MATHEMATICS OF COMPUTATION, VOLUME 26, NUMBER 117, JANUARY, 1972

A Predictor-Corrector Method for a Certain Class of Stiff Differential Equations

By Karl G. Guderley and Chen-Chi Hsu*

Abstract. In stiff systems of linear ordinary differential equations, certain elements of the matrix describing the system are very large. Sometimes, e.g., in treating partial differential equations, the problem can be formulated in such a manner that large elements appear only in the main diagonal. Then the elements causing stiffness can be taken into account analytically. This is the basis of the predictor-corrector method presented here. The truncation error can be estimated in terms of the difference between predicted and corrected values in nearly the same manner as for the customary predictor-corrector method. The question of stability, which is crucial for stiff equations, is first studied for a single equation; as expected, the method is much more stable than the usual predictor- corrector method. For systems of equations, sufficient conditions for stability are derived which require less work than a detailed stability analysis. The main tool is a matrix norm which is consistent with a weighted infinity vector norm. The choice of the weights is critical. Their determination leads to the question whether a certain matrix has a positive inverse.

1. Introduction. The present paper studies a predictor-corrector method for stiff systems of differential equations which have the following form:

$$\frac{d\mathbf{y}}{dx} + \Lambda\mathbf{y} = A\mathbf{y} + \Gamma(x) \equiv \mathbf{f}(x, \mathbf{y}), \tag{1}$$

where Λ is a diagonal matrix which may have some large elements, and the right-hand side is considered as nonstiff. The operator on the left is inverted and the right-hand side is approximated by Lagrangian interpolation polynomials at grid points. The point at which the unknown vector is to be computed is excluded in the predictor phase, but it is included in the corrector phase. The integration of the exponential functions is done analytically.

In general, it would be too costly to bring a stiff system into the form (1). However, systems of this kind arise naturally if one reduces certain partial differential equations into a system of ordinary differential equations by a Galerkin procedure which uses approximating functions closely related to the partial differential equation, see for instance [1]. The ideas of usual predictor-corrector methods can be applied almost immediately to (1). In particular, it is possible to express the truncation error in terms of the difference between predicted and corrected values and to use this relation for step control.

For stiff systems, the question of stability is crucial. We discuss it, first for a single equation, and then for a system, by means of a simplified analysis. Actually, this

Received July 21, 1971.

AMS 1969 *subject classifications.* Primary 6561, 6580.

Key words and phrases. Predictor-corrector method, stiff differential equations, interval control, stability, weighted infinity vector norm.

* This work was carried out while one of the authors (C. C. Hsu) was pursuing an NRC-OAR postdoctoral resident research associateship.

latter discussion may be more interesting for the insight which it gives than for the practical work. In practice, the step control will probably suffice. It is true that a step control is only based on accuracy requirements, but the step selected by the method usually lies close to the stability limit, if a step of this size is compatible with the desired accuracy. The reason lies in the fact that instabilities lead to a deterioration of accuracy which leads to a reduction of the step size. The method would be A-stable if the matrix A vanishes. Therefore, one expects to find rather wide stability limits if A is sufficiently small.

Whenever one wants to use a large integration step for the solution of a stiff system, one is confronted with matrix inversions. A method can be made more effective if, within a certain interval which contains a number of integration steps, the operator which determines the differential equation is decomposed into a constant stiff part and a remainder. Then the matrix inversion is needed only once in each interval. The iteration steps which would be needed to invert the complete matrix (in our case $A - \Lambda$) are combined with measures for increasing the order of the integration method; they do not appear explicitly even in the derivation of the formulae. In Gear's work [2], this idea is expressed clearly; in other methods it would probably appear if they are implemented in an efficient manner. In the present procedure the matrix inversion is, of course, trivial. Inversions of (well-conditioned) matrices will appear at values x where the functions used in Galerkin's method are changed.

The inversion of the operator on the left of (1) leads to the equation

$$\mathbf{y}(x) = e^{-\Lambda(x-x_0)}\mathbf{y}(x_0) + e^{-\Lambda x}\int_{x_0}^{x} e^{\Lambda\tau}\mathbf{f}(\tau, \mathbf{y}(\tau))\, d\tau. \tag{2}$$

This equation is the point of departure for some papers [3], [10]. If one approximates $\mathbf{f}$ by a polynomial, then, after integration by parts, one is led to matrices $\exp(-\Lambda h)$, where h is the integration step. For a nondiagonal Λ, an approximation must be used at this stage. For stability reasons, it is advisable to use a rational function for the approximation of the exponential function [3], [4]. Here, a matrix inversion is required too. This additional work does not occur in the present approach, for Λ is a diagonal matrix. In Gear's work, an approximation of this kind is not needed for he makes the stronger assumption that $\mathbf{y}$ and $\mathbf{f}$, rather than $\mathbf{f}$ only, can be approximated by a polynomial. The difference has only minor importance; it would be felt in regions where those contributions of the first term on the right of (2) which are related to large elements of Λ are important and if A is small. But this happens only rarely and can be handled by taking a smaller integration step.

In summary, we can say that the present paper deals with stiff systems of a form which allows us to use a modified version of the predictor-corrector method for nonstiff systems. One obtains a fairly simple computational procedure and a convenient characterization of the truncation error.

2. Integration Formulae. Let the dimension of the vector $\mathbf{y}$ be N, and assume that $\mathbf{f}$ is approximated by a polynomial of degree k. Assume that the integration has progressed to a station $x_n = x_0 + nh$ and that we want to compute the value of $\mathbf{y}$ at x_{n+1}. The values of $\mathbf{y}$ and $\mathbf{f}$ at equidistant stations, $\mathbf{y}_{n-k+j}$ and $\mathbf{f}_{n-k+j}$ for $j = 0, 1, \cdots, k$, have been retained in the memory. Then the prediction polynomial for

$\mathbf{f}$ is obtained by Lagrangian interpolation. One finds

$$\mathbf{f}(x, \mathbf{y}) \simeq F^P B^P \mathbf{g}(\xi), \qquad \xi = (x - x_n)/h, \tag{3}$$

where F^P is an N by $(k+1)$ matrix whose jth column is given by $\mathbf{f}_{n-k+j}$, for short $F^P = (\mathbf{f}_{n-k}, \mathbf{f}_{n-k+1}, \cdots, \mathbf{f}_n)$, B^P is a $(k+1)$ by $(k+1)$ constant matrix, and $\mathbf{g}(\xi)$ is a $(k+1)$-dimensional vector whose jth component is given by ξ^{j-1}. The value of $\mathbf{y}$ at x_{n+1} is predicted by substituting (3) into (2) and carrying out a number of integrations by parts. One obtains

$$\mathbf{y}^P_{n+1} = e^{-\Delta h}\mathbf{y}_n + h[\mathbf{q}_{n+1} - e^{-\Delta h}\mathbf{q}_n], \tag{4}$$

in which the vector $\mathbf{q}_{n+\xi}$ is given by

$$\mathbf{q}_{n+\xi} = (\Delta h)^{-1}F^P B^P \mathbf{g}(\xi) - (\Delta h)^{-2}F^P B^P \frac{d\mathbf{g}}{d\xi} + \cdots + (-1)^k(\Delta h)^{-k-1}F^P B^P \frac{d^k\mathbf{g}}{d\xi^k}. \tag{5}$$

For small Δh, including $\Delta h = 0$, an alternate expression for $\mathbf{q}_{n+\xi}$ is found by expanding $\exp(\Delta h\xi)$ in (2) before the integration is carried out. One finds

$$\mathbf{q}_{n+\xi} = e^{-\Delta h\xi} \sum_j (\Delta h)^j F^P B^P \mathbf{s}_j(\xi), \qquad j = 0, 1, \cdots, \tag{6}$$

where the column vector $\mathbf{s}_j(\xi)$ is

$$\mathbf{s}_j(\xi) = \frac{1}{j!}\int_0^\xi \tau^j \mathbf{g}(\tau)\, d\tau. \tag{7}$$

The prediction formula (4) can be written in the form

$$\mathbf{y}^P_{n+1} = e^{-\Delta h}\mathbf{y}_n + h[V_0\mathbf{f}_n + V_1\mathbf{f}_{n-1} + \cdots + V_k\mathbf{f}_{n-k}], \tag{8}$$

where the diagonal matrices V_j depend only upon Δh.

From the predicted value $\mathbf{y}^P_{n+1}$ so obtained, one computes the predicted value of $\mathbf{f}_{n+1}$. The correction polynomial of degree k for $\mathbf{f}$ is obtained in a similar manner:

$$\mathbf{f}(x, \mathbf{y}) \simeq F^C B^C \mathbf{g}(\xi), \qquad \xi = (x - x_n)/h. \tag{9}$$

Here, F^C is an N by $(k+1)$ matrix whose jth column is given by $\mathbf{f}_{n-k+j+1}$, $F^C = (\mathbf{f}_{n-k+1}, \mathbf{f}_{n-k+2}, \cdots, \mathbf{f}_{n+1})$, B^C is a $(k+1)$ by $(k+1)$ matrix. Proceeding in complete analogy, one finds

$$\mathbf{y}^C_{n+1} = e^{-\Delta h}\mathbf{y}_n + h[W_0\mathbf{f}_{n+1} + W_1\mathbf{f}_n + \cdots + W_k\mathbf{f}_{n-k+1}], \tag{10}$$

where W_j are also diagonal matrices which depend only upon Δh. Specific expressions for V_j, W_j, B^P and B^C for the case $k = 4$ are given in the Appendix.

To start or restart the integration procedure, Picard's iteration method is employed. Assume that the initial value $\mathbf{y}_i$ at the station x_i is given, then the values $\mathbf{y}_{i+j}$, for $j = 1, 2, \cdots, k$, are computed simultaneously by iteration. Assume that the mth approximation for these values of $\mathbf{y}$ has been found, one then computes the corresponding values of $\mathbf{f}$ and obtains an interpolation polynomial for the mth approximation. The recurrence relation for generating these values of $\mathbf{y}$ are obtained from (2) and (3) with n in the definition of ξ and F^P replaced by $i + k$. One finds

$$\mathbf{y}^{(m+1)}_{i+j} = e^{-\Delta h}\mathbf{y}^{(m+1)}_{i+j-1} + h[\mathbf{q}^{(m)}_{i+j} - e^{-\Delta h}\mathbf{q}^{(m)}_{i+j-1}], \qquad j = 1, \cdots, k, \tag{11}$$

where the superscript denotes the order of approximation and the function q is defined by (5). The first approximation for y_{i+j} is obtained from (11) by neglecting the last term of the equation.

3. Truncation Error. In problems of the kind considered, initial perturbations die out with increasing x. The same is true for the propagation of errors unless the numerical procedure is not stable. Accordingly, most of the truncation error is generated locally, and the local truncation error should be a useful measure for the entire truncation error. In the present method, the local truncation error is caused by the fact that $\mathbf{f}$ is approximated by a polynomial of degree k. According to [5], the error caused by this approximation in the corrector phase is given by

$$\varepsilon(\xi) = \frac{\mathbf{f}^{k+1}(\bar{\xi})}{(k+1)!}(\xi - \xi_{1-k})(\xi - \xi_{2-k}) \cdots (\xi - \xi_0)(\xi - \xi_1), \qquad \xi = \frac{x - x_n}{h}, \tag{12}$$

where $\mathbf{f}^{k+1}(\bar{\xi})$ is the $(k+1)$th derivative evaluated at some station $\bar{\xi}$ in the interval $\xi_{1-k} \leqq \xi \leqq \xi_1$. The value $\bar{\xi}$ depends on the value of ξ for which $\varepsilon(\xi)$ is evaluated. The local truncation error for the corrector is obtained from (2) and (12). One finds

$$\mathbf{y}(x_{n+1}) - \mathbf{y}^C_{n+1} = \mathbf{t} = he^{-\Lambda h} \int_0^1 e^{\Lambda h \xi} \varepsilon(\xi)\, d\xi. \tag{13}$$

To compute t, an estimate for $\mathbf{f}^{k+1}(\bar{\xi})$ is required. In general, $\mathbf{f}^{k+1}$ is not available, therefore, the additional assumption is made that $\mathbf{f}$ is exactly given by a polynomial of degree $(k+1)$. Then the derivative $\mathbf{f}^{k+1}$ is a constant vector; it can be expressed in terms of the values $\mathbf{f}_{n-k}, \mathbf{f}_{n-k+1}, \cdots, \mathbf{f}_n$ and $\mathbf{f}_{n+1}$, and (13) can be evaluated.

If $\mathbf{f}^{k+1}$ is constant, the difference between the predicted and the corrected values is simply related to the local truncation error. This relation is used for step control. One has

$$\mathbf{y}^C_{n+1} - \mathbf{y}^P_{n+1} = G\mathbf{t}, \tag{14}$$

where G is a diagonal matrix which depends on the degree of the polynomial k and Λh. This matrix G can be determined from any convenient example, since it does not depend on $\mathbf{f}$. To derive (14), we first observe that $\mathbf{f}$ as a polynomial of degree $(k+1)$ depends linearly upon $(k+2)$ parameters. We choose for these parameters the values $\mathbf{f}_{n-k+1}, \cdots, \mathbf{f}_n, \mathbf{f}_{n+1}$ and the derivative $\mathbf{f}^{k+1}$. In the corrector formula, $\mathbf{f}$ is computed from $\mathbf{f}_{n-k+1}, \cdots, \mathbf{f}_n, \mathbf{f}_{n+1}$, the correction polynomial can therefore be obtained from the exact polynomial by setting $\mathbf{f}^{k+1} = 0$. One finds by an integration that

$$\mathbf{t} = hG_1\mathbf{f}^{k+1}, \tag{15}$$

where G_1 is a diagonal matrix which depends on k and Λh. This result is, of course, already implied by (13). In the predictor phase, the approximation polynomial is determined by $\mathbf{f}_{n-k}, \mathbf{f}_{n-k+1}, \cdots, \mathbf{f}_n$. But in the "exact" polynomial the constant vector $\mathbf{f}^{k+1}$ is a linear function of $\mathbf{f}_{n-k}, \mathbf{f}_{n-k+1}, \cdots, \mathbf{f}_n$ and $\mathbf{f}_{n+1}$. Thus, the prediction polynomial can be considered as a linear function of the parameters $\mathbf{f}_{n-k+1}, \cdots, \mathbf{f}_n, \mathbf{f}_{n+1}$ and $\mathbf{f}^{k+1}$. Now, for $\mathbf{f}^{k+1} = 0$, the prediction polynomial is the same as the correction polynomial, therefore

$$\mathbf{y}^{C}_{n+1} - \mathbf{y}^{P}_{n+1} = hG_2\mathbf{f}^{k+1}. \tag{16}$$

Again, G_2 is a diagonal matrix depending on k and Λh. Equations (15) and (16) then lead to (14).

To find the matrix G, we choose for convenience $\mathbf{f}_{n-k+j} = 0$ for $j = 0, 1, \cdots, k$ and $\mathbf{f}_{n+1}$ = unit vector. One then obtains

$$G = (k+1)\frac{\int_0^1 e^{\Lambda h\xi}\xi(\xi+1)\cdots(\xi+k-1)\,d\xi}{\int_0^1 e^{\Lambda h\xi}(\xi-1)\xi(\xi+1)\cdots(\xi+k-1)\,d\xi}. \tag{17}$$

For large values of Λh one has

$$G \sim -(k+1)\Lambda h. \tag{18}$$

The values of G for $k = 1, 2, 3$ and 4 are given in Fig. 1. In practice, a simple bound for G will be sufficient. For $k = 4$, the matrix

$$\tilde{G} = -0.95(10 + 5\Lambda h)(2 + \Lambda h)/(1 + \Lambda h) \tag{19}$$

serves this purpose, it satisfies

$$0.9\,|G| \leqq |\tilde{G}| \leqq |G|. \tag{20}$$

4. Stability. We shall use the usual stability definition, namely, that the solution should remain bounded as x tends to infinity. If the matrix A on the right-hand side of (1) is zero, the present integration method is perfect, except for errors in the

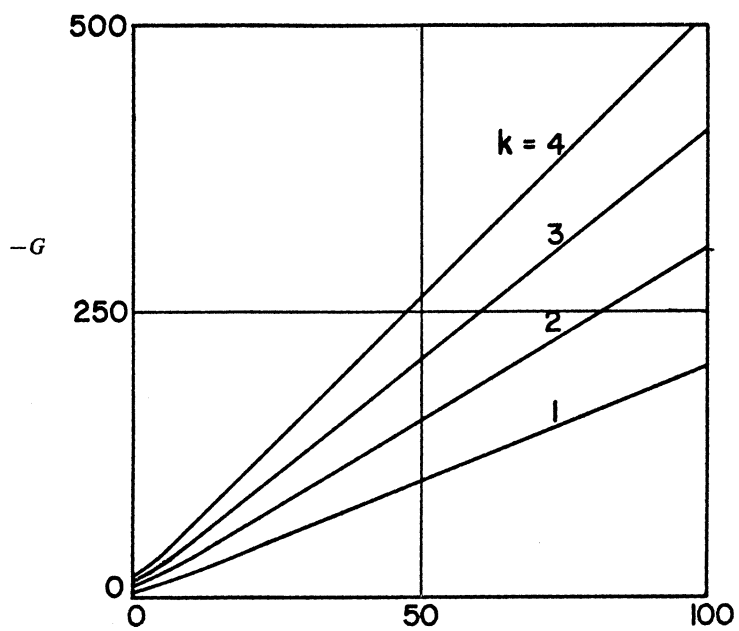


FIGURE 1. *Ratio Between the Difference of Corrected and Predicted y and the Truncation Error*

evaluation of integrals. Therefore, one would expect the method to be stable for rather large values of h if A is sufficiently small in comparison to Λ. For integration methods in which the matrix governing the system is treated as a whole, the stability analysis is greatly simplified by applying a similarity transformation which brings the matrix into its diagonal form; the results found from integrating the original system and the transformed system are exactly the same, except for round-off errors. Therefore, only a single equation needs to be considered in the stability discussion. However, if several matrices are considered simultaneously, as in the present method and also in Gear's approach [2], this simplification does not materialize. Nevertheless, we discuss the stability of a single equation in this section, for the results so obtained can be considered as an indication of the usefulness of the method. (This is the same as assuming that A is a diagonal matrix, too.) In the next section, we shall derive sufficient stability conditions for a system by a simplified analysis; even these discussions are still too complicated for practical use. In practice, one will depend upon the step control, via the control of truncation error, as a means of controlling the stability.

Assume that A in (1) is a diagonal matrix, then we examine the stability of a single equation of the form

$$y' + \lambda y = \gamma y. \tag{21}$$

In the integration formulae, (8) and (10), f_{n-k+i} is now replaced by γy_{n-k+i}. Substituting the predicted value of y into the corrector formula, one obtains

$$y^C_{n+1} = \alpha_0 y_n + \alpha_1 y_{n-1} + \cdots + \alpha_k y_{n-k}, \tag{22}$$

where the coefficients α_i depend only upon γh and λh; for $k = 4$, one has specifically

$$\begin{aligned} \alpha_0 &= e^{-\lambda h} + \gamma h(w_0 e^{-\lambda h} + w_1 + \gamma h w_0 v_0), \\ \alpha_1 &= \gamma h(w_2 + \gamma h w_0 v_1), \qquad \alpha_2 = \gamma h(w_3 + \gamma h w_0 v_2), \\ \alpha_3 &= \gamma h(w_4 + \gamma h w_0 v_3), \qquad \alpha_4 = (\gamma h)^2 w_0 v_4. \end{aligned} \tag{23}$$

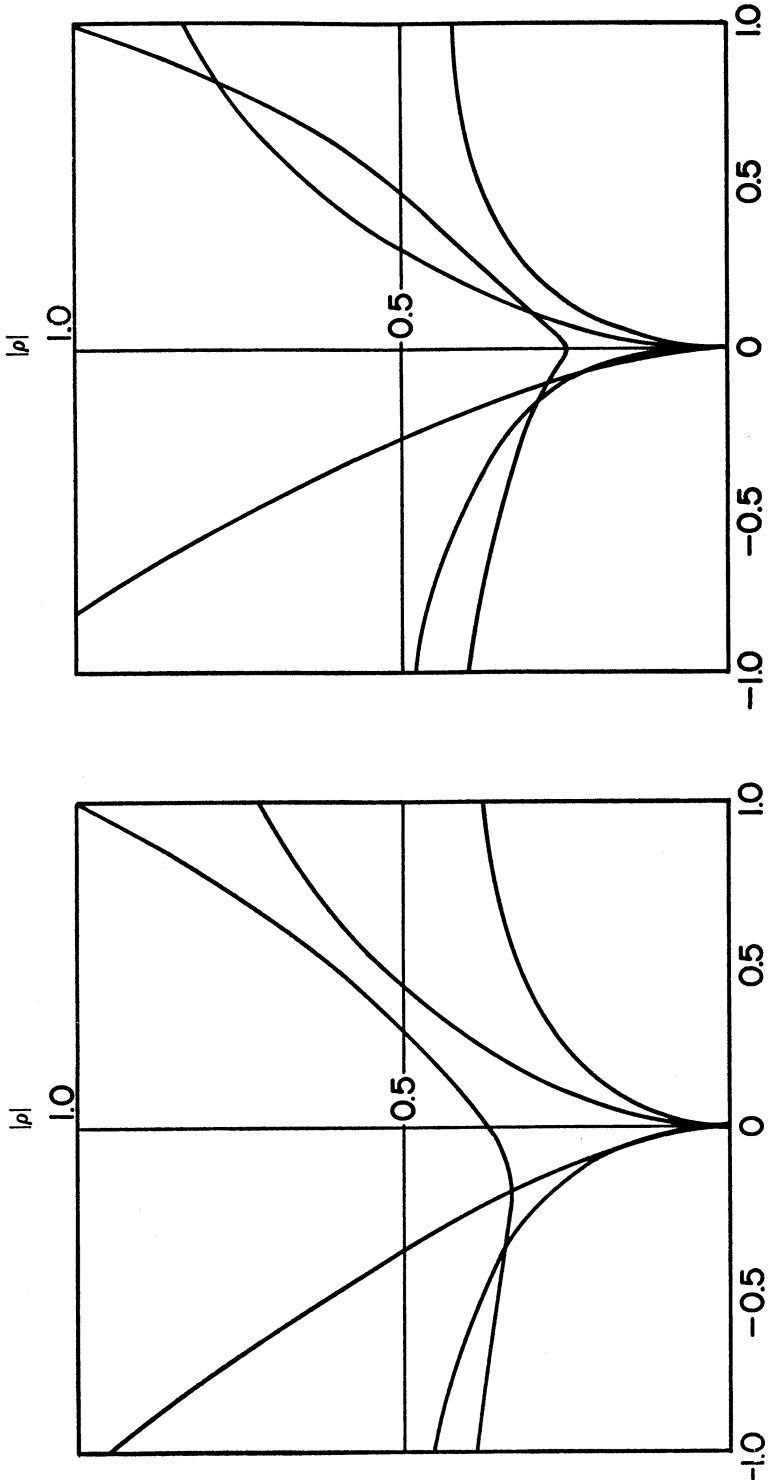
Here, w_i and v_i are elements of the matrices given in the Appendix. Now, consider the following characteristic polynomial:

$$\rho^{k+1} - \alpha_0 \rho^k - \alpha_1 \rho^{k-1} - \cdots - \alpha_{k-1}\rho - \alpha_k = 0. \tag{24}$$

The method is stable if the maximum of the absolute values of the roots of (24), $|\rho|_{\max}$, is less than one.

The function $|\rho|_{\max}$ is continuous in λh and γh, but its first derivative need not be continuous since $|\rho|_{\max}$ is not always attained by the same root. This can be seen from Fig. 2 which shows the relative magnitude of different roots in dependence upon λh and γh.

Fig. 3 presents for $k = 4$ the relation between $|\rho|_{\max}$ and the ratio γ/λ for various values of λh. For $\lambda h < 1.0$ and positive γh, $|\rho|_{\max}$ is almost exactly approximated by the principal root, $\exp[-(\lambda - \gamma)h]$. But for negative γh, $|\rho|_{\max}$ is sometimes attained by spurious roots which are not related to the ideal solution, even if λh is small. Fig. 4 gives the same curves as Fig. 3, but for larger values of λh. Here, no resemblance with the curve $|\rho| = \exp[-(\lambda - \gamma)h]$ exists. For $\gamma = 0$, the method is, of course,



(a) $\lambda h = 1.0$ and $k = 4$ (b) $\lambda h = 1.4$ and $k = 4$

FIGURE 2. *Effect of λh and γh on the Magnitude of the Roots*

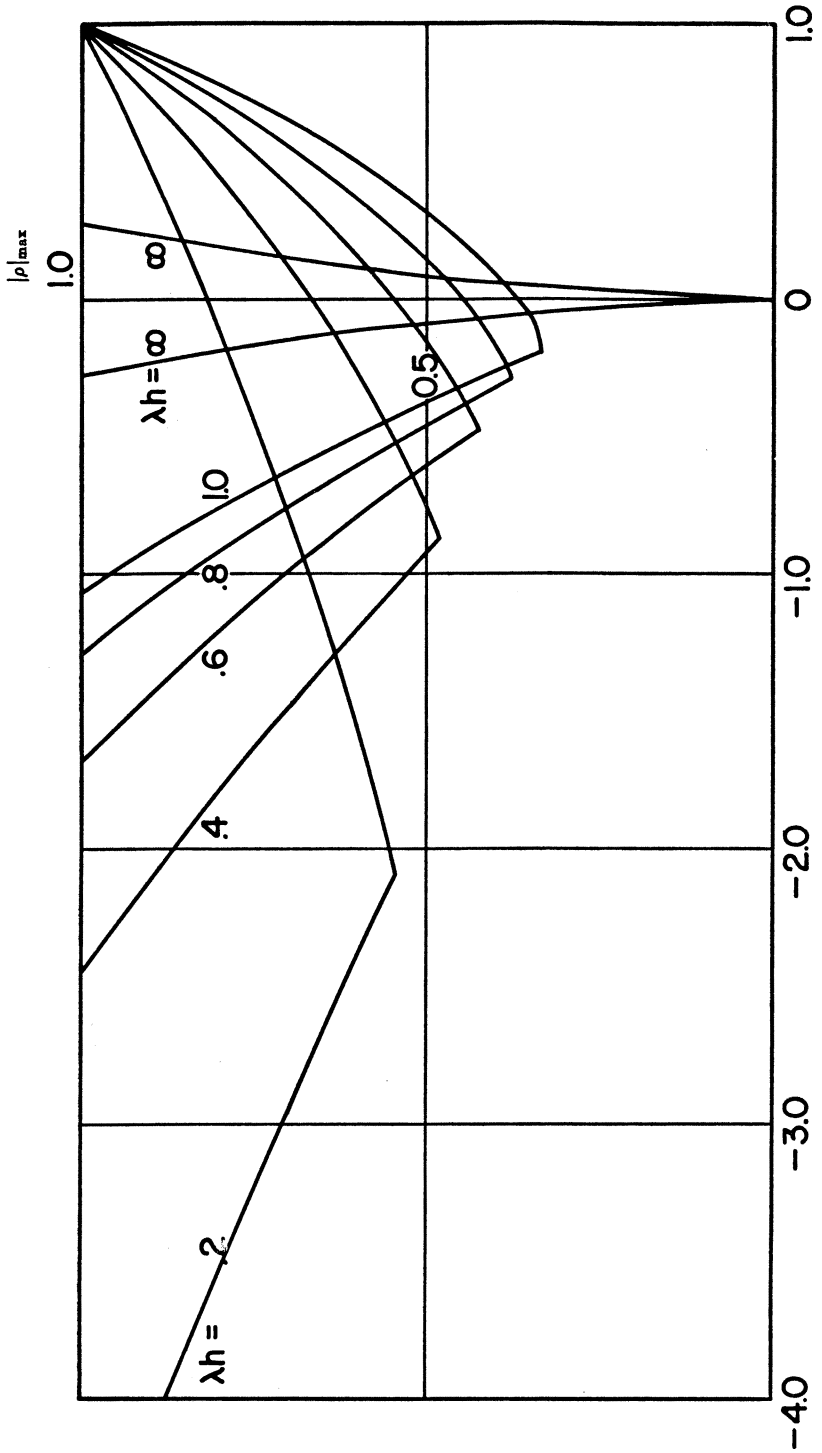


FIGURE 3. $|\rho|_{max}$ *versus* γ/λ *for* $k = 4$ *and Various* λh

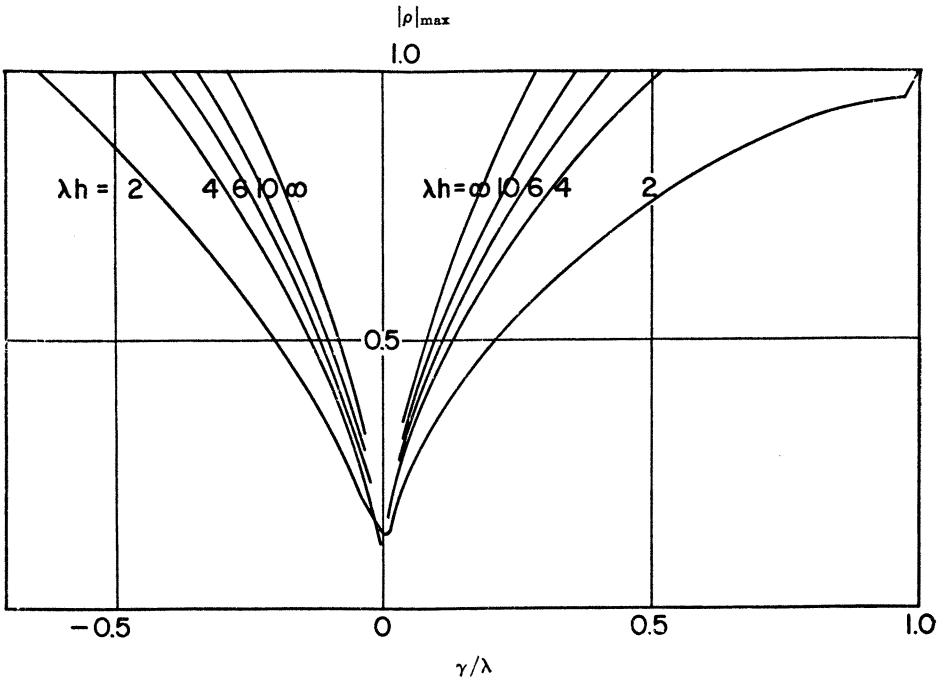


FIGURE 4. $|\rho|_{max}$ *versus* γ/λ *for* $k = 4$ *and* λh *Larger than One*

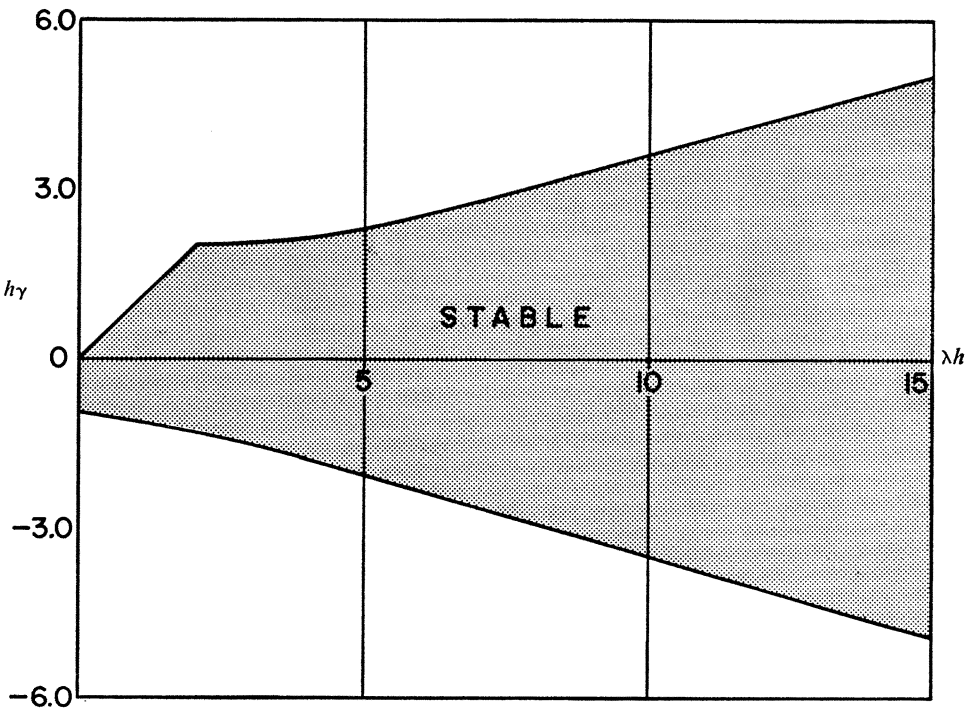


FIGURE 5. *Region of Stability for* $k = 4$

exact and $|\rho|_{max}$ is given by the principal root, $\exp(-\lambda h)$. But, even in a close vicinity of $\gamma = 0$, very marked deviations from the correct principal root are encountered. For the exact solution, the right-hand side of (21) is $\gamma \exp[-(\lambda - \gamma)h]$; in the integration scheme, this expression is replaced by a polynomial of degree k. The deviation between the correct principal root and the actual value of $|\rho|_{max}$ must be attributed to this approximation. For $k = 4$, the present method is stable for $|\gamma| \leqq 0.28\lambda$, even if λ approaches infinity.

Fig. 5 shows the region of stability in λ-γ plane for $k = 4$. Assume that one treats a number of equations simultaneously and that γ has about the same magnitude in all equations. Then, according to Fig. 5, the step size which is admissible from the point of view of stability is determined by those equations with small values of λ; under these circumstances, the stiffness of the system, which expresses itself by the presence of large values of λ, is no longer critical.

Figs. 6–9 show the effects of the degree k of the polynomial on the stability. The admissible step size is smaller for higher values of k. For $\lambda = 0$, the method reduces to the usual predictor-corrector method and the results shown in Fig. 6 are the stability criteria.

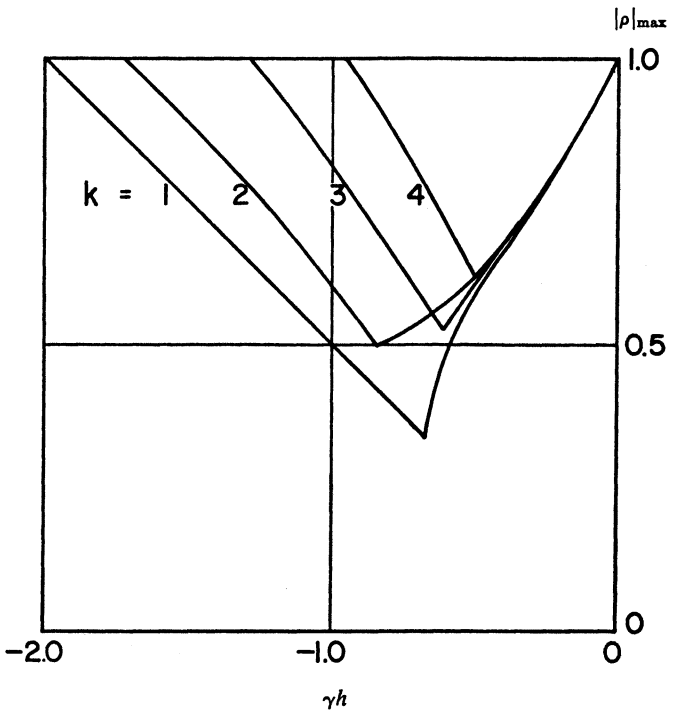


FIGURE 6. *Effect of k on* $|\rho|_{max}$ *for* $\lambda h = 0$

5. Stability Criteria for a System. For a more general stability discussion of the system (1) we consider A as a constant matrix. The inhomogeneous term $\mathbf{r}(x)$ has no influence in the stability analysis, therefore $\mathbf{f}_i$ in (8) and (10) is replaced by $A\mathbf{y}_i$. In the following analysis, the degree of the predictor and corrector polynomials is

assumed to be $k = 4$. Substituting the predictor formula (8) into the corrector formula (10), one finds

$$\mathbf{y}_{n+1}^{C} = Q_0\mathbf{y}_n + Q_1\mathbf{y}_{n-1} + Q_2\mathbf{y}_{n-2} + Q_3\mathbf{y}_{n-3} + Q_4\mathbf{y}_{n-4}, \tag{25}$$

where the matrices Q_j are given by

$$\begin{aligned} Q_0 &= e^{-\Lambda h} + hW_0A(e^{-\Lambda h} + hV_0A) + hW_1A, \\ Q_1 &= hW_2A + h^2W_0AV_1A, \qquad Q_2 = hW_3A + h^2W_0AV_2A, \\ Q_3 &= hW_4A + h^2W_0AV_3A, \qquad Q_4 = h^2W_0AV_4A. \end{aligned} \tag{26}$$

Now one sets

$$\mathbf{y}_j = \rho^j\mathbf{z}, \qquad \mathbf{y}_{n+1}^{C} = \mathbf{y}_{n+1}, \tag{27}$$

where $\mathbf{z}$ is a vector which does not depend on j, and ρ is a complex number. Substitution of (27) into (25) then yields a linear, homogeneous system

$$Q\mathbf{z} = 0, \tag{28}$$

with the N by N matrix Q given by

$$Q = \rho^5 I - \rho^4 Q_0 - \rho^3 Q_1 - \rho^2 Q_2 - \rho Q_3 - Q_4. \tag{29}$$

The system (28) has a nontrivial solution for $\mathbf{z}$ if and only if the determinant $|Q|$ vanishes. This requirement leads to a characteristic polynomial of degree $5N$ in ρ. The integration method is stable if all zeros of this polynomial lie within the unit circle.

For an assumed value of the step size h, the elements of the matrices $Q_0, \cdots, Q_4$ can be computed without too much effort. However, if N is not small, the evaluation of the determinant requires many multiplications of polynomials. A test which avoids the explicit evaluation of the characteristic polynomial would, therefore, be preferable. We make certain simplifications which allow us to derive a sufficient condition for the stability of the system which is somewhat easier to apply. Because of the simplification made, the test may sometimes fail, although the system is stable. Assume that the diagonal part of the matrix Q is nonsingular and let

$$Q = Q_{\mathrm{d}} + Q_{\mathrm{nd}}, \tag{30}$$

where Q_{d} and Q_{nd}, respectively, are the diagonal part and the nondiagonal part of Q. Then the system (28) can be written as

$$(I + Q_{\mathrm{d}}^{-1}Q_{\mathrm{nd}})\mathbf{z} = 0. \tag{31}$$

Stability is guaranteed if we can show that for $|\rho| \geqq 1$ only the trivial solution for $\mathbf{z}$ exists. Because of the triangular inequality, this is the case if, for $|\rho| \geqq 1$, the following condition is satisfied

$$||Q_{\mathrm{d}}^{-1}Q_{\mathrm{nd}}|| < 1. \tag{32}$$

Of course, the matrix norm used here must be consistent with the definition used for the vector norm.

For the vector norm, we propose to use a weighted infinity norm where the weights will be determined such that (32) is as permissive as possible. For such a norm defi-

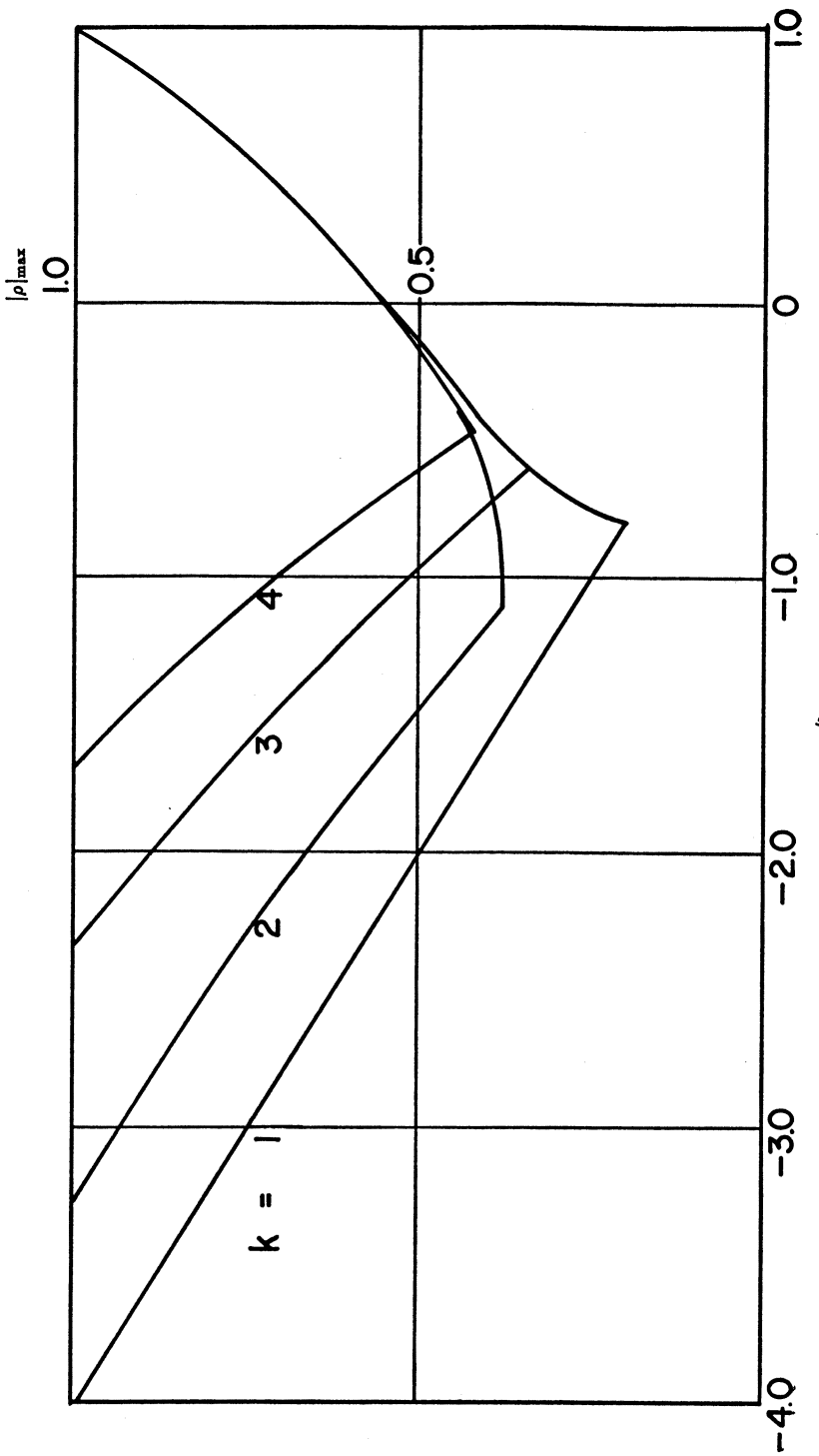


FIGURE 7. *Effect of k on* $|\rho|_{max}$ *for* $\lambda h = 0.6$

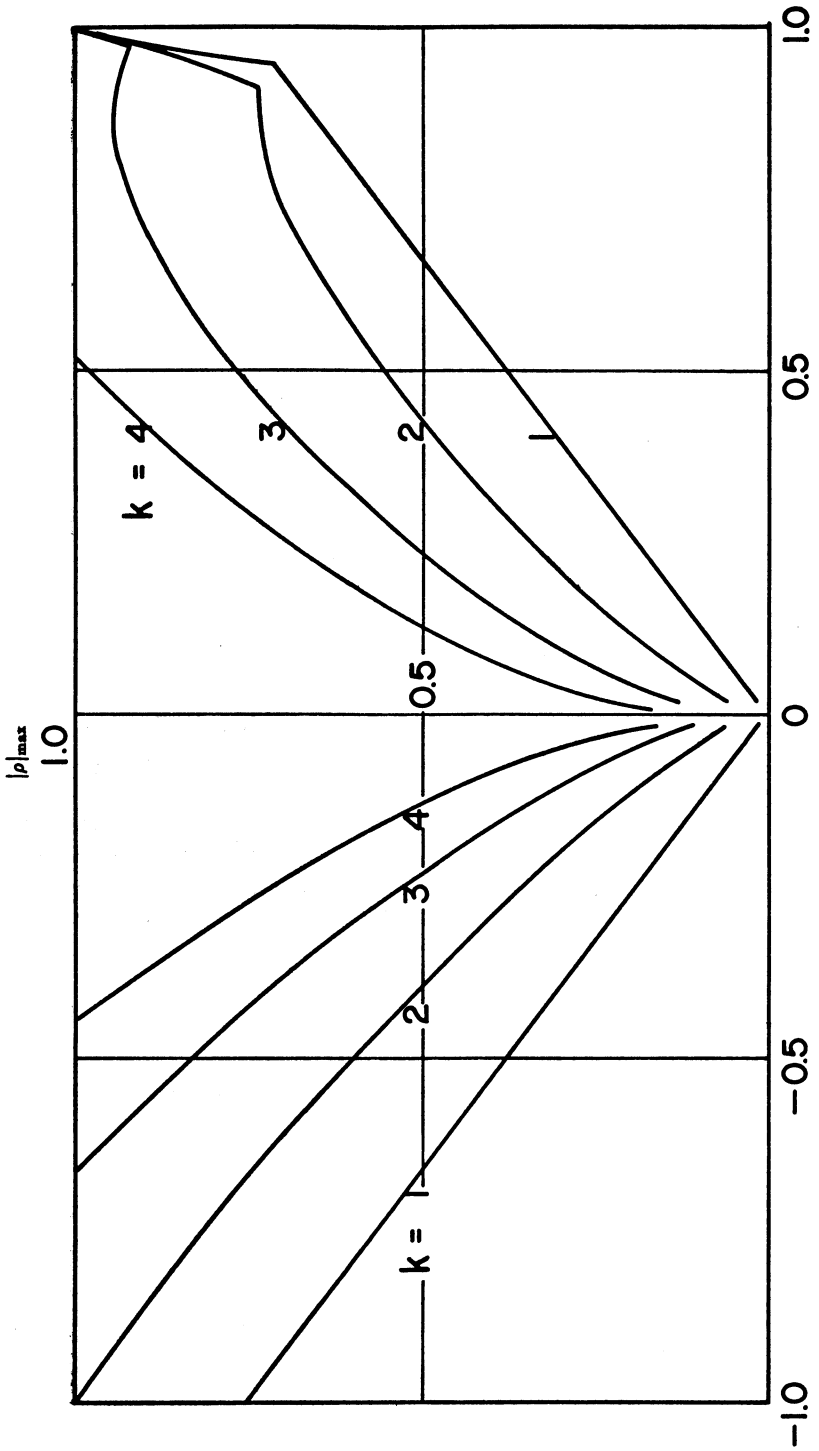


FIGURE 8. *Effect of k on* $|\rho|_{max}$ *for* $\lambda h = 4.0$

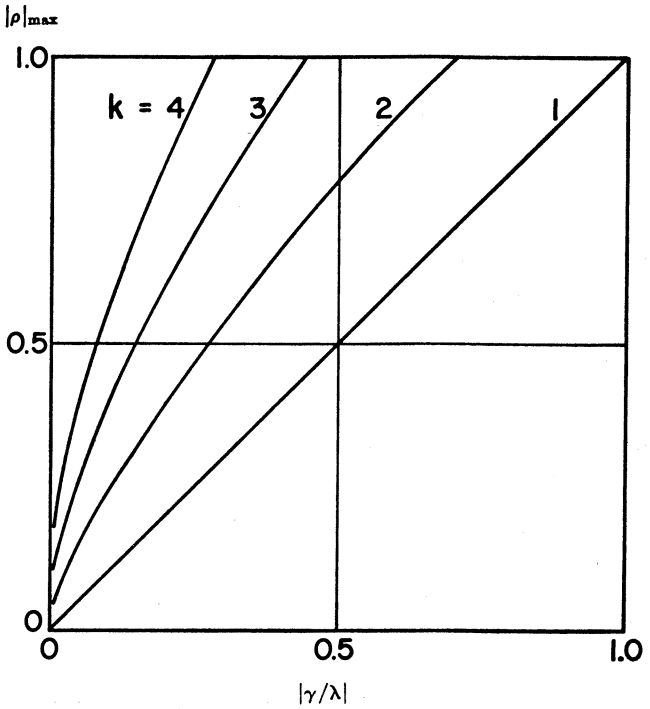


FIGURE 9. *Effect of k on* $|\rho|_{max}$ *for* λh *equal to Infinity*

nition, only the absolute values of the element of $Q_d^{-1}Q_{nd}$ are important. These elements are rational functions of ρ, which vanish at infinity; according to (29) the degree of the polynomials occurring in Q_d exceeds that of the polynomials for Q_{nd} by at least 1. We now make the assumption that zeros of all elements of Q_d lie within the unit circle. To check this, one can use the methods of [6], [7]. Then the elements of $Q_d^{-1}Q_{nd}$ are regular functions in the region $|\rho| \geqq 1$ (including the point at infinity) and assume their maximum along the boundary $|\rho| = 1$. Let H be a matrix whose elements are given by

$$H_{ik} = \max_{\theta}[|(Q_d^{-1}Q_{nd})_{ik}|, \rho = e^{i\theta}]. \tag{33}$$

We note that the diagonal elements of the matrix $Q_d^{-1}Q_{nd}$ are zero.

We introduce an auxiliary vector $\mathbf{u}$ with positive components u_i and define the vector norm by

$$||z|| = \max_{i}[(u_i)^{-1} |z_i|]. \tag{34}$$

It follows that

$$|z_i| \leqq u_i ||z||. \tag{35}$$

In order for the definition of a matrix norm to be consistent with the given vector norm (34), one must choose

$$||R|| = \max_{i}\left[(u_i)^{-1} \sum_{k} |R_{ik}| u_k\right]. \tag{36}$$

One has, according to the definition (34),

$$||Rz|| = \max_i \left[(u_i)^{-1} \left| \sum_k R_{ik} z_k \right| \right] \leqq \max_i \left[(u_i)^{-1} \sum_k |R_{ik}|\, |z_k| \right].$$

Using (35), one arrives at the following relation, which implies the definition of the matrix norm (36):

$$||Rz|| \leqq \max_i \left[(u_i)^{-1} \sum_k |R_{ik}| u_k ||z|| \right].$$

The definition of the matrix norm (36) is now applied in (32), and (33) is substituted. This gives the following stability condition:

$$(u_i)^{-1} \sum_k H_{ik} u_k < 1.$$

This condition can be written in matrix notation

$$(I - H)\mathbf{u} > 0, \tag{37}$$

where the inequality sign is to be applied componentwise. The method is stable for step h if (37) is satisfied for vector $\mathbf{u}$ with positive components. This requirement is not completely identical with the definition of matrices of the positive kind introduced in Section 23.1 of [8]. According to definition, the matrix $(I - H)$ is of the positive kind if the vector $\mathbf{u}$ has positive components whenever (37) is satisfied. We have the weaker requirement that there should exist at least one vector $\mathbf{u}$ for which (37) is satisfied. The following theorem shows that $(I - H)$ is actually a matrix of the positive kind.

THEOREM. *Let $H = (H_{ij})$ be a matrix for which $H_{ij} = 0$ if $i = j$ and $H_{ij} > 0$ if $i \neq j$. Then a vector $\mathbf{u} = (u_i)$, $u_i > 0$, for which $(I - H)\mathbf{u} > 0$ will exist if and only if the matrix $(I - H)^{-1}$ exists and has all positive elements.*

Remark. The assumptions made for H are rather strong. It is likely that the proof can be carried out for irreducible matrices in nearly the same form. From a practical point of view one also should consider reducible matrices. For the purpose at hand, one can evade this problem by a continuity argument, i.e. we replace the zeros in the off-diagonal elements of H by small positive quantities and then apply the present theorem.

Proof. First we discuss the case that $(I - H)^{-1}$ does not exist. We note that the matrix H satisfies the conditions of the Perron-Frobenius theorem for an irreducible matrix [9]. Accordingly, the maximum eigenvalue of H and its transpose H^t is positive and the respective eigenvectors $\mathbf{n}$ and ζ are positive vectors.

LEMMA 1. *The real (and the imaginary) parts of the components of all other eigenvectors cannot have the same sign.*

The eigenvectors of H which do not belong to the maximum eigenvalue must be orthogonal to ζ and the components of ζ are all positive.

LEMMA 2. *The vectors of the null space of $(I - H)$ can be considered as real.*

LEMMA 3. *A vector of the null space of $(I - H)$ cannot have some zero components if all other components are positive.*

Assume that the kth component of a vector $\mathbf{n}$ is zero but that all others are positive, then the kth component of $(I - H)\mathbf{n}$ is negative, which contradicts the assumption that $\mathbf{n}$ is a vector in the null space.

LEMMA 4. *If the null space of $(I - H)$ contains one vector $\mathbf{n}_1$ whose components are*

all positive, then the null space is one dimensional.

Assume that within the null space there is a second vector $\mathbf{n}_2$ with some positive and some negative components, then $\mathbf{n}_3 = \mathbf{n}_1 + \alpha\mathbf{n}_2$ is also a vector of the null space. One can choose α such that all components are nonnegative and at least one is zero. This contradicts Lemma 3. The same holds if $\mathbf{n}_1$ and $\mathbf{n}_2$ are linearly independent positive vectors.

LEMMA 5. *If $(I - H)$ has zero for an eigenvalue, then there is no vector $\mathbf{n}_1 > 0$ for which $(I - H)\mathbf{n}_1 > 0$.*

Case 1. The null space of $(I - H)$ is one dimensional and contains one positive vector $\mathbf{n}_1$. Then $\mathbf{n}_1$ is identical with the eigenvector of H which belongs to the maximum eigenvalue, for according to Lemma 1 all other eigenvectors of H have components with mixed signs. It follows that the vector adjoint to $\mathbf{n}_1$, to be denoted by ζ_1, is also positive. In order for the equation $(I - H)\mathbf{n} = \mathbf{z} > 0$ to be solvable for $\mathbf{n}$, one must have $(\mathbf{z}, \zeta_1) = 0$. But this condition cannot be satisfied for $\zeta_1 > 0$.

Case 2. The null space of $(I - H)$ contains one vector $\mathbf{n}_1$ with positive and negative components. Assume that there exists a vector $\mathbf{n}_2 > 0$ for which $(I - H)\mathbf{n}_2 > 0$. One can then form a vector

$$\mathbf{n}_3 = \alpha\mathbf{n}_1 + (1 - \alpha)\mathbf{n}_2, \qquad 0 < \alpha < 1,$$

for which at least one component is zero and all others are positive. Then one has on the one hand

$$(I - H)\mathbf{n}_3 = (1 - \alpha)(I - H)\mathbf{n}_2 > 0,$$

since $(I - H)\mathbf{n}_1 = 0$. On the other hand, assume that the kth component of $\mathbf{n}_3$ is zero. Then one finds by direct evaluation that the kth component of $(I - H)\mathbf{n}_3$ is negative. Because of this contradiction, the assumption that a vector $\mathbf{n}_2 > 0$ exists is wrong.

This concludes the discussion of cases where $(I - H)^{-1}$ does not exist.

If $(I - H)^{-1}$ exists, then the sufficiency of the theorem is trivial. Since $(I - H)^{-1}$ has all positive elements, an admissible vector $\mathbf{u} > 0$ can be constructed as $(I - H)^{-1}\mathbf{d}$, where $\mathbf{d} > 0$.

The necessity is next shown by contradiction. First, assume that there exists a vector $\mathbf{u}_I > 0$ for which

$$(I - H)\mathbf{u}_I > 0. \tag{38}$$

Next, assume that $(I - H)^{-1}$ has some nonpositive components:

$$[(I - H)^{-1}]_{mn} \leqq 0, \quad \text{for some suitable } (m, n). \tag{39}$$

Denote the nth column vector of $(I - H)^{-1}$ by $\mathbf{u}_{II}$, then the components of $\mathbf{u}_{II}$ are

$$(u_{II})_i = [(I - H)^{-1}]_{in}. \tag{40}$$

For some values of j, say $j = k', k'', m$,

$$(u_{II})_j \leqq 0. \tag{41}$$

We can now construct a vector

$$\mathbf{u}_{III} = (1 - \alpha)\mathbf{u}_I + \alpha\mathbf{u}_{II}. \tag{42}$$

For $\alpha = 0$, $\mathbf{u}_{III} > 0$; for $\alpha = 1$, the components $(u_{III})_j$, for $j = k', k'', m$, are nonpositive. Therefore, one can find a value of α, $0 < \alpha \leqq 1$, such that all components of $\mathbf{u}_{III}$ are nonnegative and at least one is zero. Assume that

$$(u_{III})_{k'} = 0 \quad \text{and} \quad (u_{III})_j \geqq 0 \quad \text{for } j \neq k'. \tag{43}$$

Based on (43), we can now compute $(I - H)\mathbf{u}_{III}$ by two different methods. If we write

$$(I - H)\mathbf{u}_{III} = (1 - \alpha)(I - H)\mathbf{u}_I + \alpha(I - H)\mathbf{u}_{II}, \qquad 0 < \alpha \leqq 1, \tag{44}$$

then all the components of $(I - H)\mathbf{u}_{III}$ are nonnegative because of (38) and since the components $(I - H)\mathbf{u}_{II}$ are given by δ_{jn}. In particular, one has

$$[(I - H)\mathbf{u}_{III}]_{k'} \geqq 0. \tag{45}$$

On the other hand, we can also write

$$[(I - H)\mathbf{u}_{III}]_{k'} = (u_{III})_{k'} - \sum_j H_{k'j}(u_{III})_j < 0, \tag{46}$$

which is negative because of (43) and the definition of H. The contradiction between (45) and (46) shows that the assumption (39) is false. This completes the proof of the theorem.

Instead of determining the inverse of $(I - H)$, one can solve the system

$$(I - H)\mathbf{u}_1 = \mathbf{x}_1 \tag{47}$$

with respect to $\mathbf{u}_1$ for some $\mathbf{x}_1 > 0$. A vector $\mathbf{u} > 0$ for which $(I - H)\mathbf{u} > 0$ will exist only if $\mathbf{u}_1 > 0$. The particular choice of $\mathbf{x}_1$ is unessential. This can be shown by replacing the vector $\mathbf{u}_{II}$ in the above argument by $\mathbf{u}$.

If a vector $\mathbf{u} > 0$ satisfying (37) exists, then $||H|| < 1$ and (32) will hold outside of the unit circle. The explicit form of $\mathbf{u}$ need not be known. Sometimes, in particular if H is small, it will be easiest to find $\mathbf{u}$ by inspection.

It is crucial that the matrix H give a rather narrow bound for $Q_d^{-1}Q_{nd}$. Such a bound would be obtained if one searches along the unit circle for the maxima of the individual elements of $Q_d^{-1}Q_{nd}$. If H is smaller, then rougher estimations will be sufficient. One can, for instance, find upper bounds for the elements of Q_{nd} and lower bounds for Q_d along the unit circle from the coefficient of the polynomials. If the determination of the lower bound of Q_d is critical, it might be preferable to search for it along the unit circle. The following examples show the effects of different approximations for $Q_d^{-1}Q_{nd}$.

We consider a system of two equations in which the diagonal matrix Λ is given by

$$\Lambda = \begin{bmatrix} 1 & 0 \\ 0 & 100 \end{bmatrix}.$$

For the matrix A, we have assumed three different forms

$$A_1 = \begin{bmatrix} \frac{1}{2} & 1 \\ 1 & 30 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} \frac{1}{2} & 1 \\ 1 & 20 \end{bmatrix}, \qquad A_3 = \begin{bmatrix} \frac{1}{2} & 1 \\ 1 & 10 \end{bmatrix}.$$

An exact stability analysis gives the corresponding admissible step sizes as

$$h_1 = 0.55, \qquad h_2 = 3.30, \qquad h_3 = 3.66.$$

If H is obtained by searching the maximum elements of $Q_d^{-1}Q_{nd}$ along the unit circle, then (47) gives the admissible step size for stability as

$$h_1 = 0.40, \qquad h_2 = 2.55, \qquad h_3 = 3.15.$$

These results show that the present approach gives good approximations to the actual stability limits. If H is found by bounding the elements of Q_d and Q_{nd} separately, the results obtained are less satisfactory. The values of h for which stability can then be guaranteed are

$$h_1 < 0.1, \qquad h_2 = 0.2, \qquad h_3 = 1.31.$$

The results summarized in Fig. 5 would suggest that a method is stable for a certain h if stability has been established for a larger h but this conjecture has not been proven.

Appendix. The constant matrices B^P and B^C, for $k = 4$, are

$$B^P = \begin{bmatrix} 0 & -1/4 & 11/24 & -1/4 & 1/24 \\ 0 & 4/3 & -7/3 & 7/6 & -1/6 \\ 0 & -3 & 19/4 & -2 & 1/4 \\ 0 & 4 & -13/3 & 3/2 & -1/6 \\ 1 & -25/12 & 35/24 & -5/12 & 1/24 \end{bmatrix},$$

$$B^C = \begin{bmatrix} 0 & -1/12 & -1/24 & 1/12 & 1/24 \\ 0 & 1/2 & 1/6 & -1/2 & -1/6 \\ 0 & -3/2 & 1/4 & 1 & 1/4 \\ 1 & 5/6 & -5/6 & -5/6 & -1/6 \\ 0 & 1/4 & 11/24 & 1/4 & 1/24 \end{bmatrix}.$$

Let $M \equiv \Lambda h$ and $E \equiv e^{-M}$, then the diagonal matrices V_i and W_i for $k = 4$ are given by

$$V_0 = M^{-5}\left[M^4(5 - E) - M^3\left(\frac{77}{12} - \frac{25}{12}E\right) + M^2\left(\frac{71}{12} - \frac{35}{12}E\right) - M\left(\frac{7}{2} - \frac{5}{2}E\right) + (1 - E)\right],$$

$$V_1 = -M^{-5}\left[10M^4 - M^3\left(\frac{107}{6} - 4E\right) + M^2\left(\frac{59}{3} - \frac{26}{3}E\right) - M(13 - 9E) + 4(1 - E)\right],$$

$$V_2 = M^{-5}\left[10M^4 - M^3\left(\frac{39}{2} - 3E\right) + M^2\left(\frac{49}{2} - \frac{19}{2}E\right) - M(18 - 12E) + 6(1 - E)\right],$$

$$V_3 = -M^{-5}\Big[5M^4 - M^3\Big(\frac{61}{6} - \frac{4}{3}E\Big) + M^2\Big(\frac{41}{3} - \frac{14}{3}E\Big) - M(11 - 7E) + 4(1 - E)\Big],$$

$$V_4 = M^{-5}\Big[M^4 - M^3\Big(\frac{25}{12} - \frac{1}{4}E\Big) + M^2\Big(\frac{35}{12} - \frac{11}{12}E\Big) - M\Big(\frac{5}{2} - \frac{3}{2}E\Big) + (1 - E)\Big];$$

$$W_0 = V_4,$$

$$W_1 = -M^{-5}\Big[M^4E - M^3\Big(4 + \frac{5}{6}E\Big) + M^2\Big(\frac{26}{3} - \frac{5}{3}E\Big) - M(9 - 5E) + 4(1 - E)\Big],$$

$$W_2 = M^{-5}\Big[-M^3\Big(3 + \frac{3}{2}E\Big) + M^2\Big(\frac{19}{2} - \frac{1}{2}E\Big) - M(12 - 6E) + 6(1 - E)\Big],$$

$$W_3 = -M^{-5}\Big[-M^3\Big(\frac{4}{3} + \frac{1}{2}E\Big) + M^2\Big(\frac{14}{3} + \frac{1}{3}E\Big) - M(7 - 3E) + 4(1 - E)\Big],$$

$$W_4 = M^{-5}\Big[-M^3\Big(\frac{1}{4} + \frac{1}{12}E\Big) + M^2\Big(\frac{11}{12} + \frac{1}{12}E\Big) - M\Big(\frac{3}{2} - \frac{1}{2}E\Big) + (1 - E)\Big].$$

We note that V_i and W_i are finite for $M = 0$.

Aerospace Research Laboratories
Wright-Patterson Air Force Base
Ohio 45433

Ohio State University Research Foundation
Columbus, Ohio 43210

1. K. G. GUDERLEY & C. C. HSU, "A special form of Galerkin's method applied to heat transfer in plane Couette-Poiseuille flows." (In prep.)

2. C. W. GEAR, "The automatic integration of stiff ordinary differential equations. (With discussion)," *Proc. IFIP Congress Information Processing* 68 (Edinburgh, 1968), vol. 1: *Mathematics, Software*, North-Holland, Amsterdam, 1969, pp. 187–193. MR **41** #4808.

3. D. A. CALAHAN, "Numerical solution of linear systems with widely separated time constants," *Proc. IEEE*, v. 55, 1967, pp. 2016–2017.

4. J. L. BLUE & H. K. GUMMEL, "Rational approximations to matrix exponential for systems of stiff differential equations," *J. Computational Physics*, v. 5, 1970, pp. 70–83. MR **40** #8267.

5. P. HENRICI, *Discrete Variable Methods in Ordinary Differential Equations*, Wiley, New York, 1962. MR **24** #B1772.

6. F. JOHN, *Lectures on Advanced Numerical Analysis*, Gordon and Breach, New York, 1967. MR **36** #4773.

7. M. MARDEN, *Geometry of Polynominals*, 2nd ed., Math. Surveys, no. 3, Amer. Math. Soc., Providence, R. I., 1966. MR **37** #1562.

8. L. COLLATZ, *Funktionalanalysis und numerische Mathematik*, Die Grundlehren der math. Wissenschaften, Band 120, Springer-Verlag, Berlin, 1964; English transl., Academic Press, New York, 1966. MR **29** #2931; MR **34** #4961.

9. R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, N. J., 1962. MR **28** #1725.

10. J. CERTAINE, "The solution of ordinary differential equations with large time constants," in *Mathematical Methods for Digital Computers*, A. Ralston and H. S. Wilf (Editors), Wiley, New York, 1960, pp. 128–132. MR **22** #8691.