Boundary Value Techniques for Initial Value Problems in Ordinary Differential Equations
Author(s): A. O. H. Axelsson and J. G. Verwer
Source: *Mathematics of Computation*, Vol. 45, No. 171 (Jul., 1985), pp. 153-171
Published by: American Mathematical Society
Stable URL: https://www.jstor.org/stable/2008054
Accessed: 04-04-2019 14:53 UTC

REFERENCES
Linked references are available on JSTOR for this article:
https://www.jstor.org/stable/2008054?seq=1&cid=pdf-reference#references_tab_contents
You may need to log in to JSTOR to access the linked references.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at https://about.jstor.org/terms

American Mathematical Society is collaborating with JSTOR to digitize, preserve and extend access to *Mathematics of Computation*

This content downloaded from 131.211.22.192 on Thu, 04 Apr 2019 14:53:59 UTC
All use subject to https://about.jstor.org/terms

MATHEMATICS OF COMPUTATION
VOLUME 45, NUMBER 171
JULY 1985, PAGES 153–171

Boundary Value Techniques for Initial Value Problems in Ordinary Differential Equations

By A. O. H. Axelsson and J. G. Verwer

Abstract. The numerical solution of initial value problems in ordinary differential equations by means of boundary value techniques is considered. We discuss a finite-difference method which was already investigated by Fox in 1954 and Fox and Mitchell in 1957. Hereby we concentrate on explaining the fundamentals of the method because for initial value problems the boundary value method seems to be fairly unknown. We further propose and discuss new Galerkin methods for initial value problems along the lines of the boundary value approach.

1. Introduction. Traditionally, methods used for the numerical integration of initial value problems in ordinary differential equations

$$\dot{y}(x) = f(x, y(x)), \qquad a \leqslant x \leqslant b, y(a) \text{ given}, \tag{1.1}$$

are *step-by-step methods*. Familiar step-by-step methods, which are also called *forward-step methods*, are the Runge-Kutta and linear multistep method (see, e.g., Henrici [12], Lambert [16], Stetter [23]). The latter, in its most simple form, is defined by the so-called k-step formula

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f(x_{n+j}, y_{n+j}), \qquad \alpha_j, \beta_j, h \in \mathbf{R}, h > 0, k \in \mathbf{N}^+, \tag{1.2}$$

where y_{n+j} represents the approximation to the exact solution value $y(x_{n+j})$ defined by (1.1). The positive real h is called the step size. Assuming that h is constant, it is given by $h = (b - a)/N$, N being some positive integer. The points x_{n+j} are called grid points and belong to the uniform grid

$$G_h = \{ x_j : x_j = a + jh, j = 0(1)N \}. \tag{1.3}$$

In the forward-step approach, the numerical solution is obtained by stepping through this grid in the direction from a to b, i.e., given approximations y_{n+j} for some integer n and $j = 0(1)k - 1$, the approximation y_{n+k} at the next grid point x_{n+k} is computed by solving (1.2) for y_{n+k}. In fact, all results on convergence and numerical stability which emanate from the pioneering work of Dahlquist [5] are based on this forward-step application.

Received February 1, 1983; revised September 25, 1984.
1980 *Mathematics Subject Classification*. Primary 65L05, 65L10.
Key words and phrases. Numerical analysis, initial value problems for ordinary differential equations, stiffness, boundary value techniques.

©1985 American Mathematical Society
0025-5718/85 $1.00 + $.25 per page

This content downloaded from 131.211.22.192 on Thu, 04 Apr 2019 14:53:59 UTC
All use subject to https://about.jstor.org/terms

In this paper we will tackle the numerical solution of (1.1) in a completely different way than in the step-by-step approach. For its numerical solution we will consider (1.1) as a two-point boundary value problem with a given value at the left endpoint and an implicitly defined value, by the equations $\dot{y}(x) = f(x, y(x))$, at the right endpoint. In this approach formula (1.2) ought to be considered as a *finite-difference formula* as is the practice in the numerical solution of genuine two-point boundary value problems for systems of first-order differential equations (see Keller [14], [15]). One of the aims of this *boundary value approach* is to circumvent the known *Dahlquist-barriers* on convergence and stability which are a direct consequence of the step-by-step application of (1.2). In this respect boundary value methods for (1.1) bear a relationship with the iterative algorithms of Cash [4] for the stable solution of recurrence relations and with Olver's algorithm [18], [19].

Up to now, boundary value methods for initial value problems have hardly been discussed in the numerical literature. Perhaps because the step-by-step application of formulas of type (1.2) is invariably easier to perform. As far as we know, the first contributions have been made by Fox [9] in 1954 and Fox and Mitchell [10] in 1957. They discuss a simple finite-difference formula for (1.1) and for the derived second-order equation

$$\ddot{y}(x) = g(x, y(x)) = \frac{\partial f}{\partial x}(x, y(x)) + \frac{\partial f}{\partial y}(x, y(x))f(x, y(x)). \tag{1.4}$$

A feature of the boundary value method is that all approximations on the grid G_h are generated simultaneously. In 1964 Axelsson [1] proposed a quadrature type method for the integrated form of (1.1) which also computes all approximations over the interval $[a, b]$ simultaneously. This method has been called a global integration method. It is best characterized as a huge implicit Runge-Kutta method which performs just one step with step size $b - a$. A special feature of this global method is that the global errors at the end of the interval are particularly small, even when the problem is mathematically unstable. On the other hand, the errors of step-by-step methods have a tendency to grow, owing to accumulation at every step, especially when the problem itself is unstable.

Two recent contributions on boundary value methods for initial value problems are due to Rolfes [20] and Rolfes and Snyman [21]. They consider a finite-difference method which has also been proposed by Fox [9] and apply it to stiff equations. Rolfes and Snyman report that the finite-difference method performs satisfactorily on stiff problems. Fox considered nonstiff equations, but was not satisfied with the method because of an oscillating error behavior which prevents the application of difference correction for improving the accuracy.

The present contribution consists of two parts. The first part deals with *finite-difference methods*, while the second one is devoted to *Galerkin methods*. When discussing boundary value techniques for initial value problems it is, of course, obvious to consider Galerkin methods because of their use in the numerical solution of genuine two-point boundary value problems. We shall comment on a relation between the two approaches.

To a certain extent this paper is of an expository nature, especially in its first part on finite-difference methods (Section 2). There, we have concentrated on describing

This content downloaded from 131.211.22.192 on Thu, 04 Apr 2019 14:53:59 UTC
All use subject to https://about.jstor.org/terms

the fundamentals of the boundary value approach, because for initial value problems this approach seems to be fairly unknown. For that purpose Section 2 reports on a case study of a straightforward combination of the explicit midpoint rule with the first-order backward difference formula. Among others, this case study clearly reveals that with respect to stability, an essential difference exists between the standard forward-step and the boundary value approach. We emphasize that the phenomena involved are typical for the boundary value approach, rather than accidental for our case study.

Finally, we should like to mention three serious applications of the boundary value method in situations where the forward-step method may be less appropriate (not further treated in the present paper). Firstly, the numerical solution of initial value problems where the right-hand side function $f(x, y)$ is not available in analytic form but merely in the form of discrete data.* Such a situation frequently arises in simulation processes. These problems might be tackled by fitting the data so as to generate functions which can be evaluated anywhere such that Runge-Kutta methods or multistep methods can be applied. This approach involves the difficulty of avoiding too large errors in the generated functions. An alternative is to employ a method which uses only the discrete data available. Shampine [22] examines such a method. The boundary value methods of this paper can also be applied to the problems discussed by Shampine [22].

The second application we have in mind lies in the control of the global error. When integrating in a forward-step manner, direct global error control cannot be theoretically justified since the behavior of the global error in time depends on the stability of the problem and on all previous global errors. The only justifiable procedure here is simply reintegration over the whole integration interval with a smaller step size, in case the estimation of the global error has turned out to be too crude. By its very nature, the boundary value method is better adapted for global error control, because now the numerical solutions are computed simultaneously as if we were solving a boundary value problem. This implies that, for global error control purposes, one could implement sophisticated adaptive mesh techniques from currently available boundary value codes.

Thirdly, the boundary value methods can also be used as step-by-step methods but with much larger steps than for an ordinary step-by-step method. A possible application of this is for ill-posed problems of the form (1.1). If the solution to such a problem is smooth, one may approximate it well by a boundary value technique using large time steps, and the inherent instability will not be noticed as much as for an ordinary step-by-step method. This situation is similar to the effect of using parallel shooting instead of just simple shooting in boundary value problems.

Naturally we can also envision problems where a boundary value technique would be less appropriate. This occurs, for instance, in certain nonlinear problems where the solution suddenly becomes very unsmooth. In a step-by-step method one can more easily adapt the step-lengths in order to better approximate the steep gradients when they occur.

---

*This application has been brought to our attention by Larry F. Shampine.

This content downloaded from 131.211.22.192 on Thu, 04 Apr 2019 14:53:59 UTC
All use subject to https://about.jstor.org/terms

2. A Finite-Difference Boundary Value Method.

2.1. *Outline of the Method.* Consider the initial value problem (1.1). Let us discretize the differential equation $\dot{y} = f(x, y)$ on the grid (1.3) by means of the *explicit midpoint rule*

$$y_{n+1} - y_{n-1} - 2hf(x_n, y_n) = 0. \tag{2.1}$$

When we apply (2.1) as a step-by-step method we need two initial values, one at the left endpoint $x = a$, and one at $x = a + h$. The first initial value is known from the problem, while the second one has to be computed by another method. When we apply (2.1) as a boundary value method it is applied at each of the points $x_n \in G_h$ for $n = 1(1)N - 1$. In addition to the initial value at the left endpoint $x = a$, we now need a boundary condition at the right endpoint $x = b$. For that purpose, one can use the most simple *backward-difference formula* (Backward Euler)

$$y_N - y_{N-1} - hf(x_N, y_N) = 0. \tag{2.2}$$

Thus we arrive at the discrete boundary value problem

$$\begin{aligned} &y_0 \text{ given,} \\ &y_{n+1} - y_{n-1} - 2hf(x_n, y_n) = 0, \qquad n = 1(1)N - 1, \\ &y_N - y_{N-1} - hf(x_N, y_N) = 0, \end{aligned} \tag{2.3}$$

whose solution values $y_1, \ldots, y_N$ must be generated simultaneously. Since f may be nonlinear in y, the discrete problem (2.3) must be solved by iteration. A Newton-type iteration is feasible because of the *tridiagonal structure* (block-tridiagonal for systems).

As an alternative for formula (2.2), we mention the more accurate *trapezoidal rule*

$$y_N - y_{N-1} - \tfrac{1}{2}hf(x_{N-1}, y_{N-1}) - \tfrac{1}{2}hf(x_N, y_N) = 0, \tag{2.4}$$

or the *second-order backward-difference formula*

$$y_N - \tfrac{4}{3}y_{N-1} + \tfrac{1}{3}y_{N-2} - \tfrac{2}{3}hf(x_N, y_N) = 0. \tag{2.5}$$

The use of (2.4) or (2.5) instead of (2.2) does not increase the order of accuracy of the method. Both combinations are of order two. Normally, method (2.3) will be somewhat less accurate. Convergence questions are further discussed in Section 2.3.

Combination (2.1), (2.5) has already been proposed by Fox [9] and Fox and Mitchell [10]. Rolfes [20] and Rolfes and Snyman [21] have applied this combination to stiff problems. A slight disadvantage is that by using (2.5), the tridiagonal coupling is lost. This might be overcome, however, by eliminating y_{N-2} from (2.5) and the particular equation

$$y_N - y_{N-2} - 2hf(x_{N-1}, y_{N-1}) = 0. \tag{2.6}$$

This yields

$$\begin{aligned} &y_0 \text{ given,} \\ &y_{n+1} - y_{n-1} - 2hf(x_n, y_n) = 0, \qquad n = 1(1)N - 1, \\ &\tfrac{4}{3}(y_N - y_{N-1}) - \tfrac{2}{3}hf(x_{N-1}, y_{N-1}) - \tfrac{2}{3}hf(x_N, y_N) = 0, \end{aligned} \tag{2.7}$$

which is just method (2.1), (2.4).

Finally we observe that methods like (2.3) can be directly applied to problems with periodic solutions. The last line of (2.3) then should read $y_N = y_0$. In what follows we concentrate on the pure initial value problem.

This content downloaded from 131.211.22.192 on Thu, 04 Apr 2019 14:53:59 UTC
All use subject to https://about.jstor.org/terms

2.2. *The Test Model.* In this section we consider the standard test model

$$\dot{y} = \delta y, \qquad \delta \in \mathbf{C}, a \leqslant x \leqslant b, y(a) \text{ given.} \tag{2.8}$$

We observe that this model plays an important role in the stability of step-by-step integration methods. The notion of absolute stability (see, e.g., [16]) is based on this simple problem which is also very suitable for becoming acquainted with the boundary value approach and for comparison with the step-by-step approach. In Section 2.3 the model is linked with a constant-coefficient linear system. We will concentrate on method (2.3), i.e., explicit midpoint combined with Backward Euler.

Our discrete boundary value problem (2.3) now reads

$$\begin{aligned} &y_0 = y(a), \\ &y_{n+1} - y_{n-1} - 2zy_n = 0, \qquad z = h\delta, n = 1,\ldots,N-1, \\ &y_N - y_{N-1} - zy_N = 0, \end{aligned} \tag{2.9}$$

i.e., we have to solve the linear algebraic system

$$A(z)Y = R, \tag{2.10}$$

where $Y = [y_1,\ldots,y_N]^T$, $R = [y(a),0,\ldots,0]^T$ and $A(z)$ is given by

$$A(z) = \begin{pmatrix} -2z & 1 & & & \\ -1 & -2z & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & -2z & 1 \\ & & & -1 & 1-z \end{pmatrix}. \tag{2.11}$$

The first question which arises is, for which z-values is Y a well-defined vector of approximations y_n to $e^{nz}y(a)$, $n = 1,\ldots,N$, i.e., for which z-values is $A(z)$ regular. In what follows, we call z *a regular point for* $A(z)$ if $A(z)$ is regular. Otherwise, z is called a *singular point*.

Define $\tilde{A}(z) = \operatorname{diag}(1,\ldots,1,2)A(z)$, and write $\tilde{A}(z) = E - 2zI$, i.e.

$$E = \begin{pmatrix} 0 & 1 & & & \\ -1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 0 & 1 \\ & & & -2 & 2 \end{pmatrix}. \tag{2.12}$$

$A(z)$ is singular, iff $\tilde{A}(z)$ is singular. Hence we can use $\tilde{A}(z)$, and in turn the constant matrix E to find the singular points for $A(z)$. Obviously, the location of the eigenvalues λ_j of E is decisive, since z is a singular point, iff $z = \lambda_j/2$.

LEMMA 1. *All eigenvalues* λ_j *of* E *satisfy* $0 < \operatorname{Re}(\lambda_j) \leqslant 2$, $-2 \leqslant \operatorname{Im}(\lambda_j) \leqslant 2$.

Proof. The inequality $-2 \leqslant \operatorname{Im}(\lambda_j) \leqslant 2$ is a direct consequence of Geršgorin's circle theorem. To prove the inequality for the real part we first perform the similarity transformation

$$\tilde{E} = \operatorname{diag}(1,\ldots,1,d)E\operatorname{diag}(1,\ldots,1,d^{-1})$$

which leaves the spectrum invariant. Let λ and μ be the real and imaginary parts of an arbitrary eigenvalue and let u and v be the real and imaginary parts of the corresponding eigenvector. Then we easily derive

$$\tfrac{1}{2}\left[u^T(\tilde{E} + \tilde{E}^T)u + v^T(\tilde{E} + \tilde{E}^T)v\right] = \lambda(u^Tu + v^Tv). \tag{2.13}$$

This content downloaded from 131.211.22.192 on Thu, 04 Apr 2019 14:53:59 UTC
All use subject to https://about.jstor.org/terms

Now we take $d = 1/\sqrt{2}$ for which $\frac{1}{2}(\tilde{E} + \tilde{E}^T) = \text{diag}(0,\ldots,0,2)$. Hence,

$$0 \leqslant \tfrac{1}{2}u^T(\tilde{E} + \tilde{E}^T)u \leqslant 2u^Tu, \quad \text{all } u \in \mathbf{R}^N,$$

so that $0 \leqslant \lambda \leqslant 2$. Finally, assume $\lambda = 0$ and let u_i, v_i denote the ith component of u and v, respectively. From (2.13) it then follows that $u_N = v_N = 0$. By using the relations $\tilde{E}u = -\mu v$, $\tilde{E}v = \mu u$ and the specific form of $\tilde{E}$ it is now easy to verify that $u_i = v_i = 0$, all $i = 1(1)N$. This leads to a contradiction, showing that $\lambda \neq 0$. □

We thus have the following result:

THEOREM 2. *All singular points z for $A(z)$ satisfy* $0 < \text{Re}(z) \leqslant 1$, $-1 \leqslant \text{Im}(z) \leqslant 1$. □

We cannot determine the eigenvalues of E explicitly. Note that if in E the last row elements are replaced by -1 and 0, respectively, the eigenvalues become $2i\cos(j\pi/(N+1))$, $j = 1(1)N$. Figure 1 shows all numerically computed eigenvalues of E and $E/2h$ for some values of $h = N^{-1}$. The eigenvalues of $E/2h$ play an important role in the convergence analysis (cf. Section 2.3). We see that when N increases, a pair of eigenvalues of E approaches $\pm 2i$. This means that for N large, the points $\pm i$ will act numerically as singular points for $A(z)$.

The second question we now wish to discuss is, how well are the decaying exponentials e^{nz} approximated. From diagonal dominance properties it easily follows that for $\text{Re}(z) \ll 0$ (stiff eigenvalues) $|y_n|$ is an excellent approximation to $|e^{nz}y(a)|$. More precisely, if $z \neq 0$ is a regular point, then (2.10) can be rewritten as $Y = -(2z)^{-1}(I - (2z)^{-1}E)^{-1}R$, which implies

$$y_1 = -(2z)^{-1}y(a) + O\big(|z|^{-2}\big), \quad y_n = O\big(|z|^{-2}\big), \qquad n = 2(1)N,\ |z| \to \infty.$$

Observe that the method cannot approximate positive exponentials if $\text{Re}(z) \gg 0$. Roughly speaking, for $|\text{Re}(z)|$ large, the approximations for the negative and positive exponential e^{nz} are of the same magnitude.

To get more insight into the question of how decaying exponentials are approximated, we now proceed with the analytical solution of the recurrence equation $y_{n+1} - y_{n-1} - 2zy_n = 0$ defined by the explicit midpoint rule when applied to test-model (2.8):

$$y_n = C_1\mu_1^n + C_2\mu_2^n, \qquad n = 1,2,\ldots,N, \tag{2.14}$$

where $\mu_1 = z + \sqrt{z^2+1}$, $\mu_2 = z - \sqrt{z^2+1}$ and C_1, C_2 are constants to be determined by boundary conditions. Note that $\mu_1 = e^z + O(z^3)$, $z \to 0$, whereas μ_2 has no relation to e^z, i.e. μ_2 is the parasitic root.

Solution (2.14) can be adapted to our discrete problem (2.9) via C_1 and C_2 by requiring

$$\begin{aligned} C_1 + C_2 &= y(a), \\ (1-z)\big(C_1\mu_1^N + C_2\mu_2^N\big) &= C_1\mu_1^{N-1} + C_2\mu_2^{N-1}. \end{aligned} \tag{2.15}$$

Solving for C_1 and C_2 yields $C_2 = \delta C_1$, $C_1 = y(a)/(1+\delta)$, where

$$\delta = \left(\frac{-1}{\mu_2^2}\right)^{N-1}\eta, \qquad \eta = \frac{1 - \mu_1(1-z)}{(1-z)\mu_2 - 1}, \tag{2.16}$$

This content downloaded from 131.211.22.192 on Thu, 04 Apr 2019 14:53:59 UTC
All use subject to https://about.jstor.org/terms

and where it is assumed that $\mathrm{Re}(\delta) \neq -1$. $\mathrm{Re}(\delta) = -1$ means singularity of the 2×2 system (2.15). Like for system (2.10), one thus must distinguish singular and regular points z. We emphasize that the set of singular points for (2.15) is not identical with that of $A(z)$. For example, $z = \pm i$ is a singular point for (2.15) for all N, but not for $A(z)$ according to Theorem 2. Nevertheless, as observed before, for numerical computations, the points $z = \pm i$ must be regarded also as singular points for $A(z)$. Of course, if z is a singular point for $A(z)$ and not for (2.15), (2.14) defines a particular solution for system (2.10).

Let us consider the behavior of the principal solution component $C_1\mu_1^n$ and the parasitic component $C_2\mu_2^n$ for varying n and z, where we restrict ourselves to $z \leqslant 0$ and N even. We observe that for N even, $z \in \mathbf{R}$, the quantity $\delta \geqslant 0$, since $\eta(0) = 0$ and $\eta(z) < 0$ if $z \neq 0$. Hence, for $z \leqslant 0$ and N even, the solution (2.14) is well-defined and is just the unique solution of system (2.10).

We distinguish between $z = 0$ and $z < 0$. The case $z = 0$ corresponds to $\dot{y}(x) = 0$, i.e. $y(x) = y(a)$, $a \leqslant x \leqslant b$. It is readily seen that for $z = 0$, $y_n = y(a)$ for all $n = 1, \ldots, N$. Hence the constant solution is computed without error. For $z < 0$, i.e., decaying exponentials, we have $0 < \mu_1 < 1$, $\mu_2 < -1$ and the limit behavior

$$\mu_1 \sim 1 + z, \quad \mu_2 \sim -1 + z, \quad \eta \sim -\tfrac{1}{4}z^2 \quad \text{as } z \uparrow 0,$$
$$\mu_1 \to 0, \quad \mu_2 \to -\infty, \quad \eta \sim -\tfrac{1}{4}z^{-2} \quad \text{as } z \to -\infty.$$

Taking this into consideration, the behavior of $C_1\mu_1^n$ and $C_2\mu_2^n$ is best described as follows. $C_1\mu_1^n$ approximates the decaying solution for z close to zero and vanishes if $z \to -\infty$. This is true for all $1 \leqslant n \leqslant N$. For z close to zero, the parasitic component $C_2\mu_2^n$ is negligibly small (up to the discretization order in z). For h fixed, $C_2\mu_2^n$ increases with n. However, for all $z < 0$, its contribution to y_n is negligible for all n, $1 \leqslant n \leqslant N$. We once more note that for $\mathrm{Re}(z) \ll 0$ (stiff eigenvalues) the strongly decaying exponential e^{nz} is well approximated. A similar description can be given for $z > 0$.

At this point it is appropriate to make a comparison with the standard step-by-step approach. Suppose that the explicit midpoint rule is applied that way. Consider the general solution (2.14). In order to obtain *absolute stability* μ_1 and μ_2 now must satisfy *the root condition*, i.e., none of the characteristic roots has modulus greater than one and every root with modulus one is simple. The root condition is satisfied if and only if z is purely imaginary and $|z| < 1$. Hence, as is well-known, the step-by-step explicit midpoint rule has no real interval of absolute stability, which shows that with respect to stability the boundary value method is just opposite to the step-by-step method. In fact, from the investigation of equations (2.14)–(2.16), it can be seen that the boundary value method can be applied for $\mathrm{Re}(z) < 0$, just because there $|\mu_1| < 1$ and $|\mu_2| > 1$. This conclusion, which is valid for other difference schemes as well, has been drawn before by Rolfes [20]. She considers the tridiagonal infinite Toeplitz matrix with rows (–1 0 1) and shows that the forward-backward substitution of the LU-decomposed Toeplitz matrix can be interpreted as a stable forward recursion ($|\mu_1| < 1$) followed by a stable backward recursion ($|\mu_2| > 1$) (see also [18], [19]).

This content downloaded from 131.211.22.192 on Thu, 04 Apr 2019 14:53:59 UTC
All use subject to https://about.jstor.org/terms

2.3. *Convergence Properties.* This section is devoted to convergence properties of the finite-difference boundary value method. As in the preceding section we concentrate on method (2.3). It will be assumed that the vector function f: $[a, b] \times \mathbf{R}^s \to \mathbf{R}^s$ is as smooth as our analysis requires.

We introduce the conventional operators $\mathcal{N}$ and $\mathcal{N}_h$ (see, e.g., [14], [15]):

$$\mathcal{N}y \equiv \dot{y}(x) - f(x, y(x)) = 0, \qquad a \leqslant x \leqslant b, y(a) \text{ given},$$
$$\mathcal{N}_h y_n \equiv \frac{y_{n+1} - y_{n-1}}{2h} - f(x_n, y_n) = 0, \qquad n = 1, \ldots, N-1, y_0 = y(a),$$
$$\mathcal{N}_h y_N \equiv \frac{y_N - y_{N-1}}{h} - f(x_N, y_N) = 0.$$

Next, for any sufficiently smooth function $v(x)$, we define the *local truncation errors* $\tau_n[v] \equiv \mathcal{N}_h v(x_n) - \mathcal{N}v(x_n)$, $n = 1(1)N$, and observe that

$$\tau_n[v] = \tfrac{1}{6}h^2 \dddot{v}(x_n) + O(h^3), \qquad n = 1(1)N-1,$$
$$\tau_N[v] = -\tfrac{1}{2}h\ddot{v}(x_N) + O(h^2).$$

Let e_n be the *global error vector* at x_n, i.e., $e_n \equiv y_n - y(x_n)$, $n = 1(1)N$. By subtracting $\mathcal{N}_h y(x_n)$ from $\mathcal{N}_h y_n$ and by using the mean value equation

$$f(x_n, y(x_n) + e_n) - f(x_n, y(x_n)) = M(x_n)e_n,$$
$$M(x_n) \equiv \int_0^1 f'(x_n, y(x_n) + \theta e_n)\, d\theta, \qquad f'(x, u) \equiv \frac{\partial f}{\partial u}(x, u),$$

it can be seen that e_n satisfies the difference scheme

$$\begin{aligned} &\mathcal{L}_h e_n \equiv \frac{e_{n+1} - e_{n-1}}{2h} - M(x_n)e_n = -\tau_n[y], \qquad n = 1(1)N-1, \\ &\mathcal{L}_h e_N \equiv \frac{e_N - e_{N-1}}{h} - M(x_N)e_N = -\tau_N[y], \end{aligned} \tag{2.17}$$

where e_0 is the zero vector and $y = y(x)$ denotes the exact solution of the initial value problem (1.1). Hence method (2.3) is convergent, for a given vector function f, if for this function $\mathcal{L}_h$ is a *stable* difference operator (cf., [14], [15]).

Let us reformulate (2.17) in the block matrix form

$$\begin{pmatrix} -2hM(x_1) & I & & & & \\ -I & -2hM(x_2) & I & & & \\ & \ddots & \ddots & \ddots & & \\ & & -I & -2hM(x_{N-1}) & I & \\ & & & -I & I - hM(x_N) \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_{N-1} \\ e_N \end{pmatrix} = \begin{pmatrix} -2h\tau_1[y] \\ -2h\tau_2[y] \\ \vdots \\ -2h\tau_{N-1}[y] \\ -h\tau_N[y] \end{pmatrix} \tag{2.18}$$

which we denote by

$$\mathcal{A}_h \vec{e} \equiv (E_1 \otimes I - 2h\mathcal{M})\vec{e} = -2h\vec{\tau}, \tag{2.19}$$

This content downloaded from 131.211.22.192 on Thu, 04 Apr 2019 14:53:59 UTC
All use subject to https://about.jstor.org/terms

where E_1 is given by (2.12) with the last row divided by two, and where $\otimes$ denotes the direct matrix product. The definitions of $\mathscr{M}$, $\vec{e}$, and $\vec{\tau}$ are obvious. Stability of $\mathscr{L}_h$ is equivalent to the existence and *uniform boundedness* of the inverses of the family of matrices $h^{-1}\mathscr{A}_h$.

Example 3. To gain some feeling for how the local errors τ_n accumulate in the global error we now first consider the scalar equation $\dot{y}(x) = f(x)$, i.e., f does not depend on y. Then $\vec{e}$ satisfies

$$(2.20) \qquad (2h)^{-1}E_1\vec{e} = -\tau.$$

From the computation of E_1^{-1} one finds the global errors

$$(2.21) \qquad \begin{aligned} e_n &= -\sum_{j=1}^{n/2} 2h\tau_{2j-1}, \qquad n \text{ even}, \\ e_n &= e_{n-1} - \sum_{j=n}^{N} 2h(-1)^{j-1}\tau_j + h(-1)^{N-1}\tau_N, \qquad n \text{ odd}, e_0 = 0. \end{aligned}$$

It follows that for all n, $e_n = O(h^2)$. Note that $\tau_N = O(h)$ occurs only once in each e_n, n odd, and not in e_n if n is even. We also see a distinction between even and odd numbered errors, implying that e_n is not smooth when considering all grid points. □

In Example 3 we considered an over-simplified problem. It nicely illustrates, however, the role of the matrix E, or E_1, in the convergence process, which, as we will show below, plays a similar role for the general problem.

Let us proceed with Eq. (2.19). Since E_1 is nonsingular, we can write

$$(2.22) \qquad \left(I - 2h\left(E_1^{-1} \otimes I\right)\mathscr{M}\right)\vec{e} = \vec{\gamma} \equiv -2h\left(E_1^{-1} \otimes I\right)\vec{\tau}.$$

Note that we use I to denote the $s \times s$ unit matrix, as well as the $sN \times sN$ unit matrix. The sN-vectors $\vec{\tau}$ and $\vec{\gamma}$ consist of N blocks, each of length s. Let $\vec{\tau}_j$ and $\vec{\gamma}_j$ denote the N-vector composed of the jth element from each block. These vectors are associated with the jth component of the solution vector $y(x)$. Then, for $j = 1(1)s$, we have $(2h)^{-1}E_1\vec{\gamma}_j = -\vec{\tau}_j$ as in Eq. (2.20), implying that each nth element of $\vec{\gamma}_j$ satisfies relation (2.21). This in turn implies that each element of the whole vector $\vec{\gamma}$ is $O(h^2)$, or, equivalently,

$$(2.23) \qquad \|\vec{\gamma}\|_\infty \leqslant Ch^2, \quad C \text{ a constant not depending on } h \leqslant h_0.$$

THEOREM 4. *Let* $\|\mathscr{M}(x)\|_\infty \leqslant \frac{1}{2}$ *for all* $x \in [a, b]$. *Then method* (2.3) *is convergent in the maximum norm with order two.*

Proof. Consider Eq. (2.22) and observe that $hE_1^{-1} \otimes I$ is uniformly bounded. In fact, from the equation for e_1 in (2.21) it follows that $\|hE_1^{-1} \otimes I\|_\infty = 1$. The proof is now easily completed by applying the perturbation lemma to the left-hand side matrix of Eq. (2.22) and by using inequality (2.23). □

This result covers only a rather narrow class of problems on account of the norm inequality on $M(x)$. For example, stiff problems do not satisfy this inequality. The above derivation indicates, however, through the introduction of $\vec{\gamma}$, that for the general problem $\dot{y} = f(x, y)$, the global errors show a similar behavior as described in Example 3. In fact, we observed this behavior in all our numerical experiments,

This content downloaded from 131.211.22.192 on Thu, 04 Apr 2019 14:53:59 UTC
All use subject to https://about.jstor.org/terms

with nonstiff, as well as stiff problems. In the next theorem we will prove convergence in the spectral norm for a much broader class of problems:

THEOREM 5. *Define*

$$\mu_2(f'(x,u)) \equiv \max_i \lambda_i\left(\frac{f'(x,u)+f'^{\mathrm{T}}(x,u)}{2}\right),$$

where $\lambda_i(\cdot)$ *denotes the ith eigenvalue and assume that* $\mu_2(f'(x,u)) \leqslant \nu < 0$ *for all* $(x,u) \in [a,b] \times \mathbf{R}^s$. *Method* (2.3) *is then convergent in the spectral norm.*

Proof. We consider the matrix $\mathscr{A}_h^* = (-2h)^{-1}\mathscr{A}_h$ (cf. (2.19)). By definition,

$$\mu_2(\mathscr{A}_h^*) = \max_i \lambda_i\left(\operatorname{diag}\left(\frac{M_1+M_1^{\mathrm{T}}}{2},\ldots,\frac{M_{N-1}+M_{N-1}^{\mathrm{T}}}{2},\frac{-2I+h(M_N+M_N^{\mathrm{T}})}{4h}\right)\right),$$

where $M_n = M(x_n)$. For all $h > 0$ we have

$$\mu_2(\mathscr{A}_h^*) \leqslant \max_n \mu_2(M_n) \leqslant \nu.$$

The first inequality is trivial, while the second is a direct consequence of the definition of M_n and of a result given by Dahlquist [5, p. 11]. Since $\nu < 0$ does not depend on h, but only on the problem, and since

$$\max_i \operatorname{Re}\lambda_i(\mathscr{A}_h^*) \leqslant \mu_2(\mathscr{A}_h^*),$$

it is immediate that $\mathscr{A}^{*-1}_h$ exists and is uniformly bounded in $\|\cdot\|_2$. More precisely, $\|\mathscr{A}^{*-1}_h\|_2 \leqslant -\nu^{-1}$, so that

$$\|\vec{e}\|_2 \leqslant -\nu^{-1}\|\vec{\tau}\|_2. \quad \square \tag{2.24}$$

We observe that the method of proof of this theorem cannot be used to deal with Eq. (2.22). This prevents us from proving order two convergence in the spectral norm. In Section 3, however, we are able to prove second-order convergence in the spectral norm by considering method (2.3) as a particular Galerkin method.

The inequality $\mu_2(f'(x,u)) \leqslant \nu < 0$ is satisfied by all differential equations which possess strictly contractive solutions in the Euclidean vector norm (see Dahlquist [5, p. 13] and [6, Chapter 2]). Hence Theorem 5 covers a broad and interesting class of problems, including many stiff ones. Furthermore, for these problems the stiffness, i.e., the magnitude of the stiff eigenvalues of $f'(x,u)$, does not enter into the one-sided Lipschitz constant ν. This constant ν is related to the smooth, nonstiff solution components (see [6, Chapter 2] for a clarifying discussion). Inequality (2.24) thus shows that if the solution to be computed is smooth, the global error will not suffer from the stiffness of the problem. Rolfes and Snyman [20], [21] observed this in their experiments.

If ν is very close to zero, inequality (2.24) is useless. We emphasize, however, that the algorithm then still may perform quite satisfactorily, even if ν is larger than zero. We will explain this from the constant-coefficient linear model system

$$\dot{y}(x) = My(x) + g(x), \qquad M \text{ a normal matrix, } M = XDX^{-1}. \tag{2.25}$$

This content downloaded from 131.211.22.192 on Thu, 04 Apr 2019 14:53:59 UTC
All use subject to https://about.jstor.org/terms

Consider for (2.25) the matrix $\mathscr{A}_h$ given by (2.18), but with the last row again multiplied by two. We then can write $\mathscr{A}_h^* \equiv (2h)^{-1}\mathscr{A}_h$ in the form

$$\mathscr{A}_h^* = (I \otimes X)\left(\frac{E \otimes I}{2h} - I \otimes D\right)(I \otimes X^{-1}),$$

E as in (2.12). The eigenvalues of $\mathscr{A}_h^*$ are the sN numbers (cf. [17, p. 259])

$$\lambda_j/2h - \delta_k, \qquad j = 1(1)N, k = 1(1)s, \tag{2.26}$$

where λ_j and δ_k are the eigenvalues of E and M, respectively (each eigenvalue δ_k of M plays the role of δ in the test-model (2.8)). Hence method (2.3) will perform satisfactorily on problem (2.25), for a certain h, if the eigenvalues (2.26) stay away from zero. Figure 1 shows all numerically computed eigenvalues of $E/2h$ for some values of the step size h. Note that some of the eigenvalues remain close to the imaginary axis if h decreases. Further, $\max \mathrm{Re}(\lambda_j/2h)$ slowly increases as h decreases. Figure 1 is useful to ascertain for which spectra of M the method will converge. For example, if M has positive eigenvalues δ_k, i.e., the problem is unstable: the method will perform satisfactorily for $h \leqslant h_0$ if $\max \delta_k < \max \mathrm{Re}(\lambda_j/2h_0)$. See also Fox and Mitchell [10], where it is pointed out that boundary value methods may have an advantage over step-by-step methods if the problem to be integrated is unstable.

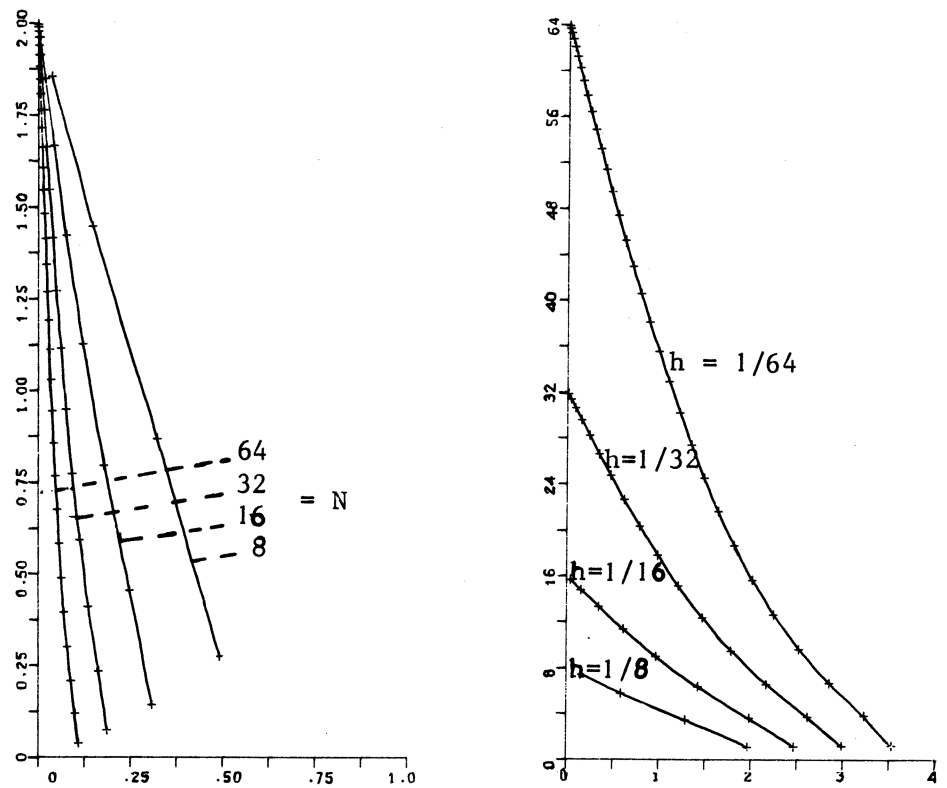


FIGURE 1

Eigenvalues of E (left plot) and E/2h (right plot) for $h = \frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \frac{1}{64}$. *We have only plotted eigenvalues with nonnegative imaginary part.*

This content downloaded from 131.211.22.192 on Thu, 04 Apr 2019 14:53:59 UTC
All use subject to https://about.jstor.org/terms

2.4. *A Numerical Illustration.* This section deals with a numerical example which serves to illustrate the convergence results derived in the previous section. For that purpose we selected the simple scalar problem

$$\dot{y}(x) = \delta\left(y(x) - \frac{1}{x+1}\right) - \frac{1}{(x+1)^2}, \qquad 0 \leqslant x \leqslant 1, y(0) = 1, \delta \in \mathbf{R}, \tag{2.27}$$

whose general solution is given by $y(x) = e^{\delta x}(y(0) - 1) + 1/(x + 1)$. Since $y(0) = 1$, only the smooth solution component $1/(x + 1)$ has to be computed. If $\delta \ll -1$, (2.27) is an example of a stiff problem where $e^{\delta x}y(0)$ represents the strongly varying solution component. In order to give sufficient insight into the error behavior, which has been predicted in Example 3, results will be shown for various choices of h and δ. We wish to emphasize that these results are not isolated. On the contrary, in a qualitative sense they are valid for systems as well. We refer to [20], [21] for extensive experiments with a known collection of stiff problems.

Table 2 contains results of method (2.3) for $h = 1/4, 1/8, 1/16$, and $\delta = -1, -5, -10, -100$. Table 3 shows results for $\delta = 1, 5, 10, 100$. The following observations are relevant. The lack of smoothness over the grid is clearly observable. However, when we consider either even grid points, or odd ones, the error behaves smoothly. Recall that we only have to compute the smooth solution of (2.27). For $\delta < 0$ the algorithm nicely shows its order two convergence at even-numbered grid points. Observe that after halving h the absolute error should decrease by a factor 4 because the method is of order two and that $-\log_{10}(\frac{1}{4}) \simeq 0.6$. At odd grid points the order behavior is much less pronounced as expected from Example 3. For $\delta > 0$ the algorithm yields more or less comparable results, though the second order not always shows up. This is because δ comes too close to the spectrum of $E/2h$ (cf. Figure 1).

TABLE 2

Results of method (2.3) *for problem* (2.27) *with* $\delta < 0$.
The table contains the value $-\log_{10}$ (*absolute error*).

δ	−1			−5			−10			−100		
x_n \ h	1/4	1/8	1/16	1/4	1/8	1/16	1/4	1/8	1/16	1/4	1/8	1/16
1/16			4.56			3.76			3.83			4.54
2/16		3.18	3.42		3.01	3.56		3.15	3.70		4.02	4.60
3/16			3.54			3.51			3.69			4.69
4/16	2.33	2.64	3.23	2.41	2.91	3.48	2.63	3.15	3.72	3.58	4.18	4.78
5/16			3.39			3.52			3.78			4.87
6/16		2.77	3.15		2.97	3.52		3.26	3.84		4.34	4.95
7/16			3.34			3.60			3.91			5.03
8/16	1.96	2.53	3.12	2.44	3.00	3.58	2.78	3.37	3.97	3.88	4.50	5.10
9/16			3.34			3.73			4.07			5.17
10/16		2.76	3.10		3.24	3.65		3.57	4.09		4.64	5.24
11/16			3.36			3.94			4.28			5.31
12/16	2.25	2.51	3.10	2.98	3.07	3.67	3.33	3.50	4.12	4.20	4.76	5.37
13/16			3.41			4.49			5.03			5.43
14/16		2.86	3.10		4.20	3.62		4.00	3.96		5.14	5.44
15/16			3.48			4.10			4.05			5.61
16/16	1.94	2.51	3.11	2.35	2.91	3.49	2.57	3.05	3.59	3.46	3.81	4.16

This content downloaded from 131.211.22.192 on Thu, 04 Apr 2019 14:53:59 UTC
All use subject to https://about.jstor.org/terms

TABLE 3

Results of method (2.3) *for problem* (2.27) *with* $\delta > 0$.
The table contains the values $-\log_{10}$ (*absolute error*).

δ	1			5			10			100		
x_n \ h	1/4	1/8	1/16	1/4	1/8	1/16	1/4	1/8	1/16	1/4	1/8	1/16
1/16			2.96			3.07			3.42			4.49
2/16		2.43	3.61		2.62	3.83		2.95	4.02		3.99	4.63
3/16			3.11			3.20			3.70			4.72
4/16	1.95	2.81	3.40	2.20	3.27	3.52	2.54	3.45	4.01	3.57	4.21	4.81
5/16			3.28			3.21			3.92			4.89
6/16		2.76	3.31		2.84	3.28		3.39	4.07		4.37	4.97
7/16			3.49			3.09			4.04			5.05
8/16	2.11	2.66	3.24	2.64	2.91	3.04	3.00	3.56	4.08	3.91	4.52	5.12
9/16			3.81			2.89			4.01			5.19
10/16		3.38	3.20		2.65	2.78		3.48	3.91		4.66	5.26
11/16			5.20			2.65			3.74			5.33
12/16	3.06	2.58	3.16	2.25	2.46	2.52	2.83	3.27	3.54	4.14	4.78	5.39
13/16			3.88			2.39			3.31			5.44
14/16		3.18	3.12		2.21	2.26		2.90	3.06		4.72	5.43
15/16			3.58			2.13			2.81			5.00
16/16	1.97	2.50	3.09	1.88	1.97	2.00	2.30	2.48	2.56	3.43	3.75	4.03

3. A Variational Approach.

3.1. *Preliminaries.* We consider nonlinear systems of ODE's

$$\dot{U} = \tilde{F}(t, U), \qquad 0 < t \leqslant T,\ U(\cdot) \in \mathbf{R}^m,\ U(0) \text{ prescribed.} \tag{3.1}$$

We first make a transformation of this equation to a more suitable form. In problems to be considered, there may exist positive stiffness parameters ε_i, such that parts of $\tilde{F}$ and the corresponding parts of the Jacobian matrix $\partial\tilde{F}/\partial U$ are unbounded as $O(\varepsilon_i^{-1})$, $\varepsilon_i \to 0$. We then multiply the corresponding equations by this parameter to get

$$\varepsilon\dot{U} = F(t, U), \qquad 0 < t \leqslant T, \tag{3.2}$$

where ε is a diagonal matrix with entries ε_i, $0 < \tilde{\varepsilon} \leqslant \varepsilon_i \leqslant 1$, and F and $\partial F/\partial U$ are bounded with respect to ε. A typical example is given by $\tilde{F}(t, U) = \tilde{A}U + \tilde{C}$

$$\tilde{A} = \begin{pmatrix} -1800 & 900 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1000 & -2000 \end{pmatrix}, \qquad \tilde{C} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1000 \end{pmatrix}$$

found in Enright et al. [8]. Here $\varepsilon = \operatorname{diag}(\frac{1}{900}, 1, \ldots, 1, \frac{1}{1000})$ is an obvious choice. In more general problems we may have to multiply by a more general positive-definite matrix ε, in order to get a bounded F and $\partial F/\partial U$. We further assume that F satisfies

$$(F(t, U) - F(t, V), U - V) \leqslant \rho(t)\|U - V\|^2 \quad \forall U, V \in \mathbf{R}^m,\ t > 0, \tag{3.3}$$

This content downloaded from 131.211.22.192 on Thu, 04 Apr 2019 14:53:59 UTC
All use subject to https://about.jstor.org/terms

where ρ: $[0, T] \to \mathbf{R}$ is at least piecewise continuous and independent of ε and $\rho(t) \leqslant -\rho_0$, $t \geqslant t_0 \geqslant 0$, $\rho_0 > 0$. Further $\|V\| = (V, V)^{1/2}$, where $(\cdot, \cdot)$ is the inner product in $\mathbf{R}^m$. As is well-known and easily seen, this means that, if U, V are two solutions of (3.2) corresponding to different initial values, then

$$\frac{1}{2}\frac{d}{dt}(\varepsilon(U - V), U - V) = (F(t, U) - F(t, V), U - V)$$

$$\leqslant \rho(t)\|U - V\|^2 \leqslant \rho(t)(\varepsilon(U - V), U - V), \qquad t \geqslant t_0,$$

so

$$\|U(t) - V(t)\|_\varepsilon^2 \leqslant \exp\left(\int_{t_0}^{t} 2\rho(s)\, ds\right)\|U(t_0) - V(t_0)\|_\varepsilon^2$$

$$\leqslant \|U(t_0) - V(t_0)\|_\varepsilon^2, \qquad t_0 \leqslant t \leqslant T,$$

where $\|V\|_\varepsilon = (\varepsilon V, V)^{1/2}$. This means that the system is contractive for $t \geqslant t_0$ if condition (3.3) holds. We further assume that F is Lipschitz continuous, i.e., there exists a constant C such that

$$\|F(t, U) - F(t, V)\| \leqslant C\|U - V\| \quad \forall U, V \in \mathbf{R}^m. \tag{3.4}$$

In the initial phase $(0, t_0)$, the system does not have to be contractive, i.e., the eigenvalues of the Jacobian may have positive real parts. In this interval we may choose to use a step-by-step method with very small step sizes, if it is of importance to follow the transients.

3.2. *The Galerkin Method.* We first describe the global Galerkin method to be used in the interval (t_0, T). We divide this interval into a number of subintervals (t_{i-1}, t_i), $i = 1, 2, \ldots, N$, where $t_N = T$. The length of the intervals, $t_i - t_{i-1}$, may vary smoothly with some function $h(t_i)$, but for ease of presentation, we assume that the intervals have equal length, i.e., $t_i - t_{i-1} = h$, $i = 1, 2, \ldots, N$. We consider each interval as an element on which we place some nodal points, $t_{i,j}$, $j = 0, 1, \ldots, p$, and $t_{i,j} = t_i + \xi_j h$, where ξ_j are the Lobatto quadrature points which satisfy $0 = \xi_0 < \xi_1 < \cdots < \xi_p = 1$, and $\xi_j + \xi_{p-j} = 1$. Hence the endpoints of the interval are always nodal points and (if $p > 1$) we choose also $p - 1$ disjoint nodal points in the interior of each element.

To each nodal point we associate a basis function $\phi_{i,j}$. The basis functions may be exponential or trigonometric functions and may also be discontinuous, but in this paper we only consider the most common choice where they are continuous and polynomials over each element. Basis functions corresponding to interior nodes have support only in the element to which they belong, and those corresponding to endpoints have support over the two adjacent elements (except those at t_0 and at t_N). The number of nodal points in each closed interval then equals the degree p of the polynomial plus one.

Let $\mathring{S}_h$ be the subspace of test functions which are zero at t_0, i.e.,

$$\mathring{S}_h = \text{SPAN}\{\phi_{i,j}, i = 0, 1, \ldots, N - 1, j = 1, 2, \ldots, p\}.$$

This content downloaded from 131.211.22.192 on Thu, 04 Apr 2019 14:53:59 UTC
All use subject to https://about.jstor.org/terms

Let

$$a(U;V) \equiv \int_{t_0}^{T} (\varepsilon \dot{U} - F(t,U), V)\, dt, \qquad U, V \in \left[H^1(t_0, T)\right]^m,$$

where $H^1(t_0, T)$ is the first-order Sobolev space of functions with square-integrable derivatives. To get an approximation $\tilde{U}$ of the solution of (3.2), we take a test function (vectorial function) $V = \phi_{i,j}^{[r]}$, and multiply the equation with V to get, after integration,

$$a\left(\tilde{U}; \phi_{i,j}^{[r]}\right) = \int_{t_{i-1}}^{t_{i+1}} \left(\varepsilon \dot{\tilde{U}} - F(t, \tilde{U}), \phi_{i,j}^{[r]}\right) dt = 0, \tag{3.5a}$$

$$j = 0,\ i = 1,2,\ldots,N-1,$$

$$a\left(\tilde{U}; \phi_{i,j}^{[r]}\right) = \int_{t_i}^{t_{i+1}} \left(\varepsilon \dot{\tilde{U}} - F(t, \tilde{U}), \phi_{i,j}^{[r]}\right) dt = 0, \tag{3.5b}$$

$$j = 1,2,\ldots,p-1,\ i = 0,1,\ldots,N-1.$$

At $t_N = T$, we get

$$a\left(\tilde{U}; \phi_{N,0}^{[r]}\right) = \int_{t_{N-1}}^{t_N} \left(\varepsilon \dot{\tilde{U}} - F(t, \tilde{U}), \phi_{N,0}^{[r]}\right) dt = 0. \tag{3.5c}$$

Here we choose in turn $\phi_{i,j}^{[r]} = \phi_{i,j} e_r$, where $\phi_{i,j}$ is the corresponding scalar basis function and e_r the rth coordinate vector. This defines the Galerkin approximation $\tilde{U}$ corresponding to $\mathring{S}_h$, where

$$\tilde{U} = U(t_0)\phi_{0,0} + \sum_{i=0}^{N-1} \sum_{j=1}^{p} d_{i,j}\phi_{i,j}, \qquad d_{i,j} \in \mathbf{R}^m,$$

i.e., we have imposed the essential boundary condition at t_0. Clearly,

$$a(U;V) = 0 \quad \forall V \in \left[H^1(t_0, T)\right]^m.$$

We then get from (3.5a)

$$a(U;V) - a(\tilde{U};V) = \int_{t_{i-1}}^{t_{i+1}} (\varepsilon[\dot{U} - \dot{\tilde{U}}] - [F(t,U) - F(t,\tilde{U})], V)\, dt = 0, \tag{3.6}$$

$$V = \phi_{i,j}^{[r]}, j = 0, i = 1,2,\ldots,N-1, r = 1,2,\ldots,m,$$

and similarly for (3.5b, c).

To estimate the Galerkin discretization error $U - \tilde{U}$, we let $U_I \in S_h$ be the interpolant to U on $\{t_{i,j}\}, j = 0,1,2,\ldots,p, i = 0,1,\ldots,N-1$, and we write

$$U - \tilde{U} = \eta - \theta,$$

where $\eta = U - U_I$ is the interpolation error and

$$\theta = -U + \tilde{U} + \eta = \tilde{U} - U_I.$$

This content downloaded from 131.211.22.192 on Thu, 04 Apr 2019 14:53:59 UTC
All use subject to https://about.jstor.org/terms

Note that $\theta \in \mathring{S}_h$. Assuming that the solution U is sufficiently smooth, from the interpolation error expansion in integral form we get the usual Sobolev norm estimates

$$\begin{aligned} &\int_{t_0}^{T} \|U - U_I\|^2 \, dt \leqslant C_0 h^{2(p+1)} \int_{t_0}^{T} \|U\|_{p+1}^2 \, dt, \\ &\int_{t_0}^{T} \|\dot{U} - \dot{U}_I\|^2 \, dt \leqslant C_1 h^{2p} \int_{t_0}^{T} \|\dot{U}\|_{p+1}^2 \, dt, \end{aligned} \tag{3.7}$$

for the interpolation error. Here,

$$\|U\|_{p+1}^2 = \int_{t_0}^{T} \sum_{k=0}^{p+1} \left(\frac{\partial^k U}{\partial t^k}, \frac{\partial^k U}{\partial t^k} \right) dt$$

is the norm in the Sobolev space $H^{p+1}(t_0, T)$.

THEOREM 6. *Let U be the solution of* (3.2) *where* (3.3), (3.4) *are satisfied. Then the Galerkin solution $\tilde{U}$, in the space of piecewise polynomial continuous functions of degree p, defined by* (3.5a, b, c) *satisfies*

$$|||U - \tilde{U}||| = O(h^{p+\nu}) \left\{ \|\varepsilon U\|_{p+2}^2 + \|U\|_{p+1}^2 \right\}^{1/2}, \qquad h \to 0,$$

where $\nu = 1$ if $p = 1$, $1 \geqslant \nu \geqslant \frac{1}{2}$ if $p = 3, 5, \ldots$ and $\nu = 0$ if p is even, and

$$|||V|||^2 = \frac{1}{2}(\varepsilon V(T), V(T)) - \int_{t_0}^{T} \rho(t) \|V(t)\|^2 \, dt.$$

(Note that this estimate implies both a least-square estimate as well as a pointwise estimate at the endpoint of the interval.)

For a proof, see [3].

3.3. *Difference Schemes.* In order to get a fully discretized scheme we have to use numerical quadrature, which results in various difference schemes. We shall consider this only for the case $p = 1$. Then $\phi_{i,p} = \phi_i$ are the usual hat functions and there are no interior nodes. With

$$\tilde{U} = U(t_0)\phi_0 + \sum_{i=1}^{N} U_i \phi_i,$$

(3.5a) and (3.5c) imply

$$\begin{cases} \varepsilon(\tilde{U}_{i+1} - \tilde{U}_{i-1}) = 2 \int_{t_{i-1}}^{t_{i+1}} F(t, \tilde{U}_{i-1}\phi_{i-1} + \tilde{U}_i \phi_i + \tilde{U}_{i+1}\phi_{i+1}) \phi_i \, dt, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i = 1, 2, \ldots, N-1, \\ \varepsilon(\tilde{U}_N - \tilde{U}_{N-1}) = \int_{t_{N-1}}^{t_N} F(t, \tilde{U}_{N-1}\phi_{N-1} + \tilde{U}_N \phi_N) \phi_N \, dt. \end{cases} \tag{3.8}$$

We call this the *generalized midpoint rule difference scheme.* Let $F_i = F(t, \tilde{U})|_{t=t_i}$. If we use numerical integration by the trapezoidal rule, i.e.,

$$\int_{t_{i-1}}^{t_i} F\phi_i \, dt \simeq \frac{1}{2} h \left[F_{i-1}\phi_i(t_{i-1}) + F_i \phi_i(t_i) \right] = \frac{1}{2} h F_i,$$

This content downloaded from 131.211.22.192 on Thu, 04 Apr 2019 14:53:59 UTC
All use subject to https://about.jstor.org/terms

we recover the difference method (2.1), (2.2). As we know, this scheme is of $O(h^2)$, see Section 2.3. We consider now a more accurate difference scheme which we may derive from (3.8). For this purpose let

$$F(t) \simeq \frac{1}{2}[F_{i-1} + F_i] + \left(t - t_i + \frac{h}{2}\right)\frac{1}{h}(F_i - F_{i-1}), \qquad t_{i-1} \leqslant t \leqslant t_i,$$

except that for the last formula in (3.8) we use

$$F(t) \simeq \tfrac{1}{2}[F_{N-1} + F_N], \qquad t_{N-1} \leqslant t \leqslant t_N.$$

Then

$$\int_{t_{i-1}}^{t_i} F(t)\phi_i\,dt \simeq \frac{h}{4}(F_{i-1} + F_i) + \frac{h}{12}(F_i - F_{i-1}) = \frac{h}{6}(F_{i-1} + 2F_i),$$

$$i = 1,2,\ldots,N-1,$$

and similarly

$$\int_{t_i}^{t_{i+1}} F(t)\phi_i\,dt \simeq \frac{h}{6}(F_{i+1} + 2F_i).$$

Hence, the generalized midpoint rule (3.8) takes the form

$$\text{(3.9)} \qquad \begin{cases} \varepsilon(\tilde{U}_{i+1} - \tilde{U}_{i-1}) = \dfrac{h}{3}(F_{i-1} + 4F_i + F_{i+1}), & i = 1,2,\ldots,N-1, \\ \varepsilon(\tilde{U}_N - \tilde{U}_{N-1}) = \dfrac{h}{2}(F_{N-1} + F_N). \end{cases}$$

We notice that this is a combination of the Simpson and trapezoidal rules.

For this combination, numerical tests (see Tables 4 and 5) indicate very accurate results. Note that already on a very coarse mesh ($h = \frac{1}{4}$) the accuracy is high. For $\delta < 0$ (Table 4), the order of convergence seems to be $\simeq 3.5$.

Finally some remarks about methods for the solution of the algebraic systems. These have block-tridiagonal form. If we use a special starting scheme for the calculation of $\tilde{U}_1$, we may use a "shooting method" for the solution of (3.1), i.e.,

$$\tilde{U}_{i+1} = \tilde{U}_{i-1} + 2hF(t_i, \tilde{U}_i), \qquad i = 1,2,\ldots.$$

This is, of course, nothing but the two-step midpoint rule, which, as is well-known, is unstable for stiff problems (and of order $O(h^2)$ for nonstiff problems). If the order of the systems (3.1, 3.2) is large and $\partial F/\partial U$ is sparse we may, however, apply an iterative method, which would preserve sparsity. There exist methods, such as preconditioned generalized conjugate gradient methods, for which convergence of the iterations is fast; see, for instance, Axelsson [2], and Hageman and Young [11].

Hence the large size of the matrices which arise should not be detrimental for the application of the methods described in this paper.

From the analyses and the numerical experiments it is concluded that the global method is a robust reliable method for both stiff systems and systems with increasing fundamental solutions. It is particularly efficient when moderate accuracy is desired. It does not seem to be very sensitive to stiffness.

This content downloaded from 131.211.22.192 on Thu, 04 Apr 2019 14:53:59 UTC
All use subject to https://about.jstor.org/terms

In the case of high accuracy, large or nonlinear systems, the efficiency depends on the availability of good algebraic systems solvers.

TABLE 4

Results of method (3.9) *for problem* (2.27) *with* $\delta < 0$.

The table contains the value $-\log_{10}$ (*absolute error*).

δ	−1			−5			−10			−100		
x_n \ h	1/4	1/8	1/16	1/4	1/8	1/16	1/4	1/8	1/16	1/4	1/8	1/16
1/16			5.28			5.98			7.08			7.05
2/16		4.44	5.89		5.52	5.92		5.37	6.17		5.95	7.30
3/16			5.34			6.14			6.96			7.42
4/16	3.73	4.59	5.69	4.69	4.72	5.78	4.34	5.02	6.13	5.03	6.42	7.56
5/16			5.38			6.04			6.78			7.69
6/16		4.57	5.59		5.21	5.72		5.74	6.09		6.69	7.80
7/16			5.39			5.89			6.29			7.94
8/16	3.40	4.47	5.53	3.62	4.66	5.66	3.39	4.98	6.00	5.19	6.54	7.98
9/16			5.40			5.75			6.03			8.41
10/16		4.58	5.49		4.79	5.60		4.98	5.86		6.44	7.84
11/16			5.39			5.63			5.82			7.75
12/16	3.77	4.43	5.46	3.73	4.56	5.53	3.84	4.74	5.70	4.76	5.85	7.22
13/16			5.39			5.53			5.64			6.85
14/16		4.54	5.43		4.55	5.45		4.61	5.53		5.39	6.42
15/16			5.38			5.43			5.45			6.02
16/16	3.36	4.40	5.41	3.46	4.42	5.37	3.52	4.43	5.36	4.21	4.90	5.61

TABLE 5

Results of method (3.9) *for problem* (2.27) *with* $\delta > 0$.

The table contains the value $-\log_{10}$ (*absolute error*).

δ	1			5			10			100		
x_n \ h	1/4	1/8	1/16	1/4	1/8	1/16	1/4	1/8	1/16	1/4	1/8	1/16
1/16			4.96			4.59			5.98			6.99
2/16		4.08	7.70		4.08	4.96		4.93	7.57		5.92	7.49
3/16			4.97			4.44			6.12			7.40
4/16	3.23	5.16	6.33	3.45	4.21	4.53	3.96	5.85	6.73	4.99	6.63	7.64
5/16			4.96			4.24			6.03			7.71
6/16		4.10	5.94		3.75	4.20		5.21	6.02		6.57	7.86
7/16			4.95			4.00			5.65			7.95
8/16	3.72	6.84	5.69	3.43	3.56	3.91	4.72	5.40	5.44	5.36	6.83	8.07
9/16			4.93			3.74			5.15			8.17
10/16		4.09	5.52		3.26	3.62		4.69	4.89		7.20	8.27
11/16			4.90			3.48			4.62			8.40
12/16	3.31	4.93	5.38	2.77	3.00	3.35	3.98	4.19	4.35	4.79	6.10	8.28
13/16			4.88			3.21			4.07			7.97
14/16		4.05	5.27		2.72	3.08		3.64	3.80		5.44	7.00
15/16			4.84			2.94			3.53			6.15
16/16	4.40	4.56	5.16	2.24	2.45	2.80	2.79	3.09	3.26	4.13	4.74	5.28

Mathematical Institute
Catholic University
Toernooiveld
6525 ED Nijmegen, The Netherlands

Center for Mathematics and Computer Science
Kruislaan 413
1098 SJ Amsterdam, The Netherlands

This content downloaded from 131.211.22.192 on Thu, 04 Apr 2019 14:53:59 UTC
All use subject to https://about.jstor.org/terms

1. A. O. H. AXELSSON, "Global integration of differential equations through Lobatto quadrature," *BIT*, v. 4, 1964, pp. 69–86.

2. A. O. H. AXELSSON, "Conjugate gradient type methods for unsymmetric and inconsistent systems of linear equations," *Linear Algebra Appl.*, v. 29, 1980, pp. 1–16.

3. A. O. H. AXELSSON & J. G. VERWER, "Boundary value techniques for initial value problems in ordinary differential equations," Appendix, *Math. Comp.*, Supplements section, this issue.

4. J. R. CASH, *Stable Recursions*, Academic Press, London, 1979.

5. G. DAHLQUIST, *Stability and Error Bounds in the Numerical Integration of Ordinary Differential Equations*, Trans. Roy. Inst. Tech., No. 130, Stockholm, 1959.

6. K. DEKKER & J. G. VERWER, *Stability of Runge-Kutta Methods for Stiff Nonlinear Differential Equations*, North-Holland, Amsterdam, 1984.

7. M. DELFOUR, W. HAGER and F. TROCHU, "Discontinuous Galerkin methods for ordinary differential equations," *Math. Comp.*, v. 36, 1981, pp. 455–473.

8. W. H. ENRIGHT, T. E. HULL & B. LINDBERG, "Comparing numerical methods for stiff systems of ODEs," *BIT*, v. 15, 1985, pp. 10–48.

9. L. FOX, "A note on the numerical integration of first order differential equations," *Quart. J. Mech. Appl. Math.*, v. 7, 1954, pp. 367–378.

10. L. FOX & A. R. MITCHELL, "Boundary value techniques for the numerical solution of initial value problems in ordinary differential equations," *Quart. J. Mech. Appl. Math.*, v. 10, 1957, pp. 232–243.

11. L. A. HAGEMAN & D. M. YOUNG, *Applied Iterative Methods*, Academic Press, New York, 1981.

12. P. HENRICI, *Discrete Variable Methods for Ordinary Differential Equations*, Wiley, New York, 1962.

13. B. L. HULME, "Discrete Galerkin and related one-step methods for ordinary differential equations," *Math. Comp.*, v. 26, 1972, pp. 881–891.

14. H. B. KELLER, *Numerical Methods for Two-Point Boundary-Value Problems*, Blaisdell, Waltham, Mass., 1968.

15. H. B. KELLER, *Numerical Solution of Two-Point Boundary Value Problems*, SIAM Regional Conference Series in Applied Mathematics, No. 24, 1976.

16. J. D. LAMBERT, *Computational Methods in Ordinary Differential Equations*, Wiley, London, 1973.

17. P. LANCASTER, *Theory of Matrices*, Academic Press, New York, 1969.

18. F. W. J. OLVER, "Numerical solution of second order linear difference relations," *J. Res. Nat. Bur. Standards*, v. 71B, 1967, pp. 111–129.

19. F. W. J. OLVER & D. J. SOOKNE, "Note on backward recurrence algorithms," *Math. Comp.*, v. 26, 1972, pp. 941–947.

20. L. ROLFES, *A Global Method for Solving Stiff Differential Equations*, NRIMS Special Report, TWISK 228, CSIR, NRIMS, Pretoria, 1981.

21. L. ROLFES & J. A. SNYMAN, *An Evaluation of a Global Method Applied to Stiff Ordinary Differential Equations*, preprint, University of Pretoria, 1982.

22. L. F. SHAMPINE, "Solving ODEs with discrete data in SPEAKEASY," *Recent Advances in Numerical Analysis* (C. de Boor and G. H. Golub, eds.), Academic Press, New York, 1978.

23. H. J. STETTER, *Analysis of Discretization Methods for Ordinary Differential Equations*, Springer-Verlag, Berlin and New York, 1973.

This content downloaded from 131.211.22.192 on Thu, 04 Apr 2019 14:53:59 UTC
All use subject to https://about.jstor.org/terms