平方根の計算においてビット精度、表式および丸め方式の違いにより現れる３種のアトラクタ

Three Different Attractors Dependent on Bit Precision, Mathematical Expressions and Rounding Modes in Square-Root Calculations

矢吹道郎
YABUKI, Michiro
明星大学情報学部

土屋 尚
TSUCHIYA, Takashi
明星大学名誉教授

要旨

収束解を求める数値計算の代表としてニュートン法による平方根の求解を例にとり、仮数部の bit 長、表式、まるめ方式をパラメータとして実験を行う。その結果、LSB(Least Significant Bit) の値の距離を持つ２つの値の間の振動、１つの値に収束、LSB の値の距離を持つ２つの値の一方に収束の３種のアトラクタがあり、これらは仮数部の bit 長、表式、まるめ方式により決定されることを示す。

1　はじめに

収束解を求める計算に代表されるような数値計算においてはその解は誤差を持ち、数学的に真の値とはならないため、計算におけるメモリのビット数、演算精度のビット数、あるいは物理的な要件から許容誤差を定め、解と見なすことができる値が得られた時点で計算を打ち切っている。一般に許容誤差は 10 進数で与えられ、C 言語の型で言うならば、float (32bit)、double (64bit)、long double (processor 及び処理系依存) などの変数のメモリビット長に依存する精度に合わせて決定される。そして、一般に、指定された許容誤差より小さい計算結果の違いは議論されない。

本論文で我々はこの許容誤差を設定せずに計算を進めるとどのようなことが起きるかを、変数のメモリのビット数を変化させながら観察する。このことを可能とするため、多倍長計算ライブラリ MPFR(Multiple-Precision Binary Floating-Point Reliable Library)[1] を用いる。これは GMP(GNU Multiple Precision Arithmetic

Library)[2] を拡張したもので、仮数部の精度を 10 億 bit のオーダーまで設定できる特徴をを持つ。我々は、既に MPFR の基礎的性能の評価を行った [3]。

とりあげる計算は、数値計算で最も基本的なニュートン法による 2 の平方根の計算である。変化させるのはメモリの bit 数だけでなく、ニュートン公式の表現の方式 (表式)、まるめの方式、初期条件である。実験の結果、興味深いことに bit 精度とまるめ方式に依存して、 1 つの値に収束する場合、 2 つの値の間を振動する場合があることが分かった。これは力学系の言葉でいえば、固定点と 2 周期という質の異なるアトラクタが現れたことになる。さらに興味深いことに、初期値に依存して 2 つの異なる固定点に収束する場合もあることが分かった。力学的には 2 つの固定点の共存ということになる。

次節でまるめ方式と MPFR について解説する。第 3 節で許容誤差について述べる。第 4 節でニュートン法の表式の種類を示し、第 5 節では典型的な計算結果を例示する。第 6 節に我々の行ったすべての実験結果をまとめて報告する。第 7 節はまとめである。

2　浮動小数と多倍長計算

2.1　浮動少数の形式

浮動小数は一般に IEEE-754[4][5] に定められた形式に準拠した形で利用される。最も広く利用されている Intel の現在の CPU の浮動小数点プロセッサも IEEE-754 に準拠している。IEEE-754 では演算結果のまるめとして、

- 正の無限大へのまるめ。いわゆる切り上げ。
- 負の無限大へのまるめ 。いわゆる切り捨て。
- 最近接偶数まるめ 。いわゆる 10 進の四捨五入。ちょうど中点となる場合 (10 進の 5 にあたる) に、まるめの結果が偶数になるように切り上げと切り捨てが選択される。
- 0 へのまるめ 。絶対値が小さくなるような切り捨て。

の 4 つのモードが定義されている。

一般の数値計算プログラムで利用される 32bit 浮動小数点 (float:単精度)、64bit 浮動小数点 (double:倍精度) は、それぞれ、仮数部が 23bit ビット、52bit であ。ただし、正規化による先頭 bit の 1 の bit は隠れ bit とされ値として格納されないため、実質的にはそれぞれ 24bit、53bit の仮数部を持つ。(本論文では float、double の仮数部の bit 長を、それぞれ実質的な値である、24bit、53bit として扱う。) float、double は 10 進でそれぞれおよそ 7 桁、16 桁の精度を持ち、特に指定しない限り最近接偶数まるめが用いられる。

2.2　GMP と MPFR

GMP は GNU Multiple Precision arithmetic library の略で、仮数部として自由な bit 数を設定可能な C 言語で書かれた多倍長計算のための数値計算ライブラリである。単なるライブラリであり、計算式解釈のためのメカニズム、あるいは言語仕様を持たないため、加減乗除等の演算、代入のための関数等を使用しなくてはならない。

MPFR は基本的に GMP を拡張したライブラリで、ポータビリティを向上させたこと、仮数部の bit 数を広範囲に設定できること、IEEE754-1985 標準に準拠したまるめモードを持つ事等の違いがある。MPFR は理論上仮数部として最大約 21 億 bit まで設定可能である。

我々は既に GMP と MPER の四則演算の計算速度についての基礎的性能の評価を行い、実用上の可能性を示した [3]。

3　収束解を求める数値計算における許容誤差

一般に収束解を求める数値計算においては、許容誤差を定義し、計算結果が許容誤差範囲に入った時に解とし、計算を終了する。数値計算プログラムでは用いる変数の限界 (すなわち仮数部 bit 数) と問題固有の必要性を考慮して許容誤差を 10 進の値で指定する。許容誤差の値を利用する浮動小数点の表現の限界 (ハードウェア的に言うならばマシンイプシロン) に対して安全側に与えれば、与えられた初期条件に対して解 (沈点、アトラクター) の吸引領域が広がり、解を得やすくなる。

しかしながら計算機による演算はすべて 2 進で行われるため、10 進で許容誤差を表現した場合には、許容誤差が浮動小数による 2 進の表現の限界に対して一定でないため、その収束に対する評価を一様に行う事はできない。

4　ニュートン法の表式

数値演算における誤差を議論するために、最も代表的な数値計算の例としてニュートン法により平方根を求める例を取り上げる。ニュートン法による計算式は、

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{1}$$

である。当然のことながら実際に使われる式は $f(x)$ に依存して変形されるが、値 r の平方根を求める場合には、以下の 3 つの式のいずれかが用いられる。

$$x_{n+1} = x_n - \frac{x_n^2 - r}{2x_n} \tag{2}$$

$$x_{n+1} = \frac{x_n^2 + r}{2x_n} \tag{3}$$

$$x_{n+1} = \frac{x_n + \frac{r}{x_n}}{2} \tag{4}$$

数学的にはどの式も同じであるが、プログラミング言語を用いて記述した場合は、計算の順序、２項演算の回数、ならびに中間的な演算結果のまるめが行われるため、演算の過程の違いにより最終計算結果も異なってくる。

上記の 3 式を C 言語を用いて記述した例を以下に示す。

$$\mathtt{x} = \mathtt{x} - (\mathtt{x} * \mathtt{x} - \mathtt{r})/(2.0 * \mathtt{x}); \tag{5}$$

$$\mathtt{x} = (\mathtt{x} * \mathtt{x} + \mathtt{r})/(2.0 * \mathtt{x}); \tag{6}$$

$$\mathtt{x} = (\mathtt{x} + \mathtt{r}/\mathtt{x})/2.0; \tag{7}$$

式 (5) 式 (6) 式 (7) は、それぞれ式 (2) 式 (3) 式 (4) に対応した記述で、演算の回数は 5 回、4 回、3 回となっている。

5　実験結果

5.1　仮数部 53bit(double) を用いた計算の場合

実際に 2 の平方根を求める計算を行ってみる。初期値を 2.0 として、double(仮数部 53 ビット) を用い、まるめ方式を一般的な最近接偶数まるめとしてとした結果を示す。ここでは結果を判断するに十分なステップだけを表示する。当然のことながら計算ステップをさらに進めても結果は変わらない。

式 (6) を用いた場合は、

```
ステップ              値
   1  1.50000000000000000000
   2  1.41666666666666407815
   3  1.41421568627450988664 0
   4  1.41421356237468986982 71
   5  1.41421356237309492343 00
   6  1.41421356237309492343 00
   7  1.41421356237309492343 00
```

となり、5 ステップ目で収束していることが分かる。同様に式 (7) を用いた場合には、

```
ステップ              値
   1 1.50000000000000000000
   2 1.41666666666666651863 69
   3 1.41421568627450966459 94
   4 1.41421356237468986982 71
   5 1.41421356237309492343 00
   6 1.41421356237309492343 00
   7 1.41421356237309492343 00
```

となり、途中の過程は異なるものの、 5 ステップ目で同じ値に収束している。

一方、式 (5) を用いると、

```
ステップ              値
   1  1.50000000000000000000
   2  1.41666666666666407815
```

```
  3  1.414215686274509866440
  4  1.414213562374689869827 1
  5  1.414213562373095145474 6
  6  1.414213562373094923430 0
  7  1.414213562373095145474 6
  8  1.414213562373094923430 0
  9  1.414213562373095145474 6
 10  1.414213562373094923430 0
```

となり、結果は収束せずに振動する。

式 (5) を用い、まるめのモードを切捨てとすると、同じ初期値 2.0 では、

```
ステップ              値
  1  1.500000000000000000000
  2  1.416666666666666518636 9
  3  1.414215686274509866440
  4  1.414213562374689869827 1
  5  1.414213562373094923430 0
  6  1.414213562373094923430 0
  7  1.414213562373094923430 0
```

となり、式 (6)(7) で最近接偶数まるめを用いた場合と同じ値に収束する。ところが、式 (5) で丸めのモードを切捨てのまま初期値 3.0 とすると、

```
ステップ              値
  1  1.833333333333333481363 0
  2  1.462121212121212154855 2
  3  1.414998429894802844941 6
  4  1.414213780047197666078 7
  5  1.414213562373111798819 9
  6  1.414213562373095145474 6
  7  1.414213562373095145474 6
  8  1.414213562373095145474 6
```

となる。すなわち、式 (5) で最近接偶数まるめを用い初期値 2.0 の場合の振動解の大きい方の値に収束する。２つの収束値は振動する場合の２つの値のそれぞれと同じになっている。

5.2　仮数部 7bit を用いた計算の場合

前項では double(仮数部の 53bit) による計算結果で３種類の結果が生じることを示したが、結果を分かりやすくするため、ここでは MPFR を用いて仮数部の bit 長を 7 とし、結果が振動、１つの収束値、２つの収束値となることを示す。

5.2.1　振動

式 (6)、最近接まるめ方式、初期値 1.55〜1.70、初期値の刻み幅.015625 の場合の結果を図 1 に示す。

振動している２つの値は、1.421875(仮数部の値は 2 進で 1011010) と 1.406250(仮数部の値は 2 進で 1011011) であり、7bit の値として bit 距離 1 である。初期値によ

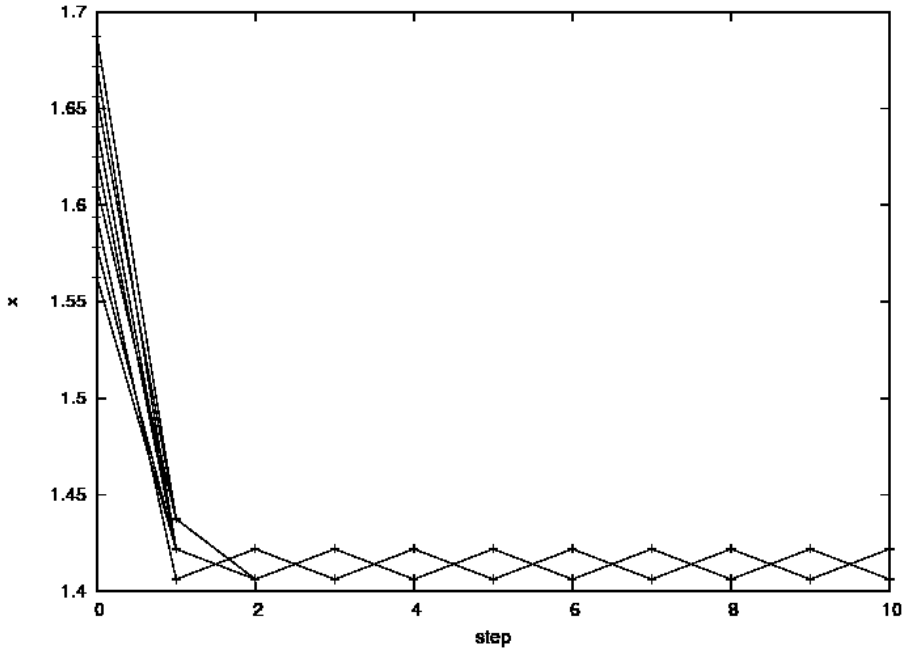

図 1　振動となる場合 (式 (6)、最近接まるめ、初期値 1.55〜1.70、初期値の刻み幅.015625)

り 2 つの位相を持っている。力学系の言葉で言えば、LSB である 7bit 目で 1bit の距離を持つ 2 つの値を取る周期 2 の振動に陥ったことになる。

ここでは仮数部 7bit の例を示したが、この LSB である bit 距離 1 の値で振動する事実は、変数の仮数部 bit 長に依存しない。

5.2.2　2 つの収束値

収束する場合は、変数の bit 長、計算式、まるめ方式の要因が同じでも、初期値により 2 つの値 (収束しない場合の 2 つの値) のどちらかに収束する場合と、初期値によらず 1 つの値に収束する場合がある。ここでは 2 つの収束値を持つ場合を示す。

変数の bit 長 7、初期値 1.55〜1.70、初期値の刻み幅.015625、式 (5)、負の無限大へのまるめ方式の場合で、初期値に依存して 2 つの値に収束する結果を図 2 に示す。

ここでも収束する 2 つの値は、1.421875(仮数部の値は 2 進で 1011010) と 1.406250(仮数部の値は 2 進で 1011011) であり、振動の場合の 2 つの値と同じである。力学系の言葉で言えば、初期値に依存して 2 つの固定点が共存することになる。

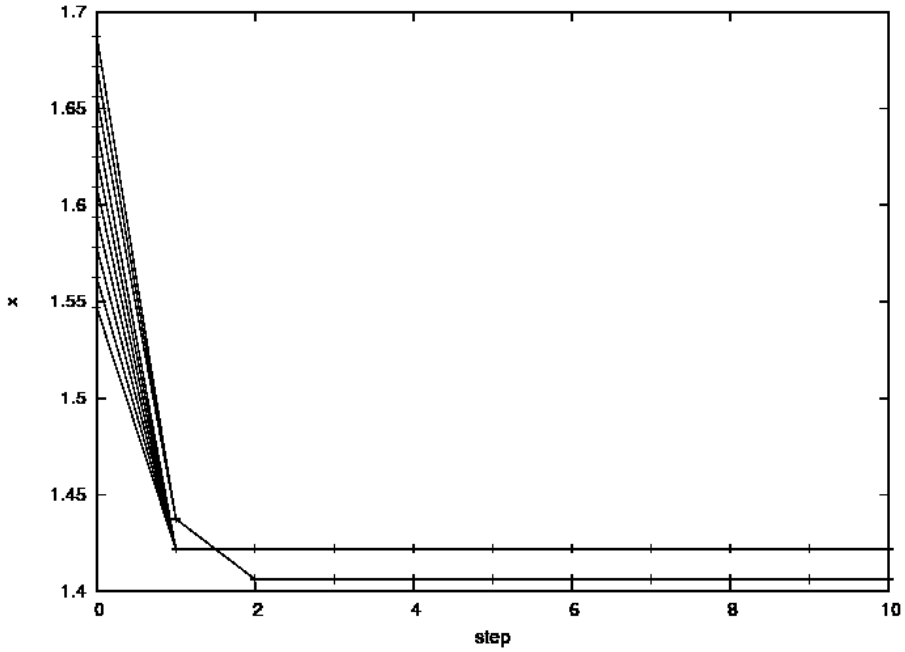

図 2　２つの収束値となる場合 (式 (5)、負の無限大へのまるめ、初期値 1.55〜1.70、初期値の刻み幅.015625)

5.2.3　１つの収束値

変数の bit 長 7、初期値 1.55〜1.70、初期値の刻み幅.015625、式 (5)、最近接まるめ方式の場合で、初期値によらず収束する場合の結果を図 3 に示す。収束値は 1.406250(仮数部の値は 2 進で 1011011) である。

6　bit 長と表式とまるめの影響

前節では 53bit の場合と 7bit の場合の幾つかの場合について示したが、ここでは仮数部の bit 長、まるめ方式、式の形について、MPFR を用いることにより以下のパラメータの範囲

- 計算式 式 (5)(6)(7)
- 変数の bit 長 6〜150bit
- まるめ方式 IEEE-754 で定められた、0 へのまるめ、正の無限大へのまるめ、負の無限大へのまるめ、最近接まるめ (偶数へのまるめ) の 4 つのモード
- 初期値 2〜3000

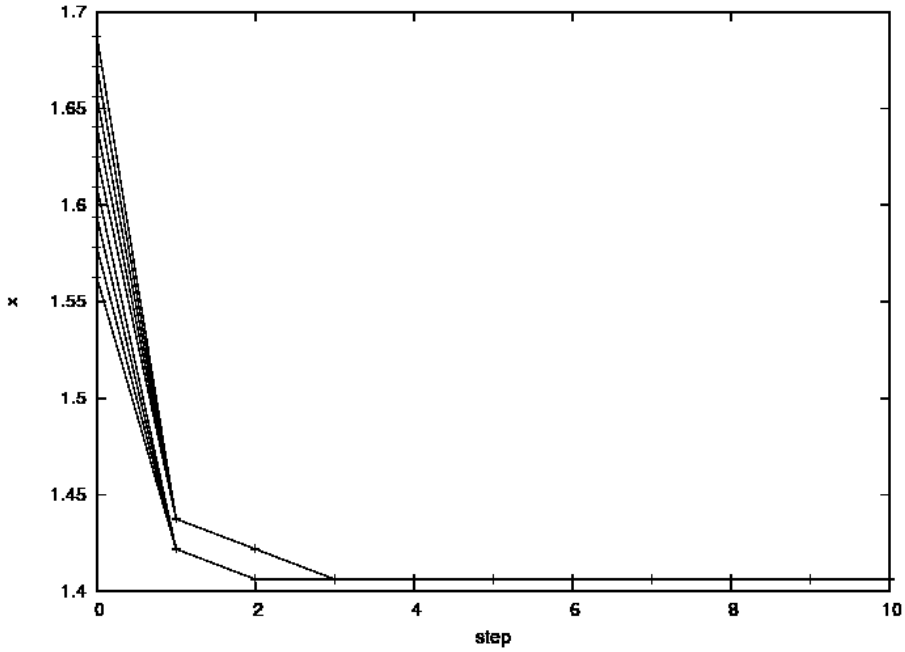

図 3　1 つの収束値となる場合 (式 (5)、最近接まるめ、初期値 1.55〜1.70、初期値の刻み幅.015625)

を設定し、実験を行った結果を報告する。

これまでの結果と同じく、すべてのパラメータの範囲において計算の結果は、

1. 収束せず bit 距離 1 で振動する
2. 初期値に依存して 2 つの値のどちらかに収束する (2 つの値は同じ bit 長の演算で収束しない場合の振動する 2 つの値のどちらかである)
3. 初期値によらず 1 つの値に収束する

の 3 つパターンのいずれかに分類された。

式 (5) とし、初期値を 2〜3、初期値の変化幅を 1/1000 とした場合の結果を図 4 に示す。

ビットの影響を見やすくするため、以下 6〜32bit の範囲で示した図が図 5 である。図において、To zero は 0 へのまるめ、To plus は正の無限大へのまるめ、To minus は負の無限大へのまるめ 、Nearest は最近接偶数まるめ である。2 cycles は振動、2 values は 2 つの収束値、stable は 1 つの収束値である。

同じく、式 (5) とし、初期値を 5.0〜100.0、初期値の変化幅を (100-5)/637 とした場合の結果を図 6 に示す。図 5 と図 6 はまったく同じ結果となった。すなわち、どの

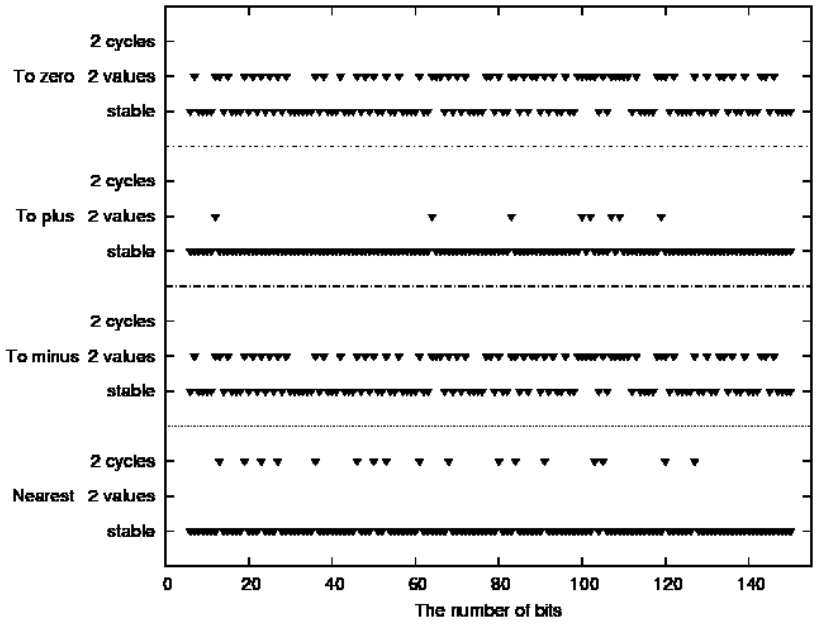

図 4　式 (5) による結果のパターン (初期値 2.0〜3.0、仮数部 bit 長 6〜150))

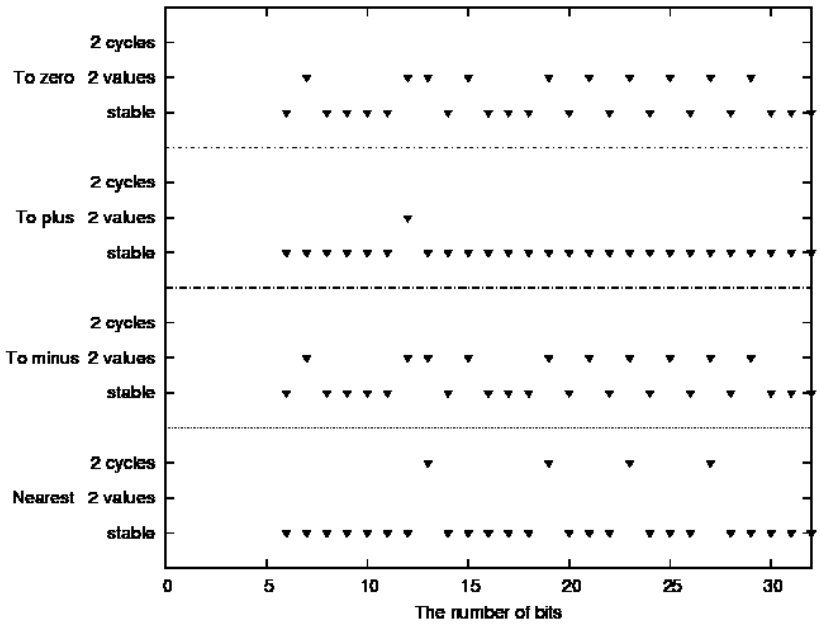

図 5　式 (5) による結果のパターン (初期値 2.0〜3.0、仮数部 bit 長 6〜32)

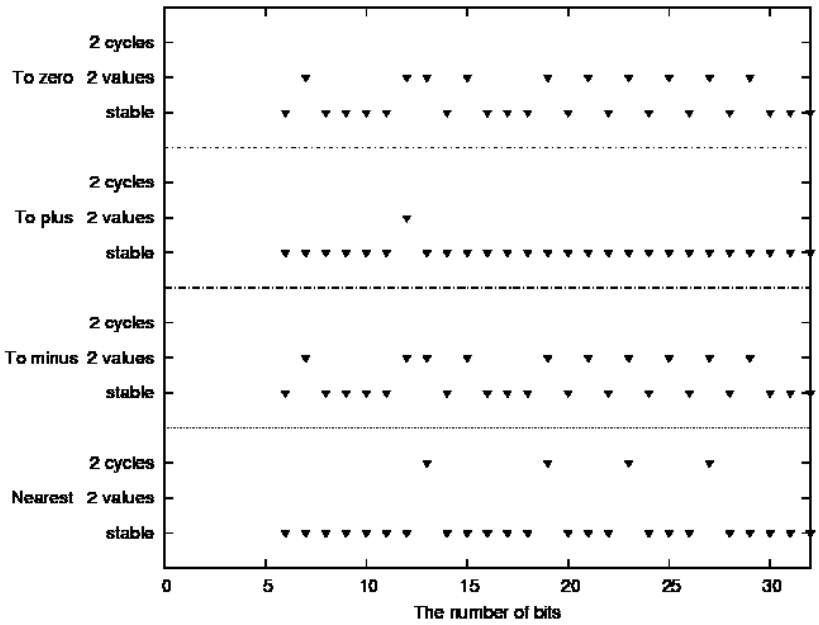

図 6　式 (5) による結果のパターン (初期値 5.0〜100.0、仮数部 bit 長 6〜32)

ような結果が得られるかは、初期値によらないことが分かる。

一方、式 (6) とし、初期値を 2.0〜3.0、初期値の変化幅を 1/1000 とした場合の結果を図 7 に示す。すなわち、計算結果がどのパターンになるかは、初期値、初期値の変化幅に関係なく、計算式、変数のビット長、まるめ方式にのみ依存することが分かった。このことは変数の bit 長を 150 ビットまで変化させても変わらない。式 (6) においても、初期値、初期値の変化幅を変えても結果は変わらない。

この計算では $r = 2.0$ すなわち 2 の平方根を求めているが、3 の平方根を求める場合もパターンは異なるが、初期値によらず式とまるめによりパターンが決定されることは同じである。

式 (7) については、結果がすべて 1 つの収束値となった。結果を図 8 に示しておく。すなわち、bit 長やまるめ方式に影響されず、安定的に 1 つの収束値を持つ式と言える。

7　終りに

本研究では、最も基礎的な数値計算法であるニュートン法による 2 つの平方根を求める問題を例に、許容誤差を設定せず、変数の記憶領域の bit 長を精度としてその精度を変化させることにより、どのような現象が現れるかを調べた。その結果、表式およびまるめ方式に依存し 1 つの固定点に収束する場合と、 2 つの値を振動的にとる場

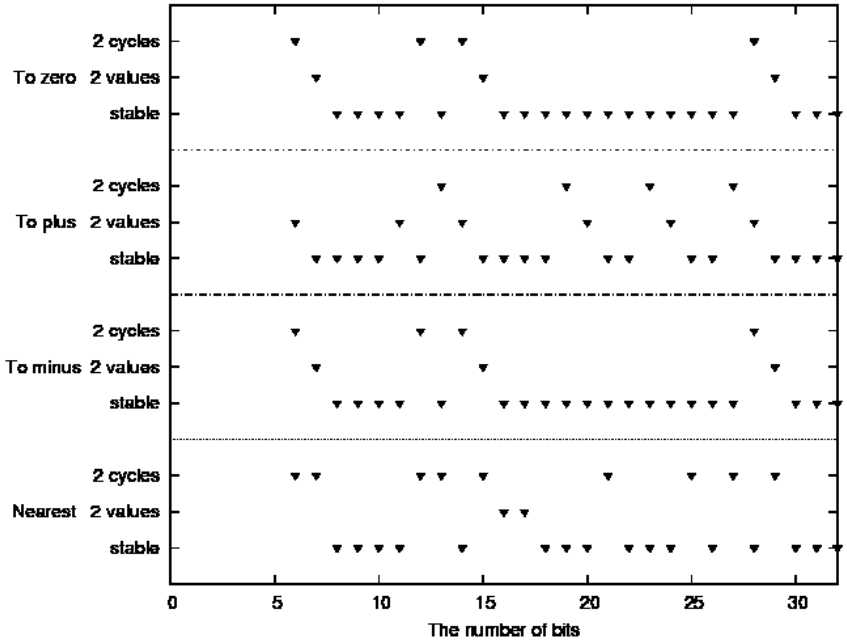

図 7　式 (6) による結果のパターン (初期値 2.0～3.0、仮数部 bit 長 6～32)

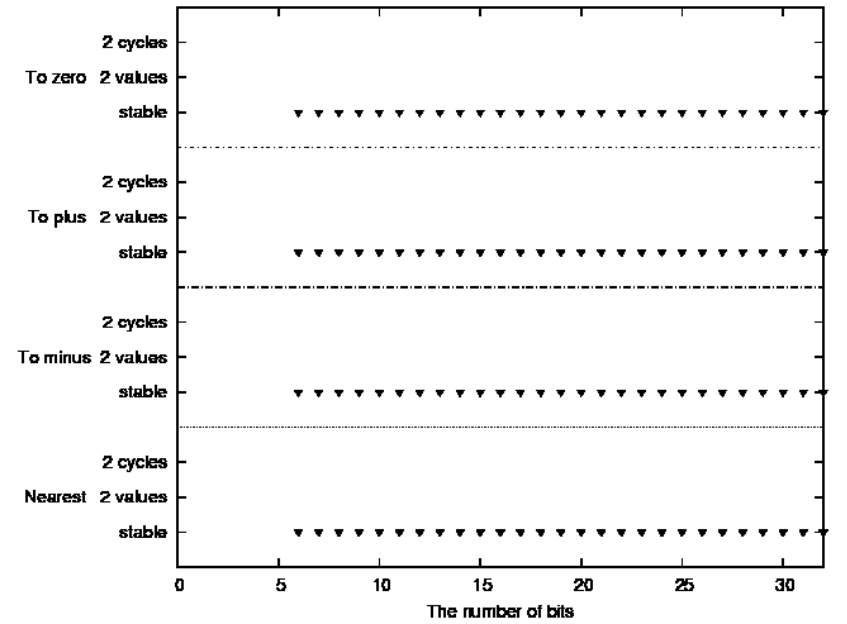

図 8　式 (7) による結果のパターン (初期値 5.0～100.0、仮数部 bit 長 6～32)

合があることが分かった。さらに初期条件に依存して、２つの異なる固定点に収束する固定帝の共存という場合もあることを発見した。

我々は、これらの振る舞いの違いを生む原因が、カオスに代表される収束しない計算を計算機で行う際のアーティファクトに影響を与えていると考えている。

参考文献

[1] Laurent Fousse, Guillaume Hanrot, Vincent Lefèvre, Patrick Pélissier and Paul Zimmermann, “MPFR: A Multiple-Precision Binary Floating-Point Reliable Library With Correct Rounding”, ACM Transactions on Mathematical Software, volume 33, issue 2, article 13, 15 pages, (2007), ‘http://doi.acm.org/10.1145/1236463.1236468’.

[2] Torbjőrn Granlund, “GNU MP: The GNU Multiple Precision Arithmetic Library”, version 4.2.2 (2007), ‘http://gmplib.org’.

[3] 矢吹道郎、土屋尚 “Bit 長指定の多倍長計算ライブラリ GMP と MPFR の評価”, 明星大学情報学部研究紀要, no.19, (2011), 1-8.

[4] Technical Report ANSI-IEEE Standard 754-1985, “IEEE standard for binary floating-point arithmetic”, American National Standards Institute, (1985).

[5] ANSI-IEEE Standard 754-2008, “IEEE Standard for Floating-Point Arithmetic”, IEEE Standards Board, (2008).