This is a peer-reviewed, accepted author manuscript of the following research article: Amiri, A, Barrenechea, G & Pryer, T 2024, 'A nodally bound-preserving finite element method for reaction-convection-diffusion equations', Mathematical Models and Methods in Applied Sciences.

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

Abdolreza Amiri[1] Gabriel R. Barrenechea[1] Tristan Pryer[2]

[1] Department of Mathematics and Statistics, University of Strathclyde, 26 Richmond Street, Glasgow G1 1XH, Scotland
{abdolreza.amiri, gabriel.barrenechea}@strath.ac.uk
[2]Department of Mathematical Sciences, University of Bath, Claverton down, Bath BA2 7AY, UK
tmp38@bath.ac.uk

Abstract

This paper introduces a novel approach to approximate a broad range of reaction-convection-diffusion equations using conforming finite element methods while providing a discrete solution respecting the physical bounds given by the underlying differential equation. The main result of this work demonstrates that the numerical solution achieves accuracy of $O(h^k)$ in the energy norm, where k represents the underlying polynomial degree. To validate the approach, a series of numerical experiments is conducted for various problem instances. Comparisons with the linear continuous interior penalty stabilised method, and the algebraic flux-correction scheme (for the piecewise linear finite element case) have been carried out, where we can observe the favourable performance of the current approach.

1 Introduction

In numerous fields of mathematical modelling, the partial differential equations (PDEs) used to model the phenomenon in question respect the basic laws that were used to derive them, and their solutions satisfy these *first principles*. For example, in the modelling of incompressible flows, the PDEs involved provide divergence-free velocities, the PDEs modelling dissipative systems provide solutions that are energy stable, and in phase-field modelling the solution has strict global maxima and minima, just to give a few examples. From a numerical approximation perspective, it is very desirable that the finite element methods used to discretise these PDEs do respect the respective laws. Unfortunately, in most instances this is not the case. For example, the most popular finite element methods for incompressible flows are not pointwise divergence-free (see, e.g., [20] for a recent review on the topic), and proving that schemes are energy stable is far from trivial (see, e.g., [7, 15]).

In the particular case of bound-preservation, it was understood very early that the solution of a standard finite element method does not, in general, respect the physical bounds. This was first formalised in [9] in the finite element context where it was shown that the approximation using piecewise linear finite elements respects such bounds only if the mesh satisfies certain assumptions about its internal angles, and how refined it is. In addition to the mesh restriction, for conservation laws finite element methods that respect the physical bounds of the problem can either be first order accurate, or nonlinear, due to the notorious Godunov order barrier theorem (see, e.g., [16]).

As a consequence of the above discussion, in the last few decades numerous methods that respect global bounds have been proposed. A special attention has been given, in fact, to a stronger property. Namely, numerous methods respecting the *Discrete Maximum Principle (DMP)* have been proposed, especially for

convection-dominated problems (see [3, 5, 24, 27, 30], just to name a few, and [4] for a recent review). These methods are, for the most part, nonlinearly stabilised methods. That is, methods that add a nonlinear stabilising term to the Galerkin scheme in a way that diffusion is added locally, thus making the problem locally diffusion-dominated, and avoiding the spurious oscillations and local violations of the maximum principle. Interestingly, despite the above-mentioned discretisations being nonlinear, in most cases the finite element methods proposed are based on piecewise linear elements. In fact, the extension to higher order elements imposes even stronger mesh conditions (for example, in [18] it is proven that for the Poisson equation in two space dimensions a monotone discretisation using quadratic elements can be built if the mesh is either equilateral, or consists of squares in which the squares are divided by arbitrary diagonals), and, on the other hand, there is not much analysis available for nonlinear schemes using higher order polynomials.

In many situations, the stability of the numerical method does not require the discrete solution to be devoid of local spurious oscillations, but only needs to satisfy the global bounds. In such a case, the problem is actually simpler and there are several methodologies available. The first option one might think of is simply cutting off the values that lie outside the admissible range. This approach is perhaps used in many simulations without being explicitly mentioned, and it has been analysed for a linear reaction-diffusion equation in [23], and for parabolic problems in [25]. A more radical approach consists on *reformulating the problem* in such a way that the discrete solution always respects the corresponding bounds (see, e.g. [19] for the application to Chemotaxis, or [14] for an extremely widely used reformulation in non-Newtonian fluid mechanics). Alternatively, global bounds can be enforced by introducing them as inequality constraints and then approximate a control problem as it has been done, e.g., in [12]. Similarly, inequality constraints can be dealt with using Lagrange multipliers and solving an extended system, as has been recently done in [8], or [29],where a semi-smooth Newton method has also been introduced to deal with the non-smoothness.

In the very recent work [2] a different strategy was followed to impose global bounds in the solution. The first step is to define the set, denoted $V_{\mathcal{P}}^{+}$, of *admissible* finite element functions as those satisfying the global bounds *at their degrees of freedom* (nodal values in the case of Lagrangian elements); then, introduce an algebraic projection onto the admissible set, denote by u_h^{+} the projection of u_h onto $V_{\mathcal{P}}^{+}$ and write a finite element problem for the projected object. Since this process introduces a kernel, as the projection is not injective, a stabilising term is added to remove the singularity. This last step allowed, in particular, to avoid the introduction of Lagrange multipliers. Moreover, when applied to the linear reaction-diffusion equation, it turns out that u_h^{+} is the orthogonal projection onto $V_{\mathcal{P}}^{+}$, and thus is independent of the stabilisation used (this last property is lost when applied to more complicated problems, such as problems with nonlinear reaction, also addressed in [2]). So, the purpose of the present work is to extend the methodology presented in [2] to the convection-diffusion equation. Although the driving principles are similar, this work presents significant novelities with respect to [2], namely:

1. the starting point in the construction of the method is not the plain Galerkin scheme, but a stabilised finite element method instead. The reason for this is twofold: first, in regions where the constraint is not active, local oscillations might still appear (and linear stabilisation helps with this), and second, our numerical experiments show that the addition of linear stabilisation in u_h^{+} helps tremendously the good behaviour of the nonlinear solver;

2. the definition of the stabilisation form is different, since we need to control the convective term;

3. the analysis differs greatly from that of [2]. On the one hand, the well-posedness analysis is very different, as the discretisation is not driven by a monotone operator. On the other hand, since the problem is non-symmetric the solution u_h^{+} is no longer the orthogonal projection of u onto $V_{\mathcal{P}}^{+}$, and thus the error analysis follows an alternative path.

The remainder of the paper is organised as follows: In Section 2 we introduce the notation, the model problem, and all the preliminary material for the setup of the method (including the choice of linearly sta-

bilised method to be used). In Section 3 we present the finite element method and show its well-posedness. The error analysis is carried out in Section 4, and in Section 5 we test the performance of the method via numerical experiments, comparing also with previously existing alterantives. Finally, some conclusions and future directions are drawn in Section 6.

2 General Setting And The Model Problem

We will adopt standard notations for Sobolev spaces, in line with, e.g., [10]. For $D \subseteq \mathbb{R}^d$, we denote by $\|\cdot\|_{0,p,D}$ the $L^p(D)$-norm; when $p = 2$ the subscript p will be omitted and we only write $\|\cdot\|_{0,D}$. In addition, for $s \geq 0$, $p \in [1, \infty]$, we denote by $\|\cdot\|_{s,p,D}$ ($|\cdot|_{s,p,D}$) the norm (seminorm) in $W^{s,p}(D)$; when $p = 2$, we will again omit the subscript p and only write $\|\cdot\|_{s,D}$ ($|\cdot|_{s,D}$). In addition, we denote by $H^{-1}(D)$ the dual of $H^1_0(D)$ while identifying $L^2(D)$ with its dual. Thus, writing $\langle\cdot,\cdot\rangle_D$ for the duality pairing, we have

$$\langle f, v\rangle_D = \int_D f(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \qquad \forall\, v \in H^1_0(D)\,,$$

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.

2.1 The Model Problem

Let Ω be an open bounded Lipschitz domain in $\mathbb{R}^d$ ($d = 2, 3$) with polyhedral boundary $\partial\Omega$. For a given $f \in H^{-1}(\Omega)$, we consider the following convection-diffusion problem:

$$\begin{aligned} -\mathrm{div}(\mathcal{D}\nabla u) + \boldsymbol{\beta}\cdot\nabla u + \mu u &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \partial\Omega\,, \end{aligned} \tag{1}$$

where $\mathcal{D} = (d_{ij})_{i,j=1}^d \in L^\infty(\Omega)^{d\times d}$, $\boldsymbol{\beta} = (\beta_i)_{i=1}^d \in L^\infty(\Omega)^d$, and $\mu \in \mathbb{R}^+$, respectively, are the diffusion tensor, the convective field, and the reaction coefficient. We will assume that $\mathrm{div}\boldsymbol{\beta} = 0$, and the diffusion tensor $\mathcal{D}$ is symmetric and uniformly strictly positive definite in Ω; in other words, there exists a positive constant $d_0 > 0$ such that for almost all $\boldsymbol{x} \in \Omega$, we have

$$\sum_{i,j=1}^d y_i d_{ij}(\boldsymbol{x}) y_j \geq d_0 \sum_{i=1}^d y_i^2 \qquad \forall (y_1, \ldots, y_d) \in \mathbb{R}^d. \tag{2}$$

Remark 2.1. We have assumed $\mathrm{div}\boldsymbol{\beta} = 0$ only to avoid technical diversions. The results presented in this paper remind essentially unchanged under the milder condition $\mu - \mathrm{div}\boldsymbol{\beta}/2 > 0$. Moreover, all the results presented below also are applicable to the case when $\mu \in L^\infty(\Omega)$ is a function that is strictly positive in $\overline{\Omega}$.

The standard weak formulation of (1) reads as follows: find $u \in H^1_0(\Omega)$, such that

$$a(u, v) = \langle f, v\rangle_\Omega \qquad \forall v \in H^1_0(\Omega), \tag{3}$$

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

$$a(w, v) := (\mathcal{D}\nabla w, \nabla v)_\Omega + (\boldsymbol{\beta}\cdot\nabla w, v)_\Omega + (\mu w, v)_\Omega \qquad \forall v, w \in H^1_0(\Omega). \tag{4}$$

The bilinear form $a(\cdot.\cdot)$ induces the following "energy" norm in $H^1_0(\Omega)$

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

and thus (3) is well-posed thanks to the Lax-Milgram Lemma (see, e.g. [11, Lemma 25.2]).

As it was mentioned in the introduction, our aim is to look for discrete solutions that respect the same bounds as the solution of (3). Thus, we make the following assumption.
Assumption (A1): We will suppose that the weak solution of (3) satisfies

$$0 \leq u(\boldsymbol{x}) \leq \kappa, \qquad \text{for almost all } \boldsymbol{x} \in \Omega, \tag{5}$$

where κ is a known positive constant.

Remark 2.2. Assumption (A1) is, in fact, a re-statement of one of the consequences of the maximum principle for elliptic partial differential equations, see e.g. [13]. The lower bound in (5) is not required to be equal to zero, but we set it as zero for clarity of exposition. In addition, the same results proven in this work hold if κ is replaced by a non-negative continuous function $\kappa(\boldsymbol{x})$. In general, sharp bounds for the constant κ are not available, but in some cases they can be obtained. For example, as a consequence of maximum and comparison principles (see e.g, [28, Corollary 4.4]) the following bounds can be proven: for almost all $\boldsymbol{x} \in \Omega$ the solution u of (3) satisfies

$$-\frac{\|f\|_{0,\infty,\Omega}}{\mu} \leq u(\boldsymbol{x}) \leq \frac{\|f\|_{0,\infty,\Omega}}{\mu}\,. \tag{6}$$

Moreover, if $f \geq 0$ in Ω we can sharpen the above bound to:

$$0 \leq u(\boldsymbol{x}) \leq \frac{\|f\|_{0,\infty,\Omega}}{\mu}\,, \tag{7}$$

for almost all $\boldsymbol{x} \in \Omega$. Hence, a reasonable estimate for κ is $\frac{\|f\|_{0,\infty,\Omega}}{\mu}$ (and this is, in fact, the estimate we have used in our numerical experiments).

2.2 Triangulations, finite element spaces, and preliminary results

Let $\mathcal{P}$ be a conforming, shape-regular, quasi-uniform partition of Ω into closed simplices (or affine quadrilateral/hexahedra). Over $\mathcal{P}$, and for $k \geq 1$, we define the finite element spaces

$$\tilde{V}_{\mathcal{P}} := \{v_h \in C^0(\overline{\Omega}) : v_h|_K \in \mathfrak{R}(K) \quad \forall K \in \mathcal{P}\}\,, \tag{8}$$

$$V_{\mathcal{P}} := \tilde{V}_{\mathcal{P}} \cap H_0^1(\Omega)\,, \tag{9}$$

where

$$\mathfrak{R}(K) = \begin{cases} \mathbb{P}_k(K), & \text{if } K \text{ is a simplex,} \\ \mathbb{Q}_k(K), & \text{if } K \text{ is an affine quadrilateral/hexahedral,} \end{cases} \tag{10}$$

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

Remark 2.3. The results presented below can, in principle, be extended to more general quadrilateral meshes. Nevertheless, that would require technical diversions due to the need to prove norm-equivalences (that are classical for mapped elements). To avoid these diversions and keep the presentation focused on the bound-preservation aspects, we restrict the presentation to affine simplices and quadrilateral/hexahedral meshes.

For a mesh $\mathcal{P}$, the following notations are used:

- let $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N\}$ denote the set of internal nodes, and the usual Lagrangian basis functions associated to these nodes, spanning the space $V_{\mathcal{P}}$, are denoted by $\phi_1, \dots, \phi_N$;
- let $\mathcal{F}_I$ denote the set of internal facets, $\mathcal{F}_\partial$ denotes the set of boundary facets, and $\mathcal{F}_h = \mathcal{F}_I \cup \mathcal{F}_\partial$ denotes the set of all facets of $\mathcal{P}$; for an element $K \in \mathcal{P}$ the set of its facets is denoted by $\mathcal{F}_K$;
- for $K \in \mathcal{P}$, $F \in \mathcal{F}_h$, and a node $\boldsymbol{x}_i$, we define the following neighbourhoods:

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

- for a facet $F \in \mathcal{F}_I$, $[\![\cdot]\!]$ denotes the jump of a function across F.

The diameter of a set $G \subset \mathbb{R}^d$ is denoted by h_G, and $h = \max\{h_K : K \in \mathcal{P}\}$ stands for the mesh size. We also define the mesh function $\mathfrak{h}$ as a continuous, element-wise linear function defined as a local average of element diameters commonly used in finite element analysis [26]. For this, we introduce the set of *vertices* of the mesh, $\boldsymbol{v}_1, \dots, \boldsymbol{v}_M$, and define $\mathfrak{h}$ as the piecewise linear function prescribed by the nodal values

$$\mathfrak{h}(\boldsymbol{v}_i) = \frac{\sum_{K : \boldsymbol{v}_i \in K} h_K}{\#\{K : \boldsymbol{v}_i \in K\}}. \tag{11}$$

In the construction of the method, and its analysis, the following mass-lumped L^2-inner product will be of importance: for every $v_h, w_h \in V_{\mathcal{P}}$, we define

$$(v_h, w_h)_h = \sum_{i=1}^{N} \mathfrak{h}(\boldsymbol{x}_i)^d v_h(\boldsymbol{x}_i) w_h(\boldsymbol{x}_i), \tag{12}$$

which induces the norm $|v_h|_h = (v_h, v_h)_h^{\frac{1}{2}}$ in $V_{\mathcal{P}}$. This norm is, in fact, equivalent to the standard $L^2(\Omega)$-norm. More precisely, the following result, whose proof can be found in [11, Propositions 28.5, 28.6], will be used repeatedly in our analysis below: There exist $C, c > 0$, independent of h, such that

$$c \sum_{i : \boldsymbol{x}_i \in K} h_K^d v_h^2(\boldsymbol{x}_i) \leq \|v_h\|_{0,K}^2 \leq C \sum_{i : \boldsymbol{x}_i \in K} h_K^d v_h^2(\boldsymbol{x}_i) \qquad \forall\, K \in \mathcal{P}, \tag{13}$$

and thus, as a consequence of the shape-regularity of the mesh, the following holds:

$$c\, |v_h|_h^2 \leq \|v_h\|_{0,\Omega}^2 \leq C\, |v_h|_h^2, \tag{14}$$

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

Next, we recall that the Lagrange interpolation operator is defined by (see, e.g., [10, Chapter 11])

$$\begin{aligned} i_h &: C^0(\overline{\Omega}) \cap H_0^1(\Omega) \longrightarrow V_{\mathcal{P}}, \\ v &\longmapsto i_h v = \sum_{j=1}^{N} v(x_j) \phi_j. \end{aligned} \tag{15}$$

In addition, the $L^2(\Omega)$ -orthogonal projection operator (see [10, Chapter 22]) $\pi : L^2(\Omega) \longrightarrow V_{\mathcal{P}}$ is defined as follows

$$\begin{aligned} &\pi : L^2(\Omega) \longrightarrow V_{\mathcal{P}}, \\ &\qquad w \longmapsto \pi(w) \text{ where } \left(\pi(w), v_h\right)_\Omega = (w, v_h)_\Omega \quad \forall\, v_h \in V_{\mathcal{P}}\,. \end{aligned} \tag{16}$$

With the above ingredients, we now state some inequalities and properties that will be useful in what follows:

a) Inverse inequality: ([10, Lemma 12.1]) For all $m, \ell \in \mathbb{N}_0$, $0 \le m \le \ell$ and all $p, q \in [1, \infty]$, there exists a constant C, independent of h, such that

$$|v_h|_{\ell,p,K} \le C h_K^{m-\ell+d\left(\frac{1}{p}-\frac{1}{q}\right)} |v_h|_{m,q,K} \qquad \forall\, v_h \in V_{\mathcal{P}}\,. \tag{17}$$

b) Discrete Trace inequality: ([10, Lemma 2.15]) There exists $C > 0$ independent of h such that, for every $v \in H^1(K)$ the following holds

$$\|v\|^2_{0,\partial K} \leqslant C\left(h_K^{-1}\|v\|^2_{0,K} + h_K|v|^2_{1,K}\right). \tag{18}$$

c) Approximation property of the Lagrange interpolant: ([10, Proposition 1.12]) Let $1 \le \ell \le k$ and i_h be the Lagrange interpolant. Then, there exists $C > 0$, independent of h, such that for all h and $v \in H^{\ell+1}(\Omega) \cap H^1_0(\Omega)$ the following holds:

$$\|v - i_h v\|_{0,\Omega} + h|v - i_h v|_{1,\Omega} \le C h^{\ell+1}|v|_{\ell+1,\Omega}\,. \tag{19}$$

d) Approximation property of the L^2-orthogonal projection operator: ([10, Section 22.5]) Let $0 \le \ell \le k$ and π be the $L^2(\Omega)$-orthogonal projection. Then, there exists $C > 0$, independent of h, such that for all h and $v \in H^{\ell+1}(\Omega) \cap H^1_0(\Omega)$ the following holds:

$$\|v - \pi(v)\|_{0,\Omega} + h\,|v - \pi(v)|_{1,\Omega} \le C h^{\ell+1}|v|_{\ell+1,\Omega}\,. \tag{20}$$

2.3 The algebraic projection onto the admissible set

With Assumption (A1) in mind, we define the following subset of finite element functions that satisfy the bound (5) at the degrees of freedom:

$$V_{\mathcal{P}}^+ := \{v_h \in V_{\mathcal{P}} : v_h(\boldsymbol{x}_i) \in [0, \kappa] \text{ for all } i = 1, \dots, N\}. \tag{21}$$

Every element $v_h \in V_{\mathcal{P}}$ can be split as the sum $v_h = v_h^+ + v_h^-$, where v_h^+ and v_h^- are given by

$$v_h^+ = \sum_{i=1}^{N} \max\left\{0, \min\{v_h(\boldsymbol{x}_i), \kappa\}\right\}\phi_i\,, \tag{22}$$

and

$$v_h^- = v_h - v_h^+. \tag{23}$$

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

$$(\cdot)^+ : V_{\mathcal{P}} \to V_{\mathcal{P}}^+ \quad , \quad v_h \mapsto v_h^+\,. \tag{24}$$

Remark 2.4. If κ is not a constant value, but a non-negative continuous function, the only difference in the definition of the projection is that in such a case the constrained part is given by

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

To avoid technical diversions we will not detail such a case, but the results proven in this paper remain unchanged.

The following results concerning this projection will be used repeatedly.

Lemma 2.5. *Let the operator* $(\cdot)^+$ *be defined in* (24)*. There exists a constant* $C > 0$*, independent of* h*, such that*

$$\|w_h^+ - v_h^+\|_{0,\Omega} \leq C\|w_h - v_h\|_{0,\Omega}, \tag{25}$$

$$\|v_h^+\|_{0,\Omega} \leq C\kappa, \tag{26}$$

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

Proof. During this proof we drop the subindex h to lighten the notation. Let $w, v \in V_{\mathcal{P}}$. We start noticing 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 $v^+(\boldsymbol{x}_i) - w^+(\boldsymbol{x}_i) \leq -(v(\boldsymbol{x}_i) - w(\boldsymbol{x}_i))$. So, $|v^+(\boldsymbol{x}_i) - w^+(\boldsymbol{x}_i)| \leq |v(\boldsymbol{x}_i) - w(\boldsymbol{x}_i)|$. Then, using (14) we obtain

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

concluding the proof. □

2.4 A linear stabilised method

As it was mentioned in the introduction, in the method's definition we need to introduce a linear stabilising term aimed at dampening the oscillations caused by the dominating convection. In this work we have chosen to use the Continuous Interior Penalty (CIP) method originally proposed in [6]. This method adds the following stabilising term to the Galerkin scheme:

$$J(u_h, v_h) = \gamma \sum_{F\in\mathcal{F}_I} \int_F \|\boldsymbol{\beta}\|_{0,\infty,F} h_F^2\, [\![\nabla u_h]\!] \cdot [\![\nabla v_h]\!]\,\mathrm{d}s\,. \tag{27}$$

Here, $\gamma \geq 0$ is a non-dimensional constant. Using this stabilising term, the CIP stabilised method proposed in [6] reads as follows: find $u_h \in V_{\mathcal{P}}$ such that

$$a_J(u_h, v_h) := a(u_h, v_h) + J(u_h, v_h) = \langle f, v_h\rangle_\Omega \qquad \forall v_h \in V_{\mathcal{P}}\,. \tag{28}$$

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

$$\|v_h\|_h := a_J(v_h, v_h)^{\frac{1}{2}} = \left(\|\mathcal{D}^{\frac{1}{2}}\nabla v_h\|_{0,\Omega}^2 + \|\mu^{\frac{1}{2}} v_h\|_{0,\Omega}^2 + J(v_h, v_h)\right)^{\frac{1}{2}}. \tag{29}$$

The following result will be of use in the error analysis.

Lemma 2.6. *There exists $C > 0$ such that for $v_h \in V_{\mathcal{P}}$, the penalty term (27) satisfies the following property*

$$\frac{\gamma}{\|\boldsymbol{\beta}\|_{0,\infty,\Omega}} \left\| h^{\frac{1}{2}} \left(\boldsymbol{\beta} \cdot \nabla v_h - \pi(\boldsymbol{\beta} \cdot \nabla v_h) \right) \right\|_{0,\Omega}^2 \leq C\, J(v_h, v_h). \tag{30}$$

Moreover, the stabilising term (27) *satisfies the following bounds: there exists $C > 0$, independent of h and any physical constant, such that for all $w_h, v_h \in V_{\mathcal{P}}$ the following holds*

$$J(v_h, w_h) \leq C\gamma\, h\, \|\boldsymbol{\beta}\|_{0,\infty,\Omega} |v_h|_{1,\Omega} |w_h|_{1,\Omega}, \tag{31}$$

$$J(v_h, w_h) \leq C\gamma \left(\sum_{K \in \mathcal{P}} h_K^{-1} \|\boldsymbol{\beta}\|_{0,\infty,K} \|v_h\|_{0,K}^2 \right)^{\frac{1}{2}} \left(\sum_{K \in \mathcal{P}} h_K^{-1} \|\boldsymbol{\beta}\|_{0,\infty,K} \|w_h\|_{0,K}^2 \right)^{\frac{1}{2}}. \tag{32}$$

Proof. The inequality (30) is a direct consequence of the result proven in [6, Lemma 5]. To prove (31) we use the Cauchy-Schwarz inequality, the local trace result (18) and the inverse inequality (17) to obtain

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

which proves (31). The proof of (32) follows from the last inequality above and one further the application of the inverse inequality (17). □

Remark 2.7. An alternative definition of the CIP stabilising term, that we will also use in some of our numerical results, is given by penalising the upwind gradient jumps rather than the normal gradient, that is,

$$J(u_h, v_h) = \gamma \sum_{F \in \mathcal{F}_I} \int_F \frac{\gamma_\beta}{\|\boldsymbol{\beta}\|_{0,\infty,F}} h_F^2 [\![\boldsymbol{\beta} \cdot \nabla u_h]\!] [\![\boldsymbol{\beta} \cdot \nabla v_h]\!] \mathrm{d}s\,, \tag{33}$$

where $\gamma_\beta \geq 0$ is a non-dimensional constant. In all our proofs we have used the term given by (27), but the proofs remain valid if we use (33) instead.

Remark 2.8. The choice of CIP stabilisation has been made for convenience and simplicity of the presentation. In the numerical experiments we will show that the addition of the linear stabilising term has a positive effect on the performance of the method, more specifically, it will improve the performance of the nonlinear solver greatly. From a stability/error estimates point of view, it is also worth mentioning that the exact same results proven in this work are also valid for other choices of linear stabilisation, e.g., local projection stabilisation [22], or subgrid viscosity [17], for example.

3 The finite element method

The finite element method proposed in this work reads as follows: find $u_h \in V_{\mathcal{P}}$ such that

$$a_h(u_h; v_h) = \langle f, v_h \rangle_\Omega \qquad \forall v_h \in V_{\mathcal{P}}\,, \tag{34}$$

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

$$a_h(u_h; v_h) := a_J(u_h^+, v_h) + s(u_h^-, v_h)\,. \tag{35}$$

Here, $a_J(\cdot,\cdot)$ is the bilinear form defined in (28), u_h^+ and u_h^- are defined in (22),(23). The bilinear form $s(\cdot,\cdot)$ is added in order to control the complementary part u_h^-, and is defined as follows:

$$s(v_h, w_h) = \alpha \sum_{i=1}^{N} \left(\|\mathcal{D}\|_{0,\infty,\omega_i} \mathfrak{h}(\boldsymbol{x}_i)^{d-2} + \|\boldsymbol{\beta}\|_{0,\infty,\omega_i} \mathfrak{h}(\boldsymbol{x}_i)^{d-1} + \mu \mathfrak{h}(\boldsymbol{x}_i)^d \right) v_h(\boldsymbol{x}_i) w_h(\boldsymbol{x}_i)\,, \tag{36}$$

where the parameter $\alpha > 0$ is a non-dimensional constant. The stabilising form $s(\cdot,\cdot)$ induces the following norm in $V_{\mathcal{P}}$:

$$\|v_h\|_s = \sqrt{s(v_h, v_h)}\,. \tag{37}$$

The following result, that appears as a consequence of (13), shows that the stabilising bilinear form $s(\cdot,\cdot)$ indeed controls u_h^-, more specifically it controls the kernel of the projection $(\cdot)^+$.

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

$$\|v_h\|_h^2 \le \frac{C_{\text{equiv}}}{\alpha} \|v_h\|_s^2 \qquad \forall v_h \in V_{\mathcal{P}}\,, \tag{38}$$

where $\|\cdot\|_h$ is the norm defined in (29)*.*

Proof. Using the inverse inequality (17), (13), and the mesh regularity we obtain

$$\|\mathcal{D}^{\frac{1}{2}} \nabla v_h\|_{0,\Omega}^2 + \|\mu^{\frac{1}{2}} v_h\|_{0,\Omega}^2 \le C \sum_{i=1}^{N} \left(\|\mathcal{D}\|_{0,\infty,\omega_i} \mathfrak{h}(\boldsymbol{x}_i)^{d-2} + \mu\, \mathfrak{h}(\boldsymbol{x}_i)^d \right) v_h(\boldsymbol{x}_i)^2. \tag{39}$$

Also, (32) and (13) yield

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

Gathering the last two bounds proves (38) with $C_{\text{equiv}} = (1+\gamma)C$. □

Remark 3.2. The result of Lemma 3.1 explains the scaling factors chosen for defining $s(\cdot,\cdot)$. Additionally, our formula (36) differs from the one used in reaction diffusion equations (as seen in [2]) by including a specific term, $\|\boldsymbol{\beta}\|_{0,\infty,\omega_i} \mathfrak{h}(\boldsymbol{x}_i)^{d-1}$. This inclusion is crucial for proving (38), which is important for both guaranteeing the problem is well-posed and for the error analysis. From our practical experience, this term also enhances the performance of the nonlinear solver. As for the structure of the stabilisation, we opted for a mass-lumped approach in defining $s(\cdot,\cdot)$, largely because of the monotonicity established in Lemma 3.3 below.

3.1 Well-posedness

In this section, we analyse the existence and uniqueness of solutions for (35). The first step is given by the following monotonicity result, whose proof is identical to that of [2, Lemma 3.1].

Lemma 3.3. *The bilinear form $s(\cdot,\cdot)$ defined in* (36) *satisfies the following inequalities:*

$$s(v_h^- - w_h^-, v_h^+ - w_h^+) \geq 0 \qquad \forall v_h, w_h \in V_{\mathcal{P}}, \tag{40}$$

$$s(v_h^-, w_h - v_h^+) \leq 0 \qquad \forall v_h \in V_{\mathcal{P}}, w_h \in V_{\mathcal{P}}^+. \tag{41}$$

Despite the fact that $s(\cdot,\cdot)$ is monotone, the discrete problem (34) is not driven by a monotone nonlinear mapping. So, the well-posedness of (35) needs to be proven using different arguments to those used in [2]. We will first prove existence that will appear as a consequence of Brouwer's fixed point theorem, and only after linking any solution u_h^+ of (34) to a variational inequality, we will be able to prove a uniqueness result for u_h.

Theorem 3.4. *Suppose that $\alpha \geq C_{\text{equiv}}$. Then, there exists $u_h \in V_{\mathcal{P}}$ that solves* (34).

Proof. We begin by defining the bilinear form

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

and the mapping

$$\begin{aligned} T : V_{\mathcal{P}} &\longrightarrow V_{\mathcal{P}}, \\ \widehat{u}_h &\longrightarrow u_h = T(\widehat{u}_h), \end{aligned}$$

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

$$\tilde{a}_J(u_h^+, v_h) + s(u_h^-, v_h) = \langle f, v_h\rangle_\Omega - (\boldsymbol{\beta}\cdot\nabla\widehat{u}_h^+, v_h)_\Omega \qquad \forall v_h \in V_{\mathcal{P}}. \tag{42}$$

We observe that u_h solves (34) if and only if $T(u_h) = u_h$. So, the proof will consist on proving that T satisfies the hypotheses of Brouwer's fixed point Theorem [28, Theorem 10.41].
i) T is well-defined: To prove that T is well-defined, we see that (42) is a particular example of the method proposed in [2]. So, using [2, Theorem 3.2], there exists a unique solution $u_h \in V_{\mathcal{P}}$ of (42), and thus T is well-defined.
ii) T is continuous: Since we have supposed that α is large enough, we can use the monotonicity result proven in [2, Theorem 3.2] and obtain 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, suppose that for $\widehat{v}_h, \widehat{w}_h \in V_{\mathcal{P}}$ and let $v_h = T(\widehat{v}_h)$ and $w_h = T(\widehat{w}_h)$. Integrating by parts, using Hölder's inequality, Lemma 2.5, and (38) we obtain

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

Therefore

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

and T is Lipschitz continuous.

iii) There exists $R > 0$, such that $T(B(0,R)) \subseteq B(0,R)$: Let $\widehat{z}_h \in V_{\mathcal{P}}$ be arbitrary and $z_h = T(\widehat{z}_h)$. By using $v_h = z_h^+$ in (42), Cauchy-Schwarz's and Hölder's inequalities, and (26) we get

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

and so z_h^+ satisfies

$$\|z_h^+\|_h \leq C \left\{ \frac{\|f\|_{0,\Omega}}{\mu^{\frac{1}{2}}} + \frac{\|\boldsymbol{\beta}\|_{0,\infty,\Omega} \, \kappa}{d_0^{\frac{1}{2}}} \right\}. \tag{43}$$

Next, we take $v_h = z_h^-$ in (42). Integrating by parts and using Hölder's inequality we get

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

Now using (26) we have

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

and applying (38) and Young's inequality we get

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

Using Young's and Cauchy-Schwarz's inequalities for $\tilde{a}_J(z_h^+, z_h^-)$, and (43), yields

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

for any $\delta > 0$. Then, choosing δ small enough, and using Lemma 3.1 we get

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

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

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

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

Hence, using Brouwer's fixed point theorem, there exists $u_h \in V_{\mathcal{P}}$ such that $T(u_h) = u_h$. In other words, problem (34) has at least one solution. □

The proof of the last result does not imply uniqueness of solutions. The next two results will close that gap, whilst at the same time providing a very useful characterisation for u_h^+.

Lemma 3.5. *Let $u_h \in V_{\mathcal{P}}$ solve* (34)*. Then, $u_h^+ \in V_{\mathcal{P}}^+$ satisfies*

$$a_J(u_h^+, v_h - u_h^+) \geq \langle f, v_h - u_h^+\rangle_\Omega \qquad \forall v_h \in V_{\mathcal{P}}^+, \tag{44}$$

where $a_J(\cdot,\cdot)$ is defined in (28). In addition, u_h^- is the unique solution of

$$s(u_h^-, v_h) = \langle f, v_h\rangle - a_J(u_h^+, v_h) \qquad \forall v_h \in V_{\mathcal{P}}. \tag{45}$$

Proof. Testing (34) with $v_h \in V_{\mathcal{P}}^+$ and u_h^+ as test functions gives

$$\begin{aligned} a_J(u_h^+, v_h) + s(u_h^-, v_h) &= \langle f, v_h\rangle_\Omega, \\ a_J(u_h^+, u_h^+) + s(u_h^-, u_h^+) &= \langle f, u_h^+\rangle_\Omega. \end{aligned}$$

Subtracting the second equation from the first one we arrive at

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

and then, using (41) in Lemma 3.3 we get that $u_h^+ \in V_{\mathcal{P}}^+$ satisfies (44). Finally, since $s(\cdot,\cdot)$ is an elliptic bilinear form in $V_{\mathcal{P}}$, u_h^- is the unique solution of (45), thus concluding the proof. □

The last result provides a characterisation of any solution of (34) as the solution of the two successive problems (44), (45).

Corollary 3.6. *Problem* (34) *has a unique solution.*

Proof. If $u_1, u_2 \in V_{\mathcal{P}}$ satisfy (34), then u_1^+ and u_2^+ satisfy (44). But the solution of (44) is unique, thanks to Stampacchia's Theorem (see [21, Chapter II, Theorem 2.1]). So, $u_1^+ = u_2^+$. Therefore, (45) holds for u_1^- and u_2^-, but $u_1^+ = u_2^+$ shows that the right hand side of both problems are the same, and since $s(\cdot,\cdot)$ is an elliptic bilinear form, then $u_1^- = u_2^-$. Thus, $u_1 = u_1^+ + u_1^- = u_2^+ + u_2^- = u_2$. □

Remark 3.7. We finish this section by remarking that the complementary part u_h^- of u_h has a local support. In fact, we first remark 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 $u_h^+(\boldsymbol{x}_i) = 0$. So, the support of u_h^- is contained in the region where $u_h^+ = 0$ or $u_h^+ = \kappa$. In other words, u_h^- has a localised support, restricted to the regions where the constraint imposed in the definition of $V_{\mathcal{P}}^+$ is active.

4 Error analysis

This section is devoted to the error analysis of the method (34). Since our interest in this work is to provide a discrete solution that respects the bounds given by the continuous problem, error estimates will be proven for the constrained part u_h^+.

Theorem 4.1. *Let $u \in H^{k+1}(\Omega) \cap H_0^1(\Omega)$ be the solution of* (1) *and $u_h \in V_{\mathcal{P}}$ be the solution of* (34)*. Then, there exists $C > 0$ independent of $\mathcal{D}$, μ, $\boldsymbol{\beta}$, and h, such that*

$$\|u - u_h^+\|_h \leq Ch^k \left(\|\mathcal{D}\|_{0,\infty,\Omega}^{\frac{1}{2}} + \mu^{-\frac{1}{2}} \|\boldsymbol{\beta}\|_{0,\infty,\Omega} + h^{\frac{1}{2}} \|\boldsymbol{\beta}\|_{0,\infty,\Omega}^{\frac{1}{2}} + h\mu^{\frac{1}{2}} \right) |u|_{k+1,\Omega}. \tag{46}$$

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

$$u - u_h^+ = (u - \pi(u)) + (\pi(u) - u_h^+) =: \eta_h + e_h\,, \tag{47}$$

where π is the $L^2(\Omega)$-orthogonal projection defined in (16).

The bound for η_h is a direct consequence of (20) and (31). In fact, using the Cauchy-Schwarz and Young inequalities we get

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

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

$$\|e_h\|_h^2 = -a_J(\eta_h, e_h) + a_J(u - u_h^+, \pi(u) - u_h^+) =: \mathrm{I} + \mathrm{II}. \tag{48}$$

We start decomposing I as follows:

$$\mathrm{I} = (\mathcal{D}\nabla\eta_h, \nabla e_h)_\Omega + (\boldsymbol{\beta} \cdot \nabla\eta_h, e_h)_\Omega + \mu(\eta_h, e_h)_\Omega + J(\eta_h, e_h) =: \text{(a)} + \text{(b)} + \text{(c)} + \text{(d)}\,, \tag{49}$$

and bound each one of the above terms separately. Firstly, by Cauchy-Schwarz inequality, then using (20) we have

$$\text{(a)} \leq \|\mathcal{D}\|_{0,\infty,\Omega}^{\frac{1}{2}} |\eta_h|_{1,\Omega} \|\mathcal{D}^{\frac{1}{2}} \nabla e_h\|_{0,\Omega} \leq Ch^k \|\mathcal{D}\|_{0,\infty,\Omega}^{\frac{1}{2}} |u|_{k+1,\Omega} \|e_h\|_h\,. \tag{50}$$

To bound (b) we first integrate by parts, use the orthogonality of π, and Lemma 2.6.

$$\begin{aligned}
\text{(b)} &= (\pi(u) - u, \boldsymbol{\beta} \cdot \nabla e_h)_\Omega \\
&= \big(\pi(u) - u, \boldsymbol{\beta} \cdot \nabla e_h - \pi(\boldsymbol{\beta} \cdot \nabla e_h)\big)_\Omega \\
&\leq \|\eta_h\|_{0,\Omega} \|\boldsymbol{\beta} \cdot \nabla e_h - \pi(\boldsymbol{\beta} \cdot \nabla e_h)\|_{0,\Omega} \\
&\leq Ch^{k+\frac{1}{2}} \|\boldsymbol{\beta}\|_{0,\infty,\Omega}^{\frac{1}{2}} |u|_{k+1,\Omega} \|e_h\|_h\,.
\end{aligned} \tag{51}$$

The term (c) is handled analagously to (a)

$$\text{(c)} \leq \mu^{\frac{1}{2}} \|\eta_h\|_{0,\Omega} \|\mu^{\frac{1}{2}} e_h\|_{0,\Omega} \leq C h^{k+1} \mu^{\frac{1}{2}} |u|_{k+1,\Omega} \|e_h\|_h \,. \tag{52}$$

Finally, for (d), since $J(\cdot,\cdot)$ is semi-positive definite and symmetric we apply Cauchy-Schwarz's inequality followed by Lemma 2.6 and (20)

$$\begin{aligned} \text{(d)} &\leq J(\eta_h, \eta_h)^{\frac{1}{2}} J(e_h, e_h)^{\frac{1}{2}} \\ &\leq C h^{k+\frac{1}{2}} \|\boldsymbol{\beta}\|_{0,\infty,\Omega}^{\frac{1}{2}} |u|_{k+1,\Omega} \|e_h\|_h \,. \end{aligned} \tag{53}$$

Substituting (50), (51), (52), and (53) into (49) we obtain the following bound for I:

$$\text{I} \leq C h^k \left(\|\mathcal{D}\|_{0,\infty,\Omega}^{\frac{1}{2}} + h^{\frac{1}{2}} \|\boldsymbol{\beta}\|_{0,\infty,\Omega}^{\frac{1}{2}} + h\mu^{\frac{1}{2}} \right) |u|_{k+1,\Omega} \|e_h\|_h \,. \tag{54}$$

To bound II we first recall that $(\pi(u))^- = \pi(u) - (\pi(u))^+$ and then

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

Now, thanks to the regularity of u, we note that $J(u, (\pi(u))^+ - u_h^+) = 0$. Hence, using that u_h^+ solves (44), we have

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

Therefore,

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

The term in the right-hand side of the above inequality is, in essence, a consistency error, and will require special treatment. Let i_h be the Lagrange interpolant defined in (15). Since $u(x) \in [0, \kappa]$ a.e. in Ω, then $i_h(u) \in V_{\mathcal{P}}^+$. So, $i_h(u) = (i_h(u))^+ \in V_{\mathcal{P}}^+$, which implies $(i_h(u))^- = 0$. Therefore,

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

So, using the definition of $a_J(\cdot,\cdot)$ we bound II as follows

$$\begin{aligned} \text{II} &\leq a_J(u - u_h^+, (\pi(u))^- - (i_h(u))^-) \\ &= (\mathcal{D}\nabla(u - u_h^+), \nabla((\pi(u))^- - (i_h(u))^-)_\Omega + (\boldsymbol{\beta} \cdot \nabla(u - u_h^+), (\pi(u))^- - (i_h(u))^-)_\Omega \\ &\quad + \mu(u - u_h^+, (\pi(u))^- - (i_h(u))^-)_\Omega + J(u - u_h^+, (\pi(u))^- - (i_h(u))^-) \\ &= \text{(e)} + \text{(f)} + \text{(g)} + \text{(h)}. \end{aligned} \tag{55}$$

We begin with (e), which can be bounded using the Cauchy-Schwarz inequality

$$\begin{aligned} \text{(e)} &\leq \|\mathcal{D}\|_{0,\infty,\Omega}^{\frac{1}{2}} \|\mathcal{D}^{\frac{1}{2}} \nabla(u - u_h^+)\|_{0,\Omega} |(\pi(u))^- - (i_h(u))^-|_{1,\Omega} \\ &\leq C h^{-1} \|\mathcal{D}\|_{0,\infty,\Omega}^{\frac{1}{2}} \|\mathcal{D}^{\frac{1}{2}} \nabla(u - u_h^+)\|_{0,\Omega} \|(\pi(u))^- - (i_h(u))^-\|_{0,\Omega}, \end{aligned} \tag{56}$$

where we have used an inverse inequality. Then, as $(\cdot)^-$ is Lipschitz continuous by Lemma 2.5

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

by the approximation properties for i_h and π given in (19) and (20). For (f) we begin by integrating by parts

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

using the Cauchy-Schwarz. inequality Then again, by an inverse estimate

$$\begin{aligned}(\mathrm{f}) &\leq Ch^{-1}\|\boldsymbol{\beta}\|_{0,\infty,\Omega}\|u-u_h^+\|_{0,\Omega}\|(\pi(u))^- - (i_h(u))^-\|_{0,\Omega}\\ &\leq Ch^k\|\boldsymbol{\beta}\|_{0,\infty,\Omega}\,\mu^{-\frac{1}{2}}|u|_{k+1,\Omega}\|u-u_h^+\|_h,\end{aligned}\tag{57}$$

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

$$\begin{aligned}(\mathrm{g}) &\leq \mu^{\frac{1}{2}}\|\mu^{\frac{1}{2}}(u-u_h^+)\|_{0,\Omega}\|(\pi(u))^- - (i_h(u))^-\|_{0,\Omega}\\ &\leq Ch^{k+1}\mu^{\frac{1}{2}}|u|_{k+1,\Omega}\|u-u_h^+\|_h,\end{aligned}\tag{58}$$

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

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

using Lemma 2.6. Now by an inverse inequality

$$\begin{aligned}(\mathrm{h}) &\leq C\|\boldsymbol{\beta}\|_{0,\infty,\Omega}^{\frac{1}{2}}\,h^{-\frac{1}{2}}\|(\pi(u))^- - (i_h(u))^-\|_{0,\Omega}\,J(u-u_h^+,u-u_h^+)^{\frac{1}{2}}\\ &\leq Ch^{k+\frac{1}{2}}\|\boldsymbol{\beta}\|_{0,\infty,\Omega}^{\frac{1}{2}}|u|_{k+1,\Omega}\|u-u_h^+\|_h,\end{aligned}\tag{59}$$

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

Combining (56), (57), (58) and (59) we arrive to the following bound for II:

$$\mathrm{II} \leq Ch^k\left(\|\mathcal{D}\|_{0,\infty,\Omega}^{\frac{1}{2}} + \mu^{-\frac{1}{2}}\|\boldsymbol{\beta}\|_{0,\infty,\Omega} + h^{\frac{1}{2}}\|\boldsymbol{\beta}\|_{0,\infty,\Omega}^{\frac{1}{2}} + h\mu^{\frac{1}{2}}\right)\|u-u_h^+\|_h|u|_{k+1,\Omega}.\tag{60}$$

Hence, inserting (54) and (60) into (48), and using Young's inequality, we obtain the following bound for $\|e_h\|_h$:

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

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

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

and (46) follows rearranging terms. □

4.1 The extension to problems with non-homogeneous boundary conditions

Although the presentation of the method and its analysis have been done assuming that the Dirichlet boundary conditions are homogeneous, many problems of practical interest involve non-homogeneous boundary conditions. So, in this section we briefly describe the method for a non-homogeneous Dirichlet problem. Let us assume that, instead of (1) we are interested in solving

$$\begin{aligned} -\mathrm{div}(\boldsymbol{D}\nabla u) + \boldsymbol{\beta}\cdot\nabla u + \mu u &= f \qquad \text{in } \Omega, \\ u &= g \qquad \text{on } \partial\Omega\,, \end{aligned} \tag{61}$$

where $g \in H^{\frac{1}{2}}(\partial\Omega)$, $g \geq 0$ on $\partial\Omega$. Thanks to the definition of κ and the maximum and comparison principles for partial differential equations, $\kappa \geq \|g\|_{0,\infty,\partial\Omega}$. For simplicity, we will assume that g is the trace of a function belonging to $\tilde{V}_{\mathcal{P}}$.

Let us consider the set of nodes of the triangulation (including boundary nodes): $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_P$, and, as earlier in the manuscript, we denote the interior ones by $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, $N < P$. Let us define the following extension of g into Ω: $u_{h,g} \in \tilde{V}_{\mathcal{P}}$ defined as follows

$$u_{h,g}(\boldsymbol{x}_i) = \begin{cases} g(\boldsymbol{x}_i) & \text{if } i \in \{N+1, \ldots, P\}\,, \\ 0 & \text{else}\,. \end{cases} \tag{62}$$

With these ingredients the analogue of (34) for the non-homogeneous boundary data is: Find $\tilde{u}_h \in V_{\mathcal{P}}$ such that

$$a_J\big((\tilde{u}_h + u_{h,g})^+, v_h\big) + s\big((\tilde{u}_h + u_{h,g})^-, v_h\big) = \langle f, v_h\rangle_\Omega \qquad \forall\, v_h \in V_{\mathcal{P}}\,. \tag{63}$$

The fundamental reason for choosing $u_{h,g}$ as extension of g, rather than any other, is that at each node of $\mathcal{P}$ either $\tilde{u}_h$ or $u_{h,g}$ are zero, and then the following important property holds:

$$(\tilde{u}_h + u_{h,g})^+ = \tilde{u}_h^+ + u_{h,g}\,, \tag{64}$$

and, as a consequence $(\tilde{u}_h + u_{h,g})^- = \tilde{u}_h^-$. Thus, (63) can be rewritten as: Find $\tilde{u}_h \in V_{\mathcal{P}}$ such that

$$a_J\big(\tilde{u}_h^+, v_h\big) + s\big(\tilde{u}_h^-, v_h\big) = (f, v_h)_\Omega - a_J\big(u_{h,g}, v_h\big) \qquad \forall\, v_h \in V_{\mathcal{P}}\,. \tag{65}$$

Now the proof of Theorem 3.4 remains unchanged, and thus (65) has a solution. For the uniqueness result, the same arguments used in the proof of Lemma 3.5 lead to the fact that $\tilde{u}_h^+$ solves the following variational inequality: $\tilde{u}_h^+ \in V_{\mathcal{P}}^+$, and

$$a_J(\tilde{u}_h^+, v_h - \tilde{u}_h^+) \geq \langle f, v_h - \tilde{u}_h^+\rangle_\Omega - a_J\big(u_{h,g}, v_h - \tilde{u}_h^+\big) \qquad \forall\, v_h \in V_{\mathcal{P}}^+\,. \tag{66}$$

Since (66) has a unique solution thanks to Stampacchia's Theorem, the existence and uniqueness of solution for (63) follows using exactly the same arguments as those for (34). Finally, to analyse the error we notice that, assuming enough regularity for the exact solution we get

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

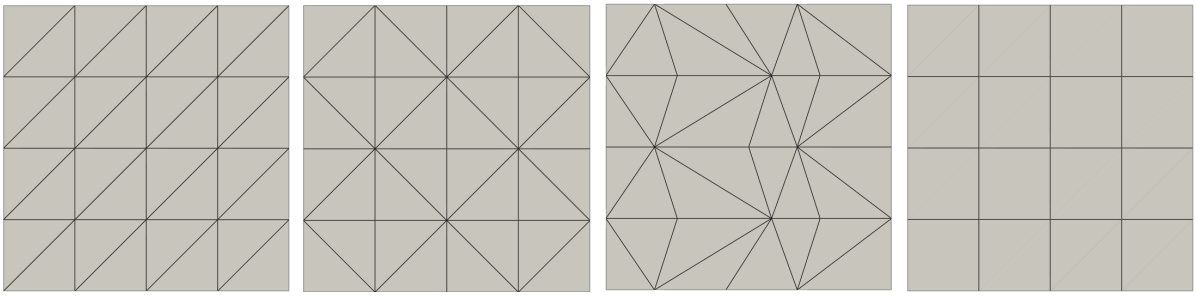
for all $v_h \in V_{\mathcal{P}}$. So, the following variational inequality holds

$$a_J((\tilde{u}_h + u_{h,g})^+ - u, v_h - u_h^+) \geq 0 \qquad \forall\, v_h \in V_{\mathcal{P}}^+\,, \tag{67}$$

which is instrumental in the bound for the bound of II in the proof of Theorem 4.1. Hence, the error analysis follows very similar arguments as those presented in the proof of Theorem 4.1.

5 Numerical experiments

In this section we present three series of numerical results testing the performance of the finite element method (34). In all numerical experiments in this section $\Omega = (0,1)^2$, and we have used the value $\alpha = 1$ in the stabilising bilinear form $s(\cdot,\cdot)$. We have selected three different types of meshes, the coarsest level of them are depicted in Figure 1. While the family depicted in Figure 1a and 1b are symmetric and Delaunay, the mesh depicted in Figure1c is a non-Delaunay one, and the one depicted in 1d is quadrilateral. The mesh represented by 1c is generated by displacing some interior nodes of the mesh in 1b to the right, thus creating obtuse angles. The reason for this choice is motivated by the fact that the discrete maximum principle fails to hold for most finite element methods in such meshes (see, e.g., [4]). In particular, the initial datum u_h^0 defined below will not, in general, belong to $V_{\mathcal{P}}^+$.



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

Figure 1: Three coarse level indicative meshes used in the experiments all with $N = 5$.

To solve the nonlinear system associated to (35) we use the following Richardson-like iterative method: Set $u_h^0 \in V_{\mathcal{P}}$ such that

$$a_J(u_h^0, v_h) = \langle f, v_h\rangle_\Omega \qquad \forall v_h \in V_{\mathcal{P}}\,. \tag{68}$$

Then, for $n = 0, 1, 2, \ldots$, find $u_h^{n+1} \in V_{\mathcal{P}}$ such that:

$$a_J(u_h^{n+1}, v_h) = a_J(u_h^n, v_h) + \omega\left(\langle f, v_h\rangle_\Omega - a_J((u_h^n)^+, v_h) - s((u_h^n)^-, v_h)\right) \qquad \forall v_h \in V_{\mathcal{P}}, \tag{69}$$

where $\omega \in (0,1]$ is a damping parameter. The iterations are terminated when

$$\|u_h^{n+1} - u_h^n\|_{0,\Omega} \le 10^{-8}. \tag{70}$$

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 in the triangular meshes, and $\mathbb{Q}_1$ and $\mathbb{Q}_2$ elements in the quadrilateral mesh.

Example 1 (Convergence of a problem with smooth solution)**.** We consider $\mu = 1$, $\mathcal{D} = \epsilon \begin{bmatrix} 100 & \cos(x) \\ \cos(x) & 1 \end{bmatrix}$ ($\epsilon = 10^{-5}$), $\boldsymbol{\beta} = (2,1)$, and set f such that the function $u(x,y) = 100\sin(\pi x)\sin(\pi y)$ is the analytical solution of (1). Notice that $u(x) \in [0,100]$, and thus we choose $\kappa = 100$. The CIP stabilisation parameter is $\gamma = 0.025$ in the penalty term (27), and we have used $\omega = 1$ in the iterative method (69).

In Tables 1-5 we report the convergence results in both the $\|\cdot\|_{0,\Omega}$ and $\|\cdot\|_h$ norms for $u - u_h^+$, and the $\|\cdot\|_s$-norm for u_h^-, as well as the number of iterations needed to reach convergence for the nonlinear system. The results show an optimal order of convergence for the constrained part u_h^+, thus confirming the results from Section 4. In addition, they show a higher order of convergence (to zero) for 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 1: Numerical results for Example 1 using $\mathbb{P}_1$ elements and Mesh 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: Numerical results for Example 1 using $\mathbb{Q}_1$ elements and Mesh 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 3: Numerical results for Example 1 using $\mathbb{P}_2$ elements and Mesh 1c.

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 4: Numerical results for Example 1 using $\mathbb{Q}_2$ elements and Mesh 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 5: Numerical results for Example 1 using $\mathbb{P}_3$ elements and Mesh 1c.

Example 2 (A problem with two inner layers)**.** For this test case and the following one the diffusion in (1) is given by $\mathcal{D} = \epsilon \mathcal{I}$, where $\epsilon > 0$. We now approximate the solution of (1) for $f = 0$, $\mu = 0$, $\epsilon = 10^{-5}$, and $\boldsymbol{\beta} = (-y, x)$. Homogeneous Neumann boundary conditions are imposed on exit (that is, at the lines $x = 0$ and $y = 1$), while the following (discontinuous) Dirichlet datum is imposed on entry ($x = 1$ and $y = 0$):

$$g(x,y) = \begin{cases} 0 & \text{if } x \leq \frac{1}{3} \text{ and } y = 0 \\ \frac{1}{2} & \text{if } x \in \left(\frac{1}{3}, \frac{2}{3}\right) \text{ and } y = 0, \\ 1 & \text{otherwise.} \end{cases} \tag{71}$$

The goal of this numerical experiment is twofold. First, we aim at testing the capabilities of the current bound preserving method (BPM) (34) to suppress over- and under-shoots in regions where the constraint is not needed. In fact, for this example $\kappa = 1$ in the whole domain, while there are two inner layers inside the domain where the solution varies rapidly from 0 to approximately 0.5, and a second one where it does from 0.5 to 1. Now, around those layers the BPM will control the undershoot at $u_h^+ = 0$, but there is no explicit control on the region where the solution is approximately 0.5. Our intent is to assess the capability of the current method to suppress the possible overshoot at that region, even if the nonlinear stabilisation is not active in it. The second goal of this numerical experiment is to provide numerical evidence that the addition of the (linear) CIP stabilisation on u_h^+ has a positive impact on the performance of the method.

We start addressing the latter of our objectives. In Tables 6 and 7 we report the number of iterations required by the nonlinear solver. In all our simulations we have set a maximum number of iterations to $3,000$, and whenever we have reached this value, the solver stops and we report "NC" which represents non-convergence. As can be seen in Tables 6 and 7, if the linear CIP stabilisation is not added to the formulation then the nonlinear iteration process is more prone to non-convergence, thus giving one more argument in favour of adding linear stabilisation to (34). For this particular example we have chosen to use the stabilising term (33) with stabilising parameter $\gamma_{\boldsymbol{\beta}} = 0.05$, as it is the one that has provided the best numerical results in terms of sharpness of the interior layers.

We have run the experiments using meshes from Figure 1a and Figure 1c for a variety of values for N. For these experiments, for $\mathbb{P}_1$, $\mathbb{Q}_1$ and $\mathbb{Q}_2$ we have used $\omega = 0.1$ in (69) for the BPM and $\omega = 0.05$ if CIP is removed, that is, choosing $\gamma_{\boldsymbol{\beta}} = 0$ (it is worth mentioning that making ω smaller makes the convergence of the iterative solver more likely).

	N	5	9	17	33	65	129
Mesh 1a	$\gamma_\beta = 0.05$	82	96	122	124	113	98
	$\gamma_\beta = 0$	228	1702	NC	NC	NC	NC
Mesh 1c	$\gamma_\beta = 0.05$	140	148	174	137	123	111
	$\gamma_\beta = 0$	126	NC	NC	NC	NC	NC

Table 6: Number of iterations for the fixed point linearisation (69) needed to reach convergence using $\mathbb{P}_1$ elements and the meshes given in Figures 1a and 1c.

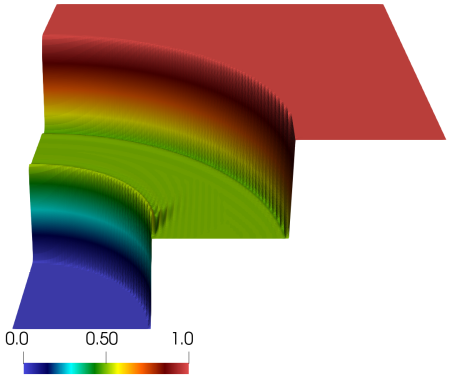
	N	5	9	17	33	65	129
$\mathbb{Q}_1$	$\gamma_\beta = 0.05$	72	128	136	151	159	190
	$\gamma_\beta = 0$	1901	NC	NC	NC	NC	NC
$\mathbb{Q}_2$	$\gamma_\beta = 0.05$	283	243	360	315	339	258
	$\gamma_\beta = 0$	2034	NC	NC	NC	NC	NC

Table 7: Number of iterations for the fixed point linearisation (69) needed to reach convergence using $\mathbb{Q}_1$ and $\mathbb{Q}_2$ elements and the mesh given in Figure 1d.

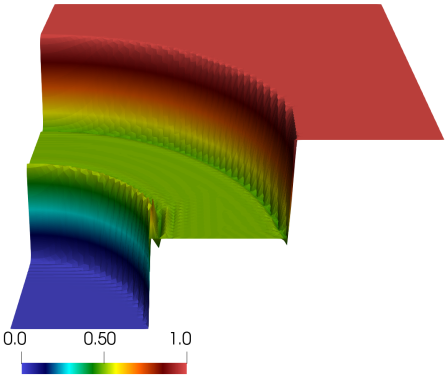
We next address the sharpness in the approximation of the interior layers. In Figure 2-4 we depict the approximate solution for the meshes from Figure 1a and 1c and observe the lack of significant oscillations in the vicinity of the layers, even for mesh 1c, which is non-Delaunay.

For comparison purposes, we have also approximated the same problem using the (linear) CIP method (with the stabilising term (33) and $\gamma_\beta = 0.05$), and the Algebraic Flux Correction (AFC) scheme, as written in [1] (the latter only for the case of $\mathbb{P}_1$ elements). For this last method it is known that it respects the discrete maximum principle (at least in Delaunay meshes in two space dimensions), and thus the results are expected to lie within the bounds, at least for the mesh from Figure 1a. In our experiments we have used the values $p = 8$ and $\gamma_0 = 0.75$, (see [1] for details on the formulation of the method). We have carried out the comparison by taking a cross-section along the line $y = x$. We focus our attention in two main points, namely, suppression of over and undershoots in the numerical solution, and in how diffused the interior layers are. We depict zooms of the cross-sections at the onset of the layers. We observe that, as expected, the CIP method by itself presents over- and under-shoots, while the BPM and AFC method do not. In fact, the function u_h^+ presents significantly smaller oscillations than CIP, while showing an accuracy comparable to that of the AFC method regarding the sharpness of the layers.

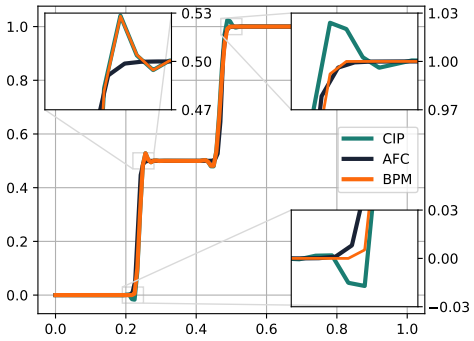
The same comparison has been carried out using the non-Delaunay mesh from Figure 1c, and very similar conclusions are drawn. Interestingly, when employing the BPM method for $\mathbb{P}_2$, $\mathbb{Q}_2$ and higher-order elements, even if the bounds are only imposed at the nodes, no noticeable undershoots (in terms of violations of the physical bounds) have been observed in the numerical solution.



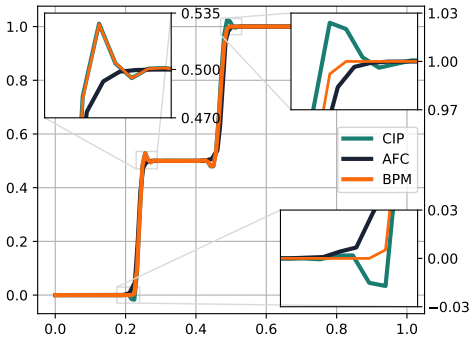
(a) Approximation using mesh 1a



(b) Approximation using mesh 1c

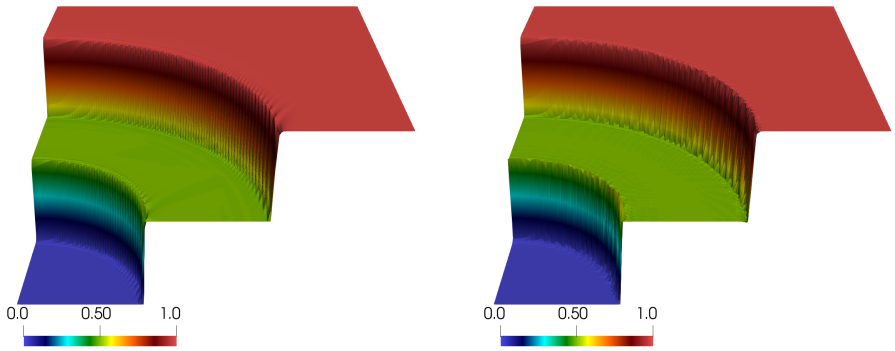


(c) Cross-section using mesh 1a

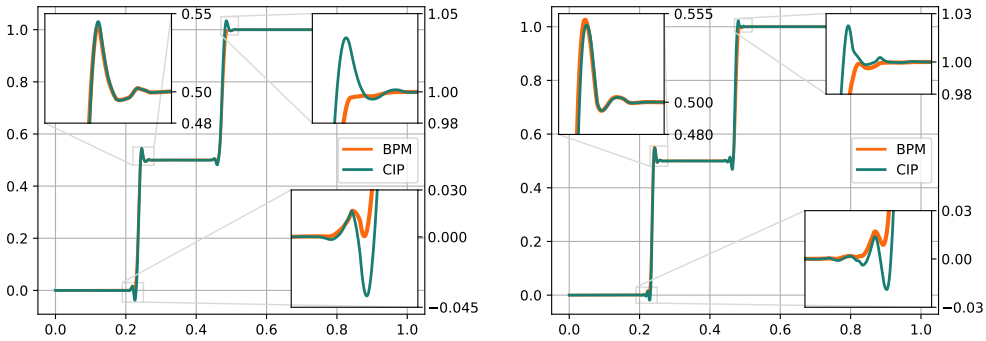


(d) Cross-section using mesh 1c

Figure 2: The approximation of the solution of Example 2 by the bound preserving method (BPM), using $\mathbb{P}_1$ elements and the meshes given in Figures 1a and 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 (33) $\gamma_\beta = 0.05$ and $\omega = 0.1$ has been used.

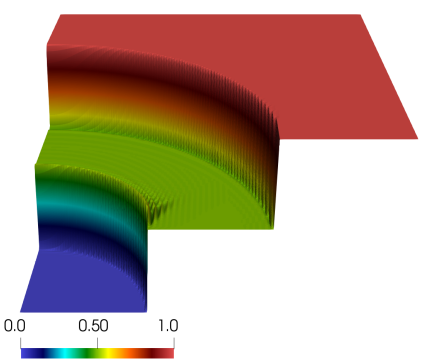


(a) Approximation using $\mathbb{P}_2$ elements and mesh 1c(b) Approximation using $\mathbb{P}_3$ elements and mesh 1c

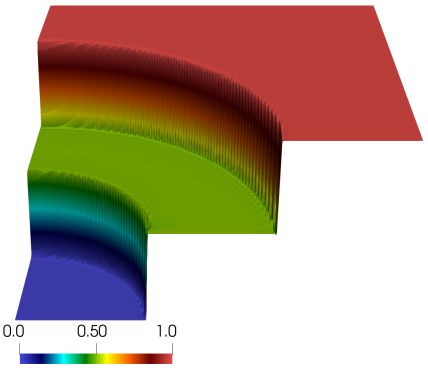


(c) Cross-section using $\mathbb{P}_2$ elements and mesh 1c (d) Cross-section using $\mathbb{P}_3$ elements and mesh 1c

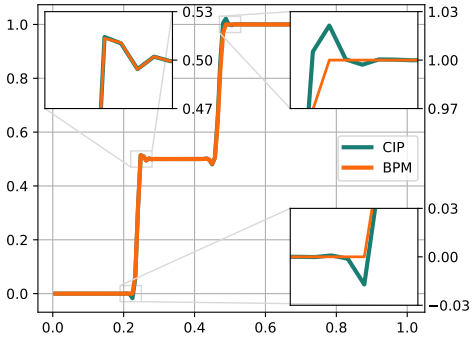
Figure 3: The approximation of the solution of Example 2 by the bound preserving method (BPM), using $\mathbb{P}_2$ and $\mathbb{P}_3$ elements and the meshes given in Figure 1c with $N = 129$. Cross-sections taken along the line $y = x$. For both methods the penalty (33) with $\gamma_\beta = 0.05$ was 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.



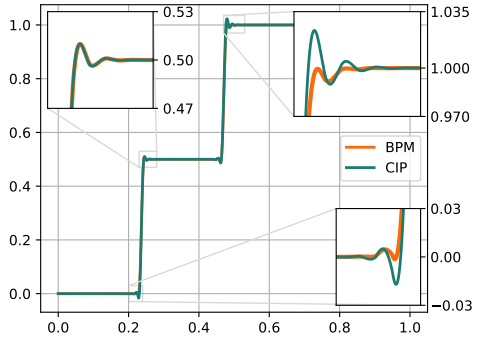
(a) Approximation using $\mathbb{Q}_1$ elements.



(b) Approximation using $\mathbb{Q}_2$ elements.



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



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

Figure 4: The approximation of the solution of Example 2 by the bound preserving method (BPM), using $\mathbb{Q}_1$ and $\mathbb{Q}_2$ elements and the mesh given in Figure 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 (33) $\gamma_\beta = 0.05$ 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.

Example 3 (A problem with an inner and a boundary layer)**.** For this last example we consider $f = 0$, $\mu = 0$, $\epsilon = 10^{-5}$, $\boldsymbol{\beta} = (\cos(\frac{\pi}{3}), \sin(\frac{\pi}{3}))^T$, and the Dirichlet boundary condition $u = g$ on Γ, where g is given by

$$g(x, y) = \begin{cases} 1 & \text{if } x = 0 \text{ or } y = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{72}$$

The problem consists of propagating a discontinuous entry condition to the interior, thus generating an interior layer that meets a boundary layer at $y = 1$. We have approximated this problem using the meshes depicted in Figures 1a–1c. For this experiment, especially the approximation of the outflow layer, the best results were provided by the method enhanced with the CIP stabilising term (27) and $\gamma = 0.01$. So, we only report the results obtained for this choice.

For the iterative method (69) we use $\omega = 0.1$, and we now report the number of fixed-point iterations needed to convergence:

	N	5	9	17	33	65	129
Mesh 1a	Itr.	109	143	177	212	249	249
Mesh 1c	Itr.	123	152	186	218	245	240

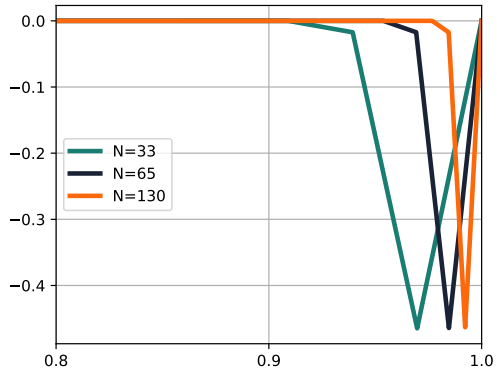
Table 8: Iterations needed to reach convergence using $\mathbb{P}_1$ elements and the meshes given in Figures 1a–1c, and the penalty term (27) with $\gamma = 0.01$ ($\omega = 0.1$).

	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 9: Iterations needed to reach convergence using $\mathbb{Q}_1$ and $\mathbb{Q}_2$ elements and the mesh given in Figure 1d, and the penalty term (33) with $\gamma_\beta = 0.01$ ($\omega = 0.1$).

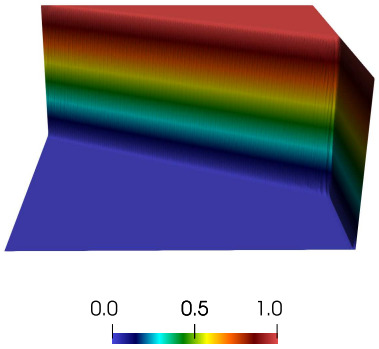
We now validate the statement made in Remark 3.7, by depicting in Figure 5 elevations of u_h^- using $\mathbb{P}_1$ elements. In Figure 5 we depict a zoom near the boundary of the cross-section along the line $x = 0.9$ of u_h^- for different levels of refinment. We can observe that as the mesh gets refined, the magnitute of u_h^- decreases slowly, and the support of u_h^- gets more and more localised, confirming what is stated in Remark 3.7.

Finally, in Figures 6-8 we depict the approximate solutions using BPM, AFC, and CIP methods for this problem as well as cross sections showing the nature of the interior layer. We also present cross-sections of u_h^+ across $y = 1 - x$ for Mesh 1a, and $N = 129$. Once again, we have performed comparisons with the linear CIP method (with the stabilising term (27) and $\gamma = 0.01$), and the AFC scheme as described in the last example. The results are depicted in Figure 6-8. Similar comments can be made about both sets of results, namely, that the current method removes the oscillations from the CIP solution successfully, while presenting a similar behaviour to AFC in terms of sharpness of the layers.

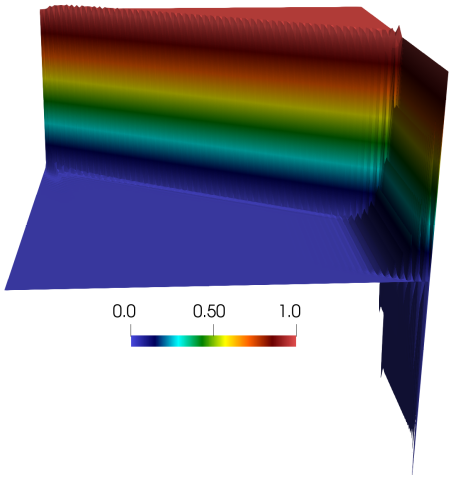


(a) Cross-section taken of u_h^- along $x = 0.9$.

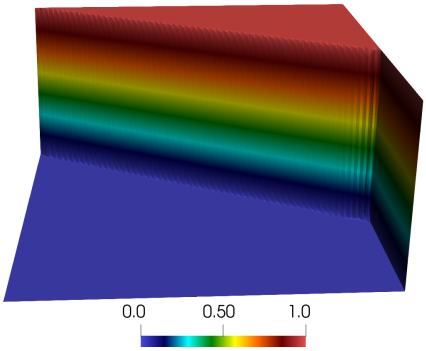
Figure 5: Cross-sections of u_h^- for Example 3 illustrating the behaviour at the boundary layers using $\mathbb{P}_1$ elements and the mesh given in Figures 1a.



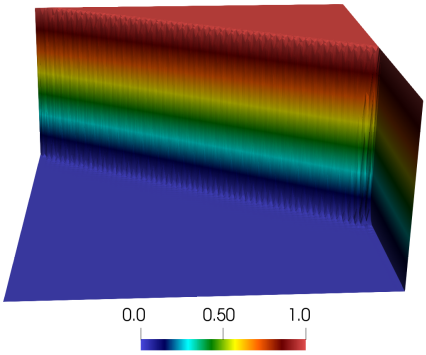
(a) AFC approximation using mesh 1a



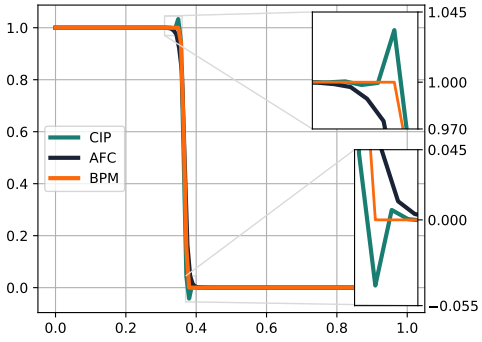
(b) CIP approximation using mesh 1a



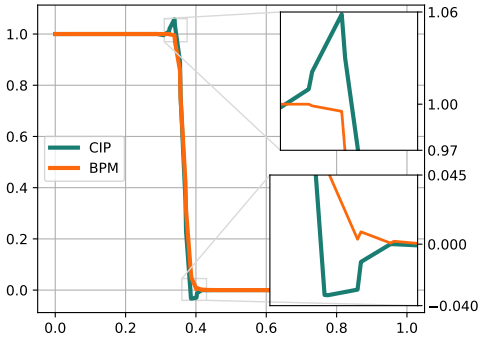
(c) BPM approximation using mesh 1a



(d) BPM approximation using mesh 1c

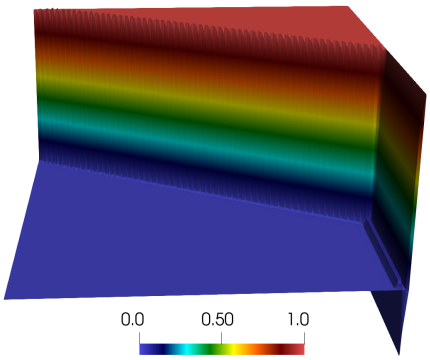


(e) Cross-section using mesh 1a

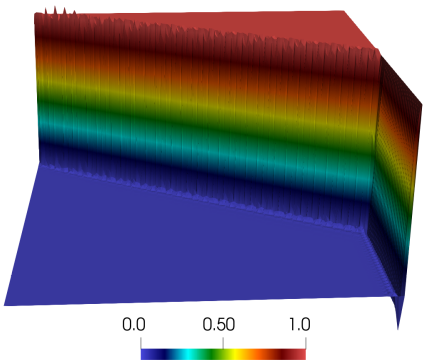


(f) Cross-section using mesh 1c

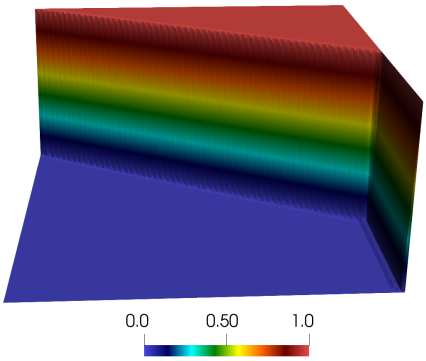
Figure 6: The approximation of the solution of Example 3 by the bound preserving method (BPM), using $\mathbb{P}_1$ elements and the meshes given in Figures 1a and 1c with $N = 129$. Cross-sections of the discrete solution of the BPM, CIP, and AFC methods taken about the line $y = x$. For AFC $p = 8$ and for BPM and CIP the penalty (33) $\gamma_\beta = 0.01$ was used ($\omega = 0.1$). For plotting the cross-sections we used linear interpolation between the nodes.



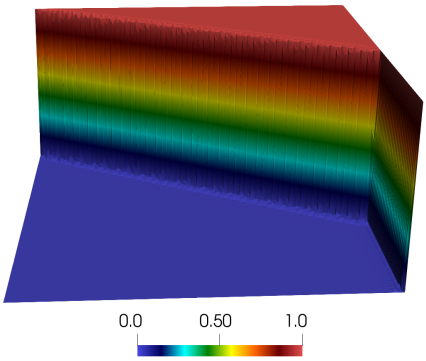
(a) CIP approximation using mesh 1a



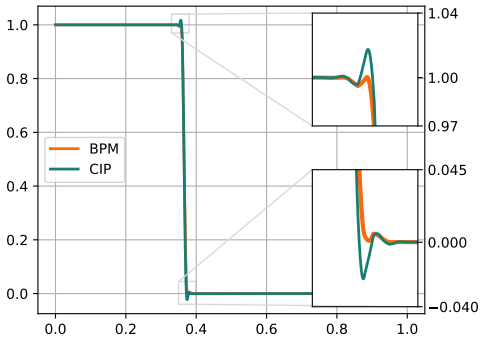
(b) CIP approximation using mesh 1c



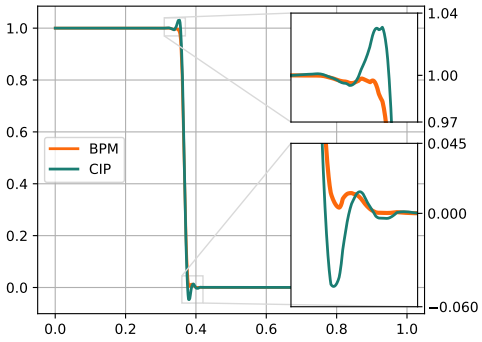
(c) BPM approximation using mesh 1a



(d) BPM approximation using mesh 1c

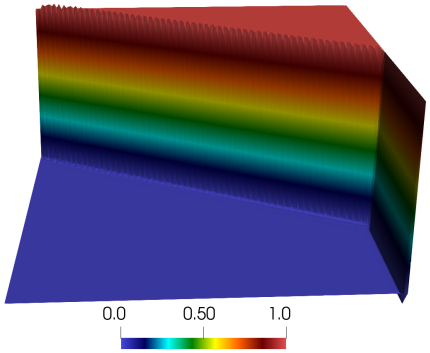


(e) Cross-section using mesh 1a

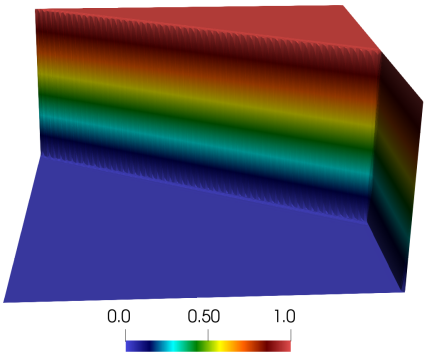


(f) Cross-section using mesh 1c

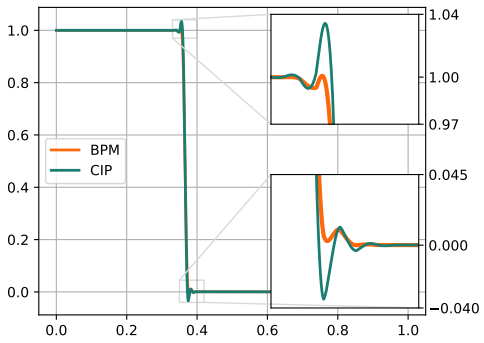
Figure 7: The approximation of the solution of Example 3 by the bound preserving method (BPM), using $\mathbb{P}_2$ elements and the meshes given in Figures 1a and 1c with $N = 129$. Cross-sections around the line $y = x$ of the solution of the BPM and CIP methods. For both methods the penalty (33) with $\gamma_\beta = 0.01$ 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.



(a) CIP approximation using mesh 1d.



(b) BPM approximation using mesh 1d.



(c) Cross-section taken using mesh 1d.

Figure 8: The approximation of the solution of Example 3 by the bound preserving method (BPM), using $\mathbb{Q}_2$ elements and the mesh given in Figure 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 (33) $\gamma_\beta = 0.01$ was used ($\omega = 0.1$).

6 Conclusions and outlook

The method proposed in this work constitutes an inexpensive and simple way to impose global bounds in the numerical solution to convection-diffusion equations (at least at the nodes of the triangulation, and globally for the lowest order case $k = 1$). Other than its well-posedness, we proved optimal order error estimates for the constrained part u_h^+. It is worth stressing that optimal-order error estimates are not common for this type of method (even in the piecewise linear case, see [4] for more details), so we believe this result constitutes a highlight of this paper. The numerical experiments presented show that the method provides an approximation that respects the global bounds of the continuous solution, while the inner and boundary layers are not excessively smeared.

Looking forward, the complementary part u_h^- has a localised support. This has been checked numerically, but the important question about whether this fact can be exploited in a posteriori error analysis remains open. In addition, the generality of this approach makes the applications of the method to unsteady nonlinear fluids problems, for example of porous media and Allen-Cahn type, quite natural. These, amongst other areas, are topics of ongoing research that shall be reported in upcoming publications.

Acknowledgements The work of AA, GRB, and TP has been partially supported by the Leverhulme Trust Research Project Grant No. RPG-2021-238. TP is also partially supported by EPRSC grants EP/W026899/2, EP/X017206/1 and EP/X030067/1. The authors also want to thank Emmanuil Geourgoulis and Andreas Veeser for numerous very helpful and stimulating discussions.

References

[1] G. R. Barrenechea, E. Burman, and F. Karakatsani, *Edge-based nonlinear diffusion for finite element approximations of convection-diffusion equations and its relation to algebraic flux-correction schemes*, Numer. Math., 135 (2017), pp. 521–545.

[2] G. R. Barrenechea, E. H. Georgoulis, T. Pryer, and A. Veeser, *A nodally bound-preserving finite element method*, IMA Journal of Numerical Analysis, (2023), p. drad055.

[3] G. R. Barrenechea, V. John, and P. Knobloch, *An algebraic flux correction scheme satisfying the discrete maximum principle and linearity preservation on general meshes*, Math. Models Methods Appl. Sci., 27 (2017), pp. 525–548.

[4] ——, *Finite element methods respecting the discrete maximum principle for convection-diffusion equations*, SIAM Review, 66 (2024), pp. 3–88.

[5] E. Burman and A. Ern, *Stabilized Galerkin approximation of convection-diffusion-reaction equations: discrete maximum principle and convergence*, Math. Comp., 74 (2005), pp. 1637–1652 (electronic).

[6] E. Burman and P. Hansbo, *Edge stabilization for Galerkin approximations of convection-diffusion-reaction problems*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 1437–1453.

[7] E. Celledoni and J. Jackaman, *Discrete conservation laws for finite element discretisations of multisymplectic PDEs*, Journal of Computational Physics, 444 (2021), p. 110520.

[8] Q. Cheng and J. Shen, *A new Lagrange multiplier approach for constructing structure preserving schemes, I. Positivity preserving*, Computer Methods in Applied Mechanics and Engineering, 391 (2022), p. 114585.

[9] P. G. CIARLET AND P.-A. RAVIART, *Maximum principle and uniform convergence for the finite element method*, Comput. Methods Appl. Mech. Engrg., 2 (1973), pp. 17–31.

[10] A. ERN AND J.-L. GUERMOND, *Finite Elements I*, Springer, 2021.

[11] ——, *Finite Elements II*, Springer, 2021.

[12] J. A. EVANS, T. J. R. HUGHES, AND G. SANGALLI, *Enforcement of constraints and maximum principles in the variational multiscale method*, Comput. Methods Appl. Mech. Engrg., 199 (2009), pp. 61–76.

[13] L. C. EVANS, *Partial differential equations*, vol. 19 of Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, second ed., 2010.

[14] R. FATTAL AND R. KUPFERMAN, *Constitutive laws for the matrix-logarithm of the conformation tensor*, Journal of Non-Newtonian Fluid Mechanics, 123 (2004), pp. 281–285.

[15] J. GIESSELMANN, C. MAKRIDAKIS, AND T. PRYER, *Energy consistent discontinuous Galerkin methods for the Navier–Stokes–Korteweg system*, Mathematics of Computation, 83 (2014), pp. 2071–2099.

[16] S. K. GODUNOV, *A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics*, Mat. Sb. (N.S.), 47 (89) (1959), pp. 271–306.

[17] J.-L. GUERMOND, *Stabilization of Galerkin approximations of transport equations by subgrid modeling*, ESAIM: Mathematical Modelling and Numerical Analysis, 33 (1999), pp. 1293–1316.

[18] W. HÖHN AND H.-D. MITTELMANN, *Some remarks on the discrete maximum-principle for finite elements of higher order*, Computing, 27 (1981), pp. 145–154.

[19] X. HUANG AND J. SHEN, *Bound/positivity preserving SAV schemes for the Patlak-Keller-Segel-Navier-Stokes system*, Journal of Computational Physics, 480 (2023), p. 112034.

[20] V. JOHN, A. LINKE, C. MERDON, M. NEILAN, AND L. G. REBHOLZ, *On the divergence constraint in mixed finite element methods for incompressible flows*, SIAM Rev., 59 (2017), pp. 492–544.

[21] D. KINDERLEHRER AND G. STAMPACCHIA, *An introduction to variational inequalities and their applications*, SIAM, 2000.

[22] P. KNOBLOCH AND G. LUBE, *Local projection stabilization for advection–diffusion–reaction problems: One-level vs. two-level approach*, Applied Numerical Mathematics, 59 (2009), pp. 2891–2907.

[23] C. 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 (2014), pp. 994–1002.

[24] D. KUZMIN, *Algebraic flux correction for finite element discretizations of coupled systems*, in Proceedings of the Int. Conf. on Computational Methods for Coupled Problems in Science and Engineering, M. Papadrakakis, E. Oñate, and B. Schrefler, eds., CIMNE, Barcelona, 2007, pp. 1–5.

[25] C. LU, W. HUANG, AND E. 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 (2013), pp. 24–36.

[26] C. G. MAKRIDAKIS, *On the Babuška–Osborn approach to finite element analysis: L 2 estimates for unstructured meshes*, Numerische Mathematik, 139 (2018), pp. 831–844.

[27] A. MIZUKAMI AND T. 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 (1985), pp. 181–193.

[28] M. RENARDY AND R. C. ROGERS, *An introduction to partial differential equations*, vol. 13 of Texts in Applied Mathematics, Springer-Verlag, New York, second ed., 2004.

[29] J. J. W. VAN DER VEGT, Y. XIA, AND Y. XU, *Positivity preserving limiters for time-implicit higher order accurate discontinuous Galerkin discretizations*, SIAM Journal on Scientific Computing, 41 (2019), pp. A2037–A2063.

[30] J. XU AND L. ZIKATANOV, *A monotone finite element scheme for convection-diffusion equations*, Math. Comp., 68 (1999), pp. 1429–1446.