NASA TECHNICAL NOTE

NASA TN D-4703

STABLE IMPLICIT AND EXPLICIT NUMERICAL METHODS FOR INTEGRATING QUASI-LINEAR DIFFERENTIAL EQUATIONS WITH PARASITIC-STIFF AND PARASITIC-SADDLE EIGENVALUES

by Harvard Lomax

Ames Research Center
Moffett Field, Calif.

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION • WASHINGTON, D. C. • JULY 1968

STABLE IMPLICIT AND EXPLICIT NUMERICAL METHODS FOR INTEGRATING QUASI-LINEAR DIFFERENTIAL EQUATIONS WITH PARASITIC-STIFF AND PARASITIC-SADDLE EIGENVALUES

By Harvard Lomax

Ames Research Center
Moffett Field, Calif.

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

For sale by the Clearinghouse for Federal Scientific and Technical Information
Springfield, Virginia 22151 – CFSTI price $3.00

STABLE IMPLICIT AND EXPLICIT NUMERICAL METHODS FOR INTEGRATING QUASI-LINEAR DIFFERENTIAL EQUATIONS WITH PARASITIC-STIFF AND PARASITIC-SADDLE EIGENVALUES

By Harvard Lomax

Ames Research Center

SUMMARY

Certain classes of coupled, quasi-linear, ordinary, differential equations contain eigenvalues in their associated matrix which make them difficult to integrate by means of conventional numerical differencing schemes, even when the solutions are continuous and nonsingular. Two classes of such "parasitic" eigenvalues are defined and general ways in which their effects can be suppressed are discussed.

INTRODUCTION

Special classes of differential equations (both ordinary and partial), termed quasi-linear, are defined by the condition that the highest order derivative terms appear explicitly and to the first power only. For ordinary differential equations these can be written

$$[B]\vec{w}' = \vec{C} \tag{1}$$

where the elements of $[B]$ and $\vec{C}$ can depend upon $\vec{w}$ and upon the independent variable t. This equation can be written formally

$$\vec{w}' = [B]^{-1}\vec{C} = \vec{F}(\vec{w},t) \tag{2}$$

Cases in which $[B]^{-1}$ does not exist ($\det(B) = 0$) are of special interest and lead to the study of critical points, and in particular, for the discussion in this paper, to the study of saddle points.

We are principally concerned with the practical situation when $\vec{F}$ has a nonlinear dependence upon $\vec{w}$. However, we assume this dependence is continuous at least through the first derivative. In such cases, we can make a local Taylor series expansion of $\vec{F}$ about some reference point n, where $t = nh$ and h is a small (step) interval. Equation (2) then becomes

$$\vec{w}' = [A_n]\vec{w} + \vec{f}_n + h^2\vec{er}_t \tag{3a}$$

where $[A_n]$ is the Jacobian of $\vec{F}$ with respect to $\vec{w}$ and

$$\vec{f}_n = \vec{F}_n - [A_n]\vec{w}_n \tag{3b}$$

and $\vec{er}_t$ is bounded as $h \to 0$.

If the term $h^2\vec{er}_t$ is neglected in the interval $nh \le t \le h(n+1)$, the result is, in that step, a set of coupled, ordinary, linear differential equations. Further, if equations (2) are autonomous (i.e., $\vec{F}$ does not depend explicitly on t), they are linear equations with constant coefficients. In carrying out practical numerical calculations, the elements of $\vec{F}$ (or $[A_n]$ and $\vec{f}_n$, if that form is used) are varied from step to step. This "captures" the nonlinear effects by embedding local polynomials of $O(h^2)$ in the calculations as they proceed. If equations (3) are used, $\vec{w}$ is approximated to $O(h^3)$. Further, if in each successive step a method is chosen that is stable for the local linearized form in that step, overall or "global" stability is assured in the sense described below.

There are two common approaches to the numerical integration of differential equations. One is to use explicit differencing formulas and apply them directly to equations (2), without ever actually formulating equations (3). These methods can have unlimited accuracy (since the local linearization is never carried out) and, in the author's experience, have stability properties that correlate extremely well with the eigenvalues of $[A_n]$ (as if the local linearization actually had been performed). However, all explicit methods have a finite (and, practically speaking, rather limited) stability boundary. The optimization of explicit methods, from the point of view of stability, is treated in reference 1, and briefly reviewed in the section Highly Stable Explicit Methods.

Another approach is to use implicit methods. A general discussion of this approach is the intended contribution of this paper. In the implicit case, equations (2) are in fact put in the form of equations (3) and treated as coupled linear equations in a single step. The elements of $[A_n]$ must be reevaluated at each step, although it is consistent with the accuracy involved to do this numerically rather than analytically. For example, use of the formula

$$a_{ij} \equiv \frac{\partial F_i}{\partial w_j} \approx \frac{F_i(1.01\ w_j) - F_i(0.99\ w_j)}{0.02\ w_j}$$

has given good practical results. Since calculating all the a_{ij} for each step can be quite time-consuming, it is advisable to know when it is and when it is not worthwhile. This leads directly to the subject discussed next.

GENERAL COMMENTS ON THE CONSTRUCTION OF NUMERICAL METHODS FOR INTEGRATING ORDINARY DIFFERENTIAL EQUATIONS

It is unlikely that "new" combinations of linear equations that connect a function, u, and its derivative, u', at a series of reference points,

equispaced or not, will improve existing methods for the numerical integration of general sets of coupled ordinary differential equations having the form given by equations (2). If nothing special is known a priori about the differential equations, the standard fourth order Runge-Kutta method is probably the "best." It is self-starting, has low storage capabilities, is easy to program, has good accuracy $O(h^5)$, and, as we shall presently see, is more stable than any of the standard predictor-corrector processes (e.g., Hamming's, Adams-Moulton, etc.).

Nevertheless, it is still popular to publish numerical integration methods for general classes of ordinary differential equations (see, e.g., refs. 2-5). These methods are meant to compete with, or improve upon, the classical ones mentioned above. Under what qualifications they do, or do not, is a subject discussed in reference 6. For the most part, their development seems to be motivated by a popular misconception that one can "break through" the Dahlquist stability "barrier." Now the Dahlquist stability theorem is the result of a rigorous proof (see ref. 7) based upon certain premises. One of these premises is that the function and its derivative are evaluated at, and only at, a series of equispaced points. Another is that a predictor is followed by only one corrector. Under these conditions Dahlquist shows that no linear stable equations relate u_{n+i} and u'_{n+i}, $i = 1, 2, \cdots, k$ (i.e., k-step methods) with an accuracy of order greater than $k + 2$, if k is even, or $k + 1$, if k is odd. But if the conditions described above are violated in any way, the theorem no longer applies.

In order to avoid the confusion and ambiguity that can arise when we try to classify explicit methods according to (computational) step size and step number, the representative step size is introduced such that:

$$H \equiv \text{the distance a solution is advanced after two evaluations of the derivative.} \tag{4}$$

If both the accuracy and stability of a method are referenced to this parameter, truly significant comparisons can be made amongst all types of numerical integration procedures involving linear connections between the function and its derivative. Let us next consider two examples.

The standard, fourth order, Runge-Kutta method is usually referred to as a one-step method. When it is cast in predictor-corrector terminology, it can be written

$$u^{(1)}_{n+\frac{1}{2}} = u_n + \left(\frac{h}{2}\right) u'_n \tag{5a}$$

$$u^{(2)}_{n+\frac{1}{2}} = u_n + \left(\frac{h}{2}\right) u^{(1)'}_{n+\frac{1}{2}} \tag{5b}$$

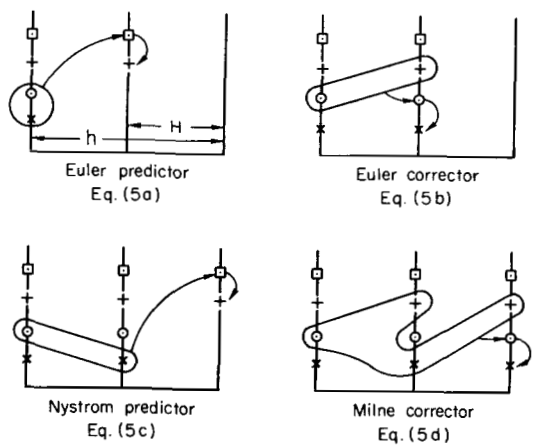
$$u^{(3)}_{n+1} = u_n + 2\left(\frac{h}{2}\right) u^{(2)'}_{n+\frac{1}{2}} \tag{5c}$$

$$u_{n+1} = u_n + \frac{1}{3}\left(\frac{h}{2}\right)\left(u^{(3)'}_{n+1} + 2u^{(2)'}_{n+\frac{1}{2}} + 2u^{(1)'}_{n+\frac{1}{2}} + u'_n\right) \tag{5d}$$

Graphically (see definition of symbols in table I) it appears as shown in sketch (a). From the sketch we see that the reference step size, H, is half that usually used in the computations. The example shows clearly that

TABLE I.- DEFINITIONS OF SYMBOLS USED IN SKETCHES

Symbol	Represents
□	Value of function predicted from data at previous steps.
+	Derivative calculated using □ .
○	1. Value of corrected function using data at previous steps and predicted data at this step, or 2. Value of function calculated using implicit method.
×	Derivative calculated using ○ .



Sketch (a).- Standard, fourth-order, Runge-Kutta method.

references to accuracy and stability on the basis of a calculation step size, h, and step number, k, (for eqs. (5) is k equal to 2 or 1?) are untrustworthy since the step designation is arbitrary. On the basis of H, equations (5) have the error terms

$$er_\lambda = \frac{1}{15}(\sigma H)^5 \tag{6a}$$

$$er_\mu = -\frac{\mu^2 H^5}{180}(\mu^3 - 5\sigma\mu^2 + 10\sigma^2\mu - 30\sigma^3) \tag{6b}$$

for the complementary and particular solutions (respectively) to the representative equation

$$u' = \sigma u + e^{\mu t} \tag{7}$$

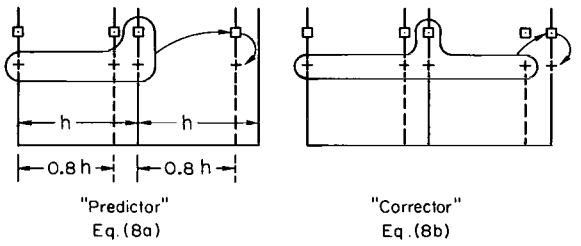
There is, generally, no connection between step number, k, and the accuracy and stability of a method. This has been discovered by various authors (see refs. 2-5), and has led to the introduction of terms such as "hybrid" and "combined" methods, terms which indicate (simply) that the premises required for the validity of Dahlquist's theorem have been abandoned.

Typical of the hybrid or combined methods (for a complete analysis see ref. 6, pp. 93-95) is

$$u^{(1)}_{n+1.8} = u_{n+1} + \frac{h}{75}\left(268u'_{n+1} + 22u'_n - 230u^{(1)'}_{n+0.8}\right) \tag{8a}$$

$$u_{n+2} = u_{n+1} + \frac{h}{48}\left(47u'_{n+1} + u'_n + 25u^{(1)'}_{n+1.8} - 25u^{(1)'}_{n+0.8}\right) \tag{8b}$$

which has the graphical representation shown in sketch (b). Its error, when applied to equation (7), is given by



Sketch (b).- A typical "hybrid" or "combined" method.

$$er_\lambda = \frac{1}{720}(\sigma H)^5 \tag{9a}$$

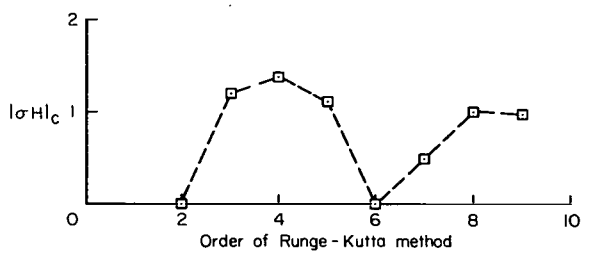
$$er_\mu = \frac{1}{720}(\mu H)^5 \tag{9b}$$

Clearly, equations (9) are much more accurate (when σH is small enough for the first missed term in a Taylor series expansion to give a valid approximation of the total error) than equations (5); but they pay the usual price: they are less stable.

Now the stability of a method depends upon the roots to its characteristic equation when that method is used to difference equation (7). The stability of Runge-Kutta methods is analyzed in many places (see, e.g., ref. 6). The characteristic polynomial for equations (5) (in terms of the displacement operator $E \equiv e^{h(d/dx)}$) is

$$P(E) = E - 1 - \sigma h - \frac{1}{2}(\sigma h)^2 - \frac{1}{6}(\sigma h)^3 - \frac{1}{24}(\sigma h)^4 \tag{10}$$

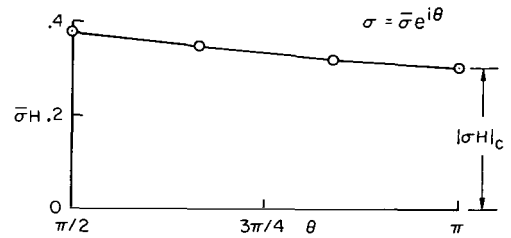
It has only one root (the principal one) and this root represents $e^{\sigma h}$ exactly through the fourth order. The P(E) for the ith order Runge-Kutta method represents $e^{\sigma h}$ exactly through the ith order, at which term it ends. The detailed behavior of these roots, as σh ranges from zero through a variety of complex values, is shown on page 83 of reference 6 for the second, third, and fourth order methods. The stability boundary $|\sigma H|_c$ (defined precisely in the next section) is shown in sketch (c). From this sketch we see that the fourth order Runge-Kutta method is the best of all Runge-Kutta methods (i.e., minimizes machine computing time) when applied to coupled ordinary differential equations with parasitic (stiff) eigenvalues (a term defined precisely in the next section).



Sketch (c).- General stability boundary of Runge-Kutta methods.

The characteristic equation for the method defined by equations (8) is a cubic, so that it contains two spurious roots. However, both of these roots go to zero when h goes to zero. Because of this the method is said to have "Adams-Moulton stability" (this

property is characteristic of all Adams-Moulton methods), and equations (8) have guaranteed stability for small enough h. The detailed behavior of all the roots is shown on page 96 of reference 6, and the general stability boundary is shown in sketch (d). Although the method is quite accurate, it is not suitable for coupled equations with parasitic eigenvalues because of its low stability boundary.



Sketch (d).- General stability boundary for equations (8).

An analysis of the above examples (and many others presented in ref. 6) leads to the conclusion expressed in the first sentence of this part. This does not mean, however, that valuable numerical methods (of the type being considered) cannot still be formulated. What appears to be needed are studies of special methods designed for special classes of equations (among which the quasi-linear equations (2) are already a special case). For example, if something is known about the eigenvalues in the local associated matrix in equations (3), particular numerical methods can be constructed which are superior to the classical ones for the particular parameters involved. Such cases are considered below.

DEFINITIONS AND TERMINOLOGY

We wish now to introduce some general notation and define precisely some terms used in the later discussion. First, consider M coupled, linear, ordinary differential equations with constant coefficients and (for simplicity, though it is not essential) distinct eigenvalues. Then

$$\vec{w}' \equiv \frac{d\vec{w}}{dt} = [A]\vec{w} + \vec{f} \tag{11}$$

where, if c_{m_j} are constants dependent on the initial conditions, t = nh, and σ_j are the eigenvalues of [A],

$$w_{mn} = \sum_{j=1}^{M} c_{m_j}(e^{\sigma_j h})^n + \text{Particular S.}\,, \qquad m = 1, 2, \cdots, M \tag{12}$$

Next choose any set of linear, difference-differential equations with constant coefficients (e.g., Runge-Kutta, Adams-Moulton, Hamming's, etc.) and symbolize it by the operator L. By itself L represents any linear set of differencing operations composing a method; on the other hand, L_e refers specifically to linear explicit methods, and L_i to linear implicit ones. Thus if L_e represents the predictor $u^{(1)}_{n+1} = u_n + hu'_n$ followed by the corrector $u_{n+1} = u_n + \frac{1}{2}h\left(u^{(1)'}_{n+1} + u_n\right)$, the operation $L_e(u' = \sigma u + f)$ results in the difference equations

$$u_{n+1}^{(1)} = (1 + \sigma h)u_n + hf_n$$

$$u_{n+1} = u_n + \frac{1}{2}\sigma h\left(u_{n+1}^{(1)} + u_n\right) + \frac{1}{2}h(f_{n+1} + f_n)$$

having the characteristic polynomial

$$P(E) = E - 1 - \sigma h - \frac{1}{2}\sigma^2 h^2 = 0$$

For arbitrary L the operation $L(u' = \sigma u + f)$ results in a set of difference equations, the solution to which can always be written

$$u_n = \sum_{i=1}^{k} B_i(\lambda_i)^n + P.S. \tag{13}$$

where λ_i are the roots to the characteristic polynomial $P(E) = 0$. The value of these roots depends on the choice of L and, in particular,

$$\lambda_i = g_i(\sigma H) , \quad i = 1, 2, \cdots, k \tag{14a}$$

$$\lambda_1 = e^{\sigma H} + O(H^{l}) \tag{14b}$$

λ_i are the spurious roots of the numerical method if $i > 1$

B_i are constants dependent upon the initial conditions

H is the representative step size defined in equation (4)

It should be noted that:

The value of l depends upon the order of the local Taylor series which L embeds in the calculations. In general, l is independent of k even for stable L. Dahlquist imposes special conditions under which l and k are connected if stability is imposed. (15)

The function g_i depends entirely upon the choice of L. (16)

Now it can be shown (see ref. 6) that the operation $L(\vec{w} = [A]\vec{w} + \vec{f})$ results in a set of coupled difference equations, the solutions to which can always be written

$$w_{mn} = \sum_{j=1}^{M}\sum_{i=1}^{k} \bar{c}_{m_j}(\lambda_{ji})^n + P.S. , \quad m = 1, 2, \cdots, M \tag{17}$$

where

$$\lambda_{ji} = g_i(\sigma_j H), \quad i = 1, 2, \cdots, k; \quad j = 1, 2, \cdots, M \tag{18a}$$

$$\lambda_{j1} = e^{\sigma_j H} + O(H^l) \tag{18b}$$

λ_{ji} are the spurious roots of the numerical method if $i > 1$

$\bar{c}_{m_j}$ are constants dependent upon the initial conditions

It should be noted that:

If, and only if,[1] the same L is applied in a given calculation step, h, to all the equations in a coupled set, the numerical accuracy and stability do not depend upon the elements in $[A_n]$ except as those elements affect the eigenvalues σ_j. (19)

Under the same conditions as in (19), the functions g_i in equations (14a) and (18a) are identical. (20)

The following definitions can now be formulated.

A set of differential equations is inherently stable if $\mathrm{Re}(\sigma_j) \leq 0$, $j = 1, 2, \cdots, M$. (21)

If a set of differential equations is inherently stable, the set of difference equations formed by $L(\vec{w} = [A]\vec{w} + \vec{f})$ has an induced instability if $|\lambda_{ji}| = |g_i(\sigma_j H)| > 1$, $j = 1, 2, \cdots, M$; $i = 1, 2, \cdots, k$. (Note that $i = 1$ is included.) (22)

If $\mathrm{Re}(\sigma_j H) \leq 0$, the value of $\sigma_j H$ for which any increase in H makes $|\lambda_{ji}| = |g_i(\sigma_j H)| > 1$ is labeled $|\sigma_j H|_c$ and called the general stability boundary, or simply, the stability boundary. (23)

If we set $\sigma_j = \bar{\sigma}_j e^{i\theta}$ (where $\bar{\sigma}_j$ and θ are real), and let $\bar{\sigma}_j H < 0$, the value of $\bar{\sigma}_j H$ for which any increase in H makes $|\lambda_{ji}| = |g_i(\bar{\sigma} H)| > 1$ is labeled $|\bar{\sigma} H|_c$ and called the real stability boundary. Similarly, the value of $i\bar{\sigma}_j H$ for which any increase in H makes $|\lambda_{ji}| = |g_i(i\bar{\sigma}_j H)| > 1$ is labeled $|i\bar{\sigma} H|_c$ and called the imaginary stability boundary. (24)

[1] Proof of (19) is given in reference 6. It should be noted that roundoff effects (which may be important) are neglected in all statements made in this report.

Methods for which $|\sigma H|_c < \infty$ are called conditionally stable. (25)

All explicit methods are conditionally stable. A proof is given in reference 1. (26)

Certain classes (containing any order of accuracy) of implicit methods are unconditionally stable. This is discussed in the last section. (27)

The σ_j can be divided into two classes:

1. Driving eigenvalues, q of them with subscript d, say, and
2. Parasitic eigenvalues, M-q of them with subscript p

and the parasitic eigenvalues can be subdivided into two groups: parasitic-stiff and parasitic-saddle. The definitions follow:

Driving combined with parasitic-stiff eigenvalues result when

$$|\sigma_d| \ll |\sigma_p|$$

$$\sum_{j=1}^{q} c_j (e^{\sigma_j H})_d^n \gg \sum_{j=q}^{M} |c_j (e^{\sigma_j H})_p| \qquad (28)$$

Driving combined with parasitic-saddle eigenvalues result when

$$\mathrm{Re}(\sigma)_p > 0$$

$$\mathrm{Re}(\sigma)_d < 0$$

$$\sum_{j=1}^{q} c_j (e^{\sigma_j H})_d \neq 0 \qquad (29)$$

$$\sum_{j=q}^{M} c_j (e^{\sigma_j H})_p = 0 \quad \text{(analytically)}$$

THE CONTROL OF PARASITIC EIGENVALUES

Consider the following numerical problem: How can one integrate coupled, quasi-linear, differential equations with both driving and parasitic eigenvalues, letting the step size be determined by the driving ones, and, at the same time, introducing no significant errors due to the parasitic ones?

Of course, one could always locally linearize the equations and uncouple them. But this, in effect, requires not only finding the elements of [A], but also the eigenvalues and, what is more drastic, the eigenvectors themselves. The remarkable thing is that, although there are cases for which finding the elements of [A] is advisable, actually calculating the eigenvalues or eigenvectors is unnecessary. It is a direct consequence of (19) that, under the qualifications noted, any standard, linear, differencing scheme integrates the coupled equations by seeking out each individual eigenvalue and integrating it as if the others did not exist (see footnote 1). Under these conditions, a step size can be chosen that will integrate one group of eigenvalues completely inaccurately but another group with high accuracy, and if the two groups are coupled together, the high accuracy would remain for the one in spite of the large errors in the others. (An example is given in ref. 8, pp. 35-36.) However, in general, all the coupled solutions would be inaccurate, and the highly accurate results could only be recovered by uncoupling the final answers.

While the last remark is interesting, it is not quite at the heart of the matter. Consider instead the following logic which is, in fact, the "solution" to the problem posed at the beginning of this section.

1. Let the coefficients c_j of $(e^{\sigma_j H})_p$ in equations (28) and (29) be made small with respect to the coefficients of $(e^{\sigma_j H})_d$ by choosing appropriate initial conditions at the commencement of the numerical integration.

2. Choose any numerical method, L, that is stable for all $(\sigma_j H)_p$ with total disregard as to its accuracy for them.

3. Let the same L be both stable and accurate for all $(\sigma_j H)_d$.

4. Then, in applying L to a set of differential equations in which $(\sigma_j H)_p$ and $(\sigma_j H)_d$ are coupled in any fashion, all of the coupled solutions will be accurately resolved and will represent the solution due to the driving eigenvalues as if the parasitic ones had been removed.

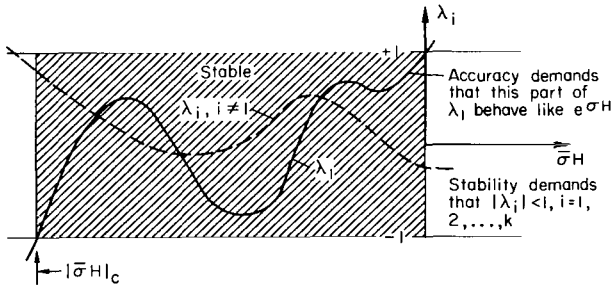
HIGHLY STABLE EXPLICIT METHODS

Some highly stable explicit methods for integrating coupled, autonomous, quasi-linear, differential equations with real, parasitic-stiff eigenvalues were developed in reference 1. For the sake of completeness, the subject is briefly reviewed in this section.

The characteristic equation for any explicit method, L_e, is a monic polynomial of the general form

$$E^k + P_{k-1}(\sigma H)E^{k-1} + \cdots + P_0(\sigma H) = 0 \tag{30}$$

where the $P_j(\sigma H)$ are themselves polynomials in σH. Highly stable explicit methods for real, parasitic-stiff eigenvalues can be constructed, if L_e is chosen so that the roots to its characteristic polynomial, equation (30), behave like those shown in sketch (e). The principal root must correspond to the expansion of $e^{\sigma H}$ through order l if the accuracy of L_e is to be $O(H^l)$. Aside from this, it is only necessary to make the magnitude of λ_1 and all the spurious roots λ_i, $i > 1$ (if there are any) less than one for as large a range of $-\bar{\sigma}H$ as possible.



Sketch (e).

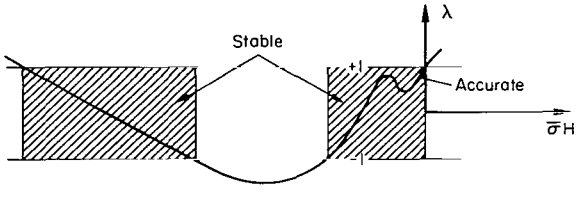
In reference 1 one-root methods were developed for autonomous equations by finding the b_n in

$$\lambda_1 = 1 + \bar{\sigma}h + \frac{1}{2}(\bar{\sigma}h)^2 + \sum_{n=3}^{N} b_n(\bar{\sigma}h)^n = e^{\bar{\sigma}h} + O(h^3) \tag{31}$$

such that $|\bar{\sigma}H|_c$ in sketch (e) was maximized in a least squares sense. (The actual optimum polynomials are unknown for $N > 3$.) A summary of the results is contained in the following chart.

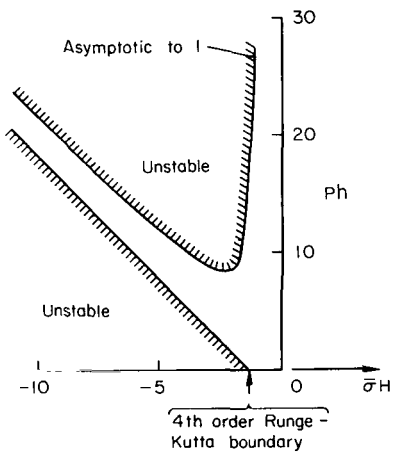
N	4	5	6	7	8	9	10
$\lvert\sigma h\rvert_c$	12	17	25	35	45	57	70
$\lvert\sigma H\rvert_c$	5.5	7.2	8.6	10.1	11.4	12.7	14.1

Notice that methods can be constructed that are highly stable for complex or even imaginary σ_j. Furthermore, they can be constructed such that they are stable in certain bands of σH and unstable in other bands. For example, taking again the case for real σ (not only for simplicity, but since this is a practical situation), the b_n in equation (31) could be chosen so that the curve in sketch (f) would result. The method L_e which had such a (single) root in its



Sketch (f).

characteristic polynomial would be stable when used to integrate differential equations with rather large negative eigenvalues, provided the intermediate smaller set did not appear in the coupled group. This kind of method has not been thoroughly exploited, but it is not impractical. An example of its use is contained in the method suggested by Treanor (ref. 9). If Treanor's method is applied to the representative equation (7), its stability bounds would be those illustrated in sketch (g). The parameter P can either be fixed,[2] or varied as the integration proceeds. Maximum stability for all $\bar{\sigma}H$ in the range $0 > \bar{\sigma}H > -|\bar{\sigma}H|_c$ is found when Ph = 8 and gives $|\bar{\sigma}H|_c = 5$, a considerable improvement over the Runge-Kutta methods shown in sketch (c). For higher values of Ph stability can be attained for larger $-\bar{\sigma}H$, provided they lie in the stability corridor shown in the sketch.



Sketch (g).- Real stability boundaries for Treanor's method.

HIGHLY STABLE IMPLICIT METHODS

The explicit methods just discussed have the advantage that they can be applied directly to equations (2) and have the theoretical capability of integrating those equations with arbitrary accuracy. There are many practical cases, however, for which their limited stability boundary makes them very costly to apply. In such instances the use of implicit methods may be advisable. To compare the efficiency (measured with regard to machine running time) of implicit methods with explicit ones is not a simple matter because the former require the step by step conversion of equations (2) to equations (3), and the subsequent solution of simultaneous algebraic equations. An attempt at a rational comparison is presented in reference 8. In this report the capabilities of the implicit methods are considered without regard to their efficiency. With these words of caution, we return to the use of the calculation step size, h, as the reference parameter in estimates of stability and accuracy.

From the point of view of the computer, the principal difference between explicit and implicit methods is that the latter require the solution of M simultaneous algebraic equations at each step. From a more abstract point of view, the fundamental difference between the two is that the characteristic polynomial for the implicit methods is no longer monic, and can always be written in the form

[2]Treanor takes P at each step to be the ratio of certain terms in the first two steps of a standard fourth-order Runge-Kutta process. For an analysis of Treanor's method see reference 1.

$$P_k(\sigma h)E^k + P_{k-1}(\sigma h)E^{k-1} + \cdots + P_0(\sigma h) = 0 \tag{32}$$

where the $P_j(\sigma h)$ are polynomials in σh. This difference is extremely important.

Consider, for example, the implicit modified Euler method. Let L_i represent

$$u_{n+1} = u_n + \frac{1}{2} h(u'_{n+1} + u'_n) \tag{33}$$

Then $L_i(u' = \sigma u + f)$ results in the difference equation

$$\left(1 - \frac{1}{2}\sigma h\right) u_{n+1} = \left(1 + \frac{1}{2}\sigma h\right) u_n + \frac{1}{2} h(f_{n+1} + f_n)$$

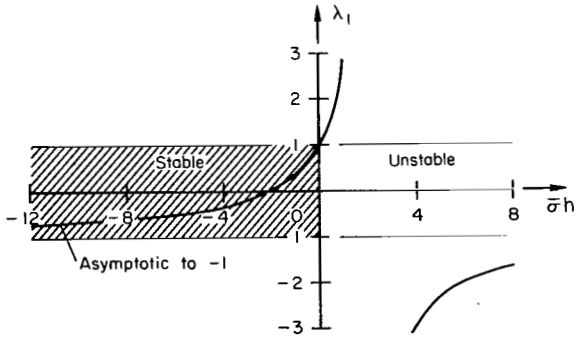
having the characteristic polynomial

$$\left(1 - \frac{1}{2}\sigma h\right) E - \left(1 + \frac{1}{2}\sigma h\right) = 0 \tag{34}$$

The single root

$$\lambda_1 = \frac{1 + \frac{1}{2}\sigma h}{1 - \frac{1}{2}\sigma h} \tag{35}$$

has a magnitude that is less than 1 (see sketch (h)) for all $\mathrm{Re}(\sigma h) < 0$, and the method is, as is well known, unconditionally stable.



Sketch (h).- Real stability boundary for implicit modified Euler method (eq. (33)).

Not all implicit methods are unconditionally stable. For example, if L_i represents the Adams-Moulton, two-step-corrector used implicitly

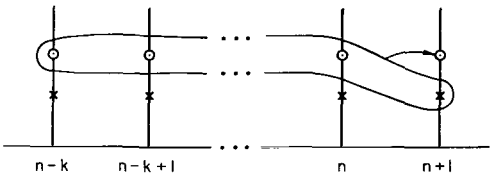
$$u_{n+2} = u_{n+1} + \frac{1}{12} h(5u'_{n+2} + 8u'_{n+1} - u'_n) \tag{36}$$

the operation $L_i(u' = \sigma u + f)$ results in the characteristic polynomial

$$\left(1 - \frac{5}{12}\sigma h\right) E^2 - \left(1 + \frac{8}{12}\sigma h\right) E + \frac{1}{12}\sigma h = 0 \tag{37}$$

which is unstable if $\bar{\sigma}h < -6$.

On the other hand, unconditionally stable methods with any order of accuracy can be constructed. The graphic description of a class of them is shown in sketch (i) for equispaced data. These are sometimes referred to as the backward difference formulas and the first five of them are given in reference 10, pages 96-98. These methods can be "more stable" than the modified Euler method in the sense that the roots to their characteristic polynomials are smaller for large $|\sigma h|$. For example, the two roots derived from the use of



Sketch (i).- Unconditionally stable implicit methods.

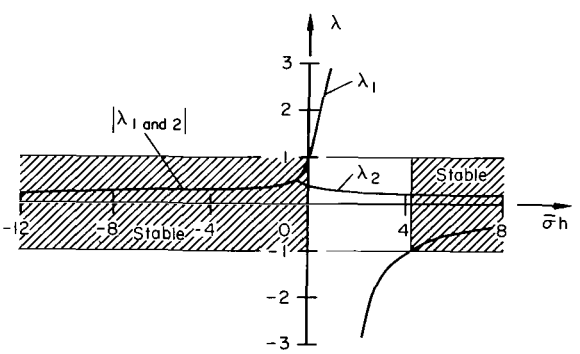
$$u_{n+2} = \frac{1}{3}\left(4u_{n+1} - u_n + 2hu'_{n+2}\right) \tag{38}$$

are

$$\lambda_1 = \frac{2 + \sqrt{1 + 2\sigma h}}{3 - 2\sigma h} \tag{39a}$$

$$\lambda_2 = \frac{2 - \sqrt{1 + 2\sigma h}}{3 - 2\sigma h} \tag{39b}$$

Both have magnitudes less than 1 for all $\mathrm{Re}(\sigma h) < 0$, but behave asymptotically as $1/\sqrt{|\sigma h|}$, rather than as 1 for the modified Euler case. Their variation for real σh is shown in sketch (j). Equation (38) has been used in boundary-layer studies (ref. 11), and equation (33) is extremely popular (ref. 12) for use in studying parabolic partial differential equations, in which discipline they are known as the Crank-Nicholson equations.



Sketch (j).- Real stability boundary for implicit method given by equation (38).

Dahlquist has shown (ref. 13) that only one L_i with $p \geq k + 1$ is unconditionally stable, and in fact, the method is that given by equation (33). In this statement p refers to the highest order of the embedded Taylor series that accurately represents $e^{\sigma h}$ in the expansion of $\lambda_1 = g_1(\sigma h)$ (i.e., $p = l - 1$ in eq. (14b)), and k is the step number in an equispaced, multistep, linear method. Dahlquist is sometimes misquoted as saying that only one implicit method is unconditionally stable, namely, equation (33) which has an error $O(h^3)$. However, his theorem states nothing about the order of accuracy that can be achieved in unconditionally stable implicit methods. What it does state is that if, for example, an unconditionally stable, four-step, equispaced,

linear, implicit method is used, its embedded Taylor series expansion can be accurate, at most, through the term $(1/24)(\sigma h)^4$.

Aside from the hypothesis of equispaced steps in L_i, there is another qualification underlying Dahlquist's theorem discussed in the previous paragraph; namely, that $[A_n]$ is used only to the first power. Let us consider this next in more detail.

The operator L_i converts the (locally) linear differential equations into a set of M algebraic equations that must be solved by the methods of linear algebra. For example, if L_i refers to equation (33), $L_i(\vec{w} = [A_n]\vec{w} + \vec{f}_n)$ results in

$$\vec{w}_{n+1} = \vec{w}_n + \frac{1}{2} h([A_n]\vec{w}_{n+1} + \vec{f}_n + [A_n]\vec{w}_n + \vec{f}_n) \tag{40}$$

which can be rearranged to form

$$\left([I] - \frac{1}{2} h[A_n]\right)(\vec{w}_{n+1} - \vec{w}_n) = h[A_n]\vec{w}_n + h\vec{f}_n \tag{41}$$

Now it is clear that equation (33) is a special difference-differential relationship from the class

$$u_{n+1} = u_n + h(a_1 u'_n + a_2 u'_{n+1}) + h^2(b_1 u''_n + b_2 u''_{n+1}) + \cdots \tag{42a}$$

which is, in turn, a special case of <u>multistep</u> groups (note, eq. (42a) only relates data between one step) with higher order derivatives. In this report, however, attention is restricted to equations (42a) and, in particular, to

$$u_{n+1} = u_n + hau'_n + ha_2 u'_{n+1} + h^2 b_2 u''_{n+1} \tag{42b}$$

since it displays all the essential points to be made. Now if $u' = \sigma u + f$, where σ is a constant and f is also a constant (i.e., <u>autonomous equations, the only kind we will consider in the following</u>), $u'' = \sigma u' = \sigma^2 u + \sigma f$. Let L_i represent equation (42b). Then the operation $L_i(u' = \sigma u + f)$ results in the difference equation

$$u_{n+1} = (1 + a\sigma h)u_n + \sigma h(a_2 + b_2\sigma h)u_{n+1} + hf(a + a_2 + \sigma h b_2) \tag{43}$$

which has the characteristic polynomial

$$(1 - a_2\sigma h - b_2\sigma^2 h^2)E - (1 + a\sigma h) = 0 \tag{44}$$

The single root expands to give

$$\lambda_1 = g_1(\sigma h) = \frac{1 + a\sigma h}{1 - a_2\sigma h - b_2\sigma^2 h^2} = 1 + (a + a_2)\sigma h + [b_2 + a_2(a + a_2)]\sigma^2 h^2$$

$$+ [a_2^2(a + a_2) + b_2(a + 2a_2)]\sigma^3 h^3 + \cdots \tag{45}$$

Insisting that λ_1 represent the expanded $e^{\sigma h}$ through the term $(1/2)\sigma^2 h^2$ gives

$$\left.\begin{aligned} a_2 &= 1 - a \\ b_2 &= a - \frac{1}{2} \end{aligned}\right\} \tag{46}$$

for which equation (45) reduces to

$$\lambda_1 = \frac{1 + a\sigma h}{1 + (a - 1)\sigma h + \left(\frac{1}{2} - a\right)\sigma^2 h^2} = 1 + \sigma h + \frac{1}{2}\sigma^2 h^2 + \frac{a}{2}\sigma^3 h^3 + \cdots \tag{47}$$

The exact solution to the difference equation (43) combined with equations (46) is

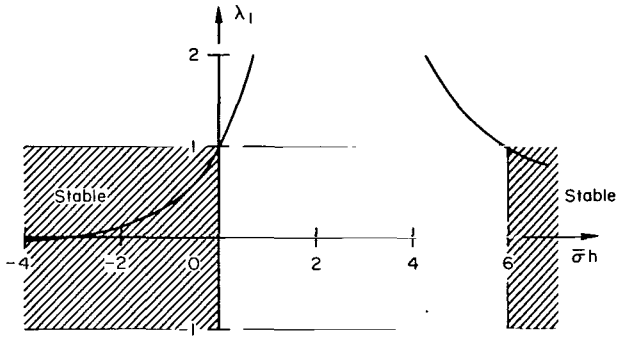
$$u_n = c_o\left(\frac{1 + a\sigma h}{1 + (a - 1)\sigma h + \left(\frac{1}{2} - a\right)\sigma^2 h^2}\right)^n - \frac{f}{\sigma} \tag{48a}$$

and the exact solution of $u' = \sigma u + f$, with constant σ and f, is

$$u_n = \bar{c}_o(e^{\sigma h})^n - \frac{f}{\sigma} \tag{48b}$$

The particular solution is calculated exactly and the complementary solution has an error $O(h^3)$, consistent with the error in equations (3).

There remains the discussion of stability for the method L_i defined by equations (42b) and (46). Let us first consider the <u>real</u> stability boundary in detail and later present the results for complex and imaginary σ. When $a = 1/3$, the error is minimized (the method is then $O(h^4)$) and λ_1 has the



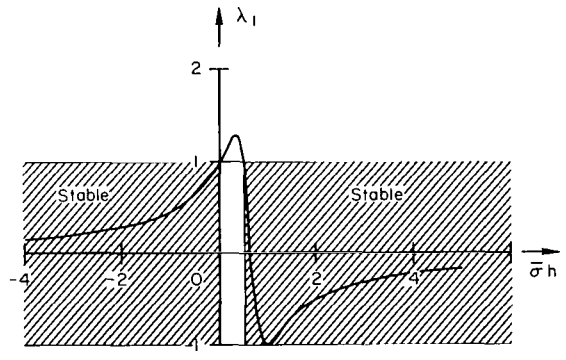
Sketch (k).- Real stability boundary for method given by equation (52b).

variation shown in sketch (k). This form of the method is not only unconditionally stable, it is also stable for all positive $\bar{\sigma}h \geq 6$. In general, the value of $\lambda_1(\sigma h)$ given by equation (47) crosses the line +1 when

$$\left.\begin{aligned} \bar{\sigma}h &= 0 \\ \bar{\sigma}h &= 2/(1 - 2a) \end{aligned}\right\} \qquad (49)$$

and the line -1 when

$$\bar{\sigma}h = 1 \pm \sqrt{1 - \left(\frac{4}{1 - 2a}\right)} \qquad (50)$$



Sketch (l).- Real stability boundary for method given by equation (52c).

Thus, between $-3/2 \leq a \leq 1/2$ the value of λ_1 is never less than -1 for any $\bar{\sigma}h$, and is above the line 1 only in the range $0 < \bar{\sigma}h < 2/(1 - 2a)$. Hence, the maximum real stability of the method is achieved (see sketch (l)) when $a = -3/2$, in which case the error is $\sim(11/12)(\bar{\sigma}h)^3$, and the $|\lambda_1| > 1$ only for $0 < \bar{\sigma}h < 1/2$. Of course, the accuracy for $\bar{\sigma}h > 1/2$ is completely lost, but if the positive eigenvalues are parasitic-saddle, this is immaterial.

The complete stability properties for all complex σh are shown, for the case $a = -3/2$, in figure 1 in the $(\lambda, \sigma h)$ plane, where the lines ± 1 represent the stability boundaries. Figure 2 shows some of the same results in the complex plane where $\mathrm{Re}(\lambda_1)$ is plotted against $\mathrm{Im}(\lambda_1)$, and the unit circle is the stability boundary.

When L_i represents

$$u_{n+1} = u_n + h[au'_n + (1 - a)u'_{n+1}] + \left(a - \frac{1}{2}\right) u''_{n+1} \qquad (51)$$

the operation $L_i(\vec{w} = [A_n]\vec{w} + \vec{f}_n)$ results in the difference equation

$$\left([I] + h(a - 1)[A_n] + h^2\left(\frac{1}{2} - a\right)[A_n]^2\right)(\vec{w}_{n+1} - \vec{w}_n)$$

$$= h\left([I] + h\left(a - \frac{1}{2}\right)[A_n]\right)\left([A_n]\vec{w}_n + \vec{f}_n\right) \qquad (52a)$$

For the most accurate case, a = 1/3 and

$$\left([I] - \frac{2}{3}h[A_n] + \frac{1}{3}h^2[A_n]^2\right)(\vec{w}_{n+1} - \vec{w}_n) = h\left([I] - \frac{1}{3}h[A_n]\right)([A_n]\vec{w}_n + \vec{f}_n) \tag{52b}$$

and for the most stable case, a = -3/2 and

$$\left([I] - \frac{5}{2}h[A_n] + 2h^2[A_n]^2\right)(\vec{w}_{n+1} - \vec{w}_n) = h([I] + h[A])([A_n]\vec{w}_n + \vec{f}_n) \tag{52c}$$

Equations (52) are actually programmed when the method is used on a digital computer. They can be said to violate the Dahlquist theorem discussed above; that is, equation (42b) is an unconditionally stable method for which (in the notation used in presenting the theorem) p = 3 and k = 1. But, as has been pointed out, such a violation does not really occur, since Dahlquist's theorem is based on the premise that $[A_n]$ is used only to the first power.

The practical application of methods such as these must, of course, be made with care. They are designed for quasi-linear differential equations for which the standard general methods are unsatisfactory and something is known, or can be determined about, the eigenvalue structure of the associated $[A_n]$ matrix. An example of such a case is given by the quasi-linear equation

$$(1 - w)\frac{dw}{dt} = 2w - t + 0.5 \tag{53}$$

with the initial conditions

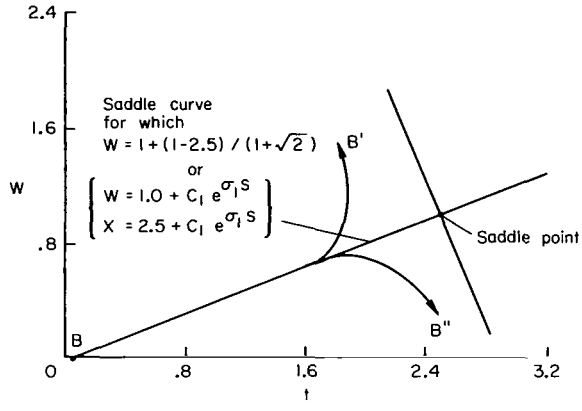
$$\left.\begin{aligned} t_o &= 0 \\ w_o &= (7 - 5\sqrt{2})/2 \end{aligned}\right\} \tag{54}$$

It is easy to show that equation (53) has a saddle point (see, e.g., ref. 14) at w = 1 and t = 2.5. Introduce the new independent variable such that

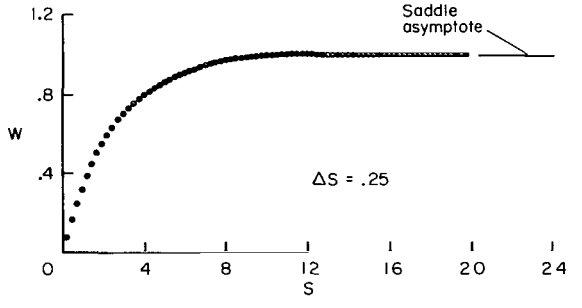
$$\frac{dt}{ds} = 1 - w \tag{55}$$

and equation (53) becomes

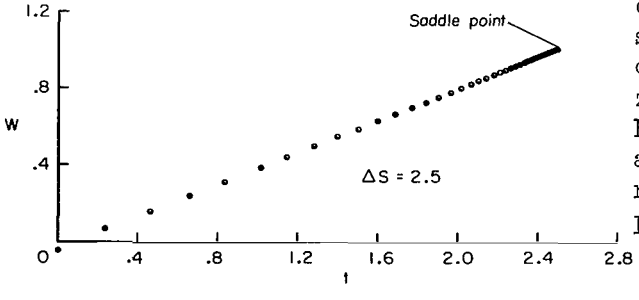
$$\frac{d\vec{w}}{ds} = [A]\vec{w} + \vec{f} \tag{56a}$$



Sketch (m).- Saddle behavior for equation (53).



Sketch (n).- Locus of computed points for w in (w, s) plane.



Sketch (o).- Locus of computed points in (t, w) plane.

where

$$\vec{w}^T = [w,s] \qquad (56b)$$

$$[A] = \begin{bmatrix} 2 & -1 \\ -1 & 0 \end{bmatrix} \qquad (56c)$$

and

$$\vec{f}^T = [0.5,1.0] \qquad (56d)$$

Notice that the introduction of equation (55) has transformed the nonautonomous equation (53) into the autonomous form (56). The latter may, therefore, be integrated using the methods defined by equations (52).

The eigenvalues of [A] are

$$\sigma_1 = -(\sqrt{2} - 1) \approx -0.414$$

$$\sigma_2 = +(\sqrt{2} + 1) \approx 2.414$$

The initial values are located at point (B) in sketch (m) and fall exactly on the saddle curve shown. Any standard numerical integrating scheme applied to equations (56) will diverge in typical fashion, following the curves (BB') or (BB") shown in the sketch. This is because the coefficient of the term with $e^{\sigma_2 h}$ (which is exactly zero along the saddle curve) cannot be held to zero in finite place arithmetic, and an instability will ensue if the numerical method itself is unstable for positive eigenvalues in this range.

Using the method described by equation (52c) with a step size $h = 0.25$ gave the results shown in sketches (n) and (o). The use of equation (55) transforms the saddle point to infinity in terms of s, the independent variable used in the numerical integration, so the saddle point is approached asymptotically in the (w,s) and (t,s) planes as is shown in sketch (n). (The behavior for t(s) was identical.) When w is plotted as a function of x, however, the "physical" nature of the variation is regained and shown in sketch (o). Notice that the value of σh is about -0.103 for $\sigma_1 h$, which reduces the local error to about 0.001. The value of the parasitic-saddle

eigenvalue is about 0.603. According to sketch (l), this gives a value of $|\lambda(\sigma_2 h)| < 1$, so the application of the method should be (and was) stable.

Ames Research Center
National Aeronautics and Space Administration
Moffett Field, Calif. 94035, May 15, 1968
129-01-02-05-00-21

REFERENCES

1. Lomax, Harvard: On the Construction of Highly Stable, Explicit, Numerical Methods for Integrating Coupled Ordinary Differential Equations With Parasitic Eigenvalues. NASA TN D-4547, 1968.

2. Butcher, J. C.: A Modified Multistep Method for the Numerical Integration of Ordinary Differential Equations. J. Assoc. Comput. Mach., vol. 12, no. 1, Jan. 1965, pp. 124-135.

3. Gragg, William B.; and Stetter, Hans J.: Generalized Multistep Predictor-Corrector Methods. J. Assoc. Comput. Mach., vol. 11, no. 2, April 1964, pp. 188-209.

4. Gear, C. W.: The Numerical Integration of Ordinary Differential Equations. Math. Comp., vol. XXI, no. 98, April 1967, pp. 146-156.

5. Brush, D. G.; Kohfeld, J. J.; and Thompson, G. T.: Solution of Ordinary Differential Equations Using Two "Off-Step" Points. J. Assoc. Comput. Mach., vol. 14, no. 4, Oct. 1967, pp. 769-784.

6. Lomax, Harvard: An Operational Unification of Finite Difference Methods for the Numerical Integration of Ordinary Differential Equations. NASA TR R-262, 1967.

7. Dahlquist, Germund G.: Convergence and Stability in the Numerical Integration of Ordinary Differential Equations. Math. Scand., vol. 4, 1956, pp. 33-53.

8. Lomax, Harvard; and Bailey, Harry E.: A Critical Analysis of Various Numerical Integration Methods for Computing the Flow of a Gas in Chemical Nonequilibrium. NASA TN D-4109, 1967.

9. Treanor, Charles E.: A Method for the Numerical Integration of Coupled First-Order Differential Equations With Greatly Different Time Constants. Math. Comp., vol. 20, no. 93, Jan. 1966, pp. 39-45.

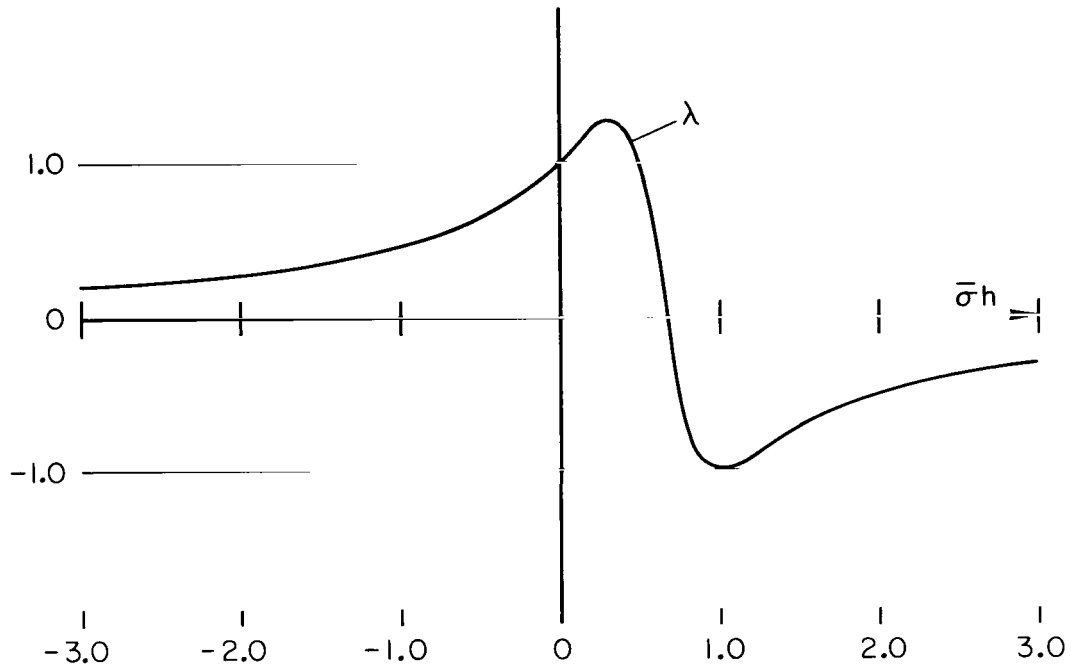
10. Milne, W. E.: Numerical Calculus. Princeton University Press, 1949.

11. Flügge-Lotz, I.; and Davis, R. T.: Laminar Compressible Flow Past Axisymmetric Blunt Bodies. Tech. Rep. 143, Div. of Eng. Mech., Stanford Univ., 1964.

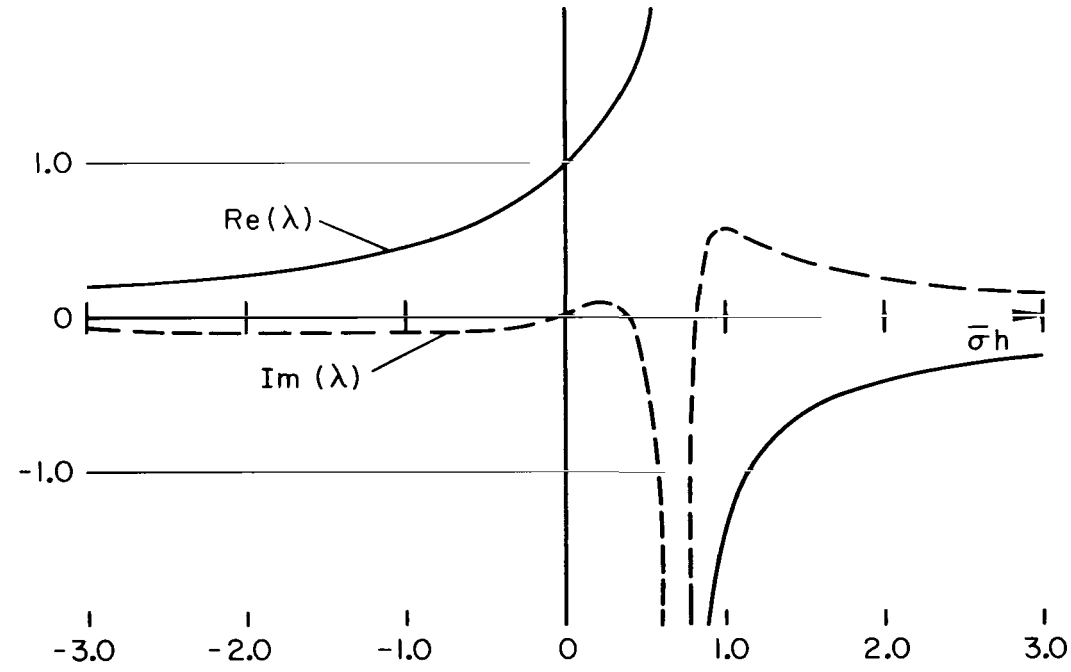
12. Varga, R. S.: Matrix Iterative Analysis. Prentice-Hall, Inc., 1962.

13. Dahlquist, Germund G.: Stability Questions for Some Numerical Methods for Ordinary Differential Equations. Proc. Symp. Appl. Math., vol. XV, 1963, pp. 147-158.

14. Saaty, T. L.; and Bram, J.: Nonlinear Mathematics. McGraw-Hill Book Co., Inc., 1964.

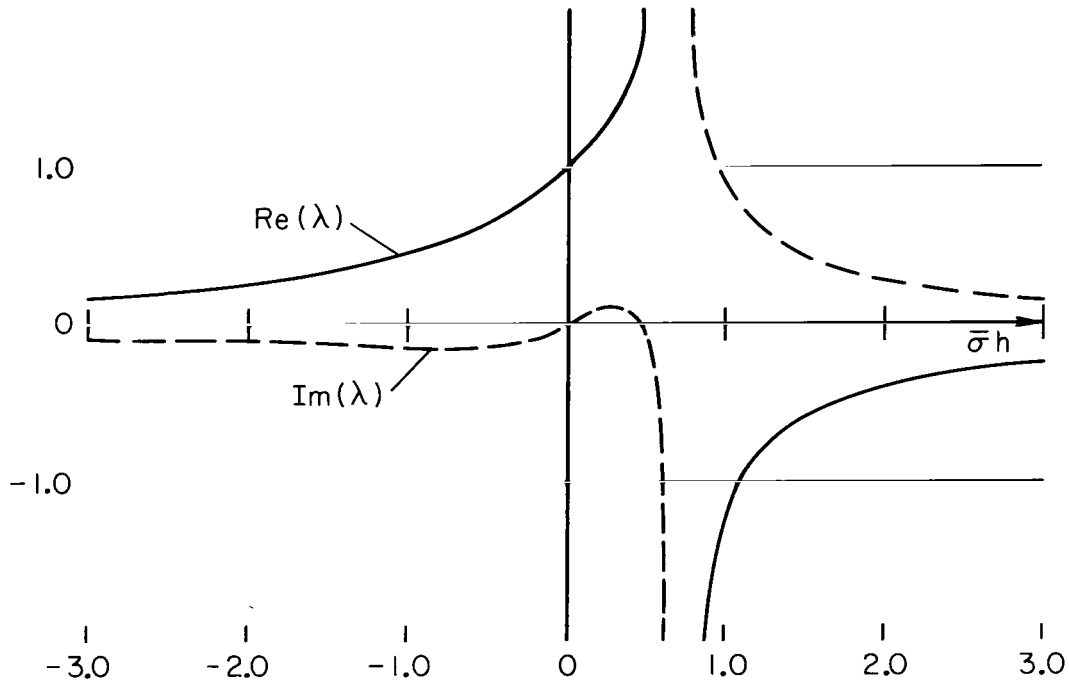


(a) $\sigma = \bar{\sigma}e^{i\theta}$, $\theta = 0^{\circ}$

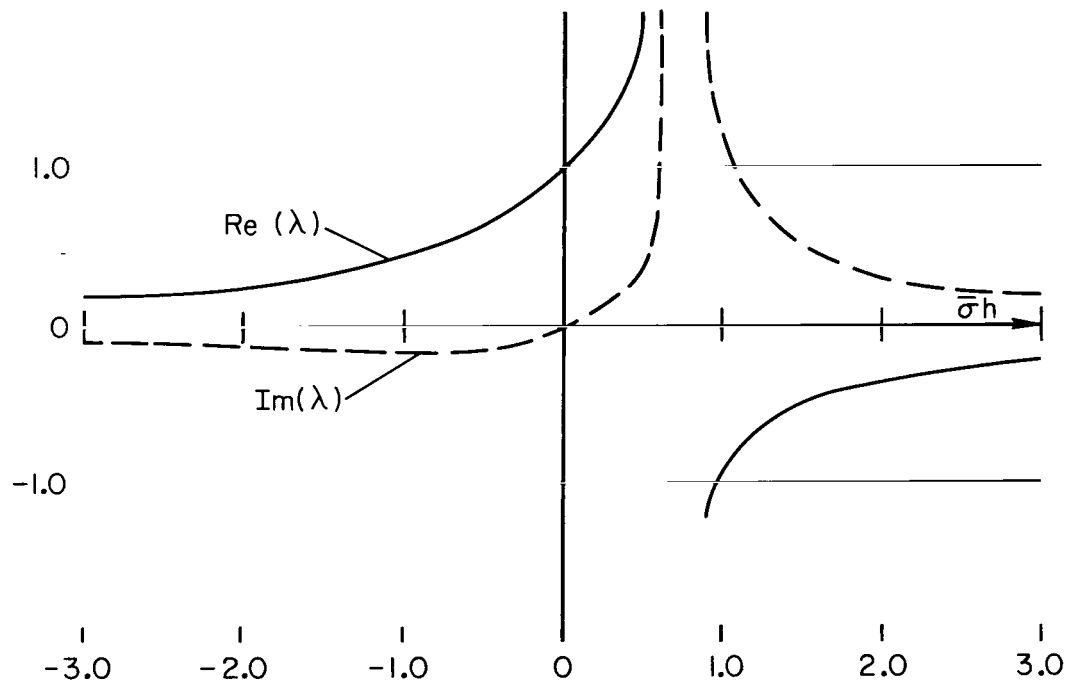


(b) $\sigma = \bar{\sigma}e^{i\theta}$, $\theta = 23^{\circ}$

Figure 1.- Variation with $\bar{\sigma}h$ of the single root, λ_1, in the characteristic polynomial for the method defined by equation (52c).

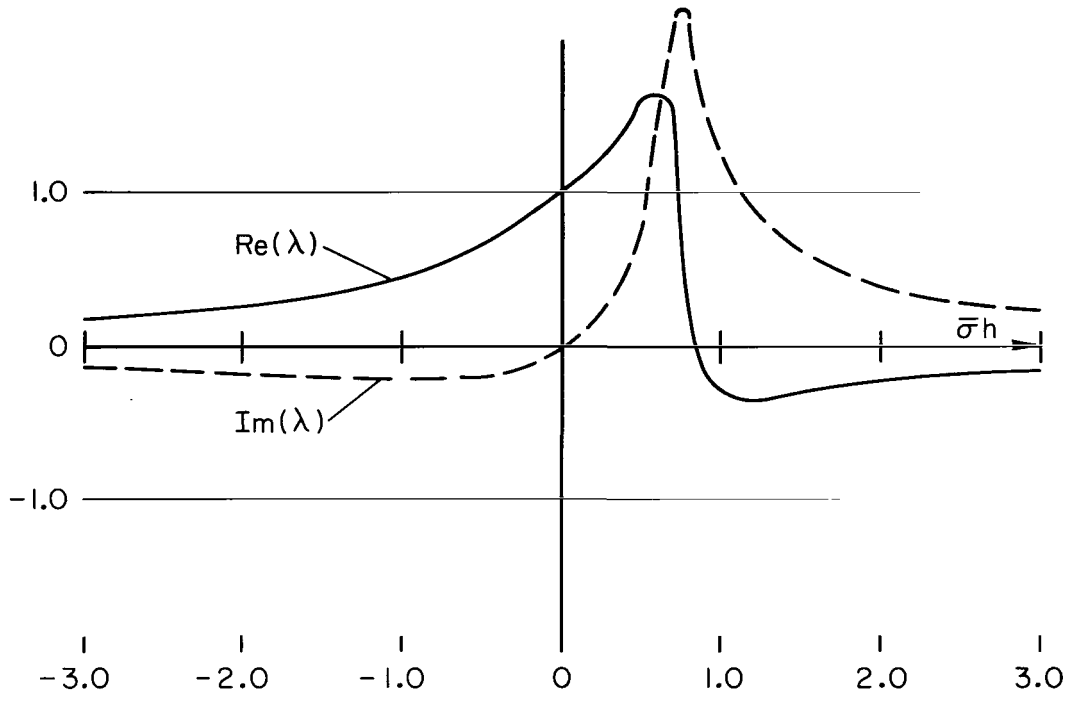


(c) $\sigma = \bar{\sigma}e^{i\theta}$, $\theta = 28^{\circ}$

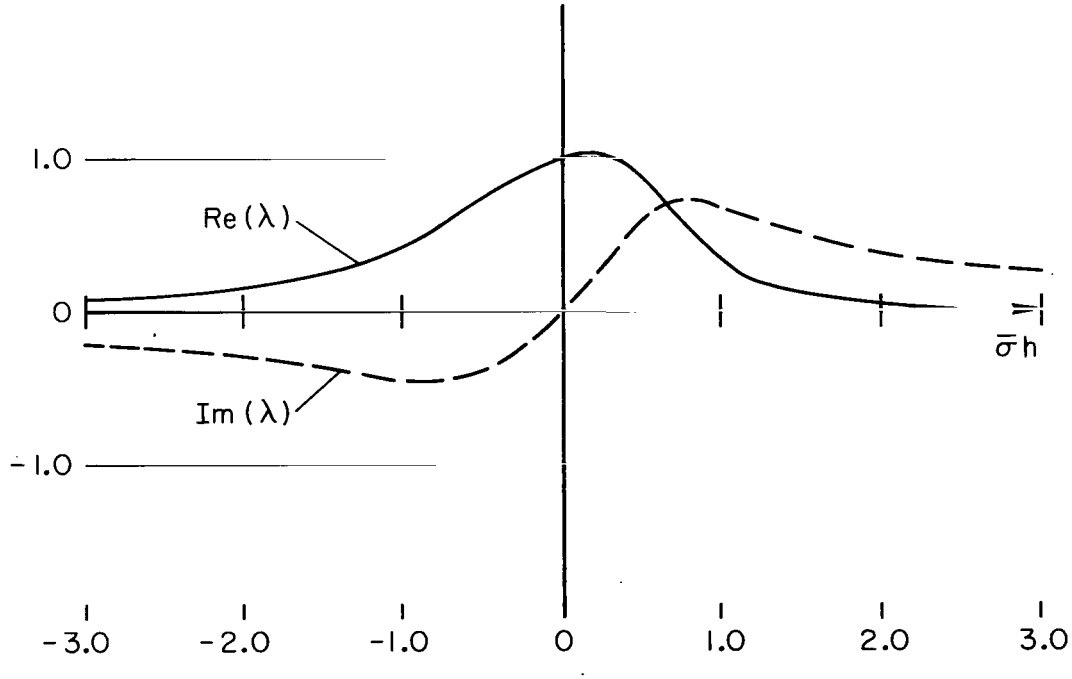


(d) $\sigma = \bar{\sigma}e^{i\theta}$, $\theta = 34^{\circ}$

Figure 1.- Continued.

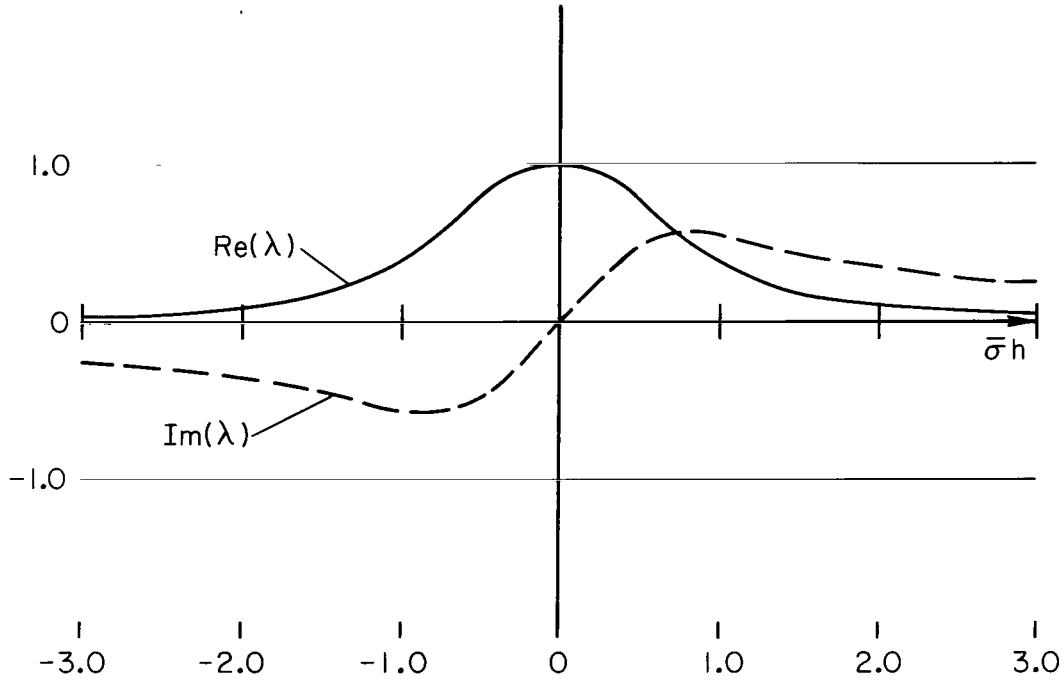


(e) $\sigma = \bar{\sigma}e^{i\theta}$, $\theta = 45^{\circ}$

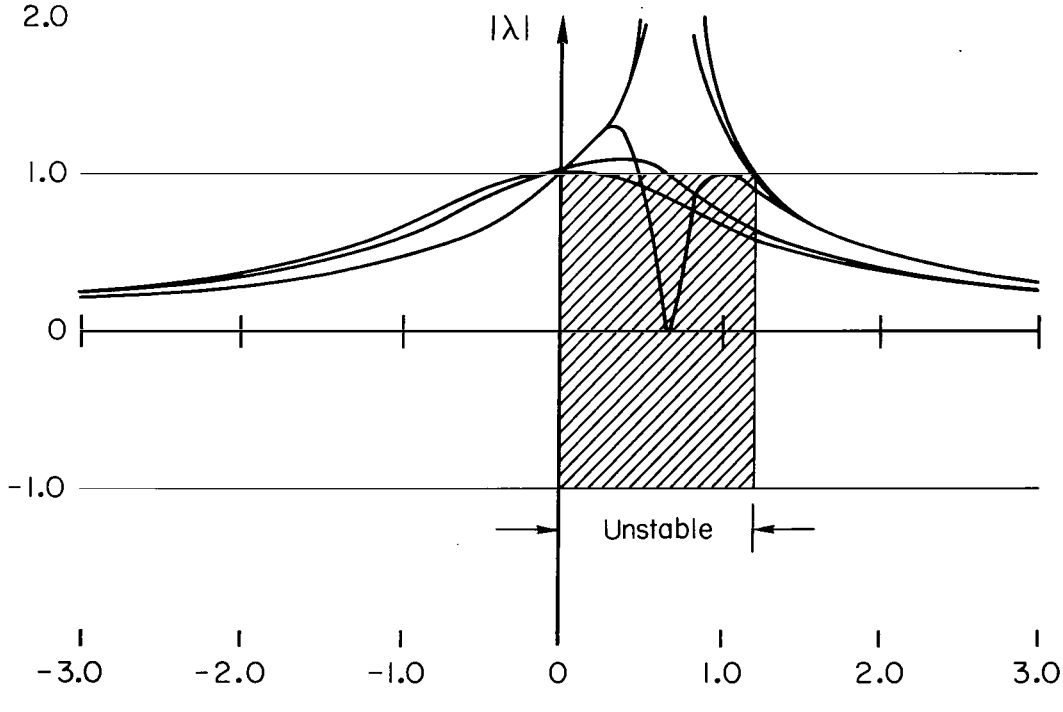


(f) $\sigma = \bar{\sigma}e^{i\theta}$, $\theta = 78^{\circ}$

Figure 1.- Continued.



(g) $\sigma = \bar{\sigma}e^{i\theta}$, $\theta = 90^{\circ}$



(h) $\sigma = \bar{\sigma}e^{i\theta}$, $0 \leq \theta \leq 90^{\circ}$

Figure 1.- Concluded.

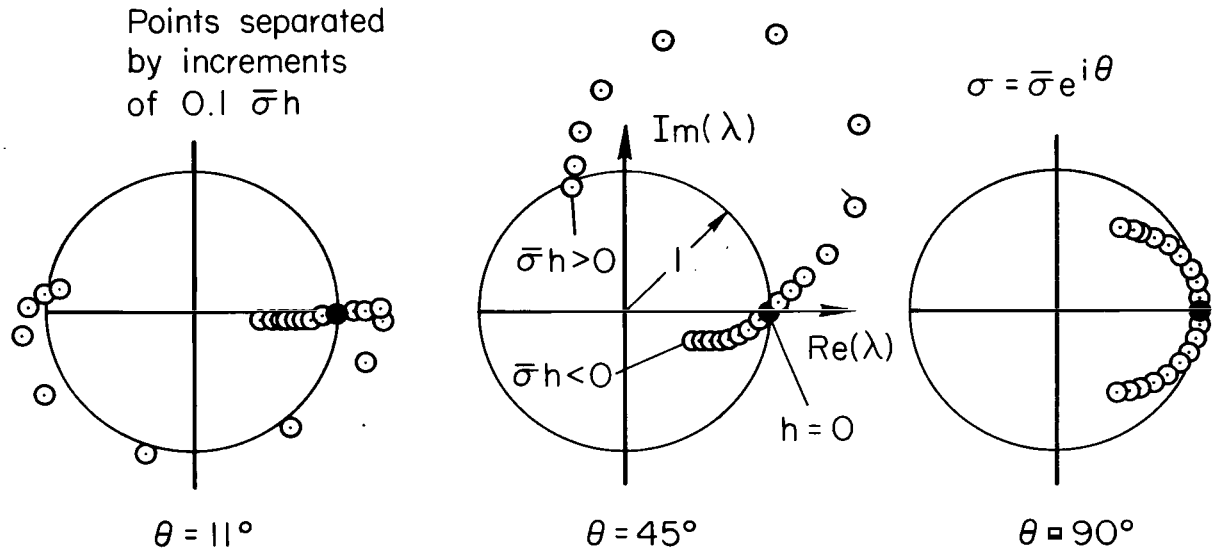


Figure 2.- Some results shown in figure 1 redisplayed in the complex plane.

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
WASHINGTON, D.C. 20546

OFFICIAL BUSINESS

FIRST CLASS MAIL

POSTAGE AND FEES PAID
NATIONAL AERONAUTICS AND
SPACE ADMINISTRATION

"The aeronautical and space activities of the United States shall be conducted so as to contribute . . . to the expansion of human knowledge of phenomena in the atmosphere and space. The Administration shall provide for the widest practicable and appropriate dissemination of information concerning its activities and the results thereof."

—NATIONAL AERONAUTICS AND SPACE ACT OF 1958

NASA SCIENTIFIC AND TECHNICAL PUBLICATIONS

TECHNICAL REPORTS: Scientific and technical information considered important, complete, and a lasting contribution to existing knowledge.

TECHNICAL NOTES: Information less broad in scope but nevertheless of importance as a contribution to existing knowledge.

TECHNICAL MEMORANDUMS: Information receiving limited distribution because of preliminary data, security classification, or other reasons.

CONTRACTOR REPORTS: Scientific and technical information generated under a NASA contract or grant and considered an important contribution to existing knowledge.

TECHNICAL TRANSLATIONS: Information published in a foreign language considered to merit NASA distribution in English.

SPECIAL PUBLICATIONS: Information derived from or of value to NASA activities. Publications include conference proceedings, monographs, data compilations, handbooks, sourcebooks, and special bibliographies.

TECHNOLOGY UTILIZATION PUBLICATIONS: Information on technology used by NASA that may be of particular interest in commercial and other non-aerospace applications. Publications include Tech Briefs, Technology Utilization Reports and Notes, and Technology Surveys.

Details on the availability of these publications may be obtained from:

SCIENTIFIC AND TECHNICAL INFORMATION DIVISION

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Washington, D.C. 20546