MATHEMATICS OF COMPUTATION
VOLUME 44, NUMBER 169
JANUARY 1985, PAGES 81-92

Difference Methods for Problems With Different Time Scales*

By Robert E. Scheid, Jr.

Abstract. We consider the use of difference methods for weakly nonlinear systems of ordinary differential equations with rapidly oscillating solutions and develop a general approach which depends only on the smoothness of the coefficients and the nonlinearities. In particular, one is led to a strategy which is suitable for the detection and resolution of turning points in such systems. A computational example is presented.

1. Introduction. Mathematical modeling of many physical phenomena often leads to the study of differential systems whose stiffness gives rise to rapidly oscillating solutions. Physical examples as well as a detailed bibliography can be found in [5] and [6], where we developed an asymptotic theory for weakly nonlinear, highly oscillatory systems of ordinary differential equations and introduced corresponding methods which are suitable away from turning points for accurate computation with large time steps. Here we extend our results to achieve a general theory on difference methods for such problems. Our approach depends only on the smoothness of the coefficients and the nonlinearities and leads to a strategy which is suitable for the detection and resolution of turning points in such systems.

To illustrate the ideas, we consider the scalar problem:

$$z' = ia(t)z + z^2, \qquad z(0) = z_0, \qquad 0 \leqslant t \leqslant T, \tag{1.1}$$

where $a(t)$ is a smooth real function with

$$\inf_t |a(t)| \geqslant M > 1, \tag{1.2a}$$

$$\sup_t |d^\nu a/dt^\nu|/|a(t)| \leqslant K \qquad (\nu = 1,2,\ldots,p). \tag{1.2b}$$

Here K is a constant of moderate size. This expression is somewhat vague but often it is satisfactory to say that K is a constant of moderate size if

$$Kh \leqslant .1, \tag{1.3}$$

where h is the maximal step size to be used in the computation (cf. Kreiss [2]); that is, we approximate the solution on the grid $\{t_j\}$, where

$$t_0 = 0, \quad t_N = T, \quad h_j = t_{j+1} - t_j, \quad h = \max_j |h_j|. \tag{1.4}$$

Received May 16, 1983; revised March 16, 1984.

1980 *Mathematics Subject Classification.* Primary 65L05, 34E20.

Key words and phrases. Difference methods, oscillatory, numerical solution of ordinary differential equations, stiff equations, turning point, resonance.

*This research was carried out in part at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

©1985 American Mathematical Society
0025-5718/85 $1.00 + $.25 per page

We assume that the coefficients of the system (1.1) are known at the grid points.

With the simple change of variables

$$z = \exp(ip(t))x, \qquad p(t) = \int_0^t a(t)\,dt, \tag{1.5}$$

we reach a formulation in which the right-hand side is bounded independently of M:

$$x' = \exp(ip(t))x^2, \qquad x(0) = x_0 := z_0, \qquad 0 \leqslant t \leqslant T. \tag{1.6}$$

This system, whose coefficients are also assumed to be known at the grid points, is separable, and so the solution is readily obtained by integration:

$$x = x_0\{1/[\,1 - x_0F(t)]\} = x_0 \sum_{k=0}^{\infty} [x_0F(t)]^k, \tag{1.7}$$

$$F(t) = \int_0^t (\exp(ip(t)))\,dt.$$

For sufficiently large M integration by parts gives an asymptotic expansion for $F(t)$:

$$F(t) \sim \exp(ip(t))\left\{[-i/a(t)] + \left[a'(t)/(a(t))^3\right] + \cdots\right\}\Big|_0^t. \tag{1.8}$$

And by (1.7) we have a corresponding expansion for $x(t)$. The kth term of the expansion (1.8) has the form

$$\exp(ip(t))(1/(a(t)))^k f_k(\{(d^\nu a/dt^\nu)/a(t)\}), \qquad 1 < \nu < k-1, \tag{1.9}$$

and the remainder $R_k(t)$ after k terms can be bounded in accordance with (1.2):

$$|R_k(t)| \leqslant \hat{R}_k(K)/M^{k+1}. \tag{1.10}$$

Here $\hat{R}_k$ is bounded in terms of constants of moderate size. To approximate $x(t)$, one need only compute with a time step sufficient for the resolution of the $f_k(\cdot)$, and if the corresponding errors are dominated by those arising from the truncation of the expansion then, for sufficiently large M and for $x(t_j)$ given, a bound of the form (1.10) gives the error after a time step h_j. Clearly, the infimum and supremum in (1.2) can be taken locally since we need only develop a principle which allows us to calculate from grid point to grid point. Thus, as $|a(t)|$ becomes smaller, more terms must be included in the expansion to ensure an adequate approximation.

In [5] we demonstrated that even for more general systems such expansions can be generated automatically by means of a strategy of successive linearization combined with integration by parts. However, without an assumption such as (1.2a) this procedure is unworkable because the mathematical structure of $F(t)$ changes significantly over any interval where $|a(t)|$ approaches zero. This behavior characterizes the difficulty of the general theory at a turning point, where the expansions first become nonuniform and eventually break down entirely due to the failure of integration by parts.

When $|a(t)|$ becomes sufficiently small, however, it is reasonable to attempt to resolve it fully. Let us assume that $|a(t)|$ is of moderate size as specified by (1.3). On the grid defined by (1.4) we introduce the difference operator

$$\Delta a(t_j) = a(t_{j+1}) - a(t_j). \tag{1.11}$$

For $\nu < p$ we then have

$$|\Delta^{\nu}\exp(ip(t_j))| \leqslant R(K)h^{\nu}|a(t_*)|^{\nu} \qquad (t_j < t_* < t_{j+\nu}), \tag{1.12}$$

where R is bounded in accordance with (1.2b). Thus, the system (1.6) is suitable for approximation by standard difference methods. If $x(t_j)$ is known exactly, then the error with a standard qth order method after the step h_j has a bound of the form

$$|a(t_j)|^{q}h^{q+1}\tilde{R}(K), \tag{1.13}$$

where $\tilde{R}$ is bounded in accordance with (1.2b). A similar bound holds when (1.2a) is violated as long as $|a(t)|$ remains bounded (see, for example, Lambert [4]).

Thus, we have two different procedures for resolving the solution of the differential equation (1.6). To combine the methods we first must replace (1.2) with an assumption that remains appropriate as $|a(t)|$ approaches zero. Thus, we assume

$$\sup_{t}\left[|d^{\nu}a/dt^{\nu}|/\max\{|a(t)|,1\}\right] \leqslant K \qquad (\nu = 1,2,\ldots,p), \tag{1.14}$$

where K is a constant of moderate size. A similar bound was introduced by Kreiss [2] to develop a theory of difference methods for stiff problems which are essentially nonoscillatory. This requirement is apparently necessary for any approximation scheme based on finite differences, for otherwise the point values of the function cannot provide an adequate representation at the grid points. Thus, from (1.11) and (1.14) we have

$$|\Delta^{\nu}a(t_j)| \leqslant h^{\nu}K \sup_{t_j < s < t_{j+\nu}} \left[\max\{|a(s)|,1\}\right] \qquad (\nu \leqslant p). \tag{1.15}$$

Replacing assumption (1.2) by assumption (1.14) yields an appropriate modification of the estimate (1.13):

$$\max\{|a(t_j)|^{q},1\}h^{q+1}\tilde{R}(K). \tag{1.16}$$

Here $\tilde{R}$ is bounded in accordance with (1.14). Unlike (1.13) this estimate for the error after a time step h_j remains valid even as $|a(t)|$ approaches zero. The estimate (1.10) for the truncation of the asymptotic expansion remains valid as long as (1.14) and (1.2a) hold.

When $|a(t)|$ is sufficiently large one proceeds as in (1.8), where the derived asymptotic expansion decays in reciprocal powers of $|a(t)|$ and the error is bounded in accordance with the estimate (1.10). As $|a(t)|$ decreases in magnitude, more terms in the expansion must be included to ensure the same accuracy. Comparing the bounds (1.10) and (1.16) and assuming the amount of work for each procedure is comparable for some given k and q, we conclude that an appropriate point of transition from the first method to the second occurs when

$$|a(t)| \approx \left(x_0^2\hat{R}_k/\tilde{R}\right)^{1/(k+q+1)}h^{-(q+1)/(k+q+1)} \tag{1.17}$$

since the local errors are approximately equal. If the switching is made according to this rule, then, for fixed K and sufficiently small h, the solution is fully resolved only when $|a(t)|$ is of moderate size as specified by (1.3). Of course this procedure can be reversed if $|a(t)|$ again becomes large as in a passage through resonance (cf. Example 4.1).

The changeover from one method to the other then marks the transition between a fast mode which requires asymptotic analysis and a slow mode which requires full

resolution, and correspondingly the values of k, h and q determine a trade-off between the number of integrations needed in one region and the number of grid points needed in another. One might compare this procedure to a perturbation theorist's use of matched asymptotic expansions.

In the subsequent sections of this paper we demonstrate how this approach extends very naturally for the general case, where the crucial bound (1.14) must be enforced for all of the relevant frequencies of the problem. As the following example demonstrates, an appropriate stretching of the independent variable can ensure the bound (1.14) if it is not met in the original formulation of the problem.

Example. The solution of the system

$$dy/dt = (it^p/\varepsilon)y, \qquad y(0) = 1, t \geqslant 0, 0 < \varepsilon \ll 1, p \in \{1,2,3,\dots\}$$

is given by

$$y(t) = \exp\bigl(it^{p+1}/\varepsilon(p+1)\bigr).$$

To enforce the bound (1.14) in a neighborhood of $t = 0$, we introduce a new variable $t = \alpha s$, where α is a positive constant, and obtain

$$dy/ds = i\bigl(\alpha^{p+1}/\varepsilon\bigr)s^p y =: a(s)y, \qquad y(0) = 1.$$

For this example, as in many cases, it is sufficient to bound only the first derivative. We determine the largest α such that

$$\sup_{0<s<1} \bigl(|da/ds|/\max\{|a(s)|,1\}\bigr) \leqslant K,$$

where K is a constant of moderate size. For simplicity we take $K = p$ and find $\alpha = \varepsilon^{1/(p+1)}$. Thus we have obtained the stretched variables for $0 < t < t_1 := \varepsilon^{1/(p+1)}$. For $t = t_1 + \alpha s$ the obvious modification of the previous bound gives $\alpha = t_1$, and we have the appropriate stretching for $0 < t < t_2 := 2t_1$. This process can be continued, and after n steps we have obtained the proper stretching for the interval $0 < t < t_n := 2^{n-1}\varepsilon^{1/(p+1)}$. Thus the stretching is appropriately reduced as the distance from the turning point increases. For $t > 1$ the bound is achieved with no additional stretching.

2. The General Problem. Let $\mathbf{z} = (z^{(1)}, z^{(2)},\dots,z^{(n)})^T$ be an n-dimensional vector, and let A be an $n \times n$ matrix.**

We consider the general system

$$\mathbf{z}' = A(t)\mathbf{z} + \mathbf{h}(\mathbf{z}, t) + \mathbf{f}(t), \qquad \mathbf{z}(T_1) = \mathbf{z}_0, \qquad T_1 \leqslant t \leqslant T_2, \tag{2.1}$$

and we assume

$$\sup_t |A(t) - A(t)^*| \gg 1, \tag{2.2a}$$

**If $\mathbf{z}$ is a vector, then $\mathbf{z}^T$ denotes its transpose and $\mathbf{z}^*$ its adjoint. The vector norm is defined by $|\mathbf{z}| = \max_k |z^{(k)}|$. Similar notations hold for matrix norms. For example, $|A| = \sup_{\mathbf{z}} |A\mathbf{z}|/|\mathbf{z}|$. " ′ " denotes differentiation with respect to t, and $a^{[\nu]}(t) = d^\nu a/dt^\nu$. $C^p(t)$ denotes the class of functions which have p continuous derivatives on the interval of interest.

$$\sup_t |A(t) + A(t)^*| \leqslant K, \tag{2.2b}$$

$$\sup_t |\mathbf{h}(\mathbf{z}, t)| \leqslant K, \qquad |\mathbf{z}| = 1, \tag{2.2c}$$

$$\sup_t |(A(t) - cI)^{-1}\mathbf{f}(t)| \leqslant K, \tag{2.2d}$$

$$|\mathbf{z}_0| \leqslant K, \tag{2.2e}$$

$$|T_2 - T_1| \leqslant K, \tag{2.2f}$$

where K and c are constants of moderate size. Conditions for smoothness and their significance are outlined in the following assumptions and discussion.

Assumption A. The matrix $A(t)$ is in diagonal form with purely imaginary entries, each of which is in $C^p(t)$ and satisfies the bound (1.14):

$$A(t) = \operatorname{diag}(i \cdot a_{11}(t), i \cdot a_{22}(t), \ldots, i \cdot a_{nn}(t)), \qquad A(t) = -A(t)^*. \tag{2.3}$$

For convenience we introduce the vector $\Lambda(t)$ defined by

$$\Lambda(t) = (i \cdot a_{11}(t), i \cdot a_{22}(t), \ldots, i \cdot a_{nn}(t))^T. \tag{2.4}$$

The entries of $\Lambda(t)$ are called the *fundamental frequencies* of the system (2.1).

According to the restrictions (2.2), we need only consider the case where $A(t)$ is antisymmetric since the symmetric part of $A(t)$ can be absorbed by the other terms of the right-hand side. If $A(t)$ is an analytic function of t in a domain which contains the interval of interest and $A(t)$ is antisymmetric on that interval, then there exists an analytic unitary transformation which reduces $A(t)$ to diagonal form on the interval (see Kato [1, p. 121]), and thus Assumption A is justified. If the conditions for smoothness are relaxed, then pathological cases can arise. Like Kato [1, p. 11] we quote an example due to Rellich. The symmetric matrix

$$A(t) = \begin{cases} \exp(-1/t^2)\begin{bmatrix} \cos(2/t) & \sin(2/t) \\ \sin(2/t) & -\cos(2/t) \end{bmatrix}, & t \neq 0, \\ \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, & t = 0, \end{cases}$$

is infinitely differentiable for all real values of t and has eigenvalues

$$\lambda_{\pm}(t) = \begin{cases} \pm\exp(-1/t^2), & t \neq 0, \\ 0, & t = 0. \end{cases}$$

However, there does not exist a continuous parametrization of an eigenvector in a neighborhood of $t = 0$.

Other difficulties are also possible in highly oscillatory systems. For example, the matrix

$$\frac{1}{\varepsilon}\begin{bmatrix} i & 1 \\ 0 & i \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \qquad (0 < \varepsilon \ll 1) \tag{2.5}$$

has eigenvalues

$$\lambda_{\pm} = i/\varepsilon \pm 1/\sqrt{\varepsilon}.$$

Thus, if $A(t)$ has Jordan structure as in the first term of (2.5), then a $O(1)$ perturbation of the system will cause the solutions to become unbounded as ε approaches zero.

Assumption B. The components of $\mathbf{h}(\mathbf{z}, t)$ are polynomial in the components of $\mathbf{z}$ with coefficients $\{c_m\}$ which are in $C^p(t)$ and which satisfy the bound

$$\sup_t |c_m^{[\nu]}(t)| \leqslant K \qquad (\nu = 0, 1, \ldots, p).$$

Here K is a constant of moderate size.

If the nonlinearities of (2.1) are sufficiently smooth, then a polynomial approximation can be made locally or globally. For example, let $\mathbf{r}(\mathbf{z}, t)$ be a polynomial approximation to $\mathbf{h}(\mathbf{z}, t)$ with

$$\sup_{\mathbf{z}, t} |\mathbf{h}(\mathbf{z}, t) - \mathbf{r}(\mathbf{z}, t)| < \delta$$

for $\mathbf{z}$ and t in the range of interest. We shall demonstrate that under our assumptions the induced error is bounded by $R\delta|T_2 - T_1|$, where R is bounded in terms of constants of moderate size.

Assumption C. $\mathbf{f}(t)$ is in $C^p(t)$ and satisfies the bound

$$|\mathbf{f}^{[\nu]}(t)| \leqslant K \qquad (\nu = 0, 1, \ldots, p),$$

where K is a constant of moderate size.

Large forcing as permitted by (2.2d) also can be treated. However, we shall show that the more general problem can be reduced to this form, which is more convenient for theoretical purposes. We say that the system (2.1) satisfying Assumptions A, B, and C is in *stiff oscillatory form*.

As in the scalar problem (1.1), this system can be reduced to a more tractable formulation. Let $S(t, s)$ be the solution operator of the system with $\mathbf{h}(\mathbf{z}, t) \equiv 0$ and $\mathbf{f}(t) \equiv 0$. In fact, we can write

$$S(t, s) = \operatorname{diag}\left\{\exp\left[\left(i\int_s^t a_{kk}(r)\, dr\right)\right]\right\}. \tag{2.6}$$

Thus, with the change of variables

$$z = S(t, T_1)x \tag{2.7}$$

we reach a system in which the coefficients are bounded but some are rapidly oscillating:

$$\mathbf{x}' = \mathbf{G}(\mathbf{x}, t), \qquad \mathbf{x}(T_1) = \mathbf{v}_0, \qquad T_1 \leqslant t \leqslant T_2. \tag{2.8}$$

The components of the terms on the right-hand side of the equation have the form

$$p(\mathbf{x})c(t)\exp\left(\int_{T_1}^t \phi(t)\, dt\right), \tag{2.9}$$

where $p(\mathbf{x})$ is monomial in the components of $\mathbf{x}$, $c(t)$ is in $C^p(t)$, and $\phi(t)$ has the form

$$\phi(t) = \mathbf{N}^T\Lambda(t). \tag{2.10}$$

Here **N** is an n-dimensional vector with integral components which depend on the polynomial form of the nonlinearities. Occurring combinations of this form are called *relevant secondary frequencies*, and the system (2.8) is said to be in *nonstiff oscillatory form*. The bound (1.14) must be enforced for each of the relevant secondary frequencies, and also the partitioning suggested in Section 1 must be introduced. Thus we are led to the following assumption, which is basic for our theory.

Assumption D. Each relevant secondary frequency of the system (2.8) satisfies the bound (1.14). Moreover, the relevant secondary frequencies of the system can be divided into two groups:

nonoscillatory set:

$$\left\{\phi(t)\colon \sup_t |\phi(t)| \leqslant M_{\mathrm{II}}\right\},$$

oscillatory set:

$$\left\{\phi(t)\colon \inf_t |\phi(t)| \geqslant M_{\mathrm{I}} > 1\right\},$$

where M_{II} is a constant of moderate size.

The bound (1.14) is appropriate since the discretizations we shall employ are equivalent to polynomial interpolations of these functions, and we must insure that point values are sufficient for such representations (cf. (1.15)). The partitioning of the relative secondary frequencies is fundamental to our approach, whereby some modes are resolved fully while others are treated asymptotically.

We can rewrite system (2.8) satisfying Assumption D as

$$\mathbf{x}' = \mathbf{g}_{\mathrm{I}}(\mathbf{x}, t) + \mathbf{g}_{\mathrm{II}}(\mathbf{x}, t) + \mathbf{f}_{\mathrm{I}}(t) + \mathbf{f}_{\mathrm{II}}(t), \quad \mathbf{x}(T_1) = \mathbf{v}_0, \qquad T_1 \leqslant t \leqslant T_2, \tag{2.11}$$

where the subscripts reflect the partitioning given by Assumption D. Terms with the subscript I contain frequencies in the oscillatory set while terms with the subscript II contain frequencies in the nonoscillatory set. The forcing terms $\mathbf{f}_{\mathrm{I}}$ and $\mathbf{f}_{\mathrm{II}}$ contain all components of the form (2.9) with $p(\mathbf{x}) \equiv 1$. The following assumption is needed to guarantee the well-posedness of the system.

Assumption E. In correspondence to the system (2.11), the reduced system

$$\mathbf{v}' = \mathbf{g}_{\mathrm{II}}(\mathbf{v}, t) + \mathbf{f}_{\mathrm{II}}(t), \qquad \mathbf{v}(T_1) = \mathbf{v}_0, \qquad T_1 \leqslant t \leqslant T_2, \tag{2.12}$$

is well-posed and has a solution bounded in terms of constants of moderate size.

For the case where turning points are not allowed, we developed in [5] an asymptotic theory based on linearizations about $\mathbf{v}(t)$ and the integration by parts of rapidly oscillating forcing functions. We now give the extension of these results to the more general system (2.11). The essential point is that each linearization yields a linear system of the form (2.11) with possibly different relevant secondary frequencies to which Assumption D must apply.

THEOREM 2.1. *Let the system* (2.11) *in nonstiff oscillatory form satisfy Assumptions* D *and* E. *Assume also that Assumption* D *holds for the linear systems obtained in the first* k *linearizations as done in* [5] ($k < p$). *Then for sufficiently large* M_{I} *there exists an*

$$\tilde{\mathbf{x}}_k = \mathbf{x}_0 + (1/M_{\mathrm{I}})\mathbf{x}_1 + (1/M_{\mathrm{I}})^2\mathbf{x}_2 + \cdots + (1/M_{\mathrm{I}})^k\mathbf{x}_k \tag{2.13}$$

such that we have

$$\sup_t |\mathbf{x}_j| \leqslant R, \tag{2.14a}$$

$$\sup_t |\tilde{\mathbf{x}}_k - \mathbf{x}| < R(1/M_1)^{k+1}, \tag{2.14b}$$

where R is bounded in terms of constants of moderate size. Each x_j consists of a smooth component, which has a number of derivatives bounded in terms of constants of moderate size, and a rapidly oscillating component, which is given explicitly in terms of the coefficients of the system and the smooth components of the solution.

Proof. The theorem follows directly from the theorems of [5], where we assumed that relevant secondary frequencies were either uniformly large or else identically vanishing. Terms corresponding to vanishing frequencies were simply absorbed into the nonoscillatory terms of the system. Actually, we need only assume that the frequencies which are not large can be adequately resolved. Since our assumptions guarantee this, we have the desired result.

The results of [5] also give error estimates which justify the approximation of the nonlinearities by polynomials as in Assumption B; the original differential equation has simply been replaced by one that is nearby in a very natural sense. And also the same reference gives a formal procedure for developing the expansions so that repeated linearizations are unnecessary. The preceding theorem can be extended to give the degree of smoothness of each correction.

3. Reduction to Stiff Oscillatory Form. If the character of an oscillatory problem is essentially harmonic, then one should be able to achieve the form (2.1) so long as the coefficients and the nonlinearities are sufficiently smooth. We now extend our basic results to systems of the form

$$\begin{aligned} &\mathbf{x}' = A(t)\mathbf{x} + \mathbf{f}_1(\mathbf{y}, t) + \mathbf{f}_2(\mathbf{x}, \mathbf{y}, t), \\ &\mathbf{y}' = \mathbf{f}_3(\mathbf{x}, \mathbf{y}, t), \\ &\mathbf{x}(T_1) = \mathbf{x}_0, \qquad \mathbf{y}(T_1) = \mathbf{y}_0, \qquad T_1 \leqslant t \leqslant T_2. \end{aligned} \tag{3.1}$$

Here $\mathbf{x}$ and $\mathbf{y}$ are vectors of possibly different dimensions. We assume that the system with $\mathbf{f}_1 \equiv 0$ satisfies all the assumptions of Section 2. We are interested in the case where

$$\sup_t |(A(t) - cI)^{-1}\mathbf{f}_1(\mathbf{y}, t)| \leqslant K, \tag{3.2}$$

where c and K are constants of moderate size. We now introduce the appropriate assumption.

Assumption F. $f_1^{(k)}(\mathbf{y}, t)$ is polynomial in the components of $\mathbf{y}$ with coefficients $\{d_{mk}(t)\}$ which are in $C^p(t)$ and which satisfy the bound

$$\sup_t |d_{mk}^{[\nu]}(t)/(a_{kk}(t) - c)| \leqslant K \qquad (\nu = 0, 1, \ldots, p). \tag{3.3}$$

Here c and K are constants of moderate size.

By inspection we have the following useful result, which justifies Assumption C.

THEOREM 3.1. *If the system* (3.1) *satisfies the preceding assumption, then the system for*

$$\tilde{\mathbf{x}} = \mathbf{x} + (A(t) - cI)^{-1}\mathbf{f}_1(\mathbf{y}, t) \tag{3.4}$$

and $\mathbf{y}$ *is in stiff oscillatory form.*

A similar transformation was introduced in [3], where the goal for such systems was to characterize the solutions which have a number of bounded derivatives. Other techniques based on linearization also can lead to systems of the form (2.1).

Many important applications arise in celestial mechanics. In particular, we can apply our techniques to the calculation of orbits of satellites. For many of these problems, the dominant forces are due to a spherically symmetric Newtonian potential perturbed by small effects, and, as Laplace noted, a change of variables will reduce such a system to the form (3.1). We consider a very simple model taken from [7], where a number of examples and references can be found. In dimensionless polar coordinates, the equations for the planar motion of a satellite about a planet are

$$\begin{aligned} r'' - r(\theta')^2 &= -1/r^2 + \varepsilon f_1(r, \theta, r', \theta'), \\ r\theta'' + 2r'\theta' &= \varepsilon f_2(r, \theta, r', \theta'), \end{aligned} \tag{3.5}$$

where ε, the small parameter, measures the ratio of the small perturbing force to the gravitational force. For $\varepsilon = 0$ the solution, which is available in any text on dynamics, indicates that $u = 1/r$ is a harmonic function of θ. Like Laplace we now transform the variables from r and θ as functions of t to u and t as functions of θ. Denoting the θ-derivative of u by $u^{\cdot}$, we have

$$\begin{aligned} u^{\cdot\cdot} + u - u^4 t^{\cdot 2} &= -\varepsilon u^2 t^{\cdot 2}(f_1 + (u^{\cdot}/u) f_2), \\ (u^2 t^{\cdot})^{\cdot} &= -\varepsilon u^3 t^{\cdot 3} f_2. \end{aligned} \tag{3.6}$$

If f_1 and f_2 are smooth functions of u, $u^{\cdot}$, and $t^{\cdot}$ only, then the system with dependent variables

$$x^{(1)} = u, \quad x^{(2)} = u^{\cdot}, \quad y^{(1)} = u^2 t^{\cdot}, \tag{3.7}$$

and independent variable

$$\tilde{\theta} = \varepsilon\theta, \tag{3.8}$$

has the form (3.1). Thus, if polynomial approximations for the nonlinearities can be made, we can reduce the problem to stiff oscillatory form. The time history is given by the equation

$$t^{\cdot} = y^{(1)}/x^{(1)2}, \tag{3.9}$$

which must be included in the analysis if f_1 or f_2 depends on t; we shall discuss the details of this in a future paper.

4. Difference Methods. The constructive approach we have taken leads very naturally to methods which are suitable for accurate computation with large time steps. We consider the system (2.11) in nonstiff oscillatory form and satisfying the assumptions of Theorem 2.1. Since the relevant secondary frequencies can be determined a priori, one can make the stretching necessary to enforce the bound

(1.14) before any computations are performed. Transitions between the oscillatory and nonoscillatory sets can be controlled by a procedure analogous to the one specified by (1.17), since estimates similar to (1.10) and (1.16) also hold for the general case. Then, for sufficiently small h, frequencies in the nonoscillatory set will be of moderate size while frequencies in the oscillatory set will be sufficiently large.

If analytic forms are not available, then the integrals of the relevant secondary frequencies can be approximated by quadrature. Thus, if a qth order integrator $I_q(\cdot, t_1, t_2)$ is chosen ($q < p$), we have

$$\left| I_q(\phi, t_1, t_2) - \int_{t_2}^{t_1} \phi(t)\, dt \right| \leqslant R|t_2 - t_1| \Big(\sup_s \left[\max\{ |\phi(s)|, 1 \} \right] \Big) h^q, \tag{4.1}$$

where R is bounded in terms of constants of moderate size. If $\phi(t)$ belongs to the oscillatory set, then the phase distortion induced by this approximation does not effect the other calculations.

The t-derivatives resulting from the successive linearizations can be approximated by differences without difficulty. Then, the smooth component of each term in the expansion (2.13) can be approximated by standard numerical techniques, since the equations are not stiff, while the rapidly oscillating component is given explicitly. If some frequencies in the oscillatory set are much larger than others, then fewer integrations will be required for the corresponding terms; appropriate modifications in the theory can be made by a further partitioning of the oscillatory set. In any case, the errors induced by these approximations must remain bounded by arguments similar to those of Theorem 2.1.

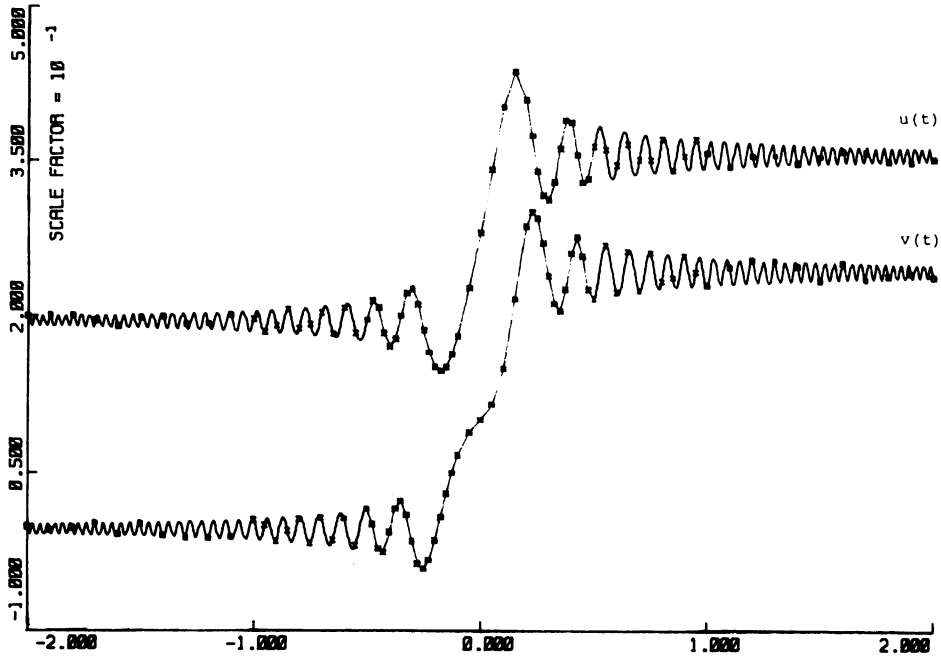

FIGURE 1

TABLE 1

h_j	j	t_j	Re{y(t_j)}	Im{y(t_j)}	ERR(t_j)
0.100	0	-2.0000	1.00000	0.00000	0.0E+00
0.100	2	-1.8000	1.00107	-0.00129	0.9E-05
0.100	4	-1.6000	0.99113	-0.00677	0.1E-03
0.100	6	-1.4000	0.99973	-0.00829	0.1E-03
0.100	8	-1.2000	0.99352	-0.01050	0.2E-03
0.100	10	-1.0000	0.99826	0.00721	0.4E-04
0.050	12	-0.9000	0.99165	-0.01260	0.4E-04
0.050	14	-0.8000	0.98892	0.00797	0.5E-04
0.050	16	-0.7000	1.00420	0.00904	0.6E-04
0.050	18	-0.6000	1.00840	0.00842	0.7E-04
0.050	20	-0.5000	0.99741	0.01743	0.1E-03
0.025	22	-0.4500	1.00870	-0.02025	0.1E-03
0.025	24	-0.4000	0.97171	-0.00420	0.1E-03
0.025	26	-0.3500	1.00153	0.02518	0.1E-03
0.025	28	-0.3000	1.02715	-0.01301	0.1E-03
0.025	30	-0.2500	0.98836	-0.03937	0.1E-03
0.025	32	-0.2000	0.95287	-0.01222	0.1E-03
0.025	34	-0.1500	0.95326	0.03179	0.1E-03
0.025	36	-0.1000	0.98220	0.06779	0.1E-03
0.050	38	0.0000	1.08236	0.10200	0.1E-03
0.050	40	0.1000	0.20317	0.15097	0.1E-03
0.050	42	0.2000	1.20994	0.28779	0.1E-03
0.025	44	0.2500	1.14090	0.29565	0.1E-03
0.025	46	0.3000	1.11363	0.24000	0.1E-03
0.025	48	0.3500	1.16308	0.20580	0.1E-03
0.025	50	0.4000	1.18836	0.26220	0.1E-03
0.025	52	0.4500	1.13042	0.25888	0.1E-03
0.025	54	0.5000	1.16518	0.21792	0.1E-03
0.050	56	0.6000	1.14675	0.22434	0.2E-03
0.050	58	0.7000	1.15234	0.22559	0.2E-03
0.050	60	0.8000	1.17186	0.23485	0.2E-03
0.050	62	0.9000	1.15608	0.26028	0.1E-03
0.050	64	1.0000	1.15907	0.23090	0.2E-03
0.100	66	1.2000	1.15570	0.25631	0.3E-03
0.100	68	1.4000	1.14887	0.25014	0.4E-03
0.100	70	1.6000	1.16079	0.25278	0.2E-03
0.100	72	1.8000	1.15089	0.24039	0.3E-03
0.100	74	2.0000	1.15296	0.23930	0.2E-03

j is the number of the grid point
t_j is the value of t at the jth grid point
h_j is the time step at the jth grid point

Thus the system (2.11) is suitable for numerical approximation with a large time step. Two computational examples were given in [5], where we precluded the possibility of turning points. Here we present an example which corresponds to such a passage through resonance.

Example 4.1. We consider the system

$$y' = \exp(it^2/2\varepsilon)y^2, \qquad y(-2) = 1, \qquad -2 \leqslant t \leqslant 2, \varepsilon = .01,$$

which one might have derived from the system

$$z' = (it/\varepsilon)z + z^2, \qquad z(-2) = \exp(2i/\varepsilon), \qquad -2 \leqslant t \leqslant 2, \varepsilon = .01,$$

after the change of variables (2.7). We propose to solve this system to within three or four digits of accuracy. Complete details of the calculations are given in [6], where

we have included the resulting algebra. Smooth components were resolved by means of a fourth-order Runge-Kutta solver with a time step sufficient so that the estimated local truncation error would be bounded by $Rh(10^{-3})$, where R is a constant of moderate size. For this example we adjusted the step size by a factor of (1/2) or (2) as required. One can estimate the local truncation error by comparing the result of two increments with the result of one double-increment (see, for example, Lambert [4]). The t-derivatives brought about by the linearizations were determined analytically rather than estimated by differences. For the range

$$1 \leqslant |t| \leqslant 2$$

only one integration was required, while for the range

$$1/2 \leqslant |t| \leqslant 1$$

several integrations were required. And finally all frequencies were fully resolved in the range

$$|t| \leqslant 1/2.$$

All calculations were done with single-precision accuracy on a VAX11/780 computer. Plots of the resulting approximations to

$$u = \mathrm{Re}\{y\} - .8 \quad \text{and} \quad v = \mathrm{Im}\{y\}$$

illustrate the passage through resonance in Figure 1, where we have linearly interpolated the amplitudes and phases between grid points. In Table 1 we compare the computed grid values with the accepted function values, which were computed with double-precision accuracy by means of a fourth-order Runge-Kutta scheme with a time step $h = 10^{-4}$. In the table we list

$$\mathrm{ERR}(t) = \max\left[|\mathrm{Re}\{y - \tilde{y}\}|, |\mathrm{Im}\{y - \tilde{y}\}|\right],$$

where $y(t)$ and $\tilde{y}(t)$ are respectively the computed and the accepted function values.

5. Acknowledgment. The author is indebted to Professor H.-O. Kreiss for his helpful criticism and advice.

Jet Propulsion Laboratory
M.C. 198-326
California Institute of Technology
Pasadena, California 91109

1. T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.

2. H.-O. Kreiss, "Difference methods for stiff ordinary differential equations," *SIAM J. Numer. Anal.*, v. 15, 1978, pp. 21–58.

3. H.-O. Kreiss, "Problems with different time scales for ordinary differential equations," *SIAM J. Numer. Anal.*, v. 16, 1979, pp. 980–998.

4. J. D. Lambert, *Computational Methods in Ordinary Differential Equations*, Wiley, London, 1973.

5. R. E. Scheid, Jr., "The accurate numerical solution of highly oscillatory ordinary differential equations," *Math. Comp.*, v. 41, 1983, pp. 487–509.

6. R. E. Scheid, Jr., *The Accurate Numerical Solution of Highly Oscillatory Ordinary Differential Equations*, Ph. D. thesis, California Institute of Technology, 1982.

7. J. Kevorkian & J. D. Cole, *Perturbation Methods in Applied Mathematics*, Springer-Verlag, New York, 1981.