数理解析研究所講究録
153巻 1972年 58-128

浮動小数点型演算による反復法のcomponent-wiseな誤差解析

九大 理　占部 実

まえがき

与えられた方程式を

$$\varphi_i(x_1, x_2, \dots, x_d)=0 \qquad (i=1, 2, \dots, d),$$

あるいはベクトル記法を用いて

$$(0.1) \qquad \varphi(x)=0$$

とする．この方程式を反復法を用いて解くのには，上の方程式を

$$(0.2) \qquad x=f(x)$$

の形に書き直し，適当な$x^{(0)}$から出発して

$$(0.3) \qquad x^{(n+1)}=f[x^{(n)}] \qquad (n=0, 1, 2, \dots)$$

により，$x^{(1)}$，$x^{(2)}$，… を逐次求め，適当なところでこの反復計算を止めて，最後に得られた $x^{(n)}$ をもって望む近似解とする．

$f[x^{(n)}]$ が固定小数点型演算で計算される場合，上の反復法が実際計算ではどのようになるかは，一応［1］，［2］で論じたので，ここでは $f[x^{(n)}]$ が浮動小数点型演算で計算される場合について考察する．

なお，連立方程式の解の誤差を考察するとき，誤差評価には普通，誤差ベクトルのノルムが用いられているが，これはしばしば過大評価を与えることになって，実際の誤差評価には必ずしも適切でない．そこで，ここでは，誤差評価を成分毎に行う方法をとった．

われわれの結果は，以下［注意］で述べているように，ノルムを用いる場合にも成り立つので，必要があれば，ノルムを用いる場合にそのまま適用して差し支えない．

§1. 予備

定義1．"任意のベクトル $v=(v_1, v_2, \dots, v_d)^T$（T は転置を表わす）に対して，ベクトル $(|v_1|, |v_2|, \dots, |v_d|)^T$ を v の絶対ベクトルとよび，$\sigma(v)$ で表わす．"

$\sigma(v)$ の第 i 成分を $\sigma_i(v)$ で表わすと，明らかに

$$\sigma_i(v) = |v_i| \quad (i=1, 2, \dots, d)$$

である．

<u>定義2.</u> "任意の行列 $A=(a_{ij})$ $(i, j=1, 2, \dots, d)$ に対して，ij - 要素が $|a_{ij}|$ である行列を A の<u>絶対行列</u>とよび，$\sigma(A)$ で表わす．"

$\sigma(A)$ の ij - 要素を $\sigma_{ij}(A)$ で表わすと，明らかに

$$\sigma_{ij}(A) = |a_{ij}| \quad (i, j=1, 2, \dots, d)$$

である．

<u>定義3.</u> "ベクトル $v=(v_1, v_2, \dots, v_d)^T$, $w=(w_1, w_2, \dots, w_d)^T$ に対して $v_i \geq w_i$ $(i=1, 2, \dots, d)$ が成り立つとき，$v \geq w$ と書く。また，行列 $A=(a_{ij})$, $B=(b_{ij})$ に対して $a_{ij} \geq b_{ij}$ $(i, j=1, 2, \dots, d)$ が成り立つとき，$A \geq B$ と書く．"

もちろん，$v \geq w$ のときには，$w \leq v$ とも書き，$A \geq B$ のときには $B \leq A$ とも書く。上の定義から明らかなように，

$$v_i \geq 0 \quad (i=1, 2, \dots, d) \quad と \quad v \geq 0,$$

$$a_{ij} \geq 0 \quad (i,j=1,2,\cdots,d) \quad と \quad A \geq 0,$$

はそれぞれ互に同値であり，また

$$v \geq w \quad と \quad v - w \geq 0,$$
$$A \geq B \quad と \quad A - B \geq 0,$$

はそれぞれ互に同値である．

<u>補題1.</u> "v, wをベクトル，A，Bを行列とするとき，

$$\sigma(v+w) \leq \sigma(v) + \sigma(w),$$
$$\sigma(A+B) \leq \sigma(A) + \sigma(B),$$
$$\sigma(Av) \leq \sigma(A)\sigma(v),$$
$$\sigma(AB) \leq \sigma(A)\sigma(B).$$

また，$\sigma(v) \geq 0$, $\sigma(A) \geq 0$ でしかも

$$\sigma(v) = 0 \quad と \quad v = 0,$$
$$\sigma(A) = 0 \quad と \quad A = 0,$$

はそれぞれ互に同値である．"

この補題は，定義1，2，3により明らかである．

つぎの補題はよく知られていることであるが，以下でしばしば用いるので，ここに一応あげておく．

補題2. "任意の正方行列Hに対して，それのspectral radiusを$\rho(H)$で表わすと，$\rho(H)<1$ならば $\lim_{n\to\infty} H^n=0$ であり，また逆に $\lim_{n\to\infty} H^n=0$ ならば $\rho(H)<1$ である．ただしここで nは正の整数である．"

補題3. "任意の正方行列Hに対して，$\rho(H)<1$ ならば，$(E-H)^{-1}$ （Eは単位行列）が存在し，

$$(E-H)^{-1}=\sum_{n=0}^{\infty} H^n$$

が成立する．"

補題3から，正方行列$H\geq 0$ に対して$\rho(H)<1$ ならば，$(E-H)^{-1}\geq 0$ であることが，すぐわかる．

［注意］ $\sigma(\cdot)$をベクトル，あるいは行列のノルムとしても，補題1はそのまま成り立つ．このことは，以下の議論において，$\sigma(\cdot)$をベクトル，あるいは行列のノルムとしても結果はそのまま成りたつことを示すことになる．

§2. 反復法の収束定理

定理1. "$f(x)$ はx－空間の集合Dで定義され，Dに属する任意のx', x'' に対して

$$(2\cdot1)\quad \sigma[f(x')-f(x'')]\leq K_0\,\sigma(x'-x'')$$

をみたす，とする．ただし K_0 は正方行列で

$$(2\cdot2)\quad K_0\geq 0,\quad \rho(K_0)<1$$

をみたすものとする．

Dに属し，次の条件をみたすような $x^{(0)}$ が存在するとする：

$$(2\cdot3)\quad S=\left\{h\,\middle|\,\sigma[h-x^{(1)}]\leq K_0(E-K_0)^{-1}\sigma[x^{(1)}-x^{(0)}]\right\}\subset D.$$

ただしここで

$$(2\cdot4)\quad x^{(1)}=f[x^{(0)}]$$

である．

このとき，

$$(2\cdot5)\quad x^{(n+1)}=f[x^{(n)}]\quad (n=0,1,2,\cdots)$$

によって定まる列 $\{x^{(n)}\}$ $(n=1,2,\cdots)$ は S の中にあって収束し，その極限

$$(2\cdot6)\quad \hat{x}=\lim_{n\to\infty}x^{(n)}$$

は，方程式

$$(2\cdot 7) \qquad x = f(x)$$

のDにおける唯一の解である."

[証明] まず,(2・5)によりSの中に無限点列$\{x^{(n)}\}$が得られることを示そう.

$x^{(1)} \in S$ は明らかであるから, $x^{(1)}, x^{(2)}, \dots, x^{(n)}$ がSの中に得られた,とする.すると, $x^{(n)} \in S \subset D$ であるから, $x^{(n+1)} = f[x^{(n)}]$ によって $x^{(n+1)}$ が得られる.問題はこの $x^{(n+1)}$ がSに属するか,ということである.まず,(2・1)により

$$\sigma[x^{(2)} - x^{(1)}] = \sigma[f(x^{(1)}) - f(x^{(0)})] \leq K_0 \sigma[x^{(1)} - x^{(0)}]$$

を得,同様にして

$$\sigma[x^{(3)} - x^{(2)}] \leq K_0 \sigma[x^{(2)} - x^{(1)}] \leq K_0^2 \sigma[x^{(1)} - x^{(0)}],$$

$$\cdots\cdots\cdots\cdots\cdots$$

$$\sigma[x^{(n+1)} - x^{(n)}] \leq K_0 \sigma[x^{(n)} - x^{(n-1)}] \leq K_0^n \sigma[x^{(1)} - x^{(0)}]$$

を得る.すると,これらの式より,つぎの不等式を得る:

$$\sigma[x^{(n+1)} - x^{(1)}] \leq \sigma[x^{(n+1)} - x^{(n)}] + \sigma[x^{(n)} - x^{(n-1)}] + \cdots + \sigma[x^{(2)} - x^{(1)}]$$

$$\leq (K_0^n + K_0^{n-1} + \cdots + K_0)\sigma[x^{(1)} - x^{(0)}]$$
$$\leq K_0(E-K_0)^{-1}\sigma[x^{(1)} - x^{(0)}].$$

Sの定義により，これは $x^{(n+1)} \in S$ を表わしている．かくて，帰納法により，(2・5)によってSの中に無限点列$\{x^{(n)}\}$が得られることがわかる．

上の結果から，任意の正の整数rに対して

$$(2\cdot 8)\quad \sigma[x^{(n+r)} - x^{(n)}] \leq \sigma[x^{(n+r)} - x^{(n+r-1)}] + \cdots + \sigma[x^{(n+1)} - x^{(n)}]$$
$$\leq (K_0^{n+r-1} + \cdots + K_0^n)\sigma[x^{(1)} - x^{(0)}]$$
$$\leq K_0^n(E-K_0)^{-1}\sigma[x^{(1)} - x^{(0)}]$$

を得る．$\rho(K_0) < 1$ であるから，補題2により，

$$\lim_{n\to\infty} K_0^n = 0$$

となり，任意の正の整数rに対して，$n\to\infty$のとき

$$\sigma[x^{(n+r)} - x^{(n)}] \to 0$$

を得る．これは任意の正の整数rに対して

$$\lim_{n\to\infty} |x_i^{(n+r)} - x_i^{(n)}| = 0 \quad (i=1, 2, \cdots, d)$$

を意味するから，

$$\hat{x}_i = \lim_{n\to\infty} x_i^{(n)} \qquad (i=1, 2, \cdots, d)$$

が存在する．$\hat{x} = (\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_d)^T$ とおけば，上式は

$$(2\cdot 9) \qquad \hat{x} = \lim_{n\to\infty} x^{(n)}$$

を意味する．$x^{(n)} \in S \quad (n=1, 2, \cdots)$ であるから，

$$(2\cdot 10) \qquad \hat{x} \in S \subset D$$

であることは明らかである．

(2・1) により $f(x)$ は D において連続であるから，(2・5) において $n \to \infty$ とすれば，(2・9) により

$$\hat{x} = f(\hat{x})$$

を得る．これは $\hat{x}$ が方程式 (2・7) の解であることを示している．

最後に，方程式 (2・7) は D の中では，上に求められた $\hat{x}$ 以外には解をもたないことを示そう．

いま，$\hat{x}$ 以外に D の中に解 $\hat{x}'$ があったとする．すると，

$$\hat{x}' = f(\hat{x}'), \quad \hat{x} = f(\hat{x})$$

より，(2・1) によって

$$\sigma(\hat{x}'-\hat{x})=\sigma[f(\hat{x}')-f(\hat{x})]\leq K_0\,\sigma(\hat{x}'-\hat{x})$$

を得る．したがって

$$(E-K_0)\,\sigma(\hat{x}'-\hat{x})\leq 0$$

を得る．ところが，$\rho(K_0)<1$ であるから，補題3により $(E-K_0)^{-1}\geq 0$ が存在する．したがって，これを上の不等式に左から掛ければ

$$\sigma(\hat{x}'-\hat{x})\leq 0$$

を得る．$\sigma(\hat{x}'-\hat{x})\geq 0$ であるから，

$$\sigma(\hat{x}'-\hat{x})=0,\quad \text{すなわち}\quad \hat{x}'-\hat{x}=0$$

を得る．これは $\hat{x}'=\hat{x}$ を意味し，解の一意性を示している．

これで定理の証明は完全にできたことになる．

［注意1］ 行列 K_0 の要素を k_{ij} $(i,j=1,2,\dots,d)$ とすると，条件（2・1）は

$$|f_i(x')-f_i(x'')|\leq\sum_{j=1}^{d}k_{ij}\,|x_j'-x_j''|$$

を表わしている．したがって，ベクトル$v=(v_1, v_2, \cdots, v_d)^T$のノルム$\|v\|$として，$\|v\|=\max_i |v_i|$を用いると，行列$K_0$の対応するノルム$\|K_0\|$は$\|K_0\|=\max_i \sum_{j=1}^{d} k_{ij}$（$k_{ij} \geq 0$であるから）となり，条件（2・1）より

$$\|f(x')-f(x'')\| \leq \|K_0\| \cdot \|x'-x''\|$$

が成り立つことになる．

［注意2］ $\sigma(v)$をベクトルvのノルムとすると，K_0はスカラーとなり，（2・2）は

$$0 \leq K_0 < 1$$

を意味する．このとき，定理1の結果がそのまま成立することは上の証明から明らかである．

［注意3］ 条件（2・3）は，$x^{(0)}$が適当な方程式（2・7）の近似解であることを要請しているものである．なぜなら，もし$x^{(0)}$がDの内部にあり，相当な精度をもつ方程式（2・7）の近似解であるならば，（2・4）により$x^{(1)}$は$x^{(0)}$に近く，したがってSは$x^{(1)}$を中心とする小さな長方形領域になり，条件（2・3）はおのずからみたされるからである．

§3．浮動小数点型演算による反復法の誤差解析

3・1　反復法の振動定理

反復計算（0・3）を計算機によって実行するとき，$f(x)$ の計算は浮動小数点型演算で行われるものとし，得られる値を $\tilde{f}(x)$ で表わす．$\tilde{f}(x)$ は一般に $f(x)$ と異なっているが，いまの場合

$$(3\cdot1)\qquad \sigma[\tilde{f}(x)-f(x)]\leq\varepsilon\cdot\sigma[f(x)]$$

が成立する，と考える．ただしこのとき，ε は対角行列で，それの対角線要素は正数 ε_i $(i=1,2,\dots,d)$ であるとする．この場合，（3・1）は

$$|\tilde{f}_i(x)-f_i(x)|\leq\varepsilon_i\cdot|f_i(x)|\qquad(i=1,2,\dots,d)$$

を意味し，これは $\tilde{f}_i(x)$ の相対誤差が ε_i を越えないことを示している．便宜上，われわれは

$$(3\cdot2)\qquad \varepsilon_i\leq\hat{\varepsilon}\qquad(i=1,2,\dots,d)$$

で，

$$(3\cdot3)\qquad 0<\hat{\varepsilon}<1$$

であるとする．行列 ε の定義，および（3・2）により明らかに

$$(3\cdot4)\qquad 0 \leq \varepsilon \leq \hat{\varepsilon} E$$

が成り立っている．

反復計算（0・3）を計算機の上で実行すると，上に述べたことから，われわれは

$$(3\cdot5)\qquad \tilde{x}^{(n+1)} = \tilde{f}[\tilde{x}^{(n)}] \quad (\tilde{x}^{(0)} = x^{(0)};\ n = 0, 1, 2, \dots)$$

によって定まる列 $\{\tilde{x}^{(n)}\}$ を得，これは（0・3）によって定まる列 $\{x^{(n)}\}$ とは一般に異なっている．この $\{\tilde{x}^{(n)}\}$ の挙動については，つぎの定理が成り立つ．

<u>定理2.</u> "$f(x)$ は定理1の条件をみたしているとする．また $\hat{\varepsilon}$ は小さくて

$$(3\cdot6)\qquad \hat{L}_0 = (1+\hat{\varepsilon})K_0$$

とおくとき，

$$(3\cdot7)\qquad \rho(\hat{L}_0) < 1$$

が成り立つとする．

Dに属し，つぎの条件をみたすような $x^{(0)}$ が存在するとする：

$$(3\cdot 8)\quad \Sigma=\{h \mid \sigma[h-\tilde{x}^{(1)}]\leq \rho_1+\rho_2\}\subset D.$$

ただしここで，

$$(3\cdot 9)\quad \tilde{x}^{(1)}=\tilde{f}[x^{(0)}],$$

$$(3\cdot 10)\quad \rho_1=K_0(E-K_0)^{-1}\sigma[\tilde{x}^{(1)}-x^{(0)}]+(E-K_0)^{-1}\varepsilon(E-\varepsilon)^{-1}\sigma[\tilde{x}^{(1)}],$$

$$(3\cdot 11)\quad \rho_2=(E-L_0)^{-1}\varepsilon\{\rho_1+\sigma[\tilde{x}^{(1)}]\}$$

で，

$$(3\cdot 12)\quad L_0=(E+\varepsilon)K_0$$

である．

このとき，方程式(2・7)はDにおいて一つただ一つの解$\hat{x}$をもち，列$\{\tilde{x}^{(n)}\}$は有限回の反復ののち有限個の値をとって振動し，振動状態に達した$\tilde{x}^{(n)}$に対しては

$$(3\cdot 13)\quad \sigma[\tilde{x}^{(n)}-\hat{x}]\leq(E-L_0)^{-1}\varepsilon\cdot\sigma(\hat{x})$$

が成り立つ．"

［証明］ 初めに，次の不等式が成り立つことを示しておこう：

$$(3\cdot 14)\quad \sigma[\tilde{x}^{(1)}-x^{(1)}]\leq\varepsilon\cdot\sigma[x^{(1)}]\leq\varepsilon(E-\varepsilon)^{-1}\sigma[\tilde{x}^{(1)}].$$

(3・1) により

$$(3\cdot 15)\qquad \sigma[\tilde{x}^{(1)}-x^{(1)}]=\sigma[\tilde{f}(x^{(0)})-f(x^{(0)})]$$
$$\leq \varepsilon\cdot\sigma[f(x^{(0)})]=\varepsilon\cdot\sigma[x^{(1)}]$$

であるから，(3・14) の初めの不等号が成立する．次に $\sigma[x^{(1)}]$ を考えると，

$$\sigma[x^{(1)}]\leq\sigma[x^{(1)}-\tilde{x}^{(1)}]+\sigma[\tilde{x}^{(1)}]$$

であるから，(3・15) より

$$\sigma[x^{(1)}]\leq\varepsilon\cdot\sigma[x^{(1)}]+\sigma[\tilde{x}^{(1)}],$$

すなわち，

$$(E-\varepsilon)\sigma[x^{(1)}]\leq\sigma[\tilde{x}^{(1)}]$$

を得る．(3・3) により $\rho(\varepsilon)<1$ であるから，補題3を用いて

$$(3\cdot 16)\qquad \sigma[x^{(1)}]\leq(E-\varepsilon)^{-1}\sigma[\tilde{x}^{(1)}]$$

を得る．すると，(3・15) より (3・14) の第2の不等号が成立する．これで (3・14) は証明された．

われわれは，まず，(3・14) を利用し，定理1で定義され

たSに対して

$$(3\cdot17)\qquad S\subset\Sigma_0=\{h\mid\sigma[h-\tilde{x}^{(1)}]\leq\rho_1\}$$

が成り立つことを示そう．Sに属する任意のhに対しては，

$$\sigma[h-x^{(1)}]\leq K_0(E-K_0)^{-1}\sigma[x^{(1)}-x^{(0)}]$$

が成り立つから，(3·14)，(3·10)により逐次つぎの不等式を得る：

$$\begin{aligned}\sigma[h-\tilde{x}^{(1)}]&\leq\sigma[h-x^{(1)}]+\sigma[x^{(1)}-\tilde{x}^{(1)}]\\&\leq K_0(E-K_0)^{-1}\sigma[x^{(1)}-x^{(0)}]+\sigma[x^{(1)}-\tilde{x}^{(1)}]\\&\leq K_0(E-K_0)^{-1}\sigma[x^{(1)}-\tilde{x}^{(1)}]+K_0(E-K_0)^{-1}\sigma[\tilde{x}^{(1)}-x^{(0)}]\\&\qquad+\sigma[x^{(1)}-\tilde{x}^{(1)}]\\&=K_0(E-K_0)^{-1}\sigma[\tilde{x}^{(1)}-x^{(0)}]+(E-K_0)^{-1}\sigma[x^{(1)}-\tilde{x}^{(1)}]\\&\leq K_0(E-K_0)^{-1}\sigma[\tilde{x}^{(1)}-x^{(0)}]+(E-K_0)^{-1}\varepsilon(E-\varepsilon)^{-1}\sigma[\tilde{x}^{(1)}]\\&=\rho_1.\end{aligned}$$

これは $h\in\Sigma_0$ を意味するから，われわれは(3·17)を得る．

(3·17)が成り立てば，われわれの仮定(3·8)により

$$(3\cdot18)\qquad S\subset\Sigma_0\subset\Sigma\subset D$$

が成り立つから，定理1の条件はすべてみたされることになり，定理1の結果がすべて成り立ち，方程式(2・7)はDにおいて一つただ一つの解$\hat{x}$をもつことになる．

さて，数値計算による反復演算(3・5)によってΣの中に無限点列$\{\tilde{x}^{(n)}\}$が得られることを示そう．

$\tilde{x}^{(1)} \in \Sigma$は明らかであるから，$\tilde{x}^{(1)}, \tilde{x}^{(2)}, \cdots, \tilde{x}^{(n)}$が$\Sigma$の中に得られた，とする．すると，$\tilde{x}^{(n)} \in \Sigma \subset D$であるから，$\tilde{x}^{(n+1)} = \tilde{f}[\tilde{x}^{(n)}]$によって$\tilde{x}^{(n+1)}$が得られる．問題はこの$\tilde{x}^{(n+1)}$が$\Sigma$に属するか，ということである．

さて，(3・1)，(2・1)，(3・12)，(3・14)により

$$
\begin{aligned}
(3\cdot 19)\quad \sigma[\tilde{x}^{(2)} - x^{(2)}] &= \sigma[\tilde{f}(\tilde{x}^{(1)}) - f(x^{(1)})] \\
&\leq \sigma[\tilde{f}(\tilde{x}^{(1)}) - f(\tilde{x}^{(1)})] + \sigma[f(\tilde{x}^{(1)}) - f(x^{(1)})] \\
&\leq \varepsilon \cdot \sigma[f(\tilde{x}^{(1)})] + K_0 \sigma[\tilde{x}^{(1)} - x^{(1)}] \\
&\leq \varepsilon \{\sigma[f(\tilde{x}^{(1)}) - f(x^{(1)})] + \sigma[f(x^{(1)})]\} \\
&\qquad + K_0 \sigma[\tilde{x}^{(1)} - x^{(1)}] \\
&\leq (E + \varepsilon) K_0 \sigma[\tilde{x}^{(1)} - x^{(1)}] + \varepsilon \sigma[x^{(2)}] \\
&\leq L_0 \varepsilon \sigma[x^{(1)}] + \varepsilon \sigma[x^{(2)}]
\end{aligned}
$$

を得る．同様に

$$\sigma[\tilde{x}^{(3)} - x^{(3)}] \leq L_0 \sigma[\tilde{x}^{(2)} - x^{(2)}] + \varepsilon \sigma[x^{(3)}]$$

を得，(3·19) より

$$\sigma[\tilde{x}^{(3)}-x^{(3)}] \leq L_0^2 \varepsilon \sigma[x^{(1)}] + L_0 \varepsilon \sigma[x^{(2)}] + \varepsilon \sigma[x^{(3)}]$$

を得る．これを続けてゆくと，

$$(3\cdot 20) \quad \sigma[\tilde{x}^{(n+1)}-x^{(n+1)}] \leq L_0^n \varepsilon \sigma[x^{(1)}] + L_0^{n-1} \varepsilon \sigma[x^{(2)}] + \cdots + \varepsilon \sigma[x^{(n+1)}]$$

が得られる．ところで，定理 1 により $x^{(n)} \in S$ $(n=1, 2, \cdots)$ であるから，

$$\sigma[x^{(n)}-x^{(1)}] \leq K_0 (E-K_0)^{-1} \sigma[x^{(1)}-x^{(0)}] \quad (n=1, 2, \cdots)$$

が成り立つ．すると，

$$\begin{aligned}
\sigma[x^{(n)}] &\leq \sigma[x^{(n)}-x^{(1)}] + \sigma[x^{(1)}] \\
&\leq K_0 (E-K_0)^{-1} \sigma[x^{(1)}-x^{(0)}] + \sigma[x^{(1)}] \\
&\leq K_0 (E-K_0)^{-1} \sigma[x^{(1)}-\tilde{x}^{(1)}] + K_0 (E-K_0)^{-1} \sigma[\tilde{x}^{(1)}-x^{(0)}] \\
&\qquad + \sigma[x^{(1)}-\tilde{x}^{(1)}] + \sigma[\tilde{x}^{(1)}] \\
&= (E-K_0)^{-1} \sigma[x^{(1)}-\tilde{x}^{(1)}] + K_0 (E-K_0)^{-1} \sigma[\tilde{x}^{(1)}-x^{(0)}] \\
&\qquad + \sigma[\tilde{x}^{(1)}]
\end{aligned}$$

となるから，(3·14) により，

$$\sigma[x^{(n)}] \leq K_0(E-K_0)^{-1}\sigma[\tilde{x}^{(1)}-x^{(0)}]+(E-K_0)^{-1}\varepsilon(E-\varepsilon)^{-1}\sigma[\tilde{x}^{(1)}] + \sigma[\tilde{x}^{(1)}]$$

$$= \rho_1 + \sigma[\tilde{x}^{(1)}]$$

を得る．かくて，(3・20) よりわれわれは次の不等式を得る:

(3・21) $\sigma[\tilde{x}^{(n+1)}-x^{(n+1)}] \leq (L_0^n + L_0^{n-1} + \cdots + E)\varepsilon\{\rho_1 + \sigma[\tilde{x}^{(1)}]\}.$

ところで，(3・4)，(3・6)，(3・12) により

(3・22) $L_0 = (E+\varepsilon)K_0 \leq (1+\varepsilon')K_0 = \hat{L}_0$

である．(3・7) から，補題2によって $\lim_{n\to\infty} \hat{L}_0^n = 0$ である．すると，(3・22) により

$$0 \leq L_0 \leq \hat{L}_0$$

であるから，$\lim_{n\to\infty} L_0^n = 0$ となり，補題2によって

(3・23) $\rho(L_0) < 1$

となる．すると，補題3により

(3・24) $E + L_0 + L_0^2 + \cdots + L_0^n \leq E + L_0 + L_0^2 + \cdots$
$= (E-L_0)^{-1}$

となるので，(3・21)より(3・11)によって

(3・25)　$\sigma[\tilde{x}^{(n+1)}-x^{(n+1)}]\leq\rho_2$

を得る．$x^{(n+1)}\in S\subset\Sigma_0$ であるから，(3・25)より

$$\sigma[\tilde{x}^{(n+1)}-\tilde{x}^{(1)}]\leq\rho_1+\rho_2$$

を得る．これはΣの定義により $\tilde{x}^{(n+1)}\in\Sigma$ であることを示している．

上の結果から，帰納法により，(3・5)によってΣの中に無限点列$\{\tilde{x}^{(n)}\}$が得られることがわかる．

さて，Σはd-次元ユークリッド空間の有界な閉領域で，$\tilde{x}^{(n)}$の成分$\tilde{x}_i^{(n)}$　$(i=1,2,\dots,d)$は計算機によって得られる有限桁，有限指数の数である．したがって，$\tilde{x}^{(n)}$がことごとく異なる値をとる，ということはできない．すると，必ず，ある正の整数pに対して

(3・26)　$\tilde{x}^{(m+p)}=\tilde{x}^{(m)}$

が成り立つようなmが存在することになる．一旦(3・26)が起ると，

$$\tilde{x}^{(m+p+r)}=\tilde{x}^{(m+r)}\qquad(r=1,2,\dots)$$

となるので，$\{\tilde{x}^{(n)}\}$ は $\tilde{x}^{(m)}$ 以後は p 個の値

$$\tilde{x}^{(m)},\ \tilde{x}^{(m+1)},\ \dots,\ \tilde{x}^{(m+p-1)}$$

をとって振動することになる．

いま，$\tilde{x}^{(m)}$ を振動状態にある任意の $\tilde{x}^{(n)}$ の値としよう．

(3・20)の証明からわかるように，任意の $\tilde{x}^{(n)}$ に対しては

$$(3\cdot 27)\quad \sigma[\tilde{x}^{(n)}-x^{(n)}]\leq L_0^{n-1}\varepsilon\sigma[x^{(1)}]+L_0^{n-2}\varepsilon\sigma[x^{(2)}]+\dots+\varepsilon\sigma[x^{(n)}]$$

が成り立つ．ところが，定理1の証明中の(2・8)において，$r\to\infty$ とすると，(2・9)により

$$(3\cdot 28)\quad \sigma[\hat{x}-x^{(n)}]\leq K_0^{n}(E-K_0)^{-1}\sigma[x^{(1)}-x^{(0)}]$$

を得る．したがって

$$\sigma[x^{(n)}]\leq\sigma[x^{(n)}-\hat{x}]+\sigma[\hat{x}]$$

より

$$(3\cdot 29)\quad \sigma[x^{(n)}]\leq K_0^{n}(E-K_0)^{-1}\sigma[x^{(1)}-x^{(0)}]+\sigma[\hat{x}]$$

を得る．すると，(3・27)より次の不等式が得られる：

$$(3\cdot30)\quad \sigma[\tilde{x}^{(n)}-x^{(n)}]\leq L_0^{n-1}\varepsilon K_0(E-K_0)^{-1}\sigma[x^{(1)}-x^{(0)}]+L_0^{n-1}\varepsilon\sigma[\hat{x}]$$
$$+L_0^{n-2}\varepsilon K_0^2(E-K_0)^{-1}\sigma[x^{(1)}-x^{(0)}]+L_0^{n-2}\varepsilon\sigma[\hat{x}]$$
$$\cdots\cdots\cdots\cdots$$
$$+\varepsilon K_0^n(E-K_0)^{-1}\sigma[x^{(1)}-x^{(0)}]+\varepsilon\sigma[\hat{x}]$$
$$=(L_0^{n-1}\varepsilon K_0+L_0^{n-2}\varepsilon K_0^2+\cdots+\varepsilon K_0^n)(E-K_0)^{-1}\sigma[x^{(1)}-x^{(0)}]$$
$$+(L_0^{n-1}+L_0^{n-2}+\cdots+E)\varepsilon\sigma[\hat{x}].$$

さて，(3·4)，(3·6)，(3·12) により次の式が成り立つ：

$$(3\cdot31)\quad L_0^{n-1}\varepsilon K_0+L_0^{n-2}\varepsilon K_0^2+\cdots+\varepsilon K_0^n$$
$$\leq\hat{\varepsilon}(\hat{L}_0^{n-1}K_0+\hat{L}_0^{n-2}K_0^2+\cdots+K_0^n)$$
$$=\hat{\varepsilon}(\hat{L}_0^{n-1}+\hat{L}_0^{n-2}K_0+\cdots+K_0^{n-1})K_0.$$

ところが，(3·6) により $\hat{L}_0K_0=K_0\hat{L}_0$ であるから，

$$(\hat{L}_0-K_0)(\hat{L}_0^{n-1}+\hat{L}_0^{n-2}K_0+\cdots+K_0^{n-1})=\hat{L}_0^n-K_0^n,$$

すなわち，

$$\hat{\varepsilon}K_0(\hat{L}_0^{n-1}+\hat{L}_0^{n-2}K_0+\cdots+K_0^{n-1})=\hat{L}_0^n-K_0^n$$

を得る．すると，(3·31) より

$$(3\cdot32)\quad 0\leq L_0^{n-1}\varepsilon K_0+L_0^{n-2}\varepsilon K_0^2+\cdots+\varepsilon K_0^n\leq\hat{L}_0^n-K_0^n$$

を得る．ところで，(2·2)，(3·7)により

$$\rho(K_0),\ \rho(\hat{L}_0)<1$$

であるから，補題2により $n\to\infty$ のとき $\hat{L}_0^n-K_0^n\to 0$ となる．かくて(3·32)より

$$(3\cdot 33)\qquad \lim_{n\to\infty}\left(L_0^{n-1}\varepsilon K_0+L_0^{n-2}\varepsilon K_0^2+\cdots+\varepsilon K_0^n\right)=0$$

が得られる．

さて，振動状態にある任意の $\tilde{x}^{(n)}$ に対しては，ある正の整数 p に対して $\tilde{x}^{(n+p)}=\tilde{x}^{(n)}$ が成り立つから，いまの $\tilde{x}^{(m)}$ に対しては $\tilde{x}^{(m+p)}=\tilde{x}^{(m)}$ が成り立つとする．すると，明らかに任意の正の整数 r に対して

$$(3\cdot 34)\qquad \tilde{x}^{(m+rp)}=\tilde{x}^{(m)}$$

が成り立つことになる．そこで(3·30)において

$$n=m+rp$$

とおき，$r\longrightarrow\infty$ とすると，(3·34)，(2·9)，(3·24)，(3·33)により

$$(3\cdot 35)\qquad \sigma[\tilde{x}^{(m)}-\hat{x}]\leq(E-L_0)^{-1}\varepsilon\sigma[\hat{x}]$$

を得る．これは望む不等式(3・13)である．

これで定理は完全に証明されたことになる．

［注意1］ 方程式(2・7)の解$\hat{x}$に対して，$\hat{x}$のある成分$\hat{x}_i$が0であるときは，$f_i(\hat{x})=\hat{x}_i=0$となり，一般には$\tilde{f}_i(\hat{x})\neq 0$であるから，(3・1)は成立しない．このようなことが予知，あるいは予想されるような場合には，もとの方程式でx_iを適当な量だけずらしておいて解くようにすることが望ましい，と考えられる．

［注意2］ 条件(3・8)は，定理1の条件(2・3)と同じく，$x^{(0)}$が適当な方程式(2・7)の近似解であることを要請しているものである．なぜなら，$x^{(0)}$がDの内部にある，相当な精度をもつ方程式(2・7)の近似解であり，$f(x)$の計算は相当精密に行われて$\hat{\varepsilon}$が小さいとすると，(3・9)により$\tilde{x}^{(1)}$は$x^{(0)}$に近くなり，(3・8)，(3・10)，(3・11)によりΣは$\tilde{x}^{(1)}$を中心とする小さな長方形領域になって，条件(3・8)はおのずからみたされるようになるからである．

［注意3］ K_0の各要素が小さくて$K_0\approx 0$である場合には，(3・12)により$L_0\approx 0$となり，(3・13)は

$$\sigma[\tilde{x}^{(n)}-\hat{x}]\leq \varepsilon\cdot\sigma(\hat{x}),$$

すなわち，

$$|\tilde{x}_i^{(n)} - \hat{x}_i| \leq \varepsilon_i |\hat{x}_i| \qquad (i=1, 2, \cdots, d)$$

となる．これは $\tilde{x}^{(n)}$ の各成分 $\tilde{x}_i^{(n)}$ の相対誤差がそれぞれ ε_i を越えないことを示している．このような誤差評価は実際問題ではきわめて便利であり，またノルムを用いての評価より一般にはるかに精密になっている．

［注意4］ $\sigma(v)$ をベクトルのノルムとすると，K_0, ε はともにスカラーになり，前に述べたように (2・2) は

$$0 \leq K_0 < 1$$

を意味し，(3・3), (3・4) は

$$0 \leq \varepsilon \leq \hat{\varepsilon} < 1$$

を意味し，(3・7) は

$$(1+\hat{\varepsilon})K_0 < 1$$

を意味する．このとき，定理2の結果が，定理1の場合と同じく，そのまま成立することは，上の証明から明らかである．

3・2 反復計算の停止

反復計算（0・3）を計算機の上で実行し，（3・5）により定まる $\tilde{x}^{(n)}$ $(n=1, 2, \cdots)$ を求めてゆくとき，$\tilde{x}^{(n)}$ が振動状態に達したかどうかを直接確めるためには，反復計算で得られる $\tilde{x}^{(n)}$ をすべて記憶装置にとっておかなければならない。これは実際計算にはあまり便利ではない。そこで，次の不等式を考えて，これがみたされると，そこで反復計算を停止する，という方法がよく行われる：

$$(3\cdot 36) \qquad \sigma[\tilde{x}^{(n+1)}-\tilde{x}^{(n)}] \leq \alpha\cdot\sigma[\tilde{x}^{(n)}].$$

ただしここで，α は $\alpha \geq 0$ である適当な行列である。

α の各要素を無暗に小さくとると，上の方法は有効に働かない。いいかえると，不等式（3・36）は決してみたされず，したがって反復計算は停止されないことになる。上の方法が有効に働くためには，α をどのように選べばよいか。次の定理はこれに答えるものである。

<u>定理3.</u> "定理2において

$$(3\cdot 37) \qquad \delta_0 = (E-L_0)^{-1}\varepsilon$$

とおく。$\hat{\varepsilon}$ が小さくて

(3・38)　$\rho(\delta_0) < 1$

が成り立つときは，行列 $\alpha \geq 0$ を

(3・39)　$\alpha \geq 2\delta_0 (E - \delta_0)^{-1}$

が成り立つように選ぶならば，(3・36) は有限の n に対して必ず成り立つ，いいかえると，(3・36) がみたされるとそこで反復計算を停止するという方法は，実際に有効に働く．"

［証明］　いま，$\tilde{x}^{(n)}$ が振動状態にあるとすると，$\tilde{x}^{(n+1)}$ も振動状態にあるから，定理2の (3・13) により (3・37) を用いて

$$\sigma[\tilde{x}^{(n)} - \hat{x}],\ \sigma[\tilde{x}^{(n+1)} - \hat{x}] \leq \delta_0 \cdot \sigma(\hat{x})$$

を得る．したがって

(3・40)　$\sigma[\tilde{x}^{(n+1)} - \tilde{x}^{(n)}] \leq 2\delta_0 \cdot \sigma(\hat{x})$

を得る．ところで，

$$\begin{aligned}\sigma(\hat{x}) &\leq \sigma[\hat{x} - \tilde{x}^{(n)}] + \sigma[\tilde{x}^{(n)}] \\ &\leq \delta_0 \cdot \sigma(\hat{x}) + \sigma[\tilde{x}^{(n)}]\end{aligned}$$

であるから，

$$(E-\delta_0)\sigma(\hat{x}) \leq \sigma[\tilde{x}^{(n)}]$$

を得，(3・38) により補題3を用いて

$$\sigma(\hat{x}) \leq (E-\delta_0)^{-1}\sigma[\tilde{x}^{(n)}]$$

を得る．かくて (3・40) により，振動状態にある $\tilde{x}^{(n)}$ に対しては

$$(3\cdot 41) \qquad \sigma[\tilde{x}^{(n+1)}-\tilde{x}^{(n)}] \leq 2\delta_0\cdot(E-\delta_0)^{-1}\sigma[\tilde{x}^{(n)}]$$

を得る．

いま，行列 $\alpha \geq 0$ は (3・39) をみたすとして，(3・36) がすべての n に対して成り立たない，としてみよう．すると，

$$(3\cdot 42) \qquad \sigma[\tilde{x}^{(n+1)}-\tilde{x}^{(n)}] > \alpha\cdot\sigma[\tilde{x}^{(n)}] \quad (n=0,1,2,\cdots)$$

となる．ところが，定理2により $\tilde{x}^{(n)}$ はある n に対し必ず振動状態になるから，そのような n に対しては (3・41) が成り立つ．すると，(3・42) により

$$2\delta_0(E-\delta_0)^{-1}\sigma[\tilde{x}^{(n)}] > \alpha\cdot\sigma[\tilde{x}^{(n)}]$$

を得る．これは (3・39) に矛盾する．これで定理は証明された．

反復計算を停止するために不等式(3·36)を用いる場合，最後に得られる $\tilde{x}^{(n+1)}$ はどの程度の誤差をもっているであろうか。次の定理はこれに答えるものである．

<u>定理4.</u> "定理2において，ある n に対し(3·36)が成り立つとする．このとき，行列 $\alpha \geq 0$ の各要素が小さくて

$$(3\cdot 43) \qquad \rho(L_0+\alpha) < 1$$

が成り立つならば，$\tilde{x}^{(n+1)}$ に対して次の不等式が成り立つ：

$$(3\cdot 44) \qquad \sigma[\tilde{x}^{(n+1)}-\hat{x}] \leq \Big\{ L_0[E-(L_0+\alpha)]^{-1}\alpha + (E-\alpha)[E-(L_0+\alpha)]^{-1}\varepsilon \Big\}\sigma(\hat{x}).$$

［証明］ (3·36)より逐次つぎの不等式を得る：

$$\begin{aligned}
\sigma[\tilde{x}^{(n)}-\hat{x}] &\leq \sigma[\tilde{x}^{(n)}-\tilde{x}^{(n+1)}]+\sigma[\tilde{f}(\tilde{x}^{(n)})-f(\hat{x})] \\
&\leq \alpha\cdot\sigma[\tilde{x}^{(n)}]+\sigma[\tilde{f}(\tilde{x}^{(n)})-f(\tilde{x}^{(n)})]+\sigma[f(\tilde{x}^{(n)})-f(\hat{x})] \\
&\leq \alpha\cdot\sigma[\tilde{x}^{(n)}-\hat{x}]+\alpha\cdot\sigma(\hat{x})+\varepsilon\cdot\sigma[f(\tilde{x}^{(n)})]+K_0\sigma[\tilde{x}^{(n)}-\hat{x}] \\
&\leq \alpha\cdot\sigma[\tilde{x}^{(n)}-\hat{x}]+\alpha\cdot\sigma(\hat{x})+\varepsilon\cdot\sigma[f(\tilde{x}^{(n)})-f(\hat{x})] \\
&\qquad +\varepsilon\cdot\sigma[f(\hat{x})]+K_0\sigma[\tilde{x}^{(n)}-\hat{x}] \\
&\leq \alpha\cdot\sigma[\tilde{x}^{(n)}-\hat{x}]+\alpha\cdot\sigma(\hat{x})+\varepsilon K_0\sigma[\tilde{x}^{(n)}-\hat{x}]+\varepsilon\cdot\sigma(\hat{x})
\end{aligned}$$

$$+K_0\sigma[\tilde{x}^{(n)}-\hat{x}]$$

$$=(L_0+\alpha)\sigma[\tilde{x}^{(n)}-\hat{x}]+(\alpha+\varepsilon)\sigma(\hat{x}).$$

これより

$$[E-(L_0+\alpha)]\sigma[\tilde{x}^{(n)}-\hat{x}]\leq(\alpha+\varepsilon)\sigma(\hat{x})$$

となるが，(3.43)により補題3を用いて

$$(3.45)\qquad \sigma[\tilde{x}^{(n)}-\hat{x}]\leq[E-(L_0+\alpha)]^{-1}(\alpha+\varepsilon)\sigma(\hat{x})$$

を得る．すると，

$$\begin{aligned}
\sigma[\tilde{x}^{(n+1)}-\hat{x}]&=\sigma[\tilde{f}(\tilde{x}^{(n)})-f(\hat{x})]\\
&\leq\sigma[\tilde{f}(\tilde{x}^{(n)})-f(\tilde{x}^{(n)})]+\sigma[f(\tilde{x}^{(n)})-f(\hat{x})]\\
&\leq\varepsilon\cdot\sigma[f(\tilde{x}^{(n)})]+K_0\sigma[\tilde{x}^{(n)}-\hat{x}]\\
&\leq\varepsilon\cdot\sigma[f(\tilde{x}^{(n)})-f(\hat{x})]+\varepsilon\cdot\sigma(\hat{x})+K_0\sigma[\tilde{x}^{(n)}-\hat{x}]\\
&\leq\varepsilon K_0\sigma[\tilde{x}^{(n)}-\hat{x}]+\varepsilon\cdot\sigma(\hat{x})+K_0\sigma[\tilde{x}^{(n)}-\hat{x}]\\
&=L_0\sigma[\tilde{x}^{(n)}-\hat{x}]+\varepsilon\cdot\sigma(\hat{x})
\end{aligned}$$

となるから，(3.45)により

$$(3.46)\qquad \sigma[\tilde{x}^{(n+1)}-\hat{x}]\leq\left\{\varepsilon+L_0[E-(L_0+\alpha)]^{-1}(\alpha+\varepsilon)\right\}\sigma(\hat{x})$$

を得る．これより(3.44)は容易にみちびかれる．これで

定理は証明された。

［注意 1］　α の各要素が小さく，L_0 の要素と α の要素との積が ε の対角線要素に比べて無視できるときは，(3.44) は近似的に

$$(3.47)\qquad \sigma[\tilde{x}^{(n+1)}-\hat{x}] \leq (E-L_0)^{-1}\varepsilon\cdot\sigma(\hat{x})$$

と書くことができる。(3.13) と比較すると，これは $\tilde{x}^{(n+1)}$ が振動状態にあるものと同程度の精度をもっていることを示している。

このことから，(3.39) を考慮に入れて，つぎの結論が得られる：

α としては，その要素が ε の対角線要素に比べては大きく，しかも L_0 の要素との積は ε の対角線要素に比べて小さくなるように選ぶことが望ましい。もし α がこのように選ばれるならば，最後に得られる $\tilde{x}^{(n+1)}$ は振動状態にあるものと，ほぼ同程度の精度をもつことになる。

［注意 2］　$\sigma(v)$ をベクトルのノルムとしても，定理 3，4 はそのまま成立する。ただしこのときは，K_0，L_0，ε，δ_0，α はいずれもスカラーになる。

§4. Newton法の誤差解析

4・1 Newton法

与えられた方程式を

(4・1) $\varphi(x)=0$

とすると，Newton法とは，

(4・2) $f(x)=x-J^{-1}(x)\varphi(x)$

として，適当な $x^{(0)}$ から出発して

(4・3) $x^{(n+1)}=f[x^{(n)}]$ $(n=0,1,2,\cdots)$

により $x^{(n)}$ $(n=1,2,\cdots)$ を逐次求め，適当なところでこの反復計算を止めて，最後に得られたものを与えられた方程式の解の近似とする方法である．ただし(4・2)で $J(x)$ は $\varphi(x)$ の x に関するヤコビ行列を表わしている．実際計算では，しかし，適当な行列 $H(x)$ を選び，

(4・4) $f(x)=x-H(x)\varphi(x)$

として(4・3)によって $x^{(n)}$ $(n=1,2,\cdots)$ を計算する方法がよく用いられる．ここでは，このような方法をNewton

法とよび，前者の方法を特に区別する必要があるときにはそれを本来のNewton法とよぶことにする．

$\varphi(x)$，$H(x)$ については次のことを仮定する：

(a)　$\varphi(x)$ は x－空間の凸コンパクトな集合Dにおいて x に関し2回連続微分可能である，

(b)　行列 $H(x)$ はDに属する任意の x'，x'' に対して次の条件をみたしている：

$$(4\cdot5)\quad \sigma\left[H(x')-H(x'')\right]\leq C\cdot\sigma(x'-x'').$$

ただしここで C はその成分がすべて非負である3次のテンソルである．

$H(x)$ の要素を $h_{ij}(x)$ $(i,j=1,2,\cdots,d)$，C の成分を c_{ijk} $(i,j,k=1,2,\cdots,d)$ とすると，(4・5) は次のことを表わしている：

$$|h_{ij}(x')-h_{ij}(x'')|\leq\sum_{k=1}^{d}c_{ijk}\,|x_k'-x_k''|.$$

ただしここで x_k'，x_k'' はそれぞれベクトル x'，x'' の成分である．

Dに属する任意の x'，x'' に対しては，われわれの仮定

から次の等式を得る：

$$(4\cdot 6)\quad \begin{cases} \varphi(x'')-\varphi(x')=\displaystyle\int_0^1 J[x'+\theta(x''-x')](x''-x')\,d\theta, \\ J[x'+\theta(x''-x')]-J(x')=\theta\displaystyle\int_0^1 J_x[x'+\theta_1\theta(x''-x')](x''-x')\,d\theta_1 \\ \qquad\qquad (0\leq\theta\leq 1). \end{cases}$$

ただしここで，$J_x(x)$ は成分が $\partial^2\varphi_i/\partial x_j\partial x_k$ $(i, j, k=1, 2, \cdots, d)$ であるテンソルで，φ_i，x_j はそれぞれベクトル φ，x の成分である．(4・6) から次の式が得られる：

$$(4\cdot 7)\quad \varphi(x'')-\varphi(x')=J(x')(x''-x')+\Phi(x', x'').$$

ただし

$$(4\cdot 8)\quad \Phi(x', x'')=\int_0^1\int_0^1 J_x[x'+\theta_1\theta(x''-x')]\theta\,d\theta_1\,d\theta\cdot(x''-x')(x''-x').$$

である．(4・8) から

$$(4\cdot 9)\quad \sigma[\Phi(x', x'')]\leq M_0\,\sigma(x''-x')\,\sigma(x''-x')$$

をみたす定数テンソル M_0 の存在することがわかる．ただし M_0 の成分はすべて非負である．

(4・7) を用いると，(4・4) で定義される $f(x)$ に対しては，

$$
\begin{aligned}
f(x')-f(x'') &= x'-x''-[H(x')\varphi(x')-H(x'')\varphi(x'')] \\
&= (x'-x'')-H(x')[\varphi(x')-\varphi(x'')]-[H(x')-H(x'')]\varphi(x'') \\
&= (x'-x'')-H(x')J(x')(x'-x'')+H(x')\Phi(x',x'') \\
&\qquad -[H(x')-H(x'')]\varphi(x'')
\end{aligned}
$$

を得るので，次の不等式を得る：

$$
\begin{aligned}
(4\cdot 10)\quad \sigma[f(x')-f(x'')] &\leq \sigma[E-H(x')J(x')]\,\sigma(x'-x'') \\
&\quad +\sigma[H(x')]M_0\,\sigma(x'-x'')\,\sigma(x'-x'') \\
&\quad + C\,\sigma(x'-x'')\,\sigma[\varphi(x'')] \\
&= \Big\{\sigma[E-H(x')J(x')]+\sigma[H(x')]M_0\,\sigma(x'-x'') \\
&\qquad + C\,\sigma[\varphi(x'')]\Big\}\,\sigma(x'-x'').
\end{aligned}
$$

われわれは，Dに属する任意のxに対しては，

$$
(4\cdot 11)\quad \begin{cases} \sigma[E-H(x)J(x)] \leq \kappa, \\ \sigma[H(x)]M_0 \leq M \end{cases}
$$

が成り立つとする．ただし，κは$\kappa \geq 0$である行列であり，Mは各要素が非負であるテンソルである．集合Dはコンパクトであるから，(4・11)はつねに成り立つ．(4・11)により(4・10)から次の不等式が得られる：

$$(4\cdot12)\qquad \sigma[f(x')-f(x'')]\leq K(x',x'')\,\sigma(x'-x'').$$

ただし

$$(4\cdot13)\qquad K(x',x'')=\kappa+M\sigma(x'-x'')+C\sigma[\varphi(x'')]\geq 0$$

である．

(4・13)で与えられる行列 $K(x',x'')$ に対して，われわれは次のことを仮定する：

(c)　Dに属する任意の x', x'' に対して

$$(4\cdot14)\qquad K(x',x'')\leq K_0$$

が成り立ち，

$$(4\cdot15)\qquad \rho(K_0)<1$$

であるような行列 $K_0\geq 0$ が存在する．

上の仮定は，Dが与えられた方程式(4・1)の解を含む小さな領域であるとき明らかにみたされている．なぜなら，このような場合には，Dに属する任意の x', x'' に対して $\sigma(x'-x'')\approx 0$, $\sigma[\varphi(x'')]\approx 0$ であり，また

Newton法の場合には(4・11)の第一式をみたす$\varkappa$の要素はすべて小さい，と考えてよいからである．

(c)がみたされていると，(4・12)から

$$\sigma[f(x')-f(x'')]\leq K_0\,\sigma(x'-x'')$$

が，Dに属する任意のx'，x''に対して成り立つことになる．したがって定理1，あるいは定理2に示された条件をみたす$x^{(0)}$をとれば，それぞれ定理1，あるいは定理2の結果が成り立つことになる(ただし定理2の場合には$\hat{\varepsilon}$は小さくて(3・7)が成り立つとして)．

[注意1]　(4・11)により

(4・16)　　$E-H(x)J(x)=\nu(x)$

とおくと，

$$\sigma[\nu(x)]\leq\varkappa$$

を得る．ところが，(4・13)，(4・14)により

$$\varkappa\leq K_0$$

であり，しかも(4・15)により$\rho(K_0)<1$である．すると，

$$\{\sigma[\nu(x)]\}^n \leq \kappa^n \leq K_0^n$$

で，補題2により $\lim_{n\to\infty} K_0^n = 0$ であるから，

$$\lim_{n\to\infty} \{\sigma[\nu(x)]\}^n = 0$$

を得る．すると

$$\sigma[\nu^n(x)] \leq \{\sigma[\nu(x)]\}^n$$

であるから，

$$\lim_{n\to\infty} \sigma[\nu^n(x)] = 0$$

を得，したがって

$$\lim_{n\to\infty} \nu^n(x) = 0$$

を得る．すると，補題2により

$$\rho[\nu(x)] < 1$$

を得る．すると，補題3により $[E-\nu(x)]^{-1}$ が存在し，明らかに

$$\det[E-\nu(x)] \neq 0$$

となる．これは（4・16）から

$$\det[H(x)J(x)] \neq 0,$$

すなわち，

$$\det H(x),\ \det J(x) \neq 0$$

を意味する．すると，$H^{-1}(x)$, $J^{-1}(x)$ が存在するので，与えられた方程式（4・1）と方程式

$$(4\cdot 17) \qquad x = f(x) = x - H(x)\varphi(x)$$

とは明らかに同値になる．

［注意２］　$\sigma(\cdot)$ をベクトル，あるいは行列のノルムとしても上の議論は明らかに成り立つ．ただしこのときは，C, M_0, $\varkappa$, M, K_0 はすべてスカラーになる．

4・2　Newton法の誤差解析

（4・4）で定義された $f(x)$ に対して，それの計算値を $\tilde{f}(x)$ とするとき（3・1）が成り立つとし，ε の対角線要素はすべて小さくて（3・7）が成立するとする．$x^{(0)}$ は定理２の条件（3・8）をみたす D に属する点とする．すると，定理２により，（3・5）により得られる列 $\{\tilde{x}^{(n)}\}$ はある

所から先きは振動状態になる．いまの場合，振動状態にある任意の $\tilde{x}^{(n)}$ に対しては次の定理を得る．

<u>定理 5.</u> "与えられた方程式(4・1)のDの中にある解を $\hat{x}$ とするとき，Newton法によって得られる近似解のうち振動状態にある任意のものを $\tilde{x}^{(n)}$ とすれば，

$$(4\cdot 18)\qquad \sigma[\tilde{x}^{(n)}-\hat{x}]\leq \hat{\delta}\cdot\sigma[\hat{x}]$$

が成立する．ただしここで $\hat{\delta}$ は

$$(4\cdot 19)\qquad \hat{\delta}=[E-(E+\varepsilon)\{\kappa+M\hat{\delta}\cdot\sigma(\hat{x})\}]^{-1}\varepsilon$$

によって定まる行列である．"

［証明］ 定理2により，振動状態にある $\tilde{x}^{(n)}$ に対しては

$$(4\cdot 20)\qquad \sigma[\tilde{x}^{(n)}-\hat{x}]\leq \delta_0\cdot\sigma(\hat{x})$$

が成り立つ．ただしここで

$$(4\cdot 21)\qquad \delta_0=(E-L_0)^{-1}\varepsilon$$

である．

さて，定理1により $\hat{x}\in S$ であり，しかも定理2の証明で

示したように，(3·17) が成り立つから，

(4·22) $\sigma[\hat{x}-\tilde{x}^{(1)}]\leq\rho_1$

が成立する．すると，

$$\sigma(\hat{x})\leq\rho_1+\sigma[\tilde{x}^{(1)}]$$

となるから，(4·21) により

$$\delta_0\cdot\sigma(\hat{x})\leq(E-L_0)^{-1}\varepsilon\{\rho_1+\sigma[\tilde{x}^{(1)}]\}$$

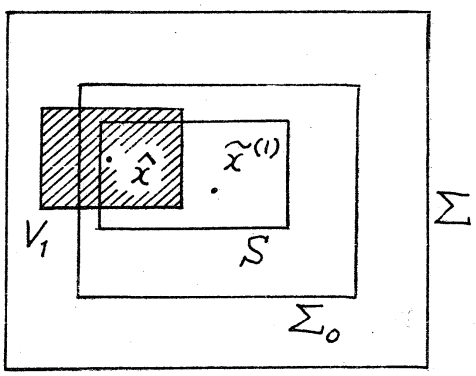
が成り立ち，(3·11) により

(4·23) $\delta_0\cdot\sigma(\hat{x})\leq\rho_2$

を得る．すると，(4·22), (4·23), (3·8) により

(4·24) $V_1=\{h\mid\sigma(h-\hat{x})\leq\delta_0\cdot\sigma(\hat{x})\}\subset\Sigma\subset D$

を得る．



すると，(4·13)，(4·14) により，

$$(4\cdot 25)\quad K_1=\max_{x\in V_1} K(x,\hat{x})=\kappa+M\delta_0\sigma(\hat{x})\leq K_0$$

が成り立つ．さて，$\tilde{x}^{(n)}$が振動状態にあれば，(4·20)，(4·24) により $\tilde{x}^{(n)}\in V_1$ となるから，(3·1)，(4·12) により逐次つぎの不等式を得る：

$$\begin{aligned}(4\cdot 26)\quad \sigma[\tilde{x}^{(n+1)}-\hat{x}]&=\sigma[\tilde{f}(\tilde{x}^{(n)})-f(\hat{x})]\\
&\leq\sigma[\tilde{f}(\tilde{x}^{(n)})-f(\tilde{x}^{(n)})]+\sigma[f(\tilde{x}^{(n)})-f(\hat{x})]\\
&\leq\varepsilon\cdot\sigma[f(\tilde{x}^{(n)})]+K(\tilde{x}^{(n)},\hat{x})\sigma[\tilde{x}^{(n)}-\hat{x}]\\
&\leq\varepsilon\cdot\sigma[f(\tilde{x}^{(n)})-f(\hat{x})]+\varepsilon\cdot\sigma(\hat{x})+K(\tilde{x}^{(n)},\hat{x})\sigma[\tilde{x}^{(n)}-\hat{x}]\\
&\leq(E+\varepsilon)K(\tilde{x}^{(n)},\hat{x})\sigma[\tilde{x}^{(n)}-\hat{x}]+\varepsilon\cdot\sigma(\hat{x})\\
&\leq(E+\varepsilon)K_1\sigma[\tilde{x}^{(n)}-\hat{x}]+\varepsilon\cdot\sigma(\hat{x})\\
&=L_1\sigma[\tilde{x}^{(n)}-\hat{x}]+\varepsilon\cdot\sigma(\hat{x}).\end{aligned}$$

ただし

$$(4\cdot 27)\qquad L_1=(E+\varepsilon)K_1$$

である．(4·25) により明らかに

$$(4\cdot 28)\qquad 0\leq L_1\leq L_0$$

であるから，(3·7) より補題2によって

$$(4\cdot29)\qquad \rho(L_1)<1$$

である．さて，$\tilde{x}^{(n)}$ が振動状態にあるときは，$\tilde{x}^{(n+1)}$ も振動状態にあるから，(4・26) と同様にして

$$\sigma[\tilde{x}^{(n+2)}-\hat{x}]\leq L_1\sigma[\tilde{x}^{(n+1)}-\hat{x}]+\varepsilon\cdot\sigma(\hat{x})$$

を得る．すると，(4・26) を用いて

$$\sigma[\tilde{x}^{(n+2)}-\hat{x}]\leq L_1^2\sigma[\tilde{x}^{(n)}-\hat{x}]+(L_1+E)\varepsilon\cdot\sigma(\hat{x})$$

を得る．この手続きを続けてゆくと，任意の正の整数 r に対して

$$(4\cdot30)\quad \sigma[\tilde{x}^{(n+r)}-\hat{x}]\leq L_1^r\sigma[\tilde{x}^{(n)}-\hat{x}]+(E+L_1+\cdots+L_1^{r-1})\varepsilon\cdot\sigma(\hat{x})$$

を得る．ところが，$\tilde{x}^{(n)}$ は振動状態にあるから，ある正の整数 p に対して $\tilde{x}^{(n+p)}=\tilde{x}^{(n)}$ となる．そこで (4・30) において $r=p$ とすれば，

$$(4\cdot31)\quad (E-L_1^p)\sigma[\tilde{x}^{(n)}-\hat{x}]\leq(E+L_1+\cdots+L_1^{p-1})\varepsilon\cdot\sigma(\hat{x})$$

を得る．さて，

$$(E+L_1+\cdots+L_1^{p-1})(E-L_1)=E-L_1^p$$

で，(4・29) より補題3によって

$$(E-L_1)^{-1},\ (E-L_1^p)^{-1} \geq 0$$

が存在するから，(4.31) より

$$(4.32)\qquad \sigma[\tilde{x}^{(n)}-\hat{x}] \leq (E-L_1)^{-1}\varepsilon\cdot\sigma(\hat{x})$$

を得る．

$$(4.33)\qquad \delta_1 = (E-L_1)^{-1}\varepsilon$$

とおけば，(4.28) より補題3によって

$$(4.34)\qquad \delta_1 \leq \delta_0$$

が成立し，(4.32) より

$$(4.35)\qquad \sigma[\tilde{x}^{(n)}-\hat{x}] \leq \delta_1\cdot\sigma(\hat{x})$$

を得る．

上の手続きを繰り返してゆくと，

$$(4.36)\qquad \begin{cases} K_r = \kappa + M\delta_{r-1}\cdot\sigma(\hat{x}), & L_r = (E+\varepsilon)K_r, \\ \delta_r = (E-L_r)^{-1}\varepsilon \end{cases}$$

$$(r=1,2,\cdots)$$

で定まる行列の列 $\{K_r\}$, $\{\delta_r\}$ を得，これらは

$$(4\cdot37)\qquad \begin{cases} K_0 \geq K_1 \geq K_2 \geq \cdots \quad \geq 0, \\ \delta_0 \geq \delta_1 \geq \delta_2 \geq \cdots \quad \geq 0 \end{cases}$$

をみたし，しかもすべての r に対して

$$(4\cdot38)\qquad \sigma[\tilde{x}^{(n)} - \hat{x}] \leq \delta_r \cdot \sigma(\hat{x})$$

が成り立つようになる．(4・37)により

$$(4\cdot39)\qquad \hat{K} = \lim_{r\to\infty} K_r \geq 0, \quad \hat{\delta} = \lim_{r\to\infty} \delta_r \geq 0$$

が存在し，(4・38)により

$$(4\cdot40)\qquad \sigma[\tilde{x}^{(n)} - \hat{x}] \leq \hat{\delta} \cdot \sigma(\hat{x})$$

を得る．(4・36)より明らかに

$$(4\cdot41)\qquad \hat{K} = \kappa + M\hat{\delta}\cdot\sigma(\hat{x}), \quad \hat{\delta} = [E - (E+\varepsilon)\hat{K}]^{-1}\varepsilon$$

が成り立つ．(4・41)の第1式を第2式に代入すれば，(4・19)の式が得られる．これで定理は証明された．

［注意1］ $\hat{\varepsilon}$ が小さいときは，(4・19)からわかるように，行列 $\hat{\delta}$ の各要素の絶対値はすべて小さくなり，したがって，

(4·42) $\hat{\delta} \approx (E-\varkappa)^{-1}\varepsilon$

となり，本来のNewton法の場合にはとくに

(4·43) $\hat{\delta} \approx \varepsilon$

となる．(4·43)の場合には，(4·18)は

(4·44) $|\tilde{x}_i^{(n)} - \hat{x}_i| \leq \varepsilon_i |\hat{x}_i| \quad (i=1, 2, \cdots, d)$

を意味することになる．これは得られた近似解$\tilde{x}^{(n)}$の各成分$\tilde{x}_i^{(n)}$の相対誤差がそれぞれε_iを越えないことを表わしていて，これは実際問題ではきわめて重要なことである，と思われる．なお，本来のNewton法でなくても，行列$\varkappa$の各要素が十分小さいときは，(4·42)は(4·43)のようになるので，上と同じ結論が得られる．

［注意2］ $\sigma(v)$をベクトルのノルムとしても，定理5はそのまま成り立つ．ただしこのときは，ε, $\varkappa$, M, $\hat{\delta}$, はすべてスカラーになる．

［注意3］ $\hat{\varepsilon}$が十分小さいときは，$\hat{\delta}$は(4·19)によってつねに一意的に定まる．

これを示すために，(4·19)を

$$(4\cdot45)\qquad \delta=[E-(E+\varepsilon)(\kappa+N\delta)]^{-1}\varepsilon$$

と書いて，この方程式が $\varepsilon\to0$ のとき $\delta\cdot\varepsilon^{-1}\to(E-\kappa)^{-1}$ となるような解を一つただ一つもつことを示そう．

(4·13)，(4·14) により $\kappa\leq K_0$ であるから，(4·15) により補題2を用いて

$$(4\cdot46)\qquad \rho(\kappa)<1$$

を得る．κ の Jordan 標準形を κ' とすると，

$$(4\cdot47)\qquad S^{-1}\kappa S=\kappa'=\sum\oplus\begin{bmatrix}\lambda & & 0\\ \zeta & \lambda & \\ & \ddots & \ddots\\ 0 & & \zeta & \lambda\end{bmatrix}$$

の形の式が成り立ち，λ は κ の固有値で ζ は任意の正数である．ここで ζ を十分小さくとると，$|\lambda|<1$ であるから，

$$(4\cdot48)\qquad \|\kappa'\|<1$$

を得る．ただしここで，$\|\cdot\|$ は，任意の行列 $A=(a_{ij})$ $(i,j=1,2,\cdots,d)$ に対して

$$(4\cdot49)\qquad \|A\|=\max_i\sum_{j=1}^{d}|a_{ij}|$$

で定義される行列のノルムを表わしている．(4·47) に対応

して

(4·50) $S^{-1}\delta S=\delta'$, $S^{-1}NS=N'$, $S^{-1}\varepsilon S=\varepsilon'$

とおくと，方程式(4·45)は次のようになる：

(4·51) $\delta'=[E-(E+\varepsilon')(\kappa'+N'\delta')]^{-1}\varepsilon'$.

さて，$\varepsilon\to 0$ のとき $\delta\cdot\varepsilon^{-1}\to(E-\kappa)^{-1}$ となる δ に対しては，(4·47)，(4·50) より，$\varepsilon'\to 0$ のとき $\delta'\cdot\varepsilon'^{-1}\to(E-\kappa')^{-1}$ となる．そこで

$$\delta'\cdot\varepsilon'^{-1}=(E-\kappa')^{-1}+R'$$

とおけば，$\varepsilon'\to 0$ のとき $R'\to 0$ で，上の式より

(4·52) $\|\delta'\|\leq[\|(E-\kappa')^{-1}\|+\|R'\|]\|\varepsilon'\|$

を得る．ところが，よく知られているように，(4·48) により

$$(E-\kappa')^{-1}=\sum_{n=0}^{\infty}\kappa'^{n}$$

となるから，

(4·53) $\|(E-\kappa')^{-1}\|\leq(1-\|\kappa'\|)^{-1}$

を得，(4.52) より

$$(4.54)\qquad \|\delta'\| \leq [(1-\|\kappa'\|)^{-1} + \|R'\|]\|\varepsilon'\|$$

を得る．$\|\varepsilon'\| \to 0$ のとき $\|R'\| \to 0$ であるから，

$$(4.55)\qquad l_1 > (1-\|\kappa'\|)^{-1}$$

をみたす正数 l_1 をとれば，$\|\varepsilon'\|$ が十分小さいとき，(4.54) をみたす δ' は次の不等式をみたすようになる：

$$(4.56)\qquad \|\delta'\| \leq l_1 \|\varepsilon'\|.$$

以上のことから，われわれは，$\|\varepsilon'\|$ が十分小さいとき，(4.56)，(4.51) をみたす δ' が一つただ一つ存在することを示せばよいことになる．

(4.51) を簡単に

$$(4.57)\qquad \delta' = F(\delta')$$

と書こう．明らかに

$$(4.58)\qquad F(\delta') = [E-(E+\varepsilon')(\kappa'+N'\delta')]^{-1}\varepsilon'$$

である．(4.56) をみたす δ' の集合を Ω とすれば，Ω に属する任意の δ' に対しては

(4·59) $\|(E+\varepsilon')(\kappa'+N'\delta')\| \leq (1+\|\varepsilon'\|)(\|\kappa'\|+l_1\|N'\|\cdot\|\varepsilon'\|)$

となる．したがって，(4.48)により

(4·60) $\|\kappa'\| < l_2 < 1$

をみたす正数 l_2 をとるとき，十分小さな正数 ζ_1 をとれば $\|\varepsilon'\| < \zeta_1$ をみたす ε' に対してつねに

(4·61) $(1+\|\varepsilon'\|)(\|\kappa'\|+l_1\|N'\|\cdot\|\varepsilon'\|) \leq l_2 < 1$

が成立する．すると，(4·59)，(4·61)により，(4·53)と同様にして

$$\|[E-(E+\varepsilon')(\kappa'+N'\delta')]^{-1}\| \leq [1-(1+\|\varepsilon'\|)(\|\kappa'\|+l_1\|N'\|\cdot\|\varepsilon'\|)]^{-1}$$

が成立する．上の式の右辺は，$\|\varepsilon'\| \to 0$ のとき $(1-\|\kappa'\|)^{-1}$ に収束するから，(4·55)により，十分小さな正数 $\zeta_2 \leq \zeta_1$ をとれば $\|\varepsilon'\| < \zeta_2$ のときつねに

(4·62) $\|[E-(E+\varepsilon')(\kappa'+N'\delta')]^{-1}\| \leq l_1$

が成立する．このことは，(4·58)により，$\|\varepsilon'\| < \zeta_2$ であれば

$$(4\cdot 63)\qquad F(\Omega)\subset\Omega$$

が成立することを意味している．

次に，Ωに属する任意のδ_1'，δ_2'をとると，(4·58)から次の等式が得られる：

$$\begin{aligned}
&F(\delta_1')-F(\delta_2')\\
&=[E-(E+\varepsilon')(\kappa'+N'\delta_1')]^{-1}\{[E-(E+\varepsilon')(\kappa'+N'\delta_2')]\\
&\qquad-[E-(E+\varepsilon')(\kappa'+N'\delta_1')]\}[E-(E+\varepsilon')(\kappa'+N'\delta_2')]^{-1}\varepsilon'\\
&=[E-(E+\varepsilon')(\kappa'+N'\delta_1')]^{-1}(E+\varepsilon')N'(\delta_1'-\delta_2')\times\\
&\qquad\times[E-(E+\varepsilon')(\kappa'+N'\delta_2')]^{-1}\varepsilon'.
\end{aligned}$$

すると，$\|\varepsilon'\|<\zeta_2$ならば(4·62)により次の不等式が成り立つことになる：

$$(4\cdot 64)\quad \|F(\delta_1')-F(\delta_2')\|\le l_1^2(1+\|\varepsilon'\|)\|N'\|\cdot\|\varepsilon'\|\cdot\|\delta_1'-\delta_2'\|.$$

さて，1より小さい任意の正数kをとるとき，十分小さい正数$\zeta_3\le\zeta_2$をとれば，$\|\varepsilon'\|<\zeta_3$のとき

$$l_1^2(1+\|\varepsilon'\|)\|N'\|\cdot\|\varepsilon'\|\le k$$

がつねに成り立つようにすることができる．すると，(4·64)により，$\|\varepsilon'\|<\zeta_3$のとき，Ωに属する任意のδ_1'，δ_2'に

対して

$$(4\cdot 65)\quad \|F(\delta_1')-F(\delta_2')\| \leq k\|\delta_1'-\delta_2'\|$$

が成立することになる.

(4·63), (4·65)から, よく知られている縮少写像の原理により, $\|\varepsilon'\|<\zeta_3$ ならば方程式(4·57)はΩにおいて一つただ一つの解をもつことになる. これは(4·56), (4·51)をみたすδ'が, $\|\varepsilon'\|$が十分小さければ, 一つただ一つ存在することを意味している. これによって, われわれの望む結果は証明されたことになる.

4·3 Newton法における反復計算の停止

Newton法の場合に, 反復計算(3·5)を不等式(3·36)を用いて停止することを考える.

Newton法の場合, 振動状態にある$\tilde{x}^{(n)}$に対しては, 定理5により(4·18)が成り立つので, 定理3と同様にして次の定理が得られる.

<u>定理6.</u> "定理5において, $\hat{\varepsilon}$は小さく,

$$L_0=(E+\varepsilon)K_0,\quad \delta_0=(E-L_0)^{-1}\varepsilon$$

とおくとき，

(4・66)　　$\rho(\delta_0) < 1$

が成り立つとする．このとき，行列 α を

(4・67)　　$\alpha \geq 2\hat{\delta}(E-\hat{\delta})^{-1}$

が成り立つように選ぶならば，不等式(3・36)を用いて，たしかに反復計算を停止することができる．"

定理5の証明に示されているように，

$$\delta_0 \geq \delta_1 \geq \delta_2 \geq \cdots \geq 0$$

で $\lim_{r\to\infty} \delta_r = \hat{\delta}$ であるから，(4・66)により

$$\rho(\hat{\delta}) < 1$$

が成り立つ．したがって定理3と全く同様にして定理6は証明される．

Newton法において，反復計算を不等式(3・36)を用いて停止した場合，最後に得られる $\tilde{x}^{(n+1)}$ はどの程度の誤差をもっているであろうか．次の定理はこれに答えるものである．

<u>定理7.</u> "定理5において，あるnに対して(3·36)が成り立つとする．ただしこのとき，行列$\alpha \geq 0$の各要素は小さくて，

(4·68) $$\rho(L_0+\alpha)<1$$

で，

(4·69) $$\rho_3=[E-(L_0+\alpha)]^{-1}(\alpha+\varepsilon)\{\rho_1+\sigma[\tilde{x}^{(1)}]\}$$

とするとき

(4·70) $$\Sigma_\alpha=\{h \mid \sigma[h-\tilde{x}^{(1)}]\leq\rho_1+\rho_3\}\subset D$$

が成り立つとする．

すると，$\tilde{x}^{(n+1)}$に対して次の不等式が成り立つ：

(4·71) $$\sigma[\tilde{x}^{(n+1)}-\hat{x}]\leq(\hat{\eta}-\alpha\hat{\eta}-\alpha)\sigma(\hat{x}).$$

ただしここで$\hat{\eta}$は

(4·72) $$\hat{\eta}=[E-\{\alpha+(E+\varepsilon)(\kappa+M\hat{\eta}\sigma(\hat{x}))\}]^{-1}(\alpha+\varepsilon)$$

によって定まる行列である．"

［証明］ (3·36)がみたされているから，(3·1)，(4·

12）により逐次つぎの不等式を得る：

$$(4\cdot 73)\quad \sigma[\tilde{x}^{(n)}-\hat{x}]\leq\sigma[\tilde{x}^{(n)}-\tilde{x}^{(n+1)}]+\sigma[\tilde{f}(\tilde{x}^{(n)})-f(\tilde{x}^{(n)})]$$
$$+\sigma[f(\tilde{x}^{(n)})-f(\hat{x})]$$
$$\leq\alpha\cdot\sigma[\tilde{x}^{(n)}]+\varepsilon\cdot\sigma[f(\tilde{x}^{(n)})]+K(\tilde{x}^{(n)},\hat{x})\sigma[\tilde{x}^{(n)}-\hat{x}]$$
$$\leq\alpha\cdot\sigma[\tilde{x}^{(n)}-\hat{x}]+\alpha\cdot\sigma(\hat{x})$$
$$+\varepsilon\cdot\sigma[f(\tilde{x}^{(n)})-f(\hat{x})]+\varepsilon\cdot\sigma(\hat{x})+K(\tilde{x}^{(n)},\hat{x})\sigma[\tilde{x}^{(n)}-\hat{x}]$$
$$\leq[\alpha+(E+\varepsilon)K(\tilde{x}^{(n)},\hat{x})]\sigma[\tilde{x}^{(n)}-\hat{x}]+(\alpha+\varepsilon)\sigma(\hat{x}).$$

(4·14)により，すると，

$$\sigma[\tilde{x}^{(n)}-\hat{x}]\leq[\alpha+(E+\varepsilon)K_0]\sigma[\tilde{x}^{(n)}-\hat{x}]+(\alpha+\varepsilon)\sigma(\hat{x}),$$

すなわち，

$$(4\cdot 74)\quad [E-(L_0+\alpha)]\sigma[\tilde{x}^{(n)}-\hat{x}]\leq(\alpha+\varepsilon)\sigma(\hat{x})$$

を得る．(4·68)により $[E-(L_0+\alpha)]^{-1}\geq 0$ が存在するから（補題3参照），(4·74)より

$$(4\cdot 75)\qquad \sigma[\tilde{x}^{(n)}-\hat{x}]\leq\eta_0\sigma(\hat{x})$$

が得られる．ただしここで

$$(4\cdot 76)\qquad \eta_0=[E-(L_0+\alpha)]^{-1}(\alpha+\varepsilon)$$

である．

さて，$\hat{x}$ に対しては (4·22) が成り立つから，(4·76) より

(4·77) $\eta_0 \cdot \sigma(\hat{x}) \leq [E-(L_0+\alpha)]^{-1}(\alpha+\varepsilon)\left\{\rho_1 + \sigma[\tilde{x}^{(1)}]\right\} = \rho_3$

が成り立つ．なお，(4·69) を (3·11) と比べると

$$\rho_2 \leq \rho_3$$

であるから，(3·17)，(3·8)，(4·70) より

(4·78) $S \subset \Sigma_0 \subset \Sigma \subset \Sigma_\alpha \subset D$

が成り立つことになる．

さて，(4·22) により

$$\sigma[\hat{x} - \tilde{x}^{(1)}] \leq \rho_1$$

であるから，(4·77)，(4·70) により

(4·79) $W_1 = \left\{ h \mid \sigma(h - \hat{x}) \leq \eta_0 \cdot \sigma(\hat{x}) \right\} \subset \Sigma_\alpha \subset D$

が成り立つ．すると，(4·13)，(4·14) により次の式が得られる：

(4·80) $R_1 = \max_{x \in W_1} K(x, \hat{x}) = \varkappa + M\eta_0 \sigma(\hat{x}) \leq K_0.$

すると，(4·75) により $\tilde{x}^{(n)} \in W_1$ であるから，(4·73) より

(4·81) $\sigma[\tilde{x}^{(n)} - \hat{x}] \leq [\alpha + (E+\varepsilon)R_1]\sigma[\tilde{x}^{(n)} - \hat{x}] + (\alpha+\varepsilon)\sigma(\hat{x})$

を得る．このとき

(4·82) $\alpha + (E+\varepsilon)R_1 \leq \alpha + (E+\varepsilon)K_0 = \alpha + L_0$

であるから，(4·68) により $[E - \{\alpha + (E+\varepsilon)R_1\}]^{-1} \geq 0$ が存在する（補題3参照）．すると，(4·81) から次の不等式が得られる：

(4·83) $\sigma[\tilde{x}^{(n)} - \hat{x}] \leq \eta_1 \cdot \sigma(\hat{x})$.

ただし

(4·84) $\eta_1 = [E - \{\alpha + (E+\varepsilon)R_1\}]^{-1}(\alpha+\varepsilon)$

である．(4·82), (4·76) により

(4·85) $\eta_1 \leq \eta_0$

であることは明らかである．

上の手続きを繰り返してゆくと，

$$(4\cdot 86)\qquad \begin{cases} R_r = \varkappa + M\eta_{r-1}\,\sigma(\hat{x}), \\ \eta_r = [E - \{\alpha + (E+\varepsilon)R_r\}]^{-1}(\alpha+\varepsilon) \end{cases}$$
$$(r = 1, 2, \cdots)$$

によって定まる行列の列 $\{R_r\}$, $\{\eta_r\}$ を得る．これらは明らかに

$$(4\cdot 87)\qquad \begin{cases} K_0 \geq R_1 \geq R_2 \geq \cdots \geq 0, \\ \eta_0 \geq \eta_1 \geq \eta_2 \geq \cdots \geq 0 \end{cases}$$

をみたし，しかもすべての r に対して

$$(4\cdot 88)\qquad \sigma[\tilde{x}^{(n)} - \hat{x}] \leq \eta_r\,\sigma(\hat{x})$$

が成り立っている．

(4・87) により

$$(4\cdot 89)\qquad \hat{R} = \lim_{r\to\infty} R_r \geq 0, \quad \hat{\eta} = \lim_{r\to\infty} \eta_r \geq 0$$

が存在し，(4・88) により

$$(4\cdot 90)\qquad \sigma[\tilde{x}^{(n)} - \hat{x}] \leq \hat{\eta}\cdot\sigma(\hat{x})$$

が成り立つようになる．(4・86) より明らかに

$$(4\cdot 91)\qquad \begin{cases} \hat{R} = \kappa + M\hat{\eta}\,\sigma(\hat{x}), \\ \hat{\eta} = [E - \{\alpha + (E+\varepsilon)\hat{R}\}]^{-1}(\alpha+\varepsilon) \end{cases}$$

が成り立つ．(4・91) の第1式を第2式に代入すると，(4・72) が得られる．

(4・90) が成り立つと，(4・12)，(4・13) により次の不等式が得られる：

$$\begin{aligned}(4\cdot 92)\qquad \sigma[\tilde{x}^{(n+1)} - \hat{x}] &= \sigma[\tilde{f}(\tilde{x}^{(n)}) - f(\hat{x})] \\ &\le \sigma[\tilde{f}(\tilde{x}^{(n)}) - f(\tilde{x}^{(n)})] + \sigma[f(\tilde{x}^{(n)}) - f(\hat{x})] \\ &\le \varepsilon\cdot\sigma[f(\tilde{x}^{(n)})] + K(\tilde{x}^{(n)}, \hat{x})\,\sigma[\tilde{x}^{(n)} - \hat{x}] \\ &\le \varepsilon\cdot\sigma[f(\tilde{x}^{(n)}) - f(\hat{x})] + \varepsilon\cdot\sigma(\hat{x}) \\ &\qquad + K(\tilde{x}^{(n)}, \hat{x})\,\sigma[\tilde{x}^{(n)} - \hat{x}] \\ &\le (E+\varepsilon)[\kappa + M\hat{\eta}\,\sigma(\hat{x})]\,\hat{\eta}\,\sigma(\hat{x}) + \varepsilon\cdot\sigma(\hat{x}) \\ &= [\varepsilon + (E+\varepsilon)\{\kappa + M\hat{\eta}\,\sigma(\hat{x})\}\hat{\eta}\,]\,\sigma(\hat{x}).\end{aligned}$$

ところが，(4・72) より

$$[E - \{\alpha + (E+\varepsilon)(\kappa + M\hat{\eta}\,\sigma(\hat{x}))\}]\,\hat{\eta} = \alpha + \varepsilon,$$

すなわち，

$$\hat{\eta} - \alpha\hat{\eta} - (E+\varepsilon)\{\kappa + M\hat{\eta}\,\sigma(\hat{x})\}\hat{\eta} = \alpha + \varepsilon$$

が成り立つ．したがって，(4・92) より次の不等式が得られる：

$$\sigma[\tilde{x}^{(n+1)}-\hat{x}] \leq (\hat{\eta}-\alpha\hat{\eta}-\alpha)\sigma(\hat{x}).$$

これは望む (4・71) である．これで定理は証明された．

［注意 1］ (4・72) は補題 3 により次のように書くことができる：

$$(4\cdot 93)\quad \hat{\eta}=(\alpha+\varepsilon)+\{\alpha+(E+\varepsilon)[\kappa+M\hat{\eta}\sigma(\hat{x})]\}(\alpha+\varepsilon)$$
$$+\{\alpha+(E+\varepsilon)[\kappa+M\hat{\eta}\sigma(\hat{x})]\}^2(\alpha+\varepsilon)+\cdots.$$

さて，(4・72) より，κ, α, $\varepsilon \to 0$ のとき

$$\sigma(\hat{\eta})=O[\sigma(\alpha)+\sigma(\varepsilon)]$$

であるから，(4・93) より

$$(4\cdot 94)\quad \hat{\eta}=\alpha+\varepsilon+[\kappa,\alpha,\varepsilon]_2$$

を得る．ただしここで $[\kappa,\alpha,\varepsilon]_2$ は κ, α, ε の要素に関する 2 次以上の項の和を表わしている．(4・94) を (4・93) の右辺に代入すると次の式が得られる：

$$(4\cdot 95)\quad \hat{\eta}=\alpha+\varepsilon+\{\alpha+\kappa+M(\alpha+\varepsilon)\sigma(\hat{x})\}(\alpha+\varepsilon)$$
$$+[\kappa,\alpha,\varepsilon]_3.$$

ただしここで$[\kappa,\alpha,\varepsilon]_3$は$\kappa$, α, εの要素に関する3次以上の項の和を表わしている.

対角行列εの対角線要素ε_i $(i=1,2,\dots,d)$は相当に小さいので, われわれはκあるいはαの要素とε_iとの積はε_iに比べて無視できるものとしよう. すると, (4・95)からわれわれは次の式を得る:

$$(4\cdot 96)\quad \hat{\eta}\approx\alpha+\varepsilon+\alpha^2+\kappa\alpha+M\alpha\sigma(\hat{x})\alpha.$$

すると,

$$(4\cdot 97)\quad \hat{\eta}-\alpha\hat{\eta}-\alpha\approx\varepsilon+\kappa\alpha+M\alpha\sigma(\hat{x})\alpha$$

となるので, (4・71)から近似的に次の式を得る:

$$(4\cdot 98)\quad \sigma[\hat{x}^{(n+1)}-\hat{x}]\leq[\varepsilon+\kappa\alpha+M\alpha\sigma(\hat{x})\alpha]\sigma(\hat{x})$$
$$=[E+\kappa\alpha\varepsilon^{-1}+M\alpha\sigma(\hat{x})\alpha\varepsilon^{-1}]\varepsilon\sigma(\hat{x}).$$

もし, κの各要素の絶対値が小さくて

$$(4\cdot 99)\qquad \kappa\alpha\varepsilon^{-1}\approx 0$$

であり，またαの各要素の絶対値が小さくて

(4.100)　$M\alpha\sigma(\hat{x})\alpha\varepsilon^{-1}\approx 0$

である場合には，(4.98)より近似的に

(4.101)　$\sigma[\tilde{x}^{(n+1)}-\hat{x}]\leq\varepsilon\cdot\sigma(\hat{x})$

を得る．この場合には，(4.44)と比べてわかるように，$\tilde{x}^{(n+1)}$は振動状態にあるものと同程度の精度をもつことになる．本来のNewton法の場合には，$\varkappa=0$としてよいから，(4.99)はおのずからみたされており，必要な条件は(4.100)のみとなる．

本来のNewton法の場合，あるいは行列$H(x)$が$J^{-1}(x)$に近くて(4.11)の第1式をみたす行列$\varkappa$の各要素の絶対値が小さい場合には，対角行列εの対角線要素ε_i $(i=1,2,\cdots,d)$がすべて小さければ，4.2の注意1で述べたように(4.19)で定まる$\hat{\delta}$に対しては$\hat{\delta}\approx\varepsilon$となるので，定理6の条件(4.67)は

(4.102)　$\alpha\geq 2\varepsilon$

となる．またこの場合には条件(4.99)は上に述べたように自動的にみたされている．かくてわれわれは次の結論を得

る.

本来のNewton法の場合，あるいは行列$H(x)$が$J^{-1}(x)$に十分近い場合には，対角行列εの対角線要素ε_i $(i=1,2,\cdots,d)$がすべて小さければ，行列$\alpha=(\alpha_{ij})$を

(i) $\alpha_{ii}\varepsilon_i^{-1}$ $(i=1,2,\cdots,d)$が1に比べて大きく，

(ii) αの要素の二つの積の絶対値がすべてε_i $(i=1,2,\cdots,d)$に比べて小さく，

なるように選ぶと，不等式(3.36)を用いて反復計算を停止することができ，しかも最後に得られる$\tilde{x}^{(n+1)}$は振動状態にあるものと同程度の精度をもち，

$$(4.103)\qquad \sigma[\tilde{x}^{(n+1)}-\hat{x}]\leq\varepsilon\cdot\sigma(\hat{x})$$

がほぼ成立する.

[注意2] $\sigma(v)$をベクトルのノルムとしても本節の議論はそのまま成り立つ．ただしこのときは，ε, α, K_0, L_0, ρ_1, ρ_2, ρ_3, R_r, η_r $(r=0,1,2,\cdots)$, $\hat{R}$, $\hat{\eta}$はすべてスカラーとなり，注意1の条件(i), (ii)は次のようになる：

"αはεに比べて大きく，$\varepsilon^{1/2}$に比べて小さい."

[注意3] K, α, εの各要素の絶対値が十分小さいときは，$\hat{\eta}$は方程式(4.72)によってつねに一意的に定まる.

実際，(4.72) を $\hat{\eta}$ に関する方程式と考えると，$\kappa, \alpha, \varepsilon \to 0$ ならば $\hat{\eta} = (E+R)(\alpha+\varepsilon)$ で $R \to 0$ となるような (4.72) の解 $\hat{\eta}$ はつねに一つただ一つ存在する．このことは，方程式 (4.45) の場合と全く同じ方法で証明されるので，ここでは一切省略する．

§5. 数値例

次の方程式を考える：

$$(5.1)\quad \begin{cases} \varphi(p,q) = -p^2 + ap + q - b = 0, \\ \psi(p,q) = pq - aq + c = 0. \end{cases}$$

この方程式は，3次の多項式

$$P(x) = x^3 - ax^2 + bx - c$$

の2次因数 $x^2 - px + q$ を求めるときに現われる．すなわち，

$$P(x) = (x^2 - px + q)(x + l) - (Ax + B)$$

とおいて両辺の係数を比較すると

$$a = p - l, \quad b = q - pl - A, \quad c = B - ql$$

を得るので，これらから l を消去して

$$\begin{cases} A = -p^2 + ap + q - b, \\ B = pq - aq + c \end{cases}$$

を得る．$x^2 - px + q$ が $P(x)$ の因数になるための条件は

$$A = B = 0$$

であるから，$P(x)$ が与えられるとき，2次因数 $x^2 - px + q$ を求めるためには，方程式 (5·1) を解けばよいことになる．

いま，

$$P(x) = (x-\alpha)(x-\beta)(x-\gamma)$$

とすれば，明らかに

(5·2) $a = \alpha + \beta + \gamma, \quad b = \alpha\beta + \beta\gamma + \gamma\alpha, \quad c = \alpha\beta\gamma$

であり，また $x^2 - px + q$ が $P(x)$ の因数であるならば，それは

$$(x-\alpha)(x-\beta), \quad (x-\beta)(x-\gamma), \quad (x-\gamma)(x-\alpha)$$

のいずれかであるから，(p, q) は

(5·3) $(\alpha+\beta, \alpha\beta), \quad (\beta+\gamma, \beta\gamma), \quad (\gamma+\alpha, \gamma\alpha)$

のいずれかである．すなわち，a, b, c が（5・2）であるときは，方程式（5・1）の解は（5・3）で与えられることになる．

いま，たとえば

(5・4)　$\alpha=1,\quad \beta=10^{-3},\quad \gamma=10$

とする．すると，（5・2）により

(5・5)　$a=11.001,\quad b=10.011,\quad c=0.01$

となり，（5・3）は次のようになる：

(5・6)　$(1.001, 0.001), (10.001, 0.01), (11, 10).$

われわれは，（5・5）で与えられる a, b, c に対して，方程式（5・1）の解のうち（1.001, 0.001）を本来のNewton法を用いて計算してみた．

（5・1）で与えられる $\varphi(p, q)$, $\psi(p, q)$ の p, q に関するヤコビ行列 $J(p, q)$ は

(5・7)　$$J(p,q)=\begin{bmatrix} -(2p-a) & 1 \\ q & p-a \end{bmatrix}$$

となるので，$J^{-1}(p, q)$ は次のようになる：

$$(5\cdot 8)\qquad J^{-1}(p,q)=\frac{1}{D(p,q)}\begin{bmatrix} p-a & -1 \\ -q & -(2p-a)\end{bmatrix}.$$

ただし

$$(5\cdot 9)\qquad \begin{aligned} D(p,q)&=\det J(p,q)\\ &=-(2p^2-3ap+q+a^2)\end{aligned}$$

である．したがって

$$(5\cdot 10)\qquad \begin{cases} f(p,q)=p-\dfrac{1}{D(p,q)}\left[(p-a)\varphi(p,q)-\psi(p,q)\right], \\ g(p,q)=q-\dfrac{1}{D(p,q)}\left[-q\varphi(p,q)-(2p-a)\psi(p,q)\right]\end{cases}$$

とおけば，本来のNewton法は

$$(5\cdot 11)\qquad \begin{cases} p_n=f(p_{n-1},q_{n-1}), \\ q_n=g(p_{n-1},q_{n-1})\end{cases}$$

によって逐次 (p_n, q_n) $(n=1,2,\cdots)$ を計算することになる．

$p_0=2$，$q_0=0$ とし，10進8桁の浮動小数点型演算でこの計算を実行した結果は次の通りである：

n	p_n	$\left\lvert\frac{p_n-p_{n-1}}{p_{n-1}}\right\rvert$	q_n	$\left\lvert\frac{q_n-q_{n-1}}{q_{n-1}}\right\rvert$
0	2.0		0.0	
1	0.85843290	0.5708	0.0011109877	
2	0.99879570	0.1635	0.0011396923	0,8974
3	1.0009995	2.206×10^{-3}	0.0010000307	0,1225
4	1.0010001	5.994×10^{-7}	0.00099999998	3.072×10^{-5}
5	1.0010000	9.990×10^{-8}	0.00099999998	0
6	1.0010000	0	0.00099999998	0

表 1

$n \geq 3$ では

(5・12) $1 \leq p \leq 1.002, \quad 0.0009 \leq q \leq 0.0011$

であるから，この範囲で10進8桁浮動小数点型演算で $f(p,q)$, $g(p,q)$ を計算した場合の誤差を考える．$f(p,q)$, $g(p,q)$ の計算に含まれる丸め誤差を逐次評価してゆくと，(5・12) の範囲においては最終的に次の結果が得られる：

(5・13) $\left\{ \left|\tilde{f}(p,q)-f(p,q)\right| < 2.231\times10^{-7}\left|f(p,q)\right|, \right.$

$$\left| |\tilde{g}(p,q)-g(p,q)| < 2.477\times10^{-7}|g(p,q)| \right. .$$

ただしここで $\tilde{f}(p,q)$, $\tilde{g}(p,q)$ はそれぞれ $f(p,q)$, $g(p,q)$ の計算によって得られる値を示している．(5・13) は §3 の (3・1) が

$$(5\cdot14)\qquad \varepsilon_1=2.231\times10^{-7},\quad \varepsilon_2=2.477\times10^{-7}$$

に対して成り立っていることを示している．

さて，表1からわかるように，反復計算 (5・11) によって得られる (p_n, q_n) のうちで振動状態にあるものは

$$(5\cdot15)\qquad p_5=1.001,\quad q_5=0.00099999998$$

である．いまの場合，(p, q) の真の値 $(\hat{p}, \hat{q})$ は (5・6) により

$$(5\cdot16)\qquad \hat{p}=1.001,\quad \hat{q}=0.001$$

であるから，(5・15) の相対誤差は次のようになる：

$$(5\cdot17)\quad \begin{cases} |p_5-\hat{p}|/|\hat{p}|=0, \\ |q_5-\hat{q}|/|\hat{q}|=2\times10^{-8}. \end{cases}$$

(5・14) により，これは定理5の (4・18)，とくに (4・44)

が成り立っていることを示している．

なお，(5.14)より

$$\varepsilon_1,\ \varepsilon_2 < \hat{\varepsilon} = 2.5\times10^{-7}.$$

したがって，不等式(3.36)を用いて(5.11)の反復計算を停止するため，4・3の注意1，2により，

$$\alpha = \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{bmatrix}$$

として，たとえば

$$(5.18)\quad \begin{cases} \alpha_{11} = \alpha_{22} = 10^{-5} < \sqrt{2.5\times10^{-7}} = 5\times10^{-4} \\ \alpha_{12} = \alpha_{21} = 0 \end{cases}$$

を選ぶと，表1からわかるように反復計算は(p_5, q_5)を出して停止する．(p_5, q_5)は振動状態にある値であるから，これは4・3の結論が正しいことを示している．

αとして

$$(5.19)\quad \alpha_{11} = \alpha_{22} = 10^{-4},\quad \alpha_{12} = \alpha_{21} = 0$$

を選ぶと，表1からわかるように反復計算は(p_4, q_4)を出して停止する．(p_4, q_4)の相対誤差は

$$\begin{cases} |p_4-\hat{p}|/|\hat{p}| = 10^{-7}/1.001 \approx 0.9990\times 10^{-7}, \\ |q_4-\hat{q}|/|\hat{q}| = 2\times 10^{-8} \end{cases}$$

であるから，振動状態にあるものとほとんど違わない．

αとして

$$(5\cdot 20) \qquad \alpha_{11}=\alpha_{22}=10^{-6},\ 10^{-7}, \qquad \alpha_{12}=\alpha_{21}=0$$

を選ぶと，反復計算は(5・18)の場合と同じく(p_5, q_5)を出して停止する．

αとして

$$\alpha_{11}=\alpha_{22}=10^{-8}, \qquad \alpha_{12}=\alpha_{21}=0$$

を選ぶと，表1からわかるように反復計算は(p_6, q_6)を出して停止する．(p_6, q_6)は(p_5, q_5)と同じであるから，この場合には計算を1回余計にしたことになる．

文献

[1] Urabe, M.: Convergence of numerical iteration in solution of equations, J. Sci. Hiroshima Univ., Ser. A, 19(1956), 479-489.

[2] Urabe, M.: Error estimation in numerical solution of equations by iteration process, J. Sci. Hiroshima Univ., Ser. A-I, 26(1962), 77-91.