An adaptive stabilized finite element method based on residual minimization

Victor M Calo, Alexandre Ern, Ignacio Muga, Sergio Rojas

An adaptive stabilized finite element method based on residual minimization

Victor M. Calo[1,2], Alexandre Ern[3,4], Ignacio Muga[5], and Sergio Rojas[1]

[1]School of Earth and Planetary Sciences, Curtin University, Kent Street, Bentley, Perth, WA 6102, Australia
[2]Mineral Resources, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Kensington, Perth, WA 6152, Australia
[3]Université Paris-Est, CERMICS (ENPC), 6-8 avenue Blaise Pascal, 77455 Marne la Vallée cedex 2, France
[4]INRIA Paris, 75589 Paris, France
[5]Instituto de Matemáticas, Pontificia Universidad Católica de Valparaíso, Casilla 4059, Valparaíso, Chile

July 27, 2019

Abstract

We devise and analyze a new adaptive stabilized finite element method. We illustrate its performance on the advection-reaction model problem. We construct a discrete approximation of the solution in a continuous trial space by minimizing the residual measured in a dual norm of a discontinuous test space that has inf-sup stability. We formulate this residual minimization as a stable saddle-point problem which delivers a stabilized discrete solution and an error representation that drives the adaptive mesh refinement. Numerical results on the advection-reaction model problem show competitive error reduction rates when compared to discontinuous Galerkin methods on uniformly refined meshes and smooth solutions. Moreover, the technique leads to optimal decay rates for adaptive mesh refinement and solutions having sharp layers.

1 Introduction

Continuous Galerkin Finite Element Methods (CG-FEM) are popular solution strategies in many engineering applications. However, these methods can suffer from instability under reasonable physical assumptions if the mesh is not sufficiently refined. This limitation led to the development of several alternative methods that achieve stability differently (see, e.g., [25, 31] and references therein). Broadly speaking, and referring to [27] for a more detailed overview, there are two approaches

to enhance stability in CG-FEM: residual and fluctuation-based stabilization techniques. Residual-based stabilization is exemplified by the Least-Squares Finite Element Method (LS-FEM) which was pioneered in the late 60's and early 70's (cf., [24, 37, 7] and [32] for a more recent overview). The suboptimal convergence rate of LS-FEM in the L^2-norm was improved by the Galerkin/Least-Squares (GaLS) method [30]. The extension of the Least-Squares approach to more general *dual norms* was studied in [18] where the authors make the connection between residual minimization in test dual norms and the mixed formulation (cf., Equation (27)). In this context, the use of different test norms to stabilize convection-diffusion problems was further investigated in [14]. Interestingly, the idea of residual minimization in non-standard dual norms is also at the heart of the recent Discontinuous Petrov–Galerkin (DPG) methods (see, e.g., [40, 20, 15], and [19] for a general overview). Alternatively, several strategies based on fluctuation stabilization exist. Some methods penalize the gradient of some fluctuation or some fluctuation of the gradient as in [29, 6, 38]. Other methods penalize the gradient jumps across the mesh interfaces as in [9, 11]. Yet, other techniques enlarge the trial and test spaces with discontinuous functions and penalize the solution jumps across the mesh interfaces, as in the discontinuous Galerkin (DG) methods [39, 36, 33, 17, 8, 26] (see also the textbook [22] for a recent overview).

In this work, we introduce a new adaptive stabilized FEM and study its numerical performance when approximating advection-reaction problems. We combine the residual minimization idea with the inf-sup stability offered by a large class of DG methods. More precisely, the discrete solution we seek is the minimizer in a continuous trial space (e.g., H^1-conforming finite elements) of the residual measured in a dual norm with respect to DG test functions. This DG norm provides inf-sup stability to the formulation. In practice, such a residual minimization implies a stable saddle-point problem involving the continuous trial space and the discontinuous test space. The price paid for solving the larger linear system (compared to just forming the normal equations as in LS-FEM) is compensated by two advantages. Firstly, the present solution is more accurate, especially in the L^2-norm, than the LS-FEM solution. Actually, we prove that the present method leads to error estimates of the same quality as those delivered by DG methods, yet for a discrete solution belonging to a continuous trial space only. The second advantage is that in the present approach, the component in the DG space plays the role of an error representative that can be used to drive adaptive mesh refinement. Thus, we compute on the fly a discrete solution in the continuous trial space and an error representation in the discontinuous test space. From the practical point of view, we illustrate numerically the benefits of the adaptive mesh refinement procedure in the context of the approximation of the advection-reaction model problem with solutions possessing sharp inner layers.

We now put our work in perspective with respect to the state-of-the-art literature. Compared with the LS-FEM paradigm, we solve a larger linear system, but the advantage is that we obtain an error representation that guides the mesh adaptation and that the error decay rates with respect to the number of degrees of freedom are better. Compared to the recent developments on DPG methods, we share the goal of minimizing the residual with respect to an adequate norm, using broken test spaces,

implying discrete stability and an error representation to guide adaptivity. Nevertheless, the main difference is that DPG methods are constructed considering conforming formulations and localizable test norms. In particular, for first-order PDEs and if the trial space is continuous, the norm equipping the test space must be the L^2-norm. Stronger norms can be applied using ultra-weak formulations which only require L^2 regularity for the trial space. This in practice leads to a discontinuous approximation of the trial solution, requiring additional unknowns to represent traces of the solution over the mesh skeleton, and therefore increasing the total number of degrees of freedom in the system (however, the broken test space structure with localizable norm allows for practical static condensation procedures, which lead to linear cost solvers when combined with multigrid techniques). Here, the trial space is continuous, but the test space is equipped with a stronger norm than L^2, and the stability of the method is inherited from the DG formulation that provides the inner product and norm on top of which we build our method.

We organize the paper as follows. In Section 2, we recall some basic facts concerning the continuous and discrete settings. In particular, we present the abstract framework from [26] for the error analysis of nonconforming approximation methods (see also [22] and Strang's Second Lemma). We devise and analyze the present stabilized FEM in Section 3, where our main results are Theorems 2 and 3. In Section 4, we briefly describe the mathematical setting for the advection-reaction model problem and recall its DG discretization using both centered and upwind fluxes. In Section 5, we present numerical results illustrating the performance of our adaptive stabilized FEM. Finally, we draw conclusions in Section 6.

2 Continuous and discrete settings

2.1 Well-posed model problem

Let X (continuous trial space) and Y (continuous test space) be (real) Banach spaces equipped with norms $\|\cdot\|_X$ and $\|\cdot\|_Y$, respectively. Assume that Y is reflexive. We want to solve the following linear model problem:

$$\begin{cases} \text{Find } u \in X, \text{ such that:} \\ b(u,v) = \langle l, v\rangle_{Y^*,Y}, \quad \forall\, v \in Y, \end{cases} \tag{1}$$

where b is a bounded bilinear form on $X \times Y$, $l \in Y^*$ is a bounded linear form on Y, and $\langle\cdot,\cdot\rangle_{Y^*,Y}$ is the duality pairing in $Y^* \times Y$. Equivalently, problem (1) can be written in operator form by introducing the operator $B : X \to Y^*$ such that:

$$\langle B\,z\,,\, v\rangle_{Y^*,Y} := b(z,v), \quad \forall\,(z,v) \in X \times Y, \tag{2}$$

leading to the following problem:

$$\begin{cases} \text{Find } u \in X, \text{ such that:} \\ Bu = l \text{ in } Y^*. \end{cases} \tag{3}$$

We assume that there exists a constant $C_b > 0$ such that

$$\inf_{0\neq z\in X} \sup_{0\neq y\in Y} \frac{|b(z,y)|}{\|z\|_X \|y\|_Y} \geq C_b, \tag{4}$$

and that $\{y \in Y : b(z,y) = 0, \forall\, z \in X\} = \{0\}$. These two assumptions are equivalent to problem (1) (or equivalently (3)) being well-posed owing to the Banach–Nečas–Babuška theorem (see, e.g., [25, Theorem 2.6]). Moreover, the following a priori estimate is then satisfied:

$$\|u\|_X \leq C_b^{-1} \|l\|_{Y^*}. \tag{5}$$

Finally, we mention that when $Y = X$, a sufficient condition for well-posedness is that the bilinear form b is coercive, that is, there exists a constant $C_b > 0$ such that

$$b(v,v) \geq C_b \|v\|_X^2, \quad \forall\, v \in X. \tag{6}$$

Then problem (1) (or equivalently (3)) is well-posed owing to the the Lax–Milgram lemma (see, e.g., [35]), and the same a priori estimate (5) is satisfied.

2.2 Functional setting

For any open and bounded set $\mathscr{D} \subset \mathbb{R}^d$, let $L^2(\mathscr{D})$ be the standard Hilbert space of square-integrable functions over $\mathscr{D}$ for the Lebesgue measure, and denote by $(\cdot,\cdot)_{\mathscr{D}}$ and $\|\cdot\|_{\mathscr{D}} := \sqrt{(\cdot,\cdot)_{\mathscr{D}}}$ its inner product and inherited norm, respectively. Denote by $L^2(\mathscr{D};\mathbb{R}^d)$ the corresponding space composed of square-integrable vector-valued functions and, abusing the notation, still by $(\cdot,\cdot)_{\mathscr{D}}$ the associated inner product. Considering weak derivatives, we recall the following well-known Hilbert space:

$$H^1(\mathscr{D}) := \left\{ v \in L^2(\mathscr{D}) : \nabla v \in L^2(\mathscr{D};\mathbb{R}^d) \right\}, \tag{7}$$

equipped with the following inner product and norm (respectively):

$$(v,u)_{1,\mathscr{D}} := (v,u)_{\mathscr{D}} + (\nabla v, \nabla u)_{\mathscr{D}}, \qquad \|v\|_{1,\mathscr{D}} := \sqrt{(v,v)_{1,\mathscr{D}}}, \tag{8}$$

Let $H^{1/2}(\partial\mathscr{D})$ be the standard Dirichlet trace space of $H^1(\mathscr{D})$ over the boundary $\partial\mathscr{D}$, and $H^{-1/2}(\partial\mathscr{D})$ be its dual space with $L^2(\partial\mathscr{D})$ as pivot space. More generally, to warrant that traces are well defined at least in $L^2(\partial\mathscr{D})$, we will call upon the fractional-order Sobolev spaces $H^s(\mathscr{D})$, with $s > \frac{1}{2}$.

2.3 Discrete setting

Let $\mathscr{P}_h = \{K_m\}_{m=1}^N$ be a conforming partition of the domain $\mathscr{D}$ into N open disjoint elements K_m, such that

$$\mathscr{D}_h := \bigcup_{m=1}^{N} K_m \quad \text{satisfies} \quad \mathscr{D} = \operatorname{int}\left(\overline{\mathscr{D}_h}\right). \tag{9}$$

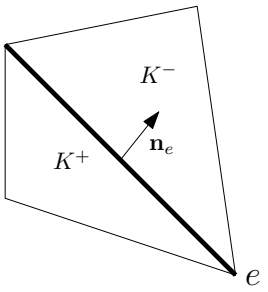


Figure 1: Skeleton orientation over the internal face $e = \partial K^+ \cap \partial K^-$.

We denote by ∂K_m the boundary of the element K_m, by $\mathscr{S}_h^0$, $\mathscr{S}_h^\partial$ the set of interior and boundary edges/faces (respectively) of the mesh, and by $\mathscr{S}_h := \mathscr{S}_h^0 \cup \mathscr{S}_h^\partial$ the skeleton of $\mathscr{D}_h$. Over $\mathscr{D}_h$, we define the following standard broken Hilbert space:

$$H^1(\mathscr{D}_h) := \left\{ v \in L^2(\mathscr{D}) : \nabla v|_{K_m} \in L^2(K_m;\mathbb{R}^d),\ \forall\, m = 1,\dots,N \right\}, \tag{10}$$

with inner product defined as

$$(v,u)_{1,\mathscr{D}_h} := \sum_{m=1}^{N} (v,u)_{1,K_m}. \tag{11}$$

For any $v \in H^1(\mathscr{D}_h)$ and any interior face/edge $e = \partial K^+ \cap \partial K^- \in \mathscr{S}_h^0$ (see Figure 1), we define the jump $[\![v]\!]_e$ and the average $\{\!\{v\}\!\}_e$ of a smooth enough function v as follows:

$$[\![v]\!]_e(x) := v^+(x) - v^-(x), \tag{12}$$

$$\{\!\{v\}\!\}_e(x) := \frac{1}{2}(v^+(x) + v^-(x)), \tag{13}$$

for a.e. $x \in e$, where v^+ and v^- denote the traces over e defined from a predefined orientation which is fixed by the unit normal vector $\mathbf{n}_e$. The subscript e is omitted from the jump and average operators when there is no ambiguity.

Finally, we denote by $\mathbb{P}^p(K_m)$, $p \geq 0$, the set of polynomials of total degree at most p defined on the element K_m, and we consider the following broken polynomial space:

$$\mathbb{P}^p(\mathscr{D}_h) := \left\{ v \in L^2(\mathscr{D}) : v|_{K_m} \in \mathbb{P}^p(K_m),\ \forall\, m = 1,\dots,N \right\}. \tag{14}$$

2.4 Abstract setting for nonconforming approximation methods

Let V_h be a finite-dimensional space composed of functions defined on $\mathscr{D}_h$ (typically a broken polynomial space). We approximate the continuous problem (1) as follows:

$$\begin{cases} \text{Find } \theta_h \in V_h, \text{ such that:} \\ b_h(\theta_h, v_h) = \langle l_h, v_h \rangle_{V_h^* \times V_h}, \quad \forall\, v_h \in V_h, \end{cases} \tag{15}$$

where $b_h(\cdot,\cdot)$ denotes a discrete bilinear form defined over $V_h \times V_h$, and $l_h(\cdot)$ a discrete linear form over V_h. We say that the approximation setting is nonconforming whenever $V_h \not\subset X$ or $V_h \not\subset Y$.

To ascertain that the discrete problem (15) is well posed and to perform the error analysis, we follow the framework introduced for DG approximations in [26] (see also [22] and Strang's Second Lemma) which relies on the following three assumptions:

Assumption 1. *(Stability): The space V_h can be equipped with a norm $\|\cdot\|_{V_h}$ such that there exists a constant $C_{\mathrm{sta}} > 0$, uniformly with respect to the mesh size, such that:*

$$\inf_{0\neq z_h\in V_h} \sup_{0\neq v_h\in V_h} \frac{b_h(z_h, v_h)}{\|z_h\|_{V_h}\|v_h\|_{V_h}} \geq C_{\mathrm{sta}}. \tag{16}$$

Assumption 2. *(Strong consistency with regularity) The exact solution u of* (1) *belongs to a subspace $X_\# \subset X$ such that the discrete bilinear form $b_h(\cdot,\cdot)$ supports evaluations in the extended space $V_{h,\#} \times V_h$ with $V_{h,\#} := X_\# + V_h$, and the following holds true:*

$$b_h(u, v_h) = \langle l_h, v_h\rangle_{V_h^*\times V_h}, \quad \forall\, v_h \in V_h, \tag{17}$$

which amounts to $b_h(u-\theta_h, v_h) = 0$, for all $v_h \in V_h$ (Galerkin's orthogonality).

Assumption 3. *(Boundedness): The stability norm $\|\cdot\|_{V_h}$ can be extended to $V_{h,\#}$ and there is a second norm $\|\cdot\|_{V_h,\#}$ on $V_{h,\#}$ satisfying the following two properties:*

(i) $\|v\|_{V_h} \leq \|v\|_{V_h,\#}$, *for all $v \in V_{h,\#}$;*

(ii) there exists a constant $C_{\mathrm{bnd}} < \infty$, uniformly with respect to the mesh size, such that:

$$b_h(z, v_h) \leq C_{\mathrm{bnd}}\|z\|_{V_h,\#}\|v_h\|_{V_h}, \quad \forall\,(z, v_h) \in V_{h,\#} \times V_h. \tag{18}$$

For a linear form $\phi_h : V_h \to \mathbb{R}$, we set

$$\|\phi_h\|_{V_h^*} := \sup_{0\neq v_h\in V_h} \frac{\langle \phi_h, v_h\rangle_{V_h^*\times V_h}}{\|v_h\|_{V_h}}, \tag{19}$$

where $\|\cdot\|_{V_h}$ is the stability norm identified in Assumption 1. The above assumptions lead to the following a priori and error estimates (see [26, 22]).

Theorem 1 (A priori and error estimates)**.** *Denote by u the solution of the continuous problem* (1) *and suppose that the assumptions 1–3 are satisfied. Then there exists a unique $\theta_h \in V_h$ solution to the discrete problem* (15)*, and the following two estimates are satisfied:*

$$\|\theta_h\|_{V_h} \leq \frac{1}{C_{\mathrm{sta}}}\|l_h\|_{V_h^*}, \tag{20}$$

and

$$\|u-\theta_h\|_{V_h} \leq \left(1 + \frac{C_{\mathrm{bnd}}}{C_{\mathrm{sta}}}\right) \inf_{v_h\in V_h} \|u - v_h\|_{V_h,\#}. \tag{21}$$

3 Residual minimization problem

We obtain the schemes we propose from the following two ingredients:

a) First, we select a finite-dimensional space V_h and a discrete bilinear form $b_h(\cdot,\cdot)$ such that Assumptions 1, 2, and 3 hold true.

b) Second, we identify a subspace $U_h \subset V_h$ such that U_h has the same approximation capacity as the original space V_h for the types of solutions we want to approximate. The main example we have in mind is to choose V_h as a broken polynomial space and U_h as the H^1-conforming subspace. We refer the reader to Remark 5 for a further discussion on the appoximation capacity of the spaces U_h and V_h in the context of advection-reaction equations.

Starting from a stable formulation of the form (15) in V_h, we use the trial subspace $U_h \subset V_h$ to solve the following residual minimization problem:

$$\left\{\begin{array}{l} \text{Find } u_h \in U_h \subset V_h, \text{ such that:} \\ u_h = \underset{z_h \in U_h}{\arg\min} \dfrac{1}{2}\| l_h - B_h z_h\|^2_{V_h^*} = \underset{z_h \in U_h}{\arg\min} \dfrac{1}{2}\| R^{-1}_{V_h}(l_h - B_h z_h)\|^2_{V_h}, \end{array}\right. \tag{22}$$

where $B_h : V_{h,\#} \to V_h^*$ is defined as:

$$\langle B_h z, v_h\rangle_{V_h^* \times V_h} := b_h(z, v_h), \tag{23}$$

and $R^{-1}_{V_h}$ denotes the inverse of the Riesz map:

$$\begin{array}{rcl} R_{V_h} & : & V_h \to V_h^* \\ & & \langle R_{V_h} y_h, v_h\rangle_{V_h^* \times V_h} := (y_h, v_h)_{V_h}, \quad \forall\ v_h \in V_h. \end{array} \tag{24}$$

The second equality in (22) follows from the fact that the Riesz map is an isometric isomorphism. Classically, the minimizer in (22) is a critical point of the minimizing functional, which translates into the following linear problem:

$$\left\{\begin{array}{l} \text{Find } u_h \in U_h, \text{ such that:} \\ (R^{-1}_{V_h}(l_h - B_h u_h), R^{-1}_{V_h} B_h \delta u_h)_{V_h} = 0, \quad \forall\, \delta u_h \in U_h. \end{array}\right. \tag{25}$$

Defining the residual representation function as

$$\varepsilon_h := R^{-1}_{V_h}(l_h - B_h u_h) \in V_h, \tag{26}$$

problem (25) is equivalent to finding the pair $(\varepsilon_h, u_h) \in V_h \times U_h$, such that:

$$\left\{\begin{array}{lll} (\varepsilon_h\,,\, v_h)_{V_h} + b_h(u_h\,,\, v_h) = l_h(v_h), & \forall\, v_h \in V_h, & \text{(27a)} \\ b_h(\varepsilon_h\,,\, z_h) = 0, & \forall\, z_h \in U_h. & \text{(27b)} \end{array}\right.$$

Conversely, if the pair $(\varepsilon_h, u_h) \in V_h \times U_h$ solves (27), then (26) holds true and u_h is the minimizer of the quadratic functional in (22).

Theorem 2 (A priori bounds and error estimates)**.** *If the assumptions 1-3 are satisfied, then the saddle-point problem* (27) *has one and only one solution* $(\varepsilon_h, u_h) \in V_h \times U_h$*. Moreover, such a solution satisfies the following a priori bounds:*

$$\|\varepsilon_h\| \leq \|l_h\|_{V_h^*} \qquad \textit{and} \qquad \|u_h\|_{V_h} \leq \frac{1}{C_{\text{sta}}} \|l_h\|_{V_h^*}, \tag{28}$$

and the following a priori error estimate holds true:

$$\|u - u_h\|_{V_h} \leq \left(1 + \frac{C_{\text{bnd}}}{C_{\text{sta}}}\right) \inf_{z_h \in U_h} \|u - z_h\|_{V_{h,\#}}, \tag{29}$$

recalling that $u \in X_\#$ *is the exact solution to the continuous problem* (1)*.*

Proof. See A. □

Assumption 4. *(Saturation) Let* $u_h \in U_h$ *be the second component of the pair* $(\varepsilon_h, u_h) \in V_h \times U_h$ *solving the saddle-point problem* (27)*. Let* $\theta_h \in V_h$ *be the unique solution to* (15)*. There exists a real number* $\delta \in [0,1)$*, uniform with respect to the mesh size, such that* $\|u - \theta_h\|_{V_h} \leq \delta \|u - u_h\|_{V_h}$*.*

Theorem 3 (Error representative)**.** *Let* $u_h \in U_h$ *be the second component of the pair* $(\varepsilon_h, u_h) \in V_h \times U_h$ *solving the saddle-point problem* (27)*. Let* $\theta_h \in V_h$ *be the unique solution to* (15)*. Then the following holds true:*

$$\|u_h - \theta_h\|_{V_h} \leq \frac{1}{C_{\text{sta}}} \|\varepsilon_h\|_{V_h}. \tag{30}$$

Moreover, if the saturation Assumption 4 is satisfied, then the following a posteriori error estimate holds true:

$$\|u - u_h\|_{V_h} \leq \frac{1}{(1-\delta) C_{\text{sta}}} \|\varepsilon_h\|_{V_h}. \tag{31}$$

Proof. We observe that

$$\|B_h(u_h - \theta_h)\|_{V_h^*} = \|R_{V_h}^{-1}(B_h u_h - l_h)\|_{V_h} = \|\varepsilon_h\|_{V_h},$$

which leads to the bound (30) owing to the inf-sup condition (16). Then, invoking the triangle inequality and the saturation assumption leads to

$$\|u - u_h\|_{V_h} \leq \|u - \theta_h\|_{V_h} + \|u_h - \theta_h\|_{V_h} \leq \delta \|u - u_h\|_{V_h} + \|u_h - \theta_h\|_{V_h}.$$

Re-arranging the terms and using (30) proves the a posteriori estimate (31). □

Remark 1 ($U_h = V_h$)**.** *In the particular case where* $U_h = V_h$*, one readily verifies that the unique solution to the saddle-point problem* (27) *is the pair* $(0, \theta_h)$*, where* $\theta_h \in V_h$ *is the unique solution to* (15)*. In this situation, the bound* (30) *is not informative.*

Remark 2 (LS-FEM)**.** *Assume that $V_h := \mathbb{P}^p(\mathscr{D}_h)$, $p \geq 1$, is the broken polynomial space defined in* (14) *and that U_h is the H^1-conforming subspace $U_h := V_h \cap H^1(\mathscr{D})$. Assume that V_h is equipped with the L^2-norm and that the inf-sup condition* (16) *holds true (in the examples we have in mind, for example, a first-order PDE such as the advection-reaction equation described in Section 4, the discrete bilinear form is L^2-coercive). Then, the residual minimization problem* (22) *coincides with the Least-Squares Finite Element Method (LS-FEM) set in $L^2(\mathscr{D})$, and the error representative $\varepsilon_h \in V_h$ is the L^2-projection onto V_h of the finite element residual. Unfortunately, the L^2-norm is too weak, leading to an error estimate* (29) *that is suboptimal, typically by one order in the mesh size (i.e., of the form $Ch^p|u|_{H^{p+1}(\mathscr{D})}$). To remedy this difficulty, we shall equip V_h with a stronger norm inspired by the DG method, thereby leading to (asymptotic) quasi-optimality in* (29)*, that is, an upper bound of the form $Ch^{p+\frac{1}{2}}|u|_{H^{p+1}(\mathscr{D})}$.*

4 Model problem: Advection-reaction equation

In this section, we present examples of the above formulations in the context of the advection-reaction model problem.

4.1 Continuous setting

Let $\mathscr{D} \subset \mathbb{R}^d$, with $d = 1,2,3$, be an open, bounded Lipschitz polyhedron with boundary $\Gamma := \partial\mathscr{D}$ and outward unit normal $\mathbf{n}$. Let $\gamma \in L^\infty(\mathscr{D})$ denote a bounded reaction coefficient, and let $\mathbf{b} \in L^\infty(\mathscr{D};\mathbb{R}^d)$ denote velocity field such that $\nabla\cdot\mathbf{b} \in L^\infty(\mathscr{D})$. Assume that the boundary Γ can be split into the three subsets $\Gamma^\pm := \{\mathbf{x} \in \Gamma : \pm\mathbf{b}(\mathbf{x})\cdot\mathbf{n}(\mathbf{x}) < 0\}$ (inflow/outflow) and $\Gamma^0 := \{\mathbf{x} \in \Gamma : \mathbf{b}(\mathbf{x})\cdot\mathbf{n}(\mathbf{x}) = 0\}$ (characteristic boundary). Finally, let $f \in L^2(\mathscr{D})$ denote a source term and $g \in L^2(|\mathbf{b}\cdot\mathbf{n}|;\Gamma)$ denote a boundary datum, where

$$L^2(|\mathbf{b}\cdot\mathbf{n}|;\Gamma) := \left\{ v \text{ is measurable on } \Gamma : \int_\Gamma |\mathbf{b}\cdot\mathbf{n}|\, v^2\, d\Gamma < \infty \right\}. \tag{32}$$

The advection-reaction problem reads:

$$\left\{ \begin{array}{l} \text{Find } u \text{ such that:} \\ \quad \mathbf{b}\cdot\nabla u + \gamma\, u = f \quad \text{in } \mathscr{D}, \\ \qquad\qquad u = g \quad \text{on } \Gamma^-. \end{array} \right. \tag{33}$$

Consider the graph space $V := \left\{ v \in L^2(\mathscr{D}) : \mathbf{b}\cdot\nabla v \in L^2(\mathscr{D}) \right\}$ equipped with the inner product $(z,v)_V := (z,v)_{\mathscr{D}} + (\mathbf{b}\cdot\nabla z, \mathbf{b}\cdot\nabla v)_{\mathscr{D}}$ leading to the norm such that $\|v\|_V^2 := (v,v)_V$. Then, V is a Hilbert space. Moreover, assuming that Γ^- and Γ^+ are well separated, that is, $d(\Gamma^-,\Gamma^+) > 0$, traces are well defined over V, in the sense that the operator:

$$\begin{array}{rcl} \tau & : & C^0(\overline{\mathscr{D}}) \to L^2(|\mathbf{b}\cdot\mathbf{n}|;\Gamma) \\ & & v \to \tau(v) := v|_\Gamma, \end{array} \tag{34}$$

extends continuously to V (cf., [26]).

One possible weak formulation of (33) with weak imposition of the boundary condition reads:

$$\begin{cases} \text{Find } u \in V, \text{ such that:} \\ b(u,v) = (f,v)_{\mathscr{D}} + \langle (\mathbf{b}\cdot\mathbf{n})^{\ominus} g\,,\, v\rangle_{\Gamma}\,, \quad \forall\, v \in V, \end{cases} \tag{35}$$

with $(\cdot)^{\ominus}$ denoting the negative part such that $x^{\ominus} := \frac{1}{2}(|x| - x)$ for any real number x, and the continuous bilinear form b is such that:

$$b(z,v) := \big(\mathbf{b}\cdot\nabla z + \gamma\, z\,,\, v\big)_{\mathscr{D}} + \langle (\mathbf{b}\cdot\mathbf{n})^{\ominus}\, z\,,\, v\rangle_{\Gamma}\,. \tag{36}$$

The model problem (35) is well-posed under some mild assumptions on γ and $\mathbf{b}$ (see [26] or [12] for the details). The well-posedness is specific to right-hand sides of the form (35) (and not to any right-hand side in V^*), so that well-posedness is not established by reasoning directly on (35), but on an equivalent weak formulation where the boundary condition is strongly enforced by invoking a surjectivity property of the trace operator.

4.2 DG discretization

Recall the discrete setting from Section 2.3. For a polynomial degree $p \geq 0$, the standard DG discretization of the model problem (35) is of the form:

$$\begin{cases} \text{Find } \theta_h \in V_h := \mathbb{P}^p(\mathscr{D}_h), \text{ such that:} \\ b_h^{\mathrm{dg}}(\theta_h, v_h) := b_h(\theta_h, v_h) + p_h(\theta_h, v_h) = l_h(z_h), \quad \forall\, v_h \in V_h, \end{cases} \tag{37}$$

where $b_h(\cdot,\cdot)$ and $p_h(\cdot,\cdot)$ are the bilinear forms over $V_h \times V_h$ such that

$$\begin{aligned} b_h(z_h, v_h) := \sum_{m=1}^{N} (\mathbf{b}\cdot\nabla z_h + \gamma\, z_h\,,\, v_h)_{K_m} + \sum_{e\in\mathscr{S}_h^{\partial}} \langle (\mathbf{b}\cdot\mathbf{n}_e)^{\ominus} z_h\,,\, v_h\rangle_e \\ - \sum_{e\in\mathscr{S}_h^{0}} \langle (\mathbf{b}\cdot\mathbf{n}_e)\, [\![z_h]\!]\,,\, \{\!\{ v_h \}\!\}\rangle_e\,, \end{aligned} \tag{38}$$

and

$$p_h(z_h, v_h) := \frac{\eta}{2} \sum_{e\in\mathscr{S}_h^{0}} \langle |\mathbf{b}\cdot\mathbf{n}_e|\, [\![z_h]\!], [\![v_h]\!]\rangle_e\,, \tag{39}$$

where $\eta \geq 0$ is the penalty parameter, and $l_h(\cdot)$ is the linear form over V_h such that

$$l_h(v_h) := \sum_{m=1}^{N} (f, v_h)_{K_m} + \sum_{e\in\mathscr{S}_h^{\partial}} \langle (\mathbf{b}\cdot\mathbf{n}_e)^{\ominus}\, g\,,\, v_h\rangle_e\,. \tag{40}$$

The choice $\eta = 0$ corresponds to the use of centered fluxes, and the choice $\eta > 0$ (typically $\eta = 1$) to the use of upwind fluxes, see [8, 22]. To verify that the assumptions 1-3 are satisfied, we need to make a reasonable regularity assumption on the exact solution.

Assumption 5. *(Partition of $\mathscr{D}$ and regularity of exact solution u) There is a partition $\mathscr{P}_{\mathscr{D}} = \bigcup_{i=1}^{N_{\mathscr{D}}} \mathscr{D}_i$ of $\mathscr{D}$ into open disjoint polyhedra $\mathscr{D}_i$ such that:*

(i) The inner part of the boundary of $\mathscr{D}_i$ is characteristic with respect to the advective field, that is, $\mathbf{b}(\mathbf{x})\cdot\mathbf{n}_{\mathscr{D}_i}(\mathbf{x}) = 0$ for all $\mathbf{x} \in \partial\mathscr{D}_i \cap \mathscr{D}$ and all $i \in \{1,\dots,N_{\mathscr{D}}\}$, where $\mathbf{n}_{\mathscr{D}_i}$ is the unit outward normal to $\mathscr{D}_i$;

(ii) The exact solution is such that

$$u \in X_\# := V \cap H^s(\mathscr{P}_{\mathscr{D}}), \quad s > \frac{1}{2}. \tag{41}$$

(iii) The mesh is aligned with this partition, that is, any mesh cell belongs to one and only one subset $\mathscr{D}_i$.

Assumption 5 is instrumental to ensure that the trace of u is meaningful on the boundary of each mesh cell, and that u can only jump across those interfaces that are subsets of some $\partial\mathscr{D}_i \cap \mathscr{D}$ with $i \in \{1,\dots,N_{\mathscr{D}}\}$. In the case where u has no jumps, we can broadly assume that $\mathscr{P}_{\mathscr{D}} = \mathscr{D}$, in which case the statement of Assumption 5-(i) is void.

The broken polynomial space V_h can be equipped with various norms. We consider the following choices:

$$\begin{aligned} \|w\|_{\mathrm{cf}}^2 &:= \|w\|_{\mathscr{D}}^2 + \frac{1}{2}\left\| |\mathbf{b}\cdot\mathbf{n}|^{\frac{1}{2}}\, w \right\|_\Gamma^2, \\ \|w\|_{\mathrm{up}}^2 &:= \|w\|_{\mathrm{cf}}^2 + \sum_{e\in\mathscr{S}_h^0} \frac{\eta}{2}\left\| |\mathbf{b}\cdot\mathbf{n}_e|^{\frac{1}{2}} [\![w]\!] \right\|_e^2 + \sum_{m=1}^N h_{K_m} \|\,\mathbf{b}\cdot\nabla w\|_{K_m}^2. \end{aligned} \tag{42}$$

In view of Assumption 3, we define $V_{h,\#} := X_\# + V_h$ with $X_\#$ defined in (41), and we consider the following extensions of the above norms:

$$\begin{aligned} \|w\|_{\mathrm{cf},\#}^2 &:= \|w\|_{\mathrm{cf}}^2 + \sum_{m=1}^N \left(\|\mathbf{b}\cdot\nabla w\|_{K_m}^2 + h_{K_m}^{-1}\|w\|_{\partial K_m}^2 \right), \\ \|w\|_{\mathrm{up},\#}^2 &:= \|w\|_{\mathrm{up}}^2 + \sum_{m=1}^N \left(h_{K_m}^{-1}\|w\|_{K_m}^2 + \|w\|_{\partial K_m}^2 \right). \end{aligned} \tag{43}$$

For the proof of the following result, we refer the reader to [22].

Proposition 1. *(Verification of the assumptions) In the above framework, Assumptions 1-3 hold true in the following situations:* (i) $\eta = 0$ *(centered fluxes) and the norms $\|\cdot\|_{\mathrm{cf}}$ and $\|\cdot\|_{\mathrm{cf},\#}$;* (ii) $\eta > 0$ *(upwind fluxes) and the norms $\|\cdot\|_{\mathrm{up}}$ and $\|\cdot\|_{\mathrm{up},\#}$.*

Remark 3. *(Coercivity) The case of centered fluxes in Proposition 1 leads to stability in the form of coercivity, whereas the case of upwind fluxes leads to an inf-sup condition, and the norm that is controlled is stronger than with centered fluxes.*

An immediate consequence of Theorem 1 is the bound

$$\inf_{y_h\in V_h} \|u - y_h\|_{V_h} \le \|u - \theta_h\|_{V_h} \le C \inf_{y_h\in V_h} \|u - y_h\|_{V_h,\#}, \tag{44}$$

with C uniform with respect to the mesh size, where the lower bound is trivial and simply results from the fact that $\theta_h \in V_h$. When comparing the decay rates with respect to the mesh size for the best-approximation errors of smooth solutions in the norms $\|\cdot\|_{V_h}$ and $\|\cdot\|_{V_h,\#}$, the following terminology is useful: the estimate (44) is said to be suboptimal if the left-hand side decays at a faster rate than the right-hand side, and it is said to be (asymptotically) quasi-optimal if both sides decay at the same rate. The inequality in (44) is suboptimal when centered fluxes are used (the left-hand side decays at a rate of order $h^{p+\frac{1}{2}}$ and the right-hand side decays at rate h^p), and it is (asymptotically) quasi-optimal when upwind fluxes are used (both sides decay at a rate of order $h^{p+\frac{1}{2}}$) (cf., [22]).

Remark 4 (Continuous trial space)**.** *If we set $U_h := V_h \cap H^1(\mathscr{D})$, that is, if U_h is composed of continuous, piecewise polynomial functions, the discrete bilinear form $b_h(\cdot,\cdot)$ reduces on $U_h \times V_h$ to*

$$b_h(z_h, v_h) = (\mathbf{b}\cdot\nabla z_h + \gamma\, z_h\,,\, v_h)_{\mathscr{D}} + \sum_{e\in\mathscr{S}_h^\partial} \langle (\mathbf{b}\cdot\mathbf{n}_e)^\ominus z_h\,,\, v_h\rangle_e\,,$$

since functions in U_h have zero jumps across the mesh interfaces and their piecewise gradient coincides with their weak gradient over $\mathscr{D}$.

Remark 5 (Best approximation)**.** *An interesting property of the present setting regarding the approximation capacity of the discrete spaces U_h and V_h is that, for all $v \in H^s(\mathscr{D})$, $s > \frac{1}{2}$,*

$$\inf_{v_h\in U_h} \|v - v_h\|_{up,\#} \le C_{\mathrm{app}} \inf_{v_h\in V_h} \|v - v_h\|_{up,\#},$$

with C_{app} uniform with respect to the mesh size (see B). The converse bound is trivially satisfied with constant equal to 1 *since $U_h \subset V_h$.*

5 Numerical examples

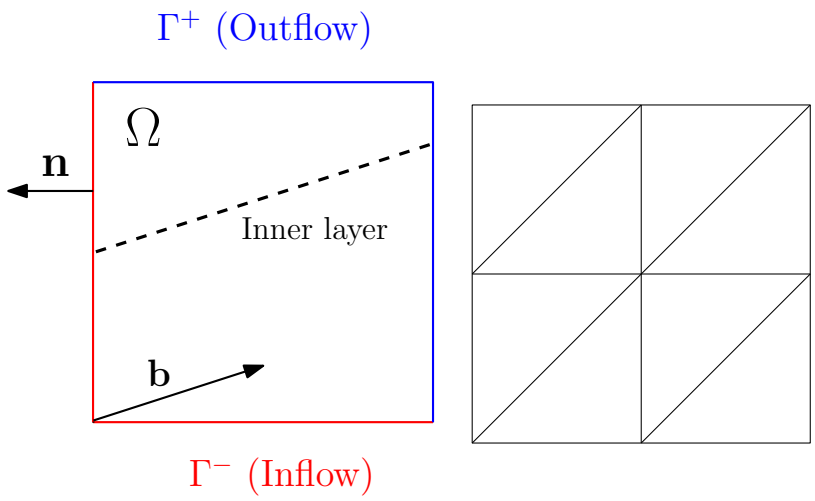
In this section, we present the 2D and 3D test cases, cover some implementation aspects, and discuss the numerical results.

5.1 Model problems

5.1.1 2D model problem

We consider the purely advective problem (33) ($\gamma = 0$) over the unit square $\mathscr{D} = (0,1)^2 \subset \mathbb{R}^2$, with a constant velocity field $\mathbf{b} = (3,1)^T$ (see Figure 2a). We consider the source term $f = 0$ in $\mathscr{D}$ and, for a parameter $M > 0$, an inflow boundary datum $g = u_{M|\Gamma^-}$, where $\Gamma^- = \{(0,y), y\in(0,1)\}\cup\{(x,0), x\in(0,1)\}$, and the exact solution u_M is given by

$$u_M(x_1,x_2) = 1 + \tanh\left(M\left(x_2 - \frac{x_1}{3} - \frac{1}{2}\right)\right). \tag{45}$$



(a) 2D problem setting (b) 2D initial mesh

Figure 2: 2D model problem and initial mesh

The parameter M tunes the width of the inner layer along the line of equation $x_2 = \frac{x_1}{3} + \frac{1}{2}$. The limit value as M grows becomes

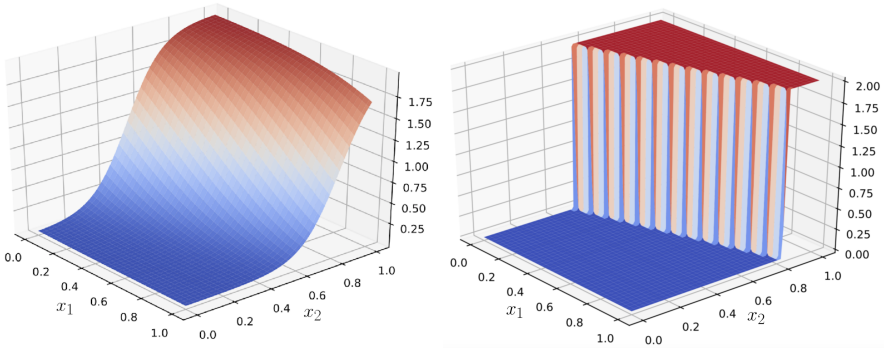
$$u_\infty(x_1, x_2) := \lim_{M \to +\infty} u_M(x_1, x_2) = 1 + \text{sign}\left(x_2 - \frac{x_1}{3} - \frac{1}{2}\right). \tag{46}$$

The inner layer can be seen in Figure 3b for the case $M = 500$ (compare with Figure 3a for the case $M = 5$). Although the inflow and outflow boundaries are not well-separated here, the exact solution matches the partition and regularity assumption 5. Indeed, $\mathscr{P}_\mathscr{D} = \Omega$ when $M < \infty$, whereas $\mathscr{P}_\mathscr{D}$ is composed of two subsets with the common characteristic interface $\{(x_1, x_2) \in \mathscr{D} \mid x_2 - \frac{x_1}{3} - \frac{1}{2} = 0\}$ when $M = \infty$ (see Figure 2a). In addition, the absence of a reactive term precludes the straightforward derivation of L^2-stability by a coercivity argument. However, L^2-stability is recovered via an inf-sup argument which remains valid whenever the advective field is filling (which is trivially the case for a constant field across a square domain); we refer the reader to [21, 3, 2, 13] for further insight into this point.

5.1.2 3D model problem

We still consider the purely advective problem (33), this time over the unit cube $\mathscr{D} = (0\,,1)^3 \subset \mathbb{R}^3$, the source term $f = 0$, the spiral-type velocity field $\mathbf{b}(x_1, x_2, x_3) = (-0.15\sin(4\pi x_3)\,, 0.15\cos(4\pi x_3)\,, 1)^T$, and the inflow boundary datum $g = u_M|_{\Gamma^-}$, where the exact solution u_M is given by

$$u_M(x_1, x_2, x_3) = 1 + \tanh\left(M\left(0.15^2 - \left(x_1 - X_1(x_3)\right)^2 - \left(x_2 - X_2(x_3)\right)^2\right)\right), \tag{47}$$



(a) 2D exact solution for $M = 5$ (b) 2D exact solution for $M = 500$

Figure 3: 2D exact solutions for $M = 5$ and $M = 500$

with $X_1(x_3) = 0.15\cos(4\pi x_3) + 0.45$ and $X_2(x_3) = 0.15\sin(4\pi x_3) + 0.5$. The limit value as M grows becomes

$$u_\infty(x_1, x_2, x_3) = 1 + \text{sign}\Big(0.15^2 - \big(x_1 - X_1(x_3)\big)^2 - \big(x_2 - X_2(x_3)\big)^2\Big). \tag{48}$$

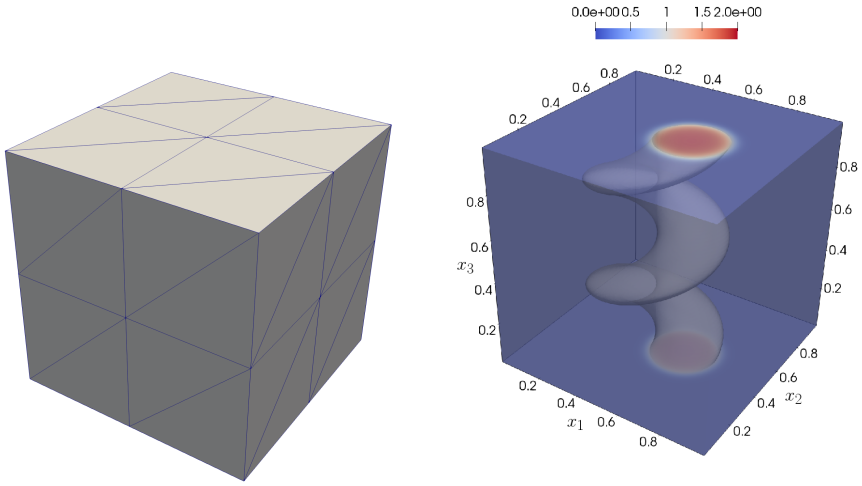
Figure 4b illustrates the inner layer for $M = 100$. The inflow boundary corresponds to the following union of portions of planes:

$$\begin{aligned}\Gamma^- = \{&(x_1, x_2) \in \Gamma_1 \cup \Gamma_2,\ x_3 \in [0, \tfrac{1}{8}] \cup [\tfrac{1}{2}, \tfrac{5}{8}]\}\\ &\cup \{(x_1, x_2) \in \Gamma_2 \cup \Gamma_3,\ x_3 \in [\tfrac{1}{8}, \tfrac{1}{4}] \cup [\tfrac{5}{8}, \tfrac{3}{4}]\}\\ &\cup \{(x_1, x_2) \in \Gamma_3 \cup \Gamma_4,\ x_3 \in [\tfrac{1}{4}, \tfrac{3}{8}] \cup [\tfrac{3}{4}, \tfrac{7}{8}]\}\\ &\cup \{(x_1, x_2) \in \Gamma_4 \cup \Gamma_1,\ x_3 \in [\tfrac{3}{8}, \tfrac{1}{2}] \cup [\tfrac{7}{8}, 1]\},\end{aligned} \tag{49}$$

where $\Gamma_1 = \{x_1 \in (0,1), x_2 = 0\}$, $\Gamma_2 = \{x_1 = 1, x_2 \in (0,1)\}$, $\Gamma_3 = \{x_1 \in (0,1), x_2 = 1\}$, and $\Gamma_4 = \{x_1 = 0, x_2 = 0\}$. As for the 2D model problem, the solution does not satisfy the regularity assumption 5 since the inflow and outflow boundaries are not well separated; moreover, whenever $M = \infty$, the exact solution is piecewise smooth, but the subsets $\mathscr{D}_i$ in the corresponding partition are not polyhedra.

5.2 Implementation aspects

We consider the broken polynomial space $V_h := \mathbb{P}^p(\mathscr{D}_h)$, with $p = 1, 2$, defined in (14), and consider the H^1-conforming subspace $U_h := V_h \cap H^1(\mathscr{D})$, that is, U_h is composed of continuous, piecewise polynomial functions of degree $p = 1, 2$. Since the trial space for the minimization problem (22) is composed of continuous functions, we use the label "CT" in our figures. We equip the space V_h with one of the two norms defined in (42), leading to the labels "CT-cf" and "CT-up". For comparison purposes, we also compute the DG solution $\theta_h \in V_h$ solving the primal problem (15). When reporting the corresponding error, we use the labels "DT-cf" and "DT-up", where



(a) 3D initial mesh

(b) 3D exact solution for $M = 100$

Figure 4: 3D initial mesh and exact solution

"DT" means discontinuous trial space, and the labels “cf” and “up” indicate the use of centered fluxes and upwind fluxes (with $\eta = 1$), respectively, as well as the norm in which the error is evaluated. The solution “CT-cf” can be loosely interpreted as a LS-FEM solution. Indeed the residual minimization is performed over the H^1-conforming finite element space, and the minimizing functional is the L^2-norm of the residual in $\mathscr{D}$ supplemented by the L^2-norm in Γ^- for the residual associated with the boundary condition (see the norm $\|\cdot\|_{\text{cf}}$ in (42)).

In all the 2D simulations, we start with the uniform triangular mesh shown in Figure 2b, whereas for 3D simulations, we start with the uniform tetrahedral mesh shown in Figure 4a. We produce subsequent mesh refinements under uniform and adaptive criteria. We obtain all the solutions of the saddle-point problem (27) and the primal formulation (15) by using FEniCS [1]. We show convergence plots of the error measured in the chosen norm of V_h as a function of the number of degrees of freedom (DOFs) (that is, $\dim(U_h) + \dim(V_h)$ for (27) and $\dim(V_h)$ for (15)).

5.2.1 Adaptive mesh refinement

Adaptive mesh refinement is possible when solving (27) and we use the error representative $\varepsilon_h \in V_h$ for that purpose. A standard adaptive procedure considers an iterative loop where each step consists of the following four modules:

$$\text{SOLVE} \rightarrow \text{ESTIMATE} \rightarrow \text{MARK} \rightarrow \text{REFINE}.$$

These four modules are applied as follows: We first solve the saddle-point problem (27). Then, we compute for each mesh cell K, the local error indicator E_K defined as

$$E_K^2 := \begin{cases} E_{\mathrm{cf},K}^2 := \|\varepsilon_h\|_K^2 + \frac{1}{2}\left\| |\mathbf{b}\cdot\mathbf{n}|^{\frac{1}{2}}\,\varepsilon_h\right\|_{\Gamma\cap\partial K}^2, & \text{if } \|\cdot\|_{V_h} = \|\cdot\|_{\mathrm{cf}}, \\ E_{\mathrm{cf},K}^2 + \dfrac{\eta}{2}\left\| |\mathbf{b}\cdot\mathbf{n}|^{\frac{1}{2}}\,[\![\varepsilon_h]\!]\right\|_{\mathscr{S}_h^0\cap\partial K}^2 + h_K\|\,\mathbf{b}\cdot\nabla\varepsilon_h\|_K^2, & \text{if } \|\cdot\|_{V_h} = \|\cdot\|_{\mathrm{up}}. \end{cases} \tag{50}$$

We mark using the Dörfler bulk-chasing criterion (see [23]) that marks elements for which the cumulative sum of the local values E_K in a decreasing order remains below a chosen fraction of the total estimated error $\|\varepsilon_h\|_{V_h}$. For the numerical examples, we consider this fraction to be one half and a quarter for the 2D and 3D examples, respectively. Finally, we use a bisection-type refinement criterion (see [4]) to obtain the refined mesh to be used in the next step.

5.2.2 Iterative solver

The algebraic realization of problem (27) takes the form

$$\begin{pmatrix} G & B \\ B^T & 0 \end{pmatrix}\begin{pmatrix} \varepsilon \\ \mathbf{u} \end{pmatrix} = \begin{pmatrix} \mathbf{l} \\ \mathbf{0} \end{pmatrix}. \tag{51}$$

We consider the iterative algorithm proposed in [5]. Denoting by $\widehat{G}$ a preconditioner for the Gram matrix G, and by $\widehat{S}$ a preconditioner for the reduced Schur complement $B^T\widehat{G}^{-1}B$, the iterative scheme can be written as

$$\begin{pmatrix} \varepsilon_{i+1} \\ \mathbf{u}_{i+1} \end{pmatrix} = \begin{pmatrix} \varepsilon_i \\ \mathbf{u}_i \end{pmatrix} + \begin{pmatrix} \widehat{G} & B \\ B^T & \widehat{C} \end{pmatrix}^{-1}\left\{\begin{pmatrix} \mathbf{l} \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} G & B \\ B^T & 0 \end{pmatrix}\begin{pmatrix} \varepsilon_i \\ \mathbf{u}_i \end{pmatrix}\right\}, \tag{52}$$

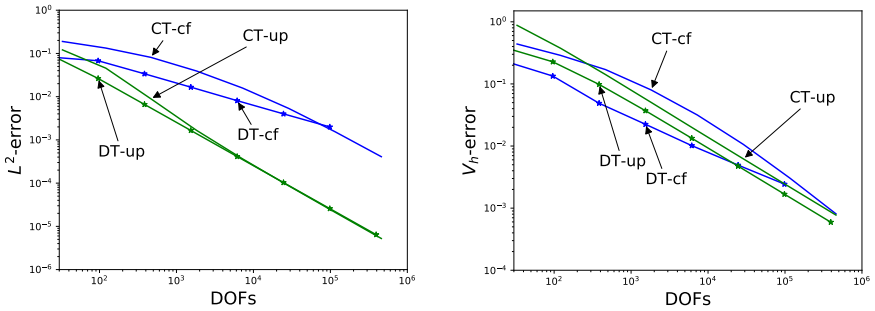
with $\widehat{C} = B^T\widehat{G}^{-1}B - \hat{S}$. Denoting by $\mathbf{r}_i = \mathbf{l} - G\varepsilon_i - B\,\mathbf{u}_i$ and by $\mathbf{s}_i = -B^T\varepsilon_i$ the residuals for ε and $\mathbf{u}$ at the outer iteration i, the scheme requires the resolution of two interior problems for the following increments:

$$\eta_{i+1} := \mathbf{u}_{i+1} - \mathbf{u}_i = \hat{S}^{-1}\left(B^T\left(\widehat{G}^{-1}\mathbf{r}_i\right) - \mathbf{s}_i\right), \tag{53}$$

and

$$\delta_{i+1} := \varepsilon_{i+1} - \varepsilon_i = \hat{G}^{-1}\left(\mathbf{r}_i - B\,\boldsymbol{\eta}_{i+1}\right). \tag{54}$$

In [5], the authors suggest to consider $\hat{G}$ as a relaxed approximation for the matrix G, for instance, a few iterations of the conjugate gradient method. However, in our context, the matrix G plays an important role in stabilizing the system. Therefore, less accurate representations worsen the conditioning of the reduced Schur complement in (53). For this reason, we consider one outer iteration and, as preconditioners, a sparse Cholesky factorization obtained using the module "sksparse.cholmod" (see [16]) for the G matrix, whereas the conjugate gradient method (available in the Scipy sparse linear algebra package) is the preconditioner $\widehat{S}$. On the coarsest mesh, we consider the initial guess to be zero vectors, whereas on the subsequent adaptive meshes, this guess becomes the solution obtained in the previous level of refinement.

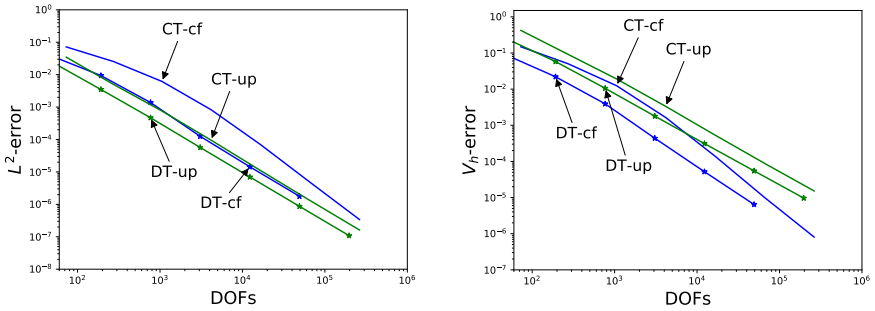


(a) L^2 error vs. DOFs (b) V_h error vs. DOFs

Figure 5: 2D model problem: L^2-error and V_h-error vs. DOFs using uniform mesh refinement for $p = 1$ and $M = 5$.

5.3 Discussion of the numerical results

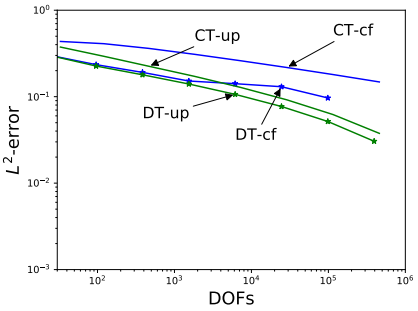
5.3.1 Discussion of the 2D results



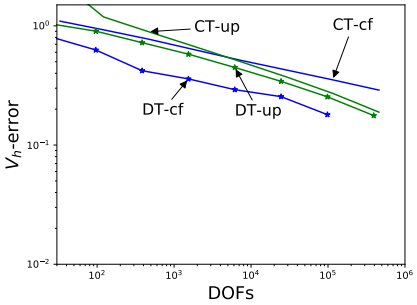
(a) L^2 error vs. DOFs (b) V_h error vs. DOFs

Figure 6: 2D model problem: L^2-error and V_h-error vs. DOFs using uniform mesh refinement for $p = 2$ and $M = 5$.

As a first example, we consider the value $M = 5$ in the 2D exact solution (45) so as to obtain a sufficiently smooth solution that allows us to appreciate the expected convergence rates. Figure 5 reports the results obtained with uniform mesh refinement and the polynomial degree $p = 1$. Figure 5a shows the error measured in the L^2-norm vs. DOFs in log-scale, whereas Figure 5b shows the error measured in the corresponding V_h-norm vs. DOFs in log-scale. The main point is that the "CT-up" solution converges at the same rate as the "DT-up" solution and with very close error values, when measuring the error using both the L^2- and $\|\cdot\|_{\text{up}}$-norms. Interestingly, the "CT-cf" solution converges with a higher rate when compared with the "DT-cf"

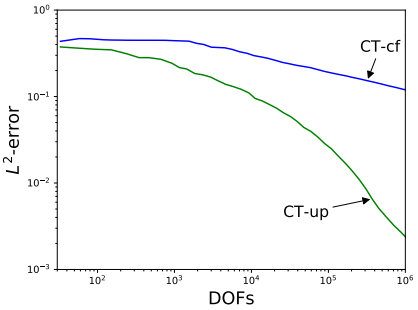


(a) L^2 error vs. DOFs; uniform

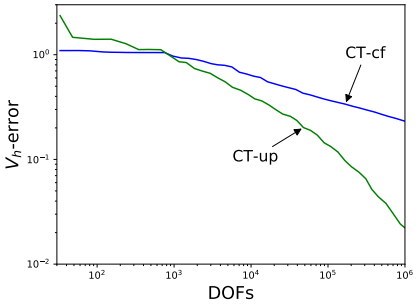


(b) V_h error vs. DOFs; uniform

Figure 7: 2D model problem: errors in the L^2- and $\|\cdot\|_{V_h}$-norms vs. DOFs for uniform meshes, $p = 1$, and $M = 500$.



(a) L^2 error vs. DOFs; adaptive

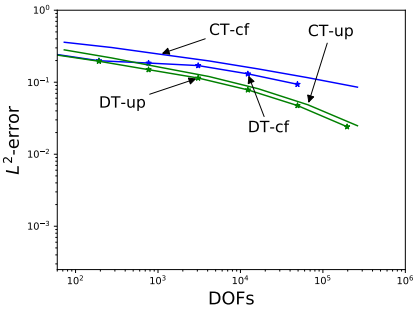


(b) V_h error vs. DOFs; adaptive

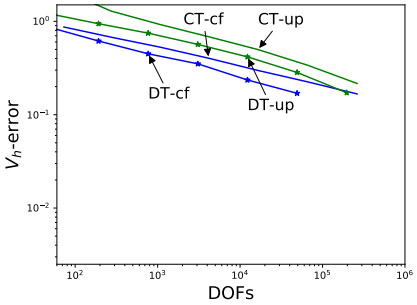
Figure 8: 2D model problem: errors in the L^2- and $\|\cdot\|_{V_h}$-norms vs. DOFs for adaptive meshes, $p = 1$, and $M = 500$.

solution in the L^2-norm, and that is also the case for this solution in the $\|\cdot\|_{\text{cf}}$-norm. In Figure 6, we consider the same type of results, but using the polynomial degree $p = 2$. The conclusions we can draw are similar. Again the most salient one is that the "CT-up" and "DT-up" solutions converge at the same rate for both the L^2- and $\|\cdot\|_{\text{up}}$-norms, both methods delivering very close error values. Incidentally, the "DT-cf" solution also converges at the same rate for $p = 2$, again in both norms.

As a second example, we consider the value $M = 500$ for the 2D exact solution u_M in (45) so that u_M is very close to the discontinuous function u_∞ defined in (46) (see Figure 3b). In this case, optimal convergence rates are not obtained when considering uniform mesh refinements as can be appreciated from the results reported in Figures 7a-7b, and 9a-9b for the polynomial degrees $p = 1$ and $p = 2$, respectively. However, as Figures 8a-8b, and 10a-10b show, the convergence improves for the "CT-up" solution by resorting to mesh adaptation. Figures 11 and 12 compare

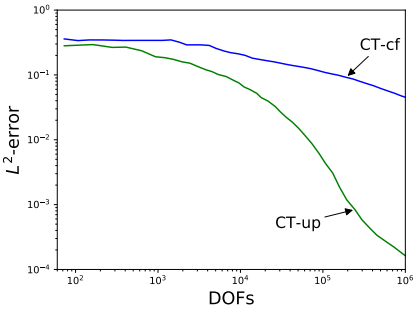


(a) L^2 error vs. DOFs; uniform

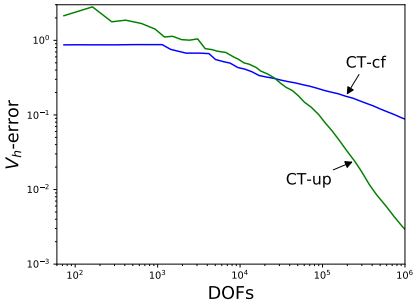


(b) V_h error vs. DOFs; uniform

Figure 9: 2D model problem: errors in the L^2- and $\|\cdot\|_{V_h}$-norms vs. DOFs for uniform meshes, $p = 2$, and $M = 500$.



(a) L^2 error vs. DOFs; adaptive



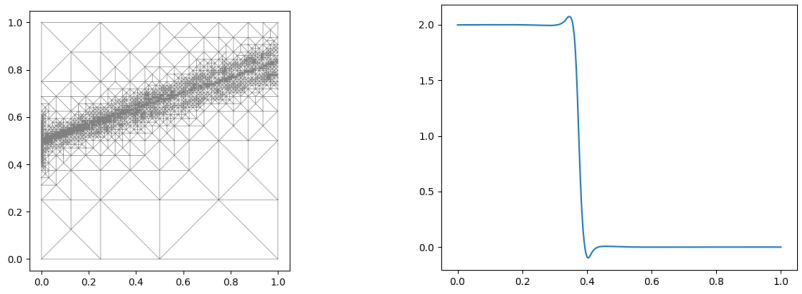
(b) V_h error vs. DOFs; adaptive

Figure 10: 2D model problem: errors in the L^2- and $\|\cdot\|_{V_h}$-norms vs. DOFs for adaptive meshes, $p = 2$, and $M = 500$.

the adaptively refined mesh and a cut of the solution over the line $\{x_1 = 1 - x_2\}$ requiring a similar number of total DOFs. The key observation is that the choice of the norm in V_h has a significant impact on the convergence of the adaptive process, as much faster convergence rates are observed if one uses the $\|\cdot\|_{\mathrm{up}}$-norm rather than the $\|\cdot\|_{\mathrm{cf}}$-norm. This not only shows that the error representative defined in (26) can be used for adaptivity, but also that the use of the stronger norm $\|\cdot\|_{\mathrm{up}}$ drives the adaptive process more efficiently than the $\|\cdot\|_{\mathrm{cf}}$-norm which would be considered in LS-FEM.

5.3.2 Discussion of the 3D results

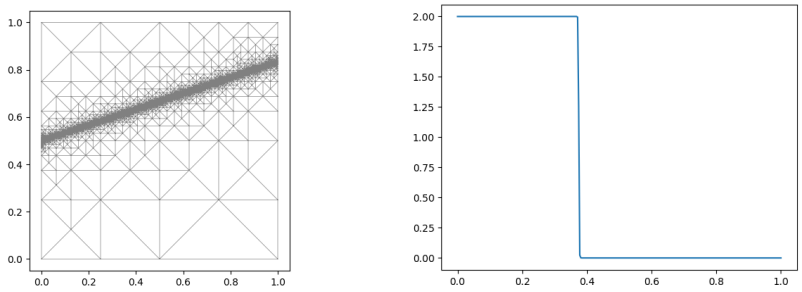
In this section, we explore the performance of the proposed adaptive method with the $\|\cdot\|_{\mathrm{up}}$-norm for the 3D model problem. We consider the value $M = 100$ in the



(a) Adapted mesh

(b) Solution cut over the line $x_1 = 1 - x_2$

Figure 11: 2D model problem: adaptively refined mesh and transversal cut of the discrete solution for $p = 2$, $M = 500$ at level 35 of refinement; CT-cf, 191,546 DOFs.

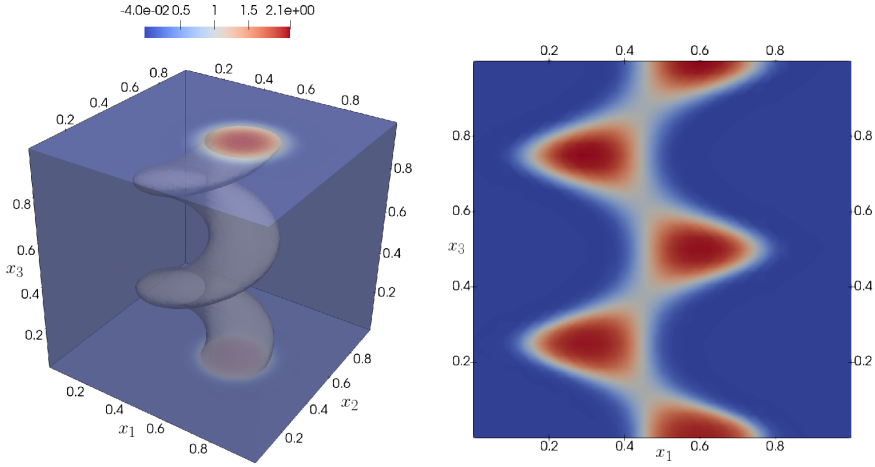


(a) Adapted mesh

(b) Solution cut over the line $x_1 = 1 - x_2$

Figure 12: 2D model problem: adaptively refined mesh and transversal cut of the discrete solution for $p = 2$, $M = 500$; CT-up, 159,255 DOFs.

exact solution (47) so that u_M is close to the discontinuous limit when $M = \infty$. The method delivers accurate solutions by refining where it is most needed, as can be appreciated in Figures 13a-13d, and 14a-14d respectively, where we show a 3D representation of the contour and cuts over the planes $\{x_2 = 0.5\}$, $\{x_3 = 0\}$ (part of inflow) and $\{x_3 = 1\}$ (part of outflow) for the solution and for the mesh obtained after 31 levels of adaptive refinement. The mesh refinement at the plane $\{x_3 = 0\}$ located at the inflow boundary is finer compared with that at the plane $\{x_3 = 1\}$ located at the outflow boundary. Finally, Figures 15a-15b show the convergence in the L^2- and V_h-norms vs the total number of DOFs for the polynomial degrees $p = 1, 2$.



(a) Contour representation of the solution

(b) Solution cut over the plane $\{x_2 = 0.5\}$

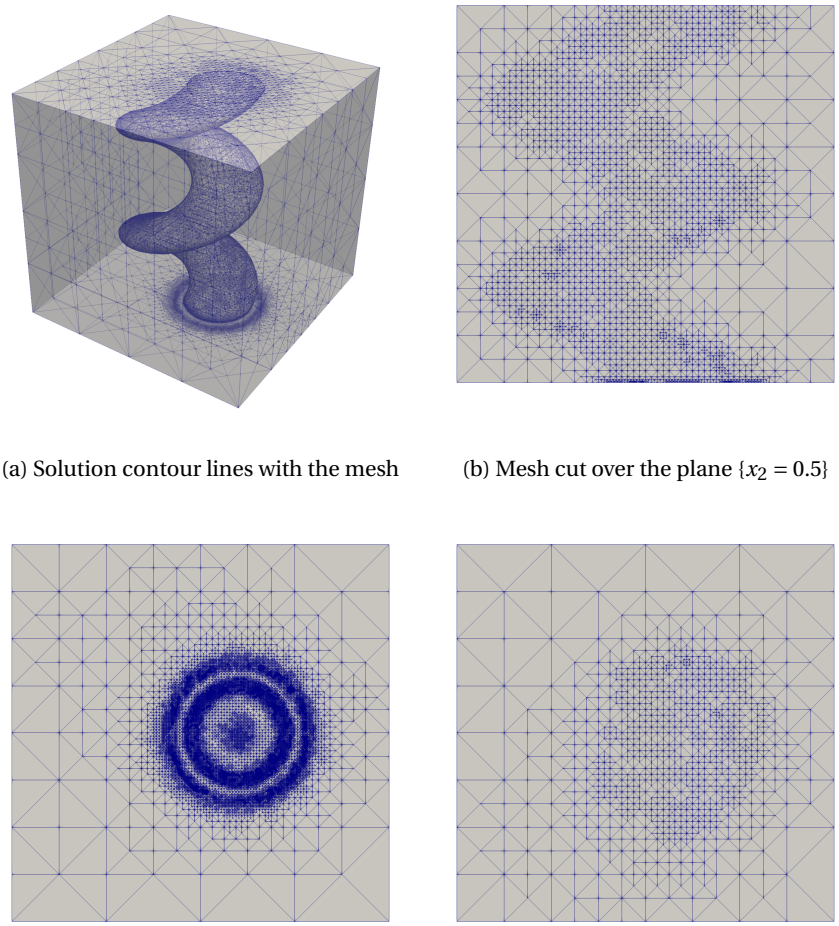
(c) Solution cut over the plane $\{x_3 = 0\}$

(d) Solution cut over the plane $\{x_3 = 1\}$

Figure 13: 3D model problem: solution contours over the whole domain and cuts over three planes at the level 31 of refinement; $p = 1$. 159,705 DOFs for the trial space, and 3,569,333 total DOFs

6 Conclusions

In this paper, we proposed a new stabilized finite element method based on residual minimization. The key idea is to obtain a residual representation using a dual norm defined over a discontinuous Galerkin space that is equipped with a norm



(a) Solution contour lines with the mesh

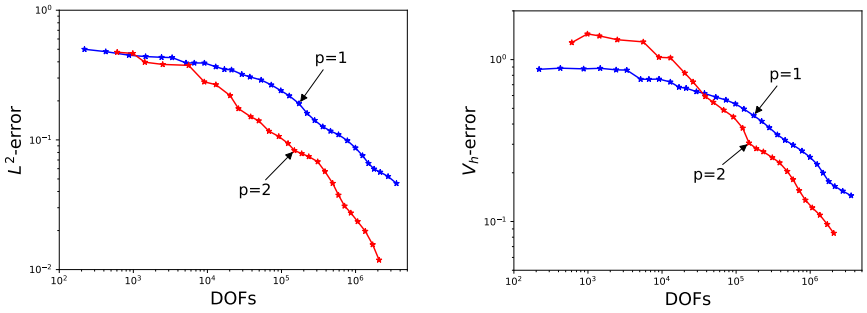
(b) Mesh cut over the plane $\{x_2 = 0.5\}$

(c) Mesh cut over the plane $\{x_3 = 0\}$

(d) Mesh cut over the plane $\{x_3 = 1\}$

Figure 14: 3D model problem: mesh over the whole domain and cuts over three planes at level 31 of refinement, $p = 1$, 159,705 DOFs for the trial space, and 3,569,333 total DOFs

that delivers inf-sup stability. The cost is that one needs to solve a stable saddle-point problem. The advantage is that one recovers at the same time a stabilized finite element solution and an error representative defined in the discontinuous Galerkin space that can be used to drive adaptive mesh refinement. Our numerical results on 2D and 3D advective model problems featuring sharp inner layers indi-



(a) L^2 error vs. DOFs

(b) V_h error vs. DOFs

Figure 15: 3D model problem: L^2-error and V_h-error vs. DOFs for the 3D spiral problem with adaptivity mesh refinement; up norm. $p = 1,2$

cate that the present method leads to competitive error decay rates on uniformly refined meshes with respect to the discontinuous Galerkin approximation while at the same time being able to deliver adaptive meshes that sharply capture inner layers. Further studies are on the way to assess the computational performance of the proposed method, especially in 3D, and on other model problems comprising, for example, systems of first-order PDEs of Friedrichs's type, as in Darcy's equations or in Maxwell's equations.

A Proof of Theorem 2.

Recall the definition of the discrete operator $B_h : U_h \to V_h^*$ from (23) (here, the domain of B_h is restricted to U_h). Let $B_h U_h \subset V_h^*$ be the range of B_h, and $(B_h U_h)^\perp \subset V_h$ be such that

$$(B_h U_h)^\perp := \{v_h \in V_h : b_h(z_h, v_h) = 0, \forall\, z_h \in U_h\},$$

where $(B_h U_h)^\perp = \ker B_h^*$ with $B_h^* : V_h \to U_h^*$.

We prove the well-posedness of (27) by establishing the equivalence between (27) and the following constrained minimization problem:

$$\inf_{v_h \in (B_h U_h)^\perp} \left\{ \frac{1}{2} \|v_h\|_{V_h}^2 - \langle l_h, v_h \rangle_{V_h^*, V_h} \right\} =: \inf_{v_h \in (B_h U_h)^\perp} F(v_h), \tag{55}$$

which has a unique solution since the functional F is strictly convex and $(B_h U_h)^\perp$ is a closed subspace of V_h. By differentiating with respect to v_h, we observe that the minimizer $\tilde{v}_h \in (B_h U_h)^\perp$ of (55) must be a critical point of F, that is,

$$(\tilde{v}_h, v_h)_{V_h} - \langle l_h, v_h \rangle_{V_h^*, V_h} = 0, \qquad \forall\, v_h \in (B_h U_h)^\perp. \tag{56}$$

Any component ε_h of a solution $(\varepsilon_h, u_h) \in V_h \times U_h$ of the saddle-point problem (27) satisfies (56). Conversely, let $\varepsilon_h = \tilde{v}_h \in (B_h U_h)^\perp$ be the unique solution of (56). Then,

$l_h - R_{V_h}\varepsilon_h$ is in $(B_hU_h)^{\perp\perp} = B_hU_h$. Hence, there must be $u_h \in U_h$ such that $B_hu_h = l_h - R_{V_h}\varepsilon_h$. In other words, $(\varepsilon_h, u_h) \in V_h \times U_h$ solves the saddle-point problem (27). Finally, the uniqueness of the solution to (27) readily follows from the injectivity of B_h (which is a consequence of the inf-sup condition (16)) and the bijectivity of the Riesz isomorphism.

Let us now prove the a priori bounds (28) for ε_h and u_h. Testing the equation (27a) with $v_h = \varepsilon_h$, we infer that

$$\|\varepsilon_h\|_{V_h}^2 = \langle l_h, \varepsilon_h\rangle_{V_h^*,V_h} \le \|l_h\|_{V_h^*}\|\varepsilon_h\|_{V_h},$$

and the first a priori bound follows by dividing the above expression by $\|\varepsilon_h\|_{V_h}$. For the second a priori bound in (28), we have

$$\begin{aligned}
\|u_h\|_{V_h} &\le \frac{1}{C_{\mathrm{sta}}} \sup_{0\neq v_h\in V_h} \frac{b_h(u_h,v_h)}{\|v_h\|_{V_h}} = \frac{1}{C_{\mathrm{sta}}} \sup_{0\neq v_h\in V_h} \frac{\big(R_{V_h}^{-1}B_hu_h, v_h\big)_{V_h}}{\|v_h\|_{V_h}} && \text{(by (16) and (23))}\\
&= \frac{1}{C_{\mathrm{sta}}} \frac{\big(R_{V_h}^{-1}B_hu_h, R_{V_h}^{-1}B_hu_h\big)_{V_h}}{\|R_{V_h}^{-1}B_hu_h\|_{V_h}} && \text{(since } v_h = R_{V_h}^{-1}B_hu_h \text{ is the supremizer)}\\
&= \frac{1}{C_{\mathrm{sta}}} \left[\frac{\big(\varepsilon_h + R_{V_h}^{-1}B_hu_h, R_{V_h}^{-1}B_hu_h\big)_{V_h}}{\|R_{V_h}^{-1}B_hu_h\|_{V_h}}\right] && \text{(since } b_h(u_h,\varepsilon_h) = 0)\\
&= \frac{1}{C_{\mathrm{sta}}} \frac{\langle l_h, R_{V_h}^{-1}B_hu_h\rangle_{V_h^*,V_h}}{\|R_{V_h}^{-1}B_hu_h\|_{V_h}} \le \frac{1}{C_{\mathrm{sta}}}\|l_h\|_{V_h^*}. && \text{(by (27a))}
\end{aligned}$$

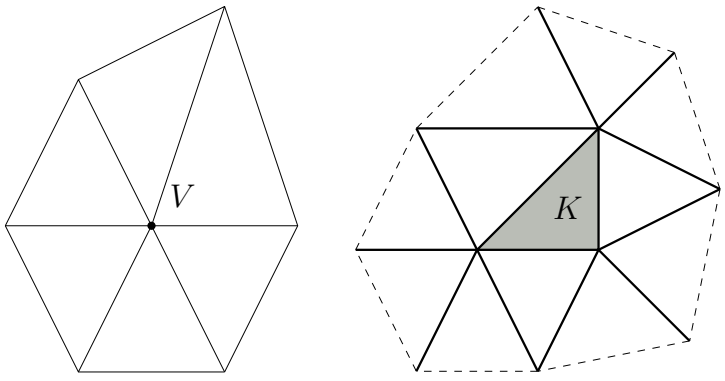
Finally, we prove the a priori error estimate (29). For any $z \in V_{h,\#}$, we define the projector $P_h : V_{h,\#} \to U_h$ by $P_hz = z_h$, where $z_h \in U_h$ is the second component of the solution (ε_h, z_h) of the saddle-point problem (27) with right-hand side $l_h(v_h) = b_h(z, v_h)$ (i.e., meaningful owing to Assumption 2). Using the a priori bound in (28) and the bound (18) in Assumption 3 leads to

$$\|P_hz\|_{V_h} = \|z_h\|_{V_h} \le \frac{1}{C_{\mathrm{sta}}}\|b_h(z,\cdot)\|_{V_h^*} \le \frac{C_{\mathrm{bnd}}}{C_{\mathrm{sta}}}\|z\|_{V_{h,\#}}. \tag{57}$$

Besides, $P_hz_h = z_h$ for any $z_h \in U_h$. Indeed, in that case, $(0, z_h)$ solves the corresponding saddle-point problem (27). To conclude, we observe that for the exact solution $u \in X_\#$ and the discrete solution $u_h \in U_h$, we have

$$\begin{aligned}
\|u - u_h\|_{V_h} &= \|(u - z_h) - P_h(u - z_h)\|_{V_h} && \text{(by definition of } P_h, \forall z_h \in U_h)\\
&\le \|u - z_h\|_{V_h} + \frac{C_{\mathrm{bnd}}}{C_{\mathrm{sta}}}\|u - z_h\|_{V_{h,\#}} && \text{(by the triangle inequality and (57))}\\
&\le \Big(1 + \frac{C_{\mathrm{bnd}}}{C_{\mathrm{sta}}}\Big)\|u - z_h\|_{V_{h,\#}}. && \text{(by Assumption 3)}
\end{aligned}$$

The result follows by taking the infimum over $z_h \in U_h$.



(a) Set D_V (b) Set $\mathscr{S}_K$ (thick lines)

Figure 16: Sets considered for the averaging operator and asspciated with a given node V (here a mesh vertex) and an element $K \in \mathscr{P}_h$

B Best-approximation in the upwinding norm

We introduce the well-known averaging operator (also known as Oswald interpolator) $\Pi_h^{\mathrm{av}} : \mathbb{P}_p(\mathscr{D}_h) \to \mathbb{P}_p(\mathscr{D}_h) \cap H_0^1(\Omega) = U_h$ such that, for any interpolation node $V \in \overline{\Omega}$,

$$\Pi_h^{\mathrm{av}}(v_h)(V) := \frac{1}{\mathrm{card}(\mathscr{D}_V)} \sum_{K \in \mathscr{D}_V} v_h|_K(V), \tag{58}$$

with $\mathscr{D}_V \subset \mathscr{D}_h$ denoting the union of the elements K sharing V as a common node (see the left panel of Figure 16a). In [34, 10, 28], it is shown that, for all $K \in \mathscr{P}_h$,

$$\left\| v_h - \Pi_h^{\mathrm{av}}(v_h) \right\|_K^2 \lesssim \sum_{e \in \mathscr{S}_K \cap \mathscr{S}_h^0} h_k \left\| [\![v_h]\!] \right\|_e^2, \tag{59}$$

with $\mathscr{S}_K$ denoting the mesh faces/edges having a non-empty intersection with ∂K (see the right panel of Figure 16b).

Let $v \in H^s(\Omega) \cap V$, $s > \frac{1}{2}$, and denote by $v_h \in V_h$ a function such that

$$\| v - v_h \|_{\mathrm{up},\#} := \inf_{z_h \in V_h} \| v - z_h \|_{\mathrm{up},\#}. \tag{60}$$

Let $v_h^* := \Pi_h^{\mathrm{av}}(v_h)$. Since $v_h^* \in U_h$, we have

$$\inf_{z_h \in U_h} \| v - z_h \|_{\mathrm{up},\#} \le \| v - v_h^* \|_{\mathrm{up},\#} \le \| v - v_h \|_{\mathrm{up},\#} + \| v_h - v_h^* \|_{\mathrm{up},\#}. \tag{61}$$

Therefore, we only need to show that $\| v_h - v_h^* \|_{\mathrm{up},\#} \lesssim \| v - v_h \|_{\mathrm{up},\#}$ to prove the claim. Using inverse and discrete trace inequalities, we infer that

$$\| v_h - v_h^* \|_{\mathrm{up},\#}^2 \lesssim \sum_{K \in \mathscr{P}_h} h_K^{-1} \| v_h - v_h^* \|_K^2. \tag{62}$$

Using (59) leads to

$$\|v_h - v_h^*\|_{\mathrm{up},\#}^2 \lesssim \sum_{e\in\mathscr{S}_h^0} \|[\![v_h]\!]\|_e^2. \tag{63}$$

Since $[\![v]\!] = 0$, for all $e \in \mathscr{S}_h^0$ (recall that $s > \frac{1}{2}$), we have $[\![v_h]\!] = [\![v_h - v]\!]$. We can now use the triangle inequality to decompose the jump into the two parts coming from the two cells sharing e, and re-arranging the terms leads to

$$\|v_h - v_h^*\|_{\mathrm{up},\#}^2 \lesssim \sum_{K\in\mathscr{P}_h} \|v - v_h\|_{\partial K}^2 \leq \|v - v_h\|_{\mathrm{up},\#}^2, \tag{64}$$

thereby completing the proof.

References

[1] Martin S Alnæs, Jan Blechta, Johan Hake, August Johansson, Benjamin Kehlet, Anders Logg, Chris Richardson, Johannes Ring, Marie E Rognes, and Garth N Wells. The fenics project version 1.5. *Archive of Numerical Software*, 3(100):9–23, 2015.

[2] Blanca Ayuso and L. Donatella Marini. Discontinuous Galerkin methods for advection-diffusion-reaction problems. *SIAM J. Numer. Anal.*, 47(2):1391–1420, 2009.

[3] Pascal Azérad and Jérôme Pousin. Inégalité de Poincaré courbe pour le traitement variationnel de l'équation de transport. *C. R. Acad. Sci. Paris Sér. I Math.*, 322(8):721–727, 1996.

[4] Randolph E Bank, Andrew H Sherman, and Alan Weiser. Some refinement algorithms and data structures for regular local mesh refinement. *Scientific Computing, Applications of Mathematics and Computing to the Physical Sciences*, 1:3–17, 1983.

[5] Randolph E Bank, Bruno D Welfert, and Harry Yserentant. A class of iterative methods for solving saddle point problems. *Numerische Mathematik*, 56(7):645–666, 1989.

[6] R. Becker and M. Braack. A finite element pressure gradient stabilization for the Stokes equations based on local projections. *Calcolo*, 38(4):173–199, 2001.

[7] J. H. Bramble and A. H. Schatz. Rayleigh-Ritz-Galerkin-methods for Dirichlet's problem using subspaces without boundary conditions. *Comm. Pure Appl. Math.*, 23:653–675, 1970.

[8] F. Brezzi, L. D. Marini, and E. Süli. Discontinuous Galerkin methods for first-order hyperbolic problems. *Math. Models Methods Appl. Sci.*, 14(12):1893–1903, 2004.

[9] E. Burman. A unified analysis for conforming and nonconforming stabilized finite element methods using interior penalty. *SIAM J. Numer. Anal.*, 43(5):2012–2033 (electronic), 2005.

[10] E. Burman and A. Ern. Continuous interior penalty hp-finite element methods for advection and advection-diffusion equations. *Math. Comp.*, 76(259):1119–1140, 2007.

[11] E. Burman and P. Hansbo. Edge stabilization for Galerkin approximations of convection-diffusion-reaction problems. *Comput. Methods Appl. Mech. Engrg.*, 193(15-16):1437–1453, 2004.

[12] P. Cantin. Well-posedness of the scalar and the vector advection–reaction problems in Banach graph spaces. *C. R. Math. Acad. Sci. Paris*, 355:892–902, 2017.

[13] Pierre Cantin and Alexandre Ern. An edge-based scheme on polyhedral meshes for vector advection-reaction equations. *ESAIM Math. Model. Numer. Anal.*, 51(5):1561–1581, 2017.

[14] J. Chan, J. A. Evans, and Weifeng Qiu. A dual Petrov–Galerkin finite element method for the convection–diffusion equation. *Comput. Math. Appl.*, 68(11):1513–1529, 2014.

[15] J. Chan, N. Heuer, T. Bui-Thanh, and L. Demkowicz. A robust DPG method for convection-dominated diffusion problems II: Adjoint boundary conditions and mesh-dependent test norms. *Comput. Math. Appl.*, 67(4):771–795, 2014.

[16] Yanqing Chen, Timothy A Davis, William W Hager, and Sivasankaran Rajamanickam. Algorithm 887: Cholmod, supernodal sparse cholesky factorization and update/downdate. *ACM Transactions on Mathematical Software (TOMS)*, 35(3):22, 2008.

[17] B. Cockburn, G. E. Karniadakis, and C.-W. Shu. *Discontinuous Galerkin Methods - Theory, Computation and Applications*, volume 11 of *Lecture Notes in Computer Science and Engineering*. Springer, 2000.

[18] A. Cohen, W. Dahmen, and G. Welper. Adaptivity and variational stabilization for convection-diffusion equations. *M2AN Math. Model. Numer. Anal.*, 46(5):1247–1273, 2012.

[19] L. Demkowicz and J. Gopalakrishnan. An overview of the discontinuous Petrov Galerkin method. In X. Feng, O. Karakashian, and Y. Xing, editors, *Recent Developments in Discontinuous Galerkin Finite Element Methods for Partial Differential Equations: 2012 John H Barrett Memorial Lectures*, volume 157 of *The IMA Volumes in Mathematics and its Applications*, pages 149–180. Springer, Cham, 2014.

[20] L. Demkowicz and N. Heuer. Robust DPG method for convection-dominated diffusion problems. *SIAM J. Numer. Anal.*, 51(5):2514–2537, 2013.

[21] Allen Devinatz, Richard Ellis, and Avner Friedman. The asymptotic behavior of the first real eigenvalue of second order elliptic operators with a small parameter in the highest derivatives. II. *Indiana Univ. Math. J.*, 23:991–1011, 1973–1974.

[22] Daniele Antonio Di Pietro and Alexandre Ern. *Mathematical aspects of discontinuous Galerkin methods*, volume 69. Springer Science, 2012.

[23] Willy Dörfler. A convergent adaptive algorithm for Poisson's equation. *SIAM Journal on Numerical Analysis*, 33(3):1106–1124, 1996.

[24] A.V. Džiškariani. The least square and Bubnov-Galerkin methods. *Ž. Vyčisl. Mat. i Mat. Fiz.*, 8:1110–1116, 1968.

[25] A. Ern and J.-L. Guermond. *Theory and practice of finite elements*, volume 159. Springer Science, 2004.

[26] A. Ern and J.-L. Guermond. Discontinuous Galerkin Methods for Friedrichs' Systems. I. General theory. *SIAM Journal on Numerical Analysis*, 44(2):753–778, 2006.

[27] A. Ern and J.-L. Guermond. Linear stabilization for first-order PDEs. In *Handbook of numerical methods for hyperbolic problems*, volume 17 of *Handb. Numer. Anal.*, pages 265–288. Elsevier/North-Holland, Amsterdam, 2016.

[28] A. Ern and J.-L. Guermond. Finite element quasi-interpolation and best approximation. *M2AN Math. Model. Numer. Anal.*, 51(4):1367–1385, 2017.

[29] J.-L. Guermond. Stabilization of Galerkin approximations of transport equations by subgrid modeling. *M2AN Math. Model. Numer. Anal.*, 33(6):1293–1316, 1999.

[30] T. J. R. Hughes, L. P. Franca, and G. M. Hulbert. A new finite element formulation for computational fluid dynamics: VIII. The Galerkin/Least-Squares method for advection-diffusive equations. *Comput. Methods Appl. Mech. Engrg.*, 73:173–189, 1989.

[31] Thomas J. R. Hughes, Guglielmo Scovazzi, and Leopoldo P. Franca. *Multiscale and Stabilized Methods*, pages 1–64. American Cancer Society, 2017.

[32] B.N. Jiang. *The Least-Squares Finite Element Method.* Springer, 1998.

[33] C. Johnson and J. Pitkäranta. An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation. *Math. Comp.*, 46(173):1–26, 1986.

[34] O. A. Karakashian and F. Pascal. A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems. *SIAM J. Numer. Anal.*, 41(6):2374–2399, 2003.

[35] Peter D Lax and Arthur N Milgram. Parabolic equations. *Selected Papers Volume I*, pages 8–31, 2005.

[36] P. Lesaint and P.-A. Raviart. On a finite element method for solving the neutron transport equation. In *Mathematical Aspects of Finite Elements in Partial Differential Equations*, pages 89–123. Publication No. 33. Math. Res. Center, Univ. of Wisconsin-Madison, Academic Press, New York, 1974.

[37] A.J. Lučka. The rate of convergence to zero of the residual and the error for the Bubnov-Galerkin method and the method of least squares. In *Proc. Sem. Differential and Integral Equations, No. I (Russian)*, pages 113–122. Akad. Nauk Ukrain. SSR Inst. Mat., Kiev, Ukraine, 1969.

[38] G. Matthies, P. Skrzypacz, and L. Tobiska. A unified convergence analysis for local projection stabilisations applied to the Oseen problem. *M2AN Math. Model. Numer. Anal.*, 41(4):713–742, 2007.

[39] W. H. Reed and T. R. Hill. Triangular mesh methods for the neutron transport equation. Technical Report LA-UR-73-0479, `http://lib-www.lanl.gov/cgi-bin/getfile?00354107.pdf`, Los Alamos Scientific Laboratory, Los Alamos, NM, 1973.

[40] Jeff Zitelli, Ignacio Muga, Leszek Demkowicz, Jayadeep Gopalakrishnan, David Pardo, and Victor M Calo. A class of discontinuous Petrov–Galerkin methods. Part IV: The optimal test norm and time-harmonic wave propagation in 1D. *Journal of Computational Physics*, 230(7):2406–2432, 2011.