MATHEMATICS OF COMPUTATION
VOLUME 60, NUMBER 201
JANUARY 1993, PAGES 49–78

# ON THE USE OF LAGRANGE MULTIPLIERS IN DOMAIN DECOMPOSITION FOR SOLVING ELLIPTIC PROBLEMS

HOWARD SWANN

ABSTRACT. The primal hybrid method for solving second-order elliptic equations is extended from finite element approximations to general bases. Variational techniques are used to show convergence of approximations to the solution of the homogeneous Dirichlet problem for selfadjoint equations. Error estimates are obtained and examples are given.

## INTRODUCTION

Lagrange multipliers have been employed to define classes of functions used to approximate solutions of elliptic partial differential equations in a number of ways. Greenstadt has described the *cell discretization* method, where the domain of a problem is partitioned into *cells*; approximations are made on each cell, and the approximations are forced to be weakly continuous across the boundaries of each cell by using Lagrange multipliers in a method called *moment collocation* [5, 6, 13–16]. These results are discussed in §4. Babuška has shown how Lagrange multipliers can be used to make finite element approximations match the boundary data in elliptic problems [1] (see also [4]). Dorr [9] has applied the methods of Babuška to force continuity across an internal interface formed by dividing a domain in $\mathbb{R}^2$ into two parts, using a finite element basis. The *primal hybrid* finite element method of Raviart and Thomas [18] shows how Lagrange multipliers can be used to ensure that nonconforming finite element approximations converge to solutions as the size of the mesh of the finite element grid becomes small. We show here that convergence of the Greenstadt method occurs in quite general situations. The cells do not diminish in size. The only requirement for convergence is that the basis functions on each cell constitute a Schauder basis in an appropriate space and that the weight functions defined on the boundary segments of each cell that are used to enforce moment collocation also be a Schauder basis. The algorithm is naturally suited for parallel computational methods.

Received by the editor November 26, 1990 and, in revised form, October 3, 1991.
1991 *Mathematics Subject Classification.* Primary 65N30; Secondary 65N35, 65N15.
*Key words and phrases.* Hybrid methods, nonconforming methods, Lagrange multipliers, domain decomposition, cell discretization.
This work was partially supported by grants from the IBM Palo Alto Scientific Center and the IBM Almaden Research Center.

In §1 we describe the setting for the problem and state some preliminary results. The main convergence result for strongly elliptic second-order selfadjoint problems is stated and proved in §2, and error estimates are given. We describe parallel methods for obtaining an approximation to the solution in §3. The preliminary results stated in §1 are proved in this section. Section 4 describes an implementation of the algorithm using polynomial bases for approximations in domains in $\mathbb{R}^2$. This example strongly resembles the $p$-version of the finite element method (see [2, 3, 8]). Two examples are given and comparisons are made with ELLPACK's Hermite collocation finite element method [19]. Results obtained by others using the algorithm are discussed.

## 1. Description of the problem and preliminary results

The task is to approximate the solution of an elliptic selfadjoint problem of the form

$$\mathbf{E}u = f \tag{1.1a}$$

over a suitable domain $\Omega$ in $\mathbb{R}^K$ (described below) with boundary $\Gamma$. If $D_i$ denotes partial differentiation with respect to $x_i$, the operator $\mathbf{E}$ is expressed as

$$\mathbf{E}u = -\sum_{i,j}^{K} D_i(A_{ij}(x)D_j u) + A_0(x)u\,.$$

We consider the homogeneous Dirichlet boundary condition expressed as

$$u|_\Gamma = 0\,. \tag{1.1b}$$

The weak formulation of the problem is the following:

Let $H_0^1(\Omega)$ be the subspace of functions in $H^1(\Omega)$ equal to zero on $\Gamma$. Define

$$a(u, v) = \int_\Omega \sum_{i,j}^{K} A_{ij}(x)D_i u D_j v + A_0(x)uv\,dx\,.$$

Find $u \in H_0^1(\Omega)$ such that

$$a(u, v) = (f, v) \tag{1.2}$$

for all $v \in H_0^1(\Omega)$, where $(\cdot, \cdot)$ denotes the $L_2$ inner product over $\Omega$.

We consider domains that have the following properties:

**Definition 1.1.** A domain $\Omega \subset \mathbb{R}^K$ has a boundary $\Gamma$ that is Lipschitz and piecewise $C^1$, denoted $\mathrm{LP}C^1$, if it satisfies the following:

(i) $\Omega$ is open, bounded, and connected;

(ii) $\Omega$ is the interior of its closure;

(iii) $\Gamma$ is Lipschitz, i.e., for any $x \in \Gamma$, there exists a neighborhood $V$ of $x$ in $\mathbb{R}^K$ and new orthogonal coordinates $(y_1, y_2, \ldots, y_K)$ such that

(a) $V = \{(y_1, \ldots, y_K)\colon -a_i < y_i < a_i,\ 1 \le i \le K\}$;

(b) There exists a uniformly Lipschitz-continuous function $g$ defined in

$$V^1 \equiv \{(y_1, \ldots, y_{K-1})\colon -a_i < y_i < a_i,\ 1 \le i \le K-1\}$$

such that $|g(y^1)| \le \frac{1}{2}a_K$ for any $y^1 \in V^1$, and

$$\Omega \cap V = \{y = (y^1, y_K) \in V\colon y_K < g(y^1)\}\,,$$

$$\Gamma \cap V = \{y = (y^1, y_K) \in V\colon y_K = g(y^1)\}\,.$$

(iv) $\Gamma = \mathscr{A}_1 \cup \cdots \cup \mathscr{A}_p \cup \mathscr{B}$, where

(a) $\mathscr{A}_i$ is a relatively open subset of $\Gamma$ and a simply-connected compact subset of a $C^1$ $(K-1)$-manifold, and

(b) $\mathscr{B}$ is a compact set contained in a finite union of $(K-2)$-manifolds and $\mathscr{A}_i \cap \mathscr{A}_j \subset \mathscr{B}$ if $i \neq j$.

This definition is that of Grisvard [17] for Lipschitz boundaries and follows Fleming [11] for the definition of piecewise $C^1$.

Let $\Omega$ be an LP$C^1$ domain. The Hilbert spaces we use are the following: Let $(\cdot, \cdot)$ denote the $L_2(\Omega)$ inner product, with norm denoted $\|\cdot\|_0$. $H^1(\Omega) \equiv \{u: \Omega \to \mathbb{R}: u \in L_2(\Omega);\ D_i u \in L_2(\Omega) \text{ for } i = 1, \ldots, K\}$, where partial derivatives $D_i u$ are distribution derivatives with respect to $x_i$. The space $H^1(\Omega)$ has inner product

$$(u, v)_{1,\Omega} = \sum_{i=1}^{K} (D_i u, D_i v) + (u, v).$$

The norm on $H^1(\Omega)$ is denoted $\|\cdot\|_{1,\Omega}$. $H_0^1(\Omega)$ is the closure of $C_0^\infty(\Omega)$ in the norm $\|\cdot\|_{1,\Omega}$.

Following Greenstadt's cell discretization method, we allow the domain $\Omega$ to be partitioned in any way into $N$ LP$C^1$ domains $\Omega_1, \ldots, \Omega_N$, with $\Omega_i \cap \Omega_j = \varnothing$ if $i \neq j$ and $\overline{\Omega} = \bigcup_{i=1}^{N} \overline{\Omega}_i$. The $\Omega_i$ are called *cells*.

Let $\Omega_0 \equiv \mathbb{R}^K \backslash \overline{\Omega}$. Let $(\cdot, \cdot)_{1,i}$ denote the $H^1(\Omega_i)$ inner product on the cell $\Omega_i$; the norm is denoted by $\|\cdot\|_{1,i}$. The $L_2(\Omega_i)$ inner product is denoted $(\cdot, \cdot)_i$; the norm is denoted by $\|\cdot\|_{0,i}$.

The parent space for much of our discussion is

$$H \equiv \{u \in L_2(\Omega): u|_{\Omega_i} \in H^1(\Omega_i);\ i = 1, \ldots, N\}.$$

The Hilbert space $H$ has inner product

$$(u, v)_H \equiv \sum_{i=1}^{N} (u, v)_{1,i}.$$

The $H$-norm is denoted by $\|\cdot\|_H$.

Let $\Gamma_{ij} = \overline{\Omega}_i \cap \overline{\Omega}_j$. Assume that $\Gamma_{ij}$ is the finite union of $\overline{\mathscr{A}_q}$, where the $\{\mathscr{A}_q\}$ have the properties of Definition 1.1(iv). Where no $\overline{\mathscr{A}_q}$ exist, $\Gamma_{ij}$ is ignored. To simplify notation, we refer to such $\overline{\mathscr{A}_q}$ as $\Gamma_{ij}$, acknowledging that there may be multiplicity involved. $\Gamma_{i0}$ is a boundary segment between $\Omega_i$ and $\Omega_0$. The inner product for $L_2(\Gamma_{ij})$ is denoted by $\langle \cdot, \cdot \rangle_{ij}$, with norm represented as $\|\cdot\|_{ij}$.

We denote by $\gamma_{ij}$ the trace operator restricting $u|_{\Omega_i}$ to its values on $\Gamma_{ij}$. From [17, p. 41] we can take $\gamma_{ij}$ to be a bounded linear operator from $H^1(\Omega_i)$ to $L_2(\Gamma_{ij})$; there are constants $C_{ij}$ such that for any $w \in H$, $\|\gamma_{ij}(w)\|_{ij} \leq C_{ij}\|w\|_{1,i}$. Since we are concerned with estimates in terms of $\|\gamma_{ij}(w)\|_{ij}$ rather than the $H^{1/2}(\Gamma_{ij})$ norm of $\gamma_{ij}(w)$ required by full use of the trace theorem [17, p. 37], constants $C_{ij}$ can be explicitly obtained for many $\Omega_i$ [20].

For each $\Gamma_{ij}$, choose $\{\omega_q^{ij}\}_{q=1}^{\infty}$ to be functions in $H^{1/2}(\Gamma_{ij})$ that are a Schauder basis for $L_2(\Gamma_{ij})$. For any $g \in L_2(\Gamma_{ij})$, there are some coefficients

$d_k$ such that $g = \sum_{k=1}^{\infty} d_k \omega_k^{ij}$. For any $n$, let $\mathscr{T}_n^{ij}(g) \equiv \sum_{k=n+1}^{\infty} d_k \omega_k^{ij}$. For any $\varepsilon > 0$, there is some $N(g, \varepsilon)$ such that $n > N(g, \varepsilon) \Rightarrow \|\mathscr{T}_n^{ij}(g)\|_{ij} < \varepsilon$.

Approximations are in $H$; weak continuity across interfaces $\Gamma_{ij}$ is enforced by Greenstadt's method called *moment collocation.*

For $u \in H$, we define the $q$th *moment* of $u$ on $\Gamma_{ij}$ to be

$$M_q^{ij}(u) \equiv \langle \gamma_{ij}(u), \omega_q^{ij} \rangle_{ij}.$$

We require that the moments of an approximation $u$ be equal on interfaces $\Gamma_{ij}$ in the following way.

Let $N_I$ be the number of interfaces $\Gamma_{ij}$. $[n]$ denotes a multi-index, an $N_I$-vector of nonnegative integers $(\ldots, n_{ij}, \ldots)$. A partial order is $[n'] \geq [n]$ if and only if for any $ij$, $n'_{ij} \geq n_{ij}$. We say that $[n^k] \to [\infty]$ if $[n^k] \leq [n^{k+1}]$ and $\inf\{n_{ij}^k\} \to \infty$ as $k \to \infty$.

Set $G[n] \equiv \{u \in H$: for any $ij$, $ij = 1, \ldots, N_I$, $j \neq 0$, and for any $q \leq n_{ij}$, we have $M_q^{ij}(u) = M_q^{ji}(u)\}$. In this case, $[n]$ is the multi-index described above, with all $n_{i0} = 0$, where the $n_{i0}$ refer to the $\Gamma_{i0}$. Thus, $G[n]$ is the set of functions $u$ in $H$ such that on any internal interface $\Gamma_{ij}$, $\gamma_{ij}(u) - \gamma_{ji}(u)$ is $L_2(\Gamma_{ij})$-orthogonal to $\omega_k^{ij}$, $k = 1, \ldots, n_{ij}$. This gives a notion of weak continuity across interfaces called *moment collocation.*

Define $G_0[n] = \{u \in G[n]$: for any $i$ and any $k \leq n_{i0}$, $M_k^{i0}(u) = 0\}$. Thus, $G_0[n]$ is the set of functions in $G[n]$ that are weakly 0 on the external interfaces $\Gamma_{i0}$ making up $\Gamma$; our approximations of solutions for problems with homogeneous boundary conditions are in this space. Owing to the continuity of the trace operator, $G_0[n]$ is a closed subspace of $H$. We have the inclusions $[n'] \geq [n] \Rightarrow G_0[n'] \subset G_0[n]$.

For each $i$th cell, choose any Schauder basis $\{B_k^i\}$ for $H^1(\Omega_i)$. For any $v$ in $H^1(\Omega_i)$, there are $b_k^i$ such that $\sum_{k=1}^{\infty} b_k^i B_k^i = v$; let $v_{\cdot,m} = \sum_{k=1}^{m} b_k^i B_k^i$. Let $\mathscr{Q}_m^i(v)$ denote the orthogonal projection (in the $H^1(\Omega_i)$ inner product) of $v$ onto the $H^1(\Omega_i)$-orthogonal complement of the span of $\{B_1^i, B_2^i, \ldots, B_m^i\}$. Thus, $\mathscr{Q}_m^i(v_{\cdot,m}) = 0$, $\mathscr{Q}_m^i(v) = \mathscr{Q}_m^i(v - v_{\cdot,m})$, and

$$\|\mathscr{Q}_m^i(v)\|_{1,i} \leq \|v - v_{\cdot,m}\|_{1,i} = \left\| \sum_{k=m+1}^{\infty} b_k^i B_k^i \right\|_{1,i}.$$

We have

$$\lim_{m \to \infty} \|\mathscr{Q}_m^i(v)\|_{1,i} = 0.$$

These properties of $\mathscr{Q}_m^i$ are independent of $[n]$.

Let $[m]$ be an $N$-dimensional multi-index indicating the number of basis functions used in the approximation on each cell; we employ the same notational conventions as those used for the multi-index $[n]$.

$H[m]$ is the subspace of $H$ such that for any $v \in H[m]$, $v|_{\Omega_i}$ is in the span of $\{B_1^i, B_2^i, \ldots, B_{m_i}^i\}$.

Given $[m]$ and any function $v$ in $H$, $\mathscr{Q}_{[m]}(v)$ is the function in $H$ such that $\mathscr{Q}_{[m]}(v)|_{\Omega_i} = \mathscr{Q}_{m_i}^i(v|_{\Omega_i})$. Thus, $\mathscr{Q}_{[m]}(\cdot)$ is the projection of $H$ onto $H[m]^{\perp}$. We have $\lim_{[m] \to [\infty]} \|\mathscr{Q}_{[m]}(v)\|_H = 0$.

Let $G_0[n][m] = \{u \in G_0[n]$: $u|_{\Omega_i} = \sum_{k=1}^{m_i} b_k^i B_k^i\}$. This is a finite-dimensional space; the moment collocation requirements are met by requiring that certain

linear equations hold among the $b_k^i$. It is shown in §3 that these equations are independent if $[m]$ is sufficiently large. Note that if $u \in G_0[n][m]$, then $\mathscr{Q}_{[m]}(u) = 0$. We have the inclusions

$$[m'] \geq [m] \Rightarrow G_0[n][m'] \supset G_0[n][m].$$

The result showing that members of $G_0[n]$ are approximated by elements in $G_0[n][m]$ is the following:

**Lemma 1.2.** *If $\mathscr{P}_m^n(\cdot)$ is the orthogonal projection operator of $G_0[n]$ onto $G_0[n][m]$, then there exists a constant $K_1$ depending on $[n]$, the choice of cells, the choice of basis functions, and the choice of moment collocation weight function such that, for any $v \in G_0[n]$,*

$$\|v - \mathscr{P}_m^n(v)\|_H \leq K_1 \|\mathscr{Q}_{[m]}(v)\|_H.$$

This lemma is proved in §3. The dependence of $K_1$ on $[n]$ is discussed there as well.

The following diagram shows the projections and the relations among these spaces:

$$\begin{array}{ccc} & \overset{\mathscr{Q}_{[m]}}{\dashrightarrow} & \\ H & = H[m] \oplus H[m]^{\perp} & \\ | & | & \\ G_0[n] & = G_0[n][m] \oplus G_0[n][m]^{\perp} & \\ | & \underset{\mathscr{P}_m^n}{\uparrow} & \\ H_0^1(\Omega) & & \end{array}$$

## 2. Statement of the theorem and proof of convergence

Our estimates require that a unique solution to (1.1a) exist that is in $H_0^2(\Omega)$. Sufficient conditions for this requirement are the following [17, p. 124]:

(2.1) We assume that $\Gamma$ is $C^1$, with Lipschitz derivatives.

(2.2) We assume that $A_{ij}(x) \in H^1(\Omega)$ with $D_k A_{ij}(x) \in L_\infty(\Omega)$, and that the $A_{ij}(x)$ are Lipschitz continuous on $\overline{\Omega}$ and $A_0(x) \in L_\infty(\Omega)$. We assume that there exists $c > 0$ such that $\sum_{i,j}^K A_{ij}(x) z_i z_j \geq c \sum_{i=1}^K z_i^2$ in $\Omega$ for any $z_i \in \mathbb{R}$, and that $A_0(x) \geq c$ a.e. for $x \in \Omega$ and $A_{ij}(x) = A_{ji}(x)$.

(2.3) We assume that $f \in L_2(\Omega)$.

Let $M' = \max\{\|A_{ij}\|_{L_\infty}, \|A_0\|_{L_\infty}\}$. For $u, v \in H^1(\Omega_k)$, we define

$$a(u, v)_k = \int_{\Omega_k} \left( \sum_{i,j}^K A_{ij}(x) D_i u D_j v + A_0(x) uv \right) dx.$$

We let $a(u, v)$ represent $\sum_{k=1}^N a(u, v)_k$. We have the inequality

$$\begin{aligned} |a(u, v)| &\leq M' \sum_{k=1}^N \left[ \sum_{i,j}^K \|D_i u\|_{0,k} \|D_j v\|_{0,k} + \|u\|_{0,k} \|v\|_{0,k} \right] \\ &\leq M' K \|u\|_H \|v\|_H. \end{aligned}$$

Let $M = M'K$. Note that $a(\cdot, \cdot)$ is coercive, for

$$a(v, v) \geq c \sum_{j=1}^N \left[ \sum_{i=1}^K \int_{\Omega_j} (D_i^2 v)\, dx + \int_{\Omega_j} v^2\, dx \right] = c\|v\|_H^2.$$

Let $D_{\mathbf{n}_{ij}}u$ be the "conormal derivative with respect to $\mathbf{E}$ of $u$ on $\Gamma_{ij}$". This is defined for sufficiently smooth $u$ as follows: If $\mathbf{n} = (n_1, n_2, \ldots, n_K)$ is the unit normal to $\Gamma_{ij}$ (pointing outward relative to the interior of $\Omega_i$), then

$$D_{\mathbf{n}_{ij}}u \equiv \sum_{p,q}^{K} \gamma_{ij}(A_{pq}D_q u)n_p .$$

Green's formula $(\mathbf{E}u, v) = a(u, v) - \langle D_{\mathbf{n}}u, \gamma(v)\rangle$ is valid for LP$C^1$ domains for $u$ in $H^2$ and $v$ in $H^1$ if the $A_{ij}$ are sufficiently smooth [17]; in particular, this holds with our assumptions concerning the $A_{ij}$ and for $\Omega$ and all $\Omega_j$.

Denote the solution in $H_0^2(\Omega)$ to (1.1a) and (1.1b) by $u$. Then $\mathbf{E}u = f$, $D_{\mathbf{n}_{ij}}u$ is in $L_2(\Gamma_{ij})$, and $u \in G_0[n]$.

A variational argument shows that a unique function $u_{n,m}$ exists in $G_0[n][m]$ that minimizes $a(u, u) - 2(f, u)$ over all $u \in G_0[n][m]$, and

$$a(u_{n,m}, v) = (f, v) \quad \text{for all } v \text{ in } G_0[n][m].$$

The function $u_{n,m}$ is obtained by solving a system of linear equations. The matrix describing the system is nonsingular if $[m]$ is sufficiently large; details are in §3.

We prove the following theorem.

**Theorem.** *Assume that* (2.1)–(2.3) *hold and that* $u$ *is the solution in* $H_0^2(\Omega)$ *to* (1.1a) *and* (1.1b). *Let* $(n_f)$ *be the largest number of faces* $\Gamma_{ij}$ *of any cell. Then*

$$c\|u - u_{n,m}\|_H \le (n_f)\sqrt{N}\sup\{C_{ij}\}\sup\{\|\mathscr{T}_{n_{ij}}^{ij}(D_{\mathbf{n}_{ij}}u)\|_{ij}\} + MK_1\|\mathscr{E}_{[m]}(u)\|_H .$$

Thus, $[n]$ is to be chosen so that the error estimated by the first term is acceptable. Lemma 1.2 shows that the error expressed by the second term is small if $[m]$ is made sufficiently large.

*Proof.* Since $u - u_{n,m} \in G_0[n]$ and $\mathscr{P}_m^n(u) - u_{n,m} = \mathscr{P}_m^n(u - u_{n,m}) \in G_0[n][m]$, we have

$$\begin{aligned}
c\|u - u_{n,m}\|_H^2 &\le a(u - u_{n,m}, u - u_{n,m}) \\
&= a(u - u_{n,m}, u - \mathscr{P}_m^n(u) + \mathscr{P}_m^n(u) - u_{n,m}) \\
&= a(u - u_{n,m}, u - \mathscr{P}_m^n(u)) + a(u - u_{n,m}, \mathscr{P}_m^n(u) - u_{n,m}) \\
&= a(u - u_{n,m}, u - \mathscr{P}_m^n(u)) + a(u, \mathscr{P}_m^n(u) - u_{n,m}) - a(u_{n,m}\mathscr{P}_m^n(u) - u_{n,m}) \\
&= a(u - u_{n,m}, u - \mathscr{P}_m^n(u)) + a(u, \mathscr{P}_m^n(u) - u_{n,m}) - (f, \mathscr{P}_m^n(u) - u_{n,m}) .
\end{aligned}$$

Let $\delta = \mathscr{P}_m^n(u) - u_{n,m}$; thus, $a(u, \mathscr{P}_m^n(u) - u_{n,m}) = a(u, \delta)$. Using Green's formula, we have

$$\begin{aligned}
a(u, \delta) = \sum_{i=1}^{N} a(u, \delta)_i &= \sum_{i=1}^{N} \left( (\mathbf{E}u, \delta)_i + \sum_j \langle D_{\mathbf{n}_{ij}}u, \gamma_{ij}(\delta)\rangle_{ij} \right) \\
&= \sum_{i=1}^{N} \left( (f, \delta)_i + \sum_j \langle D_{\mathbf{n}_{ij}}u, \gamma_{ij}(\delta)\rangle_{ij} \right) \\
&= (f, \delta) + \sum_{i=1}^{N} \left( \sum_j \langle D_{\mathbf{n}_{ij}}u, \gamma_{ij}(\delta)\rangle_{ij} \right) .
\end{aligned}$$

Using the fact that if the boundary segment is an internal interface, $D_{\mathbf{n}_{ij}}u = -D_{\mathbf{n}_{ji}}u$ and $\gamma_{ij}(u) = \gamma_{ji}(u)$, and grouping the sum of the boundary integrals above in pairs (if the boundary segment is an internal interface), we obtain

$$\sum_{i=1}^{N}\left(\sum_{j}\langle D_{\mathbf{n}_{ij}}u,\ \gamma_{ij}(\delta)\rangle_{ij}\right) = \sum_{\Gamma_{ij}}\langle D_{\mathbf{n}_{ij}}u,\ \gamma_{ij}(\delta)-\gamma_{ji}(\delta)\rangle_{ij} + \sum_{\Gamma_{i0}}\langle D_{\mathbf{n}_{i0}}u,\ \gamma_{i0}(\delta)\rangle_{i0}.$$

The first sum in the expression above is taken over $j \neq 0$ and we assume that $i < j$.

There exist $d_k$ such that $D_{\mathbf{n}_{ij}}u = \sum_{k=1}^{\infty} d_k\omega_k^{ij}$. Suppose that $n_{ij}$ moment collocations are enforced on $\Gamma_{ij}$. Then $\gamma_{ij}(\delta) - \gamma_{ji}(\delta)$ is orthogonal to the weight functions $\omega_k^{ij}$, $k = 1, \ldots, n_{ij}$, so

$$\begin{aligned}
|\langle D_{\mathbf{n}_{ij}}u,\ \gamma_{ij}(\delta)-\gamma_{ji}(\delta)\rangle_{ij}| &= \left|\left\langle\sum_{k=1}^{\infty} d_k\omega_k^{ij},\ \gamma_{ij}(\delta)-\gamma_{ji}(\delta)\right\rangle_{ij}\right| \\
&= \left|\sum_{k=1}^{\infty} d_k\langle\omega_k^{ij},\ \gamma_{ij}(\delta)-\gamma_{ji}(\delta)\rangle_{ij}\right| = \left|\sum_{k=n_{ij}+1}^{\infty} d_k\langle\omega_k^{ij},\ \gamma_{ij}(\delta)-\gamma_{ji}(\delta)\rangle_{ij}\right| \\
&= \left|\left\langle\sum_{k=n_{ij}+1}^{\infty} d_k\omega_k^{ij},\ \gamma_{ij}(\delta)-\gamma_{ji}(\delta)\right\rangle_{ij}\right| = |\langle\mathscr{T}_{n_{ij}}^{ij}(D_{\mathbf{n}_{ij}}u),\ \gamma_{ij}(\delta)-\gamma_{ji}(\delta)\rangle_{ij}| \\
&\leq \|\mathscr{T}_{n_{ij}}^{ij}(D_{\mathbf{n}_{ij}}u)\|_{ij}\|\gamma_{ij}(\delta)-\gamma_{ji}(\delta)\|_{ij} \\
&\leq \|\mathscr{T}_{n_{ij}}^{ij}(D_{\mathbf{n}_{ij}}u)\|_{ij}(\|\gamma_{ij}(\delta)\|_{ij} + \|\gamma_{ji}(\delta)\|_{ij}).
\end{aligned}$$

By the trace theorem, using the constants $C_{ij}$, this last expression is majorized by $\|\mathscr{T}_{n_{ij}}^{ij}(D_{\mathbf{n}_{ij}}u)\|_{ij}(C_{ji}\|\delta\|_{1,j} + C_{ij}\|\delta\|_{1,i})$.

Similarly, with $n_{i0}$ moment collocations enforced on $\Gamma_{i0}$, the trace $\gamma_{i0}(\delta)$ is orthogonal to the weight functions $\omega_k^{i0}$, $k = 1, \ldots, n_{i0}$, so

$$\begin{aligned}
\langle D_{\mathbf{n}_{i0}}u,\ \gamma_{i0}(\delta)\rangle_{i0} &\leq \|\mathscr{T}_{n_{i0}}^{i0}(D_{\mathbf{n}_{i0}}u)\|_{i0}\|\gamma_{i0}(\delta)\|_{i0} \\
&\leq \|\mathscr{T}_{n_{i0}}^{i0}(D_{\mathbf{n}_{i0}}u)\|_{i0}C_{i0}\|\delta\|_{1,i}.
\end{aligned}$$

Hence, using the estimates above, with $[n]$ moment collocations enforced, we get

$$\begin{aligned}
c\|u-u_{n,m}\|_H^2 &\leq a(u-u_{n,m},\ u-\mathscr{P}_m^n(u)) + a(u,\ \mathscr{P}_m^n(u)-u_{n,m}) - (f,\ \mathscr{P}_m^n(u)-u_{n,m}) \\
&= a(u-u_{n,m},\ u-\mathscr{P}_m^n(u)) + a(u,\ \delta) - (f,\ \delta) \\
&= a(u-u_{n,m},\ u-\mathscr{P}_m^n(u)) + (f,\ \delta) + \sum_{i=1}^{N}\left(\sum_{j}\langle D_{\mathbf{n}_{ij}}u,\ \gamma_{ij}(\delta)\rangle_{ij}\right) - (f,\ \delta) \\
&= a(u-u_{n,m},\ u-\mathscr{P}_m^n(u)) + \sum_{i=1}^{N}\left(\sum_{j}\langle D_{\mathbf{n}_{ij}}u,\ \gamma_{ij}(\delta)\rangle_{ij}\right).
\end{aligned}$$

Now

$$a(u-u_{n,m},\, u-\mathscr{P}_m^n(u)) \le M\|u-u_{n,m}\|_H\|u-\mathscr{P}_m^n(u)\|_H,$$

and

$$\begin{aligned}
&\sum_{i=1}^{N}\left(\sum_j \langle D_{\mathbf{n}_{ij}}u,\, \gamma_{ij}(\delta)\rangle_{ij}\right)\\
&\quad\le \sum_{\Gamma_{ij}} \|\mathscr{T}_{n_{ij}}^{ij}(D_{\mathbf{n}_{ij}}u)\|_{ij}(C_{ij}\|\delta\|_{1,j}+C_{ji}\|\delta\|_{1,i})\\
&\qquad+\sum_{\Gamma_{i0}} \|\mathscr{T}_{n_{i0}}^{i0}(D_{\mathbf{n}_{i0}}u)\|_{i0}C_{i0}\|\delta\|_{1,i}\\
&\quad\le \sup\{C_{ij}\}\sup\{\|\mathscr{T}_{n_{ij}}^{ij}(D_{\mathbf{n}_{ij}}u)\|_{ij}\}\left(\sum_{\Gamma_{ij}}(\|\delta\|_{1,j}+\|\delta\|_{1,i})+\sum_{\Gamma_{i0}}\|\delta\|_{1,i}\right)\\
&\quad\le \sup\{C_{ij}\}\sup\{\|\mathscr{T}_{n_{ij}}^{ij}(D_{\mathbf{n}_{ij}}u)\|_{ij}\}(n_f)\sum_{i=1}^{N}\|\delta\|_{1,i}\\
&\quad\le \sup\{C_{ij}\}\sup\{\|\mathscr{T}_{n_{ij}}^{ij}(D_{\mathbf{n}_{ij}}u)\|_{ij}\}(n_f)\sqrt{N}\|\delta\|_H.
\end{aligned}$$

We have used the fact that any $\|\delta\|_{1,i}$ will occur at most $(n_f)$ times in the sums over the $\Gamma_{ij}$. Note that

$$\|\delta\|_H = \|\mathscr{P}_m^n(u-u_{n,m})\|_H \le \|u-u_{n,m}\|_H.$$

Assembling these estimates, we get

$$\begin{aligned}
c\|u-u_{n,m}\|_H^2 &\le a(u-u_{n,m},\, u-\mathscr{P}_m^n u)+\sum_{i=1}^{N}\left(\sum_j\langle D_{\mathbf{n}_{ij}}u,\, \gamma_{ij}(\delta)\rangle_{ij}\right)\\
&\le M\|u-u_{n,m}\|_H\|u-\mathscr{P}_m^n(u)\|_H\\
&\quad+\sup\{C_{ij}\}\sup\{\|\mathscr{T}_{n_{ij}}^{ij}(D_{\mathbf{n}_{ij}}u)\|_{ij}\}(n_f)\sqrt{N}\|\delta\|_H\\
&\le M\|u-u_{n,m}\|_H\|u-\mathscr{P}_m^n(u)\|_H\\
&\quad+\sup\{C_{ij}\}\sup\{\|\mathscr{T}_{n_{ij}}^{ij}(D_{\mathbf{n}_{ij}}u)\|_{ij}\}(n_f)\sqrt{N}\|u-u_{n,m}\|_H.
\end{aligned}$$

Dividing both sides of the inequality by $\|u-u_{n,m}\|_H$, we get

$$c\|u-u_{n,m}\|_H \le M\|u-\mathscr{P}_m^n(u)\|_H+\sup\{C_{ij}\}\sup\{\|\mathscr{T}_{n_{ij}}^{ij}(D_{\mathbf{n}_{ij}}u)\|_{ij}\}(n_f)\sqrt{N}.$$

Since $\|u-\mathscr{P}_m^n(u)\|_H \le K_1\|\mathscr{Q}_{[m]}(u)\|_H$, we obtain our estimate. □

If we make $[n]$ sufficiently large, $\sup\{\|\mathscr{T}_{n_{ij}}^{ij}(D_{\mathbf{n}_{ij}}u)\|_{ij}\}$ is less than any $\varepsilon>0$ in view of the properties of $\mathscr{T}_n^{ij}(\cdot)$. To estimate $\|\mathscr{T}_{n_{ij}}^{ij}(D_{\mathbf{n}_{ij}}u)\|_{ij}$, we consider how well we can approximate $D_{\mathbf{n}_{ij}}u$ in the $L_2(\Gamma_{ij})$ norm by linear combinations of $\omega_k^{ij}$, $k=1,\dots,n_{ij}$. With an appropriate choice of cells, three or four moment collocations on each interface have usually been sufficient to obtain

reasonable accuracy, assuming that the basis for the approximations on the cells and the functions used to enforce moment collocation are suitably chosen (see §4).

Since, by Lemma 1.2, $\|u - \mathscr{P}_m^n(u)\|_H \to 0$ as $[m] \to [\infty]$, convergence of $u_{n,m}$ to $u$ is established.

The expression $\|u - \mathscr{P}_m^n(u)\|_H$ is majorized by $K_1 \|\mathscr{Q}_{[m]}(u)\|_H$. This partially decouples the problem of estimating errors in this term from concern with the moment collocation constraints. The definition of $\mathscr{Q}_{[m]}(\cdot)$ is independent of moment collocations $[n]$, so to estimate $\|\mathscr{Q}_{[m]}(u)\|_H$, we need only consider how well we can approximate $u$ in the $H^1$ norm by the chosen basis on any cell. Estimates of both errors in terms of $[n]$ and $[m]$ depend on the mode of convergence of the bases chosen for the cells and the interfaces $\Gamma_{ij}$. If the basis is a polynomial basis, we would expect that the error estimates would be in terms of the degree of polynomial approximation reflected in $[m]$ and the regularity of $u$. The methods used to study the $p$-version of finite element approximations are relevant here (see [2, 3, 8]). We give some examples showing the error in approximations for various values of $[m]$ and $[n]$ in §4, where polynomials provide both the cell bases and the moment collocation weight functions.

The parameter $K_1$, described in more detail in §3, is dependent on the moment collocation constraints (but not on $[m]$).

Dorr [9] has obtained some convergence results for finite element bases in domains in $\mathbb{R}^2$. Thus, if finite element solutions have been obtained on two domains, the domains can be "glued together" using this method. This may be of use in elasticity problems (see [20]).

## 3. Methods for obtaining the approximation and proofs of the preliminary results

In this section we describe the system of linear equations that generate the approximation in more detail, show that there exists a unique solution, and suggest a parallel algorithm for solving the system. We prove Lemma 1.2.

We wish to obtain the function $u$ in $G_0[n][m]$ that minimizes

$$a(u, u) - 2(f, u)$$

over all $u \in G_0[n][m]$.

On each cell $\Omega_k$, we use any Schauder basis $\{B_i^k(x)\}$ and form an approximation

$$u|_{\Omega_k} = \sum_{i=1}^{m_k} b_i^k B_i^k(x).$$

Then

$$\begin{aligned} a(u, u) - 2(f, u) &= \sum_{k=1}^{N} [a(u, u)_k - 2(f, u)_k] \\ &= \sum_{k=1}^{N} \left[ \sum_{i=1}^{m_k} b_i^k \sum_{j=1}^{m_k} b_j^k a(B_i^k, B_j^k)_k - 2 \sum_{i=1}^{m_k} b_i^k (f, B_i^k)_k \right]. \end{aligned}$$

This quadratic form is to be minimized subject to the moment collocation constraints. This is done by adding terms of form

$$-\lambda_q^{ij}(\langle \gamma_{ij}(u), \omega_q^{ij}\rangle_{ij} - \langle \gamma_{ji}(u), \omega_q^{ij}\rangle_{ij}), \qquad q = 1, \ldots, n_{ij},$$

and

$$-\lambda_q^{i0}(\langle \gamma_{i0}(u), \omega_q^{i0}\rangle_{i0}), \qquad q = 1, \ldots, n_{i0},$$

to the quadratic form for each interface $\Gamma_{ij}$, where $-\lambda_q^{ij}$ is a Lagrange multiplier. This converts the problem to that of finding the unconstrained minimum of a function $F(\mathbf{b}, \lambda)$, which produces a system of linear equations of form

$$\left(\begin{array}{c|c} \mathbf{C} & \mathbf{M}^{\mathrm{T}} \\ \hline \mathbf{M} & \mathbf{0} \end{array}\right)\begin{pmatrix} \mathbf{b} \\ -\lambda \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ \mathbf{0} \end{pmatrix}.$$

We have assumed that the elliptic equation is of Helmholtz type, with $A_0 > 0$. In this case, the matrix $\mathbf{C}$ consists of symmetric positive definite blocks along the diagonal and is zero elsewhere; each block corresponds to a cell and the number of basis functions used on the cell gives the block's size. The vector $\mathbf{b}$ contains the coefficients to be used with the basis functions to obtain the approximation, and $\mathbf{f}$ represents entries corresponding to the right-hand side of the elliptic equation $\mathbf{E}u = f$. The rectangular matrix $\mathbf{M}$, which we call the matrix of *moment collocation rows*, consists of a band of blocks, with zeros below the band; it is sparse above the band. We show below that the rows of $\mathbf{M}$ are independent if the total number of basis functions used in the approximation is sufficiently large. The vector $\lambda$ represents the Lagrange multipliers $\lambda_q^{ij}$ used to enforce the linear moment collocation requirements expressed here as $\mathbf{Mb} = \mathbf{0}$. As in [18], we expect $\lambda$ to represent an approximation to the normal derivative of the solution along the interfaces; $D_{\mathbf{n}_{ij}}u$ should be approximated by $\sum_{k=1}^{n_{ij}} \lambda_k^{ij}\omega_k^{ij}$. This is discussed in [20] and at the end of this section.

The computations required to generate $\mathbf{M}$, $\mathbf{f}$, and the diagonal blocks comprising $\mathbf{C}$ are independent, inviting the utilization of parallel processors. In our implementation of the algorithm described in §4, the entries for the blocks of $\mathbf{C}$ are computed in parallel on a Sequent Symmetry machine. This is particularly appropriate if a cell has a curved boundary segment, for quadrature over such cells can be time-consuming (see §4).

The following block-elimination algorithm provides a parallel direct method for solution of the linear system.

Since $\mathbf{C}$ is a matrix of positive definite diagonal blocks, parallel computations can obtain the Cholesky decomposition for each block, which allows us to represent $\mathbf{C}$ by $\mathbf{LL}^{\mathrm{T}}$, where $\mathbf{L}$ is lower triangular.

We can then proceed as follows:

We wish to solve $\mathbf{Cb} - \mathbf{M}^{\mathrm{T}}\lambda = \mathbf{f}$; $\mathbf{Mb} = \mathbf{0}$.

1. Find $\mathbf{Y}$ such that $\mathbf{CY} = \mathbf{M}^{\mathrm{T}}$ (so $\mathbf{Y} = \mathbf{C}^{-1}\mathbf{M}^{\mathrm{T}}$).
2. Solve $\mathbf{Cy} = \mathbf{f}$ (so $\mathbf{y} = \mathbf{C}^{-1}\mathbf{f}$).
3. Solve $[\mathbf{MY}]\lambda = -\mathbf{My}$ $(= -\mathbf{MC}^{-1}\mathbf{f})$.
4. Compute $\mathbf{b} = \mathbf{y} + \mathbf{Y}\lambda$ (so $\mathbf{b} = \mathbf{C}^{-1}\mathbf{f} + \mathbf{C}^{-1}\mathbf{M}^{\mathrm{T}}\lambda$).

Then the pair $(\mathbf{b}, \lambda)$ is our solution:
The definition of $\mathbf{b}$ in step 4 gives $\mathbf{Cb} - \mathbf{M}^{\mathrm{T}}\lambda = \mathbf{f}$, and

$$\mathbf{Mb} = \mathbf{MC}^{-1}\mathbf{f} + \mathbf{MC}^{-1}\mathbf{M}^{\mathrm{T}}\lambda = \mathbf{MC}^{-1}\mathbf{f} + \mathbf{MY}\lambda = \mathbf{MC}^{-1}\mathbf{f} - \mathbf{MC}^{-1}\mathbf{f} = 0.$$

Because of the structure of $\mathbf{C}$, steps 1 and 2 can be done in parallel without actually assembling the matrix $\mathbf{C}$. The matrix $\mathbf{MY} = \mathbf{MC}^{-1}\mathbf{M}^{\mathrm{T}}$ is positive definite, of relatively small size equal to the number of rows of $\mathbf{M}$. This establishes the existence of a unique solution. Since the moment collocation technique deals with the problem of matching approximations across interfaces, we can concentrate on finding basis functions that make the blocks of $\mathbf{C}$ well conditioned (see §4).

This method for solving the system of equations fails when we approximate solutions to Poisson's equation $-\Delta u = f$, where $A_0 = 0$, since the blocks of $\mathbf{C}$ are then only positive semidefinite. However, it is shown in [20] that the Laplace operator is coercive over $G_0[n]$ if, for each interface $\Gamma_{ij}$, at least one of the $\omega_k^{ij}$ satisfies $\langle \omega_k^{ij}, 1\rangle_{ij} \neq 0$. In this case, although $\mathbf{C}$ is singular, a unique solution to the entire system of equations exists, owing to the presence of $\mathbf{M}$ and $\mathbf{M}^{\mathrm{T}}$. The estimates in §2 that establish convergence of the approximations hold for Poisson's equation.

We can adapt the *block-elimination algorithm with iterative refinement for bordered linear systems* of Govaerts and Pryce [12] to obtain a solution of the linear system when $\mathbf{C}$ is only positive semidefinite. This method allows us to use a nonsingular matrix close to $\mathbf{C}$ and obtain approximations to a solution using iterations employing the algorithm described above. Experiments suggest that good results are obtained with very few iterations [20]. We need only do steps 2, 3, and 4 above for each iteration; if we provide a Cholesky decomposition for the matrix $\mathbf{MY}$, such iterations are easily computed. Greenstadt has described a way to decouple the system so that each cell is treated independently, and the computations can be done in parallel [14]; this method is valid for Poisson's equation. Iterative techniques are appropriate for problems requiring large numbers of cells; in §4 we discuss an example [15] where the generalized conjugate gradient method [7] was used.

The arguments above require independence of the rows of $\mathbf{M}$. To show that this occurs for sufficiently large $[m]$, we consider a representative cell, say $\Omega_1$, and $C^1$ faces $\Gamma_{12}$ and $\Gamma_{13}$. Denote the Schauder basis for $H^1(\Omega_1)$ by $\{B_n\}$. Inner products $\langle \cdot, \cdot \rangle_{12}$ and $\langle \cdot, \cdot \rangle_{13}$ are both denoted $\langle \cdot, \cdot \rangle$; identification is carried by the subscript $1i$ on $\gamma_{1i}$ or the superscript on $\omega_j^{1i}$.

We assume that Schauder bases $\{\omega_p^{1i}\}$ for $L_2(\Gamma_{1i})$ are in $H^{1/2}(\Gamma_{1i})$. In this paper this assumption is only used to establish that the rows of $\mathbf{M}$ are independent. It means that we can use the full force of the trace theorem [17, p. 37], so that for any finite linear combination $g$ of the $\{\omega_k^{ij}\}$ there is some $v$ in $H^1(\Omega_i)$ such that $\gamma_{ij}(v) = g$.

Without loss of generality, we assume that $\|\omega_p^{1i}\|_{1i} = 1$ for all $p \leq n_{1i}$, $i = 2, 3$.

By $C$ we denote a constant such that, for any $u \in H^1(\Omega_1)$,

$$\|\gamma_{1i}(u)\|_{1i} \leq C\|u\|_{1,1}.$$

For any $m$, the segments of the collocation rows relevant to $\Omega_1$ are

$$
\begin{aligned}
&(\langle \gamma_{12}(B_1), \omega_1^{12}\rangle, \langle \gamma_{12}(B_2), \omega_1^{12}\rangle, \ldots, \langle \gamma_{12}(B_m), \omega_1^{12}\rangle)\\
&(\langle \gamma_{12}(B_1), \omega_2^{12}\rangle, \langle \gamma_{12}(B_2), \omega_2^{12}\rangle, \ldots, \langle \gamma_{12}(B_m), \omega_2^{12}\rangle)\\
&\quad\vdots\\
&(\langle \gamma_{12}(B_1), \omega_{n_{12}}^{12}\rangle, \langle \gamma_{12}(B_2), \omega_{n_{12}}^{12}\rangle, \ldots, \langle \gamma_{12}(B_m), \omega_{n_{12}}^{12}\rangle)\\
&(\langle \gamma_{13}(B_1), \omega_1^{13}\rangle, \langle \gamma_{13}(B_2), \omega_1^{13}\rangle, \ldots, \langle \gamma_{13}(B_m), \omega_1^{13}\rangle)\\
&(\langle \gamma_{13}(B_1), \omega_2^{13}\rangle, \langle \gamma_{13}(B_2), \omega_2^{13}\rangle, \ldots, \langle \gamma_{13}(B_m), \omega_2^{13}\rangle)\\
&\quad\vdots\\
&(\langle \gamma_{13}(B_1), \omega_{n_{13}}^{13}\rangle, \langle \gamma_{13}(B_2), \omega_{n_{13}}^{13}\rangle, \ldots, \langle \gamma_{13}(B_m), \omega_{n_{13}}^{13}\rangle).
\end{aligned}
$$

We first show that any of these rows is not 0 if $m$ is sufficiently large.

**Lemma 3.1.** *For any row, there is some $m$ such that the row is not* 0.

*Proof.* Take row

$$(\langle \gamma_{12}(B_1), \omega_1^{12}\rangle, \langle \gamma_{12}(B_2), \omega_1^{12}\rangle, \ldots, \langle \gamma_{12}(B_m), \omega_1^{12}\rangle)$$

as a representative row. By the trace theorem there is some $u \in H^1(\Omega_1)$, $u \neq 0$, such that $\gamma_{12}(u) = \omega_1^{12}$. Since $\{B_j\}$ is a Schauder basis for $H^1(\Omega_1)$, there exists some $m$ and $b_j \in \mathbb{R}$ such that $\| \sum_{j=1}^m b_j B_j - u\|_{1,1} < 1/C$.

First, note that

$$
\begin{aligned}
\left|\left\langle \gamma_{12}\left(\sum_{j=1}^m b_j B_j - u\right), \omega_1^{12}\right\rangle\right| &= \left|\left\langle \sum_{j=1}^m b_j \gamma_{12}(B_j) - \omega_1^{12}, \omega_1^{12}\right\rangle\right|\\
&= \left|\sum_{j=1}^m b_j \langle \gamma_{12}(B_j), \omega_1^{12}\rangle - \langle \omega_1^{12}, \omega_1^{12}\rangle\right| = \left|\sum_{j=1}^m b_j \langle \gamma_{12}(B_j), \omega_1^{12}\rangle - 1\right|.
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
\left|\left\langle \gamma_{12}\left(\sum_{j=1}^m b_j B_j - u\right), \omega_1^{12}\right\rangle\right| &\leq \left\|\gamma_{12}\left(\sum_{j=1}^m b_j B_j - u\right)\right\|_{12} \|\omega_1^{12}\|_{12}\\
&\leq C\left\|\sum_{j=1}^m b_j B_j - u\right\|_{1,1} \cdot 1 < C(1/C) = 1.
\end{aligned}
$$

Thus, $|\sum_{j=1}^m b_j \langle \gamma_{12}(B_j), \omega_1^{12}\rangle - 1| < 1$, so $\sum_{j=1}^m b_j \langle \gamma_{12}(B_j), \omega_1^{12}\rangle \neq 0$, and the result follows. □

*Assume that $m$ has been taken large enough so that none of the rows is* 0.

**Lemma 3.2.** *The integer $m$ can be made large enough so that for any $\Gamma_{1i}$, the rows corresponding to moment collocation on $\Gamma_{1i}$ are independent.*

*Proof.* We take $\Gamma_{12}$ as a representative for our argument and let $n = n_{12}$. Suppose that for any $m$, the $n$ rows are dependent. Then, for each $m$, there

exist $a_{m,j}$, $j = 1, \dots, n$, not all zero, such that

$$0 = \sum_{j=1}^{n} a_{m,j}(\langle \gamma_{12}(B_1), \omega_j^{12}\rangle, \langle \gamma_{12}(B_2), \omega_j^{12}\rangle, \dots, \langle \gamma_{12}(B_m), \omega_j^{12}\rangle).$$

Let $w_m \equiv \sum_{j=1}^{n} a_{m,j}\omega_j^{12}$. This cannot be 0, for we assume that $\{\omega_j^{12}\}$ is a Schauder basis for $L_2(\Gamma_{12})$.

By linearity,

$$0 = (\langle \gamma_{12}(B_1), w_m\rangle, \langle \gamma_{12}(B_2), w_m\rangle, \dots, \langle \gamma_{12}(B_m), w_m\rangle).$$

Thus, $w_m$ is orthogonal to the span of

$$\{\gamma_{12}(B_1), \gamma_{12}(B_2), \dots, \gamma_{12}(B_m)\}.$$

We can assume that $\|w_m\|_{12} = 1$. Hence, $\{w_m\}$ is a bounded set in the finite-dimensional span of $\{\omega_p^{12}: p = 1, \dots, n\}$. Hence, there is some subsequence $w_{m_i}$ and some $z$ in this span such that $w_{m_i} \to z$ in the $L_2(\Gamma_{12})$ norm. Choose $M$ such that $i > M$ implies $\|w_{m_i} - z\|_{12} < 1/2$. By the trace theorem there is some $u \in H^1(\Omega_1)$, $u \neq 0$, such that $\gamma_{12}(u) = z$. Since $\{B_j\}$ is a Schauder basis for $H^1(\Omega_1)$, there exists some $m_i$, with $i > M$, and $b_j \in \mathbb{R}$ such that

$$\left\| \sum_{j=1}^{m_i} b_j B_j - u \right\|_{1,1} < 1/(2C).$$

Then, using these $b_j$, we get

$$\begin{aligned}
\left\| \sum_{j=1}^{m_i} b_j \gamma_{12}(B_j) - w_{m_i} \right\|_{12} &\leq \left\| \sum_{j=1}^{m_i} b_j \gamma_{12}(B_j) - \gamma_{12}(u) \right\|_{12} + \|\gamma_{12}(u) - w_{m_i}\|_{12} \\
&= \left\| \gamma_{12}\left( \sum_{j=1}^{m_i} b_j B_j \right) - \gamma_{12}(u) \right\|_{12} + \|z - w_{m_i}\|_{12} \\
&\leq C \left\| \sum_{j=1}^{m_i} b_j B_j - u \right\|_{1,1} + 1/2 < C(1/(2C)) + 1/2 = 1.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\left\| \sum_{j=1}^{m_i} b_j \gamma_{12}(B_j) - w_{m_i} \right\|_{12}^2 \\
&\quad = \left\langle \sum_{j=1}^{m_i} b_j \gamma_{12}(B_j) - w_{m_i}, \sum_{j=1}^{m_i} b_j \gamma_{12}(B_j) - w_{m_i} \right\rangle \\
&\quad = \left\| \sum_{j=1}^{m_i} b_j \gamma_{12}(B_j) \right\|_{12}^2 + \|w_{m_i}\|_{12}^2 = \left\| \sum_{j=1}^{m_i} b_j \gamma_{12}(B_j) \right\|_{12}^2 + 1,
\end{aligned}$$

since $\sum_{j=1}^{m_i} b_j \gamma_{12}(B_j)$ is orthogonal to $w_{m_i}$. This last result shows that

$$\left\| \sum_{j=1}^{m_i} b_j \gamma_{12}(B_j) - w_{m_i} \right\|_{12}$$

is an expression greater than or equal to 1, yet the preceding estimate showed it to be less than 1, which gives the desired contradiction. □

*Assume that $m$ has been taken large enough so that all rows that represent moment collocation for a single boundary face $\Gamma_{1i}$ are independent.*

**Lemma 3.3.** *The integer $m$ can be taken large enough so that the rows corresponding to moment collocation with $\Gamma_{12}$ and $\Gamma_{13}$ are all independent.*

*Proof.* We show that $m$ can be taken large enough so that the first collocation row for $\Gamma_{13}$ and the collocation rows for $\Gamma_{12}$ comprise a linearly independent set. The process can then be iterated to obtain the proof.

Suppose the result is false for all $m$. Then, for any $m$, there exist $a_{m,j}$, $j = 1, \dots, n$, not all zero, such that

$$(\langle \gamma_{13}(B_1), \omega_1^{13}\rangle, \langle \gamma_{13}(B_2), \omega_1^{13}\rangle, \dots, \langle \gamma_{13}(B_m), \omega_1^{13}\rangle)$$
$$= \sum_{j=1}^{n} a_{m,j}(\langle \gamma_{12}(B_1), \omega_j^{12}\rangle, \langle \gamma_{12}(B_2), \omega_j^{12}\rangle, \dots, \langle \gamma_{12}(B_m), \omega_j^{12}\rangle).$$

Now, assuming that $m$ is large enough so that the collocation rows for $\Gamma_{12}$ are independent, we argue that the $a_{m,j}$ must be independent of $m$. If this were not the case, there would be some $m$ with the properties above and $m' > m$ such that

$$(\langle \gamma_{13}(B_1), \omega_1^{13}\rangle, \langle \gamma_{13}(B_2), \omega_1^{13}\rangle, \dots, \langle \gamma_{13}(B_{m'}), \omega_1^{13}\rangle)$$
$$= \sum_{j=1}^{n} a_{m',j}(\langle \gamma_{12}(B_1), \omega_j^{12}\rangle, \langle \gamma_{12}(B_2), \omega_j^{12}\rangle, \dots, \langle \gamma_{12}(B_{m'}), \omega_j^{12}\rangle).$$

This dependency relationship will hold for the first $m$ components of the vectors above. Then

$$\sum_{j=1}^{n} a_{m,j}(\langle \gamma_{12}(B_1), \omega_j^{12}\rangle, \langle \gamma_{12}(B_2), \omega_j^{12}\rangle, \dots, \langle \gamma_{12}(B_m), \omega_j^{12}\rangle)$$
$$= \sum_{j=1}^{n} a_{m',j}(\langle \gamma_{12}(B_1), \omega_j^{12}\rangle, \langle \gamma_{12}(B_2), \omega_j^{12}\rangle, \dots, \langle \gamma_{12}(B_m), \omega_j^{12}\rangle).$$

Thus

$$\sum_{j=1}^{n}(a_{m,j} - a_{m',j})(\langle \gamma_{12}(B_1), \omega_j^{12}\rangle, \langle \gamma_{12}(B_2), \omega_j^{12}\rangle, \dots, \langle \gamma_{12}(B_m), \omega_j^{12}\rangle) = 0,$$

which would produce a dependency unless $a_{m,j} = a_{m',j}$ for $j = 1, \dots, n$.

Let $z = \sum_{j=1}^{n} a_{m,j}\omega_j^{12}$. Then, by linearity,

$$(\langle \gamma_{13}(B_1), \omega_1^{13}\rangle, \langle \gamma_{13}(B_2), \omega_1^{13}\rangle, \dots, \langle \gamma_{13}(B_m), \omega_1^{13}\rangle)$$
$$= \sum_{j=1}^{n} a_{m,j}(\langle \gamma_{12}(B_1), \omega_j^{12}\rangle, \langle \gamma_{12}(B_2), \omega_j^{12}\rangle, \dots, \langle \gamma_{12}(B_m), \omega_j^{12}\rangle)$$
$$= (\langle \gamma_{12}(B_1), z\rangle, \langle \gamma_{12}(B_2), z\rangle, \dots, \langle \gamma_{12}(B_m), z\rangle)$$

for any $m$. Thus, for any $u$ in the span of $\{B_1, \dots, B_m\}$,

$$\langle \gamma_{13}(u), \omega_1^{13}\rangle = \langle \gamma_{12}(u), z\rangle.$$

We have that $\omega_1^{13} \neq 0$ a.e. is in $L_2(\Gamma_{13})$. The assumption that $\Omega_1$ is an LP$C^1$ domain (so that $\Gamma_{13}$ is $C^1$) and a partition of unity argument allows us to find some open set $\mathscr{O}$ in $\mathbb{R}^K$ with the following properties:

(a) $\mathscr{O} \cap \Gamma_{13} \neq \varnothing$ and $\omega_1^{13} \neq 0$ a.e. on $\mathscr{O} \cap \Gamma_{13}$, and

(b) $\mathscr{O} \cap \Gamma_{12} = \varnothing$.

We can find some $v$ in $C^1(\mathbb{R}^K)$ such that

(i) $v$ has support in $\mathscr{O}$;

(ii) $v \geq 0$; and

(iii) $v > 0$ on an open set in $\Omega_1$ and on a relatively open set in $\mathscr{O} \cap \Gamma_{13}$ and $\langle \gamma_{13}(v), \omega_1^{13} \rangle \neq 0$.

Thus, $|\langle \gamma_{13}(v), \omega_1^{13} \rangle| \equiv d > 0$ and $\langle \gamma_{12}(v), z \rangle = 0$.

Since $\{B_j\}$ is a Schauder basis, we can find some $m$ and $b_j$ such that $\| \sum_{j=1}^m b_j B_j - v \|_{1,1} < d/[C(1 + \|z\|_{12})]$. Let $v_m \equiv \sum_{j=1}^m b_j B_j$. So

$$\langle \gamma_{13}(v_m), \omega_1^{13} \rangle = \langle \gamma_{12}(v_m), z \rangle,$$

$$\begin{aligned} d &= |\langle \gamma_{13}(v), \omega_1^{13} \rangle - \langle \gamma_{12}(v), z \rangle| \\ &= |\langle \gamma_{13}(v), \omega_1^{13} \rangle - \langle \gamma_{13}(v_m), \omega_1^{13} \rangle + \langle \gamma_{12}(v_m), z \rangle - \langle \gamma_{12}(v), z \rangle| \\ &\leq \|\gamma_{13}(v) - \gamma_{13}(v_m)\|_{12} \|\omega_1^{13}\|_{12} + \|\gamma_{12}(v_m) - \gamma_{12}(v)\|_{12} \|z\|_{12} \\ &\leq C\|v - v_m\|_{1,1} \cdot 1 + C\|v - v_m\|_{1,1} \|z\|_{12} \\ &= C\|v - v_m\|_{1,1}(1 + \|z\|_{12}) \\ &= C \left\| \sum_{j=1}^m b_j B_j - v \right\|_{1,1} (1 + \|z\|_{12}) < C\{d/[C(1 + \|z\|_{12})]\}(1 + \|z\|_{12}) = d. \end{aligned}$$

We have obtained the inequality $d < d$, the desired contradiction. □

If there are more than two $C^1$ boundary faces $\Gamma_{1j}$, the argument in Lemma 3.3 can be continued to show independence of all moment collocation rows for the boundary faces for $\Omega_1$. Finally, since this argument holds for any cell, this suffices to show that the entire moment collocation matrix $\mathbf{M}$ consists of linearly independent rows if $[m]$ is made sufficiently large.

For efficiency of computation, it is appropriate to choose both Schauder bases $\{B_k^i\}$ and $\{\omega_k^{ij}\}$ so that for any choice of $[n]$ moment collocations, $[m]$ does not have to be very large to insure that $\mathbf{M}$ is to have full rank. An example of this is described in §4.

A crucial result is that functions in $G_0[n]$ can be approximated by functions in $G_0[n][m]$, which is Lemma 1.2.

Without loss of generality, we can assume that the Schauder basis $\{B_k^i\}$ is orthonormal in $H^1(\Omega_i)$, since the Gram-Schmidt process could produce such a basis from $\{B_k^i\}$, and the projected functions $\mathscr{P}_m^n(v)$ satisfying the moment collocation constraints such that $\mathscr{P}_m^n(v)|_{\Omega_i}$ is in the span of $\{B_1^i, \ldots, B_{m_i}^i\}$ are the same whether or not the basis has been made orthonormal using this process.

Thus, for any $u \in H$, we have $u|_{\Omega_i} = \sum_{k=1}^\infty b_k^i B_k^i$, where $b_k^i = (u, B_k^i)_{1,i}$ and

$$\|u\|_H^2 = \sum_{i=1}^N \|u\|_{1,i}^2 = \sum_{i=1}^N \left[ \sum_{k=1}^\infty (b_k^i)^2 \right].$$

Note that $\mathscr{E}_m^i(u) = \sum_{k=m+1}^\infty b_k^i B_k^i$.

We prove two representative cases of Lemma 1.2. We first consider the one-cell case.

**Lemma 3.4.** *When there is one cell, denoted $\Omega_i$, with only one external boundary segment $\Gamma_{i0}$, there is a constant $K_1$, depending on the cell structure, the choice of bases $\{B_k^i\}$ and $\{\omega_k^{ij}\}$ and $[n]$ such that, for any $v \in G_0[n]$,*

$$\|v - \mathscr{P}_m^n(v)\|_{1,i} \leq K_1 \|\mathscr{Q}_m^i(v)\|_{1,i}.$$

*Proof.* For any $v \in G_0[n]$, we represent $v$ by $\sum_{k=1}^{\infty} b_k^i B_k^i$. Since we assume that $v \in G_0[n]$, we have for $\omega_p^{i0}$ with $p = 1, \ldots, n_{i0}$ that

$$0 = \langle \gamma_{i0}(v), \omega_p^{i0} \rangle_{i0} = \sum_{k=1}^{\infty} b_k^i \langle \gamma_{i0}(B_k^i), \omega_p^{i0} \rangle_{i0}.$$

Let $v_m = \sum_{k=1}^{m} b_k^i B_k^i$. To obtain the projection $\mathscr{P}_m^n(v)$, we seek $w \in G_0[n][m]$ that minimizes $\|v - w\|_{1,i}^2$. Represent $w$ as a perturbation $v_m - e$ of $v_m$, where $e = \sum_{k=1}^{m} e_k B_k^i$. Then

$$\|v - w\|_{1,i}^2 = \|\mathscr{Q}_m^i(v) + v_m - (v_m - e)\|_{1,i}^2 = \|\mathscr{Q}_m^i(v)\|_{1,i}^2 + \|e\|_{1,i}^2.$$

Thus, we must minimize $\|e\|_{1,i}^2 = \sum_{k=1}^{m} e_k^2$ subject to the requirement that $v_m - e \in G_0[n][m]$, or

$$\begin{aligned} 0 &= \langle \gamma_{i0}(v_m - e), \omega_p^{i0} \rangle_{i0} \\ &= \sum_{k=1}^{m} b_k^i \langle \gamma_{i0}(B_k^i), \omega_p^{i0} \rangle_{i0} - \sum_{k=1}^{m} e_k \langle \gamma_{i0}(B_k^i), \omega_p^{i0} \rangle_{i0} \quad \text{for } p = 1, \ldots, n_{i0}. \end{aligned}$$

So $\sum_{k=1}^{m} e_k \langle \gamma_{i0}(B_k^i), \omega_p^{i0} \rangle_{i0} = \sum_{k=1}^{m} b_k^i \langle \gamma_{i0}(B_k^i), \omega_p^{i0} \rangle_{i0}$ must hold. Let $\alpha_p = \sum_{k=1}^{m} b_k^i \langle \gamma_{i0}(B_k^i), \omega_p^{i0} \rangle_{i0}$. We express this in terms of matrices. Let $\mathbf{a} = (\alpha_1, \ldots, \alpha_{n_{i0}})$ and $\mathbf{e} = (e_1, \ldots, e_m)$. The matrix $\mathbf{M}$ denotes the $n_{i0} \times m$ array with rows $\mathbf{m}_p \equiv (\langle \gamma_{i0}(B_1^i), \omega_p^{i0} \rangle_{i0}, \langle \gamma_{i0}(B_2^i), \omega_p^{i0} \rangle_{i0}, \ldots, \langle \gamma_{i0}(B_m^i), \omega_p^{i0} \rangle_{i0})$. The requirement is that $\mathbf{M}\mathbf{e}^{\mathrm{T}} = \mathbf{a}^{\mathrm{T}}$. We assume that $m$ is sufficiently large so that the rows of $\mathbf{M}$ are independent. We wish to minimize $\mathbf{e}\mathbf{e}^{\mathrm{T}}$ such that $\mathbf{M}\mathbf{e}^{\mathrm{T}} = \mathbf{a}^{\mathrm{T}}$. This is a well-known linear programming problem; it has the following solution: $\mathbf{A} \equiv \mathbf{M}\mathbf{M}^{\mathrm{T}}$ is symmetric and nonsingular; it is positive definite, since for any $\mathbf{x} \neq 0$, $\mathbf{x}\mathbf{A}\mathbf{x}^{\mathrm{T}} = (\mathbf{x}\mathbf{M})(\mathbf{x}\mathbf{M})^{\mathrm{T}} > 0$. Let $\mathbf{y}$ be the solution of $\mathbf{y}\mathbf{A} = \mathbf{a}$, so $\mathbf{y} = \mathbf{a}\mathbf{A}^{-1}$, and form $\mathbf{z} \equiv \sum_{k=1}^{n_{i0}} y_k \mathbf{m}_k = \mathbf{y}\mathbf{M}$. Note that $\mathbf{M}\mathbf{z}^{\mathrm{T}} = \mathbf{M}(\mathbf{y}\mathbf{M})^{\mathrm{T}} = \mathbf{A}\mathbf{y}^{\mathrm{T}} = \mathbf{a}^{\mathrm{T}}$, so $\mathbf{z}$ is a possible $\mathbf{e}$. First, $\mathbf{z}\mathbf{z}^{\mathrm{T}} = \mathbf{y}\mathbf{M}(\mathbf{y}\mathbf{M})^{\mathrm{T}} = \mathbf{y}\mathbf{A}\mathbf{y}^{\mathrm{T}} = \mathbf{a}(\mathbf{a}\mathbf{A}^{-1})^{\mathrm{T}} = \mathbf{a}\mathbf{A}^{-1}\mathbf{a}^{\mathrm{T}}$, for $\mathbf{A}^{-1}$ is symmetric. Also, for any $\mathbf{e}$ satisfying the necessary requirement, $\mathbf{z}\mathbf{e}^{\mathrm{T}} = \mathbf{y}\mathbf{M}\mathbf{e}^{\mathrm{T}} = \mathbf{y}\mathbf{a}^{\mathrm{T}} = \mathbf{a}\mathbf{A}^{-1}\mathbf{a}^{\mathrm{T}} = \mathbf{z}\mathbf{z}^{\mathrm{T}}$, so

$$\mathbf{z}(\mathbf{e} - \mathbf{z})^{\mathrm{T}} = \mathbf{0} = (\mathbf{e} - \mathbf{z})\mathbf{z}^{\mathrm{T}}.$$

With these results we show that the $\mathbf{e}$ that minimizes $\sum_{k=1}^{m} e_k^2 = \mathbf{e}\mathbf{e}^{\mathrm{T}}$ subject to the necessary condition $\mathbf{M}\mathbf{e}^{\mathrm{T}} = \mathbf{a}^{\mathrm{T}}$ is $\mathbf{z}$. Now

$$\begin{aligned} \mathbf{e}\mathbf{e}^{\mathrm{T}} &= (\mathbf{e} - \mathbf{z} + \mathbf{z})(\mathbf{e} - \mathbf{z} + \mathbf{z})^{\mathrm{T}} \\ &= (\mathbf{e} - \mathbf{z})(\mathbf{e} - \mathbf{z})^{\mathrm{T}} + \mathbf{z}(\mathbf{e} - \mathbf{z})^{\mathrm{T}} + (\mathbf{e} - \mathbf{z})\mathbf{z}^{\mathrm{T}} + \mathbf{z}\mathbf{z}^{\mathrm{T}} \\ &= (\mathbf{e} - \mathbf{z})(\mathbf{e} - \mathbf{z})^{\mathrm{T}} + 0 + 0 + \mathbf{z}\mathbf{z}^{\mathrm{T}}. \end{aligned}$$

Thus, the optimal $\mathbf{e}$ is $\mathbf{z}$. The minimal value is $\mathbf{z}\mathbf{z}^{\mathrm{T}} = \mathbf{a}\mathbf{A}^{-1}\mathbf{a}^{\mathrm{T}}$. Since $\mathscr{P}_m^n(v) = \sum_{k=1}^m (b_k^i - e_k)B_k^i$, we have

$$\|v - \mathscr{P}_m^n(v)\|_{1,i}^2 = \|\mathscr{E}_m^i(v)\|_{1,i}^2 + \|e\|_{1,i}^2 = \|\mathscr{E}_m^i(v)\|_{1,i}^2 + \mathbf{z}\mathbf{z}^{\mathrm{T}}$$
$$= \|\mathscr{E}_m^i(v)\|_{1,i}^2 + \mathbf{a}\mathbf{A}^{-1}\mathbf{a}^{\mathrm{T}}.$$

We argue that this can be made small for sufficiently large $m$.

If $\mu$ is the smallest eigenvalue of the positive definite matrix $\mathbf{A}$, then it is easily shown that $\mathbf{a}\mathbf{A}^{-1}\mathbf{a}^{\mathrm{T}} \le (1/\mu)\mathbf{a}\mathbf{a}^{\mathrm{T}}$. For any $m' > m$, the matrix of moment collocation rows $\mathbf{M}'$, when more basis functions are employed, is formed by adjoining a matrix $\mathbf{M}_1$ with rows of form

$$(\langle \gamma_{i0}(B_{m+1}^i), \omega_p^{i0}\rangle_{i0}, \langle \gamma_{i0}(B_{m+2}^i), \omega_p^{i0}\rangle_{i0}, \ldots, \langle \gamma_{i0}(B_{m'}^i), \omega_p^{i0}\rangle_{i0})$$

to $\mathbf{M}$; $\mathbf{M}'$ can be presented as $(\mathbf{M}\ \mathbf{M}_1)$. Define

$$\mathbf{A}' = (\mathbf{M}\ \mathbf{M}_1)(\mathbf{M}\ \mathbf{M}_1)^{\mathrm{T}} = \mathbf{M}\mathbf{M}^{\mathrm{T}} + \mathbf{M}_1\mathbf{M}_1^{\mathrm{T}}.$$

Let $\mu = \inf \mathbf{w}\mathbf{A}\mathbf{w}^{\mathrm{T}} = \mathbf{w}\mathbf{M}(\mathbf{w}\mathbf{M})^{\mathrm{T}}$, where the infimum is taken over all $\mathbf{w}$ such that $\mathbf{w}\mathbf{w}^{\mathrm{T}} = 1$. If $\mu'$ is the least eigenvalue for $\mathbf{A}'$, with the same assumptions about $\mathbf{w}$, then

$$\mu' = \inf \mathbf{w}\mathbf{A}'\mathbf{w}^{\mathrm{T}} = \inf(\mathbf{w}\mathbf{M}\mathbf{M}^{\mathrm{T}}\mathbf{w}^{\mathrm{T}} + \mathbf{w}\mathbf{M}_1\mathbf{M}_1^{\mathrm{T}}\mathbf{w}^{\mathrm{T}})$$
$$= \inf(\mathbf{w}\mathbf{M}(\mathbf{w}\mathbf{M})^{\mathrm{T}} + \mathbf{w}\mathbf{M}_1(\mathbf{w}\mathbf{M}_1)^{\mathrm{T}})$$
$$\ge \mu + \inf(\mathbf{w}\mathbf{M}_1(\mathbf{w}\mathbf{M}_1)^{\mathrm{T}}) \ge \mu.$$

Thus, $1/\mu' \le 1/\mu$.

We assumed that $v \in G_0[n]$, so

$$0 = \langle \gamma_{i0}(v), \omega_p^{i0}\rangle_{i0} = \sum_{k=1}^{\infty} b_k^i \langle \gamma_{i0}(B_k^i), \omega_p^{i0}\rangle_{i0}.$$

Thus,

$$\alpha_p = \sum_{k=1}^{m} b_k^i \langle \gamma_{i0}(B_k^i), \omega_p^{i0}\rangle_{i0} = -\sum_{k=m+1}^{\infty} b_k^i \langle \gamma_{i0}(B_k^i), \omega_p^{i0}\rangle_{i0}$$
$$= -\langle \gamma_{i0}(\mathscr{E}_m^i(v)), \omega_p^{i0}\rangle_{i0}.$$

Using Schwarz's inequality and the trace theorem, we get

$$\alpha_p^2 \le \|\gamma_{i0}(\mathscr{E}_m^i(v))\|_{i0}^2 \|\omega_p^{i0}\|_{i0}^2 \le C_{i0}^2 \|\mathscr{E}_m^i(v)\|_{1,i}^2 \|\omega_p^{i0}\|_{i0}^2.$$

This gives the estimate

$$\|v - \mathscr{P}_m^n(v)\|_{1,i}^2 = \|\mathscr{E}_m^i(v)\|_{1,i}^2 + \mathbf{a}\mathbf{A}^{-1}\mathbf{a}^{\mathrm{T}} \le \|\mathscr{E}_m^i(v)\|_{1,i}^2 + (1/\mu)\mathbf{a}\mathbf{a}^{\mathrm{T}}$$
$$= \|\mathscr{E}_m^i(v)\|_{1,i}^2 + (1/\mu)\sum_{p=1}^{n_{i0}} \alpha_p^2$$
$$\le \|\mathscr{E}_m^i(v)\|_{1,i}^2 + (1/\mu)\sum_{p=1}^{n_{i0}} C_{i0}^2 \|\mathscr{E}_m^i(v)\|_{1,i}^2 \|\omega_p^{i0}\|_{i0}^2$$
$$= \|\mathscr{E}_m^i(v)\|_{1,i}^2 \left(1 + (1/\mu)C_{i0}^2 \sum_{p=1}^{n_{i0}} \|\omega_p^{i0}\|_{i0}^2\right). \quad \square$$

The next case we consider is a domain partitioned into two cells with one internal interface.

**Lemma 3.5.** *When there are two cells, denoted* $\Omega_1$ *and* $\Omega_2$, *with boundaries* $\Gamma_{10}, \Gamma_{12}$ *for* $\Omega_1$ *and* $\Gamma_{20}$ *and* $\Gamma_{21} = \Gamma_{12}$ *for* $\Omega_2$, *there is a constant* $K_1$, *independent of the length of the moment collocation rows, that depends only on the moment collocation rows (assumed to be independent), constants* $C_{ij}$ *obtained by the trace theorem and the* $L_2(\Gamma_{i0})$ *norms of the first* $n_{i0}, n_{12}$, *and* $n_{20}$ *weight functions* $\omega_p^{ij}$ *such that for any* $v$ *in* $G_0[n]$, *with* $[m] \equiv (m_1, m_2)$,

$$\|v - \mathscr{P}_m^n(v)\|_H \leq K_1 \|\mathscr{C}_{[m]}(v)\|_H .$$

*Proof.* For any $v \in G_0[n]$, we represent $v|_{\Omega_i}$ by $v^i \equiv \sum_{k=1}^{\infty} b_k^i B_k^i$ for $i = 1, 2$. Since we assume that $v \in G_0[n]$, for $i = 1, 2$ and for $\omega_p^{i0}$, $p = 1, \ldots, n_{i0}$, we have

$$0 = \langle \gamma_{i0}(v^i), \omega_p^{i0} \rangle_{i0} = \sum_{k=1}^{\infty} b_k^i \langle \gamma_{i0}(B_k^i), \omega_p^{i0} \rangle_{i0}$$

and, for $\omega_p^{12}$, $p = 1, \ldots, n_{12}$, in addition

$$\begin{aligned} 0 &= \langle \gamma_{12}(v^1), \omega_p^{12} \rangle_{12} - \langle \gamma_{21}(v^2), \omega_p^{12} \rangle_{21} \\ &= \sum_{k=1}^{\infty} b_k^1 \langle \gamma_{12}(B_k^1), \omega_p^{12} \rangle_{12} - \sum_{j=1}^{\infty} b_j^2 \langle \gamma_{21}(B_j^2), \omega_p^{12} \rangle_{21} . \end{aligned}$$

We express $[m]$ as $(m_1, m_2)$. Let $v_{m_i} = \sum_{k=1}^{m_i} b_k^i B_k^i$. To obtain the projection $\mathscr{P}_m^n(v)$, we seek $w \in G_0[n][m]$ that minimizes $\|v - w\|_H^2$.

Represent $w$ as a pair $(w^1, w^2)$, where, for $i = 1, 2$,

$$w^i = w|_{\Omega_i} \equiv v_{m_i} - e^i \quad \text{and} \quad e^i = \sum_{k=1}^{m_i} e_k^i B_k^i .$$

Then

$$\begin{aligned} \|v - w\|_H^2 &= \|v^1 - w^1\|_{1,1}^2 + \|v^2 - w^2\|_{1,2}^2 \\ &= \|\mathscr{C}_{m_1}^1(v) + v_{m_1}^1 - (v_{m_1}^1 - e^1)\|_{1,1}^2 + \|\mathscr{C}_{m_2}^2(v) + v_{m_2}^2 - (v_{m_2}^2 - e^2)\|_{1,2}^2 \\ &= \|\mathscr{C}_{m_1}^1(v)\|_{1,1}^2 + \|e^1\|_{1,1}^2 + \|\mathscr{C}_{m_2}^2(v)\|_{1,2}^2 + \|e^2\|_{1,2}^2 . \end{aligned}$$

Thus, we must minimize $\|e^1\|_{1,1}^2 + \|e^2\|_{1,2}^2 = \sum_{k=1}^{m_1} (e_k^1)^2 + \sum_{k=1}^{m_2} (e_k^1)^2$ subject to the requirement that $w \in G_0[n][m]$, or, first, for $i = 1, 2$,

$$\begin{aligned} 0 &= \langle \gamma_{i0}(w^i), \omega_p^{i0} \rangle_{i0} = \langle \gamma_{i0}(v_{m_i} - e^i), \omega_p^{i0} \rangle_{i0} \\ &= \sum_{k=1}^{m_i} b_k^i \langle \gamma_{i0}(B_k^i), \omega_p^{i0} \rangle_{i0} - \sum_{k=1}^{m_i} e_k^i \langle \gamma_{i0}(B_k^i), \omega_p^{i0} \rangle_{i0} \quad \text{for } p = 1, \ldots, n_{i0} . \end{aligned}$$

So

$$\sum_{k=1}^{m_i} e_k^i \langle \gamma_{i0}(B_k^i), \omega_p^{i0} \rangle_{i0} = \sum_{k=1}^{m_i} b_k^i \langle \gamma_{i0}(B_k^i), \omega_p^{i0} \rangle_{i0}$$

must hold. Let $\alpha_p^i = \sum_{k=1}^{m_i} b_k^i \langle \gamma_{i0}(B_k^i), \omega_p^{i0} \rangle_{i0}$.

We represent the second requirement as

$$-\sum_{k=1}^{m_1} (b_k^1 - e_k^1) \langle \gamma_{12}(B_k^1), \omega_p^{12} \rangle_{12} = -\sum_{j=1}^{m_2} (b_j^2 - e_j^2) \langle \gamma_{21}(B_j^2), \omega_p^{12} \rangle_{21}$$

or

$$\sum_{k=1}^{m_1} e_k^1 \langle \gamma_{12}(B_k^1), \omega_p^{12} \rangle_{12} - \sum_{j=1}^{m_2} e_j^2 \langle \gamma_{21}(B_j^2), \omega_p^{12} \rangle_{21}$$
$$= \sum_{k=1}^{m_1} b_k^1 \langle \gamma_{12}(B_k^1), \omega_p^{12} \rangle_{12} - \sum_{j=1}^{m_2} b_j^2 \langle \gamma_{21}(B_j^2), \omega_p^{12} \rangle_{21}.$$

Let

$$\alpha_p^{12} \equiv \sum_{k=1}^{m_1} b_k^1 \langle \gamma_{12}(B_k^1), \omega_p^{12} \rangle_{12} - \sum_{j=1}^{m_2} b_j^2 \langle \gamma_{21}(B_j^2), \omega_p^{12} \rangle_{21}.$$

We express these requirements in terms of matrices. Let $\mathbf{a}^i$ denote $(\alpha_1^i, \ldots, \alpha_{n_{i0}}^i)$ and $\mathbf{a}^{12} = (\alpha_1^{12}, \ldots, \alpha_{n_{12}}^{12})$. Put $\mathbf{a} \equiv (\mathbf{a}^1\ \mathbf{a}^{12}\ \mathbf{a}^2)$. Let $\mathbf{e}^i \equiv (e_1^i, \ldots, e_{m_i}^i)$ and $\mathbf{e} \equiv (\mathbf{e}^1\ \mathbf{e}^2)$, and $\mathbf{M}$ be the $(n_{10} + n_{12} + n_{20}) \times (m_1 + m_2)$ array formed in the following fashion. Let $\mathbf{M}_{i0}$ be the $n_{i0} \times m_i$ array with rows

$$(\langle \gamma_{i0}(B_1^i), \omega_p^{i0} \rangle_{i0}, \langle \gamma_{i0}(B_2^i), \omega_p^{i0} \rangle_{i0}, \ldots, \langle \gamma_{i0}(B_{m_i}^i), \omega_p^{i0} \rangle_{i0}),$$

$\mathbf{M}_{12}$ be the $n_{12} \times m_1$ array with rows

$$(\langle \gamma_{12}(B_1^1), \omega_p^{12} \rangle_{12}, \langle \gamma_{12}(B_2^1), \omega_p^{12} \rangle_{12}, \ldots, \langle \gamma_{12}(B_{m_1}^1), \omega_p^{12} \rangle_{12}),$$

and $\mathbf{M}_{21}$ be the $n_{12} \times m_2$ array with rows

$$(-\langle \gamma_{21}(B_1^2), \omega_p^{12} \rangle_{21}, -\langle \gamma_{21}(B_2^2), \omega_p^{12} \rangle_{21}, \ldots, -\langle \gamma_{21}(B_{m_2}^2), \omega_p^{12} \rangle_{21}).$$

Then $\mathbf{M}$ is the array

$$\begin{pmatrix} \mathbf{M}_{10} & \mathbf{0} \\ \mathbf{M}_{12} & \mathbf{M}_{21} \\ \mathbf{0} & \mathbf{M}_{20} \end{pmatrix}.$$

The requirement is $\mathbf{M}\mathbf{e}^{\mathrm{T}} = \mathbf{a}^{\mathrm{T}}$.

We proceed in a manner similar to the previous proof. We assume that $m$ is sufficiently large so that the rows of $\mathbf{M}$ are independent. Form the $(n_{10}+n_{12}+n_{20}) \times (n_{10}+n_{12}+n_{20})$ matrix $\mathbf{A} \equiv \mathbf{M}\mathbf{M}^{\mathrm{T}}$. It is symmetric, nonsingular, and positive definite by the previous argument. Let $\mathbf{y}$ be the solution of $\mathbf{y}\mathbf{A} = \mathbf{a}$, so $\mathbf{y} = \mathbf{a}\mathbf{A}^{-1}$. The vector $\mathbf{y}$ has length $(n_{10} + n_{12} + n_{20})$. Represent $\mathbf{y}$ as $(\mathbf{y}^1\ \mathbf{y}^{12}\ \mathbf{y}^2)$, where $\mathbf{y}^1$ has length $n_{10}$, $\mathbf{y}^{12}$ has length $n_{12}$, and $\mathbf{y}^2$ has length $n_{20}$. Let $\mathbf{z}$ be the row vector of length $(m_1 + m_2)$ defined by

$$\mathbf{z} = \mathbf{y}\mathbf{M} = (y^1\mathbf{M}_{10} + y^{12}\mathbf{M}_{12}\ y^{12}\mathbf{M}_{21} + y^2\mathbf{M}_{20}).$$

As before, $\mathbf{z}\mathbf{z}^{\mathrm{T}} = \mathbf{y}\mathbf{M}(\mathbf{y}\mathbf{M})^{\mathrm{T}} = \mathbf{y}\mathbf{A}\mathbf{y}^{\mathrm{T}} = \mathbf{a}(\mathbf{a}\mathbf{A}^{-1})^{\mathrm{T}} = \mathbf{a}\mathbf{A}^{-1}\mathbf{a}^{\mathrm{T}}$ and $\mathbf{M}\mathbf{z}^{\mathrm{T}} = \mathbf{a}^{\mathrm{T}}$, so $\mathbf{z}$ is a possible $\mathbf{e}$.

The argument in the previous theorem shows that

$$\mathbf{e} = (\mathbf{e}^1 \vdots \mathbf{e}^2) = (\mathbf{y}^1\mathbf{M}_{10} + \mathbf{y}^{12}\mathbf{M}_{12}\ \mathbf{y}^{12}\mathbf{M}_{21} + \mathbf{y}^2\mathbf{M}_{20}) = \mathbf{z}$$

defines a vector whose components minimize $\sum_{k=1}^{m_1}(e_k^1)^2 + \sum_{k=1}^{m_2}(e_k^2)^2 = \mathbf{e}\mathbf{e}^{\mathrm{T}}$ subject to the necessary condition $\mathbf{M}\mathbf{e}^{\mathrm{T}} = \mathbf{a}^{\mathrm{T}}$. The minimal value is $\mathbf{z}\mathbf{z}^{\mathrm{T}} = \mathbf{a}\mathbf{A}^{-1}\mathbf{a}^{\mathrm{T}}$. As before, $\mathscr{P}_m^n(v)|_{\Omega_i} = \sum_{k=1}^{m_i}(b_k^i - e_k^i)B_k^i$, so

$$\|v - \mathscr{P}_m^n(v)\|_H^2 = \|\mathscr{E}_{m_1}^1(v)\|_{1,1}^2 + \|e^1\|_{1,1}^2 + \|\mathscr{E}_{m_2}^2(v)\|_{1,2}^2 + \|e^2\|_{1,2}^2$$
$$= \|\mathscr{E}_{m_1}^1(v)\|_{1,1}^2 + \|\mathscr{E}_{m_2}^2(v)\|_{1,2}^2 + \mathbf{a}\mathbf{A}^{-1}\mathbf{a}^{\mathrm{T}}.$$

To continue with the argument of the previous theorem, we must show that when the number of basis functions employed on each cell is increased to $(m_1', m_2')$, the smallest eigenvalue of the resulting matrix $\mathbf{A}'$ is greater than or equal to the smallest eigenvalue $\mu$ of $\mathbf{A}$. The matrix of moment collocation rows $\mathbf{M}'$, when more basis functions are employed, is formed by augmenting the matrix $\mathbf{M}$ with additional entries so that it has the form

$$\mathbf{M}' = \begin{pmatrix} \mathbf{M}_{10} & \mathbf{M}'_{10} & \mathbf{0} & \mathbf{0} \\ \mathbf{M}_{12} & \mathbf{M}'_{12} & \mathbf{M}_{21} & \mathbf{M}'_{21} \\ \mathbf{0} & \mathbf{0} & \mathbf{M}_{20} & \mathbf{M}'_{20} \end{pmatrix},$$

where, for example, $\mathbf{M}'_{i0}$ is a matrix whose rows are

$$(\langle \gamma_{i0}(B^i_{m_i+1}), \omega^{i0}_p \rangle_{i0}, \langle \gamma_{i0}(B^i_{m_i+2}), \omega^{i0}_p \rangle_{i0}, \ldots, \langle \gamma_{i0}(B^i_{m'_i}), \omega^{i0}_p \rangle_{i0}).$$

We represent the matrix $\mathbf{M}'$ as the sum of two matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ in the following fashion:

$$\mathbf{M}' = \mathbf{M}_1 + \mathbf{M}_2 \equiv \begin{pmatrix} \mathbf{M}_{10} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{M}_{12} & \mathbf{0} & \mathbf{M}_{21} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{M}_{20} & \mathbf{0} \end{pmatrix} + \begin{pmatrix} \mathbf{0} & \mathbf{M}'_{10} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}'_{12} & \mathbf{0} & \mathbf{M}'_{21} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{M}'_{20} \end{pmatrix}.$$

We have $\mathbf{M}_1\mathbf{M}_2^{\mathrm{T}} = 0$, and $\mathbf{M}_1\mathbf{M}_1^{\mathrm{T}} = \mathbf{A}$, so $\mathbf{M}'(\mathbf{M}')^{\mathrm{T}} = \mathbf{A} + \mathbf{M}_2\mathbf{M}_2^{\mathrm{T}}$, and the analysis in the previous theorem establishes the result.

Thus,

$$\mathbf{a}\mathbf{A}^{-1}\mathbf{a}^{\mathrm{T}} \leq (1/\mu)\mathbf{a}\mathbf{a}^{\mathrm{T}} = (1/\mu)\left(\sum_{k=1}^{n_{10}} (\alpha_k^1)^2 + \sum_{k=1}^{n_{12}} (\alpha_k^{12})^2 + \sum_{k=1}^{n_{20}} (\alpha_k^2)^2\right).$$

As in the previous lemma,

$$\begin{aligned} \alpha_p^i &= \sum_{k=1}^{m_i} b_k^i \langle \gamma_{i0}(B_k^i), \omega_p^{i0} \rangle_{i0} = -\sum_{k=m_i+1}^{\infty} b_k^i \langle \gamma_{i0}(B_k^i), \omega_p^{i0} \rangle_{i0} \\ &= -\langle \gamma_{i0}(\mathscr{E}^i_{m_i}(v)), \omega_p^{i0} \rangle_{i0}, \end{aligned}$$

so

$$(\alpha_p^i)^2 \leq \|\gamma_{i0}(\mathscr{E}^i_{m_i}(v))\|^2_{i0} \|\omega_p^{i0}\|^2_{i0} \leq C^2_{i0} \|\mathscr{E}^i_{m_i}(v)\|^2_{1,i} \|\omega_p^{i0}\|^2_{i0};$$

$$\begin{aligned} \alpha_p^{12} &\equiv \sum_{k=1}^{m_1} b_k^1 \langle \gamma_{12}(B_k^1), \omega_p^{12} \rangle_{12} - \sum_{j=1}^{m_2} b_j^2 \langle \gamma_{21}(B_j^2), \omega_p^{12} \rangle_{21} \\ &= -\sum_{k=m_1+1}^{\infty} b_k^1 \langle \gamma_{12}(B_k^1), \omega_p^{12} \rangle_{12} + \sum_{j=m_2+1}^{\infty} b_j^2 \langle \gamma_{21}(B_j^2), \omega_p^{12} \rangle_{21} \\ &= -\langle \gamma_{12}(\mathscr{E}^1_{m_1}(v)), \omega_p^{12} \rangle_{12} + \langle \gamma_{21}(\mathscr{E}^2_{m_2}(v)), \omega_p^{12} \rangle_{12} \\ &= \langle \gamma_{21}(\mathscr{E}^2_{m_2}(v)) - \gamma_{12}(\mathscr{E}^1_{m_1}(v)), \omega_p^{12} \rangle_{12}. \end{aligned}$$

Schwarz's inequality and the trace theorem give

$$\begin{aligned} (\alpha_p^{12})^2 &\leq \|\gamma_{21}(\mathscr{E}^2_{m_2}(v)) - \gamma_{12}(\mathscr{E}^1_{m_1}(v))\|^2_{12} \|\omega_p^{12}\|^2_{12} \\ &\leq (\|\gamma_{21}(\mathscr{E}^2_{m_2}(v))\|_{12} + \|\gamma_{12}(\mathscr{E}^1_{m_1}(v))\|_{12})^2 \|\omega_p^{12}\|^2_{12} \\ &\leq 2(\|\gamma_{21}(\mathscr{E}^2_{m_2}(v))\|^2_{12} + \|\gamma_{12}(\mathscr{E}^1_{m_1}(v))\|^2_{12}) \|\omega_p^{12}\|^2_{12} \\ &\leq 2(C_{21}\|\mathscr{E}^2_{m_2}(v)\|^2_{1,2} + C_{12}\|\mathscr{E}^1_{m_1}(v)\|^2_{1,1}) \|\omega_p^{12}\|^2_{12}. \end{aligned}$$

So

$$\sum_{k=1}^{n_{10}}(\alpha_k^1)^2+\sum_{k=1}^{n_{12}}(\alpha_k^{12})^2+\sum_{k=1}^{n_{20}}(\alpha_k^2)^2$$

$$\leq \sum_{k=1}^{n_{10}}(C_{10}^2\|\mathscr{Q}_{m_1}^1(v)\|_{1,1}^2\|\omega_k^{10}\|_{10}^2)$$

$$+\sum_{k=1}^{n_{12}}2(C_{21}^2\|\mathscr{Q}_{m_2}^2(v)\|_{1,2}^2+C_{12}^2\|\mathscr{Q}_{m_1}^1(v)\|_{1,1}^2)\|\omega_k^{12}\|_{12}^2$$

$$+\sum_{k=1}^{n_{20}}(C_{20}^2\|\mathscr{Q}_{m_2}^2(v)\|_{1,2}^2\|\omega_k^{20}\|_{20}^2)$$

$$=\|\mathscr{Q}_{m_1}^1(v)\|_{1,1}^2\left(C_{10}^2\sum_{k=1}^{n_{10}}\|\omega_k^{10}\|_{10}^2+2C_{12}^2\sum_{k=1}^{n_{12}}\|\omega_k^{12}\|_{12}^2\right)$$

$$+\|\mathscr{Q}_{m_2}^2(v)\|_{1,2}^2\left(C_{20}^2\sum_{k=1}^{n_{20}}\|\omega_k^{20}\|_{20}^2+2C_{21}^2\sum_{k=1}^{n_{12}}\|\omega_k^{12}\|_{12}^2\right).$$

Thus,

$$\|v-\mathscr{P}_m^n(v)\|_H^2=\|\mathscr{Q}_{m_1}^1(v)\|_{1,1}^2+\|\mathscr{Q}_{m_2}^2(v)\|_{1,2}^2+\mathbf{a}\mathbf{A}^{-1}\mathbf{a}^{\mathrm{T}}$$

$$\leq \|\mathscr{Q}_{m_1}^1(v)\|_{1,1}^2+\|\mathscr{Q}_{m_2}^2(v)\|_{1,2}^2+(1/\mu)\mathbf{a}\mathbf{a}^{\mathrm{T}}$$

$$\leq \|\mathscr{Q}_{m_1}^1(v)\|_{1,1}^2\left\{1+(1/\mu)\left(C_{10}^2\sum_{k=1}^{n_{10}}\|\omega_k^{10}\|_{10}^2+2C_{12}^2\sum_{k=1}^{n_{12}}\|\omega_k^{12}\|_{12}^2\right)\right\}$$

$$+\|\mathscr{Q}_{m_2}^2(v)\|_{1,2}^2\left\{1+(1/\mu)\left(C_{20}^2\sum_{k=1}^{n_{20}}\|\omega_k^{20}\|_{20}^2+2C_{21}^2\sum_{k=1}^{n_{12}}\|\omega_k^{12}\|_{12}^2\right)\right\}.$$

If $K_1^2$ is the maximum of the expressions in the brackets $\{\cdot\}$, we obtain the desired estimate

$$\|v-\mathscr{P}_m^n(v)\|_H^2\leq K_1^2(\|\mathscr{Q}_{m_1}^1(v)\|_{1,1}^2+\|\mathscr{Q}_{m_2}^2(v)\|_{1,2}^2)=K_1^2\|\mathscr{Q}_{[m]}(v)\|_H^2. \quad \square$$

This result generalizes to prove Lemma 1.2. The following estimate holds:

Suppose that $n(\Omega_i)$ is the total number of moment collocations employed on all interfaces $\Gamma_{ij}$ of cell $\Omega_i$. Let $n_c=\sup\{n(\Omega_i)\}$. Suppose that $C=\sup\{C_{ij}^2\}$ and $\|\omega_p^{ij}\|_{ij}^2\leq W$ for all $(ij)$; then $K_1^2$ is bounded by $1+2(1/\mu)CWn_c$. Thus,

$$\|v-\mathscr{P}_m^n(v)\|_H\leq\sqrt{1+2(1/\mu)CWn_c}\|\mathscr{Q}_{[m]}(v)\|_H.$$

The argument above is based on the assumption that $\{B_k^i\}$ is an orthonormal basis for $H^1(\Omega_i)$. The result holds for the more general case, since we have established the previous result in terms of $\mathscr{Q}_m^i(v)$. Note that this shows that there is no particular advantage in choosing an orthonormal basis for $H^1(\Omega_i)$.

In the case where the basis functions $\{B_k^i\}$ are orthonormal in $H^1(\Omega_i)$, $\mu$ is the smallest eigenvalue for the positive definite matrix $\mathbf{M}\mathbf{M}^{\mathrm{T}}$ and this eigenvalue is nondecreasing as $[m]$ increases. It is an open question whether or not there exist bases such that there is some $C>0$ such that, for any choice of $[n]$, there is a sufficiently large $[m]([n])$ so that $\mu>C$ for $[m']\geq$

$[m]([n])$. Using Legendre polynomials to both provide the weight functions $\{\omega_k^{ij}\}$ and generate a basis for a square in the manner described in §4, we have computed $1/\mu$ for at most 13 moments on each side of the square and have shown that if we make the number of basis functions $m$ sufficiently large ($m = 190$ when eight moments are used), then $1/\mu < 337$. In our experiments, we typically use three or four moments and usually no more than 28 basis functions; in this case, $1/\mu < 67$. Our proof that the Lagrange multipliers $\lambda$ can be used to approximate the conormal derivative of the solution $u$ on the $\Gamma_{ij}$ requires a similar assumption [20]. An example is presented in [6], where a good approximation to the conormal derivative is obtained using the values of the Lagrange multipliers.

## 4. Experimental results

We have written programs that produce approximations to solutions of problems with domains in $\mathbb{R}^2$ [20]. We accommodate four types of cells. Cells can be parallelograms (type 4) or triangles (type 3) in any orientation. Two kinds of cells with one curved (external) boundary segment are accepted; the first has one straight side and one curved side (type 1); the second has two straight sides and one curved side (type 2). Typical cells are shown in Figure 1.

Domains need not be simply connected and can have cracks. We treat Neumann problems and mixed problems as well as the Dirichlet problem; convergence of the approximations to the solution of the nonhomogeneous Dirichlet problem is shown in [20] as well as convergence of approximations to a Neumann and a mixed problem. The coefficients $A_{ij}(x)$ and $A_0(x)$ need not be constant.

In the first preprocessor stage of the program, the user describes the cells by indicating the corner points and the type of cell; FORTRAN formulas are entered giving the parametric representation for any curved boundary segments and the formulae for $A_{ij}$, $A_0$, and $f$.

The second part of the program determines which cells are adjacent and subdivides sides of a cell if necessary, for adjacent cells need not always share corner points. For example, a square can be decomposed into three cells, one rectangular and the other two square, so that the common boundary of the two square cells meets the rectangle in a "T". The program splits the internal boundary of the rectangle into two segments, one for each of the squares.

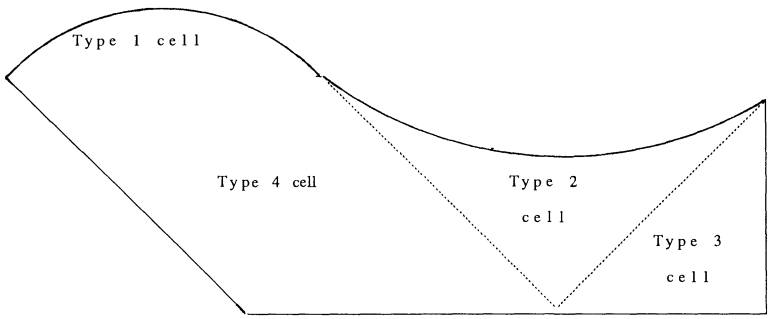


Figure 1. Typical cells

Legendre polynomials are used to generate a basis for a square, which provides a basis for any parallelogram by the use of affine transformations. An interesting $L_2$-orthonormal polynomial basis has been contrived for triangles. On the standard simplex, by use of symmetry considerations, we roughly halve the number of basis functions to be generated. Each basis function $B(x_1, x_2)$ is either symmetric (so $B(x_1, x_2) = B(x_2, x_1)$), or $B(x_2, x_1)$ is orthogonal to $B(x_1, x_2)$ and provides a further basis function. These two bases are adapted for use in type-1 and type-2 cells.

Since we use polynomial bases, if a cell is a parallelogram or a triangle, Gaussian quadrature is appropriate for computing the entries $a(B_i^k, B_j^k)_k$ for the diagonal blocks in the matrix $\mathbf{C}$ [10]. In addition, if the $A_{ij}$ and $A_0$ are constant, appropriate values for a representative square and triangle are stored and the affine transformations used to quickly generate the entries for all blocks in $\mathbf{C}$ corresponding to parallelogram or triangular cells.

The quadrature required on the two types of cells with curved boundaries is the most time consuming part of the program, although such computations are done in parallel. For a type-1 cell, we use an affine transformation to move the cell to fit inside the unit square so that the curved boundary is represented by a function. (Such a representation is required of type-1 cells.) Gaussian quadrature is then used to integrate in the vertical direction. The upper limit representing the curved boundary segment may not be expressed by a polynomial, so we may not be able to depend on the special properties of Gaussian quadrature for accurate results when integrating in the horizontal direction. Thus an adaptive Romberg scheme is used to integrate in the horizontal direction; if the Romberg method fails to meet the convergence criteria set by the program, this fact is noted as the program is executed. It is then usually necessary to return to the preprocessor to further subdivide the cell giving the difficulty. A similar method is used for type-2 cells. More experimentation is needed to find a method that is faster and equally accurate.

We use Legendre polynomials for the weight functions $\omega_k^{ij}$. The computations for moment collocation take little time. They are done using Gaussian quadrature if an interface is a straight line; a variant of Simpson's rule is used to compute the moments on curved boundary segments.

Two parameters are set by the user to determine the number of basis functions for each cell. Parameter "nmc" represents the number of moment collocations to be enforced on each interface; parameter "edf" represents extra degrees of freedom so that on any cell $\Omega_i$, the total number of basis functions is (number of sides $\Gamma_{ij}$ for cell $i$) $\times$ nmc + edf. We typically set "nmc" equal to three or four and "edf" equal to about 12 for initial approximations. Our software currently generates up to 66 basis functions, giving a full tenth-order polynomial basis. The number of basis functions for type-1 or -4 cells could be increased, for appropriate Gaussian quadrature is available.

We have had no difficulty meeting the requirement that the rows of $\mathbf{M}$ be independent, using the bases and the weight functions described above. If the rows are dependent, the program informs the user; the parameter "edf" must then be increased. Such increases could be done without user intervention. In such a case, the new matrix is formed by expanding the original matrix; the entries in the entire matrix do not have to be recomputed.

We use $L_2$-orthogonal bases for parallelograms and triangles in an attempt to

provide good structure for the matrix. For parallelograms, if $A_{ij}$ and $A_0$ are constant, there are many zeros off the diagonal in the symmetric blocks comprising $\mathbf{C}$. Our experiments have been concerned with testing the implementation, using relatively small problems defined on domains with increasingly complex boundaries; the solution of the linear system has given no difficulties. We have not made any general survey of the conditioning of the system matrix. Large problems have been solved using the cell discretization method; for example, Greenstadt [14] has obtained approximate solutions to the diffusion equations of nuclear reactor theory. The model required as many as 3,800 equations to estimate solutions for three-dimensional problems. The system was solved using the generalized conjugate gradient method of Concus, Golub, and O'Leary [7]. The results, when compared with solutions generated by other methods, are quite promising.

Although §3 shows that the system of equations has a unique solution for *any* choice of basis functions and weight functions (if $[m]$ is sufficiently large), we expect that bases and weight functions for two- and three-dimensional problems should be chosen to be particularly "compatible", so that the matrix is well structured and convergence of approximations to the solution is rapid. For example, we might expect that if trigonometric functions are used for a basis on a cell, then trigonometric functions would also be appropriate for the moment collocation weight function. We refer to [14], where approximations to the lowest eigenvalue of Laplace's equation on the unit square are obtained using both polynomial and trigonometric bases.

Once the system of linear equations is solved, we can generate the values of the approximation $u_{n,m}$ at any point in the domain. For each interface, the $L_2(\Gamma_{ij})$ norm of the difference of the traces

$$\gamma_{ij}(u_{n,m} \text{ on cell } i) \quad \text{and} \quad \gamma_{ji}(u_{n,m} \text{ on cell } j)$$

on $\Gamma_{ij}$ is computed. If any of these are unacceptably high, the user can either increase "nmc" or return to the preprocessing stage to further subdivide the domain. Such subdivision of parallelograms or triangles could be done without user intervention.

We discuss approximate solutions for two simple problems and compare the sizes of the linear systems used for our cell discretization method and the "Hermite collocation" finite element method provided by our 1985 version of "ELLPACK" [19]. The collocation method obtains Hermite bicubic piecewise polynomial approximations to a solution on rectangular domains. Both programs are run on a Sequent Symmetry machine.

The first example, problem "C" in [19], is a Helmholtz problem with a boundary layer. We approximate the solution to

$$\Delta u - 100u = 150\cosh(20y)/\cosh(20).$$

The domain is the unit square, with Dirichlet boundary data agreeing with the solution

$$u(x, y) = \cosh(10x)/(2\cosh(10)) + \cosh(20y)/(2\cosh(20)).$$

Figure 2 portrays a sketch of the solution. The maximum is 1.

For our cell discretization approximation (CDA) to the solution, we use the entire unit square as our cell. In the table shown in Figure 3, we indicate the

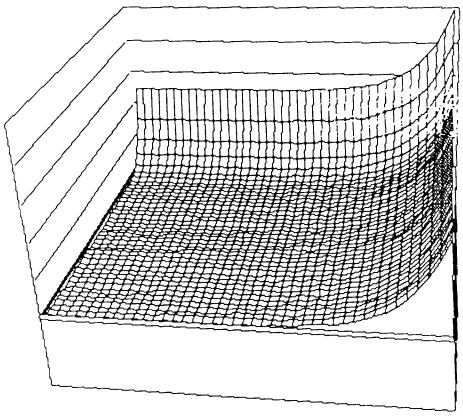

FIGURE 2. Solution of the first example

| CELL DISCRETIZATION METHOD | | | | ERRORS | | ELLPACK'S BICUBIC F.E.M. | | ERRORS | |
|---|---|---|---|---|---|---|---|---|---|
| Order | nbf | nmc | nle | max | L2 | grid | nle | max | L2 |
| 5 | 21 | 3 | 33 | .049 | .021 | 6×6 | 144 | .045 | .010 |
| 6 | 28 | 3 | 40 | .023 | .010 | 7×7 | 194 | .029 | .007 |
| 7 | 36 | 3 | 48 | .010 | .005 | 10×10 | 400 | .010 | .003 |
| 8 | 45 | 4 | 61 | .004 | .002 | 13×13 | 676 | .004 | .0008 |
| 9 | 55 | 3 | 67 | .0018 | .0017 | 17×17 | 1156 | .0016 | .0004 |

FIGURE 3. Comparison of results for the first example

order of the basis used in the CDA approximation, the number of basis functions (nbf), the number of moment collocations enforced (nmc), the number of linear equations used in the discretization (nle), and the maximum of the error (max) and the "$L_2$" estimate of the error (L2). These are compared with solutions generated by ELLPACK's finite element collocation discretization module using grid sizes that produce similar maximum error. We list the size of the grid, the number of linear equations (nle), and the maximum error and the "$L_2$" error. The errors are all estimated using ELLPACK's method, based on the points in a $41 \times 41$ grid.

Figure 4 provides a log-log plot of the errors versus the number of equations used in both approximations.

For our second example, we construct approximations to the homogeneous Dirichlet problem for Poisson's equation $\Delta u = f$, with the formula for $f$ obtained from the desired solution $u(x, y) = -e^y \sin(\pi x)\sin(\pi y)$. The domain is a square of side 2. A sketch of the solution is shown in Figure 5. The maximum is about 4.7.

The number of Gauss points used in our implementation of the cell discretization method limits us to at most a 10th-order basis for Poisson's equation. The maximum error when we use this basis and just one cell is 0.022; some results are shown in Figure 6.

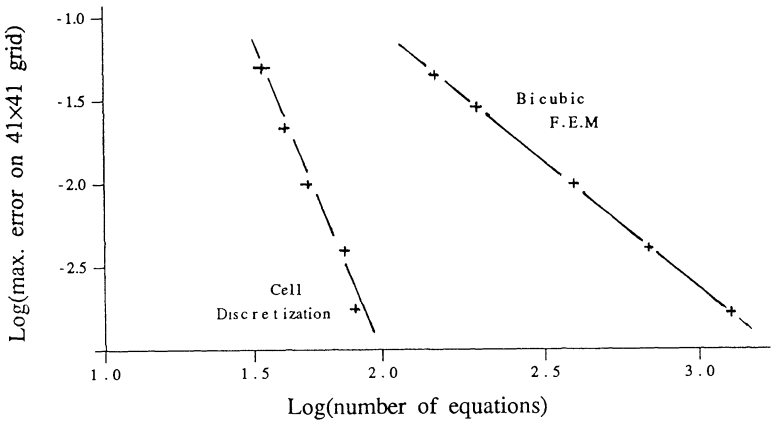


FIGURE 4. Comparison of the max. error vs. number of equations for Example 1

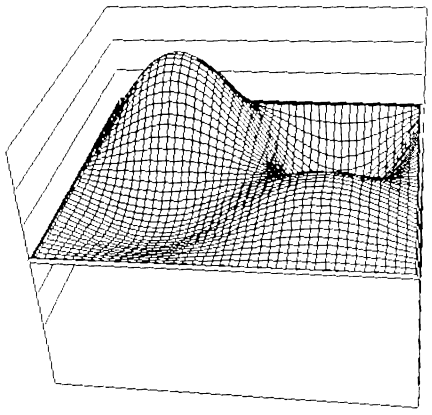

FIGURE 5. Solution of the second example

| CELL DISCRETIZATION METHOD | | | | | | ELLPACK'S BICUBIC F.E.M. | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | ERRORS | | | | ERRORS | |
| Order | nbf | nmc | nle | max | L2 | grid | nle | max | L2 |
| 9 | 55 | 6 | 79 | .050 | .018 | 6×6 | 100 | .052 | .010 |
| 10 | 66 | 6 | 90 | .022 | .009 | 7×7 | 144 | .026 | .005 |

FIGURE 6. Results for the second example using one cell

To obtain greater accuracy using the cell discretization method, we subdivide the domain into four square cells. Sample results are shown in Figure 7.

Figure 8 provides a log-log plot of the errors versus the number of equations used in both approximations; both the one-cell case and the four-cell case are shown.

| CELL DISCRETIZATION METHOD | | | | ERRORS | | ELLPACK'S BICUBIC F.E.M. | | ERRORS | |
|---|---|---|---|---|---|---|---|---|---|
| Order | nbf | nmc | nle | max | L2 | grid | nle | max | L2 |
| 6 | 28 | 4 | 160 | .056 | .009 | 6×6 | 100 | .052 | .010 |
| 7 | 36 | 5 | 204 | .0058 | .0017 | 10×10 | 324 | .0052 | .0010 |
| 8 | 45 | 6 | 252 | .0013 | .0004 | 13×13 | 576 | .0015 | .0003 |
| 9 | 55 | 7 | 304 | .00012 | .00003 | 24×24 | 2116 | .00012 | .00002 |

FIGURE 7. Results for the second example using four cells

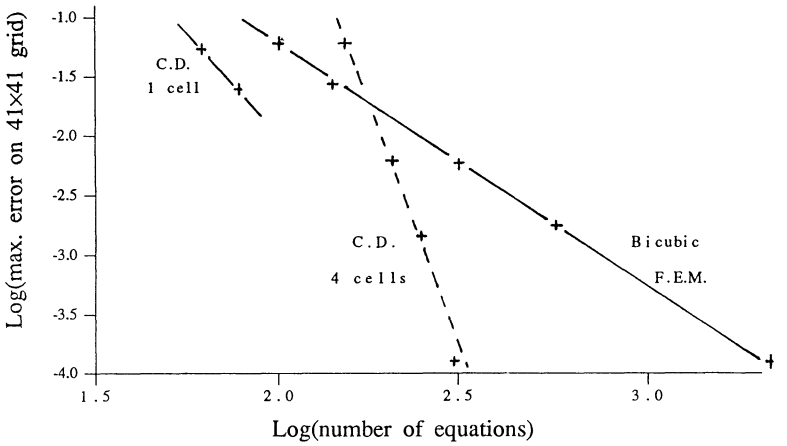


FIGURE 8. Comparison of the max. error vs. number of equations for Example 2

In these tests there is a strong linear relation between the logarithms of the errors and the number of equations for both methods ($|r| > 0.99$).

Note that in both examples, although the maximum errors are similar in the cases cited above, the $L_2$ errors for the bicubic finite element collocation method are less than the $L_2$ errors for the cell discretization method in each case.

The two examples suggest that the number of equations necessary to obtain good accuracy in many applications is similar for both methods. If higher accuracy is desired, it appears that fewer equations are required for the cell discretization method than for the collocation finite element method. The matrix representing the linear system for the Hermite collocation method is banded. The structure of the symmetric matrix generated by the cell discretization algorithm is described in §3. The band width of the array **C** of diagonal blocks is about the same size as the band width of the collocation method for approximations producing similar maximum error.

The second example illustrates the two ways that can be used to increase the accuracy of approximations using the cell discretization algorithm. The most efficient method is to increase the number of basis functions and the number of

moment collocations enforced. The size of the linear system increases relatively slowly and the entries in the matrix are just augmented; the entire matrix does not have to be recomputed to obtain improvement. However, in our implementation, the methods of quadrature limit the number of basis functions that can be used in this procedure, particularly if the $A_{ij}$ and $A_0$ are not constant.

The second method for increasing accuracy is to increase the number of cells. This is more difficult to implement than the first method and may substantially increase the number of linear equations in the discretization if all cells in a first approximation are subdivided. It may be appropriate to first identify the areas of the domain where the errors are largest and just refine the cells in such areas. The program reports the $L_2$-norm of the match-up of approximations across interfaces to help identify areas of the domain where the approximation is poor. An example of this procedure can be found in [16]. Good approximations can then be obtained using fewer basis functions on each cell. In [14], experiments were made to determine the effect of mesh refinement on the accuracy of approximations to the solution in two simple problems. The domain in the first problem in the study in [14] is a square. For a fixed number of basis functions employed on each cell and a fixed number of moment collocations, the error of the approximation on the boundary of the domain was roughly proportional to the 4th power of $h$, the length of the side of square cells of equal size. This proportionality to the 4th power of $h$ did *not* hold in the second problem defined on an $L$-shaped domain (see [14] for an interesting conjecture concerning this result).

We hope that the results cited above will encourage interest in the cell discretization method. It is fairly easy to implement and allows great flexibility in the choice of basis functions and domain decomposition.

Some of the drawbacks and unresolved questions concerning the cell discretization method are the following:

1. The implementation discussed above requires that a user provide a description of the decomposition of the domain into cells of appropriate type. Is there a way to automate an appropriate decomposition, starting, for example, from the description of an irregular domain provided by the ELLPACK definition of a problem?

2. If the coefficients of the equations are not constant, or we have chosen cells with curved boundaries, the required quadrature is time consuming.

3. We have described a general implementation of the method using polynomial bases. This could be extended to accommodate other bases that may be particularly appropriate for certain problems (see [5]).

4. No systematic study of the conditioning of the system matrix for large problems has been made.

5. The method has yet to be implemented to solve general problems in three dimensions.

6. What is the trade-off between the number of moment collocations employed and the total number of basis functions used on any cell? We have as yet only the sorts of empirical results indicated in the examples discussed above.

7. The errors are expressed in terms of projections in $H^1(\Omega_i)$ or $L_2(\Gamma_{ij})$ onto the orthogonal complement of the span of a finite number of chosen basis functions. Methods of approximation theory should enable us to characterize such errors in terms of the size of the cell and the properties of various bases,

e.g., degree of a polynomial approximation or number of trigonometric functions utilized.

8. In [16], Greenstadt extends the method to non-self-adjoint problems, using a primal-dual variational approach as the basis for the discretization process. Two examples of approximations to the solution of convection-diffusion problems are given. The convergence and error results described in this paper have not been extended to the non-self-adjoint case.

## ACKNOWLEDGMENTS

The author is indebted to Dr. John Greenstadt and Dr. Alan Karp for many suggestions concerning the problems addressed in this paper. Professor Leslie Foster proposed the methods we have described for treating the system of linear equations.

## BIBLIOGRAPHY

1. I. Babuška, *The finite element method with Lagrangian multipliers*, Numer. Math. **20** (1973), 179–192.
2. I. Babuška and M. R. Dorr, *Error estimates for the combined h and p versions of the finite element method*, Numer. Math. **37** (1981), 257–277.
3. I. Babuška, B. A. Szabo, and I. N. Katz, *The p-version of the finite element method*, SIAM J. Numer. Anal. **18** (1981), 515–545.
4. J. H. Bramble, *The Lagrange multiplier method for Dirichlet's problem*, Math. Comp. **37** (1981), 1–11.
5. M. W. Coffey, J. Greenstadt, and A. Karp, *The application of cell discretization to a "circle in the square" model problem*, SIAM J. Sci. Statist. Comput. **7** (1986), 917–939.
6. M. W. Coffey, *On the relation of the cell discretization algorithm to the primal hybrid finite element method*, Comm. Appl. Numer. Methods **8** (1992), 109–116.
7. P. Concus, G. H. Golub, and D. P. O'Leary, *A generalized conjugate gradient method for the numerical solution of elliptic partial differential equations*, Sparse Matrix Computations (J. R. Bunch and D. J. Rose, eds.), Academic Press, New York, 1976, pp. 309–332.
8. M. R. Dorr, *The approximation of solutions of elliptic boundary-value problems via the p-version of the finite element method*, SIAM J. Numer. Anal. **23** (1986), 58–76.
9. ____, *On the discretization of interdomain coupling in elliptic boundary-value problems*, Domain Decomposition Methods (T. F. Chan, R. Glowinski, J. Periaux, and O. B. Widlund, eds.), SIAM, Philadelphia, PA, 1989.
10. D. A. Dunavant, *High degree efficient symmetric Gaussian quadrature rules for the triangle*, Internat. J. Numer. Methods Engrg. **21** (1985), 1129–1148.
11. W. Fleming, *Functions of several variables*, 2nd ed., Springer-Verlag, New York, 1977.
12. W. Govaerts and J. D. Pryce, *Block elimination with iterative refinement for bordered linear systems*, Numerical Linear Algebra, Digital Signal Processing and Parallel Algorithms (G. H. Golub and P. VanDooren, eds.), NATO ASI Series, vol. F70, 514-519, Springer-Verlag, Berlin and Heidelberg, 1991.
13. J. Greenstadt, *Cell discretization*, Conference on Applications of Numerical Analysis (J. H. Morris, ed.), Lecture Notes in Math., vol. 228, Springer-Verlag, New York, 1971, pp. 70–82.
14. ____, *The cell discretization algorithm for elliptic partial differential equations*, SIAM J. Sci. Statist. Comput. **3** (1982), 261–288.
15. ____, *The application of cell discretization to nuclear reactor problems*, Nuclear Sci. Engrg. **82** (1982), 78–95.
16. ____, *Cell discretization of nonselfadjoint linear elliptic partial differential equations*, SIAM J. Sci. Statist. Comput. **12** (1991), 1074–1108.

17. P. Grisvard, *Elliptic problems in non-smooth domains*, Pitman, Boston, 1985.
18. P. A. Raviart and J. M. Thomas, *Primal hybrid finite element methods for second order elliptic equations*, Math. Comp. **31** (1977), 391–413.
19. J. Rice and R. Boisvert, *Solving elliptic problems using ELLPACK*, Springer-Verlag, New York, 1985.
20. H. Swann and M. Nishimura, *Implementation of the cell discretization algorithm for solving elliptic partial differential equations*, CAM-16-88, Center for Applied Mathematics and Computer Science, San José, CA, 1988; revision, 1990.

Department of Mathematics and Computer Science, San José State University, San José, California 95192-0103
*E-mail address*: swann@sjsumcs.sjsu.edu