NOVI SAD J. MATH.
VOL. 29, NO. 1, 1999, 123-138

A NEW APPROACH TO GLOBALIZATION OF INEXACT NEWTON METHODS

Zorana Lužanin[1]

Abstract. One of the possible modifications of the Newton method for solving nonlinear system of equations is the inexact Newton method. The basic idea underlining this type of method is to approximate solution of Newton equation by applying linear iterative method. Many results for both local and global convergence are known for this method. This paper focuses on a wider class of methods, which are the result of preconditioning of the Newton equation by some nonsingular matrix A. By introducing the relaxation parameter t_k, we achieve, under suitable assumptions, global convergence. Mutual influence of parameters t_k and η_k also be discussed.

AMS Mathematics Subject Classification (1991): 65H10

Key words and phrases: Nonlinear system; Inexact Newton methods, Global convergence

1. Introduction

Consider the system of nonlinear equations

$$F(\mathbf{x}) = 0 \tag{1.1}$$

where $F : D \subset R^n \to R^n$ is differentiable. A well-known iterative method for solving (1.1) is the Newton method. At each iteration of this method the linear system

$$\mathcal{J}(\mathbf{x^k})\mathbf{s_N^k} = -\mathbf{F}(\mathbf{x^k}) \tag{1.2}$$

is solved, and the new approximation to the solution of (1.1) is defined by

$$\mathbf{x^{k+1}} = \mathbf{x^k} + \mathbf{s_N^k}. \tag{1.3}$$

The system (1.2) is called *Newton equation* and $\mathbf{s_N^k}$ is designated as *Newton correction.* The attractiveness of the Newton method stems from its property to converge quadratically from any sufficiently good initial guess. However, this method has a few drawbacks which motivated many researchers in the past three decades to alleviate them by introducing methods with the similar local convergence properties.

Application of direct method in solving the Newton equation can be rather expensive and difficult to implement for some structures of $\mathcal{J}(\mathbf{x^k})$ matrix (large

[1]Department of Mathematics and Physics, Faculty of Engineering, University of Novi Sad

n, etc). The problem can be overcome by use of the linear iterative method or by approximating the Jacobian matrix with an alternative matrix which renders the direct method significantly simpler. The methods based on such modification of the Newton method are known as inexact Newton methods. In fact, these methods yield approximation $\mathbf{s^k}$ of the Newton correction $\mathbf{s_N^k}$, so that the step (1.3) becomes $\mathbf{x^{k+1}} = \mathbf{x^k} + \mathbf{s^k}$.

Ortega and Rheinboldt [20] considered the case when a fixed number of iterations is used by a linear iterative method for the solution of (1.2). Under certain assumptions regarding the spectral radius of the iteration matrix at the solution, linear convergence can be obtained.

The other option involves the application of linear iterative method to solving system (1.2) until a solution is produced which can be considered "satisfactory enough". One such method was proposed by Dembo, Eisenstat and Steihaug [7]

Algorithm 1: Let $\mathbf{x^0} \in R^n$ be given.

For $k = 0, 1, ...$

Step 1: Find an increment $\mathbf{s^k}$ satisfying

$$||\mathcal{J}(\mathbf{x^k})\mathbf{s^k} + F(\mathbf{x^k})|| \leq \eta_k ||F(\mathbf{x^k})|| \tag{1.4}$$

where $\eta_k \geq 0$ is a real parameter

Step 2: Compute

$$\mathbf{x^{k+1}} = \mathbf{x^k} + \mathbf{s^k}. \tag{1.5}$$

Dembo et al. proved that the method defined by Algorithm 1 is locally and linearly convergent if η_k are uniformly bounded below one and convergence is superlinear if $\lim_{k\to\infty} \eta_k = 0$. The parameter η_k is also known as *forcing term*. Some selections of the forcing terms can be found in [8, ?, 11]. For reason of brevity, the method described by Algorithm 1 is designated as the classic inexact Newton method (CIN method).

The condition (1.4) can be modified in a number of ways. One of them is to precondition the Newton equation. Considered in Martínez [18, 19] is the case of preconditioning of the system (1.2) by the matrix B_k^{-1} ($B_k^{-1} \approx \mathcal{J}(\mathbf{x^k})^{-1}$) which is simple to compute. In this case, the condition (1.4) is replaced by the following condition

$$||B_k^{-1}(\mathcal{J}(\mathbf{x^k})\mathbf{s^k} + F(\mathbf{x^k}))|| \leq \eta_k ||B_k^{-1} F(\mathbf{x^k})|| \tag{1.6}$$

The method defined by (1.6) and (1.5) is called the preconditioned inexact Newton method (PIN method).

It is well known that the Newton method is affinely invariant [9]. This means that the Newton method applied to the affinely transformed problem

$$\tilde{F}(\mathbf{x}) = 0 \quad \tilde{F} = TF$$

where $T \in R^{n\times n}$ is a nonsingular matrix, produces precisely the same sequence $\{\mathbf{x^k}\}$ as it produces when used to solve (1.1). It is obvious that the inexact Newton method is not affinely invariant. Ypma [21] replaced condition (1.4) with the affinely invariant one

$$||\mathcal{J}(\mathbf{x^k})^{-1}(\mathcal{J}(\mathbf{x^k})\mathbf{s^k} + F(\mathbf{x^k}))|| \leq \eta_k||\mathcal{J}(\mathbf{x^k})^{-1}F(\mathbf{x^k})|| \tag{1.7}$$

If the condition (1.4) is replaced in Algorithm 1 by the condition (1.7), the method thus derived is designated as the affine variant of inexact Newton method (AIN method). It is obvious that CIN and AIN methods are special cases of PIN method, for $B_k = I$ and $B_k = \mathcal{J}(\mathbf{x^k})$ respectively. The local character of the inexact Newton method can be avoided. One often modifies this method in such a way that a decrease of some level function in every iteration is guaranteed. We replace the iterates (1.5) by

$$\mathbf{x^{k+1}} = \mathbf{x^k} + t_k\mathbf{s^k} \tag{1.8}$$

where the scalar $t_k > 0$ is chosen such that monotonicity of some level function $||AF(\mathbf{x})||$ (A is nonsingular matrix) is guaranteed, e.g.

$$||AF(\mathbf{x^{k+1}})|| < ||AF(\mathbf{x^k})||. \tag{1.9}$$

A special case of the previous condition, $A = I$, is frequently considered in the literature. The affinely invariant function can also be considered as a level function, i.e., the condition (1.9) can be replaced by

$$||\mathcal{J}(\mathbf{x^k})^{-1}F(\mathbf{x^{k+1}})|| < ||\mathcal{J}(\mathbf{x^k})^{-1}F(\mathbf{x^k})||.$$

The method defined by conditions (1.4) (or (1.6), i.e. (1.7)) and (1.8) is also known as the relaxed inexact Newton method (some other terms are also in use, such as "restrained", "damping", "step size control"). The local linear convergence of the relaxed method is known for

$$0 < t_k < \frac{2}{1+\eta_k}, \quad \text{and} \quad 0 \leq \eta_k < 1,$$

while the superlinear convergence requires $\lim_{k\to\infty}\eta_k = 0$ and $\lim_{k\to\infty} t_k = 1$.

This paper shall focus on the influence of the parameter t_k on global convergence. In other words, for a well-selected parameter t_k, the method is convergent for any $\mathbf{x^0}$ from the set D or its greater part. Defined in Section 2 is a new class of inexact Newton equations for which convergence is considered. Main convergence theorems are proved in the same section. Section 3 presents an analysis of parameter selection.

2. Global convergence of a class of the inexact Newton method

Our analysis of convergence will be restricted to the cases where $\mathbf{s^k}$ is defined to satisfy the following

$$||A(\mathcal{J}(\mathbf{x^k})\mathbf{s^k} + F(\mathbf{x^k}))|| \leq \eta_k ||AF(\mathbf{x^k})||, \tag{2.1}$$

where $A \in R^{n\times n}$ is a nonsingular. In other words, we consider a method defined by the following algorithm

Algorithm 2: Let $\mathbf{x^0}$ be given.
For k=0,1,...
Step 1: Define $\mathbf{s^k}$ such that (2.1) is satisfied for $\eta_k \geq 0$.
Step 2: Compute a new iteration

$$\mathbf{x^{k+1}} = \mathbf{x^k} + t_k \mathbf{s^k},$$

where t_k is a real parameter.
It is obvious that CIN method represents a special case of the previous method ($A \equiv I$), while the method (2.1) can be considered a special case of PIN method ($B_k^{-1} \equiv A$).

Preliminaries. Through the remainder of the paper, $||\cdot||$ denotes the Euclidean norm over R^n and Frobenius norm over $R^{n\times n}$. We shall also define the term level set, introduce some additional assumptions and notations.

Definition 1. *Let $A \in R^{n\times n}$ be nonsingular and $\mathbf{x} \in D$. The* **level set** *of F with respect to A and $\mathbf{x}$, denoted by $L(\mathbf{x}, A)$, is defined to be that path-connected component of the set*

$$\{\mathbf{y} \mid \mathbf{y} \in D,\ ||AF(\mathbf{y})|| \leq ||AF(\mathbf{x})||\}$$

which contains $\mathbf{x}$.

For the nonsingular $A \in R^{n\times n}$ and $\mathbf{x} \in D$ we introduce the following assumptions:

A1: $L(\mathbf{x}, A)$ is a compact set and $L(\mathbf{x}, A) \subseteq D$,
A2: $\mathcal{J}(\mathbf{z})$ is nonsingular and continuous for $\mathbf{z} \in L(\mathbf{x}, A)$,
A3: There exists $q > 0$ such that

$$||\mathcal{J}(\mathbf{y})^{-1}\mathcal{J}(\mathbf{z}) - I|| \leq q||\mathbf{y} - \mathbf{z}|| \quad \text{for } \mathbf{y}, \mathbf{z} \in L(\mathbf{x}, A).$$

To continue the analysis we define the following quantities:

Q1: $\beta(\mathbf{x}) = \sup_{\mathbf{y} \in L(\mathbf{x}, A)} ||\mathcal{J}(\mathbf{y})^{-1} F(\mathbf{x})||,$

Q2: $\kappa(\mathbf{x}) = ||A\mathcal{J}(\mathbf{x})|| \cdot ||(A\mathcal{J}(\mathbf{x}))^{-1}||,$

Q3: $a_0(\mathbf{x}) = \frac{1}{2}\kappa(\mathbf{x})(1 + \kappa(\mathbf{x})\eta(\mathbf{x}))),$

$$a_1(\mathbf{x}) = \frac{1 + 2\eta(\mathbf{x}) + 2\kappa(\mathbf{x})\eta(\mathbf{x})^2}{1 + \kappa(\mathbf{x})\eta(\mathbf{x})},$$

$$a_2(\mathbf{x}) = \frac{1 - \kappa(\mathbf{x})\eta(\mathbf{x})}{a_0(\mathbf{x})},$$

Q4: $P_3(\mathbf{x}, t) = a_0(\mathbf{x})t((q\beta(\mathbf{x})t)^2 + a_1(\mathbf{x})q\beta(\mathbf{x})t - a_2(\mathbf{x})).$

Let us also introduce the theorems essential to our further work

Theorem 1. *[6] Let $\mathbf{x} \in D$, D be an open set, and $A \in R^{n \times n}$. Suppose $F, \mathbf{x}$ and A satisfy A1 and A2. Then, there exists a unique differentiable function $p : [0, 1] \to L(\mathbf{x}, A)$ satisfying*

$$F(p(t)) = (1 - t)F(\mathbf{x}), \quad t \in [0, 1].$$

Moreover, p satisfies

$$\begin{aligned} p'(t) &= -\mathcal{J}(p(t))^{-1} F(\mathbf{x}), \quad t \in [0, 1], \\ p(0) &= \mathbf{x}. \end{aligned}$$

Furthermore, $\mathbf{x}^* = p(1)$ is a unique solution of $F(\mathbf{x}) = 0$ in $L(\mathbf{x}, A)$.

Lemma 1. *[6] Assume that $F : D \subset R^n \to R^m$ is differentiable on a convex set $D_0 \subset D$. Then, for any $\mathbf{x}, \mathbf{y} \in D_0$,*

$$||F(\mathbf{y}) - F(\mathbf{x})|| \leq \sup_{0 < t < 1} ||\mathcal{J}(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))|| \cdot ||\mathbf{y} - \mathbf{x}||.$$

Lemma 2. *[20] Let $F : D \subset R^n \to R^m$ be continuously differentiable on a convex set $D_0 \subset D$ and $\mathbf{x} \in D_0$. Suppose that*

$$||\mathcal{J}(\mathbf{y}) - \mathcal{J}(\mathbf{x})|| \leq \gamma(\mathbf{x})||\mathbf{y} - \mathbf{x}|| \quad \textit{for all } \mathbf{y} \in D_0.$$

Then

$$||F(\mathbf{y}) - F(\mathbf{x}) - \mathcal{J}(\mathbf{x})(\mathbf{y} - \mathbf{x})|| \leq \frac{1}{2}\gamma(\mathbf{x})||\mathbf{y} - \mathbf{x}||^2 \quad \textit{for all } \mathbf{y} \in D_0.$$

2.1 Convergence theorems. The first theorem to be proved, shows that the condition (2.1) is sufficient for the existence of the parameter t_k such that the condition (1.9) is satisfied, i.e. $\mathbf{s}^k$ is a descent direction.

Theorem 2. *Let F be differentiable on a non-empty and open set D. Let $\mathbf{x^k} \in D$ and $F(\mathbf{x^k}) \neq 0$, and let us denote*

$$z(t) = \mathbf{x^k} + t\mathbf{s^k},$$

where $\mathbf{s^k}$ satisfies the relation (2.1) for $\eta_k \in [0,1)$. Then there exists $t_0 > 0$ such that

$$||AF(z(t_k))|| < ||AF(\mathbf{x^k})|| \quad \textit{for all } t_k \in [0, t_0) \tag{2.2}$$

and $z(t_k) \in D$.

Proof. As D is an open and non-empty set and $z(0) = \mathbf{x} \in D$ there exists $t' > 0$ such that $z(t) \in D$ for all $t \in [0, t')$. Let us define the function $g : R \to R_+$

$$g(t) = ||AF(z(t))||^2,$$

Differentiation of $g(t)$ yields

$$g'(t) = 2 \cdot (A\mathcal{J}(z(t))\mathbf{s^k}, AF(z(t))).$$

By declaring $t = 0$ and using the Cauchy-Schwarz inequality and (2.1), we obtain

$$\begin{aligned} g'(0) &= 2(A\mathcal{J}(\mathbf{x^k})\mathbf{s^k}, AF(\mathbf{x^k})) \pm 2(AF(\mathbf{x^k}), AF(\mathbf{x^k})) \\ &\leq 2[(A(\mathcal{J}(\mathbf{x^k})\mathbf{s^k} + F(\mathbf{x^k})), AF(\mathbf{x^k})) - (AF(\mathbf{x^k}), AF(\mathbf{x^k}))] \\ &\leq 2(\eta_k - 1)||AF(\mathbf{x^k})||^2 < 0, \end{aligned}$$

for all $\eta_k \in [0,1)$. Therefore, there exists $t_0 \leq t'$ such that

$$g(t_k) < g(0)$$

is satisfied for all $t_k \in [0, t_0)$, i.e. (2.2) holds. □

The condition $||AF(\mathbf{x^{k+1}})|| < ||AF(\mathbf{x^k})||$ is not sufficient for the convergence of the iterative sequence $\{\mathbf{x^k}\}$ to the solution $\mathbf{x^*}$ of the problem (1.1). However, this problem can be worked around by imposing stricter conditions on the mapping F, set D or starting approximation $\mathbf{x^0}$. The problem for the Newton, quasi-Newton and CIN method was discussed, among others, by [1, 2, 3, 4, 6, 10, 12, 13, 14]. The next theorem is essential to the proof of convergence for the method defined by Algorithm 2.

Theorem 3. *Assume that $\mathbf{x} \in D$ such that $F(\mathbf{x}) \neq 0$ and A1,A2,A3 hold. Let $\mathbf{s}$ satisfy*

$$||A(\mathcal{J}(\mathbf{x})\mathbf{s} + F(\mathbf{x}))|| \leq \eta(\mathbf{x})||AF(\mathbf{x})||. \tag{2.3}$$

Moreover, suppose that there exists a constant $\theta > 1$ such that

$$\eta(\mathbf{x})\mathrm{K}(\mathbf{x}) < \frac{1}{\theta} \tag{2.4}$$

then, for all t satisfying

$$0 \le t \le \min(1, \xi(t)),$$

where $\xi(x)$ is a real positive root of the equation $P_3(\mathbf{x}, t) = 0$, i.e.

$$\xi(\mathbf{x}) = \frac{-a_1(\mathbf{x}) + \sqrt{a_1(\mathbf{x})^2 + 4a_2(\mathbf{x})}}{2q\beta(\mathbf{x})} \tag{2.5}$$

we have $\mathbf{x} + t\mathbf{s} \in L(\mathbf{x}, A)$ and

$$||AF(\mathbf{x} + t\mathbf{s})|| \le (1 + P_3(\mathbf{x}, t))||AF(\mathbf{x})||. \tag{2.6}$$

Proof. According to Theorem 1, there exists a unique function $p : [0, 1] \to L$, $L = L(\mathbf{x}, A)$, such that for all $t \in [0, 1]$ holds

$$F(p(t)) = (1 - t)F(\mathbf{x}) \tag{2.7}$$

$$p'(t) = -\mathcal{J}(p(t))^{-1}F(\mathbf{x}), \quad p(0) = \mathbf{x}. \tag{2.8}$$

Let us define the functions $z : [0, 1] \to R^n$, $w : [0, 1] \times [0, 1] \to R^n$, $\delta : [0, 1] \to R$ and set L_0 as follows

$$z(t) = \mathbf{x} + t\mathbf{s}, \tag{2.9}$$

$$w(t, v) = p(t) + v(z(t) - p(t)), \tag{2.10}$$

$$\delta(t) = \sup\{v \mid v \in [0, 1], w(t, v') \in L \text{ for all } v' \in [0, v]\}, \tag{2.11}$$

$$L_0 = \{\mathbf{x} \mid \mathbf{x} = w(t, v),\ t \in [0, 1], v \in [0, \delta(t)]\}. \tag{2.12}$$

From the continuity of the function F and (2.7), there follows $\delta(t) \neq 0$ for $t \in (0, 1]$. For $t = 0$, $w(0, v) = \mathbf{x}$ which means $\delta(0) = 1$. Therefore

$$\delta(t) \neq 0 \quad \text{for all } t \in [0, 1].$$

According to (2.11) we have $w(t, v) \in L$ for all $t \in [0, 1]$ and $v \in [0, \delta(t))$, while from the condition of compactness of the set L follows $w(t, \delta(t)) \in L$ and, thus $L_0 \subset L$.

Next follows the proof of the relation (2.6). Let us select $t \in [0,1]$ and for $v \in (0, \delta(t)]$ apply the mean value theorem,

$$F(w(t,v)) = F(w(t,0)) + \frac{1}{v}\int_0^v \mathcal{J}(w(t,v'))dv' \, (w(t,v) - w(t,0)).$$

By the definition of the function w and relation (2.7) we have

$$\begin{aligned} AF(w(t,v)) &= (1-t)AF(\mathbf{x}) + A\mathcal{J}(\mathbf{x})(w(t,v) - p(t)) + \\ &\quad + \frac{1}{v}A\int_0^v (\mathcal{J}(w(t,v')) - \mathcal{J}(\mathbf{x}))dv' \, (w(t,v) - p(t)), \end{aligned}$$

that is

$$\begin{aligned} ||AF(w(t,v))|| &\leq (1-t)||AF(\mathbf{x})|| + \Big(||A\mathcal{J}(\mathbf{x})|| + \\ &\quad \frac{1}{v}\int_0^v ||A(\mathcal{J}(w(t,v')) - \mathcal{J}(\mathbf{x}))||dv' \Big) ||w(t,v) - p(t)||. \end{aligned} \tag{2.13}$$

Prior to estimation of

$$I_1 = ||w(t,v) - p(t)|| \quad \text{and} \quad I_2 = \frac{1}{v}\int_0^v ||A(\mathcal{J}(w(t,v')) - \mathcal{J}(\mathbf{x}))||dv',$$

let us derive some additional estimations. According to (2.8) and assumption A3 there follows

$$\begin{aligned} ||p'(t) - p'(0)|| &= || - \mathcal{J}(p(t))^{-1}F(\mathbf{x}) + \mathcal{J}(p(0))^{-1}F(\mathbf{x})|| \\ &= ||(\mathcal{J}(p(t))^{-1}\mathcal{J}(\mathbf{x}) - I)\mathcal{J}(\mathbf{x})^{-1}F(\mathbf{x})|| \\ &\leq q||p(t) - \mathbf{x}|| \cdot ||\mathcal{J}(\mathbf{x})^{-1}F(\mathbf{x})||. \end{aligned} \tag{2.14}$$

Using Lemma 1, we have

$$||p(t) - \mathbf{x}|| \leq t\beta(\mathbf{x}). \tag{2.15}$$

The combination of the relations (2.14) and (2.15) yields

$$||p'(t) - p'(0)|| \leq q\beta(\mathbf{x})t||\mathcal{J}(\mathbf{x})^{-1}F(\mathbf{x})||.$$

Now Lemma. 2 can be applied to the function p, which yields

$$\begin{aligned} ||p(t) - \mathbf{x} - p'(0)t|| &\leq \frac{1}{2}q\beta(\mathbf{x})||\mathcal{J}(\mathbf{x})^{-1}F(\mathbf{x})||t^2 \\ &\leq \frac{1}{2}q\beta(\mathbf{x})||(A\mathcal{J}(\mathbf{x}))^{-1}|| \cdot ||AF(\mathbf{x})||t^2. \end{aligned} \tag{2.16}$$

Estimation for I_1 : According to (2.10) and (2.8), there follows

$$\begin{aligned} w(t,v) - p(t) &= v(z(t) - p(t) \pm tp'(0)) \\ &= v\Big((\mathbf{x} - p(t) + tp'(0)) + t(\mathbf{s} + \mathcal{J}(\mathbf{x})^{-1}F(\mathbf{x})) \Big) \end{aligned}$$

that is, using (2.16) and (2.3) we have

$$\begin{aligned} I_1 &\leq v\Big(||\mathbf{x} - p(t) + tp'(0)|| + t||\mathcal{J}(\mathbf{x})^{-1}A^{-1}|| \cdot ||A(\mathcal{J}(\mathbf{x})\mathbf{s} + F(\mathbf{x}))|| \Big) \\ (2.17) \quad &\leq v\left(\frac{1}{2}q\beta(\mathbf{x})t^2 + t\eta(\mathbf{x})\right) ||(A\mathcal{J}(\mathbf{x}))^{-1}|| \cdot ||AF(\mathbf{x})|| \end{aligned}$$

Estimation for I_2 : According to A3, we have

$$\begin{aligned} ||A(\mathcal{J}(w(t,v')) - \mathcal{J}(\mathbf{x}))|| &\leq ||A\mathcal{J}(\mathbf{x})|| \cdot ||\mathcal{J}(\mathbf{x})^{-1}\mathcal{J}(w(t,v')) - I|| \\ &\leq ||A\mathcal{J}(\mathbf{x})||q||w(t,v') - \mathbf{x}||, \end{aligned}$$

while, according to (2.8), (2.10) and (2.15), there follows

$$\begin{aligned} ||w(t,v') - \mathbf{x}|| &= ||p(t) + v'(z(t) - p(t) \pm tp'(0)) - p(0)|| \\ &\leq (1-v')||p(t) - p(0)|| + v't||\mathbf{s} + \mathcal{J}(\mathbf{x})^{-1}F(\mathbf{x})|| + v't||\mathcal{J}(\mathbf{x})^{-1}F(\mathbf{x})|| \\ &\leq (1-v')t\beta(\mathbf{x}) + v't\eta(\mathbf{x})||(A\mathcal{J}(\mathbf{x}))^{-1}|| \cdot ||AF(\mathbf{x})|| + v't\beta(\mathbf{x}) \\ &\leq t\beta(\mathbf{x}) + t\eta(\mathbf{x})||(A\mathcal{J}(\mathbf{x}))^{-1}|| \cdot ||AF(\mathbf{x})||. \end{aligned}$$

Finally,

$$(2.18) \qquad I_2 \leq ||A\mathcal{J}(\mathbf{x})||qt\left(\beta(\mathbf{x}) + \eta(\mathbf{x})||(A\mathcal{J}(\mathbf{x}))^{-1}|| \cdot ||AF(\mathbf{x})||\right)$$

Using $||(A\mathcal{J}(\mathbf{x}))^{-1}|| \cdot ||AF(\mathbf{x})|| \leq \kappa(\mathbf{x})\beta(\mathbf{x})$ and the relations (2.17) and (2.18), from (2.13) there follows

$$\begin{aligned} ||AF(w(t,v))|| \leq &\Big((1-t) + \kappa(\mathbf{x})(\frac{1}{2}q\beta(\mathbf{x})t + \eta(\mathbf{x}))\cdot \\ &\cdot(1 + t\beta(\mathbf{x})(1 + \eta(\mathbf{x})\kappa(\mathbf{x}))\Big) ||AF(\mathbf{x})||. \end{aligned}$$

After rearranging the right-hand side, we have

$$(2.19) ||AF(w(t,v))|| \leq (1 + P_3(\mathbf{x},t))||AF(\mathbf{x})|| \quad t \in [0,1] \text{ and } v \in (0,\delta(t)]$$

Since $a_2(\mathbf{x}) > 0$, due to the condition (2.4) there exists a unique, positive root to the polynomial $P_3(\mathbf{x},t)$ which is given as the relation (2.5). According to the condition of compactness of the set L and the continuity of $\mathcal{J}$, as well as definitions of the functions w and δ it is possible to show that $\delta(t) = 1$ for $t \in [0, \min(1, \xi(\mathbf{x}))]$.

Let us suppose the opposite,

$$\exists t^* \in (0,1), \quad t^* < \xi(\mathbf{x}), \quad \delta(t^*) < 1.$$

Then from (2.19) follows $||AF(w(t^*,v)|| < ||AF(\mathbf{x})||$ for $v \in [0,\delta(t^*)]$, i. e. $w(t^*,\delta(t^*)) \in int(L)$. From the condition A2 F is a local homeomorphism for all $\mathbf{y} \in int(L)$, and there exists $\rho > 0$ such that

$$S(w(t^*,\delta(t^*)),\rho) = \{\mathbf{y} |\ ||\mathbf{y} - w(t^*,\delta(t^*)|| < \rho\} \subset L.$$

Accordingly, $w(t^*, v) \in int(L)$ for all $v < \delta(t^*) + \rho$ which contradicts the definition of $\delta(t)$.

Now we can assume that $v = 1$, thus from the relation (2.19) follows

$$||AF(\mathbf{x} + t\mathbf{s})|| \leq (1 + P_3(\mathbf{x}, t))||AF(\mathbf{x})|| \quad \text{for } t \in [0, \min(1, \xi(\mathbf{x}))]$$

which proves Theorem 3. □

The previous theorem provides the existence of a real parameter t such that the iterative sequence is well defined and convergent, which will be useful for proving the main convergence theorem. The next theorem will allow us to estimate the rate of convergence.

Theorem 4. *Let the conditions of Theorem 3 be satisfied, and let α satisfy*

$$0 < \alpha \leq \min(1, \frac{\theta - 1}{4.5q\beta(\mathbf{x})\mathrm{K}(\mathbf{x})(1 + \theta)}). \tag{2.20}$$

Then $\alpha \leq \min(1, \xi(\mathbf{x}) - \alpha)$, and for all t with

$$\alpha \leq t \leq \min(1, \xi(\mathbf{x}) - \alpha) \tag{2.21}$$

we have $\mathbf{x} + t\mathbf{s} \in L(\mathbf{x}, A)$ and

$$||A(\mathbf{x} + t\mathbf{s})|| \leq (1 - \frac{\alpha^2}{4}(1 - \frac{1}{\theta}))||AF(\mathbf{x})||. \tag{2.22}$$

Proof. From $\mathrm{K}(\mathbf{x}) \geq 1$, and according to the condition $\eta(\mathbf{x})\mathrm{K}(\mathbf{x}) < 1/\theta$ we have

$$2\frac{\theta - 1}{\mathrm{K}(\mathbf{x})(1 + \theta)} < a_2(\mathbf{x}) < \frac{2}{\mathrm{K}(\mathbf{x})} \leq 2.$$

Using the inequality $-b + \sqrt{b^2 + x} \geq \dfrac{x}{2b + \sqrt{x}}$ for $b, x > 0$ we obtain estimation for $\xi(\mathbf{x})$

$$\xi(\mathbf{x}) \geq \frac{a_2(\mathbf{x})}{q\beta(\mathbf{x})\left(a_1(\mathbf{x}) + \sqrt{a_2(\mathbf{x})}\right)}. \tag{2.23}$$

If we use $\mathrm{K}(\mathbf{x}) \geq 1$, $\theta > 1$, that is $\eta(\mathbf{x}) < 1$, then the estimation for $a_1(\mathbf{x})$ is obtained

$$a_1(\mathbf{x}) = \frac{1 + 2\eta(\mathbf{x}) + 2\mathrm{K}(\mathbf{x})\eta(\mathbf{x})^2}{1 + \mathrm{K}(\mathbf{x})\eta(\mathbf{x})} \leq 1 + 2\eta(\mathbf{x}) < 3.$$

Now we have

$$a_1(\mathbf{x}) + \sqrt{a_2(\mathbf{x})} < 3 + \sqrt{2} < 4.5,$$

which in combination with (2.23) yields

$$\xi(\mathbf{x}) \geq \frac{a_2(\mathbf{x})}{4.5q\beta(\mathbf{x})} \geq \frac{2(\theta-1)}{4.5q\beta(\mathbf{x})\mathrm{K}(\mathbf{x})(1+\theta)}$$

leading us to the conclusion that for α, which satisfies (2.20), we have $\alpha \leq \xi(\mathbf{x})/2$ which yields

$$\alpha \leq \min(1, \xi(\mathbf{x}) - \alpha). \tag{2.24}$$

Knowing that t from the interval (2.21) is narrower than the interval $[0, \min(1, \xi(\mathbf{x}))]$ according to Theorem 3 there follows $\mathbf{x} + t\mathbf{s} \in L(\mathbf{x}, A)$.

In order to prove the relation (2.22) it is sufficient to show

$$P_3(\mathbf{x}, t) \leq -\frac{\alpha^2}{4}(1 - \frac{1}{\theta}) \text{ for } t \in [\alpha, \min(1, \xi(\mathbf{x}) - \alpha)]. \tag{2.25}$$

Let us denote

$$P_2(\mathbf{x}, t) = (q\beta(\mathbf{x}))^2 t^2 + a_1(\mathbf{x})q\beta(\mathbf{x})t - a_2(\mathbf{x}).$$

The polynomial $P_2(\mathbf{x}, t)$ has one real positive root $\xi(\mathbf{x})$ which is given by (2.5) and a minimum at the point

$$t^* = -\frac{a_1(\mathbf{x})}{2q\beta(\mathbf{x})} < 0.$$

Accordingly, $P_2(\mathbf{x}, t)$ is negative and increasing for $t \in [0, \min(1, \xi(\mathbf{x}))]$.

Let us denote

$$\bar{\xi} = \min(1, \xi(\mathbf{x}) - \alpha),$$

then $P_2(\mathbf{x}, t) \leq P_2(\mathbf{x}, \bar{\xi})$ and

$$P_3(\mathbf{x}, t) = a_0(\mathbf{x})tP_2(\mathbf{x}, t) \leq a_0(\mathbf{x})\alpha P_2(\mathbf{x}, \xi(\mathbf{x})) \quad \text{for } t \in [\alpha, \bar{\xi}]. \tag{2.26}$$

Now the problem is reduced to proving the relation

$$P_2(\mathbf{x}, \bar{\xi}) \leq -\frac{\alpha}{4}a_2(\mathbf{x}). \tag{2.27}$$

By introducing $d = \xi(\mathbf{x}) - \bar{\xi}$, there follows

$$\begin{aligned} P_2(\mathbf{x}, \bar{\xi}) &= P_2(\mathbf{x}, \bar{\xi}) - P_2(\mathbf{x}, \xi(\mathbf{x})) \\ &= -(q\beta(\mathbf{x}))^2 d(\bar{\xi} + \xi(\mathbf{x})) - a_1(\mathbf{x})q\beta(\mathbf{x})d \end{aligned}$$

so, using (2.5) and $\bar{\xi} = \xi(\mathbf{x}) - d$ there follows

$$P_2(\mathbf{x}, \bar{\xi}) = q\beta(\mathbf{x})d\left(q\beta(\mathbf{x})d - \sqrt{a_1(\mathbf{x})^2 + 4a_2(\mathbf{x})}\right). \tag{2.28}$$

Finally, we can estimate

$$O_1 = q\beta(\mathbf{x})d \quad \text{and} \quad O_2 = q\beta(\mathbf{x})d - \sqrt{a_1(\mathbf{x})^2 + 4a_2(\mathbf{x})}.$$

We shall discern between two cases, depending on $\bar{\xi}$.

Case I: $\bar{\xi} = \xi(\mathbf{x}) - \alpha$, i. e. $d = \alpha$

Using the relations (2.24) and (2.5) we have

$$\alpha \leq \frac{1}{4q\beta(\mathbf{x})}\left(-a_1(\mathbf{x}) + \sqrt{a_1(\mathbf{x})^2 + 4a_2(\mathbf{x})}\right).$$

Multiplying the previous relation by $q\beta(\mathbf{x})$, adding $-\sqrt{a_1(\mathbf{x})^2 + 4a_2(\mathbf{x})}$ and, finally, using $\alpha = d$, we have

$$O_2 \leq -\frac{1}{4}\left(a_1(\mathbf{x}) + \sqrt{a_1(\mathbf{x})^2 + 4a_2(\mathbf{x})}\right) < 0. \tag{2.29}$$

Using $\bar{\xi} < 1$ and $d < 1$ i. e. $\xi(\mathbf{x}) < 2$ there follows

$$2 > \frac{-a_1(\mathbf{x}) + \sqrt{a_1(\mathbf{x})^2 + 4a_2(\mathbf{x})}}{2q\beta(\mathbf{x})}.$$

Multiplying the previous relation by $dq\beta(\mathbf{x})/2$ and using $d = \alpha$ there follows

$$O_1 > \frac{1}{4}\alpha\left(-a_1(\mathbf{x}) + \sqrt{a_1(\mathbf{x})^2 + 4a_2(\mathbf{x})}\right) > 0. \tag{2.30}$$

Thus, from (2.28) and according to (2.29) and (2.30) there follows the relation (2.27).

Case II: $\bar{\xi} = 1$ Using (2.5) and $\bar{\xi} = 1$ there follows

$$1 = \frac{-a_1(\mathbf{x}) + \sqrt{a_1(\mathbf{x})^2 + 4a_2(\mathbf{x})}}{2q\beta(\mathbf{x})} - d.$$

Multiplying the previous equality by $q\beta(\mathbf{x})d$, after rearranging there follows

$$dq\beta(\mathbf{x}) = \frac{d}{2(1+d)}\left(-a_1(\mathbf{x}) + \sqrt{a_1(\mathbf{x})^2 + 4a_2(\mathbf{x})}\right) \tag{2.31}$$

Since the function $h : R_+ \to R$ defined by $h(y) = y/(1+y)$ is increasing and $h(y) \leq 1$ from (2.31) there follows

$$O_2 < -\frac{1}{2}\left(a_1(\mathbf{x}) + \sqrt{a_1(\mathbf{x})^2 + 4a_2(\mathbf{x})}\right) < 0. \tag{2.32}$$

If we make use of the facts that $\alpha \leq \xi(\mathbf{x}) - \bar{\xi} = d$ and $h(\alpha) < h(d)$, from (2.31) there follows

$$O_1 \geq \frac{\alpha}{2(1+\alpha)}\left(-a_1(\mathbf{x}) + \sqrt{a_1(\mathbf{x})^2 + 4a_2(\mathbf{x})}\right) > 0. \tag{2.33}$$

Using (2.32) and (2.33) as well as the condition $\alpha \leq 1$, from the relation (2.28) follows (2.27).

Thus, from the proven relation (2.27), from (2.26) we have

$$P_3(\mathbf{x},t) \leq -\frac{\alpha^2}{4}a_0(\mathbf{x})a_2(\mathbf{x}). \tag{2.34}$$

From the definitions of $a_0(\mathbf{x})$ and $a_1(\mathbf{x})$ and the condition (2.4) it follows

$$a_0(\mathbf{x})a_2(\mathbf{x}) \leq 1 - \frac{1}{\theta},$$

i.e. the relation (2.25).

Since$[\alpha, \xi(\mathbf{x}) - \alpha] \subset [0, \xi(\mathbf{x})]$, from Theorem 3 follows (2.22), which should be proven. □

2.2 Main Theorem. We can now finally get to the proof of the theorem of global convergence for the considered iterative method. For the sake of simplicity, let us denote $\eta_k = \eta(\mathbf{x})$.

Theorem 5. *Let $\mathbf{x^0} \in D$ and $A \in R^{n\times n}$ be a nonsingular matrix such that A1, A2, A3 are satisfied. Let the sequence $\{\mathbf{x^k}\}$ be defined by Algorithm 2, $F(\mathbf{x^k}) \neq 0$, and let there exist a constant $\theta > 1$ such that*

$$\eta_k \mathrm{K}(\mathbf{x^k}) < \frac{1}{\theta} \quad k = 0, 1, 2 \ldots \tag{2.35}$$

If the constant α satisfies

$$0 < \alpha \leq \min(1, \frac{\theta - 1}{4.5q\beta(\mathbf{x^k})\mathrm{K}(\mathbf{x^k})(1+\theta)}) \tag{2.36}$$

and the parameter t_k

$$\alpha \leq t_k \leq \min(1, \xi(\mathbf{x^k}) - \alpha) \tag{2.37}$$

then the following holds

(i) the sequence $\{\mathbf{x^k}\}$ is well defined and $\{\mathbf{x^k}\} \subset L(\mathbf{x^0}, A)$

(ii) the sequence $\{\mathbf{x^k}\}$ converges to the unique solution $\mathbf{x}^* \in L(\mathbf{x^0})$

(iii) the there exists $k_0 \geq 0$ such that $t_k = 1$ satisfies (2.37) for $k \geq k_0$.

Proof. We use induction to prove the statement *(i)*. According to the definition of the level set, for $k = 0$, $\mathbf{x^0} \in L(\mathbf{x^0}, A)$. Let us suppose that $\mathbf{x^k} \in L(\mathbf{x^0}, A)$. Then, according to the definition and compactness of the set $L(\mathbf{x^0}, A)$, the level set $L(\mathbf{x^k}, A)$ is also compact and $L(\mathbf{x^k}, A) \subset L(\mathbf{x^0}, A)$ holds. Therefore, for $\mathbf{x^k}$ the conditions from Theorems 3 and 4 are satisfied and their applications yields

$$\mathbf{x^{k+1}} = \mathbf{x^k} + t_k\mathbf{s^k} \in L(\mathbf{x^k}, A) \quad \text{for all} \quad t_k \in [0, \min(1, \xi(\mathbf{x^k}))]$$

and

$$||AF(\mathbf{x^{k+1}})|| \leq (1 - \frac{\alpha^2}{4}(1 - \frac{1}{\theta}))||AF(\mathbf{x^k})|| \tag{2.38}$$

for t_k which satisfies (2.37).

As $\theta > 1$ and $\alpha \leq 1$ do not depend on k, from (2.38) follows $\lim_{k\to\infty} ||AF(\mathbf{x^k})|| = 0$.

Due to the compactness of the set $L(\mathbf{x^0}, A)$ there exists a sub-sequence of the sequence $\{\mathbf{x^k}\}$ which converges to $\mathbf{x^*} \in L(\mathbf{x^0}, A)$ such that $F(\mathbf{x^*}) = 0$. The uniqueness of the solution follows from Theorem 1. Thus, the statement *(ii)* is proved.

In order to prove the statement *(iii)* we shall use the fact that $\lim_{k\to\infty} \beta(\mathbf{x^k}) = 0$, and that there exists $k_0 > 0$ such that

$$\xi(\mathbf{x^k}) \geq 2 \quad \text{for all} \quad k \geq k_0,$$

that is, $\min(1, \xi(\mathbf{x^k}) - \alpha) = 1$, so we can assume $t_k = 1$ for all $k \geq k_0$. □

The relation (2.38) illustrates the rate of convergence. It is obvious that the relation $1 - \frac{\alpha^2}{4}(1 - \frac{1}{\theta})$ decreases with the increase of θ. For the Newton method $\eta_k \equiv 0$ which allows us to apply an arbitrarily large θ and thus generate the fastest method in the considered class of methods.

Example 1. Let $f(x) = \ln x$, $x_0 = 10$ and $A = 1$. It is obvious that the Newton method is not converging, i.e. it is not well defined. Since $L(x_0) = [0.1, 10]$ is a compact set, $f'(x) = 1/x \neq 0$, $x \in L(x_0)$, $|f'(y)^{-1}f'(z) - 1| \leq 10|y - z|$, $y, z \in L(x_0)$ and $\kappa(x) = 1$, $x \in L(x_0)$, Theorem 5 can be applied, i.e. the method defined by Algorithm 2 is convergent.

3. Applications and summary

The majority of the so far proposed methods for determination of the parameter t_k use *backtracking* which can be rather expensive, i.e., slow. The approach used in this paper allows this problem to be alleviated. According to the relation (2.35) we can determine a parameter η_k which allows simplified determination of parameter t_k. For practical purposes, we can first select a constant $\theta > 1$ and approximate the value of $\kappa(\mathbf{x^k}) \leq C_k$ and finally select η_k such that $\eta_k < \frac{1}{\theta C_k}$ is satisfied. Should we also approximate the value of $\beta(\mathbf{x^k})$ (e.g., if $||\mathcal{J}(\mathbf{x})^{-1}|| \leq M$ then $\beta(\mathbf{x^k}) \leq M||F(\mathbf{x^k})||$) then the approximation for $\xi(\mathbf{x^k})$ is obtained which yields the interval for t_k based on the relation from Theorem 3. The purpose of the matrix A is to simplify the application of linear iterative method in Step 1 of Algorithm 2 or to influence $\kappa(\mathbf{x^k})$ thus, indirectly determining η_k.

The polynomial $1 + P_3(\mathbf{x^k}, t)$ reaches minimum at the point

$$t^k_{opt} = \frac{-a_1(\mathbf{x^k}) + \sqrt{a_1(\mathbf{x^k})^2 + 3a_2(\mathbf{x^k})}}{3q\beta(\mathbf{x^k})}. \tag{3.1}$$

Applying the technique similar to that from the proof of Theorem 4, it can be shown that t^k_{opt} belongs to the interval (2.37).

Example 2. For the function from Example 1 there follow $q = 10$, $\beta(x^k) = x^k \ln x^k$, $x^k > 1$ and $\beta(x^k) = -\ln x^k / x^k$, $x^k < 1$, so that the optimal parameter t^k_{opt} can be precisely calculated for each iteration.

In order to establish parameter t_k one needs to know the values of $\beta(\mathbf{x^k})$, $\kappa(\mathbf{x^k})$ and q. This could prove to be a serious impediment in which case these values can only be approximated. The most simple method (when $M > 0$ is known, such that $||\mathcal{J}(\mathbf{x})^{-1}|| \leq M$, $\mathbf{x} \in D$) has already been mentioned, in which case we get the following estimations:

$$\begin{aligned} \kappa(\mathbf{x^k}) &\leq M\kappa_A ||\mathcal{J}(\mathbf{x^k})|| \quad (\kappa_A = ||A|| \cdot ||A^{-1}||) \\ \beta(\mathbf{x^k}) &\leq M||F(\mathbf{x^k})||. \end{aligned}$$

The second possibility lies in the application of a quasi-Newton method. Let H_k be an approximation of matrix $\mathcal{J}(\mathbf{x^k})^{-1}$. Among others, such methods can be found in Broyden [5] and Martínez[17]. The values $\beta(\mathbf{x^k})$ and $\kappa(\mathbf{x^k})$ are approximated by

$$\tilde{\beta}_k = ||H_k|| \cdot ||F(\mathbf{x^k})||, \qquad \tilde{\kappa}_k = ||A^{-1}H_k|| \cdot ||A\mathcal{J}(\mathbf{x^k})|| \tag{3.2}$$

respectively.

Let us note that such globalization of inexact Newton method could be relatively easily applied to the method defined in Martínez [18] by Algorithm 2.1.

The influence of the parameter t_k on convergence of CIN method was considered in Brown, Saad [4], Eisenstat, Walker [10], Lanzkron, Rose, Wilkes [14]. Beside focusing on a broader class of methods, the authors also changed the approach to defining the parameter η_k. This parameter is selected according to the mapping F, thus allowing simplified determination of the parameter t_k. Global convergence of AIN method was considered in Lužanin, Krejić, Herceg[15]. Future research should be aimed at considering the PIN class of methods.

Another point to be considered is the possibility of weakening or alteration of the conditions A1-A3.

Finally, the future efforts shall be focused on estimation of values which are essential to the determination of the parameters η_k and t_k. One of the solutions to be considered is the estimation of the inverse Jacobian. In this way some concrete methods will be generated which can be easily applied to a wider class of systems.

References

[1] Bank, R.E., Rose, D.J., Parameter selection for Newton-like methods applicable to nonlinear partial differential equations, SIAM J. Numer. Anal. 17 (1980), 806-822.

[2] Bank, R.E., Rose, D.J., Global approximate Newton methods, Numer. Math. 37 (1981), 279-295.

[3] Brown, P.N., Saad, Y., Hybrid Krylov methods for nonlinear systems of equations, SIAM J. Sci. Statist. Comput. 11 (1990), 450-481.

[4] Brown, P.N., Saad, ., Convergence theory of nonlinear Newton-Krylov algorithms, SIAM J.Optim. 4 (1994), 297-330.

[5] Broyden, G., A class of methods for solving nonlinear simultaneous equations, Math. Comp. 19 (1965), 577-593.

[6] Bus, J.C.P., Numerical Solution of Systems of Nonlinear Equations, Tract 122, Mathematisch Centrum, Amsterdam, 1980.

[7] Dembo, R.S., Eisenstat, S.C., Steihaug, T., Inexact Newton methods, SIAM J. Numer. Anal., 19 (1982), 400-408.

[8] Dembo, R.S., Steihaug, T., Truncated Newton algorithms for large-scale optimization, Math. Programming 26 (1983), 190-212.

[9] P. Deuflhard, P., Heindl, G., Affine invariant convergence theorems for Newton's method and extensions to related methods, SIAM J. Numer. Anal. 16 (1979), 1-10.

[10] Eisenstat, S.C., Walker, H.F., Globally convergent inexact Newton methods, SIAM J. Optim., 4 (1994), 393-422.

[11] Eisenstat, S.C., Walker, H.F., Choosing the forcing terms in an inexact Newton method, SIAM J. Sci. Comput., 17 (1996), 16-32.

[12] Gomes-Ruggiero, M.A., Martínez, J.M., Moretti, A.C., Comparing algorithms for solving sparse nonlinear system of equations, SIAM J. Sci. Comput. 13 (1992), 459-483.

[13] Gomes-Ruggiro, M.A., Kozakevich, D.N., Martínez, J.M., A numerical study on large-scale nonlinear solvers, Department of Applied Mathematics, IMECC-UNICAMP, 1995.

[14] Lanzkron, P.J., Rose, D.J., Wilkes, J.T., An analysis of approximate nonlinear elimination, SIAM J. Sci. Comput., 17 (1996), 538-559.

[15] Lužanin, Z., Krejić, N., Herceg, D., Parameter selection for inexact Newton method, The Second World Congress of Nonlinear Analysts, Athens, 1996.

[16] Martínez, J.M., Local convergence theory of inexact Newton methods based on structured least change updates, Math. Comput. 55 (1990), 143-168.

[17] Martínez, J.M., Zambaldi, M.C., An inverse column-updating method for solving large-scale nonlinear systems of equations, Optim. Methods Software, 1 (1991), 129-140.

[18] Martínez, J.M., A theory of secant preconditioners, Math. Comput. 60 (1993), 681-698.

[19] Martínez, J.M., An extension of the theory of secant preconditioners, J. Comput. and Appl. Math. 60 (1995), 115-125

[20] Ortega, J.M., Rheinboldt, W.C., Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, New York, 1970.

[21] Ypma, T.J., Local convergence of inexact Newton methods, SIAM J. Numer. Anal. 21 (1984), 583-590.

Received by the editors September 5, 1998.