

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

PAUL T. BOGGS†

Abstract. It is well known that a damped or underrelaxed Newton's method will sometimes solve a system of nonlinear equations when the full Newton's method cannot. This happens, for example, when only a poor initial approximation to the solution is known. By considering Newton's method as Euler's method applied to the corresponding differential equation, we ask if there is a better way to integrate that differential equation, *i.e.*, we seek a more rapidly converging and more stable integration technique than damped Newton's method. It is shown here by an extension of the Dahlquist A -stability theory that the A -stable methods allow us to achieve our goals.

1. Introduction. The primary purpose of this paper is to generate a class of practical numerical methods for finding the solution \mathbf{x}^* to the nonlinear system of equations $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ when only a poor initial approximation, \mathbf{x}_0 , is known. The procedure here is to examine the class of Davidenko [9] differential equations, to choose a practical member from this class and to attempt to find the "best" way to integrate that differential equation. The ordinary differential equation (O.D.E.) which we choose is such that its solution $\mathbf{x}(t) \rightarrow \mathbf{x}^*$ as $t \rightarrow \infty$, where t is the independent variable. This form enables us to consider the class of linear multistep integration techniques as iteration methods to solve nonlinear equations and to search this class for members having certain desirable properties.

Several other authors, *e.g.*, Meyer [25] and Bosarge [3], have considered integrating a Davidenko differential equation, but have chosen a finite interval, usually $[0, 1]$, for the independent variable. In this way, they are forced to consider accuracy of prime importance since the approximate value of $\mathbf{x}(1)$ will be the best approximation obtainable for \mathbf{x}^* without further refinement. By choosing the infinite interval we can, as is shown below, avoid the necessity of being overly concerned with accuracy.

A large number of authors consider the use of solution by continuation, from which the Davidenko equations are derived, to solve nonlinear problems, but their work has not resulted in the creation of practical general purpose algorithms.

In § 2 of this paper we derive the particular differential equation with which we shall be concerned in the sequel. We then consider the characteristics of the solution of that differential equation and heuristically justify the use of the A -stable methods of Dahlquist [8] as the primary integration schemes. In order to theoretically justify this choice, we sharpen some of the results of Dahlquist relative to our differential equation.

The results of § 2 lead to the choice of the trapezoidal rule as the basic algorithm, and in § 3 we consider the practical aspects of actually using the trapezoidal

* Received by the editors August 27, 1970, and in revised form February 18, 1971.

† Department of Computer Science, University of Kansas, Lawrence, Kansas 66044. This work was supported in part by the National Science Foundation under Grant GJ-844 to the Computer Science Department, Cornell University, Ithaca, New York, and by a NSF Traineeship to the author while at Cornell University.

rule as an integration technique. We are thereby led to develop a predictor-corrector algorithm and to consider modifications which substantially reduce the amount of computational effort at each step. Convergence theorems for several of the resulting algorithms are then given. Finally, in § 4, some numerical results are presented.

Throughout the paper, $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ is the equation whose solution is sought. Here we consider $\mathbf{F}: D \subset R^N \rightarrow R^N$ although much of the analysis extends to Hilbert space. The point \mathbf{x}^* will always denote the solution and \mathbf{x}_0 the initial approximation to \mathbf{x}^* . The function \mathbf{F} is assumed differentiable and $\mathbf{J}(\mathbf{x})$ denotes the Jacobian of \mathbf{F} at the point \mathbf{x} .

2. The differential equations and A -stability. In order to derive the O.D.E., we consider the operator homotopy $\mathbf{H}(t, \mathbf{x}) = \mathbf{0}$ which imbeds the real parameter t into the original equation in such a way that $\mathbf{H}(0, \mathbf{x}) = \mathbf{0}$ has the solution \mathbf{x}_0 and $\mathbf{H}(t', \mathbf{x}) = \mathbf{F}(\mathbf{x})$. We note here that the interval of definition of the parameter t may be infinite, i.e., t' may be ∞ . For example, we could use

$$(2.1) \quad \mathbf{H}(t, \mathbf{x}) = \mathbf{F}(\mathbf{x}) - e^{-t}\mathbf{F}(\mathbf{x}_0)$$

so that $\mathbf{H}(t, \mathbf{x}) \rightarrow \mathbf{F}(\mathbf{x})$ as $t \rightarrow \infty$.

Sufficient conditions that the homotopy be effective are that the curve $\mathbf{x}(t)$ be differentiable with respect to t , that \mathbf{H} be differentiable with respect to \mathbf{x} and that the linear operator $\mathbf{H}_{\mathbf{x}}(t, \mathbf{x})$ be nonsingular for all values of \mathbf{x} and t in some domain. Then, since $\mathbf{H}(t, \mathbf{x}(t)) = \mathbf{0}$, we have, by differentiating with respect to t and using the chain rule,

$$(2.2) \quad \mathbf{x}'(t) = -\mathbf{H}_{\mathbf{x}}^{-1}(t, \mathbf{x}(t))\mathbf{H}_t(t, \mathbf{x}(t))$$

For the example of (2.1) we obtain the initial value problem

$$(2.3) \quad \mathbf{x}'(t) = -\mathbf{J}^{-1}(\mathbf{x})\mathbf{F}(\mathbf{x}), \quad \mathbf{x}(0) = \mathbf{x}_0.$$

This is the problem with which we shall be primarily concerned in the remainder of the paper. We note that Euler's method with a step size of one applied to (2.3) is equivalent to Newton's method applied to the original equation $\mathbf{F}(\mathbf{x}) = \mathbf{0}$. Many other differential equations arising from different imbedding functions could be used, but, as long as $\mathbf{x}(t) \rightarrow \mathbf{x}^*$ as $t \rightarrow \infty$, the development which follows could be applied. (See Gavurin [19] for an alternate derivation of these differential equations.)

The derivation of (2.3) was under the hypothesis of a nonsingular Jacobian over some region. It now seems reasonable to reverse the question and ask under what conditions on (2.3) we can guarantee the existence of a solution which tends to \mathbf{x}^* . In other words, when can we show that $\mathbf{x}(t) \equiv \mathbf{x}^*$ is an asymptotically stable solution of (2.3)? We would certainly expect that the Kantorovich hypotheses for Newton's method would be sufficient to guarantee this and indeed they are; the author in [2] has shown this by extending the concept of weak majorization for iteration functions to a class of differential equations. (See Dennis [13] for the development of weak majorization.)

At this point we consider an example which illustrates the stability problem which we hope to overcome and provides the motivation for our approach. Consider the scalar equation $F(x) = \sin x$ and the corresponding equation of

the form (2.3) with Euler's method with step size one (Newton's method applied to $F(x) = \sin x$) used to integrate the equation. This example is illustrated in Fig. 1 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. Note that the iterates "bounce" from one side of the root to the other. The type of "bouncing" which occurs here also occurs in higher dimensions and is the cause of instability when it arises while using some numerical integration techniques.

From this example, it is clear that we are only interested in the asymptote of the solution and that we are not overly concerned with accuracy. High accuracy methods would merely keep the iterates near the true solution curve and thus converge more slowly to the asymptote. Thus we would like to strike a compromise between accuracy and the rate of convergence while maintaining stability, and we are therefore led to consider the A -stable methods of Dahlquist [8].

We digress here to give a brief account of Dahlquist's work and to indicate where his results can be strengthened relative to a particular class of differential equations

Dahlquist motivates his concept of A -stability by considering the linear O.D.E.

$$(2.4) \quad \mathbf{y}'(t) = q\mathbf{y},$$

where q is a complex constant with negative real part. The solution of this

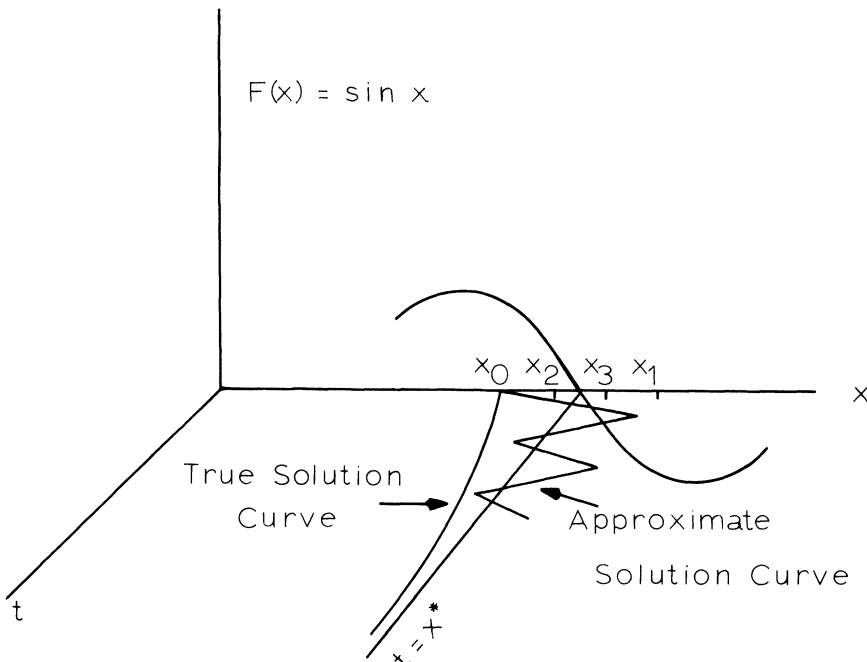


FIG. 1

differential equation for any initial value \mathbf{y}_0 is given by $e^{at}\mathbf{y}_0$. Clearly $\mathbf{y}(t)$ tends to zero as $t \rightarrow \infty$. Dahlquist thus desires a method which, when applied to (2.4) with any initial value, also tends to zero as $n \rightarrow \infty$. 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.

DEFINITION 2.1. For the initial value problem $\mathbf{y}' = \mathbf{f}(t, \mathbf{y})$, $\mathbf{y}(0) = \mathbf{y}_0$, the *general linear k -step method* is defined by

$$(2.5) \quad \sum_{i=0}^k \alpha_i \mathbf{y}_{n+i} = h \sum_{i=0}^k \beta_i \mathbf{f}_{n+i}, \quad n = 0, 1, 2, \dots,$$

where k is an integer and $\alpha_j, \beta_j, j = 0, 1, \dots, k$, are real constants independent of n . $\mathbf{f}_m = \mathbf{f}(t_m, \mathbf{y}_m)$, $m = 0, 1, \dots$, and h is the step size.

Henrici [21] is a standard reference for the general theory of such methods.

DEFINITION 2.2. A k -step method is said to be *A-stable* if all solutions of (2.5) tend to zero as $n \rightarrow \infty$ when the method is applied with fixed positive step size h to any equation of the form (2.4).

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

LEMMA 2.3 (Dahlquist). *For a k -step method, define the polynomials*

$$\rho(\zeta) = \sum_{j=0}^k \alpha_j \zeta^j, \quad \sigma(\zeta) = \sum_{j=0}^k \beta_j \zeta^j.$$

Then the method is A-stable if and only if $\rho(\zeta)/\sigma(\zeta)$ is regular and has nonnegative real part for $|\zeta| > 1$.

THEOREM 2.4 (Dahlquist). *An explicit k -step method cannot be A-stable.*

THEOREM 2.5 (Dahlquist). *The order p of an A-stable linear multistep method cannot exceed 2. The smallest error constant $c^* = \frac{1}{12}$ is obtained for the trapezoidal rule*

$$(2.6) \quad \mathbf{y}_{n+1} = \mathbf{y}_n + (h/2)[\mathbf{f}_{n+1} + \mathbf{f}_n].$$

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 (2.4), i.e., the equation

$$(2.7) \quad \mathbf{x}'(t) = \mathbf{f}(t, \mathbf{x})$$

should have a uniformly asymptotically stable (u.a.s.) solution. Based on these derived conditions, Dahlquist is able to show that in approximating this u.a.s. solution using the trapezoidal rule with fixed positive step size, the computation results in a bounded error for an infinite computation. That is, for an equation of the form (2.7) satisfying these derived conditions,

$$\|\mathbf{x}(t_n) - \mathbf{x}_n\| \leq r$$

for all $n > 0$, where r is an upper bound on the truncation error and \mathbf{x}_n is the approximation to $\mathbf{x}(t_n)$ generated by the trapezoidal rule.

In our case, however, we know that if we are close enough, all solutions to our differential equation tend to the constant solution $\mathbf{x}(t) \equiv \mathbf{x}^*$. Thus we would like to strengthen Dahlquist's results relative to a class of differential equations whose solutions in some region tend to a constant solution.

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 (2.3) can be written as a perturbed linear equation by expanding the right-hand side through three terms of a Taylor series to obtain

$$(2.8) \quad \mathbf{x}' = -\mathbf{x} + \mathbf{R}(\mathbf{x}),$$

where $\mathbf{R}(\mathbf{x})$ is the remainder term and the origin has been translated to \mathbf{x}^* .

In the theorem which follows, we consider perturbed linear differential equations and integration techniques satisfying a weaker stability condition than A -stability. The theorem is more general than necessary for the present work, but its generality is useful in other applications. Corollary 2.8 specializes the result to an equation of the form (2.8) in which case we obtain convergence to zero. We first give the following definition.

DEFINITION 2.6. A k -step method is said to be *weakly A -stable* with respect to (2.4) if there exists an h depending on q in that equation and on the method such that all solutions of (2.5) tend to zero as $n \rightarrow \infty$ when the method is applied to that equation.

Remark. Definition 2.6 is equivalent to saying that for a method to be weakly A -stable there must exist an h such that all of the roots of $\rho(\zeta) - hq\sigma(\zeta) = 0$ are strictly less than one. Not all consistent methods are weakly A -stable: consider $\rho(\zeta) = \zeta - 1$ and $\sigma(\zeta) = -\zeta^3 + 2$.

THEOREM 2.7. For the equation

$$(2.9) \quad \mathbf{x}' = q\mathbf{x} + \mathbf{f}(\mathbf{x}) + \mathbf{g}(\mathbf{x}, \mu)$$

let q be a complex constant with negative real part and let \mathbf{f} and \mathbf{g} be continuous for $\|\mathbf{x}\|$ small. Assume that for every $\varepsilon > 0$ there exists a $\delta > 0$ such that if $\|\mathbf{x}\| < \delta$, then $\|\mathbf{f}(\mathbf{x})\| < \varepsilon\|\mathbf{x}\|$ and $\|\mathbf{g}(\mathbf{x}, \mu)\| \leq G_\varepsilon(\mu)$. Assume further that $G_\varepsilon(\mu) < \infty$ for all μ and ε and that $G_\varepsilon(\mu)$ is continuous in μ with $\lim_{\mu \rightarrow 0} G_\varepsilon(\mu) = 0$. Choose any weakly A -stable method and an h such that Definition 2.6 is satisfied with respect to the equation $\mathbf{x}' = q\mathbf{x}$. Choose ε such that $1 - h\varepsilon \sum_{i=0}^k |\beta_i|K > 0$, where K is a positive constant depending on h, q and the particular method used. Then there exists a $\delta' \leq \delta$ (δ depends on ε) such that if $\{\mathbf{x}_n\}$ is generated by the application of this method to (2.9) from initial points $\{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{k-1}\} \in \{\mathbf{x} : \|\mathbf{x}\| < \delta'\}$, then

$$\limsup_{n \rightarrow \infty} \|\mathbf{x}_n\| \leq \frac{hG_\varepsilon(\mu) \sum_{i=0}^k |\beta_i|K}{1 - h\varepsilon \sum_{i=0}^k |\beta_i|K}.$$

Proof. Let

$$\sum_{i=0}^k \alpha_i \mathbf{x}_{n+i} = h \sum_{i=0}^k \beta_i (q\mathbf{x}_{n+i} + \mathbf{f}_{n+i} + \mathbf{g}_{n+i})$$

be the weakly A -stable method applied to (2.9). Then

$$(2.10) \quad \sum_{i=0}^k \alpha_i \mathbf{x}_{n+i} - hq \sum_{i=0}^k \beta_i \mathbf{x}_{n+i} = \gamma_{n+k} + \Phi_{n+k},$$

where $\gamma_{n+k} = h \sum_{i=0}^k \beta_i \mathbf{f}_{n+i}$, $\Phi_{n+k} = h \sum_{i=0}^k \beta_i \mathbf{g}_{n+i}$. Thus we have a nonhomogeneous linear difference equation with constant coefficients. It is well known, cf. Henrici [21, p. 212], that the solution of such equations may be written as $\mathbf{x}_n = \mathbf{s}_n + \mathbf{z}_n$, where \mathbf{s}_n denotes the solution to the homogeneous equation ((2.10) with $\gamma_n = \Phi_n = 0$ for all n) with the given initial values and \mathbf{z}_n is a particular solution to (2.10) with zero initial data. Following Henrici [21, Theorem 5.2, p. 212] we construct a representation for \mathbf{z}_n as follows: Define $\{y_j\}$ to be a sequence which satisfies the homogeneous equation for all $j > 0$ and $y_0 = 1/(\alpha_k - hq\beta_k)$. Then

$$\mathbf{z}_n = \sum_{m=k}^n \gamma_m y_{n-m} + \sum_{m=k}^n \Phi_m y_{n-m}$$

has the desired property.

Now, for $\|\mathbf{x}_{m+i}\| \leq \delta$, $i = 0, 1, \dots, k$,

$$(2.11) \quad \begin{aligned} \|\gamma_{m+k}\| &= \left\| h \sum_{i=0}^k \beta_i \mathbf{f}_{m+i} \right\| \leq h \sum |\beta_i| \cdot \|\mathbf{f}(\mathbf{x}_{m+i})\| \\ &\leq h \sum |\beta_i| (\varepsilon \|\mathbf{x}_{m+i}\|) \leq \varepsilon' \bar{x}_m, \end{aligned}$$

where $\bar{x}_m = \max_{0 \leq i \leq k} \|\mathbf{x}_{m+i}\|$ and $\varepsilon' = h\varepsilon \sum_{i=0}^k |\beta_i|$. We also have that

$$(2.12) \quad \|\Phi_{m+k}\| \leq h \sum_{i=0}^k |\beta_i| \cdot \|\mathbf{g}_{m+i}\| < h \sum_{i=0}^k |\beta_i| G_\varepsilon(\mu) \leq G'_\varepsilon(\mu)$$

where $G'_\varepsilon(\mu) = h \sum_{i=0}^k |\beta_i| G_\varepsilon(\mu)$. Thus by (2.11) and (2.12),

$$(2.13) \quad \begin{aligned} \|\mathbf{x}_n\| &= \|\mathbf{s}_n + \mathbf{z}_n\| = \left\| \mathbf{s}_n + \sum_{m=k}^n [\gamma_m + \Phi_m] y_{n-m} \right\| \\ &\leq \|\mathbf{s}_n\| + \varepsilon' \sum_{m=k}^n \bar{x}_{m-k} |y_{n-m}| + G'_\varepsilon(\mu) \sum_{m=k}^n |y_{n-m}|. \end{aligned}$$

Let $X_N = \max \{\|\mathbf{x}_n\| : n \leq N\}$ and $S_N = \max \{\|\mathbf{s}_n\| : n \leq N\}$. Then

$$X_N \leq S_N + \varepsilon' X_N \sum_{m=k}^N |y_{n-m}| + G'_\varepsilon(\mu) \sum_{m=k}^N |y_{n-m}|$$

or

$$(2.14) \quad X_N \leq \frac{S_N}{1 - \varepsilon' K} + \frac{G'_\varepsilon(\mu) K}{1 - \varepsilon' K},$$

where $K = \sum_{m=0}^\infty |y_m|$. If $K < \infty$, then ε can indeed be chosen such that $1 - \varepsilon' K > 0$. The fact that $K < \infty$ follows from Henrici [21, Theorem 5.3, p. 214] and the remark following Definition 2.6. Now if μ is chosen sufficiently small, the second term of (2.14) can be made $< \delta/2$. Even though $\|\mathbf{s}_n\| \rightarrow 0$ as $n \rightarrow \infty$ by hypothesis, it does not necessarily tend monotonically to zero. S_N can however be made $< \delta/2$ by choosing δ' sufficiently small. Thus $\|\mathbf{x}_n\|$ will remain $< \delta$ for all n .

Let $X = \limsup_{n \rightarrow \infty} \|\mathbf{x}_n\|$. We have that $0 \leq X \leq \delta < \infty$ and (2.13) holds. For every integer I , let $\tilde{I} = \{i : i \geq I\}$. Now for every $\xi > 0$, there exists an I_ξ and an infinite set $\tilde{I}_\xi \subset \tilde{I}$ such that $\|\mathbf{x}_j\| > X - \xi$ for all $j \in \tilde{I}_\xi$. Also for every $\xi > 0$ there exists an integer M_ξ such that $X + \xi > \|\mathbf{x}_j\|$ for all $j > M_\xi$. We rewrite (2.13) as

$$\|\mathbf{x}_j\| \leq \|\mathbf{s}_j\| + \sum_{m=k}^{j/2} \|\boldsymbol{\gamma}_m\| \cdot |y_{j-m}| + \sum_{m=j/2+1}^j \|\boldsymbol{\gamma}_m\| \cdot |y_{j-m}| + \sum_{m=k}^j \|\boldsymbol{\Phi}_m\| \cdot |y_{j-m}|,$$

and thus for all $j \in \tilde{I}_\xi$ such that $j/2 > M_\xi$,

$$X - \xi < \|\mathbf{x}_j\| \leq \|\mathbf{s}_j\| + \varepsilon' \delta \sum_{m=k}^{j/2} |y_{j-m}| + \varepsilon'(X + \xi)K + G'_\varepsilon(\mu)K.$$

As $j \rightarrow \infty$, $\sum_{m=k}^{j/2} |y_{j-m}| \rightarrow 0$ and $\mathbf{s}_j \rightarrow \mathbf{0}$. Therefore, in the limit as $j \rightarrow \infty$,

$$X - \xi \leq \varepsilon'(X + \xi)K + G'_\varepsilon(\mu)K$$

and

$$(2.15) \quad X \leq \frac{1 + \varepsilon'K}{1 - \varepsilon'K} \xi + \frac{G'_\varepsilon(\mu)K}{1 - \varepsilon'K}.$$

But (2.15) must hold for all ξ so that

$$X \leq \frac{G'_\varepsilon(\mu)K}{1 - \varepsilon'K}.$$

This completes the proof.

We first specialize Theorem 2.7 to the case where $\mathbf{g}(\mathbf{x}, \mu) \equiv \mathbf{0}$ to correspond to (2.8). In that event, we have immediately the following corollary.

COROLLARY 2.8. *If $\mathbf{g}(\mathbf{x}, \mu) \equiv \mathbf{0}$, then $\|\mathbf{x}_n\| \rightarrow 0$ as $n \rightarrow \infty$.*

Remark. It should be noted here that a form of (2.8) can be derived by only assuming that $\|\mathbf{J}^{-1}(\mathbf{x})\|$ exists and is bounded, say by B , over an appropriate region around $\|\mathbf{x}\| = 0$ and that $\mathbf{J}(\mathbf{x})$ is Lipschitz continuous with constant L over the same region. Then, if $\mathbf{F}(\mathbf{0}) = \mathbf{0}$,

$$\begin{aligned} \mathbf{x}' &= -\mathbf{J}^{-1}(\mathbf{x})\mathbf{F}(\mathbf{x}) = -\mathbf{x} - \mathbf{J}^{-1}(\mathbf{x})\mathbf{F}(\mathbf{x}) + \mathbf{x} \\ &= -\mathbf{x} - \mathbf{J}^{-1}(\mathbf{x})(\mathbf{F}(\mathbf{x}) - \mathbf{J}(\mathbf{x})(\mathbf{x})) \\ &= -\mathbf{x} + \mathbf{J}^{-1}(\mathbf{x})(\mathbf{F}(\mathbf{0}) - \mathbf{F}(\mathbf{x}) - \mathbf{J}(\mathbf{x})(-\mathbf{x})) \end{aligned}$$

satisfies the hypotheses of Theorem 2.7 relative to the differential equation: $\mathbf{f}(\mathbf{x}) = \mathbf{J}^{-1}(\mathbf{x})(\mathbf{F}(\mathbf{0}) - \mathbf{F}(\mathbf{x}) - \mathbf{J}(\mathbf{x})(-\mathbf{x}))$, is continuous for $\|\mathbf{x}\|$ small, and by Davis [11, Proposition 7, p. 11],

$$\|\mathbf{f}(\mathbf{x})\| \leq B \cdot \|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{0}) - \mathbf{J}(\mathbf{x})(\mathbf{x})\| \leq \frac{1}{2}BL\|\mathbf{x}\|^2.$$

The advantage of selecting an A -stable method is that we may choose h independent of q and the particular method. There is also the advantage that the A -stable methods are low order, and hence we may expect reasonably rapid convergence near the solution. The preceding discussion and the following result, due to Dennis and Sweet [16], make the trapezoidal rule a very reasonable and

attractive choice for our problem. Theorem 2.9 says that the trapezoidal rule is the best in the sense of Sard.

THEOREM 2.9 (Dennis and Sweet). *Let $T(\alpha_0, \dots, \alpha_k, \beta_0, \dots, \beta_k)\mathbf{y}$ denote the truncation error functional associated with a k -step formula. Define $\|T(\cdot)\|_p \equiv \|T\|_p$, $p \in [1, \infty]$, to be the minimum number such that*

$$|T(\cdot)\mathbf{y}| \leq \|T\|_p \left(\int_0^1 |\mathbf{f}'(t, \mathbf{y})|^p dt \right)^{1/p}.$$

Then the minimum over all consistent methods of $\|T\|_p$ is $\frac{1}{2}(p+1)^{-1/p}$, and this is achieved only by the trapezoidal rule.

Gear [20] considers methods for integrating the linear equation (2.4) but with the relaxed condition that $\text{Re}(q) \leq \text{const.} < 0$ 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. The technique for deriving the stiffly stable methods in [20] is such that these methods are also weakly A -stable and thus Theorem 2.7 proves convergence for them relative to (2.8). In our case, however, 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

$$(3.1) \quad \mathbf{x}'(t) = -\mathbf{J}^{-1}(\mathbf{x})\mathbf{F}(\mathbf{x}), \quad \mathbf{x}(0) = \mathbf{x}_0,$$

and recall that any solution of (3.1) must satisfy

$$(3.2) \quad \mathbf{F}(\mathbf{x}(t)) = e^{-t}\mathbf{F}(\mathbf{x}_0).$$

The basic trapezoidal rule now becomes

$$(3.3) \quad \mathbf{x}_{n+1} = \mathbf{x}_n - (h/2)[\mathbf{J}_n^{-1}\mathbf{F}_n + \mathbf{J}_{n+1}^{-1}\mathbf{F}_{n+1}]$$

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 \mathbf{x} , \mathbf{F} and \mathbf{J} at t_{n+1} given the values at t_n , and iterating (3.3) to obtain final values of \mathbf{x} , \mathbf{F} and \mathbf{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 [4] and [5] 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 \mathbf{H}_{n+1} to the inverse Jacobian \mathbf{x}_{n+1} should satisfy

$$(3.4) \quad \mathbf{H}_{n+1}(\mathbf{F}_{n+1} - \mathbf{F}_n) = \mathbf{x}_{n+1} - \mathbf{x}_n.$$

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

$$\mathbf{H}_{n+1} = \mathbf{H}_n - \frac{[\mathbf{H}_n(\mathbf{F}_{n+1} - \mathbf{F}_n) - h(\mathbf{x}_{n+1} - \mathbf{x}_n)](\mathbf{H}_n^T \mathbf{H}_n \mathbf{F}_n)^T}{(\mathbf{H}_n^T \mathbf{H}_n \mathbf{F}_n)^T (\mathbf{F}_{n+1} - \mathbf{F}_n)}.$$

For the purposes at hand, we shall 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 \mathbf{x}_{n+1} from \mathbf{x}_n . We first need predicted values of \mathbf{F} and \mathbf{J}^{-1} at \mathbf{x}_{n+1} . We assume that $\mathbf{F}_n = \mathbf{F}(\mathbf{x}_n)$ and $\mathbf{H}_n = \mathbf{J}^{-1}(\mathbf{x}_n)$, or an approximation to it, are available. Then the most natural means of predicting \mathbf{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 \mathbf{p} be the predicted value of \mathbf{x}_{n+1} , we have $\mathbf{p} = \mathbf{x}_n - h\mathbf{H}_n\mathbf{F}_n$. Having obtained \mathbf{p} , we evaluate $\mathbf{F}(\mathbf{p})$ and then evaluate $\mathbf{J}^{-1}(\mathbf{p})$ by either direct computation or by a Broyden update of \mathbf{H}_n .

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

$$(3.5) \quad \begin{aligned} \mathbf{H}(\mathbf{p})\mathbf{F}(\mathbf{p}) &\approx \left[\mathbf{H}_n - \frac{[\mathbf{H}_n(e^{-h}\mathbf{F}_n - \mathbf{F}_n) + h\mathbf{H}_n\mathbf{F}_n](\mathbf{H}_n^T \mathbf{H}_n \mathbf{F}_n)^T}{(\mathbf{H}_n^T \mathbf{H}_n \mathbf{F}_n)^T (e^{-h}\mathbf{F}_n - \mathbf{F}_n)} \right] e^{-h}\mathbf{F}_n \\ &= \frac{he^{-h}}{1 - e^{-h}} \mathbf{H}_n \mathbf{F}_n. \end{aligned}$$

This method clearly requires very little computation, but has some serious disadvantages which will be discussed later.

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 \mathbf{J}^{-1} and \mathbf{F} at the iterates. We clearly must evaluate \mathbf{F} , and we could then evaluate \mathbf{J}^{-1} by either directly computing it or by use of the Broyden approximation using the computed value of \mathbf{F} . We note here that the Broyden evaluation requires only an evaluation of \mathbf{F} whereas the evaluation of \mathbf{J}^{-1} requires, by finite differences, an additional N evaluations of \mathbf{F} .

Having found \mathbf{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 \mathbf{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. Hull and Creemer [22] have shown in practice over a wide class of problems that, in general, final evaluations are neither necessary nor desirable. In our case, however, we intend to use rather large step sizes, and our numerical experiments indicate that final evaluations are desirable. Hull and Creemer also conclude that, in general, one or at most two iterations of the corrector are sufficient, a practice which we do follow. If the corrector is not

iterated and if $\mathbf{F}(\mathbf{p}) \approx e^{-h}\mathbf{F}_n$ is used to predict, then a final evaluation is necessary. The reasons for this will be discussed later.

The class of methods thus developed can now be described compactly as a predictor-corrector algorithm with the usual notation, i.e., $\text{P(EC)}^N[\mathbf{E}]$, $N \geq 1$. This means that we first predict, and then evaluate and correct as many times as we please, and finally do an optional final evaluation. As noted above, however, we have the choice in the evaluation phase of using Broyden's method to approximate \mathbf{J}^{-1} . To denote this choice, we shall use \mathbf{E}_B . Then for example, PE_BCE will mean that the first evaluation consists of an evaluation of $\mathbf{F}(\mathbf{p})$ and a Broyden update of \mathbf{J}_n^{-1} to approximate $\mathbf{J}^{-1}(\mathbf{p})$ whereas the second evaluation consists of actual evaluations of both \mathbf{F} and \mathbf{J}^{-1} at \mathbf{x}_{n+1} .

We note that all of the above methods involve an actual evaluation of \mathbf{F} and no use of the approximation $\mathbf{F}(\mathbf{p}) \approx e^{-h}\mathbf{F}_n$. This is because, cf. (3.5), the resulting approximation would be merely Newton's (Euler's) method with a different step size. In conjunction with this observation, it is interesting to consider the method of using (3.2) to approximate $\mathbf{F}(\mathbf{p})$, not iterating the corrector and not doing a final evaluation. With this choice, it would appear that the method would "solve" the equation having only evaluated the function and its derivative at \mathbf{x}_0 . It is no surprise that this does not work. An induction argument using (3.5) shows that the method converges to

$$\mathbf{x} = \mathbf{x}_0 - \frac{c}{1-d} \mathbf{J}_0^{-1} \mathbf{F}_0,$$

where $c = (h/2)[1 + d]$ and $d = he^{-h}/(1 - e^{-h})$. Obviously this is just an over-relaxed Newton's method with the step size dependent on h . Values of $r = c/(1 - d)$ were computed as a function of h with the result that r never decreased below 1.88, a relaxation factor much too large in many problems.

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

$$(3.6) \quad \|\mathbf{F}_{n+1} - e^{-h_{n+1}}\mathbf{F}_n\| \leq 10^{-s}e^{-h_{n+1}}\|\mathbf{F}_n\|,$$

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.

We now obtain 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 \mathbf{F} to always increase. We thus obtain the following lemma.

LEMMA 3.1. *Let $s \geq 0$ and assume that h_i , $i = 1, 2, \dots$, was chosen so that (3.6) holds for each i . Then, convergence will be guaranteed if $\liminf_{i \geq 1} h_i > \ln(1 + 10^{-s})$.*

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

$$\|\mathbf{F}_1\| = e^{-h_1}\|\mathbf{F}_0\|(1 + \varepsilon_1),$$

where $|\varepsilon_1| \leq 10^{-s}$. Again,

$$\begin{aligned}\|\mathbf{F}_2\| &= e^{-h_2}\|\mathbf{F}_1\|(1 + \varepsilon_2) \\ &= e^{-h_2}(e^{-h_1}\|\mathbf{F}_0\|(1 + \varepsilon_1))(1 + \varepsilon_2) \\ &= e^{-(h_1+h_2)}\|\mathbf{F}_0\|(1 + \varepsilon_1)(1 + \varepsilon_2).\end{aligned}$$

Clearly now

$$\begin{aligned}\|\mathbf{F}_n\| &= \exp\left(-\sum_{i=1}^n h_i\right)\|\mathbf{F}_0\| \prod_{i=1}^n (1 + \varepsilon_i) \\ (3.7) \quad &= \|\mathbf{F}_0\| \prod_{i=1}^n \frac{1 + \varepsilon_i}{e^{h_i}} \\ &\leq \|\mathbf{F}_0\| \prod_{i=1}^n \frac{1 + 10^{-s}}{e^{h_i}}.\end{aligned}$$

We may guarantee convergence if all of the terms of the product in (3.7) are eventually strictly less than one, i.e.,

$$e^{h_i} > 1 + 10^{-s} \quad \text{for all } i > I$$

or

$$h_i > \ln(1 + 10^{-s}) \quad \text{for all } i > I,$$

where I is a nonnegative integer. This completes the proof.

Note that since s determines the h_i , it may not be possible to satisfy the hypotheses.

We are now ready to consider a Kantorovich analysis for the algorithms developed above. It does not seem reasonable to attempt the whole class in such an analysis and so only three of the algorithms are analyzed. These three, PECE, PE_BCE and PE_BCE_B , contain the three basic types of iteration which can occur in the class. The proofs are not presented here (they may be found in [2]) because they do not differ substantially from the usual Kantorovich proof for the convergence of Newton's method. The essential idea in the proof is to avoid using the Cauchy inequality but rather use the law of cosines to express the norm of the sum and difference of two vectors. The theorems for convergence of the two methods involving Broyden's method also require the use of a theorem of Dennis [14] which bounds the rate of deterioration of the approximation to the Jacobian.

In the following theorems, it is not surprising that the PECE algorithm converges in the full Kantorovich region ($h_0 \leq \frac{1}{2}$, cf. Theorem 3.2 below), but it is surprising that the PE_BCE algorithm also converges in the full Kantorovich region. The explanation for this is that the final evaluation prevents the deterioration allowed by Dennis' theorem to take place. For the PE_BCE_B algorithm, because of difficulties arising from the use of that theorem we can only show convergence in a region "half" as large as expected. (See the note after Theorem 3.3.)

THEOREM 3.2. Assume $\mathbf{F} \in C'(D)$,

$$(i) \quad \|\mathbf{J}_0^{-1}\mathbf{F}_0\| \leq \eta_0,$$

$$(ii) \quad \|J(\mathbf{x}) - J(\mathbf{y})\| \leq K\|\mathbf{x} - \mathbf{y}\| \text{ for all } \mathbf{x}, \mathbf{y} \in \Omega$$

$$= \{\mathbf{x}: \|\mathbf{x} - \mathbf{x}_0\| \leq 2\eta_0\} \subset D,$$

$$(iii) \quad \|J_0^{-1}\| \leq \beta_0,$$

$$(iv) \quad \beta_0 K \eta_0 \equiv h_0 \leq \frac{1}{2}.$$

Then the PECE and PE_B CE algorithms converge uniquely to a root $\mathbf{x}^* \in \Omega$.

THEOREM 3.3. Assume the hypotheses of Theorem 3.2 except that $h_0 \leq 1/16$. Then the PE_B CE $_B$ algorithm converges to a unique root $\mathbf{x}^* \in \Omega$.

Note. Dennis [14] has proved that Broyden's method will converge for $h_0 \leq 1/8$, thus the use of the term "half" in the above discussion.

4. Numerical results. Several of the algorithms derived in § 3 have been programmed and tested on a set of examples. It is the purpose of this section 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 § 3. The two non- A -stable routines provide counter examples to the Gavurin claim that numerical stability is implied by the mathematical stability of the problem. The behavior of the methods of § 3 on the two examples is rather typical and thus further examples are not reported.

The algorithms of § 3 which we use in our tests are the three basic algorithms, PECE, PE_B CE and PE_B CE $_B$, and the three basic algorithms with a final correction, PECEC, PE_B CEC and PE_B CE $_B$ C. Each of these methods is run with and without the accuracy test of § 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 [26] (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 $e^{-h}\mathbf{F}_n$ as the approximation to \mathbf{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.

The example we choose to analyze in depth is the two \times two equation

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

with initial guess $\mathbf{x}_0 = (1, 0)$. The solution we seek is at $\mathbf{x}^* = (0, 1)$. The Jacobian of this system is

$$\mathbf{J}(\mathbf{x}) = \begin{bmatrix} 2x_1 & -1 \\ 1 & \frac{\pi}{2} \sin\left(\frac{\pi}{2}x_2\right) \end{bmatrix},$$

and the determinant of the Jacobian is given by

$$\det(\mathbf{J}(\mathbf{x})) = x_1 \pi \sin\left(\frac{\pi}{2} x_2\right) + 1.$$

The Jacobian is thus singular for

$$\sin\left(\frac{\pi}{2} x_2\right) = -\frac{1}{\pi x_1}.$$

The lines of singularity are plotted in Fig. 2. Also plotted in Fig. 2 are \mathbf{x}_0 , \mathbf{x}^* , another root \mathbf{x}^{**} and the solution curve of the initial value problem projected onto the x_1 , x_2 -plane. Since this solution curve tends to \mathbf{x}^* , \mathbf{x}^* is the root we are seeking. Newton's method, however, does not converge to \mathbf{x}^* , but rather, it crosses a line of singularities and converges in eight iterations to $\mathbf{x}^{**} = (-\frac{1}{2}\sqrt{2}, \frac{3}{2})$. The iterates generated by Newton's method display the classical form of instability reported by Gear in [20]; i.e., the first two iterates oscillate unstably around the true solution curve. The second iterate is outside of the asymptotic stability region of \mathbf{x}^* . From there, the third iterate crosses a singularity and enters the asymptotic stability region of \mathbf{x}^{**} . See Fig. 3 for a two-dimensional representation of this process.

The damped Newton's method was applied to this problem and it converged to the root \mathbf{x}^* in 107 iterations or 321 function evaluations. (Compare to Table 1.)

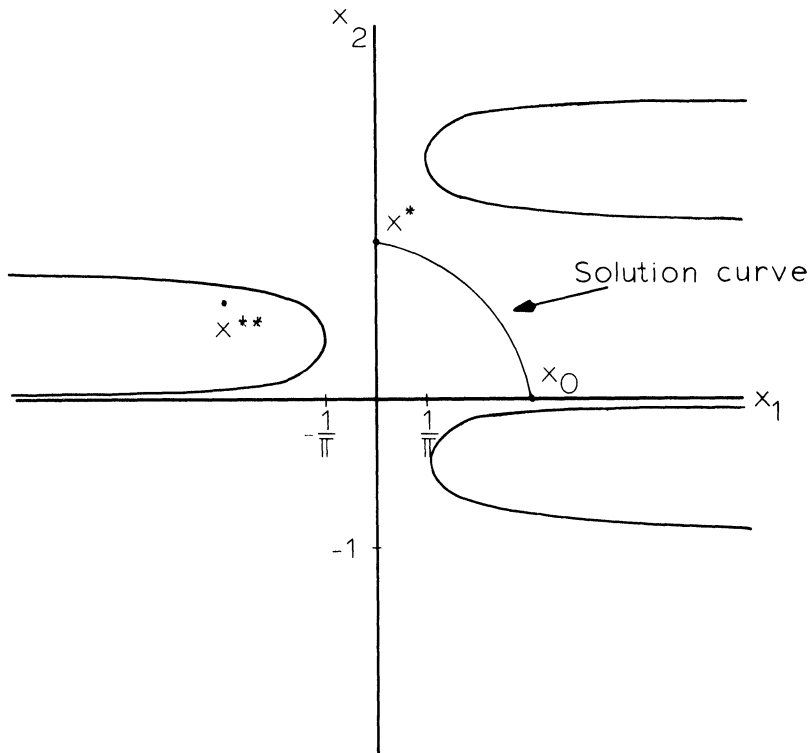


FIG. 2

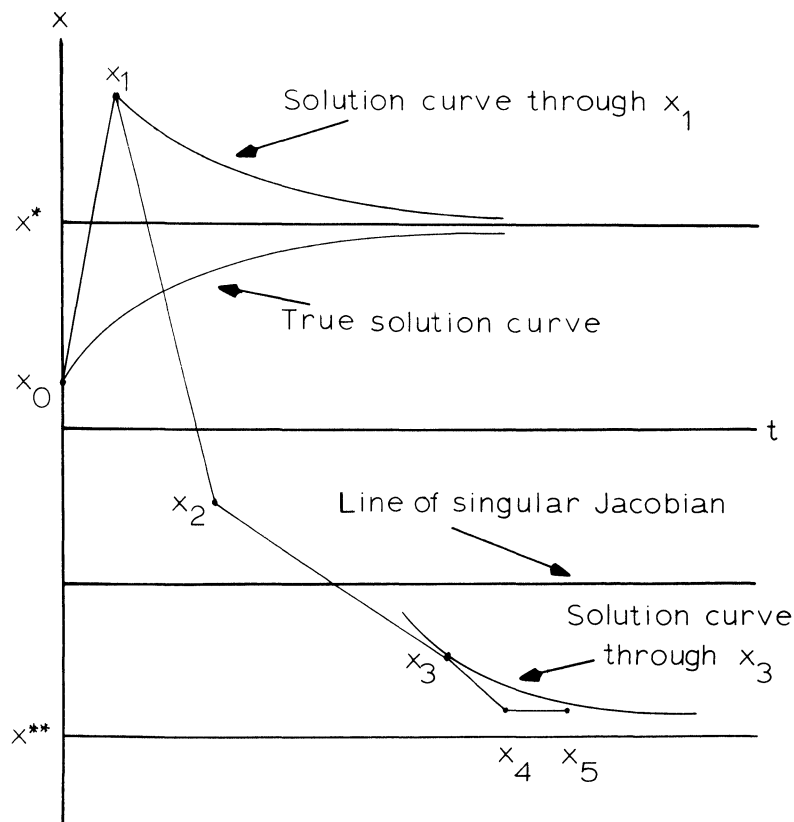


FIG. 3

TABLE I

Algorithm	Accuracy	No. Funct.	No.	Total No.
	Test	Evals./Iter.	Iterations	Funct. Evals.
PECE	Yes	6	47	282
PECE	No	6	23	138
PE _B CE	Yes	4	101 +	—
PE _B CE	No	4	17	68
PE _B CE _B	Yes	2	101 +	—
PE _B CE _B	No	2	17	34 + 2 = 36
PECEC	Yes	6	26	156
PECEC	No	6	40	240
PE _B CEC	Yes	4	28	112
PE _B CEC	No	4	17	68
PE _B CE _B C	Yes	2	101 +	—
PE _B CE _B C	No	2	101 +	—

When Nordsieck's method was applied to the problem, the iterates were able to get very close to \mathbf{x}^* , i.e., to (0.06, 0.99), but not consistently closer. The procedure was allowed to go on for another 275 iterations and it only reached the point (0.007, 0.98). The modified Nordsieck method was able to generate iterates which "almost" converged and then "blew-up." The iterates moved towards \mathbf{x}^* until finally the 40th iterate crossed the asymptote. The next iterate recrossed the asymptote and was an equally good approximation to \mathbf{x}^* . Each iterate after that again recrossed the asymptote but with ever increasing error. Figure 4 graphically illustrates this situation.

We next report the results of the application of the algorithms of § 3 to the problem. These results are summarized in Table 1 and discussed below. For each algorithm, with and without the accuracy test, Table 1 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 $1/32$. When no accuracy test was used, the step size was kept constant at one.

The first observation is that many of these methods were successful and that the method requiring the least amount of work was the $PE_B CE_B$ 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,

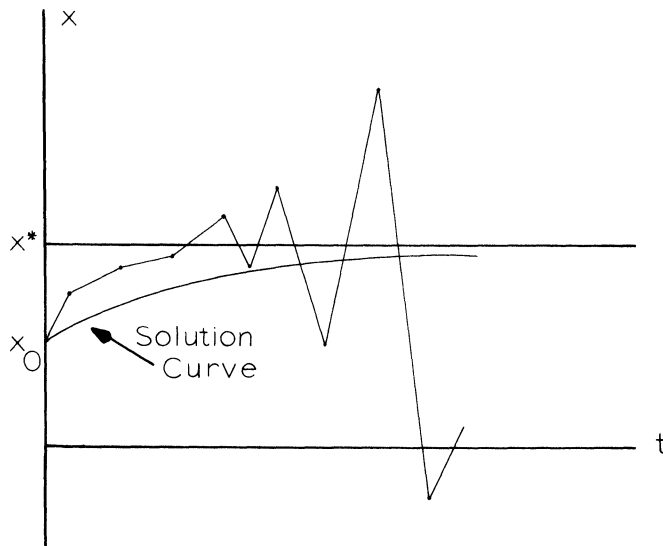


FIG. 4

to Fig. 4, 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 27 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 [6]:

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

with initial approximation (0.4, 3). There is a solution at $(0.5, \pi)$, but this is not the solution we are seeking. Newton's method converges to the root $(-0.26, 0.62)$ in five iterations, but this also is not the correct root. The asymptote of the solution of the initial value problem

$$\mathbf{x}' = -\mathbf{J}^{-1}(\mathbf{x})\mathbf{F}(\mathbf{x}), \quad \mathbf{x}(0) = (0.4, 3)$$

is $\mathbf{x}^* = (0.30, 2.8)$. Application of the methods of §3 to the problem yield the results in Table 2. (Table 2 is organized exactly as Table 1.)

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 $(-0.26, 0.62)$.

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

TABLE 2

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

the change that the Jacobian was reevaluated after every 4 iterations. In this case, we obtained convergence in 14 iterations or 38 function evaluations.

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 § 3 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 that the region of convergence of the PE_BCE algorithm is the same as the region for the PECE algorithm. An examination of Theorem 3.5 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_0 of Theorem 3.4 can be made by using an approximation to the Lipschitz constant K based on two successive iterates ($K \approx \|\mathbf{J}_n - \mathbf{J}_{n-1}\|/\|\mathbf{x}_n - \mathbf{x}_{n-1}\|$) and bounding the norm of \mathbf{J}_n^{-1} and $\mathbf{J}_n^{-1}\mathbf{F}_n$. When the approximation to h_0 is less than $\frac{1}{2}$, 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.

5. Conclusion. In summary we have shown that the linear A -stable methods of Dahlquist, in particular the trapezoidal rule, have proved to be effective new ways of solving nonlinear systems of equations. There are, however, several questions which present themselves as possible sources of other algorithms. In particular, is there any advantage in considering the stiffly stable methods of Gear mentioned in § 3? 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? Also, is there any advantage in considering nonlinear A -stable methods such as those developed by Loscalzo [24] using spline functions?

6. Acknowledgments. The author would like to thank his thesis advisor J. E. Dennis, Jr. for many helpful suggestions and much encouragement throughout the preparation of this work. He would also like to thank R. A. Sweet, R. J. Walker and K. M. Brown for several discussions about the material.

The author also gratefully acknowledges the careful reading and excellent suggestions of the referees, in particular, those suggestions regarding Theorem 2.7.

REFERENCES

- [1] P. M. ANSELONE AND R. H. MOORE, *An extension of the Newton-Kantorovich method for solving nonlinear equations with an application to elasticity*, Tech. Rep. 520, Mathematics Research Center, Madison, Wisc., 1965.
- [2] P. T. BOGGS, *The solution of nonlinear operator equations by A -stable integration techniques*, Doctoral thesis, Cornell University, Ithaca, N.Y., 1970.
- [3] W. E. BOSARGE, JR., *Infinite dimensional iterative methods and applications*, Publication 320-2347, IBM, Houston, Texas.
- [4] C. G. BROYDEN, *A class of methods for solving nonlinear simultaneous equations*, Math. Corp., 19 (1965), pp. 577–593.
- [5] ———, *Quasi-Newton methods and their application to function minimization*, Math. Comp., 21 (1967), pp. 368–381.
- [6] ———, *A new method of solving nonlinear simultaneous equations*, Comput. J., 12 (1969), pp. 94–99.
- [7] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.
- [8] G. G. DAHLQUIST, *A special stability problem for linear multistep methods*, B.I.T., 3 (1963), pp. 27–43.
- [9] D. F. DAVIDENKO, *On a new method of numerical solution of systems of nonlinear equations*, Dokl. Akad. Nauk USSR (N.S.), 88 (1953), pp. 601–602.
- [10] W. C. DAVIDON, *Variable metric method for minimization*, A.E.C. Research and Development Rep. ANL-5990 (Rev. TID-4500, 14th ed.), 1959.
- [11] J. DAVIS, *The solution of nonlinear operator equations with critical points*, Doctoral thesis, Oregon State University, Corvallis, 1966.
- [12] F. H. DEIST AND L. SEFOR, *Solution of systems of nonlinear equations by parameter variation*, Comput. J., 10 (1967), pp. 78–82.
- [13] J. E. DENNIS, *On the convergence of Newton-like methods*, Numerical Methods for Nonlinear Algebraic Equations, P. Rabinowitz, ed., Gordon and Breach, London, 1970.
- [14] ———, *On the convergence of Broyden's method for nonlinear systems of equations*, Tech. Rep. 69-48, Cornell Univ., Ithaca, N.Y., 1969.
- [15] ———, *On the local convergence of Broyden's method for nonlinear systems of equations*, Tech. Rep. 69-46, Cornell Univ., Ithaca, N.Y., 1969.
- [16] J. E. DENNIS AND R. A. SWEET, *Some minimal properties of the trapezoidal rule*, Tech. Rep. 70-61, Cornell Univ., Ithaca, N.Y., 1970.
- [17] F. A. FICKEN, *The continuation method for functional equations*, Comm. Pure Appl. Math., 4 (1951), pp. 435–456.
- [18] F. FREUDENSTEIN AND B. ROTH, *Numerical solutions of systems of nonlinear equations*, Assoc. Comput. Mech., 10 (1963), pp. 550–556.
- [19] M. K. GAVURIN, *Nonlinear functional equations and continuous analogs of iterative methods*, Izv. Vyss. Uchebn. Zaved. Matematika, 6 (1958), pp. 18–31; English Transl., Tech. Rep. 68-70, Univ. of Maryland, College Park, 1968.
- [20] C. W. GEAR, *Numerical integration of stiff ordinary differential equations*, Tech. Rep. 221, Univ. of Illinois, Urbana, 1967.

- [21] P. HENRICI, *Discrete Variable Methods in Ordinary Differential Equations*, John Wiley, New York, 1962.
- [22] T. E. HULL AND A. L. CREEMER, *Efficiency of predictor-corrector procedures*, Assoc. Comput. Mach., 10 (1963), pp. 291–301.
- [23] W. KIZNER, *A numerical method for finding solutions of nonlinear equations*, SIAM Appl. Math., 12 (1964), pp. 424–428.
- [24] F. R. LOSCALZO, *An introduction to the applications of spline functions to initial value problems*, Theory and Applications of Spline Functions, T. N. E. Grenville, ed., Academic Press, New York, 1969, pp. 37–64.
- [25] G. H. MEYER, *On solving nonlinear equations with a one-parameter operator imbedding*, this Journal, 5 (1968), pp. 739–752.
- [26] A. NORDSIECK, *On numerical integration of ordinary differential equations*, Math. Comp., 16 (1962), pp. 22–49.
- [27] N. N. YAKOVLEV, *On the solution of systems of nonlinear equations by differentiation with respect to a parameter*, U.S.S.R. Comput. Math. and Math. Phys., 4 (1964), pp. 146–149.
- [28] ———, *On certain methods of solution of nonlinear equations*, Trudy Mat. Inst. Steklov., 84 (1965), pp. 8–40; English Transl., Tech. Rep. 68-75, Univ. of Maryland, College Park, 1968.