N O T I C E

THIS DOCUMENT HAS BEEN REPRODUCED FROM MICROFICHE. ALTHOUGH IT IS RECOGNIZED THAT CERTAIN PORTIONS ARE ILLEGIBLE, IT IS BEING RELEASED IN THE INTEREST OF MAKING AVAILABLE AS MUCH INFORMATION AS POSSIBLE

JSC-16302

Applying Integrals of Motion to the Numerical Solution of Differential Equations

Mission Planning and Analysis Division

December 1979

NASA
National Aeronautics and Space Administration

Lyndon B. Johnson Space Center
Houston, Texas

79-FM-47

JSC-16302

SHUTTLE PROGRAM

APPLYING INTEGRALS OF MOTION TO THE NUMERICAL SOLUTION OF DIFFERENTIAL EQUATIONS

By Donald J. Jezewski · FM15
Software Development Branch

Approved: Elric N McHenry
Elric N. McHenry, Chief
Software Development Branch

Approved: Ronald L. Berry
Ronald L. Berry, Chief
Mission Planning and Analysis Division

Mission Planning and Analysis Division

National Aeronautics and Space Administration

Lyndon B. Johnson Space Center

Houston, Texas

December 1979

CONTENTS

FIGURES

1.0 SUMMARY

In this document, a method is developed for using the integrals of systems of nonlinear, ordinary, differential equations in a numerical integration process to (1) control the local errors in these integrals and (2) reduce the global errors of the solution. The method is general and can be applied to either scalar or vector integrals. A number of example problems , with accompanying numerical results, are used to verify the analysis and support the conjecture of global error reduction.

2.0 INTRODUCTION

Whittaker (ref. 1) defines the integral of a system of differential equations as a function of the state and time when the total time derivative is zero and the state variables are any functions of time. In solving nonlinear differential equations by a numerical integration process, these integrals will be used along with their associated numerical errors to (1) control the local errors in these integrals and (2) reduce the global error of the solution.

Integrals of systems of differential equations are constraints on the solution and, as such, can be used to reduce the number of degrees of freedom in the problem. Topologically, as more integrals of a system are introduced into the problem, the solution is constrained to lie on a larger manifold of the solution space with a resulting reduction in the global errors. The limiting case is, of course, an analytic solution to the differential equations where the solution is known at any time and the integral and global error are zero.

The direct approach (analytical substitution) to using integrals of systems of differential equations to reduce the number of degrees of freedom in a problem and to reduce the errors introduces other types of difficulties (such as singularities and switching logic in the remaining unsolved equations). Invariably, the overhead of calculating the right-hand side (RHS) of the remaining differential equations increases significantly; thus, any overall advantage of such an approach is nullified.

In a direct analytical approach, Szebehely (ref. 2) linearized certain nonlinear differential equations by utilizing the integrals of the system and by introducing a new independent variable. This was an effort to formalize the results of Steifel and Scheifele (ref. 3), Burdet (ref. 4), Szebehely (ref. 5), and Sperling (ref. 6) who attempted to linearize, and in some cases stabilize, certain nonlinear differential equations by transformations of the dependent and independent variables of the problem.

A direct numerical approach was made by Nacozy (ref. 7). The numerical errors in the integrals of the system were used to rectify the solution at each integration step in an attempt to stabilize the solution and reduce the errors. For some integrals, a linear expansion was necessary to compute the correction vector. Using a fourth-order predictor-corrector integration routine with a variable stepsize, global error reduction of two or three orders of magnitude was obtained.

An optimization technique is the most general, indirect approach to controlling the error in a system described by nonlinear differential equations. A performance function is defined as a function of the error in the integrals, and the differential equations of the system are adjoined to this performance function by Lagrange multipliers. Through the use of variational calculus or some other similar technique, differential equations for the multipliers and optimality and boundary conditions can be developed. However, this is not an advisable procedure for controlling the errors for the following reasons: (1) the number of degrees of freedom in the problem typically increases twofold, (2) an iteration procedure is introduced to satisfy the developed boundary conditions, and (3) the overhead in the solution of the problem increases significantly. However, the error in the integral would be minimized and would be indirectly used to reduce the global error of the solution.

A functional, indirect approach to introducing integrals of systems in the solution of differential equations was advanced by Baumgarte. In a number of studies (refs. 8 through 10) he used the integrals of the system to stabilize certain nonlinear differential equations and reduce the errors. The procedure was based on the principle of adjoining a form of the constraint (the coefficient of the second derivative) by a Lagrange multiplier to the original second-order system of differential equations. The technique was applied to several problems and the results were encouraging. The error in the integral (almost always the energy) or constraint was substantially reduced by using this control process. However, several characteristics of these studies were disappointing: (1) global error results of the solution were almost always absent although some analytic solutions were available, (2) the technique required a particular formulation for the problem, (3) the parameters that were introduced were not mathematically defined, and (4) a lack of generality existed when applying the approach to any system of equations and constraints.

This document does not pretend to advance the best technique for using and controlling the errors in the integrals of a system of differential equations. It does propose a general technique for incorporating the integrals of a system of nonlinear differential equations and their associated errors in a process that will control the integral errors and reduce the global error of the solution. The process requires no increase in the number of degrees of freedom in the solution. The technique has two disadvantages: (1) it requires some premathematical analysis to formulate the control vector, and (2) it generates some additional overhead and complexity in the solution.

3.0 STATEMENT OF PROBLEM

This study examines the numerical solution of a first-order, nonlinear system of ordinary differential equations of the form

$$\dot{X} = F(X,t) \qquad X(0) = X_0 \text{ at } t = 0 \tag{1}$$

which possess integrals of motion of the form

$$J(X,t) = K \tag{2}$$

where X and F are n-vectors and J and K are m-vectors; $m < n$ and K is a constant defined by the initial conditions. The system of equation (1) is assumed not to be analytically integrable in terms of known functions and, hence, must be solved by a numerical integration process. Specifically, this method presents a numerical solution to equation (1) with the additional criteria that the solution lies "arbitrarily close" to the (m + 1)-dimensional manifold described in equation (2).

One obvious solution to this problem is to use equation (2) to eliminate m of the X's and thus reduce the problem to integrating only the remaining (n-m)-first-order differential equations. However, past experiences and the examination of a particular system of differential equations and their associated integrals indicate that this is not an advisable procedure. There is information about the solution embedded in equation (2), and it should not be used only as a check on the numerical integration of equation (1), but it should be used either directly or indirectly in the solution to reduce the errors and possibly the number of degrees of freedom in the problem. For example, if a problem-free procedure could be devised for introducing the m integrals of motion into a system of n differential equations, and the m integrals were introduced into the solution one at a time, the number of degrees of freedom in the solution would be reduced by one each time, requiring the solution to lie on a manifold with a dimension increasing by one. If there were actually n integrals of motion, then as m approaches n, the solution would be constrained to a larger subspace of the problem, and would also have global errors that tend to zero-vanish when m equals n.

Because the direct approach to using the integrals of motion in the solution of a system of differential equations introduces other mathematical difficulties, an indirect approach shall be proposed to solve equation (1) while attempting to satisfy the constraints expressed by the integrals of the motion.

4.0 SOLUTION WITH ERRORS

If equation (1) is solved by a numerical integration process, the solution will certainly contain errors. Consider these errors as due to the inability of the numerical integration process to correctly evaluate the right-hand side (RHS) of the differential equations. Defining the numerical errors in the RHS from the integration process as δF, the differential equation (eq. (1)) can be expressed as

$$\dot{X} = F(X,t) + \delta F \tag{3}$$

where at the initial time ($t = 0$), $\delta F(0) = 0$. The vector $F(X,t)$ in equation (3) is the exact representation of the RHS of the differential equations. Now consider the term δF as a perturbation to the original system of equation (1). The integrals of motion (eq. (2)) are not conserved (K is not equal to a

constant) and the differential equations for their rate of change can be expressed as

$$\dot{K} = G(X, \delta F, t) \tag{4}$$

where G is a vector function of the indicated arguments. The numerical error in the integral of motion is defined as

$$\epsilon = K - K_o \quad , \quad K_o = K(0) \tag{5}$$

The differential equation for the time rate of change of the error is simply

$$\dot{\epsilon} = \dot{K} \tag{6}$$

However, δF is not known (except at the initial time) since the exact solution of the RHS in equation (1), $F(X,t)$ is not known. Hence, it appears that this development is an interesting but insignificant exercise.

5.0 SOLUTION WITH CONTROL

In this section, a solution philosophy used in optimal control theory (ref. 11) will be adopted. In controlling a system, it is fundamental to have a process that is (1) observable and (2) controllable. The first criterion is certainly fullfilled, for it is noted in the numerical integration process that the value of the integral is not constant but grows in some manner characteristic to the particular numerical integrator, stepsize, etc. The second criterion, however, is not fullfilled.

The process is not controllable because the error vector δF is an unknown output of the numerical integration process and not an input.

A control vector λ (an n-vector) is added to the RHS of equation (1) (the exact equation) in an attempt to control the numerical error δF of the integration process.

$$\dot{X} = F(X,t) + \lambda \tag{7}$$

The justification for the addition of the vector λ to the original equation (eq. (1)) is: if m integrals of the motion or constraints exist to a system of n equations, then the solution to this system of equations is constrained to lie on a m-dimensional manifold of the solution space and there are only $n-m$ independent degrees of freedom in the problem. The introduction of

a control vector λ is an attempt to numerically reduce the n-order dependent system of equations to a (n-m)-order independent system of equations.

The introduction of the control vector λ to the RHS of the exact equation (eq. (1)) implies that the numerically integrated equation (eq. (3)) has a similar term added to its RHS. Of course, it is understood that with the addition of the control vector λ, the term δF appearing in equation (3) will have a different meaning since the error vector will certainly be disturbed by the introduction of this control vector. The introduction of the control vector λ is an attempt to introduce a term to cancel all or part of the error vector δF such that when the numerical integration process is applied to equation (7), the error (defined in eq. (5)) will be arbitrarily close to zero. Thus, if

$$\delta F + \lambda = 0 \tag{8}$$

and equations (5) and (6) are used to obtain a "stable" solution for the control vector λ, the numerical integration process will produce a value of the state X, which nulls the error. Also, since the number of degrees of freedom in the system of equations to be integrated have been numerically reduced, it is conjectured that the global error of the solution will also be reduced. A solution for the n-vector λ from the system of m-constraint equations must now be obtained.

Since the error rate and the error are now controllable, a stable differential equation for the desired functional relationship between these two errors is introduced as

$$\dot{\epsilon} = -\gamma\epsilon \tag{9}$$

where γ is a positive function (defined in the appendix). Now, any error arising in the integral of motion due to a dissatisfaction of equation (5) will be critically damped by the control vector λ obtained from equation (9). Using equations (4), (5), (6), and (8) in equation (9) gives

$$G(X,-\lambda,t) = -\gamma\left(K(X,t) - K_o\right) \tag{10}$$

Since the vector λ must span-the-space defined by the state vector X, a solution for λ is assumed to be

$$\lambda = \Lambda X \tag{11}$$

where Λ is an undefined matrix of appropriate dimensions. Using equation (11) in equation (10) yields

$$G(X, -\Lambda X, t) = -\gamma \left(K(X,t) - K_o\right) \tag{12}$$

The problem of controlling the integral errors is reduced to determining a solution of an algebraic equation. Any solution for the undefined matrix Λ that satisfies equation (12) will produce a control vector (from eq. (11)) that, when used in the numerical integration process, will control the errors in the integrals of motion. The difficulty in determining a matrix Λ that satisfies equation (12) is, of course, problem dependent. In two of the example problems (linear oscillator and two-body problem) with a scalar integral of motion (energy), the matrix Λ was obtained by inspection. In a third example (two-body problem with an angular momentum integral), some manipulation was required to obtain a matrix Λ that satisfied equation (12). In a final example, a solution is developed to the two-body problem when both the energy and the angular momentum integral errors are present.

6.0 ONE-DIMENSIONAL HARMONIC OSCILLATOR

The first-order linear, differential equations describing the one-dimensional harmonic oscillator state are

$$\begin{aligned} \dot{X}_1 &= X_2 \\ \dot{X}_2 &= -X_1 \end{aligned} \tag{13}$$

Since an analytic solution to this problem exists, there are two integrals of motion. For this exercise, however, it is assumed that equation (13) cannot be solved analytically and that only one integral of motion (energy) exists and is defined as

$$J(X) = 1/2\ X^T X = k$$

where the superscript ^{T} refers to the transpose.

The numerical error in the integral is defined as

$$\varepsilon = 1/2\ X^T X - k_o \qquad , k_o = k(0) \tag{14}$$

Adding the control vector λ to equation (13), the controlled equation to be integrated is

$$\dot{X}_1 = X_2 + \lambda_1$$
$$\dot{X}_2 = -X_1 + \lambda_2 \tag{15}$$

Developing the total time derivative of the integral of motion, using equation (15) yields

$$\dot{k} = G(X,\lambda) = X^T\lambda \tag{16}$$

A stable differential equation for the functional relationship between the error and error rates is defined as

$$\dot{\epsilon} = -\gamma\, \epsilon \tag{17}$$

where γ is a positive scalar function. From equations (14), (16), and (17), the control vector λ is required to satisfy the following equation.

$$X^T(\lambda + \frac{\gamma}{2} X) = \gamma\, k_o \tag{18}$$

If γ were a constant, equation (18) would represent a new integral of motion; however, one that is functionally dependent on the control vector.

A solution is assumed for the control of the form

$$\lambda + \frac{\gamma}{2} X = \gamma k_o\, \alpha X \tag{19}$$

where α in an undefined coefficient. Using equation (19) in equation (18), a necessary condition is

$$\alpha\, X^T X = 1$$

One value of the coefficient satisfying the equation is

$$\alpha = \frac{1}{2k}$$

which results in a control vector (from eq. (19)) of

$$\lambda = -\frac{\gamma}{2}\frac{\varepsilon}{k} X \tag{20}$$

As expected, the control vector for the energy integral error is in the form of a feedback control law directly proportional to the error. It should be noted that the vector λ is never singular unless the energy constant is zero (trivial case).

6.1 COMPARATIVE RESULTS

Equation (15) was integrated with a fixed step, fourth order, Runge-Kutta integrator using the control vector defined by equation (20). Since the period of the uncontrolled solution is equal to 2π, solutions were obtained for a constant integration stepsize defined as

$$h = 2\pi/N$$

where N is the number of steps in the integration process. The function γ was determined from a solution of the equation $\varepsilon(t + h) = 0$ at each integration step (see appendix). For this linear problem, the function γ for each value of N was found to be a constant.

In the solution, it was noted that the uncontrolled error in the energy grew in a linear manner with time. The controlled error remained essentially constant at a value five to ten orders of magnitude less (depending on the value of N and t) than the uncontrolled error. Thus, the method described in the analysis of controlling the error in the integral of motion appears to be valid for the linear problem.

Because the analytic solution to equation (13) is known, the global error of the numerically integrated solution can also be computed. The global error at a given time is defined as

$$\Delta X = |X^I - X^A| \tag{21}$$

where the superscripts I and A on the vector X refer to the integrated and analytic values, respectively. The global error of the controlled solution was found to be always less than that of the uncontrolled solution. This indicates that the correct information from the error in the integral of motion was entering the solution via the control vector λ. However, although the integral errors were reduced substantially by the control, the global error showed only an infinitesimal reduction. This agrees with the results reported by Nacozy (ref. 7) for the harmonic oscillator problem.

Conclusion: For this linear and stable problem, controlling the error in the integral of motion has only a negligible effect on the global error of the solution.

7.0 TWO-BODY PROBLEM WITH ENERGY INTEGRAL

The first-order, nonlinear differential equations describing the two-body problem are

$$\dot{R} = V \tag{22a}$$

$$\dot{V} = -\mu \frac{R}{r^3} \tag{22b}$$

where R and V are the position and velocity vectors (respectively), μ is the gravitational constant and $r = |R|$. There are three integrals of motion to this system of equations (not all independent): two vector integrals - Laplace and angular momentum, and one scalar integral - energy. This example is concerned only with the energy integral that is formally obtained by scalar multiplying equation (22a) by $\dot{V}$, and equation (22b) by $\dot{R}$, then taking the difference and noting the exact differential. This integral can be expressed as

$$J(R,V) = 1/2\ V^T V - \mu/r = k \tag{23}$$

where k is a constant of the motion defined by the initial conditions. If equation (22) is solved with a numerical integrator, the solution will contain errors. This numerical error in the energy integral is defined as

$$\varepsilon = k - k_o \qquad , k_o = k(0) \tag{24}$$

and the differential equation for its time rate of change is defined as

$$\dot{\epsilon} = \dot{k}$$

Now a control vector

$$\lambda^T = (\lambda_R{}^T, \lambda_V{}^T)$$

is added to the RHS of equation (22) in an attempt to control the numerical error in the integral of motion. For convenience of notation, a state vector is defined as

$$S^T = \left(\frac{\mu R^T}{r^3} , V^T \right)$$

and an associated matrix is defined as

$$A = \begin{pmatrix} -\frac{r^3 I}{\mu} & 0 \\ 0 & 1/2\ I \end{pmatrix}$$

where 0 and I are appropriate null and identity matrices, respectively. Using these identities, the numerical error and error rate can be expressed as

$$\epsilon = S^T A S - k_o$$

$$\dot{\epsilon} = -\lambda^T S \qquad (25)$$

The stable differential equation (eq. (17)) is now used for the functional relationship between the error in the integral of motion and its time rate of change. Using equation (25) in equation (17) yields

$$S^T(\lambda - \gamma\ AS) = -\gamma k_o \qquad (26)$$

Since the vector λ must span the space defined by the state vector S, an assumed solution for λ is

$$\lambda - \gamma\ AS = -\gamma k_o\ \Lambda\ S \qquad (27)$$

where Λ is an undefined matrix. Using equation (27) in equation (26), a necessary condition is

$$S^T \Lambda S = 1 \tag{28}$$

Comparing equations (28) and (24), it is noted that one solution for the undefined matrix Λ is

$$\Lambda = \frac{A}{k}$$

which results in a control vector (from eq. (27)) of

$$\lambda = \gamma \frac{\varepsilon}{k} AS \tag{29}$$

Thus, the control vector for the energy integral of the two-body problem can be cast in the same form as that of the control vector for the harmonic oscillator problem. It is only singular for the special case when the energy constant is zero (parabolic orbit).

7.1 NUMERICAL RESULTS

Equation (22) with the added control vector defined by equation (29) was integrated with the same processor as the previous example. However, the constant stepsize was determined as

$$h = \frac{P}{N}$$

where P is the period of the orbit defined by the initial conditions.

Figure 1 illustrates a comparison of controlled and uncontrolled solutions for a two-body problem. The solutions were obtained using a fourth-order, Runge-Kutta integrator with a constant stepsize of N. The abscissa label of NORBIT refers to the number of orbits the equations were integrated, while the ordinate is the logarithm to the base 10 of either the absolute value of the position error or velocity error. (These two errors were approximately equal for this problem.) Hence, the ordinate is a measure of the number of significant digits

of information remaining in the solution at a given time (a measure of the global error of the solution). The solid curves in figure 1 refer to solutions obtained with the addition of the control vector λ, while the dashed curves refer to no control.

In figure 1(a), the initial conditions were specified by a circular orbit that is an orbit with an eccentricity, $e = 0$ and semimajor axis $a = 1$. For $N = 20$, the uncontrolled solution has lost all information content at NORBIT equals 20, whereas the controlled solution still retains approximately two digits of accuracy. For $N = 40$, the controlled solution has approximately two aditional digits of accuracy at NORBIT equals 40. For a given N, the two solutions are diverging; that is, the uncontrolled solutions have a steeper slope or are approaching zero more rapidly than the controlled solutions.

In figure 1(b) and 1(c), similar results are obtained for initial orbits with eccentricities of $e = 0.1$ and $e = 0.2$, respectively . For $N = 20$, the uncontrolled solution for $e = 0.1$ and $e = 0.2$ has zero-significant digits of accuracy at NORBIT = 15 and 9, respectively.

When overlaying figures 1(a) and 1(b), it is noted that the controlled solution shows almost no loss in accuracy for this change in eccentricity, whereas the uncontrolled solution shows significant losses. Similar (but not as significant) results are obtained when overlaying figures 1(b) and 1(c).

The errors in the integral of motion (energy) in figures 1(a) through 1(c) behave similar to that of the harmonic oscillator problem. The uncontrolled integral error grows almost linearly with time; the slope is determined by the initial conditions and the number of steps N. The controlled integral error remains nearly constant in value, with a number of orders in magnitude less than the uncontrolled solution. The actual number of orders in magnitude depends on the initial conditions, integration stepsize, and integration time.

In figure 2, the γ function is plotted versus time for one orbit for the three sets of initial conditions defined by $e = 0.0$, 0.1, and 0.2. For $e = 0.0$, which is similar to the previous linear problem, the γ function is nearly constant. The γ functions for $e \neq 0$ are periodic in nature. For $e = 0.2$, there is a portion of the solution where $\gamma < 0$. Since there are no constraints imposed on either the sign or the magnitude of the function γ, a negative value is certainly possible (although disconcerting) when viewed from a solution to equation (9).

In the next two example problems (concerned in part with controlling the error in the angular momentum vector), only the control vector and an outline of the solution are developed.

8.0 TWO-BODY PROBLEM WITH ANGULAR MOMENTUM INTEGRAL

The differential equations to be integrated are the same as in the previous example; however, the angular momentum integral shall now be considered, and it is expressed as

$$J(R,V) = R \times V = K \tag{30}$$

where the vector K is a constant of the motion defined by the initial conditions. Similar to the previous example, the numerical error vector in this integral is defined as

$$\epsilon = K - K_o \qquad , K_o = K(0) \tag{31}$$

and after introducing a control vector $\lambda^T = (\lambda_R^T , \lambda_V^T)$ into the RHS of equation (22), the time rate of change of the error may be determined as

$$\dot{\epsilon} = R \times \lambda_V - V \times \lambda_R \tag{32}$$

A stable vector differential equation for the functional relationship between the vector error and its rate is defined as

$$\dot{\epsilon} = - \gamma \epsilon \tag{33}$$

where γ is again a positive function. The control vector λ must now span a space defined by the vectors R, V, and K or

$$\lambda = \begin{bmatrix} \alpha_R & \alpha_V & \alpha_K \\ \beta_R & \beta_R & \beta_K \end{bmatrix} \begin{bmatrix} R \\ V \\ K \end{bmatrix} \tag{34}$$

where the α's and β's are undetermined scalars. But from equation (32), any component of the vector λ_R parallel to the vector V and, similarly, any component of the vector λ_V parallel to the vector R will have no effect on the error rate vector $\dot{\epsilon}$.

Hence, the vectors λ_R and λ_V have the simpler form

$$\lambda_R = \alpha_R R + \alpha_K K$$

$$\lambda_V = \beta_V V + \beta_K K \tag{35}$$

Using equation (35) in equation (32), the error rate vector is

$$\dot{\epsilon} = (\alpha_R + \beta_V) K + (\beta_K R - \alpha_K V) \times K \tag{36}$$

The first and second terms on the right-hand side of this equation are the rates of change of the angular momentum error vector along and perpendicular to the vector K, respectively. Taking the inner product of the vectors K, U = R x K and W = V x K (respectively) with equation (36) and using equation (33) yields the necessary conditions on the coefficients α and β:

$$\begin{aligned} (\alpha_R + \beta_V) &= -\gamma q \\ \beta_K\, U^T U - \alpha_K\, U^T W &= -\gamma\, U^T \varepsilon \\ \beta_K\, U^T W - \alpha_K\, W^T W &= -\gamma\, W^T \varepsilon \end{aligned} \tag{37}$$

where

$$q = \frac{\varepsilon^T K}{K^T K}$$

The first equation implies that α_R and β_V are homogenous in the term q. One solution for these coefficients is

$$\alpha_R = -\gamma\, c\, q\ , \quad \beta_V = -\gamma(1 - c)q \tag{38}$$

where the coefficient c is to be determined. The latter two equations (37) can be solved for α_K and β_K as

$$\begin{aligned} \alpha_K &= \frac{\gamma}{\Delta}\left[(\varepsilon^T W)(U^T U) - (\varepsilon^T U)(U^T W)\right] \\ \beta_K &= \frac{\gamma}{\Delta}\left[(\varepsilon^T W)(U^T W) - (\varepsilon^T U)(W^T W)\right] \end{aligned} \tag{39}$$

where the determinant, Δ, is

$$\Delta = (U^T U)(W^T W) - (U^T W)^2$$

It should be noted that the determinant is only zero for rectilinear motion.

The control vector for the vector angular momentum error is

$$\lambda_R = \alpha_R R + \alpha_K K$$
$$\lambda_V = \beta_V V + \beta_K K \qquad (40)$$

where α and β are defined by equations (38) and (39) and γ is determined from the condition $|\varepsilon(t + h| = 0$ (see appendix).

The coefficient c in the angular momentum error control vector is a scaling parameter between the components of error along and normal to the vector K (eq. (36)). Since the physics of the problem dictates that the error normal to the vector K will be extremely small, the actual value of the coefficient c used in the solution will have a minimal effect on the global error. One difficulty should be mentioned. Since the angular momentum error is formed by a vector product rather than as in the energy error (a scalar product) a considerable amount of algebraic manipulation will be required to obtain the function γ.

9.0 TWO-BODY PROBLEM WITH ANGULAR MOMENTUM AND ENERGY INTEGRALS

In this section the scalar energy integral is added to the vector angular momentum integral, and a control vector λ for both of these errors is determined. From equations (28) and (37), the necessary conditions on the coefficients α and β are

$$\frac{\mu\alpha_R}{r} + \beta_V V^T V = \gamma_k \varepsilon_k$$
$$\alpha_R + \beta_V = -\gamma_K q \qquad (41)$$

where α_K and β_K are given by equation (39) and the subscripts k, K refer to variables associated with the energy and angular momentum integrals, respectively. Solving equation (41) for the coefficients α_R and β_V yields

$$\alpha_R = -(\gamma_k \varepsilon_k + \gamma_K q V^T V)/(2k + \mu/r)$$
$$e \neq 0 \qquad (42a)$$
$$\beta_V = (\gamma_k \varepsilon_k + \gamma_K q \mu/r)/(2k + \mu/r)$$

However, this solution is singular when the eccentricity of the orbit $e = 0$. A solution valid for $e = 0$ is

$$\alpha_R = \frac{-\gamma_k \epsilon_k}{k}$$

$$e = 0 \qquad (42b)$$

$$\beta_V = \frac{\gamma_k \epsilon_k}{2k}$$

with the additional condition

$$\frac{\gamma_k \epsilon_k}{2k} = \gamma_K q \qquad (43)$$

Thus, for orbits with $e = 0$, the function γ_K is not directly determined by the integrator but by the function γ_k and the integral errors. But, the coefficients given by equation (42b) will produce the energy error control vector given by equation (29). Then if the coefficients α_K and β_K are small and if $e \sim 0$, the control vector that was used to reduce the energy integral error will also reduce the angular momentum integral error. The conclusion has been numerically verified.

For the problem illustrated in figure 1(a) with $N = 20$, the energy integral control vector reduced the angular momentum integral errors by approximately five orders of magnitude. However, when the eccentricity e was equal to 0.1, the energy integral control vector only reduced the angular momentum integral errors by approximately one order of magnitude.

10.0 RECOMMENDATIONS

This method for using the integrals of systems of nonlinear differential equation and their associated numerical errors should be applied to (1) a variable step integrator, preferably the Runge-Kutta 4/5, (2) other problems where integrals or other type of constraints are satisfied through rectification at each integration step, and (3) an unstable system of nonlinear differential equations.

11.0 REFERENCES

1. Whittaker, E. T.: A Treatise On the Analytical Dynamics of Particles and Rigid Bodies. Cambridge University Press, 1960.

2. Szebehely, V.: Linearization of Dynamical Systems Using Integrals of the Motion. Celestial Mechanics, vol. 14, 1976, pp. 499-508.

3. Stiefel, E.; and Scheifele, G.: Linear and Regular Celestial Mechanics. Springer-Verlag (New York), 1971.

4. Burdet, C. A.: Regularizations of the Two-Body Problem. ZAMP, vol. 18 1967, pp. 434-438.

5. Szebehely, V.: Theory of Orbits. Academic Press (New York), 1967.

6. Sperling, H. J.: The Collision Singularity in a Perturbed Two-Body Problem. Celestial Mechanics, vol. 1, 1969, pp. 213-221.

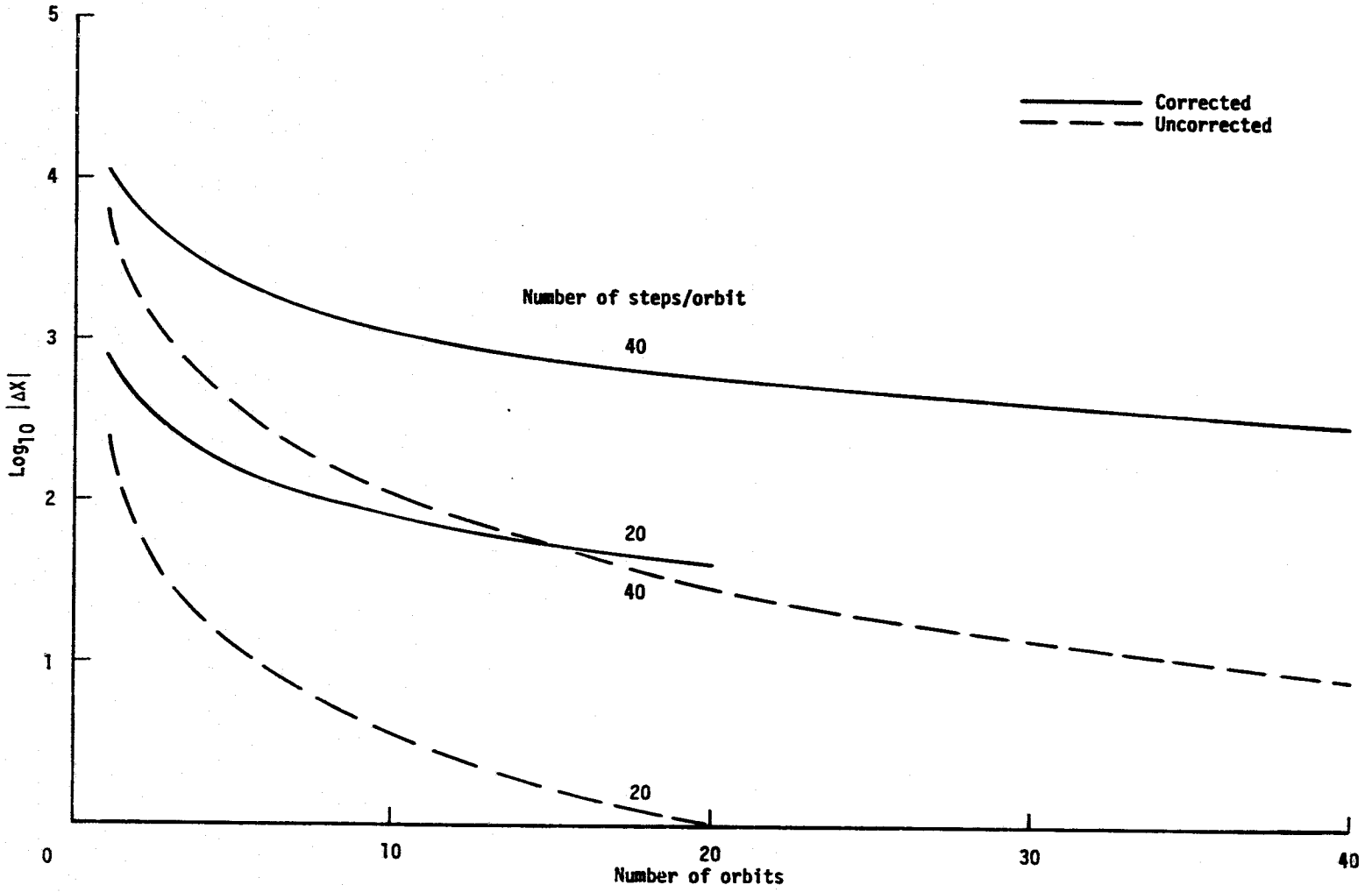
7. Nacozy, P.: The Use of Integrals in Numerical Integrations of the N-Body Problem. Astrophysics and Space Science, vol. 14, 1971, pp. 40-51.

8. Baumgarte, J.: Stabilization of Constraints and Integrals of Motion in Dynamical Systems. Computer Methods in Applied Mechanics and Engineering, 1972, pp. 1-16.

9. Baumgarte, J.: Numerical Stabilization of All Laws of Conservation in the Many Body Problem. Celestial Mechanics, vol. 8, 1973, pp. 223-228.

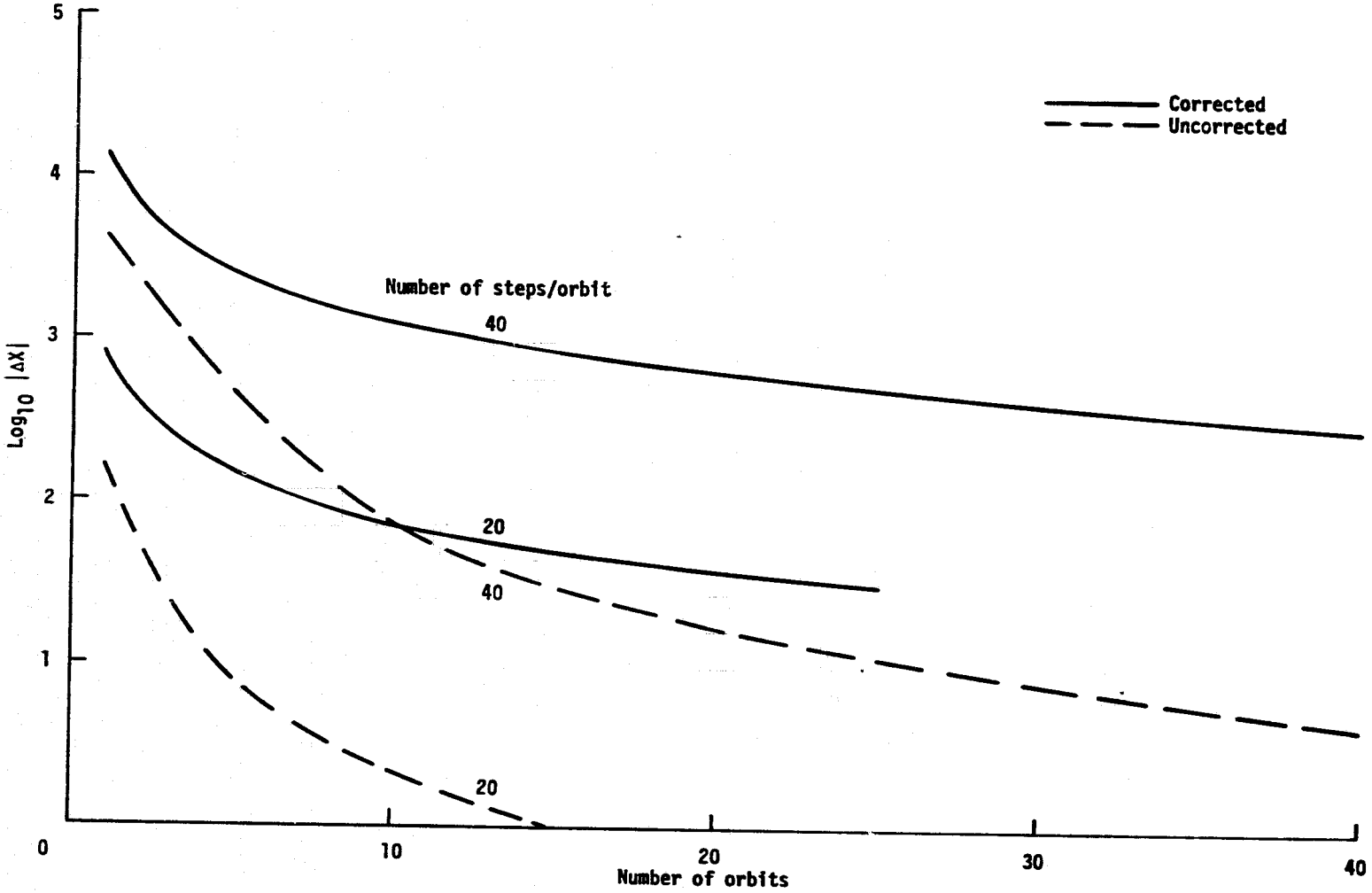
10. Baumgarte, J.: Asymptolische Stabilisierung von Integrales bis Gewohnlichen Differentialgleichunge 1. Ordnung. ZAMM 53, 1973, pp. 701-704.

11. Athans, M.; and Falb, P.L.: Optimal Control. McGraw-Hill (New York), 1966.



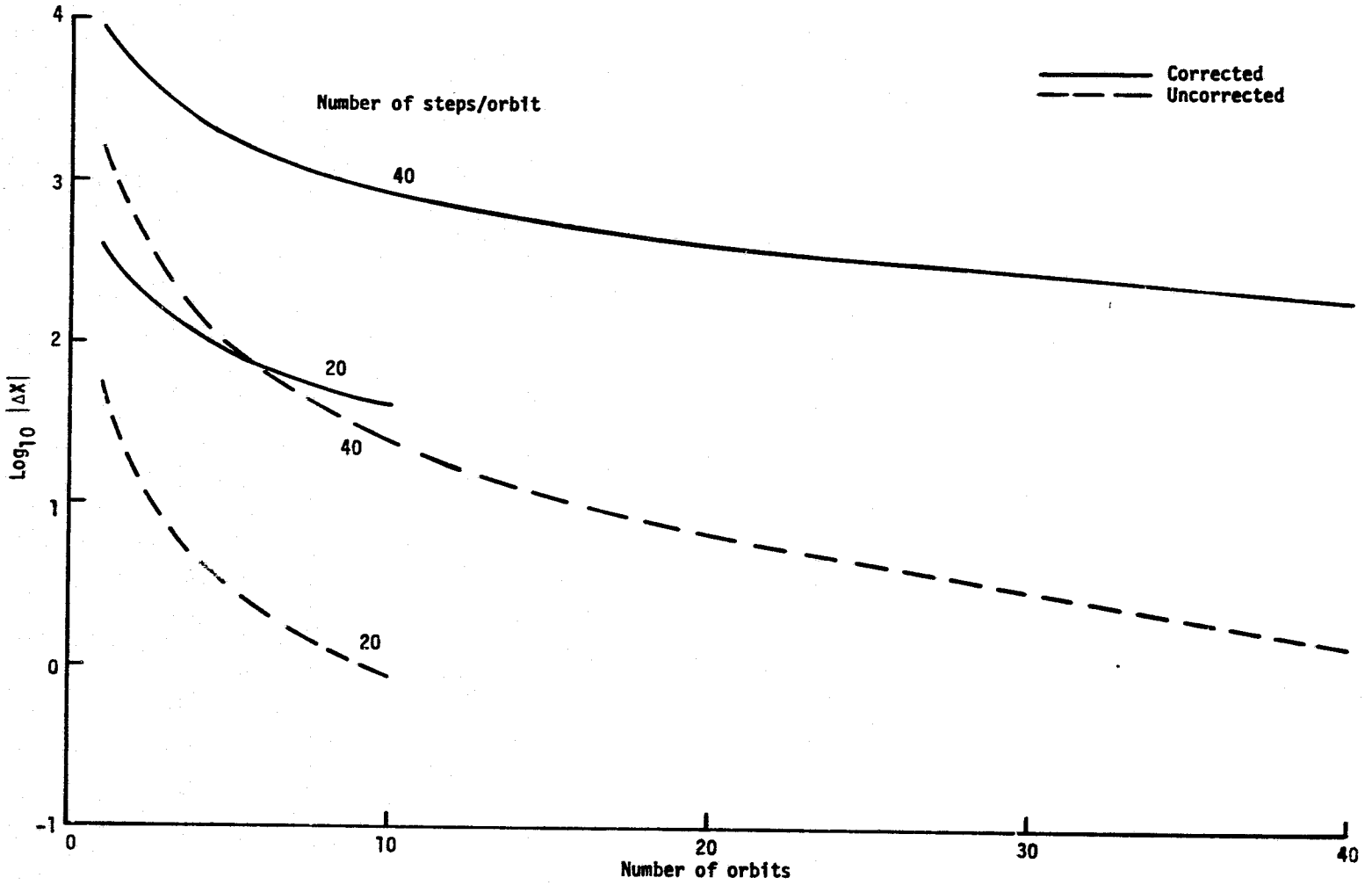
(a) e = 0.0, a = 1.0, 4th-order Runge-Kutta integrator.

Figure 1.- Two-body problem with energy control.



(b) e = 0.1, a = 1.0, 4th-order Runge-Kutta integrator.

Figure 1.- Continued.



(c) e = 0.2, a = 1.0, 4th-order Runge-Kutta integrator.

Figure 1.- Concluded.

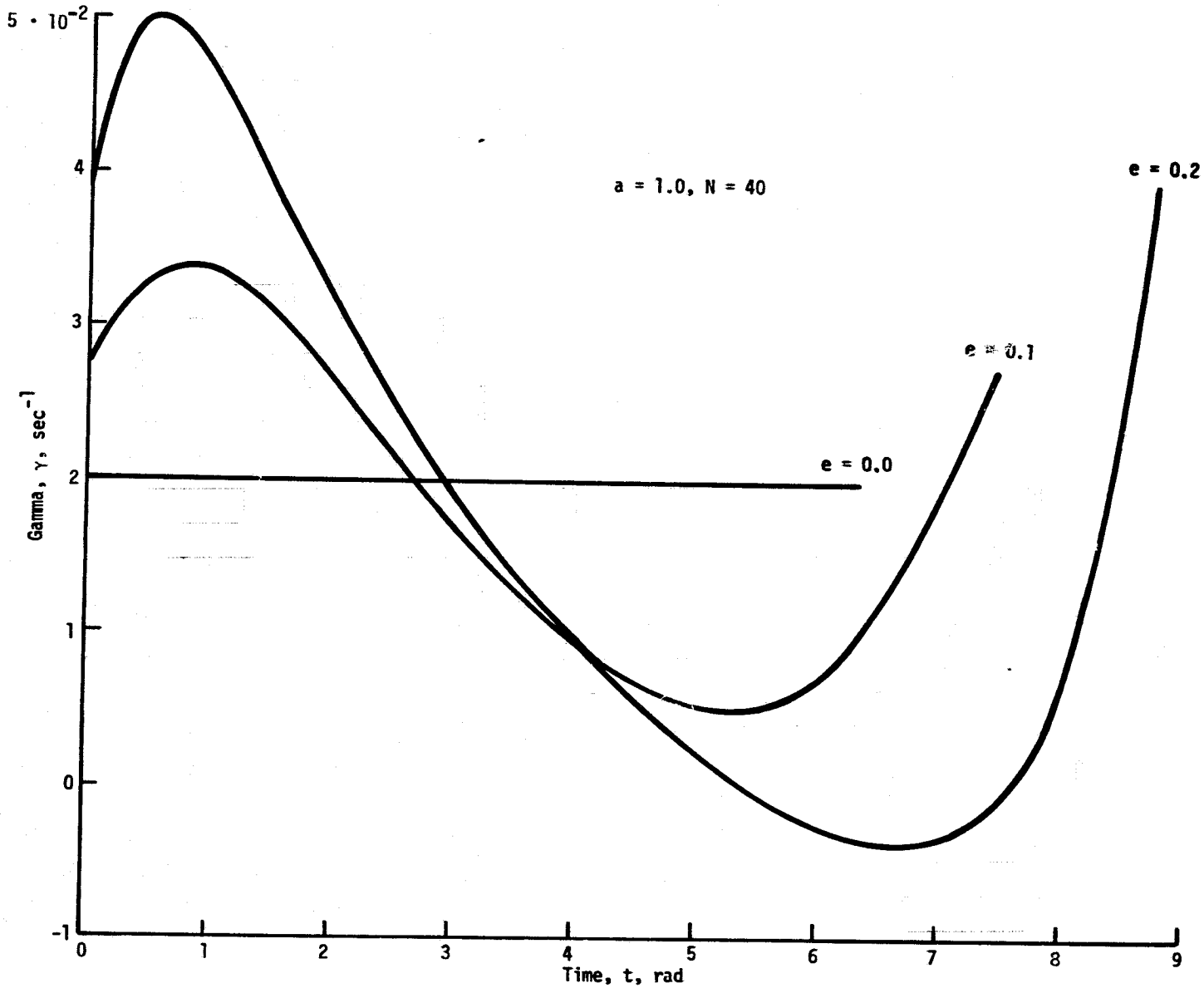


Figure 2.- Gamma function for two-body problem with energy control.

APPENDIX

CALCULATION OF THE FUNCTION γ

For a fourth-order, fixed-step Runge-Kutta integrator, the state vector is updated at a time $t + h$ by the expression

$$X(t + h) = X_o + \frac{h}{6}(F^1 + 2(F^2 + F^3) + F^4) \quad , X_o = X(t) \tag{A1}$$

where h is the integration step and the functions F^j are defined as

$$F^1 = F(X_o, t_o)$$

$$F^2 = F\left(X_o + \frac{h}{2} F^1, t_o + \frac{h}{2}\right)$$

$$F^3 = F\left(X_o + \frac{h}{2} F^2, t_o + \frac{h}{2}\right)$$

$$F^4 = F(X_o + hF^3, t_o + h)$$

For the controlled solution, the right-hand sides of the differential equations may be separated into two parts; the vector F and the vector η adjoined by the unknown function γ :

$$\dot{X} = F(X,t) + \gamma\, \eta(X,t)$$

where $\gamma\, \eta = \lambda$. For simplicity of notation, the argument t is excluded from the vectors F and η. Then the vector F^1 may be expressed as

$$F^1 = F(X_o) + \gamma\eta(X_o) = F_o + \gamma\, \eta_o \tag{A2}$$

and the vector F^2 may be expressed as

$$F^2 = F\left[X_o + \frac{h}{2}(F_o + \gamma\, \eta_o)\right] + \gamma\, \eta\left[X_o + \frac{h}{2}(F_o + \gamma\, \eta_o)\right]$$

But the vector λ represents a small perturbation to the vector F, and hence the vector F^2 may be approximated as

$$F^2 = F_1 + \gamma\,(\eta_1 + \delta_1) \quad + \ldots \tag{A3}$$

where

$$F_1 = F(X_1) \qquad , \eta_1 = \eta(X_1) \qquad , X_1 = X_o + \frac{h}{2} F_o$$

and

$$\delta_1 = \frac{h}{2} \left(\frac{\partial F}{\partial X} + \gamma \frac{\partial \eta}{\partial X}\right)\Big|_{X_1} (\eta_o + \delta_o) \qquad , \delta_o = 0$$

Similarly, the vectors F^3 and F^4 may be expressed as

$$\begin{aligned} F^3 &= F_2 + \gamma(\eta_2 + \delta_2) \\ F^4 &= F_3 + \gamma(\eta_3 + \delta_3) \end{aligned} \tag{A4}$$

where

$$X_2 = X_o + \frac{h}{2} F_1$$

$$X_3 = X_o + h\, F_2$$

Thus, the coefficient of $\frac{h}{6}$ in equation (A1) may be expressed as

$$F^1 + 2(F^2 + F^3) + F^4 = \mathcal{F} + \gamma(\mathcal{N} + \mathcal{S}) \tag{A5}$$

where

$$\begin{aligned} \mathcal{F} &= F_o + 2(F_1 + F_2) + F_3 \\ \mathcal{N} &= \eta_o + 2(\eta_1 + \eta_2) + \eta_3 \\ \mathcal{S} &= \delta_o + 2(\delta_1 + \delta_2) + \delta_3 \end{aligned}$$

For clarity, the analysis is restricted to a particular problem - the harmonic oscillator. Extensions to other problems are straightforward. The desire

is to determine the function Υ, which will result at a time $t + h$ in an integral error of zero.

$$G = \epsilon(t + h) = 1/2\ X(t + h)^T\ X(t + h) - k_o \tag{A6}$$

Using equations (A1) and (A5) in equation (A6) yields

$$\left| X_o + \frac{h}{6} \left[\mathcal{J} + \Upsilon(\mathcal{N}+\mathcal{S}) \right] \right|^2 = 2k_o \tag{A7}$$

This is a polynominal in the function Υ; however, if the small term $\frac{\partial \eta}{\partial X}$ is ignored, equation (A7) may be expressed as the quadratic equation

$$a\ \Upsilon^2 + b\ \Upsilon + c = 0$$

where

$$a = \left(\frac{h}{6}\right)^2 |\mathcal{N} + \mathcal{S}|^2$$

$$b = \frac{h}{3} \left(X + \frac{h}{6} \mathcal{J} \right)^T (\mathcal{N}+\mathcal{S})$$

$$c = \left| X + \frac{h}{6} \mathcal{J} \right|^2 - 2k_o$$

There are two solutions for the function Υ; however, it may be readily verified that the plus sign of the radical is correct.