J. Sci. Hiroshima Univ. Ser. A-I
26 (1962), 77-91

Error Estimation in Numerical Solution of Equations by Iteration Process

Minoru Urabe

(Received August 30, 1962)

1. Introduction

Let R be a linear normed space and F be a complete subset of R. Let f be a functional defined on F such that $f(F) \subset R$.

We assume that

(i)

$$\| f(x') - f(x'') \| \leqq K_0 \| x' - x'' \| \tag{1.1}$$

for any $x', x'' \in F$, *where*

$$0 < K_0 < 1; \tag{1.2}$$

(ii)

$$\| f^*(x) - f(x) \| \leqq \varepsilon \tag{1.3}$$

for any $x \in F$, *where* $f^*(x)$ *is a numerical valuation of* $f(x)$ *in actual computation with the error bound* ε (>0) *such that* $f^*(F) \subset R$ (*here, by a numerical valuation in actual computation, we mean a valuation by a set of finite numbers of the numbers rounded to a certain fixed number of decimal digits*);

(iii) *for a certain numerical value* $x_0 \in F$,

$$\sum \left\{ h : \| h - x_1^* \| \leqq \frac{K_0}{1-K_0} \| x_1^* - x_0 \| + 2\delta_0 \right\} \subset F, \tag{1.4}$$

where $x_1^* = f^*(x_0)$ *and*

$$\delta_0 = \frac{\varepsilon}{1-K_0}. \tag{1.5}$$

Then, by the author's previous paper [1],

(i) *the equation*

$$x = f(x) \tag{1.6}$$

has one and only one solution in F;

(ii) *the unique solution* $\bar{x}$ *of* (1.6) *is obtained by the ideal iteration process*

$$x_{n+1} = f(x_n) \qquad (n = 0, 1, 2, \cdots)$$

as follows:

$$\bar{x} = \lim_{n \to \infty} x_n,$$

where $\bar{x}, x_n \in \sum$ $(n = 1, 2, \cdots)$;

(iii) *the actual iteration process*

$$x_{n+1}^* = f^*(x_n^*) \qquad (n = 0, 1, 2, \cdots; x_0^* = x_0) \tag{1.7}$$

can be continued indefinitely so that

$$x_n^* \in \sum \qquad (n = 1, 2, \cdots); \tag{1.8}$$

(iv) *the iteration process* (1.7) *ends after finite numbers of repetitions in the state of numerical convergence, in which the sequence* $\{x_n^*\}$ *oscillates taking a certain finite number of values* (in the sequel, this state of numerical convergence is called *the state of oscillatory numerical convergence* and this is abbreviated as "*the state ONC*");

(v) *for any* x_n^* *in the state ONC,*

$$\| x_n^* - \bar{x} \| \leqq \delta_0. \tag{1.9}$$

The approximate solutions of the equation (1.6) *are given by any one of* x_n^* *in the state ONC and their error estimates are given by the inequality* (1.9).

In this note, we replace the assumption (i) by a little more general one that

(i′)

$$\| f(x') - f(x'') \| \leqq K(x', x'') \| x' - x'' \| \tag{1.10}$$

for any $x', x'' \in F$, where

$$0 \leqq K(x', x'') \leqq K_0 < 1. \tag{1.11}$$

This assumption is satisfied for instance by the Newton method.

One of the purpose of this note is to obtain the error estimates more precise than (1.9) under the assumption (i′).

In computation by an electronic computer, however, it is not convenient to continue the iteration process up to the state ONC, because this requires that all x_n^*'s computed successively should be stored till the oscillation is detected. Hence, in computation by an electronic computer, the iteration is stopped often by the criterion of the form

$$\| x_{n+1}^* - x_n^* \| < \alpha, \tag{1.12}$$

where α is a certain positive number. In the present note, under the assump-

tion (i'), there is derived a condition upon α so that the criterion (1.12) may be actually effective for stopping the iteration process, and further there is given an error estimate for a solution obtained by the iteration process stopped by the criterion of the form (1.12).

Lastly, in illustration, our theory is applied to the Newton method to solve a system of equations and a numerical example is cited.

2. Error estimation of x_n^* in the state ONC

Let $x_n^*, x_{n+1}^*, \cdots, x_{n+m-1}^*$ be in the state ONC and assume $x_{n+m}^* = x_n^*$. Then, by [1],

$$\| x_{n+i}^* - \bar{x} \| \leqq \delta_0 \qquad (i = 0, 1, 2, \cdots). \tag{2.1}$$

Let V_1 be the sphere $V_1\{x : \| x - \bar{x} \| \leqq \delta_0\}$, then, by [1], $V_1 \subset \sum$. The inequality (2.1) implies

$$x_{n+i}^* \in V_1 \qquad (i = 0, 1, 2, \cdots).$$

Put

$$K_1 = \underset{x \in V_1}{\text{l.u.b.}}\, K(\bar{x}, x), \tag{2.2}$$

then it is evident that

$$0 \leqq K_1 \leqq K_0 < 1. \tag{2.3}$$

Now, in V_1, it holds that

$$\begin{aligned} \| x_{n+1}^* - \bar{x} \| &\leqq \| f^*(x_n^*) - f(\bar{x}) \| \\ &\leqq \| f^*(x_n^*) - f(x_n^*) \| + \| f(x_n^*) - f(\bar{x}) \| \\ &\leqq \varepsilon + K_1 \| x_n^* - \bar{x} \|, \end{aligned}$$

$$\| x_{n+2}^* - \bar{x} \| \leqq \varepsilon + K_1 \| x_{n+1}^* - \bar{x} \|,$$

$$\vdots$$

$$\| x_{n+m}^* - \bar{x} \| \leqq \varepsilon + K_1 \| x_{n+m-1}^* - \bar{x} \|.$$

Therefore it follows that

$$\| x_{n+m}^* - \bar{x} \| \leqq \varepsilon(1 + K_1 + \cdots + K_1^{m-1}) + K_1^m \| x_n^* - \bar{x} \|. \tag{2.4}$$

Since $x_{n+m}^* = x_n^*$ by the assumption, from (2.4) follows readily

$$\| x_n^* - \bar{x} \| \leqq \frac{\varepsilon}{1 - K_1}.$$

As is readily seen from its derivation, x_n^* in the above inequality can be re-

placed by any x_{n+i}^* in the state ONC, consequently we have

$$\|x_{n+i}^* - \bar{x}\| \leqq \delta_1, \tag{2.5}$$

where

$$\delta_1 = \frac{\varepsilon}{1-K_1} \leqq \delta_0. \tag{2.6}$$

If K_1 defined by (2.2) is equal to K_0, $\delta_1 = \delta_0$ and the error estimate (2.5) is the same as the initial estimate (2.1). But, if $K_1 < K_0$, then $\delta_1 < \delta_0$. In this case, (2.5) gives an error estimate more precise than the initial estimate (2.1).

In this latter case, let us repeat our process. Namely let us consider the sphere $V_2\{x: \|x-\bar{x}\| \leqq \delta_1\}$. Then, by (2.5) and (2.6),

$$x_{n+i}^* \in V_2 \qquad (i = 0, 1, 2, \cdots) \tag{2.7}$$

and $V_2 \subset V_1$. Put

$$K_2 = \underset{x \in V_2}{\text{l.u.b.}}\, K(\bar{x}, x),$$

then it is evident from (2.2) and (2.3) that

$$0 \leqq K_2 \leqq K_1 < K_0 < 1.$$

Also, by (2.7), in like manner as (2.5), we have

$$\|x_{n+i}^* - \bar{x}\| \leqq \delta_2$$

where

$$\delta_2 = \frac{\varepsilon}{1-K_2} \leqq \delta_1 < \delta_0.$$

This process can be repeated again and again.

Thus, if we put

$$\begin{cases} V_p = V_p\{x: \|x - \bar{x}\| \leqq \delta_{p-1}\}, \\ K_p = \underset{x \in V_p}{\text{l.u.b.}}\, K(\bar{x}, x), \\ \delta_p = \dfrac{\varepsilon}{1-K_p}, \\ \qquad\qquad (p = 1, 2, 3, \cdots), \end{cases} \tag{2.8}$$

then, either

$$\begin{cases} \Sigma \supseteq V_1 \supset V_2 \supset \cdots \supset V_p \supset V_{p+1} = V_{p+2}, \\ 1 > K_0 > K_1 > K_2 > \cdots > K_p = K_{p+1} \geqq 0, \\ \delta_0 > \delta_1 > \delta_2 > \cdots > \delta_p = \delta_{p+1} \geqq \varepsilon, \end{cases} \tag{2.9}$$

or

$$\text{(2.10)}\qquad \begin{cases} \sum \supseteq V_1 \supset V_2 \supset \cdots \supset V_p \supset V_{p+1} \supset \cdots, \\ 1 > K_0 > K_1 > K_2 > \cdots > K_p > K_{p+1} > \cdots \geqq 0, \\ \delta_0 > \delta_1 > \delta_2 > \cdots > \delta_p > \delta_{p+1} > \cdots \geqq \varepsilon. \end{cases}$$

Here, in (2.10), $K_{p+1} < K_p$ for any positive integer p.

In the case of (2.9),

$$\text{(2.11)}\qquad x_{n+i}^{*} \in V_{p+1} \qquad (i = 0, 1, 2, \cdots)$$

and the desired error estimate is given by

$$\text{(2.12)}\qquad \| x_{n+i}^{*} - \bar{x} \| \leqq \delta_p.$$

In the case of (2.10), there exist

$$\text{(2.13)}\qquad K'_\omega = \lim_{p \to \infty} K_p \quad \text{and} \quad \delta'_\omega = \lim_{p \to \infty} \delta_p = \frac{\varepsilon}{1 - K'_\omega}.$$

Consider the sphere $V_\omega \{x : \| x - \bar{x} \| \leqq \delta'_\omega\}$, then evidently

$$\text{(2.14)}\qquad x_{n+i}^{*} \in V_\omega \qquad (i = 0, 1, 2, \cdots)$$

and

$$\text{(2.15)}\qquad K_\omega \overset{\text{def}}{=} \underset{x \in V_\omega}{\text{l.u.b.}}\, K(\bar{x}, x) \leqq K'_\omega.$$

If $K_\omega = K'_\omega$, then, likewise as the case $K_{p+1} = K_p$, the desired error estimate is given by

$$\text{(2.16)}\qquad \| x_{n+i}^{*} - \bar{x} \| \leqq \delta'_\omega.$$

If $K_\omega < K'_\omega$, then the same process as (2.8) can be carried out starting from V_ω instead of V_1 and we obtain the sequence of the form of either (2.9) or (2.10). When the sequence of the form of (2.10) is obtained, the above process is repeated.

In such a way, the process is continued till the relation of the form

$$\underset{x \in V_{p+1}}{\text{l.u.b.}}\, K(\bar{x}, x) = K_p \qquad (p \text{ is a transfinite number})$$

happens to hold. *Let $\hat{V}$, $\hat{K}$ and $\hat{\delta}$ be the transfinite limits of V_p, K_p and δ_p for which there holds*

$$\text{(2.17)}\qquad \underset{x \in \hat{V}}{\text{l.u.b.}}\, K(\bar{x}, x) = \hat{K},$$

where

$$\text{(2.18)}\qquad \begin{cases} \hat{V} = \hat{V}\{x : \| x - \bar{x} \| \leqq \hat{\delta}\}, \\ \hat{K} = \text{transfinite limit of } K_p, \\ \hat{\delta} = \dfrac{\varepsilon}{1 - \hat{K}}. \end{cases}$$

Then

$$\|x_{n+i}^* - \bar{x}\| \leqq \hat{\delta} \qquad (i = 0, 1, 2, \cdots) \tag{2.19}$$

and this is the desired error estimate for x_{n+i}^* $(i=0, 1, 2, \cdots)$, *namely the desired error estimate for the approximate solutions of the equation* (1.6) *given by the* x_n^**'s in the state ONC.*

Remark 1. In some cases, it may be impossible to obtain the transfinite limit $\hat{\delta}$ by actual calculation. In such a case, we have to stop our process half way without completing the necessary steps, but, even in such a case, we can obtain the error estimate of the form (2.12) which is more precise than the initial estimate (2.1).

Remark 2. The quantities K_p $(p=1, 2, \cdots)$ can be replaced by

$$K_p = \text{l.u.b.}_{x', x'' \in V_p} K(x', x'')$$

or some other similar quantities. Of course, in these cases, the regions V_p and the quantities δ_p must be replaced by the corresponding ones defined by the rule (2.8).

3. The condition for the criterion of the form (1.12) used for stopping the iteration process

As is stated in §1, the iteration process (1.7) ends in the state ONC after finite numbers of repetitions. But, by §2, the x_n^*'s in the state ONC lie all in $\hat{V}$, consequently

$$\|x_{n+1}^* - x_n^*\| \leqq 2\hat{\delta} \tag{3.1}$$

for any x_{n+1}^*, x_n^* in the state ONC.

Then, if $\alpha > 2\hat{\delta}$, there exist only a finite number of x_n^*'s which do not satisfy the criterion (1.12). This implies that, if $\alpha > 2\hat{\delta}$, we can really stop the iteration process (1.7) after finite numbers of repetitions.

If $\alpha \leqq 2\hat{\delta}$, there may not exist any x_n^* satisfying the criterion (1.12), for it may happen that

$$\|x_{n+1}^* - x_n^*\| = 2\hat{\delta}$$

for any x_{n+1}^*, x_n^* in the state ONC. In such a case, we can not stop the iteration process (1.7) in a finite number of repetitions.

Thus, *in order to stop the iteration process* (1.7) *by the criterion of the form* (1.12), *we should choose* α *so that*

$$\alpha > 2\hat{\delta}. \tag{3.2}$$

When $\hat{\delta}$ *can not be found by actual calculation, we should have to replace* $\hat{\delta}$

by some δ_p $(>\hat{\delta})$ *obtained in the midst of the process described in* §2.

4. Error estimation in case the iteration process (1.7) is stopped by the criterion of the form (1.12)

Let

$$\|x_{n+1}^* - x_n^*\| < \alpha. \tag{4.1}$$

Then it follows from (1.3), (1.10) and (4.1) that

$$\begin{aligned}\|x_n^* - \bar{x}\| &= \|x_n^* - f(\bar{x})\| \\ &\leqq \|x_n^* - x_{n+1}^*\| + \|f^*(x_n^*) - f(x_n^*)\| + \|f(x_n^*) - f(\bar{x})\| \\ &< \alpha + \varepsilon + K(\bar{x}, x_n^*)\|x_n^* - \bar{x}\|.\end{aligned}$$

Consequently we have

$$\|x_n^* - \bar{x}\| < \frac{\varepsilon+\alpha}{1-K(\bar{x}, x_n^*)}. \tag{4.2}$$

As in §2, let us put

$$\eta_0 = \frac{\varepsilon+\alpha}{1-L_0} \qquad (L_0 = K_0); \tag{4.3}$$

$$\left\{\begin{aligned} &W_p = W_p\{x: \|x - \bar{x}\| \leqq \eta_{p-1}\}, \\ &L_p = \text{l.u.b.}_{x\in W_p} K(\bar{x}, x), \\ &\eta_p = \frac{\varepsilon+\alpha}{1-L_p} \end{aligned}\right. \tag{4.4}$$

$$(p = 1, 2, 3, \cdots).$$

Then the same reasonings as in §2 prevail for the error estimate of x_n^* and there are obtained the conclusions as follows:

Let $\hat{W}$, $\hat{L}$ and $\hat{\eta}$ be the transfinite limits of W_p, L_p and η_p for which there holds

$$\text{l.u.b.}_{x\in\hat{W}} K(\bar{x}, x) = \hat{L}, \tag{4.5}$$

where

$$\left\{\begin{aligned} &\hat{W} = \hat{W}\{x: \|x - \bar{x}\| \leqq \hat{\eta}\}, \\ &\hat{L} = \text{transfinite limit of } L_p, \\ &\hat{\eta} = \frac{\varepsilon+\alpha}{1-\hat{L}}. \end{aligned}\right. \tag{4.6}$$

Then

$$\|x_n^* - \bar{x}\| \leqq \hat{\eta} \tag{4.7}$$

for x_n^* for which the criterion (4.1) is fulfilled.

By the way,

$$\|x_{n+1}^* - \bar{x}\| \leqq \|f^*(x_n^*) - f(x_n^*)\| + \|f(x_n^*) - f(\bar{x})\|$$
$$\leqq \varepsilon + K(\bar{x}, x_n^*)\|x_n^* - \bar{x}\|.$$

Then, since $x_n^* \in \hat{W}$, it follows from (4.5), (4.6) and (4.7) that

$$\|x_{n+1}^* - \bar{x}\| \leqq \varepsilon + \hat{L}\hat{\eta} = \frac{\varepsilon}{1-\hat{L}} + \frac{\hat{L}}{1-\hat{L}}\alpha. \tag{4.8}$$

This is the desired error estimate for x_{n+1}^* which is an approximate solution of the equation (1.6) *obtained by the iteration process* (1.7) *in case the iteration is stopped by the criterion* (4.1).

Remark 1. To the transfinite limits and the quantities L_p described above, the same remarks as in §2 can be made.

Remark 2. When the iteration process is stopped by the criterion of the form (1.12), the quantity $\|x_{n+1}^* - x_n^*\|$ may be far smaller than α itself used for stopping the iteration process. In such a case, it is needless to say that any quantity α_0 not smaller than $\|x_{n+1}^* - x_n^*\|$ can be used for estimation of the error of x_{n+1}^* instead of α in (4.8).

5. Error estimates for the Newton method

The Newton method to solve a system of equations

$$\varphi_i(x_1, x_2, \cdots, x_m) = 0 \qquad (i = 1, 2, \cdots, m)$$

or

$$\varphi(x) = 0 \tag{5.1}$$

in vector form, is nothing but an iteration process (1.7) where

$$f(x) = x - H(x)\varphi(x). \tag{5.2}$$

Here $H(x)=(H_{ij}(x))$ $(i, j=1, 2, \cdots, m)$ is an inverse matrix of the Jacobian matrix $J(x)=(J_{ij}(x))$ $(i, j=1, 2, \cdots, m)$ of $\varphi(x)$ with respect to x.

Let $\bar{x}$ be a solution of (5.1) and assume $\det J(\bar{x}) \neq 0$. Let F be a closed domain containing $\bar{x}$ such that $\varphi(x)$ is defined on F.

Let us assume $\varphi(x) \in C_x^3[F]$. Then, if F is sufficiently small and ε in (1.3) is chosen sufficiently small, all the conditions (i)-(iii) of §1 turn out to be fulfilled provided x_0 is taken sufficiently near $\bar{x}$. Then the results obtained in the preceding paragraphs are all valid for the present iteration process, namely for the Newton method.

From (5.2), it readily follows that

$$f_i(x) - f_i(\bar{x}) = \sum_{j,k=1}^{m} \psi_{ijk}(\xi_i)(x_j - \bar{x}_j)(x_k - \bar{x}_k) \tag{5.3}$$

for any $x \in F$, where

$$\xi_i = \bar{x} + \theta_i(x - \bar{x}) \qquad (0 < \theta_i < 1) \quad (i = 1, 2, \cdots, m) \tag{5.4}$$

and

$$\psi_{ijk}(x) = -\frac{1}{2}\sum_{l=1}^{m}\left[\frac{\partial^2 H_{il}(x)}{\partial x_j \partial x_k}\varphi_l(x) + \frac{\partial H_{il}(x)}{\partial x_j}J_{lk}(x)\right] \quad (i, j, k = 1, 2, \cdots, m). \tag{5.5}$$

Let us define the norm of a vector $x=(x_i)$ $(i=1, 2, \cdots, m)$ by

$$\|x\| = \max_i |x_i|$$

and put

$$M = \max_{x \in F} \max_i \sum_{j,k=1}^{m} |\psi_{ijk}(x)|, \tag{5.6}$$

then, by (5.3), we may suppose that

$$K(\bar{x}, x) = M\|x - \bar{x}\| \tag{5.7}$$

for any $x \in F$.

Then, by (2.8), we have

$$\begin{cases} K_p = M\delta_{p-1}, \\ \delta_p = \dfrac{\varepsilon}{1 - K_p}, \end{cases} \qquad (p = 1, 2, 3, \cdots). \tag{5.8}$$

Then evidently it holds that

$$\delta_p = \frac{\varepsilon}{1 - M\delta_{p-1}} \qquad (p = 1, 2, 3, \cdots).$$

Hence, in the limit, we have

$$\hat{\delta} = \frac{\varepsilon}{1 - M\hat{\delta}},$$

which can be solved as follows:

$$\hat{\delta} = \frac{1}{2M}[1 - (1 - 4\varepsilon M)^{1/2}] = \varepsilon + \varepsilon^2 M + o(\varepsilon^2). \tag{5.9}$$

This is a desired error estimate for the approximate solutions obtained by the Newton method in case the iteration process is carried out up to the state ONC.

The fact $\hat{\delta} \approx \varepsilon$ expresses that

(i) *the Newton method is very good,*

(ii) *the error estimate $\hat{\delta}$ is very good,*

for, in computation of a solution, we can not exceed the precision of the error bound ε with which the computation is carried out.

In case the iteration process is stopped by the criterion of the form (1.12), by (4.4), we have

$$\begin{cases} L_p = M\eta_{p-1}, \\ \eta_p = \dfrac{\varepsilon+\alpha}{1-L_p}, \end{cases} \qquad (p = 1, 2, 3, \cdots). \tag{5.10}$$

Then evidently it holds that

$$L_{p+1} = \frac{M(\varepsilon+\alpha)}{1-L_p} \qquad (p = 1, 2, 3, \cdots).$$

Hence, in the limit, we have

$$\hat{L} = \frac{M(\varepsilon+\alpha)}{1-\hat{L}},$$

which can be solved as follows:

$$\begin{aligned} \hat{L} &= \frac{1}{2}[1 - \{1 - 4M(\varepsilon+\alpha)\}^{1/2}] \\ &= M(\varepsilon+\alpha) + M^2(\varepsilon+\alpha)^2 + o\{(\varepsilon+\alpha)^2\}. \end{aligned} \tag{5.11}$$

Substituting this into (4.8), we have the error estimate as follows:

$$\begin{aligned} \|x_{n+1}^* - \bar{x}\| &\leqq \frac{2\varepsilon}{1+\sqrt{1-4M(\varepsilon+\alpha)}} + \frac{1-\sqrt{1-4M(\varepsilon+\alpha)}}{1+\sqrt{1-4M(\varepsilon+\alpha)}} \cdot \alpha \\ &= \varepsilon + M(\varepsilon+\alpha)^2 + o\{(\varepsilon+\alpha)^2\}. \end{aligned} \tag{5.12}$$

This is the desired error estimate for the approximate solution x_{n+1}^* obtained by the Newton method in case the iteration process is stopped by the criterion of the form (1.12).

The inequality (5.12) expresses that the error bound of x_{n+1}^* is nearly equal to ε provided α is chosen so that $\alpha = O(\varepsilon)$. Such choice of α is always possible since the condition on α is only $\alpha > 2\hat{\delta} \approx 2\varepsilon$. The fact that the error bound of x_{n+1}^* is nearly equal to ε, expresses, by the same reason as is remarked concerning the error bound $\hat{\delta}$, that

(i) *the Newton method is very good, because the x_{n+1}^* gives a sufficiently*

accurate approximate root (with the same accuracy as the values in the state ONC) *of the given equation even if the iteration process is stopped half way,*

(ii) *the error estimate* (5.12) *is very good.*

From the above mentioned, it is needless to say that, *if we want to get a solution as accurate as possible for the available computer, we should choose* α *so that* $\alpha = O(\varepsilon)$.

But, in actual computation, sometimes we do not need so accurate a solution but only an approximate solution having a certain accuracy. In such a case, as is shown below, we may choose α so that $\alpha \gg \varepsilon$.

Indeed, if $\alpha \gg \varepsilon$, then, by (5.12), the error bound of x_{n+1}^* is nearly equal to

$$\varepsilon + M\alpha^2.$$

This is either $O(\varepsilon)$ or $O(\alpha^2)$ according as $\alpha = O(\varepsilon^{1/2})$ or $\alpha \gg \varepsilon^{1/2}$. Namely, when $\alpha = O(\varepsilon^{1/2})$, the error bound of x_{n+1}^* is of the order of ε, namely of the same order as in the case where $\alpha = O(\varepsilon)$, and further, even if $\alpha \gg \varepsilon^{1/2}$, the error bound of x_{n+1}^* is of the order of α^2, namely of the second order of α. These facts say that *the Newton method gives a considerably accurate solution even if the iteration process is stopped by the criterion of the form* (1.12) *for* α *such that* $\alpha \gg \varepsilon$.

Example. To compute $\sqrt{0.1}$ by the Newton method.

The problem is to find a positive root of the equation

$$\varphi(x) = x^2 - 0.1 = 0 \tag{5.13}$$

by the iteration process

$$x_{n+1}^* = f^*(x_n^*) \qquad (n = 0, 1, 2, \cdots), \tag{5.14}$$

where

$$f(x) = x - \frac{x^2 - 0.1}{2x} = \frac{x^2 + 0.1}{2x} \tag{5.15}$$

and

$$|f^*(x) - f(x)| \leqq \varepsilon. \tag{5.16}$$

Let us assume that the computation in the iteration process is always carried out to eight decimal places (this means the results in each step of computation are always rounded to eight decimal places).

Let us start the computation from $x_0 = 0.4$. Let F be the closed interval $[0.2, 0.4]$. Then, as is readily seen from the form of (5.15), the inequality (5.16) is valid for

$$\varepsilon = \frac{1}{2}\left(1 + \frac{1}{2 \times 0.2}\right) \times 10^{-8} = \frac{7}{4} \times 10^{-8} \tag{5.17}$$

whenever $x \in F$.

Since

$$f'(x) = \frac{\varphi(x)\varphi''(x)}{\varphi'^2(x)} = \frac{1}{2}\left(1 - \frac{0.1}{x^2}\right),$$

it is evident that

$$|f'(x)| \leqq \frac{3}{4}$$

for $x \in F$. Therefore we see that the condition (i) is valid for

$$K_0 = \frac{3}{4}. \tag{5.18}$$

Then, by (1.5), we see that

$$\delta_0 = 7 \times 10^{-8}. \tag{5.19}$$

The actual computation shows

$$\begin{cases} x_1^* = 0.3250\ 0000, \\ x_2^* = 0.3163\ 4615. \end{cases} \tag{5.20}$$

Then, if we suppose the iteration is started from $x_1^* = 0.3250\ 0000$ rather than $x_0 = x_0^* = 0.4$, by (5.18), (5.19) and (5.20), we see that the condition (iii) is valid for the iteration process (5.14) in $F[0.2, 0.4]$, because

$$\frac{\frac{3}{4}}{1-\frac{3}{4}} \times 0.0087 + 14 \times 10^{-8} < 0.083.$$

Thus we may suppose that, for the iteration process (5.14), all the conditions (i)-(iii) are fulfilled in the interval $F[0.2, 0.4]$.

In the present example, substituting (5.13) into (5.5), we have

$$\psi(x) = \frac{0.05}{x^3},$$

from which, by (5.6), follows

$$M = 25/4. \tag{5.21}$$

Then, by (5.9), we have

$$\begin{aligned} \hat{\delta} &= \frac{2}{25}\left[1 - \left(1 - \frac{175}{4} \times 10^{-8}\right)^{1/2}\right] \\ &= \frac{7}{4} \times 10^{-8} + \frac{1225}{64} \times 10^{-16} + \cdots \\ &\approx \frac{7}{4} \times 10^{-8}. \end{aligned} \tag{5.22}$$

This is the error estimate for x_n^* in the state ONC.

If we want to stop the iteration process (5.14) by the criterion of the form (1.12), then, by the theory of §3, the number α must be chosen so that

$$\alpha > 2\hat{\delta} \approx \frac{7}{2} \times 10^{-8}.$$

Let us choose α so that

$$\alpha = 4 \times 10^{-8}. \tag{5.23}$$

Then, substituting (5.17), (5.21) and (5.23) into (5.12), we have

$$\begin{aligned} &|x_{n+1}^* - \bar{x}| \\ &\leq \frac{\frac{7}{2} \times 10^{-8}}{1+\sqrt{1-\frac{575}{4} \times 10^{-8}}} + \frac{1-\sqrt{1-\frac{575}{4} \times 10^{-8}}}{1+\sqrt{1-\frac{575}{4} \times 10^{-8}}} \cdot 4 \times 10^{-8} \\ &= \frac{7}{4} \times 10^{-8} + \frac{13225}{64} \times 10^{-16} + \cdots \\ &\approx \frac{7}{4} \times 10^{-8}. \end{aligned} \tag{5.24}$$

This is the error estimate for x_{n+1}^* which is obtained by stopping the iteration process by the criterion

$$|x_{n+1}^* - x_n^*| < 4 \times 10^{-8}. \tag{5.25}$$

The values obtained by the iteration process (5.14) in the specified manner are as follows:

$$\begin{cases} x_0 = x_0^* = 0.4000\ 0000, \\ x_1^* = 0.3250\ 0000, \\ x_2^* = 0.3163\ 4615, \\ x_3^* = 0.3162\ 2779, \\ x_4^* = 0.3162\ 2777, \\ x_5^* = 0.3162\ 2776, \\ x_6^* = 0.3162\ 2777 = x_4^*. \end{cases} \tag{5.26}$$

The true value $\bar{x} = \sqrt{0.1}$ is

$$\bar{x} = 0.3162\ 2776\ 6016\cdots.$$

Thus the values in the state ONC are

$$x_4^* = 0.3162\ 2777 \quad \text{and} \quad x_5^* = 0.3162\ 2776,$$

and their true errors are respectively

$$+0.3983\cdots\times 10^{-8} \quad \text{and} \quad -0.6016\cdots\times 10^{-8}.$$

The error estimate given by (5.22) is

$$\hat{\delta}\approx 1.75\times 10^{-8}.$$

This is far more precise than the initial error estimate

$$\delta_0 = 7\times 10^{-8}$$

given by (5.19).

In case the iteration process is stopped by the criterion (5.25), we see from (5.26) that the iteration process is stopped at $x_4^*=0.3162\ 2777$. The true error of this value is

$$+0.3983\cdots\times 10^{-8}$$

and its error estimate given by (5.24) is

$$\frac{7}{4}\times 10^{-8}+\frac{13225}{64}\times 10^{-16}+\cdots\approx\frac{7}{4}\times 10^{-8}=1.75\times 10^{-8}.$$

The above results show that the error estimates obtained in the present note are considerably good.

Lastly let us consider the case where the iteration process is stopped by the criterion either

$$|x_{n+1}^*-x_n^*|<10^{-3} \tag{5.27}$$

or

$$|x_{n+1}^*-x_n^*|<10^{-4}. \tag{5.28}$$

As is seen from (5.26), the iteration process is stopped at $x_3^*=0.3162\ 2779$ or $x_4^*=0.3162\ 2777$ according as the criterion (5.27) or (5.28) is used.

In case (5.27) is used,

$$|x_3^*-x_2^*|<1.2\times 10^{-4}.$$

Therefore, by the Remark 2 of §4, from (5.12), we have

$$\begin{aligned}|x_3^*-\bar{x}| &\leqq \frac{7}{4}\times 10^{-8}+\frac{25}{4}\left(\frac{7}{4}\times 10^{-8}+1.2\times 10^{-4}\right)^2+\cdots \\ &\approx 10.8\times 10^{-8},\end{aligned} \tag{5.29}$$

while the true error of x_3^* is $+2.3983\cdots\times 10^{-8}$.

In case (5.28) is used,

$$|x_4^* - x_3^*| \leqq 2 \times 10^{-8}.$$

Therefore, likewise, we have

$$|x_4^* - \bar{x}| \leqq 1.75 \times 10^{-8} + \cdots \approx 1.75 \times 10^{-8}, \tag{5.30}$$

while the true error of x_4^* is $+0.3983\cdots \times 10^{-8}$.

The above results show that the Newton method gives a considerably accurate solution even if the iteration process is stopped by the criterion of the form (1.12) *for* $\alpha \gg \varepsilon$ *and, in addition, that the error estimates given in the present note are considerably good.*

Reference

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

Department of Mathematics
Faculty of Science
Hiroshima University