

浮動小数点数の加減算における丸め誤差の分布について¹

久保田 光一

中央大学理工学部

浮動小数点数を用いた計算では、丸め誤差の発生は避けられないで、丸め誤差解析は重要である。個々の演算で発生する丸め誤差、すなわち、発生誤差の分布はある限界内の一様分布に従うと仮定し、丸め誤差の振舞いを統計的に捉える丸め誤差の評価方法が確立されている。実際には、発生誤差の分布は、加減算においては一様分布とは異なる少々特徴的なものになる。本稿では、2進0捨1入の浮動小数点数体系について、加減算における発生誤差の分布形について解析し、簡単なモデルを構成した。数値実験を行ない、解析の妥当性を確認した。

On distribution of rounding errors generated in additions and subtractions of floating-point numbers

Koichi KUBOTA

Faculty of Science and Engineering, Chuo University

Analyzing the behaviour of rounding errors generated in additions and subtractions of floating-point numbers and regarding them as random variables, we propose a slightly strange model for their distribution. It gives distributions which approximate the observed generated relative errors more precisely than the unit distribution which is usually adopted in probabilistic analysis of rounding errors does.

¹日本応用数理学会計算の品質研究部会 / 情報処理学会 HPC 研究会共催

1 はじめに

浮動小数点数を用いた計算では丸め誤差の発生は避けられないことから、丸め誤差解析は、打ち切り誤差の解析とならんで、重要である。そこでは、丸め誤差の累積により計算値が真値からずれるそのずれの大きさを解析する ([3],[4],[5])。

同じ値を入力して同じ計算機で実行すれば同一の結果を得るという点で、丸め誤差自身は、確率的な現象ではない。しかし、入力値に擾動を与えたときの計算結果の変動などはある程度確率的な事象として捉えることができ、実際、そのような立場で丸め誤差の統計的な処理を行ない、誤差評価の方法として成功を収めている ([3],[4],[5])。そのような場合に、個々の演算実行の際に発生する誤差（これを発生誤差という）は丸め誤差限界内の一様分布に従う確率変数であると仮定するのが普通である。入力値の確率モデルが特に与えられない限り、一様分布と仮定することは妥当であり、実際に誤差を評価するのにも支障はない。

しかし、実際には加減算に関してはもう少し精密な議論ができる、一様分布とは異なる少々変わった分布として特徴づけられる。その分布は、一見すると、Bernoulli 分布や Cantor 分布 ([1]) に似ている。ここでは、この加減算における相対的な発生誤差の分布を解析し、いくつかのプログラムについて数値実験を行ない、その解析が妥当であることを示す。

2 発生誤差

以下では、丸めの方式として 0 捨 1 入 (Round to nearest) を指定した IEEE の浮動小数点数体系の下での計算を考える（四則演算結果は誤差のない真の結果を丸めたものになることが規定されている）[2]。丸めのちょうど境界上の値は、丸めた結果の最下位ビットが 0 になるように丸められる（round to even）。仮数部のビット長を隠れビットも含めて m と表す。単精度の場合は $m = 24$ であり、倍精度の場合は $m = 53$ である。この体系では、いわゆるマシンエプソン ε は、 $\varepsilon = 2^{-m}$ と表される。他の浮動小数点数体系でも同様の解析が可能なはずであるが、それについては別の機会に譲る。

ここでの議論は、個々の演算において発生する丸めの誤差の相対的な大きさの分布である。計算の途中で

$$w = \psi(u, v) \quad (2.1)$$

という形の演算を実行するとき、計算の誤差の存在を強調するために、これを

$$\bar{w} = \bar{\psi}(\bar{u}, \bar{v}) \quad (2.2)$$

と表す。上線をつけて誤差を含んだ値であることを明示する。

発生誤差とは、被演算数の誤差とは関わりなく、それらは正確であるとして演算を実行し、結果を丸めたときの「誤差」であり、次式の δ である：

$$\bar{w} = \bar{\psi}(\bar{u}, \bar{v}) = \psi(\bar{u}, \bar{v}) + \delta. \quad (2.3)$$

最右辺の ψ の引数は \bar{u}, \bar{v} であり、 u, v ではないこと、および、一般には誤差のために $\bar{u} \neq u, \bar{v} \neq v$ であることに注意されたい。

発生誤差は、累積誤差、すなわち、最終的な計算値 \bar{f} と、誤差のない無限精度演算の結果であるところの真の値 f との差 $\bar{f} - f$ を近似するときに用いられる ([3],[4]):²

$$\bar{f} - f \approx \sum_w \frac{\partial f}{\partial w} \delta_w. \quad (2.4)$$

ただし、右辺の総和は \bar{f} を計算する過程でのすべての途中結果 \bar{w} についてのものであり、微分係数 $\partial f / \partial w$ は f を計算するための途中計算結果 w の擾動に関する関数値の「感度」を表し、 δ_w は w を計算したときの発生誤差を表す。この δ_w を確率変数とみなして、式 (2.4) を評価しようというのが、丸め誤差の解析法の一つである ([3],[4])。

²丸め誤差の議論なので、ここでは入力の誤差は無いとする。

演算結果 \bar{w} が得られたときの発生誤差 $\delta \equiv \delta_w$ は、本来未知量であるが、その大きさの限界は定められる。 \bar{w} と同一の指數部を持つ正の浮動小数点数の最小のものを

$$M(\bar{w}) = 2^{\lfloor \log_2 |\bar{w}| \rfloor} \quad (2.5)$$

と記すことにすれば、 δ の限界は、

$$-M(\bar{w})\varepsilon \leq \delta \leq M(\bar{w})\varepsilon \quad (2.6)$$

と表される ([3],[4])。 δ の相対的な大きさ η を

$$\eta = \frac{\delta}{M(\bar{w})} = \frac{\delta}{2^{\lfloor \log_2 |\bar{w}| \rfloor}} \quad (2.7)$$

とおけば、 η の限界はマシンエプシロン ε そのものになる：

$$-\varepsilon \leq \eta \leq \varepsilon. \quad (2.8)$$

従来は、この η を区間 $[-\varepsilon, \varepsilon]$ 上の一様分布 $U[-\varepsilon, \varepsilon]$ に従う確率変数とみなして、統計的解析を行なった。

実際のところは、 η は一様分布に従うと考えても構わないことは、従来の解析が成功していることからも推察できる。しかも、乗除算については、 η が一様分布に従うという仮定はほとんど不都合はない。しかし、加減算については、 η は一様分布というよりは、少々変った分布に従うのである。

3 加減算の発生誤差

加(減)算 $z = x + y$ の発生誤差が被演算数 x と y の指數部がどれだけ異なるかということに依存することは自明である。まず、指數の差を固定したときの加減算の発生誤差を調べ、つぎに、指數の差の分布の単純なモデルの下で、その発生誤差の相対的な大きさの分布を解析する。

3.1 指數の差と加減算の発生誤差

以下では、一般性を失わずに、 x, y を正の浮動小数点数とし、 x の指數部 $M(x) = 2^{\lfloor \log_2 |x| \rfloor}$ の方が y の指數部 $M(y) = 2^{\lfloor \log_2 |y| \rfloor}$ よりも小さくないことを仮定して説明する。また、アンダーフロー、オーバーフローは起きないとする。 $x + y$ と $x - y$ とを分離して考え、あとでまとめる。 x と y の指數の差とは、

$$k = \lfloor \log_2 |x| \rfloor - \lfloor \log_2 |y| \rfloor \geq 0 \quad (3.1)$$

のことである。これは、 x の小数点に合わせて y を桁移動するときに、 y の下位 k ビットが x の最下位ビットより右にくることを表す。また、 k を固定したときに発生する誤差の相対的な大きさは、 x と y を $1/M(x)$ 倍して、 x が $[1, 2)$ の範囲に、 y が $[2^{-k}, 2^{-k+1})$ の範囲に入るよう大きさを揃えたとき(スケーリングしたとき)の発生誤差の相対値で表されることに注意する。したがって、 x が $[1, 2)$ 上のすべての浮動小数点数を動き、 y が $[2^{-k}, 2^{-k+1})$ 上のすべての浮動小数点数を動くときの $x + y$ および $x - y$ の発生誤差を計算すればよいことになる。

命題 1：指數が k ($\leq m-2$) だけ異なる 2 個の正の数を x と y とする。 $x + y$ の相対発生誤差は $\delta = 2^{-m-k}$ を単位として $2^{k+1} + 1$ 種類ある：

$$j \cdot 2^{-m-k} \quad (j = 0, \pm 1, \pm 2, \dots, \pm 2^k). \quad (3.2)$$

また、 $x - y$ の相対発生誤差は $\delta = 2^{-m-k+1}$ を単位として $2^k + 1$ 種類ある：

$$i \cdot 2^{-m-k+1} \quad (i = 0, \pm 1, \pm 2, \dots, \pm 2^{k-1}). \quad (3.3)$$

(アンダーフロー、オーバーフローは起こらないと仮定したことに注意。)

例: $m = 24$, $k = 1$ のとき, 正の数の加算の際に観察される相対発生誤差は $\delta = 2^{-24-1} = 2^{-25}$ として,

$$-2\delta, -\delta, 0, \delta, 2\delta \quad (3.4)$$

の 5 種類である.

正の数と負の数の加算の際には,

$$-2\delta, 0, 2\delta \quad (3.5)$$

の 3 種類である. ■

x と y の指数の差を $k (\leq m-2)$ に固定したとき, x の仮数部のビットパターンと y の仮数部のビットパターンは起こり得るすべての組合せを同等に扱ってもよいはずである. 絶対値と関係なく, 相対的な大きさの違いを考えているからである. そこで, 次では, 加算の被演算数の x, y はある区間上の一様分布に従う互いに独立な確率変数と仮定する.

命題 2: x と y の指数が $k (\leq m-2)$ だけ異なるとする. すなわち, x を区間 $[1, 2]$ 上の一様分布に従う確率変数, y を区間 $[2^{-k}, 2^{-k+1}]$ 上の一様分布に従う x と独立な確率変数とする. これらの和 $z = x + y$ を丸めるとときに, 相対発生誤差が $j \cdot 2^{-m-k}$ である確率 $f_k^+(j \cdot 2^{-m-k})$ は, $k = 0$ のとき, $f_0^+(\pm 2^{-m}) = \frac{1}{4}$, $f_0^+(0) = \frac{1}{2}$ であり, $k \geq 1$ のとき,

$$f_k^+(j \cdot 2^{-m-k}) = \begin{cases} \frac{1}{2^{k+1}} \frac{3}{2^{k+1}}, & j(|j| < 2^k) \text{ は奇数;} \\ \frac{1}{2^k} \left(1 - \frac{3}{2^{k+1}}\right) + \frac{1}{2^{k+1}} \frac{3}{2^{k+1}}, & j(|j| < 2^k) \text{ は偶数;} \\ \frac{1}{2^{k+1}} \left(1 - \frac{3}{2^{k+1}}\right) + \frac{1}{2^{k+2}} \frac{3}{2^{k+1}}, & j = 2^k, -2^k; \end{cases} \quad (3.6)$$

である. ■

命題 3: x と y の指数が $k (\leq m-2)$ だけ異なるとする. すなわち, x を区間 $[1, 2]$ 上の一様分布に従う確率変数, y を区間 $[2^{-k}, 2^{-k+1}]$ 上の一様分布に従う x と独立な確率変数とする. これらの差 $z = x - y$ を丸めるとときに, 相対発生誤差が $j \cdot 2^{-m-k+1}$ である確率 $f_k^-(j \cdot 2^{-m-k+1})$ は, $k = 0$ のとき, $f_0^-(0) = 1$, $k = 1$ のとき $f_1^-(\pm 2^{-m}) = \frac{1}{16}$, $f_1^-(0) = \frac{7}{8}$ であり, $k \geq 2$ のとき,

$$f_k^-(j \cdot 2^{-m-k+1}) = \begin{cases} \frac{1}{2^k} \left(1 - \frac{3}{2^{k+1}}\right), & j(|j| < 2^{k-1}) \text{ は奇数;} \\ \frac{1}{2^{k-1}} \frac{3}{2^{k+1}} + \frac{1}{2^k} \left(1 - \frac{3}{2^{k+1}}\right), & j(|j| < 2^{k-1}) \text{ は偶数;} \\ \frac{1}{2^k} \frac{3}{2^{k+1}} + \frac{1}{2^{k+1}} \left(1 - \frac{3}{2^{k+1}}\right), & j = \pm 2^{k-1}; \end{cases} \quad (3.7)$$

である. ■

以上の命題 2, 命題 3 を二つまとめて, 次を得る.

命題 4: x と y の指数が $k (\leq m-2)$ だけ異なるとする. すなわち, x はその絶対値 $|x|$ が区間 $[1, 2]$ 上の一様分布(正負とも等確率であるとする)に従う確率変数, y はその絶対値 $|y|$ が区間 $[2^{-k}, 2^{-k+1}]$ 上の一様分布(正負とも等確率であるとする)に従う x と独立な確率変数とする. これらの和 $z = x + y$ を丸めるとときに, 相対発生誤差が $j \cdot 2^{-m-k+1}$ である確率 $f_k(j \cdot 2^{-m-k})$ は,

$$f_k(j \cdot 2^{-m-k}) = \frac{1}{2} (f_k^+(j \cdot 2^{-m-k}) + f_k^-(j \cdot 2^{-m-k}))$$

と表される. したがって, $k = 0$ のとき, $f_0(0) = \frac{1}{2}(\frac{1}{2} + 1) = 3/4$, $f_0(\pm 2^{-m}) = \frac{1}{2}(\frac{1}{4} + 0) = 1/8$ となる. $k = 1$ のときは, $f_1(\pm 2^{-m}) = \frac{1}{2}(\frac{5}{32} + \frac{1}{16}) = \frac{7}{64}$, $f_1(0) = \frac{1}{2}(\frac{5}{16} + \frac{7}{8}) = \frac{19}{32}$, $f_1(\pm 2^{-m-1}) = \frac{1}{2}(\frac{3}{16} + 0) = \frac{3}{32}$ で

ある。 $k \geq 2$ のときは,

$$f_k(j \cdot 2^{-m-k}) = \begin{cases} \frac{3}{8} \frac{1}{2^{2k}}, & j(|j| < 2^k) \text{ は奇数;} \\ \frac{1}{2^k} - \frac{9}{8} \frac{1}{2^{2k}}, & j = 2j' (|j'| < 2^{k-2}) j' \text{ は奇数;} \\ \frac{1}{2^k} + \frac{3}{8} \frac{1}{2^{2k}}, & j = 2j' (|j'| < 2^{k-2}) j' \text{ は偶数;} \\ \frac{1}{2^{k+1}} + \frac{3}{16} \frac{1}{2^{2k}}, & j = \pm 2^k; \end{cases} \quad (3.8)$$

である。 ■

3.2 相対発生誤差の単純モデル

計算途中で加算の被演算数として現れる数の絶対的な大きさの違いは、入力値、計算の過程に大きく左右される。一方、指數の差 k を基準にしてその相対的な大きさの違いに着目すれば、相対発生誤差に関する次のようなモデルを考えることができる。

x に比較して y が小さ過ぎて激しい情報落ちをおこす場合を除くことにして、 $k \leq m-2$ の場合を考える。そうすると、指數が k 異なる 2 数 (y の下位 k ビットが x の最下位ビットの右にくる場合) の加算が、加算の総数に占める割合を $P(k)$ とおけば、相対発生誤差が $j \cdot 2^{-m-\ell}$ (j は奇数) である確率は、

$$\sum_{i=0}^{k-\ell} f_{\ell+i}(j2^{-m-\ell}) P(\ell+i) \quad (3.9)$$

と表すことができる ($j=0$ のときは $\ell=0$ とおく)。 $P(k)$ がどのようになるかは計算の過程に依存する。

ここで、単純なモデルとして、指數の差が 0 の場合、1 の場合、…, k の場合が一様に起こり、指數の差が k ビットを越える 2 数の加算は無視できる状況を考える。すなわち、

$$P(0) = P(1) = \cdots = P(k) = \frac{1}{k+1}, \quad P(k+1) = \cdots = 0 \quad (3.10)$$

とする。

命題 5：式 (3.10) の仮定の下で、すなわち、指數の差が k 以下の 2 数 (正負問はず) の加算において観察される相対発生誤差は $j2^{-m-k}$ ($j = 0, \pm 1, \pm 2, \dots, \pm 2^k$) である。その確率 $g_k(j2^{-m-k})$ は次のように表される。

$$g_k(j2^{-m-k}) = \begin{cases} \frac{1}{k+1} \frac{3}{8} \frac{1}{4^k}, & j \text{ は奇数;} \\ \frac{1}{k+1} \left(\frac{1}{2^{k-p}} - \frac{1}{2^k} - \frac{1}{8} \frac{1}{4^k} + \frac{1}{8} \frac{1}{4^{k-p}} \right), & j = \pm j' 2^p \\ \frac{1}{k+1} \left(\frac{15}{8} - \frac{1}{8} \frac{1}{4^k} - \frac{1}{2^k} \right), & j = 0 \text{ のとき;} \\ \frac{1}{k+1} \left(\frac{1}{2} - \frac{1}{4^{k+2}} - \frac{1}{2^{k+1}} \right), & j = \pm 2^k \text{ のとき.} \end{cases}$$

この分布の平均は 0 であり、分散 V は次式で与えられる (分布は 0 に関して対称であることを用いている)。

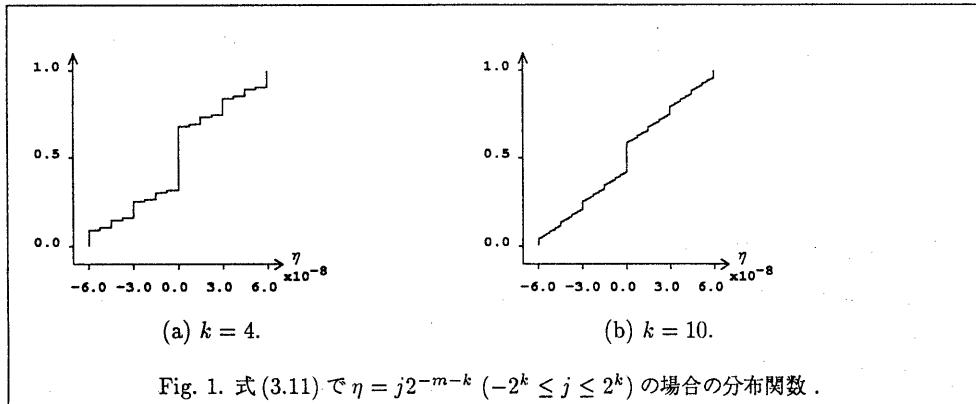
$$V = \frac{2^{-2m}}{k+1} \left(\frac{1}{3} k + \frac{65}{252} - \frac{2}{9} \frac{1}{2^{2k}} - \frac{9}{56} \frac{1}{2^{3k}} \right).$$

k が大きくなるほど一様分布の分散に近付く。

$m = 24, k = 4$ および 10 のときの分布関数 ($\eta = j2^{-m-k}$)

$$G_k(j2^{-m-k}) = \sum_{i=-2^k}^j g_k(i2^{-m-k}) \quad (3.11)$$

の様子を Fig. 1 に示す。



4 数値例

実際に、単精度計算の結果と倍精度計算の結果の差をとり、単精度計算の相対発生誤差(の近似値)を計算した。C++のオペレータオーバーロード機能を用い、単精度計算と倍精度計算を同時に実行し、発生誤差を出力するようなクラスを定義して、プログラムを組んだ。使用したコンパイラはSUN C++ version 3、計算機は富士通 S4/CL である。

相対発生誤差の分布は、プログラムのアルゴリズムとは直接的には関係なく、加減算における被演算数の指数の差がどのように分布するかということだけに依存することを確かめるための実験を行なった。選んだプログラムは、LU 分解により連立一次方程式を解くプログラム(P_{LU})、CG 法により連立一次方程式を解くプログラム(P_{CG})、入力値の総和と 2 乗和とを計算するプログラム(P_{ST})の 3 種である。これらについて、いくつかの入力値について計算を実行し、相対発生誤差を実測した。同時に、加減算の被演算数の指数の分布も調べた。

- (i) P_{LU} : $[-1, 1]$ 上の一様乱数を成分とする 20 次の行列を係数行列として持つ連立一次方程式を考える。LU 分解による解を P_{LU} で計算したときの「減算」における相対発生誤差の分布を Fig. 2 に示す。この形は、単純モデルの分布 (Fig. 1 (b)) に近い。実際、このときの減算の非演算数の指数の差の分布 (Fig. 3) を見ると、 $k \leq 4$ が全体の 8 割程度を占めていることがわかる。
- (ii) P_{LU} : 20 次の Hilbert 行列を係数行列として持つ連立一次方程式の LU 分解による解を P_{LU} で計算したときの「減算」における相対発生誤差の分布を Fig. 4 に示す。非演算数の指数の差の分布 (Fig. 5) は、 $k \leq 2$ が 9 割程度占めている。 $k = 2$ の単純モデル (Fig. 1(a)) がこれをよく近似していることがわかる。
- (iii) P_{CG} : 一様乱数を成分とする 20 次の(対称)行列を係数行列として持つ連立一次方程式の共役勾配法による解を、 P_{CG} で計算したときの「加算」における相対発生誤差の分布を Fig. 6 に示す。アルゴリズムは P_{LU} と異なるが、同様の分布を示すことがわかる。アルゴリズム自身は異なるものの、指数の差の分布 (Fig. 7) が似ていれば、結果として、相対発生誤差の分布も似たようなものになっていることが確認できる。(係数行列の正定値性はここでは問題ではないので、 P_{CG} の実行完了だけを確認した。)
- (iv) P_{CG} : 20 次の Hilbert 行列についての共役勾配法の結果を Fig. 8, Fig. 9 に示す。 P_{CG} の場合には、 P_{LU} と異なり、一様乱数による行列の場合と同様である。
- (v) P_{ST} : $[-1, 1]$ 上の一様乱数を 2500 個用意して、その和と 2 乗和を P_{ST} で計算した。このときの「加算」における相対発生誤差の分布を Fig. 10 に示す。これは一様分布に近い。指数の差の分布を Fig. 11 に示す。モデルからも明らかであるが、このように被演算数の指数の差が大きいような演算の占める割合が大きくなると、一様分布に近付くことが確かめられる。

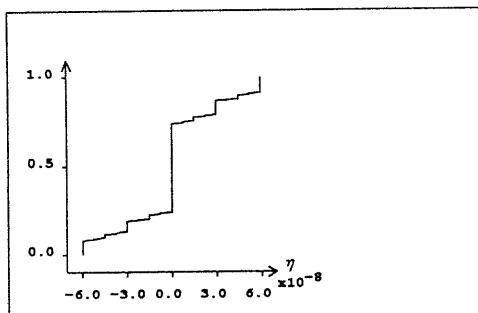


Fig. 2. 20次の1様乱数行列を係数行列に持つ連立一次方程式をLU分解により解いたときの減算の相対発生誤差の累積分布.

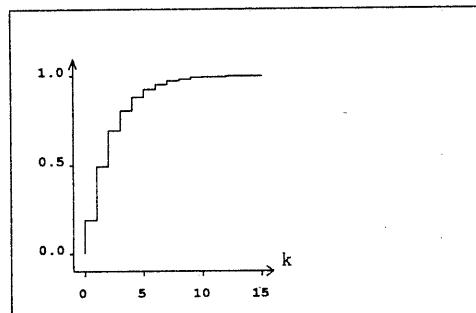


Fig. 3. Fig. 2を観測したときの指数の差の累積分布($k > 15$ は省略).

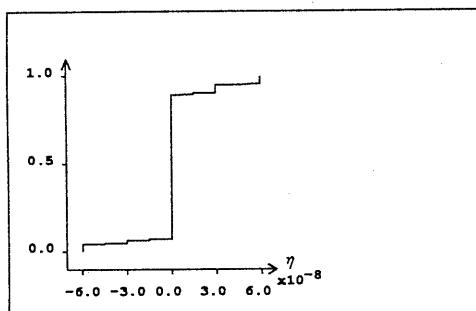


Fig. 4. 20次のHilbert行列を係数行列に持つ連立一次方程式をLU分解により解いたときの減算の相対発生誤差の累積分布.

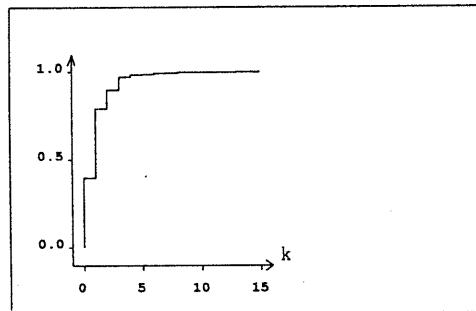


Fig. 5. Fig. 4を観測したときの指数の差の累積分布($k > 15$ は省略).

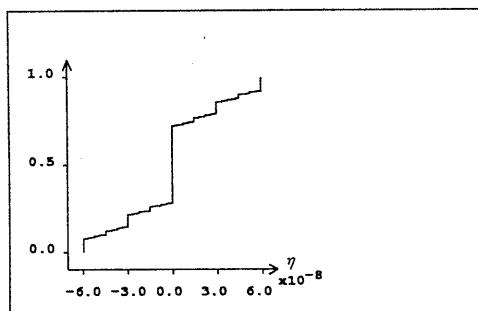


Fig. 6. 20次の1様乱数行列を係数行列に持つ連立一次方程式をCG法で解いたときの加算の相対発生誤差の累積分布.

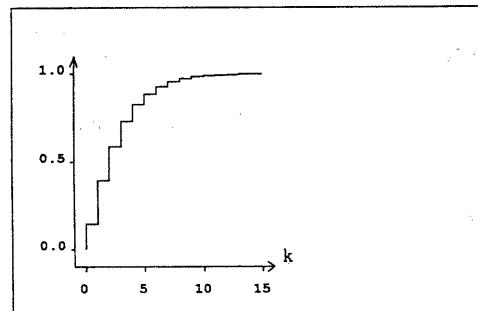


Fig. 7. Fig. 6を観測したときの指数の差の累積分布($k > 15$ は省略).

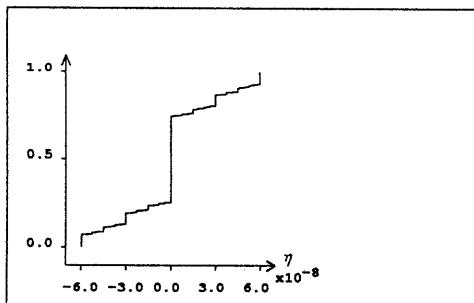


Fig. 8. 20次の Hilbert 行列を係数行列に持つ連立一次方程式を CG 法で解いたときの加算の相対発生誤差の累積分布。

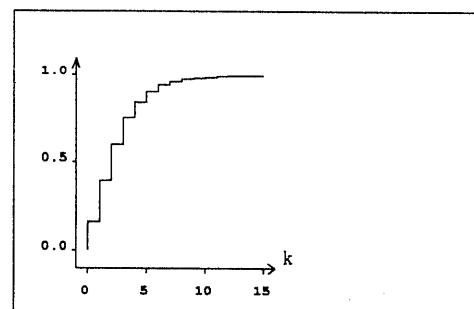


Fig. 9. Fig. 8 を観測したときの指数の差の累積分布 ($k > 15$ は省略)。

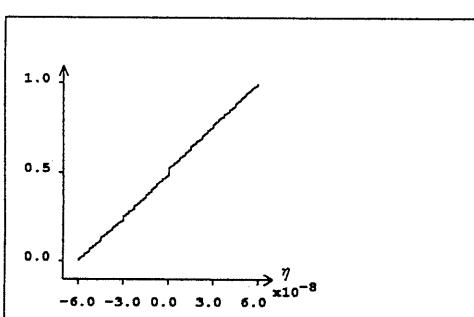


Fig. 10. 2500 個の一様乱数の和と自乗和を計算したときの加算の相対発生誤差の累積分布。

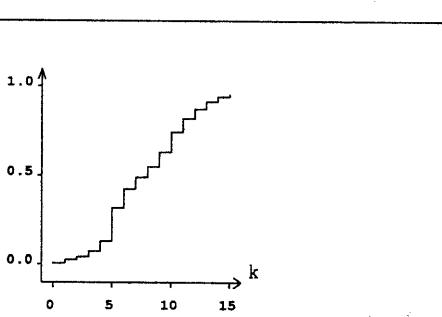


Fig. 11. Fig. 10 を観測したときの指数の差の累積分布 ($k > 15$ は省略)。

5まとめ

加減算における発生誤差の相対的な大きさの分布について考察し、加減算については特徴的な分布を得た。本論文では IEEE 標準にもとづく四則演算だけに絞って考察したが、他の演算や 16 進数演算についての解析は、今後の課題である。

なお、本研究の一部は文部省科学研究費補助金（一般研究 (B)05452120: 大規模システムの感度解析技法の総合的研究）の援助による。

参考文献

- [1] W. Feller: *An Introduction to Probability Theory and Its Applications, Volume II*. John Wiley and Sons, 1966.
- [2] IEEE: IEEE Standard 754-1985 for binary floating-point arithmetic. *ACM SIGPLAN*, Vol. 22 (1987), No. 2, pp. 9-25.
- [3] 伊理 正夫, 土谷 隆, 星 守: 偏導関数計算と丸め誤差推定の自動化の大規模非線形方程式系への応用. *情報処理*, Vol. 26 (1985), pp. 1411-1420.
- [4] M. Iri, T. Tsuchiya and M. Hoshi: Automatic computation of partial derivatives and rounding error estimates with applications to large-scale systems of nonlinear equations. *Journal of Computational and Applied Mathematics*, Vol. 24 (1988), pp. 365-392.
- [5] J. Vignes: A stochastic arithmetic for reliable scientific computation. *Mathematics and Computers in Simulation*, Vol. 35 (1993), pp. 223-261.