J. Austral. Math. Soc. 19 (Series B), (1975), 173–199.

FAST LOCAL CONVERGENCE WITH SINGLE AND MULTISTEP METHODS FOR NONLINEAR EQUATIONS

JAMES P. ABBOTT and RICHARD P. BRENT

(Received 14 November 1975)

Abstract

Methods which make use of the differential equation $\dot{x}(t) = -J(x)^{-1}f(x)$, where $J(x)$ is the Jacobian of $f(x)$, have recently been proposed for solving the system of nonlinear equations $f(x) = 0$. These methods are important because of their improved convergence characteristics. Under general conditions the solution trajectory of the differential equation converges to a root of f and the problem becomes one of solving a differential equation. In this paper we note that the special form of the differential equation can be used to derive single and multistep methods which give improved rates of local convergence to a root.

1. Introduction

Recently some interest has been shown in methods for the solution of a system of nonlinear equations $f(x) = 0$, where $f : D \subset R^n \to R^n$, when only a poor initial estimate of a zero, x^*, of f is known. One approach is to define a differential equation whose solution $x(t) \to x^*$ as $t \to \infty$, where t is an independent variable. $x(t)$ then defines a trajectory which converges to the required solution and one can consider any method for solving the differential equation as a means of following the trajectory to that solution. One such differential equation, originally suggested by Davidenko [9], is

$$\dot{x}(t) = -J(x)^{-1}f(x), \qquad x(0) = x_0, \tag{1.1}$$

where $J(x)$, the Jacobian of $f(x)$, is assumed to be nonsingular at x^*. One can look at this differential equation in various ways, see for example Ortega and Rheinboldt [24, §7.5] or Branin [5], but perhaps the simplest is as a continuous Newton's method as described by Gavurin [12], and the solution $x(t)$ of (1.1) can be considered as the continuous Newton trajectory.

An alternative, but related, approach which has been investigated by several authors (see, for example, [6], [18], [19], [24]) is to integrate the differential equation

$$\dot{x}(t) = -J(x)^{-1}f(x_0), \qquad x(0) = x_0, \tag{1.2}$$

over the interval $[0, 1]$. However, in an extensive treatment of the subject by Boggs [4], it has been noted that integrating (1.2) to find a good estimate of $x(1)$ demands a greater concern for accuracy than is necessary when integrating (1.1). This fact is confirmed in Section 5 and we find that certain methods which make use of (1.1) are more reliable and, excepting simple cases, are generally more efficient than those which use (1.2).

When the solution $x(t)$ of (1.1) converges to x^* any method which, because of small steps or high accuracy, follows the trajectory sufficiently closely will surely converge to x^* also. However this convergence will be slow since $x(t)$ converges to x^* only linearly. Thus, for an algorithm to be efficient there must be a change of emphasis at some stage from accurate representation of $x(t)$ to rapid convergence to x^*. In this paper we consider methods for the solution of the differential equation (1.1) which can, by suitable step length control, be induced to give rapid final convergence to x^*.

In Section 2 we discuss the differential equation with regard to regions of convergence of the trajectory $x(t)$ to x^*. In Section 3 we present some general results on the convergence of one step methods with variable step size and use these results to derive methods for the solution of (1.1) which can give rapid final convergence to x^*. Also we discuss briefly a conjecture of Boggs [4] that the most suitable methods for the solution of (1.1) are the A-stable methods of Dahlquist [8]. In Section 4 we present general results on the convergence of multistep methods and use the results to generate methods which can give rapid final convergence to x^*. We also discuss the stability problems involved with such methods if the step size is varied. In Section 5 we apply the methods suggested by the theory to a number of problems and draw conclusions about their relative merits.

2. The differential equation

We consider the differential equation (1.1), where $f(x)$ is continuously differentiable for all $x \in D$. There are a great many theorems on the existence and uniqueness of solutions of (1.1) (see for example [1], [3], [22], [24], [27], [29] and the references therein) but most are local in nature. Since the differential equation approach is concerned with wider convergence we present a theorem which is not local. The theorem is not new, having been proved with marginally greater assumptions on f by Gavurin [12], Deuflhard

[11] and Ortega and Rheinboldt [24], but is given for clarity and as motivation for the overall approach. First we give some definitions.

DEFINITION 2.1. $P \subset D$ is *a region of stability* of (1.1) if, for any $x_0 \in P$, the solution $x(t)$ of (1.1) is defined and unique for all $t \geqq 0$, $x(t) \in P$ for all $t \geqq 0$ and $\lim_{t\to\infty} x(t) = x^* \in P$, where x^* is a zero of f.

For any nonsingular $n \times n$ matrix A define $\phi_A : D \subset R^n \to R$ by

$$\phi_A(x) = f(x)^T A^T A f(x)$$

and, for any $\alpha > 0$, define $P_\alpha(A)$ by

$$P_\alpha(A) = \{x \mid x \in D, \phi_A(x) \leqq \alpha\}.$$

$P_\alpha(A)$ is a level set of $\phi_A(x)$, (see [11], [24]). Let $L = \{x \mid x \in D, \operatorname{Det}(J(x)) = 0\}$. Then, for some $\alpha > 0$ and $P^*_\alpha(A)$, a path connected component of $P_\alpha(A)$, condition $\mathscr{A}$ will be

$$\mathscr{A}: P^*_\alpha(A) \cap L \quad \text{and} \quad P^*_\alpha(A) \cap \partial D \quad \text{are empty,} \quad P^*_\alpha(A) \text{ is bounded.}$$

Under these conditions $P^*_\alpha(A)$ is compact and contains one and only one zero of f.

THEOREM 2.1. *Assume $f : D \subset R^n \to R^n$ is continuously differentiable on D and $\alpha > 0$ is such that condition $\mathscr{A}$ holds. If in addition $J(x)^{-1}f(x)$ is Lipschitz continuous on* $\operatorname{Int}(P^*_\alpha(A))$ *then* $\operatorname{Int}(P^*_\alpha(A))$ *is a region of stability of* (1.1).

PROOF. Standard theorems on ordinary differential equations (e.g. [16, Chapter 1]) show that, for any $x_0 \in \operatorname{Int}(P^*_\alpha(A))$, there exists a $\tau > 0$ such that (1.1) has a solution which is unique in $\operatorname{Int}(P^*_\alpha(A))$ for each $t \in [0, \tau)$. If the maximal such τ is not ∞ and $\{x(t) \mid 0 \leqq t < \tau\}$ has limit point x_τ, then $x_\tau \in \partial P^*_\alpha(A)$.

When the solution $x(t)$ of (1.1) exists it satisfies

$$f(x(t)) = e^{-t} f(x_0) = e^{-t} f_0, \tag{2.1}$$

say, because (1.1) is equivalent to the initial value problem $df/dt = -f$, $f(0) = f_0$. Thus

$$\phi_A(x(t)) = e^{-2t}\phi_A(x_0), \qquad t \in [0, \tau),$$

and so $\phi_A(x(t))$ is a decreasing function of t. Thus

$$\phi_A(x_\tau) = \lim_{t\to\tau-} \phi_A(x(t)) < \alpha.$$

Now suppose, if possible, that $x_\tau \in \partial P^*_\alpha(A)$. Since $P_\alpha(A)$ is closed and $P^*_\alpha(A) \cap \partial D$ is empty there exists an $\varepsilon > 0$ such that $S(x_\tau, \varepsilon) \subset D$ and $S(x_\tau, \varepsilon) \cap \{P_\alpha(A) \setminus P^*_\alpha(A)\}$ is empty, where $S(x, \varepsilon)$ is the open ball with centre x

https://doi.org/10.1017/S0334270000001077 Published online by Cambridge University Press

and radius ε. Let $\varepsilon_i = \varepsilon/i$, then because $x_\tau \in \partial P^*_\alpha(A)$, for each $i > 0$ there exists a $y_i \in S(x_\tau, \varepsilon_i)$ such that $\phi_A(y_i) > \alpha$. Now $\lim_{i\to\infty} y_i = x_\tau$ and, by continuity of $\phi_A(x)$, $\lim_{i\to\infty}\phi_A(y_i) = \phi_A(x_\tau) \geqq \alpha$, which is a contradiction. Thus $x_\tau \in \mathrm{Int}(P^*_\alpha(A))$ and it follows that $\tau = \infty$, so $x(t)$ is defined and $x(t) \in \mathrm{Int}(P^*_\alpha(A))$ for all $t \geqq 0$. Also, from (2.1), if x_∞ is a limit point of $\{x(t)\}$, then $f(x_\infty) = 0$. Since a zero of f is unique in $P^*_\alpha(A)$ it follows that $x_\infty = x^* = \lim_{t\to\infty} x(t)$. This completes the proof.

We note that a sufficient condition for $J(x)^{-1}f(x)$ to be Lipschitz continuous on $\mathrm{Int}(P^*_\alpha(A))$ is that, in addition to condition $\mathscr{A}$, $J(x)$ be Lipschitz continuous on $\mathrm{Int}(P^*_\alpha(A))$. This follows from the fact that $\|J(x)^{-1}\|$ and $\|f(x)\|$ are bounded on $P^*_\alpha(A)$ and $f(x)$ is continuously differentiable (and hence Lipschitz continuous) on $P^*_\alpha(A)$.

Whilst Theorem 2.1 is not practically useful it shows that around each zero at which $J(x)$ is nonsingular there is a region of stability of (1.1). Also this region will generally be larger than that predicted by the local existence theorems. We note that if x_0 is not in such a region then convergence to a root is unpredictable and the reader is referred to [5] and [7] for progress in this case.

For the remainder of this paper we assume that x_0 is contained in a region of stability and that the solution trajectory converges to a zero x^*. If this is the case then, by following the trajectory closely enough, we can guarantee convergence to x^*. For this purpose any stable method of solving an initial value problem may be employed and, for sufficiently small steps, convergence to x^* is certain. In practice, however, we would like to take large steps. Far from the zero this entails using a sophisticated step size estimator which will adapt the step according to the function behaviour and choose it to be as large as possible consistent with sufficient accuracy. Obviously the lower the accuracy the less work will be involved but the higher the probability of leaving the correct trajectory and diverging or finding the wrong solution.

Close to the solution, however, we can make use of the special characteristics of the problem to give rapid final convergence, using methods which are also suitable for following the trajectory far from the solution. In the following two sections we consider single and multistep methods, traditionally used for the standard initial value problem, which are adapted to give rapid convergence close to the zero x^*.

3. Single step methods

3.1. General theory

In this section we give some general results on iterative processes of the form

https://doi.org/10.1017/S0334270000001077 Published online by Cambridge University Press

$$x_{i+1} = G(x_i, h_i), \qquad i = 0, 1, \cdots \tag{3.1}$$

where $G : D \times D_h \subset R^n \times R \to R^n$, and in the following sections we apply these results to particular iterations. We use the results of Ostrowski [26] and Ortega and Rockoff [25] on processes of the form $x_{i+1} = G(x_i)$, $G : D \subset R^n \to R^n$, and generalize the existing theory to include the extra variable. We quote the following definitions which can be found in [24], except that here suitable modification has been made to allow for the slight generalization.

Let $C(\mathcal{I}, x^*)$ denote the set of all sequences generated by an iterative process $\mathcal{I}$ with limit point x^*. Let $\{x_k\} \subset R^n$ be any sequence that converges to x^*. Then the *R-convergence factors of the sequence* are the numbers

$$R_p\{x_k\} = \begin{cases} \limsup_{k\to\infty} \|x_k - x^*\|^{1/k}, & \text{if} \quad p = 1 \\ \\ \limsup_{k\to\infty} \|x_k - x^*\|^{1/p^k}, & \text{if} \quad p > 1. \end{cases}$$

The *R-convergence factor of* $\mathcal{I}$ *at* x^* is defined by

$$R_p(\mathcal{I}, x^*) = \sup\{R_p\{x_k\} \mid \{x_k\} \in C(\mathcal{I}, x^*)\}$$

and the quantity

$$O_R(\mathcal{I}, x^*) = \begin{cases} \infty \quad \text{if} \quad R_p(\mathcal{I}, x^*) = 0 \quad \text{for all} \quad p \in [1, \infty), \\ \\ \inf\{p \in [1, \infty) \mid R_p(\mathcal{I}, x^*) = 1\} \quad \text{otherwise} \end{cases}$$

is called the *R-order of* $\mathcal{I}$ *at* x^*. We say that the convergence of $\mathcal{I}$ at x^* is *superlinear* if $R_1(\mathcal{I}, x^*) = 0$.

Let $G : D \times D_h \subset R^n \times R \to R^n$, then x^* is *a point of attraction* of the iterative process (3.1) if there exists an open neighbourhood S of x^* and a set I, called the *h-domain of* $\mathcal{I}$, such that $S \subset D$, $I \subset D_h$ and for any $x_0 \in S$ and any $\{h_i\} \subset I$ the iterates $\{x_i\}$ remain in D and converge to x^*. Also we say that x^* is a *fixed point* of the iteration (3.1) if $x^* = G(x^*, h)$ for all $h \in D_h$.

We can now give conditions on $G(x, h)$ which are sufficient for a point x^* to be a point of attraction of (3.1).

THEOREM 3.1. *Suppose that* $G : D \times D_h \subset R^n \times R \to R^n$ *has a fixed point* x^* *and* $\partial_x G(x, h)$ *exists and is continuous in a neighbourhood of* (x^*, h) *for each* $h \in D_h$. *Let* $I_\alpha \subset D_h$ *be such that, for some norm,* $\|\partial_x G(x^*, h)\| \leqq \alpha < 1$ *for each* $h \in I_\alpha$. *If* I_α *is non empty then* x^* *is a point of attraction of iteration* (3.1) *with h-domain* I_α.

https://doi.org/10.1017/S0334270000001077 Published online by Cambridge University Press

(Note that here and subsequently $\partial_x G(x, h)$ denotes the Fréchet partial derivative of G with respect to x.)

PROOF. Using the continuity of $\partial_x G(x^*, h)$ it follows from [24, Theorem 5.2.3] that, given $\varepsilon > 0$, for any $h \in I_\alpha$ there exists a $\delta > 0$ such that

$$\| G(x, h) - G(x^*, h) - \partial_x G(x^*, h)(x - x^*) \| \leqq \varepsilon \| x - x^* \|$$

for all $x \in S(x^*, \delta)$. Now

$$\begin{aligned} \| x_{i+1} - x^* \| &= \| G(x_i, h_i) - G(x^*, h_i) \| \\ &= \| G(x_i, h_i) - G(x^*, h_i) - \partial_x G(x^*, h_i)(x_i - x^*) \\ &\quad + \partial_x G(x^*, h_i)(x_i - x^*) \| \\ &\leqq \| G(x_i, h_i) - G(x^*, h_i) - \partial_x G(x^*, h_i)(x_i - x^*) \| \\ &\quad + \| \partial_x G(x^*, h_i) \| \, \| x_i - x^* \|. \end{aligned}$$

Thus, if $h_i \in I_\alpha$ for each i,

$$\| x_{i+1} - x^* \| \leqq (\varepsilon + \alpha) \| x_i - x^* \|.$$

Since $\alpha < 1$ we may assume that ε was chosen so that $\varepsilon + \alpha < 1$ and the result follows.

The following example shows that the condition $\| \partial_x G(x^*, h) \| \leqq \alpha < 1$ cannot in general be replaced by $\eta(\partial_x G(x^*, h)) \leqq \alpha < 1$, where $\eta(\cdot)$ denotes the spectral radius. If $G(x, h)$ is defined by

$$G(x, h) = \begin{bmatrix} \alpha x_1 + x_2/(1 - h) \\ \alpha x_2 \end{bmatrix}$$

and $\alpha < 1$ then $\eta(\partial_x G(x, h)) = \alpha$ for all h, but the iteration (3.1) does not converge, even locally, if h_i converges to 1 sufficiently fast. However, a corollary to Theorem 3.1 will be useful and gives a case when $\eta(\partial_x G(x^*, h))$ can replace $\| \partial_x G(x^*, h) \|$ in the theorem.

COROLLARY 3.1. *Suppose G satisfies the conditions of Theorem* 3.1. *Suppose also that there is a set $I \subset D_h$ such that $\partial_x G(x^*, h)$ can be expressed as*

$$\partial_x G(x^*, h) = \phi(h) A$$

for each $h \in I$, where A is a fixed matrix and $\phi : D_h \subset R \to R$. If $I_\alpha = \{h \in I \mid \eta(\partial_x G(x^*, h)) \leqq \alpha < 1\}$ is non empty then x^* is point of attraction of (3.1) *with h-domain I_α.*

Theorem 3.1 gives sufficient conditions for local convergence of the iterative process (3.1) but gives no information on the rate of convergence.

https://doi.org/10.1017/S0334270000001077 Published online by Cambridge University Press

For this we require conditions on $\{h_i\}$. We begin by deriving a result on the assumption that $\lim_{i\to\infty} h_i$ exists.

THEOREM 3.2. *Suppose* $G : D \times D_h \subset R^n \times R \to R^n$ *has a fixed point* $x^* \in \mathrm{Int}(D)$ *and is continuous in an open neighbourhood* S *of* (x^*, h^*), *where* $\lim_{i\to\infty} h_i = h^*$. *Suppose also that* $\partial_x G(x, h)$ *exists in* S *and is continuous at* (x^*, h^*). *If the spectral radius of* $\partial_x G(x, h)$ *satisfies*

$$\eta = \eta(\partial_x G(x^*, h^*)) < 1$$

then x^* *is a point of attraction of the iterative process* $\mathcal{G}$ *given by* (3.1). *Moreover*

$$R_1(\mathcal{G}, x^*) = \eta$$

and if $\eta > 0$ *then* $O_R(\mathcal{G}, x^*) = 1$.

PROOF. The proof is omitted since it follows closely those given by Ortega and Rheinboldt [24, Theorems 10.1.3, 10.1.4 and 10.1.7] except for modifications to allow for the extra variable h.

To complete the theoretical background we consider the possibility of faster convergence in the case when $\eta(\partial_x G(x^*, h^*)) = 0$. For this case we require further knowledge of the sequence $\{h_i\}$.

THEOREM 3.3. *Suppose* $G : D \times D_h \subset R^n \times R \to R^n$ *satisfies the conditions of Theorem* 3.2 *and that* $\eta(\partial_x G(x^*, h^*)) = 0$. *Suppose also that* $\partial_{xx} G(x, h)$ *exists and is continuous and bounded in a neighbourhood* S *of* (x^*, h^*) *and that* $\partial_{xh} G(x, h)$ *exists and is bounded on* S. *Then* x^* *is a point of attraction of the iterative process* $\mathcal{G}$ *given by* (3.1) *and* $R_1(\mathcal{G}, x^*) = 0$. *If, in addition,* $\{h_i\}$ *converges to* h^* *with* R*-order* $q \geqq 1$ *then* $O_R(\mathcal{G}, x^*) \geqq \min(2^{1/m}, q)$, *where* $\partial_x G(x^*, h^*)^m = 0$, $\partial_x G(x^*, h^*)^{m-1} \neq 0$.

PROOF. Theorem 3.2 shows that x^* is a point of attraction of $\mathcal{G}$ and that $R_1(\mathcal{G}, x^*) = 0$.

Define $u(x, h)$ by

$$G(x, h) = G(x^*, h) + \partial_x G(x^*, h)(x - x^*) + u(x, h).$$

Then, using the existence and boundedness of $\partial_{xx} G(x, h)$, it follows from [24, Theorem 3.3.6] that there exist positive constants δ_1, δ_2 and K_1 such that

$$\|u(x, h)\| \leqq K_1 \|x - x^*\|^2$$

for all $x \in S(x^*, \delta_1)$, $h \in (h^* - \delta_2, h^* + \delta_2) = I_2$, say. Similarly, from [24, Theorem 3.2.3], with $D(h)$ defined by

$$D(h) = \partial_x G(x^*, h) - \partial_x G(x^*, h^*),$$

https://doi.org/10.1017/S0334270000001077 Published online by Cambridge University Press

then, if δ_1, δ_2 are sufficiently small, there exists a $K_2 > 0$ such that

$$\|D(h)\| \leqq K_2 |h - h^*|$$

for all $x \in S(x^*, \delta_1)$, $h \in I_2$. Finally let $A = \partial_x G(x^*, h^*)$. Then $\eta(A) = 0$ and there is an integer $m \leqq n$ such that $A^{m-1} \neq 0$, $A^m = 0$.

Now

$$\begin{aligned} G(x, h) - x^* &= G(x^*, h) + \partial_x G(x^*, h)(x - x^*) + u(x, h) - x^* \\ &= D(h)(x - x^*) + \partial_x G(x^*, h^*)(x - x^*) + u(x, h). \end{aligned}$$

Thus, if $D_i = D(h_i)$, $u_i = u(x_i, h_i)$ and $e_i = x_i - x^*$, we have

$$e_{i+1} = Ae_i + D_i e_i + u_i$$

and, by induction, for $j \geqq 0$

$$\begin{aligned} e_i = A^j e_{i-j} + A^{j-1} D_{i-j} e_{i-j} + \cdots + AD_{i-2} e_{i-2} + D_{i-1} e_{i-1} \\ + A^{j-1} u_{i-j} + \cdots + Au_{i-2} + u_{i-1}. \end{aligned}$$

With $j = m$, since $A^m = 0$, we can derive

$$\begin{aligned} \|e_i\| \leqq K_1(\gamma^{m-1} \|e_{i-m}\|^2 + \cdots + \gamma \|e_{i-2}\|^2 + \|e_{i-1}\|^2) \\ + K_2(\gamma^{m-1} \|e_{i-m}\| \varepsilon_{i-m} + \cdots + \gamma \|e_{i-2}\| \varepsilon_{i-2} + \|e_{i-1}\| \varepsilon_{i-1}) \end{aligned} \tag{3.2}$$

where $\varepsilon_j = |h_j - h^*|$ and $\gamma = \|A\|$.

Since $\{x_i\}$ converges to x^*, it follows from (3.2) that there exists an $i_0 > 0$ and constants B_1, B_2 such that, for each $i \geqq i_0$,

$$\|e_i\| \leqq B_1 \|e_{i-m}\|^2 + B_2 \|e_{i-m}\| \varepsilon_{i-m}.$$

Replacing i by mi and writing $\alpha_i = B_1 \|e_{mi}\|$ and $\beta_i = B_2 \varepsilon_{mi}$ we have

$$\alpha_i \leqq \alpha_{i-1}^2 + \alpha_{i-1} \beta_{i-1},$$

for all sufficiently large i.

Rather than give uninstructive details of the remainder of the proof, we state that, if $0 < p < \min(2, q^m)$, then there exists a $c > 0$ and $j > 0$ such that

$$\alpha_{i+1} \leqq \tfrac{1}{2} e^{-cp^i}$$

for all $i \geqq j$, where e is the base of the natural logarithm. It follows from this that the R-order of the sequence $\{\alpha_i\}$ is at least p. Since $\alpha_i = \|e_{im}\|$ and p is arbitrarily close to $\min(2, q^m)$, it follows that $O_R(\mathcal{I}, x^*) \geqq \min(2^{1/m}, q)$.

3.2. Runge–Kutta methods

Consider the general class of explicit Runge–Kutta methods for solving the differential equation

https://doi.org/10.1017/S0334270000001077 Published online by Cambridge University Press

$$\dot{x}(t) = q(x), \qquad x(0) = x_0, \tag{3.3}$$

given by

$$x_{m+1} = x_m + h_m \sum_{i=1}^{r} \alpha_i k_i(x_m), \qquad m = 0, 1, \cdots, \tag{3.4a}$$

where x_m is an approximation to $x(h_0 + h_1 + \cdots + h_{m-1})$,

$$k_i(x) = q\left(x + h_m \sum_{j=1}^{i-1} \beta_{ij} k_j(x)\right), \qquad i = 1, \cdots, r, \tag{3.4b}$$

and h_m is the step length. A discussion of stability for this method is usually based upon consideration of the linear differential equations

$$\dot{x}(t) = Ax, \qquad x(0) = x_0, \tag{3.5}$$

where A is a fixed matrix whose eigenvalues have negative real part. The true solution of (3.5) is

$$x(t + h_m) = \exp(h_m A)x(t)$$

whereas the solution given by (3.4) is

$$x_{m+1} = p(h_m A)x_m, \tag{3.6}$$

where $p(z)$ is a polynomial of degree r whose coefficients depend upon choice of the α's and β's in (3.4). The usual practice is to choose these parameters so that $p(z)$ is a good approximation to $\exp(z)$. We note that, since the true solution of (3.5) is decreasing, a requirement on the step length h_m is that the condition

$$\eta(p(h_m A)) < 1, \qquad m = 0, 1, \cdots \tag{3.7}$$

be satisfied so that the iterates in (3.6) also decrease. However, in the nonlinear case, (3.7) is of little practical use in controlling the step size.

In this section we consider (3.4) not only as a means of approximating the solution of (1.1) but also as a one-step method for finding a zero of f. For the former the theory is well known [17] and for the latter we use the results of Section 4.1. In this case we have

$$x_{m+1} = G(x_m, h_m), \qquad m = 0, 1, \cdots,$$

where

$$G(x, h) = x + h \sum_{i=1}^{r} \alpha_i k_i(x, h)$$

and, as in (3.4b),

$$k_i(x, h) = q\left(x + h \sum_{j=1}^{i-1} \beta_{ij} k_j(x, h)\right), \qquad i = 1, \cdots, r.$$

https://doi.org/10.1017/S0334270000001077 Published online by Cambridge University Press

We apply this process to the case when $q(x)$ is given by

$$q(x) = -J(x)^{-1} f(x). \tag{3.8}$$

Then, if I represents the unit matrix,

$$\partial_x G(x, h) = I + h \sum_{i=1}^{r} \alpha_i \partial_x k_i (x, h).$$

If x^* is a zero of $f(x)$ then x^* is a fixed point of (3.4) and also, from (3.8), we have

$$q'(x^*) = -I,$$

where the prime denotes differentiation with respect to x. It then follows by some simple algebra that

$$\partial_x G(x^*, h) = p(-h)I$$

where $p(z)$ is the same polynomial as appeared in (3.6). Thus, from Corollary 3.1, a sufficient condition for x^* to be a point of attraction of (3.4) is that, for some $\alpha < 1$,

$$\eta(p(-h_m)I) = |p(-h_m)| \leqq \alpha, \qquad m = 1, 2, \cdots, \tag{3.9}$$

which, unlike (3.7), provides an *explicit* bound on each h_m for ultimate convergence to x^*. It also follows from Theorem 3.2 that, if $\lim_{i\to\infty} h_i = h^*$, the iterative process can give superlinear convergence to x^* only if h^* satisfies

$$p(-h^*) = 0. \tag{3.10}$$

In the case when $f(x)$ is three times continuously differentiable it follows from Theorem 3.3 that if h_i converges to h^* with R-order $\geqq 2$, then the iterative process (3.4) has R-order at least 2.

In the application of (3.4) it is of benefit to choose the parameters so that the resulting method will follow the solution of (1.1) well enough to inhibit divergence but will also provide a fast rate of final convergence. This means choosing a method which allows h^* to be chosen so that (3.10) is satisfied. We note here that for the well-known 4th-order Runge–Kutta process $p(z)$ is defined by

$$p(z) = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \frac{z^4}{4!}$$

and $p(-z)$ has no real root. Thus no choice of h^* can furnish second order convergence. Also Heun's predictor-corrector method [17, p. 67] may be written

https://doi.org/10.1017/S0334270000001077 Published online by Cambridge University Press

$$x_{m+1} = x_m + \frac{h_m}{2}(q(x_m) + q(x_m + h_m q(x_m))). \tag{3.11}$$

This is of the class (3.4) and has $p(z)$ defined by

$$p(z) = 1 + z + \frac{z^2}{2}.$$

This is simply a Runge–Kutta method of order 2 and again $p(-z)$ has no real root, so no choice of h^* can give second order convergence to x^*.[1] In attempting to solve (1.1), Boggs [4] used this method as an explicit approximation to the trapezoidal rule.

We note that for these two methods we can use Theorem 3.2 to show that

$$O_R(\mathcal{I}, x^*) = 1$$

and

$$R_1(\mathcal{I}, x^*) = |p(-h^*)|.$$

So, assuming (3.9) is satisfied, convergence is at best linear and the fastest convergence is achieved by choosing h^* to minimise $|p(-h^*)|$. For the order two method this is $h^* = 1.0$. If the sequence $\{h_m\}$ does not satisfy (3.9), then the method will not generally converge.

Boggs [4] in his paper suggested there is a difficulty of stiffness involved in integrating (1.1). However, close to the solution at least, this is not the case, since the Jacobian matrix of the right hand side of (1.1) is close to $-I$. The symptoms of instability which Boggs ascribes to stiffness appear identical to the behaviour observed if the sequence $\{h_m\}$ contravenes (3.9). If we attempt to solve the differential equation (1.1), the standard methods tend to allow h_m to increase as the zero is approached, since the rate of change in direction of the solution trajectory is decreasing. If this happens then oscillations may occur if h_m becomes too large, as would be the case, for example, when using Newton's method with a step length greater than 2. When the step is suitably controlled no problems of instability occur and, indeed, as long as h_m satisfies (3.9) for each m, close to the zero the problem is extremely stable, simply because any zero of f is an asymptotically stable node of the autonomous differential equation (1.1) [20].

The foregoing theory shows that any method giving a polynomial $p(z)$ such that $p(-h)$ has a positive real root will be effective for producing rapid final convergence if $\{h_m\}$ is suitably chosen. For example, we consider briefly Runge–Kutta methods of orders one, three and five.

[1] Note that "order" is a term related to the accuracy of single and multistep methods in following the trajectory $x(t)$ (see [17]), while the terms "R-order" and "second order" are related to the speed of convergence of a sequence to its limit (see §3.2 and [24]).

https://doi.org/10.1017/S0334270000001077 Published online by Cambridge University Press

The simplest first-order method is Euler's method. In this case $p(z)$ is given by

$$p(z) = 1 + z$$

and from (3.9) we see that x^* is a point of attraction with h-domain $[\delta, 2\text{-}\delta]$, for δ arbitrarily small, i.e. local convergence is guaranteed if $0 < \delta \leqq h_m \leqq 2 - \delta$ for each m. Also, from (3.10), convergence to x^* can be second order only if h_m converges to 1 with R-order at least 2. This is essentially Newton's method.

There is a class of third-order Runge–Kutta methods and, for each, $p(z)$ is defined by

$$p(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6}.$$

Again, from (3.9), these methods converge locally with h-domain $[\delta, \bar{h} - \delta]$, for arbitrarily small δ, where $\bar{h} = 2.5127\cdots$. Also convergence to x^* can be second order only if h_m converges sufficiently quickly to $h_R = 1.596\cdots$ (h_R is the real root of $p(-z)$).

Finally, there exists a class of six stage fifth-order methods described by Lawson [21]. For one which he recommends, $p(z)$ is defined by

$$p(z) = \sum_{j=0}^{5} \frac{z^j}{j!} + 0.5625\frac{z^6}{6!}.$$

In this case x^* is a point of attraction with h-domain $[\delta, \bar{h} - \delta]$ for arbitrarily small δ, where $\bar{h} = 5.6039\cdots$, and again convergence to x^* is second order if h_m converges sufficiently quickly to $\hat{h} = 2.6299\cdots$, where $\hat{h}$ is a real root of $p(-z)$.

The conclusion of this section is that there exist single-step methods which can follow the solution trajectory of (1.1) sufficiently accurately and which, by suitable control of the step length, can furnish rapid convergence to x^*. In Section 5 numerical details are given for a third-order method which adapts the step length until it reaches a maximum of $h_R = 1.596\cdots$, after which it is not allowed to increase further.

4. Multistep methods

4.1. Implicit multistep methods

In this section we consider the solution of the differential equation (3.3) by means of a linear multistep method of the form

$$\rho(E)x_m - h\sigma(E)q(x_m) = 0, \qquad m = 0, 1, \cdots, \tag{4.1}$$

https://doi.org/10.1017/S0334270000001077 Published online by Cambridge University Press

where E is the displacement operator defined by

$$E^k(v(x)) = v(x + kh)$$

and $\rho(\lambda)$ and $\sigma(\lambda)$ are polynomials given by

$$\rho(\lambda) = \sum_{j=0}^{r} \alpha_j \lambda^j, \qquad \alpha_r \neq 0, \tag{4.2}$$

and

$$\sigma(\lambda) = \sum_{j=0}^{r} \beta_j \lambda^j. \tag{4.3}$$

The process (4.1) can be considered as a (possibly implicit) multistep method of the form

$$G(x_{m+r}, \cdots, x_m) = 0, \qquad m = 0, 1, \cdots \tag{4.4}$$

and we can use the following theorem, due to Voigt [28], to give conditions on the method which will guarantee local convergence to a zero of f when $q(x)$ is given by (3.8). In the following $\partial_i G(x_1, \cdots, x_m)$ denotes the Fréchet partial derivative of G with respect to x_i.

THEOREM 4.1. *Suppose that $G : D^{r+1} \subset (R^n)^{r+1} \to R^n$ is continuously differentiable on an open neighbourhood $D_0^{r+1} \subset D^{r+1}$. Assume that there is an $x^* \in D_0$ such that $G(x^*, \cdots, x^*) = 0$, $\partial_1 G(x^*, \cdots, x^*)$ is nonsingular and $\eta = \eta(H) < 1$, where H is given by*

$$H = \begin{bmatrix} H_2 & H_3 & \cdot & \cdot & & H_{r+1} \\ I & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & I & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & I & 0 \end{bmatrix} \tag{4.5}$$

and

$$H_i = -\partial_1 G(x^*, \cdots, x^*)^{-1} \partial_i G(x^*, \cdots, x^*), \qquad i = 2, \cdots, r+1. \tag{4.6}$$

Then there is an open neighbourhood S of x^* such that the sequence $\{x_k\}$ defined by the iterative process $\mathcal{S}$ given by (4.4) is well defined for any $(x_0, x_1, \cdots, x_{r-1}) \in S^r$ and converges to x^* with

$$R_1(\mathcal{S}, x^*) = \eta.$$

https://doi.org/10.1017/S0334270000001077 Published online by Cambridge University Press

PROOF. See Voigt [28].

In our application, from (4.1)–(4.3), we have

$$G(y_1, \cdots, y_{r+1}) = \sum_{j=0}^{r} \alpha_j y_{r-j+1} - h \sum_{j=0}^{r} \beta_j q_{r-j+1}, \tag{4.7}$$

where $q_k = q(y_k)$.

The first condition that Theorem 4.1 imposes is that

$$G(x^*, \cdots, x^*) = 0 \tag{4.8}$$

which, since $q(x^*) = 0$, gives

$$\sum_{j=0}^{r} \alpha_j = 0 \tag{4.9}$$

and this, in the usual notation, can be expressed as

$$\rho(1) = 0. \tag{4.10}$$

Also

$$\partial_i G(y_1, \cdots, y_{r+1}) = \alpha_{r-i+1} I - h\beta_{r-i+1} q'(y_i), \qquad i = 1, \cdots, r+1$$

and since $q'(x^*) = -I$, it follows that

$$\partial_i G(x^*, \cdots, x^*) = (\alpha_{r-i+1} + h\beta_{r-i+1})I, \qquad i = 1, \cdots, r+1.$$

For application of Theorem 4.1 we require that $\partial_1 G(x^*, \cdots, x^*)$ be nonsingular, i.e. that

$$\alpha_r + h\beta_r \neq 0 \tag{4.11}$$

and subsequently we assume this to be the case. In Section 4.3 we assume (4.1) to be an explicit method, in which case $\alpha_r \neq 0$ and $\beta_r = 0$, so (4.11) is automatically satisfied.

Define

$$\xi_i = \frac{\alpha_{r-i+1} + h\beta_{r-i+1}}{\alpha_r + h\beta_r}, \qquad i = 2, \cdots, r+1,$$

so

$$H_i = -\xi_i I, \qquad i = 2, \cdots, r+1.$$

To guarantee that the sequence $\{x_k\}$ generated by (4.4) converges to x^*, we look at $\eta(H)$ with H given by (4.5) and (4.6). Simple algebra shows that λ is an eigenvalue of H if and only if λ satisfies

$$\rho(\lambda) + h\sigma(\lambda) = 0. \tag{4.12}$$

https://doi.org/10.1017/S0334270000001077 Published online by Cambridge University Press

Thus, from Theorem 4.1, a sufficient condition for local convergence to x^* is that each root of (4.12) is less than 1 in magnitude. As in (3.9), this gives an *explicit* bound on h to ensure ultimate convergence.

We now consider the possibility of superlinear convergence of the sequence $\{x_k\}$ to x^*. Theorem 4.1 shows that this is possible only if

$$\eta(G(x^*, \cdots, x^*)) = 0,$$

i.e. if all the roots of (4.12) are zero. This is equivalent to the condition

$$\rho(\lambda) + h\sigma(\lambda) = \gamma\lambda^r$$

for some $\gamma \neq 0$. From (4.2) and (4.3) this is equivalent to

$$\alpha_r + h\beta_r = \gamma,$$

and

$$\alpha_j + h\beta_j = 0, \qquad j = 0, \cdots, r-1.$$

We have therefore proved the following theorem.

THEOREM 4.2. *For superlinear convergence of a linear multistep method applied to* (1.1) *the general iterative process*

$$\sum_{j=0}^{r} \alpha_j x_{m+j} + h \sum_{j=0}^{r} \beta_j J(x_{m+j})^{-1} f(x_{m+j}) = 0$$

must be of the form

$$\sum_{j=0}^{r} \alpha_j x_{m+j} - \sum_{j=0}^{r-1} \alpha_j J(x_{m+j})^{-1} f(x_{m+j}) + h\beta_r J(x_r)^{-1} f(x_r) = 0,$$

where

$$\sum_{j=0}^{r} \alpha_j = 0$$

and

$$\alpha_r + h\beta_r \neq 0.$$

In the explicit case, when $\beta_r = 0$, this can be considered as a weighted Newton method where, at each step, x_{r+m} is taken to be a weighted sum of Newton steps, i.e.

$$x_{r+m} = \sum_{j=0}^{r-1} \hat{\alpha}_j (x_{j+m} - J(x_{j+m})^{-1} f(x_{j+m})),$$

where $\hat{\alpha}_j = -\alpha_j/\alpha_r$ and $\Sigma_{j=0}^{r-1} \hat{\alpha}_j = 1$.

https://doi.org/10.1017/S0334270000001077 Published online by Cambridge University Press

4.2. Explicit methods

Since an implicit method requires, at each iteration, the solution of a system of nonlinear equations and since finding such a solution is our original problem, we regard implicit methods as inappropriate and do not consider them further. In this section we consider explicit multistep methods for solving (1.1) which have satisfactory stability and order properties. The results of the previous section show that, given h_0, any method for which $\rho(\lambda)$ satisfies (4.10) and

$$\rho(\lambda) = \lambda^r - h_0\sigma(\lambda), \tag{4.13}$$

(where $\sigma(\lambda)$ is a polynomial of degree $r-1$), is explicit and gives local superlinear convergence to x^* when $h = h_0$. Consider now the order, in the sense of Henrici [17], attainable by this method.

THEOREM 4.3. *Given any h_0 in* (4.13) *there exists a unique polynomial $\sigma(\lambda)$ of degree $r-1$ such that the resulting method has order $r-1$. For any r there exist at most r values of h_0 such that the method has order r.*

PROOF. The proof is an application of Lemma 5.3 of Henrici [17] which states that a method has exact order p if and only if the function

$$\phi(\zeta) = \frac{\rho(\zeta)}{\log \zeta} - \sigma(\zeta)$$

has a zero of exact order p at $\zeta = 1$. In this case, from (4.13), $\phi(\zeta)$ is given by

$$\phi(\zeta) = \frac{\rho(\zeta)}{\log \zeta} - \frac{\zeta^r - \rho(\zeta)}{h_0}.$$

Thus, a method defined by (4.13) has order p if and only if there exists a function $\psi_1(\zeta)$ such that $\psi_1(1) \neq 0$ and

$$\frac{\rho(\zeta)}{\log \zeta} - \frac{\zeta^r - \rho(\zeta)}{h_0} = (\zeta - 1)^p \psi_1(\zeta).$$

Letting $1 + \gamma = \zeta$ this is equivalent to the existence of a function $\psi_2(\gamma)$ such that $\psi_2(0) \neq 0$ and

$$\rho(1+\gamma) = \left[\frac{\log(1+\gamma)}{h_0 + \log(1+\gamma)}\right]\left[(1+\gamma)^r + \gamma^p \psi_2(\gamma)\right],$$

i.e.

$$\rho(1+\gamma) = \frac{(1+\gamma)^r \log(1+\gamma)}{h_0 + \log(1+\gamma)} + \frac{\gamma^p \log(1+\gamma)}{h_0 + \log(1+\gamma)} \psi_2(\gamma).$$

https://doi.org/10.1017/S0334270000001077 Published online by Cambridge University Press

Expanding both terms on the right hand side in powers of γ, the condition that the method has order p is that there exist constants $\pi_1, \pi_2, \cdots$, such that $\pi_1 \neq 0$ and

$$\rho(1+\gamma) = \frac{a_1(h_0)}{h_0}\gamma + \frac{a_2(h_0)}{h_0^2}\gamma^2 + \cdots + \frac{a_r(h_0)}{h_0^r}\gamma^r + \cdots + \gamma^{p+1}(\pi_1 + \pi_2\gamma + \pi_3\gamma^2 + \cdots) \quad (4.14)$$

where, for each j, $a_j(h_0)$ is a polynomial of degree $j-1$ in h_0.

For $p = r-1$ the coefficients π_j, $j = 1, 2, \cdots$, can be chosen so that $\pi_1 + a_r(h_0)/h_0^r = 1$ and

$$\frac{a_{j+r-1}(h_0)}{h_0^{j+r-1}} + \pi_j = 0, \qquad j \geqq 2, \quad (4.15)$$

in which case the right hand side of (4.14) represents a polynomial of degree r with coefficient of γ^r equal to 1 as required. The derived method is obviously unique and has order $r-1$.

If $p = r$, h_0 is such that

$$a_r(h_0)/h_0^r = 1,$$

and π_j, $j > r$, are chosen to satisfy (4.15), then the method has order r. This can only be the case when h_0 is a root of the polynomial $a_r(h_0) - h_0^r$, which is of degree r. Thus there are at most r values of h_0 for which a method satisfying (4.13) can be of order r. This completes the proof.

Next we use Theorem 3.3 to give a lower bound on the local R-convergence rate of methods satisfying (4.10) and (4.13).

THEOREM 4.4. *Suppose that $q(x) = -J(x)^{-1}f(x)$ is continuous and there exists a $\delta > 0$ such that $q''(x)$ exists and is bounded in $S(x^*, \delta)$. Then any iterative process $\mathcal{S}$ defined by* (4.1)–(4.3) *for which $\rho(\lambda)$ satisfies* (4.10) *and* (4.13) *when applied to* (1.1) *converges locally to x^* and*

$$O_R(\mathcal{S}, x^*) \geqq 2^{1/r}.$$

PROOF. Rewrite (4.1)–(4.3) in the explicit form

$$x_{m+r} = G(x_{m+r-1}, \cdots, x_m)$$

and set $z_k = (x_k, \cdots, x_{k-r+1})$, for $k = m+r-1, m+r, \cdots$, and $z^* = (x^*, \cdots, x^*)$. Define $\hat{G} : D^r \subset (R^n)^r \to (R^n)^r$ by

$$\hat{G}(y_1, \cdots, y_m) = (G(y_1, \cdots, y_m), y_1, \cdots, y_{m-1}).$$

Then $z_{k+1} = \hat{G}(z_k)$. Since G is differentiable at x^*, $\hat{G}$ is differentiable at z^* and $G'(z^*) = H$, where H is given by (4.5). However it follows from (4.13)

https://doi.org/10.1017/S0334270000001077 Published online by Cambridge University Press

that in (4.5), $H_i = 0$, $i = 2, \cdots, r+1$, and so $\eta(G'(z^*)) = 0$. Also, from the form of (4.5), $G'(z^*)^r = 0$ and $G'(z^*)^{r-1} \neq 0$.

$\hat{G}(z)$ therefore satisfies the conditions of Theorem 3.3, z^* is a point of attraction of the iteration $\mathscr{I}_2 : z_{k+1} = \hat{G}(z_k)$, and $O_R(\mathscr{I}_2, z^*) \geqq 2^{1/r}$.

Now there exists a norm such that $\|x_i - x^*\| \leqq \|z_i - z^*\|$ for each i (see [28]) and so $O_R(\mathscr{I}, x^*) \geqq O_R(\mathscr{I}_2, z^*) \geqq 2^{1/r}$. This completes the proof.

We can now look at methods suggested by Theorem 4.3 for various values of r. The relevant polynomials are

$$\rho(\lambda) = \lambda^2 - \frac{2h_0 - 1}{h_0}\lambda - \frac{(1-h_0)}{h_0}, \qquad \text{for } r = 2, \tag{4.16a}$$

$$\rho(\lambda) = \lambda^3 - \frac{(6h_0^2 - 5h_0 + 2)}{2h_0^2}\lambda^2 + \frac{(3h_0^2 - 4h_0 + 2)}{h_0^2}\lambda - \frac{(2h_0^2 - 3h_0 + 2)}{2h_0^2}, \qquad \text{for } r = 3, \tag{4.16b}$$

and

$$\rho(\lambda) = \lambda^4 - \frac{(12h_0^3 - 13h_0^2 + 9h_0 - 3)}{3h_0^3}\lambda^3 + \frac{(12h_0^3 - 19h_0^2 + 16h_0 - 6)}{2h_0^3}\lambda^2 - \frac{(4h_0^3 - 7h_0^2 + 7h_0 - 3)}{h_0^3}\lambda + \frac{(6h_0^3 - 11h_0^2 + 12h_0 - 6)}{6h_0^3} \quad \text{for } r = 4, \tag{4.16c}$$

and similar formulae, of increasing complexity, can be derived for larger values of r. The two-step method in (4.16a) is order 1, but if $h_0 = 1$ the method deflates to a one-step method, also of order 1. This is, of course, Newton's method, and is the one-step method of order 1 suggested by Theorem 4.3.

Similarly if h_0 in (4.16b) is chosen so that the constant term is zero then the resulting method would be two-step and of order 2. That the polynomial $2h_0^3 - 3h_0 + 2$ has no real root shows that there is no such method. However there exists one value of h_0 for which a three-step method of order 3 exists. This is the method obtained by setting the constant coefficient of $\rho(\lambda)$ in (4.16c) equal to zero. The equation

$$6h_0^3 - 11h_0^2 + 12h_0 - 6 = 0 \tag{4.17}$$

has only one real solution, which is approximately 0.8599, and on setting h_0 to this value (4.16c) deflates to a three-step method.

Theorem 4.4 gives information on the R-order of convergence of iterative processes specified by (4.16). For (4.16a) the R-order is $\geqq 2^{1/2}$ and for (4.16b) is $\geqq 2^{1/3}$. We note however that the inequality is not necessarily strict, for example, if $h_0 = 1$ in (4.16a) the method becomes Newton's which has R-order 2. However, Theorem 4.4 does suggest that increasing r will reduce the efficiency of final convergence to x^*.

https://doi.org/10.1017/S0334270000001077 Published online by Cambridge University Press

Two further requirements on any practical method, for small h at least, are those of consistency and stability (see Henrici [17]). Consistency is equivalent to having order at least 1, which is the case for the methods under discussion, and stability demands that no root of $\rho(\lambda)$ exceeds 1 in modulus and that the roots of modulus 1 be simple. In this case the stability condition depends upon h_0 and for $r = 2, 3, 4$ the methods are stable if

$$\left.\begin{aligned} &h_0 \geqq 1/2 && \text{for } r = 2, \\ &h_0 \geqq 2/3 && \text{for } r = 3, \\ &2/3 \leqq h_0 \leqq 2.5147\cdots && \text{for } r = 4. \end{aligned}\right\} \tag{4.18}$$

So, for each r considered, if h_0 is chosen to satisfy (4.18) the methods will be stable for small h. That this condition need not be strictly fulfilled is shown in the next section for the methods will not be used with small h but only with $h = h_0$.

4.3. Variable steps

The methods discussed in the previous section were derived with the idea of initially using a small step size which, as the zero x^* is approached, could be increased and finally fixed at h_0 to give superlinear convergence to x^*. However the foregoing theory assumes h to be fixed throughout and so is not directly applicable to variable step size. We may generate methods based upon those described in section 4.2 with varying step size, in the style of Gear [13]. These can be either of the Nordsieck type [23], where instead of using approximations to $x(ih)$ and $\dot{x}(ih)$, $i = m, m+1, \cdots, m+r-1$, we use approximations to the derivatives $x^{(k)}(mh)$, $k = 0, 1, \cdots, 2r-1$, or of the variable step type where we start with r unequally spaced points t_{m+r-i}, $r > i \geqq 0$, and compute the coefficients of the explicit multistep formula

$$\begin{aligned} y_{m+r} = {} & h_{m+r-1}\,\beta_{r-1,m}\,y_{m+r-1} + \cdots + h_m \beta_{0,m}\,y_m \\ & + h_{m+r-1}\,\beta_{r-1,m}\,q_{m+r-1} + \cdots + h_m \beta_{0,m}\,q_m \end{aligned}$$

so that the order is $r-1$, where $h_j = t_{m+j+1} - t_{m+j}$. This is the formula for variable steps (based upon (4.13)) which, if $h_j = h_0$ for $j = m, \cdots, m+r$, gives the formulae listed in (4.16).

Unfortunately these variable step methods are unstable with respect to changes in step size. When programmed the methods work well for fixed step but display obvious instability when step sizes are increased. This behaviour is explained in detail by the theory developed by Gear and Tu [14] and precludes the use of the methods with varying step. However, it is shown in [14] that the variable step methods based upon the Adams–Bashforth

formulae are stable and so the methods of section 4.3 can be combined with these to give the required characteristics. If an Adams–Bashforth variable-step method with r steps is applied to (1.1) then, as x^* is approached, the step size can be increased. Because the Adams method cannot give superlinear convergence to x^* we finally hold the step fixed at some value h_0 and when enough steps of fixed size have been taken we can switch to a method which gives fast ultimate convergence. Should a premature change to the fixed step be made then it will be necessary to reduce h and revert again to the variable step Adams method. These composite methods are thus variable formula and possibly variable order and an application of the comprehensive theory of Gear and Watanabe [15], on stability of variable order multistep methods, shows that the derived methods are stable.

In the following section we describe some numerical experience with variable formula methods of this type. A third-order Adams method is coupled with methods of order 3 as given by (4.16c).

5. Numerical results

We begin by making some general comments on the effectiveness of solving (1.1) as a means of finding a zero of f. Although it has been necessary to assume that x_0 is in a stability region of a zero x^*, for if this is not so then convergence is not guaranteed, there are applications where the approach will be effective. For example, where the usual methods continually converge to a zero which is known but where the user requires to find a different zero, which he knows to exist, and has a suitable starting point. However, one should realize that, whilst the number of evaluations required to follow the trajectory sufficiently accurately may seem reasonable to one used to solving ordinary differential equations, it may seem prohibitively large to one used to solving nonlinear equations.

Following the trajectory $x(t)$ is usually a simple matter if h can be chosen sufficiently small, but in practice the crucial part of solving (1.1) is in the step length control. Far from a zero of f all of the usual problems of step control occur and great care is required to maintain accuracy. Close to a zero of f this is not the case so long as h is controlled in a way which will guarantee convergence (see for example (3.8) or the bound on the roots of (4.12)). As x^* is approached we are less interested in accuracy in following the trajectory than in convergence to x^* and indeed, if we are to achieve fast ultimate convergence to x^*, we must relax our preoccupation with accurate representation of $x(t)$ which converges to x^* only linearly (see (2.1)). In the examples that follow we are interested only in demonstrating ways of achieving faster

final convergence and so we look only at cases when x_0 is fairly close to x^*. In this case the criterion for varying h can be simpler than would be necessary in the general case.

The basic technique depends upon the fact that the solution of (1.1) satisfies

$$f(x(t)) = e^{-t} f(x_0).$$

Let $f_i = f(x_i)$ and Z_i be given by

$$Z_i = I - \frac{f_i f_i^T}{f_i^T f_i}.$$

Then any point x, on $x(t)$, satisfies

$$Z_0 f(x) = 0.$$

Suppose x_i is our current approximation to x^*, then the solution of

$$\dot{x}(t) = -J(x)^{-1} f(x), \qquad x(0) = x_i$$

converges to x^* (under the conditions of Theorem 2.1) and $\|Z_i f_{i+1}\|$ gives a measure of the deviation of x_{i+1} from this trajectory. On this basis a suitable step change criterion was found to be $h_{i+1} = \min(h^*, \alpha h_i)$ where α is given by

$$\alpha = \begin{cases} 2 & \text{if} \quad 0 < \delta \leqq \varepsilon_1 \\ 1 & \text{if} \quad \varepsilon_1 < \delta \leqq \varepsilon_2 \\ 0.5 & \text{if} \quad \varepsilon_2 < \delta \leqq \varepsilon_3 \end{cases} \tag{5.1}$$

and where $\delta = \|Z_i f_{i+1}\|$. In addition, the point x_{i+1} was rejected and the step repeated with half the step length if either $\delta > \varepsilon_3$ or x_{i+1} crossed a region of singularity of the Jacobian $J(x)$. Finally, for each method, h_i was not allowed to increase beyond h^*, the step size required to furnish the fastest convergence for that method.

Various methods were tested on a variety of problems and the results of some of these tests are tabulated below. The methods described are a third-order Runge–Kutta method (RK3) with $h^* = h_R = 1.596\cdots$ and an Adams–Bashforth variable-step method of order 3, coupled with a multistep method of order 3, as described in section 4 (AB3). This method was tested for various values of h_0 and the results for $h_0 = 0.8598\cdots$, which is a three-step method, and for $h_0 = 0.7$, which is a four-step method, are given below. For comparison we looked also at the basic algorithm described by Boggs (PECE) given in (3.11).

Since we are advocating the use of (1.1) as opposed to (1.2), we also

https://doi.org/10.1017/S0334270000001077 Published online by Cambridge University Press

looked at a third-order Runge–Kutta method (K3) for solving equation (1.2) to find an estimate of the solution at $t = 1$. In this method a major iteration consists of integrating

$$\dot{x}(t) = -J(x)^{-1} f(x_i), \qquad x(0) = x_i, \tag{5.2}$$

giving a sequence $\{y_{i,j}\}$, $j = 1, \cdots, N_i$, such that $y_{i,j}$ is an approximation to $x(t_{i,j})$, where $t_{i,j} = \sum_{k=1}^{j-1} h_{i,k}$ and $t_{i,N_i} = 1$. Then $x_{i+1} = y_{i,N_i} = y_{i+1,1}$. It is proved by Kleinmichel [19] that, under general conditions, if the method uses step size $h^* = 1$ then the sequence $\{x_i\}$ converges to x^* with R-order 4. Despite this high rate of convergence, the greater demand on accuracy required in following the solution trajectory of (5.2) causes the algorithm to be less effective than those described in this paper.

For a fair comparison of methods we consider a similar step control to that described above. Since the solution of

$$\dot{x}(t) = -J(x)^{-1} f(y_{ij}), \qquad x(0) = y_{ij},$$

does not generally converge to x^* and may, in practice, cross a region of singularity of $J(x)$, it is necessary that each y_{ij} be close to the solution trajectory of (5.2). In this case, therefore, the most suitable criterion is that $h_{i,j+1} = \min(\alpha h_{i,j}, 1 - t_{i,j+1})$ where α is given by (5.1) and $\delta = \| Z_i f(y_{i,j+1}) \|$. Also we took $h_{i+1,1} = \min(1, 2 \max(h_{i,N_i}, h_{i,N_i-1}))$. The conditions for rejecting a step were the same as before.

In each algorithm $\varepsilon_3 = 0.5$, $\varepsilon_2 = 0.25$ and $\varepsilon_1 = 0.05$ were found to be suitable, except that $\varepsilon_1 = 0.01$ was used in AB3 since, with $\varepsilon_1 = 0.05$, that method occasionally made a premature change to step size h^* when close to the solution x^*. The initial step, in each case, was taken as $h^*/8$.

Each algorithm was applied to a variety of functions and the following eight problems gave results which were typical. In each case the solution given is the limit of the trajectory defined by (1.1) with the given value of x_0.

1. A function found in Boggs [4];

$$f_1 = x_1^2 - x_2 + 1,$$

$$f_2 = x_1 - \cos\left(\frac{\pi}{2} x_2\right),$$

with initial guess $x_0 = (1, 0)$. The correct solution is $x^* = (0, 1)$.

2. Problem 1 with initial guess $(-1, -1)$. The correct solution is $(0, 1)$ and the solution trajectory passes close to a region where $J(x)$ is singular.

3. A function found in Broyden [6];

https://doi.org/10.1017/S0334270000001077 Published online by Cambridge University Press

$$f_1 = \tfrac{1}{2}\sin(x_1x_2) - x_2/(4\pi) - x_1/2,$$

$$f_2 = (1 - 1/(4\pi))(e^{2x_1} - e) + ex_2/\pi - 2ex_1,$$

with initial guess (0.6, 3.0). The correct solution is $(1/2, \pi)$.

4. The gradient of Rosenbrock's function;

$$f_1 = 400x_1(x_1^2 - x_2) + 2(x_1 - 1),$$

$$f_2 = -200(x_1^2 - x_2),$$

with initial guess $(-1.2, 1.0)$. The correct solution is (1,1) and this problem can be considered fairly difficult since the solution trajectory is always close to the region were $J(x)$ is singular (see [5]).

5. A function found in Branin [5];

$$f_1 = 2\sin(2\pi x_1/5)\sin(2\pi x_3/5) - x_2,$$

$$f_2 = 2.5 - x_3 + 0.1x_2\sin(2\pi x_3) - x_1,$$

$$f_3 = 1 + 0.1x_2\sin(2\pi x_1) - x_3,$$

with initial guess (0,0,0). The correct solution is $(1.5, 1.809\cdots, 1.0)$.

6. A function found in Deist and Sefor [10];

$$f_i = \sum_{\substack{j=1 \\ j\neq i}}^{6} \cot \beta_i x_j, \qquad i = 1, \cdots, 6,$$

where $100\beta_i = 2.249, 2.166, 2.083, 2.0, 1.918, 1.835$, for $i = 1, \cdots, 6$ respectively. With initial guess $x_i = 75.0$, $i = 1, \cdots, 6$ the correct solution is approximately (121.9, 114.2, 93.6, 62.3, 41.3, 30.5).

7. A discretisation of

$$3\ddot{y}y + \dot{y}^2 = 0$$

with boundary conditions $y(0) = 0$, $y(1) = 20$, gives rise to the equations

$$f_1 = 3x_1(x_2 - 2x_1) + x_2^2/4,$$

$$f_i = 3x_i(x_{i+1} - 2x_i + x_{i-1}) + (x_{i+1} - x_{i-1})^2/4, \qquad i = 2, \cdots, n-1,$$

$$f_n = 3x_n(20 - 2x_n + x_{n-1}) + (20 - x_{n-1})^2/4.$$

The true solution of the boundary value problem is $y = 20t^{3/4}$. As initial guess we chose $x_i = 10$, $i = 1, \cdots, n$ and set $n = 10$.

8. Same as problem 7 with $n = 20$.

https://doi.org/10.1017/S0334270000001077 Published online by Cambridge University Press

Both of these problems have solution trajectories which pass close to a region of singularity.

Table 1 gives results on the effort required by the methods to reduce each component of f to less than 10^{-6}. For each method the first line gives the number of Jacobian evaluations, the second gives the number of function evaluations and the third the number of equivalent function evaluations counting a Jacobian evaluation as n function evaluations, except for problems 7 and 8 where the Jacobian is tridiagonal and its evaluation is counted as being equivalent to 3 function evaluations. Note that, because of the way steps were either accepted or rejected, the number of Jacobian and function evaluations are not necessarily the same.

TABLE 1

Algorithm		Problem							
		1	2	3	4	5	6	7	8
		21	29	18	110	28	24	69	69
RK3		22	31	19	114	29	25	73	73
		64	89	55	334	113	169	280	280
		23	31	14	99	27	18	54	56
	$h_0 = .859\cdots$	25	33	15	101	28	19	59	61
		71	95	43	299	109	127	221	229
AB3									
		26	35	16	95	33	21	55	58
	$h_0 = .7$	28	38	17	98	34	22	59	63
		80	108	49	288	133	148	224	237
		18	39	15		26	29		
K3		7	13	6	*	10	11	*	**
		43	91	36		88	185		
		44	52	38	109	46	44	86	88
PECE		45	53	39	119	47	45	89	91
		133	157	115	337	185	309	347	355

* – h reduced to minimum allowed, viz. $2^{-13}h^*$.
** – terminated after 200 function evaluations.

We can draw a number of conclusions from the numerical results. The first is that the PECE algorithm, which has only linear convergence to x^*, requires significantly more evaluations than the other methods. This is as we

https://doi.org/10.1017/S0334270000001077 Published online by Cambridge University Press

would expect. Of the algorithms described in the previous two sections, the multistep methods generally seem to be the most efficient. This is more obviously the case when many evaluations are required for then these methods gain by requiring only one evaluation per iteration. They are most efficient when h_0 is close to 0.75. For values larger than 1 the stability decreases since the methods have some difficulty with large steps. For this reason the Runge–Kutta methods appear more efficient when x_0 is close to x^*. Also for values of h_0 smaller than 0.6 multistep methods suffer from instability, presumably because the steps are sufficiently small for the instability predicted by (4.18) to have an effect. In general, the three-step version seemed superior since it could change to give high order convergence one iteration sooner.

Because of the high rate of ultimate convergence, the K3 algorithm is generally superior when the problem is simple, i.e. when the solution trajectory is smooth and does not approach close to regions where the Jacobian is singular. However, where this is not the case RK3 and AB3 are more efficient and in particular we note that they are more reliable in that they always succeeded in finding the desired solution in a reasonable time.

We note here that any comparison of routines is necessarily a comparison also of the step change criteria and that the criteria chosen were not necessarily the best for each routine. However we have deliberately adopted simple criteria for step size in the hope of demonstrating that the methods which use (1.1) are more robust than those which use (1.2).

6. Conclusion

Single and multistep methods, normally applied to the solution of ordinary differential equations have proved useful as a means of solving nonlinear equations. These methods work well so long as the step lengths used are strictly controlled. Although far from a solution of the equations any accurate and efficient method is satisfactory, close to a solution greater efficiency can be achieved by choosing a method which will give fast ultimate convergence to the required solution.

Acknowledgement

The authors would like to thank the referee for several useful comments.

References

[1] J. H. Avila, *Continuation methods for nonlinear equation*, (Ph.D. Thesis Computer Science Center, University of Maryland, 1971).

[2] L. Bittner, 'Einige kontinuierliche Analogien von Iterationsverfahren, in Functionalanalysis, Approximationstheorie Numerische Mathematik, *ISNM*7, (Birkhauser Verlag, Basel), (1967), 114–135.

[3] P. T. Boggs, *The solution of nonlinear operator equations by A-stable integration techniques*, (Ph.D. Thesis, Cornell University, 1970).

[4] P. T. Boggs, 'The solution of nonlinear systems of equations by A-stable integration techniques', *SIAM J. Numer. Anal.* 8 (1971), 767–785.

[5] F. H. Branin Jr., 'Widely convergent method for finding multiple solutions of simultaneous nonlinear equations', *IBM J. Res. Develop.* 16 (1972), 504–522.

[6] C. G. Broyden, 'A new method of solving nonlinear simultaneous equations', *Comput. J.*, 12 (1969), 94–99.

[7] K. S. Chao, D. K. Liu and C. T. Pan, *A systematic search method for obtaining multiple solutions of simultaneous nonlinear equations*, IEEE Transactions on Circuits and Systems, Vol. CAS–22, 9, September 1975.

[8] G. G. Dahlquist, 'A special stability problem for linear multistep methods', *BIT* 3 (1963), 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), 601–602.

[10] F. H. Deist and L. Sefor, 'Solution of systems of nonlinear equations by parameter variation', *Comput. J.* 10 (1967), 78–82.

[11] P. Deuflhard, 'A modified Newton method for the solution of ill-conditioned systems of nonlinear equations with application to multiple shooting', *Numer. Math.* 22 (1974), 289–315.

[12] M. K. Gavurin, 'Nonlinear functional equations and continuous analogs of iterative methods', *Izv. Vyss. Ucebn. Saved. Mathematika* 6 (1958), 18–31; English Transl., *Tech. Rep.* 68–70, Computer Science Center, Univ. of Maryland, 1968.

[13] C. W. Gear, *Numerical initial value problems in ordinary differential equations*, (Prentice-Hall, Englewood Cliffs, N.J. 1971).

[14] C. W. Gear and K. W. Tu, 'The effect of variable mesh size on the stability of multistep methods', *SIAM J. Numer. Anal.* 11 (1974), 1025–1043.

[15] C. W. Gear and D. S. Watanabe, 'Stability and convergence of variable order multistep methods', *SIAM J. Numer. Anal.* 11 (1974), 1044–1058.

[16] J. K. Hale, *Ordinary differential equations*, (Wiley-Interscience, N.Y., 1969).

[17] P. Henrici, *Discrete variable methods for ordinary differential equations*, (John Wiley, N.Y. 1962).

[18] W. Kizner, 'A numerical method for finding solutions of nonlinear equations', *SIAM J Appl. Math.* 12 (1964), 424–428.

[19] H. Kleinmichel, 'Stetige Analoga und Iterationsverfahren für nichtlineare Gleichungen in Banachräumen', *Math. Nachr.* 37 (1968), 313–344.

[20] L. Lasalle and S. Lefshetz, *Stability by Liapunov's direct method with applications*, (Academic Press, N.Y. 1961).

[21] J. D. Lawson, 'An order five Runge–Kutta process with extended region of stability', *SIAM J. Numer. Anal.* 3 (1966), 593–597.

[22] G. H. Meyer, 'On solving nonlinear equations with a one parameter operator embedding', *SIAM J. Numer. Anal.* 4 (1968), 739–752.

[23] A. Nordsieck, 'On the numerical integration of ordinary differential equations', *Maths. Comp.* 16 (1962), 22–49.

[24] J. Ortega and W. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, (Academic Press, N.Y., 1970).

[25] J. Ortega and M. Rockoff, 'Nonlinear difference equations and Gauss–Seidel type iterative methods', *SIAM J. Numer. Anal.* 3 (1966), 497–513.

[26] A. Ostrowski, *Solution of equations and systems of equations*, (Academic Press, N.Y., 2nd ed, 1966).

[27] W. C. Rheinboldt, 'Local mapping relations and global implicit function theorems', *Trans. Amer. Math. Soc.* 138 (1969), 183–198.

[28] R. G. Voigt, 'Rates of convergence for a class of iterative procedures', *SIAM J. Numer. Anal.* 8 (1971), 127–134.

[29] M. N. Yakovlev, 'On some methods of solving nonlinear equations', *Trudy Mat. Inst. Steklov.* 84 (1965), 8–40; English Transl. *Tech. Rep.* 68–75 (Computer Science Center, Univ. of Maryland, 1968).

Computer Centre,
Australian National University,
Canberra 2600,
Australia.