INSTITUTUL
DE
MATEMATICA

INSTITUTUL NATIONAL
PENTRU CREATIE
STIINTIFICA SI TEHNICA

AN ERROR ANALYSIS FOR THE SECANT METHOD

by

Florian A.POTRA

PREPRINT SERIES IN MATHEMATICS
No.66/1980

BUCURESTI

AN ERROR ANALYSIS FOR THE SECANT METHOD

by

Florian A.POTRA*)

December 1980

*) Department of Mathematics, National Institute for
Scientific and Technical Creation, Bd.Pacii 220,
79622 Bucharest, Romania.

AN ERROR ANALYSIS FOR THE SECANT METHOD

by

Florian A.Potra*)

*) *Department of Mathematics, National Institute for Scientific and Technical Creation, Bd.Păcii 220, 779622 Bucharest, ROMANIA*

Summary. One gives sharp apriori and aposteriori error bounds for the secant method for solving non-linear equations in Banach spaces. One also investigates the numerical stability of this method. The stability results are analoguous to those obtained by Lancaster for Newton's method.

1. Introduction

Let us consider a class $\mathcal{C}$ of pairs (f,v_o) where f is a nonlinear operator defined on a subset $\mathcal{D}_f$ of a Banach space $\mathcal{E}$ with values in a Banach space $\mathcal{F}$, and $v_o=(x_{-k+1},\dots,x_o)$ is a system of k points from $\mathcal{D}_f$. We want to attach to each pair $(f,v_o) \in \mathcal{C}$ a sequence $(x_n)_{n\geqslant 0}$ of points of $\mathcal{D}_f$ converging to a root x^* of the equation $f(x)=0$. One way of doing this is to associate with the pair (f,v_o) a mapping $F:\mathcal{D}\subset\mathcal{D}_f^p\longrightarrow\mathcal{E}$, where $k\geqslant p$, and to try to obtain a sequence $(x_n)_{n\geqslant 0}$ by the recurrent scheme:

$$x_{n+1}=F(x_{n-p+1},\dots,x_n)\ , \qquad n=0,1,2,\dots \tag{1}$$

The above scheme will actually yield a sequence $(x_n)_{n\geqslant 0}$, if $u_o=(x_{-p+1},\dots,x_o)$ is an admissible system of starting points in the sense given by the following defintion:

Definition 1. Consider a mapping $F:\mathcal{D}\subset\mathcal{E}^p\to\mathcal{E}$ and define recursively

$$\mathcal{D}_o=\mathcal{D},\quad \mathcal{D}_{n+1}=\{u=(y_1,y_2,\dots,y_p)\in\mathcal{D}_n;\ (y_2,\dots,y_p,F(u))\in\mathcal{D}_n\}\quad n=0,1,2.$$

..

Any $u_o\in\mathcal{D}_\infty=\bigcap\limits_{n\geqslant 0}\mathcal{D}_n$ will be called an admissible system of starting points for the recurrent scheme (1). ■

If $u_o=(x_{-p+1},\dots,x_o)$ is an admissible system of starting points for the recurrent scheme (1) we shall also say that the iterative algorithm (1) is well defined.

Now we can define the notion of an iterative procedure of type (p.1) for the class $\mathcal{C}$. The more general notion of an iterative procedure of type (p.m) will be given in [11]. (see also [9]and [10])

Definition 2. Let $\mathcal{C}$ be a class of pairs (f,v_o) where f is a nonlinear operator defined on a subset $\mathcal{D}_f$ of a Banach space $\mathcal{E}$ with values in a Banach space $\mathcal{F}$, and $v_o=(x_{-k+1},\dots,x_o)\in\mathcal{D}^k_f$. Let p be an integer less than or equal to k. By an iterative procedure of type (p.1) for the class $\mathcal{C}$,we mean an application which associates with any $(f,v_o)\in\mathcal{C}$ a mapping $F:\mathcal{D}\subset\mathcal{D}^p_f\to\mathcal{E}$ having the following two properties:

(i) $u_o=(x_{-p+1},\dots,x_o)$ is an admissible system of starting points for the recurrent scheme (1);

(ii) the sequence $(x_n)_{n\geqslant}$ generated by (1) converges to a root x^* of the equation $f(x)=0$. ■

Having an iterative procedure of type (p.1) for the class $\mathcal{C}$ it is important to find a function $\alpha:\mathbb{Z}_+\to\mathbb{R}_+$ and a function $\beta:\mathbb{R}^p_+\to\mathbb{R}_1$ such that the following inequalities be satisfied

$$d(x_n,x^*)\leqslant\alpha(n) \tag{2}$$

$$d(x_n,x^*)\leqslant \beta(d(x_{n-p+1}\ ,\ x_{n-p}),\ldots,d(x_n\ ,x_{n-1})) \tag{3}$$

for every pair $(f,x_o)\in \mathcal{C}$ and every positive integer n.

The inequalities (2) are generally called apriori estimates because the right hand side of (2) can be computed before obtaining the points $x_1,\ldots,x_n$ via algorithm (1), while the inequatities (3) are called aposteriori estimates because their right hand side can be computed only after obtaining these points.

The estimates (2) and/or (3) will be called sharp if there exists a pair $(f,u_o)\in \mathcal{C}$ for which these estimates are attained for all n=1,2,3,... . Using the method of nondiscrete mathematical induction, in a series of recent paper ([6]-[10]) one has obtained such estimates for some well known iterative procedures. However, these results have mainly a theoretical importance because in practical applications the iterative algorithm (1) can be performed only approximatively. Thus instead of the "theoretical" sequence $(x_n)_{n\geqslant 0}$ in practice we shall obtain a perturbed sequence $(\tilde{x}_n)_{n\geqslant 0}$ given by

$$\tilde{u}_o=u_o\ ,\ \tilde{x}_{n+1}=\tilde{F}(\tilde{x}_{n-p+1}\ ,\ldots,\tilde{x}_n)\qquad n=0,1,\ldots \tag{4}$$

The domain $\tilde{\mathcal{D}}$ of $\tilde{F}$ is included in the domain $\mathcal{D}$ of F. In practice $\tilde{\mathcal{D}}$ consists of those elements of $\mathcal{D}$ which are representable on a certain computer, being thus a finite set.

It is important to give sufficient conditions under which there exists a number $\delta>0$ and a sequence $(t_n)_{n\geqslant 0}$ such that the following inequalities be satisfied:

$$d(\tilde{x}_n,x_n)\leqslant t_n\leqslant \delta\ \ ,\qquad n=0,1,2,\ldots \tag{5}$$

In applications, the set $\tilde{\mathcal{D}}$ being finite, it follows that after a certain numbers of steps the sequence $(\tilde{x}_n)_{n\geqslant 0}$ will become periodic i.e. there exist n_o and m such that $\tilde{x}_{n+m}=\tilde{x}_n$ for every $n\geqslant n_o$. In this case (5) implies that the following estimate

$$d(\tilde{x}_n, x^*) \leqslant \delta \tag{6}$$

holds for all $n \geqslant n_o$. To see this, one has to write $d(\tilde{x}_n, x_{n+km}) = d(\tilde{x}_{n+km}, x_{n+km}) \leqslant \delta$ and to let k to tend to infinity.

The estimate (6) shows us that we can compute the root x^* with the precision given by δ. But this result is not very convenient in applications because the number n_o can be very large. In this case we note that from (2) and (5) one can obtain the following estimates

$$d(x_n, x^*) \leqslant t_n + \alpha(n), \qquad n=0,1,2,\ldots \tag{7}$$

If the function β is increasing, in the sense that $a_1 \leqslant b_1, \ldots, a_p \leqslant b_p$ imply $\beta(a_1,\ldots,a_p) \leqslant \beta(b_1,\ldots,b_p)$, then from (3) and (5) it follows that

$$d(x_n,x^*) \leqslant t_n + \beta(t_{n-p+1} + t_{n-p} + d(\tilde{x}_{n-p+1}, \tilde{x}_{n-p}) + \ldots + t_n + t_{n-1} + d(\tilde{x}_n, \tilde{x}_{n-1})). \tag{8}$$

The inequalities (7) and (8) may be interpreted as apriori, respectively aposteriori, estimates for the perturbed algorithm (4).

In this paper we shall make an analysis of the type described above for the secant method and for one of its modifications. The results obtained in the "theoretical" case constitute a slight improvement of the results contained in [6] and [7]. The results obtained in the perturbed case are new. Similar results concerning Newton's process can be found in the papers of P.Lancaster [2], J.Rokne [14] and G.Miel [3].

2. The method of nondiscrete mathematical induction

In the study of the iterative algorithm (1) we shall use the method of nondiscrete mathematical induction. This method was developped by V.Ptak by refining the closed graph theorem (see [12] or [13] for its general principles and motivation). V.Ptak used this method to investigate iterative algorithms of type (1) with p=1. In [7] the method of V.Ptak has been extended to the case where p was arbitrary. In this section we shall restate the results obtained in the above mentioned paper.

Let T denote either the set of all positive numbers, or an interval of the form $(0,b]=\{x\in\mathbb{R};\ 0<x\leq b\}$. Let ω be a mapping of the cartesian product T^p into T and let us consider the "iterates" $\omega^{(n)}$ of ω given for each $t=(t_1,\dots,t_p)\in T^p$ by the following recurrent scheme:

$$\omega^{(o)}(t)=t_p,\quad \omega^{(n+1)}(t)=\omega^{(n)}(t_2,\dots,t_p,\omega(t)),\quad n=0,1,\dots \tag{10}$$

Definition 3. A mapping $\omega:T^p\to T$, with the above iteration law, is called a rate of convergence of type (p,1) on T, if the series

$$\sigma(t)=\sum_{n=0}^{\infty}\omega^{(n)}(t) \tag{11}$$

is convergent for all $t\in T^p$.∎

In what follows we shall use this notion in the study of the iterative algorithm (1). F will be a mapping of $\mathcal{D}$ into X, where X is a complete metric space, and $\mathcal{D}$ a subset of the cartesian power X^p. We shall attach to F the mapping $\overline{F}:\mathcal{D}\to X^p$, defined for every $u=(y_1,\dots,y_p)\in\mathcal{D}$ by

$$\overline{F}(u)=(y_2,\dots,y_p,F(u)). \tag{12}$$

Denoting $u_n=(x_{n-p+1},\dots,x_n)$ we shall have

$$u_{n+1}=\overline{F}(u_n),\qquad n=0,1,2,\dots \tag{1'}$$

Similarly we shall attach to ω the mapping $\overline{\omega}:T^p\to T^p$ defined for every $t=(t_1,\dots,t_p)\in T^p$ by

$$\bar{\omega}(t)=(t_2,\dots,t_p,\omega(t)). \tag{13}$$

Let us denote by $\bar{\omega}^{(n)}$ the iterates of $\bar{\omega}$ in the sense of the usual composition of functions i.e. $\bar{\omega}^{(o)}(t)=t$, $\bar{\omega}^{(n+1)}(t)=\bar{\omega}(\bar{\omega}^{(n)}(t))$. Then relation (10) reduces to

$$\omega^{(o)}(t)=t_p\,,\quad \omega^{(n+1)}(t)=\omega(\bar{\omega}^{(n)}(t)). \tag{10'}$$

It will be convenient to introduce the notation

$$\beta(t)=\sigma(t)-t_p.$$

From (11) and (13) it follows immediately that $\beta(t)=\sigma(\bar{\omega}(t))$.

With the above notations we can state the following proposition:

<u>Proposition 1</u>. Let X be a complete metric space and let $\mathcal{D}$ be a subset of X^p. Let us consider the mappings $F:\mathcal{D}\longrightarrow X$ and $Z:T^p\longrightarrow \exp\mathcal{D}$, where $\exp\mathcal{D}$ denotes the class of all subsets of $\mathcal{D}$. Let ω be a rate of convergence of type (p,1) on T.

If there exist $u_o=(x_{-p+1},\dots,x_o)\in\mathcal{D}$ and $t_o\in T^p$ such that

$$u_o\in Z(t_o) \tag{14}$$

and if the relations

$$Fu\in Z(\bar{\omega}(t))\,, \tag{15}$$

$$d(Fu,y_p)\leqslant t_p \tag{16}$$

are satisfied for all $t=(t_1,\dots,t_p)\in T^p$ and $u=(y_1,\dots,y_p)\in Z(t)$, then:

(i) The iterative algorithm (1) is well defined.

(ii) There exists an $x^*\in X$ such that $x^*=\lim_{n\to\infty} x_n$.

(iii) The following relations are satisfied for all n=0,1,...

$$u_n \in Z(\varpi^n(t_o)), \tag{17}$$

$$d(x_{n+1},x_n) \leqslant \omega^{(n)}(t_o), \tag{18}$$

$$d(x_n,x_o) \leqslant \sigma(t_o)-\sigma(\varpi^{(n)}(t_o)), \tag{19}$$

$$d(x_n,x^*) \leqslant \sigma(\varpi^{(n)}(t_o)). \tag{20}$$

(iv) Let n be a positive integer and let $d_n \in T^p$; if $u_{n-1} \in Z(d_n)$, then

$$d(x_n,x^*) \leqslant \beta(d_n). \tag{21}$$

<u>Proof</u>. The fact that u_o is an admissible system of starting points for (1) is a consequence of (17). For n=0 this relation reduces to (14). If we suppose that (17) is true for a certain n, then according to (15) it follows that it is true for n+1 also. The inequality (18) follows then immediately from (17) and (16). Now for any $k \in \mathbb{N}$ we may write

$$d(x_n,x_{n+k}) \leqslant \sum_{j=n}^{n+k-1} \omega^{(j)}(t_o) \tag{22}$$

This shows that the sequence $(x_n)_{n\geqslant 0}$ is a fundamental one. Point (ii) of the proposition follows then from the assumption that X is complete. Letting k to tend to infinity in (22) we obtain (20). To get (19) we have only to observe that

$$d(x_n,x_o) \leqslant \sum_{k=0}^{n-1} \omega^{(k)}(t_o)=\sigma(t_o)-\sigma(\varpi^{(n)}(t_o)) \tag{23}$$

Thus we have proved the first three points of the proposition. Taking n=1 in (20) it follows that in particular we have proved the implication

$$\text{"If } u_o \in Z(t_o)\text{, then } d(Fu_o,x^*) \leqslant \beta(t_o)\text{"}. \tag{24}$$

Replacing u_o by u_{n-1} and t_o by d_n we obtain point (iv). ∎

3. Optimal error bounds for the secant method

In this section we shall study the iterative procedures

$$x_{n+1}=x_n-\delta f(x_{n-1},x_n)^{-1}f(x_n) \tag{25}$$

$$x_{n+1}=x_n-\delta f(x_{-1},x_o)^{-1}f(x_n) \tag{26}$$

where f is a nonlinear operator between two Banach spaces, x_{-1} and x_o are two points in the domain of f, and δf is a consistent approximation of f'.

The first procedure is generally called the secant method but it is also known under the name of Regula falsi or the method of chords. This procedure has been known from the time of early italian algebrists (see [5]) and it was extended for the solution of nonlinear equations in Banach spaces by Sergeev [18] and Schmidt [15]. In the above mentioned papers one has used the notion of divided difference of an operator. The possibility of using the more general notion of consistent approximation of the derivative was realized later (see [1] and [16]).

The iterative procedure (26), called the modified secant method, was first considered by S.Ulm [19].

In the sequel we shall prove that if the triplet (f,x_o,x_{-1}) belongs to a certain class $\mathcal{C}(h_o,q_o,r_o)$, then the iterative procedures (25) and (26) are convergent and we shall give sharp apriori and aposteriori estimates.

We shall consider the notion of consistent approximation of the derivative in the acception given in [1], which is more particular than the original acception given in [4]. If $\mathcal{E}$ and $\mathcal{F}$ are two Banach spaces we shall denote by $L(\mathcal{E},\mathcal{F})$ the Banacha space of all bounded linear operators from $\mathcal{E}$ into $\mathcal{F}$.

Definition 4. Let $\mathcal{E}$ and $\mathcal{F}$ be two Banach spaces and let V be a convex and open subset of $\mathcal{E}$. Let $f:V\to\mathcal{F}$ be a (nonlinear operator) which is Fréchet-differentiable on V. A mapping $\delta f:V\times V\to L(\mathcal{E},\mathcal{F})$ will be called a consistant approximation of f', if there exists a constant $H>0$ such that the following inequality be satisfied for all $x,y,z\in V$:

$$\|\delta f(x,y)-f'(z)\|\leqslant H(\|x-z\|+\|y-z\|).\blacksquare \qquad (27)$$

The above condition implies the Lipschitz continuity of f'. In this case using a standard argument (see [4; 3.2.12]) we deduce that

$$\|f(u)-f(v)-f'(v)(u-v)\|\leqslant H\|u-v\|^2;\quad u,v\in V. \qquad (28)$$

Thus, for all $x,y,u,v\in V$ we have

$$\|f(u)-f(v)-\delta f(x,y)(u-v)\|\leqslant \|f(u)-f(v)-f'(v)(u-v)\|$$

$$+\|(f'(v)-\delta f(x,y))(u-v)\|\leqslant H(\|u-v\|+\|x-v\|+\|y-v\|)\|u-v\|. \qquad (29)$$

Let $\mathcal{C}(h_o,q_o,r_o)$ be the class of all the triplets (f,x_o,x_{-1}) satisfying the following properties:

(c_1) f is a nonlinear operator having the domain of definition $\mathcal{D}_f$ included into a Banach space $\mathcal{E}$ and taking values in a Banach space $\mathcal{F}$.

(c_2) x_o and x_{-1} are two points of $\mathcal{D}_f$ such that

$$\|x_o-x_{-1}\|\leqslant q_o \qquad (30)$$

(c_3) f is Frechet differentiable in the open ball $U=S(x_o,\mu)$ and continuous on its closure $\overline{U}$.

(c_4) there exists a consistent approximation δf of f' such that $D_o:=\delta f(x_{-1},x_o)$ is invertible and

$$\|D_o^{-1}(\delta f(x,y)-f'(z))\|\leqslant h_o(\|x-z\|+\|y-z\|) \qquad (31)$$

for all $x,y,z\in U$.

(c_5) the following inequalities are satisfied:

$$\|D_o^{-1}f(x_o)\| \leqslant r_o , \tag{32}$$

$$h_oq_o+2\sqrt{h_or_o} \leqslant 1 , \tag{33}$$

$$\mu \geqslant \frac{1}{2h_o}(1-h_oq_o-\sqrt{(1-h_oq_o)^2-4h_or_o}\ ) =: \mu_o \tag{34}$$

Let us remark that the constant h_o appearing in (31) generally depends on μ. On the other hand μ has to be greater or equal to μ_o which depends on h_o. It is worth then to note the following inequality

$$r_o+\sqrt{r_o(q_o+r_o)} \geqslant \mu_o \tag{35}$$

which allows us to take for μ the estimate $r_o+\sqrt{r_o(q_o+r_o)}$ which does not depend any more on h_o.

Using the iterative procedure (26) we shall show that if $(f,x_o,x_{-1}) \in \mathcal{C}(h_o,q_o,r_o)$, then the equation $f(x)=0$ has a solution x^* which is unique in a certain neighbourhood of x_o. First let us associate with this iterative procedure a rate of convergence of type (1.1).

Lemma 1. If $h_o>0$, $q_o \geqslant 0$, $r_o \geqslant 0$ are three numbers satisfying condition (33) then the function

$$\omega_1(r)=r(h_or+1-2\sqrt{h_o^2a^2+h_or}) \tag{36}$$

is a rate of convergence of type (1,1) on the interval $T=(0,r_o]$ and the corresponding σ-function is given by

$$\sigma_1(r)=\sqrt{a^2+h_o^{-1}r}-a, \tag{37}$$

where

$$a=\frac{1}{2h_o}\sqrt{(1-h_oq_o)^2-4h_or_o} . \tag{38}$$

Proof. Let us first observe that (33) implies that the quantity under the square root sign from (38) is nonnegative. Let us consider now the real polynomial $g(s)=h_o(s^2-a^2)$ and let $(s_n^{(1)})_{n\geq 0}$ be a sequence of real numbers satisfying the following relation

$$s_{n+1}^{(1)}=s_n^{(1)}-g(s_n^{(1)}) , \qquad n=0,1,2,\ldots \qquad (39)$$

One can show that if $s_o^{(1)}\in(a,h_o^{-1}-a)$ then the sequence $(s_n^{(1)})_{n\geq 0}$ is decreasing and converges to a. Let us put now $s_o^{(1)}=s_o^{(1)}(r)=\sqrt{a^2+h_o^{-1}r}$ then, for all $r\in(0,r_o]$, we shall have $s_o^{(1)}\in(a,\frac{1-h_oq_o}{2h_o})\subset(a,h_o^{-1}-a)$. We also observe that in this case $g(s_o^{(1)})=r$, $g(s_1^{(1)})=\omega_1(r)<r$. Let us denote by $\omega_1^{(n)}$ the iterates of ω_1 in the sense of the usual function composition. We shall obviously have

$$\omega_1^{(n)}(r)=g(s_n^{(1)})=s_n^{(1)}-s_{n+1}^{(1)} , \qquad n=0,1,2,\ldots$$

It follows then that ω_1 is a rate of convergence of type (1.1) on the interval $(0,r_o]$ and that

$$\sigma_1(\omega_1^{(n)}(r))=s_n^{(1)}-a , \qquad n=0,1,2,\ldots \quad \blacksquare \qquad (40)$$

We shall pass now to the study of the iterative procedure (26). Before stating the main result let us note that from (c_4) it follows that

$$\|D_o^{-1}(f(u)-f(v)-\delta f(x,y)(u-v))\|\leq h_o(\|u-v\|+\|x-v\|+\|y-v\|)\|u-v\| \qquad (41)$$

for all $x,y,u,v\in U$. Using the above inequality, together with Proposition 1 and Lemma 1, we shall prove the following theorem:

Theorem 1. If $(f,x_o,x_{-1})\in\mathcal{C}(h_o,q_o,r_o)$, then by the iterative algorithm (26) one obtains a sequence $(x_n)_{n\geq 0}$ of points belonging to the open sphere $S(x_o,\mu_o)$, which converges to a root x^* of the equation $f(x)=0$ and the following estimates hold :

$$\|x_n - x^*\| \leqslant \sigma_1(\omega_1^{(n)}(r_o)), \qquad n=0,1,\ldots \tag{42}$$

$$\|x_n - x^*\| \leqslant \sigma_1(\|x_n - x_{n-1}\|) - \|x_n - x_{n-1}\|, \quad n=1,2,\ldots \tag{43}$$

where ω_1 and σ_1 are the functions given in Lemma 1.

Proof. Let us consider the mappings $F:S(x_o,\mu_o) \to \mathcal{E}$ and $Z:(0,r_o] \to \exp\mathcal{E}$ given by the following relations:

$$Fx = x - D_o^{-1}f(x), \tag{44}$$

$$Z(r) = \{x \in \mathcal{E};\ \|x - x_o\| \leqslant \sigma_1(r_o) - \sigma_1(r),\ \|D_o^{-1}f(x)\| \leqslant r\}. \tag{45}$$

observing that $\sigma_1(r_o) = \mu_o$, it follows that $Z(r) \subset S(x_o,\mu_o)$. If $r \in (0,r_o]$, $x \in Z(r)$ and $x' = Gx$, then we have

$$\|x' - x_o\| \leqslant \|x' - x\| + \|x - x_o\| \leqslant r + \sigma_1(r_o) - \sigma_1(r) = \sigma_1(r_o) - \sigma_1(\omega_1(r)).$$

The relation $x' = Gx$ is equivalent to $f(x) + D_o(x' - x) = 0$, so that using (41) we obtain

$$\|D_o^{-1}f(x')\| = \|D_o^{-1}(f(x') - f(x) - D_o(x' - x))\| \leqslant h_o(\|x' - x \quad \|x - x_o\| + \|x_o - y_o\|)\|x' - x\|$$

$$\leqslant h_o(2\sigma_1(r) + q_o - r)r = \omega_1(r)$$

From the above relations it follows that the hypotheses (14), (15) and (16) of Proposition 1 are satisfied. Thus the sequence $(x_n)_{n \geqslant 0}$ converges to a point $x^* \in \mathcal{E}$. The estimates (42) follow then from (20), while, corresponding to (17) and (18), we have

$$x_{n-1} \in Z(\omega_1^{(n-1)}(r_o)), \ \|x_n - x_{n-1}\| \leqslant \omega_1^{(n-1)}(r_o), \quad n=1,2,3,\ldots \tag{46}$$

Using the fact that σ_1 increasing on $(0,r_o]$ from the above relations we deduce that $x_{n-1} \in Z(\|x_n - x_{n-1}\|)$, so that according to point (iv) of Proposition 1 it follows that the aposteriori estimates (43) are true for $n=1,2,\ldots$.

To end the proof of the theorem we observe that by letting n to tend to infinity in (26) we obtain $f(x^*)=0$. ■

The following proposition contains some information about the behavior of the sequence $(\sigma_1(\omega_1^{(n)}(r_o)))_{n\geqslant 0}$ which appears in (42). The cases $a>0$ and $a=0$ are considered separately. Let us observe that $a=0$ if and only if we have equality in (33).

Proposition 2. Suppose the hypotheses of Lemma 1 are verified.

(i) If $a>0$, then for all $n=0,1,2,\ldots$ hold the inequalities :

$$\frac{2r_o}{1-q_o-2h_oa}\left[h_o(q_o+r_o)\right]^n \leqslant \sigma_1(\omega_1^{(n)}(r_o)) \leqslant \frac{r_o}{2h_oa}(1-2h_oa)^n . \qquad (47)$$

(ii) If $a=0$, then the following estimates are satisfied for every $n=1,2,3,\ldots$

$$\frac{1}{n+1}\sqrt{\frac{r_o}{h_o}} \leqslant \sigma_1(\omega_1^{(n)}(r_o)) \leqslant \frac{1}{n+2}\,\frac{1}{h_o} \qquad (48)$$

Proof. (i) Observing that for $r\in(0,r_o]$ we have

$$h_o(q_o+r_o)r \leqslant \omega_1(r) \leqslant (1-2h_oa)r$$

and using the fact that ω_1 is increasing on $(0,r_o]$ one can easily show that

$$r_o\left[h_o(q_o+r_o)\right]^n \leqslant \omega_1^{(n)}(r_o) \leqslant r_o(1-2h_oa)^n \;,\; n=0,1,2,\ldots$$

The estimates (47) follow then as a consequence of the inequalities

$$\frac{2r}{1-q_o-2h_oa} \leqslant \sigma_1(r) \leqslant \frac{r}{2ah_o} \quad ,$$

which are valid for every $r\in(0,r_o]$.

(ii) The inequalities (48) can easily be proved by induction observing that for $a=0$ from (39) and (40) we have

$$\sigma_1(\omega_1^{(n+1)}(r_o))=\sigma_1(\omega_1^{(n)}(r_o))-h_o\left[\sigma_1(\omega_1^{(n)}(r_o))\right]^2 . ■$$

The result obtained above will be essentially used in the proof of the following theorem:

Theorem 2. Let $(f,x_o,x_{-1})\in\mathcal{C}(h_o,q_o,r_o)$, and let x^* be the root of the equation $f(x)=0$ obtained in Theorem 1.

(i) If $a=0$, then x^* is the unique solution of the equation $f(x)=0$ in the set $\bar{U}\cap S(x_o,\mu_o+2a)$.

(ii) If $a=0$, then x^* is the unique root of the equation $f(x)=0$ in the closed ball $\overline{S(x_o,\mu_o)}$.

Proof. Observing that (31) implies the Lipschitz condition

$$\| D_o^{-1}(f'(u)-f'(v))\| \leq 2h_o\| u-v\| ; \quad u,v\in U \tag{49}$$

and using the equality

$$f(x)-f(y)-D_o(x-y)=\int_0^1\left[f'(x+t(y-x))-f'(x_o)\right](x-y)dt + (f'(x_o)-D_o)(x-y)$$

we deduce that for all $x,y\in U$ we have

$$\| D_o^{-1}(f(x)-f(y)-D_o(x-y))\| \leq h_o(\| x-x_o\| + \| y-x_o\| + \| x_o-y_o\|) . \tag{50}$$

(i) consider an $y^*\in\bar{U}\cap S(x_o,\mu_o+2a)$ such that $f(y^*)=0$. Using the inequality $\|x^*-x_o\|\leq\mu_o$ we may write

$$\begin{aligned}\|x^*-y^*\| &= \|D_o^{-1}(f(x^*)-f(y^*)-D_o(x^*-y^*))\| \\ &\leq h_o(\|y^*-x_o\|+\|x^*-x_o\|+\|x_o-y_o\|)\|x^*-y^*\| \\ &< h_o(2\mu_o+2a+q_o)\|x^*-y^*\| = \|x^*-y^*\|\end{aligned}$$

and thus we infer that $y^*=x^*$.

(ii) Let $(x_n)_{n\geq 0}$ be the sequence considered in Theorem 1. If $a=0$ then $\sigma_1(r_o)=\mu_o=\sqrt{h_o^{-1}r_o}$ and from (46) it follows that

$\|x_n-x_o\|\leqslant \mu_o-\sigma_1(\omega_1^{(n)}(r_o))$. If $y^*\in S(x_o,\mu_o)$ and $f(y^*)=0$, then we have succesively

$$\|x_{n+1}-y^*\| = \|D_o^{-1}(f(x_n)-f(y^*)-D_o(x_n-y^*))\|$$

$$\leqslant h_o(\|y^*-x_o\|+\|x_o-y_o\|+\|x_n-x_o\|)\|y^*-x_n\|)$$

$$\leqslant h_o(\mu_o+q_o+\mu_o-\sigma_1(\omega^{(n)}(r_o)))\|y^*-x_n\|$$

$$\leqslant(1-h_o\sigma_1(\omega_1^{(n)}(r_o)))\|y^*-x_n\|\leqslant\|x_1-y^*\|\prod_{j=1}^{n}(1-h_o\sigma_1(\omega_1^{(j)}(r_o))).$$

The series $\sum_{i=1}^{\infty}\sigma_1(\omega_1^j(r_o))$ being divergent (see (48)), from the above inequality it follows that $y^*=\lim_{n\to\infty}x_n=x^*$. ■

Let us remark that in the proofs of the results from this section condition (33) was essentialy used. This condition is fulfilled only if q_o and r_o are small enough. In practical applications q_o can be taken as small as wanted, because, having an initial approximation x_o , we can take y_o very close to it. On the other hand r_o is small only if the initial approximation is "good enough". In practical applications it is sometimes very difficult to find such an initial approximation. However one can prove that condition (33) is in some sense the weakest possible.

Proposition 3. Let $h_o>0$, $q_o\geqslant 0$, $r_o\geqslant 0$ be three numbers which do not satisfy condition (33). Then there exists a function $f:\mathbb{R}\to\mathbb{R}$ and two points $x_o,x_{-1}\in\mathbb{R}$ such that:

(i) conditions (c_1)-(c_4), as well as inequality (32) are satisfied.

(ii) the equation $f(x)=0$ has no solution.

Proof. Let us first observe that the inequality $h_o q_o+2\sqrt{h_o r_o}>1$ is equivalent to $h_o^{-1}<q_o+2r_o+2\sqrt{r_o(q_o+r_o)}$. We shall consider two cases:

If $q_o+2r_o-2\sqrt{r_o(q_o+r_o)}<h_o^{-1}<q_o+2r_o+2\sqrt{r_o(q_o+r_o)}$, then we shall take

$$f(x)=h_o x^2+\frac{1}{4h_o}(2h_o(q_o+2r_o)-1-h_o^2q_o^2),\ x_o=\frac{1-h_oq_o}{2h_o}\ ,\ x_{-1}=\frac{1+h_oq_o}{2h_o}$$

and if $0<h_o^{-1}\leqslant q_o+2r_o-\sqrt{r_o(q_o+r_o)}$, then we shall take

$$f(x)=q_o^{-1}x^2+r_o\ ,\ x_o=0,\ \ x_{-1}=q_o\ \ .\ \blacksquare$$

In the following lemma we shall obtain a rate of convergence of type (2,1), which will be then used in the study of the iterative procedure (25).

Lemma 2. Let T be the whole positive axis and let a be a positive number. Then the function

$$\omega_2(q,r)=\frac{r(q+r)}{r+2\sqrt{r(q+r)+a^2}} \tag{51}$$

is a rate of convergence of type (2,1) on T and the corresponding σ-function is given by

$$\sigma_2(q,r)=r-a+\sqrt{r(q+r)+a^2}\ . \tag{52}$$

Proof. Let us consider the real polynomial $g(s)=s^2-a^2$. The secant method (25) applied to this function reduces to the following recurrent scheme:

$$s_{n+1}^{(2)}=\frac{s_n^{(2)}s_{n-1}^{(2)}+a^2}{s_n^{(2)}+s_{n-1}^{(2)}}\ , \qquad n=0,1,2,\ldots \tag{53}$$

It is obvious that if $s_{-1}^{(2)}\geqslant s_o^{(2)}>a$, then via the above scheme one obtains a sequence $(s_n^{(2)})_{n\geqslant 0}$ which is decreasing and converges to a. For every $t=(q,r)\in T^2$ let us put

$$s_o^{(2)}=s_o^{(2)}(t)=r+\sqrt{r(q+r)+a^2}\ , \quad s_{-1}^{(2)}=s_o^{(2)}+q.$$

In this case we have

$$s_{-1}^{(2)}-s_o^{(2)}=q,\quad s_o^{(2)}-s_1^{(2)}=r\ ,\quad s_1^{(2)}-s_2^{(2)}=\omega_2(q,r)\ .$$

If we define the iterates $\omega_2^{(n)}$ of ω_2 as in (10), with p=2, then it is easy to see that

$$\omega_2^{(n)}(t)=s_n^{(2)}-s_{n+1}^{(2)}\ . \tag{54}$$

Thus it follows that ω_2 is a rate of convergence of type (2,1) on T and that $\sigma_2(t)=s_o^{(2)}-a$, which is exactly formula (52). It also follows that

$$\sigma_2(\overline{\omega}_2^{(n)}(t))\leqslant s_n^{(2)}-a\ ,\ n=0,1,2,\ldots\ ■ \tag{55}$$

In the proof of the next theorem we shall use the following well known result concerning the inversability of linear and bounded operators in Banach spaces.

<u>Lemma 3</u>. If $L_o\in L(\mathcal{E},\mathcal{F})$ is invertible and if $L\in L(\mathcal{E},\mathcal{F})$ has the property that $\| L\| < \| L_o^{-1}\|^{-1}$, then the operator L_o-L is also invertible and

$$\|(L_o-L)^{-1}\| \leqslant (1-\|L\|\,\|L_o^{-1}\|)^{-1}\,\|L_o^{-1}\| . \tag{56}$$

<u>Theorem 3</u>. If $(f,x_o,x_{-1})\in\mathscr{C}(h_o,q_o,r_o)$, then via the iterative procedure (25) one obtains a sequence $(x_n)_{n\geqslant 0}$ of points from the open ball $S(x_o,\mu_o)$ which converges to the root x^* of the equation $f(x)=0$ and the following estimates are satisfed:

$$\|x_n-x^*\| \leqslant \sigma_2(\bar{\omega}_2^{(n)}(q_o,r_o)) \ , \quad n=0,1,2,\dots \tag{57}$$

$$\|x_n-x^*\| \leqslant \left[a^2+\|x_n-x_{n-1}\|\,(\|x_{n-1}-x_{n-2}\|+\|x_n-x_{n-1}\|)\right]^{1/2}-a \ , \quad n=1,2,\dots \ , \tag{58}$$

where ω_2 is the rate of convergence obtained in Lemma 2 and the constant a is given by (38).

<u>Proof</u>. Let $\mathscr{D}=\{u=(y,x)\in U^2;\ \delta f(y,x) \text{ is invertible}\}$ and let $F:\mathscr{D}\to\mathcal{E}^2$ be the mapping given by

$$Fu=x-\delta f(y,x)^{-1}f(x) \ . \tag{59}$$

Denote $t_o=(q_o,r_o)$. From (52) and (38) it follows that $\sigma_2(t_o)=\mu_o$. For every $t=(q,r)\in T^2$ consider te set

$$Z(t)=\{u=(y,x)\in\mathcal{E}^2;\ y\in U,\ \|x-y\|\leqslant q,\ \|x-x_o\|\leqslant \mu_o-\sigma_2(t),$$
$$\text{the linear operator } D=\delta f(y,x) \text{ is invertible and } \|D^{-1}f(x)\|\leqslant r\}$$

One can easily see that $Z(t)\subset\mathscr{D}$ and $u_o=(x_{-1},x_o)\in Z(t_o)$. Let us prove now that if $u=(y,x)\in Z(t)$ then $(x,Fu)\in Z(\bar{\omega}_2(t))$. For this we shall denote $z=Fu$ and we shall prove the following relations:

$$x\in U \ , \quad \|z-x\|\leqslant r \ , \tag{60}$$

$$\|z-x_o\|\leqslant \mu_o-\sigma_2(\bar{\omega}_2(t)) \ , \tag{61}$$

the operator $D_1=\delta f(x,z)$ is invertible and $\|D_1^{-1}f(z)\|\leqslant\omega_2(t)$. (62)

Relations (60) are immediate consequences of the fact that $u\in Z(t)$, because $z-x=-D^{-1}f(x)$.

Using the fact that $\sigma_2(\bar{\omega}_2(t))=\sigma_2(t)-r$ we may write

$$\|z-x_o\| \leqslant \|z-x\| + \|x-x_o\| \leqslant r+\mu_o-\sigma_2(t)=\mu_o-\sigma_2(\bar{\omega}_2(t)).$$

In order to prove (62), let us note first that according to (31) we have:

$$\|D_o^{-1}(D_o-D_1)\| \leqslant \|D_o^{-1}(D_o-f'(x))\| + \|D_o^{-1}(f'(x)-D_1)\| \leqslant h_o(\|x-x_o\| + \|x-y_o\| + \|x-z\|)$$

$$\leqslant h_o(q_o+r+2(\mu_o-\sigma_2(t)))=1-h_o(r+2\sqrt{r(q+r)+a^2}) .$$

Applying now Lemma 3 it follows that the linear operator D_1 is invertible and

$$\|(D_o^{-1}D_1)^{-1}\| \leqslant \frac{1}{h_o(r+2\sqrt{r(q+r)+a^2})} . \tag{63}$$

On the other hand, (59) implies the identity

$$f(z)=f(z)-f(x)-\delta f(y,x)(z-x)$$

and using (41) we obtain

$$\|D_o^{-1}f(z)\| \leqslant h_o(\|z-x\| + \|x-y\|)\|z-x\| \leqslant h_o r(q+r) . \tag{64}$$

Finally from (63) and (64) we have

$$\|D_1^{-1}f(z)\| = \|(D_o^{-1}D_1)^{-1}D_o^{-1}f(z)\| \leqslant \omega_2(t) . \tag{65}$$

Thus we have checked the validity of (60), (61) and (62). It follows that the hypotheses of Proposition 1 are satisfied in our case. Consequently, the sequence $(x_n)_{n\geqslant 0}$ produced by (25) will converge to a point $x^*\in\mathcal{D}$ and the apriori estimates (57) will be satisfied. Moreover we shall have:

$$(x_{n-2},x_{n-1})\in Z(\bar{\omega}_2^{(n-1)}(t_o)) , \qquad n=1,2,3,\ldots \tag{66}$$

$$\|x_{k+1}-x_k\|\leqslant \omega_2^{(k)}(t_o) , \qquad k=0,1,2,\ldots \tag{67}$$

The function σ_2 being increasing (in the sense that $q_1 \leqslant q_2$ and $r_1 \leqslant r_2$ implies $\sigma_2(q_1,r_1)\leqslant \sigma_2(q_2,r_2)$), from the above relations one can easily deduce that

$$(x_{n-2},x_{n-1})\in Z(\|x_{n-1}-x_{n-2}\|,\|x_n-x_{n-1}\|) , \quad n=1,2,\ldots \tag{68}$$

According to point (iv) of Proposition 1 we obtain the estimates (58).

We still have to prove that x^* is a solution of the equation $f(x)=0$. This follows easily if we substitute in (64) $t=\bar{\omega}_2^{(n)}(t_o)$, $z=x_{n+1}$ and let n to tend to infinity. ■

The following proposition shows that the estimates obtained in this section are sharp in the class $\mathcal{C}(h_o,q_o,r_o)$.

Proposition 4. Let $h_o>0$, $q_o\geqslant 0$ and $r_o\geqslant 0$ be any triplet of real numers satisfying condition (33). Then there exist a function $f:\mathbb{R}\rightarrow\mathbb{R}$ and two points $x_o,x_{-1}\in\mathbb{R}$ such that $(f,x_o,x_{-1})\in\mathcal{C}(h_o,q_o,r_o)$ and for which the relations (42), (43), (57) and (58) are verified with equality for all n.

Proof. Take

$$f(x)=h_o(x^2-a^2),\ x_o=(1-h_oq_o)/(2h_o),\ x_{-1}=(1+h_oq_o)/(2h_o) \tag{69}$$

The rest follows from the proofs of Lemma 1 and Lemma 2. ■

4. Estimations in the perturbed case

In this section we shall consider that the iterative procedures studied in Section 3 are perturbed. We shall suppose that all the elements contained in the construction of these procedures are known only approximatively. Moreover we shall suppose that at each step the matrix inversion (or the solution of the respective linear system) is also performed approximatively.

Thus we shall consider that the perturbed iterative algorithms corresponding to (25) and (26) are respectively of the form

$$\tilde{x}_{-1}=x_{-1}\ ,\ \tilde{x}_o=x_o\ ,\ \tilde{x}_{n+1}=\tilde{x}_n-(\delta f(\tilde{x}_{n-1},\tilde{x}_n)+E_n)^{-1}(f(\tilde{x}_n)+e_n)+g_n\ , \qquad (70)$$

$$\tilde{x}_o=x_o\ ,\ \tilde{x}_{n+1}=\tilde{x}_n-(\delta f(x_{-1},x_o)+E_o)^{-1}(f(\tilde{x}_n)+e_n)+g_n\ , \qquad (71)$$

where $e_n\in\mathcal{F}$, $E_n\in L(\mathcal{E},\mathcal{F})$, $g_n\in\mathcal{E}$.

In what follows we shall suppose that there exist three positive numbers ε_1, ε_2 and ε_3 such that

$$\|e_n\|\leqslant\varepsilon_1\ ,\quad \|E_n\|\leqslant\varepsilon_2\ ,\quad \|g_n\|\leqslant\varepsilon_3 \qquad (72)$$

for all $n\in\mathbb{Z}_+$.

In the preceding section we have seen that if $(f,x_o,x_{-1})\in\mathcal{C}(h_o,q_o,r_o)$, then the sequences produced by (25) and (26) stay in the open ball $U_o=S(x_o,\mu_o)$ and consequently in $\mathcal{D}_f$. In the perturbed case we have to suppose that f is defined on a ball $U^*=S(x_o,\mu^*)$, with $\mu^*>\mu_o$. We have also to suppose that the mapping δf extends to $U^*\times U^*$.

In the definition of the class $\mathcal{C}(h_o,q_o,r_o)$ one has imposed condition (31) for all $x,y,z\in U$. In the perturbed case it is more convenient to suppose that the following conditions are satisfied for all $x,y,u,v\in U^*$:

$$\delta f(x,x)=f'(x) \tag{73}$$

$$\|\delta f(x,y)-\delta f(u,v)\|\leqslant H(\|x-u\|+\|y-v\|) \tag{74}$$

These conditions are more restrictive than condition (27) but they are satisfied by the usual examples of consistent approximation (see [16] and [19]).

If $(f,x_o,x_{-1})\in\mathcal{C}(h_o,q_o,r_o)$, then the linear operator $\delta f(x_{n-1},x_n)$ is invertible for all $n\geqslant 0$. In order to assure the invertibility of $\delta f(\tilde{x}_{n-1},\tilde{x}_n)+E_n$ for all $n\geqslant 0$ we shall suppose that $\delta f(x,y)$ is invertible for all $x,y\in U^*$ and that the norms $\|\delta f(x,y)^{-1}\|$ are bounded. More precisely, in the perturbed case we shall impose one, or both, of the following conditions:

(C^*) The open ball $U^*=S(x_o,\mu^*)$ is included into the domain of definition of f and conditions (73)-(74) hold for all $x,y,u,v\in U^*$.

(C^{**}) The linear operator $\delta f(x,y)$ is invertible for all $x,y\in U^*$ and there exists a positive number $\emptyset$ such that

$$1/\emptyset\geqslant\sup\{\|\delta f(x,y)^{-1}\|\ ;\ x,y\in U^*\}\ . \tag{75}$$

We can state now the following theorem concerning the iterative procedure (71).

<u>Theorem 4</u>. Suppose $(f,x_o,x_{-1})\in\mathcal{C}(h_o,q_o,r_o)$ and let $(x_n)_{n\geqslant 0}$ be the sequence generated by the iterative algorithm (26). If condition (C^*) is satisfied and if the following inequalities hold:

$$0<d_o\leqslant\|\delta f(x_{-1},x_o)^{-1}\|^{-1}\ , \tag{76}$$

$$Q_1=d_o-H(h_o^{-1}-2a)-2\varepsilon_2\geqslant 0, \tag{77}$$

$$D_1=Q_1^2-4H(\varepsilon_1+\varepsilon_2 r_o+\varepsilon_3(d_o-\varepsilon_2)), \tag{78}$$

$$\delta_1 = \frac{Q_1 - \sqrt{D_1}}{2H} \leq \mu^* - \mu_o \ , \tag{79}$$

then the iterative algorithm (71) is well defined and we shall have the estimates

$$\| \tilde{x}_n - x_n \| \leq t_n^{(1)} \leq \delta_1 \tag{80}$$

for all $n \in \mathbb{Z}_+$, where the sequence $(t_n^{(1)})_{n \geq 0}$ is given by

$$t_o^{(1)} = 0, \ s_o^{(1)} = \frac{1 - h_o q_o}{2h_o} \ , \ s_{n+1}^{(1)} = s_n^{(1)} - h_o \left[(s_n^{(1)})^2 - a^2 \right]$$

$$t_{n+1}^{(1)} = \frac{1}{d_o - \varepsilon_2} (H(t_n^{(1)})^2 + (2H(s_o^{(1)} - s_n^{(1)}) + Hq_o + \varepsilon_2) t_n^{(1)} +$$

$$+ \varepsilon_1 + \varepsilon_2 (s_n^{(1)} - s_{n+1}^{(1)}) + \varepsilon_3 (d_o - \varepsilon_2)) \ .$$

Proof. We note first that from the proofs of Lemma 1 and Theorem 1 it follows that

$$\| x_{n+1} - x_n \| \leq s_n^{(1)} - s_{n+1}^{(1)} \ , \quad \| x_n - x_o \| \leq s_o^{(1)} - s_n^{(1)} \tag{81}$$

for all $n \in \mathbb{Z}_+$.

Using Lemma 3 it follows that the linear operator $\delta f(x_{-1}, x_o)$ is invertible and

$$\| (\delta f(x_{-1}, x_o) + E_o)^{-1} \| \leq (d_o - \varepsilon_2)^{-1} \ . \tag{82}$$

This fact, together with the remark that (79) and (80) imply $\tilde{x}_n \in U^*$, show us that if (80) is satisfied, then the formulae (71) make sense. Let us prove the inequalities (80). For n=0 they are trivially satisfied. Supposing they hold for n=0,1,...,k we shall prove that they hold for n=k+1 too.

Let $D_o=\delta f(x_{-1},x_o)$ and $\tilde{D}_o=D_o+E_o$. From (26) and (71) we have:

$$x_{k+1}-\tilde{x}_{k+1}=\tilde{D}_o^{-1}(f(\tilde{x}_k)-f(x_k)-D_o(\tilde{x}_k-x_k)-E_oD_o^{-1}f(x_k)+E_o(x_k-\tilde{x}_k)+e_k)-g_k \,. \tag{83}$$

Using (74) and (81) we deduce the following inequalities

$$\begin{aligned}\|f(\tilde{x}_k)-f(x_k)-D_o(\tilde{x}_k-x_k)\| &\leq H(\|\tilde{x}_k-x_k\|+2\|x_k-x_o\|+\|x_o-x_{-1}\|)\|\tilde{x}_k-x_k\| \\ &\leq H(t_k^{(1)}+2(s_o^{(1)}-s_k^{(1)})+q_o)t_k^{(1)},\end{aligned} \tag{84}$$

$$\|E_oD_o^{-1}f(x_k)\| = \|E_o(x_k-x_{k+1})\| \leq \varepsilon_2(s_k^{(1)}-s_{k+1}^{(1)}) \;. \tag{85}$$

Finally, from (82) - (85) it follows that

$$\begin{aligned}\|x_{k+1}-\tilde{x}_{k+1}\| \leq (d_o-\varepsilon_2)^{-1}\Big[&H(t_k^{(1)})^2+(2H(s_o^{(1)}-s_k^{(1)})+Hq_o+\varepsilon_2)t_k^{(1)}+\varepsilon_1 \\ &+\varepsilon_2(s_k^{(1)}-s_{k+1}^{(1)})+\varepsilon_3(d_o-\varepsilon_2)\Big]=t_{k+1}^{(1)}\end{aligned}$$

Let us denote now $B=H(h_o^{-1}-2a)+\varepsilon_2$, $C=\varepsilon_1+\varepsilon_2r_o+\varepsilon_3(d_o-\varepsilon_2)$. Because $s_o^{(1)}-s_k^{(1)}\leq s_o^{(1)}-a$ and $s_k^{(1)}-s_{k+1}^{(1)}\leq s_o^{(1)}-s_1^{(1)}$ we shall have

$$t_{k+1}^{(1)}\leq(d_o-\varepsilon_2)^{-1}(H\delta_1^2+B\delta_1+C)=\delta_1$$

This completes the proof of the theorem. ■

The function $\beta_1(r)=\sqrt{a^2+h_o^{-1}r}-a$ is increasing only on the interval $[-a^2h_o \,,\, r_o+q_o(2-h_oq_o)/4]$, so that we cannot give, in general, aposteriori estimates of the form (8). However, related to (7), we have $\|\tilde{x}_n-x^*\|\leq s_n^{(1)}+t_n^{(1)}-a$.

Concerning the perturbed secant method (70) we have:

Theorem 4. Suppose $(f,x_o,x_{-1})\in\mathscr{C}(h_o,q_o,r_o)$ and let $(x_n)_{n\geqslant 0}$ be the sequence generated by the iterative algorithm (25). Suppose also, that conditions (C^*) and (C^{**}) are fulfilled. Denote $v_o=\max\{q_o,r_o\}$. If the following inequalities are satisfied:

$$Q_2=\emptyset-\frac{6v_oH}{1+2\sqrt{2}}-2\varepsilon_2\geqslant 0, \tag{86}$$

$$D_2=Q_2^2-4H(\varepsilon_1+\varepsilon_2v_o+\varepsilon_3(\emptyset-\varepsilon_2))\geqslant 0, \tag{87}$$

$$\delta_2=\frac{Q_2-\sqrt{D_2}}{2H}\leqslant \mu^*-\mu_o \ , \tag{88}$$

then the iterative algorithm (70) is well defined and for each $n\in\mathbb{Z}_+$ we shall have the estimates

$$\|\tilde{x}_n-x_n\|\leqslant t_n^{(2)}\leqslant\delta_2 \ , \tag{89}$$

where the sequence $(t_n^{(2)})_{n\geqslant 0}$ is given by:

$$t_{-1}^{(2)}=t_o^{(2)}=0, \quad s_{-1}^{(2)}=\frac{1+h_oq_o}{2h_o} \ , \ s_o^{(2)}=\frac{1-h_oq_o}{2h_o} \ , \ w_{-1}=q_o \ ,$$

$$s_{n+1}^{(2)}=\frac{s_n^{(2)}s_{n-1}^{(2)}+a^2}{s_n^{(2)}+s_{n-1}^{(2)}}, \quad w_n=s_n^{(2)}-s_{n+1}^{(2)}$$

$$t_{n+1}^{(2)}=\frac{1}{\emptyset-\varepsilon_2}\left[Ht_n^{(2)}t_{n-1}^{(2)}+(H(w_n+w_{n-1})+\varepsilon_2)t_n^{(2)}+Hw_nt_{n-1}^{(2)}+\varepsilon_1+\varepsilon_2w_n+\varepsilon_3(\emptyset-\varepsilon_2)\right]$$

Proof. For n=-1 and n=0 the inequalities (89) are trivially satisfied. Suppose they are satisfied for $n=-1,0,1,\ldots,k$, where $k\geqslant 0$.

From (89) it follows that $\tilde{x}_{k-1}, \tilde{x}_k\in U^*$. In this case condition (C^{**}) implies, according to Lemma 3, the invertibility of the linear operator $\delta f(\tilde{x}_{k-1},\tilde{x}_k)+E_k$, as well as the inequality

$$\|(\delta f(\tilde{x}_{k-1},\tilde{x}_k)+E_k)^{-1}\| \leqslant (\emptyset-\varepsilon_2)^{-1} . \tag{90}$$

Let us denote $D_k=\delta f(x_{k-1},x_k)$ and $\tilde{D}_k=\delta f(\tilde{x}_{k-1},\tilde{x}_k)$. Using (25) and (70) we may write:

$$\tilde{x}_{k+1}-x_{k+1}=(\tilde{D}_k+E_k)^{-1}\Big[f(x_k)-f(\tilde{x}_k)-\tilde{D}_k(x_k-\tilde{x}_k)+(\tilde{D}_k-D_k)D_k^{-1}f(x_k)+ \\ +E_kD_k^{-1}f(x_k)-E_k(x_k-\tilde{x}_k)-e_k\Big]+g_k \; . \tag{91}$$

Taking into account the proof of Lemma 2 it follows that $w_k=\omega_2^{(k)}(q_o,r_o)$ and then from the proof of Theorem 3 we have $\|D_k^{-1}f(x_k)\| = \|x_{k+1}-x_k\| \leqslant w_k$. Using this remark and inequalities (74) and (28) we obtain the following estimates:

$$\|f(x_k)-f(\tilde{x}_k)-\tilde{D}_k(x_k-\tilde{x}_k)\| \leqslant H(t_k^{(2)}+w_{k-1})t_k^{(2)},$$

$$\|(\tilde{D}_k-D_k)D_k^{-1}f(x_k)\| \leqslant H(t_{k-1}^{(2)}+t_k^{(2)})w_k$$

From (91) we can now deduce that

$$\|\tilde{x}_{k+1}-x_{k+1}\| \leqslant (\emptyset-\varepsilon_2)^{-1}\Big[Ht_k^{(2)}t_{k-1}^{(2)}+(H(w_k+w_{k-1})+\varepsilon_2)t_k^{(2)}+ \\ +Hw_kt_{k-1}^{(2)}+\varepsilon_1+\varepsilon_2w_k+\varepsilon_3(\emptyset-\varepsilon_2)\Big]=t_{k+1}^{(2)} \; .$$

Thus we have checked the first inequality (89) for n=k+1.

Remarking that

$$\omega(q,r)\leqslant\frac{r(q+r)}{r+2\sqrt{r(q+r)}}\leqslant\frac{2}{1+2\sqrt{2}}\max(q,r) ,$$

we obtain the inequalities $w_o\leqslant r_o$, $w_1\leqslant\frac{2}{1+2\sqrt{2}}v_o$, $w_2\leqslant\frac{2}{1+2\sqrt{2}}v_o$.

It follows that

$$(H(w_{k-1}+w_k)+\varepsilon_2)t_k^{(2)}+Hw_k t_{k-1}^{(2)} \leqslant (\frac{6v_oH}{1+2\sqrt{2}}+\varepsilon_2)\delta_2 .$$

(Note: for k=0 and k=-1 one has used the fact that $t_o=t_{-1}=0$). Denoting $B=(6v_oH)/(1+2\sqrt{2})+\varepsilon_2$, $C=\varepsilon_1+\varepsilon_2 v_o+\varepsilon_3(\emptyset-\varepsilon_2)$ and taking into account the definition of δ_2 it follows that $t_{k+1}^{(2)} \leqslant (\emptyset-\varepsilon_2)^{-1}(H\delta_2^2+B\delta_2+C)=\delta_2$.

In this way we have proved that the second inequality (89) is also satisfied for n=k+1. The proof is complete. ■

The function $\beta_2(q,r)=\sqrt{r(q+r)+a^2}-a$ is increasing so that for the perturbed secant method we can obtain aposteriori estimates of the form (8). Thus we have the following

Corollary. Under the hypotheses of Theorem 4 the following estimates

$$\|\tilde{x}_n-x^*\| \leqslant s_n^{(2)}+t_n^{(2)}-a \tag{92}$$

$$\|\tilde{x}_n-x^*\| \leqslant \Big[(\|\tilde{x}_n-\tilde{x}_{n-1}\|+t_n^{(2)}+t_{n-1}^{(2)})(\|\tilde{x}_n-\tilde{x}_{n-1}\|+\|\tilde{x}_{n-1}-\tilde{x}_{n-2}\|+t_n^{(2)}+2t_{n-1}^{(2)}+t_{n-2}^{(2)}) +a^2\Big]^{1/2}+t_n^{(2)}-a \tag{93}$$

are satisfied for all $n \in \mathbb{Z}_+$. ■

In the end of this paper we shall apply Theorem 4 to an "ill conditioned" example proposed by Wilkinson [21] and considered also by Lancaster [2]. We are asked to solve iteratively the equation

$$x^2-2.0288888+1.028769=0$$

using a computer characterized by the accuracy $\varepsilon_1=\varepsilon_2=\varepsilon_3=0.5x10^{-7}$.

Starting with $\xi_{-1}=1.21$, $\xi_o=1.2$ and using the secant method we have obtained :

$\xi_1=1.1105182$

$\xi_2=1.0789103$

$\xi_3=1.0550694$

$\xi_4=1.0424910$

$\xi_5=1.0358181$

$\xi_6=1.0332202$

$\xi_7=1.0326199$

$\xi_8=1.0325685$

$\xi_n=\xi_8$ for $n\geqslant 8$.

If we take $x_{-1}=\xi_6$, $x_o=\xi_7$, and $\mu_o=\mu^*=0.0016199$, then we obtain H=1, $\emptyset=0.0331112$, $q_o=0.0006004$, $r_o=0.0000517$, $t_1^{(2)}<15,7x10^{-7}$, $\delta_2<16,1x10^{-7}$.

We want to find an estimate for the distance $|\xi_8-x^*|$. The hypotheses of Theorem 4 being satisfied we can use Corrolary 1. From (92) it follows that $|\xi_8-x^*|<25x10^{-7}$, while from (93) we have $|\xi_8-x^*|<25.3x10^{-7}$.

Taking advantage of the fact that we know that the sequence (ξ_n) becomes constant beginning with n=8, we obtain according to (6) that $|\xi_8-x^*|<16,1x10^{-7}$. This is very closed to the reality because $x^*=1.0325673...$

REFERENCES

1. BURMEISTER, W., Inversionsfreie Verfahren zur Lösung nichtlinearer Operatorgleichungen, ZAMM, 52 (1972), 101-110.

2. LANCASTER, P., Error analysis for the Newton-Raphson method.Numer.Math., 9 (1966), 55-68.

3. MIEL, G.J., Unified error analysis for Newton-type methods. Numer.Math., 33 (1979), 391-396.

4. ORTEGA, J.M. and RHEINBOLDT, W.C., Iterative solution of nonlinear equations in several variables, Academic Press, New York and London, 1970.

5. OSTROWSKI, A.M. Solution of equations in Euclidian and Banach Spaces. Academic Press, New York and London, 1973.

6. POTRA, F.-A., On a modified secant method. L'Analyse numérique et la théorie de l'approximation, 8, 2 (1979), 203-214.

7. POTRA, F.-A., An application of the induction method of V.Pták to the study of Regula Falsi. Preprint INCREST no.11/1979 (to appear in Aplikace Matematiky).

8. POTRA, F.-A. and PTÁK, V., Sharp error bounds for Newton's process. Numer.Math., 34 (1980), 63-72.

9. POTRA, F.-A. and PTAK, V., On a class of modified Newton processes. Num.Func.Anal.Optim., 2, 1(1980) 107-120.

10. POTRA, F.-A. and PTÁK, V., A generalization of Regula Falsi. Preprint INCREST no.10/1980 (to appear in Numer.Math).

11. POTRA, F.A. and PTAK, V., Nondiscrete induction and iterative procedures (to appear).

12. PTÁK, V., Nondiscrete mathematical induction and iterative existence proofs. Linear algebra and its applications, 13 (1976), 223-236.

13. PTÁK, V., Nondiscrete mathematical induction, in: General Topology and its Relations to Modern Analysis and Algebra IV, pp.166-178. Lecture Notes in Mathematics 609, Springer, Berlin 1977.

14. ROKNE, J., Newton's method under mild differentiability conditions with error analysis. Numer.Math., 18 (1972), 401-412.

15. SCHMIDT, J.W., Eine Übertragung der Regula Falsi auf Gleichungen in Banach Räumen, I, II, ZAMM, 43 (1963), 1-8, 97-110.

16. SCHMIDT, J.W., Regula-Falsi Verfahren mit konsistenter Steigung und Majoranten Prinzip. Periodica Matematica Hungarica, 5, 3 (1974), 187-193.

17. SCHWETLICK, H., Numerische Lösung nichtlinearer Gleichungen, VEB, DVW, Berlin 1979.

18. SERGEEV, A.S., On the method of chords (Russian), Sibir. Mat.J. 2, 2 (1961), 282-289.

19. ULM, S., Majorant principle and secant method (Russian), I.A.N. Estonskoi S.S.R., fiz.mat., 3 (1964), 217-227.

20. ULM, S., On generalized divided differences (Russian), I, II, I.A.N. Estonskoi SSR, Fiz.mat., 12 (1967), 13-26, 146-156.

21. WILKINSON, J.H., The Algebraic eigenvalue problem. Oxford: Clarendon Press 1965 .

22. WOZNIAKOWSKI, H., Numerical stability for solving non-linear equations. Numer.Math. 27 (1977), 373-390.