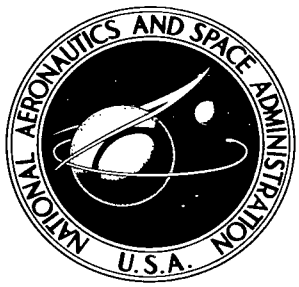
NASA TECHNICAL NOTE

NASA TN D-3696



NASA TN D-3696

AN ENGINEERING EXPOSÉ OF NUMERICAL INTEGRATION OF ORDINARY DIFFERENTIAL EQUATIONS

by John L. Engvall

Manned Spacecraft Center

Houston, Texas

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION • WASHINGTON, D. C. • NOVEMBER 1966

NASA TN D-3696

AN ENGINEERING EXPOSÉ OF NUMERICAL INTEGRATION OF ORDINARY DIFFERENTIAL EQUATIONS

By John L. Engvall

Manned Spacecraft Center
Houston, Texas

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

For sale by the Clearinghouse for Federal Scientific and Technical Information
Springfield, Virginia 22151 – Price $2.00

ABSTRACT

This paper presents error analyses for numerically integrating first order ordinary differential equations. Roundoff, truncation, mathematical instability, and numerical instability errors are the four basic errors considered. Emphasis is placed upon the catastrophic effects of numerical instability errors. Several examples of numerical and mathematical instability errors are contained in the paper.

AN ENGINEERING EXPOSÉ OF NUMERICAL INTEGRATION OF ORDINARY DIFFERENTIAL EQUATIONS

By John L. Engvall
Manned Spacecraft Center

SUMMARY

This paper presents error analyses for numerically integrating first order ordinary differential equations. Roundoff, truncation, mathematical instability, and numerical instability errors are the four basic errors considered with emphasis placed upon the catastrophic effects of numerical instability errors. Several examples of numerical and mathematical instability errors are presented.

Although instability is not a totally undesirable characteristic, in many cases the effect of instability errors is more significant than roundoff errors and truncation errors.

INTRODUCTION

Mathematical analysis concerning numerical integration of ordinary differential equations has been in progress at the Manned Spacecraft Center (MSC) for some time. This paper presents the theoretical results of a portion of this error analysis, and contains numerical examples which indicate the order of magnitude of the difficulties that may be encountered due to various errors. Emphasis is placed upon the importance of error growth due to numerical instability of multistep methods.

SYMBOLS

A	array of coefficients for infinite series associated with numerical methods
A(x)	particular solution to differential equation
C_1, C_2, C_3, C_4	constants of integration
$dy/dx = f(x, y)$	denotes ordinary differential equation

E, ER(x)	error in numerical solution
ET(x)	local truncation error
exp(x)	exponential function of x
F(x)	total solution to differential equation
G(x)	homogeneous solution to differential equation
HBAR	product of step size and $\partial f/\partial y$
h	denotes the integration step size
P	present set of data
q	array of differences T(I) - A(I)
R_1, R_2, R_3, R_4	expected relative errors for example 1
RER(x)	relative total error
RET	relative local truncation error
T	array of Taylor series constant coefficients
x_0	initial independent variable
$x(t_0)$	approximate initial condition
yN, FN	numerical solution to differential equation
$y^{(n)}(x)$	denotes the nth derivative of y with respect to x
$y(t_0)$	exact initial condition for solution to differential equation
y(x)	denotes solution in differential equation above
$\partial f/\partial y$	partial derivative of f with respect to y

Symbols not listed are used as constant parameters in the development of equations.

BASIC DEFINITIONS

Definition 1

Roundoff error is any error which is directly caused by the fact that, in computers, numbers must be expressed by a finite number of significant figures.

The term "truncation error" is often applied to error introduced by truncating a number or by retaining only the most significant part of a number. According to the above definition, this is roundoff error. The term "truncation error" will be used for a different type of error and is defined subsequently.

Roundoff error can occur before what is generally considered to be initial execution of the numerical integration. Conversion of input data from decimal to binary might generate errors of this form. For example, a number which can be represented by a finite number of significant figures in the decimal system may have an infinite repeating binary representation; for example,

$$0.1 \text{ (base ten)} = 0.000110011001100 \ldots\ldots\ldots \text{ (base two)}$$

Normally, individual roundoff errors tend to occur in the least significant figures carried within the computer. This certainly seems insignificant, and, as a matter of fact, roundoff error often receives little treatment in applied texts on numerical methods. Actually, roundoff error is the most significant error in the sense that if roundoff error can be made arbitrarily small, then almost all other errors can be made arbitrarily small; yet paradoxically, roundoff error can usually be assumed to be negligible compared to other errors. The above statement will be discussed later in detail.

Discussion of any other type errors will be governed by the following five restrictions:

a. Only differential equations which have unique bounded solutions through the initial data point and over the domain of numerical integration will be considered. The necessary and sufficient conditions for this restriction are fully discussed in reference 1.

b. The dependent variable is assumed to have derivatives of all orders, with respect to the independent variable, over the domain of any one integration step.

c. The numerical approximation given by any method considered herein can theoretically be expressed as an infinite series. This infinite series differs from Taylor's series about a given data point only in the values of the constant coefficients.

d. The numerical methods will be restricted to two basic types. These types are the single-step methods such as Runge-Kutta, and the multistep methods consisting of explicit or implicit methods utilizing linear difference equations, such as Adams-Moulton, using equally spaced points. Development of these methods can be found in reference 2.

e. Only first order systems will be considered. Any errors, which might arise from integrating through or near a point where a unique bounded solution does not exist, will not be considered. The Taylor series, in restriction c above, is expanded about the single point for single step methods and one of the points contained in the difference equation for multistep methods. The series approximates the value of the dependent variable at a distance h away from this point.

Definition 2

The distance h shall be referred to as the integration step size.

Definition 3

A present set of data is a set of data which has been obtained by conversion from decimal to binary and/or integration of the differential equation by some numerical method from some initial set of data.

Roundoff error usually precludes the possibility of zero error in a present set of data. The error can be assumed to be zero and then the error of one integration step with respect to a present set of data can be measured.

Generally, the numerical methods used to solve differential equations are based upon a mathematical model which yields an erroneous solution for each integration step. One of the primary reasons for this is that many methods are based on an exact mathematical solution expressed by using infinite difference tables or infinite series. The numerical method is then based upon a finite process which differs from the true solution but agrees with the infinite process up to some desired term. This difference is usually referred to as truncation error. For the restrictive types of methods discussed in this paper, truncation error can be described as follows.

Given a method, a differential equation, a present set of data, and an integration step size, all satisfying restrictions a through e of Definition 1; then, (1) there exists a Taylor series expansion about the present value of the independent variable, and (2) there exists an expansion about the same point which is associated with the numerical method and differs from (1) only in the values of the constant coefficients.

Definition 4

Assuming that the present set of data is exact (zero error), then the local truncation error is the difference between the Taylor series expansion and the expansion associated with the method being used. This is a measure of the error introduced in a single integration step.

Assuming the data to be exact simply means that one is defining error with respect to the solution through the present point rather than with respect to the initial point. Note that after several integration steps there might exist a large amount of

error in the present set of data with respect to the initial data. Thus the local truncation error need not reflect the size of the total error with respect to the initial data.

For the remainder of this paper the term "truncation error" refers to local truncation error unless otherwise stated.

Definition 5

The value of the independent variable associated with a present set of data shall be referred to as the point P.

For example, if integrating from the point P to P + h, the truncation error for this step is usually approximated by using two (or more) approximations to the value of the dependent variable at P + h. The difference in these two values is then used as a criterion for determining the approximate truncation error for that step. If the error is acceptable, the dependent variable at the new point can be reinitialized and the integration procedure applied again. This is effectively the equivalent of expanding a new series about the point P + h.

Definition 6

The order of a method is the largest value of n for which the series associated with the method agrees with all terms in the Taylor series through the term containing the nth derivative. This is compatible with the definitions in reference 2.

Consider the infinite series for yN(x + h) associated with a given method at a point x.

$$yN(x + h) = S = y(x) + A(1)hy^{(1)}(x) + A(2)h^2y^{(2)}(x) + \ldots$$

where A(I) are the coefficients associated with the given method.

The Taylor series at the same point is

$$y(x + h) = T = y(x) + T(1)hy^{(1)}(x) + T(2)h^2y^{(2)}(x) + \ldots$$

where:

$$T(J) = 1/J!, \quad J = 1, 2, \ldots$$

Define

$$q(I) = T(I) - A(I), \quad I = 1, 2, \ldots$$

If a method is of order K then:

$$q(I) = 0, \quad I = 1, 2, \ldots, K$$

If the truncation error is not acceptable, there is the option of decreasing the magnitude of h or using a higher order method. It might be expected that the numerical solution becomes exact in the limit as h approaches zero (for a given fixed method and order). This would be true only if an infinite precision computer were used. It might also be expected that for a fixed step size, the error would tend to decrease as higher and higher order methods are used. Although truncation error might be made arbitrarily small by decreasing h, there is again the impeding effect of roundoff error. For example, suppose h is decreased until the numerical solution agrees with the true solution to within the least significant figure which can be carried in computation (for one integration step). To reduce h beyond this point would decrease the truncation error, but it could not possibly increase the number of accurate significant figures.

Truncation error can be decreased by using higher order methods; however, if multistep methods are used, numerical instability can cause the errors to increase markedly. (This will be demonstrated later.) Higher order single-step methods (for example, Runge-Kutta methods) involve the use of parameters of large magnitudes and the calculation of more intermediate quantities. This increases roundoff error, and computation time. Thus, a low-order, single-step method can possibly achieve better accuracy than a high-order, single-step method by using a smaller step size. Also, the low-order method might use less computation time because of the difference in the number of intermediate calculations for each integration step.

Definition 7

The relative truncation error is the numerical value of the local truncation error divided by the true solution with respect to the present set of data.

Relative truncation error is a measure of the number of significant accurate figures. The transition from absolute error to relative error is essential for clarity in some of the subsequent numerical examples. Consider a method (using 20 significant decimal digits) of calculating exp(x) such that the answer is always accurate through 19 significant figures. As x increases without bound, the absolute error becomes unbounded. From this viewpoint, the process is unstable. The relative error is bounded in magnitude by 10^{-18}. In this sense, the process is stable.

At this point, the behavior of the errors discussed could be encountered in evaluating a definite Riemann integral. Beyond this point, the errors discussed present no problem in dealing with derivatives which are functions of the independent variable alone.

CONCEPT OF STABILITY

Let $y(t)$ be the exact solution to the differential equation $y' = f(y, t)$ with the initial condition $y(t_0) = y_0$. Let $x(t)$ be an approximate solution to the same differential equation using the approximate initial condition $x(t_0) = x_0$.

Definition 8

The approximate solution $x(t)$ shall be defined as stable provided that for every positive number ϵ there exists a positive number δ such that if

$$|x(t_0) - y(t_0)| < \delta$$

then

$$|x(t) - y(t)| < \epsilon$$

for all $t > t_0$.

The above definition is one of the classical definitions of stability of a solution to an initial value problem. If the approximate solution is stable, then given a tolerance ϵ, regardless of how small, there exists a positive δ such that if the initial condition $x(t_0)$ can be determined to within a distance δ of the exact initial condition $y(t_0)$, then there is assurance that the error in $x(t)$ is less than ϵ in magnitude for all time t greater than t_0. The following is the negation of the definition of stability.

Definition 9

The approximate solution $x(t)$ is said to be unstable provided there exists a positive ϵ_0 such that for every positive δ, there exists some $t_1 > t_0$ and some $x(t_0)$ such that

$$| x(t_0) - y(t_0)| < \delta$$

and

$$| x(t_1) - y(t_1)| \geq \epsilon_0$$

Stability of a solution is a desirable characteristic; however, instability is not necessarily an undesirable characteristic. For example, suppose $|x(t) - y(t)|$ can be made arbitrarily small for all t such that $t_0 < t < t_2$. The solution $x(t)$ would be acceptable although it might be unstable by the above definition. It should be noted that some authors might define stability with respect to a finite interval rather than for all $t > t_0$.

In reality, almost all numerical solutions are unstable in the strict sense of the definition. For example, in computational work, all quantities are usually restricted to a fixed maximum number of significant figures. Thus, if $y(t) = 1.1$ at $t = 1$, and a computer having a maximum word length of 27 significant binary bits is used, then obviously, $|x(1) - y(1)| \geq 2^{-28}$ if the answer $x(1)$ is stored in a single word. Although this is an unstable solution, in many cases, if five or six significant decimal figures can be retained then the solution is considered satisfactory. Thus, even though a solution is unstable, it might be satisfactory in reality. From the engineering viewpoint, a solution may be considered stable if for a fixed tolerance ϵ_0 there is a particular δ_0 such that $|x(t) - y(t)| < \epsilon_0$ for every time t such that $t_0 < t < t_1$ provided $|x(t_0) - y(t_0)| < \delta_0$. In this case, ϵ_0 and δ_0 could be derived from a reasonable interpretation of the physical problem.

Many authors define stability in different ways; some leave the definition to the intuition of the reader. The term stability shall be used in this paper as formally defined at the beginning of this section.

In dealing with numerical solutions to ordinary differential equations, there is a natural division of stability problems into two basic categories. The first deals with the properties of the differential equation and will be referred to as "dynamic stability" or "mathematical stability." This form of stability can be defined in terms of the differential equation with no reference to a numerical method. The second category is dependent upon the numerical method which is used to obtain the solution of the differential equation. It has been pointed out that, in the strict sense of the

definition, any numerical solution is unstable. This is not the form of numerical stability to be used here. Numerical stability shall be defined according to the formal definition of stability where the approximate solution $x(t)$ is the solution obtained by a numerical method, using infinite precision numerical computation. The last part of this paper discusses numerical stability of a particular class of methods for solving differential equations. The following sections of this paper give specific examples which show that instability errors can become very significant over finite domains.

A Typical Problem

Consider the difference in two methods of approach to the following problem.

$$dy/dx = f(x, y)$$

$$y = y_0 \text{ at } x = x_0$$

Find a function $F(x)$ such that $y = F(x)$ satisfies the above conditions.

Method 1. - Solve the differential equation for the family of solutions.

$$y(x) = G(x) + CA(x) \tag{1}$$

Solve for C using the initial conditions

$$C_0 = \left[y_0 - G(x_0)\right] / A(x_0) \tag{2}$$

Thus,

$$F(x) = G(x) + C_0 A(x) \tag{3}$$

Method 2. - Suppose a numerical integration method is used for one step of size h. Let us assume that after this integration step there exists a truncation error, a roundoff error, or both. This error, or the combination of the two errors, shall be designated by E. Thus the present set of data is:

$$Y = F(x_0 + h) + E \text{ at } x = x_0 + h \tag{4}$$

For simplicity, assume that no more errors are made in the numerical integration. The numerical solution will be equivalent to using the present set of data as initial conditions and solving for C in equation (1) above. Suppose C = C1.

[1]If the numerical solution is designated by FN, then

$$FN(x) = G(x) + C1 \ A(x) \tag{5}$$

Definition 10

ER(x) is the difference between the true solution (using the initial conditions) and the numerical solution of a differential equation. In this example we have

$$ER(x) = F(x) - FN(x) = \left(C1 - C_0\right) A\,x = CON \ A(x) \tag{6}$$

The major observation at this point is that only one error introduced upon the first integration step (or any other step) might very definitely affect the error in following steps. Interest shall be in two basic types of functions A(x):

a. Those which tend to amplify error.

b. Those which tend to damp out the error.

(A single function may be of type a over one domain, and type b over another domain.)

Definition 11

Error amplification of type a in Definition 10 shall be referred to as error due to the "inherent characteristics" of the differential equations.

Definition 12

Differential equations of type b in Definition 10, which are characterized by A(x), shall be referred to as being mathematically stable provided that $A(x) \to 0$ as x becomes unbounded, and A(x) bounded for all $x > x_0$.

---

[1]FN(x) is sometimes referred to as a solution corresponding to modified initial conditions.

Definition 13

RER(x) is the value of ER(x) divided by the true solution.

RER(x) shall be referred to as the relative error (of the numerical solution) with respect to the initial conditions.

EXAMPLE:

The following examples indicate the extent to which A(x) can affect accuracy. Consider the differential equations:

a. $dy/dx = e^x$

b. $dy/dx = y$

c. $dy/dx = 2e^x - y$

d. $dy/dx = -14e^x + 15y$

Consider each equation as having the same initial condition:

$$y = 1 \text{ at } x = 0$$

The solution to each equation is the same using this initial condition. The solution is $y = e^x$.

The general solutions are respectively:

a. $y = e^x + C_1$

b. $y = C_2 e^x$

c. $y = e^{x} + C_3 e^{-x}$

d. $y = e^{x} + C_4 e^{15x}$

with the given initial conditions:

$$C_1 = C_3 = C_4 = 0,\ C_2 = 1$$

The differential equations were integrated numerically in such a manner that the relative local truncation error magnitude was less than 10^{-10} for each integration step.[2] The equations were integrated separately using one method. The absolute errors in these numerical solutions are given in table I. Note that the first three solutions were well behaved. RER(x) is less than or equal to the allowable relative local truncation error per step. (See table II.) The fourth equation introduced extreme mathematical instability, forcing even the relative error to grow exponentially, as would be expected.

The relative errors which would be expected can be written as follows:

a. $R_1 = C_1 / e^{x}$

b. $R_2 = C_2 - 1 / e^{x}$

c. $R_3 = C_3 e^{-2x}$

d. $R_4 = C_4 e^{14x}$

[2] E ± n shall be used to denote $10^{\pm n}$ in most of the tables.

EXAMPLE (MORE GENERAL):

Consider the differential equation

$$dy/dx = (1 + B)e^{x} - By \tag{7}$$

$$y = 1 \text{ at } x = 0$$

$$B \geqq 0$$

The family of solutions to the above differential equation is

$$y = e^{x} + Ce^{-Bx} \tag{8}$$

The solution with the given initial conditions is

$$y = e^{x} \tag{9}$$

If integrated in the positive x direction, then truncation and roundoff errors are damped according to CON e^{-Bx}, where CON can be determined as in equation (6). Also, the larger the value of B, the faster the error is damped. Thus, as x increases, RER(x) approaches the allowable relative local truncation error. At least this is true if the errors discussed so far are the only errors present in the numerical integration.

Definition 14

In integrating from the point x_i to $x_i + h$, $ET(x_{i+1})$ shall designate the local truncation error for this step.

If a method is of order K, then

$$ET(x_{i+1}) = \sum_{n=K+1}^{\infty} q_n h^n y^n(x_i) \tag{10}$$

Definition 15

The numerical solution obtained for the dependent variable at $x = x_i$ shall be designated by $yN(x_i)$.

Assume that h is held constant. A bound will be derived for $RER(x)$ of equation (7).

$$y(x) = e^x + Ce^{-Bx} \tag{11}$$

$$y^n(x) = e^x + C(-1)^n B^n e^{-Bx} \tag{12}$$

$$ET(x_{i+1}) = \sum_{n=K+1}^{\infty} q_n h^n y^n(x_i) \tag{13}$$

$$ET(x_{i+1}) = \sum_{n=K+1}^{\infty} e^{x_i} q_n h^n + \sum_{n=K+1}^{\infty} e^{-Bx_i} C(x_i)(-1)^n B^n q_n h^n \tag{14}$$

where:

$$C(x_i) = \left[yN(x_i) - e^{x_i}\right] \Big/ e^{-Bx_i} = ER(x_i) \Big/ e^{-Bx_i}$$

Since $y = e^x$ is the solution corresponding to the exact initial conditions, it follows that:

$$ET(x_{i+1}) = e^{x_i} \sum_{n=K+1}^{\infty} q_n h^n + ER(x_i) \sum_{n=K+1}^{\infty} (-1)^n B^n q_n h^n \tag{15}$$

Definition 16

$RET(x_{i+1})$ shall designate $ET(x_{i+1})$ divided by the true solution with respect to the initial conditions.

For this case,

$$RET(x_{i+1}) = ET(x_{i+1}) \Big/ e^{x_i} \tag{16}$$

$$RET(x_{i+1}) = \sum_{n=K+1}^{\infty} q_n h^n + \left[ER(x_i) \Big/ e^{x_i} \right] \sum_{n=K+1}^{\infty} (-1)^n B^n q_n h^n \tag{17}$$

and

$$RET(x_{i+1}) = \sum_{n=K+1}^{\infty} q_n h^n + RER(x_i) \sum_{n=K+1}^{\infty} (-1)^n B^n q_n h^n \tag{18}$$

If B is held constant and assuming h is small enough such that the series are convergent, then

$$\sum_{n=K+1}^{\infty} (-1)^n B^n q_n h^n = \text{constant} \equiv CON$$

Also,

$$\sum_{n=K+1}^{\infty} q_n h^n = \text{constant} = R_0$$

(Note that $R_0 \equiv RER(x)$ for $B = 0$)

Thus,

$$RET(x_{i+1}) = R_0 + CON\ RER(x_i) \tag{19}$$

If in general, we have the true solution, yT(x), with exact initial conditions given by

$$yT(x) = G(x) + CA(x) \tag{20}$$

and at a point $x = x_i$

$$yN(x_1) = G(x_i) + C(x_i)\,A(x_i) \tag{21}$$

then,

$$^{3}RER(x_{i+1}) = \left[C - C(x_i)\right] A(x_{i+1}) / yT(x_{i+1}) + RET(x_{i+1}) \tag{22}$$

For the example in question,

$$yT(x) = e^{x} \tag{23}$$

$$A(x) = e^{-Bx}$$

$$C = 0$$

---

[3]This implies that RER(x) and RET(x) must have the same directional sense.

Thus,

$$RER(x_{i+1}) = -C(x_i)\, e^{-Bx_{i+1}} \Big/ e^{x_{i+1}} + RET(x_{i+1}) \tag{24}$$

By the triangle inequality,

$$\left| RER(x_{i+1}) \right| \leqq \left| C(x_i)\, e^{-Bx_{i+1}} \Big/ e^{x_{i+1}} \right| + \left| RET(x_{i+1}) \right| \tag{25}$$

$$\left| C(x_i)\, e^{-Bx_{i+1}} \Big/ e^{x_{i+1}} \right| = \left| \frac{\left[yN(x_i) - e^{x_i} \right] e^{-Bx_{i+1}}}{e^{-Bx_i}\, e^{x_{i+1}}} \right| \tag{26}$$

$$\frac{e^{-Bx_{i+1}}}{e^{-Bx_i}} < 1$$

$$e^{-x_{i+1}} \leqq e^{-x_i}$$

$$yN(x_i) - e^{x_i} = ER(x_i)$$

Thus,

$$\left| C(x_i)\, e^{-Bx_{i+1}} \Big/ e^{x_{i+1}} \right| \leqq \left| ER(x_i) \Big/ e^{x_i} \right| = \left| RER(x_i) \right| \tag{27}$$

Substituting equation (27) into equation (25),

$$\left|RER\left(x_{i+1}\right)\right| \leqq \left|RER\left(x_i\right)\right| + \left|RET\left(x_{i+1}\right)\right| \tag{28}$$

Substituting from equation (11),

$$\left|RER\left(x_{i+1}\right)\right| \leqq \left|RER\left(x_i\right)\right| + \left|RER\left(x_i\right) CON + R_0\right| \tag{29}$$

The final result for the error bound is

$$\left|RER\left(x_{i+1}\right)\right| \leqq \left(1 + \left|CON\right|\right) \left|RER\left(x_i\right)\right| + \left|R_0\right| \tag{30}$$

From equation (19), the necessary equation to evaluate CON is available.

$$CON = \left[RET\left(x_1\right) - R_0\right] \Big/ RER\left(x_0\right) \tag{31}$$

$$CON = \left[ET\left(x_1\right) \Big/ e^{x_1} - R_0\right] \Big/ RER\left(x_0\right) \tag{32}$$

EXAMPLE:

Suppose an error was purposely introduced into the initial conditions such that:

$$RER\left(x_0\right) = RER(0) = 20$$

This is a 2000 percent error.

$$yN(0) = 20 = e^0 + C\left(x_0\right) e^0$$

$$x_0 = 0$$

$$C(x_0) = 19$$

Thus the calculation of CON can be obtained by the following substitution into equation (31).

$$RET(x_1) = \left[yN(x_1) - \left(e^{x_1} + 19e^{-Bx_1} \right) \right] \Big/ e^{x_1}$$

For multistep methods, numerical instability errors are sometimes present. Numerical stability theory predicts exponential error growth if one integrates outside the region of stability.

The preceding theory has been applied to integration with a fourth order Runge-Kutta method. The numerical solution behaves exactly as predicted. This does not mean that the relative error is small; with a large value of B, the value for CON is very large. However, according to equation (30) the magnitude of the relative error at x_{i+1} is bounded by a linear function of the error at x_i and the theoretical results provide an error bound.

Actually, the behavior of the numerical solution can be reasonably predicted from knowledge of the first integration error alone. The results are quite interesting but too lengthy to include.

Table III demonstrates the results of integrating the equation over 70 integration steps using Runge-Kutta fourth order with $h = 0.1$ for each step.

The third column lists the results of integrating the equations.

$$B = 1 \quad y' = 2e^x - y$$

$$B = 2 \quad y' = 3e^x - 2y$$

$$\vdots$$

$$B = 19 \quad y' = 20e^x - 19y$$

Initial conditions for each equation were

$$y = 1 \text{ at } x = 0$$

The fourth column lists the results of integrating the same equations but using erroneous data for initial conditions.

$$y = 20 \text{ at } x = 0 \text{ thus } RER(0) = 20$$

Note that the errors were completely damped.

STABILITY OF MULTISTEP METHODS

Recent analysis has shown that multistep methods might be very undependable for integrating a system of simultaneous equations.

Numerical stability has more than one definition and many modes of application. The emphasis is to be placed upon the catastrophic errors rather than their cause. However, we shall point out that for multistep methods there is a direct correspondence between stability of the numerical solution and HBAR. The error of the numerical solution of the point $t = t_0 + mh$ has a term $E(t) = c\lambda^m$ where c is a constant, and λ is a function of HBAR. The "stability of the method" is defined using this error term if $|\lambda|$ is, respectively, less than, equal to, or greater than 1, then this error term tends to, respectively, zero, remains constant, or becomes unbounded as t increases without bound. The method is called strongly stable, weakly stable, or unstable. It should be pointed out that this does not imply that the numerical solution is stable. The total error is also determined by truncation error, mathematical stability errors, and perhaps many other factors all of which are not included in the term $c\lambda^m$. There are numerous texts and journal publications which give mathematical derivations for the case of a single differential equation (refs. 3, 4, 5, and 6). All of the present theory is based on linear error theory; thus, the example will be a linear equation which agrees exactly with the theory.

EXAMPLE:

Consider the previous differential equation

$$dy/dx = (1 + B)\, e^x - By \tag{33}$$

$$y = 1 \text{ at } x = 0$$

Define

$$HBAR = |h| \cdot \frac{\partial y'}{\partial y}$$

so that for

$$B = 1, 2, 3, \ldots \text{ and } h = 0.1$$

then,

$$HBAR = -0.1, -0.2, -0.3, \ldots$$

Stability theory predicts that numerical stability is a function of HBAR. The following tables indicate the sensitivity of Adams-Moulton methods[4] to the value of HBAR. Table IV shows results of integrating 70 steps with $h = 0.1$. Table V shows the results of integrating with $h = 0.01$. The results of the fourth order Adams-Moulton and the fourth order Runge-Kutta were compared. Both results were obtained with the same step size and the same number of integration steps.

Table VI is a more dynamic example of the catastrophic effect encountered when numerical instability errors occur. RER(7.0) can be obtained approximately by dividing through by 10^3.

The accuracy of the solutions in tables IV, V, and VI could have been improved by applying the Adams-Moulton corrector more than once at each integration step and successively substituting the corrected value of y into the corrector equation. The following equation was integrated from $x = 1.5$ to $x = 1.6$ with one integration step of $h = 0.1$.

$$dy/dx = 11e^x - 10y$$

with initial condition

$$y = 20 \text{ at } x = 0$$

The solution is:

$$y(x) = e^x + 19e^{-10x}$$

[4]Adams-Moulton corrector, Adams-Bashforth predictor - both the same order. Initial conditions were accurate to within 15 significant decimal digits.

The numerically integrated value for $y(1.6)$ was substituted back into the corrector equation for 10 iterations. All input points were exact to within roundoff error. Table VII gives the results for four different orders of Adams-Moulton integration.

Even after 10 iterations the sum of truncation and instability errors for the 14th order and 15th order methods were greater than that of the 3rd order method. Note that this is for one integration step. The error would grow much faster for the high-order methods as the number of integration steps increased. Table VIII shows the result of integrating the following equation from $x = 0$ to $x = 7.0$ with $h = 0.1$.

$$dy/dx = 20e^{x} - 19y$$

With initial condition

$$y = 1 \quad \text{at} \quad x = 0$$

Thus,

$$y(x) = e^{x} + Ce^{-19x}, \quad C = 0$$

Table VIII also shows a comparison for 1, 2, and 10 iterations.

If the corrector is used iteratively, then the difference between consecutive answers should be used as a criterion for determining the number of iterations. Choosing the number of iterations arbitrarily can result in hazardous errors as shown in column three of table VIII. This type of behavior was also present when four iterations were used at each integration step, but the errors using three iterations per step were about the same as using one iteration.

DISCUSSION AND RESULTS

In the absence of roundoff error, the error for any integration step could be made arbitrarily small. The truncation error could be made small by decreasing step size or by using higher order methods. If roundoff and truncation errors could be made arbitrarily small, then so could errors due to numerical and mathematical instability.

If mathematical instability errors are present, then the truncation error must be kept to a minimum at each integration step. If the step size is decreased, then more integration steps must be used thus yielding more roundoff error. If higher order methods are used, the coefficients in the integration formulas become larger thus causing more roundoff error. Roundoff error does not cause truncation error,

but it prevents effective control of mathematical instability by means of controlling truncation error.

If numerical instability errors are present, then a lower order method can be used or the step size can be decreased. If a lower order method is used, then the step size must be decreased to retain the same truncation error. Decreasing step size increases the number of integration steps which increases roundoff error.

CONCLUDING REMARKS

A low-order single-step method requires the use of a smaller integration step size than a higher order method if the truncation errors are to be the same for both methods. If a higher order multistep method is used, then numerical instability errors might cause the solution to be less accurate than that of the low-order method. Higher order single-step methods increase the size of the constant parameters used in computation and the number of times that the derivative must be calculated. This can increase roundoff error and the computation time required for the integration.

If HBAR is calculated at each integration step to eliminate numerical stability errors, then a higher order multistep method could be used successfully. However, for large systems of equations, this calculation becomes time consuming as it involves solving for the eigenvalues of a matrix and the zeros of polynomials. (For a development of this theory, see reference 7.) Thus, the use of high-order multistep methods introduces serious difficulty.

Manned Spacecraft Center
National Aeronautics and Space Administration
Houston, Texas, August 19, 1966
030-00-00-CF-72

REFERENCES

1. Murray, F. J. and Miller, K. S.: Existence Theorems for Ordinary Differential Equations. New York University Press, 1954.

2. Henrici, P.: Discrete Variable Methods in Ordinary Differential Equations. John Wiley & Sons, Inc., 1962.

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

4. Bellman, R.: Stability Theory of Differential Equations. McGraw-Hill Book Company, Inc., 1953.

5. Henrici, P.: Error Propagation for Difference Methods. John Wiley & Sons, Inc., 1963.

6. Fox, L.: Numerical Solutions of Ordinary and Partial Differential Equations. Addison Wesley Publishing Company, Inc., 1962.

7. Lea, R. N.: Stability of Multi-step Methods in Numerical Integration. NASA TN-D-2772, 1965.

TABLE I.- ABSOLUTE ERRORS

[Initial conditions for all equations: $y = 1$; $x = 0$
Solution to all equations: $y = e^x$]

Independent variable x	Absolute errors		
	Equation 1	Equation 2	Equation 3
3.0	0.134 E - 10	0.603 E - 11	0.538 E - 10
6.0	.283 E - 09	.734 E - 09	.108 E - 08
9.0	.571 E - 08	.271 E - 07	.217 E - 07
12.0	.114 E - 06	.791 E - 06	.435 E - 06
15.0	.230 E - 05	.203 E - 04	.875 E - 05
18.0	.463 E - 04	.518 E - 03	.176 E - 03
21.0	.929 E - 03	.124 E - 01	.353 E - 02
24.0	.187 E - 01	.289 E - 00	.709 E - 01
Equation 4			
0.752	0.309 E - 07		
1.5	.238 E - 02		
2.03	.621 E + 01		
2.36	.875 E + 03		
3.01	.165 E + 08		
4.04	.861 E + 14		
4.51	.975 E + 17		
5.07	.450 E + 21		
5.54	.509 E + 27		
6.0	.576 E + 27		

TABLE II. - RELATIVE ERRORS

Independent variable x	Relative errors		
	Equation 1	Equation 2	Equation 3
3.0	0.659 E - 12	0.294 E - 12	0.263 E - 11
6.0	.689 E - 12	.178 E - 11	.263 E - 11
9.0	.692 E - 12	.327 E - 11	.263 E - 11
12.0	.692 E - 12	.477 E - 11	.263 E - 11
15.0	.692 E - 12	.626 E - 11	.263 E - 11
18.0	.692 E - 12	.776 E - 11	.263 E - 11
21.0	.692 E - 12	.925 E - 11	.263 E - 11
24.0	.692 E - 12	.107 E - 10	.263 E - 11
Equation 4			
0.377	0.76 E - 10		
.752	.146 E - 07		
1.5	.530 E - 03		
2.03	.818 E - 00		
2.36	.829 E + 02		
3.01	.810 E + 06		
4.04	.151 E + 13		
4.51	.106 E + 16		
5.07	.281 E + 19		
5.54	.199 E + 22		
6.0	.141 E + 25		
6.48	.999 E + 27		

TABLE III. - MATHEMATICALLY STABLE SOLUTIONS

[Runge-Kutta]

B	$h \cdot \partial f/\partial y$	RER (7. 0)	RER (7. 0) when introducing 2000 percent error on 1st step
1	-0.1	0.81 E - 06	0.16 E - 04
2	-.2	.35 E - 05	.36 E - 05
3	-.3	.88 E - 05	.88 E - 05
4	-.4	.17 E - 04	.17 E - 04
5	-.5	.28 E - 04	.28 E - 04
6	-.6	.43 E - 04	.43 E - 04
7	-.7	.62 E - 04	.62 E - 04
8	-.8	.86 E - 04	.86 E - 04
9	-.9	.11 E - 03	.11 E - 03
10	-1.0	.15 E - 03	.15 E - 03
11	-1.1	.19 E - 03	.19 E - 03
12	-1.2	.23 E - 03	.23 E - 03
13	-1.3	.29 E - 03	.13 E - 03
14	-1.4	.36 E - 03	.36 E - 03
15	-1.5	.43 E - 03	.43 E - 03
16	-1.6	.53 E - 03	.53 E - 03
17	-1.7	.63 E - 03	.63 E - 03
18	-1.8	.76 E - 03	.76 E - 03
19	-1.9	.92 E - 03	.92 E - 03

TABLE IV.- ADAMS-MOULTON, ADAMS-BASHFORTH

[RER(x) evaluated at x = 7.0 after 70 integration steps of h = 0.1]

HBAR	1st order RER (x)	2nd order RER (x)	3rd order RER (x)	4th order RER (x)
0.0	0.83 E - 03	0.40 E - 04	0.24 E - 05	0.16 E - 06
-.1	.54 E - 03	.28 E - 04	.18 E - 04	.13 E - 06
-.2	.45 E - 03	.24 E - 04	.16 E - 05	.12 E - 06
-.3	.41 E - 03	.23 E - 04	.16 E - 05	.12 E - 06
-.4	.40 E - 03	.23 E - 04	.16 E - 05	.12 E - 06
-.5	.40 E - 03	.23 E - 04	.16 E - 05	.12 E - 06
-.6	.40 E - 03	.23 E - 04	.16 E - 05	.13 E - 06
-.7	.42 E - 03	.24 E - 04	.17 E - 05	.13 E - 06
-.8	.44 E - 03	.25 E - 04	.18 E - 05	.14 E - 06
-.9	.47 E - 03	.26 E - 04	.18 E - 05	.14 E - 06
-1.0	.50 E - 03	.28 E - 04	.19 E - 05	.15 E - 06
-1.1	.55 E - 03	.29 E - 04	.20 E - 05	.64 E - 07
-1.2	.60 E - 03	.32 E - 04	.22 E - 05	.79 E - 05
-1.3	.68 E - 03	.34 E - 04	.23 E - 05	.38 E - 03
-1.4	.78 E - 03	.37 E - 04	.24 E - 05	.13 E - 01
-1.5	.93 E - 03	.41 E - 04	.78 E - 05	.32 E - 00
-1.6	.11 E - 02	.46 E - 04	.12 E - 03	.57 E + 01
-1.7	.15 E - 02	.53 E - 04	.18 E - 02	.71 E + 02
-1.8	.22 E - 02	.61 E - 04	.71 E - 03	.54 E + 03
-1.9	.41 E - 02	.67 E - 04	.34 E - 00	.35 E + 03

TABLE V. - ADAMS-MOULTON, ADAMS-BASHFORTH

HBAR	12th order RER(.7)	13th order RER(.7)	14th order RER(.7)	15th order RER(.7)
0.0	0.11 E - 15	0.99 E - 15	0.18 E - 14	0.35 E - 14
-.01	.12 E - 14	.77 E - 15	.13 E - 14	.30 E - 14
-.02	.12 E - 14	.00 E - 15	.99 E - 15	.30 E - 13
-.03	.14 E - 14	.11 E - 15	.27 E - 14	.99 E - 13
-.04	.14 E - 14	.11 E - 15	.13 E - 14	.16 E - 12
-.05	.12 E - 14	.11 E - 15	.17 E - 14	.22 E - 15
-.06	.99 E - 15	.55 E - 15	.11 E - 15	.19 E - 14
-.07	.14 E - 14	.44 E - 15	.66 E - 15	.63 E - 13
-.08	.16 E - 14	.77 E - 15	.14 E - 14	.17 E - 11
-.09	.12 E - 14	.38 E - 14	.12 E - 12	.14 E - 11
-.10	.55 E - 15	.95 E - 14	.19 E - 11	.11 E - 09
-.11	.25 E - 14	.10 E - 12	.15 E - 10	.31 E - 09
-.12	.12 E - 13	.65 E - 12	.32 E - 10	.78 E - 08
-.13	.29 E - 13	.20 E - 12	.31 E - 09	.38 E - 07
-.14	.26 E - 12	.10 E - 10	.27 E - 08	.59 E - 07
-.15	.17 E - 11	.40 E - 10	.51 E - 08	.12 E - 05
-.16	.72 E - 11	.49 E - 09	.56 E - 07	.64 E - 05
-.17	.30 E - 12	.87 E - 09	.50 E - 06	.26 E - 04
-.18	.59 E - 10	.71 E - 08	.15 E - 05	.64 E - 04
-.19	.42 E - 09	.15 E - 07	.43 E - 05	.74 E - 04

TABLE VI. - 14th ORDER ADAMS-MOULTON VERSUS 2nd ORDER ADAMS-MOULTON

[Both sets are using same step size, same initial conditions, same number of integration steps]

HBAR	14th order ER(7.0)	2nd order ER(7.0)
0.0	0.37 E - 11	0.44 E - 01
-.1	.10 E - 10	.31 E - 01
-.2	.41 E - 03	.27 E - 01
-.3	.19 E + 02	.25 E - 01
-.4	.43 E + 05	.25 E - 01
-.5	.16 E + 08	.25 E - 01
-.6	.18 E + 10	.25 E - 01
-.7	.40 E + 11	.26 E - 01
-.8	.34 E + 13	.27 E - 01
-.9	.20 E + 15	.29 E - 01
-1.0	.38 E + 16	.30 E - 01
-1.1	.94 E + 16	.32 E - 01
-1.2	.53 E + 18	.35 E - 01
-1.3	.73 E + 19	.38 E - 01
-1.4	.27 E + 20	.41 E - 01
-1.5	.18 E + 21	.45 E - 01
-1.6	.23 E + 22	.51 E - 01
-1.7	.81 E + 22	.58 E - 01
-1.8	.10 E + 23	.60 E - 01
-1.9	.18 E + 24	.74 E - 01

TABLE VII. - ADAMS-MOULTON, ADAMS-BASHFORTH INTEGRATION WITH ITERATIONS ON THE CORRECTOR

$\left[|ER(x)| \text{ at } x = 1.6 \quad h = 0.1 \text{ one step}\right]$

Number of iterations	2nd order	3rd order	14th order	15th order
1	$0.87\ 10^{-4}$	$0.37\ 10^{-5}$	$0.50\ 10^{-3}$	$0.82\ 10^{-3}$
2	$.17\ 10^{-4}$	$.45\ 10^{-6}$	$.11\ 10^{-3}$	$.19\ 10^{-3}$
3	$.26\ 10^{-4}$	$.10\ 10^{-5}$	$.41\ 10^{-4}$	$.67\ 10^{-4}$
4	$.81\ 10^{-5}$	$.51\ 10^{-6}$	$.41\ 10^{-7}$	$.45\ 10^{-6}$
5	$.15\ 10^{-4}$	$.73\ 10^{-6}$	$.11\ 10^{-4}$	$.17\ 10^{-4}$
6	$.13\ 10^{-4}$	$.64\ 10^{-6}$	$.79\ 10^{-5}$	$.12\ 10^{-4}$
7	$.14\ 10^{-4}$	$.67\ 10^{-6}$	$.87\ 10^{-5}$	$.13\ 10^{-4}$
8	$.13\ 10^{-4}$	$.67\ 10^{-6}$	$.85\ 10^{-5}$	$.13\ 10^{-4}$
9	$.14\ 10^{-4}$	$.67\ 10^{-6}$	$.85\ 10^{-5}$	$.13\ 10^{-4}$
10	$.13\ 10^{-4}$	$.67\ 10^{-6}$	$.85\ 10^{-5}$	$.13\ 10^{-4}$

TABLE VIII.- ADAMS-MOULTON, ADAMS-BASHFORTH INTEGRATION WITH ITERATIONS ON THE CORRECTOR

[RER(x) at x = 7.0 - h = 0.1 - 70 steps]

Order	One iteration	Two iterations	Ten iterations
1	$0.42\ 10^{-2}$	$0.84\ 10^{15}$	$0.21\ 10^{-2}$
2	$.67\ 10^{-4}$	$.16\ 10^{19}$	$.30\ 10^{-4}$
3	$.34\ 10^{0}$	$.31\ 10^{22}$	$.12\ 10^{-5}$
4	$.85\ 10^{3}$	$.10\ 10^{25}$	$.11\ 10^{-2}$
5	$.21\ 10^{7}$	$.82\ 10^{26}$	$.19\ 10^{1}$
6	$.66\ 10^{10}$	$.19\ 10^{28}$	$.12\ 10^{3}$
7	$.19\ 10^{12}$	$.17\ 10^{29}$	$.29\ 10^{5}$
8	$.36\ 10^{14}$	$.67\ 10^{29}$	$.90\ 10^{6}$
9	$.34\ 10^{15}$	$.14\ 10^{30}$	$.16\ 10^{8}$
10	$.15\ 10^{16}$	$.17\ 10^{30}$	$.99\ 10^{8}$
11	$.35\ 10^{16}$	$.60\ 10^{29}$	$.16\ 10^{9}$
12	$.53\ 10^{17}$	$.22\ 10^{30}$	
13	$.18\ 10^{18}$	$.36\ 10^{30}$	
14	$.18\ 10^{20}$	$.30\ 10^{31}$	
15	$.31\ 10^{20}$	$.79\ 10^{32}$	

NASA-Langley, 1966 S-116

"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: Technical information generated in connection with a NASA contract or grant and released under NASA auspices.

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

TECHNICAL REPRINTS: Information derived from NASA activities and initially published in the form of journal articles.

SPECIAL PUBLICATIONS: Information derived from or of value to NASA activities but not necessarily reporting the results of individual NASA-programmed scientific efforts. Publications include conference proceedings, monographs, data compilations, handbooks, sourcebooks, and special bibliographies.

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