stichting
mathematisch
centrum

MC

AFDELING NUMERIEKE WISKUNDE NW 20/75 JUNE

J.C.P. BUS

AN ANALYSIS OF THE CONVERGENCE OF NEWTON-LIKE METHODS FOR SOLVING SYSTEMS OF NONLINEAR EQUATIONS

2e boerhaavestraat 49 amsterdam

Printed at the Mathematical Centre, 49, 2e Boerhaavestraat, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O), by the Municipality of Amsterdam, by the University of Amsterdam, by the Free University at Amsterdam, and by industries.

AMS(MOS) subject classification scheme (1970): 65H10

An analysis of the convergence of Newton-like methods for solving systems of nonlinear equations

by

J.C.P. Bus.

ABSTRACT

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.

KEYWORDS & PHRASES: *Systems of nonlinear equations, Newton-like methods, analysis of convergence, computational aspects.*

CONTENTS

1. INTRODUCTION

In this report we will be concerned with iterative methods for solving a system of nonlinear equations. Let

$$F: D \subset \mathbb{R}^n \to \mathbb{R}^n \tag{1.1}$$

be some function defined on some region D. Let $x_0 \in D$. Then we want to construct a sequence of points $\{x_i\}_{i=0}^{\infty}$, with $x_i \in D$ $(i = 1,2,\ldots)$ such that

$$z = \lim_{i \to \infty} x_i \tag{1.2}$$

exists and

$$F(z) = 0. \tag{1.3}$$

A very well-known method for solving this problem is Newton's method, defined by

$$x_{k+1} = \phi(x_k) = x_k - [J(x_k)]^{-1} F(x_k) , \tag{1.4}$$

where J(x) denotes the jacobian matrix of partial derivatives of F. For this method, KANTOROVICH [5] presented convergence results, known in literature as the Newton-Kantorovich theorem (see for instance ORTEGA & RHEINBOLDT [8]). However, most frequently $J(x_k)$ is not available, so that in practice an approximation to $J(x_k)$ is used. In fact, calculating on a computer with finite wordlength, $J(x_k)$ can not be obtained exactly. In order to derive results about the convergence of the very useful modifications of Newton's method, we study methods defined by

$$x_{k+1} = \psi(x_k) = x_k - M_k^{-1} F(x_k) , \tag{1.5}$$

where M_k is some approximation to $J(x_k)$. We call such a method a *Newton-like method*. 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:

$$(1.6)\qquad (B(x,h))_{ij} = \begin{cases} \frac{1}{h_{ij}}[f_i(x+h_{ij}e^j)-f_i(x)], & \text{if } h_{ij} \neq 0, \\ \frac{\partial}{\partial x_j} f_i(x), & \text{if } h_{ij} = 0, \end{cases}$$

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

$$(1.7)\qquad M_k = B(x_k,h_k).$$

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

3. $$M_k = J(x_k) + \lambda_k I,$$

where $\lambda_k \in \mathbb{R}$ and I denotes the unit-matrix. The Newton-like method obtained in this way has been proposed by LEVENBERG [6] and MARQUARDT [7].

The analysis of Newton-like methods, given in this report is essentially based on the Newton-Kantorovich theorem and its extension given by ORTEGA & RHEINBOLDT [8]. However, we use a somewhat different approach, in order to be able to deal with difficulties that arise when finite precision arithmetic is used.

Considering (1.5) we see that it is obtained by approximating the function F(x) in a neighbourhood of x_k by

$$(1.8)\qquad F_k(x) = F(x_k) + M_k(x-x_k)$$

and by solving the linear system which arises by setting $F_k(x) = 0$. Clearly, two sources of errors arise in approximating a solution z, with F(z) = 0, by $\psi(x_k)$, when a computer is used.

1. The error caused by approximating $F(x)$ by $F_k(x)$.
2. The error caused by the numerical solution of the linear system

$$M_k(x-x_k) = -F(x_k).$$

In section 2 we will discuss the properties of a Newton-like method when exact arithmetic is used, so that the second source of errors does not occur. In section 3 we will discuss the influence of rounding errors in the computation on the results given in section 2.

2. ANALYSIS OF NEWTON-LIKE METHODS

Let a function F and some region D be given by (1.1). Let $J(x)$ denote the jacobian matrix of partial derivatives of F at x and let $H(x)$ denote the tensor of partial second derivatives of F at x. Suppose $x_0 \in D$ is given and a sequence $\{x_i\}_{i=0}^{\infty}$ is constructed by a Newton-like method as given by (1.5). Let, moreover, $z \in D$ be a solution of the system of nonlinear equations defined by F, i.e. z satisfies (1.3). 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 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 where it is stated explicitly, we do not specify the norms used in this report. 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 with the norm of $L(\mathbb{R}^n)$, which in turn is subordinate with 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 with 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 lemma will be extremely useful for obtaining the desired results.

LEMMA 2.1. (Perturbation lemma)
Suppose $A \in L(\mathbb{R}^n)$. *Then* A^{-1} *exists if and only if there is a* $B \in L(\mathbb{R}^n)$ *such that* B^{-1} *exists and*

$$\|B-A\| \leq 1/\|B^{-1}\|.$$

Moreover, if A^{-1} *exists, then*

$$A^{-1} = \sum_{n=0}^{\infty} (I-B^{-1}A)^n B^{-1}, \tag{2.1}$$

$$\|A^{-1}\| \leq \frac{\|B^{-1}\|}{1 - \|I-B^{-1}A\|} \leq \frac{\|B^{-1}\|}{1 - \|B^{-1}\|\|B-A\|}, \tag{2.2}$$

where I *denotes the unit-matrix.*

PROOF. See RALL [9] Section 10. □

Our analysis of Newton-like methods is based on the analysis of Newton's method as given by KANTOROVICH [5]. See also ORTEGA & RHEINBOLDT [8], COLLATZ [2] and RALL [9]. It appears to be useful for our analysis to define a concept which expresses the relation between the jacobian matrix J(x) and its approximation M. The following definition appears to be useful.

DEFINITION 2.2. Let F be differentiable on $D \subset \mathbb{R}^n$ and let for some real number $r > 0$ and integer $m \geq 0$ an operator M be defined by

$$M: D_0 \times U_r^m \subset \mathbb{R}^n \times \mathbb{R}^m \to L(\mathbb{R}^n), \tag{2.3}$$

where $D_0 \subset D$ and $U_r^m = \{y \in \mathbb{R}^m \mid \|y\| \leq r\}$. 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

$$x \in D_0,\ h \in U_r^m \Rightarrow \|J(x)-M(x,h)\| \leq \bar{c}\|h\|. \tag{2.4}$$

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 (1.6). The following result for B(x,h) can be proved.

THEOREM 2.3. *Assume that* F *is continuously differentiable on* D. *Then, for any compact set* $D_0 \subset D$, *there exists a* $\rho > 0$, *such that* B(x,h), *given by* (1.6), *is well-defined for* $h \in U_\rho^m$ *and* $x \in D_0$. *Moreover, if*

$$\|J(x)-J(y)\| \leq \gamma\|x-y\|, \qquad \text{for all } x,y \in D_0 \tag{2.5}$$

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

PROOF. See ORTEGA & RHEINBOLDT [8], section 11.2.5. □

The following corollary is easily derived from definition 2.2 and lemma 2.1.

COROLLARY 2.4. *Let* F *be differentiable and* J(x) *nonsingular on* D, *with*

$$\sup_{x\in D} \|[J(x)]^{-1}\| \leq \alpha. \tag{2.6}$$

Let for some integer m *and real* r, *the operator* M *be defined by* (2.3) *and let* M(x,h) *be a strongly consistent approximation to* J(x) *on* D, *with consistency factor* $\bar{c}$. *Denote*

$$\rho = \min(r,\ 1/(2\bar{c}\alpha)). \tag{2.7}$$

Then $[M(x,h)]^{-1}$ *exists for all* $h \in U_\rho^m$, $x \in D$ *and*

$$[M(x,h)]^{-1} = \sum_{n=0}^{\infty} (I-[J(x)]^{-1} M(x,h))^n [J(x)]^{-1} \tag{2.8}$$

and

$$\|[M(x,h)]^{-1}\| \leq 2\alpha\ . \tag{2.9}$$

PROOF. Since $h \in U_\rho^m$, we know that $\|h\| \leq \rho$. Substituting this in (2.4) and using lemma 2.1 leads immediately to the required result. □

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

DEFINITION 2.5. We call a method as given by (1.5), for solving the nonlinear system $F(x) = 0$, where F satisfies (1.1), a *Proper Newton-like method* with consistency factor $\bar{c}$, if there exists an operator M as given by (2.3), for some integer m and some real r, and $h_k \in U_r^m$ $(k = 0,1,2,\ldots)$, such that

$$M_k = M(x_k,h_k), \qquad k = 0,1,\ldots, \tag{2.10}$$

and $M(x,h)$ is a strongly consistent approximation to $J(x)$ on D with consistency factor $\bar{c}$.

To study the convergence behaviour of proper Newton-like methods we compare them with Newton's method. Define, similar to (1.4) and (1.5)

$$\phi(x) = x - [J(x)]^{-1} F(x) \tag{2.11}$$

and

$$\psi(x) = x - [M(x,h)]^{-1} F(x). \tag{2.12}$$

Hence, $\phi(x)$ defines an iteration step of the Newton iteration and $\psi(x)$ defines an iteration step of a proper Newton-like method. Furthermore, we assume that $J(x)$ is nonsingular and satisfies (2.5) on $D \in \mathbb{R}^n$. Using the mean value theorem we obtain the following expression for the error in $\phi(x)$ as an approximation to the solution vector z:

$$\|\phi(x)-z\| = \|[J(x)]^{-1} (J(x)(x-z)-F(x))\| \leq S(x,z)\|x-z\|^2, \tag{2.13}$$

where

$$S(x,z) = \tfrac{1}{2}\Big(\sup_{y\in L[z,x]} \|H(y)\|\Big) \|[J(x)]^{-1}\| \tag{2.14}$$

and

$$L[z,x] = \{u \in \mathbb{R}^n \mid u = \theta x+(1-\theta)z,\ 0 \leq \theta \leq 1\}. \tag{2.15}$$

(2.13) expresses the well-known result, that the asymptotic order of con-

vergence of Newton's method is quadratic. In our further analysis we assume that that $F(x)$ and $M(x,h)$ satisfy the conditions of corollary 2.4. Then, the difference between $\psi(x)$ and $\phi(x)$ can be given by

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

$$= [I - \sum_{n=0}^{\infty} (I - [J(x)]^{-1} M(x,h))^n][J(x)]^{-1}\, F(x).$$

Hence, provided $h \in U_\rho^m$, where ρ is defined by (2.7), we obtain

$$\|\phi(x) - \psi(x)\| \leq \sum_{n=1}^{\infty} (\bar{c}\|h\|\, \|[J(x)]^{-1}\|)^n\, \|[J(x)]^{-1}\, F(x)\| \leq \tag{2.16}$$

$$\leq \bar{C}(x,h)\|[J(x)]^{-1}\, F(x)\|,$$

where

$$\bar{C}(x,h) = \bar{c}\|h\|\, \|[J(x)]^{-1}\|. \tag{2.17}$$

Furthermore,

$$\|[J(x)]^{-1}\, F(x)\| = \|x - \phi(x)\| \leq \|x - z\| + \|\phi(x) - z\|. \tag{2.18}$$

So, combining (2.13), (2.16) and (2.18), we obtain the following upper bound for the error in $\psi(x)$ as an approximation to z:

$$\|\psi(x) - z\| \leq \|\psi(x) - \phi(x)\| + \|\phi(x) - z\| \leq \tag{2.19}$$

$$\leq \bar{C}(x,h)\|x - z\| + (1 + \bar{C}(x,h))S(x,z)\|x - z\|^2.$$

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

The above results are summarized in the following definition.

DEFINITION 2.6. Let a nonlinear system be defined by F cf. (1.1) and let $x_0 \in D$ be an approximation to the solution z of the equation $F(x) = 0$. Then we say that this problem is *properly solvable* by a proper Newton-like method with consistency factor $\bar{c}$, if the following conditions are satisfied:

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

b. h_0 is chosen such that

$$\bar{C}(x_0,h_0) \leq \tfrac{1}{2} \tag{2.20}$$

and

$$r_0 = \bar{C}(x_0,h_0)\|\phi(x_0)-x_0\| + \|\phi(x_0)-z\| < \|x_0-z\|; \tag{2.21}$$

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

d. define

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

and h_k is chosen such that $\bar{K} \leq \frac{1}{2}$;

e.

$$\bar{\sigma}(F,z,x_0,\bar{c}) = \bar{K}+(\bar{K}+1)Sr_0 < 1, \tag{2.22}$$

where

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

If a. to d. are satisfied then $\bar{\sigma}(F,z,x_0,\bar{c})$ is called the *solvability number* of the Newton-like method with consistency factor $\bar{c}$, 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 2.7. *If a nonlinear system defined by* F *cf.* (1.1) *with initial approximation* x_0 *and solution* z *is properly solvable by a proper Newton-like method with consistency factor* $\bar{c}$, *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 (2.20) is satisfied we may use a similar argument as in corollary (2.4). With inequality (2.16) we obtain

$$\|\psi(x_0)-z\| \leq \|\psi(x_0)-\phi(x_0)\| + \|\phi(x_0)-z\|$$

$$\leq \bar{C}(x_0,h_0)\|\phi(x_0)-x_0\| + \|\phi(x_0)-z\| .$$

Because of (2.21) we know that

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

Because of c. d. and e. we can use (2.19) which gives

$$\|\psi(x)-z\| \leq \bar{\sigma}(F,z,x_0,\bar{c})\|x-z\| \leq \|x-z\| .$$

The result follows immediately. □

Although, in practice, condition e. is a rather strong condition, it gives us a clear insight in the behaviour 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.

3. THE EFFECT OF ROUNDING ERRORS

In this section we consider the effect of round off errors on the convergence behaviour 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 precision of computation ε.

When we want to apply the theory given in section 2 on 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

$$M_k = f\ell_\varepsilon(J(x_k)),$$

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

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

where $\delta \geq \varepsilon$ is some value depending on ε and the way M_k is calculated. Therefore, the notion "strongly consistent approximation" (cf. def. 2.2) 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. 2.2)
Let F be differentiable on $D \subset \mathbb{R}^n$ and let for some real number $r > 0$ and integral number $m \geq 0$ the operator M be defined by

$$M: D_0 \times U_r^m \subset \mathbb{R}^n \times \mathbb{R}^m \to L(\mathbb{R}^n), \tag{3.2}$$

where $D_0 \subset D$ and $U_r^m = \{y \in \mathbb{R}^m \mid \|y\| \leq r\}$. Then $M(x,h)$ is called a *numerically consistent approximation* to $J(x)$ on D_0, if there exist a constant c_1 and a function $c_0(\varepsilon,h)$ which is continuous in ε and h for $\varepsilon \geq 0$ and $h \in U_r^m \setminus \{0\}$, such that for all $x \in D_0$ and $h \in U_r^m \setminus \{0\}$ the following conditions are satisfied:

$$\|J(x) - f\ell_\varepsilon(M(x,h))\| \leq c_0(\varepsilon,h) + c_1\|h\|, \tag{3.3}$$

$$\lim_{\varepsilon \to 0} c_0(\varepsilon,h) = 0. \tag{3.4}$$

We call

$$c(x,h) = c_0(x,h) + c_1\|h\| \tag{3.5}$$

the *consistency function* of M.

As an example of a numerically consistent approximation we again

consider the forward difference approximation B(x,h), defined by (1.6). We prove the following theorem.

THEOREM 3.2. *Assume that* F (cf. (1.1)) *is continuously differentiable on* D. *Then, for any compact set* $D_0 \subset D$ *there exists a* $\rho > \varepsilon$ *such that* B(x,h), *given by* (1.6), *is well-defined for* $h \in U_\rho^m$ *and* $x \in D_0$. *Moreover, if* (2.5) *is satisfied, then* B(x,h) *is a numerically consistent approximation to* J(x) *on* D_0.

PROOF. We use the following relations (WILKINSON [10], DEKKER [3]).

$$(3.6) \qquad |fl_\varepsilon(a\pm b)-(a\pm b)| \le (|a|+|b|)\varepsilon ,$$

$$(3.7) \qquad |fl_\varepsilon(a/b)-(a/b)| \le |a/b|\varepsilon.$$

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

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

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

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

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

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

$$\|fl_\varepsilon(B(x,h))-J(x)\| \le (1+\varepsilon)\|B(x,h)-J(x)\| + \varepsilon\|J(x)\| +$$

$$+ \frac{3(n+1)\delta}{h_{min}} \sup_{\|y-x\| \le \|h\|} (\|F(y)\|),$$

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

From theorem 2.3 and the fact that D_0 is compact and D open we know that there exist a $\rho > 0$ and a $\bar{c}_1$ such that

$$\|B(x,h)-J(x)\| \leq \bar{c}_1 \|h\|. \tag{3.8}$$

Choose

$$\begin{aligned} c_0(\varepsilon,h) &= \frac{3(n+1)\delta}{h_{min}} \sup_{\|y-x\| \leq \|h\|} (\|F(y)\|) + \varepsilon\|J(x)\|, \\ c_1 &= (1+\varepsilon)\bar{c}_1. \end{aligned} \tag{3.9}$$

Then the theorem is proved, since

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

□

The following corollary shows the relation between the condition number

$$\kappa(J(x)) = \|J(x)\| \, \|[J(x)]^{-1}\|$$

of the jacobian matrix and the condition number of its numerically consistent approximation.

COROLLARY 3.3. *Let* F *be given (cf.* (1.1)*) and let* J(x) *be nonsingular on* D. *Suppose, for some integer* m *and real* r *the operator* M *is defined by* (3.1). *Suppose* M(x,h) *is a numerically consistent approximation to* J(x) *on* D *with consistency function* c(ε,h). *Assume, moreover, that for all* $\varepsilon > 0$ *a value* $\rho > 0$ *exists such that*

$$\|[J(x)]^{-1}\| \leq 1/(2c(\varepsilon,h)), \quad \text{for all } x \in D, \quad h \in U_\rho^m \setminus \{0\}. \tag{3.10}$$

Then $[f\ell_\varepsilon(M(x,h))]^{-1}$ *exists and*

$$[f\ell_\varepsilon(M(x,h))]^{-1} = \sum_{n=0}^{\infty} (I - [J(x)]^{-1} f\ell_\varepsilon(M(x,h)))^n \, [J(x)]^{-1}, \tag{3.11}$$

(3.12) $$\|[f\ell_{\varepsilon}(M(x,h))]^{-1}\| \leq 1/c(\varepsilon,h),$$

(3.13) $$\kappa(f\ell_{\varepsilon}(M(x,h))) \leq 3\kappa(J(x)).$$

PROOF. The proof of (3.11) and (3.12) follows easily from definition 3.1 and lemma 2.1. For proving (3.13) denote $\bar{M} = f\ell_{\varepsilon}(M(x,h))$. Using (2.2) we obtain

$$\kappa(\bar{M}) = \|\bar{M}\| \, \|M^{-1}\| \leq \frac{\|[J(x)]^{-1}\| \, \|\bar{M}\|}{1 - \|[J(x)]^{-1}\| \, \|J(x)-\bar{M}\|} .$$

Substituting $\|\bar{M}\| \leq \|\bar{M}-J(x)\| + \|J(x)\|$ and $\kappa(J(x)) \geq 1$ we obtain the required result. □

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

DEFINITION 3.4. (see def. 2.5)
We call a numerical Newton-like method for solving the nonlinear system defined by F (cf. (1.1)), a *proper numerical Newton-like method* with consistency function c, if there exists an operator M as given by (3.2) for some integer m and real r, such that

$$M_k = f\ell_{\varepsilon}(M(x_k,h_k))$$

and M(x,h) is a numerically consistent approximation to J(x) on D.

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

(3.14) $$f\ell_{\varepsilon}(M(x,h))(\bar{\psi}(x)-x) = F(x).$$

Assume that the conditions of corollary 3.3 are satisfied. Then, an upper bound for the error in $\bar{\psi}(x)$ as an approximation to $\phi(x)$ (cf. (1.4)) can be given by

$$\|\phi(x)-\bar{\psi}(x)\| \leq C(x,h,\varepsilon)\|[J(x)]^{-1} F(x)\|, \tag{3.15}$$

where

$$C(x,h,\varepsilon) = c(\varepsilon,h)\|[J(x)]^{-1}\|, \tag{3.16}$$

and $c(\varepsilon,h)$ is given by (3.5). The proof of (3.15) uses corollary 3.3 in a similar way as corollary 2.4 is used in the proof of (2.16).

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

$$f\ell_\varepsilon(\bar{\psi}(x)) = f\ell_\varepsilon(f\ell_\varepsilon(\bar{\psi}(x)-x)+x), \tag{3.17}$$

where $f\ell_\varepsilon(\bar{\psi}(x)-x)$ denotes the numerical solution of the system (3.14), where $F(x)$ is replaced by $f\ell_\varepsilon(F(x))$.

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

$$\frac{\|\bar{x}-x^*\|}{\|x^*\|} \leq \kappa(A)\left[\frac{\varepsilon g(n)}{1-\kappa(A)\varepsilon g(n)} + \frac{\|\delta b\|}{\|b\|}\right], \tag{3.18}$$

where $g(n)$ is some function depending on the order n, the norm used and the pivoting strategy used (WILKINSON [10]), and where it is assumed that

$$\kappa(A)\varepsilon g(n) < 1.$$

Applying this result to $f\ell_\varepsilon(\bar{\psi}(x)-x)$ we obtain

$$\|f\ell_\varepsilon(\bar{\psi}(x)-x)-(\bar{\psi}(x)-x)\| \leq \alpha(x,\varepsilon,n)\|\bar{\psi}(x)-x\|, \tag{3.19}$$

where

$$\alpha(x,\varepsilon,n) = 3\kappa(J(x))\left[\frac{\varepsilon g(n)}{1-3\kappa(J(x))\varepsilon g(n)} + \delta\right] \tag{3.20}$$

and δ satisfies

(3.21) $\|f\ell(F(x))-F(x)\| \le \delta\|F(x)\|$.

We assume that the conditions of corollary 3.3 are satisfied, so that (3.13) can be used, and, moreover, that

(3.22) $3\kappa(J(x))\varepsilon g(n) < 1$.

Combining (3.17) and (3.19) we obtain

$$\|f\ell_\varepsilon(\bar\psi(x))-\bar\psi(x)\| \le \varepsilon(\|f\ell_\varepsilon(\bar\psi(x)-x)\| + \|x\|) + \|f\ell_\varepsilon(\bar\psi(x)-x)-(\bar\psi(x)-x)\| \le$$

$$\le \varepsilon\|x\| + [(1+\varepsilon)\alpha(x,\varepsilon,n)+\varepsilon]\|\bar\psi(x)-x\|.$$

Thus

(3.23) $\|f\ell_\varepsilon(\bar\psi(x))-\bar\psi(x))\| \le \varepsilon\|x\| + \beta(x,\varepsilon,n)\|\bar\psi(x)-x\|$,

where

(3.24) $\beta(x,\varepsilon,n) = (1+\varepsilon)\alpha(x,\varepsilon,n) + \varepsilon$.

Finally, combining (2.13), (3.15) and (3.23) we obtain for the error in $f\ell_\varepsilon(\bar\psi(x))$ as an approximation to solution z:

(3.25) $\|f\ell_\varepsilon(\bar\psi(x))-z\| \le \|f\ell_\varepsilon(\bar\psi(x))-\bar\psi(x)\| + \|\bar\psi(x)-z\| \le$

$$\le \varepsilon\|x\| + \beta(x,\varepsilon,n)\|x-z\| + (1+\beta(x,\varepsilon,n))\|\bar\psi(x)-z\|.$$

With

$$\|\bar\psi(x)-z\| \le \|\bar\psi(x)-\phi(x)\| + \|\phi(x)-z\|$$

we obtain as the final result:

(3.26) $\|f\ell_\varepsilon(\bar\psi(x))-z\| \le \varepsilon\|x\| + L(x,\varepsilon,h,n)\|x-z\| + Q(x,\varepsilon,h,n,z)\|x-z\|^2$,

where

(3.27) $L(x,\varepsilon,h,n) = \beta(x,\varepsilon,n)+(1+\beta(x,\varepsilon,n))C(x,h,\varepsilon)$

and

(3.28) $\quad Q(x,\varepsilon,h,n,z) = (1+\beta(x,\varepsilon,n))(1+C(x,h,\varepsilon))S(x,z).$

From the first term in the right hand side of (3.26) we see that one can not 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, 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 $f\ell_\varepsilon(M(x,h))$ 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 $\kappa(J(x))$ should be small relative to $1/\varepsilon$ (cf. (3.20)).

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 3.5. (see def. 2.6)
Let a nonlinear system be defined by F (cf. (1.1)) and let $x_0 \in D$ be an approximation to the solution z of the equation $F(x) = 0$. Then, we call this problem *numerically solvable* by a proper numerical Newton-like method with consistency function $c(\varepsilon,h)$ and precision of computation ε, if the following conditions are satisfied:

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

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

where $g(n)$ depends on the method used for solving the linear system (cf. (3.18));

b. h_0 satisfies $C(x_0,h_0,\varepsilon) \leq \frac{1}{2}$, 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. $U_0 = \{y \in \mathbb{R}^n \mid \|y-z\| \leq r_0\} \subset D$

and

$$\sup_{x \in U_0} \kappa(J(x)) < 1/(3\varepsilon g(n));$$

d. define $K = \sup_{\substack{x \in U_0 \\ k=1,2,\ldots}} C(x,h_k,\varepsilon)$ and h_k is chosen such that $K \leq \frac{1}{2}$;

e. $\sigma(F,z,x_0,c,\varepsilon) = \beta + (1+\beta)C + (1+\beta)(1+C)S\, r_0 < 1,$

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

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

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

The following theorem is now easily proved.

THEOREM 3.6. (see theorem 2.7)
If a system of nonlinear equations defined by F *(cf.* (1.1)*) with initial approximation* x_0 *and solution* z *is numerically solvable by a proper numerical Newton-like method with consistency function* c *and precision of computation* ε*, then the sequence of points, generated by this method converges to a point* x^* *with* $\|x^*-z\| \leq \varepsilon\|x^*\|$.

PROOF. The proof is similar to the proof of theorem 2.7. Use of corollary 3.3 and the formulas (3.15), (3.25) and (3.26) leads immediately to the required result. □

4. SOME EXAMPLES

Consider the problem, given by GHERI & MANCINO [4]:

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

where

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

and

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

and the starting point is chosen to be

$$(4.2)\qquad x_0 = -F(0)\frac{c+K}{2cK},$$

where $c = \beta n - (\alpha+1)(n-1)$, $K = \beta n + (\alpha+1)(n-1)$. We consider the example for which

$$(4.3)\qquad n = 10, \quad \alpha = 5, \quad \beta = 14, \quad \gamma = 3.$$

Let method A be a Newton-like method with

$$(4.4)\qquad M_k^A = f\ell_{\varepsilon}(J(x_k))$$

and let method B be a Newton-like method with

$$(4.5)\qquad M_k^B = f\ell_{\varepsilon}(B(x,0.0001)),$$

where B(x,0.0001) is defined by (1.6) with $h_{ij} = h = 0.0001$, $i,j = 1,\ldots,n$. As a value of ε we use $\varepsilon = 10^{-14}$.

As is easily shown, the jacobian matrix

$$J(x) = (J_{ij}(x))$$

satisfies

$$J_{ii}(x) = \beta n, \qquad i = 1,\ldots,n$$

and

$$|J_{ij}(x)| \leq \alpha+1, \quad i,j = 1,\ldots,n, \quad i \neq j.$$

Therefore, using Gershgorin's theorem (see for instance WILKINSON [11]) we have for the smallest eigenvalue λ_{min} and the largest eigenvalue λ_{max} of $J^T(x)J(x)$:

$$\sqrt{\lambda_{min}} \geq 47.00; \qquad \sqrt{\lambda_{max}} \leq 200.$$

Hence, for the spectral norm we obtain

$$\|J(x)\| \leq 200, \quad \|[J(x)]^{-1}\| \leq 0.021. \tag{4.6}$$

$$\kappa(J(x)) \leq 4.2.$$

Furthermore,

$$H(x) = \begin{pmatrix} 0 & \cdots & 0 & & 0 \cdots 0 & h_{1n} \\ h_{21} & 0 \cdots 0 & & & \vdots & \vdots \\ \vdots & \vdots & & \cdots & \vdots & \vdots \\ \vdots & \vdots & & & \vdots & \vdots \\ h_{n1} & 0 \cdots 0 & & & 0 \cdots 0 & 0 \end{pmatrix},$$

where elementary computation shows that with the choice of $\alpha = 5$ we have

$$|h_{ij}| < 55.$$

Therefore, we have approximately

$$\|H(x)\| \leq 55\sqrt{n} < 180. \tag{4.7}$$

Using these results and assuming that gaussian elimination with complete pivoting is used, so that

$$g(n) \approx 20n^3 \tag{4.8}$$

(see WILKINSON [10]), it is easily seen that condition a. and c. of definition 3.5 is satisfied.

By (3.1) we have for the consistency function of method A:

$$c^A(x,h) = 200\delta.$$

By choosing the very reasonable value $\delta = 10^{-11}$, we obtain

(4.9) $\quad C^A(x,h,\varepsilon) \leq 4.2\ 10^{-11} < 0.5$ (condition b. of definition 3.5).

From (3.20) and (3.24) we obtain

(4.10) $\quad \beta^A(x,\varepsilon,n) < 3\ 10^{-9}.$

Hence

(4.11) $\quad \sigma^A(F,z,x_0,c,\varepsilon) \approx (0.5\times 180\times 0.021)r_0 \approx 1.89\ r_0.$

It appears from this result and from theorem 3.6 that the condition

(4.12) $\quad r_0 < 0.53$

is sufficient to guarantee convergence of method A. Numerical experiments show that this condition is easily satisfied.

For method B we use the mean value theorem to obtain a value for $\bar{c}_1$ in (3.8). Hence, with (4.7),

$$\bar{c}_1 \leq \sup_x(\|H(x)\|) < 180.$$

Using (3.9) we obtain for the consistency function of method B

(4.13) $$c^B(\varepsilon,h) = \frac{3(n+1)\delta}{h} \sup_{x\in D}(\|F(x)\|) + \varepsilon \sup_{x\in D}(\|J(x)\|) + \bar{c}_1 h,$$

where

$$D = \{x \in \mathbb{R}^n \mid \|x-z\| \leq \|x_0-z\|+h\}$$

and z is the solution of the problem.

Since the order of magnitude of $\|z\|$ and $\|x_0\|$ is about 1, we obtain after some elementary calculations

$$\sup_{x \in D} \|F(x)\| < 10^3,$$

so that, with the choice of h and δ, we obtain from (4.2)

$$c^B(\varepsilon,h) < 3\,10^{-3} + 2\,10^{-12} + 1.8\,10^{-2} \simeq 2.1\,10^{-2}$$

and

$$c^B(x,h,\varepsilon) < 4\,10^{-4} < 0.5 \qquad \text{(condition b. of definition 3.5).}$$

Hence, using (4.10), it is easily seen that

$$\sigma^B(F,z,x_0,c,\varepsilon) \simeq 4\,10^{-4} + 1.89\,r_0,$$

so that

$$r_0 < 0.53$$

is sufficient to guarantee convergence of method B.

These examples show that a rather simple analysis is sufficient sometimes to proof convergence of a numerical Newton-like method in advance, even for such complicated functions as given by (4.1). It is enough to know roughly the region in which the starting guess and the solution lies.

5. DISCUSSION

In this report, we gave an analysis of Newton-like methods for solving systems of nonlinear equations. The main results of this analysis are expressed in definition 3.5 and theorem 3.6. They establish sufficient conditions for global convergence of Newton-like algorithms when finite precision arithmetic is used. It appears from these conditions that the condition of the jacobian matrix and the consistency of the approximation to

the jacobian matrix used in the algorithm, are crucial points for the rate of convergence. A second result expressed in definition 3.5 is the introduction of a numerical solvability number. This number enables us to determine whether a problem may be expected to be easily solved. Although this is usually not useful for solving practical problems, it can be extremely useful when we have to create sets of test functions that should be used for testing algorithms for solving systems of nonlinear equations (see BUS [1]).

ACKNOWLEDGEMENTS. *The author is grateful to dr. P.J. van der Houwen, C. den Heijer and J. Kok for their careful reading of this manuscript and their suggestions about the framework of this report. I also like to thank Th. Gunsing, mrs C. Klein Velderman-Los and D. Zwarst for their effort to get this report typed and printed.*

REFERENCES

[1] BUS, J.C.P., *A comparative study of algorithms for solving nonlinear equations*, Mathematical Centre (to appear).

[2] COLLATZ, L., *Funktional Analysis und numerische Mathematik*, German, ed. Springer, Berlin (1964), English ed. Acad. Press, New York, (1966).

[3] DEKKER, T.J., *Numerical Algebra*, (Dutch), Mathematical Centre, Syllabus 12 (1971).

[4] GHERI, G. & O.G. MANCINO, *A significant example to test methods for solving systems of nonlinear equations*, Calculo, 8 (1971) 107-113.

[5] KANTOROVICH, L., *On Newton's method for functional equations*, (Russian), Dokl. Akad. Nauk. SSSR, 59 (1948) 1237-1240.

[6] LEVENBERG, K., *A method for the solution of certain nonlinear problems in least squares*, Quart. Appl. Math., 2 (1944) 164-168.

[7] MARQUARDT, D.W., *An algorithm for least-squares estimation of nonlinear parameters*, SIAM. J. 11 (1963) 431-441.

[8] ORTEGA, J.M. & W.C. RHEINBOLDT, *Iterative solution of nonlinear equations in several variables*, Acad. Press (1970).

[9] RALL, L.B., *Computational solution of nonlinear operator equations*, Wiley (1969).

[10] WILKINSON, J.H., *Rounding errors in algebraic processes*, Notes on applied science no. 32, Prentice Hall (1963).

[11] WILKINSON, J.H., *The algebraic eigenvalue problem*, Clarendon Press (1965).