NUMERICAL STABILITY FOR SOLVING NONLINEAR EQUATIONS

H. Wozniakowski
Department of Computer Science
Carnegie-Mellon University
Pittsburgh, Pa. 15213
(On leave from the University of Warsaw)
February 1975

This work was supported in part by the Office of Naval Research under Contract N00014-67-0314-0010 NR 044-422 and by the National Science Foundation under Grant GJ32111

ABSTRACT

The concepts of the condition number, numerical stability and well-behavior for solving systems of nonlinear equations $F(x) = 0$ are introduced. Necessary and sufficient conditions for numerical stability and well-behavior of a stationary iteration are given. We prove numerical stability and well-behavior of the Newton iteration for solving systems of equations and of some variants of secant iteration for solving a single equation under a natural assumption on the computed evaluation of F. Furthermore we show that the Steffensen iteration is unstable and show how to modify it to have well-behavior and hence stability.

1. INTRODUCTION

An algorithm for the solution of nonlinear equations and systems of equations should satisfy a number of criteria. Among these criteria are that it should enjoy good convergence properties, be efficient, and be numerically stable. Convergence continues to be extensively studied (see Ortega and Rheinboldt (1970)). Analytic computational complexity, which deals with the theory of efficient iteration, is under current investigation (see Traub (1973), (1974) for recent surveys).

In this paper we study numerical stability for solving nonlinear equations. We wish to solve $F(x) = 0$. Assume that the function F is sufficiently smooth and depends parametrically on a data vector d, i.e., $F(x) = F(x;d)$ (Section 2). Then we define the condition number cond(F;d) of F with respect to d which measures the relative sensitivity of the solution with respect to a small relative perturbation of the data vector (Section 3). Section 4 deals with numerical stability and well-behavior of iterative methods for the solution $F(x;d) = 0$. An iteration is said to be numerically stable if it produces a sequence $\{x_k\}$ of the approximations of the solution α such that for large k the relative error $\|x_k-\alpha\|/\|\alpha\|$ is of order $\zeta(1 + \mathrm{cond}(F;d))$ where ζ is the relative computer precision. An iteration is said to be well-behaved if a slightly perturbed x_k is an almost exact solution of a slightly perturbed problem, i.e., $F(x_k + \delta x_k;\mathbf{d} + \delta\mathbf{d}_k) = O(\zeta^2)$ where $\|\delta x_k\|/\|x_k\|$ and $\|\delta\mathbf{d}_k\|/\|\mathbf{d}\|$ are of order ζ. Note that well-behavior implies numerical stability. Next we prove necessary and sufficient conditions for a stationary iteration to be numerically stable and well-behaved.

In Section 5 we discuss numerical properties of Newton iteration for the multivariate case and secant iteration for the scalar case. We prove that Newton iteration is well-behaved under a natural assumption on the computed evaluation of F. Secant iteration is also well-behaved whenever an additional assumption holds. This assumption does not hold for the Steffensen iteration which implies that Steffensen iteration is numerically unstable. However, it is shown how to modify Steffensen iteration to get well-behavior.

2. DATA VECTOR

We consider the numerical solution of the equation

(2.1) $F(x) = 0,$

where F is in general a nonlinear function,

$$F: D_x \to \mathbb{C}^N, \ D_x \subset \mathbb{C}^N,$$

where $\mathbb{C}^N$ is the N dimensional complex space.

We want to define a condition number of the function F. The condition number should measure the sensitivity of the solution (output) with respect to the change of the data (input). The first question which arises is what we mean by data of a nonlinear function. For a particular F a data vector can be given implicitly. For instance if $N = 1$ and

$$F(x) = \sum_{i=1}^{n} a_i x^i$$

it is natural to assume that the data are all coefficients a_i of the polynomial. Next if $F(x) = Ax+b$ for a N×N matrix A and N dimensional vector b we usually mean by data all entries of A and b, while if the matrix A is sparse we can mean by data only nonzero entries of A.

In general, we shall assume that the function F from (2.1) parametrically depends on a vector "d", i.e.,

(2.2) $F(x) = F(x;d)$ where $d \in D_d \subset \mathbb{C}^m,$

This vector "d" will be called a data vector. We shall treat d as an element of a normed vector space and, in general, one should pay much attention how to choose a norm in that space to fit the problem (see Section 3). For certain F it may not be obvious how d should be chosen. An example is given by

(2.3) $F(x) = x^2 - e^x, \quad N = 1.$

We present a general idea how to define a data vector. As it was mentioned at the outset we want to solve (2.1) by iteration. Let x be a sufficiently close approximation of the solution α, $F(\alpha) = 0$. Suppose we use the value of $F(x)$ to get the next approximation. In numerical practice instead of the exact value $F(x;d)$ we only have the computed value of $F(x;d)$ in t digit floating point binary arithmetic (see Wilkinson (1963)). Let us denote that computed value by $fl(F(x;d))$. At best we can expect that a slightly perturbed computed value is the exact one of a slightly perturbed function at slightly perturbed inputs (see Kahan (1971)), i.e., (2.4) $fl(F(x;d)) = (I + \Delta F)\, F(x + \Delta x; d + \Delta d)$

where I denotes the unit N×N matrix,

$$\|\Delta F\| \le \zeta\, K_F, \quad \Delta F \text{ is a } N\times N \text{ matrix},$$
$$\|\Delta x\| \le \zeta\, K_x \|x\|,$$
$$\|\Delta d\| \le \zeta\, K_d \|d\|$$

for constants K_F, K_x and K_d which can only depend on the sizes N,m, and $\zeta = 2^{-t}$ denotes the relative computer precision. Here $\|d\|$ is a choosing norm in $\mathbb{C}^m$ which should fit the problem.

The condition (2.4) can be treated as an equation on a data vector. It means that for a given algorithm for the evaluation of F we want to define a data vector such that the condition (2.4) holds. Let us illustrate this point by an example.

Example 2.1

Let

$$F(x) = x^2 - e^x, \quad x \in D_x.$$

We assume that the result of a computer subroutine for the evaluation e^x satisfies

$$fl(e^x) = (1+\epsilon_1)e^{x+\Delta x}$$

where $|\Delta x| \le C\cdot\zeta\,|x|$ and $|\epsilon_1| \le C\zeta$ for a constant C which does not depend on $x \in D_x$. Then

(2.5) $fl(F(x)) = (1+\epsilon_3)(x^2(1+\epsilon_2) - (1+\epsilon_1)e^{x+\Delta x}) = (1+\eta_1)(1+\eta_2)((x+\Delta x)^2 - e^{x+\Delta x})$.

where

$$1+\eta_1 = (1+\epsilon_3)\cdot(1+\epsilon_1), \qquad |\eta_1| \le 2\cdot\zeta + O(\zeta^2),$$
$$1+\eta_2 = (1+\epsilon_2)[(1+\epsilon_1)(1+\Delta x/x)^2]^{-1}, \quad |\eta_2| \le 2(C+1)\zeta + O(\zeta^2).$$

The factor $1+\eta_1$ is a perturbation of the computed value and Δx is a perturbation of x. The factor $1+\eta_2$ may be interpreted as a perturbation of a data vector. Let us define

(2.6) $F(x;d) = d\,x^2 - e^x, \quad d \in \mathbb{C}, \quad m = 1.$

Hence, our problem is to solve $F(x;1) = 0$. From (2.5) and (2.6) it follows

$$fl(F(x;1)) = (1+\Delta F)F(x+\Delta x;\ 1+\Delta d)$$

where

$$\Delta F = \eta_1, \quad |\Delta F| \le 2\cdot\zeta + O(\zeta^2),$$
$$\Delta d = \eta_2, \quad |\Delta d| \le 2(C+1)\zeta + O(\zeta^2).$$

Hence (2.4) holds.

The definition of a data vector does not need to be unique. For instance, from (2.5) we can interpret the computed value as

$$\mathrm{fl}(F(x)) = \frac{1+\eta_1}{1+\eta_2}((x+\Delta x)^2 - \frac{1}{1+\eta_2} e^{x+\Delta x}).$$

Setting

$$\tilde{F}(x;d) = x^2 - de^x$$

we get

(2.7) $\mathrm{fl}(\tilde{F}(x;1)) = (1+\Delta\tilde{F})\tilde{F}(x+\Delta x;\ 1+\tilde{\Delta}d)$

where now $\tilde{F}(x;1) \equiv F(x;1) \equiv F(x)$ and

$$1+\Delta\tilde{F} = \frac{1+\eta_1}{1+\eta_2}; \quad |\Delta\tilde{F}| \le 2(C+2)\zeta + O(\zeta^2),$$

$$1+\Delta\tilde{d} = \frac{1}{1+\eta_2}; \quad |\Delta\tilde{d}| \le 2(C+1)\zeta + O(\zeta^2).$$

Hence (2.4) also holds. ∎

The lack of the uniqueness of a data vector causes no problems. In the next section we shall define the condition number of F with respect to the data vector. As the condition number measures the sensitivity of the solution when the data vector slightly changes it is reasonable to seek a data vector which minimizes the condition number and for which we can find an algorithm such that (2.4) holds.

3. CONDITION NUMBER

We want to solve the equation

(3.1) $F(x;d) = 0$

where F is now assumed that

$$F:\ D_x x D_d \to \mathbb{C}^N$$

and $D_x x D_d$ is an open subset of $\mathbb{C}^N \times \mathbb{C}^m$.

Let $\tilde{d} = rd(d)$ denote t digit representation of d in floating point arithmetic, fl. Then for all components of $\tilde{d}$ hold $\tilde{d}_i = d_i(1+\eta_i)$ where $|\eta_i| \le \zeta$, $i = 1,\dots,n$, and

(3.2) $\|\tilde{d}-d\| \le C_\zeta \|d\|.$

Here $\zeta = 2^{-t}$ is the relative computer precision and C only depends on the size m and the given norm. If the norm $\|\cdot\|_p$ is used, $1 \le p \le +\infty$, then $C = 1$.

It should be stressed the necessity of choosing a norm of d which fits the problem. For instance, in many cases we can set $D_d = \{y:\ \|y-d\| \le \Gamma\}$ where Γ is small enough (say, $\Gamma = O(\zeta)$). Then if all components of d are nonzero numbers we can define

$$\|\tilde{d}\| = \sqrt{\sum_{i=1}^{m} \gamma_i |\tilde{d}_i|^2} \quad \text{where } \gamma_i = H/|d_i|^2 \quad \text{for any } H > 0.$$

Now (3.2) holds with $C = 1$ and moreover, this norm exposes the inaccuracy in all components of $\tilde{d}$ (see Stewart (1973), p. 186).

In general we shall assume that the considered norm of d fits the problem. Note that if ζ is small enough, then $\tilde{d} \in D_d$. Usually $\tilde{d} \neq d$ which means that instead of the equation (3.1) we can at best approximate the solution of a perturbed equation

(3.3) $F(x;\tilde{d}) = 0.$

Note that this unavoidable change of the data vector does not depend on a method which uses d and solves (3.1).

Assume (3.1) has a simple root α and let F be a sufficiently smooth function of x and d. (In fact, it is sufficient for (3.4) to assume that F has a Lipschitz first derivative in a neighborhood of (α, d)). If t is sufficiently large, then it is straightforward to verify that (3.3) has a unique, simple solution $\tilde{\alpha}$ in a neighborhood of α and

(3.4) $\tilde{\alpha} - \alpha = -F_x'(\alpha;d)^{-1} F_d'(\alpha;d)(\tilde{d}-d) + O(\zeta^2),$

where F_x' and F_d' denote the derivatives with respect to x and d. The constant which appears in the "O" notation can depend on F, α and d. For $\alpha \neq 0$, from (3.2) and (3.4) it follows

(3.5) $\dfrac{\|\tilde{\alpha}-\alpha\|}{\|\alpha\|} \le C\zeta \operatorname{cond}(F;d) + O(\zeta^2)$

where

(3.6) $\operatorname{cond}(F;d) = \|F_x'(\alpha;d)^{-1} F_d'(\alpha;d)\| \dfrac{\|d\|}{\|\alpha\|}$

is <u>called the condition number of F with respect to the data vector d</u>.

Note that (3.5) is, in general, sharp. This means that an unavoidable change of the solution mainly depends on two factors:

(i) ζ which is the relative computer precision; for modern computers; $\zeta \in [10^{-16}, 10^{-6}]$ for most machines

(ii) cond(F;d) which measures the relative sensitivity of the solution with respect to small relative perturbations of the data vector.

Hence, we can at best compute an approximation of α with the relative error of order $\zeta \cdot \operatorname{cond}(F;d)$. If the problem is ill-conditioned, i.e. $\operatorname{cond}(F;d) \gg 1$, it is impossible to compute a good approximation of α no matter how sophisticated a method is used. If the problem is extremely ill-conditioned, $\operatorname{cond}(F;d) \ge \frac{1}{\zeta}$, then in general we do not compute any reasonable approximation of α. For such a case (which can be called <u>numerically singular</u>) it seems to be necessary to increase the relative precision to $\zeta_1 = \zeta^k$ for k such that $\text{cond}(F;d)\zeta_1 \ll 1$. (See a similar approach in Wilkinson (1963).)

Let us illustrate the concept of the condition number by a few examples.

Example 3.1. Solution of a Linear System

Let

$$F(x;d) = Ax + b, \quad A = [a_1,\ldots,a_N], \quad a_i, b \in \mathbb{C}^N, \ b \neq 0,$$

where the data vector $d = [a_1^T,\ldots,a_N^T]^T \in \mathbb{C}^m$, $m = N^2$. For the sake of simplicity we do not include b as a part of the data vector. Thus,

$$F'_x(x;d) = A, \ F'_d(x;d) = [x_1 I,\ldots,x_N I]$$

where $x = [x_1,\ldots,x_N]^T$ and I is the unit N×N matrix. The condition number cond(F;d) is now equal to

$$\text{cond}(F;d) = \frac{\|A^{-1}F'_d(\alpha;d)\|_2 \ \|d\|_2}{\|\alpha\|_2} = \|A^{-1}\|_2 \ \|A\|_2 .$$

Hence for a linear system cond(F;d) is the usual condition number of the matrix A, $\kappa(A) = \|A^{-1}\|_2\|A\|_2$. ∎

Example 3.2. Root of a Scalar Polynomial

Let

$$F(x;d) = \sum_{i=0}^{n} d_i x^i, \quad d = [d_0,\ldots,d_n]^T \in \mathbb{C}^m, \quad m = n+1.$$

Then, (3.4) becomes

$$\tilde{\alpha} - \alpha = \frac{-1}{F'(\alpha)} \sum_{i=0}^{n} \Delta d_i \ \alpha^i + O(\zeta^2)$$

where Δd_i is the ith component of $\Delta d = \tilde{d} - d$, $\|\Delta d\| \le C\zeta \ \|d\|$, (see Wilkinson (1963) pp. 38-41). The condition number is equal to

$$\text{cond}(F;d) = \sqrt{\sum_{i=0}^{n} \ |\alpha|^{2i-1}} \ \|d\|_2/|F'(\alpha)|.$$ ∎

It should be stressed that one can normalized the considered problems in Examples 3.1 and 3.2 by choosing a suitable norm.

Example 3.3. Solution of a Nonlinear System

Suppose we solve F(x) = 0 by Newton iteration. Let x_k be a sufficiently close approximation of α. The next point x_{k+1} is given by

$$\text{(3.7)} \quad F'(x_k)z_k = -F(x_k),$$
$$x_{k+1} = x_k + z_k.$$

It might seem that the numerical accuracy of x_{k+1} depends mainly on the condition number $\kappa(F'(\alpha)) = \|F'(\alpha)\| \, \|F'(\alpha)^{-1}\|$, which is crucial for the relative accuracy of the solution of linear equations. (Note that for $N = 1$, $\kappa(F'(\alpha)) = 1$ which might imply that all scalar nonlinear problems are perfectly well-conditioned!) We shall show that the numerical accuracy for nonlinear problems depends on the condition number cond(F;d) which is, in general, not related to $\kappa(F'(\alpha))$. An intuitive reason that $\kappa(F'(\alpha))$ does not reflect on the numerical accuracy for nonlinear problems is that the righthand side of (3.7) tends (at least in theory) to zero and we can exactly solve a homogeneous system no matter how ill-conditioned it is.

To illustrate this point we consider an idealized case of (3.7). Namely let $F(x_k)$ and $F'(x_k)$ be error free and the only one rounding-error source is the solution of the linear system (3.7). We can assume that the computed z_k is the exact solution of a slightly perturbed problem, i.e.,

$$(3.8) \quad (F'(x_k) + E_k)z_k = - F(x_k), \ \|E_k\| \le \zeta C_1 \| F'(x_k)\|$$

for a constant $C_1 = C_1(N)$.

Assume that $q = \zeta C_1 \kappa(F'(\alpha))/(1-C_1\zeta\kappa(F'(\alpha)) < 1$. Then for the "computed" x_{k+1} holds

$$e_{k+1} \le C_2 e_k^2 + q e_k = (C_2 e_k + q)e_k$$

where $e_k = \|x_k - \alpha\|$ and $C_2 = C_2(F)$. Thus if $e_0 < (1-q)/C_2$ then the "computed" sequence tends to α, although for large k the convergence is linear, $e_{k+1} \cong q e_k$ and it depends on $\kappa(F'(\alpha))$. However, if for fixed F, the relative precision ζ tends to zero the condition number $\kappa(F'(\alpha))$ gets less important.

A real case when $F(x_k)$ and $F'(x_k)$ are not error free is considered in Section 5.

We wish to finish this example by showing a problem for which $\kappa(F'(\alpha))$ is extremely large but cond(F;d) is very moderate.

Let $N = 2$, $x = [x_1, x_2]^T$ and

$$F(x) = [x_1 - x_2, \ x_1^2 + Cx_2^2 - C]^T$$

where a constant $C > 0$. The solution $\alpha = \sqrt{\frac{C}{1+C}}\,[1,1]^T$. We need to define a data vector for F. For $x = rd(x)$ we get

$$fl(F(x)) = (I - \Delta F)\begin{bmatrix} x_1 - x_2 \\ (1+\epsilon_1)(x_1^2 + Cx_2^2) - C \end{bmatrix}$$

where $\|\Delta F\| \le 2 \cdot 2^{-t}$, $|\epsilon_1| \le K \cdot 2^{-t}$, $K \cong 3$.

Setting

$$F(x;d) = [x_1 - x_2, \ d(x_1^2 + Cx_2^2) - C]^T, \quad d \in \mathbb{C}, \ m = 1,$$

our problem is to solve $F(x;1) = 0$.

Since

$$F_x'(\alpha;1)^{-1}\ F_d'(\alpha;1) = \frac{1}{2}\alpha$$

we conclude

$$\mathrm{cond}(F;1) = \frac{1}{2} \quad \forall\ C > 0,$$

which means that the problem is extremely well-conditioned. But

$$\lim_{\substack{C\to 0 \\ \text{or } C\to+\infty}} \bar{k}(F'(\alpha)) = +\infty$$

which proves that cond(F;1) and $\bar{k}(F'(\alpha))$ are not in general related.

4. NUMERICAL STABILITY AND WELL-BEHAVIOR OF ITERATIONS

Let us suppose that $F(x;d) = 0$ is solved by an iteration φ. Let $\{x_k\}$ be a computed sequence of the successive approximations of α by an iteration φ. We know we can at best approximate $\tilde{\alpha}$, the solution of $F(x;\tilde{d}) = 0$. It means that, in general, we cannot expect x_k to be closer to $\tilde{\alpha}$ than $\zeta\|\tilde{\alpha}\|$. Thus, for large k,

$$\|x_k-\alpha\| \le \|\tilde{\alpha}-\alpha\| + \|x_k-\tilde{\alpha}\| \le \|\tilde{\alpha}-\alpha\| + k_1\zeta\|\tilde{\alpha}\|,$$

k_1 is a constant. Keeping in mind that $\tilde{\alpha}_k-\alpha$ is given by (3.4) and (3.5) we get the following definitions.

<u>Definition 4.1</u>

(i) An iteration φ is called <u>numerically stable</u> if

$$\text{(4.1)}\qquad \overline{\lim_k}\ \|x_k-\alpha\| \le \zeta(k_1\|\alpha\| + k_2\ \|F_x'(\alpha;d)^{-1}F_d'(\alpha;d)\|\ \|d\|\ ) + O(\zeta^2).$$

(ii) An iteration φ is called <u>well-behaved</u> if there exist $\{\delta x_k\}$ and $\{\delta d_k\}$ such that

$$\text{(4.2)}\qquad \overline{\lim_k}\ \|F(x_k+\delta x_k;\ d+\delta d_k)\| = O(\zeta^2)$$

and $\|\delta x_k\| \le k_3\zeta\|x_k\|$, $\|\delta d_k\| \le k_4\zeta\ \|d\|$ for large k

where k_i can only depend on N and m, $i = 1,\dots,4$. (See Jankowska (1974), Kielbasinski (1974).) ∎

Well-behavior states that a slightly perturbed computed x_k, k large, is an almost exact solution of a slightly perturbed problem (see Kahan (1971).

It should be stressed that $O(\zeta^2)$ in (4.1) and (4.2) can be dropped whenever we redefined $k_i = k_i(N,m) + O(\zeta^2)$. We prefer the form of (4.1) and (4.2) as it is a simple generalization of (3.4) and (3.5).

In practice we often want to find an approximation x_k such that $\|x_k-\alpha\| \le \epsilon\ \|x_k\|$ for a moderate value of ε, say $\varepsilon \in [10^{-5}, 10^{-2}]$. This is possible if the problem is sufficiently well-conditioned, with respect to the available numerical arithmetic, namely if $\mathrm{cond}(F;d)\zeta$ is of order ε.

Note that if φ is well-behaved then it is also numerically stable but in general not vice versa. However, for scalar problems, $N = 1$, these two concepts are equivalent which is proved in Lemma 4.1.

Lemma 4.1

If $N = 1$ then numerical stability of φ is equivalent to well-behavior of φ. ■

Proof

It is enough to assume that φ is numerically stable and to prove it is well-behaved. Without loss of generality we can assume to use the second norm, $\|\cdot\| = \|\cdot\|_2$. Hence, from (4.1) it follows

$$x_k - \alpha = \zeta\, C_k\, k_1 |\alpha| + \zeta\, C_k\, k_2 \|F_d'(\alpha)\|\, \|d\| / |F'(\alpha)| + O(\zeta^2)$$

for large k, constants C_k such that $|C_k| \le 1$ and $F'(\alpha) \equiv F_x'(\alpha;d)$, $F_d'(\alpha) \equiv F_d'(\alpha;d)$. We want to show that

$$F(x_k+\delta x_k;\ d+\delta d_k) = O(\zeta^2)$$

for suitable chosen δx_k and δd_k. From numerical stability it follows

$$F(x_k+\delta x_k;\ d+\delta d_k) = F'(\alpha)(x_k-\alpha+\delta x_k) + F_d'(\alpha)\,\delta d_k + O(\|\delta d_k\|^2 + \|x_k-\alpha+\delta x_k\|^2) =$$

$$= F'(\alpha)(\zeta C_k k_1 |x_k| + \delta x_k) + F_d'(\alpha)\,\delta d_k + \zeta \bar{C}_k k_2 \|F_d'(\alpha)\|\, \|d\| + O(\|\delta d_k\|^2 + \|x_k-\alpha+\delta x_k\|^2)$$

where $|\bar{C}_k| = |C_k| \le 1$.

Setting $\delta x_k = -\zeta C_k k_1 |x_k|$ and $\delta d_k = -\zeta \bar{C}_k k_2 \|d\| \cdot u$ where $u = F_d'(\alpha)^T / \|F_d'(\alpha)\|$ we get $F(x_k+\delta x_k;\ d+\delta d_k) = O(\zeta^2)$ which means that φ is well-behaved. ■

The next part of this section deals with numerical stability and well-behavior for stationary iterative methods. Let $(x_k,\dots,x_{k-n})$ be approximations sufficiently close to the solution α, $F(\alpha) = 0$. Suppose that the next approximation $\overset{*}{x}_{k+1}$ is given by a stationary iteration φ, namely,

(4.3) $\quad \overset{*}{x}_{k+1} = \varphi(x_k;F)$

where
$$\varphi(x_k;F) = \varphi(x_k,\dots,x_{k-n};\ \mathfrak{N}(x_k,\dots,x_{k-n};F))$$

and $\mathfrak{N} = \mathfrak{N}(x_k,\dots,x_{k-n};F)$ is generalized information of F at $x_k,\dots,x_{k-n}$ points. For instance $\mathfrak{N}$ can be so called standard information given by values of F and its first derivatives,

(4.4) $\quad \mathfrak{N}(x_k,\dots,x_{k-n};F) = \{F^{(i)}(x_{k-j}):\ i = 0,1,\dots,s;\ j = 0,1,\dots,n\}$

(for details see Wozniakowski (1975a)).

Note that the nonnegative integer n is the number of iteration points at which one reuses the information of F. Next, suppose that there exists a constant $C = C(F)$ such that for all sufficiently close approximations $(x_k,\dots,x_{k-n})$ to α such that $\|x_k-\alpha\| \le \dots \le \|x_{k-n}-\alpha\|$, the next x^*_{k+1} satisfies

$$(4.5) \quad \|x^*_{k+1}-\alpha\| \le C \prod_{j=0}^{n} \|x_{k-j}-\alpha\|^{p_j}$$

where $p_j \ge 0$, $\nu \equiv \sum_{j=0}^{n} p_j \ge 1$. If $\nu = 1$ then assume $C < 1$. If (4.5) is sharp then the unique positive zero p, $p \ge 1$, of the polynomial $t^{n+1} - \sum_{j=0}^{n} p_j t^{n-j}$ is called the order of φ (for details see Wozniakowski (1974)).

Conditions (4.3) and (4.5) describe theoretical properties of a stationary iteration φ. In floating point arithmetic, instead of (4.3), we have

$$(4.6) \quad x_{k+1} = \varphi(x_k;F) + \xi_k$$

where

$$(4.7) \quad \xi_k = fl(\varphi(x_k;F)) - \varphi(x_k;F)$$

is the computed error in one iterative step. The value of ξ_k depends on the computed error of the generalized information $\mathfrak{N}$ as well on the computed error of an algorithm which is used to perform one iterative step. We want to show necessary and sufficient conditions on $\{\xi_k\}$ to get numerical stability and well-behavior.

<u>Theorem 4.1</u>

Let φ be a stationary iterative method defined by (4.3) and (4.5). Let $x_n,x_{n-1},\dots,x_0$ be initial approximations of a simple zero α of a sufficiently smooth function F, $F(\alpha;d) = 0$, $F(x) \equiv F(x;d)$. Let $\|x_n-\alpha\| \le \dots \le \|x_0-\alpha\| \le \Gamma$ where $C(F)\Gamma^{\nu-1} < 1$ for $\nu = \sum_{j=0}^{n} p_j \ge 1$ and $C(F)$ is a constant from (4.5).

Suppose that

$$(4.8) \quad \|\xi_k\| \le \Gamma(1 - C(F)\Gamma^{\nu-1}) \text{ for all } k.$$

(i) Let $\nu \ge 2$. A stationary iteration φ is numerically stable iff

$$(4.9) \quad \overline{\lim_k} \|\xi_k\| \le \beta \equiv \zeta(k_1\|\alpha\| + k_2\|F'_x(\alpha;d)^{-1}F'_d(\alpha;d)\| \, \|d\|) + O(\zeta^2).$$

(ii) A stationary iteration φ is well-behaved iff for $k \ge k_0$ there exist $\{\Delta x_k\}$ and $\{\Delta d_k\}$ such that

$$(4.10) \quad \xi_k = x_k - \varphi(x_k;F) - F'_x(x_k)^{-1}F(x_k) - F'_x(x_k)^{-1}\{F'_x(x_k)\Delta x_k + F'_d(x_k)\Delta d_k\} + O(\zeta^2)$$

where $\|\Delta x_k\| \le k_3\,\zeta\|x_k\|$, $\|\Delta d_k\| \le k_4\,\zeta\|d\|$.

(Constants k_i can only depend on N and m.)

Proof

(i) First we deal with numerical stability. Suppose that φ is numerically stable. This means that

$$e \equiv \overline{\lim_k} \|x_k-\alpha\| \le \beta.$$

From (4.6), (4.3) and (4.5) it follows

$$\overline{\lim_k}\|\xi_k\| = \overline{\lim_k} \|x_{k+1}-\varphi(x_k;F)\| \le \overline{\lim_k} (\|x_{k+1}-\alpha\| + \|x^*_{k+1}-\alpha\|) \le \beta + C(F)\beta^\nu.$$

Since $\beta = O(\zeta)$ and $\nu \ge 2$, then

$$\overline{\lim_k} \|\xi_k\| \le \beta + O(\zeta^2)$$

which completes this part of the proof.

Assume now that (4.9) holds. We want to prove that $e \le \beta + O(\zeta^2)$. First of all, suppose by induction that $e_k = \|x_k-\alpha\| \le \Gamma$. This is valid for $k = 0,1,\dots,n$ due to the assumption. Next, from (4.6), (4.5) and (4.8) it follows

$$e_{k+1} \le C(F)\Gamma^\nu + \|\xi_k\| \le C(F)\Gamma^\nu + \Gamma - C(F)\Gamma^\nu = \Gamma.$$

Thus, $e = \overline{\lim_k} e_k \le \Gamma$ and once more from (4.6) and (4.5) we get

(4.11) $\quad e \le C(F)e^\nu + \beta.$

Since $e \le \Gamma$, then $e \le \beta/(1 - C(F)\Gamma^{\nu-1}) = O(\zeta)$. From this and the fact that $\nu \ge 2$, (4.11) implies

$$e \le \beta + O(\zeta^2)$$

which completes the proof of numerical stability.

(ii) We now deal with well-behavior. Let φ be well-behaved. It means that for $k \ge k_0$ there exist $\{\delta x_k\}$ and $\{\delta d_k\}$ such that

$$O(\zeta^2) = F(x_{k+1}+\delta x_{k+1};\ d+\delta d_{k+1}) = F(x_k;d) + F'_x(x_k;d)(x_{k+1}-x_k+\delta x_{k+1}) + F'_d(x_k;d)\,\delta d_{k+1} + O(\|x_{k+1}-x_k\|^2) + O(\zeta^2),$$

for

$$\|\delta x_{k+1}\| \le k_3\ \zeta\|x_{k+1}\|,\ \ \|\delta d_k\| \le k_4\ \zeta\|d\|.$$

Since $\xi_k = x_{k+1} - \varphi(x_k;F)$ and $\|x_{k+1} - x_k\| = O(\zeta)$, we get

(4.12) $\quad \|\delta x_{k+1}\| \le k_3\ \zeta\|x_k\| + O(\zeta^2),$

(4.13) $\quad \xi_k = x_k - \varphi(x_k;F) - F'_x(x_k)^{-1}F(x_k) - F'_x(x_k)^{-1}\{F'_x(x_k)\,\delta x_{k+1} + F'_d(x_k)\,\delta d_{k+1}\} + O(\zeta^2).$

Due to (4.12) we can split δx_{k+1} as follows

$$\delta x_{k+1} = \delta^{(1)}x_{k+1} + \delta^{(2)}x_{k+1}$$

where $$\|\delta^{(1)}x_{k+1}\| \le k_3\ \zeta\|x_k\| \text{ and } \|\delta^{(2)}x_{k+1}\| = O(\zeta^2).$$

Set

$$\Delta x_k = \delta^{(1)}x_{k+1} \text{ and } \Delta d_k = \delta d_{k+1}.$$

Then (4.10) follows from (4.13) which completes this part of the proof.

Assume now that (4.10) holds. For $k \ge k_0$, (4.6) becomes

$$x_{k+1} = x_k - F_x'(x_k)^{-1}F(x_k) + O(\zeta)$$

which implies that

$$e_{k+1} = O(e_k^2) + O(\zeta) = O(\zeta).$$

We want to find $\{\delta x_k\}$ and $\{\delta d_k\}$ of order ζ such that (4.2) holds. From (4.14), (4.6) and (4.10) we get

$$(4.15)\quad F(x_{k+1}+\delta x_{k+1};\ d+\delta d_{k+1}) = F(x_k) + F_x'(x_k)(x_{k+1}-x_k+\delta x_{k+1}) + F_d'(x_k)\delta d_{k+1} +$$

$$+ O(\zeta^2) = F_x'(x_k)(\delta x_{k+1}-\Delta x_k) + F_d'(x_k)(\delta d_{k+1}-\Delta d_k) + O(\zeta^2).$$

Since $\|\Delta x_k\| \le k_3\ \zeta\|x_k\| \le k_3\ \zeta\|x_{k+1}\| + O(\zeta^2)$, we can split Δx_k, $\Delta x_k = \Delta^{(1)}x_k + \Delta^{(2)}x_k$, $\|\Delta^{(1)}x_k\| \le k_3\zeta\|x_{k+1}\|$, $\|\Delta^{(2)}x_k\| = O(\zeta^2)$.

Finally, setting

$$\delta x_{k+1} = \Delta^{(1)}x_k \text{ and } \delta d_{k+1} = \Delta d_d$$

(4.15) yields (4.2) which defines well-behavior of φ and which completes the proof. ■

Theorem 4.1 states an assumption on the vector ξ_k which implies numerical stability and well-behavior. Assumption (4.8) means that ξ_k has to be small enough. It is a natural assumption as many iterations are well-defined and have property (4.5) in a small neighborhood of the solution. In case (i) of Theorem 4.1 we assume $\nu \ge 2$. If $1 < \nu < 2$ then it is straightforward to verify that the same results hold with $O(\zeta^\nu)$ in place of $O(\zeta^2)$. However, if ν is close enough to unity one cannot neglect a term $O(\zeta^\nu)$ in the presence of $O(\zeta)$ for common used values of ζ. Thus, we prefer to assume $\nu \ge 2$ which seems to be valid for all iterations of practical interest with order higher than one.

An interesting question is numerical stability of iterations with linear convergence, $\nu = 1$ and $C(F) < 1$. It is easy to verify that

$$(4.16)\quad \overline{\lim_k}\ \|\xi_k\| \le (1 - C(F))\beta$$

assures numerical stability. Furthermore (4.16) seems to be necessary for numerical stability (see Wozniakowski (1975c), where the method of successive approximations for large linear systems $x = Bx+g$ is discussed. See also a proof of the numerical stability of Chebyshev method for large linear systems which is an example of a nonstationary iteration with linear convergence, Wozniakowski (1975b).).

Note that for well-behavior we need no assumption on ν. However, if $\nu \geq 2$ then (4.10) can be simplified. Note that

$$x_k - \varphi(x_k;F) - F_x(x_k)^{-1}F(x_k) = \alpha - \varphi(x_k;F) + O(\zeta^2) = O(\zeta^\nu+\zeta^2) = O(\zeta^2)$$

for large k. Thus, it is easy to verify the following corollary.

Corollary 4.2

Let $\nu \geq 2$. A stationary iteration φ is well-behaved iff for $k \geq k_0$ there exist $\{\Delta x_k\}$ and $\{\Delta d_k\}$ such that

$$\xi_k = \Delta x_k + F_x'(x_k)^{-1}F_a'(x_k)\Delta d_k + O(\zeta^2)$$

where

$$\|\Delta x_k\| \leq k_3\ \zeta\|x_k\|,\ \|\Delta d_k\| \leq k_4\ \zeta\|d\|$$

and

$$k_i = k_i(N,m) \quad \text{for} \quad i = 3,4.$$

5. NEWTON ITERATION

In this section we prove well-behavior of Newton iteration under natural assumptions on computed values of F. We recall that the Newton method constructs the next approximation as

(5.1) $\quad F'(x_k)(x_k-x_{k+1}^*) = F(x_k)$

and if x_k is close enough to a simple zero α of a "smooth" function F then

$$\|x_{k+1}^*-\alpha\| = O(\|x_k-\alpha\|^2).$$

An algorithm of one Newton step in fl arithmetic is given by

(i) compute $F(x_k)$, $F'(x_k)$,

(ii) solve a linear system

(5.2) $\quad F'(x_k)z_k = F(x_k)$ then

(5.3) $\quad x_{k+1} = x_k - z_k$.

Let us assume a well-behaved algorithm for the computation of F, i.e.,

(5.4) $\quad fl(F(x_k;d)) = (I+\Delta F_k)F(x_k+\Delta x_k;\ d+\Delta d_k) = F(x_k) + \delta F_k$

where $\|\Delta F_k\| \leq \zeta K_F$, $\|\Delta x_k\| \leq K_x\ \|x_k\|$, $\|\Delta d_k\| \leq K_d\ \|d\|$ (see (2.4)),

and

(5.5) $\quad \delta F_k = \Delta F_k\ F(x_k) + F_x'(x_k)\Delta x_k + F_d'(x_k)\Delta d_k + O(\zeta^2)$.

Further, let us assume that

(5.6) $fl(F'(x_k;d)) = F'(x_k) + \delta F'_k, \ \delta F'_k = O(\zeta).$

This means that we do not need a well-behaved algorithm for the evaluation of $F'(x_k)$. The constant which appears at $\delta F'_k$ in the "O" notation can be arbitrary. Finally, let us assume that a computed solution of the linear system (5.2) satisfies

(5.7) $(F'(x_k) + \delta F'_k + E_k)z_k = F(x_k) + \delta F_k$

where $E_k = O(\zeta)$.

Condition (5.7) means that z_k is the exact solution of a perturbed system, however we only claim that E_k is of order ζ and we do not specify what constant appears in the "O" notation. If one uses Gaussian elimination with pivoting or the Householder method then $\|E_k\| \le \zeta K \, \|F'(x_k)\|$ and K depends on the size N.

A computed approximation x_{k+1} from (5.3) satisfies

(5.8) $x_{k+1} = (I + \delta I_k)(x_k - z_k)$

where δI_k is a diagonal matrix and $\|\delta I_k\| \le C_1\zeta$, C_1 depends on a considered norm (if $\|\cdot\| = \|\cdot\|_p$, $1 \le p \le +\infty$ then $C_1 = 1$).

Theorem 5.1

If (5.4), (5.6) and (5.7) hold then Newton iteration is well-behaved. Specifically it produces a sequence $\{x_k\}$ such that

(5.9) $\overline{\lim_k} \, \|F(x_{k+1}+\Delta x_k - \delta I_k x_k;\ d+\Delta d_k)\| = O(\zeta^2)$

where Δx_k, δI_k and Δd_k are defined by (5.4) and (5.8). ■

Proof

Let

$$F'(x_k) + \delta F'_k + E_k = F'(x_k)(I+H_k) \quad \text{where}$$

$$H_k = F'(x_k)^{-1}\{\delta F'_k + E_k\} = O(\zeta)$$

due to (5.6) and (5.7).

Thus for small ζ, $I + H_k$ is invertible and

$$(I+H_k)^{-1} = I - H_k + O(\zeta^2).$$

From (5.8) and (5.7) the next approximation x_{k+1} is given by

$$x_{k+1} = (I+\delta I_k)(x_k - (I+H_k)^{-1}F'(x_k)^{-1}F(x_k) + \delta F_k)) = x_k - F'(x_k)^{-1}F(x_k) + \xi_k$$

where

$$(5.10)\quad \xi_k = \delta I_k(x_k - F'(x_k)^{-1}F(x_k)) - F'(x_k)^{-1}\delta F_k + H_k F'(x_k)^{-1}F(x_k) + O(\zeta^2).$$

We want to use Theorem 4.1. Condition (5.10) states that $\xi_k = O(\zeta)$ which means that for small ζ assumption (4.8) holds. Hence it is enough to show that ξ_k has the form of (4.10), e.g.,

$$(5.11)\quad \xi_k = - F'(x_k)^{-1}\{F'(x_k)\tilde{\Delta}x_k + F_d'(x_k)\tilde{\Delta}d_k\} + O(\zeta^2)$$

for suitable $\tilde{\Delta}x_k$ and $\tilde{\Delta}d_k$.

For a sufficiently good initial approximation we get

$$e_{k+1} = O(e_k^2) + \|\xi_k\| = O(e_k^2) + O(\zeta) = O(\zeta),$$

where as always $e_k = \|x_k-\alpha\|$.

From this and (5.5), condition (5.10) turns out

$$(5.12)\quad \xi_k = \delta I_k\alpha - F'(x_k)^{-1}\delta F_k + O(\zeta^2) = - F'(x_k)^{-1}\{F'(x_k)[\Delta x_k - \delta I_k x_k] + F_d'(x_k)\Delta d_k\} + O(\zeta^2)$$

which is equivalent to (5.11) with

$$\tilde{\Delta}x_k = \Delta x_k - \delta I_k x_k,\quad \|\tilde{\Delta}x_k\| \le \zeta(K_x+C_1)\|x_k\|,$$
$$\tilde{\Delta}d_k = \Delta d_k \qquad\qquad ,\quad \|\tilde{\Delta}d_k\| \le \zeta K_d \|d\|.$$

Due to Theorem 4.1 this means that the Newton method is well-behaved. To prove (5.9) it is enough to observe that

$$F(x_{k+1} + \Delta x_k - \delta I_k x_k;\ d + \Delta d_k) = F(x_k) + F'(x_k)(x_{k+1} - x_k + \Delta x_k - \delta I_k x_k) +$$
$$+ F_d'(x_k)\Delta d_k + O(\zeta^2) = F'(x_k)(\xi_k + \Delta x_k - \delta I_k x_k) + F_d'(x_k)\Delta d_k + O(\zeta^2) = O(\zeta^2)$$

which completes the proof. ■

A crucial point of the well-behavior of Newton iteration is assumption (5.4), i.e., how accurate the values of F can be computed. The accuracy of the evaluation of F' and the solution of the linear system (5.2) is not so important as long as (5.6) and (5.7) hold. To illustrate this point we assume that one wants to approximate α with high relative precision for an ill-conditioned problem, say, $\mathrm{cond}(F;d) \in \left[\frac{1}{\sqrt{\zeta}}, \frac{1}{\zeta}\right)$. From (5.9) it follows

$$\frac{\|x_{k+1}-\alpha\|}{\|\alpha\|} \le \zeta(K_x+C_1) + \mathrm{cond}(F;d)\frac{\|\Delta d_k\|}{\|d\|} + O(\zeta^2).$$

If $\|\Delta d_k\| \le K_d\zeta^2\|d\|$ then x_{k+1} is almost the best possible approximation of α in fl arithmetic. The last assumption holds if we use double precision for the function evaluations. Thus, to compute α with high relative precision for an ill-conditioned problem by Newton iteration it is sufficient to use double precision for the evaluation of F and single precision for the evaluation of F' and for the solution of the linear system (5.2).

6. SECANT ITERATION IN THE SCALAR CASE

In this section we deal with two variants of secant iteration in the scalar case. Let x_k and y_k be two different sufficiently close approximations of a simple zero α of a "smooth" scalar function F, i.e., $F(\alpha;d) = 0$ and $N = 1$. The next approximation is given by

$$(6.1) \quad \overset{*}{x}_{k+1} = x_k - \frac{x_k - y_k}{F(x_k) - F(y_k)} F(x_k).$$

Then

$$\overset{*}{x}_{k+1} - \alpha = O((x_k - \alpha)(y_k - \alpha)).$$

We shall assume that y_k is equal to $x_k + \gamma_k F(x_k)$ where $\{\gamma_k\}$ is a bounded sequence or y_k is equal to x_{k-1}. If $F(y_k)$ requires a new function evaluation then it is a two-point secant iteration. For instance, if $\gamma_k \equiv 1$ and $y_k = x_k + F(x_k)$ this variant of secant iteration is often called Steffensen iteration. If $y_k = x_{k-1}$, then it is secant iteration with memory (see Traub (1964)).

In fl arithmetic due to unavoidable rounding errors it could happen that the computed x_{k+1} is a worse apprxoimation (or even not well-defined) than x_k and y_k. Therefore we slightly modify (6.1) as follows.

<u>Algorithm</u>

(i) Let x_0 and y_0 be sufficient close approximations of α, $k = -1$;

(ii) CON: $k := k+1$;

$$z_k = fl\left(x_k - \frac{x_k - y_k}{F(x_k) - F(y_k)} F(x_k)\right);$$

(6.2) <u>if</u> $|fl(F(z_k))| < |fl(F(x_k))|$ then $x_{k+1} = z_k$, $y_{k+1} = x_{k+1} + \gamma_{k+1} F(x_{k+1})$ or x_k, go to CON;

(6.3) if $|fl(F(z_k)| \geq |fl(F(x_k))|$ and $y_k = x_{k-1}$ then $x_{k+1} = x_k$ and $y_{k+1} = x_{k+1} + \gamma_{k+1} F(x_{k+1})$, go to CON;

(6.4) if $|fl(F(z_k))| \geq |fl(F(x_k))|$ and $y_k = x_k + \gamma_k F(x_k)$ then go to END;

END: $x_{k+j} = x_k$ for all j. ∎

This means that if $|fl(F(z_k))| \geq |fl(F(x_k))|$ and $y_k = x_k + \gamma_k F(x_k)$ we terminate the iteration and formally set $x_{k+j} = x_k$. If the latter inequality holds and $y_k = x_{k-1}$ then we locally switch to a two-point secant iteration setting $y_{k+1} = x_k + \gamma_{k+1} F(x_k)$ and $x_{k+1} = x_k$. Note that in any case the computed sequence $\{|fl(F(x_d))|\}$ is non-increasing.

Let us assume a well-behaved algorithm for the computation of F, i.e.,

(6.5) $fl(F(x;d)) = (1+\Delta F)\, F(x+\Delta x; d+\Delta d) = F(x) + \delta Fx$

where $|\Delta F| \le \zeta K_F$, $|\Delta x| \le \zeta K_x |x|$, $\|\Delta d\| \le \zeta K_d \|d\|$ and δFx is given by (5.5). If z_k is well-defined then

$$(6.6)\quad z_k = (1+\eta_k)\left(x_k - \frac{(x_k-y_k)(F(x_k)+\delta Fx_k)}{F(x_k)-F(y_k)+\delta Fx_k-\delta Fy_k}(1+\epsilon_k)\right)$$

where $|\eta_k| \le \zeta$ and $|\epsilon_k| \le 3\zeta + O(\zeta^2)$.

<u>Theorem 6.1</u>

If there exists a positive constant Q independent of F such that

$$(6.7)\quad \left|\frac{F(x_k)}{F(x_k)-F(y_k)}\right| \le Q$$

for all k under consideration then secant iteration is well-behaved. ■

<u>Proof</u>

Let $q_k = \dfrac{\delta Fx_k - \delta Fy_k}{F(x_k)-F(y_k)}$ and let $Q_k = -\dfrac{q_k}{1+q_k}$. Suppose for now that $|q_k| \ge \frac{1}{2}$. Since $\delta Fx_k - \delta Fy_k = O(\zeta)$ then $y_k - x_k = O(\zeta)$. Due to (6.7) and (5.5) we get

$$|F(x_k)| \le 2Q|\delta Fx_k - \delta Fy_k| \le 4Q\zeta(K_F|F(x_k)| + K_x|x_k|\ |F_x'(x_k)| + K_d\|d\| \cdot \|F_d'(x_k)\|) + O(\zeta^2),$$

and

$$(6.8)\quad |x_k-\alpha| \le 4Q\zeta(K_x|\alpha| + K_d\|d\|\ \|F'(\alpha)^{-1}F_d'(\alpha)\|) + O(\zeta^2).$$

Since $|fl(F(x_{k+j}))| \le |fl(F(x_k))|$ for all $j \ge 0$, then $|F(x_{k+j})| \le |F(x_k)| + |\delta Fx_k| + |\delta Fx_{k+j}|$ which yields

$$|x_{k+j}-\alpha| \le 2\zeta(2Q+1)(K_x|\alpha| + K_d\|d\|\ \|F'(\alpha)^{-1}F_d'(\alpha)\|) + O(\zeta^2).$$

This means numerical stability and due to Lemma 4.1 also well-behavior of secant iteration.

Thus, without loss of generality we can assume that $|q_k| \le \frac{1}{2}$ for all k. This implies that $|Q_k| \le 1$. Now z_k is well-defined and we can rewrite (6.6) as follows.

$$(6.9)\quad z_k = (1+\eta_k)\left(x_k - \frac{(x_k-y_k)(F(x_k)+\delta Fx_k)}{F(x_k)-F(y_k)}(1+Q_k)(1+\epsilon_k)\right) = x_{k+1}^* + \xi_k$$

where

$$\xi_k = \eta_d\alpha - F'(x_k)^{-1}\{\delta Fx_k(1+\kappa_k) + \frac{F(x_k)}{F(x_k)-F(y_k)}(\delta Fy_k - \delta Fx_k)(1+\beta_k)\} + O(\zeta(x_k-\alpha))$$

for

$$1 + \varkappa_k = (1+\eta_k)(1+O(y_k-x_k))(1+Q_k)(1+\epsilon_k),$$

$$1 + \beta_k = (1+\eta_k)(1+O(y_k-x_k))(1+\epsilon_k)/(1+q_k).$$

From (5.5), (6.6) and (6.7) we have

$$(6.10)\quad |\xi_k| \le \zeta\{(1 + 2(2Q+1)K_x)|\alpha| + 2(2Q+1)K_d\|F'(\alpha)^{-1}F_d'(\alpha)\|\,\|d\|\,\} + O(\zeta|y_k-\alpha| + \zeta|x_k-\alpha| + \zeta^2).$$

Suppose for a moment that (6.4) holds, i.e., $|fl(Fz_k)| \ge |fl(F(x_k)|$ and $y_k-\alpha = O(x_k-\alpha)$. Then, it is easy to verify that

$$(6.11)\quad |x_k-\alpha| \le \zeta\{(1+4(Q+1)K_x)|\alpha| + 4(Q+1)K_d\|F'(\alpha)^{-1}F_d'(\alpha)\|\,\|d\|\,\} + O(\zeta^2).$$

Since $x_{k+j} \equiv x_k$, then (6.11) means well-behavior of secant iteration. Note that if (6.3) holds then we perform one iterative step using x_k and $x_k+\gamma_{k+1}F(x_k)$ as two approximations of α and at next iterative step we can pass to (6.4). Thus, without loss of generality, let $|fl(F(z_k))| < |fl(F(x_k))|$ for all k. This implies that

$$x_{k+1} = x_{k+1}^* + \xi_k.$$

Since $\xi_k = O(\zeta)$ then (4.8) holds for small ζ and it is straightforward to verify that (6.10) is equivalent to (4.9). Then from Theorem 4.1 and next from Lemma 4.1 follows well-behavior of secant iteration which completes the proof. ∎

We discuss assumption (6.7) for different values of y_k.

<u>Case I</u>. Let $y_k = x_k + \gamma_k F(x_k)$. This is a two-point secant method. Note that (6.7) is now equal to

$$\frac{F(x_k)}{F(x_k)-F(y_k)} \cong \frac{1}{\gamma_k F'(\alpha)}.$$

Since the lefthand side requires a bound by a constant Q which is independent of F then γ_k has to approximate $F'(\alpha)^{-1}$. It means that in Steffensen iteration with $\gamma_k \equiv 1$, (6.7) does not hold in general. This can cause instability. To prove instability of Steffensen iteration we consider the problem $cF(x;d) = 0$ where c is a small positive constant and $\alpha = 1$. The condition number of cF with respect to the data vector d does not depend on c although $cF'(\alpha)$ tends to zero as c tends to zero. In fl arithmetic

$$y_k = fl(x_k + cF(x_k)) = x_k$$

whenever $x_k \cong 1$ and $|cF(x_k)| \cong \frac{1}{2}\zeta$. Thus, the next Steffensen step is not well-defined and we can have only an approximation x_k such that $x_k-\alpha \cong \frac{1}{2c}F'(\alpha)^{-1}\zeta$. Hence, even for very well-conditioned problems Steffensen iteration can produce extremely bad approximations of α which means instability of this iteration. Numerical tests on a PDP-10 computer confirm this. However, if $\gamma_k \cong F'(\alpha)^{-1}$ then (6.7) holds and this variant of secant iteration is well-behaved. Moreover, if $\lim_k \gamma_k = -F'(\alpha)^{-1}$ then the iteration has order greater than two. Specifically, if $\gamma_k = -F'(x_k)^{-1}$ the order is equal to three while if $\gamma_k = -\gamma_{k-1}F(x_{k-1})/(F(x_{k-1} + \gamma_{k-1}F(x_{k-1})) - F(x_{k-1}))$ then the order is equal to $1 + \sqrt{2}$. (See Traub (1964), pp. 185-187.)

<u>Case II</u>. Let $y_k = x_{k-1}$. This is the secant iteration with memory. Now (6.7) becomes

$$\frac{F(x_k)}{F(x_k)-F(y_k)} \cong \frac{x_k-\alpha}{(x_{k-1}-\alpha)(1-\frac{x_k-\alpha}{x_{k-1}-\alpha})} = O((x_{k-2}-\alpha)) + O(\zeta/(x_{k-1}-\alpha)).$$

Note that at least for some initial steps $|x_{k-1}-\alpha| \gg \zeta$ and (6.7) holds. If not we can modify y_k as follows.

$$(6.12)\quad y_k = \begin{cases} x_{k-1} & \text{if } |F(x_k)/(F(x_k)-F(x_{k-1}))| \le Q \\ x_k + \gamma_k F(x_k) & \text{otherwise} \end{cases}$$

where γ_k ought to approximate $F'(\alpha)^{-1}$ and $Q \ge 2$, say.

Summarizing, modified Steffensen iteration and secant iteration with memory defined by (6.12) are well-behaved.

Numerical stability of the multivariate secant method is considered by Jankowska (1974). This method is stable under some assumptions on a suitable distance and position of successive approximations.

ACKNOWLEDGMENT

I am indebted to A. Kielbasinski, H. T. Kung and J. F. Traub for their comments and help during the preparation of this paper.

REFERENCES

J. Jankowska (1974), "Numerical Analysis of Multivariate Secant Method," in progress.

A. Kielbasinski (1974), "Basic Concepts of Numerical Analysis in Linear Algebra," <u>Mat. Stosowana</u> 4, 1974 (in Polish).

W. Kahan (1971), "A Survey of Error Analysis," <u>IFIP Congress 1971</u>, I, pp. 220-206.

J. M. Ortega and W. C. Rheinboldt (1970), <u>Iterative Solutions of Nonlinear Equations in Several Variables</u>, Academic Press, New York, 1970.

G. W. Stewart (1973), <u>Introduction to Matrix Computations</u>, Academic Press, New York, 1973.

J. F. Traub (1964), <u>Iterative Methods for the Solution of Equations</u>, Prentice-Hall, Englewood Cliffs, New Jersey, 1964.

J. F. Traub (1973), An Introduction to Some Current Research in Numerical Computational Complexity, Computer Science Department report, Carnegie-Mellon University, Pittsburgh, Pa., 1973.

J. F. Traub (1974), Theory of Optimal Algorithms, Computer Science Department report, Carnegie-Mellon

University, Pittsburgh, Pa., Software for Numerical Mathematics, edited by D. J. Evans, Academic Press, New York, 1974, pp. 1-13.

J. H. Wilkinson (1963), Rounding Errors in Algebraic Processes, Prentice-Hall, Englewood Cliffs, New Jersey, 1963.

H. Wozniakowski (1974), "Maximal Stationary Iterative Methods for the Solution of Operator Equations," SIAM J. Numer. Anal., Vol. 11, No. 5, October 1974. (Also available as a Computer Science Department report, Carnegie-Mellon University, Pittsburgh, Pa., 1973.)

H. Wozniakowski (1975a), "Generalized Information and Maximal Order of Iteration for Operator Equations," SIAM J. Numer. Anal. Vol. 12, No. 1, March 1975. (Also available as a Computer Science Department report, Carnegie-Mellon University, Pittsburgh, Pa., 1974.)

H. Wozniakowski (1975b), Numerical Stability of the Chebyshev Method for the Solution of Large Linear Systems, in progress.

H. Wozniakowski (1975c), Numerical Stability of the Successive Approximation Method for the Solution of Large Linear and Nonlinear Equations, in progress.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
NUMERICAL STABILITY FOR SOLVING NONLINEAR EQUATIONS

5. TYPE OF REPORT & PERIOD COVERED
Interim

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
H. Wozniakowski

8. CONTRACT OR GRANT NUMBER(s)
N00014-67-0314-0010
NR 044-422

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Carnegie-Mellon University
Department of Computer Science
Pittsburgh PA 15213

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
Office of Naval Research
Arlington VA 22217

12. REPORT DATE
February 1975

13. NUMBER OF PAGES
22

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report)
UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for Public Release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
The concepts of the condition number, numerical stability and well-behavior for solving systems of nonlinear equations $F(x) = 0$ are introduced. Necessary and sufficient conditions for numerical stability and well-behavior of a stationary iteration are given. We prove numerical stability and well-behavior of the Newton iteration for solving systems of equations and of some variants of secant iteration for solving a single equation under a natural assumption on the computed evaluation of F. Furthermore we show that the Steffensen iteration is unstable and show how to modify it to have well-behavior and hence stability.

DD FORM 1473 1 JAN 73 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)