。文献 1) において、仮数分布が逆数分布であるならば、 $P_{n,s}(\alpha)$ は b^{-n} のオーダーで一様分布に収束することが明らかにされた。すなわち、

$$P_\alpha(\alpha) = \frac{1}{\alpha \ln b}, \quad 1/b \leq \alpha < 1, \quad (35)$$

のとき

$$P_{n,s}(\alpha) = 1/b + O(b^{-n}), \quad (n \rightarrow \infty). \quad (36)$$

ここでは、逆数分布に限らず、 $P_\alpha(\alpha)$ の微係数 $P_\alpha'(\alpha)$ が区間 $[b^{-1}, 1)$ で有界であるならば、(36) が成立することを証明する。

定理 3

区間 $[b^{-1}, 1)$ において、 $P_\alpha'(\alpha)$ が有界であれば (36) が成り立つ。

(証明)

ここでは、 $n \geq 2$ の場合についてだけ考察する。 $\delta_{n,s}(\alpha)$ は、 a が区間 $[0, b-1]$ の整数値であるときだけ意味のある関数であるが、その定義域を区間 $[0, b-1]$ の実数値に式 (21) そのままの形で拡張する。そうすると $\Delta(\alpha)$ の連続性より、 $\delta_{n,s}(\alpha)$ は a について連続な関数になる。しかも、

$$\begin{aligned}
 \sum_{\alpha=0}^{b-1} \delta_{n,s}(\alpha) &= \sum_{\alpha=0}^{b-1} (P_{n,s}(\alpha) - 1/b) \\
 &= 0 \quad (37)
 \end{aligned}$$

が成立するから、 $\delta_{n,s}(\alpha) = 0$ となる a が区間 $[0, b-1]$ の中に必ず 1 つは存在するはずである。それを a^* とすれば、(21) より

$$\begin{aligned}
 \delta_{n,s}(\alpha) &= \delta_{n,s}(\alpha) - \delta_{n,s}(\alpha^*) \\
 &= \sum_{i=1}^{I(n)} \left(\int_{\varphi_i(\alpha)}^{\varphi_i(\alpha+1)} \Delta(\alpha) d\alpha \right. \\
 &\quad \left. - \int_{\varphi_i(\alpha^*)}^{\varphi_i(\alpha^*+1)} \Delta(\alpha) d\alpha \right). \quad (38)
 \end{aligned}$$

上式右辺に第一平均値の定理⁹⁾ を適用すると、

$$\begin{aligned}
 \delta_{n,s}(\alpha) &= \sum_{i=1}^{I(n)} [(\varphi_i(\alpha+1) - \varphi_i(\alpha)) \Delta(\xi_i) \\
 &\quad - (\varphi_i(\alpha^*+1) - \varphi_i(\alpha^*)) \Delta(\xi_i^*)] \\
 &= b^{-n} \sum_{i=1}^{I(n)} (\Delta(\xi_i) - \Delta(\xi_i^*)). \quad (39)
 \end{aligned}$$

ただし、 ξ_i, ξ_i^* は

$$\varphi_i(\alpha) < \xi_i < \varphi_i(\alpha+1), \quad \varphi_i(\alpha^*) < \xi_i^* < \varphi_i(\alpha^*+1)$$

を満たす数である。

つぎに、(39) 右辺に平均値の定理を適用すると

$$\delta_{n,s}(\alpha) = b^{-n} \sum_{i=1}^{I(n)} \Delta'(\eta) (\xi_i - \xi_i^*). \quad (40)$$

ただし、 η は ξ_i と ξ_i^* の間の数である。仮定より $\Delta'(\alpha)$ は有界であるから、区間 $[b^{-1}, 1)$ におけるその上限を U と置けば、すなわち

$$U \triangleq \sup_{b^{-1} \leq \alpha < 1} |\Delta'(\alpha)| \quad (41)$$

とすれば、

$$\begin{aligned}
 |\delta_{n,s}(\alpha)| &\leq b^{-n} \sum_{i=1}^{I(n)} |\Delta'(\eta)| |\xi_i - \xi_i^*| \\
 &\leq U b^{-n} \sum_{i=1}^{I(n)} |\xi_i - \xi_i^*|
 \end{aligned}$$

を得る。したがって、

$$\begin{aligned}
 \max_{0 \leq \alpha \leq b-1} |\delta_{n,s}(\alpha)| &< U b^{-n} \sum_{i=1}^{I(n)} (\varphi_i(b) - \varphi_i(0)) \\
 &= U b^{-n} \sum_{i=1}^{I(n)} b^{-(n-1)} \\
 &< U b^{-n} \quad (42)
 \end{aligned}$$

となる。ここで、 U は n に無関係な定数であるから $\delta_{n,s}(\alpha) = O(b^{-n})$, ($n \rightarrow \infty$) となる。したがって、(36) が得られる。 ■

4. 例題および検討

4.1 $P_\alpha(\alpha)$ が逆数分布とは異なるとき

例題 1. まず、仮数分布が一様分布になるような 2 つの独立な確率変数の間に乗算をほどこすと、その演算結果の仮数分布は

$$P_\alpha(\alpha) = \{\ln b - (b-1) \ln \alpha\} b / (b-1)^2 \quad (43)$$

となることが知られている⁹⁾。

明らかに、この $P_\alpha(\alpha)$ は定理 3 の仮定を満たしている。ここでは、この $P_\alpha(\alpha)$ を用い $b=2$ として (9), (10) より $P_{n,s}(\alpha)$ を計算し、その計算結果から最大偏差 $\max_a |\delta_{n,s}(\alpha)|$ (以下これを $D_{n,s}$ で表わす) を求め図 1 に示す。

また、区間 (0, 1) の一様乱数* を用いて電子計算機

* 区間 (0, 1) における一様乱数は、その仮数分布が一様分布になる。(付録参照)

でシミュレーションを行い、その結果も同時に図1に示す。

例題 2. つぎに、例題1と同様の条件をもつ2つの数の間に除算をほどこすと、その演算結果の仮数分布 $P_a(\alpha)$ は

$$P_a(\alpha) = (b+1/\alpha^2)/2(b-1) \quad (44)$$

となることが知られている⁹⁾。

この場合も、 $P_a(\alpha)$ は定理3の仮定を満たしている。この $P_a(\alpha)$ を用いて、例題1と同じ量を計算しそれとシミュレーション結果とを図2に示す。

ここで、2つのシミュレーションは一様乱数を30万

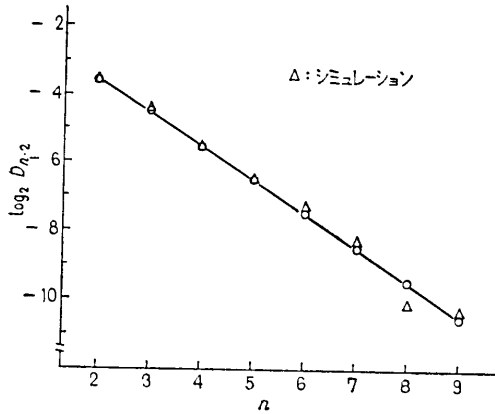


図1 仮数の確率密度関数が式(43)で記述される場合の n と $D_{n,b}$ の関係 ($b=2$).

Fig. 1 The relation between n and the maximum deviation $D_{n,b} (= \max_{0 \leq a \leq b-1} |\delta_{n,b}(a)|)$ for $b=2$ when the probability density function of mantissa is described by eq. (43).

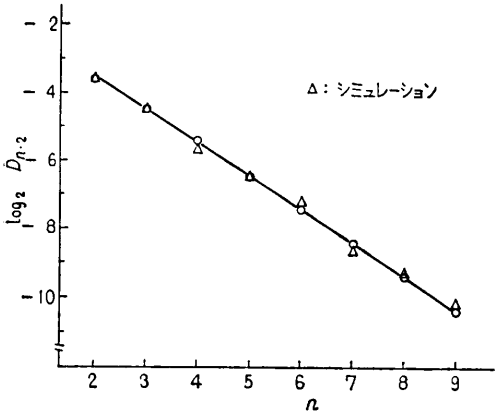


図2 仮数の確率密度関数が式(44)で記述される場合の n と $D_{n,b}$ の関係 ($b=2$).

Fig. 2 The relation between n and the maximum deviation $D_{n,b} (= \max_{0 \leq a \leq b-1} |\delta_{n,b}(a)|)$ for $b=2$ when the probability density function of mantissa is described by eq. (44).

回発生させて行った。

これ等の図では、横軸は n を表わし縦軸は最大偏差 $D_{n,2}$ を2を底とする対数の目盛で表わしている。これ等の図からわかるように、理論値において $\log_2 D_{n,2}$ は n の増加と共に直線的に減少している。これは、 $P_{n,2}(a)$ が 2^{-n} のオーダーで一様分布に近づいていることを意味している。また、この場合は理論値とシミュレーション結果はよく一致している。

さらに、参考までに $b=10, 16$ の場合理論値 $\log_b D_{n,b}$ を求め図3~6に示す。

4.2 逆数分布との比較

表1 仮数の確率密度関数がそれぞれ(43), (35), (44)で与えられる場合の $D_{n,b}$ の値 ($b=2$).

Table 1 The values of $D_{n,b}$ for $b=2$ when the probability density function of mantissa is described by eq. (43), (35) and (44), respectively.

n	$D_{n,b}$		
	eq. (43)	eq. (35)	eq. (44)
2	0.650E-1	0.850E-1	0.833E-1
3	0.431E-1	0.443E-1	0.452E-1
4	0.216E-1	0.224E-1	0.232E-1
5	0.108E-1	0.113E-1	0.117E-1
6	0.542E-2	0.563E-2	0.586E-2
7	0.271E-2	0.282E-2	0.293E-2
8	0.135E-2	0.141E-2	0.147E-2
9	0.677E-3	0.704E-3	0.732E-3
10	0.339E-3	0.352E-3	0.366E-3

表2 仮数の確率密度関数がそれぞれ(43), (35), (44)で与えられる場合の $D_{n,b}$ の値 ($b=10$).

Table 2 The values of $D_{n,b}$ for $b=10$ when the probability density function of mantissa is described by eq. (43), (35) and (44), respectively.

n	$D_{n,b}$		
	eq. (43)	eq. (35)	eq. (44)
2	0.121E-1	0.197E-1	0.294E-1
3	0.116E-2	0.178E-2	0.254E-2
4	0.115E-3	0.176E-3	0.248E-3
5	0.115E-4	0.176E-4	0.248E-4

表3 仮数の確率密度関数がそれぞれ(43), (35), (44)で与えられる場合の $D_{n,b}$ の値 ($b=16$).

Table 3 The values of $D_{n,b}$ for $b=16$ when the probability density function of mantissa is described by eq. (43), (35) and (44), respectively.

n	$D_{n,b}$		
	eq. (43)	eq. (35)	eq. (44)
2	0.566E-2	0.112E-1	0.191E-1
3	0.340E-3	0.625E-3	0.990E-3
4	0.212E-4	0.387E-4	0.609E-4

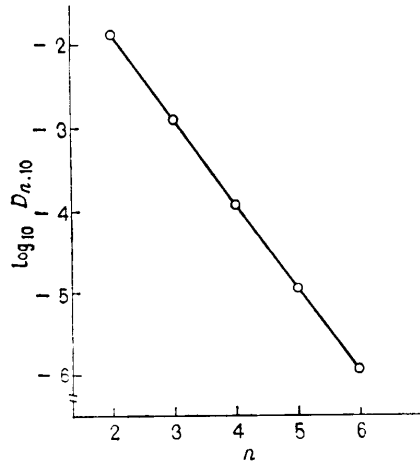


図 3 仮数の確率密度関数が式(43)で記述される場合の n と $D_{n,b}$ の関係 ($b=10$).

Fig. 3 The relation between n and the maximum deviation $D_{n,b} (= \max_{0 \leq a \leq b-1} |\delta_{n,b}(a)|)$ for $b=10$ when the probability density function of mantissa is described by eq. (43).

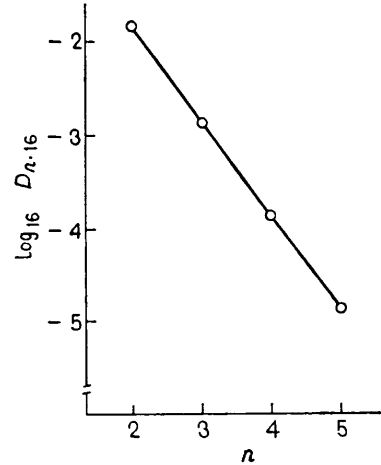


図 5 仮数の確率密度関数が式(43)で記述される場合の n と $D_{n,b}$ の関係 ($b=16$).

Fig. 5 The relation between n and the maximum deviation $D_{n,b} (= \max_{0 \leq a \leq b-1} |\delta_{n,b}(a)|)$ for $b=16$ when the probability density function of mantissa is described by eq. (43).

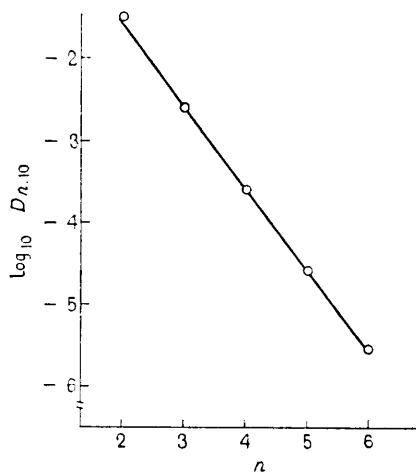


図 4 仮数の確率密度関数が式(44)で記述される場合の n と $D_{n,b}$ の関係 ($b=10$).

Fig. 4 The relation between n and the maximum deviation $D_{n,b} (= \max_{0 \leq a \leq b-1} |\delta_{n,b}(a)|)$ for $b=10$ when the probability density function of mantissa is described by eq. (44).

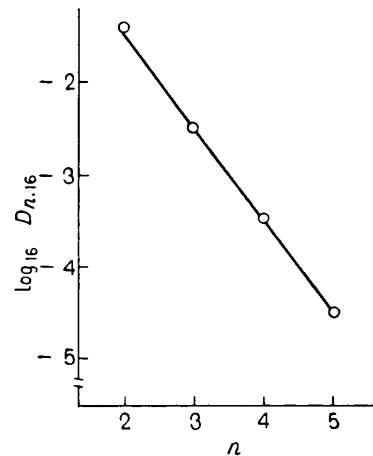


図 6 仮数の確率密度関数が式(44)で記述される場合の n と $D_{n,b}$ ($b=16$).

Fig. 6 The relation between n and the maximum deviation $D_{n,b} (= \max_{0 \leq a \leq b-1} |\delta_{n,b}(a)|)$ for $b=16$ when the probability density function of mantissa is described by eq. (44).

つぎに、仮数分布が式(43)、(44)で与えられる場合と A. Feldstein 等¹⁾によって与えられた逆数分布、すなわち式(35)で与えられる場合の三通りの $D_{n,b}$ を比較する。表1~3に三通りの $D_{n,b}$ を示す。

これ等の表からわかるように、最大偏差は3つの結果ともに b^{-n} のオーダーで0に収束しているが、その値はそれぞれ異なる。これは、 $U (= \sup_{b^{-1} \leq \alpha < 1} |A'(\alpha)|)$ の値によると思われる。何故ならば、定理3より $\delta_{n,b}(a)$

の上界は Ub^{-n} であるからである。実際、 $b=2$ のとき U の値を計算して大小関係を調べると (他の b でも大小関係は不変である)。

例題 1:

$$U = \sup_{b^{-1} \leq \alpha < 1} \{b/(b-1)\alpha\} = b^2/(b-1) = 4.00$$

例題 2:

$$U = \sup_{b^{-1} \leq \alpha < 1} \{1/(b-1)\alpha^3\} = b^3/(b-1) = 8.00$$

逆数分布:

$$U = \sup_{b^{-1} \leq \alpha < 1} \{1/\alpha^2 \ln b\} = b^2/\ln b = 5.77$$

であり、 U の大きい分布では最大偏差も大きいことがわかる。

5. 丸め誤差分布導出への応用

定理2より、 $P_{n,s}(a)$ がすべての n について一様分布になるのは、非常に限られた場合であることが明らかになった。しかし、また定理3より $P_a(\alpha)$ が微分可能で微係数が区間 $[b^{-1}, 1)$ で有界であれば、 n が大きくなるにつれて $P_{n,s}(a)$ が一様分布に近づくことも明らかになった。

したがって、この結果より仮数 α の上位 n 桁による有限桁近きを $\bar{\alpha}$ とし、仮数における丸め誤差を $\beta (= \bar{\alpha} - \alpha)$ とすれば、各桁が独立である場合以下に示す近似が成り立つ。まず $\bar{\alpha}$ については、

$$\begin{aligned} P_r\{\bar{\alpha} = (0.d_1d_2\dots d_n)_b\} \\ &= \prod_{i=1}^n P_{i,s}(d_i) \\ &\approx (b-1)^{-1} \cdot b^{-1} \dots b^{-1} \\ &= (b-1)^{-1} \cdot b^{-(n-1)}. \end{aligned} \quad (45)$$

つぎに、 β は連続量となりその密度関数 $P_\beta(\beta)$ は定理2の必要性の証明と同様の考察を行うことによって、ほぼ一様分布になることがわかる。すなわち、

$$P_\beta(\beta) \approx b^n, \quad 0 \leq \beta < b^{-n}. \quad (46)$$

これらの近似は、一様分布への収束のオーダーを考慮すると、 b が大きいときほど有効であることがわかる。

上に示した2式(45)、(46)から相対誤差分布を導出することも可能である。文献10)ではこれらを用い相対誤差分布を導出した。

6. おわりに

本論文では、実数値確率変数を b 進浮動小数点数を用いて表現したとき、その仮数の b 進表示の持っている基本的性質について考察を行った。その結果、つぎのようなことが明らかになった。

(i) 仮数の b 進表示における各桁が互いに独立かつ一様分布になるのは、 $P_a(\alpha)$ が一様分布になるとき、そのときに限る。

(ii) $P_a'(\alpha)$ が有界であれば、 $n \rightarrow \infty$ なるとき $P_{n,s}(a)$ は b^{-n} のオーダーで一様分布に近づく。したがって、(45)、(46)の近似が非常に有効である。

本論文の結果は、非常に一般的な仮定にもとづいて導出されたものでその適用範囲は広いものと思われる。

また、ここで得られた定理はすべて無限の精度をもった数値について成り立つものであるが、有限の精度しかない数値についても近似的に成り立つ。これは、シミュレーション結果が保証している。

なお、ここで用いた計算機は理論値の計算は本校電子計算機FACOM 230-28であり、シミュレーションの方は東北大学大型計算センターACOS-6をTSS端末より使用した。

最後に、日頃御指導いただいている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) Feldstein, A. and Goodman, R.: Convergence Estimates for the Distribution of Trailing Digits, J. ACM, Vol. 23, No. 2, pp. 287-296 (1976).
- 2) Pinkham, R. S.: On the Distribution of First Significant Digits, Ann. Math. Statist., Vol. 32, pp. 1223-1230 (1961).
- 3) Knuth, D. E.: The Art of Computer Programming, Vol. 2: Seminumerical Algorithm, Addison-Wesely, 218-229 (1969).
- 4) Tsao, N. K.: On the Distribution of First Significant Digits and Roundoff Errors, Comm. ACM, Vol. 17, No. 5, pp. 269-271 (1974).
- 5) Kaneko, T. and Liu, B.: On Local Roundoff Errors in Floating-point Arithmetic, J. ACM, Vol. 20, No. 3, pp. 391-398 (1973).
- 6) Hamming, R. W.: On the Distribution of Numbers, Bell Syst. Tech. J. Vol. 49, No. 8, pp. 1609-1625 (1970).
- 7) Sterbentz, P. H.: Floating-point Computation, Prentice-Hall, Inc., Englewood Cliffs, N. J. (1973).
- 8) 小平邦彦: 解析入門II, 岩波講座 基礎数学, 岩波書店, pp. 160 (1976).
- 9) Hamming, R. W.: Numerical Methods for Scientists and Engineers, McGraw-Hill, Inc., pp. 27-40 (1973).
- 10) 小沢一文: 有限桁浮動小数点近似による丸め誤差の統計的性質について, 情報処理学会投稿中.

付 録

区間 $(0, 1)$ の一様乱数の仮数は、区間 $[b^{-1}, 1)$ で一様に分布している。

〔証明〕

確率変数 x の分布関数 $F_x(x)$ は

$$F_x(x) = x, \quad (0 < x < 1)$$

で与えられるとする。

このとき、 x の仮数 α の分布関数 $F_a(\alpha)$ は

$$\begin{aligned} F_\alpha(\alpha) &= \sum_{m=0}^{\infty} \{F_x(b^{-m}\alpha) - F_x(b^{-m-1})\} \\ &= \sum_{m=0}^{\infty} b^{-m}(\alpha - b^{-1}) \\ &= (b/(b-1))(\alpha - 1/b) \end{aligned}$$

となる。また、密度関数 $P_\alpha(\alpha)$ は

$$\begin{aligned} P_\alpha(\alpha) &= dF_\alpha(\alpha)/d\alpha \\ &= b/(b-1) \end{aligned}$$

となる。したがって、一様分布になる。

(昭和53年9月18日受付)

(昭和54年7月19日採録)
