数理解析研究所講究録
第153巻 1972年 129-151



# 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が目で追って計算していたときは，桁落ちして精度が無くなっ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って，その大部分の計算が処理されている．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も悪いので，Print outは最小限に止めなければならず，数値のほとんどが電算機のブラック・ボックス中で移動変化して行く．そのため，すべての事態にそなえて対処出来るような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して，極く少数が case by case として扱われているだけで一環したものがない。一方丸め誤差は，その計算法に従って忠実に規則的に拡大，縮少，消滅して行くもので，丸め誤差を最小限にもっ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，いくつかの例によってわかりました。そこでその規則を見い出すためには，先ず丸め誤差の挙動と，その計算式の関係をパタンに分類してみたらと思い，計算の基本にもどって考えて見ました。こゝ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. 桁落ち　　2. 情報落ち　　3. 丸め誤差

§3. 精度落ちする計算パタンとその計算例

1. 精度落ちした数値での乗除算

2. 桁落ち　　3. 情報落ち

§4. 精度恢復の計算パタンとその計算例

1. 情報落ち

2. 分母，分子の誤差項の約分

§0. はじめに　数値，誤差の定義をしておく．

真値：$a$，　誤差：$\Delta a$，　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．

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

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．

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

1.3. べき乗数の相対誤差は，底の相対誤差のべき倍となる．

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

§2. 演算における精度

2.1. 桁落ちと精度

$$a \pm b = c \qquad (|a| \fallingdotseq |b|)$$

において左辺が2数の絶対値の差となるようなとき，$c$ の位いは $a$，$b$ より下る．これを桁落ちという．

$$1.234\underline{535} - 1.23678\underline{5} = 0.00225\underline{0}$$

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

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

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

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

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

$$a \pm b = c \qquad (|a| >> |b| \text{ 又は } |a| << |b|)$$

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ことがある。

(2) 情報落ちしては困るとき，高精度計算で救われることがある。 (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来る。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っている。しかし 1つ1つの数値の誤差だけを追っていたのでは，目的の解の誤差は思い出せない。計算過程での誤差の変動を，その数値との相対関係を加味して追跡して行かなければならない。

## §3. 精度落ちする計算パタンとその計算例

### 3.1. 精度落ちした数での乗除算

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

$\hat{e}$の相対誤差が$\hat{a}$, $\hat{b}$の相対誤差に比し大きいときは, $\hat{c}$, $\hat{d}$の相対誤差は$\hat{e}$の相対誤差と同じになり精度が悪くなる. しかし, 桁落ちしないときは$\hat{a}$, $\hat{b}$の情報は失われない.

### 3.2. 桁落ちの起るパタン

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

この場合は, 分子の有理化によって, 桁落ちによる精度落ちは防げる.

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

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

この計算式では複号の加減算で, 一方が桁落ちを起す.

(解法II)

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

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

(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]　　(単精度浮動10進約7桁計算)

二根を $x_1=\sqrt{3}$, $x_2=\sqrt{2}\times10^{-4}$ とし, この近似値を

X1= 0.1732050807568877D 01, X2= 0.1414213562373095D-03

でまえ倍桁計算による係数より有効7桁とって係数とし,

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

により (解法I) の解 X1= 0.1732050E 01, X2= 0.1414418E-03 *)

(解法II) の解 X1= 0.1732050E 01, X2=C/A/X1= 0.1414213E-03

*) $x_2=\frac{-b}{2a}-\sqrt{D}=0.866096-0.8659546=0.0001414\times\times\times$

となって, 減算で3桁けたおちしたので丸め誤差が下から3桁上って来た.

[数値例2]　　(単精度浮動計算)

二根を $x_1=\sqrt{3}$, $x_2=\sqrt{2}\times10^{-8}$ とし, この近似値を

X1= 0.1732050807568877D 01, X2= 0.1414213562373095D-07

でまえ倍桁計算による係数より有効7桁とって係数とし,

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

により (解法I) の解 X1= 0.1732049E 01, X2= 0.5960464E-07

(解法II) の解 X1= 0.1732049E 01, X2=C/A/X1= 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法で反復計算をすると, 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, まだまだ改善の余地のあることを暗示している.

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

$\tilde{y}=\sqrt{1.732050^2-3.000000}=\sqrt{0.000003}=0.00\underline{1732050}$

(浮動7桁計算)

(i) $|a|\fallingdotseq|b|\fallingdotseq n\times10^d$ $(1\leqq n<10)$ のとき $a^2-b^2$ の計算で $m$ 桁桁おちしたとき $\sqrt{a^2-b^2}$ の誤差は約 $10^{d-l+m/2}$ となる。($l$ は計算桁数)

(ii) $a$, $b$ が正しい値のときは，桁お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 $\sqrt{3}\pm\sqrt{2}\times10^{-6}\,i$ とし，この近似値を

0.1732050807568877D 01 ± 0.1414213562373095D-05 I とし

(a) 倍桁計算による係数より有効7桁とって係数とし

A= 0.1000000E 01, B= -0.3464101E 01, C= 0.3000000E 01

により単精度計算の解 0.1732050E 01 ± 0.1953125E-02 I

倍精度計算の解 0.1732050418853760D 01 ± 0.1160408770803387D-02 I

(b) 倍桁計算による係数 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 ± 0.1414197919868275D-05 I

i) 係数が単精度のときは，虚数部は前記の $\tilde{y}$ に示すように根号内の最後の桁まで桁おちして完全に誤差となる．このとき倍精度計算しても精度は上らず，$10^{-6}$ の誤差が周平で，$10^{-3}$ 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桁桁おちし，計算桁数17桁程度なので，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

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

$|a|>|b|$ 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って真値が出なくなる。

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

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）

$$5.000000^3 + 0.03000000^3 = 125.0000 + 0.0000270000$$
$$= 125.0000$$

［例 2］

行列の固有値解法の Jacobi 法において，回転角を θ としたとき，cosθ の計算は次式でなされる。

$$\cos\theta = \sqrt{\frac{1}{2}\left\{1 + \frac{|a_{pp}-a_{qq}|/2}{\sqrt{\{(a_{pp}-a_{qq})/2\}^2 + a_{pq}^2}}\right\}} \qquad \cdots\cdots (1)$$

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

$$\max\{|a_{pp}|, |a_{qq}|\} > |a_{pq}| \times 10^{L/2} \qquad (L\text{は計算桁数})$$

となると，$\sqrt{\{(a_{pp}-a_{qq})/2\}^2 + a_{pq}^2} = |a_{pp}-a_{qq}|/2$ となって (1) 式の $\cos\theta = 1$ となり，$\theta = 0$ となって回転は止ってしまう。すなわちこのとき $a_{pq}$ の情報は完全に落されている。

## §4. 精度恢復の計算パタンとその計算例

### 4.1. 情報落ちによる精度恢復

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

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。このとき

> 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，$\hat{y}$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ることになる。ゆえに $\hat{a} = a + \Delta a$, $\hat{b} = b + \Delta b$ として考えればよい。

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

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

[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(解法エ)で倍桁計算する．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　$x_1=\sqrt{3}$ ,　$x_2=\sqrt{2}\times10^{-4}$　の近似値

X1= 0.1732050807568877D 01,　X2= 0.1414213562373095D-03

より係数 A= 0.1000000E 01,　B= -0.1732192E 01,　C= 0.2449488E-03

の解 X1= 0.1732050618230121D 01,　X2= 0.1414212596248659D-03

二根　$x_1=\sqrt{3}$ ,　$x_2=\sqrt{2}\times10^{-8}$　の近似値

X1= 0.1732050807568877D 01,　X2= 0.1414213562373095D-07

より係数 A= 0.1000000E 01,　B= -0.1732050E 01,　C= 0.2449490E-07

の解 X1= 0.1732049927874460D 01,　X2= 0.1414214141626236D-07

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

[例3]　$a_{11}$が他の$a_{ij}$に比し絶対値が小さいとき

真値 $\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = \begin{vmatrix} 0.0010 & 2.2222 \\ 3.3333 & 4.4444 \end{vmatrix}$

$= 0.0044\,444 - 7.4072\,5926$

$= -7.4028\,1486$

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

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

$= -7.4026$

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 $x_1$ が求まったとき

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

の係数 $b_i$ を求めるとき

$$\begin{cases} b_n = 0 & \text{とおき} \\ b_i = a_{i+1} + b_{i+1} \cdot x_1 & (i = n-1, n-2, \cdots, 0) \end{cases}$$

により $b_i$ 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なって $b_i$ の精度は悪くならない。

[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 $\tilde{x}_1 = 0.1001$ により $f_1(x)$ を求める。

(浮動5桁計算)

| | | | | | |
|---|---|---|---|---|---|
| $a_i$ | 1.0 | −103.10 | 212.30 | −20.2 | |
| | | 0.1001 | −10.310 | 20.219 | |
| $b_i$ | 1.0 | −103.00 | 201.99 | 0.019 | $= f(\tilde{x}_1)$ |

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

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

「コメント」

絶対値の小さい根から求め，この[例4]の解法によっ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7桁計算で $x_3 = 101.0001$，$x_2 = 1.999534$，$x_3 = 0.1004567$ となる。

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

の絶対値最大根 $x_n$ が求まったとき

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

の係数を求めるとき

$$b_0 = a_0 \quad \text{とおき}$$

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

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

[数値例7]

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

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

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ときの割算で，絶対値最小の根により次数を下げるときには，高次の項より割って行く従来の方法によらなければならない．

$f(x)$ を $x$ の2次式で割っ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次式で割って行くのに対し，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っている．

[例6] 代数方程式の近接根の誤差のcansel

近接根をもつ代数方程式のNewton法において，その解法手つづきにおいて，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いが，その誤差は相手の近接根の誤差と相殺して，近接根以外の根の精度には影響しなくなる。

[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 | $10^{-7}$ 103.0 | 1.000 0398 | $10^{-7}$ 398.0 |
| 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，それぞれ

$\sum_{i=1}^{5} \varepsilon_i = -2.0 \times 10^{-7}$，$\sum_{i=1}^{5} \varepsilon_i^* = -3.0 \times 10^{-7}$ となっていて，ともに計算桁の最後の桁に止っている。すなわちほとんどゼロになっているといえる。

（解析）

$$f(x) = x^5 + a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0$$

$$f_1(x) = f(x) \,/\, (x - x_1) = x^4 + b_3x^3 + b_2x^2 + b_1x + b_0$$

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行く．すなわち

$b_4 = 1$ として

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

一方 $a_i$ は根と係数の関係より

$$\begin{cases} a_4 = -x_1 - (x_2 + x_3 + x_4 + x_5) \\ a_3 = x_1(x_2+x_3+x_4+x_5) + (x_2x_3 + x_2x_4 + x_2x_5 + x_3x_4 + x_3x_5 + x_4x_5) \\ a_2 = -x_1(x_2x_3 + x_2x_4 + x_2x_5 + x_3x_4 + x_3x_5 + x_4x_5) \\ \qquad\quad - (x_2x_3x_4 + x_2x_3x_5 + x_2x_4x_5 + x_3x_4x_5) \\ a_1 = x_1(x_2x_3x_4 + x_2x_3x_5 + x_2x_4x_5 + x_3x_4x_5) + x_2x_3x_4x_5 \\ a_0 = -x_1x_2x_3x_4x_5 \end{cases}$$

となり ＿＿ の部分が $b_i$ である．

今 $\widetilde{x_1} = x_1 + \varepsilon$ のとき $f_1(x)$ の係数を $\widetilde{b_i}$ とすると

$$\widetilde{b_3} = -\{(x_2 - \varepsilon) + x_3 + x_4 + x_5\}$$

$$\widetilde{b_2} = \{(x_2-\varepsilon)(x_3+x_4+x_5) + x_3x_4 + x_3x_5 + x_4x_5\} + \varepsilon(x_1 - x_2) + \varepsilon^2$$

$$\widetilde{b_1} \fallingdotseq -\{(x_2-\varepsilon)(x_3x_4 + x_3x_5 + x_4x_5) + x_3x_4x_5\} - \varepsilon(x_1-x_2)(x_3+x_4+x_5) - \varepsilon^2(x_3+x_4+x_5)$$

$$\widetilde{b_0} \fallingdotseq \{(x_2-\varepsilon)x_3x_4x_5\} + \{\varepsilon(x_1-x_2) + \varepsilon^2\}(x_3x_4 + x_3x_5 + x_4x_5)$$

となり ～～ の部分が計算打ち切られて，$f_1(x)$ はほぼ，$(x_2-\varepsilon)$，$x_3$，$x_4$，$x_5$ を根とする方程式となっている．

4.2. 分母，分子の誤差項の約分による精度恢復

$$\frac{\tilde{e}\cdot\tilde{a}}{\tilde{e}\cdot\tilde{b}} \fallingdotseq \frac{\tilde{a}}{\tilde{b}} \qquad \cdots\cdots (4.2)$$

[数値例8]

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

各項にそれぞれ誤差を入れて計算すると

$$\text{近似値} = \frac{1.2530\times 4.9875}{1.2530\times 9.9754} = \frac{6.2493}{12.499}$$

$$= 0.49998 \qquad (\text{浮動5桁計算})$$

$$= 0.4999799507 \qquad (\text{浮動10桁計算})$$

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もa, bの誤差により5桁計算と同程度の精度となる．

(証明)

$$\frac{(e+\Delta e)(a+\Delta a)}{(e+\Delta e)(b+\Delta b)} \fallingdotseq \frac{ea\left(1+\frac{\Delta a}{a}+\frac{\Delta e}{e}\right)}{eb\left(1+\frac{\Delta b}{b}+\frac{\Delta e}{e}\right)}$$

$$= \frac{a}{b}\left(1+\frac{\Delta a}{a}+\frac{\Delta e}{e}\right)\left\{1-\left(\frac{\Delta b}{b}+\frac{\Delta e}{e}\right)+\left(\frac{\Delta b}{b}+\frac{\Delta e}{e}\right)^2-\cdots\right\}$$

$$\fallingdotseq \frac{a}{b}\left(1+\frac{\Delta a}{a}-\frac{\Delta b}{b}\right)$$

となりe, Δeは完全に消滅して，結果には影響せずΔa, Δbが相対誤差で関与して来る．よってeはどのような数値が来てもよいことになる．

[例7]　$a_{11}$が他の$a_{ij}$に比し絶対値が小さいとき

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

による計算を考える.

P10／[例3]の数値で，浮動5桁計算で計算してみる.

$$\begin{vmatrix} \widetilde{a_{11}} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = 0.001053 \times \left(4.4444 - \frac{2.2222\times 3.3333}{0.001053}\right)$$

$$= 0.001053 \times (4.4444 - 7034.4)$$

$$= 0.001053 \times 7030.0 = 7.4026 \qquad \cdots\cdots (4.4)$$

§1. の誤差評価に従うと上記の～～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っている．このことは(4.3)式の右辺で$|a_{11}\cdot a_{22}| \ll \left|a_{11}\cdot\frac{a_{12}\cdot a_{21}}{a_{11}}\right|$であるため計算結果に影響するのは$a_{11}\cdot\frac{a_{12}\cdot a_{21}}{a_{11}}$である．$\widetilde{a_{11}}$に多くの誤差が入っていても，P15／(4.2)式によって，その誤差はCanselされて高々計算桁最後の丸め誤差に止まっている.

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る.

更に浮動10桁計算をこころみてみる

$$\begin{vmatrix} \widetilde{a_{11}} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = 0.001053 \times (4.4444 - 7034.434245)$$

$$= 0.001053 \times 7029.989845$$

$$= 7.402579307 \qquad \cdots\cdots (4.5)$$

§1.の誤差評価に従えば相対誤差は2桁しかないが実際は4桁出ている．この場合$|\widetilde{a_{11}}\cdot a_{22}|$が$\left|a_{11}\cdot\frac{a_{21}\cdot a_{12}}{a_{11}}\right|$の絶対誤差より優越しているため10桁計算しても5桁程度の精度に止っている.

［例 8］ 情報落ちと高精度計算

3次の行列式において，$\hat{a}_{11}$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，単精度の解が得られる．

［数値例 9］

$$|A| = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = \begin{vmatrix} 0.0011 & 2.2222 & 4.4444 \\ 3.3333 & 5.5555 & 6.6666 \\ 7.7777 & 8.8888 & 9.9999 \end{vmatrix} \fallingdotseq -19.20788\ 74414$$

$\hat{a}_{11} = 0.001153$ のとき，浮動5桁計算による計算過程

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

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

ただし $\tilde{a}'_{ij} = a_{ij} - \dfrac{a_{i1}\cdot a_{1j}}{\tilde{a}_{11}}$

$\tilde{a}''_{33} = \tilde{a}'_{33} - \dfrac{\tilde{a}'_{32}\cdot\tilde{a}'_{23}}{\tilde{a}'_{22}}$

$\hat{a}''_{33}$が完全に誤差になっているため$|A|$の値はデタラメとなる．

$\widehat{a_{11}} = 0.001153$ のとき，浮動10桁（倍桁）計算過程

| | | | 要素名 | | |
|---|---|---|---|---|---|
| 0.001153 | 2.2222 | 4.4444 | $\widehat{a_{11}}$ | $a_{12}$ | $a_{13}$ |
| 3.3333 | 5.5555<br>6424.33587 3 | 6.6666<br>12848.67174 | $a_{21}$ | $a_{22}$ | $a_{23}$ |
| 7.7777 | 8.8888<br>14990.11704 | 9.9999<br>29980.23406 | $a_{31}$ | $a_{32}$ | $a_{33}$ |
| 0 | −6418.78037 3 | −12842.00514 | | $\widehat{a'_{22}}$ | $\widetilde{a'_{23}}$ |
| 0 | −14981.22824 | −29970.23416<br>29972.82955 | | $\widehat{a'_{32}}$ | $\widehat{a'_{33}}$ |
| 0 | 0 | 2.59539 | | | $\widetilde{a''_{33}}$ |

$$|A| \fallingdotseq \widehat{a_{11}} \cdot \widetilde{a'_{22}} \cdot \widetilde{a''_{33}} = 19.20810\ 186$$

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

（解析）

1°） 2次の行列式の精度

$$\begin{vmatrix} \widehat{a_{11}} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = \widehat{a_{11}} \times \widetilde{a'_{22}}$$ の値はp14/(4.4)式で示したように，$\widetilde{a'_{22}}$ には $\widehat{a_{11}}$ の誤差が含まれ精度は悪くなつているが，$\widehat{a_{11}}$ を乗ずることによつて誤差の cansel が起り，計算結果は計算桁最後の桁に丸め誤差が含まれる程度となる．数値にて示
~~~~

すと

真値 ： $a_{11} \times a_{22}' = -7.40114\ 821$

近似値： $\tilde{a}_{11} \times \tilde{a}_{22}' = -7.4008$ （浮動5桁計算）

$= -7.40085\ 3770$ （浮動10桁計算）

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 $\tilde{a}_{11}, a_{22}$ 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る．

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

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

$$a_{33}'' = (a_{31} - M_{31} \cdot a_{13}) - \frac{(a_{32} - M_{31} \cdot a_{12})(a_{23} - M_{21} \cdot a_{13})}{(a_{22} - M_{21} \cdot a_{12})}$$

$$= a_{31} - M_{31} \cdot a_{13} - \left\{\frac{a_{32}a_{23} - M_{31} \cdot a_{12}a_{23} - M_{21}a_{32} \cdot a_{13} + M_{31} \cdot a_{13} \cdot a_{22}}{a_{22} - M_{21} \cdot a_{12}}\right\} + M_{31} \cdot a_{13} \quad \cdots (4.7)$$

$|M_{21}|$, $|M_{31}|$ は他の $|a_{ij}|$ に比し，優越して大きい．

このことは $|\tilde{a}_{ij}|$ が異状に大きくなった要因である．

また（4.7）式では，$M_{21}$, $M_{31}$ 以外の項は $\tilde{a}_{11}$ の誤差を含んでいないから，精度おちはしていない．

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

大きく桁おちして，あとに $a_{31}$ と $\{\cdots\}$ の項が残される．

次に $\{\cdots\}$ の項を考える．

分母，分子とも $M_{21}$, $M_{31}$ が掛っている項が絶対値が優越する．このことは分母，分子に同じ $1/\tilde{a}_{11}$ が乗ぜ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，分母，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 $\tilde{a}_{11}$ を乗じてみると $\{\cdots\cdots\}$ 内は

$$\frac{a_{32}\cdot a_{23}\cdot \tilde{a}_{11} - a_{31}\cdot a_{22}\cdot a_{23} - a_{21}\cdot a_{32}\cdot a_{13} + a_{31}\cdot a_{13}\cdot a_{22}}{a_{22}\cdot \tilde{a}_{11} - a_{21}\cdot a_{12}} \quad \cdots (4.8)$$

となる．$\tilde{a}_{11}$ 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 $M_{21}$ よりよくなっている．

結局 $\tilde{a}_{33}$ は，$\tilde{a}_{11}$ 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，精度のある解が得られる．

<u>この例の要点をまとめると</u>

(1) $1/\tilde{a}_{11}$ の誤差項によって $a_{ij}$ の情報が運ばれている．

(2) さらに高精度計算によって $a_{ij}$ 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おちによって，$a_{ij}$ の正しい情報が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。

「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の桁おちである。この例で示したように，Pivotが桁おちしたとき，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おけば，Pivotの桁おちによる精度おちは防げる。今入力Dataが4桁のとき8桁計算ですでに倍桁計算になっている。それを16桁計算をしても無意味である。ところが最近の計算機は10進で7桁前後しか精度がない。これでは4桁精度の入力Dataに対しては，倍桁に一寸と足りなくなるため，桁角数える誤差も救えなくなってしまう。倍桁計算で救える問題はよく出て来るので，桁数があと2，3桁多くなったら大部効率が良くなるのではないかと思う。

[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，桁おち項が分母，分子で約されて精度おちしない。右辺で計算するときは分母，分子の多項式計算で

桁みちして精度が悪くなり，結果は分母，分子の相対誤差の和の相対誤差となる．

［数値例10］

$$f(x) \equiv \frac{(x-1.0000)(x-2222.2)}{(x-1.0000)(x-3333.3)} = \frac{x^2-2223.2x+2222.2}{x^2-3334.3x+3333.3} \quad \cdots\cdots (1)$$

$x = 1.0011$ のとき

$$\text{真値} \equiv \frac{(0.0011)(-2221.1989)}{(0.0011)(-3332.2989)} = 0.66656\ 65256\cdots\cdots \quad \cdots\cdots (2)$$

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

$$\frac{1.0012-2225.7+2222.2}{1.0012-3338.0+3333.3} = \frac{-2.5}{-3.7} = 0.67568$$

b) 浮動8桁計算

$$\frac{1.0022012-2225.6455+2222.2}{1.0022012-3337.9677+3333.3} = \frac{-2.4433}{-3.6655} = 0.6665\ 6663$$

c) 浮動10桁計算

$$\frac{1.0022\ 0121 - 2225.64552 + 2222.2}{1.0022\ 0121 - 3337.96773 + 3333.3} = \frac{-2.443319}{-3.665529}$$

$$= 0.6665665447$$

(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るから，(第2項＋第3項)＋第1項の順に計算すれば分る，分母は正しい値が得られる．

a), b), c) とも3桁の桁あちのため，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桁が誤差となっている．

(4.7)

おわりに，単に計算例を示すに止まつたが，これ等以外のパタンも含めて，丸め誤差の伝播の解析が理論的になされることが望ましく，理論的態型をつくつて行きたいと思う．