[1]

BLENDING LOW-ORDER STABILISED FINITE ELEMENT METHODS: A POSITIVITY-PRESERVING LOCAL PROJECTION METHOD FOR THE CONVECTION-DIFFUSION EQUATION

GABRIEL R. BARRENECHEA, ERIK BURMAN, AND FOTINI KARAKATSANI

Abstract. In this work we propose a nonlinear blending of two low-order stabilisation mechanisms for the convection-diffusion equation. The motivation for this approach is to preserve monotonicity without sacrificing accuracy for smooth solutions. The approach is to blend a first-order artificial diffusion method, which will be active only in the vicinity of layers and extrema, with an optimal order local projection stabilisation method that will be active on the smooth regions of the solution. We prove existence of discrete solutions, as well as convergence, under appropriate assumptions on the nonlinear terms, and on the exact solution. Numerical examples show that the discrete solution produced by this method remains within the bounds given by the continuous maximum principle, while the layers are not smeared significantly.

1. Introduction

The design and analysis of stabilised finite element methods for convection–diffusion equations remains a challenging problem. In particular if the method is required to have (close to) optimal convergence where the exact solution is smooth, but to preserve the monotonicity properties of the continuous problem in the vicinity of layers. The standard approach has been to combine a linear stabilisation method that ensures accuracy in the smooth part of the solution and control of the propagation of perturbations from layers with a nonlinear so-called shock-capturing term that is designed to diminish, or ideally, eliminate the local spurious oscillations close to the layer. For an overview and a critical evaluation of such methods we refer to [22, 23] and references therein.

The design of such shock-capturing terms have typically been residual based [21, 14, 9], matching residual based stabilisation methods such as SUPG. The idea that finite element shock-capturing terms should be designed with the objective of satisfying a discrete maximum principle was pioneered by Mizukami and Hughes in [35], and further discussed in [14, 9]. However when symmetric stabilisation methods such as local projection stabilisation (LPS) or continuous interior penalty (CIP) methods are used, classical shock-capturing terms appear to be less natural. Instead, our objective in this paper is to explore the idea of designing a method that switches from a low order, but monotone, method that acts in the vicinity of layers, to an optimal, non monotone, method that is active in smooth regions. This main philosophy can be tracked back to the seminal work [6], and has been explored in different guises since, especially in the context of algebraic flux correction schemes in [40, 32, 33, 34], and more recently in the work of Kuzmin et. al. [30, 28, 38, 27, 24, 29], and Guermond et al. [20, 19]. See also the

[1]Corresponding author: gabriel.barrenechea@strath.ac.uk

1991 *Mathematics Subject Classification.* 65N12, 65N30.

recent work [26] for an idea based on a low-order local-projection type method, applied to the transport problem.

In this work we propose a particular realisation of the general approach described in the previous paragraph. More precisely, we will develop an idea introduced in [4]. Therein it was suggested that in the framework of a local projection method (or subgrid viscosity method), where the stabilisation takes the form of a penalty on the gradient of u_h minus the projection of u_h onto some smaller space, i.e. on $\nabla(u_h - \pi_H u_h)$, a nonlinear switch $\alpha(u_h) \in [0,1]$ could be introduced, $\nabla(u_h - \alpha(u_h)\pi_H u_h)$ taking the value 1 in the smooth part and 0 close to layers, hence turning off the projection part in the vicinity of layers. This makes the stabilisation degenerate to first order viscosity in the non-smooth part of the approximate solution so that the spurious oscillations are damped or even completely eliminated provided the mesh satisfies certain geometric conditions.

It turns out, however, that since the parameters for the first order viscosity and the LPS-term are of different magnitude, the idea can not be realised in this simple fashion, but instead the first order linear diffusion and the high order stabilisation term must be blended together locally using the nonlinear switch $\alpha(u_h)$ (similar approaches have been advocated recently by Ern and Guermond [16] and Badia and Hierro [1], using different stabilisation methods and slightly different focus). Below we design a nonlinear LPS method based on these ideas. We show that the method satisfies a discrete maximum principle under suitable assumptions on the mesh (depending on the diffusion operator), that the nonlinear discrete problem admits (at least) one solution and discuss what properties are required from the approximate solution and the nonlinear stabilisation in order to obtain an optimal a priori error estimate for smooth solutions, including the effect both of the linear and the nonlinear stabilisation operator. The above results are, essentially, independent of the concrete definition of the blending parameter, as long as it satisfies the basic requirements. We modify slightly two known limiters that have been applied in the context of Algebraic Flux Correction (AFC) schemes, and use them as a blending parameter. We then test them numerically, focussing on the accuracy and elimination of spurious oscillations.

The rest of the manuscript is organised as follows. The remainder of this introduction will be devoted to present the notations and necessary preliminary results. The bulk of this work is Section 2, where we describe the linear diffusion and LPS methods used in this work, and the way to blend them. An existence and convergence analysis is carried out, and the discrete maximum principle is discussed, under rather general assumptions on the nonlinear switch $\alpha(u_h)$. The two definitions used for this switch are presented in Section 3, and are tested by several numerical experiments in Section 4. Finally, we draw some conclusions and perspectives.

1.1. **The model problem, notations and preliminary results.** Throughout this work we adopt standard notation for Sobolev spaces. In particular, for $D \subset \mathbb{R}^d$ we denote by $(\cdot,\cdot)_D$ the inner product in $L^2(D)$ (or $L^2(D)^d$, if necessary). For $\ell \geq 0$, we denote by $\|\cdot\|_{\ell,D}$ ($|\cdot|_{\ell,D}$) the norm (seminorm) in $H^\ell(D)$. We will also adopt the usual convention that $H^0(D) = L^2(D)$.

Let $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, be a bounded polygonal (polyhedral) domain with a Lipschitz-continuous boundary $\partial\Omega$. We consider the steady-state convection-diffusion-reaction equation

$$\begin{aligned} -\varepsilon\,\Delta u + \boldsymbol{b}\cdot\nabla u + \sigma\, u &= f \quad \text{in } \Omega\,, \\ u &= g \quad \text{on } \partial\Omega\,, \end{aligned} \tag{1.1}$$

where $\varepsilon > 0$, $\boldsymbol{b} \in W^{1,\infty}(\Omega)$, and $\sigma \in L^\infty(\Omega)$ stand for diffusion, convection, and reaction, respectively, and $f \in L^2(\Omega)$, $g \in H^{\frac{1}{2}}(\partial\Omega)$ are given functions. Just to simplify the presentation we will suppose that $\nabla\cdot\boldsymbol{b} = 0$ in Ω. The weak problem associated to (1.1) is: Find $u \in H^1(\Omega)$

such that $u = g$ on $\partial\Omega$, and

$$a(u,v) = (f,v)_\Omega \qquad \forall\, v \in H^1_0(\Omega)\,, \tag{1.2}$$

where a is the bilinear form given by

$$a(u,v) := \varepsilon(\nabla u, \nabla v)_\Omega + (\boldsymbol{b}\cdot\nabla u, v)_\Omega + (\sigma u, v)_\Omega\,. \tag{1.3}$$

It is well-known (cf. [15]) that if the data of the problem satisfy $\sigma \geq \sigma_0 \geq 0$ in Ω, then (1.2) has a unique solution in $H^1(\Omega)$. The solution of (1.2) also satisfies the following properties (see, e.g., [17]).

Definition 1.1 (Strong maximum principle with $\sigma \geq 0$)**.** *Assume that $u \in C^2(\Omega)\cap C(\bar\Omega)$. If $f \geq 0$ in Ω (resp. $f \leq 0$ in Ω) and u attains a nonpositive minimum (resp. nonnegative maximum) over $\bar\Omega$ at an interior point, then u is constant in Ω.*

Definition 1.2 (Strong maximum principle with $\sigma = 0$)**.** *Assume that $u \in C^2(\Omega)\cap C(\bar\Omega)$. If $f \geq 0$ in Ω (resp. $f \leq 0$ in Ω) and u attains a minimum (resp. maximum) over $\bar\Omega$ at an interior point, then u is constant in Ω.*

We now describe the main notations, and the main preliminary results for the discrete problems we will study in this work. Let $\mathscr{T}_h$ be a family of shape-regular triangulations of $\bar\Omega$ into disjoint d-simplices K. We denote by h_K the diameter of K and $h := \max\{h_K : K \in \mathscr{T}_h\}$. We associate with the triangulation $\mathscr{T}_h$ the finite element spaces

$$\mathcal{V}_h := \{\chi \in H^1(\Omega) \,:\, \chi|_K \in \mathbb{P}_1(K)\,\forall K \in \mathscr{T}_h\}, \quad \text{and} \quad \mathcal{V}_h^0 := \mathcal{V}_h \cap H^1_0(\Omega), \tag{1.4}$$

where, for $\ell \geq 0$, $\mathbb{P}_\ell(D)$ is the space of polynomials of degree at most ℓ on D.

We let $\mathscr{F}_h$ be the set of the interior facets of $\mathscr{T}_h$, that is the set of $(d-1)$-simplices which are not included in the boundary $\partial\Omega$, and let h_F be the diameter of $F \in \mathscr{F}_h$. In addition, for each $F \in \mathscr{F}_h$ we denote by K_F^+ and K_F^- the two simplices in $\mathscr{T}_h$ such that $F = K_F^+ \cap K_F^-$ and we introduce the patch $K_F := K_F^+ \cup K_F^-$. For an element $K \in \mathscr{T}_h$, we define the set of its facets as $\mathscr{F}(K) := \{F \in \mathscr{F}_h : F \subset \partial K\}$. By $[\![v]\!]_F$ we will denote the jump of a function v across F, and will omit the subindex F when it is clear from the context.

For each vertex p_i of $\mathscr{T}_h$ we denote by ψ_i the associated nodal basis function, by Ω_i the set of simplices K which share p_i, i.e. $\Omega_i := \{K \in \mathscr{T}_h : K \cap p_i \neq \emptyset\}$, and denote the set of all facets of Ω_i by $\mathscr{F}(\Omega_i)$. By $\mathscr{F}(p_i)$ we denote the set of interior facets sharing p_i, that is $\mathscr{F}(p_i) := \{F \in \mathscr{F}_h : F \cap p_i \neq \emptyset\}$. We also define the set of neighboring nodes of p_i by

$$S_i := \{j \neq i \,:\, p_j \text{ shares an interior edge with } p_i\}\,.$$

Finally, for two vertices p_i and p_j belonging to the same simplex $K \in \mathscr{T}_h$, we denote by E_{ij} the edge connecting the vertices p_i and p_j, and $F_{i,K}$ and $F_{j,K}$ the facets in K opposite to p_i and p_j, respectively.

We shall assume from now on that the mesh satisfies the following property:

Assumption 1.1. *[Hypothesis of Xu and Zikatanov] For every internal edge E_{ij} the following inequality holds*

$$\frac{1}{d(d-1)} \sum_{K \supset E_{ij}} |\omega_{ij}^K| \cot(\theta_{ij}^K) \geq 0\,, \tag{1.5}$$

where ω_{ij}^K is the $(d-2)$-dimensional simplex $F_{i,K}\cap F_{j,K}$ opposite to the edge E_{ij}, θ_{ij} is the angle between the facets $F_{i,K}$ and $F_{j,K}$, and where $\sum_{K\supset E_{ij}}$ means summation over all simplexes K containing E_{ij}.

In the two-dimensional setting, this hypothesis implies that the mesh is of Delaunay type, this is, the sum of the two angles opposite to the same facet (edge) F is less than, or equal to, π (for details, see [39]).

Finally, we introduce the discrete analogues of the maximum principle.

Definition 1.3 (Strong DMP)**.** *We say that the semilinear form $\tilde{a}(\cdot;\cdot)$ satisfies the strong DMP property if the following is valid: For all $u_h \in \mathcal{V}_h$ and for all vertices $p_i \in \Omega$, if u_h has a local minimum (respectively, local maximum) on the vertex p_i over Ω_i, then there exist positive constants a_F, $F \in \mathscr{F}(p_i)$, such that*

$$\tilde{a}(u_h;\psi_i) \leq -\sum_{F\in\mathscr{F}(p_i)} a_F |\, [\![\nabla u_h]\!]_F \,| \tag{1.6}$$

(respectively, $\tilde{a}(u_h;\psi_i) \geq \sum_{F\in\mathscr{F}(p_i)} a_F |\, [\![\nabla u_h]\!]_F \,|$).

Definition 1.4 (Weak DMP)**.** *The semilinear form $\tilde{a}(\cdot;\cdot)$ is said to satisfy the weak maximum principle property if in Definition 1.3 we further assume that the local minimum is non-positive (respectively, the local maximum is non-negative).*

Observe that in the linear case the left hand side of (1.6) simply represents the scalar product of line i of the system matrix and the vector of solution coefficients. Hence, since the right-hand side of (1.6) is negative or zero, the satisfaction of this condition leads to a contradition whenever the right-hand side f is non-negative and the solution is not a constant in Ω_i.

2. A nonlinear local projection method

The method presented in Section 2.3 below will result as a natural blend of a linear diffusion method and a local projection stabilised method. Hence, we describe first its two main components.

2.1. A linear artificial diffusion method.

We start by introducing a function $u_{gh} \in \mathcal{V}_h$ such that its trace approximates the non-homogeneous Dirichlet boundary condition g in (1.1). We then propose the first discretisation method for (1.1): Find $u_h \in \mathcal{V}_h$ such that $u_h - u_{gh} \in \mathcal{V}_h^0$ and

$$a(u_h, v_h) + a_{LD}(u_h, v_h) = (f, v_h)_\Omega \qquad \forall\ v_h \in \mathcal{V}_h^0\,, \tag{2.1}$$

where $a(\cdot,\cdot)$ has been defined in (1.3), and $a_{LD}(\cdot,\cdot)$ corresponds to a linear diffusion stabilisation term given by

$$a_{LD}(u_h, v_h) = \sum_{F\in\mathscr{F}_h} \tau_F\ (\nabla u_h\,, \nabla v_h)_{K_F}\ , \tag{2.2}$$

where

$$\tau_F = c_0\,(\|\boldsymbol{b}\|_{\infty,K_F} + h_F\|\sigma\|_{\infty,K_F})\,h_F\,, \tag{2.3}$$

and $c_0 \geq 0$ is a constant. In addition to Assumption 1.1, we will assume that the mesh (and the physical coefficients) satisfy the following assumption:

Assumption 2.1. *For all interior node p_i, the following relation holds*

$$\max_{E\in\mathscr{F}(\Omega_i)} \tau_E \leq \left(1 + \frac{1}{2\tilde{C}}\right) \min_{E\in\mathscr{F}(\Omega_i)} \tau_E\,, \tag{2.4}$$

where $\tilde{C}$ is the constant used in (2.14) *below.*

In order that this method preserves positivity, we start by presenting the following preliminary result (see [11] for its proof).

Lemma 2.1. *If $u_h \in \mathcal{V}_h$ has a local extremum on the vertex p_i, then*

$$|\nabla u_h|_K| \leq \sum_{F\in\mathscr{F}(p_i)} |\,[\![\nabla u_h]\!]_F\,| \quad \forall K \subset \Omega_i. \tag{2.5}$$

We shall next show that the parameters τ_F in (2.2) may be chosen such that the bilinear form $a(\cdot,\cdot)+a_{LD}(\cdot,\cdot)$ satisfies a discrete maximum principle if the mesh satisfies Assumption 1.1. For simplicity we assume that $d=2$ throughout the proof. The result can be extended to the three-dimensional case by working as in [11].

Theorem 2.1. *There exists a positive constant $c_\star$ such that, if $c_0 > c_\star$, then the bilinear form $a(\cdot,\cdot)+a_{LD}(\cdot,\cdot)$ satisfies the weak DMP property.*

Proof. Let us assume that u_h has a local minimum on the interior vertex $p_i \in \Omega$ over the macro-element Ω_i, and that $u_h(p_i) \leq 0$. Under Assumption 1.1 it can be shown, [39, 36], that

$$(\nabla u_h, \nabla\psi_i)_\Omega = -\sum_{F\in\mathscr{F}(p_i)} \frac{h_F}{2} |\,[\![\nabla u_h]\!]_F\,|\,. \tag{2.6}$$

Furthermore, the convective term may be bounded as

$$(\boldsymbol{b}\cdot\nabla u_h, \psi_i)_K \leq \frac{|K|}{3}\|\boldsymbol{b}\|_{\infty,K}|\nabla u_h|_K|\,. \tag{2.7}$$

If u_h changes sign (or vanishes) in K, then using a Taylor expansion we deduce that $\|u_h\|_{\infty,K} \leq h_K\|\nabla u_h\|_{\infty,K}$. Since ∇u_h is constant in K, this leads to

$$(\sigma u_h, \psi_i)_K \leq \frac{|K|}{3}\|\sigma\|_{\infty,K} h_K|\nabla u_h|_K|\,. \tag{2.8}$$

Note that if u_h is negative in K, the reaction term $(\sigma u_h,\psi_i)_\Omega$ is negative and the above inequality holds trivially. In view of (2.5) and Lemma 2.1, we can thus derive the following bound for the convection-reaction term

$$\begin{aligned} |(\boldsymbol{b}\cdot\nabla u_h + \sigma u_h, \psi_i)_\Omega| &\leq \sum_{K\subset\Omega_i} |(\boldsymbol{b}\cdot\nabla u_h + \sigma u_h, \psi_i)_K| \\ &\leq \frac{1}{3}\sum_{K\subset\Omega_i} |K|(\|\boldsymbol{b}\|_{\infty,K} + \|\sigma\|_{\infty,K}h_K) \sum_{F\in\mathscr{F}(p_i)} |\,[\![\nabla u_h]\!]_F\,|. \end{aligned} \tag{2.9}$$

By using the shape regularity of the mesh, there exists a constant $\rho > 1$, independent of h, such that

$$|K| \leq \rho \min_{F\in\mathscr{F}(p_i)} h_F^2 \quad \text{and} \quad h_K \leq \rho \min_{F\in\mathscr{F}(p_i)} h_F\,, \quad \text{for all } K\in\Omega_i. \tag{2.10}$$

Also, the number of the elements K of Ω_i is bounded by a fixed number η_ρ, again independent of h. We can thus conclude that

$$|(\boldsymbol{b}\cdot\nabla u_h + \sigma u_h, \psi_i)_\Omega| \leq \frac{\rho\,\eta_\rho}{3} \sum_{F\in\mathscr{F}(p_i)} (\|\boldsymbol{b}\|_{\infty,K_F} + h_F\|\sigma\|_{\infty,K_F})\, h_F^2\, |\,[\![\nabla u_h]\!]_F\,|. \tag{2.11}$$

We finally analyse the artificial diffusion term. To do this, we define the following two (complementary) subsets of Ω_i:

$$\Omega_i^- := \{K\in\mathscr{T}_h : K\subset\Omega_i\,,\ (\nabla u_h,\nabla\psi_i)_K \leq 0\} \quad , \quad \Omega_i^+ := \{K\in\mathscr{T}_h : K\subset\Omega_i\,,\ (\nabla u_h,\nabla\psi_i)_K > 0\}\,.$$

Using these sets the artificial diffusion term can be written as

$$\begin{aligned}
&\sum_{F\in\mathscr{F}_h} \tau_F \,(\nabla u_h\,,\nabla\psi_i)_{K_F} = \sum_{K\subset\Omega_i}\Big(\sum_{F\in\mathscr{F}(K)}\tau_F\Big)\,(\nabla u_h\,,\nabla\psi_i)_K\\
&= \sum_{K\subset\Omega_i^-}\Big(\sum_{F\in\mathscr{F}(K)}\tau_F\Big)\,(\nabla u_h,\nabla\psi_i)_K + \sum_{K\subset\Omega_i^+}\Big(\sum_{F\in\mathscr{F}(K)}\tau_F\Big)\,(\nabla u_h,\nabla\psi_i)_K\\
&\le 3\Big(\min_{E\in\mathscr{F}(\Omega_i)}\tau_E\Big)\sum_{K\subset\Omega_i^-}(\nabla u_h,\nabla\psi_i)_K + 3\Big(\max_{E\in\mathscr{F}(\Omega_i)}\tau_E\Big)\sum_{K\subset\Omega_i^+}(\nabla u_h,\nabla\psi_i)_K\\
&= 3\Big(\min_{E\in\mathscr{F}(\Omega_i)}\tau_E\Big)(\nabla u_h,\nabla\psi_i)_\Omega + 3\left(\max_{E\in\mathscr{F}(\Omega_i)}\tau_E - \min_{E\in\mathscr{F}(\Omega_i)}\tau_E\right)\sum_{K\subset\Omega_i^+}(\nabla u_h,\nabla\psi_i)_K\,. \qquad (2.12)
\end{aligned}$$

Now, we analyse the last expression above term by term. First, using (2.6) the first term gives

$$3\Big(\min_{E\in\mathscr{F}(\Omega_i)}\tau_E\Big)(\nabla u_h,\nabla\psi_i)_\Omega = -3\left(\min_{E\in\mathscr{F}(\Omega_i)}\tau_E\right)\sum_{F\in\mathscr{F}(p_i)}\frac{h_F}{2}|\,[\![\nabla u_h]\!]_F\,|\,. \qquad (2.13)$$

Next, for $K\in\Omega_i^+$, Lemma 2.1 gives

$$(\nabla u_h,\nabla\psi_i)_K \le \sum_{F\in\mathscr{F}(p_i)}|\,[\![\nabla u_h]\!]_F\,|(1,|\nabla\psi_i|)_K \le \sum_{F\in\mathscr{F}(p_i)}|\,[\![\nabla u_h]\!]_F\,|\,C_0h_K\,,$$

and then, the mesh regularity leads to

$$\sum_{K\subset\Omega_i^+}(\nabla u_h,\nabla\psi_i)_K \le \sum_{F\in\mathscr{F}(p_i)}|\,[\![\nabla u_h]\!]_F\,|\sum_{K\subset\Omega_i^+}C_0h_K \le \tilde{C}\sum_{F\in\mathscr{F}(p_i)}\frac{h_F}{2}|\,[\![\nabla u_h]\!]_F\,|\,. \qquad (2.14)$$

Then, replacing (2.13) and (2.14) in (2.12), and using Assumption 2.1 we arrive at the following final bound for the artificial diffusion term

$$\begin{aligned}
\sum_{F\in\mathscr{F}_h}\tau_F\,(\nabla u_h,\nabla\psi_i)_{K_F} &\le -3\left(\min_{E\in\mathscr{F}(\Omega_i)}\tau_E\right)\sum_{F\in\mathscr{F}(p_i)}\frac{h_F}{2}|\,[\![\nabla u_h]\!]_F\,|\\
&\quad + 3\tilde{C}\left(\max_{E\in\mathscr{F}(p_i)}\tau_E - \min_{E\in\mathscr{F}(p_i)}\tau_E\right)\sum_{F\in\mathscr{F}(p_i)}\frac{h_F}{2}|\,[\![\nabla u_h]\!]_F\,|\\
&\le -\frac{3}{2}\left(\min_{E\in\mathscr{F}(\Omega_i)}\tau_E\right)\sum_{F\in\mathscr{F}(p_i)}\frac{h_F}{2}|\,[\![\nabla u_h]\!]_F\,|\,. \qquad (2.15)
\end{aligned}$$

Gathering all the above computations, we obtain the following bound

$$\begin{aligned}
&\varepsilon(\nabla u_h,\nabla\psi_i)_\Omega + (\boldsymbol{b}\cdot\nabla u_h,\psi_i)_\Omega + (\sigma u_h,\psi_i)_\Omega + \sum_{F\in\mathscr{F}_h}\tau_F\,(\nabla u_h,\nabla\psi_i)_{K_F}\\
&\le -\sum_{F\in\mathscr{F}(p_i)}\left(\varepsilon + \frac{3}{2}\min_{E\in\mathscr{F}(\Omega_i)}\tau_E - \frac{2\rho\,\eta_\rho}{3}\left(\|\boldsymbol{b}\|_{\infty,K_F} + h_F\|\sigma\|_{\infty,K_F}\right)h_F\right)\frac{h_F}{2}|\,[\![\nabla u_h]\!]_F\,|\,.
\end{aligned}$$

Thus, choosing

$$c_\star \ge \frac{4\rho\,\eta_\rho}{9}\,\frac{\max_{E\in\mathscr{F}(\Omega_i)}\left\{(\|\boldsymbol{b}\|_{\infty,K_E} + h_E\|\sigma\|_{\infty,K_E})h_E\right\}}{\min_{E\in\mathscr{F}(\Omega_i)}\left\{(\|\boldsymbol{b}\|_{\infty,K_E} + h_E\|\sigma\|_{\infty,K_E})h_E\right\}}\,, \qquad (2.16)$$

the result follows. □

Remark 2.2. *It is worth remarking that, thanks to the regularity of the mesh, the constant $c_\star$ depends on σ and $\boldsymbol{b}$, but not on the size of the elements.*

Moreover, if the mesh is supposed to be weakly acute, then Assumption 2.1 is not necessary. As a matter of fact, in this case we have $(\nabla u_h, \nabla\psi_i)_K \leq 0$ for each $K \in \Omega_i$ and (2.12) can be replaced by

$$\begin{aligned}\sum_{F\in\mathscr{F}_h} \tau_F\ (\nabla u_h\,, \nabla\psi_i)_{K_F} &= \sum_{K\subset\Omega_i}\Big(\sum_{F\in\mathscr{F}(K)}\tau_F\Big)\ (\nabla u_h, \nabla\psi_i)_K\\ &\leq 3\sum_{K\subset\Omega_i}\min_{E\in\mathscr{F}(K)}\tau_E\,(\nabla u_h, \nabla\psi_i)_K\\ &= 3\Big(\min_{E\in\mathscr{F}(\Omega_i)}\tau_E\Big)(\nabla u_h, \nabla\psi_i)_\Omega\\ &\leq -3\left(\min_{E\in\mathscr{F}(\Omega_i)}\tau_E\right)\sum_{F\in\mathscr{F}(p_i)}\frac{h_F}{2}|\,[\![\nabla u_h]\!]_F\,|\,, \end{aligned}\tag{2.17}$$

and the result follows.

Finally, another possibility for removing Assumption 2.1 is to replace the artificial diffusion introduced here by an edge oriented anisotropic diffusion of the form (in the two-dimensional case)

$$a_{ALD}(u_h, v_h) := \sum_{F\in\mathscr{F}_h}\tau_F\sum_{\hat{F}\in\mathscr{F}(K_F)} h_{\hat{F}}^2\nabla u_h\cdot\mathbf{t}_{\hat{F}}\nabla v_h\cdot\mathbf{t}_{\hat{F}},$$

where $\mathbf{t}_F$ denotes the tangential vector of the facet (i.e., edge in the two-dimensional case) F. For details see [11].

We end this section by remarking that if $\sigma = 0$ then the last result can be strengthened. In fact, the following corollary arises.

Corollary 2.1. *If $\sigma = 0$, then the bilinear form $a(\cdot,\cdot) + a_{LD}(\cdot,\cdot)$ satisfies the strong DMP 1.3 .*

2.2. **A local projection method.** For each $F \in \mathscr{F}_h$ we introduce the projection operator $G_F : H^1(K_F) \to \mathbb{P}_1(K_F)/\mathbb{R}$ defined by

$$(\nabla G_F w, \nabla v_h)_{K_F} = (\nabla w, \nabla v_h)_{K_F} \quad \forall\ v_h \in \mathbb{P}_1(K_F)/\mathbb{R}\,. \tag{2.18}$$

From its definition, G_F is an orthogonal projection, and then it enjoys the following stability

$$\|\nabla G_F v\|_{0,K_F} \leq \|\nabla v\|_{0,K_F} \quad \forall\ v \in H^1(K_F). \tag{2.19}$$

Using G_F we introduce the following local projection method: Find $u_h \in \mathcal{V}_h$ such that $u_h - u_{gh} \in \mathcal{V}_h^0$ and

$$a(u_h, v_h) + s_h(u_h, v_h) = (f, v_h)_\Omega \qquad \forall\ v_h \in \mathcal{V}_h^0\,, \tag{2.20}$$

where $s_h(\cdot,\cdot)$ is the stabilisation term given by

$$s_h(u_h, v_h) = \sum_{F\in\mathscr{F}_h}\gamma_F\ (\nabla(I - G_F)u_h\,, \nabla v_h)_{K_F}\ , \tag{2.21}$$

and

$$\gamma_F = \gamma_0\ \min\left\{h_F\,(\|\boldsymbol{b}\|_{\infty,K_F} + \|\sigma\|_{\infty,K_F}h_F), \frac{h_F^2}{\varepsilon}\right\}\,, \tag{2.22}$$

where γ_0 is a positive constant.

Remark 2.3. *We now shed some light on the LPS terminology for Method* (2.20)*. First, as written, this method can be seen as a subgrid viscosity scheme, reminiscent of the one proposed in* [18]*. A more LPS-oriented version of* (2.20) *would consist on replacing the stabilisation term* $s_h(u_h, v_h)$ *by*

$$\tilde{s}_h(u_h, v_h) = \sum_{F \in \mathscr{F}_h} \gamma_F \left((I - \tilde{G}_F)\nabla u_h \, , \nabla v_h \right)_{K_F} , \tag{2.23}$$

where $\tilde{G}_F \nabla u_h \in (\mathbb{P}_0(K_F))^d$ *is the projection defined by*

$$(\tilde{G}_F \nabla u_h, \boldsymbol{v}_h)_{K_F} = (\nabla u_h, \boldsymbol{v}_h)_{K_F} \quad \forall \, \boldsymbol{v}_h \in (\mathbb{P}_0(K_F))^d \, . \tag{2.24}$$

Since both $\tilde{G}_F \nabla u_h$ *and* $\boldsymbol{v}_h$ *are constant vectors in* K_F *we can easily see that*

$$\tilde{G}_F \nabla u_h = \frac{|K_F^+| \, \nabla u_h|_{K_F^+} + |K_F^-| \, \nabla u_h|_{K_F^-}}{|K_F|} \, . \tag{2.25}$$

On the other hand, since in (2.18) *both* $\nabla G_F u_h$ *and* ∇v_h *are constant vectors in* K_F*, the following holds*

$$\nabla G_F u_h = \frac{|K_F^+| \, \nabla u_h|_{K_F^+} + |K_F^-| \, \nabla u_h|_{K_F^-}}{|K_F|} . \tag{2.26}$$

Consequently, $\nabla G_F u_h = \tilde{G}_F \nabla u_h$ *and* $s_h(u_h, v_h) = \tilde{s}_h(u_h, v_h)$*. This means that the present method is indeed a LPS method.*

As one further rewriting of (2.20)*, the stabilisation term may be rewritten as follows*

$$s_h(u_h, v_h) = \sum_{F \in \mathscr{F}_h} \gamma_F \, \frac{|K_F^+||K_F^-|}{|K_F|} \, [\![\nabla u_h]\!]_F \, [\![\nabla v_h]\!]_F \quad \forall \, u_h, v_h \in \mathcal{V}_h . \tag{2.27}$$

Thus, Method (2.20) *is one particular realisation of the Continuous Interior Penalty (CIP) method proposed in* [13] *(see also* [4]*, where a connection of this type has been used to carry out the analysis).*

We shall now discuss the convergence of the LPS method introduced in (2.20). We introduce the norms $\| \cdot \|_{LPS}$ and $\| \cdot \|_{CIP}$, respectively, as follows

$$\|v\|_{LPS} := \left(\varepsilon \, |v|_{1,\Omega}^2 + \|\sigma^{\frac{1}{2}} \, v\|_{0,\Omega}^2 + s_h(v, v) \right)^{\frac{1}{2}} , \tag{2.28}$$

and

$$\|v\|_{CIP} := \left(\varepsilon |v|_{1,\Omega}^2 + \|\sigma^{\frac{1}{2}} \, v\|_{0,\Omega}^2 + \sum_{F \in \mathscr{F}_h} \gamma_F \, h_F^{-1} \, \frac{|K_F^+||K_F^-|}{|K_F|} \| \, [\![\nabla v]\!] \, \|_{0,F}^2 \right)^{\frac{1}{2}} . \tag{2.29}$$

Furthermore, we let $i_h : H^2(\Omega) \to \mathcal{V}_h$ denote the Lagrange interpolation operator satisfying

$$\|w - i_h w\|_{0,K} + h_K \|\nabla(w - i_h w)\|_{0,K} \leq C h_K^2 |w|_{2,K} \quad \text{for all } w \in H^2(K) \, . \tag{2.30}$$

The following result confirms the fact that the convergence of this method is optimal.

Theorem 2.2. *Let* $u \in H^2(\Omega)$ *be the solution of* (1.1) *and* $u_h \in \mathcal{V}_h$ *its finite element approximation given by* (2.20)*. Then*

$$\|u - u_h\|_{LPS} \leq C \big(\varepsilon + \|\boldsymbol{b}\|_{\infty,\Omega} h + (\sigma_0^{-1} \|\nabla \boldsymbol{b}\|_{\infty,\Omega}^2 + \|\sigma\|_{\infty,\Omega}) h^2 \big)^{\frac{1}{2}} \, h \, \|u\|_{2,\Omega} . \tag{2.31}$$

Proof. The error $u - u_h$ may be split as $u - u_h = (u - i_h u) + (i_h u - u_h) =: \rho_h + e_h$. Then, we have

$$\|u - u_h\|_{LPS} \leq \|u - i_h u\|_{LPS} + \|i_h u - u_h\|_{LPS}. \tag{2.32}$$

From (2.27) it follows that

$$\|e_h\|_{LPS} = \|e_h\|_{CIP}\,, \tag{2.33}$$

and thus, following the same steps as in [13], we have

$$\|e_h\|_{CIP} \leq C\big(\varepsilon + \|\boldsymbol{b}\|_{\infty,\Omega} h + (\sigma_0^{-1}\|\nabla \boldsymbol{b}\|_{\infty,\Omega}^2 + \|\sigma\|_{\infty,\Omega})h^2\big)^{\frac{1}{2}}\, h\, \|u\|_{2,\Omega}\,. \tag{2.34}$$

Moreover, the interpolation error $\|\rho_h\|_{LPS}$ may be bounded as follows

$$\begin{aligned}\|\rho_h\|_{LPS} &= \Big(\varepsilon\,|\rho_h|_{1,\Omega}^2 + \|\sigma^{\frac{1}{2}}\,\rho_h\|_{0,\Omega}^2 + s_h(\rho_h,\rho_h)\Big)^{\frac{1}{2}} \\ &\leq C\left((\varepsilon + \|\sigma\|_{\infty,\Omega}h^2)h^2\|u\|_{2,\Omega}^2 + \sum_{F\in\mathscr{F}_h}\gamma_F\,\|\nabla(I - G_F)\rho_h\|_{0,K_F}^2\right)^{\frac{1}{2}}.\end{aligned} \tag{2.35}$$

By using the stability (2.19), (2.30), and the regularity of the mesh, we get

$$\|\nabla(I - G_F)\rho_h\|_{0,K_F} \leq \|\nabla\rho_h\|_{0,K_F} \leq C h_F^2\,|u|_{2,K_F}^2\,. \tag{2.36}$$

This leads to the following bound

$$\sum_{F\in\mathscr{F}_h}\gamma_F\,\|\nabla(I - G_F)\rho_h\|_{0,K_F}^2 \leq C\,(\|\boldsymbol{b}\|_{\infty,\Omega} + \|\sigma\|_{\infty,\Omega}h)\,h^3\|u\|_{2,\Omega}^2\,,$$

and the desired result follows from (2.35) and (2.34). □

Remark 2.4. *The same analysis can be carried out in the case $\sigma_0 = 0$, but the error estimate will depend on ε. More precisely, let π_h denote the $L^2(\Omega)$-orthogonal projection onto $\mathcal{V}_h^0$, and let us suppose that $\boldsymbol{b}\cdot\boldsymbol{n} = 0$ on $\partial\Omega$. Then, following analogous steps as the ones presented in* [7]*, we can show that there exists a constant $C > 0$, independent of h and ε, such that for all $v_h \in \mathcal{V}_h^0$ the following holds*

$$(u - \pi_h u, \boldsymbol{b}\cdot\nabla v_h)_\Omega \leq C\left(\|\boldsymbol{b}\|_{1,\infty,\Omega}\|u - \pi_h u\|_{0,\Omega}\|v_h\|_{0,\Omega} + \|\boldsymbol{b}\|_{\infty,\Omega}^{\frac{1}{2}}\|h^{-\frac{1}{2}}(u - \pi_h u)\|_{0,\Omega}s_h(v_h,v_h)^{\frac{1}{2}}\right). \tag{2.37}$$

Then, using this last inequality, the following bound can be obtained for the discrete error

$$\|e_h\|_{CIP} \leq C\left(\varepsilon + h + \frac{\|\boldsymbol{b}\|_{1,\infty,\Omega}h^2}{\varepsilon}\right)^{\frac{1}{2}} h\|u\|_{2,\Omega}\,, \tag{2.38}$$

which leads to an estimate containing a term of the form $h\varepsilon^{-\frac{1}{2}}$. The negative power of ε gets somehow compensated by the h in the numerator. This estimate is, nevertheless, only satisfactory for sufficiently small h.

2.3. **The nonlinear LPS scheme.** As we could see in the last two sections, the linear diffusion scheme (2.1) preserves the maximum principle, while the Local Projection Stabilisation scheme (2.20) enjoys optimal convergence. Then, our purpose in this section is to blend these two approaches, in such a way that the resulting scheme preserves positivity for rough solutions, while keeping the optimal accuracy of the LPS scheme where the solution is smooth. The method proposed reads as follows: Find $u_h \in \mathcal{V}_h$ such that $u_h - u_{gh} \in \mathcal{V}_h^0$ and

$$\tilde{a}(u_h; v_h) := a(u_h, v_h) + d_h(u_h; u_h, v_h) = (f, v_h)_\Omega \qquad \forall\ v_h \in \mathcal{V}_h^0\,, \tag{2.39}$$

where the stabilisation term $d_h(\cdot\,;\cdot,\,\cdot)$ is defined as follows

$$
\begin{aligned}
d_h(w_h;u_h,v_h) = &\sum_{F\in\mathscr{F}_h} \tau_F\,\alpha_F(w_h)(\nabla u_h,\nabla v_h)_{K_F}\\
&+\sum_{F\in\mathscr{F}_h} \gamma_F\,(1-\alpha_F(w_h))(\nabla(I-G_F)u_h\,,\nabla(I-G_F)v_h)_{K_F}\,.
\end{aligned}
\tag{2.40}
$$

Here, for every $F\in\mathscr{F}_h$, $\alpha_F:\mathcal{V}_h\to[0,1]$ is a continuous function of u_h, and the parameters τ_F and γ_F are defined in (2.3) and (2.22), respectively. For the moment, the only extra requirement imposed on the functions α_F is for them to satisfy $\alpha_F(u_h)=1$ whenever u_h has a local extremum in a node of K_F. The following result is the first step towards proving the solvability of (2.39).

Lemma 2.5. *Let τ_F and γ_F be defined by* (2.3) *and* (2.22)*, respectively, and let $T_h:\mathcal{V}_h^0\to[\mathcal{V}_h^0]'$ be the nonlinear operator defined by*

$$
[T_h z_h, v_h] := a(z_h+u_{gh},v_h)+d_h(z_h+u_{gh};z_h+u_{gh},v_h)-(f,v_h)_\Omega,\quad z_h,v_h\in\mathcal{V}_h^0\,.
\tag{2.41}
$$

Then, there exist two positive constants C_1,C_2 such that

$$
[T_h z_h, z_h] \geq C_1|z_h|_{1,\Omega}^2 - C_2(\|u_{gh}\|_{1,\Omega}^2+\|f\|_{0,\Omega}^2)\,.
\tag{2.42}
$$

Proof. First, using the definition of T_h and the ellipticity of the bilinear form $a(\cdot,\cdot)$ we obtain

$$
\begin{aligned}
[T_h z_h, z_h] = &\|\sigma^{\frac12} z_h\|_{0,\Omega}^2+\varepsilon|z_h|_{1,\Omega}^2+d_h(z_h+u_{gh};z_h,z_h)+a(u_{gh},z_h)\\
&+d_h(z_h+u_{gh};u_{gh},z_h)-(f,v_h)_\Omega.
\end{aligned}
\tag{2.43}
$$

Using $0\leq\alpha_F(z_h+u_{gh})\leq 1$ we arrive at

$$
\begin{aligned}
d_h(z_h+u_{gh};z_h,z_h) = &\sum_{F\in\mathscr{F}_h}\big\{\tau_F\,\alpha_F(z_h+u_{gh})\|\nabla z_h\|_{0,K_F}^2\\
&+\gamma_F\,(1-\alpha_F(z_h+u_{gh}))\|\nabla(I-G_F)z_h\|_{0,K_F}^2\big\}\geq 0\,.
\end{aligned}
\tag{2.44}
$$

Moreover, the Poincaré inequality gives

$$
\begin{aligned}
|a(u_{gh},z_h)| &= |(\sigma u_{gh},z_h)_\Omega+\varepsilon(\nabla u_{gh},\nabla z_h)_\Omega+(\boldsymbol{b}\cdot\nabla u_{gh},z_h)_\Omega|\\
&\leq C\left(\|\sigma\|_{\infty,\Omega}+\varepsilon+\|\boldsymbol{b}\|_{\infty,\Omega}\right)\|u_{gh}\|_{1,\Omega}\,|z_h|_{1,\Omega}.
\end{aligned}
\tag{2.45}
$$

By using the orthogonality of G_F, the mesh regularity, and the definitions (2.3) and (2.22) we obtain that

$$
\begin{aligned}
|d_h(z_h+u_{gh};u_{gh},z_h)| &\leq \sum_{F\in\mathscr{F}_h}(\tau_F+\gamma_F)\,\|\nabla u_{gh}\|_{0,K_F}\|\nabla z_h\|_{0,K_F}\\
&\leq Ch\,(\|\sigma\|_{\infty,\Omega}\,h+\|\boldsymbol{b}\|_{\infty,\Omega})\|u_{gh}\|_{1,\Omega}\,|z_h|_{1,\Omega}.
\end{aligned}
$$

Summarising, T_h satisfies

$$
[T_h z_h, z_h] \geq \varepsilon|z_h|_{1,\Omega}^2 - C\left(\|\sigma\|_{\infty,\Omega}+\varepsilon+\|\boldsymbol{b}\|_{\infty,\Omega}\right)\|u_{gh}\|_{1,\Omega}\,|z_h|_{1,\Omega}-\|f\|_{0,\Omega}\|z_h\|_{0,\Omega}\,.
\tag{2.46}
$$

The claimed result can be derived by applying the Poincaré and Young inequalities to the last relation. □

The solvability of (2.39) appears then as a corollary of the previous result.

Theorem 2.3. *The discrete problem* (2.39) *has at least one solution.*

Proof. First, we notice that u_h solves (2.39) if and only if $T_h(u_h - u_{gh}) = 0$. Let K be any positive constant satisfying $K^2 > \frac{C_2}{C_1}(\|u_{gh}\|^2_{1,\Omega} + \|f\|^2_{0,\Omega})$ and $v_h \in \mathcal{V}_h^0$ such that $|v_h|_{1,\Omega} = K$. Then, (2.42) gives

$$[T_h v_h, v_h] \geq C_1 |v_h|^2_{1,\Omega} - C_2(\|u_{gh}\|^2_{1,\Omega} + \|f\|^2_{0,\Omega}) > 0. \tag{2.47}$$

Hence, since α_F are continuous functions, the operator T_h is continuous. This allows us to use a consequence of Brower's fixed point theorem (see, [37, p. 162, Lemma 1.4]) to deduce that there exists $\tilde{z}_h \in \mathcal{V}_h^0$ such that $|\tilde{z}_h|_{1,\Omega} < K$ and $T_h(\tilde{z}_h) = 0$. Then, $u_h := z_h + u_{gh} \in \mathcal{V}_h$ solves (2.39). □

We end this section by making a short comment on the satisfaction of the discrete maximum principle. If u_h has a local extremum at an interior node, then $\alpha_F(u_h) = 1$ for all $F \in \mathscr{F}(\Omega_i)$. Hence

$$\tilde{a}(u_h; \psi_i) = a(u_h, \psi_i) + a_{LD}(u_h, \psi_i)\,,$$

and the proof of the discrete maximum principle for Method (2.39) follows from the proof of Theorem 2.1.

2.4. **Convergence of the nonlinear scheme.** The error will be measured in the following mesh-dependent norm:

$$|||v|||_h := \left(\varepsilon\,|v|^2_{1,\Omega} + \|\sigma^{\frac{1}{2}}\,v\|^2_{0,\Omega} + d_h(u_h; v, v)\right)^{\frac{1}{2}}\,. \tag{2.48}$$

As usual, we split the error $e := u - u_h$ as follows

$$e = u - u_h = (u - i_h u) + (i_h u - u_h) := \rho_h + e_h, \tag{2.49}$$

and start by noting that, using the orthogonality of G_F, the definition of τ_F and γ_F, and standard interpolation estimates, ρ_h can be bounded as

$$|||\rho_h|||_h \leq C\left(\varepsilon + \|\boldsymbol{b}\|_{\infty,\Omega} h + \|\sigma\|_{\infty,\Omega} h^2\right)^{\frac{1}{2}} h\,\|u\|_{2,\Omega}\,. \tag{2.50}$$

In general it is not possible to prove optimal convergence of the nonlinear scheme even for smooth solutions regardless of the smoothness assumed. Here we will instead use the fact that numerical evidence shows that the nonlinear switch is active on a small set of the whole domain Ω. This then suggests an analysis that uses an a priori assumption the size of this subdomain and on its distance to the local extrema of the exact solution. Making these assumptions disregards the nonlinear coupling and is equivalent to considering a linear case where the switch α_F is a function of the space coordinate only. Nevertheless we believe that the resulting error analysis below gives some insights in what is required from the nonlinear stabilisation method in order for it to be both monotone and accurate.

Let

$$S_\alpha := \{K \in \mathcal{T}_h : \max_{F \in \mathscr{F}(K)} \alpha_F(u_h) > h^2\},$$

and assume that $\mathrm{meas}(S_\alpha) = O(h^s)$, with $s \in \mathbb{R}_+$. Observe that this means that the set on which the first order viscosity is active must diminish under mesh refinement. Even if this is a reasonable assumption for problems where the strong gradients are localized in space it can in general only be verified *a posteriori*. To keep $\alpha = 1$ in the vicinity of a local extremum, we must have $s \leq 1$.

Let $r \in \mathbb{R}_+$, let us define the set

$$S_{h,ext} := \{x \in \Omega : |\nabla u(x)| \leq C h^r |u|_{2,\infty,\Omega}\}\,,$$

and assume that

$$l_\alpha := \sup_{x\in S_\alpha} \inf_{y\in S_{h,ext}} |x-y| \leq Ch^r. \tag{2.51}$$

Under these assumptions there holds

$$|d_h(u_h; i_h u, i_h u)|^{\frac{1}{2}} \leq \left(\sum_{F\in\mathscr{F}_h} \tau_F \alpha_F(u_h) \|\nabla i_h u\|^2_{0,K_F} \right)^{\frac{1}{2}} + Ch^{\frac{3}{2}} |u|_{2,\Omega}\,,$$

where the second term is due to the linear stabilisation.

Since the extrema of u_h are in an $O(h^r)$, $r>0$, vicinity of the extrema of u, then we have using the mean value theorem,

$$\|\nabla u\|_{\infty,S_\alpha} \leq l_\alpha |u|_{2,\infty,\Omega} + \|\nabla u\|_{\infty,S_{h,ext}} \leq Ch^r |u|_{2,\infty,\Omega}\,,$$

and as a consequence

$$\|\nabla i_h u\|_{\infty,S_\alpha} \leq C(\|\nabla(i_h u - u)\|_{\infty,S_\alpha} + h^r \|u\|_{2,\infty,\Omega})\,,$$

and it follows that taking the parts d_h over $\Omega\setminus S_\alpha$ and S_α separately we get

$$|d_h(u_h; i_h u, i_h u)|^{\frac{1}{2}} \leq C(h^{\frac{3}{2}} \|\nabla u\|_{\infty,\Omega} + h^{\frac{1+s}{2}}(h+h^r)|u|_{2,\infty,\Omega} + h^{\frac{3}{2}}|u|_{2,\Omega}). \tag{2.52}$$

We see that to obtain optimal global convergence for smooth solutions we must have $r=\frac{1}{2}$, i.e. the approximate local extremum is in an $O(h^{\frac{1}{2}})$ neighbourhood of an exact extremum, allowing us to take $s=1$. More generally we need to assume $r+s/2\geq 1$. Note also that the above analysis shows that there is some flexibility regarding how quickly the nonlinear term has to be turned off in zones where the solution is flat. Indeed if r is big, or more precisely the associated contribution small, then s can be allowed to be smaller and hence S_α more spread out.

We also assume that there exists an interpolation operator i_h with optimal approximation properties satisfying

$$(u-i_h u, \boldsymbol{b}\cdot\nabla e_h) \leq \|h^{-\frac{1}{2}}(u-i_h u)\|_{0,\Omega} (d_h(u_h, e_h, e_h)^{\frac{1}{2}} + h^{\frac{1}{2}} \sigma_0^{-1/2} \|\nabla \boldsymbol{b}\|_{\infty,\Omega} |||e_h|||_h). \tag{2.53}$$

If, as what we did in Remark 2.4, we assume that $\boldsymbol{b}\cdot\mathbf{n}=0$ on $\partial\Omega$, this can be shown for the L^2-projection (see [7]).

Under these assumptions the convergence order of the nonlinear LPS scheme matches that of the linear LPS scheme as we show in the following Lemma.

Lemma 2.6. *Let us suppose that $u\in W^{2,\infty}(\Omega)$. Then, there exists $C>0$, independent of h and ε, such that*

$$|||e_h|||_h \leq C\big(\varepsilon + \|\boldsymbol{b}\|_{\infty,\Omega} h + \big(\|\sigma\|_{\infty,\Omega} + \sigma_0^{-1}\|\nabla\boldsymbol{b}\|^2_{\infty,\Omega}\big)\, h^2\big)^{\frac{1}{2}} h|u|_{2,\Omega} + Ch^{\frac{1+s}{2}}(h+h^r)|u|_{2,\infty,\Omega}\,. \tag{2.54}$$

If in addition the coefficients s and r introduced above satisfy $r+s/2\geq 1$, then

$$|||e_h|||_h \leq C(\varepsilon + \|\boldsymbol{b}\|_{\infty,\Omega} h + \big(\|\sigma\|_{\infty,\Omega} + \sigma_0^{-1}\|\nabla\boldsymbol{b}\|^2_{\infty,\Omega}\big)\, h^2)^{\frac{1}{2}} h|u|_{2,\infty,\Omega}\,.$$

Proof. We start by noticing that, since, for a given $w_h\in\mathcal{V}_h$, then $d_h(w_h;\cdot,\cdot)$ is a symmetric and positive semi-definite bilinear form, and the following holds

$$|d_h(u_h; e_h, i_h u)| \leq d_h(u_h; e_h, e_h)^{\frac{1}{2}} d_h(u_h; i_h u, i_h u)^{\frac{1}{2}}\,. \tag{2.55}$$

Then, using the ellipticity of the bilinear form a, the properties (2.53) and (2.55) we get

$$\begin{aligned}
|||e_h|||_h^2 =& a(e_h,e_h)+d_h(u_h;e_h,e_h) = -(f,e_h)_\Omega + a(i_hu,e_h)+d_h(u_h;i_hu,e_h)\\
=& -a(\rho_h,e_h)+d_h(u_h;i_hu,e_h)\\
\le& |||\rho_h|||_h|||e_h|||_h + (\rho_h,\boldsymbol{b}\cdot\nabla e_h)_\Omega + |d_h(u_h;i_hu,i_hu)|^{\frac12}|d_h(u_h;e_h,e_h)|^{\frac12}\\
\le& |||\rho_h|||_h|||e_h|||_h + \|h^{-\frac12}(u-i_hu)\|_{0,\Omega}(d_h(u_h,e_h,e_h)^{\frac12} + h^{\frac12}\sigma_0^{-1/2}\|\nabla\boldsymbol{b}\|_{\infty,\Omega}|||e_h|||_h)\\
&+ |d_h(u_h;i_hu,i_hu)|^{\frac12}|d_h(u_h;e_h,e_h)|^{\frac12}\,.
\end{aligned}$$

The claim follows using standard approximation of i_hu, the estimate (2.50) and (2.52) under the assumptions made on s and r. □

Remark 2.7. *Observe that trivially $r=s=0$ holds, which results in the error estimate $|||e_h|||_h \le Ch^{\frac12}$, which is the worst possible case.*

3. Definition of α_F

As was mentioned before, the only requirements imposed to the functions α_F were that they need to have a range in $[0,1]$, they need to take a value one when there is a local extremum in an node of K_F, and they need to be continuous. In the numerical results below we have used two different definitions for this parameter. Both of them are modifications of known parameters that have been used previously in the context of AFC schemes, and they respect one main designing principle for this method, namely, the method should reduce to linear diffusion in the vicinity of sharp layers and local extrema.

Since the numerical results presented in the next section are two-dimensional, we restrict the presentation to two space dimensions (then, facets are edges). Nevertheless, the definitions can be extended, without major difficulties, to the three-dimensional case.

3.1. **The Zalesak-Kuzmin limiters.** For this alternative we define the blending parameter using the Zalesak-Kuzmin limiters. These limiters have been used extensively in the context of AFC schemes (see, e.g., [25], and [3] for the analysis of the resulting method), and they have been modified in such a way they satisfy the requirements of the analysis presented in the previous sections.

Let $\mathbb{A}=(a_{ij})_{i,j=1,\dots,N}$ be the matrix arising from the Galerkin discretisation of (1.2) using $\mathcal{V}_h$ as discrete space, and considering that homogeneous Neumann conditions are imposed at the boundary. Then, for each pair i,j, we define $d_{ij}=-\max\{a_{ij},0,a_{ji}\}$, and, for a function $w_h\in\mathcal{V}_h$, we define the fluxes

$$f_{ij}=d_{ij}(w_h(p_j)-w_h(p_i))\,.$$

We then introduce the coefficients P_i^+, P_i^-, Q_i^+, Q_i^- computed for $i=1,\dots,N$ in the following way. First, these quantities are initialised by zero. Then we go through all pairs of indices $i,j\in\{1,\dots,N\}$ and perform the updates

$$\begin{aligned}
&P_i^+ := P_i^+ + \max\{0,f_{ij}\}\,,\ P_i^- := P_i^- - \max\{0,f_{ji}\} && \text{if } a_{ji}\le a_{ij}\,,\\
&Q_i^+ := Q_i^+ + \max\{0,f_{ji}\}\,,\ Q_i^- := Q_i^- - \max\{0,f_{ij}\} && \text{if } i<j\,,\\
&Q_j^+ := Q_j^+ + \max\{0,f_{ij}\}\,,\ Q_j^- := Q_j^- - \max\{0,f_{ji}\} && \text{if } i<j\,.
\end{aligned}$$

After having computed the values P_i^+, P_i^-, Q_i^+, Q_i^-, $i=1,\dots,N$, we define

$$R_i^+ := \min\left\{1,\frac{Q_i^+}{P_i^++\tau}\right\},\quad R_i^- := \min\left\{1,\frac{Q_i^-}{P_i^-+\tau}\right\},\qquad i=1,\dots,N\,.$$

Here, $\tau > 0$ is a regularisation parameter. Furthermore, these quantities are set to 1 at Dirichlet nodes, i.e.,

$$R_i^+ := 1\,, \quad R_i^- := 1\,, \qquad \text{if } p_i \text{ belongs to the Dirichlet boundary}\,.$$

Finally, for any $i, j \in \{1, \dots, N\}$ such that $a_{ji} \leq a_{ij}$, we set

$$\alpha_{ij} := \begin{cases} R_i^+ & \text{if } f_{ij} > 0\,, \\ 1 & \text{if } f_{ij} = 0\,, \\ R_i^- & \text{if } f_{ij} < 0\,, \end{cases} \qquad \alpha_{ji} := \alpha_{ij}\,. \tag{3.1}$$

It is worth mentioning that this algorithm is the one presented in [25] (that originates from the ideas of [40]), to which the symmetry condition $\alpha_{ij} = \alpha_{ji}$ has been added.

Then, adopting the convention that an internal edge F has endpoints x_i and x_j, the blending parameter has been chosen as:

$$\alpha_F(w_h) := \max\left\{1 - \alpha_{ij}(w_h) : p_i \,\text{and}\, p_j \,\text{share an edge of}\, K_F\right\}. \tag{3.2}$$

3.2. **A blending parameter based on the variation of u_h.** We use as blending parameter a slight modification of the limiter recently proposed in [2], and whose definition is as follows. First, for $w_h \in \mathcal{V}_h$, we define ξ_{w_h} as the unique element in $\mathcal{V}_h^0$ whose nodal values are given by

$$\xi_{w_h}(x_i) := \frac{\left|\sum_{j \in S_i} w_h(x_i) - w_h(x_j)\right|}{\sum_{j \in S_i} |w_h(x_i) - w_h(x_j)| + \tau}\,. \tag{3.3}$$

Again, $\tau > 0$ is a regularisation parameter. Then, on each F, α_F is defined by

$$\alpha_F(w_h) := \max\left\{\left[\xi_{w_h}(p_j)\right]^p \,:\, p_j \text{ is a vertex of } K_F\right\} \quad , \quad p \in\, [1, +\infty)\,. \tag{3.4}$$

The value for p will determine the rate of decay of the numerical diffusion with the distance to the critical points. A higher value for p will naturally reduce the amount of artificial diffusion added to the scheme, thus providing sharper layers. The flip side of that coin is that a higher value for p also makes the nonlinear system harder to solve, and it usually leads to a higher number of iterations needed, and eventually non-convergence of the fixed point iterations if p it too large. In our numerical experience, it is safe to consider values for p up to 10 or 15, but not higher. The values used for every particular case are reported in the captions of the respective figures.

It is important to remark that for both the above choices, the regularisation parameter τ is needed in order to ensure continuity of the resulting nonlinear form. The downside of that is that the blending parameter α_F is not exactly one, but rather $1/(1+\tau)$ in the vicinity of extrema. This, obviously, introduces a violation of the discrete maximum principle, but this violation is of the order of τ, which is below the tolerance for the nonlinear iterations. Then, we believe this does not affect the overall quality of the numerical results given by the method.

3.3. **Linearity preservation.** We end this section by stating that, provided with these two blending parameters, then the method is linearity preserving on structured meshes. More precisely, we will say that the mesh is symmetric with its internal edges if, for every edge $F \in \mathscr{F}_h$ with end points p_i, p_j, then there exists an edge $E \in \mathscr{F}_h$ with enpoints p_i and p_k, and $p_i - p_j = -(p_i - p_k)$. Moreover, we will define the neighbourhood $\omega_F := \{K' \in \mathscr{T}_h : K' \cap K_F \neq \emptyset\}$. Then, supposing that the mesh is symmetric with respect to its internal nodes, then for all $v \in \mathbb{P}_1(\omega_F)$ we have that $\alpha_F(v) = 0$ (for the case of the parameter given by (3.2) this is true

under the extra condition that the convection field $\boldsymbol{b}$ is constant). As a consequence, denoting $d_h(v; v, w_h) = \sum_{F\in\mathscr{F}_h} d_h^F(v; v, w_h)$, the following holds

$$d_h^F(v; v, w_h) = \gamma_F(\nabla(I - G_F)v, \nabla(I - G_F)w_h)_{K_F} = 0\,, \tag{3.5}$$

for all $w_h \in \mathcal{V}_h$, where in the last step we have used that the function v is a linear polynomial in K_F. This means that the stabilisation disappears whenever the discrete solution is a linear polynomial in a neighbourhood of the facet F. This property is believed to lead to better error convergence in unstructured meshes, and there is numerical evidence supporting this statement (see, e.g., [2], or the recent work [26] in the context of the transport equation), but a proof of this fact is lacking.

4. Numerical experiments

In this section we present the result of different numerical experiments illustrating the performance of the present method.

The nonlinear system (2.39) has been solved using the modified fixed-point algorithm proposed in [23]. Defining the discrete residual as

$$(R_h(u_h^n), v_h) = a(u_h^n, v_h) + d_h(u_h^n; u_h^n, v_h) - (f, v_h)_\Omega\,,$$

then the fixed point iterations are stopped when the relative residual (i.e., the norm of the residual divided by the norm of the right-hand-side) is smaller than, or equal to, 10^{-6} (unless otherwise specified). All iterations have been started with the solution of the LPS method (2.20).

Concerning the design of the parameters c_0 and γ_0, we have adopted an empirical approach. The parameter c_0 has to be large enough for the method to respect the discrete maximum principle. Since the formula (2.16) is not explicit, we have solved the problem using different values, and, in order to add as little numerical diffusion as possible, we have retained the smallest one for which the linear diffusion method respects the DMP. This value has changed from problem to problem, and we report them in the appropriate captions. For the parameter γ_0 more freedom is at hand, and the best results have been given using $\gamma_0 = 0.05$ in all calculations. Finally, concerning the regularisation parameter τ present in the definition of the blending parameters α_F, we have chosen them as the value $\tau = \text{DOLFIN\_EPS} = 3 \times 10^{-16}$, where DOLFIN_EPS is a preseted value in FEniCS (all of our codes have been written in FEniCS, see, e.g., [31]).

4.1. Two-dimensional examples.

In this section we will solve the following benchmark problems. We always take $\Omega = (0,1)^2$, $\varepsilon = 10^{-5}$, $\sigma = 0$, and:

Example 1 (Convection skew to the mesh)**.** *Problem* (1.1) *is considered with* $\boldsymbol{b}(x,y) = (\cos(\pi/3), \sin(\pi/3))^T$, $f = 0$*, and the boundary condition* $u = g$ *on* $\partial\Omega$*, where*

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

Our first objective was to compare both possible definitions of the blending parameters α_F. We have made numerous comparisons on this, and we just report here the result of a representative one, which helps us to draw some conclusions. We have solved Example 1 on a structured mesh containing $2 \times 64 \times 64$ elements, using both definitions for the parameter. The results are depicted in Figure 1. A more detailed view can be seen in Figure 2 where cross-sections of the solutions are depicted. From these results (and others that show the same pattern) we can observe that the blending parameters defined by (3.4) provide a solution which is sharper than the one given by the use of (3.2). Also, the fixed point iteration tends to be much slower when using

the blending parameters defined by (3.2) (in some cases we have even observed non-convergence of the iterations, or convergence after several thousands of iterations). Then, from now on, we will only consider method (2.39) provided with the blending parameters α_F defined by (3.4).

Next, we solve Example 1 using an unstructured mesh containing $7,970$ elements. In Figure 3 we present the result for the method (2.39). For comparison, we also present the results obtained using two alternative positivity preserving methods. The first one is the Algebraic Flux-Correction (AFC) scheme from [25]. This scheme is usually linked to time-dependent problems, but it was adapted to address steady-state problems in [25]. Its analysis has been recently carried out in [3], where a description of its use for steady-state problems such as the one solved in this work can be found. The second scheme is one recently proposed in [2], which is a nonlinear edge diffusion scheme, using limiters related to the ones defined in (3.4), and that can also be seen as an alternative flux limiter for the AFC schemes (see [2] for details).

For this example (and others as well) we see that the three methods provide qualitatively similar solutions. A more detailed comparison can be observed in Figure 4, where cross-sections along the lines $y = 0.9$ and $x = 0.25$ are depicted. For comparison, we have also included the results obtained by the shock-capturing method proposed in [11], and the SUPG method (the last one as an example of the reference linear method for the convection-diffusion equation). These cross-sections show that the results obtained with method (2.39) are of similar quality to both AFC and the edge diffusion method from [2]. As is to be expected though, the upwind character of the flux limiters defined in [25] makes the AFC method better suited to approximate outflow boundary layers (this has been observed in our numerical experience, not only for the problems described in this work). Also, the overlapping character of method (2.39) makes it more prone to add some more diffusivity to the problem. Despite this, the results show sharp layers for all the methods, and sharper than the ones provided by the method from [11], while respecting the discrete maximum principle.

We finally have solved this problem in a structured $2 \times 40 \times 40$ mesh. The results are depicted in Figure 5 for the same three elevations as before, and in Figure 6 for cross-sections. Essentially the same comments made before are valid in this case.

Example 2 (A rotating convective field, two inner layers)**.** *Problem* (1.1) *is considered with* $\boldsymbol{b}(x, y) = (-y, x)^T$*,* $f = 0$*, and the boundary condition*

$$u = g \quad \text{on } \Gamma^D, \qquad \partial_{\boldsymbol{n}} u = 0 \quad \text{on } \Gamma^N,$$

where $\Gamma^N = \{0\} \times (0, 1)$*,* $\Gamma^D = \partial\Omega \setminus \Gamma^N$*, and*

$$g(x, y) = \begin{cases} 1, & (x, y) \in (0.15, 0.45) \times \{0\}, \\ \cos^2\left(10\pi \frac{x-0.7}{3}\right), & (x, y) \in (0.55, 0.85) \times \{0\}, \\ 0, & \text{else on } \Gamma^D. \end{cases}$$

This is a problem recently proposed in [26]. This example poses some more issues in the convergence of the fixed-point iteration (especially for the AFC scheme). That is why we have relaxed slightly the convergence criterion for this problem to stopping the iterations when the normalised residual is smaller than 5×10^{-6}. The results (in the same order as it was done for the previous example) are depicted in Figures 7-8. In these figures we can observe that the results from Method (2.39) and the methods from [2] and the AFC scheme are very close, and the layers do not seem to have been smeared significantly, even close to the outflow. On the other hand, for the three methods that respect the DMP, the trigonometric profile that has been transported seems to have diffused slightly, as the result for the SUPG method shows. Since the problem is highly convection-dominated, then its exact solution is close to the one of the

Method	min	diff
AFC	1.16e-08	0.2066
BBK_ED	0.0	0.1248
CURRENT METHOD	0.0	0.1449
BE05_02	0.01017	0.2691
SUPG	0.034	0.0580

TABLE 1. Results for the different methods for the mesh consisting of $2 \times 64 \times 64$ elements.

transport problem (i.e., using $\varepsilon = 0$). This is why we add the profile of the solution of the tranport problem (labeled as "REFERENCE SOLUTION") in the plots for reference.

Example 3 (A problem with two inner layers)**.** *Problem* (1.1) *is considered with* $\boldsymbol{b}(x,y) = (1,0)^T$*, boundary condition* $u = 0$ *on* $\partial\Omega$*, and*

$$f(x,y) = \begin{cases} 16(1-2x) & \text{if } (x,y) \in [0.25, 0.75]^2\,, \\ 0, & \text{else.} \end{cases}$$

This problem has been proposed in [23] as a benchmark for problems for which nonlinear discretisations usually fail to provide satisfactory results. The exact solution is very close to the quadratic function $(4x-1)(3-4x)$ in the region $(0.25, 0.75)^2$, and is very close to zero (but positive) elsewhere in Ω (this function is referred to as 'REFERENCE SOLUTION' in the cross sections presented below). Despite the fact that the exact solution is non-negative, usually nonlinear methods give a numerical solution that is negative in some regions (see [23] for a detailed discussion on this topic). This does not contradict the DMP property, as the right-hand side f changes sign inside Ω, but it is somewhat upsetting. Elevations of the solution provided by different methods are depicted in Figure 9, while cross-sections along the lines $x = 0.5$ and $x = 0.8$ are depicted in Figure 10. The cross-sections along the line $x = 0.5$ show a similar behaviour of all the methods, with all of them providing values around 1.06 (rather than the "reference" value of 1). Surprisingly, as we can observe in Figure 10, bottom, the SUPG method seems to be the one that gives the most accurate results in this case (remember that the exact solution to this problem is very close to zero for $x \geq 0.8$). Out of the nonlinear schemes tested, the present method and the method from [2] seem to be the most accurate ones.

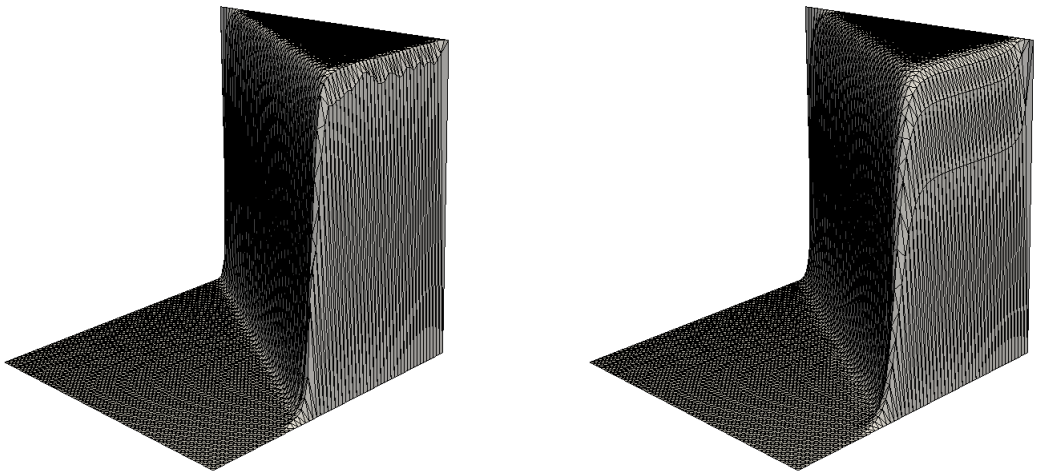
Finally, to make a more detailed analysis for this problem, we have repeated the study carrried out in [23] for this example. More precisely, we have computed the quantities

$$\text{min} := -\min_{0.4 \leq x \leq 0.6} u_h(x,y) \quad , \quad \text{diff} := \max_{x \geq 0.8} u_h(x,y) - \min_{x \geq 0.8} u_h(x,y)\,. \tag{4.1}$$

The results for the mesh of $2 \times 64 \times 64$ elements are given in Table 1, where we can observe that they confirm the discussion carried out in the last paragraph. The method that provides the smallest oscillations in the region $x \geq 0.8$ is the SUPG, followed by the current method and the method from [2] (presenting results which are very close to each other). Concerning the over and undershoots near the layers, the SUPG method does present an undershoot which is noticeably larger than the one given by the nonlinear methods.

5. Conclusions and future work

In this work we have blended two different strategies. One of them is built to satisfy the discrete maximum principle, and the other one converges optimally. This blend leads to a nonlinear scheme which has been proven to possess at least one solution, and whose solution does satisfy the DMP. The numerical results have shown that the numerical solution provided

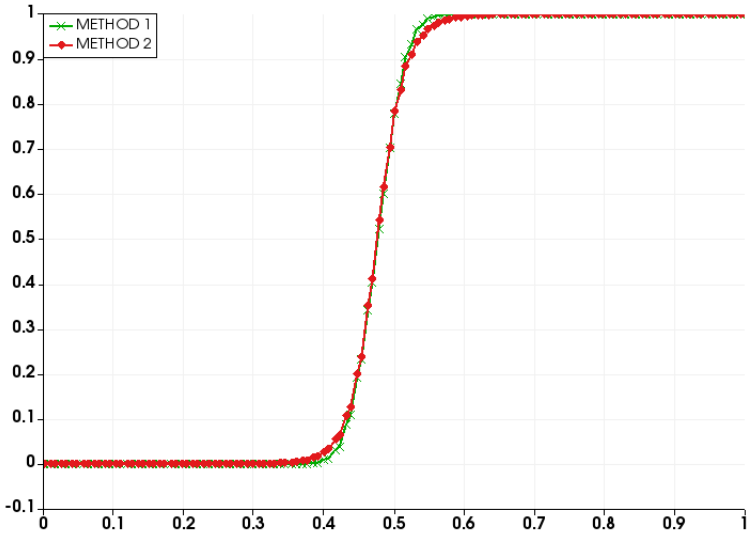


(a) (b)

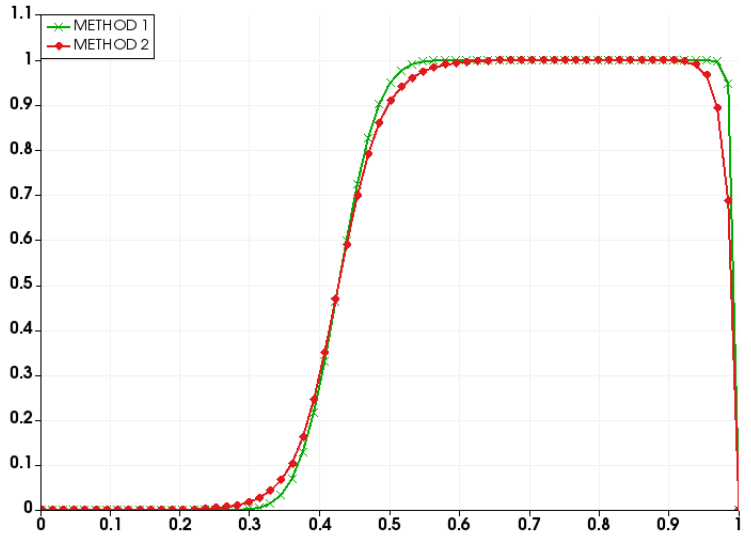
FIGURE 1. Solution for the method (2.39) using the limiters given by (3.4) (left, where 127 iterations were needed to reach convergence) and the ones given by (3.2) (right, where 323 iterations were needed to reach convergence). For both cases we used $c_0 = 0.3$.

by this method remains within the bounds provided by the continuous one, while presenting layers which are sharp. These results have also been compared to previously existing positivity-preserving methods providing results which are comparable to the previously existing ones. The final method is quite robust with respect to the convergence of the fixed point iterations. In fact, for Method (2.39) with the blending parameters (3.4) we have not found an example for which the fixed point iterative scheme does not converge. Still, for some examples the convergence can be very slow, leading to several hundreds of iterations. Then, the search for more efficient nonlinear iteration schemes than the fixed-point algorithm used in this work is an open problem. Observe also that in the transient case the above positivity results carry over for implicit time discretization using the backward Euler method following [10]. For explicit methods on the other hand the situation is less clear. However we expect the method to be stable, under a CFL condition, for both the explicit Euler method and the second order Runge-Kutta method [12, 8, 5]. Positivity under explicit time stepping is more difficult to assess due to the extended stencil of the LPS part of the stabilization. Nevertheless an upwind scheme using a similar nonlinear mechanism as the one discussed here was shown to be stable and monotonicity preserving for explicit time discretization of the transport equation in [8].

All these aspects, along with more theoretical topics such as a localised error analysis, are potential directions of future research.



(a)



(b)

FIGURE 2. Cross-sections, along the lines $y = 0.9$ (top) and $x = 0.25$ (bottom) of the solution of (2.39) for both definitions of the blending parameters. In the label, 'METHOD 1' refers to the use of the limiters given by (3.4), while 'METHOD 2' refers to the use of (3.2).

ACKNOWLEDGEMENTS

The work of GRB and FK has been partially funded by the Leverhulme Trust via the Research Project Grant No. RPG-2012-483. The authors also want to thank the anonymous referees whose very inquisitive (but fair) criticism greatly increased the quality of this paper.

References

[1] S. Badia and A. Hierro. On monotonicity-preserving stabilized finite element approximations of transport problems. *SIAM J. Sci. Comput.*, 36(6):A2673–A2697, 2014.

[2] 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.*, 2017. in press.

[3] G. R. Barrenechea, V. John, and P. Knobloch. Analysis of algebraic flux correction schemes. *SIAM J. Numer. Anal.*, 54(4):2247–2451, 2016.

[4] R. Becker, E. Burman, and P. Hansbo. A finite element time relaxation method. *C. R. Math. Acad. Sci. Paris*, 349(5-6):353–356, 2011.

[5] A. Bonito, J.-L. Guermond, and B. Popov. Stability analysis of explicit entropy viscosity methods for nonlinear scalar conservation equations. *Math. Comp.*, 83(287):1039–1062, 2014.

[6] J. P. Boris and D. L. Book. Flux-corrected transport. I. SHASTA, a fluid transport algorithm that works [J. Comput. Phys. **11** (1973), no. 1, 38–69]. *J. Comput. Phys.*, 135(2):170–186, 1997. With an introduction by Steven T. Zalesak, Commemoration of the 30th anniversary {of J. Comput. Phys.}.

[7] E. Burman. Robust error estimates in weak norms for advection dominated transport problems with rough data. *Math. Models Methods Appl. Sci.*, 24(13):2663–2684, 2014.

[8] E. Burman. A monotonicity preserving, nonlinear, finite element upwind method for the transport equation. *Appl. Math. Lett.*, 49:141–146, 2015.

[9] E. Burman and A. 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.

[10] E. Burman and A. Ern. The discrete maximum principle for stabilized finite element methods. In *Numerical mathematics and advanced applications*, pages 557–566. Springer Italia, Milan, 2003.

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

[12] E. Burman, A. Ern, and M. A. Fernández. Explicit Runge-Kutta schemes and finite elements with symmetric stabilization for first-order linear PDE systems. *SIAM J. Numer. Anal.*, 48(6):2019–2042, 2010.

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

[14] R. Codina. A discontinuity-capturing crosswind-dissipation for the finite element solution of the convection-diffusion equation. *Comput. Methods Appl. Mech. Engrg.*, 110(3-4):325–342, 1993.

[15] A. Ern and J.-L. Guermond. *Theory and practice of finite elements*, volume 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004.

[16] A. Ern and J.-L. Guermond. Weighting the edge stabilization. *SIAM J. Numer. Anal.*, 51(3):1655–1677, 2013.

[17] L. C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. AMS, Providence, Rhode Island, 1998.

[18] J. Guermond. Stabilization of Galerkin approximations of transport equations by subgrid modeling. *M2AN Math. Model. Numer. Anal.*, 33:1293–1316, 1999.

[19] J.-L. Guermond and M. Nazarov. A maximum-principle preserving C^0 finite element method for scalar conservation equations. *Comput. Methods Appl. Mech. Engrg.*, 272:198–213, 2014.

[20] J.-L. Guermond, M. Nazarov, B. Popov, and Y. Yang. A second-order maximum principle preserving Lagrange finite element technique for nonlinear scalar conservation equations. *SIAM J. Numer. Anal.*, 52(4):2163–2182, 2014.

[21] T. J. R. Hughes and M. Mallet. A new finite element formulation for computational fluid dynamics. IV. A discontinuity-capturing operator for multidimensional advective-diffusive systems. *Comput. Methods Appl. Mech. Engrg.*, 58(3):329–336, 1986.

[22] V. John and P. Knobloch. On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part I – A review. *Comput. Methods Appl. Mech. Engrg.*, 196:2197–2215, 2007.

[23] V. John and P. Knobloch. On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part II – Analysis for P_1 and Q_1 finite elements. *Comput. Methods Appl. Mech. Engrg.*, 197:1997–2014, 2008.

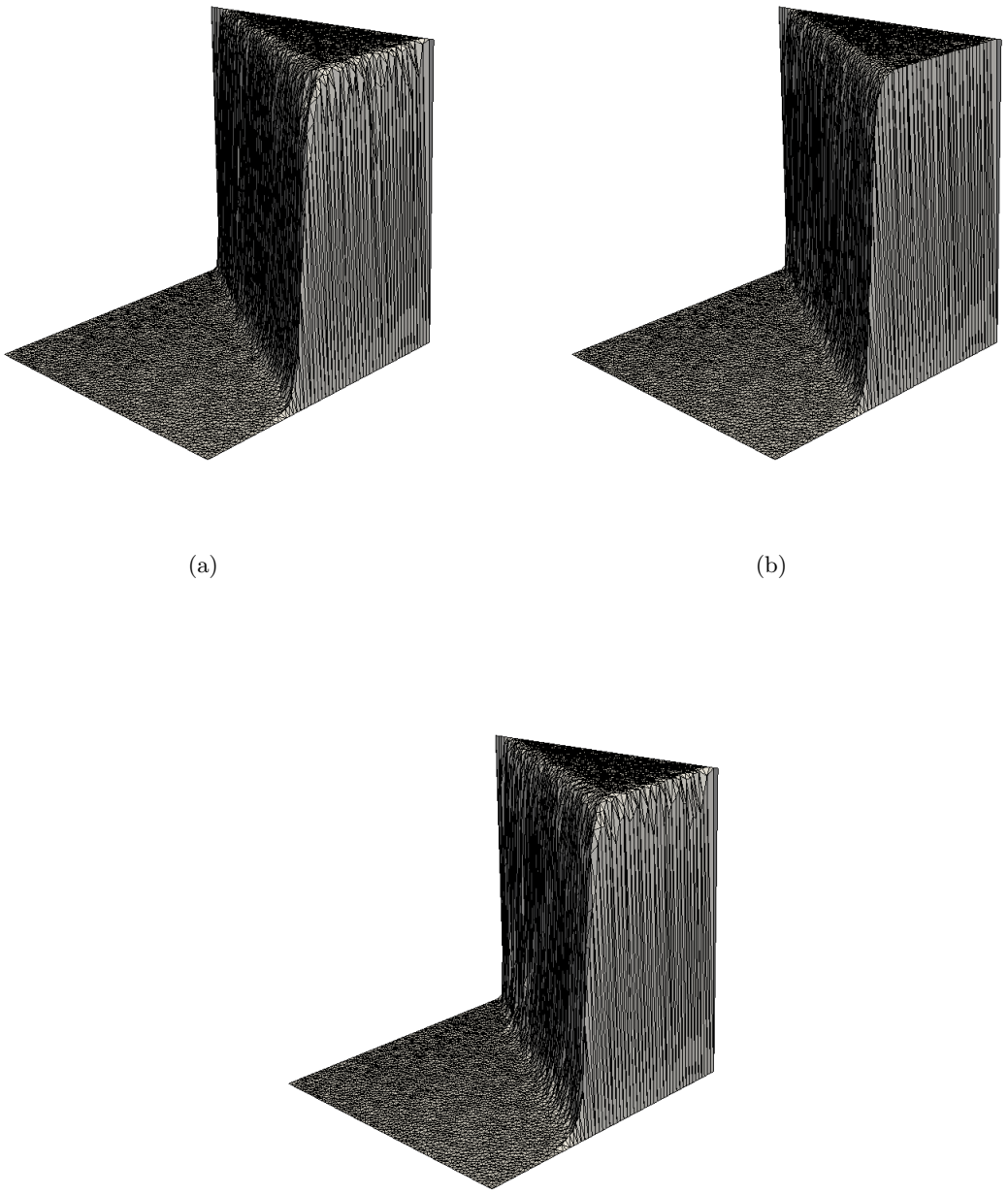
[24] D. Kuzmin. On the design of general-purpose flux limiters for finite element schemes. I. Scalar convection. *J. Comput. Phys.*, 219(2):513–531, 2006.

[25] D. 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.
[26] D. Kuzmin, S. Basting, and J. Shadid. Monotone local projection stabilization schemes for continuous finite elements. 2016. Preprint.
[27] D. Kuzmin and M. Möller. Algebraic flux correction. I. Scalar conservation laws. In *Flux-corrected transport*, Sci. Comput., pages 155–206. Springer, Berlin, 2005.
[28] D. Kuzmin, M. Möller, and S. Turek. High-resolution FEM-FCT schemes for multidimensional conservation laws. *Comput. Methods Appl. Mech. Engrg.*, 193(45-47):4915–4946, 2004.
[29] D. Kuzmin and J. N. Shadid. A new approach to enforcing discrete maximum principles in continuous Galerkin methods for convection-dominated transport equations. Technical report, UA Ruhr Zentrum für partielle Differentialgleichungen, 2015.
[30] D. Kuzmin and S. Turek. Flux correction tools for finite elements. *J. Comput. Phys.*, 175(2):525–558, 2002.
[31] A. Logg, K.-A. Mardal, and G. N. Wells, editors. *Automated solution of differential equations by the finite element method*, volume 84 of *Lecture Notes in Computational Science and Engineering*. Springer, Heidelberg, 2012. The FEniCS book.
[32] R. Löhner, K. Morgan, J. Peraire, and M. Vahdati. Finite element flux-corrected transport (FEM–FCT) for the Euler and Navier–Stokes equations. *Int. J. Numer. Meths. Fluids*, 7(10):1093–1109, 1987.
[33] R. Löhner, K. Morgan, M. Vahdati, J. P. Boris, and D. L. Book. FEM-FCT: combining unstructured grids with high resolution. *Comm. Appl. Numer. Methods*, 4(6):717–729, 1988.
[34] P. R. M. Lyra and K. Morgan. A review and comparative study of upwind biased schemes for compressible flow computation. III. *Arch. Comput. Methods Engrg.*, 9(3):207–256, 2002.
[35] 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(2):181–193, 1985.
[36] H.-G. Roos, M. Stynes, and L. Tobiska. *Robust Numerical Methods for Singularly Perturbed Differential Equations. Convection–Diffusion–Reaction and Flow Problems. 2nd ed.* Springer-Verlag, Berlin, 2008.
[37] R. Temam. *Navier-Stokes equations. Theory and numerical analysis*. North-Holland Publishing Co., Amsterdam-New York-Oxford, 1977. Studies in Mathematics and its Applications, Vol. 2.
[38] S. Turek and D. Kuzmin. Algebraic flux correction. III. Incompressible flow problems. In *Flux-corrected transport*, Sci. Comput., pages 251–296. Springer, Berlin, 2005.
[39] J. Xu and L. Zikatanov. A monotone finite element scheme for convection-diffusion equations. *Math. Comp*, 68:1429–1446, 1999.
[40] S. T. Zalesak. Fully multidimensional flux-corrected transport algorithms for fluids. *J. Comput. Phys.*, 31(3):335–362, 1979.

(G.R.B.) Department of Mathematics and Statistics, University of Strathclyde, 26 Richmond Street, Glasgow G1 1XH, Scotland. gabriel.barrenechea@strath.ac.uk

(E.B.) Department of Mathematics, University College London, Gower Street, London WC1E 6BT, UK. e.burman@ucl.ac.uk

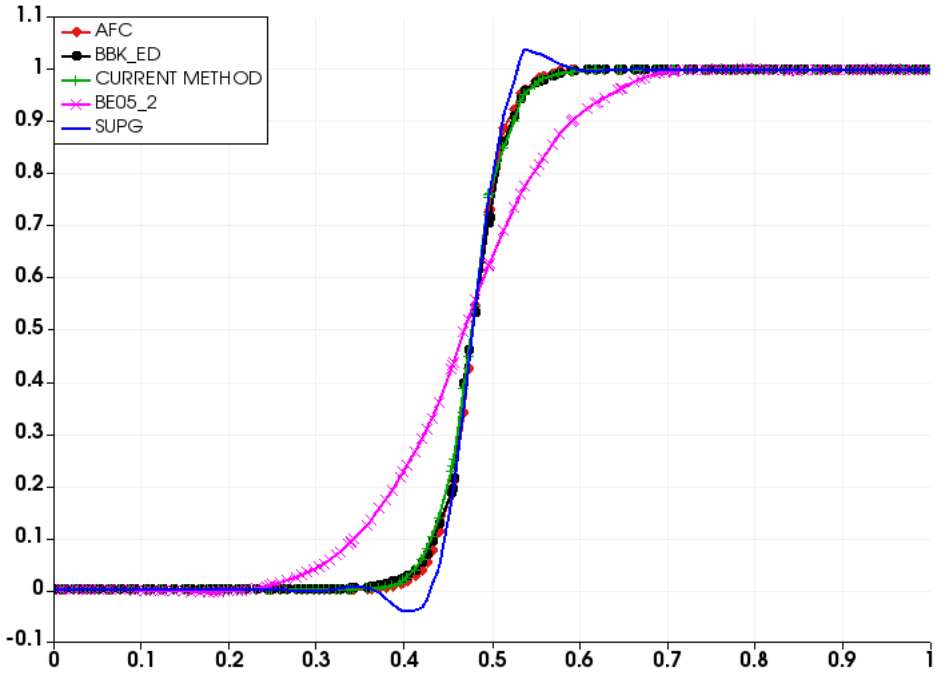
(F.K.) Department of Mathematics, University of Chester, Chester, UK. f.karakatsani@chester.ac.uk



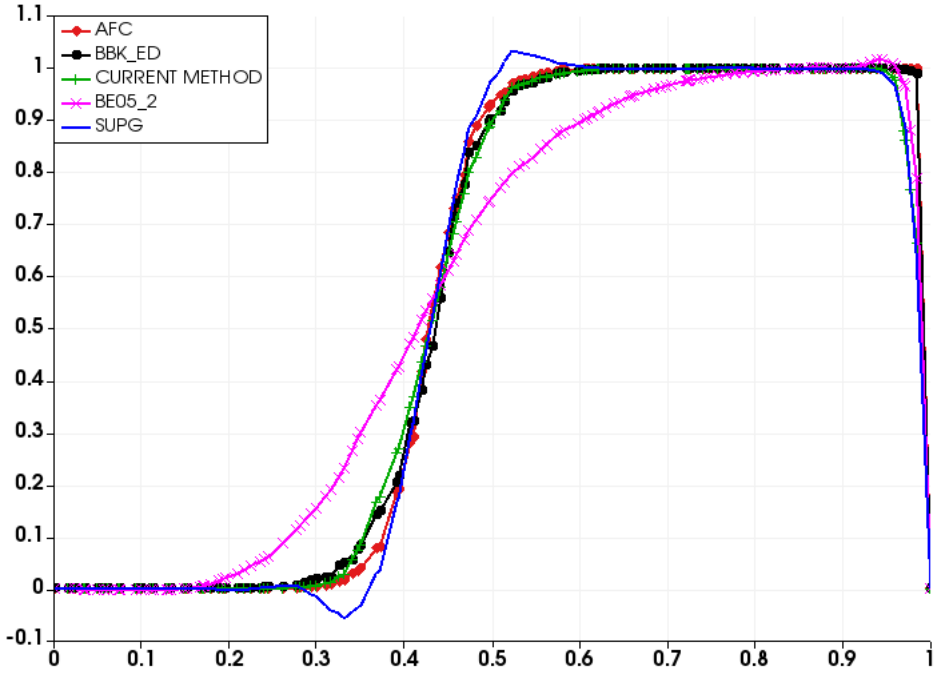
(a) (b)

(c)

FIGURE 3. Solution for Example 1 on the unstructured mesh with 7970 elements obtained by the nonlinear LPS method (2.39) using $p = 10$ and $c_0 = 0.3$ (top left, 234 iterations were needed to reach convergence). The results for the AFC scheme are depicted in the top right (where 77 iterations were needed for convergence). At the bottom, we depict the results obtained the method proposed in [2] with $p = 10$ (where 426 iterations were needed to reach convergence).

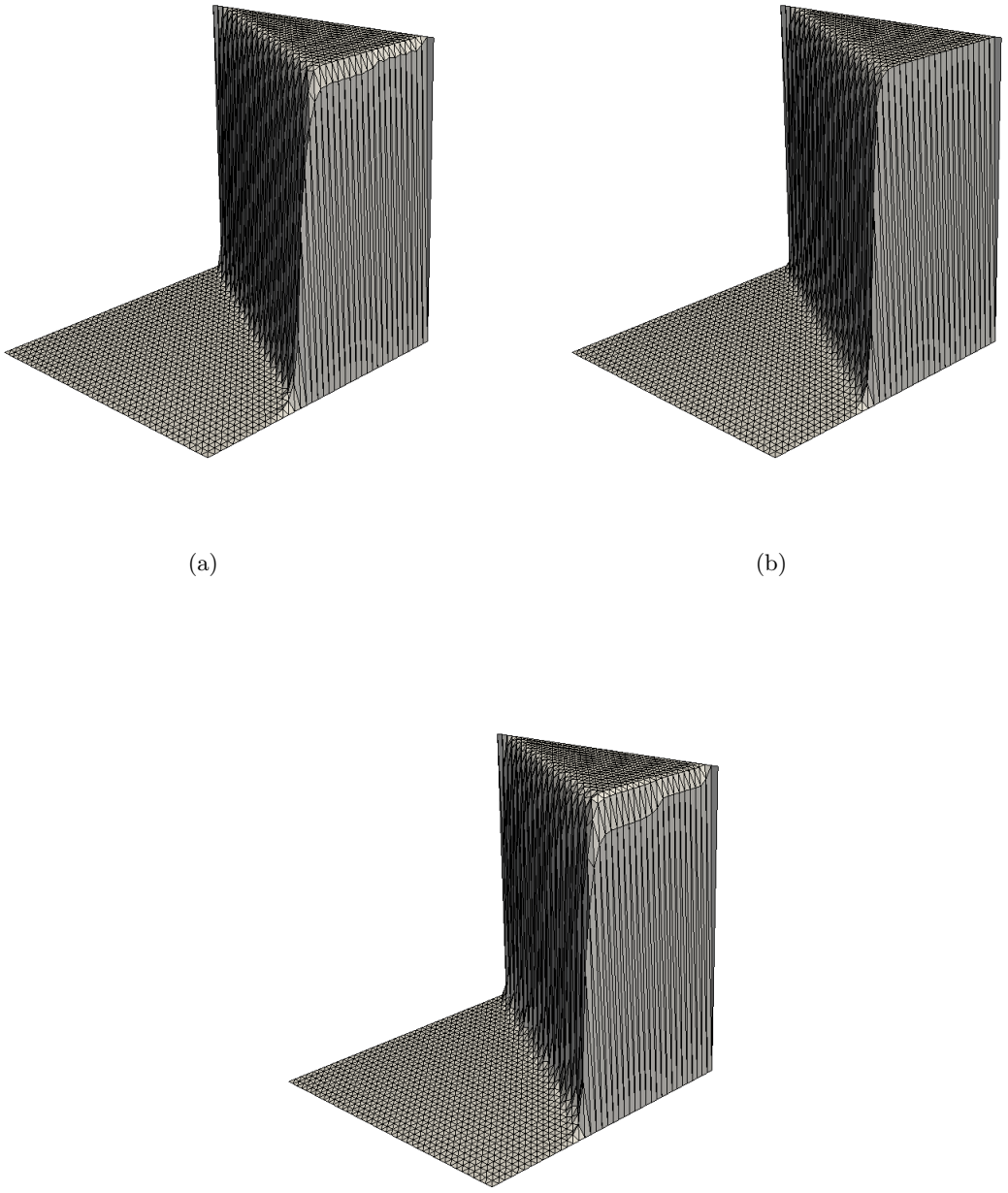


(a)



(b)

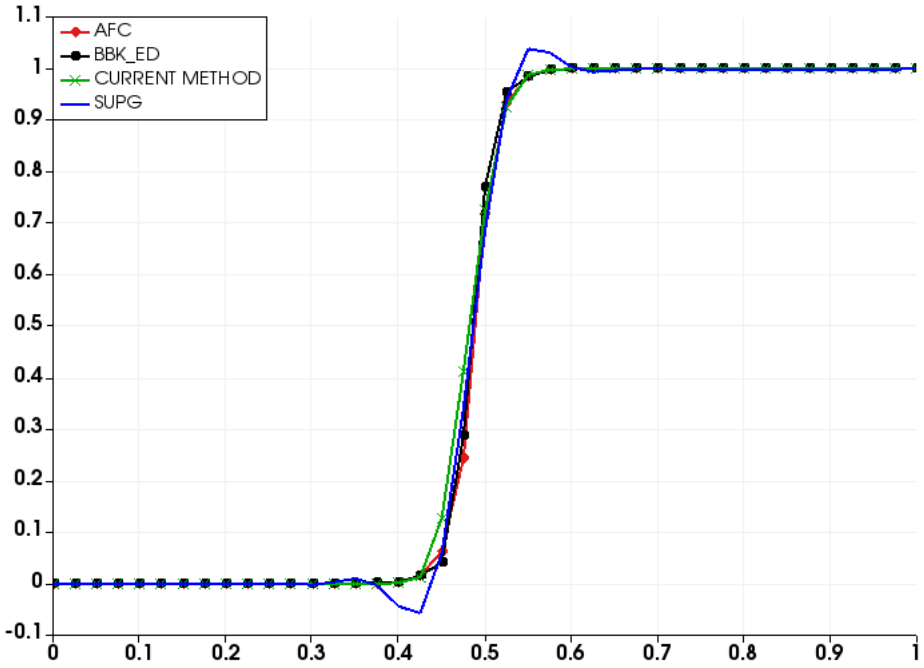
FIGURE 4. Cross-sections along the lines $y = 0.9$ (top) and $x = 0.25$ (bottom) for Example 1. We depict the results of method (2.39) ('CURRENT METHOD' in the label), the AFC scheme ('AFC' in the label), the method proposed in [2] ('BBK_ED' in the label), the method analysed in [11] ('BE05_2' in the label, using $\gamma_{cip} = 0.3$, $c_\rho = 0.9$), and the SUPG method.



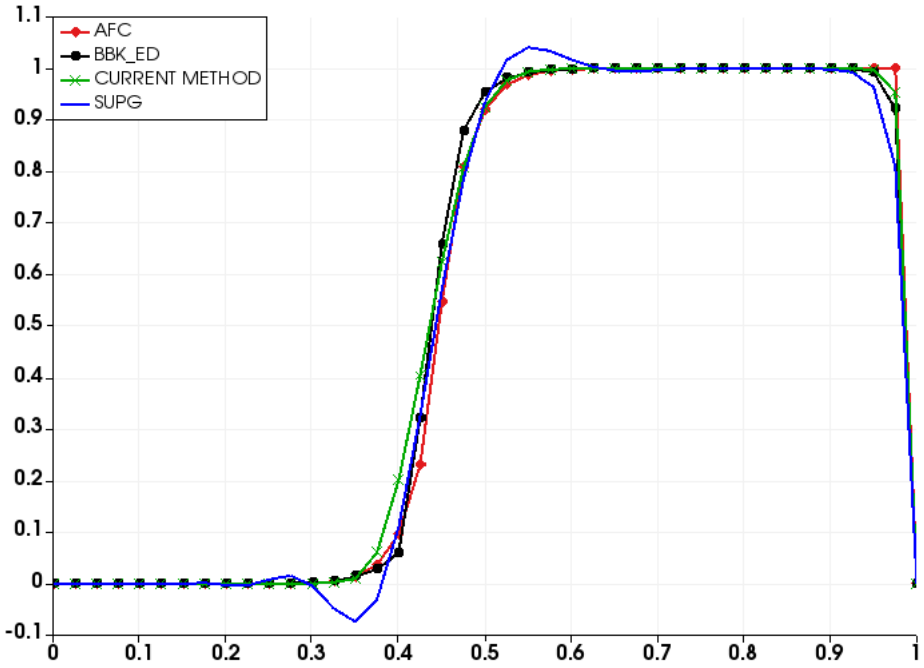
(a) (b)

(c)

FIGURE 5. Solution for Example 1 on the structured mesh with $2 \times 40 \times 40$ elements obtained by the nonlinear LPS method (2.39) using $p = 15$ and $c_0 = 0.25$ (top left, 252 iterations were needed to reach convergence). The results for the AFC scheme are depicted in the top right (where 1191 iterations were needed for convergence). At the bottom, we depict the results obtained the method proposed in [2] with $p = 15$ (where 148 iterations were needed to reach convergence).

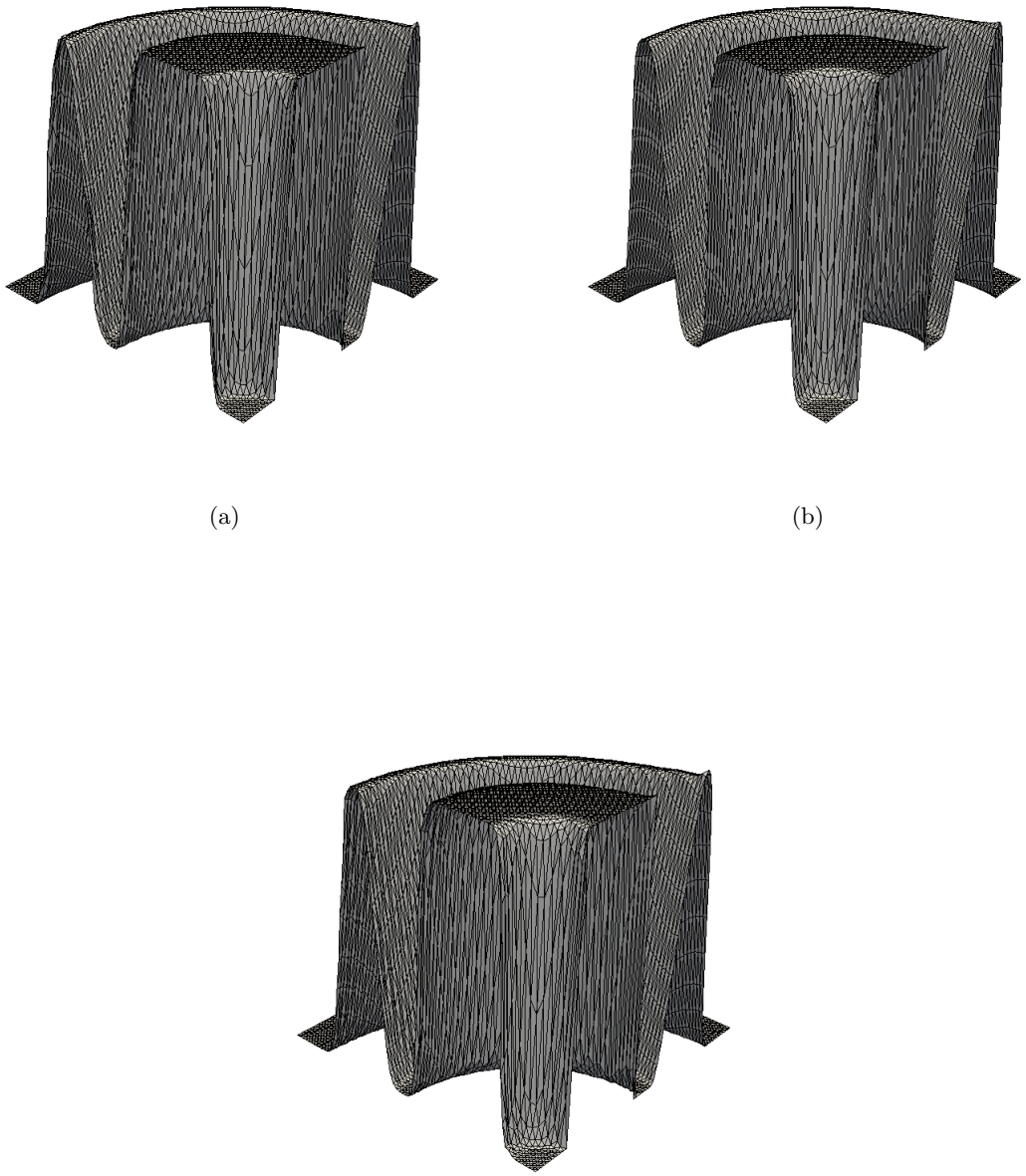


(a)



(b)

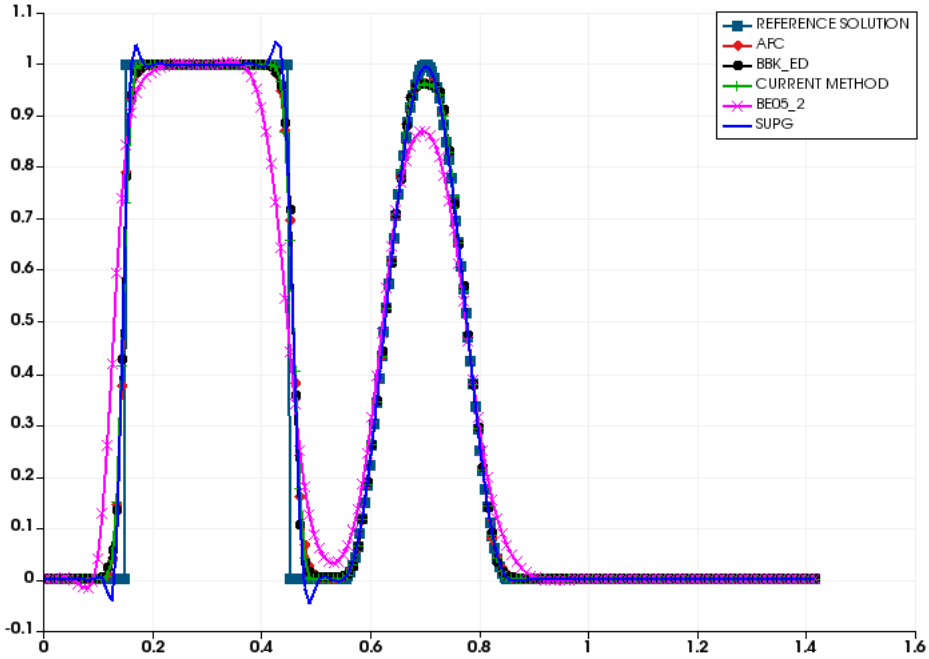
FIGURE 6. Cross-sections along the lines $y = 0.9$ (top) and $x = 0.25$ (bottom) for Example 1. We depict the results of method (2.39) ('CURRENT METHOD' in the label), the AFC scheme ('AFC' in the label), the method proposed in [2] ('BBK_ED' in the label), and the SUPG method.



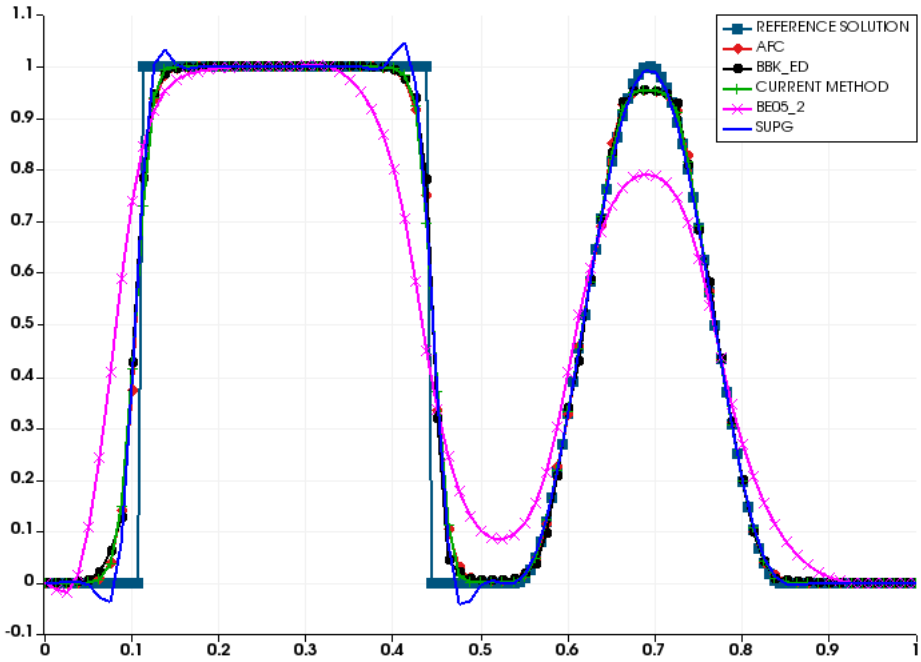
(a) (b)

(c)

FIGURE 7. Solution for Example 2 on the structured mesh using $2 \times 80 \times 80$ elements. We depict the solutions for the nonlinear LPS method (2.39) using $p = 10$ and $c_0 = 0.2$ (top left, 110 iterations were needed to reach convergence). The results for the AFC scheme are depicted in the top right (where 162 iterations were needed for convergence). At the bottom, results obtained using the edge diffusion method from [2] with $p = 10$, where 60 iterations are needed for convergence.

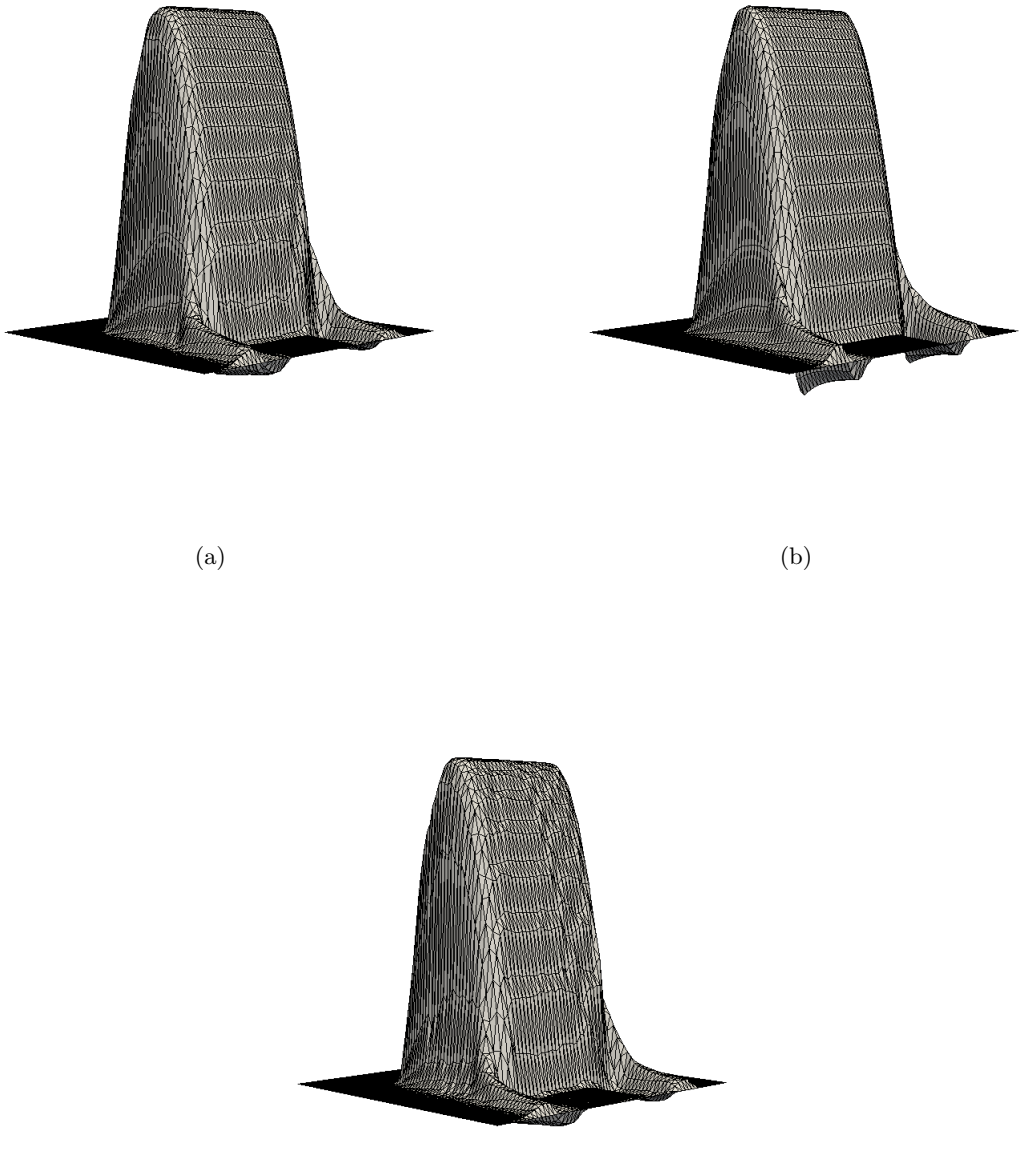


(a)



(b)

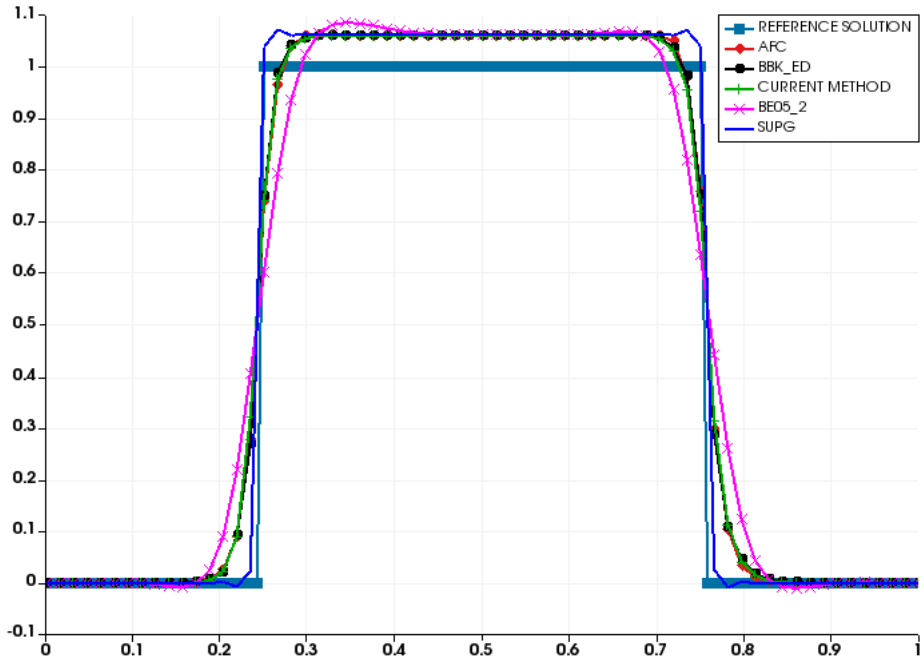
FIGURE 8. Cross-sections along the lines $y = x$ (top) and $x = 0.1$ (bottom) for Example 2. We depict the results of method (2.39) ('CURRENT METHOD' in the label), the AFC scheme ('AFC' in the label), the method proposed in [2] ('BBK_ED' in the label), the method analysed in [11] ('BE05_2' in the label, using $\gamma_{cip} = 0.1$, $c_\rho = 0.5$), and the SUPG method. Finally, 'REFERENCE SOLUTION' refers to the solution of the transport problem.



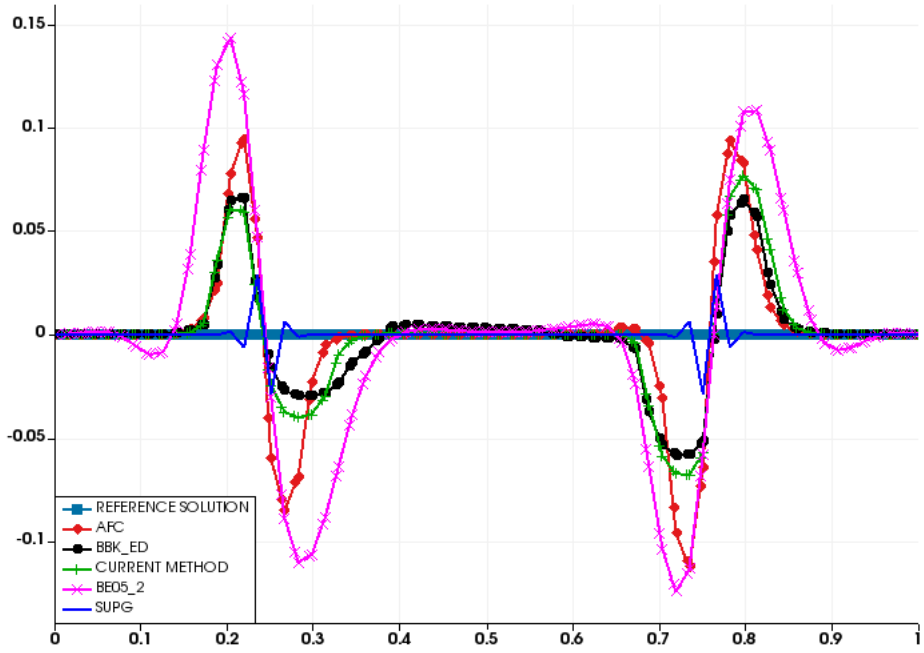
(a) (b)

(c)

FIGURE 9. Solution for Example 3 on the structured structured mesh using $2 \times 64 \times 64$ elements. We depict the results of the nonlinear LPS method (2.39) using $p = 8$ and $c_0 = 0.4$ (top left, 590 iterations were needed to reach convergence). The results for the AFC scheme are depicted in the top right (71 iterations were needed for convergence). At the bottom, results obtained using the method from [2] with $p = 8$, where 945 iterations are needed for convergence.



(a)



(b)

FIGURE 10. Cross-sections along the lines $x = 0.5$ (top) and $x = 0.8$ (bottom) for Example 3. We depict the results of method (2.39) ('CURRENT METHOD' in the label), the AFC scheme ('AFC' in the label), the method proposed in [2] ('BBK_ED' in the label), the method analysed in [11] ('BE05_2' in the label, using $\gamma_{cip} = 0.2$, $c_\rho = 0.4$), and the SUPG method. Here, 'REFERENCE SOLUTION' refers to the parabolic profile which is very close to the exact solution of the problem.