MATHEMATICS OF COMPUTATION, VOLUME 29, NUMBER 131
JULY 1975, PAGES 777–793

Computational Investigations of Least-Squares Type Methods for the Approximate Solution of Boundary Value Problems*

By Steven M. Serbin

Abstract. Several Galerkin schemes for approximate solution of linear elliptic boundary value problems are studied for such computational aspects as obtainable accuracy, sensitivity to parameters and conditioning of linear systems. Methods studied involve computing subspaces (e.g., splines) whose elements need not satisfy boundary conditions. A Poisson problem study on the square produces computed error reflective of theoretical L_2 estimates and L_∞ behavior optimal for smooth data but loss according to Sobolev's lemma for nonsmooth data. Insensitivity to parameters is evidenced. Analogous one-dimensional methods enhance the conditioning study. Studies are included for parallelogram and L-shaped domains.

1. Introduction. The purpose of this paper is to present the results of several numerical experiments which have been performed with least-squares and related methods for the approximate solution of linear elliptic boundary value problems. We consider such computational aspects as obtainable accuracy, sensitivity with respect to weighting parameters, and conditioning of resulting linear algebraic systems for each of these methods, which have the common characteristic that the elements of the finite-dimensional subspace in which the solution is approximated need not satisfy the boundary conditions of the problem.

In Section 2 we describe the class of problems under consideration and develop the notation we will use. In Section 3, we present the three approximation methods with which the studies have been performed and include a theoretical result pertaining to a quadratic form of one of these methods. In Section 4 we discuss the particular computational details of our implementation of these methods; it is believed that these details may be of interest to some members of the scientific community. In Section 5 we detail several experiments performed on the Poisson problem in the unit square and compare results with various approximating subspaces and boundary weightings. In Section 6 we look briefly at some problems on other domains, including the L-shaped region. In Section 7 we present some analogous methods for two-point boundary value problems and use these mainly to examine conditioning behavior. We conclude in Section 8 with a discussion of results and a mention of ongoing experiments.

2. The Problem. The class of boundary value problems upon which the experiments have been performed may be described as follows (we adopt the notation of [7]).

Received January 16, 1974.
AMS (MOS) subject classifications (1970). Primary 65N30, 35J25.
*This research was supported in part by National Science Foundation Grants GP-27630 and GP-22936.

Let R be a bounded domain in $\mathbf{R}^N$ with piecewise smooth boundary ∂R. (Bramble and Schatz in their original paper on least squares require that ∂R is C^∞; however, one of our aims is to study computationally domains, such as rectangles, for which this is violated.)

We shall be interested in approximating the solution of the problem

$$Au = f \quad \text{in } R, \qquad u = g \quad \text{on } \partial R, \tag{2.1}$$

where

(i) $$Au = \sum_{i,j=1}^{N} a_{ij}(x) D_i D_j(u) + \sum_{i=1}^{N} b_i(x) D_i(u) + c(x)u$$

and A is uniformly elliptic in $\bar{R}$; that is, there exists a constant $C > 0$ such that

$$C^{-1}|\xi|^2 \leqslant \left| \sum_{i,j=1}^{N} a_{ij}\xi_i\xi_j \right| \leqslant C|\xi|^2$$

for all $x = (x_1, x_2, \ldots, x_N) \in \bar{R}$ and $\xi \in \mathbf{R}^N$.

(ii) $D_i = \partial/\partial x_i$ and all coefficients a_{ij}, b_i, and c are assumed to be real-valued and C^∞ in $\bar{R}$.

(iii) The data satisfy (at least) $f \in L_2(R)$, $g \in L_2(\partial R)$—additional smoothness on g is required to obtain optimal error estimates in [7] and [4].

When R is a smooth domain, the problem (2.1) is viewed in a weak form [7], wherein the data is approximated by C^∞ data (f_n, g_n) which converge in appropriate Sobolev spaces (see below) to the data of (2.1) the problem is solved for smooth solution u_n, and a limit used to define u. However, this process fails in domains with corners. Although theoretically our problem should also be viewed in a weak sense, for computational purposes, solutions will be obtained in the classical sense.

The following notation will be used: On $L_2(R)$ and $L_2(\partial R)$ we have the respective inner products $(\phi, \psi) = \int_R \phi\psi\, dx$ and $\langle \phi, \psi \rangle = \int_{\partial R} \phi\psi\, d\sigma$.

Let Ω be a fixed open set containing $\bar{R}$. $H^m(\Omega)$ and $H^M(\partial\Omega)$ are the Sobolev spaces of order m of functions Ω and $\partial\Omega$ respectively with norms denoted by $\|\cdot\|_m^\Omega$ and $|\cdot|_m^\Omega$. (See [14] for definitions of these spaces when $\partial\Omega$ is C^∞; in the case of polygonal domains, though, a different definition of $H^m(\partial\Omega)$, due to Kellogg [21] is appropriate.) Denote $H^{(r,s)} = H^r(\Omega) \times H^s(\partial\Omega)$. $S^h_{m,k}$ is any finite-dimensional subspace of $H^m(\Omega)$ which satisfies the approximability assumption:

(2.2) For any $u \in H^k(\Omega)$, $k \geqslant m$, there exists a constant C (independent of the parameter h and of u) such that

$$\inf_{\chi \in S^h_{m,k}} \sum_{j=0}^{m} h^j \|u - \chi\|_j^\Omega \leqslant Ch^k \|u\|_k^\Omega .$$

This is not exactly the assumption of [4], but for many practical subspaces, both assumptions hold.

We shall make a particular choice of such subspaces (splines) below; we mention also some other subspaces that have been studied by others. S. Hilbert [13] details

the construction of multi-dimensional Hermite functions. Schultz [17] has studied many such spaces on rectilinear domains in $\mathbf{R}^n$. Strang [20] gives several examples and presents easily verifiable conditions for pointwise approximation of smooth functions to specified order and also for L_2 approximation. Bramble and Zlámal [8] and Di Guglielmo [9] use subspaces in which the "elements" (here, the support of the trial functions) are nonrectilinear and thus have better chance of conforming to irregular boundaries.

Finally, we require the Dirichlet integral

$$D(\psi, \chi) = \int_R \sum_{i=1}^{N} \frac{\partial \psi}{\partial x_i} \frac{\partial \chi}{\partial x_i} dx$$

and denote by ψ_n the outward normal derivative, ∇_n the outward normal derivative, $\nabla_s \psi$ the surface gradient for $\psi \in H^1(\partial R)$, and $\gamma > 0$ and $0 < h < 1$ parameters.

3. Methods. I. The least-squares method of Bramble and Schatz [7] may be described as follows: For $(f, g) \in H^{(0,0)}$, the solution u to (2.1) minimizes over $H^m(R)$ ($m = 2$ here is the order of the differential operator) the functional

$$G(\chi) = \|f - A\chi\|^2 + \gamma h^{-3} |g - \chi|^2 \tag{3.1}$$

(where the zero subscript on the norms has been omitted). Equivalently, if we define the bilinear form

$$L(\psi, \chi) = (A\psi, A\chi) + \gamma h^{-3} \langle \psi, \chi \rangle, \tag{3.2}$$

then u satisfies

$$L(u, \chi) = (f, A\chi) + \gamma h^{-3} \langle g, \chi \rangle \tag{3.3}$$

for all $\chi \in H^2(R)$.

We define $S^h_{m,k}$ to be the restriction to $\bar{R}$ of $S^h_{m,k}(\Omega)$. The approximation method, using the Galerkin idea, is to find $w \in S^h_{2,k}$ such that

$$L(w, \chi) = (f, A\chi) + \gamma h^{-3} \langle g, \chi \rangle \quad \text{for all } \chi \in S^h_{2,k}. \tag{3.4}$$

For computational purposes, we select a basis $\{\phi_s\}_{s=1}^M$ of $S^h_{2,k}$, (M is inversely proportional to a power of h) and setting $w = \Sigma_{s=1}^M c_s \phi_s$, (3.4) yields

$$\sum_{s=1}^{M} c_s L(\phi_s, \phi_r) = (f, A\phi_r) + \gamma h^{-3} \langle g, \phi_r \rangle, \qquad r = 1, \cdots, M. \tag{3.5}$$

The matrix of this problem is symmetric, positive definite, and with the choice of basis discussed below, a band matrix. Fix and Larsen [11] provide the result that the spectral condition number of the least-squares matrix behaves as $O(h^{-4})$; in view of the fact that the usual Rayleigh-Ritz methods are known to demonstrate $O(h^{-2})$ conditioning, an investigation of possible ill effects of roundoff in (3.5) is thus indicated, and those studies are presented in Section 5.

We also study what happens if ∂R is not C^∞, the sensitivity with respect to the weight γ, and compare obtainable accuracy vs. theoretical estimates. The original error estimates for domains with C^∞ boundary which were obtained by Bramble and Schatz

have been verified and their proofs simplified by Baker [4]. The particular estimate with which we shall be concerned may be stated:

Suppose R is a bounded domain with C^∞ boundary. If $u \in H^s(R)$ for $2 \leqslant s \leqslant k$ ($k \geqslant 4$) satisfies (2.1) and $w \in S^h_{2,k}$ satisfies (3.4), then the L_2 error satisfies

$$\|u - w\| \leqslant Ch^s \|u\|_s. \tag{3.6}$$

This says that the least-squares technique reproduces the order of best approximation, a condition we shall refer to as being "optimal".

In [19] we obtain a like result for $k \geqslant 4$ when R is the unit square, but for brevity's sake we shall not include this proof. Note that the proofs do not hold for the case $k = 3$, and we shall investigate this in Section 5 to see if the techniques of proof are at fault or if indeed one cannot achieve optimal accuracy.

II. Now, let $A = -\Delta$ (Δ is the Laplace operator; $\Delta u = \Sigma_{i=1}^N \partial^2 u/\partial x_i^2$). Let us write the boundary value boundary problem in weak form:

$$-(-f - \Delta u, \chi) + \langle g - u, \gamma h^{-1}\chi - \chi_n \rangle = 0 \quad \text{for all } \chi \in H^2(R) \tag{3.7}$$

which can be rearranged as

$$\begin{aligned}(f, \chi) + \langle g, \gamma h^{-1}\chi - \chi_n \rangle &= -(\Delta u, \chi) + \langle u, \gamma h^{-1}\chi - \chi_n \rangle \\ &= D(u, \chi) - \langle \chi, u_n \rangle - \langle u, \chi_n \rangle + \gamma h^{-1} \langle u, \chi \rangle\end{aligned} \tag{3.8}$$

using Green's theorem. If we define the bilinear form

$$N(\psi, \chi) = D(\psi, \chi) - \langle \psi, \chi_n \rangle - \langle \chi, \psi_n \rangle + \gamma h^{-1} \langle \psi, \chi \rangle \tag{3.9}$$

our problem becomes

$$N(u, \chi) = (f, \chi) + \langle g, \gamma h^{-1}\chi - \chi_n \rangle. \tag{3.10}$$

Nitsche's method [15] is then: Find $w \in S^h_{2,k}$ such that

$$N(w, \chi) = (f, \chi) + \langle g, \gamma h^{-1}\chi - \chi_n \rangle \quad \text{for all } \chi \in S^h_{2,k}. \tag{3.11}$$

Some of the properties of Nitsche's method for smooth domains with solutions in $H^2(R)$ are

(i) L_2 error estimates are optimal for $k > 2$.

(ii) The condition number is $O(h^{-2})$.

(iii) But, "inverse theorems" (bounding higher Sobolev norms by lower ones with appropriate loss of powers of h) are required on the computing subspaces in order to make $N(\psi, \psi)$ positive definite.

III. Bramble and Nitsche [6] have combined their methods in order to utilize the best properties of each. We define the bilinear form

$$\begin{aligned}K_0(\psi, \chi) = {} & D(\psi, \chi) - \langle \psi, \chi_n \rangle - \langle \chi, \psi_n \rangle \\ & + h^2(\Delta\chi, \Delta\psi) + \gamma h^{-1} \langle \psi, \chi \rangle + \gamma h \langle \nabla_s \psi, \nabla_s \chi \rangle.\end{aligned} \tag{3.12}$$

Method K_0 then becomes: Find $w \in S^h_{2,k}$ such that

$$\begin{aligned}K_0(w, \chi) = K_0(u, \chi) = {} & (f, -\chi + h^2\Delta\chi) + \langle g, \gamma h^{-1}\chi - \chi_n \rangle \\ & + \langle \nabla_s g, \gamma h \nabla_s \chi \rangle \quad \text{for all } \chi \in S^h_{2,k}.\end{aligned} \tag{3.13}$$

This method has $K_0(\psi, \psi)$ positive definite for all $\gamma \geqslant \gamma_0$ and hence we seek both theoretical and computational estimates for γ_0. The proof of the following may be found in [19].

PROPOSITION 1. *Let $N = 2$ and R be star-shaped. By this we mean that if $x \in \partial R$ and the unit outward normal n exists at x (at all but possibly finitely many corners), then $x \cdot n \geqslant \kappa > 0$. Suppose that $|\chi_i| \leqslant x_M$ and let $\alpha = x_M/\kappa$. Then if $\gamma \geqslant \underline{\gamma} = \frac{1}{2}(\alpha + \sqrt{\alpha^2 + (\alpha + \sqrt{1+\alpha})^2})$, $K_0(\psi, \psi) > 0$, $\psi \neq 0$. In particular, on the unit square $\kappa = 1$, $x_M = 1$, $\alpha = 1$, and $\underline{\gamma} \simeq 1.8$. Note that all constants are independent of h.*

Since a matrix is positive definite if and only if its eigenvalues are all positive, we may use the inverse power method to estimate the least eigenvalue for several values of h in $S^h_{2,4}$ when R is the unit square to obtain a comparison with $\underline{\gamma}$, which is a bound for all h. In Tables 3.1 and 3.2 we see that $K_0(\psi, \psi)$ is positive definite for $\gamma > 1$, while Nitsche's form $N(\psi, \psi)$ (for which the theoretical analysis was not done) is definite for $\gamma > 6.7$.

TABLE 3.1. *Least Eigenvalue (λ) in Modulus for Matrices of Method K_0*

$h = 1/6$	$M = 81$	$h = 1/12$	$M = 225$
γ	λ	γ	λ
4	$.162 \times 10^{-2}$	4	$.168 \times 10^{-2}$
2	$.120 \times 10^{-2}$	2	$.125 \times 10^{-2}$
1	$.952 \times 10^{-2}$	1	$.992 \times 10^{-3}$
31/32	< 0	31/32	< 0

TABLE 3.2. *Least Eigenvalue in Modulus for Matrices of Nitsche's Method*

$h = 1/6$	$M = 81$	$h = 1/10$	$M = 169$
γ	λ	γ	λ
7	$.183 \times 10^{-5}$	7	$.214 \times 10^{-5}$
6.75	$.281 \times 10^{-6}$	6.75	$.506 \times 10^{-6}$
6.71875	$.806 \times 10^{-7}$	6.6875	$.847 \times 10^{-7}$
6.6875	< 0	6.5625	< 0

IV. For practical purposes, we omit the tangential derivitives:

$$K(\psi, \chi) = K_0(\psi, \chi) - \gamma h \langle \nabla_s \psi, \nabla_s \chi \rangle. \tag{3.14}$$

The Galerkin equations are

$$\sum_{s=1}^{M} c_s K(\phi_s, \phi_r) = (f, -\phi_r + h^2 \Delta \phi_r) + \left\langle g, \gamma h^{-1} \phi_r - \frac{\partial \phi_r}{\partial n} \right\rangle, \tag{3.15}$$

where Green's theorem allows us to write

$$\begin{aligned} K(\phi_s, \phi_r) = &-(\phi_r, \Delta \phi_s) - (\phi_s, \Delta \phi_r) - D(\phi_r, \phi_s) \\ &+ h^2 (\Delta \phi_r, \Delta \phi_s) + \gamma h^{-1} \langle \phi_r, \phi_s \rangle. \end{aligned} \tag{3.16}$$

This method requires inverse theorems to get $K(\psi, \psi)$ definite, but its condition number is $O(h^{-2})$.

4. Computational Details. We have chosen to compute with subspaces whose elements are tensor products of one-dimensional spline functions referred to by Babuška [2] as "hill" functions. These coincide in fact with B-splines defined by Schoenberg [16] and the resulting tensor products are just the splines in $\mathbf{R}^N$ discussed by Bramble and Hilbert [5].

We define $\psi_1(\chi) = \chi_{[-\frac{1}{2}, \frac{1}{2}]}(x)$ where $\chi_{[a,b]}(x)$ is the characteristic function of the interval $[a, b]$. Then, define recursively $\psi_k(x) = (\psi_{k-1} * \psi)(x)$. ψ_k has support on $[-k/2, k/2]$, is a piecewise polynomial of degree $(k-1)$ and is $C^{k-2}(-\infty, \infty)$. Segethova [18] has developed a stable recursive procedure for generating representations of the hill functions up to very high order, and we adopt her expansion method. We represent the hill functions and derivatives in local coordinates with Legendre polynomials $P_l(\chi)$ (orthogonal on $[-\frac{1}{2}, \frac{1}{2}]$): Let

$$\theta_\nu^{k\beta}(x) = \sum_{l=1}^{N} \alpha_{l\nu}^{k\beta} P_l(x) \tag{4.1}$$

and then

$$D^\beta \psi_k(x) = \sum_{\nu=1}^{k} \theta_\nu^{k\beta}\left(x - \frac{2\nu - k - 1}{2}\right) \chi_{[\nu-1-k/2,\, \nu-k/2]}(x), \tag{4.2}$$

where the $\alpha_{l\nu}^{k\beta}$ are coefficients given in [19].

For the square $R_s = (0, 1) \times (0, 1)$, we impose a uniform mesh of spacing h where h^{-1} is an integer and choose

$$S_{2,k}^h = \operatorname{Span}\left\{\psi_k\left(\frac{x - ih}{h}\right) \psi_k\left(\frac{y - jh}{h}\right),\right.$$

$$\left. i, j = -\left[\frac{k-1}{2}\right], \ldots, h^{-1} + \left[\frac{k-1}{2}\right]\right\}.$$

When A is a constant coefficient operator, the representation (4.2) and the orthogonality of the Legendre polynomials allows the inner products in the matrices of (3.5), (3.11), or (3.15) to be accumulated by analytical means, in which only Euclidean inner products are performed. By suitable changes of variables, only terms of the form

$$\int_{-1/2}^{1/2} \theta_\nu^{k\sigma}(x)\, \theta_{\nu+\delta x}^{k\beta}(x)\, dx \tag{4.3}$$

need be evaluated, and from (4.1) we obtain, with $\mu = \nu + \delta x$,

$$\int_{-1/2}^{1/2} \theta_\nu^{k\sigma}(x)\, \theta_\mu^{k\beta}(x)\, dx = \sum_{l=1}^{k} \alpha_{l\nu}^{k\sigma} \alpha_{l\mu}^{k\beta}\left(\frac{1}{2l-1}\right).$$

The software has been designed to handle terms corresponding to mesh points near the boundary.

For special cases, e.g. $A = \Delta$, the matrix has a very specific structure if we order the basis functions consecutively along horizontal rows: $\phi_s = \psi_k((x - ih)/h) \cdot \psi_k((y - jh)/h)$ with $s = (j-1)\sqrt{M} + i$ and $M = (h^{-1} + k - 1)^2$. If we denote the matrix problem of (3.5) by $Tc = d$, the matrix T is a band matrix with upper band width equal to $(k-1)\sqrt{M} + k$. Dividing the matrix into a block structure with M square blocks, each block assumes the same banded structure element by element as the overall matrix assumes block by block. In particular, $t_{r+\mu,s+\mu} = t_{rs}$ if both supp $\{\phi_{r+\mu} \cap \partial R\}$ and supp $\{\phi_r \cap \partial R\}$ are empty. Borrowing from a definition of Strang [20], we might call this structure quasi-convolution form; in this regard the matrix behaves quite like a finite-difference matrix.

The data terms d must be accumulated by numerical integration. We have used Romberg quadrature to ensure as much accuracy as desired; in practice, quadratures using a small number of function evaluations would be used. Herbold [12] and Fix [10] treat the problem of selecting "consistent" quadratures for Rayleigh-Ritz schemes.

Various techniques have been used to solve the linear systems; we finally selected a Cholesky decomposition modified for band matrices as being most efficient for reasonably small ($M \leqslant 200$) problems.

The error quantities of interest in our experiments are $\|e\|_0$, the L_2 error, where $e = u - w$, and $\|e\|_\infty$, the supremum norm error for which sharp estimates do not exist. In attempting to determine the order of accuracy of a method, we assume that $\|e\| = Ch^\lambda$ as $h \longrightarrow 0$ and wish to determine λ. The quantity which we actually compute is the error reduction

$$\lambda_{ji} = \log(\|e(h_i)\|/\|e(h_j)\|)/\log(h_i/h_j),$$

where h_i and h_j are different mesh spacings. In [19], our tables present the computations of λ_{ji} for all possible combinations of i and j; for brevity here we shall only tabulate $\lambda_{j,j+1}$.

We shall present evidence that the pointwise error exhibits oscillatory behavior. This requires that we estimate $\|e\|_0$ by Simpson's rule using at least eight points between mesh points; we use these same points to estimate $\|e\|_\infty$.

5. The Poisson Problem on the Unit Square. We now present the results of several computational experiments with the methods described in Section 3. We have selected for presentation here only a portion of the experimental results found in [19]. With the exception of the least-squares method for the square, we have no theoretical foundation for any of our results, since all domains considered are polygonal. All experiments have been performed on an IBM 360/65 system. Computations have been done in double precision to minimize roundoff difficulties, unless otherwise noted (Tables 5.5 and 5.7). We consider model Problem 1:

$$\Delta u = 2e^{(x+y)} \quad \text{in } R_s, \qquad u = e^{(x+y)} \quad \text{in } \partial R_s.$$

For the least-squares method, Table 5.1 presents evidence of optimal fourth-order con-

vergence for the subspace $S^h_{2,4}$ of bicubic splines. Along with the more comprehensive sensitivity study in Table 5.2, these results indicate that the least-squares method is not very sensitive to the choice of the boundary weighting γ.

TABLE 5.1. *Problem* 1, *Least Squares, Bicubic Splines, Wide Parameter Range*

γ	h	L_2 Error ($\times 10^{-5}$)	L_2 Reduction	L_∞ Error ($\times 10^{-4}$)	L_∞ Reduction
10	1/6	.414	-------	.352	-------
	1/8	.129	4.06	.113	3.95
	1/10	.0521	4.04	.0466	3.97
	1/12	.025	4.03	.0226	3.97
100	1/6	.396	-------	.158	-------
	1/8	.125	4.00	.0517	3.87
	1/10	.0512	4.00	.0222	3.78
	1/12	.0247	4.00	.0111	3.82
1000	1/6	.392	-------	.163	-------
	1/8	.124	3.99	.0533	3.90
	1/10	.0510	3.99	.0222	3.92
	1/12	.0246	4.00	.0109	3.93

TABLE 5.2. *Problem* 1, *Least Squares, Bicubic Splines, Full Parameter Study*

h	γ	L_2 Error ($\times 10^{-5}$)	L_∞ Error ($\times 10^{-4}$)
1/5	1	1.68	2.06
	4	.979	1.02
	64	.826	.357
	256	.814	.324
	4096	.809	.331
1/8	1	.189	.325
	4	.135	.160
	64	.125	.0559
	256	.125	.0525
	4096	.124	.0536

The results presented in Table 5.3 for the subspace $S^h_{2,6}$ of biquintic splines demonstrate the greatly improved accuracy available if one is willing to pay the price of added bandwidth. The L_2 error reduction is indicative of optimal 6th-order accuracy; the L_∞ reduction does not evidence quite this high an order (using a least-squares fit to plot $\log(L_\infty$ error) as a function of $\log h$, we determine a slope $\lambda = 5.36$). We see no reason why the actual L_∞ error should not also be of 6th order and attribute our numerical results to the effect of roundoff error, which begins to contribute more significantly when our approximation method becomes more accurate. We present only the results for $\gamma = 100$; a similar result is obtained with $\gamma = 1000$.

TABLE 5.3. *Problem* 1, *Least Squares, Biquintic Splines*

γ	h	L_2 Error ($\times 10^{-8}$)	L_2 Reduction	L_∞ Error ($\times 10^{-7}$)	L_∞ Reduction
100	1/3	14.6	------	4.36	------
	1/4	2.48	6.15	.858	5.66
	1/6	.233	5.83	.104	5.20
	1/7	.0956	5.78	.0454	5.38

For the case of biquadratic splines (type $S^h_{2,3}$), we have mentioned that error estimates do not indicate optimal 3rd-order convergence, and our results in Table 5.4 seem to show that only 2nd-order convergence should be expected. Our data for the L_∞ error are anomalous; we know that the L_∞ error cannot be of higher order than that of L_2 error, hence if we take smaller meshes, we predict this order also to go toward 2. Regardless, we see little to recommend the use of biquadratic splines.

TABLE 5.4. *Problem* 1, *Least Squares, Biquadratic Splines*

γ	h	L_2 Error ($\times 10^{-3}$)	L_2 Reduction	L_∞ Error ($\times 10^{-3}$)	L_∞ Reduction
100	1/6	.295	-------	.900	-------
	1/8	.153	2.29	.415	2.69
	1/10	.094	2.17	.224	2.75

We have also used Problem 1 to make one study of the overall conditioning of the methods. Using double precision, we obtain an approximate solution and determine its error $e_d(h)$. We then assume that

$$\|e_d(h)\| = Ch^\lambda + C'\theta_d h^{-\sigma}, \tag{5.1}$$

where σ is the conditioning effect and θ_d is the double-precision unit roundoff error. Similarly, if we compute in single precision, we determine an error $e_s(h)$ satisfying

$$\|e_s(h)\| = Ch^\lambda + C'\theta_s h^{-\sigma}, \tag{5.2}$$

where θ_s is single precision unit roundoff and $\theta_s \gg \theta_d$. If we assume that $C'\theta_d h^{-\sigma}$ is negligible, the error "reduction" in Table 5.5 is essentially σ in $\Gamma(h) = \|e_s(h)\| - \|e_d(h)\| \approx C'\theta_s h^{-\sigma}$. Our results are not extensive, but they do evidence the $O(h^{-4})$ conditioning for the least-squares method. A further conditioning study will be discussed for one-dimensional problems in Section 7.

TABLE 5.5. *Problem* 1, *Least Squares, Bicubic Splines, Conditioning Study (Single Precision Computation)*

γ	h	L_2 Error ($\times 10^{-4}$)	"Reduction"	L_∞ Error ($\times 10^{-3}$)	"Reduction"
100	1/6	.807	---------	.158	---------
	1/12	16.0	−4.31	3.11	−4.30

Proceeding on to similar studies with Method K, the results in Table 5.6 again demonstrate the optimal 4th-order convergence with bicubic splines; we omit the tabulation of the results for biquintic splines but mention that 6th-order convergence in L_2 error is shown.

TABLE 5.6. *Problem* 1, *Method K, Bicubic Splines*

γ	h	L_2 Error ($\times 10^{-6}$)	L_2 Reduction	L_∞ Error ($\times 10^{-5}$)	L_∞ Reduction
10	1/6	3.96	------	4.08	------
	1/8	1.19	4.16	1.31	3.95
	1/10	.474	4.14	.541	3.96
	1/12	.224	4.12	.262	3.97
1000	1/6	3.57	-------	1.55	-------
	1/8	1.11	4.06	.502	3.91
	1/10	.449	4.06	.209	3.93
	1/12	.214	4.06	.102	3.94

The conditioning study in Table 5.7 for Method K indicates that it experiences only an $O(h^{-2})$ deterioration due to roundoff.

TABLE 5.7. *Problem* 1, *Method K, Bicubic Splines, Conditioning Study (Single Precision Computation)*

γ	h	L_2 Error ($\times 10^{-4}$)	"Reduction"
100	1/6	.266	----------
	1/8	.381	−1.25
	1/10	.696	−2.70
	1/12	1.02	−2.08

In contrast to the least-squares method, Table 5.8 shows that even with quadratic splines Method K yields optimal 3rd-order accuracy.

TABLE 5.8. *Problem* 1, *Method K, Biquadratic Splines*

γ	h	L_2 Error ($\times 10^{-4}$)	L_2 Reduction	L_∞ Error ($\times 10^{-3}$)	L_∞ Reduction
100	1/6	1.97	------	.844	-------
	1/8	.812	3.08	.385	2.73
	1/10	.404	3.14	.207	2.79

As a final study with Problem 1, we present in Table 5.9 the results obtained with Nitsche's method, and note that we get the best overall results via this technique (compare with Tables 5.1 and 5.6).

TABLE 5.9. *Problem* 1, *Nitsche's Method, Bicubic Splines*

γ	h	L_2 Error ($\times 10^{-6}$)	L_2 Reduction	L_∞ Error ($\times 10^{-5}$)	L_∞ Reduction
100	1/6	2.86	-------	1.02	-------
	1/8	.929	3.90	.368	3.55
	1/10	.387	3.93	.163	3.66

We have used Problem 2:

$$\Delta u = 6xye^{x+y}(xy + x + y - 3) \quad \text{in } R_s,$$
$$u = 0 \quad \text{on } \partial R_s,$$

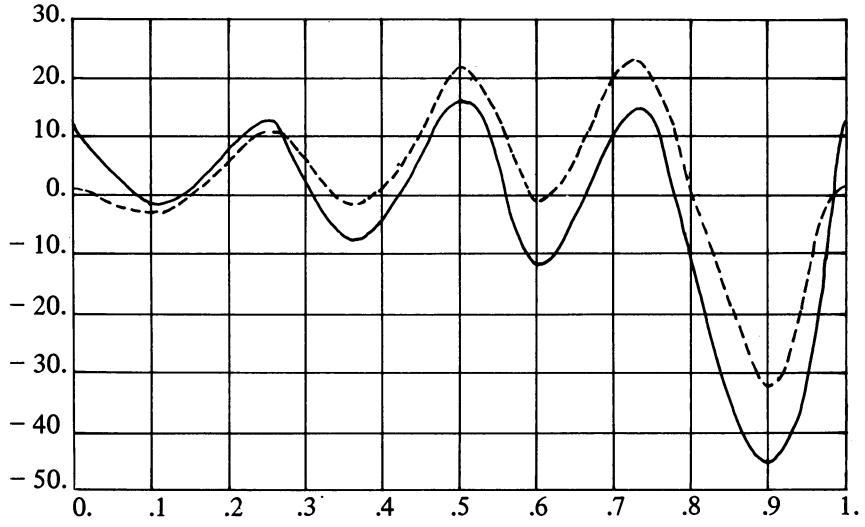
which has solution $u = 3xye^{x+y}(1-x)(1-y)$ to compare least-squares with the Rayleigh-Ritz method in which the approximating functions (bicubic splines) satisfy the homogeneous boundary conditions. In Table 5.10 the results of Herbold-Varga [12] via Rayleigh-Ritz are noted, while our computations via least squares with two boundary weightings are also included. The data indicate that with adequate weighting of boundary terms, one need not be troubled with satisfying boundary conditions to achieve accurate approximations, whereas underweighting the boundary ($\gamma = 1$) produces less desirable results.

TABLE 5.10. *Problem 2, Comparison of Rayleigh-Ritz and Least-Squares L_∞ Error* ($\times 10^{-4}$)

h	Rayleigh-Ritz	Order	Least Squares ($\gamma = 1$)	Order	Least Squares ($\gamma = 16$)	Order
1/3	10.8	-------	---------	-------	-------	-------
1/4	3.57	3.85	21.3	-------	5.36	-------
1/5	1.53	3.80	7.40	4.74	2.23	3.93
1/6	.766	3.80	---------	-------	------	-------
1/7	.419	3.91	---------	-------	------	------
1/8	---------	------	.753	4.86	.342	3.99
1/10	---------	------	.251	4.92	.144	3.89

We also use Problem 2 to illustrate the oscillatory behavior of the pointwise error in the least-squares approximation by presenting in Figure 5.1 a plot of the error on the cross-section at $y = .5$ for two different boundary weights $\gamma = 1$ (solid line) and $\gamma = 32$ (broken line). Notice how the larger weight has forced the boundary condition to be more nearly satisfied.

FIGURE 5.1. *Pointwise Error* ($\times 10^{-5}$) *at $y = .5$, Problem 2, Least Squares*)



We shall refer to Problem 3:

$$\Delta u = \pi \text{ in } R_s \quad \text{with solution}$$

$$u = xy \ln(x^2 + y^2) + (x^2 - y^2)\tan^{-1} y/x + (\pi/2)y^2$$

and boundary data determined accordingly as the "singular" problem, although the singularity actually occurs in the third derivatives of the solution (that is, $u \in H^{3-\epsilon}$). Thus, we may only anticipate from (3.6) at most 3rd-order convergence in L_2, for any $S^h_{2,k}$ space, $k \geqslant 4$. Indeed, in Table 5.11, we see that the L_2 error is clearly reduced as $O(h^{-3})$ for a wide range of parameters. Interestingly, the reduction in L_∞ is clearly second order; it appears that for nonsmooth solutions the L_∞ estimates obtainable via Sobolev's lemma (see, for example, [1]) may be sharp.

TABLE 5.11. *Problem* 3, *Least Squares, Bicubic Splines*

γ	h	L_2 Error ($\times 10^{-4}$)	L_2 Reduction	L_∞ Error ($\times 10^{-3}$)	L_∞ Reduction
4	1/5	1.23	-------	3.57	-------
	1/8	.301	3.00	1.39	2.00
	1/10	.154	3.00	.892	2.00
64	1/5	.903	-------	1.86	-------
	1/8	.221	3.00	.724	2.00
	1/10	.113	3.00	.464	2.00
1024	1/5	.410	-------	.699	-------
	1/8	.174	3.00	.273	2.00
	1/10	.089	3.00	.175	2.00

We mention that we obtain exactly the same convergence orders using quintic splines and that the same behavior is evidenced with Method K.

We also investigate the error on a subdomain (namely, (½, 1) × (½, 1)) away from the origin, at which the singularity in the 3rd derivative of the solution occurs. Interestingly, the error reduction appears (Table 5.12) to be the optimal 4th order for bicubic splines; a similar experiment with biquintics yields approximately 6th-order reduction. Hence, the effect of the singularity is not felt globally.

TABLE 5.12. *Problem* 5, *Least Squares, Bicubic Splines, Subdomain Error*

γ	h	L_2 Error ($\times 10^{-6}$)	L_2 Reduction	L_∞ Error ($\times 10^{-5}$)	L_∞ Reduction
100	1/4	5.35	-------	5.49	-------
	1/6	.978	4.19	.929	4.38
	1/8	.308	4.01	.273	4.26
	1/10	.127	3.98	.110	4.08

6. Other Domains. We define Problem 4:

$$\Delta u + \epsilon u_{xy} = (2 + \epsilon)e^{x+y} \quad \text{in } R_s,$$
$$u = e^{x+y} \quad \text{on } \partial R_s.$$

We may consider this to be merely a problem in which a mixed second partial deriva-

tive occurs, or as a simulation of a Poisson problem on a parallelogram $R_{p,\alpha}$ with angle at the origin π/α $(1 < \alpha < 2)$ in (x', y') coordinates by changing variables: $\delta = -\tan \pi/\alpha$, $x = \delta x' + y'$, $y = \sqrt{1 + \delta^2}y'$, $\epsilon = 2/\sqrt{1 + \delta^2}$. We determine that $\pi/\alpha = \sec^{-1}(2/\epsilon)$. The least-squares theory of [7] yields no error estimates. We present here only results corresponding to $\epsilon = \sqrt{2}$ $(\alpha = 3/4)$ with bicubic splines as the approximating functions. The results for least squares and Method K appear in Tables 6.1 and 6.2, respectively.

Notice that while sensitivity with respect to γ is not too marked for least squares, there is some decrease in error reduction for $\gamma = 256$, the largest weight chosen. Contrasting, the L_2 error reductions of Method K are definitely 4th order (as proved in [6]) with only slight sensitivity noted. Similar experiments with $\epsilon = \sqrt{3}$ produce even more pronounced evidence that the boundary term should not be overweighted in least squares as the operator becomes less elliptic $(\epsilon \longrightarrow 2)$, while Method K shows no such difficulty.

TABLE 6.1. *Problem 4, Least Squares, Bicubic Splines*

γ	h	L_2 Error $(\times 10^{-6})$	L_2 Reduction	L_∞ Error $(\times 10^{-5})$	L_∞ Reduction
4	1/4	24.6	--------	9.00	-------
	1/8	1.49	4.04	.583	3.95
	1/16	.0883	4.08	.040	3.86
64	1/4	21.1	-------	7.63	-------
	1/8	1.25	4.08	.520	3.87
	1/16	.0814	3.94	.034	3.93
256	1/4	20.1	-------	6.23	-------
	1/8	1.52	3.72	.562	3.47
	1/16	.146	3.39	.050	3.48

TABLE 6.2. *Problem 4, Method K, Bicubic Splines*

γ	h	L_2 Error $(\times 10^{-5})$	L_2 Reduction	L_∞ Error $(\times 10^{-4})$	L_∞ Reduction
64	1/4	1.95	-------	.720	-------
	1/6	.368	4.11	.151	3.86
	1/8	.113	4.09	.049	3.89
256	1/4	1.87	------	.623	-------
	1/6	.367	4.02	.135	3.77
	1/8	.114	4.06	.046	3.76

Our next experiment is a true departure from the square. We study two problems on an L-shaped domain R_L, i.e., a six-sided rectilinear domain with one interior angle $3\pi/2$ and five interior angles $\pi/2$. For the problem $\Delta u = 2e^{x+y}$ in $R_L = R_s \backslash \{(\frac{1}{2}, 1) \times (0, \frac{1}{2})\}$, $u = e^{x+y}$ on ∂R_L, we present in Tables 6.3 and 6.4 evidence that even for this notoriously difficult domain on which to compute we obtain optimal convergence.

TABLE 6.3. *L-Shaped Domain, Least Squares, Bicubic Splines, Exponential Data*

γ	h	L_2 Error ($\times 10^{-5}$)	L_2 Reduction	L_∞ Error ($\times 10^{-5}$)	L_∞ Reduction
100	1/4	1.98	-------	7.49	-------
	1/6	.371	4.13	1.57	3.85
	1/8	.124	3.81	.518	3.86
	1/10	.049	4.10	.223	3.78

TABLE 6.4. *L-Shaped Domain, Method K, Bicubic Splines, Exponential Data*

γ	h	L_2 Error ($\times 10^{-5}$)	L_2 Reduction	L_∞ *Error* ($\times 10^{-5}$)	L_∞ Reduction
100	1/4	1.83	-------	7.23	-------
	1/6	.335	4.19	1.48	3.91
	1/8	.105	4.03	.484	3.89
	1/10	.042	4.14	.205	3.85

In order to study a "singular" problem, we orient $R_L = R_S \backslash \{(0, \tfrac{1}{2}) \times (0, \tfrac{1}{2})\}$. If we consider the problem $\Delta u = \pi$ in R_L, $u(\tfrac{1}{2}, y) = 0$ on $0 \leqslant y \leqslant \tfrac{1}{2}$, $u(x, 0) = 0$ on $\tfrac{1}{2} \leqslant x \leqslant 1$, with the other boundary data computable from the solution

$$u(x, y) = (x - \tfrac{1}{2})y \ln[(x - \tfrac{1}{2})^2 + y^2] + [(x - \tfrac{1}{2})^2 - y^2] \tan^{-1} \frac{y}{x - \tfrac{1}{2}} + \frac{\pi}{2} y^2,$$

we have placed the singularity at $(\tfrac{1}{2}, 0)$, not the reentrant corner. Our results in Tables 6.5 and 6.6 for least squares and Method K respectively show that we obtain the same 3rd-order L_2 reductions and 2nd-order L_∞ reductions for this domain as in the case of the square (Problem 3, Section 4). Even so, it may be that for smaller meshes the error reductions will evidence a pollution due to the singularity.

TABLE 6.5. *L-Shaped Domain, Least Squares, Bicubic Splines, Singularity Away From Reentrant Corner*

γ	h	L_2Error ($\times 10^{-4}$)	L_2 Reduction	L_∞ Error ($\times 10^{-3}$)	L_∞ Reduction
100	1/4	1.79	-------	2.29	-------
	1/6	.522	3.04	1.03	1.97
	1/8	.226	2.92	.583	1.99

TABLE 6.6. *L-Shaped Domain, Method K, Bicubic Splines, Singularity Away from Reentrant Corner*

γ	h	L_2 Error ($\times 10^{-4}$)	L_2 Reduction	L_∞ Error ($\times 10^{-3}$)	L_∞ Reduction
100	1/4	1.68	--------	2.32	--------
	1/6	.496	3.02	1.05	1.97
	1/8	.208	3.02	.592	1.99

Finally, an attempt was made to place the singularity right at the reentrant corner. For both methods, the evidence is that overweighting the boundary term amplifies the ill effect of the geometry and optimal order of convergence is *not* evidenced.

7. **One-Dimensional Studies; Conditioning Effects.** Quite obviously the same methods we have been considering here can be applied to linear boundary value problems of ordinary differential equations. On these problems it is economically feasible to consider quite small mesh sizes and hence, recalling from (5.1) that $\|e_d(h)\| = Ch^{\lambda} + C'\theta_d h^{-\sigma}$, we may allow h to become small enough that the error "reductions" are actually estimates of σ.

Since we have not included any discussion of the biharmonic problem, we shall only mention here that an analogous fourth-order boundary value problem has been studied, and that least squares evidences $O(h^{-8})$ conditioning, while a scheme like Method K shows only $O(h^{-4})$ deterioration.

Our main concern is with second-order boundary value problems and we shall study

$$\begin{aligned} -u'' + Cu &= f \quad \text{on } (0, 1),\ C > 1, \\ u(0) &= u_0, \qquad u(1) = u_1. \end{aligned} \tag{7.1}$$

In particular, we take $C = 1$, $u = e^{4x}$ and $f = -15e^{4x}$. If we define $A\phi = -\phi'' + C\phi$, then by analogy to (3.2), if we define

$$L(\psi, \chi) = (A\psi, A\chi) + \gamma h^{-3}[\psi(1)\chi(1) + \psi(0)\chi(0)] \tag{7.2}$$

then u satisfies

$$L(u, \chi) = (f, A\chi) + \gamma h^{-3}[u_1\chi(1) + u_0\chi(0)] \quad \text{for all } \chi \in H^2(R) \tag{7.3}$$

and hence the least-squares method is:

Find $w \in S^h_{2,k}$ such that

$$L(w, \chi) = (f, A\chi) + \gamma h^{-3}[u_1\chi(1) + u_0\chi(0)] \quad \text{for all } \chi \in S^h_{2,k}. \tag{7.4}$$

Similarly, if we define

$$K(\psi, \chi) = h^3 L(\psi, \chi) + h[(A\psi, \chi) - (\psi, \chi'') - (\psi', \chi')] \tag{7.5}$$

then the analog of Method K is to find $w \in S^h_{2,k}$ such that

$$\begin{aligned} K(w, \chi) = (f, h^3 A\chi + \chi) &+ [u_1\chi(1) + u_0\chi(0)] \\ &+ h[u_0\chi'(0) - u_1\chi'(1)] \quad \text{for all } \chi \in S^h_{2,k}. \end{aligned} \tag{7.6}$$

The results presented in Table 7.1 for least squares with biquintic splines show optimal 6th-order error reduction before roundoff sets in, at which time the indication of $O(h^{-4})$ conditioning is evident. The data for Method K in Table 7.2 also show optimal convergence and roundoff becomes a problem only for much smaller h, reflective of the $O(h^{-2})$ conditioning. Notice that even when roundoff appears, the approximations are extremely accurate.

TABLE 7.1. *Two-Point Boundary Value Problem, Least Squares, Biquintic Splines*

γ	h	L_2 Error	L_2 Reduction	L_∞ Error	L_∞ Reduction
100	1/10	$.188 \times 10^{-5}$	------	$.730 \times 10^{-5}$	------
	1/25	$.790 \times 10^{-8}$	5.97	$.404 \times 10^{-7}$	5.67
	1/50	$.121 \times 10^{-9}$	6.03	$.691 \times 10^{-9}$	5.87
	1/75	$.172 \times 10^{-9}$	−0.86	$.248 \times 10^{-9}$	2.52
	1/100	$.560 \times 10^{-9}$	−4.12	$.791 \times 10^{-9}$	−4.02

TABLE 7.2. *Two-Point Boundary Value Problem, Method K, Biquintic Splines*

γ	h	L_2 Error	L_2 Reduction	L_∞ Error	L_∞ Reduction
100	1/25	$.786 \times 10^{-8}$	-------	$.396 \times 10^{-7}$	-------
	1/50	$.121 \times 10^{-9}$	6.03	$.679 \times 10^{-9}$	5.87
	1/75	$.110 \times 10^{-10}$	5.90	$.619 \times 10^{-10}$	5.91
	1/100	$.678 \times 10^{-11}$	1.70	$.119 \times 10^{-10}$	5.72
	1/125	$.109 \times 10^{-10}$	−2.13	$.158 \times 10^{-10}$	−1.26

8. Conclusions; Further Investigations. We have clearly demonstrated the potential of these approximation methods, perhaps especially Method K, as practically applicable schemes, and have shown that we need not worry too much about sensitivity to boundary weighting or effects of ill-conditioning. In [19], we have considered several additional classes of problems, e.g., biharmonic problems, general second-order constant coefficient equations, several methods for parabolic problems, some penalty methods of Babuška [3], [22], [23], and some methods involving indefinite bilinear forms due to Schatz. We are currently considering problems with variable coefficients via the least-squares method and the application of these ideas to periodic boundary value problems.

Department of Mathematics
University of Tennessee
Knoxville, Tennessee 37916

1. S. AGMON, *Lectures on Elliptic Boundary Value Problems*, Van Nostrand Math. Studies, no. 2, Van Nostrand, Princeton, N. J., 1965. MR 31 #2504.

2. I. BABUŠKA, "Approximation by Hill functions," *Comment. Math. Univ. Carolinae*, v. 11, 1970, pp. 787–811. MR 45 #1396.

3. I. BABUŠKA, *Numerical Solution of Boundary Value Problems by the Perturbed Variational Principle*, Univ. of Maryland Tech. Note BN-624, Oct. 1969.

4. G. BAKER, "Simplified proofs of error estimates for the least squares method for Dirichlet's problem," *Math. Comp.*, v. 27, 1973, pp. 229–235.

5. J. H. BRAMBLE & S. HILBERT, "Estimation of linear functionals on Sobolev spaces with application to Fourier transforms and spline interpolation," *SIAM J. Numer. Anal.*, v. 7, 1970, pp. 112–124. MR 41 #7819.

6. J. H. BRAMBLE & J. A. NITSCHE, "A generalized Ritz-least-squares method for Dirichlet problems," *SIAM J. Numer. Anal.*, v. 10, 1973, pp. 81–93. MR 47 #2836.

7. J. H. BRAMBLE & A. H. SCHATZ, "Rayleigh-Ritz-Galerkin methods for Dirichlet's problem using subspaces without boundary conditions," *Comm. Pure Appl. Math.*, v. 23, 1970, pp. 653–675. MR 42 #2690.

8. J. H. BRAMBLE & M. ZLÁMAL, "Triangular elements in the finite element method," *Math. Comp.*, v. 24, 1970, pp. 809–820. MR 43 #8250.

9. F. DiGUGLIELMO, "Construction d'approximations des espaces de Sobolev sur des réseaux en simplexes," *Calcolo*, v. 6, 1969, pp. 279–331.

10. G. FIX, "Effects of quadrative error in the finite element method," *Proc. of the Second Japan-U.S. Symposium on Matrix Methods in Structural Mechanics*, University of Alabama Press, 1972.

11. G. FIX & K. LARSEN, "On the convergence of SOR-iterations for finite element approximations to elliptic boundary value problems," *SIAM J. Numer. Anal.*, v. 8, 1971, pp. 536–547. MR 45 #2935.

12. R. J. HERBOLD & R. S. VARGA, "The effect of quadrature errors in the numerical solution of two-dimensional boundary value problems by variational techniques," *Aequationes Math.*, v. 7, 1972, pp. 36–58. MR 45 #8028.

13. S. HILBERT, *Numerical Methods for Elliptic Boundary Value Problems*, Ph.D Thesis, University of Maryland, College Park, Md., 1969.

14. J.-L. LIONS & E. MAGENES, *Problèmes aux Limites Non Homogènes et Applications.* Vol. 1, Travaux et Recherches Mathématiques, no. 17, Dunod, Paris, 1968. MR 40 #512.

15. J. NITSCHE, "Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die Keinen Randbedingungen unterworfen sind," *Abh. Math. Sem. Univ. Hamburg*, v. 36, 1970/71, pp. 9–15.

16. I. SCHOENBERG, "On cardinal spline interpolation and a summability method for the cardinal series," Lecture Notes, *Symposium on Approximation*, Univ. of Cincinnati, Cincinnati, Ohio, 1969.

17. M. H. SCHULTZ, "Rayleigh-Ritz-Galerkin methods for multidimensional problems," *SIAM J. Numer. Anal.*, v. 6, 1969, pp. 523–538. MR 41 #7859.

18. J. SEGETHOVÁ, "Numerical construction of the hill functions," *SIAM J. Numer. Anal.*, v. 9, 1972, pp. 199–204. MR 46 #4682.

19. S. M. SERBIN, *A Computational Investigation of Least Squares and Other Projection Methods for the Approximate Solution of Boundary Value Problems*, Ph.D Thesis, Cornell University, Ithaca, N. Y.. 1971.

20. G. STRANG, "The finite element method and approximation theory,"*Numerical Solution of Partial Differential Equations*, II (SYNSPADE, 1970) (Proc. Sympos., Univ. of Maryland, College Park, Md., 1970), Academic Press, New York, 1971, pp. 547–583. MR 44 #4926.

21. R. B. KELLOGG, "Higher order singularities for interface problems," *The Mathematical Foundations of the Finite Element Method With Applications to Partial Differential Equations*, (A. K. Aziz, editor), Academic Press, New York, 1972.

22. I. BABUŠKA, "The finite element method for elliptic equations with discontinuous coefficients," *Computing*, v. 5, 1970, pp. 207–213.

23. I. BABUŠKA, "The finite element method with penalty," *Math. Comp.*, v. 27, 1973, pp. 221–228.