1977

A High Order Difference Method for Differential Equation

Robert E. Lynch
Purdue University, rel@cs.purdue.edu

John R. Rice
Purdue University, jrr@cs.purdue.edu

Report Number:
77-244

A HIGH ORDER DIFFERENCE METHOD FOR DIFFERENTIAL EQUATIONS

Robert E. Lynch and John R. Rice
Division of Mathematical Sciences
Purdue University, West Lafayette, IN 47907

CSD-TR 244
September 1977

Abstract

This paper analyzes a high accuracy approximation to the m-th order linear ordinary differential equation $Mu = f$. At mesh points U is the estimate of u and U satisfies $M_nU = I_nf$ where M_nU is a linear combination of values of U at $m+1$ stencil points (adjacent mesh points) and I_nf is a linear combination of values of f at J auxiliary points which are between the first and last stencil points. The coefficients of M_n, I_n are obtained "locally" by solving a small linear system for each group of stencil points in order to make the approximation exact on a linear space S of dimension L. For separated two-point boundary value problems, U is the solution of an n-by-n linear system with full band-width $m+1$. For S a space of polynomials, existence and uniqueness are established and the discretization error is $O(h^{L+1-m})$. For a general set of auxiliary points one has $L = J+m$, but special auxiliary points allow larger L up to $L = 2J+m$. Comparison of operation counts for this method and five other common schemes, shows that this method is among the most efficient. A brief selection from extensive experiments is presented which supports the theoretical results and the practicality of the method.

A HIGH ORDER DIFFERENCE METHOD FOR DIFFERENTIAL EQUATIONS

Robert E. Lynch* and John R. Rice*
Division of Mathematical Sciences
Purdue Unviersity, West Lafayette, IN 47907

1. Introduction. We consider some aspects of a new flexible finite difference method which gives high accuracy approximation to solutions u of linear differential equations $Mu = f$ subject to rather general initial or boundary conditions. The approximation to u is taken as U defined at mesh points as the solution of a system of difference equations $M_n U = I_n f$ together with appropriate boundary conditions; n is used to identify a particular partition of the domain of u. M_n is a difference operator and $M_n U$ is a linear combination of values of U at a small** number of mesh points of a standard stencil; the value of $I_n f$ is equal to a linear combination of values of f at several auxiliary points close to the stencil points and always inside the domain of u.

With appropriate normalization of the coefficients of M_n and I_n, then $I_n f$ is $f + O(h)$, where h is a norm of the partition, and thus the operator I_n can be regarded as a perturbation, or an expansion, of the identity operator as is commonly done for such operators in Approximation

* Work supported in part by National Science Foundation Grants GP3290X and 7610225.

** For an m-th order ordinary differential equation, one uses m+1 adjacent mesh points as stencil points; for a second order elliptic partial differential equation, one uses a nine-point stencil for two independent variables, a 27-point stencil for three independent variables, and so on.

Theory. We have named this method High Order Difference approximtion with Identity Expansions which leads to the pronouncible acronym HODIE.

In this paper the application of the HODIE method to ordinary differential equation problems is treated. The analysis and results presented here give insight into the more complicated--and more important--application of HODIE to the solution of partial differential equations. Preliminary results about the multi-dimensional applications are given by Lynch and Rice [1975,1977a,1977b] and by Lynch [1977a,1977b] and more detailed analyses will be presented at a later time. The method was discovered by R.E. Lynch during a study of methods for approximating solutions of elliptic partial differential equations in two independent variables.

Some of the key features of the method include: (a) the small number of stencil points which leads to a matrix with small bandwidth; (b) the coefficients of the operators M_n, I_n are determined so that the approximation is exact on a linear space of functions and their values are obtained by solving a small local system of linear algebraic equation whose size is fixed independent of the mesh length; (c) high accuracy is obtained by the use of values of f at the auxiliary points rather than with additional stencil points; (d) a variety of boundary conditions can easily be approximated with high accuracy; (e) the method is computationally efficient.

Although the difference equation is similar to one obtained by the Mehrstellenverfahren (or the "Hermitian" method) of Collatz

[1960] after one replaces derivatives of f with divided differences, the method of obtaining the coefficients of the difference equation is different from that of Mehrstellenverfahren.

For ordinary differential equations, the HODIE method gives the same difference equations as obtained by Osborn [1967] who generalized the Størmer-Numerov scheme. Osborn was pessimistic about its practicality; he did not prove convergence results. More recently and independently, Doedel [1976] presented an essentially equivalent method for the ordinary differential equation case and he proved some results. Doedel also presents results about difference schemes which use more than the minmial number, $m+1$, of stencil points for an m-th order ordinary differential operator; we do not consider this case. The results presented below are more complete than those of Doedel for the cases we treat. Both Osborn's and Doedel's approaches lead to different and less efficient implementations than the one described below.

This paper is briefly summarized as follows. In Section 2, a description of the HODIE method for ordinary differential equations is presented and some simple examples are given. In Section 3, a bound on the truncation error is given when the HODIE method is exact on $\mathbf{P}_L$, the space of polynomials of degree at most L. For an m-th order operator, the order of the truncation error is $L-m+1$ and higher order ("super convergence") is obtained with special auxiliary points. In Section 4, approximation for the simple operator $M = d^m/dt^m$ is treated in detail. A direct relationship between the truncation error and quadrature

error is demonstrated and Gauss-type auxiliary points are introduced and analyzed. These Gauss-type points are the zeros of polynomials orthogonal with respect to an integral inner product with weight function a polynomial B-spline. In Section 5, we extend the results of Section 4 to the general linear variable coefficient differential operator with leading term d^m/dt^m. In Section 6, we show that the HODIE method gives a stable difference approximation and that the order of the discretization error is equal to the order of the truncation error. Section 7 contains a comparison of the computational effort for the HODIE method and five other methods; this suggests that the HODIE method is among the most efficient methods available for solving second order boundary value problems. Finally, in Section 8, we give a small sample of extensive experimental results which verify that the HODIE method works as the theory predicts and that there are no unforeseen difficulties in its implementation.

2. Approximation of differential operators. We construct and analyze high accuracy (m+1)-point difference approximation to m-th order differential equations $\mathcal{M}[u,f] = 0$ subject to appropriate initial or two-point boundary conditions $\mathcal{M}^k[u,c_k] = 0$, $k = 0,\ldots,m-1$ where

$$\mathcal{M}[u,f](t) = Mu(t) - f(t), \quad A < t < B, \tag{2-1a}$$

$$Mu(t) = D^m u(t) + \sum_{i=0}^{m-1} a_i(t)\, D^i u(t), \quad D = d/dt, \tag{2-1b}$$

$$\mathcal{M}^k[u,c_k] = M^k u(A) + M^k u(B) - c_k, \quad k = 0,\ldots,m-1, \tag{2-1c}$$

$$M^k u(t) = \sum_{i=0}^{m-1} a_{k,i}(t)\, D^i(t) \tag{2-1d}$$

For the initial value problem, $a_{k,i}(A) = 0$ if $i \neq k$, $a_{k,k}(A) = 1$, and $a_{k,i}(B) = 0$, $i,k = 0,\ldots,m-1$; for the separated two-point boundary value problem, either $M^k u(A)$ or $M^k u(B)$ is zero, $k = 0,\ldots,m-1$.

The interval $A \leq t \leq B$ is partitioned into n subintervals by $n+1$ mesh points t_k: $A = t_0 < t_1 < \ldots < t_n = B$, with $m \leq n$. The approximation U to the solution u is obtained at these mesh points as the solution of a system of $(m+1)$-st order difference equations subject to appropriate boundary conditions.

The approximation $\mathcal{M}_n[u,f] = M_n u - I_n f$ of the differential operator $\mathcal{M}$ is obtained locally by use of a pair of point sets and a set of basis functions. The $m+1$ stencil points are $m+1$ adjacent mesh points: $\bar{t}_k = (t_k, t_{k+1}, \ldots, t_{k+m})$ and we set $h_k = (t_{k+m} - t_k)/m$. The difference operator M_n with coefficients α is

$$M_n U_k = (1/h_k^m) \sum_{i=0}^{m} \alpha_{k,i}\, U_{k+i}, \qquad U_{k+i} \equiv U(t_{k+i}).$$

The second set of points comprise J distinct <u>auxiliary points</u> $\bar{\tau}_k = (\tau_{k,1},\ldots,\tau_{k,J})$ subject to the restrictions $t_k \leq \tau_{k,1} < \ldots < \tau_{k,J} \leq t_{k+m}$. The identity expansion I_n with coefficients β is

$$I_n f_k = \sum_{j=1}^{J} \beta_{k,j}\, f_{k,j}, \qquad f_{k,j} \equiv f(\tau_{k,j}).$$

For a given f, U is the solution of $\mathcal{M}_n[U,f]_k = M_n U_k - I_n f_k = 0$ subject to appropriate boundary conditions.

The coefficients α,β of the operators M_n and I_n are determined so that the approximation is exact on an $(L+1)$-dimensional linear space S of functions. A basis $s_0,s_1,\ldots,s_L$ for S is chosen and the coefficients are made to satisfy the

<u>HODIE equations</u> $\mathcal{M}_n[s_\ell,Ms_\ell]_k = 0$, $\ell = 0,\ldots,L$; that is:

$$(2\text{-}2)\quad (1/h_k^m) \sum_{i=0}^{m} \alpha_{k,i}\, s_\ell(t_{k+i}) - \sum_{j=1}^{J} \beta_{k,j}\, Ms_\ell(\tau_{k,j}) = 0, \quad \ell = 0,\ldots,L.$$

The system (2-2) is homogeneous in the coefficients α,β; hence, in addition to (2-2), we take some convenient normalization equation such as

$$(2\text{-}3)\qquad (a)\ \beta_1 = 1; \qquad (b)\ \textstyle\sum_j |\beta_j| = 1; \qquad (c)\ \sum_j \beta_j = 1.$$

The first is used in actual computation since it simplifies the calculation. The second and third are useful at various places in the theoretical treatment. It is a consequence of the analysis in Sections 4 and 5 that the third normalization can be used.

Remarks about bases and efficient methods of solving the HODIE equations (2-2) are given in Section 7.

Boundary conditions for U are obtained in a similar way. The equation $\mathcal{M}^k[u,c_k] = 0$ is approximated with

$$(1/h_k^m) \sum_{i=0}^{m-1} \{ \alpha_{A,k,i} U_i + \alpha_{B,k,i} U_{n-i} \}$$

$$- \sum_{j=1}^{J} \{ \beta_{A,k,j} f_{A,k,j} + \beta_{B,k,j} f_{B,k,j} \} - c_k = 0$$

where the values $f_{A,k,j}$ and $f_{B,k,j}$ are taken at auxiliary points near $t = A$ and $t = B$, respectively. The coefficients α,β are determined by

$$(1/h_k^m) \sum_{i=0}^{m-1} \{ \alpha_{A,k,i} s_\ell(t_i) + \alpha_{B,k,i} s_\ell(t_{n-i}) \}$$

$$(2\text{-}4) \qquad - \sum_{j=1}^{J} \{ \beta_{A,k,j} Ms_\ell(\tau_{A,k,j}) + \beta_{B,k,j} Ms_\ell(\tau_{B,k,j}) \}$$

$$- M^k s_\ell(A) - M^k s_\ell(B) = 0, \quad \ell = 0,\ldots,L,$$

and here $h_k = (t_n - t_0 + t_n - t_{n-m})/(2m)$ for the case of nonseparated two-point boundary conditions.

The truncation error is defined with respect to a space of functions Σ in terms of the <u>truncation operator</u>

$$T_n[\sigma] = \mathcal{M}_n[\sigma,M\sigma] - \mathcal{M}[\sigma,M\sigma] = M_n\sigma - I_n(M\sigma), \ \sigma \in \Sigma$$

For $\sigma \in \Sigma$, the <u>truncation error</u> is the value of the max-norm of $T_n[\sigma]$, namely, $\|T_n[\sigma]\| = \max_k |T_n[\sigma]_k|$. In Section 3, we obtain a bound on the truncation error for polynomial approximation.

The truncation error is related to the discretization error, defined as the max-norm of the error $e = u - U$ at mesh points. This is because if $u \in \Sigma$, then $M_n e = M_n u - M_n U = M_n u - I_n(Mu)$ $= T_n u$; that is, e satisfies the equation $M_n e = T_n u$. In Section 6 we show that with natrual hypotheses and appropriate boundary condition approximation, a bound on the truncation error yields a similar bound on the discretization error.

Examples. We consider a few examples for equal spaced mesh points with spacing h and the operator $Mu = D^2u + a_1Du + a_0u$. It is sufficient to consider $t_k = -h$, $t_{k+1} = 0$, $t_{k+2} = h$. For brevity, we use a single subscripted notation for the coefficients α,β and the auxiliary points τ. For approximation which is exact on the space $S = \mathbf{P}_L$ of polynomials of degree at most L, we use the Lagrange basis for quadratic interpolation together with elements which are zero at the three stencil points, specifically:

$$s_0(t) = t(t-h)/(2h^2), \quad s_1(t) = (h^2-t^2)/h^2, \quad s_2(t) = t(t+h)/(2h^2),$$

$$s_\ell(t) = t^{\ell-2}(t^2-h^2)/h^\ell, \quad \ell = 3,4,\ldots,L.$$

For normalization, we take the sum of the β's to be equal to unity. After division by appropriate powers of h, the HODIE equations (2-2) and the normalization equation become

$$\ell = 0 \qquad 0 = \alpha_0 - \sum_{j=1}^{J} \beta_j \{ 1 + a_1(\tau_j)[\tau_j - h/2] + a_0(\tau_j)[\tau_j^2 - \tau_j h]/2 \}$$

$$\ell = 1 \qquad 0 = \alpha_1 - \sum_{j=1}^{J} \beta_j \{ -2 + a_1(\tau_j)[-2\tau_j] + a_0(\tau_j)[h^2 - \tau_j^2] \}$$

$$\ell = 2 \qquad 0 = \alpha_2 - \sum_{j=1}^{J} \beta_j \{ 1 + a_1(\tau_j)[\tau_j + h/2] + a_0(\tau_j)[\tau_j^2 + \tau_j h]/2 \}$$

$$\text{normalization} \qquad 1 = \sum_{j=1}^{J} \beta_j$$

$$\ell = 3 \qquad 0 = \sum_{j=1}^{J} \beta_j \{ 6\tau_j + a_1(\tau_j)[3\tau_j^2 - h^2] + a_0(\tau_j)[\tau_j^3 - \tau_j h^2] \}$$

$$\ell = 4 \qquad 0 = \sum_{j=1}^{J} \beta_j \{ 12\tau_j^2 - 2h^2 + a_1(\tau_j)[4\tau_j^3 - 2\tau_j h^2] + a_0(\tau_j)[\tau_j^4 - \tau_j^2 h^2] \}$$

$$\ell = 5 \qquad 0 = \sum_{j=1}^{J} \beta_j \{ 20\tau_j^3 - 6\tau_j h^2 + a_1(\tau_j)[5\tau_j^4 - 3\tau_j h^2] + a_0(\tau_j)[\tau_j^5 - \tau_j^3 h^2] \}$$

and so on.

Note that the first three equations give the α's in terms of the β's. Also note that in all the equations above, the terms which involve the coefficients a_1 and a_0 are order h and h^2, respectively, compared with the leading term in each of the curly brackets.

For specific examples, we consider $M = D^2$ in which $a_1 = a_0 = 0$. One obtains immediately that $\alpha_0 = \alpha_2 = 1$, $\alpha_1 = -2$, so that the difference operator M_n is the usual divided difference approximation for the second derivative operator:

$$M_n U_k = (U_k - 2U_{k+1} + U_{k+2})/h^2 = U[t_k, t_k+h, t_k+2h]$$

However, the operator I_n changes when J or the locations of

the auxiliary points change. Below $O(h^p)$ denotes the truncation error with respect to the space of functions $\Sigma = C^{(p+2)}$ of functions with continuous $(p+2)$-nd derivative.

Example 2-1: For $J = 1$ and $-h = t_k \leq \tau_1 \leq t_{k+2} = h$, $\tau_1 \neq 0$, the equation for $\ell = 3$ is not satisfied [$a_1 = a_0 = 0$] and with $I_n f_k = f(\tau_1)$ we obtain an $O(h)$ scheme which is exact on $\mathbb{P}_2$.

Example 2-2: For $J = 1$ and $\tau_1 = t_{k+1} = 0$, the equation for $\ell = 3$ is satisfied, but the one for $\ell = 4$ is not satisfied, and with $I_n f_k = f(\tau_1)$ we obtain an $O(h^2)$ scheme which is exact on $\mathbb{P}_3$.

Example 2-3: For $J = 2$ and $-\tau_1 = \tau_2 = h(1/6)^{1/2}$ we obtain an $O(h^4)$ scheme which is exact on $\mathbb{P}_5$ with $I_n f_k = [f(\tau_1) + f(\tau_2)]/2$.

Example 2-4: $J = 3$, exact on $\mathbb{P}_5$, $O(h^4)$ Størmer-Numerov approximation: $I_n f_k = [f(-h) + 10f(0) + f(h)]/12$.

Example 2-5: $J = 3$, exact on $\mathbb{P}_7$, $O(h^6)$ approximation of Osborne [1967]: $I_n f_k = [5f(\tau_1) + 14f(0) + 5f(\tau_3)]/24$, $-\tau_1 = \tau_3 = h(2/5)^{1/2}$.

Example 2-6: $J = 5$, exact on $\mathbb{P}_{11}$, a new $O(h^{10})$ approximation:

$$I_n f_k = \beta_1 f(\tau_1) + \ldots + \beta_5 f(\tau_5),$$

$$\beta_1 = \beta_5 = 0.0516582578, \quad \beta_2 = \beta_4 = 0.2394732407, \quad \beta_3 = 0.4177370031$$

$$-\tau_1 = \tau_5 = 0.8214405997h, \quad -\tau_2 = \tau_4 = 0.4499203525h, \quad \tau_3 = 0.$$

Example 2-7: $J = 3$, exact on the space of cubic splines with joints at the equal spaced mesh points (see, for example, Birkhoff and de Boor [1965], page 189), $O(h^2)$ approximation:

$$I_n f_k = [f(-h) + 4f(0) + f(h)]/6.$$

To use these schemes for the Dirichlet problem, one solves the system

(2-5a) $$U_0 = u(A), \quad U_n(B) = u(B)$$

(2-5b) $$(U_{k-1} - 2U_k + U_{k+1})/h^2 = g_k, \quad k = 1,\ldots,n-1$$

with $h = (B-A)/n$, $t_k = A + kh$ and

$$g_k = I_n f_{k-1} = \sum_{j=1}^{J} \beta_j \, f(A+kh+\tau_j)$$

where g differs from example to example. In each case, however, the matrix formulation has the same tridiagonal $(n-1)$-by-$(n-1)$ coefficient matrix. Once g has been evaluated, the work to solve the system is independent of the particular g used. Thus, asymptotically, the higher the order of the scheme, the smaller is the work required to achieve a given accuracy. This is verified for more complicated differential equations by both operation counts (Section 7) and experimental results (Section 8).

For the initial value problem, one solves (2-4) to get the coefficients of the initial conditions for the finite difference equation (2-5b). Use of polynomial spaces S and simplification leads to

(2-5a') $$U_0 = u(A), \quad U_1 = u(A) + h\,Du(A) + h^2\,g^0/2,$$

and the following gives the value of g^0 for accuracy comparable to that for the schemes given above:

Example 2-1': $O(h)$, $g^0 = f(A)$

Example 2-2': $O(h^2)$, $g^0 = f(A+h/3)$

Example 2-4': $O(h^4)$, $g^0 = [9f(A) + 25f(A+2h/5) + 2f(A+h)]/36.$

Example 2-5': $O(h^6)$, $g^0 = \beta_1 f(A+\tau_1) + \beta_2 f(A+\tau_2) + \beta_3 f(A+\tau_3)$,

$\beta_1 = 0.4018638275$, $\beta_2 = 0.4584822127$, $\beta_3 = 0.1396539598$,

$\tau_1 = 0.0885879595h$, $\tau_2 = 0.4094668644h$, $\tau_3 = 0.7876594618h$.

Example 2-6': $O(h^{10})$, $g^0 = \beta_1 f(A+\tau_1) + \ldots + \beta_5 f(A+\tau_5)$,

$\beta_1 = 0.1935631805$, $\beta_2 = 0.3343492762$, $\beta_3 = 0.2927739742$,

$\beta_4 = 0.1478177401$, $\beta_5 = 0.0314958290$,

$\tau_1 = 0.0398098571h$, $\tau_2 = 0.1980134179h$, $\tau_3 = 0.4379748102h$,

$\tau_4 = 0.6954642734h$, $\tau_5 = 0.9014649142h$.

3. Truncation error for polynomial approximation. We only consider approximation away from boundaries and approximation which is exact on a polynomial space $\mathbb{P}_L$ for some $L \geq m$. Results for approximation of boundary conditions are obtained by an easy modification. Results for other spaces, such as those appropriate for approximation near singular points of differential equations, will be presented elsewhere.

We use $\xi_{k,j}$, $j = 0,1,\ldots,$ to denote distinct points such that $t_k \leq \xi_{k,j} \leq t_{k+m}$ and set $\bar{\xi}_{k,j} = (\xi_{k,0},\ldots,\xi_{k,j})$; we also set

$$\Delta\bar{\xi}_{k,j} = \min_{i,q=0,\ldots,j,\ i\neq q} |\xi_{k,i} - \xi_{k,q}|. \tag{3-1}$$

We use the polynomials

$$w(\bar{\xi}_{k,j};t) = \Pi_{q=0}^{j} (t - \xi_{k,q})/(j+1)!, \quad j = 0,1,\ldots, \tag{3-2a}$$

and, to simplify notation below, we set

$$w(\bar{\xi}_{k,-1};t) = 1. \tag{3-2b}$$

We also use the Lagrange polynomial interpolation basis with respect to the points in $\bar{\xi}_{k,j}$:

$$\ell_r(\bar{\xi}_{k,j};t) = w(\bar{\xi}_{k,j};t)/[(t - \xi_{k,r})\, w'(\bar{\xi}_{k,j};\xi_{k,r})], \quad r = 0,\ldots,j. \tag{3-3}$$

For fixed $\bar{\xi}_{k,j}$, there is a constant K which does not depend on h_k or $\Delta\bar{\xi}_{k,j}$ such that for all t, $t_k \leq t \leq t_{k+m}$:

$$\begin{aligned} |D^i w(\bar{\xi}_{k,j};t)| &\leq K\, h_k^{j-i+1}, \quad i = 0,\ldots,j+1, \\ |D^i \ell_r(\bar{\xi}_{k,j};t)| &\leq K\, h_k^{j-i}/\Delta\bar{\xi}_{k,j}^{\,j}, \quad i,r = 0,\ldots,j. \end{aligned} \tag{3-4}$$

Because $T_n u_k = M_n u_k - I_n[Mu]_k$ involves derivatives of u only up to order $m \leq L$, it follows (see, for example, Theorem 2.1 of de Boor and Lynch [1966]) that for

$$u \in \mathbb{F}^{L+1}[t_k,t_{k+m}] = \{\, v \mid D^L v \text{ is absolutely continuous,}$$
$$D^{L+1} v \text{ is square integrable on } t_k \leq t \leq t_{k+m} \,\} \tag{3-5}$$

we have

$$T_n u_k = \sum_{i=0}^{L} T_n[\ell_i(\bar{\xi}_{k,L};t)]_k \, u(\xi_{k,i}) + \int_{t_k}^{t_{k+m}} T_{n(t)}[q(\bar{\xi}_{k,L};t,x)]_k \, D^{L+1}u(x)\, dx \tag{3-6a}$$

where

$$q(\bar{\xi}_{k,L};t,x) = \left[\, (t-x)_+^L - \sum_{i=0}^{L} \ell_i(\bar{\xi}_{k,L};t)(\xi_{k,i}-x)_+^L \,\right]/L! \tag{3-6b}$$

$$(t-x)_+^L = \begin{cases} (t-x)^L & \text{for } t-x > 0 \\ 0 & \text{for } t-x \leq 0 \end{cases} \tag{3-6c}$$

and where the subscript (t), as in $T_{n(t)}$, denotes that the operator is applied to a function of t.

Suppose that the HODIE approximation is exact on $\mathbb{P}_L$ so that the coefficients α,β satisfy (2-2) with polynomial basis elements. Then, because $\ell_i(\bar{\xi}_{k,L};\cdot) \in \mathbb{P}_L$, the sum in (3-6a) is equal to zero. The sum in the definition (3-6b) of q is that element of $\mathbb{P}_L$ which interpolates to $(t-x)_+^L$ at the points $t = \xi_{k,j}$, $j = 0,\ldots,L$. By taking $m+1$ of the points in $\bar{\xi}_{k,L}$ to be the stencil points $\bar{t}_k$,

one has $q = 0$ on the stencil points and hence $M_{n(t)}q = 0$; (3-6a) then reduces to

$$(3\text{-}7a)\qquad T_n u_k = -\sum_{j=1}^{J} \beta_{k,j} \int_{t_k}^{t_{k+m}} M_{(t)}q(\bar{\xi}_{k,L};t,x)\Big|_{t=\tau_{k,j}} D^{L+1}u(x)\,dx,$$

where

$$(3\text{-}7b)\qquad M_{(t)}q(\bar{\xi}_{k,L};t,x) = \sum_{i=0}^{m} a_i(t)\,\{(t-x)_+^{L-i}/(L-i)! - \sum_{j=0}^{L} (\xi_{k,L}-x)_+^{L} D^{i}\ell_j(\bar{\xi}_{k,L};t)/L!\}$$

In (3-7a), points $\tau_{k,j}$, x, and those in $\bar{\xi}_{k,L}$ are between t_k and t_{k+m}. Therefore, by (3-4) we can bound the quantity in curly brackets in (3-7b) by $h_k^{L-m}(K_1 + K_2[h_k/\Delta\bar{\xi}_{k,L}]^L)$ where K_1, K_2 are constants which do not depend on h_k or $\Delta\bar{\xi}_{k,L}$. Consequently, if $D^{L+1}u$ and the coefficients a_i are continuous, then

$$|T_n u_k| \le K_3(1 + [h_k/\Delta\bar{\xi}_{k,L}]^L)\{\sum_{j=1}^{J} |\beta_{k,j}|\}\,\|D^{L+1}u\|_\infty\, h_k^{L-m+1}$$

where $\|\cdot\|_\infty$ denotes the max-norm, K_3 depends on $\max_i \|a_i\|_\infty$ but not on h_k or $\Delta\bar{\xi}_{k,L}$.

We have introduced the restriction that $m+1$ of the points in $\bar{\xi}_{k,L}$ are the stencil points in $\bar{t}_k$; the other $L-m$ points in $\bar{\xi}_{k,L}$ are arbitrary and we can choose them to maximize $\Delta\bar{\xi}_{k,L}$. Clearly, this maximum depends only on the stencil points and L. For $L \ge m$, set

$$(3\text{-}8)\qquad R_L(\bar{t}_k) = h_k/\max\,\Delta\bar{\xi}_{k,L}$$ where the maximization is over all points $\xi_{k,L}$ such that $t_k \le \xi_{k,\ell} \le t_{k+m}$, $\ell = 0,\ldots,L$ and $m+1$ of the points $\xi_{k,\ell}$ are equal to t_{k+j}, $j = 0,\ldots,m$.

Furthermore, set

$$H_n = \max_{j=0,\ldots,n-m} (t_{j+m} - t_j)/m \tag{3-9}$$

and we have the following.

THEOREM 3-1: *Suppose the coefficients* a_i *of* M *are continuous. Let* $A = t_0 < t_1 < \ldots < t_n = B$, $n \geq m$, *be a set of mesh points and* $\bar{\tau}_k$, $k = 0,\ldots,n-m$, *sets of auxiliary points. Suppose that for* $k = 0,\ldots,n-m$ *there are coefficients* $\alpha_{k,i}$, $\beta_{k,j}$ *which satisfy* (2-2) *and* (2-3b) *for* $s_0,\ldots,s_L$, $L \geq m$, *a basis for* $\mathbb{P}_L$. *Then there is a constant* K *which depends only on* B-A, *the order* m *of* M, *and the coefficients* a_i *such that for any* u *with continuous* (L+1)-*st derivative*

$$|T_n u_k| \leq K\,[1 + \max_{j=0,\ldots,n-m} R_L(\bar{t}_j)^L]\; \|D^{L+1}u\|_\infty\; H_n^{L-m+1},\quad k = 0,\ldots,n-m.$$

4. Analysis of the special case $M = D^m$. The main results about the special case $M = D^m$ carry over to the general case of the variable coefficient operator M in (2-1b). In this section, we consider in detail the special case. To distinguish between the two cases, we use the superscript 0 for quantities which apply to the special case, in particular, we use α^0, β^0, M_n^0, and I_n^0 for the coefficients and the operators when $M = D^m$.

In (2-2) set $M = D^m$, replace α,β with α^0, β^0, and use the following basis for $\mathbf{P}_L$ [see (3-2) and (3-3)]:

$$s_i(t) = \begin{cases} \ell_i(\bar{t}_k;t), & i = 0,\dots,m, \\ w(\bar{\xi}_{k,i-1};t), & i = m+1,\dots,L \end{cases} \tag{4-1a}$$

where

$$\bar{\xi}_{k,i-1} = (\xi_{k,0},\dots,\xi_{k,i-1}), \text{ the points } \xi_{k,\ell} \text{ are distinct,} \tag{4-1b}$$

$$t_k \le \xi_{k,\ell} \le t_{k+m}, \text{ and } \xi_{k,j} = t_{k+j},\ j = 0,\dots,m.$$

In this section we use the normalization (2-3c) and it is a consequence of the analysis below that this is allowed, i.e. $\sum_j \beta_j \neq 0$. Note that $D^m w(\bar{\xi}_{k,m-1};t) = D^m(t-t_k)\dots(t-t_{k+m-1})/m! = 1$ so that (2-3c) can be written as

$$\sum_{j=1}^{J} \beta_{k,j}\, D^m w(\bar{\xi}_{k,m-1};\tau_{k,j}) = 1.$$

Since the Lagrange basis element $\ell_i(\bar{t}_k;\cdot)$ is in $\mathbf{P}_m$, its m-th derivative is a constant.

The HODIE equations for the special case then become

$$\text{(4-2a)} \qquad \alpha^0_{k,i}/h_k^m - [m!/w'(\bar{t}_k;t_{k+i})] \sum_{j=1}^{J} \beta_{k,j} = 0, \quad i = 0,\ldots,m,$$

$$\text{(4-2b)} \qquad \sum_{j=1}^{J} \beta^0_{k,j} D^m w(\bar{\xi}_{k,m+\ell-2};\tau_{k,j}) = \delta_{\ell,1}, \ \ell = 1,\ldots,L-m+1,$$

where $\delta_{\ell,i}$ denotes the Kronecker delta function.

Since the sum of the β^0's is unity, (4-2a) shows that the operator M_n^0 is $m!$ times the usual divided difference approximation to $M = D^m$:

$$\text{(4-3)} \qquad \begin{aligned} M_n^0 u_k &= \sum_{i=0}^{m} \alpha^0_{k,i} \, u(t_{k+i})/h_k^m = m! \sum_{i=0}^{m} u(t_{k+i})/w'(\bar{t}_k;t_{k+i}) \\ &= m! \, u[t_k,t_{k+1},\ldots,t_{k+m}], \end{aligned}$$

that is, $M_n^0 u_k$ is the m-th derivative of the unique polynomial in $\mathbb{P}_m$ which interpolates to the values $u(t_{k+i})$ at t_{k+i}, $i = 0,\ldots,m$.

By Taylor's Theorem, any u in $\mathbb{F}^m$ can be represented as

$$u(t) = \sum_{i=0}^{m-1} D^i u(t_k) \, (t-t_k)^i/i! + \int_{t_k}^{t} (t-x)^{m-1} D^m u(x) \, dx/(m-1)!$$

Because the m-th divided difference of an element of $\mathbb{P}_{m-1}$ is zero, we have

$$\text{(4-4)} \qquad M_n^0 u_k = m! \, u[t_k,\ldots,t_{k+m}] = \int_{t_k}^{t_{k+m}} B_m(\bar{t}_k;x) \, D^m u(x) \, dx,$$

where $B_m(\bar{t}_k;x)$ is the m-th divided difference $g_m[t_k,\ldots,t_{k+m};x]$ with respect to t of

$$g_m(t;x) = (t-x)_+^{m-1} = \begin{cases} (t-x)^{m-1}/(m-1)! & \text{if } t \geq x \\ 0 & \text{if } t < x \end{cases}$$

so that $B_m(\bar{t}_k;\cdot)$ is the $(m-1)$-st degree polynomial B-spline with joints

at the stencil points in $\bar{t}_k$. This B-spline satisfies (Curry and Schoenberg [1966])

$$(4\text{-}5a) \qquad B_m(\bar{t}_k;x) = \begin{cases} > 0, & t_k < x < t_{k+m}, \\ = 0, & x \leq t_k \ \text{ or } \ t_{k+m} \leq x, \end{cases}$$

$$(4\text{-}5b) \qquad \int_{t_k}^{t_{k+m}} B_m(\bar{t}_k;x)\, dx = 1.$$

Therefore, we have

$$\begin{aligned} T_n^0 u_k &= M_n^0 u_k - I_n^0[D^m u]_k \\ &= \int_{t_k}^{t_{k+m}} B_m(\bar{t}_k;x)\, D^m u(x)\, dx - \sum_{j=1}^{J} \beta_{k,j}^0\, D^m u(\tau_{k,j}) \\ &\equiv E_n^0[D^m u]_k \end{aligned}$$

and in this we have defined the operator E_n^0. Clearly $E_n^0 v_k$ is the quadrature error in using $I_n^0 v_k$ as an approximation to the integral of $B_m(\bar{t}_k;x)v(x)$. This quadrature error is zero for any v in $\mathbf{P}_{J-1}$ if and only if

$$(4\text{-}6) \qquad \beta_{k,j}^0 = \int_{t_k}^{t_{k+m}} B_m(\bar{t}_k;x)\, \ell_{j-1}(\bar{\tau}_k;x)\, dx, \quad j = 1,\ldots,J$$

In particular, it is zero for $v(t) = 1$ and thus by (4-5b) the sum of the β^0's is unity, i.e., the normalization (2-3c) is satisfied. But then, for any u in $\mathbf{P}_{m+J-1}$, $T_n^0 u_k = E_n^0[D^m u]_k = 0$.

Consequently, for any stencil points $\bar{t}_k$ and any J auxiliary points $\bar{\tau}_k$, there is a unique HODIE scheme with normalization (2-3c) which is exact on $\mathbf{P}_{m+J-1}$. One obtains a family of HODIE schemes which are exact on $\mathbf{P}_L$ for any L such that $0 \leq L < m+J-1$. We

now show that there exist special sets of auxiliary points which make the approximation exact on $\mathbb{P}_L$ for L up to $m+2J-1$.

Since $B_m(\bar{t}_k;\cdot)$ is positive on the range of integration, we can define the following inner product:

$$(u,v) = \int_{t_k}^{t_{k+m}} B_m(\bar{t}_k;x)\, u(x)\, v(x)\, dx.$$

For fixed m, k, and $B_m(\bar{t}_k;\cdot)$, let $b_0, b_1, \ldots$ with, b_i in $\mathbb{P}_i$ denote the normalized orthogonal polynomials with respect to this inner product; we call these the B-spline orthogonal polynomials. Based on the well-known theory of orthogonal polynomials, b_i has i distinct real zeros in $t_k < t < t_{k+m}$, and, for fixed i, we call these the B-spline Gauss points.

When the J auxiliary points in $\bar{\tau}_k$ are the B-spline Gauss points for b_J, then the unique HODIE approximation which is exact on $\mathbb{P}_{J+m-1}$ is also exact on $\mathbb{P}_{2J+m-1}$. In this case, the β^0's are the coefficients of the J-point Gauss quadrature formula with weight functions $B_m(\bar{t}_k;\cdot)$ and each $\beta^0_{k,j}$, $j = 1,\ldots,J$, is positive. Since (b_J,b_J) is positive and $I^0_n[b_J^2]_k = 0$, this HODIE approximation is not exact on $\mathbb{P}_{2J+m}$.

The B-spline Gauss points and the quadrature coefficients have been tabulated by Phillips and Hanson [1974] for a number of degrees and for a normalized interval and equal spaced joints.

The preceding results are summarized below.

THEOREM 4-1: Let $M = D^m$ and let the normalization for HODIE approximation be (2-3c). For any set of m+1 stencil and $J > 0$ auxiliary points $\bar{t}_k, \bar{\tau}_k$, there is a HODIE approximation with coefficients $\alpha_{k,j} = \alpha^0_{k,j}$, $\beta_{k,j} = \beta^0_{k,j}$ which is exact on $\mathbb{P}_L$ for any L with $0 \leq L-m \leq J-1$. The operator $M_n = M^0_n$ is unique, it is m! times the divided difference operator with respect to the stencil points. There are sets of J auxiliary points for which a HODIE approximation is exact for L with $J \leq L-m \leq 2J-1$. If $L-m \geq J-1$, then the coefficients β^0 of I^0_n are unique and are given by (4-6). The J auxiliary points which give exactness on $\mathbb{P}_{2J+m-1}$ are the zeros of the J-th degree B-spline orthogonal polynomial b_J associated with the B-spline $B_m(\bar{t}_k;\cdot)$ with joints at the stencil points.

The examples for $M = D^2$ in Section 2 illustrate various special cases of the resutls stated in Theorem 4-1. Examples 2-2, 2-3, 2-5, and 2-6 use J B-spline Gauss points for $J = 1, 2, 3,$ and 5, respectively.

Example 2-4 (Størmer-Numerov) and 2-7 (exact on cubic splines) both use the same set of three auxiliary points. Both are exact on $\mathbb{P}_3$ and for this $L-m = 3-2 = 1 < J-1 = 2$; their different sets of β's illustrate the nonuniqueness for $L-m < J-1$. Since the Størmer-Numerov scheme is also exact on $\mathbb{P}_4$ and since $L-m = J-1$ for this case, the scheme is the unique HODIE scheme with those three auxiliary points which is exact on $\mathbb{P}_4$. One of the auxiliary points, $\tau_{k,2} = t_{k+1}$ is a B-spline Gauss point for all odd degree B-spline orthogonal polynomials associated with $B_{2,k}$ with equal spaced joints;

because of this (or, alternatively, symmetry), the scheme is exact on $\mathbf{P}_5$. Another set of three auxiliary points (Example 2-5) yields an approximation exact on $\mathbf{P}_7$.

We now derive bounds on the elements of the inverse of the coefficient matrix of the system in (4-2b) with $L-m+1 = J$; these are used in the next sectio

For $i = 1,\ldots,J$ consider the systems

$$\sum_{j=1}^{J} x_{j,i}\, D^m w(\bar{\xi}_{k,m+\ell-2};\tau_{k,j}) = \delta_{\ell,i}, \quad \ell = 1,\ldots,J.$$

Multiply the ℓ-th equation by the constant (determined below) $\pi_{r-1,m+\ell-2}$ and sum with respect to ℓ to obtain

$$(4\text{-}7) \qquad \sum_{j=1}^{J} x_{j,i}\, D^m \sum_{\ell=1}^{J} \pi_{r-1,m+\ell-2}\, w(\bar{\xi}_{k,m+\ell-2};\tau_{k,j}) = \pi_{r-1,m+i-2}.$$

Define the polynomial

$$p_{r-1}(t) = \sum_{\ell=1}^{J} \pi_{r-1,m+\ell-2}\, w(\bar{\xi}_{k,m+\ell-2};t)$$

$$= \sum_{\ell=1}^{J} \pi_{r-1,m+\ell-2}\, (t-\xi_{k,0})\ldots(t-\xi_{k,m+\ell-2})/(m+\ell-1)!.$$

Then the π's can be expressed in terms of divided differences of p_{r-1}:

$$\pi_{r-1,m+\ell-2} = (m+\ell-1)!\; p_{r-1}[\xi_{k,0},\ldots,\xi_{k,m+\ell-1}].$$

Choose these constants so that

$$D^m p_{r-1}(t) = \ell_{r-1}(\bar{\tau}_k;t)$$

where ℓ_{r-1} is a Lagrange basis (3-3).

Then the solution of (4-7) is given by

$$x_{r,i} = \pi_{r-1,m+i-2} = (m+i-1)!\, p_{r-1}[\xi_{k,0},\ldots,\xi_{k,m+i-1}].$$

The points $\xi_{k,\ell}$ are distinct, are between t_k and t_{k+m} and $\xi_{k,\ell} = t_{k+\ell}$ $\ell = 0,\ldots,m$. Hence it follows from (4-4) that

$$x_{r,i} = \int_{t_k}^{t_{k+m}} B_{m+i-1}(\bar{\xi}_{k,m+i-1};x)\, D^{m+i-1} p_{r-1}(x)\, dx$$

$$= \int_{t_k}^{t_{k+m}} B_{m+i-1}(\bar{\xi}_{k,m+i-1};x)\, D^{i-1} \ell_{r-1}(\bar{\tau}_k;x)\, dx,$$

where $B_{m+i-1}(\bar{\xi}_{m+i-1};\cdot)$ denotes the polynomial B-spline of degree $m+i-2$ with joints at $\xi_{k,\ell}$, $\ell = 0,\ldots,m+i-1$. For the case $i = 1$, this reduces to $x_{j,1} = \beta_{k,j}$ with $\beta_{k,j}$ given in (4-6). By (4-5) and (3-4), we have, therefore, the following result.

LEMMA 4-1: Let $\bar{\xi}_k = (\xi_{k,0},\ldots,\xi_{k,m+J-2})$ where the points $\xi_{k,\ell}$ are distinct and between t_k and t_{k+m} and $\xi_{k,\ell} = t_{k+\ell}$, $\ell = 0,\ldots,m$. Let B^0 denote the matrix with elements

$$(B^0)_{\ell,i} = D^m w(\xi_{k,m+\ell-2};\tau_{k,j}), \quad \ell,i = 1,\ldots,J$$

where w is as in (3-2a) and $\bar{\tau}_k = (\tau_{k,1},\ldots,\tau_{k,J})$ is a set of auxiliary points between t_k and t_{k+m}. There exist constants $K_{\ell,i}$ which are independent of h_k and $\Delta\bar{\tau}_k$ [see (3-1)] such that

$$|([B^0]^{-1})_{\ell,i}| \le K_{\ell,i}\, (h_k/\Delta\bar{\tau}_k)^{J-1}/h_k^{i-1}$$

5. Analysis of the variable coefficient case. Let λ and ψ denote the functions obtained by applying M to the basis element ℓ and w in (4-1a):

$$\text{(5-1a)} \qquad \lambda_i(t) = Ms_i(t) = M\ell_i(\bar{t}_k;t), \qquad i = 0,\dots,m,$$

$$\text{(5-1b)} \qquad \psi_\ell(t) = Ms_{m+\ell}(t) = Mw(\bar{\xi}_{k,m+\ell-1};t), \qquad \ell = 1,\dots,L-m,$$

and set

$$\text{(5-1c)} \qquad \psi_0(t) = Mw(\bar{\xi}_{k,m-1};t) \equiv 1.$$

We use λ_i^0, ψ_ℓ^0 to denote these functions in the special case $M = D^m$.

The HODIE equations are then

$$\text{(5-2a)} \qquad \alpha_{k,i}/h_k^m - \sum_{j=1}^{J} \beta_{k,j}\, \lambda_i(\tau_{k,j}) = 0, \qquad i = 0,\dots,m,$$

$$\text{(5-2b)} \qquad \sum_{j=1}^{J} \beta_{k,j}\, \psi_{\ell-1}(\tau_{k,j}) = \delta_{\ell,1}, \qquad = 1,\dots,L-m+1.$$

To see that λ,ψ differ from λ^0,ψ^0 by $O(h_k)$, express the variables in terms of nondimensional parameters γ, $\gamma_{k,j}$, $\rho_{k,j}$:

$$t = t_k+\gamma h_k; \quad \xi_{k,i} = t_k+\gamma_{k,i}h_k,\ i = 0,\dots,L; \quad \tau_{k,j} = t_k+\rho_{k,j}h_k,\ j = 1,\dots,J,$$

$$\bar{\gamma}_{k,\ell} = (\gamma_{k,0},\dots,\gamma_{k,\ell}), \qquad \bar{\rho}_k = (\rho_{k,1},\dots,\rho_{k,J})$$

and then $\gamma_{k,j}$ and $\rho_{k,j}$ are between 0 and m. From (3-2a) we have

$$w(\bar{\xi}_{k,\ell};t) = h_k^{\ell+1} w(\bar{\gamma}_{k,\ell};\gamma), \qquad D^r w(\bar{\xi}_{k,\ell};t) = h_k^{\ell+1-r} D^r w(\bar{\gamma}_{k,\ell};\gamma)$$

Since

$$\lambda_i^0(\tau_{k,j}) = m!/w'(\bar{t}_k;t_{k+i}) = m!/[h_k^m\, w'(\bar{\gamma}_{k,m};\gamma_{k,i})],$$

for $i = 0,\ldots,m$ we have

$$\lambda_i(\tau_{k,j}) = \lambda_i^0(\tau_{k,j})[\ 1$$

(5-3a)

$$+ \{h_k a_{m-1}(\tau_{k,j})\textstyle\sum_{q=0,q\neq i}^{m}(\rho_{k,j}-\gamma_{k,q}) + \ldots + h_k^m a_0(\tau_{k,j})\prod_{q=0,q\neq i}^{m}\gamma_{k,q}\}/m!\ ]$$

For $\ell = 1,\ldots,L-m+1$ we have

$$\psi_\ell(\tau_{k,j}) = \psi^0(\tau_{k,j}) + \textstyle\sum_{p=0}^{m-1} a_p(\tau_{k,j})\ D^p w(\bar{\xi}_{k,m+\ell-1};\tau_{k,j})$$

(5-3b)

$$= h_k^\ell[D^m w(\bar{\gamma}_{k,m+\ell-1};\rho_{k,j}) + \textstyle\sum_{p=0}^{m-1} h_k^p\ a_p(\tau_{k,j})D^p w(\bar{\gamma}_{k,m+\ell-1};\rho_{k,j})].$$

The following establishes existence and uniqueness of HODIE schemes for $L-m = J-1$.

THEOREM 5-1: Let the normalization of the HODIE approximation be (2-3c). Suppose that the coefficients a_i of M are continuous. There is a positive H such that if the stencil points $\bar{t}_k$ satisfy $0 < h_k = (t_{k+m}-t_k)/m \leq H$, then for any set of J auxiliary points $\bar{\tau}_k$, there is a unique HODIE approximation which is exact on $\mathbb{P}_{J+m-1}$. Its coefficients α,β are the solution of (5-2) with $L-m = J-1$.

Proof: By hypothesis, the coefficients a_i are continuous, hence so are the functions λ_i, ψ_ℓ in (5-1). The values of these functions differ by $O(h_k)$ at the auxiliary points from λ_i^0, ψ_ℓ^0 for the special case $M = D^m$. Because of the uniqueness of the coefficients $\alpha_{k,i}^0$, $\beta_{k,j}^0$, there is positive H so that the coefficient matrix of the system in (5-2) is nonsingular for $h_k < H$. ■

To show that HODIE approximations exist for $L-m = 2J-1$ with special auxiliary points, we need some preliminary results. After changing to nondimensional parameters, the functions ψ_ℓ in (5-1) have the same form as the functions in the next theorem. This theorem shows that the set of functions ψ_ℓ, $\ell = 0,\ldots,L-m$ is a Chebyshev set.

THEOREM 5-2: Let K and m denote positive integers. Let γ_k, $k = 0,\ldots,K+m-1$ denote distinct points in the unit interval. Let the functions Φ_ℓ have the form

$$\Phi_0(h;\gamma) \equiv 1, \quad \Phi_\ell(h;\gamma) = D^m \Pi_{k=0}^{m+\ell} (\gamma-\gamma_k) + \phi_\ell(h;\gamma), \ \ell = 1,\ldots,K-1,$$

where ϕ_ℓ is continuous and $O(h)$ on $0 \leq \gamma \leq 1$. Let $\bar{\rho} = (\rho_1,\ldots,\rho_K)$ have distinct components such that $0 \leq \rho_k \leq 1$. There is a positive H such that for any h, $0 \leq h \leq H$,

$$\sum_{\ell=0}^{K-1} c_\ell \Phi_\ell(h;\rho_k) = 0, \ k = 1,\ldots,K, \ \text{implies} \ c_\ell = 0, \ \ell = 0,\ldots,K-1.$$

The result in Theorem 5-2 is easy to prove for fixed $\bar{\rho}$. But, in addition, we must show that H can be chosen independent of $\bar{\rho}$.

Proof: First, let $\bar{\rho}$ be fixed and consider the K-by-K matrix $V(h)$ with elements $V(h)_{k,\ell} = \Phi_{\ell+1}(h;\rho_k)$. The product of $V(0)$ and a diagonal matrix is equal to a Vandermonde matrix; therefore, $V(0)$ is nonsingular. By continuity of the elements of $V(h)$, there is an $H(\bar{\rho})$ such that $V(h)$ is nonsingular for all h, $0 \leq h \leq H(\bar{\rho})$.

Second, suppose that there is no positive H_0 independent of

$\bar{\rho}$ such that $V(h)$ is nonsingular for all h, $0 \leq h \leq H_0$. Then there are sequences with index $i = 1,2,\ldots, \to \infty$,

$$H_i \to 0, \quad c_\ell(H_i),\ \ell = 0,\ldots,K\text{-}1, \text{ with } \max_\ell |c_\ell(H_i)| = 1,$$

$$\bar{\rho}(H_i) = (\rho_1(H_i),\ldots,\rho_K(H_i)), \quad P_i(\rho) = \sum_{\ell=0}^{K-1} c_\ell(H_i)\ \Phi_\ell(H_i;\rho),$$

where P_i has zeros at $\rho = \rho_j(H_i)$, $j = 1,\ldots,K$. There exist, therefore, convergent subsequences (whose elements we also denote as above) such that

$$c_\ell(H_i) \to c_\ell^*, \quad \rho_j(H_i) \to \rho_j^*, \text{ and } P_i \to P^*.$$

By continuity and the form of the functions Φ_ℓ, the limiting function P^* is a polynomial of degree at most K-1. Again by continuity, $P^*(\rho_j^*) = 0$, $j = 1,\ldots,K$. Since $\max_\ell |c^*| = 1$, P^* is not identically equal to zero; consequently, the K points ρ_j^* are not distinct. Suppose that there are exactly $N > 1$ zeros of P^* which are equal to ρ_k^* and $\rho_k^* = \rho_{k+1}^* = \ldots = \rho_{k+N-1}^*$. Then we can write

$$P_i(\rho) = p(H_i;\rho)\ \prod_{q=0}^{N-1} [\rho - \rho_{k+q}(H_i)] \to p(0;\rho)\ [\rho - \rho_k^*]^N,$$

which shows that the $(K\text{-}1)$-st degree polynomial P^* has K zeros counting multiplicities. This contradiction establishes the theorem. ∎

The application of Theorem 5-2 to HODIE approximation follows from representations of moments of a Chebyshev set. Such moments are discussed in detail in Chapter 2 of Karlin and Studden [1966]; see, especially, pages 38-46. We summarize the pertinent information in the next paragraph.

Let μ denote any nondecreasing right continuous function of bounded variation on $t_k \leq t \leq t_{k+m}$. Let ψ_ℓ, $\ell = 0,\ldots,L$ denote functions of a Chebyshev set on this interval. The ℓ-th moment q_ℓ of the set with respect to the measure $d\mu$ is

$$q_\ell = \int_{t_k}^{t_{k+m}} \psi_\ell(x)\, d\mu(x), \quad \ell = 0,\ldots,L.$$

For each measure, one gets a set of moments $\bar{q} = (q_0,\ldots,q_L)$ and the set of all such $\bar{q}$ is a subset Q of Euclidian $(L+1)$-space which is called the moment space of the Chebyshev set. This moment space is the smallest cone with vertex at the origin which contains the curve $\Psi(t) = (\psi_0(t),\ldots,\psi_L(t))$, $t_k \leq t \leq t_{k+m}$; this curve is not in Euclidean L-space. If $L = 2J-1$, $J \geq 1$, and $\bar{q} \in Q$ is an interior point of Q, then there is a unique principle representation of $\bar{q}$ which involves J points $\tau_{k,1} < \ldots < \tau_{k,J}$ in the open interval $t_k < t < t_{k+m}$; that is, there are positive values $\beta_{k,j}$ such that

$$q_\ell = \sum_{j=1}^{J} \beta_{k,j}\, \psi_\ell(\tau_{k,j}), \quad \ell = 0,\ldots,L = 2J-1.$$

Clearly the prinicple representation gives an abstract setting for Gauss quadrature.

Let Q_0 denote the moment space for the Chebshev set $1, D^m s_{m+\ell}$, $\ell = 1,\ldots,L$, where $s_{m+\ell}$ are the basis elements in (4-1). The results in Section 4 for the case $M = D^m$ show that with $d\mu(x) = B_m(\bar{t}_k;x)dx$,

then $\bar{q}_0 = (q_{0,0}, q_{0,1}, \ldots, q_{0,L})$, $q_{0,\ell-1} = \delta_{\ell,1}$, is in Q_0. Thus, the principle representation is given with $\tau_{k,j}$, the zeros of the J-th degree B-spline orthogonal polynomial and with $\beta_{k,j}$ equal to $\beta^0_{k,j}$ in (4-6). By uniqueness, $\bar{q}_0$ is an interior point of the moment space Q_0 and so there is a closed sphere S_0 with center $\bar{q}_0$ in the interior of Q_0.

It follows from Theorem 5-2 that if the coefficients a_i of M are continuous, then the functions ψ_ℓ in (5-1b) form a Chebyshev set for all h_k sufficiently small. Let Q denote the moment space for this Chebyshev set. The curve $\Psi(t) = (\psi_0, \ldots, \psi_L)$ converges uniformly to the curve $\Psi_0(t) = (1, D^m s_{m+1}(t), \ldots, D^m s_L(t))$ on $t_k \leq t \leq t_{k+m}$, hence for sufficiently small h_k, the sphere S_0 is in the interior of the moment space Q. This establishes the next theorem.

THEOREM 5-3: Suppose the coefficients a_i of M are continuous. For a HODIE approximation with J auxiliary points, there is a positive H such that for any set of m+1 stencil points $\bar{t}_k$ with $0 < h_k = (t_{k+m} - t_k)/m \leq H$, there is a set of $\beta_{k,j}$'s and a unique set of J auxiliary points $\bar{\tau}_k$ with $t_k < \tau_{k,1} < \ldots < \tau_{k,J} < t_{k+m}$ such that the HODIE approximation is exact on $\mathbb{P}_{2J+m-1}$. The $\beta_{k,j}$'s are nonzero, all have the same sign, and are unique for a given normalization.

We call the special set of J auxiliary points which makes the HODIE approximation exact on $\mathbb{P}_{2J+m-1}$ the generalized B-spline Gauss points.

We now obtain a specific uniform bound on the β's for the variable coefficient case.

The system (5-2b) with $L-m+1 = J$ can be written in matrix form as

$$(B^0 + B^1)\,\bar{\beta} = \tilde{e}_1, \quad \bar{e}_1{}^t = (1,0,\ldots,0)$$

where B^0 is the matrix in Lemma 4-1. With $\bar{\beta} = \bar{\beta}^0 + \delta\bar{\beta}$, where $\bar{\beta}^0$ has components $\beta^0_{k,j}$ from the special case $M = D^m$, we have

$$(I + [B^0]^{-1} B^1)\,\delta\bar{\beta} = -\,[B^0]^{-1} B^1 \bar{\beta}^0$$

From (5-3b) it follows that elements of B^1 are given by

$$(B^1)_{1,j} = 0$$

$$(B^1)_{i,j} = h_k^i \sum_{p=0}^{m-1} h_k^p\, a_p(\tau_{k,j})\, D^p w(\bar{\gamma}_{k,m+\ell-1};\rho_{k,j}),$$

$$i = 2,\ldots,J.$$

Thus there are constants $k_{i,j}$ which depend only on max-norm bounds on $a_0,a_1,\ldots,a_{m-1}$ but not on h_k such that for all sufficiently small h_k

$$|(B^1)_{i,j}| \le h_k^i\, k_{i,j}, \qquad i,j = 1,\ldots,J.$$

Hence, from Lemma 4-1 we obtain bounds on the elements of the product $[B^0]^{-1} B^1$:

$$|([B^0]^{-1} B^1)_{r,j}| \le (h_k/\Delta\bar{\tau}_k)^{J-1}\, h_k \sum_i K_{r,i}\, k_{i,j}$$

Consequently, for all sufficiently small h_k, $I + [B^0]^{-1}B^1$ is invertible and we have the bound

$$\|\delta\bar{\beta}\|_\infty \le \|\,[B^0]^{-1} B^1\|_\infty\, \|\bar{\beta}^0\|_\infty/(1 - \|\,[B^0]^{-1} B^1\|_\infty)$$

where the norms are the vector max-norm and the matrix row-sum-norm. For all sufficiently small h_k there is, therefore, a constant K_0 such that

$$\|\delta\bar{\beta}\|_\infty \leq h_k \left(h_k/\Delta\bar{\tau}_k\right)^{J-1} K_0 \max_j|\beta^0_{k,j}|. \tag{5-4}$$

Lemma 4-1 with $i = 1$ gives a bound on $\beta^0_{k,j}$ which yields

$$\max_j|\beta_{k,j}| \leq \left(h_k/\Delta\bar{\tau}_k\right)^{J-1} \max_r(K_{r,1}) \left[1 + h_k \left(h_k/\Delta\bar{\tau}_k\right)^{J-1} K_0\right].$$

This gives the following result.

LEMMA 5-1: Under the same hypotheses as Theorem 5-3, there is a constant K which is independent of $h_k/\Delta\bar{\tau}_k$ such that for all sufficiently small h_k

$$\max_j|\beta_{k,j}| \leq K \left(h_k/\Delta\bar{\tau}_k\right)^{J-1}$$

Equations (5-2a), (5-3a), and (5-4) yield the following result.

COROLLARY 5-1: Under the same hypotheses as Theorem 5-3,

$$\alpha_{k,j} = \alpha^0_{k,j} + O(h_k[h_k/\Delta\bar{\tau}_k]^{J-1}), \quad \beta_{k,j} = \beta^0_{k,j} + O(h_k[h_k/\Delta\bar{\tau}_k]^{J-1}).$$

6. Discretization error for polynomial approximation. We begin by obtaining a bound on the solution of a homogeneous HODIE difference equation problem with values of the first $m-1$ divided differences given at $t_0 = A$. Let V denote the solution of

$$M_n V_k = 0, \quad k = 0,1,\ldots,$$

$$V[t_0],\ V[t_0,t_1],\ \ldots\ ,\ V[t_0,\ldots,t_{m-1}] \text{ are given,}$$

where M_n is from a HODIE approximation which is exact on $\mathbb{P}_L$ with $L \geq m$.

For fixed k, let p denote that unique element in $\mathbb{P}_m$ which interpolates to $V_k, V_{k+1}, \ldots, V_{k+m}$ at $t_k, t_{k+1}, \ldots, t_{k+m}$. Writing p in the Newton form of the interpolation polynomial, we have

$$p(t) = V[t_k]\, s_{k,0}(t) + V[t_k,t_{k+1}]\, s_{k,1}(t) + \ldots + V[t_k,\ldots,t_{k+m}]\, s_{k,m}(t)$$

$$s_{k,0}(t) = 1, \quad s_{k,\ell+1}(t) = s_{k,\ell}(t)\,(t - t_{k+\ell}),\ \ell = 0,\ldots,m-1.$$

Hence

$$M_n V_k = M_n p_k = V[t_k]\, C_{k,0} + \ldots + V[t_k,\ldots,t_{k+m}]\, C_{k,m}$$

where, with $a_m(t) \equiv 1$ and for $\ell = 0,1,\ldots,m$,

$$C_{k,\ell} = M_n(s_{k,\ell})_k = I_n(Ms_{k,\ell})_k = \sum_{j=1}^{J} \beta_{k,j} \sum_{i=0}^{\ell} a_i(\tau_{k,j})\, D^i s_{k,\ell}(\tau_{k,j})$$

and the last equality holds because the HODIE approximation is exact on $\mathbb{P}_L$, $L \geq m$.

Using the normalization (2-3c) and $D^m s_{k,m}(t) = m!$, we have

$$C_{k,m} = m! + \sum_{j=1}^{J} \beta_{k,j} \sum_{i=0}^{m-1} a_i(\tau_{k,j})\, D^i s_{k,m}(\tau_{k,j})$$

Set $H_n = \max_k h_k$. Because the auxiliary points are between t_k and t_{k+m}, there are constants K_ℓ which depend on $\max_i \|a_i\|_\infty$, but not on the mesh points nor on the auxiliary points, nor on H_n such that for $H_n < 1$

$$|C_{k,\ell}| \leq \max_j |\beta_{k,j}| \sum_{i=0}^{\ell} \|a_i\|_\infty |D^i s_{k,\ell}(\tau_{k,j})|$$

$$\leq K_\ell \max_j |\beta_{k,j}| (1 - H_n^{\ell+1})/(1 - H_n)$$

$$|C_{k,m} - m!| \leq H_n K_m \max_j |\beta_{k,j}| (1 - H_n^m)/(1 - H_n)$$

By Lemma 5-1, $\max_j |\beta_{k,j}| \leq K R^{J-1}$, $R = h_k/\Delta\bar{\tau}_k$. Consequently, if the ratio R is uniformly bounded as $H_n \to 0$, then the coefficients $C_{k,\ell}$ are uniformly bounded independent of k and H_n and for all sufficiently small H_n, $C_{k,m} \simeq m!$, so that $C_{k,m}$ is uniformly bounded above zero:

$$C_{k,m} \geq \delta > 0,$$

and we can divide the difference equation $M_n V_k = 0$ by $C_{k,m}$.

Using the definition of the m-th divided difference and the difference equation, we obtain an expression for the $(m-1)$-st divided difference at t_{k+1}:

$$\begin{aligned} V[t_{k+1},\ldots,t_{k+m}] = {} & (1 - mh_k\, C_{k,m-1}/C_{k,m})\, V[t_k,\ldots,t_{k+m-1}] \\ & - (mh_k\, C_{k,m-2}/C_{k,m})\, V[t_k,\ldots,t_{k+m-2}] \\ & - \ldots - (mh_k\, C_{k,0}/C_{k,m})\, V[t_k]. \end{aligned} \tag{6-1a}$$

We also have

$$V[t_{k+1},\ldots,t_{k+i}] = V[t_k,\ldots,t_{k+i-1}] + (-t_k+t_{k+i})\, V[t_k,\ldots,t_{k+i}], \quad i = 1,\ldots,k+m-1. \tag{6-1b}$$

Let $\|V_k\|_{m-1}$ denote

$$\|V_k\|_{m-1} = |V[t_k]| + |V[t_k,t_{k+1}]| + \ldots + |V[t_k,\ldots,t_{k+m-1}]|.$$

From (6-1) we obtain

$$\|V_{k+1}\|_{m-1} \leq (1 + H_n\, K)\, \|V_k\|_{m-1}$$

where

$$K = \max_{k,i} \{ 1 + m\, C_{k,m-i}\, /C_{k,m} + (-t_k+t_{k+i})/H_n \}$$

and, with the assumptions introduced above, K can be taken independent of H_n for all sufficiently small H_n. Consequently, we have

$$\|V_k\|_{m-1} \leq (1 + H_n\, K)^k\, \|V_0\|_{m-1} \leq e^{H_n K k}\, \|V_0\|_{m-1}, \quad k = 0,\ldots,n-m$$

From this we can obtain a bound on the Green's Function for the intial value difference equation problem.

Let $G_{\ell,k}$ denote the solution of

$$G_{\ell,k} = 0, \quad k = 0,1,\ldots,\ell+m-2$$

$$M_n G_{\ell,\ell-1} = 1$$

$$M_n G_{\ell,k} = 0, \quad k = \ell,\ell+1,\ldots,n-m$$

for $\ell = 1,\ldots,m-n$. Because $G_\ell[t_{\ell-1}] = G_\ell[t_{\ell-1},t_\ell] = \ldots = G_\ell[t_{\ell-1},\ldots,t_{\ell+m-2}] = 0$,

we have

$$M_n G_{\ell,\ell-1} = C_{\ell-1,m}\, G_\ell[t_\ell,\ldots,t_{\ell+m-1}]/mh_\ell = 1.$$

Hence, for $k = \ell,\ell+1,\ldots,n-m$,

$$\|G_{\ell,k}\|_{m-1} \leq e^{H_n K(k-\ell)} \|G_{\ell,\ell}\|_{m-1} = mh\, e^{H_n K(k-\ell)}/C_{\ell,m} \leq mH_n e^{H_n K(k-\ell)}/\delta.$$

The solution W of the initial value problem

$$M_n W_k = F_k, \quad i = 0,1,\ldots,n-m,$$

$$W[t_0],\ W[t_0,t_1],\ \ldots,\ W[t_0,\ldots,t_{m-1}] \text{ given},$$

is bounded by

$$\|W_k\|_{m-1} \leq e^{H_n K k}(\|W_0\|_{m-1} + K_0 \max_\ell |F_\ell|)$$

where

$$K_0 \leq m \sum_{\ell=0}^{n-m} e^{-H_n K\ell} H_n/\delta \leq m/[\delta K(1 - H_n K/2)].$$

Consider a HODIE approximation $M_n U_k = I_n f_k$ to the differential equation of the problem in (2-1) but subject to given intial divided difference conditions. If the approximation is exact on $\mathbb{P}_L$ and if the HODIE initial conditions (2-4) are also exact on $\mathbb{P}_L$, then by the result above and Theorerm 3-1, one has

$$|(U-u)_k| \leq \|(U-u)_k\|_{m-1} = O(H_n^{L-m+1}),$$

and the discretization error has the same order as the truncation error.

One can choose a set of solutions $u^{(j)}$, $j = 0,\ldots,m-1$, which span the space of all solutions of $Mu - f = 0$ and the corresponding HODIE approximations $U^{(j)}$ subject to initial conditions converge to $u^{(j)}$ as $O(H_n^{L-m+1})$. These can be used to obtain the unique HODIE approximation of the solution of (2-1) subject to the general boundary conditions in (2-1) where the HODIE approximation satisfies the general boundary conditions in (2-4). In addition to existence, uniqueness, and smoothness of the solution u of (2-1), one needs that the boundary conditions in (2-1c) are linearly independent on the space of polynomials $\mathbb{P}_m$, that is, if $\mathcal{M}^k[p,0] = 0$, $k = 0,\ldots,m-1$, for p in $\mathbb{P}_m$, then $p = 0$. We then have the following result

THEOREM 6-1: Suppose the coefficients a_i of M in (2-1b) and continuous and that there is a unique solution u of (2-1) which has a continuous $(L+1)$-st derivative, $L \geq m$. Suppose that the boundary conditions (2-1c) are linearly independent with respect to $\mathbb{P}_m$. Consider a sequence of partitions

$$A = t_{n,0} < t_{n,1} < \ldots < t_{n,n} = B, \quad n \geq m, \quad n \to \infty,$$

and sets of J auxiliary points

$$\bar{\tau}_{n,k} = (\tau_{n,k,1},\ldots,\tau_{n,k,J}), \quad t_{n,k} \leq \tau_{n,k,1} < \ldots < \tau_{n,k,J} \leq t_{n,k+m}.$$

Set $h_{n,k} = (t_{n,k+m} - t_{n,k})/m$ and $H_n = \max_k h_{n,k}$. Suppose that $H_n \to 0$ and that

$$R_{1,n} = \max_k \{ h_{n,k} / \min_{i=1,\ldots,m} [t_{k+i} - t_{k+i-1}] \}$$

and

$$R_{2,n} = \max_k \{ h_{n,k} / \min_{i,j=1,\ldots,J,\ i\neq j} [\tau_{n,k,i} - \tau_{n,k,j}] \}$$

are bounded as $n \to \infty$. Suppose that the HODIE approximation is exact on $\mathbb{P}_L$ with $L \geq m + J - 1$ and let $U^{(n)}$ denote the HODIE approximation on the n-th partition. Then for all sufficiently small H_n

$$|u(t_{n,k}) - U_k^{(n)}| = O(H_n^{L-m+1}).$$

7. Computation analysis. In this section, we consider the computational aspects of the HODIE method. We discuss specific features of our implementation and we compare the amount of work with other available methods. The discussion is restricted to the case of second order equations subject to Dirichlet boundary conditions for four reasons: it is simple, it is the most important case, it is readily generalized, and there are detailed analyses of other methods available for comparison.

The differential equation problem is

$$Mu(t) = a_2(t)u''(t) + a_1(t)u'(t) + a_0(t)u(t) = f(t), \quad A \leq t \leq B,$$

$$u(A) \text{ and } u(B) \text{ given,}$$

where, for generality, we have taken the coefficient of u'' in M to be a positive function a_2 rather than unity. Estimates $U_k = U(t_k)$ of $u(t_k)$ at mesh points $A = t_0 < t_1 < \ldots < t_n = B$ are obtained by solving the HODIE difference equation problem for $k = 0,\ldots,n-2$:

$$M_n U_k \equiv [\alpha_{k,0}U_k + \alpha_{k,1}U_{k+1} + \alpha_{k,2}U_{k+2}]/h_k^2 = \sum_{j=1}^{J} \beta_{k,j} f(\tau_{k,j}) \equiv I_h f_k,$$

$$U_0 = u(A), \quad U_n = u(B), \quad h_k = (t_{k+2} - t_k)/2,$$

where the coefficients α,β satisfy $M_n[s_\ell]_k = I_n[Ms_\ell]_k$ for s_ℓ, $\ell = 0,\ldots,L$, a basis for $\mathbb{P}_L$. We consider two choices of the auxiliary points $\tau_{k,j}$, $j = 1,\ldots,J$:

Regular auxiliary points: $\tau_{k,j} = t_k + (j-1)h_k/J$,

Gauss-type auxiliary points: the generalized B-spline Gauss points.

There are two distinct parts in an implementation of a specific HODIE approximation. The first part consists in the determination of the values of the coefficients $\alpha_{k,i}$, $i = 0,1,2$, and $\beta_{k,j}$, $j = 1,\ldots,J$, for each $k = 0,\ldots,n-2$ and then the determination of the values $I_n f_k$, $k = 0,\ldots,n-2$. The second part is the determination of the values U_k, $k = 1,\ldots,n-2$, of the solution of the resulting $(n-1)$-by-$(n-1)$ tridiagonal system of difference equations.

In the first part, the system of algebraic equations for the α's and β's is reducible: one solve a J-by-J system for the β's and then a 3-by-3 system for the α's; this is done for each $k = 0,\ldots,n-2$. This reducibility results in significant savings of work for the special second order case, $m = 2$, as well as in the general case. Although the Lagrange basis is convenient for theoretical analysis, we have found that it is computationally more efficient to use a different basis:

$$s_0(t) = 1, \quad s_1(t) = t - t_{k+1}, \quad s_2(t) = (t - t_k)(t - t_{k+2}),$$

$$s_{3+\ell}(t) = (t - t_k)(t - t_{k+1})(t - t_{k+2})\, p_{\ell-3}(t)$$

where $p_0(t) = 1$, $p_1(t) = (t - t_{k+1})$, $p_2(t) = (t - t_{k+1})^2$,

$p_3(t) = (t - t_k)p_2(t)$, $p_4(t) = (t - t_k)^2 p_2(t)$,

$p_5(t) = (t - t_{k+2})p_3(t)$, $p_6(t) = (t - t_{k+2})^2 p_3(t)$,

and so on.

With $\delta_1 = t_{k+1} - t_k$, $\delta_2 = t_{k+2} - t_{k+1}$, this choice leads to the following system for $\alpha_{k,i}/h_k^2 = \eta_{k,i}$

$$\eta_{k,0} + \eta_{k,1} + \eta_{k,2} = \sum_j \beta_{k,j}\, a_0(\tau_{k,j})$$

$$\delta_1 \eta_{k,0} \quad + \delta_2 \eta_{k,2} = \sum_j \beta_{k,j} [a_1(\tau_{k,j}) + (\tau_{k,j}-t_{k+1})\, a_0(\tau_{k,j})]$$

$$\delta_1 \delta_2 \eta_{k,1} \quad = \sum_j \beta_{k,j} [2a_2(\tau_{k,j}) + 2(\tau_{k,j}-t_{k+1})\, a_1(\tau_{k,j}) + (\tau_{k,j}-t_k)(\tau_{k,j}-t_{k+2})\, a_0(\tau_{k,j})].$$

Use of the normalization $\beta_{k,1} = 1$ eliminates one of the J HODIE equations. The remaining equations for the β's are

$$\sum_{j=2}^{J} \nu_{\ell,j}\, \beta_{k,j} = -\nu_{\ell,1}, \quad \ell = 1,\ldots,J-1,$$

$$\nu_{\ell,j} = s''_{2+\ell}(\tau_{k,j})\, a_2(\tau_{k,j}) + s'_{2+\ell}(\tau_{k,j})\, a_1(\tau_{k,j}) + s_{2+\ell}(\tau_{k,j})\, a_0(\tau_{k,j})$$

The choice of the basis elements makes the evaluation of the coefficients simple and it also gives a structure to the system which allows it to be solved easily. Specifically, for the Regular case, three of the auxiliary points are at mesh points. Arranging the system so that its first three columns corresponds to t_{k+1}, t_k, t_{k+2}, one finds that these columns have the special form

$-\delta_1\delta_2 a_1(t_{k+1})$	$-6\delta_1 a_2(t_k)+2\delta_1\delta_2 a_1(t_k)$	$6\delta_2 a_2(t_{k+2})+2\delta_1\delta_2 a_1(t_{k+2})$
$2\delta_1\delta_2 a_2(t_{k+1})$	X	X
0	X	X
0	X	X
0	0	X
0	0	X
0	0	0
⋮	⋮	⋮
0	0	0

where the X's indicate nonzero elements. This, of course, is very advantageous for solving for the β's in the regular case.

We consider the computational effort required first for a uniform partition: $t_k = kh$, $k = 0,\ldots,n$. We measure the effort in terms of the number F of function evaluations (a_2, a_1, a_0, or f) and the number M of multiplications required. In regard to the non-function-evaluation work, we assume: the total computational effort is proportional to the number of multiplications. Table 7-1 lists the effort required for various part of an implementation of the HODIE scheme.

Computation step	Regular Case				Gauss-type Case			
J =	3	5	7	9	2	3	4	5
Compute the β-matrix elements	8	39	89	137	6	14	36	50
Solve for the β's	3	17	47	111	1	7	38	47
Evaluate right sides of the α-equations	13	21	33	43	12	18	24	30
Solve for the α's	3	3	3	3	3	3	3	3
Solve the tridiagonal system for the U's	7	9	11	13	6	7	8	9
Total number of multiplications	34M	89M	183M	307M	28M	49M	109M	139M
Total number of function evaluations	4F	12F	20F	28F	8F	12F	16F	20F

Table 7-1: Number of multiplications and function evaluations required for each interior mesh point for HODIE approximations of orders 4, 6, 8, 10 for the Regular and the Gauss-type Cases for auxiliary points.

The β-matrix elements are found from a simple examination and assuming that the values of s", s', and s have been previously computed and stored (these values are independent of k since a uniform partition is assumed). The special structure of this matrix for the

Regural Case is assumed for estimating the work to solve this matrix equation. For the Gauss-type Case, we have a general (J-1)-by-(J-1) system to solve. Note that we assume that the Gauss-type auxiliary points have been previously computed or are otherwise known. The right sides of the α-equations are of a special form and the computation is carried out by forming $\beta_{k,j}a_{\ell}(\tau_{k,j})$ and then combining these appropriately. The solution of the α-equations is trivial and the final multiplications occur in solving the large tridiagonal system plus the evaluation of its right side. In the Regular Case, the function values at the mesh points and the auxiliary points are used more than once without recomputation.

We now use these work estimates to compare, roughly, the work of the HODIE method with other methods. The comparision is presented in Table 8-2 for seven methods, three different orders of accuracy (4, 6, and 8) and for both uniform and nonuniform partitions. The data for collocation by Hermite piecewise polynomials, least squares by splines, and discrete-Ritz are derived from Russell and Varah [1975], where they are described in detail. We have had to modify the multiplication counts in order to account for the slightly different differential equations used here and to rationalize the effect of the E_L term used by Russell and Varah. Note that the discrete Ritz method is limited to self-adjoint problems and is, therefore, not strictly comparable to the other methods included in Table 7-2. Collocation by splines and extrapolation of the trapezoid rule are analyzed in detail by Russell [1977] and we have adapted his results for our particular equation. Russell also considers collocation with Hermite cubics and quintics in detail.

We emphasize that the exact values of these operations counts depend on small details of the implementation of a particular algorithm and one can trade multiplications for additions, and so on, in some instances.

Method	Order of the method and mesh type					
	Fourth		Sixth		Eighth	
	Uniform	General	Uniform	General	Uniform	General
HODIE, Regular Case	34M+4F	40M+4F	89M+12F	113M+12F	183M+20F	241M+20F
HODIE, Gauss-type Case	28M+8F	32M+8F	49M+12F	57M+12F	109M+16F	140M+16F
Collocation, piecewise Hermite	38M+8F	42M+8F	62M+12F	72M+12F	145M+16F	159M+16F
Collocation, splines	24M+4F	56M+4F	37M+ 4F	99M+ 4F	52M+ 4F	152M+ 4F
Extrapolation of the trapezoid rule	32M+8F	32M+8F	70M+16F	70M+16F	165M+32F	165M+32F
Least squares, splines	66M+8F	90M+8F	198M+16F	270M+16F	440M+24F	580M+24F
Discrete Ritz, splines or piecewise Hermite	133M+9F	157M+9F	465M+15F	525M+15F	1200M+21F	1300M+21F

Table 7-2: Summary of number of multiplications (M) and function evaluations (F) for seven different methods. The counts are given per interior mesh point or interval and one would hope that methods with the same order give comparable accuracy.

The changes for collocation, least squares, and discrete Ritz from the equal to nonequal spaced meshs come from the need to evaluate the basis functions at each point. The changes for the HODIE method come from the need to evaluate the derivatives of the basis functions in each interval and we have assumed that two more multiplications are needed for each element of the β-matrix. A minor increase also occurs in the computation of the right side of the α-matrix equation. There are only insignificant changes in the extrapolation method's work, but it is not clear how effective extrapolation is for non-uniform spacing (consider

extrapolation, even for uniform spacing, for a problem for which the error behavior is as in Figure 8-3).

Considerable caution should be taken in attaching importance to the specific numbers in Table 7-2. These give only rough comparisons and various other considerations can completely override the difference between, say, 28 and 35 multiplications per point. We can only conclude that the first five methods are generally comparable in work and the last two seem unlikely to be competitive. Collocation with splines seems to gain a work advantage as the order increases, but it is simultaneously increasingly complicated near the boundaries which may well negate this advantage somewhat.

To obtain a realistic evaluation of these methods, one needs not only actual execution times for the different methods for a range of problems and accuracies, but one also needs to consider other factors such as numerical reliability and stability, ease of programming, and memory requirements.

The operation counts for the HODIE method for ordinary differential equations given here indicate that the work is close to the work involved in a number of other available methods. But, the comparisions for partial differential equations indicate that the work for the HODIE method is significantly less than for other available methods; see Lynch and Rice [1975,1977a]

8. Experimental results. We present support for the following points: (1) The HODIE method converges as predicted by theory; there are no unforeseen numerical complications. (2) There are no unforeseen difficulties or complexities in implementation. (3) There is a definite pattern in the relationship among the accuracy actually achieved, the actual computation time, and the order of the method. Specifically, the higher the desired accuracy, the higher should the order of the method be to minimize computation time. (4) The use of Gauss-type auxiliary points gives the rate of convergence predicted by theory. (5) The use of Gauss-type auxiliary point for the operator D^2 improves the rate of convergence for a general second order operator M over that expected for a general set of auxiliary points.

The first two points must be verified for any new method; the third point applies to collections of methods with varying orders; and the last two points apply to the HODIE method and to certain other schemes, such as collocation and Galerkin which have "superconvergence" characteristics.

We note that most of the content of these five points is supported by the theory presented explicitly or implicitly in the preceding sections, or is part of the general folklore about numerical computations. Nevertheless, experience shows that points such as these must be verified experimentally for a new method and, for the rates of convergence, they must be verified in the sense of establishing that asymptotic results are valid in the range of ordinary application.

Accordingly, we have run hundreds of cases for numerous second order ordinary differential Dirichlet boundary value problems. The results of these experiments support the points listed above and we have acquired

confidence in the reliability of the HODIE method.

The Fortran program we wrote seemed to be as easy to write and to debug as a program for any other method of solving this class of problems. However, we quickly found that in order to verify the rates of convergence for very high order HODIE schemes, we had to use very high precision. In the remainder of this section, we discuss only a small subset of the experiments which we performed.

All computation was done on the Purdue University CDC6500 with double precision arithmetic which uses values with about 28 decimal digits. In each experiment, the domain of the problem was partitioned by an equal-spaced mesh with N subintervals, so the mesh spacing h was proportional to $1/N$.

Example 8-1: $u''(t) - 4u(t) = 2\cosh(1), \quad 0 < t < 1$

$u(0) = u(1) = 0$

solution: $u(t) = \cosh(2t-1) - \cosh(1)$

This problem has been used by Russell and Shampine [1972], de Boor and Swartz [1973], and others.

Figure 8-1 summarizes one set of experimental results. The logarithm of the maximum error is plotted versus the logarithm of the number of subintervals for eleven different sets of $J=5$ auxiliary points. We describe the various curves in this figure and give our interpretation of the results.

(a) The topmost curve gives the results when 5 Regular (equal spaced) auxiliary points were used. One expects at least $O(h^5)$ rate of convergence with a set of 5 auxiliary points because the approximation is locally exact on at least $\mathbb{P}_7$. The curve shows a very consistent $O(h^6)$ rate of

convergence. The central auxiliary point is the central mesh point of the three-point difference operator M_N and it is clear from the symmetry of the differential operator that this auxiliary point is a zero of every odd-degree generalized B-spline orthogonal polynomial. This (or symmetry) shows that one expects $O(h^6)$ rather than $O(h^5)$ convergence

(b) There is a set of nine curves in Figure 8-1 which have sharp downward spikes at N = 4,8,16,25,32,50,64,100, and 200, respectively. The set of 5 auxiliary points used for each one of these curves is the set of 5 Gauss-type points for that value of N at which the spike occurs. One has nine different sets of these Gauss-type points because their locations depends on $h = 1/N$. The curve with spike at N = 8 is typical and we describe some of its features. First, the spike is very abrupt, for the curve also shows the error for the cases of N = 7 and N = 9. Second, for N different from 8, the auxiliary points are not the Gauss-type points, hence one expects only $O(h^6)$--one of the points is the central mesh point of the operator M_N--and this behavior can be seen for large values of N, say N greater than about 16 for the curve with spike at N = 8.

(c) Consider the tips of the spikes from the collection of nine curves discussed in (b). If one joins the tips, one sees a very consistent $O(h^{10})$ rate of convergence for N up to 64. This is what one expects, since this new curve gives the behavior of the error when 5 Gauss-type points are used for each N. The maximum error at N=64 is about 10^{-25} and the $O(h^{10})$ rate of convergence breaks down beyond N = 64 because of roundoff error; the values of the Gauss-type points were accurate only to about one part in 10^{15} (single precision on the CDC 6500).

(d) The last curve is the one for 5 Gauss-type points for the operator $M = D^2$. Except for the central auxiliary point, these are not the Gauss-type points for the operator $D^2 - 4$. One expects at least $O(h^6)$ rate of convergence; however, a very consistent $O(h^8)$ rate of convergence is observed. As h tends to zero, the Gauss-type auxiliary points tend to those of the operator D^2, hence one expects improvement over an arbitrary set of auxiliary points, even a set which contains the central mesh point of the operator M_N.

Example 8-2: Typical of a fairly difficult problem is one taken from Rachford and Wheeler [1974]:

$$t_0 = 0.36388$$

$$\frac{d}{dt}\left[(.01 + 100(t-t_0) \frac{d}{dt} u(t) \right] = -2\{1 + 100(t-t_0)(\tan^{-1}[100(t-t_0)] - \tan^{-1}[100\ t_0]) \}$$

$$u(0) = u(1) = 0$$

$$\text{solution:}\quad u(t) = (1-t) \{ \tan^{-1}[100(t-t_0)] + \tan^{-1}[100\ t_0] \}$$

The solution has a very sharp rise near $t = 0.36$: it increases form about 0.1 at $t = 0.3$ to about 1.7 at $t = 0.4$ and then it decreases nearly linearly to 0 at $t = 1$. See Rachford and Wheeler for a graph of the solution.

Results for two sets of auxiliary points are shown in Figure 8-2: three Regular auxiliary points--which is the $O(h^4)$ Størmer-Numerov scheme--and the seven Gauss-type auxiliary points for the operator D^2. One sees that there is a considerable irregularity for N up to about 100 and then for large N, the error decreases smoothly at rates of $O(h^4)$ and

$O(h^{10})$, respectively. For a general set of seven auxiliary points, one expects $O(h^7)$ rate of convergence; the use of the Gauss-type points for the operator D^2 improves the rate of convergence to $O(h^{10})$.

To compare efficiency, we note that the Størmer-Numerov scheme with $N = 300$ required almost exactly the same amount of computation time as the seven-point scheme with 100 points. The Størmer-Numerov scheme achieved a maximum erro of .00026 which is almost exactly 100 times greater than the error for the higher order scheme.

Finally, we note that the usefulness of extrapolation techniques is doubtful for either of these schemes for N less than about 100.

Example 8-3: The final example we discuss is:

$$u''(t) + \sin(t)\, u'(t) + 4\, t^2\, u(t) = 2[1 + t \sin(t)] \cos(t^2), \quad 0 \quad t \quad 5,$$

$$u(0) = u(5) = 0,$$

solution: $u(t) = \sin(t^2)$.

The solution has several oscillations as t ranges from 0 to 5.

We solved this problem with a wide variety of HODIE schemes and Figure 8-3 summarizes the results for a selection of them. This figure shows the relationship between work, order, and accuracy. The logarithm of the execution time is plotted versus the logarithm of the maximum error. Since the error is, asymptotically, proportional to N^{-p} and the time is proportional to N, one expects straight-line graphs for large N; the slope gives p.

One sees the advantage that comes from using a higher order method for higher accuracy. All of the methods require a fairly large value of N to achieve any significant accuracy. The low order methods are competitive only for very low accuracy requirements. The 5-point Regular method and the

3-point D^2 Gauss-type method both are $O(h^6)$ methods, but the maximum error of the Regular method is about 10 times larger than the Gauss-type method for the same execution time.

REFERENCES

Birkhoff, G., and C.R. de Boor, [1965], Piecewise polynomial interpolation and approximation, in Approximation of functions, Editor H.L. Garabedian, Elsevier Publishing Co., Amsterdam, 164-190.

de Boor, C., and R.E. Lynch, [1966], On splines and their minimum properties, J. Math and Mech 15 953-970.

de Boor, C.W. and B. Swartz, [1973], Collocation at Gaussian points, SIAM J. Num. Anal., 10 582-606.

Collatz, L., [1960], The numerical treatement of differential equations, 3rd Edition, Springer-Verlag, Berlin.

Curry, H.B., and I.J. Schoenberg, [1966], On Polya frequency functions IV. The spline functions and their limits, J. Analyse Math. 17 71-107.

Doedel, E.J., [1976], The construction of finite difference approximations to ordinary differential equations, Report, Applied Mathematics, California Institute of Technology, Pasadena, California

Karlin, S.J., and W.J. Studden, [1966], Tchebycheff systems: with applications in analysis and statistics, Interscience, New York.

Lynch, R.E., [1977a], $O(h^6)$ accurate finite difference approximation to solutions of the Poisson eqaution in three variables, Department of Computer Science Report CSD-TR 221, Purdue University, Feb. 15.

Lynch, R.E., [1977b], $O(h^6)$ discretization error finite difference approximation to solutions of the Poisson equation in three variables, Department of Computer Science Report CSD-TR 230, Purdue University, April 19.

Lynch, R.E., and J.R. Rice [1975], The HODIE method: A brief introduction with summary of computational properties, Department of Computer Science Report 170, Purdue University, Nov. 18.

Lynch, R.E., and J.R. Rice [1977a], High accuracy finite difference approximation to solutions of elliptic partial differential equations, Department of Computer Science Report CSD-TR 223, Purdue University, Feb. 21.

Lynch, R.E., and J.R. Rice [1977b], High accuracy finite difference approximation to solutions of elliptic partial differential equations, (complete revision of Report CSD-TR 223) to appear.

Osborne, M.R., Minimizing truncation error in finite difference approximation to ordinary differential equations, [1967], Math. of Comp. 21 133-145.

Phillips, J.L., and R.J. Hanson, [1974], Gauss quadrature rules with B-spline weight functions, Math. Comp. 28 666 and microfiche supplement.

Rachford, H.H., and M.F. Wheeler, An H^{-1} Galerkin procedure for the two-point boundary value problem, in Mathematical Aspects of Finite Elements in Partial Differential Equations, Editor C.W. de Boor, Academic Press, New York, 253-382.

Russell, R.D., [1977], A comparison of collocation and finite differences for two-point boundary value problems, SIAM J. Numer. Anal. 14 19-39.

Russell, R.D., and L.F. Shampine, [1972], A collocation method for boundary value problems, Numer. Math. 19 1-28.

Russell, R.D., and J.M. Varah, [1975], A comparision of global methods for linear two-point boundary value problems, Math. Comp. 29 1007-1019.

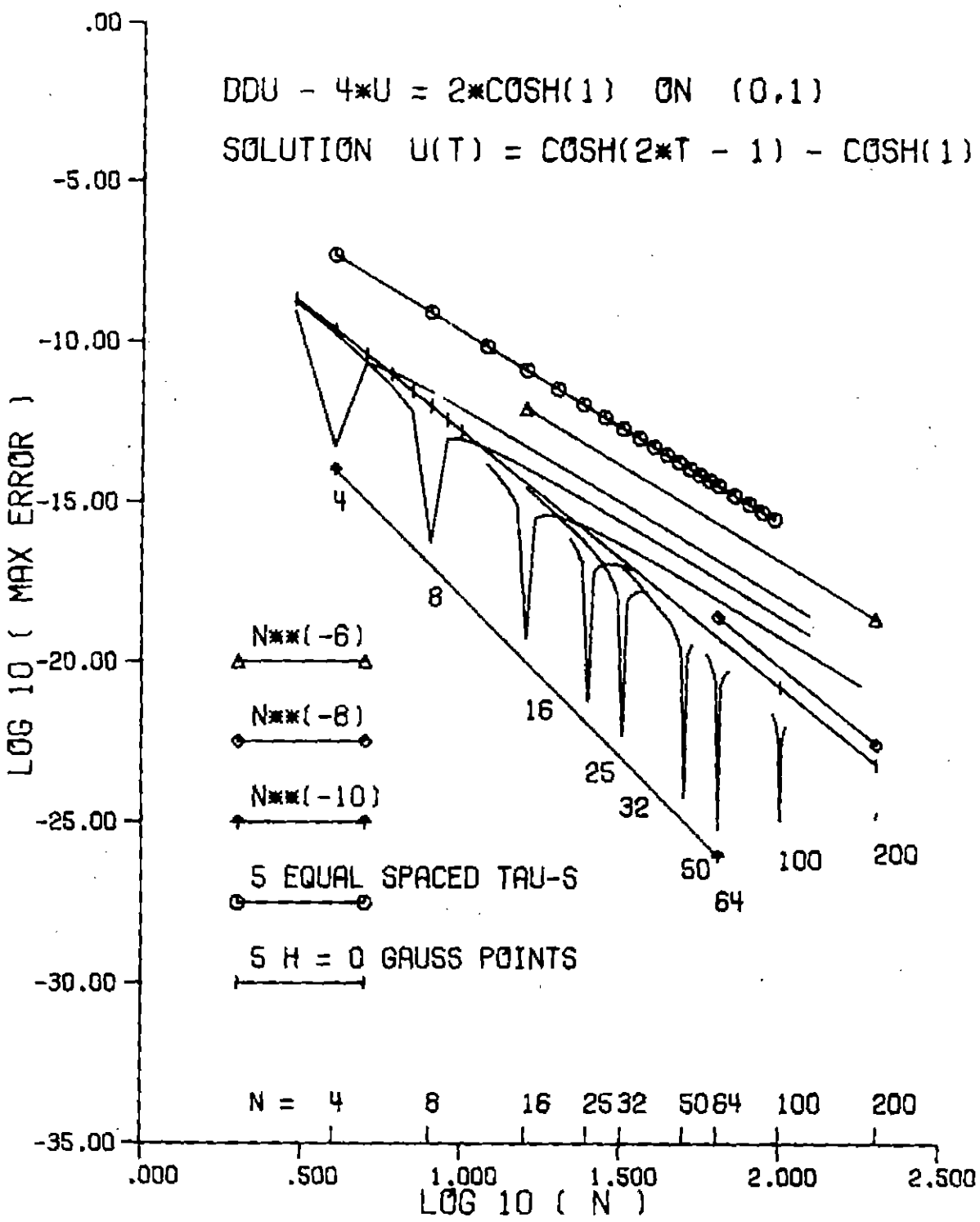

Figure 8-1: Behavior of the error as a function of number N of subintervals for eleven different 5 τ-point HODIE schemes for Example 8-1.

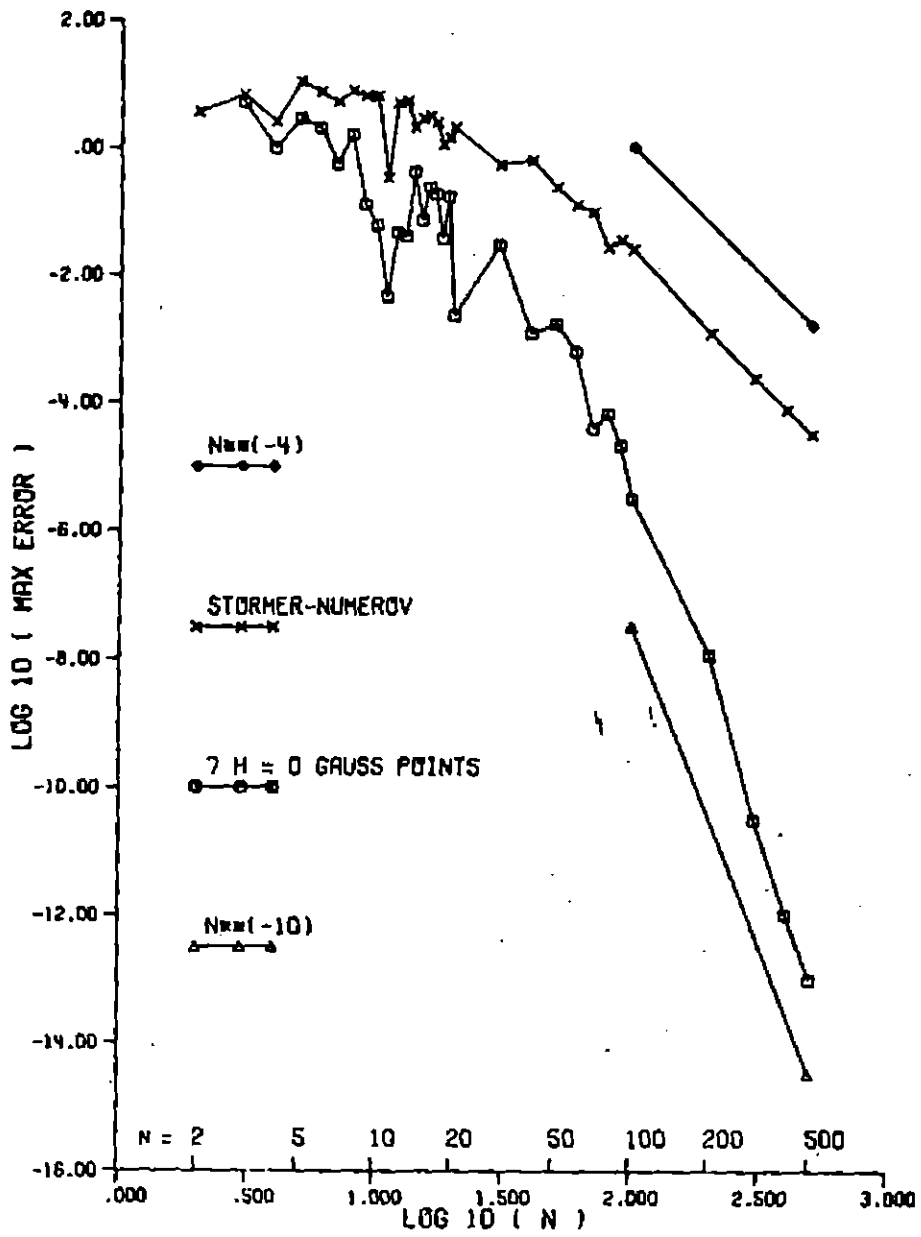

Figure 8-2: Behavior of the error as a function of the number N of subintervals for two HODIE schemes, one the Størmer-Numerov scheme, and one with 7 Gauss-type τ-points for the operator D^2 for Example 8-2.

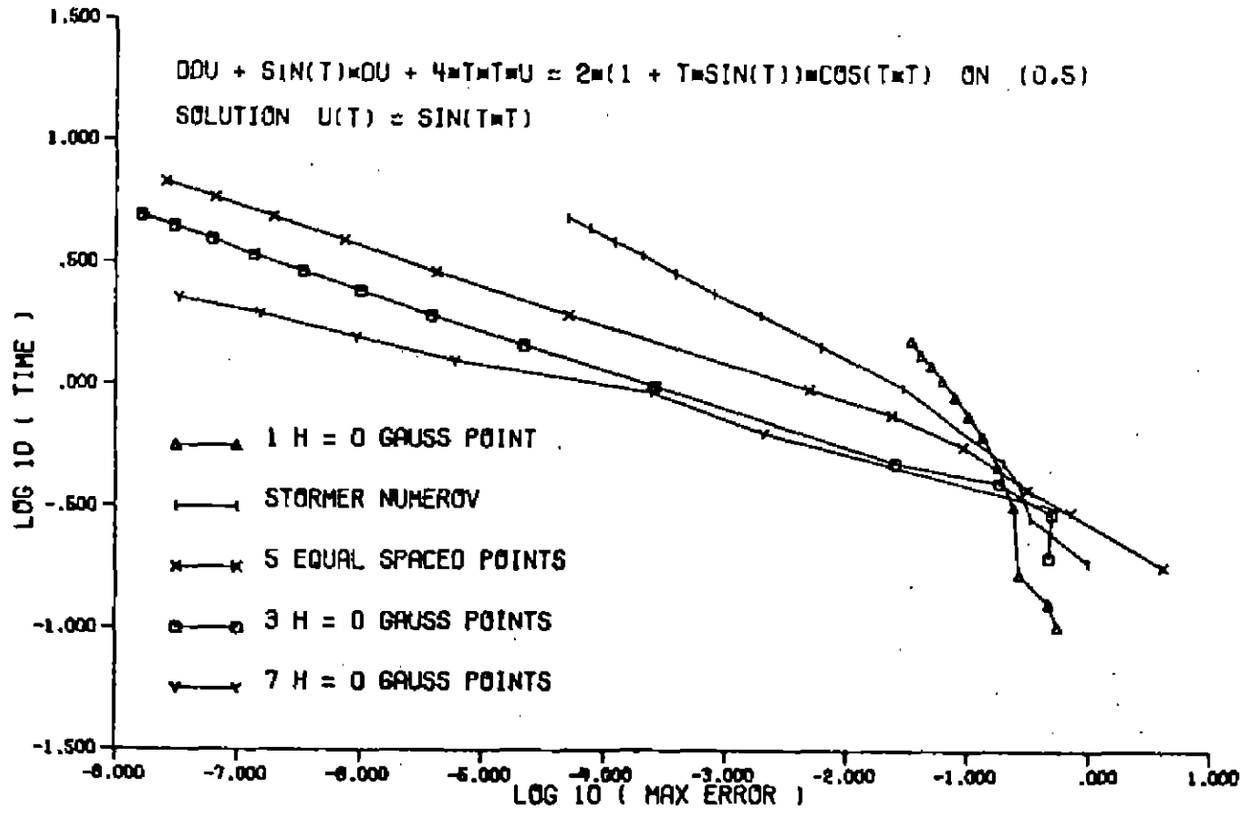

Figure 8-3: Illustration of the relationship between work (execution time), accuracy achieved, and order of the HODIE method for Example 8-3.