

数値計算中の情報喪失について

日大 理工 永坂秀子

数値計算は有限桁によって計算されることは、今も昔も変りなく、丸め誤差は無視することが出来ない。従来、紙上に数値を置いて人目で追って計算していたときは、桁落ちして精度が悪くなつた数値を見れば必要に応じて桁数を適当に増加したり、他の計算法に移行して、最適な計算を選びながら計算することが出来た。現在は電子計算機によつて、その大部分の計算が処理されていい。電算機による計算では、計算過程の数値を逐一 Print out することは、その利用効率からいって不要ので、Print out は最小限に止めなければならず、数値のはんざか電算機のブラウカ・オーフス中で移動変化していく。そのため、すべての状態に適応して対処出来るよう計算手順を考えなければならないのは周知のことである。

従来、計算法における数値解析は、無限桁数値を前提として

の誤差解析であって、丸め誤差についての解析は至難のこととして、極く少數が case by case として扱われてゐるだけの一環したものがない。一方丸め誤差は、その計算法に従つて忠実に規則的に拡大、縮小、消滅して行くもので、丸め誤差を最小限にもつて行く計算法によれば、その誤差解析も容易になることが、いくつかの例によつてわかりました。そこでこの規則を見出すためには、先づ丸め誤差の変動と、その計算式の関係をパターンに分類してみたらと思ひ、計算の基本にもどつて考えて見ました。ここで特に注目したいのは、情報落ち現象と、情報恢復現象である。

§ 0. はじめ

§ 1. 四則演算誤差

§ 2. 演算における精度

1. 衍落ち
2. 情報落ち
3. 丸め誤差

§ 3. 精度落ちする計算パターンとその計算例

1. 精度落ちした数値での乗除算

2. 衍落ち
3. 情報落ち

§ 4. 精度恢復の計算パターンとその計算例

1. 情報落ち

2. 分母、分子の誤差項の約分

§ 0. はじめに 數値、誤差の定義をしておく。

真値: a , 誤差: Δa , 近似値: $\tilde{a} = a + \Delta a$,

絶対誤差: $\Delta a = \tilde{a} - a$, 相対誤差: $\frac{\Delta a}{a} = \frac{\tilde{a} - a}{a}$

§ 1. 四則演算誤差

1. 1. 加減算の絶対誤差は、それぞれの絶対誤差の和となる。

$$\Delta(a \pm b) = \Delta a \pm \Delta b, \quad (\text{符号同順})$$

1. 2. 乗(除)算の相対誤差は、それぞれの相対誤差の和(差)となる。

1. 3.

$$\frac{\Delta(a \cdot b)}{a \cdot b} = \frac{\Delta a}{a} + \frac{\Delta b}{b}, \quad \frac{\Delta\left(\frac{a}{b}\right)}{\frac{a}{b}} = \frac{\Delta a}{a} - \frac{\Delta b}{b}$$

1. 4. ベキ乗数の相対誤差は、底の相対誤差のベキ倍となる。

$$\frac{\Delta(a^m)}{a^m} = m \cdot \frac{\Delta a}{a}$$

§ 2. 演算における精度

2. 1. 戻落ちと精度

$$a \pm b = c \quad (|a| \neq |b|)$$

において左辺が 2 数の絶対値の差となるようなどき、 c の位
には a , b より下3。これを 戻落ち という。

$$1.234\underline{5}35 - 1.23678\underline{5} = -0.002\underline{2}50$$

以下の波形を誤差桁とする \tilde{c} は 3 衡々落ちして誤差は下3桁

$\tilde{a} \pm \tilde{b} = \tilde{c}$ で \tilde{c} は m 行々落ちしたとき

\tilde{c} の精度は、 $\{\min(a, b \text{ の有効桁数}) - m\}$ 行

\tilde{c} の誤差は、 $\{\max(a, b \text{ の誤差桁数}) + m\}$ 行

桁落ちした桁数だけ丸め誤差が下の位から上って来る。

(2.1)

2.2. 情報落ち

$$a \pm b = c \quad (|a| > |b| \text{ 又は } |a| \ll |b|)$$

有限桁計算では絶対値の小さい数の情報はその order (位数) 差だけ情報が落ちる。これを情報落ちという。

(1) 誤差を含む数が情報落ちすると精度が良くなるとある。

(2) 情報落ちには困るとき、高精度計算で救われるところがある。

(2.2)

2.3. 丸め誤差と計算結果誤差

実数を扱う数値計算においては、演算毎に丸め誤差が殆んど入ってくる。前節にあげた桁落ち、情報落ちの誤差評価にあたっても §1. 節の演算誤差評価が基盤になつてゐる。しかし 1つ1つの数値の誤差だけを追つていたのでは、目的の解の誤差は全く出せない。計算過程での誤差の変動と、その数値との相対関係を知りながら跡して追跡して行かなければならぬ。

§ 3. 精度落ちする計算パターンとその計算例

3.1. 精度落ちした数での乗除算

$$\hat{e} \cdot \hat{a} + \hat{e} \cdot \hat{b} = \hat{c}, \quad \hat{a} / \hat{e} + \hat{e} \cdot \hat{b} = \hat{x}$$

この相対誤差が \hat{a} , \hat{b} の相対誤差に比べ大きいときは, \hat{c} , \hat{x} の相対誤差は \hat{a} の相対誤差と同じになり精度が悪くなる。
しかし, 術落ちしないときは \hat{a}, \hat{b} の情報は失われない。

3.2. 術落ちの起るパターン

$$(1) \quad y = \sqrt{a^2 + b^2} - |a| \quad (|a| > |b|)$$

この場合は, 分子の有理化によって, 術落ちによる精度落ちは防げない。

$$y = \frac{|b|}{\sqrt{a^2 + b^2} + |a|}$$

[例 1] 二根が絶対値で大きく異なる二次方程式

$$ax^2 + bx + c = 0$$

$$(解法 I) \quad x_{1,2} = \frac{-b}{2a} \pm \sqrt{\left(\frac{b}{2a}\right)^2 - \frac{c}{a}}$$

この計算式では複号の加減算で, 一方が術落ちを起す。

$$(解法 II) \quad \begin{cases} b > 0 \text{ のとき} & x_1 = \frac{-b}{2a} - \sqrt{\left(\frac{b}{2a}\right)^2 - \frac{c}{a}} \\ b < 0 \text{ のとき} & x_1 = \frac{-b}{2a} + \sqrt{\left(\frac{b}{2a}\right)^2 - \frac{c}{a}} \end{cases}$$

$$x_2 \text{ は分子を有理化し} \quad x_2 = \frac{c}{a} / x_1$$

(この式は, $x_1 \cdot x_2 = c/a$ より直ちに導き出せる。)

[数値例1]

(单精度浮動10進約7桁計算)

二根を $x_1 = \sqrt{3}$, $x_2 = \sqrt{2} \times 10^{-4}$ とし, この近似値を

$$X_1 = 0.1732050807568877D\ 01, \quad X_2 = 0.1414213562373095D-03$$

で支え倍桁計算による係数より有効7桁にとって係数とし,

$$A = 0.1000000E\ 01, \quad B = -0.1732192E\ 01, \quad C = 0.2449488E-03$$

により(解法I)の解 $X_1 = 0.1732050E\ 01, \quad X_2 = 0.1414418E-03$ (解法II)の解 $X_1 = 0.1732050E\ 01, \quad X_2 = C/A/X_1 = 0.1414213E-03$

$$*) \quad x_2 = \frac{-b}{2a} - \sqrt{D} = 0.866096 - 0.8659546 = 0.000141414xxx$$

となって、減算で3桁けたあちしたので丸め誤差が下から3桁上つて来た。

[数値例2]

(单精度浮動計算)

二根を $x_1 = \sqrt{3}$, $x_2 = \sqrt{2} \times 10^{-8}$ とし, この近似値を

$$X_1 = 0.1732050807568877D\ 01, \quad X_2 = 0.1414213562373095D-07$$

で支え倍桁計算による係数より有効7桁にとって係数とし,

$$A = 0.1000000E\ 01, \quad B = -0.1732050E\ 01, \quad C = 0.2449490E-07$$

により(解法I)の解 $X_1 = 0.1732049E\ 01, \quad X_2 = 0.5960464E-07$ (解法II)の解 $X_1 = 0.1732049E\ 01, \quad X_2 = C/A/X_1 = 0.1414215E-07$ 2根のorder差が計算桁数(7桁)を越えているので(解法I)では x_2 は完全に誤差のみとなってしまったが、(解法II)では正しく出ている。

「コメント」

(解法I)の解 $x_2 (|x_1| >> |x_2|)$ から Newton法で反復計算すると、計算桁数まで正しい根が得られる。このことからも(解法I)はよえられた係数の限界まで計算する解法でなく、まだ改善の余地のあることを暗示している。

$$(2) \quad y = \sqrt{a^2 - b^2}, \quad (|a| \neq |b|)$$

$$\tilde{y} = \sqrt{1.732050^2 - 3.000000} = \sqrt{0.000003} = 0.001732050$$

(浮動小数点計算)

- (i) $|a| \neq |b| \neq n \times 10^k$ ($1 \leq n < 10$) のとき $a^2 - b^2$ の計算で m 行おちしたとき $\sqrt{a^2 - b^2}$ の誤差は約 $10^{2-l+m/2}$ となる。 (l は計算桁数)
- (ii) a, b が正しい値のときは、桁うちの桁数だけ桁数を増加して計算すれば精度はよくなる。

3.1

[数値例 3] 等根に近い複素根をもつ2次方程式の解

複素根を $\sqrt{3} \pm \sqrt{2} \times 10^{-6}i$ とし、この近似値を

$0.1732050807568877D\ 01 \pm 0.1414213562373095D-05\ I$ とし

(a) 倍精度計算による係数より有効7桁にとって係数とし

$$A = 0.1000000E\ 01, \quad B = -0.3464101E\ 01, \quad C = 0.3000000E\ 01$$

により单精度計算の解 $0.1732050E\ 01 \pm 0.1953125E-02\ I$

倍精度計算の解 $0.1732050418853760D\ 01 \pm 0.1160408770803387D-02\ I$

(b) 倍精度計算による係数 DA= 0.1000000000000000D 01

$$DB = -0.3464101615137754D\ 01$$

$$DC = 0.3000000000001996D\ 01$$

により
倍精度計算の解 $0.1732050807568877D\ 01 \pm 0.1414197919868275D-05\ I$

i) 係数が单精度のときは、虚数部は前記のとおりに示すように根号内の最後の桁まで桁うちして完全に誤差となる。このときは倍精度計算しても精度は上らず、 10^{-6} の誤差が開平で、 10^{-3} まで上って来ている。

ii) 倍精度係数による倍精度計算では、虚数部は 10^{+0} から 12 行おちし、計算桁数 17 行程度なので、誤差は $10^{0-17+12/2} = 10^{-11}$ まで上って来ている。

3.3. 情報落ちで精度が悪くなる例

$$a^m \pm b^m \quad (m > 0, |a| \neq |b|)$$

$|a| > |b|$ のとき ($|a| < |b|$ のときは a と b を入れかえればよい), b の情報は有限桁計算のため下の桁は落されてしまつて真値が出なくなる。

$$|a| > |b| \times 10^{-L/m} \quad (L \text{ は計算桁数})$$

のとき, b の情報は完全に失われる。よって m が大きいときは特に注意しなければならない。

[数値例4] (浮動7桁計算)

$$\begin{aligned} 5.000000^3 + 0.030000000^3 &= 125.0000 + 0.00002700000 \\ &= 125.0000 \end{aligned}$$

[例2]

行列の固有値解法の Jacobi 法において, 回転角を θ としてとき, $\cos\theta$ の計算は次式でなされる。

$$\cos\theta = \sqrt{\frac{1}{2} \left\{ 1 + \frac{|a_{pp} - a_{gg}|/2}{\sqrt{\{(a_{pp} - a_{gg})/2\}^2 + a_{pg}^2}} \right\}} \quad \dots (1)$$

$a_{pp} \neq a_{gg}$ (対角要素) のとき, a_{pg} (非対角要素) が

$$\max\{|a_{pp}|, |a_{gg}|\} > |a_{pg}| \times 10^{L/2} \quad (L \text{ は計算桁数})$$

となると, $\sqrt{\{(a_{pp} - a_{gg})/2\}^2 + a_{pg}^2} = |a_{pp} - a_{gg}|/2$ となつて

(1)式の $\cos\theta = 1$ となり, $\theta = 0$ となつて回転は止つてしまふ。

すなわちこのとき a_{pg} の情報は完全にみとまれている。

§4. 精度恢復の計算パターンとその計算例

4.1. 情報落ちによる精度恢復

$$(1) \quad y = |a| - \sqrt{a^2 \pm b^2} \quad (|a| > |b|)$$

この形は3.2.節の(1)と全く同じであるが、 b について考えるときは、 b の情報は根号内の計算で計算桁数から落されてしまう。二つとも

a, b は単精度のまゝ、 b の情報落ちの桁数だけ桁数を増して計算すれば、 \tilde{y} は単精度の値が得られる

(4.1)

「証明」

a, b を単精度のまゝ桁数を増して計算することは、それぞれ最後の桁に誤差が入った数で高精度計算をしていくことになる。ゆえに $\hat{a} = a + \Delta a$, $\hat{b} = b + \Delta b$ として考へればよい。

$$\begin{aligned} \tilde{y} &= |a + \Delta a| - \left\{ (a + \Delta a)^2 + (b + \Delta b)^2 \right\}^{\frac{1}{2}} \\ &= |a + \Delta a| - |a + \Delta a| \left\{ 1 + \left(\frac{b + \Delta b}{a + \Delta a} \right)^2 \right\}^{\frac{1}{2}} \\ &\doteq - \frac{1}{2} \frac{(b + \Delta b)^2}{|a + \Delta a|} + \frac{1}{4} \frac{(b + \Delta b)^4}{|a + \Delta a|^3} - \dots \end{aligned}$$

この計算で2行目から最後の式に移るととき、 $|a + \Delta a|$ が完全に Δa まで含めて除去され、あとには b^2/a の形が残されていく。 b の情報が完全に保存されなければ、 Δa は a の最後の桁で丸めの誤差程度であるので、逆に Δa はどんな数でもよいことになる。そしてその Δa が b の情報の運動段級をしていく。

[数値例5] 2根の絶対値が大きくちがう2次方程式

[数値例1], [数値例2]を係数は単精度のまゝ(解法I)で倍筋計算する。係数は7桁目に誤差があるのに、二根とも解は単精度の桁数だけ精度が得られている。

二根 $x_1 = \sqrt{3}$, $x_2 = \sqrt{2} \times 10^{-4}$ の近似値

$X_1 = 0.1732050807568877D\ 01$, $X_2 = 0.1414213562373095D-03$
より係数 $A = 0.1000000E\ 01$, $B = -0.1732192E\ 01$, $C = 0.2449488E-03$

✓解 $X_1 = 0.1732050618230121D\ 01$, $X_2 = 0.1414212596248659D-03$

二根 $x_1 = \sqrt{3}$, $x_2 = \sqrt{2} \times 10^{-8}$ の近似値

$X_1 = 0.1732050807568877D\ 01$, $X_2 = 0.1414213562373095D-07$
より係数 $A = 0.1000000E\ 01$, $B = -0.1732050E\ 01$, $C = 0.2449490E-07$

✓解 $X_1 = 0.1732049927874460D\ 01$, $X_2 = 0.1414214141626236D-07$

(2) $a \pm \tilde{e}$ (\tilde{e} は誤差を含む数で $|\tilde{e}| \ll |a|$)

[例13] a_{11} が他の a_{ij} に比し絶対値が小さくとき

$$\begin{aligned} \text{真値} \quad & \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = \begin{vmatrix} 0.0010 & 2.2222 \\ 3.3333 & 4.4444 \end{vmatrix} \\ & = 0.0044444 - 7.40725926 \\ & = -7.40281486 \end{aligned}$$

a_{11} が誤差を含み $\tilde{a}_{11} = 0.001053$ となつたとき, 浮動5桁計算によると

$$\begin{vmatrix} \tilde{a}_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = \begin{vmatrix} 0.0046800 & -7.4073 \\ -7.4026 & \end{vmatrix}$$

となり相対誤差は減少する。

[例4]

代数方程式 $f(x) \equiv \sum_{i=0}^n a_i x^i = 0$
の絶対値最小の根 x_1 が求まつたとき

$$f_1(x) = f(x)/(x-x_1) = \sum_{i=0}^{n-1} b_i x^i$$

の係数 b_i を求めると

$$\begin{cases} b_n = 0 & \text{とおき} \\ b_i = a_{i+1} + b_{i+1} \cdot x_1 & (i=n, n-1, \dots, 0) \end{cases}$$

(\therefore より b_i を求めれば $|a_{i+1}| > |b_{i+1}x_1|$ となることが多く、 a_i の情報は保存され、 x_1 の誤差は情報おちとよって b_i の精度は悪くならぬ)。

[数値例6]

$$f(x) = x^3 - 103.10x^2 + 212.30x - 20.2 = (x-0.1)(x-2.0)(x-101.0)$$

$x_1 = 0.1$ の近似値 $\tilde{x}_1 = 0.100\bar{1}$ (\therefore より $f_1(x)$ を求めよ。
(浮動小数点計算))

a_i	1.0	-103.10	212.30	-20.2	
		0.100 $\bar{1}$	-10.310	20.219	
b_i	1.0	-103.00	201.99	0.019	$= f(\tilde{x}_1)$

$$f_1(x) = x^2 - 103.00x + 201.99$$

$$f(x)/(x-0.1) = x^2 - 103.00x + 202.00$$

「エメント」

絶対値の小さい根から求め、この[例4]の解法によつて次
数を下げて順次大きい根を求めて行くと三根とも正しく出る
が、絶対値の大きい根から、この解法で次数を下げて行くと
浮動小数点計算で $x_3 = 101.000\bar{1}$, $x_2 = 1.999\bar{534}$, $x_1 = 0.1004567$ となる。

[例5] 代数方程式 $f(x) = \sum_{i=0}^n a_i x^i = 0$

の絶対値最大根 x_n が求まつたとき

$$f_n^*(x) = f(x) / \{x \cdot (\frac{1}{x} - \frac{1}{x_n})\} = \sum_{i=0}^{n-1} b_i x^i$$

の係数を求めるとき

$$b_0 = a_0 \quad \text{とおき}$$

$$b_i = a_i + b_{i-1} / x_n \quad (i=1, 2, \dots, n-1)$$

により b_i を求めれば、 $|a_i| > |b_{i-1} / x_n|$ となる、 a_i の情報は落されず b_i は精度がよくなる。

[数值例17]

$$f(x) = x^3 - 103.10x^2 + 212.30x - 20.2$$

$$f_1(x) = f(x) / (x - 101.00) = x^2 - 2.10x + 0.20$$

のとき $x_n = 101.00$ の近似値を $\tilde{x}_n = 101.01$ とする。

a_i	1.0	-103.10	212.30	-20.2
	0.99990	2.0998	-0.19998	
b_i	0.00010	-101.00	212.10	-20.2

$$f(\tilde{x}_n) / \tilde{x}_n^3$$

$$f_n^*(x) = -101.00x^2 + 212.10x - 20.2$$

$$= -101.00(x^2 - 2.10x + 0.20) = -101.00 \times f_1(x)$$

「コメント」

この解法は、絶対値最大の根により次数を下げるまでの割算で、絶対値最小の根により次数を下げるときは、高次の項より割つて行く従来の方法によらなければならぬ。

$f(x)$ を x の 2 次式で割つて複素根を求める方法に McAuley 法がある。Bairstow 法が高次の項より 2 次式で割つて行くのに対し、この方法は低次の方からの割算による計算法で、1 次式の割り算のときが丁度この計算となつていて。

[例6] 代数方程式の近接根の誤差のcancel

近接根をもつ代数方程式の Newton 法において、その解法手づきにおいて、丸め誤差を最小にするようにして求めた解は、勿論近接根は精度が悪くなるが、その誤差は相手の近接根と相殺して、近接根以外の根の精度には影響しないとする。

[数値例 11]

$$f(x) \equiv x^5 - 27.001x^4 + 257.026x^3 - 997.23x^2 + 1326.766x - 560.56 = 0$$

i	真値 x_i	解 \tilde{x}_i	$\varepsilon_i = \tilde{x}_i - x_i$	解 x_i^*	$\varepsilon_i^* = x_i^* - x_i$
1	1.000	1.000 0103	103.0×10^{-7}	1.000 0398	398.0×10^{-7}
2	1.001	1.000 9897	-103.0	1.000 9601	-399.0
3	7.000	7.000 0058	58.0	7.000 0021	21.0
4	8.000	7.999 9942	-58.0	7.999 9987	-13.0
5	10.000	9.999 9998	-2.0	9.999 9990	-10.0

解 \tilde{x}_i は [例4] に示す解法に従って次数を下げた。

解 x_i^* は [例5] に示す解法に従って次数を下げた。

2つの解の誤差を見ると、ともに x_1 と x_2 の誤差、 x_3 と x_4 の誤差が互に打ち消し合って誤差の和は、それだけ $\sum_{i=1}^5 \varepsilon_i = -2.0 \times 10^{-7}$, $\sum_{i=1}^5 \varepsilon_i^* = -3.0 \times 10^{-7}$ となつていて、ともに計算結果の最後の桁に止つていい。すなはちほとんどゼロになつていいといえる。

(解析)

$$f(x) = x^5 + a_4 x^4 + a_3 x^3 + a_2 x^2 + a_1 x + a_0$$

$$f_1(x) = f(x) / (x - x_1) = x^4 + b_3 x^3 + b_2 x^2 + b_1 x + b_0$$

とおく。 x_1 が絶対値最小の根のとき [例4] に示す解法により
 x の次数の高い方から割って 3 で $\frac{1}{x_1}$ すなはち

$$b_4 = 1 \quad \text{とし} \quad$$

$$b_i = a_{i+1} + b_{i+1} \cdot x_1 \quad (i = 3, 2, 1, 0)$$

一方 a_i は根と係数の関係より

$$\begin{aligned} a_4 &= -x_1 - (\underline{x_2 + x_3 + x_4 + x_5}) \\ a_3 &= x_1 (x_2 + x_3 + x_4 + x_5) + (\underline{x_2 x_3 + x_2 x_4 + x_2 x_5 + x_3 x_4 + x_3 x_5 + x_4 x_5}) \\ a_2 &= -x_1 (x_2 x_3 + x_2 x_4 + x_2 x_5 + x_3 x_4 + x_3 x_5 + x_4 x_5) \\ &\quad - (\underline{x_2 x_3 x_4 + x_2 x_3 x_5 + x_2 x_4 x_5 + x_3 x_4 x_5}) \\ a_1 &= x_1 (x_2 x_3 x_4 + x_2 x_3 x_5 + x_2 x_4 x_5 + x_3 x_4 x_5) + \underline{x_2 x_3 x_4 x_5} \\ a_0 &= -x_1 x_2 x_3 x_4 x_5 \end{aligned}$$

とすると $\underline{\quad}$ の部分が b_i である。

今 $\tilde{x}_1 = x_1 + \varepsilon$ のとき $f_1(x)$ の係数を \tilde{b}_i とするとき

$$\tilde{b}_3 = -\{(x_2 - \varepsilon) + x_3 + x_4 + x_5\}$$

$$\tilde{b}_2 = \{(x_2 - \varepsilon)(x_3 + x_4 + x_5) + x_3 x_4 + x_3 x_5 + x_4 x_5\} + \varepsilon(x_1 - x_2) + \varepsilon^2$$

$$\tilde{b}_1 = -\{(x_2 - \varepsilon)(x_3 x_4 + x_3 x_5 + x_4 x_5) + x_3 x_4 x_5\} - \varepsilon(x_1 - x_2)(x_3 + x_4 + x_5) - \varepsilon^2(x_3 + x_4 + x_5)$$

$$\tilde{b}_0 = \{(x_2 - \varepsilon)x_3 x_4 x_5\} + \{\varepsilon(x_1 - x_2) + \varepsilon^2\}(x_3 x_4 + x_3 x_5 + x_4 x_5)$$

となる $\underline{\quad}$ の部分が計算結果より落されて、 $f_1(x)$ は $x_1^4 + (x_2 - \varepsilon)$, x_3, x_4, x_5 を根とする方程式となつてゐる。

4.2. 分母, 分子の誤差項の約分による精度恢復

$$\frac{\tilde{e} \cdot \tilde{a}}{\tilde{e} \cdot \tilde{b}} \doteq \frac{\tilde{a}}{\tilde{b}} \quad \dots (4.2)$$

[数値例 8]

$$\text{真 值} = \frac{1.2 \times 4.9876}{1.2 \times 9.9752} = 0.5$$

各項にそれぞれ誤差を入れて計算すると

$$\begin{aligned} \text{近似値} &= \frac{1.2530 \times 4.9875}{1.2530 \times 9.9752} = \frac{6.2493}{12.499} \\ &= 0.49998 \quad (\text{浮動} 8\text{桁計算}) \\ &= 0.4999799507 \quad (\text{浮動} 10\text{桁計算}) \end{aligned}$$

近似値の計算は 1. の誤差評価に従えば、分母, 分子の精度は 2 桁である。さらにこの 2 数の除算であるから結果の有効精度は 2 桁となる。ところが実際は \tilde{a}/\tilde{b} の精度 4 桁が得られている。ところで精度が出るからといって倍桁計算して \tilde{a}, \tilde{b} の誤差により 5 桁計算と同程度の精度となる。

(証明)

$$\begin{aligned} \frac{(e + \Delta e)(a + \Delta a)}{(e + \Delta e)(b + \Delta b)} &\doteq \frac{ea(1 + \frac{\Delta a}{a} + \frac{\Delta e}{e})}{eb(1 + \frac{\Delta b}{b} + \frac{\Delta e}{e})} \\ &= \frac{a}{b} \left(1 + \frac{\Delta a}{a} + \frac{\Delta e}{e}\right) \left\{1 - \left(\frac{\Delta b}{b} + \frac{\Delta e}{e}\right) + \left(\frac{\Delta b}{b} + \frac{\Delta e}{e}\right)^2 - \dots\right\} \\ &\doteq \frac{a}{b} \left(1 + \frac{\Delta a}{a} - \frac{\Delta b}{b}\right) \end{aligned}$$

となり $e, \Delta e$ は完全に消滅して、結果には影響せず $\Delta a, \Delta b$ が相対誤差で割算して来る。よって \tilde{e} はどのような数値が来てよいことになる。

[例] 7] a_{11} が他の a_{ij} に比し絶対値が小さいとき

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11} \cdot \left(a_{22} - \frac{a_{12} \cdot a_{21}}{a_{11}} \right) \quad \dots (4.3)$$

(による計算を参考).

P10 / [例] 3 の数値で、浮動 5 行計算で計算してみる.

$$\begin{aligned} \begin{vmatrix} \tilde{a}_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} &= 0.001053 \times \left(4.4444 - \frac{2.2222 \times 3.3333}{0.001053} \right) \\ &= 0.001053 \times (4.4444 - 7034.4) \\ &= 0.001053 \times 7030.0 = 7.4026 \quad \dots (4.4) \end{aligned}$$

§1. の誤差評価に従うと上記の 7.4026 が誤差となる。ところが実際は 7.4026 で最後の桁だけが誤差となる。

このことは (4.3) 式の右辺で $|a_{11} \cdot a_{22}| \ll |a_{11} \cdot \frac{a_{12} \cdot a_{21}}{a_{11}}|$ であるため計算結果に影響するだけ $a_{11} \cdot \frac{a_{12} \cdot a_{21}}{a_{11}}$ である。 \tilde{a}_{11} に多くの誤差が入っていても、P15 / (4.2) 式によつて、その誤差は Cancel されて高々計算桁最後の丸め誤差に止まつてしまつ。

この例では $|a_{11} \cdot a_{22}|$ が小さいため、その誤差は情報おちして、前記丸め誤差だけが残されてい。

更に浮動 10 行計算を二、三みてみる。

$$\begin{aligned} \begin{vmatrix} \tilde{a}_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} &= 0.001053 \times (4.4444 - 7034.434245) \\ &= 0.001053 \times 7029.989845 \\ &= 7.402579307 \quad \dots (4.5) \end{aligned}$$

§1. の誤差評価に従えば相対誤差は 2 行しかないが実際は 4 行出でる。この場合 $|\tilde{a}_{11} \cdot a_{22}|$ が $|a_{11} \cdot \frac{a_{21} \cdot a_{12}}{a_{11}}|$ の絶対誤差より優越しているため 10 行計算しても 5 行程度の精度に止まつてしまつ。

[例 8] 情報落ちと高精度計算

3次の行列式において、 \tilde{a}_{ii} が他の要素に比し絶対値が小さく、相対誤差が大きいときは、高精度計算によって、卓精度の解が得られる。

[数值例 9]

$$|A| = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = \begin{vmatrix} 0.0011 & 2.2222 & 4.4444 \\ 3.3333 & 5.5555 & 6.6666 \\ 7.7777 & 8.8888 & 9.9999 \end{vmatrix} \doteq -19.2078874414$$

$\tilde{a}_{11} = 0.001153$ のとき、浮動小数点計算による計算過程

			要素名と数值の内容		
0.001153	2.2222	4.4444	\tilde{a}_{11}	a_{12}	a_{13}
3.3333	5.5555 6424.3	6.6666 12849.	a_{21}	a_{22} $a_{21}a_{12}/\tilde{a}_{11}$	a_{23} $a_{21}a_{13}/\tilde{a}_{11}$
7.7777	8.8888 14990.	9.9999 29980.	a_{31}	a_{32} $a_{31}a_{12}/\tilde{a}_{11}$	a_{33} $a_{31}a_{13}/\tilde{a}_{11}$
0	-6418.7	-12842.		\tilde{a}'_{22}	\tilde{a}'_{23}
0	-14981. 29973.	-29970.		\tilde{a}'_{32} $\tilde{a}'_{32}\tilde{a}'_{23}/\tilde{a}'_{22}$	\tilde{a}'_{33}
0	0	3.0000			\tilde{a}''_{33}

$$|A| \doteq \tilde{a}_{11} \cdot \tilde{a}'_{22} \cdot \tilde{a}''_{33} = -22.202$$

$$\begin{aligned} T = T - L & \quad \tilde{a}'_{ij} = a_{ij} - \frac{a_{i1} \cdot a_{1j}}{\tilde{a}_{11}} \\ & \quad \tilde{a}''_{ij} = \tilde{a}'_{ij} - \frac{\tilde{a}'_{i2} \cdot \tilde{a}'_{2j}}{\tilde{a}'_{22}} \end{aligned}$$

\tilde{a}''_{33} が完全に誤差に依って 3 行めの $|A|$ の値は T と \tilde{T} で 3.

$\tilde{a}_{11} = 0.001153$ のとき、浮動10桁(倍桁)計算過程

			要素名		
a_{11}	a_{12}	a_{13}	\tilde{a}_{11}	a_{12}	a_{13}
3.3333	5.5555 <u>6424.335873</u>	6.6666 <u>12848.67174</u>	a_{21}	a_{22}	a_{23}
7.7777	8.8888 <u>14990.11704</u>	9.9999 <u>29980.23406</u>	a_{31}	a_{32}	a_{33}
0	-6418.780373	-12842.00514	\tilde{a}'_{22}	\tilde{a}'_{23}	
0	-14981.22824	-29970.23416 <u>29972.82955</u>	\tilde{a}'_{32}	\tilde{a}'_{33}	
0	0	2.59539			\tilde{a}''_{33}

$$|A| \div \tilde{a}_{11} \cdot \tilde{a}'_{22} \cdot \tilde{a}''_{33} = 19.20810186$$

の誤差範囲は、各演算毎に §1. の演算評価によつても
のであるが、行列式の値は、 $a_{11} = 0.0011$ のときの値とくら
べると、有効5桁まで一致してい3.

(解釈)

1°) 2次の行列式の精度

$$\begin{vmatrix} \tilde{a}_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = \tilde{a}_{11} \times \tilde{a}'_{22} \quad \text{の値は p14/(4.4) 式で示したよう}$$

に、 \tilde{a}'_{22} (= 1 ± \tilde{a}_{11}) の誤差が含まれ精度は悪くなつてい3か;
 \tilde{a}_{11} を乗すことによつて誤差の cancel が起り、計算結果は
計算桁最後の行に丸め誤差が含まれる程なるとなる。数値にて示

すと

$$\text{真値} : a_{11} \times a_{22}' = -7.40114821$$

$$\begin{aligned}\text{近似値} : \tilde{a}_{11} \times \tilde{a}_{22}' &= -7.4008 \quad (\text{浮動}5\text{桁計算}) \\ &= -7.400853770 \quad (\text{浮動}10\text{桁計算})\end{aligned}$$

となり、5桁目に丸め誤差が入って来るのは、5桁計算でも
10桁計算でも同じである。(P16/(4.5)式で示したように $\tilde{a}_{11} \cdot a_{22}$
の誤差でおさえられている。)

2°) \tilde{a}_{33}'' の精度

\tilde{a}_{33}'' までの計算過程を逆にたどって、はじめの要素で考えて見よ。

$$\tilde{a}_{33}'' = \tilde{a}_{33}' - \frac{\tilde{a}_{32}' \cdot \tilde{a}_{23}'}{\tilde{a}_{22}'} = \left(a_{33} - \frac{a_{31} \cdot a_{13}}{\tilde{a}_{11}} \right) - \frac{\left(a_{32} - \frac{a_{31} \cdot a_{12}}{\tilde{a}_{11}} \right) \left(a_{23} - \frac{a_{21} \cdot a_{13}}{\tilde{a}_{11}} \right)}{\left(a_{22} - \frac{a_{21} \cdot a_{12}}{\tilde{a}_{11}} \right)} \quad \dots (4.6)$$

$$M_{21} = a_{21} / \tilde{a}_{11}, \quad M_{31} = a_{31} / \tilde{a}_{11} \quad \text{とおくと上式は}$$

$$\begin{aligned}a_{33}'' &= \left(a_{31} - M_{31} \cdot a_{13} \right) - \frac{\left(a_{32} - M_{31} \cdot a_{12} \right) \left(a_{23} - M_{21} \cdot a_{13} \right)}{\left(a_{22} - M_{21} \cdot a_{12} \right)} \\ &= a_{31} - M_{31} \cdot a_{13} - \left\{ \frac{a_{32} a_{23} - M_{31} \cdot a_{12} a_{23} - M_{21} a_{32} \cdot a_{13} + M_{31} \cdot a_{13} \cdot a_{22}}{a_{22} - M_{21} \cdot a_{12}} \right\} + M_{31} \cdot a_{13} \quad \dots (4.7)\end{aligned}$$

$|M_{21}|, |M_{31}|$ は他の $|a_{ij}|$ に比べて優れて大きい。

このことは $|\tilde{a}_{ij}|$ が異常に大きくなつた要因である。

また(4.7)式では、 M_{21}, M_{31} 以外の項は \tilde{a}_{11} の誤差を含んでいいから、精度がちがつていい。

さて \tilde{a}_{33}'' の結果に影響する項を見ると、絶対値の一番大きな $M_{31} \cdot a_{13}$ の項が、+,-されてcancelされて、数値は

大きく桁落ちして、あとに a_{31} と { } の項が残される。

次に { } の項を省く。

分子、分母、分子と M_{21}, M_{31} が掛っている項が絶対値が優越する。

このことは分子、分子に同じ $1/\tilde{a}_{ii}$ が乗せられた項が優越していることになる。誤差を含んでいても、分子、分子全体にそれぞれ乗じられていれば、精度には影響しないから \tilde{a}_{ii} を乗じてみると { } 内は

$$\frac{a_{32} \cdot a_{23} \cdot \tilde{a}_{ii} - a_{31} \cdot a_{12} \cdot a_{23} - a_{21} \cdot a_{32} \cdot a_{13} + a_{31} \cdot a_{13} \cdot a_{22}}{a_{22} \cdot \tilde{a}_{ii} - a_{21} \cdot a_{12}} \quad \dots (4.8)$$

となる。 \tilde{a}_{ii} を含む項は他の項に比し絶対値が小さいため、その誤差の影響は下の桁に移動し、精度は M_{ii} よりよくなっている。

結局 \tilde{a}_{33} は、 \tilde{a}_{ii} 以外の a_{ij} の情報を途中の計算で落さないで最後まで運んで来れば、精度のよい解が得られる。

この例の要点をまとめる

- (1) $1/\tilde{a}_{ii}$ の誤差項によつて a_{ij} の情報を運ばれていく。
- (2) さうに高精度計算によつて a_{ij} の情報を忠実に保持して来たので、最後に誤差を多く含む項の桁落ちによつて、 a_{ij} の正しい情報を生きてきて精度のよい結果が得られた。
- (3) 分母、分子が同じ誤差の数値で約されて精度が恢復した。

た。

「コメント」

この例は連立一次方程式の掃き出し法の過程で起る Pivot の桁落ちである。この例で示したように、Pivot が桁落ちしたとき、その段と次の段階の消去計算を倍精度計算にすれば 3 段階目の要素は、またもとの要素と同じ程度の精度となる。このことから倍桁計算をしてみけば、Pivot の桁落ちによる精度落ちは防げる。今入力 Data が 4 行のとき 8 行計算 すでに倍桁計算になつていい。それを 16 行計算としても無意味である。ところが最近の計算機は 10 進で 7 行前後しか精度がない。これでは 4 行精度の入力 Data に対しては、倍桁に一寸足りないため、桁削減による誤差も求めなくなつてしまふ。倍桁計算で求める問題はよく出て来るが、桁数が奇数 2, 3 行多くなつたら大部動率が高くなつてはなるいかで思う。

[例 9] $x \sim C$ のとき

$$\frac{(x-c)(x-a)}{(x-c)(x-b)} = \frac{x^2 - (a+c)x + ac}{x^2 - (b+c)x + bc}$$

左辺で計算すれば、桁落ち項が分母、分子で約されて精度落ちしない。右辺で計算するときは分母、分子の多項式計算で

桁あらして精度が悪くなり、結果は分子、分子の相対誤差の和の相対誤差となる。

[数値例] 10]

$$f(x) \equiv \frac{(x-1.0000)(x-2222.2)}{(x-1.0000)(x-3333.3)} = \frac{x^2 - 2223.2x + 2222.2}{x^2 - 3334.3x + 3333.3}$$

$x = 1.0011$ のとき

$$\text{真値} \equiv \frac{(0.0011)(-2221.1989)}{(0.0011)(-3332.2989)} = 0.6665665256\cdots$$

… (2)

1°) (1) 式の左辺で計算 (浮動5桁計算)

$$\frac{(0.0011)(-2221.2)}{(0.0011)(-3332.3)} = \frac{-2.4433}{-3.6655} = 0.66657$$

2°) (1) 式の左辺で計算で, $\widetilde{x-c} = 0.001153$ と丸めの誤差を入ったとき (浮動5桁計算)

$$\frac{(0.001153)(-2221.2)}{(0.001153)(-3332.3)} = \frac{-2.5610}{-3.8421} = 0.66656$$

となり $\widetilde{x-c}$ の誤差は打ち消され, 1°) と同程度の精度となる。

3°) $x = 1.0011$ で(1)式の右辺の計算

a) 浮動5桁計算

$$\frac{1.0022 - 2225.7 + 2222.2}{1.0022 - 3338.0 + 3333.3} = \frac{-2.5}{-3.7} = 0.67568$$

b) 浮動8桁計算

$$\frac{1.0022012 - 2225.6455 + 2222.2}{1.0022012 - 3337.9677 + 3333.3} = \frac{-2.4433}{-3.6655} = 0.66656663$$

c) 浮動10桁計算

$$\begin{array}{r} 1.0022 \ 0121 - 2225.64552 + 2222.2 \\ \hline 1.0022 \ 0121 - 3337.96773 + 3333.3 \end{array} = \begin{array}{r} -2.443319 \\ \hline -3.665529 \end{array}$$

$$= 0.6665665\underline{4}47$$

(注) このときは、分母、分子の各項は変換誤差がなければ正しい値が出ていいから、(第2項+第3項)+第1項の順に計算すれば分子、分母は正しい値が得られる。

a), b), c)とも3桁の桁落ちのため、計算結果の最後の桁から3桁目で丸められるため下から4桁が誤差となる。

(4.7)

おりに、單に計算例を示すに止まつたが、これ等以外のパターンも含めて、丸め誤差の伝播の解析が理論的にはされどとが望ましく、理論的態型をつくって行きたいと思う。