REMOVING THE CELL RESONANCE ERROR IN THE MULTISCALE FINITE ELEMENT METHOD VIA A PETROV-GALERKIN FORMULATION*

THOMAS Y. HOU†, XIAO-HUI WU‡, AND YU ZHANG§

Abstract. We continue the study of the nonconforming multiscale finite element method (Ms-FEM) introduced in [17, 14] for second order elliptic equations with highly oscillatory coefficients. The main difficulty in MsFEM, as well as other numerical upscaling methods, is the scale resonance effect. It has been show that the leading order resonance error can be effectively removed by using an over-sampling technique. Nonetheless, there is still a secondary cell resonance error of $O(\varepsilon^2/h^2)$. Here, we introduce a Petrov-Galerkin MsFEM formulation with nonconforming multiscale trial functions and *linear* test functions. We show that the cell resonance error is eliminated in this formulation and hence the convergence rate is greatly improved. Moreover, we show that a similar formulation can be used to enhance the convergence of an immersed-interface finite element method for elliptic interface problems.

Key words. finite element, nonconforming, homogenization, resonance, interface

AMS subject classifications. 65F10, 65F30

1. Introduction

A multiscale finite element method has been developed in [19, 17]. The method is designed to capture accurately the averaged effect of differential operators with highly oscillatory coefficients on the large scale solutions. These solutions are typically computed on grids that are coarser than the scale of oscillations. In practice, the oscillatory coefficients are often representations of random media, e.g., in porous media flows. The general idea of MsFEM is to construct finite element base functions that capture the small scale information of the leading order differential operator. This requires solving a homogeneous equation of the differential operator. A similar method commonly used in the industry is numerical upscaling, where a random medium is first modeled by an effective medium on a coarse grid, and then the effective solution is computed. See [13, 24] and references cited therein. In MsFEM, the first step is bypassed and the focus is on the final solution.

The greatest challenge for MsFEM and numerical upscaling is the error due to scale resonance (cf. [25]), which is characterized by the ratio between the small scale of the media and the grid scale (size). Thus, numerical error becomes large when the grid size is close to the heterogeneity scale of the media. Moreover, most practical simulations are performed under this scenario. Therefore, it is highly desirable to eliminate the resonance error.

Two origins of scale resonance have been revealed in the study of MsFEM for second order elliptic equations [19, 14]. One is the mismatch between the artificial *local* boundary conditions imposed on the base functions and the *global* nature of the

oscillatory solution of the differential operator. It induces boundary layers in the first order correctors of the base functions. It was found that the error due to the boundary layers is most significant and can be effectively removed by using an over-sampling technique [17, 14]. The other origin is the mismatch between the grid size and the "perfect" sample size (e.g., for periodic structures, the "perfect" sample size would be integer multiples of the period). The resonance thus created is called "cell resonance". Though cell resonance cannot be removed by over-sampling, its effect is of lower order [14]. Both types of error exist in the upscaling methods and they are of the same order [25].

In this paper, we modify the nonconforming MsFEM formulation of [14] to remove the cell resonance error. Our main finding is that additional error cancellation can be achieved by choosing piecewise *linear* test functions while keeping the construction of the multiscale trial functions in the original method. In particular, over-sampling is used in the construction of the trial functions to remove the effect of the boundary layers. As a result, the trial space is non-conforming. Thus, our new resonance-free method is a Petrov-Galerkin [21], nonconforming, multiscale finite element method (P-G MsFEM).

We show that this error cancellation mechanism is quite subtle. It does not exist for the original Galerkin method, nor does it occur if the trial space and the test space are exchanged. Furthermore, we compare P-G MsFEM with a typical numerical upscaling method. It turns out that, although the error cancellation does exist for the upscaling method, the cell resonance is brought back by another error term.

The effect of the piecewise linear test space can be understood from another point of view. It was shown in [14] that the cell resonance error comes from the nonconforming error in the Galerkin formulation. Here, we show that P-G MsFEM *has no non-conforming error* due to the conforming piecewise linear test space. As a result, the cell resonance error vanishes. This idea of removing nonconforming error has been used in developing an MsFEM (by the authors, see [23]) for solving multiscale transport problems (cf. [15]). In this case, the nonconforming error is much larger compared to that in the present elliptic problem due to singular perturbation. The same idea can also be used to improve the convergence rate of a nonconforming finite element method for solving elliptic interface problems on non-interface-fitting grids. See Section 5.

Note that the idea of combining a nonconforming trial space and conforming test space has been used in [1] to construct a special finite element method for second order elliptic problems with rough coefficients that are constant along straight lines. The one-dimensional structure of the rough coefficients was utilized in the proof of the stability of the method. The analysis, however, is not applicable to the two-dimensional problems we are interested in. See Sections 3 and 5 for further details. The use of different test spaces to improve the solutions represented by given trial spaces has been widely used in the finite element solution of the convection-diffusion equation. See, e.g., [11, 16, 12, 3, 8] and references cited therein. The current technique is different in that the test space is chosen to eliminate the nonconforming error and bring about additional error cancellation. The cited methods for the convection-diffusion equation often lead to symmetrized discrete systems of equations for asymmetric differential operators; the present technique *asymmetrizes* the discrete equations for a symmetric differential operator. Thus, as a trade off for improving the convergence of the numerical solution, the coercive condition for the discrete formulation becomes nontrivial to analyze and sometimes it may be more restrictive. See Lemma 3.2 and

discussions in Remark 3.6 below. A similar observation was also made in [1].

The rest of the paper is organized as follows: In Section 2, we review the elliptic model problem, relevant homogenization results, and the over-sampling technique. We also introduce the formulation for P-G MsFEM. Next, in Section 3 we prove the convergence of the method. We first prove an energy norm estimate using the conventional approach and show the removal of the nonconforming error. Then we analyze the subtle cancellation of the cell resonance error and derive more estimates in both the energy and L_2 norms. In Section 4 we present supporting numerical results for our theory. Removing nonconforming error in other related problems is summarized in Section 5.

2. Model problem, over-sampling, and P-G MsFEM

In this section we introduce the model problem and the Petrov-Galerkin non-conforming multiscale finite element method (P-G MsFEM). First, we fix some notations and conventions to be used in this paper. In the following, the Einstein summation convention is used: summation is taken over repeated indices. $L_2(\Omega)$ denotes the space of square integrable functions defined in domain Ω. We use the $L_2(\Omega)$ based Sobolev spaces $H^k(\Omega)$ equipped with norms and seminorms given by:

$$\|u\|_{k,\Omega}^2 = \int_\Omega \sum_{|\alpha| \le k} |D^\alpha u|^2, \qquad |u|_{k,\Omega}^2 = \int_\Omega \sum_{|\alpha| = k} |D^\alpha u|^2.$$

$H_0^1(\Omega)$ consists of those functions in $H^1(\Omega)$ that vanish on $\partial\Omega$. $\|u\|_{\infty,\Omega}$ is the L_∞ norm of u in Ω. Throughout, C denotes a generic constant, which is independent of ε and h unless otherwise stated. By $\langle \cdot \rangle_Y$ we denote the mean value of a function in the domain Y, i.e.,

$$\langle f \rangle_Y = \frac{1}{|Y|} \int_Y f(x) dx.$$

2.1. Model problem and the multiscale expansion.   We consider solving the second order elliptic equation:

$$L_\varepsilon u_\varepsilon = f, \qquad \text{in } \Omega, \tag{2.1}$$

where $L_\varepsilon = -\nabla \cdot a_\varepsilon \nabla = -\dfrac{\partial}{\partial x_i} a_\varepsilon^{ij} \dfrac{\partial}{\partial x_j}$ is the linear elliptic operator, $f \in L_2(\Omega)$, a_ε^{ij} is a symmetric positive definite matrix satisfying $\alpha|\xi|^2 \le \xi_i a^{ij} \xi_j \le \beta|\xi|^2$, for all $\xi \in R^2$ and with $0 < \alpha < \beta < \infty$, ε is a small parameter characterizing the small scale in the physical problem. To simplify the presentation of the finite element formulation, we assume $u = 0$, on $\partial\Omega$ and that the solution domain is a unit square $\Omega = [0,1] \times [0,1]$.

The multiscale finite element method is designed mainly for physical problems that do not have scale separation or periodic structure. On the other hand, for problems that have scale separation and periodic structure, one can take advantage of such structure to construct the multiscale basis semi-analytically. To better understand the source of resonance error, we restrict ourselves to the special case when the coefficient has a periodic structure, i.e. $a_\varepsilon^{ij} = a_{ij}(x/\varepsilon)$. Furthermore, we assume that $a^{ij}(y)$ are C^3 periodic functions in y in a unit cube Y. For this special class of coeffiicients, the homogenization theory can be applied to analyze the structure of the solution for problem (2.1). It is known [5] that the solution of (2.1) can be expanded as:

$$u_\varepsilon = u_0 + \varepsilon \chi^i(\frac{x}{\varepsilon}) \frac{\partial u_0}{\partial x_i} + \varepsilon \theta_\varepsilon^u. \tag{2.2}$$

In expansion (2.2), $u_0 \in H^2(\Omega)$ is the solution of the homogenized equation

$$-\nabla \cdot a_0 \nabla u_0 = f, \qquad \text{in } \Omega,$$

satisfying $u_0 = 0$, on $\partial\Omega$. The constant homogenized coefficient matrix a_0 is given by

$$a_0^{ij} = \frac{1}{|Y|} \int_Y a^{ik}(y) \left(\delta_{kj} + \frac{\partial \chi^j(y)}{\partial y_k} \right) dy. \tag{2.3}$$

And χ^k is the periodic solution of

$$-\frac{\partial}{\partial y_i} a^{ij}(y) \frac{\partial}{\partial y_j} \chi^k(y) = \frac{\partial a^{ik}(y)}{\partial y_i} \tag{2.4}$$

in the unit cell Y and $\langle \chi^k \rangle_Y = 0$. For the correction term θ_ε^u, we have

$$
\begin{aligned}
L_\varepsilon \theta_\varepsilon^u &= \frac{1}{\varepsilon} L_\varepsilon \left(u_\varepsilon - u_0 - \varepsilon \chi^i(\frac{x}{\varepsilon}) \frac{\partial u_0}{\partial x_i} \right), \qquad \text{in } \Omega, \\
\theta_\varepsilon^u &= -\chi^i(\frac{x}{\varepsilon}) \frac{\partial u_0}{\partial x_i}, \qquad \text{on } \partial\Omega.
\end{aligned}
\tag{2.5}
$$

It is known that u_ε converges to u_0 under the L_∞ norm when ε approaches 0.

2.2. Multiscale basis and over-sampling method.   For $0 \le h \le 1$, let $\mathcal{K}^h$ be a partition of Ω of triangles K with diameter less than h. In each element $K \in \mathcal{K}^h$, we define a set of nodal basis $\{\phi_{\varepsilon,K}^i\}, i = 1, \cdots, d$ with $d \, (= 3)$ being the number of nodes of the element. We will neglect the subscript K when working in one element. The multiscale base function ϕ_ε^i satisfies

$$L_\varepsilon \phi_\varepsilon^i = 0, \qquad \text{in } K \in \mathcal{K}^h. \tag{2.6}$$

Let $x_j \in K(j = 1, \cdots, d)$ be the nodal points of K. In [19], the multiscale base function satisfies $\phi_\varepsilon^i(x_j) = \delta_{ij}$ and some boundary conditions on ∂K so that (2.6) is well-posed. Similar to (2.2), ϕ_ε^i can be expanded as:

$$\phi_\varepsilon^i = \phi_0^i + \varepsilon \chi^j(\frac{x}{\varepsilon}) \frac{\partial \phi_0^i}{\partial x_j} + \varepsilon \theta_\varepsilon^{\phi^i}. \tag{2.7}$$

If ϕ_ε^i is linear along ∂K, since the homogenized coefficient a_0 is a constant matrix; ϕ_0^i must be linear; and $\phi_0^i(x_j) = \delta_{ij}$ [19]. Thus, by (2.4), (2.5), and (2.6) the equation for $\theta_\varepsilon^{\phi^i}$ can be further simplified:

$$
\begin{aligned}
L_\varepsilon \theta_\varepsilon^{\phi^i} &= \frac{1}{\varepsilon} L_\varepsilon \left(\phi_\varepsilon^i - \phi_0^i - \varepsilon \chi^j(\frac{x}{\varepsilon}) \frac{\partial \phi_0^i}{\partial x_j} \right) \\
&= -\frac{1}{\varepsilon} \nabla \cdot a_\varepsilon \nabla \left(\phi_0^i + \varepsilon \chi^j(\frac{x}{\varepsilon}) \frac{\partial \phi_0^i}{\partial x_j} \right) \\
&= -\left(\frac{1}{\varepsilon} \frac{\partial a_\varepsilon^{kj}}{\partial x_k} + \frac{\partial}{\partial x_k} a_\varepsilon^{kl} \frac{\partial \chi^j}{\partial x_l}(\frac{x}{\varepsilon}) \right) \frac{\partial \phi_0^i}{\partial x_j} \\
&= 0.
\end{aligned}
$$

As mentioned in the introduction, the multiscale finite element method using the multiscale basis discussed above suffers from the resonance error. The main cause

of the resonance error is that the first order corrector, $\theta_\varepsilon^{\phi^i}$, has a boundary layer (see [19]). In [17] and [14], an over-sampling method was introduced to remove the boundary layer effect.

Since the boundary layer is thin, we can sample in a domain with size larger than h and use only the interior data to construct the base function. Specifically, we first construct the base functions ψ_ε^i in a sampling domain $S \supset K$ with $diam(S) = H > h$ and ∂S is away from ∂K at some distance d_S (to be specified below). In doing so, we let ψ_ε^i to satisfy the linear boundary conditions on ∂S. Then, the base functions ψ_ε^i on K are constructed from the linear superposition of ψ_ε^i:

$$\phi_\varepsilon^i = \sum_{j=1}^{d} c_{ij} \psi_\varepsilon^j.$$

The constants c_{ij} are determined by the condition $\phi_0^i(x_j) = \sum_{l=1}^{d} c_{il} \psi_0^l(x_j) = \delta_{ij}$. Here, x_j are the nodal points of K and ψ_0^l are the homogenized part of ψ_ε^l (see (2.7)) which are linear. Note that this construction follows that in [14]; it is slightly different from the original construction given in [17], where $\phi_\varepsilon^i(x_j) = \delta_{ij}$ was used to determine c_{ij}.

By this procedure, the boundary layer structure near ∂S can be avoided if d_S is sufficiently large. Noting that $\theta_\varepsilon^{\psi^i} = -\chi^j(x/\varepsilon) \partial \psi_0^i / \partial x_j$ on ∂S, ϕ_ε^i can be expressed as:

$$\phi_\varepsilon^i = \phi_0^i + \varepsilon \chi^j\left(\frac{x}{\varepsilon}\right) \frac{\partial \phi_0^i}{\partial x_j} + \varepsilon \eta(x)^j \frac{\partial \phi_0^i}{\partial x_j}, \tag{2.8}$$

with η^j being the solution of

$$\nabla \cdot a_\varepsilon \nabla \eta^j = 0 \quad \text{in } S, \qquad \eta^j(x) = -\chi^j\left(\frac{x}{\varepsilon}\right) \quad \text{on } \partial S. \tag{2.9}$$

Thus, d_S is determined by the thickness of the boundary layer of η_j. Numerically, it has been observed that the boundary layer is about $O(\varepsilon)$ thick [17, 18]. It was also observed that $d_S = h(> \varepsilon)$ is usually sufficient for eliminating the boundary layer effect. These findings are consistent with those in [4], where the boundary layer structure of the half-plane problem was analyzed. However, the boundary layer is not well understood for domains with corners. Noting that each S may contain multiple elements (which in fact has advantage in practical computations [17]) and that the shape of S can be flexible (i.e., it does not have to be triangular or rectangular), we make the following assumption:

ASSUMPTION 2.1. *The over-sampling domain S is chosen such that for any element K in S*

$$\|\nabla \eta^i\|_{L_\infty(K)} \leq C, \tag{2.10}$$

where C is a constant that is independent of ε and h.

We remark that Assumption 2.1 has not been found to be necessary in practice, see, e.g., [17, 14] and the numerical tests given later in this paper. On the other hand, the assumption can be satisfied by taking a sample domain S so that the distance from the boundary of S to K, denoted by $dist(K, \partial S)$, is large enough. In particular, using an estimate for $\nabla \chi$ from [9], we obtain that $\|\nabla \eta^i\|_{L_\infty(K)} \leq c/dist(K, \partial S)$ with

c independent of ε and h. So if there is enough separation from ∂S to K, the above assumption is satisfied.

We define two finite element spaces as follows:

$$\Phi_\varepsilon^h = \text{span}\{\phi_{\varepsilon,K}^i; \quad i = 1, \cdots, d, \quad K \subset \mathcal{K}^h\};$$
$$\Phi_0^h = \text{span}\{\phi_{0,K}^i; \quad i = 1, \cdots, d, \quad K \subset \mathcal{K}^h\} \subset H_0^1(\Omega).$$

Note that in general $\Phi_\varepsilon^h \not\subset H_0^1(\Omega)$, since the multiscale base functions constructed by the over-sampling method are no longer continuous across the internal boundaries of the elements.

We remark that when the periodic fine-scale structure is known, one can construct the multiscale basis functions by using the first two terms of (2.7), that is:

$$\phi_\varepsilon^i = \phi_0^i + \varepsilon \chi^j\left(\frac{x}{\varepsilon}\right)\frac{\partial \phi_0^i}{\partial x_j}. \tag{2.11}$$

This avoids the boundary layer effect completely. Numerical experiments by Andrew Westhead demonstrate a clear first order convergence of this method without suffering from resonance error. For more details, see www.ama.caltech.edu/ westhead/MSFEM.

2.3. Petrov-Galerkin nonconforming multiscale finite element method.
The variational problem of (2.1) is to seek $u_\varepsilon \in H_0^1(\Omega)$ such that

$$a(u_\varepsilon, v) = f(v), \qquad \forall v \in H_0^1(\Omega), \tag{2.12}$$

where

$$a(u, v) = \int_\Omega a_\varepsilon^{ij} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} dx, \qquad f(v) = \int_\Omega f v dx.$$

A Petrov-Galerkin finite element method is obtained by restricting the weak formulation (2.12) to some finite dimensional subspaces of $H_0^1(\Omega)$. We define two subspaces of $H_0^1(\Omega)$, namely U^h and V^h, on $\mathcal{K}^h$. The finite element solution $u_\varepsilon^h \in U^h$ is obtained by solving the following variational problem

$$a(u_\varepsilon^h, v) = f(v), \qquad \forall v \in V^h.$$

The subspaces U^h and V^h are the trial space and the test space, respectively. Generally, they need not be the same. Different choices of U^h and V^h lead to different methods (cf. [16]).

For the classical finite element method, $U^h = V^h = \Phi_0^h$. In Appendix A, we prove that fixing $h \ll 1$ and letting ε go to 0, the solution of this finite element method, $\overline{u}^h$, is an approximation to $\overline{u}$, the solution of the equation

$$-\nabla \cdot (\langle a \rangle_Y) \nabla \overline{u} = f \quad \text{in } \Omega \qquad \text{and} \qquad \overline{u} = 0 \quad \text{on } \partial\Omega.$$

Since $\langle a \rangle_Y \neq a_0$ in general, we have $\overline{u} \neq u_0$ and hence the classical finite element method gives a wrong solution for problem (2.1) when $h \gg \varepsilon$. See Theorem A.1.

In [17], a multiscale finite element was developed where Φ_ε^h was used instead of Φ_0^h. The method is *nonconforming* because Φ_ε^h is not a subspace of $H_0^1(\Omega)$. The bilinear form $a(\cdot, \cdot)$ needs to be modified as

$$a_h(u, v) = \sum_{K \in \mathcal{K}^h} \int_K \nabla v \cdot a_\varepsilon \nabla u dx.$$

We seek $u_\varepsilon^h \in \Phi_\varepsilon^h$ such that

$$a_h(u_\varepsilon^h, v) = f(v), \qquad \forall v \in \Phi_\varepsilon^h.$$

Following the proof of Theorem 3.1 of [14] and using Assumption 2.1, it can be shown that u_ε^h converges to u_ε when $h \gg \varepsilon$, i.e.,

$$\|u_\varepsilon^h - u_\varepsilon\|_{h,\Omega} \leq C_1 \frac{\varepsilon}{h} + C_2 \sqrt{\varepsilon} + C_3 h, \tag{2.13}$$

where $\| \cdot \|_{h,\Omega}$ is the energy norm defined as

$$\|v\|_{h,\Omega} = \left(\sum_{K \in \mathcal{K}^h} \int_K |\nabla v|^2 dx \right)^{\frac{1}{2}}.$$

(This is indeed a norm in Φ_ε^h, see Remark 3.1 of [14].) Note that estimate (2.13) obtained with Assumption 2.1 is identical to that derived in [14] without using the assumption. However, here the $O(\varepsilon/h)$ term contains solely the cell resonance effect, whereas the same term in (3.4) of [14] contains the contribution of the first order corrector (see the last term in (2.8)).

The goal is to eliminate the $O(\varepsilon/h)$ error in (2.13). We propose the following Petrov-Galerkin formulation–seek $u_\varepsilon^h \in \Phi_\varepsilon^h$ such that:

$$a_h(u_\varepsilon^h, v) = f(v), \qquad \forall v \in \Phi_0^h \subset H_0^1(\Omega). \tag{2.14}$$

Note that the trial space is oscillatory and nonconforming while the test space is piecewise linear and conforming. We will prove that the $O(\varepsilon/h)$ resonance error vanishes in this new method.

3. Convergence of P-G MsFEM

In this section we prove the convergence of P-G MsFEM. Let the finite element solution $u_\varepsilon^h \in \Phi_\varepsilon^h$ be expressed as

$$u_\varepsilon^h = \sum_{i=1}^d u_i \phi_{\varepsilon,K}^i \qquad \text{in } K \in \mathcal{K}^h,$$

and set

$$u_0^h = \sum_{i=1}^d u_i \phi_{0,K}^i \qquad \text{in } K \in \mathcal{K}^h.$$

Thus, by (2.8) u_ε^h has the following expansion

$$u_\varepsilon^h = u_0^h + \varepsilon \chi(x/\varepsilon) \cdot \nabla u_0^h + \varepsilon \theta^h \tag{3.1}$$

on each element K, where $\theta^h = \eta^i \partial u_0^h / \partial x_i$.

LEMMA 3.1. *Under assumption 2.1, the following estimates are valid:*

$$\|\theta^h\|_{0,\Omega} \leq C\|\nabla u_0^h\|_{0,\Omega}, \qquad \|\theta^h\|_{h,\Omega} \leq C\|\nabla u_0^h\|_{0,\Omega}. \tag{3.2}$$

Proof. Since $a(y)$ is a smooth function, so is χ^i. Thus, by (2.9) $\|\eta^i\|_{\infty,S} \leq \|\chi^i\|_{\infty,S} \leq C$. Therefore, the first estimate holds. The second estimate can be obtained from (2.10) and the fact that u_0^h is piecewise linear. □

It follows immediately that

$$\|u_\varepsilon^h\|_{h,\Omega} \leq \|\nabla u_0^h\|_{0,\Omega} + \varepsilon\|\nabla\chi \cdot \nabla u_0^h\|_{h,\Omega} + \varepsilon\|\theta^h\|_{h,\Omega} \leq C\|\nabla u_0^h\|_{0,\Omega}. \qquad (3.3)$$

Define

$$\tilde{a}^{ij}(y) = a^{ik}(y)\left(\delta_{kj} + \frac{\partial\chi^j(y)}{\partial y_k}\right).$$

It is clear from (2.3) and (2.4) that

$$\langle\tilde{a}\rangle_Y = a_0 \qquad \text{and} \qquad \nabla \cdot \tilde{a} = 0. \qquad (3.4)$$

The following Lemma about $\tilde{a}$ will be useful:

LEMMA 3.2. *There exists a constant C, such that when the ratio $\frac{\varepsilon}{h} < C$, $\tilde{a}$ satisfies the following inequalities*

$$C_1\|\nabla u\|_{0,K}^2 \leq \int_K \nabla u \cdot \tilde{a}(x/\varepsilon)\nabla u dx \leq C_2\|\nabla u\|_{0,K}^2, \qquad \forall u \in \Phi_0^h, \quad \forall K \in \mathcal{K}^h. \quad (3.5)$$

Proof. Since a_0 is symmetric and $(\langle a^{-1}\rangle_Y)^{-1} \leq a_0 \leq \langle a\rangle_Y$ (see [20]), there exist C_1 and C_2 such that

$$C_1|\xi|^2 \leq \xi_i a_0^{ij}\xi_j \leq C_2|\xi|^2. \qquad (3.6)$$

Moreover,

$$\int_K \nabla u \cdot \tilde{a}(\frac{x}{\varepsilon})\nabla u dx = \int_K \nabla u \cdot a_0 \nabla u dx + \int_K \nabla u \cdot (\tilde{a}(\frac{x}{\varepsilon}) - a_0)\nabla u dx. \qquad (3.7)$$

Dividing K into

$$K = \left(\bigcup_{Y_k \subset \Omega} Y_k\right)\bigcup K',$$

where Y_k is a periodic cell of $a(x/\varepsilon)$ and K' is the difference between K and the union of all Y_k in K. Since $\langle\tilde{a} - a_0\rangle_{Y_k} = 0$, we have for the second term on the right-hand side of (3.7)

$$\left|\int_K \nabla u \cdot (\tilde{a}(\frac{x}{\varepsilon}) - a_0)\nabla u dx\right| = \left|\int_{K'} \nabla u \cdot (\tilde{a}(\frac{x}{\varepsilon}) - a_0)\nabla u dx\right|$$
$$\leq \max|\tilde{a}^{ij} - a_0^{ij}|\|\nabla u\|_{0,K'}^2 \leq C\frac{\varepsilon}{h}\max|\tilde{a}^{ij} - a_0^{ij}|\|\nabla u\|_{0,K}^2. \qquad (3.8)$$

In deriving the last inequality, we have used the fact that ∇u is a constant in $K \in \mathcal{K}^h$ since $u \in \Phi_0^h$. From (3.6), (3.7) and (3.8), we see that when $\frac{\varepsilon}{h}$ is sufficiently small, (3.5) is valid. □

The inequality (3.5) is crucial for proving theorems in the next subsections. In the following, we always assume (3.5) is valid.

Remark 3.1. For some coefficients a, (3.5) is independent of h and ε. For example, it is easy to prove that this is the case if $a^{ij}(y) = \delta_{ij}(1 + \mu b(y))$, where $b(y)$ is a smooth periodic function and μ is small. Another example is that $a = a(x_1)$ only depends on one variable.

3.1. Removal of nonconforming error.   The main consequence of using a conforming test space is the following orthogonality result:

$$a_h(u_\varepsilon^h - u_\varepsilon, v) = 0, \qquad \forall v \in \Phi_0^h. \tag{3.9}$$

Eq. (3.9), together with the inf-sup condition (cf. [2]) proved below, enables one to bound the solution error by the interpolation error. Denote

$$\tilde{a}(u, v) = \sum_{K \in \mathcal{K}^h} \int_K \nabla v \cdot \tilde{a}(\frac{x}{\varepsilon}) \nabla u \, dx, \qquad v \in \Phi_0^h.$$

Thus, by (3.1) and the definition of $\tilde{a}$ we have

$$a_h(u_\varepsilon^h, v) = \tilde{a}(u_0^h, v) + \varepsilon a_h(\theta^h, v) = f(v), \qquad \forall v \in \Phi_0^h. \tag{3.10}$$

THEOREM 3.3. *Let u_ε^h be the P-G MsFEM solution. Assume that (3.5) is valid and Assumption 2.1 holds, if the homogenized part of u_ε, i.e. u_0 is in $H^2(\Omega)$, we have*

$$\|u_\varepsilon^h - u_\varepsilon\|_{h,\Omega} \le C_1 h + C_2 \varepsilon + C_3 \sqrt{\varepsilon}. \tag{3.11}$$

Proof. To estimate $\|u_\varepsilon^h - u_\varepsilon\|_{h,\Omega}$, we first show that the following inf-sup condition or coercive condition of the bilinear form $a_h(\cdot, \cdot)$ holds for sufficiently small ε. There exists $C > 0$, independent of ε and h such that

$$\sup_{v \in \Phi_0^h} \frac{|a_h(u_\varepsilon^h, v)|}{\|v\|_{1,\Omega}} \ge C \|u_\varepsilon^h\|_{h,\Omega}, \qquad \forall u_\varepsilon^h \in \Phi_\varepsilon^h. \tag{3.12}$$

Taking $v = u_0^h \in \Phi_0^h$ in (2.14) and using (3.10), we get

$$a_h(u_\varepsilon^h, u_0^h) = \tilde{a}(u_0^h, u_0^h) + \varepsilon a_h(\theta^h, u_0^h). \tag{3.13}$$

Moreover, we obtain using (3.2) that

$$|a_h(\theta^h, u_0^h)| \le C |\theta^h|_{h,\Omega} \|\nabla u_0^h\|_{0,\Omega} \le C \|\nabla u_0^h\|_{0,\Omega}^2. \tag{3.14}$$

It follows from (3.3), (3.13), (3.14), and (3.5) that when ε is sufficiently small

$$|a_h(u_\varepsilon^h, u_0^h)| \ge |\tilde{a}(u_0^h, u_0^h)| - \varepsilon |a_h(\theta^h, u_0^h)| \ge C \|\nabla u_0^h\|_{0,\Omega}^2 \ge C \|\nabla u_0^h\|_{0,\Omega} \|u_\varepsilon^h\|_{h,\Omega};$$

thus, (3.12) is valid.

Now, let $u_\varepsilon^I \in \Phi_\varepsilon^h$ be the interpolation of u_ε. Using (3.12) we have

$$
\begin{aligned}
\|u_\varepsilon^h - u_\varepsilon\|_{h,\Omega} &\le \|u_\varepsilon^I - u_\varepsilon\|_{h,\Omega} + \|u_\varepsilon^h - u_\varepsilon^I\|_{h,\Omega} \\
&\le \|u_\varepsilon^I - u_\varepsilon\|_{h,\Omega} + C \sup_{v \in \Phi_0^h} \frac{|a_h(u_\varepsilon^h - u_\varepsilon^I, v)|}{\|v\|_{1,\Omega}} \\
&= \|u_\varepsilon^I - u_\varepsilon\|_{h,\Omega} + C \sup_{v \in \Phi_0^h} \frac{|a_h(u_\varepsilon^I - u_\varepsilon, v)|}{\|v\|_{1,\Omega}} \\
&\le (1 + \beta C) \|u_\varepsilon^I - u_\varepsilon\|_{h,\Omega}.
\end{aligned}
\tag{3.15}
$$

We have used (3.9) in the third step of above derivation.

Following the derivation in [14] (page 924) and using (2.10), we can show that

$$\|u_\varepsilon^I - u_\varepsilon\|_{h,\Omega} \le C_1 h + C_2 \varepsilon + C_3 \sqrt{\varepsilon}.$$

Therefore, (3.11) follows from (3.15). □

Remark 3.2. The above proof shows that the error of the P-G MsFEM solution is bounded by the interpolation error. In fact, (3.15) is true for any $v_\varepsilon^h \in \Phi_\varepsilon^h$. Therefore, it can be viewed as a replacement of the classical Cea's lemma.

Remark 3.3. The $\sqrt{\varepsilon}$ term in (3.11) comes from the first order corrector for the global solution. This term is unavoidable. In practice, its effect is localized near the boundary; hence the solution in the interior is not affected.

As noted in [19], the conventional Aubin-Nitsche trick is not helpful in analyzing the L_2 convergence of MsFEM. A discrete method was introduced in [19] and was followed by [14]. The advantage of the method is that it can reveal the subtle error cancellations in the discrete solutions and hence gives a better convergence estimate. In the following, we use a simpler approach to show the error cancellation.

3.2. Error cancellation and more estimates. From the discussion in Section 2.1, we know that the homogenized solution $u_0 \in H_0^1(\Omega)$ satisfies the following variational formulation

$$a_0(u_0, v) = f(v), \qquad \forall v \in H_0^1(\Omega), \tag{3.16}$$

where

$$a_0(u, v) = \int_\Omega \nabla v \cdot a_0 \nabla u \, dx, \qquad u, v \in H_0^1(\Omega).$$

Therefore, by (3.10) we get

$$\tilde{a}(u_0^h, v) + \varepsilon a_h(\theta^h, v) = a_0(u_0, v), \qquad \forall v \in \Phi_0^h. \tag{3.17}$$

Below, we first derive the error estimate for $u_0 - u_0^h$. Then we apply this result to prove the L_2 and H^1 convergence of u_ε^h.

THEOREM 3.4. *If u_0 belongs to $H^3(\Omega) \bigcap W^{2,\infty}(\Omega)$ and $\partial\Omega$ is $C^{0,1}$, under (3.5) and Assumptions 2.1, the following estimate holds:*

$$\|u_0^h - u_0\|_{1,\Omega} \le C_1 \varepsilon + C_2 h + C_3 \varepsilon (\ln h)^{1/2}. \tag{3.18}$$

Proof. Let $u_0^I \in \Phi_0^h$ be the linear interpolant of u_0 and set $v = u_0^h - u_0^I$, (3.5) and (3.17) give

$$\begin{aligned}
C\|u_0^h - u_0^I\|_{1,\Omega}^2 \le \tilde{a}(v, v) &= \tilde{a}(u_0^h - u_0, v) + \tilde{a}(u_0 - u_0^I, v) \\
&= [a_0(u_0, v) - \tilde{a}(u_0, v)] - \varepsilon a_h(\theta^h, v) + \tilde{a}(u_0 - u_0^I, v).
\end{aligned} \tag{3.19}$$

For the second and the third terms on the right-hand side of the last equality, we have

$$|\varepsilon a_h(\theta^h, v)| \le C\varepsilon \|\theta^h\|_{h,\Omega} \|v\|_{1,\Omega} \le C\varepsilon \|u_0^h\|_{1,\Omega} \|v\|_{1,\Omega}; \tag{3.20}$$

$$|\tilde{a}(u_0 - u_0^I, v)| \le Ch|u_0|_{2,\Omega} \|v\|_{1,\Omega}. \tag{3.21}$$

In the above derivation we have used (3.2) and the estimate of the interpolation

$$\|u_0 - u_0^I\|_{1,\Omega} \le Ch|u_0|_{2,\Omega}.$$

Furthermore, it follows from (3.5), (3.2) and (3.10) that

$$\|u_0^h\|_{1,\Omega}^2 \le \quad C\tilde{a}(u_0^h, u_0^h) = Cf(u_0^h) - C\varepsilon a_h(\theta^h, u_0^h)$$
$$\le \quad C\|f\|_{0,\Omega}\|u_0^h\|_{1,\Omega} + C\varepsilon\|u_0^h\|_{1,\Omega}^2.$$

Since ε is small, we get

$$\|u_0^h\|_{1,\Omega} \le C\|f\|_{0,\Omega}. \tag{3.22}$$

Thus, (3.20) becomes

$$|\varepsilon a_h(\theta^h, v)| \le C\varepsilon\|f\|_{0,\Omega}\|v\|_{1,\Omega}. \tag{3.23}$$

For the first term on the right-hand side of (3.19), noting that $\nabla \cdot \tilde{a} = 0$ (Eq. (3.4)) and using integration by parts, we obtain

$$\tilde{a}(u_0, v) - a_0(u_0, v) \quad = \int_\Omega \nabla v \cdot \left(\tilde{a}(\frac{x}{\varepsilon}) - a_0\right) \nabla u_0 dx$$
$$= -\int_\Omega v\nabla \cdot \left((\tilde{a}(\frac{x}{\varepsilon}) - a_0)\nabla u_0\right) dx \tag{3.24}$$
$$= -\int_\Omega v\left[H_{u_0} \cdot \left(\tilde{a}(\frac{x}{\varepsilon}) - a_0\right)\right] dx$$

where

$$H_{u_0} = \left(\begin{array}{cc} (u_0)_{xx} & (u_0)_{yx} \\ (u_0)_{xy} & (u_0)_{yy} \end{array}\right).$$

Following the proof of Lemma 3.2 in [14], let

$$\Omega = \left(\bigcup_{Y_k \subset \Omega} Y_k\right)\bigcup\Omega',$$

where Y_k are small cells in Ω with length ε. Noting that $\langle\tilde{a} - a_0\rangle_{Y_k} = 0$ and letting $g_k = \langle g\rangle_{Y_k}$, for $g \in H^1(\Omega) \cap L^\infty(\Omega)$, we have

$$\left|\int_{Y_k} g(\tilde{a}^{ij} - a_0^{ij})dx\right| = \left|\int_{Y_k} (g - g_i)(\tilde{a}^{ij} - a_0^{ij})dx\right|$$
$$\le C\varepsilon\|\nabla g\|_{L_2(Y_k)}\|\tilde{a}^{ij} - a_0^{ij}\|_{L_2(Y_k)}. \tag{3.25}$$

Thus using (3.25), (3.24) gives

$$|\tilde{a}(u_0, v) - a_0(u_0, v)|$$
$$\le \quad \left|\sum_{Y_i \in \Omega}\int_{Y_i} v\left[H_{u_0} \cdot \left(\tilde{a}(\frac{x}{\varepsilon}) - a_0\right)\right] dx\right| + \left|\int_{\Omega'} v\left[H_{u_0} \cdot \left(\tilde{a}(\frac{x}{\varepsilon}) - a_0\right)\right] dx\right|$$
$$\le \quad C\varepsilon\|v\|_{\infty,\Omega}|u_0|_{3,\Omega} + C|\Omega'|\|v\|_{\infty,\Omega'}\|H_{u_0}\|_{\infty,\Omega'}.$$

Clearly, the area of Ω' is $O(\varepsilon)$. Therefore

$$
\begin{aligned}
|\tilde{a}(u_0, v) - a_0(u_0, v)| &\leq C\varepsilon \|v\|_{\infty,\Omega}(|u_0|_{3,\Omega} + \|H_{u_0}\|_{\infty,\Omega}) \\
&\leq C\varepsilon |\ln h|^{\frac{1}{2}} \|v\|_{1,\Omega}(|u_0|_{3,\Omega} + \|H_{u_0}\|_{\infty,\Omega}).
\end{aligned}
\tag{3.26}
$$

In the last step, we have used the following well-known inequality (see Lemma 2.3 of [7] and also [6])

$$
\|v\|_{\infty,\Omega} \leq C|\ln h|^{1/2}\|v\|_{1,\Omega}, \qquad v \in \Phi_0^h.
$$

Combining (3.23), (3.21), and (3.26) with (3.19), we get the desired estimate

$$
\|u_0^h - u_0\|_{1,\Omega} \leq C[\varepsilon(\|f\|_{0,\Omega} + (\ln h)^{1/2}\|H_{u_0}\|_{\infty,\Omega} + |u_0|_{3,\Omega}) + h|u_0|_{2,\Omega}].
$$

$\square$

Remark 3.4. In the above proof, the fact that $\tilde{a}$ is divergence free is the key to the additional error cancellation. It enables the application of Lemma 3.2 of [14] in Ω instead of K. Also, since $\varepsilon|\ln h|^{1/2} \ll \varepsilon^{1/2}$, (3.18) is a tighter bound than (3.11).

Remark 3.5. The estimate (3.18) can be used to derive the bound for $\|u_\varepsilon^h - u_\varepsilon\|_{h,\Omega}$ via the multiscale expansions. It is easy to check that the bound is essentially the same as in (3.18) with an additional term of $O(\sqrt{\varepsilon})$ due to θ^u.

THEOREM 3.5. *Under the same assumptions in Theorem 3.4, we have the following estimate*

$$
\|u_\varepsilon^h - u_\varepsilon\|_{0,\Omega} \leq C_1\varepsilon + C_2 h + C_3\varepsilon|\ln h|^{1/2},
\tag{3.27}
$$

Proof. Using Theorem 3.4 and the multiscale expansion of u_ε^h, (3.1), we get

$$
\begin{aligned}
\|u_\varepsilon^h - u_\varepsilon\|_{0,\Omega} &\leq \|u_\varepsilon^h - u_0\|_{0,\Omega} + \|u_0 - u_\varepsilon\|_{0,\Omega} \\
&\leq \|u_0^h - u_0\|_{0,\Omega} + \varepsilon\|\chi \cdot \nabla u_0^h\|_{0,\Omega} + \varepsilon\|\theta^h\|_{0,\Omega} + C\varepsilon \\
&\leq \|u_0^h - u_0\|_{1,\Omega} + C_1\varepsilon\|u_0^h\|_{1,\Omega} + C_2\varepsilon \\
&\leq C_1\varepsilon + C_2 h + C_3\varepsilon|\ln h|^{1/2}.
\end{aligned}
$$

In the above derivation, we have used the classical estimate $\|u_\varepsilon - u_0\|_{0,\Omega} \leq C\varepsilon$ [5] as well as (3.2), (3.22), and (3.18). $\square$

Remark 3.6. The divergence free structure of $\tilde{a}$ is a special feature of the above Petrov-Galerkin formulation. It has an obvious variant with the trial space and the test space switched. Thus, $\tilde{a}^T$ plays the role of $\tilde{a}$ in the new formulation. Noting that $\langle \tilde{a}^T \rangle_Y = a_0$, the solution of the variant formulation converges to u_0. Nonetheless, because $\nabla \cdot \tilde{a}^T \neq 0$, the extra error cancellation does not occur, from which worse convergence rates result. The same conclusion can be obtained by observing that the non-conforming error is not zero. This observation has been confirmed by our numerical tests.

Remark 3.7. It is also interesting to compare our P-G MsFEM formulation with traditional numerical upscaling methods. In these methods, the discrete effective coefficients are first computed on grid blocks or elements. Then standard finite volume or finite element methods are used to obtain the effective equations at the discrete

level. See [24] for an overview of the various formulations for computing the effective coefficients. Here, on each element K we define

$$a_0^h \langle \nabla \phi_\varepsilon \rangle_K = \langle a_\varepsilon \nabla \phi_\varepsilon \rangle_K, \qquad (3.28)$$

where $\phi_\varepsilon = (\phi_\varepsilon^i)$ $(i = 1, 2)$ and $\phi_\varepsilon^i \in \Phi_\varepsilon^h$ are the multiscale base functions on K. Note that the third base function on K is not used in (3.28) because its gradient depends linearly on those of the other two base functions. Term a_0^h is well defined if $\langle \nabla \phi_\varepsilon \rangle_K^{-1}$ exists (see [25] and below). Eq. (3.28) is consistent with the effective permeability defined in [24] and [25]. The effective finite element solution $\hat{u}^h \in \Phi_0^h$ can be obtained by solving

$$a_0^h(\hat{u}^h, v) \equiv \int_\Omega \nabla v \cdot a_0^h \nabla \hat{u}^h dx = \int_\Omega f v dx, \qquad \forall v \in \Phi_0^h.$$

Using the multiscale expansion of ϕ_ε, it is easy to see that $\langle \nabla \phi_\varepsilon \rangle_K^{-1}$ exists when ε/h is small. Moreover, it can be shown that, to the leading order, we have

$$a_0^h(\hat{u}^h, v) = \tilde{a}(\hat{u}^h, v) - (\tilde{a}\langle \nabla_y \chi \rangle_K \cdot \nabla \hat{u}^h, \nabla v) + \ldots = f(v).$$

The leading order term of this equation is identical to that of (3.10). However, there is a $O(\varepsilon/h)$ difference between the two formulations due to the term with $\langle \nabla_y \chi \rangle_K$. Therefore, although $\tilde{a}$ leads to the cancellation of resonance error, the $\langle \nabla_y \chi \rangle_K$ term brings the error back. It can be shown that

$$\|\hat{u}^h - u_0\|_{1,\Omega} \leq C\varepsilon/h.$$

Since $\langle \nabla_y \chi \rangle_K$ comes from the expansion of $\nabla \phi_\varepsilon$, we may avoid this additional source of error by defining a_0^h as

$$a_0^h = \langle a_\varepsilon \nabla \phi_\varepsilon \rangle_K (\nabla \phi_0)^{-1}.$$

Note that both definitions of a_0^h given in this remark are $O(\varepsilon/h)$ accurate approximations to a_0 and yet they lead to different error in the solutions.

4. Numerical Test

In this section, we present some numerical examples to demonstrate the theoretical analysis. The resolved solutions are obtained by using standard FEM with piecewise linear elements. Given ε, we solve the model problem twice on two meshes with one mesh size being twice that of the other. Then the Richardson extrapolation is used to approximate the exact solutions from the numerical solutions on the two meshes. Since both of the mesh sizes used to compute the well-resolved solution are less than $\varepsilon/10$, the error of the extrapolated solutions is less than 10^{-7}.

We use the same periodic test problem as in [17] and [14]

$$a_\varepsilon = \left(\frac{2 + P\sin(2\pi x/\varepsilon)}{2 + P\cos(2\pi y/\varepsilon)} + \frac{2 + \sin(2\pi y/\varepsilon)}{2 + P\sin(2\pi x/\varepsilon)} \right) I$$

with $P = 1.8$, $f = -1$, and I being the identity matrix. It is easy to compute the homogenized coefficient numerically from (2.4) and (2.3):

$$a_0 \approx \begin{pmatrix} 3.946 & 0 \\ 0 & 3.342 \end{pmatrix}.$$

We have also performed numerical experiments with other forms of a_ε. The results are similar and hence are not included.

The convergence of P-G MsFEM with different fixed parameters are shown in Tables 4.1-4.4. Table 4.1 shows the first order convergence of the method for fixed $h/\varepsilon = 1.5$. As shown in [14], MsFEM failed to converge for the same problem due to the cell resonance. However, the cell resonance is not a problem here. In Table 4.2, the convergence of P-G MsFEM with respect to h with fixed ε is shown. The convergence deteriorates slightly as h decreases. This is expected from Theorem 3.4. In fact, the table indicates that the error constant of the $O(\varepsilon)$ term (i.e., C_1 in (3.18)) is smaller than that of the $O(h)$ term.

The L_2 and H^1 convergence of u_ε^h to u_ε are presented in Tables 4.3 and 4.4. The convergence of the numerical solution for fixed h/ε in Table 4.3 is similar to that in Table 4.1, which is consistent with Theorem 3.5. Numerical tests for smaller h could not be carried out because the well-resolved solution of u_ε could not be obtained on the computing resource (a 512-node Intel Paragon computer) available to us. The results for fixed ε shown in Table 4.4 are also consistent with the theorem. However, compared to Table 4.2, Table 4.4 indicates a stronger $O(\varepsilon)$ term in $\|u_\varepsilon^h - u_\varepsilon\|_{0,\Omega}$ than in the $\|u_0^h - u_0\|_{1,\Omega}$.

TABLE 4.1. $\|u_0^h - u_0\|_{1,\Omega}$. *Keep* $h/\varepsilon = 1.5$.

h	$\|u_0^h - u_0\|_{1,\Omega}$	rate
1/16	$5.69e-3$	
1/32	$2.88e-3$	0.98
1/64	$1.43e-3$	1.01
1/128	$7.08e-4$	1.01
1/256	$3.54e-4$	1.00
1/512	$1.98e-4$	0.84

TABLE 4.2. $\|u_0^h - u_0\|_{1,\Omega}$.

h	$\varepsilon = 0.02$		$\varepsilon = 0.01$	
	$\|u_0^h - u_0\|_{1,\Omega}$	rate	$\|u_0^h - u_0\|_{1,\Omega}$	rate
1/16	$5.59e-3$		$5.53e-3$	
1/32	$2.83e-3$	0.98	$2.76e-3$	1.00
1/64	$1.48e-3$	0.93	$1.38e-3$	1.00
1/128	$1.41e-3$	0.07	$8.32e-4$	0.73

TABLE 4.3. $\|u_\varepsilon^h - u_\varepsilon\|_{0,\Omega}$. *Keep* $h/\varepsilon = 1.5$.

h	$\|u_\varepsilon^h - u_\varepsilon\|_{0,\Omega}$	*rate*
1/16	$1.96e-4$	
1/32	$8.31e-5$	1.24
1/64	$4.01e-5$	1.05
1/128	$2.14e-5$	0.91

TABLE 4.4. $\|u_\varepsilon^h - u_\varepsilon\|_{1,\Omega}$.

h	$\varepsilon = 0.02$		$\varepsilon = 0.01$	
	$\|u_\varepsilon^h - u_\varepsilon\|_{1,\Omega}$	rate	$\|u_\varepsilon^h - u_\varepsilon\|_{1,\Omega}$	rate
$1/16$	$1.79e-4$		$1.89e-4$	
$1/32$	$9.56e-5$	0.90	$9.35e-5$	1.01
$1/64$	$8.21e-5$	0.22	$7.20e-5$	0.37
$1/128$	$7.94e-5$	0.05	$6.74e-5$	0.10

5. Petrov-Galerkin formulation for interface problems

In this section, we apply the idea of combining non-conforming trial space and conforming test space to solving elliptic interface problems. Let Ω be a convex domain in R^2 and $\Omega_1 \subset \Omega$ be an open domain with C^2 boundary $\Gamma = \partial\Omega_1$. Let $\Omega_2 = \Omega - \Omega_1$. The elliptic interface problem is given by

$$-\nabla \cdot (\kappa \nabla u) = f \qquad \text{in } \Omega, \tag{5.1}$$

where $\kappa(x) \geq \kappa_0 > 0$ is a piecewise smooth scalar function, i.e., $\kappa = \kappa_1$ and κ_2, respectively, in Ω_1 and Ω_2 (see Figure 5.1). The following jump condition is satisfied on the interface Γ

$$[u] = 0, \quad \left[\kappa\frac{\partial u}{\partial n}\right] = g \qquad \text{across } \Gamma, \tag{5.2}$$

where $[v](x) = v_1(x) - v_2(x)$, $x \in \Gamma$, with v_1 and v_2 being the restrictions of v on Ω_1 and Ω_2, respectively. In the following we assume

$$u = 0 \qquad \text{on } \partial\Omega. \tag{5.3}$$

The weak formulation of the interface problem (5.1)-(5.3) is given as follows. Find $u \in H_0^1(\Omega)$ such that

$$a(u,v) = (f,v) + \langle g, v \rangle, \qquad \forall v \in H_0^1(\Omega), \tag{5.4}$$

where

$$a(u,v) = \int_\Omega \kappa(x)\nabla u \cdot \nabla v\, dx, \qquad \forall u, v \in H^1(\Omega),$$

and $(\cdot, \cdot)$ and $\langle \cdot, \cdot \rangle$ are used to denote the inner products of the $L^2(\Omega)$ space and the $L^2(\Gamma)$ space, respectively. In the following, we consider the problem with κ being a piecewise constant (i.e., κ_1 and κ_2 are constants). The extension to variable κ is straightforward.

Here, we are interested in solving (5.1)-(5.3) using a finite element method on grids that are not aligned with the interface (i.e., a non-interface-fitting), e.g., a Cartesian grid. The method is very useful for solving problems with moving interfaces. See [22] for a review of Cartesian grid methods and a discussion of their advantages.

Let $\mathcal{K}^h$ be a non-interface fitting triangulation of Ω. In the following we assume that the triangulation is sufficiently fine so that the interface intersects with the boundary of a triangle at either one vertex, two vertices, or two edges, each edge
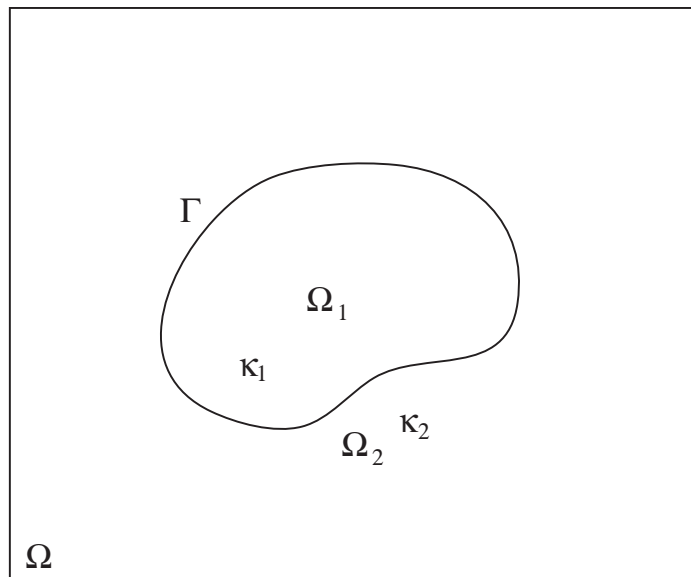
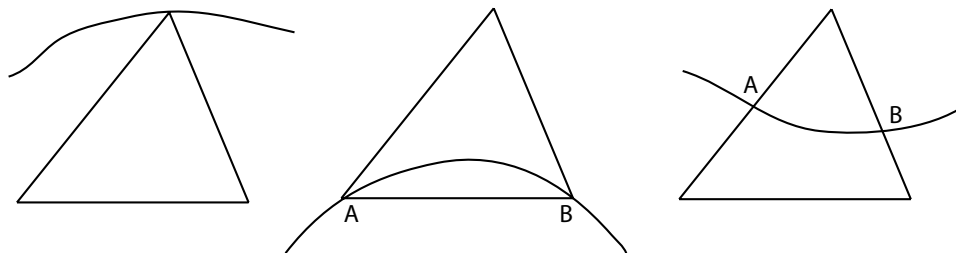FIG. 5.1. *A 2D domain Ω with a smooth interface Γ.*

FIG. 5.2. *Possible intersections between the interface Γ and the triangles in a sufficiently fine triangulation of Ω.*

having one intersection point, as shown in Fig. 5.2. The interface Γ is approximated by the line segment $\overline{AB}$ in the second and third triangles in the figure. The union of such line segments form an approximate interface Γ^h, which divides Ω into Ω_1^h and Ω_2^h, approximations of Ω_1 and Ω_2, respectively. Furthermore, we denote the set of $K \in \mathcal{K}^h$ with the third type of intersection (e.g., the right most triangle in Fig. 5.2) $\mathcal{K}_\Gamma^h$. Note that when the grid fits the interface, only the first two types of triangles exist, and the standard finite element method with piecewise linear base functions works well (see e.g., [10]). However, it is well-known that the linear base is not suitable for $K \in \mathcal{K}_\Gamma^h$; their use leads $O(h)$ convergence of the finite element solution. The reason is that the flux jump condition in (5.2) is not satisfied by the linear base functions. In [22], special base functions were constructed for $K \in \mathcal{K}_\Gamma^h$ so that that they satisfy the jump conditions approximately.

More specifically, let ϕ_Γ^i $(i = 1, 2, 3)$ be the base functions for an element $K \in \mathcal{K}_\Gamma^h$ (see Fig. 5.3). We require ϕ_Γ^i be piecewise linear on K_1 and K_2, $\phi_\Gamma^i(x_j) = \delta_{ij}$, and
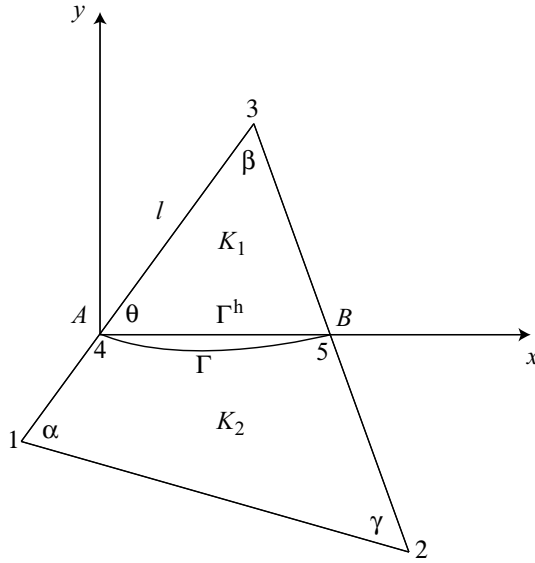
FIG. 5.3. *An element $K = \triangle_{123} \in \mathcal{K}_\Gamma^h$. Γ is approximated by the line segment $\overline{AB}$, which splits K into two pieces, $K_1 = \triangle_{345} \subset K \cap \Omega_1^h$ and $K_2 = K - K_1$.*

satisfy the following homogeneous jump condition:

$$[\phi_\Gamma^i] = 0, \quad \left[\kappa^h \frac{\partial \phi_\Gamma^i}{\partial n}\right] = 0 \qquad \text{across } \Gamma^h, \tag{5.5}$$

where $\kappa^h = \kappa_1$ and κ_2 in Ω_1^h and Ω_2^h, respectively. This construction always exists (Theorem 2.1 of [22]). Moreover, if $u(x)$ is piecewise twice differentiable in Ω_1 and Ω_2, then the maximum interpolation errors for u and ∇u using ϕ_Γ^i are $O(h^2)$ and $O(h)$, respectively (Theorem 2.2 and 2.3 of [22]). Note that the $O(h)$ estimate for ∇u was obtained outside the small region $K - (\Omega_1 \cap K_1) - (\Omega_2 \cap K_2)$ (see the region between Γ and Γ^h in Fig. 5.3), whose size is $O(h^3)$. For a picture of the bases that satisfy (5.5), see [22].

Note that the base functions ϕ_Γ^i are non-conforming; they are discontinuous at the element edges that intersect with Γ. Using these bases and linear bases for $K \notin \mathcal{K}_\Gamma^h$ for both the trial and test spaces, one obtains a non-conforming Galerkin method for the interface problem (5.4). This method is a significant improvement over the traditional method without using the special base functions; see Tables 5.1 to 5.4 below. The method does suffer from the non-conforming error, which is localized near the interface. Again, we use a conforming piecewise linear test space to remove the non-conforming error and obtain a Petrov-Galerkin formulation with higher order convergence rate.

Let S^h be the space spanned by ϕ_Γ^i for $K \in \mathcal{K}_\Gamma^h$ and by linear base functions on $K \in \mathcal{K}^h - \mathcal{K}_\Gamma^h$, and let $S_0^h = \{s \in S^h; s|_{\partial\Omega} = 0\}$. Moreover, denote $\Phi_0^h \subset H_0^1(\Omega)$ the space of piecewise linear functions. We now have the non-conforming Petrov-Galerkin finite element method for solving (5.4): find $u^h \in S_0^h$ such that

$$a_h(u^h, v^h) = (f, v^h), \qquad \forall v^h \in \Phi_0^h,$$

TABLE 5.1. *Maximum error* ($\kappa_1 = 1000, \; \kappa_2 = 1$)

h	GFEM-L	GFEM-J	ratio	PGFEM-JL	ratio
1/32	3.55e-2	1.83e-3		3.75e-3	
1/64	1.50e-2	6.40e-4	2.85	1.03e-3	3.66
1/128	7.86e-3	2.33e-4	2.74	2.26e-4	4.53
1/256	3.91e-3	8.84e-5	2.64	5.41e-5	4.13
1/512	1.97e-3	4.47e-5	1.98	1.35e-5	4.06
1/1024	9.92e-4	2.08e-5	2.14	3.58e-6	3.77

where

$$a_h(u,v) = \sum_{K \in \mathcal{K}^h} \int_K \kappa^h \nabla u \cdot \nabla v dx. \tag{5.6}$$

This method can be analyzed in a way similar to analyzing the multiscale Petrov-Galerkin method in Section 3.1. The main task is to establish the inf-sup condition, i.e., $\forall u^h \in S_0^h$, there exist $v^h \in \Phi_0^h$ and $C > 0$, independent of h, such that

$$a_h(u^h, v^h) \geq C\|u^h\|_{h,\Omega}\|v^h\|_{1,\Omega}. \tag{5.7}$$

We note that a similar method has been proposed and studied for problems with rough coefficients that vary rapidly along a given direction [1]. The inf-sup condition was proved under the assumption that all triangular elements have one edge perpendicular to the direction in which the rough coefficient varies. This assumption is invalid in the present case, and the inf-sup condition becomes difficult to obtain. We can prove (5.7) when κ^h has a mild jump. The proof is not included because it is not general and yet rather technical. On the other hand, our numerical experiments below show that large jumps are handled by the method without a problem. A more general proof is currently being studied.

In the following, we present numerical tests of the above Petrov-Galerkin formulation using the exact solution of (5.1) as given by Eq. (2.40) of [22]:

$$u(x,y) = \begin{cases} r^a/\kappa_1, & r \leq r_0, \\ r^a/\kappa_2 + (1/\kappa_1 - 1/\kappa_2)r_0^a, & r > r_0, \end{cases} \tag{5.8}$$

where $r = \sqrt{x^2 + y^2}$, $a = 3$, $r_0 = \pi/6.28$ is the radius of a circular interface centered at the origin. The computational domain is the rectangle $-1 \leq x, y \leq 1$. The triangulation of the computational domain is obtained by connecting the south-west and north-east corners of the rectangular grid cells of a uniform Cartesian grid. Using (5.8), we can obtain the source term f in (5.1) and the Dirichlet boundary condition. We compare the performance of three different methods: traditional Galerkin finite element method with linear trial and test spaces (GFEM-L), the non-conforming Galerkin finite element method with base functions satisfying the jump conditions (GFEM-J), and the above Petrov-Galerkin method (PGFEM-JL).

TABLE 5.2. l^2 error ($\kappa_1 = 1000$, $\kappa_2 = 1$)

h	GFEM-L	GFEM-J	ratio	PGFEM-JL	ratio
1/32	1.33e-2	7.04e-4		5.84e-4	
1/64	6.95e-3	1.77e-4	3.98	1.64e-4	3.57
1/128	3.60e-3	4.57e-5	3.86	3.37e-5	4.86
1/256	1.79e-3	1.19e-5	3.86	8.72e-6	3.86
1/512	8.98e-4	3.27e-6	3.62	2.18e-6	4.00
1/1024	4.51e-4	9.49e-7	3.45	5.51e-7	3.96

TABLE 5.3. Maximum error ($\kappa_1 = 1$, $\kappa_2 = 1000$)

h	GFEM-L	GFEM-J	ratio	PGFEM-JL	ratio
1/32	4.12e-2	1.18e-3		2.03e-3	
1/64	2.49e-2	3.65e-4	3.23	6.06e-4	3.36
1/128	1.31e-3	1.34e-4	2.72	2.03e-4	2.98
1/256	6.76e-3	5.81e-5	2.31	5.19e-5	3.90
1/512	3.46e-3	2.85e-5	2.04	1.41e-5	3.69
1/1024	1.74e-4	1.52e-5	1.88	3.81e-6	3.70

TABLE 5.4. l^2 error ($\kappa_1 = 1$, $\kappa_2 = 1000$)

h	GFEM-L	GFEM-J	ratio	PGFEM-JL	ratio
1/32	1.37e-2	1.87e-4		6.03e-4	
1/64	7.44e-3	6.28e-5	2.98	1.93e-4	3.12
1/128	3.38e-3	1.57e-5	3.99	4.93e-5	3.92
1/256	1.74e-3	4.57e-6	3.44	1.29e-5	3.81
1/512	8.72e-4	1.42e-6	3.22	3.21e-6	4.04
1/1024	4.36e-4	5.11e-7	2.78	8.26e-7	3.88

In Tables 5.1 to 5.4, the maximum and discrete l^2 norm errors are reported for two test cases with different κ_2/κ_1 ratios. As shown by the tables, GFEM-L is about first order accurate in both maximum and l^2 norms. GFEM-J provides a significant improvement over GFEM-L, especially in the l^2 norm. This improvement is due to the base functions that satisfy the jump conditions at the (discrete) interface. However, the convergence of GFEM-J in the maximum norm decreases to about first order as the grid refines. This behavior can be attributed to the non-conforming error near the interface. Indeed, by plotting the error distribtion, we find that the error along the interface is much larger than the error just slightly away from the interface. By removing the non-conforming error with PGFEM-JL, the error becomes more uniformly distributed and better convergence rates are obtained. We do notice that the error constant for PGFEM-JL appears to be larger than that of GFEM-J, a trade-off for using different trial and test spaces; a similar phenomenon has been observed in [1]. Finally, we remark that the oscillatory behavior in the convergence rate of PGFEM-JL is expected, since the intersection between the interface and the grid (i.e., angle and position) changes as grid refines.

Appendix A. Convergence of classical FEM for (2.1)**.**

THEOREM A.1. *Suppose $\overline{u}^h \in \Phi_0^h$ and $\overline{u}$ are the solutions of the variational problem*

$$\int_\Omega \nabla \overline{u}^h \cdot a_\varepsilon \nabla v dx = \int_\Omega f v dx, \qquad v \in \Phi_0^h, \tag{A.1}$$

and

$$\int_\Omega \nabla \overline{u} \cdot \overline{a} \nabla v dx = \int_\Omega f v dx, \qquad v \in H_0^1(\Omega), \tag{A.2}$$

respectively, where $\overline{a}$ is the mean value of a, i.e., $\overline{a} = \langle a \rangle_Y$. Moreover, assume $\overline{u} \in H^2(\Omega) \cap W^{1,\infty}(\Omega)$, then we have the estimate

$$\|\overline{u}^h - \overline{u}\|_{1,\Omega} \le C_1 \frac{\varepsilon}{h} + C_2 h.$$

Proof. Let $\overline{u}^I \in \Phi_0^h$ be the interpolation of $\overline{u}$. Assigning $\overline{u}^h - \overline{u}^I$ to v in (A.1) and (A.2), and subtracting (A.2) from (A.1), we obtain

$$\int_\Omega \nabla(\overline{u}^h - \overline{u}) \cdot a_\varepsilon \nabla(\overline{u}^h - \overline{u}^I) dx = -\int_\Omega \nabla \overline{u} \cdot (a_\varepsilon - \overline{a}) \nabla(\overline{u}^h - \overline{u}^I) dx. \tag{A.3}$$

We first examine the difference between $\overline{u}^h$ and $\overline{u}^I$. Using (A.3), Lemma 3.2 in [14], together with the Poincare-Friedrichs inequality, we get

$$\|\overline{u}^h - \overline{u}^I\|_{1,\Omega}^2 \le C \int_\Omega \nabla(\overline{u}^h - \overline{u}^I) \cdot a_\varepsilon \nabla(\overline{u}^h - \overline{u}^I) dx$$

$$= C \left(\int_\Omega \nabla(\overline{u}^h - \overline{u}) \cdot a_\varepsilon \nabla(\overline{u}^h - \overline{u}^I) dx + \int_\Omega \nabla(\overline{u} - \overline{u}^I) \cdot a_\varepsilon \nabla(\overline{u}^h - \overline{u}^I) dx \right)$$

$$= C \left(-\int_\Omega \nabla \overline{u} \cdot (a_\varepsilon - \overline{a}) \nabla(\overline{u}^h - \overline{u}^I) dx + \int_\Omega \nabla(\overline{u} - \overline{u}^I) \cdot a_\varepsilon \nabla(\overline{u}^h - \overline{u}^I) dx \right)$$

$$\le C \left(\frac{\varepsilon}{h}(|\overline{u}|_{2,\Omega} + \|\nabla \overline{u}\|_{\infty,\Omega}) + \|\overline{u} - \overline{u}^I\|_{1,\Omega} \right) \|\overline{u}^h - \overline{u}^I\|_{1,\Omega}.$$

Combining with the classical result of the finite element theory

$$\|\overline{u} - \overline{u}^I\|_{1,\Omega} \le Ch\|\overline{u}\|_{2,\Omega},$$

we get the desired result

$$\|\overline{u}^h - \overline{u}\|_{1,\Omega} \le \|\overline{u}^h - \overline{u}\|_{1,\Omega} + \|\overline{u} - \overline{u}^I\|_{1,\Omega} \le C \left(\frac{\varepsilon}{h} + h \right). \qquad \square$$

REFERENCES

[1] I. Babuška, G. Caloz, and E. Osborn, *Special finite element methods for a class of second order elliptic problems with rough coefficients*, SIAM J. Numer. Anal., 31:945–981, 1994.

[2] I. Babuska and A.K. Aziz, *Survey Lectures on the Mathematical Foundations of the Finite Element Method with Applications*, 3–362, Academic Press, 1972.

[3] J.W. Barrett and K.W. Morton, *Approximate symmetrization and Petrov-Gallerkin methods for diffusion-convection problems*, Comput. Methods Appl. Mech. Engrg., 45:97–122, 1984.

[4] A. Bensoussan, J.L. Lion, and G. Papanicolaou, *Boundary layer analysis in homogenization of diffusion equations with dirichlet conditions in the half space*, In K. Ito, editor, Proceedings of the International Symposium on Stochastic Differential Equations, 21–40, Kyoto, 1976. Wiley.

[5] A. Bensoussan, J.L. Lion, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structure*, volume 5 of Studies in Mathematics and Its Applications, North-Holland Publ., 1978.

[6] J. Bramble, *A second order finite difference analog of the first biharmonic boundary value problem*, Numer. Math., 9:236–249, 1966.

[7] J.H. Bramble and J. Xu, *Some estimates for a weighted l^2 projection*, Math. Comp., 56:463–476, 1991.

[8] M.A. Celia, T.F. Russell, I. Herrera, and R. E. Ewing, *An Eulerian-Lagrangian localized adjoint method for the advection-diffusion equation*, Adv. in Water Res., 13:187–206, 1990.

[9] Z. Chen and T.Y. Hou, *A mixed finite element method for elliptic problems with rapidly oscillating coefficients*, Math. Comput., 72(242):541–576, 2002.

[10] Z. Chen and J. Zou, *Finite element methods and their convergence for elliptic and parabolic interface problems*, Numer. Math., 79:175–202, 1997.

[11] I. Christie, D.F. Griffiths, A.R. Mitchell, and O.C. Zienkiewicz, *Finite element methods for second order differential equations with significant first derivatives*, Internat. J. Numer. Engrg., 10:1389–1396, 1976.

[12] J. Douglas and T.F. Russel, *Numerical methods for convection-dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures*, SIAM J. Numer. Anal., 19:871–885, 1982.

[13] L.J. Durlofsky, *Numerical-calculation of equivalent grid block permeability tensors for heterogeneous porous media*, Water Resour. Res., 27:699–708, 1991.

[14] Y.R. Efendiev, T.Y. Hou, and X.H. Wu, *Convergence of a nonconforming multiscale finite element method*, SIAM J. Numer. Anal., 37:888–910, 2000.

[15] A. Fannjiang and G. Papanicolaou, *Convection enhanced diffusion for periodic flows*, SIAM J. Appl. Math., 54:333–408, 1994.

[16] D.F. Griffiths and J. Lorenz, *An analysis of the Petrov-Gallerkin finite element method*, Comput. Methods Appl. Mech. Engrg., 14:39–64, 1978.

[17] T.Y. Hou and X.H. Wu, *A multiscale finite element method for elliptic problems in composite materials and porous media*, J. Comput. Phys., 134:169–189, 1997.

[18] T.Y. Hou and X.H. Wu, *A multiscale finite element method for PDEs with oscillatory coefficients*, Notes on Numerical Fluid Mechanics, 70:58–69, Vieweg-Verlag, 1999.

[19] T.Y. Hou, X.H. Wu, and Z. Cai, *Convergence of a multiscale finite element method for elliptic problems with rapidly oscillating coefficients*, Math. Comp., 68:913–943, 1999.

[20] V.V. Jikov, S.M. Kozlov, and O.A. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, 1994.

[21] C. Johnson, *Numerical Solution of Partial Differential Equations by the Finite Element Method*, Cambridge University Press, 1987.

[22] Z. Li, T. Lin, and X.H. Wu, *New cartesian grid methods for interface problems using finite element formulation*, Numer. Math., 2003. to appear.

[23] P. Park, *Multiscale finite element method for transport equations*, PhD thesis, Caltech, 1999.

[24] X.H. Wen and J.J. Gómez-Hernández, *Upscaling hydraulic conductivities in heterogeneous media: An overview*, Journal of Hydrology, 183:9–32, 1996.

[25] X.H. Wu, Y.R. Efendiev, and T.Y. Hou, *Analysis of upscaling absolute permeability*, Discrete and Continuous Dynamical Systems, Series B, 2:185–204, 2002.