이학박사 학위논문

P_1–Nonconforming Quadrilateral Finite Space with Periodic Boundary Condition and Its Application to Multiscale Problems

주기경계조건을 갖는 P_1–비순응유한요소공간과 멀티스케일 문제에 대한 응용

2018 년 2 월

서울대학교 대학원

협동과정 계산과학전공

임 재 륜

P_1–Nonconforming Quadrilateral Finite Space with Periodic Boundary Condition and Its Application to Multiscale Problems

주기경계조건을 갖는 P_1–비순응유한요소공간과 멀티스케일 문제에 대한 응용

지도교수 신 동 우

이 논문을 이학박사 학위논문으로 제출함

2017 년 11 월

서울대학교 대학원

협동과정 계산과학전공

임 재 륜

임 재 륜의 이학박사 학위논문을 인준함

2017 년 12 월

위 원 장	정 상 권	(인)
부위원장	신 동 우	(인)
위 원	전 영 목	(인)
위 원	이 종 우	(인)
위 원	박 춘 재	(인)

Abstract

P_1–Nonconforming Quadrilateral Finite Space with Periodic Boundary Condition and Its Application to Multiscale Problems

Jaeryun Yim
The Interdisciplinary Program in
Computational Science and Technology
The Graduate School
Seoul National University

We consider the P_1–nonconforming quadrilateral finite space with periodic boundary condition, and investigate characteristics of the finite space and discrete Laplace operators in the first part of this dissertation. We analyze dimension of the finite element spaces in help of concept of minimally essential discrete boundary conditions. Based on the analysis, we classify functions in a basis for the finite space with periodic boundary condition into two types. And we introduce several Krylov iterative schemes to solve second-order elliptic problems, and compare their solutions. Some of the schemes are based on the Drazin inverse, one of generalized inverse operators, since the periodic nature may derive a singular linear system of equations. An application to the Stokes equations with periodic boundary condition is considered. Lastly, we extend our results for elliptic problems to 3-D case. Some numerical results are provided in our discussion.

In the second part, we introduce a nonconforming heterogeneous multiscale method for multiscale problems. Its formulation is based on the P_1–nonconforming quadrilateral finite element, mainly with periodic boundary

condition. We analyze a priori error estimates of the proposed scheme by following general framework for the finite element heterogeneous multiscale method. For numerical implementations, we use one of the proposed iterative schemes for singular linear systems in the previous part. Several numerical examples and results are given.

Keywords: P_1–nonconforming quadrilateral finite element, periodic boundary condition, minimally essential discrete boundary conditions, singular linear system, Drazin inverse, heterogeneous multiscale method, numerical homogenization

Student Number: 2012-20414

Contents

List of Figures

List of Tables

Part I

P_1–Nonconforming Quadrilateral Finite Space with Periodic Boundary Condition

Chapter 1

Introduction

After the P_1–nonconforming quadrilateral finite element was introduced in [44], there have been a lot of studies about this finite element for fluid dynamics, elasticity, electromagnetics [35, 27, 42, 40, 43, 47, 16, 28]. Most of those works are focused on the finite element space with Dirichlet and/or Neumann boundary conditions. Altmann and Carstensen [7] show the dimension of, and a basis for the finite element space with inhomogeneous Dirichlet boundary conditions which share similar discrete nature with Neumann boundary case. On the other hand, the finite element space with periodic boundary condition has not been investigated more than other boundary conditions. For instance, it is not known that the dimension of the finite space with periodic boundary condition as well as its basis functions.

In many cases, the solution of periodic problem is unique upto additive constant. The discrete formulation of such problem yields a corresponding matrix system which is singular. In a mathematical theory, we can deal a

singular matrix system using generalized inverses. There are various kinds of generalized inverses of a matrix. We concentrate on the Drazin inverse which is one of them. One of the most important properties which the Drazin inverse of a matrix satisfies is the expressibility as a polynomial in the given matrix. As well known, the Krylov iterative method for a nonsingular matrix equation is established on this property. The Krylov scheme can be applied to a singular matrix system as well under proper consistency conditions [31, 34, 50, 15, 8, 9].

In this thesis, we mainly investigate the P_1–nonconforming quadrilateral finite element spaces with periodic boundary condition. In chapter 2, we give brief explanation for the P_1–nonconforming quadrilateral finite element and the Drazin inverse. We investigate the dimension of the finite spaces with various boundary conditions, including periodic condition which is our main concern, in chapter 3. For the analysis, we introduce the concept of *minimally essential discrete boundary conditions* to understand precise effect of given boundary condition on the dimension of the corresponding finite space. In chapter 4, we discuss a basis for the finite space, of which the majority are node based functions after identification between boundary nodes. And a complementary basis consisting of a few alternating functions is considered. After that, we propose several numerical schemes for solving a second-order elliptic problem with periodic boundary condition. Each scheme may give a solution of a singular matrix equation corresponding to the weak formulation. We use an efficient iterative method based on the Krylov space in help of the Drazin inverse of the corresponding singular matrix. The relationship between solutions of the schemes will be discussed. We apply this approach to the Stokes equations with periodic boundary condition in chapter 5. The discrete stability of the formulation is proved based on the result of the Dirichlet boundary case. Based on the Drazin inverse, we introduce a variant of Uzawa

method for a singular indefinite system with a positive semi-definite block on diagonal. Finally, we extend all our results for the elliptic problem to 3-D case in chapter 6.

Chapter 2

Preliminaries

2.1 P_1–nonconforming quadrilateral finite element

The P_1–nonconforming quadrilateral finite space in $\mathbb{R}^d$ is a set of all piecewise linear polynomials on a quadrilateral mesh ($d = 2$) or a hexahedral mesh ($d = 3$), which fulfill the integral-continuity across all $(d-1)$-dimensional interior faces. The integral-continuity is described precisely as follows: if f is a $(d-1)$-dimensional face which is shared by two adjacent elements K^+ and K^-, then every function v in the finite space satisfies $\int_f v|_{K^+} = \int_f v|_{K^-}$. Since we consider piecewise linear functions, the above relation is equivalent to the continuity of function at the midpoint ($d = 2$) or at the center point ($d = 3$, parallelepipedal mesh) of f. Thus degrees of freedom (DoFs) of the P_1–nonconforming quadrilateral finite element are function values at the midpoints (or center points) of all $(d-1)$-dimensional faces.

There are 4 midpoints in a quadrilateral, and 6 center points in a parallelepiped. As mentioned above, the value at each midpoint (or center point)

corresponds to DoF of the finite element. On the other hand, just $(d+1)$ coefficients are enough to determine a unique linear function in a d-dimensional space. Such difference concludes the existence of a linear relation between DoFs in local, so called, *the dice rule*. For a given linear function which is defined in a quadrilateral in 2-D space, the sum of two function values at the midpoints of the edge pair on opposite sides is always equal to the sum of those at the midpoints of the other edge pair. An analog relation in 3-D space holds, as an ordinary dice.

Due to the dice rule, a set of specially designed functions is used to construct a global basis for the finite space with Dirichlet or Neumann boundary conditions. Since each of them corresponds a node in the triangulation, we call them *node based functions*. The specific construction of node based functions will be explained in the section for notations.

For more details on the P_1–nonconforming quadrilateral finite element, see [44].

2.2 Drazin inverse

The Drazin inverse is a generalized inverse of linear transformations or matrices. Here, we introduce the Drazin inverse in brief.

Let A be a linear transformation on $\mathbb{C}^n$. Let k be the smallest nonnegative integer such that $\operatorname{Im} A^0 \supset \operatorname{Im} A \supset \cdots \supset \operatorname{Im} A^{k-1} \supset \operatorname{Im} A^k = \operatorname{Im} A^{k+1} = \cdots$. It is equivalent to $\ker A^0 \subset \ker A \subset \cdots \subset \ker A^{k-1} \subset \ker A^k = \ker A^{k+1} = \cdots$, due to the dimension theorem. k is called the index of A, and denoted by $\operatorname{Ind}(A)$. Then the vector space $\mathbb{C}$ can be decomposed as the sum of the image space and the kernel space of A^k:

Lemma 2.2.1 ([15]). $\mathbb{C}^n = \operatorname{Im} A^k + \ker A^k$.

It yields that, restricted on $\operatorname{Im} A^k$, the transformation A becomes an invertible linear transformation. Thus we can define a linear transformation A^D on $\mathbb{C}^n$ as follows: for $u = v + w \in \mathbb{C}^n$ where $v \in \operatorname{Im} A^k$ and $w \in \ker A^k$, $A^D u := A|^{-1}_{\operatorname{Im} A^k} v$. A^D is called the Drazin inverse of A. When A is a complex matrix in $\mathbb{C}^{n\times n}$, A^D is defined as the matrix of the Drazin inverse of induced linear transform with respect to the standard basis of $\mathbb{C}^n$.

One of the most important properties of the Drazin inverse matrix is that the Drazin inverse matrix of A is expressible as a polynomial in A:

Theorem 2.2.2 ([15]). *If $A \in \mathbb{C}^{n\times n}$, then there exists a polynomial $p(x)$ such that $A^D = p(A)$.*

We know that for given nonsingular matrix A the possibility to express its inverse as a polynomial in A is closely related with Krylov iterative methods. Similarly, even if A is a singular matrix system, a unique Drazin inverse solution can be found using Krylov iterative method under proper consistency condition.

Theorem 2.2.3 ([31]). *Let m be the degree of the minimal polynomial for A, and let i be the index of A. If $b \in \operatorname{Im} A^i$, then the linear system $Ax = b$ has a unique Krylov solution $x = A^D b \in \mathcal{K}_{m-i}(A, b)$. If $b \notin \operatorname{Im} A^i$, then $Ax = b$ does not have a solution in the Krylov space $\mathcal{K}_n(A, b)$.*

For details, see [15, 31].

2.3 Notations

Assume $\Omega \subset \mathbb{R}^d$ is a d-dimensional rectangular domain where $d = 2$ or 3. Let $\mathcal{T}_h$ be a triangulation of Ω consisting of d-dimensional cubes. h denotes the mesh parameter. N_x, N_y, and N_z are the number of elements in $\mathcal{T}_h$ along x-, y-,

and z-direction, respectively. Let $\mathcal{F}_h$, $\mathcal{F}_h^i$, $\mathcal{F}_h^b$, and $\mathcal{F}_h^{b,opp}$ denote the set of all $(d-1)$-dimensional faces, of all interior faces, of all boundary faces, and of all pairs consisting of two boundary faces on opposite position, respectively. Let $\mathcal{N}_h$ denote the set of all nodes in $\mathcal{T}_h$. We introduce several standard Sobolev spaces and discrete function spaces for the P_1–nonconforming quadrilateral finite element:

$$
\begin{aligned}
C^\infty_{per}(\Omega) &= \text{the subset of } C^\infty(\mathbb{R}^d) \text{ of } \Omega\text{-periodic functions},\\
H^1_{per}(\Omega) &= \text{the closure of } C^\infty_{per}(\Omega) \text{ in } H^1\text{-norm},\\
H^1_{per}(\Omega)/\mathbb{R} &= \{v \in H^1_{per}(\Omega) \mid \int_\Omega v = 0\},\\
V^h &= \{v_h \in L^2(\Omega) \mid \ v_h|_K \in \mathcal{P}_1(K)\, \forall K \in \mathcal{T}_h,\, \langle [v_h]_f, 1\rangle_f = 0\, \forall f \in \mathcal{F}_h^i\},\\
V_0^h &= \{v_h \in V^h \mid \langle v_h, 1\rangle_f = 0 \ \forall f \in \mathcal{F}_h^b\},\\
V_{per}^h &= \{v_h \in V^h \mid \langle v_h, 1\rangle_{f_1} = \langle v_h, 1\rangle_{f_2} \ \forall (f_1, f_2) \in \mathcal{F}_h^{b,opp}\},\\
V_{per}^h/\mathbb{R} &= \{v_h \in V_{per}^h \mid \int_\Omega v_h = 0\},
\end{aligned}
$$

where $\mathcal{P}_1(K)$ denotes the set of all linear polynomials on K and $[\cdot]_f$ the jump across $(d-1)$-dimensional face f. Let $\|\cdot\|_0$, $|\cdot|_1$, and $|\cdot|_{1,h}$ denote the standard L^2-norm, H^1-(semi-)norm, and mesh-dependent energy norm in Ω, respectively.

Here we define the concept of node based functions. For a given node z in $\mathcal{T}_h$, let $\mathcal{F}_{(z)}$ denote the set of all $(d-1)$-dimensional faces containing z. Then we can construct a function $\phi_z \in V^h$ associated with z such that

$$
\phi_z(m_f) = \begin{cases} 0.5 & \text{if } f \in \mathcal{F}_{(z)},\\ 0 & \text{else,} \end{cases}
$$

where m_f is the midpoint of $(d-1)$-dimensional face f in $\mathcal{F}_h$. We call ϕ_z

the node based function associated with z. In the case of periodic boundary condition with rectangular Ω, of course, we identify two side boundary nodes in every opposite periodic position, and four nodes at corners. Using the node based functions, we introduce a discrete function space and a set of functions which we mainly use in after:

$$V_{per}^{\mathfrak{B},h} = \{v_h \in V_{per}^h \mid v_h \in \mathrm{Span}\{\phi_z\}_{z\in\mathcal{N}_h^{per}}\},$$
$$\mathfrak{B} = \{\phi_z\}_{z\in\mathcal{N}_h^{per}} : \text{the set of all node based functions in } V_{per}^h,$$

where $\mathcal{N}_h^{per}$ denotes the set of all nodes after periodic identifying. Clearly, due to their definitions, $\mathrm{Span}\,\mathfrak{B} = V_{per}^{\mathfrak{B},h} \subseteq V_{per}^h$. But $\mathfrak{B}$ may not be linearly independent. It is worth to note that $|\mathfrak{B}| = N_xN_y$ in 2-D case, $N_xN_yN_z$ in 3-D case, due to identification between nodes on boundary.

For a given set $\mathfrak{S}$, suppose a vector $\mathbf{v}$ of size $|\mathfrak{S}|$ is given. Then we denote a linear combination of $\mathfrak{S}$, whose representation vector with respect to $\mathfrak{S}$ is $\mathbf{v}$, by $\mathbf{v}\mathfrak{S}$. If a scalar-valued (integrable) function f is given, $\int_{\mathcal{D}} f\mathfrak{S}$ denotes a vector, size of $|\mathfrak{S}|$, such that each component is the integral of the product of f and the corresponding element in $\mathfrak{S}$ over the domain $\mathcal{D}$. $\mathbf{1}_{\mathfrak{S}}$ denotes a vector, size of $|\mathfrak{S}|$, consisting of 1 for all components.

Chapter 3

Dimension of the Finite Spaces

3.1 Induced relation between boundary DoF values

We firstly consider the case of $d = 2$. The higher dimensional case will be covered in Chapter 6. Let N_Q denote the number of all elements in $\mathcal{T}_h$. Let N_V, N_V^i, and N_V^b denote the number of all vertices, of all interior vertices, and of all boundary vertices, respectively. Similarly N_E, N_E^i, and N_E^b denote the number of all edges, of all interior edges, and of all boundary edges, respectively. Our consideration starts from a partition of all vertices.

Lemma 3.1.1. *There exists a partition of all vertices in $\mathcal{T}_h$ into two groups, Red and Black, such that any two vertices connected by an edge are not contained in the same group.*

Proof. Suppose there is no such partition. It means that there are two vertices and two different paths connecting them such that one path consists of edges in even number and the other path consists of edges in odd number. Without

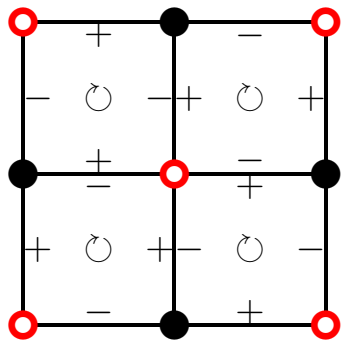

Figure 3.1. An example of dice rules on elements under the same orientation

loss of generality, we assume that these two paths do not share any edge as their common segment. Then the union of two paths composes the boundary of a simply connected domain Ω' consisting of quadrilaterals and the boundary of Ω' consists of edges in odd number. However, a counting formula for the number of edges in Ω' is

$$4\#(\text{elements}) = 2\#(\text{interior edges}) + \#(\text{boundary edges}), \tag{3.1}$$

and it implies that the number of boundary edges of Ω' must be even. This contradiction completes the claim. □

Remark 3.1.2. *Lemma 3.1.1 holds for any simply connected domain and any triangulation with quadrilaterals. If a domain is not simply connected, then such partition of vertices may not exist.*

If (3.1) is applied to the domain Ω, we easily get a simple fact that the number of boundary edges in $\mathcal{T}_h$ is always even. Each edge contains a midpoint and each midpoint is associated with DoF. Thus we have DoF values in even number along boundary edges in $\mathcal{T}_h$. We want to claim a relation between these boundary DoF values.

Choose an orientation and apply it to all elements in $\mathcal{T}_h$. On each element,

we define the direction of each edge along given orientation. If an edge has a direction from Red to Black, then we impose the plus sign on the edge. Else if from Black to Red, the minus sign will be imposed. This rule determines the sign of edges locally. Indeed, every interior edge gets two local signs corresponding two adjacent elements, respectively. It can be observed that two local signs on each interior edge are always opposite because all elements share the same orientation. Figure 3.1 shows an example of such construction with clockwise orientation.

According to the local sign on each edge, we can get a relation which is another form of the dice rule on each element. In other word, if we add 4 DoF values at edge midpoints in each element with the signs corresponding to, then it has to be 0:

$$v_h(m_1^K) - v_h(m_2^K) + v_h(m_3^K) - v_h(m_4^K) = 0 \quad \forall v_h \in V^h, \forall K \in \mathcal{T}_h.$$

Thus we can get the-number-of-elements relations by employing the local signs. Note that the value on each interior edge appears in exactly two equations, but with opposite sign. Therefore, by summing up all equations, we get a single relation which only contains DoF values on boundary with alternating sign. Note that the number of boundary edges in $\mathcal{T}_h$ is always even.

Lemma 3.1.3. *There exists a way to give alternating sign on boundary edges. Moreover, an alternating sum of boundary DoF values of $v_h \in V^h$ is always zero.*

We want to emphasize that the relation between boundary DoF values is induced by the dice rule. In other words, the characteristic of the P_1–nonconforming quadrilateral element enforces the relation on boundary, even in the case of Dirichlet boundary problems. A combination of imposing bound-

ary DoF values violating the relation on boundary is not allowed.

Conversely, this relation can help to impose discrete boundary condition. For instance, in order to impose homogeneous Dirichlet condition on the boundary we do not need to set all boundary DoF values to zero. Zero DoF values at all boundary midpoints except any one of them are just enough because the appropriate last DoF value is naturally given as zero by the relation on the boundary. Such a role of the relation leads to concept of *minimally essential discrete boundary conditions*.

3.2 Minimally essential discrete boundary conditions

As mentioned in the previous section, a combination of the dice rules on all elements induces a relation on boundary DoF values. This relation means a compatibility condition for boundary DoF values in order to be in the discrete function space appropriately. And the induced relation between boundary DoF values can help to impose boundary DoF values associated with given boundary condition. Therefore we do not need to impose given essential boundary condition to all boundary DoFs independently. A subset of essential boundary DoF values will be enough. We call a set of discrete boundary conditions *minimally essential* if essential boundary DoF values in the set induce all other essential boundary DoF values naturally, but any proper subset of the set does not.

The P_1–nonconforming quadrilateral element satisfies the dice rule on each element and inter-element continuity at each interior edge midpoint. Since the dice rule on each element is equivalent to a single relation between DoFs in 2-D case, without considering boundary conditions, the dimension of the discrete function space is equal to the number of all edges subtracted by the number of elements. When a boundary condition is considered, each essential

boundary DoF removes the dimension of the space by 1. Therefore, the number of subtracted degrees of freedom due to essential boundary conditions is just equal to the number of minimally essential discrete boundary conditions.

Lemma 3.2.1. *The following relation holds.*

$$\begin{aligned}&(\textit{dimension of finite space})\\&\quad = \#(\textit{edges}) - \#(\textit{elements})\\&\qquad - \#(\textit{minimally essential discrete boundary conditions}).\end{aligned}$$

Proposition 3.2.2. *(Neumann and Dirichlet B.C.) It holds that*

$$\begin{aligned}&\#(\textit{minimally essential discrete boundary conditions})\\&\quad = \begin{cases} 0 & \textit{if the case of Neumann B.C.,} \\ N_E^b - 1 & \textit{if the case of homogeneous Dirichlet B.C.} \end{cases}\end{aligned}$$

Consequently,

$$\dim V^h = N_E - N_Q = N_V - 1, \tag{3.2a}$$

$$\dim V_0^h = N_E - N_Q - (N_E^b - 1) = N_V^i. \tag{3.2b}$$

Now we consider the case of periodic boundary conditions. In contrast with the case of Dirichlet boundary condition, periodic boundary conditions enforce two boundary DoF values on two opposite boundary edges to be equal. Thus, in this case, the concept of minimally essential discrete boundary conditions means a smallest set of periodic relations between opposite boundary edges which induce all such periodic relations.

The behavior is quite different, which depends on the parity of N_x and N_y.

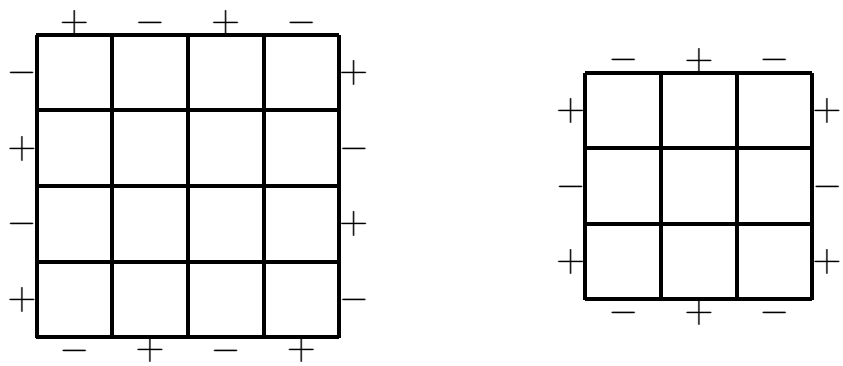

Figure 3.2. Induced relation between boundary DoF values

Suppose both N_x and N_y are even. Then we can easily derive the last periodic relation from the other periodic relations with the help of the relation between boundary DoFs in Lemma 3.1.3. It means that a set of all periodic relations except any one of them is minimally essential. On the other hand, if either N_x or N_y is odd, then we can not get such a natural induction, and a set of all periodic relations itself is minimally essential, see Figure 3.2.

Proposition 3.2.3. *(Periodic B.C.) In case of periodic B.C. on $N_x \times N_y$ rectangular mesh,*

$$\begin{aligned} &\#\textit{(minimally essential discrete boundary conditions)} \\ &\quad = \begin{cases} N_x + N_y - 1 & \textit{if both } N_x \textit{ and } N_y \textit{ are even,} \\ N_x + N_y & \textit{otherwise.} \end{cases} \end{aligned}$$

Consequently,

$$\dim V_{per}^h = \begin{cases} N_x N_y + 1 & \textit{if both } N_x \textit{ and } N_y \textit{ are even,} \\ N_x N_y & \textit{otherwise.} \end{cases} \tag{3.3}$$

Chapter 4

Deeper Look on the Finite Space with Periodic B.C.

We derive the dimension of V_{per}^h which depends on the parity of discretizations in $\mathcal{T}_h$ in Chapter 3. In the first two parts of this chapter, we investigate basis for V_{per}^h. A natural guess to basis for periodic finite space is $\mathfrak{B}$, the set of all node based functions in V_{per}^h. It is a result of natural inference from the case of Dirichlet boundary condition. The set of all interior node based functions becomes a basis for V_0^h. However, in general, $\mathfrak{B}$ may not be a basis for V_{per}^h. It may be linearly dependent and even fail to span V_{per}^h in some cases.

4.1 Linear dependence of $\mathfrak{B}$

We write $\mathfrak{B} = \{\phi_1, \phi_2, \cdots, \phi_{|\mathfrak{B}|}\}$. Define a surjective linear map $B_h^{\mathfrak{B}} : \mathbb{R}^{|\mathfrak{B}|} \to V_{per}^{\mathfrak{B},h}$ by $B_h^{\mathfrak{B}}(\mathbf{c}) = \sum_j c_j \phi_j$ where $\mathbf{c} = (c_j) \in \mathbb{R}^{|\mathfrak{B}|}$. Then $\ker B_h^{\mathfrak{B}}$ is the set of all nontrivial representations of the zero function. Before investigation on global

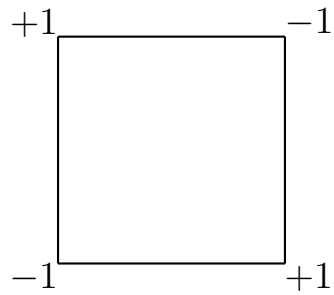

Figure 4.1. A nontrivial representation for the zero function on a square

representations, let us consider local representations in detail.

On a single element, there is a single degree of freedom for the zero representation. Figure 4.1 shows such a representation of coefficients for node based functions. The value at each vertex represents a coefficient for the corresponding node based function in $\mathfrak{B}$. By extension of local coefficients, global coefficient representations for $\ker B_h^{\mathfrak{B}}$ can be obtained. To match coefficients on adjacent elements, the only way to extend local representation is repetition of local representation with alternating sign. The extension is possible only if the number of discretization on each coordinate is even due to the periodicity. Moreover such extension is unique. On the other hand, if N_x is odd, the alternating extension along x-direction implies the trivial representation because we identify some nodes on the boundary. The case of odd N_y is similar.

Proposition 4.1.1. *(The dimension of* $\ker B_h^{\mathfrak{B}}$ *and* $V_{per}^{\mathfrak{B},h}$*) It holds that*

$$\dim \ker B_h^{\mathfrak{B}} = \begin{cases} 1 & \text{if both } N_x \text{ and } N_y \text{ are even,} \\ 0 & \text{else,} \end{cases} \tag{4.1}$$

and any $|\mathfrak{B}| - 1$ *functions in* $\mathfrak{B}$ *form a basis for* $V_{per}^{\mathfrak{B},h}$ *when both* N_x *and* N_y *are even, whereas* $\mathfrak{B}$ *itself is a basis for* $V_{per}^{\mathfrak{B},h}$ *when either* N_x *or* N_y *is odd.*

Consequently,

$$\dim V_{per}^{\mathfrak{B},h} = |\mathfrak{B}| - \dim \ker B_h^{\mathfrak{B}} = \begin{cases} N_x N_y - 1 & \textit{if both } N_x \textit{ and } N_y \textit{ are even,} \\ N_x N_y & \textit{else.} \end{cases} \tag{4.2}$$

4.2 A Basis for V_{per}^h

For the first case, we suppose both N_x and N_y are even. Propositions 3.2.3 and 4.1.1 imply that $\mathfrak{B}$ is linearly dependent and $V_{per}^{\mathfrak{B},h}$ is a proper subset of V_{per}^h. The difference between the dimensions of $V_{per}^{\mathfrak{B},h}$ and V_{per}^h is equal to 2. It means that there exist two complementary basis functions for V_{per}^h which do not belong to $V_{per}^{\mathfrak{B},h}$.

Let ψ_x denote a piecewise linear function in V_{per}^h whose DoF values on vertical edges are all 1 with alternating sign in vertical and horizontal direction, and DoF values on horizontal edges are all 0 (Figure 4.2 (a)). ψ_x is well-defined since N_x is even. Note that piecewise partial derivative of ψ_x in x-direction forms a checkerboard pattern, but piecewise partial derivative in y-direction is always zero. To show $\psi_x \notin V_{per}^{\mathfrak{B},h}$, define a linear functional $J_x^h : V_{per}^h \to \mathbb{R}$ as follows. For given $v_h \in V_{per}^h$, $J_x^h(v_h)$ is the sum of DoF values of v_h on all vertical edges with the alternating sign same to that of ψ_x. It is easily shown that, if N_y is even, J_x^h maps every node based function ϕ_j to zero. However $J_x^h(\psi_x)$ is nonzero, which means ψ_x can not be constructed by any linear combination of node based functions. In other words, $\psi_x \notin V_{per}^{\mathfrak{B},h}$. Similarly, we can find another piecewise linear function ψ_y in V_{per}^h, not belonging to $V_{per}^{\mathfrak{B},h}$ (Figure 4.2 (b)).

The second is the case either N_x or N_y is odd. Propositions 3.2.3 and 4.1.1 imply that $\mathfrak{B}$ is linearly independent and $\dim V_{per}^{\mathfrak{B},h} = \dim V_{per}^h$. Therefore

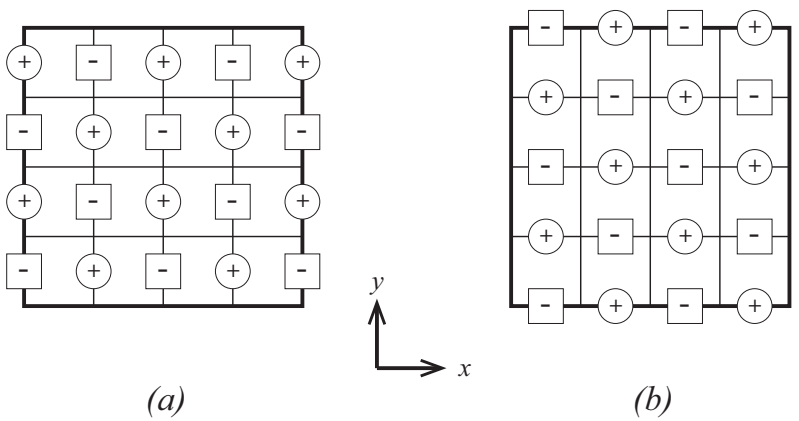

Figure 4.2. An example of two alternating functions (*a*) ψ_x and (*b*) ψ_y

$V_{per}^{\mathfrak{B},h} = V_{per}^h$ and $\mathfrak{B}$, the set of all node based functions, is a basis for V_{per}^h.

Theorem 4.2.1. *(A complementary basis for V_{per}^h)*

1. *If both N_x and N_y are even, then $V_{per}^{\mathfrak{B},h}$ is a proper subset of V_{per}^h. And $\{\psi_x, \psi_y\}$ is a complementary basis for V_{per}^h, not belonging to $V_{per}^{\mathfrak{B},h}$.*

2. *Else if either N_x or N_y is odd, then $V_{per}^{\mathfrak{B},h} = V_{per}^h$.*

4.3 Stiffness matrix associated with $\mathfrak{B}$

Even though it may not be a basis for V_{per}^h, $\mathfrak{B}$ is still a useful set of functions to understand V_{per}^h. The dimension result in previous sections claims that $V_{per}^{\mathfrak{B},h}$, the span of $\mathfrak{B}$, occupies almost of V_{per}^h. Furthermore, the node based functions are easy to handle in implementation viewpoint. We study about $\mathfrak{B}$ in this section.

Let $\mathbf{S}_h^{\mathfrak{B}}$ be the $|\mathfrak{B}|$-by-$|\mathfrak{B}|$ stiffness matrix associated with $\mathfrak{B} = \{\phi_j\}$.

$$(\mathbf{S}_h^{\mathfrak{B}})_{jk} = \sum_{K \in \mathcal{T}_h} \int_K \nabla\phi_k \cdot \nabla\phi_j \, \mathrm{d}\mathbf{x} \quad 1 \le j, k \le |\mathfrak{B}|. \tag{4.3}$$

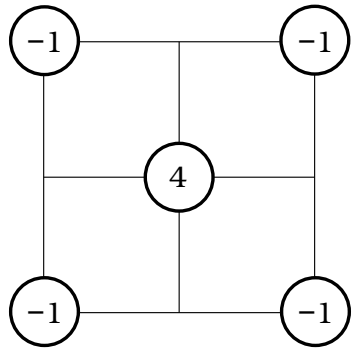

Figure 4.3. The stencil for $\mathbf{S}_h^{\mathfrak{B}}$

The local stencil for the stiffness matrix associated with $\mathfrak{B}$ is shown in Figure 4.3. Obviously, $\mathbf{S}_h^{\mathfrak{B}}$ is symmetric and positive semi-definite.

Lemma 4.3.1. *Let* $v_h = \sum_j v_j \phi_j$ *for* $\mathbf{v} = (v_j) \in \mathbb{R}^{|\mathfrak{B}|}$. *Then* $\mathbf{v} \in \ker \mathbf{S}_h^{\mathfrak{B}}$ *if and only if* v_h *is a constant function in* Ω.

Proof. By the definition of $\mathbf{S}_h^{\mathfrak{B}}$, $\sum_{K \in \mathcal{T}_h} \int_K |\nabla v_h|^2 \, \mathrm{d}\mathbf{x} = \mathbf{v}^T \mathbf{S}_h^{\mathfrak{B}} \mathbf{v}$. If $\mathbf{v} \in \ker \mathbf{S}_h^{\mathfrak{B}}$, then v_h is constant in Ω, due to the weak continuity across each edge. Conversely if v_h is constant, then $\mathbf{v}^T \mathbf{S}_h^{\mathfrak{B}} \mathbf{v} = 0$. Since $\mathbf{S}_h^{\mathfrak{B}}$ is symmetric positive semi-definite, it has its square root matrix. Thus we get $\mathbf{S}_h^{\mathfrak{B}} \mathbf{v} = \mathbf{0}$. □

Next claims reveal the relation between $\ker \mathbf{S}_h^{\mathfrak{B}}$ and $\ker B_h^{\mathfrak{B}}$.

Lemma 4.3.2. $\ker B_h^{\mathfrak{B}} \subset \ker \mathbf{S}_h^{\mathfrak{B}}$.

Proof. Let $\mathbf{v} = (v_j)$ be in $\ker B_h^{\mathfrak{B}}$, *i.e.*, $\sum_j v_j \phi_j = 0$. The claim is a simple consequence of Lemma 4.3.1. □

Proposition 4.3.3. $\ker \mathbf{S}_h^{\mathfrak{B}}$ *can be decomposed as*

$$\ker \mathbf{S}_h^{\mathfrak{B}} = \ker B_h^{\mathfrak{B}} \oplus \operatorname{Span} \mathbf{1}_{\mathfrak{B}}. \tag{4.4}$$

Consequently, $\dim \ker \mathbf{S}_h^{\mathfrak{B}} = \dim \ker B_h^{\mathfrak{B}} + 1$.

Proof. Note that both $\ker B_h^{\mathfrak{B}}$ and $\operatorname{Span} \mathbf{1}_{\mathfrak{B}}$ are subsets of $\ker \mathbf{S}_h^{\mathfrak{B}}$, and $\ker B_h^{\mathfrak{B}} \cap \operatorname{Span} \mathbf{1} = \{\mathbf{0}\}$ due to Lemmas 4.3.1 and 4.3.2. Thus it is enough to show that any $\mathbf{v}$ in $\ker \mathbf{S}_h^{\mathfrak{B}}$ can be expressed as a sum of two vectors which are in $\ker B_h^{\mathfrak{B}}$ and $\operatorname{Span} \mathbf{1}_{\mathfrak{B}}$, respectively.

Suppose $\mathbf{v} = (v_j) \in \ker \mathbf{S}_h^{\mathfrak{B}}$. Lemma 4.3.1 implies that there exists a constant $\alpha \in \mathbb{R}$ such that $\sum_j v_j \phi_j \equiv \alpha$. Note that $\mathfrak{B}$ is a partition of unity, *i.e.*, $\sum_j \phi_j \equiv 1$. We can rewrite as $\sum_j (v_j - \alpha)\phi_j = 0$, which implies that $\mathbf{v} - \alpha \mathbf{1}_{\mathfrak{B}} \in \ker B_h^{\mathfrak{B}}$. Therefore $\mathbf{v}$ can be decomposed as $\mathbf{v} = (\mathbf{v} - \alpha \mathbf{1}_{\mathfrak{B}}) + \alpha \mathbf{1}_{\mathfrak{B}}$ and it completes the proof. □

Remark 4.3.4. *Lemmas 4.3.1 and 4.3.2, and Proposition 4.3.3 are also valid in 3-D case.*

The following is a simple consequence of Propositions 4.1.1 and 4.3.3.

Proposition 4.3.5. *(The dimension of* $\ker \mathbf{S}_h^{\mathfrak{B}}$*)*

$$\dim \ker \mathbf{S}_h^{\mathfrak{B}} = \begin{cases} 2 & \text{if both } N_x \text{ and } N_y \text{ are even,} \\ 1 & \text{else.} \end{cases} \tag{4.5}$$

4.4 Numerical schemes for elliptic problems with periodic boundary condition

Consider an elliptic problem with periodic boundary condition

$$-\Delta u = f \text{ in } \Omega, \tag{4.6a}$$

$$u \text{ is periodic,} \tag{4.6b}$$

$$\int_\Omega u \, \mathrm{d}\mathbf{x} = 0, \tag{4.6c}$$

with the compatibility condition $\int_\Omega f = 0$. The zero-integral condition (4.6c) is quite natural since the governing equation is invariant to additive constant on the variable. The weak formulation is as follows: *find* $u \in H^1_{per}(\Omega)$ *such that*

$$\int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}\mathbf{x} = \int_\Omega f v \, \mathrm{d}\mathbf{x} \quad \forall v \in H^1_{per}(\Omega), \tag{4.7a}$$

$$\int_\Omega u \, \mathrm{d}\mathbf{x} = 0. \tag{4.7b}$$

And the corresponding discrete weak formulation is as follows: *find* $u_h \in V^h_{per}$ *such that*

$$a_h(u_h, v_h) = \int_\Omega f v_h \, \mathrm{d}\mathbf{x} \quad \forall v_h \in V^h_{per}, \tag{4.8a}$$

$$\int_\Omega u_h \, \mathrm{d}\mathbf{x} = 0, \tag{4.8b}$$

where $a_h(u_h, v_h) := \sum_{K \in \mathcal{T}_h} \int_K \nabla u_h \cdot \nabla v_h \, \mathrm{d}\mathbf{x}$.

Throughout this section, we assume that both N_x and N_y are even. The other case which considers odd N_x and/or N_y is easy to handle because V^h_{per} is just equal to $V^{\mathfrak{B},h}_{per}$. Due to Proposition 4.1.1, we can find $\mathfrak{B}^\flat$, a proper subset of $\mathfrak{B}$, which is a basis for $V^{\mathfrak{B},h}_{per}$. It clearly holds that $|\mathfrak{B}^\flat| = \dim V^{\mathfrak{B},h}_{per} = |\mathfrak{B}| - 1$. Without loss of generality, we take $\mathfrak{B}^\flat = \{\phi_1, \cdots, \phi_{|\mathfrak{B}|-1}\}$. We want to recall ψ_x and ψ_y, the two complementary basis functions for V^h_{per} which are not belonging to $V^{\mathfrak{B},h}_{per}$, in Section 4.2. Let $\mathfrak{A}$ denote the set consisting of these two functions, $\{\psi_x, \psi_y\}$. Consider two extended sets $\mathfrak{E} := \mathfrak{B} \cup \mathfrak{A}$, and $\mathfrak{E}^\flat := \mathfrak{B}^\flat \cup \mathfrak{A}$. Remark that $\mathfrak{E}^\flat$ is a basis for V^h_{per}. The characteristics of $\mathfrak{B}^\flat$, $\mathfrak{B}$, $\mathfrak{E}^\flat$, and $\mathfrak{E}$ are summarized in Table 4.1.

For a vector $\mathbf{v}$ of size $|\mathfrak{E}|$, let $\mathbf{v}|_{\mathfrak{B}}$ and $\mathbf{v}|_{\mathfrak{A}}$ denote vectors consisting of the first $|\mathfrak{B}|$ components, and of the last $|\mathfrak{A}|$ components, respectively. Similarly,

$\mathcal{S}$	$\lvert\mathcal{S}\rvert$	Span $\mathcal{S}$	dim Span $\mathcal{S}$
$\mathfrak{B}^\flat$	N_xN_y-1	$V_{per}^{\mathfrak{B},h}$	N_xN_y-1
$\mathfrak{B}$	N_xN_y		
$\mathfrak{E}^\flat$	N_xN_y+1	V_{per}^{h}	N_xN_y+1
$\mathfrak{E}$	N_xN_y+2		

Table 4.1. Summary of characteristics of $\mathfrak{B}^\flat$, $\mathfrak{B}$, $\mathfrak{E}^\flat$, $\mathfrak{E}$ when both N_x, N_y are even

notations $\mathbf{v}|_{\mathfrak{B}^\flat}$ and $\mathbf{v}|_{\mathfrak{A}}$ are used for a vector $\mathbf{v}$ of size $|\mathfrak{E}^\flat|$. Several properties of functions in $\mathfrak{B}$ and $\mathfrak{A}$ are observed.

Lemma 4.4.1. *Let $\mathfrak{B}$ and $\mathfrak{A}$ be as above. Then the followings hold.*

1. $a_h(\phi,\psi)=0 \quad \forall\phi\in\mathfrak{B}\ \forall\psi\in\mathfrak{A}$.

2. $a_h(\psi_\mu,\psi_\nu)=0 \quad \forall\psi_\mu,\psi_\nu\in\mathfrak{A}$ *such that* $\mu\neq\nu$.

3. $\int_\Omega\psi=0 \quad \forall\psi\in\mathfrak{A}$.

4. *There exists an h-independent constant C such that* $\|\psi\|_0\le C$ *and* $|\psi|_{1,h}\le C/h \quad \forall\psi\in\mathfrak{A}$.

Next, we introduce a stiffness matrix associated with another set of functions, and its variant. Let $\mathbf{S}_h^{\mathfrak{B}^\flat}$ be the $|\mathfrak{B}^\flat|$-by-$|\mathfrak{B}^\flat|$ stiffness matrix associated with $\mathfrak{B}^\flat$,

$$(\mathbf{S}_h^{\mathfrak{B}^\flat})_{jk}:=a_h(\phi_k,\phi_j)\quad 1\le j,k\le|\mathfrak{B}^\flat|, \tag{4.9}$$

and $\tilde{\mathbf{S}}_h^{\mathfrak{B}^\flat}$ be the matrix same as $\mathbf{S}_h^{\mathfrak{B}^\flat}$, but the last row is modified in order to impose the zero-integral condition. Because all the integrals $\int_\Omega\phi_j$ are same

for all ϕ_j in $\mathfrak{B}$, every entry in the last row is replaced by 1.

$$(\tilde{\mathbf{S}}_h^{\mathfrak{B}^\flat})_{jk} := \begin{cases} a_h(\phi_k, \phi_j) & j \neq |\mathfrak{B}^\flat|, \\ 1 & j = |\mathfrak{B}^\flat|. \end{cases} \tag{4.10}$$

Note that $\tilde{\mathbf{S}}_h^{\mathfrak{B}^\flat}$ is nonsingular whereas both $\mathbf{S}_h^{\mathfrak{B}}$ and $\mathbf{S}_h^{\mathfrak{B}^\flat}$ are singular with rank deficiency 2 and 1, respectively. For the complementary part, let $\mathbf{S}_h^{\mathfrak{A}}$ be the $|\mathfrak{A}|$-by-$|\mathfrak{A}|$ stiffness matrix associated with $\mathfrak{A}$,

$$(\mathbf{S}_h^{\mathfrak{A}})_{jk} := a_h(\psi_k, \psi_j) \quad 1 \leq j, k \leq |\mathfrak{A}|. \tag{4.11}$$

$\mathbf{S}_h^{\mathfrak{A}}$ is a nonsingular diagonal matrix due to Lemma 4.4.1. In followings we introduce 4 numerical approaches to solve (4.8).

4.4.1 Option 1: $\mathcal{S} = \mathfrak{E}^\flat$ for a nonsingular nonsymmetric system

Since $\mathfrak{E}^\flat$ is a basis for V_{per}^h, $\mathfrak{E}^\flat$ is a natural choice as a set of trial and test functions to assemble a matrix equation corresponding to (4.8). The numerical solution $u_h \in V_{per}^h$ is uniquely expressed, associated with $\mathfrak{E}^\flat$, as

$$u_h = \tilde{\mathbf{u}}^\flat \mathfrak{E}^\flat \tag{4.12}$$

where $\tilde{\mathbf{u}}^\flat$ is the solution of the system of equations associated with $\mathfrak{E}^\flat$

$$\tilde{\mathcal{L}}_h^{\mathfrak{E}^\flat} \tilde{\mathbf{u}}^\flat := \begin{bmatrix} \tilde{\mathbf{S}}_h^{\mathfrak{B}^\flat} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_h^{\mathfrak{A}} \end{bmatrix} \tilde{\mathbf{u}}^\flat = \begin{bmatrix} \tilde{\mathbf{f}}_{\mathfrak{B}^\flat} \\ \mathbf{f}_{\mathfrak{A}} \end{bmatrix} \tag{4.13a}$$

with

$$(\tilde{\mathbf{f}}_{\mathfrak{B}^\flat})_j = \begin{cases} \int_\Omega f\phi_j, & j \neq |\mathfrak{B}^\flat| \\ 0, & j = |\mathfrak{B}^\flat| \end{cases}, \quad \mathbf{f}_{\mathfrak{A}} = \int_\Omega f\mathfrak{A}. \tag{4.13b}$$

Due to Lemma 4.4.1, we get a block-diagonal system as above. The system matrix is nonsingular, but nonsymmetric due to modification of the last row of $\tilde{\mathbf{S}}^{\mathfrak{B}^\flat}$ which is derived from the zero-integral condition. We can use any known numerical scheme for general matrix systems, for instance GMRES, to solve (4.13).

4.4.2 Option 2: $\mathcal{S} = \mathfrak{E}^\flat$ for a symmetric positive semi-definite system with rank deficiency 1

In the previous approach, the zero-integral condition is imposed in a system of equations directly. In a consequence, the associated system matrix becomes nonsymmetric due to modification of just a single row. If we use a numerical scheme which conserves symmetry of the system, then we can enjoy advantages of the symmetry.

An alternative approach is a way to impose the zero-integral condition indirectly in order to conserve symmetry of the assembled system matrix. We make our solution satisfying the zero-integral condition in post-processing. On the other hand, nonsingularity of the matrix can not be maintained any longer in this approach. We have to find out a solution of a singular matrix problem. Fortunately the system matrix is at least positive semi-definite.

Consider a system of equations for (4.8) associated with $\mathfrak{E}^\flat$ without any

modification,

$$\mathcal{L}_h^{\mathfrak{E}^\flat}\mathbf{u}^\flat := \begin{bmatrix} \mathbf{S}_h^{\mathfrak{B}^\flat} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_h^{\mathfrak{A}} \end{bmatrix} \mathbf{u}^\flat = \int_\Omega f\mathfrak{E}^\flat. \tag{4.14}$$

Note that the above system matrix is singular, and symmetric positive semi-definite. We find the solution $\mathbf{u}^\flat$ of the system such that

$$\mathbf{u}^\flat|_{\mathfrak{B}^\flat} \cdot \mathbf{1}_{\mathfrak{B}^\flat} = 0 \tag{4.15}$$

since $\int_\Omega \mathbf{v}\mathfrak{E}^\flat = 0$ if and only if $\mathbf{v}|_{\mathfrak{B}^\flat} \cdot \mathbf{1}_{\mathfrak{B}^\flat} = 0$, and the numerical solution $u_h^\flat \in V_{per}^h$ of this scheme is obtained by

$$u_h^\flat = \mathbf{u}^\flat \mathfrak{E}^\flat. \tag{4.16}$$

As mentioned in Section 2.2, we can find a unique Drazin inverse solution of a singular system using Krylov iterative methods under proper condition. When a symmetric positive semi-definite system $Ax = b$ is given, as our formulation, the Conjugate Gradient method (CG) gives a unique Krylov solution if consistency condition $b \in \operatorname{Im} A$ holds. The general solution is obviously obtained upto its kernel space.

The kernel space of the system matrix in (4.14) is closely related with the kernel space of $\mathbf{S}_h^{\mathfrak{B}^\flat}$. A simple analog of Section 4.3 implies that the dimension of $\ker \mathbf{S}_h^{\mathfrak{B}^\flat}$ is 1, and $\mathbf{v} \in \ker \mathbf{S}_h^{\mathfrak{B}^\flat}$ if and only if $\mathbf{v}\mathfrak{B}^\flat$ is a constant function in Ω. Note that $\mathfrak{B}^\flat$ is not a partition of unity, whereas $\mathfrak{B}$ is. Let $\mathbf{w}_{\mathfrak{B}^\flat}$ denote a unique vector in $\mathbb{R}^{|\mathfrak{B}^\flat|}$ such that $\mathbf{w}_{\mathfrak{B}^\flat}\mathfrak{B}^\flat \equiv 1$ in Ω. Then the kernel space of the system matrix in (4.14) is simply represented by $\operatorname{Span} \mathbf{w}_{\mathfrak{E}^\flat}$ where $\mathbf{w}_{\mathfrak{E}^\flat} \in \mathbb{R}^{|\mathfrak{E}^\flat|}$ is the trivial extension of $\mathbf{w}_{\mathfrak{B}^\flat}$, as $\begin{bmatrix} \mathbf{w}_{\mathfrak{B}^\flat}^T & \mathbf{0} \end{bmatrix}^T$. Therefore in post-processing we add a multiple of $\mathbf{w}_{\mathfrak{E}^\flat}$ to the Krylov solution to satisfy (4.15).

We have the numerical solution $u_h^\flat$ as follows.

1. Take a vector $\mathbf{u}^{(0)} \in \mathbb{R}^{|\mathfrak{E}^\flat|}$ for an initial guess.

2. Solve the singular symmetric positive semi-definite system (4.15) by the CG and get the Krylov solution $\mathbf{u}' = \mathbf{u}^{(n)}$.

3. Add a multiple of $\mathbf{w}_{\mathfrak{E}^\flat}$ to $\mathbf{u}'$ in order to enforce the zero-integral condition (4.15) as

$$\mathbf{u}^\flat = \mathbf{u}' - \frac{\mathbf{u}'|_{\mathfrak{B}^\flat} \cdot \mathbf{1}_{\mathfrak{B}^\flat}}{\mathbf{w}_{\mathfrak{B}^\flat} \cdot \mathbf{1}_{\mathfrak{B}^\flat}} \mathbf{w}_{\mathfrak{E}^\flat}.$$

4. The numerical solution is obtained as $u_h^\flat = \mathbf{u}^\flat \mathfrak{E}^\flat$.

Let $\mathbf{u}^\flat$ and $\tilde{\mathbf{u}}^\flat$ be the solutions as in Sections 4.4.1 and 4.4.2, respectively. Note that two linear systems (4.13) and (4.14) coincide except $|\mathfrak{B}^\flat|$-th row. Even on $|\mathfrak{B}^\flat|$-th row,

$$\begin{aligned}
\left(\tilde{\mathcal{L}}_h^{\mathfrak{E}^\flat} \mathbf{u}^\flat\right)_{|\mathfrak{B}^\flat|} &= \mathbf{1}_{\mathfrak{B}^\flat} \cdot \left(\mathbf{u}' - \frac{\mathbf{u}'|_{\mathfrak{B}^\flat} \cdot \mathbf{1}_{\mathfrak{B}^\flat}}{\mathbf{w}_{\mathfrak{B}^\flat} \cdot \mathbf{1}_{\mathfrak{B}^\flat}} \mathbf{w}_{\mathfrak{E}^\flat}\right)\Big|_{\mathfrak{B}^\flat} \\
&= \mathbf{1}_{\mathfrak{B}^\flat} \cdot \left(\mathbf{u}'|_{\mathfrak{B}^\flat} - \frac{\mathbf{u}'|_{\mathfrak{B}^\flat} \cdot \mathbf{1}_{\mathfrak{B}^\flat}}{\mathbf{w}_{\mathfrak{B}^\flat} \cdot \mathbf{1}_{\mathfrak{B}^\flat}} \mathbf{w}_{\mathfrak{B}^\flat}\right) \\
&= 0.
\end{aligned}$$

Thus $\tilde{\mathcal{L}}_h^{\mathfrak{E}^\flat} \mathbf{u}^\flat = \begin{bmatrix} \tilde{\mathbf{f}}_{\mathfrak{B}^\flat} \\ \mathbf{f}_{\mathfrak{A}} \end{bmatrix} = \tilde{\mathcal{L}}_h^{\mathfrak{E}^\flat} \tilde{\mathbf{u}}^\flat$, and it implies $\mathbf{u}^\flat = \tilde{\mathbf{u}}^\flat$ because $\tilde{\mathcal{L}}_h^{\mathfrak{E}^\flat}$ is nonsingular. Therefore two schemes give the same numerical solution.

4.4.3 Option 3: $\mathcal{S} = \mathfrak{E}$ for a symmetric positive semi-definite system with rank deficiency 2

Although symmetry and positive semi-definiteness of the system matrix are key factors for an efficient numerical scheme, we can not enjoy full benefits in the previous scheme. We need the extra post-processing to impose the zero-integral condition. The defect in the previous approach comes from the fact that the Riesz representation vector for the integral functional does not belong to the kernel space of the system matrix. As shown above, the kernel space of the system matrix is closely related with the coefficient vector for the unity function. If these two vectors coincide, we can get our solution without any post-processing. The imbalance of $\mathfrak{B}^\flat$ for the linear independence is also a disadvantage to numerical implementation.

In this approach, we find the numerical solution $u_h^\natural \in V_{per}^h$ such that

$$u_h^\natural = \mathbf{u}^\natural \mathfrak{E} \tag{4.17}$$

where $\mathbf{u}^\natural$ is a solution of a system of equations for (4.8) associated with full $\mathfrak{E}$,

$$\mathcal{L}_h^{\mathfrak{E}} \mathbf{u}^\natural := \begin{bmatrix} \mathbf{S}_h^{\mathfrak{B}} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_h^{\mathfrak{A}} \end{bmatrix} \mathbf{u}^\natural = \int_\Omega f \mathfrak{E} \tag{4.18}$$

with

$$\mathbf{u}^\natural|_{\mathfrak{B}} \cdot \mathbf{1}_{\mathfrak{B}} = 0, \tag{4.19}$$

since $\int_\Omega \mathbf{v}\mathfrak{E} = 0$ if and only if $\mathbf{v}|_{\mathfrak{B}} \cdot \mathbf{1}_{\mathfrak{B}} = 0$. The numerical solution $u^\natural$ is unique because solution of the matrix system is unique upto additive nontrivial representation for the zero function in $\mathfrak{B}$. We want to emphasize that, unlike

the previous scheme, $\mathbf{1}_{\mathfrak{B}}$ belongs to the kernel space of $\mathbf{S}_h^{\mathfrak{B}}$ as shown in (4.3.3). It implies that, without any extra post-processing, we can find the solution of the linear system which satisfies the zero-integral condition (4.19) if an initial guess is chosen to satisfy the same condition.

We have the numerical solution $u_h^\natural$ as follows.

1. Take an initial vector $\mathbf{u}^{(0)} \in \mathbb{R}^{|\mathfrak{E}|}$ which satisfies $\mathbf{u}^{(0)}|_{\mathfrak{B}} \cdot \mathbf{1}_{\mathfrak{B}} = 0$.
2. Solve the singular symmetric positive semi-definite system (4.18) by the CG and get the Krylov solution $\mathbf{u}^\natural = \mathbf{u}^{(n)}$.
3. The numerical solution is obtained as $u_h^\natural = \mathbf{u}^\natural \mathfrak{E}$.

Let $\mathbf{u}^\natural$ and $\mathbf{u}^\flat$ be the solutions as in Sections 4.4.3 and 4.4.2, respectively. Clearly $\mathbf{u}^\natural|_{\mathfrak{A}} = \mathbf{u}^\flat|_{\mathfrak{A}}$. Therefore it is enough to show $\mathbf{u}^\natural|_{\mathfrak{B}}\mathfrak{B} = \mathbf{u}^\flat|_{\mathfrak{B}^\flat}\mathfrak{B}^\flat$ to prove the equality of two solutions $u_h^\natural$ and $u_h^\flat$. Note that $u_h^\flat$ has been already proven to be equal to u_h.

Let $\begin{bmatrix} \mathbf{u}^\flat|_{\mathfrak{B}^\flat} \\ 0 \end{bmatrix}$ be a trivial extension of $\mathbf{u}^\flat|_{\mathfrak{B}^\flat}$ into a vector in $\mathbb{R}^{|\mathfrak{B}|}$ by padding a single zero. Note that $\sum_{j=1}^{|\mathfrak{B}|}(\mathbf{S}_h^{\mathfrak{B}})_{jk} = 0$ for all $1 \le k \le |\mathfrak{B}|$. Due to the definition of $\mathbf{u}^\flat$, we have

$$
\begin{aligned}
\mathbf{S}_h^{\mathfrak{B}} \begin{bmatrix} \mathbf{u}^\flat|_{\mathfrak{B}^\flat} \\ 0 \end{bmatrix} &= \begin{bmatrix} \mathbf{S}_h^{\mathfrak{B}^\flat}\mathbf{u}^\flat|_{\mathfrak{B}^\flat} \\ [\mathbf{S}_h^{\mathfrak{B}}]_{|\mathfrak{B}|,1:|\mathfrak{B}^\flat|}\mathbf{u}^\flat|_{\mathfrak{B}^\flat} \end{bmatrix} \\
&= \begin{bmatrix} \int_\Omega f\mathfrak{B}^\flat \\ -\sum_{j\neq|\mathfrak{B}|}[\mathbf{S}_h^{\mathfrak{B}}]_{j,1:|\mathfrak{B}^\flat|}\mathbf{u}^\flat|_{\mathfrak{B}^\flat} \end{bmatrix} \\
&= \begin{bmatrix} \int_\Omega f\mathfrak{B}^\flat \\ -\sum_{j=1}^{|\mathfrak{B}^\flat|}[\mathbf{S}_h^{\mathfrak{B}^\flat}]_{j,1:|\mathfrak{B}^\flat|}\mathbf{u}^\flat|_{\mathfrak{B}^\flat} \end{bmatrix} \\
&= \begin{bmatrix} \int_\Omega f\mathfrak{B}^\flat \\ -\sum_{j=1}^{|\mathfrak{B}^\flat|}\int_\Omega f\phi_j \end{bmatrix} = \begin{bmatrix} \int_\Omega f\mathfrak{B}^\flat \\ \int_\Omega f(\phi_{|\mathfrak{B}|}-1) \end{bmatrix} = \int_\Omega f\mathfrak{B},
\end{aligned}
$$

since $\mathfrak{B}$ is a partition of unity and $\int_\Omega f = 0$. On the other hand, the definition of $\mathbf{u}^\natural$ implies $\mathbf{S}_h^{\mathfrak{B}}\mathbf{u}^\natural|_{\mathfrak{B}} = \int_\Omega f\mathfrak{B}$. Thus $\mathbf{u}^\natural|_{\mathfrak{B}} - \begin{bmatrix} \mathbf{u}^\flat|_{\mathfrak{B}^\flat} \\ 0 \end{bmatrix}$ is in the kernel space of $\mathbf{S}_h^{\mathfrak{B}}$, which is decomposed as Proposition 4.3.3. Due to the zero-integral condition in each scheme, $\left(\mathbf{u}^\natural|_{\mathfrak{B}} - \begin{bmatrix} \mathbf{u}^\flat|_{\mathfrak{B}^\flat} \\ 0 \end{bmatrix}\right) \cdot \mathbf{1}_{\mathfrak{B}} = \mathbf{u}^\natural|_{\mathfrak{B}} \cdot \mathbf{1}_{\mathfrak{B}} - \mathbf{u}^\flat|_{\mathfrak{B}^\flat} \cdot \mathbf{1}_{\mathfrak{B}^\flat} = 0$. Therefore $\mathbf{u}^\natural|_{\mathfrak{B}} - \begin{bmatrix} \mathbf{u}^\flat|_{\mathfrak{B}^\flat} \\ 0 \end{bmatrix}$ must belong to $\ker B_h^{\mathfrak{B}}$, and consequently $\left(\mathbf{u}^\natural|_{\mathfrak{B}} - \begin{bmatrix} \mathbf{u}^\flat|_{\mathfrak{B}^\flat} \\ 0 \end{bmatrix}\right) \mathfrak{B} = \mathbf{u}^\natural|_{\mathfrak{B}}\mathfrak{B} - \mathbf{u}^\flat|_{\mathfrak{B}^\flat}\mathfrak{B}^\flat$ is equal to 0. This concludes our claim.

4.4.4 Option 4: $\mathcal{S} = \mathfrak{B}$ for a symmetric positive semi-definite system with rank deficiency 2

Consider a system of equations associated only with $\mathfrak{B}$ for (4.8) with $V_{per}^{\mathfrak{B},h}$ rather than V_{per}^h,

$$\mathcal{L}_h^{\mathfrak{B}}\bar{\mathbf{u}}^\natural := \mathbf{S}_h^{\mathfrak{B}}\bar{\mathbf{u}}^\natural = \int_\Omega f\mathfrak{B}. \tag{4.20}$$

Starting from an initial vector $\mathbf{u}^{(0)} \in \mathbb{R}^{|\mathfrak{B}|}$ which satisfies $\mathbf{u}^{(0)} \cdot \mathbf{1}_{\mathfrak{B}} = 0$, let $\bar{\mathbf{u}}^\natural$ be the Krylov solution of the linear system. The numerical solution $\bar{u}_h^\natural \in V_{per}^h$ is obtained by

$$\bar{u}_h^\natural = \bar{\mathbf{u}}^\natural\mathfrak{B}. \tag{4.21}$$

Let $\mathbf{u}^\natural$ and $\bar{\mathbf{u}}^\natural$ be the solutions as in Sections 4.4.3 and 4.4.4, respectively. Note that $\mathbf{u}^\natural|_{\mathfrak{B}} = \bar{\mathbf{u}}^\natural$, and $\mathbf{u}^\natural|_{\mathfrak{A}} = diag\left(a_h(\psi_x,\psi_x), a_h(\psi_y,\psi_y)\right)^{-1} \int_\Omega f\mathfrak{A} \leq$

$Ch^2 \int_\Omega f\mathfrak{A}$ due to Lemma 4.4.1. A simple inequality

$$\int_\Omega f\psi \le C\left(\int_\Omega |f|^2\right)^{1/2}\left(\int_\Omega |\psi|^2\right)^{1/2} \le C\|f\|_0 \quad \forall \psi \in \mathfrak{A}$$

implies that each component of $\mathbf{u}^\natural|_{\mathfrak{A}}$ is bounded by $\mathcal{O}(h^2)$. It estimates the difference between $u_h^\natural$ and $\bar{u}_h^\natural$ in L^2- and H^1-(semi-)norm. The following theorem states the relation between all numerical solutions discussed above.

Theorem 4.4.2 (Relation between numerical solutions)**.** *Let u_h, $u_h^\flat$, $u_h^\natural$, $\bar{u}_h^\natural$ be the numerical solutions of* (4.6) *as* (4.12), (4.16), (4.17), (4.21)*, respectively. Then $u_h = u_h^\flat = u_h^\natural$, and*

$$\|u_h^\natural - \bar{u}_h^\natural\|_0 \le Ch^2\|f\|_0, \ |u_h^\natural - \bar{u}_h^\natural|_{1,h} \le Ch\|f\|_0.$$

4.5 Numerical results

For the scheme option 1 in numerical tests, we use the restarted GMRES scheme in MGMRES library provided by Ju and Burkardt [33]. We emphasize that we replace one of essentially linearly dependent rows of $\mathbf{S}_h^{\mathfrak{B}^\flat}$ by the zero-integral condition in order to make $\tilde{\mathbf{S}}_h^{\mathfrak{B}^\flat}$ nonsingular.

The first example is the problem (4.6) on the domain $\Omega = (0,1)^2$ with the exact solution $u(x,y) = s(x)s(y)$ where

$$s(t) = \sum_{k=1}^{3} \frac{4}{(2k-1)\pi} \sin\Big(2(2k-1)\pi t\Big), \tag{4.22}$$

a truncated Fourier series for the square wave. For each option, the error in energy norm and L^2-norm are shown in Table 4.2. We can observe that all schemes give a very similar numerical solution.

The second example is the same problem with the exact solution $u(x,y) =$

$s(x)s(y)$ where

$$s(t) = \exp\left(-\frac{1}{1-(2t-1)^2}\right) t^2(1-t) + C, \tag{4.23}$$

with a constant C satisfying $\int_{[0,1]} s = 0$. Table 4.3 shows numerical results in each option, and all options give almost the same result, as the previous example. The iteration number and elapsed time in each option in case of $h = 1/256$ are shown in Table 4.4. We can observe decrease of the iteration number and elapsed time in option 3 compared to option 2. Decrease from option 3 to option 4 is quite natural because we only use the node based functions as trial and test functions in option 4.

	Opt 1				Opt 2			
h	$\|u-u_h\|_{1,h}$	order	$\|u-u_h\|_0$	order	$\|u-u_h\|_{1,h}$	order	$\|u-u_h\|_0$	order
1/8	1.123E+01	-	4.230E-01	-	1.123E+01	-	4.230E-01	-
1/16	5.466E-00	1.039	8.607E-02	2.297	5.466E-00	1.039	8.607E-02	2.297
1/32	2.832E-00	0.949	2.216E-02	1.957	2.832E-00	0.949	2.216E-02	1.957
1/64	1.429E-00	0.987	5.585E-03	1.989	1.429E-00	0.987	5.585E-03	1.989
1/128	7.160E-01	0.997	1.399E-03	1.997	7.160E-01	0.997	1.399E-03	1.997
1/256	3.582E-01	0.999	3.499E-04	1.999	3.582E-01	0.999	3.499E-04	1.999
	Opt 3				Opt 4			
h	$\|u-u_h\|_{1,h}$	order	$\|u-u_h\|_0$	order	$\|u-u_h\|_{1,h}$	order	$\|u-u_h\|_0$	order
1/8	1.123E+01	-	4.230E-01	-	1.123E+01	-	4.230E-01	-
1/16	5.466E-00	1.039	8.607E-02	2.297	5.466E-00	1.039	8.607E-02	2.297
1/32	2.832E-00	0.949	2.216E-02	1.957	2.832E-00	0.949	2.216E-02	1.957
1/64	1.429E-00	0.987	5.585E-03	1.989	1.429E-00	0.987	5.585E-03	1.989
1/128	7.160E-01	0.997	1.399E-03	1.997	7.160E-01	0.997	1.399E-03	1.997
1/256	3.582E-01	0.999	3.499E-04	1.999	3.582E-01	0.999	3.499E-04	1.999

Table 4.2. Numerical results for the exact solution with $s(t)$ as in (4.22)

	Opt 1				Opt 2			
h	$\|u-u_h\|_{1,h}$	order	$\|u-u_h\|_0$	order	$\|u-u_h\|_{1,h}$	order	$\|u-u_h\|_0$	order
1/8	1.225E-03	-	5.649E-05	-	1.225E-03	-	5.649E-05	-
1/16	6.024E-04	1.024	1.033E-05	2.450	6.024E-04	1.024	1.033E-05	2.450
1/32	3.045E-04	0.984	1.949E-06	2.406	3.045E-04	0.984	1.949E-06	2.406
1/64	1.527E-04	0.996	4.682E-07	2.058	1.527E-04	0.996	4.682E-07	2.058
1/128	7.642E-05	0.999	1.171E-07	1.999	7.642E-05	0.999	1.171E-07	1.999
1/256	3.822E-05	1.000	2.929E-08	2.000	3.822E-05	1.000	2.929E-08	2.000
	Opt 3				Opt 4			
h	$\|u-u_h\|_{1,h}$	order	$\|u-u_h\|_0$	order	$\|u-u_h\|_{1,h}$	order	$\|u-u_h\|_0$	order
1/8	1.225E-03	-	5.649E-05	-	1.225E-03	-	5.649E-05	-
1/16	6.024E-04	1.024	1.033E-05	2.450	6.024E-04	1.024	1.033E-05	2.450
1/32	3.045E-04	0.984	1.949E-06	2.406	3.045E-04	0.984	1.949E-06	2.406
1/64	1.527E-04	0.996	4.682E-07	2.058	1.527E-04	0.996	4.682E-07	2.058
1/128	7.642E-05	0.999	1.171E-07	1.999	7.642E-05	0.999	1.171E-07	1.999
1/256	3.822E-05	1.000	2.929E-08	2.000	3.822E-05	1.000	2.929E-08	2.000

Table 4.3. Numerical results for the exact solution with $s(t)$ as in (4.23)

	solver	iter	time (sec.)
Opt 1	GMRES(20)	4944	61.52
Opt 2	CG	817	3.30
Opt 3	CG	437	1.80
Opt 4	CG	318	1.33

Table 4.4. Iteration number and elapsed time in each option when $h = 1/256$

Chapter 5

Application to Stokes Equations

Suppose $\Omega = (0,1)^2 \subset \mathbb{R}^2$. Consider the incompressible Stokes equations with periodic boundary condition:

$$-\Delta \mathbf{u} + \nabla p = \mathbf{f} \quad \text{in } \Omega, \tag{5.1a}$$

$$\nabla \cdot \mathbf{u} = 0 \quad \text{in } \Omega, \tag{5.1b}$$

$$\mathbf{u} \text{ is periodic and } \int_\Omega \mathbf{u} \, \mathrm{d}\mathbf{x} = \mathbf{0}, \tag{5.1c}$$

$$p \text{ is periodic and } \int_\Omega p \, \mathrm{d}\mathbf{x} = 0 \tag{5.1d}$$

with the compatibility condition $\int_\Omega f = 0$. The corresponding weak formulation is as follows: *find* $(\mathbf{u}, p) \in [H^1_{per}(\Omega)/\mathbb{R}]^2 \times L^2_0(\Omega)$ *such that*

$$\int_\Omega \nabla \mathbf{u} : \nabla \mathbf{v} \, \mathrm{d}\mathbf{x} - \int_\Omega p \, \nabla \cdot \mathbf{v} \, \mathrm{d}\mathbf{x} = \int_\Omega \mathbf{f} \cdot \mathbf{v} \, \mathrm{d}\mathbf{x} \quad \forall \mathbf{v} \in [H^1_{per}(\Omega)/\mathbb{R}]^2, \tag{5.2a}$$

$$\int_\Omega q \, \nabla \cdot \mathbf{u} \, \mathrm{d}\mathbf{x} = 0 \quad \forall q \in L^2_0(\Omega) \tag{5.2b}$$

where $L_0^2(\Omega) := \{v \in L^2(\Omega) \mid \int_\Omega v = 0\}$. Then we can easily show that the inf-sup stability of $[H^1_{per}(\Omega)/\mathbb{R}]^2 \times L_0^2(\Omega)$: there exists $\beta > 0$ such that

$$\inf_{q \in L_0^2(\Omega)} \sup_{\mathbf{v} \in [H^1_{per}(\Omega)/\mathbb{R}]^2} \frac{\int_\Omega q \, \nabla \cdot \mathbf{v}}{|\mathbf{v}|_1 \, \|q\|_0} \geq \beta, \tag{5.3}$$

since $H_0^1(\Omega) \subset H^1_{per}(\Omega)$ and $[H_0^1(\Omega)]^2 \times L_0^2(\Omega)$ is inf-sup stable [29]. Thus there exists a unique solution $(\mathbf{u}, p) \in [H^1_{per}(\Omega)/\mathbb{R}]^2 \times L_0^2(\Omega)$ of (5.2).

5.1 Discrete inf-sup stability

Assume that $\mathcal{T}_h$ consists of uniform squares with same even number N_x and N_y. We need to define several discrete function spaces for velocity and pressure:

$$\begin{aligned}
V_0^h/\mathbb{R} &= \{v_h \in V^h \mid v_h = w_h - \frac{1}{|\Omega|}\int_\Omega w_h, \ w_h \in V_0^h\},\\
P^h &= \{p_h \in L^2(\Omega) \mid \ p_h|_K \in \mathcal{P}_0(K) \ \forall K \in \mathcal{T}_h\},\\
P_0^h &= \{p_h \in P^h \mid \int_\Omega p_h = 0\},\\
P_c^h &= \{p_h \in P_0^h \mid \sum_{K \in \mathcal{T}_h} \int_K p_h \, \nabla \cdot \mathbf{v}_h \, \mathrm{d}\mathbf{x} = 0 \ \forall \mathbf{v}_h \in [V_0^h]^2\},\\
P_{cf}^h &= \text{the } L^2(\Omega)\text{-orthogonal complement of } P_c^h \text{ in } P_0^h.
\end{aligned}$$

We denote the standard basis of P^h by $\mathfrak{P}$. Define two bilinear forms $a_h(\cdot,\cdot) : [V^h]^2 \times [V^h]^2 \to \mathbb{R}$, and $b_h(\cdot,\cdot) : [V^h]^2 \times P^h \to \mathbb{R}$ corresponding to the Laplace operator and the divergence operator, respectively, as follows: for all $\mathbf{v}_h, \mathbf{w}_h \in [V^h]^2$ and $q_h \in P^h$,

$$a_h(\mathbf{v}_h, \mathbf{w}_h) := \sum_{K \in \mathcal{T}_h} \int_K \nabla \mathbf{v}_h : \nabla \mathbf{w}_h \, \mathrm{d}\mathbf{x}, \quad b_h(\mathbf{v}_h, q_h) := -\sum_{K \in \mathcal{T}_h} \int_K q_h \, \nabla \cdot \mathbf{v}_h \, \mathrm{d}\mathbf{x}\,.$$

Consider the following discrete weak formulation: *find* $(\mathbf{u}_h, p_h) \in [V^h_{per}/\mathbb{R}]^2 \times P_0^h$

such that

$$a_h(\mathbf{u}_h, \mathbf{v}_h) + b_h(\mathbf{v}_h, p_h) = \int_\Omega \mathbf{f} \cdot \mathbf{v}_h \, \mathrm{d}\mathbf{x} \quad \forall \mathbf{v}_h \in [V_{per}^h/\mathbb{R}]^2, \tag{5.4a}$$

$$b_h(\mathbf{u}_h, q_h) = 0 \quad \forall q_h \in P_0^h. \tag{5.4b}$$

Our goal of this section is to prove the following theorem for discrete inf-sup stability.

Theorem 5.1.1. *$[V_{per}^h/\mathbb{R}]^2 \times P_0^h$ is uniformly discrete inf-sup stable, i.e., there exists a positive constant β which is independent of h such that*

$$\beta_h := \inf_{q_h \in P_0^h} \sup_{\mathbf{v}_h \in [V_{per}^h/\mathbb{R}]^2} \frac{b_h(\mathbf{v}_h, q_h)}{|\mathbf{v}_h|_{1,h} \|q_h\|_0} \geq \beta > 0.$$

We can prove the above theorem in help of results from the discrete formulation of the Stokes equations with homogeneous Dirichlet boundary condition. For the Stokes equations with homogeneous Dirichlet boundary condition, there exists the lowest order uniformly discrete inf-sup stable space pair as follows.

Theorem 5.1.2. *(Theorem 2.2, [35]) $[V_0^h]^2 \times P_{cf}^h$ satisfies the uniform discrete inf-sup condition.*

We quote the Subspace Theorem of Qin, which is useful in the proof.

Theorem 5.1.3. *([45]) Given $\mathbf{X}^h \times M^h$, let $\mathbf{X}_1$ and $\mathbf{X}_2$ be two subspaces of $\mathbf{X}^h$, and M_1 and M_2 be two subspaces of M^h. Assume the following three conditions hold:*

1. $M^h = M_1 + M_2$,

2. there exist $\beta_j > 0$, $j = 1, 2$, which are independent of h such that

$$\sup_{\mathbf{v}_j \in \mathbf{X}_j} \frac{b_h(\mathbf{v}_j, q_j)}{|\mathbf{v}_j|_{1,h}} \geq \beta_j \|q_j\|_0 \quad \forall q_j \in M_j,$$

3. there exist $\alpha_j \geq 0$, $j = 1, 2$, such that

$$|b_h(\mathbf{v}_j, q_k)| \leq \alpha_j |\mathbf{v}_j|_{1,h} \|q_k\|_0 \quad \forall \mathbf{v}_j \in \mathbf{X}_j, \forall q_k \in M_k,\ k \neq j,$$

with

$$\alpha_1 \alpha_2 \leq \beta_1 \beta_2.$$

Then, $\mathbf{X}^h \times M^h$ satisfies the inf-sup condition with the inf-sup constant depending only on $\alpha_1, \alpha_2, \beta_1, \beta_2$.

Proof of Theorem 5.1.1. Let us consider two subspaces of $[V^h_{per}/\mathbb{R}]^2$, namely, $[V^h_0/\mathbb{R}]^2$ and $[\mathrm{Span}\,\mathfrak{A}]^2$. We use Theorem 5.1.3 where $\mathbf{X}_1 = [V^h_0/\mathbb{R}]^2$, $\mathbf{X}_2 = [\mathrm{Span}\,\mathfrak{A}]^2$ and $M_1 = P^h_{cf}$, $M_2 = P^h_c$. Since P^h_{cf} is a subspace of P^h_0 which is complementary to P^h_c, the first condition holds.

For given $\mathbf{v}_h$ in $[V^h_0]^2$, let $\tilde{\mathbf{v}}_h$ denote $\mathbf{v}_h - \frac{1}{|\Omega|}\int_\Omega \mathbf{v}_h$, a trivial correspondent of $\mathbf{v}_h$ belonging to $[V^h_0/\mathbb{R}]^2$. Since $b_h(\tilde{\mathbf{v}}_h, q_h) = b_h(\mathbf{v}_h, q_h)\ \forall q_h \in P^h$ and $|\tilde{\mathbf{v}}_h|_1 = |\mathbf{v}_h|_1$, simple modification of Theorem 5.1.2 implies that $[V^h_0/\mathbb{R}]^2 \times P^h_{cf}$ is also uniformly discrete inf-sup stable.

On the other hand, we know that the dimension of P^h_c is just equal to 1, and it is generated by a global checkerboard pattern c_h, where $c_h|_{Q_{jk}} = (-1)^{j+k}$, see [42]. Take $\mathbf{w}_h = (\psi_x, 0)$ in $[\mathrm{Span}\,\mathfrak{A}]^2$. The definition of ψ_x yields $\nabla \cdot \mathbf{w}_h = \partial\psi_x/\partial x = (-1)^{j+k} 2/h$ on Q_{jk}. Thus $|b_h(\mathbf{w}_h, c_h)| = \sum_{Q_{jk}} \int_{Q_{jk}} 2h = 2/h$. Furthermore, $\|c_h\|^2_0 = \sum_{Q_{jk}} \int_{Q_{jk}} 1 = 1$, and $|\mathbf{w}_h|^2_1 = \sum_{Q_{jk}} \int_{Q_{jk}} |\nabla\psi_x|^2 =$

$\sum_{Q_{jk}} \int_{Q_{jk}} (\partial\psi_x/\partial x)^2 = \sum_{Q_{jk}} 4 = 4/h^2$. Therefore, for c_h,

$$\sup_{\mathbf{v}_h \in [\mathrm{Span}\,\mathfrak{A}]^2} \frac{b_h(\mathbf{v}_h, c_h)}{|\mathbf{v}_h|_1 \ \|c_h\|_0} \geq \frac{|b_h(\mathbf{w}_h, c_h)|}{|\mathbf{w}_h|_1 \ \|c_h\|_0} = 1.$$

Since P_c^h is generated by c_h, it implies the uniform discrete inf-sup stability of $[\mathrm{Span}\,\mathfrak{A}]^2 \times P_c^h$.

For the last condition, recall that $b_h(\mathbf{v}_h, q_h) = 0$ for all $\mathbf{v}_h \in [V_0^h]^2$ and $q_h \in P_c^h$. Note that every function in $[V_0^h/\mathbb{R}]^2$ is represented as $\tilde{\mathbf{v}}_h = \mathbf{v}_h - \frac{1}{|\Omega|}\int_\Omega \mathbf{v}_h$ for some $\mathbf{v}_h$ in $[V_0^h]^2$. Thus $b_h(\tilde{\mathbf{v}}_h, q_h) = 0$ for all $\tilde{\mathbf{v}}_h \in [V_0^h/\mathbb{R}]^2$ and $q_h \in P_c^h$. It implies $\alpha_1 = 0$, so the last condition is satisfied. Therefore we conclude that $[V_{per}^h/\mathbb{R}]^2 \times P_0^h$ satisfies the uniform discrete inf-sup condition. □

Theorem 5.1.1 leads the following error estimates [29, 12].

Theorem 5.1.4. *There exists a unique solution* $(\mathbf{u}_h, p_h) \in [V_{per}^h/\mathbb{R}]^2 \times P_0^h$ *of* (5.4)*, and*

$$|\mathbf{u} - \mathbf{u}_h|_{1,h} + \|p - p_h\|_0 \leq Ch(|\mathbf{u}|_2 + |p|_1).$$

5.2 Numerical scheme: Uzawa variant with a semi-definite block

Consider the set of trial and test functions consisting of $\mathfrak{E}$ for each component of the velocity variable as the option 3 in Section 4.4.3, and the standard basis $\mathfrak{P}$ of P^h for the pressure variable. It leads to the following system of equations in form of

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} u \\ p \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}, \tag{5.5}$$

where A is a symmetric positive semi-definite matrix with $dim \ker A = 4$. If the incompressible Stokes equations are considered, g in the right hand side vector becomes 0. We can easily show that the linear system (5.5) satisfies the following assumptions.

Assumption 5.1. *Let assume the followings.*

1. *A is symmetric positive semi-definite.*

2. $\ker A \subset \ker B$.

3. *$f \in \operatorname{Im} A$ and $g \in \operatorname{Im} B$.*

Recall the relation between the Drazin inverse and a Krylov solution of the equation $Ax = b$. Let k be the index of A which is the smallest nonnegative integer such that $\mathbb{C}^n = \operatorname{Im} A^k + \ker A^k$. If $b \in \operatorname{Im} A^k$, then the equation has a unique Krylov solution as $x^{\mathcal{K}} = A^D b$, *i.e.*, $A^D b$ is genuinely a solution of $Ax = b$ and is belonging to the Krylov space $\mathcal{K}_n(A, b)$. When A is symmetric or diagonalizable, the index of A is equal to 1. This leads the consistency condition $b \in \operatorname{Im} A$ for the existence of the Krylov solution.

Return to the our problem. Let A^D be the Drazin inverse of A. The equation in the first block in (5.5) is simplified as

$$Au = f - B^T p. \tag{5.6}$$

For any value p, the right hand side belongs to the image space of A because $\operatorname{Im} B^T \subset \operatorname{Im} A^T = \operatorname{Im} A$ from Assumption 5.1. The matrix equation (5.6) with respect to the variable u is symmetric and consistent. Thus, starting from an initial guess in $\operatorname{Im} A$, the equation has a unique Krylov solution associated with p, namely $u^{\mathcal{K}}(p) = A^D(f - B^T p) \in \operatorname{Im} A$. We can write the general solution of

(5.6) associated with p as

$$u(p) = u^{\mathcal{K}}(p) + u^{\circ},\ u^{\circ} \in \ker A. \tag{5.7}$$

We put the above expression into the equation in the second block in (5.5). Due to Assumption 5.1, it gives an equation which is containing variable p without u°,

$$BA^{D}B^{T}p = BA^{D}f - g. \tag{5.8}$$

We can easily observe that $BA^{D}B^{T}$ is also symmetric positive semi-definite. Furthermore, we can show the consistency of (5.8). Suppose $x \in \ker BA^{D}B^{T}$. Then $x^{T}BA^{D}B^{T}x = 0$ and thus we get $B^{T}x \in (\operatorname{Im} A)^{\perp}$ due to the characteristic of A^{D}. But Assumption 5.1 implies $(\operatorname{Im} A)^{\perp} = \ker A \subset \ker B = (\operatorname{Im} B^{T})^{\perp}$. Therefore $B^{T}x = 0$ and we get $\ker BA^{D}B^{T} \subset \ker B^{T}$. Since the converse is trivial, we conclude that $\ker BA^{D}B^{T} = \ker B^{T}$. As a consequence, we also get $\operatorname{Im} BA^{D}B^{T} = \operatorname{Im} B$, and (5.8) is consistent. Therefore starting from an initial guess in $\operatorname{Im} B$, there exists a unique Krylov solution $p_{*}^{\mathcal{K}} = (BA^{D}B^{T})^{D}(BA^{D}f - g) \in \operatorname{Im} B$. The general solution of (5.8) is

$$p = p_{*}^{\mathcal{K}} + p^{\circ},\ p^{\circ} \in \ker B^{T}. \tag{5.9}$$

Let $u_{*}^{\mathcal{K}}$ denote $u^{\mathcal{K}}(p_{*}^{\mathcal{K}})$, the Krylov solution of (5.6) associated with $p_{*}^{\mathcal{K}}$.

Note that the approach discussed above is a Uzawa variant for a singular block system. The numerical scheme to get $(u_{*}^{\mathcal{K}}, p_{*}^{\mathcal{K}})$ is described in Algorithm 1.

Now we discuss about properties of the solution obtained from the scheme. Recall that $(\mathbf{u}_h, p_h) \in [V_{per}^{h}/\mathbb{R}]^2 \times P_0^h$ is the solution of (5.4). Define the nu-

Algorithm 1 Uzawa method with conjugate directions and the Drazin inverse

$p_0 \leftarrow$ initial guess in $\operatorname{Im} B$
$u_1 \leftarrow A^D(f - B^T p_0)$ ▷ Use CG with an initial guess in $\operatorname{Im} A$
$q_1 \leftarrow g - Bu_1$
$d_1 \leftarrow -q_1$
while $k = 1, 2, \cdots$ **do**
    $s_k \leftarrow B^T d_k$
    $h_k \leftarrow A^D s_k$ ▷ Use CG with an initial guess in $\operatorname{Im} A$
    $\alpha_k \leftarrow (q_k^T q_k)/(s_k^T h_k)$
    $p_k \leftarrow p_{k-1} + \alpha_k d_k$
    $u_{k+1} \leftarrow u_k - \alpha_k h_k$
    $q_{k+1} \leftarrow g - Bu_{k+1}$
    $\beta_k \leftarrow (q_{k+1}^T q_{k+1})/(q_k^T q_k)$
    $d_{k+1} \leftarrow -q_{k+1} + \beta_k d_k$
end while

merical solution $(\mathbf{u}_h^\natural, p_h^\natural)$ corresponding to $(u_*^\mathcal{K}, p_*^\mathcal{K})$ by $\mathbf{u}_h^\natural := (u_{*,x}^\mathcal{K}\mathfrak{E}, u_{*,y}^\mathcal{K}\mathfrak{E})$ and $p_h^\natural := p_*^\mathcal{K}\mathfrak{P}$ where $u_*^\mathcal{K} = (u_{*,x}^\mathcal{K}, u_{*,y}^\mathcal{K})$. Clearly, $(\mathbf{u}_h^\natural, p_h^\natural) \in [V_{per}^h]^2 \times P^h$.

Lemma 5.2.1. *$(u_*^\mathcal{K}, p_*^\mathcal{K})$ truly solves (5.5).*

Proof. Note that the symmetry of A implies $\operatorname{Ind}(A) = 1$, thus $AA^Db = b$ for all $b \in \operatorname{Im} A$ [15]. Therefore we have

$$\begin{aligned} Au_*^\mathcal{K} + B^T p_*^\mathcal{K} &= AA^D(f - B^T p_*^\mathcal{K}) + B^T p_*^\mathcal{K} \\ &= AA^D f + (I - AA^D)B^T p_*^\mathcal{K} = f, \\ Bu_*^\mathcal{K} &= BA^D(f - B^T p_*^\mathcal{K}) \\ &= BA^D f - BA^D B^T p_*^\mathcal{K} \\ &= BA^D f - (BA^D f - g) = g. \end{aligned}$$

□

The following lemma shows that (5.7) and (5.9) truly represent the general solution of (5.5).

Lemma 5.2.2. *(v, q) is in the kernel space of* $\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix}$ *if and only if* $v \in \ker A$, $q \in \ker B^T$.

Proof. Suppose $v \in \ker A$ and $q \in \ker B^T$. Since $\ker A \subset \ker B$ from Assumption 5.1, we immediately get $Av + B^T q = 0$ and $Bv = 0$.

Conversely, suppose (v, q) belongs to the kernel space of the block matrix. It leads that v is a solution of the equation $Ax = -B^T q$, which is consistent. Thus $v = -A^D B^T q + v^\circ$ for some $v^\circ \in \ker A$. If we plug it into the second equation $Bv = 0$, we get $BA^D B^T q = 0$. Therefore q belongs to $\ker BA^D B^T$, which is equal to $\ker B^T$, as mentioned in lines above (5.9). It implies that v belongs to $\ker A$. □

Lemma 5.2.3. $(\mathbf{u}_h^\natural, p_h^\natural) \in [V_{per}^h/\mathbb{R}]^2 \times P_0^h$.

Proof. It is enough to show that $\int_\Omega \mathbf{u}_h^\natural = \mathbf{0}$, and $\int_\Omega p_h^\natural = 0$. Note that these are equivalent to $\mathbf{1}_\mathfrak{B} \cdot u_{*,x}^\mathcal{K}\big|_\mathfrak{B} = 0$, $\mathbf{1}_\mathfrak{B} \cdot u_{*,y}^\mathcal{K}\big|_\mathfrak{B} = 0$, and $\mathbf{1}_\mathfrak{P} \cdot p_*^\mathcal{K} = 0$.

Since $\mathbf{1}_\mathfrak{B} \in \ker \mathbf{S}_h^\mathfrak{B}$ and $u_{*,x}^\mathcal{K}\big|_\mathfrak{B} \in \operatorname{Im} \mathbf{S}_h^\mathfrak{B}$, the first two conditions are proved immediately. Since $b_h(\mathbf{v}_h, 1) = 0$ for all $\mathbf{v}_h \in [V_{per}^h]^2$, we have $\mathbf{1}_\mathfrak{P} \in \ker B^T$. Therefore $\mathbf{1}_\mathfrak{P} \cdot p_*^\mathcal{K} = 0$ because $p_*^\mathcal{K}$ belongs to $\operatorname{Im} B$, the space which is orthogonal to $\ker B^T$. □

The system of equations (5.5) is consistent with the system of equations derived from (5.4). Lemmas 5.2.1 and 5.2.3 imply $(\mathbf{u}_h^\natural, p_h^\natural)$ is a solution of (5.4) in $[V_{per}^h/\mathbb{R}]^2 \times P_0^h$. Due to the uniqueness in Theorem 5.1.4, we have the equivalence between two solution pairs.

Theorem 5.2.4. *Let $(\mathbf{u}_h^\natural, p_h^\natural)$ be the corresponding function to the Krylov solution $(u_*^\mathcal{K}, p_*^\mathcal{K})$ of (5.5) which is derived from the incompressible Stokes equations with periodic boundary condition (5.1). Then $(\mathbf{u}_h^\natural, p_h^\natural)$ is the solution of (5.4), i.e., $\mathbf{u}_h^\natural = \mathbf{u}_h$ and $p_h^\natural = p_h$.*

Next, we describe the discrete inf-sup constant of $[V^h_{per}/\mathbb{R}]^2 \times P^h_0$ in Theorem 5.1.1 in terms of A and B in (5.5). This is an analog of the work in [41] to a singular system.

Lemma 5.2.5. *Suppose $D^T z = \mathbf{0}$. Then $\inf_{\substack{z\cdot y=0,\\ y^T y=1}} \sup_{w^T w=1} y^T D w$ is the square root of the second smallest eigenvalue of DD^T.*

Proof. Without loss of generality, we assume that z is a unit vector. For fixed y, it is easily shown that $\sup_{w^T w=1} y^T D w = (y^T D D^T y)^{1/2}$. We can find a unitary matrix U, and a nonnegative diagonal matrix Σ such that

$$\begin{aligned} DD^T &= U\Sigma U^T \\ &= \begin{bmatrix} z & u_2 & \cdots & u_n \end{bmatrix} \begin{bmatrix} 0 & & & \\ & d_2 & & \\ & & \ddots & \\ & & & d_n \end{bmatrix} \begin{bmatrix} z^T \\ u_2^T \\ \vdots \\ u_n^T \end{bmatrix}. \end{aligned}$$

Since $z \cdot y = 0$ implies $y \in \mathrm{Span}\{u_2, \cdots, u_n\}$, the claim is derived in consequence. □

Theorem 5.2.6. *Let $M \in \mathbb{R}^{|\mathfrak{P}|\times|\mathfrak{P}|}$ be the mass matrix associated with the standard basis $\mathfrak{P}$ for P^h, with the Cholesky decomposition $M = GG^T$. Then the discrete inf-sup constant of $[V^h_{per}/\mathbb{R}]^2 \times P^h_0$ is the square root of the second smallest eigenvalue of $G^{-1}BA^D B^T G^{-T}$.*

Proof. Consider a nontrivial function $\mathbf{v}_h = \begin{pmatrix} v^x_h \\ v^y_h \end{pmatrix} \in [V^h_{per}/\mathbb{R}]^2$. There exists a vector $v = \begin{pmatrix} v^x \\ v^y \end{pmatrix} \in \mathbb{R}^{2|\mathfrak{E}|}$ such that $v^x_h = v^x \mathfrak{E}$ and $v^y_h = v^y \mathfrak{E}$. Let $v^x =$

$\begin{pmatrix} v^x_{\mathfrak{B}} \\ v^x_{\mathfrak{A}} \end{pmatrix}$ and $v^y = \begin{pmatrix} v^y_{\mathfrak{B}} \\ v^y_{\mathfrak{A}} \end{pmatrix}$. The zero-integral conditions $\int_\Omega v^x_h = \int_\Omega v^y_h = 0$ imply $\mathbf{1}_{\mathfrak{B}} \cdot v^x_{\mathfrak{B}} = \mathbf{1}_{\mathfrak{B}} \cdot v^y_{\mathfrak{B}} = 0$. Furthermore, without loss of generality, we can assume that $v^x_{\mathfrak{B}}$ and $v^y_{\mathfrak{B}}$ are orthogonal to the kernel space of $B^{\mathfrak{B}}_h$ because the representation is unique upto $\ker B^{\mathfrak{B}}_h$ (Section 4.1). Due to Proposition 4.3.3, we conclude that v is orthogonal to the kernel space of A, or equivalently $v \in \operatorname{Im} A$. Conversely, any $v \in \operatorname{Im} A$ corresponds to $\mathbf{v}_h$ in $[V^h_{per}/\mathbb{R}]^2$. Similarly, for every $q_h \in P^h_0$, there exists the corresponding $q \in \mathbb{R}^{|\mathfrak{P}|}$ such that $\mathbf{1}_{\mathfrak{P}} \cdot q = 0$, and vice versa. Thus,

$$\inf_{q_h \in P^h_0} \sup_{\mathbf{v}_h \in [V^h_{per}/\mathbb{R}]^2} \frac{b_h(\mathbf{v}_h, q_h)}{|\mathbf{v}|_{1,h} \|q_h\|_0} = \inf_{\substack{q \in \mathbb{R}^{|\mathfrak{P}|}, \\ \mathbf{1}_{\mathfrak{P}} \cdot q = 0}} \sup_{\substack{v \in \mathbb{R}^{2|\mathfrak{E}|}, \\ v \in \operatorname{Im} A}} \frac{q^T B v}{(v^T A v)^{1/2} (q^T M q)^{1/2}}. \tag{5.10}$$

Let $X \Lambda X^T$ be the eigendecomposition of A, where $X \in \mathbb{R}^{2|\mathfrak{E}| \times 2|\mathfrak{E}|}$ is a unitary matrix, and $\Lambda \in \mathbb{R}^{2|\mathfrak{E}| \times 2|\mathfrak{E}|}$ is a diagonal matrix with nonnegative entries. Since $\dim \ker A = 4$, we can rewrite as

$$A = \begin{bmatrix} X_m & \tilde{X}_m \end{bmatrix} \begin{bmatrix} \Lambda_m & \\ & 0 \end{bmatrix} \begin{bmatrix} X^T_m \\ \tilde{X}^T_m \end{bmatrix} = X_m \Lambda_m X^T_m$$

where $X_m \in \mathbb{R}^{2|\mathfrak{E}| \times (2|\mathfrak{E}|-4)}$ and $\Lambda_m \in \mathbb{R}^{(2|\mathfrak{E}|-4) \times (2|\mathfrak{E}|-4)}$ with positive diagonals. Note that $\operatorname{Im} A$ is equal to the column space of X_m. Therefore,

$$\begin{aligned} (5.10) &= \inf_{\substack{q \in \mathbb{R}^{|\mathfrak{P}|}, \\ \mathbf{1}_{\mathfrak{P}} \cdot q = 0}} \sup_{\substack{v \in \mathbb{R}^{2|\mathfrak{E}|}, \\ v \in \operatorname{Im} A}} \frac{q^T B v}{(v^T X_m \Lambda_m X^T_m v)^{1/2} (q^T G G^T q)^{1/2}} \\ &= \inf_{\substack{y \in \mathbb{R}^{|\mathfrak{P}|}, \\ y = G^T q, \\ (G^{-1} \mathbf{1}_{\mathfrak{P}}) \cdot y = 0}} \sup_{\substack{v_m \in \mathbb{R}^{2|\mathfrak{E}|-4}, \\ v = X_m v_m}} \frac{y^T G^{-1} B X_m v_m}{(v^T_m X^T_m X_m \Lambda_m X^T_m X_m v_m)^{1/2} (y^T y)^{1/2}} \end{aligned}$$

$$= \inf_{\substack{y\in\mathbb{R}^{|\mathfrak{P}|},\\ (G^{-1}\mathbf{1}_{\mathfrak{P}})\cdot y=0}} \sup_{v_m\in\mathbb{R}^{2|\mathfrak{E}|-4}} \frac{y^T G^{-1} B X_m v_m}{(v_m^T \Lambda_m v_m)^{1/2}(y^T y)^{1/2}}$$

$$= \inf_{\substack{y\in\mathbb{R}^{|\mathfrak{P}|},\\ (G^{-1}\mathbf{1}_{\mathfrak{P}})\cdot y=0}} \sup_{\substack{w_m\in\mathbb{R}^{2|\mathfrak{E}|-4},\\ w_m=\Lambda_m^{1/2}v_m}} \frac{y^T G^{-1} B X_m \Lambda_m^{-1/2} w_m}{(w_m^T w_m)^{1/2}(y^T y)^{1/2}}.$$

Simple calculation shows $\mathbf{1}_{\mathfrak{P}}$ is an eigenvector of M, and also that of M^{-1}. Thus it holds that

$$\begin{aligned}(G^{-1}BX_m\Lambda_m^{-1/2})^T(G^{-1}\mathbf{1}_{\mathfrak{P}}) &= (\Lambda_m^{-1/2})^T X_m^T B^T G^{-T} G^{-1}\mathbf{1}_{\mathfrak{P}} \\ &= (\Lambda_m^{-1/2})^T X_m^T B^T M^{-1}\mathbf{1}_{\mathfrak{P}} \\ &= \frac{1}{\lambda}(\Lambda_m^{-1/2})^T X_m^T B^T \mathbf{1}_{\mathfrak{P}} = \mathbf{0}.\end{aligned}$$

In the last line, we use $B^T\mathbf{1}_{\mathfrak{P}} = \mathbf{0}$, since $b_h(\mathbf{v}_h, 1) = -\sum_{K\in\mathcal{T}_h}\int_K \nabla\cdot\mathbf{v}_h \,\mathrm{d}\mathbf{x} = -\sum_{K\in\mathcal{T}_h}\int_{\partial K}\nu\cdot\mathbf{v}_h \,\mathrm{ds} = 0$ for all $\mathbf{v}_h \in [V_{per}^h/\mathbb{R}]^2$. Due to Lemma 5.2.5, the discrete inf-sup constant in (5.10) is equal to the square root of the second smallest eigenvalue of

$$\begin{aligned}(G^{-1}BX_m\Lambda_m^{-1/2})(G^{-1}BX_m\Lambda_m^{-1/2})^T &= G^{-1}BX_m\Lambda_m^{-1}X_m^T B^T G^{-T} \\ &= G^{-1}BA^D B^T G^{-T}.\end{aligned}$$

We refer to [15] for the matrix representation of the Drazin inverse used in the last line. □

5.3 Numerical results

On $\Omega = (0,1)^2$, consider the periodic incompressible Stokes equations (5.1) with the exact solution pair for the velocity and pressure

$$\mathbf{u}(x,y) = \nabla \times \big(\sin(2\pi x)s(y)\big),$$
$$p(x,y) = \sin(2\pi x)\cos(2\pi y)$$

where $s(t) = \exp\left(-\frac{1}{1-(2t-1)^2}\right)\big(1-(2t-1)^2\big)+C$ with a constant C satisfying $\int_{[0,1]} s = 0$. The results on Table 5.1 show optimal convergence order in various norms.

We compute the discrete inf-sup constant of $[V^h_{per}/\mathbb{R}]^2 \times P^h_0$ as in Theorem 5.2.6. And, for a comparison, we also consider the trial and test functions based on the option 4 in Section 4.4.4; just $\mathfrak{B}$ instead of $\mathfrak{E}$ for each component of the velocity. This combination of functions corresponds to the space pair $[V^{\mathfrak{B},h}_{per}/\mathbb{R}]^2 \times P^h_0$. The numerically computed 4 smallest eigenvalues $\lambda_1 \le \lambda_2 \le \lambda_3 \le \lambda_4$ of $G^{-1}BA^DB^TG^{-T}$, and the discrete inf-sup constant $\beta_h = \sqrt{\lambda_2}$ for each option are shown in Table 5.2. We can observe that the discrete inf-sup constant based on the option 3 is bounded below by a positive number which does not depend on the mesh size, as expected. The results confirm our theoretical claims for the inf-sup stability in Theorem 5.1.1. On the other hand, the second smallest numerically computed eigenvalue in the scheme based on the option 4 is comparable to the machine epsilon, which means nearly zero. Thus we can conclude that the discrete inf-sup constant of $[V^{\mathfrak{B},h}_{per}/\mathbb{R}]^2 \times P^h_0$ is almost equal to zero. It is a consequence of the simple fact that $b_h(\mathbf{v}_h, c_h) = 0$ for all $\mathbf{v}_h \in [V^{\mathfrak{B},h}_{per}]^2$, where c_h is a piecewise constant function in global checkerboard pattern.

	Opt 3					
	velocity				pressure	
h	$\|\mathbf{u}-\mathbf{u}_h\|_{1,h}$	order	$\|\mathbf{u}-\mathbf{u}_h\|_0$	order	$\|p-p_h\|_0$	order
1/8	3.018E-00	-	9.105E-02	-	5.686E-01	-
1/16	1.449E-00	1.058	1.655E-02	2.460	9.541E-02	2.575
1/32	7.462E-01	0.957	4.869E-03	1.765	4.550E-02	1.068
1/64	3.733E-01	0.999	1.169E-03	2.058	2.057E-02	1.145
1/128	1.868E-01	0.999	2.937E-04	1.993	1.009E-02	1.028
1/256	9.341E-02	1.000	7.347E-05	1.999	5.019E-03	1.007

Table 5.1. Numerical results based on the option 3 for the Stokes equations

h	λ_1	λ_2	λ_3	λ_4	β_h
		Opt 3			
1/8	9.437E-16	1.000	1.000	1.000	1.000
1/16	-2.776E-16	1.000	1.000	1.000	1.000
1/32	-2.331E-15	1.000	1.000	1.000	1.000
1/64	-1.144E-14	1.000	1.000	1.000	1.000
		Opt 4			
1/8	-4.594E-16	1.527E-17	1.000	1.000	(≈ 0)
1/16	-6.708E-16	-3.284E-16	1.000	1.000	(≈ 0)
1/32	-1.703E-15	-1.516E-15	1.000	1.000	(≈ 0)
1/64	-2.223E-16	2.443E-15	1.000	1.000	(≈ 0)

Table 5.2. Numerically computed eigenvalues and discrete inf-sup constant

Chapter 6

3-D Case

In this chapter, we consider the case of $d = 3$. Following similar discussions as in 2-D case, we will get 3-D results.

6.1 Dimension of finite spaces in 3-D

The following lemma is 3-D analog of Lemma 3.2.1.

Lemma 6.1.1. *For* $\Omega \subset \mathbb{R}^3$,

$$\begin{aligned}&(\textit{dimension of finite space})\\ &\quad = \#(\textit{faces}) - 2\#(\textit{cells})\\ &\qquad - \#(\textit{minimally essential discrete boundary conditions}).\end{aligned}$$

Proof. We can rewrite the dice rule in a single 3-D cubic cell $K \in \mathcal{T}_h$ into two

separated relations:

$$v_h(m_1^K) - v_h(m_2^K) + v_h(m_6^K) - v_h(m_5^K) = 0,$$
$$v_h(m_1^K) - v_h(m_3^K) + v_h(m_6^K) - v_h(m_4^K) = 0$$

for all $v_h \in V^h$ where m_j^K is the center point of a face f_j^K of K, and the faces are arranged to satisfy that the sum of indices in opposite faces is equal to 7, as an ordinary dice. Since each relation reduces the number of degrees of freedom in the finite space by 1, same as 2-D case, the claim is derived in consequence. □

Proposition 6.1.2. *(Neumann and Dirichlet B.C. in 3-D)*

$$\#(\textit{minimally essential discrete boundary conditions}) = \begin{cases} 0 & \textit{in the case of Neumann B.C.,} \\ 2(N_xN_y + N_yN_z + N_zN_x) & \\ \qquad - (N_x + N_y + N_z) + 1 & \textit{in the case of homo. Dirichlet B.C.} \end{cases}$$

Consequently,

$$\dim V^h = N_xN_yN_z + N_xN_y + N_yN_z + N_zN_x, \tag{6.1a}$$
$$\dim V_0^h = (N_x - 1)(N_y - 1)(N_z - 1). \tag{6.1b}$$

Proof. It is enough to consider the homogeneous Dirichlet boundary case since there is nothing to prove in the Neumann case. Suppose that the homogeneous Dirichlet boundary condition is given. Similar to the argument in 2-D, we need to investigate induced relations on boundary DoF values. Consider x-direction first, and classify all cells into N_x groups by their position in x.

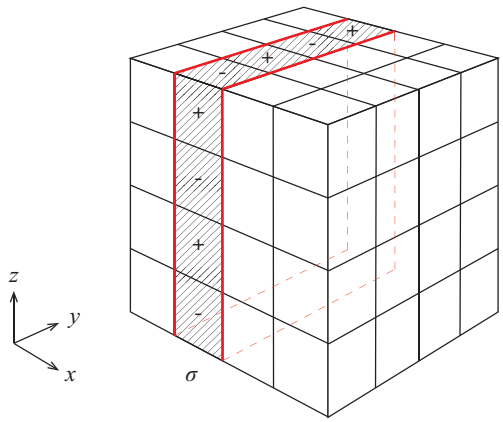

Figure 6.1. An example of a strip

Then each group consists of $N_y \times N_z$ cells which are attached in y- and z-direction. For each cell in a group, the dice rule in 3-D implies a relation between 4 DoFs on faces which are parallel to xy- or zx-plane. A collection of such relations from all cells in a group derives a single relation between DoF values on a set of boundary faces, called a *strip* perpendicular to x-axis, similarly to the 2-D case. Precisely speaking, an alternating sum of $2N_y + 2N_z$ boundary DoF values on the strip is equal to zero. This induced relation on the strip is well-defined because the number of faces in the strip is always even. Figure 6.1 shows an example of a strip perpendicular to x-axis. The signs on the strip represent the alternating sum of boundary DoF values. For x-direction, there are N_x relations between DoFs on boundary faces corresponding to N_x strips perpendicular to x-axis, respectively. We can continue to discuss similar arguments for y- and z-direction. Consequently, we can find totally $N_x + N_y + N_z$ strips and corresponding relations between boundary DoFs.

However, it is not true that these induced relations are linearly independent. Choose a cube K at one of corners in $\mathcal{T}_h$. There are three strips σ_K^x, σ_K^y, σ_K^z which are attached to K, and perpendicular to x-, y-, z-axis, respectively.

Let each of these strips call *the standard strip* for each axis. There are two options to give proper alternating sign to DoF values on each standard strip in order to make corresponding alternating relation between boundary DoFs. For each standard strip, we choose an option for alternating sign in the relation to cancel out all boundary DoFs belong to K when summing up all three relations on three standard strips. We call them *the standard choices*. Consider σ, a strip among others, which is obviously parallel to one of these standard strips, without loss of generality, σ_K^x. There are also two options for alternating sign in the relation on σ. One option is same to the standard choice on σ_K^x: in this option, the sign for each boundary DoF on σ is equal to the sign for the corresponding boundary DoF in the standard choice on σ_K^x. The other option is just opposite to the standard choice. We make a choice on σ depending on the distance from σ_K^x. If σ is adjacent to σ_K^x, or is away from σ_K^x by an even number of faces in x-direction, then we choose an option for alternating sign on σ to be opposite to the standard choice on σ_K^x. Else if σ is away from σ_K^x by an odd number of faces in x-direction, then the same alternating sign as the standard choice is chosen on σ. Under this rule, we can make all choices for alternating sign in the induced relations on all $N_x + N_y + N_z$ strips. And it can be easily shown that the sum of all induced relations on all strips with chosen alternating sign becomes a trivial relation. It implies that there is a single linear relation between those induced relations on all strips. Therefore,

$$\begin{aligned}
&\#(\text{minimally essential discrete boundary conditions}) \\
&\quad = \#(\text{boundary faces}) - \#(\text{independent relations}) \\
&\quad = 2(N_xN_y + N_yN_z + N_zN_x) - (N_x + N_y + N_z - 1).
\end{aligned}$$

□

Proposition 6.1.3. *(Periodic B.C. in 3-D) Let* $\epsilon_j := (1 + (-1)^j)/2$. *In case of periodic boundary condition,*

$$\begin{aligned}
&\#(\textit{minimally essential discrete boundary conditions}) \\
&\quad = (N_x N_y + N_y N_z + N_z N_x) \\
&\qquad - (N_x \epsilon_{N_y} \epsilon_{N_z} + N_y \epsilon_{N_x} \epsilon_{N_z} + N_z \epsilon_{N_x} \epsilon_{N_y}) + \epsilon_{N_x} \epsilon_{N_y} \epsilon_{N_z}.
\end{aligned}$$

Consequently,

$$\dim V^h_{per} = \begin{cases} N_x N_y N_z + (N_x + N_y + N_z) - 1 & \textit{if all } N_x,\ N_y,\ N_z \textit{ are even}, \\ N_x N_y N_z + N_\iota & \textit{if only } N_\iota \textit{ is odd}, \\ N_x N_y N_z & \textit{else}. \end{cases} \tag{6.2}$$

Proof. Note that ϵ_j is equal to 1 for even j and 0 for odd. Due to the same reason discussed in 2-D case, an induced relation between boundary DoFs on a strip perpendicular to x-axis can help to impose periodic boundary condition only when both N_y and N_z are even. In this case, coincidence of two DoF values of the last boundary face pair is naturally achieved by pairwise coincidence of DoF values of other boundary face pairs in the strip. Consequently, totally N_x periodic boundary conditions can hold naturally due to other periodic boundary conditions and induced boundary relations on strips perpendicular to x-axis. Similar claims hold for induced boundary relations on strips perpendicular to y-, and z-directional axis.

However, as discussed in the case of Dirichlet boundary condition, due to the linear dependence between $N_x + N_y + N_z$ induced relations on all strips we have to consider 1 redundant relation when all $N_x + N_y + N_z$ strips are meaningful, *i.e.*, all N_x, N_y and N_z are even. It completes the claims. □

6.2 Linear dependence of $\mathfrak{B}$ in 3-D

In this section, we identify a global coefficient representation for node based functions in $\mathfrak{B}$ with a vector in $\mathbb{R}^{|\mathfrak{B}|}$. With this identification, we use a vector $\mathbf{c} \in \mathbb{R}^{|\mathfrak{B}|}$ to represent a global coefficient representation on given 3-D grid $\mathcal{T}_h$. In this sense, we denote the local coefficients of $\mathbf{c}$ in a cube $Q \in \mathcal{T}_h$ by $\mathbf{c}|_Q$. For the sake of simple description, we use this abusive notation as long as there is no chance of misunderstanding. A surjective linear map $B_h^{\mathfrak{B}} : \mathbb{R}^{|\mathfrak{B}|} \to V_{per}^{\mathfrak{B},h}$ is defined as in Chapter 4, but for 3-D case.

As shown in Figure 6.2, there are exactly 4 kinds of local coefficient representation for the zero function in a single cube. The value at each vertex represents the coefficient for the corresponding node based function in $\mathfrak{B}$. If any global coefficient representation for the zero function is restricted in a cube, then it has to be a linear combination of these 4 elementary representations which are denoted by $\mathcal{A}, \mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$, respectively. In other words, any global representation for the zero function is obtained by consecutive extension of local representation in appropriate way.

Define the following subspaces consisting of global representations:

$$\begin{aligned}
\mathcal{S}_{\mathcal{X}\mathcal{Y}\mathcal{Z}\mathcal{A}} &:= \left\{ \mathbf{c} \in \mathbb{R}^{|\mathfrak{B}|} \mid \mathbf{c}|_Q \in \mathrm{Span}\{\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{A}\} \ \forall Q \in \mathcal{T}_h \right\}, \\
\mathcal{S}_{\mathcal{X}\mathcal{A}} &:= \left\{ \mathbf{c} \in \mathbb{R}^{|\mathfrak{B}|} \mid \mathbf{c}|_Q \in \mathrm{Span}\{\mathcal{X}, \mathcal{A}\} \ \forall Q \in \mathcal{T}_h \right\}, \\
\mathcal{S}_{\mathcal{Y}\mathcal{A}} &:= \left\{ \mathbf{c} \in \mathbb{R}^{|\mathfrak{B}|} \mid \mathbf{c}|_Q \in \mathrm{Span}\{\mathcal{Y}, \mathcal{A}\} \ \forall Q \in \mathcal{T}_h \right\}, \\
\mathcal{S}_{\mathcal{Z}\mathcal{A}} &:= \left\{ \mathbf{c} \in \mathbb{R}^{|\mathfrak{B}|} \mid \mathbf{c}|_Q \in \mathrm{Span}\{\mathcal{Z}, \mathcal{A}\} \ \forall Q \in \mathcal{T}_h \right\}, \\
\mathcal{S}_{\mathcal{A}} &:= \left\{ \mathbf{c} \in \mathbb{R}^{|\mathfrak{B}|} \mid \mathbf{c}|_Q \in \mathrm{Span}\{\mathcal{A}\} \ \forall Q \in \mathcal{T}_h \right\}.
\end{aligned}$$

Remark 6.2.1. *$\mathcal{S}_{\mathcal{X}\mathcal{Y}\mathcal{Z}\mathcal{A}}$, $\mathcal{S}_{\mathcal{X}\mathcal{A}}$, $\mathcal{S}_{\mathcal{Y}\mathcal{A}}$, $\mathcal{S}_{\mathcal{Z}\mathcal{A}}$, and $\mathcal{S}_{\mathcal{A}}$ are truly vector spaces.*

Remark 6.2.2. *The definition of $B_h^{\mathfrak{B}}$ implies* $\ker B_h^{\mathfrak{B}} = \mathcal{S}_{\mathcal{X}\mathcal{Y}\mathcal{Z}\mathcal{A}}$.

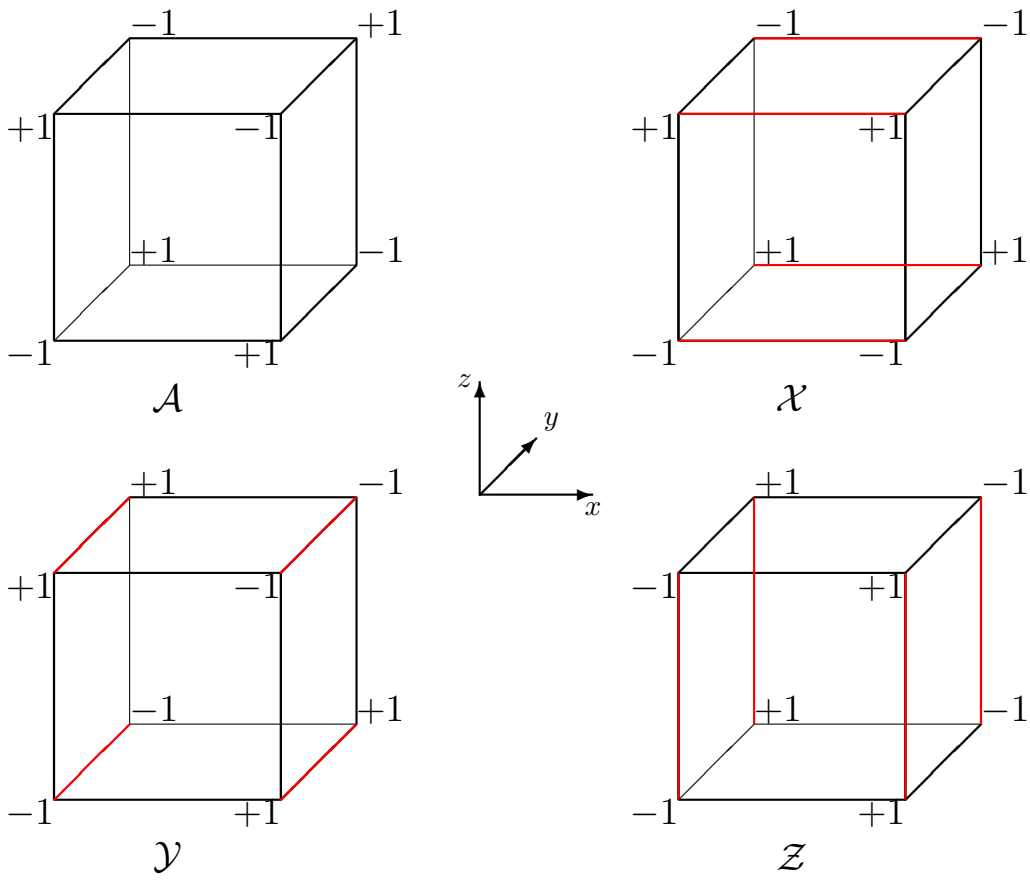

Figure 6.2. Nontrivial representations for the zero function in a cube: $\mathcal{A}$, $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$

Lemma 6.2.3. *Let* $\mathbf{c} \in \mathcal{S}_{\mathcal{X}\mathcal{Y}\mathcal{Z}\mathcal{A}}$, *and* $c^{\mathcal{X}}_{ijk}, c^{\mathcal{Y}}_{ijk}, c^{\mathcal{Z}}_{ijk}, c^{\mathcal{A}}_{ijk}$ *denote coefficients of* $\mathbf{c}$ *in a cube* $Q_{ijk} \in \mathcal{T}_h$ *for* $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$, $\mathcal{A}$, *respectively,* i.e., $\mathbf{c}|_{Q_{ijk}} = c^{\mathcal{X}}_{ijk}\mathcal{X} + c^{\mathcal{Y}}_{ijk}\mathcal{Y} + c^{\mathcal{Z}}_{ijk}\mathcal{Z} + c^{\mathcal{A}}_{ijk}\mathcal{A}$. *Then for all* $1 \le i \le N_x$, $1 \le j \le N_y$, $1 \le k \le N_z$,

$$c^{\mathcal{X}}_{ijk} - c^{\mathcal{A}}_{ijk} = c^{\mathcal{X}}_{(i+1)jk} + c^{\mathcal{A}}_{(i+1)jk}, \tag{6.3a}$$

$$c^{\mathcal{Y}}_{ijk} = -c^{\mathcal{Y}}_{(i+1)jk}, \tag{6.3b}$$

$$c^{\mathcal{Z}}_{ijk} = -c^{\mathcal{Z}}_{(i+1)jk}, \tag{6.3c}$$

$$c^{\mathcal{X}}_{ijk} = -c^{\mathcal{X}}_{i(j+1)k}, \tag{6.4a}$$

$$c^{\mathcal{Y}}_{ijk} - c^{\mathcal{A}}_{ijk} = c^{\mathcal{Y}}_{i(j+1)k} + c^{\mathcal{A}}_{i(j+1)k}, \tag{6.4b}$$

$$c^{\mathcal{Z}}_{ijk} = -c^{\mathcal{Z}}_{i(j+1)k}, \tag{6.4c}$$

$$c^{\mathcal{X}}_{ijk} = -c^{\mathcal{X}}_{ij(k+1)}, \tag{6.5a}$$

$$c^{\mathcal{Y}}_{ijk} = -c^{\mathcal{Y}}_{ij(k+1)}, \tag{6.5b}$$

$$c^{\mathcal{Z}}_{ijk} - c^{\mathcal{A}}_{ijk} = c^{\mathcal{Z}}_{ij(k+1)} + c^{\mathcal{A}}_{ij(k+1)}. \tag{6.5c}$$

Here all indices are understood in modulo N_x, N_y, N_z, respectively, due to the periodicity.

Remark 6.2.4. *Conversely, local relations* (6.3)–(6.5) *in Lemma 6.2.3 for all* $1 \le i \le N_x$, $1 \le j \le N_y$, $1 \le k \le N_z$ *imply well-definedness of* $\mathbf{c} \in \mathcal{S}_{\mathcal{X}\mathcal{Y}\mathcal{Z}\mathcal{A}}$, i.e., *on each face shared by two adjacent cubes the vertex values are matching.*

Proof of Lemma 6.2.3. These relations are nothing, but just the matching conditions on every face which is shared by two adjacent cubes.

Two cubes Q_{ijk} and $Q_{(i+1)jk}$ are adjacent in x-direction, and sharing a common face perpendicular to x-axis. Thus the vertex values on the right face of the left cube Q_{ijk} have to be matched with the vertex values on the left face of the right cube $Q_{(i+1)jk}$. Since there are 4 nodes in the common face, we have 4 equations in 8 variables:

$$-c^{\mathcal{X}}_{ijk} + c^{\mathcal{Y}}_{ijk} + c^{\mathcal{Z}}_{ijk} + c^{\mathcal{A}}_{ijk} = -c^{\mathcal{X}}_{(i+1)jk} - c^{\mathcal{Y}}_{(i+1)jk} - c^{\mathcal{Z}}_{(i+1)jk} - c^{\mathcal{A}}_{(i+1)jk}, \tag{6.6a}$$

$$c^{\mathcal{X}}_{ijk} + c^{\mathcal{Y}}_{ijk} - c^{\mathcal{Z}}_{ijk} - c^{\mathcal{A}}_{ijk} = \ c^{\mathcal{X}}_{(i+1)jk} - c^{\mathcal{Y}}_{(i+1)jk} + c^{\mathcal{Z}}_{(i+1)jk} + c^{\mathcal{A}}_{(i+1)jk}, \tag{6.6b}$$

$$c^{\mathcal{X}}_{ijk} - c^{\mathcal{Y}}_{ijk} + c^{\mathcal{Z}}_{ijk} - c^{\mathcal{A}}_{ijk} = \ c^{\mathcal{X}}_{(i+1)jk} + c^{\mathcal{Y}}_{(i+1)jk} - c^{\mathcal{Z}}_{(i+1)jk} + c^{\mathcal{A}}_{(i+1)jk}, \tag{6.6c}$$

$$-c^{\mathcal{X}}_{ijk} - c^{\mathcal{Y}}_{ijk} - c^{\mathcal{Z}}_{ijk} + c^{\mathcal{A}}_{ijk} = -c^{\mathcal{X}}_{(i+1)jk} + c^{\mathcal{Y}}_{(i+1)jk} + c^{\mathcal{Z}}_{(i+1)jk} - c^{\mathcal{A}}_{(i+1)jk}. \tag{6.6d}$$

Simple calculation shows that (6.6) are equivalent to (6.3). Similarly, considering faces perpendicular to y- and z-direction, we get (6.4) and (6.5). □

The next decomposition theorem is essential for the dimension analysis in 3-D case.

Theorem 6.2.5 (Decomposition Thoerem)**.** *The quotient space $\mathcal{S}_{\mathcal{XYZA}}/\mathcal{S}_{\mathcal{A}}$ can be decomposed as*

$$\mathcal{S}_{\mathcal{XYZA}}/\mathcal{S}_{\mathcal{A}} = \mathcal{S}_{\mathcal{XA}}/\mathcal{S}_{\mathcal{A}} \oplus \mathcal{S}_{\mathcal{YA}}/\mathcal{S}_{\mathcal{A}} \oplus \mathcal{S}_{\mathcal{ZA}}/\mathcal{S}_{\mathcal{A}}. \tag{6.7}$$

Proof. Clearly $\mathcal{S}_{\mathcal{A}} \subset \mathcal{S}_{\mathcal{XA}}, \mathcal{S}_{\mathcal{YA}}, \mathcal{S}_{\mathcal{ZA}} \subset \mathcal{S}_{\mathcal{XYZA}}$ and $\mathcal{S}_{\mathcal{XA}} \cap \mathcal{S}_{\mathcal{YA}} = \mathcal{S}_{\mathcal{YA}} \cap \mathcal{S}_{\mathcal{ZA}} = \mathcal{S}_{\mathcal{ZA}} \cap \mathcal{S}_{\mathcal{XA}} = \mathcal{S}_{\mathcal{A}}$. Thus it is enough to show that for any $\mathbf{c} \in \mathcal{S}_{\mathcal{XYZA}}$, there exist $\mathbf{u} \in \mathcal{S}_{\mathcal{XA}}$, $\mathbf{v} \in \mathcal{S}_{\mathcal{YA}}$, $\mathbf{w} \in \mathcal{S}_{\mathcal{ZA}}$ such that $\mathbf{c} \in \mathbf{u} + \mathbf{v} + \mathbf{w} + \mathcal{S}_{\mathcal{A}}$.

Let $c^{\mathcal{X}}_{ijk}$, $c^{\mathcal{Y}}_{ijk}$, $c^{\mathcal{Z}}_{ijk}$, $c^{\mathcal{A}}_{ijk}$ denote the coefficients of $\mathbf{c}$ in a cube $Q_{ijk} \in \mathcal{T}_h$ for $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$, $\mathcal{A}$, respectively, *i.e.*, $\mathbf{c}|_{Q_{ijk}} = c^{\mathcal{X}}_{ijk}\mathcal{X} + c^{\mathcal{Y}}_{ijk}\mathcal{Y} + c^{\mathcal{Z}}_{ijk}\mathcal{Z} + c^{\mathcal{A}}_{ijk}\mathcal{A}$. Due to Lemma 6.2.3, the relations (6.3)–(6.5) hold. Now we construct $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$. First, define $\mathbf{u} \in \mathbb{R}^{|\mathfrak{B}|}$ by

$$\mathbf{u}|_{Q_{ijk}} := u^{\mathcal{X}}_{ijk}\mathcal{X} + u^{\mathcal{A}}_{ijk}\mathcal{A} \quad \text{where} \quad u^{\mathcal{X}}_{ijk} = c^{\mathcal{X}}_{ijk}, u^{\mathcal{A}}_{ijk} = (-1)^{j+k} c^{\mathcal{A}}_{i11}. \tag{6.8}$$

The above definition naturally implies that $u^{\mathcal{Y}}_{ijk} = u^{\mathcal{Z}}_{ijk} = 0$. We can check the followings.

1. $\mathbf{u}$ is well-defined, and belongs to $\mathcal{S}_{\mathcal{XYZA}}$: See Remark 6.2.4. For a face shared by two adjacent cubes Q_{ijk} and $Q_{(i+1)jk}$,

$$\begin{aligned} u^{\mathcal{X}}_{ijk} - u^{\mathcal{A}}_{ijk} &= c^{\mathcal{X}}_{ijk} - (-1)^{j+k} c^{\mathcal{A}}_{i11} \\ &= (-1) c^{\mathcal{X}}_{i(j-1)k} - (-1)^{j+k} c^{\mathcal{A}}_{i11} \\ &= \cdots \end{aligned}$$

$$
\begin{aligned}
&= (-1)^{(j-1)} c^{\mathcal{X}}_{i1k} - (-1)^{j+k} c^{\mathcal{A}}_{i11} \\
&= (-1)^{(j-1)+1} c^{\mathcal{X}}_{i1(k-1)} - (-1)^{j+k} c^{\mathcal{A}}_{i11} \\
&= \cdots \\
&= (-1)^{(j-1)+(k-1)} c^{\mathcal{X}}_{i11} - (-1)^{j+k} c^{\mathcal{A}}_{i11} \\
&= (-1)^{j+k} (c^{\mathcal{X}}_{i11} - c^{\mathcal{A}}_{i11}) \\
&= (-1)^{j+k} (c^{\mathcal{X}}_{(i+1)11} + c^{\mathcal{A}}_{(i+1)11}) \\
&= (-1)^{j+k} c^{\mathcal{X}}_{(i+1)11} + (-1)^{j+k} c^{\mathcal{A}}_{(i+1)11} \\
&= \cdots \\
&= c^{\mathcal{X}}_{(i+1)jk} + (-1)^{j+k} c^{\mathcal{A}}_{(i+1)11} \\
&= u^{\mathcal{X}}_{(i+1)jk} + u^{\mathcal{A}}_{(i+1)jk}.
\end{aligned}
$$

Thus $\mathbf{u}$ is matching on all faces perpendicular to x-axis. For the faces perpendicular to y-axis, it holds that

$$
\begin{aligned}
u^{\mathcal{X}}_{ijk} &= c^{\mathcal{X}}_{ijk} = -c^{\mathcal{X}}_{i(j+1)k} = -u^{\mathcal{X}}_{i(j+1)k}, \\
-u^{\mathcal{A}}_{ijk} &= -(-1)^{j+k} c^{\mathcal{A}}_{i11} = (-1)^{j+1+k} c^{\mathcal{A}}_{i11} = u^{\mathcal{A}}_{i(j+1)k},
\end{aligned}
$$

and similar for the faces perpendicular to z-axis. Therefore $\mathbf{u}$ is also matching along y- and z-direction.

2. $\mathbf{u} \in \mathcal{S}_{\mathcal{X}\mathcal{A}}$: it is trivial due to the definition of $\mathbf{u}$ and $\mathcal{S}_{\mathcal{X}\mathcal{A}}$.

Similarly above, we define $\mathbf{v}$ and $\mathbf{w} \in \mathbb{R}^{|\mathfrak{B}|}$ by

$$
\mathbf{v}|_{Q_{ijk}} := v^{\mathcal{Y}}_{ijk} \mathcal{Y} + v^{\mathcal{A}}_{ijk} \mathcal{A} \quad \text{where} \quad v^{\mathcal{Y}}_{ijk} = c^{\mathcal{Y}}_{ijk}, v^{\mathcal{A}}_{ijk} = (-1)^{i+k} c^{\mathcal{A}}_{1j1}, \tag{6.9}
$$

$$
\mathbf{w}|_{Q_{ijk}} := w^{\mathcal{Z}}_{ijk} \mathcal{Z} + w^{\mathcal{A}}_{ijk} \mathcal{A} \quad \text{where} \quad w^{\mathcal{Z}}_{ijk} = c^{\mathcal{Z}}_{ijk}, w^{\mathcal{A}}_{ijk} = (-1)^{i+j} c^{\mathcal{A}}_{11k}. \tag{6.10}
$$

Then both $\mathbf{v}$ and $\mathbf{w}$ are well-defined, and $\mathbf{v} \in \mathcal{S}_{\mathcal{Y}\mathcal{A}}$, $\mathbf{w} \in \mathcal{S}_{\mathcal{Z}\mathcal{A}}$. Thus $\mathbf{c} - (\mathbf{u} +$

$\mathbf{v}+\mathbf{w}) \in \mathcal{S}_{\mathcal{X}\mathcal{Y}\mathcal{Z}\mathcal{A}}$, and for each cube Q_{ijk} it holds that

$$\mathbf{c}-(\mathbf{u}+\mathbf{v}+\mathbf{w})|_{Q_{ijk}} = \left(c^{\mathcal{A}}_{ijk} - (-1)^{j+k}c^{\mathcal{A}}_{i11} - (-1)^{i+k}c^{\mathcal{A}}_{1j1} - (-1)^{i+j}c^{\mathcal{A}}_{11k}\right)\mathcal{A}.$$

Therefore we conclude that $\mathbf{c}-(\mathbf{u}+\mathbf{v}+\mathbf{w}) \in \mathcal{S}_{\mathcal{A}}$. □

Corollary 6.2.6. $\dim \ker B^{\mathfrak{B}}_h = \dim \mathcal{S}_{\mathcal{X}\mathcal{A}} + \dim \mathcal{S}_{\mathcal{Y}\mathcal{A}} + \dim \mathcal{S}_{\mathcal{Z}\mathcal{A}} - 2\dim \mathcal{S}_{\mathcal{A}}$.

The following lemmas explain the dimension of subspaces which depends on parity of the discretization numbers.

Lemma 6.2.7. *(The dimension of $\mathcal{S}_{\mathcal{X}\mathcal{A}}$, $\mathcal{S}_{\mathcal{Y}\mathcal{A}}$, $\mathcal{S}_{\mathcal{Z}\mathcal{A}}$)*

$$\dim \mathcal{S}_{\mathcal{X}\mathcal{A}} = \begin{cases} N_x & \textit{if both } N_y \textit{ and } N_z \textit{ are even,} \\ 0 & \textit{else.} \end{cases} \tag{6.11a}$$

$$\dim \mathcal{S}_{\mathcal{Y}\mathcal{A}} = \begin{cases} N_y & \textit{if both } N_x \textit{ and } N_z \textit{ are even,} \\ 0 & \textit{else.} \end{cases} \tag{6.11b}$$

$$\dim \mathcal{S}_{\mathcal{Z}\mathcal{A}} = \begin{cases} N_z & \textit{if both } N_x \textit{ and } N_y \textit{ are even,} \\ 0 & \textit{else.} \end{cases} \tag{6.11c}$$

Proof. It is enough to show the claim for $\mathcal{S}_{\mathcal{X}\mathcal{A}}$, since the others can be shown similarly. Let $\mathbf{c} \in \mathcal{S}_{\mathcal{X}\mathcal{A}}$ where $\mathbf{c}|_{Q_{ijk}} = c^{\mathcal{X}}_{ijk}\mathcal{X} + c^{\mathcal{A}}_{ijk}\mathcal{A}$ in each cube $Q_{ijk} \in \mathcal{T}_h$. By applying matching conditions (6.4) and (6.5) consecutively, it is shown

$$c^{\mathcal{X}}_{ijk} = (-1)^{j+k}c^{\mathcal{X}}_{i11} \quad \text{and} \quad c^{\mathcal{A}}_{ijk} = (-1)^{j+k}c^{\mathcal{A}}_{i11}.$$

Consider N_x+1 combined surfaces such that each of them consists of $N_y \times N_z$ faces in $\mathcal{T}_h$, and is lying on the same hyperplane perpendicular to x-axis. The above relations imply that on each surface the coefficients for node based

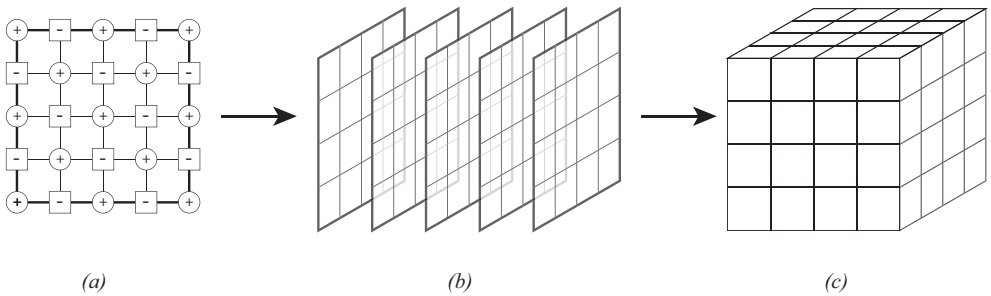

Figure 6.3. Construction of a global representation for a function in $\mathcal{S}_{\mathcal{XA}}$

functions are all the same, but with alternating sign like a checkerboard pattern at nodes, not on faces. Due to the identification between boundary nodes in y- and z-direction, all coefficients vanish unless both N_y and N_z are even.

Under the case of even N_y and N_z, we consider a basis checkerboard pattern at nodes on a combined surface consisting of $+1$ and -1, alternatively, as Figure 6.3 (a) shows. In the figure, the plus and minus sign at nodes represent the positive value one, and the negative value one, respectively. We get N_x+1 checkerboard patterns on N_x+1 combined surfaces in series (Figure 6.3 (b)). Based on the basis checkerboard pattern described in above, we can represent all coefficients on each combined surface by a single factor in real number. Due to the identification between boundary nodes in x-direction, two factors for the first and the last combined surface must be identical. Then the series of N_x+1 checkerboard patterns compose a global representation for a function in $\mathcal{S}_{\mathcal{XA}}$ (Figure 6.3 (c)). Conversely, for the N_x+1 combined surfaces which are perpendicular to x-axis and the basis checkerboard pattern at nodes on surfaces, suppose N_x+1 factors are given, where the first and the last of them are equal. Then we can determine unique $c^{\mathcal{X}}_{ijk}$ and $c^{\mathcal{A}}_{ijk}$, for all $Q_{ijk} \in \mathcal{T}_h$. Therefore, only in the case when both N_y and N_z are even, $\mathcal{S}_{\mathcal{XA}}$ is equivalent to $\{\mathbf{v} \in \mathbb{R}^{N_x+1} \mid v_1 = v_{N_x+1}\}$, and consequently $\dim \mathcal{S}_{\mathcal{XA}} = N_x$. □

Lemma 6.2.8. *(The dimension of $\mathcal{S}_{\mathcal{A}}$)*

$$\dim \mathcal{S}_{\mathcal{A}} = \begin{cases} 1 & \text{if all } N_x,\ N_y,\ N_z \text{ are even,} \\ 0 & \text{else.} \end{cases} \tag{6.12}$$

Proof. Let $\mathbf{c} \in \mathcal{S}_{\mathcal{A}}$ where $\mathbf{c}|_{Q_{ijk}} = c_{ijk}^{\mathcal{A}}\mathcal{A}$ in each cube Q_{ijk}. By applying matching conditions (6.3)–(6.5) consecutively, it is shown

$$c_{ijk}^{\mathcal{A}} = (-1)^{i+j+k+1} c_{111}^{\mathcal{A}}.$$

Due to the identification of boundary nodes in x-, y-, and z-direction, all coefficients vanish unless all N_x, N_y and N_z are even. In the case of all even N_x, N_y and N_z, it is easily shown that the coefficients form a multiple of *the 3-D checkerboard pattern* at nodes. Therefore $\dim \mathcal{S}_{\mathcal{A}} = 1$. □

Proposition 6.2.9. *(The dimension of* $\ker B_h^{\mathfrak{B}}$, $V_{per}^{\mathfrak{B},h}$ *in 3-D)*

$$\dim \ker B_h^{\mathfrak{B}} = \begin{cases} N_x + N_y + N_z - 2 & \text{if all } N_x,\ N_y,\ N_z \text{ are even,} \\ N_\iota & \text{if only } N_\iota \text{ is odd,} \\ 0 & \text{else.} \end{cases} \tag{6.13}$$

Consequently,

$$\begin{aligned} \dim V_{per}^{\mathfrak{B},h} &= |\mathfrak{B}| - \dim \ker B_h^{\mathfrak{B}} \\ &= \begin{cases} N_x N_y N_z - (N_x + N_y + N_z) + 2 & \text{if all } N_x,\ N_y,\ N_z \text{ are even,} \\ N_x N_y N_z - N_\iota & \text{if only } N_\iota \text{ is odd,} \\ N_x N_y N_z & \text{else.} \end{cases} \end{aligned} \tag{6.14}$$

Proof. Direct consequences of Corollary 6.2.6, Lemmas 6.2.7 and 6.2.8. □

6.3 A basis for V_{per}^h in 3-D

Propositions 6.1.3 and 6.2.9 imply that $V_{per}^{\mathfrak{B},h}$ is a proper subset of V_{per}^h if at most one of N_x, N_y, and N_z is odd. Furthermore, if all N_x, N_y, and N_z are even, then there exist $2(N_x + N_y + N_z) - 3$ complementary basis functions for V_{per}^h, not belonging to $V_{per}^{\mathfrak{B},h}$. If only N_ι is odd, then the number of complementary basis functions for V_{per}^h is $2N_\iota$. In other cases, $V_{per}^{\mathfrak{B},h}$ is equal to V_{per}^h. We will discuss about the complementary basis functions below.

For the first case, suppose that all N_x, N_y, and N_z are even. Consider N_x strips σ_i^x, $1 \leq i \leq N_x$, which are perpendicular to x-axis. Each strip σ_i^x defines a subdomain Ω_i^x, the union of $N_y \times N_z$ cubes which are wrapped up with σ_i^x. Let $(\psi_i^x)_y$ denote a piecewise linear function in V_{per}^h whose support is Ω_i^x as follows. Within Ω_i^x it has nonzero DoF values only on faces perpendicular to y-axis, and all the nonzero DoF values are 1 with alternating sign in y- and z-direction, as similar to the alternating function ψ_x in 2-D case (Figure 6.4 (a), (b)). The alternating function $(\psi_i^x)_y$ is obtained by trivial extending to Ω (Figure 6.4 (c)). A similar argument as in 2-D case, it is easily shown that $(\psi_i^x)_y$ is well-defined, and not belonging to $V_{per}^{\mathfrak{B},h}$ since N_y and N_z are even. A similar property holds for $(\psi_i^x)_z$, a piecewise linear function in V_{per}^h whose support is Ω_i^x and which has nonzero DoF values as 1 only on faces perpendicular to z-axis with alternating sign in y- and z-direction. Thus totally there exist $2N_x$ alternating functions $\{(\psi_i^x)_y, (\psi_i^x)_z\}_{1 \leq i \leq N_x}$ for V_{per}^h associated with strips perpendicular to x-axis. By considering other strips perpendicular to y- or z-axis, we can find out $2(N_x + N_y + N_z)$ alternating functions for V_{per}^h, not belonging to $V_{per}^{\mathfrak{B},h}$: $\{(\psi_i^x)_y, (\psi_i^x)_z, (\psi_j^y)_x, (\psi_j^y)_z, (\psi_k^z)_x, (\psi_k^z)_y\}_{1 \leq i \leq N_x, 1 \leq j \leq N_y, 1 \leq k \leq N_z}$.

However, there is a single relation between the alternating functions in each direction on subscript. An alternating sum of $(\psi_i^x)_z$ in $1 \leq i \leq N_x$ is equal to that of $(\psi_j^y)_z$ in $1 \leq j \leq N_y$. And any $N_x + N_y - 1$ among all $(\psi_i^x)_z$ and $(\psi_j^y)_z$ are

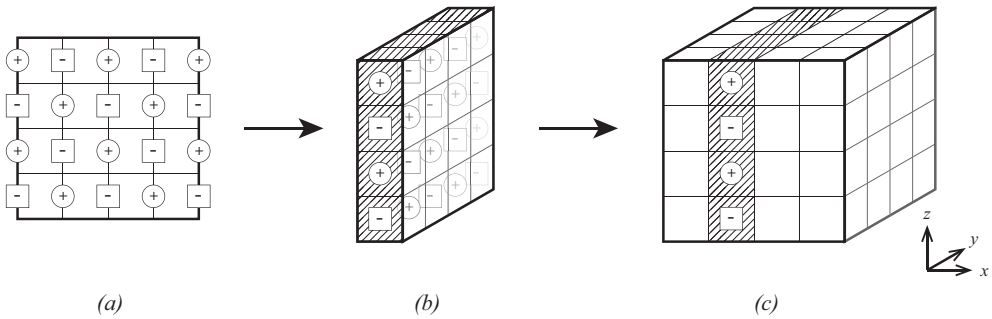

Figure 6.4. Construction of an alternating function in 3-D

linearly independent due to their supports. Similarly, any $N_y + N_z - 1$ among all $(\psi_j^y)_x$ and $(\psi_k^z)_x$ are linearly independent, and so any $N_z + N_x - 1$ among all $(\psi_k^z)_y$ and $(\psi_i^x)_y$ are. Consequently, suitably chosen $2(N_x + N_y + N_z) - 3$ alternating functions form a complementary basis for V_{per}^h.

In the case of only one odd N_ι (and even N_μ, N_ν), the set of all alternating functions associated to the strips perpendicular to ι-axis, $\{(\psi_j^\iota)_\mu, (\psi_j^\iota)_\nu\}_{1\le j\le N_\iota}$, are meaningful because N_μ and N_ν are even.

Theorem 6.3.1. *(A complementary basis for V_{per}^h in 3-D)*

1. *If all N_x, N_y, and N_z are even, then $V_{per}^{\mathfrak{B},h}$ is a proper subset of V_{per}^h. The union of*
   - *any $N_x + N_y - 1$ among $\mathfrak{A}_z := \{(\psi_i^x)_z, (\psi_j^y)_z\}_{1\le i\le N_x, 1\le j\le N_y}$,*
   - *any $N_y + N_z - 1$ among $\mathfrak{A}_x := \{(\psi_j^y)_x, (\psi_k^z)_x\}_{1\le j\le N_y, 1\le k\le N_z}$, and*
   - *any $N_z + N_x - 1$ among $\mathfrak{A}_y := \{(\psi_k^z)_y, (\psi_i^x)_y\}_{1\le i\le N_x, 1\le k\le N_z}$*

   is a complementary basis for V_{per}^h, not belonging to $V_{per}^{\mathfrak{B},h}$.

2. *If only N_ι is odd (and N_μ, N_ν are even), then $V_{per}^{\mathfrak{B},h}$ is a proper subset of V_{per}^h. And $\{(\psi_j^\iota)_\mu, (\psi_j^\iota)_\nu\}_{1\le j\le N_\iota}$ is a complementary basis for V_{per}^h, not belonging to $V_{per}^{\mathfrak{B},h}$.*

3. *Else, $V_{per}^{\mathfrak{B},h} = V_{per}^h$.*

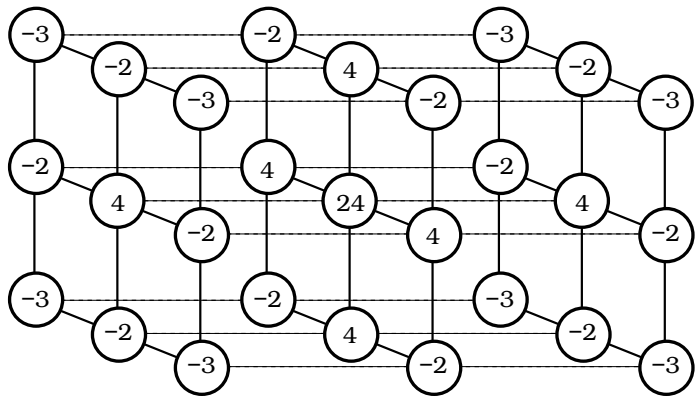

Figure 6.5. The stencil for $\mathbf{S}_h^{\mathfrak{B}}$ in 3-D

6.4 Stiffness matrix associated with $\mathfrak{B}$ in 3-D

The stiffness matrix $\mathbf{S}_h^{\mathfrak{B}}$ is defined as in (4.3) but in 3-D space. See Figure 6.5 for the 3-D local stencil for the stiffness matrix associated with $\mathfrak{B}$.

Proposition 6.4.1. *(The dimension of* $\ker \mathbf{S}_h^{\mathfrak{B}}$ *in 3-D)*

$$\dim \ker \mathbf{S}_h^{\mathfrak{B}} = \begin{cases} N_x + N_y + N_z - 1 & \text{if all } N_x,\ N_y,\ N_z \text{ are even,} \\ N_\iota + 1 & \text{if only } N_\iota \text{ is odd,} \\ 1 & \text{else.} \end{cases} \tag{6.15}$$

Proof. It is a direct consequence of Propositions 6.2.9 and 4.3.3. □

We numerically assemble $\mathbf{S}_h^{\mathfrak{B}}$ for various combinations of N_x, N_y, and N_z. The rank deficiency can be computed in help of well-known numerical tools or libraries, for instance MATLAB or LAPACK. Table 6.1 shows numerically obtained rank deficiency of the stiffness matrix associated with $\mathfrak{B}$ in 3-D space. Numbers in red represent the case of all even discretizations. Blue is for the case of odd discretization in only one direction, and black for the other cases. These numerical results confirm our theoretical result in Proposition 6.4.1.

$N_z = \mathbf{2}$		N_y						
		2	3	**4**	5	**6**	7	**8**
N_x	**2**	**5**						
	3	**4**	1					
	4	**7**	**4**	**9**				
	5	**6**	1	**6**	1			
	6	**9**	**4**	**11**	**6**	**13**		
	7	**8**	1	**8**	1	**8**	1	
	8	**11**	**4**	**13**	**6**	**15**	**8**	**17**

$N_z = 3$							
	2	3	**4**	5	**6**	7	**8**
2							
3		1					
4		1	**4**				
5		1	1	1			
6		1	**4**	1	**4**		
7		1	1	1	1	1	
8		1	**4**	1	**4**	1	**4**

$N_z = 4$							
	2	3	**4**	5	**6**	7	**8**
2							
3							
4			**11**				
5			**6**	1			
6			**13**	**6**	**15**		
7			**8**	1	**8**	1	
8			**15**	**6**	**17**	**8**	**19**

$N_z = 5$							
	2	3	**4**	5	**6**	7	**8**
2							
3							
4							
5				1			
6				1	**6**		
7				1	1	1	
8				1	**6**	1	**6**

Table 6.1. Numerically obtained rank deficiency of $\mathbf{S}_h^{\mathfrak{B}}$ in 3-D

6.5 Numerical schemes in 3-D

Consider again an elliptic problem with periodic boundary condition (4.6) with the compatibility condition $\int_\Omega f = 0$, the corresponding weak formulation (4.7), and the corresponding discrete weak formulation (4.8) in 3-D. Throughout this section, we assume that all N_x, N_y, and N_z are even. $\mathfrak{B}^\flat$ again denotes a basis for V_{per}^h, a proper subset of $\mathfrak{B}$. Note that we have known what the cardinality of $\mathfrak{B}^\flat$ is, but the way to find $\mathfrak{B}^\flat$ is not constructive yet. Let $\mathfrak{A}$ and $\mathfrak{A}^\flat$ be the set of all alternating functions, and a complementary basis for V_{per}^h which consists of alternating functions as in Theorem 6.3.1, respectively. Without loss of generality, we write $\mathfrak{B}^\flat = \{\phi_j\}_{j=1}^{|\mathfrak{B}^\flat|}$, $\mathfrak{B} = \{\phi_j\}_{j=1}^{|\mathfrak{B}|}$, $\mathfrak{A}^\flat = \{\psi_j\}_{j=1}^{|\mathfrak{A}^\flat|}$, and $\mathfrak{A} = \{\psi_j\}_{j=1}^{|\mathfrak{A}|}$. Define two extended sets $\mathfrak{E} := \mathfrak{B} \cup \mathfrak{A}$, and $\mathfrak{E}^\flat := \mathfrak{B}^\flat \cup \mathfrak{A}^\flat$. Even in 3-D case, $\mathfrak{E}^\flat$ is a basis for V_{per}^h.

Remark 6.5.1. *Unlike 2-D case, $\mathfrak{A}$ may not be linearly independent in 3-D case. Thus we use $\mathfrak{A}^\flat$, a linearly independent subset, instead of $\mathfrak{A}$ to construct*

$\mathcal{S}$	$\lvert\mathcal{S}\rvert$	Span $\mathcal{S}$	dim Span $\mathcal{S}$
$\mathfrak{B}^\flat$	$N_xN_yN_z-(N_x+N_y+N_z)+2$	$V_{per}^{\mathfrak{B},h}$	$N_xN_yN_z-(N_x+N_y+N_z)+2$
$\mathfrak{B}$	$N_xN_yN_z$		
$\mathfrak{E}^\flat$	$N_xN_yN_z+(N_x+N_y+N_z)-1$	V_{per}^{h}	$N_xN_yN_z+(N_x+N_y+N_z)-1$
$\mathfrak{E}$	$N_xN_yN_z+2(N_x+N_y+N_z)$		

Table 6.2. Summary of characteristics of $\mathfrak{B}^\flat$, $\mathfrak{B}$, $\mathfrak{E}^\flat$, $\mathfrak{E}$ in 3-D when all N_x, N_y, N_z are even

$\mathfrak{E}^\flat$ *as a basis for* V_{per}^h.

Lemma 6.5.2. *Let* $\mathfrak{B}$ *and* $\mathfrak{A}$ *be as above. Then the followings hold.*

1. $a_h(\phi,\psi)=0 \quad \forall\phi\in\mathfrak{B}\ \forall\psi\in\mathfrak{A}$.
2. $\int_\Omega \psi = 0 \quad \forall\psi\in\mathfrak{A}$.
3. *There exists an* h*-independent constant* C *such that* $\|\psi\|_0 \le Ch^{1/2}$ *and* $|\psi|_{1,h}\le Ch^{-1/2} \quad \forall\psi\in\mathfrak{A}$.

Remark 6.5.3. *The second equation in Lemma 4.4.1 does not hold in 3-D case. If* $\mu=\nu$, *then* $a_h((\psi^\iota)_\mu,(\psi^\lambda)_\nu)$ *does not vanish in general.*

For 3-D case, we define $\mathbf{S}_h^{\mathfrak{B}^\flat}$, $\tilde{\mathbf{S}}_h^{\mathfrak{B}^\flat}$, and $\mathbf{S}_h^{\mathfrak{A}}$ as in (4.9)–(4.11), respectively. Furthermore we define $\mathbf{S}_h^{\mathfrak{A}^\flat}$, the stiffness matrix associated with $\mathfrak{A}^\flat$ in similar manner. Define the linear systems $\tilde{\mathcal{L}}_h^{\mathfrak{E}^\flat}$, $\mathcal{L}_h^{\mathfrak{E}^\flat}$ as in (4.13), (4.14), with slight modification since $\mathfrak{E}^\flat$ is equal to $\mathfrak{B}^\flat\cup\mathfrak{A}^\flat$ in 3-D case. Other linear systems $\mathcal{L}_h^{\mathfrak{E}}$, $\mathcal{L}_h^{\mathfrak{B}}$ are defined as in (4.18), (4.20). The solutions $\tilde{\mathbf{u}}^\flat$, $\mathbf{u}^\flat$, $\mathbf{u}^\natural$, $\bar{\mathbf{u}}^\natural$, and the numerical solutions u_h, $u_h^\flat$, $u_h^\natural$, $\bar{u}_h^\natural$ are defined as in (4.13)–(4.15), (4.18)–(4.20), (4.12), (4.16), (4.17), (4.21).

In the following, we compare these numerical solutions as in Section 4.4. The equality between u_h and $u_h^\flat$ is clear. The next is for comparison between $u_h^\flat$ and $u_h^\natural$.

Since $\mathfrak{B}^\flat$ is a basis for $V_{per}^{\mathfrak{B},h}$, there exist $t_{\ell j} \in \mathbb{R}$ for $1 \le \ell \le |\mathfrak{B}| - |\mathfrak{B}^\flat|$ and $1 \le j \le |\mathfrak{B}^\flat|$, such that

$$\phi_{|\mathfrak{B}^\flat|+\ell} = \sum_{j=1}^{|\mathfrak{B}^\flat|} t_{\ell j}\phi_j. \tag{6.16}$$

Thus $\sum_{K\in\mathcal{T}_h} \nabla\phi_k \cdot \nabla \left(\phi_{|\mathfrak{B}^\flat|+\ell} - \sum_{j=1}^{|\mathfrak{B}^\flat|} t_{\ell j}\phi_j\right) \mathrm{d}\mathbf{x} = 0$ for all k, and it is simplified as $(\mathbf{S}_h^{\mathfrak{B}})_{|\mathfrak{B}^\flat|+\ell,k} = \sum_{j=1}^{|\mathfrak{B}^\flat|} t_{\ell j}(\mathbf{S}_h^{\mathfrak{B}})_{jk}$. Let $\mathbf{T}$ denote a matrix of size $(|\mathfrak{B}| - |\mathfrak{B}^\flat|)\times|\mathfrak{B}^\flat|$ such that $(\mathbf{T})_{\ell j} = t_{\ell j}$. Then the last equation for $1 \le \ell \le |\mathfrak{B}|-|\mathfrak{B}^\flat|$ and $1 \le k \le |\mathfrak{B}^\flat|$ can be expressed as a matrix equation

$$[\mathbf{S}_h^{\mathfrak{B}}]_{|\mathfrak{B}^\flat|+1:|\mathfrak{B}|,1:|\mathfrak{B}^\flat|} = \mathbf{T}[\mathbf{S}_h^{\mathfrak{B}}]_{1:|\mathfrak{B}^\flat|,1:|\mathfrak{B}^\flat|}. \tag{6.17}$$

Note that $[\mathbf{S}_h^{\mathfrak{B}}]_{1:|\mathfrak{B}^\flat|,1:|\mathfrak{B}^\flat|}$ is just equal to $\mathbf{S}_h^{\mathfrak{B}^\flat}$. Let $\begin{bmatrix}\mathbf{u}^\flat|_{\mathfrak{B}^\flat}\\ \mathbf{0}\end{bmatrix}$ be a trivial extension of $\mathbf{u}^\flat|_{\mathfrak{B}^\flat}$ into a vector in $\mathbb{R}^{|\mathfrak{B}|}$ by padding zeros. Then

$$\mathbf{S}_h^{\mathfrak{B}}\begin{bmatrix}\mathbf{u}^\flat|_{\mathfrak{B}^\flat}\\ \mathbf{0}\end{bmatrix} = \begin{bmatrix}\mathbf{S}_h^{\mathfrak{B}^\flat}\mathbf{u}^\flat|_{\mathfrak{B}^\flat}\\ [\mathbf{S}_h^{\mathfrak{B}}]_{|\mathfrak{B}^\flat|+1:|\mathfrak{B}|,1:|\mathfrak{B}^\flat|}\mathbf{u}^\flat|_{\mathfrak{B}^\flat}\end{bmatrix} = \begin{bmatrix}\mathbf{S}_h^{\mathfrak{B}^\flat}\mathbf{u}^\flat|_{\mathfrak{B}^\flat}\\ \mathbf{T}\mathbf{S}_h^{\mathfrak{B}^\flat}\mathbf{u}^\flat|_{\mathfrak{B}^\flat}\end{bmatrix} = \begin{bmatrix}\int_\Omega f\mathfrak{B}^\flat\\ \mathbf{T}\int_\Omega f\mathfrak{B}^\flat\end{bmatrix}$$

since $\mathbf{S}_h^{\mathfrak{B}^\flat}\mathbf{u}^\flat|_{\mathfrak{B}^\flat} = \int_\Omega f\mathfrak{B}^\flat$. We can easily derive

$$\mathbf{T}\int_\Omega f\mathfrak{B}^\flat = \mathbf{T}\begin{bmatrix}\int_\Omega f\phi_1\\ \vdots\\ \int_\Omega f\phi_{|\mathfrak{B}^\flat|}\end{bmatrix} = \begin{bmatrix}\int_\Omega f\sum_{j=1}^{|\mathfrak{B}^\flat|} t_{1j}\phi_j\\ \vdots\\ \int_\Omega f\sum_{j=1}^{|\mathfrak{B}^\flat|} t_{|\mathfrak{B}^\flat|j}\phi_j\end{bmatrix} = \begin{bmatrix}\int_\Omega f\phi_{|\mathfrak{B}^\flat|+1}\\ \vdots\\ \int_\Omega f\phi_{|\mathfrak{B}|}\end{bmatrix},$$

which implies that

$$\mathbf{S}_h^{\mathfrak{B}}\begin{bmatrix}\mathbf{u}^\flat|_{\mathfrak{B}^\flat}\\ \mathbf{0}\end{bmatrix} = \int_\Omega f\mathfrak{B}.$$

In the same way we obtain that $\mathbf{S}_h^{\mathfrak{A}} \begin{bmatrix} \mathbf{u}^\flat|_{\mathfrak{A}^\flat} \\ \mathbf{0} \end{bmatrix} = \int_\Omega f\mathfrak{A}$. Therefore we can conclude the equality of $u_h^\natural$ and $u_h^\flat$ by the same argument in 2-D case.

For the last, consider the difference between $u_h^\natural$ and $\bar{u}_h^\natural$. We can easily observe that $u_h^\natural - \bar{u}_h^\natural = \mathbf{u}^\natural\big|_{\mathfrak{A}}\,\mathfrak{A}$, and $a_h(u_h^\natural - \bar{u}_h^\natural, \psi) = \int_\Omega f\psi$ for all $\psi \in \mathfrak{A}$. Thus

$$\begin{aligned} |u_h^\natural - \bar{u}_h^\natural|_{1,h}^2 &= a_h\left(u_h^\natural - \bar{u}_h^\natural, u_h^\natural - \bar{u}_h^\natural\right) \\ &= \int_\Omega f(u_h^\natural - \bar{u}_h^\natural) \le C\|f\|_0\,\|u_h^\natural - \bar{u}_h^\natural\|_0 = Ch\,\|f\|_0\,|u_h^\natural - \bar{u}_h^\natural|_{1,h} \end{aligned}$$

due to the following lemma, and we immediately obtain the difference in mesh-dependent norm, and in L^2-norm.

Lemma 6.5.4. *Let $\mathcal{M}_h^{\mathfrak{A}}$ be the mass matrix associated with $\mathfrak{A}$. Then there exists an h-independent constant C such that $\mathcal{M}_h^{\mathfrak{A}} = Ch^2\mathcal{S}_h^{\mathfrak{A}}$. In a consequence, $\|v_h\|_0 = C^{1/2}h|v_h|_{1,h}$ for all $v_h \in \operatorname{Span}\mathfrak{A}$.*

Proof. Remind that $(\psi_j^\iota)_\mu$ is the alternating function such that the support is Ω_j^ι and the nonzero DoF values are only lying on faces perpendicular to μ-axis. Thus only μ-component of the piecewise gradient of $(\psi_j^\iota)_\mu$ survives. It implies that $a_h((\psi_j^\iota)_\mu, (\psi_k^\lambda)_\nu) = 0$ if $\mu \neq \nu$. Therefore we can consider $\mathcal{S}_h^{\mathfrak{A}}$ as a block diagonal matrix:

$$\mathcal{S}_h^{\mathfrak{A}} = \begin{bmatrix} \mathcal{S}_h^{\mathfrak{A}_x} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathcal{S}_h^{\mathfrak{A}_y} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathcal{S}_h^{\mathfrak{A}_z} \end{bmatrix},$$

where $\mathfrak{A}_x$, $\mathfrak{A}_y$, $\mathfrak{A}_z$ are defined as in Theorem 6.3.1, and $\mathcal{S}_h^{\mathfrak{A}_x}$, $\mathcal{S}_h^{\mathfrak{A}_y}$, $\mathcal{S}_h^{\mathfrak{A}_z}$ are the stiffness matrices associated with the respective sets.

We can also consider $\mathcal{M}_h^{\mathfrak{A}}$ as a block diagonal matrix in the same structure, since the following observation: if $\mu \neq \lambda$, then

$$\begin{aligned}\left((\psi_j^\iota)_\mu, (\psi_k^\lambda)_\nu\right)_\Omega &= \int_\Omega (\psi_j^\iota)_\mu \, (\psi_k^\lambda)_\nu \, \mathrm{d}\mathbf{x} = \sum_{Q \in \mathcal{T}_h(\Omega)} \int_Q (\psi_j^\iota)_\mu \, (\psi_k^\lambda)_\nu \, \mathrm{d}\mathbf{x} \\ &= \sum_{Q \in \mathcal{T}_h(\Omega)} h \int_{Q_\mu} (\psi_j^\iota)_\mu \, \mathrm{d}\mu \int_{Q_\nu} (\psi_k^\lambda)_\nu \, \mathrm{d}\nu = 0.\end{aligned}$$

We write

$$\mathcal{M}_h^{\mathfrak{A}} = \begin{bmatrix} \mathcal{M}_h^{\mathfrak{A}_x} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathcal{M}_h^{\mathfrak{A}_y} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathcal{M}_h^{\mathfrak{A}_z} \end{bmatrix},$$

where $\mathcal{M}_h^{\mathfrak{A}_x}$, $\mathcal{M}_h^{\mathfrak{A}_y}$, $\mathcal{M}_h^{\mathfrak{A}_z}$ are the mass matrices associated with the respective sets. Therefore, it is enough to show $\mathcal{M}_h^{\mathfrak{A}_\mu} = Ch^2 \mathcal{S}_h^{\mathfrak{A}_\mu}$ for each $\mu \in \{x, y, z\}$.

First, we consider the blocks associated with $\mathfrak{A}_x = \{(\psi_j^y)_x, (\psi_k^z)_x\}$ for $1 \leq j \leq N_y, 1 \leq k \leq N_z$. The proof for other blocks is similar. For any two alternating functions $(\psi_j^\iota)_x$ and $(\psi_k^\lambda)_x$ in $\mathfrak{A}_x$,

1. if $\iota = \lambda$ (let them be equal to y, without loss of generality), then

$$\begin{aligned}a_h\left((\psi_j^y)_x, (\psi_k^y)_x\right) &= \sum_{Q \in \mathcal{T}_h(\Omega)} \int_Q \nabla(\psi_j^y)_x \cdot \nabla(\psi_k^y)_x \, \mathrm{d}\mathbf{x} \\ &= \sum_{Q \in \mathcal{T}_h(\Omega_j^y \cap \Omega_k^y)} \int_Q (2/h)^2 \, \mathrm{d}\mathbf{x} = 4N_x N_z h \delta_{jk},\end{aligned}$$

since the number of cubes in Ω_j^y is $N_x N_z$. Here, δ_{jk} denotes the Kronecker delta.

2. if $\iota \neq \lambda$ (let $\iota = y$ and $\lambda = z$, without loss of generality), then

$$\begin{aligned} a_h\left((\psi_j^y)_x, (\psi_k^z)_x\right) &= \sum_{Q\in\mathcal{T}_h(\Omega)} \int_Q \nabla(\psi_j^y)_x \cdot \nabla(\psi_k^z)_x \, \mathrm{d}\mathbf{x} \\ &= \sum_{Q\in\mathcal{T}_h(\Omega_j^y\cap\Omega_k^z)} \int_Q (2/h)^2 \, \mathrm{d}\mathbf{x} = 4N_x h, \end{aligned}$$

since the number of cubes in $\Omega_j^y \cap \Omega_k^z$ is N_x.

On the other hand, we can easily observe that

$$\begin{aligned} \left((\psi_j^y)_x, (\psi_k^y)_x\right)_\Omega &= \sum_{Q\in\mathcal{T}_h(\Omega)} \int_Q (\psi_j^y)_x \, (\psi_k^y)_x \, \mathrm{d}\mathbf{x} \\ &= \sum_{Q\in\mathcal{T}_h(\Omega_j^y\cap\Omega_k^y)} \int_Q (\psi_j^y)_x \, (\psi_k^y)_x \, \mathrm{d}\mathbf{x} = \frac{1}{3} N_x N_z h^3 \delta_{jk}, \text{ and} \end{aligned}$$

$$\begin{aligned} \left((\psi_j^y)_x, (\psi_k^z)_x\right)_\Omega &= \sum_{Q\in\mathcal{T}_h(\Omega)} \int_Q (\psi_j^y)_x \, (\psi_k^z)_x \, \mathrm{d}\mathbf{x} \\ &= \sum_{Q\in\mathcal{T}_h(\Omega_j^y\cap\Omega_k^z)} \int_Q (\psi_j^y)_x \, (\psi_k^z)_x \, \mathrm{d}\mathbf{x} = \frac{1}{3} N_x h^3. \end{aligned}$$

Therefore $\mathcal{M}_h^{\mathfrak{A}_x} = \frac{1}{12} h^2 \mathcal{S}_h^{\mathfrak{A}_x}$, and the proof is completed. □

Theorem 6.5.5 (Relation between numerical solutions in 3-D)**.** *Let* u_h, $u_h^\flat$, $u_h^\natural$, $\bar{u}_h^\natural$ *be the numerical solutions of* (4.6) *in 3-D as* (4.12), (4.16), (4.17), (4.21)*, respectively, with* $\mathfrak{E}^\flat = \mathfrak{B}^\flat \cup \mathfrak{A}^\flat$*. Then* $u_h = u_h^\flat = u_h^\natural$*, and*

$$\|u_h^\natural - \bar{u}_h^\natural\|_0 \leq Ch^2\|f\|_0, \ |u_h^\natural - \bar{u}_h^\natural|_{1,h} \leq Ch\|f\|_0.$$

6.6 Numerical results

As mentioned before, we can not construct a basis $\mathfrak{B}^\flat$ for $V_{per}^{\mathfrak{B},h}$ in 3-D explicitly. We only use the scheme option 4 for our numerical test. The exact solution is $u(x,y,z) = \sin(2\pi x)\sin(2\pi y)\sin(2\pi z)$. The numerical results on Table 6.3 confirm our theoretical results.

	Opt 4			
h	$\lvert u-u_h\rvert_{1,h}$	order	$\lVert u-u_h\rVert_0$	order
1/8	1.505E-00	-	3.848E-02	-
1/16	7.550E-01	0.995	9.716E-03	1.986
1/32	3.777E-01	0.999	2.434E-03	1.997
1/64	1.889E-01	1.000	6.089E-04	1.999
1/128	9.443E-02	1.000	1.523E-04	2.000

Table 6.3. Numerical result of the elliptic problem in 3-D with the scheme option 4

Part II

Nonconforming Heterogeneous Multiscale Method

Chapter 1

Introduction

Finite element method (FEM) is one of successful methods to approximate the solution of partial differential equations derived in various fields of studies. However it has a drawback when we treat a problem containing heterogeneity. For instance, when the coefficient tensor of the problem is highly oscillatory in micro scale, we need to consider a sufficiently refined mesh consisting of elements which are comparable with the micro scale in order to get a numerical solution sufficiently close to the exact solution. Such a refinement increases the number of unknowns in the corresponding system of equations. It is, of course, a critical burden on solving the equation numerically.

To overcome this shortage of the standard FEM, several efficient methods have been proposed and developed in decades. Multiscale finite element method (MsFEM) [30, 26, 25] employs basis functions representing multiscale features whereas a local shape function in the standard FEM is just a plain polynomial. In each macro element, the multiscale shape function is con-

structed by solving a discrete harmonic equation associated with given multiscale coefficient. Generalized multiscale finite element method (GMsFEM) reduces the number of degrees of freedom in the discrete model by considering a few dominant modes of the corresponding generalized eigenvalue problem [23, 24]. Heterogeneous multiscale method (HMM) numerically estimates the homogenized coefficient using the micro scale structure. Especially, the finite element heterogeneous multiscale method (FEHMM) is a HMM framework which is based on finite element implementation [1, 2, 6, 3, 4, 21].

Most of the above works employ the conforming finite element approach, while nonconforming elements have prominence for their numerical stability in various problems [18, 14, 35, 46, 13, 20, 11, 38]. Recently, there are some works in MsFEM based on the nonconforming approach [36, 37, 19]. Lee and Sheen [39] proposed a nonconforming GMsFEM framework for elliptic problems.

In this thesis, we propose a FEHMM scheme based on nonconforming finite elements for multiscale elliptic problems. As a prototype of nonconforming elements, we employ the P_1–nonconforming quadrilateral finite element, which is the lowest-order element on quadrilateral or rectangular mesh (in 2-D case), and hexahedral mesh (in 3-D case). Thus this finite element shares the same nature of the well-known Crouzeix-Raviart element on triangular mesh. We would like to emphasize the advantage of rectangular elements over simplicial elements on mesh construction, especially in 3-D space. Each micro problem in the proposed FEHMM scheme may derive a singular linear equation due to its periodic nature. By using results from recent analysis for P_1–nonconforming quadrilateral finite element with periodic boundary condition, we formulate the singular linear equation firmly, and solve it efficiently.

This thesis is organized as follows. In chapter 2, we state in brief preliminaries and notations for our discussion. We introduce a nonconforming

FEHMM scheme in chapter 3. Chapter 4 is devoted to prove main theorem for a priori error estimates of the proposed method. We give several numerical results in chapter 5.

Chapter 2

Preliminaries

Let $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ be a bounded domain with smooth boundary $\partial\Omega$. Denote $\mathbf{x} = (x_1, \cdots, x_d) \in \mathbb{R}^d$. Consider a multiscale elliptic problem

$$-\nabla \cdot \Big(\mathbf{A}^\varepsilon(\mathbf{x}) \nabla u^\varepsilon(\mathbf{x}) \Big) = f(\mathbf{x}) \quad \text{in } \Omega, \tag{2.1a}$$

$$u^\varepsilon = 0 \qquad \text{on } \partial\Omega, \tag{2.1b}$$

where $\varepsilon \ll 1$ is a scale parameter. Here, the coefficient tensor $\mathbf{A}^\varepsilon \in [L^\infty(\Omega)]^{d\times d}$ is assumed to be symmetric, uniformly elliptic and bounded, *i.e.*, there exist $\lambda, \Lambda > 0$ which do not depend on $\mathbf{x}$ such that $\lambda|\xi|^2 \leq \mathbf{A}^\varepsilon(\mathbf{x})\ \xi \cdot \xi \leq \Lambda|\xi|^2$ for all $\xi \in \mathbb{R}^d$.

2.1 Homogenization

Let $Y = \prod_{k=1}^d [0, \ell_k]$ for given $\{\ell_k\}_{k=1}^d$ and $\mathbf{e}_j$ be the standard unit basis of $\mathbb{R}^d$ corresponding to the j-th component. Suppose that $\mathbf{A}^\varepsilon(\mathbf{x}) := \mathbf{A}(\mathbf{x}, \mathbf{x}/\varepsilon)$ for a

Y-periodic function $\mathbf{A}(\cdot,\cdot)$ with respect to the second variable, *i.e.*, a function $\mathbf{A}:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}^{d\times d}$ satisfies that $\mathbf{A}(\mathbf{x},\mathbf{y})=\mathbf{A}(\mathbf{x},\mathbf{y}+\ell_k\mathbf{e}_k)$ for $1\le k\le d$. Then, the following result is well known [17, 32].

Theorem 2.1.1 (Periodic case)**.** *Suppose that* $\mathbf{A}^\varepsilon(\mathbf{x}):=\mathbf{A}(\mathbf{x},\mathbf{x}/\varepsilon)$ *where* $\mathbf{A}(\mathbf{x},\mathbf{y})$ *is* Y*-periodic for the variable* $\mathbf{y}=(y_1,\cdots,y_d)$*. Let* $f\in L^2(\Omega)$*. Then there exists a homogenized coefficient tensor* $\mathbf{A}^0$ *such that*

$$\begin{cases} u^\varepsilon \rightharpoonup u^0 \ \textit{weakly in}\ H_0^1(\Omega),\\ \mathbf{A}^\varepsilon\nabla u^\varepsilon \rightharpoonup \mathbf{A}^0\nabla u^0 \ \textit{weakly in}\ [L^2(\Omega)]^d, \end{cases}$$

where u^0 *is a unique solution in* $H_0^1(\Omega)$ *of the homogenized problem:*

$$\begin{cases} -\nabla\cdot\Big(\mathbf{A}^0(\mathbf{x})\nabla u^0(\mathbf{x})\Big)=f(\mathbf{x}) \quad \textit{in}\ \Omega,\\ \qquad\qquad\qquad u^0=0 \qquad \textit{on}\ \partial\Omega. \end{cases} \tag{2.2}$$

In fact, the homogenized coefficient $\mathbf{A}^0=(A^0_{ij})$ *is given by*

$$A^0_{ij}(\mathbf{x})=\frac{1}{|Y|}\int_Y\left(A_{ij}(\mathbf{x},\mathbf{y})+\sum_{k=1}^d A_{ik}(\mathbf{x},\mathbf{y})\frac{\partial\chi^j}{\partial y_k}\right)\,\mathrm{d}\mathbf{y},$$

where $|Y|$ *denotes the volume of* Y*, and* $\chi^j=\chi^j(\mathbf{x},\mathbf{y})$ *the solution of the cell problem:*

$$\begin{cases} -\nabla_{\mathbf{y}}\cdot\big(\mathbf{A}(\mathbf{x},\mathbf{y})\nabla_{\mathbf{y}}\chi^j\big)=\nabla_{\mathbf{y}}\cdot\left(\mathbf{A}(\mathbf{x},\mathbf{y})\ \mathbf{e}_j\right) \quad \textit{in}\ Y,\\ \chi^j \ \textit{is}\ Y\textit{-periodic},\\ \displaystyle\int_Y\chi^j\,\mathrm{d}\mathbf{y}=0. \end{cases}$$

2.2 Notations

Let D be a bounded open domain in $\mathbb{R}^d$ $(d = 2, 3)$. Denote by $L^2(D)$, $H^1(D)$, and $H^1_0(D)$ the standard Sobolev spaces on D with the standard Sobolev norms $\|\cdot\|_{0,D}$, $\|\cdot\|_{1,D}$, and (semi-)norm $|\cdot|_{1,D}$, respectively. By $C^\infty_{per}(D)$ designate the set of smooth periodic functions on D and by $H^1_{per}(D)$ the closure of $C^\infty_{per}(D)$ with respect to the norm $\|\cdot\|_{1,D}$ in $H^1(D)$. $W^1_{per}(D)$ is a subspace of $H^1_{per}(D)$ which consists of functions whose mean value on D are zero. We will mean by $(\cdot,\cdot)_D$ the $L^2(D)$ inner product. In the case of $D = \Omega$, the subscript D on notations of norms and inner product is omitted. For $(d-1)$-dimensional face f, $\langle\cdot,\cdot\rangle_f$ indicates the $L^2(f)$ inner product.

By $|D|$ we denote the volume of the domain D. For an integrable function $v \in L^1(D)$, the mean value on D is denoted by $\mathcal{M}_D(v) := \frac{1}{|D|}\int_D v$. Throughout this thesis C denotes a generic constant and its value varies depending on the position where it appears.

Consider a family of triangulations $\{\mathcal{T}_h(D)\}_{0<h<1}$ for the domain D consisting of quadrilateral elements. Let $\mathcal{E}^i_h, \mathcal{E}^b_h$, and $\mathcal{E}^{b,opp}_h$ denote the sets of all interior edges, of all boundary edges, and of all pairs consisting of two boundary edges on opposite position, respectively. Set

$$V^{P1}_h(D) = \left\{ v \in L^2(D) \,\middle|\, v|_K \in \mathcal{P}_1(K)\ \ \forall K \in \mathcal{T}_h(D),\ \langle [v]_e, 1\rangle_e = 0\ \ \forall e \in \mathcal{E}^i_h \right\}, \tag{2.3}$$

$$V^{P1}_{h,0}(D) = \left\{ v \in V^{P1}_h(D) \,\middle|\, \langle v, 1\rangle_e = 0 \quad \forall e \in \mathcal{E}^b_h \right\}, \tag{2.4}$$

$$V^{P1}_{h,per}(D) = \left\{ v \in V^{P1}_h(D) \,\middle|\, \langle v, 1\rangle_{e_1} = \langle v, 1\rangle_{e_2}\ \forall (e_1, e_2) \in \mathcal{E}^{b,opp}_h,\ (v, 1)_D = 0 \right\}, \tag{2.5}$$

where $\mathcal{P}_1(K)$ denotes the set of linear polynomials on K, and $[\cdot]_e$ the jump

across edge e. Let $|\cdot|_{1,h,D}$ denote mesh-dependent energy norm on $V_h^{P1}(D)$. The standard error analysis for nonconforming elements implies a priori error estimate, see [44, 20],

$$|u - u_h|_{1,h} \leq Ch\|u\|_2.$$

Chapter 3

FEHMM Based on Nonconforming Spaces

In this chapter we introduce a FEHMM scheme based on nonconforming finite spaces for the multiscale elliptic problem (2.1). We follow the framework of FEHMM [1, 2] with slight modification for nonconforming function spaces. Here and in what follows, we only treat the case of $d = 2$. Let $\mathcal{T}_H := \mathcal{T}_H(\Omega)$ be a regular triangulation of Ω with quadrilaterals. Define the macro mesh parameter $H := \max_{K\in\mathcal{T}_H} \operatorname{diam}(K)$. For each macro element $K_H \in \mathcal{T}_H$, let $\mathcal{E}(K_H)$ denote the set of its edges. The set of all edges, of all interior edges and of all boundary edges are denoted by $\mathcal{E}_H$, $\mathcal{E}_H^i$ and $\mathcal{E}_H^b$, respectively. Let $F_{K_H} : \widehat{K} \to K_H$ be a bilinear transformation from the reference domain onto K_H. Set

$$V = H_0^1(\Omega) \quad \text{and} \quad V_H = V_{H,0}^{P1}(\Omega) \tag{3.1}$$

and denote the macro mesh-dependent (semi-)norm on $V + V_H$ by $|||\cdot|||_H := \left(\sum_{K_H \in \mathcal{T}_H} |\cdot|^2_{1,K_H}\right)^{1/2}$.

To formulate the FEHMM scheme, we need a quadrature formula which consists of I points with corresponding weights $(\mathbf{x}_i, \omega_i)_{i=1}^I$ on each element $K_H \in \mathcal{T}_H$ such that

$$\sum_{i=1}^{I} \omega_i |\nabla v(\mathbf{x}_i)|^2 \geq C|v|^2_{1,K_H} \qquad \forall v \in \mathcal{P}_1(K_H),$$
$$\sum_{i=1}^{I} \omega_i \nabla v(\mathbf{x}_i) \cdot \nabla w(\mathbf{x}_i) = \int_{K_H} \nabla v \cdot \nabla w \, \mathrm{d}\mathbf{x} \qquad \forall v, w \in \mathcal{P}_1(K_H).$$

Remark 3.0.1. *The above characteristics of the quadrature formula are useful to prove the existence and uniqueness of the solution as well as optimal error estimates in Chapter 4.*

On each element $K_H \in \mathcal{T}_H$ we define I sampling domains $K_{\delta,i} := \mathbf{x}_i + [-\delta/2, \delta/2]^2$ corresponding to each quadrature point $\mathbf{x}_i$ for given $\delta \ll 1$. The size of the sampling domains δ should be chosen to be comparable with ε. The most trivial case is $\delta = \varepsilon$, but not always. The effect of various δ will be mentioned in Section 4.5.2. On each sampling domain we consider a micro triangulation to deal a bundle of micro problems on it. Let $\mathcal{T}_h(K_{\delta,i})$ be a uniform triangulation of a sampling domain $K_{\delta,i}$ consisting of quadrilateral elements and $h := \max_{K \in \mathcal{T}_h(K_{\delta,i})} \mathrm{diam}(K)$ the micro mesh parameter. Each micro element $K_h \in \mathcal{T}_h(K_{\delta,i})$ has a bilinear transformation $F_{K_h} : \widehat{K} \to K_h$ such that $F_{K_h}(\widehat{K}) = K_h$. Let denote the set of all edges, of all interior edges and of all boundary edges in $\mathcal{T}_h$ by $\mathcal{E}_h$, $\mathcal{E}_h^i$ and $\mathcal{E}_h^b$, respectively. $\mathcal{E}(K_h)$ denotes the set of edges of K_h.

On each sampling domain $K_{\delta,i}$ we will consider two micro function spaces, namely, a continuous function space $W(K_{\delta,i})$ and a discrete space $W_h(K_{\delta,i})$

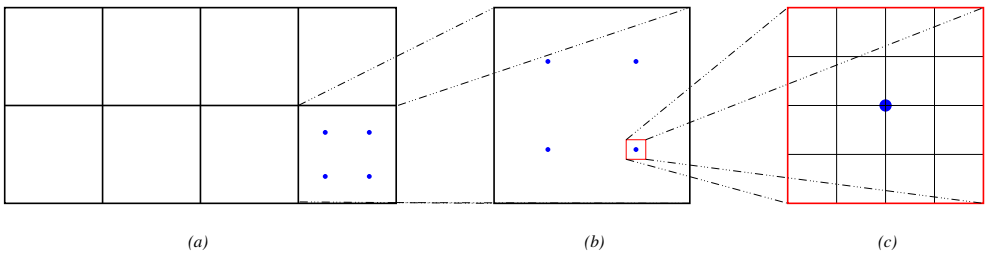

Figure 3.1. The hierarchy of geometric objects in FEHMM scheme

which are determined by a choice of macro-micro coupling condition we use. If the coefficient tensor $\mathbf{A}^\varepsilon$ in (2.1) has a periodic property, then we can impose periodic coupling condition. On the other hand, Dirichlet coupling condition can be used for general cases. Respective to the choice we define two micro function spaces by

$$W(K_{\delta,i}) = \begin{cases} W^1_{per}(K_{\delta,i}), & \text{periodic case,} \\ H^1_0(K_{\delta,i}), & \text{Dirichlet BC case,} \end{cases} \tag{3.3a}$$

$$W_h(K_{\delta,i}) = \begin{cases} V^{P1}_{h,per}(K_{\delta,i}) & \text{periodic case,} \\ V^{P1}_{h,0}(K_{\delta,i}) & \text{Dirichlet BC case.} \end{cases} \tag{3.3b}$$

The micro mesh-dependent (semi-)norm on $W(K_{\delta,i}) + W_h(K_{\delta,i})$ is defined by $|||\cdot|||_{h,K_{\delta,i}} := \left(\sum_{K_h \in \mathcal{T}_h(K_{\delta,i})} |\cdot|^2_{1,K_h}\right)^{1/2}$ in both periodic and Dirichlet BC coupling cases. The expression $K_{\delta,i}$ in notations will be omitted if there is no ambiguity of choice for sampling domains.

Figure 3.1 shows the hierarchy of geometric objects in the FEHMM scheme at a glance: (a) domain Ω and its triangulation $\mathcal{T}_H$, (b) macro element K_H, (c) sampling domain $K_{\delta,i}$ surrounding a quadrature point $\mathbf{x}_i$ and its triangulation $\mathcal{T}_h$ consisting of micro elements K_h.

For the sake of convenience, introduce the two bilinear forms, $a^{K_{\delta,i}}$:

$H^1(K_{\delta,i}) \times H^1(K_{\delta,i}) \to \mathbb{R}$ and $a_h^{K_{\delta,i}} : V_h^{P1}(K_{\delta,i}) \times V_h^{P1}(K_{\delta,i}) \to \mathbb{R}$ by

$$a^{K_{\delta,i}}(u,v) = \int_{K_{\delta,i}} \mathbf{A}^\varepsilon \nabla u \cdot \nabla v \, \mathrm{d}\mathbf{x} \quad \forall u, v \in H^1(K_{\delta,i}),$$
$$a_h^{K_{\delta,i}}(u_h,v_h) = \sum_{K_h \in \mathcal{T}_h(K_{\delta,i})} \int_{K_h} \mathbf{A}^\varepsilon \nabla u_h \cdot \nabla v_h \, \mathrm{d}\mathbf{x} \quad \forall u_h, v_h \in V_h^{P1}(K_{\delta,i}).$$

Also define two bilinear forms $\overline{a}_H$ and $a_H : V_H \times V_H \to \mathbb{R}$ as follows: for all $u_H, v_H \in V_H$,

$$\overline{a}_H(u_H,v_H) = \sum_{K_H \in \mathcal{T}_H} \sum_{i=1}^{I} \frac{\omega_i}{|K_{\delta,i}|} \int_{K_{\delta,i}} \mathbf{A}^\varepsilon \nabla u^m \cdot \nabla v^m \, \mathrm{d}\mathbf{x}, \tag{3.4a}$$
$$a_H(u_H,v_H) = \sum_{K_H \in \mathcal{T}_H} \sum_{i=1}^{I} \frac{\omega_i}{|K_{\delta,i}|} \sum_{K_h \in \mathcal{T}_h(K_{\delta,i})} \int_{K_h} \mathbf{A}^\varepsilon \nabla u_h^m \cdot \nabla v_h^m \, \mathrm{d}\mathbf{x} \tag{3.4b}$$

where u^m, v^m, u_h^m, v_h^m are the solutions of the continuous and discrete micro problems with constraints u_H and v_H, respectively, on each sampling domain $K_{\delta,i}$ in $K_H \in \mathcal{T}_H$ defined as follows: for given $w_H \in V_H$, $w^m \in w_H + W(K_{\delta,i})$ and $w_h^m \in w_H + W_h(K_{\delta,i})$ fulfill

$$a^{K_{\delta,i}}(w^m, z) = 0 \quad \forall z \in W(K_{\delta,i}), \tag{3.5a}$$
$$a_h^{K_{\delta,i}}(w_h^m, z_h) = 0 \quad \forall z_h \in W_h(K_{\delta,i}). \tag{3.5b}$$

In the above expressions $w_H + W(K_{\delta,i})$ and $w_H + W_h(K_{\delta,i})$, w_H actually means $w_H|_{K_{\delta,i}}$, the function restricted onto the domain $K_{\delta,i}$. However, here and in what follows, we use this abusive notation for the sake of simple expressions if context determines proper range of given function.

Remark 3.0.2. *By following a typical FEHMM framework, one needs to consider w_H^{lin}, a linearization of w_H at $\mathbf{x}_i$, instead of w_H itself in order to get w^m and w_h^m in* (3.5)*. In our discussion, however, such a linearization is unnec-*

essary because the finite element space which we are considering consists of piecewise linear functions.

A nonconforming FEHMM weak formulation of the problem (2.1) is now ready to be stated as follows:

(Main Weak Formulation) *find $u_H \in V_H$ such that*

$$a_H(u_H, w_H) = (f, w_H) \quad \forall w_H \in V_H. \tag{3.6}$$

For analysis in Chapter 4, we introduce several micro functions. Let $\psi_h^j = \psi_h^j(\mathbf{x}) \in W_h(K_{\delta,i}), j = 1, \cdots, d$, the solution of the following micro problem

$$a_h^{K_{\delta,i}}(\psi_h^j, z_h) = -\sum_{K_h \in \mathcal{T}_h(K_{\delta,i})} \int_{K_h} \mathbf{A}^\varepsilon \mathbf{e}_j \cdot \nabla z_h \, \mathrm{d}\mathbf{x} \quad \forall z_h \in W_h(K_{\delta,i}). \tag{3.7}$$

Also, for $j = 1, \cdots, d$, let $\psi^j = \psi^j(\mathbf{x}) \in W(K_{\delta,i})$ be the solution of

$$a^{K_{\delta,i}}(\psi^j, z) = -\int_{K_{\delta,i}} \mathbf{A}^\varepsilon \mathbf{e}_j \cdot \nabla z \, \mathrm{d}\mathbf{x} \quad \forall z \in W(K_{\delta,i}). \tag{3.8}$$

Later, $\psi_h^j, j = 1, \cdots, d$, play as basis functions for the solution space of the micro problem (3.5). We will also use the following functions denoted by $\varphi_h^j(\mathbf{x}) := \psi_h^j(\mathbf{x}) + x_j$ and $\varphi^j(\mathbf{x}) := \psi^j(\mathbf{x}) + x_j$ on each sampling domain $K_{\delta,i}$.

Remark 3.0.3. *Indeed, ψ^j and ψ_h^j are nothing but $\psi^j = (x_j)^m - x_j$ and $\psi_h^j = (x_j)_h^m - x_j$, respectively, with the superscript 'm' as in* (3.5)*. Moreover $\varphi^j = (x_j)^m$ and $\varphi_h^j = (x_j)_h^m$.*

We also introduce several weak formulations which are used for analysis.

(Weak Formulation of Homogenized Problem) A weak formulation of the homogenized problem (2.2) is given as to find $u^0 \in V$ such that

$$a^0(u^0, v) = (f, v) \quad \forall v \in V,$$

where

$$a^0(v, w) = \int_\Omega \mathbf{A}^0(\mathbf{x}) \nabla v \cdot \nabla w \, \mathrm{d}\mathbf{x} \qquad \forall v, w \in V. \tag{3.9}$$

(Weak Formulation with Quadrature Rule in Macro Scale) A weak formulation of the homogenized problem (2.2) with quadrature rule in macro scale, corresponding to (3.6), can be defined as to find $u_H^0 \in V_H$ fulfilling

$$a_H^0(u_H^0, v_H) = (f, v_H) \quad \forall v_H \in V_H,$$

where

$$a_H^0(v_H, w_H) = \sum_{K_H \in \mathcal{T}_H} \sum_{i=1}^{I} \omega_i \; \mathbf{A}^0(\mathbf{x}_i) \nabla v_H(\mathbf{x}_i) \cdot \nabla w_H(\mathbf{x}_i) \qquad \forall v_H, w_H \in V_H. \tag{3.10}$$

(Semi-discrete FEHMM) A semi-discrete FEHMM solution is defined as $\overline{u}_H \in V_H$ such that

$$\overline{a}_H(\overline{u}_H, v_H) = (f, v_H) \quad \forall v_H \in V_H. \tag{3.11}$$

Chapter 4

Fundamental Properties of Nonconforming HMM

4.1 Existence and uniqueness

For the beginning of analysis, we prove the existence and uniqueness of solution of the equation.

Lemma 4.1.1. *Let v_h^m be the solution of the micro problem* (3.5b) *with constraint v_H on a sampling domain $K_{\delta,i}$. Then*

$$|v_H|_{1,K_{\delta,i}} \leq |||v_h^m|||_{h,K_{\delta,i}} \leq \frac{\Lambda}{\lambda}|v_H|_{1,K_{\delta,i}}.$$

Proof. Utilizing the fact that v_H is linear on K for all $K \in \mathcal{T}_H$ and (3.3b), we have

$$0 \leq \sum_{K_h \in \mathcal{T}_h} \int_{K_h} \nabla(v_h^m - v_H) \cdot \nabla(v_h^m - v_H) \, \mathrm{d}\mathbf{x}$$

$$
\begin{aligned}
&= \sum_{K_h \in \mathcal{T}_h} \int_{K_h} |\nabla v_h^m|^2 - |\nabla v_H|^2 - 2\nabla(v_h^m - v_H) \cdot \nabla v_H \, \mathrm{d}\mathbf{x} \\
&= \left[\sum_{K_h \in \mathcal{T}_h} \int_{K_h} |\nabla v_h^m|^2 \, \mathrm{d}\mathbf{x} - \int_{K_{\delta,i}} |\nabla v_H|^2 \, \mathrm{d}\mathbf{x} \right] \\
&\qquad\qquad - 2\nabla v_H \cdot \sum_{K_h \in \mathcal{T}_h} \int_{\partial K_h} \mathbf{n}_{K_h} (v_h^m - v_H) \, \mathrm{ds}
\end{aligned}
$$

where $\mathbf{n}_{K_h}$ denotes the unit outward normal to K_h. Since $v_h^m - v_H \in W_h(K_{\delta,i})$, the last term in the above equation vanishes. Consequently, we get

$$
\int_{K_{\delta,i}} |\nabla v_H|^2 \, \mathrm{d}\mathbf{x} \leq \sum_{K_h \in \mathcal{T}_h} \int_{K_h} |\nabla v_h^m|^2 \, \mathrm{d}\mathbf{x} \, .
$$

On the other hand, due to the ellipticity of $\mathbf{A}^\varepsilon$, we have

$$
\begin{aligned}
0 &\leq \sum_{K_h \in \mathcal{T}_h} \int_{K_h} \mathbf{A}^\varepsilon \nabla(v_h^m - v_H) \cdot \nabla(v_h^m - v_H) \, \mathrm{d}\mathbf{x} \\
&= \sum_{K_h \in \mathcal{T}_h} \int_{K_h} \mathbf{A}^\varepsilon \nabla v_H \cdot \nabla v_H - \mathbf{A}^\varepsilon \nabla v_h^m \cdot \nabla v_h^m \\
&\qquad\qquad + \mathbf{A}^\varepsilon \nabla(v_h^m - v_H) \cdot \nabla v_h^m + \mathbf{A}^\varepsilon \nabla v_h^m \cdot \nabla(v_h^m - v_H) \, \mathrm{d}\mathbf{x} \, .
\end{aligned}
$$

Due to the definition of v_h^m in (3.5) and symmetry of $\mathbf{A}^\varepsilon$, the last two terms vanish. Thus we get

$$
\sum_{K_h \in \mathcal{T}_h} \int_{K_h} \mathbf{A}^\varepsilon \nabla v_h^m \cdot \nabla v_h^m \, \mathrm{d}\mathbf{x} \leq \int_{K_{\delta,i}} \mathbf{A}^\varepsilon \nabla v_H \cdot \nabla v_H \, \mathrm{d}\mathbf{x} \, .
$$

The uniform ellipticity and boundedness of $\mathbf{A}^\varepsilon$ imply the desired inequality.

□

Due to the properties of the quadrature formula, the bilinear form a_H is bounded and coercive in V_H. Therefore the existence and uniqueness of the

solution u_H to (3.6) is guaranteed by the Lax-Milgram Lemma. Thus, we have

Theorem 4.1.2. *There exists a unique solution u_H to the problem* (3.6).

Similarly, the coercivity and boundedness of the bilinear form $\overline{a}_H$ can be obtained immediately, and thus one also get the existence and uniqueness of the solution $\overline{u}_H$ as stated below.

Theorem 4.1.3. *There exists a unique solution $\overline{u}_H$ to the problem* (3.11).

4.2 Recovered homogenized tensors

Recovered homogenized tensors $\mathbf{A}^0_{K_{\delta,i}}$ and $\overline{\mathbf{A}}^0_{K_{\delta,i}}$ on a sampling domain $K_{\delta,i}$ are defined by

$$\mathbf{A}^0_{K_{\delta,i}} = \frac{1}{|K_{\delta,i}|} \sum_{K_h \in \mathcal{T}_h(K_{\delta,i})} \int_{K_h} \mathbf{A}^\varepsilon(\mathbf{x}) \left(I + \mathbf{J}^T_{\psi_h}\right) \, \mathrm{d}\mathbf{x}, \tag{4.1a}$$

$$\overline{\mathbf{A}}^0_{K_{\delta,i}} = \frac{1}{|K_{\delta,i}|} \int_{K_{\delta,i}} \mathbf{A}^\varepsilon(\mathbf{x}) \left(I + \mathbf{J}^T_{\psi}\right) \, \mathrm{d}\mathbf{x}, \tag{4.1b}$$

where $\mathbf{J}_{\psi_h}$ and $\mathbf{J}_{\psi}$ are $d \times d$ matrices defined by

$$[J_{\psi_h}]_{jk} = \frac{\partial \psi^j_h}{\partial x_k} \quad \text{and} \quad [J_{\psi}]_{jk} = \frac{\partial \psi^j}{\partial x_k}, \quad 1 \le j,k \le d,$$

respectively. The following proposition shows the essential characteristic of two recovered homogenized tensors.

Proposition 4.2.1. *Let u^m_h and v^m_h be the solutions of the discrete micro problem* (3.5b) *corresponding to the macro constraints u_H and v_H, respectively, on $K_{\delta,i}$. Then the following holds.*

$$\frac{1}{|K_{\delta,i}|} \sum_{K_h \in \mathcal{T}_h(K_{\delta,i})} \int_{K_h} \mathbf{A}^\varepsilon \nabla u^m_h \cdot \nabla v^m_h \, \mathrm{d}\mathbf{x} = \mathbf{A}^0_{K_{\delta,i}} \nabla u_H \cdot \nabla v_H. \tag{4.2}$$

Similarly, let u^m and v^m be the solutions of the continuous micro problem (3.5a) *corresponding to the macro constraints u_H and v_H, respectively, on $K_{\delta,i}$. Then the following also holds.*

$$\frac{1}{|K_{\delta,i}|}\int_{K_{\delta,i}} \mathbf{A}^\varepsilon \nabla u^m \cdot \nabla v^m \,\mathrm{d}\mathbf{x} = \overline{\mathbf{A}}^0_{K_{\delta,i}} \nabla u_H \cdot \nabla v_H. \tag{4.3}$$

Proof. We will show (4.2) only, since (4.3) follows immediately by a similar argument. Since u_h^m is the solution of (3.5b), and $v_h^m - v_H \in W_h(K_{\delta,i})$, it holds

$$\frac{1}{|K_{\delta,i}|}\sum_{K_h\in\mathcal{T}_h}\int_{K_h} \mathbf{A}^\varepsilon \nabla u_h^m \cdot \nabla v_h^m \,\mathrm{d}\mathbf{x} = \frac{1}{|K_{\delta,i}|}\sum_{K_h\in\mathcal{T}_h}\int_{K_h} \mathbf{A}^\varepsilon \nabla u_h^m \cdot \nabla v_H \,\mathrm{d}\mathbf{x}\,. \tag{4.4}$$

Since ∇u_H is constant, u_h^m is represented by a linear combination of the basis functions ψ_h^j as

$$u_h^m = u_H + \sum_{j=1}^{d} \psi_h^j \,\frac{\partial u_H}{\partial x_j}.$$

By plugging the above representation into (4.4), we have

$$\begin{aligned}
&\frac{1}{|K_{\delta,i}|}\sum_{K_h\in\mathcal{T}_h}\int_{K_h} \mathbf{A}^\varepsilon \left(\nabla u_H + \sum_{j=1}^{d} \nabla\psi_h^j \,\frac{\partial u_H}{\partial x_j}\right)\cdot \nabla v_H \,\mathrm{d}\mathbf{x} \\
=&\frac{1}{|K_{\delta,i}|}\sum_{K_h\in\mathcal{T}_h}\int_{K_h} \mathbf{A}^\varepsilon \left(I + J_{\psi_h}^T\right)\nabla u_H \cdot \nabla v_H \,\mathrm{d}\mathbf{x} \\
=&\mathbf{A}^0_{K_{\delta,i}} \nabla u_H \cdot \nabla v_H.
\end{aligned}$$

This completes the proof. □

Remark 4.2.2. *Proposition 4.2.1 implies that $\mathbf{A}^0_{K_{\delta,i}}$ indeed plays a role as the homogenized tensor on each sampling domain $K_{\delta,i}$ numerically.*

4.3 The case of periodic coupling

In this section, main ingredients of error analysis are provided under periodic assumptions. Recall the definitions of ψ^j and φ^j in Chapter 3. The following two assumptions will be taken into our discussion in this section.

Assumption 4.1 (H1. Periodic coupling)**.**

1. $\mathbf{A}^\varepsilon(\mathbf{x}) := \mathbf{A}(\mathbf{x}, \frac{\mathbf{x}}{\varepsilon})$ *where* $\mathbf{A}(\mathbf{x}, \cdot)$ *is* Y*-periodic with* $Y = [0,1]^2$ *and* $\mathbf{A}(\mathbf{x}, \cdot) \in W^{1,\infty}(Y)$.

2. *On each sampling domain* $K_{\delta,i}$*, solution of the micro problem* (3.8) *with periodic coupling* (3.3a) *has regularity* $\psi^j \in H^2(K_{\delta,i})$ *and* $\mathbf{A}^\varepsilon \nabla \varphi^j \in [H^1(K_{\delta,i})]^2$.

First, we have the following result.

Lemma 4.3.1. *Under Assumption 4.1,* $\nabla \cdot \left(\mathbf{A}^\varepsilon \nabla \varphi^j\right) = 0$ *on* $K_{\delta,i}$ *a.e.*

Proof. From the definition of φ^j, it holds

$$\int_{K_{\delta,i}} \mathbf{A}^\varepsilon \nabla \varphi^j \cdot \nabla z \, \mathrm{d}\mathbf{x} = 0 \qquad \forall z \in W^1_{per}(K_{\delta,i}).$$

Let $\mathbf{1}_{K_{\delta,i}}$ be the characteristic function on $K_{\delta,i}$. Since $v - \mathcal{M}_{K_{\delta,i}}(v)\mathbf{1}_{K_{\delta,i}} \in W^1_{per}(K_{\delta,i})$ for all $v \in H^1_0(K_{\delta,i})$, we have

$$\int_{K_{\delta,i}} \mathbf{A}^\varepsilon \nabla \varphi^j \cdot \nabla v \, \mathrm{d}\mathbf{x} = 0 \qquad \forall v \in H^1_0(K_{\delta,i}).$$

The integration by parts gives $\nabla \cdot \left(\mathbf{A}^\varepsilon \nabla \varphi^j\right) = 0$ *a.e.* □

Lemma 4.3.2. *Under Assumption 4.1, it holds*

$$|\mathbf{A}^\varepsilon \nabla \varphi^j|_{1,K_{\delta,i}} \leq C |K_{\delta,i}|^{1/2} \varepsilon^{-1}. \tag{4.5}$$

Proof. Taking $z = \psi^j$ in (3.8), we have

$$\int_{K_{\delta,i}} \mathbf{A}^\varepsilon \nabla \psi^j \cdot \nabla \psi^j \, \mathrm{d}\mathbf{x} = -\int_{K_{\delta,i}} \mathbf{A}^\varepsilon \mathbf{e}_j \cdot \nabla \psi^j \, \mathrm{d}\mathbf{x}\,.$$

Due to the ellipticity of $\mathbf{A}^\varepsilon$ and Assumption 4.1, it holds that

$$\begin{aligned}
\lambda |\psi^j|^2_{1,K_{\delta,i}} &\le \left| \int_{K_{\delta,i}} \mathbf{A}^\varepsilon \mathbf{e}_j \cdot \nabla \psi^j \, \mathrm{d}\mathbf{x} \right| \\
&\le \left(\int_{K_{\delta,i}} |\mathbf{A}^\varepsilon \mathbf{e}_j|^2 \, \mathrm{d}\mathbf{x} \right)^{1/2} \left(\int_{K_{\delta,i}} |\nabla \psi^j|^2 \, \mathrm{d}\mathbf{x} \right)^{1/2} \\
&\le C |K_{\delta,i}|^{1/2} \, |\psi^j|_{1,K_{\delta,i}}.
\end{aligned}$$

Thus it implies $|\psi^j|_{1,K_{\delta,i}} \le C|K_{\delta,i}|^{1/2}$, and therefore $|\varphi^j|_{1,K_{\delta,i}} \le |\psi^j|_{1,K_{\delta,i}} + |x_j|_{1,K_{\delta,i}} \le C|K_{\delta,i}|^{1/2}$. Furthermore, the regularity of the problem implies that, see also *Remark 5.1* in [2],

$$|\psi^j|_{2,K_{\delta,i}} \le C|K_{\delta,i}|^{1/2} \varepsilon^{-1}. \tag{4.6}$$

The above results give the desired bound as follows.

$$\begin{aligned}
|\mathbf{A}^\varepsilon \nabla \varphi^j|_{1,K_{\delta,i}} &\le C \left(\int_{K_{\delta,i}} \sum_{k,\ell,m} \left| \frac{\partial}{\partial x_k} \left(A^\varepsilon_{\ell m} \frac{\partial \varphi^j}{\partial x_m} \right) \right|^2 \mathrm{d}\mathbf{x} \right)^{1/2} \\
&\le C \left(\int_{K_{\delta,i}} \sum_{k,\ell,m} \left| \frac{\partial A^\varepsilon_{\ell m}}{\partial x_k} \frac{\partial \varphi^j}{\partial x_m} \right|^2 + \left| A^\varepsilon_{\ell m} \frac{\partial^2 \varphi^j}{\partial x_k \partial x_m} \right|^2 \mathrm{d}\mathbf{x} \right)^{1/2} \\
&\le C \left(\|\nabla \mathbf{A}^\varepsilon\|_{0,\infty,K_{\delta,i}} |\varphi^j|_{1,K_{\delta,i}} + \|\mathbf{A}^\varepsilon\|_{0,\infty,K_{\delta,i}} |\varphi^j|_{2,K_{\delta,i}} \right) \\
&\le C |K_{\delta,i}|^{1/2} \varepsilon^{-1}.
\end{aligned}$$

□

The following proposition is a discretization error estimation of the micro basis function ψ^j.

Proposition 4.3.3. *Under Assumption 4.1, it holds*

$$\left|\!\left|\!\left|\psi^j - \psi_h^j\right|\!\right|\!\right|_{h,K_{\delta,i}} \leq Ch|K_{\delta,i}|^{1/2}\varepsilon^{-1}. \tag{4.7}$$

Proof. The Second Strang Lemma ([48, 10]) for the micro problems (3.7) and (3.8) implies that

$$\begin{aligned}\left|\!\left|\!\left|\psi^j - \psi_h^j\right|\!\right|\!\right|_{h,K_{\delta,i}} \leq C \Bigg(& \inf_{v_h \in W_h(K_{\delta,i})} \left|\!\left|\!\left|\psi^j - v_h\right|\!\right|\!\right|_{h,K_{\delta,i}} \\ & + \sup_{w_h \in W_h(K_{\delta,i})} \frac{\left| a_h^{K_{\delta,i}}(\psi^j, w_h) - a_h^{K_{\delta,i}}(\psi_h^j, w_h) \right|}{\left|\!\left|\!\left| w_h \right|\!\right|\!\right|_{h,K_{\delta,i}}} \Bigg).\end{aligned}$$

The first term represents the best approximation error of ψ^j. It is bounded by the micro mesh parameter h due to the standard approximation property of nonconforming element spaces. The second term, so-called the consistency error, is for nonconformity of the finite element space. Let denote the numerator of the second term by $L(w_h)$. The definitions of ψ_h^j and φ^j imply

$$\begin{aligned}|L(w_h)| :&= \left| a_h^{K_{\delta,i}}(\psi^j, w_h) - a_h^{K_{\delta,i}}(\psi_h^j, w_h) \right| \\ &= \left| \sum_{K_h \in \mathcal{T}_h} \int_{K_h} \mathbf{A}^\varepsilon \nabla \psi^j \cdot \nabla w_h \, \mathrm{d}\mathbf{x} + \sum_{K_h \in \mathcal{T}_h} \int_{K_h} \mathbf{A}^\varepsilon \mathbf{e}_j \cdot \nabla w_h \, \mathrm{d}\mathbf{x} \right| \\ &= \left| \sum_{K_h \in \mathcal{T}_h} \int_{K_h} \mathbf{A}^\varepsilon \nabla \varphi^j \cdot \nabla w_h \, \mathrm{d}\mathbf{x} \right| \\ &= \left| \sum_{K_h \in \mathcal{T}_h} \int_{\partial K_h} \mathbf{n}_{K_h} \cdot \left(\mathbf{A}^\varepsilon \nabla \varphi^j \right) w_h \, \mathrm{d}s - \sum_{K_h \in \mathcal{T}_h} \int_{K_h} \nabla \cdot \left(\mathbf{A}^\varepsilon \nabla \varphi^j \right) w_h \, \mathrm{d}\mathbf{x} \right|.\end{aligned}$$

The integration by parts is used in the last equation. Due to Lemma 4.3.1, it is bounded by

$$\left|\sum_{K_h\in\mathcal{T}_h}\int_{\partial K_h}\mathbf{n}_{K_h}\cdot\mathbf{A}^\varepsilon\nabla\varphi^j w_h\ \mathrm{ds}\right| = \left|\sum_{K_h\in\mathcal{T}_h}\sum_{e\in\mathcal{E}(K_h)}\int_e \mathbf{n}_{K_h}\cdot\mathbf{A}^\varepsilon\nabla\varphi^j w_h\ \mathrm{ds}\right| \leq \sum_{e\in\mathcal{E}_h}\left|\int_e \mathbf{n}_e\cdot\mathbf{A}^\varepsilon\nabla\varphi^j[w_h]_e\ \mathrm{ds}\right|.$$

Let denote K^+ and K^- two adjacent elements that share a common interior edge e. Let define the average of w_h over the edge e as $\overline{w_h} := \frac{1}{|e|}\int_e w_h^+ = \frac{1}{|e|}\int_e w_h^-$ where $w_h^\iota = w_h|_{K^\iota}$ for $\iota = +,-$. Note that the integral value of a function in $W_h(K_{\delta,i})$ on each interior edge is well-defined due to the definition of the nonconforming finite space. The regularity in Assumption 4.1 implies

$$\begin{aligned}
&\int_e \mathbf{n}_e\cdot\mathbf{A}^\varepsilon\nabla\varphi^j[w_h]_e\ \mathrm{ds}\\
&\quad= \int_e \mathbf{n}_e\cdot\mathbf{A}^\varepsilon\nabla\varphi^j\left[w_h-\overline{w_h}\right]_e\ \mathrm{ds}\\
&\quad= \int_e \mathbf{n}_e\cdot\left(\mathbf{A}^\varepsilon\nabla\varphi^j-\mathcal{M}_e\left(\mathbf{A}^\varepsilon\nabla\varphi^j\right)\right)\left[w_h-\overline{w_h}\right]_e\ \mathrm{ds}\\
&\quad\leq \sum_{\iota=+,-}\left(\int_e\left|\mathbf{A}^\varepsilon\nabla\varphi^j-\mathcal{M}_e\left(\mathbf{A}^\varepsilon\nabla\varphi^j\right)\right|^2\mathrm{ds}\right)^{\frac{1}{2}}\left(\int_e|w_h^\iota-\overline{w_h}|^2\ \mathrm{ds}\right)^{\frac{1}{2}}.
\end{aligned}$$

Due to the trace theorem and Poincaré inequality on the reference domain $\widehat{K}$ with the standard scaling argument, the first term is bounded by

$$\begin{aligned}
\int_e\left|\mathbf{A}^\varepsilon\nabla\varphi^j-\mathcal{M}_e\left(\mathbf{A}^\varepsilon\nabla\varphi^j\right)\right|^2\mathrm{ds} &\leq Ch^{-1}\int_{\widehat{e}}\left|\widehat{\mathbf{A}}^\varepsilon\nabla\widehat{\varphi}^j-\mathcal{M}_{\widehat{e}}\left(\widehat{\mathbf{A}}^\varepsilon\nabla\widehat{\varphi}^j\right)\right|^2\mathrm{d}\widehat{\mathrm{s}}\\
&\leq Ch^{-1}\left\|\widehat{\mathbf{A}}^\varepsilon\nabla\widehat{\varphi}^j-\mathcal{M}_{\widehat{e}}\left(\widehat{\mathbf{A}}^\varepsilon\nabla\widehat{\varphi}^j\right)\right\|^2_{1,\widehat{K}}\\
&\leq Ch^{-1}\left|\widehat{\mathbf{A}}^\varepsilon\nabla\widehat{\varphi}^j\right|^2_{1,\widehat{K}} \leq Ch\left|\mathbf{A}^\varepsilon\nabla\varphi^j\right|^2_{1,K^\iota}
\end{aligned}$$

for each $\iota = +, -$. Here we use a simple fact that the bilinear transformation F_{K_h} linearly transforms an edge $\widehat{e}$ of the reference domain onto e. Note that the average value of a function is preserved by a linear transformation. The second term is bounded by

$$\begin{aligned}\int_e |w_h^\iota - \overline{w_h}|^2 \, \mathrm{ds} &\le Ch \int_{\widehat{e}} |\widehat{w}_h^\iota - \overline{w_h}|^2 \, \mathrm{d\widehat{s}} \le Ch \, \|\widehat{w}_h^\iota - \overline{w_h}\|_{1,\widehat{K}}^2 \\ &\le Ch \, |\widehat{w}_h^\iota|_{1,\widehat{K}}^2 \le Ch \, |w_h^\iota|_{1,K^\iota}^2 \, .\end{aligned}$$

Consequently, we have

$$\begin{aligned}|L(w_h)| &\le C \sum_{K_h \in \mathcal{T}_h} h |\mathbf{A}^\varepsilon \nabla \varphi^j|_{1,K_h} |w_h|_{1,K_h} \\ &\le Ch \left(\sum_{K_h \in \mathcal{T}_h} |\mathbf{A}^\varepsilon \nabla \varphi^j|_{1,K_h}^2 \right)^{1/2} \left(\sum_{K_h \in \mathcal{T}_h} |w_h|_{1,K_h}^2 \right)^{1/2} \\ &= Ch |\mathbf{A}^\varepsilon \nabla \varphi^j|_{1,K_{\delta,i}} |\!|\!| w_h |\!|\!|_{h,K_{\delta,i}}.\end{aligned}$$

Finally, Lemma 4.3.2 and (4.6) imply the desired error estimate:

$$\left|\!\left|\!\left| \psi^j - \psi_h^j \right|\!\right|\!\right|_{h,K_{\delta,i}} \le Ch \left(|\psi^j|_{2,K_{\delta,i}} + |\mathbf{A}^\varepsilon \nabla \varphi^j|_{1,h,K_{\delta,i}} \right) \le Ch |K_{\delta,i}|^{1/2} \varepsilon^{-1}.$$

□

Remark 4.3.4. *In a similar way, we can obtain* (4.7) *for the case of* $d = 3$*, under an additional assumption such that* F_{K_h} *is a linear transformation.*

The following proposition shows difference between the recovered homogenized tensors.

Proposition 4.3.5. *Under Assumption 4.1, it holds*

$$\sup_{K_{\delta,i}} \|\overline{\mathbf{A}}^0_{K_{\delta,i}} - \mathbf{A}^0_{K_{\delta,i}}\|_2 \leq C \left(\frac{h}{\varepsilon}\right)^2. \tag{4.8}$$

Proof. For given sampling domain $K_{\delta,i}$, definition of the recovered homogenized tensors (4.1) implies

$$\begin{aligned}
[A^0_{K_{\delta,i}}]_{jk} &= \frac{1}{|K_{\delta,i}|} \sum_{K_h \in \mathcal{T}_h} \int_{K_h} \sum_{\ell=1}^{d} A^\varepsilon_{j\ell} \frac{\partial \varphi^k_h}{\partial x_\ell} \, \mathrm{d}\mathbf{x} \\
&= \frac{1}{|K_{\delta,i}|} \sum_{K_h \in \mathcal{T}_h} \int_{K_h} \mathbf{A}^\varepsilon \nabla \varphi^k_h \cdot \nabla x_j \, \mathrm{d}\mathbf{x} \\
&= \frac{1}{|K_{\delta,i}|} \sum_{K_h \in \mathcal{T}_h} \int_{K_h} \mathbf{A}^\varepsilon \nabla \varphi^k_h \cdot \nabla \varphi^j_h \, \mathrm{d}\mathbf{x},
\end{aligned}$$

and a similar expression for $[\overline{\mathbf{A}}^0_{K_{\delta,i}}]_{jk}$. Thus

$$\begin{aligned}
&\left| [\mathbf{A}^0_{K_{\delta,i}}]_{jk} - [\overline{\mathbf{A}}^0_{K_{\delta,i}}]_{jk} \right| \\
&= \left| \frac{1}{|K_{\delta,i}|} \sum_{K_h \in \mathcal{T}_h} \int_{K_h} \mathbf{A}^\varepsilon \nabla \varphi^k_h \cdot \nabla \varphi^j_h - \mathbf{A}^\varepsilon \nabla \varphi^k \cdot \nabla \varphi^j \, \mathrm{d}\mathbf{x} \right| \\
&= \left| \frac{1}{|K_{\delta,i}|} \sum_{K_h \in \mathcal{T}_h} \int_{K_h} \mathbf{A}^\varepsilon \nabla \left(\varphi^k_h - \varphi^k\right) \cdot \nabla \left(\varphi^j_h - \varphi^j\right) \right. \\
&\qquad\qquad \left. + \mathbf{A}^\varepsilon \nabla \varphi^k \cdot \nabla \left(\varphi^j_h - \varphi^j\right) + \mathbf{A}^\varepsilon \nabla \left(\varphi^k_h - \varphi^k\right) \cdot \nabla \varphi^j \, \mathrm{d}\mathbf{x} \right| \\
&\leq \frac{1}{|K_{\delta,i}|} \left| \sum_{K_h \in \mathcal{T}_h} \int_{K_h} \mathbf{A}^\varepsilon \nabla \left(\varphi^k_h - \varphi^k\right) \cdot \nabla \left(\varphi^j_h - \varphi^j\right) \, \mathrm{d}\mathbf{x} \right| \\
&\quad + \frac{1}{|K_{\delta,i}|} \left| \sum_{K_h \in \mathcal{T}_h} -\int_{K_h} \nabla \cdot \left(\mathbf{A}^\varepsilon \nabla \varphi^k\right) \left(\varphi^j_h - \varphi^j\right) \mathrm{d}\mathbf{x} + \int_{\partial K_h} \mathbf{n}_{K_h} \cdot \mathbf{A}^\varepsilon \nabla \varphi^k \left(\varphi^j_h - \varphi^j\right) \mathrm{d}\mathbf{s} \right| \\
&\quad + \frac{1}{|K_{\delta,i}|} \left| \sum_{K_h \in \mathcal{T}_h} -\int_{K_h} \nabla \cdot \left(\mathbf{A}^\varepsilon \nabla \varphi^j\right) \left(\varphi^k_h - \varphi^k\right) \mathrm{d}\mathbf{x} + \int_{\partial K_h} \mathbf{n}_{K_h} \cdot \mathbf{A}^\varepsilon \nabla \varphi^j \left(\varphi^k_h - \varphi^k\right) \mathrm{d}\mathbf{s} \right|.
\end{aligned}$$

The last inequality is due to the integration by parts and the symmetry of $\mathbf{A}^\varepsilon$. By Lemmas 4.3.1 and 4.3.2, Proposition 4.3.3 and using a similar technique, it is bounded by

$$
\begin{aligned}
&\leq \frac{\Lambda}{|K_{\delta,i}|}\left|\!\left|\!\left|\varphi_h^k-\varphi^k\right|\!\right|\!\right|_{h,K_{\delta,i}}\left|\!\left|\!\left|\varphi_h^j-\varphi^j\right|\!\right|\!\right|_{h,K_{\delta,i}}\\
&\quad+\frac{1}{|K_{\delta,i}|}\left|\sum_{K_h\in\mathcal{T}_h}\int_{\partial K_h}\mathbf{n}_{K_h}\cdot\mathbf{A}^\varepsilon\nabla\varphi^k\left(\varphi_h^j-\varphi^j\right)\,\mathrm{ds}\right|\\
&\quad+\frac{1}{|K_{\delta,i}|}\left|\sum_{K_h\in\mathcal{T}_h}\int_{\partial K_h}\mathbf{n}_{K_h}\cdot\mathbf{A}^\varepsilon\nabla\varphi^j\left(\varphi_h^k-\varphi^k\right)\,\mathrm{ds}\right|\\
&\leq \frac{\Lambda}{|K_{\delta,i}|}\left|\!\left|\!\left|\varphi_h^k-\varphi^k\right|\!\right|\!\right|_{h,K_{\delta,i}}\left|\!\left|\!\left|\varphi_h^j-\varphi^j\right|\!\right|\!\right|_{h,K_{\delta,i}}\\
&\quad+\frac{h}{|K_{\delta,i}|}|\mathbf{A}^\varepsilon\nabla\varphi^k|_{1,K_{\delta,i}}\left|\!\left|\!\left|\varphi_h^j-\varphi^j\right|\!\right|\!\right|_{h,K_{\delta,i}}+\frac{h}{|K_{\delta,i}|}|\mathbf{A}^\varepsilon\nabla\varphi^j|_{1,K_{\delta,i}}\left|\!\left|\!\left|\varphi_h^k-\varphi^k\right|\!\right|\!\right|_{h,K_{\delta,i}}\\
&\leq C\left(\frac{h}{\varepsilon}\right)^2.
\end{aligned}
$$

□

4.4 The case of Dirichlet coupling

In this section we consider the following assumptions and Dirichlet coupling condition for micro problems.

Assumption 4.2 (H2. Dirichlet coupling)**.**

1. $\mathbf{A}^\varepsilon(\mathbf{x})\in W^{1,\infty}(K_H)$ *with* $|A^\varepsilon_{jk}|_{0,\infty,K_H}\leq C$ *and* $|\nabla A^\varepsilon_{jk}|_{0,\infty,K_H}\leq C/\varepsilon$ *for all* $K_H\in\mathcal{T}_H$.

2. *On each sampling domain* $K_{\delta,i}$*, solution of the micro problem* (3.8) *with Dirichlet coupling* (3.3a) *has regularity* $\psi^j\in H^2(K_{\delta,i})$ *and* $\mathbf{A}^\varepsilon\nabla\varphi^j\in[H^1(K_{\delta,i})]^2$.

The definition of φ^j implies

$$\int_{K_{\delta,i}} \mathbf{A}^\varepsilon \nabla \varphi^j \cdot \nabla z \, \mathrm{d}\mathbf{x} = 0 \quad \forall z \in H_0^1(K_{\delta,i}).$$

Applying the integration by parts, we have $\nabla \cdot \left(\mathbf{A}^\varepsilon \nabla \varphi^j\right) = 0$ *a.e.* It implies the same results in Propositions 4.3.3 and 4.3.5 under Assumption 4.2, instead of Assumption 4.1.

4.5 A priori error estimate

4.5.1 Macro error

Under sufficient regularity of u^0, for instance H^2, the standard analysis for nonconforming finite elements and approximation by quadrature formulas [49] imply that

$$\left|\left|\left| u^0 - u_H^0 \right|\right|\right|_H \leq CH \|u^0\|_2. \tag{4.9}$$

4.5.2 Modeling error

Due to the uniform ellipticity of $\overline{a}_H$, we have

$$\begin{aligned}
\left|\left|\left| u_H^0 - \overline{u}_H \right|\right|\right|_H^2 &\leq \overline{a}_H(u_H^0 - \overline{u}_H, u_H^0 - \overline{u}_H) \\
&= \overline{a}_H(u_H^0, u_H^0 - \overline{u}_H) - \overline{a}_H(\overline{u}_H, u_H^0 - \overline{u}_H) \\
&= \overline{a}_H(u_H^0, u_H^0 - \overline{u}_H) - (f, u_H^0 - \overline{u}_H) \\
&= \overline{a}_H(u_H^0, u_H^0 - \overline{u}_H) - a_H^0(u_H^0, u_H^0 - \overline{u}_H).
\end{aligned}$$

Note that the definitions of $\overline{u}_H$ and u_H^0 are applied successively in the above equations. Dividing by a factor $\left|\left|\left|u_H^0 - \overline{u}_H\right|\right|\right|_H$ implies a Strang-type inequality

$$\left|\left|\left|u_H^0 - \overline{u}_H\right|\right|\right|_H \leq \sup_{w_H \in V_H} \frac{\left|\overline{a}_H(u_H^0, w_H) - a_H^0(u_H^0, w_H)\right|}{\left|\left|\left|w_H\right|\right|\right|_H}. \tag{4.10}$$

Proposition 4.2.1 implies that the numerator is bounded as

$$\begin{aligned}
&\left|\overline{a}_H(u_H^0, w_H) - a_H^0(u_H^0, w_H)\right| \\
&= \left| \sum_{K_H \in \mathcal{T}_H} \sum_{i=1}^{I} \omega_i \, \overline{\mathbf{A}}^0_{K_{\delta,i}} \nabla u_H^0 \cdot \nabla w_H - \sum_{K_H \in \mathcal{T}_H} \sum_{i=1}^{I} \omega_i \, \mathbf{A}^0(\mathbf{x}_i) \nabla u_H^0(\mathbf{x}_i) \cdot \nabla w_H(\mathbf{x}_i) \right| \\
&\leq \sum_{K_H \in \mathcal{T}_H} \sum_{i=1}^{I} \omega_i \left| \left(\overline{\mathbf{A}}^0_{K_{\delta,i}} - \mathbf{A}^0(\mathbf{x}_i) \right) \nabla u_H^0 \cdot \nabla w_H \right| \\
&\leq \sup_{K_{\delta,i}} \|\overline{\mathbf{A}}^0_{K_{\delta,i}} - \mathbf{A}^0(\mathbf{x}_i)\|_2 \; \left|\left|\left|u_H^0\right|\right|\right|_H \; \left|\left|\left|w_H\right|\right|\right|_H,
\end{aligned}$$

and we have

$$\left|\left|\left|u_H^0 - \overline{u}_H\right|\right|\right|_H \leq C \sup_{K_{\delta,i}} \|\overline{\mathbf{A}}^0_{K_{\delta,i}} - \mathbf{A}^0(\mathbf{x}_i)\|_2. \tag{4.11}$$

The analysis in [22] reads the difference between two homogenized tensors, one from the homogenization theory and the other from micro problems with sampling domains of size δ. If we assume the local periodicity of $\mathbf{A}^\varepsilon$, then

$$\sup_{K_{\delta,i}} \|\mathbf{A}^0 - \overline{\mathbf{A}}^0_{K_{\delta,i}}\|_2 \leq \begin{cases} C\varepsilon & \text{if periodic coupling with} \\ & \delta/\varepsilon \in \mathbb{N} \text{ is used for (3.5)}, \\ C\left(\frac{\varepsilon}{\delta} + \delta\right) & \text{if Dirichlet coupling or} \\ & \delta/\varepsilon \notin \mathbb{N} \text{ is used.} \end{cases} \tag{4.12}$$

4.5.3 Micro error

Due to the uniform ellipticity of a_H, we have

$$\begin{aligned}
|||\overline{u}_H - u_H|||_H^2 &\leq a_H(\overline{u}_H - u_H, \overline{u}_H - u_H) \\
&= a_H(\overline{u}_H, \overline{u}_H - u_H) - a_H(u_H, \overline{u}_H - u_H) \\
&= a_H(\overline{u}_H, \overline{u}_H - u_H) - (f, \overline{u}_H - u_H) \\
&= a_H(\overline{u}_H, \overline{u}_H - u_H) - \overline{a}_H(\overline{u}_H, \overline{u}_H - u_H).
\end{aligned}$$

Therefore it holds

$$|||\overline{u}_H - u_H|||_H \leq \sup_{w_H \in V_H} \frac{|a_H(\overline{u}_H, w_H) - \overline{a}_H(\overline{u}_H, w_H)|}{|||w_H|||_H}. \tag{4.13}$$

Propositions 4.2.1 and 4.3.5 imply that the numerator is bounded as

$$\begin{aligned}
&|a_H(\overline{u}_H, w_H) - \overline{a}_H(\overline{u}_H, w_H)| \\
&= \left| \sum_{K_H \in \mathcal{T}_H} \sum_{i=1}^{I} \omega_i \ \mathbf{A}^0_{K_{\delta,i}} \nabla \overline{u}_H \cdot \nabla w_H - \sum_{K_H \in \mathcal{T}_H} \sum_{i=1}^{I} \omega_i \ \overline{\mathbf{A}}^0_{K_{\delta,i}} \nabla \overline{u}_H \cdot \nabla w_H \right| \\
&\leq \sum_{K_H \in \mathcal{T}_H} \sum_{i=1}^{I} \omega_i \left| \left(\mathbf{A}^0_{K_{\delta,i}} - \overline{\mathbf{A}}^0_{K_{\delta,i}} \right) \nabla \overline{u}_H \cdot \nabla w_H \right| \\
&\leq \sup_{K_{\delta,i}} \|\mathbf{A}^0_{K_{\delta,i}} - \overline{\mathbf{A}}^0_{K_{\delta,i}}\|_2 \ |||\overline{u}_H|||_H \ |||w_H|||_H \\
&\leq C \left(\frac{h}{\varepsilon} \right)^2 |||\overline{u}_H|||_H \ |||w_H|||_H,
\end{aligned}$$

and we have

$$|||\overline{u}_H - u_H|||_H \leq C \left(\frac{h}{\varepsilon} \right)^2. \tag{4.14}$$

4.6 Main theorem for error estimates

The Aubin-Nitsche duality argument gives L^2 error estimates, see [22].

Theorem 4.6.1. *Let u^0 and u_H be the solutions of (2.2) and (3.6). Then, the followings hold under various assumptions:*

1. *under Assumption 4.1, if periodic coupling with $\delta/\varepsilon \in \mathbb{N}$ is used, then*

$$\left|\left|\left|u^0 - u_H\right|\right|\right|_H \leq C\left(H + \varepsilon + \left(\frac{h}{\varepsilon}\right)^2\right), \tag{4.15a}$$

$$\|u^0 - u_H\|_0 \leq C\left(H^2 + \varepsilon + \left(\frac{h}{\varepsilon}\right)^2\right); \tag{4.15b}$$

2. *under Assumption 4.1, if periodic coupling with $\delta/\varepsilon \notin \mathbb{N}$ is used, then*

$$\left|\left|\left|u^0 - u_H\right|\right|\right|_H \leq C\left(H + \left(\frac{\varepsilon}{\delta} + \delta\right) + \left(\frac{h}{\varepsilon}\right)^2\right), \tag{4.16a}$$

$$\|u^0 - u_H\|_0 \leq C\left(H^2 + \left(\frac{\varepsilon}{\delta} + \delta\right) + \left(\frac{h}{\varepsilon}\right)^2\right); \tag{4.16b}$$

3. *under Assumption 4.1 and Assumption 4.2, if Dirichlet coupling is used, then*

$$\left|\left|\left|u^0 - u_H\right|\right|\right|_H \leq C\left(H + \left(\frac{\varepsilon}{\delta} + \delta\right) + \left(\frac{h}{\varepsilon}\right)^2\right), \tag{4.17a}$$

$$\|u^0 - u_H\|_0 \leq C\left(H^2 + \left(\frac{\varepsilon}{\delta} + \delta\right) + \left(\frac{h}{\varepsilon}\right)^2\right); \tag{4.17b}$$

4. *under Assumption 4.2, if Dirichlet coupling is used, then*

$$\left|\left|\left|u^0 - u_H\right|\right|\right|_H \leq C\left(H + \sup_{K_{\delta,i}} \|\overline{\mathbf{A}}^0_{K_{\delta,i}} - \mathbf{A}^0(\mathbf{x}_i)\|_2 + \left(\frac{h}{\varepsilon}\right)^2\right), \tag{4.18a}$$

$$\|u^0 - u_H\|_0 \leq C\left(H^2 + \sup_{K_{\delta,i}} \|\overline{\mathbf{A}}^0_{K_{\delta,i}} - \mathbf{A}^0(\mathbf{x}_i)\|_2 + \left(\frac{h}{\varepsilon}\right)^2\right). \tag{4.18b}$$

Chapter 5

Numerical Results

When one uses periodic coupling condition for the micro problems, the corresponding algebraic system of equations for each micro problem must be constructed to consider two important properties — periodicity and zero-integral property. Technically, one can enforce the solution to satisfy these properties through either the discrete function space or the formulation for the problem.

In [5] these two properties are imposed through the formulation, by use of a Lagrange multiplier and a constraint matrix. This approach is quite simple to implement. However, it requires to solve an expanded indefinite linear system of equations. Furthermore, the authors solve the linear system with a direct method because its structure is not suitable to use efficient iterative methods for the saddle point problems.

In order to overcome such disadvantages, we can alternatively use the numerical schemes recently proposed for the P_1–nonconforming quadrilateral finite element with periodic boundary condition. These alternatives are based

on a simple iterative method without any help of a Lagrange multiplier or a constraint matrix, since they enforce the discrete function space with the periodic property. The zero-integral condition is also treated in efficient ways.

For micro problems in all numerical examples, we employ one of these alternative approaches: the option 2, whose trial and test functions are $\mathfrak{C}^\flat$ for a symmetric positive semi-definite system. We will investigate efficiency of the alternative approach for micro problems in Section 5.1.1. Furthermore, we use 2-point Gauss-Legendre quadrature formula for each coordinate in all numerical examples.

5.1 Periodic diagonal example

The first example is the multiscale elliptic problem of which the coefficient tensor has anisotropic periodicity in micro scale. On $\Omega = (0,1)^2$, we consider the problem (2.1) with $\mathbf{A}^\varepsilon(\mathbf{x}) = \begin{pmatrix} \sqrt{2} + \sin(2\pi x_1/\varepsilon) & 0 \\ 0 & \sqrt{2} + \sin(2\pi x_2/\varepsilon) \end{pmatrix}$, where ε is 10^{-3}. By the homogenization theory, it can be easily shown that the associated homogenized tensor $\mathbf{A}^0$ is equal to I, the identity tensor. $f(\mathbf{x})$ is set to satisfy that the associated homogenized elliptic problem has the exact solution $u^0(\mathbf{x}) = \sin(\pi x_1)\sin(\pi x_2)$. For the sake of simplicity we use the macro and the micro mesh consisting of uniform squares. The size parameter δ of each sampling domain is set to be same as ε. We use periodic coupling for micro problems.

Table 5.1 shows error in energy norm, and in L^2-norm, and the difference between two observable homogenized tensors $\mathbf{A}^0$ and $\mathbf{A}^0_{K_{\delta,i}}$ in the Frobenius norm. Note that the matrix 2-norm for a finite dimensional matrix is equivalent to the Frobenius norm. The theoretical error estimates (4.15) depend on H as well as h. We can observe that the error is decreasing as H is decreasing,

but there is a critical H value where the error does not decrease anymore for fixed h, as particularly in L^2-norm. Furthermore, in order to observe the convergence rate as in (4.15) we have to consider simultaneous reduction of H and h in different orders, since the theorem shows the dependency on H and h with their own convergence orders. For instance, simultaneous reduction of H in the second-order and h in the first-order gives convergence order of 2 in energy norm, as observed in the table. The error in L^2-norm is similar. The numerical results confirm (4.15) in Theorem 4.6.1, the main convergence result for periodic cases.

H	h/ε=1/4	1/8	1/16	1/32	1/64
			$\|\|\|u^0 - u_H\|\|\|_H$		
1/2	1.33E-00	1.35E-00	1.36E-00	1.36E-00	1.36E-00
1/4	6.98E-01	6.99E-01	7.03E-01	7.04E-01	7.05E-01
1/8	3.75E-01	3.55E-01	3.54E-01	3.55E-01	3.55E-01
1/16	2.77E-01	1.84E-01	1.78E-01	1.78E-01	1.78E-01
1/32	1.54E-01	1.04E-01	8.98E-02	8.90E-02	8.90E-02
1/64	1.93E-01	6.79E-02	4.66E-02	4.46E-02	4.45E-02
			$\|u^0 - u_H\|_0$		
1/2	1.21E-01	1.20E-01	1.20E-01	1.21E-01	1.21E-01
1/4	4.58E-02	3.22E-02	3.04E-02	3.04E-02	3.04E-02
1/8	3.50E-02	1.41E-02	8.21E-03	7.64E-03	7.60E-03
1/16	4.96E-02	1.20E-02	3.69E-03	2.06E-03	1.91E-03
1/32	2.88E-02	1.25E-02	3.20E-03	9.31E-04	5.16E-04
1/64	4.23E-02	1.16E-02	3.17E-03	8.09E-04	2.33E-04
			$\sup_{K_{\delta,i}} \|\mathbf{A}^0 - \mathbf{A}^0_{K_{\delta,i}}\|_F$		
1/2	1.00E-01	3.47E-02	9.02E-03	2.27E-03	5.68E-04
1/4	1.07E-01	3.42E-02	9.02E-03	2.27E-03	5.68E-04
1/8	1.04E-01	3.44E-02	9.02E-03	2.27E-03	5.68E-04
1/16	1.56E-01	3.43E-02	9.02E-03	2.27E-03	5.68E-04
1/32	8.65E-02	3.62E-02	9.02E-03	2.27E-03	5.68E-04
1/64	1.44E-01	3.37E-02	9.02E-03	2.27E-03	5.68E-04

Table 5.1. Error table of the example in Section 5.1

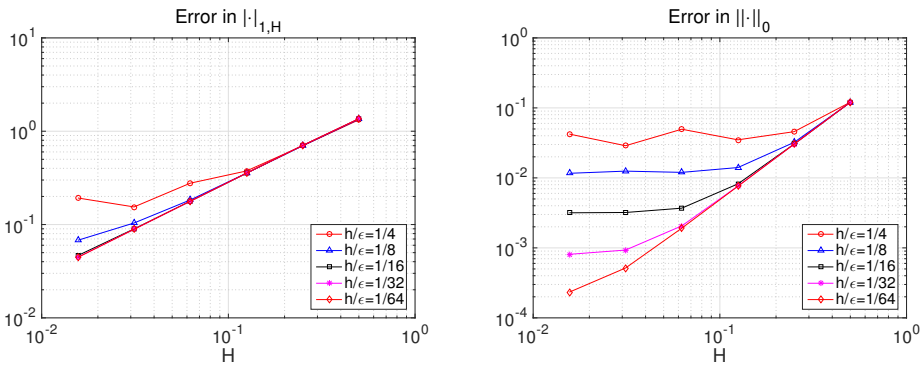

Figure 5.1. Error plots of the example in Section 5.1

5.1.1 Comparison between approaches to solve micro problem

As mentioned in the beginning of this chapter, we mainly use the alternative iterative approach based on the Conjugate Gradient method (CG) for micro problems with Dirichlet coupling as well as periodic coupling condition. Here we investigate the efficiency of the alternative iterative approach over the direct solver for the periodic coupling case.

We consider three approaches for implementation of the FEHMM scheme. They only differ in way for setting and solving linear systems corresponding to micro problems. We describe these approaches in brief.

The first approach uses the Q_1 bilinear conforming element to assemble a linear system for each micro problem. As mentioned in [5], the assembled system is indefinite due to blocks for constraints. The number of rows of the system matrix is equal to $n^2 + 4n + 3$, where n is the number of discretization in each coordinate of each sampling domain. A direct solver from LAPACK is used to solve the indefinite system numerically. We name this approach 'DirQ1'.

The second approach, denoted by 'DirP1NC', assembles a linear system using the P_1–nonconforming quadrilateral element in similar manner as the

previous approach. The only difference between two approaches is kind of used finite elements. Thus the system matrix in this approach is also indefinite, and has the size of $n^2 + 4n + 2$. This system is solved by the same direct solver as the previous approach.

The last approach, denoted by 'IterP1NC' and mainly used throughout the whole numerical implementations in our discussion, also uses the P_1–nonconforming quadrilateral element but in different manner unlike two previous approaches. This approach uses a basis for the discrete function space with periodic property, and assembles a corresponding symmetric positive semi-definite system with rank 1 deficiency. The zero-integral property is imposed as a post-processing procedure. The size of the system matrix is $n^2 + 1$, less than previous, due to the absence of constraint blocks. We solve this semi-definite system in iterative way, by use of the CG.

For the comparison between three approaches, we again consider the same multiscale elliptic problem in Section 5.1. Each of three approaches is used to solve micro problems numerically, and (sum of) the elapsed time for micro solver is measured. Table 5.2 shows the elapsed time in seconds for each approach in various combinations of macro and micro mesh size. We can observe the elapsed time in IterP1NC approach is much less than other direct approaches.

	$h/\varepsilon = 1/32$			$h/\varepsilon = 1/64$		
H	DirQ1	DirP1NC	IterP1NC	DirQ1	DirP1NC	IterP1NC
1/2	6.8	4.2	1.6	326.0	295.5	12.8
1/4	20.3	16.4	6.3	1303.3	1164.1	52.1
1/8	73.9	66.8	25.8	5147.1	5143.5	213.9
1/16	288.8	260.8	102.5	20943.3	18845.1	845.5

Table 5.2. Elapsed time for micro solvers

5.2 Periodic example with off-diagonal terms

In this example we take a tensor whose components are all nonzero with single directional periodicity. For $\varepsilon = 10^{-3}$, consider the problem (2.1) with a multiscale tensor $\mathbf{A}^\varepsilon(\mathbf{x}) = \begin{pmatrix} \sqrt{2} + \sin(2\pi x_1/\varepsilon) & \frac{1}{2} + \frac{1}{2\sqrt{2}}\sin(2\pi x_1/\varepsilon) \\ \frac{1}{2} + \frac{1}{2\sqrt{2}}\sin(2\pi x_1/\varepsilon) & 2 + \sin(2\pi x_1/\varepsilon) \end{pmatrix}$, and the associated homogenized tensor $\mathbf{A}^0(\mathbf{x}) = \begin{pmatrix} 1 & \frac{1}{2\sqrt{2}} \\ \frac{1}{2\sqrt{2}} & \frac{17-\sqrt{2}}{8} \end{pmatrix}$. We set $f(\mathbf{x})$ to satisfy that the exact homogenized solution $u^0(\mathbf{x}) = \sin(\pi x_1)\sin(\pi x_2)$. As the previous example, we use the macro and the micro mesh consisting of uniform squares, and $\delta = \varepsilon$ with periodic coupling for each micro problem. Table 5.3 shows that similar results can be obtained in more general periodic case.

5.3 Example with noninteger-ε-multiple sampling domain and Dirichlet coupling

This example, which is originated from [1], is to investigate the effect of Dirichlet coupling on micro problems. Consider the multiscale elliptic problem with mixed boundary condition

$$\begin{aligned} -\nabla \cdot \Big(\mathbf{A}^\varepsilon(\mathbf{x})\nabla u^\varepsilon(\mathbf{x})\Big) &= f(\mathbf{x}) \quad \text{in } \Omega = (0,1)^2, \\ u^\varepsilon|_{\Gamma_D} &= 0, \\ \nu \cdot \mathbf{A}^\varepsilon \nabla u^\varepsilon|_{\Gamma_N} &= 0, \end{aligned}$$

where $\Gamma_D = \{(x_1, x_2) \mid x_1 = 0 \text{ or } 1\} \cap \partial\Omega$ and $\Gamma_N = \partial\Omega \setminus \Gamma_D$. We use the multiscale coefficient tensor $\mathbf{A}^\varepsilon(\mathbf{x}) = \big(2 + \cos(2\pi x_1/\varepsilon)\big)I$ where $\varepsilon = 10^{-3}$, the associated homogenized tensor $\mathbf{A}^0(\mathbf{x}) = diag(\sqrt{3}, 2)$, and $f \equiv 1$ which admits

H	h/ε=1/4	1/8	1/16	1/32	1/64
			$\|\|\|u^0 - u_H\|\|\|_H$		
1/2	1.35E-00	1.36E-00	1.36E-00	1.36E-00	1.36E-00
1/4	6.98E-01	7.02E-01	7.04E-01	7.05E-01	7.05E-01
1/8	3.56E-01	3.54E-01	3.55E-01	3.55E-01	3.55E-01
1/16	1.96E-01	1.78E-01	1.78E-01	1.78E-01	1.78E-01
1/32	1.01E-01	9.12E-02	8.91E-02	8.90E-02	8.90E-02
1/64	8.72E-02	4.86E-02	4.48E-02	4.45E-02	4.45E-02
			$\|u^0 - u_H\|_0$		
1/2	1.20E-01	1.20E-01	1.21E-01	1.21E-01	1.21E-01
1/4	3.30E-02	3.06E-02	3.04E-02	3.04E-02	3.04E-02
1/8	1.54E-02	8.83E-03	7.68E-03	7.60E-03	7.60E-03
1/16	2.00E-02	4.91E-03	2.25E-03	1.92E-03	1.90E-03
1/32	1.13E-02	4.80E-03	1.29E-03	5.63E-04	4.81E-04
1/64	1.68E-02	4.45E-03	1.21E-03	3.25E-04	1.41E-04
			$\sup_{K_{\delta,i}} \|\mathbf{A}^0 - \mathbf{A}^0_{K_{\delta,i}}\|_F$		
1/2	7.99E-02	2.76E-02	7.17E-03	1.80E-03	4.52E-04
1/4	8.54E-02	2.72E-02	7.17E-03	1.80E-03	4.52E-04
1/8	8.26E-02	2.74E-02	7.17E-03	1.80E-03	4.52E-04
1/16	1.24E-01	2.73E-02	7.17E-03	1.80E-03	4.52E-04
1/32	6.88E-02	2.88E-02	7.17E-03	1.80E-03	4.52E-04
1/64	1.15E-01	2.68E-02	7.18E-03	1.80E-03	4.52E-04

Table 5.3. Error table of the example in Section 5.2

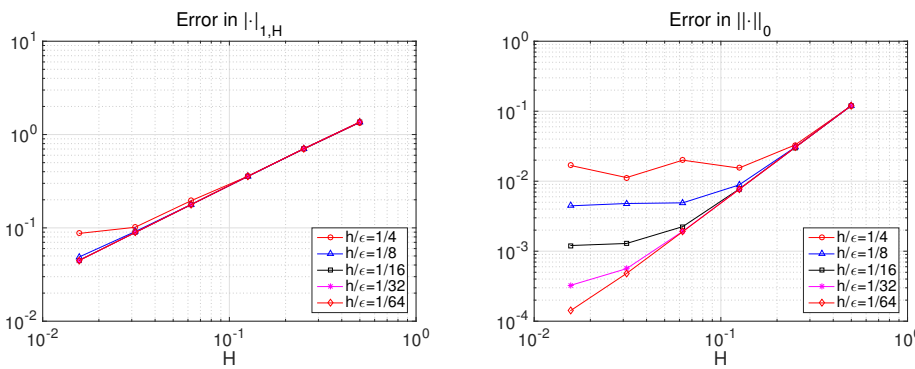

Figure 5.2. Error plots of the example in Section 5.2

the exact homogenized solution $u^0(\mathbf{x}) = -\frac{1}{2\sqrt{3}}x_1(x_1-1)$. We use Dirichlet coupling on each micro problem. We have three options for sampling domain size δ which are not multiple of ε; $\delta = 1.1\varepsilon$, 3.1ε and $\sqrt{\varepsilon}$. The last option is deduced from (4.17) for the optimal convergence. The number of micro elements is fixed sufficiently large to guarantee that the micro error (4.14) can not disrupt the tendency of the total error.

We can observe that error varies depending on size of sampling domains. As shown in Table 5.4, the bigger size of sampling domains gives the more accurate results.

H	$\delta = 1.1\varepsilon$ (Diri.)	3.1ε (Diri.)	$\sqrt{\varepsilon}$ (Diri.,512)
		$\|\|\|u^0 - u_H\|\|\|_H$	
1/2	8.41E-02	8.34E-02	8.33E-02
1/4	4.22E-02	4.17E-02	4.17E-02
1/8	2.51E-02	2.14E-02	2.09E-02
1/16	1.50E-02	1.11E-02	1.04E-02
1/32	1.14E-02	6.33E-03	5.28E-03
		$\|u^0 - u_H\|_0$	
1/2	1.60E-02	1.41E-02	1.34E-02
1/4	5.07E-03	3.91E-03	3.33E-03
1/8	5.11E-03	2.29E-03	1.20E-03
1/16	3.56E-03	1.38E-03	3.57E-04
1/32	2.84E-03	1.03E-03	2.39E-04
		$\sup_{K_{\delta,i}} \|\mathbf{A}^0 - \mathbf{A}^0_{K_{\delta,i}}\|_F$	
1/2	1.59E-01	5.34E-02	1.16E-02
1/4	8.45E-02	2.97E-02	4.82E-03
1/8	1.78E-01	6.01E-02	1.64E-02
1/16	1.42E-01	4.79E-02	8.22E-03
1/32	1.74E-01	5.88E-02	1.55E-02

Table 5.4. Error table of the example in Section 5.3 with $\delta = 1.1\varepsilon$, 3.1ε, $\sqrt{\varepsilon}$

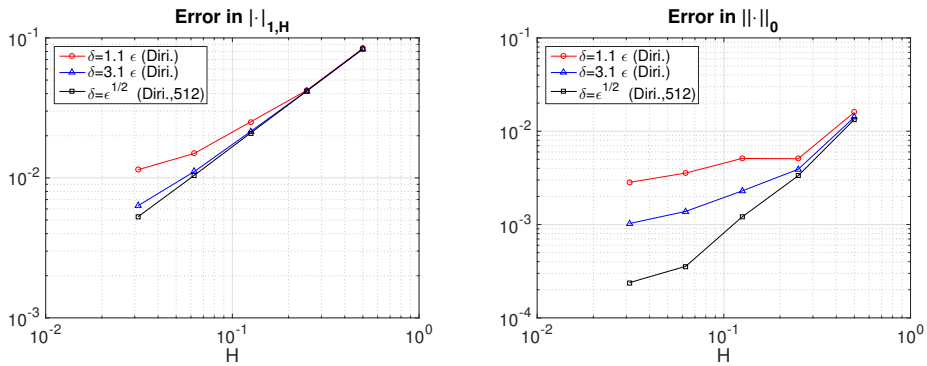

Figure 5.3. Error plots of the example in Section 5.3 with $\delta = 1.1\varepsilon$, 3.1ε, $\sqrt{\varepsilon}$

5.4 Example on mixed domain

The last example is a problem on a domain which consists of distinct coefficients. Let $\Omega = (0,1)^2$, and disjoint subdomains $\Omega_1 = \{(x_1, x_2) \in \Omega \mid x_1 > 0.5 \text{ and } x_2 < 0.5\}$ and $\Omega_2 = \Omega \setminus \Omega_1$. We consider the second-order elliptic problem with the coefficient tensor

$$\mathbf{A}^\varepsilon(\mathbf{x}) = \begin{pmatrix} 1.1 + \delta_{k,1}\sin(2\pi x_1/\varepsilon) & 0 \\ 0 & 1.1 + \delta_{k,1}\sin(2\pi x_1/\varepsilon) \end{pmatrix} \quad \text{if} \quad \mathbf{x} \in \Omega_k$$

with $\varepsilon = 10^{-3}$. Here δ_{ij} denotes the standard Kronecker delta. We impose homogeneous Neumann boundary condition on the upper and lower boundary, and Dirichlet boundary condition on the left and right boundary: value 1 on the left and 0 on the right. Any mesh used in this example consists of uniform squares. We use periodic coupling for micro problems with $\delta = \varepsilon$. By using

the associated homogenized tensor

$$\mathbf{A}^0(\mathbf{x}) = \begin{cases} \begin{pmatrix} \sqrt{0.21} & 0 \\ 0 & 1.1 \end{pmatrix} & \text{for } \mathbf{x} \in \Omega_1, \\ \begin{pmatrix} 1.1 & 0 \\ 0 & 1.1 \end{pmatrix} & \text{for } \mathbf{x} \in \Omega_2, \end{cases}$$

the reference solution u^0_{ref} on 1024×1024 mesh is obtained.

Contour plots of the solutions are drawn in Figure 5.4 for comparison. The plot on top is for the FEM solution u^ε_{ref}, and the middle plot is for the FEM solution u^0_{ref} of the homogenized problem. Both solutions are obtained on 512×512 uniform square mesh. The plot on bottom is for the FEHMM solution u_H from the macro mesh with 8×8 uniform squares, and the micro mesh with 16×16 uniform squares. The contour plots show the resemblance of the FEHMM solution to the solution of the homogenized problem as well as the solution of the original multiscale problem. Table 5.5 shows error of FEHMM solutions to the reference solution in energy norm, and in L^2-norm. We can observe the reduction of error due to decreasing H and h, but not as much as the purely periodic case.

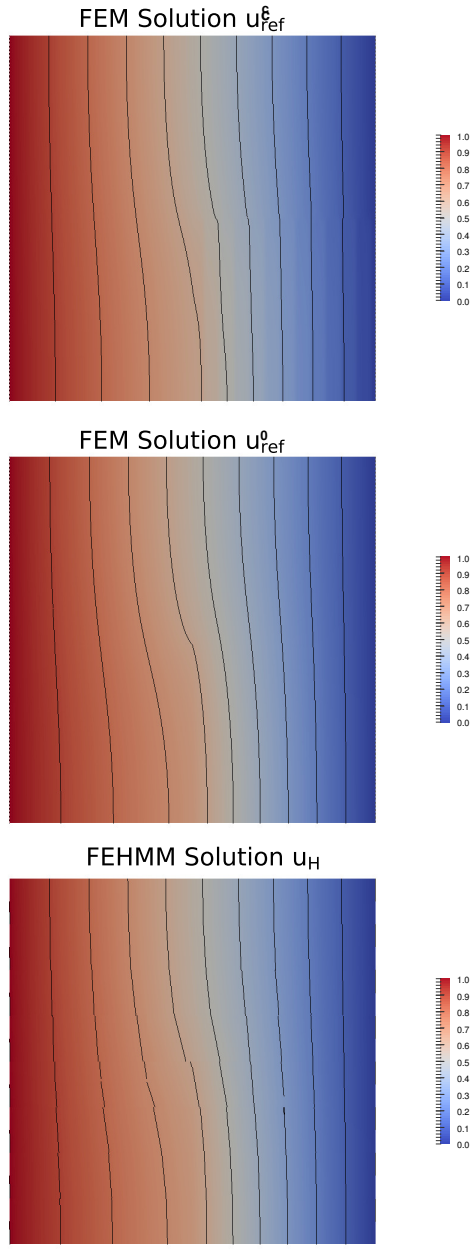

Figure 5.4. Contour plots of the solutions of the example in Section 5.4

H	h/ε=1/16	1/32	1/64
	$\|\|\|u^0_{ref} - u_H\|\|\|_H$		
1/2	9.02E-02	9.07E-02	9.09E-02
1/4	5.32E-02	5.34E-02	5.35E-02
1/8	3.07E-02	3.04E-02	3.04E-02
1/16	1.78E-02	1.69E-02	1.69E-02
1/32	1.11E-02	9.32E-03	9.21E-03
1/64	8.31E-03	5.21E-03	4.97E-03
	$\|u^0_{ref} - u_H\|_0$		
1/2	9.45E-03	9.84E-03	9.97E-03
1/4	2.86E-03	2.83E-03	2.90E-03
1/8	1.53E-03	8.31E-04	8.20E-04
1/16	1.48E-03	4.11E-04	2.32E-04
1/32	1.50E-03	3.86E-04	1.06E-04
1/64	1.58E-03	3.91E-04	9.73E-05

Table 5.5. Error of the example in Section 5.4

Bibliography

[1] A. Abdulle. The finite element heterogeneous multiscale method: a computational strategy for multiscale PDEs. *GAKUTO International Series Mathematical Sciences and Applications*, 31(EPFL-ARTICLE-182121):135–184, 2009.

[2] A. Abdulle. Discontinuous Galerkin finite element heterogeneous multiscale method for elliptic problems with multiple scales. *Mathematics of Computation*, 81(278):687–713, 2012.

[3] A. Abdulle and Y. Bai. Reduced basis finite element heterogeneous multiscale method for high-order discretizations of elliptic homogenization problems. *Journal of Computational Physics*, 231(21):7014–7036, 2012.

[4] A. Abdulle, W. E, B. Engquist, and E. Vanden-Eijnden. The heterogeneous multiscale method. *Acta Numerica*, 21:1–87, 2012.

[5] A. Abdulle and A. Nonnenmacher. A short and versatile finite element multiscale code for homogenization problems. *Computer Methods in Applied Mechanics and Engineering*, 198(37):2839–2859, 2009.

[6] A. Abdulle and A. Nonnenmacher. Adaptive finite element heterogeneous multiscale method for homogenization problems. *Computer Methods in Applied Mechanics and Engineering*, 200(37):2710–2726, 2011.

[7] R. Altmann and C. Carstensen. P_1-Nonconforming Finite Elements on Triangulations into Triangles and Quadrilaterals. *SIAM Journal on Numerical Analysis*, 50(2):418–438, 2012.

[8] O. Axelsson. *Iterative solution methods*. Cambridge University Press, 1996.

[9] P. Bochev and R. B. Lehoucq. On the finite element solution of the pure Neumann problem. *SIAM review*, 47(1):50–66, 2005.

[10] D. Braess. *Finite elements: Theory, fast solvers, and applications in solid mechanics*. Cambridge University Press, 2007.

[11] S. C. Brenner and L.-Y. Sung. Linear finite element methods for planar linear elasticity. *Mathematics of Computation*, 59(200):321–338, 1992.

[12] F. Brezzi and M. Fortin. *Mixed and hybrid finite element methods*. Springer-Verlag, 1991.

[13] Z. Cai, J. Douglas Jr., J. E. Santos, D. Sheen, and X. Ye. Nonconforming quadrilateral finite elements: a correction. *Calcolo*, 37(4):253–254, 2000.

[14] Z. Cai, J. Douglas Jr., and X. Ye. A stable nonconforming quadrilateral finite element method for the stationary Stokes and Navier-Stokes equations. *Calcolo*, 36(4):215–232, 1999.

[15] S. L. Campbell and C. D. Meyer. *Generalized inverses of linear transformations*. SIAM, 2009.

[16] C. Carstensen and J. Hu. A unifying theory of a posteriori error control for nonconforming finite element methods. *Numerische Mathematik*, 107(3):473–502, 2007.

[17] D. Cioranescu and P. Donato. An introduction to homogenization, volume 17 of Oxford Lecture Series in Mathematics and its Applications. *The Clarendon Press Oxford University Press, New York*, 4:118, 1999.

[18] M. Crouzeix and P.-A. Raviart. Conforming and nonconforming finite element methods for solving the stationary Stokes equations I. *Revue française d'automatique informatique recherche opérationnelle. Mathématique*, 7(R3):33–75, 1973.

[19] P. Degond, A. Lozinski, B. P. Muljadi, and J. Narski. Crouzeix-Raviart MsFEM with bubble functions for diffusion and advection-diffusion in perforated media. *Communications in Computational Physics*, 17(4):887–907, 2015.

[20] J. Douglas Jr., J. E. Santos, D. Sheen, and X. Ye. Nonconforming Galerkin methods based on quadrilateral elements for second order elliptic problems. *ESAIM: Mathematical Modelling and Numerical Analysis*, 33(4):747–770, 1999.

[21] W. E and B. Engquist. The heterognous multiscale methods. *Communications in Mathematical Sciences*, 1(1):87–132, 2003.

[22] W. E, P. Ming, and P. Zhang. Analysis of the heterogeneous multiscale method for elliptic homogenization problems. *Journal of the American Mathematical Society*, 18(1):121–156, 2005.

[23] Y. Efendiev, J. Galvis, and T. Y. Hou. Generalized multiscale finite element methods (GMsFEM). *Journal of Computational Physics*, 251:116–135, 2013.

[24] Y. Efendiev, J. Galvis, G. Li, and M. Presho. Generalized multiscale finite element methods: Oversampling strategies. *International Journal for Multiscale Computational Engineering*, 12(6), 2014.

[25] Y. Efendiev and T. Y. Hou. *Multiscale finite element methods: theory and applications*, volume 4. Springer Science & Business Media, 2009.

[26] Y. R. Efendiev, T. Y. Hou, and X.-H. Wu. Convergence of a nonconforming multiscale finite element method. *SIAM Journal on Numerical Analysis*, 37(3):888–910, 2000.

[27] X. Feng, I. Kim, H. Nam, and D. Sheen. Locally stabilized P_1-nonconforming quadrilateral and hexahedral finite element methods for the Stokes equations. *Journal of Computational and Applied Mathematics*, 236(5):714–727, 2011.

[28] X. Feng, R. Li, Y. He, and D. Liu. P_1-Nonconforming quadrilateral finite volume methods for the semilinear elliptic equations. *Journal of Scientific Computing*, 52(3):519–545, 2012.

[29] V. Girault and P.-A. Raviart. *Finite element methods for Navier-Stokes equations: theory and algorithms*. Springer-Verlag, 1986.

[30] T. Y. Hou and X.-H. Wu. A multiscale finite element method for elliptic problems in composite materials and porous media. *Journal of computational physics*, 134(1):169–189, 1997.

[31] I. C. Ipsen and C. D. Meyer. The idea behind Krylov methods. *American Mathematical Monthly*, pages 889–899, 1998.

[32] V. V. Jikov, S. M. Kozlov, and O. A. Oleinik. *Homogenization of differential operators and integral functionals*. Springer Science & Business Media, 2012.

[33] L. Ju and J. Burkardt. MGMRES: Restarted GMRES solver for sparse linear systems. `http://people.sc.fsu.edu/~jburkardt/f_src/mgmres/mgmres.html`. [Online; revision on 28-Aug-2012].

[34] E. F. Kaasschieter. Preconditioned conjugate gradients for solving singular systems. *Journal of Computational and Applied mathematics*, 24(1-2):265–275, 1988.

[35] S. Kim, J. Yim, and D. Sheen. Stable cheapest nonconforming finite elements for the Stokes equations. *Journal of Computational and Applied Mathematics*, 299:2–14, 2016.

[36] C. Le Bris, F. Legoll, and A. Lozinski. MsFEM à la Crouzeix-Raviart for highly oscillatory elliptic problems. In *Partial Differential Equations: Theory, Control and Approximation*, pages 265–294. Springer, 2014.

[37] C. Le Bris, F. Legoll, and A. Lozinski. An MsFEM type approach for perforated domains. *Multiscale Modeling & Simulation*, 12(3):1046–1077, 2014.

[38] C.-O. Lee, J. Lee, and D. Sheen. A locking-free nonconforming finite element method for planar linear elasticity. *Advances in Computational Mathematics*, 19(1-3):277–291, 2003.

[39] C. S. Lee and D. Sheen. Nonconforming generalized multiscale finite element methods. *Journal of Computational and Applied Mathematics*, 311:215–229, 2017.

[40] R. Lim and D. Sheen. Nonconforming finite element method applied to the driven cavity problem. *Communications in Computational Physics*, 21(4):1012–1038, 2017.

[41] D. S. Malkus. Eigenproblems associated with the discrete LBB condition for incompressible finite elements. *International Journal of Engineering Science*, 19(10):1299–1310, 1981.

[42] H. Nam, H. J. Choi, C. Park, and D. Sheen. A cheapest nonconforming rectangular finite element for the stationary Stokes problem. *Computer Methods in Applied Mechanics and Engineering*, 257:77–86, 2013.

[43] C. Park. *A study on locking phenomena in finite element methods*. PhD thesis, Department of Mathematics, Seoul National University, Seoul, Korea, 2002.

[44] C. Park and D. Sheen. P_1-nonconforming quadrilateral finite element methods for second-order elliptic problems. *SIAM Journal on Numerical Analysis*, 41(2):624–640, 2003.

[45] J. Qin. *On the convergence of some low order mixed finite elements for incompressible fluids*. PhD thesis, Pennsylvania State University, Pennsylvania, 1994.

[46] R. Rannacher and S. Turek. Simple nonconforming quadrilateral Stokes element. *Numerical Methods for Partial Differential Equations*, 8(2):97–111, 1992.

[47] D. Shi and L. Pei. Low order Crouzeix-Raviart type nonconforming finite element methods for approximating Maxwell's equations. *Int. J. Numer. Anal. Model*, 5(3):373–385, 2008.

[48] G. Strang. Variational crimes in the finite element method. In *The mathematical foundations of the finite element method with applications to partial differential equations*, pages 689–710. Elsevier, 1972.

[49] G. Strang and G. J. Fix. *An analysis of the finite element method*, volume 212. Prentice-hall Englewood Cliffs, NJ, 1973.

[50] N. Zhang, T.-T. Lu, and Y. Wei. Semi-convergence analysis of Uzawa methods for singular saddle point problems. *Journal of Computational and Applied Mathematics*, 255:334–345, 2014.

국문초록

본 학위논문의 제1부에서는 주기경계조건을 갖는 P_1–비순응유한요소공간을 고려하고, 그것과 이산 라플라스 연산자의 특성에 대해 조사한다. 최소의 필수 이산경계조건이라는 개념의 도움을 받아 유한요소공간들의 차원을 해석한다. 이 해석에 기반하여, 주기경계조건을 갖는 유한공간의 기저함수들을 두 가지 종류로 분류한다. 그리고 이차 타원형 문제를 풀기 위한 크릴로프 반복법 몇 가지를 소개하고 그 해들을 비교한다. 그중 몇몇의 방법은 일반화된 역작용소의 하나인 Drazin 역에 기반하는데, 이는 주기적 성질이 특이 선형연립방정식을 유도할 수 있기 때문이다. 주기경계조건을 갖는 스토크스 방정식으로의 응용을 다룬다. 마지막으로 타원형 문제에 대한 결과들을 3차원 경우로 확장한다. 이러한 논의에 수치적 결과들을 보여준다.

제2부에서는 멀티스케일 문제를 위한 비순응 이종 멀티스케일 방법을 소개한다. 이에 대한 공식화는 P_1–비순응유한요소에 기반을 두고 있는데, 대개는 주기경계조건을 갖는다. 이종 멀티스케일 유한요소법의 일반적인 구성을 따라서, 제안된 방법의 사전 추정오차를 분석한다. 수치적인 구현을 위해서, 우리는 앞선 제1부에서 특이 선형연립방정식을 위해 제안된 반복법 중 하나를 사용한다. 수치적 예제와 결과를 보인다.

주요어 : P_1–비순응유한요소, 주기경계조건, 최소의 필수이산경계조건, 특이 선형연립방정식, Drazin 역, 이종 멀티스케일 방법, 수치적 균질화

학번 : 2012-20414