MATHEMATICS OF COMPUTATION, VOLUME 25, NUMBER 116, OCTOBER, 1971

Matricial Difference Schemes for Integrating Stiff Systems of Ordinary Differential Equations

By W. L. Miranker*

Abstract. In this paper we give a description and analysis of a class of matricial difference schemes. This class of schemes is based in part on a generalization of the feature of classical numerical methods of being characterized by approximations at a single point in the complex plane. The schemes introduced here are effective for integrating stiff systems.

1. Introduction. In this paper we give a description and analysis of a class of difference schemes of matricial type. The special nature of this class of schemes makes them effective for the integration of stiff systems of differential equations.

The system

$$\dot{x} = Ax \tag{1.1}$$

of ordinary differential equations has for its solution

$$x_{n+1} = \exp(hA)x_n \tag{1.2}$$

where h is a mesh increment and $x_n = x(nh)$. A difference approximation to (1.1) has a solution $u(t)$ which may be written as

$$u_{n+1} = K(hA)u_n. \tag{1.3}$$

$K(z)$ is typically a polynomial or rational function of z. The control of the error $e_n = u_n - x_n$ depends on $K(hA)$ being stable (boundedness of the norm of the powers of $K(hA)$) and the closeness of $K(hA)$ to $\exp(hA)$. Existing methods usually handle these two features in the following way; $K(hA)$ will be close to $\exp(hA)$ if it is close on the spectrum $\sigma(hA)$ of hA. Since $\sigma(A)$ has some arbitrary configuration in the complex plane, making $K(hA)$ close to $\exp(hA)$ is accomplished in two steps. First, $K(z)$ is chosen close to $\exp(z)$ in a neighborhood of $z = 0$ (e.g. $K(z) = 1 + z$ for the forward Euler). Then h is reduced until $h \times \sigma(A)$ is shrunk into the neighborhood of $z = 0$ where $K(z)$ is close to $\exp(z)$. The stability of $K(hA)$ is accomplished by making $|K(z)| \leqq 1$, typically in a set containing $z = 0$ (e.g. in the $z = x + iy$ plane this set is $x^2 + y^2 - 2x \geqq 0$ for the backward Euler). Then h is reduced until $h \times \sigma(A)$ of A is shrunk into this neighborhood.

The methods introduced here are not restricted to operate in this classical way. Rather the single point, the origin, is replaced by a set of points in the complex plane and the approximation of $K(z)$ to $\exp(z)$ is arranged at this set of points. To be effective, the set of points at which the approximation is to be made should have certain characteristics (to be described later) relative to the spectrum of A. Generally,

Received September 25, 1969, revised March 28, 1971.

AMS 1969 *subject classifications.* Primary 6560, 6561; Secondary 3937.

Key words and phrases. Stiff systems, matricial difference schemes.

* This work was performed while the author was a visiting Professor of Mathematics at the Hebrew University of Jerusalem.

these characteristics can be determined at the expense of additional computations. A favorable situation for this procedure is provided by stiff systems.

A definition of a stiff system is one for which the stiffness

$$s = \max |\lambda| / \min |\lambda| \tag{1.4}$$

is a large number. The max and min are taken over the nonzero $\lambda \in \sigma(A)$.

Since a classical method of numerical integration when applied to a stiff system is enormously handicapped by the requirement that $|h\lambda|$ is small for the large $|\lambda|$ in order to gain stability and accuracy, it becomes feasible to expand the computation to gain information about $\sigma(A)$ for exploitation by the method discussed here.

There are a number of schemes in the literature which deal with integrating stiff systems, (see references [1]–[5]). A general approach to stiffness is through the approach of A-stability, (see [7]). Here, an A-stable scheme, such as the backward Euler, is used. In the initial transitional region, h must be small to get any accuracy. However, as the solution smooths out as time increases, h is increased. The A-stability of the method guarantees not to excite the modes with large λ which have become quiescent, while the smallness of these modes allows accuracy to be maintained. In this connection, there are the results of G. G. Dahlquist [7] and C. W. Gear [8]. Dahlquist's result places a strong limitation on this approach, since he shows that A-stable methods of order greater than two do not exist within the class of linear multistep methods, considered by him.

In Section 2, we describe our integration schemes and give a stability and error analysis for them. In Section 3, we describe the results of calculations based on one family of our schemes. Included is a comparison with an effective classical method. In an appendix, we give an error analysis of a scheme due to W. Liniger and R. Willoughby [1]. This scheme is a scalar version of the class of schemes presented here where, moreover, the set of points in the complex plane (referred to above) consists of two points on the real line, one of which is the origin. Even for this extremely special case, enormously effective computations were performed for large nonlinear systems (see [1]).

2. Statement of Results. Consider the system of ordinary differential equations

$$\dot{x} = Ax, \qquad t > 0. \tag{2.1}$$

Here x is an m-vector and A is an $m \times m$ matrix. Let h denote a mesh increment and let $x_n = x(nh)$. Then

$$x_n = e^{Ah} x_{n-1}. \tag{2.2}$$

Now let**

$$L(z) = \sum (\alpha_i + z\beta_i) e^{(r-i)z}, \qquad R(z) = \sum (\gamma_i + z\delta_i) e^{(r-i)z} \tag{2.3}$$

and let

$$C(z) = L(z)[R(z)]^{-1}. \tag{2.4}$$

** Unless otherwise specified all sums are taken from 0 to r.

For any solution x of (2.1), we have

$$[L(hA) - C(hA)R(hA)]x_{n-r} = 0 \tag{2.5}$$

identically. This follows from (2.4) by inserting (2.2) and (2.3) into (2.5). This suggests the following difference scheme

$$\sum \alpha_i u_{n-i} + h \sum \beta_i A u_{n-i} - P(hA)\Big[\sum \gamma_i u_{n-i} + h \sum \delta_i A u_{n-i}\Big] = 0, \tag{2.6}$$

$n = r, r + 1, \cdots$, for determining the m-vector valued mesh function u_n as an approximation to x_n. Here, $P(z)$ is an approximation to $C(z)$ to be specified.

Let H be the shift operator

$$Hf(t) = f(t + h). \tag{2.7}$$

Let $\mathfrak{L} = \mathfrak{L}(H)$ and $\mathfrak{R} = \mathfrak{R}(H)$ be the operators corresponding to $L(z)$ and $R(z)$ respectively, i.e.

$$\mathfrak{L}(H) = \sum (\alpha_i + hA\beta_i)H^{r-i}, \qquad \mathfrak{R}(H) = \sum (\gamma_i + hA\delta_i)H^{r-i}. \tag{2.8}$$

Then (2.5) and (2.6) can be written respectively as

$$[\mathfrak{L}(H) - C(hA)\mathfrak{R}(H)]x_{n-r} = 0, \tag{2.9}$$

$$[\mathfrak{L}(H) - P(hA)\mathfrak{R}(H)]u_{n-r} = 0. \tag{2.10}$$

By subtracting (2.9) from (2.10), the error $e_n = u_n - x_n$ is seen to satisfy the following equation

$$[\mathfrak{L}(H) - P(hA)\mathfrak{R}(H)]e_{n-r} = [P(hA) - C(hA)]\mathfrak{R}(H)x_{n-r}. \tag{2.11}$$

Of course,

$$\mathfrak{R}(H)x_{n-r} = R(hA)x_{n-r}. \tag{2.12}$$

In order to estimate e_n, we first solve (2.11) for e_n.

Let

$$\mathfrak{S}(H) = \mathfrak{L}(H) - P(hA)\mathfrak{R}(H) \equiv \sum s_i H^{r-i}. \tag{2.13}$$

Thus

$$s_i \equiv s_i(A) \equiv \alpha_i + hA\beta_i - P(hA)(\gamma_i + hA\delta_i). \tag{2.14}$$

Using (2.12) and (2.13), we may rewrite (2.11) as

$$\mathfrak{S}(H)e_{n-r} = [P(hA) - C(hA)]R(hA)x_{n-r}. \tag{2.15}$$

Further let

$$S(z) = \sum s_i z^{r-i}. \tag{2.16}$$

Finally, define σ_j, $j = 0, 1, \cdots$, formally through

$$[z^r S(z^{-1})]^{-1} = \sum_{i=0}^{\infty} \sigma_i z^i. \tag{2.17}$$

Now multiply (2.15) by σ_{N-n} and sum the result over n from r to N. The left member resulting from this procedure is

$$\sum_{n=r}^{N} \sigma_{N-n} \mathcal{S}(H) e_{n-r} = \sum_{n=r}^{N} \sigma_{N-n} \sum_{i=0}^{r} s_i H^{r-i} e_{n-i}$$

$$= \sigma_0 R_0 e_N + (\sigma_1 R_0 + \sigma_0 R_1) e_{N-1} \tag{2.18}$$

$$+ \cdots + (\sigma_{N-r} s_0 + \cdots + \sigma_{N-2r} s_r) e_r$$

$$+ \text{multiples of } e_0, \cdots, e_{r-1}.$$

Using the relations (2.16) and (2.17) between the s_i and the σ_j simplifies (2.18) to

$$\sum_{n=r}^{N} \sigma_{N-n} \mathcal{S}(H) e_{n-r} = e_N + \text{multiples of } e_0, \cdots, e_{r-1}. \tag{2.19}$$

Thus this summing procedure solves (2.15) and gives

$$e_N = \sum_{n=r}^{N} \sigma_{N-n} [P(hA) - C(hA)] R(hA) x_{n-r} \tag{2.20}$$

$$+ \text{multiples of } e_0, \cdots, e_{r-1}.$$

Now, to get an estimate of e_N, we need a stability and an accuracy statement. The stability statement is given by the following two lemmas.

LEMMA 1. *If $\sum s_i(\lambda) z^{n-i}$ obeys the root condition for all eigenvalues λ of A, then the determinant of $S(z)$ obeys the root condition.*

Proof. Let $f(A) = \sum s_i(A) z^{n-i}$. Suppose the determinant $|f(A)|$ vanishes for a value of z, then $|f(A) + \mu I - \mu I|$ vanishes. Then $\mu = \mu + f(\lambda)$, for λ any eigenvalue of A, or $f(\lambda) = 0$ for that value of z. Q.E.D.

LEMMA 2. *Let the determinant $|S(z)|$ of $S(z)$ obey the root condition. If the determinant of s_0 is not zero, then the matrix $[z^r S(z^{-1})]^{-1}$ is analytic in a neighborhood of $z = 0$. Furthermore, the matrices σ_j, $j = 0, 1, \cdots$, given by* (2.17), *have uniformly bounded norms.*

Proof. Since $z^r S(z^{-1}) = \sum_{i=0}^{r} s_i z^i$ and $|s_0| \neq 0$, it is clear that $[z^r S(z^{-1})]^{-1}$ is analytic in a neighborhood of the origin. Since $|z^r S(z)^{-1}| = z^{mr} |S(z^{-1})|$, the root condition locates the roots of the polynomial $|z^r S(z^{-1})|$ outside the open unit disc and those on the boundary of the unit disc are simple. (Note that the apparent root at the origin occurring from z^{mr} is annihilated by a corresponding pole of $|S(z^{-1})|$.) Since

$$[z^r S(z^{-1})]^{-1} = [\text{matrix of polynomials}]/|z^r S(z^{-1})|,$$

it suffices to show that the power series for the reciprocal polynomial $|z^r S(z^{-1})|^{-1}$ has bounded coefficients, given that its only roots are outside the open unit disc, with those on the boundary being simple. Let $mr = q$ and let

$$|z^r S(z^{-1})|^{-1} = \left[\sum_{i=0}^{q} t_i z^i \right]^{-1} = \sum_{i=0}^{\infty} u_i z^i.$$

Then

$$u_n = \frac{1}{2\pi i} \oint \left(\zeta^{n+1} \sum_{i=0}^{q} t_i \zeta^i \right)^{-1} d\zeta,$$

where the contour of integration lies inside the unit disc and encircles the origin.

If we move the contour through the unit disc and out to infinity in all directions, the integral will vanish if $q \geqq 1$ and we are left with the sum of the residues. If there is a pole at ζ_0 on the unit disc, it is simple. Let the residue from it be $\mathcal{R}_0$. Then

$$|\mathcal{R}_0| = \left|\left(\zeta_0^{n+1} \sum_{j=0}^{q} j t_j \zeta_0^{j-1}\right)^{-1}\right| = \left|\sum_{j=0}^{q} j t_j \zeta_0^{j-1}\right|^{-1}$$

which is independent of n.

If there is a pole at ζ_1 of order $\rho + 1$ outside the unit disc, let the residue from it be $\mathcal{R}_1$. Then

$$\mathcal{R}_1 = D_\zeta^\rho\left[(\zeta - \zeta_1)^{\rho+1}\left(\zeta^{n+1} \sum_{j=0}^{q} t_j \zeta^j\right)^{-1}\right]_{\zeta=\zeta_1} = D_{\zeta_1}^\rho[(\zeta_1^{n+1} Q(\zeta_1))^{-1}],$$

where the polynomial Q is given by

$$Q(\zeta) = \left(\sum_{j=0}^{\infty} t_j \zeta^j\right) \Big/ (\zeta - \zeta_1)^{\rho+1}.$$

Then

$$\begin{aligned}\mathcal{R}_1 &= \sum_{j=0}^{\rho} \binom{\rho}{j} D_{\zeta_1}^j(\zeta_1^{-n})\, D_{\zeta_1}^{\rho-j}(\zeta_1 Q(\zeta_1)) \\ &= \sum_{j=0}^{\rho} (-1)^j \binom{\rho}{j} n(n+1)\cdots(n+j-1)\zeta_1^{-(n+j)}\, D^{\rho-j}(\zeta_1 Q(\zeta_1)).\end{aligned}$$

Then

$$|\mathcal{R}_1| \leqq (n + \rho)^\rho F/|\zeta_1|^n$$

where the factor F is independent of n. Since $|\zeta_1| > 1$, this residue vanishes as $n \to \infty$. Since there are at most a finite number of residues to be accounted for, the coefficients u_n are bounded uniformly in n and the lemma is proved.

If $S(z)$ satisfies the hypothesis of Lemma 2, then (2.20) may be used to get

$$||e_N|| \leqq \text{const } ||[P(hA) - C(hA)]R(hA)|| \sum_{n=r}^{N} ||x_{n-r}||, \tag{2.21}$$

where we have assumed that the initial errors, $e_0, e_1, \cdots, e_{r-1}$ may be neglected. If we assume that $Nh = 1$, this becomes

$$||e_N|| \leqq \text{const } h^{-1}\, ||[P(hA) - C(hA)]R(hA)||. \tag{2.22}$$

Now we turn to the question of accuracy. We must introduce hypotheses so that we can estimate $[P(hA) - C(hA)]R(hA)$.

Let

$$L(z) = z^\mu + O(z^{\mu+1}), \qquad R(z) = z^\nu + O(z^{\nu+1}) \tag{2.23}$$

as $z \to 0$. This hypothesis amounts to stating that the difference scheme (2.6) in the limiting cases $P = 0$ and $P = \infty$ has order of accuracy $\mu - 1$ and $\nu - 1$, respectively.

Let $P(z)$ be chosen as the polynomial which has contact τ_i with $C(z)$ at the nonzero nodes hz_i, $i = 1, \cdots, p$, i.e.

$$P^{(m)}(hz_i) - C^{(m)}(hz_i) = 0, \qquad m = 0, \cdots, \tau_i - 1. \tag{2.24}$$

Now divide the eigenvalues of A into $p + 1$ clusters, $k_0, \cdots, k_p$. Let $z_0 = 0$. All eigenvalues in k_i are closer to z_i than to all $z_j, j \neq i, i = 0, 1, \cdots, p$. Ties are resolved randomly. Let

$$d_i = \max_{\lambda_j \in k_i} |\lambda_j - z_i|, \qquad i = 0, \cdots, p. \tag{2.25}$$

From (2.3) and (2.23), we have

$$|L(z)| \leqq \text{const min } (|z|, |z|^r), \qquad |R(z)| \leqq \text{const min } (|z|, |z|^\mu) \tag{2.26}$$

for Re $z \leqq 0$.

Now using the spectral representation theorem (see [10]) (for simplicity we treat only the case that the eigenvalues of A are distinct), we have

$$\begin{aligned}
&[P(hA) - C(hA)]R(hA) \\
&\quad = \sum_{i=0}^{p} \sum_{\lambda_j \in k_i} [P(h\lambda_j) - C(h\lambda_j)]R(h\lambda_j)z_{ij}(hA)^{***} \\
&\quad = \sum_{\lambda_j \in k_0} [P(h\lambda_j)R(h\lambda_j) - L(h\lambda_j)]z_{0j}(hA) \\
&\qquad + \sum_{i=1}^{p} \sum_{\lambda_j \in k_i} \frac{1}{\tau_i!} [h(\lambda_j - z_i)]^{\tau_i} [P^{(\tau_i)}(h\tilde{\lambda}_{ij}) - C^{(\tau_i)}(h\overset{\approx}{\lambda}_{ij})]R(h\lambda_j)z_{ij}(hA).
\end{aligned} \tag{2.27}$$

Here $\tilde{\lambda}_{ij}$ and $\overset{\approx}{\lambda}_{ij}$ arise from Taylor's theorem and represent appropriate values corresponding to λ_j and z_i and the functions P and C, respectively.

Then

$$\begin{aligned}
||[P(hA) - C(hA)]R(hA)|| \leqq C_1 \max (|hd_0|^r, |hd_0|^\mu) \\
+ C_2 \sum_{i=1}^{p} \frac{1}{\tau_i!} |hd_i|^{\tau_i}.
\end{aligned} \tag{2.28}$$

Inserting (2.28) into (2.22) gives

$$||e_N|| \leqq \frac{\text{const}}{h} \left[\max (|hd_0|^r, |hd_0|^\mu) + \sum_{i=1}^{p} \frac{1}{\tau_i!} |hd_i|^{\tau_i} \right]. \tag{2.29}$$

The constant will depend on many aspects of the difference scheme and conceivably could be quite large. For example, as $\overset{\approx}{\lambda}_{ij}$ approaches a root of $R(z)$, the constant C_2 will grow without bound.

The derivation of (2.29) may be formulated as the following theorem.

THEOREM. *For $n = 0, 1, \cdots, N$ with $Nh = 1$, let x_n be a solution of the differential equations* (2.1), *let u_n be a solution of the difference equation* (2.6), *and let $e_n = u_n - x_n$. If* (2.6) *is stable (e.g. obeys the hypothesis of Lemma* 2) *and if* (2.6) *obeys the accuracy conditions characterized by* (2.23) *and* (2.24), *then modulo the initial errors $||e_N||$ has the bound* (2.29).

Remark. Classical linear multistep methods correspond to the case where $P \equiv 0$.

*** $z_{ij}(hA)$ are the polynomials entering into the spectral representation theorem i.e. z_{ij} is the polynomial of minimal degree which vanishes on the spectrum of A except that at $\lambda_j \in k_i, z_{ij} = 1$.

At another extreme, we may use (2.6) and disregard the node z_0 completely. This would give a more symmetric treatment.

Example. A simple example of the scheme (2.6) corresponds to $r = 1$, $\alpha_0 = 1$, $\alpha_1 = -1$ and $\delta_1 = 1$. All other coefficients are zero. We select one node, i.e., $p = 1$. $P(z)$ is taken to be the constant $C(hz_1)$. The difference scheme is

$$u_n - u_{n-1} = \frac{e^{hz_1} - 1}{hz_1} h\dot{u}_{n-1}. \tag{2.30}$$

For this scheme $\mu = 1$, $\nu = 1$, and $\tau_1 = 1$. Thus the scheme has zeroth order accuracy at the origin and at z_1. This low accuracy scheme may be viewed as the forward Euler with the mesh increment scaled by $(e^{hz_1} - 1)/hz_1$.

For this scheme $S(z) = Iz - (I + ((e^{hz_1} - 1)/z_1)A)$. By Lemma 1, the determinant $|S(z)|$ obeys the root condition if $z - 1 - ((e^{hz_1} - 1)/z_1)\lambda$ does for every eigenvalue λ of A. This latter requirement is seen to be satisfied for any choice of z_1 in an interval which itself contains the interval $(-\infty, \lambda)$. (We are assuming that $\lambda < 0$.) Thus, if z_1 is chosen as any lower estimate for the spectrum of A, (2.30) will be stable. Let us choose $z_1 = \min\{\lambda\} - d$ for some $d \geqq 0$. To simplify ideas let us consider the special case corresponding to $m = 2$ and to, say, $\lambda_2 = -1$ and λ_1, some very large negative number. The difference scheme then becomes

$$\begin{aligned} u_n - u_{n-1} &= \frac{e^{h(\lambda_1 - d)} - 1}{\lambda_1 - d} Au_{n-1} \\ &\approx \frac{1}{d - \lambda_1} Au_{n-1} \end{aligned} \tag{2.31}$$

since $\lambda_1 \ll -1$. This corresponds to a forward Euler with a very small mesh increment, $1/(d - \lambda_1)$.

Now since

$$\begin{aligned} x_n = e^{Ah}x_{n-1} \quad \text{and} \quad u_n &= \left[I + \frac{e^{h(\lambda_1 - d)} - 1}{\lambda_1 - d} A\right]u_{n-1}, \\ e_n &= \left[I + \frac{e^{h(\lambda_1 - d)} - 1}{\lambda_1 - d} A - e^{Ah}\right]e_{n-1} \\ &\equiv T(hA)e_{n-1}. \end{aligned} \tag{2.32}$$

$T(h\lambda)$ is then the difference between the exponential $e^{\lambda h}$ and the straight line $1 - (e^{h(\lambda_1 - d)} - 1)\lambda/(\lambda_1 - d)$. At the eigenvalue λ_1, we have

$$T(h\lambda_1) = \frac{d}{\lambda_1} + e^{h\lambda_1}\left[1 + d\left(\frac{1}{\lambda_1} - h\right) + O\left(d^2\left(\frac{1}{\lambda_1} - h\right)^2\right)\right] + O\left(\frac{d^2}{\lambda_1^2}\right). \tag{2.33}$$

Remark on the Nonlinear Case. Although we have only discussed the linear case, the numerical method which we are considering may be applied to nonlinear differential equations in a variety of ways. Since the method is itself a nonlinear method, the analysis which we have given can only be carried over in a formal way to the case of nonlinear differential equations. (This has been done in Section 5 of [9] for a related numerical method.) One method for applying the problem to the nonlinear case is as follows.

For the nonlinear differential equation

$$\dot{y} = f(y) \tag{2.34}$$

and at the mesh point x_n, we replace y by $z = y - y_n$ to get

$$\dot{z} = J(y_n)z \tag{2.35}$$

where J is the Jacobian

$$J(y_n) = \partial f(y)/\partial y. \tag{2.36}$$

The values of $z_{n-1}, z_{n-2}, \cdots$ which are needed are chosen as $y_{n-1} - y_n, y_{n-2} - y_n, \cdots$, respectively, and A is taken as $J(y_n)$. This procedure is standard in the subject and we refer to [1] for a report of calculations on a related method using this procedure for nonlinear problems.

3. Numerical Example. We chose

$$\begin{aligned} u_n - u_{n-1} + h[\theta\dot{u}_n + (1 - \theta)\dot{u}_{n-1}] \\ - P(hA)[u_n - u_{n-1} + h(\varphi\dot{u}_n + (1 - \varphi)\dot{u}_{n-1})] = 0 \end{aligned} \tag{3.1}$$

as an example of the class of schemes discussed above with which to perform some calculations. For A we chose the matrix

$$A = -\begin{bmatrix} 10^p & 0 & 0 \\ 0 & 10^q & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.2}$$

with p and q as parameters. The function $C(z)$ is

$$C(z) = \frac{(1 - \theta z)e^z - 1 - (1 - \theta)z}{(1 - \varphi z)e^z - 1 - (1 - \varphi)z}. \tag{3.3}$$

$P(z)$ was taken as the linear polynomial which interpolates $C(z)$ at the two points $-(1 + \rho/100)10^p$ and $-(1 + \rho/100)10^q$ with ρ a parameter.

As a comparison scheme, we chose the second-order scheme

$$v_n - v_{n-1} - \frac{h}{2}[\dot{v}_n + \dot{v}_{n-1}] = 0. \tag{3.4}$$

Let w_n be the exact solution of the differential equation, $\dot{w} = Aw$, at the mesh point nh. Let $E_n(u)$ be the vector whose ith component is

$$E_n^i(u) = u_n^i/w_n^i - 1. \tag{3.5}$$

Let the length of $E_n(u)$ be $||E_n(u)||$ and let

$$E_u = \sum_{n=1}^{N} ||E_n(u)||/N. \tag{3.6}$$

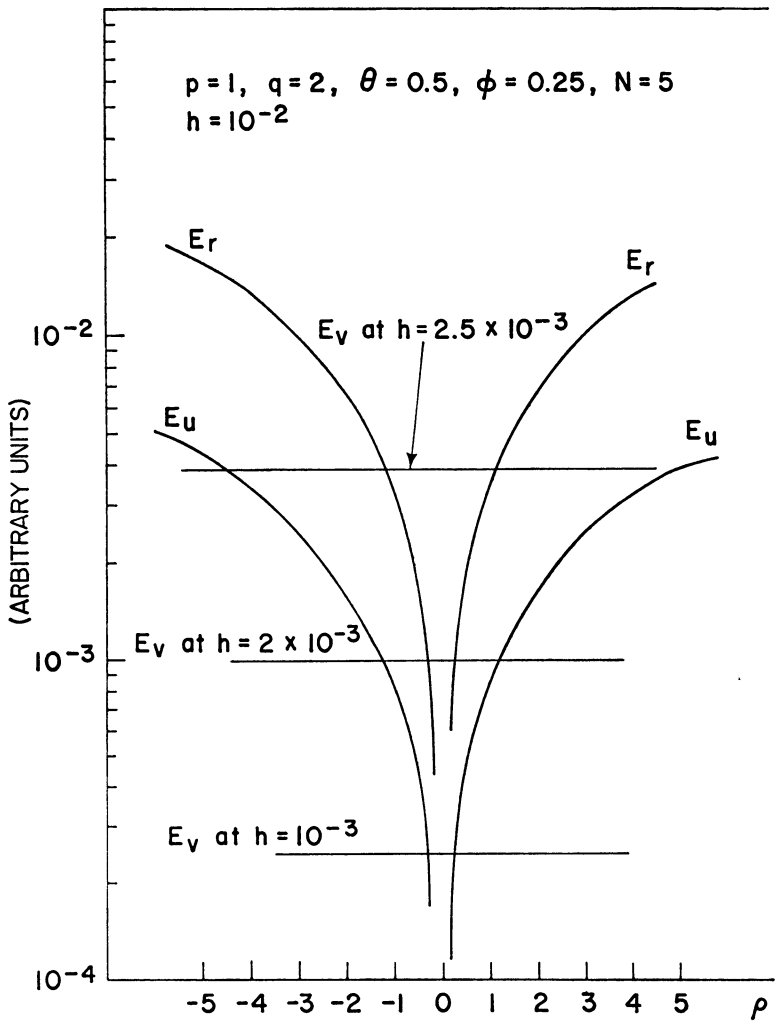
E_v is defined analogously with v_n^i replacing u_n^i in (3.5) and (3.6). E_u and E_v then are relative errors averaged over N time steps for the example scheme (3.1) and the comparison scheme (3.4), respectively. Finally, let

$$E_r = E_u/E_v. \tag{3.7}$$

In the following three graphs, we have plotted E_u, E_v, and E_r for various parameter values. The initial condition was chosen to be (1, 1, 1).

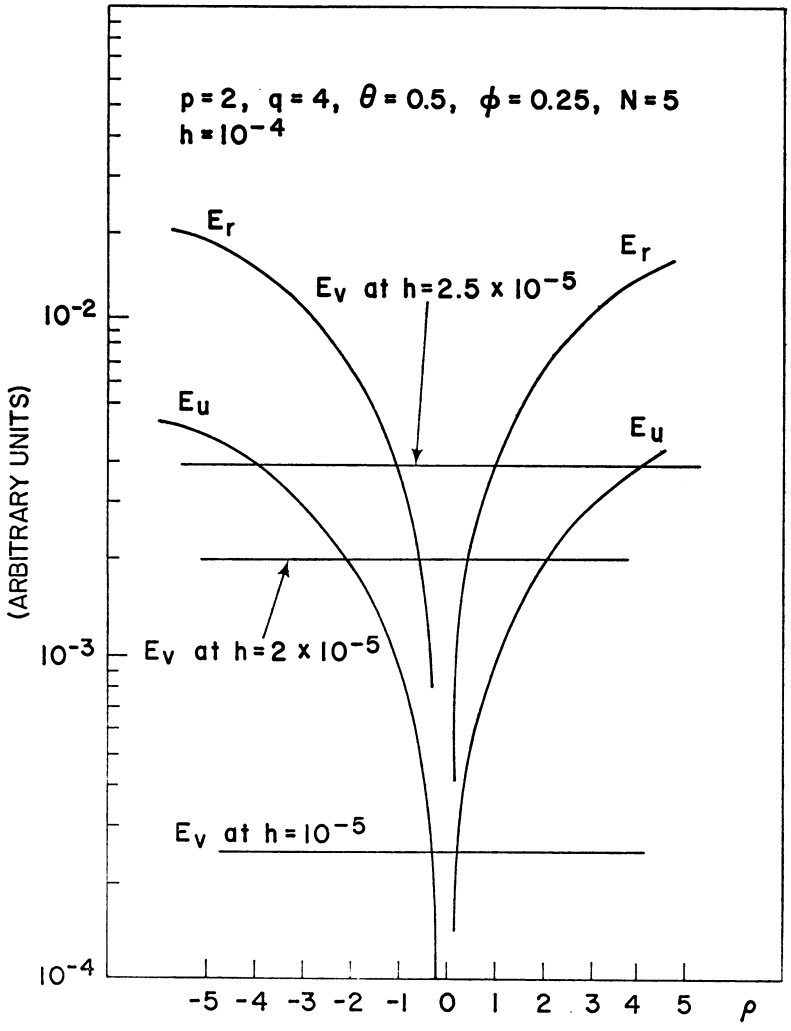
The crossings of the horizontal plot representing E_v with the curves representing E_u show with what accuracy the interpolation points must approximate the eigenvalues of A in order to match a calculation with the comparison formula with which a mesh increment, one order in magnitude finer, has been used.

All qualitative features of the graphs are as the theory above predicts.



Appendix.

The Liniger-Willoughby Scheme. This scheme is not in the class considered in this paper. However, it employs the idea of approximating the solution operator of the differential equation both at the origin as is usual and at one additional point using a free parameter. The scheme is simple and very effective for stiff systems with



one cluster of small and one cluster of large roots. It even works well in some other cases. The scheme is

(A.1) $$u_{n+1} - u_n = \mu h \dot{u}_n + (1 - \mu) h \dot{u}_{n+1}$$

with μ to be specified. In the linear case

(A.2) $$\dot{x} = -Ax,$$

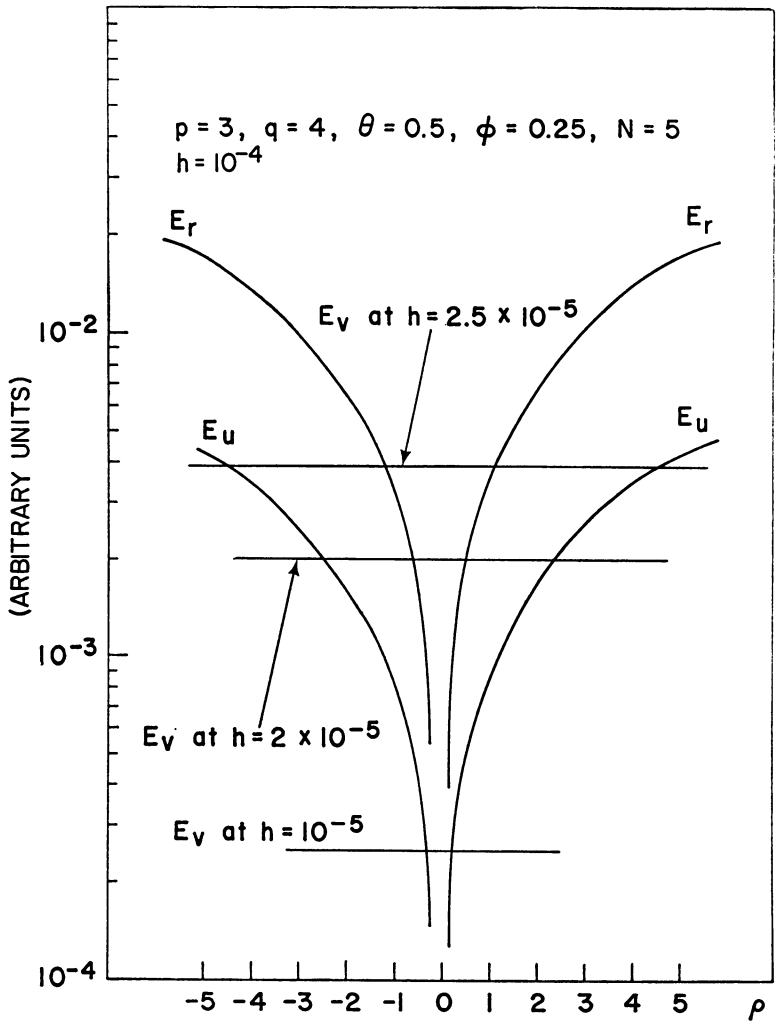
so that $\dot{u}_n$ is replaced by $-Au_n$ in (3.1).† The scheme becomes

(A.3) $$u_{n+1} = K(hA)u_n$$

with

(A.4) $$K(z) = (1 - \mu z)/(1 + (1 - \mu)z).$$

† Our analysis of this scheme makes use of the positive definiteness of A.



The error e_n obeys the equation

$$e_{n+1} = K(hA)e_n + (K(hA) - e^{-Ah})x_n. \tag{A.5}$$

Then

$$e_n = \sum_{i=0}^{n-1} K^i(hA)[K(hA) - e^{-Ah}]x_{n-i-1} + K^n(hA)e_0. \tag{A.6}$$

The function $K(z)$ is less than one in magnitude for all real $z > 0$ and all μ, $0 < \mu < \frac{1}{2}$. The theorem on the spectral resolution (quoted in the proof of Theorem 1 above) shows that

$$||K^n(hA)|| < \text{const}, \tag{A.7}$$

since the eigenvalues of A are strictly positive. Thus, from (A.6) and since the eigenvalues of A are strictly positive,

$$||e_n|| \leqq \text{const} \cdot [||K(hA) - e^{-Ah}||\,||+ ||e_0||]. \tag{A.8}$$

To estimate $||K(hA) - e^{-Ah}||$, we use the spectral resolution theorem again. We obtain the estimate

$$||K(hA) - e^{-Ah}|| \leqq \text{const} \cdot \min_{\Pi} \left[\max_{i \in I_1} |h^2\lambda_1^2| + \max_{i \in I_2} |r - \lambda|\right], \tag{A.9}$$

where I_1 and I_2 form a partition Π over the integers $1, \cdots, m$ and if $i \in I_2$ and $j \in I_2$, then $\lambda_i \neq \lambda_j$. The minimum is taken over all such partitions. r is some positive scalar, fixed for convenience. Once r is fixed, μ is chosen so that

$$K(hr) - e^{-hr} = 0. \tag{A.10}$$

The crux of the scheme is that this may be accomplished for any $r > 0$ and by a μ in the interval $(0, \frac{1}{2})$.

The estimate (A.9) then follows from the following observations.

$$|K(z) - e^{-z}| < \text{const} \cdot \min\{z^2, 1\}, \qquad z \geqq 0, \tag{A.11}$$

also

$$K(z) - e^{-z} = (r - z)[K'(\tilde{z}) + e^{-\tilde{z}}] \tag{A.12}$$

so that

$$|K(z) - e^{-z}| < \text{const} \cdot |r - z|, \qquad z > 0. \tag{A.13}$$

The boundedness of the constant here depends on $\mu \in (0, \frac{1}{2})$. Since the fit of $K(z)$ to e^{-z} at $z = r$ is only of first order, we cannot permit repeated roots in I_2 if (3.13) is to hold.

IBM
Thomas J. Watson Research Center
Yorktown Heights, New York 10598

1. W. LINIGER & R. WILLOUGHBY, "Efficient numerical integration of stiff systems of ordinary differential equations," *SIAM J. Numer. Anal.*, v. 6, 1969, pp. 47–66.

2. M. E. FOWLER & R. M. WARTEN, "A numerical integration technique for ordinary differential equations with widely separated eigenvalues," *IBM J. Res. Develop.*, v. 11, 1967, pp. 537–543. MR **35** #7586.

3. C. F. CURTISS & J. O. HIRSCHFELDER, "Integration of stiff equations," *Proc. Nat. Acad. Sci. U.S.A.*, v. 38, 1952, pp. 235–243. MR **13**, 873.

4. C. E. TREANOR, "A method for the numerical integration of coupled first-order differential equations with greatly different time constants," *Math. Comp.*, v. 20, 1966, pp. 39–45. MR **33** #889.

5. J. CERTAINE, "The solution of ordinary differential equations with large time constants," in *Mathematical Methods for Digital Computers*, Wiley, New York, 1960, pp. 128–132. MR **22** #8691.

6. W. L. MIRANKER & W. LINIGER, "Parallel methods for the numerical integration of ordinary differential equations," *Math. Comp.*, v. 21, 1967, pp. 303–320. MR **36** #6155.

7. G. G. DAHLQUIST, "A special stability problem for linear multistep methods," *Nordisk Tidskr. Informations-Behandling*, v. 3, 1963, pp. 27–43, MR **30** #715.

8. C. W. GEAR, *Numerical Integration of Stiff Ordinary Differential Equations*, Dept. of Computer Science, Report #221, University of Illinois, Urbana, Illinois.

9. W. L. MIRANKER, *Difference Schemes for the Integration of Stiff Systems of Ordinary Differential Equations*, IBM Research Center, Report #RC-1977, 1968.

10. F. R. GANTMACHER, *The Theory of Matrices*, GITTL, Moscow, 1953; English transl., Chelsea, New York, 1959. MR **16**, 438.