平成 12 年度

卒業論文

Affine Arithmeticの丸め誤差について

On the Rounding Error of the Affine Arithmetic

平成 13 年 2 月 5 日

指導教授： 柏木 雅英 助教授

早稲田大学理工学部情報学科

G97P0503 塩田 孝一

目 次

1 序論 **5**

2 区間演算 **8**

2.1 はじめに 9

2.2 区間 9

2.3 区間演算 9

2.4 むすび 11

3 Affine Arithmetic **12**

3.1 はじめに 13

3.2 Affine 形式 13

3.3 初期化 14

3.4 通常の区間表示への還元 15

3.5 Affine 演算 15

3.6 線形演算 15

3.7 非線形演算 16

3.7.1 非線形単項演算の定義 16

3.8 非線形二項演算 17

3.9 むすび 18

4 Affine Arithmetic における丸め誤差 **19**

4.1 はじめに 20

4.2 機械区間演算 20

4.3 実数の Affine 形式による表示 20

4.4 丸め誤差を評価する 3 つの方法 21
4.4.1 第一の方法 22
4.4.2 第二の方法 23
4.4.3 第三の方法 26
4.5 むすび 29

5 数値例 **30**
5.1 はじめに 31
5.2 第一の方法 31
5.3 第二の方法 31
5.4 第三の方法 31
5.5 考察 31

6 まとめ **32**

7 謝辞 **34**

参考文献 **36**

図 目 次

3.1 x と y に相関がない場合 14
3.2 x と y に相関がない場合 14
3.3 非線形関数の線形近似 17

4.1 a, b, c, r の関係 21

表 目 次

2.1 区間 [x],[y] の乗算 10
2.2 区間 [x],[y] の除算 10

第 1 章

序論

微分方程式の境界値問題などの連続数学の問題は, 一般に解析的に解くことが難しいので計算機上の離散的かつ有限精度の計算に置き換えて解くことがよく行われる. 連続数学の問題は実数の体系の上で記述されるのに対し, 計算機を援用して行う数値計算では実数を浮動小数点数に近似して計算する. しかし, 計算機の浮動小数点数は有限個であるので, 実数を厳密に表わすことができない. そこで, 2 つの浮動小数点を両端とする区間の中に連続数学の問題の解を包み込み, その区間同士の加減剰余等の演算を演算結果として有り得る集合を包含するように定義する, 区間演算という考え方が生まれた.

区間演算を生かすために区間を数の拡張として考えることが始められた. これは幅の充分狭い区間に求める解が含まれていれば実用上充分だからである. つまり, 区間を数として扱い, 四則演算を定義したものが区間演算である.

区間演算の大きな欠点として計算された区間は真の解を確実に含むものの, 区間幅が極端に増大してしまうことが挙げられる. 例として $f(x) = x^2 - 2x - 1$ を扱う. 通常の区間演算の手法で区間 $[0.9, 1.1]$ の演算を行なうと $[-2.39, -1.59]$ という結果になってしまう. 真の値は $[-2.0, -1.99]$ であるのでこれでは実に 80 倍の区間の開きである. この区間の増大の原因は二つの関数 x^2 と $2x$ が $[0.9, 1.1]$ の区間では同じ x の関数で互いに相関があり, しかも区間内で傾きが非常に近い値をとりそれの減算をするためであり, 区間演算における演算ではこの相関性を無視してそれぞれ独立な値として計算しているため例のように区間が爆発的に増大してしまうのである.

本論文では, 区間演算で起こる爆発的な区間の増大を解決するために Affine Arithmetic を取り上げる. この手法は区間演算の一種であるが, 区間演算と違うのは, 一般の区間演算では区間が他の区間に全く影響を与えないのに対し,Affine Arithmetic では変量が互いに関係を持っているという違いがある. この特徴により,Affine Arithmetic は計算式が膨大な場合や計算式が多くの変数を含んでいるような場合等に特にその効果を発揮する.

現在, 多くのパソコンやワークステーションでは IEEE 標準 754 という浮動小数点数システムを持ち, これらの計算機を援用して数値計算を行う場合は計算の各過程における丸めの向きを制御することができる [2]. そのため, 通常の区間演算の場合は, このような計算機に対して丸め誤差まで考慮したプログラムを容易に実装できるという利点がある.

反面,Affine Arithmetic で計算機を援用する場合, 丸め誤差を効率よく見積もるのが困難であるという問題点がある. すなわち、丸め誤差の見積もり方によっては、変量の初期化や計算の各過程で混入する丸め誤差が増幅して区間評価を甘くする場合がある。そこで、本論文では Affine

Arithmetic における丸め誤差の効率良い見積もり方について検討する。

本論文の構成は以下のようになっている.

- 第 2 章 精度保証付数値計算の基本的な手法である, 区間演算についての解説とその問題点を取り上げる.
- 第 3 章 Affine Arithmetic の定義とその有効性, また, 現状の解説をする.
- 第 4 章 丸め誤差を考慮した Affine 演算について解説をする.
- 第 5 章 実際に, 丸め誤差を考慮した 3 つの方法を用いて, 数値例を解き, 比較, 考察する.

第 2 章

区間演算

2.1 はじめに

本章では通常の区間演算の概念と定義の説明をする.

2.2 区間

区間演算は数値計算において最も一般的な手法であり,Affine Arithmeticの特徴は区間演算との比較で明確になる. 区間は実数値を [下限, 上限] という 2 つの浮動小数点で挟まれた区間で表現し

$$[\underline{x}, \overline{x}] = \{x \in R \mid \underline{x} \leq x \leq \overline{x}\}$$

とあらわす. ただし,$\underline{x} \leq \overline{x} \in R$ でそれぞれ区間の下端, 上端とする.

2.3 区間演算

区間演算は区間同士の加減乗除等の計算を「 演算結果として有り得る集合を包含するように 」定義することにより行なわれる.

2 つの区間 [x],[y] が与えられたときその区間の四則演算を次のように定義する.

$$[x] \circ [y] = \{x \circ y \mid x \in [x], y \in [y]\}$$

ただし $\circ \in \{+, -, \times, /\}$ とする. この定義では次が成立する.

$$\begin{aligned}
[x] + [y] &= [\underline{x} + \underline{y}, \overline{x} + \overline{y}] \\
[x] - [y] &= [\underline{x} - \overline{y}, \overline{x} - \underline{y}] \\
[x] \times [y] &= [\min\{\underline{x}\underline{y}, \overline{x}\overline{y}, \underline{x}\overline{y}, \overline{x}\underline{y}\}, \max\{\underline{x}\underline{y}, \overline{x}\overline{y}, \underline{x}\overline{y}, \overline{x}\underline{y}\}] \\
[x]/[y] &= [x] \times [1/\underline{y}, 1/\overline{y}], (0 \notin [y])
\end{aligned}$$

さらに乗算と除算については, 場合分けにより, より少ない手間で計算することができる (表 2.1, 表 2.2)

区間演算について, 包含関係における単調性,

$$[x] \subseteq [x'], [y] \subseteq [y'] \longrightarrow [x] \circ [y] \subseteq [x'] \circ [y'], \circ \in \{+, -, \times, /\}$$

	$y \geq 0$	$y \ni 0$	$y \leq 0$
$x \geq 0$	$[\underline{xy}, \overline{xy}]$	$[\overline{xy}, \overline{xy}]$	$[\overline{xy}, \underline{xy}]$
$x \ni 0$	$[\underline{x}\overline{y}, \overline{xy}]$	$[min\{\underline{x}\overline{y}, \overline{x}\underline{y}\}, max\{\underline{xy}, \overline{xy}\}]$	$[\overline{xy}, \underline{xy}]$
$x \leq 0$	$[\underline{x}\overline{y}, \overline{x}\underline{y}]$	$[\underline{x}\overline{y}, \underline{x}\overline{y}]$	$[\overline{xy}, \underline{xy}]$

表 2.1: 区間 [x],[y] の乗算

	$y \geq 0$	$y \leq 0$
$x \geq 0$	$[\underline{x}/\overline{y}, \overline{x}/\underline{y}]$	$[\overline{x}/\overline{y}, \underline{x}/\underline{y}]$
$x \ni 0$	$[\underline{x}/\underline{y}, \overline{x}/\underline{y}]$	$[\overline{x}/\overline{y}, \underline{x}/\overline{y}]$
$x \leq 0$	$[\underline{x}/\underline{y}, \overline{x}/\overline{y}]$	$[\overline{x}/\underline{y}, \underline{x}/\overline{y}]$

表 2.2: 区間 [x],[y] の除算

が成立する. また加法と乗法に関し交換則と結合則が成立する.

$$[x] \circ [y] = [y] \circ [x], \circ \in \{+, \times\}$$

$$[x] \circ ([y] \circ [z]) = ([x] \circ [y]) \circ [z], \circ \in \{+, \times\}$$

しかし, 減法と除法の逆元は存在しない. すなわち,$-[x] = [-\overline{x}, -\underline{x}]$ であるが

$$0 = [0] \subseteq [x] - [x] = [\underline{x} - \overline{x}, \overline{x} - \underline{x}]$$

$$1 = [1] \subseteq [x]/[x]$$

となる. 上式で等号は $[x]$ が点区間のときのみ成立する.

また, 分配則も区間演算に対しては成立しない. その代わり次の劣分配則が成立する.

$$[x] \times ([y] + [z]) \subseteq [x] \times [y] + [x] \times [z]$$

上式で等号は区間 [y] と [z] が同じ符号を持つときに成立する.

2.4 むすび

本章では区間演算の概説を行った. 区間演算はその考え方が単純でわかりやすい. しかし, 単純に区間演算を行っていくと, 区間の幅が大きくなってしまい, その結果の値の使うのには非常に扱いにくくなる. それは区間が単純に演算可能になっているだけで, それぞれの変量の相関性などについてはまったく考慮されないためである.Affine Arithmetic はこの区間演算の改良版であり, 区間演算よりもタイトな区間で演算を行うことが可能である. 次章ではその仕組みと演算の実装を紹介する.

第 3 章

Affine Arithmetic

3.1 はじめに

本章では Affine Arithmetic の概念と定義の説明をする.

3.2 Affine 形式

Affine Arithmetic は, 変数間の相関性を考慮することにより区間演算の区間幅の爆発的増大問題を解決する方法の一つである.

Affine Arithmetic では, 変量 x は affine 多項式

$$x = x_0 + x_1\varepsilon_1 + x_2\varepsilon_2 + \cdots + x_n\varepsilon_n \tag{3.1}$$

で表される. ここで,x_i は実数であり,ε_i はそれが区間 $[-1, 1]$ に含まれることだけが分かっているようなダミー変数であり,x の変量の部分のそれぞれ独立した構成要素である.

x_0 をこの Affine form の central value, 係数 x_i を部分偏差 (partial deviations),ε_i を noise symbols という.noise symbol である,ε_i はそれぞれ独立した値をとって変量 x を構成し, 係数 x_i がそのふれ幅を決定する.

たとえば,Affine 形式 x, y があり, それが

$$x = 1 + 0.5\varepsilon_1 \tag{3.2}$$

$$y = 1 + 0.5\varepsilon_2 \tag{3.3}$$

とする. これは x と y に相関性がない状態で, このとき (x, y) がとり得る領域は図 3.1 のようになる.

これに対し,

$$x = 1 + 0.5\varepsilon_1 \tag{3.4}$$

$$y = 1 + 0.4\varepsilon_1 + 0.1\varepsilon_2 \tag{3.5}$$

では, それぞれがとり得る範囲は $[0.5, 1, 5]$ で変わらないが ε_1 の係数を見ると分かるように両者には強い相関があり (x, y) のとり得る範囲は図 3.2 のようになる.

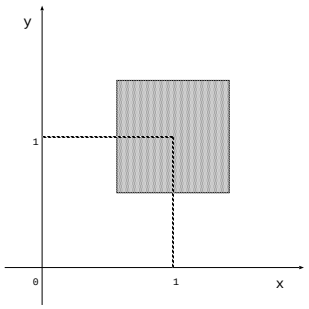

図 3.1: x と y に相関がない場合

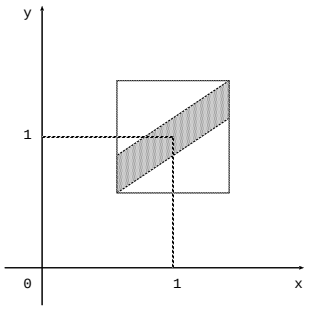

図 3.2: x と y に相関がない場合

3.3 初期化

n 変数の関数 $f(p_1, \cdots, p_n)$ の評価を行なう場合, 最初に与えられる n この変数 $p_1, \cdots, p_n$ は以下のような affine 多項式で初期化する.

$$\begin{aligned} p_1 &= \frac{\overline{p_1}+\underline{p_1}}{2} + \frac{\overline{p_1}-\underline{p_1}}{2}\varepsilon_1 \\ p_2 &= \frac{\overline{p_2}+\underline{p_2}}{2} + \frac{\overline{p_2}-\underline{p_2}}{2}\varepsilon_2 \\ &\vdots \\ p_n &= \frac{\overline{p_n}+\underline{p_n}}{2} + \frac{\overline{p_n}-\underline{p_n}}{2}\varepsilon_n \end{aligned} \tag{3.6}$$

但し, 入力変数 p_k の変域を $[\underline{p_k}, \overline{p_k}]$ とする.

3.4 通常の区間表示への還元

Affine 多項式は, 次のようにして通常の区間に変換できる.

$$[x_0 - \delta, x_0 + \delta], \delta = \sum_{i=1}^{n} |x_i| \tag{3.7}$$

3.5 Affine 演算

2 つの Affine 多項式 x, y,

$$x = x_0 + x_1\varepsilon_1 + \cdots + x_n\varepsilon_n \tag{3.8}$$

$$y = y_0 + y_1\varepsilon_1 + \cdots + y_n\varepsilon_n \tag{3.9}$$

に対して, 以下に各種演算の定義を行う. 演算とは,

$$\begin{aligned} z &= f(x, y) \\ &= f(x_0 + x_1\varepsilon_1 + \cdots + x_n\varepsilon_n, y_0 + y_1\varepsilon_1 + \cdots + y_n\varepsilon_n) \end{aligned} \tag{3.10}$$

という $f(x, y)$ を,

$$z = z_0 + z_1\varepsilon_1 + \cdots + z_n\varepsilon_n \tag{3.11}$$

のように,Affine 多項式の形を満たすように表すことである.

各種 Affine 演算は次のように定義するが, 線形演算と非線形演算にわけられる.

3.6 線形演算

Affine 多項式 x, y に対して, 加算, 減算, 定数倍は次のように定義する.

$$x \pm y = (x_0 \pm y_0) + (x_1 \pm y_1)\varepsilon_1 + \cdots + (x_n \pm y_n)\varepsilon_n \tag{3.12}$$

$$x \pm \alpha = (x_0 \pm \alpha) + x_1\varepsilon_1 + \cdots + x_n\varepsilon_n \tag{3.13}$$

$$\alpha x = (\alpha x_0) + (\alpha x_1)\varepsilon_1 + \cdots + (\alpha x_n)\varepsilon_n$$

3.7 非線形演算

演算 f が非線形な場合の演算を考える. 次の式のように,z を ε の式で表すことを考える.

$$\begin{aligned} z &= f(x_0 + x_1\varepsilon_1 + \cdots + x_n\varepsilon_n, y_0 + y_1\varepsilon_1 + \cdots + y_n\varepsilon_n) \\ &= f^*(\varepsilon_1, \cdots, \varepsilon_n) \end{aligned} \tag{3.14}$$

この場合は一般的に z を正しく表すような $z_0, \cdots, z_n$ を決定することはできない.

よって次のような f^a を与え,

$$f^a(\varepsilon_1, \cdots, \varepsilon_n) = z_0 + z_1\varepsilon_1 + \cdots + z_n\varepsilon_n \tag{3.15}$$

f^* と f^a の誤差を $z_{n+1}\varepsilon_{n+1}$ として次のように表す.

$$\begin{aligned} z &= f^a(\varepsilon_1+, \cdots, \varepsilon_n) + z_{n+1}\varepsilon_{n+1} \\ &= z_0 + z_1\varepsilon_1 + \cdots + z_n\varepsilon_n + z_{n+1}\varepsilon_{n+1} \end{aligned} \tag{3.16}$$

ここで z_{n+1} とは f^* と f^a の誤差の最大値であるから,

$$|z_{n+1}| \geq \max |f^*(\varepsilon_1, \cdots, \varepsilon_n) - f^a(\varepsilon_1, \cdots, \varepsilon_n| \tag{3.17}$$

である.

3.7.1 非線形単項演算の定義

f を非線形単項演算とすると, 関数 $f(x)$ は一般に曲線を描くからこれを Affine 形式で表すことはできない。

そこで,$f(x)$ を Affine 多項式

$$z_0 + z_1\varepsilon_1 + \cdots + z_n\varepsilon_n \tag{3.18}$$

で線形近似すると, その時に最大で

$$\delta = \max_{-1 \leq \varepsilon_i \leq 1} \left| f(x) - (z_0 + z_1\varepsilon_1 + \cdots + z_n\varepsilon_n) \right| \tag{3.19}$$

の誤差が生じるので,$f(x)$ の包含

$$\hat{f}(x) = (z_0 + z_1\varepsilon_1 + \cdots + z_n\varepsilon_n) + \delta\varepsilon_{n+1} \tag{3.20}$$

を非線形単項演算 f の結果とする.

なお, これは x の変域 X において $f(x)$ を

$$ax + b \tag{3.21}$$

で近似し, そのときの最大誤差

$$\delta' = \max_{x \in X} \left| f(x) - (ax + b) \right| \tag{3.22}$$

を用いて

$$(ax + b) + \delta' \varepsilon_{n+1} \tag{3.23}$$

を結果とした場合も, 同一の結果を得ることがわかっている.

直線 $y = ax + b$ をどのようにして設定し δ をいかに小さな値にするかが Affine の単項演算の焦点となる (図 3.3).

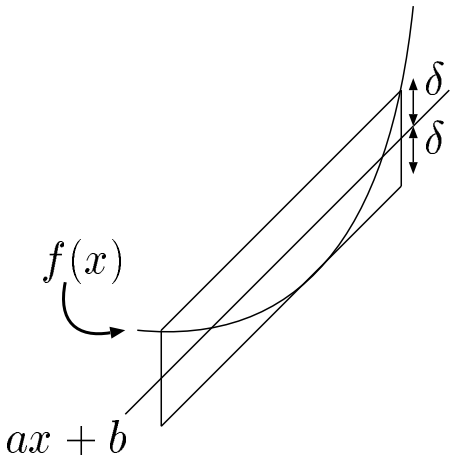

図 3.3: 非線形関数の線形近似

3.8 非線形二項演算

f が非線形二項演算であるとすると, 関数 $f(x, y)$ は一般に曲面を描くからこれを Affine 形式で表すことはできない.

そこで, 現状では x, y の変域をそれぞれ X, Y として, 長方形領域 $X \times Y$ において $f(x, y)$ を

$$ax + by + c \tag{3.24}$$

で近似し, そのときの最大誤差 δ を

$$\delta = \max_{\substack{x \in X \\ y \in Y}} \left| f(x, y) - (ax + by + c) \right| \tag{3.25}$$

で求め, ここで新たなダミー変数 ε_{n+1} を追加して得られる Affine 多項式

$$(ax + by + c) + \delta\varepsilon_{n+1} \tag{3.26}$$

を $f(x, y)$ の包含としている.

3.9 むすび

本章では,Affine Arithmetic の概説を行った.Affine Arithmetic は変量に相関性があり, 演算が面倒な分, 区間演算よりもタイトに演算を行える. 次章では,Affine Arithmetic の丸め誤差について検討してみる.

第 4 章

Affine Arithmeticにおける丸め誤差

4.1 はじめに

現在, 多くのパソコンやワークステーションではIEEE標準754という浮動小数点数システムを持ち, これらの計算機を援用して数値計算を行う場合は計算の各過程における丸めの向きを制御することができる[2]. そのため, 通常の区間演算の場合は, このような計算機に対して丸め誤差まで考慮したプログラムを容易に実装できるという利点がある. 一方,Affine Arithmeticは実数の表現が [下端 , 上端] という通常の区間演算のような形式をとらないため, 変量の初期化や計算の各過程でAffine多項式の係数に混入する計算機の丸め誤差をどのように反映させるかという問題点を抱えている.

4.2 機械区間演算

本節では計算機上で区間演算を展開するための手法について述べる. 以下,IEEE標準754に基づく浮動小数点数システムの場合を考える. 区間の両端が浮動小数点数で与えられたものを機械区間というが, 計算機上で区間演算を展開するためにはまず実区間を機械区間に置き換える必要がある. ただし, その際も数学的に正しい解を常に含んでいなければならないことに注意する. 実空間 $[a,b](a,b \in \mathbf{R})$ は $[(a$ の下向き丸め$),(b$ の上向き丸め$)]$ として機械区間表示される.
例) 実区間 $[1/7, 1/3]$ はC++で

```
volatile double a=1.,b=7.,c=3.;
down(); l=a/b;
up(); u=a/c;
```

のように $(1/7$ の下向き丸め$)l$ と $(1/3$ の上向き丸め$)u$ を計算することができ,
[0.14285714285714284921$(<1/7)$,$(1/3<)$0.3333333333333333]
のような機械区間に置き換えられる.

4.3 実数のAffine形式による表示

a, b を浮動小数点数とする.$[a,b]$ がAffine形式 $c + r\varepsilon$ に変換される時, c と r は以下のようにして決められる. 図1は a, b, c, r の関係を示している.c^* は $(a+b)/2$ の真の値である. なお、この初期化は下の三つの演算法にも適用される。

```
up();
c=(a+b)/2;
r=c-a;
```

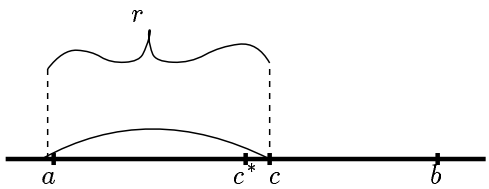

図 4.1: a, b, c, r の関係

以後は

```
up();
c=(a+b)/2;
r=c-a;
```

を

```
c=up((a+b)/2);
r=up(c-a);
```

として記述していく.

4.4 丸め誤差を評価する3つの方法

本論文では Affine 演算における丸め誤差の取り扱いに関して, 以下の 3 つの方法を実装した.

4.4.1 第一の方法

変数 x, y をそれぞれ

$$x = x_0 + x_1\varepsilon_1, \quad y = y_0 + y_2\varepsilon_2 \tag{4.1}$$

と初期化する。ここで、$z = x + y = (x_0 + y_0) + x_1\varepsilon_1 + y_2\varepsilon_2$ となるが、浮動小数点数の四則演算には丸め誤差が生じるので、そういう箇所 (この場合 $x_0 + y_0$) を区間評価し、その区間を係数とするのが第一の方法である。

Affine 形式

$$x = [\underline{x_0}, \overline{x_0}] + [\underline{x_1}, \overline{x_1}]\varepsilon_1 + [\underline{x_2}, \overline{x_2}]\varepsilon_2 + \cdots + [\underline{x_n}, \overline{x_n}]\varepsilon_n \tag{4.2}$$

通常の区間表示への還元

$$[\mathtt{down}(\underline{x_0} - \mathtt{up}(\sum_{i=1}^{n} \max(|\underline{x_i}|, |\overline{x_i}|))), \mathtt{up}(\overline{x_0} + \mathtt{up}(\sum_{i=1}^{n} \max(|\underline{x_i}|, |\overline{x_i}|)))] \tag{4.3}$$

線形演算

$$\begin{aligned}
x + y &= [\mathtt{down}(\underline{x_0} + \underline{y_0}), \mathtt{up}(\overline{x_0} + \overline{y_0})] + [\mathtt{down}(\underline{x_1} + \underline{y_1}), \mathtt{up}(\overline{x_1} + \overline{y_1})]\varepsilon_1 + \cdots && (4.4)\\
&+ [\mathtt{down}(\underline{x_n} + \underline{y_n}), \mathtt{up}(\overline{x_n} + \overline{y_n})]\varepsilon_n && (4.5)\\
x - y &= [\mathtt{down}(\underline{x_0} - \overline{y_0}), \mathtt{up}(\overline{x_0} - \underline{y_0})] + [\mathtt{down}(\underline{x_1} - \overline{y_1}), \mathtt{up}(\overline{x_1} - \underline{y_1})]\varepsilon_1 + \cdots && (4.6)\\
&+ [\mathtt{down}(\underline{x_n} - \overline{y_n}), \mathtt{up}(\overline{x_n} - \underline{y_n})]\varepsilon_n && (4.7)\\
[\underline{\alpha}, \overline{\alpha}]x &= [\min(\mathtt{down}(\underline{\alpha}\underline{x_0}), \mathtt{down}(\overline{\alpha}\underline{x_0}), \mathtt{down}(\underline{\alpha}\overline{x_0}), \mathtt{down}(\overline{\alpha}\overline{x_0})) && (4.8)\\
&\quad , \max(\mathtt{up}(\underline{\alpha}\underline{x_0}), \mathtt{up}(\overline{\alpha}\underline{x_0}), \mathtt{up}(\underline{\alpha}\overline{x_0}), \mathtt{up}(\overline{\alpha}\overline{x_0}))] && (4.9)\\
&+ [\min(\mathtt{down}(\underline{\alpha}\underline{x_1}), \mathtt{down}(\overline{\alpha}\underline{x_1}), \mathtt{down}(\underline{\alpha}\overline{x_1}), \mathtt{down}(\overline{\alpha}\overline{x_1})) && (4.10)\\
&\quad , \max(\mathtt{up}(\underline{\alpha}\underline{x_1}), \mathtt{up}(\overline{\alpha}\underline{x_1}), \mathtt{up}(\underline{\alpha}\overline{x_1}), \mathtt{up}(\overline{\alpha}\overline{x_1}))]\varepsilon_1 + \cdots && (4.11)\\
&+ [\min(\mathtt{down}(\underline{\alpha}\underline{x_n}), \mathtt{down}(\overline{\alpha}\underline{x_n}), \mathtt{down}(\underline{\alpha}\overline{x_n}), \mathtt{down}(\overline{\alpha}\overline{x_n})) && (4.12)\\
&\quad , \max(\mathtt{up}(\underline{\alpha}\underline{x_n}), \mathtt{up}(\overline{\alpha}\underline{x_n}), \mathtt{up}(\underline{\alpha}\overline{x_n}), \mathtt{up}(\overline{\alpha}\overline{x_n}))]\varepsilon_n && (4.13)
\end{aligned}$$

非線形単項演算

$$\mathtt{Min}(a, x_i) = \min(\mathtt{down}(\underline{a}\underline{x_i}), \mathtt{down}(\overline{a}\underline{x_i}), \mathtt{down}(\underline{a}\overline{x_i}), \mathtt{down}(\overline{a}\overline{x_i})) \tag{4.14}$$

$$\mathtt{Max}(a, x_i) = \max(\mathtt{up}(\underline{a}\underline{x_i}), \mathtt{up}(\overline{a}\underline{x_i}), \mathtt{up}(\underline{a}\overline{x_i}), \mathtt{up}(\overline{a}\overline{x_i})) \tag{4.15}$$

とすると

$$[\underline{a}, \overline{a}]x + [\underline{b}, \overline{b}] + \overline{\delta}\varepsilon_{n+1} = [\mathtt{down}(\mathtt{Min}(a, x_0) + \underline{b}), \mathtt{up}(\mathtt{Max}(a, x_0) + \overline{b})] \tag{4.16}$$

$$+ [\mathtt{Min}(a, x_1), \mathtt{Max}(a, x_1)]\varepsilon_1 + \cdots + [\mathtt{Min}(a, x_n), \mathtt{Max}(a, x_n)]\varepsilon_n \tag{4.17}$$

$$+ \overline{\delta}\varepsilon_{n+1} \tag{4.18}$$

これを $f(x)$ の包含とする.

非線形二項演算

$$[\underline{a}, \overline{a}]x + [\underline{b}, \overline{b}]y + [\underline{c}, \overline{c}] + \overline{\delta}\varepsilon_{n+1} = [\mathtt{down}(\mathtt{Min}(a, x_0) + \mathtt{Min}(b, y_0) + \underline{c}) \tag{4.19}$$

$$, \quad \mathtt{up}(\mathtt{Max}(a, x_0) + \mathtt{Max}(b, y_0) + \overline{c})] \tag{4.20}$$

$$+ [\mathtt{down}(\mathtt{Min}(a, x_1) + \mathtt{Min}(b, y_1)) \tag{4.21}$$

$$, \quad \mathtt{up}(\mathtt{Max}(a, x_1) + \mathtt{Max}(b, y_1))]\varepsilon_1 + \cdots \tag{4.22}$$

$$+ [\mathtt{down}(\mathtt{Min}(a, x_n) + \mathtt{Min}(b, y_n)) \tag{4.23}$$

$$, \quad \mathtt{up}(\mathtt{Max}(a, x_n) + \mathtt{Max}(b, y_n))]\varepsilon_n \tag{4.24}$$

$$+ \overline{\delta}\varepsilon_{n+1} \tag{4.25}$$

これを $f(x, y)$ の包含とする.

4.4.2 第二の方法

この方法では、変数 x, y をそれぞれ

$$x = x_0 + x_1\varepsilon_r, \quad y = y_0 + y_2\varepsilon_r \tag{4.26}$$

と初期化し、$z = x + y$ を

$$z = z_0 + (|x_1| + |y_2| + |z_r|)\varepsilon_r \tag{4.27}$$

で評価する。ただし、

$$\begin{aligned} \mathtt{down}(); \quad & \underline{\mathtt{z_0}} = \mathtt{x_0} + \mathtt{y_0}; \\ \mathtt{up}(); \quad & \overline{\mathtt{z_0}} = \mathtt{x_0} + \mathtt{y_0}; \\ & \mathtt{z_0} = (\underline{\mathtt{z_0}} + \overline{\mathtt{z_0}})/2; \\ & \mathtt{z_r} = \mathtt{z_0} - \underline{\mathtt{z_0}}; \end{aligned} \tag{4.28}$$

丸め誤差に関する dummy 変数 ε_r の係数は常に絶対値和をとる点に注意する (たとえば、$x - y$ の誤差項は $(|x_1| + |y_2| + |z'_r|)\varepsilon_r$)。また、$z^2$ のような非線形演算の場合、$z^2 = w_0 + w_1\varepsilon_1$ のように、追加する dummy 変数 ε_1 の項に丸め誤差を足し込む。z^2 を計算する過程で生じる丸め誤差は, ε_1 が打ち消される場合に限って打ち消すことができるのでこのようにしてよい.

Affine 形式

$$x_i = \mathtt{up}(\frac{x_i + \overline{x_i}}{2}), R_x = \mathtt{up}(\sum_{i=0}^{n} \mathtt{up}(\frac{\overline{x_i} - \underline{x_i}}{2})) \tag{4.29}$$

とすると

$$x = x_0 + x_1\varepsilon_1 + x_2\varepsilon_2 + \cdots + x_n\varepsilon_n + R_x\varepsilon_r \tag{4.30}$$

通常の区間表示への還元

$$[\mathtt{down}(x_0 - \mathtt{up}(\sum_{i=1}^{n} |x_i|) - R_x)), \mathtt{up}(x_0 + \mathtt{up}(\sum_{i=1}^{n} |x_i|) + R_x))] \tag{4.31}$$

線形計算

$$\underline{p_i} = \mathtt{down}(x_i + y_i) \tag{4.32}$$

$$\overline{p_i} = \mathtt{up}(x_i + y_i) \tag{4.33}$$

とすると

$$x+y = \mathtt{up}(\frac{\underline{\underline{p_0}}+\overline{p_0}}{2})+\mathtt{up}(\frac{\underline{\underline{p_1}}+\overline{p_1}}{2})\varepsilon_1+\cdots+\mathtt{up}(\frac{\underline{\underline{p_n}}+\overline{p_n}}{2})\varepsilon_n \quad (4.34)$$

$$+ \mathtt{up}(\sum_{i=0}^{n}\mathtt{up}(\frac{\overline{p_i}-\underline{\underline{p_i}}}{2})+R_x+R_y)\varepsilon_r \quad (4.35)$$

又,

$$\underline{q_i} = \mathtt{down}(x_i-y_i) \quad (4.36)$$

$$\overline{q_i} = \mathtt{up}(x_i-y_i) \quad (4.37)$$

とすると

$$x-y = \mathtt{up}(\frac{\underline{\underline{q_0}}+\overline{q_0}}{2})+\mathtt{up}(\frac{\underline{\underline{q_1}}+\overline{q_1}}{2})\varepsilon_1+\cdots+\mathtt{up}(\frac{\underline{\underline{q_n}}+\overline{q_n}}{2})\varepsilon_n \quad (4.38)$$

$$+ \mathtt{up}(\sum_{i=0}^{n}\mathtt{up}(\frac{\overline{q_i}-\underline{\underline{q_i}}}{2})+R_x+R_y)\varepsilon_r \quad (4.39)$$

次に,

$$\underline{z_i} = \min(\mathtt{down}(x_i\underline{\alpha}),\mathtt{down}(x_i\overline{\alpha})) \quad (4.40)$$

$$\overline{z_i} = \max(\mathtt{up}(x_i\underline{\alpha}),\mathtt{up}(x_i\overline{\alpha})) \quad (4.41)$$

とすると

$$[\underline{\alpha},\overline{\alpha}]x = \mathtt{up}(\frac{\underline{\underline{z_0}}+\overline{z_0}}{2})+\mathtt{up}(\frac{\underline{\underline{z_1}}+\overline{z_1}}{2})\varepsilon_1+\cdots+\mathtt{up}(\frac{\underline{\underline{z_n}}+\overline{z_n}}{2})\varepsilon_n \quad (4.42)$$

$$+ \mathtt{up}(\sum_{i=0}^{n}\mathtt{up}(\frac{\overline{z_i}-\underline{\underline{z_i}}}{2})+\max(|\underline{\alpha}|R_x,|\overline{\alpha}|R_x))\varepsilon_r \quad (4.43)$$

非線形単項計算

$$\underline{z1_i} = \min(\mathtt{down}(x_i\underline{a}),\mathtt{down}(x_i\overline{a})) \quad (4.44)$$

$$\overline{z1_i} = \max(\mathtt{up}(x_i\underline{a}),\mathtt{up}(x_i\overline{a})) \quad (4.45)$$

$$A = \mathtt{up}(\frac{\underline{\underline{z1_0}}+\overline{z1_0}}{2})+\mathtt{up}(\frac{\underline{b}+\overline{b}}{2}) \quad (4.46)$$

とすると

$$
\begin{aligned}
[\underline{a},\overline{a}]x+[\underline{b},\overline{b}]+\overline{\delta}\varepsilon_{n+1} &= \mathtt{up}(\frac{\mathtt{down}(A)+\mathtt{up}(A)}{2})+\mathtt{up}(\frac{\underline{\underline{z1_1}}+\overline{z1_1}}{2})\varepsilon_1 && (4.47)\\
&+ \cdots+\mathtt{up}(\frac{\underline{\underline{z1_n}}+\overline{z1_n}}{2})\varepsilon_n+\mathtt{up}(\sum_{i=0}^{n}\mathtt{up}(\frac{\overline{z1_i}-\underline{\underline{z1_i}}}{2}) && (4.48)\\
&+ \max(\mathtt{up}(|\underline{a}|R_x),\mathtt{up}(|\overline{a}|R_x)) && (4.49)\\
&+ \mathtt{up}(\frac{\overline{b}-\underline{b}}{2})+\overline{\delta}+\mathtt{up}(\frac{\mathtt{up}(A)-\mathtt{down}(A)}{2}))\varepsilon_{n+1} && (4.50)
\end{aligned}
$$

これを $f(x)$ の包含とする.

非線形二項計算

$$
\begin{aligned}
\underline{z2_i} &= \min(\mathtt{down}(y_i\underline{b}),\mathtt{down}(y_i\overline{b})) && (4.51)\\
\overline{z2_i} &= \max(\mathtt{up}(y_i\underline{b}),\mathtt{up}(y_i\overline{b})) && (4.52)\\
P &= \mathtt{up}(\frac{\underline{\underline{z1_0}}+\overline{z1_0}}{2})+\mathtt{up}(\frac{\underline{\underline{z2_0}}+\overline{z2_0}}{2})+\mathtt{up}(\frac{\underline{c}+\overline{c}}{2}) && (4.53)\\
Q_i &= \mathtt{up}(\frac{\underline{\underline{z1_i}}+\overline{z1_i}}{2})+\mathtt{up}(\frac{\underline{\underline{z2_i}}+\overline{z2_i}}{2}) && (4.54)
\end{aligned}
$$

とすると

$$
\begin{aligned}
[\underline{a},\overline{a}]x+[\underline{b},\overline{b}]y+[\underline{c},\overline{c}]+\overline{\delta}\varepsilon_{n+1} &= \mathtt{up}(\frac{\mathtt{down}(P)+\mathtt{up}(P)}{2})+\mathtt{up}(\frac{\mathtt{down}(Q_1)+\mathtt{up}(Q_1)}{2})\varepsilon_1 && (4.55)\\
&+ \cdots+\mathtt{up}(\frac{\mathtt{down}(Q_n)+\mathtt{up}(Q_n)}{2})\varepsilon_n && (4.56)\\
&+ \mathtt{up}(\mathtt{up}(\frac{\mathtt{up}(P)-\mathtt{down}(P)}{2})+\sum_{i=1}^{n}\mathtt{up}(\frac{\mathtt{up}(Q_i)-\mathtt{down}(Q_i)}{2}) && (4.57)\\
&+ \mathtt{up}(\sum_{i=0}^{n}\mathtt{up}(\frac{\overline{z1_i}-\underline{\underline{z1_i}}}{2})+\max(\mathtt{up}(|\underline{a}|R_x),\mathtt{up}(|\overline{a}|R_x)) && (4.58)\\
&+ \mathtt{up}(\sum_{i=0}^{n}\mathtt{up}(\frac{\overline{z2_i}-\underline{\underline{z2_i}}}{2})+\max(\mathtt{up}(|\underline{b}|R_y),\mathtt{up}(|\overline{b}|R_y)) && (4.59)\\
&+ \mathtt{up}(\frac{\overline{c}-\underline{c}}{2})+\overline{\delta})\varepsilon_{n+1} && (4.60)
\end{aligned}
$$

これを $f(x,y)$ の包含とする.

4.4.3 第三の方法

変数 x,y をそれぞれ

$$x=x_0+x_1\varepsilon_1,\quad y=y_0+y_2\varepsilon_2 \tag{4.61}$$

と初期化する。ここで、$z = x + y = (x_0 + y_0) + x_1\varepsilon_1 + y_2\varepsilon_2$ となるが、線形演算 $x_0 + y_0$ によって生じる丸め誤差を新たに誤差項として追加するのが第三の方法である。すなわち、

$$z = z_0 + x_1\varepsilon_1 + y_2\varepsilon_2 + z_r\varepsilon_3 \tag{4.62}$$

この方法では、別々の線形演算ごとに別々の dummy 変数を用意する必要がある。又, この方法は, 演算を行う度に誤差項が増えていくので, 計算時間が著しく増大する恐れがある.

Affine 形式

$$x = x_0 + x_1\varepsilon_1 + x_2\varepsilon_2 + \cdots + x_n\varepsilon_n \tag{4.63}$$

通常の区間表示への還元

$$[\mathtt{down}(x_0 - \delta), \mathtt{up}(x_0 + \delta)], \delta = \mathtt{up}(\sum_{i=1}^{n} |x_i|) \tag{4.64}$$

線形演算

$$\begin{aligned}
x + y &= \mathtt{up}(\frac{\mathtt{down}(x_0 + y_0) + \mathtt{up}(x_0 + y_0)}{2}) && (4.65)\\
&+ \mathtt{up}(\frac{\mathtt{down}(x_1 + y_1) + \mathtt{up}(x_1 + y_1)}{2})\varepsilon_1 && (4.66)\\
&+ \cdots + \mathtt{up}(\frac{\mathtt{down}(x_n + y_n) + \mathtt{up}(x_n + y_n)}{2})\varepsilon_n && (4.67)\\
&+ \mathtt{up}(\sum_{i=0}^{n} \mathtt{up}(\frac{\mathtt{up}(x_i + y_i) - \mathtt{down}(x_i + y_i)}{2}))\varepsilon_{n+1} && (4.68)\\
x - y &= \mathtt{up}(\frac{\mathtt{down}(x_0 - y_0) + \mathtt{up}(x_0 - y_0)}{2}) && (4.69)\\
&+ \mathtt{up}(\frac{\mathtt{down}(x_1 - y_1) + \mathtt{up}(x_1 - y_1)}{2})\varepsilon_1 && (4.70)\\
&+ \cdots + \mathtt{up}(\frac{\mathtt{down}(x_n - y_n) + \mathtt{up}(x_n - y_n)}{2})\varepsilon_n && (4.71)\\
&+ \mathtt{up}(\sum_{i=0}^{n} \mathtt{up}(\frac{\mathtt{up}(x_i - y_i) - \mathtt{down}(x_i - y_i)}{2}))\varepsilon_{n+1} && (4.72)\\
[\underline{\alpha}, \overline{\alpha}]x &= \mathtt{up}(\frac{\min(\mathtt{down}(\underline{\alpha}x_0), \mathtt{down}(\overline{\alpha}x_0)) + \max(\mathtt{up}(\underline{\alpha}x_0), \mathtt{up}(\overline{\alpha}x_0))}{2}) && (4.73)
\end{aligned}$$

$$+ \quad \mathtt{up}(\frac{\min(\mathtt{down}(\underline{\alpha}x_1), \mathtt{down}(\overline{\alpha}x_1)) + \max(\mathtt{up}(\underline{\alpha}x_1), \mathtt{up}(\overline{\alpha}x_1))}{2})\varepsilon_1 + \cdots \tag{4.74}$$

$$+ \quad \mathtt{up}(\frac{\min(\mathtt{down}(\underline{\alpha}x_n), \mathtt{down}(\overline{\alpha}x_n)) + \max(\mathtt{up}(\underline{\alpha}x_n), \mathtt{up}(\overline{\alpha}x_n))}{2})\varepsilon_n \tag{4.75}$$

$$+ \quad \mathtt{up}(\sum_{i=0}^{n} \mathtt{up}(\frac{\max(\mathtt{up}(\underline{\alpha}x_i), \mathtt{up}(\overline{\alpha}x_i)) - \min(\mathtt{down}(\underline{\alpha}x_i), \mathtt{down}(\overline{\alpha}x_i))}{2}))\varepsilon_{n+1} \tag{4.76}$$

$$\tag{4.77}$$

非線形単項計算

$$\underline{s_i} = \min(\mathtt{down}(\underline{a}x_i), \mathtt{down}(\overline{a}x_i)) \tag{4.78}$$

$$\overline{s_i} = \max(\mathtt{up}(\underline{\alpha}x_i), \mathtt{up}(\overline{\alpha}x_i)) \tag{4.79}$$

$$B = = \mathtt{up}(\frac{\underline{s_0} + \overline{s_0}}{2}) + \mathtt{up}(\frac{\underline{b} + \overline{b}}{2}) \tag{4.80}$$

とすると

$$[\underline{a}, \overline{a}]x + [\underline{b}, \overline{b}] + \overline{\delta}\varepsilon_{n+1} = \mathtt{up}(\frac{\mathtt{down}(B) + \mathtt{up}(B)}{2}) + \mathtt{up}(\frac{\underline{s_1} + \overline{s_1}}{2})\varepsilon_1 \tag{4.81}$$

$$+ \quad \cdots + \mathtt{up}(\frac{\underline{s_n} + \overline{s_n}}{2})\varepsilon_n + \mathtt{up}(\sum_{i=0}^{n} \mathtt{up}(\frac{\overline{s_i} - \underline{s_i}}{2}) \tag{4.82}$$

$$+ \quad \mathtt{up}(\frac{\overline{b} - \underline{b}}{2}) + \overline{\delta} + \mathtt{up}(\frac{\mathtt{up}(B) - \mathtt{down}(B)}{2}))\varepsilon_{n+1} \tag{4.83}$$

$$\tag{4.84}$$

非線形二項計算

$$\underline{t_i} = \min(\mathtt{down}(\underline{b}y_i), \mathtt{down}(\overline{b}y_i)) \tag{4.85}$$

$$\overline{t_i} = \max(\mathtt{up}(\underline{b}y_i), \mathtt{up}(\overline{b}y_i)) \tag{4.86}$$

$$M = \mathtt{up}(\frac{\underline{s_0} + \overline{s_0}}{2}) + \mathtt{up}(\frac{\underline{t_0} + \overline{t_0}}{2}) + \mathtt{up}(\frac{\underline{c} + \overline{c}}{2}) \tag{4.87}$$

$$N_i = \mathtt{up}(\frac{\underline{s_i} + \overline{s_i}}{2}) + \mathtt{up}(\frac{\underline{t_i} + \overline{t_i}}{2}) \tag{4.88}$$

とすると

$$[\underline{a}, \overline{a}]x + [\underline{b}, \overline{b}]y + [\underline{c}, \overline{c}] + \overline{\delta}\varepsilon_{n+1} = \mathtt{up}(\frac{\mathtt{down}(M) + \mathtt{up}(M)}{2}) + \mathtt{up}(\frac{\mathtt{down}(N_1) + \mathtt{up}(N_1)}{2})\varepsilon_1 \tag{4.89}$$

$$+ \quad \cdots + \mathtt{up}(\frac{\mathtt{down}(N_n) + \mathtt{up}(N_n)}{2})\varepsilon_n \tag{4.90}$$

$$+ \quad \mathtt{up}(\mathtt{up}(\frac{\mathtt{up}(M) - \mathtt{down}(M)}{2}) + \sum_{i=1}^{n} \mathtt{up}(\frac{\mathtt{up}(N_i) - \mathtt{down}(N_i)}{2}) \tag{4.91}$$

$$+ \quad \mathtt{up}(\sum_{i=0}^{n} \mathtt{up}(\frac{\overline{s_i} - \underline{s_i}}{2})) + \mathtt{up}(\sum_{i=0}^{n} \mathtt{up}(\frac{\overline{t_i} - \underline{t_i}}{2})) \tag{4.92}$$

$$+ \quad \mathtt{up}(\frac{\overline{c} - \underline{c}}{2}) + \overline{\delta})\varepsilon_{n+1} \tag{4.93}$$

これを $f(x, y)$ の包含とする.

4.5 むすび

本章では, 3 つの, 丸め誤差を考慮した Affine 演算の方法を提案した. 次章では, この 3 つの方法による数値例を取り扱う.

第 5 章

数値例

5.1 はじめに

$x = 1/3, y = 1/15, z = x + y$ として,$z^2 - 0.8z$ の値を求める ($\text{真の値は } -0.16$). また,$z \in [a, b]$ において, z^2 は $(a + b)z - \frac{a^2+6ab+b^2}{8} + \frac{(a-b)^2}{8}\varepsilon_n$ で与えられる.

5.2 第一の方法

```
z^2 - 0.8z = [ -0.16000000000000025313, -0.15999999999999978129 ]
```

5.3 第二の方法

```
z^2 - 0.8z = [ -0.16000000000000022538, -0.15999999999999983680 ]
```

5.4 第三の方法

```
z^2 - 0.8z = [ -0.16000000000000019762, -0.15999999999999986455 ]
```

5.5 考察

第一の方法より第二の方法が, 第二の方法より第三の方法が, それぞれより良い結果を与えた. この結果から,Affine 多項式の係数をそのまま区間表示する方法は区間評価を甘くすることがわかる. また, この例で線形演算 $x + y$ で生じる丸め誤差はキャンセルされるので, その分を単独の誤差項で表わした第三の方法は, 第二の方法より良い結果を与えたことがわかる.

第 6 章

まとめ

本論文では第 2 章で精度保証付数値計算の基礎である区間演算の解説を行ない, 第 3 章から Affine Arithmetic の定義の解説を行なった.

そして第 4 章では丸め誤差を考慮した Affine 演算を 3 つ提案し, 第 5 章ではその数値例を示した. その結果, 第三の方法が現状で丸め誤差の取り扱いに関して最適な結果を与える方法であることがわかった. 今後は計算コストも勘案したうえで, どの方法が実用上もっとも有用であるか摸索したい.

第 7 章

謝辞

本研究を進めるに当たり常に正しい道を示し、終始丁寧な御指導及び御激励を賜り、その他多くの面でもいろいろと御面倒を見て下さり御助言を与えて下さいました柏木雅英 助教授に深く感謝致します。

また、卒業論文中間報告の際など機会のあるごとに御指導、御鞭撻を賜り、研究に方向正を与えて下さった大石進一 教授に深く感謝致します。

柏木研において柏木研助手 宮田孝富 氏には, 非線形の分野について熱心に御指導してくださり, また, 日頃の柏木研究室内でのあらゆる面において御教示頂いたことに, 心より感謝申し上げます.

柏木研修士課程 2 年高崎大輔 氏 , 柏木研修士課程 1 年吉田直史 氏には, 中間発表の準備の際に適切なアドバイスを頂き, また私のコンピュータに関する稚拙な疑問にも親身に御指導を賜り, 心より感謝申し上げます.

最後に, 意見の交換や協力などして下さいました柏木研究室学部 4 年の皆様, 黒崎俊行氏, 石田寛氏, 菊池智行氏, 西中川英氏, 古川周一氏, 宝迫敦氏, 宮島信也氏に心より感謝致します.

参考文献

[1] Marcus Vinícius A.Andrade, João L. D. Comba and Jorge Stolfi 『Affine Arithmetic』 INTERVAL '94, St. petersburg (Russia), March 5-10, 1994.

[2] 大石進一 著『数値計算』裳華房