ADAPTIVE & MULTILEVEL STOCHASTIC GALERKIN FINITE ELEMENT METHODS

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Science and Engineering

2020

Adam James Crowder
School of Mathematics

Contents

List of Tables

List of Figures

The University of Manchester

Adam James Crowder
Doctor of Philosophy
Adaptive & Multilevel Stochastic Galerkin Finite Element Methods
January 14, 2020

Stochastic Galerkin finite element methods (SGFEMs) are a popular choice for the numerical solution of PDE problems with uncertain or random inputs that depend on countably many random variables. Standard SGFEMs compute approximations in a fixed number of random variables which are selected a priori and the rates of convergence deteriorate as the number of variables increases. The size of the associated linear systems of equations also grows rapidly with the number of input random variables and desktop computer memory is quickly exhausted. In general, it is unknown a priori which, or how many, random variables need be incorporated into the discretisation in order to estimate quantities of interest to a prescribed error tolerance. Quantities of interest which depend strongly on a high number of input random variables pose serious theoretical and computational challenges and new sophisticated algorithms are required to approximate them.

We focus on the design of efficient adaptive SGFEM algorithms for elliptic PDEs with inputs with affine dependence on a countably infinite number of random variables. Starting with an initial cheap–to–compute approximation, we employ an implicit a posteriori error estimation strategy to steer the adaptive refinement of the approximation space, ensuring that only the most important random variables with respect to the energy error are incorporated. To ensure that the correct decisions are made during the adaptive selection process, the error estimate needs to be highly accurate. Additionally, a suitable balance between the computational cost of the estimate and the desired accuracy must be struck to ensure that the algorithm is efficient and quick to run.

This thesis contains two novel contributions. First, we investigate the so–called CBS constant associated with two finite element spaces that appears in the bound relating the true error to the estimated error. With the aim of designing cheap error estimates with effectivity indices close to one, we compute the CBS constant associated with several non–standard pairs of finite element spaces. For certain pairs, we also prove new theoretical estimates for the associated CBS constants using only linear algebra arguments. Second, we design a novel adaptive multilevel SGFEM algorithm, where each solution mode is associated with a potentially different finite element space. When applied to the stochastic diffusion problem, we demonstrate that our multilevel algorithm performs optimally in that it realises the rate of convergence afforded to the underlying finite element method for the analogous deterministic problem (despite the diffusion coefficient being modelled as a function of an infinite number of random variables). We consider convex and non–convex spatial domains, the latter of which leads to solutions with spatial singularities. To realise the optimal rates of convergence on non–convex domains, we also employ a local mesh refinement strategy within our multilevel algorithm for each solution mode.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

Acknowledgements

I thank my project supervisors, Professors Catherine Powell and David Silvester, from whom I've learnt so much over the past four years. I thank them deeply for taking me on as a PhD student and for the time and effort they've invested in guiding me to the completion of my project. I am also deeply grateful for the support of my close friends in the Alan Turing building, especially my officemates James and Steven who were always available in times of need, work or otherwise.

I thank my loving family, most of all my parents Gill and Trevor who have made a number of sacrifices to support me throughout the past four years. I thank them for their patience, understanding and for not asking too often what I was stuck on. Dad, you can at last go back to making noise in the workshop.

Finally, I thank my loving wife Sally. Her endless encouragement, wisdom and unique ability to put a smile on my face I could not do without. With it, this project brought many challenges for us both and its end marks the beginning of an exciting future together.

in memory of *Alan & Muriel*

Chapter 1

Introduction

1.1 Uncertainty Quantification

Mathematicians and engineers often model physical processes using partial differential equations (PDEs). In many circumstances, we employ such models as a replacement for real world experiments regarded to be either too expensive, unethical or infeasible. Many physical processes of interest appear naturally; biological and chemical processes in organic matter for example. Others arise by design in industrial and manufacturing processes.

Most PDE models require some form of input data. For example, coefficients, boundary conditions and the geometry of the domain on which the PDE is posed. In traditional applied mathematics, such inputs are usually assumed to be deterministic, or completely known. However, when modelling real–world processes, this is often not the case. Sometimes, data and physical measurements may be collected before constructing meaningful mathematical representations of model inputs. Clearly, only a finite number of measurements or recordings can be taken. Similarly, only a limited amount of data can be stored. Models therefore exhibit a level of uncertainty caused by a lack of knowledge about their inputs, which we refer to as *epistemic* uncertainty. Often, epistemic uncertainty is overlooked and disguised as a modelling assumption. Whilst many models are still extremely effective, despite the underlying uncertainty, most state–of–the–art methods now seek to quantify that uncertainty and its impact on the model solution, thus giving rise to the field of Uncertainty Quantification (UQ).

Broadly speaking, the field of UQ can be split into two distinct research areas;

forward UQ and *inverse* UQ. Forward UQ is straightforward to comprehend, and its essence is captured above. We are simply tasked to consider the following question; given a model with uncertain inputs, what is the uncertainty in the output? Advancing this notion we seek to compute probabilistic information about the uncertain output, its expectation or variance for example, given a probability distribution for the inputs. In certain situations we may even wish to compute the probability that a rare or catastrophic event may occur. Conversely, for inverse UQ we are tasked to consider the question; given a model whose outputs have been observed (subject to noise), what *were* the inputs? Recasting the inverse problem in the Bayesian framework enables us to compute posterior probability distributions for the unknown inputs, conditioned on available data. In–depth reviews of UQ can be found in [97, 101, 67] and more information about the Bayesian framework is provided in [99]. The focus of this thesis is forward UQ and the efficient solution of PDE problems with random or uncertain inputs.

1.2 Stochastic Finite Element Methods

Our starting point is the deterministic diffusion problem: find $u(\mathbf{x}) : D \to \mathbb{R}$ such that

$$-\nabla \cdot (a(\mathbf{x})\nabla u(\mathbf{x})) = f(\mathbf{x}), \qquad \mathbf{x} \in D, \tag{1.1}$$

$$u(\mathbf{x}) = g(\mathbf{x}), \qquad \mathbf{x} \in \partial D, \tag{1.2}$$

where $D \subset \mathbb{R}^2$ is the spatial domain and $g(\mathbf{x}) : D \to \mathbb{R}$ represents some prescribed Dirichlet data associated with the boundary ∂D of D. Finite element methods (FEMs) are commonly employed for the numerical solution of PDE problems such as (1.1)–(1.2) with deterministic input data; see [91, 27, 28, 49] for example. When the model inputs are uncertain or when we lack information about them, they may be modelled or represented by functions of suitably chosen random variables. The resulting equations are often referred to as stochastic PDEs or PDEs with random inputs. The chief idea of forward UQ is to propagate the uncertainty in the inputs to the outputs (the solution or some quantity of interest that depends on the solution). In this thesis, we consider a powerful class of methods for forward UQ called *stochastic finite element methods* (SFEMs) (see [60] for a review of several different methods).

We study the case where the diffusion coefficient $a(\mathbf{x})$ in (1.1) is uncertain. Specifically, we replace $a(\mathbf{x})$ with a random field $a(\mathbf{x},\omega) : D \times \Omega \to \mathbb{R}$ of the form

$$a(\mathbf{x},\omega) = a_0(\mathbf{x}) + \sum_{m=1}^{\infty} a_m(\mathbf{x})\xi_m(\omega), \tag{1.3}$$

for a given sequence of bounded random variables $\xi_m(\omega) : \Omega \to \mathbb{R}$ and spatially varying functions $a_m(\mathbf{x}) : D \to \mathbb{R}$ for $m = 0, 1, \ldots$. We refer to Ω as the stochastic domain and study the stochastic diffusion problem: find $u(\mathbf{x},\omega) : D \times \Omega \to \mathbb{R}$ such that $\mathbb{P}$–a.s. (i.e., with probability one)

$$-\nabla \cdot (\bar{a}(\mathbf{x},\boldsymbol{\xi})\nabla u(\mathbf{x},\omega)) = f(\mathbf{x}), \qquad \mathbf{x} \in D, \tag{1.4}$$

$$u(\mathbf{x},\omega) = g(\mathbf{x}), \qquad \mathbf{x} \in \partial D, \tag{1.5}$$

where $\bar{a}(\mathbf{x},\boldsymbol{\xi})$ represents the function $a(\mathbf{x},\omega)$ with the expansion truncated after M terms with

$$\boldsymbol{\xi} := \big(\xi_1(\omega), \xi_2(\omega), \ldots, \xi_M(\omega)\big)^\top : \Omega \to \mathbb{R}^M.$$

Note that the full coefficient $a(\mathbf{x},\omega)$ in (1.3) is a primary focus of this thesis. We consider $\bar{a}(\mathbf{x},\boldsymbol{\xi})$ here to describe standard SFEMs, the simplest of which are sampling methods called Monte Carlo FEMs (MCFEMs).

The main idea of MCFEMs is to take random samples of $\boldsymbol{\xi}$ and compute approximations $u_i(\mathbf{x})$ to (1.4)–(1.5) for each sample $\bar{a}(\cdot,\boldsymbol{\xi}_i)$ using standard deterministic FEMs. Then, the MCFEM approximation for the expectation of $u(\mathbf{x},\omega)$, for example, is given by the sample average

$$\mathbb{E}_{\mathrm{MC}}[u] := \frac{1}{N}\sum_{i=1}^{N} u_i(\mathbf{x}),$$

where N is the total number of samples taken. MCFEMs are popular with practitioners because they are *non-intrusive* in that they can be implemented using existing deterministic FEM software with little–to–no modification. However, it is well–known that the approximation error for $\mathbb{E}_{\mathrm{MC}}$ decays slowly like $N^{-1/2}$, and many samples and deterministic solves are required to generate accurate approximations.

Another class of non-intrusive sampling methods are stochastic collocation FEMs (SCFEMs) [112, 6]. For N samples of $\boldsymbol{\xi}$, we construct global approximations for $u(\mathbf{x},\omega)$

of the form

$$u_N(\mathbf{x}, \boldsymbol{\xi}) = \sum_{i=1}^{N} u_i(\mathbf{x}) L_i(\boldsymbol{\xi}),$$

where $L_i(\boldsymbol{\xi})$ represents a multivariate polynomial associated with the sample $\boldsymbol{\xi}_i$. Whilst the FEM approximations $u_i(\mathbf{x})$ associated with $\boldsymbol{\xi}_i$ are simple to compute, quantities such as $\mathbb{E}[u_N]$ require numerical integration in M dimensions, which is expensive for large values of M.

In this thesis, we consider a class of intrusive methods called stochastic Galerkin FEMs (SGFEMs). They are non–sampling methods which lead to a large linear system of equations associated with a single forward solve involving $\bar{a}(\mathbf{x}, \boldsymbol{\xi})$. Akin to FEMs for deterministic problems, the main idea is to construct a polynomial basis

$$\big\{\psi_i(\boldsymbol{\xi});\ i = 1, 2, \ldots, s\big\},$$

where ψ_i denotes a multivariate polynomial in the variables ξ_m for $m = 1, 2, \ldots, M$, and define the SGFEM approximation

$$u_X(\mathbf{x}, \boldsymbol{\xi}) = \sum_{i=1}^{s} u_i(\mathbf{x}) \psi_i(\boldsymbol{\xi}), \qquad u_i(\mathbf{x}) : D \to \mathbb{R}, \tag{1.6}$$

for $u(\mathbf{x}, \omega)$. Then, we use Galerkin approximation to determine the coefficients $u_i(\mathbf{x})$ that define u_X. When the basis functions ψ_i are constructed in a particular way, the expectation $\mathbb{E}[u_X]$ and variance $\mathrm{Var}(u_X)$ admit exact analytical expressions which are straightforwardly evaluated without numerical integration, regardless of the size of M. This is a key advantage of many SGFEMs over MCFEMs. Additionally, sophisticated SGFEMs do not require the a priori truncation of $a(\mathbf{x}, \omega)$ in (1.3).

1.3 Stochastic Galerkin FEMs

Standard SGFEMs seek approximations u_X in tensor–product spaces

$$X := H_1 \otimes P, \tag{1.7}$$

where H_1 represents a FEM space of piecewise polynomials on D and P is a set of global polynomials in the variables ξ_m for $m = 1, 2, \ldots, M$. Under this construction, the coefficients u_i in (1.6) reside in H_1. It was shown in [7] that the rate of convergence

of standard SGFEMs deteriorates as we increase the truncation number M associated with $\bar{a}(\mathbf{x}, \boldsymbol{\xi})$. This phenomenon is known as the *curse of dimensionality* and appears in many areas of numerical analysis. Consequently, standard SGFEMs are inadequate for practitioners primarily interested in PDE problems with a high number of uncertain input variables.

When the full coefficient $a(\mathbf{x}, \omega)$ in (1.3) is preserved or M is large, for the stochastic problem (1.4)–(1.5), a different approach is needed. Instead of bounding the approximation error for standard choices of H_1 and P in (1.7) a priori, many recent works analyse the functions $a_m(\mathbf{x})$ in (1.3) with the aim of constructing tailored sequences of approximation spaces $\{X\}$, not necessarily of the form (1.7), such that the error decays at a rate independent of the number of input random variables. In the works [24, 23], SGFEM algorithms driven by a priori analysis of the functions a_m are proposed, where the error associated with each discretisation parameter is balanced against the total number of degrees of freedom. Alternatively, the works [103, 35, 36, 57] use a priori analysis to establish rates of convergence for so–called *best N–term* approximations of the form (1.6) where, roughly speaking, N is the smallest value of s in (1.6) such that the approximation error on Ω satisfies a prescribed tolerance. In particular, the works [36] and [57] establish the existence of sequences $\{X\}$ such that the approximation error decays to zero at the rate afforded to the underlying FEM for (1.1)–(1.2), with respect to the total number of degrees of freedom. In general however, the sequence $\{X\}$ is unknown explicitly and depends on the problem at hand.

A few algorithms have been proposed in the literature for the adaptive construction of $\{X\}$ for the stochastic problem (1.4)–(1.5). Adaptive methods steered by residual–based a posteriori error estimates are presented in the works [44, 56, 45, 46], whereas the authors of [22] and [21] employ implicit hierarchical–based error estimation to steer the adaptive process. Building on the a posteriori error estimation strategy presented in [22], the aim of this thesis to design new adaptive SGFEMs that construct optimal sequences $\{X\}$ in an efficient way, without the a priori truncation of $a(\mathbf{x}, \omega)$. That is, for appropriate test problems, our aim is to construct sequences for which the energy norm of the approximation error decays at the rate afforded to the underlying FEM for (1.1)–(1.2). Sequences of the form (1.7) are constructed in the works [22] and [21] and rates of convergence for the error independent of the number of input parameters are

reported, but they are not optimal. In this thesis we consider a *multilevel* approach, where the coefficients $u_i(\mathbf{x})$ in (1.6) reside in potentially different FEM spaces.

1.4 Thesis Outline

In Chapter 2, we review important aspects of functional analysis and probability theory that are used throughout the thesis. In Chapter 3, we discuss Galerkin approximation and error estimation from an abstract perspective. To provide a simple example, we solve the deterministic diffusion problem (1.1)–(1.2) using Galerkin finite element approximation and demonstrate how to estimate the error a posteriori. We also introduce adaptive FEMs and give several numerical examples. Chapter 4 is an extended discussion of the work published in [37], and contains novel theoretical estimates of the CBS constants in the bound relating the true errors to the estimated errors computed in Chapter 3. In Chapter 5, we introduce SGFEMs and error estimation for the parametric reformulation of the stochastic diffusion problem. Building on Chapters 3 and 4 and results from [22], we design efficient error estimators which are demonstrated through numerical experiments to be highly accurate. Chapter 6 is an extended discussion of the work published in [38], where we design adaptive multilevel SGFEMs steered by efficient a posteriori error estimates. Numerical experiments demonstrate that our novel method achieves the optimal rate of convergence for the test problems considered. Finally, in Chapter 7 we summarise our main results and offer suggestions and potential directions for future work.

Chapter 2

Background Material

In this Chapter we summarise technical results and background material from functional analysis and probability theory needed for work in Chapters 3–6.

2.1 Functional Analysis

The following results are used extensively in the weak formulation of PDEs, where the aim is to find approximations to solutions in appropriately chosen function spaces, and can be found in [76, 28, 49, 70], for example. To ensure that the weak formulations are well–posed, and to perform error analysis, we first require a *norm*, which provides a notion of the distance between two elements in a vector space.

Definition 2.1: Normed vector space.

A *norm* $||\cdot||_V$ is a functional from a vector space V to $\mathbb{R}^+$ such that

(i) $\|u\|_V = 0$ if and only if $u = 0$,

(ii) $\|\lambda u\|_V = |\lambda|\|u\|_V$ for all $u \in V$ and $\lambda \in \mathbb{R}$, and

(iii) $\|u + v\|_V \leq \|u\|_V + \|v\|_V$ for all $u, v \in V$.

We call a vector space equipped with a norm a *normed vector space*. If only conditions (ii) and (iii) hold, $\|\cdot\|_V$ is a *semi–norm* and denoted $|\cdot|_V$.

We mostly consider function spaces in this thesis. An important class of normed function spaces are $L^p(D)$ spaces where $D \subset \mathbb{R}^d$ $(d = 1, 2, 3)$ is a bounded domain.

Definition 2.2: Lebesgue spaces $L^p(D)$.

The space $L^p(D)$ for $1 \leq p < \infty$ is given by

$$L^p(D) := \big\{u : D \to \mathbb{R};\ ||u||_{L^p(D)} < \infty\big\}, \tag{2.1}$$

with norm $||u||_{L^p(D)} := \big[\int_D |u(\mathbf{x})|^p\ d\mathbf{x}\big]^{\frac{1}{p}}$. We also consider *weighted* $L^p(D)$ spaces, denoted $L^p_w(D)$, which are defined in the same way as $L^p(D)$ using the norm

$$||u||_{L^p_w(D)} := \left[\int_D |u(\mathbf{x})|^p w(\mathbf{x})\ d\mathbf{x}\right]^{\frac{1}{p}},$$

where $w(\mathbf{x})$ is some suitable non–negative weight function.

Definition 2.3: Inner–product space.

An inner product space V is a vector space equipped with an *inner–product* $\langle\cdot,\cdot\rangle_V : V \times V \to \mathbb{R}$ which satisfies the following three axioms:

(i) $\langle u, v\rangle_V = \langle v, u\rangle_V$ for all $u, v \in V$,

(ii) $\langle \alpha u + \beta v, w\rangle_V = \alpha\langle u, w\rangle_V + \beta\langle v, w\rangle_V$ for all $\alpha, \beta \in \mathbb{R}$ and $u, v, w \in V$,

(iii) $\langle u, u\rangle_V \geq 0$ for all $u \in V$ and $\langle u, u\rangle_V = 0 \iff u = 0$.

Since the inner–product $\langle\cdot,\cdot\rangle_V$ induces the norm $\|u\|_V := \langle u, u\rangle_V^{\frac{1}{2}}$ for all $u \in V$, all inner–product spaces are normed vector spaces. In addition, any *complete* inner–product space is called a *Hilbert* space.

Example 2.1: Hilbert space $L^2(D)$.

The space $L^2(D)$ is a Hilbert space with respect to the inner–product

$$\langle u, v\rangle_{L^2(D)} := \int_D u(\mathbf{x})v(\mathbf{x})\ d\mathbf{x}, \qquad \text{for all } u, v \in L^2(D). \tag{2.2}$$

The norm induced by the inner–product in (2.2) coincides with the norm in (2.1) for $p = 2$. Note that functions in $L^2(D)$ need not be continuous.

We can classify functions in $L^2(D)$ based on their smoothness or regularity. To classify functions in this way we require *multi–index* notation and *partial differential operators*.

Definition 2.4: Multi–index.

We call a sequence of non–negative integers

$$\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n) \in \mathbb{N}_0^n, \qquad n \in \mathbb{N},$$

a *multi–index* and define $|\alpha| := \sum_{i=1}^{n} \alpha_i$. We also define the support of the multi–index α by $\text{supp}(\alpha) := \{i \in \mathbb{N};\ \alpha_i \neq 0\}$.

Definition 2.5: Partial differential operator.

For a given multi–index $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$, the *partial differential operator* $\mathcal{D}^\alpha$ is given by

$$\mathcal{D}^\alpha := \prod_{i=1}^{d} \left(\frac{\partial}{\partial x_i} \right)^{\alpha_i} = \frac{\partial^{|\alpha|}}{\partial {x_1}^{\alpha_1} \partial {x_2}^{\alpha_2} \cdots \partial {x_d}^{\alpha_d}},$$

which operates on functions in the variable $\mathbf{x} = (x_1, x_2, \ldots, x_d)^\top \in D \subseteq \mathbb{R}^d$.

Using Definitions 2.2 and 2.5 we may define the *Sobolev* spaces $H^m(D)$, which categorise functions by the square–integrability of their weak derivatives.

Definition 2.6: Sobolev spaces $H^m(D)$.

For a positive integer m, the Sobolev space $H^m(D)$ is the space of all functions $u : D \to \mathbb{R}$ such that u and all weak derivatives of u up to order m are square–integrable, that is,

$$H^m(D) := \left\{ u : D \to \mathbb{R};\ \mathcal{D}^\alpha u \in L^2(D),\ |\alpha| \leq m \right\}.$$

Sobolev spaces are extremely important in the analysis of solutions to weak formulations of PDEs. For certain choices of a, f, D and g, PDE problems such as (1.1) do not admit solutions with derivatives interpreted in the classical sense. Instead, we interpret derivatives in the weak sense and seek solutions in Sobolev spaces. The spaces $H^m(D)$ are also Hilbert spaces with respect to the inner–product

$$\langle u, v \rangle_{H^m(D)} := \sum_{|\alpha| \leq m} \langle \mathcal{D}^\alpha u, \mathcal{D}^\alpha v \rangle_{L^2(D)},$$

where $\langle \cdot, \cdot \rangle_{L^2(D)}$ is given in (2.2), which induces the norm

$$\|u\|_{H^m(D)} := \langle u, u \rangle^{\frac{1}{2}}_{H^m(D)} = \Big[\sum_{|\alpha| \leq m} \langle \mathcal{D}^\alpha u, \mathcal{D}^\alpha u \rangle_{L^2(D)} \Big]^{\frac{1}{2}} = \Big[\sum_{|\alpha| \leq m} \|\mathcal{D}^\alpha u\|^2_{L^2(D)} \Big]^{\frac{1}{2}}.$$

When solving the boundary value problem (BVP) (1.1) we deal with the set of functions

$$H^1_g(D) := \big\{ u \in H^1(D);\ \gamma(u) = g \big\} \subset H^1(D), \tag{2.3}$$

where $\gamma : H^1(D) \to L^2(\partial D)$ is a *trace operator* [70, Lemma 2.37] that maps functions on D to functions on the boundary ∂D. Indeed, if $H^1_g(D)$ is non–empty we may seek weak solutions to (1.1) in $H^1_g(D)$ with Dirichlet boundary data $g : \partial D \to \mathbb{R}$. The set $H^1_g(D)$ is non–empty if $g \in H^{\frac{1}{2}}(\partial D)$ where

$$H^{\frac{1}{2}}(\partial D) := \big\{ \gamma(u);\ u \in H^1(D) \big\}.$$

An important space is $H^1_0(D)$ which is a Hilbert space with respect to the inner product

$$\langle u, v \rangle_{H^1_0(D)} := \int_D \nabla u(\mathbf{x}) \cdot \nabla v(\mathbf{x})\ d\mathbf{x}, \qquad \text{for all } u, v \in H^1_0(D).$$

The induced norm $\|u\|_{H^1_0(D)} := \langle u, u \rangle^{\frac{1}{2}}_{H^1_0(D)}$ is a semi–norm on $H^1(D)$.

The weak formulation of (1.1) involves a bilinear form and a linear functional. To prove well–posedness using the Lax–Milgram Lemma (Lemma 2.1) we require the bilinear form to be *bounded* and *coercive*, and the linear functional to be bounded.

Definition 2.7: Bounded and coercive bilinear forms.

Let V be a Hilbert space with respect to $\langle \cdot, \cdot \rangle_V$, with induced norm $\| \cdot \|_V$. A bilinear form $B(\cdot, \cdot) : V \times V \to \mathbb{R}$ is said to be *bounded* on V if there exists a constant $C_1 > 0$ such that

$$|B(u, v)| \leq C_1 \|u\|_V \|v\|_V, \qquad \text{for all } u, v \in V,$$

and *coercive* on V if there exists a constant $C_2 > 0$ such that

$$B(u, u) \geq C_2 \|u\|^2_V, \qquad \text{for all } u \in V.$$

Definition 2.8: Bounded linear functional.

Let V be a Hilbert space with respect to $\langle \cdot, \cdot \rangle_V$, with induced norm $\| \cdot \|_V$. A linear functional $F(\cdot) : V \to \mathbb{R}$ is said to be bounded on V if there exists a constant $C_3 > 0$ such that

$$F(u) \leq C_3 \|u\|_V, \qquad \text{for all } u \in V.$$

Lemma 2.1: Lax–Milgram [28, Lemma 2.7.7].

Let V be a Hilbert space with respect to $\langle \cdot, \cdot \rangle_V$ and let $B(\cdot, \cdot) : V \times V \to \mathbb{R}$ be a bounded and coercive bilinear form, and $F(\cdot) : V \to \mathbb{R}$ be a bounded linear functional. Then there exists a unique $u \in V$ that satisfies

$$B(u, v) = F(v), \qquad \textit{for all } v \in V.$$

Key to proving the boundedness of the bilinear forms $B(\cdot, \cdot)$ that we encounter is the Cauchy–Schwarz inequality.

Lemma 2.2: Cauchy–Schwarz inequality [28, 1.1.5].

Let V be a Hilbert space with respect to $\langle \cdot, \cdot \rangle_V$, with induced norm $\| \cdot \|_V$. Then

$$|\langle u, v \rangle_V| \leq \|u\|_V \|v\|_V, \qquad \textit{for all } u, v \in V.$$

In addition, the linear functionals $F(\cdot)$ we encounter involve functions in $L^2(D)$. To prove their boundedness we require the following result that relates the norms $\| \cdot \|_{L^2(D)}$ and $\| \cdot \|_{H_0^1(D)}$ for functions in $H_0^1(D)$.

Theorem 2.1: Poincaré–Friedrichs inequality [49, Lemma 1.2].

For a bounded domain $D \subset \mathbb{R}^d$, there exists a constant $C_p > 0$ such that

$$\|u\|_{L^2(D)} \leq C_p \|u\|_{H_0^1(D)}, \qquad \textit{for all } u \in H_0^1(D).$$

The final important result that we require is the *strengthened* Cauchy–Schwarz inequality, which is used extensively in error analysis for weak solutions of PDEs, and in the analysis of certain classes of hierarchical preconditioners [3, 86, 5] for linear systems

of equations associated with finite–dimensional weak problems.

Theorem 2.2: Strenghened Cauchy–Schwarz [48, Theorem 1].

Let V be a Hilbert space equipped with inner product $\langle\cdot,\cdot\rangle_V$ and induced norm $||\cdot||_V$ and let V_1, V_2 be a pair of finite–dimensional subspaces of V satisfying $V_1 \cap V_2 = \{0\}$. Then, there exists a constant $\gamma \in [0,1)$, depending only on V_1 and V_2 such that

$$|\langle v_1, v_2\rangle_V| \leq \gamma ||v_1||_V ||v_2||_V, \qquad \text{for all } v_1 \in V_1, \qquad \text{for all } v_2 \in V_2. \tag{2.4}$$

The bound (2.4) is stronger than the traditional Cauchy-Schwarz inequality in the sense that the modulus of the inner–product on the left–hand side is strictly less than the product of the norms on the right–hand side. Note that the constant $\gamma \in [0,1)$ in (2.4) is not unique. The smallest such constant satisfying (2.4) is

$$\gamma_{\min} := \sup_{v_1 \in V_1} \sup_{v_2 \in V_2} \frac{|\langle v_1, v_2\rangle_V|}{||v_1||_V ||v_2||_V}, \tag{2.5}$$

and is known as the Cauchy-Buniakowskii-Schwarz (CBS) constant. We can interpret the CBS constant as the cosine of the angle between the spaces V_1 and V_2. Indeed, if V_1 and V_2 are orthogonal with respect to $\langle\cdot,\cdot\rangle_V$, then $\gamma_{\min} = 0$.

2.2 Probability Theory

In order to formulate PDEs with random inputs in the weak sense, we must first familiarise ourselves with some essential probability theory. We start this section by introducing standard concepts such as probability spaces and random variables. We then extend the theory to random fields. The following standard results can be found in [59, 29, 70], for example.

2.2.1 Random Variables

The analysis of mathematical models with uncertain inputs is more straightforward when the sets of all possible outcomes of the inputs are known. If the sets of outcomes are unknown or potentially infinite, formulating the model and solving it can be difficult. To this end, we work in a general framework where we are not required (for now) to explicitly define each possible outcome. We work with measurable spaces of

the form $(\Omega, \mathcal{F})$ where Ω denotes an abstract set of all possible outcomes, and $\mathcal{F}$ is a collection of measurable subsets of Ω (formally a σ–algebra). We assign to $(\Omega, \mathcal{F})$ a *probability measure* $\mathbb{P}$ to form a *probability space* $(\Omega, \mathcal{F}, \mathbb{P})$, where $\mathbb{P} : F \to [0, 1]$ returns the probability that an event $F \in \mathcal{F}$ will happen (the measure of a set), and $\mathbb{P}(\Omega) = 1$. Measurable functions that map events in Ω to $\mathbb{R}$ are called real–valued random variables.

Definition 2.9: Real–valued random variable.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A measurable function $X : \Omega \to \mathbb{R}$ is called a real–valued random variable. For $\omega \in \Omega$, the observation $X(\omega)$ is called a realisation of X.

In applied mathematics, we want to choose random variables which suitably represent the uncertain features of the processes we want to model. Two important statistical properties are the *expectation* and the *variance* of a random variable.

Definition 2.10: Expectation and variance.

Let X be a real–valued random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If X is integrable and square integrable on $(\Omega, \mathcal{F}, \mathbb{P})$, then the expectation and variance of X are given by

$$\mathbb{E}[X] := \int_\Omega X(\omega)\ d\mathbb{P}(\omega), \tag{2.6}$$

$$\mathrm{Var}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2, \tag{2.7}$$

respectively. For real–valued random variables, $\mathbb{E}[X]$ and $\mathrm{Var}(X)$ are constants.

Another important property is the covariance of a pair of random variables.

Definition 2.11: Covariance.

Let X, Y be real–valued random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The covariance of X and Y is given by

$$\mathrm{Cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])], \tag{2.8}$$

where $\mathbb{E}[XY] = \int_\Omega X(\omega)Y(\omega)\ d\mathbb{P}(\omega)$.

Integrals such as (2.6), defined over the abstract set Ω, are not usually computable. For real–valued random variables with a probability density function (pdf) $p(x)$ with respect to Lebesgue measure, we may perform a change of measure and instead compute expectations over the observation space, that is, we compute

$$\mathbb{E}[X] = \int_{\mathbb{R}} xp(x)\ dx,$$

which may be evaluated using standard calculus. In this thesis we use *uniform* random variables, where all values in some prescribed bounded sub–interval of $\mathbb{R}$ are realised with equal probability.

Definition 2.12: Uniformly distributed random variable.

A random variable X is uniformly distributed on $[a, b] \subset \mathbb{R}$ if

$$p(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b], \\ 0, & \text{otherwise.} \end{cases}$$

We write $X \sim U(a, b)$ and note that $\mathbb{E}[X] = \frac{b+a}{2}$ and $\mathrm{Var}(X) = \frac{(b-a)^2}{12}$.

Example 2.2: Zero mean and unit variance.

The random variable $X \sim U(-\sqrt{3}, \sqrt{3})$ has zero mean and unit variance, and pdf $p(x) = \frac{1}{2\sqrt{3}}$ for $-\sqrt{3} \leq x \leq \sqrt{3}$.

In Chapter 5 we meet H–valued random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where H is a Hilbert space. Such functions may be classified using *Bochner* spaces, which are generalisations of Lebesgue spaces (given in Definition 2.2) for possibly other measures. In this thesis we work with the following class of Bochner spaces.

Definition 2.13: Bochner spaces $L^p(\Omega, H)$.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and H be a Hilbert space with norm $\|\cdot\|_H$. Then $L^p(\Omega, H)$ with $1 \leq p < \infty$ denotes the space of H-valued random variables $X : \Omega \to H$ with $\mathbb{E}\left[||X||_H^p\right] < \infty$ and is a Banach space with the norm

$$\|X\|_{L^p(\Omega,H)} := \left[\int_\Omega \|X(\omega)\|_H^p\ d\mathbb{P}(\omega)\right]^{\frac{1}{p}} = \mathbb{E}\left[\|X\|_H^p\right]^{\frac{1}{p}}. \tag{2.9}$$

Example 2.3: Bochner space $L^2(\Omega, H_0^1(D))$.

Let $D \subset \mathbb{R}^2$ and let $p = 2$ and $H = H_0^1(D)$ in (2.9). Then, the space $L^2(\Omega, H_0^1(D))$ of square–integrable $H_0^1(D)$-valued random variables is a Hilbert space with respect to the inner product

$$\langle X, Y \rangle_{L^2(\Omega, H_0^1(D))} := \int_\Omega \int_D \nabla X \cdot \nabla Y \; d\mathbf{x} \; d\mathbb{P}(\omega) = \mathbb{E}\left[\int_D \nabla X \cdot \nabla Y \; d\mathbf{x} \right].$$

For any $X(\mathbf{x}, \omega) \in L^2(\Omega, H_0^1(D))$, we have that $X(\mathbf{x}, \cdot) \in L^2(\Omega)$ for every $\mathbf{x} \in D$, and $X(\cdot, \omega) \in H_0^1(D)$ for every $\omega \in \Omega$.

2.2.2 Random Fields

H-valued random variables are closely related to random fields. A random field is an extension of a stochastic process which is indexed in one dimension (usually in time for $t \in \mathbb{R}$).

Definition 2.14: Random field.

A (real–valued) *random field* $\{a(\mathbf{x});\ \mathbf{x} \in D\}$ for $D \subset \mathbb{R}^d$ is a set of real–valued random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We use the notation $a(\mathbf{x}, \omega) : D \times \Omega \to \mathbb{R}$ to stress the dependence on both $\mathbf{x}$ *and* ω.

We consider *second–order* random fields, which have well–defined mean and covariance functions.

Definition 2.15: Second–order random field.

A random field $a(\mathbf{x}, \omega) : D \times \Omega \to \mathbb{R}$ is second–order if $a(\mathbf{x}, \cdot) \in L^2(\Omega)$ for every $\mathbf{x} \in D$. By $\mu(\mathbf{x}) := \mathbb{E}[a(\mathbf{x}, \omega)]$ we denote the mean of $a(\mathbf{x}, \omega)$, and by

$$C(\mathbf{x}_1, \mathbf{x}_2) := \mathrm{Cov}(a(\mathbf{x}_1, \omega), a(\mathbf{x}_2, \omega)) = \mathbb{E}[(a(\mathbf{x}_1, \omega) - \mu(\mathbf{x}_1))\,(a(\mathbf{x}_2, \omega) - \mu(\mathbf{x}_2))]$$

for $\mathbf{x}_1, \mathbf{x}_2 \in D$, its covariance, both of which are deterministic functions.

Any function in the space $L^2(\Omega, H(D))$ for some Hilbert space $H(D)$ is a second–order random field. In this thesis we seek weak solutions to PDE problems with random inputs in spaces of this form. We also use second–order random fields to represent

the inputs. Specifically, we choose inputs from $L^2(\Omega, L^2(D))$ since functions in this space admit a Fourier–like expansion called the *Karhunen–Loève* (KL) expansion [69]. KL expansions provide a practical means to implement random fields numerically in computer software. Assuming that $C(\mathbf{x}_1, \mathbf{x}_2) \in L^2(D \times D)$, where $C(\mathbf{x}_1, \mathbf{x}_2)$ is the covariance function of the random field being expanded (see Definition 2.15), a KL expansion uses the orthonormal basis provided by the normalised eigenfunctions of the integral operator $\mathcal{C} : L^2(D) \to L^2(D)$ defined by

$$(\mathcal{C}\phi)(\mathbf{x}_1) = \int_D C(\mathbf{x}_1, \mathbf{x}_2)\phi(\mathbf{x}_2)\, d\mathbf{x}_2, \qquad \phi \in L^2(D). \tag{2.10}$$

Theorem 2.3: KL expansion of random fields [70, Theorem 7.52].

Let $D \subseteq \mathbb{R}^d$. Consider a (second–order) random field $a \in L^2(\Omega, L^2(D))$ with mean function $\mu(\cdot)$ and (continuous) covariance function $C(\cdot, \cdot)$. Then

$$a(\mathbf{x}, \omega) = \mu(\mathbf{x}) + \sum_{m=1}^{\infty} \sqrt{\nu_m}\phi_m(\mathbf{x})\xi_m(\omega), \tag{2.11}$$

where the sum converges in $L^2(\Omega, L^2(D))$,

$$\xi_m(\omega) := \frac{1}{\sqrt{\nu_m}} \langle a(\mathbf{x}, \omega) - \mu(\mathbf{x}), \phi_m(\mathbf{x}) \rangle_{L^2(D)}$$

and $\{\nu_m, \phi_m\}$ are the eigenvalues and eigenfunctions of the integral operator $\mathcal{C}$ in (2.10) with $\nu_1 \geq \nu_2 \geq \cdots \geq 0$. The random variables $\xi_m(\omega)$ have mean zero, unit variance and are pairwise uncorrelated.

Rather than expand random fields $a(\mathbf{x}, \omega)$ with prescribed distributions, in this work we generate random fields by first choosing a mean $\mu(\cdot)$ and covariance function $C(\cdot, \cdot)$, as well as a set of independent random variables $\xi_m(\omega)$, and construct

$$a(\mathbf{x}, \omega) := \mu(\mathbf{x}) + \sum_{m=1}^{\infty} \sqrt{\nu_m}\phi_m(\mathbf{x})\xi_m(\omega), \tag{2.12}$$

where the eigenpairs $\{\nu_m, \phi_m\}$ are as described in Theorem 2.3. Under this construction the distribution of $a(\mathbf{x}, \omega)$ in (2.12) depends on our choices of $\xi_m, \mu(\mathbf{x})$ and $C(\mathbf{x}_1, \mathbf{x}_2)$.

The covariance function associated with a random field $a \in L^2(\Omega, L^2(D))$ resides in $L^2(D \times D)$; see Theorem 5.28 of [70]. In order to construct $a(\mathbf{x}, \omega)$ in (2.12) using the

eigenpairs associated with (2.10), we must *choose* covariance functions $C \in L^2(D \times D)$. Additionally, for expansions of the form (2.12) to be used in computer software, the infinite sum needs truncating.

One option is to truncate (2.12) a priori say, after M terms, and construct

$$a_M(\mathbf{x}, \omega) := \mu(\mathbf{x}) + \sum_{m=1}^{M} \sqrt{\nu_m}\phi_m(\mathbf{x})\xi_m(\omega). \tag{2.13}$$

Whilst we do not take this approach in our work in Chapters 5 and 6, it is useful to consider for the following analysis.

If $C(\mathbf{x}_1, \mathbf{x}_2) = c(\mathbf{x}_1 - \mathbf{x}_2)$ for some function $c(\cdot)$, the error in the total variance captured by the truncated expansion $a_M(\mathbf{x}, \omega)$ can be stated in terms of the retained eigenvalues $\{\nu_m\}_{m=1}^{M}$. Since $\mathbb{E}[\xi_m(\omega)] = 0$ we find that

$$\int_D \mathrm{Var}\big(a(\mathbf{x}, \omega)\big) - \mathrm{Var}\big(a_M(\mathbf{x}, \omega)\big)\, d\mathbf{x} = \int_D \mathbb{E}\left[\left(\sum_{m=1}^{\infty} \sqrt{\nu_m}\phi_m(\mathbf{x})\xi_m(\omega)\right)^2 - \left(\sum_{m=1}^{M} \sqrt{\nu_m}\phi_m(\mathbf{x})\xi_m(\omega)\right)^2\right] d\mathbf{x}.$$

Covariance functions are, by definition, symmetric and positive semidefinite. If $C(\cdot, \cdot)$ is also continuous and D is bounded, then, by Mercer's theorem [70, Theorem 1.80]

$$\int_D \mathrm{Var}\big(a(\mathbf{x}, \omega)\big) - \mathrm{Var}\big(a_M(\mathbf{x}, \omega)\big)\, d\mathbf{x} = c(\mathbf{0})\mathrm{leb}(D) - \sum_{m=1}^{M} \nu_m,$$

where $\mathrm{leb}(D)$ denotes the length, area, or volume of D (corresponding to $d = 1, 2, 3$ in Theorem 2.3). A useful measurement is the relative error

$$e_M := \frac{\int_D \mathrm{Var}(a(\mathbf{x}, \omega)) - \mathrm{Var}(a_M(\mathbf{x}, \omega))\, d\mathbf{x}}{\int_D \mathrm{Var}(a(\mathbf{x}, \omega))\, d\mathbf{x}} = \frac{c(\mathbf{0})\mathrm{leb}(D) - \sum_{m=1}^{M} \nu_m}{c(\mathbf{0})\mathrm{leb}(D)} \tag{2.14}$$

which we utilise later in this section. Clearly, e_M depends on the rate at which the eigenvalues $\{\nu_m\}_{m=1}^{\infty}$ decay, which in turn depends on the regularity of the chosen covariance function; see [52] or [88, 89, 64] for general results regarding the eigenvalues of kernels $C(\mathbf{x}_1, \mathbf{x}_2)$.

Most covariance functions $C(\mathbf{x}_1, \mathbf{x}_2)$ do not lead to an eigenproblem

$$\int_D C(\mathbf{x}_1, \mathbf{x}_2)\phi(\mathbf{x}_2)\ d\mathbf{x}_2 = \nu\phi(\mathbf{x}_1) \tag{2.15}$$

that can be solved analytically. In such cases, the eigenpairs $\{\nu_m, \phi_m\}$ can be approximated numerically using collocation or Galerkin approximation [94, 43, 70], both of which require the solution of a discrete eigenproblem. In this thesis we only consider the *separable exponential* covariance function, a covariance function that *does* lead to an eigenproblem with analytical solutions [54].

Definition 2.16: Separable exponential covariance function.

For $D \subseteq \mathbb{R}^d$, the separable exponential covariance function is given by

$$C(\mathbf{x}_1, \mathbf{x}_2) = \sigma^2 \exp\left(-\frac{||\mathbf{x}_1 - \mathbf{x}_2||_1}{\ell}\right), \qquad \mathbf{x}_1, \mathbf{x}_2 \in D, \tag{2.16}$$

where ℓ is the correlation length and $C(\mathbf{x}, \mathbf{x}) = c(\mathbf{0}) = \sigma^2 \in \mathbb{R}$ for $\mathbf{x} \in D$ is the variance.

Since the exponential in (2.16) can be written as the product of exponential functions corresponding to each spatial dimension, if appropriate, we may assign a different correlation length in each dimension. When D is the Cartesian product of bounded intervals in $\mathbb{R}$, solutions to the eigenproblem (2.15) associated with (2.16) can be generated by tensorising solutions of one–dimensional eigenproblems. In particular, when $D := [-a, a]^2$ for $0 < a < \infty$ we need only solve the eigenproblem

$$\int_{-a}^{a} \exp\left(-\frac{|x_1 - x_2|}{\ell}\right)\phi(x_2)\ dx_2 = \bar{\nu}\phi(x_1), \qquad x_1 \in [-a, a], \tag{2.17}$$

and tensorise the resulting eigenpairs to generate eigenpairs $(\bar{\nu}_m, \phi_m)$ on D. Note that the constant σ^2 in (2.16) has been neglected in (2.17), and thus we simply construct $a(\mathbf{x}, \omega)$ in (2.12) using $\nu_m = \sigma^2\bar{\nu}_m$.

Example 2.4: Eigenfunctions, separable exponential covariance.

Let $D = [-1, 1]^2$ and $\sigma = \ell = 1$ in (2.16). To compute eigenpairs of the corresponding eigenproblem in two dimensions we tensorise the eigenpairs $(\bar{\nu}, \phi)$ generated by solving (2.17) analytically [54]. In Figure 2.1 we plot the eigenfunctions corresponding to the sixteen largest eigenvalues of the two–dimensional problem, ordered left–to–right, top–to–bottom.

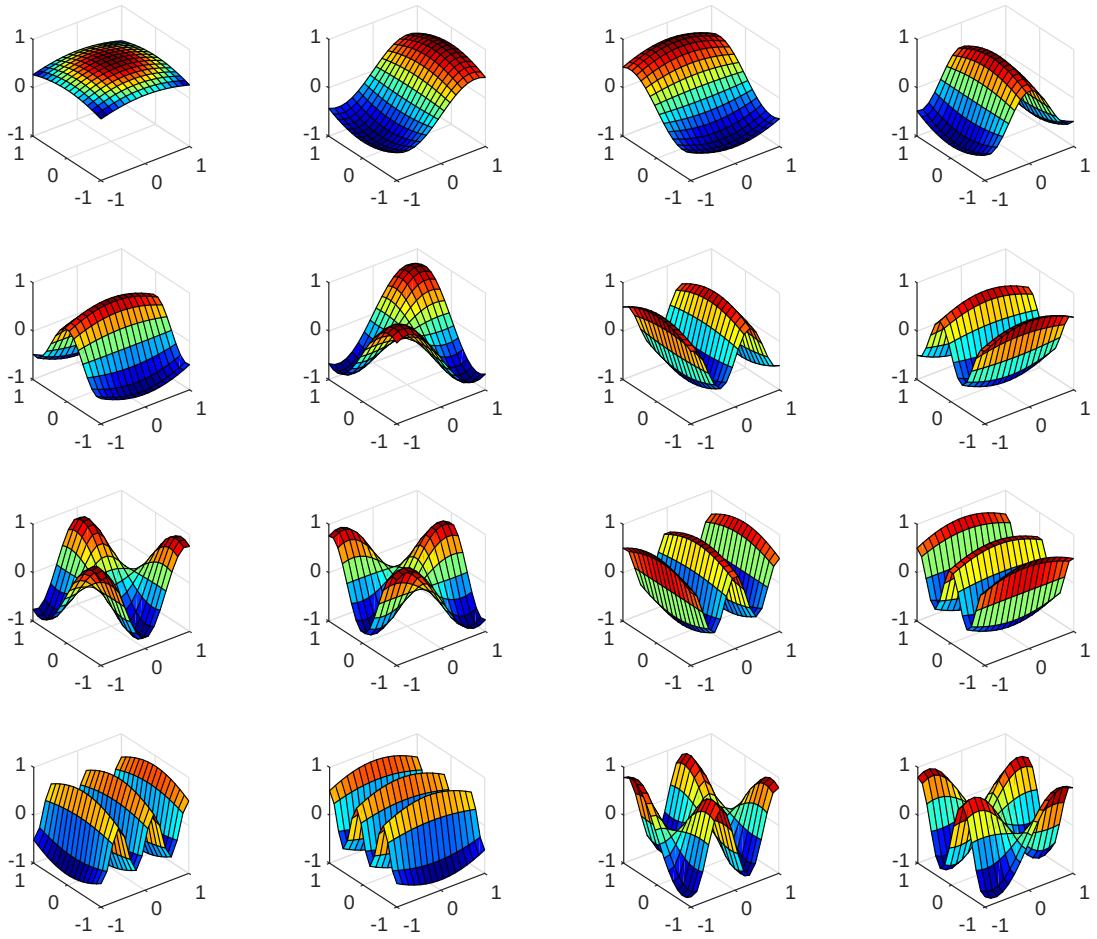


Figure 2.1: Eigenfunctions of the integral operator (2.10) when $C(\mathbf{x}_1, \mathbf{x}_2)$ is the separable exponential covariance function given by (2.16) for $D = [-1, 1]^2$ and $\sigma = \ell = 1$. The eigenfunctions correspond to the sixteen largest eigenvalues and are ordered left–to–right, top–to–bottom.

Example 2.5: Realisations, separable exponential covariance.

Let $D = [-1, 1]^2$ and define $a(\mathbf{x}, \omega)$ as in (2.12) with $\xi_m(\omega) \sim U(-\sqrt{3}, \sqrt{3})$ independent, mean $\mu = 0$, and choose the covariance function $C(\mathbf{x}_1, \mathbf{x}_2)$ in (2.16) with σ and ℓ as in Example 2.4. Once the eigenpairs associated with $C(\mathbf{x}_1, \mathbf{x}_2)$ are computed, we may generate realisations of $a_M(\mathbf{x}, \omega)$ in (2.13) by generating realisations of $\xi_m(\omega)$ for $m = 1, 2, \ldots, M$. In Figure 2.2 we plot one realisation of $a_M(\mathbf{x}, \omega)$ using the same realisations of $\xi_m(\omega)$ for $M = 10, 109, 954$ terms (left–to–right), which corresponds to retaining 75%, 95% and 99% of the total variance of $a(\mathbf{x}, \omega)$ in terms of (2.14), respectively. Note how the realisation becomes more oscillatory as M increases.

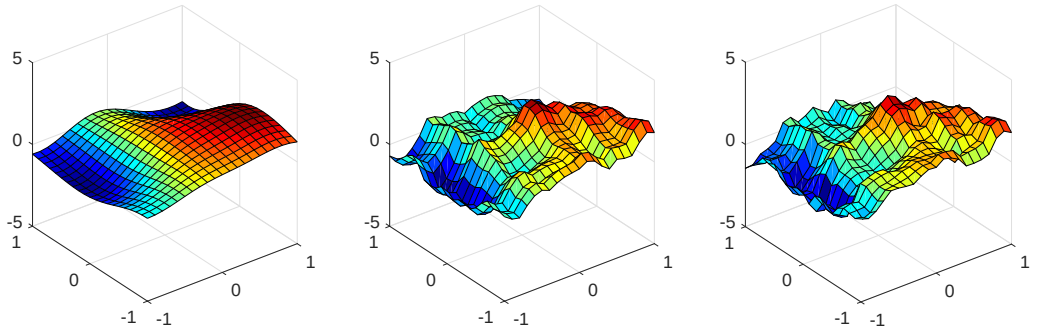


Figure 2.2: One realisation of $a_M(\mathbf{x}, \omega)$ in Example 2.5 for $M = 10,109,954$ (left–to–right). We choose $\mu = 0$, $\xi_m(\omega) \sim U(-\sqrt{3}, \sqrt{3})$ independent, and the covariance function (2.16) with $\sigma = \ell = 1$.

It is well known [70] that for the separable exponential covariance function (2.16), when D is rectangular the square roots of the eigenvalues $\{\nu_m\}_{m=1}^{\infty}$ in (2.12) decay asymptotically like m^{-1}, which is very slow. Whilst this rate is independent of the correlation length ℓ, the pre–asymptomatic regime does depend on ℓ in that larger correlation lengths result in larger initial decreases in the eigenvalues [70, Example 7.56]. As a result, when ℓ is small, it takes many terms M to reduce the relative error e_M associated with $a_M(\mathbf{x}, \omega)$ to a prescribed tolerance, and PDE problems with inputs of the form described in Example 2.5 are particularly difficult to solve numerically.

Since we are mainly interested in developing numerical methods for PDE problems with random inputs, rather than the development of PDE models themselves, we also consider *synthetic* expansions of the form (2.12), where the pairs $(\sqrt{\nu_m}, \phi_m)$ are carefully chosen to afford $a(\mathbf{x}, \omega)$ particular characteristics. That is, we consider examples not related to specific covariance functions, and therefore not strictly speaking KL expansions. In the following example, we introduce a synthetic expansion where the terms $\sqrt{\nu_m}$ decay more quickly than those in Example 2.5.

Example 2.6: Synthetic expansion (quadratic decay), realisations.

Let $D = [0, 1]^2$ and define $a(\mathbf{x}, \omega)$ as in (2.12) with $\xi_m(\omega) \sim U(-1, 1)$ independent and mean $\mu = 0$. Following [44], we choose $\sqrt{\nu_m} = 0.547m^{-2}$ and

$$\phi_m(\mathbf{x}) = \cos(2\pi\beta_m^1 x_1)\cos(2\pi\beta_m^2 x_2) \tag{2.18}$$

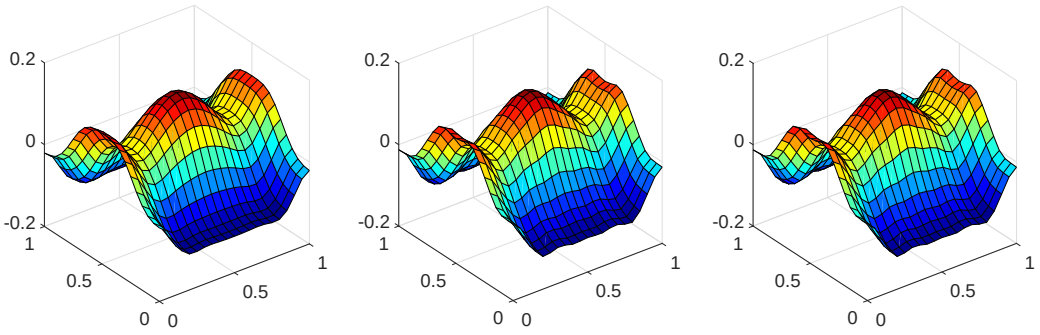


Figure 2.3: One realisations of $a_M(\mathbf{x}, \omega)$ in Example 2.6 for $M = 10,109,954$ (left–to–right). We choose $\mu = 0$ and $\xi_m(\omega) \sim U(-1, 1)$ independent.

for $\mathbf{x} = (x_1, x_2)^\top \in D$ and $m \in \mathbb{N}$, with

$$\beta_m^1 = m - \tfrac{1}{2}k_m(k_m + 1), \qquad \beta_m^2 = k_m - \beta_m^1, \qquad k_m = \lfloor -\tfrac{1}{2} + (\tfrac{1}{4} + 2m)^{\frac{1}{2}} \rfloor.$$

In Figure 2.3 we plot one realisation of $a_M(\mathbf{x}, \omega)$ using the same realisations of $\xi_m(\omega)$ for $M = 10,109,954$ terms (left–to–right). Since $\sqrt{\nu_m} = \mathcal{O}(m^{-2})$ and terms corresponding to small values of m in $a_M(\mathbf{x}, \omega)$ dominate the overall behaviour, we observe little variation in the plots as M increases.

For suitable choices of $\mu(\mathbf{x})$ and $\xi_m(\omega)$ in (2.12), the terms $\sqrt{\nu_m}$ and $\phi_m(\mathbf{x})$ defined in Example 2.6 ensure that $a_M(\mathbf{x}, \omega)$ always has positive realisations as M increases. This is necessary to guarantee that the weak formulations of the PDE problems considered in Chapters 5 and 6 are well–posed. In contrast, the expansion $a_M(\mathbf{x}, \omega)$ defined in Example 2.5 depends on our choice of σ as well, and ensuring positive realisations as M increases is not as straightforward. Expansions like the one described in Example 2.6 are therefore popular choices for designing test problems with random inputs that are guaranteed to have positive realisations. Such test problems can then be used to test new solution algorithms.

Example 2.7: Synthetic expansion (quartic decay).

This expansion is a simple variant of the one considered in Example 2.6. Let D and $\xi_m(\omega)$ be as in Example 2.6 and define $a(\mathbf{x}, \omega)$ as in (2.12). We choose $\phi_m(\mathbf{x})$ to be as in (2.18) and select the (faster) decaying coefficients $\sqrt{\nu_m} = 0.832m^{-4}$.

We consider one more synthetic expansion which was introduced in [70].

Example 2.8: Synthetic expansion (exponential decay).

Let $D = [0,1]^2$ and define $a(\mathbf{x}, \omega)$ as in (2.12) with $\xi_m(\omega) \sim U(-\sqrt{3}, \sqrt{3})$ independent. We then define

$$\nu_{ij} := \tfrac{1}{4} \exp(-\pi(i^2 + j^2)\ell^{-2}), \qquad \phi_{ij}(\mathbf{x}) := 2\cos(i\pi x_1)\cos(j\pi x_2) \tag{2.19}$$

(with correlation length ℓ) for $i, j = 1, 2, \ldots$, where $\phi_{00} = 1$, $\nu_{00} = \frac{1}{4}$ and rewrite the pairs (ν_{ij}, ϕ_{ij}) in terms of the single index m with the new sequence $\{\nu_m\}_{m=1}^{\infty}$ ordered descendingly.

Chapter 3

Galerkin Approximation & Error Estimation

Many complex and interesting PDE problems that arise in both academia and industry do not admit exact solutions, despite the often proven assertion that a unique solution exists. In such situations, practitioners employ numerical methods such as finite element methods (FEMs) to approximate the true solution. When the true solution, or a quantity of interest (QoI) that depends on the true solution, is a key ingredient in the design of, say, consumer goods, it is of paramount importance that we have an informed understanding of the magnitude of the *approximation error* incurred by the employed numerical method. A priori error analysis provides useful bounds for the asymptotic behaviour of the approximation error, but does not provide a computable estimate of the error itself. A posteriori error analysis can provide such estimates. In this chapter we review classical *Galerkin approximation* - a discretisation technique which underpins FEMs - and a well known a posteriori error estimation strategy that leads to computable estimates of the error [2, 108]. These strategies are then utilised in Chapters 5 and 6 to design a posteriori error estimators for stochastic Galerkin finite element methods (SGFEMs) as well as new efficient adaptive algorithms.

3.1 Galerkin Approximation

Our aim is to describe an a posteriori error estimation strategy that can be applied to a broad class of problems, and demonstrate how it may be implemented on test problems.

We are concerned with a class of problems that arise from Galerkin approximation of second–order elliptic PDEs, which we now describe.

Let V be a Hilbert space with norm $||\cdot||_V$ and let $B : V \times V \to \mathbb{R}$ and $F : V \times V \to \mathbb{R}$ denote a bilinear form and linear functional, respectively (recall the relevant definitions from Section 2.1). Consider the problem:

$$\text{find } u \in V : \qquad B(u, v) = F(v), \qquad \text{for all } v \in V. \tag{3.1}$$

We are interested in the class of problems for which $B(\cdot, \cdot)$ is symmetric, bounded and coercive over V, and $F(\cdot)$ is bounded over V. By Lemma 2.1 there then exists a unique solution to (3.1). Due to its symmetry, $B(\cdot, \cdot)$ is also an inner product on V and induces the so-called energy norm, which we now define.

Definition 3.1: Energy norm.

The energy norm corresponding to problem (3.1) is given by

$$||v||_B = B(v, v)^{\frac{1}{2}} \qquad \text{for all } v \in V. \tag{3.2}$$

Since V is infinite-dimensional, the true solution $u \in V$ cannot be computed. As an alternative we seek a Galerkin approximation to u. Let X denote an N_X–dimensional subspace of V and consider the new discrete problem:

$$\text{find } u_X \in X : \qquad B(u_X, v) = F(v), \qquad \text{for all } v \in X. \tag{3.3}$$

The problem in (3.3) leads to a tractable system of N_X equations that enables us to determine $u_X \in X$. Since X is a closed subspace of V, the approximation $u_X \in X$ is also unique; see [28, Corollary 2.7.13]. Our choice of $X \subset V$ determines the quality of the approximation $u_X \approx u$. In order to establish whether our Galerkin approximation is satisfactory in terms of some error tolerance, we examine the error $e = u - u_X$, which, due to the properties of the bilinear form satisfies the infinite-dimensional problem:

$$\text{find } e \in V : \qquad B(e, v) = F(v) - B(u_X, v), \qquad \text{for all } v \in V. \tag{3.4}$$

For the remainder of this section we analyse $e \in V$ satisfying (3.4), and the efficient approximation of $e \in V$ is the focus of Section 3.2. The following error analysis is well known and can be found in [49, 28], for example.

The true error $e \in V$ possesses an orthogonality property inherent to Galerkin approximation. Combining (3.3) and (3.4) it is easy to show that:

$$B(e, v) = 0, \qquad \text{for all } v \in X \qquad \text{(Galerkin orthogonality)}. \tag{3.5}$$

That is, the error is orthogonal to all functions in X with respect to the bilinear form $B(\cdot, \cdot)$. Using (3.5) we may further show that $u_X \in X$ is the best approximation to u in X with respect to the energy norm $||\cdot||_B$. Indeed, it follows from the definition of $||\cdot||_B$ that

$$\begin{aligned} ||u - u_X||_B^2 &= B(u - u_X, u - u_X) \\ &= B(u - u_X, u - v) + \underbrace{B(u - u_X, v - u_X)}_{=\ 0 \text{ due to (3.5)}} \qquad \text{(for all } v \in X), \end{aligned}$$

and thus, applying the Cauchy–Schwarz inequality to the right-hand side and cancelling out a factor of $||u - u_X||_B$ yields the desired result:

$$||u - u_X||_B \leq ||u - v||_B, \qquad \text{for all } v \in X. \tag{3.6}$$

Equivalently, $||u - u_X||_B = \inf_{v \in X} ||u - v||_B$, since $u_X \in X$. This result is extremely important. It tells us that if we were to choose a second subspace $W \subset V$ such that $X \subset W$ (i.e., W is richer than X) and compute a new Galerkin approximation $u_W \in W$ to u, then

$$||u - u_W||_B = \inf_{v \in W} ||u - v||_B \leq \inf_{v \in X} ||u - v||_B = ||u - u_X||_B, \tag{3.7}$$

and thus the approximation error cannot increase when measured in the energy norm. In addition, due to the symmetry of the bilinear form $B(\cdot, \cdot)$, the square of the energy error admits the decomposition

$$\begin{aligned} ||u - u_X||_B^2 &= B(u - u_X, u - u_X) \\ &= B(u, u) - 2B(u, u_X) + B(u_X, u_X) \\ &= B(u, u) - 2B(u - u_X, u_X) - B(u_X, u_X) \\ &= ||u||_B^2 - ||u_X||_B^2 - 2\underbrace{B(u - u_X, u_X)}_{=0 \text{ due to (3.5)}}, \end{aligned}$$

and thus

$$||u - u_X||_B = \sqrt{||u||_B^2 - ||u_X||_B^2}. \tag{3.8}$$

Generally speaking, a posteriori error estimation strategies for Galerkin approximation fall into one of two categories; explicit methods or implicit methods . Explicit methods involve direct computations using data which is available to us after solving for u_X. That is, the input data for the underlying problem and the Galerkin approximation $u_X \in X$ itself; see [9, 107, 2] for example. Typically, the *residual* (the right–hand side of (3.4), roughly speaking) is exploited to provide upper–bounds for $\|u - u_X\|_B$. In contrast, implicit strategies require the solution of some algebraic system of equations associated with the error problem (3.4) itself [16, 1, 49]. Whilst explicit estimates are often more straightforward to compute, they can lack the accuracy of a well–designed implicit estimate [49]. Additionally, implicit estimates can lead to precise estimates of the error reduction that would be achieved be performing certain enrichments of X. Thus, they are especially useful when designing adaptive FEMs (see Chapters 5 and 6 for concrete examples). Whilst explicit estimates are often used in the design of adaptive FEMs as well, the bounds relating the true error reductions to the corresponding estimates usually involve unknown constants, leading to less confidence in the efficiency of the adaptive process. For the complex stochastic problem considered in Chapter 5 it is crucial that the dimension of X is kept to a minimum. Thus, in this work, we opt to take an implicit approach.

In the next section we review a well–known implicit method [15, 2, 108], often referred to as a Hierarchical method, for the efficient estimation of $||e||_B$. Akin to our choice of X in (3.3), the success of this method depends on a finite–dimensional subspace of V of our choosing. The method is well designed if suitable balance has been struck between the accuracy of the estimate and the cost to compute it.

3.2 Implicit a Posteriori Error Estimation

Computing the error $e \in V$ satisfying (3.4) is a non–trivial task. Again, we employ Galerkin approximation to approximate it, which enables us to estimate the energy error $||e||_B$. Because of (3.5), we do not look for an approximation to e in X. Instead, we look for an approximation to e in an N_W–dimensional space $W \subset V$ that is richer than X. The quality of the resulting approximation is closely related to the quality of

the Galerkin approximation $u_W \in W$ to $u \in V$ satisfying

$$\text{find } u_W \in W : \qquad B(u_W, v) = F(v), \qquad \text{for all } v \in W. \tag{3.9}$$

By letting $e_W = u_W - u_X$ we see that

$$B(e_W, v) = B(u_W, v) - B(u_X, v) = F(v) - B(u_X, v), \qquad \text{for all } v \in W, \tag{3.10}$$

which is simply a restatement of (3.4) over W, and thus $e_W \in W$ satisfying (3.10) approximates the true error $e \in V$. Problem (3.10) leads to a tractable system of N_W equations that determines e_W, and spaces W that contain significantly improved approximations u_W to u (compared to u_X), also contain good approximations e_W to e. To analyse the quality of the energy error estimate $||e_W||_B \approx ||e||_B$, for a given choice of W, we require the following assumption.

Assumption 3.1: Saturation.

Let the functions u, u_X and u_W satisfy problems (3.1), (3.3) and (3.9) respectively. There exists a constant $\beta \in [0, 1)$ (the saturation constant) such that

$$||u - u_W||_B \leq \beta ||u - u_X||_B. \tag{3.11}$$

Note that (3.11) is a stronger property than (3.7), and always holds for $\beta \leq 1$. Whilst the constant β depends on the regularity of $u \in V$, which in turn depends on the problem at hand, in many applications Assumption 3.1 is reasonable [42, 14, 30]. The precise relationship between $||e||_B$ and $||e_W||_B$ is given in the next result.

Theorem 3.1: [2, Theorem 5.1].

Let Assumption 3.1 hold and let $e \in V$ and $e_W \in W$ satisfy (3.4) *and* (3.10) *respectively, then*

$$||e_W||_B \leq ||e||_B \leq \frac{1}{\sqrt{1 - \beta^2}} ||e_W||_B, \tag{3.12}$$

where $\beta \in [0, 1)$ is the saturation constant satisfying (3.11).

The interpretation of the bound (3.12) is as follows; $||e_W||_B$ will never overestimate the true energy error $||e||_B$, but it could underestimate it by a factor of $(1 - \beta^2)^{-1/2}$. Moreover, enlarging W such that $\beta \to 0$ ensures $|||e_W||_B \to \|e\|_B$, and thus $||e_W||_B$ can be arbitrarily accurate.

Due to the fact that N_W may be significantly larger than N_X, problem (3.10) may be too expensive to solve. We now look to approximate $e_W \in W$ by solving a cheaper problem. Suppose that the bilinear form $B_0 : V \times V \to \mathbb{R}$ is an inner product on V with induced norm $||\cdot||_{B_0} = B_0(\cdot,\cdot)^{\frac{1}{2}}$. Suppose also that the matrix representation of $||\cdot||_{B_0}$ on W is cheaper to assemble and more convenient to work with (e.g. sparser, or more structured in some way). We may then consider the alternative problem:

$$\text{find } e_0 \in W : \qquad B_0(e_0, v) = F(v) - B(u_X, v), \qquad \text{for all } v \in W. \tag{3.13}$$

Theorem 3.2: [2, Theorem 5.3].

Let $e_W \in W$ and $e_0 \in W$ satisfy (3.10) *and* (3.13) *and suppose that there exist $\lambda, \Lambda \in \mathbb{R}^+$ such that*

$$\lambda||w||_B^2 \leq ||w||_{B_0}^2 \leq \Lambda||w||_B^2, \qquad \textit{for all } w \in W, \tag{3.14}$$

(the norms are equivalent on W) then

$$\sqrt{\lambda}||e_0||_{B_0} \leq ||e_W||_B \leq \sqrt{\Lambda}||e_0||_{B_0}. \tag{3.15}$$

If $B_0(\cdot,\cdot)$ approximates $B(\cdot,\cdot)$ well, the constants λ and Λ are close to one. When choosing $B_0(\cdot,\cdot)$, we must strike a balance between the computational gain of solving (3.13) instead of (3.10), and the loss in accuracy between $||e_0||_{B_0}$ and $||e_W||_B$.

Problem (3.13) may still be too expensive to solve. By imposing a straightforward structure on W, we may approximate $e_0 \in W$ by solving a reduced problem. We insist that W is an augmented space, that is, we choose a space $Y \subset V$ of dimension N_Y satisfying $X \cap Y = \{0\}$, and construct

$$W = X \oplus Y. \tag{3.16}$$

We may than consider the lower–dimensional problem

$$\text{find } e_Y \in Y : \qquad B_0(e_Y, v) = F(v) - B(u_X, v), \qquad \text{for all } v \in Y. \tag{3.17}$$

Since X and Y are disjoint, and $B_0(\cdot,\cdot)$ induces a norm on the Hilbert space V, we deduce from Theorem 2.2 that there exists a constant $\gamma \in [0, 1)$ such that

$$|B_0(u, v)| \leq \gamma||u||_{B_0}||v||_{B_0}, \qquad \text{for all } u \in X, \qquad \text{for all } v \in Y, \tag{3.18}$$

which leads to the following relationship between $||e_Y||_{B_0}$ and $||e_0||_{B_0}$.

Theorem 3.3: [2, Theorem 5.2].

Let $e_0 \in W$ and $e_Y \in Y$ satisfy (3.13) *and* (3.17) *respectively, and suppose that* (3.16) *holds. Then*

$$||e_Y||_{B_0} \leq ||e_0||_{B_0} \leq \frac{1}{\sqrt{1-\gamma^2}}||e_Y||_{B_0}, \tag{3.19}$$

where $\gamma \in [0, 1)$ satisfies (3.18).

Theorem 3.3 tells us that the quality of the estimate $||e_Y||_{B_0}$ for $\|e_0\|_{B_0}$ depends entirely on the compatibility of the subspaces X and Y with respect to the inner product $B_0(\cdot, \cdot)$. For a fixed choice of X, choosing Y such that the associated CBS constant $\gamma_{\min}$ is small ensures that the bound (3.19) is tight. Indeed, if X and Y are mutually orthogonal with respect to the inner product $B_0(\cdot, \cdot)$, then $\gamma_{\min} = 0$ and $||e_Y||_{B_0} = ||e_0||_{B_0}$. For several choices of X and Y, methods to compute and estimate the associated CBS constant when $V = H_0^1(D)$ are discussed in Chapter 4.

Consolidating Theorems 3.1–3.3 yields the following final result[1].

Theorem 3.4.

Let $e \in V$ and $e_Y \in Y$ satisfy (3.4) *and* (3.17) *respectively, where* (3.16) *holds. If Assumption 3.1 holds, and there exist $\lambda, \Lambda \in \mathbb{R}^+$ such that* (3.14) *holds, then*

$$\sqrt{\lambda}||e_Y||_{B_0} \leq ||e||_B \leq \frac{\sqrt{\Lambda}}{\sqrt{1-\beta^2}\sqrt{1-\gamma^2}}||e_Y||_{B_0}, \tag{3.20}$$

where $\gamma \in [0, 1)$ satisfies (3.18) *and $\beta \in [0, 1)$ satisfies* (3.11).

In summary, the quality of the energy error estimate $||e_Y||_{B_0} \approx ||e||_B$ depends on the equivalency constants λ and Λ, which depend on our choice of bilinear form $B_0(\cdot, \cdot)$, β (the saturation constant) and γ (the CBS constant), both of which depend on our choice of Y. For fixed choices of $B(\cdot, \cdot)$ and X in the underlying discrete problem (3.3), choosing $B_0(\cdot, \cdot)$ and Y in (3.17) such that the constants $\sqrt{\lambda}$ and $\frac{\sqrt{\Lambda}}{\sqrt{1-\beta^2}\sqrt{1-\gamma^2}}$ are close to one ensures that the bound (3.20) is tight. For a computed estimate $||e_Y||_{B_0}$, it is then conventional to study its *effectivity index* which we define below.

[1]Note that Theorems 3.2 and 3.3 are presented in reverse order in [2]. Consequently, Theorem 5.2 in [2] is stated in terms of $\|\cdot\|_B$ rather than $\|\cdot\|_{B_0}$, and Theorem 5.3 in [2] is stated on the space Y rather than W. The proofs for both results are trivially amended to account for the new ordering presented herein.

Definition 3.2: Effectivity index.

The effectivity index of an a posteriori energy error estimate $\eta \approx ||e||_B$, where e satisfies (3.4), is the ratio

$$\theta_{\text{eff}} := \frac{\eta}{||e||_B} = \frac{\eta}{\sqrt{||u||_B^2 - ||u_X||_B^2}}. \tag{3.21}$$

If η is an accurate energy error estimate, θ_{eff} is close to one. If $\theta_{\text{eff}} < 1$, then η underestimates $||e||_B$. Of course, θ_{eff} cannot be computed exactly since it depends on the true solution $u \in V$. For a simple test problem, we demonstrate in Section 3.3 how θ_{eff} may be estimated to assess the quality of the a posteriori error estimator $e_Y \in Y$ described in this section.

Since W is finite-dimensional, for a given choice of $B_0(\cdot,\cdot)$ the constants λ and Λ can always be computed. From (3.14) we know that

$$\lambda \leq R(w) \leq \Lambda, \qquad \text{for all } w \in W,$$

where $R(w) := \frac{B_0(w,w)}{B(w,w)}$, and thus the tightest bound corresponds to $\lambda = \inf_{w\in W} R(w)$ and $\Lambda = \sup_{w\in W} R(w)$. For all $w \in W$ there exists a vector $\mathbf{w} \in \mathbb{R}^{N_W}$ such that

$$B_0(w,w) = \mathbf{w}^T B_0 \mathbf{w}, \qquad B(w,w) = \mathbf{w}^T B \mathbf{w},$$

where $B_0, B \in \mathbb{R}^{N_W \times N_W}$ are symmetric and positive definite (they induce a norm), and thus λ and Λ are the smallest and largest eigenvalues of the generalised eigenvalue problem $B_0\mathbf{w} = \theta B\mathbf{w}$. Unfortunately, the constant β cannot be computed or easily estimated. For certain problems, including the test problem considered in Section 3.3, we can verify Assumption 3.1 using a priori error bounds (which requires knowledge about the regularity of the solution).

3.3 The Deterministic Diffusion Problem

To provide a straightforward example of Galerkin approximation and the a posteriori error estimation just described, we return to the PDE problem (1.1), which we formally restate with specific boundary conditions: find $u(\mathbf{x}) : D \to \mathbb{R}$ such that

$$-\nabla \cdot (a(\mathbf{x})\nabla u(\mathbf{x})) = f(\mathbf{x}), \qquad \mathbf{x} \in D, \tag{3.22}$$

$$u(\mathbf{x}) = 0, \qquad \mathbf{x} \in \partial D, \tag{3.23}$$

where $D \subset \mathbb{R}^2$ is a bounded polygonal domain, and ∂D denotes the boundary of D. We make the following assumptions on $a(\mathbf{x})$ and $f(\mathbf{x})$.

Assumption 3.2.

There exist constants $a_{\min}, a_{\max} \in \mathbb{R}^+$ such that

$$0 < a_{\min} \leq a(\mathbf{x}) \leq a_{\max} < \infty, \qquad \text{a.e. in } D,$$

and thus $a(\mathbf{x}) \in L^\infty(D)$.

Assumption 3.3.

The function $f(\mathbf{x})$ is square integrable on D, that is, $f \in L^2(D)$.

Taking into account the boundary condition (3.23) it is well known that the weak formulation of (3.22)–(3.23) is: find $u \in V := H_0^1(D)$ such that

$$\int_D a(\mathbf{x})\nabla u(\mathbf{x}) \cdot \nabla v(\mathbf{x})\ d\mathbf{x} = \int_D f(\mathbf{x})v(\mathbf{x})\ d\mathbf{x}, \qquad \text{for all } v \in V, \tag{3.24}$$

where the functions $v \in V$ are called *test* functions and $\|\cdot\|_V = \|\cdot\|_{H_0^1(D)}$. Notice that this is an example of the abstract weak problem (3.1) with

$$B(u,v) = \int_D a(\mathbf{x})\nabla u(\mathbf{x}) \cdot \nabla v(\mathbf{x})\ d\mathbf{x}, \qquad F(v) = \int_D f(\mathbf{x})v(\mathbf{x})\ d\mathbf{x}. \tag{3.25}$$

Applying Lemma 2.2 and Theorem 2.1 in succession to $F(v)$ yields

$$|F(v)| \leq ||f||_{L^2(D)}||v||_{L^2(D)} \leq C_p||f||_{L^2(D)}||v||_V,$$

for all $v \in V$, and thus under Assumption 3.3 the linear functional $F(\cdot)$ is bounded over V. Applying Lemma 2.2 to $B(u,v)$ and noting that $B(\cdot,\cdot)$ is an inner–product on V yields

$$|B(u,v)| \leq a_{\max}\left|\int_D \nabla u(\mathbf{x}) \cdot \nabla v(\mathbf{x})\ d\mathbf{x}\right| \leq a_{\max}||u||_V||v||_V,$$

$$B(u,u) \geq a_{\min}\int_D |\nabla u(\mathbf{x})|^2\ d\mathbf{x} = a_{\min}||u||_V^2,$$

for all $u, v \in V$, and thus under Assumption 3.2 the bilinear form $B(\cdot,\cdot)$ is both bounded and coercive over V. By Lemma 2.1 it follows that there exists a unique $u \in V$ that satisfies (3.24). In the next section we approximate the true solution using finite element methods.

3.3.1 Finite Element Methods

Problem (3.24) is infinite–dimensional, and so to find a Galerkin approximation we choose a subspace $X \subset V$ and solve a finite–dimensional problem of the form (3.3). In this Section we use FEMs to construct the space X. The main idea is to place a mesh $\mathcal{T}_h$ of quadrilateral or triangular *elements* over D, and approximate $u \in V$ on each element using piecewise polynomial approximation. For now, let

$$X := \text{span}\Big\{\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots, \phi_{N_X}(\mathbf{x})\Big\},$$

and consider the finite–dimensional problem:

$$\text{find } u_X \in X: \qquad B(u_X, v) = F(v), \qquad \text{for all } v \in X, \tag{3.26}$$

where $B(\cdot,\cdot)$ and $F(\cdot)$ are given in (3.25). Now, posing

$$u_X(\mathbf{x}) = \sum_{i=1}^{N_X} u_i \phi_i(\mathbf{x}), \qquad u_i \in \mathbb{R},$$

and choosing the linearly independent test functions $v = \phi_j(\mathbf{x})$ in (3.26) yields

$$\sum_{i=1}^{N_x} u_i B(\phi_i(\mathbf{x}), \phi_j(\mathbf{x})) = F(\phi_j(\mathbf{x})), \qquad \text{for } j = 1, 2, \ldots, N_X,$$

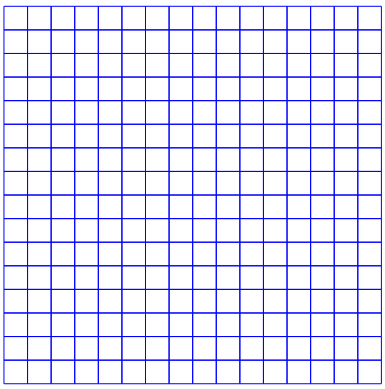
which is simply a linear system of equations $A\mathbf{u} = \mathbf{b}$ for the vector of coefficients

$$\mathbf{u} = [u_1, u_2, \ldots, u_{N_X}]^\top$$

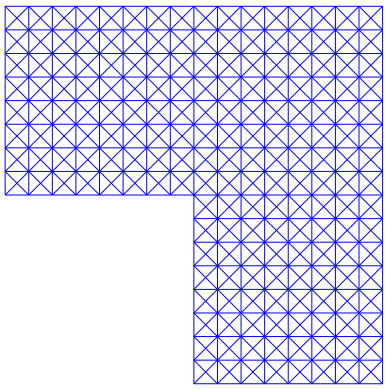
that defines $u_X \in X$. Since $\mathbf{u}^\top A\mathbf{u} = ||u_X||_B^2 > 0$ for all $u_X \neq 0$, the matrix A is positive definite. In addition, $A \in \mathbb{R}^{N_X \times N_X}$ and $\mathbf{b} \in \mathbb{R}^{N_X}$ have entries

$$[A]_{ji} = \int_D a(\mathbf{x}) \nabla\phi_i(\mathbf{x}) \cdot \nabla\phi_j(\mathbf{x})\, d\mathbf{x}, \qquad [\mathbf{b}]_j = \int_D f(\mathbf{x})\phi_j(\mathbf{x})\, d\mathbf{x},$$

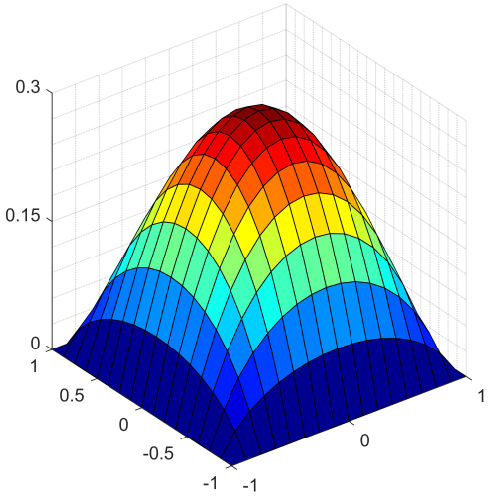
for $i, j = 1, 2, \ldots, N_X$, and thus A is symmetric.



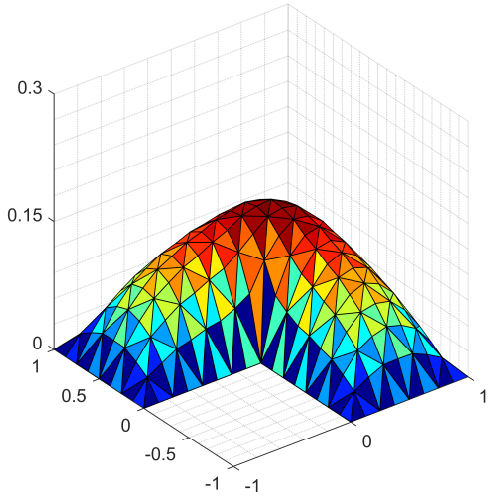
(a) mesh of square elements.



(b) mesh of triangular elements.



(c) $\mathbb{Q}_1$ FEM approximation.



(d) $\mathbb{P}_1$ FEM approximation.

Figure 3.1: FEM solutions to (3.26) using $\mathbb{Q}_1$ elements and $\mathbb{P}_1$ elements on the classical square and L–shape domains, respectively. The square mesh consists of 256 uniform elements and the triangular mesh consists of 768 uniform elements.

In this thesis, we construct X using nodal basis functions. That is, we insist that $\phi_j(\mathbf{x}_i) = \delta_{ij}$ where $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N_X}\}$ is a set of nodes placed at suitable positions on the mesh. When working with quadrilateral or triangular elements we often (if appropriate) consider spaces of continuous piecewise bilinear functions ($\mathbb{Q}_1$ elements), or continuous piecewise linear functions ($\mathbb{P}_1$ elements), respectively. Both choices are subspaces of $H^1(D)$, and, to ensure continuity, the nodes with respect to which the basis functions are defined are the vertices of $\mathcal{T}_h$. Since $X \subset H_0^1(D)$ we ignore nodes that lie on ∂D and thus N_X is the number of internal nodes. Higher order finite element spaces can be constructed which require additional nodes on each element, but for now we only consider $\mathbb{Q}_1$ and $\mathbb{P}_1$ spaces.

Example 3.1: $\mathbb{Q}_1$ and $\mathbb{P}_1$ elements.

Let $a(\mathbf{x}) = f(\mathbf{x}) = 1$ in (3.24). For our first example, let $D = [-1, 1]^2$ and let $\mathcal{T}_h$ denote a uniform square mesh over D consisting of 256 elements, as illustrated in Figure 3.1a. In Figure 3.1c we plot the corresponding $\mathbb{Q}_1$ FEM approximation $u_X \in X$ satisfying (3.26). Now, let $D = [-1, 1]^2 \backslash [-1, 0)^2$ and let $\mathcal{T}_h$ denote a uniform triangular mesh over D consisting of 768 elements (see Figure 3.1b). In Figure 3.1d we plot the corresponding $\mathbb{P}_1$ FEM approximation.

The finite element meshes presented in Figures 3.1a and 3.1b consist of relatively few elements. Since the number of elements ultimately determines the dimension N_X of X, it is reasonable to expect refined versions of these meshes (adding more elements) to lead to improved FEM approximations. In particular, since the solution in Figure 3.1d changes most rapidly at the reentrant corner, introducing additional elements in that vicinity should result in a markedly improved approximation. Whilst introducing new elements doesn't necessarily guarantee an improvement, from (3.7) we know that the approximation will not deteriorate with respect to the energy norm $\|\cdot\|_B$ (which, since $a = 1$, is equivalent to $\|\cdot\|_{H_0^1(D)}$). A key area of finite element analysis is understanding how the approximation error depends on the underlying mesh $\mathcal{T}_h$. Results of this kind require $\mathcal{T}_h$ to be admissible or *regular* in some sense. The shape regularity of rectangular meshes can be characterised in terms of the maximum edge ratios β_k (the size of the longest edge of $\square_k$ divided by that of the shortest) of each element $\square_k$ in the mesh $\mathcal{T}_h$. In a similar way, the regularity of triangular meshes can be characterised in terms of the minimum interior angles θ_k of each $\Delta_k \in \mathcal{T}_h$.

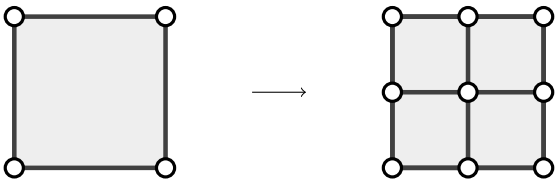
Definition 3.3: Shape regularity of rectangular meshes.

A sequence of rectangular meshes $\{\mathcal{T}_h\}$ is said to be *shape regular* if there exists an element $\square_* \in \mathcal{T}_h$ such that for every element in $\mathcal{T}_h$, $1 \leq \beta_k \leq \beta_*$.

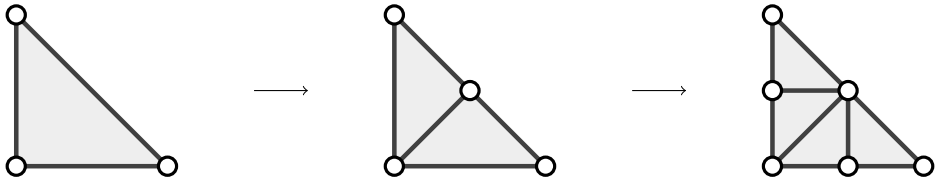
Definition 3.4: Shape regularity of triangular meshes.

A sequence of triangular meshes $\{\mathcal{T}_h\}$ is said to be *shape regular* if there exists an element $\Delta_* \in \mathcal{T}_h$ such that for every element in $\mathcal{T}_h$, $\theta_k \geq \theta_*$.

Choosing a shape regular mesh ensures that the area of each element in the mesh



(a) uniform refinement of square elements.



(b) uniform refinement of triangular elements.

Figure 3.2: Example uniform refinement of square and triangular elements. Note that for both types of element $h \to \frac{h}{2}$ (recall that h is the length of the longest edge), and in the case of trianguar elements, two iterations of longest edge bisection are employed.

is both nonzero and finite. In turn, this helps ensure that the associated FEM approximation is well defined, and also leads to the following well known convergence result (associated with (3.24) in two dimensions).

Theorem 3.5.

Let Assumptions 3.2 and 3.3 hold, and suppose that (3.26) *is solved using a* $\mathbb{Q}_m$ *or* $\mathbb{P}_m$ *FEM space* X *with* $m \geq 1$. *Provided that* $u \in H^{m+1}(D)$ *and the underlying mesh* $\mathcal{T}_h$ *is shape regular, then there exists a constant* C *such that*

$$\|u - u_X\|_B \leq Ch^m \|u\|_{H^{m+1}(D)}, \tag{3.27}$$

where h *denotes the length of the largest element edge in the mesh.*

Theorem 3.5 is a simple extension of Theorem 1.21 in [49] where $a = 1$ and $\|\cdot\|_B$ and $\|\cdot\|_{H_0^1(D)}$ coincide. When $a = a(\mathbf{x})$, it follows from (3.6) that

$$\|u - u_X\|_B^2 \leq \|u - v\|_B^2 \leq a_{\max} \|u - v\|_{H_0^1(D)}^2, \qquad \text{for all } v \in X,$$

and standard analysis for bounding $\|u - v\|_{H_0^1(D)}$ may be employed. Namely, we bound $\|u - \pi_h u\|_{H_0^1(D)}$ where $\pi_h u \in X$ denotes an appropriate interpolant of $u \in V$ that is amenable to analysis.

Theorem 3.5 states that if the true solution $u \in V$ has enough regularity (which depends on D, a, f), then $\|u - u_X\|_B \to 0$ as $h \to 0$ at an algebraic rate m that depends on our choice of X. In the case of $\mathbb{Q}_1$ or $\mathbb{P}_1$ finite element approximation, we

simply require that $u \in H^2(D)$ which results in linear convergence with respect to h, or equivalently,

$$\|u - u_X\|_B \leq \bar{C} N_X^{-1/2}, \qquad N_X \to \infty, \tag{3.28}$$

with $\bar{C} > 0$. Roughly speaking, this means that the approximation error should at least halve when we perform a *uniform refinement* of $\mathcal{T}_h$ and $h \to \frac{h}{2}$. By the term uniform refinement we mean that each element in $\mathcal{T}_h$ is divided into four smaller equally sized elements, as illustrated in Figure 3.2 (the resulting sequences $\{\mathcal{T}_h\}$ are clearly shape regular). In the case of triangular elements we define uniform refinement to coincide with performing two iterations of *longest edge bisection* [90] on each element – a strategy that forms two smaller elements by introducing a new edge that connects the midpoint of the longest edge with the opposing vertex.

Most interesting domains such as the L–shape in Figure 3.1d lead to solutions $u \in V$ with spatial singularities. These singularities mean that $u \notin H^2(D)$ and the bound (3.28) is not applicable. Due to the fact that an extremely fine mesh is required to eradicate the error near the singularity, which is then overkill on other areas of the domain, this simply means that uniform mesh refinements result in convergence at a rate worse than $-\frac{1}{2}$ with respect to N_X. Fortunately, the PDE (3.22)–(3.23) is well studied and for many situations where $u \notin H^2(D)$, it is possible to construct a sequence of FEM spaces $\{X\}$ of dimension N_X such that the bound (3.28) holds [10, 71, 72, 25]. In particular, when $\mathbb{P}_1$ FEM approximation is employed on geometries such as the L–shaped domain in Figures 3.1b and 3.1d, one can construct a sequence of meshes $\{\mathcal{T}_h\}$ such that (3.28) holds. In the next section we adaptively construct the sequence $\{\mathcal{T}_h\}$ by exploiting the a posteriori error estimator $e_Y \in Y$ described in Section 3.2.

3.3.2 Error Estimation & Adaptive Mesh Refinement

For a given $\mathbb{Q}_1$ or $\mathbb{P}_1$ approximation $u_X \in X$ satisfying (3.26) we follow the theoretical framework outlined in Section 3.2 to approximate $\|u - u_X\|_B$. The bilinear form $B(\cdot, \cdot)$ in (3.25) is simple enough to not need approximating, and thus we simply choose a FEM subspace $Y \subset H_0^1(D)$ such that $X \cap Y = \{0\}$ and solve

$$\text{find } e_Y \in Y: \qquad B(e_Y, v) = F(v) - B(u_X, v), \qquad \text{for all } v \in Y. \tag{3.29}$$

The resulting estimate $\|e_Y\|_B$ satisfies the bound

$$\|e_Y\|_B \leq \|u - u_X\|_B \leq \frac{1}{\sqrt{1-\beta^2}\sqrt{1-\gamma^2}}\|e_Y\|_B, \tag{3.30}$$

where $\beta \in [0,1)$ satisfies (3.11) for the space $W = X \oplus Y$ and $\gamma \in [0,1)$ satisfies

$$|B(u,v)| \leq \gamma \|u\|_B \|v\|_B, \qquad \text{for all } u \in X, \qquad \text{for all } v \in Y, \tag{3.31}$$

(recall Theorem 2.2).

The subspace Y may consist of local functions defined elementwise on the same mesh used to construct X, with support on only a single element (often referred to as *bubble* functions). We term the resulting estimator $e_Y \in Y$ a *local* estimator. Alternatively, Y may consist of global functions defined on D in the same way that X is constructed, and we term the resulting estimator a *global* estimator. Local estimators are often more complex to implement numerically than global estimators, whilst the cost to compute global estimators tends to grow more rapidly than for local estimators, as the mesh is refined. The best approach to take is problem dependent. We provide examples of both throughout this thesis and start in the next section with a local estimator for the deterministic diffusion problem.

The Element Residual Method

The error equation (3.29) may be decomposed elementwise over D as

$$\sum_{\square_k \in \mathcal{T}_h} B_k(e_Y, v) = \sum_{\square_k \in \mathcal{T}_h} \Big[F_k(v) - B_k(u_X, v) \Big], \tag{3.32}$$

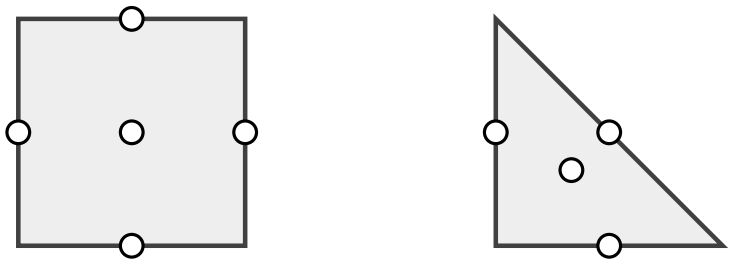
for all $v \in Y$ where

$$B_k(u,v) := \int_{\square_k} a(\mathbf{x}) \nabla u(\mathbf{x}) \cdot \nabla v(\mathbf{x}) \; d\mathbf{x}, \qquad F_k(v) := \int_{\square_k} f(\mathbf{x}) v(\mathbf{x}) \; d\mathbf{x} \tag{3.33}$$

and, for the sake of simplicity, $\square_k$ represents an element in a rectangular or triangular mesh. Assuming that $a(\mathbf{x}) \in C(D) \cap H^1(D)$, integration by parts on each element yields the residual equation

$$\sum_{\square_k \in \mathcal{T}_h} B_k(e_Y, v) = \sum_{\square_k \in \mathcal{T}_h} \Big[F_k(v) + \langle \nabla \cdot (a \nabla u_X), v \rangle_{L^2(\square_k)} - \sum_{E \in \mathcal{E}_k} \langle a \left[\!\left[\tfrac{\partial u_X}{\partial n} \right]\!\right], v \rangle_{L^2(E)} \Big],$$

for all $v \in Y$, where $\mathcal{E}_k$ is the set of edges of $\square_k$ excluding those on ∂D and



(a) $\mathbb{Q}_2$ error nodes. (b) $\mathbb{P}_2$ error nodes.

Figure 3.3: The nodes with respect to which the error estimator e_Y is computed on (a) a mesh of square elements, and (b) a mesh of triangular elements, when X is a $\mathbb{Q}_1$ and $\mathbb{P}_1$ FEM space, respectively.

$$\left[\!\left[\frac{\partial u_X}{\partial n}\right]\!\right] := \tfrac{1}{2}\big(\nabla u_X|_k - \nabla u_X|_{k'}\big)\cdot \vec{n}_{E,k}, \tag{3.34}$$

is the total flux equidistributed across the interelement edge E adjoining elements $\square_k$ and $\square_{k'}$. The above characterisation of e_Y on each element leads to cheap estimates of $e_Y|_{\square_k}$ through the solution of local subproblems. As demonstrated in [49], a good estimate to $e_Y|_{\square_k}$ can be obtained by enforcing the residual equation elementwise. That is, we compute a local estimator $e_{Y_k} \in Y_k$ for each $\square_k \in \mathcal{T}_h$ that satisfies

$$B_k(e_{Y_k}, v) = F_k(v) + \langle \nabla\cdot(a\nabla u_X), v\rangle_{L^2(\square_k)} - \sum_{E\in\mathcal{E}_k} \langle a\left[\!\left[\frac{\partial u_X}{\partial n}\right]\!\right], v\rangle_{L^2(E)}, \tag{3.35}$$

for all $v \in Y_k$. Each $Y_k \subset H_0^1(D)$ consists of functions with only compact support on $\square_k$ and must therefore be chosen carefully to ensure that $B_k : Y_k \times Y_k \to \mathbb{R}$ is coercive. We define the space $Y = \bigoplus_{\square_k\in\mathcal{T}_h} Y_k$ and construct the estimator $e_Y = \sum_{\square_k\in\mathcal{T}_h} e_{Y_k}$, for which

$$\|e_Y\|_B^2 = \sum_{\square_k\in\mathcal{T}_h} \|e_{Y_k}\|_{B_k}^2 := \sum_{\square_k\in\mathcal{T}_h} B_k(e_{Y_k}, e_{Y_k}). \tag{3.36}$$

For $\mathbb{Q}_1$ or $\mathbb{P}_1$ FEM spaces X, it is common to construct Y_k using the usual $\mathbb{Q}_2$ or $\mathbb{P}_2$ element basis functions defined with respect to the edge–midpoints and centroid of $\square_k$, as illustrated in Figures 3.3a and 3.3b. Technically speaking, the basis function associated with the centroid of the element in Figure 3.3b does not constitute a $\mathbb{P}_2$ function, rather, it is a *super–quadratic* (including the terms $x_1x_2^2$ and $x_1^2x_2$ for $(x_1, x_2)^\top \in D$) used to localise the error to $\square_k$ in the same way that its $\mathbb{Q}_2$ counterpart does. In addition, since $Y \subset H_0^1(D)$ we do not include basis functions with non–zero support on ∂D in the definition of Y_k.

Table 3.1: Estimated errors $\|e_Y\|_B$ for varying h when using the element residual method (for Example 3.2), as well as the associated approximate effectivity indices $\theta_{\text{eff}}^{\text{approx}}$.

Mesh	h	$\|e_Y\|_B$	$\|u_X\|_B^2$	$\theta_{\text{eff}}^{\text{approx}}$
8×8	2^{-2}	1.1339×10^{-1}	5.4934×10^{-1}	0.9957
16×16	2^{-3}	5.7068×10^{-2}	5.5904×10^{-1}	0.9980
32×32	2^{-4}	2.8590×10^{-2}	5.6149×10^{-1}	0.9984
64×64	2^{-5}	1.4303×10^{-2}	5.6210×10^{-1}	0.9870
128×128	2^{-6}	7.1528×10^{-3}	5.6226×10^{-1}	1.0116

Example 3.2: Rate of convergence.

Let a, f be as in Example 3.1, $D = [-1, 1]^2$ and $\mathcal{T}_h$ denote a uniform mesh of square elements over D. We compute the $\mathbb{Q}_1$ FEM approximation $u_X \in X$ to $u \in V$ satisfying (3.26) for varying h, as well as the corresponding energy error estimate $\|e_Y\|_B$ in (3.36) by solving (3.35) on each element with Y_k as described earlier (using $\mathbb{Q}_2$ element basis functions). To examine the quality of $\|e_Y\|_B$ we approximate θ_{eff} (recall Definition 3.2) by computing a reference solution $u_{\text{ref}} \in X_{\text{ref}}$ on a fine mesh of 1024×1024 elements and evaluating

$$\theta_{\text{eff}}^{\text{approx}} := \frac{\|e_Y\|_B}{\sqrt{\|u_{\text{ref}}\|_B^2 - \|u_X\|_B^2}}, \qquad \|u_{\text{ref}}\|_B^2 = 5.6231 \times 10^{-1}. \tag{3.37}$$

In Table 3.1 we observe that $\theta_{\text{eff}}^{\text{approx}}$ is close to one for each value of h, meaning that the energy error estimate is good and robust with respect to h. As predicted by the theoretical bound (3.27) for the chosen FEM space ($\mathbb{Q}_1$), we also observe the error roughly halves as $h \to \frac{h}{2}$.

For our choice of X and Y in Example 3.2 it is well known that the smallest γ satisfying (3.31) is bounded above by $\sqrt{\frac{5}{11}}$ [83, 37]. With the aim of designing error estimates with a tight associated bound (3.30), in the next chapter we compute and estimate $\gamma_{\min} \in [0, 1)$ for various other choices of X and Y.

Unfortunately, β in (3.30) cannot be easily estimated. We are interested in the smallest such constant given by $\beta_{\min} = \frac{\|u-u_W\|_B}{\|u-u_X\|_B}$. Both $u_X \in X$ and $u_W \in W$ can be computed by considering problems (3.3) and (3.9), respectively, but it goes without saying that we do not have our hands on the true solution $u \in V$. We mentioned previously that in some situations asymptotic results from a priori error analysis can

Table 3.2: Estimated errors $\|e_Y\|_B$ for varying h when using the global residual approach (for Example 3.3), as well as the associated approximate effectivity indices $\theta_{\text{eff}}^{\text{approx}}$.

# elements	$\|e_Y\|_B$	$\|u_X\|_B^2$	$\theta_{\text{eff}}^{\text{approx}}$
768	5.7852×10^{-2}	2.1028×10^{-1}	0.9388
3072	3.2274×10^{-2}	2.1288×10^{-1}	0.9324
12,288	1.8447×10^{-2}	2.1368×10^{-1}	0.9266
49,152	1.0800×10^{-2}	2.1394×10^{-1}	0.9240
196,608	6.4553×10^{-3}	2.1403×10^{-1}	0.9298

be utilised to help understand β. Suppose that X is a $\mathbb{Q}_1$ FEM space associated with a uniform mesh of square elements $\mathcal{T}_h$ with element width h (as in Example 3.2). Now, suppose that we augment X with the space of $\mathbb{Q}_2$ functions whose basis functions are defined with respect to the edge-midpoints and element centroids of $\mathcal{T}_h$. Label the resulting space W and note that W is then the usual $\mathbb{Q}_2$ FEM space associated with $\mathcal{T}_h$. If $u \in H^3(D)$, then, from (3.27) we know that for the simple diffusion problem (3.24) considered, $||u - u_X||_B = \mathcal{O}(h)$ and $||u - u_W||_B = \mathcal{O}(h^2)$. As a result, $\beta_{\min} = \mathcal{O}(h)$ and thus $\beta_{\min} \to 0$ as $h \to 0$. For even moderate mesh widths h ($h = 2^{-4} = 0.0625$ for example) we are confident that $\beta_{\min} \ll 1$.

A Global Residual Approach

Depending on the complexity of the PDE problem at hand, it is sometimes beneficial to solve the residual equation – (3.29) in this case – directly, without splitting it into many local problems. Whilst the complexity of solving the global problem may be suboptimal with respect to the number of elements in the mesh (in that it requires the solution of a large linear system of equations, the cost of which may not scale optimally), the implementation is much simpler than that of the element residual method.

Example 3.3: Rate of convergence, spatial singularity.

Let a, f be as in Example 3.1, $D = [-1,1]^2 \backslash [-1,0)^2$ and $\mathcal{T}_h$ denote a uniform mesh of triangular elements over D. Starting from the mesh given in Figure 3.1b, we compute the corresponding $\mathbb{P}_1$ FEM approximation $u_X \in X$ to $u \in V$ satisfying (3.26) for varying h (again, performing uniform refinements) as well as

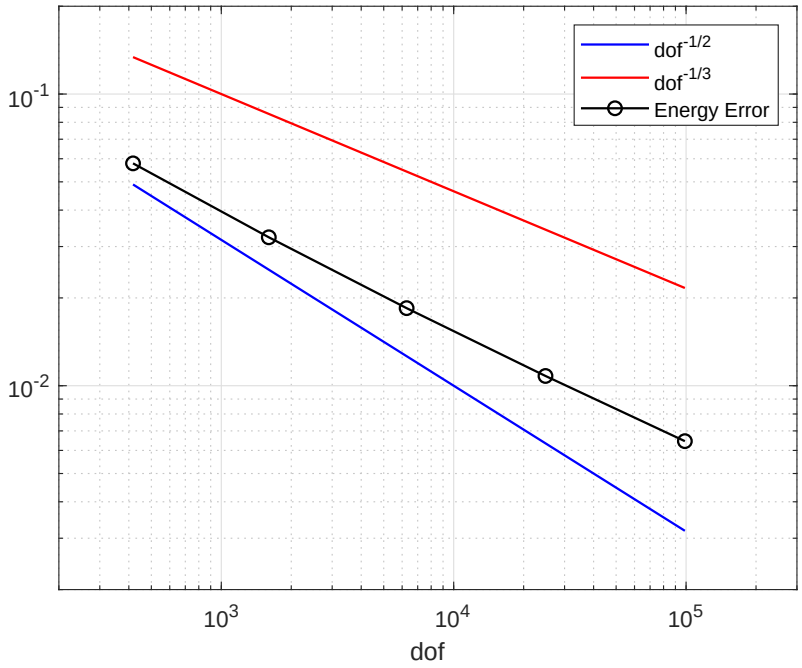


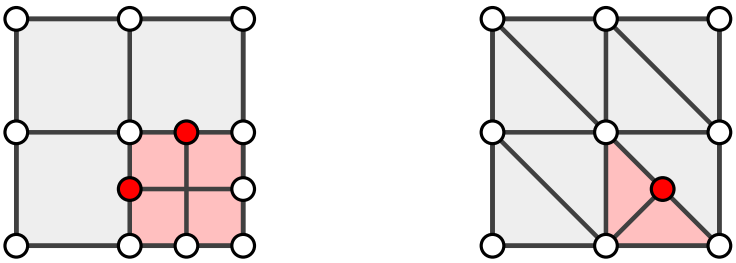
Figure 3.4: The estimated errors $\|e_Y\|_B$ given in Table 3.2 for Example 3.3 versus the corresponding number of degrees of freedom (dof) N_X.

the corresponding energy error estimate $\|e_Y\|_B$ by solving (3.29). We construct Y using the usual (global) $\mathbb{P}_2$ and super–quadratic basis functions, defined with respect to the element edge–midpoints and centroids of $\mathcal{T}_h$, respectively. We also compute $\theta_{\text{eff}}^{\text{approx}}$ using a reference solution $u_{\text{ref}} \in X_{\text{ref}}$ obtained using a fine mesh of $977,830$ elements, resulting in $\|u_{\text{ref}}\|_B^2 = 2.1407 \times 10^{-1}$. In Table 3.2 we find that the effectivity indices are close to one for each mesh, and in Figure 3.4 we plot $\|e_Y\|_B$ against N_X where, as theoretically asserted in [77], we observe that the error decays at the rate $-\frac{1}{3}$, rather than $-\frac{1}{2}$, with respect to N_X.

The cost to compute $u_X \in X$ satisfying some prescribed error tolerance, as well as the amount of computer memory required, is significantly higher when the error scales like $N_X^{-1/3}$ than when it scales like $N_X^{-1/2}$. Fortunately, for the problem considered in Example 3.3, the rate $-\frac{1}{2}$ can be recovered by employing an adaptive mesh refinement strategy where a sequence of non–uniform meshes is carefully constructed.

Adaptive Mesh Refinement

Meshes of triangular elements are considerably more flexible and amenable to adaptive refinement than meshes of square elements. This is due to the fact that *hanging nodes* –



(a) mesh of square elements. (b) mesh of triangular elements.

Figure 3.5: Example hanging nodes (red markers) following the refinement of a single (a) square element and (b) triangular element (the pink shaded regions).

nodes positioned on the edges of some elements, but at the vertices of others (resulting in a discontinuous approximation) – are much more easily resolved when triangles are employed. Hanging nodes typically occur when some elements in the mesh are refined and certain adjoining elements are not. In this thesis we refine triangular elements by performing a single iteration of longest edge bisection, and square elements by performing a uniform refinement, producing four smaller elements. The simple fix to remove hanging nodes is to refine a few more elements in the mesh, however, in the case of square meshes this can lead to large batches of elements being refined, and the benefits of adaptively constructing the mesh may be lost. In Figure 3.5 we illustrate the process of refining a single element (the pink shaded region) in an example square mesh and a single element in a triangular mesh, leading, in both cases, to the introduction of hanging nodes (represented by the red markers). Notice that resolving the hanging nodes in the square mesh would result in a complete uniform refinement of the original mesh, whereas only a single element refinement is required to resolve the hanging node in the triangular mesh. For this reason, in this thesis we only consider meshes of triangular elements when implementing an adaptive mesh refinement strategy.

Given a FEM approximation $u_X \in X$, we may sometimes exploit the corresponding a posteriori error estimator $e_Y \in Y$ to drive the adaptive refinement of the underlying mesh $\mathcal{T}_h$. Suppose that we have an error estimate $\|e_{Y_k}\|_{B_k}$ associated with each element $\square_k \in \mathcal{T}_h$. In the case of the global residual approach, provided that $B_k : Y_k \times Y_k \to \mathbb{R}$ with $Y_k := Y|_{\square_k}$ is coercive, each $\|e_{Y_k}\|_{B_k}$ is simply given by taking $\|e_{Y_k}\|_{B_k} := \sqrt{B_k(e_{Y_k}, e_{Y_k})}$ with $e_{Y_k} := e_Y|_{\square_k}$, and in the case of the element residual method we simply consider the energy norms $\|e_{Y_k}\|_{B_k}$ at hand (recall their definition in (3.36)).

The next step is to select, or *mark*, the elements that we wish to refine, based on the size of the corresponding estimates $\|e_{Y_k}\|_{B_k}$. A common approach is to employ a Dörfler [41] marking strategy where a minimal subset of marked elements $\mathcal{M} \subset \mathcal{T}_h$ satisfying

$$\sum_{\square_k \in \mathcal{M}} \|e_{Y_k}\|_{B_k}^2 \geq \theta_{\text{mark}} \sum_{\square_k \in \mathcal{T}_h} \|e_{Y_k}\|_{B_k}^2 \tag{3.38}$$

is determined, for the user–defined threshold parameter $\theta_{\text{mark}} \in (0, 1]$. The elements in $\mathcal{M}$ are then refined in the mesh $\mathcal{T}_h$, and, due to the type of refinement described, the resulting sequence is shape regular. Note that choosing $\theta_{\text{mark}} = 1$ is the same as marking all elements in the mesh, but is not equivalent to performing a uniform refinement of $\mathcal{T}_h$ (recall Figure 3.2 along with our definition of uniform refinement of triangles). Uniformly refining triangles in this way does not lead to the introduction of hanging nodes. In contrast, refining any number of elements in $\mathcal{T}_h$ once can lead to the introduction of hanging nodes, and so, as previously discussed, these are removed by refining a few more elements. Once the refined mesh has been generated, the well known [71] iterative process

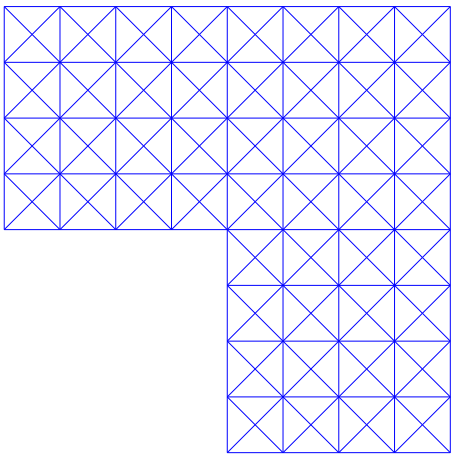
$$\texttt{SOLVE} \rightarrow \texttt{ESTIMATE} \rightarrow \texttt{MARK} \rightarrow \texttt{REFINE}, \tag{3.39}$$

is repeated until $\|e_Y\|_B < \texttt{tol}$, where $\texttt{tol}$ denotes a user–defined error tolerance.

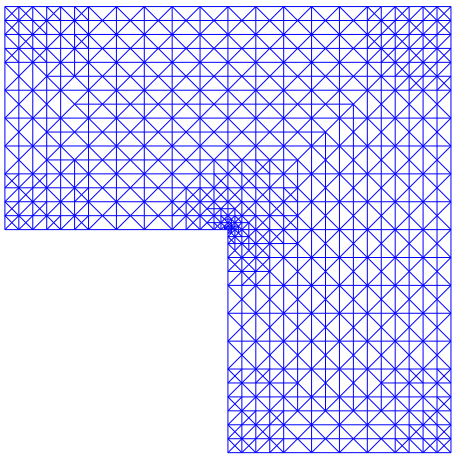
Example 3.4: Spatial adaptivity.

Let a, f be as in Example 3.1, and D and $X, Y \subset H_0^1(D)$ be as in Example 3.3. Starting from the initial uniform mesh given in Figure 3.6a, we compute $u_X \in X$ and $e_Y \in Y$ satisfying (3.26) and (3.29), respectively. Using (3.38) with $\theta_{\text{mark}} = \frac{1}{2}$ (this is a standard choice used in much of the literature, see [72, 108] for example) we then construct a minimal set of marked elements $\mathcal{M}$ and refine the mesh. We set $\texttt{tol} = 3.5 \times 10^{-2}$ and repeat (3.39) until $\|e_Y\|_B < \texttt{tol}$.

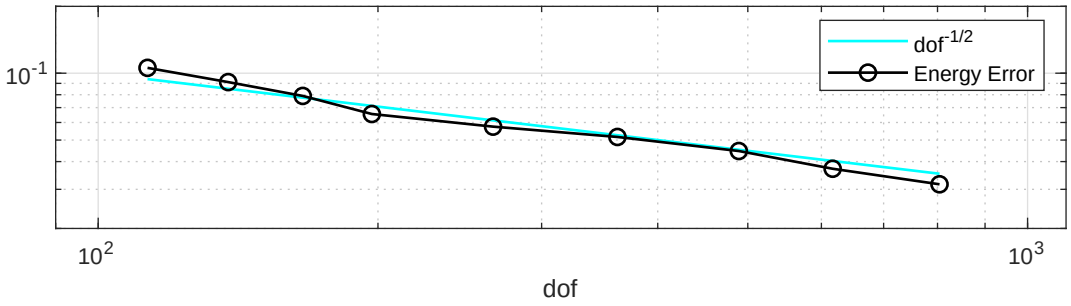
In Figure 3.6b we present the final mesh generated by the iterative process. Notice that many elements have been added at the corners of D, and most notably, at the reentrant corner, where the solution changes most rapidly. By plotting $\|e_Y\|_B$ against N_X at each step of the iterative procedure, we confirm in Figure 3.6c the assertion that the rate $-\frac{1}{2}$ can be realised for the L–shape domain. In



(a) initial mesh.



(b) adaptively constructed mesh.



(c) rate of convergence of $\|e_Y\|_B$.

Figure 3.6: Adaptive mesh construction (for Example 3.4) on the L–shape domain using a Dörfler marking strategy with $\theta_{\text{mark}} = \frac{1}{2}$ and $\texttt{tol} = 3.5 \times 10^{-2}$.

Figure 3.7 we plot the element error estimates $\|e_{Y_k}\|_{B_k}$ associated with the final mesh in Figure 3.6b. Observe that the error is most prominent at the reentrant corner of D.

The regularity of $u \in V$ satisfying (3.24) on the L–shape domain is directly related to the (270°) reflex angle at the reentrant corner [77]. In fact, the regularity of $u \in V$ diminishes as this angle increases. In Example 3.5 we investigate the crack domain

$$D = [-1,1]^2 \backslash \big\{(x_1,x_2)^\top \in \mathbb{R}^2;\ -1 < x_1 \leq 0,\ x_2 = 0\big\}, \tag{3.40}$$

for which the reflex angle is at its maximum.

Example 3.5: Spatial adaptivity, crack domain.

We rerun the experiment presented in Example 3.4 for D given in (3.40). We start from the initial uniform mesh given in Figure 3.8a and present the adaptively constructed mesh in Figure 3.8b for the same tolerance $\texttt{tol} = 3.5 \times 10^{-2}$. The

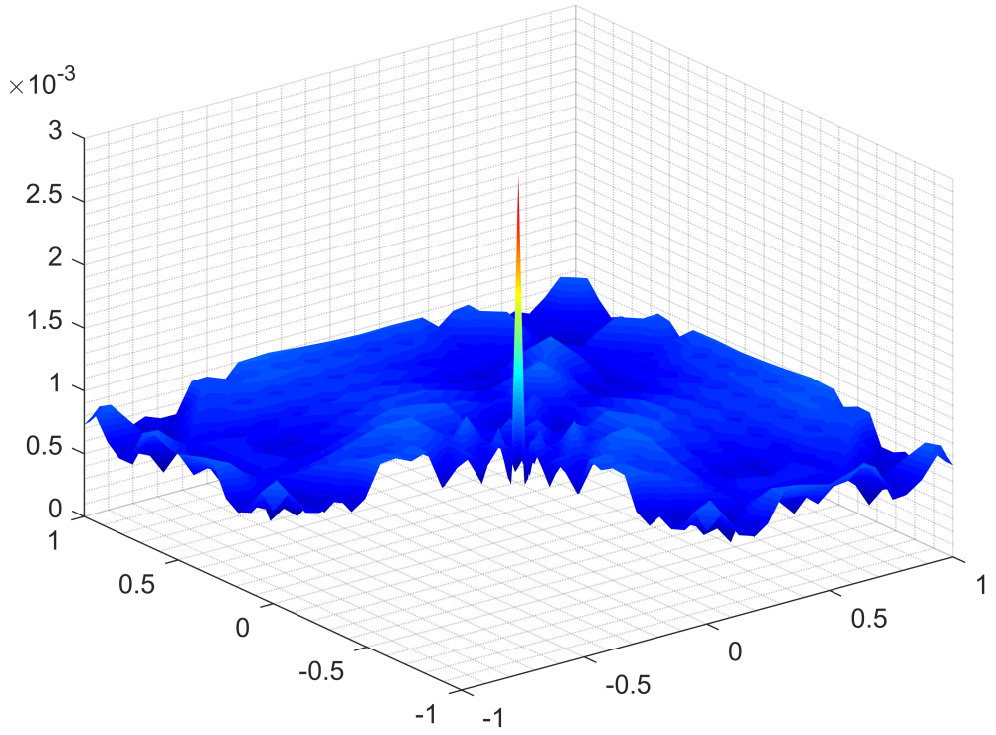


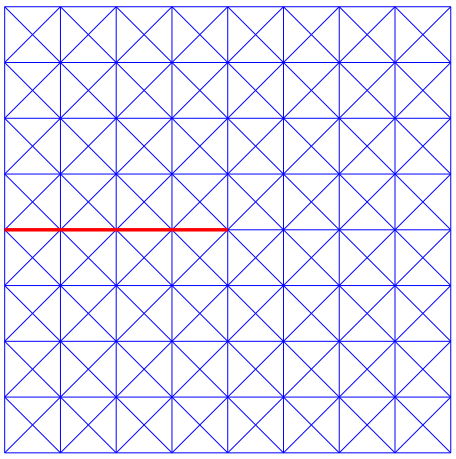
Figure 3.7: The element errors $\|e_{Y_k}\|_{B_k}$ associated with the final mesh in Figure 3.6b (for Example 3.4) when $\theta_{\text{mark}} = \frac{1}{2}$ and $\texttt{tol} = 3.5 \times 10^{-2}$.

red lines in these figures denote the "crack" in D. Using the marking parameter $\theta_{\text{mark}} = \frac{1}{2}$, we confirm in Figure 3.8c that the rate $-\frac{1}{2}$ can be realised for the tougher crack domain.

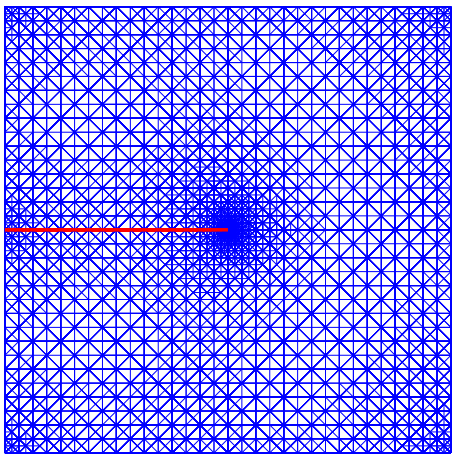
3.4 Summary

In this chapter, we introduced Galerkin approximation for a class of abstract infinite–dimensional weak problems of the form (3.1) over a Hilbert space V. For a computed Galerkin approximation $u_X \in X \subset V$ satisfying (3.3) where X is finite–dimensional, we outlined in Section 3.2 how the energy error $\|e\|_B = \|u - u_X\|_B$ can be efficiently estimated a posteriori by solving an auxiliary problem of the form (3.17) over a finite–dimensional subspace $Y \subset V$ satisfying $X \cap Y = \{0\}$. The resulting estimate $\|e_Y\|_{B_0}$ is related to the true error $\|e\|_B$ by the bound (3.20).

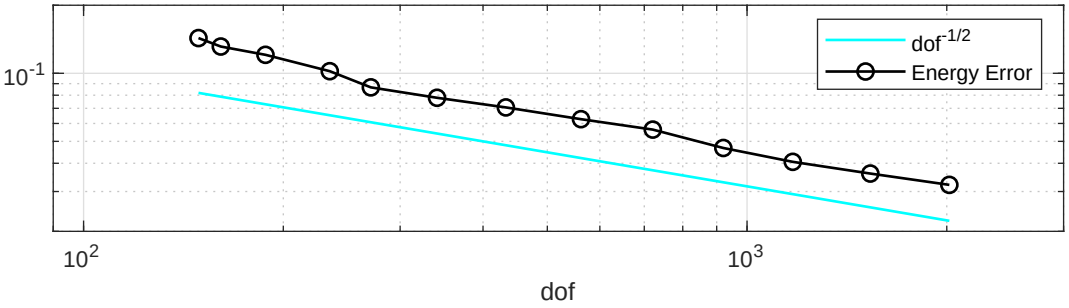
In Section 3.3, we demonstrated that the weak formulation of the diffusion problem (3.22)–(3.23) is an example of the abstract form (3.1) and discussed how Galerkin approximations $u_X \in X \subset V = H_0^1(D)$ can be generated using finite elements methods.



(a) initial mesh.



(b) adaptively constructed mesh.



(c) rate of convergence of $\|e_Y\|_B$.

Figure 3.8: Adaptive mesh construction (for Example 3.5) on the crack domain using a Dörfler marking strategy with $\theta_{\text{mark}} = \frac{1}{2}$ and $\texttt{tol} = 3.5 \times 10^{-2}$. The red line denotes the crack in D along the line $\{(x_1, x_2)^\top \in \mathbb{R}^2;\ -1 < x_1 \leq 0,\ x_2 = 0\}$.

Following the abstract framework for error estimation, in Section 3.3.2 we explained how to compute the estimate $\|e_Y\|_{B_0} \approx \|e\|_B$. We introduced the element residual method which required the solution of many local auxiliary problems as well as a global approach involving one large problem. The performance of both approaches for certain choices of $Y \subset H_0^1(D)$ was tested and in Tables 3.1 and 3.2 effectivity indices close to one were reported, meaning that the error estimates are highly accurate. Finally we demonstrated how a posteriori error estimators can be exploited to steer the adaptive refinement of finite element meshes. This is a powerful tool when the domain is non–convex and leads to spatially singular solutions. Indeed, for Examples 3.4 and 3.5, the error decays at the optimal rate $-\frac{1}{2}$ with respect to the number of DOFs N_X for the deterministic diffusion problem, using $\mathbb{P}_1$ approximation on the L–shape and crack domains.

The standard strategies outlined in this chapter will be exploited to design adaptive

SGFEMs in Chapters 5 and 6 for the more complex parametric diffusion problem. In the following chapter we investigate the constant $\gamma \in [0, 1)$ in the bound (3.30).

Chapter 4

The CBS constant

In this Chapter, we are concerned with the constant γ appearing in the error bound (3.30) associated with the problem (3.22)–(3.23). Determining the CBS constant $\gamma_{\min}$ (the smallest such constant) associated with this bound, for specified pairs of finite–dimensional subspaces $X, Y \subset H_0^1(D)$ where X is a standard FEM space ($\mathbb{Q}_1$ for example), enables us in Chapters 5 and 6 to design effective error estimators for the more complex stochastic diffusion problem discussed briefly in Chapter 1. Calculating the CBS constant associated with the splitting of a subspace $W = (X \oplus Y) \subset V$ into X and Y, where V is a Hilbert space and W is finite–dimensional, can sometimes be cast as an eigenvalue problem involving matrices. For various different pairings of subspaces $X, Y \subset H_0^1(D)$, we compute the associated CBS constants $\gamma_{\min}$ by solving the corresponding eigenvalue problem. For some of these pairings, the eigenvalue problem has a special structure which we exploit to establish novel analytical expressions for the associated CBS constants.

We begin our investigation in the next section by reformulating the strengthened Cauchy–Schwarz inequality (2.4) in terms of matrices that operate on disjoint vector spaces, and recall some standard results from [48]. Note that CBS constants associated with the splitting of vector spaces also arise in the analysis of certain preconditioners and iterative methods for linear systems of equations, see [3, 48, 86, 5, 4, 84], for example. Consequently, all results in this chapter extend beyond the field of a posteriori error analysis that we are interested in. The work presented in this chapter is based

on results from the published article [37].

4.1 Global CBS constants

First, we recall some standard results from [48]. Suppose that $M \in \mathbb{R}^{N\times N}$ is symmetric and positive definite with size $N := m + n$ for $m, n \in \mathbb{N}$. The space $\left(\mathbb{R}^N, (\cdot,\cdot)_M\right)$ is then a Hilbert space with respect to the inner product

$$(\mathbf{u}, \mathbf{v})_M = \mathbf{u}^\top M \mathbf{v}, \qquad \mathbf{u}, \mathbf{v} \in \mathbb{R}^N.$$

Next, consider the following finite–dimensional subspaces U and V of $\mathbb{R}^N$:

$$U := \left\{ \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{0} \end{pmatrix};\ \mathbf{u}_1 \in \mathbb{R}^m \right\}, \qquad V := \left\{ \begin{pmatrix} \mathbf{0} \\ \mathbf{v}_2 \end{pmatrix};\ \mathbf{v}_2 \in \mathbb{R}^n \right\}. \tag{4.1}$$

When M has a particular block structure, Theorem 2.2 coupled with the spaces (4.1) leads to the following well–known result (a discrete CBS inequality).

Corollary 4.1: [48, Corollary 1].

Let $M \in \mathbb{R}^{N\times N}$ be a symmetric and positive definite 2×2 block matrix with the structure

$$M = \begin{bmatrix} B & C^\top \\ C & A \end{bmatrix}, \tag{4.2}$$

where $N = m + n$, $B \in \mathbb{R}^{m\times m}$, $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{n\times m}$. There exists a constant $\gamma \in [0, 1)$ such that

$$\left(\mathbf{u}_1^\top C^\top \mathbf{v}_2\right)^2 \le \gamma^2 \left(\mathbf{u}_1^\top B \mathbf{u}_1\right) \left(\mathbf{v}_2^\top A \mathbf{v}_2\right), \tag{4.3}$$

for all $\mathbf{u}_1 \in \mathbb{R}^m$ and $\mathbf{v}_2 \in \mathbb{R}^n$, and thus the associated CBS constant satisfies

$$\gamma_{\min}^2 := \sup_{\mathbf{u}_1\in\mathbb{R}^m} \sup_{\mathbf{v}_2\in\mathbb{R}^n} \frac{\left(\mathbf{u}_1^\top C^\top \mathbf{v}_2\right)^2}{\left(\mathbf{u}_1^\top B \mathbf{u}_1\right)\left(\mathbf{v}_2^\top A \mathbf{v}_2\right)}.$$

Herein all supremums and infinums are assumed to exclude the zero vector or function. Note that for any matrix M which has the properties stated in Corollary 4.1, the blocks $B \in \mathbb{R}^{m\times m}$, $A \in \mathbb{R}^{n\times n}$ and the matrix $CB^{-1}C^\top \in \mathbb{R}^{n\times n}$ are all symmetric and positive definite. We also have the following known result.

Theorem 4.1: [48, Theorem 2].

Let M be as in Corollary 4.1. Any constant γ that satisfies (4.3) *also satisfies*

$$\gamma^2 \mathbf{v}_2^\top A \mathbf{v}_2 \geq \mathbf{v}_2^\top C B^{-1} C^\top \mathbf{v}_2, \qquad \textit{for all } \mathbf{v}_2 \in \mathbb{R}^n.$$

From Theorem 4.1 we learn that

$$\gamma_{\min}^2 = \sup_{\mathbf{v}_2 \in \mathbb{R}^n} \frac{\mathbf{v}_2^\top C B^{-1} C^\top \mathbf{v}_2}{\mathbf{v}_2^\top A \mathbf{v}_2}, \tag{4.4}$$

and thus the square of the CBS constant is the largest eigenvalue $\theta_{\max}$ of the generalised eigenvalue problem

$$C B^{-1} C^\top \mathbf{v}_2 = \theta A \mathbf{v}_2, \tag{4.5}$$

which is well defined and can be easily solved numerically (using the MATLAB routine `eigs`, for example).

Using the above results, we now demonstrate how to numerically compute the CBS constant associated with the strengthened Cauchy–Schwarz inequality

$$|\langle u, v \rangle_a| \leq \gamma \langle u, u \rangle_a^{\frac{1}{2}} \langle v, v \rangle_a^{\frac{1}{2}}, \qquad \text{for all } u \in Y, \qquad \text{for all } v \in X, \tag{4.6}$$

where

$$\langle u, v \rangle_a := \int_D a(\mathbf{x}) \nabla u(\mathbf{x}) \cdot \nabla v(\mathbf{x}) \, d\mathbf{x}, \tag{4.7}$$

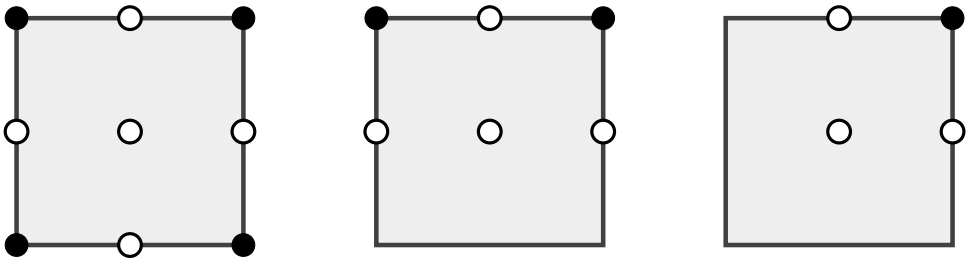
and $a(\mathbf{x})$ satisfies Assumption 3.2. We consider various FEM spaces $X, Y \subset H_0^1(D)$ satisfying $X \cap Y = \{0\}$. Notice that the inner–product in (4.7) coincides with the bilinear form $B(\cdot, \cdot)$ defined in (3.25), and thus (4.6) and (3.31) (associated with (3.25)) are equivalent. Accordingly, when X is the underlying FEM space associated with (3.26), and Y is the FEM space chosen for the error estimation problem (3.29), the CBS constant associated with (4.6) appears in the error bound (3.30).

Suppose we choose X and Y of the form

$$X = \operatorname{span}\big\{\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots, \phi_n(\mathbf{x})\big\}, \quad Y = \operatorname{span}\big\{\psi_1(\mathbf{x}), \psi_2(\mathbf{x}), \ldots, \psi_m(\mathbf{x})\big\}, \tag{4.8}$$

with $X \cap Y = \{0\}$ and define the augmented subspace

$$W := Y \oplus X = \operatorname{span}\big\{\Phi_1(\mathbf{x}), \Phi_2(\mathbf{x}), \ldots, \Phi_N(\mathbf{x})\big\},$$



(a) (b) (c)

Figure 4.1: Internal (a), edge (b), and corner (c) $\mathbb{Q}_1$ elements for Example 4.1. The black and clear markers are the nodes at which the basis functions of X and Y are defined, respectively.

of dimension $N = m + n$, where $\Phi_i = \psi_i$ for $i = 1, 2, \ldots, m$ and $\Phi_{m+j} = \phi_j$ for $j = 1, 2, \ldots, n$. For any functions $u \in Y$ and $v \in X$, there exist vectors $\mathbf{u} \in U$ and $\mathbf{v} \in V$ such that $\langle u, v \rangle_a = \mathbf{u}^\top M \mathbf{v}$, where the matrix $M \in \mathbb{R}^{N \times N}$ has the structure (4.2) with block–wise entries

$$\begin{aligned} [A]_{ij} &= \langle \phi_i, \phi_j \rangle_a, && i, j = 1, \ldots, n, \\ [B]_{ij} &= \langle \psi_i, \psi_j \rangle_a, && i, j = 1, \ldots, m, \\ [C]_{ij} &= \langle \phi_i, \psi_j \rangle_a, && i = 1, \ldots, n, \quad j = 1, \ldots, m. \end{aligned} \tag{4.9}$$

Clearly, M is also symmetric and positive definite and thus, by Corollary 4.1 there exists a constant $\gamma \in [0, 1)$ such that (4.3) holds, which is equivalent to (4.6). The CBS constant associated with (4.6) for the spaces X and Y defined in (4.8) then satisfies (4.4) for the blocks A, B and C defined in (4.9), and can thus be determined by solving the eigenvalue problem (4.5).

In the examples below, we fix a FEM space X associated with a uniform mesh $\mathcal{T}_h$ on D, and then construct Y elementwise by insisting that it admits the decomposition

$$Y = \bigoplus_{\square_k \in \mathcal{T}_h} Y_k, \qquad Y_k := \operatorname{span}\left\{ \psi_1^k(\mathbf{x}), \psi_2^k(\mathbf{x}), \ldots, \psi_{m_k}^k(\mathbf{x}) \right\} \subset H_0^1(D), \tag{4.10}$$

where $\square_k$ denotes an element in $\mathcal{T}_h$. We choose the functions $\{\psi_i^k\}_{i=1}^{m_k}$ to be bubble functions with only compact support on $\square_k$, which is the same construction of Y used in Section 3.3.2 in the implementation of the element residual method for a posteriori error estimation. Recall that the dimension $\dim(Y_k)$ of the resulting local problems (3.35) is m_k. In Chapter 5, we deal with error estimation problems for the parametric diffusion problem of dimension $m_k \times \dim(P)$, where P denotes a space

Table 4.1: Computed values of $\gamma^2_{\min}$ for Example 4.1, for varying h. The space X is the usual $\mathbb{Q}_1$ FEM space and four choices of Y are considered.

Mesh	h	N	$\mathbb{Q}_1(h/2)$	$\mathbb{Q}_2(h)$	$\mathbb{Q}_2(h/2)$	$\mathbb{Q}_4(h)$
4×4	2^{-1}	73	0.3381	0.4106	0.0401	0.0109
8×8	2^{-2}	337	0.3673	0.4454	0.0437	0.0119
16×16	2^{-3}	1441	0.3735	0.4527	0.0445	0.0121
32×32	2^{-4}	5953	0.3747	0.4541	0.0446	0.0121
64×64	2^{-5}	24193	0.3749	0.4544	0.0446	0.0121
Converged value			0.3750	0.4545	0.0446	0.0121

of polynomials associated with the parametric component of the problem. To ensure that the computational cost of solving those problems remains reasonable, we limit the size of m_k. For fixed choices of X, in Examples 4.1 and 4.2 below we investigate the CBS constant $\gamma_{\min}$ associated with (4.6) for several different choices of Y of the form (4.10) of a fixed small dimension. Our aim is to determine which spaces yield the smallest CBS constants, so that the term $\sqrt{1-\gamma^2_{\min}}$ appearing in the bound (3.30) associated with (3.22)–(3.23) is close to one. Fixing m_k in this way leads to some interesting and non–standard choices of Y. In the following examples, which were originally considered in [37], we construct the corresponding matrix M and compute the CBS constant by solving the eigenvalue problem (4.5).

Example 4.1: $X = \mathbb{Q}_1(h)$ (bilinear elements).

Let $a = 1$, $D = [-1, 1]^2$ and $\mathcal{T}_h$ denote a uniform mesh of square elements with element width h. We choose X to be the usual $\mathbb{Q}_1$ FEM space associated with $\mathcal{T}_h$ and write $X = \mathbb{Q}_1(h)$ to stress the dependence of X on h. On each $\square_k$ we then construct a space Y_k of dimension $m_k \leq 5$, and construct Y as in (4.10). That is, we limit m_k to be the maximum number of new $\mathbb{Q}_1$ nodes that would appear on each element in $\mathcal{T}_h$, if we were to perform a uniform mesh refinement. Hence, the basis functions of Y_k are constructed with respect to the edge midpoints and centroid of the element $\square_k$. In Figure 4.1 we illustrate an arbitrary internal, edge and corner $\mathbb{Q}_1$ element $\square_k \in \mathcal{T}_h$. Note the exclusion of nodes on the edge and corner elements to ensure that $Y \subset H^1_0(D)$. We start with two standard choices of Y:

1. **Piecewise bilinear bubbles**: $\mathbb{Q}_1(h/2)$. We uniformly refine each element

$\square_k \in \mathcal{T}_h$ into four smaller sub–elements. Then, for each Y_k, we construct a piecewise bilinear basis by assembling together the standard $\mathbb{Q}_1$ element basis functions defined on the four new sub–elements (there are sixteen functions; four per sub–element), with shared support at the clear markers in Figure 4.1. The functions assembled at the element centroids of $\mathcal{T}_h$ consist of four sub–element contributions, whereas the functions assembled at the midpoints consist of only two.

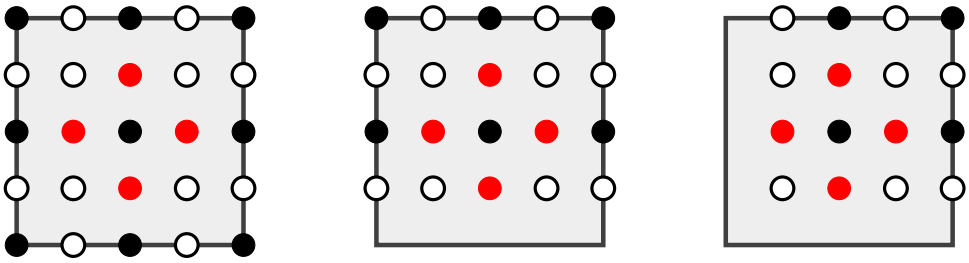
2. **Biquadratic bubbles**: $\mathbb{Q}_2(h)$. Consider the standard set of nine $\mathbb{Q}_2$ element basis functions defined on each element in $\mathcal{T}_h$, and keep only those five associated with the clear markers in Figure 4.1.

In columns four and five of Table 4.1 we record $\gamma^2_{\min}$ for the two choices of Y described above for varying h. These spaces are standard and the bounds $\gamma^2_{\min} \leq \frac{3}{8}$ and $\gamma^2_{\min} \leq \frac{5}{11}$ have been reported previously in [84] and [83], respectively. Our third and fourth choices of Y below are non–standard, however, and to the best of our knowledge CBS constants for these spaces have not been computed previously in the literature.

3. **Piecewise biquadratic bubbles**: $\mathbb{Q}_2(h/2)$. We uniformly refine each element $\square_k \in \mathcal{T}_h$. Then, as for option 1, we assemble collections of the thirty–six $\mathbb{Q}_2$ element basis functions defined across the new sub–elements (nine per sub–element) with shared support at the clear markers in Figure 4.1. Again, functions assembled at the element centroids of $\mathcal{T}_h$ consist of four contributions and functions assembled at the midpoints consist of two.

4. **Biquartic bubbles**: $\mathbb{Q}_4(h)$. Consider the standard set of twenty–five $\mathbb{Q}_4$ element basis functions defined on each element in $\mathcal{T}_h$, and keep only those five associated with the clear markers in Figure 4.1.

We record $\gamma^2_{\min}$ for these two (non–standard) choices of Y in the last two columns of Table 4.1 for varying h. Of the four choices considered, $Y = \mathbb{Q}_4(h)$ yields the smallest CBS constant and thus the term $\sqrt{1-\gamma^2_{\min}}$ appearing in the bound (3.30), for this space, is the closest to one.



(a) (b) (c)

Figure 4.2: Internal (a), edge (b), and corner (c) $\mathbb{Q}_2$ elements for Example 4.2. The black and clear/red markers are the nodes at which the basis functions of X and Y are defined, respectively.

Example 4.2: $X = \mathbb{Q}_2(h)$ (biquadratic elements).

Let a, D and $\mathcal{T}_h$ be as in Example 4.1 and let X be the usual $\mathbb{Q}_2$ space associated with $\mathcal{T}_h$. On each $\square_k$ we now construct a space Y_k of dimension $m_k \leq 16$ with basis functions defined at the additional $\mathbb{Q}_2$ nodes that would be introduced by performing a uniform refinement of $\square_k$. In Figure 4.2 we illustrate an arbitrary internal, edge and corner $\mathbb{Q}_2$ element $\square_k \in \mathcal{T}_h$. We now make the following four choices of Y (the first two of which are standard):

1. **Piecewise biquadratic bubbles**: $\mathbb{Q}_2(h/2)$. We uniformly refine each element $\square_k \in \mathcal{T}_h$. Then, for each Y_k, we construct a piecewise biquadratic basis by assembling together the standard $\mathbb{Q}_2$ element basis functions defined on the new sub–elements (of which there are thirty–six), with shared support at the clear and red markers in Figure 4.2.

2. **Biquartic bubbles**: $\mathbb{Q}_4(h)$. Consider the standard set of twenty–five $\mathbb{Q}_4$ element basis functions on each element in $\mathcal{T}_h$, and keep only those defined at the sixteen clear and red markers in Figure 4.2.

3–4. **Reduced spaces**: $\mathbb{Q}_2^r(h/2)$ and $\mathbb{Q}_4^r(h)$. For our third and fourth choices we consider some more non–standard FEM spaces. We modify the first two choices by removing the basis functions defined at the red markers in Figure 4.2, and denote the resulting number of degrees of freedom by N_r. This configuration is motived by error estimation results presented in [68], where it is demonstrated that the space $Y = \mathbb{Q}_4^r(h)$ defines a more accurate error

Table 4.2: Computed values of $\gamma^2_{\min}$ for Example 4.2, for varying h. The space X is the usual $\mathbb{Q}_2$ finite element space and four choices of Y are considered.

Mesh	h	N	$\mathbb{Q}_2(h/2)$	$\mathbb{Q}_4(h)$	N_r	$\mathbb{Q}^r_2(h/2)$	$\mathbb{Q}^r_4(h)$
2×2	2^{-0}	57	0.6764	0.3834	41	0.4904	0.3208
4×4	2^{-1}	273	0.6911	0.4341	209	0.5579	0.3565
8×8	2^{-2}	1185	0.6911	0.4391	929	0.5723	0.3595
16×16	2^{-3}	4929	0.6911	0.4399	3905	0.5758	0.3599
32×32	2^{-4}	20097	0.6911	0.4401	16001	0.5766	0.3600
Converged value			0.6911	0.4401		0.5769	0.3600

estimator $\|e_Y\|_B$ satisfying (3.30) than the richer space $Y = \mathbb{Q}_4(h)$. Note however that the associated CBS constant was not investigated in [68].

In Table 4.2 we record $\gamma^2_{\min}$ for the four choices of Y described above for varying h, where we observe that the space $Y = \mathbb{Q}^r_4(h)$ yields the smallest CBS constant. Whilst the spaces $Y = \mathbb{Q}_4(h)$ and $Y = \mathbb{Q}^r_4(h)$ have been employed previously in the construction of certain a posteriori error estimators in [68], the constants $\gamma^2_{\min}$ provided in Table 4.2 offer a new insight, and, to the best of our knowledge, have not been investigated previously.

Since the space $Y = \mathbb{Q}_4(h)$ is richer than $Y = \mathbb{Q}^r_4(h)$, the saturation constant $\beta_{\min}$ associated with the former is bounded above by that of the latter. In order for $Y = \mathbb{Q}^r_4(h)$ to define a more accurate estimator than $Y = \mathbb{Q}_4(h)$, in terms of (3.30) and as demonstrated empirically in [68], the CBS constant associated with $Y = \mathbb{Q}^r_4(h)$ must be smaller than the constant associated with $Y = \mathbb{Q}_4(h)$, which we confirm in Table 4.2.

The finite element meshes $\mathcal{T}_h$ used to compute the CBS constants presented in Tables 4.1 and 4.2 are relatively coarse, and N could in fact be much larger for the problem (3.26) (associated with (3.22)–(3.23)) of interest. For large N, the eigenvalue problem (4.5) becomes expensive to solve. To circumnavigate this issue, we show in the next section that tight upper bounds for $\gamma^2_{\min}$ can be cheaply computed using only local analysis on a single element $\square_k \in \mathcal{T}_h$. Using standard results (see [48]), we derive an eigenvalue problem for which the size of the associated matrices depends only on the total number of nodes defined on a single element (recall Figures 4.1 and 4.2).

4.2 Local Estimates of CBS constants

Element stiffness matrices tend to be only symmetric and positive semi–definite, and consequently Corollary 4.1 cannot be applied at a local level. Instead, we work with the vector spaces

$$U_k := \left\{ \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{0} \end{pmatrix};\ \mathbf{u}_1 \in \mathbb{R}^{m_k} \right\}, \qquad V_k := \left\{ \begin{pmatrix} \mathbf{0} \\ \mathbf{v}_2 \end{pmatrix};\ \mathbf{v}_2 \in \mathbb{R}^{n_k} \right\}. \tag{4.11}$$

where $U_k, V_k \subset \mathbb{R}^{N_k}$ for $N_k = m_k + n_k$, and consider the following theorem.

Theorem 4.2: [48, Theorem 3].

Let $M_k \in \mathbb{R}^{N_k \times N_k}$ be a symmetric and positive semi–definite 2×2 block matrix given by

$$M_k = \begin{bmatrix} B_k & C_k^\top \\ C_k & A_k \end{bmatrix}, \tag{4.12}$$

where $B_k \in \mathbb{R}^{m_k \times m_k}$ is invertible, $A_k \in \mathbb{R}^{n_k \times n_k}$, $C_k \in \mathbb{R}^{n_k \times m_k}$ and $\mathcal{N}(M_k) \subset V_k$. Then, there exists a constant $\gamma_k \in [0, 1)$ such that

$$\left(\mathbf{u}_1^\top C_k^\top \mathbf{v}_2\right)^2 \leq \gamma_k^2 \left(\mathbf{u}_1^\top B_k \mathbf{u}_1\right) \left(\mathbf{v}_2^\top A_k \mathbf{v}_2\right), \tag{4.13}$$

for all $\mathbf{u}_1 \in \mathbb{R}^{m_k}$ and $\mathbf{v}_2 \in \mathbb{R}^{n_k}$.

Theorem 4.3: [48, Theorem 4].

Let the vector spaces U_k, V_k and the matrix M_k be given by (4.11) and (4.12), and assume $\mathcal{N}(M_k) \subset V_k$. Any constant γ_k satisfying (4.13) then also satisfies

$$\gamma_k^2 \mathbf{v}_2^\top A_k \mathbf{v}_2 \geq \mathbf{v}_2^\top C_k B_k^{-1} C_k^\top \mathbf{v}_2, \qquad \text{for all } \mathbf{v}_2 \in \mathbb{R}^{n_k}.$$

For matrices M_k that satisfy the stated conditions in Theorem 4.2, we learn from Theorem 4.3 that

$$\gamma_{k,\min}^2 = \sup_{\mathbf{v}_2 \notin \mathcal{N}(A_k)} \frac{\mathbf{v}_2^\top C_k B_k^{-1} C_k^\top \mathbf{v}_2}{\mathbf{v}_2^\top A_k \mathbf{v}_2}, \tag{4.14}$$

and thus $\gamma_{k,\min}^2$ is the largest eigenvalue $\theta_{\max}$ satisfying the eigenvalue problem

$$C_k B_k^{-1} C_k^\top \mathbf{v}_2 = \theta A_k \mathbf{v}_2, \qquad \mathbf{v}_2 \notin \mathcal{N}(A_k). \tag{4.15}$$

To solve (4.15) numerically, deflation (removing certain rows and columns) is needed to deal with the null spaces of $C_k B_k^{-1} C_k^\top$ and A_k. Having said that, computing $\gamma^2_{k,\min}$ remains straightforward.

Given two subspaces $X, Y \subset H_0^1(D)$, the inner–product $\langle u, v\rangle_a$ of $u \in Y$ with $v \in X$ admits the elementwise decomposition

$$\langle u, v\rangle_a = \sum_{\square_k \in \mathcal{T}_h} \int_{\square_k} a_k \nabla u_k \cdot \nabla v_k \, d\mathbf{x} =: \sum_{\square_k \in \mathcal{T}_h} \langle u_k, v_k\rangle_{a_k}, \tag{4.16}$$

where $a_k = a|_{\square_k}$ and

$$u_k := u|_{\square_k} \in Y_k := Y|_{\square_k}, \qquad v_k := v|_{\square_k} \in X_k := X|_{\square_k},$$

with $Y_k, X_k \subset H^1(\square_k)$. Now, consider spaces of the form

$$X_k = \operatorname{span}\big\{\phi_1^k(\mathbf{x}), \phi_2^k(\mathbf{x}), \ldots, \phi_{n_k}^k(\mathbf{x})\big\}, \quad Y_k = \operatorname{span}\big\{\psi_1^k(\mathbf{x}), \psi_2^k(\mathbf{x}), \ldots, \psi_{m_k}^k(\mathbf{x})\big\},$$

with $X_k \cap Y_k = \{0\}$ and define the augmented subspace

$$W_k := Y_k \oplus X_k = \operatorname{span}\big\{\Phi_1^k(\mathbf{x}), \Phi_2^k(\mathbf{x}), \ldots, \Phi_{N_k}^k(\mathbf{x})\big\},$$

of dimension $N_k = m_k + n_k$, where $\Phi_i^k = \psi_i^k$ for $i = 1, 2, \ldots, m_k$ and $\Phi_{m_k+j}^k = \phi_j^k$ for $j = 1, 2, \ldots, n_k$. For any $u_k \in Y_k$ and $v_k \in X_k$, there exist vectors $\mathbf{u}_k \in U_k$ and $\mathbf{v}_k \in V_k$ such that $\langle u_k, v_k\rangle_{a_k} = \mathbf{u}_k^\top M_k \mathbf{v}_k$, where the matrix $M_k \in \mathbb{R}^{N_k \times N_k}$ has the structure (4.12) with

$$\begin{aligned} [A_k]_{ij} &= \langle \phi_i^k, \phi_j^k\rangle_{a_k}, && i, j = 1, \ldots, n_k, \\ [B_k]_{ij} &= \langle \psi_i^k, \psi_j^k\rangle_{a_k}, && i, j = 1, \ldots, m_k, \\ [C_k]_{ij} &= \langle \phi_i^k, \psi_j^k\rangle_{a_k}, && i = 1, \ldots, n_k, \quad j = 1, \ldots, m_k. \end{aligned} \tag{4.17}$$

Since $\langle \cdot, \cdot\rangle_{a_k}$ only induces a semi–norm on $H^1(\square_k)$, the matrix M_k is only positive semi–definite. However, if the subspaces X_k and Y_k are chosen such that B_k is invertible and

$$\mathcal{N}(M_k) = \left\{ \begin{pmatrix} \mathbf{0} \\ \mathbf{v}_2 \end{pmatrix};\ A_k \mathbf{v}_2 = \mathbf{0} \text{ and } C_k^\top \mathbf{v}_2 = \mathbf{0} \right\} \tag{4.18}$$

(ensuring that $\mathcal{N}(M_k) \subset V_k$), there exists a constant $\gamma_k \in [0, 1)$ such that (4.13) holds for the blocks given in (4.17), or equivalently

$$\langle u_k, v_k\rangle_{a_k}^2 \le \gamma_k^2 \langle u_k, u_k\rangle_{a_k}\langle v_k, v_k\rangle_{a_k}, \tag{4.19}$$

for all $u_k \in Y_k$ and $v_k \in X_k$, with the smallest such constant satisfying (4.14). The significance of (4.19) is that we may now estimate the global CBS constant $\gamma_{\min}$ satisfying (4.6). Indeed, combining (4.16) with (4.19) and employing the Cauchy–Schwarz inequality for sums yields

$$\langle u, v\rangle_a \le \sum_{\square_k \in \mathcal{T}_h} \gamma_{k,\min}\langle u_k, u_k\rangle_{a_k}^{\frac{1}{2}}\langle v_k, v_k\rangle_{a_k}^{\frac{1}{2}} \le \max_{\square_k \in \mathcal{T}_h} \gamma_{k,\min}\langle u, u\rangle_a^{\frac{1}{2}}\langle v, v\rangle_a^{\frac{1}{2}} \tag{4.20}$$

for all $u \in Y$ and $v \in X$. Comparing (4.20) with (4.6) we find that

$$\gamma_{\min} \le \max_{\square_k \in \mathcal{T}_h} \gamma_{k,\min} =: \gamma_*. \tag{4.21}$$

If we may assume that $a(\mathbf{x})$ is well–enough approximated by a function $a_h(\mathbf{x})$ that is constant in each element in $\mathcal{T}_h$, then, on each element $\square_k \in \mathcal{T}_h$ we have a symmetric and positive semi–definite matrix

$$M_k = \alpha_k \begin{bmatrix} B_k & C_k^\top \\ C_k & A_k \end{bmatrix}, \qquad \alpha_k := a_h|_{\square_k},$$

where B_k is invertible and B_k, A_k and C_k do not depend on α_k. The corresponding local eigenvalue problem is

$$(\alpha_k C_k)(\alpha_k B_k)^{-1}(\alpha_k C_k^\top)\mathbf{v}_2 = \theta(\alpha_k A_k)\mathbf{v}_2, \qquad \mathbf{v}_2 \notin \mathcal{N}(A_k),$$

which is equivalent to the eigenvalue problem (4.15) with $a = 1$. Thus, the associated CBS constant $\gamma_{k,\min}$ satisfying (4.14) is independent of α_k and $a_h(\mathbf{x})$. When the mesh $\mathcal{T}_h$ is uniform so that $\gamma_{k,\min}$ is independent of h (the element width), and fine enough so that $a_h(\mathbf{x})$ approximates $a(\mathbf{x})$ well, we my cheaply approximate γ_* in (4.21) by setting $a_k = 1$ in the definition of $\langle\cdot,\cdot\rangle_{a_k}$ in (4.16), and solving (4.15) for the blocks defined in (4.17) for a single internal element, as this is larger than the constant associated with the corner/edge elements (its a supremum over a superset). If the mesh is nonuniform or unstructured however, each $\gamma_{k,\min}$ depends on the geometry of the associated element in the mesh and thus $\gamma_{\min}$ cannot be approximated by computing a single constant in this manner.

We now revisit Example 4.1, and for the four choices of Y considered compute the local CBS constants $\gamma_{k,\min}$ associated with a single internal element. The space X_k is

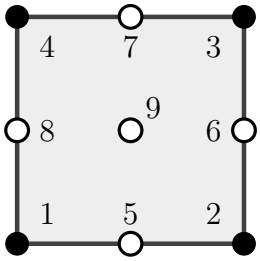


Figure 4.3: An arbitrary internal $\mathbb{Q}_1$ element $\square_k \in \mathcal{T}_h$. The numbering of the solid black and clear makers illustrates the chosen ordering of the basis functions of X_k and Y_k, respectively.

fixed to be the usual $\mathbb{Q}_1$ FEM space whose basis functions are defined with respect to the black markers in Figure 4.3, and ordered as shown. It follows that $n_k = 4$, and with $a = 1$, we have the classical $\mathbb{Q}_1$ element stiffness matrix

$$A_k = \begin{bmatrix} 2/3 & -1/6 & -1/3 & -1/6 \\ -1/6 & 2/3 & -1/6 & -1/3 \\ -1/3 & -1/6 & 2/3 & -1/6 \\ -1/6 & -1/3 & -1/6 & 2/3 \end{bmatrix}. \tag{4.22}$$

The ordering of the black markers in Figure 4.3 leads to A_k in (4.22) having a special structure. Indeed, it is a circulant matrix, meaning that all of its columns are cyclic permutations of its first column (and thus A_k is fully specified by its first column). The eigenpairs of circulant matrices can be neatly expressed analytically, which we exploit in the next section.

Lemma 4.1: [109, Corollary 5.16].

Let $T \in \mathbb{R}^{n\times n}$ be a circulant matrix with first column given by $\boldsymbol{t} = [t_0, \ldots, t_{n-1}]^\top$. The eigenvalues λ_j and eigenvectors $\boldsymbol{v}_j$ of T are

$$\lambda_j = \sum_{k=0}^{n-1} t_k \omega_j^k, \qquad \boldsymbol{v}_j = n^{(-1/2)}[1, \omega_j, \omega_j^2, \ldots, \omega_j^{n-1}]^\top \tag{4.23}$$

where $\omega_j = \exp(2\pi ij/n)$ and $i = \sqrt{-1}$.

The matrix A_k also has zero rows sums, and is thus a singular matrix with

$$\mathcal{N}(A_k) = \operatorname{span}\big\{(1,1,1,1)^\top\big\}. \tag{4.24}$$

For the four choices of Y_k considered in Example 4.1 which, by design, all have dimension $m_k = 5$, we construct the 2×2 block matrix $M_k \in \mathbb{R}^{9\times 9}$ and calculate $\gamma^2_{k,\min}$

by solving the eigenvalue problem (4.15). The ordering of the basis functions of Y_k is as illustrated by the clear markers in Figure 4.3.

Example 4.3: Local CBS constants.

For the four choices $Y_k = \mathbb{Q}_1(h/2), \mathbb{Q}_2(h), \mathbb{Q}_2(h/2), \mathbb{Q}_4(h)$, we find that

$$B_k = \left[\begin{array}{rrrr|r} 4/3 & -1/3 & 0 & -1/3 & -1/3 \\ -1/3 & 4/3 & -1/3 & 0 & -1/3 \\ 0 & -1/3 & 4/3 & -1/3 & -1/3 \\ -1/3 & 0 & -1/3 & 4/3 & -1/3 \\ \hline -1/3 & -1/3 & -1/3 & -1/3 & 8/3 \end{array}\right],$$

$$\left[\begin{array}{rrrr|r} 88/45 & -16/45 & 0 & -16/45 & -16/15 \\ -16/45 & 88/45 & -16/45 & 0 & -16/15 \\ 0 & -16/45 & 88/45 & -16/45 & -16/15 \\ -16/45 & 0 & -16/45 & 88/45 & -16/15 \\ \hline -16/15 & -16/15 & -16/15 & -16/15 & 256/45 \end{array}\right],$$

$$\left[\begin{array}{rrrr|r} 56/45 & -1/45 & 0 & -1/45 & -1/15 \\ -1/45 & 56/45 & -1/45 & 0 & -1/15 \\ 0 & -1/45 & 56/45 & -1/45 & -1/15 \\ -1/45 & 0 & -1/45 & 56/45 & -1/15 \\ \hline -1/15 & -1/15 & -1/15 & -1/15 & 112/45 \end{array}\right],$$

$$\left[\begin{array}{rrrr|r} 373/127 & -39/197 & 1/2326 & -39/197 & 84/247 \\ -39/197 & 373/127 & -39/197 & 1/2326 & 84/247 \\ 1/2326 & -39/197 & 373/127 & -39/197 & 84/247 \\ -39/197 & 1/2326 & -39/197 & 373/127 & 84/247 \\ \hline 84/247 & 84/247 & 84/247 & 84/247 & 3166/203 \end{array}\right],$$

respectively. Note that for each choice of Y_k, the matrix B_k is an invertible bordered matrix of the form

$$B_k = \begin{bmatrix} \bar{B}_k & \mathbf{b}_k \\ \mathbf{b}_k^\top & \mu_k \end{bmatrix}, \tag{4.25}$$

where, due to the ordering of the clear markers in Figure 4.3, $\bar{B}_k \in \mathbb{R}^{4\times 4}$ is a symmetric circulant matrix, $\mathbf{b}_k \in \mathbb{R}^4$ is a constant vector, and $\mu_k \in \mathbb{R}$ is a positive constant. In addition, each $C_k \in \mathbb{R}^{4\times 5}$ has the special structure

$$C_k = \alpha_k P, \qquad P := \begin{bmatrix} 1 & -1 & -1 & 1 & 0 \\ 1 & 1 & -1 & -1 & 0 \\ -1 & 1 & 1 & -1 & 0 \\ -1 & -1 & 1 & 1 & 0 \end{bmatrix}, \tag{4.26}$$

where $\alpha_k = \frac{1}{4}, \frac{1}{3}, \frac{1}{12}, \frac{1}{15}$ (each corresponding to a space Y_k, ordered as shown at the

start of the Example) and $\mathcal{N}(M_k) = \text{span}\{(\mathbf{0}^\top, 1, 1, 1, 1)^\top\}$. Since

$$\mathcal{N}(A_k) \subset \mathcal{N}(C_k^T) = \text{span}\{(1, 0, 1, 0)^\top, (0, 1, 0, 1)^\top\}$$

with $\mathcal{N}(A_k)$ defined in (4.24), all four spaces $\mathcal{N}(M_k)$ satisfy (4.18). Thus, the CBS constants associated with all four choices of Y_k satisfy (4.14) and can be found by solving the corresponding eigenvalue problem (4.15). We find that

$$\gamma_{k,\min}^2 = \frac{3}{8}, \frac{5}{11}, \frac{5}{112}, \frac{2363216}{195180975},$$

or equivalently, $\gamma_{k,\min}^2 = 0.3750, 0.4545, 0.0446, 0.0121$ (to four decimal places).

Comparing the results of Example 4.3 with those presented in Table 4.1 we confirm the relationship (4.21).

4.3 Novel Theoretical Estimates

In this section we fix X_k to be the local $\mathbb{Q}_1$ finite element space which fixes A_k to be as in (4.22), and demonstrate that if the matrices B_k and C_k have the structures (4.25) and (4.26) observed in Example 4.3, then the eigenvalues of (4.15) may be expressed analytically. Consequently, to determine the associated CBS constant $\gamma_{k,\min}$, the eigenvalue problem (4.15) need not be assembled nor solved. We begin with an abstract result, where we prove the existence of constant $\gamma_k \in [0, 1)$ satisfying (4.13) when $C_k B_k^{-1} C_k^\top$ has a certain simple structure. The following results are taken from [37], and we drop the subscript k to simplify notation.

Theorem 4.4: [37, Theorem 8].

Let $M \in \mathbb{R}^{9\times 9}$ be a symmetric and positive semidefinite matrix with the 2×2 block structure (4.12)*, where $B \in \mathbb{R}^{5\times 5}$ is symmetric and positive definite and A is given by* (4.22)*. If the matrix $CB^{-1}C^\top \in \mathbb{R}^{4\times 4}$ is of the form*

$$CB^{-1}C^\top = \delta \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix} =: \delta Q, \tag{4.27}$$

for some constant $\delta \in \mathbb{R}^+$, then there exists a constant $\gamma \in [0, 1)$ such that

$$(\mathbf{u}_1^\top C^\top \mathbf{v}_2)^2 \leq \gamma^2 \mathbf{u}_1^\top B \mathbf{u}_1 \mathbf{v}_2^\top A \mathbf{v}_2, \quad \forall \mathbf{u}_1 \in \mathbb{R}^5, \quad \forall \mathbf{v}_2 \in \mathbb{R}^4. \tag{4.28}$$

Proof.

It is sufficient to show that $\mathcal{N}(M)$ is given by (4.18). The result then follows from Theorem 4.2. Again, let $\mathbf{x}^\top = (\mathbf{u}_1^\top \mathbf{v}_2^\top) \in \mathbb{R}^9$ for $\mathbf{u}_1 \in \mathbb{R}^5$ and $\mathbf{v}_2 \in \mathbb{R}^4$ be such that $M\mathbf{x} = \mathbf{0}$. Then

$$B\mathbf{u}_1 + C^\top \mathbf{v}_2 = \mathbf{0}, \tag{4.29}$$

$$C\mathbf{u}_1 + A\mathbf{v}_2 = \mathbf{0}, \tag{4.30}$$

and $S\mathbf{v}_2 = \mathbf{0}$ for the Schur complement $S = A - CB^{-1}C^\top = A - \delta Q$. Since the matrices

$$S = \begin{bmatrix} \frac{2}{3} - \delta & -\frac{1}{6} & \delta - \frac{1}{3} & -\frac{1}{6} \\ -\frac{1}{6} & \frac{2}{3} - \delta & -\frac{1}{6} & \delta - \frac{1}{3} \\ \delta - \frac{1}{3} & -\frac{1}{6} & \frac{2}{3} - \delta & -\frac{1}{6} \\ -\frac{1}{6} & \delta - \frac{1}{3} & -\frac{1}{6} & \frac{2}{3} - \delta \end{bmatrix}$$

and A are circulant with zero row sums, we have

$$\mathcal{N}(S) = \mathcal{N}(A) = \text{span}\big\{(1, 1, 1, 1)^\top\big\} \tag{4.31}$$

and thus $\mathbf{v}_2 \in \mathcal{N}(A)$. We now show that $\mathbf{u}_1 = \mathbf{0}$ and $C^\top \mathbf{v}_2 = \mathbf{0}$ for all $\mathbf{v}_2 \in \mathcal{N}(A)$. If $\mathbf{v}_2 \in \mathcal{N}(A)$, from (4.30) it follows that $C\mathbf{u}_1 = \mathbf{0}$. Since B is invertible, (4.29) gives

$$\begin{aligned} 0 &= \mathbf{v}_2^\top C\mathbf{u}_1 = -(C^\top \mathbf{v}_2)^\top B^{-1}(C^\top \mathbf{v}_2), \\ 0 &= (B\mathbf{u}_1)^\top B^{-1}(B\mathbf{u}_1) \end{aligned}$$

Since B^{-1} is also invertible, we conclude that $B\mathbf{u}_1 = \mathbf{0}$ and $\mathbf{u}_1 = \mathbf{0}$. Finally, $\mathbf{u}_1 = \mathbf{0}$ and (4.29) gives $C^\top \mathbf{v}_2 = \mathbf{0}$.

We now show that provided the conditions of Theorem 4.4 are satisfied, then the CBS constant $\gamma_{\min}$ associated with (4.28) can be computed analytically.

Theorem 4.5: [37, Theorem 9].

Let $M \in \mathbb{R}^{9\times 9}$ be as in Theorem 4.4, then the smallest constant $\gamma \in [0,1)$ satisfying (4.28)*, denoted $\gamma_{\min}$ (the CBS constant), is given by*

$$\gamma_{\min}^2 = 2\delta, \tag{4.32}$$

where $\delta \in \mathbb{R}^+$ is the constant in (4.27)*.*

Proof.

Recall from (4.15) that $\gamma_{\min}^2$ is the largest eigenvalue $\theta_{\max}$ satisfying

$$CB^{-1}C^\top \mathbf{v}_2 = \theta A \mathbf{v}_2, \qquad \mathbf{v}_2 \notin \mathcal{N}(A). \tag{4.33}$$

By considering the expression $Q\mathbf{u} = \mathbf{0}$ it is easy to show that

$$\mathcal{N}(Q) = \operatorname{span}\left\{(1,0,1,0)^\top, (0,1,0,1)^\top\right\}, \tag{4.34}$$

and so $\mathcal{N}(A) \subset \mathcal{N}(Q)$. Under the stated assumptions, we have

$$CB^{-1}C^\top = \delta Q = \delta \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} =: \delta Q_1 \otimes Q_2,$$

and the set of eigenvalues is $\{2\delta, 2\delta, 0, 0\}$. The basis vectors of $\mathcal{N}(Q)$ in (4.34) are eigenvectors corresponding to the zero eigenvalues. In addition,

$$\mathbf{P}_1 = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \otimes \begin{bmatrix} 1 \\ 1 \end{bmatrix} =: \mathbf{p}_1 \otimes \mathbf{p}_2,$$

is an eigenvector corresponding to $\theta = 2\delta$. To see this, note that

$$\begin{aligned} CB^{-1}C^\top \mathbf{P}_1 &= \delta\,(Q_1 \otimes Q_2)\,(\mathbf{p}_1 \otimes \mathbf{p}_2) = \delta \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \otimes \mathbf{p}_2 \\ &= \delta \begin{bmatrix} 2 \\ -2 \end{bmatrix} \otimes \mathbf{p}_2 = 2\delta \mathbf{p}_1 \otimes \mathbf{p}_2 = 2\delta \mathbf{P}_1. \end{aligned}$$

The same is true for $\mathbf{P}_2 = [-1,1,1,-1]^\top$. Furthermore, the vectors $\mathbf{P}_1$ and $\mathbf{P}_2$ also satisfy $A\mathbf{P}_1 = \mathbf{P}_1$ and $A\mathbf{P}_2 = \mathbf{P}_2$ (and clearly do not belong to $\mathcal{N}(A)$, see (4.31)) and hence are eigenvectors of A with eigenvalue $\theta = 1$. Thus

$$CB^{-1}C^{\top}\mathbf{P}_i = \delta Q\mathbf{P}_i = 2\delta\mathbf{P}_i = 2\delta(1)\mathbf{P}_i = 2\delta A\mathbf{P}_i, \qquad i = 1, 2.$$

That is, $\mathbf{P}_1$ and $\mathbf{P}_2$ are eigenvectors in (4.33) with $\theta = 2\delta$. If we take $\mathbf{u}$ to be a member of $\mathcal{N}(Q)$ but not $\mathcal{N}(A)$, then (4.33) is trivially satisfied with $\theta = 0$. Hence, $\gamma^2_{\min} = \max\{0, 2\delta\} = 2\delta$.

Using the inverse for block matrices, we now demonstrate that if B has the structure (4.25) and C has the structure (4.26), the matrix $CB^{-1}C^{\top}$ always has the structure given in (4.27) and an explicit expression is available for the constant δ in (4.32).

Lemma 4.2: [37, Lemma 1].

If the matrix $C \in \mathbb{R}^{4\times 5}$ has the form $C = \alpha P$ for $\alpha \in \mathbb{R}$ and P in (4.26)*, and if $B \in \mathbb{R}^{5\times 5}$ has the form* (4.25)*, where $\bar{B} \in \mathbb{R}^{4\times 4}$ is a symmetric circulant matrix, $\mathbf{b} \in \mathbb{R}^4$ is a constant vector, and $\mu \in \mathbb{R}$, then $CB^{-1}C^{\top}$ has the form* (4.27)*.*

Proof.

First we show that if B has the form (4.25) then so does B^{-1}. We have

$$B^{-1} = \begin{bmatrix} \bar{B}^{-1} + \nu^{-1}\bar{B}^{-1}\mathbf{b}\mathbf{b}^{\top}\bar{B}^{-1} & -\nu^{-1}\bar{B}^{-1}\mathbf{b} \\ -\nu^{-1}\mathbf{b}^{\top}\bar{B}^{-1} & \nu^{-1} \end{bmatrix}$$

where $\nu := \mu - \mathbf{b}^{\top}\bar{B}^{-1}\mathbf{b} \in \mathbb{R}$ is the Schur complement. Since $\bar{B}$ is symmetric and circulant, so is its inverse (see [39], for example). Consequently, $\mathbf{q} := \bar{B}^{-1}\mathbf{b} \in \mathbb{R}^{4\times 1}$ is a constant vector and $\mathbf{q}\mathbf{q}^{\top} \in \mathbb{R}^{4\times 4}$ is a constant matrix. This is because $\mathbf{b}$ is a constant vector and the row sums of a circulant matrix are equal. Therefore

$$B^{-1} = \begin{bmatrix} \hat{B} & \hat{\mathbf{b}} \\ \hat{\mathbf{b}}^{\top} & \nu^{-1} \end{bmatrix}, \tag{4.35}$$

where $\hat{B} := \bar{B}^{-1} + \nu^{-1}\mathbf{q}\mathbf{q}^{\top} \in \mathbb{R}^{4\times 4}$ is a symmetric circulant matrix bordered by $\hat{\mathbf{b}} := -\nu^{-1}\mathbf{q} \in \mathbb{R}^{4\times 1}$ and $\nu \in \mathbb{R}$ is a constant. Hence, B^{-1} has the form

$$B^{-1} = \begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \alpha_2 & \times \\ \alpha_2 & \alpha_1 & \alpha_2 & \alpha_3 & \times \\ \alpha_3 & \alpha_2 & \alpha_1 & \alpha_2 & \times \\ \alpha_2 & \alpha_3 & \alpha_2 & \alpha_1 & \times \\ \times & \times & \times & \times & \times \end{bmatrix}, \tag{4.36}$$

for some $\alpha_1, \alpha_2, \alpha_3 \in \mathbb{R}$ and, for the rest of the proof, the elements marked with $\times$ are not important. Now, elementary matrix multiplication with C gives

$$CB^{-1} = (\alpha_1 - \alpha_3)\alpha \begin{bmatrix} 1 & -1 & -1 & 1 & \bar{\times} \\ 1 & 1 & -1 & -1 & \bar{\times} \\ -1 & 1 & 1 & -1 & \bar{\times} \\ -1 & -1 & 1 & 1 & \bar{\times} \end{bmatrix},$$

(again the elements marked with $\bar{\times}$ are not important) and

$$CB^{-1}C^\top = 4(\alpha_1 - \alpha_3)\alpha^2 \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix} = \delta Q$$

with $\delta := 4(\alpha_1 - \alpha_3)\alpha^2 \in \mathbb{R}$.

Combining the last two results, we find that $\gamma_{\min}^2 = 8(\alpha_1 - \alpha_3)\alpha^2$ and thus to compute the CBS constant we need only know α (one entry of C) and α_1 and α_3 (two entries of the first column of B^{-1} or $\hat{B}$). Applying Lemma 4.1 to $\bar{B}$ in Lemma 4.3 below enables us to determine the first column of $\bar{B}^{-1}$ in the definition of $\hat{B}$.

Lemma 4.3: [37, Lemma 3].

Let the principle minor $\bar{B}$ in (4.25) *of the matrix B be given by*

$$\bar{B} = \begin{bmatrix} b_1 & b_2 & b_3 & b_2 \\ b_2 & b_1 & b_2 & b_3 \\ b_3 & b_2 & b_1 & b_2 \\ b_2 & b_3 & b_2 & b_1 \end{bmatrix}. \tag{4.37}$$

Then the eigenvalues of $\bar{B}$ are

$$\lambda_1 = b_1 - b_3, \quad \lambda_2 = b_1 - 2b_2 + b_3, \quad \lambda_3 = \lambda_1, \quad \lambda_4 = b_1 + 2b_2 + b_3, \tag{4.38}$$

and the eigenvectors of $\bar{B}$ are given by the columns of the unitary matrix

$$F^* = \frac{1}{2} \begin{bmatrix} 1 & 1 & 1 & 1 \\ i & -1 & -i & 1 \\ -1 & 1 & -1 & 1 \\ -i & -1 & i & 1 \end{bmatrix}.$$

Moreover, $\bar{B} = F^* diag(\boldsymbol{\lambda}) F$, where $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \lambda_3, \lambda_4]^\top$.

Combining the above results, gives the following final result.

Theorem 4.6: [37, Theorem 10].

Let the assumptions of Theorem 4.4 and Lemma 4.2 hold, with the entries of $\bar{B}$ labelled as in (4.37)*. Then, the square of the CBS constant associated with* (4.28) *is given by*

$$\gamma_{\min}^2 = 8\alpha^2(b_1 - b_3)^{-1}. \tag{4.39}$$

Proof.

From Lemma 4.2, we have $CB^{-1}C^\top = \delta Q$ with $\delta = 4\alpha^2(\alpha_1 - \alpha_3)$ where α_1 and α_3 are elements of the matrix $\hat{B}$ in (4.35), which depends on the inverse of $\bar{B}$ in (4.25). By Lemma 4.3,

$$\bar{B}^{-1} = F^* \begin{bmatrix} \lambda_1^{-1} & 0 & 0 & 0 \\ 0 & \lambda_2^{-1} & 0 & 0 \\ 0 & 0 & \lambda_1^{-1} & 0 \\ 0 & 0 & 0 & \lambda_4^{-1} \end{bmatrix} F.$$

Since $\bar{B}^{-1}$ is circulant, its entries are known once we specify its first column $\bar{\mathbf{c}}$. Furthermore, since

$$F = (F^*)^* = \frac{1}{2} \begin{bmatrix} 1 & -i & -1 & i \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \\ 1 & 1 & 1 & 1 \end{bmatrix},$$

we have $\bar{\mathbf{c}} := \bar{B}^{-1}\mathbf{e}_1 = F^*\mathrm{diag}(1./\boldsymbol{\lambda})F\mathbf{e}_1 = \frac{1}{2}F^*(1./\boldsymbol{\lambda})$. It follows that

$$\bar{\mathbf{c}} = \frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \\ i & -1 & -i & 1 \\ -1 & 1 & -1 & 1 \\ -i & -1 & i & 1 \end{bmatrix} \begin{bmatrix} \lambda_1^{-1} \\ \lambda_2^{-1} \\ \lambda_1^{-1} \\ \lambda_4^{-1} \end{bmatrix} = \frac{1}{4} \begin{bmatrix} 2\lambda_1^{-1} + \lambda_2^{-1} + \lambda_4^{-1} \\ \lambda_4^{-1} - \lambda_2^{-1} \\ \lambda_2^{-1} + \lambda_4^{-1} - 2\lambda_1^{-1} \\ \lambda_4^{-1} - \lambda_2^{-1} \end{bmatrix}. \tag{4.40}$$

Now, since $\hat{B} := \bar{B}^{-1} + \nu^{-1}\mathbf{q}\mathbf{q}^\top$ in Lemma 4.2, we know that

$$\alpha_1 = [\bar{\mathbf{c}}]_1 + \tau, \qquad \alpha_3 = [\bar{\mathbf{c}}]_3 + \tau, \qquad \tau \in \mathbb{R},$$

and consequently, by considering (4.40) and the eigenvalues (4.38), we have

$$\alpha_1 - \alpha_3 = [\bar{\mathbf{c}}]_1 - [\bar{\mathbf{c}}]_3 = \frac{1}{4}\left(2\lambda_1^{-1} + 2\lambda_1^{-1}\right) = \lambda_1^{-1} = (b_1 - b_3)^{-1}.$$

Table 4.3: The constants $\alpha, b_1, b_3 \in \mathbb{R}$ required to compute $\gamma^2_{k,\min} = 8\alpha^2(b_1 - b_3)^{-1}$ when X_k is the local $\mathbb{Q}_1$ space and Y_k is chosen as in Example 4.1.

Y_k	α	b_1	b_3	$\gamma^2_{k,\min}$
$\mathbb{Q}_1(h/2)$	1/4	4/3	0	0.3750
$\mathbb{Q}_2(h)$	1/3	88/45	0	0.4545
$\mathbb{Q}_2(h/2)$	1/12	56/45	0	0.0446
$\mathbb{Q}_4(h)$	1/15	373/127	1/2326	0.0121

Since B is symmetric and positive definite, so is $\bar{B}$. Consequently, $\lambda_1 > 0$ and $\delta = 4\alpha^2 (b_1 - b_3)^{-1} > 0$. The result follows by Theorem 4.5.

For the four choices of Y_k described in Example 4.3, we record the associated values of α, b_1 and b_3, as well as the squares of the associated CBS constants given by (4.39), in Table 4.3 (to stress that these are local quantities we reintroduce the subscript k). The results match the CBS constants computed numerically in Example 4.3, which, unlike the new analytical method, required the eigenproblem (4.15) to be assembled and solved.

4.4 Summary

In this chapter, we investigated the constant $\gamma \in [0, 1)$ in the strengthened Cauchy–Schwarz inequality (4.6) for the inner product (4.7) associated with the weak diffusion problem (3.24). Recall that for two subspaces $X, Y \subset H^1_0(D)$ satisfying $X \cap Y = \{0\}$, where X is the space chosen for the discrete weak problem (3.26) and Y is the space chosen for the error problem (3.29), any constant γ satisfying (4.6) also appears in the error bound (3.30). We denoted by $\gamma_{\min}$ the smallest such constant and called it the CBS constant. Determining pairs X, Y for which the CBS constant is close to zero ensures that the term $\sqrt{1 - \gamma^2_{\min}}$ in the error bound (3.30) is close to one. In Examples 4.1 and 4.2, we solved a large eigenvalue problem to compute the CBS constant for fixed choices of X and various standard and non–standard choices of Y of a fixed small dimension. When $X = \mathbb{Q}_1(h)$, the space $Y = \mathbb{Q}_4(h)$ yields the smallest CBS constant of the four choices considered ($\gamma^2_{\min} \leq 0.0121$). When $X = \mathbb{Q}_2(h)$, $Y = \mathbb{Q}^r_4(h)$ yields the smallest constant ($\gamma^2_{\min} \leq 0.3600$). Finally, in Section 4.3 we exploited results from linear algebra to prove novel theoretical estimates for CBS constants associated with

the space $X = \mathbb{Q}_1(h)$ and certain special choices of Y (including the four considered in Example 4.1), for which no eigenvalue problem need be assembled nor solved. We presented our main and final result in Theorem 4.6.

Whilst originating from a deterministic problem, the CBS constants investigated in this chapter also play a role in a posteriori error estimation for the parametric diffusion problem, which is the focus of the next chapter.

Chapter 5

The Parametric Diffusion Problem

We begin this chapter by briefly considering the *stochastic* diffusion problem, where the diffusion coefficient $a(\mathbf{x})$ in (3.22)–(3.23) is replaced with a second–order random field $a(\mathbf{x}, \omega) \in L^2(\Omega, L^2(D))$ (recall Definitions 2.15 and 2.13). We assume that $a(\mathbf{x}, \omega)$ is a function of a countably infinite number of random variables $\xi_m(\omega) : \Omega \to \Gamma_m \subset \mathbb{R}$ for $m = 1, 2, \ldots$ associated with a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, which we collect into the vector $\boldsymbol{\xi}(\omega) = (\xi_1(\omega), \xi_2(\omega), \ldots)^\top : \Omega \to \Gamma$ (a multivariate random variable), for the observation space $\Gamma = \Gamma_1 \times \Gamma_2 \times \cdots$. More specifically, we assume that $a(\mathbf{x}, \omega)$ is of the form

$$a(\mathbf{x}, \omega) = a_0(\mathbf{x}) + \sum_{m=1}^{\infty} a_m(\mathbf{x}) \xi_m(\omega), \tag{5.1}$$

with the random variables $\{\xi_m(\omega);\ m \in \mathbb{N}\}$ bounded, mean zero and i.i.d. (independent and identically distributed). Note that the structure of $a(\mathbf{x}, \omega)$ in (5.1) is a natural choice; recall the structure of KL expansions of second–order random fields in Theorem 2.3. We consider the problem: find $u(\mathbf{x}, \omega) : D \times \Omega \to \mathbb{R}$ such that $\mathbb{P}$–a.s. (i.e., with probability one)

$$-\nabla \cdot (a(\mathbf{x}, \omega) \nabla u(\mathbf{x}, \omega)) = f(\mathbf{x}), \qquad \mathbf{x} \in D, \tag{5.2}$$

$$u(\mathbf{x}, \omega) = 0, \qquad \mathbf{x} \in \partial D, \tag{5.3}$$

where, as in Section 3.3, $D \subset \mathbb{R}^2$ is a bounded polygonal domain and $f(\mathbf{x})$ satisfies Assumption 3.3.

The solution $u(\mathbf{x}, \omega)$ to (5.2)–(5.3) is uncertain through its dependence on $\boldsymbol{\xi}(\omega)$. Once obtained, we may quantify the uncertainty in $u(\mathbf{x}, \omega)$ by computing important

statistical quantities such as its expectation and variance. It is also possible to compute probabilities involving $u(\mathbf{x}, \omega)$, such as the probability that a rare or catastrophic event may occur (for example, the probability that $u(\mathbf{x}, \cdot)$ surpasses some critical value at the point $\mathbf{x} \in D$). In the same way, the uncertainty in many other deterministic models can be quantified by incorporating random variables and/or fields into the model and solving the resulting stochastic problem. For example, stochastic convection diffusion, elasticity and Navier–Stokes problems are considered in [105, 66, 73], [62, 44, 63] and [81, 98] respectively. With them, the new stochastic components bring an additional complexity, the handling of which is often intricate and problem dependent

The stochastic problem (5.2)–(5.3) is well understood theoretically [40, 93, 7, 8, 112, 6], and its efficient numerical approximation is the focus of many works. For example, Monte Carlo, collocation and Galerkin–based approaches are taken in [34, 17, 32], [75, 74, 102, 61] and [44, 22, 21, 47], respectively. The aim of this thesis is to design new Galerkin–based methods that achieve the theoretically best known rates of convergence with respect to the number of degrees of freedom. More information about these rates will be provided at the end of this chapter and the start of Chapter 6. For now, we are interested in approximating solutions to (5.2)–(5.3) in the weak sense, but working on the abstract domain Ω is computationally inconvenient. In the next section we consider an equivalent *parametric* formulation of (5.2)–(5.3) that is more straightforward to work with.

5.1 The Parametric Formulation

Instead of working with random variables $\xi_m(\omega)$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, in this chapter we perform a change of variable and work with parameters $y_m := \xi_m(\omega)$ (the image of $\xi_m(\omega)$) on the space $(\Gamma_m, \mathcal{B}(\Gamma_m), \pi_m)$, where π_m is the probability distribution of $\xi_m(\omega)$, and $\mathcal{B}(\Gamma_m)$ denotes the Borel σ–algebra on Γ_m. Since each $\xi_m(\omega)$ is mean zero;

$$\int_{\Gamma_m} y_m \, d\pi_m(y_m) = 0, \qquad m = 1, 2, \dots. \tag{5.4}$$

Moreover, we collect the parameters $y_m \in \Gamma_m$ into the vector $\mathbf{y} = (y_1, y_2, \dots)^\top \in \Gamma$ and work on the tensor–product space $(\Gamma, \mathcal{B}(\Gamma), \pi)$, where π is the joint probability

distribution of $\boldsymbol{\xi}(\omega)$, and $\mathcal{B}(\Gamma)$ denotes the Borel σ–algebra on Γ. Since the underlying random variables $\xi_m(\omega)$ are independent, the joint probability distribution π is a product measure, given by

$$\pi(\mathbf{y}) = \prod_{m=1}^{\infty} \pi_m(y_m).$$

We now replace $a(\mathbf{x})$ in (3.22)–(3.23) with coefficients $a(\mathbf{x}, \mathbf{y}) \in L^2(\Gamma, L^2(D))$ of the form

$$a(\mathbf{x}, \mathbf{y}) = a_0(\mathbf{x}) + \sum_{m=1}^{\infty} a_m(\mathbf{x}) y_m, \tag{5.5}$$

and consider the parametric diffusion problem: find $u(\mathbf{x}, \mathbf{y}) : D \times \Gamma \to \mathbb{R}$ that satisfies

$$-\nabla \cdot (a(\mathbf{x}, \mathbf{y}) \nabla u(\mathbf{x}, \mathbf{y})) = f(\mathbf{x}), \qquad \mathbf{x} \in D,\ \mathbf{y} \in \Gamma, \tag{5.6}$$

$$u(\mathbf{x}, \mathbf{y}) = 0, \qquad \mathbf{x} \in \partial D,\ \mathbf{y} \in \Gamma. \tag{5.7}$$

We refer to Γ as the *parameter domain* and make the following important assumption on $a(\mathbf{x}, \mathbf{y})$ in (5.5).

Assumption 5.1.

The parameters y_m are images of uniformly distributed random variables $\xi_m(\omega) \sim U(-1, 1)$ so that $y_m \in \Gamma_m = [-1, 1]$ for $m = 1, 2, \ldots$. Moreover, the spatially varying terms $a_0(\mathbf{x}), a_m(\mathbf{x}) \in L^\infty(D)$ for $m = 1, 2, \ldots$ and $\|a_m\|_\infty \to 0$ sufficiently quickly as $m \to \infty$ so that

$$\sum_{m=1}^{\infty} \|a_m\|_\infty < \operatorname*{ess\,inf}_{\mathbf{x} \in D} a_0(\mathbf{x}) < \infty. \tag{5.8}$$

The affine decomposition (5.5) is not the only possibility. Many developments have been made recently concerning the solution of parametric PDE problems with lognormal or log–transformed coefficients of the form $a(\mathbf{x}, \mathbf{y}) = e^{z(\mathbf{x}, \mathbf{y})}$, where $z(\mathbf{x}, \mathbf{y})$ could be a Gaussian random field [55, 105, 31, 106, 87], but such coefficients are not the focus of this work. Additionally, $f(\mathbf{x})$ in (5.6) may feature uncertainty and be modelled by expansions like (5.5) and the boundary conditions may be nonzero and/or non–Dirichlet. The methods discussed herein are easily extended to handle such situations – we make convenient choices of f and the boundary conditions to simplify the analysis of the corresponding weak problem. In Section 5.2 we discuss the weak formulation of (5.6)–(5.7), but first we introduce some test problems.

5.1.1 Test Problems

In this work we consider the following four test problems, each with a different expansion $a(\mathbf{x}, \mathbf{y})$ in (5.6) of the form (5.5) so that the sequences $\{\|a_m\|_\infty\}_{m=1}^\infty$ decay at different rates, where $\|a_m\|_\infty \geq \|a_{m+1}\|_\infty$ for all $m > 1$. We begin with Test problem TP1 below which has the slowest decaying sequence.

Test Problem 1 (TP1)

First, we consider a problem from [18]. Let $D = [-1, 1]^2$ and

$$f(\mathbf{x}) = \tfrac{1}{8}(2 - x_1^2 - x_2^2), \qquad \mathbf{x} = (x_1, x_2)^\top \in D.$$

We choose

$$a(\mathbf{x}, \mathbf{y}) = 1 + \sqrt{3} \sum_{m=1}^{\infty} \sqrt{\nu_m}\phi_m(\mathbf{x})y_m, \tag{5.9}$$

where (ν_m, ϕ_m) are the eigenpairs of the integral operator (2.10) for the covariance function (2.16) with $\sigma = 0.15$ and $\ell = 2$. This choice is simply the parametric form of $a(\mathbf{x}, \omega)$ defined in Example 2.5 and thus $\sqrt{\nu_m}$ behaves like m^{-1} as $m \to \infty$ [70]. Note that the random variables $\xi_m(\omega) \sim U(-\sqrt{3}, \sqrt{3})$ defined in Example 2.5 are rescaled such that

$$y_m := \frac{\xi_m(\omega)}{\sqrt{3}}, \qquad m = 1, 2, \ldots,$$

takes values in $\Gamma_m = [-1, 1]$.

Test Problems 2 and 3 (TP2, TP3)

Next, we consider a problem from [44, 22] where $D = [0, 1]^2$, $f(\mathbf{x}) = 1$ and $a(\mathbf{x}, \mathbf{y})$ is the parametric form of $a(\mathbf{x}, \omega)$ defined in Examples 2.6 and 2.7. Specifically, we choose

$$a(\mathbf{x}, \mathbf{y}) = 1 + \sum_{m=1}^{\infty} \alpha_m \phi_m(\mathbf{x})y_m, \tag{5.10}$$

where $\phi_m(\mathbf{x})$ is defined in (2.18) and we set $\alpha_m = 0.547m^{-2}, 0.832m^{-4}$ for Test Problems TP2 and TP3, respectively. Note that synthetic expansions like (5.10) may be employed in conjunction with more complex geometries such as the L–shape and crack domains D discussed in Section 3.3. In contrast, the eigenpairs (ν_m, ϕ_m) in (5.9) are known only for domains D with a simple geometry.

Test Problem 4 (TP4)

Finally, we consider a problem from [70]. Let $f(\mathbf{x})$ and D be as in TP2 and let $a(\mathbf{x}, \mathbf{y})$ be the parametric form of $a(\mathbf{x}, \omega)$ defined in Example 2.8. Specifically, we choose

$$a(\mathbf{x}, \mathbf{y}) = 2 + \sqrt{3} \sum_{m=0}^{\infty} \sqrt{\nu_m} \phi_m(\mathbf{x}) y_m, \tag{5.11}$$

where $\phi_m(\mathbf{x})$ and ν_m are the terms in (2.19) with $\ell = 0.65$, but reordered in terms of a single index m such that $\nu_1 \geq \nu_2 \geq \cdots$. Note that the sequence $\{\|a_m\|_\infty\}_{m=1}^{\infty}$ decays most quickly for this test problem.

5.2 The Weak Parametric Diffusion Problem

The weak formulation of the parametric problem (5.6)–(5.7) is:

$$\text{find: } u \in V := L^2_\pi(\Gamma, H^1_0(D)) : \qquad B(u, v) = F(v), \qquad \text{for all } v \in V \tag{5.12}$$

for the symmetric bilinear form and linear functional

$$B(u, v) = \int_\Gamma \int_D a(\mathbf{x}, \mathbf{y}) \nabla u(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}, \mathbf{y}) \, d\mathbf{x} \, d\pi(\mathbf{y}), \tag{5.13}$$

$$F(v) = \int_\Gamma \int_D f(\mathbf{x}) v(\mathbf{x}, \mathbf{y}) \, d\mathbf{x} \, d\pi(\mathbf{y}), \tag{5.14}$$

respectively [40], where $H^1_0(D)$ is the same space considered in Section 3.3 for the analogous deterministic weak problem, and $L^2_\pi(\Gamma, H^1_0(D))$ is the Bochner space (recall Definition 2.13 and Example 2.3)

$$L^2_\pi(\Gamma, H^1_0(D)) = \left\{ v(\mathbf{x}, \mathbf{y}) : D \times \Gamma \to \mathbb{R}; \ \int_\Gamma \|v(\cdot, \mathbf{y})\|^2_{H^1_0(D)} \, d\pi(\mathbf{y}) < \infty \right\},$$

associated with the probability space $(\Gamma, \mathcal{B}(\Gamma), \pi)$. Note that V is a Hilbert space and thus (5.12) is an example of the abstract problem (3.1). In addition, V is equipped with the norm

$$\|v\|_V = \left[\int_\Gamma \|v(\cdot, \mathbf{y})\|^2_{H^1_0(D)} \, d\pi(\mathbf{y}) \right]^{\frac{1}{2}}, \qquad \text{for all } v \in V.$$

The following assumption on $a(\mathbf{x}, \mathbf{y})$ ensures that $B(\cdot, \cdot)$ in (5.13) is bounded and coercive over V, and induces the energy norm $\|v\|_B = B(v, v)^{1/2}$ for all $v \in V$.

Assumption 5.2.

There exist constants $a_{\min}, a_{\max} \in \mathbb{R}^+$ such that

$$0 < a_{\min} \leq a(\mathbf{x}, \mathbf{y}) \leq a_{\max} < \infty, \qquad \text{a.e. in } D \times \Gamma, \tag{5.15}$$

and thus $a(\mathbf{x}, \mathbf{y}) \in L^\infty(D \times \Gamma)$.

Note that (5.8) is a sufficient condition for Assumption 5.2 to hold. We now show that the weak problem (5.12) is well–posed. By Lemma 2.2 we find that

$$|F(v)| \leq \|f\|_{L^2(\Gamma, L^2(D))} \|v\|_{L^2(\Gamma, L^2(D))} = \|f\|_{L^2(D)} \|v\|_{L^2(\Gamma, L^2(D))},$$

for all $v \in V$, and from Theorem 2.1

$$\|v(\cdot, \mathbf{y})\|_{L^2(D)} \leq C_p \|v(\cdot, \mathbf{y})\|_{H_0^1(D)}, \qquad \mathbf{y} \in \Gamma, \qquad \text{for all } v \in V.$$

Squaring both sides and integrating over Γ yields $\|v\|^2_{L^2(\Gamma, L^2(D))} \leq C_p^2 \|v\|_V^2$ so that

$$|F(v)| \leq C_p \|f\|_{L^2(D)} \|v\|_V, \qquad \text{for all } v \in V,$$

and thus under Assumption 3.3 the linear functional $F(\cdot)$ is bounded over V. By the same arguments used in Section 3.3 for the deterministic problem, it is straightforward to show that

$$|B(u, v)| \leq a_{\max} \|u\|_V \|v\|_V, \qquad B(u, u) \geq a_{\min} \|u\|_V^2, \qquad \text{for all } u, v \in V,$$

and thus under Assumption 5.2 the bilinear form $B(\cdot, \cdot)$ is both bounded and coercive over V. Since V is a Hilbert space, by Lemma 2.1 there exists a unique $u \in V$ satisfying (5.12).

Additionally, under Assumption 5.2 the bilinear form $B(\cdot, \cdot)$ in (5.13) also inherits the decomposition in (5.5), that is

$$B(u, v) = B_0(u, v) + \sum_{m=1}^{\infty} B_m(u, v), \qquad \text{for all } u, v, \in V, \tag{5.16}$$

with the component bilinear forms $B_0, B_m : V \times V \to \mathbb{R}$ given by

$$B_0(u, v) = \int_\Gamma \int_D a_0(\mathbf{x}) \nabla u(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}, \mathbf{y}) \, d\mathbf{x} \, d\pi(\mathbf{y}), \tag{5.17}$$

$$B_m(u, v) = \int_\Gamma \int_D a_m(\mathbf{x}) y_m \nabla u(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}, \mathbf{y}) \, d\mathbf{x} \, d\pi(\mathbf{y}), \tag{5.18}$$

for $m = 1, 2, \ldots$. Under the following assumption on $a_0(\mathbf{x})$, the parameter–free bilinear form $B_0(\cdot, \cdot)$ in (5.17) also induces the norm $\|\cdot\|_{B_0} = B_0(\cdot, \cdot)^{1/2}$ on V.

Assumption 5.3.

There exist constants $a^0_{\min}, a^0_{\max} \in \mathbb{R}^+$ such that

$$0 < a^0_{\min} \leq a_0(\mathbf{x}) \leq a^0_{\max} < \infty, \qquad \text{a.e. in } D,$$

and thus $a_0(\mathbf{x}) \in L^\infty(D)$.

Under Assumptions 5.2 and 5.3 it is straightforward to show that

$$\lambda \|v\|^2_B \leq \|v\|^2_{B_0} \leq \Lambda \|v\|^2_B, \qquad \text{for all } v \in V, \tag{5.19}$$

where $0 < \lambda < 1 < \Lambda < \infty$, and

$$\lambda = a^0_{\min} a^{-1}_{\max}, \qquad \Lambda = a^0_{\max} a^{-1}_{\min}. \tag{5.20}$$

5.3 SGFEM Approximation

We now seek Galerkin approximations to $u \in V$ by projecting the problem (5.12) onto a finite–dimensional subspace $X \subset V$. The spaces $L^2_\pi(\Gamma, H^1_0(D))$ and $L^2_\pi(\Gamma) \otimes H^1_0(D)$ are isometrically isomorphic (the norm is preserved; see [92]), meaning that functions with the properties of those in V can be constructed by tensorising functions in $L^2_\pi(\Gamma)$, given by

$$L^2_\pi(\Gamma) = \left\{ v(\mathbf{y}) : \Gamma \to \mathbb{R};\ \int_\Gamma v(\mathbf{y})^2 \, d\pi(\mathbf{y}) < \infty \right\},$$

with functions in $H^1_0(D)$.

We consider the finite–dimensional problem

$$\text{find: } u_X \in X : \qquad B(u_X, v) = F(v), \qquad \text{for all } v \in X, \tag{5.21}$$

where X is the tensor–product subspace

$$X := H_1 \otimes P, \qquad H_1 \subset H^1_0(D), \qquad P \subset L^2_\pi(\Gamma). \tag{5.22}$$

The space H_1 is chosen to be a FEM space of piecewise polynomials constructed with respect to a mesh $\mathcal{T}_h$ on D. Since the solution $u \in V$ is analytic with respect to

the parameters y_m [7, 35], we construct P using global polynomials on Γ. Note that in the same way we construct H_1, we may also construct P using piecewise polynomials of a fixed degree with only local support on "elements" of Γ, see [110, 111, 67] for example, but such spaces are not considered in this work. For now, we delay making specific choices of H_1 and P and simply let

$$H_1 = \operatorname{span}\big\{\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots, \phi_n(\mathbf{x})\big\}, \qquad P = \operatorname{span}\big\{\psi_1(\mathbf{y}), \psi_2(\mathbf{y}), \ldots, \psi_s(\mathbf{y})\big\}.$$

In turn, this provides us with a basis for X, namely

$$X = \operatorname{span}\big\{\phi_i(\mathbf{x})\psi_j(\mathbf{y});\ i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, s\big\}, \quad N_X := \dim(X) = ns.$$

5.3.1 Linear Systems

Expressing the Galerkin approximation $u_X \in X$ as

$$u_X(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n}\sum_{j=1}^{s} u_{ij}\phi_i(\mathbf{x})\psi_j(\mathbf{y}), \qquad u_{ij} \in \mathbb{R},$$

and choosing the linearly independent test functions $v = \phi_q(\mathbf{x})\psi_r(\mathbf{y})$ in (5.21) yields

$$\sum_{i=1}^{n}\sum_{j=1}^{s} u_{ij}B(\phi_i(\mathbf{x})\psi_j(\mathbf{y}), \phi_q(\mathbf{x})\psi_r(\mathbf{y})) = F(\phi_q(\mathbf{x})\psi_r(\mathbf{y})),$$

for $q = 1, 2, \ldots, n$ and $r = 1, 2, \ldots, s$. This is a linear system of equations $A\mathbf{u} = \mathbf{b}$ for the block vector of coefficients

$$\mathbf{u} = \Big[[u_{11}, u_{21}, \ldots, u_{n1}]^\top, [u_{12}, u_{22}, \ldots, u_{n2}]^\top, \ldots, [u_{1s}, u_{2s}, \ldots, u_{ns}]^\top\Big]^\top.$$

Since $\mathbf{u}^\top A\mathbf{u} = ||u_X||_B^2 > 0$ for all $u_X \neq 0$, the matrix A is positive definite. Furthermore, $A \in \mathbb{R}^{N_X \times N_X}$ and $\mathbf{b} \in \mathbb{R}^{N_X}$ have the block structure

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1s} \\ A_{21} & A_{22} & \cdots & A_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ A_{s1} & A_{s2} & \cdots & A_{ss} \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_s \end{bmatrix}, \tag{5.23}$$

with entries

$$[A_{rj}]_{qi} = \int_\Gamma \psi_j(\mathbf{y})\psi_r(\mathbf{y}) \int_D a(\mathbf{x}, \mathbf{y})\nabla\phi_i(\mathbf{x}) \cdot \nabla\phi_q(\mathbf{x})\ d\mathbf{x}\ d\pi(\mathbf{y}),$$
$$[\mathbf{b}_r]_q = \int_\Gamma \psi_r(\mathbf{y})\ d\pi(\mathbf{y}) \int_D f(\mathbf{x})\phi_q(\mathbf{x})\ d\mathbf{x},$$

for $i, q = 1, 2, \ldots, n$ and $j, r = 1, 2, \ldots, s$, and thus A is symmetric since $A_{rj} = A_{jr}^{\top}$. Notice that, due to the decomposition of $a(\mathbf{x}, \mathbf{y})$ in (5.5), each block $A_{rj} \in \mathbb{R}^{n\times n}$ admits the decomposition

$$A_{rj} = \langle \psi_j, \psi_r \rangle_{L^2_\pi(\Gamma)} K_0 + \sum_{m=1}^{\infty} \langle y_m \psi_j, \psi_r \rangle_{L^2_\pi(\Gamma)} K_m, \tag{5.24}$$

where $K_m \in \mathbb{R}^{n\times n}$ has entries

$$[K_m]_{qi} = \int_D a_m(\mathbf{x}) \nabla\phi_i(\mathbf{x}) \cdot \nabla\phi_q(\mathbf{x}) \, d\mathbf{x}, \qquad m = 0, 1, \ldots,$$

for $i, q = 1, 2, \ldots, n$. Similarly, by defining the matrices $G_m \in \mathbb{R}^{s\times s}$ with entries

$$[G_0]_{rj} = \langle \psi_j, \psi_r \rangle_{L^2_\pi(\Gamma)}, \qquad [G_m]_{rj} = \langle y_m \psi_j, \psi_r \rangle_{L^2_\pi(\Gamma)}, \qquad m = 1, 2, \ldots, \tag{5.25}$$

for $j, r = 1, 2, \ldots, s$, we observe that A in (5.23) admits the Kronecker–product representation

$$A = G_0 \otimes K_0 + \sum_{m=1}^{\infty} G_m \otimes K_m. \tag{5.26}$$

The decomposition (5.26) is well–known [78, 50, 79] and is a fundamental characteristic of the stochastic Galerkin method when approximation spaces of the form (5.22) are considered for the finite–dimensional weak problem (5.21).

The sparsity pattern and spectral properties of the matrix A depend on the specific choices of H_1 and P we make; see [51] for a detailed analysis. These choices depend on the problem at hand. The vector $\mathbf{b}$ in (5.23) admits a Kronecker–product structure as well, but this is not important for the remainder of this thesis. Of course, for the matrix A to be used in computations, the infinite sum in (5.26) must be either truncated, or, have a finite number of nonzero terms. In the next section we show the latter is true for certain choices of $P \subset L^2_\pi(\Gamma)$ in (5.22).

5.3.2 Approximation Spaces

To compute SGFEM approximations $u_X \in X$ satisfying (5.21) it remains for us to choose H_1 and P in (5.22). Since D is a square for all four test problems in Section 5.1.1, in this chapter we simply choose either $H_1 = \mathbb{Q}_1(h)$ or $H_1 = \mathbb{Q}_2(h)$, constructed with respect to a uniform mesh of square elements $\mathcal{T}_h$ over D with element width h (recall Examples 4.1 and 4.2).

With the aim of constructing a basis for P comprising of global multivariate polynomials in $\mathbf{y}$ on Γ, we tensorise sets of univariate polynomials in y_m on each Γ_m. To this end, consider

$$J = \big\{\mu = (\mu_1, \mu_2, \dots) \in \mathbb{N}_0^{\mathbb{N}};\ \#\text{supp}(\mu) < \infty\big\}, \tag{5.27}$$

the set of finitely supported multi–indices, where

$$\text{supp}(\mu) := \big\{m \in \mathbb{N};\ \mu_m \neq 0\big\},$$

as well as the families of univariate polynomials

$$\Psi_m = \big\{\psi_{\mu_m}(y_m) : \Gamma_m \to \mathbb{R};\ \mu_m = 0, 1, \dots\big\}, \qquad m = 1, 2, \dots. \tag{5.28}$$

Here, $\psi_{\mu_m}(y_m)$ denotes a univariate polynomial of degree μ_m in the parameter y_m for $m = 1, 2, \dots$ with $\psi_0(y_m) = 1$. Consequently, we can define global polynomials on Γ by selecting a multi–index $\mu \in J$ and constructing

$$\psi_\mu(\mathbf{y}) = \prod_{m=1}^{\infty} \psi_{\mu_m}(y_m) = \prod_{m \in \text{supp}(\mu)} \psi_{\mu_m}(y_m).$$

For given families of polynomials $\{\Psi_m\}_{m=1}^{\infty}$, we simply choose a set of multi–indices $J_P \subset J$ with cardinality $\text{card}(J_P) = s$ to define P, that is

$$P := \text{span}\big\{\psi_\mu(\mathbf{y});\ \mu \in J_P\big\}. \tag{5.29}$$

To ensure that the basis functions of P are orthogonal with respect to the inner–product $\langle\cdot,\cdot\rangle_{L^2_\pi(\Gamma)}$, the families of univariate polynomials Ψ_m are chosen to be orthogonal with respect to the inner–product $\langle\cdot,\cdot\rangle_{L^2_{\pi_m}(\Gamma_m)}$ for $m = 1, 2, \dots$. Indeed, due to the product measure π,

$$\begin{aligned}
\langle\psi_\mu, \psi_\nu\rangle_{L^2_\pi(\Gamma)} &= \int_\Gamma \psi_\mu(\mathbf{y})\psi_\nu(\mathbf{y})\ d\pi(\mathbf{y}) \\
&= \prod_{m=1}^{\infty} \int_{\Gamma_m} \psi_{\mu_m}(y_m)\psi_{\nu_m}(y_m)\ d\pi_m(y_m) \\
&= \prod_{m=1}^{\infty} \delta_{\mu_m\nu_m} \\
&= \delta_{\mu\nu},
\end{aligned} \tag{5.30}$$

for any two multi–indices $\mu, \nu \in J_P$.

We now address the important issue that to compute SGFEM approximations $u_X \in X$ satisfying (5.21), the sums in (5.16) and (5.26) must have a finite number of nonzero terms. It was briefly touched upon that it is not necessary to truncate the expansion (5.5) a priori to achieve this. If we assume that only the first M parameters are active (i.e., we assume that for every $\mu \in J_P$, $\mu_m = 0$ for $m > M$), then, provided (5.4) holds, $B_m(u_X, v) = 0$ for $u_X, v \in X$ and all $m > M$ [18]. In other words, the projection onto $X = H_1 \otimes P$ truncates the sum in (5.16) after M terms. Equivalently, the matrix A in (5.26) becomes

$$A = G_0 \otimes K_0 + \sum_{m=1}^{M} G_m \otimes K_m, \tag{5.31}$$

where, due to (5.30), it follows that $G_0 = I$. Coefficient matrices of this form are too expensive to invert explicitly. To solve the linear system of equations associated with A in (5.31), an iterative approach is required. Many avenues have been explored; see the works [79, 50, 104, 96, 80, 82] for example.

It is well–known that if Γ_m is symmetric about zero and the measure π_m has even density with respect to Lebesgue measure, we can construct families of orthonormal univariate polynomials (5.28) satisfying a three–term recurrence of the form

$$\frac{1}{a^m_{j_m+1}}\psi_{j_m+1}(y_m) = y_m\psi_{j_m}(y_m) - \frac{1}{a^m_{j_m}}\psi_{j_m-1}(y_m), \tag{5.32}$$

for $j_m = 1, 2, \ldots$; see [58, 53] for example. The constants $a^m_{j_m+1}$ and $a^m_{j_m}$ depend on the measure π_m. Using the relationship (5.32) is is easy to show that the matrices $G_m \in \mathbb{R}^{s\times s}$ defined in (5.25) are extremely sparse. Assuming that P contains polynomials in the first M parameters only, then for two basis functions $\psi_j(\mathbf{y})$ and $\psi_r(\mathbf{y})$ of P, or equivalently $\psi_\mu(\mathbf{y})$ and $\psi_\nu(\mathbf{y})$ where $\mu, \nu \in J_P$ are multi–indices, we have

$$\begin{aligned}[G_m]_{\nu\mu} &= \langle y_m\psi_\mu(\mathbf{y}), \psi_\nu(\mathbf{y})\rangle_{L^2_\pi(\Gamma)} \\ &= \langle y_m\psi_{\mu_m}(y_m), \psi_{\nu_m}(y_m)\rangle_{L^2_{\pi_m}(\Gamma_m)} \prod_{s=1,\ s\neq m}^{M} \delta_{\mu_s\nu_s},\end{aligned}$$

and through the recurrence relation (5.32) we find that

$$[G_m]_{\nu\mu} = \left[\frac{1}{a^m_{\mu_m+1}}\delta_{(\mu_m+1)\nu_m} + \frac{1}{a^m_{\mu_m}}\delta_{(\mu_m-1)\nu_m}\right] \prod_{s=1,\ s\neq m}^{M} \delta_{\mu_s\nu_s}.$$

Consequently, $[G_m]_{\nu\mu}$ is nonzero only when the multi–indices μ and ν differ by one in the m^{th} position, and are the same in all other positions. The significance of this is two–fold. Firstly, the matrices G_m for $m = 1, 2, \ldots, M$ have at most two nonzero entries per row and column, regardless of the dimension s of P [70]. Secondly, the entries of G_m can be determined explicitly without performing any numerical integration. When y_m is the image of a uniformly distributed random variable $\xi_m(\omega)$ (with a constant pdf), the families of orthogonal polynomials on Γ_m are Legendre basis polynomials. The following example is similar to examples given in [70, 85].

Example 5.1: Legendre polynomials on Γ_m.

Let $\Gamma_m = [-1, 1]$ be the image of a mean zero and uniformly distributed random variable $\xi_m(\omega) \sim U(-1, 1)$ for $m = 1, 2, \ldots, M$. Then, $d\pi_m(y_m) = \frac{1}{2}dy_m$,

$$\frac{1}{a^m_{\mu_m}} = \frac{\mu_m}{\sqrt{(2\mu_m + 1)(2\mu_m - 1)}},$$

and the recurrence relation (5.32) generates families of Legendre polynomials with $\psi_0(y_m) = 1$ and $\psi_1(y_m) = \sqrt{3}y_m$. Furthermore,

$$[G_m]_{\nu\mu} = \begin{cases} \frac{\mu_m+1}{\sqrt{(2\mu_m+3)(2\mu_m+1)}}, & \mu \overset{m}{=} \nu,\ \mu_m = \nu_m - 1, \\ \frac{\mu_m}{\sqrt{(2\mu_m+1)(2\mu_m-1)}}, & \mu \overset{m}{=} \nu,\ \mu_m = \nu_m + 1, \\ 0, & \text{otherwise}, \end{cases}$$

where $\mu \overset{m}{=} \nu$ denotes that μ and ν differ only in the m^{th} position.

We always assume that $\mu = \mathbf{0}$ is a member of J_P and is listed first in the set. Some choices of J_P lead to standard sets of polynomials. Below, we introduce *tensor–product* polynomials.

Definition 5.1: Tensor–product polynomials.

Consider the multi–indices

$$J_T(M, k) := \big\{\mu \in J;\ \mu_m \le k \text{ for } m = 1, 2, \ldots, M,\ \mu_m = 0 \text{ for } m > M\big\},$$

then P in (5.29) for $J_P = J_T(M, k)$ is the space of tensor–product polynomials of degree $\le k$ in each of $y_1, y_2, \ldots, y_M$, and $s = (k + 1)^M$.

Example 5.2.

Let $J_P = J_T(2,2)$, then, $s = 9$ and

$$J_P = \big\{(0,0),(1,0),(0,1),(1,1),(2,0),(0,2),(2,1),(1,2),(2,2)\big\},$$

where it is understood that for each $\mu \in J_P$, $\mu_m = 0$ for all $m > M = 2$.

Another choice is *complete* polynomials.

Definition 5.2: Complete polynomials.

Consider the multi–indices

$$J_C(M,k) := \big\{\mu \in J;\ |\mu| \le k,\ \mu_m = 0 \text{ for } m > M\big\}, \qquad |\mu| := \sum_{m=1}^{\infty} \mu_m,$$

then P in (5.29) for $J_P = J_C(M,k)$ is the space of complete polynomials of *total* degree $\le k$ in each of $y_1, y_2, \ldots, y_M$, and $s = \frac{(M+k)!}{M!k!}$.

Example 5.3.

Let $J_P = J_C(2,2)$, then, $s = 6$ and

$$J_P = \big\{(0,0),(1,0),(0,1),(1,1),(2,0),(0,2)\big\} \subset J_T(2,2).$$

Again, it is understood that for each $\mu \in J_P$, $\mu_m = 0$ for all $m > 2$.

It is actually possible to construct tensor–product polynomial spaces with doubly orthogonal basis functions that also satisfy

$$\langle y_m \psi_\mu, \psi_\nu \rangle_{L^2_\pi(\Gamma)} = C^m_{\mu_m \nu_m} \delta_{\mu\nu}, \qquad m = 1, 2, \ldots, M,$$

for any $\mu, \nu \in J_T(M,k)$ [7, 51]. Computationally speaking this is very convenient; the resulting matrices G_m in (5.25) are all diagonal and the linear system $A\mathbf{u} = \mathbf{b}$ for A in (5.31) decouples into s many n–dimensional systems associated with the blocks A_{rr} in (5.24) for $r = 1, 2, \ldots, s$. Since, however, the cardinality s of $J_C(M,k)$ is often much smaller than that of $J_T(M,k)$, it is not always clear when tensor–product spaces of doubly orthogonal polynomials of degree $\le k$ in each variable are preferable to spaces of complete polynomials.

A third and less standard choice is *hyperbolic–cross* polynomials.

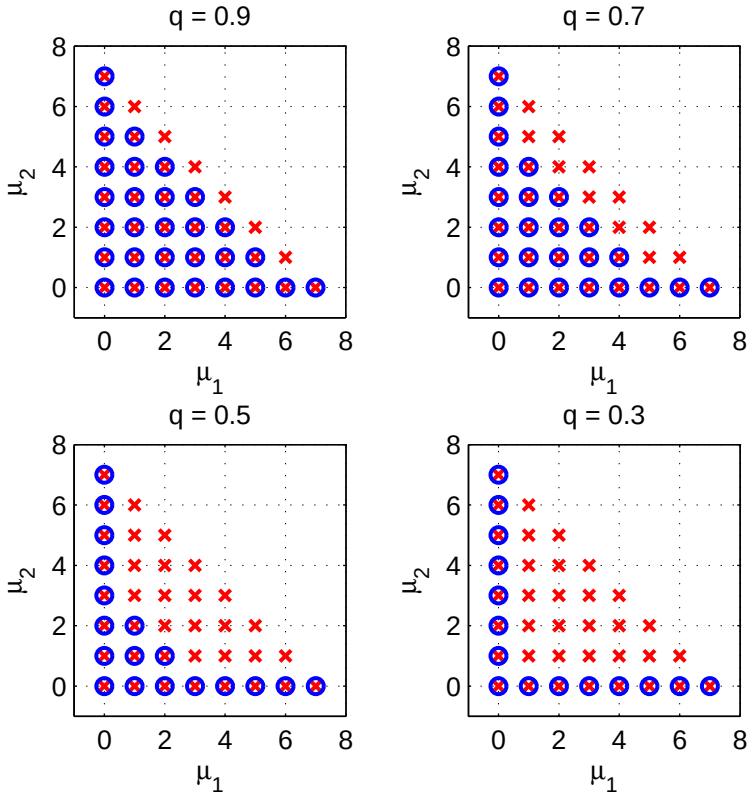


Figure 5.1: Schematic of the multi–index sets $J_C(2,7)$ (red crosses) and $J_H(2,7,q)$ (blue circles) for varying q. Each multi–index is of the form $\mu = (\mu_1, \mu_2, 0, \dots)$ and we plot μ_1 and μ_2.

Definition 5.3: Hyperbolic polynomials.

Consider the multi–indices

$$J_H(M,k,q) := \big\{\mu \in J;\ \|\mu\|_q \le k,\ \mu_m = 0 \text{ for } m > M\big\},$$

where

$$\|\mu\|_q := \Big[\sum_{m=1}^{\infty} \mu_m^q\Big]^{\frac{1}{q}}, \qquad 0 < q \le 1, \tag{5.33}$$

and q is a user–defined parameter. We say that P in (5.29) for $J_P = J_H(M,k,q)$ is the space of q–hyperbolic polynomials in the parameters $y_1, y_2, \dots, y_M$.

The norm $\|\cdot\|_q$ in (5.33) penalises multi–indices $\mu \in J$ with larger supports $\mathrm{supp}(\mu)$ (recall Definition 2.4) and absolute sums $|\mu|$ (the total degree of $\psi_\mu(\mathbf{y})$). Such spaces have been employed in the literature in the construction of polynomial chaos expansions [26, 100] and stochastic collocation methods [13] for PDE problems with random coefficients. Notice that when $q = 1$, $J_H(M,k,q) = J_C(M,k)$. To demonstrate the

influence of (5.33) we provide the following example.

Example 5.4: Complete versus hyperbolic–cross polynomials.

Consider multi–indices $\mu = (\mu_1, \mu_2, 0, \dots)$ in the sets $J_C(2,7)$ and $J_H(2,7,q)$. In Figure 5.1 we plot μ_1 and μ_2 for each multi–index in both sets for $q = 0.3, 0.5, 0.7, 0.9$. Each red cross and blue circle represents a multi–index in $J_C(2,7)$ and $J_H(2,7,q)$ respectively. Note how for each choice of q, the retained multi–indices lie under a hyperbola–like curve.

Mean & Variance of SGFEM Approximations

Once an SGFEM approximation $u_X \in X$ has been computed, we may compute important statistical quantities such as its expectation (2.6) and variance (2.7). Again, we work with the probability space $(\Gamma, \mathcal{B}(\Gamma), \pi)$ instead of the space $(\Omega, \mathcal{F}, \mathbb{P})$. When the basis of P is constructed as described, that is, when $\psi_{\mathbf{0}} = 1$ and $\langle \psi_\mu, \psi_\nu \rangle_{L^2_\pi(\Gamma)} = \delta_{\mu\nu}$ for any two multi–indices $\mu, \nu \in J_P$, $\mathbb{E}[u_X]$ and $\mathrm{Var}(u_X)$ admit neat analytical representations which do not require numerical integration over Γ.

We may express u_X as

$$u_X(\mathbf{x}, \mathbf{y}) = \sum_{\mu \in J_P} u_X^\mu(\mathbf{x}) \psi_\mu(\mathbf{y}), \qquad u_X^\mu(\mathbf{x}) = \sum_{i=1}^{n} u_i^\mu \phi_i(\mathbf{x}), \qquad u_i^\mu \in \mathbb{R},$$

so that

$$\mathbb{E}[u_X] = \int_\Gamma u_X \, d\pi(\mathbf{y}) = \sum_{\mu \in J_P} u_X^\mu(\mathbf{x}) \int_\Gamma \psi_\mu(\mathbf{y}) \times 1 \, d\pi(\mathbf{y}) = u_X^{\mathbf{0}}(\mathbf{x}), \tag{5.34}$$

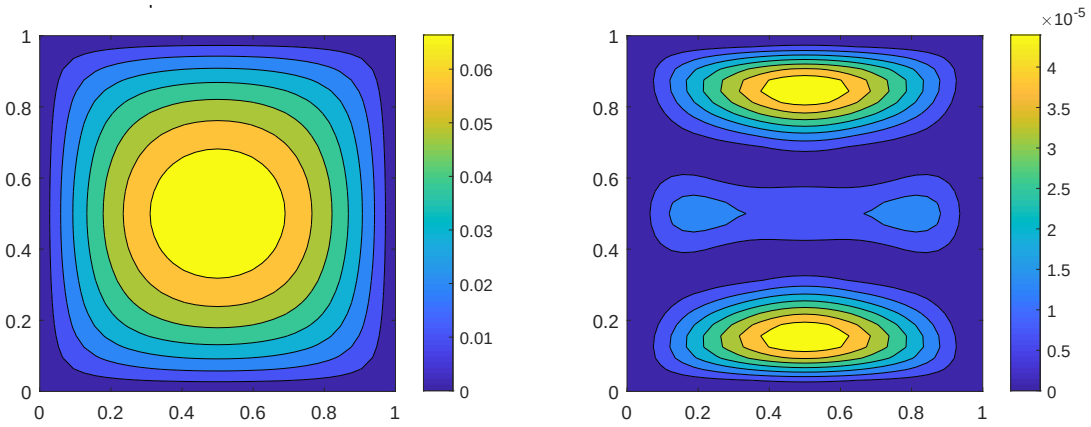
and thus the expectation of u_X is simply the mean mode $u_X^{\mathbf{0}} \in H_0^1(D)$ associated with the multi–index $\mathbf{0} \in J_P$. Likewise, we find that

$$\mathbb{E}[u_X^2] = \int_\Gamma u_X^2 \, d\pi(\mathbf{y}) = \sum_{\mu \in J_P} \sum_{\nu \in J_P} u_X^\mu(\mathbf{x}) u_X^\nu(\mathbf{x}) \int_\Gamma \psi_\mu(\mathbf{y}) \psi_\nu(\mathbf{y}) \, d\pi(\mathbf{y}) = \sum_{\mu \in J_P} u_X^\mu(\mathbf{x})^2,$$

and thus

$$\mathrm{Var}(u_X) = \mathbb{E}[u_X^2] - \mathbb{E}[u_X]^2 = \sum_{\mu \in J_P \setminus \{\mathbf{0}\}} u_X^\mu(\mathbf{x})^2, \tag{5.35}$$

which also resides in $H_0^1(D)$. The above expressions for $\mathbb{E}[u_X]$ and $\mathrm{Var}(u_X)$ are standard and can be found in [70], for example.



(a) expectation (b) variance

Figure 5.2: Expectation and variance of the SGFEM approximation $u_X \in X$ constructed in Example 5.5.

5.3.3 Numerical Experiments

In this section we construct $X = H_1 \otimes P$ in (5.22) by making specific choices of H_1 and P and compute SGFEM approximations $u_X \in X$ by solving (5.21). In Example 5.5 below we consider TP2 in Section 5.1.1 and calculate the associated expectation and variance (5.34) and (5.35).

Example 5.5: Expectation and variance, TP2.

Consider TP2 with $J_P = J_C(5,3)$ and construct $H_1 = \mathbb{Q}_1(h)$ over a uniform 32×32 mesh of square elements $\mathcal{T}_h$ (so that $h = 2^{-5}$). In Figure 5.2 we plot the expectation $\mathbb{E}[u_X]$ and the variance $\mathrm{Var}(u_X)$, where $u_X \in X$ satisfies (5.21).

For our next example, we investigate the accuracy of $u_X \in X$ when $J_P = J_H(M,k,q)$ in the definition of P in (5.29) for certain choices of M, k and q.

Example 5.6.

Consider TP1 with H_1 as in Example 5.5 (i.e., $\mathcal{T}_h$ is 32×32), and compute $u_X \in X$ satisfying (5.21) with $J_P = J_H(M,k,q)$. We fix the polynomial degree $k = 6$ and vary the number of active parameters $M = 2,\dots,7$ for each $q \in \{0.4, 0.7, 1\}$. In Figure 5.3 we plot the energy error $\|u_{\mathrm{ref}} - u_X\|_B = \sqrt{\|u_{\mathrm{ref}}\|_B^2 - \|u_X\|_B^2}$ where u_{ref} is a reference solution obtained using a uniform square mesh of 128×128 elements and the multi–indices $J_P^{\mathrm{ref}} = J_C(9,8)$. We now repeat this process for TP2 and

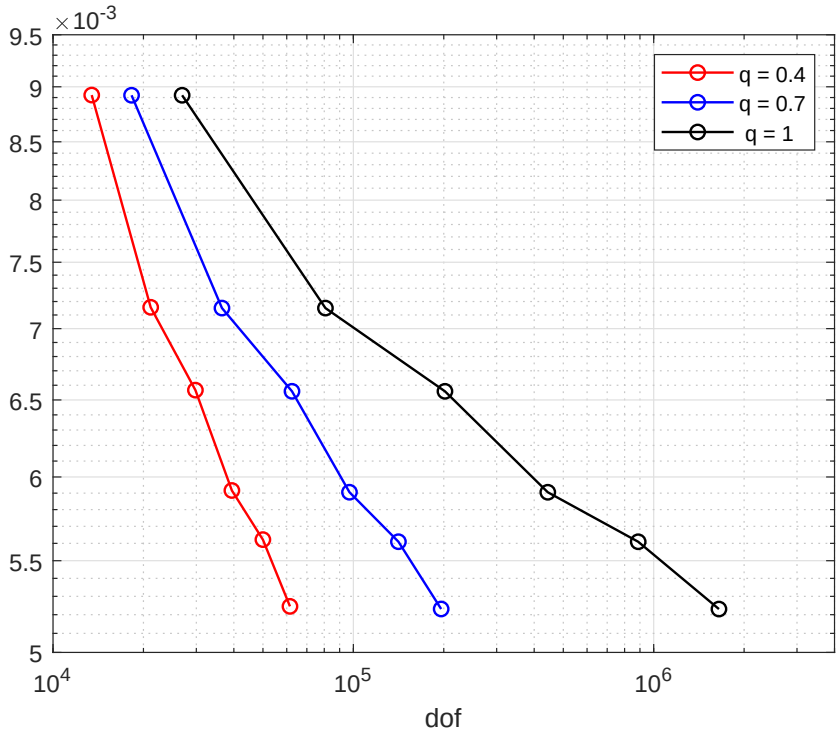


Figure 5.3: The energy errors $\|u_{\text{ref}} - u_X\|_B$ versus the corresponding number of degrees of freedom (dof) N_X for TP1 in Example 5.6. We choose $H_1 = \mathbb{Q}_1(2^{-4})$ and $J_P = J_H(M, k, q)$ in the definition of P in (5.29) with $k = 6$ fixed, and vary $M = 2, \ldots, 7$ for each $q \in \{0.4, 0.7, 1\}$.

plot the corresponding errors in Figure 5.4. Observe from Figures 5.3 and 5.4 that N_X grows much more rapidly as we increase M for larger values of q (most notably when $q = 1$ meaning that $J_P = J_C(M, 6)$).

Recall that the number of active parameters M determines how many terms in the expansion $a(\mathbf{x}, \mathbf{y})$ in (5.5) play a role in the computation of the SGFEM approximation $u_X \in X$ satisfying (5.21). Since the sequence $\{\|a_m\|_\infty\}_{m=1}^\infty$ decays more slowly for TP1 than the sequence for TP2, we expect TP1 to require a larger number of terms M to achieve high levels of accuracy. Indeed, in Figure 5.3 we observe consistent decreases in $\|u_{\text{ref}} - u_X\|_B$ as we increase M, whereas the errors in Figure 5.4 begin to stagnate for $M > 4$, meaning that an increase in the number of active parameters no longer leads to improved accuracy. To further reduce the error once it has stagnated, we need to either increase k, refine $\mathcal{T}_h$, or both.

We also observe in Figure 5.3 that for fixed values of M, the errors for all three values of q are comparable (with k and $\mathcal{T}_h$ fixed). In particular, we achieve the same

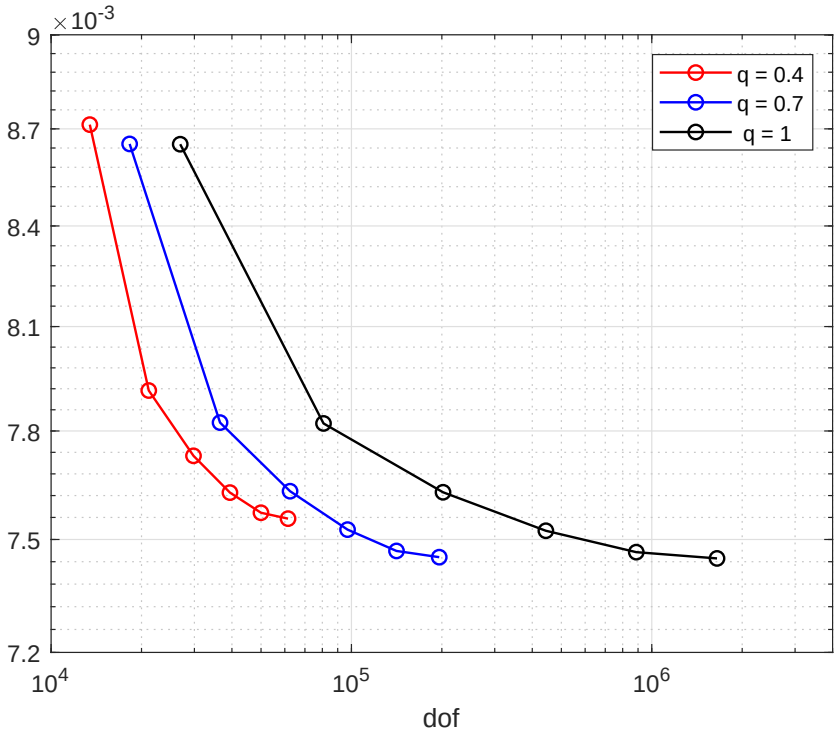


Figure 5.4: The energy errors $\|u_{\text{ref}} - u_X\|_B$ versus the corresponding number of degrees of freedom (dof) N_X for TP2 in Example 5.6. We choose $H_1 = \mathbb{Q}_1(2^{-5})$ and $J_P = J_H(M, k, q)$ in the definition of P in (5.29) with $k = 6$ fixed, and vary $M = 2, \ldots, 7$ for each $q \in \{0.4, 0.7, 1\}$.

accuracy using $q = 0.4$ as we do using $q = 1$, and choosing $J_P = J_H(M, 6, 0.4)$ over $J_P = J_H(M, 6, 1)$ is significantly more efficient since the corresponding value of N_X is much smaller (approximately two orders of magnitude smaller when $M = 7$). Likewise, in Figure 5.4 we observe that significant computational gains can be made by choosing $q = 0.7$ instead of $q = 1$ (for fixed choices of M, k and $\mathcal{T}_h$). Choosing $q = 0.4$ however leads to a drop in accuracy when comparing the errors for fixed values of M, and the computational gains are not as clear to determine.

Whilst the optimal choice of q is problem dependent and generally unknown, Figures 5.3 and 5.4 both demonstrate that, in some situations, significant computational savings can be made by carefully selecting the set of multi–indices $J_P \subset J$ that define $P \subset L^2_\pi(\Gamma)$ in (5.29). Moreover, the error associated with our choice of J_P must be carefully balanced against the error associated with $\mathcal{T}_h$ (the mesh used to construct the FEM space H_1) to prevent the total error from stagnating and avoid doing unnecessary work. In the same way a posteriori error estimation was employed to drive

the adaptive refinement of $\mathcal{T}_h$ in Section 3.3.2, we may exploit a posteriori error estimators to adaptively construct the set J_P *and* refine the mesh $\mathcal{T}_h$. The aim is to select only the most important multi–indices $\mu \in J$ for the set J_P with respect to the energy error and minimise the dimension N_X of $X \subset V = H_0^1(D) \otimes L_\pi^2(\Gamma)$ such that $\|u - u_X\|_B < \texttt{tol}$ for some prescribed error tolerance $\texttt{tol}$. To this end, in the following section we discuss efficient a posteriori error estimation for SGFEMs.

5.4 A Posteriori Error Estimation

We now look to estimate the energy error $\|u - u_X\|_B$ for a given SGFEM approximation $u_X \in X = H_1 \otimes P$ satisfying (5.21). A few explicit strategies for the model problem (5.6)–(5.7) have been proposed in the literature that directly utilise the residual

$$r := f + \nabla \cdot (a \nabla u_X),$$

Specifically, a general framework for residual–based error estimation for SGFEMs for elliptic problems is developed in [44, 45], where overestimation of the true error by up to a factor of 10 is reported for certain test problems. A similar approach is taken in [46] where residuals are computed using an equilibrated fluxes strategy and overestimation upto a factor of 5 is reported for certain test problems. We take an *implicit* approach; by following the theoretical framework outlined in Section 3.2 (provided by [15, 2, 108]) we rederive the estimation strategy developed in [18] and [22] in a way that is convenient for our analysis. In those works, effectivity indices close to one are reported for test problems similar to those considered in [44, 45, 46].

Suppose that we choose a second subspace $W \subset V$ such that $W \supset X$ and solve

$$\text{find } e_W \in W : \qquad B(e_W, v) = F(v) - B(u_X, v), \qquad \text{for all } v \in W,$$

which we previously labelled (3.10), where $B(\cdot,\cdot)$ and $F(\cdot)$ are, here, given in (5.13) and (5.14), respectively. If Assumption 3.1 holds for the chosen spaces X and W, then the bound (3.12) holds. Under Assumption 5.3, so that the norm equivalence (5.19) holds, the bound (3.15) also holds where $e_0 \in W$ satisfies

$$\text{find } e_0 \in W : \qquad B_0(e_0, v) = F(v) - B(u_X, v), \qquad \text{for all } v \in W, \tag{5.36}$$

which we previously labelled (3.13), and, here, $B_0(\cdot,\cdot)$ is given in (5.17).

There are several possible ways to construct W. Following [22], we choose

$$W = X \oplus \Big((H_2 \otimes P) \oplus (H_1 \otimes Q) \Big) =: X \oplus Y, \tag{5.37}$$

where H_2 and Q are to be chosen (recall that $X = H_1 \otimes P$ is already chosen). To ensure that $X \cap Y = \{0\}$, we choose $H_2 \subset H_0^1(D)$ and $Q \subset L^2_\pi(\Gamma)$ such that

$$H_1 \cap H_2 = \{0\}, \qquad P \cap Q = \{0\}, \tag{5.38}$$

which also gives $Y_1 \cap Y_2 = \{0\}$ for

$$Y_1 := H_2 \otimes P, \qquad Y_2 := H_1 \otimes Q. \tag{5.39}$$

In the same way as we define P in (5.29), to construct Q, we choose a subset $J_Q \subset J$ satisfying $J_P \cap J_Q = \emptyset$ and define

$$Q := \operatorname{span}\big\{ \psi_\nu(\mathbf{y});\ \nu \in J_Q \big\}. \tag{5.40}$$

In this way, the spaces P and Q are mutually orthogonal with respect to the inner–product $\langle \cdot, \cdot \rangle_{L^2_\pi(\Gamma)}$, since (5.30) holds for any two multi–indices in J. Furthermore, due to the tensor product structure of Y_1 and Y_2 and the fact that $P \cap Q = \{0\}$, it can be shown that

$$B_0(u, v) = 0, \qquad \text{for all } u \in Y_1, \qquad \text{for all } v \in Y_2. \tag{5.41}$$

To see this, expand $u \in Y_1$ and $v \in Y_2$ in the chosen bases and use (5.30).

Given Y defined as in (5.37), we can then compute the error estimate $\eta := \|e_Y\|_{B_0}$ by solving

$$\text{find } e_Y \in Y: \qquad B_0(e_Y, v) = F(v) - B(u_X, v), \qquad \text{for all } v \in Y, \tag{5.42}$$

previously labelled (3.17), and the bound (3.19) holds. Combining all the previous results yields Theorem 3.4, which, for completeness, we now restate for our parametric diffusion problem.

Theorem 5.1: [22, Theorem 4.1].

Let $u \in V$ satisfy (5.12) *and let the SGFEM approximation $u_X \in X$ satisfy* (5.21) *for $X = H_1 \otimes P$ defined as in* (5.22)*. Define Y and W as in* (5.37)*, choosing H_2 and Q such that* (5.38) *holds, and let $e_Y \in Y$ satisfy* (5.42)*. If Assumptions 3.1*

and 5.2–5.3 hold, then $\eta := \|e_Y\|_{B_0}$ *satisfies*

$$\sqrt{\lambda}\eta \leq \|u - u_X\|_B \leq \frac{\sqrt{\Lambda}}{\sqrt{1-\gamma^2}\sqrt{1-\beta^2}}\eta, \tag{5.43}$$

where λ *and* Λ *are defined in* (5.20)*, and* $\beta, \gamma \in [0, 1)$ *satisfy* (3.11) *and* (3.18)*.*

Decomposition of the Error

We are now tasked to compute $\eta = \|e_Y\|_{B_0}$. Fortunately, due to the structure we impose on Y, this problem can be simplified greatly. Indeed, when Y is chosen as in (5.37), problem (5.42) decouples into two smaller ones on Y_1 and Y_2 (rather than one large problem on Y). By construction, we have that $Y_1 \cap Y_2 = \{0\}$, and so

$$e_Y = e_{Y_1} + e_{Y_2}, \qquad \text{for some } e_{Y_1} \in Y_1,\ e_{Y_2} \in Y_2.$$

Due to the bilinearity of $B_0(\cdot, \cdot)$, and the fact that (5.42) holds for all functions in both Y_1 and Y_2, we may consider the equivalent restatement of problem (5.42): find the component errors $e_{Y_1} \in Y_1$ and $e_{Y_2} \in Y_2$ such that

$$B_0(e_{Y_1}, v_i) + B_0(e_{Y_2}, v_i) = F(v_i) - B(u_X, v_i), \qquad \text{for all } v_i \in Y_i, \qquad i = 1, 2.$$

By (5.41) it follows that $B_0(e_{Y_2}, v_1) = 0$ and $B_0(e_{Y_1}, v_2) = 0$ for all $v_1 \in Y_1$ and $v_2 \in Y_2$, and thus the above problems are reduced to the following two smaller problems;

$$\text{find } e_{Y_1} \in Y_1 : \qquad B_0(e_{Y_1}, v) = F(v) - B(u_X, v), \qquad \text{for all } v \in Y_1, \tag{5.44}$$

$$\text{find } e_{Y_2} \in Y_2 : \qquad B_0(e_{Y_2}, v) = F(v) - B(u_X, v), \qquad \text{for all } v \in Y_2, \tag{5.45}$$

of dimension $\dim(Y_1) = \dim(H_2)\dim(P)$ and $\dim(Y_2) = \dim(H_1)\dim(Q)$, respectively. In addition, since Y_2 admits the decomposition

$$Y_2 = \bigoplus_{\nu \in J_Q} Y_2^\nu, \qquad Y_2^\nu := H_1 \otimes Q^\nu, \qquad Q^\nu := \text{span}\{\psi_\nu(\mathbf{y})\}, \tag{5.46}$$

so that $e_{Y_2} = \sum_{\nu \in J_Q} e_{Y_2}^\nu$ for some functions $e_{Y_2}^\nu \in Y_2^\nu$, a similar orthogonality argument proves that (5.45) decouples into $\text{card}(J_Q)$ many smaller problems:

$$\text{find } e_{Y_2}^\nu \in Y_2^\nu : \qquad B_0(e_{Y_2}^\nu, v) = F(v) - B(u_X, v), \qquad \text{for all } v \in Y_2^\nu, \tag{5.47}$$

of dimension $\dim(H_1)$. Moreover, since $B_0(e_{Y_1}, e_{Y_2}) = 0$ and $B_0(e^{\mu}_{Y_2}, e^{\nu}_{Y_2}) = 0$ for any two multi–indices where $\mu \neq \nu$, we have

$$\eta = \|e_Y\|_{B_0} = \left[\|e_{Y_1}\|^2_{B_0} + \|e_{Y_2}\|^2_{B_0}\right]^{\frac{1}{2}} = \left[\|e_{Y_1}\|^2_{B_0} + \sum_{\nu \in J_Q} \|e^{\nu}_{Y_2}\|^2_{B_0}\right]^{\frac{1}{2}}, \tag{5.48}$$

where $\|e_{Y_1}\|_{B_0}$ estimates the energy error associated with our choice of H_1, and $\|e_{Y_2}\|_{B_0}$ estimates the energy error associated with our choice of P. This is precisely the estimator considered in [22].

As noted in [22], the coefficient matrix associated with problem (5.47) is the same for each $\nu \in J_Q$ (only the right–hand side changes). Consequently, the computation of the estimates $\|e^{\nu}_{Y_2}\|_{B_0}$ may be cheaply vectorised over the set of multi–indices $\nu \in J_Q$. Problem (5.44) also decouples further, into $\mathrm{card}(J_P)$ many problems of dimension $\dim(H_2)$, but this is not essential for now and is addressed in Chapter 6.

The CBS Constant

In [22], the augmented space W in (5.37) is rearranged as

$$W = \Big((H_1 \oplus H_2) \otimes P\Big) \oplus (H_2 \otimes Q).$$

The analysis in that work relies on the orthogonality with respect to $\langle \cdot, \cdot \rangle_{L^2_\pi(\Gamma)}$ of P and Q, rather than Y_1 and Y_2 given in (5.39), and the decoupling of (5.36) into two smaller problems over $(H_1 \oplus H_2) \otimes P$ and $H_2 \otimes Q$. A CBS constant is introduced into the analysis by splitting the former into $H_1 \otimes P$ and $H_2 \otimes P$. The approach we take is subtly different – we introduce a CBS constant by splitting the augmented space W into X and Y, as was done in Section 3.3.2 for the analogous deterministic problem. We demonstrate that the analysis of the constant γ in the bound (5.43) is the same for both approaches.

If Assumption 5.3 holds, then $H^1_0(D)$ is a Hilbert space with respect to the inner–product (4.7) with $a_0(\mathbf{x})$ in place of $a(\mathbf{x})$. In addition, since $H_1 \cap H_2 = \{0\}$, the bound (4.6) holds with H_1 and H_2 in place of X and Y, respectively, which we restate for completeness; there exists a constant $\gamma \in [0, 1)$ such that

$$|\langle u, v \rangle_{a_0}| \leq \gamma \langle u, u \rangle^{\frac{1}{2}}_{a_0} \langle v, v \rangle^{\frac{1}{2}}_{a_0}, \qquad \text{for all } u \in H_1, \qquad \text{for all } v \in H_2 \tag{5.49}$$

(we have swapped the order of H_1 and H_2). In [18] it is shown that any constant γ in (5.49) also satisfies the strengthened Cauchy–Schwarz inequality

$$|B_0(u, v_1)| \leq \gamma \|u\|_{B_0} \|v_1\|_{B_0}, \qquad \text{for all } u \in X, \qquad \text{for all } v_1 \in Y_1 = H_2 \otimes P.$$

Every $v \in Y$ admits the decomposition $v = v_1 + v_2$ for some functions $v_1 \in Y_1$ and $v_2 \in Y_2$. Since Y_1 and Y_2 are mutually orthogonal in terms of (5.41), then

$$\|v\|_{B_0}^2 = \|v_1\|_{B_0}^2 + \|v_2\|_{B_0}^2 \implies \|v_1\|_{B_0} \leq \|v\|_{B_0}.$$

In addition, since P and Q are mutually orthogonal with respect to the inner–product $\langle \cdot, \cdot \rangle_{L^2_\pi(\Gamma)}$,

$$B_0(u, v) = B_0(u, v_1), \qquad \text{for all } u \in X, \qquad \text{for all } v \in Y,$$

and thus, combining the previous results yields

$$|B_0(u, v)| = |B_0(u, v_1)| \leq \gamma \|u\|_{B_0} \|v_1\|_{B_0} \leq \gamma \|u\|_{B_0} \|v\|_{B_0}, \tag{5.50}$$

for all $u \in X$ and $v \in Y$, where γ appears in (5.49). The bound (5.50) coincides with (3.18) – the bound associated with the splitting of W into X and Y – and thus any constant in (5.50) also appears in the error bound (5.43). In other words, the CBS constants $\gamma_{\min}$ associated with the splitting of W into X and Y and $(H_1 \oplus H_2) \otimes P$ into $H_1 \otimes P$ and $H_2 \otimes P$ are *both* bounded above by the CBS constant associated with the finite element spaces H_1 and H_2. Recall that CBS constants associated with (5.49), or equivalently (4.6), are the main focus of Chapter 4 and are computed for a variety of FEM spaces $H_1, H_2 \subset H_0^1(D)$.

Estimated Error Reductions

For a computed SGFEM approximation $u_X \in X$ satisfying (5.21) with X of the form (5.22), as well as an a posteriori error estimator $e_Y \in Y$ satisfying (5.42) with Y of the form (5.37), we now briefly consider strategies for enriching X with the aim of computing enhanced SGFEM approximations. Consider the problems

$$\text{find } u_{W_1} \in W_1 : \qquad B(u_{W_1}, v) = F(v), \qquad \text{for all } v \in W_1, \tag{5.51}$$

$$\text{find } u_{W_2} \in W_2 : \qquad B(u_{W_2}, v) = F(v), \qquad \text{for all } v \in W_2, \tag{5.52}$$

where

$$W_1 = (H_1 \oplus H_2) \otimes P, \qquad W_2 = H_1 \otimes (P \oplus \bar{Q}), \tag{5.53}$$

with $\bar{Q} = \oplus_{\nu \in \bar{J}_Q} Q^\nu$ for some subset $\bar{J}_Q \subseteq J_Q$. That is, W_1 and W_2 represent spaces of the form (5.22) with enriched spatial and parametric components, respectively, and $u_{W_1} \in W_1$ and $u_{W_2} \in W_2$ represent enhanced SGFEM approximations.

Due to Galerkin orthogonality we find that

$$\|e_{W_i}\|_B^2 = \|u - u_X\|_B^2 - \|u_{W_i} - u_X\|_B^2, \qquad e_{W_i} := u - u_{W_i}, \qquad i = 1, 2,$$

where e_{W_1} and e_{W_2} are the errors associated with the augmented spaces W_1 and W_2 in (5.53) which we may rearrange as

$$W_1 = (X \oplus Y_1), \qquad W_2 := (X \oplus \bar{Y}_2), \qquad \bar{Y}_2 = \bigoplus_{\nu \in \bar{J}_Q} Y_2^\nu,$$

(recall the definition of Y_2^ν in (5.46)). In other words, $\|u_{W_1} - u_X\|_B^2$ characterises the reduction in $\|u - u_X\|_B^2$ (the square of the energy error) that would be achieved by augmenting X with Y_1 and computing an enhanced approximation $u_{W_1} \in W_1$ satisfying (5.51). Likewise, $\|u_{W_2} - u_X\|_B^2$ characterises the reduction in $\|u - u_X\|_B^2$ that would be achieved by augmenting X with $\bar{Y}_2$ and computing $u_{W_2} \in W_2$ satisfying (5.52). The following result provides estimates for these quantities, and demonstrates that the constant γ in (5.43) also plays an important role in adaptive SGFEMs. This is a simple extension of a result proved in [18]; the proof is very similar.

Theorem 5.2.

Let $u_X \in X = H_1 \otimes P$ satisfy (5.21), the approximations $u_{W_1} \in W_1$ and $u_{W_2} \in W_2$ satisfy (5.51) and (5.52) where W_1 and W_2 are defined as in (5.53) for some subset $\bar{J}_Q \subseteq J_Q$, and define $e_{\bar{Y}_2} := \sum_{\nu \in \bar{J}_Q} e_{Y_2}^\nu$ so that $\|e_{\bar{Y}_2}\|_{B_0}^2 = \sum_{\nu \in \bar{J}_Q} \|e_{Y_2}^\nu\|_{B_0}^2$. Then, the following estimates hold:

$$\lambda \|e_{Y_1}\|_{B_0}^2 \leq \|u_{W_1} - u_X\|_B^2 \leq \frac{\Lambda}{1 - \gamma^2} \|e_{Y_1}\|_{B_0}^2, \tag{5.54}$$

$$\lambda \|e_{\bar{Y}_2}\|_{B_0}^2 \leq \|u_{W_2} - u_X\|_B^2 \leq \Lambda \|e_{\bar{Y}_2}\|_{B_0}^2, \tag{5.55}$$

where λ, Λ are defined in (5.19), and $\gamma \in [0, 1)$ satisfies (3.18).

When the constant γ appearing in the bound (5.54) is small, we have more confidence that the estimates $\|e_{Y_1}\|_{B_0}$ and $\|\bar{e}_{Y_2}\|_{B_0}$ for $\|u_{W_1} - u_X\|_B$ and $\|u_{W_2} - u_X\|_B$ are accurate. Once a suitable set $\bar{J}_Q \subseteq J_Q$ has been determined, (5.54) and (5.55) provide theoretical foundations for the design of adaptive SGFEMs, which is the focus of Chapter 6. For a simple example, when $\bar{J}_Q = J_Q$ so that $\bar{Y}_2 = Y_2$, $\|e_{Y_1}\|_{B_0}$ and $\|e_{Y_2}\|_{B_0}$ provide estimates of the error reductions that would be achieved by augmenting X with $Y_1 = (H_2 \otimes P)$ or $Y_2 = (H_1 \otimes Q)$, respectively. Then, a simple enrichment strategy is to perform $X \to X \oplus Y_i$ for $i = \text{argmax}_j\{\|e_{Y_j}\|_{B_0};\ j = 1, 2\}$ and compute a new SGFEM approximation $u_X \in X$ satisfying (5.21).

5.4.1 Numerical Experiments

In this section we investigate the accuracy of the error estimate η defined in (5.48). We begin by making specific choices of $X = H_1 \otimes P$ in (5.22) and computing $u_X \in X$ satisfying (5.21). Next, choosing $H_2 \subset H_0^1(D)$ and $Q \subset L^2_\pi(\Gamma)$ so that the conditions (5.38) are satisfied, we solve problems (5.44) and (5.47) and compute η in (5.48). The dimension of the tensor product space $Y_1 = H_2 \otimes P$ associated with problem (5.44) can become unwieldy, especially if the dimension of P is high (which is fixed at the point we wish to estimate $\|u - u_X\|_B$). To keep costs reasonable, we insist on the elementwise decomposition $H_2 = \bigoplus_{\square_k \in \mathcal{T}_h} H_{2k}$ where every m_k–dimensional space $H_{2k} \subset H_0^1(D)$ contains functions with only compact support on $\square_k$, and employ the element residual method outlined in Section 3.3.2. Following [18], for each $\square_k \in \mathcal{T}_h$ we solve the local $m_k s$–dimensional problem: find $e_{Y_{1k}} \in Y_{1k} := H_{2k} \otimes P$ satisfying

$$\begin{aligned} B_{0k}(e_{Y_{1k}}, v) = F_k(v) + \int_\Gamma \langle \nabla \cdot (a(\cdot,\mathbf{y})\nabla u_X(\cdot,\mathbf{y})), v(\cdot,\mathbf{y})\rangle_{L^2(\square_k)}\, d\pi(\mathbf{y}) \\ - \sum_{E \in \mathcal{E}_k} \int_\Gamma \langle a(\cdot,\mathbf{y}) \left[\!\left[\tfrac{\partial u_X}{\partial n} \right]\!\right](\mathbf{y}), v(\cdot,\mathbf{y})\rangle_{L^2(E)}\, d\pi(\mathbf{y}), \end{aligned} \tag{5.56}$$

for all $v \in Y_{1k}$, where $\mathcal{E}_k$ is the set of edges of $\square_k$ excluding those on ∂D, $[\![\frac{\partial u_X}{\partial n}]\!](\mathbf{y})$ is given in (3.34) and

$$\begin{aligned} B_{0k}(u, v) &= \int_\Gamma \int_{\square_k} a_0(\mathbf{x})\nabla u(\mathbf{x},\mathbf{y}) \cdot \nabla v(\mathbf{x},\mathbf{y})\, d\mathbf{x}\, d\pi(\mathbf{y}), \\ F_k(v) &= \int_\Gamma \int_{\square_k} f(\mathbf{x}) v(\mathbf{x},\mathbf{y})\, d\mathbf{x}\, d\pi(\mathbf{y}). \end{aligned} \tag{5.57}$$

We then evaluate the spatial error estimate

$$\|e_{Y_1}\|_{B_0}^2 = \sum_{\square_k \in \mathcal{T}_h} \|e_{Y_{1k}}\|_{B_{0k}}^2 := \sum_{\square_k \in \mathcal{T}_h} B_{0k}(e_{Y_{1k}}, e_{Y_{1k}}), \tag{5.58}$$

in order to evaluate η defined in (5.48).

The spaces H_1 and H_2 in this section play the role of X and Y in Chapter 4 (in that both pairings are disjoint subspaces of $H_0^1(D)$ and relate to an error estimation problem). To keep notation consistent for the SGFEM problem, we now refer to X and Y in Chapter 4 as H_1 and H_2. Our construction of H_2 is identical to the construction (4.10). When H_1 and H_2 coincide with the choices made in Examples 4.1 and 4.2, and $a_0(\mathbf{x})$ in (5.5) and (5.57) is a constant, the values reported in Tables 4.1 and 4.2 are the squares of the CBS constants $\gamma_{\min}$ associated with (5.49), and hence appear in the bounds (5.43) and (5.54). All eight choices of H_2 considered in Examples 4.1 and 4.2 (four per choice of H_1) lead to CBS constants away from one. Thus, they are sensible candidates with which to define the space $Y_1 = H_2 \otimes P$ in the error problem (5.44) (since term $1/\sqrt{1-\gamma_{\min}^2}$ in the bound (5.43) doesn't blow up). Building on Chapter 4, we now investigate which choices of H_2 and Q lead to the best estimates $\eta = \|e_Y\|_{B_0}$ satisfying (5.43) (resulting in effectivity indices close to one). The following results were first published in [37].

We begin by choosing the SGFEM space $X = H_1 \otimes P$. We choose either $H_1 = \mathbb{Q}_1(h)$ or $H_1 = \mathbb{Q}_2(h)$, constructed with respect to uniform meshes of square elements $\mathcal{T}_h$ with element width h, and set

$$J_P = J_C(M, k), \tag{5.59}$$

so that P in (5.29) represents the space of complete polynomials on Γ of total degree $\leq k$ in each $y_1, y_2, \ldots, y_M$. For the error estimate $\eta = \|e_Y\|_{B_0}$ satisfying (5.43), we fix

$$J_Q = J_C(M+1, k+1) \backslash J_P \tag{5.60}$$

(for now), where $J_C(M+1, k+1)$ is the set of multi–indices associated with complete polynomials of total degree $\leq k+1$ in the first $M+1$ parameters, and consider the different choices of H_2 described in Examples 4.1 and 4.2. In order to test the accuracy of $\eta = \|e_Y\|_{B_0}$, we also compute reference solutions $u_{\text{ref}} \in X_{\text{ref}}$ where the space X_{ref} is constructed with $H_1^{\text{ref}} = \mathbb{Q}_2(h)$ on a uniform mesh of square elements with element

Table 5.1: Effectivity indices $\theta_{\text{eff}}^{\text{approx}}$ for TP1 in Example 5.7. We fix $H_1 = \mathbb{Q}_1(h)$, J_P and J_Q as in (5.59) and (5.60) with $M = 5$ and make four choices of H_2 for varying h (with k fixed) and varying k (with h fixed).

h	$\|u_{\text{ref}} - u_X\|_B$	$\mathbb{Q}_2(h)$	$\mathbb{Q}_4(h)$	$\mathbb{Q}_1(h/2)$	$\mathbb{Q}_2(h/2)$
2^{-1}	3.8125×10^{-2}	0.98	0.12	0.87	1.04
2^{-2}	1.9254×10^{-2}	0.97	0.16	0.87	1.04
2^{-3}	1.0244×10^{-2}	0.93	0.22	0.84	0.99
2^{-4}	6.2277×10^{-3}	0.80	0.31	0.73	0.85
2^{-5}	4.7187×10^{-3}	0.62	0.39	0.59	0.65
k	$\|u_{\text{ref}} - u_X\|_B$	$\mathbb{Q}_2(h)$	$\mathbb{Q}_4(h)$	$\mathbb{Q}_1(h/2)$	$\mathbb{Q}_2(h/2)$
2	1.0251×10^{-2}	0.93	0.22	0.84	0.99
3	1.0244×10^{-2}	0.93	0.22	0.84	0.99
4	1.0244×10^{-2}	0.93	0.22	0.84	0.99
5	1.0244×10^{-2}	0.93	0.22	0.84	0.99
6	1.0244×10^{-2}	0.93	0.22	0.84	0.99

width $h = 2^{-7}$ and the set $J_P^{\text{ref}} = J_C(10, 8)$. We then evaluate the effectivity index

$$\theta_{\text{eff}}^{\text{approx}} := \frac{\|e_Y\|_{B_0}}{\sqrt{\|u_{\text{ref}}\|_B^2 - \|u_X\|_B^2}}.$$

We first consider test problem TP1 in Section 5.1.1, which was initially investigated in [18]. In that work, the authors truncate the coefficient $a(\mathbf{x}, \mathbf{y})$ in (5.9) a priori after M terms. Consequently, the parametric problem (5.6)–(5.7) is instead posed on the finite–dimensional domain $D \times \bar{\Gamma}$, where $\bar{\Gamma} = \Gamma_1 \times \Gamma_2 \times \cdots \times \Gamma_M$, and the approximations $u_X \in X$ and $u_{\text{ref}} \in X_{\text{ref}}$ are both functions of only M parameters. In turn, they choose $J_Q = J_C(M, k+1) \backslash J_P$ in the definition of Q. In our experiments, $u \in V$ is a function of infinitely many parameters and u_{ref} is a function of $M_{\text{ref}} > M$ parameters. This subtle (yet crucial) variation is reflected in our choice of J_Q in (5.60). We extend the results provided in [18] by adapting the same error estimation strategy for the more complicated infinite–dimensional problem.

Example 5.7: TP1, $H_1 = \mathbb{Q}_1(h)$.

Consider TP1 with $J_P = J_C(5, k)$ and $H_1 = \mathbb{Q}_1(h)$. Initially, we compute $u_X \in X$ satisfying (5.21) for varying h with fixed $k = 4$, and varying k with fixed $h = 2^{-3}$. For each $u_X \in X$ we compute the error estimator $e_Y \in Y$ by solving (5.45) and (5.56) (on each $\square_k \in \mathcal{T}_h$) for J_Q defined in (5.60) and the four choices of H_2 described in Example 4.1. In Table 5.1 we record the effectivity indices $\theta_{\text{eff}}^{\text{approx}}$.

When $H_1 = \mathbb{Q}_1(h)$, we observe from Table 5.1 that $H_2 = \mathbb{Q}_2(h)$ and $H_2 = \mathbb{Q}_2(h/2)$ define very good estimators. We also observe that $H_2 = \mathbb{Q}_4(h)$ defines a poor estimator,

Table 5.2: Effectivity indices $\theta_{\text{eff}}^{\text{approx}}$ for TP1 in Example 5.8. We fix $H_1 = \mathbb{Q}_2(h)$, J_P and J_Q as in (5.59) and (5.60) with $M = 5$ and make four choices of H_2 for varying h (with k fixed) and varying k (with h fixed).

h	$\|u_{\text{ref}} - u_X\|_B$	$\mathbb{Q}_4(h)$	$\mathbb{Q}_2(h/2)$	$\mathbb{Q}_4^r(h)$	$\mathbb{Q}_2^r(h/2)$
2^{-0}	4.8622×10^{-3}	0.66	0.66	0.55	0.59
2^{-1}	4.1729×10^{-3}	0.48	0.49	0.46	0.47
2^{-2}	4.1003×10^{-3}	0.44	0.45	0.44	0.44
2^{-3}	4.0945×10^{-3}	0.44	0.44	0.44	0.44
2^{-4}	4.0941×10^{-3}	0.44	0.44	0.44	0.44
k	$\|u_{\text{ref}} - u_X\|_B$	$\mathbb{Q}_4(h)$	$\mathbb{Q}_2(h/2)$	$\mathbb{Q}_4^r(h)$	$\mathbb{Q}_2^r(h/2)$
2	4.1134×10^{-3}	0.45	0.45	0.45	0.45
3	4.0950×10^{-3}	0.44	0.44	0.44	0.44
4	4.0945×10^{-3}	0.44	0.44	0.44	0.44
5	4.0945×10^{-3}	0.44	0.44	0.44	0.44
6	4.0945×10^{-3}	0.44	0.44	0.44	0.44

despite the fact that its associated CBS constant is the smallest ($\gamma_{\min}^2 \leq 0.0121$) and the term $\sqrt{1-\gamma_{\min}^2}$ is closest to one. For all four estimates, the constants λ and Λ in the bound (5.43) are fixed, suggesting that when $H_2 = \mathbb{Q}_4(h)$, the saturation constant $\beta_{\min}$ is close to one, or equivalently, the enriched space $X \to X \oplus Y = W$ leads to little improvement in accuracy (recall Assumption 3.1). The same conclusion is drawn by considering the estimated error reductions $\|e_{Y_1}\|_{B_0}$ and $\|e_{Y_2}\|_{B_0}$ in Theorem 5.2 (with $\bar{J}_Q = J_Q$), which are accurate since $\gamma_{\min}$ is small. Since η is small in comparison to $\|u_{\text{ref}} - u_X\|_B$, both $\|e_{Y_1}\|_{B_0}$ and $\|e_{Y_2}\|_{B_0}$ in (5.48) are as well, and thus the true error reductions are small and $\beta_{\min}$ is close to one.

Example 5.8: TP1, $H_1 = \mathbb{Q}_2(h)$.

We now repeat the experiment conducted in Example 5.7 with $H_1 = \mathbb{Q}_2(h)$ and the four choices of H_2 described in Example 4.2 (recall that the sets of multi–indices J_P and J_Q that define P and Q are fixed in (5.59) and (5.60) with $M = 5$), and record the effectivity indices in Table 5.2.

When $H_1 = \mathbb{Q}_2(h)$, we observe from Table 5.2 that for all four choices of H_2, the true error stagnates and the error estimates are poor as we vary *both* h and k. The quality of each SGFEM approximation $u_X \in X$ depends on our choices of M, k and h (through our choices of P and H_1), and thus, to compute improved approximations, we must increase M. We observed a similar behaviour in Table 5.1, where the true errors also stagnated as we increased k. In that experiment, the spatial errors are significant enough for h–refinements to lead to improved SGFEM approximations. By

Table 5.3: Effectivity indices $\theta_{\text{eff}}^{\text{approx}}$ for TP1 in Example 5.9. We fix $H_1 = \mathbb{Q}_2(h)$, J_P and J_Q as in (5.59) and (5.61) (modified choice) with $M = 5$ and make four choices of H_2 for varying h (with k fixed) and varying k (with h fixed).

h	$\|u_{\text{ref}} - u_X\|_B$	$\mathbb{Q}_4(h)$	$\mathbb{Q}_2(h/2)$	$\mathbb{Q}_4^r(h)$	$\mathbb{Q}_2^r(h/2)$
2^{-0}	4.8622×10^{-3}	0.71	0.71	0.61	0.64
2^{-1}	4.1729×10^{-3}	0.83	0.83	0.81	0.82
2^{-2}	4.1003×10^{-3}	0.81	0.82	0.81	0.81
2^{-3}	4.0945×10^{-3}	0.81	0.81	0.81	0.81
2^{-4}	4.0941×10^{-3}	0.81	0.81	0.81	0.81
k	$\|u_{\text{ref}} - u_X\|_B$	$\mathbb{Q}_4(h)$	$\mathbb{Q}_2(h/2)$	$\mathbb{Q}_4^r(h)$	$\mathbb{Q}_2^r(h/2)$
2	4.1134×10^{-3}	0.82	0.82	0.82	0.82
3	4.0950×10^{-3}	0.81	0.81	0.81	0.81
4	4.0945×10^{-3}	0.81	0.81	0.81	0.81
5	4.0945×10^{-3}	0.81	0.81	0.81	0.81
6	4.0945×10^{-3}	0.81	0.81	0.81	0.81

choosing $H_1 = \mathbb{Q}_2(h)$ in this experiment, the spatial errors are vastly reduced and the total error stagnates with *both* k and h. We now address the reason for the poor estimated errors.

For the estimate $\eta = \|e_Y\|_{B_0}$ to be accurate, the space Y in (5.37) must comprise of functions that would substantially improve the current SGFEM approximation (to ensure that $\beta_{\min}$ is small). We noted previously that the approximation error is dominated by our choice of M in that more parameters y_m are required in the definition of P to further reduce the energy errors in Table 5.2 (the errors stagnate with both h and k). Due to our choice of J_Q in (5.60), the space Q contains polynomials in only one additional parameter. Since the sequence $\{\|a_m\|_\infty\}_{m=1}^\infty$ associated with $a(\mathbf{x}, \mathbf{y})$ in (5.9) for TP1 decays very slowly, we expect that more than one additional parameter is needed in the definition of Q in order for Y to contain functions that lead to substantial reductions in the current approximation error. To this end, we investigate in Example 5.9 below the performance of the estimate $\eta = \|e_Y\|_{B_0}$ for the modified space Q associated with the set of multi–indices

$$J_Q = J_C(M + 3, k + 1) \backslash J_P. \tag{5.61}$$

Example 5.9: TP1, $H_1 = \mathbb{Q}_2(h)$ with modified Q.

We rerun the experiment conducted in Example 5.8 for the set J_Q in (5.61) that defines Q in (5.40), and record the updated effectivity indices in Table 5.3.

We observe in Table 5.3 that the effectivity indices $\theta_{\text{eff}}^{\text{approx}}$ corresponding to the

Table 5.4: Effectivity indices $\theta_{\text{eff}}^{\text{approx}}$ for TP2 in Example 5.10. We fix $H_1 = \mathbb{Q}_1(h)$, J_P and J_Q as in (5.59) and (5.60) with $M = 5$ and make four choices of H_2 for varying h (with k fixed) and varying k (with h fixed).

h	$\|u_{\text{ref}} - u_X\|_B$	$\mathbb{Q}_2(h)$	$\mathbb{Q}_4(h)$	$\mathbb{Q}_1(h/2)$	$\mathbb{Q}_2(h/2)$
2^{-2}	6.0892×10^{-2}	0.96	0.10	1.32	0.92
2^{-3}	3.0684×10^{-2}	0.95	0.12	1.32	0.93
2^{-4}	1.5386×10^{-2}	0.95	0.14	1.32	0.94
2^{-5}	7.7745×10^{-3}	0.95	0.18	1.32	0.93
2^{-6}	4.0408×10^{-3}	0.93	0.26	1.28	0.92
k	$\|u_{\text{ref}} - u_X\|_B$	$\mathbb{Q}_2(h)$	$\mathbb{Q}_4(h)$	$\mathbb{Q}_1(h/2)$	$\mathbb{Q}_2(h/2)$
2	3.0723×10^{-2}	0.95	0.13	1.31	0.93
3	3.0686×10^{-2}	0.95	0.12	1.32	0.93
4	3.0684×10^{-2}	0.95	0.12	1.32	0.93
5	3.0684×10^{-2}	0.95	0.12	1.32	0.93
6	3.0684×10^{-2}	0.95	0.12	1.32	0.93

modified choice of J_Q are much improved (in comparison to those recorded in Table 5.2). The estimate η still underestimates $\|u - u_X\|_B$ however, and thus a large number of additional parameters must be incorporated into the definition of J_Q to ensure that $\theta_{\text{eff}}^{\text{approx}}$ is close to one. Additionally, we observe little difference between the accuracy of η for all four choices of H_2. To better compare the four spaces, we now consider test problem TP2 in Section 5.1.1. The sequence $\{\|a_m\|_\infty\}_{m=1}^\infty$ associated with $a(\mathbf{x}, \mathbf{y})$ in (5.10) for TP2 decays more quickly than that for TP1. As a result, the error associated with our choice of M is less dominant (recall that the errors stagnated in Figure 5.4 as M was increased).

Example 5.10: TP2, $H_1 = \mathbb{Q}_1(h)$ and $H_1 = \mathbb{Q}_2(h)$.

We rerun the experiments conducted in Examples 5.7 and 5.8 with J_Q in (5.60) for TP2, and record the effectivity indices in Tables 5.4 and 5.5, respectively.

We observe from Table 5.4 (when $H_1 = \mathbb{Q}_1(h)$) that the space $H_2 = \mathbb{Q}_2(h)$ yields the best estimator, very closely followed by $H_2 = \mathbb{Q}_2(h/2)$. From Table 5.5 (when $H_1 = \mathbb{Q}_2(h)$) we observe that the space $H_2 = \mathbb{Q}_4(h)$ yields the best estimator, closely followed by $H_2 = \mathbb{Q}_2^r(h/2)$ (recall that $H_2 = \mathbb{Q}_4^r(h/2)$ yields the smallest CBS constant). Note that for these experiments, we employ the original definition of Q associated with the set J_Q in (5.60), for which one additional parameter is activated. Consequently, when the sequence $\{\|a_m\|_\infty\}_{m=1}^\infty$ associated with expansions $a(\mathbf{x}, \mathbf{y})$ of the form (5.5) decays quickly enough, only a small number of additional active parameters are needed in the definition of Q to ensure that the effectivity indices $\theta_{\text{eff}}^{\text{approx}}$ are close to one. When the

Table 5.5: Effectivity indices $\theta_{\text{eff}}^{\text{approx}}$ for TP2 in Example 5.10. We fix $H_1 = \mathbb{Q}_2(h)$, J_P and J_Q as in (5.59) and (5.60) with $M = 5$ and make four choices of H_2 for varying h (with k fixed) and varying k (with h fixed).

h	$\|u_{\text{ref}} - u_X\|_B$	$\mathbb{Q}_4(h)$	$\mathbb{Q}_2(h/2)$	$\mathbb{Q}_4^r(h)$	$\mathbb{Q}_2^r(h/2)$
2^{-1}	2.7738×10^{-2}	0.94	0.92	0.66	0.76
2^{-2}	8.3254×10^{-3}	1.08	1.13	0.77	0.90
2^{-3}	2.4871×10^{-3}	1.10	1.17	0.82	0.95
2^{-4}	1.4017×10^{-3}	0.82	0.85	0.73	0.77
2^{-5}	1.2813×10^{-3}	0.71	0.71	0.70	0.70
k	$\|u_{\text{ref}} - u_X\|_B$	$\mathbb{Q}_4(h)$	$\mathbb{Q}_2(h/2)$	$\mathbb{Q}_4^r(h)$	$\mathbb{Q}_2^r(h/2)$
2	3.1896×10^{-3}	1.03	1.08	0.86	0.93
3	2.5511×10^{-3}	1.10	1.16	0.83	0.95
4	2.4871×10^{-3}	1.10	1.17	0.82	0.95
5	2.4805×10^{-3}	1.10	1.18	0.82	0.95
6	2.4798×10^{-3}	1.10	1.18	0.82	0.95

sequence decays too slowly, as for TP1 in Examples 5.7–5.8, extra care must be taken when designing Q. Since publishing these results in [37], the choice $H_2 = \mathbb{Q}_2^r(h/2)$, as well as J_Q in (5.61), has been incorporated into the software package S-IFISS [19].

5.4.2 Adaptive Single–level SGFEMs

In this section, we provide details of how the error estimator $e_Y \in Y$ satisfying (5.42) may be exploited to drive the *adaptive* enrichment of $X = H_1 \otimes P$ in (5.22) associated with (5.21). Adaptive methods for PDE problems with inputs of the form (5.5) are essential. Suppose that we truncate $a(\mathbf{x}, \mathbf{y})$ after M terms and consider the parametric problem (5.6)–(5.7) on the finite dimensional domain $D \times \bar{\Gamma}$ where $\bar{\Gamma} = \Gamma_1 \times \cdots \times \Gamma_M$. Assume that D is convex, $f(\mathbf{x}) \in L^2(D)$ and $a(\mathbf{x}, \mathbf{y}) \in L^\infty(\bar{\Gamma}, W^{1,\infty}(D))$. When H_1 is the $\mathbb{Q}_1$ or $\mathbb{P}_1$ finite element space associated with a regular mesh $\mathcal{T}_h$ of rectangular or triangular elements with maximum element edge length h and $J_P = J_T(M, k)$ in the definition of P in (5.29), it is shown in Proposition 5.1 of [7] that the energy error $\|u - u_X\|_B$ (for the truncated problem) satisfies

$$\|u - u_X\|_B \leq c_1 h + c_2 \sum_{m=1}^{M} r_m^{k+1}, \tag{5.62}$$

where we assume that $r_m := \|a_m/a\|_\infty < 1$. Here, $c_1, c_2 > 0$ depend on a, f and D but are independent of h and k. Note that $r_m^{k+1} = e^{-c_m(k+1)}$ for $m = 1, 2, \ldots, M$, where $c_m = -\log(r_m) > 0$, and thus

$$\|u - u_X\|_B = \mathcal{O}(h) + \mathcal{O}(M e^{-\min_m\{c_m\}(k+1)}).$$

Clearly, the rate of convergence of $\|u-u_X\|_B$ deteriorates as we increase M. In other words, standard SGFEMs of the type described in this chapter suffer from the curse of dimensionality and are highly inefficient when employed for the infinite–dimensional problem (5.6)–(5.7), where M is not fixed a priori or is particularly large. In order for SGFEMs to be a practical tool for tackling high–dimensional parametric PDE problems, the convergence rates must be independent of the number of input parameters. Several works establish the existence of sequences $\{X\}$ (not necessarily of the form (5.22)) such that the energy error for the infinite problem (5.12) decays to zero at an algebraic rate s independent of the number of active parameters as $N_X = \dim(X) \to \infty$; see the works [23, 35, 36, 57] for early seminal results as well as [33, 12, 11].

Notice that $X = H_1 \otimes P$ in (5.22) decomposes as

$$X = \bigoplus_{\mu \in J_P} H_1 \otimes P^\mu, \qquad P^\mu = \operatorname{span}\{\psi_\mu(\mathbf{y})\}. \tag{5.63}$$

This approach can be considered as assigning the same FEM space H_1 to each $\mu \in J_P$, and thus we refer to the tensor–product structure in (5.22) as the *single–level* structure. The design of adaptive single–level SGFEM algorithms is the main focus of the works [22, 21]. Both articles present algorithms which successfully construct sequences of spaces $\{X\}$ such that the error decays at a rate independent of the number of input parameters. In the next section, we test the performance of an algorithm from [22].

A Simple Adaptive Algorithm

The estimated error reductions $\|e_{Y_1}\|_{B_0}$ and $\|e_{\bar{Y}_2}\|_{B_0}$ in Theorem 5.2 may be exploited to design simple adaptive SGFEM algorithms. Suppose that $u_X \in X$ satisfying (5.21) with X of the form (5.22) is computed, where $H_1 = \mathbb{Q}_1(h)$ is constructed with respect to a uniform mesh of square elements $\mathcal{T}_h$ with element width h, and P is defined in (5.29) for some subset $J_P \subset J$. Suppose as well that the spatial estimator $e_{Y_1} \in Y_1$ and the parametric estimators $e^\nu_{Y_2} \in Y_2^\nu$ satisfying (5.44) and (5.47), respectively, are computed for some space H_2 and subset J_Q satisfying $H_1 \cap H_2 = \{0\}$ and $J_P \cap J_Q = \emptyset$. Recall that $\|e_{Y_1}\|_{B_0}$ and $\|e_{\bar{Y}_2}\|_{B_0}$ provide accurate estimates of the error reduction that would be achieved by performing the enrichments $X \to W_1, W_2$, where W_1 and W_2 are given in (5.53), and computing enhanced approximations satisfying (5.51) or (5.52), respectively. For suitable choices of $\bar{J}_Q \subseteq J_Q$, the enrichment strategy associated with

$\max\{\|e_{Y_1}\|_{B_0}, \|e_{\bar{Y}_2}\|_{B_0}\}$ can lead to cost–effective reductions of $\|u - u_X\|_B$.

Following [22], we choose

$$\bar{J}_Q = \big\{\nu \in J_Q;\ \|e^{\nu}_{Y_2}\|_{B_0} \geq \|e_{Y_1}\|_{B_0}\big\}.$$

If $\|e_{Y_1}\|_{B_0} > \|e_{\bar{Y}_2}\|_{B_0}$, the polynomial space P goes unchanged and a uniform refinement on the spatial domain is performed, i.e., $h \to \frac{h}{2}$ so that $H_1 \to \mathbb{Q}_1(\frac{h}{2})$. Otherwise, additional polynomials on Γ are added to P by setting $J_P \to J_P \cup \bar{J}_Q$ in the definition (5.29). As in Section 3.3.2, the iterative process

$$\texttt{SOLVE} \ \to\ \texttt{ESTIMATE} \ \to\ \texttt{MARK} \ \to\ \texttt{REFINE}, \tag{5.64}$$

is repeated until $\|e_Y\|_{B_0} < \texttt{tol}$, where `tol` denotes a user–defined error tolerance. In this loop, the module `MARK` determines the set of multi-indices $\bar{J}_Q$ and `REFINE` executes the chosen refinement strategy (spatial or parametric).

It was proven in [22] that, due to recurrence relation (5.32), $\|e^{\nu}_{Y_2}\|_{B_0} = 0$ for considerably many multi–indices $\nu \in J\backslash J_P$. To avoid unnecessary work when computing $\|e_Y\|_{B_0}$, it is essential that we identify the set of multi–indices $J^* \subset J$ that result in nonzero contributions $\|e^{\nu}_{Y_2}\|_{B_0}$ to the total error. Indeed, this set is given by

$$J^* = \big\{\mu \in J\backslash J_P;\ \mu = \nu + \epsilon^m \text{ for all } \nu \in J_P,\ \text{for all } m \in \mathbb{N}\big\}, \tag{5.65}$$

where ϵ^m is the Kronecker delta sequence given by

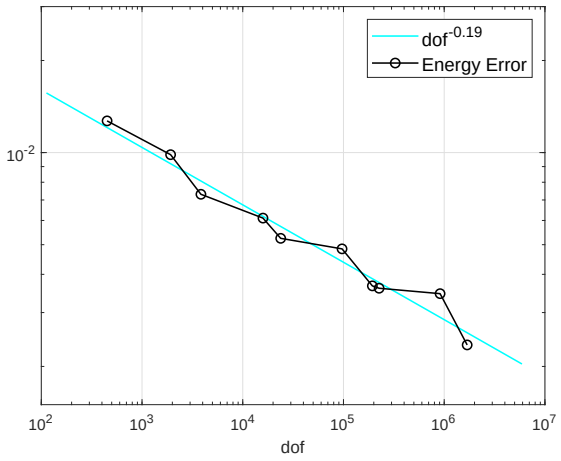
$$\epsilon^m := (\epsilon^m_1, \epsilon^m_2, \dots), \qquad \epsilon^m_j = \delta_{mj}, \qquad \text{for all } j \in \mathbb{N}.$$

We call J^* the set of *neighbouring* multi–indices and choose a subset $J_Q \subset J^*$ with which to define the error problem (5.45). Specifically, we choose

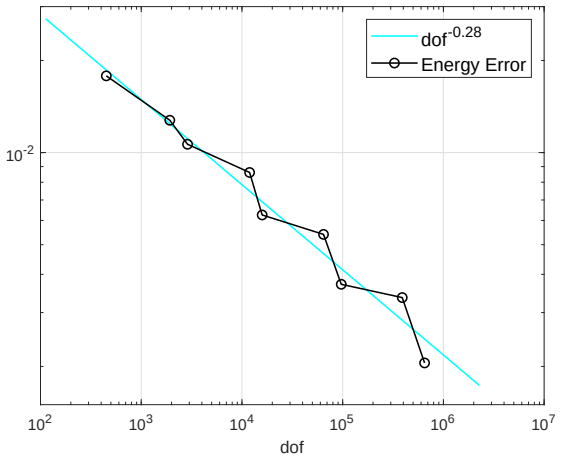
$$J_Q = \big\{\nu \in J^*;\ \max\{\text{supp}(\nu)\} \leq M + \Delta_M\big\}, \tag{5.66}$$

where $\Delta_M \in \mathbb{N}$ is the number of additional parameters we wish to activate. In other words, J_Q is the set of all multi–indices $\nu \in J$ which are nonzero in at most the first $M + \Delta_M$ positions such that $\|e^{\nu}_{Y_2}\|_{B_0} \neq 0$.

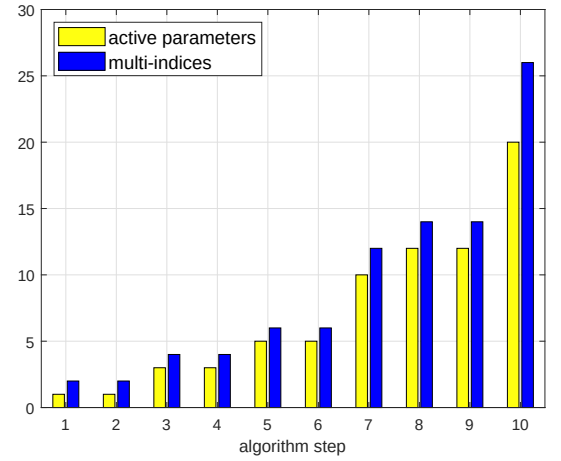
We now test the performance of the above algorithm by applying it to test problems TP1 and TP2 detailed in Section 5.1.1.



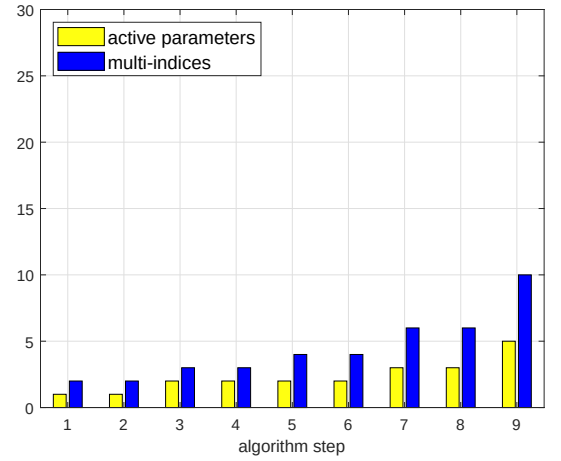
(a) convergence of $\eta = \|e_Y\|_{B_0}$ for TP1.



(b) convergence of $\eta = \|e_Y\|_{B_0}$ for TP2.



(c) M and card(J_P) for TP1.



(d) M and card(J_P) for TP2.

Figure 5.5: The convergence of $\eta = \|e_Y\|_{B_0}$ for Example 5.11 as well as the number of active parameters M and multi–indices card(J_P) in the definition of X at each step of the adaptive algorithm. We set the tolerance $\texttt{tol} = 3 \times 10^{-3}$.

Example 5.11: Adaptive single–level SGFEM.

For test problems TP1 and TP2, we start by constructing low–dimensional spaces $X = H_1 \otimes P$ using uniform 16×16 meshes $\mathcal{T}_h$ and $J_P = \{(0, 0, \dots), (1, 0, \dots)\}$ to compute initial approximations $u_X \in X$ satisfying (5.21). To compute the error estimate $\|e_Y\|_{B_0}$ in (5.48), we construct H_2 using the usual (global) $\mathbb{Q}_2$ basis functions, defined with respect to the element edge–midpoints and centroids of $\mathcal{T}_h$ (recall Example 3.3 for a similar setup), and choose J_Q as in (5.66) with $\Delta_M = 5$. For both problems, we adaptively construct a sequence of approximation spaces

$\{X\}$ by repeating (5.64) until $\|e_Y\|_B < \texttt{tol} := 3 \times 10^{-3}$. In Figures 5.5a and 5.5b we plot the convergence of $\eta = \|e_Y\|_{B_0}$. In Figures 5.5c and 5.5d we plot the number of active parameters M and multi–indices $\text{card}(J_P)$ in the definition of X at each step of the algorithm.

Figures 5.5a and 5.5b confirm that sequences of SGFEM spaces $\{X\}$ with a single–level structure can be constructed such that the energy error decays to zero at a rate independent of the number of active parameters, or, equivalently, the number of terms, in the expansion (5.5). Indeed, $\eta = \|e_Y\|_{B_0}$ decays at the rate of approximately $-\frac{1}{5}$ and $-\frac{1}{3}$ with respect to the number of DOFs N_X for TP1 and TP2, respectively. Notice that many more active parameters are required in the definition of X for TP1 than TP2 to ensure that $\eta = \|e_Y\|_{B_0}$ meets a prescribed tolerance. At the final step of the algorithm, twenty parameters are active for TP1 whereas only five are active for TP2. Again, we expect these results to relate directly to the rate of decay of the sequence $\{\|a_m\|_\infty\}_{m=1}^\infty$ associated with $a(\mathbf{x}, \mathbf{y})$ in (5.5). Since the sequence for TP1 decays more slowly than the sequence for TP2, a larger number of terms $a_m(\mathbf{x})$ are elected to play a role in the SGFEM approximation $u_X \in X$. The observed rates of convergence of $-\frac{1}{5}$ and $-\frac{1}{3}$ in Figure 5.5 are problematic in that large increases in N_X result in only small decreases in $\|u - u_X\|_B$. For small error tolerances $\texttt{tol}$, the single–level SGFEM algorithm presented in this section is very slow to converge.

5.5 Summary

In this chapter, we introduced the parametric diffusion problem (5.6)–(5.7), where the coefficient $a(\mathbf{x})$ in (3.22)–(3.23) was replaced with a coefficient $a(\mathbf{x}, \mathbf{y}) \in L^2_\pi(\Gamma, L^2(D))$ (a parameterised second–order random field) of the form (5.5). We employed standard SGFEMs in Section 5.2 to solve the associated weak formulation (5.12), where approximations were sought in tensor–product spaces of the form $X = H_1 \otimes P$ in (5.22) with $H_1 \subset H^1_0(D)$ representing a FEM space of piecewise polynomials on D and $P \subset L^2_\pi(\Gamma)$ representing a set of global polynomials on Γ. When P is the space of q–hyperbolic polynomials of degree k in the first M parameters y_m, we showed in Example 5.6 that significant computational savings can be made when solving test problems TP1 and TP2 in Section 5.1.1 by carefully selecting the parameter $q \in (0, 1]$. Indeed, the errors

associated with $q = 0.4$ in Figure 5.3 for TP1, for example, are comparable to the errors associated with $q = 1$, despite the massive reduction in the corresponding number of DOFs. Crucially, for TP1 and TP2, large numbers of multi–indices $\mu \in J$ in (5.27) are of little importance with respect to the energy error and must be overlooked when choosing the subset $J_P \subset J$ in the definition of P in (5.29).

In Section 5.4, we derived the error estimate $\|e_Y\|_{B_0} \approx \|u - u_X\|_B$ from [22] in Theorem 5.1, where $u_X \in X$ is a computed SGFEM approximation satisfying (5.21). The accuracy of $\eta = \|e_Y\|_{B_0}$ depends on the subspaces $H_2 \subset H_0^1(D)$ and $Q \subset L_\pi^2(\Gamma)$. In Section 5.4.1, we investigated which choices of H_2 and Q lead to estimates η with effectivity indices close to one. We showed in Examples 5.8 and 5.9 that if the sequence $\{\|a_m\|_\infty\}_{m=1}^\infty$ associated with $a(\mathbf{x}, \mathbf{y})$ in (5.5) decays too slowly, a large number of active parameters y_m are needed in the definition of Q. Building on those results, in Example 5.11 we tested a simple adaptive algorithm that exploits the estimate $\eta = \|e_Y\|_{B_0}$ to automatically select the most important multi–indices $\mu \in J$ for J_P and enrich the space H_1. The sets J_P were successfully tailored to the problem at hand – recall Figures 5.5c and 5.5d for some quantitative differences between the sets for test problems TP1 and TP2 – and convergence rates independent of the number of active parameters M were reported. In Chapter 6, we improve upon the rates of convergence in Figures 5.5a and 5.5b by designing adaptive SGFEMs where X has a more complex *multilevel* structure.

Chapter 6

Adaptive & Multilevel SGFEMs

In Chapter 5 we discussed so–called single–level SGFEMs for the parametric diffusion problem (5.6)–(5.7). We solved the finite–dimensional weak problem (5.21) for $B(\cdot,\cdot)$ and $F(\cdot)$ defined in (5.13) and (5.14), with approximation spaces X of the form (5.22). It was shown in Example 5.11 that sequences $\{X\}$ of single–level SGFEM spaces can be constructed adaptively such that $\|u - u_X\|_B \to 0$ at an algebraic rate independent of the number of input parameters y_m, thus breaking the curse of dimensionality associated with standard SGFEMs and the a priori error bound (5.62).

The best known theoretical rates of convergence are realised when X has a *multi-level* structure. That is, when

$$X = \bigoplus_{\mu\in J_P} H_1^\mu \otimes P^\mu, \qquad P^\mu = \operatorname{span}\big\{\psi_\mu(\mathbf{y})\big\}, \qquad J_P \subset J, \tag{6.1}$$

where $H_1^\mu \subset H_0^1(D)$ is a potentially different FEM space for each multi–index $\mu \in J_P$. It is shown in [36] that under Assumptions 3.3 and 5.2, if the sequences

$$\big\{\|a_m\|_\infty\big\}_{m=1}^\infty,\ \big\{\|\nabla a_m\|_\infty\big\}_{m=1}^\infty \in \ell^p(\mathbb{N}), \tag{6.2}$$

for some $0 < p < 1$ small enough, there exists a sequence of spaces $\{X\}$ of the form (6.1) for which the energy error $\|u - u_X\|_B$ decays to zero at the rate afforded to the chosen FEM applied to the analogous parameter–free problem. Whilst the sequence $\{X\}$ is generally unknown explicitly, it is demonstrated in [36] that the multi–indices J_P associated with each space X in the sequence correspond to the card(J_P) largest

values of $\|u^\mu\|_{H_0^1(D)}$ associated with the decomposition

$$u(\mathbf{x},\mathbf{y}) = \sum_{\mu\in J} u^\mu(\mathbf{x})\psi_\mu(\mathbf{y}), \qquad u^\mu \in H_0^1(D), \tag{6.3}$$

of $u \in V = L^2_\pi(\Gamma, H_0^1(D))$ satisfying (5.12).

Adaptive multilevel SGFEMs have been considered previously in [44] and [56], where the authors employ an explicit residual–based a posteriori error estimation strategy to steer the adaptive construction of X. In this work, we extend the implicit a posteriori error estimation strategy developed in [18, 22] (and analysed in Chapter 5) to the multilevel setting and then use this to design new SGFEMs. Our aim is to construct sequences of multilevel spaces $\{X\}$ for which the energy norm of the approximation error decays at the best possible rate. In general, to understand the optimum rate of convergence for a particular test problem beforehand, we need to know the value of p in (6.2) associated with $a(\mathbf{x},\mathbf{y})$ in (5.5). However, if our methods realise the rate afforded to the chosen FEM for the analogous parameter-free problem, the value of p in (6.2) is small enough and the optimum rate has been achieved.

6.1 Multilevel Approximation Spaces

To construct multilevel spaces of the form (6.1), we must first choose a subset $J_P \subset J$ of multi–indices and then choose an appropriate set of FEM spaces

$$\mathbf{H}_1 := \left\{H_1^\mu\right\}_{\mu\in J_P}, \qquad H_1^\mu \subset H_0^1(D). \tag{6.4}$$

Let

$$H_1^\mu = \text{span}\left\{\phi_i^\mu(\mathbf{x});\ i = 1, 2, \ldots, N_1^\mu\right\}, \qquad \mu \in J_P,$$

then, any function $v(\mathbf{x},\mathbf{y})$ can be expanded in its basis as

$$v(\mathbf{x},\mathbf{y}) = \sum_{\nu\in J_P}\sum_{j=1}^{N_1^\nu} v_j^\nu \phi_j^\nu(\mathbf{x})\psi_\nu(\mathbf{y})$$

with $v_j^\nu \in \mathbb{R}$. Note that the order of these finite sums is not interchangeable.

We now describe the construction of the set of FEM spaces $\mathbf{H}_1$. In this work, we insist that each space $H_1^\mu \in \mathbf{H}_1$ is constructed using continuous, piecewise polynomials of the same degree. Whilst this is not necessary, it enables us to characterise each

H_1^μ with only a single discretisation parameter (a level number) associated with the underlying finite element meshes.

We construct each space in the set $\mathbf{H}_1$ with respect to a finite element mesh over D. We assume that we can construct or have access to (from software packages, perhaps) a nested sequence of regular (no hanging nodes) rectangular or triangular meshes

$$\boldsymbol{\mathcal{T}} = \{\mathcal{T}_i;\ i = 0, 1, \dots\}, \tag{6.5}$$

such that $\mathcal{T}_j$ can be obtained from one or more refinements of $\mathcal{T}_i$ for $j > i$. For a fixed polynomial degree, the sequence $\boldsymbol{\mathcal{T}}$ then gives rise to a sequence of conforming finite element spaces

$$H^{(0)} \subset H^{(1)} \subset \cdots \subset H_0^1(D),$$

where $H^{(i)}$ depends on $\mathcal{T}_i$. We call i the mesh *level number* and construct each $H_1^\mu \in \mathbf{H}_1$ with respect to one of the meshes in $\boldsymbol{\mathcal{T}}$. That is, to each $\mu \in J_P$ we assign a mesh level number $\ell^\mu = i$ for some $i \in \mathbb{N}_0$ and set $H_1^\mu = H^{(i)}$. If $\ell^\mu = \ell^\nu$ for two multi–indices $\mu, \nu \in J_P$, then $H_1^\mu = H_1^\nu$. We collect the chosen levels in the set

$$\boldsymbol{\ell} := \left\{\ell^\mu;\ \mu \in J_P\right\}$$

and thus, $\mathrm{card}(\boldsymbol{\ell}) = \mathrm{card}(J_P)$. Given a fixed polynomial degree and the sequence (6.5), each $H_1^\mu \in \mathbf{H}_1$ is simply determined by its mesh level number ℓ^μ and the multilevel space X in (6.1) is completely characterised by our choices of J_P and $\boldsymbol{\ell}$.

Once J_P and $\boldsymbol{\ell}$ have been chosen, we solve the finite–dimensional weak problem:

$$\text{find: } u_X \in X: \qquad B(u_X, v) = F(v), \qquad \text{for all } v \in X, \tag{6.6}$$

where $B(\cdot,\cdot)$ and $F(\cdot)$ are defined in (5.13) and (5.14), respectively. Expanding the SGFEM approximation as

$$u_X(\mathbf{x}, \mathbf{y}) = \sum_{\mu \in J_P} u_X^\mu(\mathbf{x})\psi_\mu(\mathbf{y}), \qquad u_X^\mu(\mathbf{x}) = \sum_{i=1}^{N_1^\mu} u_i^\mu \phi_i^\mu(\mathbf{x}), \qquad u_i^\mu \in \mathbb{R}, \tag{6.7}$$

and choosing the linearly independent test functions $v = \phi_j^\nu(\mathbf{x})\psi_\nu(\mathbf{y})$ in (6.6) yields

$$\sum_{\mu \in J_P} \sum_{i=1}^{N_1^\mu} u_i^\mu B(\phi_i^\mu(\mathbf{x})\psi_\mu(\mathbf{y}), \phi_j^\nu(\mathbf{x})\psi_\nu(\mathbf{y})) = F(\phi_j^\nu(\mathbf{x})\psi_\nu(\mathbf{y})), \tag{6.8}$$

for all $\nu \in J_P$ and $j = 1, 2, \ldots, N_1^\nu$. This is a linear system of equations $A\mathbf{u} = \mathbf{b}$ where $A \in \mathbb{R}^{N_X \times N_X}$ and $\mathbf{u}, \mathbf{b} \in \mathbb{R}^{N_X}$ with

$$N_X = \dim(X) = \sum_{\mu \in J_P} \dim(H_1^\mu) = \sum_{\mu \in J_P} N_1^\mu.$$

In the next section, we derive the precise structure and entries of A. We also explain how the action of A on vectors can be efficiently computed within an iterative solver.

6.2 Multilevel SGFEM Matrices

The matrix A associated with the system of equations (6.8) is symmetric and positive definite with a block structure. Indeed, the blocks of A, as well as the blocks of $\mathbf{u}$ and $\mathbf{b}$, are indexed by the elements (multi–indices) of J_P. Namely

$$[A]_{\nu\mu} = A_{\nu\mu} \in \mathbb{R}^{N_1^\nu \times N_1^\mu}, \qquad [\mathbf{b}]_\nu = \mathbf{b}_\nu \in \mathbb{R}^{N_1^\nu}, \qquad [\mathbf{u}]_\mu = \mathbf{u}_\mu \in \mathbb{R}^{N_1^\mu},$$

for $\mu, \nu \in J_P$, with entries

$$\begin{aligned} [A_{\nu\mu}]_{ji} &= B(\phi_i^\mu(\mathbf{x})\psi_\mu(\mathbf{y}), \phi_j^\nu(\mathbf{x})\psi_\nu(\mathbf{y})), \\ [\mathbf{b}_\nu]_j &= F(\phi_j^\nu(\mathbf{x})\psi_\nu(\mathbf{y})), \\ [\mathbf{u}_\mu]_i &= u_i^\mu, \end{aligned}$$

for $i = 1, 2, \ldots, N_1^\mu$ and $j = 1, 2, \ldots, N_1^\nu$. The positions of the blocks $A_{\nu\mu}$ and $\mathbf{b}_\nu$ in A and $\mathbf{b}$ coincide with that of A_{rj} and $\mathbf{b}_r$ in (5.23) for the single–level method, however, the dimensions of $A_{\nu\mu}$ and $\mathbf{b}_\nu$ are not necessarily uniform and $A_{\nu\mu}$ may not be square. Taking the decomposition of $a(\mathbf{x}, \mathbf{y})$ in (5.5) into account yields

$$\begin{aligned} [A_{\nu\mu}]_{ji} &= \int_\Gamma \psi_\mu(\mathbf{y})\psi_\nu(\mathbf{y}) \int_D a(\mathbf{x}, \mathbf{y}) \nabla\phi_i^\mu(\mathbf{x}) \cdot \nabla\phi_j^\nu(\mathbf{x}) \, d\mathbf{x} \, d\pi(\mathbf{y}) \\ &= \sum_{m=0}^{M} [G_m]_{\nu\mu} \int_D a_m(\mathbf{x}) \nabla\phi_i^\mu(\mathbf{x}) \cdot \nabla\phi_j^\nu(\mathbf{x}) \, d\mathbf{x}, \end{aligned}$$

where the matrices G_m are defined in (5.25) and M is number of active parameters y_m incorporated into J_P. Consequently, the blocks $A_{\nu\mu}$ of A admit the decomposition

$$A_{\nu\mu} = \sum_{m=0}^{M} [G_m]_{\nu\mu} K_{\nu\mu}^m, \qquad [K_{\nu\mu}^m]_{ji} := \int_D a_m(\mathbf{x}) \nabla\phi_i^\mu(\mathbf{x}) \cdot \nabla\phi_j^\nu(\mathbf{x}) \, d\mathbf{x}, \tag{6.9}$$

for $i = 1, 2, \ldots, N_1^\mu$ and $j = 1, 2, \ldots, N_1^\nu$. The entries of the matrices $K_{\nu\mu}^m$ depend on finite element basis functions $\phi_i^\mu(\mathbf{x})$ and $\phi_j^\nu(\mathbf{x})$ associated with a pair of meshes $\mathcal{T}_{\ell^\mu}$ and $\mathcal{T}_{\ell^\nu}$, which may be different. Consequently, $K_{\nu\mu}^m \in \mathbb{R}^{N_1^\nu \times N_1^\mu}$ is nonsquare if $\ell^\mu \neq \ell^\nu$ so that $N_1^\mu \neq N_1^\nu$. As a result, A does not admit the Kronecker–product decomposition (5.31) afforded to its single–level counterpart, and efficient matrix–vector products are no longer computable through standard formulae.

6.2.1 Efficient Matrix–Vector Products

Once we have solved the linear system $A\mathbf{u} = \mathbf{b}$, the vector $\mathbf{u}$ provides the coefficients u_i^μ that define $u_X(\mathbf{x}, \mathbf{y})$ in (6.7). It should be noted, however, that we do not explicitly construct A. We need only compute the action of A on vectors when using iterative solvers to approximate $\mathbf{u}$. We exploit the structure of A by computing the matrix–vector product $\mathbf{v} = A\mathbf{x}$ blockwise via

$$[\mathbf{v}]_\nu = [A\mathbf{x}]_\nu = \sum_{\mu \in J_P} A_{\nu\mu}[\mathbf{x}]_\mu = \sum_{\mu \in J_P} \sum_{m=0}^{M} [G_m]_{\nu\mu} K_{\nu\mu}^m [\mathbf{x}]_\mu, \tag{6.10}$$

for $\nu \in J_P$.

The stiffness matrices $K_{\nu\mu}^m$ in (6.10) depend on the triplet (m, ν, μ) for $\nu, \mu \in J_P$ and $m = 0, 1, \ldots, M$. This suggests that $(M+1)s^2$ matrices need to be computed where $s = \text{card}(J_P)$. For large values of M or s, this is prohibitively expensive. Recall, the single–level method only required the computation of $(M+1)$ stiffness matrices. For multilevel SGFEMs to be cost–effective, we must be especially careful not to perform any unnecessary computations. It is clear from (6.10) that we do not need to compute $K_{\nu\mu}^m$ when $[G_m]_{\nu\mu} = 0$. It was shown in Chapter 5 that $G_0 = I$ and the matrices G_m have at most two non–zero entries per row. Hence, a sharper upper bound for the number of required stiffness matrices is $(1 + 2M)s$.

Whilst this bound takes the sparsity of each G_m into account, it does not exploit the fact that the polynomial degree of the FEM spaces H_1^μ is fixed and the same mesh $\mathcal{T}_{\ell^\mu}$ may be assigned to several multi–indices $\mu \in J_P$. In fact, we need only compute $K_{\nu\mu}^m$ for all *distinct* triplets (m, ℓ^ν, ℓ^μ) for which $[G_m]_{\nu\mu}$ is nonzero. Let $\boldsymbol{\ell}^*$ denote the set of distinct elements of $\boldsymbol{\ell}$ with cardinality $t := \text{card}(\boldsymbol{\ell}^*) \leq s$. Then, an improved

Table 6.1: Theoretical upper bound $t + M\min\{s, T\}$ for the number of required matrices $K^m_{\nu\mu}$ for test problems TP1–TP4, and the actual number when the set J_P and the mesh level numbers $\boldsymbol{\ell}$ are selected automatically using Algorithm 1 in Section 6.4.

test problem	s	t	M	$t + M\min\{s, T\}$	actual
TP1	169	4	93	934	313
TP2	36	5	13	200	53
TP3	17	5	3	50	22
TP4	21	3	8	51	31

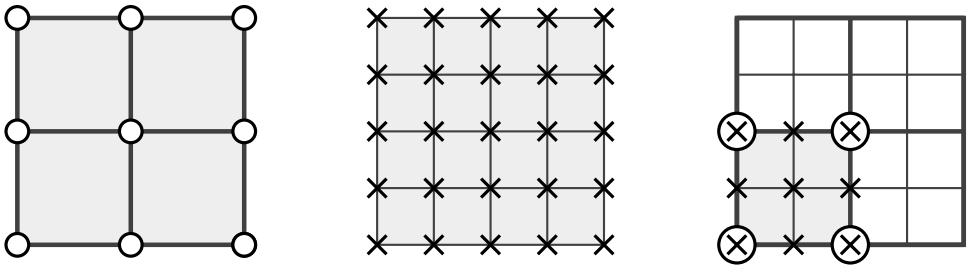
upper bound is

$$t + M\min\big\{s, T\big\}, \qquad T := \tfrac{t}{2}(t-1),$$

where T is the number of distinct pairings of elements in $\boldsymbol{\ell}^*$. Here, note that there are exactly t distinct triples $(0, \ell^\mu, \ell^\mu)$ associated with the matrix G_0 and T possible distinct triples (m, ℓ^ν, ℓ^μ) associated with G_m for $m = 1, 2, \ldots, M$. An adaptive algorithm for automatically selecting J_P as well as the associated set of mesh level numbers $\boldsymbol{\ell}$ is developed in Section 6.4. In Table 6.1 we record s, t, M and the number of matrices $K^m_{\nu\mu}$ that are required at the final step of that algorithm (when the error tolerance is set to $\epsilon = 2 \times 10^{-3}$), for the four test problems outlined in Section 5.1.1. This algorithm will be described in detail later. For now, we simply note that, since the same mesh level number is assigned to many multi–indices in J_P, the total number of stiffness matrices required is much lower than the above bounds suggest.

6.2.2 Efficient Assembly of Stiffness Matrices

When the basis functions ϕ^μ_i and ϕ^ν_j associated with the stiffness matrix $K^m_{\nu\mu}$ in (6.9) are defined with respect to two meshes $\mathcal{T}_{\ell^\mu}$ and $\mathcal{T}_{\ell^\nu}$ that have different mesh level numbers ($\ell^\mu \neq \ell^\nu$), $K^m_{\nu\mu}$ cannot be constructed using conventional element assembly methods. Computing these non–square matrices is a tough computational challenge and requires new software to be written. In [44], to avoid this, the non–square matrices are approximated using a projection technique involving only the square matrices $K^m_{\mu\mu}$ that feature in the diagonal blocks $A_{\mu\mu}$ of A. Even with this approximation step, the multilevel method considered in [44] is reported to be too computationally expensive. Furthermore, the authors report that the approximation technique does not maintain the natural symmetry of A, and thus iterative solvers such as the conjugate



(a) $\mathcal{T}_{\ell^\mu}$. (b) $\mathcal{T}_{\ell^\nu}$. (c) elements embedded in $\square_{\text{coarse}}$.

Figure 6.1: Example meshes with (a) $N_1^\mu = 9$ and level number ℓ^μ and (b) $N_1^\nu = 25$ and level number $\ell^\nu = \ell^\mu + 1$.

gradient method are technically no longer applicable. In this section we describe how the matrices $K_{\nu\mu}^m$ can be computed quickly and efficiently, without the need for the approximation step used in [44].

We describe the construction of $K_{\nu\mu}^m$ for two multi–indices $\mu, \nu \in J_P$, with $\ell^\mu \neq \ell^\nu$ (m is not important here) for a simple example. To convey the key ideas we consider only uniform meshes of square elements on a square domain D. However, the procedure is applicable to any FEM spaces H_1^μ and H_1^ν for which the associated meshes $\mathcal{T}_{\ell^\mu}$ and $\mathcal{T}_{\ell^\nu}$ are nested, that is, when $\mathcal{T}_{\ell^\nu}$ is obtained from conforming refinements of $\mathcal{T}_{\ell^\mu}$.

Example 6.1: Matrix assembly.

For simplicity, assume that $D \subset \mathbb{R}^2$ is square and that the spaces H_1^μ and H_2^μ are $\mathbb{Q}_1$ FEM spaces defined with respect to two uniform meshes of square elements $\mathcal{T}_{\ell^\mu}$ and $\mathcal{T}_{\ell^\nu}$, respectively, that are only one refinement apart. In particular, let $\mathcal{T}_{\ell^\mu}$ denote a 2×2 grid (see Figure 6.1a) with mesh level number ℓ^μ and let $\mathcal{T}_{\ell^\nu}$ denote a 4×4 grid (see Figure 6.1b) with mesh level number $\ell^\nu := \ell^\mu + 1$ (representing a uniform refinement of $\mathcal{T}_{\ell^\mu}$). As is standard practice in FEM software, we retain the boundary nodes when initially constructing $K_{\nu\mu}^m \in \mathbb{R}^{N_1^\nu \times N_1^\mu}$ so that

$$N_1^\mu = \dim(H_1^\mu) := 9, \qquad N_2^\nu = \dim(H_1^\nu) := 25.$$

To construct the matrix $K_{\nu\mu}^m \in \mathbb{R}^{25\times 9}$ defined in (6.9) we concatenate four 9×4 *coarse–element* matrices – one for each element in the coarse mesh $\mathcal{T}_{\ell^\mu}$. In Figure 6.1c we highlight one such element, which we denote $\square_{\text{coarse}}$, as well as

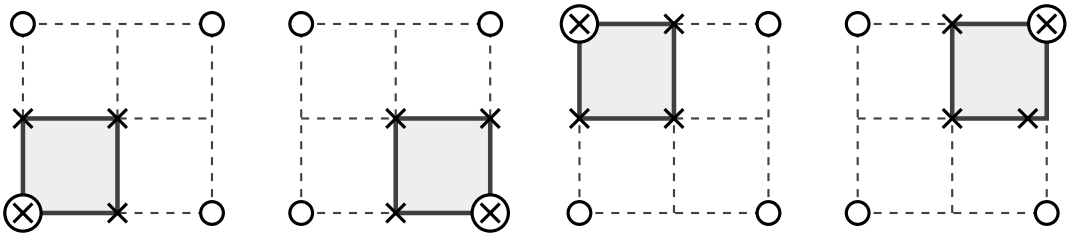


Figure 6.2: The four embedded elements in Figure 6.1c on which we construct four 4×4 local matrices.

the four elements in the finer grid that are embedded within it. The associated coarse–element matrix $K^m_{\nu\mu,\text{coarse}} \in \mathbb{R}^{9\times 4}$ has entries

$$[K^m_{\nu\mu,\text{coarse}}]_{ji} = \int_{\square_\text{coarse}} a_m \nabla\phi_i^{\mu,\text{coarse}} \cdot \nabla\phi_j^{\nu,\text{coarse}}\, d\mathbf{x}$$

for $i = 1, 2, 3, 4$ and $j = 1, 2, \ldots, 9$, where the sets of $\mathbb{Q}_1$ basis functions

$$\big\{\phi_i^{\mu,\text{coarse}};\ i = 1, 2, 3, 4\big\}, \qquad \big\{\phi_j^{\nu,\text{coarse}};\ j = 1, 2, \ldots, 9\big\},$$

are defined with respect to the round and cross markers in Figure 6.1c, respectively. Note that the functions $\phi_i^{\mu,\text{coarse}}$ are global functions with respect to $\square_\text{coarse}$ in that they are supported on (only) the whole element. In contrast, the functions $\phi_j^{\nu,\text{coarse}}$ are supported only on patches of $\square_\text{coarse}$ and are piecewise functions. As a result, we construct the coarse–element matrix $K^m_{\nu\mu,\text{coarse}} \in \mathbb{R}^{9\times 4}$ by concatenating four 4×4 *fine–element* matrices – one for each fine element in $\square_\text{coarse}$.

In Figure 6.2 we highlight the four fine elements that are embedded in $\square_\text{coarse}$. The associated fine–element matrices $K^m_{\nu\mu,\text{fine}} \in \mathbb{R}^{4\times 4}$ have entries

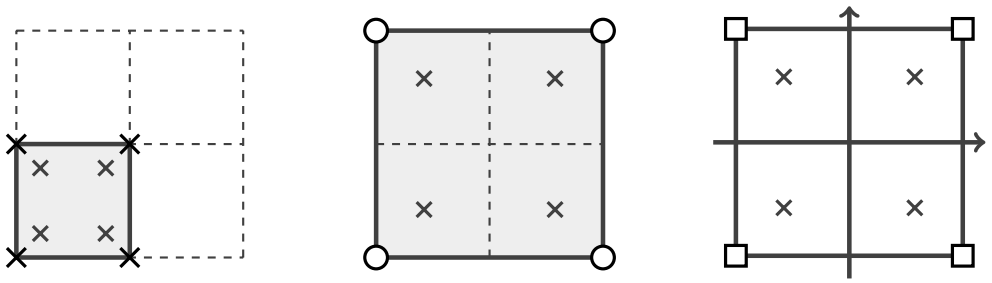
$$[K^m_{\nu\mu,\text{fine}}]_{ji} = \int_{\square_\text{fine}} a_m \nabla\phi_i^{\mu,\text{coarse}} \cdot \nabla\phi_j^{\nu,\text{fine}}\, d\mathbf{x}, \tag{6.11}$$

where $\square_\text{fine}$ is one of the four elements embedded in $\square_\text{coarse}$ and the basis functions

$$\big\{\phi_j^{\nu,\text{fine}};\ j = 1, 2, 3, 4\big\},$$

are defined with respect to the crosses in Figure 6.2, that are supported only on $\square_\text{fine}$ (the shaded regions).

For $\mathbb{Q}_1$ approximation, the construction of the matrices $K^m_{\nu\mu}$ always reduces to the concatenation of 4×4 fine–element matrices $K^m_{\nu\mu,\text{fine}}$, even if the two grids are more



(a) $\square_{\text{fine}} \in \mathcal{T}_{\ell^\nu}$ embedded in $\square_{\text{coarse}}$.

(b) $\square_{\text{coarse}} \in \mathcal{T}_{\ell^\mu}$.

(c) $\square_{\text{ref}} = [-1, 1]^2$.

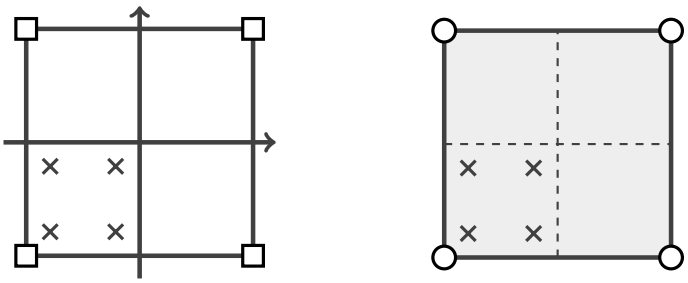
Figure 6.3: Example 2×2 Gauss quadrature points (c) on a reference element mapped to (a) a fine element in $\mathcal{T}_{\ell^\nu}$ and (b) a coarse element in $\mathcal{T}_{\ell^\mu}$.

than one refinement level apart. Similarly, for $\mathbb{P}_1$ approximation the construction of $K^m_{\nu\mu}$ reduces to the concatenation of 3×3 fine–element matrices. If $\ell^\mu = \ell^\nu$ so that $K^m_{\nu\mu}$ is square we use conventional element assembly methods. In either case, we need only perform quadrature on elements in the fine mesh.

Efficient Numerical Quadrature

The way we employ quadrature to evaluate the integrals in (6.11) plays a key role in designing computationally efficient multilevel methods. Here, we explain how quadrature can be vectorised over the coarse elements $\square_{\text{coarse}} \in \mathcal{T}_\mu$ when the meshes $\mathcal{T}_\mu$ and $\mathcal{T}_\nu$ are uniform. If appropriate, software implementations of $\mathbb{Q}_1$ or $\mathbb{Q}_2$ finite element methods map the reference element $\square_{\text{ref}} = [-1, 1]^2$ in Figure 6.3c to elements in physical space; $\square_{\text{fine}}$ or $\square_{\text{coarse}}$ in Figures 6.3a and 6.3b for example. This mapping is extremely convenient and uses the $\mathbb{Q}_1$ element basis functions associated with the square markers in Figure 6.3c. Through the mapping from $\square_{\text{ref}}$ to $\square_{\text{fine}}$, the evaluation of basis functions on $\square_{\text{fine}}$ at quadrature points in physical space can, for example, be written in terms of evaluating basis functions on $\square_{\text{ref}}$ at reference quadrature points. The same goes for the mapping from $\square_{\text{ref}}$ to $\square_{\text{coarse}}$, thus we never explicitly define basis functions on $\square_{\text{fine}}$ or $\square_{\text{coarse}}$. We illustrate by the grey crosses in Figure 6.3c the 2×2 Gauss quadrature rule on $\square_{\text{ref}}$, as well as their mapped positions in $\square_{\text{fine}}$ and $\square_{\text{coarse}}$ in Figures 6.3a and 6.3b, respectively.

Suppose now that we seek to compute the entries of the matrix $K^m_{\nu\mu,\text{fine}}$ in (6.11) associated with the shaded element in the left–most diagram in Figure 6.2. By definition, the mapping from $\square_{\text{ref}}$ to $\square_{\text{fine}}$ ensures that the quadrature points at which



(a) $\square_{\text{ref}} = [-1, 1]^2$. (b) $\square_{\text{coarse}} \in \mathcal{T}_{\ell^\mu}$.

Figure 6.4: (a) The 2×2 Gauss quadrature points on $\square_{\text{ref}}$ shifted to the lower–left quadrant as well as (b) the corresponding mapped points in $\square_{\text{coarse}}$.

$\nabla\phi_j^{\nu,\text{fine}}$ in (6.11) is implicitly evaluated lie in $\square_{\text{fine}}$. We also want to evaluate $\nabla\phi_i^{\mu,\text{coarse}}$ at the *same* quadrature points. To achieve this using a map from $\square_{\text{ref}}$ to $\square_{\text{coarse}}$, we must initially shift the reference quadrature points to the lower–left quadrant of $\square_{\text{ref}}$. Then, the quadrature points at which $\nabla\phi_i^{\mu,\text{coarse}}$ is implicitly evaluated lie in $\square_{\text{fine}}$ and coincide with those for $\nabla\phi_j^{\nu,\text{fine}}$. This is visualised in Figure 6.4. Note that all the fine–element matrices associated with, for example, the lower–left quadrants of the coarse elements, can be computed simultaneously, since these all require the same mapping of the quadrature points on $\square_{\text{ref}}$ to its lower–left quadrant.

6.3 A Posteriori Error Estimation

Given a multilevel space X of the form (6.1) and an SGFEM approximation $u_X \in X$ satisfying (6.6), we want to estimate $||u - u_X||_B$ a posteriori. Again, we use the implicit strategy outlined in Chapter 3 as a foundation for a new multilevel SGFEM strategy. Our starting point is the abstract problem (3.17), which we recall for our parametric problem:

$$\text{find } e_Y \in Y: \qquad B_0(e_Y, v) = F(v) - B(u_X, v), \qquad \text{for all } v \in Y, \tag{6.12}$$

where, $Y \subset V = L^2_\pi(\Gamma, H^1_0(D))$ satisfies $X \cap Y = \{0\}$, $B(\cdot,\cdot)$ and $F(\cdot)$ are defined in (5.13) and (5.14), and $B_0(\cdot,\cdot)$ is the parameter–free bilinear form defined in (5.17). If Assumption 3.1 holds for the space $W := X \oplus Y$, as well as Assumptions 5.2 and 5.3, then the estimate $\eta = \|e_Y\|_{B_0}$ satisfies the bound (3.20) for the constants λ and Λ defined in (5.20). In addition, β is the constant in (3.11) associated with our choices of X and W, and γ is the constant in (3.18) associated with our choices of X and Y.

Our goal in this Chapter is to design an efficient adaptive multilevel SGFEM algorithm, in which the energy error $\|u - u_X\|_B$ is estimated a posteriori at each step by solving (6.12). Suppose that the sets $\mathbf{H}_1$ and J_P which define X in (6.1) are chosen, then, one possibility (as remarked in [22]) is to choose a subset $J_Q \subset J$ in the definition of Q in (5.40) such that $J_P \cap J_Q = \emptyset$, as well as another set of FEM spaces

$$\mathbf{H}_2 := \left\{H_2^\mu\right\}_{\mu\in J_P}, \qquad H_2^\mu \subset H_0^1(D), \qquad H_1^\mu \cap H_2^\mu = \{0\}, \tag{6.13}$$

and construct

$$Y = \left(\bigoplus_{\mu\in J_P} H_2^\mu \otimes P^\mu\right) \oplus (H \otimes Q), \tag{6.14}$$

where $H \subset H_0^1(D)$ is some other FEM space to be chosen; we discuss this later. This multilevel structure is simply a more general version of the structure of Y defined in (5.37), in that we have more flexibility in the way we combine new functions on D with existing polynomials on Γ and vice versa. The quality of the estimate $\eta = \|e_Y\|_{B_0}$ depends on our choices of H_2^μ and J_Q, since these affect the constants γ and β appearing in (3.20).

Decomposition of the Error

The space Y in (6.14) admits the decomposition $Y = Y_1 \oplus Y_2$ such that $Y_1 \cap Y_2 = \{0\}$ for the spaces

$$Y_1 := \bigoplus_{\mu\in J_P} H_2^\mu \otimes P^\mu, \qquad Y_2 := H \otimes Q,$$

and thus $e_Y = e_{Y_1} + e_{Y_2}$ for some $e_{Y_1} \in Y_1$ and $e_{Y_2} \in Y_2$. By the mutual orthogonality of P given in (5.29) and Q with respect to the inner–product $\langle\cdot,\cdot\rangle_{L^2_\pi(\Gamma)}$, problem (6.12) again decouples into two lower–dimensional problems:

$$\text{find } e_{Y_i} \in Y_i: \qquad B_0(e_{Y_i}, v) = F(v) - B(u_X, v), \qquad \text{for all } v \in Y_i, \tag{6.15}$$

for $i = 1, 2$, of dimension $\dim(Y_1) = \sum_{\mu\in J_P} \dim(H_2^\mu)$ and $\dim(Y_2) = \dim(H)\dim(Q)$, respectively. Note that the only difference between these problems and those in (5.44)–(5.45) is the construction of the spaces Y_1 and Y_2.

Clearly, the spaces Y_1 and Y_2 decompose as

$$Y_1 = \bigoplus_{\mu\in J_P} Y_1^\mu, \qquad Y_1^\mu := H_2^\mu \otimes P^\mu, \qquad Y_2 = \bigoplus_{\nu\in J_Q} Y_2^\nu, \qquad Y_2^\nu := H \otimes Q^\nu, \tag{6.16}$$

where Q^ν is defined in (5.46). It follows that

$$Y_1^\mu \cap Y_1^{\bar\mu} = \{0\}, \qquad \text{for } \mu \neq \bar\mu, \qquad Y_2^\nu \cap Y_2^{\bar\nu} = \{0\}, \qquad \text{for } \nu \neq \bar\nu,$$

and thus $e_{Y_1} \in Y_1$ and $e_{Y_2} \in Y_2$ can be further decomposed as

$$e_{Y_1} = \sum_{\mu\in J_P} e_{Y_1}^\mu, \qquad e_{Y_1}^\mu \in Y_1^\mu, \qquad e_{Y_2} = \sum_{\nu\in J_Q} e_{Y_2}^\nu, \qquad e_{Y_2}^\nu \in Y_2^\nu.$$

Like the decoupled problems in (5.47) for the single–level method, another orthogonality argument shows that the first problem in (6.15) decouples into $\mathrm{card}(J_P)$ problems of dimension $N_{Y_1}^\mu := \dim(Y_1^\mu) = \dim(H_2^\mu)$, and the second into $\mathrm{card}(J_Q)$ problems of dimension $N_{Y_2}^\nu := \dim(Y_2^\nu) = \dim(H)$, namely

$$\text{find } e_{Y_1}^\mu \in Y_1^\mu: \qquad B_0(e_{Y_1}^\mu, v) = F(v) - B(u_X, v), \qquad \text{for all } v \in Y_1^\mu, \tag{6.17}$$

$$\text{find } e_{Y_2}^\nu \in Y_2^\nu: \qquad B_0(e_{Y_2}^\nu, v) = F(v) - B(u_X, v), \qquad \text{for all } v \in Y_2^\nu, \tag{6.18}$$

for $\mu \in J_P$ and $\nu \in J_Q$.

The total estimate $\|e_Y\|_{B_0}$ then admits the familiar decomposition

$$\eta = \|e_Y\|_{B_0} = \left[\|e_{Y_1}\|_{B_0}^2 + \|e_{Y_2}\|_{B_0}^2\right]^{\frac12} = \left[\sum_{\mu\in J_P} \|e_{Y_1}^\mu\|_{B_0}^2 + \sum_{\nu\in J_Q} \|e_{Y_2}^\nu\|_{B_0}^2\right]^{\frac12}. \tag{6.19}$$

As before, we refer to $\|e_{Y_1}\|_{B_0}$ as the spatial estimate and to $\|e_{Y_2}\|_{B_0}$ as the parametric estimate. Roughly speaking, each $\|e_{Y_1}^\mu\|_{B_0}$ estimates the energy lost through the approximation $\|u_X^\mu\|_{H_0^1(D)} \approx \|u^\mu\|_{H_0^1(D)}$ for $\mu \in J_P$, whereas each $\|e_{Y_2}^\nu\|_{B_0}$ estimates the energy of the mode associated with $\nu \in J_Q$ that is neglected in the definition of the SGFEM approximation $u_X \in X$.

The CBS constant

We now identify a constant $\gamma = \gamma^*$ for which (3.18) holds for the bilinear form $B_0(\cdot,\cdot)$ defined in (5.17) for the multilevel spaces X and Y in (6.1) and (6.14). Under Assumption 5.3, the space $H_0^1(D)$ is a Hilbert space with respect to the inner–product (4.7) with $a_0(\mathbf{x})$ in place of $a(\mathbf{x})$. The spaces $H_1^\mu, H_2^\mu \subset H_0^1(D)$ satisfy $H_1^\mu \cap H_2^\mu = \{0\}$. Thus, by Theorem 2.2 there exists a constant $\gamma^\mu \in [0,1)$ such that

$$|\langle u, v\rangle_{a_0}| \leq \gamma^\mu \langle u, u\rangle_{a_0}^{\frac12} \langle v, v\rangle_{a_0}^{\frac12}, \qquad \text{for all } u \in H_1^\mu, \qquad \text{for all } v \in H_2^\mu, \tag{6.20}$$

for each $\mu \in J_P$.

Expand two arbitrary functions $u \in X$ and $v \in Y$ by

$$u = \sum_{\mu \in J_P} u^\mu \psi_\mu(\mathbf{y}), \qquad u^\mu \in H_1^\mu,$$
$$v = \sum_{\bar{\mu} \in J_P} v_1^{\bar{\mu}} \psi_{\bar{\mu}}(\mathbf{y}) + \sum_{\nu \in J_Q} v_2^\nu \psi_\nu(\mathbf{y}), \qquad v_1^{\bar{\mu}} \in H_2^{\bar{\mu}},\ v_2^\nu \in H,$$

i.e., $v = v_1 + v_2$ for some functions $v_1 \in Y_1$ and $v_2 \in Y_2$. From the definition of $B(\cdot,\cdot)$ in (5.17)

$$B_0(u,v) = \sum_{\mu \in J_P} \sum_{\bar{\mu} \in J_P} \int_\Gamma \psi_\mu(\mathbf{y}) \psi_{\bar{\mu}}(\mathbf{y})\ d\pi(\mathbf{y}) \int_D a_0 \nabla u^\mu \cdot \nabla v_1^{\bar{\mu}}\ d\mathbf{x} + \sum_{\mu \in J_P} \sum_{\nu \in J_Q} \int_\Gamma \psi_\mu(\mathbf{y}) \psi_\nu(\mathbf{y})\ d\pi(\mathbf{y}) \int_D a_0 \nabla u^\mu \cdot \nabla v_2^\nu\ d\mathbf{x}.$$

Due to the orthogonality of the polynomials $\{\psi_\mu(\mathbf{y})\}_{\mu \in J_P \cap J_Q}$ with respect to the inner–product $\langle \cdot,\cdot \rangle_{L^2_\pi(\Gamma)}$, the first double–sum has nonzero contributions only when $\mu = \bar{\mu}$, and since $J_P \cap J_Q = \emptyset$, the second double–sum is equal to zero, that is

$$|B_0(u,v)| = \left| \sum_{\mu \in J_P} \int_D a_0 \nabla u^\mu \cdot \nabla v_1^\mu\ d\mathbf{x} \right| = \left| \sum_{\mu \in J_P} \langle u^\mu, v_1^\mu \rangle_{a_0} \right|.$$

Applying (6.20) as well as the Cauchy–Schwarz inequality for sums yields

$$\begin{aligned} |B_0(u,v)| &\le \gamma_* \sum_{\mu \in J_P} \langle u^\mu, u^\mu \rangle_{a_0}^{\frac{1}{2}} \langle v_1^\mu, v_1^\mu \rangle_{a_0}^{\frac{1}{2}} \\ &\le \gamma_* \left[\sum_{\mu \in J_P} \langle u^\mu, u^\mu \rangle_{a_0} \right]^{\frac{1}{2}} \left[\sum_{\mu \in J_P} \langle v_1^\mu, v_1^\mu \rangle_{a_0} \right]^{\frac{1}{2}} \end{aligned}$$

for all $u \in X$ and $v \in Y$, where

$$\gamma_* := \max_{\mu \in J_P} \{\gamma_{\min}^\mu\}, \tag{6.21}$$

and $\gamma_{\min}^\mu$ denotes the CBS constant associated with the bound (6.20). Note that

$$\|u\|_{B_0}^2 = \sum_{\mu \in J_P} \langle u^\mu, u^\mu \rangle_{a_0}, \qquad \|v_1\|_{B_0}^2 = \sum_{\mu \in J_P} \langle v_1^\mu, v_1^\mu \rangle_{a_0},$$

and $\|v_1\|_{B_0} \le \|v\|_{B_0}$ (recall Section 5.4). Thus

$$|B_0(u,v)| \le \gamma_* \|u\|_{B_0} \|v\|_{B_0}, \qquad \text{for all } u \in X, \qquad \text{for all } v \in Y,$$

and the bound (3.18) holds for the multilevel spaces X and Y with the constant $\gamma = \gamma^*$ in (6.21). Consequently, the constant γ_* appears in the error bound (3.20) and can be determined by analysing the FEM spaces H_1^μ and H_2^μ (clearly, the CBS constant $\gamma_{\min}$ associated with (3.18) also satisfies $\gamma_{\min} \leq \gamma_*$).

In Chapter 4 we showed that for two FEM spaces $H_1^\mu, H_2^\mu \subset H_0^1(D)$, there often exists a sharp upper bound for the associated CBS constant $\gamma_{\min}^\mu$ that is independent of the mesh level number ℓ^μ, or equivalently, the size of the underlying grid.

6.3.1 The Spatial & Parametric Estimates

In this section we first discuss possible choices for the FEM spaces $\mathbf{H}_2 = \{H_2^\mu\}_{\mu \in J_P}$ that define the tensor–product spaces

$$\mathbf{Y}_1 := \left\{ Y_1^\mu;\ \mu \in J_P \right\} \tag{6.22}$$

in (6.16). For each $H_1^\mu \in \mathbf{H}_1$ in the definition of X in (6.1), $H_2^\mu \subset H_0^1(D)$ must be chosen such that $H_1^\mu \cap H_2^\mu = \{0\}$. Recall that the FEM spaces H_1^μ are constructed using piecewise polynomials of the same degree, with respect to an underling mesh $\mathcal{T}_{\ell^\mu} = \mathcal{T}_i \in \boldsymbol{\mathcal{T}}$ for some $i \in \mathbb{N}_0$. In light of the single–level experiments conducted in Chapter 5, we recommend the following analogous multilevel options:

1. Construct each H_2^μ with piecewise polynomials of the same degree as those in H_1^μ. Specifically, use basis functions associated with the new nodes that would be introduced by performing the mesh refinement $\mathcal{T}_{\ell^\mu} \to \mathcal{T}_{i+1}$, i.e., by increasing the mesh level number by one.

2. Construct each H_2^μ using piecewise polynomials of a higher degree than those in H_1^μ, with respect to the same mesh $\mathcal{T}_{\ell^\mu}$. Exclude basis functions associated with nodes with respect to which basis functions of H_1^μ are defined.

For both options, the spaces H_2^μ can be constructed with locally or globally defined functions (recall similar Examples in Sections 5.4.1 and 5.4.2).

We now choose the set J_Q and the space $H \subset H_0^1(D)$ that define the tensor–product spaces

$$\mathbf{Y}_2 := \left\{ Y_2^\nu;\ \nu \in J_Q \right\} \tag{6.23}$$

in (6.16). We choose J_Q as in (5.66), where Δ_M is the number of additional parameters we wish to activate, and $H = H_1^{\bar{\mu}}$ for some $\bar{\mu} \in J_P$. That is, we choose H to be one of the FEM spaces already used in the definition of X in (6.1). We make this choice for several important reasons:

1. Choosing H in this way straightforwardly ensures that the spaces $\mathbf{Y}_2$ are automated as X changes at each step of an adaptive algorithm.

2. In an adaptive algorithm, the spaces $Y_2^\nu = H \otimes Q^\nu$ serve as candidates with which to augment X. If X is augmented by Y_2^ν for some $\nu \in J_Q$, choosing $H = H_1^{\bar{\mu}}$ maintains the structure of Y in (6.14) at the next adaptive step.

3. The cost of computing the parametric estimate $\|e_{Y_2}\|_{B_0}$ is easily balanced against the effectivity of the total estimate $\|e_Y\|_{B_0}$ by choosing a suitable multi–index $\bar{\mu} \in J_P$.

4. The coefficient matrix associated with the parametric problems (6.18) coincides with the matrix $K^0_{\bar{\mu}\bar{\mu}}$, which was computed to generate the SGFEM approximation $u_X \in X$. More generally, many of the matrices $K^m_{\nu\mu}$ appear on the right–hand side of the problems (6.18) through the term $B(u_X, v)$.

Note that through our choice of Y in (6.14), the constant β in (3.20) depends on $\bar{\mu}$. We seek a multi-index $\bar{\mu} \in J_P$ that strikes the right balance between the cost of computing the parametric error estimator and the effectivity of the total estimate $||e_Y||_{B_0}$. If we choose $\bar{\mu}$ such that $\ell^{\bar{\mu}} = \max_{\mu \in J_P} \boldsymbol{\ell}$, i.e., choose $H = H_1^{\bar{\mu}}$ to be the richest FEM space used so far, $||e_{Y_1}||_{B_0}$ may be disproportionately expensive to compute, especially if $\ell^\mu = \ell^{\bar{\mu}}$ for very few multi–indices $\mu \in J_P$. Moreover, $\dim(X)$ will grow too quickly when we augment X with spaces in $\mathbf{Y}_2$, and an adaptive algorithm is unlikely to realise the optimum rates of convergence. Conversely, if $\ell^{\bar{\mu}} = \min_{\mu \in J_P} \boldsymbol{\ell}$ the computation of $||e_{Y_1}||_{B_0}$ may be wasteful as its contribution to the effectivity of the total error estimate may be negligible. As a result, the error reduction associated with augmenting X with spaces from $\mathbf{Y}_2$ may also be negligible. To strike a balance, we will choose $\bar{\mu}$ to correspond to the FEM space H_1^μ with the smallest mesh level number ℓ^μ such that the number of spaces with level number ℓ^μ or less is greater than or equal to $\lceil \frac{1}{2}\text{card}(J_P) \rceil$. We denote this choice by $\bar{\mu} = \arg\,\text{avg}_{\mu \in J_P} \boldsymbol{\ell}$.

Example 6.2.

Suppose $\mathrm{card}(J_P) = 4$ and $\boldsymbol{\ell} = \{5, 3, 3, 4\}$, then $\ell^{\bar{\mu}} = 3$. Similarly, if $\mathrm{card}(J_P) = 5$ and $\boldsymbol{\ell} = \{5, 5, 3, 4, 3\}$, then $\ell^{\bar{\mu}} = 4$.

Although we have no formal proof, experiments provided in Section 6.4.3 suggest that the heuristic choice $\bar{\mu} = \arg\,\mathrm{avg}_{\mu \in J_P} \boldsymbol{\ell}$ performs well. Note that there are many other possibilities; $\bar{\mu}$ here ensures that the dimensions of the spaces in $\mathbf{Y}_2$ are always modest in comparison to those of the spaces $H_1^\mu \otimes P^\mu$ in the definition of X in (6.1).

6.4 Adaptive Multilevel SGFEMs

In this section we design a new adaptive multilevel SGFEM algorithm that is steered by analysing the sets of component error estimates

$$\mathbf{E}_{Y_1} := \big\{ \|e_{Y_1}^\mu\|_{B_0};\ \mu \in J_P \big\}, \qquad \mathbf{E}_{Y_2} := \big\{ \|e_{Y_2}^\nu\|_{B_0};\ \nu \in J_Q \big\}. \tag{6.24}$$

Suppose that the multi–indices $J_P \subset J$ and the finite element spaces $\mathbf{H}_1$ have been chosen and that the associated SGFEM approximation $u_X \in X$ satisfying (6.6) has been computed for the multilevel space X in (6.1). Suppose as well that the multi–indices $J_Q \subset J^*$ (for J^* defined in (5.65)) and the finite element spaces $\mathbf{H}_2$ and H have been chosen, and that the associated error estimator $e_Y \in Y$ satisfying (6.12) has been computed for the multilevel space Y in (6.14). If $\|e_Y\|_{B_0}$ is deemed too large, we need to augment X with new functions that will result in an improved SGFEM approximation to $u \in V$. In the following, we demonstrate that by analysing the available sets of component estimates $\mathbf{E}_{Y_1}$ and $\mathbf{E}_{Y_2}$ in (6.24), we can estimate the error reduction that would be achieved by augmenting X with specific subsets of $\mathbf{Y}_1$ and $\mathbf{Y}_2$ (defined in (6.22) and (6.23)), respectively, and computing a new approximation $u_X \in X$. Key to achieving the best rates of convergence is the inclusion of functions in X that lead to significant error reductions as well as the exclusion of functions that lead to negligible reductions.

Estimated Error Reduction Ratios

Suppose that we consider augmenting X with a subset of spaces $\bar{\mathbf{Y}}_1 \subseteq \mathbf{Y}_1$ or $\bar{\mathbf{Y}}_2 \subseteq \mathbf{Y}_2$. The elements of $\mathbf{Y}_1$ (Y_1^μ) are indexed by the multi–indices $\mu \in J_P$. Thus, a subset of

$\mathbf{Y}_1$ is simply determined by choosing a subset $\bar{J}_P \subseteq J_P$. Likewise, a subset of $\mathbf{Y}_2$ is determined by choosing a subset $\bar{J}_Q \subseteq J_Q$. Consider now the augmentation spaces

$$\bar{Y}_1 := \bigoplus_{\mu\in\bar{J}_P} Y_1^\mu, \qquad \bar{Y}_2 := \bigoplus_{\nu\in\bar{J}_Q} Y_2^\nu, \tag{6.25}$$

and the discrete problems

$$\text{find } u_{W_1} \in W_1 : \qquad B(u_{W_1}, v) = F(v), \qquad \text{for all } v \in W_1, \tag{6.26}$$

$$\text{find } u_{W_2} \in W_2 : \qquad B(u_{W_2}, v) = F(v), \qquad \text{for all } v \in W_2, \tag{6.27}$$

where

$$W_1 := X \oplus \bar{Y}_1 = \Bigg(\bigoplus_{\mu\in J_P\setminus\bar{J}_P} H_1^\mu \otimes P^\mu \Bigg) \oplus \Bigg(\bigoplus_{\mu\in\bar{J}_P} (H_1^\mu \oplus H_2^\mu) \otimes P^\mu \Bigg),$$

and

$$W_2 := X \oplus \bar{Y}_2 = X \oplus \Bigg(\bigoplus_{\nu\in\bar{J}_Q} H_1^{\bar{\mu}} \otimes Q^\nu \Bigg).$$

That is, W_1 corresponds to enriching FEM spaces H_1^μ associated with certain solution modes in the definition of $u_X \in X$, namely, those modes associated with the multi–indices $\mu \in \bar{J}_P$. Conversely, W_2 corresponds to adding new solution modes associated with $\bar{J}_Q$ to the definition of u_X, with associated FEM spaces $H_1^{\bar{\mu}}$.

As in Chapter 5, we find through Galerkin orthogonality that

$$\|e_{W_i}\|_B^2 = \|u - u_X\|_B^2 - \|u_{W_i} - u_X\|_B^2, \qquad i = 1, 2,$$

where $e_{W_i} = u - u_{W_i}$ denotes the error associated with the augmented space W_i. Thus, $\|u_{W_i} - u_X\|_B^2$ characterises the reduction in $\|u - u_X\|_B^2$ that would be achieved by augmenting X with $\bar{Y}_i$ and computing an enhanced approximation u_{W_i} satisfying (6.26) for $i = 1$ or (6.27) for $i = 2$. The following result provides estimates for these quantities and is a simple extension of Theorem 5.2.

Theorem 6.1.

Let $u_X \in X$ satisfy (6.6) *and let the enhanced approximations $u_{W_1} \in W_1$, $u_{W_2} \in W_2$ satisfy* (6.26) *and* (6.27)*, respectively. Define*

$$e_{\bar{Y}_1} := \sum_{\mu\in\bar{J}_P} e_{Y_1}^\mu, \qquad e_{\bar{Y}_2} := \sum_{\nu\in\bar{J}_Q} e_{Y_2}^\nu,$$

for some subsets $\bar{J}_P \subseteq J_P$ *and* $\bar{J}_Q \subseteq J_Q$ *so that*

$$\|e_{\bar{Y}_1}\|_{B_0}^2 = \sum_{\mu\in\bar{J}_P} \|e_{Y_1}^{\mu}\|_{B_0}^2, \qquad \|e_{\bar{Y}_2}\|_{B_0}^2 = \sum_{\nu\in\bar{J}_Q} \|e_{Y_2}^{\nu}\|_{B_0}^2. \tag{6.28}$$

Then, the following estimates hold:

$$\lambda\|e_{\bar{Y}_1}\|_{B_0}^2 \leq \|u_{W_1} - u_X\|_B^2 \leq \frac{\Lambda}{1-\gamma^2}\|e_{\bar{Y}_1}\|_{B_0}^2, \tag{6.29}$$

$$\lambda\|e_{\bar{Y}_2}\|_{B_0}^2 \leq \|u_{W_2} - u_X\|_B^2 \leq \Lambda\|e_{\bar{Y}_2}\|_{B_0}^2, \tag{6.30}$$

where λ, Λ *are defined in* (5.20)*, and* $\gamma \in [0,1)$ *satisfies* (3.18)*.*

We now determine an appropriate enrichment strategy for X by considering the bounds (6.29)–(6.30). A key ingredient of this strategy is the determination of suitable sets of multi–indices $\bar{J}_P$ and $\bar{J}_Q$. We only want to consider important subsets of J_P and J_Q that would result in worthwhile error reductions when augmenting X with the corresponding spaces $\bar{Y}_1$ or $\bar{Y}_2$ in (6.25). Assuming for now that $\bar{J}_P$ and $\bar{J}_Q$ are suitable, we explain how to use the estimates $\|e_{\bar{Y}_1}\|_{B_0}$ and $\|e_{\bar{Y}_2}\|_{B_0}$ to steer the enrichment of X. The appropriate selection of $\bar{J}_P$ and $\bar{J}_Q$ is the focus of Section 6.4.2.

One option is to perform the enrichment corresponding to $\max\{\|e_{\bar{Y}_1}\|_{B_0}, \|e_{\bar{Y}_2}\|_{B_0}\}$. Whilst this may lead to large reductions of $||u - u_X||_B$, it does not take into account the computational cost incurred. We want to construct sequences of SGFEM spaces $\{X\}$ for which the energy error decays to zero at the best possible rate with respect to N_X (for the underlying sequence of finite element spaces $\{H^{(i)};\ i = 0, 1, \dots\}$). This strategy is unlikely to achieve that, hence, the dimensions of the augmentation spaces $\bar{Y}_1$ and $\bar{Y}_2$ associated with the sets $\bar{J}_P$ and $\bar{J}_Q$ should be taken into account.

Recall that $N_{Y_1}^{\mu} = \dim(Y_1^{\mu})$ and $N_{Y_2}^{\nu} = \dim(Y_2^{\nu})$. The dimensions of the spaces $\bar{Y}_1$ and $\bar{Y}_2$ are hence given by

$$N_{\bar{Y}_1} := \dim(\bar{Y}_1) = \sum_{\mu\in\bar{J}_P} N_1^{\mu}, \qquad N_{\bar{Y}_2} := \dim(\bar{Y}_2) = \sum_{\nu\in\bar{J}_Q} N_2^{\nu},$$

respectively. Due to Theorem 6.1, the ratios

$$R_{\bar{Y}_i} := \frac{\|e_{\bar{Y}_i}\|_{B_0}^2}{N_{\bar{Y}_i}}, \qquad i = 1, 2, \tag{6.31}$$

provide approximations to the square of the true error reductions per additional DOF.

Algorithm 1: Adaptive multilevel SGFEM

Input : problem data $a(\mathbf{x},\mathbf{y})$, $f(\mathbf{x})$; initial index set J_P and mesh level numbers $\boldsymbol{\ell}$; energy error tolerance ϵ.

Output: final SGFEM approximation u_X and energy error estimate η.

choose `version` (1 or 2)
for $k = 0, 1, 2, \ldots$ **do**
  $u_X \leftarrow \texttt{SOLVE}\big[a, f, J_P, \boldsymbol{\ell}\big]$
  $J_Q \leftarrow \texttt{PARAMETRIC\_INDICES}\big[J_P\big]$ see (5.66)
  $\mathbf{E}_{Y_1} \leftarrow \texttt{COMPONENT\_SPATIAL\_ERRORS}\big[u_X, J_P, \boldsymbol{\ell}\big]$ (6.24)
  $\mathbf{E}_{Y_2} \leftarrow \texttt{COMPONENT\_PARAMETRIC\_ERRORS}\big[u_X, J_Q, \boldsymbol{\ell}\big]$
  $\eta = \big[\sum_{\mu\in J_P} ||e^{\mu}_{Y_1}||^2_{B_0} + \sum_{\nu\in J_Q} ||e^{\nu}_{Y_2}||^2_{B_0}\big]^{\frac{1}{2}}$ (6.19)
  if $\eta < \epsilon$ **then**
    return u_X, η
  else
    $[\texttt{enrichment\_type}, \bar{J}] \leftarrow \texttt{ENRICHMENT\_INDICES}\big[\texttt{version}, \mathbf{E}_{Y_1}, \mathbf{E}_{Y_2}, J_P, J_Q\big]$
    if `enrichment_type` $=$ `spatial` **then**
      $\boldsymbol{\ell} \rightarrow \big\{\ell^{\mu+};\ \mu \in \bar{J}\big\} \cup \big\{\ell^{\mu};\ \mu \in J_P \backslash \bar{J}\big\}$ (6.32)
    else
      $J_P \rightarrow J_P \cup \bar{J}$
      $\boldsymbol{\ell} \rightarrow \boldsymbol{\ell} \cup \big\{\ell^{\bar{\mu}};\ \nu \in \bar{J}\big\}$
    end
  end
end

That is

$$R_{\bar{Y}_i} \approx \frac{\|u_{W_i} - u_X\|^2_B}{N_{\bar{Y}_i}}, \qquad i = 1, 2.$$

For suitable sets $\bar{J}_P \subset J_P$ and $\bar{J}_Q \subset J_Q$, we are thus able to perform the enrichment strategy corresponding to the maximum estimated reduction in $||u - u_X||^2_B$ per additional DOF. In other words, we augment X with the space $\bar{Y}_1$ or $\bar{Y}_2$ corresponding to $\max\{R_{\bar{Y}_1}, R_{\bar{Y}_2}\}$.

6.4.1 An Adaptive Algorithm

We now use the error estimation strategy discussed in Section 6.3 and the resulting estimated ratios (6.31) to propose an adaptive multilevel SGFEM for the numerical solution of (5.6)–(5.7). Starting from an initial SGFEM space X of the form (6.1), our algorithm generates a sequence of spaces $\{X\}$ and terminates when $\eta = \|e_Y\|_{B_0} \leq \epsilon$ where ϵ denotes a prescribed error tolerance. Assuming that the polynomial degree of the FEM approximation on D has been fixed, to compute the initial approximation

$u_X \in X$ we need only supply an initial set of multi–indices $J_P \subset J$ and a corresponding set of mesh level numbers $\boldsymbol{\ell} = \{\ell^\mu;\ \mu \in J_P\}$. Typically, we choose an initial small level number ℓ_0 and set $\ell^\mu = \ell_0$ for all $\mu \in J_P$. Once the corresponding error estimate $||e_Y||_{B_0}$ has been computed, we then implement either a spatial or parametric enrichment of X, corresponding to $\max\{R_{\bar{Y}_1}, R_{\bar{Y}_2}\}$.

If $\max\{R_{\bar{Y}_1}, R_{\bar{Y}_2}\} = R_{\bar{Y}_1}$ we enrich a subset of the FEM spaces in $\mathbf{H}_1$. Specifically, we refine the meshes $\mathcal{T}_{\ell^\mu}$ associated with the multi–indices $\bar{J}_P$ and update the set $\boldsymbol{\ell}$. That is, if $\ell^\mu = i$ for some $\mu \in \bar{J}_P$, we set $\ell^\mu \to i+1$ or equivalently replace $\mathcal{T}_{\ell^\mu}$ with the next mesh in the sequence $\boldsymbol{\mathcal{T}}$ in (6.5). To represent this process, in our adaptive algorithm we write

$$\ell^\mu \to \ell^{\mu+}, \qquad \text{for all } \mu \in \bar{J}_P. \tag{6.32}$$

Conversely, if $\max\{R_{\bar{Y}_1}, R_{\bar{Y}_2}\} = R_{\bar{Y}_2}$, we add $\bar{J}_Q$ to the set J_P. In this case, we also update $\boldsymbol{\ell}$ and $\mathbf{H}_1$ with $\text{card}(\bar{J}_Q)$ copies of $\bar{\mu}$ and $H_1^{\bar{\mu}}$, respectively, which maintains the relationship $\text{card}(J_P) = \text{card}(\boldsymbol{\ell})$. In the spirit of (5.64), the process is repeated once updated versions of J_P and $\boldsymbol{\ell}$ are defined and a new SGFEM approximation $u_X \in X$ is computed. The general process is outlined in Algorithm 1 – at a given step k:

- `SOLVE` computes an SGFEM approximation $u_X \in X$ to $u \in V$ satisfying (6.6).
- `PARAMETRIC_INDICES` uses (5.66) to determine the subset J_Q of the neighbouring indices to J_P for a prescribed choice of Δ_M.
- `COMPONENT_SPATIAL_ERRORS` and `COMPONENT_PARAMETRIC_ERRORS` compute the sets of error estimates $\mathbf{E}_{Y_1}$ and $\mathbf{E}_{Y_2}$ in (6.24), respectively, by solving (6.17) and (6.18).
- `ENRICHMENT_INDICES` analyses the sets $\mathbf{E}_{Y_1}$ and $\mathbf{E}_{Y_2}$ in conjunction with the formulae in (6.31) to construct suitable sets $\bar{J}_P$ and $\bar{J}_Q$ and determines how to enrich the current SGFEM space X.

6.4.2 Selection of Enrichment Indices

Here, we present two versions of the module `ENRICHMENT_INDICES` which we outline in Algorithm 2. To determine suitable subsets $\bar{J}_P \subset J_P$ and $\bar{J}_Q \subset J_Q$ we utilise the

Algorithm 2: `ENRICHMENT_INDICES` versions 1 and 2

Input : `version`; $\mathbf{E}_{Y_1}$; $\mathbf{E}_{Y_2}$; J_P; J_Q.
Output: `enrichment_type`, $\bar{J}$.

$\delta_{Y_1} = \max_{\mu\in J_P} \mathbf{R}_{Y_1}$, $\delta_{Y_2} = \max_{\nu\in J_Q} \mathbf{R}_{Y_2}$
if $\delta_{Y_1} > \delta_{Y_2}$ **then**
  $\bar{J}_Q = \{\nu \in J_Q;\ R^{\nu}_{Y_2} = \delta_{Y_2}\}$
  if `version` $= 1$ **then**
    $\bar{J}_P = \{\mu \in J_P;\ R^{\mu}_{Y_1} > \delta_{Y_2}\}$
  else
    $\bar{J}_P \leftarrow$ `MARK`$[\mathbf{E}_{Y_1}, \mathbf{N}_{Y_1}, \delta_{Y_2}]$
  end
else
  $\bar{J}_P = \{\mu \in J_P;\ R^{\mu}_{Y_1} = \delta_{Y_1}\}$
  if `version` $= 1$ **then**
    $\bar{J}_Q = \{\nu \in J_Q;\ R^{\nu}_{Y_2} > \delta_{Y_1}\}$
  else
    $\bar{J}_Q \leftarrow$ `MARK`$[\mathbf{E}_{Y_2}, \mathbf{N}_{Y_2}, \delta_{Y_1}]$
  end
end
if $R_{\bar{Y}_1} > R_{\bar{Y}_2}$ **then**
  `enrichment_type` $=$ `spatial`, $\bar{J} = \bar{J}_P$
else
  `enrichment_type` $=$ `parametric`, $\bar{J} = \bar{J}_Q$
end
return [`enrichment_type`, $\bar{J}$]

dimensions of the spaces Y_1^{μ} and Y_2^{ν} in the sets $\mathbf{Y}_1$ and $\mathbf{Y}_2$ (defined in (6.22)–(6.23)). In Algorithm 2, the dimensions of Y_1^{μ} and Y_2^{ν} are stored in the sets $\mathbf{N}_{Y_1} := \{N_1^{\mu};\ \mu \in J_P\}$ and $\mathbf{N}_{Y_2} := \{N_2^{\nu};\ \nu \in J_Q\}$.

Firstly, we define the sets of estimated error reduction ratios

$$\mathbf{R}_{Y_1} := \left\{R^{\mu}_{Y_1};\ \mu \in J_P\right\}, \qquad \mathbf{R}_{Y_2} := \left\{R^{\nu}_{Y_2};\ \nu \in J_Q\right\}, \tag{6.33}$$

with elements

$$R^{\mu}_{Y_1} := \frac{\|e^{\mu}_{Y_1}\|^2_{B_0}}{N^{\mu}_{Y_1}}, \qquad \mu \in J_P, \qquad R^{\mu}_{Y_2} := \frac{\|e^{\nu}_{Y_2}\|^2_{B_0}}{N^{\nu}_{Y_2}}, \qquad \nu \in J_Q.$$

In both versions of `ENRICHMENT_INDICES` we choose the sets $\bar{J}_P$ and $\bar{J}_Q$ by considering the spaces Y_1^{μ} for $\mu \in J_P$ and Y_2^{ν} for $\nu \in J_Q$ that offer the most favourable estimated error reduction ratios $R^{\mu}_{Y_1}$ and $R^{\nu}_{Y_2}$. We simply look at the largest ratios

$$\delta_{Y_1} := \max_{\mu\in J_P} \mathbf{R}_{Y_1}, \qquad \delta_{Y_2} := \max_{\nu\in J_Q} \mathbf{R}_{Y_2},$$

and use $\min\{\delta_{Y_1}, \delta_{Y_2}\}$ as a tolerance. This principle underpins both versions of the module `ENRICHMENT_INDICES`, which we now introduce.

For Version 1, if $\delta_{Y_1} > \delta_{Y_2}$ so that δ_{Y_2} is the acting tolerance, we set

$$\bar{J}_P = \big\{\mu \in J_P;\ R^{\mu}_{Y_1} \geq \delta_{Y_2}\big\}, \qquad \bar{J}_Q = \big\{\nu \in J_Q;\ R^{\nu}_{Y_2} = \delta_{Y_2}\big\}.$$

Likewise, if $\delta_{Y_2} > \delta_{Y_1}$ so that δ_{Y_1} is the acting tolerance, we set

$$\bar{J}_P = \big\{\mu \in J_P;\ R^{\mu}_{Y_1} = \delta_{Y_1}\big\}, \qquad \bar{J}_Q = \big\{\nu \in J_Q;\ R^{\nu}_{Y_2} \geq \delta_{Y_1}\big\}.$$

The type of enrichment is determined by computing and comparing the ratios $R_{\bar{Y}_1}$ and $R_{\bar{Y}_2}$ defined in (6.31). If $R_{\bar{Y}_1} > R_{\bar{Y}_2}$ we perform spatial enrichment and set $\bar{J} = \bar{J}_P$. Thus, X is augmented with $\bar{Y}_1$ at the next step of Algorithm 1. Otherwise, we perform parametric enrichment and set $\bar{J} = \bar{J}_Q$ so that X is augmented with $\bar{Y}_2$.

For version 2, if $\delta_{Y_1} > \delta_{Y_2}$, we choose $\bar{J}_P$ to be the largest subset of J_P such that $R_{\bar{Y}_1} > \delta_{Y_2}$ (recall $R_{\bar{Y}_1}$ depends on $\bar{J}_P$). Similarly, if $\delta_{Y_2} > \delta_{Y_1}$, we choose $\bar{J}_Q$ to be the largest subset of J_Q such that $R_{\bar{Y}_2} > \delta_{Y_1}$. Again, the enrichment type chosen is the one associated with $\max\{R_{\bar{Y}_1}, R_{\bar{Y}_2}\}$. Version 2 is reminiscent of a Dörfler marking strategy [41] and so the module that generates $\bar{J}_P$ (if $\delta_{Y_1} > \delta_{Y_2}$) and $\bar{J}_Q$ (if $\delta_{Y_2} > \delta_{Y_1}$) is called `MARK`. We stress however, that for both versions of `ENRICHMENT_INDICES`, no marking or tuning parameters are required.

6.4.3 Numerical Experiments

We now test the performance of Algorithms 1 and 2 by solving the four test problems (TP1–TP4) described in Section 5.1.1. Specifically, using Algorithms 1–2 we generate sequences of SGFEM approximations $\{u_X\}$ by solving the finite–dimensional problem (6.6) for sequences of adaptively constructed multilevel SGFEM spaces $\{X\}$. To begin our investigation we initialise Algorithm 1 by constructing an initial space X for each test problem. The following results were originally published in [38].

First, we select an appropriate sequence of FEM spaces

$$\mathbf{H} := \big\{H^{(i)};\ i = 0, 1, \dots\big\}, \qquad H^{(0)} \subset H^{(1)} \subset \cdots \subset H^1_0(D).$$

Since D is square for all four test problems, we choose the underlying sequence $\boldsymbol{\mathcal{T}}$ in (6.5) to be uniform meshes of square elements. Indeed, we choose $\mathcal{T}_i$ to represent a

$2^i \times 2^i$ grid over D for $i = 0, 1, 2 \ldots$ so that $\mathcal{T}_{i+1}$ represents a uniform refinement of $\mathcal{T}_i$. Thus, $\mathcal{T}_i$ has element width $h(i) := 2^{1-i}$ for test problem TP1 and $h(i) := 2^{-i}$ for test problems TP2–TP4. We choose $\mathbf{H}$ to be the set of $\mathbb{Q}_1$ FEM spaces associated with $\boldsymbol{\mathcal{T}}$ and initialise Algorithm 1 with

$$J_P = \big\{(0, 0, \ldots), (1, 0, \ldots)\big\}, \qquad \boldsymbol{\ell} = \big\{4, 4\big\}$$

(corresponding to the polynomials $\psi_1(\mathbf{y}) = 1$ and $\psi_2(\mathbf{y}) = y_1$). In other words, for each $\mu \in J_P$ we choose $H_1^\mu = H^{(4)}$ in the definition of $\mathbf{H}_1$ in (6.4). The sequence $\boldsymbol{\mathcal{T}}$ could also be constructed using uniform meshes of triangular elements. Additionally, the meshes $\boldsymbol{\mathcal{T}}$ need not be uniform. Shishkin meshes [95, 113, 65] are a straightforward alternative to uniform meshes (for problems with boundary layers) and Algorithms 1 and 2 would not change in that case.

To compute the error estimate η in (6.19) we now choose the FEM spaces $\mathbf{H}_2$ in (6.13) that define problems (6.17). Following the adaptive single–level experiments conducted in Example 5.11, we construct each H_2^μ using the global $\mathbb{Q}_2$ basis functions defined with respect to the element–edge midpoints and centroids of $\mathcal{T}_{\ell^\mu}$ (the mesh associated with H_1^μ) for $\mu \in J_P$. We also fix $\Delta_M = 5$ in the definition of J_Q in (5.66) that defines problems (6.18). To test the performance of $\eta = \|e_Y\|_{B_0}$ at each step of Algorithm 1, we compute effectivity indices $\theta_{\text{eff}}^{\text{approx}}$ by replacing u in (3.21) with a surrogate reference solution u_{ref}, generated by applying Algorithm 1 and version 1 of Algorithm 2 with a very small tolerance ϵ_{ref}.

For our first example, we investigate the rate of convergence of the energy error.

Example 6.3: Rates of Convergence.

We solve test problems TP1–TP4 using Algorithm 1 with $\epsilon = 2 \times 10^{-3}$ and version 1 of Algorithm 2. In Figure 6.5 we plot $\eta = \|e_Y\|_{B_0}$ versus the number of DOFs N_X (left) and the corresponding effectivity indices $\theta_{\text{eff}}^{\text{approx}}$ at each step k of Algorithm 1 (right) for TP1–TP4 (top–to–bottom). We observe that the error behaves like $N_X^{-1/3}$ for TP1 and like $N_X^{-1/2}$ for TP2–TP4. Compare this with the inferior rates of $-\frac{1}{5}$ and $-\frac{1}{3}$ realised by the single–level method in Example 5.11 for TP1 and TP2. Additionally, the effectivity indices are close to one for all four problems, meaning that the error estimate $\eta = \|e_Y\|_{B_0}$ is highly accurate.

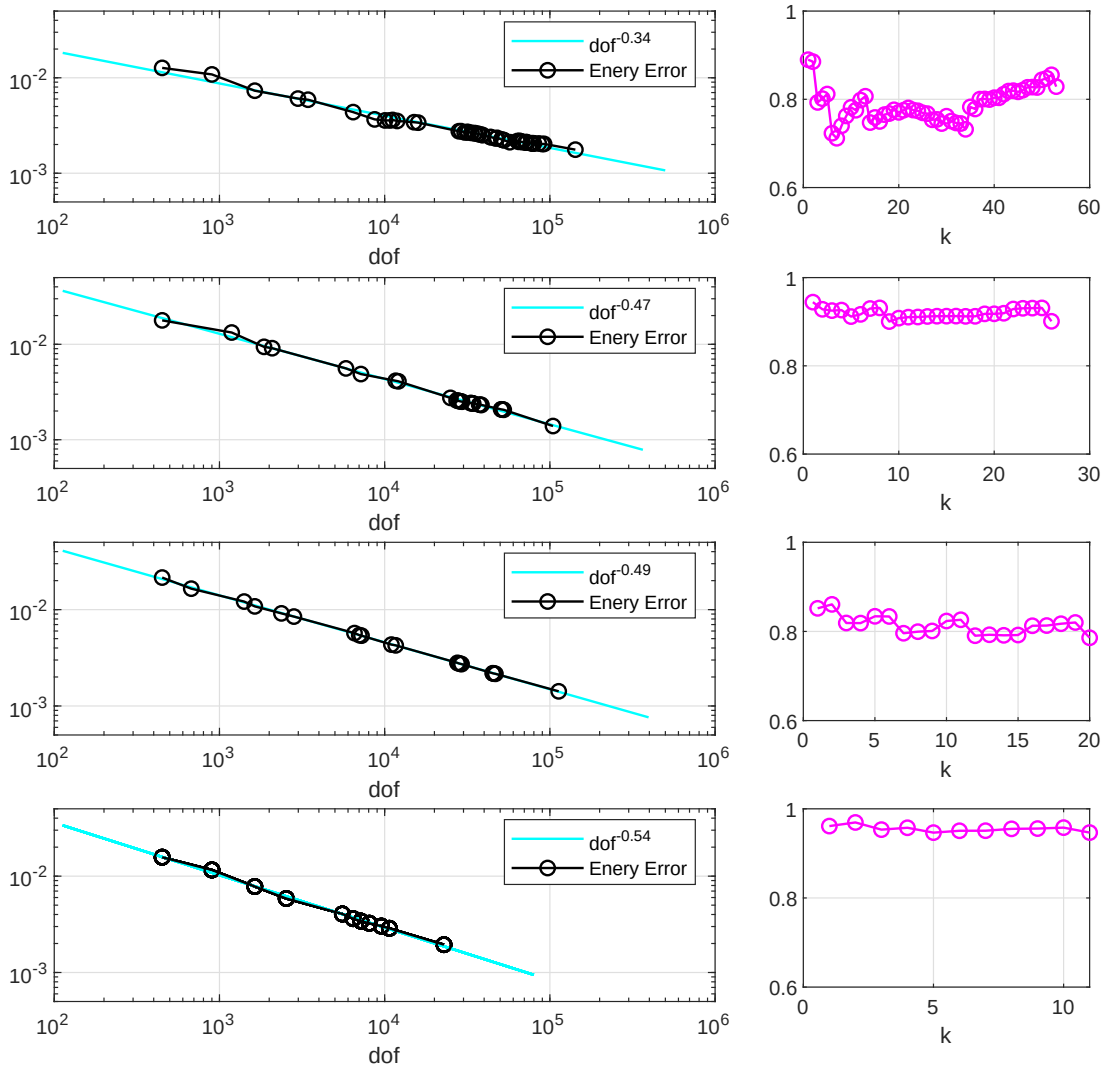
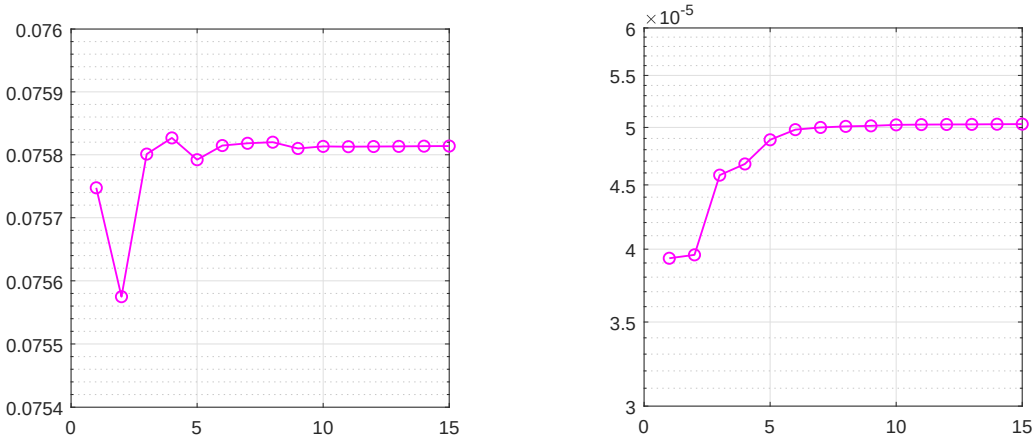


Figure 6.5: Plots of the convergence of $\eta = \|e_Y\|_{B_0}$ versus the number of DOFs N_X (left) and effectivity indices $\theta_{\text{eff}}^{\text{approx}}$ (right) for Example 6.3 when solving test problems TP1–TP4 (top–to–bottom) using Algorithms 1 and 2 (version 1).

The observed rate of $-\frac{1}{2}$ with respect to N_X in Example 6.3 is optimal for the given sequence $\mathbf{H}$ of $\mathbb{Q}_1$ FEM spaces. In other words, the sequences in (6.2) decay quickly enough and the corresponding value of p is small enough for the rate afforded to the chosen FEM for the analogous parameter–free problem to be realised. Note that for TP1, it is difficult to determine whether the observed rate of $-\frac{1}{3}$ is optimal or not. On the one hand, the true solution $u \in V$ may not be afforded enough spatial regularity and the rate may be optimal given the set $\mathbf{H}$. On the other, $u \in V$ may have enough spatial regularity, but the sequences in (6.2) decay too slowly. Additionally, Assumption 5.2 may simply not hold for TP1 (due to our choices of $\mu = 1$ and $\sigma = 0.15$), despite the well–posedness of (6.6) following the implicit truncation of the



(a) (b)

Figure 6.6: The convergence of $\max_{\mathbf{x}\in D} \mathbb{E}[u_X]$ (a) and $\max_{\mathbf{x}\in D} \mathrm{Var}(u_X)$ (b) at each step of Algorithm 1 for TP2 in Example 6.3.

sum in (5.9) after M terms.

Two potential QoIs are $\max_{\mathbf{x}\in D} \mathbb{E}[u]$ and $\max_{\mathbf{x}\in D} \mathrm{Var}(u)$. In Figure 6.6 we plot the values of $\max_{\mathbf{x}\in D} \mathbb{E}[u_X]$ and $\max_{\mathbf{x}\in D} \mathrm{Var}(u_X)$ at each step of Algorithm 1 for TP2. Notice how both QoIs converge after approximately ten steps. It is also possible to develop adaptive SGFEM algorithms which reduce $Q(u) - Q(u_X)$ in the most efficient way, instead of $\|u - u_X\|_B$, where $Q(u)$ denotes a QoI involving $u \in V$. Such methods often require an effective energy error estimate (such as those considered in this thesis) and employ a stopping criterion appropriate for the QoI considered; see [20] for an example in the single–level SGFEM framework.

We now investigate the qualitative differences between the spaces X associated with Example 6.3 at the final step of Algorithm 1 for test problems TP1–TP4. In Table 6.2 we record the final number of active parameters M and multi–indices $\mathrm{card}(J_P)$ used in the definition of X at the final step. We also record the number of multi–indices $\mu \in J_P$ that are assigned the same FEM space H_1^μ from the set $\mathbf{H}$, or equivalently, that are assigned the same mesh level number ℓ^μ with associated element width $h(\ell^\mu)$. Notice that for all four test problems, the finest mesh employed from the sequence $\boldsymbol{\mathcal{T}}$ is assigned to only a single multi–index in J_P. Conversely, the coarsest mesh employed is assigned to the greatest number of multi–indices. For TP1, approximately 70% of all multi–indices in J_P are assigned the mesh $\mathcal{T}_4 \in \boldsymbol{\mathcal{T}}$ with element width 2^{-3}. Due to the decay rates of α_m in (5.10) for TP2 and TP3, a larger number of parameters y_m are activated for TP2 ($M = 13$) than for TP3 ($M = 3$). Additionally, the slow

Table 6.2: Number of solution modes assigned the same element width $h(\ell^\mu)$ (corresponding to a mesh level number ℓ^μ in $\boldsymbol{\ell}$) at the final step of Algorithm 1 (with version 1 of Algorithm 2) as well as the final number of active parameters M and multi–indices $\mathrm{card}(J_P)$ for test problems TP1–TP4.

test problem	2^{-3}	2^{-4}	2^{-5}	2^{-6}	2^{-7}	2^{-8}	$\mathrm{card}(J_P)$	M
TP1	118	49	1	0	1	0	169	93
TP2	–	25	6	3	1	1	36	13
TP3	–	5	7	2	2	1	17	3
TP4	–	17	3	0	1	0	21	8

Table 6.3: The first twelve multi-indices from the set J_P generated by Algorithm 1 (with version 1 of Algorithm 2) and the associated element widths $h(\ell^\mu)$ assigned to those multi-indices at the final step for test problems TP1–TP4.

TP1		TP2		TP3		TP4	
μ	$h(\ell^\mu)$	μ	$h(\ell^\mu)$	μ	$h(\ell^\mu)$	μ	$h(\ell^\mu)$
(0 0 0 0 0 0 0 0 0 0)	2^{-7}	(0 0 0 0 0 0)	2^{-8}	(0 0 0)	2^{-8}	(0 0 0 0 0 0)	2^{-7}
(**1** 0 0 0 0 0 0 0 0 0)	2^{-5}	(**1** 0 0 0 0 0)	2^{-7}	(**1** 0 0)	2^{-7}	(**1** 0 0 0 0 0)	2^{-5}
(0 0 **1** 0 0 0 0 0 0 0)	2^{-4}	(0 0 **1** 0 0 0)	2^{-6}	(**2** 0 0)	2^{-7}	(0 0 **1** 0 0 0)	2^{-5}
(0 **1** 0 0 0 0 0 0 0 0)	2^{-4}	(0 **1** 0 0 0 0)	2^{-6}	(**3** 0 0)	2^{-6}	(0 **1** 0 0 0 0)	2^{-5}
(0 0 0 0 0 **1** 0 0 0 0)	2^{-4}	(**2** 0 0 0 0 0)	2^{-6}	(0 **1** 0)	2^{-5}	(0 0 0 **1** 0 0)	2^{-4}
(0 0 0 0 **1** 0 0 0 0 0)	2^{-4}	(**1 1** 0 0 0 0)	2^{-5}	(**4** 0 0)	2^{-6}	(**1** 0 **1** 0 0 0)	2^{-4}
(0 0 0 **1** 0 0 0 0 0 0)	2^{-4}	(0 0 0 0 0 **1**)	2^{-5}	(**1 1** 0)	2^{-5}	(**1 1** 0 0 0 0)	2^{-4}
(**2** 0 0 0 0 0 0 0 0 0)	2^{-3}	(0 0 0 0 **1** 0)	2^{-5}	(**5** 0 0)	2^{-5}	(**2** 0 0 0 0 0)	2^{-4}
(0 0 0 0 0 0 0 **1** 0 0)	2^{-4}	(0 0 0 **1** 0 0)	2^{-5}	(**2 1** 0)	2^{-5}	(0 0 0 0 0 **1**)	2^{-4}
(0 0 0 0 0 0 **1** 0 0 0)	2^{-4}	(**1** 0 **1** 0 0 0)	2^{-5}	(0 0 **1**)	2^{-5}	(0 0 0 0 **1** 0)	2^{-4}
(0 0 0 0 0 0 0 0 0 **1**)	2^{-4}	(**2 1** 0 0 0 0)	2^{-4}	(**3 1** 0)	2^{-5}	(**1** 0 0 **1** 0 0)	2^{-4}
(0 0 0 0 0 0 0 0 **1** 0)	2^{-4}	(**3** 0 0 0 0 0)	2^{-5}	(**6** 0 0)	2^{-5}	(0 **1 1** 0 0 0)	2^{-4}

decay of the terms ν_m in (5.9) leads to a large number of parameters being activated for TP1 ($M = 93$), and the number of multi–indices incorporated into the definition of X is significantly larger ($\mathrm{card}(J_P) = 169$) for TP1 than for TP2–TP4.

In Table 6.3 we display the first twelve multi–indices $\mu \in J$ that were chosen by Algorithms 1 and 2 (version 1) and added to the set J_P for test problems TP1–TP4. We also record the final element widths $h(\ell^\mu)$ assigned to those multi–indices. The multi–indices $\mu \in J_P$ that are selected in the early stages of Algorithm 1 are always assigned the finest meshes. This behaviour is to be expected since those multi–indices correspond to the most important solution modes with respect to the energy error. In particular, the *mean* solution mode associated with the multi–index $\mu = (0, 0, \dots)$ (recall (5.34)) is always allocated the finest mesh. Notice that the multi–indices added to J_P for TP1 mostly correspond to univariate polynomials of degree one, which activate more parameters y_m and terms in the expansion (5.9). In contrast, the multi–indices added to J_P for TP3 correspond mostly to higher degree polynomials

in the currently active parameters. Perhaps unexpectedly, the multi–indices selected for TP4 most resemble those for TP1, despite the fact that the sequence $\{\|a_m\|_\infty\}_{m=1}^\infty$ decays most quickly for this problem. That said, we observe from Table 6.2 that significantly fewer fine meshes are employed for TP4 than for TP3 and despite the increased number of active parameters ($M = 8$), TP4 may be computationally cheaper to solve than TP3.

In the next example we investigate the computational cost of solving test problems TP1–TP4 using Algorithm 1 and both versions of Algorithm 2. All computations were performed in MATLAB using new software developed from components of the S-IFISS toolbox [19] on an Intel Core i7 4770k 3.50Ghz CPU with 24GB of RAM.

Example 6.4: Computational timings.

We solve test problems TP1–TP4 using Algorithm 1 with decreasing values of ϵ and both versions of Algorithm 2. In Table 6.4 we record the cumulative time (T) in seconds and the number of adaptive steps (K) taken for $\eta = \|e_Y\|_{B_0}$ to satisfy each tolerance ϵ. At each step of Algorithm 1, the module `SOLVE` employs the preconditioned conjugate gradient (CG) method with a mean–based preconditioner [79] to solve (6.6). In particular, we employ the block–diagonal preconditioner A_0 given by

$$[A_0]_{\mu\mu} := K^0_{\mu\mu}, \qquad \mu \in J_P,$$

where $K^0_{\mu\mu}$ is parameter–free and defined in (6.9). We never construct A_0 explicitly and the action of A_0^{-1} on vectors is computed by applying the actions of $[A_0]^{-1}_{\mu\mu}$ blockwise. The number of CG iterations required is independent of the mesh level numbers $\boldsymbol{\ell}$ (the spatial discretisation) but does depend on the variance of $a(\mathbf{x}, \mathbf{y})$ [79]. Typically, more iterations are required as the constants $\lambda < 1 < \Lambda$ in (5.20) become smaller and larger, respectively.

Recall that test problems TP1–TP4 in Section 5.1.1 are ordered by the rate of decay of the associated sequences $\{\|a_m\|_\infty\}_{m=1}^\infty$ (slowest–to–fastest). We observe from Table 6.4 that for the same values of ϵ and both versions of Algorithm 2, TP4 is always the fastest problem to solve and TP1 is always the slowest. Additionally, due to the large number of matrices required, memory limitations mean that error tolerances below

Table 6.4: Solution times T (in seconds) and total adaptive step counts K required to solve test problems TP1–TP4 using Algorithms 1 and 2 (versions 1 and 2) with various choices of the error tolerances ϵ. The symbol '–' denotes that the estimated error at the previous step is already below the tolerance and the preceeding T and K are applicable.

	TP1				TP2				TP3				TP4			
	ver. 1		ver. 2		ver. 1		ver. 2		ver. 1		ver. 2		ver. 1		ver. 2	
ϵ	T	K	T	K	T	K	T	K	T	K	T	K	T	K	T	K
$4.5 \cdot 10^{-3}$	2	6	2	6	1	7	5	6	1	10	1	7	1	5	2	5
$3.0 \cdot 10^{-3}$	13	14	3	8	4	9	–	–	3	12	3	9	2	10	–	–
$1.5 \cdot 10^{-3}$	311	83	325	34	27	26	29	10	16	20	11	11	7	19	5	7
$9.0 \cdot 10^{-4}$	out of memory				236	70	167	13	87	36	62	15	23	29	22	8
$7.5 \cdot 10^{-4}$					–	–	–	–	100	38	–	–	36	38	–	–
$6.0 \cdot 10^{-4}$					881	147	–	–	147	44	92	18	110	48	80	9
$4.5 \cdot 10^{-4}$					2197	177	1306	19	484	61	340	22	158	59	95	10

$\epsilon = 1.5 \times 10^{-3}$ for TP1 could not be achieved. This is easily remedied by computing all the stiffness matrices required by the modules `SOLVE`, `COMPONENT_SPATIAL_ERRORS` and `COMPONENT_PARAMETRIC_ERRORS` at each step of Algorithm 1, rather than storing them in computer memory for the next step (recall from Table 6.1 that TP1 requires a significantly higher number of matrices than the other three test problems). However, this would inevitably increase the timings T recorded in Table 6.4.

We also observe that for TP2–TP4 and smaller tolerances ϵ, version 2 of Algorithm 2 results in considerably quicker solution times T and a lower step count K. The lower step count is due to the fact that the sets of multi–indices $\bar{J}$ produced by Algorithm 2 are usually richer for version 2 than for version 1. Note that a single step of Algorithm 1 with version 2 is also more expensive. Thus, time savings are only made when enough steps and calls to the modules `SOLVE`, `COMPONENT_SPATIAL_ERRORS` and `COMPONENT_PARAMETRIC_ERRORS` are saved. The timings for all four test problems can also be improved by computing the matrices $K^m_{\nu\mu}$ in parallel, since these are independent of each other. Similarly, the matrices G_m and the elements in the sets $\mathbf{E}_{Y_1}$ and $\mathbf{E}_{Y_2}$ in (6.24) may be computed in parallel as well.

In Figure 6.7 we plot the cumulative time T versus the number of DOFs N_X (left) at each step of Algorithm 1 with version 2 of Algorithm 2 for TP1–TP4 (top–to–bottom). The total number of markers, each reflecting a single step of Algorithm 1, equals the final value of K in Table 6.4. For example, Algorithm 1 takes 10 steps to satisfy the tolerance $\epsilon = 4.5 \times 10^{-4}$ for TP4, which corresponds to $T = 95$ and is represented by the tenth and final marker in Figure 6.7. Notice that for TP3 and TP4,

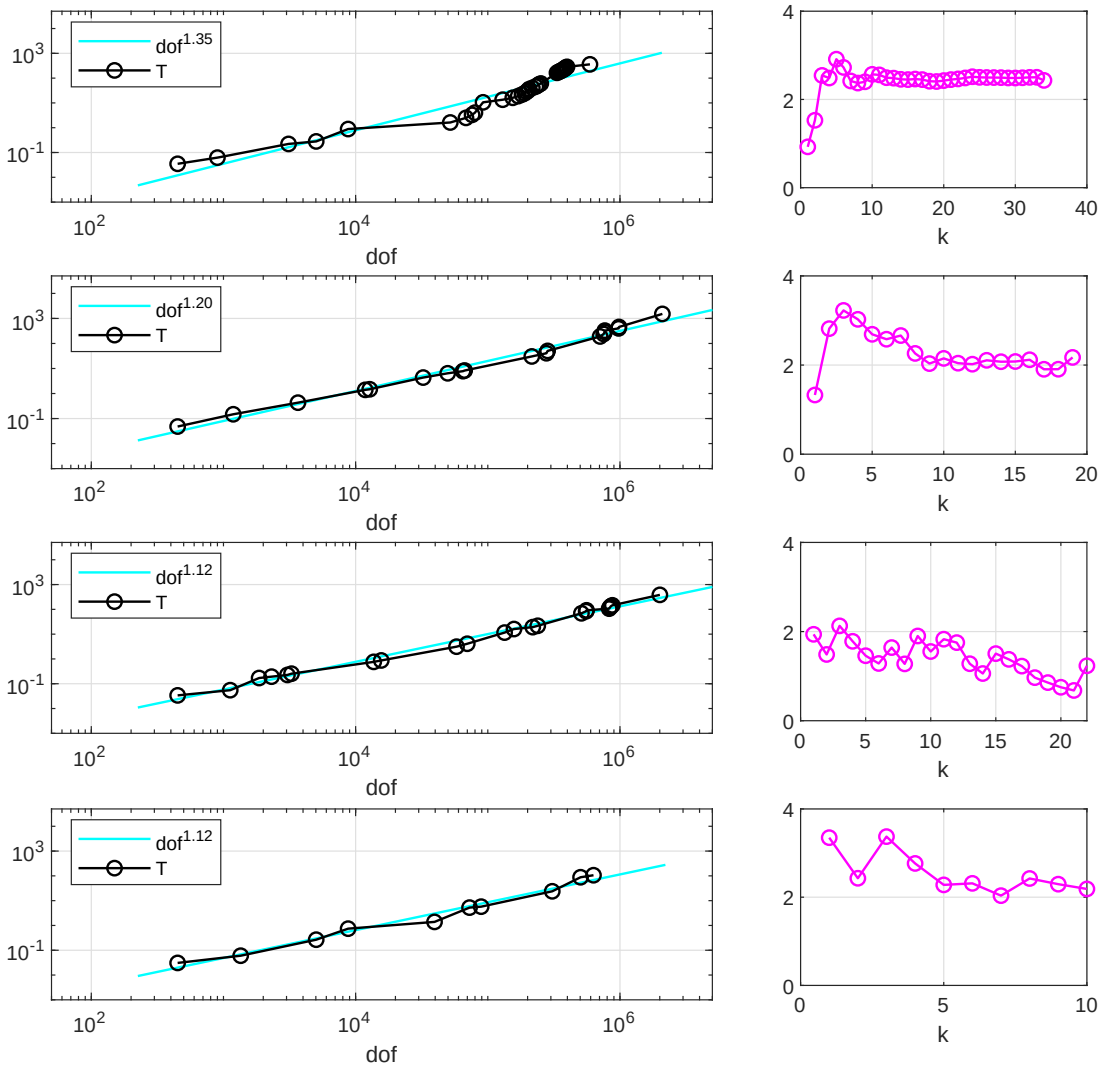


Figure 6.7: Plots of the total computational time T (left) in seconds accumulated over all refinement steps versus the number of DOFs N_X and the error estimation–solve time ratio r at each step k (right) when solving TP1–TP4 (top–to–bottom) using Algorithm 1 with version 2 of Algorithm 2.

Algorithm 1 performs almost optimally in that the time behaves almost linearly with respect to N_X. In contrast, for problems such as TP1 and TP2 where the sequence $\{\|a_m\|_\infty\}_{m=1}^\infty$ decays more slowly, the time T behaves less favourably, like $N_X^{1.35}$ and $N_X^{1.2}$, respectively. Whilst the rates 1.35 and 1.2 are not optimal, they are more than acceptable. We also plot the ratio r of the cumulative time taken to estimate the energy error (lines 5 and 6 of Algorithm 1) to the time taken to compute the corresponding approximation $u_X \in X$ (line 3 of Algorithm 1) at each step of Algorithm 1. We observe that r does not grow with N_X and thus the cost of estimating the error is proportional to the cost of computing the SGFEM approximation itself. At the final step of Algorithm 1, $r < 2.5$ for all four problems.

6.5 Extension to Localised Mesh Refinement

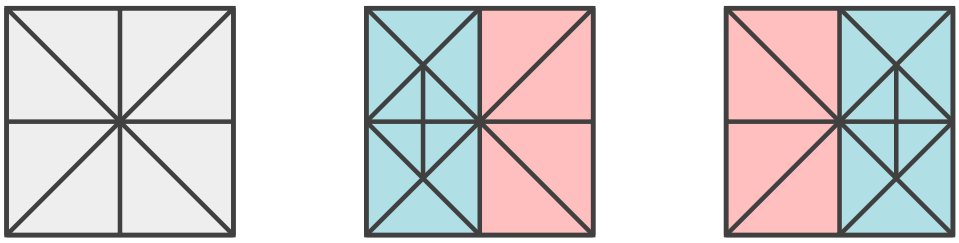
In this section, we briefly consider the parametric diffusion problem (5.6)–(5.7) with domains $D \subset \mathbb{R}^2$ that lead to spatially singular solutions. Specifically, we work with the L–shape and crack domains considered in Examples 3.3–3.5. We construct multilevel SGFEM spaces X of the form (6.1) for some subset $J_P \subset J$ (where J is defined in (5.27)), where each space $H_1^\mu \subset H_0^1(D)$ is constructed with respect to a non–uniform mesh $\mathcal{T}_\mu$ of triangular elements over D. By modifying Algorithms 1 and 2, our aim is to locally refine the meshes $\mathcal{T}_\mu$ to realise the optimum rate of convergence of $-\frac{1}{2}$ with respect to N_X on the more challenging L–shape and crack domains, when employing $\mathbb{P}_1$ approximation. Recall from Example 3.3 that the error decays at the rate of $-\frac{1}{3}$ with respect to the number of DOFs when performing uniform mesh refinements on the L–shape domain. In Examples 3.4 and 3.5, we were able to recover the rate $-\frac{1}{2}$ by employing a local mesh refinement strategy.

A new framework is required to adaptively construct sequences of multilevel SGFEM spaces $\{X\}$ of the form (6.1) with locally refined meshes. Since the sequence of meshes $\{\mathcal{T}_\mu\}$ associated with the space H_1^μ is to be constructed adaptively, it is no longer possible within an algorithm to assign a mesh from an underlying sequence such as $\boldsymbol{\mathcal{T}}$ in (6.5) to a multi–index $\mu \in J_P$. Recall that the multilevel SGFEM space X in Section 6.1 is completely determined by the set J_P and the corresponding set of mesh level numbers $\boldsymbol{\ell}$. Instead, X here is determined by J_P and the set of meshes

$$\boldsymbol{\mathcal{T}} := \big\{\mathcal{T}_\mu;\ \mu \in J_P\big\}. \tag{6.34}$$

In an adaptive algorithm, J_P is enriched and the meshes in (6.34) are refined, maintaining the relationship $\mathrm{card}(J_P) = \mathrm{card}(\boldsymbol{\mathcal{T}})$. Typically, we initialise $\boldsymbol{\mathcal{T}}$ by constructing an initial coarse triangulation $\mathcal{T}_0$ and setting $\mathcal{T}_\mu = \mathcal{T}_0$ for each $\mu \in J_P$. Note that any two meshes $\mathcal{T}_\mu$ and $\mathcal{T}_\nu$ from the set $\boldsymbol{\mathcal{T}}$ for $\mu \neq \nu$ are almost certainly different. As a result, the total number of stiffness matrices $K_{\nu\mu}^m$ required to compute an SGFEM approximation $u_X \in X$ could be significantly higher than the number required in Section 6.1 where uniform meshes are employed.

The assembly method described in Section 6.2.2 for the efficient construction of $K_{\nu\mu}^m$ straightforwardly extends to the case where $\mathcal{T}_\mu$ and $\mathcal{T}_\nu$ are non–uniform. Assume that the meshes $\mathcal{T}_\mu$ and $\mathcal{T}_\nu$ are constructed by performing different conforming refinements



(a) initial mesh $\mathcal{T}_0$. (b) $\mathcal{T}_\mu$. (c) $\mathcal{T}_\nu$.

Figure 6.8: Two conforming refinements $\mathcal{T}_\mu$ and $\mathcal{T}_\nu$ of an initial mesh $\mathcal{T}_0$. The blue elements in $\mathcal{T}_\mu$ and $\mathcal{T}_\nu$ are subsets of the pink elements in $\mathcal{T}_\nu$ and $\mathcal{T}_\mu$, respectively.

of an initial mesh $\mathcal{T}_0$ (using longest edge bisection [90]). Then, no element edge in $\mathcal{T}_\mu$ crosses an edge in $\mathcal{T}_\nu$ and every element in $\mathcal{T}_\mu$ is a subset or a superset of an element in $\mathcal{T}_\nu$ and vice versa. This is illustrated in Figure 6.8 where $\mathcal{T}_\mu$ and $\mathcal{T}_\nu$ in Figures 6.8b and 6.8c are different conforming refinements of $\mathcal{T}_0$ in Figure 6.8a. The blue elements in $\mathcal{T}_\mu$ and $\mathcal{T}_\nu$ are subsets of the pink elements in $\mathcal{T}_\nu$ and $\mathcal{T}_\mu$, respectively. We construct $K^m_{\nu\mu}$ by concatenating *coarse–element* matrices associated with each pink (coarse) element. In contrast to the construction of $K^m_{\nu\mu}$ in Example 6.1, the coarse–element matrices here are non–uniform in size and not necessarily associated with the same mesh.

Once the approximation $u_X \in X$ satisfying (6.6) is computed, we may estimate the energy error $\|u - u_X\|_B$ as before by solving problems (6.17) and (6.18) for suitable choices of $\mathbf{H}_2 = \{H_2^\mu\}_{\mu\in J_P}$, $J_Q \subset J$ and $H \subset H_0^1(D)$ (recall Section 6.3) and evaluating $\eta = \|e_Y\|_{B_0}$ in (6.19). Here onwards we fix $H = H_0$, the $\mathbb{P}_1$ FEM space associated with the initial mesh $\mathcal{T}_0$.

An Adaptive Algorithm

In this section, we explain how to incorporate local spatial refinement into Algorithm 1 by exploiting the spatial estimators $e^\mu_{Y_1} \in Y_1^\mu$ satisfying (6.17) for $\mu \in J_P$. Consider the sets of elementwise estimates

$$\boldsymbol{\mathcal{E}}_\mu := \big\{\|e^\mu_{Y_{1k}}\|_{B_{0k}};\ \Delta_k \in \mathcal{T}_\mu\big\},$$

for $\mu \in J_P$, where

$$\|e^\mu_{Y_{1k}}\|^2_{B_{0k}} := B_{0k}(e^\mu_{Y_{1k}}, e^\mu_{Y_{1k}}), \qquad e^\mu_{Y_{1k}} := e^\mu_{Y_1}|_{\Delta_k},$$

and

$$B_{0k}(v,v) = \int_\Gamma \int_{\Delta_k} a_0(\mathbf{x})|\nabla v(\mathbf{x},\mathbf{y})|^2\, d\mathbf{x}\, d\pi(\mathbf{y}), \qquad \text{for all } v \in Y_1^\mu|_{\Delta_k}.$$

We define the estimated error reduction

$$\|e_{\bar{Y}_1}\|^2_{B_0} := \sum_{\mu\in\bar{J}_P}\sum_{\Delta_k\in\mathcal{M}^*_\mu} \|e^\mu_{Y_{1k}}\|^2_{B_{0k}}. \tag{6.35}$$

for suitably chosen sets (to be discussed) $\bar{J}_P \subseteq J_P$ and $\mathcal{M}^*_\mu \subseteq \mathcal{T}_\mu$. Whilst the quantity $\|e_{\bar{Y}_1}\|^2_{B_0}$ doesn't formally satisfy an error reduction bound such as (6.29), it serves as a suitable proxy for the square of the error reduction that would be achieved by refining the elements $\mathcal{M}^*_\mu$ in the mesh $\mathcal{T}_\mu$ for each $\mu \in \bar{J}_P$ and computing an enhanced SGFEM approximation.

Recall the error reduction ratio $R_{\bar{Y}_1} = \|e_{\bar{Y}_1}\|^2_{B_0}/N_{\bar{Y}_1}$ in (6.31). To reflect the new definition of $\|e_{\bar{Y}_1}\|^2_{B_0}$ in (6.35), we redefine $N_{\bar{Y}_1}$ to be

$$N_{\bar{Y}_1} := \sum_{\mu\in\bar{J}_P} N^*_\mu, \qquad N^*_\mu := N_1^{\mu+} - N_1^\mu, \tag{6.36}$$

where N_1^μ is the dimension of H_1^μ. Suppose for now that the elements $\mathcal{M}^*_\mu$ in the mesh $\mathcal{T}_\mu$ are refined (including the resolution of hanging nodes) and write $\mathcal{T}_\mu \to \mathcal{T}_{\mu+}$. We denote by $N_1^{\mu+}$ in (6.36) the dimension of the enriched space $H_1^{\mu+} \supset H_1^\mu$ associated with the refined mesh $\mathcal{T}_{\mu+}$. Note that the definitions of $\|e_{\bar{Y}_2}\|^2_{B_0}$ and $R_{\bar{Y}_2}$ in (6.28) and (6.31) go unchanged and, as in Section 6.4.1, we implement the enrichment of X that corresponds to $\max\{R_{\bar{Y}_1}, R_{\bar{Y}_2}\}$.

If spatial enrichment is chosen, we enrich the spaces H_1^μ for $\mu \in \bar{J}_P$ by employing the refinement $\mathcal{T}_\mu \to \mathcal{T}_{\mu+}$ so that $H_1^\mu \to H_1^{\mu+}$. Otherwise, if parametric refinement is chosen we add $\bar{J}_Q$ to J_P and update $\boldsymbol{\mathcal{T}}$ with $\text{card}(\bar{J}_Q)$ copies of the initial mesh $\mathcal{T}_0$. In turn, this updates $\mathbf{H}_1$ with $\text{card}(\bar{J}_Q)$ copies of H_0. The new procedure is outlined in Algorithm 3. Notice the input of $\mathcal{T}_0$ and compare lines 13 and 16 of Algorithm 1 with those of Algorithm 3. Additionally, the module `COMPONENT_SPATIAL_ERRORS` has two additional outputs; $\boldsymbol{\mathcal{E}}^*$ and $\boldsymbol{\mathcal{M}}^*$, where

$$\boldsymbol{\mathcal{M}}^* := \big\{\mathcal{M}^*_\mu;\ \mu \in J_P\big\}.$$

When spatial enrichment is chosen, the set of elements $\mathcal{M}^*_\mu \subseteq \mathcal{T}_\mu$ is required to perform the refinement $\mathcal{T}_\mu \to \mathcal{T}_{\mu+}$ for $\mu \in \bar{J}_P$. We now explain how to choose suitable sets $\mathcal{M}^*_\mu$ (as well as $\bar{J}_P$ and $\bar{J}_Q$) and define $\boldsymbol{\mathcal{E}}^*$.

Algorithm 3: Adaptive multilevel SGFEM with local mesh refinement

Input : problem data $a(\mathbf{x},\mathbf{y})$, $f(\mathbf{x})$; initial index set J_P and mesh $\mathcal{T}_0$; energy error tolerance ϵ.

Output: final SGFEM approximation u_X and energy error estimate η.

$\boldsymbol{\mathcal{T}} = \{\mathcal{T}_0;\ \mu \in J_P\}$
choose `version` (1 or 2)
for $k = 0, 1, 2, \ldots$ **do**
  $u_X \leftarrow$ `SOLVE`$\big[a, f, J_P, \boldsymbol{\mathcal{T}}\big]$
  $J_Q \leftarrow$ `PARAMETRIC_INDICES`$\big[J_P\big]$ see (5.66)
  $[\mathbf{E}_{Y_1}, \boldsymbol{\mathcal{E}}^*, \boldsymbol{\mathcal{M}}^*] \leftarrow$ `COMPONENT_SPATIAL_ERRORS`$\big[u_X, J_P, \boldsymbol{\mathcal{T}}\big]$
  $\mathbf{E}_{Y_2} \leftarrow$ `COMPONENT_PARAMETRIC_ERRORS`$\big[u_X, J_Q, \mathcal{T}_0\big]$
  $\eta = \big[\sum_{\mu\in J_P} ||e^{\mu}_{Y_1}||^2_{B_0} + \sum_{\nu\in J_Q} ||e^{\nu}_{Y_2}||^2_{B_0}\big]^{\frac{1}{2}}$ (6.19)
  if $\eta < \epsilon$ **then**
    return u_X, η
  else
    $[$`enrichment_type`$, \bar{J}] \leftarrow$ `ENRICHMENT_INDICES`$\big[$`version`$, \boldsymbol{\mathcal{E}}^*, \mathbf{E}_{Y_2}, J_P, J_Q\big]$
    if `enrichment_type` = `spatial` **then**
      $\boldsymbol{\mathcal{T}} \rightarrow \big\{\mathcal{T}_{\mu+};\ \mu \in \bar{J}\big\} \cup \big\{\mathcal{T}_{\mu};\ \mu \in J_P \backslash \bar{J}\big\}$
    else
      $J_P \rightarrow J_P \cup \bar{J}$
      $\boldsymbol{\mathcal{T}} \rightarrow \boldsymbol{\mathcal{T}} \cup \big\{\mathcal{T}_0;\ \nu \in \bar{J}\big\}$
    end
  end
end

Selection of Enrichment Indices

Following the selection of $\mathcal{M} \subset \mathcal{T}_h$ associated with (3.38), we employ a Dörfler marking strategy. For each $\mu \in J_P$, we construct a minimal subset of marked elements $\mathcal{M}_\mu \subset \mathcal{T}_\mu$ satisfying

$$\sum_{\Delta_k \in \mathcal{M}_\mu} \|e^{\mu}_{Y_{1k}}\|^2_{B_{0k}} \geq \theta_{\text{mark}} \sum_{\Delta_k \in \mathcal{T}_\mu} \|e^{\mu}_{Y_{1k}}\|^2_{B_{0k}}. \tag{6.37}$$

Suppose that the elements $\mathcal{M}_\mu$ are refined in the mesh $\mathcal{T}_\mu$ to produce $\mathcal{T}_{\mu+}$, with any hanging nodes resolved. The total set of refined elements in $\mathcal{T}_\mu$ may be richer than $\mathcal{M}_\mu$. We denote this set by $\mathcal{M}^*_\mu$ and note that $\mathcal{M}^*_\mu = \mathcal{T}_\mu \backslash \mathcal{T}_{\mu+}$. The logical steps of Algorithm 2 (the module `ENRICHMENT_INDICES` in Algorithms 1 and 3) do not change and we only switch the input of $\mathbf{E}_{Y_1}$ with the output $\boldsymbol{\mathcal{E}}^*$ from `COMPONENT_SPATIAL_ERRORS`. Here,

$$\boldsymbol{\mathcal{E}}^* := \big\{\mathcal{E}^*_\mu;\ \mu \in J_P\big\}, \qquad \mathcal{E}^*_\mu := \sum_{\Delta_k \in \mathcal{M}^*_\mu} \|e^{\mu}_{Y_{1k}}\|^2_{B_{0k}}, \qquad \|e^{\mu}_{Y_{1k}}\|_{B_{0k}} \in \boldsymbol{\mathcal{E}}_\mu.$$

We also redefine the set of ratios $\mathbf{R}_{Y_1}$ in (6.33) by

$$\mathbf{R}_{Y_1} := \big\{R^{\mu}_{Y_1};\ \mu \in J_P\big\}, \qquad R^{\mu}_{Y_1} := \frac{\mathcal{E}^*_{\mu}}{N^*_{\mu}},$$

as well as the set $\mathbf{N}_{Y_1} := \{N^*_{\mu};\ \mu \in J_P\}$; the definitions of $\mathbf{R}_{Y_2}$ in (6.33) and $\mathbf{N}_{Y_2}$ go unchanged. With these new definitions, Algorithm 2 constructs subsets $\bar{J}_P \subseteq J_P$ and $\bar{J}_Q \subseteq J_Q$ as usual and the set $\bar{J} = \bar{J}_P$ or $\bar{J} = \bar{J}_Q$ corresponding to $\max\{R_{\bar{Y}_1}, R_{\bar{Y}_2}\}$ is outputted to Algorithm 3.

6.5.1 Numerical Experiments

Here we test the performance of Algorithm 3 by solving test problems TP4 and TP3 on the L–shape and crack domains

$$D = [-1,1]^2\backslash[-1,0)^2,$$
$$D = [-1,1]^2\backslash\big\{(x_1,x_2)^{\top} \in \mathbb{R}^2;\ -1 < x_1 \leq 0,\ x_2 = 0\big\}.$$

We initialise Algorithm 1 with the set

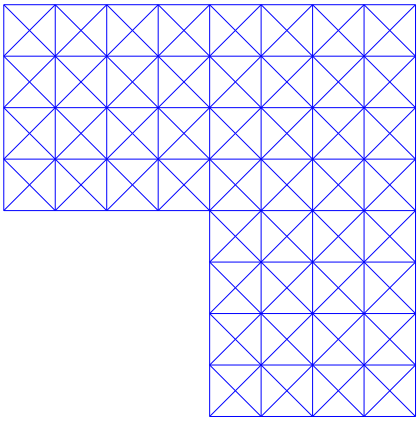
$$J_P = \big\{(0,0,\dots),(1,0\dots)\big\},$$

and $\mathcal{T}_0$ given in Figures 6.9a and 6.9b for TP4 and TP3, respectively and choose $\mathbf{H}_1$ to be the set of $\mathbb{P}_1$ spaces associated with $\boldsymbol{\mathcal{T}}$, i.e., $H^{\mu}_1 = H_0$ for each $\mu \in J_P$.

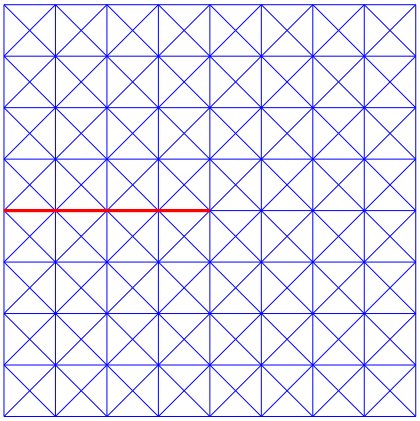
To compute the error estimate η in (6.19) we now choose the FEM spaces $\mathbf{H}_2$. Similar to Example 3.3, we construct each $H^{\mu}_2 \in \mathbf{H}_2$ using the usual (global) $\mathbb{P}_2$ and super–quadratic basis functions, defined with respect to the element edge–midpoints and centroids of $\mathcal{T}_{\mu}$. We also choose J_Q as given in (5.66) with $\Delta_M = 5$ fixed and, following the success of Examples 3.4 and 3.5, fix $\theta_{\text{mark}} = \frac{1}{2}$ in (6.37). In the following Example we compute sequences of SGFEM spaces $\{X\}$ of the form (6.1) by adaptively constructing the set J_P and locally refining each mesh in $\boldsymbol{\mathcal{T}}$.

Example 6.5: Rates of Convergence, spatial singularity.

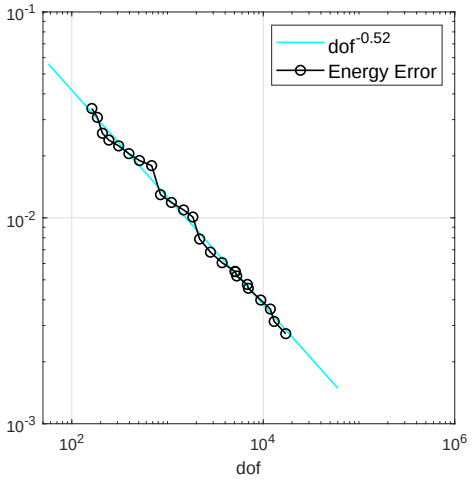
We solve TP4 and TP3 using Algorithm 3 with $\epsilon = 3 \times 10^{-3}$ and version 2 of Algorithm 2. In Figures 6.9c and 6.9d we plot the energy error estimate $\eta = \|e_Y\|_{B_0}$ versus the number of DOFs N_X at each step k of Algorithm 3. For both problems, we observe that the error decays at the optimum rate of $-\frac{1}{2}$. Similar test problems are considered in [21] where local mesh refinement is incorporated into



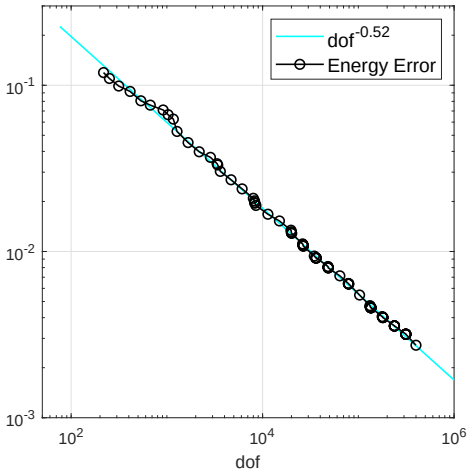
(a) $\mathcal{T}_0$ for the L–shape domain.



(b) $\mathcal{T}_0$ for the crack domain.



(c) convergence of $\eta = \|e_Y\|_{B_0}$ for TP4.



(d) convergence of $\eta = \|e_Y\|_{B_0}$ for TP3.

Figure 6.9: Initial meshes $\mathcal{T}_0$ and convergence of $\eta = \|e_Y\|_{B_0}$ for Example 6.5 when solving test problems TP4 and TP3 on the L–shape and crack domains, respectively. The red line is the crack in D along the line $\{(x_1, x_2)^\top \in \mathbb{R}^2;\ -1 < x_1 \leq 0,\ x_2 = 0\}$.

an adaptive single–level SGFEM similar to the one in Section 5.4.2. The energy error is reported in [21] to decay at rates between $-\frac{1}{3}$ and $-\frac{4}{5}$.

Figures 6.10 and 6.11 (top) display surface plots of the coefficients $u_X^\mu(\mathbf{x})$ (recall $u_X(\mathbf{x}, \mathbf{y})$ in (6.7)) associated with the first three multi–indices $\mu \in J_P$ selected by Algorithms 2 and 3 in Example 6.5. Below them, we also display the locally refined meshes $\mathcal{T}_\mu$ associated with those multi–indices after the fifteenth step of Algorithm 3. Notice how refinement is concentrated primarily at the point $(0, 0)^\top \in D$ where the true solution $u \in V$ is spatially singular and other areas where $u_X^\mu(\mathbf{x})$ has steeper gradients. At the final step of Algorithm 3, $\text{card}(J_P) = 26$ and $\text{card}(J_P) = 17$ for TP4

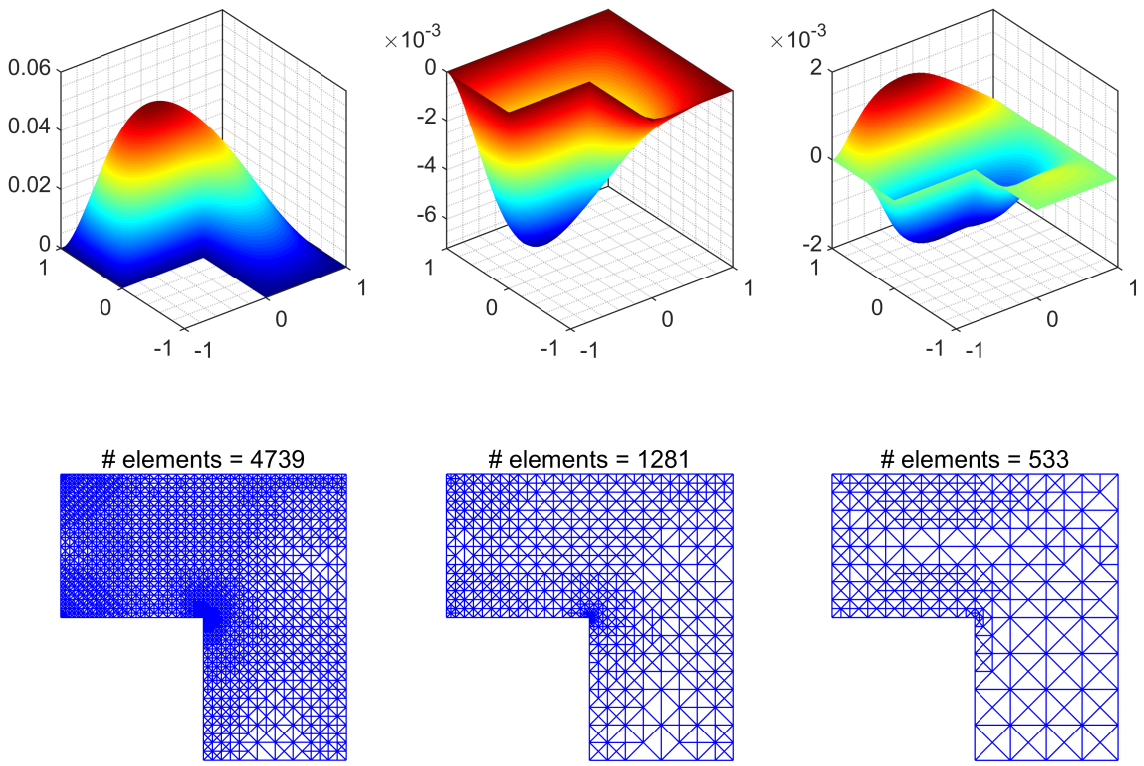


Figure 6.10: Top: surface plots of the coefficients $u_X^\mu(\mathbf{x})$ associated with the multi–indices $\mu = (0, 0, \dots), (1, 0, \dots), (2, 0, \dots) \in J_P$ for TP4 on the L–shape domain when Algorithm 3 terminates. Bottom: the corresponding adaptively constructed meshes $\mathcal{T}_\mu \in \boldsymbol{\mathcal{T}}$ after the $k = 15^{\text{th}}$ step.

and TP3, respectively, meaning that as many meshes $\mathcal{T}_\mu$ for $\mu \in J_P$ are independently and locally refined. Observe from Figures 6.9b and 6.9d that many adaptive steps (the black markers) are required for Algorithm 3 to terminate. A major reason for this is our choice $H = H_0$ associated with the parametric error problem (6.18). When parametric enrichment is performed in Algorithm 3, $\text{card}(\bar{J})$ copies of $\mathcal{T}_0$ are added to the set $\boldsymbol{\mathcal{T}}$. Algorithm 3 subsequently takes several steps to refine the newly incorporated meshes to a state commensurate with the other meshes in the set. To avoid this, $H \subset H_0^1(D)$ must be automated and depend on the current SGFEM space X. Recall in Section 6.3.1 that we chose $H = H_1^{\bar{\mu}}$ where $\bar{\mu} = \arg\,\text{avg}_{\mu \in J_P}\, \boldsymbol{\ell} \in J_P$.

6.6 Summary

In this chapter, we introduced multilevel approximation spaces X of the form (6.1) and explained how the coefficients $u_X^\mu(\mathbf{x})$ of multilevel approximations $u_X \in X$ satisfying

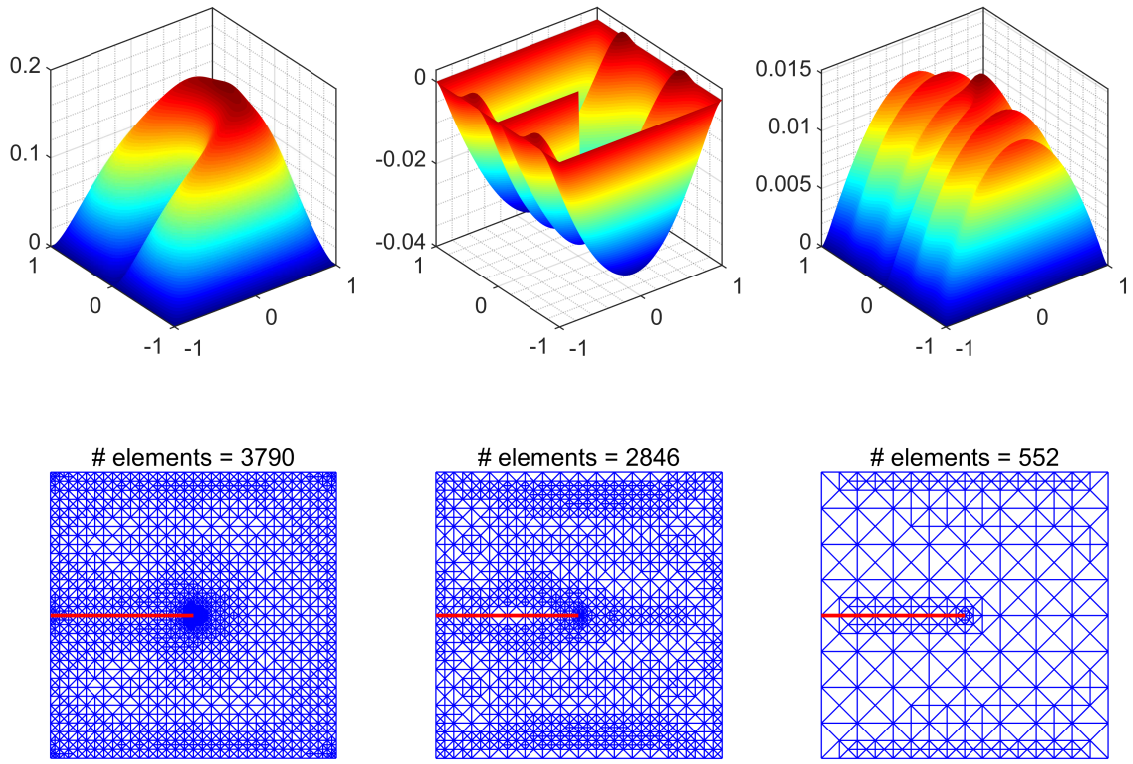


Figure 6.11: Top: surface plots of the coefficients $u_X^\mu(\mathbf{x})$ associated with the multi–indices $\mu = (0,0,\dots),(1,0,\dots),(2,0,\dots) \in J_P$ for TP3 on the crack domain when Algorithm 3 terminates. Bottom: the corresponding adaptively constructed meshes $\mathcal{T}_\mu \in \boldsymbol{\mathcal{T}}$ after the $k = 15^{\text{th}}$ step.

(6.6) reside in potentially different FEM spaces $H_1^\mu \subset H_0^1(D)$. The associated stiffness matrices $K_{\nu\mu}^m$ in (6.9) often depend on basis functions ϕ_i^μ and ϕ_j^ν associated with different meshes $\mathcal{T}_{\ell^\mu}$ and $\mathcal{T}_{\ell^\nu}$, respectively, where $\ell^\mu, \ell^\nu \in \boldsymbol{\ell}$ are mesh *level numbers*. In Section 6.2.2, we demonstrated how to construct these matrices efficiently using numerical quadrature.

In Section 6.4.1, we designed a novel adaptive multilevel SGFEM (the combination of Algorithms 1 and 2) which automatically decides whether to perform spatial or parametric enrichment of the approximation space. The adaptive process is driven by sharp estimates of the true error reduction per additional DOF for several enrichment options (the sets $\mathbf{R}_{Y_1}$ and $\mathbf{R}_{Y_2}$ in (6.33)). Our algorithm implements only the most economical options and numerical experiments in Section 6.4.3 demonstrate that the new method achieves the optimal rate of convergence for the test problems considered. A notable feature of our method is that no marking or turning parameters are required. The effectivity indices $\theta_{\text{eff}}^{\text{approx}}$ in Figure 6.5 also demonstrate that the a posteriori error

estimate $\eta = \|e_Y\|_{B_0}$ is highly accurate at each step of the algorithm. In order to realise the optimal rates of convergence on non–convex spatial domains such as the L–shape and crack domains, in Section 6.5 we introduced Algorithm 3 which employs local mesh refinement for each solution mode. Figures 6.10 and 6.11 show that Algorithm 3 successfully refines the meshes $\mathcal{T}_\mu \in \boldsymbol{\mathcal{T}}$ and resolves complex features of the modal coefficients $u_X^\mu(\mathbf{x}) \in H_1^\mu$.

Chapter 7

Conclusions

In this thesis, we considered efficient adaptive SGFEMs for the numerical solution of elliptic PDEs with uncertain or parameter dependent inputs. For a model problem, we considered the stochastic diffusion problem (5.2)–(5.3) where the uncertain diffusion coefficient $a(\mathbf{x}, \omega)$ is modelled as a KL–type expansion of the form (5.1), depending on a countably infinite number of random variables. Our motivation for this work comes from the fact that many existing methods require the a priori truncation of the infinite sum in (5.1) or do not achieve the theoretically optimal rates of convergence for the model problem. Consequently, when the chosen stopping tolerance for the approximation error is small, such methods are slow to run and quickly exhaust computer memory. We designed two novel adaptive multilevel SGFEMs – Algorithms 1 and 3 – which, when applied to various test problems, achieved the optimal rates of convergence on convex and non–convex spatial domains, respectively.

Algorithms 1 and 3 are steered by highly accurate efficient a posteriori error estimates. In Chapter 4, we investigated the so–called CBS constant for non–standard pairs of FEM spaces H_1 and H_2 that appears in the bound relating the true errors to the estimated errors. This chapter is an extended discussion of the work published in [37] and contains our first novel contributions. Our main result – Theorem 4.6 – provides a novel theoretical estimate of the CBS constant for certain special pairs of H_1 and H_2. CBS constants arise in many areas of numerical analysis and thus our results extend beyond the field of a posteriori error estimation.

In Chapter 5, we introduced standard SGFEMs and error estimation for the parametric reformulation of the stochastic diffusion problem. Building on Chapter 4, we

designed cheap–to–compute a posteriori error estimates with effectivity indices close to one and discussed the adaptive construction of SGFEM approximation spaces. Chapter 6 is an extended discussion of the work published in [38] where we married several important aspects of Chapters 3–5 to design Algorithms 1 and 3. Our novel multilevel algorithms automatically decide whether spatial or parametric enrichment of the approximation space is required, including the incorporation of more input parameters into the discretisation. We stress that Algorithms 1 and 3 do not require the a priori truncation of the infinite sum appearing in the input $a(\mathbf{x}, \mathbf{y})$ in (5.5).

We conclude this thesis by highlighting some interesting directions and opportunities for future research. The design of adaptive SGFEMs is in its infancy and the most efficient methods apply only to second–order elliptic PDEs with affine dependence on a countable number of input random variables. We envisage that several aspects of this work can stimulate the design of new adaptive multilevel methods for more complex PDE models or models with non–affine dependence on the input variables. Many practitioners also seek quantities of interest that depend on the solution to PDE models, rather than the full solution itself. For example, we may be interested in the maximum value of the solution or its value at a particular point in the spatial domain. Adaptive methods which efficiently reduce the energy error do not necessarily approximate quantities of interest in an efficient way. To this end, *goal–oriented* adaptivity enables us to tailor approximation spaces to the quantity of interest. We expect that the most efficient goal–oriented SGFEMs will include those for which a multilevel structure is imposed on the approximation space. The development of such methods will allow for a broader class of problems to be considered for quick and efficient UQ.

Bibliography

[1] Ainsworth, M. (1994). The performance of Bank-Weiser's error estimator for quadrilateral finite elements. *Numer. Methods Partial Differential Equations*, 10(5):609–623.

[2] Ainsworth, M. and Oden, J. T. (2000). *A posteriori error estimation in finite element analysis*. Pure and Applied Mathematics (New York). Wiley-Interscience [John Wiley & Sons], New York.

[3] Axelsson, O. (1994). *Iterative solution methods*. Cambridge University Press, Cambridge.

[4] Axelsson, O., Blaheta, R., Neytcheva, M., and Pultarová, I. (2015). Preconditioning of iterative methods—theory and applications [Editorial]. *Numer. Linear Algebra Appl.*, 22(6):901–902.

[5] Axelsson, O. and Vassilevski, P. S. (1990). Algebraic multilevel preconditioning methods. II. *SIAM J. Numer. Anal.*, 27(6):1569–1590.

[6] Babuška, I. M., Nobile, F., and Tempone, R. (2007). A stochastic collocation method for elliptic partial differential equations with random input data. *SIAM J. Numer. Anal.*, 45(3):1005–1034.

[7] Babuška, I. M., Tempone, R., and Zouraris, G. E. (2004). Galerkin finite element approximations of stochastic elliptic partial differential equations. *SIAM J. Numer. Anal.*, 42(2):800–825.

[8] Babuška, I. M., Tempone, R., and Zouraris, G. E. (2005). Solving elliptic boundary value problems with uncertain coefficients by the finite element method: the stochastic formulation. *Comput. Methods Appl. Mech. Engrg.*, 194(12-16):1251–1294.

[9] Babuška, I. and Miller, A. (1987). A feedback finite element method with a posteriori error estimation. I. The finite element method and some basic properties of the a posteriori error estimator. *Comput. Methods Appl. Mech. Engrg.*, 61(1):1–40.

[10] Babuška, I. and Suri, M. (1987). The *h-p* version of the finite element method with quasi-uniform meshes. *RAIRO Modél. Math. Anal. Numér.*, 21(2):199–238.

[11] Bachmayr, M., Cohen, A., DeVore, R., and Migliorati, G. (2017a). Sparse polynomial approximation of parametric elliptic PDEs. Part II: Lognormal coefficients. *ESAIM Math. Model. Numer. Anal.*, 51(1):341–363.

[12] Bachmayr, M., Cohen, A., and Migliorati, G. (2017b). Sparse polynomial approximation of parametric elliptic PDEs. Part I: Affine coefficients. *ESAIM Math. Model. Numer. Anal.*, 51(1):321–339.

[13] Bäck, J., Nobile, F., Tamellini, L., and Tempone, R. (2011). Stochastic spectral Galerkin and collocation methods for PDEs with random coefficients: a numerical comparison. In *Spectral and high order methods for partial differential equations*, volume 76 of *Lect. Notes Comput. Sci. Eng.*, pages 43–62. Springer, Heidelberg.

[14] Bank, R. E., Parsania, A., and Sauter, S. (2013). Saturation estimates for *hp*-finite element methods. *Comput. Vis. Sci.*, 16(5):195–217.

[15] Bank, R. E. and Smith, R. K. (1993). A posteriori error estimates based on hierarchical bases. *SIAM J. Numer. Anal.*, 30(4):921–935.

[16] Bank, R. E. and Weiser, A. (1985). Some a posteriori error estimators for elliptic partial differential equations. *Math. Comp.*, 44(170):283–301.

[17] Barth, A., Schwab, C., and Zollinger, N. (2011). Multi-level Monte Carlo finite element method for elliptic PDEs with stochastic coefficients. *Numer. Math.*, 119(1):123–161.

[18] Bespalov, A., Powell, C. E., and Silvester, D. (2014). Energy norm a posteriori error estimation for parametric operator equations. *SIAM J. Sci. Comput.*, 36(2):A339–A363.

[19] Bespalov, A., Powell, C. E., and Silvester, D. (2016). Stochastic IFISS (S-IFISS) version 1.1. Available online at `http://www.manchester.ac.uk/ifiss/s-ifiss1.0.tar.gz`.

[20] Bespalov, A., Praetorius, D., Rocchi, L., and Ruggeri, M. (2019). Goal-oriented error estimation and adaptivity for elliptic PDEs with parametric or uncertain inputs. *Comput. Methods Appl. Mech. Engrg.*, 345:951–982.

[21] Bespalov, A. and Rocchi, L. (2018). Efficient adaptive algorithms for elliptic PDEs with random data. *SIAM/ASA J. Uncertain. Quantif.*, 6(1):243–272.

[22] Bespalov, A. and Silvester, D. (2016). Efficient adaptive stochastic Galerkin methods for parametric operator equations. *SIAM J. Sci. Comput.*, 38(4):A2118–A2140.

[23] Bieri, M., Andreev, R., and Schwab, C. (2009). Sparse tensor discretization of elliptic SPDEs. *SIAM J. Sci. Comput.*, 31(6):4281–4304.

[24] Bieri, M. and Schwab, C. (2009). Sparse high order FEM for elliptic sPDEs. *Comput. Methods Appl. Mech. Engrg.*, 198(13-14):1149–1170.

[25] Binev, P., Dahmen, W., and DeVore, R. (2004). Adaptive finite element methods with convergence rates. *Numer. Math.*, 97(2):219–268.

[26] Blatman, G. and Sudret, B. (2011). Adaptive sparse polynomial chaos expansion based on least angle regression. *J. Comput. Phys.*, 230(6):2345–2367.

[27] Braess, D. (2007). *Finite elements*. Cambridge University Press, Cambridge, third edition. Theory, fast solvers, and applications in elasticity theory, Translated from the German by Larry L. Schumaker.

[28] Brenner, S. C. and Scott, L. R. (2008). *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, third edition.

[29] Capiński, M. and Kopp, E. (2004). *Measure, integral and probability*. Springer Undergraduate Mathematics Series. Springer-Verlag London, Ltd., London, second edition.

[30] Carstensen, C., Gallistl, D., and Gedicke, J. (2016). Justification of the saturation assumption. *Numer. Math.*, 134(1):1–25.

[31] Charrier, J. and Debussche, A. (2013). Weak truncation error estimates for elliptic PDEs with lognormal coefficients. *Stoch. Partial Differ. Equ. Anal. Comput.*, 1(1):63–93.

[32] Charrier, J., Scheichl, R., and Teckentrup, A. L. (2013). Finite element error analysis of elliptic PDEs with random coefficients and its application to multilevel Monte Carlo methods. *SIAM J. Numer. Anal.*, 51(1):322–352.

[33] Chkifa, A., Cohen, A., and Schwab, C. (2015). Breaking the curse of dimensionality in sparse polynomial approximation of parametric PDEs. *J. Math. Pures Appl. (9)*, 103(2):400–428.

[34] Cliffe, K. A., Giles, M. B., Scheichl, R., and Teckentrup, A. L. (2011). Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. *Comput. Vis. Sci.*, 14(1):3–15.

[35] Cohen, A., DeVore, R., and Schwab, C. (2010). Convergence rates of best N-term Galerkin approximations for a class of elliptic sPDEs. *Found. Comput. Math.*, 10(6):615–646.

[36] Cohen, A., Devore, R., and Schwab, C. (2011). Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDE's. *Anal. Appl.*, 9(1):11–47.

[37] Crowder, A. J. and Powell, C. E. (2018). CBS constants & their role in error estimation for stochastic Galerkin finite element methods. *J. Sci. Comput.*, 77(2):1030–1054.

[38] Crowder, A. J., Powell, C. E., and Bespalov, A. (2019). Efficient adaptive multilevel stochastic Galerkin approximation using implicit a posteriori error estimation. *SIAM J. Sci. Comput.*, 41(3):A1681–A1705.

[39] Davis, P. J. (1979). *Circulant matrices*. John Wiley & Sons, New York-Chichester-Brisbane. A Wiley-Interscience Publication, Pure and Applied Mathematics.

[40] Deb, M. K., Babuška, I. M., and Oden, J. T. (2001). Solution of stochastic partial differential equations using Galerkin finite element techniques. *Comput. Methods Appl. Mech. Engrg.*, 190(48):6359–6372.

[41] Dörfler, W. (1996). A convergent adaptive algorithm for Poisson's equation. *SIAM J. Numer. Anal.*, 33(3):1106–1124.

[42] Dörfler, W. and Nochetto, R. H. (2002). Small data oscillation implies the saturation assumption. *Numer. Math.*, 91(1):1–12.

[43] Eiermann, M., Ernst, O. G., and Ullmann, E. (2007). Computational aspects of the stochastic finite element method. *Comput. Vis. Sci.*, 10(1):3–15.

[44] Eigel, M., Gittelson, C. J., Schwab, C., and Zander, E. (2014). Adaptive stochastic Galerkin FEM. *Comput. Methods Appl. Mech. Engrg.*, 270:247–269.

[45] Eigel, M., Gittelson, C. J., Schwab, C., and Zander, E. (2015). A convergent adaptive stochastic Galerkin finite element method with quasi-optimal spatial meshes. *ESAIM Math. Model. Numer. Anal.*, 49(5):1367–1398.

[46] Eigel, M. and Merdon, C. (2016). Local equilibration error estimators for guaranteed error control in adaptive stochastic higher-order Galerkin finite element methods. *SIAM/ASA J. Uncertain. Quantif.*, 4(1):1372–1397.

[47] Eigel, M., Pfeffer, M., and Schneider, R. (2017). Adaptive stochastic Galerkin FEM with hierarchical tensor representations. *Numer. Math.*, 136(3):765–803.

[48] Eijkhout, V. and Vassilevski, P. (1991). The role of the strengthened Cauchy-Buniakowskiĭ-Schwarz inequality in multilevel methods. *SIAM Rev.*, 33(3):405–419.

[49] Elman, H. C., Silvester, D. J., and Wathen, A. J. (2014). *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics*. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, second edition.

[50] Ernst, O. G., Powell, C. E., Silvester, D. J., and Ullmann, E. (2009). Efficient solvers for a linear stochastic Galerkin mixed formulation of diffusion problems with random data. *SIAM J. Sci. Comput.*, 31(2):1424–1447.

[51] Ernst, O. G. and Ullmann, E. (2010). Stochastic Galerkin matrices. *SIAM J. Matrix Anal. Appl.*, 31(4):1848–1872.

[52] Frauenfelder, P., Schwab, C., and Todor, R. A. (2005). Finite elements for elliptic problems with stochastic coefficients. *Comput. Methods Appl. Mech. Engrg.*, 194(2-5):205–228.

[53] Gautschi, W. (2004). *Orthogonal polynomials: computation and approximation*. Numerical Mathematics and Scientific Computation. Oxford University Press, New York. Oxford Science Publications.

[54] Ghanem, R. G. and Spanos, P. D. (1991). *Stochastic finite elements: a spectral approach*. Springer-Verlag, New York.

[55] Gittelson, C. J. (2010). Stochastic Galerkin discretization of the log-normal isotropic diffusion problem. *Math. Models Methods Appl. Sci.*, 20(2):237–263.

[56] Gittelson, C. J. (2013a). An adaptive stochastic Galerkin method for random elliptic operators. *Math. Comp.*, 82(283):1515–1541.

[57] Gittelson, C. J. (2013b). Convergence rates of multilevel and sparse tensor approximations for a random elliptic PDE. *SIAM J. Numer. Anal.*, 51(4):2426–2447.

[58] Golub, G. H. and Welsch, J. H. (1969). Calculation of Gauss quadrature rules. *Math. Comp.*, 23(106):221–230.

[59] Grimmett, G. R. and Stirzaker, D. R. (2001). *Probability and random processes*. Oxford University Press, New York, third edition.

[60] Gunzburger, M. D., Webster, C. G., and Zhang, G. (2014). Stochastic finite element methods for partial differential equations with random input data. *Acta Numer.*, 23:521–650.

[61] Haji-Ali, A.-L., Nobile, F., Tamellini, L., and Tempone, R. (2016). Multi-index stochastic collocation for random PDEs. *Comput. Methods Appl. Mech. Engrg.*, 306:95–122.

[62] Khan, A., Powell, C. E., and Silvester, D. J. (2017). Robust a posteriori error estimators for mixed approximation of nearly incompressible elasticity. Available online at `https://arxiv.org/abs/1710.03328`.

[63] Khan, A., Powell, C. E., and Silvester, D. J. (2019). Robust preconditioning for stochastic Galerkin formulations of parameter-dependent nearly incompressible elasticity equations. *SIAM J. Sci. Comput.*, 41(1):A402–A421.

[64] König, H. (1986). *Eigenvalue distribution of compact operators*, volume 16 of *Operator Theory: Advances and Applications*. Birkhäuser Verlag, Basel.

[65] Kopteva, N. and O'Riordan, E. (2010). Shishkin meshes in the numerical solution of singularly perturbed differential equations. *Int. J. Numer. Anal. Model.*, 7(3):393–415.

[66] Krupski, M. o. (2019). Convection-diffusion equations with random initial conditions. *J. Math. Anal. Appl.*, 470(2):1194–1221.

[67] Le Maître, O. P. and Knio, O. M. (2010). *Spectral methods for uncertainty quantification*. Scientific Computation. Springer, New York. With applications to computational fluid dynamics.

[68] Liao, Q. (2010). *Error estimation and stabilization for low order finite elements*. PhD thesis, University of Manchester, Manchester.

[69] Loève, M. (1978). *Probability theory. II*. Springer-Verlag, New York-Heidelberg, fourth edition. Graduate Texts in Mathematics, Vol. 46.

[70] Lord, G. J., Powell, C. E., and Shardlow, T. (2014). *An introduction to computational stochastic PDEs*. Cambridge Texts in Applied Mathematics. Cambridge University Press, New York.

[71] Morin, P., Nochetto, R. H., and Siebert, K. G. (2000). Data oscillation and convergence of adaptive FEM. *SIAM J. Numer. Anal.*, 38(2):466–488.

[72] Morin, P., Nochetto, R. H., and Siebert, K. G. (2002). Convergence of adaptive finite element methods. *SIAM Rev.*, 44(4):631–658.

[73] Mu, L. and Zhang, G. (2019). A Domain Decomposition Model Reduction Method for Linear Convection-Diffusion Equations with Random Coefficients. *SIAM J. Sci. Comput.*, 41(3):A1984–A2011.

[74] Nobile, F., Tempone, R., and Webster, C. G. (2008a). An anisotropic sparse grid stochastic collocation method for partial differential equations with random input data. *SIAM J. Numer. Anal.*, 46(5):2411–2442.

[75] Nobile, F., Tempone, R., and Webster, C. G. (2008b). A sparse grid stochastic collocation method for partial differential equations with random input data. *SIAM J. Numer. Anal.*, 46(5):2309–2345.

[76] Oden, J. T. and Demkowicz, L. F. (1996). *Applied functional analysis*. CRC Series in Computational Mechanics and Applied Analysis. CRC Press, Boca Raton, FL.

[77] Oden, J. T. and O'Leary, J. (1978). Some remarks on finite element approximations of crack problems and an analysis of hybrid methods. *Journal of Structural Mechanics*, 6(4):415–436.

[78] Pellissetti, M. and Ghanem, R. (2000). Iterative solution of systems of linear equations arising in the context of stochastic finite elements. *Advances in Engineering Software*, 31(8-9):607–616.

[79] Powell, C. E. and Elman, H. C. (2009). Block-diagonal preconditioning for spectral stochastic finite-element systems. *IMA J. Numer. Anal.*, 29(2):350–375.

[80] Powell, C. E., Silvester, D., and Simoncini, V. (2017). An efficient reduced basis solver for stochastic Galerkin matrix equations. *SIAM J. Sci. Comput.*, 39(1):A141–A163.

[81] Powell, C. E. and Silvester, D. J. (2012). Preconditioning steady-state Navier-Stokes equations with random data. *SIAM J. Sci. Comput.*, 34(5):A2482–A2506.

[82] Pranesh, S. and Ghosh, D. (2018). Cost reduction of stochastic Galerkin method by adaptive identification of significant polynomial chaos bases for elliptic equations. *Comput. Methods Appl. Mech. Engrg.*, 340:54–69.

[83] Pultarová, I. (2005). The strengthened C.B.S. inequality constant for second order elliptic partial differential operator and for hierarchical bilinear finite element functions. *Appl. Math.*, 50(3):323–329.

[84] Pultarová, I. (2009). Preconditioning and a posteriori error estimates using h- and p-hierarchical finite elements with rectangular supports. *Numer. Linear Algebra Appl.*, 16(5):415–430.

[85] Pultarová, I. (2015). Adaptive algorithm for stochastic Galerkin method. *Appl. Math.*, 60(5):551–571.

[86] Pultarová, I. (2016). Hierarchical preconditioning for the stochastic Galerkin method: upper bounds to the strengthened CBS constants. *Comput. Math. Appl.*, 71(4):949–964.

[87] Pultarová, I. (2017). Block and multilevel preconditioning for stochastic Galerkin problems with lognormally distributed parameters and tensor product polynomials. *Int. J. Uncertain. Quantif.*, 7(5):441–462.

[88] Reade, J. B. (1983). Eigenvalues of positive definite kernels. *SIAM J. Math. Anal.*, 14(1):152–157.

[89] Reade, J. B. (1984). Eigenvalues of positive definite kernels. II. *SIAM J. Math. Anal.*, 15(1):137–142.

[90] Rivara, M.-C. (1984). Mesh refinement processes based on the generalized bisection of simplices. *SIAM J. Numer. Anal.*, 21(3):604–613.

[91] Schwab, C. (1998). *p- and hp-finite element methods.* Numerical Mathematics and Scientific Computation. The Clarendon Press, Oxford University Press, New York. Theory and applications in solid and fluid mechanics.

[92] Schwab, C. and Gittelson, C. J. (2011). Sparse tensor discretizations of high-dimensional parametric and stochastic PDEs. *Acta Numer.*, 20:291–467.

[93] Schwab, C. and Todor, R.-A. (2003). Sparse finite elements for elliptic problems with stochastic loading. *Numer. Math.*, 95(4):707–734.

[94] Schwab, C. and Todor, R. A. (2006). Karhunen-Loève approximation of random fields by generalized fast multipole methods. *J. Comput. Phys.*, 217(1):100–122.

[95] Shishkin, G. I. (1989). Grid approximation of singularly perturbed boundary value problems with a regular boundary layer. *Soviet J. Numer. Anal. Math. Modelling*, 4(5):397–417.

[96] Silvester, D. and Pranjal (2016). An optimal solver for linear systems arising from stochastic FEM approximation of diffusion equations with random coefficients. *SIAM/ASA J. Uncertain. Quantif.*, 4(1):298–311.

[97] Smith, R. C. (2014). *Uncertainty quantification*, volume 12 of *Computational Science & Engineering*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. Theory, implementation, and applications.

[98] Sousedík, B. and Elman, H. C. (2016). Stochastic Galerkin methods for the steady-state Navier-Stokes equations. *J. Comput. Phys.*, 316:435–452.

[99] Stuart, A. M. (2010). Inverse problems: a Bayesian perspective. *Acta Numer.*, 19:451–559.

[100] Sudret, B. (2014). Polynomial chaos expansions and stochastic finite element methods. Risk and Reliability in Geotechnical Engineering (Chap. 6), K. K. Phoon and J. Ching (Eds.), pp. 265-300, CRC Press.

[101] Sullivan, T. J. (2015). *Introduction to uncertainty quantification*, volume 63 of *Texts in Applied Mathematics*. Springer, Cham.

[102] Teckentrup, A. L., Jantsch, P., Webster, C. G., and Gunzburger, M. (2015). A multilevel stochastic collocation method for partial differential equations with random input data. *SIAM/ASA J. Uncertain. Quantif.*, 3(1):1046–1074.

[103] Todor, R. A. and Schwab, C. (2007). Convergence rates for sparse chaos approximations of elliptic problems with stochastic coefficients. *IMA J. Numer. Anal.*, 27(2):232–261.

[104] Ullmann, E. (2010). A Kronecker product preconditioner for stochastic Galerkin finite element discretizations. *SIAM J. Sci. Comput.*, 32(2):923–946.

[105] Ullmann, E., Elman, H. C., and Ernst, O. G. (2012). Efficient iterative solvers for stochastic Galerkin discretizations of log-transformed random diffusion problems. *SIAM J. Sci. Comput.*, 34(2):A659–A682.

[106] Ullmann, E. and Powell, C. E. (2015). Solving log-transformed random diffusion problems by stochastic Galerkin mixed finite element methods. *SIAM/ASA J. Uncertain. Quantif.*, 3(1):509–534.

[107] Verfürth, R. (1994). A posteriori error estimation and adaptive mesh-refinement techniques. *J. Comput. Appl. Math.*, 50(1-3):67–83.

[108] Verfürth, R. (2013). *A posteriori error estimation techniques for finite element methods*. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford.

[109] Vogel, C. R. (2002). *Computational methods for inverse problems*, volume 23 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

[110] Wan, X. and Karniadakis, G. E. (2006). Multi-element generalized polynomial chaos for arbitrary probability measures. *SIAM J. Sci. Comput.*, 28(3):901–928.

[111] Wan, X. and Karniadakis, G. E. (2009). Error control in multi-element generalized polynomial chaos method for elliptic problems with random coefficients. *Commun. Comput. Phys.*, 5(2-4):793–820.

[112] Xiu, D. and Hesthaven, J. S. (2005). High-order collocation methods for differential equations with random inputs. *SIAM J. Sci. Comput.*, 27(3):1118–1139.

[113] Zhang, Z. (2003). Finite element superconvergence on Shishkin mesh for 2-D convection-diffusion problems. *Math. Comp.*, 72(243):1147–1177.