Positivity-Preserving Discretisations on General Meshes

PhD Thesis

Abdolreza Amiri

Department of Mathematics and Statistics
University of Strathclyde, Glasgow

January 29, 2026

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Abstract

This thesis focuses on the development and analysis of bound-preserving finite element methods for solving partial differential equations (PDEs), particularly convection-diffusion and reaction-diffusion problems. Bound-preserving methods are crucial for ensuring numerical stability and accuracy, especially in models where positivity of the solution is a key physical requirement. Examples include nonlinear reaction-diffusion systems modeling chemical concentrations, phase-field equations with global extrema constraints, and turbulence models (i.e. see [53,112]). Violations of these bounds can lead to unphysical solutions and instabilities, particularly in coupled systems where errors may propagate and amplify [66, 81].

To address these challenges, we extend the bound-preserving finite element method introduced in [12] to various settings. First, we develop a method for the steady-state convection-diffusion equation and establish its well-posedness and error estimates. Next, we extend this approach to time-dependent reaction-convection-diffusion equations, proving stability and error bounds for the implicit Euler time-stepping scheme. Finally, we adapt the method for polytopic meshes within the discontinuous Galerkin framework, demonstrating its effectiveness regardless of the geometry of the mesh.

The thesis presents mathematical analysis, including well-posedness proofs and error estimates, alongside numerical experiments that validate the proposed methods. These results contribute to the ongoing development of stable and accurate finite element techniques for PDEs, ensuring solutions remain physically meaningful within computational simulations.

Contents

List of Figures

List of Tables

Preface/Acknowledgements

This thesis presents the research conducted during my PhD at the University of Strathclyde in the field of finite element methods. This thesis includes results from my published papers and ongoing research. While some sections build upon existing work, considerable effort has been made to present the material in a cohesive and self-contained manner, ensuring clarity and accessibility for the reader.

Throughout my PhD journey, I have had the privilege of working under the supervision of Prof. Gabriel Barrenechea and collaborating with Prof. Emmanuil Georgoulis and Dr. Tristan Pryer. Their invaluable guidance and constructive feedback have significantly contributed to the progress of this research. I would also like to acknowledge the support and discussions with my colleagues, which have enriched my understanding and enhanced the quality of this work.

I sincerely appreciate the financial support provided by Leverhulme Trust and the University of Strathclyde, which has enabled me to dedicate myself fully to this research.

Finally, I am deeply grateful to my family and friends for their unwavering support and encouragement throughout this journey. Their patience and belief in me have been invaluable.

Chapter 0. Preface/Acknowledgements

Chapter 1

Introduction and preliminaries

In a great many areas of study, partial differential equations (PDEs) are used to describe models, laws and systems. From the simplest of examples, the equations governing heat transfer, through to trading models for global financial markets, PDEs gives us an approach that can tackle a vast and ever growing range of real world problems. We may understand and make predictions about the behaviour of complex mechanical systems, we may study the weather, or we may gain insights into biological systems. The scope of PDEs and their relevance to our lives is beyond doubt. In many of these systems we have very complex interactions for which analytical solutions are not practicable or even possible. Direct experimentation and measurement may likewise not be practical and is generally expensive. Numerical modelling is therefore the key tool to unlock our understanding of how these systems work or how they might evolve in time. The techniques for this are well-established and constantly being refined and improved. One effective numerical technique is the use of finite elements. The Finite Element Method (FEM) is a powerful and versatile numerical technique used for solving complex engineering and mathematical problems, particularly those involving PDEs over complicated domains [75, 109]. It is widely employed in fields such as structural analysis, fluid dynamics, heat transfer, and electromagnetics. FEM works by breaking down a large, complex problem into smaller, simpler parts, called finite elements. These elements are interconnected at nodes, which are points on the boundaries or within the elements themselves. The collection of elements and nodes forms a mesh that covers the entire domain of the problem.

In this chapter, we start by introducing the Sobolev spaces which are important in studying the weak formulation of the PDEs and will be used in the next chapters. We then state the strong maximum principles, which are crucial for analysing the solution of the elliptic and parabolic problems discussed later. Additionally, we present important theorems, such as the Lax-Milgram theorem which are used to establish the

well-posedness of the weak form of the solutions. Furthermore, key inequalities and theorems essential for the discussions in the subsequent chapters are also provided.

1.1 Sobolev spaces

Sobolev spaces play a crucial role in the finite element method (FEM) as they provide the mathematical framework for defining and analysing the function spaces in which the solutions of partial differential equations (PDEs) reside. These spaces allow us to formulate the problem, prove the existence and uniqueness of solutions, and ensure the accuracy and convergence of the finite element approximations.

The construction of Sobolev spaces relies on the properties of Lebesgue spaces, which we describe in the following section.

1.1.1 Lebesgue Spaces

This section presents some Lebesgue spaces.

Definition 1.1.1. *(Lebesgue Space $L^1(D)$) [59, Definition 1.21] Let D be an open set in $\mathbb{R}^d$. The space $L^1(D)$ consists of all real-valued measurable functions that are Lebesgue integrable on D. We define the norm on $L^1(D)$ by*

$$\|f\|_{L^1(D)} := \int_D |f| \, \mathrm{d}\boldsymbol{x}.$$

Definition 1.1.2. *(Lebesgue Space $L^1_{loc}(D)$) [59, Definition 1.29] Let D be an open set in $\mathbb{R}^d$. The elements of the following space are referred to as locally integrable functions:*

$$L^1_{loc}(D) := \{v \text{ measurable } \mid \text{for any compact set } K \subset D,\ v \in L^1(K)\}.$$

Definition 1.1.3. *(Lebesgue Spaces $L^p(D)$ and $L^\infty(D)$) [59, Definition 1.33] Let D be an open set in $\mathbb{R}^d$. For all $p \in [1, \infty)$, we define the space $L^p(D)$ as the set of measurable functions for which the $L^p(D)$ norm is finite, that is*

$$\|f\|_{L^p(D)} := \left(\int_D |f|^p \, \mathrm{d}\boldsymbol{x} \right)^{1/p} < \infty,$$

and the space $L^\infty(D)$ ($p = \infty$) as the set of measurable functions for which the L^∞ norm is finite

$$\|f\|_{L^\infty(D)} := ess\ sup\{|f(x)| : x \in D\}.$$

In this thesis $\|\cdot\|_{0,p,D}$ *and* $\|\cdot\|_{0,\infty,D}$ *are used to denote* $\|\cdot\|_{L^p(D)}$ *and* $\|\cdot\|_{L^\infty(D)}$ *respectively.*

In the finite element method, Cauchy–Schwarz and Hölder's inequalities are fundamental tools used in various aspects of analysis, including error estimates, stability analysis, and proving the well-posedness of variational problems (see [59, Lemma 1.40]).

Lemma 1.1.4. *(The Cauchy-Schwarz Inequality). Let* u *and* v *belong to* $L^2(D)$*. Then,* $uv \in L^1(D)$ *and*

$$\left| \int_D u(x)v(x)\,\mathrm{d}\boldsymbol{x} \right| \leq \|u\|_{0,D}\|v\|_{0,D}.$$

Remark 1.1.5. *The space* $L^2(D)$ *is a Hilbert space when equipped with the inner product*

$$\big(f,g\big)_D := \int_D f(x)g(x)\,\mathrm{d}x, \tag{1.1}$$

for $f, g \in L^2(D)$*.*

Lemma 1.1.6. *(Hölder's Inequality) Let* $p \in [1,\infty]$ *and* q *such that* $\frac{1}{p} + \frac{1}{q} = 1$*. For any* $u \in L^p(D)$ *and* $v \in L^q(D)$*, it holds that*

$$\|u \cdot v\|_{0,1,D} \leq \|u\|_{0,q,D}\|v\|_{0,p,D}.$$

Furthermore, equality is attained if and only if u *and* v *are linearly dependent, i.e., there exists* $\lambda \geq 0$ *such that* $u = \lambda v$ *almost everywhere.*

1.1.2 Sobolev spaces

In the definition of the Sobolev spaces the following space which is called the space of test functions have an important role.

Definition 1.1.7. *(Space* $C_0^\infty(D)$ *) [59, Definition 1.31] The space* $C_0^\infty(D)$ *consists of functions* $f : D \to \mathbb{R}$ *that are infinitely differentiable, i.e.,* $f \in C^\infty(D)$*, and have compact support included in* D*, where* D *is an open set. The members of* $C_0^\infty(D)$ *are called test functions.*

Definition 1.1.8. *(Weak derivative) [59, Definition 2.3] Let* D *be an open set in* $\mathbb{R}^d$*. Let* $u, v \in L^1_{loc}(D)$*. For each* $i \in \{1,\ldots,d\}$*, we say that* v *is the weak partial derivative of* u *in the direction* x_i *if*

$$\int_D u\partial_{x_i}\varphi\,\mathrm{d}\boldsymbol{x} = -\int_D v\varphi\,\mathrm{d}\boldsymbol{x}, \quad \forall\varphi \in C_0^\infty(D). \tag{2.2}$$

We denote this weak derivative by $\partial_{x_i} u = v$. More generally, for a multi-index $\boldsymbol{\alpha} \in \mathbb{N}^d$, we say that v is the weak $\boldsymbol{\alpha}$-th partial derivative of u and write $\partial^{\boldsymbol{\alpha}} u = v$ if

$$\int_D u \, \partial_{x_1}^{\alpha_1} \dots \partial_{x_d}^{\alpha_d} \varphi \, \mathrm{d}\boldsymbol{x} = (-1)^{|\alpha|} \int_D v\varphi \, \mathrm{d}\boldsymbol{x}, \quad \forall \varphi \in C_0^\infty(D), \tag{2.3}$$

where $|\boldsymbol{\alpha}| := \alpha_1 + \cdots + \alpha_d$. Finally, we write $\partial^{\boldsymbol{\alpha}} u := \partial_{x_1}^{\alpha_1} \dots \partial_{x_d}^{\alpha_d} \varphi$, and for $\boldsymbol{\alpha} = (0, \dots, 0)$, we set $\partial^{\boldsymbol{\alpha}} u := u$.

Lemma 1.1.9. *(Uniqueness) [59, Lemma 2.4] Let $u \in L^1_{loc}(D)$. If u has a weak α-th partial derivative, then it is uniquely defined.*

Now, we are ready introduce integer-order and fractional-order Sobolev spaces.

1.1.3 Integer-order spaces

Definition 1.1.10. *($W^{m,p}$ spaces) [59, Definition 2.8] Let $m \in \mathbb{N}$ and $p \in [1, \infty]$. Let D be an open set in $\mathbb{R}^d$. We define the Sobolev space*

$$W^{m,p}(D) := \{u \in L^p(D),\, \partial^{\boldsymbol{\alpha}} u \in L^p(D),\, \forall \boldsymbol{\alpha} \in \mathbb{N}^d,\, |\boldsymbol{\alpha}| \le m\}, \tag{2.4}$$

where the derivatives are understood in the weak sense. We define $W_0^{m,p}(D) = \overline{C_0^\infty(D)}$ whenever u has compact support in D. For $p = 2$, the space $W^{m,2}(D)$ is denoted by $H^m(D)$.

We equip $W^{m,p}(D)$ with the following norm and seminorm: If $p < \infty$, we set

$$\|u\|_{W^{m,p}(D)} := \left(\sum_{|\boldsymbol{\alpha}| \le m} \|\partial^{\boldsymbol{\alpha}} u\|^p_{L^p(D)} \right)^{1/p}, \quad |u|_{W^{m,p}(D)} := \left(\sum_{|\boldsymbol{\alpha}| = m} \|\partial^{\boldsymbol{\alpha}} u\|^p_{L^p(D)} \right)^{1/p},$$

and if $p = \infty$, we set

$$\|u\|_{W^{m,\infty}(D)} := \max_{|\boldsymbol{\alpha}| \le m} \|\partial^{\boldsymbol{\alpha}} u\|_{L^\infty(D)}, \quad |u|_{W^{m,\infty}(D)} := \max_{|\boldsymbol{\alpha}| = m} \|\partial^{\boldsymbol{\alpha}} u\|_{L^\infty(D)}.$$

Here, the sums and the maxima run over multi-indices $\boldsymbol{\alpha}$.

Proposition 1.1.11. *[59, Proposition 2.9] $W^{m,p}(D)$ is a real Banach space. When $p = 2$ the Sobolev space is denoted as $H^m(D) := W^{m,2}(D)$ which is a real Hilbert space when equipped with the inner product $(u, v)_{H^m(D)} := \sum_{|\boldsymbol{\alpha}| \le m} (D^{\boldsymbol{\alpha}} u, D^{\boldsymbol{\alpha}} v)_{L^2(D)}$. The norm induced by this inner product is denoted by $\|\cdot\|_{H^m(D)}$. Furthermore, The space $H_0^m(D) := W_0^{m,2}(D)$ is defined as the closure of the space of compactly supported*

smooth functions $C_0^\infty(D)$ in the $H^m(D)$-norm, i.e.,

$$H_0^m(D) = \overline{C_0^\infty(D)},$$

which is also a Hilbert space.

Remark 1.1.12. *For $s \geq 0$, $p \in [1, \infty]$, we denote by $\|\cdot\|_{s,p,D}$ ($|\cdot|_{s,p,D}$) the norm $\|u\|_{W^{m,p}(D)}$ (seminorm $|u|_{W^{m,\infty}(D)}$) in $W^{s,p}(D)$; when $p = 2$, we define $H^s(D) = W^{s,2}(D)$, and again omit the subscript p and only write $\|\cdot\|_{s,D}$ ($|\cdot|_{s,D}$).*

1.1.4 Fractional-order spaces

Definition 1.1.13. *Let $s \in (0, 1)$ and $p \in [1, \infty]$. Let D be an open set in $\mathbb{R}^d$. We define*

$$W^{s,p}(D) := \{\, v \in L^p(D) \mid |v|_{W^{s,p}(D)} < \infty \,\},$$

where

$$|v|_{W^{s,p}(D)} := \left(\int_D \int_D \frac{|v(x) - v(y)|^p}{\|x - y\|_{\ell^2}^{sp+d}} \, dx \, dy \right)^{1/p}, \qquad p < \infty, \tag{2.6}$$

and

$$|v|_{W^{s,\infty}(D)} := \operatorname*{ess\,sup}_{x,y \in D} \frac{|v(x) - v(y)|}{\|x - y\|_{\ell^2}^s}.$$

Letting now $s > 1$, we define

$$W^{s,p}(D) := \{\, v \in W^{m,p}(D) \mid \partial^\alpha v \in W^{\sigma,p}(D),\ \forall \alpha,\ |\alpha| = m \,\}, \tag{2.7}$$

where $m := \lfloor s \rfloor$ and $\sigma := s - m$.

1.1.5 Negative-order Sobolev spaces

The notion of distribution is a powerful tool that extends the concept of integrable functions and weak derivatives. In particular, we will see that every distribution is differentiable in some reasonable sense.

Definition 1.1.14. *(Distribution) [59, Definition 4.1] Let D be an open set in $\mathbb{R}^d$. A linear map*

$$T : C_0^\infty(D) \to \mathbb{R},$$

is called a distribution in D if for every compact subset K of D, there exist an integer p, called the order of T, and a real number c (both can depend on K) such that for all $\varphi \in C_0^\infty(D)$ with $supp(\varphi) \subseteq K$, we have

$$|\langle T, \varphi \rangle| \leq c \max_{|\boldsymbol{\alpha}| \leq p} \|\partial^{\boldsymbol{\alpha}} \varphi\|_{L^\infty(K)}.$$

Example 1. *(Locally integrable functions) [59, Example 4.2] Every function $v \in L^1_{loc}(D)$ can be identified with the following distribution:*

$$T_v : C_0^\infty(D) \ni \varphi \mapsto \langle T_v, \varphi \rangle = \int_D v\varphi \, \mathrm{d}\boldsymbol{x}.$$

With the concept of distributions in hand, we can now introduce Sobolev spaces of negative order by employing duality, specifically using $W_0^{s,p}(D)$.

Definition 1.1.15. *(Negative-order Sobolev spaces $W^{-s,p}(D)$) [59, Definition 4.10] Let $s > 0$ and $p \in (1, \infty)$. Let D be an open set in $\mathbb{R}^d$. The space $W^{-s,p}(D)$ is defined as $\left(W_0^{s,p'}(D)\right)'$ (the dual of $\left(W_0^{s,p'}(D)\right)$), where $\frac{1}{p} + \frac{1}{p'} = 1$. For the case $p = 2$, we denote $W^{-s,p}(D)$ by $H^{-s}(D)$. This space is equipped with the norm*

$$\|T\|_{W^{-s,p}(D)} = \sup_{w \in W_0^{s,p'}(D)} \frac{|\langle T, w \rangle|}{\|w\|_{W_0^{s,p'}(D)}}.$$

The element T in $W^{-s,p}(D)$ is considered a distribution, because if $s = m \in \mathbb{N}$, we have

$$|\langle T, \varphi \rangle| \leq \|T\|_{W^{-m,p}(D)} |D|^{\frac{1}{p'}} \binom{m+d}{d}^{\frac{1}{p'}} \max_{|\boldsymbol{\alpha}| \leq m} \|\partial^{\boldsymbol{\alpha}} \varphi\|_{L^\infty(K)},$$

for any compact subset $K \subseteq D$ and any $\varphi \in C_0^\infty(D)$ with $\mathrm{supp}(\varphi) \subseteq K$. This reasoning can be extended to cases where $s = m + \sigma$, $\sigma \in (0, 1)$.

Remark 1.1.16. *We denote by $H^{-1}(D)$ the dual of $H_0^1(D)$ while identifying $L^2(D)$ with its dual. Thus, writing $\langle \cdot, \cdot \rangle_D$ for the duality pairing, we have*

$$\langle f, v \rangle_D = \int_D f(\boldsymbol{x}) v(\boldsymbol{x}) d\boldsymbol{x} \qquad \forall\, v \in H_0^1(D),$$

whenever $f \in H^{-1}(D)$ is regular enough. We do not distinguish between inner product and duality pairing for scalar or vector-valued functions.

1.2 Elliptic equations

In this section, we focus primarily on the boundary-value problem

$$\begin{cases} Lu &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{cases} \tag{1.2}$$

where Ω is an open and bounded subset of $\mathbb{R}^d$, and $u : \overline{\Omega} \to \mathbb{R}$ is the unknown function, with $u = u(\boldsymbol{x})$. The function $f : \Omega \to \mathbb{R}$ is given, and L represents second-order linear partial differential operator given by

$$Lu = \sum_{i,j=1}^{d} \frac{\partial}{\partial x_j} \left(d_{ij}(\boldsymbol{x}) \frac{\partial u}{\partial x_i} \right) + \sum_{i=1}^{d} \beta_i(\boldsymbol{x}) \frac{\partial u}{\partial x_i} + \mu(\boldsymbol{x}) u,$$

where the coefficient functions d_{ij}, β_i, and μ are given, with indices $i, j = 1, \dots, d$.

Definition 1.2.1. *We say the partial differential operator L is (uniformly) elliptic if there exists a constant $d_0 > 0$ such that*

$$\sum_{i,j=1}^{d} d_{ij}(\boldsymbol{x}) y_i y_j \geq d_0 \sum_{i=1}^{d} y_i^2 \tag{1.3}$$

for almost every $\boldsymbol{x} \in U$ and for all $\mathbf{y} = (y_1, y_2, \cdots, y_d) \in \mathbb{R}^d$.

The standard weak formulation of (1.2) reads as follows:

$$\begin{cases} \text{Find } u \in V \text{ such that,} \\ a(u, w) = \ell(w) \qquad \forall w \in V, \end{cases} \tag{1.4}$$

where $a : V \times V \to \mathbb{R}$, denotes the bilinear form

$$a(u, v) := (Lu, w)_\Omega \qquad \forall u, w \in V. \tag{1.5}$$

Here, V is a Banach space endowed with the norm $\| \cdot \|_V$, and $(\cdot, \cdot)_\Omega$ is the inner product which has been defined in (1.1). Often, V is a Hilbert space. We assume that a is bounded, which means

$$\|a\|_{V \times V'} := \sup_{v \in V} \sup_{w \in V} \frac{|a(v, w)|}{\|v\|_V \|w\|_V} < \infty.$$

It is implicitly understood that this supremum is taken over non-zero arguments. Also, the map $\ell : V \to \mathbb{C}$

is assumed to be a linear form. The boundedness of ℓ means

$$\|\ell\|_{V'} := \sup_{w \in V \setminus \{0\}} \frac{|\ell(w)|}{\|w\|_V} < \infty.$$

V' is the dual space of V equipped with the above norm.

Definition 1.2.2. *(Well-posedness, Hadamard) [60, Definition 25.1] A problem of the form 1.4 is said to be well-posed if it has a unique solution $u \in V$ for every $\ell \in V'$, and if there exists a constant c that is uniform with respect to ℓ, such that the following a priori estimate holds*

$$\|u\|_V \leq c\|\ell\|_{V'}.$$

We will now present the Lax-Milgram lemma, which is an important theorem for studying the well-posedness of problems like (1.4).

Lemma 1.2.3. *(Lax-Milgram) [60, Lemma 25.2] Let V be a real Hilbert space, let a be a bounded bilinear form on $V \times V$, and let $\ell \in V'$. Assume the following coercivity property: There is a real number $\alpha > 0$ such that*

$$a(v, v) \geq \alpha \|v\|_V^2, \quad \forall v \in V. \tag{1.6}$$

Then (1.4) *is well-posed with the a priori estimate $\|u\|_V \leq \frac{1}{\alpha}\|\ell\|_{V'}$.*

One of the key aspects of well-posedness, as defined by Hadamard (see Definition (1.2.2)), is the existence of a solution. The Brouwer's fixed point theorem ensures that under certain conditions, a fixed point (which often corresponds to a solution) exists.

Theorem 1.2.4. *(Brouwer's fixed point theorem) [110, Theorem 10.41] Let C be a compact, convex, nonempty subset of $\mathbb{R}^d$, and suppose f is a continuous function that maps C into C. Then f has a fixed point in C; i.e., there exists $\mathbf{u} \in C$ such that*

$$f(\mathbf{u}) = \mathbf{u}.$$

Stampacchia's Theorem is also an important result in functional analysis, particularly in the study of variational inequalities and partial differential equations (PDEs). The lemma provides crucial conditions under which a variational inequality has a solution, specifically for problems involving coercive bilinear forms on closed, convex sets.

Let $\mathcal{K} \subset H$ be closed and convex, let $f \in H'$ be a linear form, and $a(\cdot,\cdot) : H \times H \to \mathbb{R}$ be a bilinear form. Consider the following problem:

$$\begin{cases} \text{Find } u \in \mathcal{K} \text{ such that,} \\ a(u, v-u) \geq \langle f, v-u \rangle \quad \text{for all } v \in \mathcal{K}, \end{cases} \tag{1.7}$$

then, Stampacchia's Theorem provides the following result regarding the existence and uniqueness of the solution of (1.7).

Theorem 1.2.5. *(Stampacchia's Theorem) [84, Theorem 2.1] Let $a(\cdot,\cdot)$ be an elliptic bilinear form on real Hilbert space H, $\mathcal{K} \subset H$ closed and convex, and $f \in H'$. Then there exists a unique solution to Problem* (1.7)*. In addition, the mapping $f \mapsto u$ is Lipschitz, that is, if u_1, u_2 are solutions to Problem* (1.7) *corresponding to $f_1, f_2 \in H'$, then*

$$\|u_1 - u_2\| \leq \frac{1}{\alpha} \|f_1 - f_2\|_{H'}.$$

Moreover, the mapping $f \mapsto u$ is linear if $\mathcal{K}$ is a subspace of H.

1.3 Parabolic equations

We assume Ω to be an open, bounded subset of $\mathbb{R}^d$, and set space-time domain $\Omega_T = \Omega \times (0, T]$ for some fixed time $T > 0$. We will study the initial/boundary-value problem

$$\begin{cases} u_t + Lu = f & \text{in } \Omega_T, \\ u = 0 & \text{on } \partial\Omega \times [0, T], \\ u^0 = g & \text{on } \Omega \times \{t = 0\}, \end{cases} \tag{1.8}$$

where $f : \Omega_T \to \mathbb{R}$ and $g : \Omega \to \mathbb{R}$ are given, and $u : \overline{\Omega_T} \to \mathbb{R}$ is the unknown, $u = u(\boldsymbol{x}, t)$. The letter L denotes for each time t a second-order partial differential operator

$$Lu = -\sum_{i,j=1}^{d} \frac{\partial}{\partial x_j} \left(d_{ij}(\boldsymbol{x}) \frac{\partial u}{\partial x_i} \right) + \sum_{i=1}^{d} \beta_i(\boldsymbol{x}) \frac{\partial u}{\partial x_i} + \mu(\boldsymbol{x}) u.$$

We assume that L satisfies in ellipticity condition (1.3).

Remark 1.3.1. *Boundary conditions play a central role in the well-posedness and physical interpretation of partial differential equations. The two most common types are* Dirichlet *and* Neumann *conditions.*

Dirichlet boundary conditions *prescribe the value of the solution directly along the boundary, that is,*

$$u = g \quad on\ \Gamma_D \subset \partial\Omega.$$

They correspond to fixing the state of the system (for example, temperature or concentration) on the boundary.

Neumann boundary conditions *instead prescribe the normal derivative of the solution,*

$$\frac{\partial u}{\partial n} = g \quad on\ \Gamma_N \subset \partial\Omega,$$

where n is the outward unit normal. In physical terms, these conditions correspond to specifying a flux across the boundary (e.g., heat flux, mass flux).

To write the weak form of (1.8), for all $t \in [0,T]$, we consider the space V. The standard weak formulation of (1.8) reads as follows: for almost all $t \in (0,T)$ find $u \in V$, such that

$$\begin{cases} (\partial_t u, v)_\Omega + a(u,v) = (f,v)_\Omega & \forall v \in V, \\ u(\cdot,0) = u^0, \end{cases} \tag{1.9}$$

where $a(\cdot,\cdot)$, with a slight abuse of notation, denotes the bilinear form

$$a(u,v) := (Lu,v)_\Omega \qquad \forall u, v \in V \text{and } t \in (0,T), \tag{1.10}$$

where $(\cdot,\cdot)_\Omega = (\cdot,\cdot)_{L^2(\Omega)}$ refers to the inner product introduced in 1.1.11. Now, we introduce a space-time norm which we will be use in the next chapters.

Definition 1.3.2. *(space-time norm) Let $T > 0$ be a fixed time, let Ω be a domain in $\mathbb{R}^d$, and define the space-time domain $\Omega \times (0,T]$. Let u be a function defined on $\Omega \times (0,T]$. An alternative way of looking at u is to treat it as a function of t with values in a Banach space, say V, whose elements are functions depending only on the space variable:*

$$u : (0,T] \to V \quad such\ that \quad u(t) = u(\cdot,t) \in V.$$

In this spirit, we consider the following spaces:

For $1 \leq p \leq +\infty$, $L^p([0,T];V)$ is the space of V-valued functions whose norm in V is in $L^p([0,T])$; this space is a Banach space for the norm

$$\|u\|_{L^p([0,T];V)} = \begin{cases} \left(\int_0^T \|u(t)\|_V^p \mathrm{d}t\right)^{1/p} & \text{if } 1 \leq p < +\infty, \\ ess\ sup_{t\in[0,T]} \|u(t)\|_V & \text{if } p = +\infty. \end{cases}$$

So, if we set $V = W^{s,q}(D)$, for $1 \leq p, q \leq +\infty$, $L^p((0,T); W^{s,q}(D))$ is the space defined by

$$L^p((0,T); W^{s,q}(D)) = \left\{ u(t,\cdot) \in W^{s,q}(D) \text{ for almost all } t \in [0,T] : \|u\|_{s,q,D} \in L^p(0,T) \right\},$$

and the norm is modified as

$$\|u\|_{L^p((0,T);W^{s,q}(D))} = \begin{cases} \left(\int_0^T \|u\|_{s,q,D}^p \mathrm{d}t\right)^{\frac{1}{p}} & \text{if } \ 1 \leq p < \infty, \\ \text{ess sup}_{t\in(0,T)} \|u\|_{s,q,D} & \text{if } \ \ p = \infty. \end{cases}$$

Gronwall's Lemma is a pivotal mathematical tool used extensively in the analysis of time-dependent problems, particularly in proving the stability of solutions and also the error analysis in numerical methods for differential equations. Gronwall's Lemma provides bounds for functions that satisfy certain integral or differential inequalities. It comes in several forms, but in this thesis we use the following form which is proved originally in [72, Lemma 5.1].

Lemma 1.3.3. *Let k, B, a_n, b_n, c_n, γ_n, $n = 0, \ldots, m$, be non-negative numbers such that*

$$a_n + k\sum_{n=0}^{m} b_n \leq k\sum_{n=0}^{m} \gamma_n a_n + k\sum_{n=0}^{m} c_n + B \quad \text{for} \quad m \geq 0.$$

Suppose $k\gamma_n \leq 1$ for every j, and set $\sigma_n = (1 - k\gamma_n)^{-1}$. Then

$$a_m + k\sum_{n=0}^{m} b_n \leq \exp\left(k\sum_{n=0}^{m} \sigma_n\gamma_n\right)\left(k\sum_{n=0}^{m} c_n + B\right) \quad \text{for} \quad m \geq 0. \tag{1.11}$$

1.4 Lipschitz set and Lipschitz domain

A Lipschitz domain or Lipschitz set $D \subset \mathbb{R}^d$ is a domain where the boundary can be locally represented as the graph of a Lipschitz continuous function. Finite element formulations rely on Sobolev spaces. The well-posedness of problems involving these spaces requires the domain Ω to have at least a Lipschitz boundary. If

the domain is not Lipschitz (e.g., if it has sharp cusps or fractal boundaries), Sobolev embeddings and trace theorems may fail, leading to difficulties in defining weak formulations properly.

Definition 1.4.1. *(Lipschitz set and domain) (see [59, definition 3.2]) An open set $D \subset \mathbb{R}^d$, $d \geq 2$, is defined as a Lipschitz set if for all $\boldsymbol{x} \in \partial D$, there exists a neighbourhood V_x of $\boldsymbol{x}$ in $\mathbb{R}^d$, a rotation $R_x : \mathbb{R}^d \to \mathbb{R}^d$, and two real numbers $\alpha > 0$ and $\beta > 0$ (which may depend on $\boldsymbol{x}$), such that the following conditions hold:*

(i) $V_{\boldsymbol{x}} = \boldsymbol{x} + R_{\boldsymbol{x}}\left(B_\alpha \times I_\beta\right)$ *where* $B_\alpha := B_{d-1}(0,\alpha) = B_\alpha(\boldsymbol{x}) := \{\mathbf{y} \in \mathbb{R}^{d-1} \mid \|\boldsymbol{y} - \boldsymbol{x}\| < \alpha\}$, $I_\beta := (-\beta, \beta)$.

(ii) There exists a Lipschitz function $\varphi_{\boldsymbol{x}} : B_\alpha \to \mathbb{R}$ *such that* $\varphi_{\boldsymbol{x}(0)} = 0$, $\|\varphi_{\boldsymbol{x}}\|_{L^\infty(B_\alpha)} \leq \frac{1}{2}\beta$, *and (see Figure 1.1):*

$$D \cap V_{\boldsymbol{x}} = \boldsymbol{x} + R_{\boldsymbol{x}}\left(\left\{(\boldsymbol{y}', \boldsymbol{y}_d) \in B_\alpha \times I_\beta \mid \boldsymbol{y}_d < \varphi_x(\boldsymbol{y}')\right\}\right),$$

$$\partial D \cap V_{\boldsymbol{x}} = \boldsymbol{x} + R_{\boldsymbol{x}}\left(\left\{(\boldsymbol{y}', \boldsymbol{y}_d) \in B_\alpha \times I_\beta \mid \boldsymbol{y}_d = \varphi_{\boldsymbol{x}}(\boldsymbol{y}')\right\}\right).$$

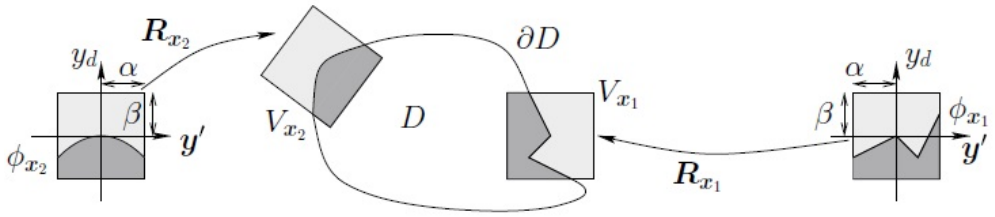


Figure 1.1: Lipschitz domain and mappings (R_{x_1}, ϕ_{x_1}), (R_{x_2}, ϕ_{x_2}).

Lemma 1.4.2. *(Poincaré–Steklov) [59, Lemma 3.24] Let D be a Lipschitz domain in $\mathbb{R}^d$. Define $\ell_D = diam(D)$. For $p \in [1, \infty)$, there exists a constant $C_{P,S,p}$ (where the subscript p is omitted when $p = 2$) such that:*

$$C_{P,S,p}\|v - v_D\|_{0,p,D} \leq \ell_D\|\nabla v\|_{0,p,D}, \quad \forall v \in W^{1,p}(D),$$

where $v_D := \frac{1}{|D|}\int_D v \, \mathrm{d}\boldsymbol{x}$. *The following also holds when D is convex:*

$$C_{P,S,1} = 2, \quad C_{P,S,2} = \pi, \quad C_{P,S,p} \geq \frac{2}{p}\left(\frac{2}{p}\right)^{\frac{1}{p}}, \quad p > 1. \tag{1.12}$$

Remark 1.4.3. *The constants presented in* (1.12) *were proven by Acosta and Durán [1] for $p = 1$, and Bebendorf [19] for $p = 2$ (see also Payne and Weinberger [106]). Also, Chua and Wheeden (see [45, Theorem 1.2]) extended the results in* (1.12) *for general p.*

Lemma 1.4.4. *(Poincaré–Steklov) [59, Lemma 3.27] Let $p \in [1, \infty)$ and let D be a Lipschitz domain. Let $\ell_D := \mathrm{diam}(D)$. There is a constant $C_{ps,p} > 0$ (the subscript p is omitted when $p = 2$) such that*

$$C_{ps,p}\|v\|_{0,p,D} \leq \ell_D \|\nabla v\|_{0,p,D}, \quad \forall v \in W_0^{1,p}(D).$$

Proof. See Brezis [27, Corollary 9.19], and Evans [61, Theorem 3 section 5.6.], □

1.5 Maximum principles

The Maximum Principle is a significant concept in the theory of partial differential equations (PDEs) and plays an important role in the analysis and application of the Finite Element Method (FEM), particularly for elliptic and parabolic PDEs. It provides crucial insights into the behavior of solutions, including bounds and uniqueness, and has implications for the numerical methods used to approximate these solutions. The principle provides a way to estimate the solution's maximum and minimum values, which are crucial for understanding the behavior of the solution and for numerical approximation.

When applying FEM to solve PDEs that satisfy a Maximum Principle, it is important that the numerical method preserves this principle. If the FEM solution violates the Maximum Principle, it might indicate instability or inaccuracy in the numerical method, especially on coarse or distorted meshes. The Maximum Principle can be used to derive error estimates for the FEM approximation. If the exact solution satisfies the Maximum Principle, then the FEM approximation should ideally exhibit similar behaviour. By comparing the maximum (or minimum) values of the FEM solution to those of the exact solution, one can assess the quality of the approximation. For example, if a finite element solution exceeds the known bounds provided by the Maximum Principle, this might suggest the presence of discretisation errors or insufficient mesh refinement.

In the following we present the Maximum principles for problem (1.2) and problem (1.8).

Theorem 1.5.1. *(weak maximum principle) [110, Theorem 4.1] Consider the problem* (1.2)*. Assume that $Lu \geq 0$ (or, respectively, $Lu \leq 0$) in a bounded domain Ω in $\mathbb{R}^d$. Then the minimum (or, respectively, the maximum) of u is achieved on $\partial\Omega$.*

We have the following corollary of Theorem 1.5.1.

Corollary 1.5.2. *[110, Corollary 4.3] Consider the problem* (1.2)*. Let Ω be bounded and assume $\mu \leq 0$ in*

Ω. *Let* $Lu \geq 0$ *(or, respectively,* $Lu \leq 0$*) in* Ω. *Then*

$$\min_{\overline{\Omega}} u = \min_{\partial\Omega} u \quad (\textit{or, resp.}, \max_{\overline{\Omega}} u = \max_{\partial\Omega} u).$$

The following corollary is typically used in applications. It yields a uniqueness result as well as a comparison principle.

Corollary 1.5.3. *[110, Corollary 4.4] Let* Ω *be bounded and* $\mu \leq 0$. *If* $Lu = Lv$ *in* Ω *and* $u = v$ *on* $\partial\Omega$, *then* $u = v$ *in* Ω. *If* $Lu \leq Lv$ *in* Ω *and* $u \leq v$ *on* $\partial\Omega$, *then* $u \leq v$ *in* Ω.

Theorem 1.5.4. *(Strong maximum principle for elliptic equations) [61, Theorem 3, Section 6.4.2] Consider the problem* (1.2). *Assume* $u \in C^2(\Omega) \cap C(\overline{\Omega})$ *(where* $C^2(\Omega) \cap C(\overline{\Omega})$ *consists of functions that are twice continuously differentiable in* Ω *and continuous up to the closure of* Ω*) and* $\mu \geq 0$ *in* Ω.

Suppose also Ω *is connected.*

(i) If $Lu \leq 0$ *in* Ω *and* u *attains a nonnegative maximum over* $\overline{\Omega}$ *at an interior point, then* u *is constant within* Ω.

(ii) Similarly, if $Lu \geq 0$ *in* Ω *and* Ω *attains a nonpositive minimum over* $\overline{\Omega}$ *at an interior point, then* u *is constant within* Ω.

Theorem 1.5.5. *(Strong Maximum Principle for parabolic equations) [62, Theorem 12, Section 7.1] Consider the problem* (1.8). *Let* u *be a function belonging to* $C_2^1(\Omega_T) \cap C(\overline{\Omega_T})$, *with* $\mu \geq 0$ *in* Ω_T. *Assume that* Ω *is connected.*

(1) If

$$\partial_t u + Lu \leq 0 \quad \textit{in } \Omega_T$$

and u *reaches a nonnegative maximum within* $\overline{\Omega_T}$ *at some point* $(x_0, t_0) \in \Omega_T$, *then* $u(\cdot, t_0)$ *must be constant on* $\Omega \times \{t_0\}$.

(2) Likewise, if

$$\partial_t u + Lu \geq 0 \quad \textit{in } \Omega_T$$

and u *reaches a nonpositive minimum within* $\overline{\Omega_T}$ *at some point* $(x_0, t_0) \in \Omega_T$, *then* $u(\cdot, t_0)$ *must be constant on* $\Omega \times \{t_0\}$.

1.6 The Galerkin method

The central idea in the Galerkin method is to replace the infinite-dimensional space V by finite-dimensional space V_h. We assume that $V_h \neq \{0\}$. The discrete problem is formulated as follows:

$$\begin{cases} \text{Find } u_h \in V_h \text{ such that ,} \\ a_h(u_h, w_h) = \ell_h(w_h) \quad \forall w_h \in V_h, \end{cases} \tag{1.13}$$

where a_h is a bounded bilinear form on $V_h \times V_h$, and ℓ_h is a bounded linear form on V_h. Note that a_h and ℓ_h may differ from a and ℓ respectively. Since V_h is finite-dimensional, problem (1.13) is referred to as the discrete problem. The space V_h is called the discrete space (or discrete solution space).

Definition 1.6.1. *(Conforming Setting) [60, Definition 26.2] The approximation* (1.13) *is said to be conforming if* $V_h \subset V$.

Remark 1.6.2. *In this thesis, Chapters 2 and 3 follow a conforming framework, whereas the final chapter employs the discontinuous Galerkin method, requiring a non-conforming setting.*

The diameter of any set $G \subset \mathbb{R}^d$ is denoted by h_G, and the mesh size is defined as $h = \max\{h_K : K \in \mathcal{P}\}$. Since the analysis of the interpolation error implicitly involves sequences of successively refined meshes, we denote by $(\mathcal{P}_h)_{h\in\mathcal{H}}$ a sequence of meshes discretising a domain $D \subset \mathbb{R}^d$, where the index h belongs to a countable set $\mathcal{H}$ with zero as its only accumulation point.

Definition 1.6.3. *(Shape-regularity) [60, Definition 11.2] A sequence of affine meshes* $(\mathcal{T}_h)_{h\in\mathcal{H}}$ *is said to be shape-regular if there is* $\sigma_t > 0$ *independent of h, such that*

$$\sigma_K := \frac{h_K}{\rho_K} \leq \sigma_t, \quad \forall K \in \mathcal{T}_h, \forall h \in \mathcal{H}, \tag{11.4}$$

where ρ_k *denotes the diameter of the largest ball contained in K (see Figure 1.2).*

When the context is unambiguous, we will say that $(\mathcal{P}_h)_{h\in\mathcal{H}}$ is regular instead of shape-regular.

Definition 1.6.4. *(Quasi-uniformity) [60, Definition 22.20] A mesh sequence* $(\mathcal{P}_h)_{h\in\mathcal{H}}$ *is said to be quasi-uniform if it is shape-regular and if there exists a constant c such that*

$$h_K \geq ch, \quad \forall K \in \mathcal{T}_h, \forall h \in \mathcal{H}. \tag{22.38}$$

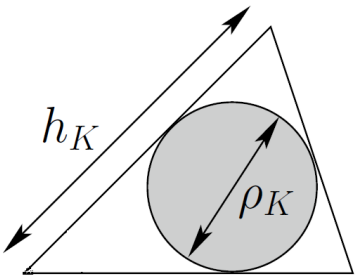


Figure 1.2: Triangular cell K and largest inscribed ball.

Remark 1.6.5. *One motivation for Definition 1.6.3 is to use the local inverse inequalities, which are presented in 1.26 and 1.27. In contrast, Definition 1.6.4 is intended for applying global inverse inequalities. To lighten the notation, throughout this thesis, we drop the index h from $(\mathcal{P}_h)_{h\in\mathcal{H}}$ and include it only when it is necessary to specify the mesh.*

We are now prepared to introduce the notations required for the conforming finite element methods that will be used in Chapters 2 and 3 below.

Let $\mathcal{P}$ be a conforming, shape-regular, quasi-uniform partition of the domain Ω into closed simplices or affine quadrilateral/hexahedral elements. For $k \geq 1$, we define the finite element spaces over $\mathcal{P}$ as follows:

$$\tilde{V}_{\mathcal{P}} := \{v_h \in C^0(\overline{\Omega}) : v_h|_K \in \mathfrak{R}(K)\ \forall K \in \mathcal{P}\}, \tag{1.14}$$

$$V_{\mathcal{P}} := \tilde{V}_{\mathcal{P}} \cap H_0^1(\Omega), \tag{1.15}$$

where the function space $\mathfrak{R}(K)$ is defined as:

$$\mathfrak{R}(K) = \begin{cases} \mathbb{P}_k(K), & \text{if } K \text{ is a simplex,} \\ \mathbb{Q}_k(K), & \text{if } K \text{ is an affine quadrilateral/hexahedral,} \end{cases} \tag{1.16}$$

with $\mathbb{P}_k(K)$ representing polynomials of total degree k on K, and $\mathbb{Q}_k(K)$ denoting the mapped space of polynomials of degree at most k in each variable.

For any mesh $\mathcal{P}$, we use the following notation:

- let $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ denote the set of internal nodes, and let $\phi_1, \ldots, \phi_N$ represent the corresponding Lagrangian basis functions associated with these nodes, which span the space $V_{\mathcal{P}}$;

- define $\mathcal{F}_I$ as the set of internal facets, $\mathcal{F}_\partial$ as the set of boundary facets, and $\mathcal{F}_h = \mathcal{F}_I \cup \mathcal{F}_\partial$ as the set of all facets in $\mathcal{P}$. For any element $K \in \mathcal{P}$, $\mathcal{F}_K$ represents the set of facets of K;
- for any element $K \in \mathcal{P}$, facet $F \in \mathcal{F}_h$, and node $\boldsymbol{x}_i$, we define the following neighbourhood:

$$\begin{aligned}\omega_K &= \bigcup \{K' \in \mathcal{P} : K \cap K' \neq \emptyset\},\\ \omega_F &= \bigcup \{K \in \mathcal{P} : F \subset K\},\\ \omega_i &= \bigcup \{K \in \mathcal{P} : \boldsymbol{x}_i \in K\};\end{aligned}$$

- define $\mathcal{E}_I$ as the set of internal edges, $\mathcal{E}_\partial$ as the set of boundary edges, and $\mathcal{E}_h = \mathcal{E}_I \cup \mathcal{E}_\partial$ as the set of all edges in $\mathcal{P}$. For any element $K \in \mathcal{P}$, $\mathcal{E}_K$ represents the set of edges of K;
- for any facet $E \in \mathcal{E}_h$, we define the following neighborhoods:

$$\omega_E = \bigcup \{K \in \mathcal{P} : E \subset K\},$$

- for any internal facet $F \in \mathcal{F}_I$, we use $[\![\cdot]\!]$ to denote the jump of a function across F.

Definition 1.6.6. *(Properties of meshes) [16, Definition Definition 2.2] A mesh $\mathcal{P}$ will be said to be connected if, for any two vertices $\boldsymbol{x}_i, \boldsymbol{x}_j$, there exists a path $j_0, \dots, j_s$ such that $E_{j_0,j_1}, \dots, E_{j_s,j}$ are all edges of $\mathcal{P}$. In addition, the mesh $\mathcal{P}$ will be said to be*

- *weakly acute if every internal dihedral angle θ of the mesh satisfies $\theta \leq \frac{\pi}{2}$;*
- *of Xu–Zikatanov (XZ) type (see [118]) if, for every $E \in \mathcal{E}_I$, the following holds (see Figure 1.3):*

$$\sum_{K \subset \omega_E} |\kappa_E^K| \cot \theta_E^K \geq 0;$$

  where $\kappa_E^K = F_i^K \cap F_j^K$, and when $d = 2$ we will adopt the convention $|\kappa_E^K| = 1$.
- *of Delaunay type if the interior of the circumscribed sphere of any simplex from the mesh $\mathcal{P}$ does not contain any vertex of $\mathcal{P}$.*

For $d = 2$, the definition of a Delaunay mesh can be equivalently stated as follows: for every $E = K \cap K' \in \mathcal{E}_I$, there holds

$$\theta_E^K + \theta_E^{K'} \leq \pi.$$

In two dimensions, the XZ-criterion and the Delaunay property are equivalent (see [119, Lemma 2.1]).

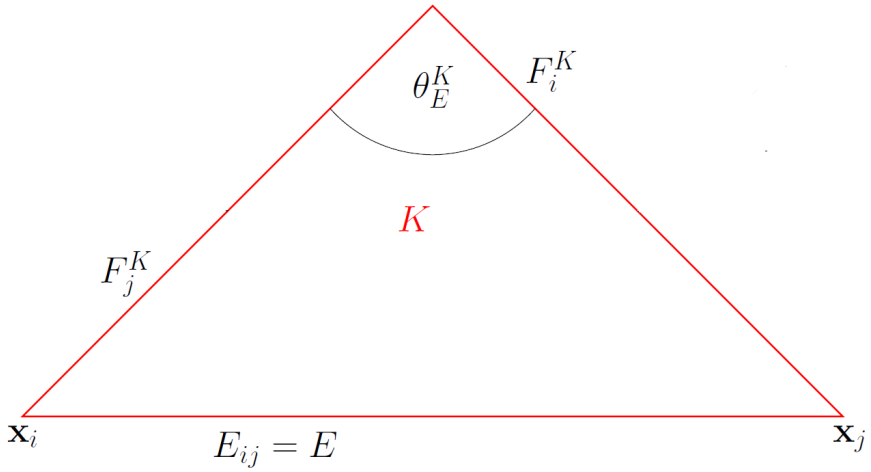


Figure 1.3: Notations for a triangle.

We also introduce the mesh function $\mathfrak{h}$, a continuous element-wise linear function often used in finite element analysis [102]. This mesh function is defined by averaging the element diameters at each vertex. For this, let $\mathbf{v}_1, \ldots, \mathbf{v}_M$ represent the set of mesh vertices. Then, $\mathfrak{h}$ is the piecewise linear function given by:

$$\mathfrak{h}(\mathbf{v}_i) = \frac{\sum_{K : \mathbf{v}_i \in K} h_K}{\#\{K : \mathbf{v}_i \in K\}}. \tag{1.17}$$

In the construction and analysis of the finite element methods proposed in this thesis, we use the mass-lumped L^2 inner product, which is defined for any $v_h, w_h \in V_{\mathcal{P}}$ as

$$(v_h, w_h)_h = \sum_{i=1}^{N} \mathfrak{h}(\boldsymbol{x}_i)^d v_h(\boldsymbol{x}_i) w_h(\boldsymbol{x}_i). \tag{1.18}$$

This inner product induces the norm:

$$|v_h|_h = (v_h, v_h)_h^{1/2}, \tag{1.19}$$

which is equivalent to the standard $L^2(\Omega)$ norm. More specifically, according to [60], Propositions 28.5 and 28.6, there exist constants $C, c > 0$, independent of h, such that:

$$c \sum_{i : \boldsymbol{x}_i \in K} h_K^d v_h^2(\boldsymbol{x}_i) \leq \|v_h\|_{0,K}^2 \leq C \sum_{i : \boldsymbol{x}_i \in K} h_K^d v_h^2(\boldsymbol{x}_i), \quad \forall K \in \mathcal{P}, \tag{1.20}$$

and as a consequence of the shape-regularity of the mesh, we also have:

$$c|v_h|_h^2 \leq \|v_h\|_{0,\Omega}^2 \leq C|v_h|_h^2, \quad \forall v_h \in V_{\mathcal{P}}. \tag{1.21}$$

With the definition of the finite element space (1.15) and the previously stated concepts, we now introduce several definitions, inequalities, and properties that will be frequently used in the subsequent chapters.

We start by introducing the Lagrange interpolation operator, which provides *interpolation error bounds* that are crucial for deriving error estimates for finite element solutions.

Definition 1.6.7. *(see [59, Chapter 11]) The Lagrange interpolation operator is defined as*

$$\begin{aligned} i_h &: C^0(\overline{\Omega}) \cap H_0^1(\Omega) \longrightarrow V_{\mathcal{P}}, \\ v &\longmapsto i_h v = \sum_{j=1}^{N} v(x_j)\phi_j, \end{aligned} \tag{1.22}$$

The error estimates of the Lagrange interpolant is crucial in understanding the convergence properties of finite element methods. Therefore, we can state the following result for the Lagrange interpolation error:

Proposition 1.6.8. *(Approximation property of the Lagrange interpolant) [59, Proposition 1.12] Let* $1 \leq \ell \leq k$ *and* i_h *be the Lagrange interpolant. Then, there exists* $C > 0$*, independent of* h*, such that for all* h *and* $v \in H^{\ell+1}(\Omega)$ *the following holds:*

$$\|v - i_h v\|_{0,K} + h|v - i_h v|_{1,K} \leq C h_K^{\ell+1} |v|_{\ell+1,K}\,. \tag{1.23}$$

Also, the orthogonal projection operator plays a significant role in the error analysis of the finite element method (FEM).

Definition 1.6.9. *(see [59], Chapter 22) The* $L^2(\Omega)$ *-orthogonal projection operator* $\pi : L^2(\Omega) \longrightarrow V_{\mathcal{P}}$ *is defined as follows*

$$\begin{aligned} \pi &: L^2(\Omega) \longrightarrow V_{\mathcal{P}}, \\ & w \longmapsto \pi(w) \text{ where } \left(\pi(w), v_h\right)_\Omega = (w, v_h)_\Omega \quad \forall\, v_h \in V_{\mathcal{P}}\,. \end{aligned} \tag{1.24}$$

Similar to the Lagrange interpolant, we use the following error bounds for the orthogonal projection in this thesis:

Proposition 1.6.10. *(Approximation property of the L^2-orthogonal projection operator) [59, Section 22.5] Let $0 \leq \ell \leq k$ and π be the $L^2(\Omega)$-orthogonal projection. Then, there exists $C > 0$, independent of h, such that for all h and $v \in H^{\ell+1}(\Omega)$ the following holds:*

$$\|v - \pi(v)\|_{0,K} + h\,|v - \pi(v)|_{1,K} \leq C h_K^{\ell+1} |v|_{\ell+1,K}\,. \tag{1.25}$$

The inverse inequality is fundamental in the error analysis of the FEM because it relates norms of finite element functions at different levels of smoothness. This is crucial in FEM because finite element spaces are typically designed to approximate solutions with certain regularity. When performing error analysis, especially in a posteriori estimates or stability analysis, controlling high derivatives is important, and the inverse inequality helps achieve this.

Lemma 1.6.11. *(Inverse inequality) (see [59, Lemma 12.1]) For all $m, \ell \in \mathbb{N}_0$, $0 \leq m \leq \ell$ and all $p, q \in [1, \infty]$, there exists a constant C, independent of h, such that*

$$|v_h|_{\ell,p,K} \leq C h_K^{m-\ell+d\left(\frac{1}{p}-\frac{1}{q}\right)} |v_h|_{m,q,K} \qquad \forall\, v_h \in V_{\mathcal{P}}\,. \tag{1.26}$$

Another important inequality is the trace inequality. The trace inequality (here locally i.e. on each element) is important in the FEM for several reasons, particularly in the context of error analysis. The trace inequality allows to estimate the norm of a function on the boundary of an element K in terms of its norms inside the same element. This provides a local control of errors, helping to manage how well the finite element function approximates the true solution near the boundaries of each element. We will repeatedly use the following trace inequality in the subsequent chapters.

Lemma 1.6.12. *(Discrete Trace inequality) (see [59, Lemma 12.8]) There exists $C > 0$ independent of h such that, for every $v \in H^1(K)$ the following holds*

$$\|v\|_{0,\partial K}^2 \leqslant C \left(h_K^{-1} \|v\|_{0,K}^2 + h_K |v|_{1,K}^2 \right). \tag{1.27}$$

1.6.1 Non-conforming approximations and discontinuous Galerkin method

In the last chapter of this thesis we study the approximation of an elliptic (reaction-convection-diffusion) problem by the discontinuous Galerkin (dG) method. The distinctive feature of dG methods is the spaces are broken finite element spaces. dG formulations are obtained by adding boundary penalty and a consistency term at all the mesh interfaces and boundary faces. Boundary conditions are weakly enforced and continuity

across the mesh interfaces is weakly enforced by penalising the jumps (see e.g. [60, Section 27.2.1 and Chapter 38]).

Before giving a short introduction and literature review on the discrete maximum principle (DMP), we first present some definitions and preliminaries.

1.6.2 Discrete maximum principle and algebraic flux correction

This section outlines the general conditions under which discrete maximum principles (DMPs) hold for both linear and nonlinear discretisations. It then briefly introduces a class of finite element methods called algebraic flux correction (AFC), which are designed to respect the discrete maximum principle.

To introduce the main idea of the algebraic flux correction method, we consider the steady-state model problem: Find $u : \overline{\Omega} \to \mathbb{R}$ such that

$$-\varepsilon \Delta u + \boldsymbol{\beta} \cdot \nabla u + \mu u = f \quad \text{in } \Omega, \tag{1.28}$$

$$u = g \quad \text{on } \partial\Omega. \tag{1.29}$$

To simplify the following presentation, we will suppose that $\varepsilon > 0$ and $\mu \geq 0$ are constants and that $\boldsymbol{\beta}$ is solenoidal. Let $\boldsymbol{\beta} \in W^{1,\infty}(\Omega)^d$, $f \in L^2(\Omega)$, and $g \in H^{1/2}(\partial\Omega)$, then the weak formulation of (1.29) reads as follows: Find $u \in H^1(\Omega)$ such that $u|_{\partial\Omega} = g$ and

$$a(u, v) = (f, v)_\Omega \qquad \forall v \in H_0^1(\Omega), \tag{1.30}$$

where $a(\cdot, \cdot)$ is the bilinear form given by

$$a(u, v) = \varepsilon(\nabla u, \nabla v)_\Omega + (\boldsymbol{\beta} \cdot \nabla u + \sigma u, v)_\Omega. \tag{1.31}$$

Linear Discretisations

Consider a matrix $A = (a_{ij})_{i=1,\dots,M}^{j=1,\dots,N} \in \mathbb{R}^{M \times N}$ and let $f_1, \dots, f_M$ and $g_1, \dots, g_{N-M}$ be given real values, where $M < N$. The discretised system leads to a linear system where we seek a vector $u = (u_1, \dots, u_N)^T \in \mathbb{R}^N$ such that:

$$\sum_{j=1}^{N} a_{ij} u_j = f_i, \quad \text{for } i = 1, \dots, M, \tag{1.32}$$

$$u_i = g_{i-M}, \quad \text{for } i = M+1, \dots, N. \tag{1.33}$$

Remark 1.6.13. *The full system matrix corresponding to equations* (1.32) *and* (1.33) *can be represented in block form as*

$$A = \begin{pmatrix} A_I & A_B \\ 0 & I \end{pmatrix}, \tag{1.34}$$

where:

- $A_I \in \mathbb{R}^{M \times M}$ *is the matrix associated with the internal (i.e., non-Dirichlet) degrees of freedom,*
- $A_B \in \mathbb{R}^{M \times (N-M)}$ *couples the boundary (Dirichlet) values to the interior,*
- $I \in \mathbb{R}^{(N-M) \times (N-M)}$ *is the identity matrix,*
- $0 \in \mathbb{R}^{(N-M) \times M}$ *is the zero matrix.*

Unless otherwise stated, the system matrix A will be considered in the form given by equation (1.34).

Definition 1.6.14 (Matrix of Nonnegative Type)**.** *Let* $A = (a_{ij})_{i=1,\dots,m}^{j=1,\dots,n} \in \mathbb{R}^{m \times n}$ *for* $m, n \in \mathbb{N}$*. We say that* A *is of* nonnegative type *if the following conditions hold*

$$a_{ij} \leq 0 \quad \textit{for all } i \in \{1, \dots, m\},\ j \in \{1, \dots, n\},\ i \neq j, \tag{1.35}$$

$$\sum_{j=1}^{n} a_{ij} \geq 0 \quad \textit{for all } i \in \{1, \dots, m\}. \tag{1.36}$$

Note that the concept of a matrix of nonnegative type should not be confused with that of a nonnegative matrix in the classical sense (see, e.g., [116, Chapter 2]).

Remark 1.6.15. *In certain contexts, such as when* $\mu = 0$ *in equation* (1.28)*, the matrix* A *may satisfy a stronger condition than* (1.36)*, specifically:*

$$\sum_{j=1}^{N} a_{ij} = 0 \quad \textit{for all } i \in \{1, \dots, M\}. \tag{1.37}$$

This stronger assumption can be leveraged to obtain more refined statements for discrete maximum principles (DMPs) than those derived from condition (1.36).

Theorem 1.6.16 (Local DMP for Nonnegative Type Matrices)**.** *Assume $a_{ii} > 0$ for all $i = 1, \dots, M$. Then, every solution u to the system* (1.32)–(1.33) *satisfies:*

$$\begin{aligned} f_i \le 0 \quad &\Rightarrow \quad u_i \le \max_{j \ne i, a_{ij} \ne 0} u_j^+, \\ f_i \ge 0 \quad &\Rightarrow \quad u_i \ge \min_{j \ne i, a_{ij} \ne 0} u_j^-. \end{aligned} \tag{1.38}$$

for all $i = 1, \dots, M$ if and only if the matrix A is of nonnegative type.

Moreover, if the stronger condition (1.37) *is satisfied, then the implications become*

$$\begin{aligned} f_i \le 0 \quad &\Rightarrow \quad u_i \le \max_{j \ne i, a_{ij} \ne 0} u_j, \\ f_i \ge 0 \quad &\Rightarrow \quad u_i \ge \min_{j \ne i, a_{ij} \ne 0} u_j. \end{aligned} \tag{1.39}$$

Now, we breifely here describe Algebraic Flux Correction methods

1.6.3 Algebraic Flux Correction methods

The diffusion matrix $\mathbf{A}_d$, the convection matrix $\mathbf{A}_c$, and the reaction matrix $\mathbf{M}_c$, which is also called consistent mass matrix, are defined by

$$\mathbf{A}_d = (\ell_{ij})_{i,j=1}^N \quad \text{where } \ell_{ij} = (\nabla\phi_j, \nabla\phi_i) \quad \text{for } i, j = 1, \dots, N, \tag{1.40}$$

$$\mathbf{A}_c = (c_{ij})_{i,j=1}^N \quad \text{where } c_{ij} = (\mathbf{b} \cdot \nabla\phi_j, \phi_i) \quad \text{for } i, j = 1, \dots, N, \tag{1.41}$$

$$\mathbf{M}_c = (m_{ij})_{i,j=1}^N \quad \text{where } m_{ij} = (\phi_j, \phi_i) \quad \text{for } i, j = 1, \dots, N. \tag{1.42}$$

The entries of the matrices can be written as a sum of local entries, e.g.,

$$\ell_{ij} = \sum_{K \subset \omega_i \cap \omega_j} \ell_{ij}^K \quad \text{with } \ell_{ij}^K = (\nabla\phi_j, \nabla\phi_i)_K,$$

and analogously for c_{ij} and m_{ij}.

Algebraic Flux Correction (AFC) methods are a class of algebraically stabilised schemes that have seen significant development in recent years; see, for example, [8, 15, 70, 86, 90, 91, 94, 95, 95, 96, 100]. AFC

stabilisation is not derived from a variational framework. Instead, the methodology begins directly from the algebraic system of equations resulting from the Galerkin finite element discretisation. A nonlinear algebraic correction term is then added to this linear system to enforce a discrete maximum principle (DMP), while avoiding undue numerical diffusion and layer smearing.

Let $\mathbb{A}_{\mathrm{N}}$ denote the matrix associated with the standard Galerkin finite element method applied to system (1.13) with Neumann boundary conditions. This matrix can be expressed as:

$$\mathbb{A}_{\mathrm{N}} = \varepsilon \mathbb{A}_{\mathrm{d}} + \mathbb{A}_{\mathrm{c}} + \mu \mathbb{M}_{\mathrm{c}}, \tag{1.43}$$

where $\mathbb{A}_{\mathrm{d}}$, $\mathbb{A}_{\mathrm{c}}$, and $\mathbb{M}_{\mathrm{c}}$ represent the diffusion, convection, and mass matrices, respectively.

The discrete problem is then reformulated as in systems (1.32) and (1.33), where $f_i = (f, \phi_i)_\Omega$ for $i = 1, \dots, M$ and $g_{i-M} = g(x_i)$ for $i = M+1, \dots, N$. To formulate an AFC scheme, we first introduce a symmetric artificial diffusion matrix $\mathbb{D} = (d_{ij})_{i,j=1}^{N}$, defined by

$$d_{ij} = -\max\{0, a_{ij}, a_{ji}\} \quad \text{for } i \neq j, \qquad d_{ii} = -\sum_{\substack{j=1 \\ j \neq i}}^{N} d_{ij}. \tag{1.44}$$

Hence, the matrix $\mathbb{D}$ has zero row and column sums, and the matrix $\mathbb{A}_{\mathrm{N}} + \mathbb{D}$ is of non-negative type. Replacing $\mathbb{A}_{\mathrm{N}}$ with $\mathbb{A}_{\mathrm{N}} + \mathbb{D}$ in system (1.32), we obtain the stabilised formulation

$$(\mathbb{A}_{\mathrm{N}} + \mathbb{D})^M \mathbf{u} = \mathbf{f}, \tag{1.45}$$

which satisfies the discrete maximum principle (DMP), where $\mathbf{f} = (f_1, \dots, f_M)^\top$.

The standard derivation of the AFC method (see, e.g., [92]) begins by adding the diffusive term ($\mathbb{D}\mathbf{u}$) to both sides of the linear system (1.32), yielding

$$(\mathbb{A}_{\mathrm{N}} + \mathbb{D})^M \mathbf{u} = \mathbf{f} + \mathbb{D}^M \mathbf{u}, \tag{1.46}$$

and then using the identity

$$(\mathbb{D}\mathbf{u})_i = \sum_{j=1}^{N} f_{ij}, \qquad \text{with} \quad f_{ij} = d_{ij}(u_j - u_i). \tag{1.47}$$

The quantities f_{ij} are referred to as fluxes, as they quantify the strength of diffusion between the nodes x_i

and x_j. These fluxes represent the intensity of information exchange and are central to understanding AFC mechanisms.

To reduce spurious oscillations in the discrete solution, the flux terms f_{ij} on the right-hand side of (1.46) are often scaled by solution-dependent correction factors. These damping factors, or *limiters*, denoted by $\alpha_{ij} \in [0, 1]$, suppress unwanted oscillations and yield the nonlinear algebraic formulation

$$\sum_{j=1}^{N} a_{ij}u_j + \sum_{j=1}^{N}(1 - \alpha_{ij}(\mathbf{u}))d_{ij}(u_j - u_i) = f_i, \qquad \text{for } i = 1, \dots, M, \tag{1.48}$$

$$u_i = g_i = g(x_i), \qquad \text{for } i = M + 1, \dots, N. \tag{1.49}$$

It is assumed that the flux limiter satisfies the symmetry condition:

$$\alpha_{ij} = \alpha_{ji}, \qquad i, j = 1, \dots, N. \tag{1.50}$$

Moreover, for any $i, j \in \{1, \dots, N\}$, the limiter function $\alpha_{ij}(\mathbf{u})(u_j - u_i)$ is assumed to be a continuous function of the discrete solution vector $\mathbf{u} \in \mathbb{R}^N$.

A theoretical analysis of the AFC scheme given by equations (1.48) and (1.49), including results on solvability, satisfaction of the local discrete maximum principle (DMP), and error estimates, can be found in [14]; see also [2, 80] for related a posteriori error analysis techniques.

The symmetry condition (1.50) plays a crucial role in the design of AFC methods for several reasons. First, it ensures that the resulting numerical scheme is conservative. Second, it implies that the matrix associated with the AFC-related correction term is positive semi-definite. This structural property significantly contributes to the overall stability of the scheme and facilitates a reliable estimation of the approximation error; see [14] for a detailed discussion.

Finally, it has been demonstrated in [13] that without enforcing the symmetry condition (1.50), the nonlinear algebraic system defined by (1.48) and (1.49) may not admit a solution in general.

1.6.4 Introduction and literature review on the discrete maximum principle (DMP)

Structure-preserving numerical methods is an important topic in computational solutions of partial differential equations. By structure-preserving, we mean methods that approximate solutions which preserve the properties of the exact solution such as local conservation, entropy inequalities, maximum principles, positivity preservation, divergence-free constraints, or exactly symmetric stress tensor approximations, to name

a few.

Numerical methods satisfying Discrete Maximum Principles (DMP) and/or monotonicity properties have been extensively studied in the finite element literature. Methods satisfying these properties often imply positivity preservation or, more generally, bound preservation of the resulting numerical solutions.

To guarantee physically consistent discretisation of equations (1.2) and (1.8), it is crucial to satisfy the discrete counterpart of the maximum principle. Discretisations that fail to meet this criterion often lead to numerical solutions exhibiting unphysical values, referred to as *spurious oscillations*. Equations of the form (1.2) and (1.8) frequently arise in coupled problems, where their numerical solutions serve as input data for additional equations. If these input values contain spurious oscillations, there is a high likelihood that the solutions to subsequent equations will also be affected, potentially leading to unphysical results. In extreme cases, numerical simulations of coupled problems may become unstable and fail, as reported in [82]. Therefore, ensuring the DMP is satisfied is essential for obtaining meaningful numerical solutions for (1.2) and (1.8) in practical applications. When this property is upheld, additional factors such as computational efficiency, the preservation of other physical properties (e.g., conservation laws), and accuracy in terms of relevant quantities, such as norms in Sobolev spaces, become key considerations for selecting an appropriate numerical method. The first proof of a maximum principle (MP) in the context of discretised partial differential equations was introduced by Gershgorin [64] in 1930. In this work, finite difference methods were considered. and Gershgorin proved that the discrete operator satisfies a maximum principle provided that the corresponding coefficient matrix is an M-matrix. A broader generalisation of this result was later provided in the monograph by Collatz in [52]. Further studies on discrete analogues of maximum principles can be found in works by Bramble and Hubbard [24, 25].

In 1970, Ciarlet [47] established the necessary and sufficient conditions for a discretisation to satisfy the discrete maximum principle (DMP). These early studies primarily focused on finite difference methods. However, the linear algebraic arguments employed in these works can be directly extended to the linear systems of equations arising from other discretisation techniques. The first explicit investigation of the DMP in the context of finite element methods was conducted by Ciarlet and Raviart in 1973 [49]. Since then, a substantial body of research has emerged, analysing the DMP for various discretisations of elliptic and parabolic boundary value problems.

A straightforward approach is to introduce sufficient numerical diffusion, ensuring that the problem becomes diffusion-dominated, thereby enforcing the DMP under suitable conditions (see, e.g. Chapter 4 and 8 of [83] respectively for elliptic and parabolic equations). However, this approach tends to introduce excessive diffusion, resulting in solutions that are highly smeared and has limited practical use. This drawback has

motivated the development of shock-capturing methods, which add a term into the numerical scheme. This term includes a viscosity coefficient that depends on the computed solution, introducing nonlinearity into the method (see [81]). The earliest known method satisfying DMP was proposed in [104], with subsequent contributions found in [8–10, 29, 31, 58].

A common feature of these methods is their reliance on first-order polynomial approximations and specific mesh constraints. In two-dimensional problems, for instance, the mesh is often assumed to be of Delaunay. This requirement dates back to early studies on the DMP, including for the Laplace equation (see [49]). Since then, several attempts have been made to relax this assumption. One such approach, presented in [30], modifies the formulation by introducing an anisotropic Laplacian, allowing the DMP to hold under more general conditions. More recently, methods proposed in [69, 70] in the context of hyperbolic equations have succeeded in reducing this constraint while ensuring convergence to the entropy solution. However, extending such ideas to diffusion-dominated problems remains a significant challenge.

As mentioned it is often necessary to impose strict conditions on the mesh, particularly when using the Galerkin method. For a detailed discussion, we refer to [16, 118], where it is proven that for the Poisson equation, the local discrete maximum principle (DMP) holds if and only if the mesh satisfies the XZ-type condition (see Definition 1.6.6). Furthermore, they have shown that satisfying the XZ-criterion ensures the validity of the global DMP.

As explained in section (1.6.3) one class of methods which have been designed to satisfy the DMP by construction is known as algebraic flux correction (AFC). The origins of AFC trace back to [69, 121], and substantial development has been made in recent years, particularly through the contributions of D. Kuzmin and colleagues (see [92–95], and the book by Kuzmin and Hajduk [89]).

The principle of algebraic flux correction (see [89,90,92]) provides a new interpretation of classical high-resolution schemes and establishes a general framework for designing multidimensional flux limiters. In this method, artificial diffusion is both added and removed at the discrete level. Given a discrete operator resulting from a linear or "multilinear" Galerkin approximation (multilinear Galerkin approximation is a Galerkin finite element method in which the discrete solution space is spanned by multilinear basis functions typically tensor products of linear polynomials in each coordinate direction), the nonoscillatory low-order part is extracted. The remainder consists of an antidiffusive correction that admits a conservative flux decomposition [89]. The discrete maximum principle holds if the antidiffusive part satisfies the *local extremum diminishing* (LED) constraint [77, 78], which is enforced by adjusting the magnitudes of the antidiffusive fluxes when necessary.

A numerical method is said to be *linearity-preserving* if it exactly reproduces linear functions over the

computational domain. This means that when the exact solution is a linear function, the numerical approximation must match it exactly. No artificial diffusion, stabilisation, or discretisation errors should alter the linearity of the solution. In other words, the constrained approximation must reduce to the underlying Galerkin scheme when the solution is linear.

In [15] introduced limiters in an algebraic flux correction (AFC) method for a convection-diffusion-reaction equation. In this work using these limiters they guarantee the satisfaction of the DMP and preserving linearity, particularly on general simplicial meshes. In fact they used limiters which satisfy and then modify the algorithm proposed in [95] to ensure that these properties hold for general meshes.

It is important to note that existing AFC methods in the literature often fail to be preserving linearity. For instance, techniques derived from [95] are proven to be linearity-preserving exclusively on symmetric meshes. The method introduced in [32] preserves the DMP solely for meshes that satisfy the condition proposed by Xu and Zikatanov [118], a criterion that remains sharp in the diffusion-dominated regime. However, its linearity preservation is also restricted to symmetric meshes. An alternative approach to ensure both monotonicity and linearity preservation for more general meshes involves solving an optimisation problem for each interior node of the mesh, thereby increasing the complexity of the method. A very recent development in this area has been proposed in [8] where the authors developed monotonicity-preserving stabilised finite element methods for transport problems by combining symmetric projection-based linear stabilisation with nonlinear shock-capturing techniques.

Bound-preservation is a weaker structural requirement than DMP. It plays a crucial role by ensuring the solution remains physically meaningful and numerically stable. In other words, the method may not generally enforce the DMP, but the guarantee of staying within prescribed bounds is already a significant structural property for many PDE models. Bound-preserving numerical solutions are crucial for the numerical stability of many schemes and, in particular, for PDEs where positivity of the solution is important. Many PDE models rely on the positivity of the solution, such as nonlinear reaction-diffusion systems modeling concentrations of reactants or turbulence-inducing fields. Additionally, phase-field PDE models often have solutions satisfying global maxima and minima. While exceeding these bounds is not typically catastrophic for scalar PDE problems, bound violations can have compounding effects in more complex coupled systems.

In addition, the *cut-off* finite element method [101] truncates the finite element function *after* it is computed at a given time step, so as to input the truncated function as approximation of the current time step; see also [120] for an application of a related idea to the Allen-Cahn equation. In the steady state case, the idea of *truncating* the finite element solution to respect given bounds has been justified for reaction-diffusion equations in [88] using energy arguments. In [99] a conservative recovery strategy is proposed and tested

numerically. Finally, in the context of the Joule heating problem a truncation of one of the variables is introduced in order to regularise a rough right-hand side in [79].

In the next section, we present the bound-preserving finite element method, originally proposed in [12].

1.6.5 The bound-preserving finite element methods

In this section, we introduce the bound-preserving finite element method, initially proposed in [12] for reaction-diffusion equations. This bound-preserving finite element method assumes a Lagrange finite element space of arbitrary polynomial degree. To simplify the formulation, the method ensures that the nodal values of the projected solution remain positive, thereby maintaining bound-preservation throughout the computation.

It is important to note that similar methods are different in key aspects from the method which has been proposed in [12]. In [22], imposing the boundary on the continuous solution are incorporated into an optimisation framework, linking this approach to nonlinear stabilisation. Also, [105] formulates a constrained optimisation problem based on a mixed weak formulation to enforce bound preservation. In [33], the relationship between positivity preservation and contact problems was used to develop a nonlinear stabilised method that enforces positivity-preservation in a weak way.

The model problem

Let Ω be an open bounded Lipschitz domain (see Definition 1.4.1) in $\mathbb{R}^d$ ($d = 2, 3$) with a polyhedral boundary $\partial\Omega$. Given a source function $f \in H^{-1}(\Omega)$, the following reaction-diffusion problem has been considered in [12] as a special case of the elliptic equation (1.2)

$$\begin{cases} -\mathrm{div}(\mathcal{D}\nabla u) + \mu u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{cases} \tag{1.51}$$

where $\mathcal{D} = (d_{ij})_{i,j=1}^{d} \in L^\infty(\Omega)^{d\times d}$ is the diffusion tensor and $\mu \in L^\infty(\Omega)$ denotes the reaction coefficient, respectively. We make the following assumptions: $\mu \geq \mu_0$ *a.e.*in Ω and the diffusion tensor $\mathcal{D}$ satisfies in (1.3) (i.e. $\mathcal{D}$ is strictly elliptic), where $[\mathcal{D}]_{ij} = d_{ij}$.
The weak formulation of (1.51) reads: find $u \in H_0^1(\Omega)$ such that

$$a(u, v) = f(v) \quad \forall v \in H_0^1(\Omega), \tag{1.52}$$

where $a(\cdot,\cdot) : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ is the bilinear form defined by

$$a(v, w) := (\mathcal{D}\nabla v, \nabla w)_\Omega + (\mu v, w)_\Omega. \tag{1.53}$$

The bilinear form $a(\cdot,\cdot)$ induces the energy norm:

$$\|v\|_a = \sqrt{a(v, v)}. \tag{1.54}$$

The well-posedness of (1.52) follows from the Lax-Milgram Lemma 1.2.3.

The core of the bound preserving method which has been proposed in [12] is the following property of the solution of (1.51). As a consequence of maximum principle Theorem 1.5.4 and comparison principles Corollary 1.5.3, the following bounds can be proven: for almost all $\boldsymbol{x} \in \Omega$, the solution u of (1.52) satisfies

$$-\frac{\|f\|_{L^\infty(\Omega)}}{\mu_0} \le u(\boldsymbol{x}) \le \frac{\|f\|_{L^\infty(\Omega)}}{\mu_0}. \tag{1.55}$$

This statement becomes more precise if $f \ge 0$ in Ω. In fact, in this case, for almost every $\boldsymbol{x} \in \Omega$, the following inequality holds:

$$0 \le u(\boldsymbol{x}) \le \frac{\|f\|_{L^\infty(\Omega)}}{\mu_0}. \tag{1.56}$$

The results given in (1.55) and (1.56) motivate the introduction of the following assumption.

Assumption (A1): We will suppose that the solution of (1.51) satisfies

$$0 \le u(\boldsymbol{x}) \le \kappa \quad \text{for almost all } \boldsymbol{x} \in \Omega, \tag{1.57}$$

where κ is a known constant. However, the lower bound in this assumption is not required to be equal to zero, and all results below hold even for the case the value κ can be replaced by a non-negative continuous function $\kappa(\boldsymbol{x})$.

Remark 1.6.17. *The upper (and lower) bounds of the solution are not chosen arbitrarily but follow from the properties of the underlying PDE. In many convection–diffusion or reaction–diffusion problems, the continuous solution satisfies a maximum principle, which implies that the solution cannot exceed certain values determined by the initial or boundary data. Alternatively, the physical interpretation of the unknown (for example, a concentration or probability) provides natural bounds, such as values lying in the interval*

$[0, 1]$. *The purpose of the bound-preserving finite element method is to ensure that the discrete solution respects these bounds.*

The bound-preserving finite element method for reaction-diffusion problem

With Assumption (A1) in mind, the following subset of finite element functions which satisfy the bounds given by (1.57) at the degrees of freedom has been considered in [12]:

$$V_{\mathcal{P}}^{+} := \{v_h \in V_{\mathcal{P}} \,:\, v_h(\boldsymbol{x}_i) \in [0, \kappa] \;\text{ for all }\; i = 1, \dots, N\}, \tag{1.58}$$

where $V_{\mathcal{P}}$ has been defined in (1.15).

Each element $v_h \in V_{\mathcal{P}}$ can be decomposed into the sum $v_h = v_h^+ + v_h^-$, where v_h^+ and v_h^- are defined as (see Figure 1.4)

$$v_h^+ = \sum_{i=1}^{N} \max\left\{0, \min\{v_h(\boldsymbol{x}_i), \kappa\}\right\} \phi_i\,, \tag{1.59}$$

and

$$v_h^- = v_h - v_h^+. \tag{1.60}$$

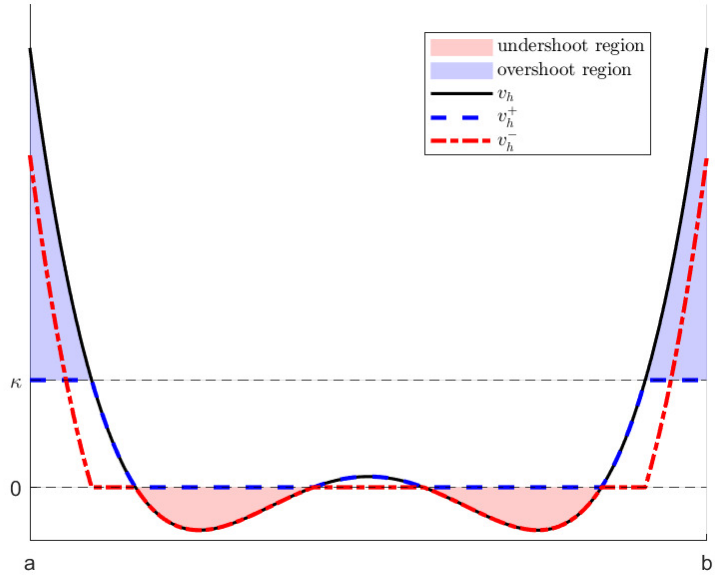


Figure 1.4: Decomposition of v_h into v_h^+ and v_h^- ($\Omega = (a, b)$).

The functions v_h^+ and v_h^- are referred as the *constrained* and *complementary* parts of v_h, respectively. Using this decomposition, the following algebraic projection is defined:

$$(\cdot)^+ : V_{\mathcal{P}} \to V_{\mathcal{P}}^+, \quad v_h \mapsto v_h^+ . \tag{1.61}$$

Remark 1.6.18. *The operator defined in* (1.61) *is a* componentwise (nodal) projection *onto the convex set $V_{\mathcal{P}}^+$, obtained by clamping each nodal value into the interval $[0, \kappa]$. If we identify a function $v_h \in V_{\mathcal{P}}$ with its coefficient vector $(U_1, \dots, U_N)$, where $U_i = v_h(\boldsymbol{x}_i)$, then this operator corresponds to the metric projection of the vector $(U_1, \dots, U_N)$ onto the box $[0, \kappa]^N \subset \mathbb{R}^N$ with respect to the standard Euclidean norm. In particular, for each node,*

$$\pi_{[0,\kappa]}(v_h(\boldsymbol{x}_i)) = \max\{0, \min\{v_h(\boldsymbol{x}_i), \kappa\}\},$$

is the closest point to $v_h(\boldsymbol{x}_i)$ in $[0, \kappa]$. The function v_h^+ is then obtained by expanding these projected coefficients in the finite element basis.

We stress that this projection is defined in the coefficient (nodal) Euclidean norm, and in general it is not an orthogonal projection in $V_{\mathcal{P}}$ *with respect to the usual $L^2(\Omega)$ inner products.*

Remark 1.6.19. *If κ is not a constant value, but instead a non-negative continuous function, the only modi-*

fication to the projection definition is that in such a case, the constrained component is given by

$$v_h^+ = \sum_{i=1}^{N} \max\left\{0, \min\{v_h(\boldsymbol{x}_i), \kappa(\boldsymbol{x}_i)\}\right\} \phi_i \,. \tag{1.62}$$

The bound-preserving finite element method proposed in [12] reads as follows: find $u_h \in V_{\mathcal{P}}$ such that

$$a_h(u_h; v_h) = \langle f, v_h \rangle_\Omega, \qquad \forall v_h \in V_{\mathcal{P}}, \tag{1.63}$$

where the nonlinear form $a_h(\cdot;\cdot)$ is defined by

$$a_h(u_h; v_h) := a(u_h^+, v_h) + s(u_h^-, v_h). \tag{1.64}$$

Here $a(\cdot,\cdot)$ defined in (1.53) and $s(\cdot,\cdot) : C(\bar{\Omega}) \times C(\bar{\Omega}) \to \mathbb{R}$ is the stabilising bilinear form defined by

$$s(v_h, w_h) = \alpha \sum_{i=1}^{N} \left(\|\mathcal{D}\|_{0,\infty,\omega_i} \mathfrak{h}(\boldsymbol{x}_i)^{d-2} + \mu \mathfrak{h}(\boldsymbol{x}_i)^d \right) v_h(\boldsymbol{x}_i) w_h(\boldsymbol{x}_i), \tag{1.65}$$

where $\alpha > 0$ is a non-dimensional constant to be determined precisely in Theorem 1.6.23, and the piecewise linear function $\mathfrak{h}(\boldsymbol{x}_i)$ has been defined in (1.20). and ω_i refers to the vertex neighbourhood of the node $\boldsymbol{x}_i$, i.e., $\omega_i := \cup_{K \in \mathcal{P}: K \cap K_{\boldsymbol{x}_i} \neq \emptyset} K$ denotes an extended patch.

Remark 1.6.20. *In finite element method* (1.63) *we use the space $V_{\mathcal{P}}$ defined in* (1.15)*. The basis functions $\{\phi_i\}_{i=1}^N$ are the nodal Lagrange basis functions associated with the set of internal nodes $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$. In the bound-preserving method* (1.63)–(1.64)*, both the projection $v_h \mapsto v_h^{\pm}$ and the stabilisation term $s(\cdot,\cdot)$ act only on these nodal values. Consequently, the nonlinear scheme is entirely characterised at the level of nodal degrees of freedom, making the bound-preserving property enforceable through nodal clipping of the coefficients into the admissible interval* $[0, \kappa]$.

The stabilisation term $s(\cdot,\cdot)$ defines the following norm:

$$\|v_h\|_s := \sqrt{s(v_h, v_h)}. \tag{1.66}$$

Lemma 1.6.21. *There exists a constant $C_{\text{equiv}} > 0$, depending only on the shape-regularity constant, such that*

$$\|v_h\|_a^2 \leq \frac{C_{\text{equiv}}}{\alpha} \|v_h\|_s^2 \quad \forall v_h \in V_{\mathcal{P}}, \tag{1.67}$$

where $\alpha > 0$ is the stabilisation parameter appearing in the definition (1.65) *of $s(\cdot,\cdot)$.*

Proof. By using the inverse inequality (1.26), (1.20), and the regularity of the mesh, we have

$$\|D^{\frac{1}{2}}\nabla v_h\|_{0,\Omega}^2 + \|\mu^{\frac{1}{2}} v_h\|_{0,\Omega}^2 \leq C \sum_{i=1}^{N} \left(\|D\|_{0,\infty,\omega_i} \mathfrak{h}(\boldsymbol{x}_i)^{d-2} + \mu\, \mathfrak{h}(\boldsymbol{x}_i)^d \right) v_h(\boldsymbol{x}_i)^2. \tag{1.68}$$

Therefore bound holds with (1.67) with $C_{\text{equiv}} = C$. □

Note that we use a mass-lumped structure in the construction of $s(\cdot,\cdot)$, primarily due to the monotonicity result stated in the following lemma.

Lemma 1.6.22. *[12, Lemma 3.1] The bilinear form $s(\cdot,\cdot)$ defined in* (1.65) *satisfies the following inequalities*

$$s(v_h^- - w_h^-, v_h^+ - w_h^+) \geq 0 \qquad \forall v_h, w_h \in V_{\mathcal{P}}, \tag{1.69}$$

$$s(v_h^-, w_h - v_h^+) \leq 0 \qquad \forall v_h \in V_{\mathcal{P}}, w_h \in V_{\mathcal{P}}^+. \tag{1.70}$$

In the following we state the well-posedness results for the finite element method (1.64) which have been proven in [12].

Theorem 1.6.23. *(Well-posedness) [12, Theorem3.2] Let $T : V_{\mathcal{P}} \to V_{\mathcal{P}}$ be the mapping defined by*

$$[T(v_h), w_h] = a(v_h^+, w_h) + s(v_h^-, w_h), \tag{1.71}$$

for all $v_h, w_h \in V_{\mathcal{P}}$. If the non-dimensional parameter α is chosen such that $\alpha \geq C_{equiv}$, then T is continuous and strongly monotone. Specifically, T satisfies the following monotonicity condition: there exists a constant $\beta > 0$, independent of h, such that

$$[T(v_h) - T(w_h), v_h - w_h] \geq \beta \|v_h - w_h\|_a^2, \tag{1.72}$$

for all $v_h, w_h \in V_{\mathcal{P}}$. Consequently, (1.64) *has a unique solution $u_h \in V_{\mathcal{P}}$.*

The following result shows that the stabilised methods (1.40) is consistent with the original problem.

Lemma 1.6.24. *(Consistency) [12, Lemma 3.3] Under Assumption (A1), the method* (1.63) *enjoys the following invariance property: if the exact solution belongs to $V_{\mathcal{P}}$, then $u_h^+ = u_h = u$.*

Theorem 1.6.25. *(Well-posedness) [12, Theorem 3.5] Under Assumption (A1), the stabilised method* (1.63) *admits a unique solution.*

One of the primary features of the method (1.63) is that u_h^+ is characterised as the unique solution of a variational inequality posed on the convex set $V_{\mathcal{P}}^+$. This is formalised in the following theorem.

Theorem 1.6.26. *(Characterisation of the constrained part) [12, Theorem 3.5] Let $u_h^+ \in V_{\mathcal{P}}^+$ be the unique solution of* (1.63)*. Then, u_h^+ satisfies the following variational inequality:*

$$a(u_h^+, v_h^+ - u_h^+) + s(u_h^+, v_h^+ - u_h^+) \geq (f, v_h^+ - u_h^+), \quad \forall v_h^+ \in V_{\mathcal{P}}^+. \tag{1.73}$$

Remark 1.6.27. *By Stampacchia's Theorem, equation* (1.73) *can be directly used to prove the uniqueness of the solution of* (1.63)*. This shows that the stabilised method* (1.63) *and the variational inequality* (1.73) *are, in fact, equivalent.*

The formulation of this method as a variational inequality allows us to establish optimal approximation error estimates using a standard approach.

Theorem 1.6.28 (Abstract error analysis)**.** *Let u be the solution of* (1.51) *and let $u_h \in V_{\mathcal{P}}$ be the unique solution of* (1.63)*. Then, we have*

$$\|u - u_h^+\|_a = \min_{v_h \in V_{\mathcal{P}}^+} \|u - v_h\|_a. \tag{1.74}$$

Moreover, let u_h^{FEM} be the solution of the standard finite element Galerkin method: find $u_h^{\mathrm{FEM}} \in V_{\mathcal{P}}$ such that

$$a(u_h^{\mathrm{FEM}}; v_h) = \langle f, v_h \rangle_\Omega, \qquad \forall v_h \in V_{\mathcal{P}},$$

then, the negative part u_h^- satisfies the following error estimate:

$$s(u_h^-, u_h^-)^{\frac{1}{2}} \leq \sqrt{\frac{C_{equiv}}{\alpha}} \min \left\{ \|u - u_h^+\|_a, \|u_h^{FEM} - u_h^+\|_a \right\}. \tag{1.75}$$

Proof. See the proof of [12, Theorem 3.7]. □

Remark 1.6.29. *These results can be interpret in two ways. First, the complementary part u_h^- implies convergence to zero that is at least as rapid as the convergence of u_h^+ to the exact solution u. Furthermore, the bound* (1.75) *implies that under certain conditions, this convergence rate can substantially exceed that of*

u_h^+. *For piecewise linear finite element approximations, when the mesh satisfies the conditions for the plain Galerkin method (as detailed in [26]), we obtain* $u_h^{\mathrm{FEM}} \in V_{\mathcal{P}}^+$. *This condition ensures that* $u_h = u_h^{\mathrm{FEM}}$ *coincides with the solution of the Galerkin method. Thus, thanks to* (1.75)*, for certain meshes, and their regular refinements, we have that* $u_h^- = 0$.

The best approximation result (1.74) *shows that no better finite element function satisfying with bound-preserving nodal values in the energy norm. Thus,* (1.74) *is the counterpart of the classical best approximation result for the solution* u_h^{FEM} *of the Galerkin method*

$$\|u - u_h^{\mathrm{FEM}}\|_a = \inf_{v_h \in V_{\mathcal{P}}} \|u - v_h\|_a. \tag{1.76}$$

Implementation

To implement the finite element method (1.63), we use the following Richardson-like iterative approximation: Given u^0 and $\omega \in (0, 1]$, for each $n = 0, 1, \ldots$ find u^{n+1} such that

$$a(u_h^{n+1}, v_h) = a(u_h^n, v_h) + \omega \left(\langle f, v_h \rangle_\Omega - a((u_h^n)^+, v_h) - s((u_h^n)^-, v_h) \right) \qquad \forall v_h \in V_{\mathcal{P}}. \tag{1.77}$$

We initialise the finite element approximation of (1.63) by the Galerkin approximation, that is we set $u_h^0 \in V_{\mathcal{P}}$ such that, for all $v_h \in V_{\mathcal{P}}$

$$a(u_h^0, v_h) = \langle f, v_h \rangle_\Omega. \tag{1.78}$$

In the experiments we used $\alpha = 1$ within the stabilisation. The linear systems arising in (1.77) are solved using an LU decomposition within the Eigen library. The linearisation was terminated when $\|u_h^{n+1} - u_h^n\|_{0,\Omega} \leq 10^{-12}$.

Remark 1.6.30. *(Discrete linear system) Let* $\{\phi_j\}_{j=1}^N$ *be the nodal basis of* $V_{\mathcal{P}}$*, and write* $u_h^n = \sum_{i=1}^N U_i^n \phi_i$ *with coefficient vector* $U^n = (U_i^n)_{i=1}^N \in \mathbb{R}^N$. *Defining the load vector and matrices*

$$b_j := \langle f, \phi_j \rangle_\Omega, \qquad (A)_{j,i} := a(\phi_i, \phi_j), \qquad (S)_{j,i} := s(\phi_i, \phi_j),$$

the iteration (2.48) *yields the linear system*

$$A\,U^{n+1} = A\,U^n + \omega\big(b - A\,U^{n,+} - S\,U^{n,-}\big), \tag{1.79}$$

where the vectors $U^{n,+}, U^{n,-} \in \mathbb{R}^N$ *correspond to the constrained and complementary parts of* U^n*, are*

respectively

$$(U^{n,+})_i = \max\{0, \min\{U_i^n, \kappa\}\}, \qquad (U^{n,-})_i = U_i^n - (U^{n,+})_i.$$

Thus each iteration consists of forming the right-hand side in (1.77) *and solving a linear system with the fixed matrix A. Note that S is diagonal and cheap to apply, while A is sparse and typically symmetric positive definite.*

When Dirichlet boundary conditions are imposed, the iteration (1.79) *is carried out only on the* free nodes*, i.e. those not fixed by the boundary data. In practice this means that the system is assembled in the full space, boundary values are prescribed directly on the corresponding entries of U^n, and the reduced linear system for the free degrees of freedom is solved at each iteration. Thus, the iteration updates only the unknown interior (or Neumann) nodes, while the boundary nodes remain fixed throughout.*

The choice of the damping parameter $\omega \in (0, 1]$ affects convergence. From a fixed-point perspective, the iteration can be written

$$U^{n+1} = U^n + \omega A^{-1}\big(b - AU^{n,+} - SU^{n,-}\big).$$

Using Clarke's generalised derivative (see [108, 114]) for the nonsmooth maps $U \mapsto U^+$ and $U \mapsto U^-$, leads to the Jacobian

$$J \approx I - \omega\big(D^+ + A^{-1}SD^-\big),$$

where D^+ is a diagonal selection matrix encoding the active set and $D^- = I - D^+$. For a discrete vector $U = (U_i)_{i=1}^N \in \mathbb{R}^N$ corresponding to the nodal values of u_h, the active set *is defined as the collection of nodes where the bound constraints are active, i.e.,*

$$\mathcal{A}(U) := \Big\{\, i \in \{1, \dots, N\} \;:\; U_i < 0 \;\text{ or }\; U_i > \kappa \,\Big\}.$$

The complement,

$$\mathcal{F}(U) := \{1, \dots, N\} \setminus \mathcal{A}(U),$$

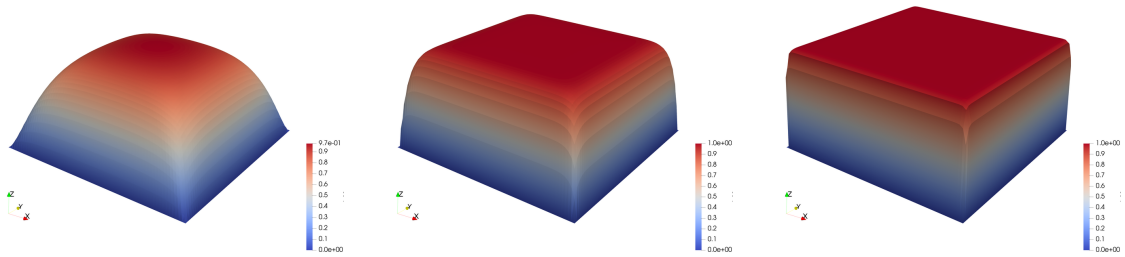
is called the free set*, corresponding to nodes for which the bound constraints are not active ($0 \le U_i \le \kappa$). At nodes in $\mathcal{A}(U)$ the values are clipped to the admissible interval $[0, \kappa]$, while the iteration proceeds only on the free set.*

Convergence is guaranteed if the spectral radius of J is strictly less than one, which in practice requires ω to be sufficiently small. Since D^+ varies with the iterate, a universal analytic bound for ω is not available. In practice, we choose ω experimentally.

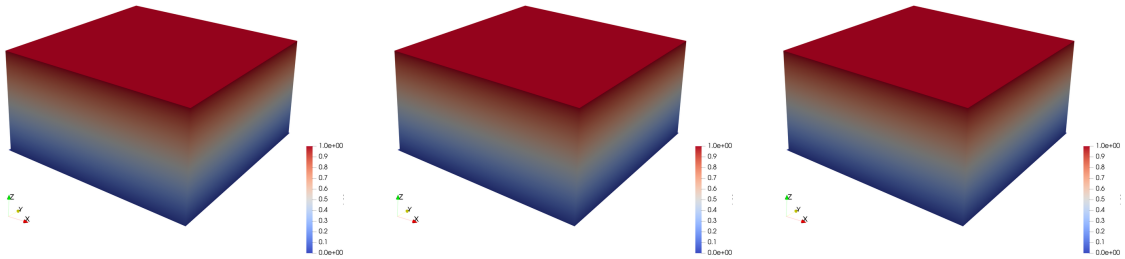
Example 2. *(Resolution of boundary layers) [12, Example 5.3] Consider the problem*

$$\begin{aligned} -\epsilon\Delta u + u &= 1 \text{ in } \Omega\,, \\ u &= 0 \text{ on } \partial\Omega. \end{aligned} \tag{1.80}$$

We fix $h \approx 0.02$ *on a criss-cross and vary* $\epsilon \in [10^{-2}, 10^{-7}]$. *For particularly small* ϵ *the Richardson iteration required dampening for convergence. With* $\epsilon > 10^{-5}$, *we use* $\omega = 1$ *and convergence was achieved within 4 iterations. When* $\epsilon \leq 10^{-5}$, $\omega = \frac{1}{2}$ *is sufficient for convergence with fewer than 46 iterations in each case. The most challenging case being the smallest value of* ϵ. *Computed solutions for different values of* ϵ *are shown in Figure 1.5.*



(a) $\epsilon = 10^{-2}$. (b) $\epsilon = 10^{-3}$. (c) $\epsilon = 10^{-4}$.



(d) $\epsilon = 10^{-5}$. (e) $\epsilon = 10^{-6}$. (f) $\epsilon = 10^{-7}$.

Figure 1.5: Elevations of the approximation to (1.80) for fixed h and different values of ϵ. We notice the absence of oscillations even for particularly small values of ϵ.

In the next chapter, we extend the bound-preserving method proposed in this section to convection–diffusion problems. Analogously to this section, the set of *admissible* finite element functions, denoted $V_{\mathcal{P}}^{+}$, is defined as the set of functions satisfying global bounds at their degrees of freedom (e.g., nodal values for Lagrangian elements). Similar to what we have done in this section, an algebraic projection onto this admissible set is introduced, and the finite element problem is written for the projected object, u_h^{+}. This

projection introduces a kernel, so a stabilising term is added to avoid singularity, allowing the method to avoid the use of Lagrange multipliers. In the case of linear reaction-diffusion equations, the projection u_h^+ is the orthogonal projection onto $V_{\mathcal{P}}^+$ (see Theorem 1.6.28), but this property is lost for more complex problems such as convection-diffusion problems and those with nonlinear reactions, as in [12]. Although the main ideas are similar, the methodology introduces several important differences. Instead of starting with the plain Galerkin method, we start with a stabilised finite element method. The stabilisation form is defined in a way that controls the convective term. Adding the stabilisation term reduces the local oscillations that may still occur. In addition, numerical experiments show that adding linear stabilisation significantly improves the performance of the nonlinear solver.

In Chapter 3, we extend this methodology to time-dependent convection–diffusion equations. At each time step, the approach follows the same ideas as in Chapter 2, adapted to the time-dependent problem.

Finally, in Chapter 4, we extend the bound-preserving finite element method to polytopic meshes using the discontinuous Galerkin method. In these meshes, the degrees of freedom do not depend on the number of vertices, edges, or faces in each element. That is, the degrees of freedom are not tied to physical points, and the basis functions are defined over the whole domain. To enforce the bound-preserving constraints, we use a sub-triangulation approach, applying the constraints at each degree of freedom within the sub-triangulated mesh.

Chapter 2

A nodally bound-preserving finite element method for convection-diffusion equations

2.1 Introduction

The aim of this chapter is to discuss the recent work [4], which extends the bound-preserving method proposed in section (1.6.5) to the convection-diffusion equation. Although the driving principles are similar, this work introduces several significant differences, because it does not start with the plain Galerkin scheme, but instead a stabilised finite element method. Using this stabilised method serves two purposes: first, in regions where the constraint is inactive, local oscillations may still occur, and linear stabilisation helps to reduce this; second, numerical experiments indicate that adding linear stabilisation to u_h^+ significantly improves the performance of the nonlinear solver. The stabilisation form is defined in a way which controls the convective term. In fact the stabilisation penalises jumps of the convective derivative across element interfaces. This suppresses spurious oscillations in convection-dominated regimes and makes the discrete operator more coercive. As a result, the linear solver (like Richardson iteration) or a nonlinear solver (where the Jacobian appears) better conditioned, which improves robustness and convergence of the nonlinear solver. Also, the analysis differs substantially from that of (1.63). The well-posedness analysis is different, as the discretisation is no longer driven by a monotone operator. Additionally, due to the non-symmetric nature of the problem, the solution u_h^+ is no longer the orthogonal projection of u onto $V_{\mathcal{P}}^+$, so the error analysis follows a different path. Much of the material in this chapter is based on Ref. [4].

The remainder of this chapter is organised as follows: In Section 2.2, we introduce the model problem, and the preliminary material for the setup of the method, including the choice of linearly stabilised methods.

Section 2.3 presents the finite element method and proves its well-posedness. The error analysis is conducted in Section 2.4, and in Section 2.5, we evaluate the performance of the method through numerical experiments, comparing it to existing alternatives.

2.2 The Model Problem

As in section (1.6.5), let Ω be an open bounded Lipschitz domain in $\mathbb{R}^d$ ($d = 2, 3$) with a polyhedral boundary $\partial\Omega$. Given a source function $f \in H^{-1}(\Omega)$, we consider the following reaction-convection-diffusion problem as a special case of the elliptic equation (1.2):

$$\begin{cases} -\mathrm{div}(\mathcal{D}\nabla u) + \boldsymbol{\beta} \cdot \nabla u + \mu u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{cases} \tag{2.1}$$

where $\mathcal{D} = (d_{ij})_{i,j=1}^d \in L^\infty(\Omega)^{d\times d}$ is the diffusion tensor, $\boldsymbol{\beta} = (\beta_i)_{i=1}^d \in L^\infty(\Omega)^d$ denotes the convective field, and $\mu \in \mathbb{R}^+$ is the reaction coefficient. Additionally, we assume that $\mathrm{div}\,\boldsymbol{\beta} = 0$ and that $\mathcal{D}$ is symmetric and uniformly strictly positive definite in Ω (in this case positive definiteness and the definition of ellipticity (1.3) are equivalent).

Remark 2.2.1. *We use the assumption* $\mathrm{div}\,\boldsymbol{\beta} = 0$ *to simplify the analysis. The results would remain valid under the standard weaker condition* $\mu - \mathrm{div}\,\boldsymbol{\beta}/2 > 0$. *Furthermore, the analysis also applies when* $\mu \in L^\infty(\Omega)$ *is a strictly positive function in* $\overline{\Omega}$.

The weak formulation of the problem (2.1) is to find $u \in H_0^1(\Omega)$ such that:

$$a(u, v) = \langle f, v\rangle_\Omega \quad \forall v \in H_0^1(\Omega), \tag{2.2}$$

where the bilinear form $a(\cdot,\cdot)$ is given by:

$$a(w, v) := (\mathcal{D}\nabla w, \nabla v)_\Omega + (\boldsymbol{\beta} \cdot \nabla w, v)_\Omega + (\mu w, v)_\Omega \quad \forall v, w \in H_0^1(\Omega). \tag{2.3}$$

This bilinear form induces the following energy norm on $H_0^1(\Omega)$:

$$\|v\|_a = \sqrt{a(v, v)}\,.$$

By the Lax-Milgram Lemma 1.2.3, the weak formulation (2.2) is well-posed.

In line with the discussion in the introduction, our goal is to develop discrete solutions that respects the bounds of the weak solution of (2.2). Thus, similar to Assumption (A1) for the solution of the weak form (2.2), we apply the same assumption to the solution of (2.2), namely, we assume that $0 \leq u(\boldsymbol{x}) \leq \kappa$ for almost every $\boldsymbol{x} \in \Omega$.

This assumption is a re-statement of a consequence of the maximum principle Theorem 1.5.4 for elliptic PDEs. The lower bound in (1.57) for the solution of (2.2) again is not necessarily zero, but is chosen as such for simplicity. The same results hold if κ is replaced by a non-negative continuous function $\kappa(\boldsymbol{x})$. In some cases, sharp bounds for κ can be derived. For example, maximum and comparison principles (see, e.g., Corollary 1.5.3) provide the following bounds: for almost all $\boldsymbol{x} \in \Omega$, the solution u of (2.2) satisfies

$$-\frac{\|f\|_{0,\infty,\Omega}}{\mu} \leq u(\boldsymbol{x}) \leq \frac{\|f\|_{0,\infty,\Omega}}{\mu}. \tag{2.4}$$

Moreover, if $f \geq 0$ in Ω, the lower bound can be sharpened to:

$$0 \leq u(\boldsymbol{x}) \leq \frac{\|f\|_{0,\infty,\Omega}}{\mu}, \tag{2.5}$$

for almost all $\boldsymbol{x} \in \Omega$. Therefore, a reasonable estimate for κ is $\frac{\|f\|_{0,\infty,\Omega}}{\mu}$, which we use in our numerical experiments.

Remark 2.2.2. *While the results that follow can, in theory, be extended to more general quadrilateral meshes, doing so would necessitate additional technical work to establish norm equivalences, which are standard for mapped elements. To maintain the focus on bound-preservation properties, we limit the discussion to affine simplices and quadrilateral/hexahedral meshes.*

2.2.1 The finite element space

To discretise the problem (2.2), we employ the conforming, shape-regular (see the definition 1.6.3), and quasi-uniform partition (see the definition 1.6.4) of the domain Ω into closed simplices or affine quadrilateral/hexahedral elements, denoted by $\mathcal{P}$, which has been defined in (1.15). Furthermore, we use the notations introduced in Section (1.6).

Remark 2.2.3. *The shape-regularity of the mesh is essential, since we need local inverse and trace inequalities in the subsequent proofs. Moreover, as global inverse inequalities are also employed, quasi-uniformity on the mesh is required throughout the analysis.*

2.2.2 The algebraic projection onto the admissible set

With Assumption (A1) in mind, we again use the subset of finite element functions (1.58) which satisfy in bound given by (1.57) at the degrees of freedom. Each element $v_h \in V_{\mathcal{P}}$ can be decomposed into the sum $v_h = v_h^+ + v_h^-$, where v_h^+ and v_h^- are defined in (1.59) and (1.60). We refer to v_h^+ and v_h^- as the *constrained* and *complementary* components of v_h, respectively. So, by this decomposition, we can use the algebraic projection defined in (1.61). Also, similar to the Remark (1.6.19), in the case that κ is not a constant the constrained component (1.58) is modified as(1.62).

The following results regarding this projection (1.61) will be used frequently in our analysis.

Lemma 2.2.4. *(see [4, Lemma 2.1]) Let the operator $(\cdot)^+$ be defined as in* (1.61)*. There exists a constant $C > 0$, independent of h, such that*

$$\|w_h^+ - v_h^+\|_{0,\Omega} \leq C\|w_h - v_h\|_{0,\Omega}, \tag{2.6}$$

$$\|v_h^+\|_{0,\Omega} \leq C\kappa, \tag{2.7}$$

for all $w_h, v_h \in V_{\mathcal{P}}$.

Proof. In this proof, we omit the subscript h to simplify the notation. Let $w, v \in V_{\mathcal{P}}$. First, note that if $w(\boldsymbol{x}_i) \leq v(\boldsymbol{x}_i)$, then $v^+(\boldsymbol{x}_i) - w^+(\boldsymbol{x}_i) \leq v(\boldsymbol{x}_i) - w(\boldsymbol{x}_i)$, and when $v(\boldsymbol{x}_i) \leq w(\boldsymbol{x}_i)$, we have $w^+(\boldsymbol{x}_i) - v^+(\boldsymbol{x}_i) \leq -(v(\boldsymbol{x}_i) - w(\boldsymbol{x}_i))$. Therefore,

$$|v^+(\boldsymbol{x}_i) - w^+(\boldsymbol{x}_i)| \leq |v(\boldsymbol{x}_i) - w(\boldsymbol{x}_i)|.$$

Using (1.21), we then obtain

$$\begin{aligned}
\|v^+ - w^+\|_{0,\Omega}^2 &\leq C\,|v^+ - w^+|_h^2 \\
&= C\sum_{i=1}^{N} \mathfrak{h}(\boldsymbol{x}_i)^d\,|v^+(\boldsymbol{x}_i) - w^+(\boldsymbol{x}_i)|^2 \\
&\leq C\sum_{i=1}^{N} \mathfrak{h}(\boldsymbol{x}_i)^d\,|v(\boldsymbol{x}_i) - w(\boldsymbol{x}_i)|^2 \\
&\leq C\|v - w\|_{0,\Omega}^2\,,
\end{aligned}$$

which concludes the proof. □

2.2.3 A Linear Stabilisation Method

In this section, we incorporate a linear stabilisation term to reduce the oscillations caused by dominant convection. The stabilisation technique used here is based on the Continuous Interior Penalty (CIP) method, originally proposed in [35], and involves augmenting the Galerkin scheme with the following stabilising term penalty term

$$J(u_h, v_h) = \gamma \sum_{F \in \mathcal{F}_I} \int_F \|\boldsymbol{\beta}\|_{0,\infty,F} h_F^2 [\![\nabla u_h]\!] \cdot [\![\nabla v_h]\!] \, \mathrm{d}s \,, \tag{2.8}$$

where $\gamma \geq 0$ is a non-dimensional constant, or alternatively we can use the upwind gradient jumps rather than the normal gradient (2.8), given by

$$J(u_h, v_h) = \gamma \sum_{F \in \mathcal{F}_I} \int_F \frac{\gamma_\beta}{\|\boldsymbol{\beta}\|_{0,\infty,F}} h_F^2 [\![\boldsymbol{\beta} \cdot \nabla u_h]\!] [\![\boldsymbol{\beta} \cdot \nabla v_h]\!] \, \mathrm{d}s \,, \tag{2.9}$$

where $\gamma_\beta \geq 0$ is a non-dimensional constant.

Remark 2.2.5. *We use the CIP stabilisation to control oscillations that appear in convection-dominated problems. The CIP term penalises jumps of the gradient across element interfaces, which reduces spurious oscillations while keeping the method consistent. It also allows us to work with continuous finite element spaces, avoiding the extra cost of discontinuous Galerkin methods.*

Other stabilisation techniques (such as SUPG, local projection, or DG-type stabilisations) could also be applied. However, CIP is chosen here because it is simple to implement, works naturally with continuous elements, and provides stability without increasing the number of degrees of freedom.

Remark 2.2.6. *A sufficiently large value of γ in* (2.8)–(2.9) *enhances the control of interelement gradient jumps and thereby improves stability in convection-dominated regimes. However, excessively large values may lead to over-diffusion, smearing layers and reducing the accuracy of the approximation. On the other hand, if γ is too small, the stabilisation becomes ineffective and spurious oscillations may persist in the numerical solution.*

In practice, there is no general rule for choosing of γ, and its optimal value may depend on the mesh, the problem parameters, and the presence of sharp layers. We select γ experimentally in our numerical tests. In particular, we observe that moderate values of γ are sufficient to suppress oscillations while still resolving sharp layers accurately, whereas too small or too large values deteriorate the quality of the solution.

So by either (2.8) or (2.9), the CIP stabilisation method can then be formulated as follows: find $u_h \in V_{\mathcal{P}}$ such that

$$a_J(u_h, v_h) := a(u_h, v_h) + J(u_h, v_h) = \langle f, v_h \rangle_\Omega \qquad \forall v_h \in V_{\mathcal{P}}\,. \tag{2.10}$$

The bilinear form $a_J(\cdot,\cdot)$ induces the following norm on $V_{\mathcal{P}}$:

$$\|v_h\|_h := a_J(v_h, v_h)^{\frac{1}{2}} = \left(\|\mathcal{D}^{\frac{1}{2}} \nabla v_h\|_{0,\Omega}^2 + \|\mu^{\frac{1}{2}} v_h\|_{0,\Omega}^2 + J(v_h, v_h) \right)^{\frac{1}{2}}. \tag{2.11}$$

The following lemma will be instrumental in the error analysis.

Lemma 2.2.7. *(see [4, Lemma 2.2]) There exists a constant $C > 0$ independent of h and other physical parameters such that for any $v_h \in V_{\mathcal{P}}$, the penalty term* (2.8) *satisfies the following upper bound*

$$\frac{\gamma}{\|\boldsymbol{\beta}\|_{0,\infty,\Omega}} \left\| h^{\frac{1}{2}} \left(\boldsymbol{\beta} \cdot \nabla v_h - \pi(\boldsymbol{\beta} \cdot \nabla v_h) \right) \right\|_{0,\Omega}^2 \leq C J(v_h, v_h)\,. \tag{2.12}$$

Moreover, above bound for all $v_h, w_h \in V_{\mathcal{P}}$ satisfies the following bounds:

$$J(v_h, w_h) \leq C\gamma h \|\boldsymbol{\beta}\|_{0,\infty,\Omega} |v_h|_{1,\Omega} |w_h|_{1,\Omega}, \tag{2.13}$$

$$J(v_h, w_h) \leq C\gamma \left(\sum_{K \in \mathcal{P}} h_K^{-1} \|\boldsymbol{\beta}\|_{0,\infty,K} \|v_h\|_{0,K}^2 \right)^{\frac{1}{2}} \left(\sum_{K \in \mathcal{P}} h_K^{-1} \|\boldsymbol{\beta}\|_{0,\infty,K} \|w_h\|_{0,K}^2 \right)^{\frac{1}{2}}, \tag{2.14}$$

where the constant $C > 0$ in these bounds, independent of h and other physical parameters.

Proof. The inequality (2.12) follows directly from Lemma 5 in [35]. To prove (2.13), we apply the Cauchy-Schwarz inequality, the local trace inequality (1.27)

$$\begin{aligned}
J(v_h, w_h) &= \sum_{F \in \mathcal{F}_I} \int_F \gamma \|\boldsymbol{\beta}\|_{0,\infty,F} h_F^2 [\![\nabla v_h]\!] \cdot [\![\nabla w_h]\!] \, \mathrm{d}s \\
&\leq \left(\sum_{F \in \mathcal{F}_I} \gamma \|\boldsymbol{\beta}\|_{0,\infty,F} h_F^2 \|[\![\nabla v_h]\!]\|_{0,F}^2 \right)^{\frac{1}{2}} \left(\sum_{F \in \mathcal{F}_I} \gamma \|\boldsymbol{\beta}\|_{0,\infty,F} h_F^2 \|[\![\nabla w_h]\!]\|_{0,F}^2 \right)^{\frac{1}{2}} \\
&\leq C\gamma \left(\sum_{K \in \mathcal{P}} \|\boldsymbol{\beta}\|_{0,\infty,K} h_K^2 \|\nabla v_h|_K\|_{0,\partial K}^2 \right)^{\frac{1}{2}} \left(\sum_{K \in \mathcal{P}} \|\boldsymbol{\beta}\|_{0,\infty,K} h_K^2 \|\nabla w_h|_K\|_{0,\partial K}^2 \right)^{\frac{1}{2}} \\
&\leq C\gamma \left(\sum_{K \in \mathcal{P}} h_K \|\boldsymbol{\beta}\|_{0,\infty,K} \|\nabla v_h\|_{0,K}^2 \right)^{\frac{1}{2}} \left(\sum_{K \in \mathcal{P}} h_K \|\boldsymbol{\beta}\|_{0,\infty,K} \|\nabla w_h\|_{0,K}^2 \right)^{\frac{1}{2}},
\end{aligned}$$

which proves (2.13). The proof of (2.14) follows from the last inequality and the application of the inverse

inequality (1.26). □

Remark 2.2.8. *Although all proofs presented here are based on the formulation in* (2.8)*, they hold for the alternative stabilisation term in* (2.9)*. The proofs only rely on $\boldsymbol{J}(\cdot,\cdot)$ being a symmetric, consistent, and coercive stabilisation that penalises gradient jumps. Both forms satisfy these properties: in* (2.8) *we control the full gradient, and in* (2.9) *we control the directional derivative along $\boldsymbol{\beta}$. Since the two terms are equivalent up to constants, all arguments in the well-posedness and error analysis go through unchanged.*

Also, CIP stabilisation is chosen for its simplicity, it is worth noting that the results proven in this thesis are equally valid for other linear stabilisation methods, such as streamline upwind Petrov-Galerkin (SUPG) [28], local projection stabilisation [87] or subgrid viscosity [67]. Numerical experiments confirm that the linear stabilisation term significantly improves the performance of the nonlinear solver.

2.3 The finite element method

The bound-preserving finite element method proposed in [4] reads as follows: find $u_h \in V_{\mathcal{P}}$ such that

$$a_h(u_h; v_h) = \langle f, v_h \rangle_\Omega, \qquad \forall v_h \in V_{\mathcal{P}}, \tag{2.15}$$

where the nonlinear form $a_h(\cdot;\cdot)$ is defined by

$$a_h(u_h; v_h) := a_J(u_h^+, v_h) + s(u_h^-, v_h). \tag{2.16}$$

Here, $a_J(\cdot,\cdot)$ refers to the bilinear form defined in (2.10), and the functions u_h^+ and u_h^- are defined in (1.59) and (1.60), respectively. The bilinear form $s(\cdot,\cdot)$ is introduced to control the complementary component u_h^- and is given by

$$s(v_h, w_h) = \alpha \sum_{i=1}^{N} \left(\|\mathcal{D}\|_{0,\infty,\omega_i} \mathfrak{h}(\boldsymbol{x}_i)^{d-2} + \|\boldsymbol{\beta}\|_{0,\infty,\omega_i} \mathfrak{h}(\boldsymbol{x}_i)^{d-1} + \mu \mathfrak{h}(\boldsymbol{x}_i)^d \right) v_h(\boldsymbol{x}_i) w_h(\boldsymbol{x}_i), \tag{2.17}$$

where $\alpha > 0$ is a non-dimensional constant and $\mathfrak{h}(\boldsymbol{x}_i)$ is the function which has been defined by (1.17). The stabilising form $s(\cdot,\cdot)$ induces the following norm in $V_{\mathcal{P}}$:

$$\|v_h\|_s = \sqrt{s(v_h, v_h)}. \tag{2.18}$$

The following result, derived from (1.20), demonstrates that the stabilising bilinear form $s(\cdot,\cdot)$ controls

u_h^-, specifically the kernel of the projection $(\cdot)^+$.

Lemma 2.3.1. *(see [4, Lemma 3.1]) There exists a constant $C_{\text{equiv}} > 0$, depending only on the shape regularity of $\mathcal{P}$, such that*

$$\|v_h\|_h^2 \leq \frac{C_{\text{equiv}}}{\alpha}\|v_h\|_s^2, \qquad \forall v_h \in V_{\mathcal{P}}, \tag{2.19}$$

where $\|\cdot\|_h$ is the norm defined in (2.11).

Proof. By using the inverse inequality (1.26), (1.20), and the regularity of the mesh, we have

$$\|\mathcal{D}^{\frac{1}{2}}\nabla v_h\|_{0,\Omega}^2 + \|\mu^{\frac{1}{2}} v_h\|_{0,\Omega}^2 \leq C\sum_{i=1}^{N}\left(\|\mathcal{D}\|_{0,\infty,\omega_i}\mathfrak{h}(\boldsymbol{x}_i)^{d-2} + \mu\,\mathfrak{h}(\boldsymbol{x}_i)^d\right) v_h(\boldsymbol{x}_i)^2. \tag{2.20}$$

Moreover, applying (2.14) and (1.20), we get

$$\begin{aligned} J(v_h, v_h) &\leq C\gamma \sum_{K\in\mathcal{P}} h_K^{-1}\|\boldsymbol{\beta}\|_{0,\infty,K}\|v_h\|_{0,K}^2 \\ &\leq C\gamma \sum_{K\in\mathcal{P}} h_K^{-1}h_K^d\|\boldsymbol{\beta}\|_{0,\infty,K}\sum_{i:\boldsymbol{x}_i\in K} v_h(\boldsymbol{x}_i)^2 \\ &\leq C\gamma \sum_{i=1}^{N}\|\boldsymbol{\beta}\|_{0,\infty,\omega_i}\mathfrak{h}(\boldsymbol{x}_i)^{d-1}v_h(\boldsymbol{x}_i)^2. \end{aligned}$$

Combining these bounds yields (2.19) with $C_{\text{equiv}} = (1+\gamma)C$. □

Remark 2.3.2. *The result of Lemma 2.3.1 justifies the scaling factors chosen in the definition of $s(\cdot,\cdot)$. Additionally, the formula* (2.17) *differs from the one typically used for the finite element method* (1.64) *by the inclusion of the term $\|\boldsymbol{\beta}\|_{0,\infty,\omega_i}\mathfrak{h}(\boldsymbol{x}_i)^{d-1}$, which plays a crucial role in proving* (2.19)*. This term is essential to ensuring the well-posedness of the problem and improving the error analysis. Furthermore, in our numerical experiments, this term enhances the performance of the nonlinear solver. Note that the monotonicity result which can been proved similar to the Lemma 1.6.22.*

2.3.1 Well-posedness

In this section, we analyse the existence and uniqueness of solutions for the discrete problem (2.16). To this end, we assume that the stabilisation term (2.17) satisfies the monotonicity conditions given in (1.69) and (1.70). These conditions are crucial for proving the well-posedness and establish the error analysis. The proofs of (1.69) and (1.70) for (2.17) are identical to the proof of Lemma 1.6.22 (see [12, Lemma 3.1]).

Despite the monotonicity of $s(\cdot,\cdot)$, the discrete problem (2.15) is not driven by a monotone nonlinear mapping. Therefore, the well-posedness of (2.16) requires different techniques than those applied for the finite element method (1.64). We first establish the existence of a solution using Brouwer's fixed-point theorem 1.2.4. Uniqueness is then proved by relating any solution u_h^+ of (2.15) to a corresponding variational inequality.

Theorem 2.3.3. *[4, Theorem 3.1] Suppose that $\alpha \geq C_{\text{equiv}}$. Then, there exists $u_h \in V_{\mathcal{P}}$ that solves* (2.15).

Proof. First, we define the bilinear form

$$\tilde{a}_J(v_h, w_h) := \left(\mathcal{D}\nabla v_h, \nabla w_h\right)_\Omega + \mu(v_h, w_h)_\Omega + \boldsymbol{J}(v_h, w_h), \quad \forall v_h, w_h \in V_{\mathcal{P}},$$

and the mapping

$$T : V_{\mathcal{P}} \longrightarrow V_{\mathcal{P}}, \quad \widehat{u}_h \longmapsto u_h = T(\widehat{u}_h),$$

where $u_h = T(\widehat{u}_h)$ satisfies the equation

$$\tilde{a}_J(u_h^+, v_h) + s(u_h^-, v_h) = \langle f, v_h\rangle_\Omega - (\boldsymbol{\beta}\cdot\nabla\widehat{u}_h^+, v_h)_\Omega, \quad \forall v_h \in V_{\mathcal{P}}. \tag{2.21}$$

A function u_h solves (2.15) if and only if $T(u_h) = u_h$. Therefore, to prove the existence of a solution, it is sufficient to show that T satisfies the conditions of Brouwer's fixed-point theorem 1.2.4.

i) T is well-defined: To show this, note that (2.21) is a particular instance of the finite element method (1.64). By applying 1.6.23, there exists a unique solution $u_h \in V_{\mathcal{P}}$ for (2.21), confirming that T is well-defined.

ii) T is continuous: Given that α is assumed sufficiently large, we apply the monotonicity result from Theorem 1.6.23, which gives that for all $v_h, w_h \in V_{\mathcal{P}}$,

$$\tilde{a}_J(v_h^+ - w_h^+, v_h - w_h) + s(v_h^- - w_h^-, v_h - w_h) \geq C\|v_h - w_h\|_h^2,$$

where $C > 0$ is independent of h.

Next, for $\widehat{v}_h, \widehat{w}_h \in V_{\mathcal{P}}$, let $v_h = T(\widehat{v}_h)$ and $w_h = T(\widehat{w}_h)$. By integration by parts, applying Hölder's

inequality, Lemma 2.2.4, and (2.19), we derive

$$\begin{aligned}
C\|v_h - w_h\|_h^2 &\le \tilde{a}_J(v_h^+ - w_h^+, v_h - w_h) + s(v_h^- - w_h^-, v_h - w_h) \\
&= -(\boldsymbol{\beta}\cdot\nabla(\widehat{v}_h^+ - \widehat{w}_h^+), v_h - w_h)_\Omega \\
&= (\widehat{v}_h^+ - \widehat{w}_h^+, \boldsymbol{\beta}\cdot\nabla(v_h - w_h))_\Omega \\
&\le C\|\boldsymbol{\beta}\|_{0,\infty,\Omega}\|\widehat{v}_h - \widehat{w}_h\|_{0,\Omega}|v_h - w_h|_{1,\Omega} \\
&\le C\|\boldsymbol{\beta}\|_{0,\infty,\Omega}\|\widehat{v}_h - \widehat{w}_h\|_{0,\Omega}\|\boldsymbol{D}^{-\frac{1}{2}}\|_{0,\infty,\Omega}\|\boldsymbol{D}^{\frac{1}{2}}\nabla(v_h - w_h)\|_{0,\Omega} \\
&\le C\frac{\|\boldsymbol{\beta}\|_{0,\infty,\Omega}}{d_0^{\frac{1}{2}}}\|\widehat{v}_h - \widehat{w}_h\|_{0,\Omega}\|v_h - w_h\|_h.
\end{aligned}$$

Therefore,

$$\|T(\widehat{v}_h) - T(\widehat{w}_h)\|_h \le C\frac{\|\boldsymbol{\beta}\|_{0,\infty,\Omega}}{d_0^{\frac{1}{2}}}\|\widehat{v}_h - \widehat{w}_h\|_{0,\Omega},$$

which shows that T is Lipschitz continuous.

iii) There exists $R > 0$, such that $T(B(0,R)) \subseteq B(0,R)$: Let $\widehat{z}_h \in V_{\mathcal{P}}$ be arbitrary and $z_h = T(\widehat{z}_h)$. By using $v_h = z_h^+$ in (2.21), and applying Cauchy-Schwarz and Hölder's inequalities, and (2.7), we obtain

$$\begin{aligned}
\tilde{a}_J(z_h^+, z_h^+) + \underbrace{s(z_h^-, z_h^+)}_{\ge 0} &= \langle f, z_h^+\rangle_\Omega - (\boldsymbol{\beta}\cdot\nabla\widehat{z}_h^+, z_h^+)_\Omega \\
&\le \|f\|_{0,\Omega}\|z_h^+\|_{0,\Omega} + (\widehat{z}_h^+, \boldsymbol{\beta}\cdot\nabla z_h^+)_\Omega \\
&\le C\|f\|_{0,\Omega}\mu^{-\frac{1}{2}}\|z_h^+\|_h + \|\boldsymbol{\beta}\|_{0,\infty,\Omega}\|\widehat{z}_h^+\|_{0,\Omega}d_0^{-\frac{1}{2}}\|z_h^+\|_h \\
&\le C\left\{\frac{\|f\|_{0,\Omega}}{\mu^{\frac{1}{2}}} + \frac{\|\boldsymbol{\beta}\|_{0,\infty,\Omega}\kappa}{d_0^{\frac{1}{2}}}\right\}\|z_h^+\|_h.
\end{aligned}$$

Thus, z_h^+ satisfies

$$\|z_h^+\|_h \le C\left\{\frac{\|f\|_{0,\Omega}}{\mu^{\frac{1}{2}}} + \frac{\|\boldsymbol{\beta}\|_{0,\infty,\Omega}\kappa}{d_0^{\frac{1}{2}}}\right\}. \tag{2.22}$$

Next, taking $v_h = z_h^-$ in (2.21), integrating by parts, and using Hölder's inequality, we obtain

$$\begin{aligned}\tilde{a}_J(z_h^+, z_h^-) + s(z_h^-, z_h^-) &= \langle f, z_h^- \rangle_\Omega - (\boldsymbol{\beta} \cdot \nabla \widehat{z}_h^+, z_h^-)_\Omega \\ &\leq \|f\|_{0,\Omega} \|z_h^-\|_{0,\Omega} + \|\widehat{z}_h^+\|_{0,\Omega} \|\boldsymbol{\beta}\|_{0,\infty,\Omega} |z_h^-|_{1,\Omega}.\end{aligned}$$

Using (2.7), we further derive

$$\begin{aligned}\tilde{a}_J(z_h^+, z_h^-) + s(z_h^-, z_h^-) &\leq C \left(\|f\|_{0,\Omega} \|z_h^-\|_{0,\Omega} + \kappa \|\boldsymbol{\beta}\|_{0,\infty,\Omega} |z_h^-|_{1,\Omega} \right) \\ &\leq C \left(\frac{\|f\|_{0,\Omega}}{\mu^{\frac{1}{2}}} + \kappa \frac{\|\boldsymbol{\beta}\|_{0,\infty,\Omega}}{d_0^{\frac{1}{2}}} \right) \|z_h^-\|_h.\end{aligned}$$

Applying (2.19) and Young's inequality, we have

$$\begin{aligned}\tilde{a}_J(z_h^+, z_h^-) + s(z_h^-, z_h^-) &\leq C \left(\frac{\|f\|_{0,\Omega}}{\mu^{\frac{1}{2}}} + \kappa \frac{\|\boldsymbol{\beta}\|_{0,\infty,\Omega}}{d_0^{\frac{1}{2}}} \right)^2 + \frac{s(z_h^-, z_h^-)}{2} \\ &=: \frac{M}{2} + \frac{s(z_h^-, z_h^-)}{2}.\end{aligned}$$

Using Young's and Cauchy-Schwarz's inequalities for $\tilde{a}_J(z_h^+, z_h^-)$, and (2.22), we get

$$\begin{aligned}-\delta \tilde{a}_J(z_h^-, z_h^-) + s(z_h^-, z_h^-) &\leq M + C\delta^{-1} \tilde{a}_J(z_h^+, z_h^+) \\ &\leq M + C\delta^{-1} \left\{ \frac{\|f\|_{0,\Omega}}{\mu^{\frac{1}{2}}} + \frac{\|\boldsymbol{\beta}\|_{0,\infty,\Omega} \kappa}{d_0^{\frac{1}{2}}} \right\},\end{aligned}$$

for any $\delta > 0$. Choosing δ small enough, and using Lemma 2.3.1, we get

$$\|z_h^-\|_h \leq C \left(-\delta \tilde{a}_J(z_h^-, z_h^-) + s(z_h^-, z_h^-) \right) \leq C_2(f, \mu, \mathcal{D}, \boldsymbol{\beta}, \kappa),$$

where

$$C_2(f, \mu, \mathcal{D}, \boldsymbol{\beta}, \kappa) = M + C \left\{ \frac{\|f\|_{0,\Omega}}{\mu^{\frac{1}{2}}} + \frac{\|\boldsymbol{\beta}\|_{0,\infty,\Omega} \kappa}{d_0^{\frac{1}{2}}} \right\}.$$

Hence, $z_h = T(\widehat{z}_h)$ satisfies the following (uniform) bound

$$\|z_h\|_h \leq \|z_h^-\|_h + \|z_h^+\|_h \leq C\left\{\frac{\|f\|_{0,\Omega}}{\mu^{\frac{1}{2}}} + \frac{\|\boldsymbol{\beta}\|_{0,\infty,\Omega}\kappa}{d_0^{\frac{1}{2}}}\right\} + C_2(f,\mu,\mathcal{D},\boldsymbol{\beta},\kappa)$$
$$=: R.$$

Thus, $z_h = T(\widehat{z}_h) \in B(0,R)$ for every $\widehat{z}_h \in V_{\mathcal{P}}$, which shows that $T(B(0,R)) \subseteq B(0,R)$.

Therefore, using Brouwer's fixed point theorem 1.2.4, there exists $u_h \in V_{\mathcal{P}}$ such that $T(u_h) = u_h$. In other words, problem (2.15) has at least one solution. □

The previous result, while establishing the existence of solutions, does not guarantee their uniqueness. To address this gap, the next two results will not only establish uniqueness but also provide a useful characterisation of u_h^+.

Lemma 2.3.4. *[4, Lemma 3.3] Let $u_h \in V_{\mathcal{P}}$ be a solution to* (2.15)*. Then, $u_h^+ \in V_{\mathcal{P}}^+$ satisfies the following variational inequality*

$$a_J(u_h^+, v_h - u_h^+) \geq \langle f, v_h - u_h^+\rangle_\Omega \qquad \forall v_h \in V_{\mathcal{P}}^+, \tag{2.23}$$

where $a_J(\cdot,\cdot)$ is as defined in (2.10)*. Additionally, u_h^- is uniquely determined as the solution to:*

$$s(u_h^-, v_h) = \langle f, v_h\rangle_\Omega - a_J(u_h^+, v_h) \qquad \forall v_h \in V_{\mathcal{P}}. \tag{2.24}$$

Proof. We begin by setting $v_h \in V_{\mathcal{P}}^+$ and then $v_h = u_h^+$ in equation (2.15) as the test function. This yields

$$a_J(u_h^+, v_h) + s(u_h^-, v_h) = \langle f, v_h\rangle_\Omega,$$
$$a_J(u_h^+, u_h^+) + s(u_h^-, u_h^+) = \langle f, u_h^+\rangle_\Omega.$$

Subtracting the second equation from the first, we obtain

$$a_J(u_h^+, v_h - u_h^+) + s(u_h^-, v_h - u_h^+) = \langle f, v_h - u_h^+\rangle_\Omega, \qquad \forall v_h \in V_{\mathcal{P}}^+.$$

Using Lemma 1.6.22 and equation (1.70), we conclude that $u_h^+ \in V_{\mathcal{P}}^+$ satisfies the variational inequality (2.23). Furthermore, since $s(\cdot,\cdot)$ is an elliptic bilinear form on $V_{\mathcal{P}}$, the uniqueness of u_h^- follows directly, and

this completing the proof. □

This lemma provides an insight on decomposing the solution u_h into two successive problems, (2.23) and (2.24). This decomposition characterises any solution of (2.15).

Corollary 2.3.5. *[4, Corollary 3.1] The problem* (2.15) *has a unique solution.*

Proof. Suppose $u_1, u_2 \in V_{\mathcal{P}}$ are two solutions to (2.15). Then, u_1^+ and u_2^+ satisfy the variational inequality (2.23). By Stampacchia's Theorem 1.2.5, the solution to this problem is unique, hence $u_1^+ = u_2^+$. Given this, equation (2.24) holds for both u_1^- and u_2^-. Since the right-hand side of both equations is identical, and $s(\cdot,\cdot)$ is an elliptic bilinear form, we conclude that $u_1^- = u_2^-$. Thus, $u_1 = u_1^+ + u_1^- = u_2^+ + u_2^- = u_2$. □

Remark 2.3.6. *[4, Remark 3.2] We close this section by noting that the complementary part u_h^- of u_h has a local support. Specifically, observe that:*

$$(u_h^+ + u_h^-)^+ = (u_h^+ + u_h - u_h^+)^+ = u_h^+.$$

This implies that $u_h^-(\boldsymbol{x}_i) \neq 0$ if and only if $u_h^+(\boldsymbol{x}_i) = \kappa$ or $u_h^+(\boldsymbol{x}_i) = 0$. Hence, the support of u_h^- is restricted to regions where $u_h^+ = 0$ or $u_h^+ = \kappa$, meaning that u_h^- has a local support in regions where the constraint in definition of $V_{\mathcal{P}}^+$ is active.

2.4 Error analysis

This section is devoted to the error analysis of the method presented in (2.15). The primary aim here is to ensure that the discrete solution respects the bounds imposed by the continuous problem. Therefore, the error estimates will be proven for the constrained component, u_h^+.

Theorem 2.4.1. *[4, Theorem 4.1] Let $u \in H^{k+1}(\Omega) \cap H_0^1(\Omega)$ be the solution of* (2.1)*, and $u_h \in V_{\mathcal{P}}$ the solution of* (2.15)*. Then, there exists a constant $C > 0$, independent of $\mathcal{D}$, μ, $\boldsymbol{\beta}$, and h, such that*

$$\|u - u_h^+\|_h \leq Ch^k \left(\|\mathcal{D}\|_{0,\infty,\Omega}^{\frac{1}{2}} + \mu^{-\frac{1}{2}} \|\boldsymbol{\beta}\|_{0,\infty,\Omega} + h^{\frac{1}{2}} \|\boldsymbol{\beta}\|_{0,\infty,\Omega}^{\frac{1}{2}} + h\mu^{\frac{1}{2}} \right) |u|_{k+1,\Omega}. \tag{2.25}$$

Proof. As usual, we decompose the error $u - u_h^+$ as follows

$$u - u_h^+ = (u - \pi(u)) + (\pi(u) - u_h^+) =: \eta_h + e_h\,, \tag{2.26}$$

where π is the $L^2(\Omega)$-orthogonal projection defined in (1.24).

The bound for η_h is a direct consequence of (1.25) and (2.13). Indeed, using the Cauchy-Schwarz and Young inequalities, we obtain

$$\begin{aligned}\|\eta_h\|_h^2 = a_J(\eta_h,\eta_h) &= (\mathcal{D}\nabla\eta_h,\nabla\eta_h)_\Omega + \mu(\eta_h,\eta_h)_\Omega + J(\eta_h,\eta_h)\\ &\le C\left(\|\mathcal{D}\|_{0,\infty,\Omega}|\eta_h|_{1,\Omega}^2 + \mu\|\eta_h\|_{0,\Omega}^2 + h\|\boldsymbol{\beta}\|_{0,\infty,\Omega}|\eta_h|_{1,\Omega}^2\right)\\ &\le Ch^{2k}\left(\|\mathcal{D}\|_{0,\infty,\Omega} + h\|\boldsymbol{\beta}\|_{0,\infty,\Omega} + h^2\mu\right)|u|_{k+1,\Omega}^2.\end{aligned}$$

To bound $\|e_h\|_h$, we use the ellipticity of $a_J(\cdot,\cdot)$ to get

$$\|e_h\|_h^2 = -a_J(\eta_h,e_h) + a_J(u-u_h^+,\pi(u)-u_h^+) =: \mathrm{I}+\mathrm{II}. \tag{2.27}$$

We first decompose I as follows

$$\mathrm{I} = (\mathcal{D}\nabla\eta_h,\nabla e_h)_\Omega + (\boldsymbol{\beta}\cdot\nabla\eta_h,e_h)_\Omega + \mu(\eta_h,e_h)_\Omega + J(\eta_h,e_h) =: \text{(a)}+\text{(b)}+\text{(c)}+\text{(d)}, \tag{2.28}$$

Each term in (2.28) is bounded separately. Using Cauchy-Schwarz inequality and (1.25), we have

$$\text{(a)} \le \|\mathcal{D}\|_{0,\infty,\Omega}^{\frac{1}{2}}|\eta_h|_{1,\Omega}\|\mathcal{D}^{\frac{1}{2}}\nabla e_h\|_{0,\Omega} \le Ch^k\|\mathcal{D}\|_{0,\infty,\Omega}^{\frac{1}{2}}|u|_{k+1,\Omega}\|e_h\|_h. \tag{2.29}$$

For (b), integrating by parts, using the orthogonality of π, and Lemma 2.2.7, we get

$$\begin{aligned}\text{(b)} &= (\pi(u)-u,\boldsymbol{\beta}\cdot\nabla e_h)_\Omega\\ &= \left(\pi(u)-u,\boldsymbol{\beta}\cdot\nabla e_h - \pi(\boldsymbol{\beta}\cdot\nabla e_h)\right)_\Omega\\ &\le \|\eta_h\|_{0,\Omega}\|\boldsymbol{\beta}\cdot\nabla e_h - \pi(\boldsymbol{\beta}\cdot\nabla e_h)\|_{0,\Omega}\\ &\le Ch^{k+\frac{1}{2}}\|\boldsymbol{\beta}\|_{0,\infty,\Omega}^{\frac{1}{2}}|u|_{k+1,\Omega}\|e_h\|_h.\end{aligned} \tag{2.30}$$

Similarly, for (c), we have

$$\text{(c)} \le \mu^{\frac{1}{2}}\|\eta_h\|_{0,\Omega}\|\mu^{\frac{1}{2}}e_h\|_{0,\Omega} \le Ch^{k+1}\mu^{\frac{1}{2}}|u|_{k+1,\Omega}\|e_h\|_h. \tag{2.31}$$

Finally, for (d), since $J(\cdot,\cdot)$ is semi-positive definite, applying Cauchy-Schwarz inequality and using Lemma 2.2.7

along with (1.25), we obtain

$$\begin{aligned}(\mathrm{d}) &\leq J(\eta_h, \eta_h)^{\frac{1}{2}} J(e_h, e_h)^{\frac{1}{2}} \\ &\leq C h^{k+\frac{1}{2}} \|\boldsymbol{\beta}\|_{0,\infty,\Omega}^{\frac{1}{2}} |u|_{k+1,\Omega} \|e_h\|_h .\end{aligned} \tag{2.32}$$

Substituting the bounds from (2.29), (2.30), (2.31), and (2.32) into (2.28), we get

$$\mathrm{I} \leq C h^k \left(\|\mathcal{D}\|_{0,\infty,\Omega}^{\frac{1}{2}} + h^{\frac{1}{2}} \|\boldsymbol{\beta}\|_{0,\infty,\Omega}^{\frac{1}{2}} + h\mu^{\frac{1}{2}} \right) |u|_{k+1,\Omega} \|e_h\|_h . \tag{2.33}$$

To bound the term II, we first recall that $(\pi(u))^- = \pi(u) - (\pi(u))^+$, which gives us

$$\mathrm{II} = a_J(u - u_h^+, e_h) = a_J(u - u_h^+, (\pi(u))^+ - u_h^+) + a_J(u - u_h^+, (\pi(u))^-).$$

Due to the regularity of u, we have $J(u, (\pi(u))^+ - u_h^+) = 0$. Thus, since u_h^+ solves the variational problem (2.23), we obtain

$$\begin{aligned}a_J(u - u_h^+, (\pi(u))^+ - u_h^+) &= a_J(u, (\pi(u))^+ - u_h^+) - a_J(u_h^+, (\pi(u))^+ - u_h^+) \\ &= \langle f, (\pi(u))^+ - u_h^+ \rangle_\Omega - a_J(u_h^+, (\pi(u))^+ - u_h^+) \leq 0.\end{aligned}$$

Therefore,

$$\mathrm{II} = a_J(u - u_h^+, e_h) \leq a_J(u - u_h^+, (\pi(u))^-).$$

The term on the right-hand side is essentially a consistency error and requires special treatment. Let i_h be the Lagrange interpolation operator defined in (1.22). Since $u(x) \in [0, \kappa]$ almost everywhere in Ω, we have $i_h(u) \in V_{\mathcal{P}}^+$, implying that $(i_h(u))^- = 0$. Thus,

$$a_J(u - u_h^+, (\pi(u))^-) = a_J(u - u_h^+, (\pi(u))^- - (i_h(u))^-).$$

Using the definition of $a_J(\cdot,\cdot)$, we now bound II

$$\begin{aligned}
\text{II} &\leq a_J(u-u_h^+, (\pi(u))^- - (i_h(u))^-) \\
&= \left(\mathcal{D}\nabla(u-u_h^+), \nabla((\pi(u))^- - (i_h(u))^-)\right)_\Omega + (\boldsymbol{\beta}\cdot\nabla(u-u_h^+), (\pi(u))^- - (i_h(u))^-)_\Omega \\
&\quad + \mu(u-u_h^+, (\pi(u))^- - (i_h(u))^-)_\Omega + J(u-u_h^+, (\pi(u))^- - (i_h(u))^-) \\
&= \text{(e)} + \text{(f)} + \text{(g)} + \text{(h)}.
\end{aligned} \tag{2.34}$$

We begin bounding the term (e) using the Cauchy-Schwarz inequality

$$\begin{aligned}
\text{(e)} &\leq \|\mathcal{D}\|_{0,\infty,\Omega}^{\frac{1}{2}} \|\mathcal{D}^{\frac{1}{2}}\nabla(u-u_h^+)\|_{0,\Omega} |(\pi(u))^- - (i_h(u))^-|_{1,\Omega} \\
&\leq Ch^{-1}\|\mathcal{D}\|_{0,\infty,\Omega}^{\frac{1}{2}} \|\mathcal{D}^{\frac{1}{2}}\nabla(u-u_h^+)\|_{0,\Omega} \|(\pi(u))^- - (i_h(u))^-\|_{0,\Omega},
\end{aligned} \tag{2.35}$$

where we have applied an inverse inequality. Since $(\cdot)^-$ is Lipschitz continuous (see Lemma 2.2.4), we can further bound (e) as

$$\begin{aligned}
\text{(e)} &\leq Ch^{-1}\|\mathcal{D}\|_{0,\infty,\Omega}^{\frac{1}{2}} \|\mathcal{D}^{\frac{1}{2}}\nabla(u-u_h^+)\|_{0,\Omega} \|\pi(u) - i_h(u)\|_{0,\Omega} \\
&\leq Ch^k \|\mathcal{D}\|_{0,\infty,\Omega}^{\frac{1}{2}} |u|_{k+1,\Omega} \|u-u_h^+\|_h,
\end{aligned}$$

using the approximation properties of i_h and π (see (1.23) and (1.25)).

Next, for term (f), we integrate by parts

$$\begin{aligned}
\text{(f)} &= -\left(u-u_h^+, \boldsymbol{\beta}\cdot\nabla((\pi(u))^- - (i_h(u))^-)\right)_\Omega \\
&\leq \|\boldsymbol{\beta}\|_{0,\infty,\Omega} \|u-u_h^+\|_{0,\Omega} |(\pi(u))^- - (i_h(u))^-|_{1,\Omega},
\end{aligned}$$

by the Cauchy-Schwarz inequality. Using an inverse estimate, we get

$$\begin{aligned}
\text{(f)} &\leq Ch^{-1}\|\boldsymbol{\beta}\|_{0,\infty,\Omega} \|u-u_h^+\|_{0,\Omega} \|(\pi(u))^- - (i_h(u))^-\|_{0,\Omega} \\
&\leq Ch^k \|\boldsymbol{\beta}\|_{0,\infty,\Omega}\, \mu^{-\frac{1}{2}} |u|_{k+1,\Omega} \|u-u_h^+\|_h,
\end{aligned} \tag{2.36}$$

again using the Lipschitz continuiuty of $(\cdot)^-$ and the approximation properties for i_h and π. Now (g) is controlled in the same way using the Cauchy-Schwarz inequality

$$\begin{aligned}
\text{(g)} &\leq \mu^{\frac{1}{2}} \|\mu^{\frac{1}{2}}(u-u_h^+)\|_{0,\Omega} \|(\pi(u))^- - (i_h(u))^-\|_{0,\Omega} \\
&\leq Ch^{k+1} \mu^{\frac{1}{2}} |u|_{k+1,\Omega} \|u-u_h^+\|_h,
\end{aligned} \tag{2.37}$$

and by Lipschitz continuiuty of $(\cdot)^-$ and the approximation properties for i_h and π. Finally for (h) Cauchy-Schwarz implies

$$\begin{aligned}(\mathrm{h}) &\leq J(u-u_h^+,u-u_h^+)^{\frac{1}{2}} J((\pi(u))^- - (i_h(u))^-,(\pi(u))^- - (i_h(u))^-)^{\frac{1}{2}} \\ &\leq C\|\boldsymbol{\beta}\|_{0,\infty,\Omega}^{\frac{1}{2}} h^{\frac{1}{2}} |(\pi(u))^- - (i_h(u))^-|_{1,\Omega} J(u-u_h^+,u-u_h^+)^{\frac{1}{2}},\end{aligned}$$

where we used Lemma 2.2.7. Now by an inverse inequality

$$\begin{aligned}(\mathrm{h}) &\leq C\|\boldsymbol{\beta}\|_{0,\infty,\Omega}^{\frac{1}{2}} h^{-\frac{1}{2}} \|(\pi(u))^- - (i_h(u))^-\|_{0,\Omega} J(u-u_h^+,u-u_h^+)^{\frac{1}{2}} \\ &\leq Ch^{k+\frac{1}{2}} \|\boldsymbol{\beta}\|_{0,\infty,\Omega}^{\frac{1}{2}} |u|_{k+1,\Omega} \|u-u_h^+\|_h,\end{aligned} \tag{2.38}$$

through the approximability of i_h and π and the Lipschitz continuiuty of $(\cdot)^-$.

Combining the bounds from (2.35), (2.36), (2.37), and (2.38), we obtain the following estimate for II

$$\mathrm{II} \leq Ch^k \left(\|\mathcal{D}\|_{0,\infty,\Omega}^{\frac{1}{2}} + \mu^{-\frac{1}{2}} \|\boldsymbol{\beta}\|_{0,\infty,\Omega} + h^{\frac{1}{2}} \|\boldsymbol{\beta}\|_{0,\infty,\Omega}^{\frac{1}{2}} + h\mu^{\frac{1}{2}} \right) \|u-u_h^+\|_h |u|_{k+1,\Omega}. \tag{2.39}$$

Substituting (2.33) and (2.39) into (2.27) and applying Young's inequality, we derive the following bound for $\|e_h\|_h$

$$\|e_h\|_h^2 \leq Ch^{2k} \left(\|\mathcal{D}\|_{0,\infty,\Omega}^{\frac{1}{2}} + h^{\frac{1}{2}} \|\boldsymbol{\beta}\|_{0,\infty,\Omega}^{\frac{1}{2}} + \mu^{-\frac{1}{2}} \|\boldsymbol{\beta}\|_{0,\infty,\Omega} + h\mu^{\frac{1}{2}} \right)^2 |u|_{k+1,\Omega}^2 + \frac{1}{2}\|e_h\|_h^2 + \frac{1}{8}\|u-u_h^+\|_h^2.$$

Finally, collecting the bounds that have been obtained for $\|e_h\|_h$ and $\|\eta_h\|_h$ gives

$$\begin{aligned}\|u-u_h^+\|_h &\leq \|e_h\|_h + \|\eta_h\|_h \\ &\leq Ch^k \left(\|\mathcal{D}\|_{0,\infty,\Omega}^{\frac{1}{2}} + \mu^{-\frac{1}{2}} \|\boldsymbol{\beta}\|_{0,\infty,\Omega} + h^{\frac{1}{2}} \|\boldsymbol{\beta}\|_{0,\infty,\Omega}^{\frac{1}{2}} + h\mu^{\frac{1}{2}} \right) |u|_{k+1,\Omega} + \frac{1}{2}\|u-u_h^+\|_h,\end{aligned}$$

and by rearranging terms (2.25) follows. □

2.4.1 The extension to problems with non-homogeneous boundary conditions

In this section, we extend the method for non-homogeneous Dirichlet conditions. Let us consider the following modified version of (2.1)

$$\begin{cases} -\mathrm{div}(\mathcal{D}\nabla u) + \boldsymbol{\beta}\cdot\nabla u + \mu u &= f \quad \text{in } \Omega, \\ u &= g \quad \text{on } \partial\Omega, \end{cases} \tag{2.40}$$

where $g \in H^{\frac{1}{2}}(\partial\Omega)$, with $g \geq 0$ on $\partial\Omega$. By using the definition of κ and applying the maximum principle 1.5.4 and comparison principles 1.5.3, we have $\kappa \geq \|g\|_{0,\infty,\partial\Omega}$. For simplicity, we assume that g is the trace of a function belongs to $\tilde{V}_{\mathcal{P}}$ (which has been defined in (1.14)).

Next, we define the set of nodes of the triangulation, including boundary nodes, as $\boldsymbol{x}_1, \dots, \boldsymbol{x}_P$, with the interior nodes being denoted as $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$, where $N < P$. We now introduce an extension of g into the domain Ω by defining $u_{h,g} \in \tilde{V}_{\mathcal{P}}$ as follows

$$u_{h,g}(\boldsymbol{x}_i) = \begin{cases} g(\boldsymbol{x}_i) & \text{if } i \in \{N+1, \dots, P\}, \\ 0 & \text{else}. \end{cases} \tag{2.41}$$

With this extension, the formulation analogous to (2.15) for the non-homogeneous boundary case is: Find $\tilde{u}_h \in V_{\mathcal{P}}$ such that

$$a_J\big((\tilde{u}_h + u_{h,g})^+, v_h\big) + s\big((\tilde{u}_h + u_{h,g})^-, v_h\big) = \langle f, v_h\rangle_\Omega \qquad \forall\, v_h \in V_{\mathcal{P}}\,. \tag{2.42}$$

The choice of $u_{h,g}$ as the extension of g is motivated by the fact that, at each node of $\mathcal{P}$, either $\tilde{u}_h$ or $u_{h,g}$ is zero. This ensures that the following property holds

$$(\tilde{u}_h + u_{h,g})^+ = \tilde{u}_h^+ + u_{h,g}\,, \tag{2.43}$$

which leads to $(\tilde{u}_h + u_{h,g})^- = \tilde{u}_h^-$. Therefore, (2.42) can be rewritten as: Find $\tilde{u}_h \in V_{\mathcal{P}}$ such that

$$a_J\big(\tilde{u}_h^+, v_h\big) + s\big(\tilde{u}_h^-, v_h\big) = \langle f, v_h\rangle_\Omega - a_J\big(u_{h,g}, v_h\big) \qquad \forall\, v_h \in V_{\mathcal{P}}\,. \tag{2.44}$$

Now, the proof of Theorem 2.3.3 remains valid, confirming the existence of a solution for (2.44). For unique-

ness, the same reasoning as in Lemma 2.3.4 applies, showing that $\tilde{u}_h^+$ solves the following variational inequality: $\tilde{u}_h^+ \in V_{\mathcal{P}}^+$ and

$$a_J(\tilde{u}_h^+, v_h - \tilde{u}_h^+) \geq \langle f, v_h - \tilde{u}_h^+ \rangle_\Omega - a_J\big(u_{h,g}, v_h - \tilde{u}_h^+\big) \qquad \forall\, v_h \in V_{\mathcal{P}}^+ \,. \tag{2.45}$$

Follows exactly the same arguments as those for (2.15) and thanks to Stampacchia's Theorem, (2.45) has a unique solution, This proves the existence and uniqueness of the solution of the problem (2.42). Finally, for the error analysis, assuming enough regularity for the exact solution, we have

$$a_J(u, v_h - \tilde{u}_h^+) = \langle f, v_h - u_h^+ \rangle_\Omega \,,$$

for all $v_h \in V_{\mathcal{P}}$. Thus, the following variational inequality holds

$$a_J((\tilde{u}_h + u_{h,g})^+ - u, v_h - u_h^+) \geq 0 \qquad \forall\, v_h \in V_{\mathcal{P}}^+ \,. \tag{2.46}$$

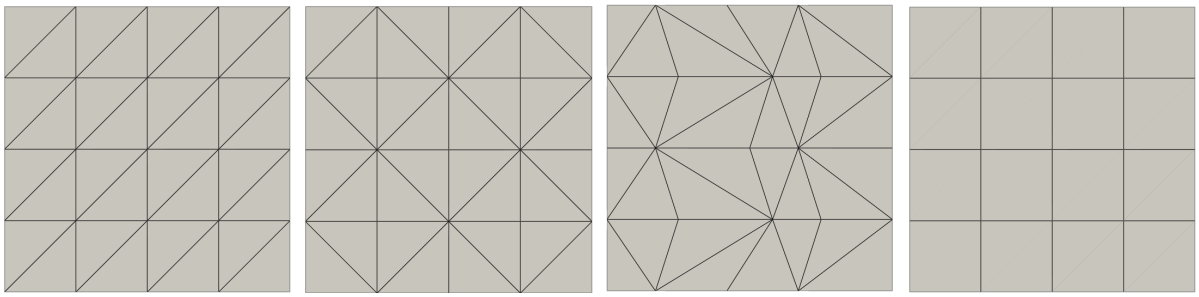
This inequality plays a key role in obtaining the bound for II in the proof of Theorem 2.4.1. Consequently, the error analysis follows a similar approach as outlined in the proof of Theorem 2.4.1.

2.5 Numerical experiments

In this section, we present three different problems to evaluate the performance of the finite element method (2.15). Throughout these experiments, the computational domain is $\Omega = (0,1)^2$, and we use $\alpha = 1$ for the stabilising bilinear form $s(\cdot,\cdot)$.

Remark 2.5.1. *In the proof of Theorem 2.3.3, we require $\alpha \geq C_{\text{equiv}}$. However, the exact value of C_{equiv} is not known in practice, and its computation is not feasible. In the numerical experiments, we therefore set $\alpha = 1$, which has been observed to provide sufficient stability for solving the nonlinear system.*

We consider three different mesh types, with the coarsest level of each shown in Figure 2.1. The meshes in Figures 2.1a and 2.1b are symmetric and Delaunay, while the mesh in Figure 2.1c is non-Delaunay, and the one in Figure 2.1d is quadrilateral. The non-Delaunay mesh in Figure 2.1c is obtained by shifting some of the interior nodes from the mesh in Figure 2.1b to the right, which results in the formation of obtuse angles. This choice of non-Delaunay mesh is motivated by this fact that the discrete maximum principle often fails for finite element methods on such meshes (see, for example, 17). As a result, the initial datum u_h^0, defined in the next section, will not generally lie in $V_{\mathcal{P}}^+$.



(a) A symmetric, Delaunay mesh. (b) A symmetric, Delaunay mesh. (c) A non-symmetric, non-Delaunay mesh. (d) A simple quadrilateral mesh.

Figure 2.1: Three coarse level indicative meshes used in the experiments all with $N = 5$.

To solve the nonlinear system associated with (2.16), we use a Richardson-like iterative method. We first set $u_h^0 \in V_{\mathcal{P}}$ such that it satisfies the following (CIP) problem

$$a_J(u_h^0, v_h) = \langle f, v_h \rangle_\Omega \qquad \forall v_h \in V_{\mathcal{P}}\,. \tag{2.47}$$

Then, for $n = 0, 1, 2, \ldots$, we compute $u_h^{n+1} \in V_{\mathcal{P}}$ by solving

$$a_J(u_h^{n+1}, v_h) = a_J(u_h^n, v_h) + \omega\left(\langle f, v_h \rangle_\Omega - a_J((u_h^n)^+, v_h) - s((u_h^n)^-, v_h)\right) \qquad \forall v_h \in V_{\mathcal{P}}, \tag{2.48}$$

where $\omega \in (0, 1]$ is a damping parameter. The iterations continues until the stopping criterion is satisfied, i.e.,

$$\|u_h^{n+1} - u_h^n\|_{0,\Omega} \leq 10^{-8}. \tag{2.49}$$

The discretisation of problem (2.48) follows the same approach as in (1.77); for a detailed discussion, see Section 1.6.30.

In all figures, $N - 1$ represents the number of divisions in the x and y directions, so the total number of vertices (including the boundary) is N^2. We test the performance of the method asymptotically in N, where we use EOC as the estimated order of convergence, and also examine the convergence of the iterative method. We have used $\mathbb{P}_1$, $\mathbb{P}_2$, and $\mathbb{P}_3$ elements with the triangular meshes, and $\mathbb{Q}_1$ and $\mathbb{Q}_2$ elements with the quadrilateral mesh.

Example 3 (Convergence for a problem with a smooth solution)**.** *Consider the case where $\mu = 1$, and the diffusion matrix is given by $\mathcal{D} = \epsilon \begin{bmatrix} 100 & \cos(x) \\ \cos(x) & 1 \end{bmatrix}$ with $\epsilon = 10^{-5}$, and the convection vector is $\boldsymbol{\beta} = (2, 1)$. The source term f is chosen such that the exact solution to* (2.1) *is the function $u(x, y) = 100 \sin(\pi x) \sin(\pi y)$.*

Note that $u(x)$ *lies within the interval* $[0, 100]$*, and hence we select* $\kappa = 100$*. For the CIP stabilisation, we use a penalty parameter of* $\gamma = 0.025$ *as defined in* (2.8)*, and the damping parameter in the iterative method* (2.48) *is* $\omega = 1$.

In Tables 2.1-2.5, we present the convergence results measured in both the $\|\cdot\|_{0,\Omega}$ and $\|\cdot\|_h$ norms for the difference $u - u_h^+$, as well as in the $\|\cdot\|_s$-norm for the complementary part u_h^-. Also, we include the number of iterations required to achieve convergence for the nonlinear system. These results indicate an optimal convergence rate for the constrained component u_h^+, aligning with the theoretical findings from Section 2.4. Moreover, they show a higher convergence (to zero) the complementary part u_h^-.

N	Itr	$\|u - u_h^+\|_{0,\Omega}$	EOC	$\|u - u_h^+\|_h$	EOC	$\|u_h^-\|_s$	EOC
5	2	8.57e+0	–	4.55e+1	–	0	–
9	7	2.12e+0	2.37	1.95e+1	1.44	3.05e-1	–
17	6	5.05e-1	2.26	7.71e+0	1.46	2.74e-1	0.17
33	7	1.23e-1	2.12	2.89e+0	1.47	6.64e-2	2.13
65	7	3.09e-2	2.03	1.06e+0	1.47	1.27e-2	2.44
129	6	7.80e-3	2.00	3.82e-1	1.49	2.29e-3	2.50

Table 2.1: Numerical results for Example 3 using $\mathbb{P}_1$ elements and Mesh 2.1c.

N	Itr	$\|u - u_h^+\|_{0,\Omega}$	EOC	$\|u - u_h^+\|_h$	EOC	$\|u_h^-\|_s$	EOC
5	15	5.51e+0	–	2.73e+1	–	4.43e+0	–
9	15	8.03e-1	3.27	9.79e+0	1.74	8.43e-1	2.82
17	13	1.38e-1	2.76	3.47e+0	1.63	1.67e-1	2.54
33	12	2.86e-2	2.37	1.23e+0	1.56	3.12e-2	2.52
65	10	6.62e-3	2.15	4.37e-1	1.53	5.70e-3	2.50
129	9	1.61e-3	2.06	1.56e-1	1.50	1.02e-3	2.51

Table 2.2: Numerical results for Example 3 using $\mathbb{Q}_1$ elements and Mesh 2.1d.

N	Itr	$\|u - u_h^+\|_{0,\Omega}$	EOC	$\|u - u_h^+\|_h$	EOC	$\|u_h^-\|_s$	EOC
5	15	2.51e+0	–	8.14e+0	–	1.37e+0	–
9	2	3.02e-1	3.60	1.52e+0	2.85	0	–
17	2	3.72e-2	3.29	3.03e-1	2.53	0	–
33	2	4.44e-3	3.20	5.78e-2	2.50	0	–
65	2	5.37e-4	3.11	1.07e-2	2.48	0	–
129	2	6.57e-5	3.03	1.97e-3	2.44	0	–

Table 2.3: Numerical results for Example 3 using $\mathbb{P}_2$ elements and Mesh 2.1d.

N	Itr	$\|u - u_h^+\|_{0,\Omega}$	EOC	$\|u - u_h^+\|_h$	EOC	$\|u_h^-\|_s$	EOC
5	2	3.77e-1	–	6.22e-1	–	0	–
9	58	4.26e-2	3.70	9.79e-2	3.14	1.85e-2	4.06
17	44	5.18e-3	3.31	1.71e-2	2.74	2.44e-3	3.18
33	28	6.36e-4	3.16	3.21e-3	2.52	2.45e-4	3.46
65	2	7.75e-5	3.10	6.43e-4	2.37	5.28e-6	5.66
129	2	9.20e-6	3.10	1.37e-4	2.25	4.35e-7	3.62

Table 2.4: Numerical results for Example 3 using $\mathbb{Q}_2$ elements and Mesh 2.1d..

N	Itr	$\|u - u_h^+\|_{0,\Omega}$	EOC	$\|u - u_h^+\|_h$	EOC	$\|u_h^-\|_s$	EOC
5	2	2.85e-1	–	8.75e-1	–	0	–
9	137	2.69e-2	4.01	9.75e-2	3.73	6.62e-3	–
17	71	2.40e-3	4.16	1.18e-2	3.32	3.75e-4	4.51
33	2	1.70e-4	3.99	1.25e-3	3.38	8.77e-7	9.13
65	2	9.66e-6	4.23	1.17e-4	3.49	0	–
129	2	5.03e-7	4.26	1.02e-5	3.51	0	–

Table 2.5: Numerical results for Example 3 using $\mathbb{P}_3$ elements and Mesh 2.1c.

Remark 2.5.2. *In our implementation of the bound-preserving finite element method, to find an initial guess, we first solve the problem using CIP method. This solution is then used to initialise the iteration. If the Richardson iteration terminates after two steps, this indicates that the initial solution already lies within the prescribed bounds. Indeed, the first iteration corresponds to the CIP solution, while the second iteration verifies the stopping criterion. In particular, if the overshoot or undershoot error of the CIP solution is already below the prescribed tolerance, the iteration halts immediately after this verification step. This behaviour is especially pronounced when higher-order finite element spaces are used, as the CIP solution tends to provide a more accurate starting point. An illustration of this phenomenon can be found in Tables 2.3 and 2.5, which correspond to the use of $\mathbb{P}_2$ and $\mathbb{P}_3$ elements, respectively, where the Richardson iteration converges in only two steps. A similar behaviour is observed when refining the mesh, as shown in Tables 2.4 and 2.5, where very fine meshes ($N = 65, 129$) are used.*

Example 4 (A problem with two inner layers)**.** *In this example, as well as in the next one, the diffusion term in* (2.1) *is given by $\mathcal{D} = \epsilon \mathcal{I}$, where $\epsilon > 0$. We approximate the solution of* (2.1) *for the parameters $f = 0$, $\mu = 0$, $\epsilon = 10^{-5}$, and $\boldsymbol{\beta} = (-y, x)$. Homogeneous Neumann boundary conditions are applied along the outflow boundaries $x = 0$ and $y = 1$. On the inflow boundaries at $x = 1$ and $y = 0$, we impose*

discontinuous Dirichlet boundary conditions:

$$g(x,y) = \begin{cases} 0 & \text{if } x \leq \frac{1}{3} \text{ and } y = 0 \\ \frac{1}{2} & \text{if } x \in \left(\frac{1}{3}, \frac{2}{3}\right) \text{ and } y = 0, \\ 1 & \text{otherwise.} \end{cases} \tag{2.50}$$

The purpose of this numerical experiment is twofold. Firstly, we aim at evaluating the effectiveness of the bound-preserving method (BPM) (2.15) in eliminating over- and under-shoots in areas where the constraint is not enforced. In this example, we set $\kappa = 1$ throughout the domain. The solution has two internal layers: one where it varies rapidly from 0 to approximately 0.5, and a second where it does from 0.5 to 1. The BPM will control any undershoot at $u_h^+ = 0$ around these layers; however, there is no explicit control in the region where the solution is approximately 0.5. So, it shows the method's ability to suppress potential overshoots in this region, even though the nonlinear stabilisation is inactive. Secondly, this experiment aims to provide numerical evidence that incorporating the continuous interior penalty (CIP) stabilisation term in the finite element method improves its overall performance.

We begin by addressing the second objective of this example. In Tables 2.6 and 2.7, we present the number of iterations required by the nonlinear solver to reach convergence. For all simulations, we set a maximum iteration of $3,000$. If this limit is reached, the solver halts, and we indicate "NC" to denote non-convergence. As observed in Tables 2.6 and 2.7, the absence of linear CIP stabilisation in the formulation results in a higher likelihood of non-convergence during the nonlinear iteration process. This provides further justification for incorporating linear stabilisation into the method (2.15).

In this experiment, we have chosen to use the stabilising term (2.9), with the stabilisation parameter $\gamma_\beta = 0.05$, as it produced the most favorable numerical results, particularly in terms of the sharpness of the interior layers.

We conducted the experiments using the meshes shown in Figures 2.1a and 2.1c for various values of N. In these experiments, for the $\mathbb{P}_1$, $\mathbb{Q}_1$, and $\mathbb{Q}_2$ elements, we set $\omega = 0.1$ in (2.48) when using the BPM, and reduced it to $\omega = 0.05$ when the CIP stabilisation was removed (i.e., setting $\gamma_\beta = 0$). It is important to note that decreasing the value of ω enhances the likelihood of convergence for the iterative solver.

Next, we examine the sharpness of the approximation in the interior layers. Figures 2.2-2.4 display the approximate solutions with the meshes from Figures 2.1a and 2.1c. We observe an absence of significant oscillations near the layers, even when using the non-Delaunay mesh in Figure 2.1c.

For comparison, we have also solved the same problem using the linear CIP method (with the stabilisation

	N	5	9	17	33	65	129
Mesh 2.1a	$\gamma_\beta = 0.05$	82	96	122	124	113	98
	$\gamma_\beta = 0$	228	1702	NC	NC	NC	NC
Mesh 2.1c	$\gamma_\beta = 0.05$	140	148	174	137	123	111
	$\gamma_\beta = 0$	126	NC	NC	NC	NC	NC

Table 2.6: Number of iterations for the fixed point linearisation (2.48) needed to reach convergence using $\mathbb{P}_1$ elements and the meshes given in Figures 2.1a and 2.1c.

	N	5	9	17	33	65	129
$\mathbb{Q}_1$	$\gamma_\beta = 0.05$	72	128	136	151	159	190
	$\gamma_\beta = 0$	1901	NC	NC	NC	NC	NC
$\mathbb{Q}_2$	$\gamma_\beta = 0.05$	283	243	360	315	339	258
	$\gamma_\beta = 0$	2034	NC	NC	NC	NC	NC

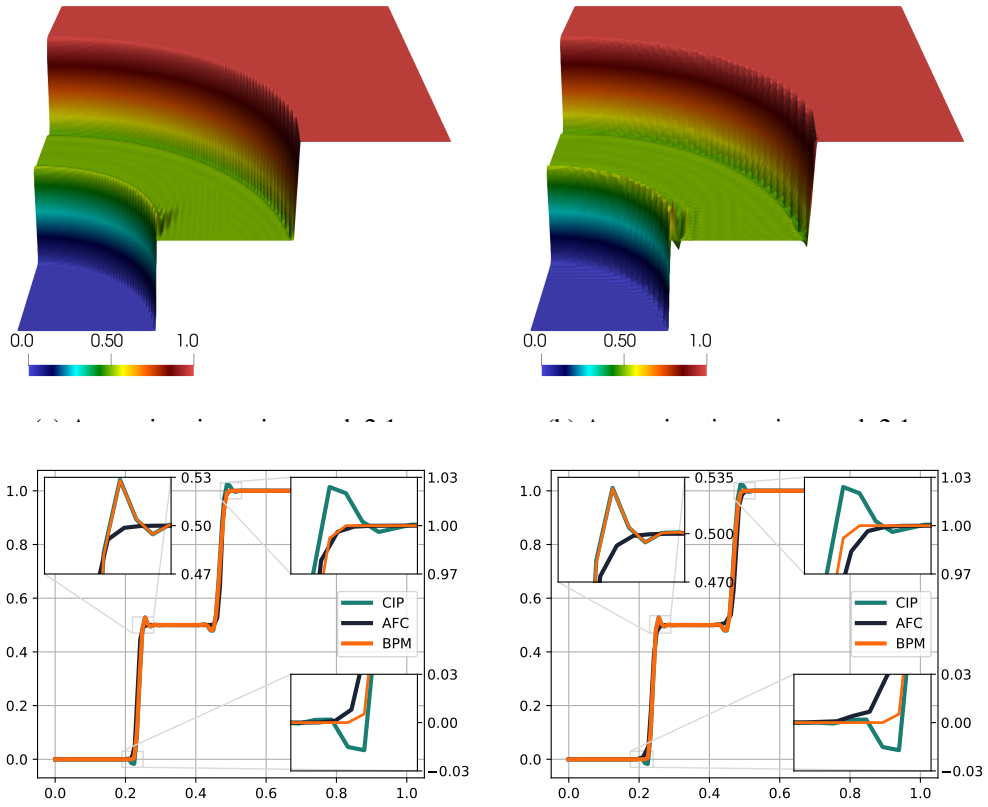
Table 2.7: Number of iterations for the fixed point linearisation (2.48) needed to reach convergence using $\mathbb{Q}_1$ and $\mathbb{Q}_2$ elements and the mesh given in Figure 2.1d.

term (2.9) and $\gamma_\beta = 0.05$) and the Algebraic Flux Correction (AFC) scheme, as described in [11] (the AFC method was applied only to $\mathbb{P}_1$ elements). The AFC method is known to preserve the discrete maximum principle, at least for Delaunay meshes in two dimensions, and thus we expect the results to respect the bounds, particularly on the mesh from Figure 2.1a. For the AFC scheme, we used the parameters $p = 8$ and $\gamma_0 = 0.75$ (see [11] for details on the method's formulation).

To compare these methods, we performed a cross-sectional analysis along the line $y = x$. Our analysis focused on two main aspects: the suppression of over and undershoots in the numerical solution, and the sharpness of the interior layers. Zoomed views of the cross-sections at the layer transitions are presented. As expected, the CIP method exhibits both over and undershoots, while the BPM and AFC methods do not. In fact, the function u_h^+ shows much smaller oscillations than the CIP method and showing a level of sharpness in the layers comparable to the AFC method.

Remark 2.5.3. *An advantage of this bound-preserving finite element method is its relatively low computational cost compared with AFC schemes. In our method, the bounds are imposed directly at the degrees of freedom, which requires only local corrections and hence adds little costs to the basic finite element solve. By contrast, AFC methods (see Section 1.6.3) involve the computation of nonlinear flux limiters, which can be significantly more expensive due to the need to evaluate limiter functions on element interfaces and to update the discrete fluxes accordingly. As a result, the present method is competitive in terms of CPU time, particularly in large-scale or higher-order computations.*

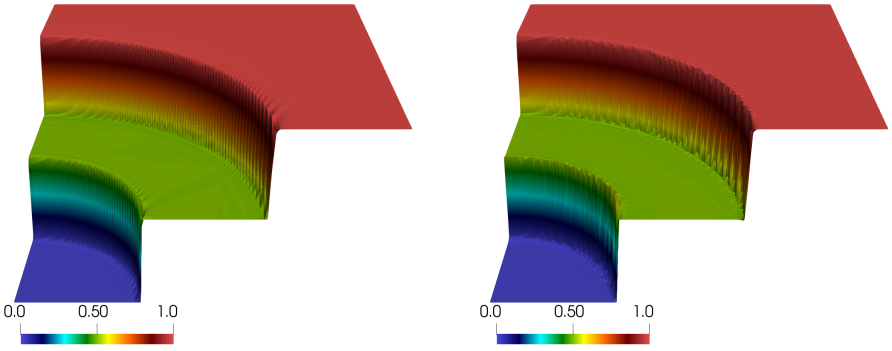
The same comparison was performed using the non-Delaunay mesh from Figure 2.1c, and we arrived at similar conclusions. Interestingly, when applying the BPM method to $\mathbb{P}_2$, $\mathbb{Q}_2$, and higher-order elements, although the bounds are enforced only at the nodes, no significant undershoots (which indicates the violations of the physical bounds) were observed in the numerical solution.



(c) Cross-section using mesh 2.1a (d) Cross-section using mesh 2.1c

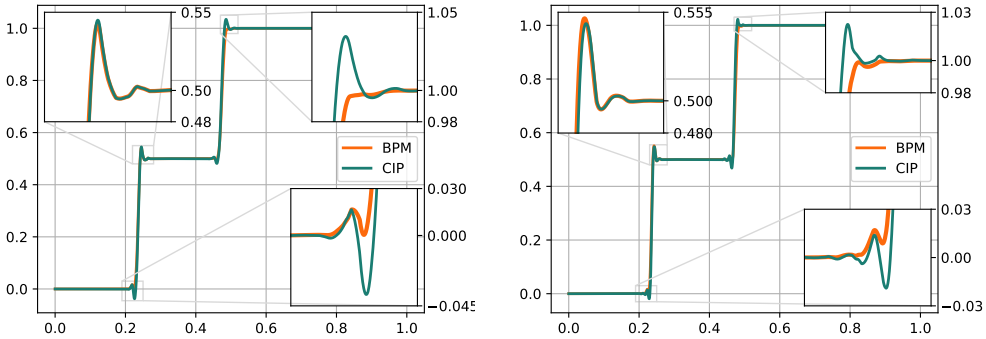
Figure 2.2: The approximation of the solution of Example 4 by the bound preserving method (BPM), using $\mathbb{P}_1$ elements and the meshes given in Figures 2.1a and 2.1c with $N = 129$. Cross-sections taken about $y = x$ plane of the solution of the BPM, CIP and AFC. For AFC $p = 8$ and for BPM and CIP the penalty (2.9) $\gamma_\beta = 0.05$ and $\omega = 0.1$ has been used.

Example 5 (A problem with an inner and a boundary layer)**.** *In this final example, we consider the problem with* $f = 0$, $\mu = 0$, $\epsilon = 10^{-5}$, *and* $\boldsymbol{\beta} = \left(\cos\left(\frac{\pi}{3}\right), \sin\left(\frac{\pi}{3}\right)\right)^T$. *The Dirichlet boundary condition* $u = g$ *is*



(a) Approximation using $\mathbb{P}_2$ elements and mesh 2.1c

(b) Approximation using $\mathbb{P}_3$ elements and mesh 2.1c



(c) Cross-section using $\mathbb{P}_2$ elements and mesh 2.1c

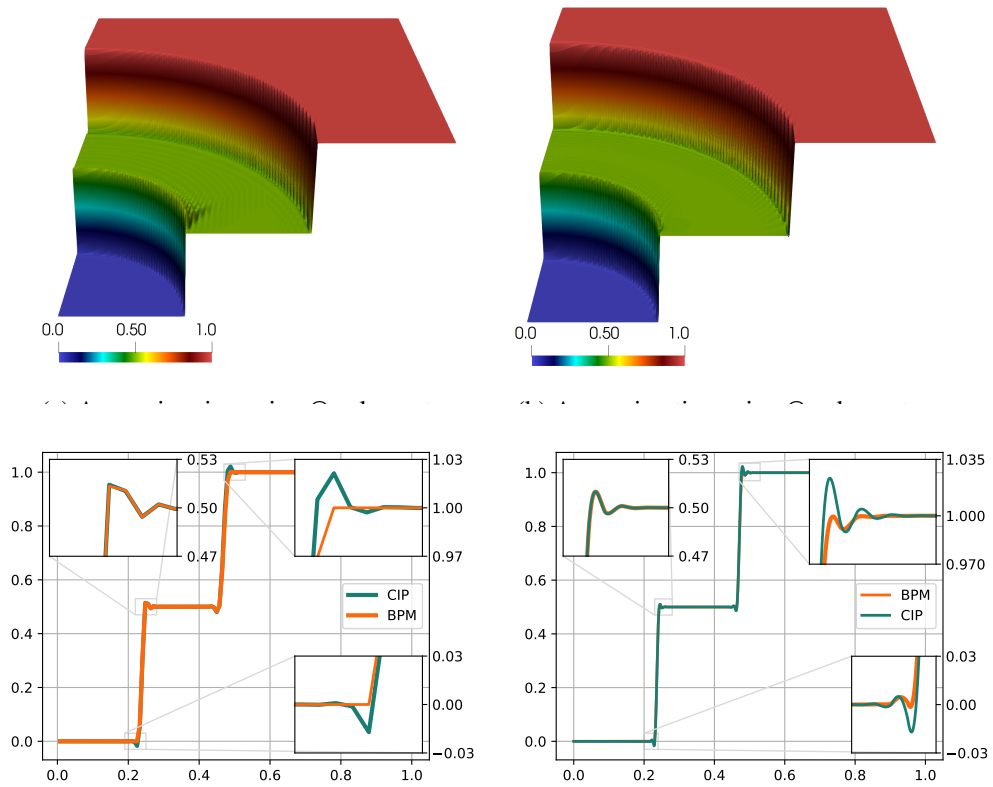
(d) Cross-section using $\mathbb{P}_3$ elements and mesh 2.1c

Figure 2.3: The approximation of the solution of Example 4 by the bound preserving method (BPM), using $\mathbb{P}_2$ and $\mathbb{P}_3$ elements and the meshes given in Figure 2.1c with $N = 129$. Cross-sections taken along the line $y = x$. For both methods the penalty (2.9) with $\gamma_\beta = 0.05$ was used ($\omega = 0.05$). For plotting these cross-sections, 10,000 equidistant points were chosen along the line $y = x$, and the values of the approximated solution have been plotted at these points.

imposed on Γ*, where* g *is defined as follows*

$$g(x, y) = \begin{cases} 1 & \text{if } x = 0 \text{ or } y = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{2.51}$$

This problem consists of propagating a discontinuous boundary condition into the interior, which creates an interior layer that intersects with a boundary layer at $y = 1$*. The solution was approximated using the meshes shown in Figures 2.1a–2.1d. In this experiment, particularly when approximating the outflow layer, the best results were achieved using the method with the CIP stabilising term* (2.8) *and* $\gamma = 0.01$*. Therefore,*



(c) Cross-section taken using $\mathbb{Q}_1$ elements. (d) Cross-section taken using $\mathbb{Q}_2$ elements.

Figure 2.4: The approximation of the solution of Example 4 by the bound preserving method (BPM), using $\mathbb{Q}_1$ and $\mathbb{Q}_2$ elements and the mesh given in Figure 2.1d with $N = 129$. Cross-sections of the discrete solution of the BPM and CIP methods taken about the line $y = x$. For BPM and CIP the penalty (2.9) $\gamma_\beta = 0.05$ was used ($\omega = 0.1$). For plotting the cross-sections with $\mathbb{Q}_2$ elements, 10,000 equidistant points were chosen along the line $y = x$, and the values of the approximated solution have been plotted at these points.

we only report the results obtained for this choice.

For the iterative method (2.48) we use $\omega = 0.1$, and we now report the number of fixed-point iterations needed to convergence:

We now validate the statement made in Remark 2.3.6 by illustrating the behavior of u_h^- using $\mathbb{P}_1$ elements. In Figure 2.5, we provide a zoomed in view near the boundary, showing the cross-section of u_h^- along the line $x = 0.9$ for various mesh refinement levels. As the mesh is refined, we observe that the magnitude of u_h^- decreases gradually, and its support becomes increasingly localised, thereby confirming the assertion in Remark 2.3.6.

Furthermore, Figures 2.6-2.9 present the approximate solutions obtained with the BPM, AFC, and CIP

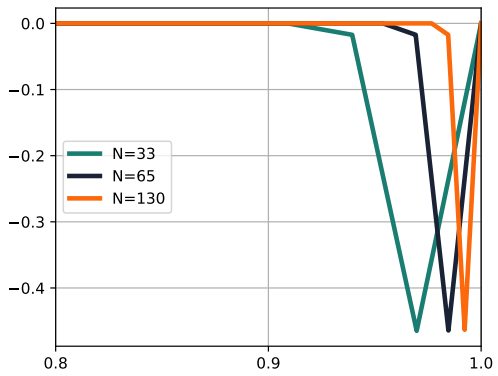
	N	5	9	17	33	65	129
Mesh 2.1a	Itr.	109	143	177	212	249	249
Mesh 2.1c	Itr.	123	152	186	218	245	240

Table 2.8: Iterations needed to reach convergence using $\mathbb{P}_1$ elements and the meshes given in Figures 2.1a–2.1c, and the penalty term (2.8) with $\gamma = 0.01$ ($\omega = 0.1$).

	N	5	9	17	33	65	129
$\mathbb{Q}_1$	Itr.	156	226	225	308	310	322
$\mathbb{Q}_2$	Itr.	375	299	291	270	236	217

Table 2.9: Iterations needed to reach convergence using $\mathbb{Q}_1$ and $\mathbb{Q}_2$ elements and the mesh given in Figure 2.1d, and the penalty term (2.9) with $\gamma_\beta = 0.01$ ($\omega = 0.1$).

methods for this problem. Additionally, we provide cross-sections illustrating the structure of the interior layer, along with a cross-section of u_h^+ along the line $y = 1 - x$ for Mesh 2.1a with $N = 129$. Once again, comparisons are made with the linear CIP method (using the stabilising term (2.8) and $\gamma = 0.01$) and the AFC scheme as discussed in the previous example. The results, displayed in Figures 2.6-2.9, show that the current method successfully eliminates the oscillations present in the CIP solution while providing similar sharpness in the layers as the AFC method.



(a) Cross-section taken of u_h^- along $x = 0.9$.

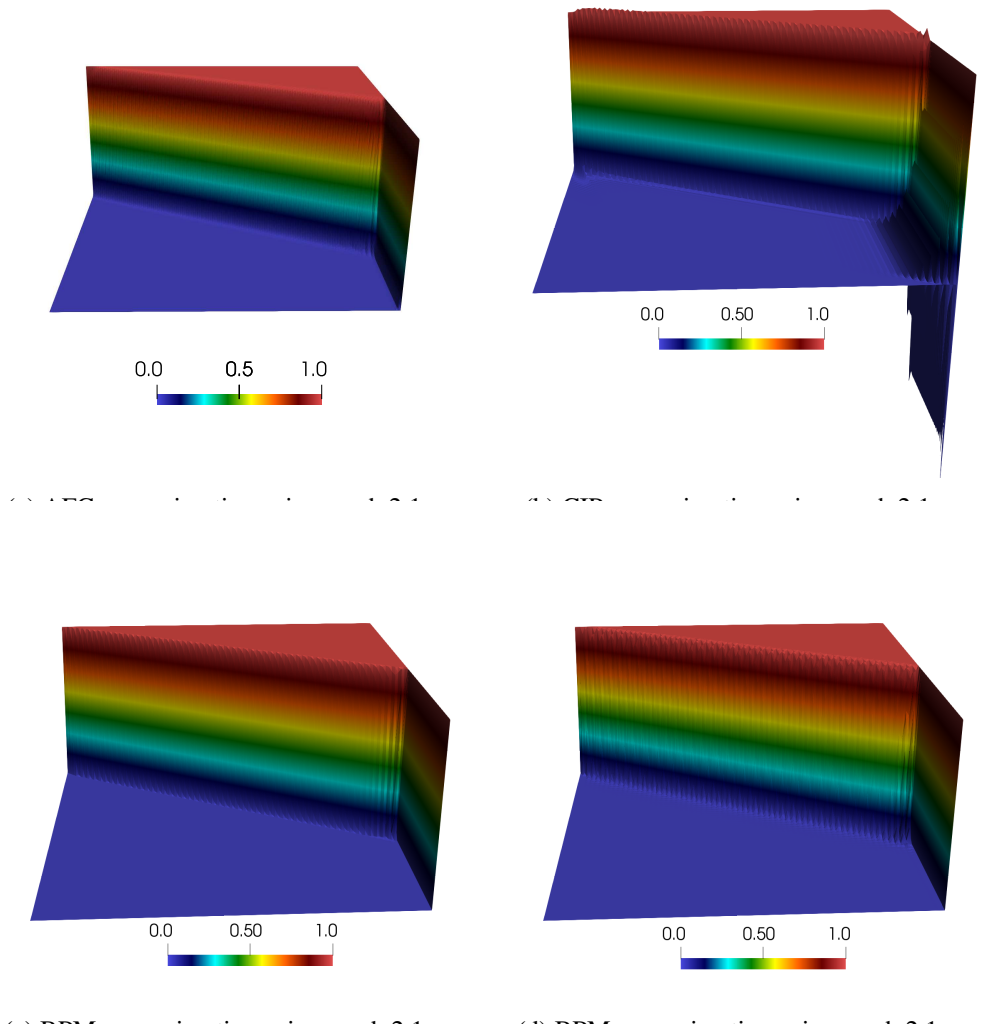
Figure 2.5: Cross-sections of u_h^- for Example 5 illustrating the behaviour at the boundary layers using $\mathbb{P}_1$ elements and the mesh given in Figures 2.1a.

Remark 2.5.4. *In this chapter, we used a simple Richardson type solver to highlight the simplicity of the scheme, but more efficient nonlinear solvers, such as localised Newton methods (see, e.g., [6]), or active set methods [7], can significantly improve convergence speed. Our preliminary results indicate that these alter-*

natives provide much faster convergence. For example in Table 2.10 we present the number of the Newton iterations to reach the convergence (one can see the implementation of the Newton's iteration method in Section 4.7). Obviously Newton's method reduced the number of iterations that is required for the convergence.

	N	5	9	17	33	65	129
Mesh 2.1a	Itr.	48	27	56	72	64	76

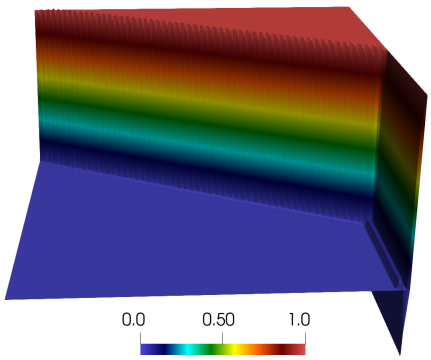
Table 2.10: Newton iterations needed to reach convergence using $\mathbb{P}_1$ elements and the mesh given in Figure 2.1a, and the penalty term (2.8) with $\gamma = 0.01$.

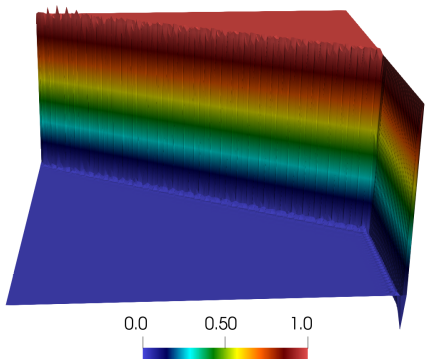


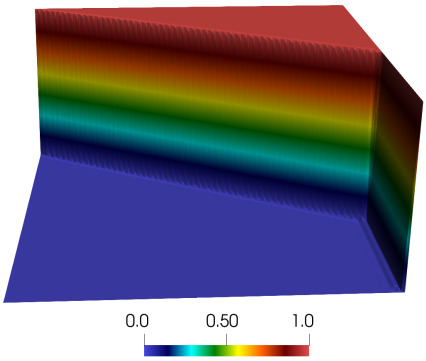
(e) Cross-section using mesh 2.1a

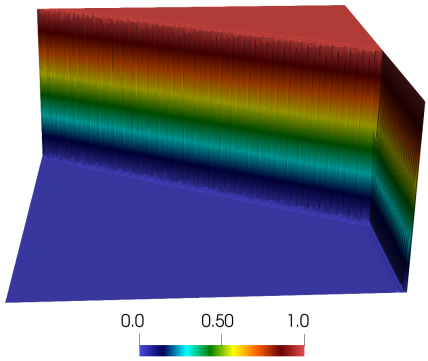
(f) Cross-section using mesh 2.1c

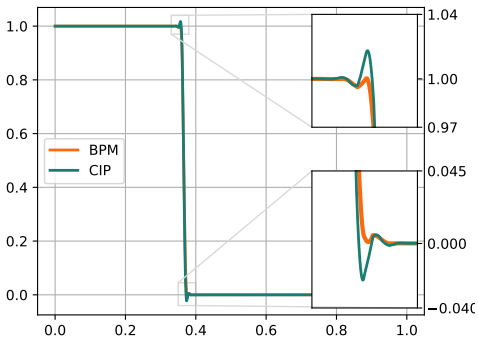
Figure 2.6: The approximation of the solution of Example 5 by the bound preserving method (BPM), using $\mathbb{P}_1$ elements and the meshes given in Figures 2.1a and 2.1c with $N = 129$. Cross-sections of the discrete solution of the BPM, CIP, and AFC methods taken about the line $y = x$. For AFC $p = 8$ and for BPM and CIP the penalty (2.9) $\gamma_\beta = 0.01$ was used ($\omega = 0.1$). For plotting the cross-sections we used linear interpolation between the nodes.



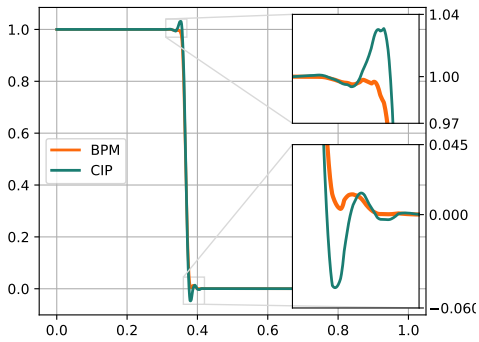






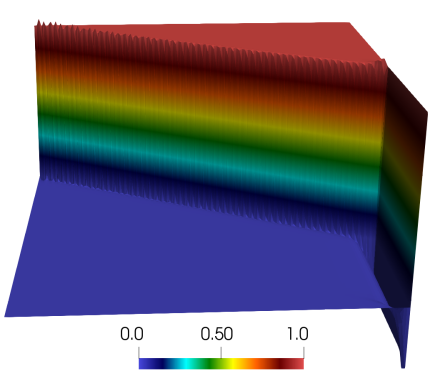


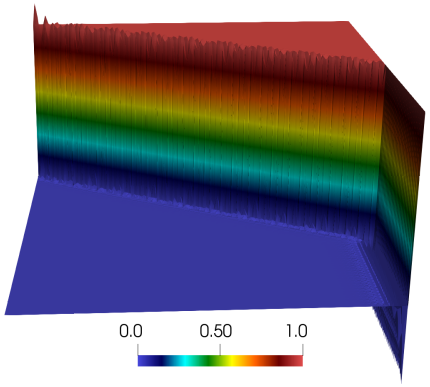
(e) Cross-section using mesh 2.1a

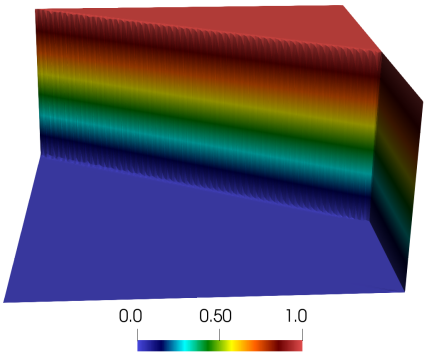


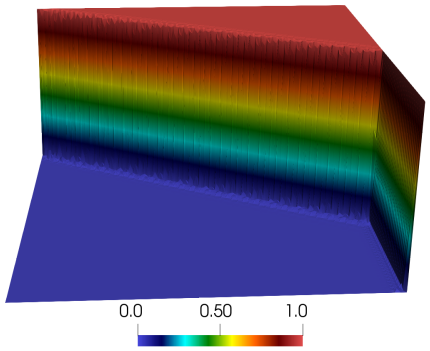
(f) Cross-section using mesh 2.1c

Figure 2.7: The approximation of the solution of Example 5 by the bound preserving method (BPM), using $\mathbb{P}_2$ elements and the meshes given in Figures 2.1a and 2.1c with $N = 129$. Cross-sections around the line $y = x$ of the solution of the BPM and CIP methods. For both methods the penalty (2.9) with $\gamma_\beta = 0.01$ was used ($\omega = 0.1$). For plotting these cross-sections, 10,000 equidistant points were chosen along the line $y = x$, and the values of the approximated solution have been plotted at these points.

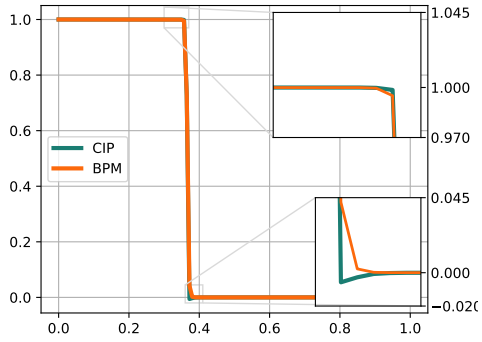


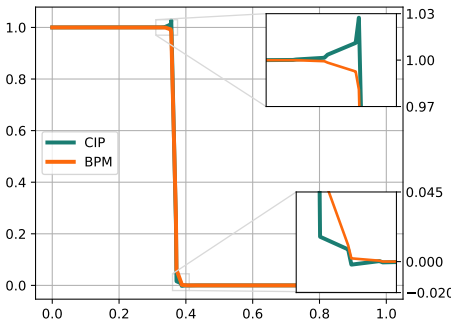






(d) BPM approximation using mesh 2.1c

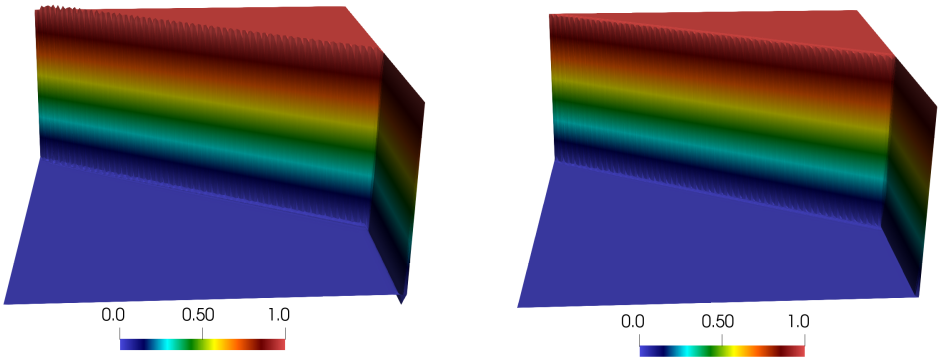




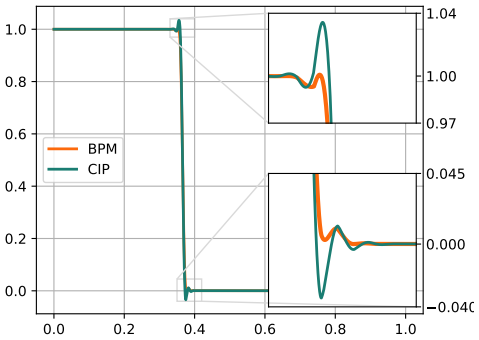
(e) Cross-section using mesh 2.1a

(f) Cross-section using mesh 2.1c

Figure 2.8: The approximation of the solution of Example 4 by the bound preserving method (BPM), using $\mathbb{P}_3$ elements and the meshes given in Figures 2.1a and 2.1c with $N = 129$. Cross-sections of the solution of the BPM and CIP taken about the line $y = x$. For BPM and CIP the penalty (2.9) $\gamma_\beta = 0.01$ was used ($\omega = 0.1$). For plotting the cross-sections we used linear interpolation between the degree of freedoms.



(a) CIP approximation using mesh 2.1d. (b) BPM approximation using mesh 2.1d.



(c) Cross-section taken using mesh 2.1d.

Figure 2.9: The approximation of the solution of Example 5 by the bound preserving method (BPM), using $\mathbb{Q}_2$ elements and the mesh given in Figure 2.1d with $N = 129$. Cross-sections of the solution of the BPM and CIP taken about the line $y = x$. For BPM and CIP the penalty (2.9) $\gamma_\beta = 0.01$ was used ($\omega = 0.1$).

Chapter 3

A nodally bound-preserving finite element method for time-dependent convection-diffusion equations

3.1 Introduction

The numerical simulation of time-dependent convection-diffusion equations is crucial in various applications, particularly in modeling transport-controlled reaction rates in shear flows and chemical reactions in flow fields. Such models typically involve solving a system of nonlinear time-dependent convection-diffusion equations describing the concentrations of reactants and products. An inaccuracy in one equation of this system can propagate and significantly affect all concentrations.

In the previous chapter, we studied the stationary convection–diffusion equation, where the convection field dominates diffusion by several orders of magnitude, this makes the numerical solution unstable and can create unwanted oscillations.

In the time-dependent convection–diffusion problem, the same difficulties appear. To reduce these oscillations, one can add stabilisation terms. Two common approaches are the Streamline Upwind Petrov–Galerkin (SUPG) method [28] and the Galerkin Least-Squares (GLS) method [76]. Both methods are residual-based stabilisations that add extra diffusion terms proportional to the residual of the equation.

In the time-dependent context, the SUPG stabilisation necessitates including the time derivative of the solution in the stabilisation term, which introduces an artificial coupling between the time step and the stabilisation parameter. Standard stability analyses suggest that improper balancing of the mesh size and time

step could lead to a loss of stability. However, some numerical evidence contradicts this claim (see [23]). Consequently, alternative stabilisation strategies using symmetric, non-residual-based approaches have been extensively studied. Examples of such methods include the subgrid viscosity method [68], the orthogonal subscale method [36], and the continuous interior penalty (CIP) method [20]. A detailed discussion of symmetric stabilisation techniques and their advantages can be found in [34].

In this chapter, similar to the previous one, we reduce the oscillations generated by the convection term by adding a CIP term. In fact, after the space discretisation of the problem, we include a CIP stabilisation term, which may also depend on time.

Similar to the previous chapters and Section 1.6.5, the strategy is directly integrates bounds into the finite element formulation. As explained in chapters 1 and 2, this approach is based on the key observation that imposing bounds on the numerical solution is equivalent to solving the problem within a convex subset of the finite element space, consisting of discrete functions that satisfy the prescribed bounds at their degrees of freedom. In this chapter, we build on this methodology and extend the bound-preserving finite element methodology to time-dependent convection-diffusion equations. The core idea is as follows: for each time step, we define a set $V_{\mathcal{P}}^{+}$ of *admissible* finite element functions satisfying the global bounds at their degrees of freedom (e.g., nodal values for Lagrangian elements). We then introduce an algebraic projection onto this admissible set, denoted by u_h^{+}, and formulate a finite element problem for the projected variable. To eliminate the non-trivial kernel introduced by this projection and to prevent singularity in the discrete system, a stabilisation term is added at each time step. Numerical experiments indicate that including a linear stabilisation term, such as CIP stabilisation, leads to more robust results.

The structure of this chapter is as follows: Section 3.2 introduces the model problem and preliminary material necessary for the method. Section 3.3 presents the finite element formulation and proves its well-posedness. Stability and error analysis are discussed in Section 3.4, and Section 3.5 illustrates the performance of the method through numerical experiments.

3.2 General setting and the model problem

Let Ω be a bounded Lipschitz domain in $\mathbb{R}^d$ $(d = 2, 3)$ with a polyhedral boundary $\partial\Omega$, and let $T > 0$. Given a function $f \in L^2((0,T); L^2(\Omega))$, we consider the following convection-diffusion problem, which is

a particular case of problem (1.8):

$$\begin{cases} \partial_t u - \varepsilon\,\Delta u + \boldsymbol{\beta}\cdot\nabla u + \mu u = f & \text{in } (0,T]\times\Omega, \\ u(\boldsymbol{x},t) = 0 & \text{on } (0,T]\times\partial\Omega, \\ u(\cdot,0) = u^0 & \text{in } \Omega, \end{cases} \tag{3.1}$$

In this formulation, $\varepsilon \in \mathbb{R}^+$ denotes the diffusion coefficient, $\boldsymbol{\beta} = (\beta_i)_{i=1}^d \in L^\infty((0,T); W^{1,\infty}(\Omega))^d$ represents the convective field, and $\mu \in \mathbb{R}_0^+$ is the reaction coefficient. Furthermore, we assume that the convective field $\boldsymbol{\beta}$ satisfies $\operatorname{div}\boldsymbol{\beta} = 0$ in $\Omega\times[0,T]$.

The weak formulation of equation (3.1) can be written as: find $u \in L^\infty((0,T), H_0^1(\Omega)) \cap H^1((0,T), H^{-1}(\Omega))$ such that, for almost every $t \in (0,T)$, the following holds:

$$\begin{cases} (\partial_t u, v)_\Omega + a(u,v) = (f,v)_\Omega & \forall v \in H_0^1(\Omega), \\ u(\cdot,0) = u^0, \end{cases} \tag{3.2}$$

where the bilinear form $a(\cdot,\cdot)$ is defined as

$$a(w,v) := \varepsilon\,(\nabla w, \nabla v)_\Omega + (\boldsymbol{\beta}\cdot\nabla w, v)_\Omega + \mu(w,v)_\Omega \qquad \forall v \in H_0^1(\Omega), \quad t \in (0,T). \tag{3.3}$$

In the above definition, we have slightly abused the notation, as the convective term $\boldsymbol{\beta}$ may depend on t. However, unless the context requires it, we will denote this bilinear form by $a(\cdot,\cdot)$. Given that $\boldsymbol{\beta}$ is assumed to be solenoidal, for each $t \in (0,T)$, the bilinear form $a(\cdot,\cdot)$ induces the following "energy" norm in $H_0^1(\Omega)$

$$\|v\|_a = \sqrt{a(v,v)} \quad t \in [0,T].$$

The well-posedness of (3.2) is a classical result. In fact, it follows directly from the Lions–Magenes Theorem (see, e.g., [98]), which guarantees existence, uniqueness, and continuous dependence of the solution.

Furthermore, based on the maximum principle for parabolic partial differential equations Theorem 1.5.5, the solution to (3.2) attains its extrema on $\big[\Omega\times\{0\}\big] \cup \big[\partial\Omega\times(0,T]\big]$. Motivated by this observation, we use the following assumption regarding on u, solution of (3.2).

Assumption (C1): We assume that the weak solution of (3.2) satisfies the following condition:

$$0 \le u(\boldsymbol{x},t) \le \kappa(t) \quad \text{for almost all } (\boldsymbol{x},t) \in \Omega\times[0,T], \tag{3.4}$$

where $\kappa(t)$ is a known positive function that depends on t.

In this assumption, the lower bound in (3.4) does not necessarily need to be zero; however, we set it to zero here for simplicity. Moreover, the results discussed in this work remain valid if $\kappa(t)$ is replaced by a positive function $\kappa(\boldsymbol{x}, t)$.

3.2.1 Space discretisation and a stabilisation Galerkin method

To discretise problem (3.2), we use a conforming, shape-regular, and quasi-uniform partition of the domain Ω into closed simplices or affine quadrilateral/hexahedral elements, denoted by $\mathcal{P}$, as defined in (1.15). Also, we use the notations established in Section (1.6).

The standard Galerkin semi-discretisation of (3.2) using the finite element space (1.15) reads:

$$\begin{cases} \text{For almost all } t \in (0,T), \text{ find } u_h \in V_{\mathcal{P}} \text{ such that} \\ (\partial_t u_h, v_h)_\Omega + a(u_h, v_h) = (f, v_h)_\Omega \qquad \forall v_h \in V_{\mathcal{P}}, \\ u_h(\cdot, 0) = i_h u^0. \end{cases} \tag{3.5}$$

Similar to the Galerkin finite element method for steady-state problems, it is well-known that applying the standard Galerkin finite element method to (3.2) in the convection-dominated regime results in discrete solutions that are affected by global spurious oscillations. Consequently, these solutions often fail to satisfy Assumption (C1) (see, e.g., [111] for a comprehensive overview).

To enhance stability, a common approach is add a linear stabilising term to suppress the oscillations induced by the dominant convection. Several stabilisation techniques exist, with those based on adding symmetric semi-positive-definite terms being particularly popular for time-dependent problems.

In this work, we use the continuous interior penalty (CIP) method, initially proposed in [36] and thoroughly analysed for time-dependent problems in [34]. The CIP method incorporates the following stabilising term into the Galerkin scheme (3.5)

$$J(u_h, v_h) = \gamma \sum_{F \in \mathcal{F}_I} \int_F \|\boldsymbol{\beta}\|_{0,\infty,F} h_F^2 [\![\boldsymbol{\nabla} u_h]\!] \cdot [\![\boldsymbol{\nabla} v_h]\!] \, \mathrm{d}s \,, \tag{3.6}$$

where $\gamma \geq 0$ is a non-dimensional constant, and thus it reads as follows:

$$\begin{cases} \text{For almost all } t \in (0,T), \text{ find } u_h \in V_{\mathcal{P}} \text{ such that} \\ (\partial_t u_h, v_h)_\Omega + a_J(u_h, v_h) = (f, v_h)_\Omega \qquad \forall v_h \in V_{\mathcal{P}}, \\ u_h(\cdot, 0) = i_h u^0. \end{cases} \tag{3.7}$$

where

$$a_J(u_h, v_h) := a(u_h, v_h) + J(u_h, v_h) \qquad \forall v_h \in V_{\mathcal{P}}\,. \tag{3.8}$$

It is important to note that, although we have opted to use CIP stabilisation in this work, the results presented here remain valid if any symmetric stabilisation is applied to the convective term. In particular, the results presented herein hold for any of the stabilised methods analysed in [34].

While the addition of (3.8) helps remove spurious oscillations and ensures a stable solution, the resulting discrete solution does not preserve the physical bounds given by (3.4). In the next section, we present the key components for constructing a finite element method that enforces the bound (3.4) on its solution.

3.2.2 The admissible set

Assumption (C1) is analogous to Assumption (A1), with the distinction that (3.4) applies to time-dependent problems. Assumption (C1) leads to the introduction of the following *admissible set*, which consists of finite element functions that satisfy the bound (3.4) at their degrees of freedom

$$V_{\mathcal{P}}^+ := \{ v_h(\boldsymbol{x}_i) \in V_{\mathcal{P}} \,:\, v_h \in [0, \kappa(t)] \;\text{ for all }\; i = 1, \dots, N \}. \tag{3.9}$$

Every element $v_h \in V_{\mathcal{P}}$ can be decomposed into the sum $v_h = v_h^+ + v_h^-$, where v_h^+ and v_h^- are defined as follows

$$v_h^+ = \sum_{i=1}^{M} \max\left\{0, \min\{v_h(\boldsymbol{x}_i), \kappa(t)\}\right\} \phi_i \quad \text{ for } \; t \in (0,T], \tag{3.10}$$

and

$$v_h^- = v_h - v_h^+. \tag{3.11}$$

We refer to v_h^+ and v_h^- as the *constrained* and *complementary* parts of v_h, respectively. Using this decomposition, we define the following algebraic projection

$$(\cdot)^+ : V_{\mathcal{P}} \to V_{\mathcal{P}}^+ \quad , \quad v_h \to v_h^+ . \tag{3.12}$$

Remark 3.2.1. *Strictly speaking* $(\cdot)^+$ *should be denoted by* $(\cdot)^{+,t}$*, as* $\kappa(t)$ *depends on* t*. To lighten the notation, we will simply use* $(\cdot)^+$ *unless it is necessary to specify the time.*

The following result, which has a proof identical to that of (2.2.4), but adapted for the time-dependent $(\cdot)^+$ which has been defined in (3.12), will be useful in the analysis presented below.

Lemma 3.2.2. *[5, Lemma 2.2] Let the operator* $(\cdot)^+$ *be defined as in* (3.12)*. There exists a constant* $C > 0$*, independent of* h*, such that for all* $t \in [0, T]$*, the following inequalities hold*

$$\|w_h^+ - v_h^+\|_{0,\Omega} \leq C\|w_h - v_h\|_{0,\Omega}, \tag{3.13}$$

$$\|v_h^+\|_{0,\Omega} \leq C\kappa(t), \tag{3.14}$$

for all $w_h, v_h \in V_{\mathcal{P}}$*.*

3.3 The finite element method

In this section, we propose a time-space discretisation of (3.7). Let $N > 0$ be a given positive integer. We partition the time interval $[0, T]$ as $t_0 = 0 < t_1 < t_2 < \cdots < t_N = T$, with the time step size defined as $\Delta t_n := t_n - t_{n-1}$. For simplicity, we assume a uniform time step size, i.e., $\Delta t_n = \Delta t = \frac{T}{N}$. The discrete value $u_h^n \in V_{\mathcal{P}}$ represents the approximation of $u^n = u(t_n)$ in $V_{\mathcal{P}}$ for $0 \leq n \leq N$. We define

$$\delta u_h^n := \frac{u_h^n - u_h^{n-1}}{\Delta t}, \qquad t_{n-1+\theta} = \theta t_n + (1-\theta) t_{n-1} \quad ,$$
$$u_h^{n-1+\theta} := \theta u_h^n + (1-\theta) u_h^{n-1}, \qquad f^{n-1+\theta} := \theta f^n + (1-\theta) f^{n-1}$$

.

With these notations, the finite element method used in this work reads as follows:

$$\begin{cases} \text{For } 1 \le n \le N, \text{ find } u_h^n \in V_{\mathcal{P}} \text{ such that} \\ (\delta(u_h^n)^+, v_h)_\Omega + a_h(u_h^{n-1+\theta}; v_h) = (f^{n-1+\theta}, v_h)_\Omega \qquad \forall v_h \in V_{\mathcal{P}}, \\ u_h^0 = i_h u^0. \end{cases} \tag{3.15}$$

Here, $a_h(\cdot;\cdot)$ is defined as

$$a_h(u_h^{n-1+\theta}; v_h) := \theta a_J((u_h^n)^+, v_h) + (1-\theta) a_J((u_h^{n-1})^+, v_h) + s((u_h^n)^-, v_h), \tag{3.16}$$

where the stabilisation term $s(\cdot,\cdot)$ is defined as

$$s(v_h, w_h) := \alpha \sum_{i=1}^{M} \left(\varepsilon \mathfrak{h}(\boldsymbol{x}_i)^{d-2} + \|\boldsymbol{\beta}(\boldsymbol{x}, t_n)\|_{0,\infty,\omega_i} \mathfrak{h}(\boldsymbol{x}_i)^{d-1} + \left(\frac{1}{\Delta t} + \mu\right) \mathfrak{h}(\boldsymbol{x}_i)^d \right) v_h(\boldsymbol{x}_i) w_h(\boldsymbol{x}_i). \tag{3.17}$$

Setting $\tilde{\varepsilon} = \Delta t \theta \varepsilon$, $\tilde{\boldsymbol{\beta}} = \Delta t \theta \boldsymbol{\beta}$, $\tilde{\mu} = (\mu \Delta t \theta + 1)$ and $\tilde{J}(\cdot,\cdot) = \Delta t \theta J(\cdot,\cdot)$, we can define the following norm at each time step t^n:

$$\|v_h\|_{h,\theta\Delta t} := \left(\tilde{\varepsilon} \|\nabla v_h\|_{0,\Omega}^2 + \tilde{\mu} \|v_h\|_{0,\Omega}^2 + \tilde{J}(v_h, v_h) \right)^{\frac{1}{2}}. \tag{3.18}$$

In addition, the stabilising form $s(\cdot,\cdot)$ for $0 \le n \le N$ induces the following norm on $V_{\mathcal{P}}$

$$\|v_h\|_s := \sqrt{s(v_h, v_h)}. \tag{3.19}$$

Remark 3.3.1. *This stabilisation term* (3.17) *differs from the stabilisation term* (2.17) *used in the previous chapter. The inclusion of the factor* $\frac{1}{\Delta t}$ *in the stabilisation was primarily motivated by the performance of the nonlinear solver. Without this factor, the nonlinear solver exhibited significantly slower convergence. Additionally, as demonstrated below, the factor* $\frac{1}{\Delta t}$ *ensures the dominance of the* $\|\cdot\|_s$ *norm over the* $\|v_h\|_{h,\theta\Delta t}$ *norm.*

The following result is a direct consequence of Lemma (1.6.21), and the proof is very similar, so we omit it.

Lemma 3.3.2. *[5, Lemma 3.1] There exists a constant* $C_{\text{equiv}} > 0$*, depending only on the shape regularity*

of $\mathcal{P}$, such that

$$\|v_h\|_{h,\theta\Delta t}^2 \leq \Delta t \frac{C_{\text{equiv}}}{\alpha} \|v_h\|_s^2 \qquad \forall v_h \in V_{\mathcal{P}}\,. \tag{3.20}$$

At each time-level $0 \leq n \leq N$, the finite element method (3.15) is a particular case of the finite element method proposed in chapter 2 (see also [4]). In fact, at each time step $1 \leq n \leq N$, (3.15) can be written as

$$((u_h^n)^+, v_h)_\Omega + \Delta t\theta a_J((u_h^n)^+, v_h) + \Delta t s((u_h^n)^-, v_h) = F^n(v_h) \qquad \forall v_h \in V_{\mathcal{P}}, \tag{3.21}$$

where $a_J(\cdot,\cdot)$ is defined in (3.8), and

$$\begin{aligned} F^n(v_h) := {} & \Delta t(f^{n-1+\theta}, v_h)_\Omega - \Delta t(1-\theta)\varepsilon\left(\nabla(u_h^{n-1})^+, \nabla v_h\right)_\Omega - \Delta t(1-\theta)(\boldsymbol{\beta}\cdot\nabla\left(u_h^{n-1})^+, v_h\right)_\Omega \\ & - (\mu\Delta t(1-\theta)-1)\left((u_h^{n-1})^+, w_h\right)_\Omega - \Delta t(1-\theta)J((u_h^{n-1})^+, v_h) \qquad \forall v_h \in V_{\mathcal{P}}. \end{aligned} \tag{3.22}$$

The realisation that at each time step the method (3.15) is related to the method proposed in [4] will be instrumental in the well-posedness result presented in the next section.

3.3.1 Well-posedness

In this section, we analyse the well-posedness of (3.15). The first step is given by the following monotonicity result, whose proof is analogous to that of Lemma 1.6.22, but adapted for (3.17).

Lemma 3.3.3. *[5, Lemma 3.3] The bilinear form $s(\cdot,\cdot)$ defined in* (3.17) *satisfies the following inequalities:*

$$s(v_h^- - w_h^-, v_h^+ - w_h^+) \geq 0\,, \tag{3.23}$$

$$s(v_h^-, w_h^+ - v_h^+) \leq 0\,, \tag{3.24}$$

for every $v_h, w_h \in V_{\mathcal{P}}$.

We now address the well-posedness of (3.15). To this end, we leverage the connection highlighted at the end of the previous section. Specifically, we employ a similar approach to the one used in Theorem 2.3.3 to show that for each $n = 1, \ldots, N$, the problem (3.21) has a unique solution. This, in turn, implies that the problem (3.15) is well-posed.

Theorem 3.3.4. *[5, Theorem 3.4] Let $n = 1, \ldots, N$, then,*

a. There exists $u_h^n \in V_{\mathcal{P}}$ that solves (3.21).

b. $(u_h^n)^+ \in V_{\mathcal{P}}^+$ *satisfies*

$$((u_h^n)^+, v_h - (u_h^n)^+)_\Omega + \Delta t\theta a_J((u_h^n)^+, v_h - (u_h^n)^+) \geq F^n(v_h - (u_h^n)^+) \qquad \forall v_h \in V_{\mathcal{P}}^+. \tag{3.25}$$

c. $(u_h^n)^-$ *is the unique solution of*

$$\Delta ts((u_h^n)^-, v_h) = F^n(v_h) - ((u_h^n)^+, v_h)_\Omega + \Delta t\theta a_J((u_h^n)^+, v_h) \qquad \forall v_h \in V_{\mathcal{P}}. \tag{3.26}$$

d. The solution of (3.21) *is unique.*

Proof. a. We begin by defining the following bilinear form

$$B(v_h, w_h) := \Delta t\theta\varepsilon \left(\nabla v_h, \nabla w_h\right)_\Omega + (\mu\Delta t\theta + 1)(v_h, w_h)_\Omega + \Delta t\theta J(v_h, w_h) \quad \forall v_h, w_h \in V_{\mathcal{P}},$$

and the mapping

$$\begin{aligned} T &: V_{\mathcal{P}} \longrightarrow V_{\mathcal{P}}, \\ \hat{u}_h^n &\longrightarrow u_h^n = T(\hat{u}_h^n), \quad n = 1, \cdots, N \end{aligned}$$

where $u_h^n = T(\hat{u}_h^n)$ solves the following equation

$$B((u_h^n)^+, v_h) + \Delta ts((u_h^n)^-, v_h) = F(v_h) - \Delta t\theta(\boldsymbol{\beta} \cdot \nabla((\hat{u}_h^n)^+, v_h)_\Omega, \tag{3.27}$$

at each time-level $1 \leq n \leq N$, $F(\cdot)$ is defined in (3.21-2). It can be observed that u_h^n satisfies (3.21) if and only if $T(u_h^n) = u_h^n$. Therefore, the proof proceeds by showing that the operator T satisfies the conditions required by Brouwer's Fixed Point Theorem (Theorem 1.2.4).

i) T is well-defined: To prove that T is well-defined, we see that (3.27) is a particular example of the finite element method (1.64). So, applying 1.6.23, there exists a unique solution $u_h^n \in V_{\mathcal{P}}$ of (3.27), and thus T is well-defined.

ii) T is continuous: Using the monotonicity result from Theorem 1.6.23 for all $v_h, w_h \in V_{\mathcal{P}}$, we have

$$B(v_h^+ - w_h^+, v_h - w_h) + \Delta ts(v_h^- - w_h^-, v_h - w_h) \geq C\|v_h - w_h\|_{h,\theta\Delta t}^2.$$

Next, suppose that for $\hat{v}_h, \hat{w}_h \in V_{\mathcal{P}}$ and let $v_h = T(\hat{v}_h)$ and $w_h = T(\hat{w}_h)$. Then, using integrating by parts,

Lemma 3.2.2, and (3.18), we get

$$\begin{aligned} C\|v_h - w_h\|_{h,\theta\Delta t}^2 &\le B(v_h^+ - w_h^+, v_h - w_h) + s(v_h^- - w_h^-, v_h - w_h) \\ &= -\theta\Delta t(\boldsymbol{\beta}\cdot\nabla(\hat{v}_h^+ - \hat{w}_h^+), v_h - w_h)_\Omega \\ &= \theta\Delta t(\hat{v}_h^+ - \hat{w}_h^+, \boldsymbol{\beta}\cdot\nabla(v_h - w_h))_\Omega \\ &\le C\theta\Delta t\|\boldsymbol{\beta}\|_{0,\infty,\Omega}\|\hat{v}_h - \hat{w}_h\|_{0,\Omega}|v_h - w_h|_{1,\Omega} \\ &\le C\theta\Delta t\tilde{\varepsilon}^{-\frac{1}{2}}\|\boldsymbol{\beta}\|_{0,\infty,\Omega}\|\hat{v}_h - \hat{w}_h\|_{0,\Omega}|\tilde{\varepsilon}^{\frac{1}{2}}(v_h - w_h)|_{1,\Omega} \\ &\le C\theta\Delta t\tilde{\varepsilon}^{-\frac{1}{2}}\|\boldsymbol{\beta}\|_{0,\infty,\Omega}\|\hat{v}_h - \hat{w}_h\|_{0,\Omega}\|v_h - w_h\|_{h,\theta\Delta t}. \end{aligned}$$

Therefore

$$\|T(\hat{v}_h) - T(\hat{w}_h)\|_{h,\theta\Delta t} \le C\theta\Delta t\tilde{\varepsilon}^{-\frac{1}{2}}\|\boldsymbol{\beta}\|_{0,\infty,\Omega}\|\hat{v}_h - \hat{w}_h\|_{0,\Omega},$$

and T is Lipschitz continuous.

iii) There exists $R > 0$, such that $T(B(0,R)) \subseteq B(0,R)$: Let $\hat{z}_h \in V_{\mathcal{P}}$ be arbitrary and $z_h = T(\hat{z}_h)$. By using $v_h = z_h^+$ in (3.27), we get

$$B(z_h^+, z_h^+) + \underbrace{\Delta t s(z_h^-, z_h^+)}_{\ge 0} = F(z_h^+) - \theta\Delta t(\boldsymbol{\beta}\cdot\nabla\hat{z}_h^+, z_h^+)_\Omega \le M\|z_h^+\|_{h,\theta\Delta t}. \tag{3.28}$$

In fact, using the Cauchy-Schwarz, since $\tilde{\mu} = 1 + \mu\theta\Delta t \neq 0$ and if $\theta \geq \frac{1}{2}$, we have

$$
\begin{aligned}
F(z_h^+) \leq C \Big(& \Delta t \parallel f^{n-1+\theta} \parallel_{0,\Omega} \parallel z_h^+ \parallel_{0,\Omega} \\
& + (\Delta t(1-\theta)\varepsilon)^{\frac{1}{2}} |(u_h^{n-1})^+|_{1,\Omega} (\Delta t(1-\theta)\varepsilon)^{\frac{1}{2}} |z_h^+|_{1,\Omega} \\
& + \Delta t(1-\theta) \|\boldsymbol{\beta}\|_{0,\infty,\Omega} |(u_h^{n-1})^+|_{1,\Omega} \parallel z_h^+ \parallel_{0,\Omega} \\
& + \parallel (1+\mu\Delta t(1-\theta))^{\frac{1}{2}} (u_h^{n-1})^+ \parallel_{0,\Omega} \parallel (1+\mu\Delta t(1-\theta))^{\frac{1}{2}} z_h^+ \parallel_{0,\Omega} \\
& + \Delta t(1-\theta) J((u_h^{n-1})^+, (u_h^{n-1})^+)^{\frac{1}{2}} J(z_h^+, z_h^+)^{\frac{1}{2}} \Big) \\
\leq C \Big(& \Delta t \parallel f^{n-1+\theta} \parallel_{0,\Omega} \parallel \tilde{\mu}^{-\frac{1}{2}} \parallel \tilde{\mu}^{\frac{1}{2}} z_h^+ \parallel_{0,\Omega} + \tilde{\varepsilon}^{\frac{1}{2}} |(u_h^{n-1})^+|_{1,\Omega} \tilde{\varepsilon}^{\frac{1}{2}} |z_h^+|_{1,\Omega} \\
& + \Delta t\theta \|\boldsymbol{\beta}\|_{0,\infty,\Omega} \tilde{\varepsilon}^{-\frac{1}{2}} \tilde{\mu}^{-\frac{1}{2}} |\tilde{\varepsilon}^{\frac{1}{2}} (u_h^{n-1})^+|_{1,\Omega} \parallel \tilde{\mu}^{\frac{1}{2}} z_h^+ \parallel_{0,\Omega} \\
& + \parallel \tilde{\mu}^{\frac{1}{2}} (u_h^{n-1})^+ \parallel_{0,\Omega} \parallel \tilde{\mu}^{\frac{1}{2}} z_h^+ \parallel_{0,\Omega} + \Delta t\theta J((u_h^{n-1})^+, (u_h^{n-1})^+)^{\frac{1}{2}} J(z_h^+, z_h^+)^{\frac{1}{2}} \Big) \\
\leq C \Big(& \Delta t \tilde{\mu}^{-\frac{1}{2}} \parallel f^{n-1+\theta} \parallel_{0,\Omega} + \Delta t\theta \|\boldsymbol{\beta}\|_{0,\infty,\Omega} \tilde{\varepsilon}^{-\frac{1}{2}} \tilde{\mu}^{-\frac{1}{2}} \|(u_h^{n-1})^+\|_{h,\theta\Delta t} \\
& + \|(u_h^{n-1})^+\|_{h,\theta\Delta t} \Big) \|z_h^+\|_{h,\theta\Delta t}
\end{aligned}
\tag{3.29}
$$

also, by integrating by parts, we have

$$
\begin{aligned}
\theta\Delta t(\boldsymbol{\beta} \cdot \nabla \hat{z}_h^+, z_h^+)_\Omega &\leq \Delta t\theta \|\boldsymbol{\beta}\|_{0,\infty,\Omega} \tilde{\varepsilon}^{-\frac{1}{2}} \parallel \hat{z}_h^+ \parallel_{0,\Omega} |\tilde{\varepsilon}^{\frac{1}{2}} z_h^+|_{1,\Omega} \\
&\leq \theta\Delta t \tilde{\varepsilon}^{-\frac{1}{2}} \|\hat{z}_h^+\|_{0,\Omega} \|\boldsymbol{\beta}\|_{0,\infty,\Omega} \parallel z_h^+ \parallel_{h,\theta\Delta t} .
\end{aligned}
$$

So, if we use (3.14) and set

$$
\begin{aligned}
M := C \Big(& \Delta t \tilde{\mu}^{-\frac{1}{2}} \parallel f^{n-1+\theta} \parallel_{0,\Omega} + \Delta t\theta \|\boldsymbol{\beta}\|_{0,\infty,\Omega} \tilde{\varepsilon}^{-\frac{1}{2}} \tilde{\mu}^{-\frac{1}{2}} \|(u_h^{n-1})^+\|_{h,\theta\Delta t} \\
& + \|(u_h^{n-1})^+\|_{h,\theta\Delta t} + \theta\Delta t \tilde{\varepsilon}^{-\frac{1}{2}} \kappa(t) \|\boldsymbol{\beta}\|_{0,\infty,\Omega} \Big),
\end{aligned}
\tag{3.30}
$$

then, (3.28) leads to

$$
\|z_h^+\|_{h,\theta\Delta t} \leq M .
$$

Next, we take $v_h = z_h^-$ in (3.27). Using computations analogous to those in (3.28) for $F(z_h^-)$, i.e. by

integrating by parts, using Cauchy-Schwarz and (3.14), we get

$$\begin{aligned} B(z_h^+, z_h^-) + \Delta t s(z_h^-, z_h^-) &= F(z_h^-) - \theta\Delta t(\boldsymbol{\beta}\cdot\nabla\hat{z}_h^+, z_h^-)_\Omega \\ &\le C\left(\Delta t\tilde{\mu}^{-\frac{1}{2}} \parallel f^{n-1+\theta}\parallel_{0,\Omega} + \Delta t\theta\|\boldsymbol{\beta}\|_{0,\infty,\Omega}\tilde{\varepsilon}^{-\frac{1}{2}}\tilde{\mu}^{-\frac{1}{2}}\|(u_h^{n-1})^+\|_{h,\theta\Delta t}\right. \\ &\quad \left. + \|(u_h^{n-1})^+\|_{h,\theta\Delta t} + \theta\Delta t\tilde{\varepsilon}^{-\frac{1}{2}}\kappa(t)\|\boldsymbol{\beta}\|_{0,\infty,\Omega}\right) \parallel z_h^- \parallel_{h,\theta\Delta t} \\ &\le M\|z_h^-\|_{h,\theta\Delta t}, \end{aligned}$$

and by (3.20) and Young's inequality

$$\begin{aligned} B(z_h^+, z_h^-) + \Delta t s(z_h^-, z_h^-) &\le \frac{C_{\text{equiv}}}{\alpha} M\Delta t\|z_h^-\|_s^2 \\ &\le \frac{\Delta t}{2}\left(\frac{C_{\text{equiv}}}{\alpha}M\right)^2 + \frac{\Delta t s(z_h^-, z_h^-)}{2} \\ &=: \frac{M_1}{2} + \frac{\Delta t s(z_h^-, z_h^-)}{2}. \end{aligned}$$

Using Cauchy-Schwarz's and Young's inequalities for $B(z_h^+, z_h^-)$ yields

$$-\lambda B(z_h^-, z_h^-) + \Delta t s(z_h^-, z_h^-) \le M_1 + C\lambda^{-1}B(z_h^+, z_h^+) \le M_1 + C\,M^2,$$

for any $\lambda > 0$. Then, choosing λ small enough, and using Lemma 3.3.2, we get

$$\|z_h^-\|_{h,\theta\Delta t} \le C\big(-\lambda B(z_h^-, z_h^-) + \Delta t s(z_h^-, z_h^-)\big) \le C_2(f^{n-1+\theta}, u_h^{n-1}, \tilde{\varepsilon}, \tilde{\mu}, \boldsymbol{\beta}, \kappa(t_n), h, \Delta t),$$

where $C_2(f^{n-1+\theta}, u_h^{n-1}, \tilde{\varepsilon}, \tilde{\mu}, \boldsymbol{\beta}, \kappa(t), h, \Delta t) = M_1 + CM$. Hence, $z_h = T(\hat{z}_h)$ satisfies the following (uniform) bound

$$\|z_h\|_{h,\theta\Delta t} \le \|z_h^-\|_{h,\theta\Delta t} + \|z_h^+\|_{h,\theta\Delta t} \le M + C_2(f^{n-1+\theta}, u_h^{n-1}, \tilde{\varepsilon}, \tilde{\mu}, \boldsymbol{\beta}, \kappa(t_n), h, \Delta t) =: R.$$

Therefore, $z_h = T(\hat{z}_h) \in B(0, R)$, for every $\hat{z}_h \in V_{\mathcal{P}}$, which shows that $T(B(0, R)) \subseteq B(0, R)$. Hence, using Brouwer's fixed point theorem, there exists one $u_h^n \in V_{\mathcal{P}}$ such that $T(u_h^n) = u_h^n$. In other words, problem (3.21) has at least one solution.

The proofs of (b) and (c) are identical to those of Lemma 2.3.4. Finally, the proof of (d) is identical to that of Corollary (2.3.5). □

Remark 3.3.5. *It is worth mentioning that, due to the equivalence between (3.15) and the variational inequality (3.25), and the well-posedness of the latter, $(u_h^n)^+$ is independent of the choice of the stabilisation, as long as it satisfies (3.20) and (3.23). In particular, all the mentioned methods proven in this work are remained valid if $s(\cdot,\cdot)$ defined in (3.17) is replaced by*

$$s(v_h, w_h) := \alpha \sum_{i=1}^{M} \left(\varepsilon \mathfrak{h}(\boldsymbol{x}_i)^{d-2} + \|\boldsymbol{\beta}(\boldsymbol{x}, t_n)\|_{0,\infty,\omega_i} \mathfrak{h}(\boldsymbol{x}_i)^{d-1} + \mu \mathfrak{h}(\boldsymbol{x}_i)^d \right) v_h(\boldsymbol{x}_i) w_h(\boldsymbol{x}_i).$$

In this case, the solution $(u_h^n)^+$ remains unchanged, as it still satisfies (3.25)*, meaning the overall analysis remains the same.*

3.4 Stability and Error analysis

This section focuses on establishing a stability result and deriving optimal error estimates for the method (3.15) in the particular case where $\theta = 1$, which corresponds to the use of the implicit Euler method for time discretisation. The analysis for $\frac{1}{2} \leq \theta < 1$ involves substantial technical obstacles, and we were unable to establish the required estimates within this work.

For $\theta = 1$, the method (3.15) takes the following form:

$$\begin{cases} \text{For } 1 \leq n \leq N, \text{ find } u_h \in V_{\mathcal{P}} \text{ such that} \\ (\delta(u_h^n)^+, v_h)_\Omega + a_h(u_h^n; v_h) = (f^n, v_h)_\Omega \qquad \forall v_h \in V_{\mathcal{P}}, \\ u_h^0 = i_h u^0. \end{cases} \tag{3.31}$$

To prove the stability, we use the test function $v_h = (u_h^n)^+$ in (3.31), and obtain

$$\begin{aligned} &(\delta(u_h^n)^+, (u_h^n)^+)_\Omega + \varepsilon(\nabla(u_h^n)^+, \nabla(u_h^n)^+)_\Omega + (\boldsymbol{\beta} \cdot \nabla(u_h^n)^+, (u_h^n)^+)_\Omega \\ &\qquad + \mu((u_h^n)^+, (u_h^n)^+)_\Omega + \boldsymbol{J}((u_h^n)^+, (u_h^n)^+) + s((u_h^n)^-, (u_h^n)^+) = (f^n, (u_h^n)^+)_\Omega, \end{aligned}$$

or, equivalently, using that $\left(\boldsymbol{\beta} \cdot \nabla(u_h^n)^+, (u_h^n)^+\right)_\Omega = 0$,

$$\begin{aligned} &((u_h^n)^+ - (u_h^{n-1})^+, (u_h^n)^+)_\Omega + \Delta t \Big\{ \varepsilon \, |(u_h^n)^+|_{1,\Omega}^2 \\ &+ \mu \, \|(u_h^n)^+\|_{0,\Omega}^2 + \boldsymbol{J}((u_h^n)^+, (u_h^n)^+) \Big\} + \Delta t s((u_h^n)^-, (u_h^n)^+) = \Delta t (f^n, (u_h^n)^+)_\Omega. \end{aligned}$$

The relation $2p(p-q) = p^2+(p-q)^2-q^2$, the Cauchy-Schwarz inequality, and the fact that $s((u_h^n)^-, (u_h^n)^+) \geq 0$ (by Lemma 1.6.22) lead to

$$\| (u_h^n)^+ \|_{0,\Omega}^2 - \| (u_h^{n-1})^+ \|_{0,\Omega}^2 + \| (u_h^n)^+ - (u_h^{n-1})^+ \|_{0,\Omega}^2 + 2\Delta t \left\{ \varepsilon |(u_h^n)^+|_{1,\Omega}^2 + \mu \| (u_h^n)^+ \|_{0,\Omega}^2 + J((u_h^n)^+, (u_h^n)^+) \right\} \leq 2\Delta t \| f^n \|_{0,\Omega} \| (u_h^n)^+ \|_{0,\Omega} . \tag{3.32}$$

Using Young's inequality for the right hand side and then summing through n, $n = 0, \dots, m$, we get that

$$\| (u_h^m)^+ \|_{0,\Omega}^2 + \sum_{n=0}^{m} \| (u_h^n)^+ - (u_h^{n-1})^+ \|_{0,\Omega}^2 + 2 \sum_{n=0}^{m} \Delta t \left\{ \varepsilon |(u_h^n)^+|_{1,\Omega}^2 + \mu \| (u_h^n)^+ \|_{0,\Omega}^2 + J((u_h^n)^+, (u_h^n)^+) \right\} \leq \| u_h^0 \|_{0,\Omega}^2 + \sum_{n=0}^{m} \Delta t \left(T \| f^n \|_{0,\Omega}^2 + \frac{1}{T} \| (u_h^n)^+ \|_{0,\Omega}^2 \right).$$

If we set $a_n = \| (u_h^n)^+ \|_{0,\Omega}^2$, $B = \| u_h^0 \|_{0,\Omega}^2$, $k = 1$, $\gamma_n = \frac{\Delta t}{T}$ and $\sigma_n = \left(1 - \frac{\Delta t}{T}\right)^{-1}$, $c_n = \Delta t T \| f^n \|_{0,\Omega}^2$ and

$$b_n = 2\Delta t \left(\varepsilon |(u_h^n)^+|_{1,\Omega}^2 + \mu \| (u_h^n)^+ \|_{0,\Omega}^2 + J((u_h^n)^+, (u_h^n)^+) \right),$$

then using the Grönwall's inequality Lemma 1.3.3, we get

$$\| (u_h^m)^+ \|_{0,\Omega}^2 + 2\Delta t \sum_{n=0}^{m} \left(\varepsilon |(u_h^n)^+|_{1,\Omega}^2 + \mu \| (u_h^n)^+ \|_{0,\Omega}^2 + J((u_h^n)^+, (u_h^n)^+) \right) \leq \exp\left(\sum_{n=0}^{m} \frac{\Delta t}{T} \left(1 - \frac{\Delta t}{T}\right)^{-1} \right) \left(\| u_h^0 \|_{0,\Omega}^2 + \Delta t T \sum_{n=0}^{m} \| f^n \|_{0,\Omega}^2 \right).$$

In this way we have proved the following stability result for the scheme (3.31).

Lemma 3.4.1. *Let $u_h^n \in V_{\mathcal{P}}$, for $n = 1, \dots, N$ solve* (3.31)*. Then the following stability estimates holds true:*

$$\max_{1 \leq m \leq N} \| (u_h^m)^+ \|_{0,\Omega}^2 + 2\Delta t \sum_{n=0}^{N} \left(\varepsilon |(u_h^n)^+|_{1,\Omega}^2 + \mu \| (u_h^n)^+ \|_{0,\Omega}^2 + J((u_h^n)^+, (u_h^n)^+) \right) \leq e^2 \left(\| u_h^0 \|_{0,\Omega}^2 + \Delta t T \sum_{n=0}^{N} \| f^n \|_{0,\Omega}^2 \right).$$

Remark 3.4.2. *In the particular case $f = 0$,* (3.32) *implies that*

$$\| (u_h^m)^+ \|_{0,\Omega} \leq \| u_h^0 \|_{0,\Omega},$$

and then (3.31) *is strongly stability preserving.* □

The next result states optimal order error estimates for the method (3.31).

Theorem 3.4.3. *Let $u^0 \in H^{k+1}(\Omega)$, $u \in L^\infty((0,T); H^{k+1}(\Omega)) \cap H^1((0,T); H^{k+1}(\Omega)) \cap H^2((0,T); L^2(\Omega))$ be the solution of (3.1), $u(\cdot,t) \in H_0^1(\Omega)$ for almost all $t \in [0,T]$, and $u_h^n \in V_{\mathcal{P}}$ be the solution of (3.31) at the time step n. Then, defining $E^m = (u_h^m)^+ - u^m$, there exists a constant $C > 0$, independent of $h, \Delta t$ and any physical parameter, such that*

$$\begin{aligned}
&\max_{1\le m\le N} \| E^m \|_{0,\Omega}^2 + \Delta t \sum_{n=1}^{N} \left(\varepsilon |E_h^n|_{1,\Omega}^2 + \mu \| E_h^n \|_{0,\Omega}^2 + J(E_h^n, E_h^n) \right) \\
&\qquad \le Ce^2 \left[h^{2k} \left\{ \Delta t \sum_{n=0}^{N} \left(\varepsilon + T\|\boldsymbol{\beta}\|_{0,\infty,\Omega}^2 + h\gamma \|\boldsymbol{\beta}\|_{0,\infty,\Omega} + h^2 T\mu^2 \right) |u^n|_{k+1,\Omega}^2 \right. \right. \\
&\qquad\qquad \left. \left. + h^2 \max_{1\le m\le N} |u^m|_{k+1,\Omega}^2 + Th^2 \int_0^T |\partial_t u(t)|_{k+1,\Omega}^2 \mathrm{d}t \right\} + \Delta t^2 T \int_0^T \| \partial_{tt} u(t) \|_{0,\Omega}^2 \, \mathrm{d}t \right].
\end{aligned} \tag{3.33}$$

Proof. For $n = 1, \dots, N$ we decompose $E^n = (u_h^n)^+ - u^n$ as

$$E^n = (u_h^n)^+ - u^n = \left((u_h^n)^+ - i_h u^n \right) + \left(i_h u^n - u^n \right) =: E_h^n + \eta_h^n, \tag{3.34}$$

where we recall that $i_h u^n$ is the Lagrange interpolant (1.22) of u^n. Subtracting (3.3) from the method (3.31) we arrive at the following error equation

$$\begin{aligned}
&\left(\delta (u_h^n)^+ - \partial_t u^n, v_h \right)_\Omega + \varepsilon \left(\nabla ((u_h^n)^+ - u^n), \nabla v_h \right)_\Omega + \left(\boldsymbol{\beta} \cdot \nabla ((u_h^n)^+ - u^n), v_h \right)_\Omega \\
&\qquad + \mu \left((u_h^n)^+ - u^n, v_h \right)_\Omega + J((u_h^n)^+, v_h) + s((u_h^n)^-, v_h) = 0.
\end{aligned} \tag{3.35}$$

Rearranging and using (3.34),we get

$$\begin{aligned}
&(\delta E_h^n, v_h)_\Omega + \varepsilon \left(\nabla E_h^n, \nabla v_h \right)_\Omega + (\boldsymbol{\beta} \cdot \nabla E_h^n, v_h)_\Omega + \mu (E_h^n, v_h)_\Omega + J((u_h^n)^+, v_h) + s((u_h^n)^-, v_h) \\
&\qquad = -(\delta (i_h u^n) - \partial_t u^n, v_h)_\Omega - \varepsilon (\nabla \eta_h^n, \nabla v_h)_\Omega - (\boldsymbol{\beta} \cdot \nabla \eta_h^n, v_h)_\Omega - \mu (\eta_h^n, v_h)_\Omega.
\end{aligned} \tag{3.36}$$

Since $u^n \in H^2(\Omega)$, then $J(u^n, v) = 0$, and so we deduce that

$$J((u_h^n)^+, v) = J((u_h^n)^+ - u^n, v) = J(E_h^n, v) + J(\eta_h^n, v).$$

Using the test function $v_h = E_h^n$ in (3.36) and $(\boldsymbol{\beta}\cdot\nabla E_h^n, E_h^n)_\Omega = 0$, we get

$$\begin{aligned}&(\delta E_h^n, E_h^n)_\Omega + \varepsilon\,|E_h^n|_{1,\Omega}^2 + \mu\,\|E_h^n\|_{0,\Omega}^2 + J(E_h^n, E_h^n) + s((u_h^n)^-, E_h^n)\\ &= -(\delta(i_h u^n) - \partial_t u^n, E_h^n)_\Omega - \varepsilon\left(\nabla\eta_h^n, \nabla E_h^n\right)_\Omega - (\boldsymbol{\beta}\cdot\nabla\eta_h^n, E_h^n)_\Omega - \mu(\eta_h^n, E_h^n)_\Omega - J(\eta_h^n, E_h^n).\end{aligned} \tag{3.37}$$

Since $(i_h u^n)^- = 0$, the monotonicity inequality (3.24) yields

$$s((u_h^n)^-, E_h^n) = s((u_h^n)^-, (u_h^n)^+ - i_h u^n) = s((u_h^n)^- - (i_h u^n)^-, (u_h^n)^+ - (i_h u^n)^+) \geq 0. \tag{3.38}$$

So, using the relation $2p(p-q) = p^2 + (p-q)^2 - q^2$ for the first term of (3.37), applying the inequality (3.38), next using the Cauchy–Schwarz inequality and then Young's inequality for the terms on the right hand side, we get

$$\begin{aligned}&\|\,E_h^n\,\|_{0,\Omega}^2 - \|\,E_h^{n-1}\,\|_{0,\Omega}^2 + \|\,E_h^n - E_h^{n-1}\,\|_{0,\Omega}^2 + 2\Delta t\left(\varepsilon|E_h^n|_{1,\Omega}^2 + \mu\,\|\,E_h^n\,\|_{0,\Omega}^2 + J(E_h^n, E_h^n)\right)\\ &\quad\leq 2\Delta t\Big(\,\|\,\delta(i_h u^n) - \partial_t u^n\,\|_{0,\Omega}\|\,E_h^n\,\|_{0,\Omega} + \varepsilon|\eta_h^n|_{1,\Omega}|E_h^n|_{1,\Omega} + \|\,\boldsymbol{\beta}\cdot\nabla\eta_h^n\,\|_{0,\Omega}\|\,E_h^n\,\|_{0,\Omega}\\ &\qquad + \mu\,\|\,\eta_h^n\,\|_{0,\Omega}\|\,E_h^n\,\|_{0,\Omega} + J(\eta_h^n, \eta_h^n)^{\frac{1}{2}} J(E_h^n, E_h^n)^{\frac{1}{2}}\Big)\\ &\quad\leq \Delta t\Big(\varepsilon|\eta_h^n|_{1,\Omega}^2 + \varepsilon|E_h^n|_{1,\Omega}^2 + T\left(\|\,\delta(i_h u^n) - \partial_t u^n\,\|_{0,\Omega} + \|\,\boldsymbol{\beta}\cdot\nabla\eta_h^n\,\|_{0,\Omega} + \mu\,\|\,\eta_h^n\,\|_{0,\Omega}\right)^2\\ &\qquad + \frac{1}{T}\,\|\,E_h^n\,\|_{0,\Omega}^2 + J(E_h^n, E_h^n) + J(\eta_h^n, \eta_h^n)\Big).\end{aligned}$$

Rearranging terms on the both sides of the inequality yields

$$\begin{aligned}&\|\,E_h^n\,\|_{0,\Omega}^2 - \|\,E_h^{n-1}\,\|_{0,\Omega}^2 + \|\,E_h^n - E_h^{n-1}\,\|_{0,\Omega}^2 + \Delta t\left(\varepsilon|E_h^n|_{1,\Omega}^2 + 2\mu\,\|\,E_h^n\,\|_{0,\Omega}^2 + J(E_h^n, E_h^n)\right)\\ &\quad\leq \Delta t\left(\varepsilon|\eta_h^n|_{1,\Omega}^2 + T\left(\|\,\delta(i_h u^n) - \partial_t u^n\,\|_{0,\Omega} + \|\,\boldsymbol{\beta}\,\|_{0,\infty,\Omega}\,|\eta_h^n|_{1,\Omega} + \mu\,\|\,\eta_h^n\,\|_{0,\Omega}\right)^2\right.\\ &\qquad \left. + \frac{1}{T}\,\|\,E_h^n\,\|_{0,\Omega}^2 + J(\eta_h^n, \eta_h^n)\right)\\ &\quad\leq C\Delta t\left(\left(\varepsilon + T\|\boldsymbol{\beta}\|_{0,\infty,\Omega}^2\right)|\eta_h^n|_{1,\Omega}^2 + T\,\|\,\delta(i_h u^n) - \partial_t u^n\,\|_{0,\Omega}^2 + T\mu^2\,\|\,\eta_h^n\,\|_{0,\Omega}^2\right.\\ &\qquad \left. + \frac{1}{T}\,\|\,E_h^n\,\|_{0,\Omega}^2 + J(\eta_h^n, \eta_h^n)\right).\end{aligned}$$

Summing for $n = 0$ to m and using $E_h^0 = 0$ leads to

$$\begin{aligned}\| E_h^m \|_{0,\Omega}^2 + \Delta t \sum_{n=1}^{m} \left(\varepsilon |E_h^n|_{1,\Omega}^2 + 2\mu \| E_h^n \|_{0,\Omega}^2 + J(E_h^n, E_h^n) \right) \\ \leq C\, \Delta t \sum_{n=0}^{m} \left\{ \left(\varepsilon + T\|\boldsymbol{\beta}\|_{0,\infty,\Omega}^2 \right) |\eta_h^n|_{1,\Omega}^2 + T \| \delta(i_h u^n) - \partial_t u^n \|_{0,\Omega}^2 \right. \\ \left. + T\mu^2 \| \eta_h^n \|_{0,\Omega}^2 + \frac{1}{T} \| E_h^n \|_{0,\Omega}^2 + J(\eta_h^n, \eta_h^n) \right\}.\end{aligned}$$

We are now ready to use Grönwall's Lemma 1.3.3 with $k = 1$, $\gamma_n = \frac{\Delta t}{T}$, $\sigma_n = (1 - \frac{\Delta t}{T})^{-1}$, $a_n = \| E_h^n \|_{0,\Omega}$, $B = 0$ and

$$c_n = \Delta t \left(\left(\varepsilon + T\|\boldsymbol{\beta}\|_{0,\infty,\Omega}^2 \right) |\eta_h^n|_{1,\Omega}^2 + T \| \delta(i_h u^n) - \partial_t u^n \|_{0,\Omega}^2 + T\mu^2 \| \eta_h^n \|_{0,\Omega}^2 + J(\eta_h^n, \eta_h^n) \right),$$

which gives

$$\begin{aligned}\| E_h^m \|_{0,\Omega}^2 + \Delta t \sum_{n=1}^{m} \left(\varepsilon |E_h^n|_{1,\Omega}^2 + 2\mu \| E_h^n \|_{0,\Omega}^2 + J(E_h^n, E_h^n) \right) \\ \leq Ce^2 \left[\Delta t \sum_{n=0}^{m} \left\{ \left(\varepsilon + T\|\boldsymbol{\beta}\|_{0,\infty,\Omega}^2 \right) |\eta_h^n|_{1,\Omega}^2 + T \| \delta(i_h u^n) - \partial_t u^n \|_{0,\Omega}^2 \right. \right. \\ \left. \left. + T\mu^2 \| \eta_h^n \|_{0,\Omega}^2 + J(\eta_h^n, \eta_h^n) \right\} \right].\end{aligned} \tag{3.39}$$

Next, to reach the error estimate (3.33) we bound each term of the right hand side of (3.39). First, using the triangle inequality, yields

$$\| \delta(i_h u^n) - \partial_t u^n \|_{0,\Omega} \leq \| \delta(i_h u^n) - \delta u^n \|_{0,\Omega} + \| \delta u^n - \partial_t u^n \|_{0,\Omega}.$$

For the first term in the above inequality, using (1.23), the Taylor's Theorem and the Cauchy-Schwarz in-

equality, we have

$$\begin{aligned}
\| \delta(i_h u^n) - \delta u^n \|_{0,\Omega}^2 &\leq Ch^{2k+2} |\delta u^n|_{k+1,\Omega}^2 \\
&\leq Ch^{2k+2} \left| \frac{1}{\Delta t} \int_{t_{n-1}}^{t_n} \partial_t u(t) \mathrm{d}t \right|_{k+1,\Omega}^2 \\
&\leq Ch^{2k+2} \left| \frac{1}{\Delta t} \left(\int_{t_{n-1}}^{t_n} \mathrm{d}t \right)^{\frac{1}{2}} \left(\int_{t_{n-1}}^{t_n} |\partial_t u(t)|^2 \mathrm{d}t \right)^{\frac{1}{2}} \right|_{k+1,\Omega}^2 \\
&\leq C \frac{h^{2k+2}}{\Delta t} \int_{t_{n-1}}^{t_n} |\partial_t u(t)|_{k+1,\Omega}^2 \mathrm{d}t.
\end{aligned}$$

For the second term, one further use of Taylor's Theorem and the Cauchy-Schwarz inequality gives

$$\begin{aligned}
\| \delta u^n - \partial_t u^n \|_{0,\Omega}^2 &= \int_\Omega (\delta u^n - \partial_t u^n)^2 \mathrm{d}\boldsymbol{x} \\
&\leq \int_\Omega \left(\int_{t_{n-1}}^{t_n} |\partial_{tt} u(t)| \mathrm{d}t \right)^2 \mathrm{d}\boldsymbol{x} \\
&\leq \int_\Omega \int_{t_{n-1}}^{t_n} \mathrm{d}t \int_{t_{n-1}}^{t_n} |\partial_{tt} u(t)|^2 \mathrm{d}t \, \mathrm{d}\boldsymbol{x} \\
&= \Delta t \int_{t_{n-1}}^{t_n} \| \partial_{tt} u(t) \|_{0,\Omega}^2 \, \mathrm{d}t.
\end{aligned}$$

Next, using (1.23), we have

$$|\eta_h^n|_{1,\Omega}^2 \leq Ch^{2k} |u^n|_{k+1,\Omega}^2 \quad , \quad \|\eta_h^n\|_{0,\Omega}^2 \leq C\, h^{2k+2} |u^n|_{k+1,\Omega}^2. \tag{3.40}$$

Furthermore, using the inverse inequality (1.26) and the approximation inequality (1.23), we have

$$J(\eta_h^n, \eta_h^n) = \gamma \sum_{F \in \mathcal{F}_I} \int_F \|\boldsymbol{\beta}\|_{0,\infty,F} h_F^2 [\![\nabla \eta_h^n]\!] \cdot [\![\nabla \eta_h^n]\!] \mathrm{d}s \leq C\gamma h^{2k+1} \|\boldsymbol{\beta}\|_{0,\infty,\Omega} |u^n|_{k+1,\Omega}^2. \tag{3.41}$$

Gathering all the above bounds, we arrive at

$$\begin{aligned}
&\| E_h^m \|_{0,\Omega}^2 + \Delta t \sum_{n=1}^{m} \left(\varepsilon |E_h^n|_{1,\Omega}^2 + 2\mu \| E_h^n \|_{0,\Omega}^2 + J(E_h^n, E_h^n) \right) \\
&\qquad \leq Ce^2 \left[h^{2k} \left\{ \Delta t \sum_{n=0}^{m} \left(\varepsilon + T\|\boldsymbol{\beta}\|_{0,\infty,\Omega}^2 + h\gamma \|\boldsymbol{\beta}\|_{0,\infty,\Omega} + h^2 T\mu^2 \right) |u^n|_{k+1,\Omega}^2 \right. \right. \\
&\qquad\qquad \left. \left. + Th^2 \sum_{n=0}^{m} \int_{t_{n-1}}^{t_n} |\partial_t u(t)|_{k+1,\Omega}^2 \mathrm{d}t \right\} + T\Delta t^2 \sum_{n=0}^{m} \int_{t_{n-1}}^{t_n} \| \partial_{tt} u(t) \|_{0,\Omega}^2 \, \mathrm{d}t \right].
\end{aligned}$$

Finally, using the triangle inequality and (1.23) once again, we arrive at the final estimate

$$\begin{aligned}
&\max_{1\leq m\leq N} \| E^m \|_{0,\Omega}^2 + \Delta t \sum_{n=1}^{N} \left(\varepsilon |E_h^n|_{1,\Omega}^2 + \mu \| E_h^n \|_{0,\Omega}^2 + J(E_h^n, E_h^n) \right) \\
&\qquad \leq Ce^2 \left[h^{2k} \left\{ \Delta t \sum_{n=0}^{N} \left(\varepsilon + T\|\boldsymbol{\beta}\|_{0,\infty,\Omega}^2 + h\gamma \|\boldsymbol{\beta}\|_{0,\infty,\Omega} + h^2 T\mu^2 \right) |u^n|_{k+1,\Omega}^2 \right. \right. \\
&\qquad\qquad \left. \left. + h^2 \max_{1\leq m\leq N} |u^m|_{k+1,\Omega}^2 + Th^2 \int_0^T |\partial_t u(t)|_{k+1,\Omega}^2 \mathrm{d}t \right\} + T\Delta t^2 \int_0^T \| \partial_{tt} u(t) \|_{0,\Omega}^2 \, \mathrm{d}t \right],
\end{aligned}$$

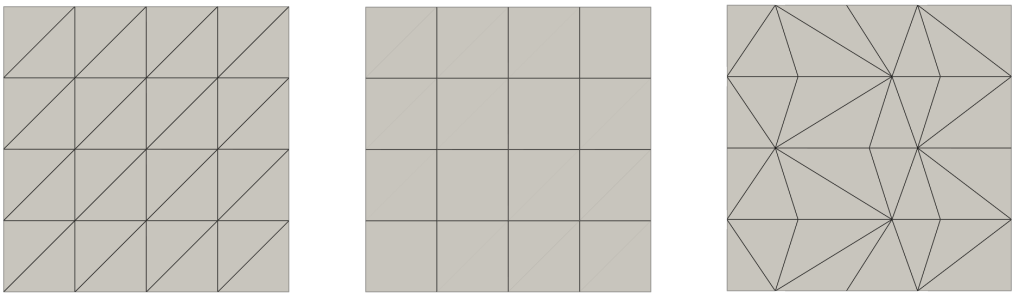
which proves the result. □

3.5 Numerical experiments

In this section we present two experiments to test the numerical performance of (3.15). In these experiments we have used $\Omega = (0,1)^2$ and value $\alpha = 1$ in the stabilising bilinear form $s(\cdot,\cdot)$. Except for the very last numerical result in this section, we have used three types of meshes, a three-directional triangular mesh, regular quadrilateral one and a non-Delaunay mesh. The non-Delaunay mesh in Figure 3.1c is obtained by shifting some of the interior nodes from the mesh in Figure 2.1b to the right, which results in the formation of obtuse angles. The coarsest level of each is depicted in Figure 3.1.

To solve the nonlinear problem (3.15) at each discrete time t_n, $n = 1, 2, \ldots, N$, first we set $\tilde{u}^0 := u_h^{n-1}$. Next, by choosing an appropriate damping parameter $\omega \in (0, 1]$ the following fixed point Richardson-like iterative method is used to find $\tilde{u}^{m+1} \in V_p$, such that

$$\begin{aligned}
&\left(\tilde{u}^{m+1}, v_h\right)_\Omega + \Delta t\theta a_J\left(\tilde{u}^{m+1}, v_h\right) = \left(\tilde{u}^m, v_h\right)_\Omega + \Delta t\theta a_J\left(\tilde{u}^m, v_h\right) \\
&\qquad + \omega \left\{ F^n(v_h) - \left[\left((\tilde{u}^m)^+, v_h\right)_\Omega + \Delta t \Big(\theta a_J\left((\tilde{u}^m)^+, v_h\right) + s\left((\tilde{u}^m)^-, v_h\right) \Big) \right] \right\} \quad \forall v_h \in V_p,
\end{aligned} \tag{3.42}$$



(a) A symmetric, Delaunay mesh.

(b) A simple quadrilateral mesh.

(c) A non-symmetric, non-Delaunay mesh.

Figure 3.1: Three coarse level indicative meshes used in the experiments all with $N = 5$.

for $m = 1, 2, \ldots, N_{\max}$, or until the following stopping criterion is achieved

$$\| \tilde{u}^{m+1} - \tilde{u}^{m} \|_{0,\Omega} \leq 10^{-8}. \tag{3.43}$$

Finally, for m_0 which satisfies in (3.43), we set $u_h^n = \tilde{u}^{m_0+1}$.

In all figures, $P - 1$ indicates the number of divisions in the x and y directions, resulting in a total of P^2 vertices, including the boundary. We evaluate the method's asymptotic performance in the $\| \cdot \|_{0,\Omega}$-norm at the final step, i.e., $\| e_h^N \|_{0,\Omega}$, and to verify the result from Theorem 4.3.7 we examine the asymptotic behaviour of the error by the following norm

$$\| e^N \|_h^2 := \| e^N \|_{0,\Omega}^2 + \sum_{n=0}^{N} \Delta t \left(\varepsilon \| \nabla e_h^n \|_{0,\Omega}^2 + \mu \| e_h^n \|_{0,\Omega}^2 + J(e_h^n, e_h^n) \right). \tag{3.44}$$

We have used $\mathbb{P}_1$ and $\mathbb{P}_2$ elements in the triangular meshes, and $\mathbb{Q}_1$ elements in the quadrilateral mesh. In the numerical experiments we use the bound preserving Euler (BP-Euler) (3.31) and the bound preserving Crank-Nicholson (BP-CN), i.e., the method (3.15) with $\theta = \frac{1}{2}$, even though stability and error estimates for BP-CN has not been proven.

Example 6 (A problem with a smooth solution)**.** *We consider $\mu = 1$, $\varepsilon = 10^{-6}$, $\boldsymbol{\beta} = (2, 1)$, and set f and u^0 such that the function*

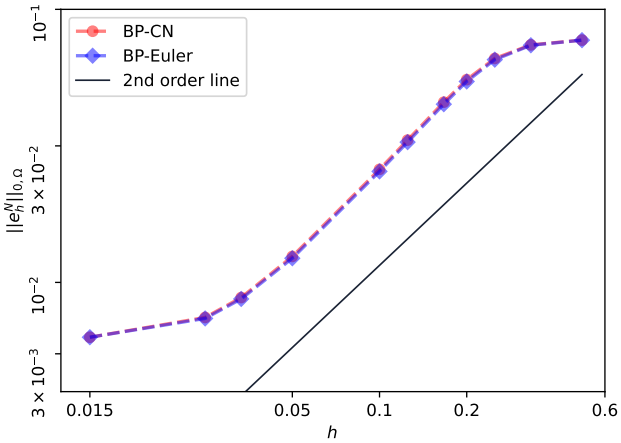
$$u(x, y, t) = \exp(t) \sin(\pi x) \sin(\pi y),$$

is the analytical solution of (2.1). Notice that $u(x, y, t) \in [0, \exp(t)]$, and thus we set $\kappa(t) = \exp(t)$ as the upper bound at time t. The CIP stabilisation parameter $\gamma = 0.05$ has been used in (2.8) *and we set $\omega = 0.1$ for the damping parameter in all the time steps.*

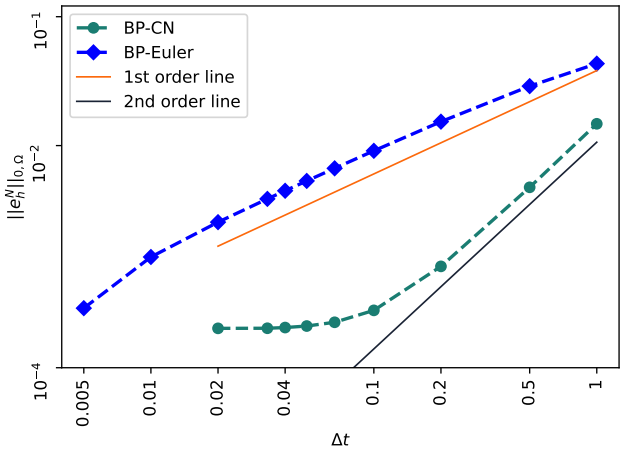
Figure 3.2 illustrates the asymptotic behaviour of the error $\| e_h^N \|_{0,\Omega}$ using $\mathbb{P}_1$ and $\mathbb{P}_2$ elements. These results align with the theoretical findings we established in Theorem 4.3.7. By fixing $\Delta t = 4 \times 10^{-4}$ and decreasing the mesh size as depicted in Figures 3.2a and 3.2c, we observe second-order and third-order convergence when using $\mathbb{P}_1$ and $\mathbb{P}_2$ elements, respectively, for both BP-Euler and BP-CN. Also, by fixing the mesh size $h = 5 \times 10^{-3}$ and varying the length of the time-step Δt, as shown in Figures 3.2b and 3.2d, we obtained first-order convergence for the Euler method and second-order convergence for Crank-Nicholson for both $\mathbb{P}_1$ and $\mathbb{P}_2$ elements, as expected.

Figure 3.3 depicts the asymptotic behaviour of the error $\| e^N \|_h^2$ using $\mathbb{P}_1$ and $\mathbb{P}_2$ elements. These results also corroborate the theoretical results we proved in Theorem 4.3.7. By fixing the time step and decreasing the mesh size as shown in Figures 3.3a and 3.3c, we observed second-order and third-order convergence when using $\mathbb{P}_1$ and $\mathbb{P}_2$ elements, respectively for the BP-Euler method. This extra order of convergence is, most likely, due to the fact that the small value of ε makes that the $\| \cdot \|_h$ norm is dominated by the $L^2(\Omega)$-norm. Additionally, we achieved first-order convergence for the BP-Euler method for both $\mathbb{P}_1$ and $\mathbb{P}_2$ elements when the size of the time step is decreased. Figures 3.3b and 3.3d show the asymptotic behaviour with respect to time for the BP-Euler method.

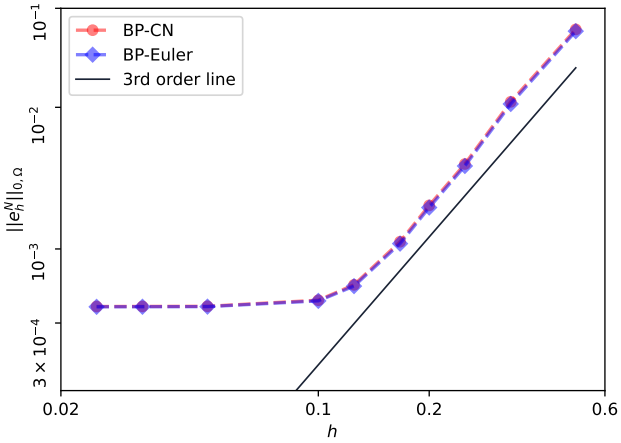
To assess the computational cost of the nonlinear algorithm at each time step, we depicted in Figure 3.4 the average number of iterations per step over 1000 time steps for a sequence of meshes with decreasing mesh size. The results indicate that there is no significant increase in the average number of iterations, which remains low regardless of the mesh size.



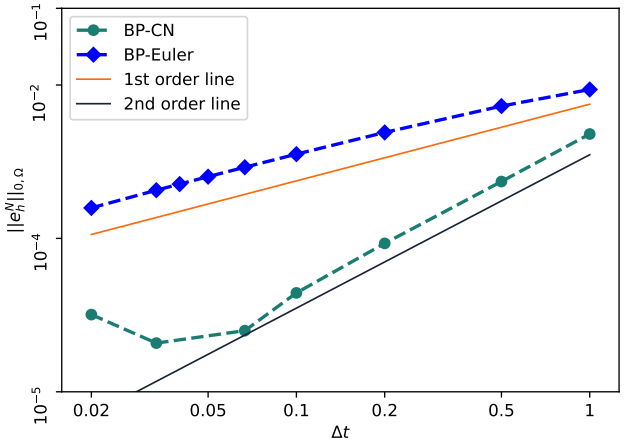
(a) $\Delta t = 4 \times 10^{-4}$, T=0.2, $\mathbb{P}_1$ elements.



(b) $h = 5 \times 10^{-3}$, $T = 1$, $\mathbb{P}_1$ elements.

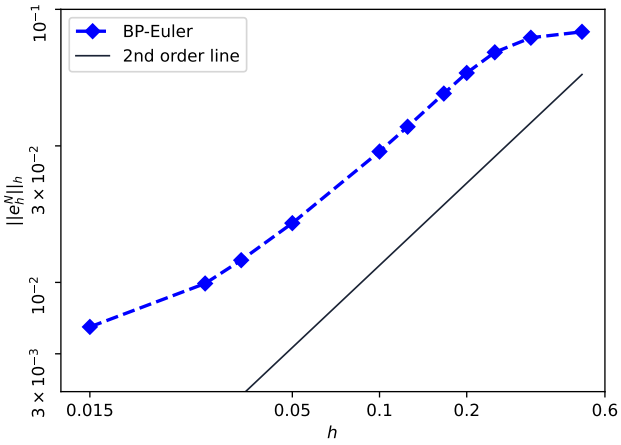


(c) $\Delta t = 4 \times 10^{-4}$, T=0.2, $\mathbb{P}_2$ elements.

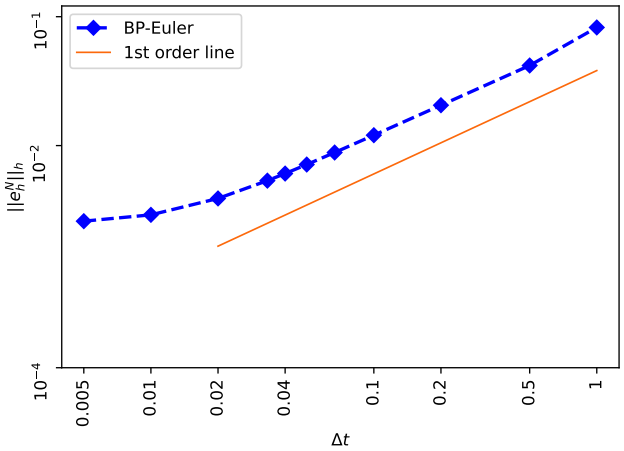


(d) $h = 5 \times 10^{-3}$, $T = 1$, $\mathbb{P}_2$ elements

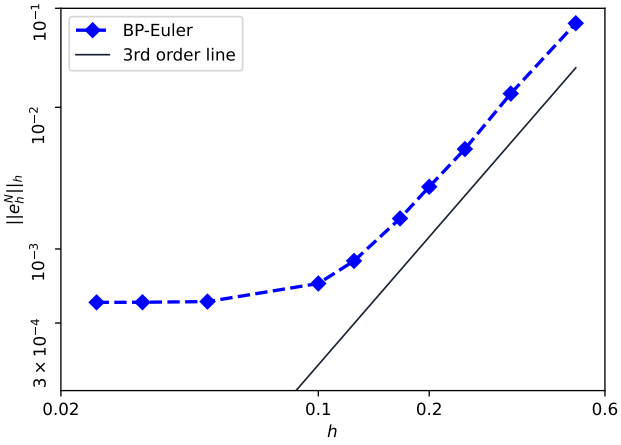
Figure 3.2: Comparison of the error of the approximated solution by the BP-Euler method and BP-CN method with the exact solution in $\|\cdot\|_{0,\Omega}$-norm (using mesh 3.1a).



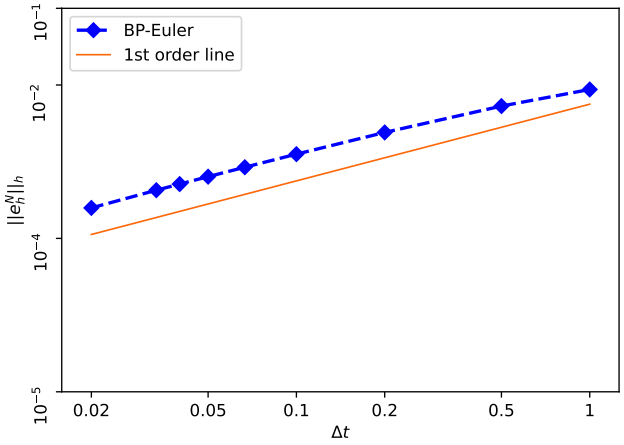
(a) $\Delta t = 4 \times 10^{-4}$, T=0.2, $\mathbb{P}_1$ elements.



(b) $h = 5 \times 10^{-3}$, $T = 1$, $\mathbb{P}_1$ elements.

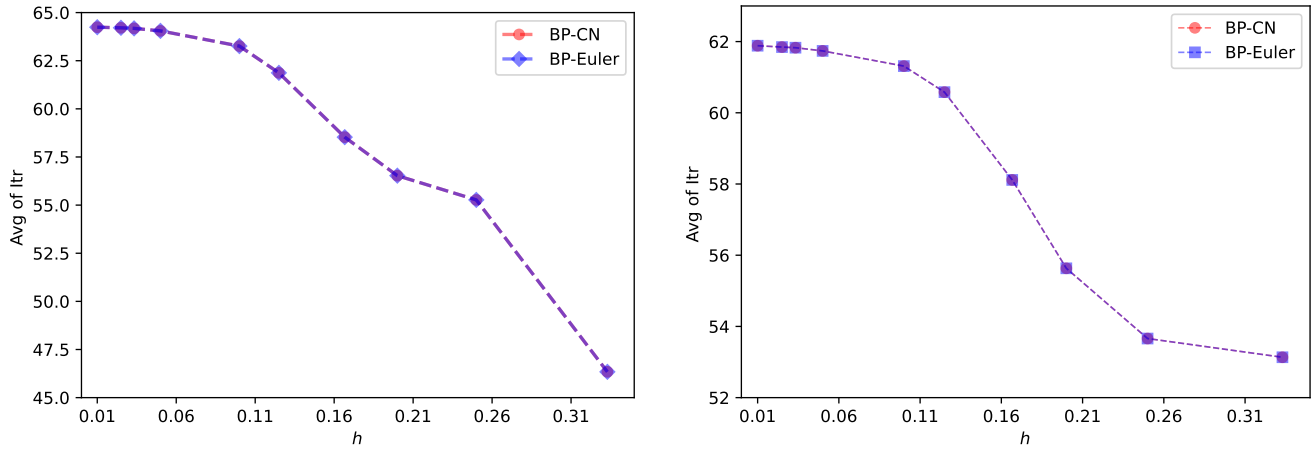


(c) $\Delta t = 4 \times 10^{-4}$, T=0.2, $\mathbb{P}_2$ elements.



(d) $h = 5 \times 10^{-3}$, $T = 1$, $\mathbb{P}_2$ elements

Figure 3.3: Using norm (3.44) for the comparison of the error of the approximated solution by the BP-Euler method with the exact solution (using mesh 3.1a).



(a) $\mathbb{P}_1$ elements

(b) $\mathbb{Q}_1$ elements

Figure 3.4: The average number of the Richardson iterations of 1000 time steps ($T = 1$) needed to reach convergence using $\mathbb{P}_1$ and $\mathbb{Q}_1$ elements and BP-Euler and BP-CN methods and the meshes 3.1a and 3.1b.

Remark 3.5.1. *As mentioned in the previous chapter, improvements to the nonlinear solver itself can enhance performance. In this chapter, we again used a simple Richardson type solver to highlight the simplicity of the scheme, but more efficient nonlinear solvers, such as localised Newton methods (see, e.g., [6]), or active set methods [7], can significantly improve convergence speed. Our preliminary results suggest that these alternatives lead to significantly faster convergence. As shown in the previous chapter, applying a localised Newton method reduced the number of iterations for the stationary convection–diffusion problem. Employing a similar approach at each step of the finite element method* (3.15) *likewise yields a substantial reduction in the average number of iterations.*

Example 7 (Three body rotation)**.** *This example is a modified version of the three body rotation transport problem from [97]. We used* $\boldsymbol{\beta} = (0.5 - y, x - 0.5)$, $\varepsilon = 10^{-12}$ *and* $\mu = f = 0$. *The initial setup involves three separate bodies, as depicted in Figure 3.5. Each body's position is defined by its centre at coordinates* (x_0, y_0). *Every body is contained within a circle of radius* $r_0 = 0.15$ *cantered at* (x_0, y_0). *Outside these three bodies, the initial condition is zero.*

Let

$$r(x, y) = \frac{1}{r_0}\sqrt{(x - x_0)^2 + (y - y_0)^2}. \tag{3.45}$$

The center of the slotted cylinder is in $(x_0, y_0) = (0.5, 0.75)$ *and its geometry is given by*

$$u(0; x, y) = \begin{cases} 1 \quad \text{if} \quad r(x, y) \leq 1, |x - x_0| \geq 0.0225 \quad \text{or} y \geq 0.85; \\ 0 \quad \text{else}, \end{cases} \tag{3.46}$$

The conical body at the bottom side is described by $(x_0, y_0) = (0.5, 0.25)$ *and*

$$u(0; x, y) = 1 - r(x, y). \tag{3.47}$$

Finally, the hump at the left hand side is given by $(x_0, y_0) = (0.25, 0.5)$ *and*

$$u(0; x, y) = \frac{1}{4}(1 + \cos(\pi \min\{r(x, y), 1\})). \tag{3.48}$$

The rotation of the bodies occurs counter-clockwise. A full revolution takes $t = 2\pi$*. We use* $P = 130$ *so a regular grid consisting of* 130×130 *mesh cells for* $\mathbb{P}_1$*,* $\mathbb{P}_2$*, and* $\mathbb{Q}_1$ *elements.*

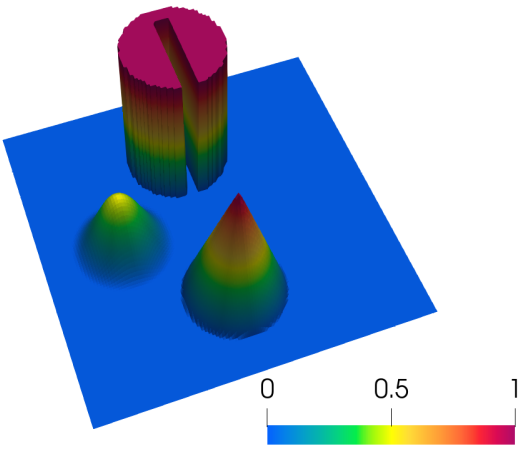


Figure 3.5: Initial data u^0 for rotating body problem.

The simulations were performed with the final time $T = 2\pi$ and the time step $\Delta t = 10^{-3}$. Figure 3.6 depicts the approximation solution for the BP-Euler method and BP-CN method using $\mathbb{P}_1$, $\mathbb{P}_2$ and $\mathbb{Q}_1$ elements. In both methods the CIP term (2.8) was used with the parameter $\gamma = 0.001$. As noted in Remark 2.2.6, there is no universal rule for selecting γ, since its optimal value may depend on the mesh, problem parameters, and the presence of sharp layers. In our numerical experiments, we determine γ empirically. We observe

that moderate values of γ are generally sufficient to suppress oscillations while still resolving sharp layers accurately, whereas excessively small or large values tend to deteriorate the quality of the solution.

Our numerical experiments show that the optimal value of ω (relative to the number of iterations needed to reach convergence) is approximately 0.07 when using the quadrilateral mesh, while for $\mathbb{P}_1$ and $\mathbb{P}_2$ elements, it is around 0.12. So, we report the results using these values.

For comparison purposes, we also approximated the same problem with CIP-Euler and CIP-Crank-Nicolson (CIP-CN) (the method that only use the CIP term i.e., the full time-space discretisation θ scheme of the method (3.7)) with the same value for the parameter γ. To compare the numerical solution of different methods, a cross section along the line $y = 0.75$ was taken of initial data (ID) u^0, BP-Euler, BP-CN, CIP-Euler and CIP-CN methods. The results are shown in Figure 3.7. Since we perform a full rotation, we can compare the solution from each method with the initial condition to assess the diffusive properties of each method. As shown in Figure 3.7, among all the methods, BP-CN exhibits the best performance, both in terms of preserving the initial data (ID) and capturing the magnitude of its deformation, regardless of the type of elements used.

The experiment which has been shown in (3.8) aims at assessing the effect of adding CIP stabilisation to the method (3.15). For this, we set $\gamma = 0$ in $J(\cdot, \cdot)$ and the results are shown in Figure 3.8 for the BP-CN method using $\mathbb{P}_1$ elements. In Figure 3.8 we can observe the solution u_h^+, while respecting the bounds of the exact solution, exhibits spurious oscillations near the layers. This justifies the need for CIP stabilisation in (3.15).

To test the performance of the method in the case when the mesh used is not Delaunay, we have approximated this example also in the non-Delaunay mesh depicted in Figure 3.7, using $P = 130$. The same cross-sections of the approximate solutions for the BP-Euler and BP-CN methods are depicted in Figure 3.9, alongside the cross-sections for the CIP stabilised finite element method. In both cases $\gamma = 0.001$ has been used in the simulation. From the results we can observe, once again, that u_h^+ respects the bounds of Assumption (C1), while the CIP solution presents noticeable over and undershoots.

Finally, to study mass conservation we use the relative mass, i.e. the ratio of the mass at time t to the initial mass defined by

$$M_r(t) = \frac{M(t)}{M(0)},$$

where $M(t)$ is the total mass at time t, and is defined as

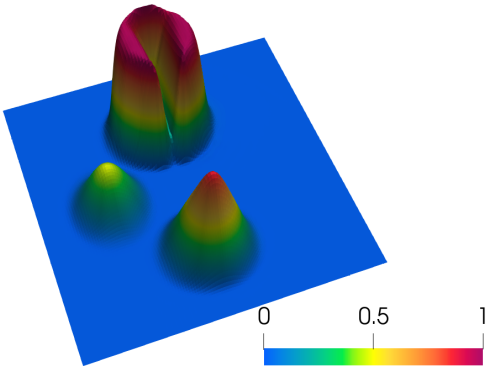
$$M(t) = \int_\Omega u(\boldsymbol{x}, t)\, \mathrm{d}\boldsymbol{x}.$$

The evolution of mass over time for BP-Euler and BP-CN methods is presented in Figure 3.10. The plot depicts the evolution of the relative mass. We observe that, despite the fact that the scheme does not preserve mass, the mass loss/gain remains low throughout the simulation.

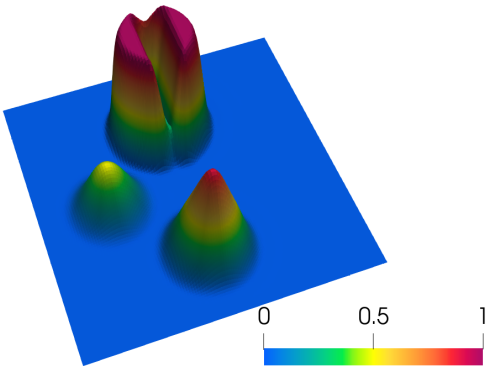
It is important to mention that more economical alternatives, such as linearised flux-corrected transport (FCT) methods (see, e.g., [94]), are also available to ensure bound preservation. Nevertheless, several factors should be considered. One aspect is the CFL condition, as most linear flux-corrected transport methods require such a condition to guarantee bound preservation, whereas our approach does not impose this restriction (see also Remark 2.5.3). Another consideration is the applicability to higher-order elements. FCT methods have primarily been developed for linear finite elements, and bound preservation is not guaranteed for higher-order elements, as the necessary analysis has not yet been carried out.

Several problems remain open at this point. The extension of the stability and error analysis to higher-order time discretisation is, at the moment, an open problem. In addition, the extension of this framework to the transport equation is also of interest. A parallel development is the extension of this methodology to discontinuous Galerkin scheme in space, which is the topic of the companion paper [18] and next chapter. These, and other topics will be the subject of future research.

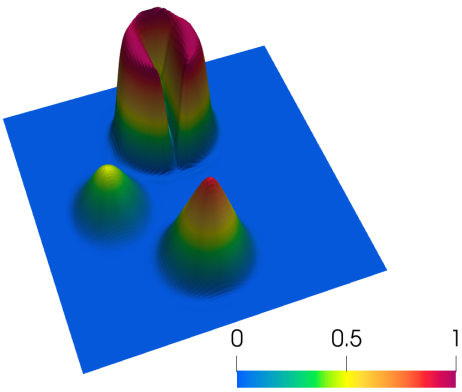
Remark 3.5.2. *Setting* $\epsilon = 0$ *in the convection–diffusion equation* (3.1) *reduces it to the pure transport equation. In this case, the problem becomes hyperbolic, but the finite element method* (3.15) *can still be applied. Moreover, most of the theoretical results proved in this chapter remain valid for the transport equation.*



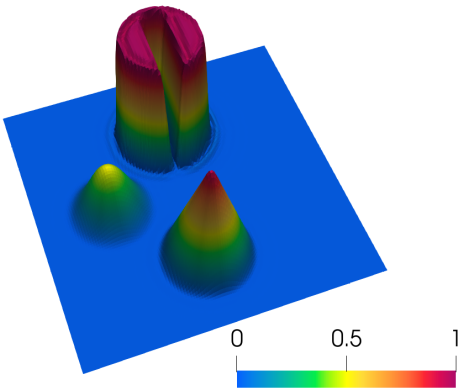
(a) BP-Euler-$\mathbb{P}_1$ elements



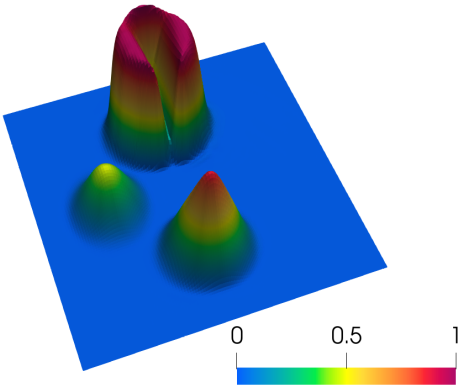
(b) BP-CN,$\mathbb{P}_1$ elements



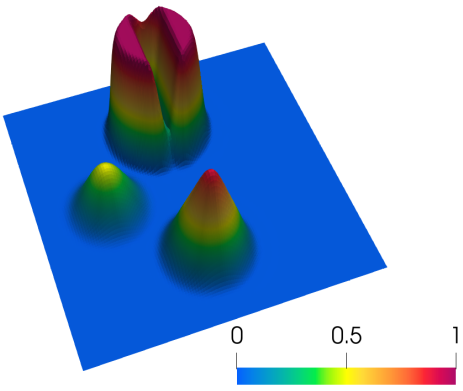
(c) BP-Euler-$\mathbb{P}_2$ elements



(d) BP-CN-$\mathbb{P}_2$ elements

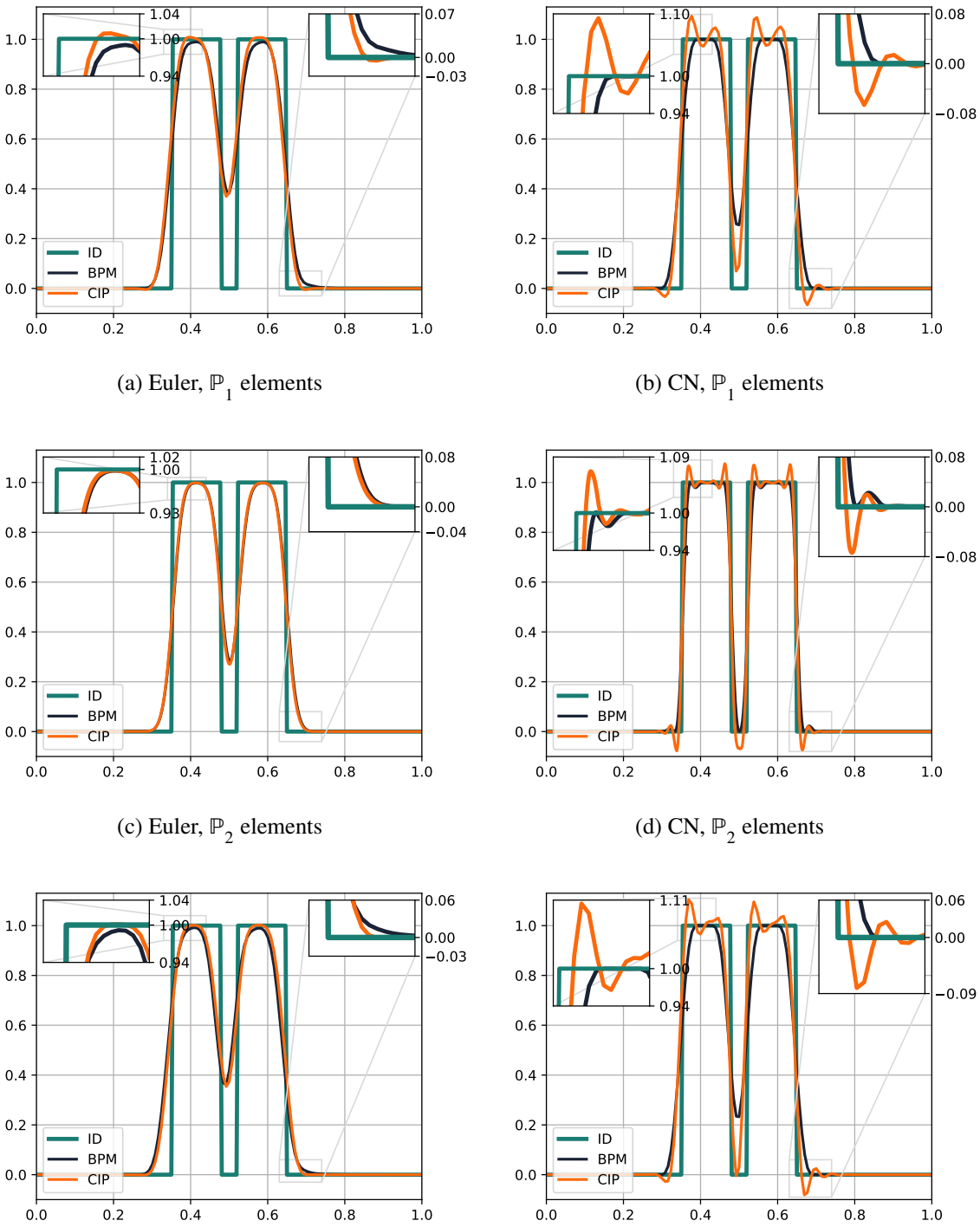


(e) BP-Euler-$\mathbb{Q}_1$ elements



(f) BP-CN-$\mathbb{Q}_1$ elements

Figure 3.6: The approximation of the solution of Example 2 for BP-Euler method and BP-CN method at $T = 6.28$ ($\gamma = 0.001$, $P = 130$).



(a) Euler, $\mathbb{P}_1$ elements

(b) CN, $\mathbb{P}_1$ elements

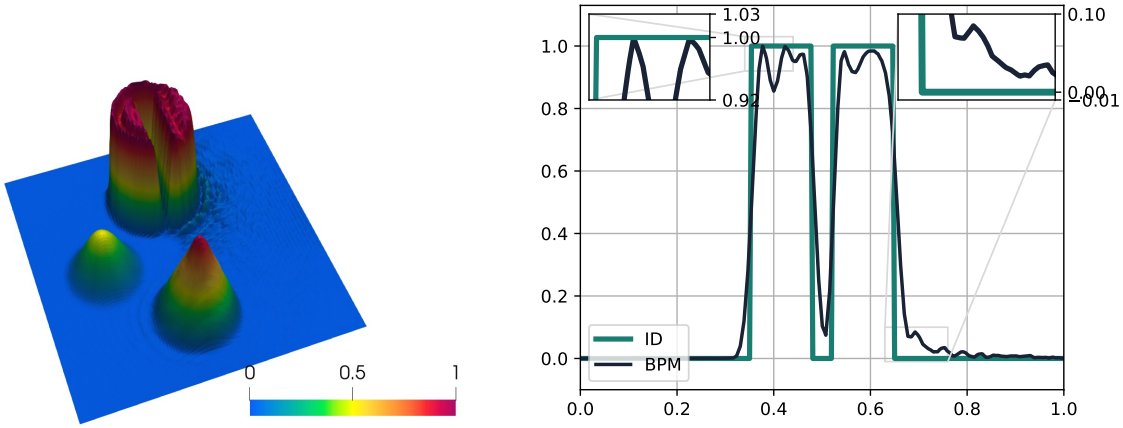
(c) Euler, $\mathbb{P}_2$ elements

(d) CN, $\mathbb{P}_2$ elements

(e) Euler, $\mathbb{Q}_1$ elements

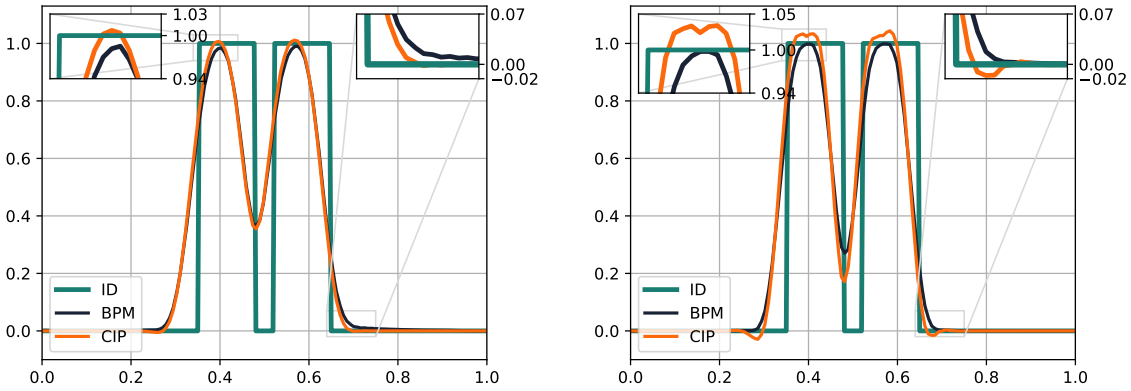
(f) CN, $\mathbb{Q}_1$ elements

Figure 3.7: Cross sections were taken along the line $y = 0.75$ of Initial data (ID) u^0, BP-Euler, BP-CN, CIP-Euler and CIP-CN methods at $T = 6.28$ ($\gamma = 0.001$, $P = 130$). For plotting these cross-sections, when $\mathbb{P}_2$ elements are used 10,000 equidistant points were chosen along the line $y = 0.75$, and the values of the approximated solution have been plotted at these points.



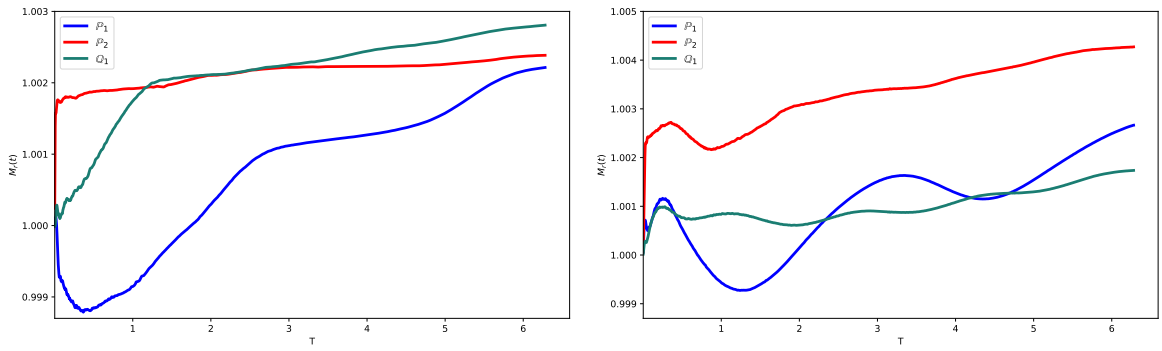
(a) CN, $\mathbb{P}_1$ elements (b) CN, $\mathbb{P}_1$ elements

Figure 3.8: **Left:** The approximation of the solution of Example 2 for BP-CN method without CIP term ($\gamma = 0$) using $\mathbb{P}_1$ elements and mesh 3.1a ($P = 130$) **Right:** Cross sections were taken along the line $y = 0.75$ of Initial data u^0 and BP-CN method without CIP term at $T = 6.28$.



(a) Euler, $\mathbb{P}_1$ elements (b) CN, $\mathbb{P}_1$ elements

Figure 3.9: Cross sections were taken along the line $y = 0.75$ of Initial data u^0, BP-Euler, BP-CN, CIP-Euler and CIP-CN methods at $T = 6.28$ ($\gamma = 0.001$, $P = 130$) on the non-Delaunay mesh 3.1c.



(a) Euler (b) CN

Figure 3.10: The evolution of mass over time employing the BP-Euler and BP-CN schemes. These methods have been implemented with $\mathbb{P}_1$ and $\mathbb{P}_2$ elements on Mesh 3.1a, and $\mathbb{Q}_1$ elements on Mesh 3.1b.

Chapter 4

Bound-preserving composite discontinuous Galerkin method on polytopic meshes

4.1 Introduction

In this chapter, we extend the bound-preserving finite element method to polytopic meshes, where the set of degrees of freedom remains *independent* of the number of vertices, edges, or faces within each element.

Over the past years, there has been growing interest in the development of Galerkin-type numerical methods for meshes composed of general polygons in two dimensions and general polyhedra in three dimensions, collectively referred to as *polytopic* meshes. Unlike classical Galerkin and discontinuous Galerkin approaches that rely on simplicial or structured box-type meshes, these methods provide greater flexibility in mesh design. A key motivation behind this shift is the potential reduction in the overall number of degrees of freedom needed to solve PDE problems efficiently.

This aspect is particularly relevant in adaptive computations for evolution PDEs, where dynamic mesh modification plays a crucial role in reducing computational costs. Such techniques are widely utilised in both Eulerian and Lagrangian frameworks to enhance efficiency. Additionally, numerical methods on polytopic meshes have found applications in problems involving interfaces—such as porosity distributions and material discontinuities—as well as in multilevel solvers for elliptic boundary-value problems, where they contribute to coarse correction strategies.

Popular approaches for polytopic meshes include the virtual element method [21, 122], which originates from the development of mimetic finite difference methods [54], as well as polygonal finite element methods [113], composite finite element methods [71, 107], and various discontinuous Galerkin (dG) formulations.

The latter range from one-field interior penalty dG methods [39–41, 55] to hybridised approaches [50, 51].

A key advantage of dG methods is that they allow for independent control over the global numerical degrees of freedom, irrespective of the mesh topology (i.e., the connectivity of nodes, faces, and elements). In contrast, polygonal finite element and virtual element methods enforce conformity by constructing approximation spaces that inherently depend on the mesh topology. More specifically, even for the lowest-order cases, these methods require a number of basis functions proportional to the number of mesh nodes, which may limit the potential computational efficiency when using polytopic elements with a large number of faces.

In this chapter, inspired by the work in [57], we aim to extend the bound-preserving method to polytopic meshes. The approach introduced in [57] developed a recovered finite element method for polygonal elements, constructing *conforming* schemes over polytopic meshes. As mentioned earlier, since the set of degrees of freedom on polytopic meshes is *independent* of the number of vertices, edges, or faces of each element, we instead select a sub-triangulation of the polytopic meshes. To implement the bound-preserving method, we then impose the bounds at each degree of freedom of this sub-triangulation.

The remainder of this chapter is organised as follows: In Section 4.2, we introduce the elliptic problem under consideration, an overview of the finite element and discontinuous Galerkin methods. Section 4.3 presents the finite element method (FEM) and establishes the well-posedness of the problem. In Section 4.4 and 4.5, we derive the optimal error estimate for the FEM solution. Finally, in Section 4.6, we propose a way for implementation of this finite element method and finally in the last Section 4.7, we evaluate the practical performance of the proposed FEM through a series of numerical experiments.

4.2 Model problem and its discretisation by the discontinuous Galerkin method

Let Ω be an open bounded polygonal/polyhedral domain in $\mathbb{R}^d$ ($d = 2, 3$) with boundary $\partial\Omega$. For given $f \in H^{-1}(\Omega)$, we consider the elliptic problem:

$$\begin{cases} -\mathrm{div}(\mathcal{D}\nabla u) + \mu u &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \partial\Omega, \end{cases} \tag{4.1}$$

here, $\mu \in L^\infty(\Omega)$ satisfies $\mu \geq 0$ a.e, and $\mathcal{D} = (d_{ij})_{i,j=1}^d \in [L^\infty(\Omega)]^{d\times d}$ is a symmetric, uniformly strictly positive definite matrix a.e. in Ω. That is, there exists a constant $\mathcal{D}_0 > 0$ such that for almost all $\boldsymbol{x} \in \Omega$ and for all $\boldsymbol{y} \in \mathbb{R}^d \setminus \{0\}$, we have:

$$\boldsymbol{y}^\top \mathcal{D}\boldsymbol{y} \geq \mathcal{D}_0 \boldsymbol{y}^\top \boldsymbol{y}.$$

By the Lax-Milgram Lemma 1.2.3, the above problem admits a unique weak solution. We note that the framework presented here can be extended to more general settings, such as mixed boundary conditions or the inclusion of first-order terms in the elliptic equation. However, for the sake of clarity, we restrict our focus to this formulation to highlight the key ideas.

In this section, following the discussion in Section 1.6.5 and the results in (1.55) and (1.56), which are derived from the maximum principle in Theorem 1.5.4 and the comparison principles in Corollary 1.5.3, we apply the Assumption (A1) to the solution of (4.1), namely, we assume $0 \leq u(\boldsymbol{x}) \leq \kappa$ for almost every $\boldsymbol{x} \in \Omega$.

Remark 4.2.1. *Including convection in the problem* (4.1) *introduces additional challenges, particularly due to the need for proper treatment of inflow boundary conditions. These issues become more pronounced when using discontinuous Galerkin methods. Therefore, in this chapter, we focus exclusively on reaction–diffusion problems. The analysis and numerical study of convection–diffusion equations will be addressed in future work.*

4.2.1 Finite element spaces

Let $\mathcal{P}$ be a subdivision of Ω into disjoint polygonal elements for $d = 2$ or polyhedral elements for $d = 3$. From now on, we will refer to both cases as polytopic elements. Below, we outline some mild assumptions on the admissible geometry of these elements.

For a nonnegative integer k, we denote by $\mathbb{P}_k(K)$ the set of all polynomials of order k on each element $K \in \mathcal{P}$. For $k \geq 1$, we consider the element-wise discontinuous space

$$V_{\mathcal{P}} := \{v_H \in L^2(\Omega) : v_H|_K \in \mathbb{P}_k(K) \quad \forall K \in \mathcal{P}\}. \tag{4.2}$$

We denote by $\Gamma_{\mathcal{P}}$ the skeleton of $\mathcal{P}$, defined as $\Gamma_{\mathcal{P}} := \bigcup_{K\in\mathcal{P}} \partial K$, and the interior skeleton as $\Gamma_{\mathcal{P}}^{\text{int}} := \Gamma_{\mathcal{P}} \backslash \partial\Omega$.

Also, we define the mesh function $H_{\mathcal{P}} : \bar{\Omega} \to \mathbb{R}_+$ for the polytopic mesh $\mathcal{P}$, where on the interior of each polytopic element, it is given by $H_{\mathcal{P}}|_{\mathring{K}} := \text{diam}(K)$.

Assumption 4.2.2. *We assume that the partition $\mathcal{P}$ can be further subdivided into a conforming, shape-regular, and quasi-uniform simplicial triangulation $\mathcal{T}$* on each $K \in \mathcal{P}$*, such that for every $K \in \mathcal{P}$, we have*

$$K = \bigcup_{T\in\mathcal{T}, T\subset K} T.$$

For convenience, we denote by $\mathcal{T}_K$ the set of simplicial elements of $\mathcal{T}$ contained within the polytopic element $K \in \mathcal{P}$, i.e.,

$$\mathcal{T}_K := \{T \in \mathcal{T} : T \subset K\}.$$

Furthermore, on each $K \in \mathcal{P}$, we consider the local piecewise polynomial space

$$W_K := \{v_h \in C^0(K) : v_h|_T \in \mathbb{P}_k(T),\ T \in \mathcal{T}_K\}, \tag{4.3}$$

and let $W_{\mathcal{T}} := \oplus_{K\in\mathcal{P}} W_K$.

Although the solution is represented by a function belonging to the finite element space $W_{\mathcal{T}}$, the test space is chosen from $V_{\mathcal{P}}$, which is defined on the polygonal mesh $\mathcal{P}$. In other words, in the finite element method presented below, the solution is interpolated in $W_{\mathcal{T}}$, while the resulting algebraic system is formed by testing against functions in $V_{\mathcal{P}}$. Consequently, the total number of global numerical degrees of freedom in the final system corresponds to the dimension of $V_{\mathcal{P}}$, rather than that of the finer space $W_{\mathcal{T}}$.

4.2.2 Interior penalty discontinuous Galerkin method

Let K_+ and K_- be two adjacent elements of $\mathcal{P}$ that share an interface $F \subset \partial K_+ \cap \partial K_- \subset \Gamma_{\mathcal{P}}^{\text{int}}$. For an element-wise continuous scalar function v and vector-valued function $\mathbf{q}$, we define the *weighted average* across F as

$$\{\!\{v\}\!\}|_F := \frac{1}{2}\big(v_+|_F + v_-|_F\big), \quad \{\!\{\mathbf{q}\}\!\}|_F := \frac{1}{2}\big(\mathbf{q}_+|_F + \mathbf{q}_-|_F\big),$$

where $v_\pm|_F$ represents the trace of v from within $K_\pm$ on F, and similarly for $\mathbf{q}$. Moreover, the *jump* across F is defined as

$$[\![v]\!]|_F := v_+\mathbf{n}_+ + v_-\mathbf{n}_-, \quad [\![\mathbf{q}]\!]|_F := \mathbf{q}_+ \cdot \mathbf{n}_+ + \mathbf{q}_- \cdot \mathbf{n}_-,$$

where $\mathbf{n}_\pm$ are the unit outward normals of $K_\pm$ on F. On a boundary face $F \subset \partial\Omega \cap \partial K$, we set $\{\!\{v\}\!\} := v$, $\{\!\{\mathbf{q}\}\!\} := \mathbf{q}$, $[\![v]\!] := v\mathbf{n}$, and $[\![\mathbf{q}]\!] := \mathbf{q} \cdot \mathbf{n}$, respectively, with $\mathbf{n}$ the unit outward normal to $\partial\Omega$.

Finally, for brevity, we denote by $\nabla_{\mathcal{P}} v$ the broken gradient of a function $v : \Omega \to \mathbb{R}$ with $v|_K \in H^1(K)$, $K \in \mathcal{P}$, defined element-wise by $(\nabla_{\mathcal{P}} v)|_K = \nabla(v|_K)$, $K \in \mathcal{P}$.

The (classical) interior penalty discontinuous Galerkin (IPDG) method reads: find $u_H \in V_{\mathcal{P}}$, such that

$$a_{DG}(u_H, v_H) = \ell(v_H), \qquad \text{for all } v_H \in V_{\mathcal{P}}, \tag{4.4}$$

with $a_{DG} : (H_0^1(\Omega) + V_{\mathcal{P}}) \times (H_0^1(\Omega) + V_{\mathcal{P}}) \to \mathbb{R}$, given by

$$\begin{aligned} a_{DG}(u_H, v_H) := & \int_\Omega \left(\mathcal{D}\nabla_{\mathcal{P}} u_H \cdot \nabla_{\mathcal{P}} v_H + \mu u_H v_H \right) \mathrm{d}x + \int_{\Gamma_{\mathcal{P}}} \sigma_{\mathcal{P}} [\![u_H]\!] \cdot [\![v_H]\!] \, \mathrm{d}s \\ & - \int_{\Gamma_{\mathcal{P}}} \left(\{\!\{ \mathcal{D}\nabla u_H \}\!\} \cdot [\![v_H]\!] + \theta \{\!\{ \mathcal{D}\nabla v_H \}\!\} \cdot [\![u_H]\!] \right) \mathrm{d}s, \end{aligned} \tag{4.5}$$

and $\ell(v_H) := \langle f, v_H \rangle_\Omega$, with $\theta \in [-1, 1]$ and $\sigma_{\mathcal{P}} : \Gamma \to \mathbb{R}_{\geq 0}$ the, so-called, *discontinuity-penalisation* or *penalty* function, whose precise definition depends on the assumptions on the geometry of the polytopic elements K; this will be discussed below. The choice $\theta = 1$ yields the *symmetric* version, while the choice $\theta = -1$ the non-symmetric version of the IPDG method.

The choice of the discontinuity-penalisation function $\sigma_{\mathcal{P}}$ has been widely investigated, particularly for meshes consisting of extremely general polytopic or even curved elements; see [37,38,42,43,56] for further discussion. In this work, we impose stronger restrictions on the admissible polytopic meshes $\mathcal{P}$ compared to these studies, owing to the two-scale nature of the positivity-preserving method introduced below.

Assumption 4.2.3 (Admissible polytopic meshes)**.** *Let $(\mathcal{P}_i)_{i \in I}$ be a family of polytopic meshes, indexed by some set I. We impose the following conditions:*

(a) each element K in $(\mathcal{P}_i)_{i \in I}$ is star-shaped with respect to an inscribed ball of radius ρ_K. Moreover, there exists a global constant $C_{\mathrm{star}} > 0$, independent of the mesh, such that $H_K \leq C_{\mathrm{star}} \rho_K$ for all $K \in \mathcal{P}_i$ and for all $i \in I$.

(b) The meshes $(\mathcal{P}_i)_{i \in I}$ satisfy a local quasi-uniformity condition, meaning there exists a constant $C_{\mathrm{qu}} \geq 1$, independent of the mesh, such that for any polytopic element $K \in \mathcal{P}_i$ and its face-neighbor K', the element diameters satisfy $H_{K'} \leq C_{\mathrm{qu}} H_K$, for all $K \in \mathcal{P}_i$ and for all $i \in I$.

Note that Assumption 4.2.3(b) implies the equivalence

$$C_{\mathrm{qu}}^{-1} H_K \leq H_{K'} \leq C_{\mathrm{qu}} H_K$$

for all face-neighboring elements K' of a polytopic element $K \in \mathcal{P}$.

Remark 4.2.4. *By the subtriangulation assumption stated below, each polygonal element K is subdivided into a shape-regular simplicial submesh. This requirement implies that the faces of K must also be of reasonable size and shape. In particular, very small faces on the boundary of the polygonal elements are not allowed.*

For any $v_H \in (H_0^1(\Omega) + V_{\mathcal{P}})$, we define the DG-norm

$$|||v_H|||_{\mathcal{P}} := \left(\|\sqrt{\mathcal{D}}\nabla_{\mathcal{P}} v_H\|_{0,\Omega}^2 + \|\sqrt{\mu} v_H\|_{0,\Omega}^2 + \|\sqrt{\sigma_{\mathcal{P}}}[\![v_H]\!]\|_{0,\Gamma_{\mathcal{P}}}^2 \right)^{1/2}.$$

Using this norm, we establish the following coercivity result for $a_{DG}(\cdot,\cdot)$.

Lemma 4.2.5. *[3, Lemma 2.4] Let $F \subset \partial K_+ \cap \partial K_-$ be a generic planar simplicial face shared by two elements $K_+, K_- \in \mathcal{P}$. If F is not simplicial, we further partition it into a union of simplices and denote each resulting simplicial (sub)face as F. When $F \subset \partial\Omega$, we set $K_- = \emptyset$. We define*

$$\sigma_{\mathcal{P}}|_F := 8C_{\text{star}} k(k-1+d) d^{-1} \max_{*\in\{+,-\}} \delta_{K_*} H_{K_*}^{-1}, \tag{4.6}$$

with $\delta_{K_*} := \|\mathcal{D}\|_{0,\infty,K_*}^2 \|\mathcal{D}^{-1}\|_{0,\infty,K_*}$, for each simplicial (sub)face $F \subset \Gamma_{\mathcal{P}}$, we have

$$a_{DG}(v,v) \geq \frac{3}{4}|||v|||_{\mathcal{P}}^2 \qquad \text{for all } v \in V_{\mathcal{P}}. \tag{4.7}$$

Proof. For $v \in V_{\mathcal{P}}$, we have

$$a_{DG}(v,v) \geq |||v|||_{\mathcal{P}}^2 - 2\|\sigma_{\mathcal{P}}^{-1/2}\{\!\{\mathcal{D}\nabla v\}\!\}\|_{0,\Gamma_{\mathcal{P}}} \|\sqrt{\sigma_{\mathcal{P}}}[\![v]\!]\|_{0,\Gamma_{\mathcal{P}}}. \tag{4.8}$$

We prove the result for interior simplicial (sub)faces; the case of boundary (sub)faces is a special case if we set $K_- = \emptyset$. On each planar simplicial interior (sub)face $F \subset \Gamma_{\mathcal{P}}^{\text{int}}$, standard estimation and a trace-inverse inequality of the form $\|v_h\|_{0,F}^2 \leq \frac{(k+1)(k+d)|F|}{d|T|}\|v_h\|_{0,T}^2$ for $v_h \in \mathbb{P}_k(T)$ for F face of a simplex T, (see [117], and also [38] for an extension,) imply

$$\begin{aligned}
\|\sigma_{\mathcal{P}}^{-1/2}\{\!\{\mathcal{D}\nabla v\}\!\}\|_{0,F}^2 &\leq \frac{1}{2\sigma_{\mathcal{P}}} \sum_{*\in\{+,-\}} \|\mathcal{D}|_{K_*}\|_{0,\infty,F}^2 \|\nabla v|_{K_*}\|_{0,F}^2 \\
&\leq \frac{1}{2\sigma_{\mathcal{P}}} \sum_{*\in\{+,-\}} \|\mathcal{D}|_{K_*}\|_{0,\infty,F}^2 \frac{k(k-1+d)|F|}{d|K_*^F|} \|\nabla v|_{K_*}\|_{0,K_*^F}^2
\end{aligned}$$

respectively, since $\nabla v \in [\mathbb{P}_{k-1}(K_*)]^d$, upon deciding that $\sigma_{\mathcal{P}}$ is constant on each simplicial (sub)face, whereby $K_*^F \subset K_*$, $*\in\{+,-\}$, denotes the simplex with face F and vertex the centre of the ball with respect to which K_* is star-shaped. Note that the simplices K_*^F are disjoint by construction and we have $\cup_{F\subset\partial K} K_*^F = K_*$. Therefore, since $\frac{|K_*^F|}{|F|} := H_F^{\perp}$ is the 'height' of the simplex with base F, and we have, by construction

that $H_F^{\perp} \geq \rho_K$, we deduce

$$\frac{|F|}{|K_*^F|} \leq \rho_K^{-1} \leq C_{star} H_{K_*}^{-1}.$$

Therefore, we can arrive at

$$\|\sigma_{\mathcal{P}}^{-1/2}\{\{\mathcal{D}\nabla v\}\}\|_{0,\Gamma_{\mathcal{P}}}^2 \leq \frac{C_{\text{star}}}{2} \sum_{F\subset\Gamma_{\mathcal{P}}} \sigma_{\mathcal{P}}^{-1}|_F \sum_{*\in\{+,-\}} \|\mathcal{D}|_{K_*}\|_{0,\infty,F}^2 \|\mathcal{D}^{-1}\|_{0,\infty,K_*} \frac{k(k-1+d)}{d H_{K_*}} \|\sqrt{\mathcal{D}}\nabla v\|_{0,K_*^F}^2.$$

Selecting, now, $\sigma_{\mathcal{P}}$ as in (4.6), we deduce

$$\|\sigma_{\mathcal{P}}^{-1/2}\{\{\mathcal{D}\nabla v\}\}\|_{0,\Gamma_{\mathcal{P}}}^2 \leq 16^{-1} \sum_{F\subset\Gamma_{\mathcal{P}}} \sum_{*\in\{+,-\}} \|\sqrt{\mathcal{D}}\nabla v\|_{0,K_*^F}^2.$$

Then, upon observing that

$$\sum_{F\subset\partial K} \|\sqrt{\mathcal{D}}\nabla v\|_{0,K_*^F}^2 = \|\sqrt{\mathcal{D}}\nabla v\|_{0,K_*}^2,$$

since $\cup_{F\subset\partial K} K_*^F = K_*$, we arrive at the bound

$$\|\sigma_{\mathcal{P}}^{-1/2}\{\{\mathcal{D}\nabla v\}\}\|_{0,\Gamma_{\mathcal{P}}}^2 \leq 16^{-1} \|\sqrt{\mathcal{D}}\nabla_{\mathcal{P}} v\|_{0,\Omega}^2, \tag{4.9}$$

which, combined with (4.8), gives the result. □

Note that the above (classical) IPDG method on polytopic meshes involves the same number of elemental basis functions on each element, irrespectively of its shape. The key property allowing to achieve this is that the local basis functions are defined on by restricting polynomials defined the physical space to each element, and *not* through element mappings. We refer to [42] for details on the implementation of the method.

4.3 A nodally bound-preserving composite discontinuous Galerkin method

The classical IPDG method described above is generally *not* bound-preserving. This means that even if the exact solution satisfies $u(\boldsymbol{x}) \in [0, \kappa]$ for almost all $\boldsymbol{x} \in \Omega$, the numerical solution u_H may exceed these bounds in a region with positive d-dimensional measure. The use of polytopic elements adds another level of complexity to this issue.

However, polytopic elements offer several advantages in numerical simulations. They allow for an exact representation of complex geometries without requiring overly refined meshes, as long as the solution remains locally smooth. Also, these meshes can improve computational efficiency by reducing the total num-

ber of global degrees of freedom. This flexibility makes polytopic elements particularly useful in adaptive methods for handling singularities, sharp gradients/layers, and layer structures more effectively.

With this in mind, using polytopic elements with large diameters is particularly useful, as they allow the solution to be represented over larger regions of the computational domain while reducing the number of numerical degrees of freedom. However, modifying the polynomial basis on these large elements to enforce bound-preservation could introduce larger errors. Additionally, in the IPDG framework, local polynomial spaces are not constructed using nodal basis functions mapped from a reference domain, making it unclear how to directly adjust nodal values to keep the solution within the bounds of the solution.

To address all these challenges at once, we propose enforcing bound preservation on the nodal basis functions associated with the simplicial sub-mesh $\mathcal{T}$. We start by noting that the number of Lagrange basis functions (and nodes) of order k in d dimensions for a simplicial element is given by $m_{k,d} := \binom{k+d}{d}$. We consider the Lagrange basis functions $\{\phi_i^T\}_{i=1}^{m_{k,d}}$ corresponding to the Lagrange nodes $\{\boldsymbol{x}_i^T\}_{i=1}^{m_{k,d}}$ for a simplex $T \subset K$, where $K \in \mathcal{P}$. Then, defining $C(K)$ as the space of continuous functions over K, we introduce the element-wise nodally bound-preserving recovery operator $\mathcal{E}_K^+ : C(K) \to W_K$, given by

$$\mathcal{E}_K^+(v) = \sum_{T\in\mathcal{T}_K} \sum_{i=1}^{m_{k,d}} \max\left\{0, \min\{v(\boldsymbol{x}_i^T), \kappa\}\right\}\phi_i^T. \tag{4.10}$$

Thus, by construction, $\mathcal{E}_K^+(v) \in W_K^+$, with the cone W_K^+ defined by

$$W_K^+ := \{v \in W_K \,:\, v(\boldsymbol{x}_i^T) \in [0,\kappa],\;\; i = 1,\dots,m_{k,d}, T \in \mathcal{T}_K\}. \tag{4.11}$$

In other words, $\mathcal{E}_K^+(v)$ is specifically constructed to stay within the predefined range $[0,\kappa]$ at each node $\boldsymbol{x}_i^T$, for $i = 1,\dots,m_{k,d}$ and $T \in \mathcal{T}_K$. We further define the global nodally bound-preserving recovery operator $\mathcal{E}^+ : C(\Omega) \to W_{\mathcal{T}}$ as

$$\big(\mathcal{E}^+(v)\big)|_K := \mathcal{E}_K^+(v|_K), \quad K \in \mathcal{P}.$$

A few remarks about this construction are important. Let $v_H \in V_{\mathcal{P}}$. If $v_H(\boldsymbol{x}_i^T)$ already lies within $[0,\kappa]$ for all $i = 1,\dots,m_{k,d}$ and $T \in \mathcal{T}_K$, then $\mathcal{E}_K^+(v_H) = v_H$ on K, since it is a polynomial of degree k in K. This means that the recovery operator $\mathcal{E}^+$ does not effect functions that are already within the required bounds. Correspondingly, if $v_H(\boldsymbol{x}_i^T) \in [0,\kappa]$, for all $i = 1,\dots,m_{k,d}$, $T \in \mathcal{T}_K$, and for all $K \in \mathcal{P}$, we have $\mathcal{E}^+(v_H) = v_H$. On the other hand, if at least for one point $\boldsymbol{x}_i^T$, we have $v_H(\boldsymbol{x}_i^T) \notin [0,\kappa]$, then $\mathcal{E}_K^+(v_H) \neq v_H$ on K. Finally, we note that if $v_H(\boldsymbol{x}_i^T) < 0$, for all $i = 1,\dots,m_{k,d}$, $T \in \mathcal{T}_K$, for some $K \in \mathcal{P}$, we get

$\mathcal{E}_K^+(v_H) = 0$, i.e., the (nonlinear) map $\mathcal{E}^+$ has a non-trivial kernel.

For notational convenience, we also set

$$\mathcal{E}_K^-(v_H) := v_H|_K - \mathcal{E}_K^+(v_H), \tag{4.12}$$

for $K \in \mathcal{P}$ and, correspondingly, $\mathcal{E}^-(v_H) := v_H - \mathcal{E}^+(v_H)$.

To alleviate the presence of the non-trivial kernel, we define a *stabilisation* bilinear form as follows: for $w_h, v_h \in W_{\mathcal{T}}$, we set

$$s(w_h, v_h) = \sum_{K\in\mathcal{P}} \alpha \sum_{T\in\mathcal{T}_K} \sum_{i=1}^{m_{k,d}} \left(\mathcal{D}_{\omega_K} h_T^{d-2} + \mu_T h_T^d\right) w_h(\boldsymbol{x}_i^T) v_h(\boldsymbol{x}_i^T)\,, \tag{4.13}$$

with $h_T := \operatorname{diam}(T)$ and $\mu_T := \|\mu\|_{0,\infty,T}$, $T \in \mathcal{T}$, while $\mathcal{D}_{\omega_K} := \|D\|_{0,\infty,\omega_K}$, for $\omega_K := \{K' \in \mathcal{P} \text{ shares face with } K\}$, for each $K \in \mathcal{P}$, for some piecewise constant function α, with $\alpha|_K := \alpha_K > 0$, $K \in \mathcal{P}$, to be determined below. The bilinear form $s(\cdot,\cdot)$ induces the norm $\|v_h\|_s := \sqrt{s(v_h, v_h)}$ in $W_{\mathcal{T}}$.

We are now ready to introduce the composite bound-preserving discontinuous Galerkin method, which reads: find $u_H \in V_{\mathcal{P}}$ such that

$$a_h(u_H; v_H) = \ell(v_H) \qquad \forall v_H \in V_{\mathcal{P}}\,, \tag{4.14}$$

with the semilinear form given by

$$a_h(u_H; v_H) := a_{DG}(\mathcal{E}^+(u_H), v_H) + s(\mathcal{E}^-(u_H), v_H)\,. \tag{4.15}$$

We observe that, if $u_H(\boldsymbol{x}_i^T) \in [0, \kappa]$, for all $i = 1, \ldots, m_{k,d}$, $T \in \mathcal{T}$, then (4.14) is just the (classical) interior penalty discontinuous Galerkin method (4.4) with solution $u_H \in V_{\mathcal{P}}$, i.e, the non-linearity in the first argument of $a_h(\cdot,\cdot)$ disappears. The stabilisation term $s(\cdot,\cdot)$ is non-trivial for polytopic elements K on which u_H violates the predetermined bounds on the range of the numerical solution.

Before discussing the well-posedness, stability, and convergence of the bound-preserving composite discontinuous Galerkin method above, we make some further assumptions on the sub-mesh $\mathcal{T}$ and its relationship with the polytopic mesh $\mathcal{P}$.

Assumption 4.3.1 (admissible simplicial submeshes)**.** *Consider a family of simplicial (sub)meshes $(\mathcal{T}_j)_{j\in J}$, for some index set J, that are constructed as refinements of a given polytopic mesh $\mathcal{P}$. We assume that:*

(a) each simplex T in $(\mathcal{T}_j)_{j\in J}$, is shape-regular, i.e., there exists a global constant $C_{\mathrm{sh}} > 0$, such that $h_T \le C_{\mathrm{sh}}\rho_T$, for all $T \in \mathcal{T}_j$, $j \in J$, with ρ_T denoting the radius of the largest inscribed ball in T.

(b) the diameter h_T of each simplex $T \subset K$, $K \in \mathcal{P}$, having a face f contained in the boundary of the polytopic element K (i.e., $\partial T \cap \partial K$ has positive $(d-1)$-dimensional measure) is smaller than or equal to the diameter $H_{K'}$ of the adjacent to f polytopic element $K' \in \mathcal{P}$, (that is the element $K' \in \mathcal{P}$ with $f \subset \partial K \cap \partial K'$).

We note that (a) in Assumption 4.3.1 is standard, while (b) is a technical (very mild) assumption requiring that the submesh of a polytopic element is not "coarser" than the neighbouring polytopic elements themselves.

Remark 4.3.2. *The mesh admissibility conditions in Assumption 4.2.3 do not exclude the presence of elements with very small faces. Indeed, the size of an individual face has no effect on either the star-shapedness requirement in Assumption 4.2.3(a) or the local quasi-uniformity condition in Assumption 4.2.3(b). If an element K possesses a small face relative to its diameter H_K, the star-shapedness condition ensures that K still admits shape-regular simplicial subtriangulation $\mathcal{T}_j$. In this case, the small face simply forces the subtriangulation $\mathcal{T}_j$ to be locally finer, but it does not violate any part of Assumption 4.2.3.*

The following result shows that $s(\cdot,\cdot)$ indeed controls the kernel of the projection $\mathcal{E}^+(\cdot)$.

Lemma 4.3.3. *[3, Lemma 3.3] There exists a constant $C_{\mathrm{equiv}} > 0$, depending only on C_{star}, C_{sh}, k and on the problem dimension d, such that, for every $v_h \in W_{\mathcal{T}}$, we have*

$$|||v_h|||_{\mathcal{P}}^2 \le C_{\mathrm{equiv}} \|\alpha^{-1/2} v_h\|_s^2. \tag{4.16}$$

Proof. Observing that the number of Lagrange basis functions (and nodes) of order k in d dimensions is given by $m_{k,d} := \binom{k+d}{d}$, we consider the Lagrange basis functions $\{\phi_i^T\}_{i=1}^{m_{k,d}}$, associated with the Lagrange nodes $\{\boldsymbol{x}_i^T\}_{i=1}^{m_{k,d}}$, for the simplex $T \subset K$, where $K \in \mathcal{P}$.

For an affine transformation $F_T : \hat{T} \to T$, mapping a reference element $\hat{T}$ to each $T \in \mathcal{T}$, we introduce the family of Lagrange basis functions $\{\psi_i\}_{i=1}^{m_{k,d}}$ on the reference simplex $\hat{T}$, such that $\phi_i^T \circ F_T = \psi_i$. Then, we obtain

$$\begin{aligned}
\|v_h\|_{0,T}^2 &\le m_{k,d} \sum_{i=1}^{m_{k,d}} v_h^2(\boldsymbol{x}_i^T) \int_T \left(\phi_i^T(\boldsymbol{x})\right)^2 \mathrm{d}\boldsymbol{x} \\
&\le m_{k,d} \max_{i=1,\ldots,m_{k,d}} \|\phi_i^T\|_{0,\infty,T}^2 |T| \sum_{i=1}^{m_{k,d}} v_h^2(\boldsymbol{x}_i^T) = C_{k,d}^{L^2} |T| \sum_{i=1}^{m_{k,d}} v_h^2(\boldsymbol{x}_i^T),
\end{aligned}$$

with $C_{k,d}^{L^2} := m_{k,d} \max_{i=1,\dots,m_{k,d}} \|\psi_i\|_{0,\infty,\hat{T}}^2$, since $\|\psi_i\|_{0,\infty,\hat{T}} = \|\phi_i^T\|_{0,\infty,T}$ for all $T \in \mathcal{T}$.

Using the last bound along with a standard inverse estimate of the form $\|\nabla v_h\|_{0,T}^2 \leq C_{\text{inv}} k^4 \rho_T^{-2} \|v_h\|_{0,T}^2$, with $C_{\text{inv}} > 0$ a constant independent of T, k, v_h, and ρ_T denoting the radius of the largest inscribed ball in T, gives

$$\|\sqrt{D}\nabla v_h\|_{0,T}^2 \leq C_{\text{inv}} k^4 \|D\|_{0,\infty,T} \rho_T^{-2} \|v_h\|_{0,T}^2 \leq C_{\text{inv}} C_{k,d}^{L^2} k^4 \|D\|_{0,\infty,T} \rho_T^{-2} |T| \sum_{i=1}^{m_{k,d}} v_h^2(\boldsymbol{x}_i^T).$$

Since $|T| \leq h_T^d$, from Assumption 4.3.1 we get $|T| \leq C_{\text{sh}}^2 \rho_T^2 h_T^{d-2}$ and, thus, we conclude

$$\|\sqrt{D}\nabla v_h\|_{0,T}^2 \leq C_{k,d}^{H^1} \|D\|_{0,\infty,T} h_T^{d-2} \sum_{i=1}^{m_{k,d}} v_h^2(\boldsymbol{x}_i^T), \tag{4.17}$$

with $C_{k,d}^{H^1} := C_{\text{inv}} C_{\text{sh}}^2 C_{k,d}^{L^2} k^4$.

Finally, let $f \subset F \subset K_- \cap K_+$ (as before), where $K_-, K_+ \in \mathcal{P}$. Using the trace–inverse estimate (see [41, Chapter 3, Lemma 6])

$$\|v_h\|_{0,F}^2 \leq \frac{(k+1)(k+d)\,|F|}{d\,|S|} \|v_h\|_{0,S}^2,$$

which holds for each face $F \subset \partial S$ of a simplex S (F lies in a generic planar simplicial face shared by two elements $K_+, K_- \in \mathcal{P}$), we proceed as follows. Consider a face $f \subset \partial T \cap \partial K$ for a polytopic element K. Then f is a (sub)face of the simplicial element T, and it is contained in the corresponding simplicial face F of the polytopic element K that contains T. Hence, the above trace–inverse estimate can be applied on f through its associated face F of K, therefore we have

$$\begin{aligned}
\|\sqrt{\sigma_{\mathcal{P}}} v_h\|_{0,f}^2 &\leq m_{k,d} \sigma_{\mathcal{P}} \sum_{i=1}^{m_{k,d}} v_h^2(\boldsymbol{x}_i^T) \int_f \left(\phi_i^T(\boldsymbol{x})\right)^2 \mathrm{d}s \\
&\leq 4 C_{\text{star}} m_{k,d} \max_{i=1,\dots,m_{k,d}} \|\psi_i\|_{0,\infty,\hat{T}} k(k-1+d) d^{-1} \|D\|_{0,\infty,K_- \cup K_+} \Big(\min_{*\in\{-,+\}} H_{K_*} \Big)^{-1} |f| \sum_{i=1}^{m_{k,d}} v_h^2(\boldsymbol{x}_i^T).
\end{aligned}$$

Upon considering only the nodes and the respective basis functions which are nontrivial on the given face, which is a $(d-1)$-dimensional simplex itself and, hence, the numbering is modified accordingly. From Assumption 4.3.1(b), we have that $h_T \leq \min_{*\in\{-,+\}} H_{K_*}$ and, since $|f| \leq h_T^{d-1}$, we conclude

$$\|\sqrt{\sigma_{\mathcal{P}}} v_h\|_{0,f}^2 \leq C_{k,d}^{\sigma} \|D\|_{0,\infty,K_- \cup K_+} h_T^{d-2} \sum_{i=1}^{m_{k,d}} v_h^2(\boldsymbol{x}_i^T), \tag{4.18}$$

with $C^{\sigma}_{k,d} := 4d^{-1}C_{\mathrm{star}}m_{k,d}k(k-1+d)\max_{i=1,\dots,m_{k,d}}\|\psi_i\|_{0,\infty,\hat{T}}$.

For convenience in the implementation of the stabilisation term, we bound each term further. To that end, we recall the notation set $\mathcal{D}_{\omega_K}$, and of μ_T, from the definition of the bilinear form $s(\cdot,\cdot)$ and we combine the above estimates, to deduce

$$|||v_h|||^2_{\mathcal{P}} \leq C_{\mathrm{equiv}} \sum_{K\in\mathcal{P}} \sum_{T\in\mathcal{T}_K} \sum_{i=1}^{m_{k,d}} \left(\mathcal{D}_{\omega_K} h_T^{d-2} + \mu_T h_T^d\right) v_h^2(\boldsymbol{x}_i^T),$$

with $C_{\mathrm{equiv}} := \max\{C^{H^1}_{k,d}, 2C^{\sigma}_{k,d}\}$, since $C^{L^2}_{k,d} \leq C^{H^1}_{k,d}$. The result already follows from the definition of $\|\cdot\|_s$. □

4.3.1 Well-posedness

We now discuss the existence and uniqueness of solutions to (4.14). As in the previous chapters, the first step we state the monotonicity result for $s(\cdot,\cdot)$. The proof of this lemma is identical as Lemma 1.6.22 (see [12, Lemma 3.1]) and is therefore omitted for brevity.

Lemma 4.3.4. *The bilinear form $s(\cdot,\cdot)$, defined in* (4.13)*, satisfies:*

$$s(\mathcal{E}^-(v_H) - \mathcal{E}^-(w_H), \mathcal{E}^+(v_H) - \mathcal{E}^+(w_H)) \geq 0 \qquad \forall v_H, w_H \in V_{\mathcal{P}}, \tag{4.19}$$

$$s(\mathcal{E}^-(v_H), w_h - \mathcal{E}^+(v_H)) \leq 0 \qquad \forall v_H \in V_{\mathcal{P}}, w_h \in W^+_{\mathcal{T}}. \tag{4.20}$$

This monotonicity result will be used below to prove the well-posedness of (4.15).

Next, we prove the coercivity and continuity properties for the stabilised bilinear form $a_h(\cdot,\cdot)$ under a specific choice of trial and test functions, which will be essential in the subsequent analysis. A key aspect of the proof is that the discontinuity-penalisation parameter $\sigma_{\mathcal{P}}$, defined in (4.6), is *sufficient* to ensure stability, even when the chosen trial functions are polynomials on the submesh $\mathcal{T}$.

Lemma 4.3.5. *[3, Lemma 3.5] Let $v_H, w_H \in V_{\mathcal{P}}$, and set $r_h^{\pm} := \mathcal{E}^{\pm}(v_H) - \mathcal{E}^{\pm}(w_H)$ for brevity. Define the set*

$$\mathcal{T}_K^{\partial} := \{T \in \mathcal{T}_K : \exists\, \text{face } f \subset \partial T \cap \partial K\}$$

of simplices T in K touching the boundary of K, select $\alpha > 0$ in (4.13)*, such that*

$$\alpha|_K := \gamma \max\left\{1, \frac{H_K}{8C_{\mathrm{star}}} \max_{T\in\mathcal{T}_K^{\partial}} \frac{|f|}{|T|}\right\}, \quad K \in \mathcal{P}, \tag{4.21}$$

with $\gamma \geq 25C_{\text{equiv}}$. *Then, we have the bound*

$$a_{DG}(r_h^+, z_H) + s(r_h^-, z_H) \leq \frac{7}{4}\Big(|||r_h^+|||_{\mathcal{P}}^2 + \|r_h^-\|_s^2\Big)^{\frac{1}{2}}\Big(|||z_H|||_{\mathcal{P}}^2 + \|z_H\|_s^2\Big)^{\frac{1}{2}}, \tag{4.22}$$

for any $z_H \in V_{\mathcal{P}}$.

Proof. Through standard estimation, we have

$$a_{DG}(r_h^+, z_H) \leq |||r_h^+|||_{\mathcal{P}}|||z_H|||_{\mathcal{P}} - \int_{\Gamma_{\mathcal{P}}} \{\!\{\mathcal{D}\nabla r_h^+\}\!\} \cdot [\![z_H]\!]\,\mathrm{d}s - \int_{\Gamma_{\mathcal{P}}} \{\!\{\mathcal{D}\nabla z_H\}\!\} \cdot [\![r_h^+]\!]\,\mathrm{d}s. \tag{4.23}$$

We continue by estimating the indefinite terms from above. For the last term on the right-hand side of last bound, we employ Cauchy-Schwarz inequality and (4.9) to deduce

$$\Big|\int_{\Gamma_{\mathcal{P}}} \{\!\{\mathcal{D}\nabla z_H\}\!\} \cdot [\![r_h^+]\!]\,\mathrm{d}s\Big| \leq 4^{-1}\|\sqrt{\mathcal{D}}\nabla_{\mathcal{P}} z_H\|_{0,\Omega}\|\sqrt{\sigma_{\mathcal{P}}}[\![r_h^+]\!]\|_{0,\Gamma_{\mathcal{P}}}. \tag{4.24}$$

For the remaining term, we begin by decomposing into two contributions that will, in turn, be estimated using the two different stabilisation terms, i.e., the discontinuity-penalisation and $s(\cdot,\cdot)$, respectively:

$$\int_{\Gamma_{\mathcal{P}}} \{\!\{\mathcal{D}\nabla r_h^+\}\!\} \cdot [\![z_H]\!]\,\mathrm{d}s = \int_{\Gamma_{\mathcal{P}}} \{\!\{\mathcal{D}\nabla(v_H - w_H)\}\!\} \cdot [\![z_H]\!]\,\mathrm{d}s - \int_{\Gamma_{\mathcal{P}}} \{\!\{\mathcal{D}\nabla r_h^-\}\!\} \cdot [\![z_H]\!]\,\mathrm{d}s =: (I) - (II), \tag{4.25}$$

since $r_h^+ + r_h^- = v_H - w_H$. For (I), we employ Cauchy-Schwarz inequality and (4.9) to get

$$\begin{aligned}|(I)| &\leq 4^{-1}\|\sqrt{\mathcal{D}}\nabla_{\mathcal{P}}(v_H - w_H)\|_{0,\Omega}\|\sqrt{\sigma_{\mathcal{P}}}[\![z_H]\!]\|_{0,\Gamma_{\mathcal{P}}}\\ &\leq 4^{-1}\big(\|\sqrt{\mathcal{D}}\nabla_{\mathcal{P}} r_h^+\|_{0,\Omega} + \|\sqrt{\mathcal{D}}\nabla_{\mathcal{P}} r_h^-\|_{0,\Omega}\big)\|\sqrt{\sigma_{\mathcal{P}}}[\![z_H]\!]\|_{0,\Gamma_{\mathcal{P}}}.\end{aligned}$$

For (II), we employ a trace-inverse estimate on every $T \in \mathcal{T}_K^{\partial}$, $K \in \mathcal{P}$, giving

$$\begin{aligned}|(II)| &\leq \sum_{K\in\mathcal{P}}\sum_{\substack{f\subset\partial T\cap\partial K\\ T\in\mathcal{T}_K^{\partial}}} \|\sigma_{\mathcal{P}}{}^{-1/2}\mathcal{D}\nabla r_h^-|_T\|_{0,f}\|\sqrt{\sigma_{\mathcal{P}}}[\![z_H]\!]\|_{0,f}\\ &\leq \Bigg(\sum_{K\in\mathcal{P}}\sum_{T\in\mathcal{T}_K^{\partial}} \frac{\delta_K|f|}{8C_{\text{star}}|T|}\min_{*\in\{+,-\}} H_{K_*}\delta_{K_*}^{-1}\|\sqrt{\mathcal{D}}\nabla r_h^-\|_{0,T}^2\Bigg)^{\frac{1}{2}}\|\sqrt{\sigma_{\mathcal{P}}}[\![z_H]\!]\|_{0,\Gamma_{\mathcal{P}}}.\end{aligned}$$

Since $\min_{*\in\{+,-\}} H_{K_*}\delta_{K_*}^{-1} \leq H_K\delta_K^{-1}$ (one of the K_* is K itself), and $\alpha|_K = \gamma\max\Big\{1, \frac{H_K}{8C_{\text{star}}}\max_{T\in\mathcal{T}_K^{\partial}}\frac{|f|}{|T|}\Big\}$, we

deduce

$$|(II)| \leq \Big(\sum_{K\in\mathcal{P}} \sum_{T\in\mathcal{T}_K^\partial} \|\sqrt{D}\nabla r_h^-\|_{0,T}^2 \Big)^{\frac{1}{2}} \|\sqrt{\sigma_{\mathcal{P}}}[\![z_H]\!]\|_{0,\Gamma_{\mathcal{P}}} \leq \gamma^{-1/2} \|\sqrt{\alpha D}\nabla_{\mathcal{P}} r_h^-\|_{0,\Omega} \|\sqrt{\sigma_{\mathcal{P}}}[\![z_H]\!]\|_{0,\Gamma_{\mathcal{P}}}. \quad (4.26)$$

Returning to (4.25), the above estimates and Lemma 4.3.3, upon observing that $r_h^- \in W_{\mathcal{T}}$ and $\sqrt{\alpha} r_h^- \in W_{\mathcal{T}}$, imply

$$\begin{aligned}
&\Big| \int_{\Gamma_{\mathcal{P}}} \{\!\{ D\nabla r_h^+ \}\!\} \cdot [\![z_H]\!] \,\mathrm{d}s \Big| \\
&\leq 4^{-1} \Big(\|\sqrt{D}\nabla_{\mathcal{P}} r_h^+\|_{0,\Omega} + \|\sqrt{D}\nabla_{\mathcal{P}} r_h^-\|_{0,\Omega} + 4\gamma^{-1/2} \|\sqrt{\alpha D}\nabla_{\mathcal{P}} r_h^-\|_{0,\Omega} \Big) \|\sqrt{\sigma_{\mathcal{P}}}[\![z_H]\!]\|_{0,\Gamma_{\mathcal{P}}} \\
&\leq \frac{\sqrt{2}}{4} \Big(\|\sqrt{D}\nabla_{\mathcal{P}} r_h^+\|_{0,\Omega}^2 + 25 C_{\mathrm{equiv}} \gamma^{-1} \|r_h^-\|_s^2 \Big)^{\frac{1}{2}} \|\sqrt{\sigma_{\mathcal{P}}}[\![z_H]\!]\|_{0,\Gamma_{\mathcal{P}}},
\end{aligned}$$

respectively, since $\alpha^{-1/2} \leq \gamma^{-1/2}$ by construction. Since $\gamma \geq 25 C_{\mathrm{equiv}}$, we arrive at the estimate:

$$\Big| \int_{\Gamma_{\mathcal{P}}} \{\!\{ D\nabla r_h^+ \}\!\} \cdot [\![z_H]\!] \,\mathrm{d}s \Big| \leq \frac{\sqrt{2}}{4} \big(\|\sqrt{D}\nabla_{\mathcal{P}} r_h^+\|_{0,\Omega}^2 + \|r_h^-\|_s^2 \big)^{\frac{1}{2}} \|\sqrt{\sigma_{\mathcal{P}}}[\![z_H]\!]\|_{0,\Gamma_{\mathcal{P}}}. \quad (4.27)$$

Applying (4.24) and (4.27) into (4.23), along with the trivial bound $\sqrt{2} \leq 2$ yields (4.22). □

Lemma 4.3.6. *[3, Lemma 3.6] With the assumptions and notation of Lemma 4.3.5, we also have*

$$a_{DG}(r_h^+, v_H - w_H) + s(r_h^-, v_H - w_H) \geq \frac{1}{2} \big(|||v_H - w_H|||_{\mathcal{P}}^2 + \|r_h^-\|_s^2 \big). \quad (4.28)$$

Proof. Since $v_H - w_H = r_h^+ + r_h^-$, we obtain

$$a_{DG}(r_h^+, v_H - w_H) + s(r_h^-, v_H - w_H) \geq a_{DG}(v_H - w_H, v_H - w_H) - a_{DG}(r_h^-, v_H - w_H) + s(r_h^-, r_h^-), \quad (4.29)$$

using $s(r_h^-, r_h^+) \geq 0$ from (1.69). Since $v_H - w_H \in V_{\mathcal{P}}$, Lemma 4.2.5 yields

$$a_{DG}(v_H - w_H, v_H - w_H) \geq \frac{3}{4} |||v_H - w_H|||_{\mathcal{P}}^2. \quad (4.30)$$

Next, arguing exactly as in the proofs of (4.24) and (4.26), we have

$$\Big| \int_{\Gamma_{\mathcal{P}}} \{\!\{ D\nabla(v_H - w_H) \}\!\} \cdot [\![r_h^-]\!] \mathrm{d}s \Big| \leq 4^{-1} \|D^{1/2}\nabla_{\mathcal{P}}(v_H - w_H)\|_{0,\Omega} \|\sqrt{\sigma_{\mathcal{P}}}[\![r_h^-]\!]\|_{0,\Gamma_{\mathcal{P}}},$$

and

$$\Big| \int_{\Gamma_{\mathcal{P}}} \{\!\{ D\nabla r_h^- \}\!\} \cdot [\![v_H - w_H]\!] \mathrm{d}s \Big| \leq \gamma^{-1/2} \| \sqrt{\alpha} D^{1/2} \nabla_{\mathcal{P}} r_h^- \|_{0,\Omega} \| \sqrt{\sigma_{\mathcal{P}}} [\![v_H - w_H]\!] \|_{0,\Gamma_{\mathcal{P}}}.$$

Applying Lemma 4.3.3 to $\sqrt{\alpha} r_h^- \in W_{\mathcal{T}}$ gives

$$\gamma^{-1/2} \| \sqrt{\alpha} D^{1/2} \nabla_{\mathcal{P}} r_h^- \|_{0,\Omega} \leq 4^{-1} \| r_h^- \|_s,$$

for the stated choices of α and γ. Using these three estimates together with the trivial bounds $\| D^{1/2} \nabla_{\mathcal{P}} v \|_{0,\Omega} \leq |||v|||_{\mathcal{P}}$ and $\| \sqrt{\sigma_{\mathcal{P}}} [\![v]\!] \|_{0,\Gamma_{\mathcal{P}}} \leq |||v|||_{\mathcal{P}}$ for any $v \in W_{\mathcal{T}}$, we obtain

$$|a_{DG}(r_h^-, v_H - w_H)| \leq \frac{5}{4} |||r_h^-|||_{\mathcal{P}} |||v_H - w_H|||_{\mathcal{P}} + \frac{1}{4} \| r_h^- \|_s |||v_H - w_H|||_{\mathcal{P}}. \tag{4.31}$$

Moreover, Lemma 4.3.3 applied to $r_h^- \in W_{\mathcal{T}}$ yields

$$16 |||r_h^-|||_{\mathcal{P}}^2 \leq 16 C_{\mathrm{equiv}} \| \alpha^{-1/2} r_h^- \|_s^2 \leq \| r_h^- \|_s^2, \tag{4.32}$$

since $\alpha^{-1} \leq \gamma^{-1} \leq (16 C_{\mathrm{equiv}})^{-1}$ by construction. Hence, from (4.31) we conclude that

$$|a_{DG}(r_h^-, v_H - w_H)| \leq \frac{21}{64} \| r_h^- \|_s |||v_H - w_H|||_{\mathcal{P}} < \frac{1}{4} \| r_h^- \|_s^2 + \frac{1}{4} |||v_H - w_H|||_{\mathcal{P}}^2.$$

Substituting this into (4.29), together with (4.30), establishes (4.28). □

The preceding two lemmas are sufficient to guarantee the applicability of monotone operator theory, thereby establishing the existence and uniqueness of a solution to (4.15). Furthermore, they provide the foundation for an iterative scheme that converges to the solution. In particular, we obtain the following result.

Theorem 4.3.7 (Well-posedness)**.** *[3, Theorem 3.7] Let $T : V_{\mathcal{P}} \to [V_{\mathcal{P}}]'$ be defined by*

$$[Tu_H, v_H] := a_h(u_H; v_H) = a_{DG}(\mathcal{E}^+(u_H), v_H) + s(\mathcal{E}^-(u_H), v_H), \qquad v_H \in V_{\mathcal{P}}.$$

Then, under the hypotheses of Lemma 4.3.5, T is continuous and strongly monotone. Consequently, the problem (4.15) *has a unique solution $u_H \in V_{\mathcal{P}}$.*

Proof. To show that T is continuous, we recall (4.32) and use (4.22), which gives

$$[Tv_H - Tw_H, v_H - w_H] \leq \frac{7}{4}\Big(|||\mathcal{E}^+(v_H) - \mathcal{E}^+(w_H)|||_{\mathcal{P}}^2 + \frac{26}{25}\|\mathcal{E}^-(v_H) - \mathcal{E}^-(w_H)\|_s^2 \Big)^{1/2} \Big(|||z_H|||_{\mathcal{P}}^2 + \|z_H\|_s^2 \Big)^{1/2}.$$

Thus, T is continuous on $V_{\mathcal{P}}$.

Moreover, (4.28) implies

$$[Tv_H - Tw_H, v_H - w_H] \geq \frac{1}{2}\big(|||v_H - w_H|||_{\mathcal{P}}^2 + \|\mathcal{E}^-(v_H) - \mathcal{E}^-(w_H)\|_s^2 \big),$$

showing that T is strongly monotone on $V_{\mathcal{P}}$. Existence and uniqueness therefore follow from the Browder-Minty Theorem (see, e.g., [110, Theorem 10.41]) together with strong monotonicity. □

4.4 Bound-preserving best approximation

For the error analysis, we require the construction of suitable nodally bound-preserving piece-wise polynomial best approximations to the exact weak solution u of (4.1). The starting point of the construction is to define the Lagrange interpolant on a simplex $K^\sharp$ containing a polytopic element $K \in \mathcal{P}$. To eventually guarantee bound-preservation in the construction below, we make the following technical assumption on the admissible families of d-simplices $K^\sharp \supset K$, $K \in \mathcal{P}$.

Assumption 4.4.1. *For each member $\mathcal{P}_i$ of a family of polytopic meshes $(\mathcal{P}_i)_{i\in I}$, we define the covering $\mathcal{P}_i^\sharp = \{K^\sharp\}$, related to the polytopic mesh $\mathcal{P}$ (see Figure 4.1), as the set of open, shape-regular d-simplices $K^\sharp$, of smallest possible in diameter $H_{K^\sharp}$, such that:*

(a) for each $K \in \mathcal{P}_i$, there exists a unique $K^\sharp \in \mathcal{P}_i^\sharp$ such that $K \subset K^\sharp$;

(b) there exists a (global) constant $C_{\mathrm{sh}}^\sharp \geq 1$, such that $H_{K^\sharp} \leq C_{\mathrm{sh}}^\sharp H_K$, for all $K \in \mathcal{P}_i$, $i \in I$;

(c) each $\mathcal{P}_i^\sharp$ covers exactly *the domain Ω, that is, we have $\Omega = \cup_{K^\sharp \in \mathcal{P}_i^\sharp} K^\sharp$.*

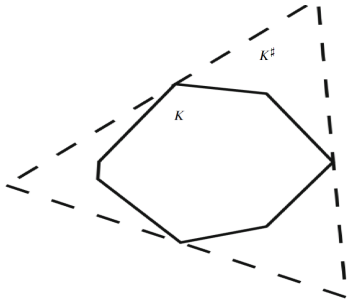


Figure 4.1: Polygonal element K and its covering $K^\sharp$.

Of course, the simplices $K^\sharp$ are allowed to overlap. Note that from Definition 4.4.1(b) and Assumption 4.2.3(a), we have

$$|K^\sharp| \leq H_{K^\sharp}^d \leq \left(C_{\text{sh}}^\sharp H_K\right)^d \leq \left(C_{\text{sh}}^\sharp C_{\text{star}}\right)^d \rho_K^d \leq \pi^{-1}\left(C_{\text{sh}}^\sharp C_{\text{star}}\right)^d |K|,$$

respectively, since ρ_K is the radius of a ball in K. Setting $C_{\text{cov}} := \pi^{-1}\left(C_{\text{sh}}^\sharp C_{\text{star}}\right)^d$, we deduce

$$\sum_{K^\sharp \in \mathcal{P}^\sharp} |K^\sharp| \leq C_{\text{cov}} \sum_{K \in \mathcal{P}} |K| = C_{\text{cov}} |\Omega|, \tag{4.33}$$

since there is correspondence between K and $K^\sharp$. The last bound shows that the assumptions made on the mesh $\mathcal{P}$ and on the simplicial covering $\mathcal{P}^\sharp$ are sufficient to control the extend of overlap in the computational domain Ω caused by the use of $\mathcal{P}^\sharp$.

We now proceed to establish a best approximation result for a nodally bound-preserving approximant in $V_\mathcal{P}$. The proof is inspired by a similar construction introduced in [85]. The result in [85] concerns the best approximation by bounds-constrained polynomials in the continuous setting. They proved that if $K \subset \mathbb{R}^n$ be a compact domain and $f : K \to \mathbb{R}$ be a continuous function with range $[m, M]$, and V_h be a vector space of continuous functions mapping K into $\mathbb{R}$ such that constant-valued functions on K are included. For any $g \in V_h$, there exists $q \in V_h$ such that the range of q is contained in $[m, M]$ and

$$\|f - q\|_\infty \leq 2\|f - g\|_\infty.$$

In the present work, we extend this idea to the finite element space $V_\mathcal{P}$, where the approximation not only preserves the nodal bounds but also satisfies optimal-order approximation properties. The following theorem provides this discrete extension of the continuous result presented in [85].

Theorem 4.4.2. *[3, Theorem 4.2] Let $K \in \mathcal{P}$, $k \in \mathbb{N}$ the polynomial degree of $V_{\mathcal{P}}$, and $v \in H^{k+1}(K^\sharp)$ a function with range in $[0,\kappa]$. Then, there exists $\pi_H v \in V_{\mathcal{P}}$ whose range is contained in $[0,\kappa]$, such that*

$$|v - \pi_H v|_{m,K} \leq C_m H_K^{k+1-m} |v|_{k+1,K^\sharp}, \tag{4.34}$$

$m = 0,1,2$, for $K^\sharp \in \mathcal{P}^\sharp$ as in Definition 4.4.1, with $C_m > 0$ depending on the shape-regularity of $K^\sharp$, on r and on k; moreover for $m = 1,2$, $C_m > 0$, depends also on C_{star} and on the upper range κ, as well as on $|v|_{1,d,\Omega}$ for $m = 1$, or on $|v|_{2,\Omega}$ for $m = 2$.

Proof. Let $k \geq 2$ and $i_H : H^2(K^\sharp) \to \mathbb{P}_k(K^\sharp)$ be the Lagrange interpolation operator, producing the interpolating polynomial of total degree k on the Lagrange nodes of the simplex $K^\sharp$. Then, as $v \in H^{k+1}(K^\sharp)$, we have the best approximation estimates

$$\|v - i_H v\|_{0,K^\sharp} + H_{K^\sharp}|v - i_H v|_{1,K^\sharp} + H_{K^\sharp}^2 |v - i_H v|_{2,K^\sharp} + |K^\sharp|^{1/2}\|v - i_H v\|_{0,\infty,K^\sharp} \leq \tilde{C}_{\mathrm{app}} H_{K^\sharp}^{k+1} |v|_{k+1,K^\sharp}, \tag{4.35}$$

for each $K \in \mathcal{P}$, with $\tilde{C}_{\mathrm{app}} > 0$ independent of $H_{K^\sharp}$ and of v; [48, Theorem 3.1.4]. From the properties stated in Definition 4.4.1, we deduce

$$\|v - i_H v\|_{0,K} + H_K |v - i_H v|_{1,K} + H_K^2 |v - i_H v|_{2,K} + |K|^{1/2}\|v - i_H v\|_{0,\infty,K} \leq C_{\mathrm{app}} H_K^{k+1} |v|_{k+1,K^\sharp}, \tag{4.36}$$

with $C_{\mathrm{app}} := \tilde{C}_{\mathrm{app}} \big(C_{\mathrm{sh}}^\sharp\big)^{k+1}$.

By construction, the nodal values in $K^\sharp$ of the interpolant $i_H v$ are within the range $[0,\kappa]$, since the range of v itself is contained in $[0,\kappa]$. When restricted to the polytopic element $K \subset K^\sharp$, however, the range of $i_H v|_K$ is *not* necessarily contained in $[0,\kappa]$, since Lagrange basis functions on a simplex do *not* all reside within the range $[0,1]$ for $k \geq 2$.

To construct an element-wise polynomial approximant of v that is also bound-preserving, we work as follows. On each $K \in \mathcal{P}$, we define the function

$$(\pi_H v)|_K := \frac{\kappa}{2} + \beta_\kappa(v)\left(i_H v - \frac{\kappa}{2}\right), \quad \text{with} \quad \beta_\kappa(v) := \kappa\big(\kappa + 2\|v - i_H v\|_{0,\infty,K}\big)^{-1};$$

cf., also, [85, Theorem 2]. Since $\pi_H v$ is constructed by vector space operations and $V_{\mathcal{P}}$ contains the constants, we have $\pi_H v \in V_{\mathcal{P}}$. To show that the range of $\pi_H v$ is contained in $[0,\kappa]$, we first observe that any function

w whose range is contained in $[0, \kappa]$ admits the bound

$$\left\| w - \frac{\kappa}{2} \right\|_{0,\infty,K} \leq \frac{\kappa}{2}. \tag{4.37}$$

Thus, for $\pi_H v$, we have, respectively,

$$\left\| \pi_H v - \frac{\kappa}{2} \right\|_{0,\infty,K} \leq \beta_\kappa(v) \left\| i_H v - \frac{\kappa}{2} \right\|_{0,\infty,K} \leq \frac{\kappa}{2},$$

since

$$\left\| i_H v - \frac{\kappa}{2} \right\|_{0,\infty,K} \leq \left\| i_H v - v \right\|_{0,\infty,K} + \left\| v - \frac{\kappa}{2} \right\|_{0,\infty,K} \leq \left\| i_H v - v \right\|_{0,\infty,K} + \frac{\kappa}{2}.$$

Having shown that $0 \leq \pi_H v \leq \kappa$, we now estimate its approximation capabilities. To that end, we have

$$v - \pi_H v = v - \frac{\kappa}{2} - \beta_\kappa(v) \left(i_H v - \frac{\kappa}{2} \right) = \left(v - \frac{\kappa}{2} \right)(1 - \beta_\kappa(v)) + \beta_\kappa(v)(v - i_H v), \tag{4.38}$$

and so, triangle inequality, (4.37), and (4.36), respectively, give

$$\begin{aligned} \left\| v - \pi_H v \right\|_{0,K} &\leq \frac{2\|v - i_H v\|_{0,\infty,K}}{\kappa + 2\|v - i_H v\|_{0,\infty,K}} \left\| v - \frac{\kappa}{2} \right\|_{0,K} + \beta_\kappa(v) \left\| v - i_H v \right\|_{0,K} \\ &\leq \sqrt{|K|} \left\| v - i_H v \right\|_{0,\infty,K} + \left\| v - i_H v \right\|_{0,K} \\ &\leq 2 C_{\mathrm{app}} H_K^{k+1} |v|_{k+1,K^\sharp}. \end{aligned} \tag{4.39}$$

For the H^1-seminorm best approximation estimate, differentiating (4.38), taking norm and employing the triangle inequality, gives

$$\left\| \nabla(v - \pi_H v) \right\|_{0,K} \leq (1 - \beta_\kappa(v)) \|\nabla v\|_{0,K} + \beta_\kappa(v) \|\nabla(v - i_H v)\|_{0,K}. \tag{4.40}$$

Noting that Hölder's inequality implies $\|\nabla v\|_{0,K} \leq |K|^{(d-2)/(2d)} \|\nabla v\|_{0,d,K}$, for $d = 2, 3$, (this is allowed as $H^{k+1}(K^\sharp)$, $k \geq 2$, embeds continuously into $W^{1,d}(K^\sharp)$), and employing (4.36), the last bound can be further estimated as follows:

$$\begin{aligned} \left\| \nabla(v - \pi_H v) \right\|_{0,K} &\leq 2\kappa^{-1} |K|^{(d-2)/(2d)} \left\| \nabla v \right\|_{0,d,K} \left\| v - i_H v \right\|_{0,\infty,K} + \left\| \nabla(v - i_H v) \right\|_{0,K} \\ &\leq C_{\mathrm{app}} \left(2\kappa^{-1} |K|^{-1/d} H_K \left\| \nabla v \right\|_{0,d,K} + 1 \right) H_K^k |v|_{k+1,K^\sharp}. \end{aligned}$$

From Assumption 4.2.3(a), we deduce $|K|^{-1/d} H_K \leq \rho_K^{-1} H_K \leq C_{\mathrm{star}}^{-1}$, and, thus, the result follows with

$C_1 := C_{\mathrm{app}}(2C_{\mathrm{star}}^{-1}\kappa^{-1}\|\nabla v\|_{0,d,K} + 1)$.

For the H^2-seminorm best approximation estimate, working as before gives

$$|v - \pi_H v|_{2,K} \leq (1 - \beta_\kappa(v))|v|_{2,K} + \beta_\kappa(v)|v - i_H v|_{2,K}. \tag{4.41}$$

Employing (4.36), the last bound can be further estimated as follows:

$$\begin{aligned} |v - \pi_H v|_{2,K} &\leq 2\kappa^{-1}|v|_{2,K}\|v - i_H v\|_{0,\infty,K} + |v - i_H v|_{2,K} \\ &\leq C_{\mathrm{app}}\big(2\kappa^{-1}H_K^2|v|_{2,K} + 1\big)H_K^{k-1}|v|_{k+1,K^\sharp}. \end{aligned}$$

The result follows upon setting $C_2 := C_{\mathrm{app}}(2\kappa^{-1}|v|_{2,K}H_K^2 + 1)$. The above conclude the proofs for $k \geq 2$.

For $k = 1$, we can set $\pi_H v = i_H v$ as the range of linear Lagrange basis functions over the simplex they are defined on is contained in $[0, 1]$, and, so, standard interpolation estimates are valid. Note that we can take $C_2 = 1$, for $k = 1$.

□

4.5 *A priori* error analysis

We are now in position to prove *a priori* error bounds between sufficiently smooth exactly solutions u to (4.1) and their nodally bound-preserving approximations $\mathcal{E}^+(u_H)$ provided by (4.14).

Theorem 4.5.1. *[3, Theorem 5.1] Let $u \in H^{k+1}(\Omega) \cap H_0^1(\Omega)$ be the solution to* (4.1) *and $\mathcal{E}^+(u_H) \in W_{\mathcal{T}}$ the nodally bound-preserving approximation produced by* (4.14)*, with $\sigma_{\mathcal{P}}$ as in* (4.6) *and α as in* (4.21)*, computed on families of polytopic meshes and respective simplicial subdivisions satisfying Assumptions 4.2.3 and 4.3.1, respectively. Then, we have the error bound:*

$$|||u - \mathcal{E}^+(u_H)|||_{\mathcal{P}} \leq C_{\mathrm{apr}}\Bigg(\sum_{K\in\mathcal{P}}\big(\delta_{\omega_K} + H_K^2\mu_K\big)H_K^{2k}|u|_{k+1,K^\sharp}^2\Bigg)^{1/2}, \tag{4.42}$$

where $\mu_K := \|\mu\|_{0,\infty,K}$, with $\{K^\sharp\}$ denoting the simplicial covering from Definition 4.4.1, and $C_{\mathrm{apr}} > 0$ dependent only on the polynomial degree $k \in \mathbb{N}$ and on the constants appearing in Assumptions 4.2.3 and 4.3.1 and in Definition 4.4.1.

Proof. We begin by decomposing the error $e := u - \mathcal{E}^+(u_H)$ as

$$e = \big(u - \pi_H u\big) + \big(\pi_H u - \mathcal{E}^+(u_H)\big) =: \eta + \xi,$$

noting that $\xi \in W_{\mathcal{T}}$. The triangle inequality implies $|||e|||_{\mathcal{P}} \leq |||\eta|||_{\mathcal{P}} + |||\xi|||_{\mathcal{P}}$.

We shall now estimate $\xi \in W_{\mathcal{T}}$. We observe that $\xi = \mathcal{E}^+(\pi_H u) - \mathcal{E}^+(u_H)$ since $0 \leq \pi_H u \leq \kappa$. Applying Lemma 4.3.6, we get

$$a_{DG}(\xi, \pi_H u - u_H) - s(\mathcal{E}^-(u_H), \pi_H u - u_H) \geq \frac{1}{2}\big(|||\pi_H u - u_H|||_{\mathcal{P}}^2 + \|\mathcal{E}^-(u_H)\|_s^2\big). \tag{4.43}$$

Assuming that $u \in H^{3/2+\epsilon}(\Omega)$, $\epsilon > 0$, with $0 \leq u \leq \kappa$ in Ω, (4.4) is consistent in the sense that $\alpha_{DG}(u, v_H) = \ell(v_H)$, for all $v_H \in V_{\mathcal{P}}$. Combining the last identity with (4.14), gives the orthogonality relation

$$\alpha_{DG}(u, v_H) = \ell(v_H) = a_{DG}(\mathcal{E}^+(u_H), v_H) + s(\mathcal{E}^-(u_H), v_H),$$

for all $v_H \in V_{\mathcal{P}}$, which, in turn implies

$$\alpha_{DG}(\xi, v_H) - s(\mathcal{E}^-(u_H), v_H) = -\alpha_{DG}(\eta, v_H). \tag{4.44}$$

Using (4.44) on (4.43), results into

$$\frac{1}{2}\big(|||\pi_H u - u_H|||_{\mathcal{P}}^2 + \|\mathcal{E}^-(u_H)\|_s^2\big) \leq -a_{DG}(\eta, \pi_H u - u_H). \tag{4.45}$$

Setting $\xi_H := \pi_H u - u_H$ for brevity, we embark on estimating the right-hand side of (4.45). Straightforward estimation gives

$$|a_{DG}(\eta, \xi_H)| \leq |||\eta|||_{\mathcal{P}} |||\xi_H|||_{\mathcal{P}} + \int_{\Gamma_{\mathcal{P}}} |\{\{\mathcal{D}\nabla\eta\}\}|\,|[\![\xi_H]\!]|\,\mathrm{d}s + \int_{\Gamma_{\mathcal{P}}} |\{\{\mathcal{D}\nabla\xi_H\}\}|\,|[\![\eta]\!]|\,\mathrm{d}s. \tag{4.46}$$

We proceed for each of the last two terms of (4.46) separately. For the last term we work as in (4.24) to obtain the bound

$$\int_{\Gamma_{\mathcal{P}}} |\{\{\mathcal{D}\nabla\xi_H\}\}|\,|[\![\eta]\!]|\,\mathrm{d}s \leq 4^{-1}\|\sqrt{\mathcal{D}}\nabla_{\mathcal{P}}\xi_H\|_{0,\Omega}\|\sqrt{\sigma_{\mathcal{P}}}[\![\eta]\!]\|_{0,\Gamma_{\mathcal{P}}}.$$

To further estimate the term involving η, for each simplicial face F shared by two elements $K_+, K_- \in \mathcal{P}$, we recall the definition of the simplex $K_*^F \subset K_*$, $*\in \{+,-\}$, having as a face F a simplicial face of K_*, (possibly after further subdivision of a polytopic face,) and opposite vertex the centre of the ball with respect to which K_* is star-shaped; cf., also the proof of Lemma 4.2.5 above. We apply the Trace Theorem on K_*^F,

getting

$$\begin{aligned}\|v\|_{0,F}^2 \le C_{\mathrm{tr}}\|v\|_{0,K_*^F}\|v\|_{1,K_*^F} &\le C_{\mathrm{tr}}(1+H_{K_*}^{-1})\|v\|_{0,K_*^F}^2 + C_{\mathrm{tr}}H_{K_*}|v|_{1,K_*^F}^2 \\ &\le C_{\mathrm{tr}}(1+H_{K_*})\big(H_K^{-1}\|v\|_{0,K_*^F}^2 + H_{K_*}|v|_{1,K_*^F}^2\big),\end{aligned} \tag{4.47}$$

with C_{tr} depending only on the star-shapedness parameter C_{star} from Assumption 4.2.3(a); we refer, e.g., to [38, Lemma 4.7] for a proof of the Trace Theorem with explicit value of C_{tr}.

Hence, (4.47) implies

$$\|\sqrt{\sigma_{\mathcal{P}}}[\![\eta]\!]\|_{0,\Gamma_{\mathcal{P}}}^2 \le \tilde{c}_1 \sum_{F\subset\Gamma_{\mathcal{P}}}\sum_{*\in\{+,-\}} \delta_{\omega_{K_*}}\big(H_{K_*}^{-2}\|\eta\|_{0,K_*^F}^2 + |\eta|_{1,K_*^F}^2\big) \le \tilde{c}_1 \sum_{K\in\mathcal{P}} \delta_{\omega_K}\big(H_K^{-2}\|\eta\|_{0,K}^2 + |\eta|_{1,K}^2\big),$$

with $\tilde{c}_1 := 16(1+\max_{K\in\mathcal{P}} H_K)C_{\mathrm{tr}}C_{\mathrm{star}}k(k-1+d)d^{-1}$, since we have $\cup_{F\subset\partial K_*} K_*^F = K_*$ by construction. Theorem 4.4.2 then implies

$$\|\sqrt{\sigma_{\mathcal{P}}}[\![\eta]\!]\|_{0,\Gamma_{\mathcal{P}}}^2 \le c_1 \sum_{K\in\mathcal{P}} \delta_{\omega_K} H_K^{2k}|v|_{k+1,K^\sharp}^2, \tag{4.48}$$

with $c_1 = \tilde{c}_1 \max\{C_0^2, C_1^2\}$.

Now we turn to the penultimate term on the right-hand side of (4.46), for which we have

$$\int_{\Gamma_{\mathcal{P}}} |\{\!\{D\nabla\eta\}\!\}|\,|[\![\xi_H]\!]|\,\mathrm{d}s \le \|\sigma_{\mathcal{P}}^{-1/2}\{\!\{D\nabla\eta\}\!\}\|_{0,\Gamma_{\mathcal{P}}}\|\sqrt{\sigma_{\mathcal{P}}}[\![\xi_H]\!]\|_{0,\Gamma_{\mathcal{P}}}.$$

We further estimate the term involving η as follows:

$$\begin{aligned}\|\sigma_{\mathcal{P}}^{-1/2}\{\!\{D\nabla\eta\}\!\}\|_{0,\Gamma_{\mathcal{P}}}^2 &\le \tilde{c}_2 \sum_{F\subset\Gamma_{\mathcal{P}}}\sum_{*\in\{+,-\}} \delta_{K_*}^{-1}\|D\|_{0,\infty,K_*}^2\big(|\eta|_{1,K_*^F}^2 + H_{K_*}^2|\eta|_{2,K_*^F}^2\big) \\ &\le \tilde{c}_2 \sum_{K\in\mathcal{P}} \|D^{-1}\|_{0,\infty,K}^{-1}\big(|\eta|_{1,K}^2 + H_K^2|\eta|_{2,K}^2\big)\end{aligned}$$

for $\tilde{c}_2 := \big(8C_{\mathrm{star}}k(k-1+d)\big)^{-1}d^2C_{\mathrm{tr}}(1+\max_{K\in\mathcal{P}} H_K)$, using the Trace Theorem (4.47) for $v = |\nabla\eta|_{K_*}|$, the definition of δ_{K_*}, and working as before to collect the contributions to each K_*. Noting that $\|D^{-1}\|_{0,\infty,K}^{-1} \le \|D\|_{0,\infty,K}$, we apply Theorem 4.4.2 to arrive at

$$\|\sigma_{\mathcal{P}}^{-1/2}\{\!\{D\nabla\eta\}\!\}\|_{0,\Gamma_{\mathcal{P}}}^2 \le c_2 \sum_{K\in\mathcal{P}} \|D\|_{0,\infty,K} H_K^{2k}|u|_{k+1,K^\sharp}^2, \tag{4.49}$$

with $c_2 := \tilde{c}_2 \max\{C_1^2, C_2^2\}$.

Finally, using (4.48), along with straightforward estimation, also yields

$$|||\eta|||_{\mathcal{P}}^2 \leq \sum_{K\in\mathcal{P}} \left(C_1^2\|\mathcal{D}\|_{0,\infty,K} + C_0^2 H_K^2\|\mu\|_{0,\infty,K} + c_1\delta_{\omega_K}\right) H_K^{2k}|u|_{k+1,K^\sharp}^2. \tag{4.50}$$

Using (4.48), (4.49) and (4.50) to estimate further (4.46), along with the discrete version of Cauchy-Schwarz inequality gives the following bound:

$$\begin{aligned} |a_{DG}(\eta,\xi_H)| &\leq 2\big(|||\eta|||_{\mathcal{P}}^2 + \|\sigma_{\mathcal{P}}^{-1/2}\{\!\{\mathcal{D}\nabla\eta\}\!\}\|_{0,\Gamma_{\mathcal{P}}}^2 + \|\sqrt{\sigma_{\mathcal{P}}}[\![\eta]\!]\|_{0,\Gamma_{\mathcal{P}}}^2\big)^{1/2}|||\xi_H|||_{\mathcal{P}} \\ &\leq c_3\Bigg(\sum_{K\in\mathcal{P}}\left(\delta_{\omega_K} + H_K^2\mu_K\right)H_K^{2k}|u|_{k+1,K^\sharp}^2\Bigg)^{1/2}|||\xi_H|||_{\mathcal{P}}, \end{aligned} \tag{4.51}$$

with $c_3 := 2(\max\{2c_1 + c_2 + C_1^2, C_0^2\})^{1/2}$ and $\mu_K := \|\mu\|_{0,\infty,K}$.

Returning, now, to (4.45), we use (4.51), (along with the inequality $ab \leq a^2 + b^2/4$ on the right-hand side of (4.51)) to deduce

$$\frac{1}{4}|||\xi_H|||_{\mathcal{P}}^2 + \frac{1}{2}\|\mathcal{E}^-(u_H)\|_s^2 \leq c_3^2\sum_{K\in\mathcal{P}}\left(\delta_{\omega_K} + H_K^2\mu_K\right)H_K^{2k}|u|_{k+1,K^\sharp}^2. \tag{4.52}$$

From the above estimate , we can arrive at an upper bound for $|||\xi|||_{\mathcal{P}}$, as follows. Since $\xi_H = \xi - \mathcal{E}^-(u_H)$, triangle inequality and (4.32) give

$$\begin{aligned} |||\xi_H|||_{\mathcal{P}}^2 &\geq \left(|||\xi|||_{\mathcal{P}} - |||\mathcal{E}^-(u_H)|||_{\mathcal{P}}\right)^2 \geq \left(|||\xi|||_{\mathcal{P}} - 5^{-1}\|\mathcal{E}^-(u_H)\|_s\right)^2 \\ &\geq |||\xi|||_{\mathcal{P}}^2 + \frac{1}{25}\|\mathcal{E}^-(u_H)\|_s^2 - \frac{2}{5}|||\xi|||_{\mathcal{P}}\|\mathcal{E}^-(u_H)\|_s \geq \frac{4}{5}|||\xi|||_{\mathcal{P}}^2 - \frac{4}{25}\|\mathcal{E}^-(u_H)\|_s^2, \end{aligned} \tag{4.53}$$

respectively, using the inequality $2ab \leq a^2/5 + 5b^2$ in the last step. Combining (4.52) with (4.53) provides us with the bound

$$|||\xi|||_{\mathcal{P}}^2 \leq 5c_3^2\sum_{K\in\mathcal{P}}\left(\delta_{\omega_K} + H_K^2\mu_K\right)H_K^{2k}|u|_{k+1,K^\sharp}^2, \tag{4.54}$$

ignoring the non-negative term $\|\mathcal{E}^-(u_H)\|_s^2$. The triangle inequality completes the proof. □

Remark 4.5.2. *Some remarks are in order.*

1. *The presence of the covering simplices $K^\sharp$ on the right-hand side of* (4.42) *does not affect the inferred of the order of the method proved. Indeed, estimating further* (4.42) *from above and using* (4.33)*, results in the bound*

$$|||u - \mathcal{E}^+(u_H)|||_{\mathcal{P}} \leq C\max_{K\in\mathcal{P}} H_K^k|u|_{k+1,\Omega},$$

for $C > 0$ depending on C_{apr}, C_{cov}, $\mathcal{D}$, and on μ only.

2. *The proof of Theorem 4.5.1 offers also an* a priori *error bound for the 'non-compliant' approximate solution $u_H \in V_{\mathcal{P}}$, which may* not *be nodally bound-preserving. Indeed, from* (4.52)*, we obtain a bound on $|||\pi_H u - u_H|||_{\mathcal{P}}$, which, combined with* (4.50) *results to the error bound*

$$|||u - u_H|||_{\mathcal{P}} \leq \tilde{C}_{\text{apr}} \left(\sum_{K \in \mathcal{P}} \left(\delta_{\omega_K} + H_K^2 \mu_K \right) H_K^{2k} |v|^2_{k+1,K^\sharp} \right)^{1/2},$$

for a $\tilde{C}_{\text{apr}} > 0$ having the same dependence on the constants as $C_{\text{apr}} > 0$.

3. *The bound* (4.52) *also gives some insight on the rate of decay of $|\mathcal{E}^-(u_H)| \to 0$, i.e., of the, 'non-compliant' to the bounds on the range, part of the approximate solution, as* $\max_{K \in \mathcal{P}} H_K \to 0$ *and/or* $\max_{T \in \mathcal{T}} h_T \to 0$.

   Assume for simplicity that $\mathcal{T}$ is quasi-uniform so that h is the representative submesh simplex diameter for $\mathcal{T}$ and that $\mathcal{P}$ is also quasi-uniform, so that H is the representative polytopic element diameter. Then, (4.21) *implies that each elemental component is proportional to Hh^{-1}.*

   We first consider the scenario of H fixed and $h \to 0$. Then, the right-hand side of (4.52) *is bounded by $C_{\text{up}} H^k$, for a constant $C_{\text{up}} > 0$, depending on the exact solution u. Then, from the definition of the stabilisation norm and the fact that $\alpha \sim Hh^{-1}$, we get*

$$C_{\text{up}} H^{2k} \geq \|\mathcal{E}^-(u_H)\|_s^2 \geq C_{\text{down}} H h^{-3} \sum_{K \in \mathcal{P}} \sum_{T \in \mathcal{T}_K} \sum_{i=1}^{m_{k,d}} h^d \left(\mathcal{E}^-(u_H)(\boldsymbol{x}_i^T) \right)^2 \sim C_{\text{down}} H h^{-3} \|\mathcal{E}^-(u_H)\|_{0,\Omega}^2,$$

   for $C_{\text{down}} > 0$ depending on $\mathcal{D}_0$. The above imply that $\|\mathcal{E}^-(u_H)\|_{0,\Omega} \sim h^{3/2}$ for fixed H. Also, since $h^{-3} \geq H^{-3}$, we have $\|\mathcal{E}^-(u_H)\|_{0,\Omega} \sim H^{k+1}$, which is expected from (4.52).

4.6 Implementation and matrix structure

We can construct the R-FEM through a transformation starting from the discontinuous Galerkin method on the following space that is define on the submesh $\mathcal{T}$

$$V_{\mathcal{T}} := \{ v_h \in L^2(\Omega) : v_h|_h \in \mathbb{P}_k(T) \quad \forall T \in \mathcal{T} \}.$$

Then, we consider the following discontinuous Galerkin finite element method: Find $u_h \in V_{\mathcal{T}}$ such that

$$a_{\mathrm{DG}}^{\mathcal{T}}(u_h; v_h) = \ell(v_h) \qquad \text{for all} \quad v_h \in V_{\mathcal{T}}, \tag{4.55}$$

with $a_{DG}^{\mathcal{T}} : (H_0^1(\Omega) + V_{\mathcal{T}}) \times (H_0^1(\Omega) + V_{\mathcal{T}}) \to \mathbb{R}$, given by

$$\begin{aligned} a_{\mathrm{DG}}^{\mathcal{T}}(u_h; v_h) = &\int_\Omega \left(\mathcal{D} \nabla_{\mathcal{T}} u_h \nabla v_h + \mu u_h v_h \right) \mathrm{d}\boldsymbol{x} + \int_{\Gamma_{\mathcal{T}}} \sigma_{\mathcal{P}} [\![u_h]\!] \cdot [\![v_h]\!] \, \mathrm{d}s \\ &- \int_{\Gamma_{\mathcal{T}}} \left(\{\!\{ \mathcal{D} \nabla u_h \}\!\} \cdot [\![v_h]\!] + \theta \{\!\{ \mathcal{D} \nabla v_h \}\!\} \cdot [\![u_h]\!] \right) \mathrm{d}s, \end{aligned}$$

and $\ell(v_h) := \langle f, v_h \rangle_\Omega$. Here, $\theta \in [-1, 1]$ is the same parameter which has been defined in (4.5), the parameter $\sigma_{\mathcal{P}}$ is the penalty parameter (4.6) and $\Gamma_{\mathcal{T}} := \cup_{T \in \mathcal{T}} \partial T$ is the the skeleton of $\mathcal{T}$.

Furthermore, consider the following finite element method on the space $W_{\mathcal{T}}$ which has been defined in Section 4.2.1: Find $w_h \in W_{\mathcal{T}}$ such that

$$a_{\#}(w_h; v_h) = \ell(v_h) \qquad \text{for all} \quad v_h \in W_{\mathcal{T}}, \tag{4.56}$$

where

$$\begin{aligned} a_{\#}(w_h; v_h) := &\int_\Omega \left(\mathcal{D} \nabla_{\mathcal{P}} w_h \cdot \nabla_{\mathcal{P}} v_h + \mu w_h v_h \right) \mathrm{d}x + \int_{\Gamma_{\mathcal{P}}} \sigma_{\mathcal{P}} [\![w_h]\!] \cdot [\![v_h]\!] \, \mathrm{d}s \\ &- \int_{\Gamma_{\mathcal{P}}} \left(\{\!\{ \mathcal{D} \nabla w_h \}\!\} \cdot [\![v_h]\!] + \theta \{\!\{ \mathcal{D} \nabla v_h \}\!\} \cdot [\![w_h]\!] \right) \mathrm{d}s. \end{aligned}$$

It is easy to see that the following relation holds between the basis functions of the finite element space W_K, $K \in \mathcal{P}$ and the basis functions of the space $V_{\mathcal{T}}$

$$\phi_i^K = \sum_{l=1}^{N_K^i} \varphi_l^K \quad i = 1, \dots, N_K, \tag{4.57}$$

where N_K is the total number of degrees of freedom of the space W_K, ϕ_i^K is a basis of W_K associate to $\boldsymbol{x}_i^K$, N_K^i represents the number of degrees of freedom in $V_{\mathcal{T}}$ that coincide with the degree of freedom $\boldsymbol{x}_i^K$ in W_K, and φ_l^K is a basis function of the finite element space $V_{\mathcal{T}}$ whose degree of freedom is located at $\boldsymbol{x}_i^K$.

Thus, the finite element method (4.56) can be expressed as a transformation of the finite element method

(4.55). In other words

$$a_{\#}(\phi_i^{K_1};\phi_j^{K_2}) = \sum_{k=1}^{N_{K_2}^j}\sum_{l=1}^{N_{K_1}^i} a_{\mathrm{DG}}^{\mathcal{T}}(\varphi_l^{K_1},\varphi_k^{K_2}) \;\; K_1, K_2 \in \mathcal{P}. \tag{4.58}$$

Note that the summation on the right-hand side of (4.58) also pertains to the penalty term in the finite element method (4.56). because when $\Gamma_{\mathcal{P}} \cap \Gamma_{\mathcal{T}} = \phi$, we have $[\![\phi_i^{K_1}]\!] = 0$ and $[\![\phi_j^{K_2}]\!] = 0$, and therefore

$$\int_{\Gamma_{\mathcal{P}}} [\![\phi_i^{K_1}]\!] \cdot [\![\phi_j^{K_2}]\!] \mathrm{d}\boldsymbol{s} = \int_{\Gamma_{\mathcal{T}}} [\![\phi_i^{K_1}]\!] \cdot [\![\phi_j^{K_2}]\!] \mathrm{d}\boldsymbol{s},$$

this means that,

$$\int_{\Gamma_{\mathcal{P}}} \sigma_{\mathcal{P}} [\![\phi_i^{K_1}]\!] \cdot [\![\phi_j^{K_2}]\!] \mathrm{d}\boldsymbol{s} = \sum_{k=1}^{N_{K_2}^j}\sum_{l=1}^{N_{K_1}^i} \int_{\Gamma_{\mathcal{T}}} [\![\varphi_l^{K_1}]\!] \cdot [\![\varphi_k^{K_2}]\!] \mathrm{d}\boldsymbol{s}. \tag{4.59}$$

In fact, (4.59) indicates that when the values of the basis functions in the space $V_{\mathcal{T}}$, corresponding to the identical degrees of freedom in $W_{\mathcal{T}}$ are aligned, the internal jumps across the polygons vanish.

Assume that $N_{\mathcal{T}} := \dim(V_{\mathcal{T}})$ and $N_{\#} := \dim(W_{\mathcal{T}})$ represent the dimensions of the spaces $V_{\mathcal{T}}$ and $W_{\mathcal{T}}$, respectively. Clearly, (4.58) implies that the bilinear forms $a_{\mathrm{DG}}^{\mathcal{T}}(\cdot,\cdot)$ can be transformed into $a_{\#}(\cdot,\cdot)$ by modifying the basis functions. This transformation is represented by the matrix $\mathbf{O} \in \mathbb{R}^{N_{\#} \times N_{\mathcal{T}}}$ (see Figure 4.2a). Therefore, (4.58) can be expressed in algebraic form as

$$\mathbf{A}_{\#} = \mathbf{O}\mathbf{A}_{\mathrm{DG}}\mathbf{O}^T,$$

where $[\mathbf{A}_{\#}]_{ij} = a_{\#}(\phi_i, \phi_j)$ is the stiffness matrix related to the finite element method (4.56) and $[\mathbf{A}_{\mathrm{DG}}]_{ij} = a_{\mathrm{DG}}^{\mathcal{T}}(\varphi_i, \varphi_j)$ is the stiffness matrix of the finite element method (4.55).

By the above relations the algebraic form of the finite element method (4.56) can be written by the transformation matrix $\mathbf{O}$ and is given by seeking $\mathbf{W} \in \mathbb{R}^{N_{\#}}$ such that

$$\mathbf{A}_{\#}\mathbf{W} = \mathbf{O}\mathbf{A}_{\mathrm{DG}}\mathbf{O}^T\mathbf{W} = \mathbf{O}\mathbf{b}, \tag{4.60}$$

where $\mathbf{b} \in \mathbb{R}^{N_{\mathcal{T}}}$ given by $b_i = \langle f, \varphi_i \rangle_{\Omega}$, and $\varphi_i \in V_{\mathcal{T}}$ being a basis of $V_{\mathcal{T}}$.

Let ϑ_j be a basis function of the polygonal element K and define the finite element space

$$V_{\mathcal{P}}^K = \text{span}\{\vartheta_1, \vartheta_2, \cdots, \vartheta_{N_K}\},$$

where $V_{\mathcal{P}} = \oplus V_{\mathcal{P}}^K$. To compute the global stiffness matrix for the finite element method (4.14), assume that the basis functions ϑ_j^K of the finite element space $V_{\mathcal{P}}^K$, associated to the element K, can be expressed as a summation of the basis functions $\phi_1^K, \ldots, \phi_{N_K}^K$ corresponding to the degrees of freedom $\{\boldsymbol{x}_1^K, \boldsymbol{x}_2^K, \ldots, \boldsymbol{x}_{N_K}^K\}$ of the space W_K

$$\vartheta_j^K = \sum_{i=1}^{N_K} a_i^K \phi_i, \tag{4.61}$$

then, we have

$$a_{\text{DG}}(\vartheta_i^{K_1}, \vartheta_j^{K_2}) = \sum_{l=1}^{N_{K_2}} \sum_{k=1}^{N_{K_1}} a_l^{K_1} a_k^{K_2} a_{\#}(\phi_l^{K_1}, \phi_k^{K_2}) \quad K_1, K_2 \in \mathcal{P}. \tag{4.62}$$

Let $[A_{\text{DG}}]_{ij} = a_{\text{DG}}(\vartheta_i, \vartheta_j)$, $i, j = 1, \cdots, N_{\mathcal{P}}$ where $N_{\mathcal{P}} = \dim(V_{\mathcal{P}})$ and $[A_{\#}]_{ij} = a_{\#}(\phi_i, \phi_j)$, then, we can represent the connection between A_{DG} and $A_{\#}$ via a transformation matrix $\mathbf{Q} \in \mathbb{R}^{N_{\mathcal{P}} \times N_{\#}}$. Therefore, the global stiffness matrix $\mathbf{A}_{\mathcal{P}}$ can be written as

$$\mathbf{A}_{\mathcal{P}} = \mathbf{Q}\mathbf{A}_{\#}\mathbf{Q}^T.$$

The block diagonal matrix $\mathbf{Q}$ is depicted in Figure 4.2b.

Therefore, the algebraic form of the finite element method (4.14) is given by

$$\begin{aligned} \mathbf{A}_{\mathcal{P}}(\mathbf{Q}\mathbf{W}^+) + \mathbf{S}(\mathbf{Q}\mathbf{W}^-) &= \left(\mathbf{Q}\mathbf{A}_{\#}\mathbf{Q}^T\right)(\mathbf{Q}\mathbf{W}^+) + \mathbf{S}(\mathbf{Q}\mathbf{W}^-) \\ &= \left(\mathbf{Q}\mathbf{O}\mathbf{A}_{\text{DG}}\mathbf{O}^T\mathbf{Q}^T\right)(\mathbf{Q}\mathbf{W}^+) + \mathbf{S}(\mathbf{Q}\mathbf{W}^-) = \mathbf{F}_{\mathcal{P}}. \end{aligned} \tag{4.63}$$

where $\mathbf{W} \in \mathbb{R}^{N_{\#}}$, and $\mathbf{F}_{\mathcal{P}} = \mathbf{Q}\mathbf{O}\mathbf{b}$. The stabilisation matrix $\mathbf{S}$ in (4.63) similar to the matrix $\mathbf{A}_{\mathcal{P}}$ can be constructed through a transformation starting from the space $V_{\mathcal{T}}$, first we can write the following stabilisation

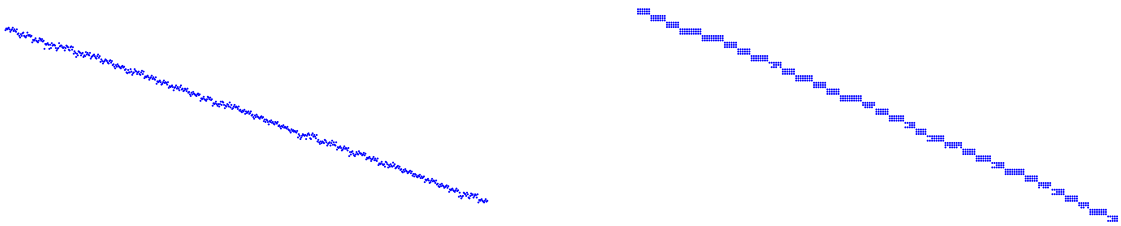
$$s_{\mathcal{T}}(v_h, u_h) = \sum_{K \in \mathcal{P}} \alpha \sum_{T \in \mathcal{T}_K} \sum_{i=1}^{m_{k,d}} \left(\mathcal{D}_{\omega_K} h_T^{d-2} + \mu_T h_T^d\right) v_h(\boldsymbol{x}_i) u_h(\boldsymbol{x}_i) \qquad u_h, v_h \in V_{\mathcal{T}},$$

where α is the same parameter which has been defined in (4.21). So, similar to $\mathbf{A}_{\mathcal{P}}$ we can write the algebraic

form of the stabilisation term $s(\cdot,\cdot)$ in finite element method (4.14) as

$$\mathbf{S}(\mathbf{Q}\mathbf{W}^{-}) = \left(\mathbf{O}\mathbf{Q}\mathbf{S}_{\mathcal{T}}\mathbf{Q}^{T}\mathbf{O}^{T}\right)(\mathbf{Q}\mathbf{W}^{-}).$$

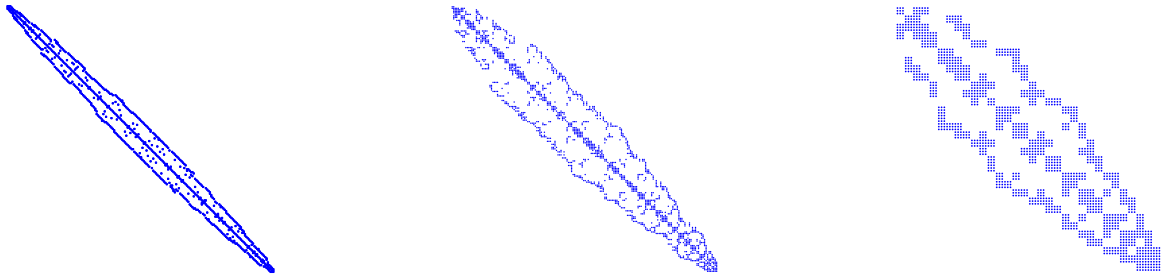
In Figure 4.2 the structure of the transportation matrices $\mathbf{O}$ and $\mathbf{Q}$ which are constructed on Mesh 4.4c have been illustrated.



(a) $\mathbf{O} \in \mathbb{R}^{216\times594}$ (b) $\mathbf{Q} \in \mathbb{R}^{96\times216}$

Figure 4.2: Illustration of the structure of the matrices $\mathbf{O}$ and $\mathbf{Q}$ constructed on Mesh 4.4c and using $\mathbb{P}_1$ elements,

Also, the structure of the matrices $\mathbf{A}_{\mathrm{DG}}$, $\mathbf{A}_{\#}$ and $\mathbf{A}_{\mathcal{P}}$ on Mesh 4.4c and using $\mathbb{P}_1$ elements have been shown in Figure 4.3.

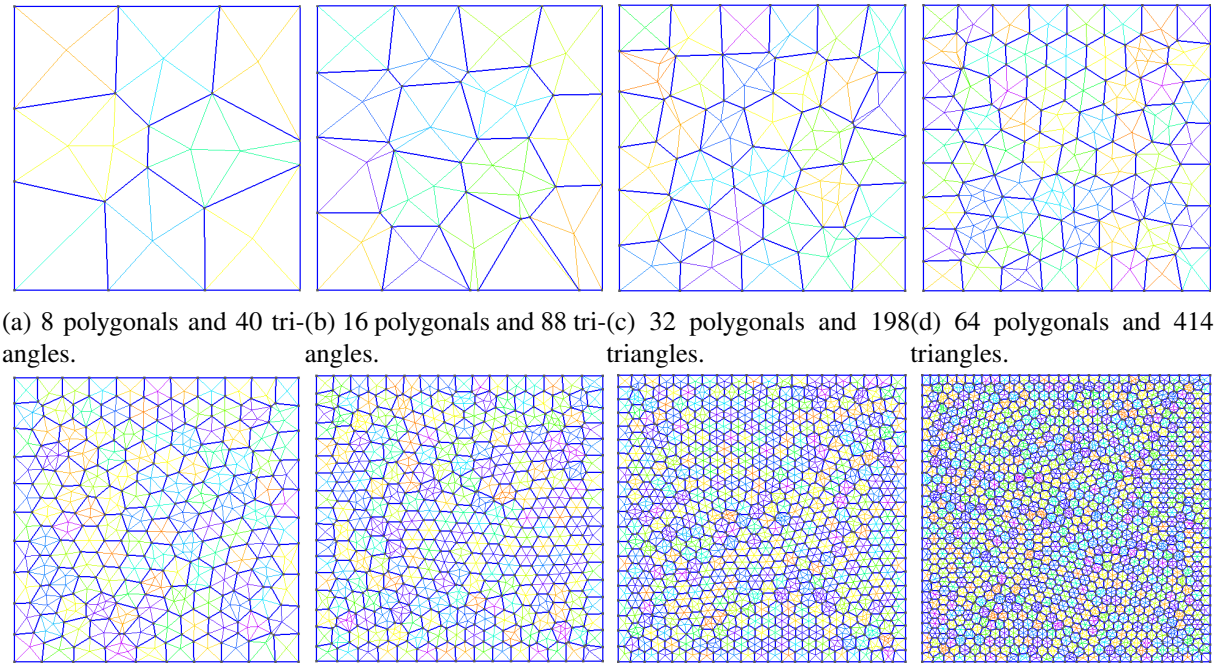


(a) $\mathbf{A}_{\mathrm{DG}} \in \mathbb{R}^{594\times594}$ (b) $\mathbf{A}_{\#} \in \mathbb{R}^{216\times216}$ (c) $\mathbf{A}_{\mathcal{P}} \in \mathbb{R}^{96\times96}$

Figure 4.3: Illustration of the structure of the matrices $\mathbf{A}_{\mathrm{DG}}$, $\mathbf{A}_{\#}$ and $\mathbf{A}_{\mathcal{P}}$ on Mesh 4.4c and using $\mathbb{P}_1$ elements.

4.7 Numerical experiments

In this section we present a set of numerical results testing the performance of the finite element method (4.14). In all numerical experiments in this section $\Omega = (0,1)^2$, and we have used the value $\gamma = 1$ in (4.21). We have selected two sets of meshes, different levels of them are depicted in Figures 4.4 and 4.5. The family illustrated in Figure 4.5 represents two levels of submesh refinement for mesh 4.4c. In this family mesh, each polygonal subdomain of mesh 4.4c has been uniformly refined, i.e., every triangular element in the subdomains is subdivided into four smaller elements by connecting the midpoints of the sides of the triangles.



(a) 8 polygonals and 40 triangles. (b) 16 polygonals and 88 triangles. (c) 32 polygonals and 198 triangles. (d) 64 polygonals and 414 triangles.

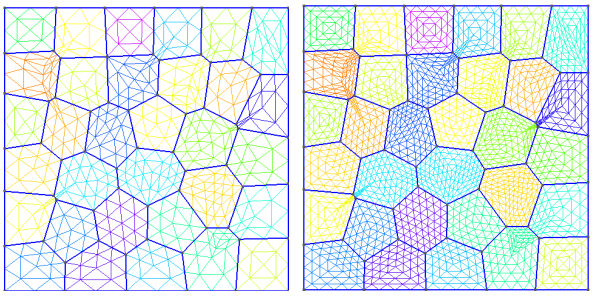
(e) 128 polygonals and 873 triangles. (f) 256 polygonals and 1745 triangles. (g) 512 polygonals and 3443 triangles. (h) 1024 polygonals and 7024 triangles.

Figure 4.4: Different levels of polygonal meshes with their corresponding triangular submeshes used in the numerical experiments.

To solve the nonlinear system associated to (4.14) as initial guess first, find $w_h^0 \in W_{\mathcal{T}}$ such that

$$a_{\#}(w_h^0, v_h) = \langle f, v_h \rangle_{\Omega}, \qquad \text{for all } v_h \in W_{\mathcal{T}}\,. \tag{4.64}$$

Now, let denote the value of w_h^0 at degree of freedom of the space $W_{\mathcal{T}}$ by $\mathbf{W}^0 = (w_1^0, w_2^0, \dots, w_{N_{\#}}^0)$, and consider the algebraic form of the finite element (4.14) which has been given by (4.63). Then, for $n = 1, 2, \dots$, find the solution $\mathbf{U}^{n+1}$ by the following semi-smooth Newton's which uses Clarke's generalized derivative



(a) 32 polygonals and 792 triangles.

(b) 32 polygonals and 3168 triangles.

Figure 4.5: Two levels of mesh refinement for mesh 4.4c.

method (for more details of Clarke's derivative see Remark 1.6.30 and see also [108, 114])

$$\mathbf{U}^{(n+1)} = \mathbf{Q}\mathbf{W}^{(n)} - J_{\mathbf{F}}(\mathbf{U}^{(n)})^{-1}\mathbf{F}(\mathbf{W}^{(n)}). \tag{4.65}$$

Here, $\mathbf{F}(\mathbf{W}) = \mathbf{F}_{\mathcal{P}} - \mathbf{A}_{\mathcal{P}}\mathbf{Q}(\mathbf{W})^{+} - \mathbf{S}\mathbf{Q}(\mathbf{W})^{-}$, and $J_{\mathbf{F}}(\mathbf{U})$ is computed as

$$J_F(\mathbf{U}) = J_F(\mathbf{W})\mathbf{Q}^T = -\big(\mathbf{A}_{\mathcal{P}}\mathbf{Q}\mathrm{diag}(I^1(w_1), \dots, I^1(w_{N_\#})) + \mathbf{S}\mathbf{Q}\mathrm{diag}(I^2(w_1), \dots, I^2(w_{N_\#}))\big)\mathbf{Q}^T, \tag{4.66}$$

where I^1 and I^2 are indicator functions and are defined as

$$I^1(w_i) = \begin{cases} 0, & \text{if} \quad w_i < 0, \\ 1, & \text{if} \quad w_i \geq 0, \end{cases} \qquad i = 1, 2, \dots, N_\#,$$

and

$$I^2(w_i) = \begin{cases} 0, & w_i \in [0, \kappa], \\ 1, & \text{otherwise}, \end{cases} \qquad i = 1, 2, \dots, N_\#.$$

Note that the relation (4.66) holds between $J_{\mathbf{F}}(\mathbf{W})$ and $J_{\mathbf{F}}(\mathbf{U})$ because

$$J_{\mathbf{F}}(\mathbf{W})\mathbf{W} = J_{\mathbf{F}}(\mathbf{W})\mathbf{Q}^T\mathbf{Q}\mathbf{W} = J_{\mathbf{F}}(\mathbf{U})\mathbf{U}.$$

The Newton's iterations (4.65) terminated once the following stopping criterion is satisfied

$$\|\mathcal{E}(u_H^{n+1}) - \mathcal{E}(u_H^n)\|_{0,\Omega} \leq 10^{-8}, \tag{4.67}$$

where $\mathcal{E}(u_H^n) = w_h^n = \sum_{i=1}^{N_\#} w_i^n \phi_i$.

We test the performance of the method asymptotically, where we use EOC as the estimated order of convergence, and we also examine the convergence of the iterative method. To compute the estimated order of convergence EOC of two different levels we use

$$\frac{\log(\frac{\|\cdot\|_{\text{level 1}}}{\|\cdot\|_{\text{level 2}}})}{\log(\frac{\sqrt{\text{Number of the polygons in level 1}}}{\sqrt{\text{Number of the polygons in level 2}}})}. \tag{4.68}$$

Note that in the numerical results, $\mathcal{E}^+(u_H)$ is always used for plotting the solutions and calculating the error.

Example 8 (Convergence of a problem with smooth solution)**.** *We consider the parameters* $\mu = 1$ *and* $\mathcal{D} = \epsilon \begin{bmatrix} 100 & \cos(x) \\ \cos(x) & 1 \end{bmatrix}$, *where* $\epsilon = 10^{-6}$. *The function* f *is chosen such that the analytical solution of* (2.1) *is given by* $u(x, y) = \sin(c\pi x)\sin(c\pi y)$. *Note that* $u(x, y) \in [0, 1]$, *and therefore we set* $\kappa = 1$. *In all the experiments, we use the penalty parameter* $\gamma = 1$ *in* (4.21).

To show the convergence results, $\mathbb{P}_1$, $\mathbb{P}_2$, and $\mathbb{P}_3$ elements on the meshes 4.4a-4.4h have been used. The convergence results are reported in Tables 4.1-4.3 for the norms $\|\cdot\|_{0,\Omega}$, $\|\cdot\|_{\text{DG}}$, and the seminorm $|\cdot|_{1,\Omega}$ for $u - \mathcal{E}^+(u_H)$. Additionally, the $\|\cdot\|_s$-norm for $\mathcal{E}^-(u_H)$ and the number of iterations required to achieve convergence for the nonlinear system are included. Note that each level of the meshes 4.4a-4.4h is not a refinement of the previous one. Therefore, the order of convergence was approximately computed using (4.68). The results in this example show optimal rates for all degrees when we do the refinement of the polygonal mesh. Additionally, refinement leads to a monotone decay of $\|\mathcal{E}^-(u_H)\|_s$.

Mesh	Itr.	$\|u - \mathcal{E}^+(u_H)\|_{0,\Omega}$	EOC	$\|u - \mathcal{E}^+(u_H)\|_{1,\Omega}$	EOC	$\|u - \mathcal{E}^+(u_H)\|_{\text{DG}}$	EOC	$\|\mathcal{E}^-(u_H^n)\|_s$	EOC
Mesh 4.4c	3	5.21e-1	–	1.80e-2	–	5.22e-1	–	0	–
Mesh 4.4d	3	3.98e-1	0.77	1.73e-2	0.11	4.09e-1	0.70	0	–
Mesh 4.4e	3	2.98e-1	0.83	1.48e-2	0.45	3.24e-1	0.67	0	–
Mesh 4.4f	10	1.73e-1	1.57	1.16e-2	0.70	2.15e-1	1.18	9.07e-3	–
Mesh 4.4g	7	8.76e-2	1.96	8.41e-3	0.93	1.27e-1	1.52	1.63e-2	–
Mesh 4.4h	7	4.33e-2	2.03	5.97e-3	0.99	7.71e-2	1.44	1.30e-3	7.29

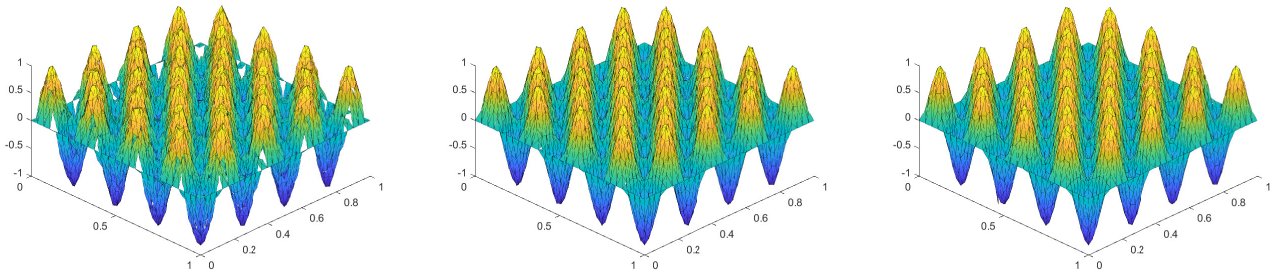
Table 4.1: Numerical results using $\mathbb{P}_1$ elements and $c = 8$ when the R-BP-FEM is used.

Mesh	Itr.	$\|u-\mathcal{E}^+(u_H)\|_{0,\Omega}$	EOC	$\|u-\mathcal{E}^+(u_H)\|_{1,\Omega}$	EOC	$\|u-\mathcal{E}^+(u_H)\|_{\mathrm{DG}}$	EOC	$\|\mathcal{E}^-(u_H^n)\|_s$	EOC
Mesh 4.4c	3	4.20e-1	0.47	1.69e-2	–	4.40e-1	0.35	0	–
Mesh 4.4d	9	2.32e-1	1.71	1.28e-2	0.80	2.70e-1	1.41	3.73e-3	–
Mesh 4.4e	10	1.12e-1	1.88	7.96e-3	1.37	1.35e-1	2.00	4.12e-2	–
Mesh 4.4f	13	4.41e-2	2.91	4.45e-3	1.68	5.93e-2	2.37	1.58e-2	2.26
Mesh 4.4g	13	1.34e-2	3.43	2.33e-3	1.87	2.12e-2	2.96	3.12e-3	5.18
Mesh 4.4h	14	4.67e-3	3.04	1.12e-3	2.11	9.66e-3	2.27	1.63e-3	1.87

Table 4.2: Numerical results using $\mathbb{P}_2$ elements and $c = 8$ when the R-BP-FEM is used.

Mesh	Itr.	$\|u-\mathcal{E}^+(u_H)\|_{0,\Omega}$	EOC	$\|u-\mathcal{E}^+(u_H)\|_{1,\Omega}$	EOC	$\|u-\mathcal{E}^+(u_H)\|_{\mathrm{DG}}$	EOC	$\|\mathcal{E}^-(u_H^n)\|_s$	EOC
Mesh 4.4c	11	2.83e-1	1.30	1.32e-2	0.80	3.18e-1	1.53	4.54e-2	–
Mesh 4.4d	13	1.04e-1	2.99	7.48e-3	1.64	1.45e-1	2.27	3.03e-2	1.17
Mesh 4.4e	14	2.72e-2	3.77	2.85e-3	2.78	3.47e-2	4.13	4.25e-3	5.67
Mesh 4.4f	17	6.52e-3	4.12	1.04e-3	2.91	9.56e-3	3.71	2.73e-3	1.28
Mesh 4.4g	13	1.73e-3	3.82	3.61e-4	3.05	2.72e-3	3.15	4.01e-4	5.53
Mesh 4.4h	14	4.31e-4	4.01	1.27e-4	3.01	8.00e-4	3.53	5.71e-5	5.62

Table 4.3: Numerical results using $\mathbb{P}_3$ elements and $c = 8$ when the R-BP-FEM is used.



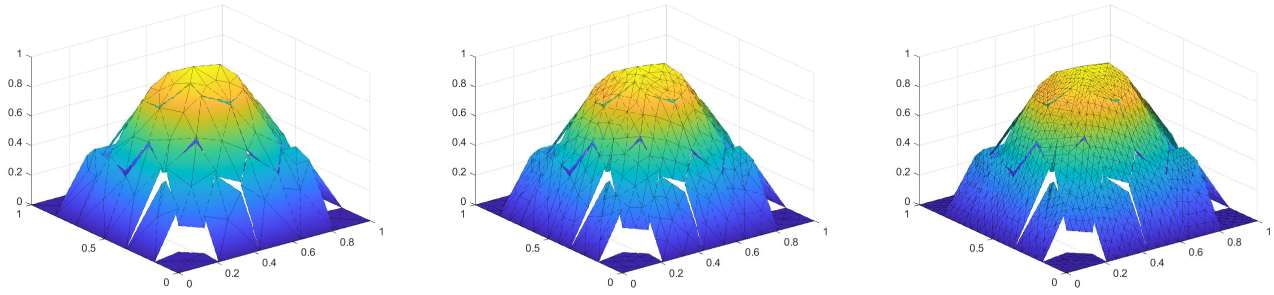
(a) $\mathbb{P}_1$ elements (b) $\mathbb{P}_2$ elements (c) $\mathbb{P}_3$ elements

Figure 4.6: Discrete solution $\mathcal{E}^+(u_H)$ of Example 8 for $c = 8$ and Mesh 4.4h using R-BP-FEM.

To examine the effect of refining the triangular submesh, we used two levels of refinement on the Mesh 4.4c which have been depicted in Figure 4.5. The results are presented in Tables 4.4 and 4.5, and the approximated solution is illustrated in Figures 4.7 and 4.8 for $\mathbb{P}_1$ and $\mathbb{P}_2$ elements. These results indicate that refining the submesh does not improve the convergence outcomes, as all convergence results almost remain consistent across different refinements of the fixed Mesh 4.4c. The first level of the mesh (i.e., Mesh 4.4c) has been denoted by R1, with subsequent refinements represented as R2, R3, and so on.

Mesh	Itr.	$\|u-\mathcal{E}^+(u_H)\|_{0,\Omega}$	$\|u-\mathcal{E}^+(u_H)\|_{0,\partial\Omega}$	$\|u-u_H\|_{0,\partial\Omega}$	$\|u-\mathcal{E}^+(u_H)\|_{1,\Omega}$	$\|u-\mathcal{E}^+(u_H)\|_{\mathrm{DG}}$	$\|\mathcal{E}^-(u_H^n)\|_s$
R1	11	3.77e-2	4.31e-4	1.48e-2	6.86e-4	1.38e-2	7.34e-2
R2	10	3.43e-2	3.64e-4	3.35e-2	6.46e-4	3.65e-2	8.15e-2
R3	9	3.19e-2	2.59e-4	5.24e-2	6.23e-4	3.38e-2	7.82e-2

Table 4.4: Numerical results using R-BP-FEM and $\mathbb{P}_1$ elements with $c = 1$ on Mesh 4.4c and two levels of its refinement.

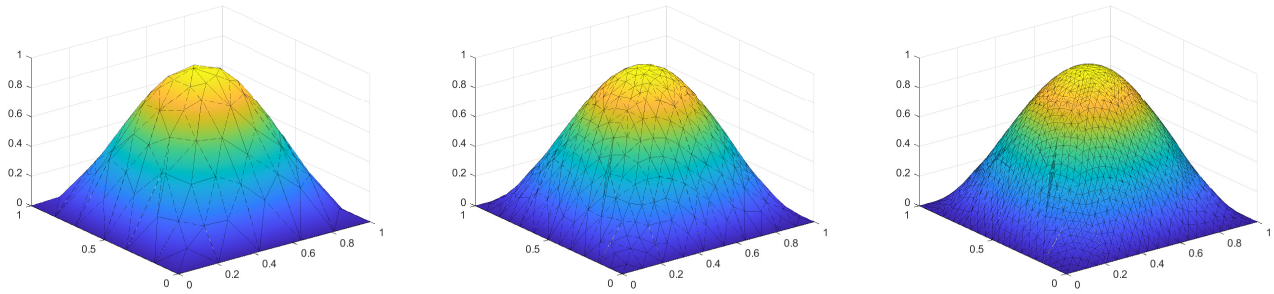


(a) R1 (b) R2 (c) R3

Figure 4.7: Discrete solution $\mathcal{E}^+(u_H)$ for Example 8 using the R-BP-FEM and $\mathbb{P}_1$ elements with $c = 1$ on Mesh 4.4c and two levels of its refinement.

Mesh	Itr.	$\|u-\mathcal{E}^+(u_H)\|_{0,\Omega}$	$\|u-\mathcal{E}^+(u_H)\|_{0,\partial\Omega}$	$\|u-u_H\|_{0,\partial\Omega}$	$\|u-\mathcal{E}^+(u_H)\|_{1,\Omega}$	$\|u-\mathcal{E}^+(u_H)\|_{\mathrm{DG}}$	$\|\mathcal{E}^-(u_H^n)\|_s$
R1	10	2.17e-3	2.96e-5	3.78e-4	9.84e-5	2.42e-3	2.18e-2
R2	12	1.96e-3	0	1.21e-3	8.98e-5	2.20e-3	2.28e-2
R3	12	1.68e-3	0	2.52e-3	8.03e-5	1.92e-3	2.09e-2

Table 4.5: Numerical results using the R-BP-FEM and $\mathbb{P}_2$ elements with $c = 1$ on Mesh 4.4c and two levels of its refinement.



(a) R1 (b) R2 (c) R3

Figure 4.8: Approximation of the solution $\mathcal{E}^+(u_H)$ for Example 8 using the R-BP-FEM and $\mathbb{P}_2$ elements with $c = 1$ on Mesh 4.4c and two levels of its refinement.

Newton's method significantly reduces the number of iterations required to solve the nonlinear problem.

In contrast, Richardson iteration often requires several hundred iterations for this problem, and this number increases further when higher-order finite element methods are used. As shown in Tables 4.1–4.5 and in the subsequent examples, Newton's method achieves a substantial reduction in iteration counts. For comparison, Table 4.6 reports the number of Richardson iterations needed when using $\mathbb{P}_1$ elements in Example 8.

Mesh	4.4f	4.4g	4.4h
Itr.	146	148	149

Table 4.6: Richardson's iterations needed to reach convergence using $\mathbb{P}_1$ elements in Example 8.

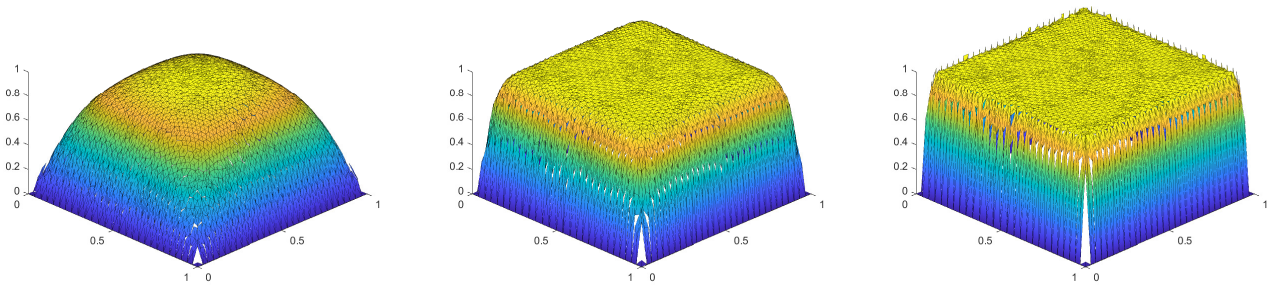
Example 9. ***Resolution of boundary layers.*** *Consider the problem*

$$\begin{cases} -\epsilon\Delta u + \mu u = 1, & \text{in } \Omega; \\ u = 0, & \text{on } \partial\Omega, \end{cases}$$

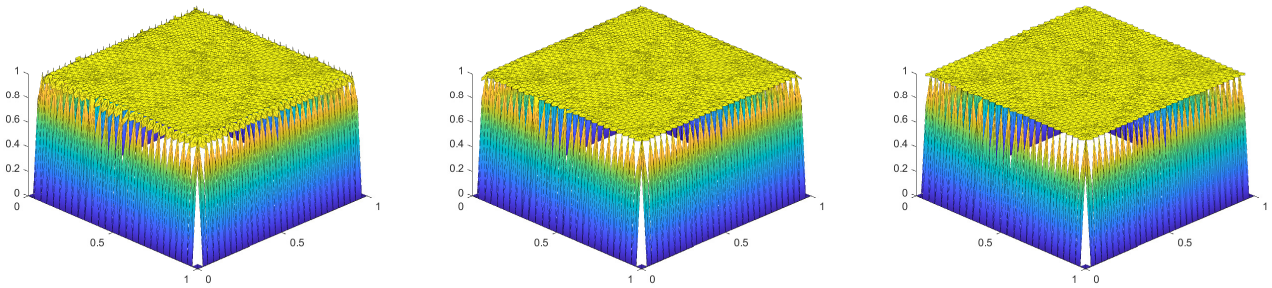
We use $\epsilon \in [10^{-7}, 10^{-2}]$. *Notice that* $u(x) \in [0, 1]$*, and thus we choose* $\kappa = 1$.

The approximations of the solution obtained using $\mathbb{P}_1$ and $\mathbb{P}_2$ elements on Mesh 4.4h are shown in Figures 4.9 and 4.10. From these figures, it is evident that as the diffusion parameter ϵ decreases, the boundary layer becomes sharper. Nevertheless, the bound-preserving finite element method successfully resolves the problem without exhibiting any noticeable oscillations near the boundary. This behaviour remains consistent when increasing the polynomial order of the elements, as both $\mathbb{P}_1$ and $\mathbb{P}_2$ discretisations yield similar results.

The number of Newton's iterations required to satisfy the stopping criterion (4.67) is listed in Table 4.7.

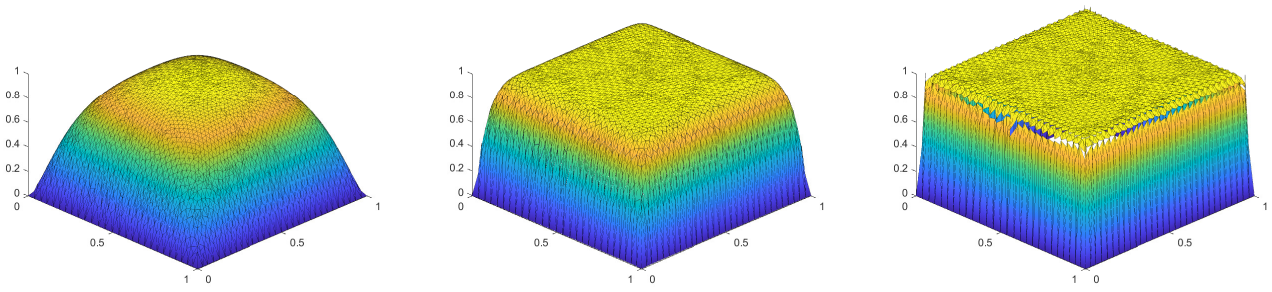


(a) $\epsilon = 10^{-2}$ (b) $\epsilon = 10^{-3}$ (c) $\epsilon = 10^{-4}$

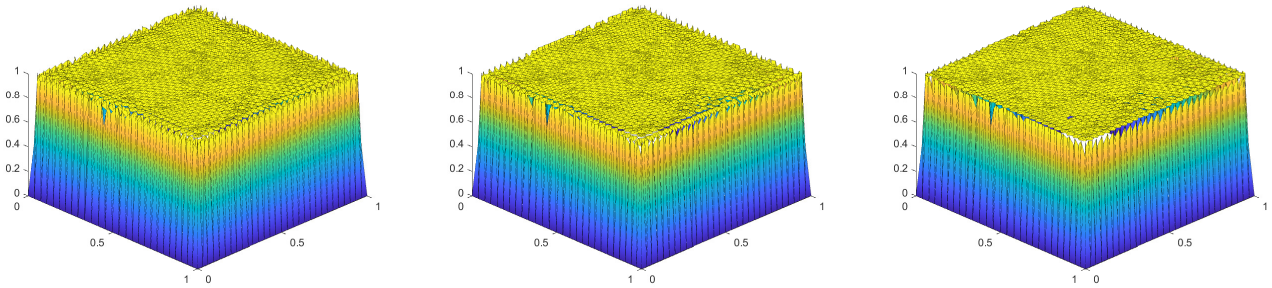


(d) $\epsilon = 10^{-5}$ (e) $\epsilon = 10^{-6}$ (f) $\epsilon = 10^{-7}$

Figure 4.9: Elevations of the approximation solution $\mathcal{E}^+(u_H)$ to Example 9 using $\mathbb{P}_1$ elements and Mesh 4.4h.



(a) $\epsilon = 10^{-2}$ (b) $\epsilon = 10^{-3}$ (c) $\epsilon = 10^{-4}$



(d) $\epsilon = 10^{-5}$ (e) $\epsilon = 10^{-6}$ (f) $\epsilon = 10^{-7}$

Figure 4.10: Elevations of the approximation solution $\mathcal{E}^+(u_H)$ to Example 9 using $\mathbb{P}_2$ elements and Mesh 4.4h.

	ϵ	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}
$\mathbb{P}_1$	Itr.	13	14	15	16	6	7
$\mathbb{P}_2$	Itr.	14	14	14	20	7	8

Table 4.7: Iterations required to satisfy the stopping criterion (4.67) using $\mathbb{P}_1$ and $\mathbb{P}_2$ elements, and Mesh 4.4h obtained with the R-BP-FEM.

Example 10. A solution with an interior layer *Consider the problem*

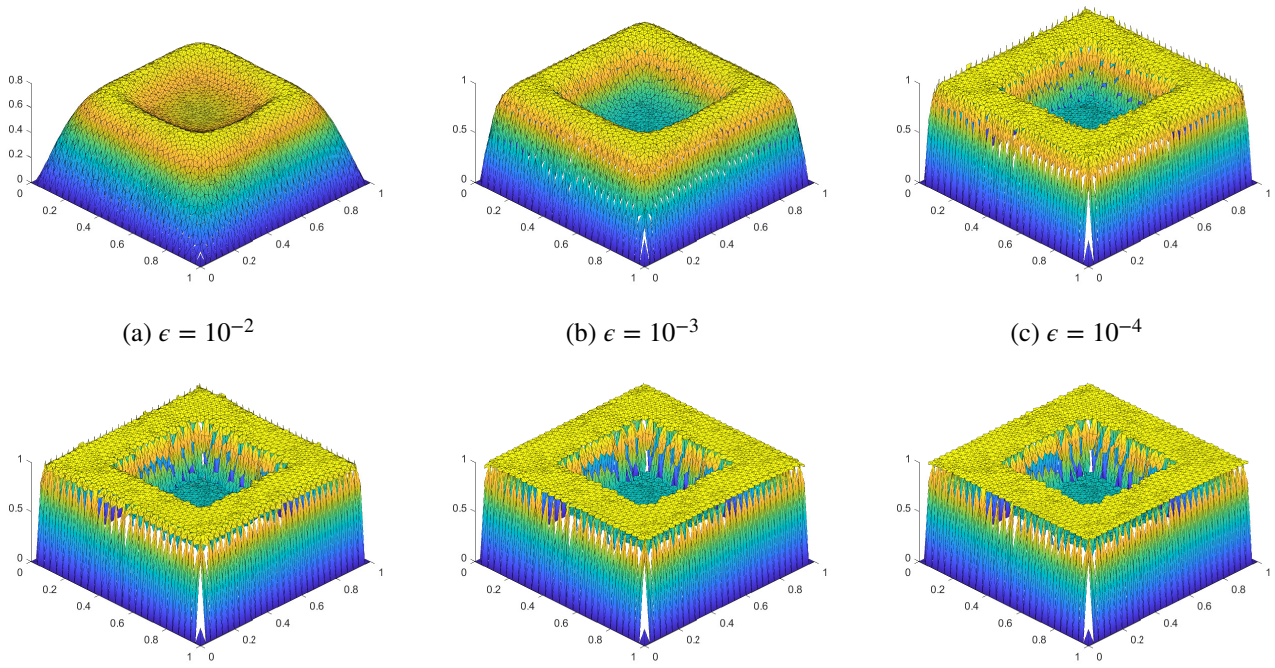
$$\begin{cases} -\epsilon\Delta u + u = f, & \text{in } \Omega; \\ u = 0, & \text{on } \partial\Omega, \end{cases} \tag{4.69}$$

with

$$f = \begin{cases} \frac{1}{2}, & \text{in } \left[\frac{1}{4}, \frac{3}{4}\right]^2; \\ 1, & \text{otherwise.} \end{cases}$$

In this case, the solution is expected to achieve a local minimum on the interior. We examine the solution for $\epsilon \in [10^{-7}, 10^{-2}]$. *Notice that* $u(x) \in [0, 1]$, *and thus we choose* $\kappa = 1$.

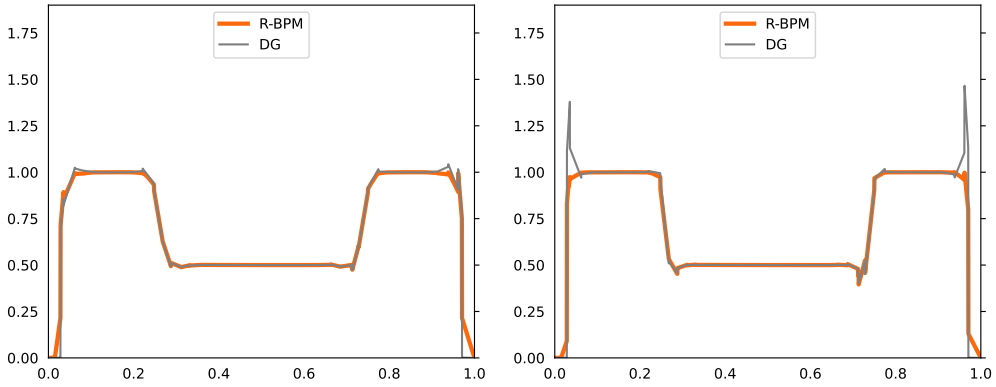
The approximations of the solution using $\mathbb{P}_1$ and $\mathbb{P}_2$ elements are shown in Figures 4.11 and 4.14. Additionally, numerical results, including the number of iterations required to satisfy the stopping criterion (4.67) is reported in Table 4.8. A cross-section of the solution along the $y = -x$ plane, obtained using the BP finite element method and the DG finite element method (4.4) on Mesh 4.4h, is presented in Figures 4.12 and 4.15. It is noteworthy that the solution of (4.4) exhibits oscillations near the boundary layer, which become severe for $\epsilon \ll 1$. These oscillations are completely eliminated by the current method. Similar numerical results for the coarser mesh 4.4e are presented in Figures 4.13 and 4.16.



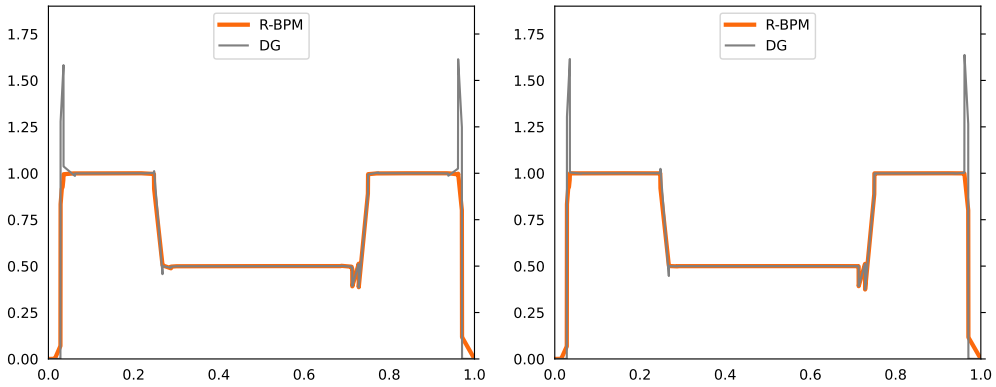
(a) $\epsilon = 10^{-2}$ (b) $\epsilon = 10^{-3}$ (c) $\epsilon = 10^{-4}$

(d) $\epsilon = 10^{-5}$ (e) $\epsilon = 10^{-6}$ (f) $\epsilon = 10^{-7}$

Figure 4.11: Elevations of the approximation to Example 10 using $\mathbb{P}_1$ elements and Mesh 4.4h.

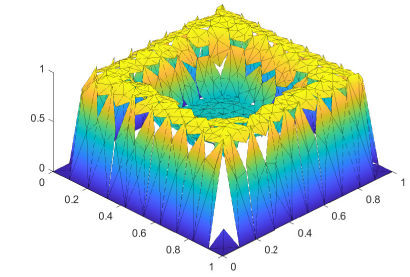


(a) $\epsilon = 10^{-4}$ (b) $\epsilon = 10^{-5}$

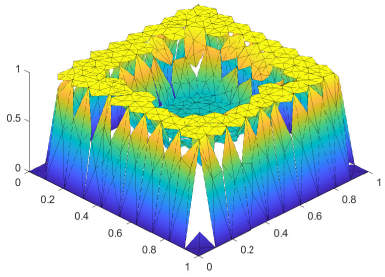


(c) $\epsilon = 10^{-6}$ (d) $\epsilon = 10^{-7}$

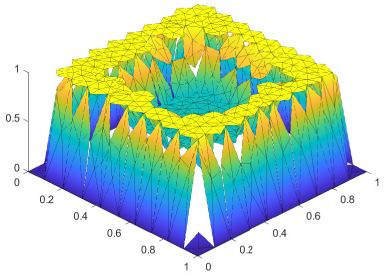
Figure 4.12: Cross-section of the solution of Example 10 along the $y = -x$ plane obtained by the R-BP-FEM and the DG method (4.4), and using $\mathbb{P}_1$ elements on Mesh 4.4h.



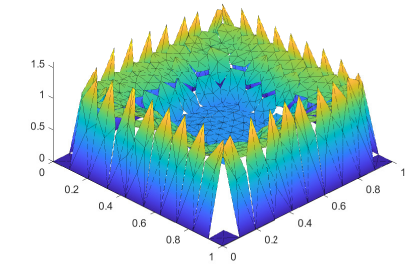
(a) $\epsilon = 10^{-4}$, R-BP-FEM



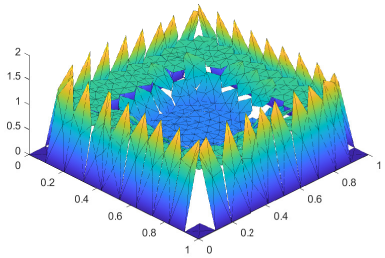
(b) $\epsilon = 10^{-5}$, R-BP-FEM



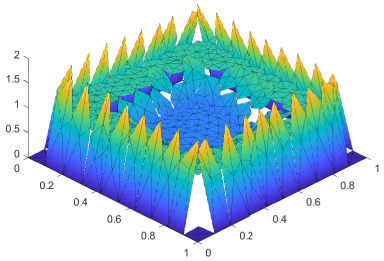
(c) $\epsilon = 10^{-7}$, R-BP-FEM



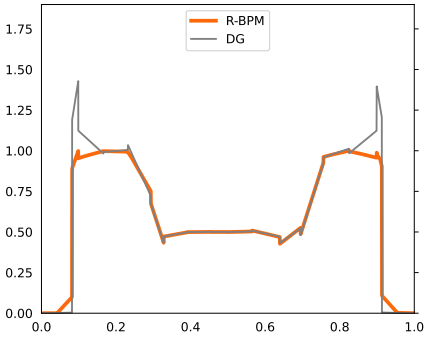
(d) $\epsilon = 10^{-4}$, DG-FEM



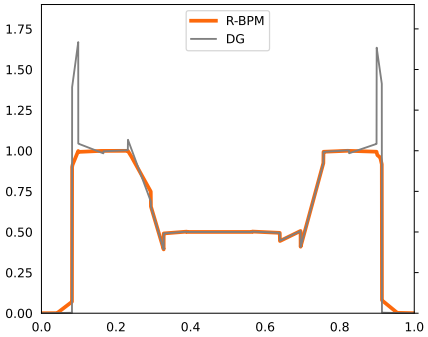
(e) $\epsilon = 10^{-5}$, DG-FEM



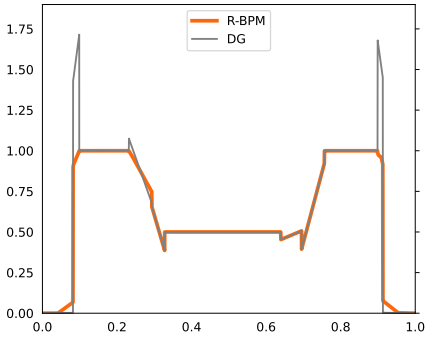
(f) $\epsilon = 10^{-7}$, DG-FEM



(g) $\epsilon = 10^{-4}$

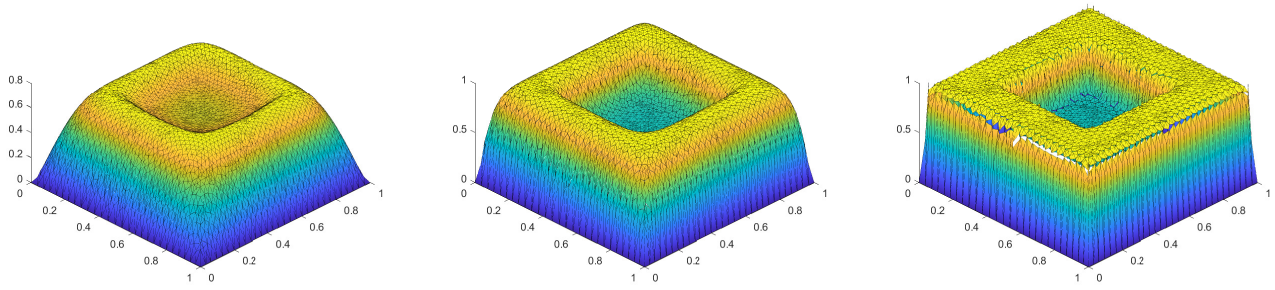


(h) $\epsilon = 10^{-5}$

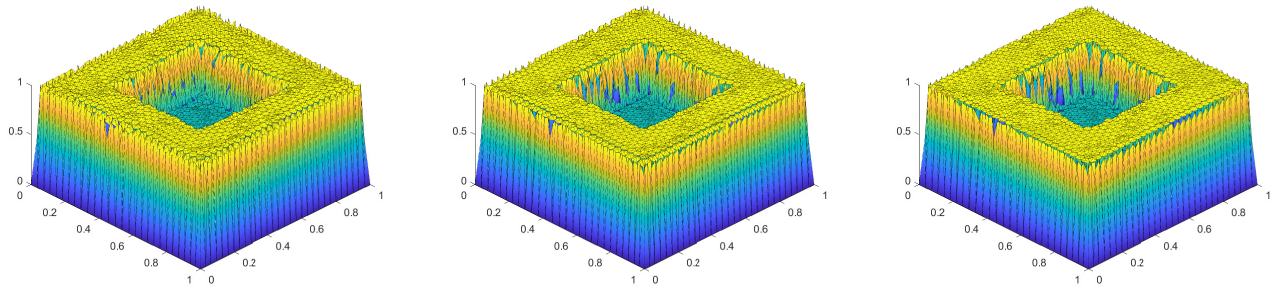


(i) $\epsilon = 10^{-7}$

Figure 4.13: Cross-section of the solution of Example 10 along the $y = -x$ plane obtained by the R-BP-FEM and the DG method (4.4), and using $\mathbb{P}_1$ elements on Mesh 4.4f.



(a) $\epsilon = 10^{-2}$ (b) $\epsilon = 10^{-3}$ (c) $\epsilon = 10^{-4}$

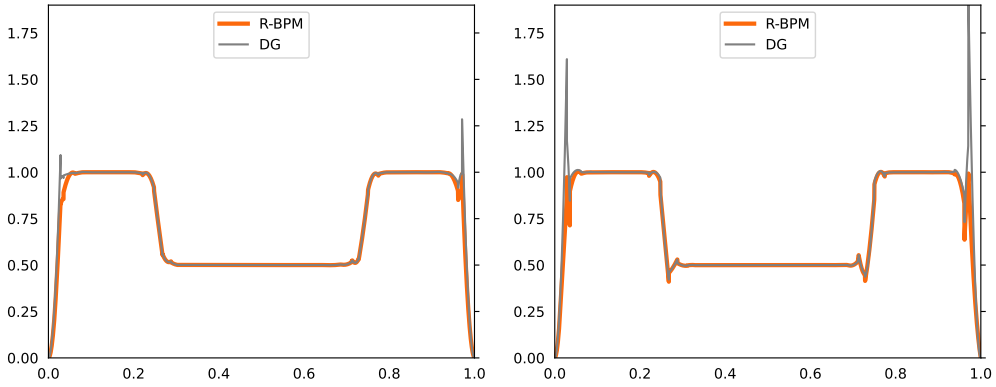


(d) $\epsilon = 10^{-5}$ (e) $\epsilon = 10^{-6}$ (f) $\epsilon = 10^{-7}$

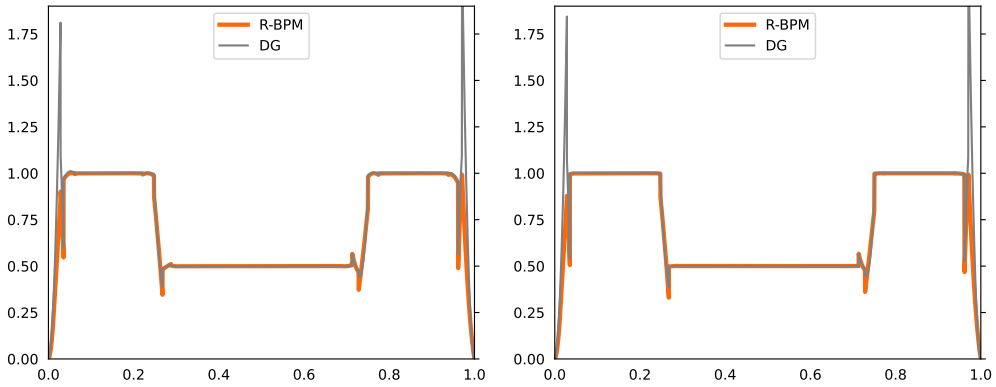
Figure 4.14: Elevations of the approximation to Example 10 using $\mathbb{P}_2$ elements and Mesh 4.4h.

	ϵ	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}
$\mathbb{P}_1$	Itr.	13	14	14	17	13	7
$\mathbb{P}_2$	Itr.	9	7	21	14	13	9

Table 4.8: Iterations required to satisfy the stopping criterion (4.67) using $\mathbb{P}_1$ and $\mathbb{P}_2$ elements on Mesh 4.4h, obtained with the R-BP-FEM.

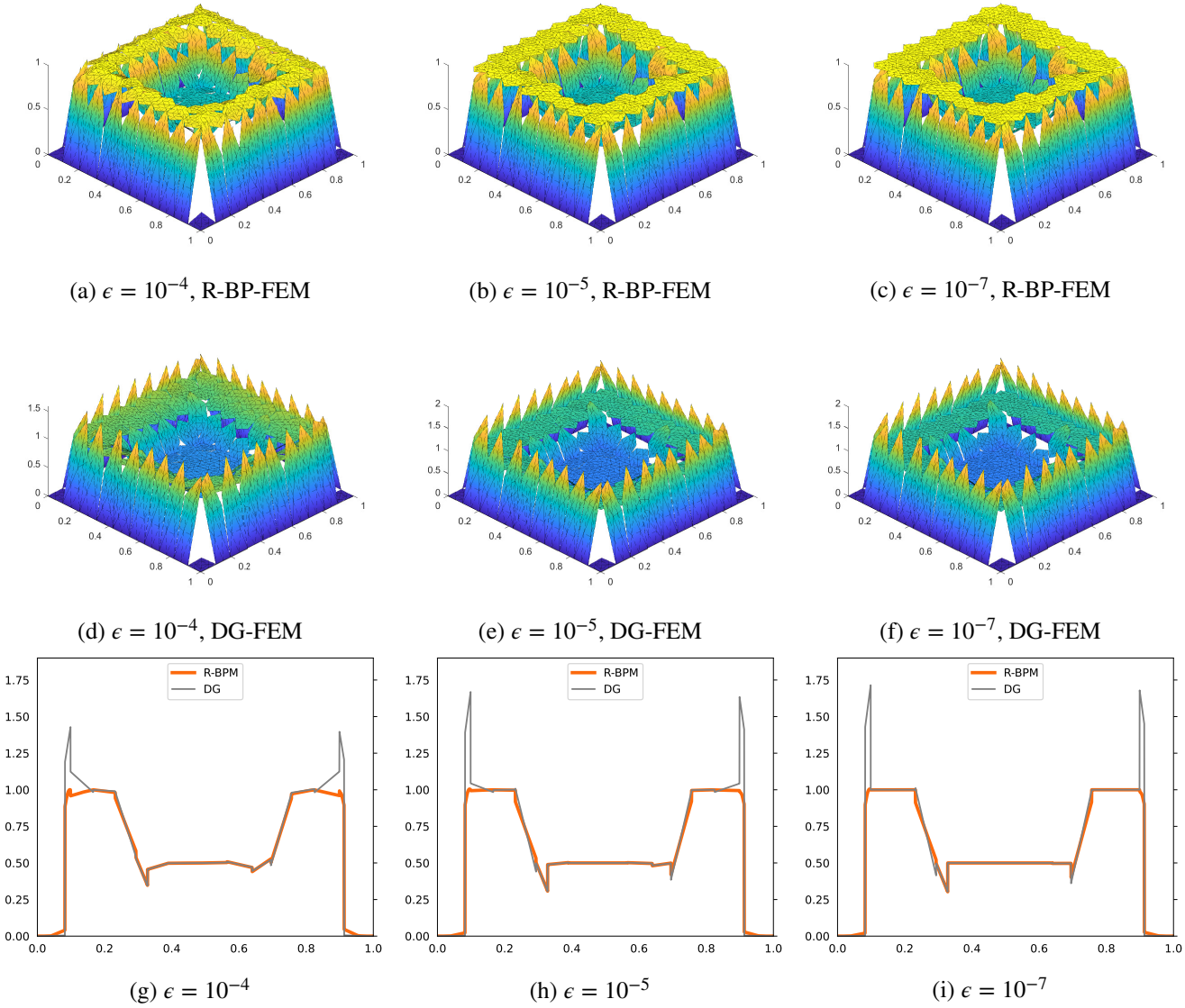


(a) $\epsilon = 10^{-4}$ (b) $\epsilon = 10^{-5}$



(c) $\epsilon = 10^{-6}$ (d) $\epsilon = 10^{-7}$

Figure 4.15: Cross-section of the solution of Example 10 along the $y = -x$ plane obtained by the R-BP-FEM and the DG method (4.4), and using $\mathbb{P}_2$ elements on Mesh 4.4h.



(a) $\epsilon = 10^{-4}$, R-BP-FEM (b) $\epsilon = 10^{-5}$, R-BP-FEM (c) $\epsilon = 10^{-7}$, R-BP-FEM

(d) $\epsilon = 10^{-4}$, DG-FEM (e) $\epsilon = 10^{-5}$, DG-FEM (f) $\epsilon = 10^{-7}$, DG-FEM

(g) $\epsilon = 10^{-4}$ (h) $\epsilon = 10^{-5}$ (i) $\epsilon = 10^{-7}$

Figure 4.16: Cross-section of the solution of Example 10 along the $y = -x$ plane obtained by the R-BP-FEM and the DG method (4.4), and using $\mathbb{P}_2$ elements on Mesh 4.4f.

The numerical results presented in this section demonstrate that the proposed finite element method is robust in enforcing bounds on the solution. Furthermore, the tests confirm optimal convergence rates for smooth solutions, robustness of the scheme across a wide range of diffusion parameters, and the ability of the method to accurately capture both boundary and interior layers without producing spurious oscillations or out-of-range over- or undershoots. Comparisons with a standard DG discretisation show that the proposed method effectively eliminates non-physical behaviour in challenging, singularly perturbed regimes. A notable feature of the approach is the flexibility to freely choose the number and location of nodes at which

range bounds are enforced, without increasing the overall computational complexity of the method.

Chapter 5

Conclusion

The development of finite element methods (FEMs) that respect the underlying physical principles of partial differential equations (PDEs) is an important aspect of numerical analysis. The preceding chapters of this thesis have addressed challenges associated with maintaining these physical principles in numerical approximations, with a particular focus on bound-preserving schemes for convection-diffusion and reaction-diffusion problems. This concluding chapter summarises the main findings, highlights the contributions of this research, and discusses potential future directions.

In Chapter 2, we extended the methodology proposed in [12] to convection-diffusion problems, addressing the challenge of bound preservation in numerical methods. It has long been recognised that standard finite element methods do not inherently preserve physical bounds in this context. This issue was formally established in the finite element setting in [46], where it was shown that piecewise linear finite element approximations maintain such bounds only under specific mesh conditions, particularly concerning internal angles and refinement levels. Additionally, for conservation laws, finite element methods that guarantee physical bound preservation are either limited to first-order accuracy or must incorporate nonlinearity due to the constraints imposed by the Godunov order barrier theorem (see, e.g., [65]).

As a result, over the past few decades, various approaches have been developed to enforce global bound preservation. A particular focus has been given to a stronger property: the *Discrete Maximum Principle* (DMP). Several methods that satisfy DMP have been proposed, particularly for convection-dominated problems (see [15, 31, 91, 103, 119], among others, and the review in [17]). Most of these methods introduce nonlinear stabilisation, modifying the standard Galerkin scheme by adding localised diffusion terms to prevent spurious oscillations and local violations of the maximum principle. Interestingly, despite their nonlinear nature, these finite element methods are often formulated using piecewise linear elements. Extending

such approaches to higher-order elements poses additional challenges, including stronger mesh restrictions. For example, it has been demonstrated in [73] that a monotone discretisation for the Poisson equation in two dimensions can be achieved with quadratic elements only if the mesh consists of equilateral triangles or squares with arbitrarily chosen diagonals. Moreover, there is limited analytical work on nonlinear finite element methods with higher-order polynomials.

In many practical scenarios, numerical stability does not require the discrete solution to be entirely free of spurious oscillations, but only to satisfy global bounds. This relaxation simplifies the problem, allowing for a wider range of methodologies. A straightforward approach is *cut-off filtering*, where values outside the admissible range are truncated. This method, frequently used in practice, has been rigorously analysed for linear reaction-diffusion equations in [88] and for parabolic problems in [101]. Another strategy is *problem reformulation*, where the numerical scheme is designed to inherently satisfy bounds, as seen in applications to chemotaxis [74] and non-Newtonian fluid mechanics [63]. Alternatively, bounds can be imposed via *inequality constraints*, treating the problem as a constrained optimisation formulation [61]. Such constraints can also be handled using *Lagrange multipliers*, leading to an extended system, as in [44, 115], where a semi-smooth Newton method was introduced to handle non-smoothness.

The methodology which has been proposed in [12] employs a bound-preserving framework for reaction-diffusion equations that relies on defining an admissible set of functions and projecting the numerical solution onto this set via an algebraic projection. This ensures that the computed solution remains within physical bounds.

Extending this methodology to the convection-diffusion equations introduces additional challenges due to the non-symmetric and nonlinear nature of the convection–diffusion equations. In particular, the presence of a convection term weakens the stability of the finite element method applied to the PDE, necessitating the use of suitable stabilisation techniques to ensure numerical stability. To address this issue, we add a linear stabilisation term to the equations i.e. continuous interior penalty (CIP) term. This stabilisation plays a crucial role in improving the robustness and efficiency of the nonlinear solver (which is used to solve the nonlinear system of the equations) used after discretisation, while also mitigating spurious oscillations, particularly within the domain.

The theoretical analysis establishes the well-posedness of the proposed approach. Also, an optimal error estimate has been proven for this finite element method.

Numerical experiments confirm the effectiveness of the method, showing that it is competitive with existing approaches in preserving solution bounds while incurring lower computational costs. This is especially evident where standard Galerkin methods and the continuous interior penalty (CIP) method failed to suppress

spurious oscillations close to the boundary. Furthermore, unlike previous approaches, the proposed method remains applicable to non-Delaunay meshes and exhibits robust performance under stringent conditions, such as non-acute meshes, even without additional mesh refinement.

In Chapter 3, we have built upon the framework established in Chapter 2 and extended it to the time-dependent convection-diffusion equations. Our theoretical analysis focused on stability and error estimates within the implicit Euler scheme, while numerical experiments shows that the method also performs well with the Crank-Nicolson time discretisation. In both cases, the numerical results confirm that the solution remains within the physical bounds without excessive smearing of layers, thereby preserving the bounds of the exact solution.

It is worth noting that alternative bound-preserving strategies for time-dependent convection-diffusion problems, such as linearised explicit and implicit flux-corrected transport (FCT) methods [94], offer more economical approaches. However, these methods impose additional constraints, such as the CFL condition, which is required to ensure bound preservation—an issue that our approach does not face. Another important consideration is the method's applicability to higher-order elements. While FCT methods are primarily designed for linear finite elements, bound preservation for higher-order elements remains an open problem due to the lack of rigorous analysis for the FCT methods.

Furthermore, enhancements to the nonlinear solver can improve computational efficiency. In Chapter 3, we employed a simple Richardson-type solver to emphasise the method's straightforward implementation. However, more advanced nonlinear solvers, such as localised Newton methods [6] and active set strategies [7], have shown the potential to significantly accelerate convergence. Preliminary numerical experiments suggest that these solvers can drastically reduce computational costs while maintaining the robustness of the proposed approach.

Overall, the method developed in this chapter provides a stable and flexible framework for solving time-dependent convection-diffusion problems, with strong theoretical guarantees and promising numerical performance. Finally, the performance of the scheme is illustrated through numerical experiments presented at the end of this chapter. These experiments demonstrate the efficiency of the method in enforcing bounds on the solution and confirm the robustness of our finite element method.

In Chapter 4, we extended the bound-preserving finite element method to polytopic meshes by using a discontinuous Galerkin method. With these meshes, the set of degrees of freedom remains independent of the number of vertices, edges, or faces within each element. In other words, the degrees of freedom are not associated with physical points, and the basis functions are defined over the entire domain. In this finite lement method we employ a sub-triangulation approach, allowing bound-preserving constraints to be

enforced at each degree of freedom within the sub-triangulated mesh.

The use of polytopic meshes offers significant advantages in terms of computational efficiency and flexibility. By allowing elements with an arbitrary number of faces, polytopic meshes can represent complex geometries with fewer elements, reducing both computational cost and memory requirements. Moreover, they provide greater adaptability for adaptive mesh refinement, interface handling, and dynamic mesh modifications. These properties make them particularly well-suited for applications in Eulerian and Lagrangian frameworks, multilevel solvers, and problems with evolving interfaces.

As mentioned above, one of the main challenges in extending the bound-preserving methodology to polytopic elements is the selection of appropriate degrees of freedom for enforcing constraints. Inspired by the recovered finite element method (RFEM) introduced in [57], we addressed this challenge by defining an underlying sub-triangulation that enables a structured implementation of the bound-preserving method. This approach ensures that numerical solutions remain within physically bounds of the solution.

Theoretical analysis provided rigorous justification for the well-posedness of the method for reaction-diffusion equations, and an optimal error estimate for the solution of the finite element was proved. The numerical experiments demonstrated its robustness and effectiveness of the finite element method. Compared to standard approaches, DG offers a practical and efficient alternative that eliminates the limitations imposed by the topology-dependent approximation spaces in polygonal finite element and virtual element methods.

The results presented in this thesis contribute to the ongoing development of robust and accurate finite element methods for convection-diffusion and reaction-diffusion problems. The bound-preserving approach introduced here offers a systematic framework for enforcing physical bound in numerical approximations. Moreover, the incorporation of stabilisation techniques enhances the reliability of the methods in convection-dominated regimes.

The theoretical analysis presented in this thesis has been supported by numerical experiments, which confirm the stability and accuracy of the proposed methods. These results highlight the potential of bound-preserving finite element techniques proposed in this thesis in improving the robustness of numerical schemes for PDEs, particularly in cases where standard methods fail to maintain physical bounds of the solutions.

Future research directions include extending these methods to discontinuous Galerkin schemes for convection-diffusion problems, applying bound-preserving techniques to tensor PDEs to preserve the range of the eigenvalues, and tackling more complex PDEs such as coupled multi-physics problems. Another important area is developing adaptive strategies that improve computational efficiency while ensuring the solution remains within the correct bounds. Also, creating new stabilisation techniques and incorporating bound-preserving methods into existing finite element frameworks could further enhance the accuracy and stability of numer-

ical simulations. Also, future work will focus on extending the analysis to more general time discretisations, investigating adaptive refinement strategies, and incorporating more efficient nonlinear solvers to further enhance performance.

Bibliography

[1] Gabriel Acosta and Ricardo Durán. An optimal Poincaré inequality in l^1 for convex domains. *Proceedings of the American Mathematical Society*, 132(1):195–202, 2004.

[2] Alejandro Allendes, Gabriel R Barrenechea, and Richard Rankin. Fully computable error estimation of a nonlinear, positivity-preserving discretization of the convection-diffusion-reaction equation. *SIAM Journal on Scientific Computing*, 39(5):A1903–A1927, 2017.

[3] Abdolreza Amiri, Gabriel R Barrenechea, Emmanuil H Georgoulis, and Tristan Pryer. A nodally bound-preserving composite discontinuous galerkin method on polytopic meshes. *arXiv preprint arXiv:2510.02094*, 2025.

[4] Abdolreza Amiri, Gabriel R. Barrenechea, and Tristan Pryer. A nodally bound-preserving finite element method for reaction–convection–diffusion equations. *Math. Models Methods Appl. Sci.*, 34(8):1533–1565, 2024.

[5] Abdolreza Amiri, Gabriel R. Barrenechea, and Tristan Pryer. A nodally bound-preserving finite element method for time-dependent convection–diffusion equations. *Math. Models Methods Appl. Sci.*, 34(8):1533–1565, 2024.

[6] Ioannis K Argyros. On newton's method and nondiscrete mathematical induction. *Bulletin of the Australian Mathematical Society*, 38(1):131–140, 1988.

[7] Ben S. Ashby, Abdalaziz Hamdan, and Tristan Pryer. A nodally bound-preserving finite element method for hyperbolic convection-reaction problems, 2025.

[8] Santiago Badia and Alba Hierro. On monotonicity-preserving stabilized finite element approximations of transport problems. *SIAM Journal on Scientific Computing*, 36(6):A2673–A2697, 2014.

[9] Santiago Badia and Alba Hierro. On discrete maximum principles for discontinuous Galerkin methods. *Computer Methods in Applied Mechanics and Engineering*, 286:107–122, 2015.

[10] Gabriel R Barrenechea, Erik Burman, and Fotini Karakatsani. Edge-based nonlinear diffusion for finite element approximations of convection–diffusion equations and its relation to algebraic flux-correction schemes. *Numerische Mathematik*, 135:521–545, 2017.

[11] Gabriel R. Barrenechea, Erik Burman, and Fotini Karakatsani. Edge-based nonlinear diffusion for finite element approximations of convection-diffusion equations and its relation to algebraic flux-correction schemes. *Numer. Math.*, 135(2):521–545, 2017.

[12] Gabriel R Barrenechea, Emmanuil H Georgoulis, Tristan Pryer, and Andreas Veeser. A nodally bound-preserving finite element method. *IMA Journal of Numerical Analysis*, 44(4):2198–2219, 2024.

[13] Gabriel R Barrenechea, Volker John, and Petr Knobloch. Some analytical results for an algebraic flux correction scheme for a steady convection–diffusion equation in one dimension. *IMA Journal of Numerical Analysis*, 35(4):1729–1756, 2015.

[14] Gabriel R Barrenechea, Volker John, and Petr Knobloch. Analysis of algebraic flux correction schemes. *SIAM Journal on Numerical Analysis*, 54(4):2427–2451, 2016.

[15] Gabriel R. Barrenechea, Volker John, and Petr Knobloch. An algebraic flux correction scheme satisfying the discrete maximum principle and linearity preservation on general meshes. *Math. Models Methods Appl. Sci.*, 27(3):525–548, 2017.

[16] Gabriel R Barrenechea, Volker John, and Petr Knobloch. Finite element methods respecting the discrete maximum principle for convection-diffusion equations. *SIAM Review*, 66(1):3–88, 2024.

[17] Gabriel R. Barrenechea, Volker John, and Petr Knobloch. Finite element methods respecting the discrete maximum principle for convection-diffusion equations. *SIAM Review*, 66(1):3–88, 2024.

[18] Gabriel R Barrenechea, Tristan Pryer, and Alex Trenam. A nodally bound-preserving discontinuous Galerkin method for the drift-diffusion equation. *arXiv preprint arXiv:2410.05040*, 2024.

[19] Mario Bebendorf. A note on the Poincaré inequality for convex domains. *Zeitschrift für Analysis und ihre Anwendungen*, 22(4):751–756, 2003.

[20] Roland Becker and Malte Braack. A two-level stabilization scheme for the Navier-Stokes equations. In *Numerical mathematics and advanced applications*, pages 123–130. Springer, Berlin, 2004.

[21] L Beirão da Veiga, Franco Brezzi, Andrea Cangiani, Gianmarco Manzini, L Donatella Marini, and Alessandro Russo. Basic principles of virtual element methods. *Mathematical Models and Methods in Applied Sciences*, 23(01):199–214, 2013.

[22] Pavel Bochev, Denis Ridzal, Marta D'Elia, Mauro Perego, and Kara Peterson. Optimization-based, property-preserving finite element methods for scalar advection equations and their connection to algebraic flux correction. *Computer Methods in Applied Mechanics and Engineering*, 367:112982, 2020.

[23] Pavel B. Bochev, Max D. Gunzburger, and John N. Shadid. Stability of the supg finite element method for transient advection–diffusion problems. *Computer Methods in Applied Mechanics and Engineering*, 193(23):2301–2323, 2004.

[24] James H Bramble and BE Hubbard. On the formulation of finite difference analogues of the dirichlet problem for Poisson's equation. *Numerische Mathematik*, 4(1):313–327, 1962.

[25] James H Bramble and Bert E Hubbard. New monotone type approximations for elliptic problems. *Mathematics of Computation*, 18(87):349–367, 1964.

[26] Jan H Brandts, Sergey Korotov, and Michal Křížek. The discrete maximum principle for linear simplicial finite element approximations of a reaction–diffusion problem. *Linear Algebra and its Applications*, 429(10):2344–2357, 2008.

[27] Haim Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011.

[28] Alexander N. Brooks and Thomas J. R. Hughes. Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations. *Comput. Methods Appl. Mech. Engrg.*, 32(1-3):199–259, 1982. FENOMECH "81, Part I (Stuttgart, 1981).

[29] Erik Burman and Alexandre Ern. Nonlinear diffusion and discrete maximum principle for stabilized Galerkin approximations of the convection–diffusion-reaction equation. *Computer Methods in Applied Mechanics and Engineering*, 191(35):3833–3855, 2002.

[30] Erik Burman and Alexandre Ern. Discrete maximum principle for Galerkin approximations of the Laplace operator on arbitrary meshes. *Comptes Rendus Mathematique*, 338(8):641–646, 2004.

[31] Erik Burman and Alexandre Ern. Stabilized Galerkin approximation of convection-diffusion-reaction equations: discrete maximum principle and convergence. *Math. Comp.*, 74(252):1637–1652 (electronic), 2005.

[32] Erik Burman and Alexandre Ern. Stabilized Galerkin approximation of convection-diffusion-reaction equations: discrete maximum principle and convergence. *Mathematics of computation*, 74(252):1637–1652, 2005.

[33] Erik Burman and Alexandre Ern. A nonlinear consistent penalty method weakly enforcing positivity in the finite element approximation of the transport equation. *Computer Methods in Applied Mechanics and Engineering*, 320:122–132, 2017.

[34] Erik Burman and Miguel A. Fernández. Finite element methods with symmetric stabilization for the transient convection–diffusion–reaction equation. *Computer Methods in Applied Mechanics and Engineering*, 198(33):2508–2519, 2009.

[35] Erik Burman and Peter Hansbo. Edge stabilization for Galerkin approximations of convection-diffusion-reaction problems. *Comput. Methods Appl. Mech. Engrg.*, 193(15-16):1437–1453, 2004.

[36] Erik Burman and Peter Hansbo. Edge stabilization for Galerkin approximations of convection-diffusion-reaction problems. *Comput. Methods Appl. Mech. Engrg.*, 193(15-16):1437–1453, 2004.

[37] Andrea Cangiani, Zhaonan Dong, and Emmanuil H. Georgoulis. *hp*-version space-time discontinuous Galerkin methods for parabolic problems on prismatic meshes. *SIAM J. Sci. Comput.*, 39(4):A1251–A1279, 2017.

[38] Andrea Cangiani, Zhaonan Dong, and Emmanuil H. Georgoulis. *hp*-version discontinuous Galerkin methods on essentially arbitrarily-shaped elements. *Math. Comp.*, 91(333):1–35, 2021.

[39] Andrea Cangiani, Zhaonan Dong, Emmanuil H Georgoulis, and Paul Houston. hp-version discontinuous galerkin methods for advection-diffusion-reaction problems on polytopic meshes. *ESAIM: Mathematical Modelling and Numerical Analysis*, 50(3):699–725, 2016.

[40] Andrea Cangiani, Zhaonan Dong, Emmanuil H. Georgoulis, and Paul Houston. *hp-version discontinuous Galerkin methods on polygonal and polyhedral meshes*. SpringerBriefs in Mathematics. Springer, Cham, 2017.

[41] Andrea Cangiani, Zhaonan Dong, Emmanuil H Georgoulis, and Paul Houston. *hp-Version Discontinuous Galerkin Methods on Polygonal and Polyhedral Meshes*. Springer, 2017.

[42] Andrea Cangiani, Zhaonan Dong, Emmanuil H. Georgoulis, and Paul Houston. *hp-version discontinuous Galerkin methods on polygonal and polyhedral meshes*. SpringerBriefs in Mathematics. Springer, Cham, 2017.

[43] Andrea Cangiani, Emmanuil H Georgoulis, and Paul Houston. hp-version discontinuous galerkin methods on polygonal and polyhedral meshes. *Mathematical Models and Methods in Applied Sciences*, 24(10):2009–2041, 2014.

[44] Qing Cheng and Jie Shen. A new Lagrange multiplier approach for constructing structure preserving schemes, I. Positivity preserving. *Computer Methods in Applied Mechanics and Engineering*, 391:114585, 2022.

[45] Seng-Kee Chua and Richard L Wheeden. Estimates of best constants for weighted Poincaré inequalities on convex domains. *Proceedings of the London Mathematical Society*, 93(1):197–226, 2006.

[46] P. G. Ciarlet and P.-A. Raviart. Maximum principle and uniform convergence for the finite element method. *Comput. Methods Appl. Mech. Engrg.*, 2:17–31, 1973.

[47] Philippe G Ciarlet. Discrete maximum principle for finite-difference operators. *Aequationes mathematicae*, 4(3):338–352, 1970.

[48] Philippe G. Ciarlet. *The finite element method for elliptic problems*, volume Vol. 4 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam-New York-Oxford, 1978.

[49] Philippe G Ciarlet and P-A Raviart. Maximum principle and uniform convergence for the finite element method. *Computer methods in Applied Mechanics and Engineering*, 2(1):17–31, 1973.

[50] Bernardo Cockburn, Daniele A Di Pietro, and Alexandre Ern. Bridging the hybrid high-order and hybridizable discontinuous galerkin methods. *ESAIM: Mathematical Modelling and Numerical Analysis*, 50(3):635–650, 2016.

[51] Bernardo Cockburn, Jayadeep Gopalakrishnan, and Raytcho Lazarov. Unified hybridization of discontinuous galerkin, mixed, and continuous galerkin methods for second order elliptic problems. *SIAM Journal on Numerical Analysis*, 47(2):1319–1365, 2009.

[52] Lothar Collatz. *The numerical treatment of differential equations*, volume 60. Springer Science & Business Media, 2012.

[53] Mark C Cross and Pierre C Hohenberg. Pattern formation outside of equilibrium. *Reviews of Modern Physics*, 65(3):851, 1993.

[54] Lourenço Beirao da Veiga, Konstantin Lipnikov, and Gianmarco Manzini. *The mimetic finite difference method for elliptic problems*, volume 11. Springer, 2014.

[55] Daniele Antonio Di Pietro and Alexandre Ern. *Mathematical aspects of discontinuous Galerkin methods*, volume 69 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer, Heidelberg, 2012.

[56] Zhaonan Dong and Emmanuil H. Georgoulis. Robust interior penalty discontinuous Galerkin methods. *J. Sci. Comput.*, 92(2):Paper No. 57, 23, 2022.

[57] Zhaonan Dong, Emmanuil H Georgoulis, and Tristan Pryer. Recovered finite element methods on polygonal and polyhedral meshes. *ESAIM: Mathematical Modelling and Numerical Analysis*, 54(4):1309–1337, 2020.

[58] Alexandre Ern and Jean-Luc Guermond. Weighting the edge stabilization. *SIAM Journal on Numerical Analysis*, 51(3):1655–1677, 2013.

[59] Alexandre Ern and Jean-Luc Guermond. *Finite Elements I*. Springer, 2021.

[60] Alexandre Ern and Jean-Luc Guermond. *Finite Elements II*. Springer, 2021.

[61] John A. Evans, Thomas J. R. Hughes, and Giancarlo Sangalli. Enforcement of constraints and maximum principles in the variational multiscale method. *Comput. Methods Appl. Mech. Engrg.*, 199(1-4):61–76, 2009.

[62] Lawrence C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010.

[63] Raanan Fattal and Raz Kupferman. Constitutive laws for the matrix-logarithm of the conformation tensor. *Journal of Non-Newtonian Fluid Mechanics*, 123(2):281–285, 2004.

[64] von S Gerschgorin. Fehlerabschätzung für das differenzenverfahren zur lösung partieller differentialgleichungen. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 10(4):373–382, 1930.

[65] S. K. Godunov. A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics. *Mat. Sb. (N.S.)*, 47 (89):271–306, 1959.

[66] Céline Grandmont and Sébastien Martin. Existence of solutions and continuous and semi-discrete stability estimates for 3d/0d coupled systems modelling airflows and blood flows. *ESAIM: Mathematical Modelling and Numerical Analysis*, 55(5):2365–2419, 2021.

[67] Jean-Luc Guermond. Stabilization of Galerkin approximations of transport equations by subgrid modeling. *ESAIM: Mathematical Modelling and Numerical Analysis*, 33(6):1293–1316, 1999.

[68] Jean-Luc Guermond. Stabilization of Galerkin approximations of transport equations by subgrid modeling. *M2AN Math. Model. Numer. Anal.*, 33(6):1293–1316, 1999.

[69] Jean-Luc Guermond and Murtazo Nazarov. A maximum-principle preserving c0 finite element method for scalar conservation equations. *Computer Methods in Applied Mechanics and Engineering*, 272:198–213, 2014.

[70] Jean-Luc Guermond, Murtazo Nazarov, Bojan Popov, and Yong Yang. A second-order maximum principle preserving lagrange finite element technique for nonlinear scalar conservation equations. *SIAM Journal on Numerical Analysis*, 52(4):2163–2182, 2014.

[71] Wolfgang Hackbusch and Stefan A Sauter. Composite finite elements for the approximation of pdes on domains with complicated micro-structures. *Numerische Mathematik*, 75:447–472, 1997.

[72] John G Heywood and Rolf Rannacher. Finite-element approximation of the nonstationary Navier-Stokes problem. part IV: error analysis for second-order time discretization. *SIAM Journal on Numerical Analysis*, 27(2):353–384, 1990.

[73] W. Höhn and H.-D. Mittelmann. Some remarks on the discrete maximum-principle for finite elements of higher order. *Computing*, 27(2):145–154, 1981.

[74] Xueling Huang and Jie Shen. Bound/positivity preserving SAV schemes for the Patlak-Keller-Segel-Navier-Stokes system. *Journal of Computational Physics*, 480:112034, 2023.

[75] Kenneth H Huebner, Donald L Dewhirst, Douglas E Smith, and Ted G Byrom. *The finite element method for engineers*. John Wiley & Sons, 2001.

[76] Thomas J.R. Hughes, Leopoldo P. Franca, and Gregory M. Hulbert. A new finite element formulation for computational fluid dynamics: Viii. the Galerkin/least-squares method for advective-diffusive equations. *Computer Methods in Applied Mechanics and Engineering*, 73(2):173–189, 1989.

[77] Antony Jameson. Computational algorithms for aerodynamic analysis and design. *Applied Numerical Mathematics*, 13(5):383–422, 1993.

[78] Antony Jameson. Analysis and design of numerical schemes for gas dynamics, 1: artificial diffusion, upwind biasing, limiters and their effect on accuracy and multigrid convergence. *International Journal of Computational Fluid Dynamics*, 4(3-4):171–218, 1995.

[79] Max Jensen and Axel Målqvist. Finite element convergence for the joule heating problem with mixed boundary conditions. *BIT Numerical Mathematics*, 53:475–496, 2013.

[80] Abhinav Jha. A residual based a posteriori error estimators for afc schemes for convection-diffusion equations. *Computers & Mathematics with Applications*, 97:86–99, 2021.

[81] Volker John and Petr Knobloch. On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part I–A review. *Computer Methods in Applied Mechanics and Engineering*, 196(17-20):2197–2215, 2007.

[82] Volker John, Teodora Mitkova, Michael Roland, Kai Sundmacher, Lutz Tobiska, and Andreas Voigt. Simulations of population balance systems with one internal coordinate using finite element methods. *Chemical Engineering Science*, 64(4):733–741, 2009.

[83] Claes Johnson. *Numerical solution of partial differential equations by the finite element method*. Courier Corporation, 2009.

[84] David Kinderlehrer and Guido Stampacchia. *An introduction to variational inequalities and their applications*. SIAM, 2000.

[85] Robert C. Kirby and Daniel Shapero. High-order bounds-satisfying approximation of partial differential equations via finite element variational inequalities. *Numer. Math.*, 156(3):927–947, 2024.

[86] Petr Knobloch. A new algebraically stabilized method for convection–diffusion–reaction equations. In *Numerical Mathematics and Advanced Applications ENUMATH 2019: European Conference, Egmond aan Zee, The Netherlands, September 30-October 4*, pages 605–613. Springer, 2020.

[87] Petr Knobloch and Gert Lube. Local projection stabilization for advection–diffusion–reaction problems: One-level vs. two-level approach. *Applied Numerical Mathematics*, 59(12):2891–2907, 2009.

[88] Christian Kreuzer. A note on why enforcing discrete maximum principles by a simple a posteriori cutoff is a good idea. *Numer. Methods Partial Differential Equations*, 30(3):994–1002, 2014.

[89] Dmitri Kuzmin. Algebraic flux correction I: Laws, scalar conservation. *Scientific Computation*, page 145.

[90] Dmitri Kuzmin. On the design of general-purpose flux limiters for finite element schemes. I. scalar convection. *Journal of Computational Physics*, 219(2):513–531, 2006.

[91] Dmitri Kuzmin. Algebraic flux correction for finite element discretizations of coupled systems. In M. Papadrakakis, E. Oñate, and B. Schrefler, editors, *Proceedings of the Int. Conf. on Computational Methods for Coupled Problems in Science and Engineering*, pages 1–5. CIMNE, Barcelona, 2007.

[92] Dmitri Kuzmin. Algebraic flux correction for finite element discretizations of coupled systems. *Computational Methods for Coupled Problems in Science and Engineering II, CIMNE, Barcelona*, pages 653–656, 2007.

[93] Dmitri Kuzmin. On the design of algebraic flux correction schemes for quadratic finite elements. *Journal of Computational and Applied Mathematics*, 218(1):79–87, 2008.

[94] Dmitri Kuzmin. Explicit and implicit FEM-FCT algorithms with flux linearization. *Journal of Computational Physics*, 228(7):2517–2534, 2009.

[95] Dmitri Kuzmin. Linearity-preserving flux correction and convergence acceleration for constrained galerkin schemes. *Journal of Computational and Applied Mathematics*, 236(9):2317–2337, 2012.

[96] Dmitri Kuzmin and John N Shadid. Gradient-based nodal limiters for artificial diffusion operators in finite element schemes for transport equations. *International Journal for Numerical Methods in Fluids*, 84(11):675–695, 2017.

[97] Randall J. Leveque. High-resolution conservative algorithms for advection in incompressible flow. *SIAM J. Numer. Anal.*, 33(2):627–665, 1996.

[98] J.-L. Lions and E. Magenes. *Non-Homogeneous Boundary Value Problems and Applications, Vol. I.* Springer, 1972.

[99] Richard Liska and Mikhail Shashkov. Enforcing the discrete maximum principle for linear finite element solutions of second-order elliptic problems. *Commun. Comput. Phys.*, 3(4):852–877, 2008.

[100] Christoph Lohmann, Dmitri Kuzmin, John N Shadid, and Sibusiso Mabuza. Flux-corrected transport algorithms for continuous galerkin methods based on high order bernstein finite elements. *Journal of Computational Physics*, 344:151–186, 2017.

[101] Changna Lu, Weizhang Huang, and Erik S. Van Vleck. The cutoff method for the numerical computation of nonnegative solutions of parabolic PDEs with application to anisotropic diffusion and lubrication-type equations. *Journal of Computational Physics*, 242:24–36, 2013.

[102] Charalambos G Makridakis. On the Babuška–Osborn approach to finite element analysis: L 2 estimates for unstructured meshes. *Numerische Mathematik*, 139(4):831–844, 2018.

[103] Akira Mizukami and Thomas J. R. Hughes. A Petrov-Galerkin finite element method for convection-dominated flows: an accurate upwinding technique for satisfying the maximum principle. *Comput. Methods Appl. Mech. Engrg.*, 50(2):181–193, 1985.

[104] Akira Mizukami and Thomas JR Hughes. A Petrov-Galerkin finite element method for convection-dominated flows: an accurate upwinding technique for satisfying the maximum principle. *Computer methods in Applied Mechanics and Engineering*, 50(2):181–193, 1985.

[105] Maruti Kumar Mudunuru and KB Nakshatrala. On enforcing maximum principles and achieving element-wise species balance for advection–diffusion–reaction equations under the finite element method. *Journal of Computational Physics*, 305:448–493, 2016.

[106] Lawrence E Payne and Hans F Weinberger. An optimal Poincaré inequality for convex domains. *Archive for Rational Mechanics and Analysis*, 5(1):286–292, 1960.

[107] Daniel Peterseim and Stefan A. Sauter. The composite mini element-coarse mesh computation of Stokes flows on complicated domains. *SIAM J. Numer. Anal.*, 46(6):3181–3206, 2008.

[108] Liqun Qi and Jie Sun. A nonsmooth version of Newton's method. *Mathematical Programming*, 58(1-3):353–367, 1993.

[109] Junuthula Narasimha Reddy. An introduction to the finite element method. *New York*, 27(14), 1993.

[110] Michael Renardy and Robert C. Rogers. *An introduction to partial differential equations*, volume 13 of *Texts in Applied Mathematics*. Springer-Verlag, New York, second edition, 2004.

[111] Hans-Görg Ross, Martin Stynes, and Lutz Tobiska. *Robust Numerical Methods for Singularly Perturbed Differential Equations*. SSCM, volume 24, Springer, 2008.

[112] Francesc Sagués and Irving R Epstein. Nonlinear chemical dynamics. *Dalton transactions*, (7):1201–1217, 2003.

[113] Natarajan Sukumar and Alireza Tabarraei. Conforming polygonal finite elements. *International Journal for Numerical Methods in Engineering*, 61(12):2045–2066, 2004.

[114] Michael Ulbrich. *Semismooth Newton methods for variational inequalities and constrained optimization problems in function spaces*. SIAM, 2011.

[115] J. J. W. van der Vegt, Yinhua Xia, and Yan Xu. Positivity preserving limiters for time-implicit higher order accurate discontinuous Galerkin discretizations. *SIAM Journal on Scientific Computing*, 41(3):A2037–A2063, 2019.

[116] Richard S Varga et al. Matrix iterative analysis [electronic resource].

[117] T. Warburton and J. S. Hesthaven. On the constants in hp-finite element trace inverse inequalities. *Comput. Methods Appl. Mech. Engrg.*, 192(25):2765–2773, 2003.

[118] Jinchao Xu and Ludmil Zikatanov. A monotone finite element scheme for convection-diffusion equations. *Mathematics of Computation*, 68(228):1429–1446, 1999.

[119] Jinchao Xu and Ludmil Zikatanov. A monotone finite element scheme for convection-diffusion equations. *Math. Comp.*, 68(228):1429–1446, 1999.

[120] Jiang Yang, Zhaoming Yuan, and Zhi Zhou. Arbitrarily high-order maximum bound preserving schemes with cut-off postprocessing for allen–cahn equations. *Journal of Scientific Computing*, 90(2):76, 2022.

[121] Steven T Zalesak. Fully multidimensional flux-corrected transport algorithms for fluids. *Journal of Computational Physics*, 31(3):335–362, 1979.

[122] Yuping Zeng, Liuqiang Zhong, Mingchao Cai, Feng Wang, and Shangyou Zhang. Conforming and nonconforming virtual element methods for Signorini problems. *J. Sci. Comput.*, 100(1):Paper No. 18, 25, 2024.

Bibliography