INTERNATIONAL JOURNAL OF
NUMERICAL ANALYSIS AND MODELING
Volume 4, Number 2, Pages 241–279

FOUNDATION OF FAST NON-LINEAR FINITE ELEMENT SOLVERS, PART II

PETER L. SHI

Abstract. The author establishes a finite element solver algorithm of optimal speed for a class of quasi-linear equations with large stiffness variations and oscillations. In particular, the algorithm can successfully handle soft inclusions of negative stiffness. Besides the convergence analysis, large number of numerical examples are presented.

Key Words. Finite elements, non-linear solver algorithm, optimal speed

1. Introduction

This is the first of a series of papers supplementing the long article of the author [11] which has established a general algorithmic architecture for solving nonlinear finite element models with linear speed. The focus here is to demonstrate a particular implementation of the methodology to handle the Galerkin formulation of linear and nonlinear finite element models with large stiffness variation and oscillation that frequently arise from composite materials. After a briefing on the general theory and algorithm, we center our discussion around two benchmark problems. The first is concerned with the elasto-plastic deformation of a membrane in which the Young's modulo in the elastic region greatly exceeds that in the plastic region, constituting large jumps in coefficients in unknown regions. The second case is concerned with soft inclusions typically seen in making a composite, in which the included soft material is distributed as mesoscale tiny blocks of much softer stiffness in the scale of $10^{-2} \sim 10^{-6}$ compared to the hard material matrix. In particular, we demonstrate the effectiveness of the algorithm in treating soft inclusions with negative stiffness, a challenging issue that has not been tackled in prior art. Large amount of numerical examples are demonstrated.

In this paper, the author only presents the method in a two dimensional setting. Its generalization to three dimensional domains requires more elaborated technicalities that deserve a separate discussion.

Let Ω a bounded polygonal domain in R^2. In order to deal with soft inclusions, we let Ω_1 and Ω_2 be sub-domains of Ω such that

$$\overline{\Omega} = \overline{\Omega}_1 \cup \overline{\Omega}_2, \quad \Omega_1 \cap \Omega_2 = \emptyset.$$

Received by the editors March 5, 2006.
2000 *Mathematics Subject Classification.* 65N30, 65J15.

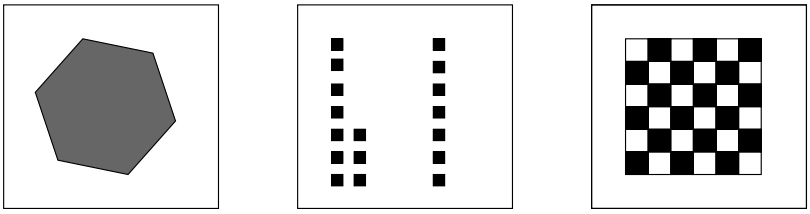


FIGURE 1. In (a) the shaded region is the unknown plastic region. In (b) and (c), the material on Ω_1 (black) is soft while the material on Ω_2 (white) is hard. The variation or oscillation in stiffness has risen to the extent that the standard algorithms are impeded in speed or accuracy.

We assume that Ω_1 is a union of small polygons that is not necessarily connected. Three examples of the domain are illustrated in Figure 1 that foreshadow the challenge we will face in the computation. In (a), the shaded area represents an unknown plastic region. In (b), the domain can be used to model material defects or random inclusions. Domain (c) is a familiar semi-periodic situation in composite material, often seen in the homogenization theory.

In order to avoid excessive technical details, we further simplify the partial differential operator to include only the principle part, given by

$$Lu = -\sum_{j=1}^{2} \big[a_j(x, \nabla u)\big]_{x_j}$$

where each a_j is a measurable function on $\Omega \times R^2$. While the dependence on the solution itself in a_j and lower order terms can also be considered, we will omit such complications. Throughout the paper, we will make the following assumptions on the coefficients of L.

(A0) For each $x \in \Omega$, $a_j(x, 0, 0) = 0$ for $j = 1, 2$.

(A1) For each $k = 1, 2$, there exists a constant $\alpha_k \geq 0$ such that for all $x \in \Omega_k$ and for all $\xi, \eta \in R^2$

$$\sum_{j=1}^{2} \big[a_j(x, \xi) - a_j(x, \eta)\big](\xi_j - \eta_j) \geq \alpha_k |\xi - \eta|^2,$$

where $|\cdot|$ denotes the Euclidean norm of R^2.

(A2) For each $k = 1, 2$, there exists a constant β_k such that for all $x \in \Omega_k$ and for all $\xi, \eta \in R^2$

$$\sum_{j=1}^{2} |a_j(x, \xi) - a_j(x, \eta)| \leq \beta_k |\xi - \eta|.$$

The sharp jumps in the stiffness coefficients are not explicitly expressed in the assumptions (A1)-(A2), but rather, embedded as a special case. In the event that

$$\alpha_1 = \delta \alpha_2, \quad \beta_1 = \delta \beta_2 \tag{1.1}$$

for a sufficiently small δ, such situations will occur. A typical range of δ can be $10^{-2} \sim 10^{-6}$ for example.

The oscillatory assumption on the coefficients a_j is embedded in the construction of the sub-domain Ω_1, which later will be discussed with greater detail.

The speed of a solver is defined as the number of float point operations needed to achieve the inequality

$$||u_h - u_c|| < \epsilon. \tag{1.2}$$

Here u_h denotes the nodal values of the exact solution to the finite element model, u_c denotes the computed approximation to u_h, ϵ denotes the the prescribed error bound, and $\|\cdot\|$ is the standard Euclidean norm of the nodal values. The optimal speed referenced above is given by $c|\ln\epsilon|N_d$, where N_d is the total degrees of freedom in the system, and c is a positive constant independent of N_d.

When (1.1) is present coupled with high frequency oscillations as illustrated in (c) of Figure 1, the Galerkin method for the boundary value problem requires a small mesh size in order to resolve the solution to a satisfactory scale, which in turn forces a large number of unknowns in the discrete system. In this paper, the author demonstrates that the linear speed of the solver algorithm when aided by the power of modern computer hardware will significantly reduce the time cost associated with the small mesh size without sacrificing the accuracy and the robustness of the algorithm. This further allows the author to unveil new computational results on a low end laptop which have been otherwise difficult to achieve even on high end computing equipments. In particular, author is able to show some interesting computational results on composite materials with inclusions of negative stiffness.

In response to the need of reducing the computational cost associated with large stiffness variation and oscillation, multi-scale modeling has become a new frontier. On the computational level, various multi-scale finite element approximations have also been introduced [6][8][2][1]. These approaches use a grid coarser than what is required for the fine scale resolution coupled with problem-dependent basis functions or built-in structures at finer level. For non-linear problems, more sophisticated mapping is constructed to replace the basis functions [7]. The resulting approximation error is typically given in the form

$$\|u - u_{h,\epsilon}\| = O(\epsilon + h + \sqrt{\epsilon/h}) \tag{1.3}$$

and its close variants. Here u is the homogenized solution, $u_{h,\epsilon}$ is the theoretical solution of the multi-scale finite element model, ϵ is the typical length of the fine scale (for example, the size of the black-white squares in Figure 1), and h is the coarse grid size in the multi-scale model. For a recent study of the performance of multi-scale finite element models we refer the reader to [1].

We point out that multi-scale finite element models are not complete solvers — $u_{h,\epsilon}$ remains to be solved. When ϵ is sufficiently small, say 10^{-6}, while the physical domain Ω is relatively large, the multi-scale finite element model itself may still have millions of unknowns. It is also unclear how the large magnitude in the stiffness variation is resolved in the in the estimate (1.3). These difficulties when further enhanced by the nonlinearity of the equations are yet to be adequately addressed at the computational level. In addition, if no periodic structure is present, it is unclear if (1.3) is valid.

Numerical challenges caused by large jumps in coefficients have been addressed independent of the homogenization theory and multi-scale modeling. The balancing domain decomposition method [10] and the references therein, also known as BDD, reduces the original Galerkin formulation to solving a problem on the interface across the sub-domains. Its feasibility depends critically on the efficiency of

the interface solver, and ultimately on the conditioning of the interface operator. The BDD algorithm uses the Neumann-Neumann preconditioner to condition the interface operator, which in turn requires solving a coarser grid problem to guarantee that the Neumann problems on sub-domains are consistent. The bulk of success of BDD is restricted to linear problems.

The reader must not view the algorithm and the spirit of the current paper as a competing device against multi-scale finite element models or homogenization theory. Instead, it is a compliment to these established mechanisms. For example, with moderate modifications, the algorithm can also serve as a linear speed solver for multi-scale finite element models. While this is highly desirable, it will not be addressed in the current paper.

2. A List of Notations

Group 1: Interpolation Spaces

$H_h^1(E_k)$	.....	Local interpolation space on kth element E_k
Π_h	.....	Direct product of local interpolation spaces
V_h	.....	Standard interpolation space for Galerkin method

Group 2: Spaces on Sub-nodes

$\mathscr{B}$	.....	The set of all sub-nodes
$\Im$	.....	The topology on $\mathscr{B}$
$\mathscr{E}_k$	.....	kth elemental construct (sub-nodes on E_k)
μ	.....	The standard generic counting measure
ν	.....	A weighted measure on $\mathscr{B}$
$L^2(\mathscr{B})$	.....	Topological finite element spaces
$\mathrm{C}[\mathscr{B}]$	.....	The continuous subspace of $L^2(\mathscr{B})$
P	.....	The orthogonal projection from $L^2(\mathscr{B})$ onto $\mathrm{C}[\mathscr{B}]$
$\mathrm{C}_\omega[\mathscr{B}]$	.....	The weighted space of continuity
P_ω	.....	The orthogonal projection from $L^2(\mathscr{B})$ onto $\mathrm{C}_\omega[\mathscr{B}]$
$\mathrm{Cnst}[\mathscr{B}]$	.....	A piece-wise constant space — the kernel space
$\mathscr{K}$	.....	The orthogonal projection from $L^2(\mathscr{B})$ onto $\mathrm{Cnst}[\mathscr{B}]$
$\mathrm{C}_q[\mathscr{B}]$	.....	The quotient space, equal to $(I-\mathscr{K})\mathrm{C}_\omega[\mathscr{E}]$
$\mathscr{P}_q$	.....	The orthogonal projection from $L^2(\mathscr{B})$ onto $\mathrm{C}_q[\mathscr{B}]$
$\mathscr{I}$	.....	The natural isomorphism from $L^2(\mathscr{B})$ onto Π_h

Group 3: Spaces on Abstract Graph

$\mathscr{E}$	.....	(Directed) abstract graph of sub-nodes
$\mathscr{G}$	.....	(Non-directed) abstract graph of sub-nodes
$[\mathscr{E}_k]$	.....	Local edges of $\mathscr{E}_k$
$[\overline{\mathscr{E}}_k]$	.....	Closure of local edges of $\mathscr{E}_k$
$[\mathscr{E}]$	.....	The set of all local edges, disjoint union of all $[\mathscr{E}_k]$
$[\mathscr{G}]$	.....	The set of all closures of local edges, union of all $[\overline{\mathscr{E}}_k]$
$\overline{\mu}$	.....	A discrete measure on $[\mathscr{E}]$
$L^2[\mathscr{E}]$	.....	The L^2-space of functions defined on $[\mathscr{E}]$
$\mathrm{C}[\mathscr{E}]$	.....	The continuous subspace of $L^2[\mathscr{E}]$
$\mathrm{C}_\omega[\mathscr{E}]$	.....	The weighted continuous subspace of $L^2[\mathscr{E}]$
$\mathrm{C}^0_\omega[\mathscr{E}]$	.....	The weighted space of zero mean
$\mathscr{P}_0$	.....	The orthogonal projection from $L^2[\mathscr{E}]$ onto $\mathrm{C}^0_\omega[\mathscr{E}]$
$\mathrm{C}^0[\mathscr{E}]$	.....	The (non-weighted) space of zero mean

$R^0_\omega[\mathscr{E}]$	.....	An isomorphic image of $\mathrm{C}^0_\omega[\mathscr{E}]$ into $\mathrm{C}[\mathscr{E}]$
$E^0_\omega[\mathscr{E}]$	.....	Orthogonal projection from $\mathrm{C}[\mathscr{E}]$ onto $R^0_\omega[\mathscr{E}]$
$L^2[\mathscr{G}]$	.....	The set of functions defined on $[\mathscr{G}]$, isometric to $\mathrm{C}[\mathscr{E}]$
$R^0_\omega[\mathscr{G}]$	.....	An isomorphic image of $\mathrm{C}^0_\omega[\mathscr{E}]$ into $L^2[\mathscr{G}]$
$R^0[\mathscr{G}]$	.....	The (non-weighted) space of zero mean on $[\mathscr{G}]$.

The language used in the paper is a direct descent from the author's article [11] characterized by the use of topology on sub-nodes and discrete measures. This inevitably invokes notations that are non-standard in the numerical literature. These notations are not just a matter of style or formalism, without which it is almost impossible to present the algorithm at the current level of theoretical rigor and algorithmic clarity. Not only the spaces provide a concrete data structure for programming — they are naturally close to C structures — but also they serve as a framework for further generalization of the algorithm to higher dimensions.

3. The Galerkin Formulation

The shape functions used for the discussion of the algorithm in the current paper are from the iso-parametric family as illustrated in Figure 2. Although more

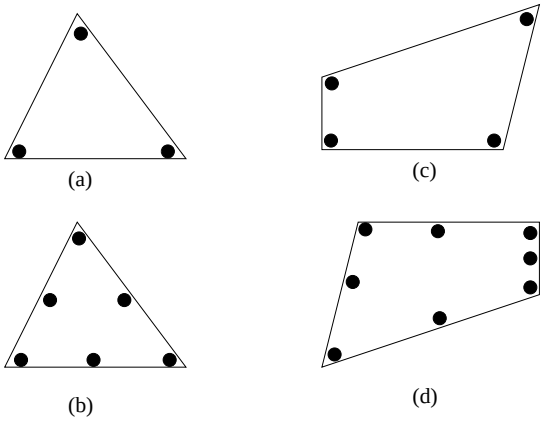


FIGURE 2. Examples of local degrees of freedom associated with shape functions. (a) Linear shape functions on a triangle (b) Quadratic shape functions on a triangle (c) Bi-linear shape functions on a quadrilateral (d) Serendipity shape functions on a quadrilateral.

complicated shape functions will further enhance the fine scale resolution, it will significantly increase the technical level of the presentation. Our goal is to keep the exposition of the main idea as simple as possible and refer the reader to [11] for generality.

We will only consider the homogeneous Neumann boundary condition, for which the interpolation space $V_h \subset H^1(\Omega)$ will be constructed accordingly. Throughout the rest of the discussion, h will denote the so-called typical element size. N will be the total number of elements in the mesh, and N_d will be the total degrees of freedom in the discrete system.

The weak formulation of the boundary value problem is defined as follows. Given $f_j \in L^2(\Omega)$, $j = 1, 2$, find $u \in H^1(\Omega)$ such that for all $v \in H^1(\Omega)$

$$\sum_{j=1}^{2} \int_\Omega a_j(x, \nabla u) v_{x_j} dx = \sum_{j=1}^{2} \int_\Omega f_j v_{x_j} dx. \tag{3.1}$$

Our task is to find $u_h \in V_h$ in a linear computational count such that

$$a(u_h, v_h) = \sum_{j=1}^{2} \int_\Omega f_j v_{hx_j} dx, \quad \forall v_h \in V_h. \tag{3.2}$$

Here $a(\cdot, \cdot)$ is the quasi-linear form defined by the left hand side of (3.1). Without loss of generality, we have eliminated the term

$$\int_\Omega f_0 v_h dx$$

from (3.1)-(3.2). This can be achieved by a relatively easy Z-matrix technique extensively described in [11], or by solving a discrete Laplacian.

4. Architecture of Discrete Spaces

We begin by making a general remark on the space environment in which a finite element solver algorithm should be analyzed. Except for algebraic multigrid method (AMG) that deliberately separates itself from the underlying Sobolev spaces describing the partial differential equations, most other algorithms carry out their algorithmic design and convergence analysis between V_h and the continuous Sobolev space setting. Here we must separate the two different notations between the convergence of the approximation scheme and the convergence of the solver algorithm. The latter refers to the estimation of $\|u_h - u_c\|$ as described in (1.2) in the Euclidean environment. Therefore it is no longer natural to analyze $\|u_h - u_c\|$ in the Sobolev space setting. However, the generic Euclidean space lacks the data structure to capture its relevance to finite elements. It is therefore necessary to restore the finite element structure in a Euclidean space, and this is the primary contents of this section, and is the foundation of the author's algorithmic design.

4.1. The Piece-wise Interpolation Space. We denote the elements in the mesh by E_k for $k = 1, 2, \ldots, N$, and denote the shape functions on E_k by

$$\varphi_l^k, \quad l = 1, 2, \ldots, m. \tag{4.1}$$

For the examples in Figure 2, $m = 3, 6, 4, 8$ respectively for the cases (a)-(d). Let

$$\Pi_h = \bigoplus_{k=1}^{N} H_h^1(E_k).$$

Here $H_h^1(E_k)$ is the span of the shape functions φ_l^k on the element E_k. Note that Π_h is no longer a subspace of $H^1(\Omega)$. However, $V_h \subset \Pi_h$ remains valid.

4.2. The Topological Space of Degrees of Freedom. The concept of a *local degrees of freedom* has been traditionally heuristic. Its topological treatment is the foundation for the design of the current algorithm. Let m be a fixed integer discussed in the above examples. We simply call the point set

$$\mathscr{E}_k = \{p_l^k,\ l = 1, 2, \ldots m\}$$

the *local degrees of freedom* on the element E_k. Each p_l^k must be understood as an abstract singleton whose content is unimportant. We call each p_l^k a *sub-node* and call each $\mathscr{E}_k$ an *elemental construct*. The set of all sub-nodes is denoted by $\mathscr{B}$. Thus

$$\mathscr{B} = \bigcup_{k=1}^{N} \mathscr{E}_k, \quad \mathscr{E}_k \cap \mathscr{E}_j = \emptyset \quad \text{if } k \neq j.$$

We call a topology $\Im$ on $\mathscr{B}$ a *conforming finite element topology* if it satisfies the following. For each $j = 1, 2, \ldots, N$

T1. $\partial_p \mathscr{E}_j \neq \emptyset$

T2. If $p, q \in \mathscr{E}_j$ with $p \neq q$, then $\overline{p} \cap \overline{q} = \emptyset$.

T3. If $p, q \in \mathscr{B}$ and $\overline{p} \cap \overline{q} \neq \emptyset$, then $\overline{p} = \overline{q}$.

Here $\partial_p \mathscr{E}_j = \mathscr{E}_j \setminus \mathrm{Int}(\mathscr{E}_j)$ denotes the so-called inner boundary[1] of $\mathscr{E}_j$ and the overline symbol denotes the closure. The actual construction of $\Im$ varies in concrete applications. Within the scope of the current paper, this will be a simple event.

Theorem 4.1. *For any topology on $\mathscr{B}$ that satisfies T3 the following properties hold.*

a. *Let $p \in \mathscr{B}$ and let $q \in \overline{p}$. Then $\overline{p} = \overline{q}$. They are both the smallest closed subset of $\mathscr{B}$ that contains p.*

b. *There exists a unique set of closed subsets of the form*

$$\{\overline{p}_j;\ p_j \in \mathscr{B},\ j = 1, 2, \ldots, M\}$$

such that

$$\mathscr{B} = \bigcup_{j=1}^{M} \overline{p}_j, \quad \overline{p}_j \cap \overline{p}_k = \emptyset, \quad i \neq k.$$

c. *Let $p \in \mathscr{B}$ and $q \in \overline{p}$. Then $\overline{p}$ is the smallest open subset of $\mathscr{B}$ that contains q.*

Theorem 4.1 allows us to model the notion of *global degrees of freedom* by the closure of a point in $\mathscr{B}$. Thus we call each $\overline{p}$ a *nodal construct* of $\mathscr{B}$. Accordingly, $\mathscr{B}$ is also a disjoint union of nodal constructs.

4.3. Topological Finite Element Space. The discrete space holding all possible programming data will be $L^2(\mathscr{B})$, the space of all functions defined on $\mathscr{B}$ equipped with the inner product

$$(u, v) = \int_{\mathscr{B}} uv \, d\mu. \tag{4.2}$$

Here μ is the standard *counting measure* on $\mathscr{B}$. At first glance, the space $L^2(\mathscr{B})$ is nothing but an Euclidean space R^n. However, a closer look reveals that the added topological and algebraic structure in $\mathscr{B}$ have united the local and global degrees of freedom, the nodal values of the interpolation and the connectivity of the mesh, all in one elegant setting, to form an ideal computational environment. We call $L^2(\mathscr{B})$ the topological finite element space.

[1]For the shape functions considered in the current paper, the inner boundary is empty

4.4. Continuity and Weighted Continuity. We denote by $\mathrm{C}[\mathscr{B}]$ the space of continuous functions defined on $\mathscr{B}$ under the topology $\Im$. In light of Theorem 4.1, it is easy to show the following.

Proposition 4.2. *Let $f \in L^2(\mathscr{B})$. Then $f \in C[\mathscr{B}]$ if and only if f is a constant on each nodal construct.*

Let $[p]_j$ for $j = 1, 2, \ldots, \sigma$ denote all nodal constructs in $\mathscr{B}$. Proposition 4.2 implies that a function in $\mathrm{C}[\mathscr{B}]$ has a unique representation

$$f = \sum_{j=1}^{\sigma} c_j \chi_{[p]_j}, \quad c_j \in R^1. \tag{4.3}$$

Here $\chi_{[p]_j}$ denotes the characteristic function of $[p]_j$. Hence the orthogonal projection from $L^2(\mathscr{B})$ onto $\mathrm{C}[\mathscr{B}]$ is simply given by

$$Pf = \sum_{j=1}^{\sigma} \frac{\chi_{[p]_j}}{\mu(\chi_{[p]_j})} \int_{\mathscr{B}} \chi_{[p]_j} f \, d\mu, \quad f \in L^2(\mathscr{B}). \tag{4.4}$$

Important to our discussion is a *weighted space of continuity*, denoted by $\mathrm{C}_\omega[\mathscr{B}]$, which is designed to compensate for the jumps in the operator coefficients while keeping the shape functions unchanged. More precisely, we decompose $\mathscr{B}$ into a disjoint union of two subsets. Let $\mathscr{B}_l$, for $l = 1, 2$, denote the union of all elemental constructs corresponding to elements E_k such that $\overline{E}_k \subset \overline{\Omega}_l$. We say that $f \in \mathrm{C}_\omega[\mathscr{B}]$ if

$$\omega f \in \mathrm{C}[\mathscr{B}], \quad \omega = \frac{\chi_{\mathscr{B}_1}}{\sqrt{\delta}} + \chi_{\mathscr{B}_2}. \tag{4.5}$$

Here δ is the same quantity that appeared in (1.1). It is clear that $\mathrm{C}_\omega[\mathscr{B}]$ is the image of $\mathrm{C}[\mathscr{B}]$ under the linear mapping defined by multiplication by ω^{-1}. Note that

$$\omega^{-1} = \sqrt{\delta}\chi_{\mathscr{B}_1} + \chi_{\mathscr{B}_2}.$$

Proposition 4.3. *Let ν be the discrete measure on $\mathscr{B}$ defined by $d\nu = \omega^{-2} d\mu$. Then the orthogonal projection from $L^2(\mathscr{B})$ onto $C_\omega[\mathscr{B}]$ is given by*

$$P_\omega f \quad = \quad \omega^{-1} \sum_{j=1}^{\sigma} \frac{\chi_{[p]_j}}{\nu(\chi_{[p]_j})} \int_{\mathscr{B}} \chi_{[p]_j} \omega f \, d\nu, \quad f \in L^2(\mathscr{B}) \tag{4.6}$$

Proof. Let $f \in L^2(\mathscr{B})$ and $u = P_\omega f$. Then $\overline{u} = \omega u \in \mathrm{C}[\mathscr{B}]$ and the following equalities hold.

$$\int_{\mathscr{B}} |u - f|^2 \, d\mu = \int_{\mathscr{B}} \omega^{-2} |\omega u - \omega f|^2 \, d\mu = \int_{\mathscr{B}} |\overline{u} - \omega f|^2 \, d\nu.$$

On the other hand,

$$\int_{\mathscr{B}} |u - f|^2 \, d\mu = \min_{v \in \mathrm{C}[\mathscr{B}]} \int_{\mathscr{B}} \omega^{-2} |v - \omega f|^2 \, d\mu = \min_{v \in \mathrm{C}[\mathscr{B}]} \int_{\mathscr{B}} |v - \omega f|^2 \, d\nu.$$

Hence $\overline{u}$ is the orthogonal projection of ωf from $L^2(\mathscr{B})$ onto $\mathrm{C}[\mathscr{B}]$, provided that the measure on $\mathscr{B}$ was given by ν. In light of (4.4), it follows that

$$\overline{u} = \sum_{j=1}^{\sigma} \frac{\chi_{[p]_j}}{\nu(\chi_{[p]_j})} \int_{\mathscr{B}} \chi_{[p]_j} \omega f \, d\nu.$$

This proves (4.6). □

4.5. The Natural Isomorphism. In this section we make the first link between the topological finite element space $L^2(\mathscr{B})$ and the piece-wise interpolation space

$$\Pi_h = \bigoplus_{k=1}^{N} H_h^1(E_k).$$

This is done via the isomorphism $\mathscr{I} : L^2(\mathscr{B}) \to \Pi_h$ defined by

$$\mathscr{I} : f \to \sum_{k=1}^{N} \sum_{l=0}^{m-1} f(p_l^k)\varphi_l^k, \quad f \in L^2(\mathscr{B}). \tag{4.7}$$

Some basic properties of $\mathscr{I}$ are given in the following.

Proposition 4.4. *Let χ_{kl} be the characteristic function of the l^{th} sub-node in $\mathscr{E}_k$. Let χ_k be the characteristic function of $\mathscr{E}_k$. The following properties hold.*

$$\begin{cases} \mathscr{I}\chi_{kl} & = & \varphi_l^k \\ \mathscr{I}\chi_k & = & \chi_{E_k} \\ \mathscr{I}(C[\mathscr{B}]) & = & V_h \end{cases} \tag{4.8}$$

Proof. These properties follow directly from (4.7) and the definitions of the quantities involved. □

Proposition 4.5. *Suppose that $g \in L^2(\mathscr{B})$ is a constant on each elemental construct. Then*

$$\mathscr{I}(gf) = (\mathscr{I}g)(\mathscr{I}f), \quad f \in L^2(\mathscr{B}) \tag{4.9}$$

In particular, $\mathscr{I}(C_\omega[\mathscr{B}]) = (\mathscr{I}\omega^{-1})V_h$.

Proof. To prove (4.9), we observe that

$$g = \sum_{k=1}^{N} c_k\chi_k, \quad gf = \sum_{k=1}^{N}\sum_{l=1}^{m} c_k f(p_l^k)\chi_l^k, \quad \mathscr{I}g = \sum_{k=1}^{N} c_k\chi_{E_k}.$$

Hence

$$\mathscr{I}(gf) = \sum_{k=1}^{N}\left\{c_k \sum_{l=1}^{m} f(p_l^k)\varphi_l^k\right\} = (\mathscr{I}g)(\mathscr{I}f).$$

Note that $\omega^{-1} = \sqrt{\delta}\chi_{\mathscr{B}_1} + \chi_{\mathscr{B}_2}$. Thus $\mathscr{I}(\mathrm{C}_\omega[\mathscr{B}]) = (\mathscr{I}\omega^{-1})V_h$ following (4.9) and the definitions of the spaces involved. □

Finally, in order to handle the homogeneous pure Neumann boundary condition, we define $\mathrm{C}_\omega^0[\mathscr{B}]$ as the set of all functions u in $\mathrm{C}_\omega[\mathscr{B}]$ such that

$$\int_{\mathscr{B}} \omega u \, d\mu = 0. \tag{4.10}$$

4.6. The Quotient Space. In order to implement element-wise conditioning in the general spirit described in [11], it is necessary to introduce the piece-wise constant space $\mathrm{Cnst}[\mathscr{B}]$, which is the span of characteristic functions of all elemental constructs. Following the argument in [11], it is easy to prove that $\mathrm{C}_\omega[\mathscr{B}] \cap \mathrm{Cnst}[\mathscr{B}] = \mathrm{span}\{\omega\}$. Consequently

$$\mathrm{C}^0_\omega[\mathscr{B}] \cap \mathrm{Cnst}[\mathscr{B}] = \{0\}. \tag{4.11}$$

Let $\mathscr{K}$ be the orthogonal projection from $L^2(\mathscr{B})$ onto $\mathrm{Cnst}[\mathscr{B}]$ and let $\mathrm{C}_q[\mathscr{B}]$ denote the quotient space

$$\mathrm{C}_q[\mathscr{B}] = (I - \mathscr{K})\mathrm{C}^0_\omega[\mathscr{B}]. \tag{4.12}$$

In light of (4.11), it follows that $I - \mathscr{K}$ is a one-to-one mapping from $\mathrm{C}^0_\omega[\mathscr{B}]$ onto $\mathrm{C}_q[\mathscr{B}]$. We will denote the orthogonal projection from $L^2(\mathscr{B})$ onto $\mathrm{C}_q[\mathscr{B}]$ by $\mathscr{P}_q$.

There is a simple but very useful relation between $\mathrm{C}_q[\mathscr{B}]$ and the non-weighted quotient space $(I - \mathscr{K})\mathrm{C}[\mathscr{B}]$, which we summarize as follows.

Proposition 4.6. *A function $f \in C_q[\mathscr{B}]$ if and only if $f = \omega^{-1}g$ for some $g \in (I - \mathscr{K})\,C[\mathscr{B}]$.*

Next we characterize the space $\mathrm{C}_q[\mathscr{B}]$ in a more favorable form for computation. For this, it is useful to construct a graph $\mathscr{G}$ in the topological structure

$$\mathscr{B} \times \Im \times \{\mathscr{E}_j;\ j = 1, 2, \ldots, N\}.$$

The resulted abstract graph $\mathscr{G}$ has richer contents than the mesh-induced graph because the so-called vertexes and edges in $\mathscr{G}$ are no longer just abstract singletons — they are subsets of $\mathscr{B}$. The additional structures within each vertex and edge in $\mathscr{G}$ will later be used to characterize $\mathrm{C}_q[\mathscr{B}]$.

As is always the case in graph theory, terminologies have become an excessive burden. Unfortunately, the author does not have a mechanism to avoid these "excessiveness". After all, the finite element structure is a elaborated graph[2]. Given two different elemental constructs $\mathscr{E}_i$ and $\mathscr{E}_j$, we define

$$e_{ij} = \overline{\mathscr{E}}_i \cap \overline{\mathscr{E}}_j \tag{4.13}$$

as the interior edge between $\mathscr{E}_i$ and $\mathscr{E}_j$ provided that e_{ij} consists of at least two nodal constructs. The rigorous definition of boundary edges are more involved in the general setting. However, in the restricted cases considered in the current paper, the boundary edges of $\mathscr{G}$ are self-explanatory (see Figure 3). We define the *degree of a nodal construct* as the number of abstract edges to which it belongs. We say that a nodal construct in $\mathscr{B}$ is a *master nodal construct* if its degree is greater than 1. Otherwise we call it a *slave nodal construct*. The topology $\Im$ considered in the current paper is of two dimensional simple *Lagrangian type*, which has the following simple properties.

a. Each abstract edge contains exactly two master nodal constructs.
b. Two different sub-nodes in a nodal construct belong to two different elemental constructs respectively.

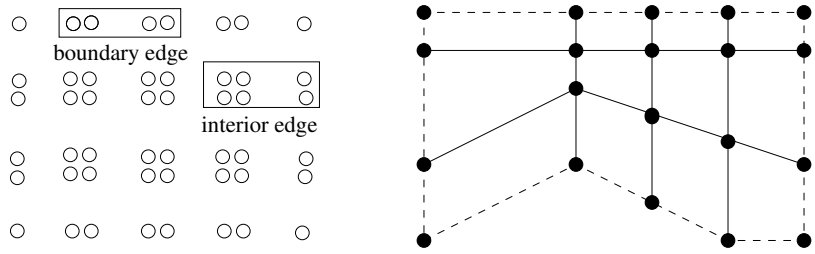
In light of the first property, we also deploy the more pronounced notation e_{ijkl} for e_{ij} when the two master nodal constructs contained in e_{ij} are indexed by k and l.

[2]In higher dimensions, algebraic topology will be naturally involved as well

At this point, we have introduced a graph $\mathscr{G}$ in a topological finite element space. The conversion into the standard graph terminologies is given as follows.

elemental construct	.....	face
e_{ijkl}	.....	edge
master nodal construct	.....	vertex

Note that slave nodal constructs are not included in the graph objects for the moment in order to simplify the graph structure. More importantly, the graph $\mathscr{G}$



(a) Abstract graph of sub–nodes (b) Original mesh of the domain

FIGURE 3. The isommetry between the abstract graph of sub-nodes (a) and the finite element mesh on Ω (b). In (a) each cluster of sub-nodes is a nodal construct, also defined as a vertex in $\mathscr{G}$. Edges in $\mathscr{G}$ are subsets of $\mathscr{B}$. The usual line segments connecting vertexes can be eliminated.

is isometric to the finite element mesh while being free from the individual shape of the elements. This will lead to significant advantage for the computation.

We also need a finer concept which we call *terminal construct* of an edge. A pair of sub-nodes $\{p, q\}$ is called a terminal construct of the edge e_{ijkl} if

$$p = \mathscr{E}_i \cap N_k,\ q = \mathscr{E}_j \cap N_k, \quad \text{or} \quad p = \mathscr{E}_i \cap N_l,\ q = \mathscr{E}_j \cap N_l.$$

Here N_k and N_l are the two nodal constructs contained in e_{ijkl} in accordance of (4.13). If N_k (N_l respectively) is a master nodal construct, then we say that

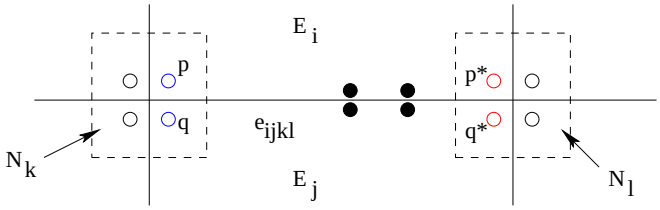


FIGURE 4. Master and slave terminal constructs of the abstract edge e_{ijkl}. Each pair of blue and red circles is a master terminal construct. Each pair of black circles is a slave terminal construct.

$\{p, q\}$ ($\{p^*, q^*\}$ respectively) is a master terminal construct, otherwise it is called a slave terminal construct. Figure 4 illustrates the typical terminal constructs of an abstract edge.

The consistency is a local criterion we will use to characterize the space $\mathrm{C}_q[\mathscr{B}]$. Let e_{ijkl} be a given edge in $\mathscr{G}$ and let $\{p, q\}$ and $\{p^*, q^*\}$ be the pair of terminal constructs of e_{ijkl}. We say that $f \in L^2(\mathscr{B})$ is *consistent on* e_{ijkl} if

$$\omega(p)f(p) - \omega(q)f(q) = \omega(p^*)f(p^*) - \omega(q^*)f(q^*). \tag{4.14}$$

We say that f is consistent on $\mathscr{B}$ if it is consistent on all edges of $\mathscr{G}$.

Theorem 4.7. *Let $f \in L^2(\mathscr{B})$. Then $f \in C_\omega[\mathscr{B}] + Cnst[\mathscr{B}]$ if and only if f is consistent on $\mathscr{B}$.*

Proof. The necessity follows from a direct verification of (4.14) by using the definition of $\mathrm{C}_\omega[\mathscr{B}]+\mathrm{Cnst}[\mathscr{B}]$. To prove the sufficiency, we consider the function $F = \omega f$. Then (4.14) implies that

$$F(p) - F(q) = F(p^*) - F(q^*). \tag{4.15}$$

As a direct application of Theorem 8.1 of [11] and (4.15), it follows that there exists $c \in \mathrm{Cnst}[\mathscr{B}]$ such that the values of c on $\mathscr{E}_i$ and $\mathscr{E}_j$ are given by c_i and c_j respectively, and

$$F(p) - F(q) = F(p^*) - F(q^*) = c_i - c_j.$$

Rearranging the above equality we obtain

$$F(p) - c_i = F(q) - c_j, \quad F(p^*) - c_i = F(q^*) - c_j. \tag{4.16}$$

This states that the function $F - c$ is a constant on each terminal construct of an edge, which in turn implies that $F - c \in \mathrm{C}[\mathscr{B}]$ when all edges in $\mathscr{G}$ are taken into account. Hence, by writing

$$f = \omega^{-1}(F - c) + \omega^{-1}c$$

we obtain $f \in \mathrm{C}_\omega[\mathscr{B}] + \mathrm{Cnst}[\mathscr{B}]$. This completes the proof. □

The equations in (4.16) suggest a convenient solution procedure for the determination of c. We refer the reader to [11] for detail.

4.7. The Space of Zero Mean. It turns out that computing $\mathscr{P}_q$ can be most effectively carried out in a space defined on edges of $\mathscr{G}$ when it is viewed as a bi-directional graph. This space is denoted by $L^2[\mathscr{E}]$, which we describe in the following.

Recall (4.13), in which the edge between two neighboring elemental constructs $\mathscr{E}_k$ and $\mathscr{E}_l$ is defined by $e_{kl} = \overline{\mathscr{E}}_k \cap \overline{\mathscr{E}}_l$. This further induces the notion of local edges of each $\mathscr{E}_k$ defined by $e_{kl} \cap \mathscr{E}_k,\ e_{kl} \neq \emptyset$. In order to ease the exposition, we will assume that $\mathscr{G}$ is isometric to a quadrilateral mesh with bi-linear iso-parametric shape functions[3]. In this case, $\mathscr{G}$ is *bipartite* in terms of faces. That is, we can associate with each elemental construct a unique $\pm$ sign such that neighboring elemental constructs have a different sign. We denote the four different local edges of an elemental construct $\mathscr{E}_k$ by

$$[\mathscr{E}_k] = \{e_{kdn},\ e_{krt},\ e_{kup},\ e_{klf}\}$$

in such a way that if $\mathscr{E}_l$ is a neighboring elemental construct of $\mathscr{E}_k$ then one and only one of the following holds.

$$\overline{e}_{kdn} = \overline{e}_{lup},\ \ \overline{e}_{klf} = \overline{e}_{lrt},\ \ \overline{e}_{ldn} = \overline{e}_{kup},\ \ \overline{e}_{llf} = \overline{e}_{krt}.$$

[3]This assumption is made only for the ease of exposition. Following the presentation is [11], the entire theory can be carried out in terms of the topology on $\mathscr{B}$ without reference to the mesh.

Here the over-line symbol denotes the topological closure. While each local edge is a set of sub-nodes, whereby topological closure applies, it must also be regarded as a singleton when the context is required. Parallel to the local edges, we also need the so-called local edge closures

$$[\overline{\mathscr{E}}_k] = \{\overline{e}_{kdn},\ \overline{e}_{krt},\ \overline{e}_{kup},\ \overline{e}_{klf}\}$$

to represent the edges of the graph $\mathscr{G}$. This notation emphasizes the aspect that each element in $[\overline{\mathscr{E}}_k]$ belongs to the topological closure of $\mathscr{E}_k$. It is easy to see that any element of $[\overline{\mathscr{E}}_k]$ can be uniquely written in the form

$$e_{kl} = \overline{\mathscr{E}}_k \cap \overline{\mathscr{E}}_l, \quad \text{for some } l.$$

At this point, we have implicitly introduced a an incomplete directed graph $\mathscr{E}$ associated with the graph $\mathscr{G}$ as illustrated in Figure 5. The exceptions are the boundary edges which remain non-directional. Let

$$[\mathscr{E}] = \bigcup_{k=1}^{N} [\mathscr{E}_k], \quad [\mathscr{G}] = \bigcup_{k=1}^{N} [\overline{\mathscr{E}}_k]$$

be the set of all local edges of $\mathscr{E}$ and the edges of $\mathscr{G}$ respectively. Given e and e_* in $[\mathscr{E}]$ with $e \neq e_*$, we say that $e \sim e_*$ if $\overline{e} = \overline{e}_*$. Hence the mapping $e \to e_*$ (with $e \sim e_*$) is one-to-one. Moreover, $(e_*)_* = e$. If e is a boundary edge, we define $e = e_*$. In general, we have $e = e_*$ if and only if e is a boundary edge. It is useful to classify the local edges in $[\mathscr{E}_k]$ via the disjoint union.

$$[\mathscr{E}_k] = \bigcup_{j=0}^{j=2} [\mathscr{E}_{kij}], \quad i = 1 \text{ or } 2.$$

For $j = 1, 2$ and $e \in [\mathscr{E}_k]$, we say that $e \in [\mathscr{E}_{kij}]$ if $e \in \mathscr{B}_i$ and $e_* \in \mathscr{B}_j$ with $e \neq e_*$. We say that $e \in [\mathscr{E}_{ki0}]$ if $e \in \mathscr{B}_i$ and $e = e_*$. Recall that $\mathscr{B}_1$ and $\mathscr{B}_2$ are unions of elemental constructs. Hence if $e \in [\mathscr{E}_{kij}]$ for some i and j, then all local edges of $[\mathscr{E}_k]$ belong to $[\mathscr{E}_{kim}]$ for some $m = 0, 1, 2$. Notice that the graph $\mathscr{G}$ can be re-generated by $\mathscr{E}$ via a pure topological procedure.

The space $L^2[\mathscr{E}]$ is the set of all functions defined on $[\mathscr{E}]$ equipped with the inner product

$$(f, g) = \int_{\mathscr{E}} fg \ d\overline{\mu}, \quad f, g \in L^2[\mathscr{E}].$$

Here $\overline{\mu}$ is a discrete measure on $[\mathscr{E}]$ such that

$$\overline{\mu}(e) = \begin{cases} \frac{1}{2} & \text{if } e \neq e^*, \\ 1 & \text{if } e = e^*. \end{cases}$$

The particular choice of $\overline{\mu}$ above will induce a useful isometry between a subspace of $L^2[\mathscr{E}]$ and $L^2[\mathscr{G}]$, the space of functions defined on $[\mathscr{G}]$, which we discuss later.

In a parallel manner, we also classify $[\overline{\mathscr{E}}_k]$ into the disjoint union

$$[\overline{\mathscr{E}}_k] = \bigcup_{j=0}^{j=2} [\overline{\mathscr{E}}_{kij}], \quad i = 1 \text{ or } 2.$$

An element is in $[\overline{\mathscr{E}}_{kij}]$ if and only if it is a topological closure of an element in $[\mathscr{E}_{kij}]$.

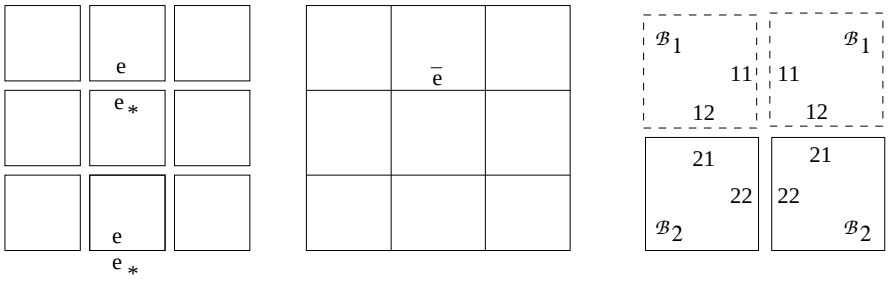


FIGURE 5. Illustration of local edges of the bi-directional graph $\mathscr{E}$, the edges of the graph $\mathscr{G}$, and the index theme for the classification of edges.

As is $L^2[\mathscr{E}]$ is parallel to $L^2(\mathscr{B})$, a space $\mathrm{C}_\omega[\mathscr{E}]$ we shall introduce is parallel to $\mathrm{C}_\omega[\mathscr{B}]$. It is the set of all functions in $L^2[\mathscr{E}]$ such that $f \in \mathrm{C}_\omega[\mathscr{E}]$ if and only if

$$\omega(e)f(e) = \omega(e_*)f(e_*) \quad \text{whenever } e \sim e_*. \tag{4.17}$$

Here $\omega(e)$ (respectively $\omega(e_*)$) stands for the value of ω on any sub-node in e (respectively e_*). The weighted space of zero mean, denoted by $\mathrm{C}^0_\omega[\mathscr{E}]$, consists of all functions in $f \in \mathrm{C}_\omega[\mathscr{E}]$ such that for each elemental construct $\mathscr{E}_k$

$$\sum_\alpha f(\alpha) = 0, \quad \alpha \text{ ranges over all local edges of } \mathscr{E}_k. \tag{4.18}$$

When $\omega = 1$, we simply denote $\mathrm{C}_\omega[\mathscr{E}]$ by $\mathrm{C}[\mathscr{E}]$, and denote $\mathrm{C}^0_\omega[\mathscr{E}]$ by $\mathrm{C}^0[\mathscr{E}]$, and call $\mathrm{C}^0[\mathscr{E}]$ the space of (non-weighted) zero mean. A much larger space $L^2_0[\mathscr{E}]$ consists of all functions in $L^2[\mathscr{E}]$ that satisfy (4.18).

There is a basis for the space $\mathrm{C}_\omega[\mathscr{E}]$ useful computing the orthogonal projection from $L^2[\mathscr{E}]$ onto $\mathrm{C}^0_\omega[\mathscr{E}]$, which we describe in the following. For each pair $\{e, e_*\} \subset [\mathscr{E}]$ with $e_* \sim e$, let

$$\xi_{ee_*} = \begin{cases} \omega^{-1}(e)\chi_e + \omega^{-1}(e_*)\chi_{e_*} & \text{if} \quad \omega(e) \neq \omega(e_*), \\ \chi_e + \chi_{e_*} & \text{if} \quad \omega(e) = \omega(e_*),\ e \neq e_*, \\ \chi_e & \text{if} \quad e = e_*. \end{cases} \tag{4.19}$$

It is clear that each ξ_{ee_*} satisfies (4.17). Here and in the subsequent discussion, it is understood that ξ_{ee_*} and ξ_{e_*e} will not appear simultaneously. In case that $\omega(e)) \neq \omega(e_*)$, it is assumed automatically that $e \in \mathscr{B}_1$ and $e_* \in \mathscr{B}_2$. From the definition of $\overline{\mu}$, it is also clear that

$$\|\xi_{ee_*}\|^2 = \begin{cases} \dfrac{\delta + 1}{2} & \text{if} \quad \omega(e) \neq \omega(e_*), \\ 1 & \text{otherwise.} \end{cases} \tag{4.20}$$

Theorem 4.8. *For any function $f \in C_\omega[\mathscr{E}]$, there exists a unique $u \in C[\mathscr{E}]$ such that*

$$f = \sum_{e \in [\mathscr{E}]} u(e)\xi_{ee_*}. \tag{4.21}$$

Conversely, any function in the form of (4.21) with $u \in C[\mathscr{E}]$ belongs to $C_\omega[\mathscr{E}]$. Given (4.21) with $u \in C[\mathscr{E}]$, then $f \in C^0_\omega[\mathscr{E}]$ if and only if the following holds. For

each $k = 1, 2, \ldots, N$, *if* $\mathscr{E}_k \subset \mathscr{B}_1$, *then*

$$\sum_{e\in[\mathscr{E}_{k10}]} u(e) + \sum_{e\in[\mathscr{E}_{k11}]} u(e) + \sum_{e\in[\mathscr{E}_{k12}]} \omega^{-1}(e)u(e) = 0; \tag{4.22}$$

if $\mathscr{E}_k \subset \mathscr{B}_2$, *then*

$$\sum_{e\in[\mathscr{E}_{k20}]} u(e) + \sum_{e\in[\mathscr{E}_{k21}]} u(e) + \sum_{e\in[\mathscr{E}_{k22}]} u(e) = 0. \tag{4.23}$$

Proof. It is easy to see that f in the form of (4.21) is another expression of (4.17). Given (4.21), (4.22)-(4.23) are exactly (4.18). This completes the proof. □

A useful point of view on Theorem 4.8 is that the mapping

$$\Phi : u \to f \tag{4.24}$$

defines an isomorphism from $\mathrm{C}[\mathscr{E}]$ onto $\mathrm{C}_\omega[\mathscr{E}]$. The space $\mathrm{C}^0_\omega[\mathscr{E}]$ is the image of a subspace $R^0_\omega[\mathscr{E}] \subset \mathrm{C}[\mathscr{E}]$ that satisfies (4.22)-(4.23). This subspace is important in subsequent discussions.

Corollary 4.9. *Suppose that for each* $k = 1, 2, \ldots, N$, *either* $[\mathscr{E}_{k12}] = [\mathscr{E}_k]$ *or* $[\mathscr{E}_{k12}] = \emptyset$, *and suppose that* $f \in C_\omega[\mathscr{E}]$ *is given by (4.21). Then* $f \in C^0_\omega[\mathscr{E}]$ *if and only if* $u \in C^0[\mathscr{E}]$, *the (non-weighted) space of zero mean. In other words,* $C^0[\mathscr{E}] = R^0_\omega[\mathscr{E}]$ *in this case.*

Proof. The scaling factor ω^{-1} in (4.22) will be canceled in the current situation. Hence (4.22) implies that

$$\sum_{e\in[\mathscr{E}_k]} u(e) = 0, \quad \mathscr{E}_k \subset \mathscr{B}_1.$$

This together with (4.23) implies that

$$\sum_{e\in[\mathscr{E}_k]} u(e) = 0, \quad \forall\ k = 1, 2, \ldots, N,$$

which states that the function u satisfies (4.18). □

We remark the above Corollary is extremely useful if each individual "soft inclusion" is isolated. It is also applicable to the checker-board pattern as illustrated in Figure 1.

The rest of this section is devoted to computing the orthogonal projection from $L^2[\mathscr{E}]$ onto $\mathrm{C}^0_\omega[\mathscr{E}]$. The most pivotal part is computing the orthogonal projection from $\mathrm{C}[\mathscr{E}]$ onto $R^0_\omega[\mathscr{E}]$, which we denote by E^0_ω. However, there are several smaller steps that must be resolved first. To this end, we first recall the space $L^2[\mathscr{G}]$, which is the space of all functions defined on the edges of the graph $\mathscr{G}$ equipped with the inner product

$$\int_{\mathscr{G}} fg\ d\mu, \quad f, g \in L^2[\mathscr{G}],$$

where μ is the standard counting measure on $[\mathscr{G}]$. It is easy to see that $\mathrm{C}[\mathscr{E}]$ as a subspace of $L^2[\mathscr{E}]$ is isometric to $L^2[\mathscr{G}]$. The isometric mapping between the two spaces, denoted by U, is simply given by

$$(Uf)(\overline{e}) = f(e), \quad \forall f \in \mathrm{C}[\mathscr{E}],\ \forall e \in [\mathscr{E}].$$

Accordingly, the isometric image of $R^0_\omega[\mathscr{E}]$ is a subspace of $L^2[\mathscr{G}]$, which we denote by $R^0_\omega[\mathscr{G}]$. In particular, we denote the isometric image of $\mathrm{C}^0[\mathscr{E}]$, the space of zero mean (non-weighted), by $R^0[\mathscr{G}]$, also called the space of zero mean.

Before proceeding further, it is helpful to summarize various spaces that we have directly or implicitly introduced. There are two basic environments, the sub-nodes environment and the graph environment.

$$
\begin{array}{ccccccc}
L^2(\mathscr{B}) & \supset & \mathrm{C}^0_\omega[\mathscr{B}] & \leftrightarrow & \mathrm{C}_q[\mathscr{B}] & & \text{if } \omega = 1 \\
\updownarrow & & & & \updownarrow & & \downarrow \\
L^2[\mathscr{E}] & \supset & \mathrm{C}_\omega[\mathscr{E}] & \supset & \mathrm{C}^0_\omega[\mathscr{E}] & \equiv & \mathrm{C}^0[\mathscr{E}] \\
 & & \updownarrow & & \updownarrow & & \updownarrow \\
 & & \mathrm{C}[\mathscr{E}] & \supset & R^0_\omega[\mathscr{E}] & \equiv & R^0[\mathscr{E}] \\
 & & \| & & \| & & \| \\
 & & L^2[\mathscr{G}] & \supset & R^0_\omega[\mathscr{G}] & \equiv & R^0[\mathscr{G}]
\end{array}
\tag{4.25}
$$

In the above illustration, the bi-directional arrows indicate isomorphism. The equal signs indicate isometry. The sign $\equiv$ means definition. Except for the relation between $\mathrm{C}_q[\mathscr{B}]$ and $\mathrm{C}^0_\omega[\mathscr{E}]$, which will be discussed in the next section, all have been discussed in detail.

Theorem 4.10. *Let $f \in L^2[\mathscr{E}]$ and let $f^0_\omega = \mathscr{P}_0 f$, the orthogonal projection from $L^2[\mathscr{E}]$ onto the space $C^0_\omega[\mathscr{E}]$ be expressed in the form*

$$f^0_\omega = \sum_{e \in [\mathscr{E}]} u(e)\xi_{ee_*}, \quad u \in R^0_\omega[\mathscr{E}].$$

Then u satisfies the generalized Wiener-Hopf equation

$$T_{E^0_\omega}(D_\omega)u = E^0_\omega F. \tag{4.26}$$

Here D_ω is the mapping from $C[\mathscr{E}]$ onto itself such that

$$D_\omega u(e) = D_\omega u(e_*) = \|\xi_{ee_*}\|^2 u(e), \quad \forall u \in C[\mathscr{E}],\ \forall e \in [\mathscr{E}].$$

$F \in C[\mathscr{E}]$ is defined by

$$F(e) = \int_{[\mathscr{E}]} f\xi_{ee_*}\, d\overline{\mu}, \quad e \in [\mathscr{E}].$$

Proof. Let $g \in \mathrm{C}^0_\omega[\mathscr{E}]$ be expressed in the form

$$g = \sum_{e \in [\mathscr{E}]} v(e')\xi_{e'e'_*}, \quad v \in R^0_\omega[\mathscr{E}].$$

Then the function u satisfies the variational equality

$$\int_{[\mathscr{E}]} f^0_\omega g\, d\overline{\mu} = \int_{[\mathscr{E}]} fg\, d\overline{\mu}, \quad \text{or} \quad \int_{[\mathscr{E}]} [D_\omega u]v\, d\overline{\mu} = \int_{[\mathscr{E}]} Fv\, d\overline{\mu}.$$

which is exactly (4.26). The theorem is proved. □

Lemma 4.11. *Suppose that V is subspace of R^n and let*

$$\omega_k = (a_{k1}, a_{12} \dots a_{kn}), \quad k = 1, 2 \dots m$$

be given vectors in R^n. Then

$$V = span\{\omega_k;\ k = 1, 2 \dots m\} \tag{4.27}$$

if and only if $V^\perp$ consists of all vectors $x = (x_1, x_2 \dots x_n)$ that satisfies the following linear constraints.

$$\sum_{j=1}^{n} a_{kj} x_j = 0, \quad k = 1, 2 \dots m. \tag{4.28}$$

Proof. Let A be the $n \times m$ matrix whose column vectors are given by (4.27). Then $y = \sum_{k=1}^{m} \xi_k \omega_k$ if and only if $y = A\xi$ for some $\xi = (\xi_1, \xi_2 \dots \xi_m)$. Hence the equivalence between (4.27) and (4.28) is the expression of the well-known formula $\mathrm{Im}(A)^\perp = \mathrm{Ker}(A^*)$. □

Theorem 4.12. *The space $R^0_\omega[\mathscr{G}]^\perp$, the orthogonal complement of $R^0_\omega[\mathscr{G}]$ with respect to $L^2[\mathscr{G}]$, is spanned by the basis functions*

$$\begin{cases} \xi_k &= \pm \left[\chi_{[\overline{\mathscr{E}}_k]} (1 - \chi_{[\overline{\mathscr{E}}_{k12}]}) + \sqrt{\delta} \chi_{[\overline{\mathscr{E}}_{k12}]} \right] & \text{if } [\mathscr{E}_{k12}] \neq [\mathscr{E}_k], \\ \xi_k &= \pm \chi_{[\overline{\mathscr{E}}_k]} & \text{if } [\mathscr{E}_{k12}] = [\mathscr{E}_k], \end{cases} \tag{4.29}$$

where $\pm$ is chosen according to the alternating sign assignment. Let $f \in C[\mathscr{E}]$ and let $F \in R^N$ be defined by

$$F_k = \int_{[\mathscr{G}]} (Uf) \xi_k \, d\mu, \quad k = 1, 2, \dots, N. \tag{4.30}$$

Let M be the $N \times N$ matrix with its ij-entries equal to

$$\int_{[\mathscr{G}]} \xi_i \xi_j \, d\mu, \quad i, j = 1, 2, \dots N. \tag{4.31}$$

Let $x \in R^N$ be the solution of the linear system of equations

$$Mx = F \tag{4.32}$$

Then $E^0_\omega f$, the orthogonal projection of f onto $R^0_\omega[\mathscr{E}]$, is given by

$$E^0_\omega f = f - \sum_{k=1}^{N} x_k U^{-1} \xi_k. \tag{4.33}$$

Proof. Recall that $R^0_\omega[\mathscr{E}]$ is described by (4.22)-(4.23). Applying Lemma 4.11 to (4.22)-(4.23), it follows that $U^{-1}\xi_k$ for $k = 1, 2, \dots N$ form a basis of the space $R^0_\omega[\mathscr{E}]^\perp$, the orthogonal complement being taken in the master space $\mathrm{C}[\mathscr{E}]$. Since U is an isometry between $\mathrm{C}[\mathscr{E}]$ and $L^2[\mathscr{G}]$, and $R^0_\omega[\mathscr{G}]^\perp$ is the isometric image of $R^0_\omega[\mathscr{E}]^\perp$, the rest of the result is self-explanatory. □

Corollary 4.13. *Suppose that for each $k = 1, 2, \dots, N$, either $[\mathscr{E}_{k12}] = [\mathscr{E}_k]$ or $[\mathscr{E}_{k12}] = \emptyset$. Then the matrix M in (4.32) is a 5-point finite difference Laplacian with the homogeneous Dirichlet boundary condition. The finite difference grid (excluding the boundary nodes) is graph-isometric to the graph obtained by identifying the center of each element in the mesh as a vertex, establishing the edges by connecting centers from neighboring elements.*

Proof. Under the assumptions of the corollary, (4.29) becomes

$$\xi_k = \pm \chi_{[\overline{\mathscr{E}}_k]}, \quad k = 1, 2, \dots, N.$$

The rest of the proof is self-explanatory. □

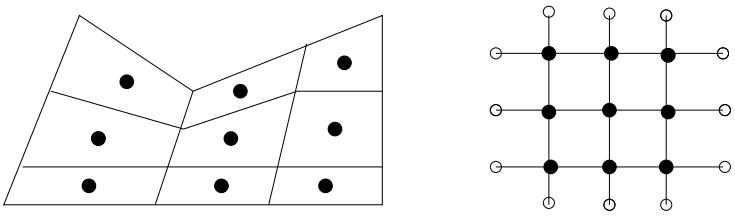


FIGURE 6. The resulting finite difference grid.

Figure 6 illustrates a situation described in Corollary 4.13. The general situation is subtle, where the matrix M has a Laplacian-like structure, but corresponding to the index k with $[\mathscr{E}_{k12}] \neq [\mathscr{E}_k]$ and $[\mathscr{E}_{k10}] = \emptyset$ (the interface between $\mathscr{B}_1$ and $\mathscr{B}_2$), the non-zero entries of the kth row of M can be given by

$$\text{diagonal: } 3+\delta, \quad \text{non-diagonal: } -1,\ -1,\ -1,\ -\sqrt{\delta}$$

which is not diagonally dominant. On the other hand, corresponding to an immediate neighbor of $\mathscr{E}_k$, the non-zero row entries of M can be given by

$$\text{diagonal: } 4, \quad \text{non-diagonal: } -1,\ -1,\ -1,\ -\sqrt{\delta}$$

which is strongly diagonally dominant. This new type of matrices is not direct discretizations of a standard partial differential operator. Although the general philosophy of algebraic multi-grid method [5][12][13][14] is likely applicable to the situation, special treatment must be given to justify its suitability.

4.8. The Filter of Local Constants. There is an intimate connection between the quotient space $\mathrm{C}_q[\mathscr{B}]$ and the space of zero mean $\mathrm{C}^0_\omega[\mathscr{E}]$. This connection can be revealed by a transform Ψ from $L^2(\mathscr{B})$ into $L^2_0[\mathscr{E}]$ which define in the following[4].

Let $y \in L^2(\mathscr{B})$. We denote by the vector $[y_{ksw},\ y_{kse},\ y_{kne},\ y_{knw}]$ the values of y on the elemental construct $\mathscr{E}_k$ and define the values of Ψy on the local edges of $\mathscr{E}_k$ by the matrix multiplication

$$\begin{bmatrix} e_{kdn} \\ e_{krt} \\ e_{kup} \\ e_{klf} \end{bmatrix} = \pm \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \\ -1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_{ksw} \\ y_{kse} \\ y_{kne} \\ y_{knw} \end{bmatrix} \tag{4.34}$$

Here the $\pm$ is chosen according to the alternating sign assignment of the faces. It is clear that the rows of the matrix in (4.34) sum to zero. This ensures that the mapping is into $L^2_0[\mathscr{E}]$. It is important to notice the pairing

$$\begin{cases} e_{kdn} &= \pm(y_{ksw} - y_{kse}) \\ e_{kup} &= \mp(y_{knw} - y_{kne}) \end{cases} \qquad \begin{cases} e_{klf} &= \pm(y_{knw} - y_{ksw}) \\ e_{krf} &= \mp(y_{kne} - y_{kse}) \end{cases}$$

which show that Ψ maps $\mathrm{C}_\omega[\mathscr{B}]$ into $\mathrm{C}^0_\omega[\mathscr{E}]$. The geometrical relevance associated with (4.34) is indicated in Figure 7. It is clear that Ψ can be written in the diagonal form

$$\Psi = \bigoplus_{k=1}^{N} \Psi_k,$$

[4]A similar filter called filter of continuity is introduced in [11] following a vertex oriented transform

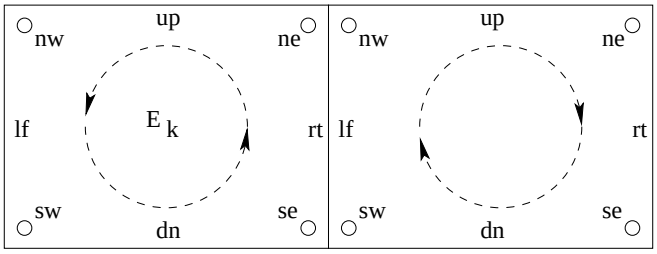


FIGURE 7. The local image of Ψ on an interior nodal space

where Ψ_k maps $L^2(\mathscr{E}_k)$ into the functions defined on the local edges of $\mathscr{E}_k$. We simply identify Ψ_k as the matrix in (4.34). While Ψ_k is not invertible, it is useful to introduce what we call the restricted inverse of Ψ_k, given by

$$\Psi_k^{-1} = \frac{\pm 1}{4}\begin{bmatrix} 3 & 2 & 1 & 0 \\ -1 & 2 & 1 & 0 \\ -1 & -2 & 1 & 0 \\ -1 & -2 & -3 & 0 \end{bmatrix} \tag{4.35}$$

and we let

$$\Psi^{-1} = \bigoplus_{k=1}^{N} \Psi_k^{-1}.$$

Straightforward calculation shows that

$$\Psi_k \Psi_k^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -1 & -1 & -1 & 0 \end{bmatrix} \tag{4.36}$$

$$\Psi_k^{-1} \Psi_k = \frac{1}{4}\begin{bmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \end{bmatrix} \tag{4.37}$$

$$\Psi_k^{-*} \Psi_k^{-1} = \frac{1}{4}\begin{bmatrix} 3 & 2 & 1 & 0 \\ 2 & 4 & 2 & 0 \\ 1 & 2 & 3 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{4.38}$$

While (4.36) and (4.37) are not the identity matrix, it is important to realize they are identity mappings on the space $\mathrm{C}_q[\mathscr{B}]$ and $\mathrm{C}^0_\omega[\mathscr{E}]$ respectively. In fact, for any vector $y = [y_1, y_2, y_3, y_4]$ it follows that

$$\sum_{k=1}^{4} y_k = 0 \quad \Rightarrow \quad \Psi_k \Psi_k^{-1} y = \Psi_k^{-1} \Psi_k y = y. \tag{4.39}$$

Similar to the definition of Ψ, we define

$$\Psi^{-1} = \bigoplus_{k=1}^{N} \Psi_k^{-1}. \tag{4.40}$$

Theorem 4.14. *Ψ maps $C_q[\mathscr{B}]$ onto $C^0_\omega[\mathscr{E}]$ in a one-to-one manner. Conversely, its restricted inverse Ψ^{-1} maps $C^0_\omega[\mathscr{E}]$ onto $C_q[\mathscr{B}]$.*

Proof. The discussion preceding the theorem has shown that Ψ maps $\mathrm{C}_q[\mathscr{B}]$ into $\mathrm{C}^0_\omega[\mathscr{E}]$. Using the matrix representation of Ψ_k^{-1}, we have

$$\begin{bmatrix} y_{ksw} \\ y_{kse} \\ y_{kne} \\ y_{knw} \end{bmatrix} = \frac{\pm 1}{4} \begin{bmatrix} 3 & 2 & 1 & 0 \\ -1 & 2 & 1 & 0 \\ -1 & -2 & 1 & 0 \\ -1 & -2 & -3 & 0 \end{bmatrix} \begin{bmatrix} e_{kdn} \\ e_{krt} \\ e_{kup} \\ e_{klf} \end{bmatrix} \tag{4.41}$$

Direct calculations from (4.41) give rise to

$$\begin{cases} y_{kse} - y_{ksw} &= \pm(-e_{kdn}) \\ y_{kne} - y_{kse} &= \pm(-e_{krt}) \\ y_{knw} - y_{kne} &= \pm(-e_{kup}) \\ y_{ksw} - y_{knw} &= \pm(e_{kdn} + ekrt + e_{kup}) = \pm(-e_{lf}). \end{cases} \tag{4.42}$$

The definition of $\mathrm{C}_\omega[\mathscr{E}]$ given in (4.17) is translated via (4.42) into the the consistency criterion as described in (4.14) . In light of Theorem 4.7, it follows that Ψ^{-1} maps $\mathrm{C}^0_\omega[\mathscr{E}]$ into $\mathrm{C}_\omega[\mathscr{B}] + \mathrm{Cnst}[\mathscr{B}]$. Moreover, the rows of the matrix in (4.41) sum to zero. Therefore, Ψ^{-1} maps $\mathrm{C}^0_\omega[\mathscr{E}]$ into $\mathrm{C}_q[\mathscr{B}]$.

The final property that $\Psi : \mathrm{C}_q[\mathscr{B}] \to \mathrm{C}^0_\omega[\mathscr{E}]$ is one-to-one and onto is now a consequence of (4.36)-(4.37). This completes the proof. □

Another surprising property of the mapping $\Psi^{-*}\Psi^{-1}$ is given in the following.

Theorem 4.15. *Let $x = [x_1, \ldots, x_4]$ and $y = [y_1, \ldots, y_4]$ satisfy*

$$\sum_{j=1}^{4} x_j = 0, \quad \sum_{j=1}^{4} y_j = 0. \tag{4.43}$$

Let M be the 3×3 matrix given by the non-zero entries of $\Psi^{-*}\Psi^{-1}$ in (4.38). For each fixed $m = 1, 2, 3, 4$, let $\overline{x}$ and $\overline{y}$ represent the three dimensional vectors obtained from x and y by deleting their mth components respectively. Then for each k

$$(\Psi_k^{-*}\Psi_k^{-1}x, y)_4 = (M\overline{x}, \overline{y})_3. \tag{4.44}$$

Here $(\cdot,\cdot)_4$ and $(\cdot,\cdot)_3$ represent the standard Euclidean inner product in R^4 and R^3 respectively.

Proof. We only prove the case when $m = 1$. Other cases are similar. In light of (4.43), we have

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} -y_2 - y_3 - y_4 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} -1 & -1 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} y_2 \\ y_3 \\ y_4 \end{bmatrix} \tag{4.45}$$

Similar properties hold for x. Let T denote the matrix that appeared in (4.45). Then it is straightforward to verify that $M = T^* M T$. This will establish (4.44). □

4.9. Computing the Orthogonal Projection $\mathscr{P}_q$. At this point, the only remaining issue is to compute $\mathscr{P}_q g$ for a given $g \in L^2(\mathscr{B})$, which we will resolve in this section.

By definition, $x = \mathscr{P}_q g$ is governed by the variational equation

$$x \in \mathrm{C}_q[\mathscr{B}] \ \ni \ \int_{\mathscr{B}} xy \, d\mu = \int_{\mathscr{B}} gy \, d\mu, \quad y \in \mathrm{C}_q[\mathscr{B}]. \tag{4.46}$$

By making the substitution $x = \Psi^{-1}\overline{x}$ and $y = \Psi^{-1}\overline{y}$ as discussed in §3.8, it is equivalent to solve

$$\overline{x} \in \mathrm{C}^0_\omega[\mathscr{E}] \ \ni \ \int_{[\mathscr{E}]} [\Psi^{-*}\Psi^{-1}\overline{x}]\overline{y} \, d\mu = \int_{[\mathscr{E}]} [\Psi^{-*}g]\overline{y} \, d\mu, \quad \overline{y} \in \mathrm{C}^0_\omega[\mathscr{E}]. \tag{4.47}$$

Next, we transform (4.47) into the setting of $L^2[\mathscr{E}]$ that requires the change from the counting measure $d\mu$ to the discrete measure $d\overline{\mu}$. This is necessary because the orthogonal projection $\mathscr{P}_0$ is defined in the space $L^2[\mathscr{E}]$, which is defined in terms of $d\overline{\mu}$. To this end, we let $\rho(e) = 1/\overline{\mu}(e)$ so that $d\mu = \rho d\overline{\mu}$. We obtain

$$\overline{x} \in \mathrm{C}^0_\omega[\mathscr{E}] \ \ni \ \int_{[\mathscr{E}]} \rho(\Psi^{-*}\Psi^{-1}\overline{x})\overline{y} \, d\overline{\mu} = \int_{[\mathscr{E}]} (\rho\Psi^{-*}g)\overline{y} \, d\overline{\mu}, \quad \overline{y} \in \mathrm{C}^0_\omega[\mathscr{E}], \tag{4.48}$$

which is exactly the generalized Wiener-Hopf equation

$$\overline{x} \in \mathrm{C}^0_\omega[\mathscr{E}] \ \ni \ T_{\mathscr{P}_0}(\rho\Psi^{-*}\Psi^{-1})\overline{x} = \mathscr{P}_0(\rho\Psi^{-*}g). \tag{4.49}$$

By a careful application of Theorem 4.15, we find that ρ does not change the characteristic nature of $\Psi^{-*}\Psi^{-1}$.

In light of (4.38)-(4.40) and the fact that $\mathrm{C}^0_\omega[\mathscr{E}] \subset L^2_0[\mathscr{E}]$, the calculation of the upper and the lower bounds of $\Psi^{-*}\Psi^{-1}$ on the space $\mathrm{C}^0_\omega[\mathscr{E}]$ can be done by maximizing and minimizing the functional

$$J(y) = (My, y), \quad \text{subject to } \ y_1^2 + y_2^2 + y_3^3 + (y_1 + y_2 + y_3)^2 = 1, \tag{4.50}$$

where M is the fixed 3×3 matrix taken from the non-zero entries of (4.38). By letting

$$w = Wy, \quad W = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix},$$

it is easy to see that the constraint in (4.50) becomes $\|w\|^2 = 1$. Direct calculation shows that

$$W^{-1} = \frac{1}{4}\begin{bmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{bmatrix}, \quad W^{-*}MW^{-1} = \frac{1}{4}\begin{bmatrix} 5 & -1 & -3 \\ -1 & 5 & -1 \\ -3 & -1 & 5 \end{bmatrix}.$$

Hence by calculating the eigen-values of $W^{-*}MW^{-1}$ we obtain the following inequalities. For all $x \in \mathrm{C}^0_\omega[\mathscr{E}]$ with $\|x\| = 1$

$$\int_{[\mathscr{E}]} [\Psi^{-*}\Psi^{-1}x]x \, d\mu \le 0.5,$$

$$\int_{[\mathscr{E}]} [\Psi^{-*}\Psi^{-1}x]x \, d\mu \ge 0.089903,$$

which ensure that the operator

$$\|I - 2T_{\mathscr{P}_0}(\Psi^{-*}\Psi^{-1})\| \le 0.83.$$

Thus, the generalized Wiener-Hopf equation (4.49) can be solved by the Banach contraction mapping principle

$$\overline{x}_{k+1} = \overline{x}_k - 2T_{\mathscr{P}_0}(\Psi^{-*}\Psi^{-1})\overline{x}_k + 2\mathscr{P}_0(\rho\Psi^{-*}g), \quad k = 0, 1, \ldots.$$

Recall that the evaluation of $\mathscr{P}_0$ has been discussed in Theorem 4.10, 4.12, and Corollary 4.13.

5. Generalized Wiener-Hopf Equations

This section marks the initial departure of methodology used in the current paper from those used in prior art. The Galerkin formulation (3.2) will be equivalently reformulated via a generalized Wiener-Hopf equation. The contents in previous sections will then serve as the basic data structure that allows us to solve the generalized Wiener-Hopf equation in linear computational count.

5.1. The Basic Reformulation. Under the traditional view of the Galerkin formulation (3.2), the domain of the quasi-linear form $a(\cdot,\cdot)$ is the product space $H^1(\Omega) \times H^1(\Omega)$, which contains $V_h \times V_h$. Associated with the inclusion $V_h \subset \Pi_h$ there exists a natural extension of $a(\cdot,\cdot)$ from $V_h \times V_h$ onto $\Pi_h \times \Pi_h$, which is outside of $H^1(\Omega) \times H^1(\Omega)$. This can be done by using the identity

$$a(u, v) = \sum_{k=1}^{N} \int_{E_k} \sum_{j=1}^{2} a_j(x, \nabla u) v_{x_j} dx. \tag{5.1}$$

Each term on the right hand side of (5.1) is naturally defined on $\Pi_h \times \Pi_h$. Hence the left hand side of (5.1) is accordingly extended. As usual, we also define a bounded (non-linear in general) operator $T : \Pi_h \to \Pi_h$ such that

$$a(u, v) = (Tu, v), \quad \forall u, v \in \Pi_h \tag{5.2}$$

where the right hand side of (5.2) is the inner product for $L^2(\Omega)$.

Along the same line, we also extend the right hand side of (3.2) as an inner product in Π_h. First, we extend the right hand side of (3.2) as a bounded functional acting on Π_h by viewing it as

$$\sum_{k=1}^{N} \int_{E_k} \big\{ f_1 v_{x_1} + f_2 v_{x_2} \big\} dx, \quad \forall v \in \Pi_h. \tag{5.3}$$

In turn, (5.3) can be represented in the inner product form

$$(f, v) = \sum_{k=1}^{N} \int_{E_k} \big\{ f_1 v_{x_1} + f_2 v_{x_2} \big\} dx, \quad \forall v \in \Pi_h. \tag{5.4}$$

for some $f \in \Pi_h$. At this point, the explicit form of f as an element in Π_h is unimportant.

The extension of $a(\cdot,\cdot)$ via (5.1)-(5.2) is nothing new from the programming point of view, which is exactly reverse procedure of the standard *assembling process* of the global stiffness operator. Unfortunately, this particular aspect was not sufficiently exploited in the past. Much effort has been directed to resolving the global stiffness operator after it is formed.

In light of (5.1)-(5.4), we are now in a position to rewrite the Galerkin discretization (3.2) in a variational form of Wiener-Hopf equations.

$$u \in V_h \ \ni \ (Tu, v) = (f, v), \quad \forall v \in V_h. \tag{5.5}$$

The difference between (5.5) and the Galerkin formulation (3.2) are more philosophical than technical. In (5.5), the space V_h is viewed as a subspace of Π_h instead of $H^1(\Omega)$. The operator T acts on the space Π_h even though the solution u is still to be found in the original V_h.

Let P_h be the orthogonal projection from Π_h onto V_h. Then we can write (5.5) as a generalized Wiener-Hopf equation

$$u \in \mathrm{Im}(P_h) \ \ni \ P_h T\big|_{\mathrm{Im}(P_h)} u = P_h f. \tag{5.6}$$

Neither of (5.5) nor (5.6) can be used as a final solution platform because so far we have done nothing about the conditioning of the system. It is not difficult to show that the condition of the system deteriorates as the dimension of V_h becomes large in exactly the same rate as the original Galerkin formulation (3.2).

For now, we turn our attention to the numerical evaluation of the operator T. Let the local quasi-linear form $\mathring{\mathrm{a}}_k(\cdot,\cdot)$ be defined by

$$\mathring{\mathrm{a}}_k(u,v) = \sum_{j=1}^{2} \int_{E_k} a_j(x, \nabla u) v_{x_j} dx \quad \forall u, v \in \Pi_h. \tag{5.7}$$

Let A_k map $L^2(\mathscr{E}_k) = \{f\chi_{\mathscr{E}_k};\ f \in L^2(\mathscr{B})\}$ into itself, and map $L^2(\mathscr{E}_j)$ to $\{0\}$ for $j \neq k$, defined by the following. Given $f \in L^2(\mathscr{E}_k)$, let $u = \sum_{l=1}^{m} f(p_l^k)\varphi_l^k$. Then $g = A_k f$ if and only if $g \in L^2(\mathscr{E}_k)$, and

$$g(p_l^k) = \sum_{j=1}^{2} \int_{E_k} a_j(x, \nabla u)[\varphi_l^k]_{x_j} dx, \quad l = 1, 2, \ldots, m \tag{5.8}$$

By definition, for all $x, \xi \in L^2(\mathscr{E}_k)$ with $u = \mathscr{J}(x)$ and $v = \mathscr{J}(\xi)$ we have

$$\begin{aligned} \mathring{\mathrm{a}}_k(u,v) &= \sum_{j=1}^{2} \int_{E_k} a_j(z, \nabla u) v_{x_j} dz \\ &= \sum_{j=1}^{2} \sum_{l=1}^{m} \xi(p_l^k) \int_{E_k} a_j(z, \nabla u)[\varphi_l^k]_{x_j} dz \\ &= \int_{\mathscr{E}_k} \xi A_k x \ d\mu. \end{aligned} \tag{5.9}$$

This is the relation between the local quasi-linear form $\mathring{\mathrm{a}}_k(\cdot,\cdot)$ and the local stiffness operator A_k.

We are now in a position to rewrite the Galerkin formulation (3.2) in the space $L^2(\mathscr{B})$.

Theorem 5.1. *Suppose that $f \in L^2(\mathscr{B})$ is defined by the following. For each $j = 1, 2, \ldots, m$ and $k = 1, 2 \ldots N$,*

$$f(p_l^k) = \int_{E_k} \left\{ f_1 \varphi_{jx_1}^k + f_2 \varphi_{jx_2}^k \right\} dx. \tag{5.10}$$

Here p_l^k is the l^{th} nodal construct in $\mathscr{E}_k$ and φ_j^k is the l^{th} shape function on E_k. Let

$$A = \bigoplus_{k=1}^{N} A_k. \tag{5.11}$$

Then $x \in C[\mathscr{B}]$ be the solution to the generalized Wiener-Hopf equation

$$x \in C[\mathscr{B}] \ni \int_{\mathscr{B}} yAx \; d\mu = \int_{\mathscr{B}} fy \; d\mu, \quad \forall y \in C[\mathscr{B}]. \tag{5.12}$$

Then $u = \mathscr{J} x$ is the solution to the Galerkin formulation (3.2).

We remark that the Wiener-Hopf operator $T_P(A) = PA\big|_{\mathrm{Im}(P)}$ have separated the differential operator from the mesh connectivity. The block diagonal structure of A is a reflection of the local property of the differential operator. The mesh connectivity is integrated into the projection operator P. We are no longer concerned with the traditional assembling process at the solver's level. Such process is embedded in the computation of P, which separates from the operator action of A. However, the operator $T_P(A)$ is still ill-conditioned.

5.2. The Outer Conditioning by Scaling. In the context considered in the current paper, there are two different sources responsible for the ill-conditioned state of (5.12). The first originates from (1.1), the large variation in coefficients. The second comes from the traditional difficulty associated with the Laplacian. The purpose of the outer conditioning is for eliminating the large variations in the coefficients.

Theorem 5.2. *Let A_ω the operator mapping $L^2(\mathscr{B})$ into $L^2(\mathscr{B})$ be defined by*

$$A_\omega x = \omega A(\omega x), \quad x \in L^2(\mathscr{B}).$$

Then y is the solution of the generalized Wiener-Hopf equation

$$y \in C_\omega[\mathscr{B}] \;\ni\; T_{P_\omega}(A_\omega)y = P_\omega(\omega f) \tag{5.13}$$

if and only if $x = \omega y \in C[\mathscr{B}]$ is a solution of (5.12). Moreover, for each $k = 1, 2, \ldots, N$ and for all $x, y \in L^2(\mathscr{B})$,

$$\int_{\mathscr{E}_k} |A_\omega x - A_\omega y|^2 \; d\mu \;\; \leq \;\; \beta_2^2 \int_{E_k} |\nabla \mathscr{J} x - \nabla \mathscr{J} y|^2 \; dz \tag{5.14}$$

$$\int_{\mathscr{E}_k} (A_\omega x - A_\omega y)(x - y) \; d\mu \;\; \geq \;\; \alpha_2^2 \int_{E_k} |\nabla \mathscr{J} x - \nabla \mathscr{J} y|^2 \; dz. \tag{5.15}$$

Proof. We rewrite the generalized Winer-Hopf equation (5.12) in the form

$$\int_{\mathscr{B}} (\omega^{-1} y)(\omega A \omega)(\omega^{-1} x) \; d\mu = \int_{\mathscr{B}} (\omega f)(\omega^{-1} y) \; d\mu, \quad \forall y \in \mathrm{C}[\mathscr{B}].$$

Hence, the equivalence between (5.13) and (5.12) is obvious under the relation $x = \omega y$. To prove (5.14)-(5.15), we let $x, y \in L^2(\mathscr{B})$ together with $u = \mathscr{J} x$ and $v = \mathscr{J} y$. In light of the definition of the local stiffness operator A_k, the values of $A_\omega x$ and $A_\omega y$ at the sub-node p_l^k are given by

$$\omega \sum_{j=1}^{2} \int_{E_k} a_j(z, (\mathscr{J}\omega)\nabla u)[\varphi_l^k]_{x_j} dz \quad \text{and} \quad \omega \sum_{j=1}^{2} \int_{E_k} a_\ell(z, (\mathscr{J}\omega)\nabla v)[\varphi_l^k]_{x_j} dz$$

respectively. We proceed in two different cases. First we assume that $\mathscr{E}_k \subset \mathscr{B}_1$. In this case, the assumption A2 implies that

$$\begin{aligned} \|A_\omega x - A_\omega y\| &\leq \frac{\beta_1}{\sqrt{\delta}} \left(\int_{E_k} \left| \frac{\nabla u - \nabla v}{\sqrt{\delta}} \right|^2 \, dz \right)^{1/2} \\ &= \beta_2 \left(\int_{E_k} |\nabla u - \nabla v|^2 \, dz \right)^{1/2}. \end{aligned} \tag{5.16}$$

Obviously, the inequality (5.16) also holds when $\mathscr{E}_k \subset \mathscr{B}_2$ where $\omega = 1$. This proves (5.14). The proof of (5.15) is identical. □

The scaled operator A_ω no longer suffer large stiffness variations which is reflected by (5.14)-(5.15).

5.3. The Inner Conditioning by Quotient. The concept of conditioning is implemented under a different philosophy in the author's approach — we will consider the generalized Wiener-Hopf equation (5.13) in the quotient space $\mathrm{C}_q[\mathscr{B}]$.

The discussion from §5.6 ensures that the operator $I - \mathscr{K}$ when restricted onto $\mathrm{C}_q[\mathscr{B}]$ is invertible. We denote its inverse by $(I - \mathscr{K})_c^{-1}$. In the actual implementation, $(I - \mathscr{K})_c^{-1}$ can be replaced by a simple operation associated with the recovery operator discussed in [11].

Theorem 5.3. *Let $\overline{y}$ be the solution of the generalized Wiener-Hopf equation*

$$\overline{y} \in C_q[\mathscr{B}] \ \ni \ T_{\mathscr{P}_q}(A_\omega)\overline{y} = \mathscr{P}_q(\omega f). \tag{5.17}$$

Then $y = (I - \mathscr{K})_c^{-1}\overline{y}$ is a solution of (5.13). Conversely, if $y \in C^0_\omega[\mathscr{E}]$ is a solution to (5.13), then $\overline{y} = (I - \mathscr{K})y$ is a solution to (5.17). Moreover, the operator $T_{\mathscr{P}_q}(A_\omega)$ is Lipschitz continuous and strongly monotone on $C_q[\mathscr{B}]$. More precisely, for all $x, y \in C_q[\mathscr{B}]$,

$$\int_{\mathscr{B}} |T_{\mathscr{P}_q}(A_\omega)x - T_{\mathscr{P}_q}(A_\omega)y|^2 \, d\mu \leq c_{l0}\beta_2^2 \int_{\mathscr{B}} |x - y|^2 \, d\mu, \tag{5.18}$$

$$\int_{\mathscr{B}} (T_{\mathscr{P}_q}(A_\omega)x - T_{\mathscr{P}_q}(A_\omega)y)(x - y) \, d\mu \geq c_{m0}\alpha_2^2 \int_{\mathscr{B}} |x - y|^2 \, d\mu. \tag{5.19}$$

Here α_2 and β_2 are the constants appeared in the assumptions A1-A2 of §1; c_{l0} and c_{m0} are constants depending only on the regularity of the mesh.

Proof. Following the definition of A_ω and f, it is easy to see that the following identities hold. For all $x, y \in L^2(\mathscr{B})$,

$$\begin{aligned} \int_{\mathscr{B}} (A_\omega x)y \, d\mu &= \int_{\mathscr{B}} [A_\omega(I - \mathscr{K})x](I - \mathscr{K})y \, d\mu, \\ \int_{\mathscr{B}} \omega f y \, d\mu &= \int_{\mathscr{B}} (\omega f)(I - \mathscr{K})y \, d\mu. \end{aligned}$$

Therefore the equivalence between (5.13) and (5.17) becomes trivial. The estimates (5.18)-(5.19) directly follow from (5.14)-(5.15). This completes the proof. □

6. Summary of the Main Algorithm

According to Theorems 5.1 - 5.3, we will solve the generalized Wiener-Hopf equation (5.17) in the quotient space $\mathrm{C}_q[\mathscr{B}]$. In theory, the algorithm can be summarized as 2 nested iterations, the outer iteration and the inner iteration. The outer iteration is given by the Banach contraction mapping principle

$$y_{k+1} = y_k - \lambda T_{\mathscr{P}_q}(A_\omega) y_k + \lambda \mathscr{P}_q(\omega f), \quad k = 0, 1, \ldots. \tag{6.1}$$

where $\lambda = c_{m0}\alpha_2 / c_{l0}\beta_2$. By Theorem 5.3, the contraction constant is bounded by $\sqrt{1-\lambda}$, which is independent of δ and N_d. Thus the cost of performing iterations in (6.1) is linear provided that of evaluating $\mathscr{P}_q$ is linear.

The outer iteration must be followed by a post-processing step since it is operated in the quotient space $\mathrm{C}_q[\mathscr{B}]$. Suppose the iteration stops for $k = s$. Then by Theorem 5.3 again, $y = (I - \mathscr{K})_c^{-1} y_s$ yields the desired approximation to (5.13). The exact procedure for selecting s can be done by the usual method of checking residuals. The evaluation of $(I - \mathscr{K})_c^{-1} y_s$ has been discussed with great detail in [11].

The inner iteration is for the evaluation of $\mathscr{P}_q$ that appears in the outer iteration (6.1). This has been discussed in §3.9, where the task of computing $\mathscr{P}_q$ is converted to computing $\mathscr{P}_0$ by the local filter of constants discussed in §3.8 that connects the problem with a weighted space of zero mean defined on an abstract graph. The core step of computing $\mathscr{P}_q$ rests on the effective evaluation of the operator E_ω^0, which is equivalent to solving a linear system similar to the 5-point finite difference Laplacian. Such perspectives have been discussed in 4.12 Corollary 4.13, and the end of §3.7. Thus, the cost of evaluating $\mathscr{P}_q$ is also linear if we choose to handle E_ω^0 by a linear speed solver such as multigrid.

In the actual programming, there is no need to follow the nested iterations stemming from (6.1). Instead, the nested iterations can be equivalently cllapsed into a single layer of iterations by solving the generalized Wiener-Hopf equation

$$\xi \in R_\omega^0[\mathscr{E}] \ \ni \ T_{E_\omega^0}(\rho \Phi^* \Psi^{-*} A_\omega \Psi^{-1} \Phi \xi) = E_\omega^0(\rho \Phi^* \Psi^{-*} \omega f) \tag{6.2}$$

followed by the post-processing procedure $x = \omega \Psi^{-1} \Phi \xi$. In the special case as described by Corollary 4.9, careful calculation reduces (6.2) even further to

$$\xi \in R_\omega^0[\mathscr{E}] \ \ni \ T_{E_\omega^0}(\rho \Psi^{-*} A \Psi^{-1} \xi) = E_\omega^0(\rho \Psi^{-*} f), \quad \omega \equiv 1, \tag{6.3}$$

in which the large jumps in coefficients are completely cured without scaling.

7. Benchmark Test 1: Elasto-plastic Membrane

By using the algorithm described in the current paper, a variety of new benchmarks can be established in terms of computational performance. The first benchmark is concerned with the quasi-linear elliptic equation

$$-\sum_{j=1}^{2} \left[a_j(u_{x_j}) \right]_{x_j} = f \tag{7.1}$$

on the unit square $\Omega = (0,1)^2$ by prescribing the homogeneous Neumann boundary condition. The right hand side load f is assumed to satisfy the compatibility condition

$$\int_\Omega f dx = 0. \tag{7.2}$$

While the equation models the vertical deflection of an idealized elasto-plastic membrane under a balanced load, its mechanical origin here is of less relevance. The coefficient $a_j(\cdot)$ as a function of u_{x_j} is defined by

$$a_1(\epsilon) = a_2(\epsilon) = \begin{cases} \epsilon & \text{if } |\epsilon| \leq 5 \\ 0.1(\epsilon - 5) + 5 & \text{if } \epsilon > 5 \\ 0.1(\epsilon + 5) - 5 & \text{if } \epsilon < -5 \end{cases} \tag{7.3}$$

At issue here is the computational performance of the algorithm in handling the

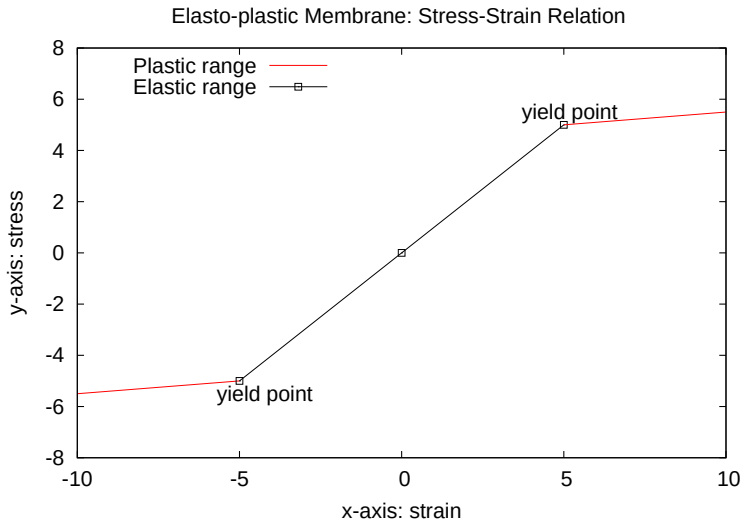


FIGURE 8. Piece-wise linear stress-strain relation, resulting a monotone, non-differentiable operator in the Sobolev space.

large variation of the Young's modulo on the elastic region and plastic region respectively, represented by the slopes of the black and red linear segments in Figure 8. The interface of the regions is a free-boundary which is not given a priori. In the traditional engineering approach, such problems are often treated by the method of incremental loading [?] in order to separate the regions during the solution process. The above problem also presents several difficult aspects for solver algorithms in modern times such Newton-Krylov method and FAS. In particular,

a. Since $a_1(\cdot)$ and $a_1(\cdot)$ are not differentiable, it would be difficult to analyze the convergence of a Newton-Krylov type scheme in this case.
b. The solution to the boundary value problem has little regularity beyond $W^{1,p}(\Omega)$ for some $p > 2$. In this regard, any solver algorithm whose convergence is based on the regularity of solution is likely to loose its theoretical footing.
c. The homogeneous Neumann boundary condition also adds a moderate singularity to the problem that has not been carefully addressed in the traditional methodology.

In the following numerical examples, the right hand force f is always given by

$$f = 4\pi^2 \cos(2\pi x_1) + 4\pi^2 \cos(2\pi x_2). \tag{7.4}$$

Both Figure 9 and Figure 10 are computed by using 65,536 bi-linear quadratic elements. Besides demonstrating the speed of the solver algorithm, our goal is also to accurately reveal the net effect of the plastic deformation. To this end, we first compare the solution to the Poisson equation with the solution to the elasto-plasticity model in Figure 9. Notice that

$$-\Delta u = f, \quad \text{where } u = \cos(2\pi x_1) + \cos(2\pi x_2). \tag{7.5}$$

The presence of the plasticity in the model not only changed the characteristic detail of the deformation such as the shape of contour lines, but also doubled the magnitude of the deformation. This perfectly fits into the physics beyond the model: as the stress becomes greater than the yield value, the material becomes much softer which leads to larger deformation.

Figure 10 (top) is a plot of the elasto-plastic deformation minus that of the elastic deformation (the solution to the Poisson equation) in order to further reveal their differences. Here the non-smoothness of the elasto-plastic deformation becomes evident.

Figure 10 (bottom) shows the distribution of the elastic region (blue) and the plastic region (red). We observe that the location of the plastic region does not necessarily occur where the magnitude of the force is large, but rather, it occurs where gradient of the force is large. This can be explained heuristicly by the Poisson equation

$$-\Delta u_{x_j} = f_{x_j} \quad \text{in the elastic region}$$

where f_{x_j} must be relatively small so that u_{x_j} does not go beyond the yield point. To a large extent, the elaso-plasticity model is a free boundary problem for the identification of the elastic-plastic interface. Once determined accurately, the model reduces to a linear problem with discontinuous coefficients. By comparing the top and the bottom in Figure 10, it also clearly shows that the non-smoothness of the solution occurs exactly on the interface.

Figure 11 summarizes the linear speed of the algorithm. Five recordings are made with the number of elements ranging from 65,536 to 4,194,304. All of them are executed in the author's laptop with a single Intel Celeron processor, whose CPU speed is 2.80GHz; cache size is 128 KB; memory size is 512 MB. At each (outer) iteration step, the maximum norm of the residual is checked. The iteration stops when the residual becomes less than 10^{-4}. Notice that with over 4 millions of elements, the clock time is less than 8 minutes on the low end laptop. It would be extremely interesting to see if an upper end computer can solve the same elasto-plastic model with a less clock time by using other algorithms in prior art.

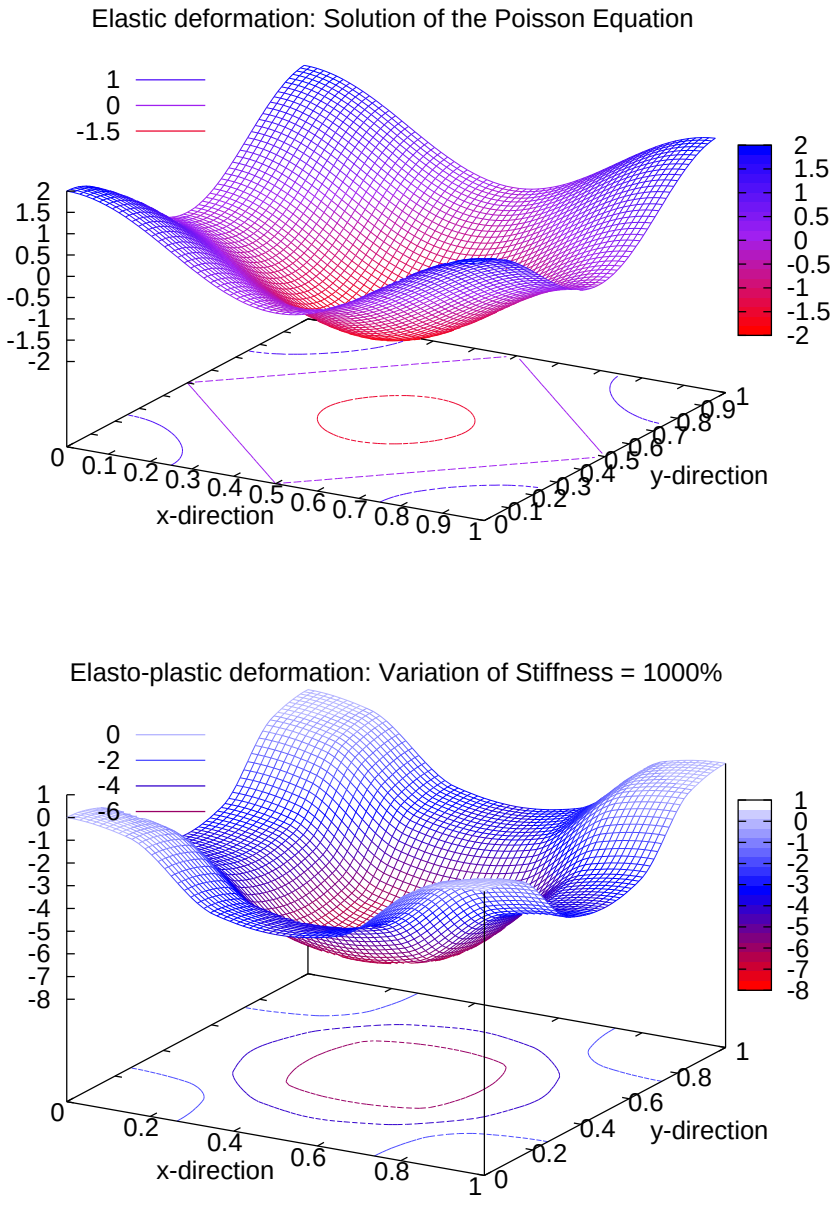


FIGURE 9. The exact solution to the Laplacian (top) and the computed solution to the elasto-plastic model (bottom)

.

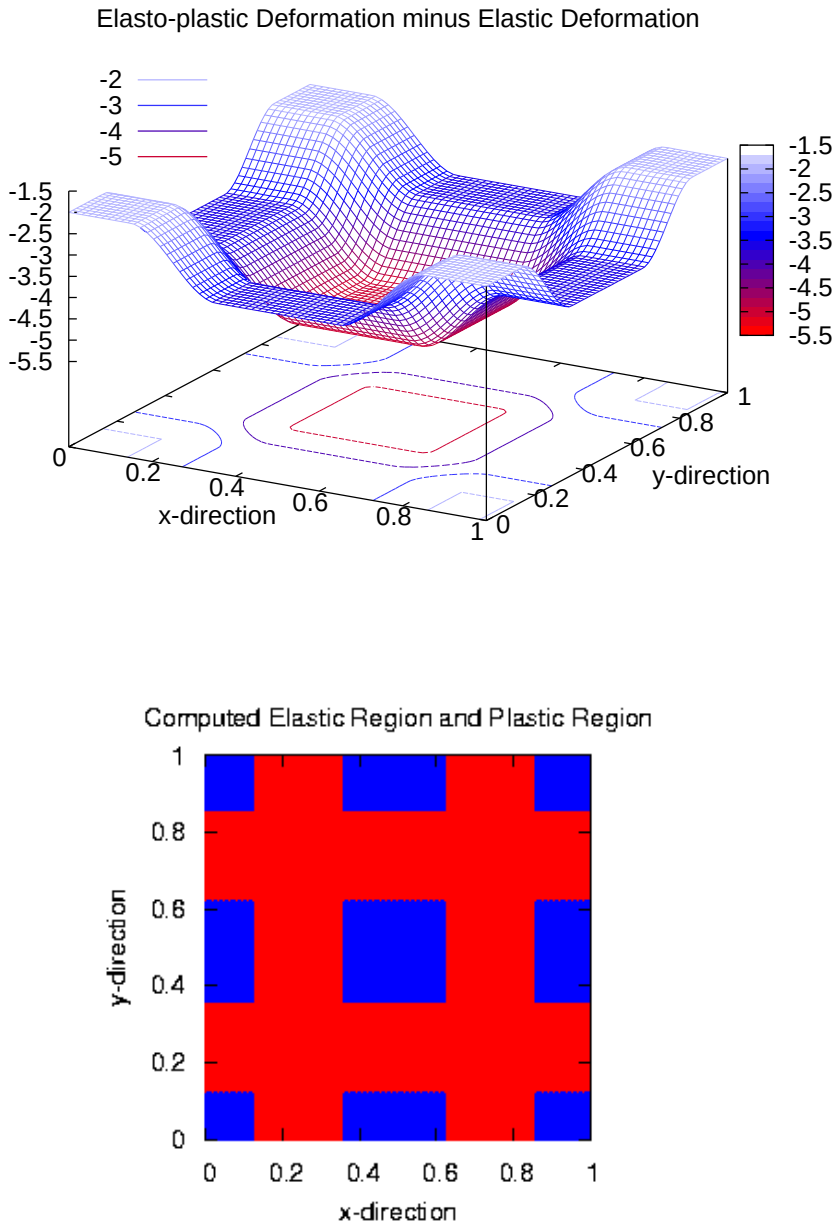


FIGURE 10. The solution to elasto-plastic model minus that of Laplacian (top), and the computed elastic region and the plastic model (bottom)

.

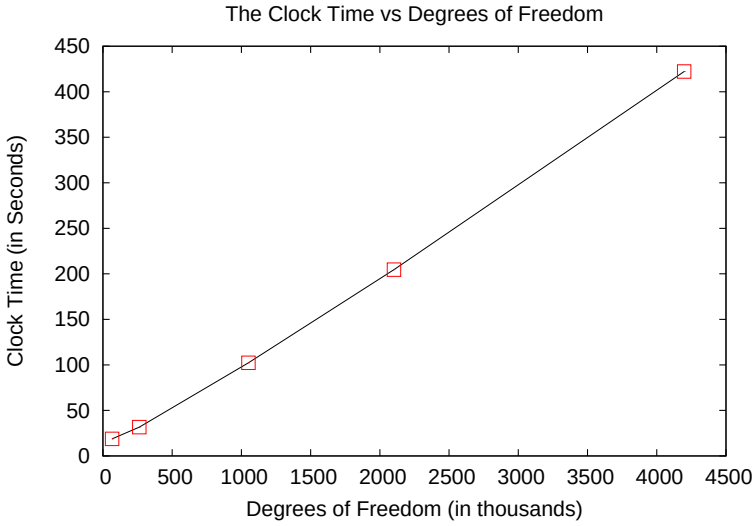


FIGURE 11. The linear speed of the algorithm. We assume that the CPU clock time is linearly related to the number of float point operations.

8. Benchmark Test 2: Soft Inclusions

The second benchmark is concerned with a typical scenario in material sciences: soft inclusions — tiny soft material blocks are mixed into the originally harder material matrix in order to increase the product's flexibility. In the idealized situation, the new material can be modeled by the linear problem

$$-\sum_{j=1}^{2}\left[a_\delta(x)u_{x_j}\right]_{x_j} = f \quad \text{in } \Omega, \tag{8.1}$$

where $a_\delta(x)$ is rapidly alternating between 1 and δ with $1 \gg \delta$. Again, only homogeneous Neumann boundary condition is considered. On the finite element level, there are additional issues involved. The most prominent one is how we model the locations of the soft inclusions. Here we take a simplified approach by assuming that $a_\delta(\cdot)$ is semi-periodic on the mesh. More precisely, we assume that the mesh and the values of $a_\delta(\cdot)$ is described in Figure 12 (see also Figure 1). At issue here is the computational speed of the algorithm with respect to the mesh-size h and its robustness with regard to the smallness of δ. In particular, we reveal that, through the computational evidence, for a fixed mesh size the finite element solution converges exponentially as $\delta \to 0^+$. Such phenomena has neither been proved theoretically nor observed computationally in the past. Also of interest is the performance of the algorithm as $h \to 0^+$ while δ is fixed. This is to numerically simulate the homogenization procedure without using homogenized equation[5].

Finally, we present some computational result for soft inclusions with negative stiffness ($\delta < 0$). This is a new frontier for computational material sciences. Finite element solver algorithms in the past have not effectively covered this matter.

[5]If $1 \gg \delta$, the so-called cell problem itself has large jumps in coefficients which is as difficult to solve as problem (8.1)

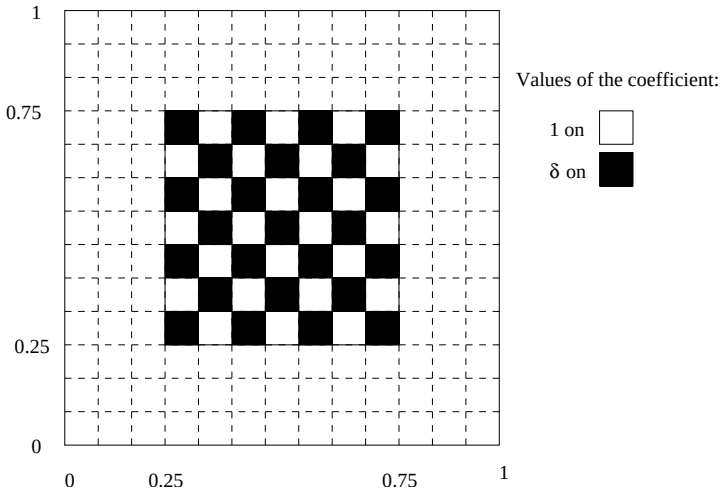


FIGURE 12. The finite element model of semi-periodic soft inclusions. The coefficient $a_\delta(\cdot)$ takes the form $a_{\delta,h}(\cdot)$, which is 1 on elements in the hard region (white) and is δ is on elements in the soft region (black).

In the following discussion, we let $u_{\delta,h}$ denote the finite element solution to problem (8.1) and let u_h denote the finite element solution to the Poisson equation respectively. Figure 13 illustrates $u_{\delta,h} - u$ as $\delta = 0.1,\ 0.01$. The plot of $u_{\delta,h} - u$ with a much smaller value $\delta = 10^{-4}$ is illustrated in Figure 14 (top). Here the number of elements, and hence the number of soft blocks are fixed. Our goal is to answer the following basic questions.

(1) By blending the soft inclusions "evenly" as illustrated in Figure 12, how softer will the new material become as $\delta \to 0^+$?
(2) How much will soft inclusions affect the solution globally beyond the concentric square $(0.25, 0.75) \times (0.25, 0.75)$?

The answer to question (1) is given in Figure 14 (bottom), where the maximum value of $|u_{\delta,h} - u|$ is plotted against $1/\delta$ in log-scale, which clearly shows that $u_{\delta,h} - u$, and hence $u_{\delta,h}$ itself, converges exponentially as $\delta \to 0^+$. In fact, at $\delta = 0$, the computed result is identical with the result when $\delta = 10^{-6}$ under the single precision float point operations. In the current situation, it happens that $\max |u_{\delta,h} - u|$ occurs at the center of the domain $(0.5, 0.5)$, so the similar limit behavior can also be observed from Figure 13 to Figure 14 (top).

We point out that Figure 13 and Figure 14 are computed using 65,536 number of elements. The clock time is virtually independent of the values of δ which slightly varies around 27 seconds.

Figure 15 shows the effect of soft inclusions of negative stiffness with $\delta = -0.3$. It is computed by using about 16,384 elements with mesh size $h = 0.0078125$. It has been experimentally shown that a composite material made from inclusions of negative stiffness can be stabilized if such inclusions are bounded by a material matrix of positive stiffness [9].

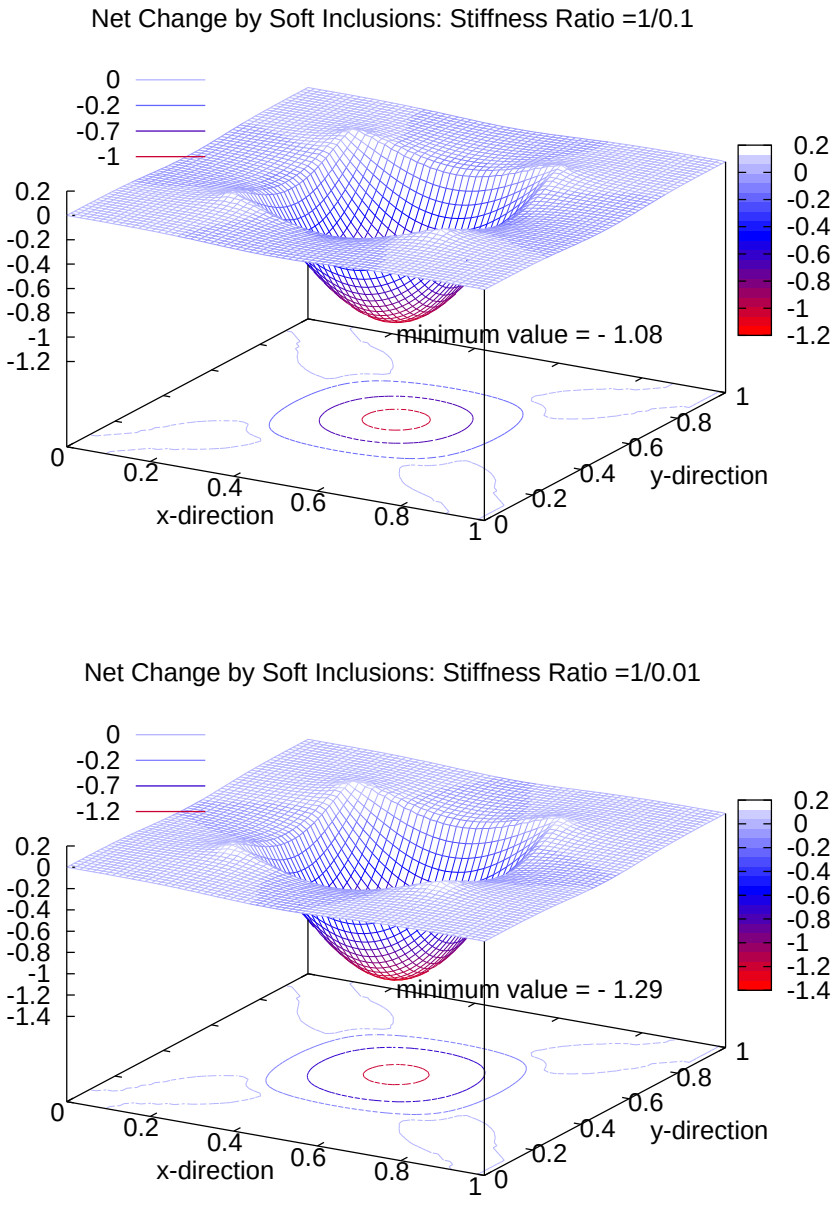


FIGURE 13. The computed finite element solution of equation (8.1) minus the solution to the Laplacian. Top: $\delta = 0.1$. Bottom: $\delta = 0.01$

.

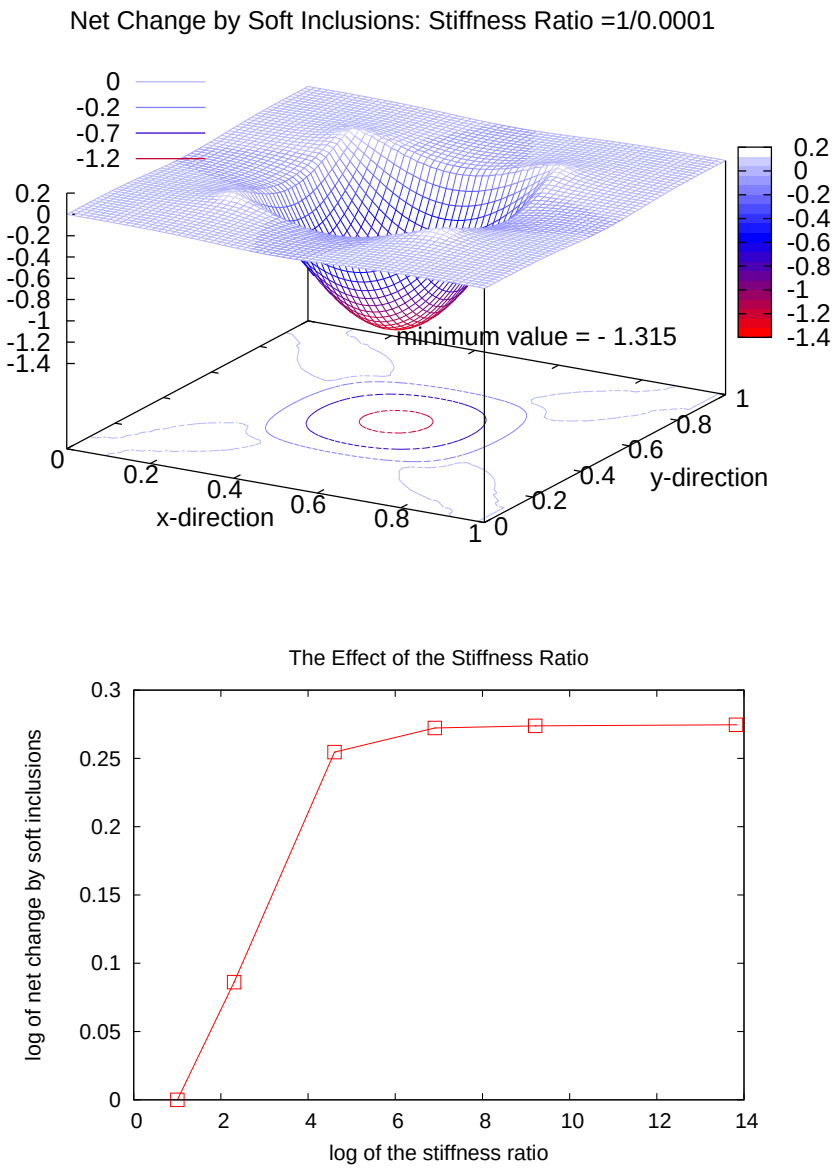


FIGURE 14. (Top: the computed solution minus the solution to the Laplacian at the stiffness ratio $1/\delta = 10^4$. Bottom: the rapid converging behavior of the net change as the stiffness ratio increase from 1 to 10^6 (in log-log scale)

In the current computational scenario, the soft inclusions are not bounded by, but rather, alternatingly connected with the material of positive stiffness, and such setting is apparently difficult to be arranged for experimentation. We must carefully observe Figure 15 (bottom) in order to understand Figure 15 (top). Intuitively, one would imagine that inclusions of negative stiffness would make the composite material stiffer as described in [9]. However, the computed result in Figure 15

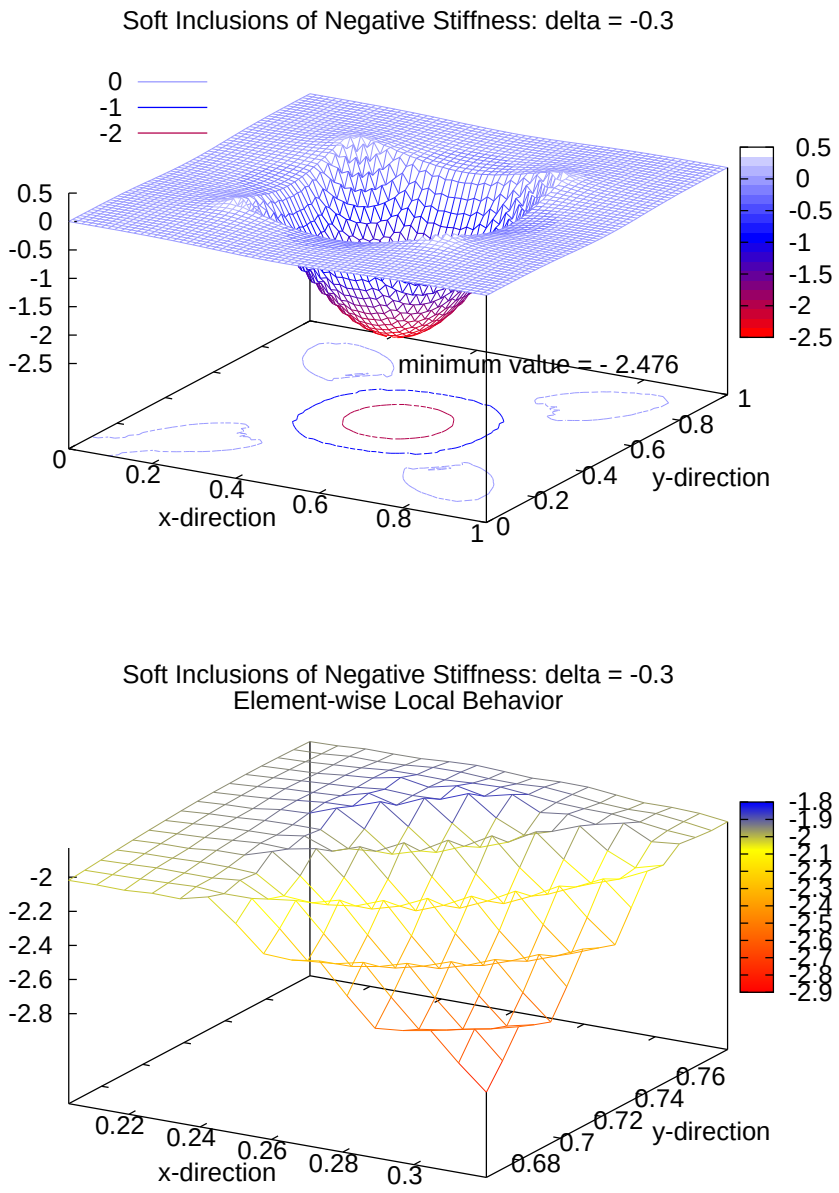


FIGURE 15. (Top: the computed solution minus the solution to the Laplacian with soft inclusions of negative stiffness $\delta = -0.45$. Bottom: local element-wise plot of the top to show non-smoothness

(top) seems to suggest the opposite — a much less stiff composite with much larger deformation. First of all, this is not in contradiction to the findings in [9] since the arrangement of the inclusions in our setting is different. Secondly, the notion of stiffness for the new composite remains to be rigorously defined, and for this, it is

worth separate articles for detailed discussion. Figure 15 (bottom) shows a highly non-smooth local behavior of the deformation. It forms a stair-like configuration in contrast to smooth variations in Figure 13 where only inclusions of positive stiffness are present. While it is not surprising that on elements of negative stiffness the local deformation is opposite to the direction of the applied force, what is truly surprising is that such opposition induces a sharp descend highly non-proportional to the applied force on the adjacent elements of positive stiffness.

We now comment on the convergence of the algorithm when δ is negative. For the example considered in this section, it is not difficult to show that the algorithm is a contractive iteration if $\delta > -0.2111325$, and the number 0.2111325 is the smallest eigenvalue of the local stiffness matrix of the Laplacian on the unit square when bi-linear shape functions are used. Thus $\delta = -0.3$ in Figure 15 is out of the theoretical range of convergence for the algorithm. Through large number of numerical experiments, the following conjecture seems to be valid. In the context of the example considered in this section, for each $-1 < \delta \leq -0.2111325$, there exists a mesh size h_δ such that the algorithm converges for all mesh-size $h \leq h_\delta$. In this case the iteration is no longer contractive. The convergence rate deteriorates as the residual becomes small. This is illustrated in Figure 16 with $\delta = -0.3$ and $h = 0.0078125$. The following table records the last 6 iterations before it is forced

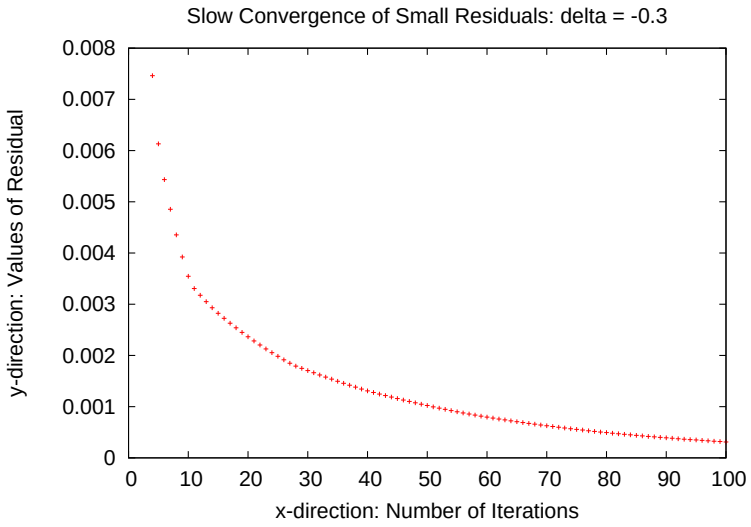


FIGURE 16. The convergence is rapid when the residual is large, and deteriorates as the residual becomes small.

to stop. It is not difficult to see that for such small residuals, the convergence rate is similar to what we suffered from the classical iterations such as Jacobi or Gauss-Seidel.

Iteration	Residual
95	0.000348985
96	0.000341028
97	0.000333190
98	0.000325590
99	0.000318184
100	0.000310913

In contract, as we take $\delta = 0.2$ and $h = 0.0078125$, which is unconditionally in the theoretical range of convergence with fixed rate, the computational performance is significantly better. This time it only takes 37 iterations to obtain an much smaller residual at 9.28119×10^{-5}. Figure 17 records the last 7 iterations that shows each of these 7 iterations reduces the residual about 8.59%, which amounts to say that the contraction constant is about 0.914.

Figure 18 shows the local behavior of $u_{\delta,h} - u_h$ with $\delta = -0.2$. As we compare it with Figure 15 (bottom) which corresponds to $\delta = -0.3$, the smoothness is significantly improved, and the oscillation is not as severe. However, it is important to notice they share the same characteristic: the included blocks of negative stiffness serve as local dampers to deform in opposition to the applied force while inducing much larger deformation along the direction of applied force on the adjacent material matrix of positive stiffness. Whether this is a realistic (rather than numerical only) phenomena remains to be seen by actual experiments.

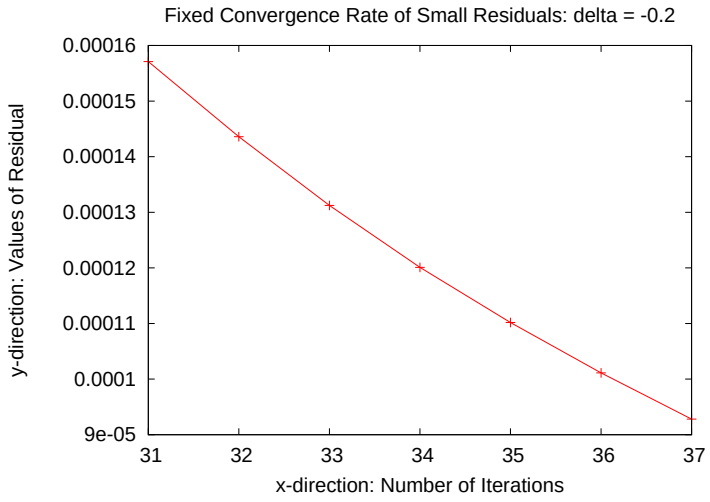


FIGURE 17. Fixed convergence rate when $\delta = -0.2$

We further comment that for $\delta = -0.1$, it takes 20 iterations to achieve a residual at 9.78783×10^{-5}. With each iteration, the the residual reduces about 17.42% vs 8.59% for the case when $\delta = -0.2$. In addition, the convergence rate improves as the mesh size h becomes smaller. For example, with $\delta = -0.1$ and $h = 0.00390625$ the residual reduces by about 20.89% each iteration vs 17.42% for the case when $h = 0.0078125$.

In concluding our numerical experiments, we summarize the performance of the algorithm as a tool for numerical homogenization. The following table is generated with a fixed $\delta = 0.001$.

# of elements	time (seconds)	$\min\{u_{\delta,h}\}$
4,096	3.09	-5.422
16,384	11.09	-5.348
65,536	27.55	-5.313
262,144	62.12	-5.296
1,048,576	244.73	-5.287
4,194,304	998.25	-5.283

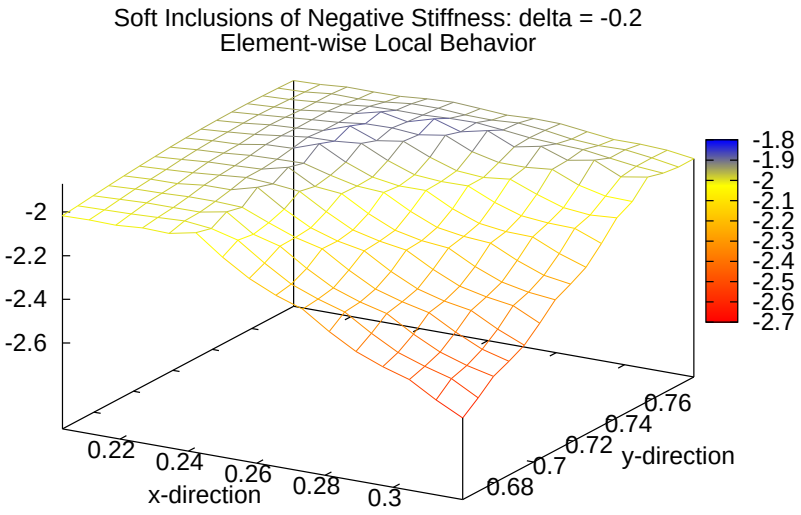


FIGURE 18. Local behavior of $u_{\delta,h} - u_h$ with $\delta = -0.2$, demonstrating much more smoothness than the case for $\delta = -3.0$

The second column shows the linear speed of the algorithm. The third column suggests that the overall stiffness of the composite increases as the frequency of the oscillatory coefficient $a_{\delta,h}$ increases. This is to say that a finer mix of the soft material increases the stiffness. However, the result must not be interpreted as a general phenomena since we have modeled each soft block by a single element, and within each element the only permissible deformations are those dictated by the shape functions on the vertexes. Whether it is more appropriate to model each soft block by higher order elements depends on the fine scale nature of the block, and is a matter of separate discussion.

References

[1] G. Allaire and Robert Brizzi, *A multiscale finite element method for numerical homogenization*, R.I.N^o 545, July 2004.

[2] T. Arbogast, *Analysis of a two-scale, local ly conservative subgrid upscaling for el liptic problems*, SIAM J. Numer. Anal., **42** (2004), 576-598.

[3] A. Brandt, S.F. McCormick, and J.W. Ruge, *Algebraic multigrid (AMG) for automatic multigrid solution with application in geo detic computations*, Technical Rep ort CO POB 1852, Inst. Comp. Studies State Univ., 1982.

[4] A. Brandt, S.F. McCormick, J. Ruge, *Algebraic multigrid (AMG) for sparse matrix equations*, in Sparsity and its Applications (D.J. Evans Ed.), Cambridge University Press, Cambridge, 1984, 257-284.

[5] W.L. Briggs, V.E. Henson, S.F. McCormick, *A Multigrid Tutorial, Second Edition*, SIAM, 2000

[6] T.Y. HOU and X.H. WU, A multiscale finite element method for el liptic problems in composite materials and porous media, *Journal of computational physics* **134** (1997), 169–189.

[7] Y. Efendiev , T. Hou and V. Ginting, *Multiscale finite element methods for nonlinear problems and their applications*, preprint

[8] T. J. R. Hughes, G. R. Feijoo, L. Mazzei, and J.B. Quincy, *The variational multiscal method–a paradigm for computational mechanics*, Comput. Methods Appl. Mech. Engrg., **166** (1998), 3-24.

[9] R. Lakes, *Extreme damping in composite materials with a negative stiffness phase*, Physical review letters, **86** (2001), 2897–2900

[10] J. Mandel and M. Brezina, *Balancing domain decomposition for problems with large jumps in coefficients*, Math. Comp., **65** (1996), 1387–1401.

[11] P. Shi, *Foundation of fast finite element solvers, Part I*, To appear in Advances in Computational Mathematics.

[12] K. Stüben, *A review of algebraic multigrid*, Journal of Computational and Applied Mathematics, **128** (2001), 281-309

[13] J. Wang, *Convergence analysis without regularity assumptions for multigrid algorithms based on SOR smoothing*, SIAM J. Numer. Anal., **29** (1992), no. 4, 987-1001

[14] J. Xu, *Iterative methods by space decomposition and subspace correction*, SIAM Review **34** (1992), 581-613.

Department of Mathematics and Statistics, Oakland University, Rochester, MI 48309-4401, USA

E-mail: pshi@oakland.edu