GLASNIK MATEMATIČKI
Vol. 18 (38) (1983), 179-185.

NEWTON'S METHOD AND ROUND-OFF ERRORS

B. Guljaš, Zagreb

Abstract. In this paper we deal with the Newton's method perturbed by the rounding off error. Using the majorization technique, we give an upper bound for $\lim_n \sup \frac{\|x_n - x^*\|}{\|x^*\|}$, where $(x_n)_{n \in N}$ is perturbed Newton's sequence and x^* is a solution of equation $F(x) = 0$. This upper bound is expressed in terms of computing accuracy, condition number $K(F'(x^*))$, dimension of the problem and relative error bounds in computing $F(x)$ and $F'(x)$.

1. Introduction

Due to evident practical interest, in many papers dealing with Newton's iterative method, the analysis of convergence of perturbed sequences plays a great role. We find very elegant semilocal analysis of convergence in J. Rockne [4], for perturbed Newton's method, and in G. J. Miel [2], for perturbed Newton-type methods. The only computational disadvantage of these analyses is that error bounds in hypotheses and bounds in convergence results are expressed in absolute form. This is to say, in almost every computer the floating point aritmetics is used, and the natural way of expressing computing error is relative error bound.

From the computational point of view, an interesting analysis of Newton-type methods is given in J. C. P. Bus [1]. This author treats a class of numerically consistent methods and gives a local covergence result, which is actually based on perturbed contraction technic (see Ortega and Rheinboldt [3]; 12.2.3.). The linearity of this technique seems to be the main difficulty in determination of relatistic final relative error bound for iterative sequence.

In the present paper we shall avoid this difficulty by restricting our consideration to the pure Newton's method, using all the advantages of quadratic convergence property.

In the second section we treat a perturbed Newton's iterative method in an arbitrary Banach space, and in the third section this result is applied to the numerical Newton's method in R^m, giving to all the parameters clear computational meaning.

Mathematics subject classifications (1980): Primary 65 G 05, 65 H 10

Key words and phrases: Newton's method, rounding off error, relative error

Ovaj rad financirala je Samoupravna interesna zajednica za znanstveni rad u društveno-ekonomskoj djelatnosti SRH — SIZ VI.

2. General perturbed iterative method

In this section X is a Banach space with an arbitrary norm $\|\cdot\|$, D is an open convex subset of X and $F : D \to X$ is Fréchet differentiable on D such that

$$\|F'(x) - F'(y)\| \leqslant K\|x - y\| \text{ for all } x, y \in D. \tag{1}$$

We define Newton's method iterative function G as $G(x) = x - [F'(x)^{-1}]F(x)$ whenever $F'(x)$ is nonsingular. The object of our interest is an iterative method defined by iterative function H such that inequality

$$\|H(x) - G(x)\| \leqslant a\|x\| + \beta\|G(x) - x\| \tag{2}$$

holds for some nonnegative constants a and β.

Convergence analysis of the method of this type is based on majorization technique described in the next lemma.

LEMMA. *Let $\psi : [0,1) \to R$ be of the form*

$$\psi(s) = aa + \left(a + \frac{\beta + \frac{1}{2}(1-\beta)s}{1-s}\right)s$$

and $(y_{n-1})_{n\in N}$ a sequence such that $\|y_n\| \leqslant \psi(\|y_{n-1}\|)$, $n \in N$.
If

$$a \geqslant 0,\ \beta \geqslant 0,\ a \geqslant 0,\ a + \beta < 1 \tag{3a}$$

$$aa < (2-a)\left(1 - \left(1 - \left(\frac{1-a-\beta}{2-a}\right)^2\right)^{1/2}\right) \tag{3b}$$

and $\|y_0\| \leqslant s_0 \in [0, \gamma)$, then $\|y_n\| \leqslant s_n$, $n \in N$, and $\limsup_n \|y_n\| \leqslant \varepsilon$, where $s_n = \psi(s_{n-1})$, $n \in N$, and $0 \leqslant \varepsilon < \gamma \leqslant \frac{2}{3}$ are the roots of the equation

$$\left(\frac{3}{2} - a - \frac{1}{2}\beta\right)s^2 - (1 - a - \beta + aa)s + aa = 0. \tag{4}$$

Furthermore, if $\|y_0\| \geqslant \varepsilon$ then for $\Delta_n = \|y_n\| - \varepsilon$ the following inequality

$$\Delta_n \leqslant \frac{3}{2}\Delta_{n-1}^2 + \omega\Delta_{n-1} \tag{5}$$

holds, where $\omega \leqslant a + 9\beta + 6\varepsilon$.

Proof. Function ψ is strictly isotone for $s \in (0, 1)$. Condition (3) ensures existence of the roots of equation (4) such that $0 \leqslant \varepsilon < \gamma \leqslant \frac{2}{3}$ and $\psi(\varepsilon) = \varepsilon$, $\psi(\gamma) = \gamma$. Convergence of the sequence $(s_{n-1})_{n \in N}$ and $\lim_n s_n = \varepsilon$ is ensured by the condition $s_0 \in [0, \gamma)$, because $s < \psi(s)$ for $s \in (0, \varepsilon)$ and $\psi(s) < s$ for $s \in (\varepsilon, \gamma)$. Istonity of ψ and the induction argument ensure $\|y_n\| \leqslant s_n$, $n \in N$, and $\limsup_n \|y_n\| \leqslant \varepsilon$. Estimate (5) can be obtained from

$$\psi(s) - \varepsilon = \psi(s) - \psi(\varepsilon) = \left(\alpha + \frac{\beta + (1-\beta)\left(1 - \frac{1}{2}\varepsilon\right)}{(1-\varepsilon)(1-s)}\right)(s-\varepsilon) + \frac{\frac{1}{2}(1-\beta)}{1-s}(s-\varepsilon)^2.$$

In what follows, the open ball $\{x : \|x - x_0\| < r\}$ is denoted by $S(x_0, r)$ and its closure by $\bar{S}(x_0, r)$.

THEOREM 1. *Assume that there exists an* $x^* \in D$ *such that* $F(x^*) = 0$ *and* $\|F'(x^*)^{-1}\| \leqslant B^*$. *Let* $S_0 = S\left(x^*, \frac{1}{KB^*}\right)$ *be a subset of* D, *and* $H : S_0 \to X$ *such that estimate* (2) *holds for* $x \in S_1 = \bar{S}\left(x^*, \frac{2}{3KB^*}\right)$.

If the condition (3) *of lemma holds for* $a = KB^* \|x^*\|$, *and* $0 \leqslant \varepsilon < \gamma \leqslant \frac{2}{3}$ *are the roots of equation* (4), *then for* $x_0 \in S_2 = S\left(x^*, \frac{\gamma}{KB^*}\right)$ *the sequence* $(x_{n-1})_{n \in N}$ *defined by the iterative method* $x_n = H(x_{n-1})$, $n \in N$, *is contained in* S_2 *and* $\limsup_n \|x_n - x^*\| \leqslant \frac{\varepsilon}{KB^*}$.

Furthermore, for $d_n = \|x_n - x^*\| - \frac{\varepsilon}{KB^*}$ *with* $d_0 \geqslant 0$ *the estimate*

$$d_n \leqslant \frac{3KB^*}{2} d_{n-1}^2 + \omega d_{n-1}, \quad (\omega \leqslant \alpha + 9\beta + 6\varepsilon) \tag{6}$$

holds.

Proof. For $x \in S_0$ using (1) we have

$$\|F'(x) - F'(x^*)\| \leqslant K\|x - x^*\| < \frac{1}{B^*} \leqslant \frac{1}{\|F'(x^*)^{-1}\|}$$

and perturbation lemma (see [3]; 2.3.2.) gives the existence of $F'(x)^{-1}$ and

$$\|F'(x)^{-1}\| \leqslant \frac{\|F'(x^*)^{-1}\|}{1 - K\|F'(x^*)^{-1}\| \, \|x - x^*\|} \leqslant \frac{B^*}{1 - KB^* \|x - x^*\|}. \tag{7}$$

We have

$$G(x) - x^* = -F'(x)^{-1} [F(x) - F(x^*) - F'(x)(x - x^*)]$$

and by (7) and mean-value theorem (see [3], 3.2.12.) it follows

$$\|G(x) - x^*\| \leqslant \frac{1}{2} K \|F'(x)^{-1}\| \, \|x - x^*\|^2 \leqslant \frac{\frac{1}{2} KB^* \|x - x^*\|}{1 - KB^* \|x - x^*\|}. \tag{8}$$

Also, we have

$$\|G(x) - x\| \leqslant \|x - x^*\| + \|G(x) - x^*\| \leqslant \frac{1 - \frac{1}{2} KB^* \|x - x^*\|}{1 - KB^* \|x - x^*\|} \|x - x^*\|. \tag{9}$$

Now, for $x \in S_1$, because of (2), it follows

$$\begin{aligned} \|H(x) - x^*\| &\leqslant \|H(x) - G(x)\| + \|G(x) - x^*\| \leqslant \\ &\leqslant \alpha \|x\| + \beta \|G(x) - x\| + \|G(x) - x^*\| \leqslant \\ &\leqslant \alpha \|x^*\| + \alpha \|x - x^*\| + \beta \|G(x) - x\| + \|G(x) - x^*\|. \end{aligned} \tag{10}$$

Using (8) and (9) in (10) we have

$$\|H(x) - x^*\| \leqslant \alpha \|x^*\| + \left(\alpha + \frac{\beta + \frac{1}{2}(1 - \beta) KB^* \|x - x^*\|}{1 - KB^* \|x - x^*\|} \right) \|x - x^*\|. \tag{11}$$

Multiplying (11) with KB^*, putting $x = x_{n-1}$ and $H(x_{n-1}) = x_n$, we see that the conditions of Lemma are fulfilled, implying that $y_n = KB^* \|x_n - x^*\|$ and $a = KB^* \|x^*\|$.

Remark. One can easily notice that the smaller root of equation (4) is bounded by

$$\varepsilon \leqslant \frac{\alpha a}{1 - \left(\alpha + \beta + \frac{2\alpha a}{1 - \alpha - \beta + \alpha a} \right)}. \tag{12}$$

3. Numerical Newton's method

Each step of Newton's method, in actual computing process, can be separated into two parts. The first part is determination of the solution of the system of linear equations $F'(x)\,h = -F(x)$, and the second one is the correction $G(x) = x + h$. Therefore we can write the numerical Newton's iterative function H as

$$H(x) = \mathrm{fl}_\alpha(\mathrm{fl}_\alpha(G(x) - x) + x) \tag{13}$$

where α is the precision of computation used in the numerical process and $\mathrm{fl}_\alpha(\cdot)$ means that an expression inside the brackets is calculated in floating point aritmetics with the precision α (see J. C. P. Bus [1]). It is important to notice that the inequality

$$\|\mathrm{fl}_\alpha(x) - x\| \leqslant \alpha\|x\|,\ x \in R^m, \tag{14}$$

holds for all standard norms in R^m. Before stating our final result we have to analyse the computational solution of linear system $Ax = b$, where A is $m \times m$ matrix and x, b are m dimensional vectors. This solution behaves as the exact solution of the system $(A + E_1 + E_2)(x + h) = b + k$, where E_2 and k are errors in computing A and b, respectively, and E_1 is a consequence of the use of particular system solving method. For Gaussian method we have the estimate $\|E_1\| \leqslant \alpha g(m)\|A\|$, where g depends on dimension of the system, and its form depends on $\|\cdot\|$ and pivoting strategy (see Wilkinson [5]).

For the relative error the inequality

$$\frac{\|h\|}{\|x\|} \leqslant \frac{\varkappa(A)\left(\dfrac{\|E_1\| + \|E_2\|}{\|A\|} + \dfrac{\|k\|}{\|b\|}\right)}{1 - \varkappa(A)\dfrac{\|E_1\| + \|E_2\|}{\|A\|}},\quad \varkappa(A) = \|A^{-1}\|\,\|A\|, \tag{15}$$

holds.

Now, we have the following theorem for numerical Newton's method:

THEOREM 2. *Let* $X = R^m\ (m \in N)$ *and let the function* F *be as in Theorem* 1. *Further, assume that for* $x \in S_1$

$$\|\mathrm{fl}_\alpha(F(x)) - F(x)\| \leqslant \alpha\delta\|F(x)\| \tag{16}$$

$$\|\mathrm{fl}_\alpha(F'(x)) - F'(x)\| \leqslant \alpha\eta\|F'(x)\| \tag{17}$$

and denote by $\varkappa^* = \varkappa(F'(x^*)) = \|F'(x^*)^{-1}\|\,\|F'(x^*)\|$, $g = g(m)$,

$$\xi = 1 + (1 + \alpha)\frac{(3\varkappa^* + 2)(\delta + \eta + g)}{1 - \alpha(3\varkappa^* + 2)(\eta + g)}. \tag{18}$$

If $(3\mathscr{K}^* + 2)(\eta + g) < \frac{1}{\alpha}$, $1 + \xi < \frac{1}{\alpha}$ *and*

$$\alpha \|x^*\| < \frac{2-\alpha}{KB^*}\left(1 - \left(1 - \left(\frac{1-\alpha(1+\xi)}{2-\alpha}\right)^2\right)^{1/2}\right)$$

then for the numerical iterative function H, *defined by* (13), *all conditions of Theorem* 1 *are valid, assuming constants* α *and* $\beta = \alpha\xi$.

Proof. For $x \in S_1$

$$\mathscr{K}(F'(x)) \leqslant \frac{\mathscr{K}(F'(x^*)) + KB^* \|x - x^*\|}{1 - KB^* \|x - x^*\|} \leqslant 3\mathscr{K}^* + 2 \tag{19}$$

and from (15), using (16), (17) and (19), we get

$$\|\mathrm{fl}_\alpha(G(x) - x) - (G(x) - x)\| \leqslant \frac{\mathscr{K}(F'(x))\,\alpha(\delta + \eta + g)}{1 - \alpha\mathscr{K}(F'(x))(\eta + g)} \|G(x) - x\|$$

$$\leqslant \frac{\alpha(3\mathscr{K}^* + 2)(\delta + \eta + g)}{1 - \alpha(3\mathscr{K}^* + 2)(\eta + g)} \|G(x) - x\|. \tag{20}$$

Now, using (14), (18) and (20) we obtain

$$\begin{aligned}
\|H(x) - G(x)\| &\leqslant \|\mathrm{fl}_\alpha(\mathrm{fl}_\alpha(G(x) - x) + x) - \mathrm{fl}_\alpha(G(x) - x)\| + \\
&+ \|\mathrm{fl}_\alpha(G(x) - x) - (G(x) - x)\| \leqslant \\
&\leqslant \alpha \|\mathrm{fl}_\alpha(G(x) - x) + x\| + \|\mathrm{fl}_\alpha(G(x) - x) - (G(x) - x)\| \\
&\leqslant \alpha \|x\| + \alpha \|G(x) - x\| + (1 + \alpha) \|\mathrm{fl}_\alpha(G(x) - x) - (G(x) - x)\| \\
&\leqslant \alpha \|x\| + \alpha\xi \|G(x) - x\|.
\end{aligned}$$

Combining results of Theorem 2 and estimate (12) we get, in computational sense, very acceptable error bound:

$$\limsup_n \frac{\|x_n - x^*\|}{\|x^*\|} \leqslant \frac{\alpha}{1 - \alpha C},$$

where

$$C = 1 + \xi + \frac{2KB^* \|x^*\|}{1 - \alpha(1 + \xi - KB^* \|x^*\|)}.$$

It is evident that the influence of condition number $\mathscr{K}^*$, relative error bounds in computing F and F', and other constants included in C is very weak, because in actual computing process the precision is very high, so $\alpha C << 1$. This is justified by the experience.

REFERENCES:

[1] *J. C. P. Bus*, Convergence of Newton-like methods for solving systems of nonlinear equations, Numer. Math. **27** (1977), 271—281.

[2] *G. J. Miel*, Unified error analysis for Newton-type methods, Numer. Math. **33** (1979), 391—396.

[3] *J. M. Ortega* and *W. C. Rheinboldt*, Iterative solution of nonlinear equations in several variables, (in russian), MIR, Moscow 1975.

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

[5] *J. H. Wilkinson*, Rundungsfehler, Springer-Verlag, Berlin-Heidelberg 1969.

(Received October 9, 1981)

NEWTONOVA METODA I GREŠKE ZAOKRUŽIVANJA

B. Guljaš, Zagreb

Sadržaj

U ovom članku se proučava konvergencija Newtonove metode uz utjecaj greške zaokruživanja. Tehnikom majorizacije dobije se gornja međa za $\limsup_n \|x_n - x^*\| / \|x^*\|$, gdje je $(x_n)_{n \in N}$ perturbirani Newtonov niz, a x^* je rješenje jednadžbe $F(x) = 0$. Ta je međa izražena pomoću konstante stroja, uvjetovanosti $\varkappa(F'(x^*))$, dimenzije sistema i relativnih grešaka kod računanja $F(x)$ i $F'(x)$.