COMPONENT-WISE ERROR ANALYSIS OF ITERATIVE METHODS PRACTICED ON A FLOATING-POINT SYSTEM

Minoru URABE

(Received Sept. 30, 1972)

Abstract

In the present paper the author gives a component-wise error estimate for an approximate solution obtained by practicing an iterative method on a floating-point system. Further, in case of Newton's method, he gives a more precise error estimate and also a simple convenient rule for stopping the iterative process.

§0. Introduction

Given a system of equations

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

or, in vector form,

$$(0.1) \quad \varphi(x) = 0.$$

In order to obtain a solution to (0.1) numerically by the use of an iterative method, we rewrite equation (0.1) in the form

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

and calculate $x^{(1)}, x^{(2)}, \dots$ successively by

$$(0.3) \quad x^{(n)} = f[x^{(n-1)}] \quad (n=1, 2, \dots)$$

starting from an appropriate $x^{(0)}$. An approximate solution to (0.1) then is given by $x^{(n)}$ obtained lastly by stopping the iterative process at an appropriate step.

The paper was prepared while the author was at Mathematics Research Center, University of Wisconsin, U.S.A.

In his previous papers [2] and [3], the author studied the case where the iterative process (0.3) is carried out on a computer on a fixed-point system, and in terms of norms of vectors, he gave an error estimate for an approximate solution obtained in the above way. In the present paper, the author studies the case where the iterative process (0.3) is carried out on a computer on a floating-point system as is usual in practical computations, and making use of some special vectors instead of norms, he gives a component-wise error estimate for an approximate solution obtained numerically. Making use of these results, in the present paper, the author shows that *if the function $f(x)$ in (0.2) is chosen appropriately, for instance, it is chosen so that the iterative process (0.3) may become Newton's method, then every component of a desired solution can be computed with the accuracy of the evaluation of $f(x)$ in the iterative process (0.3) even if the magnitudes of components of a desired solution differ considerably from each other.*

An error estimate given in the present paper is an apriori one. A method of getting an aposteriori component-wise error estimate will be published subsequently elsewhere.

§1. Preliminaries

DEFINITION 1. *For any vector v with components $v_1, v_2, \dots, v_d$, a vector with components $|v_1|, |v_2|, \dots, |v_d|$ is called a magnitude-vector of v and is denoted by $\sigma(v)$.*

It is clear that

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

where $\sigma_i(v)$ denotes the i -th component of $\sigma(v)$.

DEFINITION 2. *For any square matrix A with elements a_{ij} ($i, j=1, 2, \dots, d$), a square matrix with elements $|a_{ij}|$ ($i, j=1, 2, \dots, d$) is called a magnitude-matrix of A and is denoted by $\sigma(A)$.*

It is clear that

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

where $\sigma_{ij}(A)$ denotes the ij -element of $\sigma(A)$.

DEFINITION 3. For any d -dimensional vectors v and w ,

$$v \geq w \text{ or } w \leq v$$

denotes that

$$v_i \geq w_i \quad (i=1, 2, \dots, d)$$

where v_i and w_i denote respectively the i -th components of v and w . For any d -th order square matrices A and B ,

$$A \geq B \text{ or } B \leq A$$

denotes that

$$a_{ij} \geq b_{ij} \quad (i, j=1, 2, \dots, d),$$

where a_{ij} and b_{ij} denote respectively the ij -elements of A and B .

It is clear that for vector v and matrix A ,

$$v \geq 0 \text{ and } A \geq 0$$

mean respectively

$$v_i \geq 0 \text{ for all } i \text{ and } a_{ij} \geq 0 \text{ for all } i \text{ and } j,$$

where v_i and a_{ij} denote respectively components of v and elements of A . Also it is clear that for vectors v, w and matrices A, B ,

$$v \geq w \text{ and } A \geq B$$

are respectively equivalent to

$$v - w \geq 0 \text{ and } A - B \geq 0.$$

By our definition, it is also clear that for any vector v and any matrix A ,

$$\sigma(v) \geq 0, \sigma(A) \geq 0$$

and

$$\sigma(v)=0 \text{ if and only if } v=0,$$

$$\sigma(A)=0 \text{ if and only if } A=0.$$

LEMMA 1. *Let v and w be d -dimensional vectors and A and B be d -th order square matrices. Then*

$$\sigma(v+w) \leq \sigma(v) + \sigma(w),$$

$$\sigma(A+B) \leq \sigma(A) + \sigma(B),$$

$$\sigma(Av) \leq \sigma(A)\sigma(v),$$

$$\sigma(AB) \leq \sigma(A)\sigma(B).$$

This lemma can be proved easily from Definitions 1, 2 and 3.

In the present paper the following well-known lemmas will be used frequently.

LEMMA 2. *For any square matrix H .*

$$\lim_{n \rightarrow \infty} H^n = 0 \quad (n \text{ a positive integer})$$

if and only if $\rho(H) < 1$, where $\rho(H)$ denotes the spectral radius of H .

LEMMA 3. *For any square matrix H , $(E-H)^{-1}$ exists and moreover*

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

if $\rho(H) < 1$, where E denotes the unit matrix.

The following corollaries to Lemmas 2 and 3 will be used also frequently.

COROLLARY 1. *For square matrices A and H , suppose that $0 \leq A \leq H$. If $\rho(H) < 1$, then $\rho(A) < 1$.*

Proof. Since $\rho(H) < 1$, by Lemma 2 $H^n \rightarrow 0$ as $n \rightarrow \infty$. Then by $0 \leq A \leq H$ and Lemma 1, $A^n \rightarrow 0$ as $n \rightarrow \infty$. By Lemma 2 this implies $\rho(A) < 1$. This completes the proof.

COROLLARY 2. *For square matrix $H \geq 0$, suppose that $\rho(H) < 1$. Then $(E-H)^{-1}$ exists and moreover $(E-H)^{-1} \geq 0$.*

This corollary is clear from Lemma 3.

§ 2. Convergence theorem of an iterative process

On the convergence of the iterative process (0.3), modifying the theorem proved by Collatz in his paper [1], we get

THEOREM 1. *For vector-valued function $f(x)$ defined on a set D of the x -space, suppose that*

$$(2.1) \quad \sigma[f(x') - f(x'')] \leq K_0 \cdot \sigma(x' - x'')$$

for any $x', x'' \in D$, where K_0 is a square matrix such that

$$(2.2) \quad K_0 \geq 0, \rho(K_0) < 1.$$

Assume that there is an $x^{(0)} \in D$ such that

$$(2.3) \quad S = \{h \mid \sigma[h - x^{(1)}] \leq K_0(E - K_0)^{-1} \cdot \sigma[x^{(1)} - x^{(0)}]\} \subset D,$$

where

$$(2.4) \quad x^{(1)} = f[x^{(0)}].$$

Then an infinite sequence $\{x^{(n)}\}_{(n=1, 2, \dots)}$ can be obtained in S by the iterative process

$$(2.5) \quad x^{(n)} = f[x^{(n-1)}] \quad (n=1, 2, \dots)$$

and, as $n \rightarrow \infty$, it converges to a limit $\hat{x}$, which is a unique solution to the equation

$$(2.6) \quad x = f(x)$$

in D .

Proof. First we shall prove by the induction that an infinite sequence $\{x^{(n)}\}_{(n=1, 2, \dots)}$ can be obtained in S by the iteration process (2.5).

Since $x^{(1)} \in S$ by (2.3), let us assume that $x^{(1)}, x^{(2)}, \dots, x^{(n-1)}$ have been obtained in S . Then $x^{(n)}$ is obtained by $x^{(n)} = f[x^{(n-1)}]$ since $x^{(n-1)} \in S \subset D$. In order to complete the induction, it is sufficient to prove that $x^{(n)} \in S$. Now, by (2.1) we have

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

and in a similarly way we successively have:

$$\begin{aligned}\sigma[x^{(3)} - x^{(2)}] &\leq K_0 \cdot \sigma[x^{(2)} - x^{(1)}] \leq K_0^2 \cdot \sigma[x^{(1)} - x^{(0)}], \\ &\dots\dots\dots \\ \sigma[x^{(n)} - x^{(n-1)}] &\leq K_0 \cdot [x^{(n-1)} - x^{(n-2)}] \leq K_0^{n-1} \cdot \sigma[x^{(1)} - x^{(0)}].\end{aligned}$$

Then from these inequalities we successively have:

$$\begin{aligned}\sigma[x^{(n)} - x^{(1)}] &\leq \sigma[x^{(n)} - x^{(n-1)}] + \sigma[x^{(n-1)} - x^{(n-2)}] + \dots + \sigma[x^{(2)} - x^{(1)}] \\ &\leq (K_0^{n-1} + K_0^{n-2} + \dots + K_0) \sigma[x^{(1)} - x^{(0)}] \\ &\leq K_0(E - K_0)^{-1} \cdot \sigma[x^{(1)} - x^{(0)}].\end{aligned}$$

By the definition of S , this means that $x^{(n)} \in S$. This completes the induction.

For the infinite sequence $\{x^{(n)}\}(n=1, 2, \dots)$, we then have

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

for any non-negative integer n . Then for any positive integer r we have

$$\begin{aligned}(2.7) \quad \sigma[x^{(n+r)} - x^{(n)}] &\leq (K_0^{n+r-1} + \dots + K_0^n) \cdot \sigma[x^{(1)} - x^{(0)}] \\ &\leq K_0^n (E - K_0)^{-1} \cdot \sigma[x^{(1)} - x^{(0)}].\end{aligned}$$

Since $\rho(K_0) < 1$, $\lim_{n \rightarrow \infty} K_0^n = 0$ by Lemma 2. This implies that

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

uniformly in positive integer r as $n \rightarrow \infty$. This implies that the infinite sequence $\{x^{(n)}\}(n=0, 1, 2, \dots)$ is convergent.

Let

$$(2.8) \quad \hat{x} = \lim_{n \rightarrow \infty} x^{(n)},$$

then clearly

$$(2.9) \quad \hat{x} \in S \subset D$$

because $x^{(n)} \in S(n=1, 2, \dots)$.

Since $f(x)$ is continuous with respect to x by (2.1), letting $n \rightarrow \infty$ in (2.5) we have

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

by (2.8). This proves that $\hat{x}$ is a solution to equation (2.6).

Lastly we shall show that equation (2.6) has no other solution than $\hat{x}$ in D . Suppose that equation (2.6) has a solution $\hat{x}' \in D$ besides $\hat{x}$. Then from

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

we have

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

that is,

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

However by Corollary 2, $(E - K_0)^{-1} \geq 0$ since $\rho(K_0) < 1$ and $K_0 \geq 0$. Then from the above inequality follows

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

which implies

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

that is,

$$\hat{x}' = \hat{x}.$$

This is a contradiction. This completes the proof.

REMARK 1. As is seen from the proof, in the above theorem one can replace magnitude-vectors by corresponding norms of vectors and the matrix K_0 by a scalar constant. The theorem obtained by such replacement is nothing else the theorem proved by Collatz in [1].

REMARK 2. The condition (2.3) requires that $x^{(0)}$ is an appropriate approximate solution to equation (2.6). In fact, if $x^{(0)}$ is an approximate solution with some accuracy to (2.6) lying in the interior of D , then $x^{(1)} \approx x^{(0)}$ by (2.4), consequently S becomes a small rectangular region with the center at $x^{(1)}$ and hence $S \subset D$ since $x^{(0)}$ lies in the interior of D .

§ 3. Error analysis of an iterative method practiced on a floating-point system 3.1. Oscillation theorem of an iterative process

When the iterative process (0.3) is carried out on a computer, we have a sequence $\{\tilde{x}^{(n)}\} (n=0, 1, 2, \dots; \tilde{x}^{(0)}=x^{(0)})$ defined successively by

$$(3.1) \quad \tilde{x}^{(n)} = \tilde{f}[\tilde{x}^{(n-1)}] \quad (n=1, 2, \dots),$$

where $\tilde{f}(x)$ denotes the value of $f(x)$ obtained on a computer. Due to round-off and truncation errors arising in computations, $\tilde{f}(x)$ will usually differ from $f(x)$. Hence the sequence $\{\tilde{x}^{(n)}\}$ obtained on a computer will usually differ from the ideal sequence $\{x^{(n)}\}$ defined successively by (0.3) mathematically.

We suppose that the computations are carried out on a floating-point system as is usual. Then we may assume that

$$(3.2) \quad \sigma[\tilde{f}(x) - f(x)] \leq \varepsilon \cdot \sigma[f(x)]$$

for x appearing in the computations, where ε is a diagonal matrix with diagonal elements $\varepsilon_i > 0$. Needless to say, (3.2) means that

$$(3.3) \quad |\tilde{f}_i(x) - f_i(x)| \leq \varepsilon_i \cdot |f_i(x)|$$

for all components $\tilde{f}_i(x)$ and $f_i(x)$ of $\tilde{f}(x)$ and $f(x)$, that is, a relative error of each component of $\tilde{f}(x)$ does not exceed the corresponding diagonal element of the matrix ε . We suppose that

$$(3.4) \quad \varepsilon_i \leq \hat{\varepsilon} \text{ for all } i.$$

Naturally we may assume that

$$(3.5) \quad 0 < \hat{\varepsilon} < 1.$$

for x appearing in the computations. Then it is clear that

$$(3.6) \quad 0 \leq \varepsilon \leq \hat{\varepsilon} E.$$

For the sequence $\{\tilde{x}^{(n)}\} (n=0, 1, 2, \dots)$ obtained practically on a computer, modifying the result proved by the author in [2], we get

THEOREM 2. *Let $f(x)$ be a vector-valued function with the property stated in Theorem 1. Suppose that $\hat{\varepsilon}$ is so small that*

$$(3.7) \quad \rho[(1+\varepsilon)K_0] > 1.$$

Assume that there is an $x^{(0)} \in D$ such that

$$(3.8) \quad \Sigma = \{h \mid \sigma[h - \tilde{x}^{(1)}] \leq \rho_1 + \rho_2\} \subset D,$$

where

$$(3.9) \quad \tilde{x}^{(1)} = \tilde{f}[x^{(0)}],$$

$$(3.10) \quad \rho_1 = K_0(E - K_0)^{-1} \cdot \sigma[\tilde{x}^{(1)} - x^{(0)}] + (E - K_0)^{-1} \varepsilon (E - \varepsilon)^{-1} \cdot \sigma[\tilde{x}^{(1)}],$$

$$(3.11) \quad \rho_2 = (E - L_0)^{-1} \varepsilon \{\rho_1 + \sigma[\tilde{x}^{(1)}]\}$$

and

$$(3.12) \quad L_0 = (E + \varepsilon)K_0.$$

Then the equation

$$(3.13) \quad x = f(x)$$

has a unique solution $\hat{x}$ in D and an infinite sequence $\{\tilde{x}^{(n)}\} (n=1, 2, \dots)$ can be obtained in Σ by the practical iterative process (3.1). The sequence $\{\tilde{x}^{(n)}\} (n=1, 2, \dots)$ obtained oscillates taking a finite number of values after a certain $\tilde{x}^{(n)}$ and for $\tilde{x}^{(n)}$ in such an oscillatory state it holds that

$$(3.14) \quad \sigma[\tilde{x}^{(n)} - \hat{x}] \leq (E - L^0)^{-1} \varepsilon \cdot \sigma(\hat{x}).$$

Proof. To begin with, let us note that

$$(3.15) \quad \sigma[\tilde{x}^{(1)} - x^{(1)}] \leq \varepsilon \cdot \sigma[x^{(1)}] \leq \varepsilon(E - \varepsilon)^{-1} \cdot \sigma[\tilde{x}^{(1)}],$$

where $x^{(1)} = f[x^{(0)}]$. In fact, by (3.2) we have

$$(3.16) \quad \begin{aligned} \sigma[\tilde{x}^{(1)} - x^{(1)}] &= \sigma[\tilde{f}(x^{(0)}) - f(x^{(0)})] \\ &\leq \varepsilon \cdot \sigma[f(x^{(0)})] \\ &= \varepsilon \cdot \sigma[x^{(1)}], \end{aligned}$$

which proves the first part of the inequalities (3.15). Since

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

by (3.16) we then have

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

that is,

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

Since $\rho(\varepsilon) < 1$ by (3.5), by Corollary 2 we then have

$$(3.17) \quad \sigma[x^{(1)}] \leq (E - \varepsilon)^{-1} \cdot \sigma[\tilde{x}^{(1)}].$$

From (3.16) then follows the second part of the inequalities (3.15).

First we shall prove that

$$(3.18) \quad S \subset \Sigma_0 = \{h \mid \sigma[h - \tilde{x}^{(1)}] \leq \rho_1\},$$

where S is the set defined by (2.3) in Theorem 1. Let h be an arbitrary point belonging to S . Then by (2.3), we have

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

By the use of (3.15) and (3.10), we then successively have:

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

This proves that $h \in \Sigma_0$ for any $h \in S$. Hence this proves (3.18).

From (3.18), by (3.8), it readily follows that

$$(3.19) \quad S \subset \Sigma_0 \subset \Sigma \subset D.$$

This says that the assumptions in Theorem 1 are all fulfilled. Hence by Theorem 1 equation (3.13) has a unique solution $\hat{x}$ in D .

Next we shall prove by the induction that an infinite sequence $\{\tilde{x}^{(n)}\}$ ($n=1, 2, \dots$) can be obtained in Σ by the practical iterative process (3.1). Since $\tilde{x}^{(1)} \in \Sigma$ by (3.8), let us assume that $\tilde{x}^{(1)}, \tilde{x}^{(2)}, \dots, \tilde{x}^{(n-1)}$ have been obtained in Σ . Then $\tilde{x}^{(n)}$ is obtained by $\tilde{x}^{(n)} = \tilde{f}[\tilde{x}^{(n-1)}]$ since $\tilde{x}^{(n-1)} \in \Sigma \subset D$. In order to complete the induction, it is sufficient to prove that $\tilde{x}^{(n)} \in \Sigma$.

Let $\{x^{(n)}\}$ ($n=1, 2, \dots$) be the infinite sequence described in Theorem 1, that is, the sequence obtained in S by the iterative process (2.5). Then by (3.2), (2.1), (3.12) and (3.15), we successively have:

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

In a similar way, we have

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

therefore by (3.20) we have

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

Continuing the above process, we finally have

$$\begin{aligned}
 (3.21) \quad \sigma[\tilde{x}^{(n)} - x^{(n)}] &\leq L_0^{n-1} \varepsilon \cdot \sigma[x^{(1)}] + L_0^{n-2} \varepsilon \cdot \sigma[x^{(2)}] \\
 &\quad + \dots + \varepsilon \cdot \sigma[x^{(n)}].
 \end{aligned}$$

However, since $x^{(n)} \in S$ ($n=1, 2, \dots$) by Theorem 1,

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

therefore we have

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

Hence by (3.15) and (3.10) we have

$$\begin{aligned}
\sigma[x^{(n)}] &\leq (E - K_0)^{-1} \varepsilon (E - \varepsilon)^{-1} \cdot \sigma[\tilde{x}^{(1)}] \\
&\quad + K_0(E - K_0)^{-1} \cdot \sigma[\tilde{x}^{(1)} - x^{(0)}] + \sigma[\tilde{x}^{(1)}] \\
&= \rho_1 + \sigma[\tilde{x}^{(1)}] \quad (n = 1, 2, \dots).
\end{aligned}$$

Thus by (3.21) we have

$$(3.22) \quad \sigma[\tilde{x}^{(n)} - x^{(n)}] \leq (L_0^{n-1} + L_0^{n-2} + \dots + E) \varepsilon \{\rho_1 + \sigma[\tilde{x}^{(1)}]\}.$$

Now by (3.12) and (3.6),

$$0 \leq L_0 = (E + \varepsilon)K_0 \leq (1 + \hat{\varepsilon})K_0$$

and by (3.7),

$$\rho[(1 + \hat{\varepsilon})K_0] < 1.$$

Hence by Corollary 1 we have

$$(3.23) \quad \rho(L_0) < 1.$$

Then by Lemma 3,

$$E + L_0 + \dots + L_0^{n-1} \leq (E - L_0)^{-1},$$

therefore from (3.22), by (3.11), we have

$$(3.24) \quad \sigma[\tilde{x}^{(n)} - x^{(n)}] \leq \rho_2.$$

Since $x^{(n)} \in S \subset \Sigma_0$, we thus have

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

This proves that $\tilde{x}^{(n)} \in \Sigma$. This completes the induction.

Now $\tilde{x}^{(n)}$'s are values obtained on a computer, therefore mutual distances of two distinct $\tilde{x}^{(n)}$'s cannot be smaller than a certain fixed positive constant. Then, since Σ is bounded, the number of distinct $\tilde{x}^{(n)}$'s must be finite. Thus it necessarily happens that

$$(3.25) \quad \tilde{x}^{(m+p)} = \tilde{x}^{(m)}$$

for a certain m and a certain positive integer p . Once (3.25) has happened, then

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

Hence the sequence $\{\tilde{x}^{(n)}\} (n=1, 2, \dots)$ oscillates taking p values

$$(3.26) \quad \tilde{x}^{(m)}, \tilde{x}^{(m+1)}, \dots, \tilde{x}^{(m+p-1)}$$

after $\tilde{x}^{(m)}$.

Lastly we shall prove that (3.14) holds for $\tilde{x}^{(n)}$ in an oscillatory state, that is, any one of the values given by (3.26).

As is seen from the proof of (3.21), we have

$$(3.27) \quad \sigma[\tilde{x}^{(n)} - x^{(n)}] \leq L_0^{n-1} \varepsilon \cdot \sigma[x^{(1)}] + L_0^{n-2} \varepsilon \cdot \sigma[x^{(2)}] \\ + \dots + \varepsilon \cdot \sigma[x^{(n)}]$$

for all $n=1, 2, \dots$. However, letting $r \rightarrow \infty$ in (2.7) in the proof of Theorem 1, we have

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

Therefore we have

$$(3.28) \quad \sigma[x^{(n)}] \leq \sigma[x^{(n)} - \hat{x}] + \sigma(\hat{x}) \\ \leq K_0^n (E - K_0)^{-1} \cdot \sigma[x^{(1)} - x^{(0)}] + \sigma[\hat{x}].$$

Hence from (3.27) we have:

$$(3.29) \quad \begin{aligned} & \sigma[\hat{x}^{(n)} - x^{(n)}] \\ & \leq (L_0^{n-1}\varepsilon K_0 + L_0^{n-2}\varepsilon K_0^2 + \cdots + \varepsilon K_0^n)(E - K_0)^{-1} \cdot \sigma[x^{(1)} - x^{(0)}] \\ & \quad + (L_0^{n-1} + L_0^{n-2} + \cdots + E)\varepsilon \cdot \sigma[\hat{x}]. \end{aligned}$$

However by (3.6) it holds that

$$(3.30) \quad \begin{aligned} & L_0^{n-1}\varepsilon K_0 + L_0^{n-2}\varepsilon K_0^2 + \cdots + \varepsilon K_0^n \\ & \leq \hat{\varepsilon}(\hat{L}_0^{n-1}K_0 + \hat{L}_0^{n-2}K_0^2 + \cdots + K_0^n) \\ & = \hat{\varepsilon}(\hat{L}_0^{n-1} + \hat{L}_0^{n-2}K_0 + \cdots + K_0^{n-1})K_0, \end{aligned}$$

where $\hat{L}_0 = (1 + \hat{\varepsilon})K_0$. Since $\hat{L}_0 K_0 = K_0 \hat{L}_0$, it is clear that

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

that is,

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

Hence from (3.30) we have

$$(3.31) \quad 0 \leq L_0^{n-1}\varepsilon K_0 + L_0^{n-2}\varepsilon K_0^2 + \cdots + \varepsilon K_0^n \leq \hat{L}_0^n - K_0^n.$$

However by (2.2) and (3.7)

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

Therefore by Lemma 2 we see that

$$(3.32) \quad \lim_{n \rightarrow \infty} (L_0^{n-1}\varepsilon K_0 + L_0^{n-2}\varepsilon K_0^2 + \cdots + \varepsilon K_0^n) = 0.$$

Let $\tilde{x}^{(l)}$ be any one of the values given by (2.26), that is, an arbitrary $\tilde{x}^{(n)}$ in an oscillatory state. Then clearly $\tilde{x}^{(l+p)} = \tilde{x}^{(l)}$, therefore for any positive integer r we have

$$(3.33) \quad \tilde{x}^{(l+rp)} = \tilde{x}^{(l)}.$$

Now put $n = l + rp$ in (3.29) and let $r \rightarrow \infty$. Then by (3.33), (2.8), (3.32) and (3.23), we have

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

which is nothing else than the desired inequality (3.14). This completes the proof.

REMARK 1. This remark is concerned with the assumption (3.2). If some component, say $\hat{x}_i$, of the solution $\hat{x}$ to equation (3.13) is zero, then $f_i(\hat{x}) = \hat{x}_i = 0$. However it is usual that $\tilde{f}_i(\hat{x}) \neq 0$ due to errors. In such a case, as is seen from (3.3), inequality (3.2) is not valid for $x = \hat{x}$. When some components of a desired solution is foreseen to be zero, it is thus recommended to shift x by some amount.

REMARK 2. The condition (3.8) in Theorem 2 is of the similar character as the condition (2.3) in Theorem 1. In fact, if $\hat{\varepsilon}$ is small and $x^{(0)}$ is an approximate solution with some accuracy to equation (3.13) lying in the interior of D , then it is easily seen in a similar way as the condition (2.3) that the condition (3.8) is fulfilled.

REMARK 3. The number of values taken by $\tilde{x}^{(n)}$ in an oscillatory state is usually greater than one as one sees frequently in numerical experiments (for example, see Example in [3]). The infinite sequence $\{\tilde{x}^{(n)}\} (n=1, 2, \dots)$ obtained by practicing the iterative process (0.3) on a computer thus does not usually converge while the infinite sequence $\{x^{(n)}\} (n=1, 2, \dots)$ defined mathematically by (0.3) does converge.

REMARK 4. If $K_0 \approx 0$ in (2.1), then by (3.12) $L_0 \approx 0$. In this case (3.14) implies that the following inequality holds approximately:

$$(3.34) \quad \sigma[\tilde{x}^{(n)} - \hat{x}] \leq \varepsilon \cdot \sigma(\hat{x}),$$

that is,

$$(3.35) \quad |\tilde{x}_i^{(n)} - \hat{x}_i| \leq \varepsilon_i |\hat{x}_i|$$

for every components $\tilde{x}_i^{(n)}$ and $\hat{x}_i$ of $\tilde{x}^{(n)}$ and $\hat{x}$. Inequality (3.35) means that a relative error of each component of $\tilde{x}^{(n)}$ in an oscillatory state does not exceed the corresponding diagonal element of the matrix ε . From this result we see that if one can choose $f(x)$ in (0.2) so that K_0 appearing in (2.1) may be small, then one can compute every component of a solution with the accuracy of the evaluation of $f(x)$ even if the

magnitudes of components of a desired solution differ considerably from each other.

3.2. A stopping rule of an iterative process

In case the iterative process (0.3) is carried out on a computer, in order to detect whether the practical iterative process (3.1) has attained an oscillatory state or not, it is necessary to store all $\tilde{x}^{(n)}$'s in the memory until equality (3.25) is detected to be valid for some m and some positive integer p . However this is usually inconvenient in practical computations. Hence in practical computations we frequently stop the iterative process by the inequality

$$(3.36) \quad \sigma[\tilde{x}^{(n)} - \tilde{x}^{(n-1)}] \leq \alpha \cdot \sigma[\tilde{x}^{(n-1)}]$$

choosing an appropriate matrix $\alpha \geq 0$, in other words, we stop the iterative process (3.1) at the step where (3.36) is detected to be valid.

However, if α is chosen too small, the criterion (3.36) may not work well for stopping the iterative process, in other words, inequality (3.36) may never be fulfilled and the practical iterative process (3.1) may be continued indefinitely, because the infinite sequence $\{\tilde{x}^{(n)}\}$ ($n=1, 2, \dots$) oscillates taking a finite number of values after a certain $\tilde{x}^{(n)}$ and hence $\sigma[\tilde{x}^{(n)} - \tilde{x}^{(n-1)}]$ does not always tend to zero as $n \rightarrow \infty$. The following theorem gives a sufficient condition for matrix α , by means of which one can stop the iterative process by the criterion (3.36).

THEOREM 3. *Suppose that the assumptions of Theorem 2 are all valid, and further that ε is so small that in addition to (3.37) the following inequality holds:*

$$(3.37) \quad \rho(\delta_0) < 1,$$

where

$$(3.38) \quad \delta_0 = (E - L_0)^{-1} \varepsilon.$$

Then one can surely stop the iterative process (3.1) by the criterion (3.36), if one chooses a matrix $\alpha \geq 0$ so that

$$(3.39) \quad \alpha \geq 2\delta_0(E - \delta_0)^{-1}.$$

Proof. Suppose that $\tilde{x}^{(n-1)}$ is in an oscillatory state. Then $\tilde{x}^{(n)}$ is also in an oscillatory state. Hence by (3.14) in Theorem 2, we have

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

consequently we have

$$(3.40) \quad \sigma[\tilde{x}^{(n)} - \tilde{x}^{(n-1)}] \leq 2\delta_0 \cdot \sigma(\hat{x}).$$

However, since

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

we have

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

Due to (3.38), by Corollary 2 we then have

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

Hence from (3.40) we get

$$(3.41) \quad \sigma[\tilde{x}^{(n)} - \tilde{x}^{(n-1)}] \leq 2\delta_0 (E - \delta_0)^{-1} \cdot \sigma[\tilde{x}^{(n-1)}].$$

Then by (3.39) we have

$$(3.42) \quad \sigma[\tilde{x}^{(n)} - \tilde{x}^{(n-1)}] \leq \alpha \cdot \sigma[\tilde{x}^{(n-1)}].$$

This shows that inequality (3.36) is indeed valid for $\tilde{x}^{(n-1)}$ in an oscillatory state. Since an oscillatory state is attained after a finite number of repetitions of the iterative process (3.1), we thus see that inequality (3.36) is surely fulfilled after a finite number of repetitions of the iterative process. This completes the proof.

The following theorem gives an error estimate for $\tilde{x}^{(n)}$ obtained finally by stopping the iterative process (3.1) by the criterion (3.36).

THEOREM 4. *Suppose that the assumptions of Theorem 2 are all valid, and that inequality (3.36) holds for certain $\tilde{x}^{(n)}$ of a sequence $\{\tilde{x}^{(n)}\}$ ($n=0, 1, 2, \dots$) obtained by the iterative process (3.1). If elements of the matrix $\alpha \geq 0$ are all so small that*

$$(3.43) \quad \rho(L_0 + \alpha) < 1,$$

then for $\tilde{x}^{(n)}$ satisfying (3.36) it holds that

$$(3.44) \quad \sigma[\tilde{x}^{(n)} - \hat{x}] \leq \{L_0[E - (L_0 + \alpha)]^{-1}\alpha \\ + (E - \alpha)[E - (L_0 + \alpha)]^{-1}\varepsilon\} \cdot \sigma(\hat{x}).$$

Proof. From (3.36) we successively have:

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

Hence we have

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

Due to (3.43), by Corollary 2 we then have

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

Then for $\sigma[\tilde{x}^{(n)} - \hat{x}]$ we successively have:

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

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

Hence by (3.45) we get

$$(3.46) \quad \sigma[\tilde{x}^{(n)} - \hat{x}] \leq \{\varepsilon + L_0[E - (L_0 + \alpha)]^{-1}(\alpha + \varepsilon)\} \cdot \sigma(\hat{x}),$$

from which readily follows (3.44). This completes the proof.

REMARK. If elements of the matrix $\alpha \geq 0$ are all so small that all products of elements of L_0 and those of α may be neglected as compared with diagonal elements of ε , then from (3.44) we see that the following inequality holds approximately:

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

because in this case

$$L_0[E - (L_0 + \alpha)]^{-1} \alpha \approx 0$$

and

$$\begin{aligned} E - (L_0 + \alpha) &= (E - L_0)[E - (E - L_0)^{-1} \alpha] \\ &\approx (E - L_0)(E - \alpha). \end{aligned}$$

Inequality (3.47) is the same as (3.14). This means that $\tilde{x}^{(n)}$ has the same degree of the accuracy as $\tilde{x}^{(n)}$ in an oscillatory state.

From this result, taking into account Theorem 3, we get the following conclusion about the choice of the matrix $\alpha \geq 0$.

It is recommended to choose α so that all elements of α may be large as compared with diagonal elements of the matrix ε but products of elements of α and those of L_0 may be small as compared with diagonal elements of the matrix ε . If the matrix α can be chosen in such a way, then by stopping the iterative process (3.1) by the criterion (3.36) one can get an approximate solution $\tilde{x}^{(n)}$ having the same degree of the accuracy as $\tilde{x}^{(n)}$ in an oscillatory state.

Clearly such choice of the matrix α is always possible when the elements of K_0 are all very small.

§ 4. Error analysis of a generalized Newton's method

4.1. Reformulation of a generalized Newton's method

Given an equation

$$(4.1) \quad \varphi(x) = 0.$$

Newton's method is an iterative method to find a solution to (4.1) by calculating $x^{(n)}$ ($n=1, 2, \dots$) successively by (0.3) with $f(x)$ defined by

$$(4.2) \quad f(x) = x - J^{-1}(x)\varphi(x),$$

where $J(x)$ is a Jacobian matrix of $\varphi(x)$ with respect to x . In practical computations, however, on account of the inconvenience of calculating $J^{-1}(x)$, one sometimes makes use of

$$(4.3) \quad f(x) = x - H(x)\varphi(x)$$

instead of $f(x)$ given by (4.2) choosing an appropriate matrix $H(x)$. In the present paper, such a method will be called a *generalized Newton's method*.

For $\varphi(x)$ and $H(x)$, we assume that

(i) $\varphi(x)$ is differentiable with respect to x on a convex compact set D of the x -space and the Jacobian matrix $J(x)$ of $\varphi(x)$ with respect to x satisfies a Lipschitz condition

$$(4.4) \quad \sigma[J(x') - J(x'')] \leq M \cdot \sigma(x' - x'')$$

for any $x', x'' \in D$, where M is a constant 3rd-order tensor with non-negative components,

(ii) $H(x)$ satisfies a Lipschitz condition

$$(4.5) \quad \sigma[H(x') - H(x'')] \leq C \cdot \sigma(x' - x'')$$

for any $x', x'' \in D$, where C is a constant 3rd-order tensor with nonnegative components.

Here by (4.4) is meant that

$$(4.6) \quad |J_{ij}(x') - J_{ij}(x'')| \leq \sum_k M_{ijk} |x'_k - x''_k|$$

for all i and j , where $J_{ij}(x)$ are elements of $J(x)$, M_{ijk} are components of the 3rd-order tensor of M , and x'_k and x''_k are respectively components of vectors x' and x'' . Inequality (4.5) means that inequalities similar to (4.6) holds for $H(x)$. For Newton's method, since $H(x) = J^{-1}(x)$, it is easily seen that (ii) follows from (i) if $\det J(x)$ never vanishes on D , in other words, (ii) may be replaced by the assumption:

$$(ii') \quad \det J(x) \neq 0 \quad \text{for any } x \in D.$$

Now for any $x', x'' \in D$, by the assumption (i) we have

$$(4.7) \quad \begin{cases} \varphi(x'') - \varphi(x') = \int_0^1 J[x' + \theta(x'' - x')] (x'' - x') d\theta, \\ \sigma\{J[x' + \theta(x'' - x')] - J(x')\} \leq M \cdot \sigma(x'' - x') \quad (0 \leq \theta \leq 1). \end{cases}$$

Hence if we put

$$(4.8) \quad \varphi(x'') - \varphi(x') = J(x')(x'' - x') + \Psi(x'', x'),$$

then we have

$$(4.9) \quad \sigma[\Psi(x'', x')] \leq M \cdot \sigma(x'' - x') \cdot \sigma(x'' - x')$$

for any $x', x'' \in D$. Then for $f(x)$ defined by (4.3), we have

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

from which by (4.9) and (4.5) follows

$$(4.10) \quad \begin{aligned} \sigma[f(x') - f(x'')] &\leq \sigma[E - H(x')J(x')] \cdot \sigma(x' - x'') \\ &\quad + \sigma[H(x')] \cdot M \cdot \sigma(x' - x'') \cdot \sigma(x' - x'') \\ &\quad + C \cdot \sigma(x' - x'') \cdot \sigma[\varphi(x'')] \end{aligned}$$

$$= \{\sigma[E - H(x')J(x')] + \sigma[H(x')] \cdot M \cdot \sigma(x' - x'') \\ + C \cdot \sigma[\varphi(x'')]\} \cdot \sigma(x' - x'').$$

Since D is a compact set, we may suppose that

$$(4.11) \quad \begin{cases} \sigma[E - H(x)J(x)] \leq \kappa, \\ \sigma[H(x)] \cdot M \leq N \end{cases}$$

for all $x \in D$, where $\kappa \geq 0$ and N is a constant 3rd-order tensor with non-negative components. By (4.11), from (4.10) we have

$$(4.12) \quad \sigma[f(x') - f(x'')] \leq K(x', x'') \cdot \sigma(x' - x'')$$

for any $x', x'' \in D$, where $K(x', x'')$ is a matrix such that

$$(4.13) \quad K(x', x'') = \kappa + N \cdot \sigma(x' - x'') + C \cdot \sigma[\varphi(x'')] \geq 0.$$

For matrix $K(x', x'')$ we assume that

(iii) *there is a constant matrix $K_0 \geq 0$ such that*

$$(4.14) \quad K(x', x'') \leq K_0 \quad \text{for all } x', x'' \in D$$

and

$$(4.15) \quad \rho(K_0) < 1.$$

As seen from (4.13), assumption (iii) is fulfilled if D is a small region containing a solution to the given equation (4.1), because in such a case $\sigma(x' - x''), \sigma[\varphi(x'')] \approx 0$ and in a generalized Newton's method one may suppose that all elements of κ are small so that $\rho(\kappa) < 1$.

By (4.14), from (4.12) we have

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

for any $x', x'' \in D$. By (4.15), this means that $f(x)$ given by (4.3) satisfies the requirement in Theorem 1. Thus we see that if $x^{(0)} \in D$ is taken so that the condition in Theorem 1 or 2 may be fulfilled and, in case of Theorem 2, further ε is assumed to be small so that (3.7) may be valid, then the conclusions of Theorem 1 or 2 are all valid in a generalized Newton's method.

REMARK. If we put

$$(4.16) \quad \nu(x) = E - H(x)J(x),$$

then by (4.11) and (4.14) we have

$$0 \leq \sigma[\nu(x)] \leq \kappa \leq K_0,$$

from which follows

$$0 \leq \sigma[\nu^n(x)] \leq \{\sigma[\nu(x)]\}^n \leq K_0^n \quad (n=1, 2, \dots).$$

Since $\lim_{n \rightarrow \infty} K_0^n = 0$ by Lemma 2 due to (4.15), we then have

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

that is,

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

By Lemma 2, this implies that

$$(4.17) \quad \rho[\nu(x)] > 1.$$

Then by Lemma 3, we have

$$\det [H(x)J(x)] = \det [E - \nu(x)] \neq 0,$$

which clearly implies

$$(4.18) \quad \det H(x), \det J(x) \neq 0 \quad \text{for any } x \in D.$$

By this it is clear that the given equation (4.1) is equivalent to

$$(4.19) \quad x = f(x)$$

for $f(x)$ given by (4.3).

4.2. Error analysis of Newton's method

For $f(x)$ given by (4.3), let us denote by $\tilde{f}(x)$ the value of $f(x)$ obtained on a computer. For $\tilde{f}(x)$, we suppose that (3.2) is valid in D for a diagonal matrix $\varepsilon \geq 0$ and diagonal elements ε_i of ε are so small that (3.7) may be valid. Let $x^{(0)} \in D$ be a point for which (3.8) is valid. Then as is shown in Theorem 2, we obtain an infinite sequence $\{\tilde{x}^{(n)}\}$ ($n=1, 2, \dots$) by the practical iterative process (3.1), that is, by practicing the iterative process (0.3)– that is, a generalized Newton's method– on a

computer. By Theorem 2, the sequence $\{\tilde{x}^{(n)}\}$ ($n=1, 2, \dots$) obtained oscillates taking a finite number of values, but in the present case, for $\tilde{x}^{(n)}$ in an oscillatory state we can get an error estimate more precise than the one given in Theorem 2. The following theorem gives indeed such an error estimate.

THEOREM 5. *Let $\hat{x}$ be a unique solution to the given equation (4.1) lying in D . Let $\tilde{x}^{(n)}$ be an approximate solution in an oscillatory state obtained by a generalized Newton's method, that is, an $\tilde{x}^{(n)}$ in an oscillatory state obtained by repeating the iterative process (3.1) with $f(x)$ defined by (4.3). Then it holds that*

$$(4.20) \quad \sigma[\tilde{x}^{(n)} - \hat{x}] \leq \hat{\delta} \cdot \sigma(\hat{x}),$$

where $\hat{\delta}$ is a matrix such that $\hat{\delta} \geq 0$ and

$$(4.21) \quad \hat{\delta} = \{E - (E + \varepsilon)[\kappa + N\hat{\delta} \cdot \sigma(\hat{x})]\}^{-1}\varepsilon.$$

Proof. Since $\tilde{x}^{(n)}$ is in an oscillatory state, by Theorem 2 we have

$$(4.22) \quad \sigma[\tilde{x}^{(n)} - \hat{x}] \leq \delta_0 \cdot \sigma(\hat{x}),$$

where

$$(4.23) \quad \delta_0 = (E - L_0)^{-1}\varepsilon = [E - (E + \varepsilon)K_0]^{-1}\varepsilon.$$

Now, since $\hat{x} \in S$ by Theorem 1, we have

$$(4.24) \quad \sigma[\hat{x} - \tilde{x}^{(1)}] \leq \rho_1$$

by (3.18) in the proof of Theorem 2. Then

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

therefore by (4.23) we have

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

which by (3.11) means that

$$(4.25) \quad \delta_0 \cdot \sigma(\hat{x}) \leq \rho_2.$$

Then as is seen from Figure 1, by (4.24), (4.25) and (3.8), we have

$$(4.26) \quad V_1 = \{h \mid \sigma(h - \hat{x}) \leq \delta_0 \cdot \sigma(\hat{x})\} \subset \Sigma \subset D.$$

Figure 1

Then by (4.13) and (4.14), we have

$$(4.27) \quad K_1 = \max_{x \in V_1} K(x, \hat{x}) = \kappa + N\delta_0 \cdot \sigma(\hat{x}) \leq K_0.$$

Now, since $\tilde{x}^{(n)}$ is in an oscillatory state, by (4.22) and (4.26), we see that $\tilde{x}^{(n)} \in V_1$. Then by (3.2) and (4.12), we successively have:

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

Hence putting

$$(4.28) \quad L_1 = (E + \varepsilon)K_1,$$

we have

$$(4.29) \quad \sigma[\tilde{x}^{(n+1)} - \hat{x}] \leq L_1 \cdot \sigma[\tilde{x}^{(n)} - \hat{x}] + \varepsilon \cdot \sigma(\hat{x}).$$

By (4.27), it is clear that

$$(4.30) \quad L_0 \geq L_1 \geq 0.$$

Then due to (3.23) in the proof of Theorem 2, by Corollary 1 we have

$$(4.31) \quad \rho(L_1) < 1.$$

Now, since $\tilde{x}^{(n)}$ is in an oscillatory state, $\tilde{x}^{(n+1)}$ is also in an oscillatory state. Then similarly to (4.29), we have

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

Then by the use of (4.29) we have

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

This process can be continued indefinitely and thus we have

$$(4.32) \quad \begin{aligned} \sigma[\tilde{x}^{(n+r)} - \hat{x}] &\leq L_1^r \cdot \sigma[\tilde{x}^{(n)} - \hat{x}] \\ &\quad + (E + L_1 + \cdots + L_1^{r-1}) \varepsilon \cdot \sigma(\hat{x}) \end{aligned}$$

for any positive integer r . However, since $\tilde{x}^{(n)}$ is in an oscillatory state, there is a positive integer p such that $\tilde{x}^{(n+p)} = \tilde{x}^{(n)}$. Hence, if we put $r = p$ in (4.32), then we have

$$(4.33) \quad \begin{aligned} (E - L_1^p) \cdot \sigma[\tilde{x}^{(n)} - \hat{x}] \\ \leq (E + L_1 + \cdots + L_1^{p-1}) \varepsilon \cdot \sigma(\hat{x}). \end{aligned}$$

However,

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

and, due to (4.31), by Corollary 2 $(E - L_1)^{-1} \geq 0$ and $(E - L_1^p)^{-1} \geq 0$ both exist. Thus from (4.33) we get

$$(4.34) \quad \sigma[\tilde{x}^{(n)} - \hat{x}] \leq \delta_1 \cdot \sigma(\hat{x}),$$

where

$$(4.35) \quad \delta_1 = (E - L_1)^{-1} \varepsilon = [E - (E + \varepsilon)K_1]^{-1} \varepsilon.$$

From (4.30), by Lemma 3, it is clear that

$$(4.36) \quad \delta_0 \geq \delta_1 \geq 0.$$

The above procedure can be continued indefinitely. Hence we have a pair of infinite sequences of matrices $\{\delta_r\}$ and $\{K_r\}$ ($r=0, 1, 2, \dots$) such that

$$(4.37) \quad \begin{cases} \delta_r = [E - (E + \varepsilon)K_r]^{-1}\varepsilon, \\ K_{r+1} = \kappa + N\delta_r \cdot \sigma(\hat{x}) \end{cases} \quad (r=0, 1, 2, \dots),$$

$$(4.38) \quad \rho[(E + \varepsilon)K_r] > 1 \quad (r=0, 1, 2, \dots),$$

$$(4.39) \quad \begin{cases} \delta_0 \geq \delta_1 \geq \delta_2 \geq \dots \geq 0, \\ K_0 \geq K_1 \geq K_2 \geq \dots \geq 0, \end{cases}$$

and

$$(4.40) \quad \sigma[\hat{x}^{(n)} - \hat{x}] \leq \delta_r \cdot \sigma(\hat{x}) \quad (r=0, 1, 2, \dots).$$

By (4.39) we have

$$(4.41) \quad \hat{\delta} = \lim_{r \rightarrow \infty} \delta_r \geq 0, \quad \hat{K} = \lim_{r \rightarrow \infty} K_r \geq 0.$$

Since $0 \leq (E + \varepsilon)\hat{K} \leq (E + \varepsilon)K_0 = L_0$, we have

$$(4.42) \quad \rho[(E + \varepsilon)\hat{K}] < 1$$

similarly to (4.31). For $\hat{\delta}$ and $\hat{K}$, by (4.40) and (4.37) it holds that

$$(4.43) \quad \sigma[\hat{x}^{(n)} - \hat{x}] \leq \hat{\delta} \cdot \sigma(\hat{x})$$

and

$$(4.44) \quad \begin{cases} \hat{\delta} = [E - (E + \varepsilon)\hat{K}]^{-1}\varepsilon, \\ \hat{K} = \kappa + N\hat{\delta} \cdot \sigma(\hat{x}). \end{cases}$$

Inequality (4.43) proves (4.20) and substituting the second of (4.44) into the first of (4.44) we get (4.21). This completes the proof.

REMARK 1. If diagonal elements of the diagonal matrix ε are all small as is usual in practical computations, then by (4.21) all elements of the matrix $\hat{\delta}$ are also small, therefore (4.21) implies

$$(4.45) \quad \hat{\delta} \approx (E - \kappa)^{-1} \varepsilon.$$

For Newton's method, as is seen from (4.11), we may suppose that

$$(4.46) \quad \kappa = 0.$$

Then (4.45) implies

$$(4.47) \quad \hat{\delta} \approx \varepsilon.$$

Hence by (4.20) we have

$$(4.48) \quad \sigma[\tilde{x}^{(n)} - \hat{x}] \leq \varepsilon \cdot \sigma(\hat{x})$$

approximately. This is of the same form as (3.34). Thus we see that *if one makes use of Newton's method, one can usually compute every component of a solution with the accuracy of the evaluation of $f(x)$ in the iterative process (0.3).*

From the above reasonings, we see also that *even if a generalized Newton's method is employed, every component of a solution will be computed with the accuracy of the evaluation of $f(x)$ provided all elements of the matrix κ are small, in other words, $H(x)$ in (4.3) is close to $J^{-1}(x)$.*

REMARK 2 Matrix $\hat{\delta}$ is determined uniquely by equality (4.21) provided all diagonal elements of the diagonal matrix ε are sufficiently small. For the proof, see the appendix.

4.3 A stopping rule in a generalized Newton's method

In a generalized Newton's method a criterion of the form (3.36) is also used frequently for stopping the iterative process.

In case Newton's method is employed, by Theorem 5 we have (4.20) for $\tilde{x}^{(n)}$ in an oscillatory state. Hence analogously to Theorem 3 we have the following theorem.

THEOREM 6. *In a generalized Newton's method described in 4.2, put*

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

and suppose that diagonal elements of the diagonal matrix ε are all so small that

$$(4.49) \quad \rho(\delta_0) < 1.$$

If a matrix $\alpha \geq 0$ is chosen so that

$$(4.50) \quad \alpha \geq 2\hat{\delta}(E - \hat{\delta})^{-1}$$

for $\hat{\delta}$ in Theorem 5, then one can surely stop the iterative process in a generalized Newton's method by the criterion (3.36).

As is seen from the proof of Theorem 5,

$$\hat{\delta} = \lim_{r \rightarrow \infty} \delta_r,$$

where $\{\delta_r\}$ ($r=0, 1, 2, \dots$) is an infinite sequence of matrices such that

$$\delta_0 \geq \delta_1 \geq \delta_2 \geq \dots \geq 0.$$

Hence due to (4.49), by Corollary 1, we have

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

Theorem 6 then can be proved in the same way as Theorem 3.

The following theorem corresponds to Theorem 4 and it gives an error estimate for $\tilde{x}^{(n)}$ obtained finally by stopping the iterative process in a generalized Newton's method by the criterion (3.36).

THEOREM 7. *In a generalized Newton's method described in 4.2, suppose that inequality (3.36) holds for certain n . Putting*

$$L_0 = (E + \varepsilon)K_0,$$

suppose that elements of the matrix $\alpha \geq 0$ are all so small that

$$(4.51) \quad \rho(L_0 + \alpha) < 1$$

and

$$(4.52) \quad \Sigma_\alpha = \{h \mid \sigma[h - \tilde{x}^{(1)}] \leq \rho_1 + \rho_3\} \subset D,$$

where $\tilde{x}^{(1)} = \tilde{f}[x^{(0)}]$, ρ_1 is the vector given by (3.10) and

$$(4.53) \quad \rho_3 = [E - (L_0 + \alpha)]^{-1}(\alpha + \varepsilon)\{\rho_1 + \sigma[\tilde{x}^{(1)}]\}.$$

For $\tilde{x}^{(n)}$ satisfying (3.36), it then holds that

$$(4.54) \quad \sigma[\tilde{x}^{(n)} - \hat{x}] \leq (\hat{\eta} - \alpha\hat{\eta} - \alpha) \cdot \sigma(\hat{x}),$$

where $\hat{x}$ is a unique solution to the given equation (4.1) lying in D and $\hat{\eta}$ is a matrix such that $\hat{\eta} \geq 0$ and

$$(4.55) \quad \hat{\eta} = \{E - [\alpha + (E + \varepsilon)(\kappa + N\hat{\eta} \cdot \sigma(\hat{x}))]\}^{-1}(\alpha + \varepsilon).$$

Proof. From (3.36), by (3.2) and (4.12), we successively have:

$$(4.56) \quad \begin{aligned} \sigma[\tilde{x}^{(n-1)} - \hat{x}] &\leq \sigma[\tilde{x}^{(n-1)} - \tilde{x}^{(n)}] + \sigma\{\tilde{f}[\tilde{x}^{(n-1)}] - f[\tilde{x}^{(n-1)}]\} \\ &\quad + \sigma\{f[\tilde{x}^{(n-1)}] - f(\hat{x})\} \\ &\leq \alpha \cdot \sigma[\tilde{x}^{(n-1)}] + \varepsilon \cdot \sigma\{f[\tilde{x}^{(n-1)}]\} \\ &\quad + K(\tilde{x}^{(n-1)}, \hat{x}) \cdot \sigma[\tilde{x}^{(n-1)} - \hat{x}] \\ &\leq \alpha \cdot \sigma[\tilde{x}^{(n-1)} - \hat{x}] + \alpha \cdot \sigma(\hat{x}) \\ &\quad + \varepsilon \cdot \sigma\{f[\tilde{x}^{(n-1)}] - f(\hat{x})\} + \varepsilon \cdot \sigma(\hat{x}) \\ &\quad + K(\tilde{x}^{(n-1)}, \hat{x}) \cdot \sigma[\tilde{x}^{(n-1)} - \hat{x}] \\ &\leq [\alpha + (E + \varepsilon)K(\tilde{x}^{(n-1)}, \hat{x})] \cdot \sigma[\tilde{x}^{(n-1)} - \hat{x}] + (\alpha + \varepsilon) \cdot \sigma(\hat{x}). \end{aligned}$$

Then by (4.14) we have

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

that is,

$$(4.57) \quad [E - (L_0 + \alpha)] \cdot \sigma[\tilde{x}^{(n-1)} - \hat{x}] \leq (\alpha + \varepsilon) \cdot \sigma(\hat{x}).$$

Due to (4.51), by Corollary 2, $[E - (L_0 + \alpha)]^{-1} \geq 0$ exists, therefore from (4.57) we have

$$(4.58) \quad \sigma[\tilde{x}^{(n-1)} - \hat{x}] \leq \eta_0 \cdot \sigma(\hat{x}),$$

where

$$(4.59) \quad \eta_0 = [E - (L_0 + \alpha)]^{-1}(\alpha + \varepsilon).$$

Now for $\hat{x}$, (4.24) holds, therefore from (4.59) we have

$$(4.60) \quad \eta_0 \cdot \sigma(\hat{x}) \leq [E - (L_0 + \alpha)]^{-1}(\alpha + \varepsilon) \{\rho_1 + \sigma[\tilde{x}^{(1)}]\} = \rho_3.$$

Since $\rho_2 \leq \rho_3$ as is seen from (3.11) and (4.53), by (3.18), (3.8) and (4.52), we have

$$(4.61) \quad S \subset \Sigma_0 \subset \Sigma \subset \Sigma_\alpha \subset D.$$

Now by (4.24), (4.60) and (4.52) we have

$$(4.62) \quad W_1 = \{h \mid \sigma(h - \hat{x}) \leq \eta_0 \cdot \sigma(\hat{x})\} \subset \Sigma_\alpha \subset D.$$

Then by (4.13) and (4.14) we have

$$(4.63) \quad R_1 = \max_{x \in W_1} K(x, \hat{x}) = \kappa + N\eta_0 \cdot \sigma(\hat{x}) \leq K_0.$$

Since $\tilde{x}^{(n-1)} \in W_1$ as is seen from (4.58), from (4.56) we then have

$$(4.64) \quad \sigma[\tilde{x}^{(n-1)} - \hat{x}] \leq [\alpha + (E + \varepsilon)R_1] \cdot \sigma[\tilde{x}^{(n-1)} - \hat{x}] + (\alpha + \varepsilon) \cdot \sigma(\hat{x}).$$

Since

$$(4.65) \quad 0 \leq \alpha + (E + \varepsilon)R_1 \leq \alpha + (E + \varepsilon)K_0 = \alpha + L_0,$$

due to (4.51) by Corollary 1 we have

$$(4.66) \quad \rho[\alpha + (E + \varepsilon)R_1] < 1,$$

and hence by Corollary 2 we see that

$$\{E - [\alpha + (E + \varepsilon)R_1]\}^{-1} \geq 0$$

exists. Then from (4.64) we have

$$(4.67) \quad \sigma[\tilde{x}^{(n-1)} - \hat{x}] \leq \eta_1 \cdot \sigma(\hat{x}),$$

where

$$(4.68) \quad \eta_1 = \{E - [\alpha + (E + \varepsilon)R_1]\}^{-1}(\alpha + \varepsilon).$$

By (4.65) and (4.59) it is clear that

$$(4.69) \quad \eta_0 \geq \eta_1 \geq 0.$$

The above procedure can be continued indefinitely. Hence we have a pair of infinite sequences of matrices $\{\eta_r\}$ and $\{R_r\}$ ($r=0, 1, 2, \dots$) such that

$$(4.70) \quad \begin{cases} \eta_r = \{E - [\alpha + (E + \varepsilon)R_r]\}^{-1}(\alpha + \varepsilon), \\ R_{r+1} = \kappa + N\eta_r \cdot \sigma(\hat{x}) \end{cases} \quad (r=0, 1, 2, \dots; R_0 = K_0),$$

$$(4.71) \quad \rho[\alpha + (E + \varepsilon)R_r] < 1 \quad (r=0, 1, 2, \dots),$$

$$(4.72) \quad \begin{cases} \eta_0 \geq \eta_1 \geq \eta_2 \geq \dots \geq 0, \\ R_0 \geq R_1 \geq R_2 \geq \dots \geq 0, \end{cases}$$

and

$$(4.73) \quad \sigma[\hat{x}^{(n-1)} - \hat{x}] \leq \eta_r \cdot \sigma(\hat{x}) \quad (r=0, 1, 2, \dots).$$

By (4.72) we have

$$(4.74) \quad \hat{\eta} = \lim_{r \rightarrow \infty} \eta_r \geq 0, \quad \hat{R} = \lim_{r \rightarrow \infty} R_r \geq 0.$$

Since $0 \leq \alpha + (E + \varepsilon)\hat{R} \leq \alpha + (E + \varepsilon)R_0 = \alpha + L_0$, we have

$$(4.75) \quad \rho[\alpha + (E + \varepsilon)\hat{R}] < 1$$

similarly to (4.66). For $\hat{\eta}$ and $\hat{R}$, by (4.73) and (4.70), it holds that

$$(4.76) \quad \sigma[\hat{x}^{(n-1)} - \hat{x}] \leq \hat{\eta} \cdot \sigma(\hat{x})$$

and

$$(4.77) \quad \begin{cases} \hat{\eta} = \{E - [\alpha + (E + \varepsilon)\hat{R}]\}^{-1}(\alpha + \varepsilon), \\ \hat{R} = \kappa + N\hat{\eta} \cdot \sigma(\hat{x}). \end{cases}$$

If we substitute the second of (4.77) into the first of (4.77), then we have (4.55)

Now for $\sigma[\tilde{x}^{(n)} - \hat{x}]$, by means of (4.12), (4.13) and (4.76), we successively have:

$$\begin{aligned}
 (4.78) \quad \sigma[\tilde{x}^{(n)} - \hat{x}] &= \sigma\{\tilde{f}[\tilde{x}^{(n-1)}] - f(\hat{x})\} \\
 &\leq \sigma\{\tilde{f}[\tilde{x}^{(n-1)}] - f[\tilde{x}^{(n-1)}]\} + \sigma\{f[\tilde{x}^{(n-1)}] - f(\hat{x})\} \\
 &\leq \varepsilon \cdot \sigma\{f[\tilde{x}^{(n-1)}]\} + K(\tilde{x}^{(n-1)}, \hat{x}) \cdot \sigma[\tilde{x}^{(n-1)} - \hat{x}] \\
 &\leq \varepsilon \cdot \sigma\{f[\tilde{x}^{(n-1)}] - f(\hat{x})\} + \varepsilon \cdot \sigma(\hat{x}) \\
 &\quad + K(\tilde{x}^{(n-1)}, \hat{x}) \cdot \sigma[\tilde{x}^{(n-1)} - \hat{x}] \\
 &\leq (E + \varepsilon)[\kappa + N\hat{\eta} \cdot \sigma(\hat{x})]\hat{\eta} \cdot \sigma(\hat{x}) + \varepsilon \cdot \sigma(\hat{x}) \\
 &= \{\varepsilon + (E + \varepsilon)[\kappa + N\hat{\eta} \cdot \sigma(\hat{x})]\hat{\eta}\} \cdot \sigma(\hat{x}).
 \end{aligned}$$

However from (4.55) it follows that

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

that is,

$$\hat{\eta} - \alpha\hat{\eta} - (E + \varepsilon)[\kappa + N\hat{\eta} \cdot \sigma(\hat{x})]\hat{\eta} = \alpha + \varepsilon.$$

Hence from (4.78) we get (4.54). This completes the proof.

REMARK 1. Due to (4.75), by Lemma 3, equality (4.55) can be written as follows:

$$\begin{aligned}
 (4.79) \quad \hat{\eta} &= (\alpha + \varepsilon) + \{\alpha + (E + \varepsilon)[\kappa + N\hat{\eta} \cdot \sigma(\hat{x})]\}(\alpha + \varepsilon) \\
 &\quad + \{\alpha + (E + \varepsilon)[\kappa + N\hat{\eta} \cdot \sigma(\hat{x})]\}^2(\alpha + \varepsilon) + \dots.
 \end{aligned}$$

However, as is seen from (4.55),

$$\sigma(\hat{\eta}) = O[\sigma(\alpha) + \sigma(\varepsilon)] \text{ as } \alpha, \varepsilon \rightarrow 0.$$

Hence from (4.79) we have

$$(4.80) \quad \hat{\eta} = \alpha + \varepsilon + [\kappa, \alpha, \varepsilon]_2,$$

where $[\kappa, \alpha, \varepsilon]_2$ denotes a sum of terms of the second and higher order

in elements of κ , α and ε . Then substituting (4.80) into (4.79), we finally obtain

$$(4.81) \quad \hat{\gamma} = \alpha + \varepsilon + [\alpha + \kappa + N(\alpha + \varepsilon) \cdot \sigma(\hat{x})](\alpha + \varepsilon) + [\kappa, \alpha, \varepsilon]_3,$$

where $[\kappa, \alpha, \varepsilon]_3$ denotes a sum of terms of the third and higher order in elements of κ , α and ε .

Let us now suppose that as is usual in practical computations, elements of κ , α and ε are small and products of elements of ε with those of κ or α may be neglected as compared with diagonal elements of ε . Then from (4.81) we have

$$(4.82) \quad \hat{\gamma} \approx \alpha + \varepsilon + \alpha^2 + \kappa\alpha + N\alpha \cdot \sigma(\hat{x}) \cdot \alpha.$$

Then we have

$$(4.83) \quad \hat{\gamma} - \alpha\hat{\gamma} - \alpha \approx \varepsilon + \kappa\alpha + N\alpha \cdot \sigma(\hat{x}) \cdot \alpha.$$

If elements of κ and α are so small that $\kappa\alpha$ and $N\alpha \cdot \sigma(\hat{x}) \cdot \alpha$ may be neglected as compared with diagonal elements of ε , then from (4.83) we have

$$\hat{\gamma} - \alpha\hat{\gamma} - \alpha \approx \varepsilon,$$

consequently from (4.54) we have

$$(4.84) \quad \sigma[\tilde{x}^{(n)} - \hat{x}] \leq \varepsilon \cdot \sigma(\hat{x})$$

approximately. This is of the same form as (4.48). This means that $\tilde{x}^{(n)}$ has the same degree of the accuracy as $\tilde{x}^{(n)}$ in an oscillatory state.

Let us consider the case where Newton's method or a generalized Newton's method having $H(x)$ close to $J^{-1}(x)$ is employed as is usual in practical computations and the computations are carried out accurately as is usual so that diagonal elements of ε may be all small. In such a case, as is stated in Remark 1 in 4.2, $\hat{\delta} \approx \varepsilon$, therefore, from the above results, taking into account Theorem 6, we get the following conclusion about the choice of the matrix $\alpha \geq 0$.

It is recommended to take a diagonal matrix $\alpha \geq 0$ so that its diagonal elements may be all large as compared with corresponding diagonal elements of ε but their mutual products may be all small as compared with diagonal elements of ε . If the matrix α is chosen in such a way, then one

can stop the iterative process by the criterion (3.36) using such α and moreover one can get an approximate solution having the same degree of the accuracy as the one in an oscillatory state, that is, an approximate solution with the accuracy of the evaluation of $f(x)$ in the iterative process (0.3).

By this result it will be seen how a generalized Newton's method is convenient in practical computations.

REMARK 2. Matrix $\hat{\eta}$ is determined uniquely by equality (4.55) provided all elements of matrices α and ε are sufficiently small. This can be proved likewise as for $\hat{\delta}$ in (4.21). The proof, therefore, is omitted in the present paper.

§5. A numerical example

We consider a system of equations

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

where p and q are unknowns. A solution to equations (5.1) gives a quadratic factor $x^2 - px + q$ of the 3rd-degree polynomial $P(x) = x^3 - ax^2 + bx - c$. In fact, if we put

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

then comparing the coefficients of both sides, we have

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

from which by elimination of l follows

$$A = -p^2 + ap + q - b, \quad B = pq - aq + c.$$

This shows that $x^2 - px + q$ is a factor if and only if $A = B = 0$, that is, p and q satisfy equations (5.1).

Let α , β , and γ be roots of the algebraic equation $P(x) = 0$, then it is clear that

$$(5.2) \quad a = \alpha + \beta + \gamma, \quad b = \alpha\beta + \beta\gamma + \gamma\alpha, \quad c = \alpha\beta\gamma.$$

If $x^2 - px + q$ is a factor of $P(x)$, then clearly it must be one of the products

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

therefore (p, q) must be one of the sets

$$(5.3) \quad (\alpha + \beta, \alpha\beta), (\beta + \gamma, \beta\gamma), (\gamma + \alpha, \gamma\alpha).$$

This means that solutions to equations (5.1) are given by (5.3) if a, b and c are given by (5.2).

We now suppose that

$$(5.4) \quad \alpha = 1, \beta = 10^{-3}, \gamma = 10.$$

Then by (5.2) we have

$$(5.5) \quad a = 11.001, b = 10.011, c = 0.01$$

and by (5.3) we have

$$(5.6) \quad (p, q) = (1.001, 0.001), (10.001, 0.01), (11, 10).$$

Hence we see that equations (5.1) have solutions given by (5.6) for a, b and c given by (5.5).

We have computed a solution $(1.001, 0.001)$ by applying Newton's method to the system of equations (5.1) starting from $(2, 0)$. The iterative process to be performed is

$$(5.7) \quad p_n = f(p_{n-1}, q_{n-1}), q_n = g(p_{n-1}, q_{n-1}) \quad (n = 1, 2, \dots),$$

where

$$(5.8) \quad \begin{cases} f(p, q) = p - [(p - a)\varphi(p, q) - \phi(p, q)]/D(p, q), \\ g(p, q) = q - [-q\varphi(p, q) - (2p - a)\phi(p, q)]/D(p, q) \end{cases}$$

and

$$(5.9) \quad D(p, q) = -(2p^2 - 3ap + q + a^2).$$

We have carried out computations on the 8 decimal floating-point system. The results are shown in Table 1, where $\tilde{p}_n$ and $\tilde{q}_n$ denote the values of p_n and q_n obtained in practical computations.

n	$\tilde{p}_n$	$\left \frac{\tilde{p}_n - \tilde{p}_{n-1}}{\tilde{p}_{n-1}} \right $	$\tilde{q}_n$	$\left \frac{\tilde{q}_n - \tilde{q}_{n-1}}{\tilde{q}_{n-1}} \right $
0	2.0		0.0	
1	0.8584 3290	0.5708	0.001110 9877	
2	0.9987 9570	0.1635	0.001139 6923	0.8974
3	1.000 9995	2.206×10^{-8}	0.001000 0307	0.1225
4	1.001 0001	5.994×10^{-7}	0.0009999 9998	3.072×10^{-5}
5	1.001 0000	9.990×10^{-8}	0.0009999 9998	0
6	1.001 0000	0	0.0009999 9998	0

Table 1

For $n \geq 3$,

$$1 \leq \tilde{p}_n \leq 1.002, 0.0009 \leq \tilde{q}_n \leq 0.0011.$$

Hence we consider the region D such that

$$(5.10) \quad D = \{(p, q) | 1 \leq p \leq 1.002, 0.0009 \leq q \leq 0.0011\}.$$

By elementary calculations, we then see that

$$(5.11) \quad \begin{cases} |\tilde{f}(p, q) - f(p, q)| < 2.231 \times 10^{-7} |f(p, q)|, \\ |\tilde{g}(p, q) - g(p, q)| < 2.477 \times 10^{-7} |g(p, q)| \end{cases}$$

for $(p, q) \in D$, where $\tilde{f}(p, q)$ and $\tilde{g}(p, q)$ denote respectively the values of $f(p, q)$ and $g(p, q)$ obtained by numerical computations on the 8 decimal floating-point system. Inequalities (5.11) show that (3.2) in 3.1 is valid for

$$(5.12) \quad \varepsilon = \begin{bmatrix} \varepsilon_1 & 0 \\ 0 & \varepsilon_2 \end{bmatrix} = \begin{bmatrix} 2.231 \times 10^{-7} & 0 \\ 0 & 2.477 \times 10^{-7} \end{bmatrix}.$$

As is seen from Table 1, $(\tilde{p}_n, \tilde{q}_n)$ in an oscillatory state is

$$(5.13) \quad (\tilde{p}_5, \tilde{q}_5) = (1.001 \ 0000, 0.0009999 \ 9998).$$

The exact value $(\hat{p}, \hat{q})$ of the solution under consideration is clearly

$$(5.14) \quad (\hat{p}, \hat{q}) = (1.001, 0.001).$$

Hence the relative errors of $\tilde{p}_5$ and $\tilde{q}_5$ are respectively

$$(5.15) \quad \begin{cases} |\tilde{p}_5 - \hat{p}|/|\hat{p}| = 0 \\ |\tilde{q}_5 - \hat{q}|/|\hat{q}| = 2 \times 10^{-8}. \end{cases}$$

By (5.12) this shows that not only (4.20) in Theorem 5 but also stronger inequality (4.48) in Remark 1 in 4.2 is valid in the present case.

Let us now investigate at which step the iterative process (5.7) is stopped when a criterion of the form (3.36) is employed.

First let us consider the case where, according to the conclusion got in Remark 1 in 4.3, a matrix

$$(5.16) \quad \alpha = \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{bmatrix}$$

is chosen so that

$$(5.17) \quad \begin{aligned} \alpha_{11} &= \alpha_{22} = 10^{-5}, \\ \alpha_{12} &= \alpha_{21} = 0. \end{aligned}$$

Since $\varepsilon_1, \varepsilon_2 \approx 2.5 \times 10^{-7}$ by (5.12), the above α_{11} and α_{22} are both large as compared with ε_1 and ε_2 and at the same time their products are all small as compared with ε_1 and ε_2 . If one makes use of a criterion (3.36) with α given above for stopping the iterative process, then as is seen from Table 1, the iterative process is stopped at $n=5$ and one gets $(\tilde{p}_5, \tilde{q}_5)$ in an oscillatory state. This shows the correctness of the conclusion stated in Remark 1 in 4.3.

If one chooses α so that

$$(5.18) \quad \alpha_{11} = \alpha_{22} = 10^{-4}, \alpha_{12} = \alpha_{21} = 0$$

instead of (5.17), then as is seen from Table 1 the iterative process is stopped at $n=4$ and one gets $(\tilde{p}_4, \tilde{q}_4)$. The relative errors of $\tilde{p}_4$ and $\tilde{q}_4$ are respectively

$$\begin{cases} |\tilde{p}_4 - \hat{p}|/|\hat{p}| \approx 0.9990 \times 10^{-7}, \\ |\tilde{q}_4 - \hat{q}|/|\hat{q}| = 2 \times 10^{-8}. \end{cases}$$

Referring to (5.12) and (5.15), $(\tilde{p}_4, \tilde{q}_4)$ then is regarded as having the same degree of the accuracy as $(\tilde{p}_n, \tilde{q}_n)$ in an oscillatory state. This shows again the correctness of the conclusion stated in Remark 1 in 4.3.

If one choose α so that

$$(5.19) \quad \alpha_{11}=\alpha_{22}=10^{-6} \text{ or } 10^{-7}, \alpha_{12}=\alpha_{21}=0,$$

then the iterative process is stopped at $n=5$.

If one chooses α so that

$$(5.20) \quad \alpha_{11}=\alpha_{22}=10^{-8}, \alpha_{12}=\alpha_{21}=0,$$

then the iterative process is stopped at $n=6$ and it turns out that the iterative process is repeated once more than needed. Such choice of very small α_{11} and α_{22} , however, should be refrained in practical computations, because it may lead to inefficient stopping of the iterative process as is seen above and more seriously it may lead to failure of stopping the iterative process as is mentioned in the beginning of 3.2.

APPENDIX

We shall show that matrix $\hat{\delta}$ is uniquely determined by (4.21) provided all diagonal elements of ε are sufficiently small.

Writting (4.21) briefly as

$$(A.1) \quad \delta = [E - (E + \varepsilon)(\kappa + N_1\delta)]^{-1}\varepsilon,$$

we shall prove that equation (A.1) has a unique solution δ such that

$$(A.2) \quad \delta \cdot \varepsilon^{-1} \rightarrow (E - \kappa)^{-1} \text{ as } \varepsilon \rightarrow 0.$$

Since $0 \leq \kappa \leq K_0$ by (4.13) and (4.14), due to (4.15) by Corollary 1 we have

$$(A.3) \quad \rho(\kappa) < 1.$$

Let κ' be Jordan's canonical form of κ such that

$$(A.4) \quad \kappa' = S^{-1}\kappa S = \Sigma \oplus \begin{bmatrix} \lambda & 0 & \cdots & 0 \\ \gamma & \lambda & \ddots & \vdots \\ 0 & \gamma & \lambda & \ddots \\ \vdots & \ddots & \ddots & \gamma & 0 \\ 0 & & 0 & \gamma & \lambda \end{bmatrix}$$

where λ is an eigenvalue of κ and γ is an arbitrary positive number. By taking γ sufficiently small, we may suppose that

$$(A.5) \quad \|\kappa'\| < 1,$$

where $\|\cdot\|$ denotes a norm of a matrix such that

$$\|A\| = \max_i \sum_j |a_{ij}|$$

for a matrix A with elements a_{ij} . Put

$$(A.6) \quad S^{-1}\delta S = \delta', \quad S^{-1}N_1 S = N'_1, \quad S^{-1}\varepsilon S = \varepsilon'.$$

Then equation (A.1) is equivalent to the equation

$$(A.7) \quad \delta' = [E - (E + \varepsilon')(\kappa' + N'_1 \delta')]^{-1} \varepsilon'$$

and to a solution to (A.1) satisfying (A.2) corresponds a solution δ' to (A.7) such that

$$(A.8) \quad \delta' \cdot \varepsilon'^{-1} \rightarrow (E - \kappa')^{-1} \text{ as } \varepsilon' \rightarrow 0.$$

Now for κ' satisfying (A.5) it is well known that

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

from which readily follows

$$(A.9) \quad \|(E - \kappa')^{-1}\| \leq \sum_{n=0}^{\infty} \|\kappa'\|^n = (1 - \|\kappa'\|)^{-1}.$$

Let l_1 be a positive number such that

$$(A.10) \quad l_1 > (1 - \|\kappa'\|)^{-1},$$

then for δ' satisfying (A.8) we have

$$(A.11) \quad \|\delta'\| \leq l_1 \|\varepsilon'\|$$

provided $\|\varepsilon'\|$ is sufficiently small. Hence we see that for our purpose it is sufficient to prove that equation (A.7) has a unique solution δ' satisfying (A. 11) for sufficiently small $\|\varepsilon'\|$.

Let us write (A.7) briefly as

$$(A.12) \quad \delta' = F(\delta'),$$

where

$$(A.13) \quad F(\delta') = [E - (E + \varepsilon')(\kappa' + N_1' \delta')]^{-1} \varepsilon'.$$

Let $\mathcal{Q}$ be the set such that

$$(A.14) \quad \mathcal{Q} = \{\delta' \mid \|\delta'\| \leq l_1 \|\varepsilon'\|\}$$

and l_2 be a positive number such that

$$(A.15) \quad \|\kappa'\| < l_2 < 1.$$

Then for any $\delta' \in \mathcal{Q}$ we have

$$\begin{aligned} \|(E + \varepsilon')(\kappa' + N_1' \delta')\| &\leq (1 + \|\varepsilon'\|)(\|\kappa'\| + l_1 \|N_1'\| \cdot \|\varepsilon'\|) \\ &\leq l_2 < 1 \end{aligned}$$

whenever $\|\varepsilon'\| < \zeta_1$ for sufficiently small $\zeta_1 > 0$. Then we have

$$\begin{aligned} &\|[E - (E + \varepsilon')(\kappa' + N_1' \delta')]^{-1}\| \\ &\leq [1 - (1 + \|\varepsilon'\|)(\|\kappa'\| + l_1 \|N_1'\| \cdot \|\varepsilon'\|)]^{-1} \end{aligned}$$

for any $\|\varepsilon'\| < \zeta_1$. However the right-hand side of the above inequality tends to $(1 - \|\kappa'\|)^{-1}$ as $\|\varepsilon'\| \rightarrow 0$. Therefore by (A.10) we see that

$$(A.16) \quad \|[E - (E + \varepsilon')(\kappa' + N_1' \delta')]^{-1}\| \leq l_1$$

whenever $\|\varepsilon'\| < \zeta_2$ for sufficiently small positive number $\zeta_2 \leq \zeta_1$. By

(A.13) and (A.14), this means that

$$(A.17) \quad F(\mathcal{Q}) \subset \mathcal{Q}$$

whenever $\|\varepsilon'\| < \zeta_2$.

Next for any $\delta'_1, \delta'_2 \in \mathcal{Q}$, from (A.13) we have

$$\begin{aligned} F(\delta'_1) - F(\delta'_2) &= [E - (E + \varepsilon')(\kappa' + N'_1 \delta'_1)]^{-1} (E + \varepsilon') N'_1 (\delta'_1 - \delta'_2) \\ &\quad \times [E - (E + \varepsilon')(\kappa' + N'_1 \delta'_2)]^{-1} \varepsilon'. \end{aligned}$$

Hence by (A.16) we have

$$(A.18) \quad \|F(\delta'_1) - F(\delta'_2)\| \leq l_1^2 (1 + \|\varepsilon'\|) \|N'_1\| \cdot \|\varepsilon'\| \cdot \|\delta'_1 - \delta'_2\|$$

for $\|\varepsilon'\| < \zeta_2$. Let k be an arbitrary number such that

$$(A.19) \quad 0 < k < 1.$$

Then

$$l_1^2 (1 + \|\varepsilon'\|) \|N'_1\| \cdot \|\varepsilon'\| \leq k$$

whenever $\|\varepsilon'\| < \zeta_3$ for sufficiently small positive number $\zeta_3 \leq \zeta_2$. By (A.18) this implies

$$(A.20) \quad \|F(\delta'_1) - F(\delta'_2)\| \leq k \cdot \|\delta'_1 - \delta'_2\|$$

for any $\delta'_1, \delta'_2 \in \mathcal{Q}$ provided $\|\varepsilon'\| < \zeta_3$.

From (A.17), (A.19) and (A.20), by the well-known principle of contractive mapping, we see that equation (A.12) has a unique solution in $\mathcal{Q}$ for $\|\varepsilon'\| < \zeta_3$. This proves the desired result.

References

- [1] Collatz, L.: Einige Anwendungen funktionalanalytischer Methoden in der praktischen Analysis, Z. Angew. Math. Phys., 4 (1953), 327-357.
- [2] Urabe, M.: Convergence of numerical iteration in solution of equations, J. Sci. Hiroshima Univ., Ser. A, 19 (1956), 479-489.
- [3] Urabe, M.: Error estimation in numerical solution of equations by iteration process, J. Sci. Hiroshima Univ., Ser. A-I, 26 (1962), 77-91.