SHARP ERROR BOUNDS FOR A NEWTON-MOSER TYPE METHOD (*)

by IGOR MORET (in Trieste) (**)

SOMMARIO. - *Nello studio della convergenza di un metodo iterativo di tipo Newtoniano, noto come metodo di tipo Newton-Moser, viene impiegata una tecnica di analisi che consente di ottenere delle maggiorazioni, a posteriori, degli errori in senso stretto.*

SUMMARY. - *The convergence of a Newton-Moser type method is examined. Moreover the analysis performed allows us to obtain sharp a posteriori error bounds.*

1. - Introduction

The term Newton-Type methods usually denotes a class of iterative processes to solve an equation $F(x) = 0$; these methods have the general form

$$x_{n+1} = x_n - A_n F(x_n) \quad n = 0, 1, 2, \ldots$$

were A_n is a linear operator somehow associated with x_n.

In Newton's Method A_n is the inverse of $F'(x_n)$; in other methods (e.g. Newton-SOR, [4] p. 215) A_n is the inverse of a linear operator which approximates $F'(x_n)$ in such a way as to simplify the process of inversion; in other methods, still, A_n is the n-th element of a suitable sequence of linear operators.

The method that we consider belongs to the last class; the

(*) Pervenuto in Redazione il 31 novembre 1983.
(**) Indirizzo dell'Autore: Dipartimento di Elettrotecnica, Elettronica ed Informatica dell'Università degli Studi di Trieste - Via Valerio, 10 - 34100 Trieste.

computation of A_n is carried out using the iterative formula

$$A_n = A_{n-1} - A_{n-1}\,(F'(x_n)\,A_{n-1} - I) \quad n = 0, 1, 2, \ldots$$

On the origin of this process, called a Newton-Moser Type method by Hald in [1], see, for instance, Zehnder [7]. Moreover, in the paper by Hald one can find an analysis of the speed of convergence and other interesting computational remarks.

In this note we will examine the convergence of the method by characterizing the initial quatities x_0 and A_{-1}. Moreover the analysis we propose allows us to obtain aposteriori upper bounds for the errors. Similar results for other Newton Type methods can be found in [2], [3], [4] and [5]. In section 2 we study some properties of certain functions, which are useful tools to be used in section 3 in order to prove our main results. For the sake of simplicity we consider the method on R^n. The extension to Banach spaces can be obtained by easy changes.

2. - Preliminary results

According to the usual notations, let us denote by (a, b) the open interval $a < x < b$ and by $[a, b)$ the interval $a \leq x < b$.

Let c be a non negative real number and let us consider the following functions:

$\psi : (c, \infty) \times [0, 1) \to R$ so defined:

$$\psi(s, q) = \frac{(1 - q^2)\,(s^2 - c^2)}{2s};\ s \in (c, \infty), q \in [0, 1) \tag{2.1}$$

$\varphi, u, v : (0, \infty) \times [0, 1) \to R$ so defined:

$$\varphi(p, q) = \frac{p}{(1 - q^2)} + \sqrt{\frac{p^2}{(1 - q^2)^2} + c^2}, \tag{2.2}$$

$$u(p, q) = q^2 + \frac{(1 - q^2)\,p}{\varphi(p, q)}, \tag{2.3}$$

$$v(p, q) = \frac{(1 - q^2)\,(1 + u(p, q))}{2\varphi(p, q)}\left[\, p^2 + \frac{2q^2 p\,\varphi(p, q)}{(1 - q^2)}\,\right], \tag{2.4}$$

where $p \in (0, \infty), q \in [0, 1)$.

It is easy to verify that for any $s \in (c, \infty), p \in (0, \infty)$ and $q \in [0, 1)$ we have:

$$\varphi(\psi(s, q), q) = s, \tag{2.5}$$

$$\psi(\varphi(p,q),q) = p, \tag{2.6}$$

$$u(p,q) < 1. \tag{2.7}$$

Moreover, by a simple analysis, one can see that, if $p_1, p_2 \in (0, \infty)$ and $q_1, q_2 \in [0,1)$ with $p_1 \leq p_2$ and $q_1 \leq q_2$, the relations

$$\begin{aligned} u(p_1,q_1) &\leq u(p_2,q_2) \\ v(p_1,q_1) &\leq v(p_2,q_2) \end{aligned} \tag{2.8}$$

hold.

Now, let us consider the following iterative scheme:

$$s_{n+1} = s_n - \psi(s_n, q_n) \tag{2.9}$$

$$q_{n+1} = 1 - \frac{(1-q_n^2)\, s_{n+1}}{s_n} \tag{2.10}$$

for $n = 0, 1, 2, \ldots$

THEOREM 2.1 - *For any $s_0 \in (c, \infty)$ and $q_0 \in [0,1)$ the iterative procedure defined by* (2.9) *and* (2.10) *yields two sequence $\{s_n\}_{n=0}^{\infty}$ and $\{q_n\}_{n=0}^{\infty}$ with the following properties:*

1) *$\lim s_n = c$; $\lim q_n = 0$*

2) *For any index $n \geq 0$, if we set $p_n = s_n - s_{n+1}$, we have*

$$s_n = \varphi(p_n, q_n), \tag{2.11}$$

$$q_{n+1} = u(p_n, q_n), \tag{2.12}$$

$$p_{n+1} = v(p_n, q_n). \tag{2.13}$$

Proof. Let $s_n \in (c, \infty)$ and $q_n \in [0,1)$. From (2.9) we obtain

$$s_{n+1} - c = (s_n - c)\left[\, 1 - \frac{(1-q_n^2)\,(s_n + c)}{2\, s_n} \,\right] \tag{2.14}$$

so that $s_{n+1} \in (c, s_n)$. Moreover, if $p_n = s_n - s_{n+1}$, using properties (2.5) and (2.6) we have (2.11) and we can express (2.10) in the form

$$q_{n+1} = 1 - \frac{(1-q_n^2)\,(\varphi(p_n,q_n) - p_n)}{\varphi(p_n,q_n)}$$

that is (2.12). From (2.7) it follows, by induction, that (2.14) holds for all n's. Therefore $\{s_n\}_{n=0}^{\infty}$ is a decreasing and converging sequence. It is easy to see that $\lim s_n = c$ and $\lim q_n = 0$.

At last, since

$$p_{n+1} = s_{n+1} - s_{n+2} = \psi(s_n - p_n, q_{n+1})$$

using (2.11), (2.12) and (2.3) one can obtain (2.13) by simple computations.

As a first consequence of theorem 2.1 we have the following.

COROLLARY - Let $p \in (0, \infty)$ and $q \in [0, 1)$. Then

$$\varphi(v(p, q), u(p, q)) = \varphi(p, q) - p. \tag{2.15}$$

For $p \in (0, \infty)$ and $q \in [0, 1)$ we define

$$u^{(0)}(p, q) = q \quad ; \quad v^{(0)}(p, q) = p$$

$$u^{(k)}(p, q) = u(v^{(k-1)}(p, q), u^{(k-1)}(p, q)) ;$$

$$v^{(k)}(p, q) = v(v^{(k-1)}(p, q), u^{(k-1)}(p, q)) ;$$

for $k = 1, 2, \ldots$

THEOREM 2.2 - *For each couple (p, q) with $p \in (0, \infty)$ and $q \in [0, 1)$ the series $\sum_{k=0}^{\infty} v^{(k)}(p, q)$ converges and we have*

$$\sum_{k=0}^{\infty} v^{(k)}(p, q) = \varphi(p, q) - c \tag{2.16}$$

Proof. Let us consider the iterative scheme (2.9) - (2.10) with $s_0 = \varphi(p, q)$ and $q_0 = q$. From (2.6) we have $p_0 = s_0 - s_1 = p$.

For any index $n > 0$, from (2.12) and (2.13) we obtain

$$s_0 - s_n = \sum_{k=0}^{n-1} p_k = \sum_{k=0}^{n-1} v^{(k)}(p, q)$$

and consequently (2.16).

COROLLARY - *Let $p \in (0, \infty)$ and $q \in [0, 1)$. Then*

$$\sum_{k=0}^{\infty} v^{(k)}(v(p, q), u(p, q)) = \varphi(p, q) - p - c. \tag{2.17}$$

Proof. One obtains (2.17) recalling (2.15).

3. - Main results

In what follows, D is an open convex subset of R^n and F is a mapping from D into R^n. The space of the bounded linear operators of R^n into itself (space of the $n \times n$ real matrices) is denoted by $L(R^n)$. We assume that the norm on $L(R^n)$ is the matrix norm

induced by the vector norm on R^n. Moreover we suppose that F is Fréchet differentiable in each point of D and that its Fréchet derivative satisfies a Lipschitz condition with constant K, *i.e.*

$$\| F'(x) - F'(y) \| \leqslant K \| x - y \|, x, y \in D. \tag{3.1}$$

It is well known (cf. [4] p. 73) that from the above condition we have

$$\| F(y) - F(x) - F'(x)\,(y - x) \| \leqslant (K/2) \| y - x \|^2. \tag{3.2}$$

DEFINITION 3.1 - *We denote by the term φ-couple for F any couple (x, A) with $x \in D$ and $A \in L(R^n)$ such that there exist two real numbers $p \in (0, \infty)$ and $q \in [0, 1)$ for which the following relations are satisfied:*

$$\| I - AF'(x) \| \leqslant q \tag{3.3}$$

$$\| \bar{A} F(x) \| \leqslant p \tag{3.4}$$

$$\| \bar{A} \| \leqslant \frac{(1 - q^2)}{K \varphi(p, q)} \tag{3.5}$$

where

$$\bar{A} = A - A\,[F'(x)\,A - I] \tag{3.6}$$

We say that the couple (p, q) is associated to (x, A).

LEMMA 3.2 - *Let (x, A) be a φ-couple for F with associated (p, q) and let $z \in D$. If*

$$\| z - x \| \leqslant \| \bar{A} F(x) \| \tag{3.7}$$

we have

$$\| I - \bar{A} F'(z) \| \leqslant u(p, q) \tag{3.8}$$

Proof. From (3.6) we obtain, after simple computations,

$$I - \bar{A} F'(z) = [I - AF'(x)]^2 + \bar{A}\,[F'(x) - F'(z)]$$

and from (3.7) using (3.1), (3.3), (3.4) and (3.5), we have (3.8).

LEMMA 3.3 - *Let the hypotheses of Lemma 3.2 be valid. Moreover let the following relation hold*

$$\| \bar{\bar{A}} F(z) \| \leqslant v(p, q) \tag{3.9}$$

where

$$\bar{\bar{A}} = \bar{A} - \bar{A}\,[F'(z)\,\bar{A} - I], \tag{3.10}$$

then $(z, \bar{A})$ is a φ-couple for F with associated $(v(p, q), u(p, q))$.

Proof. We can write (3.10) in the form

$$\bar{\bar{A}} = [I + (I - \bar{A}F'(z))]\,\bar{A} \tag{3.10)'}$$

and by Lemma 3.2 we have

$$\|\bar{\bar{A}}\| \leqslant (1 + u(p,q))\|\bar{A}\|. \tag{3.11}$$

Using (2.3) it is easy to verify that

$$q^2 = \frac{\varphi(p,q)\,u(p,q) - p}{\varphi(p,q) - p}$$

so that we can express (3.5) in the form

$$\|\bar{A}\| \leqslant \frac{1 - u(p,q)}{K(\varphi(p,q) - p)}.$$

Therefore, according to (3.11) and to (2.15), we obtain

$$\|\bar{\bar{A}}\| \leqslant \frac{1 - u^2(p,q)}{K\,\varphi(v(p,q), u(p,q))}.$$

The above relation with (3.8) and (3.9) proves the Lemma.

Now we shall explain our main result.

THEOREM 3.4 - *Let $x_0 \in D$ and $A_{-1} \in L(R^n)$. Suppose that there exist $\beta > 0$, $q_0 \in [0,1)$ and $p_0 \in (0,\infty)$ such that following inequalities are satisfied:*

$$\|A_{-1}\| \leqslant \beta \tag{3.12}$$

$$\|I - A_{-1}F'(x_0)\| \leqslant q_0 \tag{.3.13}$$

$$\|A_0 F(x_0)\| \leqslant p_0, \tag{3.14}$$

where

$$A_0 = A_{-1} - A_{-1}[F'(x_0)\,A_{-1} - I], \tag{3.15}$$

and

$$m = \frac{2K\beta p_0}{(1 - q_0)^2(1 + q_0)} \leqslant 1. \tag{3.16}$$

Moreover, after setting,

$$c = \frac{(1 - q_0)\sqrt{1 - m}}{K\beta}$$

and

$$r = \varphi(p_0, q_0) - c$$

suppose that

$$B = \{x : \|x - x_0\| \leqslant r\} \subset D. \tag{3.17}$$

Then the sequence $\{x_n\}_{n=0}$ *generated by the iterative scheme*

$$A_n = A_{n-1} - A_{n-1}\,[F'(x_n)\,A_{n-1} - I] \tag{3.18}$$

$$x_{n+1} = x_n - A_n\,F(x_n), \tag{3.19}$$

remains in B *and converges to a root* x^* *of the equation* $F(x) = 0$. *Moreover, if we set, for* $n > 0$,

$$p_n = \| x_n - x_{n+1} \|$$

$$q_n = \| I - A_{n-1}\,F'(x_n) \|$$

the following relations hold

$$\| x_n - x^* \| \leqslant \varphi(p_n, q_n) - c \tag{3.20}$$

$$\| x_n - x^* \| \leqslant \varphi(p_{n-1}, q_{n-1}) - p_{n-1} - c\,. \tag{3.21}$$

Proof. First let us observe that (x_0, A_{-1}) is a φ-couple for F with associated (p_0, q_0). Indeed we have

$$\varphi(p_0, q_0) = \frac{(1 - q_0)}{K\beta}$$

so that from (3.12)

$$\| A_{-1} \| \leqslant \frac{1 - q_0}{K\,\varphi(p_0, q_0)}\,. \tag{3.22}$$

Since

$$A_0 = [I + (I - A_{-1}\,F'(x_0))]\,A_{-1}$$

according to (3.13) and (3.22) it follows that

$$\| A_0 \| \leqslant \frac{1 - q_0^2}{K\,\varphi(p_0, q_0)}\,.$$

Now, let us denote by $U = U(x_0, p_0, q_0)$ the set of all the φ-couples (x, A) for F which have an associated couple (p, q) such that the following relation is satisfied:

$$\| x - x_0 \| \leqslant \varphi(p_0, q_0) - \varphi(p, q)\,. \tag{3.23}$$

Obviously (x_0, A), with associated (p_0, q_0), belongs to U. Let (x, A), with (p, q), belong to U and let us consider the couple $(z, \bar{A})$ so defined:

$$\bar{A} = A - A(F'(x)\,A - I) \tag{3.24}$$

$$z = x - \bar{A}\,F(x). \tag{3.25}$$

Now we shall prove that $(z, \bar{A})$, with associated $(v(p,q), u(p,q))$, belongs to U.

From (3.25), (3.4) and (3.20) we have

$$\| z - x_0 \| \leqslant \| z - x \| + \| x - x_0 \| \leqslant \varphi(p_0, q_0) - (\varphi(p,q) - p)$$

and from (2.15), since $c = \varphi(0,0)$ and the function φ is increasing both in p and in q, it follows that

$$\| z - x_0 \| \leqslant \varphi(p_0, q_0) - \varphi(v(p,q), u(p,q)) < \varphi(p_0, q_0) - c. \tag{3.26}$$

Hence $z \in B$ and, by hypothesis (3.17), $z \in D$.

According to a well known result (cf. [4] p. 45), from (3.3) and (3.24) we have that both A and $\bar{A}$ are invertible. Using (3.25), we can write

$$F(z) = F(z) - F(x) - F'(x)(z - x) + \bar{A}^{-1}(\bar{A}F'(x) - I)(z - x). \tag{3.27}$$

From the above identity and from (3.10)', using Lemma 3.2, (3.2), (3.3), (3.4) and (3.5) one obtains

$$\| \bar{A} F(z) \| \leqslant \frac{(1 + u(p,q))(1 - q^2) p^2}{2\varphi(p,q)} + (1 + u(p,q)) p q^2$$

that is (3.9).

Therefore, according to Lemma 3.3, $(z, \bar{A})$ is a φ-couple for F with associated $(v(p,q), u(p,q))$; moreover from (3.26) we can conclude that it belongs to U.

We have so proved, by induction, that, for each n, x_n belongs to B and that the couple (x_n, A_{n-1}) belongs to U. Moreover, if we set $p_n = \| x_n - x_{n+1} \| = \| A_n F(x_n) \|$ and $q_n = \| I - A_{n-1} F'(x_n) \|$, we have

$$\begin{aligned} p_{n+1} &\leqslant v(p_n, q_n) \\ q_{n+1} &\leqslant u(p_n, q_n). \end{aligned} \tag{3.28}$$

Hence, according to properties (2.8), for each n and for each positive integer k, we have

$$p_{n+k} \leqslant v^{(k)}(p_n, q_n)$$

$$q_{n+k} \leqslant u^{(k)}(p_n, q_n).$$

Thus, by Theorem 2.2 and its corollary, we can conclude that the sequence $\{x_n\}_{n=0}$ converges and that (3.20) and (3.21) hold.

Finally, taking in (3.27) $z = x_{n+1}$, $x = x_n$, $\bar{A} = A_n$ and using the continuity of F, we observe that $F(x^*) = 0$.

So, the theorem just proved characterizes the initial quantities x_0 and A_{-1} in such a way the convergence of the method is assured.

Moreover, its hypotheses also guarantee the existence of a root of $F(x) = 0$ in D. With reference to this fact we shall now show that the crucial condition which has to be satisfied is (3.16).

THEOREM 3.5 - *For any* $K > 0, \beta > 0, p_0 > 0$ *and* $q_0 \in [0,1)$ *such that the condition* (3.16) *is not satisfied there exists a function* $F : R \to R$ *such that:*

1) *condition* (3.1) *is satisfied;*
2) *there exist two real values* x_0 *and* A_{-1} *such that hypotheses* (3.12), (3.13) *and* (3.14) *hold;*
3) *the equation* $F(x) = 0$ *has no real solution.*

Proof. If one takes the function

$$F(x) = \frac{K x^2}{2} + \frac{p_0}{(1 - q_0)} - \frac{(1 - q_0)^2}{2 K p_0 \beta^2}$$

and $A_{-1} = \beta, x_0 = \dfrac{(1 - q_0)}{\beta K}$, by simple computations, one can prove the theses.

Finally, the following result shows that we can consider (3.20) and (3.21) as sharp a posteriori error bounds.

THEOREM 3.6 - *For any* $K > 0, \beta > 0, p_0 > 0$ *and* $q_0 \in [0,1)$ *which satisfie condition* (3.16) *there exist a function* $F : R \to R$ *and two real values* x_0 *and* A_{-1} *such that all the other hypotheses of Theorem* 3.4 *are satisfied and* (3.20) *and* (3.21) *hold with equality for each n.*

Proof. Consider $F(x) = (K/2)(x^2 - c^2)$, were c is defined as in Theorem 3.4, and take $x_0 = (1 - q_0) / \beta K, A_{-1} = \beta$. Performing easy computations, one can see that in this case the method gives the iterative scheme (2.9) - (2.10).

REFERENCES

[1] HALD, O. H., *On a Newton-Moser type method.* Numer. Math. 23, 411-426 (1975).
[2] MIEL, G. J., *Majorizing sequences and error bounds for iterative methods.* Math. of Comp. 34, 185-202 (1980).
[3] MORET, I., *Sulla convergenza di certi metodi iterativi.* Rendiconti dell'Istituto di Matematica dell'Università di Trieste, XV, 108-115 (1983).
[4] ORTEGA, J. M., RHEINBOLDT, W. C., *Iterative solution of nonlinear equations in several variables.* Academic Press, New York, 1970.
[5] POTRA, F. A., PTAK, V., *Sharp error bounds for Newton's process.* Numer. Math. 34, 63-72 (1980).
[6] POTRA, F. A., PTAK, V., *A generalization of Regula Falsi.* Numer. Math. 36, 333-346 (1981).
[7] ZEHNDER, E. J., *A remark about Newton's method.* Communications on Pure and Applied Mathematics 27, 361-366 (1974).