A Comparison of Global Methods for Linear Two-Point Boundary Value Problems

R. D. Russell
Mathematics Dept.
Simon Fraser University
Burnaby, B. C.

and

J. M. Varah
Computer Science Dept.
University of British Columbia
Vancouver, B. C.

The work of the first author was supported by NRC(Canada) grant #A8781
The work of the second author was supported by NRC(Canada) grant #A8240

I. Introduction

Consider the $(2m)^{-th}$ order linear two-point boundary value problem

$$Lu(x) = \sum_{i=0}^{m} (-1)^i D^i(a_i(x)D^i u) = f(x), \quad a \le x \le b \tag{1.1}$$

$$D^i u(a) = D^i u(b) = 0, \quad 0 \le i \le m-1 . \tag{1.2}$$

Three well-known methods which give global continuous approximate solutions to this problem are the methods of collocation, Galerkin, and least squares. In this paper we shall relate and compare these methods from the point of view of practical machine computation.

First, in Section II, we show how these methods are related in general, using arbitrary basis functions to determine the finite-dimensional subspace in which the approximate solutions are constructed. Then in Section III, we relate the available error estimates for the common choice of piecewise polynomial bases. Finally in Section IV, we compare the amount of work required to compute these solutions by forming and solving the relevant matrix equations.

II. Description of the Methods

All the methods we consider find approximate solutions of the form

$$\sum_{1}^{N} c_i \phi_i(x),$$

i.e. the solutions are elements of a finite-dimensional subspace

$$S_n = \text{span } \{\phi_1(x), \ldots, \phi_N(x)\}$$

whose basis elements all satisfy the boundary conditions (1.2). The methods only differ in the way the coefficients $\{c_i\}$ are chosen.

1. collocation

Here the approximate solution $w^{(N)}(x) = \sum_{1}^{N} w_j \phi_j(x)$ is

determined by satisfying (1.1) exactly at N points, i.e.

$$L\, w^{(N)}(x_i) = f(x_i) \quad , \; i=1,\ldots,N. \tag{2.1}$$

The resulting linear system to solve for the coefficients $\{w_j\}$ is

$$\underline{Cw}=\underline{f} \ , \ c_{ij}=L\phi_j(x_i) \ , \ f_i=f(x_i) \ . \tag{2.2}$$

2. Galerkin

The Galerkin solution $u^{(N)}(x) = \sum_1^N u_j\phi_j(x)$ is determined by forcing the residual $(Lu^{(N)}-f)$ to be orthogonal to each basis function:

$$\int_a^b Lu^{(N)}(x)\phi_i(x)dx = \int_a^b f(x)\phi_i(x)dx \ , \ i=1,\ldots,N. \tag{2.3}$$

This gives the linear system

$$\underline{Au}=\underline{g} \ , \ a_{ij}=\int_a^b (L\phi_j)\ \phi_i \ dx, \ g_i =\int_a^b f\phi_i \ dx \quad . \tag{2.4}$$

Since integration by parts gives

$$a_{ij} = \int_a^b (L\phi_j)\phi_i \ dx = \int_a^b M(\phi_j,\phi_i) \ dx \tag{2.5}$$

where

$$M(u,v) = \sum_{i=0}^m a_i(x) \ D^iu \ D^iv \ ,$$

the Galerkin solution is equivalent to the so-called Ritz solution derived from the variational principle for (1.1),(1.2). This also shows that the matrix A is symmetric, and in fact it is positive definite when the operator in (1.1) is elliptic.

Computationally of course, these integrals must be replaced by quadrature sums in all but the most trivial problems. This can be done in a variety of ways: we assume in what follows only that the integrals on both sides of (2.3) are evaluated by the same quadrature rule, namely

$$\int_a^b p(x) \ dx \tilde{=} \sum_{k=1}^Q \omega_k \ p(x_k) \ .$$

The resulting discretized problem depends on whether we choose the Galerkin or Ritz form of the integral in (2.5), so the two formulations are no longer equivalent. We prefer to distinguish them by the terms discrete Galerkin and discrete Ritz.

(a) discrete Ritz

Using the Ritz form of the integrals leads to $\overline{A}\underline{u} = \overline{g}$, where

$$\overline{a}_{ij} = \sum_{k=1}^{Q} \omega_k \, M(\phi_j(x_k), \phi_i(x_k))$$
$$\overline{g}_i = \sum_{k=1}^{Q} \omega_k \, f(x_k) \, \phi_i(x_k) \, . \tag{2.6}$$

This is the form normally used since it retains the matrix symmetry, and we refer to Strang and Fix [7] for estimates of the number of quadrature points Q necessary to ensure no loss of accuracy from the discretization (for piecewise polynomial bases).

(b) discrete Galerkin

This gives $\hat{A}\underline{u} = \hat{g}$ where

$$\hat{a}_{ij} = \sum_{k=1}^{Q} \omega_k \, I\phi_j(x_k)\phi_i(x_k)$$
$$\hat{g}_i = \sum_{k=1}^{Q} \omega_k \phi_i(x_k) \, f(x_k) \, . \tag{2.7}$$

If we define the matrix B by

$$b_{ik} = \phi_i(x_k), \; i=1,\ldots,N, k=1,\ldots,Q,$$

then (2.7) can be expressed as

$$BDC \, \hat{\underline{u}} = BD\underline{f} \, , \tag{2.8}$$

where C and f are defined in (2.2) and D = diag (ω_i). This gives easily

Theorem 2.1: If N=Q, the discrete Galerkin method (2.7) is equivalent to the collocation method (2.2) using the same points, provided

(i) none of the quadrature weights ω_k are zero

(ii) the matrix B is nonsingular .

Note: (ii) is guaranteed if the functions $\{\phi_i(x)\}$ are unisolvent.

Thus collocation can be viewed as a discrete Galerkin method using the same set of points, and of course is much less work since C is

easier to evaluate than $\hat{A}=BDC$. Normally however, to obtain the same order of accuracy as the undiscretized Galerkin method (2.3), we need $Q>N$. But for some special choices of piecewise polynomial bases and quadrature points, $Q=N$ is sufficient; we shall discuss this in Section III.

3. least squares

This solution $v^{(N)}(x) = \sum_1^N v_j \phi_j(x)$ is found by minimizing $\int_a^b (Lv^{(N)}-f)^2 dx$ with respect to the coefficients $\{v_j\}_1^N$. Again the solution is characterized by an orthogonality condition:

$$\int_a^b Lv^{(N)}(x)\ (L\phi_i)\ dx = \int_a^b f\ (L\phi_i)\ dx\ ,\ i=1,\ldots,N\ . \tag{2.9}$$

Discretizing with the same quadrature rule on both sides, we obtain

$$C^T D C\ \hat{\underline{v}} = C^T D \underline{f}. \tag{2.10}$$

From this we easily obtain

Theorem 2.2: If $N=Q$, the discrete least squares method (2.10) is equivalent to the collocation method (2.2) using the same points, provided

(i) the quadrature weights $\{\omega_k\}$ are nonzero

(ii) the collocation matrix C is nonsingular .

Again we normally use $Q>N$ here, but even in this case we can consider discrete least squares as an extension of collocation: if the weights $\{\omega_k\}$ are all positive, (2.10) is precisely the set of normal equations for the discrete linear least squares problem

$$\min_{\underline{v}} \left|\left| D^{\frac{1}{2}}(C\underline{v}-\underline{f}) \right|\right|_2^2 \tag{2.11}$$

Thus we merely "collocate" at more points (Q) than functions (N) giving an overdetermined set of linear equations; scale these by $D^{\frac{1}{2}}$ and solve by the familiar linear least squares method. We will return to this idea later.

III Convergence Results for Piecewise Polynomial Bases

Now we specialize the choice of basic functions $\{\phi_i(x)\}$ to piecewise polynomials: given a mesh $a=x_0<x_1<\ldots<x_N=b$ and $h = \max|x_{i+1}-x_i|$, we demand that each basis function be a polynomial of degree $2n-1$ in each subinterval, with k derivatives matching at the knots $\{x_i\}$ so the functions are globally $C^{(k)}$. This is $(N-1)(k+1)$ continuity conditions, and counting the $2m$ boundary conditions (1.2), there are $[(2n-1-k)N+k+1-2m]$ free parameters left, and thus the same number of basis functions. Computationally, it is important that the basis functions used have support over as small an interval as possible; we refer to de Boor [11] for a discussion of the B-spline basis for this space of functions, which has minimal support. Particular choices of interest are

(i) splines ($Sp_o^{(n)}$): degree $2n-1$, globally $C^{(2n-2)}$; support $2n$ subintervals

(ii) Hermites ($H_o^{(n)}$): degree $2n-1$, globally $C^{(n)}$; support 2 subintervals with either B-spline or natural Hermite bases.

Our purpose here is to give a uniform presentation of the known convergence results; for more details the reader is referred to other papers. Before giving the convergence results, we mention some standard preliminary results. For $u(x), v(x)$ satisfying (1.2), define

$$a(u,v) = \int_a^b \sum_{i=0}^{m} a_i(x) D^i u(x)\, D^i v(x)\, dx \quad . \tag{3.1}$$

Integrating by parts and using (1.2), we have

$$a(u,v) = \int_a^b u(x)\, Lv(x)\, dx = \int_a^b (Lu(x))v(x)\, dx \quad .$$

We also define the norm

$$\| v \|_D^2 = \int_a^b \sum_{i=0}^{m} (D^i v(x))^2\, dx.$$

It is well-known that if (1.1) is sufficiently smooth and <u>elliptic</u> (e.g. $a_i(x) \geq 0$ $(0 \leq i \leq m-1)$ and $a_m(x) \geq \delta > 0$) we have

$$C\,||v||_D^2 \le a(v,v) \le C'||v||_D^2 \tag{3.2}$$

and

$$|a(u,v)| \le C'||u||_D||v||_D \quad . \tag{3.3}$$

We also need the bilinear form

$$b(u,v) = \int_a^b Lu(x)\ \ Lv(x)\ dx \tag{3.4}$$

and the norm

$$||v||_E^2 = \int_a^b \sum_{i=0}^{2m} (D^i v(x))^2 dx \ .$$

Again if (1.1) is sufficiently smooth and elliptic,

$$K\,||v||_E^2 \le b(v,v) \le K'||v||_E^2 \tag{3.5}$$

and

$$|b(u,v)| \le K'\,||u||_E\,||v||_E \tag{3.6}$$

1. collocation

Convergence for the collocation method is given by the following theorem of de Boor and Swartz [2].

Theorem 3.1:

Suppose (1.1),(1.2) has a unique solution u(x) and the coefficients of (1.1) are sufficiently smooth. Then using a B-spline basis of degree 2n-1 and continuity $C^{(2m-1)}$, and collocating at the 2n-2m Gaussian points in each subinterval, produces a unique solution $w^{(N)}(x)$ for sufficiently small h, which satisfies

$$||u-w^{(N)}||_2 = O(h^{\min(2n,4n-4m)}) \tag{3.7}$$

and

$$|u(x_i) - w^{(N)}(x_i)| = O\,(h^{4n-4m}) \ , \ 1 \le i \le N \ . \tag{3.8}$$

The rather unusual continuity class required here (e.g. only $C^{(1)}$ for a second-order problem) is necessary because this gives exactly 2n-2m collocation points in each subinterval (see the formula in the first paragraph of this section). From our point of view, (3.7) is natural from Theorem 2.1: collocation at the 2n-2m Gaussian points is equivalent to a discrete Galerkin method using Gaussian quadrature (error bound $0(h^{4n-4m})$) and the corresponding Galerkin method, at least for a smooth basis, has error bound $0(h^{2n})$, as we shall see later. With this in mind, we give our own proof of part of Theorem 3.1.

Proof of (3.7):

Let ϕ be the solution of $L\phi=v\equiv \dfrac{u-w^{(N)}}{||u-w^{(N)}||_2}$ as in Nitsche [4].

The Green's function for L is sufficiently smooth that $||\phi^{(j)}||_\infty \le K, 0\le j\le 2m$.

Then $$||u-w^{(N)}||_2 = \int_a^b v\,(u-w^{(N)})dx = a(\phi,u-w^{(N)}) = \int_a^b \phi(f-\hat{f})\,dx$$

where $\hat{f}=Lw^{(N)}$ satisfies $\hat{f}(t_j) = f(t_j)$ at the 2n-2m Gaussian points in each subinterval $[x_i,x_{i+1}]$. If $p_i(x) = \prod_{j=1}^{2n-2m}(x-t_j)$, then

$$\int_{x_i}^{x_{i+1}} p_i(x)\, r(x)\, dx = 0 \qquad (3.9)$$

for all polynomials r(x) of degree less than 2n-2m. On $[x_i,x_{i+1}]$, $f-\hat{f}=p_i(x)q_i(x)$ and $||p_i(x)||_\infty = 0(h^{2n-2m})_{x_{i+1}}$. Thus

$$||u-w^{(N)}||_2 = \int \phi\,(f-\hat{f})dx = \sum_{i=0}^{N-1} \int_{x_i} p_i(x)\,[\phi\,(x)q_i(x)]dx\,.$$

Now expanding $[\phi q_i]$ in a Taylor series about x_i with $k=\min\{2m,2n-2m\}$ terms and using (3.9),

$$||u-w^{(N)}||_2 = \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} p_i(x)\,\frac{[\phi q_i]^{(k)}(\xi_x)}{k!}\,(x-x_i)^k dx = 0(h^{\min(2n,4n-4m)}).$$

QED.

This collocation at Gaussian points has proven very successful in practice especially for n=2m, in which case we are working with the Hermite space $H_o^{(n)}$ which has a very convenient natural basis (see [9] for some numerical comparisons with finite difference methods). However for n≠ 2m the computations require a B-spline basis; it might be more attractive computationally to use the Hermite space $H_o^{(r)}$ with continuity $C^{(r-1)}$ rather than $C^{(2m-1)}$, with r chosen so the order of accuracy is the same. For this space, the number of collocation points required is [rN+r-2m], so we can use r in each subinterval except for r-2m intervals where we use one less point if r<2m or one more if r>2m.

If we use Gaussian points in each subinterval, and assume that the collocation solution exists, the above proof of (3.7) shows

$$||u-w^{(N)}||_2 = O(h^{\min(r+2m,2r)})$$

for the Hermite space $H_o^{(r)}$. (One fewer point in some intervals only affects the local error by h, leaving the same global error.) Notice that we get the same convergence as with the B-spline basis of degree 2n-1, continuity 2m-1, if we take r=2n-2m. That is, we can collocate at the same 2n-2m Gaussian points, but with a natural Hermite basis rather than B-splines, and obtain just as much accuracy. In Section IV, we show that the amount of computation involved is the same.

2. Galerkin

The convergence rates for the (continuous) Galerkin method are well-known (see for example Varga [10]):

Theorem 3.2 : Suppose (1.1), (1.2) has a unique solution u(x), the coefficients of (1.1) are sufficiently smooth, and (3.2), (3.3) hold. Then using piecewise polynomials of degree 2n-1 and continuity at least $C^{(n-1)}[a,b]$, there exists a unique Galerkin solution $u^{(N)}(x)$ for h sufficiently small, and it satisfies $||u-u^{(N)}||_2 = O(h^{2n})$.

For the discrete Ritz method, the number of quadrature points required to maintain this accuracy is still not completely understood. Discretizations of only the right-hand side of (2.3) have been considered by Herbold, et al [3] and Schultz [5]. More recently, Strang and Fix [7] have considered the more realistic problem of discretizing both sides

as in (2.6). They show that using (2n-1) Gaussian points in each subinterval maintains the $O(h^{2n})$ accuracy.

3. Least Squares

Convergence of the (continuous) least squares method for very general problems has been analyzed by Bramble and Schatz [1]. For the sake of completeness, we give a convergence proof for our particular problem (i.e. (2.9) applied to (1.1)).

Theorem 3.3: Suppose (1.1),(1.2) has a unique solution u(x), the coefficients of (1.1) are sufficiently smooth, and (3.5), (3.6) hold. Then using piecewise polynomials of degree 2n-1 and continuity at least $C^{(n-1)}[a,b]$, the least squares solution $v^{(N)}(x)$ exists for h sufficiently small, and satisfies $||u-v^{(N)}||_2 = O(h^{2n})$.

Proof: (The proof models Schultz [6] for the Galerkin solution.) From (2.9) and (3.4), the exact solution u(x) satisfies

$$b(u,v) = \int_a^b (Lu)(Lv)\,dx = \int_a^b f(Lv)\,dx$$

for all v, and the least squares solution $v^{(N)}(x)$ satisfies

$$b(v^{(N)},v) = \int_a^b f(Lv)\,dx$$

for all $v \in S_N$. So for any $v, w \in S_N$,

$$b(w-v^{(N)},v) = b(w-u,v).$$

Take $v=w-v^{(N)}$ and use (3.5):

$$||w-v^{(N)}||_E^2 \le \frac{1}{K}\,|b(w-v^{(N)},\, w-v^{(N)})| \le \frac{K'}{K}\,||u-w||_E\,||v^{(N)}-w||_E .$$

Now let w be the interpolate of u in S_N; it is well-known that

$$||u^{(j)}-w^{(j)}||_2 = O(h^{2n-j})\,||u^{(2n)}||_2 , \quad 1 \le j \le 2m .$$

Thus

$$||u-v^{(N)}||_E \le ||u-w||_E + ||w-v^{(N)}||_E \le (1+\frac{K'}{K})\,||u-w||_E = O(h^{2n-2m}) .$$

To get the higher order convergence in the L_2 norm, we again use the device of Nitsche: let ϕ, ψ be defined by (1.2) and

$$L\phi = \psi \ , \ L\psi = \frac{u-v^{(N)}}{||u-v^{(N)}||_2}$$

From continuity of the Green's function for L, we know $||\phi^{(j)}||_2 \leq K$, $0 \leq j \leq 2m$. Now

$$||u-v^{(N)}||_2 = \int_a^b (L(L\phi))\ (u-v^{(N)})\ dx = b(u-v^{(N)}, \phi)$$

$$= b(u-v^{(N)}, \phi-w)$$

for all $w \in S_N$. Let w be the interpolate of ϕ in S_N; since we know $||\phi^{(j)}||_2 \leq K$ for $0 \leq j \leq 2m$, we have

$$||u-v^{(N)}||_2 \leq K||u-v^{(N)}||_E\ ||\phi-w||_E \leq K\ h^{2n-2m}\ h^{2m}||\phi^{(2m)}||_2$$ QED.

If we discretize the least squares problem as in (2.10), we need to ensure that this convergence rate is maintained. As we mentioned at the end of Section II, this amounts to collocating at more points than there are functions and solving the resulting discrete linear least squares problem (2.11) by familiar methods. For example, we could use piecewise polynomials of degree 2n-1 and continuity $C^{(k)}$, k>2m-1, and "collocate" at the 2n-2m Gaussian points in each subinterval. We conjecture this keeps $O(h^{\min(2n,4n-4m)})$ accuracy. The advantages are that we obtain higher global continuity of the approximate solution, and we can use other basis functions than B-splines (e.g. the natural Hermite basis for $H_o^{(n)}$) without going to higher degree as was necessary with collocation. One can even use splines (ie. continuity $C^{(2n-2)}$); as we shall see in Section IV, this is more economical and can even be less work than collocation. Experiments of P. Sammon have shown $O(h^4)$ convergence with cubic splines, using the two Gaussian points in each subinterval as data points, and solving the resulting overdetermined linear system by familiar linear least squares methods.

IV Comparison of Methods

Here we compare the computational work involved for methods of the same global accuracy on problem (1.1), (1.2), using piecewise polynomial bases. All the methods compared have global error $0(h^{2n})$; normally the polynomials have degree 2n-1, and we assume a fixed mesh $a=x_0<x_1<\ldots<x_N=b$.

1. collocation

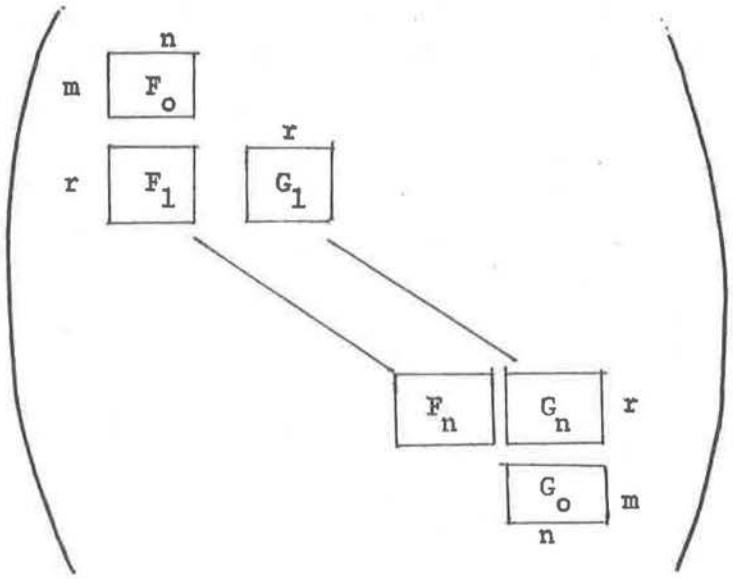
As we saw in Section III, we can get $0(h^{2n})$ global error by collocating with the B-splines of de Boor-Swartz of degree 2n-1, or by using the Hermite space $H_o^{(r)}$, r=2n-2m. In what follows, we assume $n\geq 2m$ so $r\geq n$.

(a) B-splines of deBoor-Swartz

These functions have degree 2n-1, continuity $C^{(2m-1)}$, and we collocate at the r=2n-2m Gaussian points in each subinterval. With the boundary conditions, this gives a total of rN+2m equations. There are r basis functions associated with each interior knot and n at each endpoint, giving the same number of unknown coefficients to solve for .

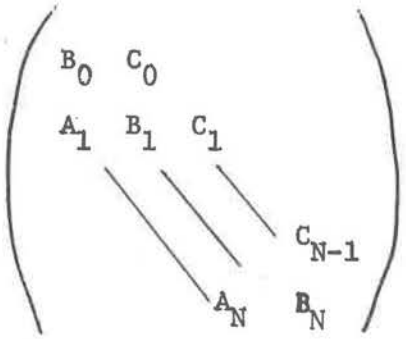
Two components to the work are involved in any of these methods: forming the matrix elements and solving the resulting linear system. For collocation, each matrix element involves an evaluation of (1.1); we denote this by E_L . Although the evaluation time depends to some extent on the basic function used (since we need $\phi_j(x_i)$, $\phi_j'(x_i)$, etc.), this work does not depend on N (i.e. on h) since the evaluations are always at the Gaussian points, and these coefficients can be stored beforehand, no matter what h is. Thus we can assume E_L is only a function of m, the order of the differential equation. Also, we do not consider the work involved in evaluating the approximate solution for given values of x after it has been computed; this also depends on the basis used.

These B-splines have support over something less than two subintervals (their continuity is less than the Hermite basis $H_o^{(n)}$ for n>2m) but to simplify the matrix analysis, we assume it is in fact two subintervals. Then the matrix has the form



(4.1)

We solve (4.1) by block-tridiagonal factorization; i.e. we write it in the form



(4.2)

with B_o nxn, A_1 rxn, C_o nxr, B_N nxn, A_N nxr, C_{N-1} rxn, and the other blocks rxr. Since r/2 = n-m, this cuts each F_i, G_i block in half horizontally, so the top half of C_i and bottom half of A_i are zero. Thus we can use the anaysis of [8, pg 867] to show the solution time is $(\frac{13}{12}r^3 + 2r^2)NM$, where M denotes the average time for a multiplication/division. This together with the setup time for the matrix elements gives a total time estimate of

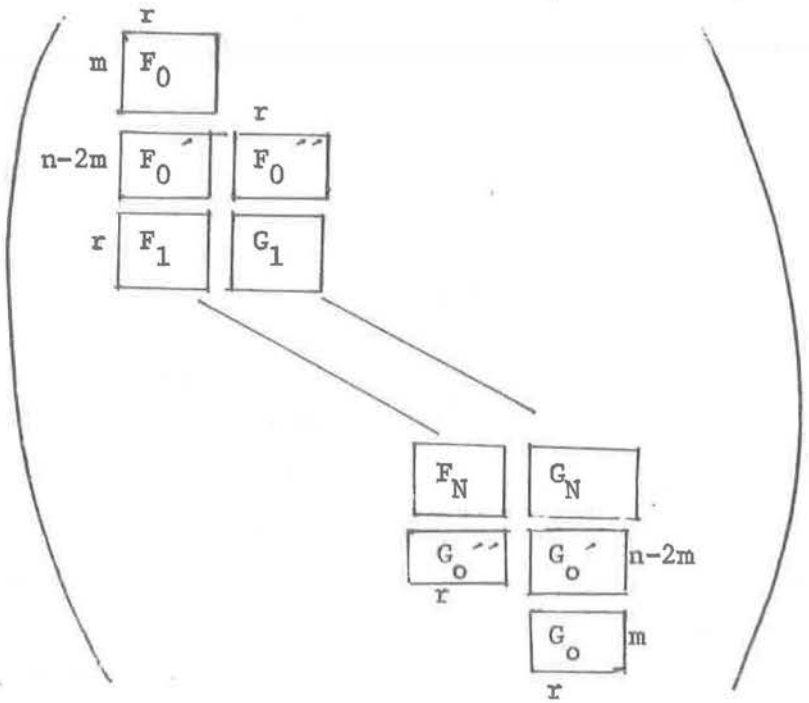
$$(\frac{13}{12}\, r^3N+2r^2N)M + 2r^2\, NE_L. \tag{4.3}$$

b) <u>collocation with</u> $H_o^{(r)}$

These are piecewise polynomials of degree 2r-1, continuity $C^{(r-1)}$, and we need r=2n-2m to give $O(h^{2n})$ convergence. We use the natural Hermite basis, having r basis functions associated with each interior

knot and r at each endpoint, giving r(N+1) functions in all and the same number of coefficients to determine. We again collocate at the r Gaussian points in each interval; this and the boundary conditions make (rN+2m) equations. Thus we need r-2m=2(n-2m) extra equations when $n \geq 2m$; we get these by collocating at (n-2m) extra points in the first and last interval. This maintains $0(h^{2n})$ accuracy and makes the matrix analysis easier than using one more point in each of several intervals, as we did for the convergence results in Section III. (However if $m<n<2m$, this messier approach would be necessary as there would be fewer points in some intervals.)

Since these basis functions have support over exactly two subintervals, the matrix has the form



Again we put this in the form (4.2), this time with B_0 rxr. Again exactly half of the A_i, C_i matrices are zero, so the solution time is the same as for the B-splines of deBoor-Swartz. The setup time is also the same, so we again get (4.3) as our work estimate. The only difference in computation time will be in evaluation of the approximate solution as we alluded to earlier. This will probably be less for the Hermite basis, as the B-splines are somewhat cumbersome to evaluate.

2. discrete Ritz

For the discrete Ritz method (see (2.6)) the two obvious choices for bases giving $0(h^{2n})$ accuracy are the piecewise Hermite polynomials of degree 2n-1 ($H_0^{(n)}$) and splines of degree 2n-1 ($S_{p_o}^{(n)}$).

(a) $H_o^{(n)}$:

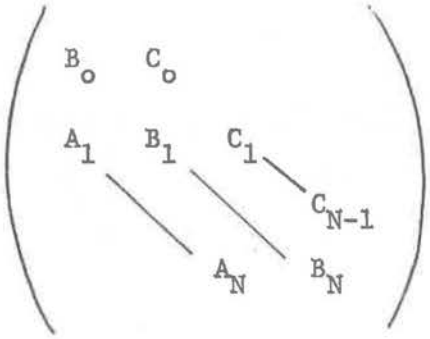
Since these are $C^{(n-1)}$ at the knots, there are n basis functions associated with each of the (N-1) interior knots and n at each endpoint. Thus we find $u^{(nN+n)}(x)=\sum_{i=1}^{nN+n} \bar{u}_i \phi_i(x)$ by solving $\bar{A}\,\bar{u} = \bar{b}$,

$$\bar{a}_{ij} = \sum_{k=1}^{qN} \omega_k M(\phi_j(\xi_k), \phi_i(\xi_k))$$
$$\bar{b}_i = \sum_{k=1}^{qN} \omega_k \phi_i(\xi_k) f(\xi_k) \quad . \tag{4.4}$$

The homogeneous boundary conditions (1.2) imply $\bar{u}_i=0$ for $1\le i\le m$ and $nN+1\le i\le nN+m$ and we include these as the first and last m equations of the linear system to simplify the matrix analysis.

We assume the quadrature rule uses q points in each subinterval; as we mentioned earlier, the value of q necessary to maintain $0(h^{2n})$ accuracy is not completely understood, but for example one could use q=2n-1 Gaussian points in each subinterval (see Strang and Fix [7]). We believe q=2n-m is in fact sufficient for the general problem (1.1). Using fewer points seems not to work: in fact solving the problem $y''=f(x)$ using a 2-point Gauss rule for cubic Hermite polynomials (n=2,m=1) with equally spaced knots leads to a singular matrix $\bar{A}$ in (2.6).

These natural Hermite basis functions have support over two subintervals and the quadrature sums are only over the intersection of the supports of the functions used. Thus $\phi_{in+j}(x)$ has support (x_{i-1},x_{i+1}) for $1\le j\le n$ and $\bar{A}$ has the block tridiagonal form



(4.5)

with each block nxn.

Now consider the setup time for (4.5). It is symmetric, so we need only consider the upper triangle. For an element of B_i, both functions in (4.4) have support (x_{i-1}, x_{i+1}) so the quadrature sum is over 2q points; for C_i the sum is only over the q points in (x_i, x_{i+1}). This makes a total of $(2n^2+n)Nq$ evaluations of $M(\phi_j, \phi_i)$ (denoted E_M) and subsequent multiplications to form $\overline{A}$ from (4.4). For the right hand side $\overline{b}$ we have an additional $4qnN$ multiplications (and qnN evaluations of ϕ_i, f which we ignore).

Solution time for a matrix like (4.5) using a block-Cholesky factorization is essentially $\frac{5}{3}n^3N$ multiplications, giving a total for discrete Ritz using Hermite functions of

$$[(2n^2+5n)qN+\tfrac{5}{3}n^3N]\ M + (2n^2+n)\ qNE_M \qquad (4.6)$$

(ii) $S_{p_o}^{(n)}$:

For regular splines of degree 2n-1, continuity $C^{(2n-2)}$, there is just one B-spline basis function associated with each interior knot and n at each endpoint. Each has support over 2n intervals so for the n additional functions at x=a we use B-splines centred at $x_o=a, x_{-1}, \ldots, x_{-n+1}$ (defined by reflection through a), and similarly for x=b. The m boundary conditions at each endpoint involve linear combinations of all 2n B-splines which are nonzero there. We take care of these implicitly by using as our basis the 2n-m appropriate linear combinations of these

2n B-splines which automatically satisfy the boundary conditions.

Thus in general ϕ_i has support (x_{i+m-2n}, x_{i+m}) and the computation of $\bar{a}_{ij}$ from (4.4) involves the intersection of the supports of ϕ_i and ϕ_j which is (x_{j+m-2n}, x_{i+m}) for $j \geq i$. So $\bar{a}_{ij} \neq 0$ for $|j-i|<2n$, or $\bar{A}$ has half-bandwidth 2n-1. Assuming q quadrature points per subinterval, this means the ith row of $\bar{A}$ takes $q(\sum_{1}^{2n} k) = qn(2n+1)$ evaluations of $M(\phi_j, \phi_i)$ and subsequent multiplications. Each $\bar{b}_i$ requires $4nq$ multiplications, and solving $\bar{A}\bar{u}=\bar{b}$ by band Cholesky takes $2n^2N$ multiplications, giving a total for discrete Ritz using splines of

$$[(2n^2+5n)qN+2n^2N]M+(2n^2+n)qNE_M \qquad (4.7)$$

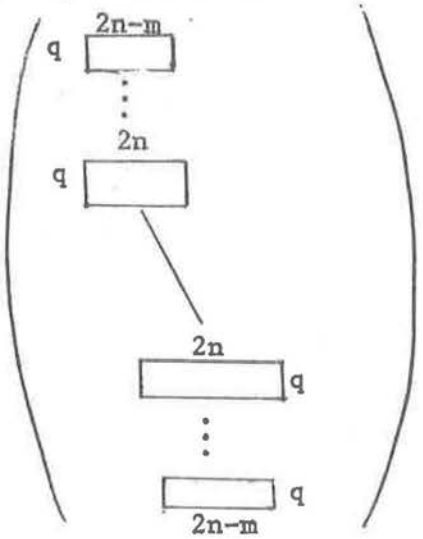
Notice that (4.7) and (4.6) are almost identical, except that the solution time with the spline matrix is less.

3. least squares

In Section II, we saw that the discrete least squares method generalizes the collocation procedure when more collocation points than functions are desired. In particular, this provides a viable alternative to the Galerkin method when a smooth spline basis is used. We consider only this smooth spline basis (i.e. degree 2n-1, continuity $C^{(2n-2)}$) because, as we saw with the Galerkin method, it is the most economical.

Assuming q quadrature points (or "collocation" points) in each subinterval, the matrix C of (2.10) looks like

$C =$



(4.8)

where there are N horizontal blocks of q rows each. Thus formation of $D^{\frac{1}{2}}C$ takes $(2nqN)E_L + (2nqN)M$. Now to solve the discrete linear least squares problem, we form the normal matrix $(D^{\frac{1}{2}}C)^T(D^{\frac{1}{2}}C)$. This has precisely the same form as the spline Galerkin matrix $\bar{A}$; namely, banded with half-bandwidth 2n. Formation of a general row of the normal matrix takes $q(2n)+q(2n-1)+\ldots+q(1)$ multiplications. For the upper half of the whole matrix, this amounts to $Nq(\sum_1^{2n} k)=Nqn(2n+1)$, plus 2nqN for the right hand side. Finally, solving by band Cholesky takes $2n^2N$ multiplications, giving a total for least squares of

$$[(2n^2+5n)qN+2n^2N]M+(2nqN)E_L \quad , \tag{4.9}$$

We can draw the following conclusions about the relative efficiencies of these methods: comparing (4.7) and (4.9) we see that discrete least squares is more efficient than discrete Ritz, assuming (as we shall) $E_L=E_M$. The number of quadrature points for discrete Ritz is at most q=2n-1 and is probably q=2n-m. On the other hand, for discrete least squares we believe q=2n-2m is sufficient, as conjectured in Section III. However even if we take the same q for both methods, least squares with splines is always more efficient than discrete Ritz with splines, because of fewer function evaluations.

The comparison with collocation is a bit more difficult; from (4.3) we see that collocation (with either $H_0^{(r)}$ or the B-splines of de Boor-Swartz) is cheaper than discrete Ritz because of fewer function evaluations. However the relative merits of collocation and least squares depend on the value of n and m (see the table below where we assume q=2n-m for discrete Ritz, q=2n-2m for discrete least squares).

	m=1,n=2	m=1,n=3	n large, m small
collocation	$(16\frac{2}{3}M+8E_L)N$	$(101\frac{1}{3}M+32E_L)N$	$(\frac{26}{3}n^3M + 8n^2E_L)N$
discrete Ritz	$(62M + 30E_M)N$	$(183M + 105E_M)N$	$(4n^3M + 4n^3E_M)N$
discrete least squares	$(44M + 8E_L)N$	$(150M + 24E_L)N$	$(4n^3M + 4n^2E_L)N$

Thus for small values of n, collocation is cheaper; however for large n least squares takes about half the time of collocation, and both are an order of magnitude better than discrete Ritz (because of fewer function evaluations).

References

1. J.H. Bramble and A.H. Schatz, On the numerical solution of elliptic boundary value problems by least squares approximation of the data.
Proceedings of SYNSPADE Conference, B. Hubbard, ed. Acad.Press 1971.

2. C. de Boor and B. Swartz, Collocation at Gaussian points.
SIAM Journal Num. Anal. 10 (1973), 582-606.

3. R.J. Herbold, M.H. Schultz, and R.S.Varga, Quadrature schemes for the numerical solution of boundary value problems by variational techniques.
Aeq. Math. 3 (1969), 96-119.

4. J. Nitsche, Ein Kriterium fur die Quasi-Optimalital des Ritzschen Verfahrens. Num. Math. 11 (1968), 346-348.

5. M.H. Schultz, Quadrature-Galerkin approximations to solutions of elliptic differential equations. Proc. AMS 33 (1972), 511-515.

6. M.H. Schultz, Spline Analysis. Prentice-Hall, New York, 1973.

7. G. Strang and G. J. Fix, An Analysis of the Finite Element Method. Prentice-Hall, New York, 1973.

8. J. M. Varah, On the solution of block-tridiagonal systems arising from certain finite-difference equations. Math. Comp. 26 (1972), 859-868.

9. J.M. Varah, A comparison of some numerical methods for two-point boundary value problems. Submitted to Math. Comp.

10. R. S. Varga, Functional analysis and approximation theory in numerical analysis. SIAM Regional Conference Series, vol. 3.

11. C. de Boor, On calculating with B-splines. J. Approx. Theory 6 (1972), 5062.