NUMERICAL LINEAR ALGEBRA WITH APPLICATIONS
Numer. Linear Algebra Appl. 2000; **00**:1–6 *Prepared using* nlaauth.cls *[Version: 2002/09/18 v1.02]*

Further Results on Error Estimators for Local Refinement with First-Order System Least Squares (FOSLS)

Thomas Manteuffel, Steven McCormick, Joshua Nolting[1†], John Ruge, and Geoff Sanders

[1] *Department of Applied Mathematics, Campus Box 526, University of Colorado at Boulder, Boulder, CO 80302, USA*

SUMMARY

Adaptive local refinement can substantially improve the performance of simulations that involve numerical solution of partial differential equations. In fact, local refinement capabilities are one of the attributes of First-Order System Least Squares (FOSLS) in that it provides an inexpensive but effective a-posteriori local error bound that accurately identifies regions that require further refinement. Previous theory on FOSLS established the effectiveness of its local error estimator, but only under the assumption that the local region is not too "thin". This paper extends this theory to the case of a rectangular domain by showing that the estimator's effectiveness holds even for certain "thin" local regions. Further, we prove that when the approximation satisfies a local *saturation* property, convergence of an adaptive local refinement scheme is guaranteed. Copyright © 2000 John Wiley & Sons, Ltd.

KEY WORDS: convergence, finite element, FOSLS, adaptive local refinement, harmonic, functional

1. Introduction

First-Order System Least Squares (FOSLS) is a methodology that can be used as a basis for optimal discretization and solution of many Partial Differential Equations (PDEs). One of the important benefits of FOSLS is the natural effective a-posteriori local error indicator provided by its local functional values. The theory in [1] shows that this estimator is reliable and effective for a class of PDEs, provided the local region under consideration is not too "thin" in the sense that the ratio between the interior diameter and exterior diameters is not small. We lift this restriction in this paper for the case of a rectangular domain in two dimensions and a special local region.

To be more specific, consider the unit square depicted in Figure 1, with the potential refinement region indicated by the "thin" vertical strip, Ω^-, on its left side. Its internal boundary is denoted by Γ and the complementary right strip is denoted by Ω^+. Consider

†josh.nolting@colorado.edu

Received
Revised

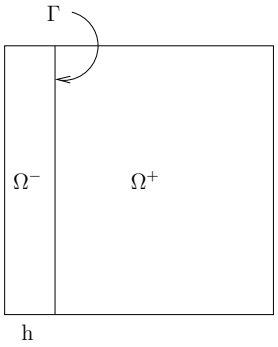

Figure 1. Unit square partitioned into two sub-domains, a thin region, $\Omega^- = (0, h) \times (0, 1)$, and the remaining domain, $\Omega^+ = (h, 1) \times (0, 1)$. The finite element mesh-size is given by h. The line segment Γ is interface between Ω^- and Ω^+. $x = h$.

an abstract quadratic FOSLS functional, $G(\mathbf{u})$, with the property that its minimum value is precisely zero and that it is attained by a unique minimizer (which, presumably, is also a solution to a first-order system of PDEs). Assume that $G(\mathbf{u})$ can be written in terms of its "element" contributions from Ω^- and Ω^+ as follows:

$$G(\mathbf{u}) = G^-(\mathbf{u}) + G^+(\mathbf{u}). \tag{1}$$

Our aim is to show that if $G^-(\mathbf{u})$ is large compared to $G^+(\mathbf{u})$ for some approximation, $\mathbf{u}$, to the minimizer, then refinement in Ω^- alone is enough to reduce $G(\mathbf{u})$ by a substantial amount, no matter how thin Ω^- is. The key here is in the treatment of the error in $\mathbf{u}$. If it consists primarily of components that are local to Ω^- and Ω^+ (i. e., that are zero on Γ), then this conclusion is easy to prove. The difficulty occurs when the error is predominantly "functional-harmonic" ("$\mathcal{F}$-harmonic"), by which we mean that it is possibly nonzero on Γ but otherwise minimizes $G(\mathbf{u})$ in the sense that no other $\mathbf{u}$ with the same trace on Γ yields a smaller value of $G(\mathbf{u})$. Error of this type with a trace that is oscillatory with respect to the scale of the elements in Ω^+ may not be substantially reduced by even infinite refinement only in Ω^-. Thus, our principal task is to show that, for such error, $G^-(\mathbf{u})$ cannot be substantially larger than $G^+(\mathbf{u})$. This would imply that the refinement decision-making would not be distracted by a relatively large value of the estimator in Ω^- that is due to irreducible error.

We establish this result from three different points of view. We begin with a specific proof of our assertions in the continuum setting. We follow this with an abstract algebraic proof that is then applied to the unit square case specified above. We then provide numerical evidence of our assertions by reporting on numerical tests with the ACE (Accuracy per Computational cost Efficiency) refinement algorithm developed in [6]. Finally, with the error estimator proofs in hand, we extend the theory in [1] by allowing arbitrarily thin refinement regions. That is, on a mesh with mesh size h, the refinement region may have inner diameter h.

The following is restricted to the unit square case introduced above. It should be clear, however, that the results hold for a general rectangle. Moreover, while the continuum theory is specific to the refinement structure specified above, the abstract algebraic theory is considerably more general. It should be fairly easy to confirm its applicability to more complicated refinement regions and more general domains, although such an extension is

beyond the current scope presented here.

With the aforementioned theory proved we end by showing a convergence proof for adaptive local refinement (ALR) schemes based on FOSLS finite elements. If a local *saturation* condition is satisfied, these schemes are guaranteed to converge. Although much of the paper is dedicated to proving that the "thinness" constraint is unnecessary, the convergence proof is critical to the viability of ALR schemes that use the local FOSLS functional as the error indicator.

This paper is organized as follows. We provide a brief introduction to FOSLS and its notation in the next section. The continuum proof is given in Section 3. In Section 4, we develop an abstract algebraic proof that is then applied to the unit square example specified above. Section 5 contains numerical results and Section 6 provides a proof of convergence of a FOSLS adaptive refinement strategy. We conclude with some remarks in Section 7.

2. FOSLS Background

A brief description of the FOSLS finite element method is introduced in this section. For a more complete description, see [3]. As the acronym FOSLS suggests, a **f**irst-**o**rder **s**ystem of equations is solved using **l**east **s**quares techniques. Throughout this paper, PDEs that have second-order derivatives are considered, so they require conversion into a first-order system. Since least-squares techniques result in a bilinear form in which the operators in the system of equations are squared, one benefit of having a first-order system is that the resulting bilinear form in the variational formulation is, at most, second-order. The condition number of the associated algebraic system is then order $O(h^{-2})$, where h is the grid mesh-size.

Assume that the PDE is of the form $L\mathbf{v} = \mathbf{g}$, where L is a system of differential operators, $\mathbf{v}$ is the continuous solution of n variables, and $\mathbf{b}$ is the given right hand side. Consider, for example, the following scalar two-dimensional PDE:

$$\Delta v = g \qquad \text{in } \Omega, \tag{2}$$

$$v = 0 \qquad \text{on } \partial\Omega, \tag{3}$$

where $v = v(x, y)$, $g = g(x, y)$, and $\partial\Omega$ is the boundary of Ω. The above second-order system is transformed into the following equivalent first-order system of equations:

$$\mathbf{U} - \nabla v = \mathbf{0} \qquad \text{in } \Omega, \tag{4}$$

$$\nabla \cdot \mathbf{U} = g \qquad \text{in } \Omega, \tag{5}$$

$$\nabla \times \mathbf{U} = 0 \qquad \text{in } \Omega, \tag{6}$$

$$v = 0 \qquad \text{on } \partial\Omega. \tag{7}$$

Let,

$$\mathbf{u} = \begin{bmatrix} v \\ \mathbf{U} \end{bmatrix}. \tag{8}$$

In general, the original PDE, $L\mathbf{v} = \mathbf{g}$, is transformed into the PDE $\mathcal{L}\mathbf{u} = \mathbf{f}$, where $\mathcal{L}$ is the first-order system of differential operators and $\mathbf{f}$ is the given right-hand side, such that $L\mathbf{v} = \mathbf{g}$ is equivalent to $\mathcal{L}\mathbf{u} = f$. Because L contains second-order or higher derivatives, m new variables are defined in $\mathbf{u}$, and $\mathcal{L}\mathbf{u} = \mathbf{f}$ will contain more equations. In the above example, a

new variable $\mathbf{U}$ was introduced. Since $\mathbf{U}$ was introduced as a gradient, its curl is equal to zero, and we often add this consistent equation. Letting $\mathcal{L}_i$ be the ith equation of the first-order system $\mathcal{L}$, then we define

$$\mathcal{G}(\mathbf{u};\mathbf{f}) = \sum_{i=1}^{M} \|w_i(\mathcal{L}_i\mathbf{u} - f_i)\|^2_{(0,\Omega)}, \tag{9}$$

where w_i is an appropriate weighting and M is the number of equations in the first-order system. Typically, $w_i = 1$. However, w_i is kept general to allow un-weighting near singularities. Although other squared norms could be applied in the FOSLS functional, this paper only considers the sum of L^2 norms squared. The minimization problem is then written as

$$\mathbf{u} = \arg\min_{\mathbf{v}\in W} \mathcal{G}(\mathbf{v};\mathbf{f}), \tag{10}$$

where W is a product of H^1 spaces. Since $\mathbf{u}$ is a minimizer of the quadratic functional $\mathcal{G}(\mathbf{v};\mathbf{f})$, the variational formulation is given as follows: find $\mathbf{u} \in W$, such that

$$\mathcal{F}(\mathbf{u};\mathbf{v}) = \sum_{i=1}^{M} \langle f_i, \mathcal{L}_i\mathbf{v}\rangle_{w_i}, \qquad \text{for all } \mathbf{v} \text{ in } W, \tag{11}$$

where $< \cdot,\cdot >_{w_i}$ represents a weighted inner product, and

$$\mathcal{F}(\mathbf{u};\mathbf{v}) := \sum_{i=1}^{M} \langle \mathcal{L}_i\mathbf{u}, \mathcal{L}_i\mathbf{v}\rangle_{w_i}. \tag{12}$$

Existence and uniqueness of $\mathbf{u}$ is attained if the bilinear form (12) is continuous and coercive in a weighted H^1 norm:

$$\mathcal{F}(\mathbf{u};\mathbf{v}) \le C_c\|\mathbf{u}\|_{(w,1,\Omega)}\|\mathbf{v}\|_{(w,1,\Omega)} \tag{13}$$

and

$$\mathcal{F}(\mathbf{u};\mathbf{u}) \ge C_e\|\mathbf{u}\|^2_{(w,1,\Omega)}, \tag{14}$$

respectively.

Definition 1. *If $\mathcal{F}(\mathbf{u};\mathbf{u})$ is coercive and continuous then the FOSLS formulation is H^1-elliptic.*

In what follows, we address the Poisson equation in domains without singularities as our model problem. Thus, we set $w_i = 1$ and, with inclusion of the curl constant, assume continuity and coercivity in the standard H^1 norm.

3. Continuum Theory

The goal is to show that after solving the discrete problem on a given finite element space, $\mathcal{V}^h$, there exists no error with large element contributions in one part of the domain, Ω^-, relative to that of its complement, Ω^+, that is not amenable to significant reduction by refinement in

Ω^-. Otherwise, the refinement algorithm could possibly concentrate resources strictly in Ω^-, offering no improvement in the accuracy of the approximation. For the continuous proof, we assume that the global problem has been solved on a uniform grid with mesh-size h. Then, without loss of generality, we restrict our attention to error that is orthogonal in the $\mathcal{F}$-inner product, which is denoted below in (15). The $\mathcal{F}$-inner product is often considered an energy inner product. We further assume that $\mathcal{F}(u;v)$ is the scalar bilinear form

$$\mathcal{F}(u;v) = \langle \nabla u, \nabla v \rangle . \tag{15}$$

We believe that the following development can be extended to more general problems in which the bilinear form, $\mathcal{F}(u;v)$, corresponds to an H^1-elliptic functional for systems of equations. However, that extension is beyond the scope of this paper. Thus, our assumption on the error becomes

$$\langle \nabla E, \nabla v_h \rangle_\Omega = 0 \quad \text{for all } v_h \text{ in } \mathcal{V}^h, \tag{16}$$

where $E = v_h - v$. For the following proofs, we decompose the error of our FOSLS discretization into four components. Keep in mind that the FOSLS minimization problem yields

$$u = \arg\min_{v \in \mathcal{V}} \mathcal{G}(v;f), \tag{17}$$

where $\mathcal{G}(v;f)$ was defined in the FOSLS background, Section 2. Note that $\mathcal{G}(v_h;f) = \mathcal{G}(E;0)$. An $\mathcal{F}$-orthogonal decomposition of E into $\mathcal{F}$-*harmonic* and $\mathcal{F}$-*local* components is given by

$$E(x,y) = \mathcal{H} + \eta.$$

To formalize $\mathcal{H}$ and η, consider $E^- := E(x,y)$ restricted to the region Ω^-, and let $\mathcal{H}^-$ and η^- be the associated error components. We define a set of local functions having support in Ω^-.

$$\mathcal{V}^- := \{u \in \mathcal{V} : u = 0 \text{ on } \Omega^+ \cup \Gamma\}.$$

Then,

$$\eta^- := \arg\min_{u \in \mathcal{V}^-} \mathcal{G}(E-u;0)$$

and

$$\mathcal{H}^- := \arg\min_{u=E \text{ on } \Omega^+ \cup \Gamma} \mathcal{G}_{\Omega^-}(u;0).$$

For the remainder of the paper we consider $\mathcal{F}$-*harmonic* and $\mathcal{F}$-*local* error to be simply *harmonic* and *local* error. Note, infinite refinement of Ω^- does not eliminate the harmonic error component, $\mathcal{H}^-$, and completely eliminates η^-. There is an analogous definition for the error components on Ω^+. Let E^+, $\mathcal{H}^+$ and η^+ represent the error restricted to the region Ω^+. The harmonic error functions also satisfy Laplace's equation, $\Delta\mathcal{H}^+ = \Delta\mathcal{H}^- = 0$, in the interior of their respective domains and $\mathcal{H}^+ = \mathcal{H}^- = E$ on the interface, Γ, and domain boundary $\partial\Omega$. This yields the following relationships:

$$\left\langle \nabla E^-, \nabla E^- \right\rangle_{\Omega^-} = \left\langle \nabla \mathcal{H}^-, \nabla \mathcal{H}^- \right\rangle_{\Omega^-} + \left\langle \nabla \eta^-, \nabla \eta^- \right\rangle_{\Omega^-} \tag{18}$$

$$\left\langle \nabla E^+, \nabla E^+ \right\rangle_{\Omega^+} = \left\langle \nabla \mathcal{H}^+, \nabla \mathcal{H}^+ \right\rangle_{\Omega^+} + \left\langle \nabla \eta^+, \nabla \eta^+ \right\rangle_{\Omega^+} \tag{19}$$

$$\langle \nabla E, \nabla E \rangle_\Omega = \left\langle \nabla E^-, \nabla E^- \right\rangle_{\Omega^-} + \left\langle \nabla E^+, \nabla E^+ \right\rangle_{\Omega^+} . \tag{20}$$

Also, applying Green's theorem yields

$$\left\langle \nabla \mathcal{H}^-, \nabla \eta^- \right\rangle_{\Omega^-} = 0, \qquad \left\langle \nabla \mathcal{H}^+, \nabla \eta^+ \right\rangle_{\Omega^+} = 0. \tag{21}$$

The overall proof has many small interesting aspects, so it will be broken-down into subsections each describing a specific facet. We first formalize the the problem setup for the proof.

Hypothesis 1.

$$\left\langle \nabla E^+, \nabla E^+ \right\rangle_{\Omega^+} \ll \left\langle \nabla E^-, \nabla E^- \right\rangle_{\Omega^-} \tag{22}$$

and

$$\left\langle \nabla \eta^-, \nabla \eta^- \right\rangle_{\Omega^-} \ll \left\langle \nabla E^-, \nabla E^- \right\rangle_{\Omega^-} \tag{23}$$

are mutually possible.

Notice that (19) and (22) implies

$$\left\langle \nabla \eta^+, \nabla \eta^+ \right\rangle_{\Omega^+} \leq \left\langle \nabla \eta^+, \nabla \eta^+ \right\rangle_{\Omega^+} + \left\langle \nabla \mathcal{H}^+, \nabla \mathcal{H}^+ \right\rangle_{\Omega^+} = \left\langle \nabla E^+, \nabla E^+ \right\rangle_{\Omega^+} \tag{24}$$

$$\ll \left\langle \nabla E^-, \nabla E^- \right\rangle_{\Omega^-}. \tag{25}$$

The goal of the remainder of this section is to show that if u_h is the exact solution on grid $\mathcal{V}^h$, then Hypothesis 1 is contradictory. In the first subsection, we develop theory with an assumption that the *local* error is exactly equal to zero everywhere. Because we cannot guarantee both $\mathcal{F}$-orthogonality and purely *harmonic* error, we then extend this theory to a more general situation. The proofs for some of the theorems and lemmas are shortened to avoid distraction and maintain a concise paper. For all complete proofs, see [6].

3.1. Case I: Purely Harmonic Error

Let $e(x, y) \in \Omega$ be a function that is defined by extending an *oscillatory* function, $t \in H^{\frac{1}{2}}(\Gamma)$, harmonically into Ω^- and Ω^+. Let $\partial\Omega$ be a rectangular closed boundary of Ω. Assume $e(x, y) = 0$ for $(x, y) \in \partial\Omega$. Then, t is the trace of e on Γ (see Fig. 1). Throughout this section, the set of *oscillatory* functions is given by the following definition:

$$\mathcal{O} := \{t \in H^{\frac{1}{2}}(\Gamma) : \|t\|_{(0,\Gamma)} \leq C_1 h^{\frac{1}{2}} \|t\|_{(\frac{1}{2},\Gamma)}\}, \tag{26}$$

where C_1 is a constant independent of h and t. For conciseness later in this section, we also define a set to capture the properties given for $e(x, y)$.

$$\mathcal{S} := \{e \in C^0(\Omega), t \in \mathcal{O} : \tag{27}$$

$$e \text{ is a } \textit{harmonic} \text{ extension of } t \text{ into } \Omega^+ \text{ and } \Omega^-\}. \tag{28}$$

Since $t \in H^{\frac{1}{2}}(\Gamma)$ and $t(y = 0) = 0$ and $t(y = 1) = 0$, t has the following sine series expansion:

$$t = \sum_{1}^{\infty} a_k \sin(k\pi y). \tag{29}$$

Then define $e^+(x,y)$ (e in Ω^+) and $e^-(x,y)$ (e in Ω^-) as the harmonic extension of the error with the following Fourier expansions:

$$e^-(x,y) = \sum_1^\infty a_k \sin(k\pi y)\frac{\sinh(k\pi x)}{\sinh(k\pi h)} \tag{30}$$

$$e^+(x,y) = \sum_1^\infty a_k \sin(k\pi y)\frac{\sinh(k\pi(1-x))}{\sinh(k\pi(1-h))}. \tag{31}$$

The denominators ($\sinh(\cdot)$) in each series force continuity across Γ. We first show that if the trace of e is oscillatory, then $\|\nabla e^+\|_{(0,\Omega^+)}$ cannot be small compared to $\|\nabla e^-\|_{(0,\Omega^-)}$.

Theorem 1. *Let $e \in \mathcal{S}$ be decomposed into e^+ and e^-, representing the extension of $t \in \mathcal{O}$ into Ω^+ and Ω^- respectively, then there exists a constant $C_2 > 0$, independent of h and t, such that*

$$\|\nabla e^-\|_{(0,\Omega^-)} \le C_2 \|\nabla e^+\|_{(0,\Omega^+)}.$$

Proof Define a function $r \in \Omega^r := (h,2h)\times(0,1)$, as the reflection of e^- about Γ. Then r is *harmonic* in Ω^r. Let $\phi(x) = (2 - \frac{x}{h})$ for $x \in (h,2h)$ be a cutoff function with support only

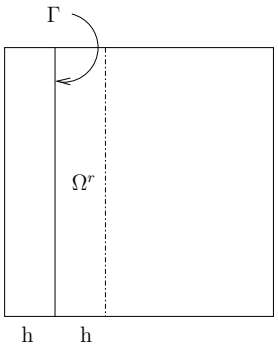

Figure 2. $\Omega^r = (h,2h)\times(0,1)$. The function r is defined in $\Omega^r \cup \Gamma$.

in $(h,2h)$. Multiplying the cutoff function with e^+, and noting that r is harmonic in Ω^r and $\phi e^+ = r$ on $\partial\Omega^r \setminus \partial\Omega^+$, results in the following inequality:

$$\|\nabla e^-\|_{(0,\Omega^-)} = \|\nabla r\|_{(0,\Omega^r)} \le \|\nabla(\phi e^+)\|_{(0,\Omega^r)}. \tag{32}$$

Using (32) and the triangle inequality we get a bound on the energy (L^2-norm of the gradient) of e^-.

$$\begin{aligned}\|\nabla e^-\|_{(0,\Omega^-)} &\le \|e^+\nabla\phi + (\nabla e^+)\phi\|_{(0,\Omega^r)} \\ &\le \|e^+\nabla\phi\|_{(0,\Omega^r)} + \|\phi\nabla e^+\|_{(0,\Omega^r)} \\ &\le \frac{1}{h}\|e^+\|_{(0,\Omega^r)} + \|\nabla e^+\|_{(0,\Omega^r)}. \end{aligned}\tag{33}$$

Therefore, the theorem is proved if $\frac{1}{h}\|e^+\|_{(0,\Omega^r)} \leq C_4\|\nabla e^+\|_{(0,\Omega^+)}$ for some constant, C_4, independent of h and t. Many steps are needed to validate the above setting for our scenario. This is the objective for the rest of this proof.

Since e^+ and t are represented by the Fourier expansions given in (29) and (30), respectively, solving for the square of the L^2 norm of e^+ in Ω^r yields

$$\|e^+\|^2_{(0,\Omega^r)} \leq h\|t\|^2_{(0,\Gamma)}. \tag{34}$$

From trace theory given in, for example, [7], there exist a constant C_3 such that

$$\|t\|_{(\frac{1}{2},\Gamma)} \leq C_3\|e^+\|_{(1,\Omega^+)} = C_3(\|e^+\|^2_{(0,\Omega^+)} + \|\nabla e^+\|^2_{(0,\Omega^+)})^{\frac{1}{2}}.$$

Therefore, using the Poincaré inequality (c.f. [5]) yields

$$\begin{aligned}\|t\|_{(\frac{1}{2},\Gamma)} &\leq C_3(c_1^2\|\nabla e^+\|^2_{(0,\Omega^+)} + \|\nabla e^+\|^2_{(0,\Omega^+)})^{\frac{1}{2}},\\ &\leq c_2\|\nabla e^+\|_{(0,\Omega^+)},\end{aligned} \tag{35}$$

where $c_2 = C_3(c_1^2+1)^{\frac{1}{2}}$. Combining (34), (35), and the fact that t is *oscillatory* yields

$$\|e^+\|_{(0,\Omega^r)} \leq h^{\frac{1}{2}}\|t\|_{(0,\Gamma)} \leq C_1 h\|t\|_{(\frac{1}{2},\Gamma)} \leq c_4 h\|\nabla e^+\|_{(0,\Omega^+)}, \tag{36}$$

where $c_3 = C_1c_2$. The proof is completed by combining (33) and (36) to yield

$$\|\nabla e^-\|_{(0,\Omega^-)} \leq c_4\|\nabla e^+\|_{(0,\Omega^+)} + \|\nabla e^+\|_{(0,\Omega^r)} \leq C_2\|\nabla e^+\|_{(0,\Omega^+)}, \tag{37}$$

with $C_2 = 1 + c_3$. □

In order to use Theorem 1, we prove that if the error is $\mathcal{F}$-orthogonal to the grid of mesh-size h and purely *harmonic*, then the *trace*$(E) = t$ along Γ must be oscillatory enough to satisfy

$$\|t\|_{(0,\Gamma)} \leq C_1 h^{\frac{1}{2}}\|t\|_{(\frac{1}{2},\Gamma)}. \tag{38}$$

In order to prove this statement, we first cover some preliminaries. After discussing the preliminary concepts, a theorem addressing the above oscillatory assumption is developed.

Recall the sine series expansion, $t = \sum_1^\infty a_k \sin(k\pi y)$. Define $\mathcal{H}^+(x,y)$ and $\mathcal{H}^-(x,y)$ as the harmonic part of the error with Fourier expansions identical to (30) and (31), respectively. For t to satisfy (38), $\sum_{k=1}^\infty a_k^2$ must be less than or equal to $C_1 h\sum_{k=1}^\infty a_k^2 k$. Since E is $\mathcal{F}$-orthogonal to the space associated with the grid,

$$\left\langle \nabla E^+, \nabla v_h\right\rangle_{\Omega^+} + \left\langle \nabla E^-, \nabla v_h\right\rangle_{\Omega^-} = 0. \tag{39}$$

Separating E into *harmonic* and *local* components yields

$$\begin{aligned}\left\langle \nabla E^+, \nabla v_h\right\rangle_{\Omega^+} + \left\langle \nabla E^-, \nabla v_h\right\rangle_{\Omega^-} &\\ = \left\langle \nabla \mathcal{H}^+, \nabla v_h\right\rangle_{\Omega^+} + \left\langle \nabla \mathcal{H}^-, \nabla v_h\right\rangle_{\Omega^-} &\\ + \left\langle \nabla \eta^+, \nabla v_h\right\rangle_{\Omega^+} + \left\langle \nabla \eta^-, \nabla v_h\right\rangle_{\Omega^-} = 0,&\end{aligned} \tag{40}$$

for any $v_h \in \mathcal{V}^h$. Using Green's first identity [4] and recalling that $\Delta\mathcal{H}^+ = \Delta\mathcal{H}^- = 0$ on the interiors of Ω^+ and Ω^-, respectively, we get

$$\int_{\partial\Omega^+}(n^+\cdot\nabla\mathcal{H}^+)v_h ds + \int_{\partial\Omega^-}(n^-\cdot\nabla\mathcal{H}^-)v_h ds = -\left\langle \nabla\eta, \nabla v_h\right\rangle_\Omega, \tag{41}$$

where n^+ and n^- are the outward unit normal vectors and $\partial\Omega^+$ and $\partial\Omega^-$ are the boundaries of Ω^+ and Ω^-, respectively. The boundary conditions on Ω are homogeneous Dirichlet, so v_h is zero along $\partial\Omega$. We are then left with a jump in the normal derivative across Γ. To be precise, define

$$f(y) := \left[\!\left[\left(\frac{\partial}{\partial x}\mathcal{H} \right) |_{x=h} \right]\!\right] = \left(\lim_{\delta x \to 0^+} \frac{\mathcal{H}^-(h-\delta x, y) - \mathcal{H}^-(h,y)}{-\delta x} \right) - \left(\lim_{\delta x \to 0^+} \frac{\mathcal{H}^+(h+\delta x, y) - \mathcal{H}^+(h,y)}{\delta x} \right), \quad (42)$$

to be the jump in the x-derivative of $\mathcal{H}$ across Γ. Equality (41) can be written as follows:

$$\int_\Gamma \left[\!\left[\frac{\partial}{\partial x}\mathcal{H}|_{x=h} \right]\!\right] v_h dy = -\langle \nabla\eta, \nabla v_h \rangle_\Gamma \qquad \text{for all } v_h \in \mathcal{V}^h. \quad (43)$$

Since $f(0) = f(1) = 0$, the Fourier expansion for this jump function is given by

$$f(y) = \left[\!\left[\left(\frac{\partial}{\partial x}\mathcal{H} \right) |_{x=h} \right]\!\right] = \sum_{k=1}^{\infty} b_k \sin(k\pi y), \quad (44)$$

where

$$b_k = a_k k\pi \delta_k, \quad (45)$$

with

$$\delta_k = \left[\frac{\cosh(k\pi h)}{\sinh(k\pi h)} + \frac{\cosh(k\pi(1-h))}{\sinh(k\pi(1-h))} \right]. \quad (46)$$

For convenience later, let $g_k := k\pi h \delta_k$. Before moving on, let us look at some important characteristics associated with g_k. We present two technical lemmas whose proofs appear in [6].

Lemma 1. *For $1 \le k \le N = \frac{1}{h}$ and $h \le \frac{1}{2}$, then g_k is a monotone non-decreasing sequence, with*

$$1 \le g_k \le 2\pi \frac{e^\pi + e^{-\pi}}{e^\pi - e^{-\pi}} =\approx 6.31.$$

Further, note that $\delta_k \to 2$ as $k \to \infty$ and is $O(1)$ for $k \ge N$.

Proof See [6].

Let I_h be a one-dimensional piecewise interpolation operator on a grid with mesh-size h. Let $I_h \sin(k\pi y) = \sin(k\pi y) + \varepsilon_k(y)$ represent the linear interpolant of the kth sine mode on a grid with mesh-size h. Then, the interpolation error is

$$\varepsilon_k(y) = I_h \sin(k\pi y) - \sin(k\pi y). \quad (47)$$

Assume, again, that $h = \frac{1}{N}$. We give a sine series expansion for $\varepsilon_k(y)$, which is the interpolation error:

$$\varepsilon_k(y) = \sum_{\ell=1}^{\infty} r_{k\ell} \sin(\ell\pi y).$$

Lemma 2. *Let $\varepsilon_k(y) = I_h \sin(k\pi y) - \sin(k\pi y)$. Then, the Fourier expansion coefficients, $r_{k\ell}$, of $\varepsilon_k(y)$ are zero, except for $\ell = k$ and $\ell = 2iN \pm k$, for $i = 1, 2, \cdots$.*

Proof See [6]. For the remainder of this subsection let ε_k denote $\varepsilon_k(y)$.

Corollary 1. *For all $j, k \leq N$ and $j \neq k$,*

$$\langle \varepsilon_k, \varepsilon_j \rangle_\Gamma = \frac{1}{2} \sum_{\ell=1}^{\infty} r_{k\ell} \cdot r_{j\ell} = 0, \qquad \text{for } j \neq k. \tag{48}$$

Proof Corollary 1 is a direct consequence of the zero structure of the coefficients $r_{k\ell}$ given in Lemma 2.

Corollary 2. *For all $j, k \leq N$ and $j \neq k$,*

$$\left\langle \varepsilon_k', \varepsilon_j' \right\rangle_\Gamma = 0. \tag{49}$$

Proof Since

$$\left\langle \varepsilon_k', \varepsilon_j' \right\rangle_\Gamma = \pi^2 \sum_{\ell=1}^{\infty} \ell^2 (r_{k\ell} \cdot r_{j\ell}),$$

the zero structure of the coefficients $r_{k\ell}$ given in Lemma 2 completes the proof. □

A standard interpolation bound, given in [2], yields a bound on the coefficients $r_{k\ell}$ as follows:

$$\|\varepsilon_k\|^2_{(0,\Gamma)} = r_{kk}^2 + \sum_{\ell > N} r_{k\ell}^2 \leq C_4 h^4 k^4. \tag{50}$$

We also have a bound, given in [2], on the semi-norm as follows:

$$|\varepsilon_k|_{(1,\Gamma)} = |I_h \sin(k\pi y) - \sin(k\pi y)|_{(1,\Gamma)} \leq c_1 h \| \sin(k\pi y)\|_{(2,\Gamma)} \leq C_5 h k^2. \tag{51}$$

We take advantage of Lemma 2 by using a special combination of sine interpolants as a test function, v_h. We also want a v_h that gives information on the behavior of the trace. A logical choice is the interpolant of $\sum_{k=1}^{\infty} b_k \sin(k\pi y)$ because the coefficients, b_k, will be included. The interpolant of the partial sum is also in $\mathcal{V}^h$, which is easier to work with for our purpose. Therefore, we let $g(y) := \sum_{k=1}^{N} b_k I_h \sin(k\pi y)$, where I_h is a linear interpolation operator on Γ and let

$$\hat{v}_h = \begin{cases} \frac{x}{h} g(y) & \text{when } x \leq h, \\ (\frac{-x}{h} + 2) g(y) & \text{when } h < x \leq 2h, \\ 0 & \text{otherwise.} \end{cases} \tag{52}$$

Lemma 3. *Let $\hat{v}_h$ be defined in (52). Let $q := -\langle \nabla \eta, \nabla \hat{v}_h \rangle_\Gamma$, from (43). Then,*

$$\sum_{k=1}^{\infty} a_k^2 \leq C_1 h \sum_{k=1}^{\infty} (a_k^2 k) + h^2 |q|,$$

where C_1 is independent of h.

Proof Since $\hat{v}_h \in \mathcal{V}^h$, we have $\langle f(y), g(y)\rangle_\Gamma = q$. Then, from (42),

$$\begin{aligned}\langle f(y), g(y)\rangle_\Gamma &= \left\langle f(y), \sum_{k=1}^{N} b_k I_h \sin(k\pi y)\right\rangle_\Gamma \\ &= \sum_{k=1}^{N}(b_k^2 + r_{kk}b_k^2) + \sum_{k=1}^{N} b_k \sum_{\ell>N} r_{k\ell}b_\ell = -\langle \nabla\eta, \nabla\hat{v}_h\rangle_\Omega . \end{aligned} \tag{53}$$

Rearranging (53) and taking absolute values results in the following inequalities:

$$\sum_{k=1}^{N} b_k^2 \le \sum_{k=1}^{N} |r_{kk}|b_k^2 + \sum_{k=1}^{N} |b_k| \sum_{\ell>N} |r_{k\ell}b_\ell| + |q|, \tag{54}$$

$$= \sum_{k=1}^{N} |r_{kk}|b_k^2 + \sum_{k=1}^{N} |b_k| \sum_{\ell>N} |r_{k\ell}\ell| \left|\frac{b_\ell}{\ell}\right| + |q|. \tag{55}$$

Note, we have multiplied and divided by ℓ in the second term in (55). Now, switching the indices in the double sum of inequality (55) and also replacing $|r_{kk}|$ with bound (50) yields

$$\sum_{k=1}^{N} b_k^2 \le \sum_{k=1}^{N} b_k^2\sqrt{C_4}h^2k^2 + \sum_{\ell>N}\left|\frac{b_\ell}{\ell}\right| \sum_{k=1}^{N} |b_k||r_{k\ell}\ell| + |q|. \tag{56}$$

For $k = 1, \ldots, N$, define

$$\underline{r}_k = [\cdots r_{k\ell} \cdots], \qquad \text{for } \ell > N, \tag{57}$$

$$\underline{\hat{r}}_k = [\cdots |r_{k\ell}\ell| \cdots], \qquad \text{for } \ell > N. \tag{58}$$

From Lemma 2, $r_{k\ell}$ is zero whenever $r_{j\ell}$ is nonzero for $j \neq k$. Then, we have $\left\langle \underline{r}_k, \underline{r}_j \right\rangle_{\ell^2} = 0$ and $\left\langle \underline{\hat{r}}_k, \underline{\hat{r}}_j \right\rangle_{\ell^2} = 0$. We also define

$$\underline{\beta} = \left[\cdots \left|\frac{b_\ell}{\ell}\right| \cdots\right], \qquad \text{for } \ell > N. \tag{59}$$

Then, the second term on the right-hand side of (56) is bounded by

$$\sum_{\ell>N}\left|\frac{b_\ell}{\ell}\right| \sum_{k=1}^{N} |b_k||r_{k\ell}\ell| = \left\langle \underline{\beta}, \sum_{k=1}^{N} |b_k|\underline{\hat{r}}_k \right\rangle_{\ell^2} \tag{60}$$

$$\le \|\underline{\beta}\|_{\ell^2} \left\| \sum_{k=1}^{N} |b_k|\underline{\hat{r}}_k \right\|_{\ell^2} \tag{61}$$

$$= \|\underline{\beta}\|_{\ell^2} \left(\sum_{k=1}^{N} |b_k|^2 \|\underline{\hat{r}}_k\|_{\ell^2}^2 \right)^{\frac{1}{2}} \tag{62}$$

$$= \left(\sum_{\ell>N} \left(\frac{b_\ell}{\ell}\right)^2 \right)^{\frac{1}{2}} \left(\sum_{k=1}^{N} b_k^2 \left(\sum_{\ell>N} r_{k\ell}^2\ell^2 \right) \right)^{\frac{1}{2}} \tag{63}$$

Thus, (56) becomes

$$\sum_{k=1}^{N} b_k^2 \leq \sum_{k=1}^{N} b_k^2 \sqrt{C_4} h^2 k^2 + \left(\sum_{\ell > N} \frac{b_\ell^2}{\ell^2} \right)^{\frac{1}{2}} \left(\sum_{k=1}^{N} b_k^2 \left(\sum_{\ell > N} (r_{k\ell} \ell)^2 \right) \right)^{\frac{1}{2}} + |q|. \tag{64}$$

Applying bound (51) and rearranging, gives way to

$$\begin{aligned} \sum_{k=1}^{N} b_k^2 &\leq \sqrt{C_4} h^2 \sum_{k=1}^{N} b_k^2 k^2 + \left(\sum_{\ell > N} \frac{b_\ell^2}{\ell^2} \right)^{\frac{1}{2}} \left(\sum_{k=1}^{N} b_k^2 |\varepsilon_k|_1^2 \right)^{\frac{1}{2}} + |q|, \\ &\leq \sqrt{C_4} h^2 \sum_{k=1}^{N} b_k^2 k^2 + \left(\sum_{\ell > N} \frac{b_\ell^2}{\ell^2} \right)^{\frac{1}{2}} \left(\sum_{k=1}^{N} b_k^2 C_5^2 h^2 k^4 \right)^{\frac{1}{2}} + |q|. \end{aligned} \tag{65}$$

Now consider the norm of the trace. Let C_6=max$\{k\pi h \delta_k$, for $k \leq N\}$ and let $c_5 = \sqrt{C_4} C_6^2$. Using Lemma 1, (45) and (65), we have

$$\begin{aligned} \sum_{k=1}^{N} a_k^2 = \sum_{k=1}^{N} \left[\frac{b_k h}{(k\pi h)\delta_k} \right]^2 &\leq h^2 \sum_{k=1}^{N} b_k^2, \\ &\leq \sqrt{C_4} h^4 \sum_{k=1}^{N} b_k^2 k^2 + \left(\sum_{\ell > N} \left(\frac{b_\ell}{\ell} \right)^2 \right)^{\frac{1}{2}} \left(\sum_{k=1}^{N} b_k^2 C_5^2 k^4 h^6 \right)^{\frac{1}{2}} + h^2 |q| \\ &\leq c_5 h^2 \sum_{k=1}^{N} a_k^2 k^2 + \left(\sum_{\ell > N} \left(\frac{b_\ell}{\ell} \right)^2 \right)^{\frac{1}{2}} \left(\sum_{k=1}^{N} b_k^2 C_5^2 k^4 h^6 \right)^{\frac{1}{2}} + h^2 |q|. \end{aligned}$$

Using an $\epsilon-$inequality ($ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$), (45), and Lemma 1, yields

$$\begin{aligned} \left(\sum_{\ell > N} \left(\frac{b_\ell}{\ell} \right)^2 \right)^{\frac{1}{2}} \left(\sum_{k=1}^{N} b_k^2 C_5^2 k^4 h^6 \right)^{\frac{1}{2}} &\leq \frac{1}{2} \sum_{\ell > N} \left(\frac{b_\ell}{\ell} \right)^2 + \frac{1}{2} \sum_{k=1}^{N} b_k^2 C_5^2 k^4 h^6, \\ &\leq \frac{1}{2} \sum_{\ell > N} a_\ell^2 \pi^2 \delta_\ell^2 + \frac{1}{2} \sum_{k=1}^{N} a_k^2 (C_6^2) C_5^2 h^4 k^4. \end{aligned}$$

For $k \leq N$, $0 < kh \leq 1$; therefore, dividing each term by kh from a sum ranging from $1 \to N$, increases the sum. Also, for $\ell > N$, $\ell h > 1$; therefore multiplying by ℓh in each term of a series ranging from $N+1 \to \infty$ increases the sum. Let $c_6 = \frac{1}{2} C_6^2 C_5^2$. Since $\delta_\ell \sim 2$, for $\ell > N$, we let $c_7 = \frac{1}{2}\pi^2 (\max \delta_\ell^2)$. Putting things back together yields

$$\sum_{k=1}^{N} a_k^2 \leq c_5 h \sum_{k=1}^{N} a_k^2 k + c_7 h \sum_{\ell > N} a_\ell^2 \ell + c_6 h \sum_{k=1}^{N} a_k^2 k + h^2 |q|. \tag{66}$$

Now add $\sum_{\ell > N} a_\ell^2$ to the left and $h \sum_{\ell > N} a_\ell^2 \ell$ to the right. We obtain

$$\sum_{k=1}^{\infty} a_k^2 \leq C_1 h \sum_{k=1}^{\infty} a_k^2 k + h^2 |q| \tag{67}$$

where $C_1 = \max\{c_5 + c_6, c_7 + 1\}$, which completes the proof. □

Thus, if the error is $\mathcal{F}$-orthogonal and *harmonic*, the *trace* of the error at $x = h$ is *oscillatory* in the sense defined above. Finally, we present the main theorem for this section.

Theorem 2. *Let E be error that is $\mathcal{F}$-orthogonal to v_h, as defined in (39). If E is harmonic in Ω^- and Ω^+, and continuous, then there exists a constant $D > 0$ such that*

$$\|\nabla E^-\|_{(0,\Omega^-)} \leq D\|\nabla E^+\|_{(0,\Omega^+)},$$

where D is independent of h and E^- and E^+ represent E restricted to Ω^- and Ω^+, respectively.

Proof Since $|q| = 0$ in Lemma 3, the proof follows directly from Theorem 1 and Lemma 3. □

3.2. Case II: Error Contains Small Local Component

We now look at the case in which the error contains a small *local* component. First, let us quantify small *local* error. We assume that the ratio of the local error in Ω versus the *harmonic* component of the error in Ω^- is small, that is,

$$\left(\|\nabla\eta^-\|_{(0,\Omega^-)} + \|\nabla\eta^+\|_{(0,\Omega^+)}\right) \leq c_1\|\nabla\mathcal{H}^-\|_{(0,\Omega^-)}. \tag{68}$$

The difficulty in this section is to show that the $trace(E) = t$ still satisfies the oscillatory assumption even if there exists a small *local* component. Therefore, we just need to properly bound $|q|$in Lemma 3.

Lemma 4. *If $|q| := |\langle\nabla\eta, \nabla\hat{v}_h\rangle|$, where $\hat{v}_h$ is defined in (52) and η satisfies (68). Then, there exists a constant c_2, independent of h, such that*

$$h^2|q| \leq c_2 h\sum_{k=1}^{\infty} a_k^2 k.$$

Proof Assume that the harmonic error has the Fourier expansion defined in (30) and (31). We start by calculating the L^2-norm of the gradient of the harmonic error in Ω^-,

$$\|\nabla\mathcal{H}^-\|_{(0,\Omega^-)} = \left(\int_0^h\int_0^1\left[\sum_{k=1}^{\infty}\frac{a_k k\pi}{\sinh(k\pi h)}(\cos(k\pi y)\sinh(k\pi x) + \sin(k\pi y)\cosh(k\pi x))\right]^2 dydx\right)^{1/2}.$$

Computing the integral yields

$$\|\nabla\mathcal{H}^-\|_{(0,\Omega^-)} = c_3 h^{-\frac{1}{2}}\left(\sum_{k=1}^{\infty} a_k^2 k\frac{h\sinh(2k\pi h)}{\sinh^2(k\pi h)}\right)^{\frac{1}{2}}, \tag{69}$$

For more details, see [6]. Splitting the sum in (69) yields the following bound.

$$\begin{aligned}\|\nabla\mathcal{H}^-\|_{(0,\Omega^-)} &= c_3\left(h^{-1}\sum_{k=1}^{N}a_k^2\frac{kh\sinh(2k\pi h)}{\sinh^2(k\pi h)}+\sum_{k>N}a_k^2k\frac{\sinh(2k\pi h)}{\sinh^2(k\pi h)}\right)^{\frac{1}{2}},\\ &\leq c_3\left(c_4h^{-1}\sum_{k=1}^{N}a_k^2+c_5\sum_{k>N}a_k^2k\right)^{\frac{1}{2}},\\ &\leq \hat{c}_3\left(h^{-1}\sum_{k=1}^{N}a_k^2+\sum_{k>N}a_k^2k\right)^{\frac{1}{2}},\end{aligned}\tag{70}$$

where $\hat{c}_3 \leftarrow c_3\max\{c_4, c_5\}$.

Now bound $\|\nabla\hat{v}_h\|_{(0,\Omega)}$, which only has support in $x < 2h$. We introduce some new functions to ease visualization. Since $\hat{v}_h$ is symmetric about $x = h$, we only define $u(x, y)$ and $z(x, y)$ (below) for $x \leq h$:

$$\frac{\partial}{\partial x}\hat{v}_h = u(x,y) = \frac{1}{h}\sum_{k=1}^{N}b_k(\sin k\pi y+\varepsilon_k(y))$$

and

$$\frac{\partial}{\partial y}\hat{v}_h = z(x,y) = \frac{x}{h}\sum_{k=1}^{N}b_k(k\pi\cos k\pi y+\varepsilon_k'(y)).$$

Then,

$$\|\nabla\hat{v}_h\|^2_{(0,\Omega^-\cup\Omega^r)} = 2\int_0^1\int_0^h u^2dxdy+2\int_0^1\int_0^h z^2dxdy.\tag{71}$$

Integrating the first term in (71) and using the orthogonality of ε_k's, as established in Corollary 1, yields

$$\begin{aligned}2\int_0^1\int_0^h u^2dxdy = &\frac{2}{h}\sum_{k=1}^{N}b_k^2\int_0^1\sin^2(k\pi y)dy+\frac{2}{h}\sum_{k=1}^{N}b_k^2\int_0^1\varepsilon_k^2(y)dy\\ &+\frac{4}{h}\int_0^1\left(\sum_{k=1}^{N}b_k\sin(k\pi y)\sum_{j=1}^{N}b_j\varepsilon_j(y)\right)dy.\end{aligned}$$

After examination of each term on the right-hand side (see [6]), we get the following bound.

$$2\int_0^1\int_0^h u^2dxdy \leq \frac{c_6}{h}\sum_{k=1}^{N}b_k^2.\tag{72}$$

Integrating the second term in (71) and using Corollary 2 yields

$$2\int_0^1\int_0^h z^2dxdy = \frac{2h}{3}\sum_{k=1}^N b_k^2(k\pi)^2\int_0^1\cos^2(k\pi y)dy+ \frac{2h}{3}\sum_{k=1}^N b_k^2\int_0^1[\epsilon_k'(y)]^2dy+ \frac{4h}{3}\int_0^1\left(\sum_{k=1}^N b_k(k\pi)\cos(k\pi y)\sum_{j=1}^N b_j\epsilon_j'(y)\right)dy$$

Again, after examination of each term on the right-hand side (see [6]), we get the following bound.

$$2\int_0^1\int_0^h z^2dxdy \le \frac{c_7}{h}\sum_{k=1}^N b_k^2. \tag{73}$$

Let $\hat{c}_7 = (c_6 + c_7)$. Then,

$$\|\nabla v_h\|_{0,\Omega}^2 \le \frac{\hat{c}_7}{h}\sum_{k=1}^N b_k^2 = \frac{\hat{c}_7}{h}\sum_{k=1}^N a_k^2k^2\pi^2\delta_\ell^2 \le \frac{\hat{c}_7(\pi\hat{\delta})^2}{h^2}\sum_{k=1}^N a_k^2k, \tag{74}$$

where $\hat{\delta} = \max\delta_\ell$ for $\ell \le N$.

With bounds on $\|\nabla v_h\|_{0,\Omega}$ and $\|\nabla\mathcal{H}^-\|_{(0,\Omega^-)}$, we now bound $|q|$. First, separate $|q|$ into regions Ω^- and Ω^+ and then use Cauchy-Schwarz to get

$$\begin{aligned} h^2|q| &= h^2|\left\langle\nabla\eta^-,\nabla v^h\right\rangle_{\Omega^-} + \left\langle\nabla\eta^+,\nabla\hat{v}_h\right\rangle_{\Omega^+}| \\ &\le h^2\left(\|\nabla\eta^-\|_{(0,\Omega^-)} + \|\nabla\eta^+\|_{(0,\Omega^+)}\right)\|\nabla\hat{v}_h\|_{(0,\Omega)}. \end{aligned}$$

Substituting bound (68) for $\left(\|\nabla\eta^-\|_{(0,\Omega^-)} + \|\nabla\eta^+\|_{0,\Omega^+}\right)$ yields

$$h^2|q| \le h^2c_1\|\nabla v_h\|_{(0,\Omega)}\|\nabla\mathcal{H}^-\|_{(0,\Omega^-)}$$

Substituting equations (70) and (74) yields

$$h^2|q| \le c_1\pi\hat{\delta}\sqrt{\hat{c}_7}\sqrt{\hat{c}_3}\left(h\sum_{k=1}^N a_k^2\right)^{\frac{1}{2}}\left(\sum_{k=1}^N a_k^2 + h\sum_{k>N} a_k^2k\right)^{\frac{1}{2}}.$$

Letting $c_8 = c_1\pi\hat{\delta}\sqrt{\hat{c}_3\hat{c}_7}$ and applying an ε-inequality ($ab \le \frac{1}{4p^2}a^2 + p^2b^2$, $\forall p \in \mathbb{R}$) gives way to

$$h^2|q| \le \frac{c_8}{4p^2}\sum_{k=1}^N a_k^2 + \frac{c_8h}{4p^2}\sum_{k>N} a_k^2k + c_8p^2h\sum_{k=1}^N a_k^2k.$$

Choosing p such that $c_9 := \dfrac{c_8}{4p^2} < 1$ and letting $c_{10} := c_8p^2$ yields

$$\begin{aligned} h^2|q| &\le c_9\sum_{k=1}^N a_k^2 + c_9h\sum_{k>N} a_k^2k + c_{10}h\sum_{k=1}^N a_k^2k \\ &\le c_9\sum_1^\infty a_k^2 + c_{11}h\sum_1^\infty a_k^2k, \end{aligned}$$

where $c_{11} = \max\{c_9, c_{10}\}$. Finally,

$$\sum_{k=1}^{\infty} a_k^2 \leq \frac{c_0 + c_{11}}{1 - c_9} h \sum_{k=1}^{\infty} a_k^2 k. \tag{75}$$

Since $c_9 < 1$, the lemma is proved with $c_2 = \dfrac{c_0 + c_{11}}{1 - c_9}$. □

The main result of this subsection is given in the following theorem.

Theorem 3. *Let E be error that is $\mathcal{F}$-orthogonal, as defined in (39). Assume the local components of E in Ω^- and Ω^+ are bounded as follows:*

$$\left(\|\nabla \eta^-\|_{(0,\Omega^-)} + \|\nabla \eta^+\|_{(0,\Omega^+)}\right) \leq c_1 \|\nabla \mathcal{H}^-\|_{(0,\Omega^-)}.$$

If E has a harmonic component $\mathcal{H}^-$ in Ω^- and a harmonic component $\mathcal{H}^+$ in Ω^+, and if $\mathcal{H}^- = \mathcal{H}^+$ on Γ, then there exists a constant $D > 0$, independent of h such that

$$\|\nabla \mathcal{H}^-\|_{(0,\Omega^-)} \leq D \|\nabla \mathcal{H}^+\|_{0,\Omega^+}.$$

Proof The combination of Theorem 1, and Lemma 3 completes the proof. □

To summarize, if E is minimum on grid h, then E is oscillatory on any Γ, as stated in (26). If E is oscillatory, then harmonic error is balance, as concluded in Theorem 3. It is important to note that the constant D is reasonable on a square domain when η^+ and η^- are equal to zero. Also, the constant D depends on the constant c_1 when η^+ and η^- are non-zero. Therefore, when c_1 is large, D is large. Here we assume c_1 is sufficiently small to insure a D that is not too large, however, in Section 6 we address the case for large c_1.

The following corollary, which determines a bound on the total energy, completes this subsection.

Corollary 3. *Let $\mathcal{H}^-$ and $\mathcal{H}^+$ be the harmonic components of the error, E, restricted to Ω^- and Ω^+, respectively. Let η^- and η^+ be the local components of E restricted to Ω^- and Ω^+, respectively. Assume the local component of E in Ω^- is bounded as follows:*

$$\|\nabla \eta^-\|_{(0,\Omega^-)} \leq c_3 \|\nabla E^-\|_{(0,\Omega^-)}$$

for some $c_3 < 1$, independent of h. If

$$\|\nabla \mathcal{H}^-\|_{(0,\Omega^-)} \leq D \|\nabla \mathcal{H}^+\|_{0,\Omega^+},$$

then there exists a constant $\hat{D} > 0$, independent of h, such that

$$\|\nabla E^-\|_{(0,\Omega^-)} \leq \hat{D} \|\nabla E^+\|_{0,\Omega^+}.$$

Proof First bound the energy in Ω^- of the *harmonic* error by the energy of the total error in Ω^+:

$$\|\nabla \mathcal{H}^-\|^2_{(0,\Omega^-)} \leq D^2 \left(\|\nabla \mathcal{H}^+\|_{(0,\Omega^+)} + \|\nabla \eta^+\|_{(0,\Omega^+)}\right)^2 = D^2 \|\nabla E^+\|^2_{(0,\Omega^+)}.$$

We also know that

$$\|\nabla \eta^-\|^2_{(0,\Omega^-)} \leq c_3^2 \|\nabla E^-\|^2_{(0,\Omega^-)} \leq c_3^2 \left(\|\nabla \mathcal{H}^-\|^2_{(0,\Omega^-)} + \|\nabla \eta^-\|^2_{(0,\Omega^-)}\right).$$

This implies that

$$(1-c_3)\|\nabla\eta^-\|^2_{(0,\Omega^-)} \le c_3^2\|\nabla\mathcal{H}^-\|^2_{(0,\Omega^-)}.$$

Now sum the energy in Ω^- of the *harmonic* and *local* error.

$$\begin{aligned}\|\nabla\eta^-\|^2_{(0,\Omega^-)} &\le \left(\frac{c_3^2}{1-c_3^2}\right) D^2\|\nabla E^+\|^2_{0,\Omega^+}\\ +\|\nabla\mathcal{H}^-\|^2_{(0,\Omega^-)} &\le D^2\|\nabla E^+\|^2_{0,\Omega^+}\\ \hline \|\nabla E^-\|^2_{(0,\Omega^-)} &\le \hat{D}^2\|\nabla E^+\|^2_{0,\Omega^+},\end{aligned}$$

where $\hat{D}^2 = D^2\left(1+\frac{c_3^2}{1-c_3^2}\right)$. □

This corollary is developed for relatively small *local* error components, thus c_3 is bounded away from 1. The impact of Corollary 3 is described in the following statement. If the local error in the refinement region is small compared to the total error in the refinement region, then the total error in the refinement region is small compared to the total error in the whole domain.

4. Algebraic Theory

This section takes an algebraic approach to establishing sharpness of our error measure. The aim is to show algebraically that our refinement indicator cannot be distracted by error that is large but locally irreducible in a subregion. To this end, assume that the problem has been solved on the current level, a global grid of mesh-size $2h$. We are thus able to restrict our attention to error that is algebraically energy-orthogonal, as shown below, to the space associated with this level.

$$< Aw, v >= v^t Aw = 0, \tag{76}$$

where A is a linear operator, w and v are vectors, and t denotes the transpose. This algebraic assumption enables us to define *discrete harmonic* error as error whose grid-h residuals are zero in the interior of the refinement region and its complement, that is, the residuals are nonzero only at the interface between the region and its complement. The discrete error can then be written as a linear combination of a *discrete harmonic* and error that is *local* in the sense that it is a grid-h vector that is nonzero only in the interior of the refinement region.

In this section, we first develop an algebraic result for the case when the error is exactly a *discrete harmonic*. We then extend the result to the case that the error is an approximate *discrete harmonic* (nonzero local error). A finite element discretization of a planar boundary value problem is then explored. Analogous to the continuous case, we have in mind that the bounded planar domain is partitioned by a closed curve (corresponding to Γ) into two subdomains (corresponding to Ω^- and Ω^+). For the algebraic proof, we consider Γ, Ω^-, and Ω^+ to be point sets denoted by $\tilde{\mathcal{R}}_1$, $\tilde{\mathcal{R}}_2$, and $\tilde{B}$, respectively.

4.1. Algebraic Theory for Discrete Harmonics

Consider an $n \times n$ real symmetric positive definite (SPD) matrix, $\mathcal{A}$, whose entries have been partitioned into point sets $\tilde{\mathcal{R}}_1, \tilde{\mathcal{R}}_2$, and $\tilde{B}$. Although $\mathcal{A}$ is presented with the above generality,

it is convenient, for this paper, to consider $\mathcal{A}$ the FOSLS discretization of the Laplace operator. Assume that the resulting block representation of $\mathcal{A}x$ is given as follows:

$$\mathcal{A}x = \begin{bmatrix} \mathcal{R}_1 & 0 & \mathcal{C}_1 \\ 0 & \mathcal{R}_2 & \mathcal{C}_2 \\ \mathcal{C}_1^t & \mathcal{C}_2^t & B \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ y \end{bmatrix},$$

where x_1, x_2, and y are variables at the points in $\tilde{\mathcal{R}}_1$, $\tilde{\mathcal{R}}_2$, and $\tilde{B}$, respectively. This matrix partitioning is analogous to a non-overlapping domain decomposition with 2 sub-domains. Point sets $\tilde{\mathcal{R}}_1$ and $\tilde{\mathcal{R}}_2$ are considered to be interior points; these points have no connections outside of their respective domains. The interface points between $\tilde{\mathcal{R}}_1$ and $\tilde{\mathcal{R}}_2$ are given by $\tilde{B}$. The entries in blocks $\mathcal{C}_1$ and $\mathcal{C}_2$ account for the connections from the interiors of the respective domains to the interface, and the entries in blocks $\mathcal{C}_1^t$ and $\mathcal{C}_2^t$ contain the connections from the interface to the sub-domains. The interface block is expanded into its finite element contribution from $\tilde{\mathcal{R}}_1$ and $\tilde{\mathcal{R}}_2$ as follows:

$$B = B^l + B^r. \tag{77}$$

The Schur complement of the interior blocks in $\mathcal{A}$ is often used as a tool to determine the inverse of $\mathcal{A}$. We use it so that the energy in a sub-domain can be computed, where by the energy of an n-vector, x, we mean $\langle \mathcal{A}x, x \rangle \equiv x^t \mathcal{A}x$. Let

$$S = B - \mathcal{C}_1^t \mathcal{R}_1^{-1} \mathcal{C}_1 - \mathcal{C}_2^t \mathcal{R}_2^{-1} \mathcal{C}_2 \tag{78}$$

denote the Schur complement of the interior blocks in $\mathcal{A}$, which we also write as

$$S = S^l + S^r, \tag{79}$$

where

$$S^l = B^l - \mathcal{C}_1^t \mathcal{R}_1^{-1} \mathcal{C}_1 \tag{80}$$

and

$$S^r = B^r - \mathcal{C}_2^t \mathcal{R}_2^{-1} \mathcal{C}_2. \tag{81}$$

Note that the inverses that define these Schur complements are guaranteed to exist because $\mathcal{A}$ is SPD. Furthermore, S is SPD (see Theorem 1.2 in [8]). We now introduce a few auxiliary matrices. Let n_B be the number of points on the interface or, equivalently, the dimension of B. Matrices defined with script notation have dimensions not related to n_B. Let Z be an $n_B \times n_B$ matrix formed by deleting the even-numbered columns of the identity matrix, where the column indices start with 1 (as apposed to 0):

$$Z = \begin{bmatrix} 1 & & & & & \\ & 0 & & & & \\ & & 1 & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 1 \end{bmatrix}.$$

Z is used below to characterize oscillatory error.

In what follows, we use the Löwner partial order, $U \geq V$, for symmetric matrices, U and V, to mean that $U-V$ is nonnegative definite and $U > V$ to mean that $U-V$ is positive definite. Let T be an $n_B \times n_B$ SPD matrix defined on y; it can be thought of as a one-dimensional finite element discretization of the Poisson PDE along the interface, $\tilde{B}$. For this general algebraic proof, we assume there exist constants c_0, c_1, $\gamma \in \mathbb{R}^+$ such that the following hold:

1. $S^r \geq c_0 T$;
2. $B \leq c_1 I$;
3. $Z^t T^3 Z \geq \gamma Z^t T Z$.

If $\mathcal{A}$ is an SPD matrix resulting from a finite element discretization of the model Poisson PDE presented in Section 2, then Assumptions 1 and 2 are readily satisfied. If T is constructed as a one-dimensional finite element discretization of the Poisson PDE with a $[-1\ 2\ -1]$ stencil, then Assumption 3 is also readily satisfied. For now, consider $\mathcal{A}$ and T to be general matrices satisfying the above conditions and decomposition. In Subsection 4.3, we demonstrate how this general theory is applied to an actual PDE discretization.

Let $u \in \mathbb{R}^n$ be a *discrete harmonic* in the sense that its residual, $\mathcal{A}u$, is 0 in $\tilde{\mathcal{R}}_1$ and $\tilde{\mathcal{R}}_2$. In the continuum theory we exploited the oscillatory nature of the function that defined the error along the line segment Γ (we denoted this function as the trace of the error on Γ). For the proof below, we also exploit the notion of an oscillatory trace. Let $\bar{u}_B$ be an n_B-vector restricted to the points on $\tilde{B}$. We call u oscillatory on $\tilde{B}$ if its restriction to $\tilde{B}$, u_B, satisfies $u_B = S^{-1}TZ\bar{v}$ for some $\bar{v}$ defined on the odd-numbered points of $\tilde{B}$. If T is represented by the stencil $[-1\ 2\ -1]$, then

$$Su_B = TZ \begin{bmatrix} v_1 \\ * \\ v_2 \\ \vdots \\ * \\ v_{\frac{n}{2}} \end{bmatrix} = \begin{bmatrix} 2v_1 \\ -v_1 - v_2 \\ 2v_2 \\ \vdots \\ -v_{\frac{n}{2}-2} - v_{\frac{n}{2}-1} \\ 2v_{\frac{n}{2}} \end{bmatrix}$$

Fortunately, as was the case in the continuous setting, this condition follows as a consequence of the discrete harmonic, u, being energy-orthogonal to the range of an interpolation operator. This is important because, for actual PDE applications, we are interested in error that is energy-orthogonal to a given global grid of mesh-size $2h$. Interpolation in this setting maps a vector on a grid with mesh-size $2h$ to a vector on a grid with mesh-size h, which we assume has the standard form

$$P = \begin{bmatrix} \mathcal{P}_{11} & \mathcal{P}_{12} & \mathcal{P}_{13} \\ \mathcal{P}_{21} & \mathcal{P}_{22} & \mathcal{P}_{23} \\ 0 & 0 & Q \end{bmatrix}.$$

Here, P is an interpolation operator that takes values in the grid-level $2h$ subspace and maps them to the grid-level h subspace. Energy orthogonality of u then allows us to conclude that $Q^T S u_B = Q^T T Z \bar{v} = 0$.

We now prove that the energy of u on $\tilde{\mathcal{R}}_2$ is bounded from below by a positive constant times the total energy, $\langle \mathcal{A}u, u \rangle$. This is equivalent to bounding the energy in $\tilde{\mathcal{R}}_1$ by a constant times the energy in $\tilde{\mathcal{R}}_2$. Note that the discrete harmonic property of u implies that $\langle \mathcal{A}u, u \rangle = \langle Su_B, u_B \rangle$. The energy of u on $\tilde{\mathcal{R}}_2$ is then defined as $\langle S^r u_B, u_B \rangle$.

Theorem 4. *Under Assumptions 1, 2, and 3, any u_B in the range of $S^{-1}TZ$ has the property that*

$$\langle S^r u_B, u_B \rangle \geq \gamma \left(\frac{c_0}{c_1} \right)^2 \langle S u_B, u_B \rangle . \tag{82}$$

Proof Any u_B is in the range of $S^{-1}TZ$, so (82) is equivalent to

$$Z^t T S^{-1} S^r S^{-1} T Z \geq \gamma \left(\frac{c_0}{c_1} \right)^2 Z^t T S^{-1} T Z, \tag{83}$$

which we now prove.

First, since $B - S$ is SPD, then $S \leq B$. Combining this with Assumption 2 ($B \leq c_1 I$) yields $S^{-1} \geq c_1^{-1} I$. Thus,

$$S^{-1} S^r S^{-1} = (S^r)^{-1/2} [(S^r)^{1/2} S^{-1} (S^r)^{1/2}]^2 (S^r)^{-1/2} \geq \left(\frac{1}{c_1} \right)^2 S^r,$$

implying

$$Z^t T S^{-1} S^r S^{-1} T Z \geq \left(\frac{1}{c_1} \right)^2 Z^t T S^r T Z. \tag{84}$$

Second, applying Assumption 1 ($S^r \geq c_0 T$) to the right-hand side of (84) yields

$$Z^t T S^{-1} S^r S^{-1} T Z \geq \left(\frac{c_0}{c_1^2} \right) Z^T T^3 Z. \tag{85}$$

Third, by Assumption 3 ($Z^t T^3 Z \geq \gamma Z^t T Z$) we have

$$Z^t T S^{-1} S^r S^{-1} T Z \geq \gamma \left(\frac{c_0}{c_1^2} \right) Z^T T Z. \tag{86}$$

Last, $S^l = S - S^r$ being SPD implies $S \geq S^r$. Combining this with Assumption 1 gives $T^{-1} \geq c_0 S^{-1}$, thus yielding

$$T \geq c_0 T S^{-1} T. \tag{87}$$

Using this in (86) yields the result in (83). □

4.2. Algebraic Theory for Approximate Discrete Harmonics

Now consider the case of small but not necessarily zero *local* error. For convenience, let $\tilde{\mathcal{R}} := \tilde{\mathcal{R}}_1 \bigcup \tilde{\mathcal{R}}_2$. The block partition of $\mathcal{A}$ is then given by

$$\mathcal{A} = \begin{bmatrix} \mathcal{R} & \mathcal{C} \\ \mathcal{C}^T & B \end{bmatrix},$$

where $\mathcal{C}^T$ and $\mathcal{C}$ represent the connections from the sub-domains to and from the interface, respectively. We similarly represent $\mathcal{P}$ and $\mathcal{P}^T$ as follows:

$$\mathcal{P} = \begin{bmatrix} \mathcal{P}_1 & \mathcal{P}_2 \\ 0 & Q \end{bmatrix}, \quad \mathcal{P}^T = \begin{bmatrix} \mathcal{P}_1^T & 0 \\ \mathcal{P}_2^T & Q^T \end{bmatrix}.$$

As was the case for the continuum theory, showing that the trace is oscillatory is key. For this section, we show that an approximate discrete harmonic that is orthogonal in energy to the range of $\mathcal{P}$ must be approximately oscillatory. This means that

$$u \approx \begin{bmatrix} -\mathcal{R}^{-1}\mathcal{C} \\ I \end{bmatrix} S^{-1}TZ\bar{v}.$$

We use the $\mathcal{P}$ defined above to construct a coarse-grid matrix:

$$\mathcal{A}_c = \mathcal{P}^T\mathcal{A}\mathcal{P} = \begin{bmatrix} \mathcal{R}_c & \mathcal{C}_c \\ \mathcal{C}_c^T & B_c \end{bmatrix}.$$

Assume that this coarse-grid operator gives way to the following relationship between the Schur complement on the fine-grid and the Schur complement on the coarse-grid:

$$\langle S_c\bar{w}_c, \bar{w}_c\rangle \leq \lambda \langle SQ\bar{w}_c, Q\bar{w}_c\rangle, \tag{88}$$

where $\bar{w}_c$ is an n_{B_c}-vector. Equation (88) is satisfied for the 2D model problem (see [6]) containing the two strip sub-domains depicted in Figure 1.

Theorem 5. *Suppose that an n-vector, u, is orthogonal in energy to Range($\mathcal{P}$) and that it has a relatively small local component in the sense that*

$$u = \begin{bmatrix} -\mathcal{R}^{-1}\mathcal{C} \\ I \end{bmatrix} \bar{u} + \begin{bmatrix} \ell \\ 0 \end{bmatrix},$$

where $\langle \mathcal{R}\ell, \ell\rangle \leq \epsilon \langle \mathcal{A}u, u\rangle$ for some small $\epsilon \in \mathbb{R}^+$. Then u is an approximate oscillatory discrete harmonic in the sense that

$$u = \begin{bmatrix} -\mathcal{R}^{-1}\mathcal{C} \\ I \end{bmatrix} S^{-1}TZ\bar{v} + v,$$

where $\langle \mathcal{A}v, v\rangle \leq (1+\lambda)\epsilon \langle \mathcal{A}u, u\rangle$.

Proof See [6].

Since u is an approximate oscillatory discrete harmonic, we use Theorem 4 to bound the energy in $\tilde{\mathcal{R}}_2$ from below. We first bound the discrete harmonic component of u in $\tilde{\mathcal{R}}_2$:

$$\begin{aligned} &\langle \mathcal{A}v, v\rangle \leq (1+\lambda)\epsilon \langle \mathcal{A}u, u\rangle \\ \Rightarrow &\langle \mathcal{A}u, u\rangle + \langle \mathcal{A}v, v\rangle \leq (1+\lambda)\epsilon \langle \mathcal{A}u, u\rangle + \langle Su_B, u_B\rangle + \langle \mathcal{A}v, v\rangle \\ \Rightarrow &\langle Su_B, u_B\rangle \geq (1-(1+\lambda)\epsilon) \langle \mathcal{A}u, u\rangle. \end{aligned}$$

Using Theorem 4, we therefore have

$$\begin{aligned} \langle S^r u_B, u_B\rangle &\geq \gamma \left(\frac{c_0}{c_1}\right)^2 \langle Su_B, u_B\rangle, \\ \Rightarrow \langle S^r u_B, u_B\rangle &\geq \hat{C} \langle \mathcal{A}u, u\rangle, \end{aligned} \tag{89}$$

where $\hat{C} = \gamma(\frac{c_0}{c_1})^2(1-(1+\lambda)\epsilon)$. Thus, we are left with the energy in $\tilde{\mathcal{R}}_2$ bounded from below by a constant times the energy in the whole domain, which was the desired goal.

4.3. Application to PDEs

Consider now the general case that A results from a conforming finite element discretization of a PDE on a bounded open domain, $\Omega \subset \mathbb{R}^2$, with rectilinear boundary $\partial\Omega$. Consider a closed curve, $\bar{B} \subset \bar{\Omega}$, that aligns with element boundaries. Next, assume that $\mathcal{A}$ is assembled from a corresponding finite element discretization of the PDE on two domains, $\Omega^-, \Omega^+ \subset \Omega$, that $\bar{B}$ defines. To be more precise, first let Ω^- be the sub-domain of $\mathbb{R}^2$ that is bounded by $\bar{B}$, let Ω^+ be the interior of the complement of Ω^- in Ω, and let $\tau = \bar{B} \cap \Omega$ (the segment of $\bar{B}$ that is not coincident with $\partial\Omega$). The nodal partition is now defined by letting $\tilde{B}$, $\tilde{\mathcal{R}}_1$, and $\tilde{\mathcal{R}}_2$ be the nodes in τ, Ω^-, and Ω^+, respectively.

As an example, consider a simple model Poisson problem with homogeneous Dirichlet boundary conditions on a rectangle. We use standard linear elements corresponding to a uniform $(n+1) \times (n+1)$ grid h of triangles, with $n+1$ an even positive integer (to facilitate coarsening by a factor of two). The partitioning geometry is slab-like, defined by letting τ be a conforming vertical line segment, $\{y = 2kh\} \cap \Omega$, for some positive integer $k < (n+1)/2$, and $\tilde{\mathcal{R}}_1$ and $\tilde{\mathcal{R}}_2$ the corresponding respective left and right rectangular subregions of Ω with common boundary τ. Note that this discretization results in the usual five-point stencil corresponding to $\mathcal{A}$, thus yielding

$$B = T + 2I,$$

where T is the matrix defined on $\tilde{B}$ (the nodes in τ) that corresponds to the three-point stencil $[-1\ 2\ -1]$ and I is the identity matrix on $\tilde{B}$. We are thus led to make the following choices for the decomposition in (77):

$$B^r = B^l = \frac{1}{2}T + I.$$

Note that this choice corresponds to the individual discretization of the Laplacian on each slab, that is, $\frac{1}{2}T + I$ is the contribution to B from the elements abutting the left (right) side of τ.

Assumptions 1 and 2, from the previous subsection, are now easily verified. To confirm Assumption 3, let $T = \frac{1}{2}T$ and note that

$$T^3 = \frac{1}{8}\begin{bmatrix}
14 & -14 & 6 & -1 & & & & & & \\
-14 & 20 & -15 & 6 & -1 & & & & & \\
6 & -15 & 20 & -15 & 6 & -1 & & & & \\
-1 & 6 & -15 & 20 & -15 & 6 & -1 & & & \\
 & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & & \\
 & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \\
 & & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & -1 \\
 & & & & \ddots & \ddots & \ddots & \ddots & -15 & 6 \\
 & & & & & \ddots & \ddots & -15 & 20 & -14 \\
 & & & & & & -1 & 6 & -14 & 14
\end{bmatrix}.$$

We then compute

$$Z^tT^3Z = \frac{1}{8}\begin{bmatrix} 14 & 6 & & & & & & \\ 6 & 20 & 6 & & & & & \\ & 6 & 20 & 6 & & & & \\ & & . & . & . & & & \\ & & & . & . & . & & \\ & & & & 6 & 20 & 6 & \\ & & & & & 6 & 14 & \end{bmatrix}$$

and

$$Z^tTZ = \frac{1}{2}\begin{bmatrix} 14 & & & & & & & \\ & 20 & & & & & & \\ & & . & & & & & \\ & & & . & & & & \\ & & & & . & & & \\ & & & & & 20 & & \\ & & & & & & 14 & \end{bmatrix}. \tag{90}$$

It is then easy to see by Gerschgorin's theorem that Assumption 3 is satisfied for $\gamma \leq \frac{1}{7}$, since all of the eigenvalues of $Z^tT^3Z - \frac{1}{7}Z^tTZ$ are nonnegative.

5. Numerical Results

We now test the impact of the theory developed in the previous sections by numerically applying our refinement algorithm [6] to a problem that is simple enough to focus on the results we have developed. Although the error that we describe below is very specific, this example illustrates the importance of the theory presented in the previous sections. This example also helps to validate the findings of the theory. We show that with a scenario consistent to that presented in the previous sections, if the refinement algorithm lead us to only refine on one side of an oscillatory trace, significant reduction of the functional was not achieved. However, even when one side of an oscillatory trace is very "thin", the refinement algorithm chooses to refine elements on both sides of the trace, and therefore significantly reduces the functional.

With Ω a square partitioned as in Subsection 4.3, we seed the FOSLS functional with a purely harmonic error defined by extending a nonzero trace on the line segment $x = h$ harmonically into $x < h$ and $x > h$. We assume an idealized FOSLS functional defined as the squared H^1 seminorm of the error. In other words, we are minimizing $\|\nabla(e(x,y) - e_h(x,y))\|^2_{0,\Omega}$, where

$$e(x,y) = \begin{cases} \sin(16\pi y)\frac{\sinh(16\pi x)}{\sinh(\pi)} & \text{when } y \leq h, \\ \sin(16\pi y)\frac{\sinh(16\pi(1-x))}{\sinh(16\pi(1-1/16))} & \text{when } y > h, \end{cases}$$

$h = 1/16$, and boundary conditions are homogeneous Dirichlet.

Three key questions arise. First, where is the local FOSLS functional large? Second, does refinement substantially decrease the functional value if there is only refinement to the left of $x = h$? Last, does the algorithm mark the proper elements for refinement?

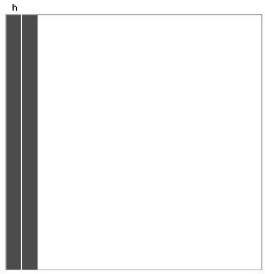

Figure 3. Functional values after solving minimization problem on a grid with mesh size $h = \frac{1}{16}$. The darker the element the larger the functional value.

To answer the first question, we simply plot the functional values. It is easy to see by looking at Figure 3 that the local functional values are largest for elements located just to the left and right of the line $x = h$, as our theory asserted.

Since the error is hrmonic and oscillatory, we expect that refining strictly in the region to the left of $x = h$ reduces the functional marginally, at best. We test this assumption by first locally refining 2 elements in the region to the left of the line $x = h$. We then increase the the number of elements refined to 4, 8, and 16.

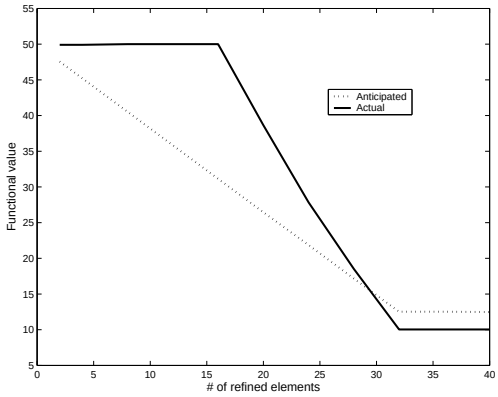

Figure 4. The functional values are given for refinement of different numbers of elements. The two plots represent the actual computed functional versus the anticipated functional.

Figure 4 shows that, even when all of the elements to the left of $x = h$ are locally refined, there is no reduction in the functional. When we start refining directly to the right of $x = h$, the functional starts to drop and, by the time we refine 32 elements, the reduction in the functional is similar to the anticipated value.

The final question is answered by looking at the plot of the mesh after local refinement. Figure 5 shows that the refinement algorithm (described in [6]) marks the optimal (if work is ignored and equations are solved exactly) number and location of elements to refine for effective

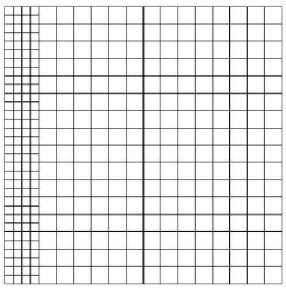

Figure 5. Finite element mesh after one level of local adaptive refinement. The original mesh-spacing is h. Local refinement occurs on both sides of $x = h$.

functional reduction. For this example optimal first implies that refining more elements would not significantly decrease the global functional. Second, since the elements that were refined contributed the identical reduction in the global functional, no specific element is less important to refine.

6. Convergence

In this section, we use the theorems in Section 3 and 4 to show convergence for adaptive refinement schemes that use the FOSLS functional as the local error indicator (for example, see [6]).

We first introduce some new notation. Let $e_h := u - u_h$, where u is the exact solution to a specified PDE and u_h is the FOSLS discrete solution. Then $\mathcal{G}(e_h; 0) = \mathcal{G}(u_h; f)$ denotes the value of the FOSLS functional at the discrete solution on level h. For ease of notation, let $\mathcal{G}(e_h) = \mathcal{G}(e_h; 0)$. Then $\mathcal{G}_{\Omega^-}(e_h)$ denotes the squared functional norm restricted to the refinement region, Ω^-. Consider, as in the continuum theory, the $\mathcal{F}$-orthogonal decomposition

$$e_h = \eta^- + \eta^+ + \mathcal{H},$$

where η^- and η^+ are local errors in Ω^- and its complement, Ω^+, respectively, and $\mathcal{H}$ is "Ω-harmonic" in the sense that it is $\mathcal{F}$-orthogonal to any errors local to Ω^- or Ω^+. Also consider the $\mathcal{F}$- orthogonal decomposition for the $h/2$-grid:

$$e_{(h/2)} = \eta^-_{(h/2)} + \eta^+_{(h/2)} + \mathcal{H}_{(h/2)}, \tag{91}$$

where $\eta^-_{(h/2)}$ and $\eta^+_{(h/2)}$ are the best $h/2$- Ω^- and Ω^+-local approximations to η^- and η^+, respectively, and $\mathcal{H}_{(h/2)}$ is $\mathcal{F}$-orthogonal to $\eta^-_{(h/2)}$ or $\eta^+_{(h/2)}$. Let

$$\delta^-_{(h/2)} = \eta^- - \eta^-_{(h/2)}, \tag{92}$$

$$\delta^+_{(h/2)} = \eta^+ - \eta^+_{(h/2)}, \tag{93}$$

and

$$\delta^{\mathcal{H}}_{(h/2)} = \mathcal{H} - \mathcal{H}_{(h/2)}. \tag{94}$$

We now show that schemes that use the FOSLS functional as the local error indicator are guaranteed to converge if the following condition is satisfied.

Conditions 1. *There exists a constant $\zeta < 1$, independent of h, such that*

$$\mathcal{G}(e_h - \delta^-_{(h/2)}) = \mathcal{G}(\eta^- - \delta^-_{(h/2)}) \leq \zeta \mathcal{G}(\eta^-). \tag{95}$$

Condition (95) implies that local refinement in Ω^- results in reduction of the local error, η^-. This property is defined as *local saturation* for this paper. We now show that convergence is guaranteed if the above condition is satisfied.

Theorem 6. *If condition (95) is satisfied, then there exists constants ϵ and μ such that*

$$\left(\mathcal{G}(\eta^+) + \mathcal{G}(\mathcal{H}^+)\right) \leq \epsilon \left(\mathcal{G}(\eta^-) + \mathcal{G}(\mathcal{H}^-)\right) \tag{96}$$

and

$$\mathcal{G}(e_h - \delta^-_{(h/2)}) \leq \mu \mathcal{G}(e_h), \tag{97}$$

where ϵ and μ are independent of h and strictly less than 1.

Proof This theorem is beneficial in showing convergence when ϵ is not arbitrarily small. We proceed through the proof with this in mind. Consider the following two cases:

- $\mathcal{G}(\mathcal{H}^-) \leq (c_1 \hat{c} + \hat{D}) \mathcal{G}(\mathcal{H}^+)$,
- $\mathcal{G}(\mathcal{H}^-) \leq \frac{1}{c_1} \left(\mathcal{G}(\eta^-) + \mathcal{G}(\eta^+)\right)$.

The constants $\hat{c}$ and $\hat{D}$ are determined in the proof of Theorem 3. The constant c_1 is chosen so that case 1 and/or 2 is valid. If c_1 is small or zero relative to the harmonic error in a possible refinement region, then the first bullet is valid, as determined in Section 3. It is important to note that $\hat{D}$ does not depend on h and depends largely on the shape of the domain. The second bullet covers the situation in which c_1 is large. This implies that the local error is large in a possible refinement region.

We first prove Theorem 6 applying the first bullet. Let $D = c_1 \hat{c} + \hat{D}$. The *local saturation* condition and the first bullet above yields

$$\begin{aligned} \mathcal{G}(e_h - \delta^-_{(h/2)}) &\leq \zeta \mathcal{G}(\eta^-) + \mathcal{G}(\eta^+) + (1 + D) \mathcal{G}(\mathcal{H}^+) \\ &\leq \zeta \mathcal{G}(\eta^-) + \mathcal{G}(\eta^+) + (1 + D)(\mathcal{G}(\mathcal{H}^+) + \mathcal{G}(\eta^+)). \end{aligned} \tag{98}$$

Thus, using equation (96) yields the following desired result

$$\begin{aligned} \mathcal{G}(e_h - \delta^-_{(h/2)}) &\leq \zeta \mathcal{G}(\eta^-) + \epsilon(1 + D)(\mathcal{G}(\mathcal{H}^-) + \mathcal{G}(\eta^-)), && (99) \\ &\leq (\zeta + \epsilon(1 + D))(\mathcal{G}(\mathcal{H}^-) + \mathcal{G}(\eta^-)), && (100) \\ &\leq (\zeta + \epsilon(1 + D)) \mathcal{G}(e). && (101) \end{aligned}$$

Letting $\mu = (\zeta + \epsilon(1 + D))$ completes this part of the proof. Notice that reasonable values for ϵ are obtained if D is not too large.

Now we utilize the second bullet. Let $\hat{c_1} = \frac{1}{c_1}$. The second bullet above and the *local saturation* condition yields

$$\begin{aligned} \mathcal{G}(e_h - \delta^-_{(h/2)}) &\leq \zeta \mathcal{G}(\eta^-) + \hat{c_1} \left(\mathcal{G}(\eta^-) + \mathcal{G}(\eta^+)\right) + \mathcal{G}(\eta^+) + \mathcal{G}(\mathcal{H}^+) \\ &\leq (\zeta + \hat{c_1}) \mathcal{G}(\eta^-) + (1 + \hat{c_1}) \mathcal{G}(\eta^+) + \mathcal{G}(\mathcal{H}^+). \end{aligned} \tag{102}$$

Finally, applying equation (96) yields

$$\begin{aligned}\mathcal{G}(e_h - \delta^-_{(h/2)}) &\leq (\zeta + \hat{c_1})\mathcal{G}(\eta^-) + \epsilon(1 + \hat{c_1})(\mathcal{G}(\mathcal{H}^-) + \mathcal{G}(\eta^-)) \\ &\leq ((\zeta + \hat{c_1}) + \epsilon(1 + \hat{c_1}))\mathcal{G}(\eta^-) + \epsilon(1 + \hat{c_1})\mathcal{G}(\mathcal{H}^-) \\ &\leq ((\zeta + \hat{c_1}) + \epsilon(1 + \hat{c_1}))\left(\mathcal{G}(\eta^-) + \mathcal{G}(\mathcal{H}^-)\right) \\ &\leq ((\zeta + \hat{c_1}) + \epsilon(1 + \hat{c_1}))\mathcal{G}(e_h). \end{aligned} \tag{103}$$

Letting $\mu = ((\zeta + \hat{c_1}) + \epsilon(1 + \hat{c_1}))$ completes this part of the proof. In the end, whether local error in the refinement region is small or large relative to the harmonic error in that region, we guarantee convergence.

7. Conclusion

Under certain assumptions, we proved that using the local FOSLS functional as a local error indicator does not lead to strict refinement of elements exhibiting error that is large but locally irreducible. This, on its own, serves as an important result, but we further applied our findings to make an overall statement concerning guaranteed convergence of an adaptive local refinement scheme applied to a FOSLS discretization. To this end, if the marked elements encompass a certain percentage of the error and the *local saturation* property is satisfied, convergence is guaranteed.

REFERENCES

1. M. Berndt, T. Manteuffel, and S. McCormmick, *Local error estimates and adaptive refinement for first-order system least squares (FOSLS)*, E.T.N.A., 6 (1998), pp. 35–43.
2. S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer, New York, second ed., 2002.
3. Z. Cai, T. Manteuffel, and S. McCormick, *First-order system least squares for the stokes equations, with application to linear elasticity*, SIAM J. Numer. Anal., 34 (1997), pp. 1727–1741.
4. R. B. Guenther and J. W. Lee, *Partial Differential Equations of Mathematical Physics and Integral Equations*, Dover, New York, 1988.
5. J. Hunter and B. Nachtergaele, *Applied Analysis*, World Scientific, Singapore, 2005.
6. J. Nolting, *Efficiency-based Local Adaptive Refinement for FOSLS Finite Elements*, PhD thesis, University of Colorado, Applied Mathematics Department, 2008.
7. C. Schwab, *p- and hp-Finite Element Methods*, Clarendon Press, Oxford, 1998.
8. F. Zhang, *The Schur Complement and its Applications*, Springer, New York, 2005.