1976

The Hodie Method for Ordinary Differential Equations

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

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

Report Number:
76-188

THE HODIE METHOD FOR ORDINARY DIFFERENTIAL EQUATIONS

R. E. Lynch and J. R. Rice
Mathematical Sciences
Purdue University

CSD-TR 188

April 30, 1976

Abstract

This paper describes a method of generating high order difference approximations to an m^{th} order differential operator L using (m + 1) points. It uses certain auxiliary points for each (m + 1)-tuple in order to achieve an arbitrary specified order of truncation error in the difference approximation. In the context of solving Lu = f, this method may be interpreted as using $L_h u = I_h f$ where L_h is an (m + 1)-step difference operator and I_h is an expansion of the identity. The acronym HODIE (High Order Differences via Identity Expansion) comes from this interpretation. Under natural assumptions we prove the existance of such approximations, and establish the order of their truncation error and discretization error. We also show the existance of certain Gauss-type points where even higher orders of convergence are achieved. We give an operations count comparison of this new method with five others which shows the HODIE method to be among the most efficient. A brief selection from extensive experiments is given which supports the practicality of the method.

THE HODIE METHOD FOR ORDINARY DIFFERENTIAL EQUATIONS

Robert E. Lynch and John R. Rice

1. Introduction. We consider the equation $Lu = f$ where L is a linear m^{th} order differential operator and rather general boundary conditions at one or two points are given. We derive a difference approximation L_h to L based on a set $\bar{t}$ of points in $[0,1]$. The right side of the equation is modified to $I_h f$ so that the truncation error in the approximation $L_h U = I_h f$ to $Lu=f$ is any desired order of accuracy. The key points are (a) the difference approximation, L_h, is based on the minimal number $(m+1)$ of points and (b) the coefficients of L_h are calculated locally on the basis of certain values of f i.e.

$$\frac{1}{h_n^m} \sum_{i=0}^{m} \alpha_{n,i} U_{n+i} = \sum_{i=1}^{J} \beta_{n,j} f(\tau_{n,j}) \qquad n=0,\ldots,N-m$$

are the approximating equations for $Lu=f$. As the notation suggests, the modified right side may be interpreted as an expansion of the identity applied to f; hence the acronym HODIE from "High Order Differences via Identity Expansion". The method is presented in detail in Section 2.

A thorough survey of methods for 2-point boundary value problems is given by Keller [6]. The method which most resembles the HODIE method is the Mehrstellenverfahren, see Collatz [2]. Indeed, the Stormer-Numerov method is a special case of both the HODIE method and Mehrstellenverfahren. However, the general use of the Mehrstellenverfahren requires information about the derivatives of f, etc. which often renders it impractical. The HODIE method has no such constraint. Another closely related piece of work is that of Birkhoff and Gulati [1] who exhaustively investigate difference approximations for second order ordinary differential operators and the Laplacian. Various aspects of their results were

extended slightly during the summer of 1975 by G. Birkhoff, R.E. Lynch, and S. Gulati. The first instances of the HODIE method were discovered, Fall 1975, by R.E. Lynch while extending this work.

As seen below the HODIE is quite competetive for 2-point boundary value problems, but its real potential seems to lie with applications to elliptic boundary value problems in 2 and 3 dimensions, see Lynch and Rice [7].

The results presented in this paper are briefly summarized as follows. In Section 3 we establish the existance and, in certain cases, the uniqueness of HODIE difference approximations and the order of their truncation error is found. We also show that there are certain Gauss-type points which increase the truncation error in a manner analogous to Gauss quadrature. The existance and nature of these points is established. In Section 4 we show the HODIE method gives stable difference approximations and that the convergence of the discretization error is given by the order of the truncation error. Section 5 contains a analysis of computational effort via an operations count. A comparison is made with five other methods (collocation with piecewise Hermite polynomials, collocation with splines, extrapolation of the trapezoidal rule, least squares method for splines and discrete Ritz with splines or piecewise Hermite polynomials). These results should be viewed as only indications of actual efficiency, but they suggest that the HODIE method is among the most efficient. Finally in Section 6 we give a small sample of extensive experiments which verify that HODIE method works as the theory predicts and that there are no unforeseen difficulties in its implementation.

2. HODIE approximation of differential operators. We construct high accuracy m+1 point difference approximation to m-th order differential operators, M:

$$M[u,f] = Lu - f \tag{2-1a}$$

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

as well as high accuracy approximations to boundary conditions of the form

$$\begin{aligned} &L^{k,0}\, u(0) + L^{k,1}\, u(1) = c_n, \quad k = 0,\ldots,m-1 \\ &L^{k,n}\, u(n) = \sum_{i=0}^{\ell} a_{k,n,i}\, D^i u(n) \quad , \quad n=0,1, \ \ell \leq m-1 \end{aligned} \tag{2-1c}$$

The analysis has two phases: local analysis of truncation error (Section 3) and global analysis of discretization error (Section 4). We use the same letters $\bar{t}$ and $\bar{\tau}$ for point sets involved but in the local analysis we index them locally (e.g. 0 to m or 1 to J) while in the global analysis they fill out the interval [0,1].

The HODIE approximations are based on a pair of point sets, $\bar{t}$ and $\bar{\tau}$. The $m+1$ points in $\bar{t}$ are mesh points and the J points in $\bar{\tau}$ are used to obtain the high accuracy. For the analysis of the local truncation error, it is sufficient to take

$$\bar{t} = \{t_0 < t_1 < \ldots < t_m\}$$

$$\bar{\tau} = \{t_0 \leq \tau_1 < \ldots < \tau_J \leq t_m\}$$

except at the boundary where, we take $\bar{\tau}$ as above and

$$\bar{t} = \{t_o = t_1 = \ldots = t_\ell < t_{\ell+1} < \ldots < t_m\}$$

or

$$\bar{t} = \{t_0 < t_1 < \ldots < t_{m-\ell} = t_{m-\ell-1} = \ldots = t_m\}$$

Throughout, we use the subscript i for an element of $\bar{t}$ and the subscript j for an element of $\bar{\tau}$. When we treat a limit, $h \equiv (-t_0 + t_m)/m \to 0$, we consider

$$t_i = \gamma_i h, \quad \tau_j = \rho_j h$$

with γ_i, ρ_j fixed.

The HODIE difference operator, M_h, used to approximate M is given by

$$M_h[u,f] = L_h u - I_h f \tag{2-2a}$$

$$L_h u = \frac{1}{h^m} \Sigma_{i=0}^{m} \alpha_i u_i, \qquad u_i = u(t_i) \tag{2-2b}$$

$$I_h f = \Sigma_{j=1}^{J} \beta_j f_j, \qquad f_j = f(\tau_j) \tag{2-2c}$$

For the boundary conditions, we use (2-2a) with L_h given by

$$L_h u = \frac{1}{h^m} \{\Sigma_{i=0}^{\ell} h^i \alpha_i D^i u(0) + \Sigma_{i=\ell+1}^{m} \alpha_i u(t_i)\}$$

or

$$L_h u = \frac{1}{h^m} \{\Sigma_{i=0}^{m-\ell-1} \alpha_i u(t_i) + \Sigma_{i=m-\ell}^{m} \alpha_i h^{i-m-\ell} D^{i-m-\ell} u(t_m)\} \tag{2-2d}$$

The truncation operator, T_h, and the truncation error $||T_h||$ with respect to a space of functions S are,

$$T_h[s] = M_h[s,Ls] - M[s,Ls] = M_h[s,Ls],$$

where $s \in S$ and

$$||T_h|| = \sup_{s \in S,\ ||s|| \neq 0} |T[s]|/||s||$$

As usual, we say that the approximation is exact, is consistent, or is of order p with respect to S if:

$$||T_h|| = 0, \quad ||T_h|| \to 0 \text{ as } h \to 0, \quad ||T_h|| = 0(h^p) \text{ as } h \to 0.$$

respectively.

The truncation error is related to the discretization error e. Suppose that u,v are such that $M[u,f] = 0$, $M_h[v,f] = 0$. Then $e = v-u$ and

$$L_h e = L_h u - I_h f = L_h u - I_h Lu = M_h[u,Lu] = T_h[u]$$

Consequently, if $u \in S$ and if the truncation error is order p, then $|L_h e| = 0(h^p)$. In Section 4 we show that under suitable hypotheses, the discretization error is $0(h^p)$ if the truncation error is $0(h^p)$.

The coefficients, α_i, β_j, are determined by making the approximation exact on a linear space, S, of dimension K. Specifically, if $s_0, s_1, \ldots, s_{K-1}$ is a basis for S, then the coefficients satisfy the system of equations

$$\text{(2-3a)} \qquad \Sigma_{j=1}^{J} \beta_j = 1$$

$$\text{(2-3b)} \qquad M_h[s_k, L s_k] = \frac{1}{h^m} \Sigma_{i=0}^{m} \alpha_i s_{k,i} - \Sigma_{j=1}^{J} \beta_j L s_{k,j} = 0, \quad k = 0,\ldots,K-1$$

and similarly for the boundary conditions. Because of the normalization (2-3a) the operator I_h in (2-2c) can be regarded as an approximation to the identity operator.

After normalization, there remain $m+J$ coefficients determined by setting $T_h[s] = 0$ for all $s \in S$ or, equivalently, $T_h[s_k] = 0$, $k = 0,1,2,\ldots,K$. If the points in $\bar{\tau}$ are specified, then one expects to be able to take $K = m+J$ and the system is linear. If the points in $\bar{\tau}$ are regarded as parameters, then these can be chosen to increase K; one expects to be able to take $K = m+2J$. In this case,

one has to solve a system of nonlinear equations analogous to those for Gauss quadrature and in Section 3 we show this can be done.

Examples: We illustrate the HODIE approximation in a simple case where the operator L is $D^2 = d^2/dt^2$; the mesh points are equal spaced: $t_i = ih$, $i=0,1,2$; and $\mathbb{P}_k$ denotes the set of polynomials of degree at most k. The truncation error is $O(h^p)$ for sufficiently smooth functions.

Scheme 1: Exact on $\mathbb{P}_2$, $O(h)$ uncentered divided difference approximation:

$$\frac{1}{h^2}[u_0 - 2u_1 + u_2] - f(t_0) = 0$$

Scheme 2: Exact on $\mathbb{P}_3$, $O(h^2)$ divided central difference approximation: the same as Scheme 1, except f is evaluated at t_1 instead of t_0.

Scheme 3: Exact on cubic splines with joints at the mesh points, $O(h^2)$ approximation:

$$\frac{1}{h^2}[u_0 - 2u_1 + u_2] - \frac{1}{6}[f(t_0) + 4f(t_1) + f(t_2)] = 0$$

Scheme 4: Exact on $\mathbb{P}_5$, $O(h^4)$ Størmer-Numerov approximation:

$$\frac{1}{h^2}[u_0 - 2u_1 + u_2] - \frac{1}{12}[f(t_0) + 10\, f(t_1) + f(t_2)] = 0$$

Scheme 5: Exact on $\mathbb{P}_7$, a new $O(h^6)$ approximation:

$$\frac{1}{h^2}[u_0 - 2u_1 - u_2] - [5f(\tau_1) + 14\, f(\tau_2) + 5f(\tau_3)]/24 = 0$$

$$\tau_1 = h - (2/5)^{1/2}h, \quad \tau_2 = h, \quad \tau_3 = h + (2/5)^{1/2}h$$

The points τ_j in Scheme 5 are the zeros of the cubic polynomial b_3, of the set $b_k \in \mathbb{P}_k$, orthogonal with respect to the inner product

$$(r,s) = \int_{t_0}^{t_2} G(t_1,t)\, r(t)\, s(t)\, dt$$

where G is the Green's function for the problem $u'' = f$, $u(t_0) = u(t_2) = 0$. The use of the J zeros, $\tau_1,\ldots,\tau_J$, of b_J gives $0(h^{2J})$ truncation error. The value of β_j is given by

$$\beta_j = \int_{t_0}^{t_1} G(t_1,x)\, \ell_j(x)dx,$$

where ℓ_j is the jth Lagrange polynomial for the points $\bar{\tau}$. Values of the zeros and the β's are given in [8].

To use these schemes to estimate the solution of the two-point boundary value problem: $u''(t) = f(t)$, $t \in (0,1)$, $u(0)$ and $u(1)$ given, one subdivides $[0,1]$ into N equal subintervals and obtains the system:

$$U_0 = u(0), \quad U_N = u(1)$$

$$\frac{1}{h^2}[U_{i-1} - 2U_i + U_{i+1}] = g_i, \quad i=1,\ldots,N-1, \quad U_i = U(ih) \tilde{=} u(ih)$$

where g differs from scheme to scheme. In each case, however, the matrix formulation has the same tridiagonal $(N-1)$-by-$(N-1)$ matrix. Once g is evaluated, the work to solve the system is independent of the scheme. The maximum principle applies to the difference equation (or, equivalently, the matrix is of monotone type), hence the discretization error is $0(h^p)$ where p is the order of the truncation error. Thus, assymptotically, the higher the order of the scheme, the smaller is the work required to achieved a given accuracy.

Specifically, we compare the $0(h^6)$ new scheme (NEW) and the $0(h^4)$ Størmer-Numerov scheme (SN). Assume $4N$ units of work are needed to solve the $(N-1)$-by-$(N-1)$ tridiagonal system and let W denote the units of work in evaluating f at a single point. Then the total work for these schemes is $(4+2W)N_{NEW}$ and $(4+W)N_{SN}$. For equal accuracy we have: error $= 1/N_{NEW}^6 = 1/N_{SN}^4$ and consequently, the SN scheme requires $(4+W)(N_{NEW})^{3/2}$ units of work. If $N_{NEW}=100$ is required, then the SN scheme requires between 5 and 10 times the work required by the NEW scheme. If the number of points in $\bar{\tau}$ is increased to, say, 5, then one obtains

a new $O(h^{10})$ scheme and in terms of its value of N_{NEW}, the work involved in the SN scheme for equal accuracy is $(4+W)(N_{NEW})^{5/2}$. This is a factor between 200 and 1000 more than the work, $(4+5W)N_{NEW}$, for the NEW 5 $\bar{\tau}$ point scheme if $N_{NEW} = 100$.

Application of these schemes to the initial value problem:

$u''(t) = f(t)$, $t > 0$, $u(0)$ and $u'(0)$ given, leads to the system

(2-4a) $$U_0 = u(0)$$

(2-4b) $$\frac{2}{h^2}\left[-u(0) - hu'(0) + U_1\right] = g_0$$

(2-4c) $$\frac{1}{h^2}\left[U_{i-1} - 2U_i + U_{i+1}\right] = g_i, \quad i = 1,2,\ldots$$

If the order of (2-4c) is p, then one wants the order of (2-4b) to be p also. This leads to the following boundary approximations for the four schemes above which are exact on polynomial subspaces:

Scheme 1, $O(h)$; $g_0 = f(0)$

Scheme 2, $O(h^2)$: $g_0 = f(h/3)$

Scheme 4, $O(h^4)$: $g_0 = \frac{1}{36}\left[9\ f(0) + 25\ f(24/5) + 2\ f(h)\right]$

Scheme 5, $O(h^6)$: $g_0 = \beta_1\ f(\tau_1) + \beta_2\ f(\tau_2) + \beta_1\ f(\tau_3)$

$\beta_1 = .4018638275$ $\beta_2 = .4584822127$ $\beta_3 = .1396539598$

$\tau_1 = .08858795951$ $\tau_2 = .4094668644$ $\tau_3 = .7876594618$

3. Approximation with polynomial subspaces. We derive specific results only for the HODIE approximation for interior approximation. The details for the boundary approximation, (2-2c), are obtained with easy modifications.

Let $\bar{\xi} = \{\xi_1 <\cdots< \xi_K\}$ denote a set distinct points (such as $\bar{t}$ or $\bar{\tau}$) and define w and ℓ by

$$w(\bar{\xi};t) = \pi_{k=1}^{K} (t - \xi_u), \quad \ell_k(\bar{\xi};t) = w(\bar{\xi};t)/(t-\xi_k)w'(\bar{\xi};\xi_k)$$

so that $\ell_k(\bar{\xi};\cdot)$ is a Lagrange basis element for $\mathbb{P}_{K-1}$. We use the following basis for the polynomials:

$$s_i(t) = \ell_i(\bar{t};t), \qquad i = 0,\ldots,m$$

$$s_{m+k}(t) = t^{k-1} w(\bar{t};t)/h^k, \quad k = 1,2,\ldots$$

then s_{m+k} has a zero at each mesh point t_i.

To determine the behavior as $h \to 0$, we take $t_i = \gamma_i h$, $\tau_j = \rho_j h$, with

$$\bar{\gamma}_0 = \{\gamma_0 < \gamma_1 <\cdots<\gamma_m\}$$

$$\bar{\rho} = \{\rho_1 < \rho_2 <\cdots<\rho_J\}$$

fixed as $h \to 0$. L applied to the basis is given by

(3-1a) $$L\, s_i(t) = \frac{m!}{w'(t;t_i)} [1 + O(h)] \equiv \psi_i(h;\gamma) \quad i = 0,\cdots,m$$

(3-1b) $$L\, s_{m+k}(t) = [\gamma^{k-1} w(\bar{\gamma};\gamma)]^{(m)} [1 + O(h)] \equiv \sigma_k(h;\gamma), \; k = 1,2,\ldots$$

$$[\gamma^{k-1} w(\bar{\gamma};\gamma)]^{(m)} = d^m[\gamma^{k-1} w(\bar{\gamma};\gamma)]/d\gamma^m$$

Here we have defined ψ_i, σ_k ; we take $\sigma_0(h;\gamma) \equiv 1$.

The system of equations for the coefficients α,β obtained by setting $T(s_i) = 0$ and $T(s_{m+k}) = 0$ is called the HODIE equations:

$$\frac{1}{h^m}\ \alpha_i(h) = \Sigma_{j=1}^{J}\ \beta_j(h)\ \psi_i\ (h;\ \tau_j),\quad i = 0,\cdots,m \tag{3-2a}$$

$$\Sigma_{j=1}^{J}\ \sigma_k(h;\tau_j)\beta_j(h) = \begin{cases} 1 & k = 0 \\ 0 & k = 1,2,\cdots,K-1 \end{cases} \tag{3-2b}$$

Thus, the α's are determined once the β's are. (Below we consider various values of K in $1 \leq K \leq 2J$).

We first treat the special case that the differential operator L is equal to D^m. The coefficients for the HODIE approximation are obtained from (3-2) after setting $h = 0$.

The normalization $\Sigma\ \beta_j=1$ together with (3-1a) and (3-2a) give the usual difference approximation to L for the points $\bar{t}$

$$L_h\ u = \frac{1}{h^m}\ \Sigma_{i=0}^{m}\ \alpha_i(0)\ u(t_i) = m!\ \ u[t_0,\cdots,t_m]$$

where $u[t_0,\ldots,t_m]$ denotes the m-th divided difference of u.

Consider the space

$$F^m : \{u \mid u^{(m-1)} \text{ is absolutely continuous, } u^{(m)} \text{ is in } L_2 \text{ on } [0,\ 1]\}$$

Any $u \in F_m$ can be represented (by Taylor's Theorem) as

$$u(t) = \Sigma_{i=0}^{m-1}\ u^{(i)}\ (t_0)\ [t-t_0]^i/i! + \int_{t_o}^{t_m}\ g(t;x)u^{(m)}(x)dx$$

$$g(t;x) = (t-x)_+^{m-1}/(m-1)! = \begin{cases} (t-x)^m/(m-1)! & t \geq r \\ o & t < r \end{cases}$$

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

$$L_h\, u = m!\, u\,[t_0,\ldots,t_m] = \int_{t_0}^{t_m} B_m(x)\, u^{(m)}(x)\,dx$$

$$B_m(x) = g[t_0,\ldots,t_m;x]$$

where B_m is an $(m-1)$-st degree polynomial B-spline with joints at the mesh points $t_0,\ldots,t_m$; it satisfies (Curry and Schoenberg [3])

(3-3a) $$B_m(x) \begin{cases} >0 & x \in (t_0,t_m) \\ =0 & x \notin (t_0,t_m) \end{cases}$$

(3-3b) $$\int_{t_0}^{t_m} B_m(x)\,dx = 1$$

The truncation error is, therefore, given by

$$T_h[u] = L_h u - \Sigma\, \beta_j(0)\, Lu_j$$

$$= \int_{t_0}^{t_m} B_m(x)\, u^{(m)}(x)\,dx - \Sigma_{j=1}^{J}\, \beta_j(0)\, u^{(m)}(\tau_j)$$

$$\equiv E_h[u^{(m)}]$$

where we have defined E_h. Clearly, $E_h[v]$ is the quadrature error in approximating the integral of $B_m v$ by $\Sigma\, \beta_j v$. Take $v = \ell_j(\bar{\tau};\cdot) \in \mathbb{P}_{J-1}$, a Lagrange basis polynomial and set

(3-4) $$\beta_j(0) = \int_{t_0}^{t_m} B_m(x)\,\ell_j(\bar{\tau};x)\,dx \qquad j = 0,1,2,\ldots,J-1$$

The quadrature is then exact on $\mathbb{P}_{J-1}$ and the normalization $\Sigma\, \beta_j(0) = 1$ holds (take $v(t) = 1$ and see (3-10b)). Furthermore, this gives the only choice for $\beta_j(0)$ which makes E_h zero on $\mathbb{P}_{J-1}$.

Consequently, we have proved that for any $\bar{\tau} = \{\tau_1 < \tau_2 < \cdots < \tau_J\}$ the system (3-2a) with $K = J$ has a unique solution for the special case that $L = D^m$.

It follows that for any $\bar{\tau}$, there is a family of solutions of (3-2a) with K such that $1 \leq K < J$.

We see from (3-3a) that B_m is a positive weight function and, therefore, Gauss quadrature applies. Specifically, let $b_0, b_1, \ldots$ denote a sequence of polynomials, $b_k \in \mathbb{P}_k$, orthogonal with respect to the inner product

$$(u,v) = \int_{t_0}^{t_m} B_m(r)u(r)v(r)dr$$

then, as is well-known, b_k has k distinct real zeros in (t_0,t_m). We call these the B-spline orthogonal polynomials. Their zeros and the Gauss quadrature weights have been tabulated by Phillips and Hanson [8]. Consequently, the system (3-2a) has a solution for K up to $K = 2J$. In particular, by taking τ_j, $j = 1,\ldots,J$, to be the zeros of b_J, K is $2J$ and $\beta_j > 0$, $j = 1,\ldots,J$.

To determine the behavior of the truncation error as $h \to 0$, suppose the HODIE equations (3-2) are satisfied for some value of K, $1 \leq K \leq 2J$. Let $\bar{\xi} = \{\xi_1,\ldots,\xi_K\}$, $\xi_j = \tau_j$ if $j \leq J$, $\xi_j = \tau_{j-J}$ if $j > J$, and write $u^{(m)}$ as

$$u^{(m)}(t) = p_{K-1}(t) + u^{(m)}[\xi_1,\ldots,\xi_K,t]w(\bar{\xi};t)$$

where $p_{K-1} \in \mathbb{P}_{K-1}$ interpolates to $u^{(m)}$ at the distinct points in $\bar{\xi}$ and to $u^{(m+1)}$ at repeated points. Then

$$T_h[u] = E_h[u^{(m)}] = \int_{t_0}^{t_m} B_m(x)\, u^{(m)}[\xi_1,\ldots,\xi_K,x]w(\bar{\xi};x)dx$$

The integral exists if $u^{(m)} - p_{K-1}$ is integrable. We make the following assumption so that this is the case: $u \in F^{m+1}$ if $K \leq J$ and $u \in F^{m+2}$ if

$J < K \leq 2J$. Let $[v]$ denote the dimensions (like "inches" or "seconds", etc.) of v. Then $[w(\bar{\xi};x)] = [x]^K$ and (3-3b) shows that $[B_m] = [x]^{-1}$. Hence $[T_h[u]] = [u^{(m)}[\xi_1,\ldots,\xi_K,x]][x]^K$ which shows that $T_h[u] = O(h^K)$ with our assumption on u.

The above analysis has proved the following theorem which summarizes the results for the special case $L = D^m$.

THEOREM 3-1: Let $L = D^m$. There is a HODIE approximation, M_h to M which is exact on $\mathbb{P}_{m+K-1}$ for any K with $1 < K < 2J$. If $K \geq J$, then the approximation is unique. If $K > J$ and the approximation is not exact on $\mathbb{P}_{m+K}$, then $k = J-K$ of the points in $\bar{\tau}$ must be zeros of the B-spline orthogonal polynomial b_k. If $u \in F^{m+1}$ and $1 \leq K \leq J$, then the truncation error is order $O(h^k)$ and it is order $O(h^J)$ if $J < K \leq 2J$. If $u \in F^{m+2}$ and $J < K \leq 2J$, then the truncation error is order $O(h^k)$.

Various aspects of these results are illustrated by the five schemes given in the examples in Section 2: Scheme 1 uses $J = K = 1$ and it gives the lowest order, 1, because τ_1 has no special features. Scheme 2 also has $J = K = 1$, but $\tau_1 = t_1$ is the zero of b_1 and therefore the scheme has order $2J = 2$. Scheme 3 is one of a family of order $K = 2$ schemes which uses $J = 3$. Scheme 4 is one with $J = K = 3$; one ordinarily expects such a scheme to be of order $K = 3$, but $\tau_2 = t_1$ is the zero of b_1, so the order is 4. Scheme 5 has maximal order for $J = 3$; it requires the use of the three zeros of b_3 for the points in $\bar{\tau}$.

We now treat a general linear differential operator with continuous coefficients. Our next theorem is primarily a direct consequence of Theorem 3-1.

THEOREM 3-2: Let $L = D^m + \sum_{i=0}^{m-1} a_i D^i$, let $\bar{\tau}$ be any set of points with $\tau_j = \rho_j h$. If the a_i are continuous, then there exists an $h_0 \geq 0$ such that there is one and only one HODIE approximation M_h to M exact on $\mathbb{P}_{m+J-1}$ for all $h \in [0,h_0]$. Its coefficients are given as solutions of (3-5) with $K = J$. Moreover,

$$\frac{1}{h^m}\ \alpha_i\ (h) = \frac{1}{h^m}\ \alpha_i(0)\ [1 + O(h)]$$

$$\beta_j(h) = \beta_j(0) + O(h)$$

Proof: The existence and uniqueness follows from the same reasoning as for Theorem 3-1. Since the a_i are continuous so are the functions $\sigma_k(h;\gamma)$ whose values appear in the HODIE equations (3-2). This system is nonsingular for $h = 0$ and hence, by continuity, it is also nonsingular for h in a neighborhood of zero, i.e. for $h < h_0$. Thus for $h < h_0$ the previous argument applies.

With a_i continuous, ψ_i, σ_k in (3-4) can be written as

$$\psi_i(h;\gamma) = \frac{m!}{w'(\bar{t};t_i)}\ [1 + h\ \ \chi_i(h;\gamma]$$

$$\sigma_k(h;\gamma) = [\gamma^{k-1}\ w(\bar{\gamma};\gamma)]^{(m)}\ \ [1 + h\ \ \Theta_k(h;\gamma)]$$

where χ_i and Θ_k are continuous functions of h and γ. Then the final conclusion follows from the fact that the solution of a nonsingular linear system, (3-2a) with $K = J$, is a continuous function of its coefficients. ∎

The functions $\sigma_k(h;\gamma)$ can be written as linear combinations of functions of the form

(3-5) $$\psi_k(h;\lambda) = \lambda^k/k!\ (1 + \Psi_k(h;\lambda)),$$

where $\Psi\ (h;\lambda)$ is continuous and $O(h)$. That the system (3-2a) can then be solved

for K up to 2J requires the next result.

THEOREM 3-3: Let $\bar{\psi}(h)$ denote a set of functions as in (3-5). There exists an $h_0 > 0$ such that $\bar{\psi}(h)$ is a Chebyshev set on [0,1] for any $h \in [0,h_0]$.

To prove this, we show that if $\Sigma_{k=0}^{K} c_k \psi_k(h;\lambda_j) = 0$, then $c_k = 0$, $k = 0,\ldots,K$. It is easy to show that for any fixed $\bar{\lambda} = \{0 \leq \lambda_1 < \ldots < \lambda_k \leq 1\}$, $\bar{\psi}(h)$ is linearly independent on $\bar{\lambda}$ for $h \in [0,h_0]$; but, in addition, we must show that h_0 can be chosen independent of $\bar{\lambda}$.

Proof: Let $\bar{\lambda}$ be fixed and let V(h) denote the (K+1)-by-(K+1) matrix with elements $V(h)_{k,j} = \psi_{k-1}(h;\lambda_j)$ then V(0) is the transpose of a Vandermond matrix and non-singular. Hence, by continuity of the elements of V(h), there is an h_0, which depends on $\bar{\lambda}$, such that V(h) is non-singular for all $h \in [0,h_0]$. Hence $\bar{\psi}(h)$ is linearly independent on $\bar{\lambda}$ for $h \in [0,h_0]$.

Assume that the set $\bar{\psi}(h)$ is not a Chebyshev set for all h_0 and arbitrary $\bar{\lambda}$. Then, there are sequences,

$$\{h_n\}_{n=1}^{\infty}, \quad \{c_k(h_n)\}_{k=0,n=1}^{K,\infty}, \quad \max|c_k(h_n)| = 1, \quad \{\bar{\lambda}(h_n)\}_{n=1}^{\infty}$$

with $h_n \neq 0$ such that

$$P_n = \Sigma_{k=0}^{K} c_k(h_n)\, \psi_k(h_n;\cdot),$$

has a zero at each point in $\bar{\lambda}(h_n)$. There exists, therefore, a subsequence (which we also denote by $\{h_n\}$) such that

$$c_k(h_n) \to c_k^*, \quad \lambda_j(h_n) \to \lambda_j^*, \quad P_n \to P^*$$

By the form of ψ_k in (3-5), P^* is a polynomial of degree at most K. By continuity, $P^*(\lambda_j^*) = 0, j = 1,\ldots,K+1$. Since max $c_k^* = 1$, P^* is not identically equal to zero. Hence the K+1 points in $\overline{\lambda}^*$ are not distinct. Suppose

$$\lambda_j^* < \lambda_{j+1}^* = \lambda_{j+2}^* = \ldots = \lambda_{j+m}^* < \lambda_{j+m+1}^*$$

We can write

$$P_n(\lambda) = \pi_{k=1}^{m} (\lambda - \lambda_{j+k}(h_n))\, q(h_n;\lambda) \to (\lambda - \lambda_{j+1}^*)^m\, q(0;\lambda)$$

which shows that P^* has K+1 zeros counting multiplicities. This is a contradiction and establishes the theorem. ∎

COROLLARY 3-3: Let ψ_k, $k = 0,\ldots,2J-1$, denote functions as in (3-5) which have continuous first derivatives. Let v denote a differentiable function on $[0,h_0]$. If $h_0 > 0$ is sufficiently small, then there exists one and only one linear combination p of the ψ_k which satisfies the interpolation conditions $p(\tau_j) = u(\tau_j)$, $p'(\tau_j) = v'(\tau_j)$, $j = 1,\ldots,J$, for any set $0 \le \tau_1 < \ldots < \tau_J < h_0$

Proof: Let p_J denote a linear combination of the first J of the functions ψ_k. By Theorem 3-3, ψ_k, $k = 0,\ldots,J-1$ is a Chebyshev set. Therefore, there is a unique p_J which satisfies the interpolation conditions $p_J(\tau_j) = u(\tau_j)$. Similarly, ψ_k', $k = J,\ldots,2J-1$ is a Chebyshev set, so there is a unique linear combination of them which equals $u'(\tau_j) - p_J'(\tau_j)$ at τ_j, $j = 1,\ldots,J$. ∎

The application of Theorem 3-3 and its corollary to HODIE approximation follows from representation of moments of a Chebyshev set. This is discussed in detail in Chapter 2 of Karlin and Studden [5]. We summarize the pertinent information in the next paragraph.

Let μ denote any nondecreasing right continuous function of bounded variation on $[t_0,t_m]$. Let $\{\psi_0,\ldots,\psi_{2J-1}\}$ denote a Chebyshev set on $[t_0,t_m]$. The k-th moment, m_k, of the set with respect to μ is,

$$m_k = \int_{t_0}^{t_m} \psi_k(r)\, d\mu(r)$$

All of the moments can be represented in terms of J positive values, β_j, and J points $\tau_j \in [t_0,t_m]$ as

$$m_k = \sum_{j=1}^{J} \beta_j \psi_k(\tau_j) \quad , \quad k = 0,\ldots,2J-1$$

It is clear that this can be taken as the abstract theoretical basis for weighted Gauss quadrature in which case the points in are the zeros of ψ_k for $k = J+1,\ldots,2J-1$

For the special case $L = D^m$, the appropriate measure μ is given by $d\mu(r) = B_m(r)\, dr$, where B_m is the B-spline. As we now show, this same measure applies to the general operator L.

For fixed sets, $\bar{t},\bar{\tau}$, of points, the only values which enter a HODIE approximation are $u(t_i)$, $i = 0,\ldots,m$, and $Lu(\tau_j)$, $j = 1,\ldots,J$. These can be expressed in terms of a suitable polynomial. Let N denote the number of distinct points in $\bar{t} \cup \bar{\tau}$ so that $m+1 \le N \le m+1+J$. Define $q \in \mathbb{P}_{N+mJ-1}$ by

$$q(t_i) = 1, \quad i = 0,\ldots,m$$

$$q(\tau_j) = 1 = a_m(\tau_j), \quad m!\, q^{(k)}(\tau_j)/k! = a_{m-k}(\tau_j), \quad j = 1,..,J; k = 1,\ldots,m$$

then for any sufficiently smooth function u we have

$$L_n u = \sum_{i=0}^{m} \alpha_i\, u(t_i)/h^m = L_n(qu)$$

$$L\, u(\tau_j) = \sum_{i=0}^{m} a_i(\tau_j) u^{(j)}(\tau_j) = \sum_{i=0}^{m} \frac{m!}{i!}\, q^{(m-i)}(\tau_j) u^{(i)}(\tau_j) = (qu)^{(m)}(\tau_j)$$

Let $p_{m-1} \in \mathbb{P}_{m-1}$ interpolate to $u(t_i) = q(t_i)\, u(t_i)$, $i = 0,\ldots,m-2,m$, in which the point t_{m-1} has been omitted by design. We can write

$$qu(t) = p_{m-1}(t) + qu[t_0,\ldots,t_{m-2},t,t_m]\ (t-t_m)\ \pi_{i=0}^{m-2}\ (t - t_i)$$

and since $p_{m-1}^{(m)} = 0$, we have

$$L_h(u - p_{m-1}) = \frac{\alpha_{m-1}}{h^m}\ w'(\bar{t};t_{m-1})\ qu[t_0,\ldots,t_{m-2},t_{m-1},t_m]$$

$$= \frac{\alpha_{m-1}\ w'(\bar{t};t_{m-1})}{m!h^m}\ \int_{t_0}^{t_m} B_m(r)\ (qu)^{(m)}(r)dr$$

$$= \Sigma_{j=1}^{J}\ \beta_j (qu)^{(m)}(\tau_j) + E[(qu)^{(m)}] = \Sigma_{j=1}^{J}\ \beta_j Lu(\tau_j) + E[(qu)^{(m)}]$$

By Theorem 3-2 we have $\alpha_{m-1} = \alpha_{m-1}(h) = \alpha_{m-1}(0)+O(h)$ so, for sufficiently small h, $\alpha_{m-1} \neq 0$. Hence again we have weighted quadrature; the weight is negative if $\alpha_{m-1}\ w'(\bar{t};t_{m-1}) < 0$. Theorem 3-3 establishes that the set of functions $L\ t^{m+j}$, $j = 0,1,\ldots$, is a Chebyshev set for sufficiently small h. Hence the results given in Karlin and Studden [5] apply and there exist a set of points τ_j for which the quadrature is exact for any $u \in \mathbb{P}_{m+2J-1}$. These Gauss-type points are the zeros of orthogonal polynomials which call the <u>generalized B-spline orthogonal polynomials for L</u>. Note that they also depend on h. Their nature is discussed more explicitly in Theorem 3-5 below.

Thus, we can solve the HODIE equations (3-2a) for any K, $1 \le K \le 2J$ and the solution is unique for $J \le K \le 2J$. We can write $u = p_{m+K-1} + v$, $p_{m+K-1} \in \mathbb{P}_{m+K-1}$ with $v(t) = \pi_{i=0}^{m+K-1}(t - \xi_i)\ V(t)$; hence $(qu)^{(m)} = (qp_{m+K-1})^{(m)} + O(h^K)$. Note that pairs of points ξ_i, ξ_{i+1}, i odd, $\xi_{i+1} < \xi_{i+2}$ can be repeated provided that the coefficients of the differential operator are differentiable (see Corollary 3-3).

The analysis above establishes the next theorem which is the analog of Theorem 3-1 but for a general variable coefficient operator L.

THEOREM 3-4: If the coefficients a_i of L are continuous then for sufficiently small h there is a HODIE approximation M_h which is exact on P_{m+K-1} for any $1 \le K \le 2J$. The approximation is unique if $J \le K \le 2J$. Any set of points, $\bar{\tau}$, can be used for $K \le J$. For $J < K \le 2J$, if the approximation is not exact on $\mathbb{P}_{m+K}$, then $K-J$ of the points in $\bar{\tau}$ must be the $k = K-J$ zeros of b_k, the generalized B-spline orthogonal polynomials for L. If $K \le J$, and $u \in F^{m+1}$, then the truncation error is $O(h^K)$ and order $O(h^J)$ if $J < K \le 2K$. If the coefficients, a_i, have continuous first derivatives, $u \in F^{m+2}$ and $J < K \le 2J$, then the truncation error is $O(h^K)$.

We can also establish a relationship between the zeros of the generalized B-spline orthogonal polynomials for L and those of the B-spline orthogonal polynomials ($L=D^m$). The proof of the next theorem also gives some indication of the nature of these polynomials

THEOREM 3-5. Let $\tau_{k,j}(h)$, $j = 1,2,\ldots,k$ denote the zeros of the generalized B-spline polynomial b_k for L and $\tau_{k,j}(0)$ denote the zeros of the B-spline orthogonal polynomial for $L = D^m$. Then we have

$$\tau_{k,j}(h) = \tau_{k,j}(0) + O(h)$$

Proof: Elements of the orthogonal sequence b_k are given by

$$b_k(t) = \sum_{j=0}^{k-1} c_{k,j} b_j(t) + L\, t^{m+k} = \frac{(m+k)!}{k!} + \sum_{j=0}^{k-1} c_{k,j} b_j(t) + O(h)$$

By induction, the coefficients $c_{k,j}$ are, therefore, order $O(h)$ different from those for the special case $L = D^m$. Hence as h tends to zero, the zeros of b_k tend to $\tau_{k,j}(0)$ as stated since all the zeros are simple. ■

4. Discretization error for approximation with polynomial subspaces. We consider the problem $Lu = f$ on $[0,1]$ together with initial conditions or boundary conditions (2-1c). We assume that these conditions are linearly independent on the null space of L as well as on the space of polynomials, $\mathbb{P}_m$.

Let π denote a partition of $[0,1]$ with N points:

$$\pi : \quad 0 = t_0 < t_1 < \ldots < t_N = 1$$

Note that the t_i now are all points used to discretize L on $[0,1]$ rather than just those used for one difference approximation. Similarly we obtain a multiple sets of τ-points, one set for each difference approximation to L. For each set of $m+1$ successive mesh points,

$$t_n < t_{n+1} < \ldots < t_{n+m}$$

we select J points

$$t_n \leq \tau_{n,1} < \tau_{n,2} < \ldots < \tau_{n,J} \leq t_{n+m}$$

and construct a HODIE approximation exact on $\mathbb{P}_{m+K-1}$ (K as in Section 3). Its coefficients are denoted by $\alpha_{n,j}$, $\beta_{n,j}$. The solution u at mesh points is approximated by values of the solution of

$$(4\text{-}1) \qquad L_\pi U_n = \frac{1}{h_n^m} \sum_{i=0}^{m} \alpha_{n,i} U_{n+i} = \sum_{j=1}^{J} \beta_{n,j} f(\tau_{n,j})$$

$$\doteq F_n \quad , \quad n = 0,\ldots,N-m,$$

where

$$h_n = (-t_n + t_{n+m})/m$$

subject to appropriate initial or boundary conditions. Here we have defined the operator L_π and the function F. We have surpressed indicating the dependence of various quantities on the partition π.

Under mild restrictions on the mesh and the τ-points, we show that the operator L_π is stable. Define

$$||U_n||_\pi = \max \{U[t_n], U[t_n,t_{n+1}],\ldots, U[t_n,\ldots,t_{n+m-1}]\}$$

$$||F||_{\pi\infty} = \max_n |F_n|$$

$$||a_i||_\infty = \max_{t\ \varepsilon\ [0,1]} |a_i(t)|$$

$$h = \max_n h_n$$

THEOREM 4-1: Let the HODIE approximation be exact on $\mathbb{P}_{m+K-1}$, if the coefficients a_i of L are continuous, then for all sufficiently small h, there are constants K_1, K_2 which depend only on a_i and

$$\rho = \min_{n,j} (-\tau_{n,j} + \tau_{n,j+1})/h$$

such that the HODIE solution U_n of the initial value problem satisfies

$$||U_n||_\pi \leq e^{K,Nh} \{ ||U_0||_\pi + K_2|| L_\pi U||_{\pi\infty}\}$$

Proof: Let $p_n \varepsilon \mathbb{P}_m$ interpolate to U at $t_n, t_{n+1},\ldots,t_{n+m}$. Represent p_n as the Newton form of the interpolation polynomial to get

$$L_\pi U_n = L_\pi p_n = A_{n,0} U[t_n] + A_{n,1} U[t_n,t_{n+1}] +\ldots+ A_{n,m} U[t_n,\ldots,t_{n+m}]$$

Set $s_{n,0}(t) = 1$, $s_{n,k}(t) = \pi_{\ell=1}^{k} (t - t_{n+\ell})$, $k = 1,2,\ldots$ Then

$$A_{n,k} = \Sigma_{i=0}^{m} \alpha_{n,i}\, s_{n,k}(t_i)/h_n^m = \Sigma_{j=1}^{J} \beta_{n,j}\, L\, s_{n,k}(\tau_{n,j})$$

Because of the spacing of the points $\bar{t}$ and $\bar{\tau}$, it follows that

$$|D^{\ell} s_{n,k}(\tau_{n,j})| \leq k!\,(mh)^{k-\ell}/(k-\ell)!$$

Consequently

$$(4\text{-}2a)\quad |A_{n,k}| \leq \max_j |\beta_{n,j}|\, \Sigma_{\ell=0}^{k} ||a_\ell||_\infty\, k!\,(mh)^{k-\ell}/(k-\ell)!$$

By Theorem 3-2, $\beta_{n,j}$ differs by $0(h_n)$ from the β given in (3-4) for the special case that $L = D^m$. It follows directly from the properties of the B-spline and the spacing of the points that

$$(4\text{-}2b)\quad |\beta_{n,j}| \leq \max_{t\,\epsilon\,(t_n,t_{n+m}]} \left|\frac{\pi_{k=1,k\neq j}^{J}(t-\tau_{n,k})}{\pi_{k=1,k\neq j}^{J}(\tau_{n,j}-\tau_{n,k})}\right| + 0(h) < \left(\frac{m}{\rho}\right)^{J-1} + 0(h)$$

where ρ is as in the theorem. Because of the normalization $\Sigma\, \beta_i = 1$ we have, therefore,

$$m! - (m/\rho)^{J-1}\, mh\, ||a_{m-1}||_\infty\, m! + 0(h)$$

$$\leq A_{n,m} \leq$$

$$m! + (m/\rho)^{J-1}\, mh\, ||a_{m-1}||_\infty\, m! + 0(h)$$

Consequently, if h is sufficiently small, then $A_{n,m}$ is close to $m!$ and positive.

We first treat the homogeneous equation, $L_\pi\ U_n = 0$. This can be written as

$$U[t_{n+1},\ldots,t_{n+m}] = U[t_n,\ldots,t_{n+m-1}]$$

$$- (-t_n + t_{n+m}) \{A_{n,m-1}\, U[t_n,\ldots,t_{n+m-1}] + \ldots + A_{n,0}\, U[t_n]\}/A_{n,m}$$

Choose

$$K_1 \geq \max \{m,\ \max_n\{m/A_{n,m}\},\ \max_{n,i}\{m|A_{n,i}|/A_{n,m}\}\}$$

which is seen to be bounded independent of h because of the bounds (4-2a) and (4-2b). We then obtain

$$|U[t_{n+1},\ldots,t_{n+m}]| < (1 + K_1\, h)\ ||U_n||_\pi$$

and

$$|U[t_{n+1},\ldots,t_{n+i}]| = |U[t_n,\ldots,t_{n+i-1}] + (-t_n + t_{n+i})\ U[t_n,\ldots,t_{n+i}]|$$

$$\leq (1 + K,\ h)\ ||U_n||_\pi$$

Consequently we obtain the following bound on any solution of the homogeneous equation:

$$(4\text{-}3) \qquad ||U_{n+1}||_\pi \leq (1 + K_1\, h)\ ||U_n|| \leq (1 + K_1\, h)^{n+1}\ ||U_0||_\pi$$

$$\leq e^{K,hN}\ ||U_0||_\pi$$

which is the desired bound in the special case that $L_\pi\ U_n = 0$

The solution of $L_\pi\ U_n = F_n$, subject to homogeneous initial conditions, $U[t_0] = U[t_0,t_1] = \ldots = U[t_0,\ldots,t_{m-1}] = 0$, is given by

$$(4\text{-}4) \qquad U_{n+m} = \Sigma_{n=0}^{m}\ G_{n+m,k}\ F_k h_k$$

where G is the Green's function for L_π subject to homogeneous initial conditions. Thus, G satisfies

$$G_{n,k} = V_n :$$

$$V[t_{k-1}] = V[t_k] = V[t_k, t_{n+1}] = \ldots = V[t_k, \ldots, t_{k+m-2}] = 0$$

$$V[t_k, \ldots, t_{n+m-1}] = m/A_{k,m}$$

$$L_\pi V_n = 0, \quad n \neq k$$

Consequently, if follows from the derivation of (4-3) that G is bounded according to

$$||G_{n+m,k}||_\pi \leq K_1 e^{K_1 h(N-k)}, \quad k = 0,1,\ldots,N-m$$

Hence from (4-4) we obtain

$$|U_{n+m}| \leq K_1 ||F||_\infty \Sigma_{k=0}^{m} e^{K_1 h(N-k)} h$$

$$\leq K_1 ||F||_\infty e^{K_1 h(N+1)} \int_0^\infty e^{-K_1 x} dx$$

$$= ||F||_\infty e^{K_1 h(N+1)}$$

Similarly, from (4-4) we have

$$|U[t_{n+m}, \ldots, t_{n+m+i}]| = |\Sigma_{k=0}^{m} G[t_{n+m}, \ldots, t_{n+m+i}]_k F_k h_k|$$

$$\leq ||F||_\infty \Sigma_{k=0}^{m} |G[t_{n+m}, \ldots, t_{n+m+i}]_k| h_k$$

and, therefore, the bound given in the theorem. ∎

Bounds on solutions of (4-1) subject to other boundary conditions can be obtained from the bound on the initial value problem. The general solution of (4-1) is

$$U_n = g_{n,0}\, U[t_0] + g_{n,1}\, U[t_0,t_1] + \ldots$$

$$+ g_{n,m-1}\, U[t_0,\ldots,t_{m-1}] + \Sigma_{k=0}^{n-m}\ G_{n,k}\, F_k\, h_k$$

where the g's are solutions of homogeneous problem and G is the Green's function for the initial value problem. This can be used to obtain expressions for $U[t_N]$, $U[t_{N-1},t_N],\ldots,$ $U[t_{N-m+1},\ldots,t_N]$. These can be substituted into boundary conditions of the form

$$(4\text{-}5)\qquad \Sigma_{k=0}^{m-1}\ B_{\ell,0,k}\ U[t_0,\ldots,t_k] + B_{\ell,1,k}\, U[t_{N-i},\ldots,t_N] = c_\ell,\quad \ell = 0,\ldots,m-1$$

to obtain an algebraic system for $U[t_0],\ldots,U[t_0,\ldots,t_{m-1}]$. These values are bounded provided the B's are bounded. We assume the system has a unique solution.

To approximate conditions of the form

$$(4\text{-}6)\qquad \Sigma_{k=0}^{m-1}\ b_{\ell,0,k}\, D^\ell u(0) + b_{\ell,1,k}\, D^\ell u(1) = c_\ell$$

we use

$$L_\pi^{\ell,0}\ U_0 + L_\pi^{\ell,1}\, U_N = c_\ell$$

where, with I such that $b_{\ell,0,k} = 0$ for $k > I$,

$$L_\pi^{\ell,0}\ U_0 = \frac{1}{h_0^m}\ (\Sigma_{i=0}^{I}\ \alpha_{0,i}^{\ell}\ U_i + \Sigma_{i=I+1}^{m}\ h^{i-I}\ \alpha_{0,i}\, U[t_0,\ldots,t_{i-I}])$$

$$= \Sigma_{j=1}^{J}\ B_{o,j}^{\ell}\ f(\tau_{0,j}^{\ell})$$

is an exact approximation on $\mathbb{P}_{m+K-1}$ to

$$\Sigma_{k=0}^{m-1} b_{\ell,0,k} D^{\ell}u(0) = \Sigma_{k=0}^{I} b_{\ell,0,k} D^{\ell}u(0)$$

for u such that $Lu = f$ and similarly for $L_{\pi}^{\ell,1}$

An analysis similar to the proof of Theorem 4-1 shows that the B's in (4-5) are then bounded like the A's in the proof, and one obtains

THEOREM 4-2. Consider $Lu = f$ with general boundary conditions (4-6). Let L_{π_N} be exact on $\mathbb{P}_{m+k-1}$ and assume the discretized boundary conditions (4-4) are linearly independent over the solutions of the discrete initial value problem. Then, with the terminology of Theorem 4-1, we have

$$||U_n|| \leq e^{K_1 hN} (K_2 \max_k |c_k| + K_3 ||F||_{\infty})$$

Let $E = u - U$ denote the discretization error. We have

$$L_{\pi} E_n = L_{\pi} u_n - F_u = L_{\pi} u_n - \Sigma_{j=1}^{J} \beta_{n,j} f(\tau_{n,j})$$

$$= L_{\pi} u_n - \Sigma_{j=1}^{J} Lu(\tau_{n,j}) = T_{\pi}[u]_n$$

where T_{π} is the truncation operator. Also

$$\Sigma_{k=0}^{m-1} B_{\ell,0,k} E[t_0,\ldots,t_k] + B_{\ell,1,k} E[t_{N-i},\ldots,t_N]$$

$$= L_{\pi}^{\ell,0} u_0 + L_{\pi}^{\ell,1} u_N - \Sigma_{j=1}^{J} \{\beta_{0,j}^{\ell} Lu(\tau_{0,j}^{d}) + \beta_{1,j}^{\ell} Lu(\tau_{n,j}^{\ell})\}$$

$$= T_{\pi}^{\ell,0}[u] + T_{\pi}^{\ell,1}[u]$$

Consequently, we have the following result for the general operator L and boundary conditions (2-1):

THEOREM 4-3: *In the notation of Theorem 4-1; if Nh is bounded as $h \to 0$ and if ρ is positive and bounded away from zero as $h \to 0$, then the discretization error for a HODIE approximation exact on $\mathbf{P}_{m+K-1}$ is $0(h^p)$ where p is the order of the truncation error.*

5. Computational analysis. In this section we consider the computational aspects of the HODIE method. Here we discuss specific features which were incorporated into our implementation and compare the amount of work with other available methods.

We restrict our discussion to the case of a second order equation subject to Dirichlet conditions for four reasons: it is simple, it is the most important, it is readily generalized, and there are detailed analyses of other methods available for comparison. Thus we have the problems

$$u(0) = A,\ u(1) = B, \quad Lu(t) = f(t), \quad 0 < t < 1$$

$$Lu(t) = a_2(t)\, u''(t) + a_1\, u'(t) + a_0(t)u(t), \quad a_2 > 0,$$

$$U_0 = A,\ U_N = B, \quad L_\pi\, U_n = I_\pi\, f_n\ , \quad n = 0,1,\ldots, N-2$$

$$L_\pi\ U_n = [\alpha_{n,0}\, U_n + \alpha_{n,1}\, U_{n+1} + \alpha_{n,2}\, U_{n+2}]/h_n^2. \quad h_n = (-t_n + t_{n+2})/2$$

$$I_\pi\ f_n = \Sigma_{j=1}^{J}\, \beta_{n,j}\, f(\tau_{n,j})$$

where, for generality, we have taken the coefficient of D^2 in L to be a positive function, a_2, rather than unity.

We consider the two most interesting choices of the τ-points to be

$$\underline{\text{Regular}}: \quad \tau_{n,j} = t_n + (j-1)\, h_n,$$

$$\underline{\text{Gauss-type}}: \quad \tau_{n,j} \text{ such that } b_J(\tau_{n,j}) = 0$$

where the Regular τ-points are equally spaced and the Gauss-type τ-points are the zeros of b_J, the appropriate generalized B-spline orthogonal polynomial for L.

The computation in an implementation of a specific high accuracy 3-point HODIE approximation consists of two distinct parts. The first is the solution, for each interior mesh point, of the HODIE equations, (3-2), for the coefficients α's and β's. The second part involves the solution of an (N-1)-by-(N-1) tridiagonal linear system. The HODIE equations as presented in (3-2) are reducible: that is, one solves a J x J system for the β's and then a 3 x 3 system for the α's. This reducibility results in a significant saving of work both in general and for the special case of m = 2 considered in this section.

Although the Lagrange basis is convenient for theoretical analysis, we have found that it is computationally more efficient to use a different basis.

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

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

$$p_0(t) = 1, \quad p_1(t) = (t - t_{n+1}), \quad p_2(t) = (t - t_{n+1})^2$$

$$p_3(t) = (t - t_n)\, p_2(t), \quad p_4(t) = (t - t_n)^2\, p_2(t), \quad p_5(t) = (t - t_{n+2})\, p_3(t)$$

$$p_6(t) = (t - t_{n+2})^2\, p_3(t), \quad \text{and so on}$$

With $s_1 = - t_n + t_{n+1}$, $s_2 = - t_{n+1} + t_{n+2}$, this choice leads to the following system for $\alpha_{n,i}/h_n^2 = \eta_{n,i}$

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

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

$$\delta_1\delta_2\eta_{n,1} = \Sigma_j \beta_{n,j}[2a_2(\tau_{n,j}) + 2(\tau_{n,j}-t_{n+1})a_1(\tau_{n,j}) + (\tau_{n,j}-t_n)(\tau_{n,j}-t_{n+2})a_0(\tau_{n,j})]$$

One equation can be eliminated from the J equations for the β's by replacing the normalization equation with $\beta_{n,1} = 1$. The remaining equations are

$$\Sigma_{j=2}^{J} \sigma_{k,j} \beta_j = -\sigma_{k,1}, \quad k = 1,\ldots,J-1$$

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

The choice of the basis functions makes the evaluation of the coefficients simple and it also gives a structure to the system which makes it easy to solve. Specifically, for the Regular case, three of the τ's are at mesh points. Arranging the system so that its first three columns correspond to t_{n+1}, t_n, and t_{n+2}, one finds that these columns have the special form

$-h_1h_2a_1(t_{n+1})$	$-6h_1a_2(t_n)+2h_1h_2a_1(t_n)$	$6h_2a_2(t_{n+2})+2h_1h_2a_1(t_{n+2})$
$2h_1h_2a_2(t_{n+1})$	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 non-zero elements. This, of course, is very advantageous for solving for the $\beta_{n,j}$ in the Regular case.

We first consider the computational effort required for an uniform partition, $t_n = nh$, $n = 0,\ldots,N-1$. We measure the effort in terms of the number, F, of function evaluations (a_2,a_1,a_0, or f) required 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 multiplication. Table 5-1 presents the effort required for some cases of interest.

Computation Step	Regular Case J=3	J=5	J=7	J=9	Gauss-type Case J=2	J=3	J=4	J=5
The β-matrix entries	8	39	89	137	6	14	36	50
Solve the β-matrix	3	17	47	111	1	7	38	47
Right side of the α-matrix	13	21	33	43	12	18	24	30
Solve the α-matrix	3	3	3	3	3	3	3	3
Solve the tridiagonal matrix	7	9	11	13	6	7	8	9
TOTALS Multiplies	34M	89M	183M	307M	28M	49M	109M	139M
Function Evaluations	4F	12F	20F	28F	8F	12F	16F	20F

Table 5-1 Breakdown of the multiplications and function evaluations required per interior mesh point for two cases of the HODIE method of orders 4, 6, 8 and 10. A uniform partition is assumed.

The β-matrix entries are found from a simple examination and assuming that the factors $s''(\tau_{n,j})$, $s'(\tau_{n,j})$, $s(\tau_{n,j})$ of (5-4) have been previously computed and stored (these values are independent of n since a uniform partition is assumed). The special structure of this matrix for the Regular case is assumed for estimating

the work to solve the β-matrix equations. For the Gauss-type case we have a general (J-1)x(J-1) system to solve. Note that we assume the Gauss-type points $\tau_{n,j}$ have been previously computed or are otherwise known. The right sides of the α-equations are of a special form and the computation is made by forming $\beta_{n,j}a_{\ell}(\tau_{n,j})$ and then combining these appropriately. The solution of the α-equations is trivial and the final multiplications occur in solving the large tridiagonal system (4M) plus evaluating its right side (J multiplies). In the regular case the function evaluation at the t_i points are used more than once for a resulting saving in computation.

We now use these work estimates to compare roughly the work of the HODIE method with others. The comparison is presented in Table 5.2 for seven methods, three different orders of accuracy (4, 6 and 8) and both uniform and nonuniform partitions. The data for three methods (collocation by Hermite piecewise polynomials, least squares by splines and discrete-Ritz) are derived from Russell and Varah [12], where they are described in detail. We have had to modify the multiplication counts in order to account for the slightly different differential equation used here and to rationalize the effect of the E_L term used there. Note that the discrete Ritz method is limited to self-adjoint problems and hence is not strictly comparable to the other methods included here. The other two methods (collocation by splines and extrapolation of the trapezoid rule) are analyzed in detail by Russell [10] and we have adapted his results for our particular equation. Russell also considers collocation with Hermite cubics and quintics in detail. One must emphasize that the exact values of these counts depend on small details of the implementation of an algorithm and one can trade multiplications for additions, etc., in some instances.

The changes for collocation, least squares and discrete Ritz from the equally to non-equally spaced computation 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 (4-1) in each interval. We have assumed that two more multiplications are needed for each entry in 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 6-3.)

Considerable caution should be taken in attaching importance to the specific numbers in Table 5-2. These are 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 comparisons for partial differential equations indicate that the work for HODIE is significantly less than for other available

methods. Theoretical and experimental results have been obtained for the partial differential equation case and these are being prepared for publication. Preliminary results are given in Lynch and Rice [7].

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	187M+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 - Trapezoidal 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 5-2 Summary of multiplication (M) counts and function evaluations (F) counts 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.

6. Experimental results. We present support for the following points:

I. The HODIE method converges as predicted by theory; there are no unforeseen numerical complications.

II. There are no unforeseen difficulties or complexities in implementation.

III. 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 be in order to minimize computation time.

IV. The use of Gauss-type τ-points gives the rate of convergence predicted by theory.

V. The use of Gauss-type τ-points for the operator D^2 improves the rate of convergence for a general operator L over that expected for a general set of τ-points.

The first two points must be supported for any new method; the third point applies to collections of methods with varying orders; and the last two points apply to HODIE 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 considerable 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 of solving this class of problems. However, we quickly found that in order to verify the rates of convergence for very high order HODIE methods, we had to use very high precision (if one has a 1% error with $N = 2$ with an $O(h^{14})$ method, then the error is $1/4 \times 10^{12}$ times smaller when $N = 16$).

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 CDC 6500 with double precision arithmetic. A double precision floating point arithmetic operation has relative error of about one part in 10^{28}.

In each of the experiments, the domain of the problem (typically [0,1] or [0,5]) was partitioned by an equal-spaced mesh with N subintervals, so the mesh spacing, h, was proportional to $1/N$.

Example 6-1: The very simple problem

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

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

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

has been used by Russell and Shampine [11], de Boor and Swartz [4], and others.

Figure 6-1 summarizes one set of experimental results. The logarithm of the maximum error is plotted versus the logarithm of the number of subdivisions for eleven different sets of 5 τ-points. We now describe the various curves in this figure and give our interpretation of the results.

a. The topmost curve gives the results when 5 Regular τ-points (equal spaced τ-points) were used. One expects at least $0(h^5)$ rate of convergence with a set of 5 τ-points (because the approximation is locally exact on $\mathbb{P}_7$). The curve shows a very consistent $0(h^6)$ rate of convergence. The central τ-point is the central mesh point of the three point difference operator and it is clear from the symmetry of the differential operator that this τ-point is a zero of the linear B-spline orthogonal polynomial for the differential operator. Consequently, one expects $0(h^6)$ from this set of τ-points.

b. There is a set of nine curves in Figure 6-1 which have sharp downward spikes at $N = 4,8,16,25,32,50,64,100,$ and 200, respectively. The set of 5 τ-points used for each one of these curves is the set of 5 Gauss points for that value of N at which the peak of the spike occurs. One has different sets of τ-points because their locations depends on $h = 1/N$. The curve with the 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 τ-points are not the Gauss points, hence one expects only $0(h^6)$ [one of the points is the central mesh point of the three point difference operator], and this behavior can be seen for large values of N, say N greater than about 16.

c. Consider the tips of the spikes from this collection of nine curves. If one joins the tips, one sees a very consistent $0(h^{10})$ 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. The maximum error at

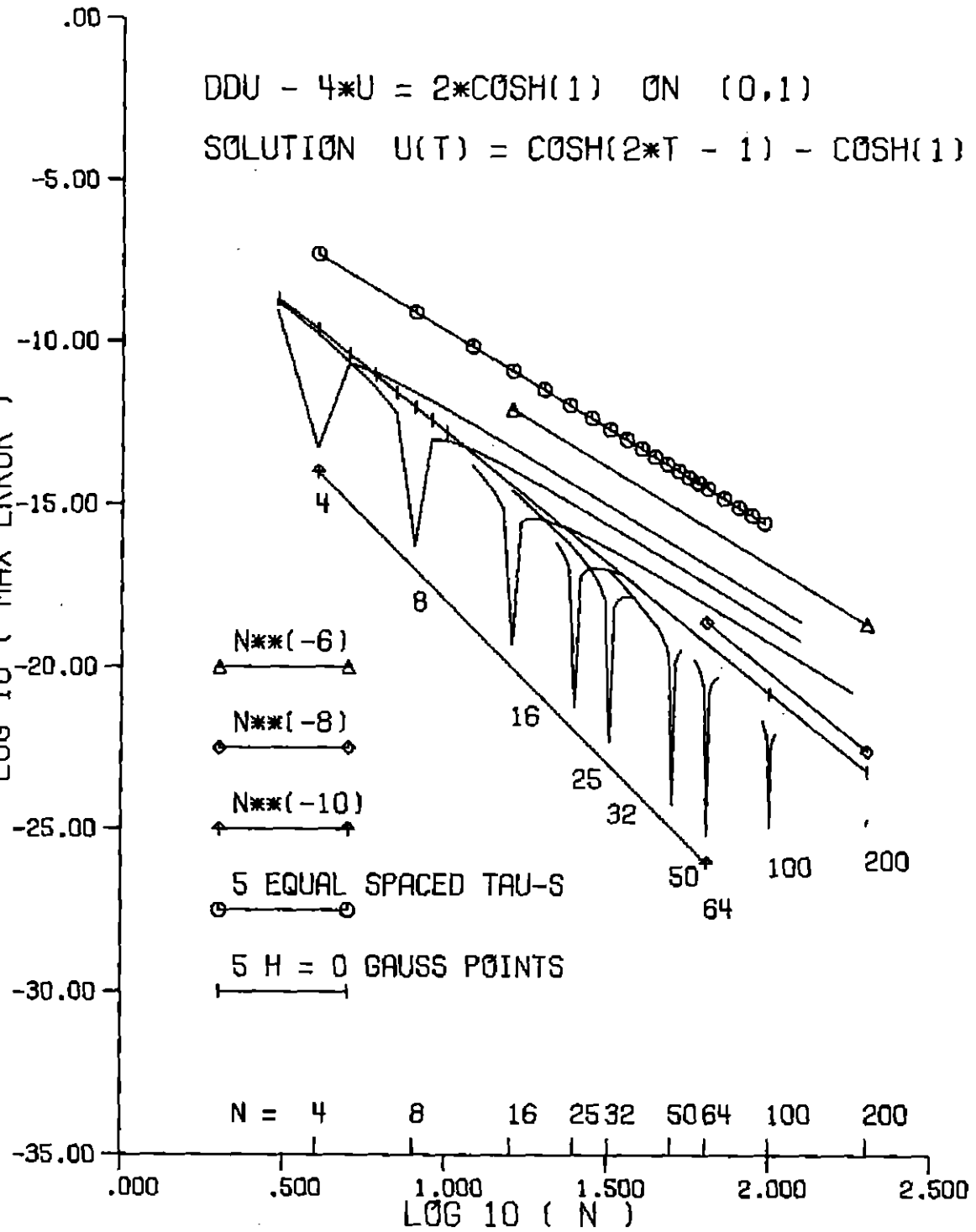

Figure 6-1: Behavior of the error as a function of number N of subintervals for eleven different 5 τ-point HODIE schemes for Example 6-1.

$N = 64$ is about 10^{-25} and this $O(h^{10})$ behavior breaks down beyond this because of round-off. We used values of the τ-points accurate only to about one part in 10^{15} (single precision).

d. The last curve is the one for 5 Gauss-type points for the operator D^2. One of these is the central mesh point of the operator, hence one expects at least $O(h^6)$. However, a very consistent $O(h^8)$ rate of convergence is observed. Recall that the Gauss points of this operator differ by $O(h)$ from those of the operator D^2 (see Theorem 3-4). Hence one expects improvement over an arbitrary set of τ-points which contain the central mesh point of the difference operator.

Example 6-2. Typical of a fairly difficult problem is, for $t_0 = 0.36388$,

$$\frac{d[(.01 + 100(t-t_0))]\ du/dt]}{dt} = -2[1 + 100(t - t_0)(\tan^{-1}100(t - t_0) + \tan^{-1}100\ t_o)]$$

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

$$u(t) = (1 - t)\ (\tan^{-1}\ 100(t - t_0) + \tan^{-1}\ 100\ t_0)$$

which was taken from Rachford and Wheeler [9]. The solution has a very sharp rise near $t = 0.36$: it increases from 0.1 at $t = 0.3$ to 1.7 at $t = 0.4$ and then it falls nearly linearly to 0 at $t = 1$ (see [9] for a graph of the solution).

Results for two sets of τ-points are shown in Figure 6-2: the 3-point Regular case--which is the $O(h^4)$ Størmer-Numerov scheme--and the 7 τ-point Gauss-type set for the operator D^2. One sees that there is considerable irregularity for N up to about 100 and then for larger N, the error decreases smoothly at the rates of $O(h^4)$ and $O(h^{10})$ respectively. For a general set

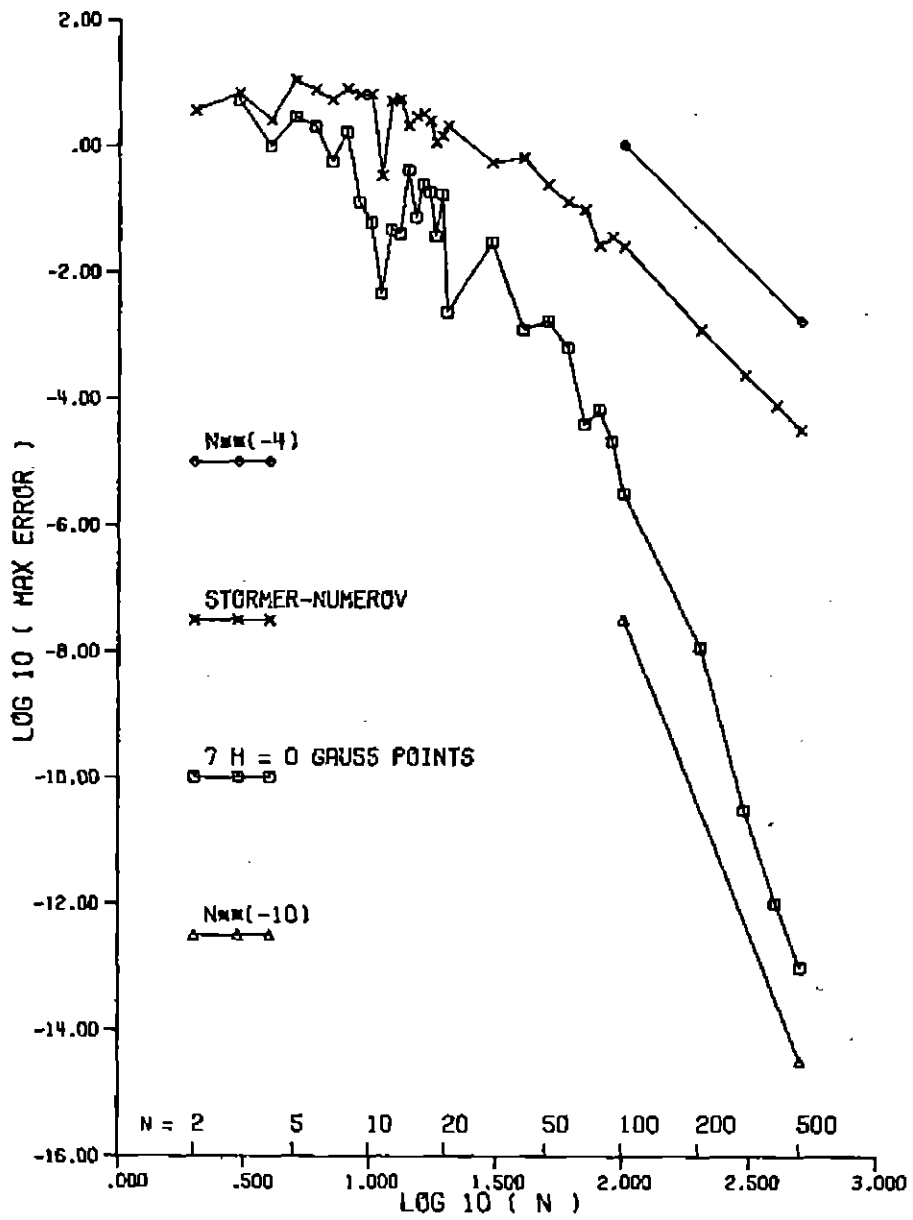

Figure 6-2: Behavior of the error as a function of the number N of subintervals for two HODIE schemes, one the Størmer-Numerov scheme, and one with 7 Gauss-type τ-points for the operator D^2 for Example 6-2.

of 7 τ-points, one expects $O(h^7)$, but use of the D^2 Gauss-points improves, as expected, the rate of convergence.

To compare efficiency, we note that the Størmer-Numerov with N = 300 required almost exactly the same amount of computation time as the 7-point scheme with 100 points. The Størmer-Numerov scheme achieved a maximum error 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 6-3: The relationship between work, order, and accuracy is seen in more detail for the problem

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

$$u(0) = 0,\quad u(5) = \sin(25)$$

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

Note that there are several oscillations of the solution as t ranges from 0 to 5.

We solved this problem with a wide variety of HODIE schemes and Figure 6-3 summarizes the results for a selection of them. 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 again 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.

Figure 6-4 displays results from a second set of experiments for the problem given above. It illustrates the strong effect of the Gauss-type points as well as the sensitivity of the error to small changes from these points. For several values of N between 25 and 400 a number of cases were run for different sets of 3 τ-points: $\tau_j = t_{n+1} + (j-2)\rho$, $j = 1,2,3$. By symmetry of the differential operator, there is a unique value of ρ which make these point Gauss-type points. Note the large downward spike at that value of ρ.

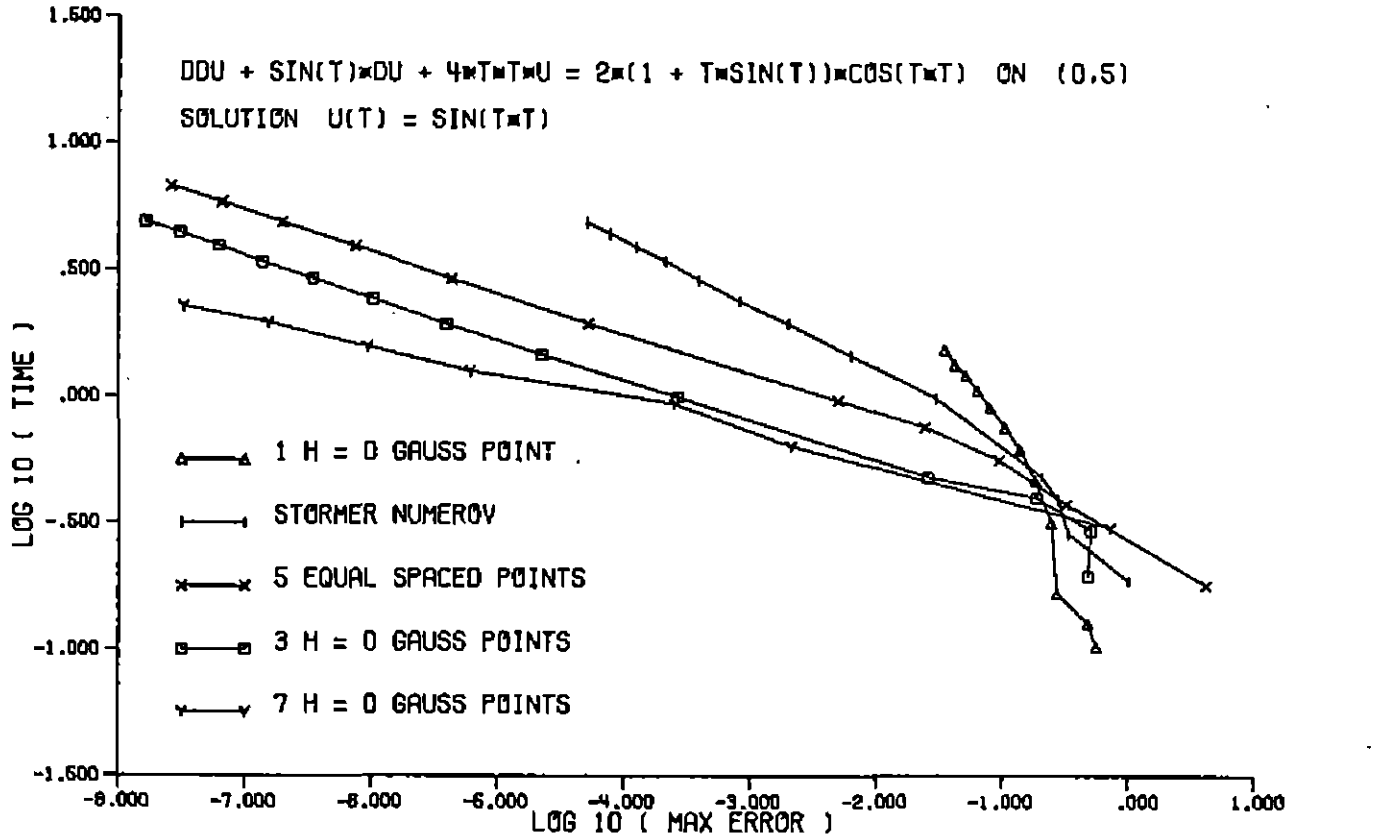

Figure 6-3: Illustration of the relationship between work (execution time), accuracy achieved, and order of the HODIE method for Example 6-3.

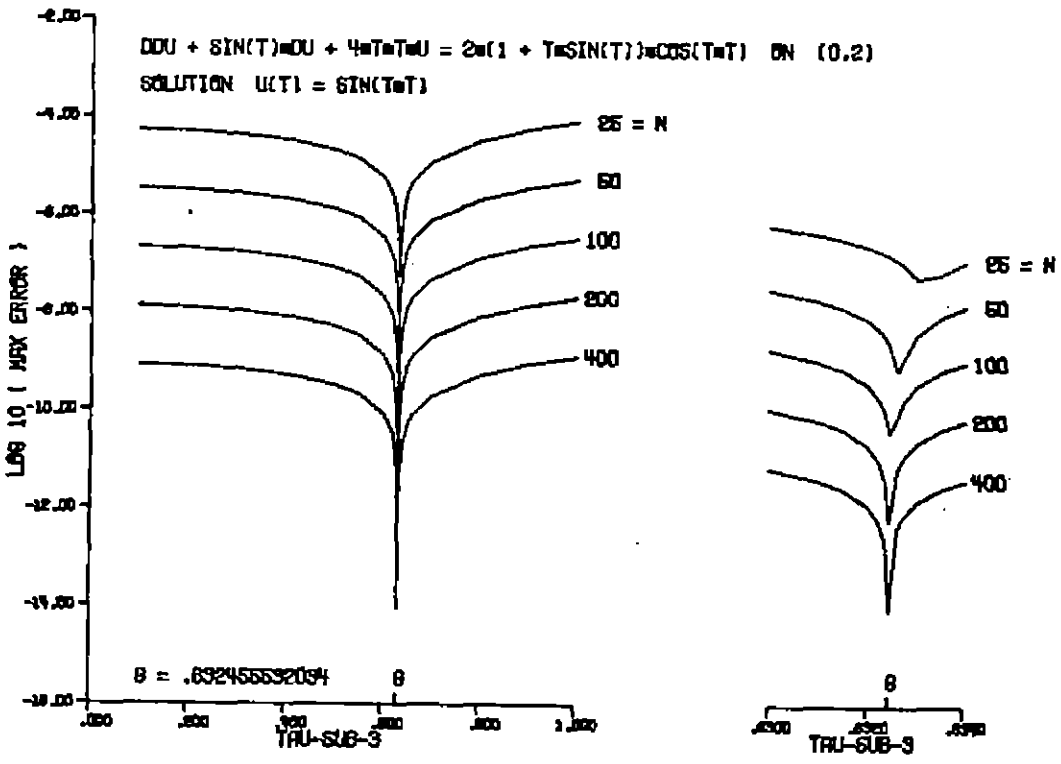

Figure 6-4: Behavior of the error as a function of a τ-point variation for Example 6-2. Three τ-points are used with $\tau_{n,j} = t_{n+1} + (j-2)\rho h$, $j = 1,2,3$. The value of ρ is varied, $\rho = G$ gives the Gauss-type points for $h = 0$. Note the dependence on $h = 2/N$.

REFERENCES

1. Birkhoff, G. and Gulati, S., Optimal few-point discretizations of linear source problems, SIAM J. Numer. Anal. 11 (1974) 700-728.

2. Collatz, L., The numerical treatment of differential equations. 3rd Edit. Springer-Verlag, Berlin, (1966).

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

4. de Boor, C.W. and Swartz, B., Collocation at Gaussion points, SIAM J. Num. Anal., 10 (1973) 582-606.

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

6. Keller, H.B., Numerical solution of boundary value problems for ordinary differential equations: Survey and some recent results on difference methods, in "Numerical Solution of Boundary Value Problems for Ordinary Differential Equations" (A.K. Aziz, ed.) Academic Press, New York, (1975) 27-88.

7. Lynch, R.E. and Rice, J.R. The HODIE method: A brief introduction with summary of computational properties, CSD-TR 170, Dept. Computer Science, Purdue University, Nov. 18, 1975, 12 pages.

8. Phillips, J.L. and Hanson, R.J., Gauss quadrature rules with B-spline weight function, Math. Comp., 28 (1974) 666 and microfiche supplement (32 pages).

9. Rachford, H.H. and Wheeler, M.F., An H^{-1} Galerkin procedure for the two-point boundary value problem, in "Mathematical Aspects of Finite Elements in Partial Differential Equations" (C.W. de Boor, ed.) Academic Press, New York, (1974) 253-382.

10. Russell, R.D. A comparison of collocation and finite differences for two-point boundary value problems, SIAM J. Numer. Anal., to appear.

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

12. Russell, R.D. and Varah, J.M., A comparison of global methods for linear two-point boundary value problems, Math. Comp. 29 (1975) 1007-1019.

13. Young, D.M. and Dauwalder, J.H., Discrete representations of partial differential operators, in "Errors in digital computation 2" (L.B. Rall, ed.) John Wiley and Sons, Inc. New York 181-217 (1965).