MATHEMATICS OF COMPUTATION, VOLUME 30, NUMBER 135
JULY 1976, PAGES 383–399

An Averaging Method for the Stiff Highly Oscillatory Problem

By W. L. Miranker* and G. Wahba**

Abstract. We show how to replace the point functionals of numerical analysis with types of stable functionals of highly oscillatory solutions of differential equations. This replacement leads to the development of effective numerical methods for the stiff highly oscillatory problem.

1. Introduction. Numerical methods for approximating the solutions of the initial value problem

$$\begin{aligned} \dot{x} &= f(x, t), \qquad t > 0, \\ x(0) &= x_0, \end{aligned} \tag{1.1}$$

where $x(t)$ and f are n-vectors, proceed for the most part by generating an approximation y_i to $x(ih)$ on the mesh $t_i = ih$, $i = 0, 1, \ldots$. (Here h is the mesh increment.) A well-known class of such methods is the class of linear multistep methods

$$\sum_{i=0}^{p} \alpha_i y_{n-i} + h \sum_{i=0}^{p} \beta_i \dot{y}_{n-i} = 0. \tag{1.2}$$

In this recurrence relation, $\dot{y}_{n-i}$ stands for $f(y_{n-i}, t_{n-i})$ when the relation is used to generate the approximation to x.

If the Jacobian, $f_x(x, t)$, (at least for some values of t and for x the solution of (1.1) of interest) is such that one of its eigenvalues is large in magnitude, then (1.1) is usually referred to as a stiff system. For such systems most of the numerical methods of this pointwise type lose their effectiveness. Many studies have recently appeared characterizing methods which are still effective for stiff equations. These studies include many new methods, not necessarily of the type (1.2), but resembling (1.2) in the sense that values y_i are obtained as approximations to x_i at the points of the mesh t_i, $i = 0, 1, \ldots$. (For a review of such methods see [1] and [5].) Most of these special methods are effective only if the eigenvalues of the Jacobian f_x which are large in magnitude are indeed large in magnitude because their real parts are large and negative.

When the stiffness of the system (1.1) is not of this usual type, the solutions of (1.1) are highly oscillatory. Recently, some studies have appeared which deal

Received July 21, 1975.
AMS (MOS) subject classifications (1970). Primary 65L05.
*This work was performed while the author was a visiting Professor of Mathematics at the University of Paris-Sud.
**The work of this author was supported by the U. S. Air Force Office of Scientific Research under grant AF-AFOSR-2363-B.

Copyright © 1976, American Mathematical Society

numerically with this oscillatory case (cf. [3], [4] and [6]).

The difficulty in dealing with this case is illustrated by the following example.

$$f = \lambda \begin{pmatrix} y \\ -x + \lambda \sin t \end{pmatrix}, \quad x(0) = \begin{pmatrix} 0 \\ \lambda a + \dfrac{1}{1 - 1/\lambda^2} \end{pmatrix}. \tag{1.3}$$

This corresponds to the single second order scalar equation

$$\ddot{x} + \lambda^2 x = \lambda^2 \sin t \tag{1.4}$$

with the solution

$$x(t) = a \sin \lambda t + \frac{\sin t}{1 - 1/\lambda^2}. \tag{1.5}$$

The eigenvalues of f_x are $\pm i\lambda$, and when λ is large, the solution consists of the high frequency carrier wave $a \cdot \sin \lambda t$, modulated by the slow wave, $(\sin t)/(1 - 1/\lambda^2)$. In fact, for large λ, the solution is a curve which practically is space filling. The specification of the value at a point (of a mesh) of such a solution is an ill-posed problem. We may expect numerical methods which furnish approximations to the value of the solution at a point to be ill posed as well (i.e. ill-conditioned or unstable).

In this paper, we present a preliminary study of numerical methods for approximating the solution at discrete times, which are properly posed even for highly oscillatory problems. We preserve the linear multistep form (1.2) of the numerical methods since this form provides desirable computational and analytic properties.

Our point of departure is to note that (1.2) is a linear combination of linear functionals of the solution which are its values and the values of its derivative at mesh points. This type of functional is unstable for solutions of stiff systems. What we do is to replace these functionals by stable ones so that the corresponding numerical method is well conditioned.

There is a wide choice of functionals which might be used, but they seem to consist of those stable functionals which together supply the following two features. First, the functionals are to give information about the solution of the differential equation which is acceptable as a description of the solution. (This is the purpose of the point evaluations, y_n, in (1.2).) Second, the functionals convey constraints imposed on the solution. (This is the purpose of the functional $\dot{y}_n$ in (1.2) and its replacement by $f(y_n, t_n)$.) The choice of appropriate functionals may depend on the problem and the solution being calculated.

We do not deal here with the questions of characterizing these types of stable functionals of the solution of a system of differential equations. Rather, we select two special functionals which are an averaging functional and an appropriate evaluation functional, which ought to be stable in the sense discussed. Then, we show how to construct an effective class of numerical methods of a linear multistep type out of these two functionals.

In Section 2 we start with the scalar case. Sections 2.2–2.8 deal with the purely

linear problem. In Section 2.2 we specify the choice of functionals. Then in Section 2.3 we introduce appropriate reproducing kernel Hilbert spaces to obtain representers of these functionals. Using these representers in Section 2.4, we derive generalized moment conditions and give a local error analysis. Then, in Section 2.5, we study the stability of these methods by a root condition approach; and we subsequently develop a global error analysis. This is followed in Section 2.6 by a set of examples of our methods. Next, in Section 2.7 we give the results of computations based on the sample problem (1.4) for the various sets of the example methods. Then, in Section 2.8 we describe a class of so-called optimal methods of the type being considered. Finally, in Section 2.9 we develop an error analysis for the nonlinear problem.

In Section 3 we describe to what extent our ideas may be carried over to systems of differential equations.

2. The Scalar Case.

2.1. *The Problem Treated.* In this section we develop our method in the context of the problem,

$$\ddot{x} + \lambda^2 x = f(x, t), \qquad t \in [0, T],$$
$$x(0) = x_0, \tag{2.1}$$

where x and f are scalars.

The solution of this problem will be required to exist on the larger interval $I = [-\tau, T]$, where the quantity $\tau > 0$ will be specified in (2.7). Thus, we assume that $f(x, t)$ is continuous in t, $t \in I$ and Lipschitz continuous in x for all such t, with Lipschitz constant K. In particular, $f(x, t)$ is uniformly bounded for $t \in I$ and x restricted to any compact real set including in particular the set of values taken on by the solutions $x(t)$ for $t \in I$.

In Sections 2.2–2.8, we restrict our attention to the linear problem in which $f(x, t) = f(t)$. Then in Section 2.9, we discuss the full nonlinear case.

2.2. *Choice of Functionals.* Let $N > 0$ be an integer, let $h = T/N$ and let $t_i = ih$, $i = 0, \pm 1, \ldots$, be the points of a mesh. We seek the functional $y(t)$ of x at points of this mesh. Let $z(t)$ be a functional of x which can be calculated at each mesh point. Then, we seek to determine $y_n = y(t_n)$, $n > 0$, in terms of y_{n-i}, $i = 1, \ldots, r$, and $z_{n-i} = z(t_{n-i})$, $i = 0, 1, \ldots, s$, by means of the linear multistep formula

$$\sum_{i=0}^{r} a_i y_{n-i} + \sum_{i=0}^{s} b_i z_{n-i} = 0, \qquad n = 0, 1, \ldots, N. \tag{2.2}$$

The initial values y_i and z_i, $i = -1, \ldots, -r$, are assumed to be furnished by some independent means.

In the case (2.1) of interest and λ large we choose $y(t)$ to be

$$y(t) = \int_{-\infty}^{\infty} k(t - s)x(s)ds, \tag{2.3}$$

where

$$k(z) = \frac{1}{\Delta}\begin{cases} 1, & 0 < z < \Delta, \\ 0, & \text{otherwise.} \end{cases} \tag{2.4}$$

Thus, $y(t)$ represents the average of $x(t)$ over the interval $[t - \Delta, t]$.

The functional $z(t)$ is chosen to be $[d^2/dt^2 + \lambda^2]\, x(t)$, i.e., $f(t)$, which can be calculated at each mesh point. Thus, with a change in normalization (2.2) may be written as

$$y_n = \sum_{i=1}^{r} c_i y_{n-i} + h^2 \sum_{i=0}^{s} d_i f_{n-i}. \tag{2.5}$$

2.3. *Representers.* We introduce the reproducing kernel space, $H \equiv H_m$ which is the Sobolev space $W_m^2[-\infty, \infty]$ with the inner product

$$\langle f, g\rangle = \sum_{j=0}^{m} \binom{m}{j} (f^{(j)}, g^{*(j)}), \tag{2.6}$$

where

$$(f, g) = \int_{-\infty}^{\infty} f(t) g^*(t)\, dt.$$

An asterisk is used to denote the complex conjugate throughout. Since we are interested in solutions of (2.1) on the interval

$$I = [-h\Delta, T], \tag{2.7}$$

we may identify both a solution of (2.1) and $f(t)$ appearing in (2.1) with the unique functions of minimal norm in H with which they agree on I, respectively. Of course, on I it is sufficient for f to have $m - 1$ absolutely continuous derivatives and an mth derivative a.e. which is square integrable.

We use a caret to denote the Fourier transform, viz.

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{i\omega t} \hat{f}(\omega)\, d\omega, \quad \hat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-i\omega t} f(t)\, dt. \tag{2.8}$$

Then, the inner product in H may be written as

$$\langle f, g\rangle = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(\omega) \hat{g}^*(\omega) |p_m(\omega)|^2\, d\omega, \tag{2.9}$$

where

$$p_m(\omega) = (1 - i\omega)^m. \tag{2.10}$$

The reproducing kernel in H is

$$R_t \equiv R_t^m(s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{e^{i(s-t)\omega}}{p_m(\omega)^2}\, d\omega. \tag{2.11}$$

A second Hilbert space, $\hat{H}$ is introduced as follows:

$$\hat{H} \equiv \hat{H}_m = \{\hat{f} \mid \hat{f} p_m \in L_2\}. \tag{2.12}$$

The inner product in $\hat{H}$ is

$$\langle \hat{f}, \hat{g} \rangle_{\wedge} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}\hat{g}^* |p_m|^2 \, d\omega. \tag{2.13}$$

(2.9) defines an isometric isomorphism between H and $\hat{H}$. The symbol $\sim$ will denote this isomorphism. Then from (2.11) we see that the isomorphism between R_t and its image in $\hat{H}$ is expressed by

$$R_t \sim e^{-i\omega t}/|p_m(\omega)|^2. \tag{2.14}$$

For the representer, η_t of $d^2/dt^2 + \lambda^2$, we have

$$\eta_t \equiv R_t'' + \lambda^2 R_t \sim (-\omega^2 + \lambda^2) e^{-i\omega t}/|p_m(\omega)|^2. \tag{2.15}$$

For the representer k_t of $y(t)$ given by (2.3) and (2.4), we have

$$\begin{aligned} k_t \equiv k_t(s) &= \frac{1}{\Delta} \int_{t-\Delta}^{t} R_u'(s)\, du \sim \frac{1}{\Delta} \int_{t-\Delta}^{t} \frac{e^{-i\omega u}}{|p_m(\omega)|^2} \, du \\ &= \frac{1}{|p_m(\omega)|^2} e^{-i\omega t} \left[\frac{1 - e^{-i\omega\Delta}}{-i\omega\Delta} \right] = \frac{e^{-i\omega t}}{|p_m(\omega)|^2} \sqrt{2\pi}\, \hat{k}(\omega), \end{aligned} \tag{2.16}$$

where $\hat{k}(\omega)$ is the Fourier transform of $k(z)$ given in (2.4).

With these representers, the formula (2.5) leads us to introduce the following linear functional g_n.

$$g_n \equiv g_n[x] \equiv \left\langle k_{t_n} - \sum_{i=1}^{r} c_i k_{t_{n-i}} - h^2 \sum_{i=0}^{s} d_i \eta_{t_i}, x \right\rangle. \tag{2.17}$$

g_n will be zero if x is the numerical solution. In general, g_n is not zero and is the analogue of the local truncation error for classical linear multistep schemes.

2.4. *Local Error and Generalized Moment Conditions.* g_n is characterized in the following definition.

Definition 2.1. Using (2.17) as a definition, we call the linear functional, g_n appearing there, the local truncation error of the method (2.5).

To estimate the local truncation error we write

$$|g_n[x]| \leqslant \left\| k_{t_n} - \sum_{j=1}^{r} c_j k_{t_{n-j}} - h^2 \sum_{j=0}^{s} d_j \eta_{t_{n-j}} \right\|_{\wedge}^2, \tag{2.18}$$

where, as usual,

$$\|x\|^2 = \langle x, x \rangle \quad \text{and} \quad \|x\|_{\wedge}^2 = \langle x, x \rangle \ . \tag{2.19}$$

We will drop the subscript, $\wedge$, since no confusion should result.

Now using (2.13), (2.15) and (2.16), we find for the right member of (2.18) that

$$\left\| k_{t_n} - \sum_{j=1}^{r} c_j k_{t_{n-j}} - h^2 \sum_{j=0}^{s} d_j \eta_{t_{n-j}} \right\|^2 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |t(\omega)|^2 \frac{d\omega}{|p_m(\omega)|^2}, \tag{2.20}$$

where

$$t(\omega) = \sqrt{2\pi}\,\hat{k}(\omega)\sum_{j=0}^{r} s_j e^{ij\omega h} - h^2(\lambda^2 - \omega^2)\sum_{j=0}^{s} d_j e^{ij\omega h}. \tag{2.21}$$

Here

$$s_0 = 1 \quad \text{and} \quad s_j = -c_j, \qquad j = 1, \ldots, r. \tag{2.22}$$

Expanding $t(\omega)$, formally in a Taylor series with remainder, gives

$$t(\omega) = \sum_{l=0}^{p-1} (ih\omega)^l m_l + R_p, \tag{2.23}$$

where from (2.21) and (2.22) we obtain

$$\begin{aligned} m_l = {} & \frac{1}{(l+1)!}\sum_{k=1}^{l+1}\binom{l+1}{k} L^{k-1}\sum_{j=0}^{r} j^{1+l-k}s_j \\ & - \frac{h^2\lambda^2}{l!}\sum_{j=0}^{s} j^l d_j - \frac{1}{(l-2)!}\sum_{j=0}^{s} j^{l-2}d_j \end{aligned} \tag{2.24}$$

and

$$\begin{aligned} R_p = {} & \frac{(ih\omega)^p}{p!}\left[-\frac{1}{L(p+1)}\sum_{j=0}^{r} s_j\left(j^{p+1}e^{ijh\omega_{j,1}} - (j+L)^{p+1}e^{ijh(1+L)\omega_{j,2}}\right)\right. \\ & \left. - h^2\lambda^2\sum_{j=0}^{s} j^p d_j e^{ij\omega_{j,3}} - p(p-1)\sum_{j=0}^{s} j^{p-2}d_j e^{ij\omega_{j,4}}\right]. \end{aligned} \tag{2.25}$$

In (2.24) and (2.25) we have used

$$L = \Delta/h. \tag{2.26}$$

That is in terms of the functional k of (2.3) and (2.4), the interval, Δ, over which the average is taken is a multiple, L, of the mesh increment h. In (2.25) the quantities $\omega_{j,1}$ and $\omega_{j,2}, j = 0, \ldots, r$, and $\omega_{1,3}$ and $\omega_{j,4}, j = 0, \ldots, s$, are values of ω which arise from the calculation of the remainder in Taylor's theorem.

The quantities m_l are characterized in the following definition.

Definition 2.2. We call the $m_l, l = 0, 1, \ldots$, the (generalized) moments (of the coefficients). Analogously, $m_l = 0, l = 0, 1, \ldots$, will be called the (generalized) moment conditions.

Consider the following remark.

Remark 2.1. View the equations $m_l = 0, l = 0, \ldots, r-1$, as r equations for the r unknowns $s_j, j = 1, \ldots, r$. The lth row of the resulting coefficient matrix which has

$$\frac{1}{(l+1)!}\sum_{k=1}^{l}\binom{l+1}{k} L^{k-1}j^{1+l-k}, \tag{2.27}$$

for its jth term is a linear combination of the first l rows of the Vandermonde matrix. Thus, the system of r equations has a solution in this case. Indeed, by choosing the

$d_j, j = 0, \ldots, s$, to be proportional to λ^{-2}, we obtain a solution for the $s_j, j = 1, \ldots, r$, which is $O(1) + O(\lambda^{-2})$.

From the form of $t(\omega)$ given in (2.21) we may make the following remark, the assertion of which follows from a familiar argument which proceeds by breaking up the range of integration in (2.20), appropriately.

Remark 2.2. If p is chosen less than m and the coefficients $s_j, j = 1, \ldots, r$, and $d_j, j = 0, \ldots, s$, are chosen as solutions of the generalized moment equations $m_l = 0, l = 0, 1, \ldots, p$, we may obtain an estimate of the local truncation error of the form

$$\|g_n\| \equiv \max_{x \in H} |g_n| \leqslant O(h^{p+1}), \qquad p < m, \tag{2.28}$$
$$\|x\| \leqslant 1.$$

We collect these remarks into the following theorem.

THEOREM 2.1. *There exists a choice of coefficients $s_j, j = 1, \ldots, r$, and d_j, $j = 0, \ldots, s$, such that the local truncation error has a bound of the form* (2.28). *Moreover, this bound is uniform in λ for $|\lambda| \geqslant \lambda_0 > 0$.*

2.5. *Stability and Global Error Analysis.* y_n, $n = 0, 1, \ldots$, denotes the values obtained by the multistep formula, (2.5), from the initial values, y_n, $n = -r, \ldots, -1$. Let $Y_n, n = -r, -r+1, \ldots$, denote the exact values of these functionals. Let

$$e_n = y_n - Y_n, \qquad n = -r, -r+1, \ldots, \tag{2.29}$$

denote the cumulative error. For convenience, assume that the initial functionals $e_n = 0, n = -r, -r+1, \ldots, -1$.

Subtracting the following identity,

$$Y_n = \sum_{j=1}^{r} c_j Y_{n-j} + h^2 \sum_{j=0}^{s} d_j f_{n-j} + Y_n - \sum_{j=1}^{r} c_j Y_{n-j} - h^2 \sum_{j=0}^{s} d_j f_{n-j}, \tag{2.30}$$

from (2.5), we get

$$e_n = \sum_{j=1}^{r} c_j e_{n-j} + g_n. \tag{2.31}$$

Here,

$$g_n = -Y_n + \sum_{j=1}^{r} c_j Y_{n-j} + h^2 \sum_{j=0}^{s} d_j f_{n-j} \tag{2.32}$$

is the value of the linear functional, g_n, of (2.17) applied to x, the exact solution of the initial value problem, (2.1). To solve (2.31) for e_n, we use the polynomial $S(z)$:

$$S(z) = \sum_{j=0}^{r} s_j z^{r-j}. \tag{2.33}$$

Since $s_0 = 1$, $[z^r S(z^{-1})]^{-1}$ is an analytic function of z in a neighborhood of $z = 0$. Then, let its power series be given by

$$[z^r S(z^{-1})]^{-1} = \sum_{j=0}^{\infty} \sigma_j z^j. \tag{2.34}$$

Now multiply (2.31) by σ_{N-n} and sum the result over n from 0 to N. The result is the solution of (2.31):

$$e_N = \sum_{n=0}^{N} \sigma_{N-n} g_n. \tag{2.35}$$

We use the following definition.

Definition 2.3 (*Stability*). If the sequence $\{\sigma_j, j = 0, 1, \ldots\}$ is bounded, then the method is said to be stable.

We recall the following definition.

Definition 2.4. $S(z)$ is said to obey the root condition if all of its roots lie in the closed unit disc while those of its roots which lie on the boundary of that disc are simple.

With this definition we may state the following lemma which characterizes the stability of the method.

LEMMA 2.1. *If the polynomial $S(z)$ obeys the root condition then the sequence $\{\sigma_j, j = 0, 1, \ldots\}$ is bounded, i.e. the method is stable.*

(For a proof of this lemma see [2, p. 720].)

If this lemma is applicable, (2.35) gives

$$|e_N| \leqslant \text{const } N \max_{r \leqslant n \leqslant N} \|g_n\| \, \|x\|, \tag{2.36}$$

where x is the exact solution of (2.1).

Combining this with (2.28) gives the following theorem.

THEOREM 2.2. *If the choice of coefficients characterized in Theorem* 2.1 *give rise to a stable method, then for the method* (2.5) *with those coefficients,*

$$\|e_N\| \leqslant O(h^p), \quad p < m, \tag{2.37}$$

uniformly in λ *for* $|\lambda| \geqslant \lambda_0 > 0$.

2.6. *Examples.* We now consider some examples of methods of the type (2.5) in which the coefficients are determined by the generalized moment conditions.

From (2.24) we have for $l = 0$, 1, and 2, respectively,

$$0. \qquad m_0 \equiv \sum_{j=0}^{r} s_j - h^2\lambda^2 \sum_{j=0}^{s} d_j,$$

$$1. \qquad m_1 \equiv \sum_{j=0}^{r} j s_j + \frac{L}{2} \sum_{j=0}^{r} s_j - h^2\lambda^2 \sum_{j=0}^{s} j d_j, \tag{2.38}$$

$$2. \quad m_2 \equiv \frac{1}{2} \sum_{j=0}^{r} j^2 s_j + \frac{L}{2} \sum_{j=0}^{r} j s_j + \frac{L^2}{6} \sum_{j=0}^{r} s_j - \frac{h^2\lambda^2}{2} \sum_{j=0}^{s} j^2 d_j - \sum_{j=0}^{s} d_j.$$

Consider the case:

A. $m_0 = m_1 = 0$.

For $r = s = 1$, we get

$$c_1 = 1 - \frac{2}{L} + \frac{2}{L} h^2\lambda^2 d_0, \quad d_1 = \frac{2}{h^2\lambda^2 L} - \left(\frac{2}{L} + 1\right) d_0. \tag{2.39}$$

In the special case $d_0 = 0$, (2.39) becomes

$$\text{I} \quad \begin{cases} c_1 = 1 - \dfrac{2}{L}, \\ d_1 = \dfrac{2}{h^2\lambda^2 L}. \end{cases} \tag{2.40}$$

These coefficients (i.e. c_1) obey the root condition if and only if

$$L \geqslant 1. \tag{2.41}$$

In the special case $d_0 = d_1$, (2.39) becomes

$$\text{II} \quad \begin{cases} c_1 = 1 - \dfrac{2}{L+1}, \\ d_0 = d_1 = \dfrac{1}{h^2\lambda^2}\dfrac{1}{L+1}. \end{cases} \tag{2.42}$$

Under the restriction $L \geqslant 0$, the root condition is equivalent to $L \geqslant 0$, for the coefficients (2.42). For $r = s = 2$,

$$\begin{aligned} c_1 &= 1 - \frac{2}{L} - \left(1 + \frac{2}{L}\right)c_2 + \frac{2}{L}\lambda^2 h^2 (d_0 - d_1), \\ d_1 &= \frac{2}{L\lambda^2 h^2}(1 + c_2) - \left(1 + \frac{2}{L}\right)d_0 - \left(1 - \frac{2}{L}\right)d_2. \end{aligned} \tag{2.43}$$

In the special case $d_0 = 0, c_1 = c_2, d_1 = d_2$, (2.43) becomes

$$\text{III} \quad \begin{cases} c_1 = c_2 = \dfrac{L-3}{2L}, \\ d_1 = d_2 = \dfrac{3}{2\lambda^2 h^2 L}. \end{cases} \tag{2.44}$$

In this case

$$S(z) = z^2 - \frac{L-3}{2L}z - \frac{L-3}{2L}$$

and this polynomial, $S(z)$, obeys the root condition for a set of values of L which includes $L \geqslant 0$.

In the special case $c_1 = c_2, d_1 = d_2 = 0$, (2.43) becomes

$$\text{IV} \quad \begin{cases} c_1 = c_2 = \dfrac{1}{2}\dfrac{L}{3+L}, \\ d_0 = \dfrac{1}{\lambda^2 h^2}\dfrac{3}{3+L}. \end{cases} \tag{2.45}$$

Here

$$S(z) = z^2 - \frac{1}{2}\frac{L}{3+L}z - \frac{1}{2}\frac{L}{3+L}.$$

This polynomial obeys the root condition for a set of values of L which includes $L \geqslant 0$.

In the special case $c_1 = c_2$, $d_0 = d_1 = d_2$, (2.43) becomes

$$\text{V} \begin{cases} c_1 = c_2 = \dfrac{1}{2}\dfrac{L-1}{L+1}, \\ d_0 = d_1 = d_2 = \dfrac{2}{3\lambda^2 h^2}\dfrac{1}{1+L}. \end{cases} \tag{2.46}$$

In this case the root condition is obeyed for $L \geqslant 0$. Now we consider a case corresponding to three moment conditions:

B. $m_0 = m_1 = m_2 = 0$.

For $r = s = 1$, we get

$$\text{VI} \begin{cases} c_1 = 1 - L \Big/ \left(\dfrac{L^2}{3} + \dfrac{L}{2} - \dfrac{2}{h^2\lambda^2} \right), \\ d_0 = \dfrac{1}{\lambda^2 h^2}\left[1 - L + L^2 \Big/ \left(\dfrac{2}{3}L^2 + L - \dfrac{4}{h^2\lambda^2} \right) \right], \\ d_1 = \dfrac{1}{\lambda^2 h^2}\left[-1 + L + (2L - L^2) \Big/ \left(\dfrac{2}{3}L^2 + L - \dfrac{4}{h^2\lambda^2} \right) \right]. \end{cases} \tag{2.47}$$

Notice that the root condition is obeyed for L large and positive but is violated for $h\lambda$ small compared to L.

Remark 2.3. In all of these examples and in the general case, we see that the coefficients obtained as solutions of the moment equations depend on λ^2. At first sight this seems to be more restrictive than the case of the classical linear multistep formulas where the coefficients of the formula do not depend on the coefficients of the differential equation. In the classical case the coefficients of the differential equations enter into the method when it is used to approximate the differential equation, e.g., when in (1.2) $\dot{y}_{n-i}$ is replaced by $f(y_{n-i}, t_{n-i})$. It is essential after all that the numerical method at some point be dependent on the equation to be solved. In our case this dependence occurs at the outset in the determination of coefficients and in the error analysis. In the classical case it enters in the error analysis and in the use of the method. The difference seems formal and in fact it may be that a treatment of the present problem may yet be found which resembles this classical feature but retains the more general functional aspects discussed here.

2.7. *Illustrative Computations.* We now apply the six sets of methods, labeled I, II, . . . , VI in Section 2.6 respectively, to the sample problem

$$\ddot{x} + \lambda^2 x = \lambda^2 \sin t, \qquad x(0) = 0, \quad x'(0) = \frac{\lambda}{2} + \frac{1}{1 - 1/\lambda^2}. \tag{2.48}$$

Runs are made over the interval $[0, T] = [0, \pi]$. In the following Table 2.1, we display the $h^{1/2} \cdot l_2$-norm of the cumulative error,

$$\|e\|_{l_2} \equiv \left[h \sum_{n=0}^{[\pi/h]} e_n^2 \right]^{1/2}, \tag{2.49}$$

for a set of various combinations of $h = .1, .01$, $\lambda = 10, 10^3, 10^5$ and $L = 1, 2, 3$ and for each of the six methods cited.

Method	$\lambda \backslash L$	1	2	3	1	2	3
I	10	.273	.108	.112	.133	.126	.126
	10^3	.113	.00217	.0611	.0283	.00683	.0083
	10^5	.112	.00209	.0611	.0111	.000106	.00627
II	10	.122	.133	.155	.126	.127	.128
	10^3	.00125	.0622	.177	.0241	.00926	.0136
	10^5	.00104	.0621	.177	.000118	.00627	.0125
III	10	.242	.111	.0872	.136	.126	.126
	10^3	.0032	.00422	.00317	.0294	.00684	.00546
	10^5	.0034	.00419	.00313	.00023	.00112	.89E-6
IV	10	.123	.111	.0938	.126	.126	.126
	10^3	.00627	.0144	.0244	.0241	.00684	.00546
	10^5	.00623	.0144	.0244	.000133	.000179	.000264
V	10	.144	.152	.156	.127	.127	.128
	10^3	.0657	.094	.119	.0249	.0116	.0136
	10^5	.0657	.0939	.119	.0063	.00942	.0125
VI	10	.758E4	.66E11	.124	.195E1	.471E1	.11E2
	10^3	.0447	.0639	.244	.0246	.00901	.0253
	10^5	.0447	.0639	.244	.00421	.00629	.0251
h			.1			.01	

$\|e\|_{l_2}$

TABLE 2.1

To illustrate both the favorable and unfavorable effects in our methods Table 2.1 contains cases for which the methods are designed to operate well, along with cases to which correspond poor or nonsensical results. For example, although the cases corresponding to $\lambda = 10$ give fair results, these cases are not stiff; and we should not expect good results. When h is decreased improvement should occur but only for the stiff cases. The cases $\lambda = 10^3$ and $h = .01$ are not stiff and improvement with decreasing h does not always occur in these cases. Method VI is used in some unstable cases. Examining (2.25), we see that R_p is proportional to L^p. Thus, in some cases as L increases, we see an improvement due to improving the averaging (i.e. increasing Δ), but ultimately a degradation due to the L dependence of R_p. The stiff cases for moderate L give extremely good results, as we expect.

2.8. *Optimal Methods.* The choice of coefficients characterized by the moment conditions and the root condition gives rise to a stable method with the error estimate (2.37). We now turn to the question of characterizing those coefficients which give the best possible (local) error estimates, and we call the corresponding methods optimal methods.

These methods are obtained by determining the best approximation to k_{t_n} in the span of

$$\{k_{t_{n-i}}, h^2\eta_{t_{n-j}} | i = 1, \ldots, r; j = 0, \ldots, s\}. \tag{2.50}$$

This, in turn, corresponds to minimizing the norm of the local truncation error, g_n.

$$\begin{aligned} \|g_n\|^2 &= \left\| k_{t_n} - \sum_{j=1}^{r} c_j k_{t_{n-j}} - h^2 \sum_{j=0}^{s} d_j \eta_{t_{n-j}} \right\|^2 \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |t(\omega)|^2 \frac{d\omega}{|p_m(\omega)|^2}. \end{aligned} \tag{2.51}$$

(See (2.18), (2.20) and (2.21).)

The solution of this minimization problem may be described as follows:

Let

$$\begin{aligned} \theta = (\theta_\nu) &= (c_1, \ldots, c_r, d_0, \ldots, d_s), \\ K = (K_\nu) &= (k_{t_{n-1}}, \ldots, k_{t_{n-r}}, \eta_{t_n}, \ldots, \eta_{t_{n-s}}) \end{aligned} \tag{2.52}$$

be $r + s + 1$ vectors of scalars and functionals, respectively. Let

$$\Sigma = (\Sigma_{i,j}) = (\langle K_i, K_j \rangle) \tag{2.53}$$

be the $(r + s + 1) \times (r + s + 1)$ Grammian matrix of the functionals composing K.

The optimization problem is

$$\min_{\theta} \left\| k_{t_n} - \sum_{\nu=1}^{r+s+1} \theta_\nu K_\nu \right\|^2. \tag{2.54}$$

Let $\theta = \bar{\theta}$ be the solution of this problem. Then

$$\bar{\theta} = \Sigma^{-1} \upsilon, \tag{2.55}$$

where υ is the $r + s + 1$ vector whose νth component is

$$\upsilon_\nu = \langle k_{t_n}, I_\nu \rangle, \quad \nu = 1, \ldots, r + s + 1. \tag{2.56}$$

The question of whether or not the optimal method is stable is open. However, since the local error for the optimal method is by definition the minimal local error, then the estimate (2.28) is valid for the optimal method as well.

As an example of an optimal method, consider the explicit optimal method, ($d_0 = 0$) in the case $r = s = 1$. In this case the optimal coefficients $\bar{c}_1$ and $\bar{d}_1$ are solutions of the following linear system:

$$\begin{bmatrix} \int_{-\infty}^{\infty} \frac{|\hat{k}(\omega)|^2}{|p_m(\omega)|^2}\,d\omega & \int_{-\infty}^{\infty} \hat{k}(\omega)\frac{\lambda^2-\omega^2}{|p_m(\omega)|^2}\,d\omega \\ \int_{-\infty}^{\infty} \hat{k}^*(\omega)\frac{\lambda^2-\omega^2}{|p_m(\omega)|^2}\,d\omega & \int_{-\infty}^{\infty} \frac{\lambda^2-\omega^2}{|p_m(\omega)|^2}\,d\omega \end{bmatrix} \begin{bmatrix} \bar{c}_1 \\ h^2\bar{d}_1 \end{bmatrix}$$

(2.57)

$$= \begin{bmatrix} \int_{-\infty}^{\infty} e^{-i\omega h}\frac{|\hat{k}(\omega)|^2}{|p_m(\omega)|^2}\,d\omega \\ \int_{-\infty}^{\infty} e^{-i\omega h}\hat{k}(\omega)\frac{\lambda^2-\omega^2}{|p_m(\omega)|^2}\,d\omega \end{bmatrix}.$$

Notice that by expanding $e^{-i\omega h}$ in Taylor's series with remainder, we get

$$\int_{-\infty}^{\infty} e^{-i\omega h}\frac{|\hat{k}(\omega)|^2}{|p_m(\omega)|^2}d\omega = \int_{-\infty}^{\infty}\left[1-\frac{\omega^2h^2}{2}e^{-i\omega_1 h}\right]\frac{|\hat{k}(\omega)|^2}{|p_m(\omega)|^2}\,d\omega, \tag{2.58}$$

since $|\hat{k}|$ and $|p_m(\omega)|^2$ are even functions. Similarly,

$$\int_{-\infty}^{\infty} e^{-i\omega h}\hat{k}(\omega)\frac{\lambda^2-\omega^2}{|p_m(\omega)|^2}\,d\omega = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left[1-\frac{\omega^2h^2}{2}(L-1)^2e^{-i\omega_3 h(L-1)}\right] \cdot \frac{\sin\omega\Delta/2}{\omega\Delta/2}\frac{\lambda^2-\omega^2}{|p_m(\omega)|^2}\,d\omega \tag{2.59}$$

and

$$\int_{-\infty}^{\infty} e^{-i\omega h}\hat{k}^*(\omega)\frac{\lambda^2-\omega^2}{|p_m(\omega)|^2}\,d\omega = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left[1-\frac{\omega^2h^2}{2}(L-1)^2\,e^{-i\omega_3 h(L-1)}\right] \cdot \frac{\sin\omega\Delta/2}{\omega\Delta/2}\frac{\lambda^2-\omega^2}{|p_m(\omega)|^2}\,d\omega. \tag{2.60}$$

Thus, if $m > 2$,

$$\bar{c}_1 = 1 + O(h^2). \tag{2.61}$$

While this does not imply stability in the strict sense, it does lead to the global error estimate,

$$\|e_N\| \leqslant \text{const}\, N \max_{1\leqslant n\leqslant N}\|g_n\|\,e^{hT}, \tag{2.62}$$

by a well-known argument, as an alternate to the estimate (2.36). Inspection of (2.57) shows that $\bar{d}_1 = O(\lambda^{-2})$. Thus, the estimate (2.62) is uniform in λ for $|\lambda| \geqslant \lambda_0 > 0$. Thus, we find that the optimal method is convergent in the case in question, uniformly in λ for $|\lambda| \geqslant \lambda_0 > 0$.

2.9. *The Nonlinear Case.* Here we return to the nonlinear problem (2.1). We describe how the method discussed above is modified to handle this case and then we obtain an error estimate.

Since in the nonlinear case $f(x, t)$ cannot be computed as we proceed along the

mesh,we replace (2.5) by

$$y_n = \sum_{i=1}^{r} c_i y_{n-i} + h^2 \sum_{i=0}^{s} d_i f_{n-i}(y_{n-i}). \tag{2.63}$$

Here we use $f_{n-i}(y_{n-i})$ to denote $f(y_{n-i}, t_{n-i})$. We replace (2.30) as well by the following expression

$$\begin{aligned} Y_n &= \sum_{j=1}^{r} c_j Y_{n-j} + h^2 \sum_{j=0}^{s} d_j f_{n-j}(x_{n-j}) \\ &\quad + Y_n - \sum_{j=1}^{r} c_j Y_{n-j} - h^2 \sum_{j=0}^{s} d_j f_{n-j}(x_{n-j}). \end{aligned} \tag{2.64}$$

Substracting (2.64) from (2.63), we find the following equation for the global error, e_n.

$$e_n = \sum_{j=1}^{r} c_j e_{n-j} + g_n + h^2 \sum_{j=0}^{s} d_j [f_{n-j}(y_{n-j}) - f(x_{n-j})]. \tag{2.65}$$

Now the linear functional $x(t) - Y_t$ has a representer in H which we denote by v_t. Then

$$v_t \sim \frac{e^{-i\omega t}}{|p_m(\omega)|^2} \; (1 - \sqrt{2\pi}\, \hat{k}(\omega)). \tag{2.66}$$

Then

$$\|v_t\|^2 = \int_{-\infty}^{\infty} \left| 1 + \frac{1 - e^{-i\omega\Delta}}{i\omega\Delta} \right|^2 \frac{d\omega}{|p_m(\omega)|^2}. \tag{2.67}$$

Using Taylor's theorem with remainder, $e^{i\omega\Delta} = 1 - i\omega\Delta - \omega^2\Delta^2 e^{-i\tilde{\omega}\Delta}/2$, (2.67) becomes

$$\|v_t\| = v_m L h, \tag{2.68}$$

where

$$v_m = \frac{1}{2}\left[\int_{-\infty}^{\infty} \frac{|\omega|^2}{|p_m(\omega)|^2}\, d\omega\right]^{1/2}, \tag{2.69}$$

when $m > 1$.

Using the mean value theorem,

$$\begin{aligned} f_{n-j}(y_{n-j}) - f(x_{n-j}) &= f_{n-j}(y_{n-j}) - f_{n-j}(Y_{n-j}) + f_{n-j}(Y_{n-j}) - f_{n-j}(x_{n-j}) \\ &= f_{\tilde{x}_{n-j}}\, e_{n-j} + f_{\tilde{\tilde{x}}_{n-j}} \langle v_t, x \rangle. \end{aligned} \tag{2.70}$$

Here $f_{\tilde{x}_k}$ is used to denote $f_x(\tilde{x}_k, t)$ with $f_{\tilde{\tilde{x}}_k}$ analogously defined. Inserting (2.70) into (2.65) gives

$$\begin{aligned} \left(1 - h^2 \sum_{j=0}^{s} d_j f_{\tilde{x}_{n-j}}\right) e_n &= \sum_{j=1}^{\max(r,s)} (c_j + h^2 d_j f_{\tilde{x}_{n-j}})\, e_{n-j} \\ &\quad + g_n + h^2 \sum_{j=0}^{s} d_j f_{\tilde{\tilde{x}}_{n-j}} \langle v_t, x \rangle. \end{aligned} \tag{2.71}$$

Additional coefficients c_j or d_j needed in (2.71) are presumed to have the value zero.

From (2.71) we may derive an error estimate characterized in the following theorem. This theorem makes use of the quantity

$$\epsilon_1 = \max_j |h^2 d_j K|. \tag{2.72}$$

THEOREM 2.3. *Let the roots of the polynomial* $S(z)$, (*cf.* (2.33)) *lie in the open disc,* $|z| < 1$. *If* ϵ_1 *is sufficiently small, then the method* (2.63) *is stable. If moreover the coefficients* (c_j, d_j), $j = 0, 1, \ldots, \max(r, s)$, *are chosen so that* (2.28) *holds, then*

$$\|e_N\| \leqslant \text{const}\,[h^p + L\epsilon_1 v_m]. \tag{2.73}$$

Proof. The proof of the theorem follows exactly as the proof of Theorem 2.2. If ϵ_1 is sufficiently small, (2.71) is a stable method and may be solved for e_N as in (2.35); however, with g_n in (2.35) replaced by the sum of the last two terms in (2.71). The term const h^p in (2.73) then bounds $\|g_n\|$, while the term const $L\epsilon_1 v_m$ is a bound for the last term in the right member of (2.71), the latter bound obtained by using (2.65).

Remark 2.4. The two terms in the estimate (2.73) are not comparable in orders of h. The first term corresponding to the local truncation error is small for h small. The second term is the error by which a function may be approximated by its average. (2.73) may be viewed as the statement that modulo the error made in replacing a function by its average, the numerical method is globally h^p. Using the coefficients in Section 2.6, we see that this second term, $L\epsilon_1 v_m$, is proportional to L/λ^2 and is thus small with this quantity.

Remark 2.5. Similarly, ϵ_1 may be expected to be small, as required in the hypothesis of Theorem 2.3, when λ is large (if coefficients such as those in Section 2.6 are used).

3. Systems. In this section, we will indicate how the results of Section 2 may be carried over to the more general case of second order systems.

We replace the scalar equation (2.1) by the system

$$\ddot{x} + \Lambda^2 x = f(x, t). \tag{3.1}$$

Here x and f are q-vectors and Λ is a $q \times q$ matrix. There is no explicit requirement that Λ be large in any sense, although for our ideas to be useful we imagine that at least one of the eigenvalues of Λ has a large imaginary component. The functional $y(t)$ defined in (2.3) and the linear multistep formula are formally the same except that the y's which appear are q-vectors and the kernel $k(t - s)$ in (2.3) and coefficients c_i and d_i in (2.5) are $q \times q$ matrices. Similarly, the error equation (2.35) is composed of vectors and matrices as appropriate, viz.

$$e_N = \sum_{n=r}^{N} \sigma_{N-n} g_n \tag{3.2}$$

with

$$g_n = \sum_{j=0}^{r} s_j y_{n-j} + h^2 \sum_{j=0}^{s} d_j f_{n-j}. \tag{3.3}$$

For the local error analysis we replace $\|g_n\|$ in (2.15) by $|||g_n|||$ which simply denotes the q-dimensional norm of the q-vector of function space norms of the components of g_n. For example, if the components of g_n are $g_{n,l}$, $l = 1, \ldots, q$, we could take

$$|||g_n||| = \sum_{l=1}^{q} \|g_{n,l}\|. \tag{3.4}$$

Thus

$$|||g_n||| = \left|\left|\left| \left(\sum_{j=0}^{r} s_j k_{t_{n-j}} - h^2 \sum_{j=0}^{s} d_j \eta_{t_{n-j}} \right) \zeta_q \right|\right|\right|, \tag{3.5}$$

where ζ_q is the q-vector all of whose components are unity. For simplicity we may take k_t to be the scalar function in (2.16) times the $q \times q$ identity matrix, I_q, i.e.

$$k_t \sim \frac{e^{-i\omega t}}{|p_m(\omega)|^2} \sqrt{2\pi}\, \hat{k}(\omega) I_q.$$

However, for η_t we must take

$$\eta_t \equiv R_t'' + \Lambda^2 R_t \sim (\Lambda^2 - \omega^2 I_q) e^{-i\omega t} / |p_m(\omega)|^2.$$

The moments m_l and the remainder R_p are direct analogues of m_l and R_p given in (2.24) and (2.25). For example

$$m_0 \equiv \left(\sum_{j=0}^{r} s_j - h^2 \Lambda^2 \sum_{j=0}^{s} d_j \right) \zeta_q,$$

$$m_1 \equiv \left(\sum_{j=0}^{r} j s_j + \frac{L}{2} \sum_{j=0}^{r} s_j - \Lambda^2 \sum_{j=0}^{s} j d_j \right) \zeta_q.$$

Thus, the local error analysis proceeds along lines similar to the scalar case. The Remark 2.1 is valid also if the matrices s_j, $j = 0, \ldots, r$, are scalars times I_q. Otherwise, the observations about the Vandermonde matrix made in that remark may not be valid.

The global error analysis follows analogously if we use the following lemma concerning the matrix valued polynomial $S(z)$, the analogue of (2.35).

LEMMA 3.1. *Let the determinant $|S(z)|$ of $S(z)$ obey the root condition. If the determinant of s_0 is not zero, then the matrix $[z^r S(z^{-1})]^{-1}$ is analytic in a neighborhood of $z = 0$. Furthermore, the matrices, σ_j, $j = 0, 1, \ldots,$ given by* (2.34) *have bounded norms.*

In our case the determinant $|s_0| = 1$. For a proof of this lemma and other details concerning the global error analyses, cf. [2].

Remark 3.1. Referring to Remark 2.3 and the dependence of the coefficients of the numerical method on the coefficients of the differential equation, we see from (3.8) the way in which the dependence appears in terms of the matrix Λ^2, for the coefficients

determined by generalized moment conditions. It is important to take note that the coefficients depend on the matrix Λ^2 and not explicitly on eigenvalues of Λ^2. Thus, if we know that a system is stiff, with highly oscillatory components, we may use the methods described here without having to calculate the eigenvalues of Λ^2 which cause this stiffness.

Remark 3.2. In the usual systems case for the numerical treatment of differential equations the methods frequently used are the scalar methods with the scalar coefficients simply multiplied by I_q. We suspect that the methods of Section 2 would work in the same way with the simple additional requirements of replacing λ or λ^{-1} by Λ or Λ^{-1}, respectively. At present this remark is only a conjecture, and we defer for a further study its verification.

Mathematical Sciences Department
IBM Thomas J. Watson Research Center
Yorktown Heights, New York 10598

Department of Mathematical Statistics
University of Wisconsin
Madison, Wisconsin 53706

1. G. BJUREL et al., *Survey of Stiff Ordinary Differential Equations*, Royal Inst. Tech., Stockholm, Report NA 70.11.

2. W. L. MIRANKER, "Matricial difference schemes for integrating stiff systems of ordinary differential equations," *Math. Comp.*, v. 25, 1971, pp. 717–728. MR 46 #1094.

3. W. L. MIRANKER & F. HOPPENSTEADT, *Numerical Methods for Stiff Systems of Differential Equations Related with Transistors and Tunnel Diodes*, Lecture Notes in Computer Science, vol. 10, Springer-Verlag, Berlin and New York, 1974, pp. 413 ff.

4. A. D. SNIDER & G. C. FLEMING, "Approximation by aliasing with application to ' Certaine ' stiff differential equations," *Math. Comp.*, v. 28, 1974, pp. 465–473. MR 49 #8377.

5. *Stiff Differential Systems*, IBM Symposia Series, Plenum Press, New York, 1974. (Edited by R. A. Willoughby.)

6. V. AMDURSKY & A. ZIV, *On the Numerical Treatment of Stiff, Highly Oscillatory Systems*, IBM Israel Science Center, Report 015, Haifa, 1974.