Componentwise Error Estimates for Approximate Solutions of Nonlinear Equations

TETSURO YAMAMOTO*

Let $x^{(0)}$ be an approximate solution of a system of nonlinear equations. On the basis of the theory of pseudometric space due to Schröder, Collatz and others, a theorem is first proved which determines an existence region of a solution of the system. Next, the result is applied to derive a theorem which may be useful for finding a sharp error bound of each component of $x^{(0)}$. Further, after proving a uniqueness theorem, an algorithm is presented for finding the best uniqueness domain based upon it. Finally, the results are illustrated with a system of two nonlinear equations.

1. Introduction

There is abundant literature concerning the convergence of iterative methods for finding a solution of a nonlinear equation

$$f(x)=0, \tag{1}$$

defined on a domain D in R^n. Many of the results developed there are applicable for estimating the error of an approximate solution $x^{(0)}$ which was obtained by some method. However, the usual norm estimates may not be suitable in general, when one wants to find a sharp error bound of each or specified component of $x^{(0)}$. In this case, it is desirable to apply the convergence theorems in pseudometric space which were obtained by Schröder [7], [8], [9], Collatz [2] and others.

They are formulated in a certain set $D_0 \subseteq D$ containing $x^{(0)}$ (cf., for example, Theorems 1–3 below). However, no criterion for the choice of D_0 has been given, while the estimates depend, in general, on the set D_0. In fact, for finding a sharp error bound, the set D_0 should be chosen as small as possible. On the other hand, for finding a large uniqueness region, D_0 should be chosen large.

In this paper, after stating notation and definitions in §2, we first establish, in §3, a theorem giving a sharp existence region of a solution, which is independent of an initial choice of the set D_0. It slightly improves Schröder's one. Next, it is shown how the result is applied in practical computation to find a sharp error bound of each component of $x^{(0)}$. Further, in §4, we prove a uniqueness theorem which somewhat generalizes a theorem of Kantorovich and Akilov [3] and describe a practical algorithm for finding the best uniqueness domain based upon the theorem. Finally, in §5, the results are illustrated with a system of two nonlinear equations.

*Department of Mathematics, Faculty of Science, Ehime University, Matsuyama 790, Japan.

2. Notation and Definitions

Let $x=(x_1,\ldots,x_n)^t$ and $y=(y_1,\ldots,y_n)^t$ be two column vectors of R^n. We write $x \geqslant y$ or $y \leqslant x$ to signify that $x_i \geqq y_i$ for all i. Thus $x \geqslant 0$ means that all the elements of x are nonnegative. We put $v[x]=(|x_1|,\ldots,|x_n|)^t$ and $\rho(x,y)=v[x-y]$. Then R^n becomes a complete pseudometric space with a pseudodistance $\rho(x,y)$, where the limit of a sequence of vectors are naturally defined (see Collatz [2]). A symbol $x>y$ or $y<x$ means that $x \geqslant y$ but $x \neq y$. The same notation is used for two matrices $A=(a_{ij})$ and $B=(b_{ij})$ of the same type. For example, $A \geqslant B$ means that $a_{ij} \geqq b_{ij}$ for all i, j, and we put $v[A]=(|a_{ij}|)$, etc.

The spectral radius of an $n \times n$ matrix A is denoted by $\sigma(A)$.

As a vector norm, we adopt the sum norm $\|x\|=|x_1|+\cdots+|x_n|$. Observe that, if $x \geqslant 0$, $y \geqslant 0$ are two vectors, then

$$\|x+y\|=\|x\|+\|y\|. \tag{2}$$

The corresponding matrix norm is defined by

$$\|A\|=\max_{\|x\|=1}\|Ax\|=\max_{j}\sum_{i=1}^{n}|a_{ij}|.$$

If $C=(c_{ijk})$, $i, j, k=1, 2, \cdots, n$ is a bilinear operator (a tensor of the third order), then, analogously, we have

$$\|C\|=\max_{\|x\|=\|y\|=1}\|Cxy\|=\max_{j,k}\sum_{i=1}^{n}|c_{ijk}|.$$

For a nonnegative vector $d \geqslant 0$ and a positive number r, we put

$$\bar{U}(x^{(0)}, d)=\{x \in R^n \mid \rho(x, x^{(0)}) \leqslant d\},$$

$$S(x^{(0)}, r)=\{x \in R^n \mid \|x-x^{(0)}\|<r\},$$

and

$$\bar{S}(x^{(0)}, r)=\{x \in R^n \mid \|x-x^{(0)}\| \leqq r\}.$$

$J(x)$ stands for the Jacobian matrix of $f(x)=(f_1(x),\cdots, f_n(x)^t$, provided that it exists on D.

Finally, we shall call an operator $G: D \subseteq R^n \to R^n$ a

P-contraction on a set $D_0 \subseteq D$ if there exists a matrix P with $P \geqq 0$ and $\sigma(P) < 1$ such that, for every $x, y \in D$, $\rho(Gx, Gy) \leqq P\rho(x, y)$.

3. Componentwise Error Estimates

Let $x^{(0)} = (x_1^{(0)}, \cdots, x_n^{(0)})^t$ be an approximate solution of (1). In this section, we shall prove an existence theorem (Theorem 4) of a solution and gives a practical method (Theorem 5) for finding a sharp error bound of each component $x_i^{(0)} (1 \leqq i \leqq n)$.

We first refer to two important results which are special cases of the theorems originally obtained by Schröder.

Theorem 1 (Schröder [7], Ortega and Rheinboldt [4] and Collatz [2]). *Let G be a P-contraction on a closed set D_0 and*

$$GD_0 \subseteq D_0. \tag{3}$$

Then G has exactly one fixed point x^* in D_0. If $x^{(0)} \in D_0$, the iterates $x^{(k+1)} = Gx^{(k)}$, $k = 0, 1, 2, \cdots$, belong to D_0 and converge to x^*. The error estimate

$$\rho(x^{(k)}, x^*) \leqq (I-P)^{-1} P^k \rho(x^{(0)}, x^{(1)}), \quad k = 0, 1, 2, \cdots, \tag{4}$$

holds. Further, if the condition (3) is replaced by

$$\overline{U} \equiv \overline{U}(x^{(1)}, (I-P)^{-1} P\rho(x^{(0)}, x^{(1)})) \subseteq D_0, \tag{5}$$

then $G\overline{U} \subseteq \overline{U}$ and the same conclusions hold in $\overline{U}$: the sequence $\{x^{(k)}\}$ belongs to $\overline{U}$ and converges to the unique fixed point x^* in $\overline{U}$, and the estimate (4) holds.

Theorem 2 (Schröder [8], [9] and Rheinboldt [6]). *Suppose that there exist a convex set $D_0 \subseteq D$ which contains $x^{(0)}$ and a symmetric bilinear operator $B(D_0)$ such that, for every x, y in D_0,*

$$v[J(x) - J(y)] \leqq B(D_0)\rho(x, y). \tag{6}$$

Let A be a matrix and put $K = v[I - AJ(x^{(0)})]$ and $\varepsilon = v[Af(x^{(0)})]$. If the sequence

$$d^{(0)} = 0, \quad d^{(k+1)} = \frac{1}{2} v[A] B(D_0) d^{(k)2} + Kd^{(k)} + \varepsilon, \quad k = 0, 1, 2, \cdots,$$

converges to a vector $d^* \geqq 0$ and, if $\overline{U}(x^{(0)}, d^*) \subseteq D_0$, then (1) has exactly one solution x^* in $\overline{U}(x^{(0)}, d^*)$.

On the other hand, the following theorem was proved byUrabe [10].

Theorem 3 *Given an equation (1) where $f(x)$ is continuously differentiable with respect to x in the set D of the x-space. Let $x^{(0)} \in D$ and suppose that there are an $n \times n$ matrix A, an $n \times n$ matrix $P \geqq 0$ with the property $\sigma(P) < 1$, and an n-dimensional vector $\delta > 0$ such that*

(i) $D_\delta = \overline{U}(x^{(0)}, \delta) \subseteq D$,

(ii) $v[I - AJ(x)] \leqq P$ for any $x \in D_\delta$,

(iii) $(I-P)^{-1} v \leqq \delta$,

where v is an n-dimensional vector satisfying $\varepsilon = v[Af(x^{(0)})] \leqq v$. The equation (1) then possesses a unique solution x^* in D_δ, and it is valid that $det[J(x^*)] \neq 0$ and $\rho(x^{(0)}, x^*) \leqq (I-P)^{-1} v$.

Obviously, this theorem follows from Theorem 1. In fact, if we define an operator G by $Gx = x - Af(x)$, then it follows from the condition (ii) of Theorem 3 that G is a P-contraction on D, since G has the first Fréchet derivative $I - AJ(x)$ at every x in D. Further, the conditions (i), (iii) imply (5) with $D_0 = D_\delta$. We thus obtain from Theorem 1

$$\rho(x^{(0)}, x^*) \leqq (I-P)^{-1} \rho(x^{(0)}, x^{(1)}) \leqq (I-P)^{-1} v.$$

The estimates obtained from these theorems depend on the choice of D_0 or D_δ. However, the theorems tell us nothing about how to choose them. For example, in Theorem 3, we should choose δ so small that the conditions (i), (ii) are satisfied. But, if δ is too small, then the condition (iii) may not hold.

To overcome this difficulty, we now set the following assumptions:

Assumption 1. There is an $n \times n$ matrix A such that the matrix $K = v[I - AJ(x^{(0)})]$ has the spectral radius which is smaller than unity.

Assumption 2. Given any convex set $D_0 \subseteq D$, there exists a symmetric tensor of the third order $H(D_0) = (h_{ijk}(D_0))$, $i, j, k = 1, 2, \ldots, n$ such that $h_{ijk}(D_0) \geqq 0$ for all i, j, k and

$$\rho(AJ(x), AJ(y)) \leqq H(D_0)\rho(x, y), \tag{7}$$

holds for every x, y in D_0.

Assumption 3. The tensor $H(D_0)$ is monotonically increasing and continuous. That is, if $D_0 \subseteq D_0'$, then $H(D_0) \leqq H(D_0')$ (i.e., $h_{ijk}(D_0) \leqq h_{ijk}(D_0')$ for all i, j, k) and $H(D_0) \to H(D_0')$ as $D_0 \to D_0'$. Observe that Assumption 1 implies the nonsingularity of $J(x^{(0)})$ and that Assumptions 2 and 3 are satisfied if the second Fréchet derivative of $f(x)$ exists and is continuous in D.

Further, we set

$$\varepsilon = v[Af(x^{(0)})], \quad e = (I-K)^{-1}\varepsilon,$$
$$C(d) = (I-K)^{-1} H(\overline{U}(x^{(0)}, d)),$$
$$C(e + \|e\|(1, \cdots, 1)^t) = (c_{ijk}), \quad c_i = \max_{j,k} c_{ijk},$$

and

$$c = (c_1, \cdots, c_n)^t.$$

Then, under the above notation and assumptions, we first have the following theorem.

Theorem 4 *Let $\overline{U}(x^{(0)}, 2\|e\|(1, \cdots, 1)^t) \subseteq D$. If $2\|c\| \cdot \|e\| \leqq 1$ or $2\|C(2\|e\|(1, \cdots, 1)^t)\| \cdot \|e\| \leqq 1$, then the monotonically increasing sequence*

$$\delta^{(0)} = 0, \quad \delta^{(k+1)} = \frac{1}{2} C(\delta^{(k)}) \delta^{(k)2} + e, \quad k = 0, 1, 2, \cdots, \tag{8}$$

converges to a vector δ^* and the equation (1) *has a solution x^* in $\bar{U}(x^{(0)}, \delta^*)$. There is no other solution in $\bar{U}(x^{(0)}, \delta^*)$.*

Proof. We first consider the case where $2\|c\| \cdot \|e\| \leqq 1$. For every vector $u \geqslant 0$, we have $C(e+\|e\|(1, \cdots, 1)^t)u^2 \leqslant \|u\|^2 c$. But, as is easily verified, the equation

$$\frac{1}{2}\|u\|^2 c - u + e = 0, \quad u \geqslant 0,$$

has a nonnegative solution

$$\alpha = e + \frac{\|e\|^2}{1-\|c\| \cdot \|e\| + \sqrt{1-2\|c\| \cdot \|e\|}} c \leqslant e + \|e\|(1, \cdots, 1)^t.$$

Therefore, by Assumption 3 and by induction on k, it is shown that the sequence

$$u^{(0)} = 0, \quad u^{(k+1)} = \frac{1}{2}\|u^{(k)}\|^2 c + e, \quad k = 0, 1, 2, \cdots,$$

satisfies

$$\delta^{(k)} \leqslant u^{(k)} \leqslant \frac{1}{2}\|\alpha\|^2 c + e = \alpha.$$

Since the sequence $\{\delta^{(k)}\}$ is monotonically increasing, it converges to a vector $\delta^* \geqslant 0$. It is a solution of the equation $C(\delta)\delta^2 - 2\delta + 2e = 0$, since, by Assumption 3, $C(\delta^{(k)}) \to C(\delta^*)$ as $k \to \infty$.

Now, to prove the existence of a solution in $\bar{U}(x^{(0)}, \delta^*)$, we apply the majorant principle due to Schröder [8], [9] and Rheinboldt [6]. For every vector v, let $H = H(\bar{U}(x^{(0)}, \delta^*))$ and

$$\psi(v) = \frac{1}{2}Hv^2 + Kv + \varepsilon.$$

Then it is known [8], [9] that a sequence of vectors $v^{(k)} = \psi(v^{(k-1)})(k \geqq 1)$, $v^{(0)} = 0$, majorizes the sequence $x^{(k)} = x^{(k-1)} - Af(x^{(k-1)})(k \geqq 1)$:

$$\rho(x^{(k+1)}, x^{(k)}) \leqslant v^{(k+1)} - v^{(k)}, \quad k = 0, 1, 2, \cdots. \tag{9}$$

In fact, we have

$$x^{(k+1)} - x^{(k)} = x^{(k)} - x^{(k-1)} - AJ(x^{(0)})(x^{(k)} - x^{(k-1)}) - A\int_0^1 \{J(x^{(k-1)} + t(x^{(k)} - x^{(k-1)})) - J(x^{(0)})\}(x^{(k)} - x^{(k-1)})\, dt,$$

and, by induction on k,

$$\begin{aligned}
\rho(x^{(k+1)}, x^{(k)}) &\leqslant K\rho(x^{(k)}, x^{(k-1)}) + \frac{1}{2}H\{\rho(x^{(k)}, x^{(0)}) + \rho(x^{(k-1)}, x^{(0)})\}\rho(x^{(k)}, x^{(k-1)}) \\
&\leqslant K(v^{(k)} - v^{(k-1)}) + \frac{1}{2}H\{(v^{(k)} - v^{(0)}) + (v^{(k-1)} - v^{(0)})\}(v^{(k)} - v^{(k-1)}) \\
&= K(v^{(k)} - v^{(k-1)}) + \frac{1}{2}(Hv^{(k)2} - Hv^{(k-1)2}) = \psi(v^{(k)}) - \psi(v^{(k-1)}) = v^{(k+1)} - v^{(k)},
\end{aligned}$$

where we have used Assumption 2.

Further, if $v \leqslant \delta^*$, then $\psi(v) \leqslant \psi(\delta^*) = \delta^*$ since $C(\delta^*)\delta^{*2} - 2\delta^* + 2e = 0$. This shows that $v^{(k)} \leqslant \delta^* (k \geqq 0)$ and the monotonically increasing sequence $\{v^{(k)}\}$ converges to a vector v^*. It is easy to see that $v^* = \delta^*$. (See Remark 1 below.) We thus conclude from (9) that $\{x^{(k)}\}$ converges to a vector x^* in $\bar{U}(x^{(0)}, \delta^*)$, which is a solution of (1). The uniqueness follows from the proof of Rheinboldt's theorem [6; Theorem 2.6]. In fact, we can show that if y^* is any solution in $\bar{U}(x^{(0)}, \delta^*)$, then

$$\rho(y^*, x^{(k)}) \leqslant \delta^* - v^{(k)}, \quad k = 0, 1, 2, \cdots,$$

which implies that $y^* = x^*$. We should remark here that

$$\bar{U}(x^{(0)}, \delta^*) \subseteq \bar{U}(x^{(0)}, 2\|e\|(1, \cdots, 1)^t) \subseteq D,$$

since, by (2),

$$\begin{aligned}
\|\delta^*\| \leqq \|\alpha\| &= \|e\| + \frac{\|e\|^2}{1-\|c\| \cdot \|e\| + \sqrt{1-2\|c\| \cdot \|e\|}}\|c\| \\
&= \frac{2\|e\|}{1+\sqrt{1-2\|c\| \cdot \|e\|}} \leqq 2\|e\|. \qquad (10)
\end{aligned}$$

Next, we consider the case where $2\|C(2\|e\|(1, \cdots, 1)^t)\| \cdot \|e\| \leqq 1$. Let $C = C(2\|e\|(1, \cdots, 1)^t)$ and

$$r = \frac{1-\sqrt{1-2\|C\| \cdot \|e\|}}{\|C\|} = \frac{2\|e\|}{1+\sqrt{1-2\|C\| \cdot \|e\|}}.$$

Then the sequence

$$w^{(0)} = 0, \quad w^{(k+1)} = \frac{1}{2}Cw^{(k)2} + e, \quad k = 0, 1, 2, \cdots,$$

remains in $\bar{S}(0, r)$ and converges to a vector w^*, since, by induction on k,

$$\|w^{(k+1)}\| \leqq \frac{1}{2}\|C\| \cdot \|w^{(k)}\|^2 + \|e\| \leqq \frac{1}{2}\|C\|r^2 + \|e\| = r.$$

This implies that

$$\delta^{(k)} \leqslant w^{(k)} \leqslant w^* \leqslant r(1, \cdots, 1)^t \leqslant 2\|e\|(1, \cdots, 1)^t.$$

Hence, $\{\delta^{(k)}\}$ again converges to δ^*. This means the unique existence of the solution x^* in $\bar{U}(x^{(0)}, \delta^*)$. Q.E.D.

Remark 1. The sequence $\{v^{(k)}\}$ converges to δ^*. For, if $v^* < \delta^*$, then $\|v^*\| < \|\delta^*\|$ and

$$\begin{aligned}
\delta^* - v^* &= \psi(\delta^*) - \psi(v^*) \\
&= \frac{1}{2}H(\bar{U}(x^{(0)}, \delta^*))(\delta^{*2} - v^{*2}) + K(\delta^* - v^*),
\end{aligned}$$

so that

$$\begin{aligned}
\delta^* - v^* &= \frac{1}{2}(I-K)^{-1}H(\bar{U}(x^{(0)}, \delta^*))(\delta^{*2} - v^{*2}) \\
&= \frac{1}{2}C(\delta^*)(\delta^* + v^*)(\delta^* - v^*) \\
&\leqslant \frac{1}{2}\|\delta^* + v^*\|c(1, \cdots, 1)(\delta^* - v^*) \\
&< \|\delta^*\|c(1 \cdots, 1)(\delta^* - v^*),
\end{aligned}$$

where we have used the facts that $\delta^* \leqslant e + \|e\|(1, \cdots, 1)^t$ and

$$C(\delta^*)u \leqslant C(e+\|e\|(1,\cdots,1)^t)u \leqslant \|u\| \cdot \begin{pmatrix} c_1 \cdots c_1 \\ \cdots\cdots \\ c_n \cdots c_n \end{pmatrix} = \|u\| c(1, \cdots, 1),$$

for every vector $u \geqslant 0$. Hence, by (10),

$$0 < \|\delta^* - v^*\| < \|\delta^*\| \cdot \|c\| \cdot \|\delta^* - v^*\| \leqq 2\|e\| \cdot \|c\| \cdot \|\delta^* - v^*\| \leqq \|\delta^* - v^*\|,$$

which is a contradiction and we must have $v^* = \delta^*$.

Remark 2. If $2\|c\| \cdot \|e\| < 1$ or $2\|C\| \cdot \|e\| < 1$ where $C = C(2\|e\|(1, \cdots, 1)^t)$, then the matrix $J(x^*)$ is nonsingular, that is, the solution x^* is simple. To prove this, let $L = I - AJ(x^{(0)})$. Then $\sigma(L) \leqq \sigma(K) < 1$ so that $J(x^{(0)})$ is nonsingular and $J(x^{(0)})^{-1} = (I-L)^{-1}A$. Therefore, putting $M = I - J(x^{(0)})^{-1}J(x^*)$, we have

$$\begin{aligned} v[M] &= v[(I-L)^{-1}(AJ(x^{(0)}) - AJ(x^*))] \\ &\leqslant v[I + L + L^2 + \cdots]H(\overline{U}(x^{(0)}, \delta^*))\rho(x^{(0)}, x^*) \\ &\leqslant (I + K + K^2 + \cdots)H(\overline{U}(x^{(0)}, \delta^*))\delta^* \\ &= C(\delta^*)\delta^* \leqslant \|\delta^*\| c(1, \cdots, 1), \end{aligned}$$

so that

$$\|M\| = \|v[M]\| \leqq \|\delta^*\| \cdot \|c\| \leqq 2\|e\| \cdot \|c\|.$$

Similarly we have $v[M] \leqslant C\delta^*$ and $\|M\| \leqq \|C\| \cdot \|\delta^*\| \leqq 2\|C\| \cdot \|e\|$. Consequently, if $2\|c\| \cdot \|e\| < 1$ or $2\|C\| \cdot \|e\| < 1$, then $\|M\| < 1$ and $J(x^*) = J(x^{(0)})(I - M)$ is nonsingular. But, if $2\|c\| \cdot \|e\| = 1$ or $2\|C\| \cdot \|e\| = 1$, then $J(x^*)$ may be singular. For example, let $n = 1$, $f(x) = x^2$, $x^{(0)} = a > 0$ and $A = (2x^{(0)})^{-1}$. Then a simple computation yields $c = C = H = a^{-1}$, $2\varepsilon = 2e = \delta^* = a$ and $2\|c\| \cdot \|e\| = 1$. Thus Theorem 4 asserts that there exists a unique solution x^* of (1) in the closed interval $[0, 2a]$. But, $x^* = 0$ and $J(x^*) = f'(0) = 0$.

Remark 3. It would be interesting to compare Theorem 4 with Theorems 2 and 3. If the convex set D_0 in Theorem 2 includes the region $\overline{U}(x^{(0)}, \delta^*)$, then, obviously, we have $\delta^* \leqslant d^*$. Next, as the region D_δ in Theorem 3, we take $D^* = \overline{U}(x^{(0)}, \delta^*)$, a sharp region whose a priori choice can not be expected in a practical computation. Then, for any x in $D_\delta = D^*$, we have

$$\begin{aligned} v[I - AJ(x)] &= v[I - AJ(x^{(0)}) - A(J(x) - J(x^{(0)}))] \\ &\leqslant K + H(D^*)\delta^*, \end{aligned}$$

provided that (7) holds. (We may take $H(D^*) = v[A]B(D^*)$ if (6) holds.) So we can take $P = K + H(D^*)\delta^*$. Then we have

$$\varepsilon + P\delta^* - \delta^* = \frac{1}{2}H(D^*)\delta^* > 0,$$

which means that $(I-P)^{-1}\varepsilon > \delta^*$ if $\sigma(P) < 1$. Therefore, in order that Theorem 3 gives a sharper estimate than Theorem 4, δ should be chosen so small that $\delta < \delta^*$.

Remark 4. If the equation (1) is defined on a domain D in C^n and $x^{(0)} \in C^n$, then our theorem holds by putting

$$\overline{U}(x^{(0)}, d) = \{x \in C^n \mid \rho(x, x^{(0)}) \leqslant d\}, \text{ etc.}$$

Corollary. *Let $f(z) = 0$ be a single equation in C^1 or R^1 and apply the Newton method $z_{k+1} = z_k - f(z_k)/f'(z_k)$, $k = 0, 1, \cdots$. Set*

$$\varepsilon_{k+1} = |f(z_{k+1})/f'(z_{k+1})|, \quad D_{k+1} = \{z \mid |z - z_{k+1}| \leqq 2\varepsilon_{k+1}\},$$

$$M_{k+1} = \sup_{z \in D_{k-1}} |f''(z)| \quad \text{and} \quad h_{k+1} = M_{k+1}/|f'(z_{k+1})|,$$

provided that z_{k+1} is defined, $f'(z_{k+1}) \neq 0$ and $f''(z)$ exists in D_{k+1}. If $2h_{k+1}\varepsilon_{k+1} \leqq 1$, then a solution z^* exists and is unique in

$$D^*_{k+1} = \left\{ z \in C^1 \mid |z_{k+1} - z| \leqq \frac{2\varepsilon_{k+1}}{1 + \sqrt{1 - 2h_{k+1}\varepsilon_{k+1}}} \right\}. \quad (11)$$

Remark 5. Let

$$\hat{D}_{k-1} = \{z \mid |z - z_k| \leqq \varepsilon_{k-1}\} \quad \text{and} \quad \hat{M}_{k-1} = \sup_{z \in \hat{D}_{k-1}} |f''(z)|.$$

Then, Ostrowski's theorem [5; Theorem 7.2] asserts that, if $2\varepsilon_{k-1}\hat{M}_{k-1} \leqq |f'(z_{k-1})|$, then

$$\begin{aligned} &|z_{k+1} - z^*| \leqq \frac{\hat{M}_{k-1}}{2|f'(z_k)|}\varepsilon^2_{k-1}, \quad (12) \\ &2\varepsilon_k |f'(z_k)| \leqq \hat{M}_{k-1}\varepsilon^2_{k-1}, \\ &\varepsilon_{k+1} \leqq \frac{1}{2}\varepsilon_k \leqq \frac{1}{4}\varepsilon_{k-1}, \quad \text{and} \quad 2\varepsilon_k M_{k-1} \leqq |f'(z_k)|. \end{aligned}$$

Obviously we have $D_{k+1} \subseteq \hat{D}_{k-1}$ and $M_{k+1} \leqq \hat{M}_{k-1}$. Hence

$$\frac{2\varepsilon_{k+1}}{1 + \sqrt{1 - 2h_{k+1}\varepsilon_{k+1}}} \leqq 2\varepsilon_{k+1} \leqq \varepsilon_k \leqq \frac{\hat{M}_{k-1}}{2|f'(z_k)|}\varepsilon^2_{k-1}.$$

That is, the estimate (11) is sharper than (12).

Under Assumptions 1–3, we now obtain the following theorem which may be useful when one wants to find a sharp error bound of each or specified component of $x^{(0)}$.

Theorem 5 *Let $\overline{U}(x^{(0)}, 2\|e\|(1, \cdots, 1)^t) \subseteq D$.*

(i) *If $2\|c\| \cdot \|e\| \leqq 1$, then there exists a solution of (1) in $\overline{U}(x^{(0)}, \alpha)$, where*

$$\alpha = e + \frac{\|e\|^2}{1 - \|c\|\|e\| + \sqrt{1 - 2\|c\| \cdot \|e\|}}c.$$

(ii) *Let $\{\delta^{(k)}\}$ be the iterated sequence of vectors defined in (8) and set $\xi^{(k)} = 2(\delta^{(k+1)} - \delta^{(k)})$ and $\eta^{(k)} = \delta^{(k)} + \xi^{(k)}$, provided that $2\|c\| \cdot \|e\| \leqq 1$ or $2\|C(2\|e\|(1, \cdots, 1)^t)\| \cdot \|e\| \leqq 1$. If the conditions*

$$\eta^{(k)} \leqslant 2\|e\|(1, \cdots, 1)^t,$$

and

$$\{C(\eta^{(k)}) - C(\delta^{(k)})\}\delta^{(k)2} + 2C(\eta^{(k)})\delta^{(k+1)}\xi^{(k)} \leqslant \xi^{(k)},$$

are satisfied for some $k \geqq 0$, then there exists a solution of (1) in $\overline{U}(x^{(0)}, \eta^{(k)})$.

Proof. It remains to prove (ii). Assume now that $\delta^{(l)} \leqq \eta^{(k)}$ for some $l > k$. Then we have

$$\begin{aligned}\delta^{(l+1)} &= \frac{1}{2}C(\delta^{(l)})\delta^{(l)2} + e \leqq \frac{1}{2}C(\eta^{(k)})\eta^{(k)2} + e \\ &= \frac{1}{2}C(\eta^{(k)})(\delta^{(k)} + \xi^{(k)})^2 + e \\ &= \frac{1}{2}\{C(\eta^{(k)}) - C(\delta^{(k)})\}\delta^{(k)2} + C(\eta^{(k)})\left(\delta^{(k)} + \frac{1}{2}\xi^{(k)}\right)\xi^{(k)} \\ &\quad + \frac{1}{2}C(\delta^{(k)})\delta^{(k)2} + e \\ &= \frac{1}{2}\{C(\eta^{(k)}) - C(\delta^{(k)})\}\delta^{(k)2} + C(\eta^{(k)})\delta^{(k+1)}\xi^{(k)} + \delta^{(k+1)} \\ &\leqq \frac{1}{2}\xi^{(k)} + \delta^{(k+1)} = \eta^{(k)}.\end{aligned}$$

Therefore, by induction on l, we have $\delta^{(l)} \leqq \eta^{(k)}$ for every $l > k$. Letting $l \to \infty$, we now obtain $\delta^* \leqq \eta^{(k)} \leqq 2\|e\|(1, \cdots, 1)^t$, which means the existence of a solution in $\bar{U}(x^{(0)}, \eta^{(k)})$. Q.E.D.

4. Determination of a Uniqueness Domain

Given an approximate solution $x^{(0)}$ of (1), it is also of importance to find a uniqueness domain as large as possible. In this section, taking account that the parallelotope $\bar{U}(x^{(0)}, d)$ with a vector $d \geqq 0$ is included in the corresponding ball $\bar{S}(x^{(0)}, \|d\|)$, we shall present a method for finding a uniqueness domain $S(x^{(0)}, r)$ as large as possible. *We keep the assumptions of §3.*

We first prove the following theorem which slightly generalizes a theorem of Kantorovich and Akilov [3], [6; Corollary 3.3].

Theorem 6. *Let $D_0 = \bar{S}(x^{(0)}, r) \subseteq D$ for some $r > 0$ and set $\hat{C}(r) = (I-K)^{-1}H(D_0)$. If $2\|\hat{C}(r)\| \cdot \|e\| < 1$, then* (1) *has at most one solution in $S(x^{(0)}, r^{**}) \cap D_0$, where*

$$r^{**} = \frac{1 + \sqrt{1 - 2\|\hat{C}(r)\| \cdot \|e\|}}{\|\hat{C}(r)\|}.$$

Proof. For simplicity, let $H = H(D_0)$ and $Gx = x - Af(x)$. Then, for every x, y in D_0,

$$\begin{aligned}Gx - Gy &= x - y - AJ(x^{(0)})(x-y) \\ &\quad - A\int_0^1 \{J(y + t(x-y)) - J(x^{(0)})\}(x-y)\,dt,\end{aligned}$$

and, by (7),

$$\rho(Gx, Gy) \leqq K\rho(x, y) + \frac{1}{2}H\{\rho(x, x^{(0)}) + \rho(y, x^{(0)})\}\rho(x, y). \tag{13}$$

Therefore, if x^* is a solution in $S(x^{(0)}, r^{**}) \cap D_0$, then

$$\begin{aligned}\rho(x^{(0)}, x^*) &\leqq \rho(x^{(0)}, Gx^{(0)}) + \rho(Gx^{(0)}, Gx^*) \\ &\leqq \varepsilon + K\rho(x^{(0)}, x^*) + \frac{1}{2}H\rho(x^{(0)}, x^*)^2,\end{aligned}$$

so that

$$\begin{aligned}\rho(x^{(0)}, x^*) &\leqq (I-K)^{-1}\varepsilon + \frac{1}{2}(I-K)^{-1}H\rho(x^{(0)}, x^*)^2 \\ &= e + \frac{1}{2}\hat{C}(r)\rho(x^{(0)}, x^*)^2.\end{aligned}$$

In particular, we have

$$\|x^{(0)} - x^*\| \leqq \|e\| + \frac{1}{2}\|\hat{C}(r)\| \cdot \|x^{(0)} - x^*\|^2.$$

This gives us

$$\|x^{(0)} - x^*\| \leqq r^* = \frac{1 - \sqrt{1 - 2\|\hat{C}(r)\| \cdot \|e\|}}{\|\hat{C}(r)\|},$$

since, by assumption, $\|x^{(0)} - x^*\| < r^{**}$.

Now, to prove the uniqueness, let $\hat{x}$ be another solution in $S(x^{(0)}, r^{**}) \cap D_0$. Then, it follows from (13) that

$$\begin{aligned}\rho(\hat{x}, x^*) &= \rho(G\hat{x}, Gx^*) \\ &\leqq K\rho(\hat{x}, x^*) + \frac{1}{2}H\{\rho(\hat{x}, x^{(0)}) + \rho(x^*, x^{(0)})\}\rho(\hat{x}, x^*),\end{aligned}$$

or

$$\rho(\hat{x}, x^*) \leqq \frac{1}{2}(I-K)^{-1}H\{\rho(\hat{x}, x^{(0)}) + \rho(x^*, x^{(0)})\}\rho(\hat{x}, x^*).$$

Since $\|\hat{x} - \hat{x}^{(0)}\| \leqq r^*$, we thus obtain

$$\begin{aligned}\|\hat{x} - x^*\| &\leqq \|\hat{C}(r)\| r^* \|\hat{x} - x^*\| \\ &= \{1 - \sqrt{1 - 2\|\hat{C}(r)\| \cdot \|e\|}\}\|\hat{x} - x^*\|,\end{aligned}$$

which means $\hat{x} = x^*$ since $2\|\hat{C}(r)\| \cdot \|e\| < 1$. Consequently, (1) has at most one solution in $S(x^{(0)}, r^{**}) \cap D_0$. Q.E.D.

Corollary. *Let $\bar{U}(x^{(0)}, 2\|e\|(1, \cdots, 1)^t) \subseteq D$ and $2\|C(2\|e\|(1, \cdots, 1)^t)\| \cdot \|e\| < 1$. Then the solution of* (1) *(which exists in $\bar{U}(x^{(0)}, \delta^*)$ where δ^* is the limit of the sequence* (8)) *is unique in $\bar{S}(x^{(0)}, 2\|e\|)$.*

Proof. Take $r = 2\|e\|$ in Theorem 6, by noting that $\|\hat{C}(r)\| \leqq \|C(r(1, \cdots, 1)^t)\|$ and $r^{**} > 2\|e\|$. (The solution of (1) does exist in $\bar{U}(x^{(0)}, \delta^*)$ by Theorem 4.) Q.E.D.

The best uniqueness domain based upon Theorem 6 is obtained, if the radius r of the closed ball D_0 is chosen so that

$$r = \frac{1 + \sqrt{1 - 2\|\hat{C}(r)\| \cdot \|e\|}}{\|\hat{C}(r)\|}.$$

Such an r can be found by the following procedures, provided that the domain D is sufficiently large.

Procedure 1. Let $2p_0 = 2\|\hat{C}(2\|e\|)\| \cdot \|e\| < 1$ and set $r_0 = 2\|e\|$ and

$$s_0 = \frac{1 + \sqrt{1 - 2p_0}}{\|\hat{C}(2\|e\|)\|}.$$

Procedure 2. Assume that $S(x^{(0)}, s_0) \subseteq D$ and define the bounded sequences $\{r_i\}$, $\{s_i\}(i = 1, 2, \cdots)$ as follows:

If $2p_i=2\|\hat{C}(s_{i-1})\|\cdot\|e\|<1$, then set

$$\omega_i=\frac{1+\sqrt{1-2p_i}}{\|\hat{C}(s_{i-1})\|},$$

$$r_i=\max\{r_{i-1}, \min\{s_{i-1}, \omega_i\}\},$$

$$s_i=\frac{1}{2}\{r_i+\max\{s_{i-1}, \omega_i\}\}.$$

If $2p_i\geqq 1$, then set $r_i=r_{i-1}$ and $s_i=\frac{1}{2}(r_{i-1}+s_{i-1})$.

Then, the solution is unique in $S(x^{(0)}, r_i)$ and the sequence $\{r_i\}$ is monotonically increasing and converges to a positive number r, the desired one. Note that $r_i\leqq s_i\leqq s_0$ for each i.

For automatic computation, however, it may be convenient to replace $\hat{C}(s_i)$ by $C(s_i(1,\cdots,1)^t)$, although the result obtained becomes rather rough.

5. A Numerical Example

To illustrate our results, we take up the system

$$\begin{aligned} f_1&=3x_1^2x_2+x_2^3-1=0,\\ f_2&=x_1^4+x_1x_2^3-1=0, \end{aligned} \tag{14}$$

which was considered by Kantorovich and Akilov [3]. By using the Newton method starting from an initial value (0.98, 0.32), they obtained an approximate solution $x^{(0)t}=(0.991189, 0.327382)$ and concluded that there is a solution x^* such that

$$\begin{aligned} &0.991173\leqq x_1^*\leqq 0.991205,\\ &0.327366\leqq x_2^*\leqq 0.327398. \end{aligned} \tag{15}$$

We shall apply Theorem 5 to (14) and $x^{(0)}$, by choosing $A=J(x^{(0)})^{-1}$. The second Fréchet derivative of $f(x)$ at x is

$$J'(x)=\begin{pmatrix} f_{111} & f_{112} & f_{121} & f_{122}\\ f_{211} & f_{212} & f_{221} & f_{222}\end{pmatrix}=\begin{pmatrix} 6x_2 & 6x_1 & 6x_1 & 6x_2\\ 12x_1^2 & 3x_2^2 & 3x_2^2 & 6x_1x_2\end{pmatrix},$$

where

$$f_{ijk}=\frac{\partial^2 f_i(x)}{\partial x_j\partial x_k}.$$

Therefore, $B=B(\bar{U}(x^{(0)}, d))$ defined in Theorem 2 is obtained if we replace each $x_i^{(0)}$ in the expression of $J'(x^{(0)})$ by $|x_i^{(0)}|+d_i$. Set $H(\bar{U}(x^{(0)}, d))=\nu[A]B$. Then, it is easily verified that $2\|c\|\cdot\|e\|=0.21\cdots E-4<1$ and

$$\alpha^t=(0.5215503\cdots E-6, 0.1331679\cdots E-5).$$

Thus, by (i) of Theorem 5, we can assert that the solution x^* does exist in $\bar{U}(x^{(0)}, \alpha)$ and

$$\begin{aligned} &0.9911884\cdots\leqq x_1^*\leqq 0.9911895\cdots,\\ &0.3273806\cdots\leqq x_2^*\leqq 0.3273833\cdots. \end{aligned} \tag{16}$$

This estimate is sharper than (15). Further, if we execute the iteration (8) with a stopping criterion $\delta^{(k+1)}-\delta^{(k)}\leqslant 10^{-13}(1,\cdots,1)^t$, then this condition is satisfied at $k=2$, as well as the conditions in (ii) of Theorem 5, and we obtain

$$\eta^{(2)t}=(0.5215459\cdots E-6, 0.1331677\cdots E-5),$$

which slightly improves the estimate (16). Observe that, if $\|c\|\cdot\|e\|$ is sufficiently small, then

$$e\leqslant\delta^*\leqslant\alpha\leqslant e+\|e\|^2c\leqq e.$$

Therefore, in this case, there is little difference between α and $\eta^{(k)}$. The result of the modified procedure described at the end of the previous section is also shown in Table 1, which shows that the solution is unique in $S(x^{(0)}, r_i)$ for each i.

The computation in this example was done on a FACOM 230–28 computer of Ehime University, with double precision arithmetic chopping to hexadecimal 14-digit numbers. The values of r_i and s_i in Table 1 are respectively chopped and rounded to six digits.

Table 1 The radii of the uniqueness domain $S(x^{(0)}, r_i)$.

i	r_i	s_i
0	0.370643E−05	0.350150E+00
1	0.188274E+00	0.269212E+00
2	0.213309E+00	0.241261E+00
3	0.223187E+00	0.232224E+00
4	0.226535E+00	0.229380E+00
5	0.227606E+00	0.228493E+00

Acknowledgement

The author wishes to express his thanks to Professor Lothar Collatz of Hamburg University for his many helpful comments and to Dr. M. T. Noda of Ehime University for discussions during the preparation of this paper. Thanks also go to the referees for their constructive comments.

References

1. BOHL, E., MONOTONIE: Lösbarkeit und Numerik bei Operatorgleichungen. New York: Springer (1974).
2. COLLATZ, L. Functional analysis and numerical mathematics. New York: Academic Press (1966).
3. KANTOROVICH, L. V. and AKILOV, G. P. Functional analysis in normed spaces. Oxford: Pergamon Press (1964).
4. ORTEGA, J. M. and RHEINBOLDT, W. C. Iterative solution of nonlinear equations in several variables. New York: Academic Press (1970).
5. OSTROWSKI, A. M. Solution of equations in Euclidean and Banach spaces. New York: Academic Press (1973).
6. RHEINBOLDT, W. C. A unified convergence theory for a class of iterative processes. SIAM J. Numer. Anal. 5 (1968) 42–63.
7. SCHRÖDER, J. Das Iterationsverfahren bei allgemeinerem Abstandsbegiff. Math. Z. 66 (1956) 111–116.
8. SCHRÖDER, J. Nichtlineare Majoranten beim Verfahren der schrittweisen Näherung. Arch. Math. 7 (1956) 471–484.
9. SCHRÖDER, J. Über das Newtonsche Verfahren. Arch. Rat. Mech. Anal. 1 (1957) 154–180.
10. URABE, M. A posteriori component-wise error estimation of approximate solutions to nonlinear equations. Lect. Notes in Computer Sci. 29 (Springer) (1975) 99–117.

(Received July 10, 1978: revised June 20, 1979)