Comp. & Maths. with Appls. Vol. 10, No. 6, pp. 419–432, 1984
Printed in the U.S.A.

0097–4943/84 $3.00 + .00
Pergamon Press Ltd.

STABILITY OF EXPLICIT RUNGE–KUTTA METHODS†

L. F. SHAMPINE
Numerical Mathematics Division, Sandia National Laboratories, Albuquerque, NM 87185, U.S.A.

(*Received March* 1984)

Abstract—The classical theory of stability of explicit Runge–Kutta methods is concerned with Lipschitzian problems. It is not useful for stable problems with "large" Lipschitz constants. The classical theory of absolute stability considers some very special problems of this kind. The problems treated arise when a general problem is linearized. It is hoped that the behavior in the case of the special problem provide guidelines as to the behavior in the case of a general problem. A synthesis is proposed here which responds to this unsatisfactory state of affairs in the classical theory.

1. INTRODUCTION

One-step methods, such as explicit Runge–Kutta, for the numerical solution of the initial value problem

$$y' = f(t, y), \quad a \leq t \leq b \tag{1.1}$$

$$y(a) \quad \text{given} \tag{1.2}$$

start with the approximation $y_0 \doteq y(a)$ and then successively step from a to b producing approximations y_n of $y(t_n)$ on a mesh $a = t_0 < t_1 < \ldots < t_n < \ldots$ At each (t_n, y_n) a recipe of the form

$$y_{n+1} = y_n + h\Phi(t_n, y_n, h)$$

is used to advance a step of length h to $t_{n+1} = t_n + h$. For such methods the stability of the differential equation is of fundamental importance. This is easy to understand: Suppose that the approximation $y_n \neq y(t_n)$. Considering the information used by the method, the best one can hope to achieve at future steps is to obtain exactly that solution $\tilde{y}(t)$ of (1.1) with $\tilde{y}(t_n) = y_n$. In this ideal situation the error at time $t > t_n$ is $y(t) - \tilde{y}(t)$. How fast integral curves of (1.1) can separate tells us, then, how much an isolated error can be amplified as the integration progresses. The stability of an effective numerical method must imitate that of the problem itself, at least for small step sizes, but there are, of course, new issues to be addressed. A stability result is an essential ingredient of a convergence proof because such a proof is basically a statement about the error made in a single step and how this error can be amplified at subsequent steps.

Traditionally it is assumed that the function f of (1.1) is continuous and satisfies a Lipschitz condition with constant L,

$$\|f(t, u) - f(t, v)\| \leq L\|u - v\|. \tag{1.3}$$

The stability of solutions of (1.1) in this class of problems is easily established. It will be sketched in a form analogous to developments which follow. Any solution $y(t)$ of (1.1) can be represented in the form

$$y(t) = y(t_n) + \int_{t_n}^{t} f(\theta, y(\theta))\, d\theta. \tag{1.4}$$

†This work performed at Sandia National Laboratories supported by the U.S. Department of Energy under Contract Number DE-AC04-76DP00789.

Suppose $\tilde{y}(t)$ is another solution. Let

$$\phi(t) = \|\tilde{y}(t) - y(t)\|.$$

Using the representation (1.4) and the Lipschitz condition (1.3), it is found easily that

$$\phi(t) \leqslant \phi(t_n) + \int_{t_n}^{t} L\phi(\theta)\,\mathrm{d}\theta.$$

Gronwall's inequality (see, e.g. [1], p. 252) then yields

$$\|\tilde{y}(t) - y(t)\| \leqslant \|\tilde{y}(t_n) - y(t_n)\| e^{L(t-t_n)}. \tag{1.5}$$

The scalar problem $y' = Ly$ shows that the bound (1.5) is sharp. When L is not large compared to the length of the interval of interest, all is well. The unhappy fact is that there are problems of great interest for which L is "large." These interesting problems are stable; the bound (1.5) just does not reveal the fact. The scalar problem $y' = -Ly$ is such an example as all its solution curves actually come together exponentially fast.

The difficulty with the traditional theory is that the class of Lipschitzian problems is too large. A theory allowing stable problems with "large" Lipschitz constants is needed. A traditional response has been to consider as a separate matter problems of the form

$$y' = Jy + g(t), \tag{1.6}$$

or the still more restricted form with the same stability properties,

$$y' = Jy.$$

The analysis of the behavior of formulas on stable problems of this kind with large Lipschitz constants is called the theory of absolute stability. One reason for considering such problems is that they are so simple that the analysis can be carried out in considerable detail. Another reason is that the general problem (1.1) can be approximated locally by a problem of the form (1.6). It is hoped that the behavior in the case of the simpler problem provide guidelines as to the behavior in the general case.

The traditional treatments sketched leave several things to be desired. The results for problems with large Lipschitz constants are restricted to very special problems. The relationship between the behavior of the methods for the general problem (1.1) and a local linearization of the form (1.6) is heuristic. The two theories are not only treated separately, they seem to have little in common. A theory is presented here which is a synthesis responding to these defects of the traditional approaches.

2. STABILITY OF THE PROBLEMS

In contrast to the form (1.1) we shall consider equations of the form

$$y' = Jy + g(t, y). \tag{2.1}$$

It is assumed that g satisfies a Lipschitz condition with constant l,

$$\|g(t, u) - g(t, v)\| \leqslant l\|u - v\|. \tag{2.2}$$

This form obviously includes the ones traditionally considered. Associated with (2.1) is the "model problem,"

$$y' = Jy. \tag{2.3}$$

Because the model problem is comparatively easy to analyze in detail, we shall view (2.1) as arising from the addition of a (possibly) nonlinear term to the equation (2.3). We shall continually ask what happens for the model problem and then ask to what degree the general problem shares this behavior.

For the sake of simplicity and definiteness, it is assumed that the matrix J can be diagonalized by a similarity transformation

$$MJM^{-1} = \Lambda = \operatorname{diag}\{\lambda_i\}. \tag{2.4}$$

It is then easy to analyze the stability of the model problem. There are two ways of going about this. A traditional way is to introduce the change of variables

$$My = w$$

so that (2.3) is equivalent to

$$w' = \Lambda w.$$

The analysis of these uncoupled equations is very easy and the results are subsequently translated into statements about y. This approach can be used to develop all our results, but we have preferred an equivalent approach based on a change of norm involving M.

It is assumed that the vector norm implicit in (2.2) is the Euclidean norm. The associated subordinate matrix norm will be used. Our analysis is based on another vector norm, the M-norm, defined by

$$\|v\|_M = \|Mv\|,$$

where M is the matrix appearing in (2.4). Its subordinate matrix norm is

$$\|R\|_M = \|MRM^{-1}\|.$$

Inequalities in one norm are connected to inequalities in the other by the condition number of M, $\kappa(M) = \|M\|\|M^{-1}\|$, viz.,

$$\|v\| \leqslant \|u\| \quad \text{implies} \quad \|v\|_M \leqslant \kappa(M)\|u\|_M,$$

$$\|v\|_M \leqslant \|u\|_M \quad \text{implies} \quad \|v\| \leqslant \kappa(M)\|u\|.$$

In particular, (2.2) implies

$$\|g(t,u) - g(t,v)\|_M \leqslant l\kappa(M)\|u - v\|_M. \tag{2.5}$$

With these preliminary observations it is now easy to obtain a stability result for the model problem.

LEMMA 1

Let $\tilde{y}(t)$, $y(t)$ be two solutions of (2.3). If J satisfies (2.4), then

$$\|\tilde{y}(t) - y(t)\|_M \leqslant \|\tilde{y}(t_n) - y(t_n)\|_M e^{\mu(t-t_n)} \tag{2.6}$$

where

$$\mu = \max_i \operatorname{Re}(\lambda_i). \tag{2.7}$$

Proof. Any solution $y(t)$ of (2.3) can be written as

$$y(t) = \exp((t - t_n)J)y(t_n)$$

from which it is obvious that

$$\|\tilde{y}(t) - y(t)\|_M \leqslant \|\exp((t - t_n)J)\|_M \|\tilde{y}(t_n) - y(t_n)\|_M.$$

At this point the roles played by the assumption (2.4) and the M-norm are seen:

$$\|\exp((t-t_n)J)\|_M = \|M \exp((t-t_n)J)M^{-1}\|$$

$$= \|\text{diag}\{\exp((t-t_n)\lambda_i)\}\| = e^{\mu(t-t_n)}.$$

The standard theory applied to the model problem (2.3) leads to (1.5) with

$$L = \|J\|_M = \|\text{diag}\{\lambda_i\}\| = \max|\lambda_i| = \rho(J).$$

The result (2.6) is advantageous when $\mu < \rho(J)$. A difference between the two approaches is made clear on considering the two scalar problems $y' = Ly$ and $y' = -Ly$. The classical theory does not distinguish these problems and Lemma 1 does.

According to (2.6), the integral curves of the model problem do not spread apart at all if $\mu \leqslant 0$ and actually come together if $\mu < 0$. More generally, an equation (2.1) is said to be contractive in the M-norm if

$$\|\tilde{y}(t) - y(t)\|_M < \|\tilde{y}(t_n) - y(t_n)\|_M \quad \text{for} \quad t > t_n$$

for all pairs of distinct solutions $\tilde{y}(t)$, $y(t)$. Because this favorable behavior can appear with L arbitrarily large, it is singled out for special study. We have the

LEMMA 2

If J satisfies (2.4), the model problem (2.3) is contractive in the M-norm if, and only if, $\mu < 0$.

Proof. One direction is an immediate consequence of (2.6). Suppose now that $\tilde{y}(t_n) - y(t_n) = v_i$ is an eigenvector of J corresponding to the eigenvalue λ_i. Then

$$\|\tilde{y}(t) - y(t)\|_M = \|v_i e^{(t-t_n)\lambda_i}\|_M = |e^{(t-t_n)\lambda_i}| \, \|\tilde{y}(t_n - y(t_n)\|_M.$$

Clearly these two solutions do not approach one another unless $Re(\lambda_i) < 0$ and we conclude that necessarily $\mu < 0$.

Now let us derive a stability result like (1.6) for the whole class (2.1). The classical variation of constants formula provides a generalization of (1.4) which exposes the roles of J and g:

$$y(t) = \exp((t - t_n)J)y(t_n) + \int_{t_n}^{t} \exp((t - \theta)J)g(\theta, y(\theta)) \, d\theta. \tag{2.8}$$

THEOREM 1

Let $\tilde{y}(t)$, $y(t)$ be two solutions of (2.1). If J satisfies (2.4) and g satisfies (2.2), then

$$\|\tilde{y}(t) - y(t)\|_M \leqslant \|\tilde{y}(t_n) - y(t_n)\|_M e^{(t-t_n)(\mu + l\kappa(M))}. \tag{2.9}$$

Proof. Let

$$\phi(t) = \|\tilde{y}(t) - y(t)\|_M.$$

Then the Lipschitz condition (2.5) and the representation (2.8) lead easily to

$$\phi(t) \leqslant \|\exp((t-t_n)J)\|_M \phi(t_n) + \int_{t_n}^{t} \|\exp((t-\theta)J)\|_M l\kappa(M)\phi(\theta)\,\mathrm{d}\theta.$$

On bounding the norms of the exponential matrices as in the proof of Lemma 1 and on introducing

$$\zeta(t) = \phi(t)\,\mathrm{e}^{-\mu t},$$

it is found that

$$\zeta(t) \leqslant \zeta(t_n) + \int_{t_n}^{t} l\kappa(M)\zeta(\theta)\,\mathrm{d}\theta.$$

Gronwall's inequality then states that

$$\zeta(t) \leqslant \zeta(t_n)\,\mathrm{e}^{(t-t_n)l\kappa(M)}$$

which is equivalent to the result desired.

COROLLARY 1

With the assumptions of Theorem 1, the problem (2.1) is contractive in the M-norm if

$$\mu + l\kappa(M) < 0.$$

Theorem 1 includes the classical result when $J = 0$. It is obviously more informative and is obtained at a price of an analysis only a little more complicated. The most interesting situation is when J has eigenvalues λ_i with $|\lambda_i|$ large and $Re(\lambda_i) \leqslant 0$ because we obtain then a much more realistic stability bound. We have an attractive formulation of contractivity: If the model problem is contractive in the M-norm, so is the general problem for all sufficiently small l. In any case, we see that for "small" l, the stability bound is much the same as that for the model problem.

3. EXPLICIT RUNGE–KUTTA METHODS

Now that we have explored the stability of equations of the form (2.1) we take up their numerical solution by explicit Runge–Kutta methods. We model the analysis of the stability of the numerical scheme after that of the equation itself and ask to what degree the methods imitate the behavior of the underlying problem.

First we define explicit Runge–Kutta formulas. A formula of $s+1$ stages advances the solution of (1.1) one step of length h from an approximation y_n of $y(t_n)$ to an approximation y_n of $y(t_{n+1})$, where $t_{n+1} = t_n + h$, by a recipe of the form:

$$\begin{aligned}
x_0 &= t_n \\
u_0 &= y_n \\
f_0 &= f(x_0, u_0) \\
&\text{for } j = 1, \ldots, s \\
x_j &= x_0 + \alpha_j h \\
u_j &= u_0 + h \sum_{k=0}^{j-1} \beta_{j,k} f_k \\
f_j &= f(x_j, u_j) \\
y_{n+1} &= u_{s+1} = u_0 + h \sum_{j=0}^{s} \gamma_j f_j.
\end{aligned} \tag{3.1}$$

Here the constants α_j, $\beta_{k,j}$, γ_j define the method.

It is supposed that the method is of order $r \geq 1$, meaning that for any sufficiently smooth f,

$$y_{n+1} = y(t_n + h) + 0(h^{r+1}). \tag{3.2}$$

A result useful later is that

$$\sum_{j=0}^{s} \gamma_j = 1 \tag{3.3}$$

which follows from (3.2) and consideration of the particular problem $f(t, y) \equiv 1$.

Notice that each u_k is itself the result of an explicit Runge–Kutta formula for approximating $y(x_k) = y(t_n + \alpha_k h)$. These formulas are generally of low order, and it is only the last, u_{s+1}, which is constructed to be of a high order of accuracy.

It is the form (2.1) of f which concerns us. For our investigation we require a representation of y_{n+1} which exposes the roles of J and g in a manner analogous to the variation of constants formula (2.8). The next theorem provides this tool.

THEOREM 2

When a Runge–Kutta method (3.1) is applied to an equation of the form (2.1), it has the representation

$$u_j = P_j(hJ)u_0 + \sum_{k=0}^{j-1} p_{j,k}(hJ)g(x_k, u_k) \quad j = 1, \ldots, s$$

$$y_{n+1} = u_{s+1} = P(hJ)u_0 + h \sum_{k=0}^{s} p_{s+1,k}(hJ)g(x_k, u_k) \tag{3.4}$$

where the polynomials $P_j(z)$, $P(z)$, and $p_{j,k}(z)$ are defined by

$$P_j(z) = 1 + z \sum_{k=0}^{j-1} \beta_{j,k} P_k(z) \quad j = 0, \ldots, s$$

$$P(z) = 1 + z \sum_{k=0}^{s} \gamma_j P_j(z)$$

$$p_{j,k}(z) = \beta_{j,k} + z \sum_{l=k+1}^{j-1} \beta_{j,l} p_{l,k}(z) \quad 0 \leq k < j \leq s$$

$$p_{s+1,k}(z) = \gamma_k + z \sum_{l=k+1}^{s} \gamma_l p_{l,k}(z) \quad k = 0, \ldots, s$$

and empty sums are interpreted as zero.

Proof. Because verification of the stated relationships is straightforward, the details are omitted.

To fully appreciate the analogy between (3.4) and (2.8), some observations are necessary. The representation shows that when applied to the model problem (2.3),

$$u_j = P_j(hJ)u_0 = P_j(hJ)y_n$$

$$y_{n+1} = u_{s+1} = P(hJ)u_0 = P(hJ)y_n. \tag{3.5}$$

From this we recognize that $P(z)$ is the familiar stability polynomial of the method and that the $P_j(z)$ are the (less) familiar internal stability polynomials. The order condition (3.2)

and (3.5) imply that

$$P(hJ) = \exp(hJ) + 0(h^{r+1}).$$

Recall that $u_k \doteq y(t_n + \alpha_k h)$. Finally, a little manipulation of (2.8) gives the equivalent form

$$y(t_{n+1}) = \exp(hJ)y(t_n) + h\int_0^1 \exp((1-\tau)hJ)g(t_n + \tau h, y(t_n + \tau h))\,d\tau.$$

which is closely analogous to (3.4).

How well the Runge–Kutta method approximates the solution of the model problem is essentially a question of how well the stability polynomial $P(z)$ approximates $\exp(z)$. The order condition says that the approximation is good for $|z|$ "small." In considering the situation for other z, let us first suppose that all eigenvalues of J lie in the left half complex plane, i.e. $\mu \leqslant 0$. According to Lemma 1, the integral curves do not spread apart in the M-norm when $\mu \leqslant 0$ and actually approach when $\mu < 0$. Behavior like this is not possible for an explicit Runge–Kutta method for $|z|$ large because $|P(z)| \to \infty$ as $|z| \to \infty$. Useful results can be obtained if we do not seek to imitate the contractivity at an exponential rate, rather merely to imitate the qualitative property

$$\|\tilde{y}(t) - t(t)\|_M \leqslant \|\tilde{y}(t_n) - y(t_n)\|_M \quad \text{for} \quad t \geqslant t_n,$$

and realize that we cannot achieve even this for all z with $Re(z) \leqslant 0$.

For the statement of the next lemma we need the concept of the stability region S of the method. This (compact) region in the complex plane is defined by

$$S = \{z | Re(z) \leqslant 0 \quad \text{and} \quad |P(z)| \leqslant 1\},$$

where $P(z)$ is the stability polynomial of the method. Notice that the origin belongs to S because $P(0) = 1$ and that S is finite.

LEMMA 3

Let $\tilde{y}_{n+1}$, y_{n+1} be the results of applying the explicit Runge–Kutta method (3.1) to the model problem (2.3) starting with the approximations $\tilde{y}_n$, y_n, respectively. Suppose that J satisfies (2.4) and that $\mu \leqslant 0$. Then

$$\|\tilde{y}_{n+1} - y_{n+1}\|_M \leqslant \|\tilde{y}_n - y_n\|_M$$

for all $\tilde{y}_n$, y_n if, and only if, $h\lambda_i \in S$ for all eigenvalues of J.

Proof. From (3.5)

$$\|\tilde{y}_{n+1} = y_{n+1}\|_M = \|P(hJ)(\tilde{y}_n - y_n)\|_M \leqslant \|P(hJ)\|_M \|\tilde{y}_n - y_n\|_M.$$

Now

$$\|P(hJ)\|_M = \|\text{diag}\,\{P(h\lambda_i)\}\| = \max |P(h\lambda_i)| \leqslant 1$$

if each $h\lambda_i \in S$. This proves the sufficient part. If $\tilde{y}_n - y_n$ is an eigenvector v_i of J corresponding to the eigenvalue λ_i, then

$$\|\tilde{y}_{n+1} - y_{n+1}\|_M = \|P(h\lambda_i)(\tilde{y}_n - y_n)\|_M = |P(h\lambda_i)|\,\|\tilde{y}_n - y_n\|_M$$

from which it is seen that $|P(h\lambda_i)| \leqslant 1$ is necessary.

Clearly, if we wish to obtain a stability result for the numerical method analogous to the stability of the general problem (2.1) when $\mu \leqslant 0$, we must require that $h\lambda_i \in S$. Given this necessary requirement we can derive a very similar result:

THEOREM 3

Let $\{\tilde{u}_j\}$, $\tilde{y}_{n+1}$ and $\{u_j\}$, y_{n+1} be the results of applying the explicit Runge–Kutta method (3.1) to the equation (2.1) starting with the approximations $\tilde{y}_n$ and y_n, respectively. Suppose that J satisfies (2.4) and g satisfies (2.2). Suppose that $Re(\lambda_i) \leq 0$ for all eigenvalues λ_i of J and that $Re(\lambda_i) = 0$ only if $\lambda_i = 0$. Suppose further that $h\lambda_i \in S$ for all λ_i. Then

$$\|\tilde{u}_j - u_j\|_M \leq (P_j^* + hL_j)\|\tilde{u}_0 - u_0\|_M \quad j = 1, \ldots, s$$

$$\|\tilde{y}_{n+1} - y_{n+1}\|_M \leq (P^* + hL)\|\tilde{y}_n - y_n\|_M \tag{3.6}$$

$$\|\tilde{y}_{n+1} - y_{n+1}\|_M \leq (1 + hl\kappa(M)\omega(h))\|\tilde{y}_n - y_n\|_M \tag{3.7}$$

where

$$p_{j,k}^* = \max_{z \in S} |p_{j,k}(z)| \quad 0 \leq k < j \leq s+1$$

$$P_j^* = \max_{z \in S} |P_j(z)| \quad 0 \leq j \leq s$$

$$P^* = \max_{z \in S} |P(z)| = 1$$

$$L_j = l\kappa(M) \sum_{k=0}^{j-1} p_{j,k}^*(P_k^* + hL_k) \quad 0 \leq j \leq s$$

$$L = l\kappa(M)\omega(h)$$

$$\omega(h) = \sum_{k=0}^{s} p_{s+1,k}^*(P_k^* + hL_k)$$

and the $p_{j,k}(z)$, $P_j(z)$, $P(z)$ are defined in Theorem 2.

Proof. From the representation of Theorem 2,

$$\tilde{u}_1 - u_1 = P_1(hJ)(\tilde{u}_0 - u_0) + hp_{1,0}(hJ)[g(x_0, \tilde{u}_0) - g(x_0, u_0)].$$

Using the Lipschitz condition (2.5) satisfied by g, one finds

$$\|\tilde{u}_1 - u_1\|_M \leq (\|P_1(hJ)\|_M + hl\kappa(M)\|p_{1,0}(hJ)\|_M)\|\tilde{u}_0 - u_0\|_M.$$

Then

$$\|P_1(hJ)\|_M = \|\text{diag}\,\{P_1(h\lambda_i)\}\| = \max_i |P_1(h\lambda_i)| \leq P_1^*$$

and similarly,

$$\|p_{1,0}(hJ)\|_M \leq p_{1,0}^*.$$

Because $P_0 \equiv 1$, $L_0 = 0$, the conclusion of the theorem is seen to hold for $j = 1$.

Suppose now that the statement holds through index $j - 1$. Then

$$\tilde{u}_j - u_j = P_j(hJ)(\tilde{u}_0 - u_0) + h \sum_{k=0}^{j-1} p_{j,k}(hJ)[g(x_k, \tilde{u}_k) - g(x_k, u_k)],$$

hence

$$\|\tilde{u}_j - u_j\|_M \leqslant \|P_j(hJ)\|_M \|\tilde{u}_0 - u_0\|_M + h\sum_{k=0}^{j-1} \|p_{j,k}(hJ)\|_M lK(M)\|\tilde{u}_k - u_k\|_M$$

$$\leqslant (P_j^* + hlK(M)\sum_{k=0}^{j-1} p_{j,k}^* (P_k^* + hL_k))\|\tilde{u}_0 - u_0\|_M$$

$$\leqslant (P_j^* + hL_j)\|\tilde{u}_0 - u_0\|_M.$$

The result for y_{n+1} is verified in the same way.

It is worth remarking that the stability result (3.7) can be translated into a result in the Euclidean norm, viz.,

$$\|\tilde{y}_{n+1} - y_{n+1}\| \leqslant \kappa(M)(1 + hlK(M))\|\tilde{y}_n - y_n\|,$$

or, indeed, into any other norm. However, if one is interested in the stability after a number of steps, the result should first be obtained in the M-norm and then translated so that powers of the condition number do not appear. The next corollary provides such a result.

COROLLARY 3.1

Let $\{\tilde{y}_n\}$, $\{y_n\}$ be the result of n steps, each of length h, of the explicit Runge–Kutta method (3.1) applied to (2.1) with the initial values $\tilde{y}_0$, y_0, respectively. If the hypotheses of the theorem hold, then

$$\|\tilde{y}_n - y_n\|_M \leqslant \|\tilde{y}_0 - y_0\|_M \, e^{(t_n - t_0)lK(M)\omega(h)}. \tag{3.8}$$

Proof. The result follows immediately from (3.7) and the inequality

$$1 + x \leqslant e^x \quad \text{for} \quad x \geqslant 0.$$

Evidently the stability bound for the numerical method has quite a lot of resemblance to that for the problem itself. The role of the method appears in the factor $\omega(h)$ and in the requirement that the $h\lambda_i$ be in the stability region of the method. This is a considerably sharper result than the classical one because the magnitudes of the eigenvalues do not appear in the bound. Of course they do play a role in that h must be small enough that the $h\lambda_i$ be in the stability region.

For convergence results we need a stability bound uniform in h. This is available because $\omega(h)$ is obviously a monotonely increasing function of h. Notice that the internal stability functions make their appearance indirectly via this function. A lower bound for $\omega(h)$ is easily obtained from $\omega(h) \geqslant \omega(0)$ where

$$\omega(0) = \sum_{k=0}^{s} p_{s+1,k}^* P_k^*.$$

Because $0 \in S$,

$$P_k^* \geqslant |P_k(0)| = 1$$

$$p_{s+1,k}^* \geqslant |p_{s+1,k}(0)| = |\gamma_k|.$$

Then, on using (3.3),

$$\omega(0) \geqslant \sum_{k=0}^{s} |\gamma_k| \geqslant \sum_{k=0}^{s} \gamma_k = 1.$$

Because $\omega(h) \geqslant 1$, we see that, according to the bounds, the numerical method is less stable than the equation itself.

For some of our results, and certainly for a convergence theory, we are interested in "small" h. In this connection we note that

$$\omega(h) = \omega(0) + 0(h).$$

The next lemma says that for small h, the hypothesis about the stability region in the stability bound falls away.

LEMMA 4

Suppose that all the eigenvalues λ_i of the matrix J are such that $Re(\lambda_i) \leqslant 0$ and $Re(\lambda_i) = 0$ only if $\lambda_i = 0$. If the method (3.1) has order $r \geqslant 1$, then for all sufficiently small h, all the $h\lambda_i \in S$.

Proof. Because of the order condition

$$P(z) = 1 + z + 0(|z|^2),$$

from which one readily finds that

$$|P(h\lambda_i)| = 1 + h\mathrm{Re}(\lambda_i) + 0(h^2).$$

If $\mathrm{Re}(\lambda_i) < 0$, then obviously $|P(h\lambda_i)| < 1$ for all sufficiently small h. If $\mathrm{Re}(\lambda_i) = 0$, we use the fact that $0 \in S$.

According to this lemma, the hypothesis that $h\lambda_i \in S$ of the theorem is satisfied for all sufficiently small h. It must be appreciated, however, that the step size might have to be very small. It is easy enough to determine a bound explicitly in the case of the forward Euler method

$$y_{n+1} = y_n + hf(t_n, y_n).$$

Obviously

$$S = \{z \mid |z + 1| \leqslant 1\}.$$

If we let

$$\zeta = \max_i |\mathrm{Im}(\lambda_i)/\mathrm{Re}(\lambda_i)|,$$

it is necessary that

$$0 < h \leqslant 2/(1 + \zeta^2).$$

The closer λ_i is to the imaginary axis, the smaller h must be to get $h\lambda_i$ into the stability region.

Let us now investigate the contractivity of the numerical method. Theorem 3 cannot be used for this purpose because we allowed $\mu = 0$, hence problems which are not contractive, and also took $P^* = 1$. It is easy enough to sharpen the result to

THEOREM 4

In addition to the assumptions of Theorem 3, suppose that

$$\mu + l\kappa(M)\omega(0) < 0. \tag{3.9}$$

Then for all sufficiently small h,

$$\|\tilde{y}_{n+1} - y_{n+1}\|_M < \|\tilde{y}_n - y_n\|_M.$$

Proof. Examination of the proof of Theorem 3 shows tht it was not necessary to take $P^* = 1$; any number P^* such that

$$P^* \geqslant |P(h\lambda_i)|$$

suffices. As in the proof of Lemma 4, each

$$|P(h\lambda_i)| \leqslant 1 + hRe(\lambda_i) + 0(h^2)$$

$$\leqslant 1 + h\mu + 0(h^2).$$

Thus with a suitable definition of P^*, the factor in the bound (3.7) can be sharpened to

$$1 + h\mu + hl\kappa(M)\omega(0) + 0(h^2)$$

whence the conclusion of the theorem is obvious.

Along with other results we have derived, this gives an attractive description of the special situation of contractivity: If the model problem (2.3) is contractive in the M-norm, then for all l small enough that (3.9) hold, the general problem (2.1) is also contractive. Furthermore, for all sufficiently small step sizes the numerical solution is also contractive. One should not get too enthusiastic about this result because the step size might have to be very small. Explicit Runge–Kutta methods simply do not have stability regions large enough to make them very interesting in this context. A more realistic variation of the last result is

THEOREM 5

In addition to the assumptions of Theorem 3, suppose that $\mu < 0$ and that all $h\lambda_i$ are in the interior of S for a given h. Then for this h,

$$\|\tilde{y}_{n+1} - y_{n+1}\|_M < \|\tilde{y}_n - y_n\|_M$$

if l is sufficiently small.

Proof. The condition on the $h\lambda_i$ implies that

$$P^* = \max |P(h\lambda_i)| < 1.$$

Then for l sufficiently small

$$P^* + l\kappa(M)\omega(h) < 1$$

and the result follows as in the proof of Theorem 4.

This result says that if the numerical method is contractive in the M-norm for the model problem, then it is also contractive for the general problem provided that l is not too large.

Now let us consider a general distribution of eigenvalues of J and, in particular, allow positive real parts. The bound (2.6) for the model problem says that the stability of the problem depends only on the real parts of the eigenvalues of J. Just as we saw earlier, the magnitude also plays a role in the stability of an explicit Runge–Kutta method. A useful result depends on the step size being small enough that the numerical approximation be accurate. This is not different from the treatment of the nonlinear term $g(t, y)$ in (2.1). With this insight as to what is reasonable, we can reduce the general case to a previous theorem by a judicious decomposition.

We need to collect those eigenvalues of J with $\mathrm{Re}(\lambda_i)\geq 0$ and it turns out to be convenient to deal with certain eigenvalues with $\mathrm{Re}(\lambda_i)<0$ at the same time. Let us define

$$\tau = \max_{\mathrm{Re}(\lambda_i)\geq 0} |\lambda_i|.$$

Of course, any λ_i with $\lambda_i > \tau$ has $\mathrm{Re}(\lambda_i)<0$.

Recalling that

$$MJM^{-1} = \Lambda = \mathrm{diag}\,\{\lambda_i\},$$

define the diagonal matrix Λ_2 by

$$(\Lambda_2)_{i,i} = \lambda_i \quad \text{if} \quad |\lambda_i| \leq \tau,$$

$$= 0 \quad \text{otherwise,}$$

and the diagonal matrix Λ_1 by

$$\Lambda_1 = \Lambda - \Lambda_2.$$

Then

$$J = J_1 + J_2$$

where

$$J_1 = M^{-1}\Lambda_1 M, \quad J_2 = M^{-1}\Lambda_2 M.$$

We rewrite (2.1) as

$$y' = Jy + g(t, y) = J_1 y + G(t, y)$$

where

$$G(t, y) = J_2 y + g(t, y).$$

Notice that

$$\|G(t, y) - G(t, v)\|_M \leq (\|J_2\|_M + l)\|u - v\|_M$$

and

$$\|J_2\|_M = \|\Lambda_2\| = \tau.$$

Now we can apply Theorem 3 to this modified problem to obtain

THEOREM 6

Let $\{\tilde{u}_j\}$, $\tilde{y}_{n+1}$ and $\{u_j\}$, y_{n+1} be the results of applying the explicit Runge–Kutta method (3.1) to the equation (2.1) starting with the approximations $\tilde{y}_n$ and y_n, respectively. Suppose that J satisfies (2.4) and g satisfies (2.2). Suppose that $h\lambda_i \in S$ for all eigenvalues λ_i such that $|\lambda_i| > \tau$. Then

$$\|\tilde{y}_{n+1} - y_{n+1}\|_M \leq (1 + h(\tau + l)\kappa(M)\omega(h))\|\tilde{y}_n - y_n\|_M.$$

Let $\{\tilde{y}_n\}$, $\{y_n\}$ be the result of n steps, each of length h, starting with the approximations $\tilde{y}_0$, y_0, respectively. Then

$$\|\tilde{y}_n - y_n\|_M \leq \|\tilde{y}_0 - y_0\|_M e^{(t_n - t_0)(\tau + l)\kappa(M)\omega(h)}.$$

This last result makes the crucial role of the distribution of eigenvalues clear. In general

$$f(t, y) = Jy + g(t, y)$$

satisfies a Lipschitz condition in the M-norm with constant

$$L = (\|J\|_M + l)\kappa(M) = (\rho(J) + l)\kappa(M).$$

If $\tau = \rho(J)$, we gain nothing over the classical theory. The big gain is when $\tau \ll \rho(J)$, i.e., when eigenvalues λ_i with $\mathrm{Re}(\lambda_i) \geq 0$ are small in magnitude compared to some eigenvalue λ_j with $\mathrm{Re}(\lambda_j) < 0$.

4. LINEAR STABILITY

Let us suppose that the differential equation (1.1) is in autonomous form

$$y' = f(y), \tag{4.1}$$

which can, if necessary, be achieved by introducing an additional dependent variable. The linear stability theory is based on the fact that for y near $y(t_0)$,

$$f(y) = f(y(t_0)) + f_y(y(t_0))(y - y(t_0)) + 0(\|y - y(t_0)\|^2).$$

This suggests that the solutions of (4.1) will behave much like the solutions of

$$y' = Jy + f(y(t_0)) - Jy(t_0), \tag{4.2}$$

where we define

$$J = f_y(y(t_0)), \tag{4.3}$$

and that the stability of (4.1) will be much like that of (4.2) and the model problem

$$y' = Jy. \tag{4.4}$$

It is hoped that in addition, the stability of Runge–Kutta methods will be much the same for (4.1) and (4.4).

Justification of this approach requires that $\|y - y(t_0)\|$ be "small." This is true for t near t_0, but not interesting because this situation corresponds to small step sizes and the classical theory is applicable. The approach is interesting only when the solution of (4.1) does not change rapidly near t_0—a matter often not fully appreciated. In our approach, we confine out attention to a ball about $y(t_0)$, namely, to those w such that

$$\|w - y_0\| \leq \zeta.$$

We define $g(y)$ as

$$g(y) = f(y) - Jy$$

so that

$$y' = f(y) = Jy + g(y). \tag{4.5}$$

This is *not* an approximation like (4.2); it is an exact decomposition of (4.1). Assuming f is sufficiently smooth, g satisfies a Lipschitz condition: Using a mean value theorem ([2], p. 70)

$$\begin{aligned}\|g(u) - g(v)\| &= \|f(u) - f(v) - f_y(y(t_0))(u - v)\| \\ &\leq \sup_{0 \leq t \leq 1} \|f_y(v + t(u - v)) - f_y(y(t_0))\| \cdot \|u - v\|,\end{aligned}$$

so if we define

$$l(\zeta) = \sup_{\|w - y(t_0)\| \leq \zeta} \|f_y(w) - f_y(y(t_0))\|,$$

we have

$$\|g(u) - g(v)\| \leq l(\zeta)\|u - v\|.$$

If we suppose that the local Jacobian (4.3) can be diagonalized, the theory developed in the last section can be applied to (4.5). A key point is that the Lipschitz constant $l(\zeta) \to 0$ as $\zeta \to 0$. Our stability theory, roughly speaking, says that the stability of the problem (4.1) and of the Runge–Kutta method applied to (4.1) are essentially the same as for the model problem (4.4) when l is small enough. By confining our attention to a sufficiently small neighborhood of $y(t_0)$, l will be small enough. Again, this is interesting only when the solution $y(t)$ does not change rapidly near $y(t_0)$. In this way we obtain a theoretical justification for the application of local linearization to the study of the stability of Runge–Kutta methods.

5. ACKNOWLEDGEMENTS

In a semi-expository paper like this one, it is not appropriate to provide references for well-known results that are easily proved in the text. We do, however, want to acknowledge some intellectual debts. Exploiting the form (2.1) is basic to treatments of stiff problems, but even for the non-stiff problems which concern us, the idea is old. The landmark paper[3] of Dahlquist is by no means the first to use the form for the investigation of stability, but it was a major stimulus for our own work. Recently Hairer, Bader and Lubich[4] have made important use of the form in analyzing the stability of semi-implicit formulas for stiff problems. A basic tool of their analysis is a variation of constants formula of which Theorem 2 is a special case; indeed, we followed their notation in this matter as closely as we could. In large measure the present work is the result of asking what happens when explicit Runge–Kutta formulas are considered instead of semi-implicit formulas. Because the formulas have entirely different qualitative properties, we had to consider different classes of differential equations, investigate $h \to 0$ instead of h fixed, study a different kind of stability, and employ different analytical techniques. We mention these details because the approach of [4] is not applicable to semi-implicit formulas with restricted linear stability. The formulas arising from extrapolation of the semi-explicit midpoint rule are important examples in [4]. Unfortunately, many cannot be treated. It seems to this author that a reassessment of the matter with assumptions and questions more in the style of the present work might well be fruitful.

REFERENCES

1. F. Brauer and J. A. Nohel, *Ordinary Differential Equations*, W. A. Benjamin, Inc., New York (1967).
2. J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York (1970).
3. G. Dahlquist, Stability and error bounds in the numerical integration of ordinary differential equations, *Trans. Royal Inst. Tech., Stockholm*, No. 130 (1959).
4. E. Hairer, G. Bader and Ch. Lubich, On the stability of semi-implicit methods for ordinary differential equations, *BIT* **22**, 211–232 (1982).