COMPUTATIONAL MATHEMATICS
BANACH CENTER PUBLICATIONS, VOLUME 13
PWN–POLISH SCIENTIFIC PUBLISHERS
WARSAW 1984

ON A CLASS OF ITERATIVE PROCEDURES FOR SOLVING NONLINEAR EQUATIONS IN BANACH SPACES

FLORIAN-ALEXANDRU POTRA

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

1. Introduction

In a joint paper with V. Pták [14], we have given optimal convergence conditions as well as sharp error bounds for the following iterative procedure for solving nonlinear equations in Banach spaces:

$$(1)\qquad \begin{cases} x_n = x_n^0 = x_{n-1}^m, \quad y_n = x_{n-1}^{m-1}, \\ x_n^{k+1} = x_n^k - \delta f(y_n, x_n)^{-1} f(x_n^k), \quad k = 0, 1, \ldots, m-1, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad n = 1, 2, 3, \ldots \end{cases}$$

For $m = 1$ this reduces to the secant method.

In the above formulae f was a nonlinear operator between two Banach spaces, and $\delta f(y, x)$ was a divided difference of f at the points y and x (see [24]).

It is known that the order of convergence of this procedure is $(m+\sqrt{m^2+4})/2$ (see [21]). The natural number m can be chosen, according to the dimension of the space, to maximize the efficiency (see [9] for the definition of the efficiency of an iterative procedure).

For example, if the dimension of the space is respectively equal to 1, 2, 3, then the optimal m is respectively equal to 1, 3, 4.

In what follows we intend to show that the results of [14] remain valid if instead of divided difference one considers the more general notion of consistent approximation of the derivative. We also intend to study the case which appears in numerical applications, where the iterative procedure (1) can be performed only approximately. More precisely, we shall investigate the following "perturbed version" of the pro-

cedure (1):

$$
\begin{aligned}
&\tilde{x}_0 = x_0, \quad \tilde{y}_0 = y_0, \\
(2) \qquad &\tilde{x}_n = \tilde{x}_n^0 = \tilde{x}_{n-1}^m, \quad \tilde{y}_n = \tilde{x}_{n-1}^{m-1}, \\
&\tilde{x}_n^{k+1} = \tilde{x}_n^k - \big(\delta f(\tilde{y}_n, \tilde{x}_n) + E_n\big)^{-1}\big(f(\tilde{x}_n^k) + e_{n,k}\big) + g_{n,k}, \\
&\qquad\qquad k = 0, 1, \ldots, m-1; \quad n = 1, 2, 3, \ldots
\end{aligned}
$$

In the above formulae $\delta f(\tilde{y}_n, \tilde{x}_n) + E_n$ and $f(\tilde{x}_n^k) + e_{n,k}$ represent our estimates for $\delta f(\tilde{y}_n, \tilde{x}_n)$ and $f(\tilde{x}_n^k)$, while the vector $g_{n,k}$ contains the errors made in the matrix inversion (or in the solution of the corresponding linear system) occurring in (2).

Supposing that there exist three positive numbers $\varepsilon_1, \varepsilon_2, \varepsilon_3$ such that

$$
(3) \qquad \|e_{n,k}\| \leqslant \varepsilon_1, \quad \|E_n\| \leqslant \varepsilon_2, \quad \|g_{n,k}\| \leqslant \varepsilon_3,
$$

for all $n \in N$ and $k = 0, 1, \ldots, m-1$, we shall prove, under appropriate hypotheses, that there exists a number δ such that

$$
(4) \qquad \|\tilde{x}_n^k - x_n^k\| \leqslant \delta
$$

for all $k = 0, 1, \ldots, m$ and $n = 0, 1, 2, \ldots$

2. Iterative procedures of type $(2, m)$ and nondiscrete induction

In the study of the iterative procedure (1), we shall use the method of nondiscrete induction.

For the motivation and the general principles of this method see [15] or [16]. The iterative procedure (1) being an iterative procedure of type $(2, m)$ we shall reproduce, in what follows, the results obtained in [14] concerning the application of the nondiscrete mathematical induction to the investigation of this type of iterative procedures.

First, let us give the definition of an iterative procedure of type $(2, m)$. Roughly speaking an iterative procedure of type $(2, m)$ is an iterative procedure which produces, at each step, from the last two points, m new points. To be more precise let us introduce some notations.

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 $z = (z_0, z_1, \ldots, z_m)$, with $z_j \in X$. For each $j = 0, 1, \ldots$ $\ldots, m$ we denote by P_j the mapping which assigns to each $z \in X^{m+1}$ its jth coordinate; thus

$$
z = (P_0 z, P_1 z, \ldots, P_m z).
$$

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

$$Pz = (P_{m-1}z, P_m z).$$

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

$$F_j = P_j F, \quad j = 0, 1, \ldots, m,$$

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

Let G be a mapping from $\mathscr{D}_F$ into X^m and let F be the mapping from $\mathscr{D}_F$ into X^{m+1} defined by setting

$$F(y, x) = \big(x, G(y, x)\big). \tag{5}$$

The mapping F will evidently satisfy the relation

$$P_0 FPz = P_m z \quad \text{for all } z \in P^{-1}\mathscr{D}_F. \tag{6}$$

Conversely, any mapping F: $\mathscr{D}_F \subset X^2 \to Y$ satisfying (6) will be of the form (5).

Let now F: $D_F \subset X^2 \to Y$ be a mapping which satisfies (6) and let $u_0 \in \mathscr{D}_F$ be given. The recurrent scheme

$$x_1 = Fu_0; \quad x_{n+1} = FPx_n, \quad n = 1, 2, 3, \ldots \tag{7}$$

will be called an *iterative procedure of type* $(2, m)$.

Set $\mathscr{D}_0 = \mathscr{D}_F$ and define recursively

$$\mathscr{D}_{n+1} = \{u \in \mathscr{D}_n;\ PFu \in \mathscr{D}_n\}, \quad n = 0, 1, 2, \ldots$$

The set $\mathscr{D} = \bigcap_{n \geq 0} \mathscr{D}_n$ will be called the set of *admissible starting points* for the iterative procedure (7). If $u_0 \in \mathscr{D}$, then we shall say that the iterative procedure (7) is *well defined*.

Now, let us see how the method of nondiscrete induction applies to the study of iterative procedures of type $(2, m)$. First, let us introduce the notion of a rate of convergence of type $(2, m)$:

Let T be either the set of all positive real numbers or a half open interval of the form $(0, s_0]$, for some $s_0 > 0$. 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

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

It will be convenient to introduce also the functions ω_{-1} and ω_0 by the formulae:

$$\omega_{-1}(s) = \omega_{m-1}^{(0)}(s) = q, \quad \omega_0(s) = \omega_m^{(0)}(s) = r; \quad s = (q, r) \in T^2.$$

Let us define the functions $\omega_k^{(n)}\colon T^2 \to T$ by the recursive formula

$$\omega_k^{n+1}(s) = \omega_k\big(\omega_{m-1}^{(n)}(s), \omega_m^{(n)}(s)\big), \quad k = -1, 0, 1, \ldots, m;\ n = 0, 1, 2, \ldots$$

We shall attach to the mapping $\omega\colon T^2 \to T^m$ the mapping $\overline{\omega}\colon T^2 \to T^2$ defined by

$$\overline{\omega}(s) = \big(\omega_{m-1}(s), \omega_m(s)\big).$$

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

$$\overline{\omega}^{(n)}(s) = \big(\omega_{m-1}^{(n)}(s), \omega_m^{(n)}(s)\big) \quad \text{for all } s \in T^2 \text{ and } n = 0, 1, 2, \ldots$$

Considering now for each $n = 1, 2, \ldots$ the mapping $\omega^{(n)}\colon T^2 \to T^m$ with components $\omega_1^{(n)}, \omega_2^{(n)}, \ldots, \omega_m^{(n)}$, it follows that

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

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

A function $\omega\colon T^2 \to T^m$ with the law of iteration described above will be called a *rate of convergence of type* $(2, m)$ on T if the series

$$\sigma(s) = \sum_{n=1}^{\infty} \sum_{j=0}^{m-1} \omega_j^{(n)}(s) \tag{8}$$

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

Since $\omega_0^{(n+1)} = \omega_m^{(n)}$ for all $n = 0, 1, \ldots$, the above expression for σ may be replaced by the following one

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

It will be convenient to introduce the functions $\sigma_0, \sigma_1, \ldots, \sigma_m$ by setting

$$\sigma_0 = \sigma; \quad \sigma_k = \sigma - (\omega_0 + \ldots + \omega_{k-1}), \quad k = 1, 2, \ldots, m.$$

We note the following important functional equation:

$$\sigma\overline{\omega}(s) = \sigma_m(s), \quad s \in T^2. \tag{9}$$

With the above notation we are able to state the following result:

LEMMA 1. *Let X be a complete metric space and let $F\colon \mathscr{D}_F \subset X^2 \to X^{m+1}$ be a mapping which satisfies condition* (6). *Let Z be a mapping which assigns to each $t \in T^2$ a set $Z(t) \subset \mathscr{D}_F$. Let ω be a rate of convergence of type $(2, m)$ on T. Let $u_0 \in \mathscr{D}_F$ and $t_0 \in T^2$ be given.*

If the following conditions are fulfilled:

$$u_0 \in Z(t_0), \tag{10}$$

$$PFZ(t) \subset Z\overline{\omega}(t), \tag{11}$$

$$d(F_k u,\ F_{k+1} u) \leqslant \omega_k(t), \tag{12}$$

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

(i) *the iterative procedure* (7) *is well defined and it yields a sequence* $(x_n)_{n\geqslant 1}$ *of points of* $P^{-1}\mathscr{D}_F$;

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

(iii) *the following relations hold for each* $n = 1, 2, 3, \ldots$

$$Px_n \in Z\overline{\omega}^{(n)}(t_0), \tag{13}$$

$$d(P_k x_n,\ P_{k+1} x_n) \leqslant \omega_k^{(n)}(t_0), \qquad 0 \leqslant k \leqslant m-1, \tag{14}$$

$$d(P_k x_n, x^*) \leqslant \sigma_k \overline{\omega}^{(n-1)}(t_0), \qquad 0 \leqslant k \leqslant m; \tag{15}$$

(iv) *suppose that, for some natural number* n, *we have*

$$Px_{n-1} \in Z(d_n), \tag{16}$$

where $d_n = \big(d(P_{m-1}x_{n-1}, P_m x_{n-1}), d(P_m x_{n-1}, P_1 x_n)\big) \in T^2$ *and* $Px_0 = u_0$; *then*

$$d(P_k x_n, x^*) \leqslant \sigma_k(d_n), \qquad 0 \leqslant k \leqslant m. \ \blacksquare \tag{17}$$

The proof of the above lemma is very simple and will be omitted. The interested reader may find the proof in [14].

In what follows we shall construct a rate of convergence of type $(2, m)$ which will then be used in the study of the iterative procedure (1).

There are some differences between the cases $m = 1$ and $m \geqslant 2$, 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 $k = 0, 1, \ldots, p$, and if p happens to be negative, ignore this instruction and pass to the next one; in the same sense the sum $a_0 + a_1 + \ldots + a_p$ will be taken equal to zero if p is negative.

LEMMA 2. *Let* T *denote the set of all positive real numbers, let* a *be a non-negative real number, and let* m *be a positive integer. For all* $q, r \in T$ *consider the functions:*

$$\varphi(q, r) = r + \sqrt{r(q+r) + a^2}, \tag{18}$$

$$\omega_{-1}(q, r) = q, \qquad \omega_0(q, r) = r, \tag{19}$$

and define

$$(20)\quad \omega_{k+1} = \frac{\omega_k\big(\omega_{-1}+\omega_k+2(\omega_0+\ldots+\omega_{k-1})\big)}{2\varphi+\omega_{-1}}, \qquad k = 0, 1, \ldots, m-2,$$

$$(21)\qquad \omega_m = \frac{\omega_{m-1}\big(\omega_{-1}+\omega_{m-1}+2(\omega_0+\ldots+\omega_{m-2})\big)}{2\varphi-2(\omega_0+\ldots+\omega_{m-2})-\omega_{m-1}}.$$

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

$$(22)\qquad \sigma(q, r) = r+\sqrt{r(q+r)+a^2}-a.$$

Proof. For the proof let us apply the iterative procedure (1) to the real polynomial $f(x) = x^2-a^2$ and initial points $x_0 = s_0^m = \varphi(q, r)$, $y_0 = s_0^{m-1} = \varphi(q, r)+q$. We shall obtain $m+1$ sequences of positive numbers $(s_n^k)_{n\geqslant 1}$, $0 \leqslant k \leqslant m$, related by the following formulae:

$$(23)\qquad s_n^0 = s_{n-1}^m, \qquad s_n^{k+1} = s_n^k - \frac{(s_n^k)^2-a^2}{s_{n-1}^{m-1}+s_{n-1}^m},$$
$$k = 0, 1, \ldots, m-1;\ n = 1, 2, 3, \ldots$$

From the convexity of f, it follows that

$$s_n^m < s_n^{m-1} < \ldots < s_n^1 < s_n^0 = s_{n-1}^m.$$

From the definition of x_0 and y_0 we have

$$s_0^{m-1}-s_0^m = q = \omega_{-1}(q, r), \qquad s_1^0-s_1^1 = s_0^m-s_1^1 = r = \omega_0(q, r).$$

One can prove that

$$s_1^k-s_1^{k+1} = \omega_k(q, r), \qquad k = 0, 1, \ldots, m-1,$$

and more generally

$$(24)\qquad s_n^k-s_n^{k+1} = \omega_k^{(n)}(q, r), \qquad k = 0, 1, \ldots, m-1;\ n = 1, 2, 3, \ldots$$

It follows that ω is a rate of convergence of type $(2, m)$ and that

$$\sigma(q, r) = s_0^m-a = \varphi(q, r)-a.$$

Moreover, we shall have

$$(25)\qquad \sigma_k\overline{\omega}^{(n-1)}(q, r) = s_n^k-a, \qquad 0 \leqslant k \leqslant m. \ \blacksquare$$

3. Sharp error bounds for the iterative procedure (1)

In this section we shall make a semilocal analysis, in the sense of Ortega and Rheinboldt [8], for the iterative procedure (1).

First, let us explain what the symbol $\delta f(y, x)$ means. Let X and Y be two Banach spaces and let $\mathscr{D}_f$ be an open convex subset of X. Let

$f\colon \mathscr{D}_f \subset X \to Y$ be a nonlinear operator which is Fréchet differentiable in $\mathscr{D}_f$. Denote by $L(X, Y)$ the Banach space of all bounded linear operators from X to Y. A mapping $\delta f\colon D_f \times D_f \to L(Y, Y)$ will be called a *(strongly) consistent approximation* of the derivative of f, if there exists a constant $H > 0$ such that

$$\|\delta f(x, y) - f'(z)\| \leqslant H(\|x-z\| + \|y-z\|) \tag{26}$$

for all $x, y, z \in \mathscr{D}_f$.

Let us note that condition (26), which was also considered in [1] and [20], is slightly stronger than the condition defining the notion of strongly consistent approximation from [8].

In (1) the linear operator $\delta f(y_n, x_n)$ appears to the power -1. In the sequel we shall use the following well known result concerning the inversion of linear operators in Banach spaces:

LEMMA 3. *If $L_0 \in L(X, Y)$ is invertible and if $L \in L(X, Y)$ satisfies the condition*

$$\|L\| < \|L_0^{-1}\|^{-1},$$

then

$$\|(L_0 - L)^{-1}\| \leqslant (1 - \|L\| \cdot \|L_0^{-1}\|)^{-1} \|L_0^{-1}\|. \ \blacksquare$$

We shall investigate the convergence of the iterative procedure (1) within the class defined below.

Let h_0 be a positive number and let q_0 and r_0 be nonnegative numbers. We denote by $\mathscr{C}(h_0, q_0, r_0)$ the class of all triplets (f, x_0, y_0) satisfying the properties:

(C_1) f is a nonlinear operator defined on a subset $\mathscr{D}_f$ of a Banach space X and with values in a Banach space Y;

(C_2) y_0 belongs to the sphere $U = \{x \in X;\ \|x - x_0\| < \mu\}$;

(C_3) f is Fréchet differentiable in U;

(C_4) f is continuous on $\overline{U} = \{x \in X;\ \|x - x_0\| \leqslant \mu\}$;

(C_5) there exists a mapping $\delta f\colon U \times U \to L(X, Y)$ such that the linear operator $D_0 = \delta f(y_0, x_0)$ is invertible and

$$\left\|D_0^{-1}\big(\delta f(x, y) - f'(z)\big)\right\| \leqslant h_0(\|x-z\| + \|y-z\|) \quad \text{for all } x, y, z \in U; \tag{27}$$

(C_6) the following inequalities are satisfied:

$$\|x_0 - y_0\| \leqslant q_0, \tag{28}$$

$$\|D_0^{-1} f(x_0)\| \leqslant r_0, \tag{29}$$

$$h_0 q_0 + 2\sqrt{h_0 r_0} \leqslant 1, \tag{30}$$

$$\mu \geqslant \mu_0 := \frac{1}{2h_0}\left(1 - h_0 q_0 - \sqrt{(1 - h_0 q_0)^2 - 4h_0 r_0}\right). \tag{31}$$

In the following theorem we shall show that if $(f, x_0, y_0) \in \mathscr{C}(h_0, q_0, r_0)$, then each of the m sequences $(x_n^k)_{n\geqslant 1}$ $(1 \leqslant k \leqslant m)$, produced by (1) converges to a root x^* of the equation $f(x) = 0$. Before stating this theorem let us make some remarks on the conditions defining the class $\mathscr{C}(h_0, q_0, r_0)$.

The constant h_0 appearing in (27) generally depends on μ. In (31) we ask μ to be greater than μ_0 which depends on h_0. It is then useful to note that $\mu_0 \leqslant r_0 + \sqrt{r_0(q_0+r_0)}$ for all $h_0 > 0$, so that we could take $\mu = r_0 + \sqrt{r_0(q_0+r_0)}$. The most restrictive condition from the definition of the class $\mathscr{C}(h_0, q_0, r_0)$ seems to be inequality (30). This inequality is satisfied only if q_0 and r_0 are small enough. In practical applications q_0 can be taken as small as wanted, because having an initial point x_0 we can choose y_0 very close to it, but r_0 can be taken small only if the initial approximation is "good enough" (see (29)). It is not so easy to find such an initial point! However it turns out that condition (30) is optimal in some sense. Indeed one can show that if this condition is not satisfied then one cannot assure any more the existence of a root of the equation $f(x) = 0$ (see [13] or [14]).

Let us state now the main result of this section.

THEOREM 1. *If $(f, x_0, y_0) \in \mathscr{C}(h_0, q_0, r_0)$ then the iterative procedure* (1) *is well defined and it yields $m+1$ sequences $(x_n^j)_{n\geqslant 1}$, $0 \leqslant j \leqslant m$, with the following properties: there exists a point $x^* \in X$ for which $f(x^*) = 0$, each of these sequences converges to x^*, and the following estimates*

$$\|x_n^j - x^*\| \leqslant \sigma_j\, \overline{\omega}^{(n-1)}(q_0, r_0), \tag{32}$$

$$\|x_n^j - x^*\| \leqslant \sigma_j(\|x_{n-1}^{m-1} - x_{n-1}^m\|,\ \|x_n^0 - x_n^1\|), \tag{33}$$

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

$$a = \frac{1}{2h_0}\sqrt{(1-h_0q_0)^2 - 4h_0r_0}. \tag{34}$$

Proof. The proof is based on Lemma 1. If $u = (y, x) \in U^2$ set $F_0(u) = x$, $F_{j+1}(u) = F_j(u) - \delta f(y, x)^{-1} f F_j(u)$, $j = 0, 1, \ldots, m-1$. Let us denote by $\mathscr{D}_F$ the set of those u for which the above formulae make sense (i.e., $\delta f(y, x)$ is invertible and $F_j(u) \in U$ for $j = 0, 1, \ldots, m-1$) and let us define a mapping $F\colon \mathscr{D}_F \to X^{m+1}$ by setting

$$F(u) = \big(F_0(u),\ F_1(u), \ldots, F_m(u)\big).$$

This function clearly satisfies the properties

$$P_0FPz = P_m z, \quad P_kFu = F_k u, \quad \text{for all } z \in P^{-1}\mathscr{D}_F \text{ and } u \in \mathscr{D}_F.$$

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

Let us assign to each $t = (q, r) \in T^2$ a subset of X^2 defined as follows

$$(35)\qquad Z(t) = \{(y, x) \in X^2;\ y \in U,\ \|y-x\| \leqslant q,\ \|x-x_0\| \leqslant \sigma(t_0)-\sigma(t),$$
$$D = \delta f(y, x) \text{ is invertible, } \|D^{-1}f(x)\| \leqslant r,$$
$$\|(D_0^{-1}D)^{-1}\| \leqslant 1/h_0\big(2\varphi(t)+q\big)\}.$$

In the above definition of $Z(t)$, t_0 stands for the pair (q_0, r_0). Hence using (31) it follows that $Z(t) \subset U^2$. Consider now the rate of convergence ω described in Lemma 2, the constant a being given by (34). Our theorem will be proved if we show that $Z(t) \subset \mathscr{D}_F$ and that conditions (10), (11), (12) and (16) from Lemma 1 are satisfied. First of all, if u_0 stands for (y_0, x_0) we clearly have $u_0 \in Z(t_0)$. Let us prove now that $u \in Z(t)$ implies

$$(36)\qquad F_k(u) \in U \quad \text{for} \quad -1 \leqslant k \leqslant m$$

and

$$(37)\qquad \|F_k(u)-F_{k+1}(u)\| \leqslant \omega_k(t) \quad \text{for} \quad -1 \leqslant k \leqslant m-1.$$

For $k = -1$ these relations reduce to $y \in U$ and $\|y-x\| \leqslant q$; for $k = 0$ they follow from $x \in U$ and $\|\delta f(y, x)^{-1}f(x)\| \leqslant r$. Consider now an i, $0 \leqslant i \leqslant m-1$, and suppose that (36) and (37) hold for $k = -1, 0, \ldots, i$. We have then

$$\|F_{i+1}(u)-x_0\| \leqslant \|F_{i+1}(u)-x\| + \|x-x_0\| \leqslant \sum_{j=0}^{i} \|F_{j+1}(u)-F_j(u)\| + \|x-x_0\|$$
$$\leqslant \sum_{j=0}^{i} \omega_j(t) + \sigma(t_0)-\sigma(t) = \sigma(t_0)-\sigma_{i+1}(t),$$

so that $F_{i+1}(u) \in U$ as well; this establishes (36). Let us remark that from (35) and (36) it follows that $Z(t) \subset \mathscr{D}_F$. To simplify the formulae let $D_m = \delta f\big(F_{m-1}(u),\ F_m(u)\big)$, $f_j = f\big(F_j(u)\big)$. The relation defining $F_{i+1}(u)$ may be thus written in the form $f_i = D\big(F_i(u)-F_{i+1}(u)\big)$. Hence we may write

$$(38)\qquad \begin{aligned} F_{i+1}(u)-F_{i+2}(u) &= D^{-1}f_{i+1} \\ &= (D_0^{-1}D)^{-1}D_0^{-1}\Big(f_{i+1}-f_i-D\big(F_{i+1}(u)-F_i(u)\big)\Big). \end{aligned}$$

At this stage let us note that condition (27) implies that

$$\big\|D_0^{-1}\big(f'(y_1)-f'(y_2)\big)\big\| \leqslant 2h_0\|y_1-y_2\| \quad \text{for all } y_1, y_2 \in U.$$

Using a standard argument (see [8], 3.2.12) we deduce that

$$\big\|D_0^{-1}\big(f(y_1)-f(y_2)-f'(y_2)(y_1-y_2)\big)\big\| \leqslant h_0\|y_1-y_2\|^2 \quad \text{for all } y_1, y_2 \in U.$$

It follows that for all $y_1, y_2, z_1, z_2 \in U$ we have

$$(39)\quad \begin{aligned} &\|D_0^{-1}(f(y_1)-f(y_2)-\delta f(z_1,z_2)(y_1-y_2))\| \\ &\qquad \leqslant \|D_0^{-1}(f(y_1)-f(y_2)-f'(y_2)(y_1-y_2))\| + \\ &\qquad\quad + \|D_0^{-1}(f'(y_2)-\delta f(z_1,z_2))(y_1-y_2)\| \\ &\qquad \leqslant h_0(\|y_2-z_1\|+\|y_2-z_2\|+\|y_1-y_2\|)\,\|y_1-y_2\|. \end{aligned}$$

Now from (38) we obtain

$$(40)\quad \begin{aligned} &\|F_{i+1}(u)-F_{i+2}(u)\| \\ &\leqslant \|(D_0^{-1}D)^{-1}\|h_0\big(\|F_i(u)-y\|+\|F_i(u)-x\|+\|F_i(u)-F_{i+1}(u)\|\big)\times \\ &\qquad\qquad \times\|F_i(u)-F_{i+1}(u)\| \\ &\leqslant \frac{\omega_i(t)}{2\varphi(t)+q}\big(\omega_i(t)+2\big(\omega_0(t)+\ldots+\omega_{i-1}(t)\big)+q\big) = \omega_{i+1}(t). \end{aligned}$$

In this manner we have established (37). Now we intend to show that $u \in Z(t)$ implies

$$\big(F_{m-1}(u), F_m(u)\big) \in Z\,\overline{\omega}(t).$$

It will suffice to prove the following inequalities:

$$(41)\qquad \|F_{m-1}(u)-F_m(u)\| \leqslant \omega_{m-1}(t),$$

$$(42)\qquad \|F_m(u)-x_0\| \leqslant \sigma(t_0)-\sigma_m(t),$$

$$(43)\qquad \|D_m^{-1}f_m\| \leqslant \omega_m(t),$$

$$(44)\qquad \|(D_0^{-1}D_m)^{-1}\| \leqslant \frac{1}{h_0\big(2\varphi\,\overline{\omega}(t)+\omega_{m-1}(t)\big)}.$$

The first inequality is a consequence of (37) and so is (42), which follows from (40) for $i = m-1$. By (9), (27) and (37) we have

$$\begin{aligned} \|D_0^{-1}(D_m-D_0)\| &\leqslant \|D_0^{-1}(D_m-f'(x_0))\| + \|D_0^{-1}(f'(x_0)-D_0)\| \\ &\leqslant h_0\big(\|F_m(u)-x_0\|+\|F_{m-1}(u)-x_0\|+\|x_0-y_0\|\big) \\ &\leqslant h_0\big(2\sigma(t_0)-2\sigma_m(t)+q_0-\omega_{m-1}(t)\big) \\ &= 1-h_0\big(2\varphi\,\overline{\omega}(t)+\omega_{m-1}(t)\big). \end{aligned}$$

According to Lemma 3 this implies the invertibility of D_m and the inequality (44).

Using the identity $f_{m-1} = -D\big(F_m(u) - F_{m-1}(u)\big)$ we obtain

$$\begin{aligned}\|D_m^{-1} f_m\| &= \big\|(D_0^{-1}D_m)^{-1}D_0^{-1}(f_m - f_{m-1}) - D\big(F_m(u) - F_{m-1}(u)\big)\big\| \\ &\leqslant h_0 \|(D_0^{-1}D_m)^{-1}\| \big(\|F_{m-1}(u) - y\| + \|F_{m-1}(u) - x\| + \\ &\qquad + \|F_m(u) - F_{m-1}(u)\|\big) \|F_m(u) - F_{m-1}(u)\| \\ &\leqslant \frac{\omega_{m-1}(t)}{2\varphi\overline{\omega}(t) + \omega_{m-1}(t)} \Big(\omega_{m-1}(t) + 2\big(\omega_0(t) + \ldots + \omega_{m-2}(t)\big) + \omega_{-1}(t)\Big) \\ &= \omega_m(t).\end{aligned}$$

Until now we have proved that conditions (10), (11) and (12) of Lemma 1 are satisfied. Our next task is to prove (16), that is to show that the inclusion

$$(x_{n-1}^{m-1}, x_{n-1}^{m}) \in Z(\|x_{n-1}^{m-1} - x_{n-1}^{m}\|, \|x_{n-1}^{m} - x_n^1\|) \tag{45}$$

holds for each $n = 1, 2, \ldots$

But according to (13) and (35) we already know that

$$(x_{n-1}^{m-1}, x_{n-1}^{m}) \in Z\big(\omega_{m-1}^{(n-1)}(t_0), \omega_m^{(n-1)}(t_0)\big), \tag{46}$$

$$\|x_{n-1}^{m-1} - x_{n-1}^{m}\| \leqslant \omega_{m-1}^{(n-1)}(t_0), \tag{47}$$

$$\|x_{n-1}^{m} - x_n^1\| \leqslant \omega_0^{(n)}(t_0) = \omega_m^{(n-1)}(t_0). \tag{48}$$

It is easy to see that the function σ given by (22) is monotone in the sense that if $q_1 \leqslant q_2$ and $r_1 \leqslant r_2$ then $\sigma(q_1, r_1) \leqslant \sigma(q_2, r_2)$. Using this property, from (46), (47) and (48) it follows that

$$\|x_{n-1}^{m} - x_0\| \leqslant \sigma(t_0) - \sigma(\|x_{n-1}^{m-1} - x_{n-1}^{m}\|, \|x_{n-1}^{m} - x_n^1\|),$$

$$\begin{aligned}&\big\|\big(D_0^{-1}\delta f(x_{n-1}^{m-1}, x_{n-1}^{m})\big)^{-1}\big\| \\ &\qquad \leqslant \big[h_0\big(2\varphi(\|x_{n-1}^{m-1} - x_{n-1}^{m}\|, \|x_{n-1}^{m} - x_n^1\|) + \|x_{n-1}^{m-1} - x_{n-1}^{m}\|\big)\big]^{-1}.\end{aligned}$$

The above relations together with (46) imply (45).

Then Lemma 1 implies that there exists a point $x^* \in U$ which is the common limit of the sequences $(x_n^j)_{n\geqslant 1}$ $(1 \leqslant j \leqslant m)$ and that estimates (32) and (33) 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 (27) implies

$$\begin{aligned}\|D_0^{-1}(x_{n+1}^1)\| &= \big\|D_0^{-1}\big(f(x_{n+1}^1) - f(x_n^m) - \delta f(x_n^{m-1}, x_n^m)(x_{n+1}^1 - x_n^m)\big)\big\| \\ &\leqslant h_0 \|x_{n+1}^1 - x_n^{m-1}\| \, \|x_{n+1}^1 - x_n^m\|.\end{aligned} \tag{49}$$

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

We shall conclude this section remarking that the estimates (32) and (33) obtained in Theorem 1 are *sharp in the class* $\mathscr{C}(h_0, q_0, r_0)$. Indeed, let $h_0 > 0$, $q_0 \geqslant 0$ and $r_0 \geqslant 0$ be any numbers satisfying condition (30), and consider the real polynomial $f(x) = h_0(x^2 - a^2)$, with a given by (34). Take $x_0 = 1/2h_0 - q_0/2$ and $y_0 = 1/2h_0 + q_0/2$. We have $(f, x_0, y_0) \in \mathscr{C}(h_0, q_0, r_0)$ and from the proof of Lemma 2 it follows that in this particular case the estimates (32) and (33) are attained for all $n = 1, 2, 3, \ldots$

4. An error analysis in the perturbed case

In this section we shall investigate the iterative procedure (2) trying to find estimates for the distances $\|\tilde{x}_n^k - x_n^k\|$ for $k = 0, 1, \ldots, m$ and $n = 1, 2, 3, \ldots$ We shall do this for triplets $(f, x_0, y_0) \in \mathscr{C}(h_0, q_0, r_0)$ satisfying the following two conditions:

(C*) The nonlinear operator f is Fréchet differentiable in the sphere $U^* = \{x \in X;\ \|x - x_0\| < \mu^*\}$ and the relations

$$\|\delta f(x, y) - \delta f(u, v)\| \leqslant H(\|x - u\| + \|y - v\|), \qquad \delta f(x, x) = f'(x) \tag{50}$$

are satisfied 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 φ such that

$$1/\varphi \geqslant \sup\{\|\delta f(x, y)^{-1}\|;\ x, y \in U^*\}. \tag{51}$$

Let us remark that condition (50) is stronger than condition (26) but it is satisfied by the most used examples.

THEOREM 2. *Let* $(f, x_0, y_0) \in \mathscr{C}(h_0, q_0, r_0)$ *and let* $(x_n^j)_{n \geqslant 1}$, $0 \leqslant j \leqslant m$ *be the sequences generated by the iterative procedure* (1). *Suppose that conditions* (3), (C*) *and* (C**) *are satisfied. Let* $v_0 = \max\{q_0, r_0\}$.

If the inequalities

$$Q = \varphi - H(2m+1)v_0 - 2\varepsilon_2 \geqslant 0, \tag{52}$$

$$D = Q^2 - 12H\big(\varepsilon_1 + \varepsilon_2 v_0 + \varepsilon_3(\varphi - \varepsilon_2)\big) \geqslant 0, \tag{53}$$

$$\delta = (Q - \sqrt{D})/6H \leqslant \mu^* - \mu_0, \tag{54}$$

hold, then the iterative procedure (2) *is well defined and the estimates*

$$\|\tilde{x}_n^j - x_n^j\| \leqslant t_n^j \leqslant \delta \tag{55}$$

are satisfied for all $n \in N$ *and* $j = 0, 1, \ldots, m$, *where*

$$t_0^{m-1} = t_0^m = 0,\ s_0^{m-1} = (1+h_0 q_0)/(2h_0),\ s_0^m = s_0^{m-1} - q_0,\ w_{m-1}^0 = q_0,$$

$$w_m^0 = r_0,\ s_n^0 = s_{n-1}^m,\ s_n^{k+1} = s_n^k - (s_{n-1}^{m-1} + s_{n-1}^m)^{-1}\big((s_n^k)^2 - a^2\big),\ w_k^n = s_n^k - s_n^{k+1},$$

$$t_n^0 = t_{n-1}^m,\ t_n^{k+1} = (\varphi - \varepsilon_2)^{-1}\Big(H(t_{n-1}^{m-1} + t_{n-1}^m + t_n^k)\, t_n^k + \big(2H(w_0^n + \ldots + w_{k-1}^n) + \\ + w_{m-1}^{n-1} + \varepsilon_2\big) t_n^k + H w_k^n (t_{n-1}^{m-1} + t_{n-1}^m) + \varepsilon_1 + \varepsilon_2 w_k^n + \varepsilon_3(\varphi - \varepsilon_2)\Big).$$

Proof. The inequalities (55) are trivially satisfied for $n = 0$ and $j = m-1, m$. Consider an $i \geqslant 1$ and suppose that they are satisfied for $n = i-1$ and $j = m-1, m$. In this case, according to (54), the points $\tilde{x}_i = \tilde{x}_{i-1}^m$ and $\tilde{y}_i = \tilde{x}_{i-1}^{m-1}$ will belong to U^*. Condition (C**) and Lemma 3 imply then the invertibility of the linear operator $\delta f(\tilde{y}_i, \tilde{x}_i) + E_i$ and the fact that

$$\big\|\big(\delta f(\tilde{y}_i, \tilde{x}_i) + E_i\big)^{-1}\big\| \leqslant (\varphi - \varepsilon_2)^{-1}.$$

We shall prove now that inequalities (55) are verified for $n = i$ and $j = 0, 1, \ldots, m$. Because $x_{i-1}^m = x_i^0$ and $t_{i-1}^m = t_i^0$ we may consider formally that they are also verified for $n = i$ and $j = 0$. Suppose they are satisfied for $n = i$ and $j = 0, 1, \ldots, k$, where $0 \leqslant k \leqslant m-1$. Write $D_i = \delta f(y_i, x_i)$ and $\tilde{D}_i = \delta f(\tilde{y}_i, \tilde{x}_i)$. From (1) and (2) we deduce the equality

$$\tilde{x}_i^{k+1} - x_i^{k+1} = (\tilde{D}_i + E_i)^{-1}[f(x_i^k) - f(\tilde{x}_i^k) - \tilde{D}_i(x_i^k - \tilde{x}_i^k) + \\ + (\tilde{D}_i - D_i) D_i^{-1} f(x_i^k) + E_i D_i^{-1} f(x_i^k) - E_i(x_i^k - \tilde{x}_i^k) - e_{i,k}] + q_{i,k}. \tag{56}$$

From the proofs of Lemma 2 and Theorem 1 it follows that

$$\|D_i^{-1} f(x_i^k)\| = \|x_i^{k+1} - x_i^k\| \leqslant w_k^i.$$

Using (50) we obtain the inequalities:

$$\|f(x_i^k) - f(\tilde{x}_i^k) - \tilde{D}_i(x_i^k - \tilde{x}_i^k)\| \leqslant H(\|x_i^k - \tilde{x}_{i-1}^{m-1}\| + \|\tilde{x}_i^k - \tilde{x}_{i-1}^m\|)\, \|x_i^k - \tilde{x}_i^k\| \\ \leqslant H\big(t_{i-1}^{m-1} + t_{i-1}^m + t_i^k + w_{m-1}^{i-1} + \\ + 2(w_0^i + \ldots + w_{k-1}^i)\big) t_i^k,$$

$$\|(\tilde{D}_i - D_i) D_i^{-1} f(x_i^k)\| \leqslant H(\|\tilde{x}_{i-1}^{m-1} + x_{i-1}^{m-1}\| + \|\tilde{x}_{i-1}^m - x_{i-1}^m\|)\, w_i^k \\ \leqslant H(t_{i-1}^{m-1} + t_{i-1}^m)\, w_i^k.$$

Now, (56) implies

$$\|\tilde{x}_i^{k+1} - x_i^{k+1}\| \leqslant (\varphi - \varepsilon_2)^{-1}\Big(H(t_{i-1}^{m-1} + t_{i-1}^m + t_i^k)\, t_i^k + \\ + \big(2H(w_0^i + \ldots + w_{k-1}^i) + w_{m-1}^{i-1} + \varepsilon_2\big) t_i^k + \\ + H w_k^i (t_{i-1}^{m-1} + t_{i-1}^m) + \varepsilon_1 + \varepsilon_2 w_k^i + \varepsilon_3(\varphi - \varepsilon_2)\Big) = t_i^{k+1}.$$

Let us denote $B = H(2m+1)v_0$ and $C = \varepsilon_1 + \varepsilon_2 v_0 + \varepsilon_3(\varphi - \varepsilon_2)$. Because $w_j^n \leqslant v_0$ for all n and j it follows that

$$t_i^{k+1} \leqslant (\varphi - \varepsilon_2)^{-1}(3H\delta^2 + B\delta + C) = \delta.$$

Thus inequalities (55) are satisfied for $n = i$ and $j = k+1$, so that they will be satisfied for all $n \in N$ and $j = 0, 1, \ldots, m$. ■

The error analysis made above for the iterative procedure (2) is similar to the error analysis made for Newton's method by Lancaster [6], Rokne [17] and Miel [7]. For the case $m = 1$ more precise results have been obtained in [12].

References

[1] W. Burmeister, *Inversions freie Verfahren zur Lösung nichtlinearer Operatorgleichungen*, ZAMM 52 (1972), 101–110.

[2] L. Collatz, *Funktionalanalysis und Numerische Mathematik*, Springer, Berlin 1964.

[3] H. P. Helfrich, *Ein modifiziertes Newtonsches Verfahren*. "Funktionalanalytische Methoden d. numer. Math"., Internat Schriftenr. z. num. Math. 12, Basel 1969. Birkhäuser Verlag, 61–70.

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

[5] P. Laasonen, *Ein überquadratisch konvergenter iterativer Algoritmus*, Annales Ac. Sci. Fenicae, Series A, Mathematica 450 (1969), 1–10.

[6] P. Lancaster, *Error analysis for the Newton Raphson-Method*, Numer. Math. 9 (1966), 55–68.

[7] G. J. Miel, *Unified error analysis for Newton-type methods*, ibid. 33 (1979), 391–396.

[8] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, New York and London 1970.

[9] A. M. Ostrowski, *Solution of equations in Euclidean and Banach spaces*, New York and London 1973.

[10] F.-A. Potra, *On a modified secant method*, L'Analyse numérique et la Théorie de l'approximation, 8.2 (1979), 203–214.

[11] —, *An application of the Induction Method of V. Pták to the study of Regula Falsi*, Aplikace Matematiky 26 (1981), 13–17.

[12] —, *An error analysis for the secant method*, Numer. Math. 38 (1982), 427–44.5

[13] F.-A. Potra, and V. Pták, *Nondiscrete induction and a double step secant method*, Math. Scand. 46 (1980), 107–120.

[14] —, —, *A generalization of Regula Falsi*, Numer. Math. 36 (1981), 333–346.

[15] V. Pták, *Nondiscrete mathematical induction and iterative existence proofs*, Linear algebra and its applications, 13 (1979), 223–236.

[16] —, *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.

[17] J. Rokne, *Newton's method under mild differentiability conditions with error analysis*, Numer. Math. 18 (1972), 401–412.

[18] А. С. Сергеев, *О методе хорд*, Сибир. матем. ж. 2.2 (1961), 282–289.

[19] J. W. Schmidt, *Eine übertragung der Regula Falsi auf Gleichungen in Banachraum I, II*, Z. Angew. Math. Mech. 43 (1963), 1–8, 97–110.

[20] J. W. Schmidt, *Regula-Falsi. Verfahren mit konsistenter Steigung und Majoranten Prinzip*, Periodica Matematica Hungarica, 5.3 (1974), 187–193.

[21] J. W. Schmidt and H. Schwetlick, *Ableitungsfreie Verfahren mit höherer Konvergenzgeschwindigkeit*, Computing 3 (1968), 215–226.

[22] J. Schröder, *Nichtlineare Majoranten beim Verfahren der schrittweisen Näherung*, Arch. Math. (Basel), 7 (1956), 471–484.

[23] С. Ульм, *Принцип мажорант и метод хорд*, Изв. АН Эст. ССР, Сер. физ.-матем. и техн. н. 13.3 (1964), 217–227.

[24] —, *Об обобщенных разделенных разностях, I, II*, ibid. 16 (1967), 13–26, 146–156.

Presented to the Semester
Computational Mathematics
February 20 – May 30, 1980