INSTITUTUL
DE
MATEMATICA

INSTITUTUL NATIONAL
PENTRU CREATIE
STIINTIFICA SI TEHNICA

ISSN 0250-3638

A GENERALISATION OF REGULA FALSI

by

Florian A.POTRA and Vlastimil PTÁK

PREPRINT SERIES IN MATHEMATICS

No.10/1980

BUCURESTI

A GENERALISATION OF REGULA FALSI

by

Florian A.POTRA*) and Vlastimil PTÁK**)

March 1980

*) *Department of Mathematics, National Institute for Scientific and Technical Creation, Bd.Pacii 220, 77538 Bucharest, Romania*

**) *Institute of Mathematics, Czechoslovak Academy of Sciences, Zitna 25, 11567 Praha 1, Czechoslovakia*

A GENERALISATION OF REGULA FALSI

by

Florian A.POTRA

and

Vlastimil PTÁK

Summary. The method of nondiscrete mathematical induction is applied to a multistep variant of the secant method. Optimal conditions for convergence as well as error estimates, sharp in every step, are obtained.

Subject classifications. AMS (MOS): 65H10; CR: 5.15

1. Introduction

Although less fast than the Newton method the secant method is - in some cases - more convenient because it does not involve the calculation of the derivative. Using the notion of the divided difference of an operator (see definition (3.1) below) A.Sergeev [11] and J.Schmidt [12] have extended the secant method to the case of nonlinear equations in Banach spaces. Let $\mathcal{E}$ and $\mathcal{F}$ be two Banach spaces and let f be a (nonlinear) operator with domain of definition in $\mathcal{E}$ and with values in $\mathcal{F}$. If x_0 and x_{-1} are two given points in the domain of f, the generalized secant method consists in the following algorithm

$$x_{n+1}=x_n-[x_{n-1}, x_n; f]^{-1}f(x_n), \quad n=0,1,2,... \qquad (1)$$

where $[x_{n-1}, x_n; f]$ denotes a divided difference of the operator f in the points x_{n-1} and x_n. In the papers mentioned above sufficient conditions are given for the convergence of x_n to a root x^* of the equation $f(x)=0$. These results were improved by S.Ulm [15]. The conditions imposed by him to the initial data are in some sense the best possible (see Proposition (3.6) below). However his hypotheses imply the symmetry of the divided difference (i.e. $[x,y;f]=[y,x;f]$ for all x and y). This symmetry property, which is rather restrictive, was also supposed by Sergeev but it does not follow from the hypotheses of Schmidt (see also Helfrich [1]). In the above mentioned paper S.Ulm investigates also a procedure of the form

$$x_{n+1}=x_n-[x_{-1}, x_o; f]^{-1} f(x_n), \qquad n=0,1,2,... \tag{2}$$

which is slower than the preceding one but does not require the inversion of a linear operator at each step.

A procedure intermediate between these two consists in fixing a natural number m and keeping the same linear operator for sections of the process consisting of m steps each. It may be described as follows: given two points $x_o=x_o^m$ and $y_o=x_o^{m-1}$, construct m sequences $(x_n^j)_{n\geq 1}$, $1\leq j\leq m$ by the algorithm:

$$\begin{aligned} x_{n+1}^o &= x_n^m \\ x_{n+1}^j &= x_{n+1}^{j-1}-[x_n^{m-1}, x_n^m; f]^{-1} f(x_{n+1}^{j-1}), \quad j=1,2,...,m \\ & \qquad\qquad\qquad\qquad\qquad\qquad n=0,1,2,... \end{aligned} \tag{3}$$

J.W. Schmidt and H.Schwethick [13] have shown that the order of convergence of this procedure is equal to $\frac{1}{2}(m+\sqrt{m^2+4})$. In the

particular case m=2, P.Laasonen [3] was able to obtain a more precise result concerning the sufficient conditions for convergence. Optimal conditions for convergence as well as sharp error bounds were given later in [6].

In the present note we apply the method of nondiscrete matematical induction to the study of the iterative procedure (3). In the particular case m=1 the present results contain those of [5] and for m=2 the results from [6].

2. The application of the method of nondiscrete induction to the study of a class of iterative procedures

The method of nondiscrete mathematical inductive was developped over a number of years in a series of papers; the general principles of its application are explained in the Gathinburg Lecture [9] or in the survey [10]. Approximate sets depending on a two-dimensional parameter were first considered in [5]. The corresponding rate of convergence was one of type (2.1), The natural rate of convergence for [6] was of type (2.2). In the present paper we work with a rate of convergence of type (2,m) which constitutes a generalisation of the notions mentioned above.

Let T be either the set of all positive real numbers or an open interval $(0,s_o)$ for some $s_o>0$. Further let m be a fixed positive integer and let ω be a mapping of T^2 into T^m; its components will be denoted by $\omega_1,\omega_2,\ldots,\omega_m$ so that

$$\omega(s)=(\omega_1(s),\omega_2(s),\ldots,\omega_m(s)) \quad \text{for each } s=(q,r)\in T^2$$

It will be convenient to introduce also the functions ω_{-1} and ω_o by the formulae

$$\omega_{-1}(s)=\omega^{(o)}_{m-1}(s)=q,\ \omega_o(s)=\omega^{(o)}_m(s)=r;\quad s=(q,r)\in T^2 \qquad (4)$$

Let us define the functions $\omega^{(n)}_k : T^2 \longrightarrow T$ by the reccursive formula

$$\omega^{(n+1)}_k(s)=\omega_k(\omega^{(n)}_{m-1}(s),\ \omega^{(n)}_m(s)),\quad k=-1,0,1,\ldots,m \qquad (5)$$
$$n=0,1,2,\ldots$$

We shall attach to the mapping $\omega : T^2 \longrightarrow T^m$, the mapping $\bar{\omega} : T^2 \longrightarrow T^2$ defined by

$$\bar{\omega}(s)=(\omega_{m-1}(s),\ \omega_m(s)) \qquad (6)$$

If we denote by $\omega^{(n)}$ the n'th iterate of ω in the sense of the usual composition of functions (i.e. $\bar{\omega}^o(s)=s$, $\bar{\omega}^{(n+1)}(s)=\bar{\omega}(\bar{\omega}^{(n)}(s))$, $n=0,1,2,\ldots$) then we have obviously

$$\bar{\omega}^{(n)}(s)=(\omega^{(n)}_{m-1}(s),\ \omega^{(n)}_m(s)),\ \text{for all } s\in T^2 \text{ and } n=0,1,2,\ldots$$

Considering now for each $n=1,2,\ldots$ the mapping $\omega^{(n)} : T^2 \longrightarrow T^m$ with components $\omega^{(n)}_1, \omega^{(n)}_2, \ldots, \omega^{(n)}_m$, it follows that

$$\omega^{(n+1)}(s)=\omega(\bar{\omega}^{(n)}(s)),\quad \text{for all } s\in T^2 \text{ and } n=0,1,2,\ldots$$

In what follows we shall omit the brackets or the sign "$\circ$" for indicating the composition of functions . For example we shall simply write $\omega\bar{\omega}^{(n)}(s)$ instead of $\omega(\bar{\omega}^{(n)}(s))$ or $\omega\circ\bar{\omega}^{(n)}(s)$.

(2.1). Definition. The function $\omega: T^2 \longrightarrow T^m$ with the law of iteration described above will be called a rate of convergence of type (2,m) on T if the series

$$G(s)=\sum_{n=1}^{\infty} \sum_{j=0}^{m-1} \omega_j^{(n)}(s) \tag{9}$$

is convergent for each $s \in T^2$.

Since $\omega_o^{(n+1)} = \omega_m^{(n)}$ for all n=0,1,..., the above expression for G may be replaced by the following one

$$G(s)=r+\sum_{n=1}^{\infty} \sum_{k=1}^{m} \omega_k^{(n)}(s) \quad , \quad s=(q,r) \in T^2$$

It will be convenient to introduce the functions $G_o, G_1, \ldots, G_m$ by setting

$$\begin{aligned} G_o &= G \\ G_k &= G-(\omega_o+\ldots+\omega_{k-1}) \quad \text{if} \quad 1 \leq k \leq m \ . \end{aligned} \tag{10}$$

In the sequel we intend to show how the notion of a rate of convergence of type (2,m) may be used in the study of a class of iterative procedures.

Let X be a complete metric space. If k is a natural number, X^k will stand for the cartesian product of k copies of X. In the whole paper m will be a fixed positive integer; the elements of X^{m+1} will be finite sequences of the form $z=(z_o, z_1, \ldots, z_m)$ with $z_j \in X$. For each j=0,1,...,m we denote by P_j the mapping which assingns to each $z \in X^{m+1}$ its j-th coordinate; thus

$$z=(P_0z,\ P_1z,\ldots,P_mz)$$

We shall also use the mapping P from X^{m+1} onto X^2 defined by

$$Pz=(P_{m-1}z,\ P_mz)$$

Let $\mathfrak{D}_F$ be a subset of X^2 and let F be a mapping of $\mathfrak{D}_F$ into X^{m+1}. To simplify some of the formulae it will be convenient to use the abbreviations

$$F_j=P_jF\quad,\qquad j=0,1,\ldots,m$$

and to introduce the mapping $F_{-1}:X^2\longrightarrow X$ defined for $u=(y,x)$ by the formula

$$F_{-1}u=y$$

Let G be a mapping from $\mathfrak{D}_F$ into X^m and let F be the mapping from $\mathfrak{D}_F$ into X^{m+1} defined by setting $F(y,x)=(x,G(y,x))$. Then for every $z\in P^{-1}\mathfrak{D}_F$ we have $P_0FPz=P_mz$.

For our purposes the following particular form of the induction theorem is most suitable:

(2.2) Lemma. Let F be a mapping from $\mathfrak{D}_F\subset X^2$ into X^{m+1} satisfying the condition

$$P_0FPz=P_mz\quad,\qquad \text{for } z\in P^{-1}\mathfrak{D}_F \tag{12}$$

Let Z be a mapping which assigns to each $t\in T^2$ a set $Z(t)\subset\mathfrak{D}_F$.

Let ω be a rate of convergence of type (2,m) on T.

Let $u_o \in \mathcal{D}_F$ and $t_o \in T^2$ be given.

If the following conditions are fulfilled:

$$u_o \in Z(t_o) \tag{13}$$

$$PFZ(t) \subset Z\,\overline{\omega}(t) \tag{14'}$$

$$d(F_k u,\ F_{k+1} u) \leq \omega_k(t) \tag{14"}$$

for all $t \in T^2$, $u \in Z(t)$ and $k=-1,0,\ldots,m-1$, then:

1^o the iterative procedure

$$\begin{aligned} x_1 &= Fu_o \\ x_{n+1} &= FPx_n , \qquad n=1,2,\ldots \end{aligned} \tag{15}$$

yields a sequence $(x_n)_{n\geq 1}$ of points of $P^{-1}\mathcal{D}_F$;

2^o there exists a point $x^* \in X$ such that each of the m+1 sequences $(P_j x_n)_{n\geq 1}$, $0 \leq j \leq m$, converges to x^* ;

3^o the following relations hold for each n=1,2,3,...

$$Px_n \in Z\,\overline{\omega}^{(n)}(t_o) \tag{16}$$

$$d(P_k x_n\ ,\ P_{k+1} x_n) \leq \omega_k^{(n)}(t_o) , \quad 0 \leq k \leq m-1 \tag{17}$$

$$d(P_k x_n\ ,\ x^*) \leq \sigma_k\,\overline{\omega}^{(n-1)}(t_o), \quad 0 \leq k \leq m \tag{18}$$

4^o suppose that, for some natural number n, we have

$$Px_{n-1} \in Z(d_n) \tag{19}$$

where $d_n = (d(P_{m-1}x_{n-1}\ ,\ P_m x_{n-1}),\ d(P_m x_{n-1}\ ,\ P_1 x_n)) \in T^2$ and $Px_o = u_o$;

Then:

$$d(P_k x_n , x^*) \leq \sigma_k(d_n) , \qquad 0 \leq k \leq m \tag{20}$$

Proof. Let n be given and suppose that the point x_n has already been obtained via the iterative procedure. Then (16) implies $x_n \in P^{-1} Z\, \bar{\omega}^{(n)}(t_o) \subset P^{-1}\mathcal{D}_F$ so that we may apply (15) to obtain a new point x_{n+1} which in its turn will belong to $P^{-1}\mathcal{D}_F$. Thus part 1° of the lemma is a consequence of (16). In the sequel we shall prove the relations (16)-(18).

It follows from (14') that

$$Px_1 = PFu_o \in PFZ(t_o) \subset Z\,\bar{\omega}(t_o)$$

so that (16) holds for n=1. Assuming (16) to be true for n, we have

$$Px_{n+1} = PFPx_n \in PFZ\,\bar{\omega}^{(n)}(t_o) \subset Z\,\bar{\omega}\bar{\omega}^{(n)}(t_o) = Z\,\bar{\omega}^{(n+1)}(t_o) .$$

(17) is also true for n=1 because according to (14") and (15)

$$d(P_k x_1 , P_{k+1} x_1) = d(F_k u_o , F_{k+1} u_o) \leq \omega_k(t_o)$$

If (17) holds for a certain $n \geq 1$ then, applying (14") for $u = Px_n$ and $t = \bar{\omega}^{(n)}(t_o)$, we obtain

$$d(P_k x_{n+1} , P_{k+1} x_{n+1}) = d(F_k Px_n , F_{k+1} Px_n) \leq \omega_k \bar{\omega}^{(n)}(t_o) = \omega_k^{(n+1)}(t_o)$$

To prove the estimates (18) we observe first that relation (12) implies $P_m x_n = P_o x_{n+1}$ for all n. It follows that:

$$d(P_m x_n , P_m x_{n+p}) \leq \sum_{j=1}^{p} \sum_{k=0}^{m-1} d(P_k x_{n+j} , P_{k+1} x_{n+j}) \leq \sum_{j=1}^{p} \sum_{k=0}^{m-1} \omega_k^{(n+j)}(t_o) ;$$

the rest follows if we allow p to tend to infinity.

In this manner the first three parts of the lemma are established. In particular, for n=1, the estimate (18) assumes the following form

$$\text{if } u_o \in Z(t_o) \text{ then } d(P_k F u_o, x^*) \leq \sigma_k(t_o), \quad \text{for } k=0,1,\ldots,m$$

To prove 4^o suppose n is a natural number for which (19) is satisfied. Replacing in the implication above u_o , t_o respectively by Px_{n-1} , d_n we obtain (20). The proof is complete.

In what follows we shall construct a rate of convergence of type (2,m) which will be then used in the study of the iterative procedure (3).

There are some differences between the cases m=1 and $m \geq 2$ but we can study them together if we make the following convention: if an algorithm requires, at a certain stage, the computation of a quantity Q_k for k=0,1,...,p, and if p happens to be negative, ignore this instruction and pass to the next one; in the same sense the sum $a_o+a_1+\ldots+a_p$ will be taken equal to zero if p is negative.

(2.3) Lemma. Let T denote the set of all positive real numbers, let a be a nonnegative real number and let m be a positive integer. For all $q,r \in T$ consider the functions:

$$\varphi(q,r)=r+\sqrt{r(q+r)+a^2} \quad , \tag{21}$$

$$\omega_{-1}(q,r)=q, \quad \omega_o(q,r)=r \quad , \tag{22}$$

and define

$$\omega_{k+1}=\frac{\omega_k(\omega_{-1}+\omega_k+2(\omega_o+\ldots+\omega_{k-1}))}{2\varphi+\omega_{-1}}\ ,\quad k=0,1,\ldots,m-2 \qquad (23)$$

$$\omega_m=\frac{\omega_{m-1}(\omega_{-1}+\omega_{m-1}+2(\omega_o+\ldots+\omega_{m-2}))}{2\varphi-2(\omega_o+\ldots+\omega_{m-2})-\omega_{m-1}}\ . \qquad (24)$$

Then the function $\omega=(\omega_1,\omega_2,\ldots,\omega_n)$ is[2] rate of convergence of type (2,m) and the corresponding σ-function is given by:

$$\sigma(q,r)=r+\sqrt{r(q+r)+a^2}-a \qquad (25)$$

Proof. Consider the real polynomial $f(x)=x^2-a^2$.

For any positive numbers q and r set:

$$x_o=x_o^m=\varphi(q,r)\ ,\quad y_o=x_o^{m-1}=\varphi(q,r)+q\ .$$

The iterative procedure (3) reduces in this particular case to the scheme

$$x_{n+1}^o=x_n^m$$
$$x_{n+1}^{k+1}=x_{n+1}^k-\frac{f(x_{n+1}^k)}{x_n^{m-1}+x_n^m}\ ,\quad k=0,1,2,\ldots,m-1,\ n=0,1,2,\ldots$$

From the convexity of f it follows that

$$x_{n+1}^m<x_{n+1}^{m-1}<\ldots<x_{n+1}^1<x_{n+1}^o=x_n^m\ . \qquad (35)$$

By definition we have

$$x_o^{m-1}-x_o^m=q=\omega_{-1}(q,r)$$

and by direct calculation we obtain

$$x_1^o - x_1^1 = x_o^m - x_1^1 = \frac{f(x_o^m)}{x_o^{m-1} + x_o^m} = r = \omega_o(q,r)$$

For $k=0,1,\ldots,m-1$ set by definition

$$\omega_k(q,r) = x_1^k - x_1^{k+1} = \frac{f(x_1^k)}{x_o^{m-1} + x_o^m} \tag{37}$$

and, finally, set

$$\omega_m(q,r) = x_2^1 - x_1^m = \frac{f(x_1^m)}{x_1^{m-1} + x_1^m} \tag{38}$$

Equalities (37) imply that for every $k=1,2,\ldots,m$ the following relation is satisfied:

$$x_1^k = \varphi - (\omega_o + \ldots + \omega_{k-1})$$

According to our convention the above relation is also true for $k=0$, because in this case it reduces to $x_1^o = x_o^m = \varphi$. Thus we may write for all $k=0,1,\ldots,m-1$:

$$f(x_1^{k+1}) = f(x_1^k - \omega_k) = (x_1^k)^2 - 2x_1^k\omega_k + \omega_k^2 - a^2 =$$

$$= \omega_k^2 - 2x_1^k\omega_k + f(x_1^k) = \omega_k^2 - 2x_1^k\omega_k + \omega_k(x_o^{m-1} + x_o^m) =$$

$$= \omega_k^2 - 2x_1^k\omega_k + \omega_k(2\varphi + \omega_{-1}) = \omega_k(\omega_k + \omega_{-1} + 2(\omega_o + \ldots + \omega_{k-1}))$$

Hence we get the formulae

$$\omega_{k+1} = \frac{f(x_1^{k+1})}{x_o^{m-1} + x_o^m} = \frac{\omega_k(\omega_k + \omega_{-1} + 2(\omega_o \ldots + \omega_{k-1}))}{2\varphi + \omega_{-1}} \qquad k=0,1,\ldots,m-2$$

$$\omega_m = \frac{f(x_1^m)}{x_1^m + x_1^{m-1}} = \frac{\omega_{m-1}(\omega_{m-1}+\omega_{-1}+2(\omega_o+\ldots+\omega_{m-2}))}{2\varphi - 2(\omega_o+\ldots+\omega_{m-2}) - \omega_{m-1}}$$

which are exactly the formulae (23) and (24).

The fact that the function $\omega=(\omega_1,\ldots,\omega_m)$ defined as above constitutes a rate of convergence of type (2,m) follows from the monotone convergence of the sequences $(x_n^k)_{n\geq 1}$ (see (35)). In fact we have

$$\omega_k^{(n)}(q,r) = x_n^k - x_n^{k+1} \quad ,$$

for all k=0,1,...,m-1 and n=1,2,... and

$$G(q,r) = x_o^m - a = \varphi(q,r) - a \quad \blacksquare$$

We shall use the above two lemmas in the proof of the main theorem of the next section.

3. Convergence conditions and error estimates

The generalisation of the secant method for solving nonlinear equations in Banach spaces is based on the notion of divided difference of an operator, notion introduced by J.Schröder [14] . If $\mathcal{E}$ and $\mathcal{F}$ are two Banach spaces we denote by $L(\mathcal{E},\mathcal{F})$ the space of all linear and bounded operators defined on $\mathcal{E}$ and with values in $\mathcal{F}$

(3.1) Definition. Let f be a nonlinear operator defined on a subset $\mathcal{D}$ of the Banach space $\mathcal{E}$ and with values in the Banach space $\mathcal{F}$. If x and y are two destinct points of $\mathcal{D}$ we call a divided difference of the operator f on the points x and y a bounded

linear operator $[x, y; f] \in L(\mathcal{E}, \mathcal{F})$ which satisfies the condition

$$[x, y; f](x-y) = f(x) - f(y) \tag{40}$$

Of course the above requirement does not determine the divided difference uniquely, except in case $\mathcal{E}$ has dimension one.

In many important particular cases, concrete methods for constructing such divided differences are known (see [12] and [16])

(3.2) Theorem. Let $\mathcal{E}$ and $\mathcal{F}$ be two Banach spaces and x_o a given point of $\mathcal{E}$. Let μ be a positive number and let U be the open ball $U = \{x \in \mathcal{E} ; \| x - x_o \| < \mu\}$. Let f be a mapping defined and continuous on the closure of U and with values in $\mathcal{F}$. Suppose that for each pair of distinct points x and y in U, a divided difference $[x,y;f]$ is given. Furthermore suppose there exists a point $y_o \in U$ such that the linear operator $D_o = [y_o, x_o; f]$ is invertible and that

$$\| D_o^{-1} ([x,y;f] - [x',y';f]) \| \leq h_o (\| x-x' \| + \| y-y' \|) \tag{42}$$

for all $x, y, x', y' \in U$ with $x \neq y$ and $x' \neq y'$.

If the following conditions are satisfied:

$$\| x_o - y_o \| \leq q_o \quad , \quad \| D_o^{-1} f(x_o) \| \leq r_o \ , \tag{43}$$

$$h_o q_o + 2 \sqrt{h_o r_o} \leq 1 \ , \tag{44}$$

$$\mu \geq \frac{1}{2h_o} (1 - h_o q_o - \sqrt{(1-h_o q_o)^2 - 4 h_o r_o}) = G(q_o, r_o) \tag{45}$$

then the iterative procedure (3) with starting points

$x_o^{m-1}=y_o$, $x_o^m=x_o$ yields m+1 sequences $(x_n^j)_{n\geq 1}$, $(0\leq j\leq m)$ with the following properties: there exists a point $x^*\in U$ for which $f(x^*)=0$, each of these sequences converges to x^*, and the following estimates

$$\|x_n^j-x^*\| \leq G_j\, \bar{\omega}^{(n-1)}(q_o,r_o) \tag{46}$$

$$\|x_n^j-x^*\| \leq G_j(\|x_{n-1}^{m-1}-x_{n-1}^m\|\ ,\ \|x_n^o-x_n^1\|) \tag{47}$$

hold for each $j=0,1,\ldots,m$ and $n=1,2,3,\ldots$, where ω is the rate of convergence defined in Lemma (2.3), the constant a being given by

$$a=\frac{1}{2h_o}\sqrt{(1-h_oq_o)^2-4h_or_o} \tag{48}$$

Proof. Let us first remark that (42) implies that for each $x\in U$ we have $\lim_{x';y'\to x}[x',y';f]=f'(x)$ where $f'(x)$ denotes the Fréchet derivative of f at the point x. Thus, setting for each $x\in U$ $[x,x;f]=f'(x)$ we may assume that (42) holds for all $x,y,x',y'\in U$.

If $u=(y,x)\in U^2$ set

$$\begin{aligned} F_o(u)&=x \\ F_{j+1}(u)&=F_j(u)-[y,x;f]^{-1}f(F_j(u)) \quad \text{for } j=0,1,\ldots,m-1 \end{aligned} \tag{49}$$

Let us denote by $\mathcal{D}_F$ the set of those u for which the above formulae make sense (i.e. $[y,x;f]$ is invertible and $F_j(u)\in U$ for $j=0,1,\ldots,m-1$) and let us define a mapping $F:\mathcal{D}_F\to\mathcal{E}^{m+1}$ by setting

$$F(u)=(F_o(u),\ F_1(u),\ldots,F_m(u))$$

This function clearly satisfies the properties

$$P_oFPz=P_mz\ ,\ P_kFu=F_ku \quad \text{for all } z\in P^{-1}\mathcal{D}_F \text{ and } u\in\mathcal{D}_F$$

It will be convenient to introduce a mapping F_{-1} as well, by setting $F_{-1}(u)=y$.

The proof will be based on lemma (2.2). To this end we assign to each $t=(q,r)\in T^2$ a subset of $\mathcal{E}^2$ defined as follows

$$Z(t)=\{(y,x)\in\mathcal{E}^2;\ y\in U,\ \|y-x\|\leq q,\ \|y-y_o\|\leq \mathcal{G}(t_o)-\mathcal{G}(t)+q_o-q, \quad (50)$$

$$\|x-x_o\|\leq \mathcal{G}(t_o)-\mathcal{G}(t),\ D=[y,x;f] \text{ is invertible and } \|D^{-1}f(x)\|\leq r\}$$

In the above definition of $Z(t)$, t_o stands for the pair (q_o,r_o). Hence using (45) it follows that $Z(t)\subset U^2$. Consider now the rate of convergence ω described in lemma (2.3), the constant a being given by (48). Our theorem will be proved if we show that $Z(t)\subset\mathcal{D}_F$ and that the conditions (13), (14) and (19) from lemma (2.2) are satisfied. First of all, if u_o stands for (y_o,x_o) we clearly have $u_o\in Z(t_o)$. Let us prove now that $u\in Z(t)$ implies

$$F_k(u)\in U \quad \text{for } -1\leq k\leq m \qquad (51')$$

and

$$\|F_k(u)-F_{k+1}(u)\|\leq \omega_k(t) \quad \text{for } -1\leq k\leq m-1 \qquad (51")$$

For $k=-1$ these relations reduce to $y\in U$ and $\|y-x\|\leq q$; for $k=0$ they follow from $x\in U$ and $\|[y,x;f]^{-1}f(x)\|\leq r$.

Consider now an i, $0\leq i\leq m-1$, and suppose that (51') and (51") hold for $k=-1,0,\ldots,i$. We have then:

$$\|F_{i+1}(u)-x_o\|\leq \|F_{i+1}(u)-x\|+\|x-x_o\|\leq \sum_{j=0}^{i}\|F_{j+1}(u)-F_j(u)\|+\|x-x_o\|\leq \quad (53)$$

$$\leq \sum_{j=0}^{i}\omega_j(t)+\mathcal{G}(t_o)-\mathcal{G}(t)=\mathcal{G}(t_o)-\mathcal{G}_{i+1}(t)$$

so that $F_{i+1}(u)\in U$ as well; this establishes (51'). Let us remark that from (50) and (51') it follows that $Z(t)\subset \mathcal{D}_F$. To simplify the formulae let $D=[y,x;f]$, $D_{i+1}=[F_{i+1}(u), F_i(u);f]$ and let f_j stand for the value $fF_j(u)$. The relation defining $F_{i+1}(u)$ may thus be rewritten in the form $f_i=D(F_i(u)-F_{i+1}(u))$. Now

$$F_{i+1}(u)-F_{i+2}(u)=D^{-1}f_{i+1}=D^{-1}(f_{i+1}-f_i-D(F_{i+1}(u)-F_i(u))= \tag{54}$$

$$=D^{-1}(D_{i+1}-D)(F_{i+1}(u)-F_i(u))=(1-D_o^{-1}(D_o-D))^{-1}D_o^{-1}(D_{i+1}-D)(F_{i+1}(u)-F_i(u))$$

provided $\| D_o^{-1}(D_o-D)\| < 1$. This is true, however, since - according to (42) - we have

$$\| D_o^{-1}(D_o-D)\| \leq h_o(\|y-y_o\|+\|x-x_o\|) \leq h_o(2\sigma(t_o)-2\sigma(t)+q_o-q)=$$

$$=1-h_o(2\varphi(t)+q)<1.$$

This estimate and another application of (42) yield

$$\| F_{i+1}(u)-F_{i+2}(u) \| \leq \frac{1}{h_o(2\varphi(t)+q)} \| D_o^{-1}(D_{i+1}-D) \| \, \| F_{i+1}(u)-F_i(u) \| \leq$$

$$\leq \frac{1}{h_o(2\varphi(t)+q)} \, h_o(\| F_{i+1}(u)-y \| + \| F_i(u)-x \|) \, \omega_i(t) \leq$$

$$\leq \frac{1}{2\varphi(t)+q}(\omega_i(t)+q+2(\omega_o(t)+\ldots+\omega_{i-1}(t)))\omega_i(t)=\omega_{i+1}(t)$$

In this manner we have established (14"). If we show that $u \in Z(t)$ implies

$$(F_{m-1}(u) \ , \quad F_m(u)) \in Z\,\bar{\omega}(t)$$

we shall have (14') as well. It will suffice to prove the following inequalities

$$\| F_{m-1}(u)-F_m(u) \| \leq \omega_{m-1}(t) \tag{58}$$

$$\| F_{m-1}(u)-y_o \| \leq \sigma(t_o)-\sigma_m(t)+q_o-\omega_{m-1}(t) \tag{59}$$

$$\| F_m(u)-x_o \| \leq \sigma(t_o)-\sigma_m(t) \tag{60}$$

$$\| D_m^{-1} f_m \| \leq \omega_m(t) \tag{61}$$

The first inequality is a consequence of (51") and so is (60) which follows from (53) for i=m-1. To obtain (59) we write

$$F_{m-1}(u)-y_o \qquad F_{m-1}(u)-x \; + \; x-y \; + \; y-y_o$$

$$\sum_{j=0}^{m-2} \; {}_j(t)+q+ \; (t_o)- \; (t)+q_o-q= \; (t_o)- \; {}_m(t)+q_o- \; {}_{m-1}(t)$$

As in (54) we obtain

$$D_m^{-1} f_m = (1-D_o^{-1}(D_o-D_m))^{-1} D_o^{-1}(D_m-D)(F_m(u)-F_{m-1}(u))$$

provided $\| D_o^{-1}(D_o-D_m) \| < 1$. By (42), (59) and (60) we have

$$\| D_o^{-1}(D_m-D_o) \| \leq h_o(\| F_{m-1}(u)-y_o \| + \| F_m(u)-x_o \|) \leq$$

$$\leq h_o(2\sigma(t_o)-2\sigma_m(t)+q_o-\omega_{m-1}(t)) = 1-h_o(2\varphi(t)-2(\omega_o(t)+\ldots+\omega_{m-2}(t))-\omega_{m-1}(t))$$

Hence

$$\| D_m^{-1} f_m \| \leq (h_o(2\varphi(t)-2(\omega_o(t)+\ldots+\omega_{m-2}(t))-\omega_{m-1}(t))^{-1} \| D_o^{-1}(D_m-D) \| \omega_{m-1}(t)$$

Since

$$\| D_o^{-1}(D_m-D) \| \leq h_o(\| F_{m-1}(u)-y \| + \| F_m(u)-x \|) \leq h_o(\sum_{j=-1}^{m-2} \omega_j(t) + \sum_{j=0}^{m-1} \omega_j(t))$$

$$= h_o(\omega_{m-1}(t)+q+2(\omega_o(t)+\ldots+\omega_{m-2}(t)))$$

we have $\| D_m^{-1} f_m \| \leq \omega_m(t)$

Until now we have proved that conditions (13) and (14) of Lemma (2.2) are satisfied. Our next task is to prove (19), that is to show that the inclusion

$$(x_{n-1}^{m-1},\ x_{n-1}^{m}) \in Z(\| x_{n-1}^{m-1}-x_{n-1}^{m} \| ,\ \| x_{n-1}^{m}-x_n^1 \|) \qquad (65)$$

holds for each n=1,2,... . But, according to (16) and (50) we already know that

$$(x_{n-1}^{m-1}\ ,\ x_{n-1}^{m}) \in Z(\omega_{m-1}^{(n-1)}(t_o), \omega_{m}^{(n-1)}(t_o)) \tag{66}$$

$$\| x_{n-1}^{m-1} - x_{n-1}^{m} \| \leq \omega_{m-1}^{(n-1)}(t_o) \tag{67}$$

$$\| x_{n-1}^{m} - x_{n}^{1} \| \leq \omega_{o}^{(n)}(t_o) = \omega_{m}^{(n-1)}(t_o) \tag{68}$$

It is easy to see that the function G given by (25) is monotone in the sense that if $q_1 \leq q_2$ and $r_1 \leq r_2$ then $G(q_1, r_1) \leq G(q_2, r_2)$. Using this property from (66), (67) and (68) it follows that

$$\| x_{n-1}^{m} - x_o \| \leq G(t_o) - G(\| x_{n-1}^{m-1} - x_{n-1}^{m} \| , \| x_{n-1}^{m} - x_{n}^{1} \|)$$

$$\| x_{n-1}^{m-1} - y_o \| \leq G(t_o) - G(\| x_{n-1}^{m-1} - x_{n-1}^{m} \| , \| x_{n-1}^{m} - x_{n}^{1} \|) + q_o - \| x_{n-1}^{m-1} - x_{n-1}^{m} \|$$

The above relations together with (66) imply (65).

Then Lemma 2.2 implies that there exists a point $x^* \in U$ which is the common limit of the sequences $(x_n^j)_{n \geq 1}$ $(1 \leq j \leq m)$ and that estimates (46) and (47) are satisfied. Thus the proof of our theorem will be complete if we demonstrate that x^* is a root of the equation $f(x)=0$. To show this let us observe that (42) implies

$$\| D_o^{-1} f(x_{n+1}^1) \| = \| D_o^{-1} (f(x_{n+1}^1) - f(x_n^m) - [x_n^{m-1}, x_n^m; f](x_{n+1}^1 - x_n^m) \| =$$

$$= \| D_o^{-1} ([x_{n+1}^1, x_n^m; f] - [x_m^{m-1}; x_n^m; f])(x_{n+1}^1 - x_n^m) \| \leq h_o \| x_{n+1}^1 - x_n^{m-1} \| \| x_{n+1}^1 - x_n^m \|$$

From the above inequality, using the continuity of f on $\bar{U}$ we deduce that $f(x^*)=0$.

If we compare the estimates (46) and (47) obtained in

the above theorem we see that the estimates (46) can be computed before performing the iterative procedure ( ) while the estimates (47) can be computed only after obtaining the points x_{n-1}^{m-1} , x_{n-1}^{m} and x_n^1 . That's why we shall call estimates (46) apriori estimates and estimates (47) aposteriori estimates. The aposteriori estimates are in general more accurate than the apriori ones.

In what follows we shall particularize the result stated in theorem (3.2) for the cases m=1 and m=2 obtaining in this way some improvements of the results obtained respectively in [5] and [6]. For m=1 from lemma (2.3) we obtain a rate of convergence of type (2.1) given by

$$\omega(q,r)=\omega_1(q,r)=\frac{r(q+r)}{r+2\sqrt{r(q+2)+a^2}} \tag{70}$$

The associate function $\bar{\omega}:T^2\longrightarrow T^2$ will be then defined by

$$\bar{\omega}(q,r)=(r,\omega(q,r)) \tag{71}$$

With the above notation we can state the following corollary of theorem (3.2).

(3.3) COROLLARY. If the hypotheses of Theorem (3.2) are satisfied and if one takes $x_{-1}=y_o$ then the iterative procedure (1) will produce a sequence $(x_n)_{n\geq 1}$ of points of U converging to the root x^* of the equation $f(x)=0$ and the following estimates will be fulfilled:

$$\| x_n-x^*\| \leq \mathcal{G}(\bar{\omega}^{(n)}(q_o,r_o)) \tag{73}$$

$$\| x_n - x^* \| \leq \sigma(\| x_{n-1} - x_{n-2} \|, \| x_{n-1} - x_n \|) - \| x_{n-1} - x_n \| \tag{74}$$

where the functions $\sigma, \omega, \bar{\omega}$, are given respectively by (34), (70), (71) the constant a being chosen as in (48). ▨

For m=2 we obtain a rate of convergence ω of type (2.2) defined as follows: for any $q \geq 0$, $r > 0$ set $t=(q,r)$ and

$$\omega_1(t) = \frac{r(q+r)}{q+2r+2\sqrt{r(q+r)+a^2}} \tag{75}$$

$$\omega_2(t) = \omega_1(t) \frac{q+2r+\omega_1(t)}{2\sqrt{r(q+r)+a^2} - \omega_1(t)} \tag{76}$$

$$\omega(t) = (\omega_1(t), \omega_2(t)) \tag{77}$$

In this case we have of course $\omega(t)=\bar{\omega}(t)$. We shall also use the following three functions:

$$\alpha(t) = \sqrt{r(q+2)+a^2} - a, \quad \beta(t)=\alpha(t)+r, \quad \gamma(t)=\alpha(t)+q+r \tag{78}$$

Now we can state another corollary of Theorem (3.2).

(3.4) COROLLARY. If the hypotheses of Theorem (3.2) are satisfied then the iterative procedure (3) produces two sequences $(y_n)_{n\geq 1}$, $(x_n)_{n\geq 1}$ of points of U such that

1° The sequences $(y_n)_{n\geq 1}$, $(x_n)_{n\geq 1}$ converge to the same limit point x^* which is a root of the equation $f(x)=0$.

2^{o} The following estimates hold:

$$\|y_n - x^*\| \leq \vartheta(\omega^{(n)}(q_o, r_o)) \tag{81}$$

$$\|x_n - x^*\| \leq \beta(\omega^{(n)}(q_o, r_o)) \tag{82}$$

$$\|y_n - x^*\| \leq \alpha(\|y_{n-1} - x_{n-1}\|, \|x_{n-1} - y_n\|) \tag{83}$$

$$\|x_n - x^*\| \leq \alpha(\|y_{n-1} - x_{n-1}\|, \|x_{n-1} - y_n\|) - \omega_1(\|y_{n-1} - x_{n-1}\|, \|x_{n-1} - y_n\|) \tag{84}$$

where the functions ω, ω_1, ω_2, α, β, ϑ are defined by (75)-(78) with the constant a chosen as in (48).

In the following proposition we show that estimates (46) and (47) obtained in theorem (3.2) and, consequently, estimates (63), (64), (81)-(84) from the above corollaries are in some sense the best possible.

(3.5) PROPOSITION. For any triplet of positive numbers h_o, q_o, r_o which verifies the inequality (44) there exists a function $f:\mathbb{R} \to \mathbb{R}$ and two points $x_o, y_o \in \mathbb{R}$ which satisfy the hypothesis of Theorem (3.2) and for which the estimates (46) and (47) are attained for all n=1,2,... .

Proof. The proof is a consequence of the proof of Lemma (2.3) observing that the iterative procedure (3) applied to the function $x \longmapsto h_o(x^2 - a^2)$ and initial points $x_o^{m-1} = \dfrac{1+h_o q_o}{2h_o}$,

$x_o^m = \frac{1-h_o q_o}{2h_o}$ produces the same sequences $(x_n^j)_{n\geq 1}$, $1\leq j\leq m$, as if applied to the function $x \mapsto x^2 - a^2$ and initial points $x_o^{m-1} = (1+q_o)/2$, $x_o^m = (1-q_o)/2$.

Analysing the hypotheses of the Theorem (3.2) we observe that inequality (44) plays a key role. This inequality is satisfied if q_o and r_o are small enough. The number q can be chosen very small from the very begining because having an initial approximation x_o we can choose the point y_o close enough to x_o. In order to have a small r_o we must have a good initial approximation. This requirement is not so easy to be fulfilled in practical applications. However we can show that condition (44) imposed to the initial data is in some sense the weakest possible.

(3.6) PROPOSITION. For any triplet of positive numbers h_o , q_o , r_o which do not satisfy (44) there exists a function $f:\mathbb{R}\to\mathbb{R}$ and two points $x_o, y_o \in \mathbb{R}$ such that:

1^o conditions (42) and (43) of Theorem (32) are satisfied (U can be taken the whole real axis);

2^o the equation $f(x)=0$ has no solution.

Proof. Take

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

$$\text{if } q_o + 2r_o - 2\sqrt{r_o(q_o+r_o)} < \frac{1}{h_o} < q + 2r_o + 2\sqrt{r_o(q_o+r_o)}$$

$$\text{and } f(x) = \frac{1}{q_o}x^2 + r_o \quad , \quad x_o = 0, \quad y_o = q$$

$$\text{if } \frac{1}{h_o} < q_o + 2r_o - 2\sqrt{r_o(q_o+r_o)} \ .$$

REFERENCES

[1] HELFRICH, H.P. Ein modifiziertes Newtonsches Verfahren . "Funktionalanalytische Methoden d.numer.Math", Internat Schriftenr.z.num.Math.12, Basel 1969. Birkhäuser Verlag, 61-70.

[2] HOFMANN, W., Konvergenzsätze für Regula-Falsi-Verfahren, Archive for Rational Mechanics and Analysis, 44, 4 (1972), 296-309.

[3] LAASONEN, P., Ein überquadratisch konvergenter iterativer Algoritmus, Annales Ac.Sci.Femicae, Series A, Mathematica 450 (1969), 1-10.

[4] POTRA, F.-A., On a modified secant method, Preprint INCREST no.8/1979.

[5] POTRA, F.-A., An application of the Induction Method of V.Pták to the study of Regula Falsi, Preprint INCREST no.11/1979.

[6] POTRA, F.-A. and PTÀK, V., Nondiscrete induction and Laasonen's method, Preprint INCREST no.12/1979.

[7] PTÁK, V., Deux théorèmes de factorisation, C.R.Acad.Sci.Paris 278 (1974), 1091-1094.

[8] PTÁK, V., A theorem of the closed graph type, Manuscripta Math. 13 (1974), 109-130.

[9] PTÁK, V., Nondiscrete mathematical induction and iterative existence proofs, Linear algebra and its applications, 13 (1979), 223-236.

[10] 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 (1977).

[11] SERGEEV, A.S., O metode chord, Sibir Matem. Z. 2 (1961), 282-289

[12] SCHMIDT, J.W., Eine übertragung der Regula Falsi auf Gleichungen in Banachraum I, II, Z.Angew.Math.Mech.

[13] SCHMIDT, J.W. and SCHWETLICK, H., Able tungsfreie Verfahren mit öherer Konvergenzgeschwindigkeit, Computing 3 (1968), 215-226.

[14] SCHRÖDER, J., Nichtlineare Majoranten beim Verfahren der schrittweisen Näherung, Arch.Math. (Basel), 7 (1956), 471-484.

[15] ULM, S., Printzip majorant i metod chord. IAN ESSR, ser fiz-matem i tehn., 3 (1964), 217-227.

[16] ULM, S., Ob obobscennych razdelennych raznostjach, I, II, IAN ESSR, ser fiz-matem u tehn, 16 (1967), 13-26, 146-156.