Numer. Math. 27, 271—281 (1977)

Convergence of Newton-Like Methods for Solving Systems of Nonlinear Equations

J. C. P. Bus

Received February 25, 1976

Summary. An analysis is given of the convergence of Newton-like methods for solving systems of nonlinear equations. Special attention is paid to the computational aspects of this problem.

1. Introduction

A well-known iterative method for solving a system of equations

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

where $F\colon \bar{D}\subset\mathbb{R}^n\to\mathbb{R}^n$ is some continuous function on some region $\bar{D}$, is Newton's method, which can be defined by

$$x_{k+1}=\phi(x_k)=x_k-[J(x_k)]^{-1}F(x_k), \tag{1.2}$$

where $J(x)$ denotes the jacobian matrix of partial derivatives of F. However, usually $J(x)$ is not available, so that in practice an approximation to $J(x)$ is used. In fact, calculating on a computer with finite wordlength, $J(x)$ cannot be obtained exactly. We will therefore consider a class of methods, including Newton's method which we will call *Newton-like methods*. These methods are defined by

$$x_{k+1}=\psi(x_k)=x_k-M_k^{-1}F(x_k), \tag{1.3}$$

where M_k is some approximation to $J(x_k)$.

We mention the following examples:

1. The approximation to $J(x)$ obtained by using forward difference formulas. Define the (i,j)-th element of a matrix $B(x,h)$ by:

$$(B(x,h))_{ij}=\begin{cases}\dfrac{1}{h_{ij}}\,[f_i(x+h_{ij}e^j)-f_i(x)], & \text{if } h_{ij}\neq 0,\\[2ex] \dfrac{\partial}{\partial x_j}f_i(x), & \text{if } h_{ij}=0,\end{cases} \tag{1.4}$$

where $h=(h_{11},h_{12},\ldots,h_{1n},h_{21},\ldots,h_{nn})^T\in\mathbb{R}^{n^2}$, $F(x)=(f_1(x),\ldots,f_n(x))^T$, and e^j denotes the j-th unit vector in $\mathbb{R}^n$. Then M_k is obtained by

$$M_k=B(x_k,h_k). \tag{1.5}$$

2. The approximation to $J(x)$ obtained by evaluating the analytic expressions for the partial derivatives on a computer with finite wordlength.

The analysis of Newton-like methods given in this paper is essentially based on the Newton-Kantorovich Theorem (Kantorovich [5]) and its extension given by Ortega and Rheinboldt [6]. In our approach the dependence of the asymptotic order of convergence on the error in M_k as an approximation to $J(x_k)$ is clearly expressed. This approach enables us to incorporate the influence of rounding errors in the computation and leads to conditions for convergence of a Newton-like method when all computation is done in finite precision. Moreover, we introduce a quantity, the solvability number, which can be used as a measure of the degree of difficulty of a problem when solving it with a Newton-like method on a computer. To calculate this number one needs information about the first as well as the second derivative of the problem function, which is not available for many practical problems. Therefore, this quantity is of limited value for solving systems of nonlinear equations. However, this notion may be extremely useful for testing programs implementing Newton-like methods. Test functions can be created such that all information required for calculating the solvability number is available. We can therefore create a representative set of test functions of various degrees of difficulty (see Bus [1]).

2. Analysis of Newton-Like Methods

Let F be given as in Section 1 and let $H(x)$ denote the tensor of partial derivatives of F at x. Suppose $x_0 \in \bar{D}$ is given and $\{x_k\}_{k=0}^{\infty}$ is defined by (1.3). Moreover, let $z \in \bar{D}$ be a solution of the system of nonlinear equations defined by F. Then the aim of this section is to derive sufficient conditions such that $\{x_i\}_{i=0}^{\infty}$ converges to z. We assume that exact arithmetic is used. To simplify the notation we omit, whenever possible, the subscripts denoting the iteration index, and we denote the current iterate by x and the new one by $\psi(x)$.

Furthermore, except for some cases in which it is stated explicitly, we do not specify the norms used in this paper. When $\|\cdot\|$ is used, the reader may think of any norm, provided it is used throughout and provided that the norm of $L(L(\mathbb{R}^n))$ is subordinate to the norm of $L(\mathbb{R}^n)$, which in turn is subordinate to the norm of $\mathbb{R}^n$. Here $L(A)$ denotes the linear space of linear operators from A to A, for some space A, and a norm $\|\cdot\|_L$ in $L(A)$ is called subordinate to some given norm $\|\ \|$ in A if it is defined, for $G \in L(A)$ and $x \in A$, by

$$\|G\|_L = \sup_{x \neq 0} \frac{\|Gx\|}{\|x\|}.$$

The following definition appears to be useful (cf. Ortega and Rheinboldt [6]).

Definition 2.1. Let F be differentiable on $D \subset \bar{D} \subset \mathbb{R}^n$ and let, for some real number $r > 0$ and integer $m > 0$, an operator M be defined by

$$M: D_0 \times U_r^m \subset \mathbb{R}^n \times \mathbb{R}^m \to L(\mathbb{R}^n), \tag{2.1}$$

where $D_0 \subset D$ and $U_r^m = \{y \in \mathbb{R}^m \mid \|x\| \leqq r\}$. Then $M(x, h)$ is called a *strongly consistent approximation* to the jacobian matrix $J(x)$ on D_0 if a constant $\bar{c}$, called the *consistency factor*, exists such that

$$x \in D_0, \qquad h \in U_r^m \Rightarrow \|J(x) - M(x, h)\| \leqq \bar{c}\|h\|. \tag{2.2}$$

An example of a strongly consistent approximation to the jacobian matrix of F is given by the forward difference approximation $B(x, h)$ defined by (1.4). The following result for $B(x, h)$ can be proved.

Theorem 2.2. *Assume that F is continuously differentiable on some open set $D \subset \bar{D}$. Then for any compact set $D_0 \subset D$ there exists a $\varrho > 0$ such that $B(x, h)$, given by (1.4), is well defined for $h \in U_\varrho^{n^2}$ and $x \in D_0$. Moreover, if*

$$\|J(x) - J(y)\| \leqq \gamma \|x - y\|, \quad \text{for all } x, y \in D \tag{2.3}$$

and some constant $\gamma > 0$, then $B(x,h)$ is a strongly consistent approximation to $J(x)$ on D_0.

Proof. See Ortega and Rheinboldt [6], Section 11.2.5. □

We are now ready to define precisely the class of methods that we are going to analyze.

Definition 2.3. We call a method as given by (1.3), for solving (1.1), *a proper Newton-like method* if there exist an operator M, as given by (2.1), for some integer m and some real r, and $h_k \in U_r^m (k = 0, 1, 2, \ldots)$, such that

$$M_k = M(x_k, h_k), \quad k = 0, 1, \ldots, \tag{2.4}$$

and $M(x, h)$ is a strongly consistent approximation to $J(x)$ on D_0. We will denote such a method by $N(M)$.

To study the convergence behavior of proper Newton-like methods we compare them with Newton's method. Define, in a manner similar to (1.2) and (1.3),

$$\phi(x) = x - [J(x)]^{-1} F(x) \tag{2.5}$$

and

$$\psi(x) = x - [M(x, h)]^{-1} F(x), \tag{2.6}$$

where $\psi(x)$ defines a proper Newton-like method. From now on we assume that $D \subset \bar{D}$ is convex, $z, x_0 \in D$, F is twice differentiable, and $J(x)$ is nonsingular on D. Then, using the mean value theorem we obtain the following expression for the error in $\phi(x)$ as an approximation to the solution vector z:

$$\|\phi(x) - z\| = \|[J(x)]^{-1}(J(x)\,(x - z) - F(x))\| \leqq S(x,z)\,\|x - z\|^2, \tag{2.7}$$

where

$$S(x,z) = \tfrac{1}{2}\Big(\sup_{y \in L[z,\,x]} \|H(y)\|\Big)\,\|[J(x)]^{-1}\| \tag{2.8}$$

and

$$L[z,x] = \{u \in \mathbb{R}^n \,|\, u = \theta x + (1 - \theta) z, \quad 0 \leqq \theta \leqq 1\}. \tag{2.9}$$

(2.7) expresses the well-known result that the asymptotic order of convergence of Newton's method is quadratic. Using the well-known perturbation lemma (e.g. Rall [7], Section 10) we obtain for the difference between $\psi(x)$ and $\phi(x)$:

$$\begin{aligned}\phi(x) - \psi(x) &= ([J(x)]^{-1} - [M(x,h)]^{-1})\,F(x)\\ &= \left[I - \sum_{n=0}^{\infty} (I - [J(x)]^{-1} M(x,h))^n\right] [J(x)]^{-1} F(x).\end{aligned}$$

Hence, if

$$\|h\| < 1/(2\bar{c}\,\|[J(x)]^{-1}\|),$$

where $\bar{c}$ is the consistency factor of M, then

$$\|\phi(x)-\psi(x)\| \leqq \bar{C}(x,h)\|[J(x)]^{-1}F(x)\|, \tag{2.10}$$

where

$$\bar{C}(x,h)=2\bar{c}\|h\|\,\|[J(x)]^{-1}\|. \tag{2.11}$$

Furthermore,

$$\|[J(x)]^{-1}F(x)\|=\|x-\phi(x)\| \leqq \|x-z\|+\|\phi(x)-z\|. \tag{2.12}$$

So, combining (2.7), (2.10) and (2.12) we obtain the following upper bound for the error in $\psi(x)$ as an approximation to z:

$$\begin{aligned}\|\psi(x)-z\| &\leqq \|\psi(x)-\phi(x)\|+\|\phi(x)-z\| \\ &\leqq \bar{C}(x,h)\|x-z\|+(1+\bar{C}(x,h))\,S(x,z)\|x-z\|^2.\end{aligned} \tag{2.13}$$

Since $\bar{C}(x,h)=0(\|h\|)$, we can only expect that the asymptotic order of convergence of a proper Newton-like method is quadratic if $\|h\|=0(\|x-z\|)$.

The above results are summarized in the following definition.

Definition 2.4. Let a nonlinear system be defined by (1.1) and let $x_0 \in D$ be an approximation to the solution z of (1.1). Then we say that this problem is *solvable* by a proper Newton-like method $N(M)$ if the following conditions are satisfied:

a) $J(x)$ and $H(x)$ exist on D and $J(x_0)$ is nonsingular.

b) If $C(x,h)$ is defined by (2.11), then h_0 satisfies

$$\bar{C}(x_0,h_0)<1 \tag{2.14}$$

and

$$r_0=\bar{C}(x_0,h_0)\|\phi(x_0)-x_0\|+\|\phi(x_0)-z\|<\|x_0-z\|. \tag{2.15}$$

c) $U_0=\{y\in\mathbb{R}^n \mid \|y-z\| \leqq r_0\} \subset D$ and $J(x)$ is nonsingular on U_0.

d) If $\bar{K}$ is defined by

$$\bar{K}=\sup_{\substack{x\in U_0 \\ k=1,2,\ldots}} \bar{C}(x,h_k),$$

then h_k satisfies $\bar{K}<1$.

e)

$$\bar{\sigma}\,(F,z,x_0,M)=\bar{K}+(\bar{K}+1)\,Sr_0<1, \tag{2.16}$$

where

$$S=\sup_{x\in U_0} S(x,z).$$

If a) to d) are satisfied, then $\bar{\sigma}(F,z,x_0,M)$ is called the *solvability number* of the Newton-like method $N(M)$ for solving the nonlinear system $F(x)=0$ with x_0 as initial guess and z as solution. If a) to d) are not all satisfied, then the solvability number is defined to be infinite.

The following theorem is now easily proved.

Theorem 2.5. *If a nonlinear system defined by* (1.1) *with initial approximation* x_0 *and solution* z *is solvable by a proper Newton-like method, then the sequence of points*

generated by this method converges to z. If, moreover, the method is such that $\|h_k\| = O(\|x_k - z\|)$ *for* $k \to \infty$, *then the asymptotic order of convergence is quadratic.*

Proof. Since (2.14) is satisfied we obtain from (2.10)

$$\begin{aligned}\|\psi(x_0) - z\| &\leqq \|\psi(x_0) - \phi(x_0)\| + \|\phi(x_0) - z\| \\ &\leqq \bar{C}(x_0, h_0)\|\phi(x_0) - x_0\| + \|\phi(x_0) - z\|.\end{aligned}$$

Because of (2.15) we know that

$$\|\psi(x_0) - z\| < \|x_0 - z\|.$$

Because of c), d) and e) we can use (2.13), so that with condition (2.16) the result follows immediately. □

Although in practice condition e) is a rather strong condition, it gives us a clear insight into the behavior of a certain Newton-like method, provided one can derive results about the consistency factor of the method. In fact $\bar{\sigma}$ gives us a possibility of measuring the degree of difficulty for solving the problem with the method. Furthermore, condition d) shows that the larger $\sup_{x \in U_0} \|[J(x)]^{-1}\|$ is, the smaller h_k should be chosen. Note that conditions c), d) and e), with U_0 replaced by

$$\bar{U} = \{y \in \mathbb{R}^n \mid \|y - z\| \leqq \|x_0 - z\|\} \subset D,$$

together with the condition $J(x)$ and $H(x)$ exist on $\bar{U}$, are sufficient to prove convergence. However, the relevance of Definition 2.4 lies in its use for calculating the solvability number of problems used for testing programs, and using condition b) we may considerably reduce this solvability number. We expect the given definition to be more realistic, which is illustrated by the examples given in Section 4.

3. The Effect of Rounding Errors

In this section we consider the effect of round-off errors on the convergence behavior of Newton-like methods. We use the following notation:

ε: the precision of computation used;

$fl_\varepsilon(\cdot)$: the expression inside the parentheses calculated with the precision of computation ε.

If we wish to apply the theory given in Section 2 to a Newton-like method where all computation is done in finite precision (such a method is called a *numerical Newton-like method* in this section) we are immediately confronted with the problem that a numerical Newton-like method will, in general, not be a proper Newton-like method. Even when we choose

$$M_k = fl_\varepsilon(J(x_k)),$$

which is the best we can do anyhow, we can, in general, only guarantee that

$$\|M_k - J(x_k)\| \leqq \delta \|J(x_k)\|, \tag{3.1}$$

where $\delta \geqq \varepsilon$ is some value depending on ε and the way in which M_k is calculated. Therefore, the notion "strongly consistent approximation" (cf. Def. 2.1) is not

a useful concept when dealing with numerical Newton-like methods. We give an extension of the theory given in Section 2, which is applicable to numerical Newton-like methods. First we introduce a more general concept for measuring the consistency of M_k as an approximation to $J(x_k)$.

Definition 3.1. (see Def. 2.1). Let F be differentiable on $D \subset \bar{D} \subset \mathbb{R}^n$ and let the operator M be defined by (2.1) for some real number $r > 0$ and integral number $m \geqq 0$. Then $M(x, h)$ is called a *numerically consistent approximation* to $J(x)$ on $D_0 \subset D$ if there exist a constant c_1 and a function $c_0(\varepsilon, h)$ which is continuous in ε for fixed $h \neq 0$ and $\varepsilon \geqq 0$, such that the following conditions are satisfied:

$$\|J(x) - fl_\varepsilon(M(x,h))\| \leqq c_0(\varepsilon,h) + c_1\|h\|, \quad \text{for all } x \in D_0, \quad h \in U_r^m \setminus \{0\}, \tag{3.2}$$

$$\lim_{\varepsilon \to 0} c_0(\varepsilon, h) = 0, \quad \text{for } h \neq 0. \tag{3.3}$$

We call

$$c(\varepsilon, h) = c_0(\varepsilon, h) + c_1\|h\| \tag{3.4}$$

the *consistency function* of M on D_0.

As an example of a numerically consistent approximation we again consider the forward difference approximation $B(x,h)$, defined by (1.4). We prove the following theorem.

Theorem 3.2. *Assume that F is continuously differentiable on some open set $D \subset \bar{D}$. Then for any compact set $D_0 \subset D$ there exists a $\varrho > 0$ such that $B(x,h)$, given by (1.4), is well defined for $h \in U_\varrho^{n^2}$ and $x \in D_0$. Moreover, if (2.3) is satisfied, then $B(x,h)$ is a numerically consistent approximation to $J(x)$ on D_0.*

Proof. We use the following relations (Wilkinson [8]; Dekker [3]):

$$|fl_\varepsilon(a \pm b) - (a \pm b)| \leqq (|a| + |b|)\varepsilon, \qquad |fl_\varepsilon(a/b) - (a/b)| \leqq |a/b|\,\varepsilon.$$

We assume that for some $\delta = \delta(\varepsilon) \geqq \varepsilon$

$$|fl_\varepsilon(f_i(x)) - f_i(x)| \leqq |f_i(x)|\,\delta, \quad \forall x \in D, \quad i = 1, 2, \ldots, n,$$

where $F(x) = (f_1(x), \ldots, f_n(x))^T$. Now suppose $h_{ij} \neq 0$. Then some simple algebra shows that the error in the forward difference approximation to an element of the jacobian matrix can be bounded by

$$\left| fl_\varepsilon((B(x,h))_{ij}) - \frac{\partial f_i}{\partial x_j} \right|$$

$$\leqq \left| (B(x,h))_{ij} - \frac{\partial f_i}{\partial x_j} \right| + \varepsilon |(B(x,h))_{ij}| + \frac{\delta + 2\varepsilon}{|h_{ij}|} (|f_i(x)| + |f_i(x + h_{ij}e^j)|),$$

where we assumed that $\delta < \frac{1}{4}$, which seems reasonable. Hence, using the l_1-norm, we obtain

$$\|fl_\varepsilon(B(x,h)) - J(x)\| \leqq (1+\varepsilon)\|B(x,h) - J(x)\| + \varepsilon\|J(x)\| + \frac{3(n+1)\delta}{h_{\min}} \sup_{\|y-x\| \leqq \|h\|} (\|F(y)\|),$$

where $h_{\min} = \min(|h_{ij}|,\ i, j = 1, \ldots, n, |h_{ij}| \neq 0)$.

From Theorem 2.2 and the fact that D_0 is compact and D open we know that there exist a $\varrho > 0$ and a $\bar{c}_1$ such that

$$\|B(x,h) - J(x)\| \leqq \bar{c}_1 \|h\|, \quad \text{for } h \in U_\varrho^{n^2}.$$

Choose

$$c_0(\varepsilon, h) = \frac{3(n+1)\delta}{h_{\min}} \left(\sup_{x \in D} (\|F(x)\|) \right) + \varepsilon \sup_{x \in D_0} (\|J(x)\|), \tag{3.5}$$

$$c_1 = (1+\varepsilon)\bar{c}_1.$$

Then the theorem is proved, since

$$\lim_{\varepsilon \to 0} |\delta(\varepsilon)| = 0. \quad \square$$

We are now ready to define whether we may expect a numerical Newton-like method to behave like Newton's method.

Definition 3.3. (see Def. 2.3). We call a numerical Newton-like method for solving (1.1) a *proper numerical Newton-like method* if there exist an operator M as given by (2.1), for some integer m and real r, and $h_k \in U_r^m \setminus \{0\}$ $(k = 0, 1, 2, \ldots)$, such that

$$M_k = fl_\varepsilon(M(x_k, h_k))$$

and $M(x,h)$ is a numerically consistent approximation to $J(x)$ on D_0. Such a method is denoted by $N(M, \varepsilon)$.

We give an analysis of proper numerical Newton-like methods which is analogous to the analysis of a proper Newton-like method. Denote by $\bar{\psi}(x)$ the vector which exactly satisfies the equation

$$fl_\varepsilon(M(x,h))\ (\bar{\psi}(x) - x) = F(x). \tag{3.6}$$

With the same assumptions as in Section 2, we obtain (cf. (2.10)):

$$\|\phi(x) - \bar{\psi}(x)\| \leqq C(x,h,\varepsilon) \|[J(x)]^{-1} F(x)\|, \tag{3.7}$$

where it is assumed that

$$C(x,h,\varepsilon) = 2c(\varepsilon,h) \|[J(x)]^{-1}\| < 1 \tag{3.8}$$

and $c(\varepsilon,h)$ is given by (3.4).

Let $fl_\varepsilon(\bar{\psi}(x))$ be the numerical approximation to $\bar{\psi}(x)$. Then

$$fl_\varepsilon(\bar{\psi}(x)) = fl_\varepsilon(fl_\varepsilon(\bar{\psi}(x) - x) + x), \tag{3.9}$$

where $fl_\varepsilon(\bar{\psi}(x) - x)$ denotes the numerical solution of the system (3.6), where $F(x)$ is replaced by $fl_\varepsilon(F(x))$.

Now, suppose we want to solve with gaussian elimination, on a computer with precision of arithmetic ε, the linear system

$$Ax = b,$$

where A is given exactly but b is not. Let the error in b be bounded by $\|\delta b\|$.

Then the error in the numerical solution $\bar{x}$ as an approximation to the exact solution x^* is bounded by

$$\frac{\|\bar{x}-x^*\|}{\|x^*\|} \leqq \varkappa(A)\left[\frac{\varepsilon g(n)}{1-\varkappa(A)\,\varepsilon g(n)}+\frac{\|\delta b\|}{\|b\|}\right], \tag{3.10}$$

where $\varkappa(A)=\|A\|\,\|A^{-1}\|$ and $g(n)$ is some function depending on the order n, the norm used and the pivoting strategy used (Wilkinson [8]), and where it is assumed that

$$\varkappa(A)\,\varepsilon g(n)<1.$$

Using the perturbation lemma it is easily shown that $\varkappa(fl_\varepsilon(M(x,h))) \leqq 3\varkappa(J(x))$. Applying these results to $fl_\varepsilon(\bar{\psi}(x)-x)$ we obtain

$$\|fl_\varepsilon(\bar{\psi}(x)-x)-(\bar{\psi}(x)-x)\| \leqq \alpha(x,\varepsilon,n)\,\|\bar{\psi}(x)-x\|, \tag{3.11}$$

where

$$\alpha(x,\varepsilon,n)=3\varkappa(J(x))\left[\frac{\varepsilon g(n)}{1-3\varkappa(J(x))\,\varepsilon g(n)}+\delta\right] \tag{3.12}$$

and δ satisfies

$$\|fl_\varepsilon(F(x))-F(x)\| \leqq \delta\|F(x)\|. \tag{3.13}$$

We assumed that $3\varkappa(J(x))\,\varepsilon g(n)<1$. Combining (3.9) and (3.11) we obtain

$$\begin{aligned}\|fl_\varepsilon(\bar{\psi}(x))-\bar{\psi}(x)\| &\leqq \varepsilon(\|fl_\varepsilon(\bar{\psi}(x)-x)\|+\|x\|)+\|fl_\varepsilon(\bar{\psi}(x)-x)-(\bar{\psi}(x)-x)\| \\ &\leqq \varepsilon\|x\|+\beta(x,\varepsilon,n)\,\|\bar{\psi}(x)-x\|,\end{aligned} \tag{3.14}$$

where

$$\beta(x,\varepsilon,n)=(1+\varepsilon)\,\alpha(x,\varepsilon,n)+\varepsilon. \tag{3.15}$$

Finally, combining (2.7), (3.7) and (3.14) we obtain for the error in $fl_\varepsilon(\bar{\psi}(x))$ as an approximation to solution z:

$$\begin{aligned}\|fl_\varepsilon(\bar{\psi}(x))-z\| &\leqq \|fl_\varepsilon(\bar{\psi}(x))-\bar{\psi}(x)\|+\|\bar{\psi}(x)-z\| \\ &\leqq \varepsilon\|x\|+\beta(x,\varepsilon,n)\,\|x-z\|+(1+\beta(x,\varepsilon,n))\,\|\bar{\psi}(x)-z\|.\end{aligned} \tag{3.16}$$

With

$$\|\bar{\psi}(x)-z\| \leqq \|\bar{\psi}(x)-\phi(x)\|+\|\phi(x)-z\|$$

we obtain as the final result:

$$\|fl_\varepsilon(\bar{\psi}(x))-z\| \leqq \varepsilon\|x\|+L(x,\varepsilon,h,n)\,\|x-z\|+Q(x,\varepsilon,h,n,z)\,\|x-z\|^2, \tag{3.17}$$

where

$$L(x,\varepsilon,h,n) = \beta(x,\varepsilon,n)+(1+\beta(x,\varepsilon,n))\,C(x,h,\varepsilon) \tag{3.18}$$

and

$$Q(x,\varepsilon,h,n,z)=(1+\beta(x,\varepsilon,n))\,(1+C(x,h,\varepsilon))\,S(x,z). \tag{3.19}$$

From the first term in the right-hand side of (3.17) we see that one cannot expect to find a solution of a nonlinear system with a proper numerical Newton-like method within a relative precision which is higher than the precision of computation. Furthermore, whether there is convergence at all depends on the quantities:

$S(x,z)$, the convergence factor of the exact Newton method;

$C(x,h,\varepsilon)$, which is a measure for the error in $fl_\varepsilon(M(x,h))$ as a numerical approximation to $J(x)$; this quantity depends on the method;

$\beta(x,\varepsilon,n)$, which reflects the condition number of the linear subproblem; the condition number $\varkappa(J(x))$ should be small relative to $1/\varepsilon$.

In either case, $L(x,\varepsilon,h,n)+Q(x,\varepsilon,h,n,z)\|x-z\|$ has to be less than 1 in order to be able to guarantee convergence.

We summarize these results in the following definition:

Definition 3.4 (see Def. 2.4). Let a nonlinear system be defined by (1.1) and let $x_0 \in D$ be an approximation to the solution z of (1.1). Then we call this problem *solvable* by a proper numerical Newton-like method $N(M,\varepsilon)$ if the following conditions are satisfied:

a) $J(x)$ and $H(x)$ exist on D and

$$\varkappa(J(x_0)) < 1/(3\,\varepsilon\, g(n)),$$

where $g(n)$ depends on the method used for solving the linear system (cf. (3.10)).

b) h_0 satisfies $C(x_0,h_0,\varepsilon)<1$, and if

$$r_0 = \varepsilon\|x_0\| + \|\phi(x_0)-z\| + [\beta(x_0,\varepsilon,n) + (1+\beta(x_0,\varepsilon,n))\, C(x_0,h_0,\varepsilon)]\,\|\phi(x_0)-x_0\|,$$

then

$$r_0 < \|x_0 - z\|.$$

c) $$U_0 = \{y \in \mathbb{R}^n \mid \|y-z\| \leqq r_0\} \subset D$$

and

$$\sup_{x\in U_0} \varkappa(J(x)) < 1/(3\,\varepsilon g(n)).$$

d) If K is defined by

$$K = \sup_{\substack{x\in U_0 \\ k=1,2,\ldots}} C(x,h_k,\varepsilon), \qquad \text{then } h_k \text{ satisfies } K<1.$$

e) $$\sigma(F,z,x_0,M,\varepsilon) = \beta + (1+\beta)K + (1+\beta)\,(1+K)\,S r_0 < 1,$$

where

$$S = \sup_{x\in U_0} S(x,z), \qquad \beta = \sup_{x\in U_0} \beta(x,\varepsilon,n).$$

If a) to d) are satisfied, then $\sigma(F, z, x_0, M, \varepsilon)$ is called the *solvability number* of the proper numerical Newton-like method $N(M, \varepsilon)$ for solving the nonlinear system $F(x)=0$ with x_0 as initial guess and z as solution. If a), b), c) or d) are not satisfied, then the solvability number is defined to be infinite.

The following theorem is now easily proved.

Theorem 3.5 (see Theorem 2.5). *If a system of nonlinear equations defined by* (1.1) *with initial approximation* x_0 *and solution* z *is solvable by a proper numerical Newton-like method* $N(M, \varepsilon)$, *then the sequence of points generated by this method converges to a point* x^* *with* $\|x^*-z\| \leqq \varepsilon\|x^*\|$.

Proof. The proof is similar to the proof of Theorem 2.5. □

4. Some Examples

Consider the problem given by Gheri and Mancino [4]:

$$f_i(x)=\beta\, n\, x_i+(i-n/2)^\gamma+\sum_{\substack{j=1\\ j\neq i}}^{n}[z_{ij}(\sin^\alpha(\log(z_{ij}))+\cos^\alpha(\log(z_{ij})))], \qquad (4.1)$$

where

$$F(x)=(f_1(x),\ldots,f_n(x))^T$$

and

$$z_{ij}=\sqrt{x_j^2+i/j}.$$

Elementary computation leads to the following inequalities for the elements $J_{ij}(x)$ of the jacobian matrix and $H_{ijk}(x)$ of the hessian tensor:

$$\begin{aligned} J_{ii}(x)&=\beta n, \quad |J_{ij}(x)|\leqq\alpha+1 && \text{for } i\neq j;\\ H_{ijk}(x)&=0 \quad \text{if} && i=j \quad \text{or} \quad j\neq k,\\ |H_{ijk}(x)|&\leqq(\alpha+1)^2 && i\neq j \quad \text{and} \quad j=k. \end{aligned}$$

Use of Gershgorin's Theorem for bounding the eigenvalues of a matrix leads to

$$\begin{aligned} \|J(x)\|&\leqq[\beta^2n^2+(2\beta n+(n-1)\,(\alpha+1))\,(\alpha+1)\,(n-1)]^{\frac{1}{2}},\\ \|[J(x)]^{-1}\|&\leqq[\beta^2n^2-(2\beta n+(n-2)\,(\alpha+1))\,(\alpha+1)\,(n-1)]^{-\frac{1}{2}}. \end{aligned} \qquad (4.2)$$

Furthermore

$$\|H(x)\|\leqq\sqrt{n-1}\,(\alpha+1)^2. \qquad (4.3)$$

Now let methods A and B be Newton-like methods with

$$M_k^A=fl_\varepsilon(J(x_k)), \qquad M_k^B=fl_\varepsilon(B(x_k, 0.0001)),$$

where $B(x, 0.0001)$ is defined by (1.4) with $h_{ij}=h=0.0001$ $(i, j=1, \ldots, n)$. The precision of arithmetic is chosen to be

$$\varepsilon=10^{-14},$$

and we assume that in both methods gaussian elimination with complete pivoting is used (see Wilkinson [8]), so that

$$g(n)\approx 20n^3.$$

We use (3.1) and (3.5) to obtain the consistency functions $c^A(\varepsilon, h)$ and $c^B(\varepsilon, h)$ of methods A and B respectively, where δ in (3.1) is assumed to be 10^{-11}, which is very reasonable.

We consider the problem for which

$$\alpha=5, \quad \beta=14, \quad \gamma=3, \quad n=10. \qquad (4.4)$$

Our goal is not to solve this problem (in fact we assume that it is already solved), but we want to know whether it is solvable by the given methods A and B, and what the values of the solvability numbers are.

Let z denote the solution and suppose x_0 is chosen such that

$$x_0^{(i)}=z^{(i)}+1, \qquad (4.5)$$

where $x^{(i)}$ denotes the i-th component of the vector x. Then computation of $\|x_0-z\|$, $\|\phi(x_0)-z\|$ and $\|\phi(x_0)-x_0\|$ with a computer delivers approximately:

$$\|x_0-z\|=3.2, \quad \|\phi(x_0)-z\|=0.023, \quad \|\phi(x_0)-x_0\|=3.1.$$

Using the above results it is easy to verify the conditions of Definition 3.4. We obtain

$$r_0^A \leqq 0.023, \quad r_0^B \leqq 0.025,$$

and finally for the solvability numbers

$$\sigma(F,z,x_0,M^A,\varepsilon) \leqq 1.24 \times r_0 \leqq 0.029,$$
$$\sigma(F,z,x_0,M^B,\varepsilon) \leqq 1.24 \times r_0 < 0.031.$$

Therefore, we may conclude that Problem 4.1 with α, β, γ and n given by (4.4) and the starting point given by (4.5) is an excellent test problem which should be solved easily by each program implementing algorithm $N(M^A, 10^{-14})$ or $N(M^B, 10^{-14})$.

Note that if we had simplified Definition 3.4 such that U_0 was chosen to be $\{y \in \mathbb{R}^n \mid \|y-z\| \leqq \|x_0-z\|\}$ and condition b) deleted, then the solvability numbers would have been 4.0 approximately and convergence would not have been proved. Hence, the given Definition is preferable (see also the end of Section 2).

References

1. Bus, J. C. P.: A comparative study of programs for solving nonlinear equations. Report NW 25/75, Mathematisch Centrum, Amsterdam (1975)
2. Collatz, L.: Funktionalanalysis und numerische Mathematik. Berlin-Göttingen-Heidelberg-New York: Springer 1964. English edition. New York: Academic Press 1966
3. Dekker, T. J.: Numerieke Algebra. MC Syllabus 12, Mathematisch Centrum, Amsterdam (1971)
4. Gheri, G., Mancino, O. G.: A significant example to test methods for solving systems of nonlinear equations. Calcolo **8**, 107–113 (1971)
5. Kantorovich, L.: On Newton's method for functional equations [Russian]. Dokl. Akad. Nauk. SSSR **59**, 1237–1240 (1948)
6. Ortega, J. M., Rheinboldt, W. C.: Iterative solution of nonlinear equations in several variables. New York: Academic Press 1970
7. Rall, L. B.: Computational solution of nonlinear operator equations. New York: Wiley 1969
8. Wilkinson, J. H.: Rounding errors in algebraic processes, Notes on Applied Science no. 32. Englewood Cliffs, N. J.: Prentice Hall 1963
9. Wilkinson, J. H.: The algebraic eigenvalue problem. Oxford: Clarendon Press 1965

J. C. P. Bus
Mathematical Centre
Tweede Boerhaavestraat 49
Amsterdam, The Netherlands