W. Liniger
F. Odeh

A-stable, Accurate Averaging of Multistep Methods for Stiff Differential Equations

Abstract: Several low-order numerical solutions of stiff systems of ordinary differential equations are computed by repeated integration, using a multistep formula with parameters. By forming suitable linear combinations of such solutions, higher-order solutions are obtained. If the parameters are properly chosen the underlying solutions, and thus the higher-order one, can be made A-stable and strongly damping with respect to the stiff components of the system. A detailed description is given of an algorithmic implementation of the method, which is computationally efficient. Numerical experiments are carried out on some test problems, confirming the validity of the method.

Introduction

We are concerned with the numerical integration of stiff systems of ordinary differential equations

$$\dot{x} = f(x) , \tag{1}$$

i.e., systems with widely separated eigenvalues, λ, of the Jacobian matrix $J = f_x$ [as evaluated, for example, along a particular solution $x = F(t)$ of Eq. (1)]. Stiff systems are encountered in many areas of applications, e.g., reactor calculations, circuit analysis, chemical kinetics, flight dynamics, constrained optimization, etc. For a survey on stiff systems, see Bjurel et al. [1].

In the presence of stiffness, a numerical integration method must have very strong stability properties if it is to be efficient. Typically, one wishes to integrate the system of Eq. (1) with a fixed integration step h such that, for one or more of the eigenvalues λ of J, one has $|q| \gg 1$, where $q = \lambda h$. It is thus desirable for a formula to have unbounded regions of fixed-h stability in the complex q-plane.

One stability concept that is widely used in connection with stiff systems, called *A-stability*, was introduced by Dahlquist [2]. It demands that all numerical solutions, generated by applying a numerical integration method, with an arbitrary integration step h, to the equation $\dot{x} = \lambda x$, $\mathrm{Re}\ \lambda < 0$ must be (strictly) asymptotically stable. In this case, the unbounded region of fixed-h stability mentioned above is the left-half q-plane. It is easy to see that for linear multistep formulae (LMF),

$$\sum_{j=0}^{k} (\alpha_j x_{n+j} - h\beta_j \dot{x}_{n+j}) = 0 , \tag{2}$$

a necessary and sufficient condition for A-stability is that the roots z_i, $1 \le i \le k$, of the characteristic polynomial

$$\chi_q(z) = \rho(z) - q\sigma(z) , \ \rho(z) = \sum_{j=1}^{k} \alpha_j z^j , \ \sigma(z) = \sum_{j=1}^{k} \beta_j z^j \tag{3}$$

be inside the unit circle, for all q in the left half-plane.

However, the class of A-stable LMF is quite small. In fact, for an LMF to be A-stable, it is necessary that it be implicit ($\beta_k \neq 0$) and that its order of accuracy [3] should not exceed two [2]. In order to find a broader class of LMF, useful for solving stiff equations, one may employ a somewhat weaker stability concept, such as $A(\alpha)$-stability introduced by Widlund [4], or stiff stability defined by Gear [5]. There do exist LMF, with $p > 2$, which have either one of these properties. On the other hand, it is possible to find strictly A-stable methods with $p > 2$ if one considers a class of formulae larger than that described by Eq. (2), e.g., LMF involving second derivatives [6].

Another way of generating A-stable integration methods with $p > 2$ was proposed by Dahlquist [2]. It consists in integrating Eq. (1) more than once by an A-stable LMF, using different step sizes h, and then extracting from the solutions thus obtained a higher-order solution

by Richardson extrapolation. For example, if $x^{(1)}$ and $x^{(2)}$ are solutions of (1) obtained by the second-order A-stable Trapezoidal Rule with step sizes h and $2h$ respectively, then $x^* = (4x^{(1)} - x^{(2)})/3$ represents, wherever defined, i.e., at multiples of $2h$, an A-stable solution of order four. This procedure, however, has some disadvantages. In principle, it gives higher-order values of the solution only every two steps. The missing solution values have to be obtained by interpolation of appropriate degree. Moreover, the Trapezoidal Rule is only marginally stable for $q \to \infty$, i.e., rapidly decaying solutions of a stiff system give rise to slowly decaying oscillatory numerical solutions. To reduce the effect of such undesirable oscillations, it seems necessary for the integration by the Trapezoidal Rule to be followed by a filtering process [7].

In this paper we consider an approach similar to the one mentioned in the preceding paragraph. Starting from a special LMF, which has an order of accuracy p and contains intrinsic parameters other than h, we integrate Eq. (1) repeatedly using the same formula and the same h but with different values of those other parameters. We then show how to obtain a solution of higher-order accuracy by forming suitable linear combinations of such parametrized solutions. The method has several useful properties. The higher-order solution is produced directly at every time step and no interpolation is necessary, since we use the same step size h for each integration. It can be arranged that the underlying integration formulae are strongly damping with respect to the rapid transients of stiff systems, thus avoiding the type of oscillations accompanying the Trapezoidal Rule. Also, by starting from special formulae with order $p \leq 2$, we can assure the A-stability of the integration procedure as well as its higher accuracy. Finally, this procedure can be implemented in a computationally efficient way. We now outline the plan of the paper.

In the second section we introduce an Adams-type, k-step differential-difference operator Ω, of order $p < k$, which contains $d = k - p$ free "primary" parameters and one "secondary" parameter, in addition to the time-step h. A simple recipe is then given for constructing, from Ω, another operator Ω' of higher order $p + m$, $1 \leq m \leq d$ which contains $(d - m)$ primary parameters. The recipe consists of forming a suitable linear combination of $(m + 1)$ versions, Ω_ρ, $1 \leq \rho \leq m + 1$, of Ω corresponding to $(m + 1)$ different sets of values of the primary parameters. We then prove that if $m \leq p$ and x_ρ is the numerical solution of Eq. (1) associated with Ω_ρ, then the corresponding linear combination of the x_ρ's, which we sometimes loosely call their "average," represents a numerical solution of order $p + m$. The secondary parameter is held fixed throughout the averaging process for saving computer operations, which is discussed in the final section. In the third section, we discuss the result of averaging Adams-type solutions from the point of view of global error analysis. The nonlinear difference equation governing the accumulated truncation error is first derived, and a perturbation method is then used to obtain the explicit dependence of this error on the primary parameters. By examining the form of this dependence we generalize the averaging recipe mentioned above and remove the restriction $m \leq p$. We also indicate, via an example, how this approach leads to averaging procedures for solutions obtained from integration formulae other than of Adams type. Next three specific operators are introduced and the domains of A-stability in their parameter spaces are explicitly determined. To do this we apply a convenient criterion given earlier by one of the authors [8]. Then a finite algorithmic implementation of the averaging method is discussed for the case of one of the specific operators mentioned above. The algorithm, designated "A4," consists in computing one solution by using cubic extrapolation and a single Newton-step, then computing two other solutions from linearization around the first, and finally "averaging" the three solutions. We prove that the averaging process is correct for the finite algorithm, i.e., it produces fourth-order solutions. Also, the A-stability analysis of this finite algorithm is identical to that of the implicit difference equation obtained by applying Eq. (2) to $\dot{x} = \lambda x$ (see Section 4.2 of [6]). In the final section we give numerical results obtained by applying, to two test problems, the algorithms associated with the three specific operators mentioned earlier. These algorithms are A4, for which $p = 2$, $k = 4$, $m = d = 2$; A3, for which $p = 2$, $k = 3$, and $m = d = 1$; and A2, for which $p = 1$, $k = 2$, and $m = d = 1$. We also offer some remarks about the computational efficiency of A4 as compared to an implicit Runge-Kutta algorithm [9].

Averaging of Adams-type solutions

Let $x(t)$ be a real analytic function on the real line and $\{t_n\}$, $t_n = nh$, n integer, a discrete set of values of t with increment – or step – $h > 0$. Consider the linear differential-difference expression

$$l(t_n) = x(t_{n+1}) - x(t_n) - h\left[c\,\dot{x}(t_{n+1}) + (1-c)\dot{x}(t_n) + \sum_{j=1}^{k-1} (b_j^* - c)\nabla^j \dot{x}(t_n)\right], \tag{4}$$

where ∇^j denotes the jth backward difference, the b_j^* and c are, for the moment, undetermined constants (later, we refer to c by the term "secondary parameter") and the dot denotes differentiation with respect to t. It is convenient to associate with the expression (4), an operator

$$\Omega^* = \Delta - D\left[c\Delta + 1 + \sum_{j=1}^{k-1} (b_j^* - c)\nabla^j\right], \tag{5}$$

where Δ is the first forward difference and $D = h(d/dt)$. An operator whose form is indicated by (5) is said to be of the *Adams type*. When the differences in Ω^* are expanded, it can be written in the perhaps more familiar ordinate form that is more suitable for the analysis of stability questions.

A procedure for defining the accuracy of Ω^* is the following: Let $x(t)$ and its differences be expanded in Taylor series about $t = t_n$. If by the symbol $\simeq$ we denote equality with error $o(h^k)$ where k is the step number of Ω^*, then

$$\Delta \simeq \sum_{i=1}^{k} (i!)^{-1} D^i$$

$$D\nabla^j \simeq \sum_{i=j}^{k} \gamma_{ij} D^{i+1}, \qquad j = 1, \cdots, k-1, \tag{6}$$

where γ_{ij} are known constants (which may be derived, for example, from the tables in Abramowitz and Stegun [10]). Upon substitution in (5) from (6), one obtains

$$\Omega^* \simeq \sum_{i=2}^{k} v_i D^i, \tag{7}$$

where

$$v_i = (i!)^{-1} - \sum_{j=1}^{i-1} \gamma_{i-1,j} b_j^*, \qquad 2 \le i \le k. \tag{8}$$

Note that if the secondary parameter c is introduced as in (5), then the relation (8) between the quantities v_i and b_i^* becomes independent of c.

Since

$$\nabla^i = \sum_{j=i}^{\infty} \gamma_{ji} D^j = (D + \frac{1}{2!} D^2 + \cdots)^i = D^i + \cdots$$

one has $\gamma_{ii} = 1$ for all i. Hence, noting that the coefficient matrix defined by (8) is triangular, one can solve these equations uniquely for the b_j^*, for any given set of values v_i. If, for example, the constants b_j^* were chosen to make v_i, $i = 2, \cdots, p$, $p \le k$ vanish, then the operator defined by (7) is $O(h^{p+1})$ and is said to be of the order of accuracy p. If, starting from the class of all operators of order $p \le k-1$, we impose the additional constraints $v_i = 0$, $i = p+1, \cdots, p+m$, $1 \le m \le d$, $d = k - p$, then we raise the order of Ω^* from p to $p+m$. In particular, let b_i denote the solution of (8) when all the v_i vanish, then $\Omega^* \simeq 0$ and it has maximum order $p = k$.

We now describe an alternate method for raising the order of accuracy of an operator Ω. Consider

$$\Omega = \Delta - D\left[c\Delta + 1 + \sum_{j=1}^{p-1} (b_j - c)\nabla^j + \sum_{j=p}^{k-1} (b_j - c + u_{j-p+1})\nabla^j\right], \tag{9}$$

where $u_1, \cdots, u_d$ are, for now, arbitrary constants we call *primary parameters*. It is easy to see that in this case the expansion (7) for Ω takes the form

$$\Omega \simeq D^p \sum_{i=1}^{d} \mu_i D^i, \tag{10}$$

where

$$\mu_i = -\sum_{j=1}^{i} \gamma_{p-1+i,p-1+j} u_j, \qquad i = 1, \cdots, d. \tag{11}$$

Thus Ω is of order $\ge p$ for any set of parameters $u_1, \cdots, u_d$. These parameters form a d-dimensional vector $\mathbf{u}$. Consider $m+1$ such vectors, $\mathbf{u}_\rho = (u_{1,\rho}, \cdots, u_{d,\rho})$, $\rho = 1, \cdots, m+1$, and for each ρ the operator Ω_ρ defined by (9) and associated with $\mathbf{u}_\rho$. Form the operator

$$\Omega' = \sum_{\rho=1}^{m+1} \nu_\rho \Omega_\rho. \tag{12}$$

Given the vectors $\mathbf{u}_\rho$, $\rho = 1, \cdots, m+1$, we show how to choose the "weights" ν_ρ to make Ω' have order $p+m$. From (10), (11), and (12) one gets

$$\Omega' \simeq D^p \sum_{i=1}^{d} \pi_i D^i,$$

where

$$\pi_i = \sum_{\rho=1}^{m+1} \nu_\rho \mu_{i,\rho}. \tag{13}$$

An easy calculation shows that $\pi_i = 0$, $i = 1, \cdots, m$, if and only if

$$\sum_{\rho=1}^{m+1} \nu_\rho u_{i,\rho} = 0, \qquad i = 1, \cdots, m. \tag{14a}$$

Adjoin to the system (14a) the normalizing condition

$$\sum_{\rho=1}^{m+1} \nu_\rho = 1, \tag{14b}$$

the significance of which will appear later. If the ν_ρ satisfy (14), the operator $\Omega' = O(h^{p+m+1})$. More specifically, if

$$w_i = \sum_{\rho=1}^{m+1} u_{i,\rho} \nu_\rho, \qquad i = m+1, \cdots, d \tag{15}$$

then

$$\Omega' \simeq \Delta - D\left[c\Delta + 1 + \sum_{j=1}^{p+m-1} (b_j - c)\nabla^j + \sum_{j=p+m}^{k-1} (b_j - c + w_j)\nabla^j\right], \tag{16}$$

which has order of accuracy $(p+m)$.

The unique solvability of (14) is equivalent to the condition that the $(m+1)$ truncated vectors $\mathbf{u}_\rho' = (u_{1,\rho}, \cdots, u_{m,\rho})$, $\rho = 1, \cdots, m+1$, span an m-dimensional linear space. To see this, note that the ρth column vector of the coefficient matrix M defined by the system (14) is the composite vector $(1, \mathbf{u}_\rho')^\dagger$. Now if, for ex-

ample, the vectors $\mathbf{u}_\rho'$, $\rho = 1, \cdots, m$, are independent and $\mathbf{v}_\rho = \mathbf{u}_\rho' - \mathbf{u}_{m+1}'$, then

$$|M|_{(m+1)\times(m+1)} = |\mathbf{v}_1, \cdots, \mathbf{v}_m|_{m\times m} \neq 0$$

and thus M is nonsingular. The necessity of the condition is obvious. For the special case $m = 2$, which we shall consider later, the solvability condition is that the 2-vectors $\mathbf{u}_1'$, $\mathbf{u}_2'$, $\mathbf{u}_3'$ are the position vectors of points forming a nondegenerate triangle.

We now turn to the effect of the averaging procedure defined by (12) and (14) on the approximate numerical solution of the system (1). Let N denote the nonlinear difference operator derived from (9) by substituting the smooth function $f(x)$ for the derivative $\dot{x}$. Similarly, define N_ρ and N'. Let x_ρ be the solution of $N_\rho x_\rho = 0$, $\rho = 1, \cdots, m+1$, at an arbitrary time $t = t_n$ and define the averaged solution $z = z_n$ by

$$z = \sum_{\rho=1}^{m+1} \nu_\rho x_\rho, \tag{17}$$

where the ν_ρ are defined by (14). Then we have the following theorem.

• *Theorem*
If $k \leq 2p$ and $m \leq p$, then the approximate solution z satisfies the difference equation $N'z = 0$ globally to within an error $O(h^{k+1})$.

The following corollary is an immediate consequence of this theorem:

Corollary: The global discretization error of the approximate solution z is $O(h^{p+m})$.

Proof. Let $y(t)$ denote the exact solution of (1), defined on an interval (a, b) and $y = y(t_n)$ its value at the grid point $t_n = a + nh$, $n = 0, 1, \cdots, H$, where $Hh = b - a$. Since x_ρ is defined by a stable difference operator of order p, the global truncation error $(x_\rho - y) = O(h^p)$ [11]. Therefore, one has

$$z - y = \sum_{\rho=1}^{m+1} \nu_\rho (x_\rho - y) = O(h^p)$$

and thus

$$x_\rho - z = (x_\rho - y) + (y - z) = O(h^p). \tag{18}$$

From the definition of z it follows that

$$\Delta z - h\left\{\left[c\Delta + 1 + \sum_{j=1}^{p-1} (b_j - c)\nabla^j\right]\left[\sum_{\rho=1}^{m+1} \nu_\rho f_\rho\right] + \sum_{j=p}^{k-1} \nabla^j \sum_{\rho=1}^{m+1} (b_j - c + u_{j-p+1,\rho})\nu_\rho f_\rho\right\} = 0, \tag{19}$$

where $f_\rho = f(x_\rho)$. In the following, it is convenient to think of all discrete values as being interpolated by C^k functions. For simplicity in notation, we write $\phi = f(z)$, $J = f_x(z)$ and, for a generic ρ, $x = x_\rho$ and $f = f_\rho$. Then using Taylor's theorem, together with the assumption $k \leq 2p$, one has

$$f = \phi + J[x - z] + O(h^{2p})$$

so that

$$hf \simeq h\phi + hJ[x - z].$$

Similarly,

$$h\Delta f \simeq h\Delta\phi + h\Delta\{J[x - z]\},$$

$$h\nabla^j f \simeq h\nabla^j\phi + h\nabla^j\{J[x - z]\}, \qquad j = 1, \cdots, (p-1).$$

The last term in the above relation is $o(h^k)$ if $j \geq p$, and therefore

$$h\nabla^j f \simeq h\nabla^j\phi, \qquad p \leq j \leq k - 1.$$

Hence, if Γ denotes any one of the operators Δ, I or ∇^j, $j = 1, \cdots, p-1$, the above relations imply

$$h\Gamma\left[\sum_{\rho=1}^{m+1} \nu_\rho f_\rho\right] \simeq h\left[\sum_{\rho=1}^{m+1} \nu_\rho\right]\Gamma\phi + h\Gamma\left\{J\left[\sum_{\rho=1}^{m+1} \nu_\rho (x_\rho - z)\right]\right\} = h\Gamma\phi \tag{20a}$$

and

$$h\nabla^j\left[\sum_{\rho=1}^{m+1} (b_j - c + u_{j-p+1,\rho})\nu_\rho f_\rho\right] \simeq h\nabla^j\phi\left[\sum_{\rho=1}^{m+1} \nu_\rho (b_j - c + u_{j-p+1,\rho})\right], \tag{20b}$$

$$p \leq j \leq k - 1.$$

Substituting the above into (19) and using the relations (14) and definition (15) one obtains

$$\Delta z - h\left[c\Delta + 1 + \sum_{j=1}^{p+m-1} (b_j - c)\nabla^j + \sum_{j=p+m}^{k-1} (b_j - c + w_j)\nabla^j\right]\phi \simeq 0. \tag{21}$$

Therefore z is governed by the relation $N'z \simeq 0$, which completes the proof.

Global error analysis and an extension of the averaging method

This section is concerned with a direct error analysis approach that shows how to obtain any order of accuracy in the approximate solution by combining, linearly, a certain number of approximate solutions of lower order which depend on sufficiently many parameters. The analysis is applicable to general difference operators with arbitrary k, but we restrict the discussion, for the moment, to the Adams-type operator (9) with $k \leq 3p$. The approx-

imate solution z is again defined by (17) and our aim is to derive the equations analogous to (14) which make z an approximate solution of order k. Naturally, for $k \le 2p$, the conditions (14) will be reproduced. As before, for a parameter vector $\mathbf{u} = (u_1, \cdots, u_d)$ where $d = k - p$, let x be the solution of the difference equation

$$N(x) = \Delta x - h\Big[c\Delta f + f + \sum_{j=1}^{p-1} (b_j - c)\nabla^j f + \sum_{j=p}^{k-1} (b_j - c + u_{j-p+1})\nabla^j f\Big] = 0\,, \tag{22}$$

where $f = f(x)$. If y is the exact solution of (1) then, since $N(y) = \Omega(y)$, we obtain from (10) that

$$N(y) \simeq \sum_{i=1}^{d} \mu_i h^{p+i} y^{(p+i)}\,, \tag{23}$$

where $\simeq$ means that terms of order $(k+1)$ are neglected and where the μ_i are defined by (11). In the following, and in defining $N(y)$, we shall use the notation $F = f(y)$. Let $\varepsilon = x - y$ be the global discretization error at an arbitrary grid point then, as stated earlier, $\varepsilon = O(h^p)$. Subtracting (23) from (22) and neglecting terms $o(h^k)$ one obtains

$$\Delta\varepsilon - h\Big\{\Big[c\Delta + 1 + \sum_{j=1}^{p-1} (b_j - c)\nabla^j\Big]\Big[F_{(1)}\varepsilon + \tfrac{1}{2}F_{(2)}\varepsilon^{[2]}\Big] + \sum_{j=p}^{k-1} (b_j - c + u_{j-p+1})\nabla^j (F_{(1)}\varepsilon)\Big\} \simeq -\sum_{i=1}^{d} \mu_i h^{p+i} y^{(p+i)}\,, \tag{24}$$

where $F_{(1)}$, $F_{(2)}$ are the first and second (Fréchet) derivatives of F at the solution y and $\varepsilon^{[2]}$ denotes the pair $(\varepsilon, \varepsilon)$. Let $\varepsilon = h^p\eta$, substitute in (24) and divide by h^{p+1} to obtain

$$\frac{1}{h}\Delta\eta - \Big[c\Delta + 1 + \sum_{j=1}^{p-1} (b_j - c)\nabla^j\Big]\Big[F_{(1)}\eta + \tfrac{1}{2}h^p F_{(2)}\eta^{[2]}\Big] - \sum_{j=p}^{k-1} (b_j - c + u_{j-p+1})\nabla^j (F_{(1)}\eta) \approx \sum_{i=0}^{d-1} \mu_{i+1} h^i y^{(p+i+1)}\,, \tag{25}$$

where $\approx$ denotes that terms $o(h^{d-1})$ are neglected. To obtain the asymptotic expansion, in powers of h, of η we let

$$\eta \approx \sum_{l=0}^{d-1} h^l \eta_l\,, \tag{26}$$

expand the left side of (25) to order h^{d-1}, and use regular perturbation theory [12]. A straightforward but rather long calculation shows that the η_l are determined recursively by the equations

$$\dot\eta_j - F_{(1)}\eta_j = -\mu_{j+1} y^{(p+j+1)} - \sum_{i=1}^{j} [(i+1)!]^{-1} \frac{d^i}{dt^i}[\dot\eta_{j-i} - F_{(1)}\eta_{j-i}]\,, \quad 0 \le j \le p-1\,, \tag{27a}$$

$$\dot\eta_j - F_{(1)}\eta_j = -\mu_{j+1} y^{(p+j+1)} - \sum_{i-1}^{j} \frac{d^i}{dt^i}\{[(i+1)!]^{-1}[\dot\eta_{j-i} - \zeta_i F_{(1)}\eta_{j-i}]\} + \tfrac{1}{2}\sum_{i=0}^{j-p} [(i+1)!]^{-1} \frac{d^i}{dt^i}\theta_{j-p-i}\,, \quad p \le j \le d-1 \tag{27b}$$

where, if $\lambda_i = \sum_{j=0}^{i-p} (j!)^{-1}\gamma_{i-j,p}$ and $\xi_i = \sum_{j=p}^{i} \gamma_{ij}$, $\quad i > p$,

we have

$$\zeta_i = \begin{cases} [(i+1)!]^{-1}\,, & 0 \le i \le p-1 \\ [(i+1)!]^{-1} + c(\lambda_i - \xi_i) + \sum_{j=p}^{i} \gamma_{ij} u_{j-p+1}\,, & p \le i \le d-1 \end{cases} \tag{27c}$$

and

$$\theta_l = \sum_{j=0}^{l} F_{(2)}\eta_j\, \eta_{l-j}\,, \qquad 0 \le l \le d-1\,. \tag{27d}$$

The above equations, together with initial values which may be assumed to vanish or to be $o(h^{3p})$, define the functions η_j. In particular, the dependence of η_j on the parameters $u_1, \cdots, u_d$ (or equivalently $\mu_1, \cdots, \mu_d$) can be determined recursively. In fact, for $0 \le j \le p-1$, it is clear that η_j is a linear function of the μ_l, $1 \le l \le j+1$, with coefficients depending only on the solution y. Explicitly one has

$$\eta_j = \sum_{l=0}^{j} \mu_{l+1}\eta_{jl}\,, \qquad 0 \le j \le p-1\,, \tag{28}$$

where the η_{jl} satisfy

$$L\eta_{jj} = -y^{(p+j+1)}\,,$$

$$L\eta_{jl} = -\sum_{i=1}^{j-l} \Lambda_i \eta_{j-i,l}\,, \qquad 0 \le l \le j-1\,; \tag{29}$$

here $L = (d/dt) - F_{(1)}$ and

$$\Lambda_i = [(1+i)!]^{-1} \frac{d^i}{dt^i}\Big[\frac{d}{dt} - F_{(1)}\Big]\,.$$

We can now (re)prove that the approximate solution defined by (17) and (14), with $m = p$, has a global error $O(h^{2p})$. To see this, let $\varepsilon = z - y$. Then, in an obvious notation,

$$\varepsilon = \sum_{\rho=1}^{p+1} \nu_\rho (x_\rho - y) = \sum_{\rho=1}^{p+1} \nu_\rho \varepsilon_\rho = h^p \sum_{\rho=1}^{p+1} \nu_\rho \eta_\rho\,. \tag{30}$$

Substituting in (30) from (26) and (28) one obtains

$$\varepsilon = h^p \left[\sum_{\rho=1}^{p+1} \nu_\rho \sum_{j=0}^{p-1} h^j \sum_{l=0}^{j} \mu_{l+1,\rho} \eta_{jl} + O(h^p) \right]$$

$$= \sum_{j=0}^{p-1} h^{p+j} \left[\sum_{l=0}^{j} \left(\sum_{\rho=1}^{p+1} \nu_\rho \mu_{l+1,\rho} \right) \eta_{jl} \right] + O(h^{2p}) . \qquad (31)$$

Because of (14a) and (11), the first term on the right-hand side of (31) vanishes, which proves the assertion.

In order to find an approximate solution with accuracy higher than $2p$ we need to calculate the dependence of the errors η_j, $p \le j \le d-1$, on the parameter vector $\mathbf{u}$ explicitly. Taking into account the form of the right-hand side of (27b), and noting that the ζ_i depend linearly on $\mathbf{u}$, an induction argument shows that η_j has the form

$$\eta_j = \sum_{l=0}^{j} \mu_{l+1} \eta_{jl} + \sum_{l=0}^{j-p} \sum_{r=0}^{j-p-l} \mu_{l+1} \mu_{r+1} \eta_{jlr} ,$$

$$p \le j \le 2p-1 , \qquad (32)$$

where the coefficients η_{jl}, η_{jlr} depend only on the solution y.

As before, let x_ρ be the solution to $N_\rho x_\rho = 0$, $z = \sum_{\rho=1}^{\mathcal{N}} \nu_\rho x_\rho$, where $\mathcal{N}$ is to be specified later and $\varepsilon = z - y$. Then, using $\sum_{\rho=1}^{\mathcal{N}} \nu_\rho = 1$ and (32) we get

$$\varepsilon = \sum_{j=0}^{d-1} h^{p+j} \left\{ \sum_{l=0}^{j} \left(\sum_{\rho=1}^{\mathcal{N}} \nu_\rho \mu_{l+1,\rho} \right) \eta_{jl} \right\}$$

$$+ \sum_{j=p}^{d-1} h^{p+j} \left\{ \sum_{l=0}^{i-p} \sum_{r=0}^{i-p-l} \left(\sum_{\rho=1}^{\mathcal{N}} \nu_\rho \mu_{l+1,\rho} \mu_{r+1,\rho} \right) \eta_{jlr} \right\}$$

$$+ O(h^{p+d}) .$$

Hence, if the coefficients ν_ρ satisfy the equations

$$\sum_{\rho=1}^{\mathcal{N}} \nu_\rho = 1 ,$$

$$\sum_{\rho=1}^{\mathcal{N}} \nu_\rho \mu_{l+1,\rho} = 0 , \qquad 0 \le l \le d-1 \qquad (33)$$

$$\sum_{\rho=1}^{\mathcal{N}} \nu_\rho \mu_{l+1,\rho} \mu_{r+1,\rho} = 0 \qquad \begin{cases} 0 \le l \le d-p-1 \\ 0 < r \le d-p-1-l , \end{cases}$$

then $\varepsilon = O(h^{p+d})$, as desired. Note here that (33) represents a system of $\mathcal{N} = (d+1) + (1/2)(d-p)(d-p+1)$ linear equations for the $\mathcal{N}$ unknowns ν_ρ. The unique solvability of these equations, for almost all parameter vectors $\boldsymbol{\mu}_\rho$ (or $\mathbf{u}_\rho$), follows from the analyticity in $\mathbf{u}$, of the coefficient matrix.

For the sake of completeness we now record the equations defining the coefficients η_{jl}, η_{jlr}.

$$L\eta_{jl} = \begin{cases} -\sum_{i=1}^{p-1} \Lambda_i \eta_{j-i,l} - \sum_{i=p}^{j-l} \Lambda_{i,0} \eta_{j-i,l} & 0 \le l \le j-p \\ -\sum_{i=1}^{j-l} \Lambda_i \eta_{j-i,l} , & j-p < l \le j-1 \end{cases}$$

$$L\eta_{jj} = -\, y^{(p+j+1)} \qquad (34)$$

and

$$L\eta_{jlr} = \begin{cases} -\sum_{i=1}^{j-p-l-r} \Lambda_i \eta_{j-i,l,r} + (\Lambda^*_{jlr} - \Lambda_{jlr}) , & \begin{cases} 0 \le l \le j-p-1 \\ 0 \le r \le j-p-1-l \end{cases} \\ \Lambda^*_{jlr} - \Lambda_{jlr} & \begin{cases} 0 \le l \le j-p \\ r = j-p-l . \end{cases} \end{cases}$$

The operators in (34) are defined as follows. Let

$$\zeta_{i,0} = [(i+1)!]^{-1} + c(\lambda_i - \xi_i) \qquad p < i \le d-1 , \quad 0 \le j \le i-p$$

$$\zeta_{i,j+1} = \sum_{l=j}^{i-p+1} \gamma_{i,l+p} \hat{\gamma}_{p+l,p+j} ,$$

where $\hat{\gamma}$ is the triangular matrix inverse of γ,

$$R_i = [2(i+1)!]^{-1} \, d^i/dt^i .$$

$$\left.\begin{aligned} \Lambda_{i,0} &= \frac{d^i}{dt^i} \left\{ [(i+1)!]^{-1} \frac{d}{dt} - \zeta_{i,0} F_{(1)} \right\} \\ \Lambda_{i,j} &= \frac{d^i}{dt^i} [\zeta_{i,j} F_{(1)}] \end{aligned}\right\} \quad \begin{cases} p \le i \le d-1 \\ 0 \le j \le i-p \end{cases}$$

$$\left.\begin{aligned} \Lambda^*_{jlr} &= \sum_{i=p+l}^{j-r} \Lambda_{i,l+1} \eta_{j-i,r} \\ \Lambda_{jlr} &= \sum_{i=0}^{j-p-l-r} \sum_{n=l}^{j-p-i-r} \\ &\quad \times R_i [F_{(2)} \eta_{nl} \eta_{j-p-i-n,r}] \end{aligned}\right\} \quad \begin{cases} p \le j \le d-1 \\ 0 \le l \le j-p \\ 0 \le r \le j-p-l . \end{cases} \qquad (35)$$

We remark here that if we impose the constraints (33) with d replaced by d', $d' \le d-1$, then the global error $\varepsilon = O(h^{p+d'} \eta_{d'})$. The equations (32) and (34) would then allow for computing an estimate for $\eta_{d'}$ which could be used for error control.

It is clear that this perturbation method can be used to arrive at solutions of accuracy $(p+d)$ for any d, but we defer a complete discussion to a future paper since the procedure of the second section will suffice for the following sections. It should also be remarked that the perturbation procedure is applicable to any consistent scheme [13] and we end this section by stating the results of its application to the two-parameter operator

$$\Omega = 6\Delta + (3 + 12u_1 - 2u_2)\Delta^2$$

$$- D\{6 + (6 + 12u_1 - 2u_2)\Delta + (1 + 5u_1 - u_2)\Delta^2\} ,$$

which is of order $p \ge 2$ for any $\mathbf{u} = (u_1, u_2)$. For $\mathbf{u} = 0$, Ω represents Simpson's Rule. The general operator Ω is

equivalent to that associated with formula (12) of Ref. 14 via the parameter transformation

$$\left.\begin{aligned} u_1 &= (-2 - 2\beta_0 + \beta_1)(1 + \beta_0 + \beta_1)^{-1} \\ u_2 &= (-15 - 9\beta_0 + 6\beta_1)(1 + \beta_0 + \beta_1)^{-1} \end{aligned}\right\}$$

$$\Longleftrightarrow \left\{\begin{aligned} \beta_0 &= (1 - 7u_1 + u_2)(1 + 5u_1 - u_2)^{-1} \\ \beta_1 &= (4 + 2u_1)(1 + 5u_1 - u_2)^{-1} \end{aligned}\right\}.$$

The operator Ω was shown in Ref. 8 to be A-stable in an appropriate parameter domain. In terms of the present parameters, this domain is $\{u_2 < 6u_1\,,\quad u_1 < -1/2\}$.

If $z = \sum_{\rho=1}^{4} \nu_\rho x$ as before, where

$$\sum_{\rho=1}^{4} \nu_\rho = 1\,,$$

$$\sum_{\rho=1}^{4} \nu_\rho u_{i,\rho} = 0\,, \qquad i = 1\,,2\,,$$

$$\sum_{\rho=1}^{4} \nu_\rho (2u_{1,\rho}{}^2 - \tfrac{1}{3} u_{1,\rho} u_{2,\rho}) = 0\,,$$

then z is an approximate solution of order four.

A-stability analysis

In the rest of this paper we consider in detail three operators of the form (9) and the algorithms for producing approximate solutions associated with them. The operators are defined by

$$\Omega_1 = \Delta - D[c\Delta + 1 + (\tfrac{1}{2} - c + r)\nabla]\,; \qquad p = 1\,,\ k = 2\,, \tag{36}$$

$$\Omega_2 = \Delta - D[c\Delta + 1 + (\tfrac{1}{2} - c)\nabla + (\tfrac{5}{12} - c + r)\nabla^2]\,; \qquad p = 2\,,\ k = 3\,, \tag{37}$$

$$\Omega_3 = \Delta - D[c\Delta + 1 + (\tfrac{1}{2} - c)\nabla + (\tfrac{5}{12} - c + r)\nabla^2 + (\tfrac{3}{8} - c + s)\nabla^3]\,; \qquad p = 2\,,\ k = 4\,, \tag{38}$$

where r, s are the primary parameters.

This section is concerned with determining the domains, in the parameter spaces, of A-stability of Ω_1, Ω_2, Ω_3. To this end we transform (36), (37) and (38) to the ordinate form and use a criterion derived by one of the authors [8]. If $q(z) = \rho(z)/\sigma(z)$, where ρ, σ are the familiar polynomials associated with linear multistep operators, then the conditions for A-stability are N1: that the zeros of σ lie inside the unit circle, and N2: that $P(x) \geq 0$, $-1 \leq x \leq 1$, where P is a certain polynomial depending on the coefficients of ρ and σ. This latter condition is equivalent to the requirement that Re $q(z) \geq 0$ for all $|z| = 1$. For a k-step operator, the condition N1 is equivalent to demanding that the roots of $\sigma(\zeta) = (\zeta - 1)^k \sigma(z(\zeta))$, $z(\zeta) = (\zeta + 1)/(\zeta - 1)$, lie in the left half of the ζ-plane.

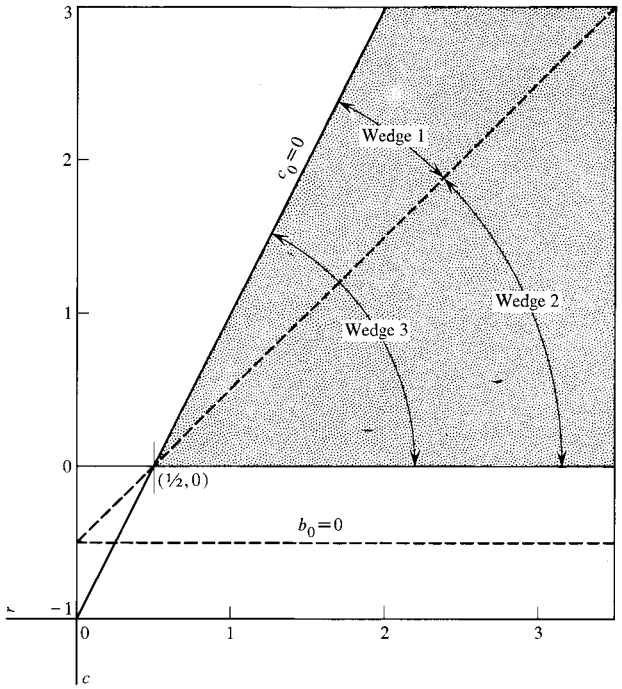

Figure 1 A-stability domain of Ω_1.

• *A-stability of* Ω_1

In the case of the operator Ω_1, we find that $2\sigma(z) = 2cz^2 + (3 - 4c + 2r)z - (1 - 2c + 2r)$ and $\hat{\sigma}(\zeta) = \zeta^2 + (1 + 2r)\zeta + (-2 + 4c - 2r)$. By the Routh criterion [15] the roots of $\hat{\sigma}$ are in the left half-plane (and thus N1 is satisfied) if and only if the inequalities $1 + 2r > 0$ and $1 - 2c + r < 0$ hold. For the operator Ω_1, condition N2 requires that $(x - 1)[(-1 + 2c - 2r)x + (1 - 2c)] \geq 0$, $-1 \leq x \leq 1$, be satisfied, which is the case if and only if either $-1 + 2c \geq r \geq -\frac{1}{2} + c$ or $-\frac{1}{2} + c \geq r \geq 0$ is true; i.e., if $(r\,, c)$ is either in wedge 1 or wedge 2 of Fig. 1. Except on a portion of the boundary, these inequalities imply those associated with condition N1 and, thus, Ω_1 is A-stable if $0 \leq r < -1 + 2c$; i.e., if the point $(r\,, c)$ is in the partially open wedge 3 of Fig. 1.

• *A-stability of* Ω_2

In the Ω_2-case, $3\hat{\sigma}(\zeta) = 3\zeta^3 + 6\zeta^2 + (2 + 12r)\zeta + (-11 + 24c - 12r)$ and N1 is satisfied if and only if $-\frac{5}{12} + \frac{2}{3}c < r < -\frac{11}{12} + 2c$ holds. The condition N2 requires that $(x - 1)^2[(\frac{5}{3} - 4c + 4r)x - (\frac{1}{6} - 2r)] \geq 0$ for $-1 \leq x \leq 1$, which is true if and only if either $2c - \frac{11}{12} \geq r \geq c - \frac{5}{12}$ or $c - \frac{5}{12} \geq r \geq \frac{2}{3}c - \frac{1}{4}$ holds (wedges 1 and 2 of Fig. 2). As in the previous case, N2 thus essentially implies N1 and Ω_2 is A-stable in $\frac{2}{3}c - \frac{1}{4} \leq r < 2c - \frac{11}{12}$ (wedge 3 of Fig. 2).

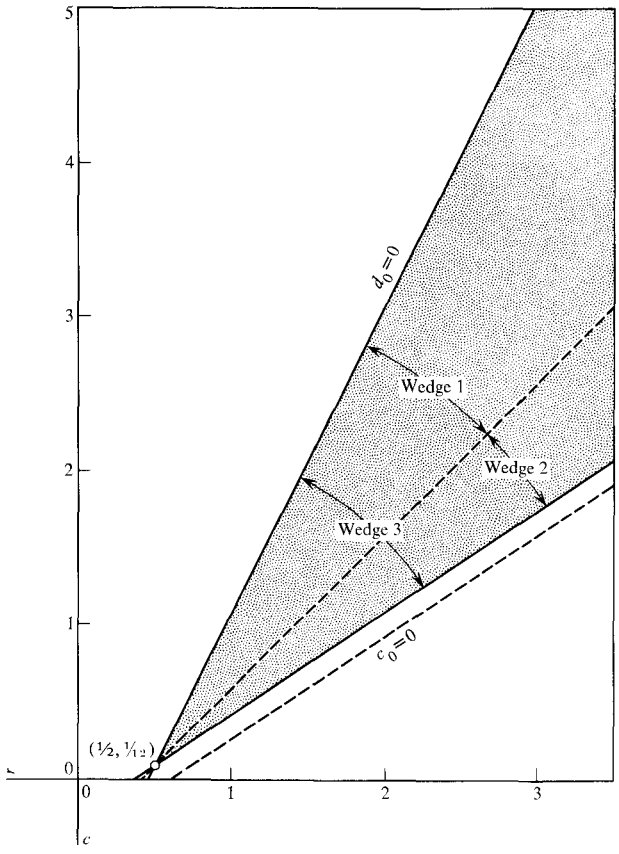

Figure 2 A-stability domain of Ω_2.

• *A-stability of* Ω_3

In the case of Ω_3,

$$3\hat{\sigma}(\zeta) = 3\zeta^4 + 9\zeta^3 + (8 + 12r)\zeta^2 + 24s\zeta + (-20 + 48c - 12r - 24s)\,. \tag{39}$$

In the notation of Ref. 15, the Routh criterion implies the inequalities

$$\tfrac{1}{4}c_0 = 2 + 3r - 2s > 0 \tag{40}$$

$$\tfrac{1}{3}d_0 = (15 - 36c + 9r + 34s + 24rs - 16s^2)/(2 + 3r - 2s) > 0 \tag{41}$$

$$\tfrac{1}{4}e_0 = -5 + 12c - 3r - 6s > 0\,. \tag{42}$$

For any $c \neq 0$ and $c_0 \neq 0$, $d_0 = 0$ represents a nondegenerate hyperbola with center $(-\frac{23}{12}, -\frac{3}{8})$ and asymptotes As1: $s = -\frac{3}{8}$, As2: $s = \frac{3}{2}r + \frac{5}{2}$; the latter is parallel to the line $c_0 = 0$ (see Fig. 3). The hyperbola intersects both $c_0 = 0$ and $e_0 = 0$ at the point $P_2 = (-\frac{11}{12} + c, -\frac{3}{8} + \frac{3}{2}c)$ and $e_0 = 0$ once more at $P_1 = (-\frac{5}{3} + 4c, 0)$. P_1 and P_2 coalesce for $c = \frac{1}{4}$. If $c > 0$, that branch of the hyperbola $d_0 > 0$ which lies in $s < -\frac{3}{8}$ (the bottom left-hand part) may be disregarded since it does not intersect $c_0 > 0$. It will be shown later on that N2 cannot be satisfied if $c < \frac{1}{2}$; we thus restrict ourselves to the case $c \geq \frac{1}{2}$, even though N1 can be satisfied for $c < 0$, which may be useful for integrating unstable problems.

In the case of Ω_3, condition N2 requires that $(x-1)^2 Q(x) \geq 0$, $-1 \leq x \leq 1$, where

$$Q(x) = (-9 + 24c - 24s)x^2 + (5 - 12c + 12r)x + (4 - 12c + 6r + 12s)\,; \tag{43}$$

which is true if and only if $Q(x) \geq 0$, $-1 \leq x \leq 1$.

The envelope of the lines $Q(x) = 0$, for $-1 \leq x \leq 1$, is that part of the ellipse

$$6r^2 + 24rs + 48s^2 + (14 - 36c)r + (34 - 96c)s + \tfrac{1}{24}(13 - 36c)^2 = 0\,, \tag{44}$$

which lies "between" $P_4 = (-\frac{13}{60} + \frac{3}{5}c, -\frac{13}{40} + \frac{9}{10}c)$, the tangent point on the tangent t_1: $s = 3r/2$ corresponding to $x = +1$, and $P_3 = (-\frac{17}{12} + 3c, -\frac{1}{8} + \frac{1}{2}c)$, the tangent point on the tangent t_2 (corresponding to $x = -1$); t_2 coalesces with $e_0 = 0$. This may be seen by calculating $s = s(x)$, the s-coordinate of the tangent point, and showing that it decreases between $P_4(x = 1)$ and $P_7 = (-\frac{17}{12} + 3c, \frac{1}{8})$, corresponding to $x = -\frac{1}{2}$, and increases between P_7 and $P_3(x = -1)$, so that the derivative ds/dx vanishes only once in $-1 \leq x \leq 1$. The ellipse lies between the two parallel tangents t_2 and t_3, the latter corresponding to $x = 0$ and having the tangent point $P_6 = (-\frac{5}{12} + c, -\frac{1}{8} + \frac{1}{2}c)$. From (43) with $x = 0$ and $x = -1$, one obtains $s \geq -\frac{1}{2}r + c - \frac{1}{3}$ and $s \leq (-\frac{5}{6} + 2c) - \frac{1}{2}r$, respectively, which are compatible only if $c \geq \frac{1}{2}$, as mentioned earlier. In this case, N2 is satisfied in the closed set $\bar{S}$ represented by the shaded area of Fig. 3, bounded by the lower part of the ellipse and the tangents t_1 and t_2. We remark here that the center of the ellipse is $(-\frac{11}{12} + 2c, -\frac{1}{8} + \frac{1}{2}c)$ which, as c varies, lies on the line $s = \frac{1}{4}r + \frac{5}{48}$ passing through the center of the hyperbola $d_0 = 0$. Also, independently of c, the angle between the positive r-axis and the major axis of the ellipse is $-14°50'$.

We now show that Ω_3 is A-stable in the half-closed set S, obtained by deleting from $\bar{S}$ the closed boundary segment $[P_3, P_5]$. We do this by proving that all points of S satisfy the strict inequalities imposed by N1. First, the inequality (42) defines the same open half-plane as the strict inequality $Q(-1) > 0$ (the one below the tangent t_2), containing all of S. Second, all points of S lie in the half-plane $s \geq \frac{3}{2}r$ bordered by t_1 and thus satisfy the weaker inequality (40). Third, to see that all of S lies in $d_0 > 0$, it is sufficient (as P_3, for example, is in $d_0 > 0$) to show that the boundary of S does not intersect the hyperbola. Now, if $P_5(-\frac{5}{12} + c, -\frac{5}{8} + \frac{3}{2}c)$ is the intersection point of t_1 and t_2, it is very easy to show that the segment $[P_3P_5]$ lies strictly between the two intersection points P_1 and P_2 of t_2 with the hyperbola and thus all of

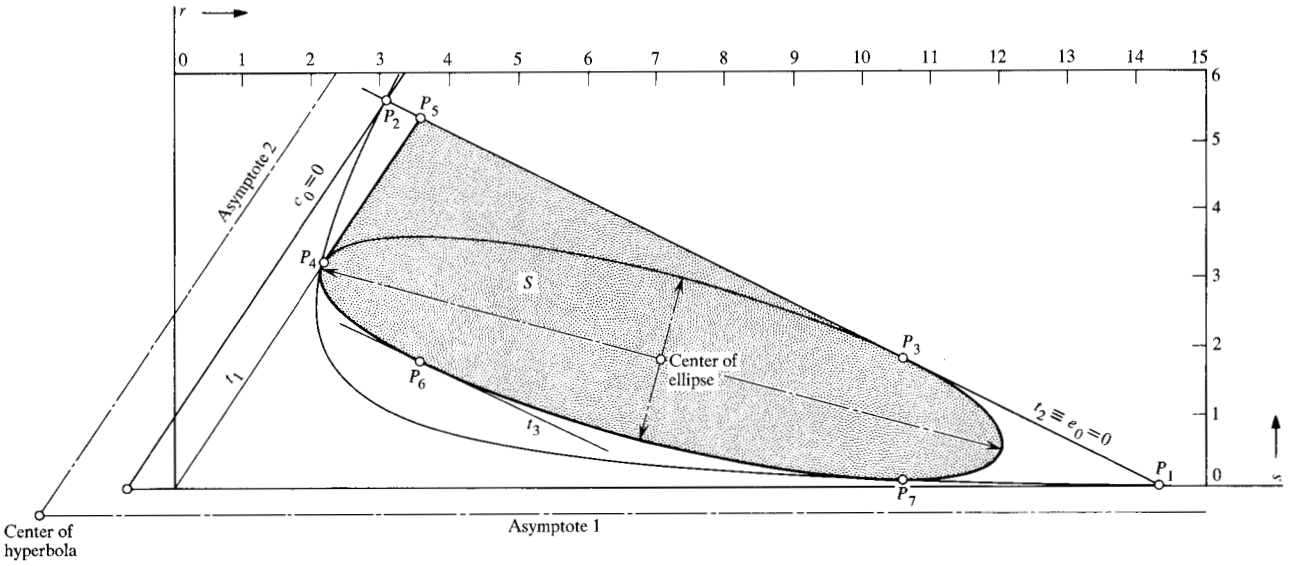

Figure 3 A-stability domain of Ω_3 for $c = 4$.

$[P_3P_5]$ is in $d_0 > 0$. Similarly, both P_4 and P_5, and thus the whole segment $[P_4P_5]$, lie on the $(d_0 > 0)$-side of the single intersection point of t_1 with the hyperbola (independently of c, the r-coordinates of P_4 and of that point differ by $+\frac{1}{30}$). Finally, to demonstrate that the elliptical part of the boundary of S does not intersect the hyperbola, we prove that this is true for the ellipse as a whole. Substituting from (44) into $d_0 = 0$ and letting $s = 3\sigma/4$, the points of intersection would have to correspond to real solutions, with respect to σ, of

$$(72\sigma^2 - 24\sigma + 2)c^2 + (-120\sigma^3 - 28\sigma^2 + 14\sigma - 1)c + (50\sigma^4 + 40\sigma^3 + 3\sigma^2 - 2\sigma + 1/8) = 0. \quad (45)$$

But the discriminant, with respect to c, of (45) equals $-216\sigma^2$ and thus, for any real σ, Eq. (45) is not satisfied by any real value of c. This is equivalent to saying that, for any real value of c, this equation has no real solutions with respect to σ (except for $\sigma = 0$ which corresponds to the double root $c = \frac{1}{4}$), as required.

Algorithm A4

Consider the nonlinear difference equation

$$N(w) \equiv \Delta w - h[c\Delta f + f + (\tfrac{1}{2} - c)\nabla f + (\tfrac{5}{12} - c + r)\nabla^2 f + (\tfrac{3}{8} - c + s)\nabla^3 f] = 0, \quad (46)$$

associated with Ω_3 [Eq. (38)] of the previous section, with $f = f(w)$. According to the results of the second section, three different solutions of (46), corresponding to three noncollinear points (r, s) chosen from the A-stability domain of Ω_3, are needed to produce a fourth-order A-stable approximate solution. As it stands, the procedure for computing this higher-order solution is costly from a computational point of view, except possibly in a parallel mode. It would involve, at each integration step, the solution of three systems of nonlinear algebraic equations.

We now describe a finite algorithm, which we call "A4," for achieving the same purpose much more efficiently. It consists of two steps. First, instead of solving (46) exactly, we show that an approximation, obtained via cubic extrapolation of previous data [16] followed by a single Newton-Raphson step around this "prediction," is sufficient to maintain the validity of the theorem in the second section. Second, only one approximate solution is calculated in this way; the other two—more precisely approximations thereof—are calculated from linearizations around the first one.

• *Step 1, finite Newton-Raphson approximation*

Starting with initial data with error at most $O(h^5)$, we calculate the approximate solution x^+, at $t^+ = (n+1)h$, from previous data as follows. Let x be the Newton solution at $t = nh$, then compute a "predicted" value $\tilde{x}^+$ as follows:

$$\tilde{x}^+ = x + \Delta\tilde{x},$$
$$\Delta\tilde{x} = \nabla x + \nabla^2 x + \nabla^3 x. \quad (47)$$

The correction θx^+ is obtained from linearizing N about $\tilde{x}^+$ and is defined by

$$N^*[x] = [I - hc\tilde{f}^+_{(1)}]\theta x^+ + \{\Delta\tilde{x} - h[c\tilde{f}^+ + (1-c)f + (\tfrac{1}{2} - c)\nabla f + (\tfrac{5}{12} - c + r)\nabla^2 f + (\tfrac{3}{8} - c + s)\nabla^3 f]\} = 0, \quad (48)$$

where $f=f(x)$, $\tilde{f}^{+}=f(\tilde{x}^{+})$ etc. Define

$$x^{+}=x+\Delta x\,,\qquad \Delta x=\Delta\tilde{x}+\theta x^{+}\,. \tag{49}$$

We now derive an equation governing the global discretization error $\varepsilon=x-y$ at an arbitrary point $t=nh$. To this effect we calculate $N^{*}(y)$, where y is the exact solution of (1), denoting again by the symbol "$\simeq$" that we are neglecting errors which are $O(h^{5})$. The notation ϕ, $\tilde{\phi}^{+}$, etc. denotes $f(y)$, $f(\tilde{y}^{+})$, etc. First, let

$$\tilde{y}^{+}=(1+\nabla+\nabla^{2}+\nabla^{3})y\simeq[1+D+\tfrac{1}{2}D^{2}+\tfrac{1}{6}D^{3}-\tfrac{23}{24}D^{4}]y\,. \tag{50}$$

Then

$$\Delta\tilde{y}\simeq\left[\left(\sum_{i=1}^{3}\frac{1}{i!}D^{i}\right)-\tfrac{23}{24}D^{4}\right]y \tag{51}$$

and hence

$$hc\tilde{\phi}^{+}\simeq hc\phi+hc\phi_{(1)}\,[Dy+\tfrac{1}{2}D^{2}y+\tfrac{1}{6}D^{3}y]$$
$$+\tfrac{1}{2}hc\,\phi_{(2)}[\,(Dy)^{[2]}+(Dy)\,(D^{2}y)\,]+\tfrac{1}{6}hc\,\phi_{(3)}(Dy)^{[3]}\,.$$

Using the differential equation, one obtains

$$hc\tilde{\phi}^{+}\simeq c\sum_{i=1}^{4}\frac{1}{(i-1)!}D^{i}y\,. \tag{52}$$

Also

$$h[(1-c)+(\tfrac{1}{2}-c)\nabla+(\tfrac{5}{12}-c+r)\nabla^{2}+(\tfrac{3}{8}-c+s)\nabla^{3}]\phi$$
$$\simeq[(1-c)D+(\tfrac{1}{2}-c)D^{2}+(\tfrac{1}{6}-\tfrac{1}{2}c+r)D^{3}+$$
$$(\tfrac{1}{24}-\tfrac{1}{6}c-r+s)D^{4}]y\,. \tag{53}$$

From (51), (52), and (53) the $\{\quad\}$ in $N^{*}(y)$ is given by

$$\{\quad\}\simeq-rh^{3}y^{(3)}-(1-r+s)h^{4}y^{(4)}\,. \tag{54}$$

By definition $\theta y^{+}=y^{+}-\tilde{y}^{+}\simeq h^{4}y^{(4)}$, and therefore

$$[1-hc\tilde{\phi}^{+}{}_{(1)}]\theta y^{+}\simeq h^{4}y^{(4)}\,. \tag{55}$$

From (54) and (55) one gets

$$N^{*}(y)\simeq-rh^{3}y^{(3)}-(s-r)h^{4}y^{(4)}\,. \tag{56}$$

Write equation (48) in the form

$$[I-hc\tilde{f}^{+}{}_{(1)}]\theta x^{+}-hM(x\,,\theta x^{+})=-\Delta\tilde{x}\,, \tag{57}$$

where M is some nonlinear difference operator. Then

$$[I-hc\tilde{\phi}^{+}{}_{(1)}]\theta y^{+}-hM(y\,,\theta y^{+})\simeq-\Delta\tilde{y}-rh^{3}y^{(3)}$$
$$-(s-r)h^{4}y^{(4)}\,. \tag{58}$$

Subtracting, one gets

$$\theta\varepsilon^{+}=-\Delta\tilde{\varepsilon}+O(h^{3})=O(h^{3})\,, \tag{59}$$

where we have used the fact that $\tilde{\varepsilon}=O(h^{2})$ since the global error ε_{ν} for $0\le\nu\le n$, is $O(h^{2})$. This latter fact is a consequence of the stability and the second order accuracy of N^{*} (see, for example, Spijker [17]). Now, since $\theta x^{+}=\theta y^{+}+\theta\varepsilon^{+}$ and θy^{+} is $O(h^{4})$ by (55), it follows that

$$\theta x^{+}=O(h^{3})\,. \tag{60}$$

Subtracting (56) from (48) and using (60) and the relation

$$\Delta\varepsilon=\Delta\tilde{\varepsilon}+\theta\varepsilon^{+}\,, \tag{61}$$

one is led to

$$[I-hc\tilde{\phi}^{+}{}_{(1)}]\Delta\varepsilon-h[c\tilde{\phi}^{+}{}_{(1)}\varepsilon+(1-c)\phi_{(1)}\varepsilon$$
$$+(\tfrac{1}{2}-c)\nabla\,(\phi_{(1)}\varepsilon)\,]\simeq rh^{3}y^{(3)}+(s-r)h^{4}y^{(4)}\,, \tag{62}$$

which is similar to the Eq. (24) governing the global error between y and the exact solution $N(w)=0$. It is now clear that if the approximate solution x^{+} is used — in place of the exact solution of $N=0$ — in forming the linear combination (17), then the global error, being a linear homogenous function of the parameters r, s because of (62), will be of $O(h^{4})$, as asserted.

The above conclusion could have been arrived at in an indirect manner. Since N^{*} is of second-order accuracy, one has initially, with sufficiently accurate starting values for x and w, $|w^{+}-y^{+}|=O(h^{3})$ and $|\tilde{x}^{+}-y^{+}|=O(h^{4})$. Then, $|\tilde{x}^{+}-w^{+}|=O(h^{3})$ and thus $|x^{+}-w^{+}|=O(h^{7})$ (Liniger, [18]). The stability of N^{*} then shows that $|x^{+}-w|=O(h^{6})$ globally. Therefore, to $O(h^{4})$, one may use x in lieu of the exact solution w of $N=0$ in forming the averaged solution (17).

• *Step II, perturbation about one solution*

Let $x=x_{1}$ be the solution of (48) corresponding to the point $(r_{1}\,,s_{1})$ and $f=f_{1}$, etc. Let x_{ρ}, $\rho=2\,,3$, be the solutions defined by $N^{*}{}_{\rho}\,[x_{\rho}]=0$ in an obvious notation. We note that

1. if $\xi_{\rho}=x_{\rho}-x$, then $\xi_{\rho}=O(h^{2})$ globally, and
2. the curly bracket in (48) is $O(h^{3})$ — see (59); hence one can replace the term $\tilde{f}^{+}{}_{(1)\rho}$ in $N^{*}{}_{\rho}$ by $\tilde{f}^{+}{}_{(1)}$ as this causes an error $O(h^{6})$ in the solution.

Linearizing $N^{*}{}_{\rho}(x_{\rho})=0$ around x and using the fact that, on any sufficiently smooth function, ∇^{j} raises the order by j, one thus obtains

$$[I-hc\,\tilde{f}^{+}{}_{(1)}]\theta x^{+}{}_{\rho}+\Delta\tilde{x}+\Delta\tilde{\xi}_{\rho}-h\{c\tilde{f}^{+}+c\tilde{f}^{+}{}_{(1)}\tilde{\xi}^{+}{}_{\rho}$$
$$+(1-c)f+(1-c)f_{(1)}\xi_{\rho}$$
$$+(\tfrac{1}{2}-c)\nabla f+(\tfrac{1}{2}-c)\nabla[f_{(1)}\xi_{\rho}]+(\tfrac{5}{12}-c+r_{\rho})\nabla^{2}f$$
$$+(\tfrac{3}{8}-c+s_{\rho})\nabla^{3}f\}\simeq 0\,, \tag{63}$$

where $\tilde{\xi}^{+}{}_{\rho}=\tilde{x}^{+}{}_{\rho}-\tilde{x}^{+}$, $\Delta\tilde{\xi}_{\rho}=\Delta\tilde{x}_{\rho}-\Delta\tilde{x}$, etc.

Subtracting (63) from the corresponding equation for θx^{+}, we get

$$[I - hc\tilde{f}^{+}_{(1)}]\theta\xi^{+}_{\rho} + \Delta\tilde{\xi}_{\rho} - h\{c\tilde{f}^{+}_{(1)}\tilde{\xi}^{+}_{\rho} + (1-c)f_{(1)}\xi_{\rho} + (\tfrac{1}{2}-c)\nabla(f_{(1)}\xi_{\rho}) + (r_{\rho}-r_1)\nabla^2 f + (s_{\rho}-s_1)\nabla^3 f\} \simeq 0\,. \tag{64}$$

Solving (64) for $\theta\xi^{+}_{\rho}$, defining $\xi^{+}_{\rho} = \tilde{\xi}_{\rho} + \theta\xi^{+}_{\rho}$ and $x^{+}_{\rho} = x^{+} + \xi^{+}_{\rho}$, and using (14b), one finds that the averaged solution is given by

$$z^{+} = x^{+} + \nu_2\xi^{+}_{2} + \nu_3\xi^{+}_{3}\,. \tag{65}$$

Two further remarks are appropriate here:

1. The updating of the differences $\nabla^3 x$, $\nabla^2 x$, $\cdots$ is particularly simple. In fact, by using the definition of $\bar{x}^{+}$, one finds that

$$\nabla^3 x^{+} = \theta x^{+} + \nabla^3 x\,; \tag{66}$$

thus $\nabla^3 x$ updates the same way as x itself. The other differences are obtained by "integrating," i.e.,

$$\nabla^2 x^{+} = \nabla^2 x + \nabla^3 x^{+}\,,$$
$$\nabla x^{+} = \nabla x + \nabla^2 x^{+}\,. \tag{67}$$

Similar relations hold for the ξ_{ρ}, $\rho = 2, 3$. Updating the differences in this manner may help reduce the effect of rounding errors.

2. Each of the differences of f is computed from lower order differences, rather than from the values of f themselves, which reduces the storage requirements.

We conclude this section by summing up the results in the form of the finite algorithm A4. Similar algorithms, designated A2 and A3, have been derived in conjunction with Ω_1, Ω_2 (36) and (37). They correspond, in the notation of the second section, to the case $m = d = 1$ and produce solutions with global errors of orders two and three, respectively. All three algorithms were tested on the problems of the next section.

• *Algorithm A4*

Given (r_1, s_1), (r_2, s_2), (r_3, s_3), representing three non-collinear points in the A-stability domain of Fig. 3, let ν_1, ν_2, ν_3 satisfy (14). Given [19] $z\,;\ x, \nabla x, \nabla^2 x, \nabla^3 x\,;\ f, \nabla f, \nabla^2 f, \nabla^3 f\,;\ \xi_{\rho}, \nabla\xi_{\rho}, \nabla^2\xi_{\rho}, \nabla^3\xi_{\rho}\,;\ (\tilde{f}_x\xi_{\rho}), \nabla(\tilde{f}_x\xi_{\rho})$, $\rho = 2, 3$:

Predict first solution:

$$\Delta\bar{x} = \nabla x + \nabla^2 x + \nabla^3 x\,,\quad \bar{x}^{+} = x + \Delta\bar{x}\,,\quad \tilde{f}^{+} = f(\bar{x}^{+})\,,$$
$$\tilde{f}^{+}_{x} = f_x(\bar{x}^{+})\,. \tag{68}$$

Compute Newton correction by solving

$$(I - hc\tilde{f}^{+}_{x})\theta x^{+} = -\Delta\bar{x} + h[c\tilde{f}^{+} + (1-c)f + (\tfrac{1}{2}-c)\nabla f + (\tfrac{5}{12}-c+r_1)\nabla^2 f + (\tfrac{3}{8}-c+s_1)\nabla^3 f]\,. \tag{69}$$

Correct first solution:

$$x^{+} = \bar{x}^{+} + \theta x^{+}\,,\ \nabla^3 x^{+} = \nabla^3 x + \theta x^{+}\,,\ \nabla^2 x^{+} = \nabla^2 x + \nabla^3 x^{+}\,,\ \nabla x^{+} = \nabla x + \nabla^2 x^{+}\,. \tag{70}$$

Predict perturbations:

$$\Delta\tilde{\xi}_{\rho} = \nabla\xi_{\rho} + \nabla^2\xi_{\rho} + \nabla^3\xi_{\rho}\,,\ \tilde{\xi}^{+}_{\rho} = \xi_{\rho} + \Delta\tilde{\xi}_{\rho}\,,\quad \rho = 2, 3\,. \tag{71}$$

Compute Newton corrections by solving:

$$(I - hc\tilde{f}^{+}_{x})\theta\xi^{+}_{\rho} = -\Delta\tilde{\xi}_{\rho} + h[c\tilde{f}^{+}_{x}\tilde{\xi}^{+}_{\rho} + (1-c)\tilde{f}_x\xi_{\rho} + (\tfrac{1}{2}-c)\nabla(\tilde{f}_x\xi_{\rho}) + (r_{\rho}-r_1)\nabla^2 f + (s_{\rho}-s_1)\nabla^3 f]\,,\quad \rho = 2, 3\,. \tag{72}$$

Correct perturbations:

$$\xi^{+}_{\rho} = \tilde{\xi}^{+}_{\rho} + \theta\xi^{+}_{\rho}\,,\ \nabla^3\xi^{+}_{\rho} = \nabla^3\xi_{\rho} + \theta\xi^{+}_{\rho}\,,\ \nabla^2\xi^{+}_{\rho} = \nabla^2\xi_{\rho} + \nabla^3\xi^{+}_{\rho}\,,\ \nabla\xi^{+}_{\rho} = \nabla\xi_{\rho} + \nabla^2\xi^{+}_{\rho}\,. \tag{73}$$

Update auxiliary quantities: Compute

$$\tilde{f}^{+}_{x}\xi^{+}_{\rho}\,,\ \nabla(\tilde{f}_x\xi_{\rho})^{+} = \tilde{f}^{+}_{x}\xi^{+}_{\rho} - f_x\xi_{\rho}\,;\qquad \rho = 2, 3\,. \tag{74}$$

Reevaluate "derivatives":

$$f^{+} = f(x^{+})\,,\ \nabla f^{+} = f^{+} - f\,,\ \nabla^2 f^{+} = \nabla f^{+} - \nabla f\,,$$
$$\nabla^3 f^{+} = \nabla^2 f^{+} - \nabla^2 f\,, \tag{75}$$

and update the averaged solution

$$z^{+} = x^{+} + \nu_2\xi^{+}_{2} + \nu_3\xi^{+}_{3}\,. \tag{76}$$

Numerical results and remarks

The algorithms A2, A3, A4 were used to solve the following two test problems P1 and P2:

$$\text{P1:}\qquad \dot{x} = -2000x + 1000y + 1000\,,$$
$$\dot{y} = x - y\,. \tag{77}$$

This is the (slightly rescaled) first example in the survey paper of Bjurel [1], section 4, p. 1, and Example A of Ref. 20. It is a linear problem with constant coefficients having the eigenvalues $\lambda_1 = -2000.500125$, $\lambda_2 = -0.499875$ with a "stiffness ratio" of 4000. The exact solution of this problem with $x(0) = y(0) = 0$ is $x = 1 - 0.49975{,}00000\, e^{-\lambda_1 t} - 0.50025{,}00000\, e^{-\lambda_2 t}$; $y = 1 + 0.00024{,}99374{,}688\, e^{-\lambda_1 t} - 1.00024{,}99374{,}688\, e^{-\lambda_2 t}$. To stay away from the boundary layer [16] of amplitude 0.5 in the x-component at $t = 0$, the integration by A2-A4 was carried out from $t = 1$ to $t = 4$, using the exact solution values near $t = 1$ as starting values in the multistep methods. The results are plotted on the doubly logarithmic graph of Fig. 4. Specifically, the maximum relative accumulated truncation error in absolute value in

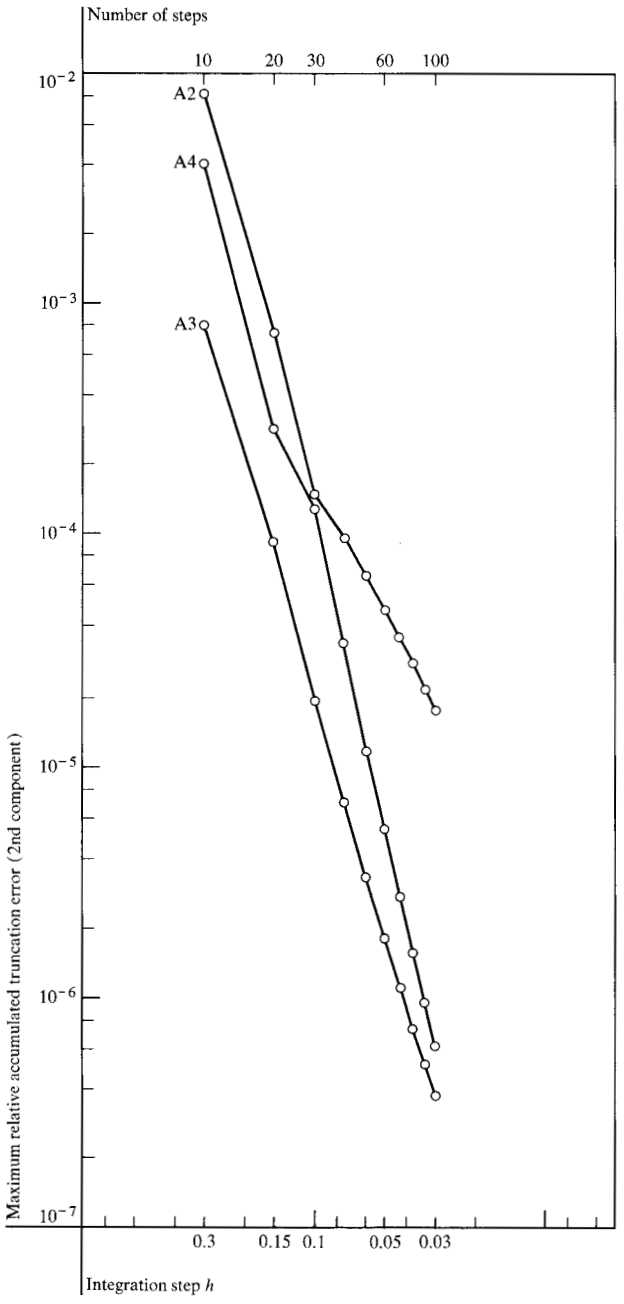

Figure 4 Accumulated truncation error for test problem P1.

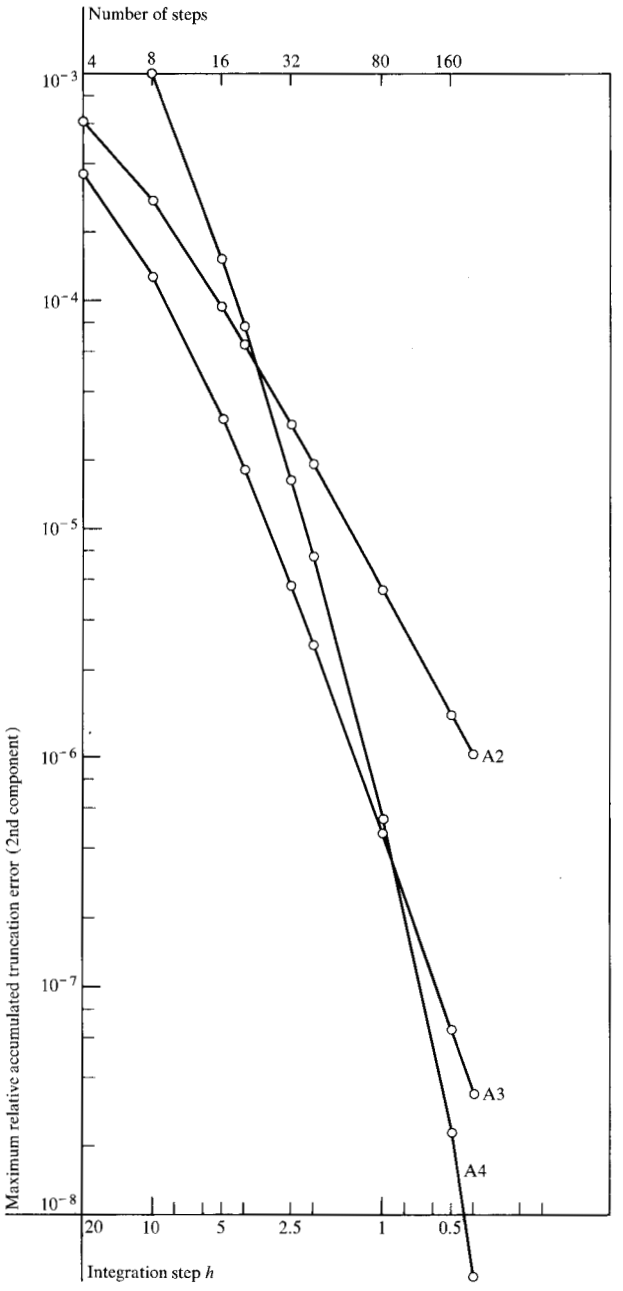

Figure 5 Accumulated truncation error for test problem P2.

the y-component is shown (the results for the other component follow very closely the same curves) as a function of h or of the total number of steps. The lower parts of the curves show straight-line behavior with slopes approximately as predicted by the theory (-2, -3, or -4, corresponding to the orders 2, 3, and 4, respectively). The "kinks" in the upper parts of the curves for A2 and A4 are due to a sign reversal in the error. The (large) q-value, $q_1 = \lambda_1 h$, is 2 to 3 orders of magnitude larger – and thus so is h – than what one could use, for example, with standard Runge-Kutta methods.

P2:
$$\dot{x} = 0.01 - [1 + (x + 1000)(x + 1)](0.01 + x + y)$$
$$\dot{y} = 0.01 - (1 + y^2)(0.01 + x + y)\,. \qquad (78)$$

This is problem no. 2 in the paper by Liniger [21] and stems from an application in chemical kinetics. Again there is a boundary layer effect near $t = 0$, and near $t = 100$ the eigenvalues begin to vary rapidly. Therefore, an interval of integration was chosen in the "smooth part," from $t = 1$ to $t = 81$. A numerically exact solution was computed by the standard fourth-order Runge-Kutta method with an extremely small step, $h = 0.002$, using 40,000 steps on this interval. Appropriately spaced values of this solution near $t = 1$ were used to start the integration by A2-A4. The eigenvalues range from $\lambda_1 \approx -981$, $\lambda_2 \approx -2 \times 10^{-5}$ at $t = 1$ to $\lambda_1 \approx -185$, $\lambda_2 \approx -10^{-3}$ at $t = 81$. Thus the stiffness ratio ranges from $\approx 5 \times 10^7$ at $t = 1$ to $\approx 2 \times 10^5$ at $t = 81$. Note that $h = 0.002$ for Runge-Kutta is just barely stable with respect to the eigenvalue $\lambda_1 \approx 1000$, since the stability limit is $q = \lambda h \approx 2.78$.

In Fig. 5, the errors, defined as in P1, are plotted and show the same general qualitative behavior. The error in the y-component only of (78) is shown, since the error in the x-component is very nearly the same.

Some remarks about the choice of parameters can be made. Thus far, such a choice is restricted only by a) the requirement of the dimensionality of the space spanned by the parameter points to insure the unique solvability of (14) for the weights ν_ρ, and b) the A-stability requirements. This leaves one with a lot of freedom, and other criteria could be used in selecting these parameters; e.g., maximizing the possible damping of solutions at $q = \infty$ or minimizing the local truncation error in some sense. In the above problems, when using A4, the parameter points were $(r_1, s_1) = (7,2)$; $(r_2, s_2) = (5,2)$ and $(r_3, s_3) = (7,1)$. The secondary parameter was chosen as $c = 4$, which is a compromise giving a reasonably large A-stability domain without too adversely affecting the local truncation error constant.

We conclude by commenting on the "amount of work" involved in one "pass" of algorithm A4 and comparing it to a similar algorithm. Two function evaluations ($\tilde{f}^+$ in step 1 and f^+ in step 8), one Jacobian evaluation ($\tilde{f}^+_x$ in step 1) and one L\U matrix factorization (that of $(I - hc\tilde{f}^+_x)$ in step 5) are needed [22]. The algorithm A4 is rather similar to implicit Runge-Kutta methods of Rosenbrock type [23]. Specifically, if we compare it to case no. 9 of the three-level, fourth-order methods of Allen-Pottle [9], we find this latter method requires three function evaluations $[f(x), f(x + b_1k), f(x - b_2k + d_1l)]$, one Jacobian evaluation, and two L\U matrix factorizations [those of $(I - ha_1f_x)$ and $(I - ha_3f_x)$]. The relatively low number of operations per pass in A4 is due to

1. The use of one single Newton-Raphson step in computing the basic solution [24].
2. The introduction of an auxiliary parameter c, which makes the Adams formula slightly longer than would be necessary for accuracy purposes, but can be used to improve the stability properties and to fix the coefficient matrix in step 5 when calculating the other two solutions, thereby saving two L\U factorizations.
3. The application of a perturbation method, which enables one to use the function and Jacobian evaluations associated with the first solution in computing solutions 2 and 3.

Aside from the major items of computation mentioned thus far there is, in A4, more data handling of other types than in the Allen method, but we expect that this should not have a great influence on the relative performance of the two methods.

References and notes

1. G. Bjurel et al., "Survey of stiff ordinary differential equations," Report NA70.11, Royal Institute of Technology, Stockholm, Sweden, 1970.
2. G. G. Dahlquist, "A special stability criterion for linear multistep methods," *BIT* **3,** 22–43 (1963).
3. P. Henrici, *Error Propagation for Difference Methods*, J. Wiley & Sons, New York, 1963, p. 23.
4. O. Widlund, "A note on unconditionally stable linear multistep methods," *BIT* **7,** 65–70 (1967).
5. C. W. Gear, "The automatic integration of stiff ordinary differential equations," *Information Processing* (North Holland Publishing Co., Amsterdam) **68,** 187–193 (1969).
6. W. Liniger and R. A. Willoughby, "Efficient integration methods for stiff systems of ordinary differential equations," *SIAM J. Num. Anal.* **7,** 47–66 (1970).
7. B. Lindberg, "On smoothing and extrapolation for the Trapezoidal Rule," *BIT* **11,** 29–52 (1971).
8. W. Liniger, "A criterion for A-stability of linear multistep integration formulae," *Computing* **3,** 280–285 (1968).
9. R. Allen and C. Pottle, "Stable integration methods for electronic circuit analysis with widely separated time constants," *Proc. Sixth Annual Allerton Conference on Circuit and Systems Theory*, Univ. of Illinois, Oct. 1967, pp. 534–543.
10. M. Abramowitz and I. A. Stegun (Ed.), *Handbook of Mathematical Functions*, National Bureau of Standards, *Appl. Math. Ser.* **55,** 883 (1964).
11. P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, J. Wiley & Sons, New York, 1962.
12. J. B. Keller, "*Perturbation Theory*" Lectures presented at the Department of Mathematics, Michigan State University, East Lansing, Michigan, 1968.
13. Henrici, Ref. 3, p. 18.
14. Liniger, Ref. 8, p. 283.
15. F. R. Gantmacher, "*The Theory of Matrices*," Chelsea Publishing Co., New York, 1959, vol. 2, pp. 177–180.
16. Obviously, with large time steps, such a procedure is meaningful only during the smooth asymptotic phase of the solution. In the initial boundary layer, it is natural to use small steps in order to sample the solution properly.
17. M. N. Spijker, *Stability and Convergence of Finite Difference Methods*, Doctoral thesis, Department of Mathematics, University of Leiden, Leiden, 1968.
18. W. Liniger, "A stopping criterion for the Newton-Raphson method in implicit multistep integration algorithms for nonlinear systems of ordinary differential equations," *Comm. ACM* **14,** 600–601 (1971).
19. The quantity z represents an output and is not used in the calculation.

20. M. E. Fowler and R. M. Warten, "A numerical integration technique for ordinary differential equations with widely separated eigenvalues," *IBM J. Res. Dev.* **11,** 537–543 (1967).
21. W. Liniger and R. A. Willoughby, "Efficient numerical integration of stiff systems of ordinary differential equations," report RC-1970, IBM Thomas J. Watson Research Center, Yorktown Heights, N.Y., 1967.
22. F. G. Gustavson et al., "Symbolic generation of an optimal Crout algorithm for sparse systems of linear equations," *J. Assoc. Comput. Mach.* **17,** 87–109 (1970).
23. H. H. Rosenbrock, "Some general implicit processes for numerical solution of differential equations," *Comput. J.* **5,** 329–330 (1963).
24. For strongly nonlinear situations it may be preferable to carry out additional Newton steps. In the test problems described here the extra iterations did not improve the accuracy appreciably.

Received December 17, 1971

The authors are located at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York, 10598.