数 値 解 析 31－2
（1989．12．8）

計算桁数と精度の関係について

山下真一郎
富士通株式会社

連続量上の数学的問題を浮動小数点演算装置の計算機で数値的に解く場合、有限桁計算のために厳密解は一般に得られない．このことは数値解の精度が計算の有効桁数に依存しているとして、経験的にもよく知られている．しかし、その依存関係は必ずしも明らかではない．

本資料では、代数方程式、線形方程式、固有値問題および微分の近似などの基本的な数学的問題に対して、数値解の精度と計算桁数との依存関係を統一的に論ずる．特に、計算の有効桁数をp、数値解の精度の有効桁数をdとするとき、これらの両者の間に次のような線形関係が成り立つことを示す:

$$d=(p-\alpha)/m$$

ここで、αは損失桁数、mは問題に依存する縮率である．

On Relations between Numerical Accuracy and Mantissa Length in Computation with Floating-point Arithmetic

Shin-ichiro Yamashita
Software Div., Fujitsu Limited
140 Miyamoto, Numazu-shi Shizuoka, 410-03 JAPAN

When a mathematical problem defined on a continuous quantity is solved by a computer with floationg-point arithmetic operations, an exact solution is not obtained in general because the computation is done by a finite number of digits. In fact, it is experimentally well-known that accuracy of a numerical solution depends on a number of significant digits for computation. But the dependence relation is not necessary clear.

In this paper, a unified discussion of the relation is given for some mathematical problems, i.e., algebraic equations, linear equations, eigen-value problems and an approximation of derivatives. In particular, we show the following linear relation between the number d of significant digits of a numerical solution and the number p of significant digits for computation with floationg-point arithmetic operations;

$$d=(p-\alpha)/m,$$

where α is a number of canceling digits and m is a rate of reduction depended on the problem.

§１　まえがき

計算桁数を少しづつ増やして行と，計算結果の精度はどうなるか？　著者はこの事に大変興味を持つて，これまでに，代数方程式，数値微分，連立一次方程式，固有値の結果を知った。この度固有ベクトルの事が分かった。従来判明している事および固有ベクトルの事を報告する。

§２　相対零／ベクトル・行列の相対零

従来の零を絶対零といい，浮動小数点数の零を相対零という。相対零を明確にする為に，浮動小数点数を二つに分ける。いま，任意の零でない浮動小数点数Ａを固定して考える。第２の数値をＢとする。このとき

１つは，$A \pm B \neq A$　となる数値Ｂ

２つは，$A \pm B = A$　となる数値Ｂ

のように，浮動小数点数を二つに分ける。このような２つの種類の数値は，それぞれ最大最小を有する。任意の零でない浮動小数点数Ａの相対零は，１の数値を考え，正の数値の時には，Ｂの最小の数値をＡの相対零という。負の数値の時には，Ｂの最大の数値をＡの相対零という。

この相対零を定量的に表せば，Ａの相対零は，$A\varepsilon$である。ここに，$|\varepsilon| \leqq u$　である。ただし，ｕは【丸め誤差単位】である。　同様に考えれば，ベクトルの相対零は，個々の要素の相対零を要素とするベクトルである。従って，ベクトルＸを

$$X = (X_1, X_2, X_3, \cdots, X_n)^T$$

のように表すと，Ｘの相対零は，

$$(X_1\varepsilon_1, X_2\varepsilon_2, X_3\varepsilon_3, \cdots, X_n\varepsilon_n)^T$$

ただし，$|\varepsilon_k| \leqq u$　, $k = 1, 2, 3, \cdots, n$;

である。簡略算として，ノルムを考える。簡略算のＸの相対零は，

$\|X\|\varepsilon$　;　ただし，$\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3, \cdots, \varepsilon_n)^T$

$|\varepsilon_k| \leqq u$　, $k = 1, 2, 3, \cdots, n$;

である。このεは，【ベクトルの丸め誤差単位】である。

同様に，行列の相対零は，行列の要素のε倍を要素とする行列で定義できる。

§３　精度公式

一般的に相対誤差を評価すれば，精度（の桁数）ｄは，

$$d = -\log_M |\text{相対誤差}|$$

と成る。そして，『計算桁数』と『解精度桁数』との関係は，多くの場合，次の式が成立つ。

$$d = (P - \alpha) / m \quad \cdots\cdots\cdots\cdots (1)$$

ここに，

ｄ：解の精度（桁）　　　　ｐ：演算桁数（桁）

α：問題と解法に固有な損失桁数　　ｍ：問題と解法に固有な縮率

この式は可変長の演算桁数のシステムで，いろいろ経験した経験則である。そして，この式を『精度公式』と呼ぶ。この式は，何時でも成立する訳ではない。この精度公式は

『多くの場合，この式が成立する』

と言うべきである。

さて，演算桁数と解精度の関係は，多くの場合，（１）式の様に一次式である。αは，損失桁で，問題と解法に依存した値である。mは，問題と解法に依存した値で，縮率と言い，通常は1である。一般的に

〔演算の精度の高い方の結果の精度〕 ＞ 〔演算の精度の低い方の結果の精度〕

が成り立つのが普通である。しかし，成立しない場合もある。普通は（１）式が成立して，演算桁数を漸増すれば解精度（桁）も漸増する。

§４　精度公式の成立しない例

ここで（２・１）が成立しない例を示す。

（ａ）データや定数に誤差がある場合。

（ｂ）計算式が近似式である場合。

（ｃ）反復計算等を適当に打切る場合。

（ｄ）その他。

等である。ａ～ｃは，具体的には示さない。ここには，その他の場合を示めそう。精度公式が素直に成立しないのは，計算方法として問題がある，と見るべきである。そのような計算は，信頼性の観点から再検討してみる価値がある。

次の例は，演算桁数が少ない方が精度が良く成る。いま，

$$A = 2.3456110000000234567$$
$$B = 1.1111109999999111111$$
$$A-B = 1.2345000000001123456$$

と置いて，１０進法p桁で，演算桁数を種々変えながら，ｙ＝（Ａ－Ｂ）を計算する。その時の演算桁数と解精度の関係は，図1の様に成る。

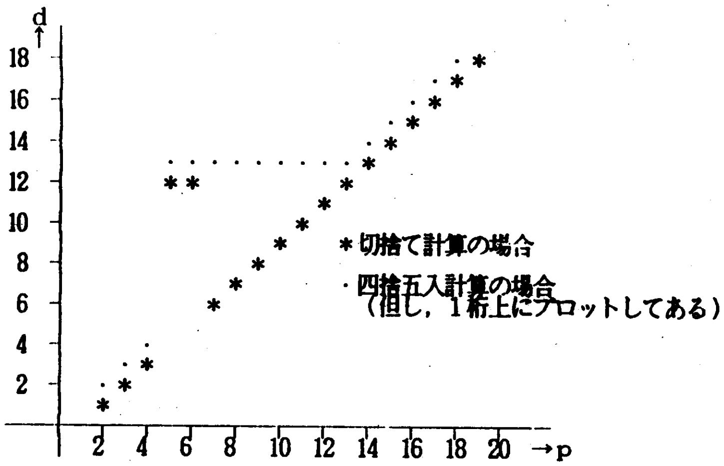

図1　切り捨て算の場合の例

これは簡単な例であるが，実際的な計算でも起きていると思うべきである。

この他の例と言うのは，例えば，

$$y = x + (Ax - [Ax]) / B, \quad \cdots\cdots\cdots\cdots\cdots\cdots (1)$$

ここに，$A=10^m$，　$B=10^n$，

$x=1/3$，$1 \leq n < m$　の整数

〔 〕はガウス記号

（yの真値）＝1／3＋1／（3B）

は，M進法p桁で計算すれば，p≦m のとき，カッコ内が整数以上になり，結局0になるので，精度は，図2のようになる。

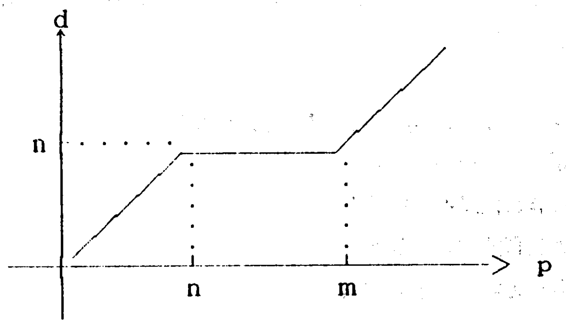

図2 y＝x＋（Ax－［Ax］）／B の桁数（p）と精度（d）の関係

これなどは，大変おもしろいが，如何にも作った感じがする。次の例は，良い例である。

$y = x + \sin(Ax)/B$, ・・・・・・・・・・・・・・・・・・・・・・（2）

x，A，Bは（1）式と同じとする。

結果のグラフも図2と大体同じである。

§5 代数方程式

高次代数方程式の0点を求める場合を考える。

$$f(X) = \sum_{k=0}^{n} A_k X^k \quad \cdots\cdots\cdots\cdots\cdots \quad (1)$$

の根をRとし，これがm（m≧1）重根であるとする。そうすれば，

$$f(X) = (X-R)^m g(X) \quad , \quad g(X) \neq 0 \quad \cdots \quad (2)$$

と考える事が出来る。X＝Rの近傍で，（1）からf（X）の0点を求めると言う事は，摂動項を考えに入れて，

$$f(X) - K(X)\varepsilon = 0 \quad , \quad |\varepsilon| \leq u$$

uは丸め単位，Computer εとも言う。

を正確に解く事を意味する。K（X）は，少し大き目であるが，

$$K(X) = \sum_{k=0}^{n} \mathrm{abs}(A_k X^k) \quad \cdots\cdots\cdots\cdots\cdots \quad (3)$$

と置く。（3）の意味は，関数f（X）をこれ以上小さくして，0点を確定出来ない事を意味する。（2），（3）から，

$$\mathrm{abs}(X-R) = (\ \mathrm{abs}(\ K(X)\varepsilon / g(X)\ )\ )^{1/m} \quad \cdots\cdots \quad (4)$$

と置けば，これは根Rからのズレを表わす。M進法p桁で計算し，丸めの単位を$u = M^{1-p}$とすれば，精度（桁）dは，

$$
\begin{aligned}
d &= -\log \mathrm{abs}(\text{相対誤差}) \quad \cdots\cdots\cdots\cdots \quad (5)\\
&= -\log \mathrm{abs}((X-R)/R)\\
&\leqq -\log \mathrm{abs}(K(X)u/(R^m g(X)))/m\\
&\sim -\log \mathrm{abs}(K(X)u/(X^m g(X)))/m\\
&= (p-\alpha)/m
\end{aligned}
$$

ただし，$\alpha = 1 + \log \mathrm{abs}(K(X)/(X^m g(X)))$

log の底はM

と成る。根の近傍で，この式が大きく変わるとすれば，g（X）であるが，これも落ち着く程度に根の近傍だとする。敏感な問題と言うのは，Xの変化で，1／g（X）が大きく変化するが，しかし，通常は，このdは根が確定しない内に確定する。

重根と近接根は区別できない，と通常言われるが，確かにその通りで，その為に，g（X）の大きさが落ち着く程度に根の近傍である事と言わねばならない。永年，高次代数方程式の数値計算に関心を払っているが，計算誤差がこれ程綺麗な形に表せるとは思わなかった。高次代数方程式の残る問題は，g（X）に集約される。

§6　数値微分

mは，原理的には，m≧1なる任意の有理数に成り得る。そのようなのは，極限操作を理想的に行った時，現われる。例えば，関数F（X）の数値微分を行うとして，q個の関数値を使って，r階の微分値を出すとする。この時，q≧r＋1≧2 でなければならない。そうすると，数値微分と言うのは，微分点Xを中心に形式的に，$F(X+B_k H)$ を展開して，この両辺に重率Ak を掛けて，和を作る。

$$\sum_{k=1}^{q} A_k F(X+B_k H) = \sum_{k=1}^{q} A_k \left[\sum_{i=0}^{\infty} \frac{(B_k)^i H^i}{i!} F^{(i)}(X)\right]$$

これを和の順序を変えて，

$$
\begin{aligned}
\sum_{k=1}^{q} A_k F(X+B_k H) &= \sum_{i=0}^{r-1} \left[\sum_{k=1}^{q} A_k (B_k)^i\right] \frac{H^i}{i!} F^{(i)}(X)\\
&+ H^r F^{(r)}(X)\\
&+ \sum_{i=r+1}^{q-1} \left[\sum_{k=1}^{q} A_k (B_k)^i\right] \frac{H^i}{i!} F^{(i)}(X)\\
&+ \sum_{i=q}^{\infty} \left[\sum_{k=1}^{q} A_k (B_k)^i\right] \frac{H^i}{i!} F^{(i)}(X)
\end{aligned}
$$

とする。ここで，第1項と第3項が0になり，第2項は丁度この様になる様にA_k，B_kを決めて，入れ換え，H^r で割ると，

$$
F^{(r)}(X) = \frac{\sum_{k=1}^{q} A_k F(X+B_k H)}{H^r} - \sum_{i=q}^{\infty} \left[\sum_{k=1}^{q} A_k (B_k)^i\right] \frac{H^{i-r}}{i!} F^{(i)}(X)
$$

と成る。第１項のΣで桁落ちが起きて，丸め誤差が顕著に成る。そこで，どの程度の誤差を見積もれば良いか，と言うと，ｍａｘ〔｜A_k Ｆ（Ｘ＋B_k Ｈ）｜〕（これをＡと置く）の末尾程度に見る訳である。言葉を変えれば，Ａｕを丸め誤差と見る訳である。第２項は打ち切り誤差であるが，最初の項だけで良い。（このＨの他のものをＢと置く。） 誤差は丸め誤差と打ち切り誤差から成るから，結局，

$$E \fallingdotseq Au/H^{r} + BH^{q-r}$$

と表せる。ここで，この誤差を最小にする様にキザミＨを取る。これは最適キザミと言うが，Ｈについて微分して，０と置くと，

$$H^{q} = rAu/[(q-r)B] = C\cdot u \quad \text{（Ｃは桁数に関係しない定数）}$$

と成る。最適キザミは，演算桁数の１／ｑ桁付近の値と言う事である。これをＥの右辺に代入して，定数（損失桁の算出には必要であるが，今は省略する）を別にすれば，Ｅはｕの（ｑ－ｒ）／ｑ乗に成る。これは，縮率では，ｍ＝ｑ／（ｑ－ｒ）に成る事を意味する。この様に原理的には，縮率は，１より大きな任意の有理数になる。

ｒ階の数値微分値は精度が悪いと言われるが，この為である。精度公式に当てはめると，

$$d = (L-\alpha)[(q-r)/q]$$
$$= (L-\alpha)(1-r/q)$$

ここに，ｒ階の微分値
ｑ個の関数値
αは関数の操作での損失桁
Ｌは計算桁数
ｄはｒ階の微分値の精度桁

と成るから，微分が出るギリギリに関数値を取ると，最適キザミを取っても，１／（ｒ＋１）倍桁以下しか残らない事に成る。例えば，５階の微分値を６個の関数値で計算するとすれば，倍精度（１６桁）の計算で３桁も残らない事に成る。最適キザミを取り，損失桁がある事を考慮に入れると，数値微分の精度は非常に悪いと言える。数値微分は精度を考慮して関数値を何個の公式を選ぶべきか，良く検討する必要がある。

§７　連立１次方程式（１）

連立１次方程式などは，どう成っているかと言えば，連立１次方程式は，ｍ＝１ である。
連立１次方程式

$$Ax=b \quad \cdots\cdots\cdots\cdots (1)$$

を考える。連立１次方程式の演算桁数と解精度の関係は，良く知られているように

$$d = p - \alpha \quad \cdots\cdots\cdots\cdots (2)$$

ここに，$\alpha = \log_{M}\{\mathrm{cond}(A)\}$

である。

§８　行列分解

仮定1

ベクトルXに正規直交行列を乗じても，Xの精度は，あまりかわらない。

仮定2

行列Aに正規直交行列を乗じても，行列Aの大きさは，あまりかわらない。

任意の正規直交行列Uは，任意のベクトルXに対して

$$\| XU \| = \| X \|$$

であるので，ベクトルXは，正規直交行列を乗ずることで，あまり大きく変わらない。同様に，行列と正規直交行列を乗じても，行列は，あまり大きく変わらない。

行列Aに正規直交行列Uを乗じで，行列を単純な形，即ち，対角行列に変形する。固有ベクトルは正規直交行列を成す。対称行列$A^T A$の固有値（対称行列の固有値は非負）の非負平方根を行列Aの特異値と言うが，この特異値を

$$\mu_1 \geqq \mu_2 \geqq \mu_3 \geqq \cdots \geqq \mu_r > \mu_{r+1} = \cdots = \mu_n = 0 \quad \cdots\cdots (1)$$

とする。Aが正則ならば，$r=n$ である。U_i を固有ベクトルとして，式で表わせば，

$$A^T A U_i = \mu_i{}^2 U_i \quad \cdots\cdots (2)$$

である。さらに，V_i を次の様に定義する。

$$V_i = A U_i / \mu_i \quad \cdots\cdots (3)$$

これを上の式に代入すれば，

〔左辺〕$= A^T A U_i = A^T (\mu_i V_i) = \mu_i A^T V_i$

〔右辺〕$= \mu_i{}^2 U_i$

であるから，直ちに

$$U_i = A^T V_i / \mu_i \quad \cdots\cdots (4)$$

が得られる。U_i を正規化して置き，V_i の内積を作ると，

$$\begin{aligned}(V_i, V_j) &= (AU_i/\mu_i, AU_i/\mu_j) \\ &= (AU_i, AU_i)/(\mu_i \mu_j) \\ &= (A^T AU_i, U_i)/(\mu_i \mu_j) \\ &= (\mu_i{}^2 U_i, U_j)/(\mu_i \mu_j) \\ &= \mu_i (U_i, U_j)/\mu_j \\ &= \mu_i \delta_{ij}/\mu_j \\ &= \delta_{ij} \quad \cdots\cdots (5)\end{aligned}$$

と成る。V_i も正規直交行列である。U_i は，行列Aの右固有ベクトル，V_i は，行列Aの左固有ベクトルとも言う。行列Aが対称行列ならば，特異値は，固有値である。

§9　連立1次方程式（2）

連立1次方程式を

$$AX = F \quad \cdots\cdots (1)$$

とし，ベクトルFをV_k で分解すると

$$F = \Sigma c_k V_k \quad \cdots\cdots (2)$$

となる。FとV_k の内積を作れば，

$$(F, V_k) = c_k \quad \cdots\cdots\cdots\cdots\cdots \quad (3)$$

と成る。解Xは，U_k で分解すると

$$X = \Sigma d_k U_k \quad \cdots\cdots\cdots\cdots\cdots \quad (4)$$

となる。それぞれ，（１）式の連立１次方程式に代入して，

$$A \Sigma d_k U_k = \Sigma c_k V_k$$

$$\therefore \ \Sigma d_k AU_k = \Sigma c_k V_k \quad \cdots\cdots\cdots\cdots\cdots \quad (5)$$

と成る。$AU_k = \mu_k V_k$ から，

$$\Sigma d_k \mu_k V_k = \Sigma c_k V_k \quad \cdots\cdots\cdots\cdots\cdots \quad (6)$$

が得られ，従って，

$$d_k \mu_k = c_k \quad \therefore \ d_k = c_k / \mu_k \quad \cdots\cdots\cdots\cdots \quad (7)$$

が得られる。直交変換に依る計算は，余り誤差が入らず，c_k の損失桁は，

$$\alpha_k = \log_M \{ \| F \| / c_k \}$$

程度で，特異値 μ_k の損失桁は，

$$\alpha_k = \log_M \{ \mu_1 / \mu_k \}$$

程度である。従って，d_k の損失桁は，

$$\alpha_k = \log_M \{ (\| F \| / c_k) \cdot (\mu_1 / \mu_k) \} \quad \cdots\cdots \quad (8)$$

と成る。連立１次方程式で最も問題なのは，小さい特異値μ_n に対するd_n である。解もU_n の方へずれる。

§１０　対称行列の固有値

基本的な考え方は，ヤコビ法を考えれば良い。対称行列Aの固有値を大きさの順に並べて

$$\lambda_1 \geqq \lambda_2 \geqq \lambda_3 \geqq \cdots \geqq \lambda_{n-1} \geqq \lambda_n \quad \cdots\cdots\cdots\cdots \quad (1)$$

とする。ヤコビ法は，行列Aの右から正規直交行列U_k を，左から U_k^T を乗じ，遂には，非対角項を十分な精度で零にする。次の仮定は，今考えている計算桁数に対して，十分成り立つ。

仮定１：正規直交行列の乗算は誤差はあまり入らない。

仮定２：最大固有値もあまり誤差は入らない。

従って，最大固有値に対して，問題にしている固有値λの大きさが問題である。即ち，損失桁数は，λ_1 / λ の対数を取ったものである。

固有値の精度（桁）：　$d = L - \alpha$

損失桁数　：　$\alpha = \log_M (\lambda_1 / \lambda)$

$= \log_M$（〔最大固有値〕／〔今考えている固有値〕）

計算桁数　：　L 桁

§１１　対称行列の固有ベクトル

基本的な考え方は，収束状況に於ける逆反復法を考えれば良い。

対称行列Aの固有値をλ_k，固有ベクトルをX_k とする。また，$U = (X_1, X_2, X_3,$

・・・，X_n）とする。

基本的な考え方は，収束状況に於ける逆反復を考える事である。

固有値と固有ベクトルの関係は，

$$(A-\lambda_k E) X_k = 0 \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (1)$$

ただし，Eは単位行列，$k=1, 2, \cdots n$

である。今，考えている固有値をλとし，固有ベクトルをXとする。固有ベクトルは誤差を含んでいるから，それを

$$Y = X + \delta X \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (2)$$

とする。固有値は正確に求まっていると考える。この時，関係式は

$$(A-\lambda E) Y = \Delta X \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (3)$$

とする。ΔXは，左辺の【相対ゼロ】である。その大きさは，（A－λE）を対角化し，Yを正規化して考えれは，次のように考える事が出来る。

$$\Delta X = \max\{\ |\lambda_k - \lambda|\ \}\varepsilon \quad \cdots\cdots\cdots\cdots\cdots \quad (4)$$

ただし，εは【相対ゼロ】ベクトルで，$\|\varepsilon\| \leqq u$

このように考えて，（3）式を連立1次方程式として解く。

右辺のΔXをX_kに分解する。

$$\Delta X = \Sigma c_k X_k \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (5)$$

ここに，$c_k = (\Delta X, X_k)$

また，YもX_kで分解する。

$$Y = \Sigma d_k X_k \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (6)$$

（5）式，（6）式を（3）式に代入して，

$$(A-\lambda E)\Sigma d_k X_k = \Sigma c_k X_k$$

$$\therefore\ \Sigma d_k (A-\lambda E) X_k = \Sigma c_k X_k$$

ここで，$AX_k = \lambda_k X_k$　を考えれば，$(A-\lambda E)\Sigma d_k X_k = \Sigma d_k (AX_k - \lambda X_k)$
$= \Sigma d_k (\lambda_k X_k - \lambda X_k) = \Sigma d_k (\lambda_k - \lambda) X_k$ となるから

$$\Sigma d_k (\lambda_k - \lambda) X_k = \Sigma c_k X_k$$

$$\therefore\ d_k (\lambda_k - \lambda) = c_k \quad ; k = 1, 2, 3, \cdots, n$$

$$\therefore\ d_k = c_k / (\lambda_k - \lambda) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (7)$$

ここで，$|\lambda_k - \lambda|$の小さい値から，番号を付ける。1番小さいλの所は0であるが，数値的には，【絶対零】はないから，最小の値は，【相対零】を考えるべきで

$$|\lambda_1 - \lambda| = |\lambda| u$$

である。2番目に小さいのは

$$|\lambda_2 - \lambda|$$

である。それ以上は，今の場合，考えなくても良い。（6）式は，結局，（7）式を代入して次のような大きさと考える事が出来る。

$$Y = c_1 X_1 / \{|\lambda| u\} + c_2 X_2 / \{|\lambda_2 - \lambda|\}$$

c_kは，おおよそ，$\max\{|\lambda_k - \lambda|\} u$　であり，これを代入すれば

$$Y = \max\{|\lambda_k - \lambda|\} u [X_1 / \{|\lambda| u\} + X_2 / \{|\lambda_2 - \lambda|\}]$$

$$= \max\{|(\lambda_k - \lambda)/\lambda|\} [X_1 + X_2 / \{|\lambda u / (\lambda_2 - \lambda)|\}]$$

更に，$\max\{|(\lambda_k - \lambda)/\lambda|\} \sim 1$と考えれば，

$$Y \sim [X_1 + X_2 / \{ | \lambda u / (\lambda_2 - \lambda) | \}] \quad \cdots\cdots\cdots (8)$$

となる。第1項X_1は，今，考えている固有値の固有ベクトルであり，第2項は誤差項である。$\| X_1 \| = 1$であり，$\| X_2 \| = 1$であるから，精度は誤差項の逆数の対数を取れば良い。

固有ベクトルの精度は，次の通りである。

精度桁： $d = L - \alpha \quad \cdots\cdots\cdots\cdots\cdots\cdots (9)$

損失桁： $\alpha = \log_M | \lambda / (\lambda_2 - \lambda) |$

計算桁： L桁

結局，固有ベクトルの損失桁は，一致する固有値の桁数である。

§12 あとがき

関心がある算法で精度公式が成立つのは，今の所，反復式が成立つ場合である。今後も他の算法の損失桁や縮率を確定したいと考えている。

本文を草するにあたり富士通（株）国際情報社会科学研究所の鈴木千里氏に何かとお世話になった。ここに甚大なる謝意を表明する。

§13 文献

〔1〕G. E. Forsythe : Singularity and Near Singularity in Numerical Analysis , Amer. Math. Mohthly 65, April 1958

〔2〕P. H. Sterbenz : Floating-point Computation , Prentice-Hall , 1974.

〔3〕J. H. Wilkinson／一松信，四条忠雄 共訳： 基本的演算における丸め誤差解析，培風館，１９７４（原著１９６３）.

〔4〕G. E. Forsythe & C. B. Moler ／渋谷政昭，田辺国士 共訳 線形計算の基礎： 培風館，１９６９（原著１９６７）

〔5〕宇野利雄： 計算機のための数値計算， 朝倉書店， １９６３.

〔6〕宇野利雄： 誤差伝播の問題，数学，Vol. 15, No. 1, １９６３・７

〔7〕永坂秀子，福井義成： 数値微分の誤差，情報処理学会・論文誌，Vol. 22, No ，１９８１.

〔8〕永坂秀子： 計算機と数値計算， 朝倉書店， １９８０・３

〔9〕山下： $\sum_{i=1}^{n} X_i$の精度判定とその応用， 電子協，NA委員会資料，No.8 ，１９６４.

〔10〕山下，佐竹： 高次代数方程式の根の計算限界について，情報処理， Vol.7, No. 4 , pp.197〜201. １９６６.

〔11〕山下： 浮動小数点演算における誤差評価について，情報処理， Vol. 15, No. 12, pp.935〜939 , Dec.１９７４.

〔12〕山下： 浮動小数点演算における誤差評価と誤差消失について 京大数解研，「計算の手間と能率」研究集会予稿集，１９７４・２

〔13〕山下： 演算桁数と解の精度の関係について，情報処理学会・第１５回大会（1974）.

〔14〕山下： 浮動小数点演算の誤差，共立出版，bit 臨時増刊，数値計算における誤差，pp. 14〜26, 12（1975）.