Finite Element Methods Are Not Always Optimal

ARTHUR G. WERSCHULZ

Division of Science and Mathematics
Fordham University / College at Lincoln Center
113 West 60th Street
New York, NY 10023

Department of Computer Science
Columbia University
New York, NY 10027

CUCS-178-84

June, 1984

ABSTRACT. Consider a regularly-elliptic $2m$-th order boundary problem $Lu = f$ with $f \in H^r(\Omega)$, $r \geq -m$. In previous work, we showed that the finite-element method (FEM) using piecewise polynomials of degree k is asymptotically optimal when $k \geq 2m - 1 + r$. In this paper, we show that the FEM is not asymptotically optimal when this inequality is violated. However, there exists an algorithm, the Traub-Wasilkowski-Woźniakowski spline algorithm, which is always optimal. Moreover, the error of the FEM can be arbitrarily larger than the error of the spline algorithm. We also obtain a necessary and sufficient condition for a Galerkin method (or a generalized Galerkin method) to be a spline algorithm.

Keywords and phrases. Elliptic problems, variational methods, finite element methods, optimal algorithms, computational complexity.

1980 Mathematics subject classifications (Amer. Math. Soc.): Primary: 65N30, 68C25. Secondary: 65J10, 65N05, 65N15.

Division of Science and Mathematics, Fordham University / College at Lincoln Center, 113 West 60th Street, New York, NY 10023

Department of Computer Science, Columbia University, New York, NY 10027

This research was supported in part by the National Science Foundation under Grants MCS-8203271 and MCS-8303111.

1. Introduction

This paper deals with the optimal solution of 2m-th order regularly-elliptic boundary-value problems $Lu = f$ with $f \in H^r(\Omega)$, $\Omega \subseteq \mathbb{R}^N$. We consider the variational form of such problems having homogeneous boundary conditions (see Section 2). We wish to solve such problems using information of cardinality at most n. (In this Introduction, we have to use words such as information, cardinality, algorithm, etc., without definition; they are defined rigorously in Section 3.)

In [10], we showed that the optimal energy-norm error of an algorithm using information of cardinality n is $\Theta(n^{-(m+r)/N})$ as $n \to \infty$.* Moreover, this optimal error is achieved by a finite-element method (FEM) using piecewise polynomials of degree k, where $k \geq 2m - 1 + r$. Suppose that this inequality is violated. For instance, suppose we have a program written using piecewise-linear polynomials to solve a second-order problem in a planar region Ω. For $f \in L_2(\Omega)$, this FEM has H^1-error $\Theta(n^{-1/2})$, which is optimal. What happens when we use this program to solve a problem with (say) $f \in H^1(\Omega)$? Does the error of the FEM improve when f has additional smoothness, and if not, is there a method using the same information as the FEM, but with better error?

---

*Here and in what follows, we use the Ω- and Θ-notations of Knuth [5], as well as the usual O-notation. That is,

$$f = \Omega(g) \quad \text{if} \quad g = O(f)$$

and

$$f = \Theta(g) \quad \text{if} \quad f = O(g) \quad \text{and} \quad f = \Omega(g).$$

In Section 4, we show that the error of the FEM is $\Theta(n^{-\mu/N})$ as $n \to \infty$, where $\mu = \min(k + 1 - m, m + r)$, so that $k \geq 2m - 1 + r$ is a necessary and sufficient condition for the FEM to be asymptotically optimal. On the other hand, we analyze the Traub-Wasilkowski-Wozniakowski spline algorithm (see [9]) which uses the same information as the FEM. We show that the error of the spline algorithm is $\Theta(n^{-(m+r)/N})$ as $n \to \infty$, regardless of whether $k \geq 2m - 1 + r$; it is therefore always asymptotically optimal. Moreover, unlike the FEM, the proof of the error estimate for the spline algorithm does not require the "shift theorem"; hence, the spline algorithm is applicable to a wider range of problems.

The optimality result mentioned above is for a worst-case f. Also of interest is the local error, i.e., the error for any particular f. The spline algorithm is known to be strongly optimal, that is, it enjoys optimal local error. It is well-known that the FEM is a Galerkin method. Furthermore, as we show in Section 5, the spline algorithm is a generalized Galerkin method. (Indeed, given the spline algorithm, we show how to realize it as a generalized Galerkin method.) This motivates our interest in the local error of generalized Galerkin methods.

The ratio of the local error of an algorithm to the optimal local error is called the deviation of the algorithm. We pose and solve the question of when a generalized Galerkin method has finite deviation. We show the deviation is finite if and only if the generalized Galerkin method is the spline algorithm.

Do FEM's always have finite deviation? We show the answer is no, by exhibiting an FEM which is not a spline algorithm. We conjecture that no convergent FEM has finite deviation.

In Section 6, we discuss the complexity of obtaining ϵ-approximations. We show that the penalty for using the FEM when $k < 2m - 1 + r$ is unbounded as $\epsilon \to 0$. Since this is an asymptotic measure, it is useful to know whether the spline algorithm has lower complexity than the FEM for moderate values of ϵ. We show that this is indeed the case, by exhibiting a model problem for which the spline algorithm has lower complexity than the FEM for any ϵ roughly less than one-half.

Finally, in Section 7, we briefly discuss implementation of the spline algorithm, and ask whether it is practical to use.

2. The Variational Boundary-Value Problem

In what follows, we use the standard notations for Sobolev spaces, inner products, and norms, multi-indices, etc. found in Ciarlet [2]. Fractional- and negative-order Sobolev spaces are defined by Hilbert-space interpolation and duality, respectively; see Chapter 2 of [1] and Chapter 4 of [6] for details.

Let Ω be a bounded C^∞ region in $\mathbb{R}^N$. Define the properly elliptic operator

$$Lv := \sum_{|\alpha|,|\beta|\le m} (-1)^{|\alpha|} D^\alpha (a_{\alpha\beta} D^\beta v) \tag{2.1}$$

(with real coefficients $a_{\alpha\beta} \in C^\infty(\bar{\Omega})$ such that $a_{\alpha\beta} = a_{\beta\alpha}$) and a normal family of operators

$$B_j v := \sum_{|\alpha|\le q_j} b_{j\alpha} D^\alpha v \qquad (0 \le j \le m-1), \tag{2.2}$$

(with real coefficients $b_{j\alpha} \in C^\infty(\partial\Omega)$), where

$$0 \le q_0 \le q_1 \le \ldots \le q_{m-1} \le 2m-1, \tag{2.3}$$

which covers L on $\partial\Omega$. Setting

$$m^* := \min\{j : q_j \ge m\}, \tag{2.4}$$

we additionally assume that

$$\{q_j\}_{j=0}^{m^*-1} \cup \{2m-1-q_j\}_{m^*}^{m-1} = \{0,\ldots,m-1\}. \tag{2.5}$$

(See Chapter 3 of [1], Chapter 5 of [6] for further definitions and illustrative examples.)

Let

$$(2.6)\quad H_E^m(\Omega) := \{v \in H^m(\Omega) : B_j v = 0 \quad \text{for} \quad 0 \le j \le m^* - 1\}$$

denote the space of $H^m(\Omega)$-functions satisfying the essential boundary conditions of order at most q_{m^*-1}. We define a symmetric, continuous bilinear form B on $H_E^m(\Omega)$ by

$$(2.7)\qquad B(v,w) := \sum_{|\alpha|,|\beta| \le m} \int_\Omega a_{\alpha\beta} D^\alpha v D^\beta w.$$

We additionally assume that B is $H_E^m(\Omega)$-coercive, so that B is an inner-product on $H_E^m(\Omega)$, yielding a norm $\|\cdot\|_B$ defined by

$$(2.8)\qquad \|v\|_B := B(v,v)^{1/2}$$

which is equivalent to the norm $\|\cdot\|_m$ on $H_E^m(\Omega)$.

We now define the <u>variational boundary-value problem</u> as follows. Let $r \ge -m$. Given $f \in H^r(\Omega)$, find $u = Sf \in H_E^m(\Omega)$ satisfying

$$(2.9)\qquad B(u,v) = (f,v)_0 = \int_\Omega fv \qquad \forall\ v \in H_E^m(\Omega).$$

From the Lax-Milgram lemma, S is a Hilbert space isomorphism of $H^{-m}(\Omega)$ onto $H_E^m(\Omega)$, so that $S : H^r(\Omega) \to H_E^m(\Omega)$ is a bounded linear operator.

It is useful to recall the "shift theorem" (Chapter 3 of [1], Chapter 5 of [6]) which states that since $f \in H^r(\Omega)$, we have $Sf \in H_E^m(\Omega) \cap H^{2m+r}(\Omega)$; moreover, there exists a positive constant σ, independent of f, such that

(2.10) $$\sigma^{-1}\|Sf\|_{2m+r} \le \|f\|_r \le \sigma\|Sf\|_{2m+r}.$$

If $r > N/2$, the shift theorem, Sobolev's embedding theorem, and an m-fold integration by parts yield that $u = Sf$ is a classical solution to the problem of finding $u : \overline{\Omega} \to \mathbb{R}$ satisfying

(2.11) $$\begin{aligned} Lu &= f \quad \text{in } \Omega \\ B_j u &= 0 \quad \text{on } \partial\Omega \quad (0 \le j \le m-1). \end{aligned}$$

3. Information and Algorithms

In this section, we define a number of the concepts mentioned in the Introduction. Most of the terminology and results are from [9]. As we state these definitions and results, we will illustrate them for the finite element method and Galerkin information.

Recall that we are trying to approximate the transformation $S : H^r(\Omega) \to H_E^m(\Omega)$ with $r \geq -m$. Since S is not of finite rank and we wish to use finite algorithms, we are only allowed to sample a finite amount of information about problem elements $f \in H^r(\Omega)$. Here (linear) information of cardinality n is a surjective linear mapping $\mathcal{N} : H^r(\Omega) \to \mathbb{R}^n$, so that we may write

$$\mathcal{N}f = [\lambda_1(f) \ \ldots \ \lambda_n(f)]^T \qquad \forall\, f \in H^r(\Omega) \tag{3.1}$$

where $\lambda_1,\ldots,\lambda_n$ are linearly independent linear functionals on $H^r(\Omega)$. (See Chapter 7 of [9] for a discussion of why we consider only linear information.)

Example 3.1. Let $\mathcal{S}$ be a subspace of $H_E^m(\Omega)$ of dimension n and having basis $\{s_1,\ldots,s_n\}$. Define $\mathcal{N}_{\mathcal{S}} : H^r(\Omega) \to \mathbb{R}^n$ by

$$\mathcal{N}_{\mathcal{S}}f := [(f,s_1)_0 \ \ldots \ (f,s_n)_0]^T \qquad \forall\, f \in H^r(\Omega). \tag{3.2}$$

Then $\mathcal{N}_{\mathcal{S}}$ is linear information of cardinality n. (Conversely, given any linear information $\mathcal{N}$ of cardinality n, one can show that there exists a subspace $\mathcal{S} \subset H_E^m(\Omega)$ of dimension n such that $\mathcal{N} = \mathcal{N}_{\mathcal{S}}$ if and only if $\mathcal{N}$ has an extension to all of $H^{-m}(\Omega)$ which is bounded in the $\|\cdot\|_{-m}$ norm.) We call $\mathcal{N}_{\mathcal{S}}$ the Galerkin information generated by $\mathcal{S}$. ■

In the remainder of this paper, for any Hilbert space H, we denote the unit ball of H by BH, i.e.,

$$BH := \{f \in H : \|f\|_H \leq 1\}. \tag{3.2}$$

In particular, we will let

$$\mathfrak{F}_0 := BH^r(\Omega). \tag{3.3}$$

By an <u>algorithm</u> φ <u>using</u> n, we mean a (possibly nonlinear) mapping $\varphi : D_\varphi \subset n(\mathfrak{F}_0) \to H_E^m(\Omega)$. The (worst-case) <u>error</u> $e(\varphi)$ <u>of</u> φ is given by

$$e(\varphi) := \sup_{f \in \mathfrak{F}_0} \|Sf - \varphi(nf)\|_B. \tag{3.4}$$

(The restriction to $f \in \mathfrak{F}_0$, rather than considering the sup over all $f \in H^r(\Omega)$, is a normalization which is necessary for the error to be finite.) We use the norm $\|\cdot\|_B$ rather than the equivalent norm $\|\cdot\|_m$ for technical reasons, as illustrated in

<u>Example 3.1 (continued)</u>. Define the <u>Galerkin method</u> φ_S <u>using</u> n_S by

$$\varphi_S(n_S f) := u_S, \tag{3.5}$$

where $u_S \in S$ satisfies

$$B(u_S, s) = (f, s)_0 \qquad \forall\ s \in S. \tag{3.6}$$

Then standard results ([1], [2], [7]) yield

$$e(\varphi_S) = \sup_{f \in \mathfrak{F}_0} \inf_{s \in S} \|Sf - s\|_B. \tag{3.7}$$

In particular, let $S = S_n$, where $\{S_n\}_{n=1}^{\infty}$ is a regular family of finite element subspaces of degree k, i.e., S_n is an n-dimensional subspace of $H_E^m(\Omega)$ consisting of piecewise polynomials of degree k over a triangulation $\mathcal{T}_n$ of Ω. Here, $\{\mathcal{T}_n\}_{n=1}^{\infty}$ is regular in the sense of page 132 of [2], which (roughly) means that the subregions do not become geometrically degenerate, and that their diameters tend to zero as $n \to \infty$. (Of course, since Ω is C^{∞}, we must make an additional assumption about boundary elements to guarantee that $S_n \subset H_E^m(\Omega)$; for instance, we may decide to use curved elements as in [3].)

For this case, we write φ_n^* and Λ_n^* rather than φ_S and Λ_S, calling φ_n^* the finite element method (FEM) using S_n. Suppose now that $\{\mathcal{T}_n\}_{n=1}^{\infty}$ is quasi-uniform (see pg. 272 of [6]), which means that the ratio of the diameters of any two subregions in $\mathcal{T}_n$ is bounded, independent of n. Then the standard results ([1],[2].[6]) yield

$$(3.8) \qquad e(\varphi_n^*) = O(n^{-\mu/N}) \text{ as } n \to \infty \qquad \mu = \min(k + 1 - m, m + r).$$

Moreover, results of Strang and Fix [8] imply that the "O" may be changed to "Θ" when $H_E^m(\Omega) = H^m(\Omega)$ (i.e., no essential boundary conditions), the triangulations $\mathcal{T}_n$ are uniform, and $k \leq 2m - 1 + r$. In Section 4 of this paper, we will remove these three restrictions, so that the bound (3.8) is always sharp. ■

Given information Λ of cardinality n, we wish to find the minimum error of an algorithm φ using Λ. In order to do this, let

(3.9) $$Vf := \{\tilde{f} \in \mathfrak{F}_0 : \Lambda\tilde{f} = \Lambda f\} \qquad \forall\ f \in \mathfrak{F}_0.$$

If φ uses Λ, then knowing only Λf, it is impossible for φ to determine which of the elements of the set

(3.10) $$Uf := SVf$$

is being approximated, so that

(3.11) $$e(\varphi) = \sup_{f\in\mathfrak{F}_0} e(\varphi, f)$$

where the <u>local error</u> $e(\varphi, f)$ is given by

(3.12) $$e(\varphi, f) := \sup_{\tilde{f}\in V(f)} \|S\tilde{f} - \varphi(\Lambda f)\|_B \qquad \forall\ f \in \mathfrak{F}_0.$$

Define the <u>local radius</u> rad Uf by

(3.13) $$\text{rad } Uf := \inf_{a\in H_E^m(\Omega)} \sup_{\tilde{f}\in Vf} \|a - S\tilde{f}\|_B \qquad \forall\ f \in \mathfrak{F}_0.$$

As in Chapter 1 of [9], we have

(3.14) $$\inf_{\varphi} e(\varphi, f) = \text{rad } Uf,$$

so that

(3.15) $$\inf_{\varphi} e(\varphi) = r(\Lambda) := \sup_{f\in\mathfrak{F}_0} \text{rad } Uf,$$

where $r(\Lambda)$ is called the <u>radius of information</u> Λ. In our Hilbert space setting, one can show that

(3.16) $$r(\Lambda) = \sup_{z\in\mathfrak{F}_0 \cap \ker \Lambda} \|Sz\|_B$$

(see Chapter 2 of [9]).

Example 3.1 (continued). In Section 4, we will show that for quasi-uniform $\{\mathcal{T}_n\}_{n=1}^{\infty}$,

$$r(n_n^*) = \Theta(n^{-(m+r)/N}) \text{ as } n \to \infty. \tag{3.17}$$

Hence the FEM φ_n^* has (asymptotically) optimal error using n_n (as $n \to \infty$) if and only if $k \geq 2m - 1 + r$. ■

Remark 3.1. Now that we know the minimal local and worst-case errors of algorithms using n, it is useful to find algorithms achieving these minima. Let $P : H^r(\Omega) \to H^r(\Omega)$ denote the orthogonal projector onto $(\ker n)^{\perp}$. Define the (Traub-Wasilkowski-Woźniakowski) spline algorithm φ^s (Chapter 4 of [9]) by

$$\varphi^s(nf) := Spf \qquad \forall\ f \in \mathcal{F}_0. \tag{3.18}$$

One may check that φ^s is well-defined, and that

$$e(\varphi^s, f) = \text{rad } Uf \qquad \forall\ f \in \mathcal{F}_0, \tag{3.19}$$

which implies that φ^s is an optimal error algorithm, i.e., for any φ using n,

$$e(\varphi^s) = r(n) \leq e(\varphi). \tag{3.20}$$

Not only is φ^s an optimal error algorithm, but it is a strongly optimal error algorithm, that is,

$$e(\varphi^s, f) \leq e(\varphi, f) \qquad \forall\ f \in \mathcal{F}_0. \tag{3.21}$$

We will discuss FEM's and spline algorithms more fully in Section 5. ■

Just as we can ask which algorithm makes optimal use of given information, one can ask which information of a given cardinality is best. Let

$$r(n) := \inf\{r(\mathcal{N}) : \mathcal{N} \text{ is of cardinality at most } n\} \tag{3.22}$$

denote the nth minimal radius of information; we say that $\mathcal{N}$ of cardinality at most n is an nth optimal information if

$$r(\mathcal{N}) = r(n). \tag{3.23}$$

Then (Chapter 2 of [9])

$$r(n) = d_n(S(\mathcal{F}_0), H_E^m(\Omega)), \tag{3.24}$$

where the Kolmogorov n-width of a balanced subset X of a Hilbert space H with norm $\|\cdot\|_H$ is given by

$$d_n(X,H) := \inf\{\sup_{x\in X} \inf_{y\in A_n} \|x - y\|_H : A_n \text{ subspace of } H, \dim A_n \le n\}. \tag{3.25}$$

Example 3.1 (continued). Results from [10] yield that

$$r(n) = \Theta(n^{-(r+m)/N}) \text{ as } n \to \infty. \tag{3.26}$$

Hence, the results in Section 4 will imply that $\mathcal{N}_n^*$ is (asymptotically, as $n \to \infty$) an nth optimal information. ■

4. Worst-Case Error of the FEM and the Spline Algorithm

In this section, we show that $k \geq 2m - 1 + r$ is a necessary and sufficient condition for an FEM to have optimal error to within a constant, independent of n. We also show that the spline method is an optimal error algorithm using the nth optimal (to within a constant) information $\mathcal{N}_n^*$, regardless of whether $k \geq 2m - 1 + r$.

Recall that S_n is an n-dimensional subspace of $H_E^m(\Omega)$ consisting of piecewise polynomials of degree k from a triangulation $\mathcal{T}_n$ of Ω. We first show

Lemma 4.1. $k \geq m$.

Proof: Suppose on the contrary that $k \leq m - 1$. Since, for any $s \in S_n$, $s \in H^m(\Omega)$ and $s|_K \in C^\infty(K)$ for each $K \in \mathcal{T}_n$, an obvious extension of Theorem 4.2.1 of Ciarlet [2] yields that $S_n \subset C^{m-1}(\Omega)$. Choose $s \in S_n$. Let K_1, K_2 be adjacent elements in the triangulation, let

$$F := \partial K_1 \cap \partial K_2 \tag{4.1}$$

and let

$$s_i := s|_{K_i} \quad (i = 1,2). \tag{4.2}$$

Let $s^* \in P_k$ satisfy $s^* = s_1$ on K_1; that is, s^* is s_1, but treated as a polynomial over Ω rather than over K_1. Pick a point p on F, and draw a ξ-axis G_p perpendicular to F through p. Hence there is an affine transformation $F_p : \mathbb{R} \to \mathbb{R}^N$ which is a bijection of $\mathbb{R}$ onto G_p, such that $p = F_p(0)$.

For $i = 1$ and $i = 2$, define

$$\sigma_i(\xi) := s_i(F_p(\xi)) \tag{4.3}$$

so that

$$\sigma_i(\xi) = s_i(x) \text{ for } x = F_p(\xi). \tag{4.4}$$

Since

$$s_i \in P_k \text{ and } F_p \text{ is affine.} \tag{4.5}$$

we see that

$$\sigma_i(\xi) \text{ is a polynomial of degree at most } k \text{ in } \xi. \tag{4.6}$$

On the other hand, since $s \in C^{m-1}(\Omega)$, we must have

$$\sigma_1^{(j)}(0) = \sigma_2^{(j)}(0) \qquad (0 \leq j \leq m-1). \tag{4.7}$$

Using (4.6), (4.7), and $k \leq m - 1$, we see that

$$\sigma_1(\xi) = \sigma_2(\xi) \quad \forall\, \xi \in \mathbb{R}, \tag{4.8}$$

i.e.,

$$s_1(x) = s_2(x) \qquad \forall\, x \in G_p \cap (K_1 \cup K_2). \tag{4.9}$$

Since $p \in F$ is arbitrary, we let p vary along F to find

$$s_1(x) = s_2(x) \qquad \forall\, x \in K_1 \cup K_2, \tag{4.10}$$

i.e.,

$$s|_{K_i}(x) = s^*(x) \qquad \forall\, x \in K_1 \cup K_2, \quad i = 1,2. \tag{4.11}$$

Repeating this argument, we see that for any $K \in \mathcal{T}_n$,

$$s|_K = s^*|_K, \tag{4.12}$$

and so $s = s^* \in P_k$. Thus s arbitrary in $\mathcal{S}_n$ yields

$$\mathcal{S}_n \subseteq P_K, \tag{4.13}$$

so that

$$n = \dim \mathcal{S}_n \leq \dim P_k = \binom{k+N}{N}, \tag{4.14}$$

which is impossible, since k and N are fixed, while n is an arbitrary positive integer. Hence $k \geq m$. ∎

We are now able to establish the sharpness of the usual estimate for this error of the FEM φ_n^*, generalizing the work of Strang and Fix [8].

Theorem 4.1. Let $r \geq -m$, and define

$$\mu = \min(k + 1 - m, m + r). \tag{4.15}$$

Then

(i) $$e(\varphi_n^*) = \Omega(n^{-\mu/N}) \text{ as } n \to \infty,$$

and

(ii) $e(\varphi_n^*) = \Theta(n^{-\mu/N})$ as $n \to \infty$ for quasi-uniform $\{\mathcal{S}_n\}_{n=1}^{\infty}$.

Proof: First note that (3.15), (3.22), (3.23), and (3.26) yield

$$e(\varphi_n^*) \geq r(n) = \Theta(n^{-(r+m)/N}). \tag{4.16}$$

It remains to show

$$e(\varphi_n^*) = \Omega(n^{-(k+1-m)/N}) \text{ as } n \to \infty, \tag{4.17}$$

since (4.15), (4.16), and (4.17) imply (i), while (i) and the usual estimate (3.8) yield (ii).

In order to show (4.17), we will rely heavily on the notation found in [2]. First, let Ω° be the interior of a hypercube such that $\overline{\Omega^\circ} \subset \Omega$,

$$\mathcal{T}_n^\circ := \{K \in \mathcal{T}_n : K \subset \overline{\Omega^\circ}\}, \tag{4.18}$$

and

$$\Omega_n := \text{int} \cup \{K : K \in \mathcal{T}_n^\circ\}. \tag{4.19}$$

For any element $K \in \mathcal{T}_n$, let

$$\rho_K := \sup\{\text{diam}(S) : S \text{ a ball}, S \subset K\} \tag{4.20}$$

and

$$h_K := \text{diam } K. \tag{4.21}$$

Then $\{\mathcal{T}_n\}_{n=1}^\infty$ regular means that

$$\lim_{n\to\infty} \sup_{K\in\mathcal{T}_n} h_K = 0 \tag{4.22}$$

and there is a constant $\sigma > 0$ such that

$$h_K \le \sigma \rho_K \qquad \forall\, K \in \mathcal{T}_n, \qquad \forall\, n \ge 1. \tag{4.23}$$

Using (4.18), (4.19), (4.21), and (4.22), we find that

$$\bar{\Omega}_n \subset \overline{\Omega^o} \qquad \forall\ n \geq 1, \tag{4.24}$$

but that

$$\lim_{n\to\infty} \mathrm{vol}(\overline{\Omega_n}) = \mathrm{vol}(\overline{\Omega^o}). \tag{4.25}$$

We next choose u to be any function in $H^m_E(\Omega) \cap H^{2m+r}(\Omega)$ such that

$$u(x) = \frac{1}{(k+1)!}\, x_1^{k+1} \qquad \forall\ x \in \overline{\Omega^o}. \tag{4.26}$$

Let $K \in \mathfrak{I}_n^o$. We now claim that there is a constant $C_1 > 0$, independent of K and n, such that

$$\inf_{s\in P_k(K)} |u - s|^2_{m,K} \geq C_1^2\, \mathrm{vol}(K)^{2(k+1-m)/N + 1}, \tag{4.27}$$

$P_k(K)$ denoting polynomials of degree k over K. To show (4.27), there is an affine bijection $F_K : \hat{K} \to K$ with

$$F_K\hat{x} = B_K\hat{x} + b_K, \tag{4.28}$$

where $\hat{K}$ is a reference element independent of K, so that K is the F_K-image of a "reference element" $\hat{K}$ independent of n and K. Then Theorem 3.1.2 of [2] yields the existence of a constant $c_1 = c_1(k,m) > 0$ such that

$$\inf_{s\in P_k(K)} |u - s|^2_{m,K} \geq c_1 |\det B_K|\, \|B_K\|^{-2m} \inf_{\hat{s}\in P_k(\hat{K})} |\hat{u} - \hat{s}|^2_{m,\hat{K}}, \tag{4.29}$$

where $\|\cdot\|$ is the Euclidean matrix norm and where, for any function $v : K \to \mathbb{R}$, we define $\hat{v} : \hat{K} \to \mathbb{R}$ by

$$\hat{v}(\hat{x}) := v(x) \quad \text{for} \quad x = F_K\hat{x}. \tag{4.30}$$

On the other hand, one may check that the functionals

$$\hat{v} \mapsto |\hat{v}|_{k+1,\hat{K}} \tag{4.31}$$

and

$$\hat{v} \mapsto \inf_{\hat{s}\in P_k(\hat{K})} |\hat{v} - \hat{s}|_{m,\hat{K}} \tag{4.32}$$

are seminorms on $P_{k+1}(\hat{K})$. Since $k \geq m$, we find that these two seminorms have the same kernel, namely the space $P_k(\hat{K})$. Since $P_{k+1}(\hat{K})$ is finite dimensional, there is a constant $c_2 = c_2(k,m,\hat{\Omega}) > 0$ such that

$$\inf_{\hat{s}\in P_k(\hat{K})} |\hat{v} - \hat{s}|_{m,K} \geq c_2 |\hat{v}|_{k+1,\hat{K}} \qquad \forall\ \hat{v} \in P_{k+1}(\hat{K}). \tag{4.33}$$

Hence, we may use (4.27) and (4.33) with $\hat{v}$ replaced by $\hat{u}$ to see that

$$\begin{aligned} \inf_{s\in P_k(K)} |u - s|^2_{m,K} &\geq c_1 c_2^2 |\det B_K| \|B_K\|^{-2m} |\hat{u}|^2_{k+1,\hat{K}} \\ &\geq c_3 \frac{(\|B_K^{-1}\|^{-1})^{2(k+1)}}{\|B_K\|^{2m}} |u|^2_{k+1,K} \end{aligned} \tag{4.34}$$

for $c_3 = c_3(k,m) > 0$ by Theorem 3.1.2 of [2]. Using (4.23) and Theorem 3.1.3 of [2], we find that there exist $c_4, c_5 > 0$ such that

$$\begin{aligned} \|B_K^{-1}\|^{-1} &\geq c_4 \rho_K \geq c_4 \sigma^{-1} h_K \\ \|B_K\| &\leq c_5 h_K \end{aligned} \tag{4.35}$$

Since

$$\text{vol}(K) \leq \sigma_N h_K^N, \tag{4.36}$$

where σ_N is the volume of the unit ball in $\mathbb{R}^N$, we have

$$\inf_{s \in P_k(K)} |u - s|^2_{m,K} \geq C_1^2 \, \text{vol}(K)^{2(k+1-m)/N} |u|^2_{k+1,K}, \tag{4.37}$$

where $C_1 = C_1(k,m) > 0$. Finally, not that (4.26) yields

$$|u|^2_{k+1,K} = \sum_{|\alpha|=k+1} \int_K |D^\alpha u|^2 = \int_K 1 = \text{vol}(K), \tag{4.38}$$

and so (4.27) now follows from (4.37) and (4.38).

Hence, $\mathcal{T}_n^0 \subset \mathcal{T}_n$ yields

$$\begin{aligned} \inf_{s \in S_n} |u - s|^2_m &\geq \sum_{K \in \mathcal{T}_n^0} \inf_{s \in P_k(K)} |u - s|^2_{m,K} \\ &\geq C_1^2 \sum_{K \in \mathcal{T}_n^0} \text{vol}(K)^{2(k+1-m)/N + 1}. \end{aligned} \tag{4.39}$$

Since

$$\sum_{K \in \mathcal{T}_n^0} \text{vol}(K) = \text{vol}(\overline{\Omega_n}), \tag{4.40}$$

we may use calculus to find that

$$\sum_{K \in \mathcal{T}_n^0} \text{vol}(K)^{2(k+1-m)/N + 1} \geq \left[\frac{\text{vol}(\overline{\Omega_n})}{\# \mathcal{T}_n^0} \right]^{2(k+1-m)/N}. \tag{4.41}$$

From (4.24) and (4.25), there is an $n_0 > 0$ such that

$$\text{(4.42)} \qquad \mathrm{vol}(\overline{\Omega_n}) \geq \frac{1}{2}\,\mathrm{vol}(\overline{\Omega^o}) \qquad \forall\ n \geq n_0.$$

Hence, (4.39), (4.41), and (4.42) yield that there is a $C_2 > 0$, independent of n, for which

$$\text{(4.43)} \qquad \inf_{s \in \mathcal{S}_n} |u - s|_m \geq C_2 (\#\mathcal{T}_n^o)^{-(k+1-m)/N} \qquad \forall\ n \geq n_0.$$

We now claim that there is a $C_3 > 0$, independent of n, such that

$$\text{(4.44)} \qquad \#\mathcal{T}_n^o \leq C_3 n.$$

We first consider the case $m = 0$. In this case, the functions

$$\text{(4.45)} \qquad \{\chi_K : K \in \mathcal{T}_n^o\}$$

are linearly independent elements of $\mathcal{S}_n$, χ_K denoting the characteristic function of K. Since $\dim \mathcal{S}_n = n$, we have

$$\text{(4.46)} \qquad \#\mathcal{T}_n^o = \#\{\chi_K : K \in \mathcal{T}_n^o\} \leq n$$

for the case $m = 0$. We now assume that $m \geq 1$. Let $\mathcal{S}_n(\overline{\Omega_n})$ denote the restrictions of functions in $\mathcal{S}_n$ to $\overline{\Omega_n}$, so that

$$\text{(4.47)} \qquad \dim \mathcal{S}_n(\overline{\Omega_n}) \leq \dim \mathcal{S}_n = n.$$

In the case $N = 1$, we may count free parameters to see that

$$\text{(4.48)} \qquad k + 1 + (\#\mathcal{T}_n^o - 1)(k + 1 - m) = \dim \mathcal{S}_n(\overline{\Omega_n}),$$

so that

$$\text{(4.49)} \qquad \#\mathcal{T}_n^o \leq \frac{n - m}{k + 1 - m},$$

which implies (4.44) for the case $N = 1$. To establish the case for $N \geq 2$, we first claim that there is a $c_6 > 0$, independent of n, for which

$$\#\mathcal{T}_n^0 \leq c_6 v(\mathcal{T}_n^0), \tag{4.50}$$

where $v(\mathcal{T}_n^0)$ is the number of vertices in the triangulation $\mathcal{T}_n^0$. Indeed, regularity of $\{\mathcal{T}_n\}_{n\geq 1}$ (and hence $\{\mathcal{T}_n^0\}_{n\geq 1}$) yields that there is a $c_6 > 0$, independent of n, such that if v is a vertex of $\mathcal{T}_n^0$, then v can belong to at most c_6 simplices in $\mathcal{T}_n^0$, which implies (4.50). We need only show that there is a $c_7 > 0$, independent of n, for which

$$v(\mathcal{T}_n^0) \leq c_7 \dim \mathcal{S}_n(\overline{\Omega_n}) \tag{4.51}$$

(in the case $m \geq 1$, $N \geq 2$); (4.44) then follows from (4.50), (4.51), and (4.47). Now $\mathcal{S}_n \subset C^{m-1}(\Omega)$ (see proof of Lemma 4.1). In the case $N = 2$, Theorem 1 of Ženíšek [11] states that at each vertex v of $\mathcal{T}_n^0$,

$$\varphi \mapsto D^\alpha \varphi(v) \text{ for } |\alpha| \leq 2(m-1) \tag{4.52}$$

must be degrees of freedom, while the case for $N > 2$ may be reduced to the case $N = 2$ by considering restrictions of functions in $\mathcal{S}_n(\overline{\Omega_n})$ to 2-faces of simplices $K \subset \mathcal{T}_n^0$. Hence, (4.51) holds with $c_7 = \binom{N + 2(m-1)}{2(m-1)}^{-1}$, which finally completes the proof of (4.44).

As a result of (4.43) and (4.44), and $\|\cdot\|_m \geq |\cdot|_m$, we see that there is a $C_4 > 0$, independent of n, and an $n_0 > 0$,

such that

$$\inf_{s\in S_n} \|u - s\|_m \geq C_4 n^{-(k+1-m)/N} \qquad \forall\ n \geq n_0. \tag{4.53}$$

Now let $f = Lu$. Since $0 \neq u \in H^{2m+r}(\Omega) \cap H_E^m(\Omega)$, we see that $0 \neq f \in H^r(\Omega)$. Let

$$f^* := f/\|f\|_r \tag{4.54}$$

and

$$u^* := Sf^* = u/\|f\|_r. \tag{4.55}$$

Recall that there exists a finite $C_5 > 0$ such that $\|\cdot\|_m \leq C_5\|\cdot\|_B$ on $H_E^m(\Omega)$. Since $\|f^*\|_r \leq 1$, we have

$$\begin{aligned} C_5 e(\varphi_n^*) &\geq \sup_{\|f\|_r \leq 1} \inf_{s\in S_n} \|Sf - s\|_m \\ &\geq \inf_{s\in S_n} \|Sf^* - s\|_m = \inf_{s\in S_n} \|u^* - s\|_m \\ &= \frac{1}{\|f\|_r} \inf_{s\in S_n} \|u - s\|_m \qquad \text{(since } S_n \text{ is a subspace)} \\ &\geq \frac{C_4}{\|f\|_r} n^{-(k+1-m)/N} \qquad \forall\ n \geq n_o, \end{aligned} \tag{4.56}$$

which establishes (4.17) and the theorem. ■

We now ask whether the FEM is asymptotically optimal using the information n_n^*. We find that this is the case if and only if $k \geq 2m - 1 + r$ from

Theorem 4.2.

(i) $r(\hat{n}_n^*) = \Omega(n^{-(m+r)/N})$ as $n \to \infty$.

(ii) If $\{\mathfrak{I}_n\}_{n=1}^{\infty}$ is quasi-uniform, then

$$e(\varphi_n^s) = r(\hat{n}_n^*) = \Theta(n^{-(r+m)/N}) \text{ as } n \to \infty, \tag{4.57}$$

where φ_n^s is the spline algorithm using the information $\hat{n}_n^*$.

Proof: Using (3.15), (3.22), (3.23), and (3.26), we find

$$r(\hat{n}_n^*) \geq r(n) = \Theta(n^{-(r+m)/N}) \text{ as } n \to \infty, \tag{4.58}$$

establishing (i). To establish (ii), let $z \in \mathfrak{I}_0 \cap \ker \hat{n}_n^*$. Then

$$z \in \ker \hat{n}_n^* \Rightarrow (z,s)_0 = 0 \qquad \forall\ s \in \mathcal{S}_n \tag{4.59}$$

and

$$z \in \mathfrak{I}_0 \Rightarrow z \in H^r(\Omega) \text{ and } \|z\|_r \leq 1. \tag{4.60}$$

From (2.8) and (2.9), we see that (4.59) yields

$$\begin{aligned} \|Sz\|_B^2 &= B(Sz,Sz) = (z,Sz)_0 \\ &= (z,Sz - s)_0 \qquad \forall\ s \in \mathcal{S}_n \\ &\leq \|z\|_r \|Sz - s\|_{-r} \qquad \forall\ s \in \mathcal{S}_n. \end{aligned} \tag{4.61}$$

By Theorem 4.1.1 of [1] and the equivalence of $\|\cdot\|_B$ and $\|\cdot\|_m$, there exists $s_n \in \mathcal{S}_n$, such that

$$\|Sz - s_n\|_{-r} \leq C_1 n^{-\lambda/N} \|Sz\|_m \leq C_2 n^{-\lambda/N} \|Sz\|_B, \tag{4.62}$$

where C_1 and C_2 are positive constants independent of n, and

(4.63) $$\lambda = \min(k + 1 + r, m + r) = m + r$$

by Lemma 4.1. Hence (4.61)-(4.63) yield that

(4.64) $$\|Sz\|_B \le C_2 n^{-(m+r)/N} \qquad \forall\ z \in \mathfrak{F}_0 \cap \ker \eta_n,$$

and so (3.16) yields

(4.65) $$r(\eta_n) = \sup_{z \in \mathfrak{F}_0 \cap \ker \eta_n} \|Sz\|_B \le C_2 n^{-(m+r)/N}.$$

Using (4.58), (4.65), and (3.20), we find (ii). ∎

Hence, the information η_n^* is (asymptotically) an nth optimal information. In the case that $k \ge 2m - 1 + r$, the FEM is (asymptotically) an optimal error algorithm using η_n^*; when this inequality no longer holds, the FEM is no longer an asymptotically optimal error algorithm.

Remark 4.1. In Section 2, some rather stringent assumptions were made concerning the smoothness of the region and the coefficients appearing in the differential operators $L, B_0, \ldots, B_{m-1}$. If these smoothness assumptions are violated, the shift theorem no longer holds; that is, although the second inequality in (2.10) holds for all $r \ge -m$, the first inequality may only hold for all r in some subinterval $[-m, r_0)$. Since the shift theorem no longer holds for all r, the error of the FEM is now $\Omega(n^{-(m+r_0)/N})$ as $n \to \infty$, no matter how big r is, and no matter how k is chosen. On the other hand, the proof of the error estimate of the spline algorithm does not use the shift theorem.

Hence, Theorem 4.2 holds, even if the smoothness conditions imposed in Section 2 are drastically weakened.

As an example, we consider a problem with mixed Dirichlet-Neumann boundary conditions. Let

$$\partial\Omega = \Gamma_D \cup \Gamma_N \tag{4.66}$$

be a partition of $\partial\Omega$ such that Γ_D is of positive boundary measure. Let

$$H_E^1(\Omega) := \{v \in H^1(\Omega) \ : \ v = 0 \text{ on } \Gamma_D\}. \tag{4.67}$$

For $r \geq -1$, we consider the problem of finding, for any $f \in H^r(\Omega)$, a function $u = Sf \in H_E^1(\Omega)$ satisfying

$$\int_\Omega \nabla u \cdot \nabla v = \int_\Omega fv \qquad \forall\ v \in H_E^1(\Omega). \tag{4.68}$$

This is the weak form of the problem

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \Gamma_D \\ \frac{\partial u}{\partial \nu} &= 0 \quad \text{on } \Gamma_N \end{aligned} \tag{4.69}$$

($\frac{\partial}{\partial \nu}$ denoting the normal derivative).

We wish to put (4.68) into the notation of Section 2. Let χ_D and χ_N denote the characteristic functions of Γ_D and Γ_N, respectively. Define

$$\begin{aligned} Lv &:= -\Delta v \\ B_0 v &:= \chi_D v + \chi_N \frac{\partial v}{\partial \nu} \end{aligned} \tag{4.70}$$

.

Then $u = Sf$ is the weak solution to

$$\begin{aligned} Lu &= f \qquad \text{in } \Omega \\ B_0 u &= 0 \qquad \text{on } \partial\Omega \end{aligned} \tag{4.71}$$

However, the coefficients appearing in B_0 are discontinuous, i.e., the smoothness assumptions of Section 2 are violated.

From results of Grisvard [4], the first inequality in (2.10) holds only for $r \in [-1, \frac{1}{2})$, i.e., $r_0 = \frac{1}{2}$ above. Hence, the FEM has error $\Omega(n^{-3/(2N)})$ as $n \to \infty$, while the spline algorithm has error $\Theta(n^{-(m+r)/N})$ as $n \to \infty$. ■

5. Local Error and Generalized Galerkin Methods

In this section, we discuss generalized Galerkin methods (i.e., with different spaces of test and trial functions) for the variational boundary-value problem. We wish to determine when such a method has local error which is bounded by a constant multiple of the optimal local error. In particular, we will exhibit a FEM which is (worst-case) asymptotically optimal, yet whose local error is arbitrarily worse than the optimal local error.

To measure the amount by which the local error of an algorithm varies from the optimal local error, Traub, Wasilkowski, and Wozniakowski (Chapter 4 of [9]) introduced the concept of "deviation." Let φ be an algorithm using n. Then the deviation $\mathrm{dev}(\varphi)$ of φ is defined to be

$$\mathrm{dev}(\varphi) := \sup_{f \in \mathfrak{F}_0} \frac{e(\varphi, f)}{\mathrm{rad}\ Uf} \quad . \tag{5.1}$$

Clearly

$$\mathrm{dev}(\varphi) \geq 1, \tag{5.2}$$

with

$$\mathrm{dev}(\varphi) = 1 \quad \text{iff} \quad \varphi \quad \text{is strongly optimal error.} \tag{5.3}$$

Moreover, Traub et al. showed that if φ is homogeneous, i.e.,

$$\varphi(\alpha y) = \alpha \varphi(y) \qquad \forall\ \alpha \in \mathbb{R},\ y \in \mathbb{R}^n \tag{5.4}$$

then (in our Hilbert space setting)

(5.5) $\qquad \mathrm{dev}(\varphi) < \infty \quad \text{iff} \quad \varphi = \varphi^S.$

So, the spline method is the only homogeneous algorithm having finite deviation.

We wish to determine when the FEM has finite deviation. Since the FEM is linear (and hence homogeneous), the FEM has finite deviation if and only if the FEM is a spline algorithm. Hence, we wish to investigate when the FEM is a spline algorithm. In order to expedite this investigation, we now define generalized Galerkin methods (which include standard Galerkin methods and hence FEM's).

Let $\{s_i\}_{i=1}^n$ and $\{t_i\}_{i=1}^n$ each be linearly independent sets of functions in $H_E^m(\Omega)$. Let

(5.6) $\qquad \mathcal{S} := \mathrm{span}\{s_i\}_{i=1}^n \quad \text{and} \quad \mathcal{T} := \mathrm{span}\{t_i\}_{i=1}^n$

denote the subspaces of test and trial functions (respectively). We define the generalized Galerkin method $\varphi_{\mathcal{S},\mathcal{T}}$ using $\mathcal{S}$ and $\mathcal{T}$ by

(5.7) $\qquad \varphi_{\mathcal{S},\mathcal{T}}(n_{\mathcal{S}} f) := u_{\mathcal{S},\mathcal{T}},$

where $u_{\mathcal{S},\mathcal{T}} \in \mathcal{T}$ satisfies

(5.8) $\qquad B(u_{\mathcal{S},\mathcal{T}}, s) = (f,s)_0 \qquad \forall\, s \in \mathcal{S}$

and $n_{\mathcal{S}}$ is the Galerkin information (3.2) generated by $\mathcal{S}$.

Remark 5.1. The (standard) Galerkin method $\varphi_{\mathcal{S}}$ is a generalized Galerkin method with $\mathcal{T} = \mathcal{S}$. The FEM is a generalized Galerkin method with $\mathcal{T} = \mathcal{S} = \mathcal{S}_n$, with $\mathcal{S}_n$ an n-dimensional subspace of $H_E^m(\Omega)$ consisting of piecewise polynomials of degree k. ■

Remark 5.2. It is more common to use Galerkin methods with different spaces of test and trial functions in the solution of variational problems associated with a bilinear form on $\mathcal{U} \times \mathcal{V}$, where $\mathcal{U}$ and $\mathcal{V}$ are different Hilbert spaces. See e.g. pp. 217-220 in Section 6.3 of [1] for such a method. ■

In what follows, we let $S^* : H_E^m(\Omega) \to H^r(\Omega)$ denote the Hilbert space adjoint to S, remembering that $H_E^m(\Omega)$ is a Hilbert space under the inner product given by the bilinear form B. Hence (2.9) yields

$$(5.9) \quad (g,v)_0 = B(Sg,v) = (g,S^*v)_r \quad \forall\, v \in H_E^m(\Omega), \quad g \in H^r(\Omega).$$

We then have

Lemma 5.1. $(\ker \mathcal{N})^\perp = S^*\mathcal{S}$.

Proof: Let $s \in \mathcal{S}$. Then for any $h \in \ker \mathcal{N}$,

$$(5.10) \qquad (S^*s,h)_r = B(Sh,s) = (h,s)_0 = 0.$$

Hence $S^*\mathcal{S} \subset (\ker \mathcal{N})^\perp$. Now S is a dense injection, so that S^* is an injection. So $\#\mathcal{N} = n$ yields

$$(5.11) \qquad \dim S^*\mathcal{S} = \dim \mathcal{S} = n = \dim(\ker \mathcal{N})^\perp,$$

which, along with $S^*\mathcal{S} \subset (\ker \mathcal{N})^\perp$, yields the desired result. ■

Lemma 5.2. Given n-dimensional subspaces $\mathcal{S}$ and $\mathcal{T}$ of $H_E^m(\Omega)$, suppose that bases $\{s_i\}_{i=1}^n$ and $\{t_i\}_{i=1}^n$ of $\mathcal{S}$ and $\mathcal{T}$, respectively, are chosen such that

$$(5.12) \qquad (S^*s_j,S^*s_i)_r = \delta_{ij} \qquad (1 \le i, j \le n)$$

and

$$B(t_j, s_i) = \delta_{ij} \qquad (1 \le i, j \le n). \tag{5.13}$$

Then

$$\varphi^S(\eta_g f) = \sum_{i=1}^{n} (f, s_i)_0 SS^* s_i \tag{5.14}$$

and

$$\varphi_{g,\mathcal{T}}(\eta_g f) = \sum_{i=1}^{n} (f, s_i)_0 t_i. \tag{5.15}$$

Proof: Let $f_i = S^* s_i$. Then Lemma 3.1 yields $f_i \in (\ker \eta_g)^{\perp}$. In addition, (5.9) (with $g = S^* s_i$ and $v = s_j$) and (5.12) yield

$$(f_i, s_j)_0 = (S^* s_i, s_j)_0 = (S^* s_i, S^* s_j)_r = \delta_{ij}. \tag{5.16}$$

The representation formula (5.14) for the spline algorithm φ^S now follows from (5.1) of Chapter 4 of [9].

To see (5.15), write $u_{g,\mathcal{T}}$ in the form

$$u_{g,\mathcal{T}} = \sum_{j=1}^{n} \alpha_j t_j. \tag{5.17}$$

Then (5.8) and (5.13) yield

$$(f, s_i)_0 = B(u_{g,\mathcal{T}}, s_i) = \sum_{j=1}^{n} \alpha_j B(t_j, s_i) = \alpha_i, \tag{5.18}$$

establishing (3.15). ■

We now give the main result of this section, which tells us the unique choice of trial function space $\mathcal{T}$ (corresponding to

the given test function space $\mathcal{S}$) for which the generalized Galerkin method is the spline method, i.e., for which the generalized Galerkin method has finite deviation.

Theorem 5.1. Let $\mathcal{S}$ and $\mathcal{T}$ be n-dimensional subspaces of $H_E^m(\Omega)$. Then the following are equivalent:

(i) $\operatorname{dev}(\varphi_{\mathcal{S},\mathcal{T}}) < \infty$.

(ii) $\varphi_{\mathcal{S},\mathcal{T}} = \varphi^S$.

(iii) $\mathcal{T} = SS^*\mathcal{S}$.

Proof: Since $\varphi_{\mathcal{S},\mathcal{T}}$ is linear and thus homogeneous, (i) and (ii) are equivalent by (5.5). We show that (ii) and (iii) are equivalent. Let $\mathcal{S}$ and $\mathcal{T}$ be n-dimensional subspaces of $H_E^m(\Omega)$; choose a basis $\{s_i\}_{i=1}^n$ for $\mathcal{S}$ such that (5.12) holds.

Suppose first that (ii) holds. Choosing a basis $\{t_i\}_{i=1}^n$ for $\mathcal{T}$ such that (5.13) holds, Lemma 5.1 yields (5.14) and (5.15). Using (5.9), we have

$$
\begin{aligned}
t_i &= \sum_{j=1}^{n} (S^*s_i, s_j)_0 t_j = \varphi_{\mathcal{S},\mathcal{T}}(n_{\mathcal{S}} S^* s_i) \\
&= \varphi^S(n_{\mathcal{S}} S^* s_i) = \sum_{j=1}^{n} (S^*s_i, s_j)_0 SS^*s_j = SS^*s_i
\end{aligned}
\tag{5.19}
$$

for $1 \le i \le n$, so that (5.6) yields $\mathcal{T} = SS^*\mathcal{S}$. So, (ii) implies (iii).

Now suppose that (iii) holds. Let

$$
t_i = SS^*s_i \qquad (1 \le i \le n). \tag{5.20}
$$

Then (iii) and the injectivity of SS^* show that $\{t_i\}_{i=1}^n$ is a basis for $\mathcal{T}$. Using (5.9) and (5.12), we have (for

$1 \leq i, j \leq n$)

$$(5.21) \qquad B(t_j, s_i) = B(SS^* s_j, s_i) = (S^* s_j, S^* s_i)_r = \delta_{ij},$$

so that (5.13) and (5.14) hold. Using (5.13), (5.14), and (5.15), we see that (ii) holds. Thus (iii) implies (ii). □

Hence, given any finite-dimensional subspace $\mathcal{S}$ of $H_E^m(\Omega)$ we see how to choose the unique subspace $\mathcal{T}$ of $H_E^m(\Omega)$ with $\dim \mathcal{T} = \dim \mathcal{S}$ such that $\varphi^S = \varphi_{\mathcal{S},\mathcal{T}}$. On the other hand, the most natural choice of subspace is to pick $\mathcal{T} = \mathcal{S}$, so that we get the standard Galerkin method $\varphi_{\mathcal{S}}$. When is $\mathrm{dev}(\varphi_{\mathcal{S}})$ finite, i.e., when is $\varphi_{\mathcal{S}}$ the spline method?

Theorem 5.2. Let $\mathcal{S}$ be an n-dimensional subspace of $H_E^m(\Omega)$. Then the following conditions are equivalent:

(i) $\mathrm{dev}(\varphi_{\mathcal{S}}) < \infty$.

(ii) $\varphi_{\mathcal{S}} = \varphi^S$.

(iii) $\mathcal{S} = SS^*\mathcal{S}$.

(iv) $\mathcal{S}$ is an eigenspace of SS^*.

(v) $\mathcal{S} = S\mathcal{T}$, where $\mathcal{T}$ is an n-dimensional subspace of $H^r(\Omega)$ such that $\mathcal{T} = S^*S\mathcal{T}$.

(vi) $\mathcal{S} = S\mathcal{T}$, where $\mathcal{T}$ is an n-dimensional eigenspace of S^*S.

Proof: From Theorem 5.1, we have (i), (ii), and (iii) are equivalent. Suppose that (iii) holds. Then $SS^* : \mathcal{S} \to \mathcal{S}$ is self-adjoint, so that $\mathcal{S}$ (being finite-dimensional) has a basis of eigenvectors of SS^*, i.e., $\mathcal{S}$ is an eigenspace of SS^*, i.e., (iii) implies (iv). On the other hand, an eigenspace

of an operator is always invariant under that operator, i.e., (iv) implies that $SS^*\mathcal{S} \subset \mathcal{S}$; the inclusion $\mathcal{S} \subset SS^*\mathcal{S}$ follows from the injectivity of SS^* and the finite-dimensionality of $\mathcal{S}$. So (iii) and (iv) are equivalent.

Suppose that (iv) holds. Let $\mathcal{T} = S^*\mathcal{S}$. Using the equivalence of (iii) and (iv), we have $S^*S\mathcal{T} = S^*SS^*\mathcal{S} = S^*\mathcal{S} = \mathcal{T}$, while injectivity of S^* yields $\dim \mathcal{T} = \dim \mathcal{S} = n$. Moreover, $S\mathcal{T} = SS^*\mathcal{S} = \mathcal{S}$. So (iv) implies (v). If (v) holds, then $\mathcal{S} = S\mathcal{T} = SS^*S\mathcal{T} = SS^*\mathcal{S}$, so (v) implies (iii), which in turn yields (iv). So (iv) and (v) are equivalent.

Finally, $\mathcal{T} = S^*S\mathcal{T}$ if and only if $\mathcal{T}$ is an eigenspace of S^*S, the argument being similar to that in the preceding paragraph. Hence (v) and (vi) are equivalent. ∎

We now consider two examples for which one of the conditions in Theorem 5.2 holds, so that the Galerkin method and the spline method are one and the same.

Example 5.1. Let $r = -m$. Then S is the Riesz map, which is an isometric isomorphism of $H^{-m}(\Omega)$ (under the norm $\|S\cdot\|_B$, which is equivalent to $\|\cdot\|_{-m}$) onto $H_E^m(\Omega)$ (under the norm $\|\cdot\|_B$), see Section 4.4 of [7]. Hence $SS^* = I$, the identity map on $H_E^m(\Omega)$, and so $\mathcal{S} = SS^*\mathcal{S}$ for any subspace $\mathcal{S}$ of $H_E^m(\Omega)$. So when $r = -m$, the standard Galerkin method is the spline algorithm, no matter what the choice of $\mathcal{S}$. Of course in this case, (3.26) shows that $\lim_{n\to\infty} r(n) \neq 0$, i.e., there is no convergent sequence of algorithms, each of which uses finite information; see also Corollary 5.1 in Chapter 2 of [9]. ∎

Example 5.2. Let $r > -m$. Then $H^r(\Omega)$ has a complete basis of eigenvectors for S^*S, i.e., there exists a basis $\{e_i\}_{i=1}^{\infty}$ for $H^r(\Omega)$ and real numbers $\lambda_1 \geq \lambda_2 \geq \ldots > 0$ with $\lim_{i\to\infty} \lambda_i = 0$, such that for any positive integers i and j,

$$S^*Se_i = \lambda_i e_i \tag{5.22}$$

and

$$(e_j,e_i)_r = \delta_{ij}. \tag{5.23}$$

Define the information $n_n : H^r(\Omega) \to \mathbb{R}^n$ by

$$n_n f := [(f,e_1)_r \ \ldots \ (f,e_n)_r]^T. \tag{5.24}$$

Then $\#n_n = n$. Moreover, n_n is the nth optimal information, and

$$r(n_n) = r(n) = \sqrt{\lambda_{n+1}} = \Theta(n^{-(r+m)/N}) \quad \text{as} \quad n \to \infty \tag{5.25}$$

(Theorem 5.3 of Chapter 2 of [9]). Letting

$$\mathcal{J} = \text{span}\{e_1,\ldots,e_n\}, \tag{5.26}$$

we see that (vi) of Theorem 5.2 holds. Setting

$$\mathcal{S} = \text{span}\{s_1,\ldots,s_n\}, \quad s_i := \frac{1}{\lambda_i} Se_i \qquad (1 \leq i \leq n), \tag{5.27}$$

we find $n_n = n_{\mathcal{S}}$, since (4.27), (4.32), and (4.5) yield

$$(f,e_i)_r = \left(f, \frac{S^*Se_i}{\lambda_i}\right)_r = (f,S^*s_i)_r = (f,s_i)_0 \qquad (1 \leq i \leq n). \tag{5.28}$$

Hence the spline method and the standard Galerkin method coincide

for the nth optimal information Λ_n. Since

$$(5.29)\quad (S^*s_j, S^*s_i)_r = \frac{1}{\lambda_i\lambda_j}(S^*Se_j, S^*Se_i)_r = (e_j,e_i)_r = \delta_{ij},$$

we see that the formula for the spline method and standard Galerkin method is given by

$$(5.30)\quad \varphi^S(\Lambda_n f) = \varphi_S(\Lambda_n f) = \sum_{i=1}^{n}(f,s_i)_0 SS^*s_i = \sum_{i=1}^{n}\lambda_i(f,e_i)_r Se_i$$

in this case. ■

We now turn to an example $S \neq SS^*S$. This example is of particular interest because it gives an instance of an FEM which has optimal worst-case error (to within a constant, independent of n), but has infinite deviation.

Example 5.3. We consider the L_2-approximation problem for H^1-functions on the interval (0,1). Choose $N = 1$, $m = 0$, $r - 1$, and let $S : H^1(0,1) \to L_2(0,1)$ be the canonical injection

$$(5.31)\qquad Su := u \qquad \forall\ u \in H^1(0,1).$$

The variational form of the problem is to define $B : L_2(0,1) \times L_2(0,1) \to \mathbb{R}$ by

$$(5.32)\qquad B(u,v) := \int_0^1 uv \qquad \forall\ u,v \in L_2(0,1).$$

Then for any $f \in H^1(0,1)$, we wish to find $u = Sf \in L_2(0,1)$ such that

$$(5.33)\qquad B(u,v) = \int_0^1 fv \qquad \forall\ v \in L_2(0,1).$$

(Of course, $u = f$.)

We let S_n be an n-dimensional subspace of $L_2(0,1)$ consisting of piecewise constants, so that $k = 0$. Let

$$0 = x_0 < x_1 < \ldots < x_{n-1} < x_n = 1 \tag{5.34}$$

be a partition of $(0,1)$. Then S_n is the span of the functions $s_1,\ldots,s_n$ where

$$s_i(x) = \delta_{ij} \qquad x_{j-1} \le x \le x_j \qquad (1 \le j \le n,\ 1 \le i \le n). \tag{5.35}$$

Using an integration by parts, one can show that for any $s \in L_2(\Omega)$, $w := SS^*s$ is the (weak) solution to

$$\begin{aligned} -w'' + w &= s \qquad \text{in } (0,1) \\ w'(0) &= w'(1) = 0 \end{aligned} \tag{5.36}$$

so that

$$w(x) = \frac{\int_0^1 s(\xi)\cosh(1-\xi)\,d\xi}{\sinh 1}\cosh x - \int_0^x s(\xi)\sinh(x-\xi)\,d\xi. \tag{5.37}$$

Hence SS^*S_n is the span of $\{w_1,\ldots,w_n\}$, where

$$\begin{aligned} w_i(x) &= \frac{\kappa_i}{\sinh 1}\cosh x - \int_0^1 s_i(\xi)\sinh(x-\xi)\,d\xi \\ &= \frac{\kappa_i}{\sinh 1}\cosh x - [\cosh(x-x_{i-1}) - \cosh(x-x_i)] \end{aligned} \tag{5.38}$$

and

$$\kappa_i = \int_0^1 s_i(\xi)\cosh(1-\xi)\,d\xi = \sinh(1-x_{i-1}) - \sinh(1-x_i). \tag{5.39}$$

Since none of the w_i is piecewise constant on (0,1), we have $w_i \notin S_n$, so that $SS^* S_n \neq S_n$.

Hence, the FEM is not the spline method in this case, and thus has infinite deviation. This is interesting for the following version. Suppose that $\{S_n\}_{n=1}^{\infty}$ is quasi-uniform. Then the FEM has worst-case error $\Theta(n^{-1})$, and is (to within a constant, independent of n) an optimal-error algorithm. Hence, we have an example of an almost optimal-error algorithm that has infinite deviation, i.e., is not strongly optimal error. ■

Examples 5.1 and 5.3 suggest the following

Conjecture 5.1. Let $r > -m$ and let S_n be a finite-element subspace of $H_E^m(\Omega)$. Then the FEM using S_n has infinite deviation. ■

From the results of Section 4, it is clear that Conjecture 5.1 holds when $k < 2m - 1 + r$. Hence, it remains only to prove the conjecture for the case $k \geq 2m - 1 + r$.

6. Complexity Analysis

In this section, we discuss the complexity of finding ϵ-approximations to the solution of the variational boundary-value problem, as well as the penalty for using the FEM when $k < 2m - 1 + r$.

Let $\epsilon > 0$. An algorithm φ using Λ furnishes an ϵ-approximation to the problem if

$$e(\varphi) \leq \epsilon. \tag{6.1}$$

The complexity $\mathrm{comp}(\varphi)$ of an algorithm φ using Λ is defined in the model of computation discussed in Chapter 5 of [9]. (Informally, we assume that linear functionals can be evaluated in finite time and that the cost of an arithmetic operation is unity.) It then turns out that for any algorithm φ using Λ of cardinality n,

$$\mathrm{comp}(\varphi) \geq nc_1 + n - 1, \tag{6.2}$$

c_1 being the complexity of evaluating a linear functional, while if φ is a linear function of the information used, then

$$\mathrm{comp}(\varphi) \leq nc_1 + 2n - 1. \tag{6.3}$$

(See Section 2, Chapter 5 of [9] for further details.) We then define, for $\epsilon > 0$, the ϵ-complexity $\mathrm{COMP}(\epsilon)$ of the problem to be

$$\mathrm{COMP}(\epsilon) := \inf\{\mathrm{comp}(\varphi) : e(\varphi) \leq \epsilon\}, \tag{6.4}$$

the infimum being taken over all such φ using information of finite cardinality.

Remark 6.1. Note that we distinguish between algorithmic complexity and problem complexity. For an algorithm φ, $\mathrm{comp}(\varphi)$ denotes the complexity of the algorithm φ, while for $\epsilon > 0$, $\mathrm{COMP}(\epsilon)$ denotes the (minimal) complexity of finding an ϵ-approximation. To tie these two concepts togehter, let $\epsilon > 0$. Suppose that φ^* is an algorithm with

$$e(\varphi^*) \leq \epsilon \tag{6.5}$$

and such that for any other algorithm φ,

$$e(\varphi) \leq \epsilon \Rightarrow \mathrm{comp}(\varphi) \geq \mathrm{comp}(\varphi^*). \tag{6.6}$$

Then φ^* is an optimal complexity algorithm for finding an ϵ-approximation, and

$$\mathrm{COMP}(\epsilon) = \mathrm{comp}(\varphi^*). \tag{6.7}$$ ■

Let $\{S_n\}_{n=1}^{\infty}$ be a quasi-uniform family of finite element subspaces of $H_E^m(\Omega)$ consisting of piecewise polynomials of degree k. Let φ_n^* be the FEM based on the space S_n; that is, for $f \in \mathcal{F}_0$, we let $u_n^* \in S_n$ satisfy

$$B(u_n^*, s) = (f,s)_0 \qquad \forall\ s \in S_n, \tag{6.8}$$

and then set

$$\varphi_n^*(h_n^* f) := u_n^*. \tag{6.9}$$

We wish to measure the algorithmic complexity of using the FEM to find ϵ-approximations, i.e.,

(6.10) $$\text{FEM}(\epsilon) := \inf\{\text{comp}(\varphi_n^*) : e(\varphi_n^*) \le \epsilon\},$$

and compare FEM(ϵ),

(6.11) $$\text{SPLINE}(\epsilon) := \inf\{\text{comp}(\varphi_n^s) : e(\varphi_n^s) \le \epsilon\}$$

(φ_n^s being the spline algorithm using the Galerkin information $\mathcal{N}_n^*$ generated by S_n), and COMP(ϵ).

Using (6.2), (6.4), and the results in Section 4, we find

(6.12) $$\text{FEM}(\epsilon) = \Theta(\epsilon^{-N/\mu}) \text{ as } \epsilon \to 0,$$

where

(6.13) $$\mu = \min(k + 1 - m, m + r),$$

while

(6.14) $$\text{SPLINE}(\epsilon) = \Theta(\epsilon^{-N/(m+r)}) \text{ as } \epsilon \to 0$$

and

(6.15) $$\text{COMP}(\epsilon) = \Theta(\epsilon^{-N/(m+r)}) \text{ as } \epsilon \to 0.$$

This yields

Theorem 6.1.

(i) The spline algorithm is asymptotically optimal, i.e.,

$$\text{SPLINE}(\epsilon) = \Theta(\text{COMP}(\epsilon)) = \Theta(\epsilon^{-N/(m+r)}) \text{ as } \epsilon \to 0.$$

(ii) If $k \ge 2m - 1 + r$, the FEM is asymptotically optimal, i.e.,

$$\text{FEM}(\epsilon) = \Theta(\text{COMP}(\epsilon)) = \Theta(\epsilon^{-N/(m+r)}) \quad \text{as} \quad \epsilon \to 0.$$

(iii) If $k < 2m - 1 + r$, then

$$\frac{\text{FEM}(\epsilon)}{\text{COMP}(\epsilon)} = \Theta\left(\frac{\text{FEM}(\epsilon)}{\text{SPLINE}(\epsilon)}\right) = \Theta\left((\tfrac{1}{\epsilon})^{\lambda N}\right) \quad \text{as} \quad \epsilon \to 0,$$

where

$$\lambda = \frac{1}{k + 1 - m} - \frac{1}{m + r} > 0,$$

so that

$$\lim_{\epsilon \to 0} \frac{\text{FEM}(\epsilon)}{\text{COMP}(\epsilon)} = \lim_{\epsilon \to 0} \frac{\text{FEM}(\epsilon)}{\text{SPLINE}(\epsilon)} = +\infty. \tag{6.16}$$

■

Thus when k is too small for a given value of r, the asymptotic penalty for using the FEM instead of the spline algorithm is infinite. Clearly (6.16) tells us that there exists $\epsilon_0 > 0$ for which

$$\text{SPLINE}(\epsilon) < \text{FEM}(\epsilon) \quad \text{for} \quad 0 < \epsilon < \epsilon_0. \tag{6.17}$$

What is the value of ϵ_0? If ϵ_0 is unreasonably small, it may turn out that it is more reasonable to use the FEM for "practical" values of ϵ. We determine the value of ϵ_0 for a model problem in

Example 6.1. Let $N = 1$, $\Omega = (0,\pi)$, $m = 1$, $r = 1$, $H_E^1(\Omega) = H_0^1(0,\pi)$, and consider the bilinear form $B : H_0^1(0,\pi) \times H_0^1(0,\pi) \to \mathbb{R}$ defined by

$$B(v,w) := \int_0^\pi v'w' \qquad \forall\ v,w \in H_0^1(0,\pi). \tag{6.18}$$

Hence for $f \in H^1(0,\pi)$, $u = Sf$ is the variational solution to the problem

$$-u''(x) = f(x) \qquad 0 < x < \pi$$
$$u(0) = u(\pi) = 0 \tag{6.19}$$

We choose $\mathcal{S}_n$ to be a subspace of $H_n^1(0,\pi)$ consisting of piecewise linear polynomials with nodes at $x_j = \frac{j\pi}{n+1}$ $(0 \le j \le n+1)$. Hence $k = 1$; moreover, since any function in $\mathcal{S}_n$ must vanish at the endpoints of $[0,\pi]$, we see that $\dim \mathcal{S}_n = n$.

We first give a lower bound on $e(\varphi_n^*)$, φ_n^* being the nth FEM. Let

$$f(x) := \frac{1}{\sqrt{\pi}} \tag{6.20}$$

Then

$$\|f\|_1 = 1 \tag{6.21}$$

and $Sf = u$, where

$$u(x) := \frac{1}{2}\, x\left(\sqrt{\pi} - \frac{x}{\sqrt{\pi}}\right) \tag{6.22}$$

Let $\tilde{u}_n$ be the $\mathcal{S}_n$-interpolate of u, i.e., $\tilde{u}_n$ is the unique function in $\mathcal{S}_n$ for which

$$\tilde{u}_n(x_j) = u(x_j) \qquad (1 \le j \le n). \tag{6.23}$$

Then using Newton's interpolation formula on each subinterval $[x_j, x_{j+1}]$ along with the fact that u'' is a constant, one can show that

$$\inf_{s \in S_n} |u - s|_1 = |u - \tilde{u}_n|_1 = \frac{\pi}{\sqrt{12}\,(n+1)} . \tag{6.24}$$

Using (3.7), (6.21), (6.24), and $u = Sf$, we conclude that

$$e(\varphi_n^*) \geq \frac{\pi}{\sqrt{12}\,(n+1)} . \tag{6.25}$$

Now we can find a lower bound on $\mathrm{FEM}(\epsilon)$. Let $e(\varphi_n^*) \leq \epsilon$. Then (6.25) yields

$$n \geq \frac{\pi}{\sqrt{12}}\,\epsilon^{-1} - 1, \tag{6.26}$$

which, along with (6.2), gives the lower bound

$$\mathrm{FEM}(\epsilon) \geq (c_1 + 1)\left(\frac{\pi}{\sqrt{12}}\,\epsilon^{-1} - 1\right) - 1. \tag{6.27}$$

Next, we wish to give an upper bound on $e(\varphi_n^s)$, where φ_n^s is the spline algorithm using N_n^*. Since $e(\varphi_n^s) = r(N_r^*)$, it suffices to compute the radius of information. Let $z \in \ker N_n^* \cap BH^1$. Let P_n denote the orthogonal projector of $L_2(0,\pi)$ onto S_n. Using (2.8), (2.9), the fact that $z \in \ker N_n^*$, and properties of the orthogonal projector, we find

$$\begin{aligned} \|Sz\|_B^2 = B(Sz,Sz) &= (z,Sz)_0 = (z, Sz - P_n Sz)_0 \\ &= (z - \tilde{z}_n, Sz - P_n Sz)_0 \\ &\leq \|z - \tilde{z}_n\|_0 \|Sz - P_n Sz\|_0 \\ &\leq \|z - \tilde{z}_n\|_0 \|Sz - (\widetilde{Sz})_n\|_0 \end{aligned} \tag{6.28}$$

(where for $v \in H_0^1(0,\pi)$, $\tilde{v}_n$ is the S_n-interpolate of v as given by (6.23)). Since for any $v \in H_0^1(0,\pi)$, Theorem 2.4 of [7] states that

$$\|v - \tilde{v}_n\|_0 \leq \frac{1}{n+1} |v|_1, \tag{6.29}$$

(6.28) becomes

$$\|Sz\|_B^2 \leq \left(\frac{1}{n+1}\right)^2 |z|_1 |Sz|_1 \leq \left(\frac{1}{n+1}\right)^2 \|Sz\|_B, \tag{6.30}$$

where we have used $z \in BH^1$ (so that $|z|_1 \leq 1$) and $\|\cdot\|_B = |\cdot|_1$. Hence

$$\|Sz\|_B \leq \left(\frac{1}{n+1}\right)^2 \qquad \forall\ z \in \ker \eta_n^* \cap BH^1, \tag{6.31}$$

so that

$$e(\varphi_n^s) = r(\eta_n^*) \leq \left(\frac{1}{n+1}\right)^2 \tag{6.32}$$

by (3.16).

Using (6.32), we now find an upper bound on SPLINE(ϵ). Let

$$n \leq \epsilon^{-1/2} - 1. \tag{6.33}$$

Then (6.32) yields that $e(\varphi_n^s) \leq \epsilon$. From (6.3), we find that

$$\text{SPLINE}(\epsilon) \leq (c_1 + 2)(\epsilon^{-1/2} - 1) - 1. \tag{6.34}$$

We now wish to find $\epsilon_0 = \epsilon_0(c_1)$ such that (6.17) holds. From (6.17), (6.27), and (6.34), we see that we may choose ϵ_0 to be the smallest positive solution of

$$(c_1 + 1)\left(\frac{\pi}{\sqrt{12}}\,\epsilon_0^{-1} - 1\right) = (c_1 + 2)(\epsilon_0^{-1/2} - 1). \tag{6.35}$$

Some algebra yields

$$(6.36)\quad \epsilon_0(c_1) = \left[\frac{1}{2} c_1 + 1 - \sqrt{\left(\frac{1}{2} c_1 + 1\right)^2 - \frac{\pi}{\sqrt{12}} (c_1 + 1)} \right]^2 .$$

Some elementary calculus tells us that ϵ_0 is an increasing function. Since $c_1 \geq 0$, we thus have

$$(6.37)\quad \epsilon_0(c_1) \leq \epsilon_0(0) = \left(1 - \sqrt{1 - \frac{\pi}{\sqrt{12}}} \right)^2 \doteq 0.482853424.$$

Thus (6.17) holds for all ϵ roughly less than one-half.

On the other hand, if we are willing to assume that evaluating a linear functional is at least as hard as an arithmetic operation, we have $c_1 \geq 1$ and so

$$(6.38)\quad \epsilon_0(c_1) \leq \epsilon_0(1) = \left(\frac{3}{2} - \sqrt{\frac{9}{4} - \frac{\pi}{\sqrt{3}}} \right)^2 \doteq 0.7048360247.$$

Of course, it is reasonable to suspect that $c_1 >> 1$ (see e.g. pg. 85 of [9]). One may check that

$$(6.39)\quad \lim_{c_1 \to \infty} \epsilon_0(c_1) = \frac{\pi^2}{12} \doteq 0.8224670334,$$

giving an estimate of $\epsilon_0(c_1)$ for large values of c_1. ■

Based on this example, it seems reasonable to conjecture that for any regularly-elliptic boundary-value problem, (6.17) will hold, where ϵ_0 is sufficiently large to be of interest. We suspect that such a result will be difficult to establish. There are two reasons for this. The first reason is that "sufficiently large" may be a subjective criterion. That is,

$\epsilon_0 = 10^{-1000}$ is obviously too small to be of practical interest, while $\epsilon_0 = 10^{-1}$ is not so absurdly small; where does one draw the line separating the reasonable values of ϵ_0 from the unreasonable values?

The second reason is perhaps more crucial. In order to determine ϵ_0, we have to change the order-of-magnitude estimates in Theorems 4.1 and 4.2 to sharp bounds involving constants whose values are explicitly known. Since the values of the constants appearing in these theorems are not explicitly known in general, we suspect that this task will be very difficult for a general problem, making it very difficult to determine, for a general problem, a value of ϵ_0 such that (6.17) will hold.

7. Some Practical Considerations

In this section, we discuss the implementation of the spline algorithm. In particular, we ask when it is of practical use.

The main problem is that the spline algorithm is hard to implement. This is mainly due to the presence of the adjoint S^* to the solution operator S. As we saw in Example 5.3, even for very simple solution operators S and subspaces $\mathcal{S}_n$ of test functions, the space $SS^*\mathcal{S}_n$ of trial functions can be very complicated.

To a certain extent, this problem may be solved by preconditioning. Generally speaking, the problem S, the class $\mathcal{F}_0$ of problem elements, and the family $\{\mathcal{S}_n\}_{n=1}^{\infty}$ of subspaces of test functions will be fixed. Suppose, for a given $\epsilon > 0$, we wish to compute ϵ-approximations to Sf for many $f \in \mathcal{F}_0$. Then we may determine a fixed cardinality n, depending on ϵ, such that $e(\varphi_n^s) \leq \epsilon$, i.e., the spline method φ_n^s yields ϵ-approximations for any $f \in \mathcal{F}_0$. Now, we may precondition: instead of finding $t_i = SS^*s_i$, we find approximations $\tilde{t}_i$ to t_i for $1 \leq i \leq n$. (This may perhaps be done via an FEM.) Moreover, we may use standard techniques (e.g., the Q - R method) to biorthonormalize $\{s_i\}_{i=1}^n$ and $\{\tilde{t}_i\}_{i=1}^n$. Although this may be a lot of work, it is independent of the choice of f. Hence, if we wish to compute ϵ-approximations to Sf for many different $f \in \mathcal{F}_0$, this may be a feasible technique. (But note that since this is a linear method which does not exactly coincide with the spline algorithm, its deviation is infinite, no matter how close $\tilde{t}_i$ and t_i are!)

On the other hand, suppose we wish to compute a sequence of approximations to Sf for a fixed $f \in \mathfrak{F}_0$. In this case, the preconditioning will be prohibitively expensive, because as ϵ changes, the value of n such that $e(\varphi_n^*) \leq \epsilon$ changes, which implies that the algorithm φ_n^s changes. Since SS^*s_i $(1 \leq i \leq n)$ cannot be explicitly computed for general S, it appears that the spline method will not be practical in this case. Using Theorem 4.1, it appears that the best advice is to use an FEM of sufficiently high degree, unless the problems involved in implementing such a method are so great (or ϵ is so large) that one doesn't mind the penalty of $\Theta(\epsilon^{-\lambda N})$ as $\epsilon \to 0$, where $\lambda = (k + 1 - m)^{-1} - (m + r)^{-1}$, which will result from using an FEM of degree $k < 2m - 1 + r$.

In summary, we see that for the case of solving problems $Lu = f$ with many different f to within a fixed error criterion ϵ, the spline algorithm may be of practical interest. On the other hand, we do not currently know how to efficiently implement the spline algorithm when solving a single problem with greater and greater accuracy.

Acknowledgments

I would like to thank Professors A. K. Aziz and T. I. Seidman (University of Maryland Baltimore County) and J. F. Traub (Columbia University) for their comments and suggestions.

References

[1] I. Babuska and A. K. Aziz, "Survey lectures on the mathematical foundations of the finite element method," in The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, (A. K. Aziz, ed.), Academic Press, New York, 1972, pp. 3-359.

[2] P. G. Ciarlet, The Finite Element Method for Elliptic Problems, North-Holland, Amsterdam, 1978.

[3] P. G. Ciarlet and P. A. Raviart, "Interpolation theory over curved elements," Comput. Methods Appl. Mech. Engrg., v. 1, 1972, pp. 217-249.

[4] P. Grisvard, "Behavior of the solutions of an elliptic boundary-value problem in a polygonal or polyhedral domain," in Numerical Solutions of Partial Differential Equations -- III (SYNSPADE 1975), (Bert Hubbard, ed.), Academic Press, New York, 1976, pp. 207-274.

[5] D. E. Knuth, "Big Omicron and Big Omega and Big Theta," SIGACT News, ACM, April, 1976.

[6] J. T. Oden and J. N. Reddy, An Introduction to the Mathematical Theory of Finite Elements, Wiley-Interscience, New York, 1976.

[7] M. Schultz, Spline Analysis, Prentice-Hall, Englewood Cliffs, 1973.

[8] G. Strang and G. Fix, "A Fourier analysis of the finite element variational method," in Constructive Aspects of Functional Analysis, Part II, C.I.M.E., Rome, 1973.

[9] J. F. Traub and H. Woźniakowski, A General Theory of Optimal Algorithms, Academic Press, New York, 1980.

[10] A. G. Werschulz, "Optimal error properties of finite element methods for second order elliptic Dirichlet problems," Math. Comp., v. 38, No. 158, April, 1982, pp. 401-413.

[11] A. Ženíšek, "Hermite interpolation on simplexes in the finite element method," in Proceedings EquaDiff 3, J. E. Purkyně University, Brno, pp. 271-277.