THE SOLUTION OF NONLINEAR OPERATOR EQUATIONS BY A-STABLE INTEGRATION TECHNIQUES*

by

Paul Thomas Boggs

Technical Report
No. 70-72
June 1970

A Thesis

Department of Computer Science
Cornell University
Ithaca, New York 14850

*Research supported in part
by NSF Grant GJ-844

THE SOLUTION OF NONLINEAR OPERATOR EQUATIONS BY A-STABLE INTEGRATION TECHNIQUES

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University for the Degree of

Doctor of Philosophy

by

Paul Thomas Boggs

June, 1970

The author was born on April 14, 1944, in Columbus, Ohio. He attended the University of Akron where he was elected to Phi Sigma Alpha liberal arts honorary and received his B.S. degree cum laude in 1966 with a major in mathematics. In September, 1966, he entered Cornell University in the Computer Science Department. While at Cornell, the author was elected to Phi Kappa Phi honor society and was a member of the Association for Computing Machinery and the Society for Industrial and Applied Mathematics. The author is married and has one child.

Acknowledgements

The author would like to thank Professors J. Hartmanis and S. Levin for serving as members of his special committee. He would also like to thank Professor K.M. Brown for chairing his special committee for three years. The author is particularly indebted to Professor J.E. Dennis, Jr. for directing his theses and for chairing his special committee during the author's final year. Professor Dennis has always given freely of his time and abilities and was especially helpful with the convergence analysis of Chapter III dealing with Broyden's method. Professor Dennis is also responsible for introducing the author to the concept of A-stability.

Special acknowledgements are due to Professors R.J. Walker and R.A. Sweet for several stimulating discussions. A suggestion of Professor Walker's prompted the author to develop the improved A-stability results of Chapter III.

The author would like to thank the National Science Foundation for supporting him during his final year on an N.S.F. Traineeship and for sponsoring part of his research under Grant GJJ-844. Finally, thanks are due to Mrs. Camille Taylor for her patience in typing the manuscript.

To Judy

Table of Contents

CHAPTER I

Introduction

1. The Problem

The nonlinear system of equations $F(x)=0$ is commonly solved by using one of several standard iteration procedures beginning from an initial approximation x_0. In order for these methods to be effective, however, x_0 must in general be "close" to the root x^*. In this thesis, we consider initial approximations x_0 which are not sufficiently close to x^* for these algorithms to be effective.

The basis for this study is the rather old idea of "solution by continuation" and the major effort is towards the development of practical techniques for exploiting this idea. Before we describe the present work, we develop the historical context into which it fits. Any history which is viewed from such close range, however, must of course be somewhat subjective and any selection of major papers somewhat risky, but the papers mentioned here are often referenced and have had a profound influence on the present work.

The process of solution by continuation can be conveniently described as the replacement of the existing problem by a sequence of problems of which the first problem in the sequence is easily solved and each succeeding problem can be solved by knowing the solution to the previous problem. The last problem in the sequence is then the original equation. The sequence may be infinite in which case we never really return to the original equation, but only to an arbitrarily good approximation to it. There are clearly many ways of carrying out the above process, but the most common way is by imbedding a parameter into the equation and then choosing a sequence of values of that parameter so that the resulting sequence of problems satisfies the above criteria. One example of this technique is to define the imbedding function

$$H(t,x) = F(x)-(1-t)F(x_0) \tag{1.1}$$

and to vary t from zero to one. For another example, consider the case where $F(x) = Q(x)+Ax+C$ where Q is a nonlinear function, A is a linear function and C is a vector constant. Then $H(t,x) = tQ(t)+Ax+C$ for t varying from zero to one will again accomplish the objective since the linear equation $Ax = -C$ can be easily solved. For some practical discussion of these parameter imbeddings and some computational results, see, e.g., Freudenstein and Roth [19], Deist and Sefor [12], Kizner [25], and Broyden [5].

Since we do not wish to assume any structure on the function F, the type of imbedding we consider is of the form of (1.1) and is such that for the first value of t, say t_o, $H(t_o,x) = 0$ has the solution x_o.

The study of parameter imbedding has generally been attacked from two points of view: the functional analytic point of view and the differential equations point of view. Both of these approaches, of course, seek to answer the same essential questions of existence of a unique value of x for each t and of continuity of the curve x(t). These conditions, existence, uniqueness and continuity, are sufficient for the construction of a sequence of intermediate problems which satisfy the above description of the solution by continuation procedure.

As indicated above, this idea of solution by continuation is rather old. Ficken [18], writing in 1951, traced its origin to Schwarz in 1869 and developed its history to 1951. The interested reader is referred to Ficken for his historical remarks and an extensive bibliography. All of this earlier work was from the functional analytic point of view and Ficken also is of this school. Ficken proves that if the parameter t is in a finite domain, then there need be only a finite sequence of intermediate problems created. Davis [11] (1966), in an excellent piece of work, proved the existence and uniqueness of an x for each t where t is a vector parameter and is imbedded in a general way into the equation. For the special case of t in a real interval, he shows that x(t) is a differentiable arc. Davis is also a good reference for more recent work using the functional analytic approach.

The differential equations approach stems from the observation that if we assume that we can differentiate a function such as (1.1) with respect to the parameter t, we can obtain an initial value problem with initial condition $x(t_o)=x_o$. This observation is generally credited to Davidenko [9] in 1953. Gavurin [20] in 1958 also derived differential equations but from a different point of view. He considers under-relaxed iterative methods of the form $x_{n+1}=x_n+hG(x_n)$ where $h<1$. By taking the limit as $h\to 0$, one obtains what Gavurin calls the "continuous analog" of an iterative method. By analyzing these differential equations, Gavurin is able to obtain existence and uniqueness theorems for zeros of functions. Yakovlev [32] and [33] (1964 and 1965) considers the equations derived by differentiating w.r.t. the parameter and obtains some existence and uniqueness results for several specific forms of parameter imbedding. Meyer [27] (1968) considers general imbeddings and very nicely generalizes and improves the work of Yakovlev. Meyer's approach is to find a scaler differential equation whose solution bounds the solution of the vector differential equation. He also investigates the numerical problem of integrating one of these O.D.E.'s for the parameter t in a finite interval [a,b]. He obtains a maximum step-size, h_{max}, such that if the equation is integrated by Euler's method with a constant step-size of h_{max}, then the approximation to x(b), which is the zero of F, is guaranteed to be within the radius of convergence of Newton's method applied to $F(x)=0$. The same type of analysis is also applied to Runge-Kutta methods. Bosarge [2] (1968) achieved the same numerical results as Meyer and applied the techniques to boundary value problems.

In this thesis, we also use the differential equations approach and obtain some improved theoretical results and some practical algorithms for integrating the differential equation. Chapter II deals with the development and analysis of the differential equation; our inspiration comes from the Gavurin idea of a continuous analog and the majorizing function theory of Kantorovich and Akilov [24] and Dennis [13]. Our goal is to develop the continuous analog of a majorizing function and to use this to improve a certain weakness in the approach and results of Meyer. We also obtain, as an added dividend, an increase in the size of the region of uniqueness of a zero of F.

The construction of practical numerical methods is carried out in Chapter III. We seek integration methods which have certain stability properties deemed necessary in the light of practical experience. The type of stability needed turns out to be the A-stability of Dahlquist [8]. We consider in detail the work of Dahlquist and sharpen his results relative to a wide class of problems of which our O.D.E.'s are members. Finally we generate a class of algorithms based on A-stability and prove convergence for several representative members.

In Chapter IV, we apply some of the algorithms of Chapter III to a set of examples and report the results. In particular, we find counter-examples to Gavurin's claim that asymptotic stability of a solution of the O.D.E. implies numerical stability in the algorithm. We thus show the desirability of A-stability.

In Chapter V, we summarize the work and indicate areas of future research.

2 Conventions and Useful Results

Throughout the thesis, $F(x)$ will be the function for which a zero is sought. We assume $F:D\subset L\to L$ where D is an open convex region of L. L is a Banach space unless otherwise stated. We will always denote the Frechet derivative of $F(x)$ (Jacobian if $L=E^N$) by $J(x)$ and its inverse by $J^{-1}(x)$. When dealing with iterative procedures, an integer subscript, say n, will denote evaluation at the n^{th} iteration point, e.g., $F_n=F(x_n)$. The point $x^*\in D$ will always denote the solution and x_o the initial approximation.

We will use the symbol "$\equiv$" to mean "defined equal to." We define $N(S,r)$ to be the open sphere of center S and radius r, relative to the norm in use, and $\bar{N}(S,r)$ to be the closure of $N(S,r)$. I is always the identity operator.

We now state several results which will be used extensively in the sequel. A reference for their proofs is supplied.

<u>Lemma 1.2.1.</u> Davis [11].

Suppose $x,x'\in D, F\in C^1(D)$ and that

$||J(y)-J(z)|| \leq K||y-z||$ for all $y_1 z \in D$. Then

$||F(x)-F(x')-J(x')|| \leq 1/2K||x-x'||^2$

Lemma 1.2.2.

Kantorovich and Akilov [24]. (Banach's theorem.) Let $A:L \to L$ be a linear operator. If $||A|| \leq B < 1$ then $(I-A)^{-1}$ exists and $||(I-A)^{-1}|| \leq \frac{1}{1-B}$,

Lemma 1.2.3.

Broyden [6]. Let x,y E^N (N-dimentional Enclidean space) with $x^T y=1$. Then

$$||I-xy^T||_2 = ||x||_2 \cdot ||y||_2 .$$

Lemma 1.2.4.

Kantorovich and Akilov [24]. Let $[L \to L]$ denote the space of linear operators from L onto L. Let $R:[x_o,x_o,+\Delta x] \to [L \to L]$ where $(x_o,x_o+\Delta x) \subset L$ is the line segment connecting the points x_o and $x_o+\Delta x$. Suppose there exists a real valued function $\phi(t)$ such that $||R(x)|| \leq \phi(t)$ for all x and t where $||x-x_o|| \subseteq t-t_o$. Then

$$|| \int_{x_o}^{x_1} R(x)dx || \leq \int_{t_o}^{t_1} \phi(t)dt$$

where x_1 is any element satisfying $||x_1-x_o|| \leq t_1-t_o$.

The following theorem is the Kantorovich convergence theorem for Newton's method. Many versions of this theorem exist, but the version we cite here, due to Dennis [14], contains improved error bounds which are shown to be the best obtainable by Tapia [31]. Rall and Tapia [29] obtain error bounds of a different form.

Theorem 1.2.5.

Dennis [14]. Let $\Omega = N(x_o, S)$. Assume F is differentiable on Ω and

(i) $||J_o^{-1} F_o|| \leq \eta_o$

(ii) $||J_o[J(x)-J(y)]|| \leq K||x-y||$ for all $x,y \in \Omega$

(iii) $K\eta_o \equiv h_o \leq 1/2$

(iv) $\Omega_o = \dfrac{1-\sqrt{1-2h_o}}{h_o} \eta_o \leq S$

Then the Newton sequence $x_{n+1} = x_n - J_n^{-1} F_n$ is well-defined, $x_n \in \Omega$ for all n and $x_n \to x^* \in \bar{\Omega}$.

also

$$||x^* - x_n|| \leq 2^{-n}(Kr_o)^{2n} \frac{\eta_o}{h_o}$$

Furthermore x^* is the only zero of F in $N(x_o, r_1) \cap \bar{\Omega}$

where $r_1 = \dfrac{1+\sqrt{1-2h_o}}{h_o} \eta_o$.

CHAPTER II

The Initial Value Problem and D-Majorization

1. Introduction

The purpose of this chapter is to derive the class of differential equations mentioned in the introduction and to explore some of their characteristics. Several authors, e.g., Gavurin [20], Yakovlev [32] and [33], Meyer [27] and Bosarge [2], have examined these O.D.E.'s from various points of view and we first review some of their results and then indicate where some improvement is possible.

We derive the differential equations by the standard technique of imbedding a parameter into the original equation $F(x)=0$ and then differentiating w.r.t. the parameter, cf. Meyer. We examine the O.D.E.'s thus developed by extending the concept of a majorizing function for an iteration scheme to a differential equation. The relationship between the two types of majorization will be clearly brought out. For contrast, we also consider the work of Davis [11] who obtained the same results by a functional analytic argument. Finally, we use the "backed-up" solution curve of a particular O.D.E. to enlarge the region of uniqueness of a zero of F.

2. The Differential Equations

In order to derive the O.D.E., we consider the operator homotopy $H(t,x)=0$ which imbeds the real parameter t into the original equation in such a way that $H(0,x)=0$ has the solution x_0 and $H(t',x)=F(x)$. We recall from Chapter I that the interval of definition of the parameter t may be infinite, i.e., t' may be ∞. For example, we could use

$$H(t,x) = F(x)-e^{-t}F(x_0) \qquad (2.1)$$

so that $H(t,x)\to F(x)$ as $t\to\infty$.

A sufficient condition that the homotopy be effective is that the curve $x(t)$ be continuous. However, in order to obtain the differential equation we seek, we need that $x(t)$ be differentiable. (It will be seen later that this restriction is not overly severe, as the conditions which guarantee it are no stronger than those necessary to show that standard iterative procedures converge.) The essential condition we must assume is that H is a differentiable function of x and that the linear operator $H_x(t,x)$ is non-singular for all values of x and t in some domain. Then, since $H(t,x(t))=0$, we have, by differentiating w.r.t. t and using the chain rule,

$$x'(t) = -H_x^{-1}(t,x(t))H_t(t,x(t)) \qquad (2.2)$$

For the basic existence theorem concerning (2.2), Meyer first assumes that $H_x(x,t)$ is invertible on the entire region $[0,1] \times D$ and then assumes that there exists a bounding function which ensures that the solution of (2.2) will remain in D for all $t \in [0,1]$. (Recall that D is the domain of F.) Davis, on the othere hand, does not seek to construct a differential equation, but only to show that the curve x(t) is continuous. He assumes, however, that H_x exists and is invertible on the set $C_o = \{(t,x) : H(t,x) = 0\}$ from which he concludes that x(t) is continuously differentiable. Also, Davis obtains that if $\Pi : C_o \to [0,1]$, then Π is a homeomorphism of C_o onto an open interval of $[0,1]$.

The connection between these homotopies and standard iterative procedures is illustrated by the well-known example of (2.1). The resulting I.V.P. derived from (2.1) is

$$x' = - J^{-1}(x)F(x), \ x(o)=x_o \qquad (2.3)$$

Integration of this equation by Euler's method with a step-size of one yields Newton's method applied to the equation $F(x)=0$.

As mentioned above, Gavurin also derives differential equations, but his method is the reverse of the process of the preceeding paragraph, i.e., he assumes that an iterative procedure of the form $x_{n+1}=x_n-hG(x_n)$ is Euler's method applied to $x' = G(x)$. Then the solution of the differential equation becomes the "continuous analog" of the iterative procedure. Gavurin presents some existence and uniqueness results, but his hypotheses are much stronger than necessary, e.g., for the Newton's analog (2.3), he assumes that for all $x \in N(x_o, B||F(x_o)||)$, $||J^{-1}(x)|| \leq B$ and that $\frac{d^2F}{dx^2}$ exists and is locally bounded in the same sphere. In the case of (2.3), however, one would expect that the weaker Kantorovich hypotheses for Newton's method, Theorem 1.2.5, would guarantee the existence of a solution of (2.3) which would tend asymptotically to x^*. Meyer and Bosarge attacked the problem but, because of their approach, could only prove the result using the stronger assumption that $h_o < \frac{1}{2}$, cf., Theorem 1.2.5. Since the case $h_o = \frac{1}{2}$ is the only case that allows the Jacobian to be singular at the root, this is a major deficiency in their approach. (For $h_o > 1/2$, it cannot be shown in general that F has a root and thus $h_o \leq 1/2$ is the best we could hope to obtain.) Davis has proved the theorem, but again by functional analytic techniques. Since we intend to make extensive use of the differential equation, we obtain the result by using differential equation theory. To the author's knowledge, this is the first time this theorem has been obtained by the differential equations approach.

3. D-Majorization

In this section, we consider the Gavurin idea of a "continuous analog" and conjecture that if an iteration function of the form $x=x+G(x)$ has a majorizing function $t=t+g(t)$, then the continuous analog of the majorizing function will majorize in some sense the continuous analog of the iteration scheme. If this is true, then we could prove the result claimed at the end of section 2. Indeed, this is true and we now proceed to make precise these ideas and to prove the conjecture. We then explicitly obtain the result for the Newton's analog (2.3).

We first present the definitions of majorization and weak majorization and then extend these ideas to the differential equations. Note on notation: The h_o of Theorem 1.2.5 and the step-size h used throughout the remainder of this section are in no way related and should not be confused.

Definition 2.3.1:

(Kantorovich and Akilov [24]) Let t_0 and t' be non-negative real numbers. Let $g: [t_0, t_0+t'] \to [t_0, t_0+t']$ and $G: \bar{N}(x_0, t') \to L$ and suppose g and G are both C^1 on their respective domains. Then $t=g(t)$ will be said to <u>majorize</u> $x = G(x)$ on $\bar{N}(x_0, t')$ if

$$||G(x_0)-x_0|| \leq g(t_0)-t_0 \qquad (3.1)$$

and $$||G'(x)|| \leq g'(t)$$

whenever $$||x-x_0|| \leq t-t_0 < t'$$

The existence of a majorizing function with a unique fixed point allows one to prove the existence and uniqueness, in some region, of a fixed point of the majorized iteration function. The requirement of differentiability of G and g in Definition 2.3.1 is, however, rather severe and we seek a weaker form of majorization not requiring differentiability of G and g. The sacrifice for omitting this requirement is the loss of uniqueness, but this weakness can, in some cases, be overcome. The following is an example of one of these cases and it motivates Dennis' definition of weak majorization. Suppose one wishes to prove Theorem 1.2.5. One could first show that F has a unique zero in $N(x_0, r_1)$ by finding a majorizing function for $x_{n+1} = x_n - J_0^{-1} F_n$.

Then, by finding a weak majorizing function for the full sequence, $x_{n+1}=x_n-J_n^{-1}F_n$, one could show that Newton's method converges to a zero of F in $N(x_o,r_o)$. The uniqueness of the zero established first gives the result. Here the assumption of differentiability of J(x) was avoided. In the following definition, note that g and G are not even required to be continuous.

Definition 2.3.2: (Dennis [13])

Let t_o, t', g and G be defined as in Definition 2.3.1 except that the differentiability of g and G is not required. Then t=g(t) will be said to weakly majorize x=G(x) if (3.1) holds and

$$||G(G(x)) - G(x)|| \leq g(g(t)) - g(t) \quad (3.2)$$

when $||x-x_o|| \leq t-t_o < t'$ and $||G(x)-x|| \leq g(t)-t$.

These two definitions and their respective theories along with the above example provide the inspiration for the following definition:

Definition 2.3.3:

Let t_o, t', g and G be defined as in Definition 2.3.2. Then g will be said to d-majorize G if for the differential equations with real independent variable s

$$x'(s) = G(x),\ x(o) = x_o \qquad (3.3)$$

$$t'(s) = g(t),\ t(o) = t_o \qquad (3.4)$$

the relation

$$||x(s_2) - x(s_1)|| \leq t(s_2) - t(s_1)$$

holds for all $s_2 < s_1$. The equations (3.3) and (3.4) are assumed to have unique solutions for all $s > o$.

The following theorem gives the relationship between weak majorization and d-majorization for a particularly useful form of a weak majorizing function and it proves part of the conjecture.

<u>Theorem 2.3.4</u>:

Assume the function $r(t,h) = t+hg(t)$ weakly majorizes $R(x,h)=x+hG(x)$ as a function of t for all $h \in (o,c]$ for initial values t_o and x_o respectively. Then g(t) d-majorizes G(x).

<u>Proof</u>

We note first that R(x,h) is Euler's method applied to (3.3) and similarly for r(t,h) and (3.4). Thus we define $x(s)_i$ to be the approximation to x(s) by Euler's method with a step-size s/i, i.e.,

$$x(s)_i = \underbrace{R(R(\ldots(R}_{i\text{ times}}(x_o,s/i),s/i)\ldots)s/i).$$

Also $t(s)_i$ is defined analogously. For $s>c$, the approximation is not defined for some small i's, but for every s, there exists an integer n such that $s/n \leq c$ and the approximation to $x(s/n)$ exists for all $i>o$.

Choose any value of s and an appropriate n so that $s/n<c$. By the definition of weak majorization

$$||x(\tfrac{s}{n})_1 - x_o|| \leq t(\tfrac{s}{n})_1 - t_o$$

and by (3.2)

$$||x(\tfrac{2s}{n})_2 - x(\tfrac{s}{n})_1|| \leq t(\tfrac{2s}{n})_2 - t(\tfrac{s}{n})_1. \quad (3.5)$$

Thus

$$\begin{aligned}||x(\tfrac{2s}{n})_2-x_o|| &\leq ||x(\tfrac{2s}{n})_2-x(\tfrac{s}{n})_1||+||x(\tfrac{s}{n})_1-x_o|| \\ &\leq t(\tfrac{2s}{n})_2-t(\tfrac{s}{n})_1+t(\tfrac{s}{n})_1-t_o \\ &\leq t(\tfrac{2s}{n})_2 - t_o.\end{aligned}$$

Clearly now n iterations yield

$$||x(s)_n-x_o|| \leq t(x)_n-t_o.$$

Choosing $h = \frac{s}{n+1}$ and iterating n+1 times, we obtain

$$||x(s)_{n+1} - x_o|| \leq t(s)_{n+1}-t_o.$$

This procedure can be continued indefinitely and thus

$$\lim_{p\to\infty} ||x(s)_{n+p} - x_o|| \leq \lim_{p\to\infty} (t(s)_{n+p} - t_o)$$

which, since Euler's method is convergent, implies that

$$||x(s)-x_o|| \leq t(s)-t_o. \qquad (3.6)$$

We now show that $||x(s_2)-x(s_1)|| \leq t(s_2)-t(s_1)$ for $s_2>s_1$. Choose any value of s_1. Then by (3.6) $||x(s_1)-x_o||\leq t(s_1)-t_o$. If we can show that equation (3.1) continues to hold for t_o of (3.1) equal to $t(s_1)$ and x_o of (3.1) equal to $x(s_1)$, then the hypotheses of this theorem will hold for these initial conditions and, by repeating the above argument for $s_2>s_1$, the conclusion will follow. To show this, we note that equation (3.5) is equivalent to $||G(x(\frac{s}{n})_1)|| \leq g(t(\frac{s}{n})_1)$. The next iteration implies $||G(x(\frac{2s}{n})_2|| \leq g(t(\frac{2s}{n})_2)$. As above $||G(x(s)_{n+p})||\leq g(t(s)_{n+p})$. Thus in the limit $||G(x(s))||\leq g(t(s))$ and for $s=s_1$, $||G(x(s_1))||\leq g(t(s_1))$. Equation (3.1) is now obviously satisfied w.r.t. $x(s_1)$ and $t(s_1)$ and the theorem is established.

Q.E.D.

The next theorem provides the same type of result as is available for majorizing functions. In this theory, the concept of convergence of an iteration function is replaced by the analogous concept of the existence of a solution of the differential equation which tends asymptotic to a particular value. The idea of uniqueness of a fixed point is replaced by the corresponding idea of asymptotic stability of the constant solution which always has the value of the asymptote. (A solution $\phi(t)$ of the differential equation $x'=f(x,t)$ is said to be stable if for every $\epsilon>o$ there exists a $\delta>o$ such that if any solution $\psi(t)$ of the O.D.E. satisfies $||\phi(o)-\psi(o)||\leq\delta$, then $||\phi(t)-\psi(t)|| \leq \epsilon$ for all $t\geq o$. A solution $\phi(t)$ is said to be asymptotically stable if it is stable and $||\phi(t)-\psi(t)||\to o$ as $t\to\infty$.) We will say that the constant solution $\phi(t)=x^{**}$ is <u>asymptotically stable in the region E</u> if the solution $\psi(t)$ of the O.D.E. satisfying $\psi(o)=x_o$ is such that $\psi(t)\to x^{**}$ as $t\to\infty$ for any $x_o\in E$.

<u>Theorem 2.3.5:</u>

If

i) g d-majorizes G

ii) $t(s)\to t^{**}\in [t_o,t_o+t^1]$ as $s\to\infty$

Then $x(s) \to x^{**} \in \bar{N}(x_o, t')$ as $s \to \infty$. If, in addition

iii) $g \in c^1[t_o, t_o+t']$ and $G \in c^1(\bar{N}(x_o, t'))$

iv) the solution $t(s) \equiv t^{**}$ is asymptotically stable for all $t \in [t_o, t_o+t']$

v) the function $r(t,h) = t+hg(t)$ majorizes $R(x,h)=x+hG(x)$ for $h \in (o,c]$ where c is a positive constant and the initial conditions are t_o and x_o respectively

then $x(t) \equiv x^{**}$ is asymptotically stable for all $x \in \bar{N}(x_o, t')$.

Proof

For the first part, pick any strictly increasing sequence $\{s_i\}$. Then since $t(s) \to t^{**}$, $t(s_k)-t(s_j)$, $k>j$ is a Cauchy sequence. Thus $||x(s_k)-x(s_j)||$ is a Cauchy sequence and the conclusion follows.

For the second part, we have only to show asymptotic stability of x^{**}. To do this, we pick new initial values $\bar{x}_o \in \bar{N}(x_o, t')$ and $\bar{t}_o = t'$ and show that $||\bar{x}(s) - x(s)|| \le \bar{t}(s)-t(s)$ where $\bar{x}(s)$ and $\bar{t}(s)$ are the solutions to (3.3) and (3.4) with initial values $\bar{x}_o$ and $\bar{t}_o$ respectively. Then by letting $s \to \infty$, the conclusion follows. Choose any $s>0$ and an n such that $s/n \le c$. Then, since $r(t,h)$ majorizes $R(x,h)$, we have by Lemma 1.2.4

$$||\bar{x}(\tfrac{s}{n})_1 - x(\tfrac{s}{n})_1|| = || \int_{x_o}^{\bar{x}_o} R'(x, \tfrac{s}{n})dx||$$

$$\leq \int_{t_o}^{\bar{t}_o} ||R'(x, \tfrac{s}{n})||dt$$

$$\leq \int_{t_o}^{\bar{t}_o} r'(t, \tfrac{s}{n})dt$$

$$= \bar{t}\ (\tfrac{s}{n})_1 - t(\tfrac{s}{n})_1 \quad .$$

Assume by way of induction that

$$||\bar{x}(\tfrac{is}{n})_j - x(\tfrac{is}{n})_j|| \leq \bar{t}(\tfrac{is}{n})_j - t(\tfrac{is}{n})_j \tag{3.7}$$

and

$$||\bar{x}(\tfrac{is}{n})_j - x_o|| \leq t' \tag{3.8}$$

hold for $j=1,2,\ldots,k-1$. Then, as above,

$$||\bar{x}(\tfrac{ks}{n})_k - x(\tfrac{ks}{n})_k||=|| \int_{x(\frac{(k-1)s}{n})_{k-1}}^{\bar{x}(\frac{(k-1)s}{n})_{k-1}} R'(x,\tfrac{s}{n})ax||$$

$$\leq \int_{t(\frac{(k-1)s}{n})_{k-1}}^{\bar{t}(\frac{(k-1)s}{n})_{k-1}} r'(t,\tfrac{s}{n})dt$$

$$= \bar{t}(\tfrac{ks}{n})_k - t(\tfrac{ks}{n})_k$$

and

$$||\bar{x}(\tfrac{ks}{n})_k - x_o|| \le || \bar{x}(\tfrac{ks}{n})_k - x(\tfrac{ks}{n})_k||+||x(\tfrac{ks}{n})_k-x_o||$$

$$\le \bar{t}(\tfrac{ks}{n})_k - t_o \le t' \ .$$

Thus (3.7) and (3.8) hold for all j, but we are only concerned with the case j=n. For this case $||\bar{x}(s)_n-x(s)_n|| \le \bar{t}(s)_n-t(s)_n$. Similarly, for any integer $p\geq o$, $||\bar{x}(s)_{n+p}-x(s)_{n+p}|| \le \bar{t}(s)_{n+p}-t(s)_{n+p}$. Then as in Theorem 2.3.4,

$$\lim_{p\to\infty} ||\bar{x}(s)_{n+p}- x(s)_{n+p}|| \le \lim_{p\to\infty} (\bar{t}(s)_{n+p}-t(s)_{n+p})$$

and

$$||\bar{x}(s)-x(s)|| \le \bar{t}(s)-t(s) \tag{3.9}$$

Now, because s was arbitrarily chosen (3.9) holds for all $s>0$ and

$$\lim_{s\to\infty} ||\bar{x}(s)-x(s)|| \le \lim_{s\to\infty} (\bar{t}(s)-t(s)) = t^{**}-t^{**} = 0 .$$

Thus, the asymptotic stability is established.

Q.E.D.

Theorem 2.3.4 and Theorem 2.3.5 thus prove the conjecture stated at the beginning of this section. We now proceed to describe the behavior of a class of O.D.E.'s with quadratic right hand sides of which a particular member will be useful in constructing a d-majorizing function for the Newton O.D.E. (2.3).

Lemma 2.3.6 :

For the differential equation $t' = g(t)$ where $g(t) = at^2 - bt + c$, a,b,c real positive constants such that $a \neq o$ and $b^2 - 4ac \geq 0$, the solution $t(s) \equiv r_o = \frac{b-\alpha}{2a}$ is asymptotically stable in the region $(-\infty, r_1)$ where $r_1 = \frac{b+\alpha}{2a}$ and $\alpha = \sqrt{b^2 - 4ac}$.

Proof

We show that the O.D.E. can be integrated exactly and that for any choice of the initial condition $t_o \in (-\infty, r_1)$, the solution tends to r_o.

Rewrite $g(t)$ as

$$g(t) = a[(t - \tfrac{b}{2a})^2 - (\tfrac{\alpha}{2a})^2]$$

Thus

$$\frac{1}{a} \int \frac{dt}{(t-\frac{b}{2a})^2 - (\frac{\alpha}{2a})^2} = \int ds \qquad (3.10)$$

where s is the independent variable. Integration of (3.10) assuming $\alpha \neq o$ yields

$$\frac{1}{\alpha}\log\frac{t-r_1}{t-r_o} = s+C \tag{3.11}$$

$$\frac{1}{\alpha}\log\frac{r_1-t}{t-r_o} = s+C' \tag{3.11'}$$

where C and C' are constants and the solution depends on the sign of the quantity to which the log is applied. We distinguish three cases for t_o. (The case for $\alpha=o$ will be covered later.)

<u>Case 1</u>. $t_o>r_1$

This corresponds to (3.11). Since $r_1>r_o$, $\frac{t_o-r_1}{t_o-r_o} < 1$ and thus C<0. Solving for t yields

$$t = \frac{r_1-r_oe^{\alpha(s+C)}}{1-e^{\alpha(s+C)}}$$

Thus as $s\to -C$, $t\to\infty$ and t does not behave asymptotically.

<u>Case 2.</u> $r_o<t_o<r_1$

This corresponds to (3.11'). Again we solve for t and obtain

$$t = \frac{r_1-r_oe^{\alpha(s+C)}}{1-e^{\alpha(s+C)}} \tag{3.12}$$

which clearly exists for all s. Using L'Hospital's rule to take the limit of (3.12) as $s\to\infty$, we obtain

$$\lim_{s\to\infty} t = \lim_{s\to\infty} \frac{\alpha r_o e^{\alpha(s+C)}}{\alpha_e \alpha(s+C)} = r_o$$

Case 3. $t_o<r_o$

Again this corresponds to (3.11) but now $\frac{t-r_1}{t-r_o} > 1$

implying $\dot{C}>0$. Thus we have

$$t = \frac{r_1-r_o e^{\alpha(s+C)}}{1 - e^{\alpha(s+C)}}$$

which now exists for all s. As in case 2,

$$\lim_{s\to\infty} t = \lim_{s\to\infty} \frac{-\alpha r_o e^{\alpha(s+C)}}{-\alpha e^{\alpha(s+C)}} = r_o$$

Clearly the asymptotes themsealeves are solutions and we have only to consider the case $\alpha=0$. Here $r_o=r_1= \frac{b}{2a}$ and $g(t) = a(t-r_o)^2$. Again we can integrate exactly to obtain.

$$\frac{1}{a} \int \frac{dt}{(t=r_o)^2} = s+C$$

or

$$\frac{-1}{a(t-r_o)} = s+C$$

and

$$t = \frac{-1}{a(s+C)} + r_o.$$

The cases are then as follows:

For $t_o > r_o$, $C<0$ and $t \to \infty$ as $s \to -C$.

For $t_o > r_o$, $C<0$ and $t \to r_o$ as $s \to \infty$.

Q.E.D.

We are now ready to obtain the result desired for the Newton analog.

<u>Theorem 2.3.7</u> :

Assume the hypotheses of Theorem 1.2.5. Then (a) the initial value problem $x'(s) = -J_o^{-1}F(x)$, $x(o)=x_o$, has a solution $x(s)$ such that $x(s) \to x^*$ as $s \to \infty$, $x^* \in \bar{N}(x_o, r_o)$ and the solution $x(s) \equiv x^*$ of the O.D.E. is asymptotically stable in $N(x_o, r_1)$;

(b) the I.V.P.

$$x'(s) = -J^{-1}(x)F(x), x(o) = x_o \tag{3.13}$$

has a solution $x(s)$ such that $x(s) \to x^*$ as $s \to \infty$.

<u>Proof</u>

The proof will be to use the function $g(t) = 1/2Kt^2 - t + \eta_o$ and to show that Theorem 2.3.5 and Lemma 2.3.6 apply. Then all of the results af part (a) follow. For part (b), we use the functions $-g(t)/g'(t)$ and $-J^{-1}(x)F(x)$ and show that (i) and (ii) of Theorem 2.3.5 are satisfied.

(a) We construct, as usual, the functions $R(x,h)=x-hJ_o^{-1}F(x)$ and $r(t,h) = t+hg(t) = t-hg(t)/g'(o)$ and show that $r(t,h)$ majorizes $R(x,h)$ as a function of t for all $h\in(0,1]$. First $||R(x_o,h)-x_o|| = h||J_o^{-1}F_o|| \leq h\eta_o = r(t_o,h)-t_o$ for $t_o=0$. Now using Lemma 1.2.1

$$\begin{aligned} ||R'(x,h)|| &= ||I-hJ_o^{-1}J(x)|| \\ &= ||I-hJ_o^{-1}(J(x)-J_o)-hI|| \\ &\leq (1-h) + hK||x-x_o|| \\ &\leq (1-h) + hKt = r'(t,h) \end{aligned}$$

whenever $||x-x_o|| \leq t-t_o = t < r_1$. Thus Theorem 2.3.4 applies since majorization implies weak majorization and $g(t)$ d-majorizes $G(x) = -J_o^{-1}F(x)$. The remainder of the conditions needed to apply Theorem 2.3.5 follow from Lemma 2.3.6 and so part (a) is established. Note that this implies that at x^*, $F(x^*) = 0$ and that this zero of F is unique in $N(x_o,r_1)$.

For part (b), we redefine $R(x,h)=x-hJ^{-1}(x)F(x)$ and $r(t,h)=t-hg(t)/g'(t)$ and show that $r(t,h)$ weakly majorizes $R(x,h)$ as a function of t for all $h\in(0,1]$. We have immediately that

$$||R(x_o,h)-x_o||=h||J_o^{-1}F_o||\leq h\eta_o = g(t_o,h)-t_o$$

for $t_o=0$. Assume that $||R(x,h)-x|| \leq r(t,h)-t$ when $||x-x_o|| \leq t-t_o=t<r_o$. For weak majorization we need that $||R(R(x,h),h)-R(x,h)|| \leq r(r(t,h),h) - r(t,h)$. To this end we first bound $||J^{-1}(x)J_o||$ using Lemma 1.2.2.

$$||J_o^{-1}(J_o-J(x))|| \leq K||x-x_o||<1$$

for all $||x-x_o||<1/K \leq r_o$. Thus $[I-J_o^{-1}(J_o-J(x))]^{-1}=J^{-1}(x)J_o$ exists and $||J^{-1}(x)J_o|| \leq [1-K||x-x_o||]^{-1} \leq (1-Kt)^{-1} = -1/g'(t)$ when $||x-x_o|| \leq t<r_o$. Next, assume $||J_o^{-1}F(x)|| \leq g(t)$ when $||R(x,h)-x|| \leq r(t,h)-t<r_o$. Then

$$\begin{aligned}||J_o^{-1}F(R(x,h))|| &\leq ||J_o^{-1}[F(R(x,h))-F(x)-J(x)(R(x,h)-x)]|| \\ &\quad +||J_o^{-1}[F(x)-J(x)hJ^{-1}(x)F(x)]|| \\ &\leq 1/2K||R(x,h)-x||^2 + ||J_o^{-1}F(x)||(1-h) \\ &\leq 1/2K[r(t,h)-t]^2 + g(t)(1-h)\end{aligned}$$

Now, since g is a quadratic,

$$\begin{aligned}g(r(t,h)) &= g(t)+g'(t)[r(t,h)-t]+1/2K[r(t,h)-t]^2 \\ &= g(t)+g'(t)\ (-hg(t)/g'(t)) + 1/2K[r(t,h)-t] \\ &= (1-h)g(t) + 1/2K[r(t,h)-t]^2\end{aligned}$$

Thus $||J_o^{-1}F(R(x,h))|| \leq g(r(t,h))$ when $||J_o^{-1}F(x)|| \leq g(t)$ and $||R(x,h)-x|| \leq r(t,h)-t<r_o$ and finally

$$||J^{-1}(R(x,h))F(R(x,h))|| \leq ||J^{-1}(R(x,h))J_o|| \cdot ||J_o^{-1}F(R(x,h))||$$

$$\leq -g(r(t,h))/g'(r(t,h)).$$

Therefore

$$||R(R(x,h),h)-R(x,h)|| \leq h||J^{-1}(R(x,h))F(R(x,h))||$$
$$\leq hg(r(t,h))/g'(r(t,h))$$
$$= r(r(t,h),h)-r(t,h)$$

when $||x-x_o|| \leq t<r_o$ and $||R(x,h)-x|| \leq r(t,h)-t$ and weak majorization is established. Theorem 2.3.4 thus implies that $-g(t)/g'(t)$ d-majorizes $-J^{-1}(x)F(x)$. Now the solution $t(s)$ of $t'(s)=-g(t)/g'(t)$, $t(o)=0$ exists for all $t<1/K$ since the derivative is non-zero for $t\neq 1/K$. Since $t(s)$ also satisfies $g(t(s)) = e^{-s}\eta_o$, $t(s)$ must go to r_o as $s\to\infty$. Thus Theorem 2.3.5 applies and the solution $x(s)$ of (3.13) is such that $x(s)\to x^{**}$. We also have that $x(s)\subset N(x_o,r_o)$ for all s and $F(x,s)) = e^{-s}F(x_o)\to 0$ and, by the uniqueness of the zero of F in $N(x_o,r_1)$, $x^{**} = x^*$.

Q.E.D.

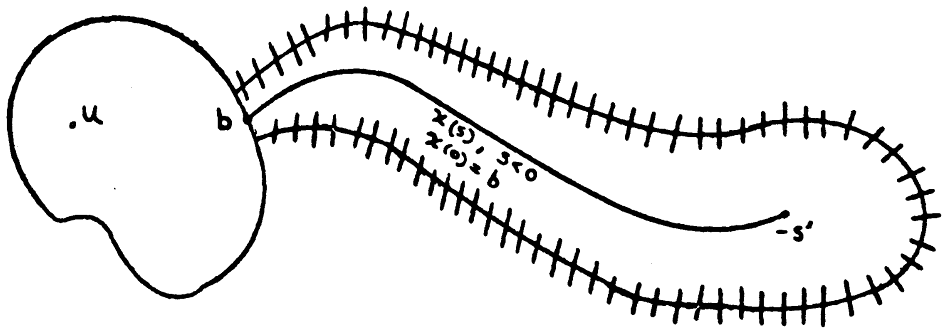
4. The Unicity Region

As indicated above, Theorem 2.3.7 is the strongest Kantorovich type result we could hope to obtain. It provides us with a maximal spherical region of uniqueness of a root of F. For most problems, however, one would anticipate a much larger region of uniqueness though the larger region would not in general be spherical. As a possible method of enlarging the region of uniqueness we consider "backing-up" the solutions of the O.D.E.'s. A particularly practical equation to work with is the Newton analog (2.3) for we again can use the fact that $F(x(s)) = e^{-s}F(x_o)$. Clearly for s decreasing $||F(x(s))||$ is an increasing function of s. Thus, if we can extend the solution outside of the original uniqueness region, we can enlarge the uniqueness region.

In particular, assume we can extend the solution of $x'(s)=-J^{-1}(x)F(x)$ from $s=0$ to $s=-s', s'>0$. Then for any point $\bar{x}$ on the solution curve for $s\in[-s',0]$ we have

$$||F(x)|| \leq ||F(\bar{x})|| - B||x-\bar{x}||$$

where $B = \sup||J(x)||$ and the sup is taken over all points x in an open region R containing the solution curve $x(s)$ for $s\in[-s',0]$. Thus we can construct a sphere of radius $r=||F(\bar{x})||/B$ around every point on $x(s)$. Taking the envelope of all of these spheres, we obtain an "ice-cream" cone in which $F(x)\neq0$ as long as the cone is contained in R. The "ice-cream" comes from the fact that even if the solution fails to exist for $s<-s'$, the sphere can still be constructed. See figure 1.



envelope of spheres around solution curve x(s)

Figure 1

In the above development, we need not make the assumption that if (2.3) were integrated for $s>0$, the solution would tend to x^*. It may happen that the solution would fail to exist due to a singularity of the Jacobian inside the uniqueness region. Thus we can pick any point b on the boundary ∂U of the (enlarged) uniqueness region and, if the Jacobian is non-singular at that point, use it as an initial value for equation (2.3) and integrate backwards to enlarge the region by growing another "cone" out of the one we already have.

We conclude this section with a theorem which gives sufficient conditions for a point not in U to be on the solution curve of an I.V.P. which has its initial value on ∂U.

Theorem 2.4.1:

Let there exist an open region R such that $||J^{-1}(x)|| \leq B$ for all $x \in R$ and $||J(x)-J(y)|| \leq K||x-y||$ and $||J(x)-J(y)|| \geq k||x-y||$ for all $x,y \in F(R)$. Let U be a closed connected region of existence and uniqueness of a solution x^* to $F(x)=0$. Then $\bar{x} \notin U$ lies on a solution curve of

$$x'(s) = -J^{-1}(x)F(x) \tag{4.1}$$

which intersects U if

i) there exists a constant $C>1$ and a point $\tilde{x} \in \partial U$ such that $F(\bar{x}) = CF(\tilde{x})$; and

ii) $\bar{x} \in E = \{x: ||x-\tilde{x}|| \leq (C-1)\, B \cdot ||F(\tilde{x})||\} \subset R$.

Proof

The combination of the normal Lipschitz condition with the nonsingularity of the Jacobian yields the existence and uniqueness of a solution of (4.1) through all $x \in E$. For any $x_o \in E$ and any s such that $x(s)$ is also in E, we have

$$||x(s)-x_o|| = ||\int_o^s -J^{-1}(x(u))F(x(u))du||$$

$$< \int_o^s ||J^{-1}(x(u))|| \cdot ||F(x(u))||du$$

$$\leq B \int_0^s ||F(x(u))||du$$

$$\leq B\left|1-e^{-s}\right| \; ||F(x_0)||$$

Thus for $x_0=\tilde{x}$ and $s=-\log C$, $\bar{x}$ is a candidate for the point $x(-\log C)$. We now show that $\bar{x}$ is the only point in E satisfying (i).

Suppose there is a point $y \in E$ such that $F(y)=F(\bar{x})$. Then $||J(F(y))-J(F(\bar{x}))||=0$ contradicting the hypotheses.

Q.E.D.

CHAPTER III

Numerical Stability and Convergence

1. Introduction.

In this chapter, we turn our attention to the practical numerical problem of integrating one of the differential equations obtained in Chapter II. Before outlining the material contained in this chapter, we make a few observations concerning the properties of an integration scheme applied to one of our problems. In this way, we hope to motivate the approach used and to obtain some further insight into the relationship between the iteration function approach to solving non-linear equations and the differential equations approach.

To make our observations, we consider again the equation

$$x'(t) = -J^{-1}(x)F(x) \tag{1.1}$$

and recall from Chapter II that Euler's method with a step-size of one applied to (1.1) is equivalent to Newton's method applied to the equation $F(x)=0$. We first observe that the convergence of Newton's method does not imply that the iterates are approximating the solution of the initial value problem with any accuracy at all. For example, consider the scaler function $F(x) = \sin x$. Here, when Newton's method converges, the iterates "bounce" from one side of the root to the other while the solution of the I.V.P. always remains on the same side of the root. This exsmple is illustrated in figure 2 where the sine function is plotted in the x,F(x)-plane and the true solution curve is extended in the x,t-plane. The first few Newton iterates are plotted and their values extended into the x,t-plane. The broken line segment is the approximate solution. The type of "bouncing" which occurs here also occurs in higher dimensions and is the cause of instability when it arises while using some higher order numerical itegration techniques.

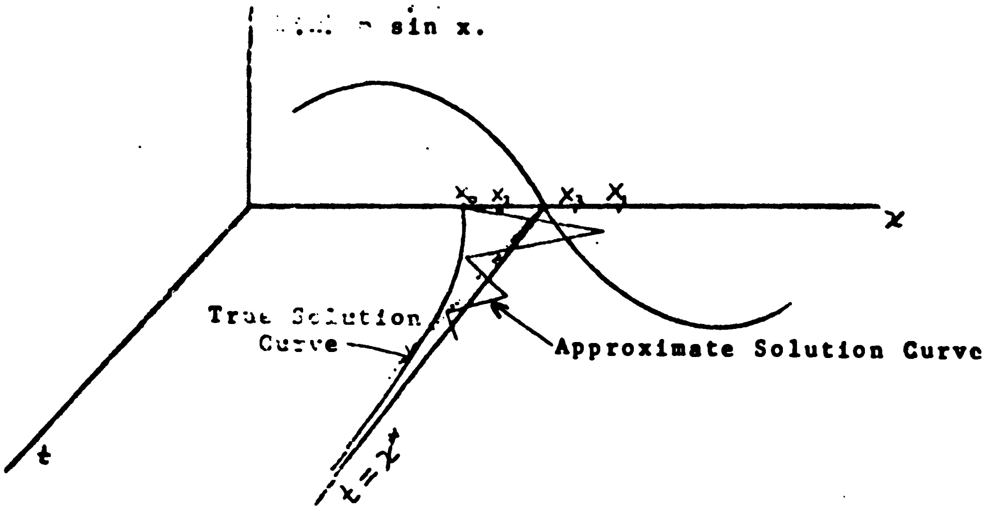


Figure 2

Just as convergence is no indication of accuracy so high accuracy is also no indication of a high order of convergence. In fact, just the opposite is true: roughly speaking, when Newton's method converges, it converges quickly because it makes large errors, but in the "right" direction. The assumption here, however, is that Newton's method does not converge so that the large errors are in the "wrong" direction.

These two observations lead us to conclude that in choosing methods to use on our differential equation, we would like to strike a compromise between accuracy and the rate of convergence while maintaining stability. Rephrasing these requirements gives us a clue as to which methods to use: we seek methods which remain stable in the sense described above, permit relatively large values of the step-size and tend rapidly to the steady state x^*. The problem is now reminicient of equilibrium problems encountered in engineering, physics or chemistry where only the steady state is of interest. These problems are considered in the work of Dahlquist [8].

In section 2, we examine in detail the work of Dahlquist on methods which are stable in roughly the above sense. This concept of stability, which Dahlquist calls

A-stability, will be made precise and examined closely. We first give Dahlquist's results and then, by using a different approach, strengthen in a very important sense the results which he obtained. The conclusions reached here allow us to write the basic formula which we will use for integration of our O.D.E.

In section 3, this formula is examined and, by the use of Broyden's method [3] and [4], modifications are made which significantly reduce the numerical work at each step. Other numerical problems are considered and finally, in section 4, convergence proofs for several standard forms of the algorithm are given.

2. A-Stability

We recall that we are interested in integrating a differential equation whose solution behaves asymptotically and that Dahlquist [8] has considered a stability question relative to this problem. In order to motivate his definition of A-stability, Dahlquist considers the linear O.D.E.

$$y'(t) = qy \tag{2.1}$$

where q is a complex constant with negative real part. The solution of this differential equation for any initial value y_o is given by $e^{qt}y_o$. Clearly $y(t)$ tends to zero as $t\to\infty$. Dahlquist thus desires a method which, when applied to (2.1) with any initial value, also tends to zero as $n\to\infty$. The class of methods over which methods having this property are sought is the class of linear multistep methods. To be precise, we give the following definitions:

<u>Definition 3.2.1</u>:

For the initial value problem $y'=f(t,y), y(o)=y_o$, the <u>general linear k-step method</u> is defined by

$$\sum_{i=o}^{k} \alpha_i y_{n+i} = h \sum_{i=o}^{k} \beta_i f_{n+i} \qquad n=0,1,2,\ldots \quad (2.2)$$

where k is an integer and α_j, β_j $(j=0,1,\ldots,k)$ are real constants independent of n. $f_m=f(t_m,y_m)$ $(m=0,1,\ldots)$ and h is the step-size.

Henrici [22] is a standard reference for the general theory of such methods.

<u>Definition 3.2.2</u>:

A k-step method is said to be <u>A-stable</u> if all solutions of (2.2) tend to zero as $n\to\infty$ when the method is applied to any equation of the form (2.1)

The properties of an A-stable method are characterized by Dahlquist in the following lemma and theorems:

<u>Lemma 3.2.3</u>: (Dahlquist)

For a k-step method, define the polynomials

$$\rho(\zeta) = \sum_{j=o}^{k} \alpha d_j \zeta^j \ , \ \sigma(\zeta) = \sum_{j=o}^{k} \beta_j \zeta^j \ .$$

Then the method is A-stable if and only if $\rho(\zeta)/\sigma(\zeta)$ is regular and has non-negative real part for $|\zeta|>|$.

<u>Theorem 3.2.4:</u> (Dahlquist)

An explicit k-step method cannot be A-stable.

<u>Theorem 3.2.5</u>: (Dahlquist)

The order p of an A-stable linear multistep method cannot exceed 2. The smallest error constant, $c^* = 1/12$, is obtained for the trapezoidal rule

$$y_{n+1} = y_n + \frac{h}{2} [f_{n+1} + f_n] \qquad (2.3)$$

Dahlquist also considers applying an A-stable method to a nonlinear problem and gives his generalization of the A-stable property. In order to make this generalization, Dahlquist first derives some properties that the differential equation should have so that it has a solution which is similar to the zero solution of (2.1), i.e., the equation

$$x'(t) = f(t,x) \tag{2.4}$$

should have a uniformly asymptotically stable (u.a.s) solution. (A solution $\phi(t)$ is said to be uniformly stable if for every $\epsilon>0$ there exists a $\delta>0$ such that if any solution $\psi(t)$ satisfies $||\phi(T)-\psi(T)||\leq\delta$ for any $T\geq 0$ then $||\phi(t)-\psi(t)||\leq\epsilon$ for all $t\geq T$. A solution $\phi(t)$ is said to be uniformly asymptotically stable if it is uniformly stable and $||\phi(t)-\psi(t)||\to 0$ as $t\to\infty$.) Dahlquist rejects the outright assumption of a u.a.s. solution as too general and instead looks for a condition which implies this. He rejects as too restructive the assumption that the origin is u.a.s. for the first approximation to (2.4), i.e., the origin is u.a.s. for

$$y'=A(t)y \tag{2.5}$$

where $y = x-x(t)$ and $A(t) = \left.\frac{\partial f}{\partial x}\right|_{x=x(t)}$. The condition that he chooses is

<u>Condition A</u>. For $y=x-x(t)$, the origin is u.a.s. for the equation

$$y' = f(t,x(t)+y) - f(t,x(t)). \tag{2.6}$$

Dahlquist phrases this requirement in terms of the existence of a quadratic Liopunov function for (2.6) whose total derivative is negative definite on

$$G_\delta = \{(t,x):\ t\in(0,\infty),\ ||x-x(t)||<\delta\}.$$

The origin being u.a.a. for (2.5) implies Condition A which in turn implies the existence of a u.a.s. solution for (2.4). The implications are irreversible.

Next Dahlquist applies the trapezoidal rule to (2.4) and derives a condition on the difference equation similar to Condition A on the differential equation. He obtains the sequence $x_o, x_1, \ldots$ from

$$x_{n+1} - x_n = 1/2h\ [f(t_{n+1}, x_{n+1}) + f(t_n, x_n)] + p_n$$

where $p_o, p_1, \ldots$ are perturbations which, for example, may be due to roundoff error. Assuming the existence of another sequence $x'_o, x'_1, \ldots$ and a corresponding perturbation sequence $p'_o, p'_1, \ldots$ which satisfy

$$x'_{n+1} - x'_n = 1/2h\ [f(t_{n+1}, x'_{n+1}) - f(t_n, x'_n)] + p'_n$$

Dahlquist obtains the difference equation

$$y'_{n+1} - y_n = 1/2h[f(t_{n+1}, x'_{n+1}+y_{n+1}) - f(t_{n+1}, x'_{n+1}) + f(t_n, x'_n+y_n) - f(t_n, x'_n)] + p_n - p'_n \quad (2.7)$$

where $y_n = x_n - x_n'$. If we let $x_n' = x(t_n)$, the true solution, then p_n' becomes the local truncation error and (2.7) becomes the difference analog of (2.6) relative to the trapezoidal rule. Thus Dahlquist assumes that the difference analog to the Liapunov function has the same negative definite property as the Liapunov function for (2.6). This property must hold for all h. Under these assumptions, Dahlquist is able to show that in using the trapezoidal rule with any fixed h on an equation of the form (2.4) which satisfies Condition A, the error incurred remains bounded in an unlimited computation, i.e.,

$$||x(t_n) - x_n|| \leq q^* \qquad (2.8)$$

for all $n>0$ where q^* is the upper bound on the truncation error. This is Dahlquist's generalization of the A-stable property.

This generalization, however, does not seem to be entirely satisfactory. Recall that the zero solution of (2.1) is u.a.s. for all values of y_0 and that the solution of any A-stable method applied to (2.1) tends to zero for any fixed value of h. It would thus seem that the solution of any A-stable method applied to any equation of the form (2.4) which satisfies Condition A would be such that (2.8) holds and $||x(t_n)-x_n||\to 0$ as $n\to\infty$ for any $x_o \in G_\delta$ and for any fixed valve of h. This generalization indeed appears to be attainable, but for the present, we consider only a special class of differential equations of which the equation of interest, (1.1), is a member and obtain the desired generalization for this class. Our methods seem to be more generally applicable, but only preliminary observations on the extension have thus far been made.

Our approach to the generalization is to view the problem as a perturbation problem and to apply the techniques developed by Coddington and Levinson [7] for perturbed linear differential equations to the difference equations. We note first that equation (1.1) can be written as a purturbed linear equation by expanding the right hand side through three terms of a Taylor series to obtain $x' = -x+R(x)$ where $R(x)$ is the remainder term and the origin has been translated to x^*.

Theorem 3.2.7:

For the equation

$$x'(t) = qx + g(x) \tag{2.9}$$

let q be a complex constant with negative real part and g be continuous for small $||x||$. For every $\epsilon>0$, let there exist a $\delta>0$ such that $||g(x)|| \leq \epsilon\ ||x||$ for all $||x|| < \delta$. Let $\{x_n\}$ be generated by the application of an A-stable linear k-step method to (2.9). If $\{x_o, x_1, \ldots, x_{k-1}\} \subset N(0,\delta)$ and h is some fixed positive number then $x_n \to 0$ as $n \to \infty$.

Proof

Let

$$\sum_{i=o}^{k} \alpha_i x_{n+i} = h \sum_{i=o}^{k} \beta_i (qx_{n+i} + g(x_{n+i}))$$

define an A-stable method applied to (2.9). Then we can write

$$\sum_{i=o}^{k} \alpha_i x_{n+i} - hq \sum_{i=o}^{k} \beta_i x_{n+i} = h\gamma_{n+k} \tag{2.10}$$

where $\gamma_{n+k} = \sum_{i=o}^{k} \beta_i g_{n+i}$ with $g_{n+i} = g(x_{n+i})$. Thus we have a non-homogeneous linear difference equation with constant coefficients. It is well-known, cf. Henrici [22, p. 212], that the solution of such equations may be written as $x_n = s_n + z_n$ where s_n denotes the solution to the homogeneous equation ((2.10) with $\gamma_n = 0$ for all n) with the given initial values and z_n is a particular solution of (2.10) with zero initial data. Note that since we have used an A-stable method, $s_n \to 0$ as $n \to \infty$. Following Henrici, we construct a representation for z_n as follows: let the sequence $\{y_i\}$ satisfy

$$y_i = 0 \qquad i<0$$

$$y_o = 1/\alpha_k$$

and for $i>0$, the sequence $\{y_i\}$ satisfies the homogeneous equation. Then by Henrici [22,p.212,Th. 5.2]

$$z_n = \sum_{m=k}^{n} \gamma_m y_{n-m} \tag{2.11}$$

has the desired property.

Now, for $||x_{m+i}|| \leq \delta$, $i=0,1,\ldots,k-1$,

$$||\gamma_m|| = ||h \sum_{i=o}^{k} \beta_i g_{m+i}|| \leq h \;\Sigma|\beta_i|\; ||g(x_{m+i})||$$

$$\leq h \;\Sigma|\beta_i|(\epsilon||x_{m+i}||)$$

$$\leq h\epsilon' \bar{x}_m \tag{2.12}$$

where $\bar{x}_m = \max\limits_{o\leq i\leq k-1} ||x_{m+i}||$ and $\epsilon' = \epsilon \sum\limits_{i=o}^{k} |\beta_i|$. Thus by (2.11) and (2.12)

$$||x_n|| = ||a_n+z_n|| = ||a_n + \sum_{m=k}^{n} \gamma_m y_{n-m}|| \tag{2.13}$$

$$\leq ||a_n|| + \sum_{m=k}^{n} ||\gamma_m|| \cdot |y_{n-m}|$$

$$\leq ||a_n|| + h\epsilon' \sum_{m=k}^{n} \bar{x}_m |y_{n-m}| \tag{2.14}$$

Let $X_N = \max \{||x_n|| : n \leq N\}$ and $S_N = \max \{||a_n|| : n \leq N\}$.

Then by (2.14)

$$X_N \leq S_N + h\epsilon' X_N \sum_{m=k}^{n} |y_{n-m}|$$

or

$$X_N - \frac{S_N}{1-h\epsilon' \sum_{m=k}^{n} |y_{n-m}|} . \tag{2.15}$$

Now if $\sum_{m=1}^{\infty} |y_m|$ is convergent, then h and ϵ can be chosen and fixed such that $1-h\epsilon' \sum_{m=1}^{\infty} |y_m| > 0$ and thus

$||x_n||$ will remain $< \delta$ for all n. (Recall that this is Dahlquist's generalization of the A-stable property.)

To obtain convergence of the series, we first recall from the theory of linear difference equations, that any solution of the homogeneous problem is a linear combination of the members of a fundamental solution set. The typical fundamental solution set is given by Henrici [22, p. 214, Th. 5.3]. By Lemma 3.2.3, all of the roots of the characteristic equation are strictly contained in the unit circle and therefore all of the fundamental solutions eventually decrease at the rate of a geometric sequence. Clearly then, any linear combination of the fundamental solution set must eventually decrease at least as fast as the minimum of these rates. Then, since a geometric series converges, we have established that $||x_n|| \leq \delta$ for all $n>0$.

We now must show that $||x_n|| \to 0$. Let

$$X = \limsup_{n\to\infty} ||x_n||$$

We have $o \leq X \leq \delta < \infty$ and (2.13) holds. For every integer I, let $\bar{I} = \{i: i \geq I\}$. Now for every $\xi>0$, there exists an I_ξ and an infinite set $\tilde{I}_\xi \subset \bar{I}_\xi$ such that $||x_j||>X-\xi$ for all $j \in \tilde{I}_\xi$. We rewrite (2.13) as

$$||x_n|| \leq ||s_n|| + \sum_{m=k}^{j/2} ||\gamma_m|| \cdot |y_{j-m}| + \sum_{m=j/2+1}^{j} ||\gamma_m|| \cdot |y_{j-m}|$$

and thus for all $j \in \tilde{I}_\xi$

$$X-\xi < ||x_j|| \leq ||s_j|| + h\epsilon'\delta \sum_{m=k}^{j/2} |y_{j-m}| + h\epsilon' XK$$

where $K = \sum_{i=0}^{\infty} |y_i| < \infty$ by the above argument. As $j \to \infty$

$\sum_{m=k}^{j/2} |y_{j-m}| \to 0$ since it is the tail end of a convergence series. Therefore

$$X - \xi \leq h\epsilon' XK.$$

But we have that $h\epsilon' K < 1$ and so $X \leq \frac{1}{1-h\epsilon' K} \xi$.

Thus, since ξ was arbitrary, $X = 0$.

Having obtained this result, we would now like to apply it to the case of the trapezoidal rule. First note that since the series $\sum_{m=1}^{\infty} |y_m|$ converges, the series $\sum_{m=1}^{\infty} y_m$ also converges and, as long as $\sum_{m=1}^{\infty} y_m \neq 0$, we have that

$$\sum_{m=0}^{\infty} |y_m| \leq K' \left| \sum_{m=0}^{\infty} y_m \right| \qquad (2.16)$$

where K' is some positive constant. We now apply the trapezoidal rult to (2.9) and obtain the improved result:

<u>Corollary 3.2.8</u>:

If the Trapezoidal rule is used in integrate (2.9) then, for δ sufficiently small, it will generate a sequence $\{x_n\}$ such that $x_n \to 0$ as $n \to \infty$ for any fixed $h > 0$

and for any $x_o \in N(0,\delta)$.

Proof

All we need show is that from (2.13) we can derive a relation similar to (2.15) but independent of h. We have from (2.13) and (2.16)

$$
\begin{aligned}
||x_n|| &\leq ||s_n|| + \sum_{m=k}^{n} \gamma_m y_{n-m}|| \\
&\leq ||s_n|| + \max_{k\leq m\leq n} ||\gamma_m|| \sum_{m=k}^{n} |y_{n-m}| \\
&\leq ||s_n|| + K' \max_{k\leq m\leq n} ||\gamma_m|| \; |\sum_{m=k}^{\infty} y_{n-m}| . \qquad (2.17)
\end{aligned}
$$

The trapezoidal rule applied to the homogeneous problem becomes

$$x_{n+1} - x_n = 1/2hq(x_{n+1} - x_n)$$

or

$$x_{n+1} = \frac{1+1/2hq}{1-1/2hq} x_n \quad .$$

Now for the sequence $\{y_i\}$ as defined in Theorem 3.2.7, we have $y_o = \frac{1}{1-1/2hq}$, $y_{n+1} = \frac{1+1/2hq}{1-1/2hq} y_n$. Thus

$$|\sum_{i=o}^{\infty} y_i| = |\frac{1}{1-1/2hq} \sum_{i=o}^{\infty} (\frac{1+1/2hq}{1-1/2hq})^i|$$

$$= |\frac{1}{1-1/2hq} \qquad \frac{1}{1 - \frac{1+1/2hq}{1-1/2hq}}|$$

$$= \left| -\frac{1}{hq} \right| = \frac{1}{h|q|} \qquad (2.18)$$

Thus by using (2.12), (2.17) and (2.18), we obtain

$$||x_n|| \leq ||s_n|| + K'h\epsilon' \max ||x_m|| \frac{1}{h|q|}$$

$$= ||s_n|| + \frac{K'\epsilon'}{|q|} \max_{k\leq m\leq n} ||x_m|| .$$

Then as in Theorem 3.2.7,

$$X_N \leq \frac{S_N}{1-\frac{K'\epsilon'}{|q|}}$$

and so if δ is chosen small enough, ϵ can then be taken so small that $\frac{K'\epsilon'}{|q|} < 1$. Therefore $X_N < \delta$ for all N>0. The convergence of x_n to zero follows from Theorem 3.2.7.

Q.E.D.

The preceeding discussion gives a very reasonable answer to our question of which method to use to integrate our differential equation. To make the trapezoidal rule even more attractive, we present a result due to Dennis and Sweet [17] concerning the truncation error of a linear multistep method.

<u>Theorem 3.2.9</u>: (Dennis and Sweet)

Let $T(\alpha_o,\ldots,\alpha_k,\beta_o,\ldots,\beta_k)y$ denote the truncation error functional assoicated with a k-step formula. Define $||T(\cdot)|| \equiv ||T||$ to be the minimum number such that

$$|T(\cdot)y| \le ||T||(\int_0^1 (f''(y,t))^2 dt)^{1/2}$$

Then the minimum over all consistent methods of $||T||$ is $1/\sqrt{12}$ and this is achieved only by the trapezoidal rule.

This theorem says that the trapezoidal rule is best in the sense of Sard.

Gear [21] considers methods for integrating the linear equation 2.1 but with the relaxed condition that $Re(q) \le \text{constant} < 0$ where the constant is method dependent. With this easing of the condition on the equation, he is able to obtain linear methods which are stable in a relaxed A-stable sense and have order up to at least six. (Gear, in private communication, has announced development of methods up to order eleven.) He calls these methods "stiffly stable." These methods have the property that for small values of the independent variable t, relatively high accuracy is maintained while for larger values of t, stability is guaranteed. In our case, we are not primarily interested in accuracy and we are always interested in stability so that stiffly stable becomes A-stable.

3. Basic Algorithm and Modifications

In this section, we consider only the initial value problem

$$x'(t) = -J^{-1}(x)F(x), \quad x(o)=x_o \qquad (3.1)$$

and recall that any solution of (3.1) must satisfy

$$F(x(t)) = e^{-t}F(x_o). \qquad (3.2)$$

The basic trapezoidal rule now becomes

$$x_{n+1} = x_n - h/2[J_n^{-1}F_n + J_{n+1}^{-1}F_{n+1}] \qquad (3.3)$$

The trapezoidal rule is now considered from the point of view of actually using it as a numerical scheme. We thus encounter the same problems which arise in using any implicit integration formula, i.e., the problem of predicting x,F and J at t_{n+1} given the values at t_n; and iterating (3.3) to obtain final values of x,F and J. The additional question of using (3.2) as an accuracy test is also considered.

In many nonlinear problems, computing the Jacobian and the inverse Jacobian entail a substantial amount of work and techniques have been devised which seek to eliminate much of this effort. One very successful attack on this problem has been carried out by Broyden and we consider the use of Broyden's method in conjunction with the trapezoidal rule. We digress here to give a brief discussion of Broyden's method.

Broyden's method [3] and [4] is a technique based on an idea due to Davidon [10] for solving nonlinear systems of equations. The technique seeks to avoid the computation of the inverse Jacobian and the essence of the scheme is that the approximation H_{n+1} to the inverse Jacobian at x_{n+1} should satisfy

$$H_{n+1}(F_{n+1}-F_n) = x_{n+1} - x_n \tag{3.4}$$

This method should give a reasonable approximation to the inverse Jacobian at least in the direction $x_{n+1}-x_n$. Broyden's method is to choose

$$H_{n+1} = H_n - \frac{[H_n(F_{n+1}-F_n)-h(x_{n+1}-x_n)](H_n^T H_n F_n)^T}{(H_n^T H_n F_n)^T(F_{n+1}-F_n)}$$

Since this method will be often used in the sequel, we formalize it in the following definition:

Definition 3.3.1:

The function

$$\mathrm{Broy}(F,H_n,h,\Delta x,G)=H_n-\frac{[H_n(G-F_n)-h\Delta x](H_n^T H_n F_n)^T}{(H_n^T H_n F_n)^T(F_{n+1}-F_n)}$$

is called the <u>Broyden update function</u> of the inverse Jacobian of the function F where H_n is the actual or approximate inverse Jacobian at x_n, $F_n = F(x_n)$, h is the step-size $\Delta x_n = x_{n+1} - x_n$ and G is either $F_{n+1} = F(x_n - hH_nF_n)$ or an approximation to it.

The analysis of section 4 requires an estimate for the norm of the difference between the Jacobian and the Broyden approximation to it at the iteration points. Let B_n be the Broyden approximation to the Jacobian at x_n. Then

$$B_{n+1} = B_n - \frac{[F_{n+1} - F_n - h\Delta x]B_n^{-1}F_n}{h(B_n^{-1}F_n)^T B_n^{-1}F_n} \tag{3.5}$$

The following theorem due to Dennis [15] bounds the rate of deterioration of the approximation to the Jacobian. We specialize his result to Broyden's method (the theorem is applicable to the entire class of single-rank update methods) and to the case where $B_o = J_o$.

<u>Theorem 3.3.2:</u> (Dennis)

Let $x_o, x_1, \ldots, x_{n+1}$ be generated by some iterative method and let $B_o, B_1, \ldots, B_{n+1}$ be computed by Broyden's method starting from x_o with $J_o=B_o$. If $J\in Lip_K(D_o)$ where D_o is a convex set such that it is the closure of its interior and it contains $x_o,x_1,\ldots,x_{n+1}$, then

$$||B_{n+1}-J_{n+1}|| \leq 3K/2 \sum_{j=o}^{n} ||x_{j+1}-x_j||.$$

For the purposes at hand, we will use Broyden's method merely as a method of obtaining an approximation to the inverse Jacobian and consider the questions raised concerning the use of (3.3). Thus suppose we wish to compute x_{n+1} from x_n. We first need predicted values of F and J^{-1} at x_{n+1}. We assume that $F_n=F(x_n)$ and $H_n=J^{-1}(x_n)$, or an approximation to it, are available. Then the most natural means of predicting x_{n+1} is to use the explicit method of the next lowest order, i.e., Euler's (Newton's) method with a step-size h. If we let p be the predicted value of x_{n+1}, we have $p=x_n-hH_nF_n$. Having obtained p, we evaluate $F(p)$ and then evaluate $J^{-1}(p)$ by either direct computation or by a Broyden update of H_n.

Another method, however, is available for computing the predicted value of $J^{-1}(p)F(p)$. From (3.2) we can estimate $F(p)$ by $e^{-h}F_n$ and then use Broyden's method to compute $H(p)$ without ever having computed p ,i.e.,

$$H(p)F(p) = \text{Broy}\ (F, H_n, h, -H_nF_n, e^{-h}F_n)e^{-h}F_n$$

$$= \frac{he^{-h}}{1-e^{-h}}\ H_nF_n. \qquad (3.6)$$

This method clearly requires very little computation.

Since accuracy of the computed solution curve is not of prime importance and the limitation of the number of function evaluations is of considerable importance, it is not desirable to iterate the corrector. If, however, for stability reasons, the corrector must be iterated, we again have two choices of how to evaluate J^{-1} and F at the iterates. We clearly must evaluate F and we could then evaluate J^{-1} by either directly computing it or by use of the Broyden approximation using the computed value of F. We note here that the Broyden evaluation requires only an evaluation of F whereas the evaluation of $J^{-1}F$ requires, by finite differences, an additional N evaluations of F.

Having found x_{n+1} to our satisfaction, we now must decide whether to evaluate the function and/or the Jacobian before proceeding to the computation of x_{n+2}. If Broyden's method has been used exclusively, it may be profitable to reevaluate the Jacobian after several steps especially if the correction vectors are not sweeping out a basis often enough. In most cases, however, final evaluations are not needed. Hull and Creemer [23] have shown in practice over a wide class of problems that, in general, final evaluations are neither necessary nor desirable. Hull and Creemer also conclude that, in general, one or at most two iterations of the corrector are sufficient. If, however, the corrector is not iterated and if Broyden's method with $F(p) \simeq e^{-h}F_n$ is used to predict, then a final evaluation is necessary. The reasons for this will be discussed later.

For determination of the step-size h, we use equation (3.2) and form the relative error test

$$||F_{n+1}-e^{-h}F_n|| \leq 10^{-s}\, e^{-h}||F_n|| \qquad (3.7)$$

where s is some integer. This test will,hopefully, keep us reasonably close to the solution curve. Too much accuracy, i.e., s large, will necessitate unreasonably small values of h. Computational experience has shown that values of s of one or zero or even minus one are sufficient in most cases to keep the iteration from blowing up.

One rather interesting choice of the combination of methods described above would be to predict with

Broyden's method using (3.2) to approximate F(p), not to iterate the corrector and not to evaluate the function at the final value of x_{n+1}. (This is the method mentioned above.) With this choice, it would appear that the method would "solve" the equation having only evaluated the function and its derivative at x_o. It is no surprise that this does not work. We can, however, analyze this method to determine to what it converges . Define $G_n = J_n^{-1}F_n$ or the approximation to it. From (3.6) we have that $G_{n+1} = dG_n$ where $d = \frac{he^{-h}}{1-e^{-h}}$. Clearly then $G_{n+1} = d^{n+1}G_o$. (3.3) then becomes $x_{n+1}=x_n-h/2[1+d]G_n = x_n-cG_n$ where $c = h/2[1+d]$. By induction $x_{n+1} = x_o-[c \sum_{i=o}^{n-1} d^i]G_o$. A little elementary calculus shows that $d<1$ for all $h>0$ so that in the limit

$$\lim_{n\to\infty} x_n = x_o - [c \sum_{i=o}^{\infty} d^i]G_o$$

$$= x_o - \frac{c}{1-d}G_o \ .$$

Thus we can define the algorithm

$$x_{n+1} = x_n - \frac{c}{1-d} G_n \qquad (3.8)$$

where G_n here is $J^{-1}(x_n)F(x_n)$ with the x_n generated by

(3.8). This is obviously just an over-relaxed Newton's method. Unfortunately, the function $r(h) = \frac{c(h)}{1-d(h)}$ behaves rather badly. Taking the limit as $h \to 0$, we get $\lim_{n \to 0} r(h) = 2$. For h increasing away from zero, $r(h)$ decreases monotonically to approximately 1.88 at $h = 1.3555$ and thereafter increases monotonically to ∞. A relaxation parameter of as much as 1.88 in many problems would be much too large.

We conclude this section by obtaining a result showing the relationship between s in (3.6) and the step-size h needed to obtain convergence at that accuracy. One would suspect that a step-size which is too small would allow undesirable effects in an accuracy test such as (3.6) since this test, for certain combinations of s and h could allow the norm of F to always increase. We thus obtain the following lemma:

<u>Lemma 3.3.3:</u> Let $s \geq 0$ and assume that h_i, $i=1,2,\ldots$, was chosen so that (3.7) holds for each i. Then, convergence will be guaranteed if $\lim_{i \geq 1} \inf h_i > \log_e(1+10^s)$.

<u>Proof</u>

Assume the ∞-norm throughout the proof. By (3.6)

$$||F_1|| = e^{-h_1}||F_0||(1+\epsilon_1) \text{ where } |\epsilon_1| \le 10^{-8}.$$

Again
$$\begin{aligned} ||F_2|| &= e^{-h_2}||F_1||(1+\epsilon_2) \\ &= e^{-h_2}(e^{-h_1}||F_0||(1+\epsilon_1))(1+\epsilon_2) \\ &= e^{-(h_1+h_2)}||F_0||(1+\epsilon_1)(1+\epsilon_2) \end{aligned}$$

Clearly now

$$\begin{aligned} ||F_r|| &= e^{-\sum_{i=1}^{r} h_i}||F_0|| \prod_{i=1}^{n}(1+\epsilon_i) \\ &= ||F_0|| \prod_{i=1}^{n} \frac{1+\epsilon_i}{e^{h_i}} \\ &\le ||F_0|| \prod_{i=1}^{n} \frac{1+10^{-8}}{e^{h_i}} \end{aligned} \tag{3.9}$$

We may guarantee convergence if all of the terms of the product in (3.9) are eventually strictly less than one, i.e.,

$e^{h_i} > 1+10^{-8}$ for all $i>I$ or

$h_i > \log_e(1+10^{-8})$ for all $i>I$ where I is a non-negative integer.

Q.E.D.

4. Convergence Theory

The techniques of the previous section give rise to a large number of procedures for solving nonlinear systems of equations. An examination of these methods reveals, however, that there are only three basic types of iterations performed depending on the type of evaluation of the inverse Jacobian used. These three types will be defined in conjunction with a basic predictor-corrector algorithm and convergence proofs for them will then be given.

Definition 3.4.1:

The PECE algorithm is defined to be the algorithm based on (3.3) which uses an Euler prediction, an actual evaluation, a correction and a final actual evaluation. The predicted value is designated by p_n:

$$p_n = x_n - hJ_n^{-1}F_n$$
$$x_{n+1} = x_n - h/2[J_n^{-1}F_n + J^{-1}(p_n)F(p_n)].$$

Definition 3.4.2:

The PE_BCE algorithm is the algorithm based on (3.3) defined as in Definition 4.1 except that the first evaluation consists of an actual evaluation of F, but a Broyden approximation to the Jacobian at the predicted value:

$$p_n = x_n - hJ_n^{-1}F_n$$

$$H(p_n) = \mathrm{Broy}(F, J_n^{-1}, h, H_n F_n, F(p_n))$$

$$x_{n+1} = x_n - h/2[J_n^{-1} F_n + H(p_n)F(p_n)]$$

Definition 3.4.3:

The PE_BCE_B algorithm is the algorithm based on (3.3) defined as above except that both of the evaluations consist of actual evaluations of F and Broyden approximations to the Jacobian. Assume that $H_o \equiv J_o^{-1}$:

$$p_n = x_n - hH_nF_n$$

$$H(p_n) = \mathrm{Broy}(F, H_n, h, H_nF_n, F(p_n))$$

$$x_{n+1} = x_n - h/2[H_nF_n + H(p_n)F(p_n)]$$

$$H_{n+1} = \mathrm{Broy}(F, H_n, h, \Delta x_n, F_{n+1})$$

For the following theory, assume $F: D \subset L \to L$ where L is a Hilbert space and D is an open convex region. Assume F is Frechet-differentiable on D and that the ℓ_2 norm is used throughout. For the theorems, the initial point is x_o. First we prove a lemma which isolates a computation needed in the sequel.

<u>Lemma 3.4.4:</u>

Let $v_1, v_2 \in L$ with v_1 fixed. Let $\nu_1 = ||v_1|| \neq 0$ and $\nu_2 = ||v_2|| \leq b < \nu_1$. Define A to be the supplement of the angle between v_1 and v_2 in the usual sense. Choose real positive constants c_1, c_2, and c_3 such that

(i) $2c_1 - \dfrac{c_2}{\nu_1+\nu_2} < 0$

(ii) $1-c_3(\nu_1+\nu_2) > 0$

Then for integer $i \geq 0$, the function

$$f(A,v_2) = \frac{c_1||v_1+v_2||^2 + c_2||v_1-v_2||}{(1-c_3||v_1+v_2||)^i} \tag{4.1}$$

is maximized for $A = \pm 2\Pi j, \ldots j=0,1,\ldots$ and for v_2 such that $\nu_2=0$ or $\nu_2=b$.

<u>Proof</u>

Using the law of cosines which is valid in a Hilbert space, we rewrite (4.1) as

$$f(A,v_2) = \frac{c_1(\nu_1^2+\nu_2^2-2\nu_1\nu_2\cos A)+c_2(\nu_1^2+\nu_2^2+2\nu_1\nu_2\cos A)^{1/2}}{(1-c_3(\nu_1^2+\nu_2^2-2\nu_1\nu_2\cos A)^{1/2})^i} \tag{4.2}$$

Let $N(A)$ and $D^i(A)$ be the numerator and denominator respectively of (4.2). Then to maximize w.r.t. A, we set $f_A(A,v_2) = 0$ and obtain, after simplification,

$$\{D^i(A)[2c_1-c_2/||v_1-v_2||]-N(A)[iD^{i-1}(A)\cdot c_3/||v_1+v_2||]\}$$
$$\cdot \sin A = 0$$

The sin A factor implies $A = \pm 2\Pi j$. $j=0,1,\ldots$ are solutions and all that is now required is to show that these are the only roots. If there are other roots, then, for $i>0$,

$$D^i(A)[2c_1-c_2/||v_1-v_2||]=N(A)[iD^{i-1}(A)\cdot c_3/||v_1+v_2||]. \tag{4.3}$$

We have that $D^i(A)$,$N(A)$ and $iD^{i-1}(A)\cdot c_3/||v_1+v_2||$ are all strictly positive and by (i)

$$2c_1-c_2/||v_1+v_2|| \leq 2c_1-c_2/(\nu_1+\nu_2) < 0$$

implying a contradiction and thus there are no other zeroes for $i>0$. For $i=o$, $D^i(A) = 1$ and the right hand side of (4.3) is zero so that (i) still yields a contradiction. Thus f is maximized w.r.t. A for $A=\pm 2\Pi j$, $j=0,1,\ldots$. Having established this, we can write (4.2) as

$$f(A_M, v_2) = \frac{c_1(\nu_1 \mp \nu_2)^2 + c_2(\nu_1 \pm \nu_2)}{(1 - c_3(\nu_1 \mp \nu_2))^i}$$

$$= \frac{c_1\nu_2^2 + (\mp 2c_1\nu_1 \pm c_2)\nu_2 + c_1\nu_1^2 + c_1\nu_1}{(\pm\, c_3\nu_2 - c_3\nu_1 + 1)^i} \qquad (4.3)$$

where the top sign corresponds to $A_M = 2\Pi j$, j even, and the bottom sign to $A_M = 2\Pi j$, j odd. We now differentiate (4.3) w.r.t. ν_2 and show that in the interval $(o,b]$, the derivative always has the same sign. Let $N(\nu_2)$ be the numerator of (4.3) and $D^i(\nu_2)$ the denominator of (4.3).

Then

$$\frac{d}{d\nu_2}\left(\frac{N(\nu_2)}{D^i(\nu_2)}\right) = \frac{D(\nu_2)N'(\nu_2) - N(\nu_2)\, i\, D'(\nu_2)}{D^{i+1}(\nu_2)} \qquad (4.4)$$

The denominator of (4.4) is always positive so we need only show that the numerator of (4.4) does not change sign. After simplification, the numerator of (4.4) becomes

$$(\mp 2c_1c_3 \pm ic_1c_3)\nu_2^2 + (\pm 2c_1c_3 \pm ic_3c_1)\nu_1^2 - 2ic_1c_3\nu_1\nu_2$$

$$+ (2c_1 - c_2c_3 + ic_2c_3)\nu_2 + (\mp 2c_1 \mp c_2c_3 \pm ic_2c_3)\nu_1 \pm c_2 \quad (4.5)$$

Taking the top sign in (4.5) and using condition (i) yields, for $i \geq 1$,

$$(-2c_1c_3 + ic_1c_3)\nu_2^2 + (2c_1c_3 + ic_1c_3)\nu_1^2 - 2ic_1c_3\nu_1\nu_2$$

$$+ (2c_1 - c_2c_3 + ic_2c_3)\nu_2 + (-2c_1 - c_2c_3 + ic_2c_3)\nu_1 + c_2$$

$$= 2c_1c_3(\nu_1^2 - \nu_2^2) + ic_1c_3(\nu_1 - \nu_2)^2 + c_2c_3(i-1)(\nu_2 + \nu_1)$$

$$+ c_2 - 2c_1(\nu_1 - \nu_2) > 0.$$

Taking the bottom sign in (4.5) and again using (i) with $i \geq 1$ yields

$$(2c_1c_3 - ic_1c_3)\nu_2^2 + (2c_1c_3 - ic_1c_3)\nu_1^2 - 2ic_1c_3\nu_1\nu_2$$

$$+ (2c_1 - c_2c_3 + ic_2c_3)\nu_2 + (2c_1 + c_2c_3 - ic_2c_3)\nu_1 - c_2$$

$$= 2c_1c_3(\nu_2^2 - \nu_1^2) - ic_1c_3(\nu_2 + \nu_1)^2 - c_2 + 2c_1(\nu_2 + \nu_1)$$

$$+ c_2c_3(1-i)(\nu_1 - \nu_2) < 0.$$

For $i = o$, the result is obvious. Therefore the maximization w.r.t. v_2 established and the lemma is proved.

Q.E.D.

<u>Theorem 3.4.5:</u>

Assume the hypotheses of Theorem 1.2.5; call the Lipschitz constant of condition (ii) K_o instead of K. Then the PECE algorithm converges to a unique root $x^* \in \Omega = \bar{N}(x_o, 2\eta_o)$

<u>Proof</u>

From Theorem 1.2.5 we already know that F has a unique root in Ω. Thus we have to show that the algorithm converges and that the iterates remain in Ω. The proof will proceed by induction: we show that from x_o we can compute x_1 and that at x_1, conditions (i),(ii) and (iii) continue to hold with the index increased by one. The general step is clearly the same.

As in Definition 2.4.1 $p = x_o - J_o^{-1} F_o$ and

$$||p-x_o|| = ||J_o^{-1} F_o|| = \eta_o \tag{4.6}$$

From (ii), (iii) and (4.6) we have that

$$||J_o^{-1}(J_o - J(p))|| \le K_o \; ||p-x_o|| \le K_o \eta_o = h_o < 1$$

so from lemma 1.2.2 (Banach's lemma)

$$G_p^{-1} \equiv [I - J_o^{-1}(J_o - J(p))]^{-1} = J(p)^{-1} J_o$$

exists and is bounded in norm by

$$||G_p^{-1}|| \le (1-h_o)^{-1} \tag{4.7}$$

Thus $J_{(p)}^{-1}$ exists and we can compute x_1. Now from (4.7) and lemma 1.2.1,

$$\begin{aligned}
\eta_p &\equiv ||J_{(p)}^{-1}F(p)|| = ||G_p^{-1}J_o^{-1}F(p)|| \\
&\le ||G_p^{-1}||\cdot||J_o^{-1}F(p)|| \\
&\le (1-h_o)^{-1}[\ \ J_o^{-1}F(p)-J_o^{-1}F_o-J_o^{-1}J_o(p-x_o)|| \\
&\qquad + ||J_o^{-1}F_o + J_o^{-1}J_o(p-x_o)||] \\
&\le (1-h_o)^{-1}\ (1/2K_o||p-x_o||^2) \\
&\le (1-h_o)^{-1}\ (1/2K_o\eta_o{}^2) \\
&\le (1-h_o)^{-1}\ 1/2h_o\eta_o \le 1/2\eta_o.
\end{aligned} \tag{4.8}$$

Similarly we can construct a bound for $||J_o^{-1}F_o||$:

$$||J_o^{-1}F_1|| \le ||J_o^{-1}F_1-J_o^{-1}F_o-(x_1-x_o)|| + ||J_o{}^{-1}F_o$$
$$+(x_1\cdot x_o)||$$

$$\le K_o/2\ ||x_1-x_o||^2 + ||J_o^{-1}F_o-1/2J_o^{-1}F_o$$
$$-1/2J_{(p)}^{-1}F(p)||$$

$$\leq K_o/8\ ||J_o^{-1}F_o + J_{(p)}^{-1} F(p)||^2 + 1/2||J_o^{-1}F_o - J_{(p)}^{-1} F(p)||.$$

For $c_1 = K_o/8$, $c_2 = 1/2$ and $i = o$, $2c_1 - \dfrac{c_2}{\eta_o + \eta_p} = \dfrac{K_o\eta_o + K_o\eta_p - 2}{4(\eta_o + \eta_p)}$

$\leq \dfrac{-5/4}{4(\eta_o + \eta_p)} < 0$ and thus by Lemma 3.4.4

$$||J_o^{-1}F_1|| \leq K_o/8(\eta_o \mp \eta_p)^2 + 1/2\ (\eta_o \pm \eta_p).$$

The existence of J_1^{-1} is again established by Banach's lemma and (4.8):

$$||J_o^{-1}(J_o - J_1)|| \leq K_o||x_1 - x_o|| \leq 1/2K_o||J_o^{-1}F_o + J_{(p)}^{-1}F(p)||$$

$$\leq 1/2K_o(\eta_o + \eta_p)$$

$$\leq 3/4h_o < 1 \qquad (4.9)$$

We choose here to write the bound for

$$G_1^{-1} = [I - J_o^{-1}(J_o - J_1)]^{-1} = J_1^{-1}J_o \text{ as}$$

$$||G_1^{-1}| \leq (1 - K/2||J_o^{-1}F_o + J_{(p)}^{-1}\ \ F(p)||)^{-1}$$

We then have

$$\eta_1 \equiv ||J_1^{-1}F_1|| = ||G_1^{-1}J_o^{-1}F_1||$$

$$\leq \frac{K_o/8\,||J_o^{-1}F_o + J(p)^{-1}\ F(p)\,||^2 + 1/2\,||J_o^{-1}F_o - J(p)^{-1}\ F(p)\,||}{1 - K_o/2\ ||J_o^{-1}F_o + J(p)^{-1}\ F(p)\,||}\ .$$

Again for $c_o = K_o/8$, $c_2 = 1/2$, $c_3 = K_o/2$ and $i=1$,

we have $2c_1 - c_2/(\eta_o+\eta_p) < 0$. By (4.9) $1-c_3(\eta_o+\eta_p) > 0$.

Thus Lemma 3.4.4 is again applicable and

$$\eta_1 \leq [K_o/8(\eta_o + \eta_p)^2 + 1/2(\eta_o \pm \eta_p)]/[1-K_o/2(\eta_o \pm \eta_p)]. \tag{4.10}$$

The three cases of (4.10) are then

Case 1. $\eta_p = 1/2\eta_o$ with the upper sign used.

$$\eta_1 \leq [K_o/8(3/2\eta_o)^2 + 1/2(1/2\eta_o)]/[1-K_o/2(3/2\eta_o)]$$

$$= 8/5\ (25/64)\eta_o = 5/8\eta_o$$

Case 2. $\eta_p = 1/2\eta_o$ with the lower sign used.

$$\eta_1 \leq 8/7(49/64)\eta_o = 7/8\eta_o$$

Case 3. $\eta_p = o$.

$$\eta_1 \leq 4/3(9/16)\eta_o = 3/4\eta_o$$

Note: $\eta_p = o$ implies $F(p)=0$, so this case is to be regarded as the limiting case for small $||J^{-1}(p)F(p)||$.

We must now compute the constant K_1 for $||J_1^{-1}(J(x)-J(y))||$ and show that $K_1\eta_1 \leq 1/2$.

$$||J_1^{-1}(J(x)-J(y)|| = ||G_1^{-1}J_o^{-1}(J(x)-J(y))||$$
$$\leq (1-K_o/2||J_o^{-1}F_o+J^{-1}(p)F(p)||)^{-1}$$
$$\cdot K_o||x-y||$$

Defining $K_1 \equiv K_o(1-K_o/2||J_o^{-1}F_o+J^{-1}(p)F(p)||)^{-1}$ we have

$$K_1\eta_1 - \frac{K_o/8||J_o^{-1}F_o+J^{-1}(p)F(p)||^2+1/2||J_o^{-1}F_o-J^{-1}(p)\ F(p)||}{(1-K_o/2\ \ J_o^{-1}F_o+J^{-1}(p)F(p)||)^2} K_o.$$

By Lemma 3.4.4 with $c_1=K_o/8$, $c_2=1/2$, $c_3=K_o/2$ and $i=2$, verified as above,

$$K_1\eta_1 \leq \frac{K_o/8(\eta_o\pm\eta_p)^2+1/2\ (\eta_o\mp\eta_p)}{(1-K_o/2\ (\eta_o\pm\eta_p))^2}\ K_o.$$

Combining all three cases, we have

$$h_1 \equiv K_1\eta_1 \leq \begin{Bmatrix} 8/5(5/8)\eta_o \\ 8/7(7/8)\eta_o \\ 4/3(3/4)\eta_o \end{Bmatrix} K_o \leq K_o\eta_o \leq 1/2 .$$

We now show that the iterates remain in Ω. Taking the three cases in order, we obtain:

Case 1.

$$||x_n-x_o|| \leq \sum_{i=1}^{n} ||x_i-x_{i-1}||$$

$$\leq 3/4 \sum_{i=1}^{n} \eta_{i-1} \leq 3/4\eta_o \sum_{i=1}^{n} (5/8)^i.$$

$$\lim_{n\to\infty} ||x_n-x_o|| \leq 3/4\eta_o \sum_{i=1}^{\infty} (5/8)^i = 3/4\eta_o(8/3) = 2\eta_o$$

Case 2.

$$\lim_{n\to\infty} ||x_n-x_o|| \leq 1/4\eta_o \sum_{i=1}^{\infty} (7/8)^i = 1/4\eta_o(8) = 2\eta_o.$$

Case 3.

$$\lim_{n\to\infty} ||x_n-x_o|| \leq 1/2\eta_o \sum_{i=1}^{\infty} (3/4)^i = 1/2\eta_o(4) = 2\eta_o$$

Note: Since $\eta_p=0$ implies convergence at x_p, case 3 is to be considered as the limiting case as $\eta_p\to 0$.

Finally we show that the iterates converge to x^*. From the three cases of (4.10) we know that $\sum_{i=o}^{\infty} ||x_{i+1}-x_i||$ converges and this implies that the sequence $\{x_n\}$ is a Cauchy sequence. Thus $x_n \to x^{**} \in \bar{\Omega}$. If we can show that $F(x^{**})=0$, they by the uniqueness of the root in $\bar{\Omega}, x^{**}=x^*$.

Now

$$||J_n|| = ||J_o(I+J_o^{-1}(J_n-J_o))||$$

$$\leq ||J_o||[1+K||x_n-x_o||]$$

$$\leq ||J_o||[1+2K\eta_o]$$

$$\leq 2||J_o|| \equiv c < \infty.$$

Rewriting the PECE algorithm, we obtain

$$F_n = -J_n[J^{-1}(p_{n+1})F(p_{n+1}) - 2(x_{n+1}-x_n)]$$

Thus

$$||F_n|| \leq c[\eta_{p_{n+1}} - 2||x_{n+1} - x_n||]$$

and

$$\lim_{n\to\infty} ||F_n|| \leq c \lim_{n\to\infty} [\eta_{p_{n+1}} - 2||x_{n+1}-x_n||] = 0$$

Q.E.D.

For the convergence of the PE_BCE algorithm, we would expect that, since we are only approximating the inverse Jacobian at the predicted value, we could only obtain convergence in a smaller region. But this is not the case and we can prove convergence in essentially the same region. We have, however, restated the hypotheses slightly with the result that the theorem is weaker by a Cauchy inequality.

Theorem 3.4.6:

Assume $F\in C'(D)$

(i) $||J_o^{-1}F_o|| \leq \eta_o$

(ii) $||J(x) - J(y)|| \leq K||x-y||$ for all $x,y\in\Omega$

$= \bar{N}(x_o,2\eta_o) \subset D$

(iii) $||J_o^{-1}|| \leq \beta_o$

(iv) $\beta_o K\eta_o \equiv h_o \leq 1/2$

Then the PE_BCE algorithm converges uniquely to a root $x^*\in\Omega$.

Proof

The idea of the proof is the same as in Theorem 3.4.5 and so only the essential details will be given.

First, for $B_o=J_o$ and $B_p=H_p^{-1}$ and by equation (3.5) and Lemma 1.2.1, we have

$$||J_o^{-1}B_p-I|| \leq \beta_o||B_p-J_o||$$

$$= \beta_o||J_o-(F_p-F_o-J_o(p-x_o))\frac{F_o^T(J_o^{-1})^T}{F_o^T(J_o^{-1})^T J_o^{-1}F_o} -J_o||$$

$$\leq \beta_o(1/2K||J_o^{-1}F_o||^2 \frac{||J_o^{-1}F_o||}{||J_o^{-1}F_o||^2}$$

$$\leq 1/2\beta_o K\eta_o = 1/2h_o<1$$

Thus $G_p^{-1} = [I+J_o^{-1}B_p-I]^{-1} = H_pJ_o$ exists and

$$||G_p^{-1}|| \leq (1-1/2h_o)^{-1} \leq 4/3.$$

as in the previous theorem

$$||J_o^{-1}F_p|| \leq 1/2h_o\eta_o \leq 1/4\eta_o$$

and

$$||H_pF_p|| = ||G_p^{-1}J_o^{-1}F_p|| \leq \frac{1/2h_o\eta_o}{1-1/2h_o} \equiv \eta_p \leq 1/3\eta_o .$$

Again we show the existence of J_1^{-1} by Banach's lemma:

$$||J_o{}^{-1}J_1-I|| \leq \beta_o||J_1-J_o|| \leq \beta_oK||x_1-x_o||$$

$$= 1/2\beta_oK||J_o^{-1}F_o + H_pF_p||$$

$$\leq 1/2\beta_oK||\eta_o+\eta_p|| \leq 2/3h_o<1.$$

Thus

$$G_1^{-1}[I-(I-J_o^{-1}J_1)]^{-1} = J_1^{-1}J_o \text{ exists and}$$

$$||G_1^{-1}|| \le (1-1/2\beta_o K(J_o^{-1}F_o+H_pF_p))^{-1} \qquad (4.11)$$

and

$$||J_1^{-1}F_1|| \le \frac{1/2||J_o^{-1}F_o-H_pF_p||+1/8\beta_o K||J_o^{-1}F_o+H_pF_p||^2}{1-1/2\beta_o K||J_o^{-1}F_o+H_pF_p||}$$

$$\equiv \eta_1.$$

We again have that Lemma 3.4.4 is applicable and

$$\eta_1 \le \frac{1/2(\eta_o \mp \eta_p) + 1/8\beta_o K(\eta_o \pm \eta_p)^2}{1 - 1/2\beta_o K(\eta_o \pm \eta_p)}$$

which has three cases:

Case 1. $\eta_p = 1/3\eta_o$ with the upper sign used.

Then $\eta_1 \le 2/3\eta_o$.

Case 2. $\eta_p = 1/3\eta_o$ with the lower sign used.

Then $\eta_1 \le 5/6\eta_o$.

Case 3. $\eta_p = 0$.

Then $\eta_1 \le 3/4\eta_o$.

To form $h_1 \equiv \beta_1 K\eta_1$, we first have that, by (4.11),

$$||J_1^{-1}|| \le \beta_o(1-1/2\beta_o K(J_o^{-1}F_o+H_pF_p))^{-1} \equiv \beta_1.$$

and then by using Lemma 3.4.4 and combining the three cases, we have

$$h_1 = \beta_1 K\eta_1 \leq \left\{ \begin{array}{l} 3/2\beta_o K(2/3\eta_o) \\ 6/5\beta_o K(5/6\eta_o) \\ 4/3\beta_o K(3/4\eta_o) \end{array} \right\} = h_o \leq 1/2$$

The rest of the proof is exactly as in Theorem 3.4.5.

Q.E.D.

Before stating the convergence theorem for the PE_BCE_B algoritm, we remark that we would only expect the algorithm to converge in a region, relative to h_o, in which Broyden's method convergea. However, due to the complications which arise because of the use of Theorem 3.3.2 we can no longer use Lemms 3.4.4 and we can only prove convergence in a region half as large as expected. Dennis [15] has proved convergence for Broyden's method for $h_o \leq 1/8$, but we can only prove convergence for the PE_BCE_B algorithm for $h_o \leq 1/16$. The essential idea of the proof is the same as in the above two theorems, but the details are somewhat different.

<u>Theorem 3.4.7:</u>

Assume the hypotheses of Theorem 3.4.6 expect that $h_o \leq 1/16$. Then the PE_BCE_B algorithm converges to a unique root $x^* \in \Omega$.

<u>Proof</u>

For the proof of this theorem, we take a slightly different tack and assume that $\sum_{i=o}^{n} ||x_{i+1}-x_i|| + ||p-x_n|| < 2\eta_o$. Then using Lemma 1.2.2 and Theorem 3.3.2

$$||H_oB_p - I|| = ||H_o(B_p-B_o)||$$

$$\leq \beta_o||B_p-J_o||$$

$$\leq \beta_o||B_o-J_p||+||J_p-J_o||$$

$$\leq \beta_o[3/2K(\sum_{i=o}^{n-1}||x_{i+1}-x_i||+||p-x_n||)$$

$$+K||p-x_o||]$$

$$\leq \beta_o(5K\eta_o) = 5h_o<1$$

Thus H_p exists and $||\bar{G}_p^1|| = ||H_pJ_o|| \leq (1-5h_o)^{-1}$.

Now, using Lemma 1.2.1 and Theorem 3.3.2

$$||J_o^{-1}F_p|| \le \beta_o\ [||F_n - J_nH_nF_n|| + ||F_p - F_n - J_n(p-x_n)||]$$

$$\le \beta_o[||B_n - J_n||\eta_n + 1/2K||p-x_n||^2]$$

$$\le \beta_o[3/2K \sum_{i=o}^{n-1} ||x_{i+1}-x_i|| + 1/2K||p-x_n||]\eta_n$$

$$\le 3h_o\eta_n$$

Thus we have

$$||H_pF_p|| \le \frac{3h_o}{1-5h_o}\,\eta_n \equiv \eta_p$$

We need now that $\sum_{i=o}^{n} ||x_{i+1}-x_i|| \le 2\eta_o$, but we defer this proof until later. Assume for the present that it is true. We now show that H_{n+1} exists in the usual way:

$$||H_oB_{n+1} - I|| \le \beta_o||B_{n+1} - B_o||$$

$$\le 5h_o < 1$$

and H_{n+1} exists. Next we compute a bound for $||J_o^{-1}F_{n+1}||$. First note that, by (3.4), $x_n - p = H_p(F_n - F_p)$. Thus $H_nF_n = H_pF_n - H_pF_p$ and

$$H_pF_n = H_nF_n + H_pF_p. \tag{4.12}$$

Using (4.12) and Lemma 1.2.1

$$||J_o^{-1}F_{n+1}|| = ||J_o^{-1}[F_n+J_n(x_{n+1}-x_n)+F_{n+1}-F_n-J_n(x_{n+1}-x_n)]||$$

$$\le ||J_o^{-1}[1/2F_n+1/2F_n-1/2J_n(H_nF_n+H_pF_p)]||$$

$$+ \ \beta_o(1/2K||x_{n+1}-x_n||^2)$$

$$\le 1/2[||J_o^{-1}F_n||+\beta_o||B_p-J_n||\cdot||H_pF_n||$$

$$+ \ \beta_oK||x_{n+1}-x_n||^2] \qquad (4.13)$$

For $||B_p-J_n||$, we have, by Lemma 1.2.1, Lemma 1.2.3, Theorem 3.3.2, and (3.5),

$$||B_p-J_n|| = ||\ B_n-(F_p-F_n+F_n)\ \frac{(H_nF_n)^T}{||H_nF_n||^2} - J_n||$$

$$= ||B_n-J_n - F_p\ \frac{(H_nF_n)^T}{||H_nF_n||^2}\ ||$$

$$= ||B_n-J_n-(F_n-J_nH_nF_n+F_p-F_n+J_nH_nF_n)\frac{(H_nF_n)^T}{||H_nF_n||^2}||$$

$$= ||(B_n-J_n)-[(B_n-J_n)\frac{(H_nF_n)(H_nF_n)^T}{||H_nF_n||2}\ ]$$

$$+ \ (F_p-F_n-J_nH_nF_n)\ \frac{(H_nF_n)^T}{||H_nF_n||^2}\ ||$$

$$= \left\| (B_n - J_n)\left[I - \frac{(H_nF_n)(H_nF_n)^T}{\|H_nF_n\|^2}\right]\right.$$

$$\left. + (F_p - F_n + J_nH_nF_n)\frac{(H_nF_n)^T}{\|H_nF_n\|^2}\right\|$$

$$\leq 3/2K\sum_{i=o}^{n-1}\|x_{i+1}-x_i\| + 1/2K\|H_nF_n\|^2\frac{\|H_nF_n\|}{\|H_nF_n\|^2}$$

$$\leq \ 3/2K\sum_{i=o}^{n-1}\|x_{i+1}-x_i\| + 1/2K\|p-x_n\|$$

$$\leq \ 3K\eta_o \qquad (4.14)$$

Combining (4.13) and (4.14) we obtain

$$\|J_o^{-1}F_{n+1}\| \leq 1/2[\|J_o^{-1}F_n\| + 3h_o(\eta_n+\eta_p)$$

$$+ \beta_o K\|x_{n+1}-x_n\|^2] \qquad (4.15)$$

and finally

$$\|H_{n+1}F_{n+1}\| \leq 1/2(1-5h_o)^{-1}[\|J_o^{-1}F_n\| + 3h_o(\eta_n+\eta_p)$$

$$+\beta_o K\|x_{n+1}-x_n\|^2] \equiv \eta_{n+1} \qquad (4.16)$$

In order to bound η_{n+1}, we first bound $\|J_o^{-1}F_{n+1}\|$. Recalling that the first step of the PE_BCE_B algorithm is the same as the PE_BCE algorithm and that $\|J_o^{-1}F_o\| \leq \eta_o$, we have

$$||J_o^{-1}F_1|| \le 1/[||J_o^{-1}F_o||+(\eta_o+\eta_p)1/2h_o+1/4\beta_o K(\eta_o+\eta_p)^2]\eta_o$$

$$\le 1/2[1+(1+1/31)1/2\cdot 1/16+1/4\cdot 1/16(\tfrac{32}{31})^2]\eta_o$$

$$\le .53\eta_o$$

Thus by (4.16), $\eta_1 \le .77\eta_o$ and $||J_o^{-1}F_1|| \le 11/16\eta_1$. We also clearly have that for $i \ge 1$, $||J_o^{-1}F_i|| \le 11/16\eta_i$. We compute the first several bounds for $||J_o^{-1}F_i||$, however by using closer bounds on $\Sigma||x_{i+1}-x_i||$ then $2\eta_o$, i.e., we bound the sum using only the terms involved. To be precise, we compute $||J_o^{-1}F_j||$ $j=2$ and 3 as $.40\eta_1$ and $.44\eta_2$ respectively. for $j \ge 4$, we use the full bound of $2\eta_o$ and obtain that, using (4.15) and (4.16), $||J_o^{-1}F_j|| \le .47\eta_{j-1}$ and thus $||H_jF_j|| \le .69\eta_{j-1}$.

As in Thoerem 3.4.5, the strictly decreasing sequence of η_i's yields convergence and it remains to show that the assumption $\sum_{i=o}^{\infty}||x_{i+1}-x_i|| \le 2\eta_o$ is valid. We have, for p_n defined to be the prediced value for the (n+1)st step,

$$\sum_{i=o}^{\infty} ||x_{i+1}-x_i|| \leq 1/2\ \Sigma\ ||H_i F_i + H_{p_i} F_{p_i}||$$

$$\leq 1/2\ \Sigma(\eta_i + \eta_{p_i})$$

$$\leq 1/2\ \Sigma 14/11\eta_i$$

$$\leq 7/11 \sum_{i=o}^{\infty} \eta_i$$

$$\leq 7/11\eta_o[1+.77+.77 \times .59$$

$$+ .77\times.59\times.63 \sum_{i=o}^{\infty} (.69)^i]$$

$$\leq 2\eta_o$$

Therefore, since $\eta_{p_n} \leq 3/11\eta_n \doteq .27\eta_n$, the assumption that $\sum_{i=o}^{n-1} ||x_{i+1}-x_i|| + ||p_n - x_n|| \leq 2\eta_o$ is also valid and the theorem is proved.

Q.E.D.

CHAPTER IV

Numerical Results

1. Introduction

Several of the algorithms derived in Chapter III have been programmed and tested on a set of examples. It is the purpose of this chapter to report the results of two examples and to give some tentative conclusions based on the observed behavior. Our study will consist of a detailed analysis of the first problem and a presentation of the results of the second. For the first example, we attempt to solve it by Newton's method, a damped Newton's method, two non-A-stable integration routines and finally by several of the algorithms of Chapter III. The two non-A-stable routines provide counter-examples to the Gavurin claim that numerical stabiltiy is implied by the mathematical stability of the problem. The behavior of the methods of Chapter III on the two examples is rather typical and thus further examples are not reported.

2. The Choice of Algorithms

The algorithms of Chapter III which we use in our tests are the three basic algorithms, PECE, PE_BCE and PE_BCE_B, and the three basic algorithms with a final correction, PECEC, PE_BCEC and PE_BCE_BC. Each of these methods is run with and without the accuracy test of Chapter III, section 3. The damped version of Newton's method is Newton's method in conjunction with that same accuracy test. The two non-A-stable integration techniques are Nordsieck's method [28] (equivalent to the fifth order Adams-Moulton method) and a modified Nordsieck's method. The modification consists of always using Broyden's method for the Jacobian evaluation with $e^{-h}F_n$ as the approximation to F_{n+1}. All programs are written in WATFOR using double precision and run on the IBM 360 model 65. The convergence criteria are that each component of two succeeding iterates must agree to within a relative error of 10^{-5} and each component of the function must be less than 10^{-5} in absolute value.

3. The examples and the Results

The example we choose to analyze in depth is the two-by-two equation

$$F(x) = \begin{bmatrix} x_1^2 - x_2 + 1 \\ \\ x_1 - \cos \frac{\Pi}{2} x_2 \end{bmatrix}$$

with initial guess $x_0 = (1,0)$. The solution we seek is at $x^* = (0,1)$. The Jacobian of this system is

$$J(x) = \begin{bmatrix} 2x_1 & -1 \\ 1 & \frac{\Pi}{2} \sin \frac{\Pi}{2} x_2 \end{bmatrix}$$

and the determinant of the Jacobian is given by

$$\mathrm{Det}(J(x)) = x_1 \sin \frac{\Pi}{2} x_2 + 1$$

The Jacobian is thus singular for

$$\sin \frac{\Pi}{2} x_2 = -\frac{1}{\Pi x_1} .$$

The lines of singularity are plotted in figure 3.

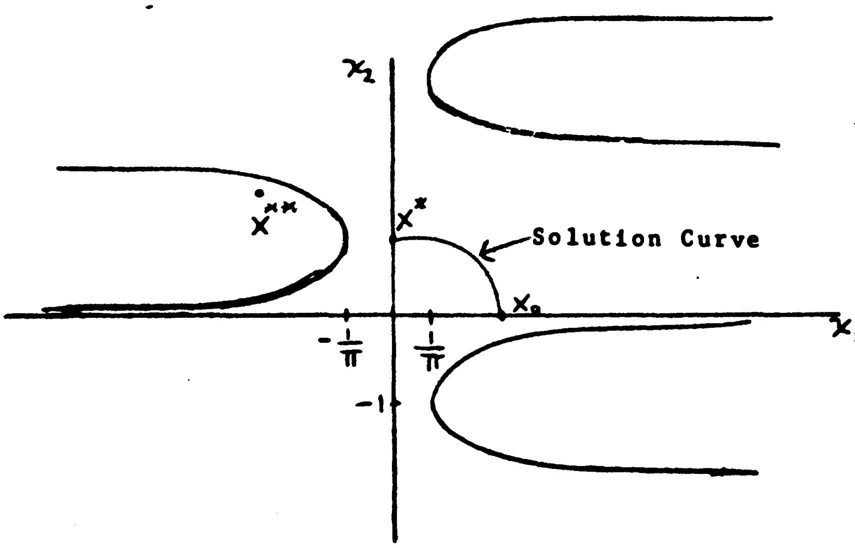


Figure 3

Also plotted in figure 3 are x_0, x^*, another root x^{**} and the solution curve of the I.V.P. projected onto the x_1,x_2-plane. Since this solution curve tends to x^*, x^* is the root we are seeking. Newton's method, however, does not converge to x^*, but rather, it crosses a line of singularities and converges in eight iterations to $x^{**}=(-\frac{1}{2}\sqrt{2}, \frac{3}{2})$. The iterates generated by Newton's method display the classical form of instability reported by Gear in [21], i.e., the first two iterates oscilate unstably around the true solution curve.
The second iterate is outside of the asymptotic stability region of x^*. From there, the third iterate crosses a singularity and enters the asymptotic stability region of x^{**}. Thereafter, the iterates converge very rapidly to x^{**}. See figure 4 for a two-dimensional representation of this process.

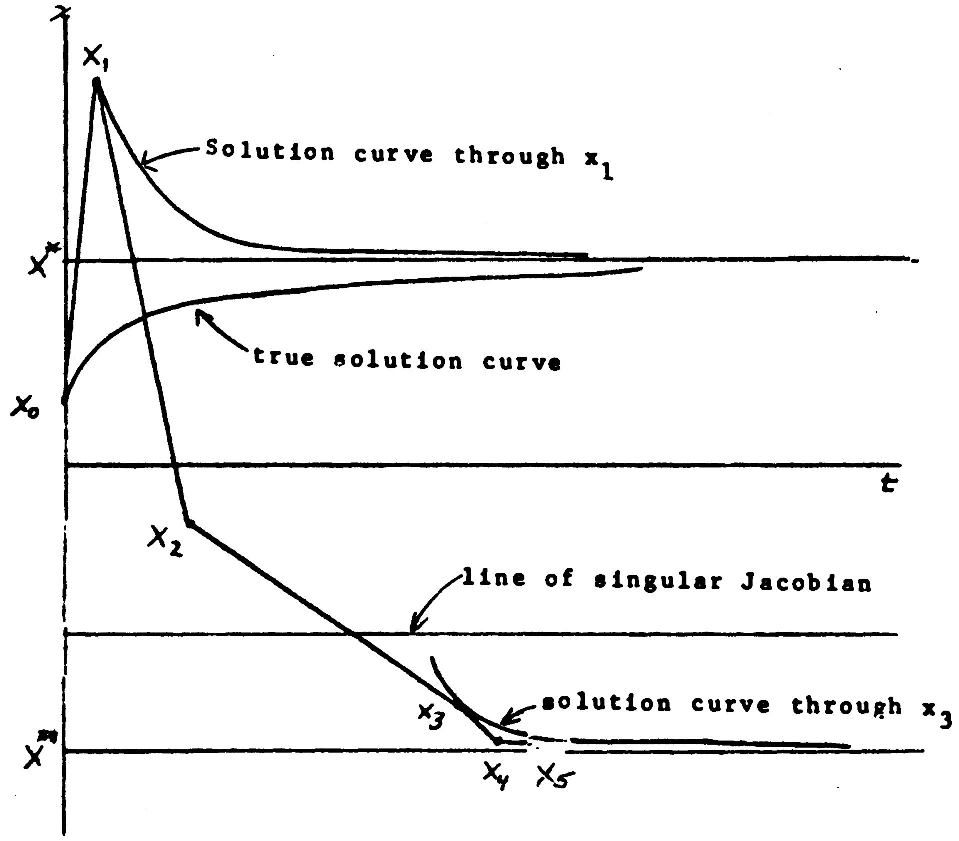


Figure 4

The damped Newton's method was applied to this problem and it converged to the root x^* in 107 iterations or 321 function evaluations. (Compare to Table I.)

When Nordsieck's method was applied to the problem, the iterates were able to get very close to x*, i.e., to (.06, .99), but not consistently closer. The procedure was allowed to go on for another 275 iterations and it only reached the point (.007, .98). The modified Nordsieck method was able to generate iterates which "almost" converged and then "blew-up". The iterates moved towards x* until finally the 40th iterate crossed the asymptote. The next iterate recrossed the asymptote and was an equally good approximation to x*. Each iterate after that again recrossed the asymptote but with ever increasing error. Figure 5 graphically illustrates this situation.

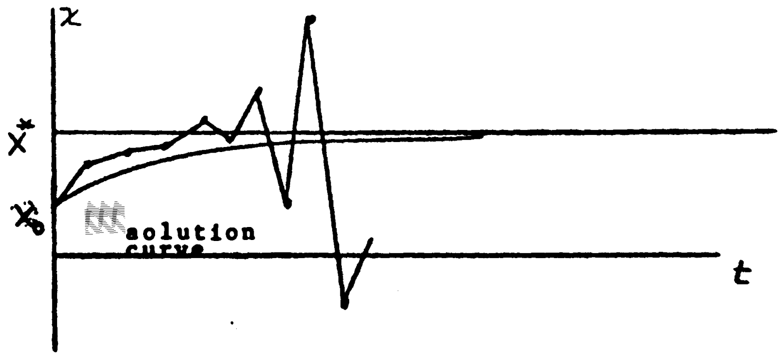


Figure 5

The last two examples show the two types of trouble that can arise when using methods which are not A-stable. Recall, however, that the stiffly stable methods of Gear [21] were also designed for this type of problem and they would probably also converge, but, to the author's knowledge, this has not been proved.

We next report the results of the application of the algorithms of Chapter III to the problem. These results are summarized in Table I and discussed below. For each algorithm, with and without the accuracy test, Table I gives the number of function evaluations per iteration, the number of iterations to convergence and the total number of function evaluations including initialization evaluations where necessary. The maximum number of iterations allowed was set at 100 and the minimum step-size allowed was 1/32. When no accuracy test was used, the step-size was kept constant at one.

Table I

Algorithm	Accuracy Test	No. Funct. Evals/Iter.	No. Iterations	Total No. Funct. Evals.
PECE	Yes	6	47	282
PECE	No	6	23	138
PE_BCE	Yes	4	101+	——
PE_BCE	No	4	17	68
PE_BCE_B	Yes	2	101+	——
PE_BCE_B	No	2	17	34+2 = 36
PECEC	Yes	6	26	156
PECEC	No	6	40	240
PE_BCEC	Yes	4	28	112
PE_BCEC	No	4	17	68
PE_BCE_BC	Yes	2	101+	——
PE_BCE_BC	No	2	101+	——

The first observation is that many of these methods were successful and that the method requiring the least amount of work was the PE_BCE_B algorithm without the accuracy test. It has been observed that this algorithm, when it converges, is usually the best of those tested. More will be said about this algorithm later. Next we note that when Broyden updates were used in an algorithm in conjunction with the accuracy test, the number of iterations often exceeded the allowable limit. This seems to indicate that the accuracy test for this problem is too sensitive and that the Broyden updates are not good approximations to the Jacobian.

As we would expect, the methods not using the accuracy test usually converged faster than those with the accuracy test. This is explained by referring, for example, to figure 5, and noting that the accuracy test merely ensures that the iterates remain close to the solution curve. Without the accuracy test, the algorithms can generate iterates with larger errors but in the "right" direction.

When adding the extra correction to the basic algorithm, the behavior is not always predictable. The expected result is that it may help an algorithm which has not converged due to the necessity of an excessively small step-size by providing more accuracy. Note, for example, that the PECE algorithm with the accuracy test converged in 47 iterations and the PECEC algorithm with the accuracy test converged in 26 iterations. An examination of the iterates and the step-sizes shows that the step-size was generally allowed to be twice as large in the latter method as in the former. Without the accuracy test, just the opposite effect was noticed: the PECE algorithm converged almost twice as fast as the PECEC algorithm. Again the reason for this is that the PECEC algorithm kept the iterates closer to the true solution curve and hence further from the asymptote. The constant step-size of one in both methods did not allow the increased accuracy to help.

In the case of the PE_BCE_B algorithm with no accuracy test and the PE_BCE_BC algorithm with no accuracy test, we note that the final correction caused the iteration not to converge. This is the same phenomenon as above. With the accuracy test, the PE_BCE_BC was still not able to converge in 100 iterations.

The final example reported here is a problem found in Broyden [5]:

$$F(x) = \begin{bmatrix} \frac{1}{2}[\sin(x_1x_2) - x_2/(2\pi) - x_1] \\ [1 - 1/(4\pi)]\ (e^{2x_1} - e) + ex_2/\pi - 2ex_1 \end{bmatrix}$$

with initial approximation (.4,3). There is a solution at $(.5,\pi)$, but this is not the solution we are seeking. Newton's method converges to the root (-.26,.62) in five iterations, but this also is not the correct root. The asymptote of the solution of the I.V.P. $x' = -J^{-1}(x)F(x)$, $x(0) = (.4,3)$ is $x^* = (.30,.28)$.

Application of the methods of Chapter III to the problem yield the following results: (Table II is organized exactly as Table I).

Table II

Algorithm	Accuracy Test	No. Funct. Evals/Iter.	No. Iterations	Total No. Funct. Evals.
PECE	Yes	6	29	174
PECE	No	6	16	96
PE_BCE	Yes	4	33	132
PE_BCE	No	4	15	60
PE_BCE_B	Yes	2	55	110+2 = 112
PE_BCE_B	No	2	singular	——
PECEC	Yes	6	17	102
PECEC	No	6	wrong root	——
PE_BCEC	Yes	4	18	72
PE_BCEC	No	4	wrong root	——
PE_BCE_BC	Yes	2	18	36
PE_BCE_BC	No	2	diverged	——

This problem was chosen because of the many roots contained in the region around the correct solution. We notice that many of the algorithms tested were able to converge to the correct root even without the accuracy test and in no case while using the accuracy test did we fail to obtain convergence. In the three cases without the accuracy test and with the extra correction, we either obtained divergence or convergence to the wrong root - the PECEC algorithm converged to (1.48, -8.38) and the PE_BCEC algorithm converged to (-.26,.62).

The PE_BCE_B had obtained the desired root to two significant figures in 11 iterations, but was then unable to significantly improve the estimate before the Broyden approximation became singular. An examination of the iterates shows that they were changing only in the fifth significant digit after 15 iterations, suggesting that the Jacobian approximation had become very poor. The remedy for this situation seems to be to reevaluate the Jacobian if the iterates have essentially converged, but the function values are still relatively large. We reran this case with the change that the Jacobian was reevaluated after every 4 iterations. In this case, we obtained convergence in 14 iterations or 38 function evaluations.

4. Conclusions. The methods described above have been tested on several other examples with results similar to those reported here. Although no really firm conclusions can be drawn from so few examples, some tentative statements based on these runs and the

convergence theory of Chapter III can be made. In the first place, the most economical and efficient algorithm of the set seems to be the PE_BCE_B algorithm without the accuracy test, but with the modification described above. This can be theoretically justified by recalling from Theorem 3.4.6 that the region of convergence of the PE_BCE algorithm is the same as the region for the PECE algorithm. An examination of Theorem 3.4.7 indicates that by using basically the PE_BCE_B algorithm, but reevaluating the Jacobian when necessary, we could prove convergence in a wider region than the region for the strict PE_BCE_B method.

The next conclusion is that the algorithms converge slowly even when very close to the root. The obvious remedy for this situation is to switch to Newton's method or Broyden's method when close to the root. An estimate for h_o of Theorem 3.4.6 can be made by using an approximation to the Lipschitz constant K based on two successive iterates $(K \sim ||J_n - J_{n-1}||/||x_n - x_{n-1}||)$ and bounding the norm of J_n^{-1} and $J_n^{-1}F_n$. When the approximation to h_o is less than 1/2, one may tentatively switch to a one-step method.

We finally conclude that these algorithms are, in general, able to find solutions when Newton's method is unable to and are thus useful tools for solving difficult nonlinear problems.

CHAPTER V

Conclusion

In summary, we can say that the principal objectives of this study have been achieved. We have been successful in creating a new class of algorithms of which some members have proved to be effective where certain standard algorithms have not been. The theory of d-majorization has been developed and shown to be a useful tool in working with the differential equations derived by differentiating w.r.t. a parameter.
By using a particular O.D.E., we have enlarged the uniqeness region of a zero of the function F and by the use of perturbation techniques, we have made some contributions to the theory of A-stability.

Although we have met with a measure of success, several questions present themselves as areas for future research. In particular, is there any advantage in considering the stiffly-stable methods of Gear mentioned in Chapter III? We would expect that these methods might be useful in problems where regions of nonsingularity of the Jacobian occur very close to the true solution curve. If we do use a stiffly-stable method, when, if ever, is it desirable to switch to an A-stable method?

It would seem that the techniques used in Chapter III to sharpen Dahlquist's results for a particular class of problems could be used to obtain similar results for a much wider class. In particular, suppose the u.a.s. solution is not a constant solution. Could we still apply the same type of perturbation theory? A preliminary analysis seems to indicate that the answer is yes. Also, is there any advantage in considering nonlinear A-stable methods such as those developed by Loscalzo [26] using spline functions?

There are still several questions raised by J. Davis which are yet unanswered. For example, how does one proceed when the parameter is a vector? Would a complex parameter be advantageous?

Finally we ask if these techniques can be used effectively in the neighborhood of clustered roots and if they can be used successfully with deflation techniques. It would seem that these questions could be answered in the affirmative since we have great freedom with regards to the choice of the imbedding function and the integration technique.

In conclusion, we have developed a sound theoretical and practical basis for the use of the differential equations approach with solution by continuation. We have not, however, as indicated above, exhausted this fertile area.

BIBLIOGRAPHY

[1] Anselone, P.M. and Moore, R.H.,"An Extension of the Newton-Kantorovich Method for Solving Nonlinear Equations with an Application to Elasticity," Math. Research Ctr. Tech. Report 520 (Madison, Wisc.).

[2] Bosarge, W.E., Jr.,"Infinite Dimentional Iterative Methods and Applications,"IBM Publication 320-2347 (Houston, Texas).

[3] Broyden, C.G.,"A Class of Methods for Solving Nonlinear Simultaneous Equations," Math. Comp., V. 19, 1965, pp. 577-593.

[4] ————'"Quasi-Newton Methods and Their Application to Function Minimization," Math. Comp., V. 21, 1967, pp. 368-381.

[5] ————, "A New Method of Solving Nonlinear Simultaneous Equations," Computer Journal, V. 12, 1969, pp. 94-99.

[6] ————, "The Convergence of Single-Rank Quasi-Newton Methods,"(Accepted for publication in Math. Comp.)

[7] Coddington, E.A. and Levinson, N., Theory of Ordinary Differential Equations, McGraw-Hill, New York, 1955.

[8] Dahlquist, G.G., "A Special Stability Problem for Linear Multistep Methods, " B.I.T., V. 3, 1963, pp. 27-43.

[9] Davidenko, D.F., "On a New Method of Numerical Solution of Systems of Nonlinear Equations," Doklady Adad. Nauk SSSR (N.S.) 88, 1953, pp. 601-602. (Russian)

[10] Davidon, W.C., "Variable Metric Method for Minimization," A.E.C. Research and Development Report, ANL-5990 (Rev. TID-4500, 14th ed.), 1959.

[11] Davis, J.,"The Solution of Nonlinear Operator Equations with Critical Points," Ph.D. thesis, Oregon State Univ., 1966.

[12] Deist, F.H., and Sefor, L., "Solution of Systems of Nonlinear Equations by Parameter Variation," Computer Journal, V. 10, 1967, pp. 78-82.

[13] Dennis, J.E., "On the Convergence of Newton-Like Methods," Numerical Methods for Nonlinear Algebraic Equations, P. Rabinowitz (editor), Gordon and Breach, London, 1970.

[14] ———'"On the Kantorovich Hypotheses for Newton's Method," SIAM J. Numer. Anal., V. 6, 1969, pp. 493-507.

[15] ———,"On the Convergence of Broyden's Method for Nonlinear Systems of Equations," Cornell Univ. Tech. Report No. 69-48, 1969.

[16] ———, "On the Local Convergence of Broyden's Method for Nonlinear Systems of Equations," Cornell Univ. Tech. Report No. 69-46, 1969.

[17] ———, and Sweet, R.A., "Some Minimal Properties of the Trapezoidal Rule" (to appear).

[18] Ficken, F.A., "The Continuation Method for Functional Equations," Comm. Pure Appl. Math., V. 4, 1951, pp. 435-456.

[19] Freudenstein, F. and Roth, B., "Numerical Solutions of Systems of Nonlinear Equations," JACM, V. 10, 1963, pp. 550-556.

[20] Gavurin, M.K., "Nonlinear Functional Equations and Continuous Analogs of Iterative Methods," Izv. Vyss. Ucebn. Zaved. Mathematika, V. 6, 1958, pp. 18-31, (Russian); Transl. Kocho, M., Ed. Avila, J.H., Univ. of Maryland Tech. Report No. 68-70.

[21] Gear, C.W., "Numerical Integration of Stiff Ordinary Differential Equations," Univ. of Illinois Tech. Report No. 221.

[22] Henrici, P.,Discrete Variable Methods in Ordinary Differential Equations, John Weley, New York, 1962.

[23] Hull, T.E. and Creemer, A.L.,"Efficiency of Predictor-Corrector Procedures," JACM, V. 10, 1963, pp. 291-301.

[24] Kantorovich, L.V. and Akilov, G.P., Functional Analysis in Normed Spaces, Macmillan, New York, 1964.

[25] Kizner, W., "A Numerical Method for Finding Solutions of Nonlinear Equations," J. SIAM, V. 12, 1964, pp. 424-428.

[26] Loscalzo, F.R., "An Introduction to the Applications of Spline Functions to Initial Value Problems," Theory and Applications of Spline Functions, T.N.E. Greville (editor), Academic Press, New York, 1969.

[27] Meyer, G.H., "On Solving Nonlinear Equations with a One-Parameter Operator Imbedding," SIAM J. Numer. Anal., V. 5, 1968, pp. 739-752.

[28] Nordsieck, A., "On Numerical Integration of Ordinary Differential Equations," Math. Comp. V. 16, 1962, pp. 22-49.

[29] Rall, L.B. and Tapia, R.A., "The Kantorovich Theorem and Error Estimates for Newton's Method, Math. Research Ctr. Tech. Report 1043 (Madison, Wisconsin.)

[30] Rosenbloom, P.C., "The Method of Steepest Descent," AMS Proc. Symp. Appl. Math., V. 6, 1956, pp. 127-176.

[31] Tapia, R.A.,"The Kantorovich Theorem," Math. Research Ctr. Tech. Report 1007 (Madison, Wisc.)

[32] Yakovlev,N.N., "On the Solution of Systems of Nonlinear Equations by Differentiation with Respect to a Parameter," USSR Comput. Math. and Math. Phys., V. 4, 1964, pp. 146-149.

[33] ————, "On Certain Methods of Solution of Nonlinear Equations," Trudy Mat. Inst. Steklov., v. 84, 1965, pp. 8-40, (Russian); Transl. Kocho, M., Ed. Avila, J.H., Univ. of Maryland Tech. Report No. 68-75.