A NEW DISCONTINUOUS FINITE ELEMENT METHOD BASED ON A LEAST-SQUARES STABILIZATION FOR ELLIPTIC AND CONVECTION-DIFFUSION PROBLEMS

by

Yongjun Yang

A thesis submitted to the Faculty and the Board of Trustees of the Colorado School of Mines in partial fulfillment of the requirements for the degree of Doctor of Philosophy (Mathematical and Computer Sciences).

Golden, Colorado
Date 06-11-2003

Signed: Yongjun Yang
Yongjun Yang

Approved:
Dr. Junping Wang
Thesis Advisor

Golden, Colorado
Date 6/11/03

Graeme Fairweather
Dr. Graeme Fairweather
Professor and Head
Department of Mathematical and
Computer Sciences

ABSTRACT

We are concerned with numerical methods for partial differential equations by finite element methods. The finite element method (FEM) is an efficient and widely used numerical technique for complex systems governed by a system of partial differential equations. The FEM consists of two basic steps: *derivation of variational formulations* and *construction of finite element approximating functions.* Although the quality of a finite element scheme is equally affected by each of the two steps, the derivation of variational formulations often plays a decisive role in the discovery of a new and innovative FEM for the complex system under consideration.

A new finite element method is proposed and analyzed for second order elliptic equations using discontinuous piecewise polynomials on a finite element partition consisting of general polygons. The new method is based on a stabilization of the well-known primal hybrid formulation by using some least-squares forms imposed on the boundary of each element. Two finite element schemes are presented. The first one is a non-symmetric formulation and is absolutely stable in the sense that no parameter selection is necessary for the scheme to converge. The second one is a symmetric formulation, but is conditionally stable in that a parameter has to be selected in order to have an optimal order of convergence. Optimal-order error estimates in some H^1-equivalence norms are established for the proposed discontinuous finite element methods. For the symmetric formulation, an optimal-order error estimate is also derived in the L^2 norm. The new method features a finite element partition consisting of general polygons as opposed to triangles or quadrilaterals in the standard finite element Galerkin method.

Our method using discontinuous finite elements with stabilization is applied to convection dominated convection-diffusion problems. In general, the standard Galerkin finite element methods applied to such problems exhibit a variety of deficiencies, including high oscillations and poor approximation of the derivatives of the solutions. A new stabilization technique, which features a non-symmetric formulation using discontinuous piecewise polynomials, is presented and analyzed for such problems. Error estimates in some H^1-equivalence norm is established for the proposed discontinuous finite element methods.

The construction of stiffness matrices of the finite element schemes is presented. For the symmetric formulation, a system of linear equations with symmetric and positive definite coefficient matrix is derived. Implementation of the finite element schemes is carried out. An numerical solver is developed using C++, and tested on some examples. The resulting numerical solutions have shown the desired accuracy and properties of the true solutions.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I give my sincere gratitude to my advisor, Dr. Junping Wang, for his persistent and thorough guidance. It was his vision that led me into this research problem which has showed me more intrinsic beauty of mathematics. As my mentor, he motivates and supports me wholeheartedly.

Secondly, my sincere thanks go to my committee members, Dr. Bernard Bialecki, Dr. Graham Davis, Dr. Graeme Fairweather, Dr. Paul Martin, and Dr. Ruichong Zhang for their supervision and support during my graduate study. I would like to express my appreciation to Dr. Fairweather, our department head, for his help and encouragement.

Last but certainly not the least, my special thanks go to my wife, Xiaohan, and my son, Eric "beibei", for their love, faith, and support. I could not have done this without them.

To my beloved father

Chapter 1

INTRODUCTION

1.1 Problem statement and main approach

The finite element method (FEM) is an efficient and widely used numerical technique for complex systems governed by a system of partial differential equations. The FEM consists of two basic steps: *derivation of variational formulations* and *construction of finite element approximating functions.* Although the quality of a finite element scheme is equally affected by each of the two steps, the derivation of variational formulations often plays a decisive role in the discovery of a new and innovative FEM for the complex system under consideration.

Many complex systems in practice involve systems of partial differential equations for which the solutions either are discontinuous or have a sharp-changing front. The standard continuous Galerkin FEM may fail to provide any physically-meaningful numerical solution for such systems. As a result, various discontinuous Galerkin methods [2, 4, 5, 16, 17, 26, 40] have evolved over the last several years to tackle physical problems with non-smooth solutions.

The discontinuous Galerkin methods use discontinuous piecewise polynomials on finite element partitions to approximate the solutions. They are often derived from testing the governing equations locally on each element. To illustrate this, we first consider a model second order elliptic problem which seeks $u = u(x)$ satisfying

$$
\begin{aligned}
-\nabla \cdot (\kappa \nabla u) &= f \qquad \text{in } \Omega, && (1.1)\\
u &= g \qquad \text{on } \partial\Omega, && (1.2)
\end{aligned}
$$

where $\kappa = (\kappa_{ij})_{d\times d}$ is symmetric and uniformly positive definite over an open bounded domain $\Omega \subseteq \mathbb{R}^d (d = 2, 3)$, $f \in L^2(\Omega)$ is given, and $g \in H^{\frac{1}{2}}(\partial\Omega)$ is the given boundary data.

Let $\mathcal{T}_h = \{K\}$ be a quasi-uniform partition of Ω with non-overlapping elements K where h is the mesh-size of the partition $\mathcal{T}_h$. To obtain a weak formulation of (1.1)-(1.2), we introduce two function spaces,

$$
\begin{aligned}
\mathcal{V} &= \left\{v : v|_K \in H^1(K), \forall K \in \mathcal{T}_h\right\}, && (1.3)\\
\mathcal{M} &= \left\{\mu : \mu|_{\partial K} \in H^{-\frac{1}{2}}(\partial K), \forall K \in \mathcal{T}_h, \exists \mathbf{q} \in H(div;\Omega), \mathbf{q} \cdot n_K = \mu|_{\partial K}\right\}, && (1.4)
\end{aligned}
$$

where n_K is the outward normal direction on ∂K. $H(div;\Omega)$ is defined as

$$
H(div;\Omega) = \left\{v \in (L^2(\Omega))^d; \nabla \cdot v \in L^2(\Omega)\right\}.
$$

It can be verified (see [15] Ch. 7) that there exists a unique pair $(\bar{u}, \lambda) \in \mathcal{V} \times \mathcal{M}$ such that

$$
\sum_{K\in\mathcal{T}_h} \int_K \kappa \nabla \bar{u} \nabla v \, dK - \sum_{K\in\mathcal{T}_h} \int_{\partial K} \lambda v \, ds = \sum_{K\in\mathcal{T}_h} \int_K f v \, dK, \quad \forall v \in \mathcal{V}, \qquad (1.5)
$$

$$
\sum_{K\in\mathcal{T}_h} \int_{\partial K} \bar{u}\mu \, ds = \sum_{K\in\mathcal{T}_h} \int_{\partial K\cap\partial\Omega} g\mu \, ds, \quad \forall \mu \in \mathcal{M}. \qquad (1.6)
$$

It is not hard to show that

$$\bar{u} = u, \quad \text{and} \quad \lambda|_{\partial K} = \kappa \nabla u \cdot n_K, \ \forall K \in \mathcal{T}_h, \tag{1.7}$$

where u is the solution of (1.1)-(1.2).

The variational form in (1.5)-(1.6) gives a natural mixed finite element formulation. To simplify the notation, let

$$\begin{aligned}
a(u,v) &= \sum_{K\in\mathcal{T}_h} \int_K \kappa \nabla u \nabla v \, dK, && u, v \in \mathcal{V}, \\
b(v,\lambda) &= \sum_{K\in\mathcal{T}_h} \int_{\partial K} \lambda v \, ds, && v \in \mathcal{V}, \lambda \in \mathcal{M}, \\
f(v) &= \sum_{K\in\mathcal{T}_h} \int_K f v \, dK, && v \in \mathcal{V}, \\
g(\mu) &= \sum_{K\in\mathcal{T}_h} \int_{\partial K \cap \partial\Omega} g\mu \, ds && \mu \in \mathcal{M}
\end{aligned} \tag{1.8}$$

be four bilinear/linear forms. Then (1.5)-(1.6) can be rewritten as the following mixed form:

$$\begin{aligned}
a(u,v) - b(v,\lambda) &= f(v), && \forall v \in \mathcal{V}, \\
b(u,\mu) &= g(\mu), && \forall \mu \in \mathcal{M}.
\end{aligned} \tag{1.9}$$

The *primal hybrid finite element method* is a Galerkin procedure based on the variational form (1.9). In this method, the unknown u is approximated by discontinuous finite elements over the partition $\mathcal{T}_h$, and the normal component of the flux $\kappa\nabla u$ is approximated on the element boundaries by a different set of discontinuous finite elements. Due to the saddle-point property of the formulation (1.9), the two discontinuous finite element spaces must be constructed so that the inf-sup condition of Brezzi [12] and Babuška [2] is satisfied. The requirement of the inf-sup condition imposes two difficulties. The first is that the two finite element spaces are not naturally correlated and thus are hard to construct. The second difficulty involves the saddle-point nature of the resulting matrix problem. Solving a matrix problem with a strong saddle-point property is a difficult task from a computational point of view. Details of this approach can be found in [33, 34]. In fact, Raviart and Thomas [34] showed the following results in the two-dimensional case when the model problem (1.1)-(1.2) has a homogeneous Dirichlet boundary condition. The finite element spaces are defined as:

$$\begin{aligned} \mathcal{V}_h = &\ \{v \in \mathcal{V} : v|_K \in P_k(K), \forall K \in \mathcal{T}_h\}, \\ \mathcal{M}_h = &\ \{\mu \in \mathcal{M} : \mu|_{e,K} \in P_l(e), \forall e \subset \partial K, K \in \mathcal{T}_h, \\ &\ \text{and } \mu|_{e,K_1} + \mu|_{e,K_2} = 0, \text{if } e = K_1 \cap K_2, \forall K_1, K_2 \in \mathcal{T}_h\}, \end{aligned} \tag{1.10}$$

where e denotes a boundary edge of element K. $k \geq 1$ and $l \geq 0$ are integers, and $P_k(K)$ and $P_l(e)$ stand for the spaces of polynomials of degree no more than k and l on the element K and the boundary edge e, respectively. Let (u_h, λ_h) be the finite element approximation satisfying

$$\begin{aligned} a(u_h, v) - b(v, \lambda_h) &= f(v), \qquad && \forall v \in \mathcal{V}_h, \\ b(u_h, \mu) &= 0, && \forall \mu \in \mathcal{M}_h. \end{aligned} \tag{1.11}$$

Then, one has the following estimate once Brezzi's inf-sup condition [12] is satisfied:

$$\left(\sum_{K \in \mathcal{T}_h} \|u - u_h\|^2_{H^1(K)} + h \sum_{K \in \mathcal{T}_h} \|\lambda - \lambda_h\|^2_{L^2(\partial K)} \right)^{1/2} \leq C h^s \|u\|_{s+1,\Omega}, \tag{1.12}$$

with $s = \min(k, l+1)$. In particular, this happens when $k \geq l+1, l$ even, or $k \geq l+2, l$ odd.

Notice that in the spaces $\mathcal{V}$ and $\mathcal{V}_h$, no continuity on u or u_h is required *a priori.*

However, the exact solution u has to be in $H_0^1(\Omega)$. The finite element approximations can be discontinuous along ∂K. Since the finite element space is not contained in $H^1(\Omega)$, this method is regarded as *non-conforming.*

One of the main objective of this thesis is to present and analyze a new and innovative discontinuous finite element scheme by using the stabilization method. The method of stabilization [18, 20, 25] is a systematic procedure that provides new and stable variational formulations by using various least-squares forms associated with the governing equation.

Our stabilized discontinuous finite element method will be demonstrated first for the second order elliptic problem (1.1)-(1.2). As a matter of fact, we shall show that the saddle-point problem (1.9) can be stabilized by using the following least-squares term:

$$\alpha h \sum_{K\in\mathcal{T}_h} \int_{\partial K} (\lambda - \kappa\nabla u\cdot n_K)(\mu - \kappa\nabla v\cdot n_K)\, ds, \tag{1.13}$$

where $\alpha > 0$ is an arbitrary, but fixed, real number. By adding this term into the finite element formulation, we can show that the resulting finite element scheme has a unique solution. Two finite element schemes will be presented. The first is a non-symmetric formulation and is absolutely stable in the sense that no parameter selection is necessary for the scheme. The second is a symmetric formulation, but is conditionally stable in that a parameter has to be chosen in order to produce a convergent scheme. Both schemes preserve the mass conservation property locally on each element. We also introduce stabilized discontinuous finite element formulations with the least-squares term (1.13) as well as the following jump term:

$$\beta h^{-1} \sum_{K} \int_{\partial K} [\![u]\!][\![v]\!]\, ds, \tag{1.14}$$

where $[\![\cdot]\!]$ is the notation for the jump across an edge e that will be defined later, and $\beta > 0$ is an arbitrary, but fixed, real number.

Optimal-order error estimates in some H^1-equivalence norms are established for the proposed discontinuous finite element methods. For the symmetric formulation, an optimal-order error estimate is also derived in the L^2 norm. The new method features a finite element partition consisting of general polygons as opposed to triangles or quadrilaterals in the standard finite element Galerkin method.

1.2 The convection-diffusion problem

Another objective of the thesis is to consider efficient numerical solution of the convection-diffusion problem

$$\begin{aligned} -\nabla \cdot (a\nabla u - \mathbf{b}u) + cu &= f \qquad \text{in } \Omega, \\ u &= g \qquad \text{on } \partial\Omega. \end{aligned} \tag{1.15}$$

If the coefficient of the diffusion term is small, the problem is said to be convection-dominated. Such convection-diffusion problem has a lot of applications in practice. Many physical processes, in particular those arising from fluid flow problems, can be modeled by convection-dominated convection diffusion problems. For instance, analysis and simulation of oil and gas reservoir, ground water transportation, and weather modeling, require good and accurate solutions of such problem. But it has been well known that standard numerical methods often fail to work well because of the sharp-changing front and/or discontinuity of the true solution. In this section we give a brief review of previous work in the area. Extensive discussions are given in the books of [30] and [35].

The *Finite Difference Methods* approximate the solution of the strong form of the equation by replacing the derivatives of the unknown by some approximation involving values of the unknown at some node points. An equation is generated at each node and then the solution is found by solving a linear system of these equations. The classical central difference scheme is to replace the derivatives by the central difference formulas. It is well known that when the mesh size of the partition is greater than the so-called Péclet number, which is almost always the case for computational reasons, the scheme produces wild and non-physical oscillations. The *Upwind Methods* employ a common technique to overcome the numerical instability by taking a one-sided approximation of the convection term in the upstream direction. Analysis [37] shows that the resulting scheme is stable independent of the mesh size. However, one order of accuracy is lost comparing to the central difference scheme. The *Artificial Diffusion Method* is an approach to solve a modified version of the equation in which the diffusion coefficient is replaced by some term of order h.

The *Finite Element Methods* aim to find the best approximations in certain finite element spaces, defined over finite element partitions. Finite element methods have been a very efficient and widely used numerical technique for solving systems governed by partial differential equations. The standard *Galerkin finite element method* looks for finite element approximations of piecewise polynomials satisfying the weak forms of the equations. It is well-known that when convection-diffusion problems are discretized using the standard Galerkin finite element method, non-physical oscillations can occur in the discrete solution whenever convection is the dominating term. To remedy this convective instability, different approaches have been developed. The method of *artificial diffusion* modifies the Galerkin finite element scheme, where the diffusion parameter is replaced by some term of order h. This method, which results in extra diffusion that smears out the sharp fronts in the solution, is at best first order accurate due to this order h perturbation. The *Streamline Diffusion Methods* is an extension of the artificial diffusion idea. Extra diffusion is added only in the streamline direction, and therefore introduces less crosswind diffusion. We refer to [24] and [27] for more details. The *generalized hierarchical basis multigrid methods* deals with the convection-diffusion problem in a multiscale approach. The idea is to construct solvers based on special (and problem-dependent) hierarchical multiscale decompositions of the trial and test function spaces. This approach [3, 22, 32, 38, 39] can give robust yet efficient solver to the convection-diffusion problems.

The *Finite Volume Methods* is a technique to solve the equation in conservation (i.e. integral) form. The domain is divided into subdomains and the integral form of the equation is posed on each of the subdomain. Then the volume integrals in these equations are converted to surfaces integrals by the Gauss Theorem. Finite volume methods have been very successful in the numerical solution of partial differential solutions and is highly suitable for diffusion problems. One class of finite volume methods, the cell-vertex methods have been very useful for the convection-diffusion problems. However, using such methods encounters difficulties such as 'counting' problems. For more details, we refer to [29].

The *Transient Methods* is a general approach to the solution of steady state problems by solving a spatially discretized transient equation (i.e., with a u_t term). The accuracy of the time stepping is not too important long as the convergence to the steady state solution is achieved. Methods that do not perform well on steady state problems can be treated in this way. However, it is obvious the cost will be much greater if a transient method is used on a steady state problem.

Our approach here is to apply the idea of stabilized discontinuous finite element method to the convection-diffusion problem. This is a non-conforming finite element scheme which shares the idea of upwinding. A new stabilization technique, which features a non-symmetric formulation using discontinuous piecewise polynomials, is presented and analyzed for such problems. Error estimates in some H^1-equivalence norm are established for the proposed discontinuous finite element methods.

Discussions on the matrix problems and results of numerical computations are presented. The construction of stiffness matrices of the finite element schemes is proposed. For the symmetric formulation, a system of linear equations with a symmetric and positive definite coefficient matrix is derived. Implementation of the finite element schemes is carried out and is tested on some numerical examples. Some numerical solutions are compared with true solutions to show the desired accuracy of our finite element schemes.

1.3 Thesis outline

We outline the thesis as follows. In Chapter 2, we give some preliminaries in Finite Element Methods. In Chapter 3, we introduce symmetric and non-symmetric formulations for elliptic problems. We also establish the error estimates for some H^1-equivalence norms as well as the usual L^2 norm for the symmetric formulation. In Chapter 4, we present the application of stabilized discontinuous finite element method to convection-diffusion problems. A non-symmetric formulation is proposed and analyzed, and an H^1-equivalence norm error estimate is derived. In Chapter 5, we discuss the matrix problems and present the numerical technique that we use to solve the resulting systems of linear equations. In Chapter 6, several numerical examples are given. Numerical solutions with high accuracy will be illustrated. We summarize and give some future research directions in Chapter 7.

Chapter 2

PRELIMINARIES IN FINITE ELEMENTS

In this chapter, we give some preliminaries in finite element methods, especially in applications to second-order elliptic problems. Some useful inequalities are also presented.

2.1 Interpolation theory in Sobolev spaces

The Sobolev space $W^{m,p}(\Omega)$, where integer $m \geq 0$ and $1 \leq p \leq \infty$, is defined as

$$W^{m,p}(\Omega) = \{v \in L^p(\Omega);\ \ \partial^\alpha v \in L^p(\Omega),\ \forall |\alpha| \leq m\}\,. \tag{2.1}$$

It is a Banach space, equipped with norm

$$\begin{aligned} ||v||_{m,p,\Omega} &= \left(\sum_{|\alpha|\leq m} \int_\Omega |\partial^\alpha v|^p \, da \right)^{1/p}, & 1 \leq p < \infty, \\ ||v||_{m,\infty,\Omega} &= \max_{|\alpha|\leq m} \{\text{esssup} |\partial^\alpha v|\}\,, & p = \infty. \end{aligned} \tag{2.2}$$

We also use the semi-norms

$$\begin{aligned} |v|_{m,p,\Omega} &= \left(\sum_{|\alpha|= m} \int_\Omega |\partial^\alpha v|^p \, da \right)^{1/p}, & 1 \leq p < \infty, \\ |v|_{m,\infty,\Omega} &= \max_{|\alpha|= m} \{\text{esssup} |\partial^\alpha v|\}\,, & p = \infty. \end{aligned} \tag{2.3}$$

Notice that $H^m(\Omega) = W^{m,2}(\Omega)$. For $H^m(\Omega)$, $\|\cdot\|_{m,\Omega} = \|\cdot\|_{m,p,\Omega}$, and $|\cdot|_{m,\Omega} = |\cdot|_{m,p,\Omega}$. For more details on the Sobolev spaces, we refer to [1].

Now we give a general definition of a finite element. A *finite element* in R^d is a triple (K, P, Σ) where:

- K is a closed subset of R^n with nonempty interior and a Lipschitz-continuous boundary.
- P is a space of real valued functions defined over K.
- Σ is a finite set of linearly independent linear forms defined over the space P.

Two finite elements $(\hat{K}, \hat{P}, \hat{\Sigma})$ and (K, P, Σ) are said to be *affine equivalent* if there exists an invertible affine mapping:

$$F : \hat{x} \in R^n \rightarrow F(\hat{x}) = B\hat{x} + b \in R^n,$$

such that the following relations hold:

$$\begin{aligned}
&K = F(\hat{K}),\\
&P = \{p : K \rightarrow R;\ \ p = \hat{p} \cdot F^{-1}, \hat{p} \in \hat{P}\},\\
&a_i^r = F(\hat{a}_i^r),\ \ r = 0, 1, 2,\\
&\xi_{ik}^1 = B\hat{\xi}_{ik}^1,\ \ \xi_{ik}^2 = B\hat{\xi}_{ik}^2,\ \ \xi_{il}^2 = B\hat{\xi}_{il}^2,
\end{aligned} \tag{2.4}$$

whenever the nodes a_i^r, $\hat{a}_i^r$ and vectors ξ_{ik}^1, ξ_{ik}^2, ξ_{il}^2, and $\hat{\xi}_{ik}^1$, $\hat{\xi}_{ik}^2$, $\hat{\xi}_{il}^2$ occur in the definition of Σ and $\hat{\Sigma}$.

We now proceed to the most important result in this section. It can be found in [15] (Theorem 3.1.5).

Theorem 2.1 *Let $(\hat{K}, \hat{P}, \hat{\Sigma})$ be a finite element, for which s denotes the highest order of the partial derivatives occurring in the definition of $\hat{\Sigma}$.*

$$\begin{aligned} &W^{k+1,p}(\hat{K}) \hookrightarrow C^s(\hat{K}), \\ &W^{k+1,p}(\hat{K}) \hookrightarrow W^{m,q}(\hat{K}), \\ &P_k(\hat{K}) \subset \hat{P} \subset W^{m,q}(\hat{K}), \end{aligned} \tag{2.5}$$

hold for some integers $m \geq 0$ and $k \geq 0$ and for some numbers $p, q \in [1, \infty]$. Then there exists a constant $C(\hat{K}, \hat{P}, \hat{\Sigma})$ such that, for all affine-equivalent finite elements (K, P, Σ), and all functions $v \in W^{k+1,p}(K)$,

$$|v - \Pi_K v|_{m,q,K} \leq C(\hat{K}, \hat{P}, \hat{\Sigma})(\text{meas}(K))^{1/q-1/p} \frac{h_K^{k+1}}{\rho_K^m} |v|_{k+1,p,K}, \tag{2.6}$$

where $\Pi_k v$ denotes the interpolation of a function v in P, and

$$\begin{aligned} &\text{meas}(K) = \textit{measure of } K, \\ &h_K = \textit{diam}(K), \\ &\rho_K = \sup\{\textit{diam}(S); S \textit{ is a ball contained in } K\}. \end{aligned}$$

We say that a family of finite elements (K, P_K, Σ_K), where K is viewed as the parameter of the family, is *shape-regular* if there exists a constant $\sigma > 0$ such that for all K, $h_K/\rho_K \leq \sigma$. Moreover, the partition $\mathcal{T}_h$ is said to be *quasi-uniform* if it is shape-regular and there is a constant $C > 0$ such that

$$h \leq C h_K, \qquad \forall K \in \mathcal{T}_h.$$

We shall assume that our partition is quasi-uniform in the remainder of this thesis.

For such families, the interpolation error estimates in Theorem 2.1 can be given as in the following theorem.

Theorem 2.2 *Assume that the reference finite element* $(\hat{K}, \hat{P}, \hat{\Sigma})$ *of a shape-regular affine family of finite elements* (K, P_K, Σ_K) *satisfies (2.5). Then there exists a constant* $C(\hat{K}, \hat{P}, \hat{\Sigma})$ *such that, for all elements* K*, and all functions* $v \in W^{k+1,p}(K)$,

$$||v - \Pi_K v||_{m,q,K} \leq C(\hat{K}, \hat{P}, \hat{\Sigma})(\text{meas}(K))^{1/q-1/p} h_K^{k+1-m} |v|_{k+1,p,K}. \tag{2.7}$$

In the special case in which $p = q = 2$ and $m = 0$, we obtain

$$||v - \Pi_K v||_{m,K} \leq C h_k^{k+1} |v|_{k+1,K}. \tag{2.8}$$

Proofs of Theorem 2.1 and 2.2 can be found in [15]. For more details of the material in this section, we refer to [8], [9], and [10].

2.2 Application to second order elliptic problems

Now let us consider the following abstract linear variational problem arising from the weak formulation of some second order elliptic problems (e.g. (1.1)-(1.2)) that seeks $u \in V$ satisfying

$$a(u, v) = f(v), \quad \forall v \in V, \tag{2.9}$$

where V is a Hilbert space, a is a continuous V-elliptic bilinear form on $V \times V$, and f is a linear form on V. On the finite element space $V_h \subset V$, the *discrete solution* $u_h \in V_h$ is an approximation of u that satisfies

$$a(u_h, v_h) = f(v_h), \quad \forall v_h \in V_h. \tag{2.10}$$

We denote by $||\cdot||$ the norm on V. We have the following basic error estimate, which is due to Céa [14].

Theorem 2.3 *(Céa's lemma) There exists a constant C independent of V_h such that*

$$||u - u_h|| \leq C \inf_{v_h \in V_h} ||u - v_h||. \tag{2.11}$$

Proof: It follows from (2.9) and (2.10) that $a(u - u_h, w_h) = 0$ for all $w_h \in V_h$. For any $v_h \in V_h$, set $w_h = u_h - v_h$. Therefore,

$$a(u - u_h, u - u_h) = a(u - u_h, u - u_h) + a(u - u_h, u_h - v_h) = a(u - u_h, u - v_h).$$

By the fact that $a(\cdot,\cdot)$ is V-elliptic and continuous, one can easily see that there exist some constants α and M such that

$$\begin{aligned} \alpha||u - u_h||^2 &\leq a(u - u_h, u - u_h) = a(u - u_h, u - v_h) \\ &\leq M||u - u_h||\,||u - v_h||. \end{aligned}$$

Then the theorem follows with $C = M/\alpha$. □

The following theorem establishes the estimate of $||u - u_h||_{1,\Omega}$.

Theorem 2.4 *Assume that for a shape-regular affine family of finite elements, there exists an integer $k \geq 1$ such that $P_k(\hat{K}) \subset \hat{P} \subset H^1(\hat{K})$ and $H^{k+1}(\hat{K}) \subset \mathcal{C}^0(\hat{K})$, where s is the maximal order of partial derivatives occurring in the definition of set $\hat{\Sigma}$. Then if the solution u of the variational problem is also in $H^{k+1}(\Omega)$, there exists a constant*

C independent of h such that

$$\|u - u_h\|_{1,\Omega} \leq Ch^k |u|_{k+1,\Omega}, \tag{2.12}$$

where u_h is the corresponding discrete solution.

Proof: First we set $p = q = 2$ and $m = 1$ in (2.8) with $v = u$. Then in addition we use Céa's lemma (Theorem 2.3), which yields

$$\begin{aligned} \|u - u_h\|_{1,\Omega} &\leq C \inf_{v_h \in V_h} \|u - v_h\|_{1,\Omega} \leq C\|u - \Pi_h u\|_{1,\Omega} \\ &\leq Ch^k |u|_{k+1,\Omega}. \end{aligned}$$

□

2.3 Useful inequalities

The first inequality that we introduce is the inverse inequality.

Theorem 2.5 *Assume that the reference finite element $(\hat{K}, \hat{P}, \hat{\Sigma})$ of a quasi-uniform affine family of finite elements (K, P_K, Σ_K) in R^d satisfies (2.5). Let (l, r) and (m, q) be two pairs with $l, m \leq 0$ and $(r, q) \in [1, \infty]$ such that*

$$l \leq m \text{ and } \hat{P} \subset W^{l,r}(\hat{K}) \cup W^{m,q}(\hat{K}).$$

Then there exists a constant $C = C(\sigma, \nu, l, r, m, q)$ such that for all v_h belonging to the finite element space,

$$\left(\sum_K |v_h|^q_{m,q,k} \right)^{1/q} \leq \frac{C}{h^{d \max\{0, 1/r - 1/q\} + m - 1}} \left(\sum_K |v_h|^r_{l,r,K} \right)^{1/r} \tag{2.13}$$

if $p, q < \infty$, with

$$\max_K |v_h|_{m,\infty,K} \text{ in lieu of } \left(\sum_K |v_h|^q_{m,q,k}\right)^{1/q} \quad \text{if } q = \infty$$
$$\max_K |v_h|_{l,\infty,K} \text{ in lieu of } \left(\sum_K |v_h|^q_{l,r,k}\right)^{1/r} \quad \text{if } r = \infty.$$

In the special case in which $r = q = 2$, $m = 1$ and $l = 0$, one has that for all v_h in the finite element space,

$$||v_h||_{1,\Omega} \leq Ch^{-1}|v_h|_{0,\Omega}. \tag{2.14}$$

The inverse inequality can be found in many places. Here we refer to [11] and [15].

Next we give a useful trace inequality in Sobolev spaces. For an H^1 function, the inequality bounds its L^2 norm on the boundary of an element by some norms inside the element.

Lemma 2.1 *If the triangulation $\mathcal{T}_h$ is shape-regular, we have*

$$||\psi||^2_{0,\partial K} \leq Ch^{-1}||\psi||^2_{0,K} + Ch||\nabla\psi||^2_{0,K}, \quad \forall\psi \in H^1(K), \tag{2.15}$$

where $C > 0$ is a constant independent of h. Moreover, for $\psi \in P^k(K)$,

$$||\psi||^2_{0,\partial K} \leq Ch^{-1}||\psi||^2_{0,K}, \tag{2.16}$$

with again constant C independent of h.

Proof: Let $A = (\alpha_i)$ be an arbitrary point in K and $B = (\beta_i)$ be an arbitrary point on ∂K. By the Fundamental Theorem of Calculus,

$$\psi(A) - \psi(B) = \sum_{i=1}^{d} \int_{\alpha_i}^{\beta_i} \frac{\partial \psi}{\partial x_i}(\beta_1, \cdots, \beta_{i-1}, s, \alpha_{i+1}, \cdots, \alpha_d)\, ds.$$

Squaring both sides, and using the Cauchy-Schwarz inequality,

$$\begin{aligned}\psi^2(A) + \psi^2(B) - 2\psi(A)\psi(B) = & \left(\sum_{i=1}^{d} \int_{\alpha_i}^{\beta_i} \frac{\partial \psi}{\partial x_i}\, ds \right)^2 \\ & \leq \sum_{i=1}^{d} \left(\int_{\alpha_i}^{\beta_i} \frac{\partial \psi}{\partial x_i}\, ds \right)^2 \sum_{i=1}^{d} 1^2 \\ & \leq d \sum_{i=1}^{d} \int_{\alpha_i}^{\beta_i} \left| \frac{\partial \psi}{\partial x_i} \right|^2 ds\, |\beta_i - \alpha_i| \\ & \leq dh \sum_{i=1}^{d} \int_{\alpha_i}^{\beta_i} \left| \frac{\partial \psi}{\partial x_i} \right|^2 ds.\end{aligned}$$

Integrating B over ∂K and then integrating A over K, we obtain that

$$\text{meas}(\partial K)\|\psi\|_{0,K}^2 + \text{meas}(K)\|\psi\|_{0,\partial K}^2 - 2\int_K \psi \int_{\partial K} \psi \leq dh \times \text{meas}(K)\|\nabla\psi\|_{0,K}^2. \quad (2.17)$$

Since the triangulation is shape-regular, meas(K) is of order h^d and meas(∂K) is of order h^{d-1}. By using the Cauchy-Schwarz inequality,

$$\int_K \psi \leq \left(\int_K \psi^2 \right)^{1/2} \left(\int_K 1^2 \right)^{1/2} \leq Ch^{d/2}\|\psi\|_{0,K}, \quad (2.18)$$

and similarly,

$$\int_{\partial K} \psi \leq Ch^{(d-1)/2}\|\psi\|_{0,\partial K}, \quad (2.19)$$

where C is a constant independent of h. Substituting (2.18) and (2.19) into (2.17) gives

$$h^{d-1}||\psi||^2_{0,K} + h^d||\psi||^2_{0,\partial K} \leq Ch^{d-\frac{1}{2}}||\psi||_{0,K}||\psi||_{0,\partial K} + Ch^{d+1}||\nabla\psi||^2_{0,K}. \tag{2.20}$$

Using the Cauchy-Schwarz inequality again,

$$Ch^{d-\frac{1}{2}}||\psi||_{0,K}||\psi||_{0,\partial K} \leq \frac{1}{2}h^d||\psi||^2_{0,\partial K} + Ch^{d-1}||\psi||^2_{0,K}$$

Therefore, it is not hard to see from (2.20) that

$$\frac{1}{2}h^d||\psi||^2_{0,\partial K} \leq Ch^{d-1}||\psi||^2_{0,K} + Ch^{d+1}||\nabla\psi||^2_{0,K},$$

and (2.15) is obtained. In the case where ψ belongs to the finite element subspace, one can use the following inverse inequality (see (2.14)):

$$||\nabla\psi||_{0,K} \leq Ch^{-1}||\psi||_{0,K},$$

and (2.16) is obtained directly from (2.15). □

A similar version of the trace inequality and a different proof can be found in [11].

Chapter 3

STABILIZED DISCONTINUOUS FEM FOR ELLIPTIC PROBLEMS

In this chapter, two stabilized discontinuous finite element schemes for elliptic problems are presented. The first one is a non-symmetric formulation and is absolutely stable in the sense that no parameter selection is necessary for the scheme to converge. The second one is a symmetric formulation, but is conditionally stable in that a parameter has to be selected in order to have an optimal order of convergence. Optimal-order error estimates in some H^1-equivalence norms are established for the proposed discontinuous finite element methods. For the symmetric formulation, an optimal-order error estimate is also derived in the L^2 norm. We also discuss briefly finite element schemes using a jump term. The main results of this chapter is summarized in Ewing, Wang, and Yang [19].

3.1 A stabilized non-symmetric formulation

Let us recall that we are concerned with stabilized discontinuous finite element procedure for the model problem (1.1)-(1.2). $\{K\} = \mathcal{T}_h$ is a non-overlapping partition of the region under consideration Ω into polygonal elements that has a mesh size of h and is quasi-uniform. For simplicity, we only discuss the case in which Ω is in $\mathbb{R}^2$. The results can be extended to the three dimensional case without any difficulty.

In addition, we assume that the common boundary of any two adjacent elements is a straight line segment; i.e., $e = \partial K_1 \cap \partial K_2$ is either an empty set or a line segment for any $K_1, K_2 \in \mathcal{T}_h$. Thus, the boundary of each element $K \in \mathcal{T}_h$ consists of line segments as follows:

$$\partial K = \bigcup_{i=1}^{m(K)} e_{i,K}. \tag{3.1}$$

We emphasize that each element K may not necessarily be a triangle or quadrilateral as commonly seen in the standard Galerkin finite element method.

In Chapter 1, the model problem (1.1)-(1.2) was rewritten in a weak form given in (1.9). We show that the saddle-point problem (1.9) can be stabilized by using the following least-squares term:

$$\alpha h \sum_{K \in \mathcal{T}_h} \int_{\partial K} (\lambda - \kappa \nabla u \cdot n_K)(\mu - \kappa \nabla v \cdot n_K)\, ds, \tag{3.2}$$

where $\alpha > 0$ is an arbitrary, but fixed real number. To this end, let us introduce two functional spaces as follows:

$$\begin{aligned} X &= \left\{ v : \ v|_K \in H^{\frac{3}{2}}(K), \forall K \in \mathcal{T}_h \right\}, & (3.3)\\ Y &= \left\{ \mu \in \prod_{K \in \mathcal{T}_h} L^2(\partial K), \mu|_{\partial K_1} + \mu|_{\partial K_2} = 0 \quad \text{on } \partial K_1 \cap \partial K_2 \right\}. & (3.4) \end{aligned}$$

On the space $X \times Y$, we define a bilinear form as follows:

$$\begin{aligned} \mathcal{L}^{(sn)}(u, \lambda; v, \mu) &= \sum_{K \in \mathcal{T}_h} \int_K \kappa \nabla u \nabla v \, dK - \sum_{K \in \mathcal{T}_h} \int_{\partial K} \lambda v \, ds + \sum_{K \in \mathcal{T}_h} \int_{\partial K} u \mu \, ds \\ &\quad + \alpha h \sum_{K \in \mathcal{T}_h} \int_{\partial K} (\lambda - \kappa \nabla u \cdot n_K)(\mu - \kappa \nabla v \cdot n_K)\, ds, \end{aligned} \tag{3.5}$$

where the superscript $^{(sn)}$ stands for stabilized non-symmetric. It is easy to see that $\mathcal{L}^{(sn)}(\cdot; \cdot)$ is non-symmetric. The non-negativity can be seen by letting $v = u$ and

$\mu = \lambda$ in the bilinear form $\mathcal{L}^{(sn)}(\cdot;\cdot)$:

$$\mathcal{L}^{(sn)}(v,\mu;v,\mu) = \sum_{K\in\mathcal{T}_h}\int_K \kappa\nabla v\nabla v\, dK + \alpha h \sum_{K\in\mathcal{T}_h}\int_{\partial K}(\mu - \kappa\nabla v\cdot n_K)^2\, ds. \tag{3.6}$$

The bilinear form $\mathcal{L}^{(sn)}(\cdot;\cdot)$ is a modification of the forms associated with (1.9) by adding the least-squares term which plays the role of stabilization. The stabilized problem seeks $(w,\lambda)\in X\times Y$ such that

$$\mathcal{L}^{(sn)}(w,\lambda;v,\mu) = \ell(v,\mu), \qquad \forall v\in X, \mu\in Y, \tag{3.7}$$

where

$$\ell(v,\mu) = \sum_{K\in\mathcal{T}_h}\left(\int_K f(x)v(x)dK + \int_{\partial K\cap\partial\Omega} g(x)\mu(x)ds\right) \tag{3.8}$$

is a continuous linear functional on the space $X\times Y$. The derivation of (3.7) can be shown as follows. First, we start with the elliptic equation

$$-\nabla\cdot(\kappa\nabla u) = f.$$

Multiplying by v and then integrating over K we obtain

$$\int_K \kappa\nabla u\nabla v\, dK - \int_{\partial K}\kappa\nabla u\cdot n_K v\, ds = \int_K fv\, dK.$$

Using $\lambda = \kappa\nabla u\cdot n_K$ and sum for all K we have

$$\sum_K\int_K \kappa\nabla u\nabla v\, dK - \sum_K\int_{\partial K}\lambda v\, ds = \sum_K\int_K fv\, dK. \tag{3.9}$$

Next, when $e = K_1\cap K_2$ is an interior edge, $\int_{e,K_1} u\mu + \int_{e,K_2} u\mu = 0$. Therefore, by the boundary condition (1.2) we have

$$\sum_K \int_{\partial K} u\mu \, ds = \sum_K \int_{\partial K \cap \partial\Omega} g\mu \, ds. \tag{3.10}$$

Finally, from $\lambda = \kappa \nabla u \cdot n_K$ we obtain

$$\alpha h \sum_K \int_{\partial K} (\lambda - \kappa \nabla u \cdot n_K)(\mu - \kappa \nabla v \cdot n_K) ds = 0. \tag{3.11}$$

Adding (3.9), (3.10), and (3.11) together we arrive at (3.7).

The problem (3.7) is called a *stabilized non-symmetric formulation* for the model problem (1.1) and (1.2).

Lemma 3.1 *If the variational problem (3.7) is solvable, then the solution is unique.*

Proof: If $(u_i, \lambda_i) \in X \times Y$ are two solutions of (3.7) for $i = 1, 2$, then the difference $(w, \lambda) = (u_1 - u_2, \lambda_1 - \lambda_2)$ is a solution of the following homogeneous problem:

$$\mathcal{L}^{(sn)}(w, \lambda; v, \mu) = 0, \qquad \forall v \in X, \mu \in Y. \tag{3.12}$$

By letting $v = w, \mu = \lambda$ in (3.12), we have from (3.6) that

$$\sum_{K \in \mathcal{T}_h} \left(\int_K \kappa \nabla w \nabla w \, dK + \alpha h \int_{\partial K} (\lambda - \kappa \nabla w \cdot n_K)^2 \, ds \right) = 0.$$

The above equality implies that $\lambda = 0$ and w has constant value on each element K. Now we let $v = 0$ in (3.12) and obtain:

$$\sum_K \int_{\partial K} w\mu \, ds = 0, \qquad \forall \mu \in Y. \tag{3.13}$$

Recall that the values of μ differ only in sign on the interior edges. The equation (3.13) then indicates that the jump of w across each interior edge must be zero. Thus, w is a constant on the domain Ω. Since $w = 0$ on the boundary of Ω, then we have $w = 0$ on Ω. This completes the proof of the lemma. □

Let $u = u(x)$ be the exact solution of (1.1)-(1.2) such that $u \in H^{3/2}(\Omega)$. By letting $w = u$ and $\lambda = \kappa \nabla u \cdot n_K$, we see that $(w, \lambda) \in X \times Y$ solves the variational problem (3.7). The following lemma shows that the converse is true also.

Lemma 3.2 *Let $(w, \lambda) \in X \times Y$ be a solution of the variational problem (3.7). Then the pair (w, λ) also solves the saddle-point problem (1.9).*

Proof: Assume that $(w, \lambda) \in X \times Y$ is a solution pair of (3.7). For any element $K \in \mathcal{T}_h$, by letting $\mu = 0$ and $v \in C_c^\infty(K)$ we have from (3.7) that

$$-\nabla \cdot (\kappa \nabla w) = f \qquad \text{in } K, \tag{3.14}$$

where $C_c^\infty(K)$ is the set of $C^\infty(K)$ functions with proper compact support. Next, let $e = \partial K_1 \cap \partial K_2$ be any interior edge of the partition $\mathcal{T}_h$. In (3.7), we choose μ such that $\mu = 0$ everywhere except on e to obtain

$$\alpha h \left(\lambda|_{e-} - \lambda|_{e+} - \kappa \nabla w \cdot n_{K_1} + \kappa \nabla w \cdot n_{K_2}\right) + w|_{e,K_1} - w|_{e,K_2} = 0, \tag{3.15}$$

where $\lambda|_{e-}$ and $\lambda|_{e+}$ are the values of λ as seen from the element K_1 and K_2, respectively. Similarly, $w|_{e,K_1}$ ($w|_{e,K_2}$) stands for the trace of w on the edge e as taken from the element K_1 (K_2). Recall that $\lambda|_{e-} = -\lambda|_{e+}$. Thus,

$$\alpha h \left(2\lambda|_{e-} - \kappa \nabla w \cdot n_{K_1} + \kappa \nabla w \cdot n_{K_2}\right) + [\![w]\!] = 0, \tag{3.16}$$

where

$$[\![w]\!] = w|_{e,K_1} - w|_{e,K_2}$$

is the jump of the function $w = w(x)$ on the edge e. If $e \subset \partial K_1 \cap \partial\Omega$ is a boundary edge, then (3.15) must be modified as follows:

$$\alpha h \left(\lambda|_{e-} - \kappa\nabla w \cdot n_{K_1}\right) + w|_{e,K_1} - g|_e = 0. \tag{3.17}$$

Let X_c be a subspace of X consisting of functions with the following properties:

- $v \in H^1(\Omega)$
- $\kappa\nabla v$ is continuous in the normal direction across each interior edge e.

By letting $\mu = 0$ and $v \in X_c$ in (3.7), we arrive at

$$\sum_{K\in\mathcal{T}_h} \left(\int_K \kappa\nabla w\nabla v dK - \alpha h \int_{\partial K} (\lambda - \kappa\nabla w \cdot n_K)(\kappa\nabla v \cdot n_K) ds \right) = \int_\Omega fv \; d\Omega \tag{3.18}$$

for all $v \in X_c$. Substituting (3.16) and (3.17) into (3.18) yields

$$\sum_{K\in\mathcal{T}_h} \left(\int_K \kappa\nabla w\nabla v dK + \int_{\partial K} w \; \kappa\nabla v \cdot n_K ds \right) = \int_\Omega fv d\Omega + \int_{\partial\Omega} g \; \kappa\nabla v \cdot n ds.$$

Furthermore, we apply the Green's formula to the first term

$$\sum_{K\in\mathcal{T}_h} \left(\int_K (-\nabla \cdot \kappa\nabla w) v \; dK + \int_{\partial K} \kappa\nabla w \cdot n_K \; v ds + \int_{\partial K} w \; \kappa\nabla v \cdot n_K \; ds \right)$$
$$= \int_\Omega fv d\Omega + \int_{\partial\Omega} g \; \kappa\nabla v \cdot n \; ds,$$

which together with (3.14) gives

$$\sum_{K\in\mathcal{T}_h}\left(\int_{\partial K}\kappa\nabla w\cdot n_K\ vds+\int_{\partial K}w\ \kappa\nabla v\cdot n_K\ ds\right)=\int_{\partial\Omega}g\ \kappa\nabla v\cdot n\ ds. \tag{3.19}$$

The last equation implies that $[\![w]\!]=0$ on each interior edge and $w=g$ on the boundary of Ω. In addition, the flux $\kappa\nabla w\cdot n$ is continuous across each interior edge. Thus, equation (1.1) and the boundary condition (1.2) are both satisfied. □

The stabilized problem (3.7) can be approximated by a finite element method using discontinuous elements. To this end, we introduce two finite element spaces as follows:

$$\begin{aligned} X_h &= \{v:\ v|_K\in P_r(K)\}, \\ Y_h &= \left\{\mu\in Y:\ \mu|_{\partial K}\in\prod_{i=1}^{m(K)}P_s(e_{i,K}),\forall K\in\mathcal{T}_h\right\}, \end{aligned} \tag{3.20}$$

where $P_r(K)$ and $P_s(e_{i,K})$ denote the space of polynomials of degree no more than $r\geq 1$ and $s\geq 0$ on K and its boundary piece $e_{i,K}$ respectively. Our stabilized discontinuous finite element method seeks $u_h\in X_h$ and $\lambda_h\in Y_h$ satisfying

$$\mathcal{L}^{(sn)}(u_h,\lambda_h;v,\mu)=\ell(v,\mu),\qquad\forall v\in X_h,\ \mu\in Y_h. \tag{3.21}$$

Theorem 3.1 *The discontinuous finite element scheme (3.21) has one and only one solution in the finite element space $X_h\times Y_h$.*

Proof: The numerical scheme (3.21) comprises a system of linear equations where the number of equations is the same as the number of unknowns. Therefore, it is sufficient to show the uniqueness of solution for (3.21). To this end, let $\ell(v,\mu)=0$ for all $(v,\mu)\in X_h\times Y_h$ and $(u_h,\lambda_h)\in X_h\times Y_h$ be the corresponding solution. By letting $v = u_h$ and $\mu = \lambda_h$ in (3.21) we arrive at

$$\mathcal{L}^{(sn)}(u_h, \lambda_h; u_h, \lambda_h) = 0.$$

Using the definition of the bilinear form $\mathcal{L}^{(sn)}(\cdot;\cdot)$, it is easy to see that

$$\sum_K \int_K \kappa \nabla u_h \nabla u_h \, dK + \sum_K \int_{\partial K} |\lambda_h - \kappa \nabla u_h \cdot n_K|^2 \, ds = 0. \tag{3.22}$$

Equation (3.22) implies that $\lambda_h = 0$ and u_h is a constant on each element K. Next, we let $v = 0$ in (3.21) to obtain

$$\sum_K \int_{\partial K} u_h \mu \, ds = 0, \qquad \forall \mu \in Y_h. \tag{3.23}$$

Since the values of μ differ only in sign on the interior edge of element boundaries. The equation (3.23) then shows that the jump of u_h is zero across each interior edge and $u_h = 0$ on each boundary edge. Thus, $u_h = 0$ on the domain Ω. This completes the proof of the theorem. □

3.2 A stabilized symmetric formulation

In this section, our objective is to present a stabilized symmetric formulation for (1.1)-(1.2) which is conditionally stable. To this end, we consider the following symmetric bilinear form on the space $X \times Y$:

$$\begin{aligned} \mathcal{L}^{(ss)}(u, \lambda; v, \mu) \; &= \sum_{K \in \mathcal{T}_h} \int_K \kappa \nabla u \nabla v \, dK - \sum_{K \in \mathcal{T}_h} \int_{\partial K} \lambda v \, ds - \sum_{K \in \mathcal{T}_h} \int_{\partial K} u \mu \, ds \\ &\quad - \alpha h \sum_{K \in \mathcal{T}_h} \int_{\partial K} (\lambda - \kappa \nabla u \cdot n_K)(\mu - \kappa \nabla v \cdot n_K) \, ds, \end{aligned} \tag{3.24}$$

where $\alpha > 0$ is an arbitrary, but fixed real number. In this case the superscript $^{(ss)}$ stands for stabilized symmetric. The bilinear form $\mathcal{L}^{(ss)}(\cdot;\cdot)$ is clearly symmetric and is a stabilized version of the bilinear form for the saddle-point problem (1.9). The stability term is given by

$$-\alpha h \sum_{K \in \mathcal{T}_h} \int_{\partial K} (\lambda - \kappa \nabla u \cdot n_K)(\mu - \kappa \nabla v \cdot n_K)\, ds,$$

which vanishes if (u, λ) is the exact solution of (1.9) and is sufficiently smooth.

The variational problem associated with $\mathcal{L}^{(ss)}(\cdot;\cdot)$ seeks $(w, \lambda) \in X \times Y$ satisfying

$$\mathcal{L}^{(ss)}(w, \lambda; v, \mu) = \ell(v, \mu), \qquad \forall v \in X, \mu \in Y, \tag{3.25}$$

where

$$\ell(v, \mu) = \sum_{K \in \mathcal{T}_h} \left(\int_K f(x)v(x)dK - \int_{\partial K \cap \partial \Omega} g(x)\mu(x)ds \right) \tag{3.26}$$

is a continuous linear functional on the space $X \times Y$. The derivation is similar to the derivation of the stabilized non-symmetric formulation (3.7). In fact, (3.25) can be obtained by subtracting (3.10) and (3.11) from (3.9). The problem (3.25) is said to be a *stabilized symmetric formulation* for the model problem (1.1)-(1.2). Similar to Lemma 3.2, the following result can be proved without any difficulty.

Lemma 3.3 *If (u, λ) is the exact solution of (1.9) such that $u \in H^{\frac{3}{2}}(\Omega)$, then $(u, \lambda) \in X \times Y$ is a solution of the variational problem (3.25). Conversely, if $(w, \lambda) \in X \times Y$ is a solution of the variational problem (3.25), then the pair (w, λ) also solves the saddle-point problem (1.9).*

Let X_h and Y_h be the pair of finite element spaces defined in (3.20). The stabilized discontinuous finite element method seeks $u_h \in X_h$ and $\lambda_h \in Y_h$ such that

$$\mathcal{L}^{(ss)}(u_h, \lambda_h; v, \mu) = \ell(v, \mu), \qquad \forall (v, \mu) \in X_h \times Y_h. \tag{3.27}$$

To study the finite element scheme (3.27), we need the trace inequality established in (2.15). Let $K \in \mathcal{T}_h$ be an element in the finite element partition $\mathcal{T}_h$. There exists a constant $C > 0$ such that for any $\psi \in H^1(K)$, we have

$$\|\psi\|_{0,\partial K}^2 \leq C \left(h^{-1} \|\psi\|_{0,K}^2 + h \|\nabla \psi\|_{0,K}^2 \right). \tag{3.28}$$

The following lemma characterizes a useful property of the bilinear form (3.25) in the finite element space $X_h \times Y_h$.

Lemma 3.4 *There exists a constant $\alpha_0 > 0$ independent of the mesh size h such that, for any $\alpha \in (0, \alpha_0)$,*

$$\mathcal{L}^{(ss)}(v, \mu; v, -\mu) \geq C \sum_{K \in \mathcal{T}_h} \left(\int_K \kappa \nabla v \nabla v dK + \alpha h \int_{\partial K} \mu^2 ds \right), \tag{3.29}$$

for all $(v, \mu) \in X_h \times Y_h$.

Proof: From the definition of $\mathcal{L}^{(ss)}(\cdot;\cdot)$, we have

$$\mathcal{L}^{(ss)}(v, \mu; v, -\mu) = \sum_{K \in \mathcal{T}_h} \left(\int_K \kappa \nabla v \nabla v dK + \alpha h \int_{\partial K} (\mu^2 - (\kappa \nabla v \cdot n_K)^2) ds \right). \tag{3.30}$$

Now using the trace inequality (3.28), we can estimate the boundary integral of (3.30) as follows:

$$\int_{\partial K} (\kappa \nabla v \cdot n_K)^2 ds \le C \left(h^{-1} \int_K \kappa \nabla v \nabla v dK + h \int_K |D^2 v|^2 dK \right),$$

where $D^2 v$ represents all the partial derivatives of v of order 2. It then follows from the inverse inequality that

$$\int_{\partial K} (\kappa \nabla v \cdot n_K)^2 ds \le C h^{-1} \int_K \kappa \nabla v \nabla v dK.$$

Substituting the above estimate into (3.31) yields

$$\mathcal{L}^{(ss)}(v, \mu; v, -\mu) = (1 - \alpha C) \sum_{K \in \mathcal{T}_h} \int_K \kappa \nabla v \nabla v dK + \alpha h \sum_{K \in \mathcal{T}_h} \int_{\partial K} \mu^2 ds,$$

which implies the existence of an α_0 with the desired property. □

One important application of the inequality (3.29) is in the solvability of the numerical scheme (3.27). The result is stated as follows.

Theorem 3.2 *There exists a constant $\alpha_0 > 0$ independent of the mesh size h such that for any $\alpha \in (0, \alpha_0)$, the discontinuous finite element scheme (3.27) has a unique solution in the finite element space $X_h \times Y_h$.*

Proof: The numerical scheme (3.27) is a system of linear equations where the number of equations equals the number of unknowns. Thus, it is sufficient to prove the uniqueness for (3.27). To this end, let $\ell(v, \mu) = 0$ for all $(v, \mu) \in X_h \times Y_h$ and $(u_h, \lambda_h) \in X_h \times Y_h$ be the corresponding solution. By setting $v = u_h$ and $\mu = -\lambda_h$ in (3.27), we obtain

$$\mathcal{L}^{(ss)}(u_h, \lambda_h; u_h, -\lambda_h) = 0. \tag{3.31}$$

It follows from Lemma 3.4 that there exists a constant α_0 such that, for any $\alpha \in (0, \alpha_0)$, the inequality (3.29) is satisfied. Thus, with this choice of α, we have from (3.31) that

$$\sum_K \left(\int_K \kappa \nabla u_h \nabla u_h \, dK + \int_{\partial K} |\lambda_h|^2 \, ds \right) \leq 0.$$

This implies that $\lambda_h = 0$ and u_h is a constant on each element K. Next, we let $v = 0$ in (3.27) and obtain:

$$-\sum_K \int_{\partial K} u_h \mu \, ds = 0, \qquad \forall \mu \in Y_h. \tag{3.32}$$

Since the values of μ differ only in sign on the interior edge of element boundaries. The equation (3.32) shows that the jump of u_h is zero across each interior edge and $u_h = 0$ on any boundary edge. Thus, $u_h = 0$ on the domain Ω. □

3.3 Error estimates, I: non-symmetric formulation

In this section, we derive some error estimates for the stabilized non-symmetric finite element scheme (3.21). For simplicity of notation, we introduce some $\mathcal{T}_h$-dependent norms. Let $H^j(K)$ be the standard Sobolev space of order $j \geq 0$ over $K \in \mathcal{T}_h$. For each $\phi \in \prod_{K \in \mathcal{T}_h} H^j(K)$, define

$$||\phi||_{j;h} = \left(\sum_{K \in \mathcal{T}_h} ||\phi||_{j,K}^2 \right)^{1/2}, \tag{3.33}$$

where $||\cdot||_{j,K}$ denotes the usual Sobolev norm in $H^j(K)$. For each $\chi \in \prod_{K \in \mathcal{T}_h} L^2(\partial K)$, define

$$||\chi||_{0;\partial h} = \left(\sum_{K \in \mathcal{T}_h} ||\chi||_{0,\partial K}^2 \right)^{1/2}, \tag{3.34}$$

where $\|\chi\|_{0,\partial K}$ is the standard L^2-norm of χ on $L^2(\partial K)$. We also introduce a semi-norm on the space $X \times Y$ as follows:

$$|\!|\!|(v,\mu)|\!|\!| = \left(\sum_{K\in\mathcal{T}_h} \int_K \kappa\nabla v\nabla v dK + \alpha h\|\mu - \kappa\nabla v\cdot n_K\|_{0;\partial h}^2 \right)^{\frac{1}{2}}. \qquad (3.35)$$

Let $u = u(x)$ be the exact solution of the problem (1.1)-(1.2). Assume that u is sufficiently regular that $u \in X$. By letting $\lambda = \kappa\nabla u\cdot n_K$, we see that $(u,\lambda) \in X\times Y$ is a solution of the stabilized variational problem (3.7). Consequently, if (u_h, λ_h) is the finite element solution given by (3.21), then the following *error equation* is satisfied:

$$\mathcal{L}^{(ns)}(u - u_h, \lambda - \lambda_h; v, \mu) = 0, \qquad \forall (v,\mu) \in X_h \times Y_h. \qquad (3.36)$$

Let Q_h be the L^2 projection from Y to Y_h. The following is our first error estimate for the finite element solution (u_h, λ_h).

Lemma 3.5 *Assume that the solution u of (1.1)-(1.2) is sufficiently smooth such that $u \in X\cap H^1(\Omega)$. Let $\lambda \in Y$ be given by $\lambda|_{\partial K} = \kappa\nabla u\cdot n_K$. If $(u_h, \lambda_h) \in X_h\times Y_h$ is the stabilized discontinuous finite element approximation obtained by solving (3.21), then there is a constant C such that*

$$\begin{aligned} |\!|\!|(u-u_h, \lambda-\lambda_h)|\!|\!| &\le \inf_{\phi\in X_h} |\!|\!|(u-\phi, \lambda - Q_h\lambda)|\!|\!| \qquad (3.37)\\ &+C\inf_{\phi\in X_h}\left(\|u-\phi\|_{1;h}^2 + h^{-2}\|u-\phi\|_{0;h}^2 + h\|\lambda - Q_h\lambda\|_{0,\partial h}^2\right)^{\frac{1}{2}}. \end{aligned}$$

Proof: Let ϕ be any function in the finite element space X_h. Using the triangle inequality, we have

$$|||(u-u_h,\lambda-\lambda_h)||| \le |||(u-\phi,\lambda-Q_h\lambda)||| + |||(u_h-\phi,\lambda_h-Q_h\lambda)|||. \tag{3.38}$$

It suffices to establish an estimate for the second term of the right-hand side of (3.38). To this end, we use the non-negativity relation (3.6) to obtain

$$|||(u_h-\phi,\lambda_h-Q_h\lambda)|||^2 = \mathcal{L}^{(sn)}(u_h-\phi,\lambda_h-Q_h\lambda;u_h-\phi,\lambda_h-Q_h\lambda). \tag{3.39}$$

By letting

$$(v,\mu)=(u_h-\phi,\lambda_h-Q_h\lambda), \tag{3.40}$$

we have from (3.39) and the error equation (3.36) that

$$\begin{aligned}
&|||(u_h-\phi,\lambda_h-Q_h\lambda)|||^2 = \mathcal{L}^{(sn)}(v,\mu;v,\mu) = \mathcal{L}^{(sn)}(u-\phi,\lambda-Q_h\lambda;v,\mu)\\
&= \sum_{K\in\mathcal{T}_h}\int_K \kappa\nabla(u-\phi)\nabla v\, dK + \sum_{K\in\mathcal{T}_h}\int_{\partial K} v(\lambda-Q_h\lambda)\, ds\\
&\quad -\sum_{K\in\mathcal{T}_h}\int_{\partial K}(u-\phi)\mu\, ds\\
&\quad +\alpha h\sum_{K\in\mathcal{T}_h}(\lambda-Q_h\lambda-\kappa\nabla(u-\phi)\cdot n_K)(\mu-\kappa\nabla v\cdot n_K)\, ds.\\
&= I_1+I_2+I_3+I_4,
\end{aligned} \tag{3.41}$$

where I_i are defined accordingly for $i=1,\cdots,4$.

The two terms I_1 and I_4 can be estimated using the Cauchy-Schwarz inequality as follows:

$$|I_1+I_4| \le C|||(u-\phi,\lambda-Q_h\lambda)|||\ |||(v,\mu)|||. \tag{3.42}$$

As for I_3, the Cauchy-Schwarz inequality can be employed to give

$$\begin{aligned} |I_3| &\leq \sum_{K\in\mathcal{T}_h} \|u-\phi\|_{0,\partial K}\|\mu\|_{0,\partial K} \\ &\leq \left(\sum_{K\in\mathcal{T}_h} \|u-\phi\|_{0,\partial K}^2\right)^{\frac{1}{2}} \left(\sum_{K\in\mathcal{T}_h} \|\mu\|_{0,\partial K}^2\right)^{\frac{1}{2}}. \end{aligned} \tag{3.43}$$

Next, we use the trace inequality (3.28) to obtain

$$\|u-\phi\|_{0,\partial K}^2 \leq C(h^{-1}\|u-\phi\|_{0,K}^2 + h\|u-\phi\|_{1,K}^2). \tag{3.44}$$

It follows from the triangle inequality and the trace inequality (3.28) that

$$\begin{aligned} \|\mu\|_{0,\partial K}^2 &\leq 2\|\mu - \kappa\nabla v\cdot n_K\|_{0,\partial K}^2 + 2\|\kappa\nabla v\cdot n_K\|_{0,\partial K}^2 \\ &\leq Ch^{-1}\left(h\|\mu-\kappa\nabla v\cdot n_K\|_{0,\partial K}^2 + \int_K \kappa\nabla v\cdot\nabla v\, dK\right). \end{aligned} \tag{3.45}$$

Substituting (3.44) and (3.45) into (3.43) yields

$$|I_3| \leq C\left(\sum_{K\in\mathcal{T}_h}(h^{-2}\|u-\phi\|_{0,K}^2 + \|u-\phi\|_{1,K}^2)\right)^{\frac{1}{2}} |\!|\!|(v,\mu)|\!|\!|. \tag{3.46}$$

It remains to deal with $I_2 = \sum_{K\in\mathcal{T}_h}\int_{\partial K} v(\lambda - Q_h\lambda)\, ds$. Since Q_h is the L^2-projection operator onto Y_h and Y_h consists of discontinuous piecewise polynomials, then

$$\int_{\partial K} v(\lambda - Q_h\lambda)\, ds = \int_{\partial K}(I-Q_h)v\,(\lambda - Q_h\lambda)\, ds.$$

By using the Cauchy-Schwarz inequality and the standard interpolation error estimate

(2.8) with $k = 1$, we obtain

$$\begin{aligned} |I_2| &\leq \left(\sum_{K\in\mathcal{T}_h} \|v - Q_h v\|_{0,\partial K}^2\right)^{\frac{1}{2}} \left(\sum_{K\in\mathcal{T}_h} \|\lambda - Q_h\lambda\|_{0,\partial K}^2\right)^{\frac{1}{2}} \\ &\leq Ch\left(\sum_{K\in\mathcal{T}_h} |v|_{1,\partial K}^2\right)^{\frac{1}{2}} \left(\sum_{K\in\mathcal{T}_h} \|\lambda - Q_h\lambda\|_{0,\partial K}^2\right)^{\frac{1}{2}}, \end{aligned} \tag{3.47}$$

where $|v|_{1,\partial K}^2$ stands for the H^1 semi-norm of v on the boundary ∂K. Since v is a polynomial on the element K, there exists a constant C such that

$$|v|_{1,\partial K}^2 \leq Ch^{-1} \int_K \kappa \nabla v \cdot \nabla v dK. \tag{3.48}$$

Substituting (3.48) into (3.47) gives

$$\begin{aligned} |I_2| &\leq C\left(\sum_{K\in\mathcal{T}_h} \int_K \kappa\nabla v \cdot \nabla v dK\right)^{\frac{1}{2}} \left(h\sum_{K\in\mathcal{T}_h} \|\lambda - Q_h\lambda\|_{0,\partial K}^2\right)^{\frac{1}{2}} \\ &\leq C|\!|\!|(v,\mu)|\!|\!| \left(h\sum_{K\in\mathcal{T}_h} \|\lambda - Q_h\lambda\|_{0,\partial K}^2\right)^{\frac{1}{2}}. \end{aligned} \tag{3.49}$$

Finally, we combine (3.41) with (3.42), (3.46), and (3.49) to obtain

$$\begin{aligned} &|\!|\!|(u_h - \phi, \lambda_h - Q_h\lambda)|\!|\!|^2 \\ &\leq C\left(\|u-\phi\|_{1;h}^2 + h^{-2}\|u-\phi\|_{0;h}^2 + h\|\lambda - Q_h\lambda\|_{0;\partial h}^2)\right)^{\frac{1}{2}} \; |\!|\!|(v,\mu)|\!|\!|, \end{aligned} \tag{3.50}$$

which, together with (3.40), implies (3.37) and therefore, completes the proof of the lemma. □

The error estimate (3.37) provides a measure of the error $u - u_h$ in the H^1 semi-norm. Since the finite element solution u_h is discontinuous, the estimate (3.37) does not give any information about the level of continuity of u_h. The next lemma intends to address this concern.

Lemma 3.6 *Under the assumptions of Lemma 3.5, there exists a constant C independent of the mesh size h such that*

$$\begin{aligned} h^{-1} \sum_{K\in\mathcal{T}_h} \int_{\partial K} [\![Q_h(u-u_h)]\!]^2 ds \leq & \; 4\alpha^2 \inf_{\phi\in X_h} |\!|\!|(u-\phi, \lambda - Q_h\lambda)|\!|\!|^2 \\ & + C\alpha^2 \inf_{\phi\in X_h} \left(\|u-\phi\|_{1;h}^2 + h^{-2}\|u-\phi\|_{0;h}^2 + h\|\lambda - Q_h\lambda\|_{0;\partial h}^2 \right), \end{aligned} \tag{3.51}$$

where $[\![\cdot]\!]$ stands for the jump on interior edges. On boundary edges, $[\![\cdot]\!]$ should be understood as a one-sided trace of the function under consideration.

Proof: By taking $v = 0$ in the error equation (3.36), we obtain

$$\sum_{K\in\mathcal{T}_h} \int_{\partial K} (u - u_h + \alpha h(\lambda - \lambda_h - \kappa\nabla(u-u_h)\cdot n_K))\mu \, ds = 0 \tag{3.52}$$

for all $\mu \in Y_h$. This implies that

$$[\![Q_h(u-u_h)]\!] = -\alpha h \; Q_h[\![\lambda - \lambda_h - \kappa\nabla(u-u_h)\cdot n_K]\!]. \tag{3.53}$$

Using the fact that Q_h is the L^2 projection operator, we have from (3.53) that

$$\sum_{K\in\mathcal{T}_h} \int_{\partial K} [\![Q_h(u-u_h)]\!]^2 ds \leq 2\alpha^2 h^2 \sum_{K\in\mathcal{T}_h} \|\lambda - \lambda_h + \kappa\nabla(u-u_h)\cdot n_K\|_{0,\partial K}^2. \tag{3.54}$$

The right-hand side of (3.54) is related to the semi-norm given by (3.35). In particular, it is easy to see that

$$h^{-1} \sum_{K \in \mathcal{T}_h} \int_{\partial K} [\![Q_h(u-u_h)]\!]^2 ds \leq 2\alpha^2 |\!|\!| (u-u_h, \lambda - \lambda_h) |\!|\!|^2, \tag{3.55}$$

which, together with Lemma 3.5, completes the proof of the lemma. □

In the rest of this section, we are concerned with applications of (3.37) and (3.51). Assume that the finite element partition $\mathcal{T}_h$ is regular such that there exists a constant C satisfying

$$\inf_{\phi \in X_h} \left(h^{-1} \|u-\phi\|_{0;h} + \|u-\phi\|_{1;h} + h\|u-\phi\|_{2;h} \right) \leq C h^m \|u\|_{m+1;h}, \tag{3.56}$$

for $0 \leq m \leq r$ and $u \in \prod_{K \in \mathcal{T}_h} H^{m+1}(K)$. Next, we provide an estimate for $\|\lambda - Q_h\lambda\|_{0,\partial K}$ if $\lambda = \kappa \nabla w \cdot n_K$ for some smooth function w defined on K.

Lemma 3.7 *Let $K \in \mathcal{T}_h$ be an element in the finite element partition $\mathcal{T}_h$. Let $\lambda = \kappa \nabla w \cdot n_K$ be the normal component of the flux $q = \kappa \nabla w$. Then there exists a constant $C > 0$ such that*

$$\|\lambda - Q_h \lambda\|_{0,\partial K} \leq C h^{\ell - \frac{1}{2}} \|w\|_{\ell+1,K}, \qquad 0 \leq \ell \leq s+1.$$

Proof: Let $\Pi_h w \in P_{s+1}(K)$ be the L^2-projection of w in the polynomial space $P_{s+1}(K)$ and

$$\tilde{\lambda} = \kappa \nabla (\Pi_h w) \cdot n_K. \tag{3.57}$$

Then,

$$\lambda - Q_h\lambda = (\lambda - \tilde{\lambda}) + (\tilde{\lambda} - Q_h \tilde{\lambda}) + Q_h(\tilde{\lambda} - \lambda). \tag{3.58}$$

It follows that

$$\begin{aligned} \|\lambda - Q_h\lambda\|_{0,\partial K} &\leq \|\lambda - \tilde{\lambda}\|_{0,\partial K} + \|\tilde{\lambda} - Q_h\tilde{\lambda}\|_{0,\partial K} + \|Q_h(\tilde{\lambda} - \lambda)\|_{0,\partial K} \\ &\leq 2\|\lambda - \tilde{\lambda}\|_{0,\partial K} + \|\tilde{\lambda} - Q_h\tilde{\lambda}\|_{0,\partial K}. \end{aligned} \tag{3.59}$$

Since $\lambda = \kappa\nabla w \cdot n_K$ and $\tilde{\lambda} = \kappa\nabla(\Pi_h w) \cdot n_K$, the trace inequality (3.28) and the standard interpolation error estimate (2.8) implies

$$\begin{aligned} \|\lambda - \tilde{\lambda}\|_{0,\partial K}^2 &= \int_{\partial K} |\lambda - \tilde{\lambda}|^2 ds = \int_{\partial K} |\kappa\nabla(w - \Pi_h w) \cdot n_K|^2 ds \\ &\leq C\left(h^{-1}\|\nabla(w - \Pi_h w)\|_{0,K}^2 + h\|D^2(w - \Pi_h w)\|_{0,K}^2\right) \\ &\leq Ch^{2\ell-1}\|w\|_{\ell+1,K}^2, \qquad \forall 1 \leq \ell \leq s+1. \end{aligned} \tag{3.60}$$

As for the second term of (3.59), we use the interpolation error estimate to obtain

$$\|\tilde{\lambda} - Q_h\tilde{\lambda}\|_{0,\partial K} \leq Ch^\ell \sum_{i=1}^{m(K)} \|\tilde{\lambda}\|_{\ell,e_{i,K}}, \qquad 0 \leq \ell \leq s+1, \tag{3.61}$$

where $\|\cdot\|_{\ell,e_{i,K}}$ is the norm in the Sobolev space $H^\ell(e_{i,K})$. To estimate each term $\|\tilde{\lambda}\|_{\ell,e_{i,K}}$, we assume that the coefficient tensor κ in (1.1) is sufficiently smooth on the element K. Using the trace inequality (3.28) and (3.57), we obtain

$$\|\tilde{\lambda}\|_{\ell,e_{i,K}}^2 \leq C\left(h^{-1}\|\Pi_h w\|_{\ell+1,K}^2 + h|\Pi_h w|_{\ell+2,K}^2\right), \tag{3.62}$$

where $|\Pi_h w|_{\ell+2,K}$ is the Sobolev semi-norm of $\Pi_h w$ on the element K. It follows from the standard inverse inequality (2.14) that

$$h|\Pi_h w|_{\ell+2,K}^2 \leq h^{-1}\|\Pi_h w\|_{\ell+1,K}^2. \tag{3.63}$$

Thus, substituting (3.63) into (3.62) we have

$$\|\tilde{\lambda}\|^2_{\ell,e_i,K} \leq Ch^{-1}\|\Pi_h w\|^2_{\ell+1,K} \leq Ch^{-1}\|w\|^2_{\ell+1,K}, \tag{3.64}$$

where we have used the boundedness of the L^2-projection Π_h in $H^{\ell+1}(K)$. Substituting (3.64) into (3.61) leads to

$$\|\tilde{\lambda} - Q_h\tilde{\lambda}\|_{0,\partial K} \leq Ch^{\ell-\frac{1}{2}}\|w\|_{\ell+1,K}, \qquad 0 \leq \ell \leq s+1. \tag{3.65}$$

Now combining (3.59) with (3.60) and (3.65) we arrive at

$$\|\lambda - Q_h\lambda\|_{0,\partial K} \leq Ch^{\ell-\frac{1}{2}}\|w\|_{\ell+1,K}, \qquad 0 \leq \ell \leq s+1, \tag{3.66}$$

which completes the proof. □

We are now in a position to state the main result of this section regarding the accuracy of the finite element method (3.21).

Theorem 3.3 *Under the assumptions of Lemma 3.5, there exists a constant C such that*

$$\begin{aligned} &|||(u-u_h, \lambda-\lambda_h)||| + h^{-\frac{1}{2}}\left(\sum_{K\in\mathcal{T}_h}\int_{\partial K}[\![Q_h(u-u_h)]\!]^2 ds\right) \\ &\quad \leq C\left(h^m\|u\|_{m+1;h} + h^\ell\|u\|_{\ell+1;h}\right), \end{aligned} \tag{3.67}$$

for any $1 \leq m \leq r$ and $1 \leq \ell \leq s+1$, provided that the solution u is sufficiently smooth.

Proof: The proof is essentially an application of the interpolation error estimate

(3.56) and the results developed in Lemmas 3.5–3.7. In fact, it follows from Lemmas 3.5 and 3.6 that

$$\begin{aligned}
&|\!|\!|(u-u_h,\lambda-\lambda_h)|\!|\!| + h^{-\frac{1}{2}}\left(\sum_{K\in\mathcal{T}_h}\int_{\partial K}[\![Q_h(u-u_h)]\!]^2 ds\right) \\
&\le \inf_{\phi\in X_h} |\!|\!|(u-\phi,\lambda-Q_h\lambda)|\!|\!| \\
&+C\inf_{\phi\in X_h}\left(\|u-\phi\|_{1;h}+h^{-1}\|u-\phi\|_{0;h}+h^{\frac{1}{2}}\|\lambda-Q_h\lambda\|_{0;\partial h}\right).
\end{aligned} \tag{3.68}$$

Since

$$\begin{aligned}
|\!|\!|(u-\phi,\lambda-Q_h\lambda)|\!|\!|^2 &= \sum_{K\in\mathcal{T}_h}\int_K \kappa\nabla(u-\phi)\cdot\nabla(u-\phi)dK \\
&+ \alpha h\|\lambda-Q_h\lambda-\kappa\nabla(u-\phi)\cdot n_K\|_{0;\partial h}^2,
\end{aligned}$$

then

$$\begin{aligned}
|\!|\!|(u-\phi,\lambda-Q_h\lambda)|\!|\!|^2 &\le C(\|\nabla(u-\phi)\|_{0;h}^2+\alpha h\|\lambda-Q_h\lambda\|_{0;\partial h}^2 \\
&+ \alpha h\|\kappa\nabla(u-\phi)\cdot n_K\|_{0;\partial h}^2).
\end{aligned} \tag{3.69}$$

Using the trace inequality (3.28), we obtain

$$\|\kappa\nabla(u-\phi)\cdot n_K\|_{0,\partial K}^2 \le C\left(h^{-1}\|\nabla(u-\phi)\|_{0,K}^2 + h\|\nabla^2(u-\phi)\|_{0,K}^2\right). \tag{3.70}$$

Substituting (3.70) into (3.69) yields

$$\begin{aligned}
|\!|\!|(u-\phi,\lambda-Q_h\lambda)|\!|\!|^2 &\le C(\|u-\phi\|_{1;h}^2+h^2\|u-\phi\|_{2;h}^2 \\
&+ \alpha h\|\lambda-Q_h\lambda\|_{0;\partial h}^2),
\end{aligned}$$

which, together with (3.68), leads to

$$\begin{aligned}&|||(u-u_h,\lambda-\lambda_h)||| + h^{-\frac{1}{2}}\left(\textstyle\sum_{K\in\mathcal{T}_h}\int_{\partial K}[\![Q_h(u-u_h)]\!]^2 ds\right)^{\frac{1}{2}}\\ &\leq Ch^{\frac{1}{2}}\|\lambda-Q_h\lambda\|_{0;\partial h}^2\\ &+C\inf_{\phi\in X_h}\left(h^{-1}\|u-\phi\|_{0;h}+\|u-\phi\|_{1;h}+h\|u-\phi\|_{2;h}\right).\end{aligned}\tag{3.71}$$

Finally, the desired error estimate (3.67) can be obtained by combining (3.71) with Lemma 3.7 and the interpolation error estimate (3.56). □

3.4 Error estimates, II: symmetric formulation

Our objective in this section is to derive some error estimates for the symmetric finite element scheme (3.27). Due to the inequality established in Lemma 3.4, a natural "energy" norm associated with the symmetric bilinear form $\mathcal{L}^{(ss)}(\cdot;\cdot)$ can be defined by

$$|||(v;\mu)|||_s = \left(\sum_{K\in\mathcal{T}_h}\int_K \kappa\nabla v\nabla v dK + \alpha h\|\mu\|_{0;\partial h}^2\right)^{\frac{1}{2}}. \tag{3.72}$$

In fact, following the proof of Theorem 3.3, it is not hard to establish an error estimate for the symmetric formulation in the "energy" norm $|||\cdot|||_s$. The result is stated as follows.

Theorem 3.4 *Assume that the solution u of (1.1)-(1.2) is sufficiently smooth such that $u\in X\cap H^1(\Omega)$. Let $\lambda\in Y$ be given by $\lambda|_{\partial K}=\kappa\nabla u\cdot n_K$. Let $(u_h,\lambda_h)\in X_h\times Y_h$ be the stabilized discontinuous finite element approximation obtained by solving (3.27). There exists a constant $\alpha_0>0$ independent of the mesh size h such that, for any*

$\alpha \in (0, \alpha_0)$,

$$\begin{aligned}&|||(u-u_h, \lambda-\lambda_h)|||_s + h^{-\frac{1}{2}}||[\![Q_h(u-u_h)]\!]||_{0;\partial h} \leq \inf_{\phi \in X_h} |||(u-\phi, \lambda - Q_h\lambda)|||_s \\ &+C \inf_{\phi \in X_h} \left(||u-\phi||_{1;h}^2 + h^{-2}||u-\phi||_{0;h}^2 + h||\lambda - Q_h\lambda||_{0;\partial h}^2\right)^{\frac{1}{2}},\end{aligned}$$

where C is a constant independent of the mesh size h. Moreover, there exists a constant C such that, for any $1 \leq m \leq r$ and $1 \leq \ell \leq s+1$,

$$\begin{aligned}&|||(u-u_h, \lambda-\lambda_h)|||_s + h^{-\frac{1}{2}}||[\![Q_h(u-u_h)]\!]||_{0;\partial h} \\ &\leq C\left(h^m||u||_{m+1;h} + h^{\ell}||u||_{\ell+1;h}\right), \qquad (3.73)\end{aligned}$$

provided that the exact solution u is sufficiently smooth.

The inequality (3.73) provides an optimal-order error estimate for $u-u_h$ in the H^1-norm and $\lambda-\lambda_h$ in the L^2-norm on interior edges. In the rest of this section, we establish an error estimate for $u-u_h$ in the L^2-norm by using the well-known duality argument. To this end, we consider a dual problem which seeks $\psi \in H_0^1(\Omega)$ such that

$$\begin{aligned}-\nabla\cdot(\kappa\nabla\psi) &= u-u_h, \qquad \text{in } \Omega, \qquad (3.74)\\ \psi &= 0, \qquad \text{on } \partial\Omega\end{aligned}$$

Assume that the H^2-regularity holds true for the dual problem (3.74). In other words, the solution of (3.74) is in $H^2(\Omega)\cap H_0^1(\Omega)$ and there exists a constant C satisfying

$$||\psi||_{2,\Omega} \leq C||u-u_h||_{0,\Omega}. \qquad (3.75)$$

Using the Green's formula, we obtain

$$\|u-u_h\|_{0,\Omega}^2 = \sum_{K\in\mathcal{T}_h}\left(\int_K \kappa\nabla\psi\cdot\nabla(u-u_h)dK - \int_{\partial K}(u-u_h)\chi ds\right), \tag{3.76}$$

where $\chi = \kappa\nabla\psi\cdot n_K$ is the normal component of the flux variable $q = \kappa\nabla\psi$. Let u be the exact solution of (1.1)-(1.2) and $\lambda = \kappa\nabla u\cdot n_K$. Let (u_h,λ_h) be their finite element approximations arising from (3.27). Since $\psi = 0$ on $\partial\Omega$, we have

$$\sum_{K\in\mathcal{T}_h}\int_{\partial K}\psi(\lambda-\lambda_h)ds = 0. \tag{3.77}$$

In addition, the fact that $\chi = \kappa\nabla\psi$ implies that

$$\sum_{K\in\mathcal{T}_h}\int_{\partial K}(\lambda-\lambda_h-\kappa\nabla(u-u_h)\cdot n_K)(\chi-\kappa\nabla\psi\cdot n_K)ds = 0. \tag{3.78}$$

It follows from (3.76), (3.77), and (3.78) that

$$\begin{aligned}
\|u-u_h\|_{0,\Omega}^2 &= \sum_{K\in\mathcal{T}_h}\left(\int_K \kappa\nabla\psi\cdot\nabla(u-u_h)dK - \int_{\partial K}(u-u_h)\chi ds\right)\\
&- \sum_{K\in\mathcal{T}_h}\int_{\partial K}\psi(\lambda-\lambda_h)ds\\
&- \alpha h\sum_{K\in\mathcal{T}_h}\int_{\partial K}(\lambda-\lambda_h-\kappa\nabla(u-u_h)\cdot n_K)(\chi-\kappa\nabla\psi\cdot n_K)ds\\
&= \mathcal{L}^{(ss)}(u-u_h,\lambda-\lambda_h;\psi,\chi).
\end{aligned}$$

Using an analogue of the error equation (3.36), we obtain

$$\|u-u_h\|_{0,\Omega}^2 = \mathcal{L}^{(ss)}(u-u_h,\lambda-\lambda_h;\psi-\phi,\chi-Q_h\mu) \tag{3.79}$$

for any $\phi \in X_h$. In particular, we choose $\phi = R_h\psi$ where R_h is the L^2 projection operator onto the finite element space X_h. The right-hand side of (3.79) can be split into four terms as follows:

$$\begin{aligned} J_1 &= \sum_{K\in\mathcal{T}_h} \int_K \kappa\nabla(\psi - R_h\psi)\cdot\nabla(u-u_h)dK, && (3.80)\\ J_2 &= -\sum_{K\in\mathcal{T}_h} \int_{\partial K} (u-u_h)(\chi - Q_h\chi)ds, && (3.81)\\ J_3 &= -\sum_{K\in\mathcal{T}_h} \int_{\partial K} (\psi - R_h\psi)(\lambda-\lambda_h)ds, && (3.82)\end{aligned}$$

and

$$J_4 = \alpha h \sum_{K\in\mathcal{T}_h} \int_{\partial K} (\lambda - \lambda_h - \kappa\nabla(u-u_h)\cdot n_K)(\chi - Q_h\chi - \kappa\nabla(\psi - R_h\psi)\cdot n_K)ds. \quad (3.83)$$

These J-terms are handled by the following four lemmas.

Lemma 3.8 *Let J_1 be given by (3.80). There exists a constant C such that*

$$|J_1| \leq Ch||\nabla(u-u_h)||_{0;h}||u-u_h||_{0,\Omega}. \quad (3.84)$$

Proof: Using the Cauchy-Schwarz inequality we obtain

$$|J_1| \leq C||\nabla(u-u_h)||_{0;h}\ ||\nabla(\psi - R_h\psi)||_{0;h}. \quad (3.85)$$

The standard interpolation error estimate (2.8) implies that

$$||\nabla(\psi - R_h\psi)||_{0;h} \leq Ch||\psi||_{2,\Omega} \leq Ch||u-u_h||_{0,\Omega}, \quad (3.86)$$

where we have used the a priori estimate (3.75) in the last inequality. The proof is then completed by substituting (3.86) into (3.85). □

Lemma 3.9 *Let J_2 be given by (3.81). There exists a constant C such that*

$$|J_2| \leq Ch \left(\|\nabla(u-u_h)\|_{0;h} + h\|D^2(u-u_h)\|_{0;h} \right) \|u-u_h\|_{0,\Omega}. \tag{3.87}$$

Proof: Since Q_h is the L^2 projection operator onto the discontinuous finite element space Y_h, then

$$J_2 = -\sum_{K\in\mathcal{T}_h} \int_{\partial K} [(I-Q_h)(u-u_h)](\chi - Q_h\chi)ds. \tag{3.88}$$

Using the Cauchy-Schwarz inequality and the interpolation error estimate we obtain

$$|J_2| \leq Ch \sum_{K\in\mathcal{T}_h} |u-u_h|_{1,\partial K}\|\chi - Q_h\chi\|_{0,\partial K}, \tag{3.89}$$

where $|u-u_h|_{1,\partial K}$ denotes the $H^1(\partial K)$ semi-norm. The trace inequality (3.28) can be applied to yield

$$|u-u_h|_{1,\partial K}^2 \leq C\left(h^{-1}\|\nabla(u-h_h)\|_{0,K}^2 + h\|D^2(u-u_h)\|_{0,K}^2\right). \tag{3.90}$$

Furthermore, with $\mu = \chi, w = \psi$, and $\ell = 1$, we have from Lemma 3.7 that

$$\|\chi - Q_h\chi\|_{0,\partial K} \leq Ch^{\frac{1}{2}}\|\psi\|_{2,K}. \tag{3.91}$$

Substituting (3.90) and (3.91) into (3.89) yields

$$|J_2| \leq Ch \left(\|\nabla(u - u_h)\|_{0;h} + h\|D^2(u - u_h)\|_{0;h} \right) \|\psi\|_{2,\Omega}, \tag{3.92}$$

which, together with the a priori estimate (3.75), completes the proof of the lemma.

□

Lemma 3.10 *Let J_3 be given by (3.82). There exists a constant C such that*

$$|J_3| \leq Ch^{\frac{3}{2}} \|\lambda - \lambda_h\|_{0;\partial h} \; \|u - u_h\|_{0,\Omega}. \tag{3.93}$$

Proof: From the Cauchy-Schwarz inequality, we have

$$|J_3| \leq \sum_{K \in \mathcal{T}_h} \|\lambda - \lambda_h\|_{0,\partial K} \; \|\psi - R_h\psi\|_{0,\partial K}. \tag{3.94}$$

The trace inequality (3.28) can be applied to yield

$$\|\psi - R_h\psi\|^2_{0,\partial K} \leq C(h^{-1}\|\psi - R_h\psi\|^2_{0,K} + h\|\nabla(\psi - R_h)\psi\|^2_{0,K}) \leq Ch^3\|D^2\psi\|^2_{0,K}. \tag{3.95}$$

Substituting (3.95) into (3.94), we obtain

$$|J_3| \leq Ch^{\frac{3}{2}} \|\psi\|_{2,\Omega} \; \|\lambda - \lambda_h\|_{0;\partial h}, \tag{3.96}$$

which, together with (3.75), completes the proof. □

Lemma 3.11 *Let J_4 be given by (3.83). There exists a constant C such that*

$$|J_4| \leq Ch\left(h^{\frac{1}{2}}\|\lambda - \lambda_h\|_{0;\partial h} + \|\nabla(u - u_h)\|_{0;h} + h\|D^2(u - u_h)\|_{0;h}\right)\|u - u_h\|_{0,\Omega}. \tag{3.97}$$

Proof: From the Cauchy-Schwarz inequality, we have

$$\begin{aligned} |J_4| \quad \leq \quad & \alpha h \sum_{K \in \mathcal{T}_h} (\|\lambda - \lambda_h\|_{0,\partial K} + \|\kappa\nabla(u - u_h) \cdot n_K\|_{0,\partial K}) \qquad (3.98) \\ & \cdot(\|\chi - Q_h\chi\|_{0,\partial K} + \|\kappa\nabla(\psi - R_h\psi) \cdot n_K\|_{0,\partial K}). \end{aligned}$$

The trace inequality (3.28) and Lemma 3.7 can be applied to yield

$$\|\chi - Q_h\chi\|_{0,\partial K} + \|\kappa\nabla(\psi - R_h\psi) \cdot n_K\|_{0,\partial K} \leq Ch^{\frac{1}{2}}\|\psi\|_{2,K}. \tag{3.99}$$

Similarly, the trace inequality (3.28) gives

$$\|\kappa\nabla(u - u_h) \cdot n_K\|_{0,\partial K}^2 \leq C\left(h^{-1}\|\nabla(u - u_h)\|_{0,K}^2 + h\|D^2(u - u_h)\|_{0,K}^2\right). \tag{3.100}$$

Substituting (3.99) and (3.100) into (3.98), we obtain

$$|J_4| \leq Ch\left(h^{\frac{1}{2}}\|\lambda - \lambda_h\|_{0;\partial h} + \|\nabla(u - u_h)\|_{0;h} + h\|D^2(u - u_h)\|_{0;h}\right)\ \|\psi\|_{2,\Omega}, \tag{3.101}$$

which, together with the a priori estimate (3.75), completes the proof of the lemma.
□

We are now in a position to prove our L^2-error estimate for the symmetric finite element method (3.27).

Theorem 3.5 *Under the assumptions of Theorem 3.4, there exists a constant $\alpha_0 > 0$ independent of the mesh size h such that, for any $\alpha \in (0, \alpha_0)$,*

$$\|u - u_h\|_{0,\Omega} \leq C\left(h |\!|\!| (u - u_h, \lambda - \lambda_h) |\!|\!|_s + h^2 \|D^2(u - u_h)\|_{0;h}\right), \tag{3.102}$$

where C is a generic constant independent of the mesh size h. Moreover, there exists a constant C such that, for any $1 \leq m \leq r$ and $1 \leq \ell \leq s + 1$,

$$\|u - u_h\|_{0,\Omega} \leq C\left(h^{m+1} \|u\|_{m+1;h} + h^{\ell+1} \|u\|_{\ell+1;h}\right), \tag{3.103}$$

provided that the solution u is sufficiently smooth.

Proof: Recall that from (3.79) we have

$$\|u - u_h\|_{0,\Omega}^2 = J_1 + J_2 + J_3 + J_4. \tag{3.104}$$

Thus, by the estimates derived in Lemmas 3.8-3.11 we obtain

$$\|u - u_h\|_{0,\Omega}^2 \leq C\left(h |\!|\!| (u - u_h, \lambda - \lambda_h) |\!|\!|_s + h^2 \|D^2(u - u_h)\|_{0;h} \| u - u_h\|_{0,\Omega}. \tag{3.105}$$

This proves (3.102). The estimate (3.103) is a direct application of (3.102) and (3.73).
□

3.5 Bilinear forms with jump terms

In this section, we introduce non-symmetric and symmetric formulations that use the least-squares term (1.13) and the jump term (1.14).

Let us first define a non-symmetric bilinear form on $X \times Y$ as follows:

$$\begin{aligned}\mathcal{L}^{(snj)}(u,\lambda;v,\mu) &= \sum_{K\in\mathcal{T}_h}\int_K \kappa\nabla u\nabla v\, dK - \sum_{K\in\mathcal{T}_h}\int_{\partial K}\lambda v\, ds \\ &+\sum_{K\in\mathcal{T}_h}\int_{\partial K} u\mu\, ds + \beta h^{-1}\sum_{K\in\mathcal{T}_h}\int_{\partial K}[\![u]\!][\![v]\!]\, ds \\ &+\alpha h\sum_{K\in\mathcal{T}_h}\int_{\partial K}(\lambda-\kappa\nabla u\cdot n_K)(\mu-\kappa\nabla v\cdot n_K)\, ds\end{aligned} \tag{3.106}$$

where $\alpha > 0$ and $\beta > 0$ are arbitrary but fixed real numbers. Our stabilized finite element approximation consists in seeking $u_h \in X_h$ and $\lambda_h \in Y_h$ such that

$$\mathcal{L}^{(snj)}(u_h,\lambda_h;v,\mu) = \ell(v,\mu), \qquad \forall v\in X_h,\ \mu\in Y_h, \tag{3.107}$$

where $\ell(v,\mu)$ is defined in (3.8). The problem (3.107) is called a *stabilized non-symmetric formulation with jump term* for the model problem (1.1)- (1.2).

The linear system in (3.107) is uniquely solvable once we show that bilinear form (3.106) is coercive in $\mathcal{V}_h \times \mathcal{M}_h$. A bilinear form $\mathcal{L}(\cdot;\cdot)$ is said to be coercive on V (or V-elliptic) if, there exists a constant $\alpha > 0$ such that,

$$\mathcal{L}(v,v) \geq \alpha\|v\|_V, \qquad \forall v\in V$$

. In fact, it is easy to see that

$$\mathcal{L}^{(snj)}(v,\mu;v,\mu) = |\!|\!|(v,\mu)|\!|\!|^2 + \beta h^{-1}\sum_{K\in\mathcal{T}_h}\int_{\partial K}[\![v]\!]^2\, ds,$$

where the semi-norm $|\!|\!|\cdot|\!|\!|$ is defined in (3.35). Therefore, the bilinear form $\mathcal{L}^{(snj)}(\cdot;\cdot)$ is coercive for any α and β.

A symmetric bilinear form on $X \times Y$ can be defined as follows:

$$\begin{aligned}\mathcal{L}^{(ssj)}(u,\lambda;v,\mu) &= \sum_{K\in\mathcal{T}_h}\int_K \kappa\nabla u\nabla v\, dK - \sum_{K\in\mathcal{T}_h}\int_{\partial K}\lambda v\, ds\\ &-\sum_{K\in\mathcal{T}_h}\int_{\partial K} u\mu\, ds + \beta h^{-1}\sum_{K\in\mathcal{T}_h}\int_{\partial K}[\![u]\!][\![v]\!]\, ds\\ &+\alpha h\sum_{K\in\mathcal{T}_h}\int_{\partial K}(\lambda-\kappa\nabla u\cdot n_K)(\mu-\kappa\nabla v\cdot n_K)\, ds\end{aligned} \tag{3.108}$$

where again $\alpha > 0$ and $\beta > 0$ are arbitrary but fixed. Our stabilized finite element approximation consists in seeking $u_h \in X_h$ and $\lambda_h \in Y_h$ such that

$$\mathcal{L}^{(ssj)}(u_h,\lambda_h;v,\mu) = \ell(v,\mu), \qquad \forall v \in X_h,\ \mu \in Y_h, \tag{3.109}$$

where $\ell(v,\mu)$ is defined in (3.26). The problem (3.109) is called a *stabilized symmetric formulation with jump term* for the model problem (1.1)-(1.2).

The linear system in (3.109) is uniquely solvable once we show that bilinear form (3.108) is coercive in $\mathcal{V}_h \times \mathcal{M}_h$. This is established in the following lemma:

Lemma 3.12 *For any α, there exist a constant β_0 such that, for any $\beta \geq \beta_0$, the bilinear form in (3.108) is coercive.*

Proof: We see from (3.108) that

$$\mathcal{L}^{(ssj)}(v,\mu;v,\mu) = |\!|\!|(v,\mu)|\!|\!|^2 + \beta h^{-1}\sum_{K\in\mathcal{T}_h}\int_{\partial K}[\![v]\!]^2\, ds - 2\sum_{K\in\mathcal{T}_h}\int_{\partial K} v\mu\, ds. \tag{3.110}$$

Since the values of μ differ only in sign on an edge e in two adjacent elements, we have

$$\left|\sum_{K\in\mathcal{T}_h}\int_{\partial K} v\mu\, ds\right| \leq \sum_{e\in\mathcal{E}_h}\int_e |\mu||[\![v]\!]|\, ds, \tag{3.111}$$

where $\mathcal{E}_h$ denotes the set of all interior edges (or faces) associated with the partition $\mathcal{T}_h$. By using the Cauchy-Schwarz inequality,

$$\sum_{e\in\mathcal{E}_h}\int_e |\mu||[\![v]\!]|\,ds \le \left(\beta^{-1}h\sum_{K\in\mathcal{T}_h}\int_{\partial K}|\mu|^2\right)^{\frac{1}{2}}\left(\beta h^{-1}\sum_{K\in\mathcal{T}_h}\int_{\partial K}[\![v]\!]^2\,ds\right)^{\frac{1}{2}}. \tag{3.112}$$

By using the triangle inequality and the trace inequality (3.28), we have

$$\begin{aligned}
&\beta^{-1}h\sum_{K\in\mathcal{T}_h}\int_{\partial K}|\mu|^2\,ds\\
&\le 2\beta^{-1}h\sum_{K\in\mathcal{T}_h}\left(\int_{\partial K}|\mu-\kappa\nabla v\cdot n_K|^2\,ds+\int_{\partial K}|\kappa\nabla v\cdot n_K|^2\,ds\right)\\
&\le C\beta^{-1}\sum_{K\in\mathcal{T}_h}\left(h\int_{\partial K}|\mu-\kappa\nabla v\cdot n_K|^2\,ds+\int_K\kappa\nabla v\nabla v\,dK\right)\\
&\le C\beta^{-1}\max(\alpha^{-1},1)\sum_{K\in\mathcal{T}_h}\left(\alpha h\int_{\partial K}|\mu-\kappa\nabla v\cdot n_K|^2\,ds+\int_K\kappa\nabla v\nabla v\,dK\right),
\end{aligned} \tag{3.113}$$

where C is an constant independent of h, α and β. For any fixed α, we choose β_0 to be sufficiently large so that

$$C\beta_0^{-1}\max(\alpha^{-1},1)\le\frac{1}{4}.$$

For $\beta\ge\beta_0$, it follows from (3.111)-(3.113) that

$$\mathcal{L}^{(ssj)}(v,\mu;v,\mu)\ge\frac{1}{2}\left(|\!|\!|(v,\mu)|\!|\!|^2+\beta h^{-1}\sum_{K\in\mathcal{T}_h}\int_{\partial K}[\![v]\!]^2\,ds\right). \tag{3.114}$$

This completes the proof the coercivity of (3.108). □

The H^1-equivalence norm error estimates for the symmetric and non-symmetric formulations with jump terms as well as the L^2 error estimate for symmetric formulation can be established in a similar way as in Section 3.3 and 3.4. In fact, these forms were studied before we discovered formulations without jump terms.

Chapter 4

CONVECTION-DIFFUSION PROBLEMS

The objective of this chapter is to apply stabilized discontinuous finite element procedure to convection-dominated convection-diffusion problems. In general, the standard Galerkin finite element methods applied to such problems exhibit a variety of deficiencies, including high oscillations and poor approximation of the derivatives of the solutions. A new stabilization technique, which features a non-symmetric formulation using discontinuous piecewise polynomials, is presented and analyzed for such problems. Error estimates in some H^1-equivalence norm is established for the proposed discontinuous finite element scheme.

4.1 A variational formula

We consider the application of the stabilized discontinuous finite element procedure to a convection-diffusion problem which seeks $u \in H^1(\Omega)$ satisfying

$$\begin{aligned} -\nabla \cdot (a\nabla u - \mathbf{b}u) + cu &= f \qquad \text{in } \Omega, \\ u &= g \qquad \text{on } \partial\Omega, \end{aligned} \tag{4.1}$$

where Ω is an open bounded domain in $\mathbb{R}^d$ $(d = 2, 3)$. For simplicity, we only discuss the case in which Ω is a polygonal domain in $\mathbb{R}^2$. The results can be extended to three dimensional case and curve boundary regions without any difficulty. $c \in L^\infty(\Omega)$ and $\mathbf{b} \in [W^{1,\infty}(\Omega)]^d$ are the coefficient functions, $f \in L^2(\Omega)$ and $g \in H^{1/2}(\Omega)$ are given.

Also, $a = (a_{ij})_{d\times d}$ is a symmetric matrix such that there are two positive constants a_1 and a_2 satisfying

$$a_1\|\xi\|^2 \leq \sum_{i,j} a_{ij}\xi_i\xi_j \leq a_2\|\xi\|^2, \quad \forall\xi = (\xi_1, \xi_2, \cdots, \xi_d) \in \Omega. \tag{4.2}$$

We impose the following assumptions:

H1. a_1 and a_2 are proportional and small:

$$a_2 \ll 1, \qquad a_2/a_1 = \mathcal{O}(1). \tag{4.3}$$

H2. There exists a positive c_0 such that

$$\frac{1}{2}\nabla \cdot \mathbf{b} + c \geq c_0 \qquad \text{in } \Omega. \tag{4.4}$$

To derive a variational form, let us first introduce some notation. Let $\{K\} = \mathcal{T}_h$ be a non-overlapping partition of Ω into polygonal elements which is quasi-uniform. For notational simplicity, we separate the boundary of Ω into two parts. $\partial\Omega_+$ denotes the part of the boundary where $\mathbf{b} \cdot n \geq 0$ and $\partial\Omega_-$ denotes the part of the boundary where $\mathbf{b} \cdot n < 0$ where n is the outward normal direction. This notation is illustrated in Fig. 4.1.

To obtain a weak formulation of (4.1), we introduce two function spaces

$$\mathcal{V} = \left\{v : v|_K \in H^1(K), \forall K \in \mathcal{T}_h\right\}, \tag{4.5}$$

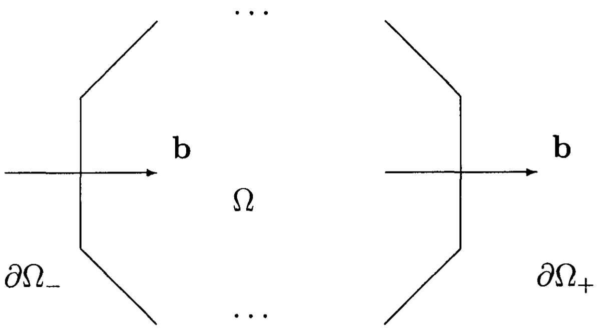

FIG. 4.1. Boundary $\partial\Omega_-$ and $\partial\Omega_+$

$$\begin{aligned}\mathcal{M} = \Big\{\mu : \mu|_{\partial K} \in H^{-\frac{1}{2}}(\partial K), \forall K \in \mathcal{T}_h, \exists \mathbf{q} \in H(div;\Omega), \mathbf{q}\cdot n_K = \mu|_{\partial K}, \\ \text{and } \mu|_{\partial\Omega_-} = 0\Big\},\end{aligned} \tag{4.6}$$

where n_K is the outward normal direction on ∂K. We also denote

$$\begin{aligned}\mathcal{V}_g &= \{v \in \mathcal{V} : v|_{\partial\Omega_-} = g\}, \\ \mathcal{V}_0 &= \{v \in \mathcal{V} : v|_{\partial\Omega_-} = 0\}.\end{aligned} \tag{4.7}$$

Therefore functions in $\mathcal{V}_g$ satisfies the boundary condition strongly on $\partial\Omega_-$. It can be shown that there exists a unique pair $(\bar{u}, \lambda) \in \mathcal{V}_g \times \mathcal{M}$ such that

$$\sum_{K\in\mathcal{T}_h} \int_K (a\nabla\bar{u} - \mathbf{b}\bar{u})\nabla v + c\bar{u}v dK - \sum_{K\in\mathcal{T}_h} \int_{\partial K} \lambda v ds = \sum_{K\in\mathcal{T}_h} \int_K f v dK, \tag{4.8}$$

$$\sum_{K\in\mathcal{T}_h} \int_{\partial K} \bar{u}\mu ds = \sum_{K\in\mathcal{T}_h} \int_{\partial K\cap\partial\Omega_+} g\mu ds, \tag{4.9}$$

for any $(v, \mu) \in \mathcal{V}_0 \times \mathcal{M}$. Moreover,

$$\bar{u} = u, \qquad \text{and} \qquad \lambda|_{\partial K} = (a\nabla u - \mathbf{b}u)\cdot n_K,\ \forall K \in \mathcal{T}_h, \tag{4.10}$$

where u is the solution of (4.1).

4.2 A stabilized non-symmetric formulation

In this section, we show that the weak form (4.8)-(4.9) can be stabilized by using the following procedure. Let us first introduce two function spaces:

$$\begin{aligned} X &= \left\{ v : \; v|_K \in H^{\frac{3}{2}}(K), \forall K \in \mathcal{T}_h \right\}, & (4.11) \\ Y &= \left\{ \mu \in \prod_{K \in \mathcal{T}_h} L^2(\partial K), \mu|_{\partial K_1} + \mu|_{\partial K_2} = 0 \quad \text{on } \partial K_1 \cap \partial K_2 \right\}. & (4.12) \end{aligned}$$

Similarly we denote

$$\begin{aligned} X_g &= \{ v \in X : v|_{\partial\Omega_-} = g \}, \\ X_0 &= \{ v \in X : v|_{\partial\Omega_-} = 0 \}. \end{aligned} \tag{4.13}$$

On the space $X \times Y$, we define a bilinear form:

$$\begin{aligned} \Phi(u, \lambda; v, \mu) &= \sum_{K \in \mathcal{T}_h} \int_K (a\nabla u - \mathbf{b}u)\nabla v + cuv \, dK - \sum_{K \in \mathcal{T}_h} \int_{\partial K} (\lambda v - u\mu) \, ds \\ &+ \alpha h^\beta \sum_{K \in \mathcal{T}_h} \int_{\partial K} z(\lambda - (a\nabla u - \mathbf{b}u) \cdot n_K)(\mu - a\nabla v \cdot n_K) \, ds, \end{aligned} \tag{4.14}$$

where $\alpha > 0$ and β are arbitrary but fixed numbers independent of h. The quantity z is an "inflow-outflow" indicator defined on the boundary of each element K as follows:

$$z = \begin{cases} 1, & \mathbf{b} \cdot n_K \geq 0, \\ 0, & \mathbf{b} \cdot n_K < 0. \end{cases} \tag{4.15}$$

As shown in Fig. 4.2, $z = 1$ for edge e on K_1 and $z = 0$ for e on K_2.

The stabilized problem seeks $w \in X_g$ and $\lambda \in Y$ such that

$$\Phi(w, \lambda; v, \mu) = \ell(v, \mu), \qquad \forall (v, \mu) \in X_0 \times Y, \tag{4.16}$$

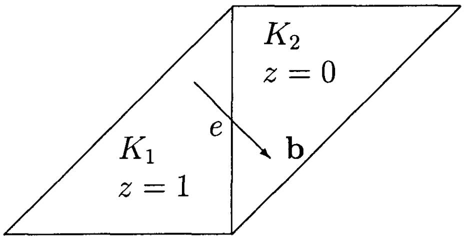

FIG. 4.2. The "inflow-outflow" indicator z and function $\tilde{w}$

where

$$\ell(v, \mu) = \sum_{K \in \mathcal{T}_h} \left(\int_K fv \, dK + \int_{\partial K \cap \partial \Omega_+} g\mu \, ds \right) \tag{4.17}$$

is a continuous linear functional on the space $X \times Y$.

The derivation of (4.16) can be shown in a similar way to the derivation of the schemes for elliptic problems. First, we start with the equation

$$-\nabla \cdot (a \nabla u - \mathbf{b} u) + cu = f.$$

Multiplying by v and then integrating over K we obtain

$$\int_K (a\nabla u - \mathbf{b}u)\nabla v + cuv \, dK - \int_{\partial K} (a\nabla u - \mathbf{b}u) \cdot n_K v \, ds = \int_K fv \, dK.$$

Using $\lambda = (a\nabla u - \mathbf{b}u) \cdot n_K$ and summing for all K, we have

$$\sum_K \int_K (a\nabla u - \mathbf{b}u)\nabla v \, dK - \sum_K \int_{\partial K} \lambda v \, ds = \sum_K \int_K fv \, dK. \tag{4.18}$$

Next, when $e = K_1 \cap K_2$ is an interior edge,

$$\int_{e,K_1} u\mu + \int_{e,K_2} u\mu = 0.$$

Therefore, by the boundary condition, we have

$$\sum_K \int_{\partial K} u\mu \, ds = \sum_K \int_{\partial K \cap \partial\Omega} g\mu \, ds. \tag{4.19}$$

Finally, from $\lambda = (a\nabla u - \mathbf{b}u) \cdot n_K$ we obtain

$$\alpha h^\beta \sum_K \int_{\partial K} z(\lambda - (a\nabla u - \mathbf{b}u) \cdot n_K)(\mu - a\nabla v \cdot n_K) ds = 0. \tag{4.20}$$

Adding (4.18), (4.19), and (4.20), we arrive at (4.16).

The next lemma shows the uniqueness of problem (4.16), which is one of the important results of this section.

Lemma 4.1 *If the variational problem (4.16) has a solution, then the solution is unique.*

Proof: It suffices to show that if

$$\Phi(w, \lambda; v, \mu) = 0, \qquad \forall (v, \mu) \in X_0 \times Y, \tag{4.21}$$

then $w = 0$ and $\lambda = 0$. We choose $v = w$ and $\mu = \lambda + (\mathbf{b} \cdot n_K)\tilde{w}$, where

$$\tilde{w} = \begin{cases} w_i, & \mathbf{b} \cdot n_K \geq 0, \\ w_o, & \mathbf{b} \cdot n_K < 0, \end{cases} \tag{4.22}$$

and $w_i(w_o)$ stands for the trace of w taken from interior (exterior) of element K. As shown in Fig. 4.2, $\tilde{w} = w|_{e,K_1}$. Therefore, it is clear that $\tilde{w}$ assumes the same trace on both sides of any edge in the partition and therefore $\mu = \lambda + (\mathbf{b} \cdot n_K)\tilde{w}$ belongs to

Y. By selecting such v and μ in (4.21), we have from (4.14) that

$$\begin{aligned}&\sum_K \int_K a\nabla w \nabla w + cw^2 dK - \sum_K \int_K \mathbf{b} w \nabla w dK + \sum_K \int_{\partial K} (\mathbf{b}\cdot n_K)\tilde{w} w ds \\ &+\alpha h^\beta \sum_K \int_{\partial K} z(\lambda - (a\nabla w - \mathbf{b}w)\cdot n_K)(\lambda + (\mathbf{b}\cdot n_K)\tilde{w} - a\nabla w \cdot n_K) ds = 0.\end{aligned} \tag{4.23}$$

The last term vanishes when $\mathbf{b}\cdot n_K < 0$ by the definition of the indicator z. When $\mathbf{b}\cdot n_K \geq 0$, $\tilde{w} = w_i = w|_K$. Therefore, the last term can be written as

$$\alpha h^\beta \sum_K \int_{\partial K, \mathbf{b}\cdot n_K \geq 0} (\lambda - (a\nabla w - \mathbf{b}w)\cdot n_K)^2 ds, \tag{4.24}$$

which is nonnegative.

To establish the non-negativity of the other terms, we introduce the following lemma:

Lemma 4.2 *Under the assumptions in Lemma 4.1, we have*

$$\begin{aligned}&-\sum_K \int_K \mathbf{b} w \nabla w dK + \sum_K \int_{\partial K} (\mathbf{b}\cdot n_K)\tilde{w} w ds \\ &= \frac{1}{2}\sum_K \int_K (\nabla \cdot \mathbf{b}) w^2 dK + \frac{1}{2}\sum_{e\in\mathcal{E}_h} \int_e |\mathbf{b}\cdot n_e| [\![w]\!]^2 ds,\end{aligned} \tag{4.25}$$

where

$$[\![w]\!] = w|_{e,K_1} - w|_{e,K_2} \tag{4.26}$$

is the jump of the function w across the edge e.

Proof: By using integration by parts,

$$\begin{aligned} -\int_K \mathbf{b}w\nabla w dK &= \int_K w\nabla \cdot (\mathbf{b}w)dK - \int_{\partial K}(\mathbf{b}\cdot n_K)w^2 ds \\ &= \int_K (\nabla\cdot\mathbf{b})w^2 dK + \int_K \mathbf{b}w\nabla w dK - \int_{\partial K}(\mathbf{b}\cdot n_K)w^2 ds. \end{aligned}$$

We now have

$$-\int_K \mathbf{b}w\nabla w dK = \frac{1}{2}\int_K (\nabla\cdot\mathbf{b})w^2 dK - \frac{1}{2}\int_{\partial K}(\mathbf{b}\cdot n_K)w^2 ds.$$

Summing for all K, we obtain that

$$\begin{aligned} &-\sum_K \int_K \mathbf{b}w\nabla w dK + \sum_K \int_{\partial K}(\mathbf{b}\cdot n_K)\tilde{w}w ds \\ &= \frac{1}{2}\sum_K \int_K (\nabla\cdot\mathbf{b})w^2 dK + \frac{1}{2}\sum_K \int_{\partial K}(\mathbf{b}\cdot n_K)(2w\tilde{w} - w^2)ds. \end{aligned} \tag{4.27}$$

The last term is nonnegative. To illustrate this, let $e = \partial K_1 \cap \partial K_2$ be the common boundary of elements K_1 and K_2. (See Fig. 4.2.) We denote by e^+ (e^-) the part of e where $\mathbf{b}\cdot n_{e,K_1} \geq 0$ (< 0). Therefore, on e^+, $\mathbf{b}\cdot n_{e,K_2} \leq 0$ and on e^-, $\mathbf{b}\cdot n_{e,K_2} > 0$.

When integrating on K_1, we have from the definition of $\tilde{w}$ that

$$\begin{aligned} \int_{e,K_1}(\mathbf{b}\cdot n_{e,K_1})(2w\tilde{w} - w^2)ds = &\int_{e^+}(\mathbf{b}\cdot n_{e,K_1})w|^2_{e,K_1}ds \\ + &\int_{e^-}(\mathbf{b}\cdot n_{e,K_1})\left(2w|_{e,K_1}w|_{e,K_2} - w|^2_{e,K_1}\right)ds. \end{aligned} \tag{4.28}$$

Similarly, when integrating on K_2,

$$\begin{aligned} \int_{e,K_2}(\mathbf{b}\cdot n_{e,K_2})(2w\tilde{w} - w^2)ds = &\int_{e^-}(\mathbf{b}\cdot n_{e,K_2})w|^2_{e,K_2}ds \\ + &\int_{e^+}(\mathbf{b}\cdot n_{e,K_2})\left(2w|_{e,K_1}w|_{e,K_2} - w|^2_{e,K_2}\right)ds. \end{aligned} \tag{4.29}$$

We combine (4.28) and (4.29) and obtain

$$\begin{aligned}&\sum_K \int_{\partial K} (\mathbf{b}\cdot n_K)(2w\tilde{w} - w^2)ds \\ &= \sum_{e\in\mathcal{E}_h} \left[\int_{e^+} (\mathbf{b}\cdot n_{e,K_1})[\![w]\!]^2 + \int_{e^-} (\mathbf{b}\cdot n_{e,K_2})[\![w]\!]^2\right] ds \\ &= \sum_{e\in\mathcal{E}_h} \int_e |\mathbf{b}\cdot n_e|[\![w]\!]^2 ds.\end{aligned} \tag{4.30}$$

In (4.30), $[\![w]\!]$ is understood as the one-sided trace on boundary edges. The result of the lemma is then obtained from (4.27) and (4.30). □

We now continue our proof of Lemma 4.1. In fact, we can see from (4.23), (4.24) and (4.25) that if (w, λ) solves the homogeneous problem (4.21), then

$$\begin{aligned}&\sum_K \int_K a\nabla w\nabla w + (\frac{1}{2}\nabla\cdot\mathbf{b} + c)w^2 dK + \frac{1}{2}\sum_{e\in\mathcal{E}_h}\int_e |\mathbf{b}\cdot n_e|[\![w]\!]^2 ds \\ &+\alpha h^\beta \sum_K \int_{\partial K, \mathbf{b}\cdot n_K \geq 0} (\lambda - (a\nabla w - \mathbf{b}w)\cdot n_K)^2\, ds = 0.\end{aligned} \tag{4.31}$$

Using the assumption **H2** (4.4), we know $w = 0$ everywhere in Ω. Also, $\lambda = 0$ on the part of ∂K where $\mathbf{b}\cdot n_K \geq 0$. This includes all boundary edges belonging to $\partial\Omega_+$. Remember that the values of λ differ only in sign for adjacent edges in the interior part of Ω. Thus $\lambda = 0$ on all interior edges. From (4.21),

$$\sum_K \int_{\partial K \cap \partial\Omega_-} \lambda v\, ds = 0 \qquad \forall v \in X.$$

This shows that $\lambda = 0$ on $\partial\Omega_-$. Therefore $\lambda = 0$ everywhere on ∂K for all element K. This completes the proof of uniqueness. □

The stabilized problem (4.16) can be approximated by a finite element method by using discontinuous elements. Let us introduce two finite element spaces:

$$\begin{aligned} X_h &= \{v \in X : \; v|_K \in P_r(K)\}, \\ Y_h &= \left\{ \mu \in Y : \; \mu|_{\partial K} \in \prod_{i=1}^{m(K)} P_s(e_{i,K}), \forall K \in \mathcal{T}_h \right\}, \end{aligned} \tag{4.32}$$

where $P_r(K)$ and $P_s(e_{i,K})$ denote the spaces of polynomials of degree no more than r and s on K and its boundary piece $e_{i,K}$, respectively. We make the following assumption:

H3. The degree of the polynomials on edges is no less than the degree of the polynomials on the element, i.e.,

$$s \geq r \geq 1. \tag{4.33}$$

Our stabilized discontinuous finite element method then seeks $u_h \in X_h \cap X_g$ (so that the boundary condition is satisfied on $\partial\Omega_-$ strongly) and $\lambda_h \in Y_h$ satisfying

$$\Phi(u_h, \lambda_h; v, \mu) = \ell(v, \mu), \qquad \forall v \in X_h, \; \mu \in Y_h. \tag{4.34}$$

Theorem 4.1 *The discontinuous finite element scheme has one and only one solution in the finite element space $X_h \times Y_h$.*

Proof: The numerical scheme (4.34) comprises a system of linear equations where the number of equations is the same as the number of unknowns. Therefore it is sufficient to show the uniqueness of the solution for (4.34). To this end, let $\ell(v, \mu) = 0$ and $(w_h, \lambda_h) \in X_h \times Y_h$ be the corresponding solution. As in the proof of Lemma 4.1, we select $v = w_h$ but here $\mu = \lambda_h + (\mathbf{b} \cdot n_K)\tilde{w_h}$ does not belong to Y_h in general. So we select $\mu = \lambda_h + Q_h[(\mathbf{b} \cdot n_K)\tilde{w_h}] \in Y_h$ where Q_h stands for the L^2-projection operator onto Y_h. We can rewrite

$$\mu = \lambda_h + (\mathbf{b} \cdot n_K)\tilde{w_h} + (Q_h - I)[(\mathbf{b} \cdot n_K)\tilde{w_h}] \tag{4.35}$$

where I denotes the identity operator. We obtain

$$\Phi(w_h, \lambda_h; w_h, \lambda_h + Q_h[(\mathbf{b} \cdot n_K)\tilde{w_h}]) = 0. \tag{4.36}$$

In fact, it follows from (4.14), (4.36) and the proof of Lemma 4.1 that

$$\begin{aligned}
&\Phi(w_h, \lambda_h; w_h, \lambda_h + Q_h[(\mathbf{b} \cdot n_K)\tilde{w_h}]) \\
&= \sum_K \int_K a\nabla w_h \nabla w_h + (\frac{1}{2}\nabla \cdot \mathbf{b} + c)w_h^2 + \frac{1}{2}\sum_{e\in\mathcal{E}_h}\int_e |\mathbf{b} \cdot n_e|[\![w_h]\!]^2 \\
&+\alpha h^\beta \sum_K \int_{\partial K, \mathbf{b}\cdot n_K \geq 0} (\lambda_h - (a\nabla w_h - \mathbf{b}w_h) \cdot n_K)^2 \\
&+\sum_K \int_{\partial K} w_h(Q_h - I)[(\mathbf{b} \cdot n_K)\tilde{w_h}] \\
&+\alpha h^\beta \sum_K \int_{\partial K} z\left(\lambda_h - (a\nabla w_h - \mathbf{b}w_h) \cdot n_K\right)(Q_h - I)[(\mathbf{b} \cdot n_K)\tilde{w_h}].
\end{aligned} \tag{4.37}$$

The second to last term vanishes because, on the boundary ∂K, w_h is a polynomial of degree $r \leq s$. The last term also vanishes for a similar reason because the degree of the polynomial $\lambda_h - (a\nabla w_h - \mathbf{b}w_h) \cdot n_K \in Y_h$ is no more than s. Therefore, we have

$$
\begin{aligned}
&\Phi(w_h, \lambda_h; w_h, \lambda_h + Q_h[(\mathbf{b}\cdot n_K)\tilde{w_h}]) \\
&= \sum_K \int_K a\nabla w_h \nabla w_h + (\frac{1}{2}\nabla\cdot\mathbf{b} + c)w_h^2 + \frac{1}{2}\sum_{e\in\mathcal{E}_h}\int_e |\mathbf{b}\cdot n_e|[\![w_h]\!]^2 \\
&+\alpha h^\beta \sum_K \int_{\partial K, \mathbf{b}\cdot n_K \geq 0} (\lambda_h - (a\nabla w_h - \mathbf{b}w_h)\cdot n_K)^2 .
\end{aligned}
\tag{4.38}
$$

It follows from (4.36) and (4.38) that $w_h = 0$ and $\lambda_h = 0$. This completes the proof of Theorem 4.1. □

4.3 Error Estimates

In this section, we derive an error estimate for the stabilized finite element scheme (4.34). We introduce a norm on $X \times Y$ as follows:

$$
\begin{aligned}
|\!|\!|(v,\mu)|\!|\!| \;=\; &\Bigg(\sum_K \int_K a\nabla v\nabla v + (\frac{1}{2}\nabla\cdot\mathbf{b} + c)v^2 + \frac{1}{2}\sum_{e\in\mathcal{E}_h}\int_e |\mathbf{b}\cdot n_e|[\![v]\!]^2 \\
&+ \alpha h^\beta \sum_K \int_{\partial K, \mathbf{b}\cdot n_K\geq 0} (\mu - (a\nabla v - \mathbf{b}v)\cdot n_K)^2\Bigg)^{\frac{1}{2}} .
\end{aligned}
\tag{4.39}
$$

Let $u = u(x)$ be the exact solution of (4.1). Assume that u is sufficiently regular such that $u \in X$. By letting $\lambda = (a\nabla u - \mathbf{b}u)\cdot n_K$, we see that $(u, \lambda) \in X \times Y$ is the solution of the stabilized variational problem (4.16). Consequently, if (u_h, λ_h) is the finite element solution of (4.34), then the following error equation is satisfied:

$$
\Phi(u - u_h, \lambda - \lambda_h; v, \mu) = 0, \qquad \forall (v,\mu) \in X_h \times Y_h. \tag{4.40}
$$

Let Π_h be the L^2 projection operator from X to X_h. By using the triangle inequality we have

$$|\!|\!|(u-u_h,\lambda-\lambda_h)|\!|\!| \le |\!|\!|(u-\Pi_h u,\lambda-Q_h\lambda)|\!|\!| + |\!|\!|(u_h-\Pi_h u,\lambda_h-Q_h\lambda)|\!|\!|. \tag{4.41}$$

It is sufficient to establish an estimate for the second term of the right-hand side of (4.41). To simplify the notation, let

$$(\phi,\eta) = (u-\Pi_h u,\lambda-Q_h\lambda), \tag{4.42}$$

$$(v,\mu) = (u_h-\Pi_h u,\lambda_h-Q_h\lambda), \tag{4.43}$$

and

$$\xi = (\mathbf{b}\cdot n_K)\tilde{v}. \tag{4.44}$$

We use the non-negativity relation (4.38) to obtain

$$|\!|\!|(u_h-\Pi_h u,\lambda_h-Q_h\lambda)|\!|\!|^2 = \Phi(u_h-\Pi_h u,\lambda_h-Q_h\lambda;v,\mu+Q_h\xi). \tag{4.45}$$

We have from (4.45) and (4.40) that

$$\begin{aligned}
&|\!|\!|(u_h-\Pi_h u,\lambda_h-Q_h\lambda)|\!|\!|^2\\
&=\Phi(u-\Pi_h u,\lambda-Q_h\lambda;v,\mu+Q_h\xi)\\
&=\sum_{K\in\mathcal{T}_h}\int_K (a\nabla\phi-\mathbf{b}\phi)\nabla v + c\phi v\,dK\\
&-\sum_{K\in\mathcal{T}_h}\int_{\partial K}\eta v\,ds + \sum_{K\in\mathcal{T}_h}\int_{\partial K}\phi(\mu+Q_h\xi)\,ds\\
&+\alpha h^{\beta}\sum_{K\in\mathcal{T}_h}\int_{\partial K} z(\eta-(a\nabla\phi-\mathbf{b}\phi)\cdot n_K)(\mu+Q_h\xi-a\nabla v\cdot n_K)\,ds\\
&=I_1+I_2+I_3+I_4,
\end{aligned} \tag{4.46}$$

where I_i are defined accordingly for $i = 1, \cdots, 4$.

The $I-$terms can be handled by the following lemmas.

Lemma 4.3 *Let I_1 be as in (4.46). There exists a constant C such that*

$$|I_1| \leq C \left(|||(u - \Pi_h u, \lambda - Q_h \lambda)||| + h^{-1} \|u - \Pi_h u\|_{0;h} \right) |||(v, \mu)|||. \tag{4.47}$$

Proof: By using the Cauchy-Schwarz inequality, we obtain

$$\left| \sum_{K \in \mathcal{T}_h} \int_K \mathbf{b} \phi \nabla v dK \right| \leq C \|\phi\|_{0;h} \|\nabla v\|_{0;h}. \tag{4.48}$$

Since v is a piecewise polynomial, we can use the standard inverse inequality (2.14) and obtain

$$\left| \sum_{K \in \mathcal{T}_h} \int_K \mathbf{b} \phi \nabla v dK \right| \leq C h^{-1} \|\phi\|_{0;h} \|v\|_{0;h}. \tag{4.49}$$

The terms involving coefficients a and c can be handled by the Cauchy-Schwarz inequality directly. This completes the proof of the lemma. □

Lemma 4.4 *Let I_2 be as in (4.46). There exists a constant C such that*

$$|I_2| \leq C h^{-\frac{1}{2}} \|\lambda - Q_h \lambda\|_{0;\partial h} |||(v, \mu)|||. \tag{4.50}$$

Proof: By using the Cauchy-Schwarz inequality, we obtain

$$|I_2| \leq \left(\sum_K \|\eta\|_{0,\partial K}^2 \right)^{\frac{1}{2}} \left(\sum_K \|v\|_{0,\partial K}^2 \right)^{\frac{1}{2}}. \tag{4.51}$$

By using the trace inequality (3.28) and the standard inverse inequality (2.14), we have

$$\|v\|_{0,\partial K}^2 \le C(h^{-1}\|v\|_{0,K}^2 + h\|\nabla v\|_{1,K}^2) \le Ch^{-1}\|v\|_{0,K}^2. \qquad (4.52)$$

This completes the proof of the lemma. □

Lemma 4.5 *Let I_3 be as in (4.46). There exists a constant C such that*

$$|I_3| \le C(h + h^{\frac{1-\beta}{2}} + a_2 h^{-1})(h^{-2}\|\phi\|_{0;h}^2 + \|\phi\|_{1;h}^2)^{\frac{1}{2}} |||(v,\mu)|||. \qquad (4.53)$$

Proof: Write

$$\phi((\mu + Q_h\xi) = \phi(\mu + \xi) + \phi(Q_h - I)\xi, \qquad (4.54)$$

where I is the identity operator. The trace values of $\mu + \xi$ only differ in sign on a common edge. Therefore,

$$\left|\sum_{K\in\mathcal{T}_h}\int_{\partial K}\phi(\mu+\xi)\,ds\right| \le \sum_{K\in\mathcal{T}_h}\int_{\partial K,\mathbf{b}\cdot n_K\ge 0} |[\![\phi]\!](\mu + (\mathbf{b}\cdot n_K v)|. \qquad (4.55)$$

Write

$$\mu + \mathbf{b}\cdot n_K v = \mu - (a\nabla v - \mathbf{b}v)\cdot n_K + a\nabla v\cdot n_K.$$

By using the Cauchy-Schwarz inequality and trace inequality (3.28), we obtain

$$\begin{aligned}\sum_{K\in\mathcal{T}_h}\int_{\partial K} |[\![\phi]\!](\mu - (a\nabla v - \mathbf{b}v)\cdot n_K)|\,ds &\le C\left(h^{-\beta}\sum_{K\in\mathcal{T}_h}\int_{\partial K}\phi^2\right)^{\frac{1}{2}} |||(v,\mu)||| \\ &\le Ch^{-\frac{\beta}{2}}(h^{-1}\|\phi\|_{0;h}^2 + h\|\phi\|_{1;h}^2)^{\frac{1}{2}} |||(v,\mu)|||,\end{aligned} \qquad (4.56)$$

and

$$\begin{aligned}\sum_{K\in\mathcal{T}_h}\int_{\partial K}|[\![\phi]\!]a\nabla v\cdot n_K|\,ds &\le Ca_2\left(\sum_{K\in\mathcal{T}_h}\int_{\partial K}\phi^2\right)^{\frac{1}{2}}\left(\sum_{K\in\mathcal{T}_h}\|\nabla v\|_{0,\partial K}^2\right)^{\frac{1}{2}}\\ &\le Ca_2h^{-\frac{1}{2}}(h^{-1}\|\phi\|_{0;h}^2+h\|\phi\|_{1;h}^2)^{\frac{1}{2}}\|\nabla v\|_{0;h}\\ &\le Ca_2h^{-\frac{3}{2}}(h^{-1}\|\phi\|_{0;h}^2+h\|\phi\|_{1;h}^2)^{\frac{1}{2}}|\!|\!|(v,\mu)|\!|\!|.\end{aligned}\tag{4.57}$$

In the last step, we used the standard inverse inequality (2.14) to obtain

$$\|\nabla v\|_{0;h}\le Ch^{-1}\|v\|_{0;h}\le Ch^{-1}|\!|\!|(v,\mu)|\!|\!|.$$

We now establish an estimate for the last term in (4.54). Similar to (4.55), we have

$$\left|\sum_{K\in\mathcal{T}_h}\int_{\partial K}\phi(Q_h-I)\xi\,ds\right|\le\sum_{K\in\mathcal{T}_h}\int_{\partial K,\mathbf{b}\cdot n_K\ge 0}|[\![\phi]\!](Q_h-I)(\mathbf{b}\cdot n_Kv)|.\tag{4.58}$$

Using the fact that if v is a piecewise polynomial of degree $r\ge 1$ and $\mathbf{b}$ is smooth, then

$$\|(Q_h-I)(\mathbf{b}\cdot n_Kv)\|_{0,\partial K}\le Ch\|v\|_{0,\partial K}.\tag{4.59}$$

Then by using the Cauchy-Schwarz inequality and the trace inequality (3.28), we obtain

$$\left|\sum_{K\in\mathcal{T}_h}\int_{\partial K}\phi(Q_h-I)\xi\,ds\right|\le Ch^{\frac{1}{2}}(h^{-1}\|\phi\|_{0;h}^2+h\|\phi\|_{1;h}^2)^{\frac{1}{2}}|\!|\!|(v,\mu)|\!|\!|.\tag{4.60}$$

Finally, the estimate in (4.53) is obtained by combining (4.56), (4.57) and (4.60). □

Lemma 4.6 *Let I_4 be as in (4.46). There exists a constant C such that*

$$|I_4| \le C(1 + h^{\frac{\beta+1}{2}}) |\!|\!|(\phi, \eta)|\!|\!| \; |\!|\!|(v, \mu)|\!|\!|. \tag{4.61}$$

Proof: We rewrite

$$\mu + Q_h\xi - a\nabla v \cdot n_K = \mu + \xi - a\nabla v \cdot n_K + (Q_h - I)\xi.$$

By the definition of z and using the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}
&\alpha h^{\beta} \left| \sum_{K\in\mathcal{T}_h} \int_{\partial K} z(\eta - (a\nabla\phi - \mathbf{b}\phi)\cdot n_K)(\mu + \xi - a\nabla v \cdot n_K)\, ds \right| \\
&\le \alpha h^{\beta} \sum_{K\in\mathcal{T}_h} \left| \int_{\partial K, \mathbf{b}\cdot n_K \ge 0} (\eta - (a\nabla\phi - \mathbf{b}\phi)\cdot n_K)(\mu + -(a\nabla v - \mathbf{b}v)\cdot n_K) \right| \\
&\le |\!|\!|(\phi, \eta)|\!|\!| \; |\!|\!|(v, \mu)|\!|\!|.
\end{aligned} \tag{4.62}$$

Similar to (4.58), we can use (4.59) to obtain

$$\alpha h^{\beta} \left| \sum_{K\in\mathcal{T}_h} \int_{\partial K} z(\eta - (a\nabla\phi - \mathbf{b}\phi)\cdot n_K)(Q_h - I)\xi\, ds \right| \le C h^{\frac{\beta+1}{2}} |\!|\!|(\phi, \eta)|\!|\!| \; |\!|\!|(v, \mu)|\!|\!|. \tag{4.63}$$

Combining (4.62) and (4.63), we obtain (4.61). □

Combining Lemma 4.3-4.6, we obtain the following first error estimate for the finite element solution (u_h, λ_h).

Lemma 4.7 *Assume that the solution u of (4.1) is sufficiently smooth such that $u \in X \cap H^1(\Omega)$. Let $\lambda \in Y$ be given by $\lambda|_{\partial K} = \kappa\nabla u \cdot n_K$. If $(u_h, \lambda_h) \in X_h \times Y_h$ is the stabilized discontinuous finite element approximation obtained by solving (4.34), then* *there is a constant C such that*

$$\begin{aligned}|||(u-u_h,\lambda-\lambda_h)||| &\le C(1+h^{\frac{\beta+1}{2}})|||(u-\Pi_h u,\lambda-Q_h\lambda)|||\\ &+C(1+h^{\frac{1-\beta}{2}}+a_2h^{-1})\left(\|u-\Pi_h u\|_{1;h}^2+h^{-2}\|u-\Pi_h u\|_{0;h}^2)\right)^{\frac{1}{2}}\\ &+Ch^{-\frac{1}{2}}\|\lambda-Q_h\lambda\|_{0;\partial h}.\end{aligned} \tag{4.64}$$

Assume that the finite element partition $\mathcal{T}_h$ is regular. Then there exists a constant C such that the interpolation error estimate

$$h^{-1}\|u-\Pi_h u\|_{0;h}+\|u-\Pi_h u\|_{1;h}+h\|u-\Pi_h u\|_{2;h}\le Ch^m\|u\|_{m+1;h} \tag{4.65}$$

holds for $0\le m\le r$ and $u\in\prod_{K\in\mathcal{T}_h}H^{m+1}(K)$. Next, we provide an estimate for $\|\lambda-Q_h\lambda\|_{0,\partial K}$ if $\lambda=(a\nabla w-\mathbf{b}w)\cdot n_K$ for some smooth function w defined on K.

Lemma 4.8 *Let $K\in\mathcal{T}_h$ be an element in the finite element partition $\mathcal{T}_h$. Let $\lambda=(a\nabla w-\mathbf{b})\cdot n_K$. Then there exists a constant $C>0$ such that*

$$\|\lambda-Q_h\lambda\|_{0,\partial K}\le C(a_2+h)h^{\ell-\frac{1}{2}}\|w\|_{\ell+1,K},\qquad 0\le\ell\le s+1.$$

Proof: Let $\Pi_h w\in P_{s+1}(K)$ be the L^2-projection of w in the polynomial space $P_{s+1}(K)$ and

$$\tilde{\lambda}=\lambda_1+\lambda_2=a\nabla(\Pi_h w)\cdot n_K-\Pi_h(\mathbf{b}w)\cdot n_K, \tag{4.66}$$

where λ_i are defined accordingly for $i=1,2$. Then,

$$\lambda-Q_h\lambda=(\lambda-\tilde{\lambda})+(\tilde{\lambda}-Q_h\tilde{\lambda})+Q_h(\tilde{\lambda}-\lambda). \tag{4.67}$$

It follows that

$$\begin{aligned}\|\lambda - Q_h\lambda\|_{0,\partial K} &\le \|\lambda - \tilde{\lambda}\|_{0,\partial K} + \|\tilde{\lambda} - Q_h\tilde{\lambda}\|_{0,\partial K} + \|Q_h(\tilde{\lambda} - \lambda)\|_{0,\partial K} \\ &\le 2\|\lambda - \tilde{\lambda}\|_{0,\partial K} + \|\tilde{\lambda} - Q_h\tilde{\lambda}\|_{0,\partial K}. \qquad (4.68)\end{aligned}$$

To estimate the first term, we see that

$$\|\lambda - \tilde{\lambda}\|^2_{0,\partial K} \le 2\|a\nabla(w - \Pi_h w)\cdot n_K\|^2_{0,\partial K} + 2\|(\mathbf{b}w - \Pi_h\mathbf{b}w)\cdot n_K\|^2_{0,\partial K}. \qquad (4.69)$$

By using the trace inequality (3.28) and the standard interpolation error estimate (4.65), we have

$$\begin{aligned}\|\lambda - \tilde{\lambda}\|^2_{0,\partial K} &\le Ca_2^2\left(h^{-1}\|\nabla(w - \Pi_h w)\|^2_{0,K} + h\|D^2(w - \Pi_h w)\|^2_{0,K}\right) \\ &+ C\left(h^{-1}\|\mathbf{b}w - \Pi_h\mathbf{b}w\|^2_{0,K} + h\|\mathbf{b}w - \Pi_h\mathbf{b}w\|^2_{0,K}\right) \qquad (4.70) \\ &\le Ca_2^2h^{2\ell-1}\|w\|^2_{\ell+1,K} + Ch^{2\ell+1}\|w\|^2_{\ell+1,K},\end{aligned}$$

for $1 \le \ell \le s+1$. As for the second term of (4.68), we again use (4.65) to obtain

$$\|\tilde{\lambda} - Q_h\tilde{\lambda}\|_{0,\partial K} \le C\sum_{i=1}^{m(K)}(h^{\ell}\|\lambda_1\|_{\ell,e_{i,K}} + h^{\ell+1}\|\lambda_2\|_{\ell+1,e_{i,K}}), \qquad 0 \le \ell \le s+1, \quad (4.71)$$

where $\|\cdot\|_{\ell,e_{i,K}}$ is the norm in the Sobolev space $H^{\ell}(e_{i,K})$. To estimate each norm $\|\tilde{\lambda}\|_{\ell,e_{i,K}}$, we assume that the coefficients a and $\mathbf{b}$ are sufficiently smooth on the element K. Using the trace inequality (3.28) and the relation (4.66), we obtain

$$\|\lambda_1\|^2_{\ell,e_{i,K}} \le Ca_2^2\left(h^{-1}\|\Pi_h w\|^2_{\ell+1,K} + h|\Pi_h w|^2_{\ell+2,K}\right), \qquad (4.72)$$

where $|\Pi_h w|_{ell+1,K}$ is the Sobolev semi-norm of $\Pi_h w$ on the element K. It follows from the standard inverse inequality (2.14) that

$$h|\Pi_h w|^2_{\ell+2,K} \le h^{-1}||\Pi_h w||^2_{\ell+1,K}. \tag{4.73}$$

Thus, substituting (4.73) into (4.72), we have

$$||\lambda_1||^2_{\ell,e_i,K} \le Ca_2^2 h^{-1}||\Pi_h w||^2_{\ell+1,K} \le Ca_2^2 h^{-1}||w||^2_{\ell+1,K},$$

where we have used the boundedness of the L^2-projection Π_h in $H^{\ell+1}(K)$. Similarly,

$$||\lambda_2||^2_{\ell+1,e_i,K} \le Ca_2^2\left(h^{-1}||\Pi_h w||^2_{\ell+1,K} + h|\Pi_h w|^2_{\ell+2,K}\right) \le Ch^{-1}||w||^2_{\ell+1,K}. \tag{4.74}$$

Substituting (4.74) into (4.71), we obtain

$$||\tilde{\lambda} - Q_h\tilde{\lambda}||_{0,\partial K} \le C(a_2 + h)h^{\ell-\frac{1}{2}}||w||_{\ell+1,K}, \qquad 0 \le \ell \le s+1. \tag{4.75}$$

Now combining (4.68) with (4.70) and (4.75), we obtain

$$||\lambda - Q_h\lambda||_{0,\partial K} \le C(a_2 + h)h^{\ell-\frac{1}{2}}||w||_{\ell+1,K}, \qquad 0 \le \ell \le s+1, \tag{4.76}$$

which completes the proof. □

We are now in a position to prove the main result of this section regarding the accuracy of the finite element method (4.34).

Theorem 4.2 *Under the assumptions of Lemma 4.7, there exists a constant C such* *that*

$$\begin{aligned}
&|||(u-u_h,\lambda-\lambda_h)||| \\
&\leq C\left(h^{1-\beta}+h+h^{\beta+1}+a_2^2h^{-2}+a_2^2h^{\beta-1}\right)^{\frac{1}{2}}h^m\|u\|_{m+1;h} \\
&+C(1+h^{\frac{\beta+1}{2}})(1+a_2h^{-1})h^\ell\|u\|_{\ell+1;h},
\end{aligned} \tag{4.77}$$

for $1\leq m\leq r$ *and* $1\leq \ell\leq s+1$*, provided that the solution* u *is sufficiently smooth.*

Proof: It is easy to see that

$$\left(\sum_K\int_K a\nabla(u-\Pi_hu)\nabla(u-\Pi_hu)\right)^{\frac{1}{2}}\leq Ca_2^{\frac{1}{2}}\|u-\Pi_hu\|_{1;h},$$

and

$$\left(\sum_K\int_K(\frac{1}{2}\nabla\cdot\mathbf{b}+c)(u-\Pi_hu)^2\right)^{\frac{1}{2}}\leq C\|u-\Pi_hu\|_{0;h}.$$

By using the trace inequality (3.28), we obtain

$$\left(\frac{1}{2}\sum_{e\in\mathcal{E}_h}\int_e|\mathbf{b}\cdot n_e|[\![u-\Pi_hu]\!]^2\right)^{\frac{1}{2}}\leq C(h^{-1}\|u-\Pi_hu\|_{0;h}^2+h\|u-\Pi_hu\|_{1;h}^2)^{\frac{1}{2}},$$

and

$$\begin{aligned}
&\left(\alpha h^\beta\sum_K\int_{\partial K,\mathbf{b}\cdot n_K\geq 0}(\lambda-Q_h\lambda-(a\nabla(u-\Pi_hu)-\mathbf{b}(u-\Pi_hu))\cdot n_K)^2\right)^{\frac{1}{2}} \\
&\leq Ch^{\frac{\beta}{2}}(\|\lambda-Q_h\lambda\|_{0;h}^2+a_2^2(h^{-1}\|u-\Pi_hu\|_{1;h}+h\|u-\Pi_hu\|_{2;h}) \\
&+(h^{-1}\|u-\Pi_hu\|_{0;h}+h\|u-\Pi_hu\|_{1;h}))^{\frac{1}{2}}.
\end{aligned}$$

Therefore, there exists a constant C such that

$$\begin{aligned} \| (u - \Pi_h u, \lambda - Q_h\lambda) \| & \leq C(h + h^{\beta+1})^{\frac{1}{2}} h^{-1} \|u - \Pi_h u\|_{0;h} \\ & + C(a_2 + h + a_2^2 h^{\beta-1} + h^{\beta+1})^{\frac{1}{2}} \|u - \Pi_h u\|_{1;h} \\ & + C(a_2^2 h^{\beta-1})^{\frac{1}{2}} h \|u - \Pi_h u\|_{2;h} \\ & + C h^{\frac{\beta}{2}} \|\lambda - Q_h \lambda\|_{0;h}. \end{aligned} \tag{4.78}$$

Omitting all higher order terms, we obtain from Lemma 4.7, Lemma 4.8 and the interpolation error estimate (4.65) that

$$\begin{aligned} & \| (u - u_h, \lambda - \lambda_h) \| \\ & \leq C \left(h^{1-\beta} + h + h^{\beta+1} + a_2^2 h^{-2} + a_2^2 h^{\beta-1} \right)^{\frac{1}{2}} h^m \|u\|_{m+1;h} \\ & + (1 + h^{\frac{\beta+1}{2}})(1 + a_2 h^{-1}) h^{\ell} \|u\|_{\ell+1;h}, \end{aligned} \tag{4.79}$$

for $1 \leq m \leq r$ and $1 \leq \ell \leq s + 1$. This completes the proof of the theorem. □

We can see from (4.79) that in order to get the best estimate, we need to choose $\beta = 0$. Using the assumption that $a_2 \ll h$, the main estimate in (4.77) can be written as follows:

$$\| (u - u_h, \lambda - \lambda_h) \| \leq C(h^{m+\frac{1}{2}} \|u\|_{m+1;h} + h^{\ell} \|u\|_{\ell+1;h}). \tag{4.80}$$

Chapter 5

ANALYSIS OF THE MATRIX PROBLEMS

In this chapter, we construct the stiffness matrices for stabilized discontinuous finite element schemes. Symmetric and non-symmetric formulations for the elliptic problem as well as the non-symmetric formulation for the convection-diffusion problem are discussed. We also discuss briefly the conjugate gradient method for the resulting linear system.

5.1 Matrix of the symmetric formulation for elliptic problems

Recall that the stabilized symmetric formulation (3.27) of the elliptic problem (1.1)-(1.2) seeks (u_h, λ_h) in $X_h \times Y_h$ such that

$$\mathcal{L}^{(ss)}(u_h, \lambda_h; v, \mu) = \ell(v, \mu), \qquad \forall (v, \mu) \in X_h \times Y_h, \tag{5.1}$$

where

$$\begin{aligned} \mathcal{L}^{(ss)}(u, \lambda; v, \mu) &= \sum_{K \in \mathcal{T}_h} \int_K \kappa \nabla u \nabla v \, dK - \sum_{K \in \mathcal{T}_h} \int_{\partial K} \lambda v \, ds - \sum_{K \in \mathcal{T}_h} \int_{\partial K} u \mu \, ds \\ &\quad -\alpha h \sum_{K \in \mathcal{T}_h} \int_{\partial K} (\lambda - \kappa \nabla u \cdot n_K)(\mu - \kappa \nabla v \cdot n_K) \, ds, \end{aligned} \tag{5.2}$$

and

$$\ell(v, \mu) = \sum_{K \in \mathcal{T}_h} \left(\int_K f(x) v(x) dK - \int_{\partial K \cap \partial \Omega} g(x) \mu(x) ds \right). \tag{5.3}$$

Finite element spaces are defined as

$$\begin{aligned} X_h &= \{v:\ v|_K \in P_r(K)\}, \\ Y_h &= \left\{\mu \in Y:\ \mu|_{\partial K} \in \prod_{i=1}^{m(K)} P_s(e_{i,K}), \forall K \in \mathcal{T}_h\right\}, \end{aligned} \tag{5.4}$$

where $P_r(K)$ and $P_s(e_{i,K})$ stand for the space of polynomials of degree no more than $r \geq 1$ and $s \geq 0$ on K and its boundary piece $e_{i,K}$, respectively, and

$$Y = \left\{\mu \in \prod_{K \in \mathcal{T}_h} L^2(\partial K), \mu|_{\partial K_1} + \mu|_{\partial K_2} = 0 \quad \text{on } \partial K_1 \cap \partial K_2\right\}. \tag{5.5}$$

The formulation (5.1) can be written as the following:

$$\begin{aligned} \sum_K \int_K \kappa \nabla u_h \nabla v dK - \sum_K \int_{\partial K} \lambda_h v ds & \\ +\alpha h \sum_K \int_{\partial K} (\lambda_h - \kappa \nabla u_h \cdot n_K) \kappa \nabla v \cdot n_K ds &= \sum_K \int_K f v dK, \\ -\sum_K \int_{\partial K} u_h \mu ds - \alpha h \sum_K \int_{\partial K} (\lambda_h - \kappa \nabla u_h \cdot n_K) \mu ds &= -\sum_K \int_{\partial K \cap \partial \Omega} g \mu ds, \end{aligned} \tag{5.6}$$

for all $(v, \mu) \in X_h \times Y_h$.

Assume that the basis functions of X_h and Y_h are given by $\{v_i\}_{i=1}^{N_1}$ and $\{\mu_j\}_{j=1}^{N_2}$ respectively. We also write $u_h = \sum_{i=1}^{N_1} u^i v_i$ and $\lambda_h = \sum_{j=1}^{N_2} \lambda^j \mu_j$. Then (5.6) is equivalent to the following linear system:

$$\begin{aligned} &\sum_K \left(\sum_{i=1}^{N_1} u^i (\kappa \nabla v_i, \nabla v_k)_K - \sum_{j=1}^{N_2} \lambda^j (\mu_j, v_k)_{\partial K} \right. \\ &\left. +\alpha h \left(\sum_{j=1}^{N_2} \lambda^j \mu_j - \sum_{i=1}^{N_1} u^i \kappa \nabla v_i \cdot n_K, \kappa \nabla v_k \cdot n_K \right)_{\partial K} \right) = \sum_K (f, v_k)_K, \end{aligned} \tag{5.7}$$

$$
\begin{aligned}
&-\sum_K \left(\sum_{i=1}^{N_1} u^i (v_i, \mu_l)_{\partial K} + \alpha h \left(\sum_{j=1}^{N_2} \lambda^j \mu_j - \sum_{i=1}^{N_1} u^i \kappa \nabla v_i \cdot n_K, \mu_l \right)_{\partial K} \right) \\
&= -\sum_K (g, \mu_l)_{\partial K \cap \partial \Omega},
\end{aligned}
\tag{5.8}
$$

for $k = 1, 2, \cdots, N_1$ and $l = 1, 2, \cdots, N_2$. $(\cdot, \cdot)_R$ denotes the L^2 inner product in the region R. The linear system (5.7)-(5.8) can be further written as

$$
\begin{bmatrix} B & C \\ D & E \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \Lambda \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ -\mathbf{g} \end{bmatrix}, \qquad \text{i.e., } A\mathbf{x} = \mathbf{b}, \tag{5.9}
$$

where

$$
\begin{aligned}
&\mathbf{u} = \left(u^1, u^2, \cdots, u^{N_1}\right)^T, \\
&\Lambda = \left(\lambda^1, \lambda^2, \cdots, \lambda^{N_2}\right)^T, \\
&\mathbf{f} = (f_1, f_2, \cdots, f_{N_1})^T, \ f_k = \textstyle\sum_K (f, v_k)_K, \\
&\mathbf{g} = (g_1, g_2, \cdots, g_{N_2})^T, \ g_l = \textstyle\sum_K (g, \mu_l)_{\partial K \cap \partial \Omega},
\end{aligned}
\tag{5.10}
$$

where T denotes the transpose operator. The block matrices in (5.9) are defined as follows:

$$
\begin{aligned}
B &= (b_{ki})_{N_1 \times N_1}, \ b_{ki} = \sum_K \left((\kappa \nabla v_i, \nabla v_k)_K - \alpha h \left(\kappa \nabla v_i \cdot n_K, \kappa \nabla v_k \cdot n_K \right)_{\partial K} \right), \\
C &= (c_{kj})_{N_1 \times N_2}, \ c_{kj} = \sum_K \left(-(v_k, \mu_j)_{\partial K} + \alpha h (\mu_j, \kappa \nabla v_k \cdot n_K)_{\partial K} \right), \\
D &= (d_{li})_{N_2 \times N_1}, \ d_{li} = \sum_K \left(-(\mu_l, v_i)_{\partial K} + \alpha h (\kappa \nabla v_i \cdot n_K, \mu_l)_{\partial K} \right), \\
E &= (e_{lj})_{N_2 \times N_2}, \ e_{lj} = \sum_K -\alpha h (\mu_l, \mu_j)_{\partial K}.
\end{aligned}
\tag{5.11}
$$

It can be seen easily from (5.11) that B and E are symmetric and $D^T = C$. Therefore the coefficient matrix in (5.9) is symmetric.

Now we discuss the construction of the coefficient matrix in (5.9) in the two dimensional case. We choose $r = 1$ and $s = 0$ in (5.4). Therefore the finite element space X_h is the space of piecewise linear functions on the elements and Y_h is space of piecewise constant functions on the boundary pieces of the elements. The stiffness matrix can be computed locally on each element and local stiffness matrices can be combined to obtain the global coefficient matrix. The one and only relation that connects elements together is given in the definition of Y in (5.5):

$$\mu|_{\partial K_1} + \mu|_{\partial K_2} = 0 \quad \text{on } \partial K_1 \cap \partial K_2. \tag{5.12}$$

In order to impose this condition, we artificially put signs on two sides of an edge. For each common edge of two adjacent element, we define one side to be positive and the other side to be negative. In the local matrices, we need to multiply the sign of an edge on all entries corresponding to that edge.

Let K be any triangle in the partition. We denote the vertices of K by (x_i, y_i) and the edges by e_i where $i = 1, 2, 3$. The length of edge e_i is denoted by h_i. We denote by s_i the sign defined on its edge e_i. The local basis function v_i is the linear function that takes on the value 1 at vertex (x_i, y_i) and 0 at the other two vertices. The local basis function μ_i is the piecewise constant function that takes on the value 1 on edge e_i and 0 on the other two edges. The local stiffness matrix on K can be obtained from (5.9) and (5.11) as follows:

$$A^{\text{loc}} = \begin{bmatrix} B^{\text{loc}} & C^{\text{loc}} \\ C^{\text{loc}^T} & E^{\text{loc}} \end{bmatrix}, \tag{5.13}$$

Where

$$\begin{aligned}
B^{\text{loc}} &= (b_{ij})_{3\times 3},\ b_{ij} = (\kappa\nabla v_i, \nabla v_j)_K - \alpha h\,(\kappa\nabla v_i\cdot n_K, \kappa\nabla v_j\cdot n_K)_{\partial K}\,, \\
C^{\text{loc}} &= (c_{ij})_{3\times 3},\ c_{ij} = (-(v_i, \mu_j)_{\partial K} + \alpha h(\mu_j, \kappa\nabla v_i\cdot n_K)_{\partial K})\, s_j, \\
E^{\text{loc}} &= (e_{ij})_{3\times 3},\ e_{ij} = -\alpha h(\mu_i, \mu_j)_{\partial K} s_i s_j.
\end{aligned} \tag{5.14}$$

We can easily see that $(\mu_i, \mu_j)_{\partial K}$ is non-zero only when $i = j$. In fact,

$$E^{\text{loc}} = -\alpha h \ \text{diag}\{h_1, h_2, h_3\} \tag{5.15}$$

because $s_i^2 = 1$. This also shows that the global matrix E in (5.9) is a diagonal matrix with negative diagonal entries. Therefore E is symmetric and negative definite (and thus invertible). The global coefficient matrix A in (5.9) can then be obtained by putting all local matrices on all elements together. In fact, it is not necessary to build up the global matrix as long as we know the global indices of all local basis functions.

The global coefficient matrix A in (5.9) is not positive definite because the matrix E is negative definite. Therefore we cannot solve the linear system by efficient linear solvers like the conjugate gradient method. The following is a discussion on how we transform the problem into a linear system with symmetric and positive definite coefficient matrix.

We showed in Lemma 3.4 that, for all $(v, \mu) \in X_h \times Y_h$,

$$\mathcal{L}^{(ss)}(v, \mu; v, -\mu) \geq C \sum_{K\in\mathcal{T}_h} \left(\int_K \kappa\nabla v\nabla v dK + \alpha h \int_{\partial K} \mu^2 ds \right) \geq 0 \tag{5.16}$$

for constant $\alpha \leq \alpha_0$. This implies that, in the matrix form,

$$[\mathbf{u}^T, -\Lambda^T] \begin{bmatrix} B & C \\ C^T & E \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \Lambda \end{bmatrix} \geq 0 \tag{5.17}$$

for any vectors $\mathbf{u} \in \mathbb{R}^{N_1}$ and $\Lambda \in \mathbb{R}^{N_2}$. A direct calculation shows that

$$\mathbf{u}^T B\mathbf{u} - \Lambda^T E\Lambda \geq 0, \qquad \forall \mathbf{u} \in \mathbb{R}^{N_1}, \Lambda \in \mathbb{R}^{N_2}. \tag{5.18}$$

This also shows that B should be at least non-negative definite.

On the other hand, the matrix problem in (5.9) can be written as:

$$B\mathbf{u} + C\Lambda = \mathbf{f}, \tag{5.19}$$

$$C^T\mathbf{u} + E\Lambda = -\mathbf{g}. \tag{5.20}$$

Solving Λ from (5.20) and substituting into (5.19), we obtain the following linear system for $\mathbf{u}$:

$$(B - CE^{-1}C^T)\mathbf{u} = \mathbf{f} + CE^{-1}\mathbf{g}. \tag{5.21}$$

Matrix $B - CE^{-1}C^T$ is symmetric by the symmetry of B and E. It is also non-negative definite because B is non-negative definite and E is negative definite. The uniqueness result that we obtained in Theorem 3.2 implies that the matrix $B - CE^{-1}C^T$ is non-singular. Therefore the coefficient matrix in (5.21) is symmetric and positive definite. This will be the linear system that we use for the symmetric formulation.

5.2 Matrix of the non-symmetric formulation for elliptic problems

Recall that the stabilized non-symmetric formulation (3.21) of the elliptic problem (1.1)-(1.2) seeks (u_h, λ_h) in $X_h \times Y_h$ such that

$$\mathcal{L}^{(sn)}(u_h, \lambda_h; v, \mu) = \ell(v, \mu), \qquad \forall v \in X_h, \mu \in Y_h, \tag{5.22}$$

where

$$\begin{aligned} \mathcal{L}^{(sn)}(u, \lambda; v, \mu) &= \sum_{K\in\mathcal{T}_h} \int_K \kappa\nabla u \nabla v \; dK - \sum_{K\in\mathcal{T}_h} \int_{\partial K} \lambda v \; ds + \sum_{K\in\mathcal{T}_h} \int_{\partial K} u\mu \; ds \\ &+\alpha h \sum_{K\in\mathcal{T}_h} \int_{\partial K} (\lambda - \kappa\nabla u \cdot n_K)(\mu - \kappa\nabla v \cdot n_K) \; ds, \end{aligned} \tag{5.23}$$

and

$$\ell(v, \mu) = \sum_{K\in\mathcal{T}_h} \left(\int_K f(x)v(x)dK + \int_{\partial K\cap\partial\Omega} g(x)\mu(x)ds \right). \tag{5.24}$$

The formulation (5.22) can be written as the following:

$$\begin{aligned} \sum_K \int_K \kappa\nabla u_h \nabla v \; dK - \sum_K \int_{\partial K} \lambda_h v \; ds & \\ -\alpha h \sum_K \int_{\partial K} (\lambda_h - \kappa\nabla u_h \cdot n_K)\kappa\nabla v \cdot n_K \; ds &= \sum_K \int_K fv \; dK, \\ \sum_K \int_{\partial K} u_h\mu \; ds + \alpha h \sum_K \int_{\partial K} (\lambda_h - \kappa\nabla u_h \cdot n_K)\mu \; ds &= \sum_K \int_{\partial K\cap\partial\Omega} g\mu \; ds, \end{aligned} \tag{5.25}$$

for all $v \in X_h$ and $\mu \in Y_h$.

By using the same notation as in the previous section, (5.25) can be written as the following linear system:

$$\begin{aligned}
&\sum_K \left(\sum_{i=1}^{N_1} u^i(\kappa\nabla v_i, \nabla v_k)_K - \sum_{j=1}^{N_2} \lambda^j(\mu_j, v_k)_{\partial K} \right. \\
&\left. -\alpha h \left(\sum_{j=1}^{N_2} \lambda^j \mu_j - \sum_{i=1}^{N_1} u^i \kappa\nabla v_i \cdot n_K, \kappa\nabla v_k \cdot n_K \right)_{\partial K} \right) = \sum_K (f, v_k)_K,
\end{aligned} \tag{5.26}$$

$$\begin{aligned}
&\sum_K \left(\sum_{i=1}^{N_1} u^i(v_i, \mu_l)_{\partial K} + \alpha h \left(\sum_{j=1}^{N_2} \lambda^j \mu_j - \sum_{i=1}^{N_1} u^i \kappa\nabla v_i \cdot n_K, \mu_l \right)_{\partial K} \right) \\
&= \sum_K (g, \mu_l)_{\partial K \cap \partial\Omega},
\end{aligned} \tag{5.27}$$

for $k = 1, 2, \cdots, N_1$ and $l = 1, 2, \cdots, N_2$. The linear system (5.26)-(5.27) is again written as

$$\begin{bmatrix} B & C \\ D & E \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \Lambda \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix}, \qquad \text{i.e., } A\mathbf{x} = \mathbf{b}. \tag{5.28}$$

The block matrices in (5.28) are defined as follows:

$$\begin{aligned}
B &= (b_{ki})_{N_1 \times N_1}, \; b_{ki} = \sum_K \left((\kappa\nabla v_i, \nabla v_k)_K - \alpha h \left(\kappa\nabla v_i \cdot n_K, \kappa\nabla v_k \cdot n_K \right)_{\partial K} \right), \\
C &= (c_{kj})_{N_1 \times N_2}, \; c_{kj} = \sum_K \left(-(v_k, \mu_j)_{\partial K} - \alpha h(\mu_j, \kappa\nabla v_k \cdot n_K)_{\partial K} \right), \\
D &= (d_{li})_{N_2 \times N_1}, \; d_{li} = \sum_K \left((\mu_l, v_i)_{\partial K} - \alpha h(\kappa\nabla v_i \cdot n_K, \mu_l)_{\partial K} \right), \\
E &= (e_{lj})_{N_2 \times N_2}, \; e_{lj} = \sum_K \alpha h(\mu_l, \mu_j)_{\partial K}.
\end{aligned} \tag{5.29}$$

It can be seen from (5.29) that $D^T \neq C$. Therefore the coefficient matrix in (5.28) is non-symmetric.

The local stiffness matrix on K can be obtained from (5.28) and (5.29) as follows:

$$A^{\mathrm{loc}} = \begin{bmatrix} B^{\mathrm{loc}} & C^{\mathrm{loc}} \\ D^{\mathrm{loc}} & E^{\mathrm{loc}} \end{bmatrix}, \tag{5.30}$$

Where

$$\begin{aligned}
B^{\mathrm{loc}} &= (b_{ij})_{3\times 3},\ \ b_{ij} = \left(\kappa\nabla v_i, \nabla v_j\right)_K - \alpha h \left(\kappa\nabla v_i \cdot n_K, \kappa\nabla v_j \cdot n_K\right)_{\partial K}, \\
C^{\mathrm{loc}} &= (c_{ij})_{3\times 3},\ \ c_{ij} = \left(-(v_i, \mu_j)_{\partial K} - \alpha h(\mu_j, \kappa\nabla v_i \cdot n_K)_{\partial K}\right) s_j, \\
D^{\mathrm{loc}} &= (d_{ij})_{3\times 3},\ \ c_{ij} = \left((v_j, \mu_i)_{\partial K} - \alpha h(\mu_i, \kappa\nabla v_j \cdot n_K)_{\partial K}\right) s_i, \\
E^{\mathrm{loc}} &= (e_{ij})_{3\times 3},\ \ e_{ij} = \alpha h(\mu_i, \mu_j)_{\partial K} s_i s_j.
\end{aligned} \tag{5.31}$$

The global coefficient matrix A in (5.28) can be obtained by putting all local matrices on all elements together. Again, it is not necessary to build up the global matrix as long as we know the global indices of all local basis functions. This will save memory space in the implementation and execution of our code.

5.3 Matrix for the convection-diffusion problem

For the convection-diffusion problem (4.1), the stabilized discontinuous finite element method (4.34) seeks $(u_h, \lambda_h) \in X_h \times Y_h$ satisfying

$$\Phi(u_h, \lambda_h; v, \mu) = \ell(v, \mu), \qquad \forall v \in X_h,\ \mu \in Y_h, \tag{5.32}$$

where

$$\begin{aligned}\Phi(u,\lambda;v,\mu) &= \sum_{K\in\mathcal{T}_h}\int_K (a\nabla u-\mathbf{b}u)\nabla v+cuv\,dK-\sum_{K\in\mathcal{T}_h}\int_{\partial K}(\lambda v-u\mu)\,ds\\ &+ \alpha h^{\beta}\sum_{K\in\mathcal{T}_h}\int_{\partial K} z(\lambda-(a\nabla u-\mathbf{b}u)\cdot n_K)(\mu-a\nabla v\cdot n_K)\,ds,\end{aligned}\tag{5.33}$$

and

$$\ell(v,\mu)=\sum_{K\in\mathcal{T}_h}\left(\int_K fv\,dK+\int_{\partial K\cap\partial\Omega} g\mu\,ds\right).\tag{5.34}$$

We choose $\beta=0$ in our computation because, as we showed in (4.80), it gives the best estimate. The notation z denotes the “inflow-outflow” indicator. Then (5.32) is equivalent to the following system:

$$\begin{aligned}\sum_K\int_K (a\nabla u_h-\mathbf{b}u_h)\nabla v+cuv\,dK-\sum_K\int_{\partial K}\lambda_h v\,ds&\\ -\alpha\sum_K\int_{\partial K} z(\lambda_h-(a\nabla u_h-\mathbf{b}u_h)\cdot n_K)a\nabla v\cdot n_K\,ds &= \sum_K\int_K fv\,dK,\\ \sum_K\int_{\partial K}(u_h\mu+\alpha z(\lambda_h-(a\nabla u_h-\mathbf{b}u_h)\cdot n_K)\mu)\,ds &= \sum_K\int_{\partial K\cap\partial\Omega} g\mu\,ds,\end{aligned}\tag{5.35}$$

for all $v\in X_h$ and $\mu\in Y_h$.

Assume that the basis functions of X_h and Y_h are given by $\{v_i\}_{i=1}^{N_1}$ and $\{\mu_j\}_{j=1}^{N_2}$ respectively. We write $u_h=\sum_{i=1}^{N_1}u^iv_i$ and $\lambda_h=\sum_{j=1}^{N_2}\lambda^j\mu_j$. Then (5.35) can be written as the following system of linear equations:

$$\begin{aligned}&\sum_K\left(\sum_{i=1}^{N_1}u^i((a\nabla v_i-\mathbf{b}v_i,\nabla v_k)_K+(cv_i,v_k)_K)-\sum_{j=1}^{N_2}\lambda^j(\mu_j,v_k)_{\partial K}\right.\\ &\left.-\alpha\left(\sum_{j=1}^{N_2}\lambda^j\mu_j-\sum_{i=1}^{N_1}u^i(a\nabla v_i-\mathbf{b}v_i)\cdot n_K,a\nabla v_k\cdot n_K\right)_{\partial K}\right)=\sum_K(f,v_k)_K,\end{aligned}\tag{5.36}$$

$$\sum_K \left(\sum_{i=1}^{N_1} u^i (v_i, \mu_l)_{\partial K} + \alpha \left(\sum_{j=1}^{N_2} \lambda^j \mu_j - \sum_{i=1}^{N_1} u^i (a\nabla v_i - \mathbf{b} v_i) \cdot n_K, z\mu_l \right)_{\partial K} \right) \tag{5.37}$$
$$= \sum_K (g, \mu_l)_{\partial K \cap \partial \Omega},$$

for $k = 1, 2, \cdots, N_1$ and $l = 1, 2, \cdots, N_2$. The linear system (5.36)-(5.37) is again written as

$$\begin{bmatrix} B & C \\ D & E \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \Lambda \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix}, \qquad \text{i.e., } A\mathbf{x} = \mathbf{b}. \tag{5.38}$$

The block matrices in (5.38) are defined as follows:

$$\begin{aligned}
B &= (b_{ki})_{N_1 \times N_1}, & b_{ki} &= \sum_K \; [(a\nabla v_i - \mathbf{b} v_i, \nabla v_k)_K + (cv_i, v_k)_K \\
& & &\quad +\alpha((a\nabla v_i - \mathbf{b} v_i) \cdot n_K, za\nabla v_k \cdot n_K)_{\partial K}], \\
C &= (c_{kj})_{N_1 \times N_2}, & c_{kj} &= \sum_K \; (-(v_k, \mu_j)_{\partial K} - \alpha(z\mu_j, a\nabla v_k \cdot n_K)_{\partial K}), \\
D &= (d_{li})_{N_2 \times N_1}, & d_{li} &= \sum_K \; ((\mu_l, v_i)_{\partial K} - \alpha((a\nabla v_i - \mathbf{b} v_i) \cdot n_K, z\mu_l)_{\partial K}), \\
E &= (e_{lj})_{N_2 \times N_2}, & e_{lj} &= \sum_K \; \alpha(z\mu_l, \mu_j)_{\partial K}.
\end{aligned} \tag{5.39}$$

We can easily see that the coefficient matrix in (5.38) is non-symmetric.

Now we construct the local stiffness matrices. In the simplest case of assumption **H3** (4.33), we choose $s = r = 1$. Let K be any triangle in the partition. We denote the vertices of K by (x_i, y_i) and the edges by e_i where $i = 1, 2, 3$. We denote by s_i the sign defined on its edge e_i. The local basis function v_i is the linear function that takes on the value 1 at vertex (x_i, y_i) and 0 at the other two vertices. The local basis function μ_j is the piecewise *linear* function that takes on the value 1 at one endpoint of an edge, 0 at the other endpoint, and 0 on the other two edges. Thus there are 6 basis functions μ_j on a single element. The local stiffness matrix on K can be obtained from (5.38) and (5.39) as follows:

$$A^{\text{loc}} = \begin{bmatrix} B^{\text{loc}} & C^{\text{loc}} \\ D^{\text{loc}} & E^{\text{loc}} \end{bmatrix}, \tag{5.40}$$

where

$$\begin{aligned} B^{\text{loc}} = (b_{ki})_{3\times 3}, \quad b_{ki} = {} & (a\nabla v_i - \mathbf{b}v_i, \nabla v_k)_K + (cv_i, v_k) \\ & -\alpha\left((a\nabla v_i - \mathbf{b}v_i)\cdot n_K, za\nabla v_k \cdot n_K\right)_{\partial K}, \\ C^{\text{loc}} = (c_{kj})_{3\times 6}, \quad c_{kj} = {} & \left(-(v_k, \mu_j)_{\partial K} - \alpha(z\mu_j, a\nabla v_k \cdot n_K)_{\partial K}\right) s_j, \\ D^{\text{loc}} = (d_{li})_{6\times 3}, \quad d_{li} = {} & \left((v_i, \mu_l)_{\partial K} - \alpha(z\mu_l, (a\nabla v_i - \mathbf{b}v_i)\cdot n_K)_{\partial K}\right) s_l, \\ E^{\text{loc}} = (e_{lj})_{6\times 6}, \quad e_{lj} = {} & \alpha(z\mu_l, \mu_j)_{\partial K} s_l s_j, \end{aligned} \tag{5.41}$$

where s_l and s_k are the signs defined on the edges of consideration.

The global coefficient matrix A in (5.38) can obtained by putting all local matrices on all elements together. Again, it is not necessary to construct the global stiffness matrix as long as we know the global indices of all local basis functions.

5.4 Computation of local stiffness matrices on reference elements

The computation of the local stiffness matrices are performed on the reference finite element, not on generic elements. Let K be a triangle in the finite element partition with vertices (x_i, y_i). The reference triangle $\hat{K}$ is a triangle in the $\hat{x}\hat{y}$-plane with vertices $(0,0)$, $(1,0)$, and $(0,1)$. An affine map (see Fig. 5.4) from $\hat{K}$ to K is given by

$$\begin{pmatrix} x \\ y \end{pmatrix} = J \begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix} + \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \tag{5.42}$$

where

$$J = \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix}. \tag{5.43}$$

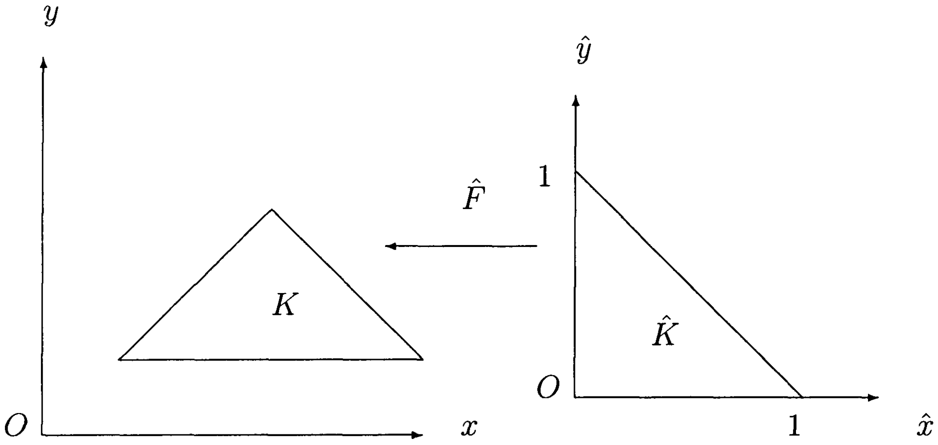

FIG. 5.1. Affine map from an arbitrary element to the reference element

Let $w = w(x, y)$ be a function on K. It is easy to see that

$$\int_K w(x,y)\ dxdy = \int_{\hat{K}} \hat{v}(\hat{x},\hat{y})\ d\hat{x}d\hat{y}, \tag{5.44}$$

where $\hat{v}(\hat{x},\hat{y}) = w(x(\hat{x},\hat{y}), y(\hat{x},\hat{y})|\text{det}J|$. The integral in (5.44) is then calculated by the following *quadrature* formula:

$$\int_{\hat{K}} \hat{v}(\hat{x},\hat{y})\ d\hat{x}d\hat{y} \approx \frac{1}{6}(\hat{v}(0.5,0) + \hat{v}(0.5,0.5) + \hat{v}(0,0.5),$$

with an error of order h^2. This formula is used to compute entries of local stiffness matrices involving integrations over element K.

For integrations over boundary pieces of K, since they are equivalent to single integrations, Simpson's rule is used.

5.5 The conjugate gradient iteration

The conjugate gradient method (CG) [28] is an iterative method for solving the system of linear equations

$$Ax = b, \tag{5.45}$$

where A is a symmetric and positive definite $N \times N$ matrix, $b \in \mathbb{R}^N$ is given, and

$$x = A^{-1}b \in \mathbb{R}^N \tag{5.46}$$

is the solution to be found. The goal of the conjugate gradient method is to find, for a give tolerance ϵ, a vector x such that $||b - Ax||_2 \leq \epsilon ||b||_2$. The procedure of the algorithm is given as follows:

CG Algorithm

```
1. For initial guess $x$, compute $r = b - Ax$ and $\rho_0 = ||r||_2^2$. Set $k = 1$.
2. While $\sqrt{\rho_{k-1}} > \epsilon ||b||_2$ do
    • if $k = 1$ then $p = r$
      else $\beta = \rho_{k-1}/\rho_{k-2}$ and $p = r + \beta p$
    • $w = Ap$
    • $\alpha = \rho_{k-1}/p^T w$
    • $x = x + \alpha p$
    • $r = r - \alpha w$
    • $\rho_k = ||r||_2^2 = r^T r$
    • $k = k + 1$
```

Note that the matrix A itself need not be formed or stored; only a routine for matrix-vector products is required. This is one big advantage of this method. Another advantage is that CG has an acceptable convergence rate. In fact, it is well known that the rate of convergence depends on the condition number of the matrix A, defined by $\kappa(A) = \dfrac{\lambda_{\text{max}}}{\lambda_{\text{min}}}$, where λ_{max} and λ_{min} are the maximum and minimum eigenvalues of A, respectively. The closer the condition number is to 1, the better the convergence rate will be.

The conjugate gradient method becomes a more efficient method when it is coupled with preconditioning. The combination is called the preconditioned conjugate gradient method (PCG). Let S be a symmetric and positive definite matrix, where $S^2 = M \approx A^{-1}$. Then the matrix SAS is symmetric and positive definite and its eigenvalues are clustered near one. Moreover, the preconditioned system

$$SASy = Sb, \tag{5.47}$$

has $y = S^{-1}x$ as a solution, where $Ax = b$. In fact, using CG algorithm on the preconditioned system (5.47), we have the following PCG algorithm:

PCG Algorithm

```
1. For initial guess $x$, compute $r = b - Ax$ and $\rho_0 = ||r||_2^2$. Set $k = 1$.
2. While $\sqrt{\rho_{k-1}} > \epsilon ||b||_2$ do
     • $z = Mr$
     • $\tau_{k-1} = z^T r$
     • if $k = 1$ then $p = z$
       else $\beta = \tau_{k-1}/\tau_{k-2}$ and $p = z + \beta p$
```

- $w = Ap$
- $\alpha = \tau_{k-1}/p^T w$
- $x = x + \alpha p$
- $r = r - \alpha w$
- $\rho_k = \|r\|_2^2 = r^T r$
- $k = k + 1$

If A is non-symmetric but nonsingular, we consider solving $Ax = b$ by applying CG to the normal equation

$$A^T A x = A^T b. \tag{5.48}$$

This is called conjugate gradient on the normal equations to minimize the residual (CGNR). Alternatively, one can solve

$$A A^T y = b, \tag{5.49}$$

and then set $x = A^T y$. This is called conjugate gradient on the normal equations to minimize the error (CGNE). The advantages of these approaches are that all the theory for CG carries over and a simple implementation for both CG and PCG can be used. There are a couple disadvantages. The first is that the condition number of the coefficient matrix $A^T A$ is the square of that of A. The second is that two matrix-vector products are needed for each step of CG iteration. Therefore the convergence is slow.

In the numerical experiments reported in this thesis, CGNR is used if the global stiffness matrix is non-symmetric.

Chapter 6

NUMERICAL EXPERIMENTS AND RESULTS

The main purposes of this chapter is to present numerical examples and results. In the first section we present experiments and results for elliptic problems. Examples with continuous and discontinuous diffusion coefficients are illustrated. In the case when true solutions are known, the rate of convergence of the numerical approximations are given. In the second section, we present some examples for the convection-diffusion problem.

6.1 Numerical examples for the elliptic problem

In this section, we give some examples for the elliptic problem (1.1)-(1.2). The stabilized symmetric finite element method (SSFEM) and the stabilized non-symmetric finite element method (SNFEM) are implemented in the two dimensional case. For simplicity, we choose $\Omega = [0,1]^2$. We also choose $r = 1$ and $s = 0$, which means that the solution u is approximated by piecewise linear functions and the normal component of the flux $\lambda = \kappa \nabla u \cdot n_K$ is approximated by piecewise constants. Uniform partitions are used, though the main code is written for quasi-uniform partitions. We consider the following examples.

Example 1. We choose $u = \sin \pi x \cos \pi y$ to be the solution to test the convergence of the schemes. The coefficient κ is chosen to be the identity matrix I. Therefore, $f = 2\pi^2 u$ in Ω and $g = u$ on $\partial\Omega$.

Let N be the number of subintervals on each side of Ω. The code of the schemes of SSFEM and SNFEM are tested for $N = 8$, 16, 32, and 64. The $|\!|\!|\cdot|\!|\!|_s$ of $(u-u_h, \lambda-\lambda_h)$, the L^2 norm of $u - u_h$, the maximum error of $u - u_h$ at the nodal points, and the maximum error of $\lambda - \lambda_h$ at midpoints of the edges are computed. The results are listed in Table 6.1 and 6.2. The rate terms show the rate of convergence, calculated by the logarithm of the ratio between consecutive errors base 2.

Table 6.1. Errors and convergence rates for SSFEM

Errors and convergence rates for SSFEM

	$\alpha = 10$		$\alpha = 1$		$\alpha = 0.1$	
N	$\vert\vert\vert(\cdot,\cdot)\vert\vert\vert_s$	Rate	$\vert\vert\vert(\cdot,\cdot)\vert\vert\vert_s$	Rate	$\vert\vert\vert(\cdot,\cdot)\vert\vert\vert_s$	Rate
8	4.849E-1		4.384E-1		4.335E-1	
16	2.246E-1	1.11	2.184E-1	1.01	2.178E-1	0.99
32	1.099E-1	1.03	1.091E-1	1.00	1.090E-1	1.00
64	5.463E-2	1.01	5.453E-2	1.00	5.452E-2	1.00
N	$\Vert u - u_h\Vert_{L^2}$	Rate	$\Vert u - u_h\Vert_{L^2}$	Rate	$\Vert u - u_h\Vert_{L^2}$	Rate
8	3.143E-1		2.941E-2		3.454E-3	
16	7.907E-2	1.99	7.369E-3	2.00	8.618E-4	2.00
32	1.980E-2	2.00	1.843E-3	2.00	2.154E-4	2.00
64	4.952E-3	2.00	4.609E-4	2.00	5.385E-5	2.00
N	$\Vert u - u_h\Vert_\infty$	Rate	$\Vert u - u_h\Vert_\infty$	Rate	$\Vert u - u_h\Vert_\infty$	Rate
8	6.551E-1		9.515E-2		3.864E-2	
16	1.662E-1	1.98	2.349E-2	2.02	9.644E-3	2.00
32	4.167E-2	2.00	5.846E-3	2.01	2.410E-3	2.00
64	1.042E-2	2.00	1.455E-3	2.01	6.024E-4	2.00
N	$\Vert\lambda - \lambda_h\Vert_\infty$	Rate	$\Vert\lambda - \lambda_h\Vert_\infty$	Rate	$\Vert\lambda - \lambda_h\Vert_\infty$	Rate
8	5.159E-2		5.159E-2		5.159E-2	
16	1.403E-2	1.88	1.403E-2	1.88	1.403E-2	1.88
32	3.648E-3	1.94	3.648E-3	1.94	3.648E-3	1.94
64	9.292E-4	1.97	9.292E-4	1.97	9.292E-4	1.97

In Table 6.1, We first notice that we have first order convergence for semi-norm $|||\cdot|||_s$. This is consistent with the error estimate we obtained in (3.73) with $r = 1$ and $s = 0$. we also see that the L^2 error of $u - u_h$ is of second order. This is consistent with our result of the L^2 error estimate for SSFEM given in (3.102). The maximum error of $u - u_h$ at nodal points and the maximum error of $\lambda - \lambda_h$ at midpoints of edges are also of second order. In fact, the second-order convergence rate of the maximum error of $\lambda - \lambda_h$ is due to superconvergence. The reason of this superconvergence is interesting and open, which requires some future work. We also want to mention that the table does show better errors for smaller parameter α, but due to the nature of our linear solver, the errors bottom out for some small α which is problem-dependent.

In Table 6.2, we see similar results to those from SSFEM. However, since CGNR is used on the normal equations, the computational cost is much greater. In fact, for $\alpha = 0.1$ and $n = 64$, the number of iteration was 6857 and CPU (Pentium III running RedHat Linux 6.2) time was 873 seconds in SNFEM comparing to 501 and 40 seconds in SSFEM. The numerical solutions are shown in Fig. 6.3.

Tests of this type have been done on $u = \sin \pi x \sin \pi y$, $u = x(1-x)y(1-y)$, $u = x^2 + y^2$, and $u = e^{x-y}$. The same convergence rates are observed.

Example 2. In this example, we test our code for which κ is not the identity matrix. In fact, κ is defined as

$$\kappa = \begin{bmatrix} 10 + x^2 & xy \\ xy & 10 + y^2 \end{bmatrix}.$$

We choose the exact solution to be $u = e^{x-y}$. Then $f = -20 - (x-y)(x-y+3))e^{x-y}$ is obtained by substituting u into the equation. The following Table 6.3 shows the errors and convergence rates of SSFEM and SNFEM for such a problem.

We can see from the result that the L^2 error of $u - u_h$ and the maximum error of

Table 6.2. Errors and convergence rates for SNFEM

	$\alpha = 10$		$\alpha = 1$		$\alpha = 0.1$	
N	$\vert\vert\vert(\cdot,\cdot)\vert\vert\vert$	Rate	$\vert\vert\vert(\cdot,\cdot)\vert\vert\vert$	Rate	$\vert\vert\vert(\cdot,\cdot)\vert\vert\vert$	Rate
8	5.000E-1		4.360E-1		4.331E-1	
16	2.267E-1	1.14	2.181E-1	1.00	2.177E-1	0.99
32	1.101E-1	1.04	1.090E-1	1.00	1.090E-1	1.00
64	5.466E-2	1.01	5.452E-2	1.00	5.452E-2	1.00
N	$\Vert u - u_h\Vert_{L^2}$	Rate	$\Vert u - u_h\Vert_{L^2}$	Rate	$\Vert u - u_h\Vert_{L^2}$	Rate
8	3.101E-1		2.615E-2		1.276E-3	
16	7.803E-2	1.99	6.585E-3	1.99	3.178E-4	2.01
32	1.954E-2	2.00	1.649E-3	2.00	7.9389E-5	2.00
64	4.887E-3	2.00	4.125E-4	2.00	1.984E-5	2.00
N	$\Vert u - u_h\Vert_\infty$	Rate	$\Vert u - u_h\Vert_\infty$	Rate	$\Vert u - u_h\Vert_\infty$	Rate
8	6.526E-1		9.177E-2		3.793E-2	
16	1.662E-1	1.97	2.325E-2	1.98	9.599E-3	1.98
32	4.1734E-2	1.99	5.833E-3	2.00	2.411E-3	1.99
64	1.045E-2	2.00	1.459E-3	2.00	6.031E-4	2.00
N	$\Vert\lambda - \lambda_h\Vert_\infty$	Rate	$\Vert\lambda - \lambda_h\Vert_\infty$	Rate	$\Vert\lambda - \lambda_h\Vert_\infty$	Rate
8	7.761E-2		5.323E-2		2.783E-2	
16	1.976E-2	1.97	1.355E-2	1.97	7.147E-3	1.96
32	4.962E-3	1.99	3.403E-3	1.99	1.807E-3	1.98
64	1.242E-3	2.00	8.519E-4	2.00	4.539E-4	1.99

$u - u_h$ at nodal points are of second order. The maximum error of $\lambda - \lambda_h$ at midpoints of edges is of first order. The numerical solutions are shown in Fig. 6.4.

Example 3. In this example, we consider discontinuous diffusion coefficient function κ. Let κ be defined as

$$\kappa = \kappa(x, y) = \begin{cases} I & x < 0.5, \\ \varepsilon I & x > 0.5, \end{cases}$$

Table 6.3. Errors on a problem with variable coefficient matrix

SSFEM ($\alpha = 0.1$)			SNFEM ($\alpha = 0.1$)	
N	$\vert\vert\vert(\cdot,\cdot)\vert\vert\vert_s$	Rate	$\vert\vert\vert(\cdot,\cdot)\vert\vert\vert$	Rate
8	6.609E-2		6.046E-2	
16	3.081E-2	1.1	3.005E-2	1.01
32	1.510E-2	1.03	1.500E-2	1.00
64	7.510E-3	1.01	7.498E-3	1.00
N	$\vert\vert u - u_h\vert\vert_{L^2}$	Ratio	$\vert\vert u - u_h\vert\vert_{L^2}$	Ratio
8	6.413E-3		6.922E-3	
16	1.605E-3	2.00	1.731E-3	2.00
32	4.014E-4	2.00	4.328E-4	2.00
64	1.003E-4	2.00	1.082E-4	7 2.00
N	$\vert\vert u - u_h\vert\vert_\infty$	Ratio	$\vert\vert u - u_h\vert\vert_\infty$	Ratio
8	2.483E-2		2.648E-2	
16	6.531E-3	1.93	6.895E-3	1.94
32	1.676E-3	1.96	1.760E-3	1.97
64	4.247E-4	1.98	4.445E-4	1.99
N	$\vert\vert\lambda - \lambda_h\vert\vert_\infty$	Ratio	$\vert\vert\lambda - \lambda_h\vert\vert_\infty$	Ratio
8	1.175E-1		2.481E-2	
16	5.301E-2	1.15	1.460E-2	0.77
32	2.429E-2	1.13	7.781E-3	0.91
64	1.145E-2	1.09	3.995E-3	0.96

where $\varepsilon > 0$ is a constant (See Figure 6.1). Such kind of regions has useful practical meanings (e.g. region filled with two media). Experiments have been done in the following cases.

1. Choose $f = 1$ and consider a solution $u = u(x)$ which depends on x only. The equation becomes $-u'' = 1$ for $x < 0.5$ and $-\varepsilon u'' = 1$ for $x > 0.5$. It is easy to

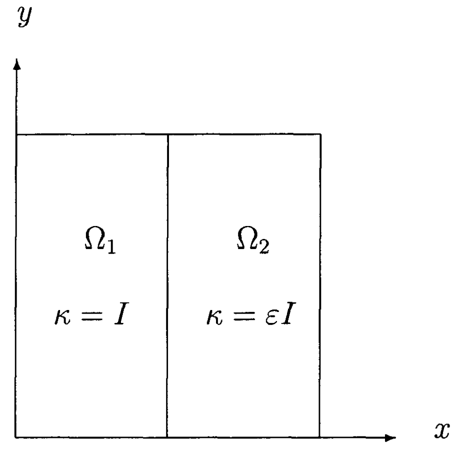

FIG. 6.1. Region Ω with discontinuous κ

see by integrating twice that

$$u(x) = \begin{cases} -0.5x^2, & x < 0.5, \\ \dfrac{1-\varepsilon-4x^2}{8\varepsilon}, & x > 0.5, \end{cases}$$

is the solution satisfying $u(0) = u'(0) = 0$. On the interface $x = 0.5$ the continuity of u and the flux $\kappa \nabla u \cdot n$ are required. We choose $\varepsilon = 0.1$. The results from SSFEM and SNFEM are listed in Table 6.4. We can see that the L^2 error of $u - u_h$ and the maximum error of $u - u_h$ at nodal points are of second order. The maximum error of $\lambda - \lambda_h$ at midpoints of edges is of first order. This is consistent with the results in the previous examples.

2. Choose $f = 1$ and $g = 0$. The true solution is not known in this case. The contour graphs of the SSFEM and SNFEM approximations are shown in Fig. 6.5-6.8. Fig. 6.5 and 6.6 show numerical solutions for symmetric and non-symmetric schemes in the case where $\varepsilon = 0.1$. The graphs are almost identical.

Table 6.4. Errors on a problem with discontinuous coefficient

SSFEM ($\alpha = 0.1$)			SNFEM ($\alpha = 0.1$)	
N	$\|\|\|(\cdot,\cdot)\|\|\|_s$	Rate	$\|\|\|(\cdot,\cdot)\|\|\|$	Rate
8	2.565E-1		2.649E-1	
16	1.282E-1	1.00	1.351E-1	0.97
32	6.411E-2	1.00	6.757E-2	1.00
64	3.205E-2	1.00	3.382E-2	1.00
N	$\|\|u - u_h\|\|_{L^2}$	Ratio	$\|\|u - u_h\|\|_{L^2}$	Ratio
8	6.773E-3		1.601E-2	
16	1.798E-3	1.91	4.088E-3	1.97
32	4.582E-4	1.97	1.031E-3	1.99
64	1.152E-4	1.99	2.589E-4	7 2.00
N	$\|\|u - u_h\|\|_\infty$	Ratio	$\|\|u - u_h\|\|_\infty$	Ratio
8	6.494E-2		8.191E-2	
16	1.635E-2	1.99	2.105E-2	1.96
32	4.090E-3	2.00	5.258E-3	2.00
64	1.022E-3	2.00	1.315E-3	2.00
N	$\|\|\lambda - \lambda_h\|\|_\infty$	Ratio	$\|\|\lambda - \lambda_h\|\|_\infty$	Ratio
8	2.142E-2		3.258E-2	
16	1.048E-2	1.03	1.659E-2	0.97
32	5.217E-3	1.01	8.288E-3	1.00
64	2.605E-3	1.00	4.145E-3	1.00

We can see that the solution behaves differently on different sides of the interface $x = 0.5$. Fig. 6.7 and 6.8 plot the numerical solutions for $\varepsilon = 0.0001$. In each of these two cases the solution in region Ω_2 dominates the solution in Ω_1.

Example 4. In this example, we also consider a discontinuous coefficient function

κ, defined by

$$\kappa = \kappa(x,y) = \begin{cases} \varepsilon_1 I & x < 0.5 \text{ and } y < 0.5 \\ \varepsilon_2 I & x > 0.5 \text{ and } y > 0.5 \\ I & \text{elsewhere in } \Omega, \end{cases}$$

where ε_1 and ε_2 are constants (See Fig. 6.2). Experiments have been done for different choices of ε_1 and ε_2. Some results are shown in Fig. 6.9-6.12. Fig. 6.9 and 6.10 plot the numerical solutions of the symmetric and the non-symmetric schemes for $\varepsilon_1 = 0.1$ and $\varepsilon_2 = 0.2$. Fig. 6.11 and 6.12 plot the numerical solutions for $\varepsilon_1 = 0.0001$ and $\varepsilon_2 = 0.0002$. Each pair of the contour graphs are almost identical. The dominance of the solution is also shown in the region of small coefficient κ.

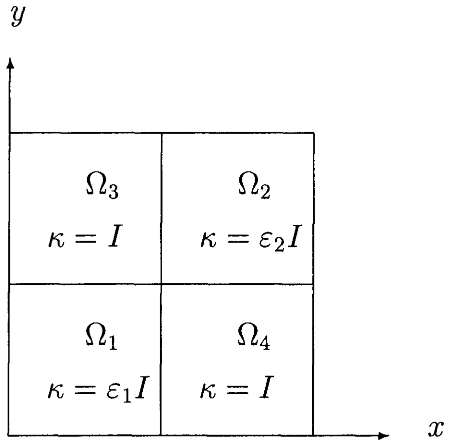

FIG. 6.2. Another region Ω with discontinuous κ

6.2 Numerical examples for the convection-diffusion problem

In this section, we show a few numerical experiments and results for the convection-diffusion problem. Again we choose $\Omega = [0,1]^2$. We also choose $r = 1$ and $s = 1$, which means that the solution u and the normal component of the flux $\lambda = (a\nabla u - \mathbf{b}u) \cdot n_K$ are approximated by piecewise linear functions.

In the first test problem, we choose diffusion coefficient $a = \varepsilon I$, convection term $\mathbf{b} = (1, 0)$, absorption term $c = 1$, source $f = 1$, and boundary value $g = 0$. We consider two different cases. In the first case, let $\varepsilon = 0.1$. The numerical solution is shown in Figure 6.13-6.14. Without the convection term, the solution should be symmetric around the center of the region. The impact of the convection pushed the solution in the direction of $\mathbf{b} = (1, 0)$. In the second case, let $\varepsilon = 0.0001$. The numerical solution is shown in Figure 6.15-6.16. The convection term plays a more significant role than in the previous case. As seen from the 3D plot of our solution, a boundary layer is observed near the outflow boundary of $x = 1$. This is consistent with the known results for such a problem.

The second test problem is chosen from [21]. We choose diffusion coefficient $a = \varepsilon I$, with $\varepsilon = 0.01$, convection term $\mathbf{b} = (0, 1)$ and absorption term $c = 0$, and source $f = 0$. The boundary value is given by $g = \sin \pi x$ if $y = 0$ and $g = 0$ elsewhere. The problem has an exact solution given by

$$u(x, y) = \frac{\sin \pi x}{e^{\lambda_2 - \lambda_1}}(e^{\lambda_2 - \lambda_1} e^{\lambda_1 y} - e^{\lambda_2 y}),$$

where

$$\lambda_1 = \frac{1 - \sqrt{1 + 4\pi^2 \varepsilon^2}}{2\varepsilon} \text{ and } \lambda_2 = \frac{1 + \sqrt{1 + 4\pi^2 \varepsilon^2}}{2\varepsilon}.$$

From the numerical solution in Fig.6.17-6.18, we can see the sharp boundary layer at the outflow boundary. As a comparison, we display the exact solution in Fig.6.19-6.20.

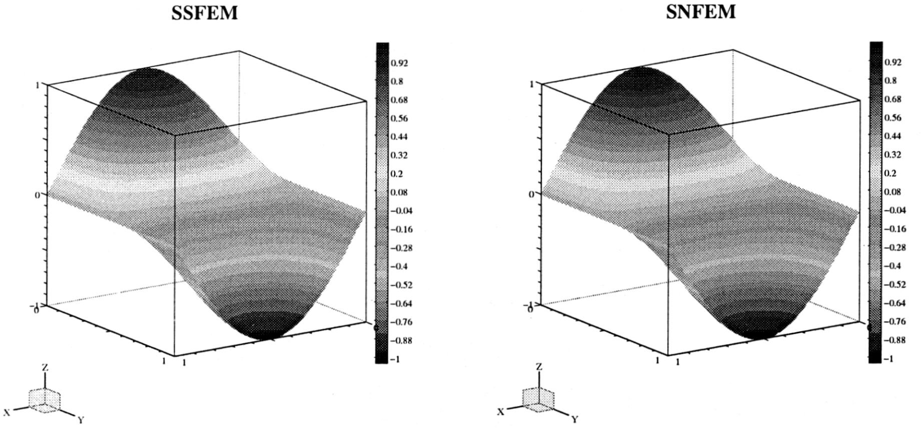

FIG. 6.3. Numerical solutions of a test problem with $\kappa = I$

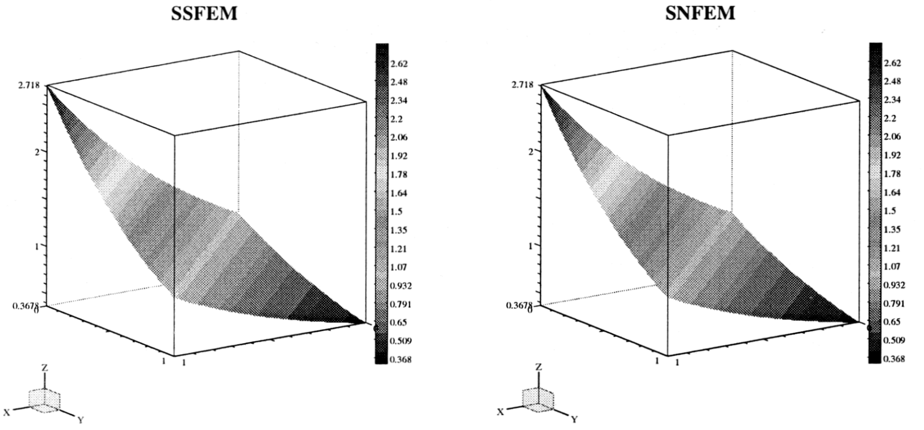

FIG. 6.4. Numerical solutions of a test problem with non-constant κ

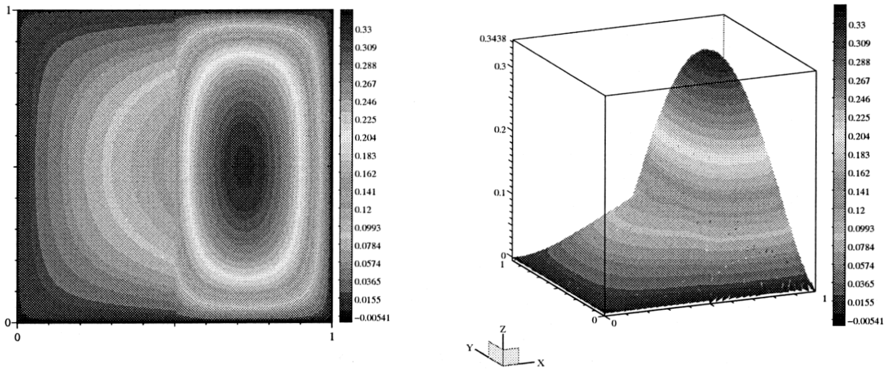

FIG. 6.5. SSFEM approximations with discontinuous κ ($\varepsilon = 0.1$)

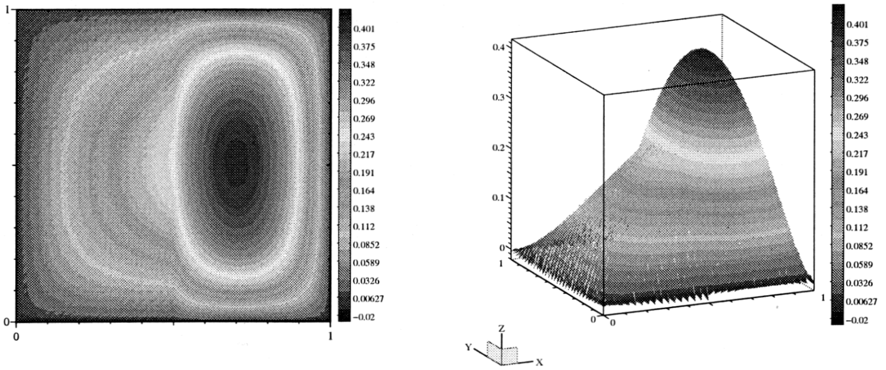

FIG. 6.6. SNFEM approximations with discontinuous κ ($\varepsilon = 0.1$)

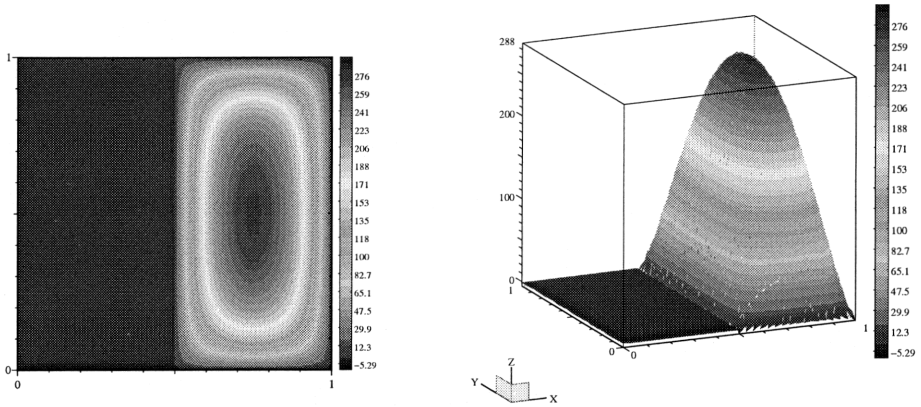

FIG. 6.7. SSFEM approximations with discontinuous κ ($\varepsilon = 0.0001$)

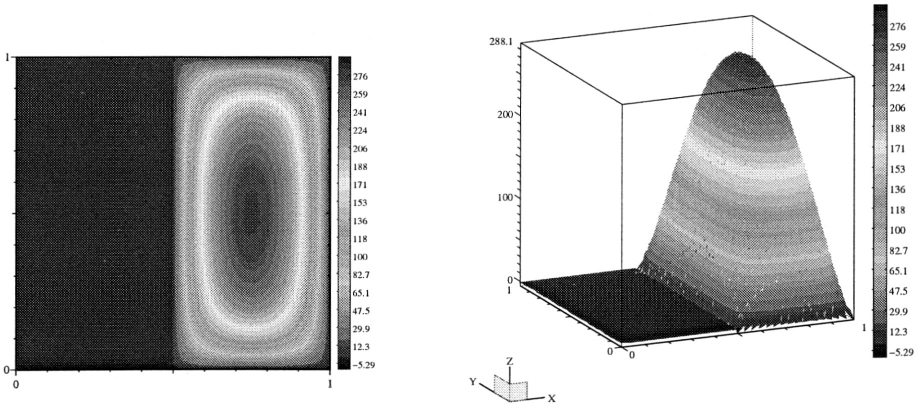

FIG. 6.8. SSFEM approximations with discontinuous κ ($\varepsilon = 0.0001$)

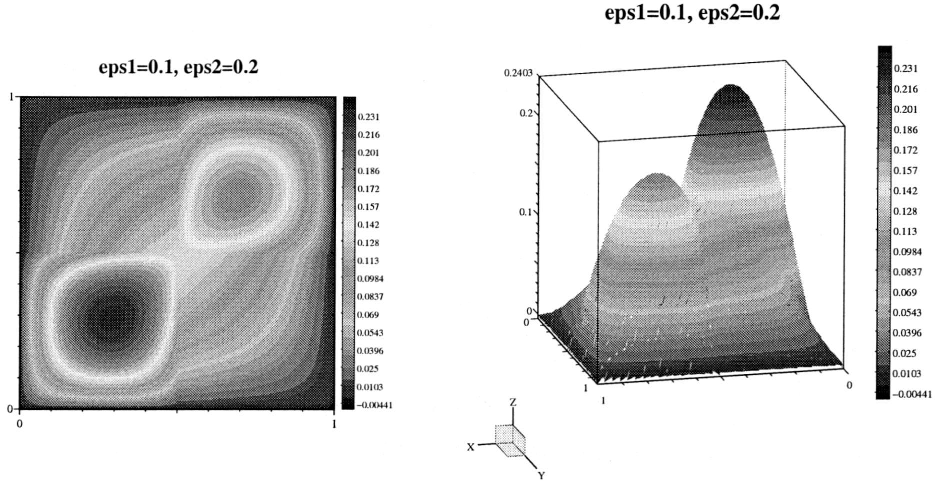

FIG. 6.9. SSFEM approximation with discontinuous κ ($\varepsilon_1 = 0.1$, $\varepsilon_2 = 0.2$)

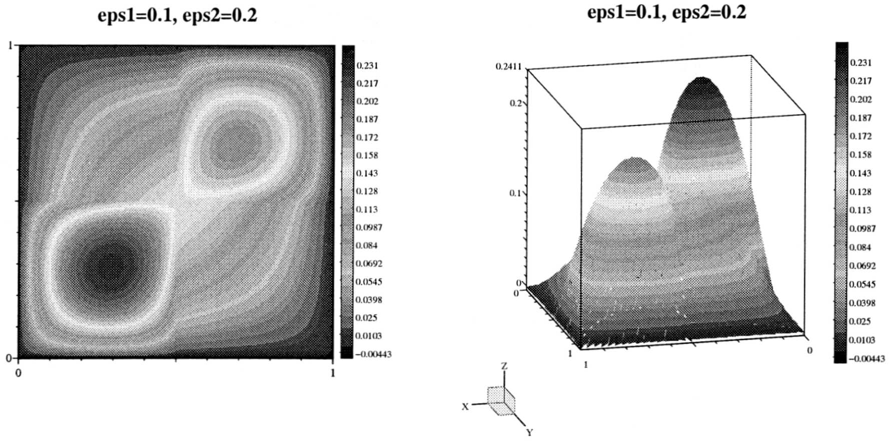

FIG. 6.10. SNFEM approximation with discontinuous κ ($\varepsilon_1 = 0.1$, $\varepsilon_2 = 0.2$)

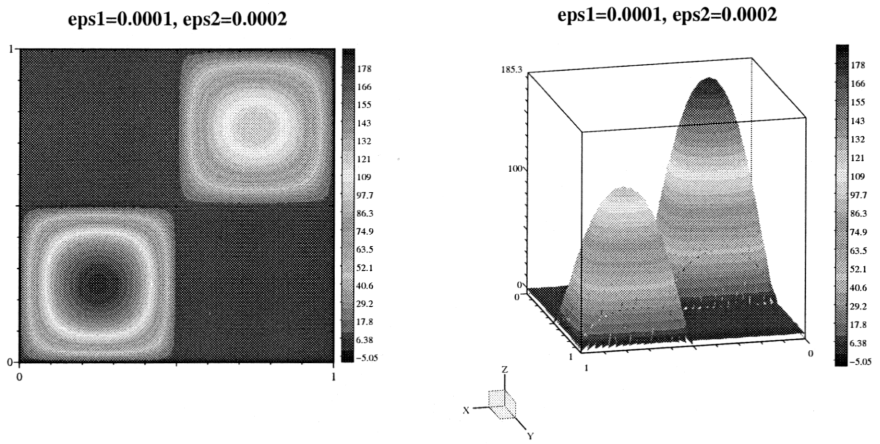

FIG. 6.11. SSFEM approximation with discontinuous κ ($\varepsilon_1 = 0.0001$, $\varepsilon_2 = 0.0002$)

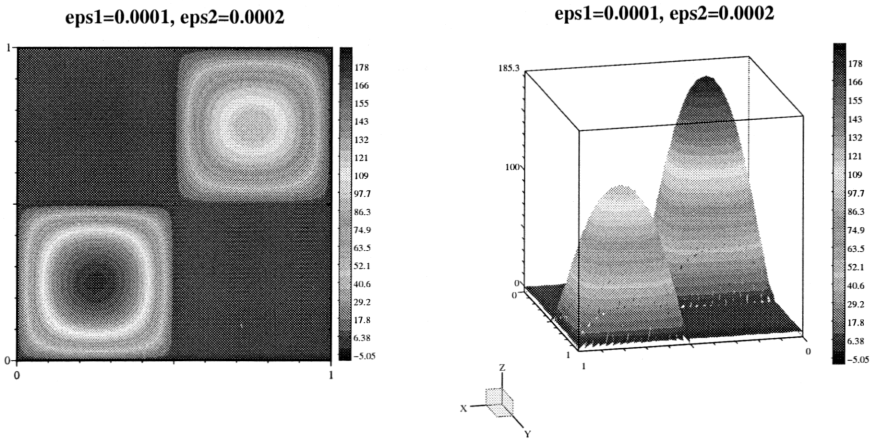

FIG. 6.12. SNFEM approximation with discontinuous κ ($\varepsilon_1 = 0.0001$, $\varepsilon_2 = 0.0002$)

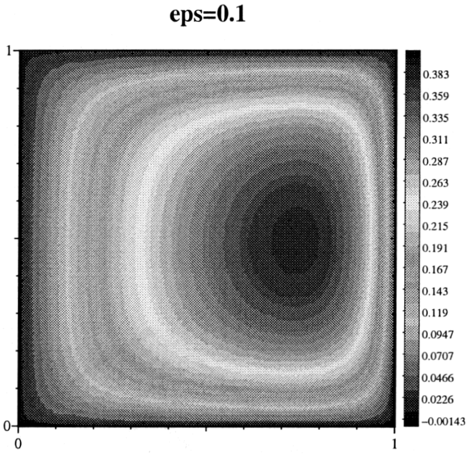

FIG. 6.13. Solution of a convection-diffusion problem

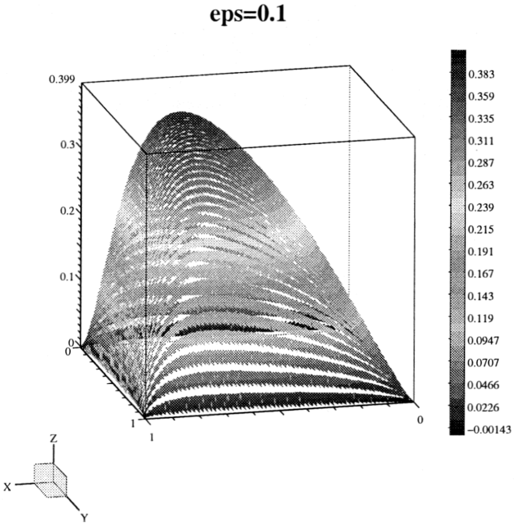

FIG. 6.14. Solution of a convection-diffusion problem

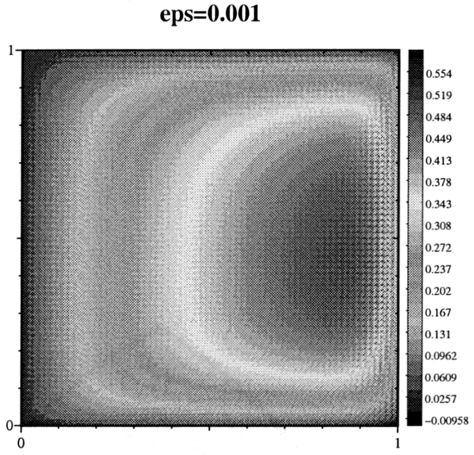

FIG. 6.15. Solution of a convection-dominated convection-diffusion problem

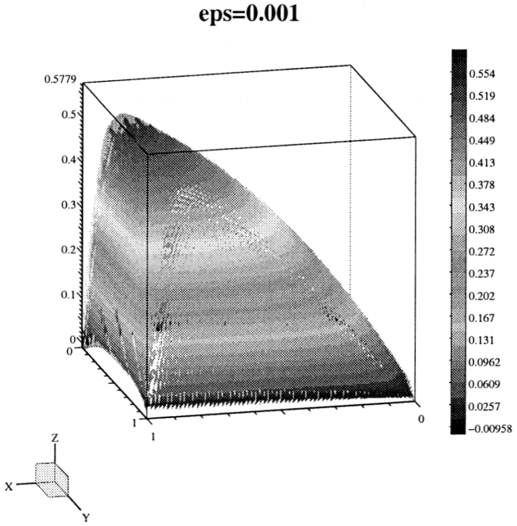

FIG. 6.16. Solution of a convection-dominated convection-diffusion problem

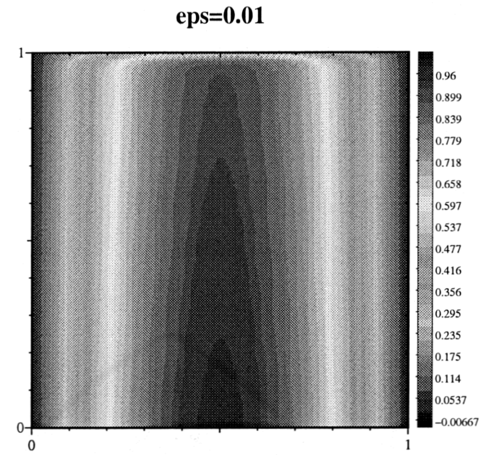

FIG. 6.17. Sharp boundary layer at outflow boundary $y = 1$

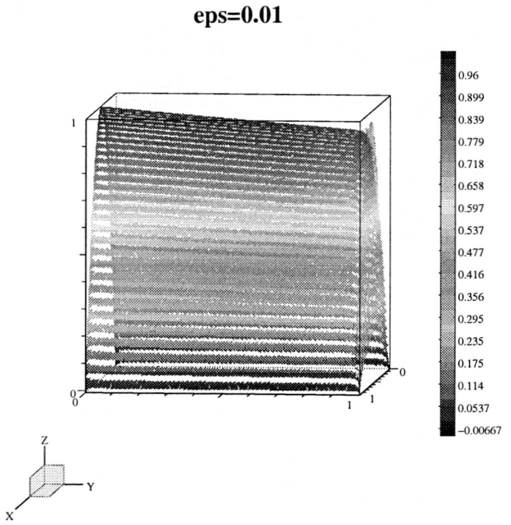

FIG. 6.18. Sharp boundary layer at outflow boundary $y = 1$

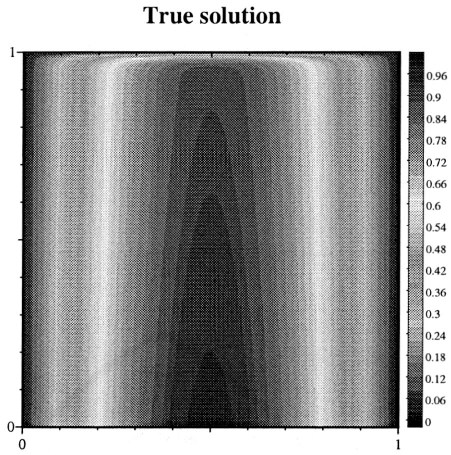

FIG. 6.19. Exact solution of a test problem

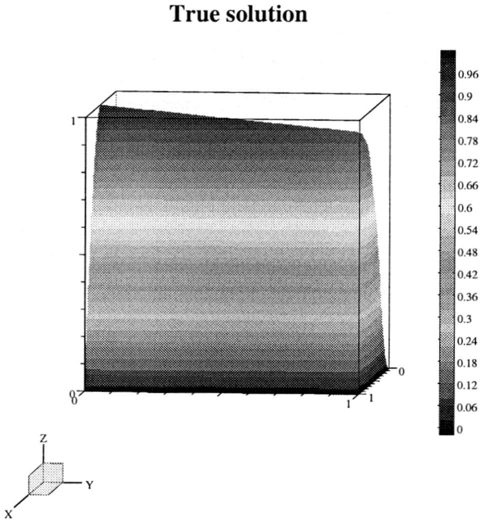

FIG. 6.20. Exact solution of a test problem

Chapter 7

SUMMARY AND FUTURE WORK

7.1 Summary

One of the main contributions of this thesis is that we proposed and analyzed a new and innovative finite element method for second order elliptic equations using discontinuous piecewise polynomials on a finite element partition consisting of general polygons. The new method is based on a stabilization of the well-known primal hybrid formulation by using some least-squares forms imposed on the boundary of each element. Two variational formulations of the new method are provided. The equivalence of the weak solution and the classical solution is shown. Corresponding symmetric and a non-symmetric finite element scheme are presented. The non-symmetric formulation is absolutely stable in the sense that no parameter selection is necessary for the scheme to converge. The symmetric formulation is conditionally stable in that a parameter has to be selected in order to have an optimal order of convergence. Optimal-order error estimates in some H^1-equivalence norms are established for the proposed discontinuous finite element methods. For the symmetric formulation, an optimal-order error estimate is also derived in the L^2 norm.

Another contribution of the thesis is the application of the stabilized discontinuous finite element method to a convection-dominated convection-diffusion problem. The standard Galerkin finite element methods applied to such problems exhibit a variety of deficiencies, including high oscillations and poor approximation of the derivatives of the solutions. A new stabilization technique, which features a non-symmetric formulation using discontinuous piecewise polynomials, is presented and analyzed for such problems. Existence and uniqueness of the finite element approximation is established. Error estimates in some H^1-equivalence norm are established for the proposed discontinuous finite element method.

Discussion on the construction of stiffness matrices of the finite element schemes is presented. For the symmetric formulation, a system of linear equations with symmetric and positive definite coefficient matrix is derived. Implementation of the finite element schemes is carried out. An efficient numerical solver is developed using C++. The solver is tested on a lot of examples. The resulting numerical solutions have showed some properties of the true solutions and desired accuracy.

7.2 Future work

The stabilization method presented and analyzed in this thesis is based on the coupled system of the solution u and the normal component of the flux (e.g. $\lambda = \kappa \nabla u \cdot n_K$ for elliptic problems). In the future, we plan to carry out the analysis to the system coupling u and the flux variable (e.g. $\mathbf{p} = \kappa \nabla u$ for elliptic problems) on all element boundaries. A similar analysis can also be applied to coupled system of u and the flux variable in all elements. The advantage of these systems is that they can provide more information on the derivatives of the solution u. The objective is to construct finite element schemes that are stable and accurate for these systems, and to give a systematic study of the error.

Another planned future research direction is in the computational aspects of our method. The linear solver that is used in the thesis is basically the conjugate gradient method. The convergence is relatively slow especially for non-symmetric systems where CG for normal equations is used. In the future, we plan to investigate fast linear solvers, such as domain decomposition and multigrid methods, to speed up the convergence and improve the efficiency of our method. We also plan to extend our numerical experiments to the three dimensional case.

The idea of stabilization can be applied in many areas. For example, in the first-order system least-squares finite element methods, second-order convection-diffusion equations can be written as first-order systems. Again applying the standard Galerkin finite element methods to elliptic equations with significant convection terms will result in non-physical oscillations. Employing the least-squares principle can overcome these deficiencies. In the future, I also want to focus on the numerical modeling of fluid flow problems. I intend to apply the idea of stabilized discontinuous finite element to the Navier-Stokes equation. In fact, this approach has been applied to the Stokes equation and some very promising results have been derived. Overall, I think the stabilization idea can be applied to many difficult problems in the area of applied partial differential equations.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] I. Babuška, *The finite element method with Lagrangian multiplier*, Numer. Math., 20(1973), pp. 179-192.

[3] R. E. Bank and M. Benbourenane, *The hierarchical basis multigrid method for convection-diffusion equations*, Numer. Math., 61(1992), pp. 7-37.

[4] C. E. Baumann and J. T. Oden, *A discontinuous hp finite element method for convection-diffusion problems.* Comput. Method Appl. Mech. Engrg., 175(1999), pp. 311-341.

[5] C. E. Baumann and J. T. Oden, *A discontinuous hp finite element method for the Euler and Navier-Stokes equations.* Internat. J. Numer. Methods Fluids., 31(1999), pp. 79-95.

[6] D. Braess, *Finite Elements: Theory, fast solvers, and applications in solid mechanics*, Cambridge University Press, New York, 2001.

[7] J. H. Bramble, J. Pasciak, J. Wang, and J. Xu, *Convergence estimates for product iterative methods with applications to domain decomposition*, Math. Comp., 57(1991), pp. 1-21.

[8] J. H. Bramble and S. R. Hilbert, *Estimation of linear functionals on Sobolev spaces with application to Fourier transforms and spline interpolation*, SIAM J. Numer. Anal., 7(1970), pp. 113-124.

[9] J. H. Bramble and S. R. Hilbert, *Bounds for a class of linear functionals with applications to Hermite interpolation*, Numer. Math., 16(1971), pp. 362-369.

[10] J. H. Bramble and M. Zlámal, *Triangular elements in the finite element method*, Math. Comp., 24(1970), pp. 809-820.

[11] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.

[12] F. Brezzi, *On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers*, RAIRO, Anal. Numér., 2(1974), pp. 129-151.

[13] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[14] J. Céa, *Approximation variationelle; Convergence dés aux limites*, Ann. Inst. Fourier (Grenoble) 14(1964), pp. 345-444.

[15] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, vol 4, 1978, North-Holland.

[16] B. Cockburn, *Discontinuous Galerkin methods for convection-dominated problems, High-Order Methods for Computational Physics (T. Barth and H. Deconink, eds.)*, Lecture Notes in Computational Science and Engineering, vol. 9, Springer Verlag, New York, 1999, pp. 69-224.

[17] B. Cockburn and C. W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for scalar conservation laws II: General framework*, Math. Comp., 52(1989), pp. 411-435.

[18] J. Douglas, Jr. and J. Wang, *An absolutely stabilized finite element method for the Stokes problem*, Math. Comp., 52(1989), pp. 495-508.

[19] R. Ewing, J. Wang, and Y. Yang, *A stabilized discontinuous finite element method for elliptic problems*, Numer. Lin. Algebra Appl., 10(2003), pp. 83-104.

[20] L. P. Franca, S. L. Frey, and T. J. R. Hughes, *Stabilized finite element methods: I. Application to the advective-diffusive model*, Comput. Methods Appl. Mech. Engrg., 95(1992), pp. 253-276.

[21] L.P. Franca, A. Nesliturk, and M. Stynes, *On the stability of residual-free bubbles for convection-diffusion problems and their approximation by a two-level finite element method*, Comput. Methods Appl. Mech. Engrg., 166(1998), pp. 35-49.

[22] M. Griebel and F. Kiefer, *Generalized hierarchical basis multigrid methods for convection-diffusion problems*, SFB Preprint 720, Sonderforschungsbereich 256, Institut für Angewandte Mathematik, Universität Bonn, 2001.

[23] M. Hanke, *Conjugate Gradient Type Methods for Ill-Posed Problems*, Longman Scientific & Technical, New York, 1995.

[24] T. J. R. Hughes and A. Brooks, *A theoretical framework for Petrov-Galerkin methods with discontinuous weighting functions: Application to the streamline-upwind procedure (R. H. Gallagher, D. H. Norrie, J. T. Oden, and O. C. Zienkiewicz eds.)*, Finite elements in fluids, vol. 4 , John Wiley & Sons Ltd., 1982, pp. 47-65.

[25] T. J. R. Hughes and L. P. Franca, *A new finite element formulation for computational fluid dynamics: VII. The Stokes problem with various well-posed boundary conditions: Symmetric formulations that converge for all velocity/pressure spaces*, Comput. Methods Appl. Mech. Engrg., 65(1987), pp. 85-96.

[26] C. Johnson and J. Pitkäranta, *An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation*, Math. Comp., 46(1986), pp. 1-26.

[27] C. Johnson, A. H. Schatz, and L. B. Wahlbin, *Crosswind smear and pointwise errors in streamline diffusion finite element methods*, Math. Comp., 49(1987) pp. 25-38.

[28] C. T. Kelley, *Iterative Methods for Linear and Nonlinear Equations*, SIAM, Philadelphia, 1995.

[29] J. A. Mackenzie and K. W. Morton, *Finite volume solutions of convection-diffusion test problems*, Math. Comp,, 60(1996), pp. 189-220.

[30] K. W. Morton, *Numerical Solution of Convection-Diffusion Problems*, Chapman & Hall, London, 1996.

[31] J. T. Oden, I. Babuska, and C E. Baumann, *A discontinuous hp finite element method for diffusion problems.* J. Comput. Phys., 146(1998), pp. 491-519.

[32] C. Pflaum, *Robust convergence of multilevel algorithms for convection-diffusion equations*, SIAM J. Numer. Anal., 37(2000), pp. 443-469.

[33] P. A. Raviart, *Hybrid finite element methods for solving 2nd order elliptic equations (J.J.H Miller, Editor)*, Topics in Numerical Analysis, II, Academic Press, New York, 1975, pp. 141-155.

[34] P. A. Raviart and J. M. Thomas, *Primal hybrid finite element methods for 2nd order elliptic problems*, Math. Comp., 31(1977), pp. 391-413.

[35] H. G. Roos, M. Stynes, and L. Tobiska, *Numerical Methods for Singularly Perturbed Equations*, Springer-Verlag, New York, 1996.

[36] A. H. Schatz, V. Thomée, and W. L. Wendland, *Mathematical Theory of Finite and Boundary Element Methods*, Birkhäuser Verlag, Boston, 1990.

[37] M. Stynes, *Numerical solution of convection-diffusion problems*, Bull. Irish Math. Soc. 30 (1993), pp. 41-55.

[38] P. S. Vassilevski and J. Wang, *Stabilizing the hierarchical basis by approximate wavelets. I: Theory*, Numer. Lin. Algebra Appl., 4(1997), pp. 103-126.

[39] P. S. Vassilevski and J. Wang, *Stabilizing the hierarchical basis by approximate wavelets, II: Implementation and numerical results*, SIAM J. Sci. Comput., 20(1998), pp. 490-514.

[40] B. Riviere, M.F. Wheeler, and V. Girault, *A priori error estimates for finite element methods based on discontinuous approximation spaces for elliptic problems*, SIAM J. Numer. Anal., 39(2001), pp. 902-931.

[41] O. C. Zienkiewicz and R. L. Taylor, *The Finite Element Method*, vol 1, McGraw-Hill, 1989.