MODELLISTICA E CALCOLO SCIENTIFICO

MODELING AND SCIENTIFIC COMPUTING

MOX–Report No. 07/2007

A Discontinuous Galerkin method with weighted averages for advection-diffusion equations with locally vanishing and anisotropic diffusivity

Alexandre Ern, Annette F. Stephansen,
Paolo Zunino

MOX, Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano, Via Bonardi 29 - 20133 Milano (Italy)

mox@mate.polimi.it http://mox.polimi.it

A Discontinuous Galerkin method with weighted averages for advection-diffusion equations with locally vanishing and anisotropic diffusivity

Alexandre Ern*, Annette F. Stephansen* and Paolo Zunino**

20th March 2007

* Cermics, Ecole des Ponts, ParisTech
6 et 8 avenue Blaise Pascal, Champs sur Marne
77455 Marne la Vallée Cedex 2, France.

** MOX– Modellistica e Calcolo Scientifico
Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano
via Bonardi 9, 20133 Milano, Italy.

Keywords: discontinuous Galerkin, vanishing diffusion, anisotropic diffusion

Abstract

We consider Discontinuous Galerkin approximations of advection-diffusion equations with anisotropic and discontinuous diffusivity, and propose the symmetric weighted interior penalty (SWIP) method for better coping with locally vanishing diffusivity. The analysis yields convergence results for the natural energy norm that are optimal (with respect to mesh-size) and robust (fully independent of the diffusivity). The convergence results for the advective derivative are optimal with respect to mesh-size and robust for isotropic diffusivity, as well as for anisotropic diffusivity in the dominant advection regime. In the dominant diffusivity regime, an optimal convergence result for the the L^2-norm is also recovered. Numerical results are presented to illustrate the performance of the scheme.

1 Introduction

Since their introduction over thirty years ago [19, 16], Discontinuous Galerkin (DG) methods have emerged as an attractive tool to approximate numerous PDEs in the engineering sciences. Here we are primarily interested in advection–diffusion equations with anisotropic (e.g., tensor-valued) and heterogeneous (e.g., non-smooth) diffusivity. Such equations are encountered, for instance, in groundwater flow models which constitute the motivation for the present work.

The analysis of DG methods to approximate advection–diffusion equations is extensively covered in [15]. This work already addresses anisotropic and heterogeneous diffusivity. However, one particular aspect that deserves further attention is the locally vanishing diffusivity case, i.e., the limiting case where the diffusivity becomes arbitrarily small in *some* parts of the computational domain. Indeed, in this case it is well-known that the presence of an advective field can trigger internal layers. Specifically, in the locally vanishing diffusivity limit, the solution becomes discontinuous on the interfaces where the normal component of the advective field measured from the vanishing-diffusivity region towards the nonvanishing-diffusivity region is nonnegative. This situation has been analyzed in [10] and, more recently, in [5]. In the presence of internal layers resulting from vanishing diffusivity, all the usual DG methods meet with difficulties since they have been designed to weakly enforce continuity of the discrete solution across mesh interfaces. One possible remedy is to modify the DG method at the interfaces affected by internal layers, as already proposed in [15] and, more recently, in [9]. However, this approach is not fully satisfactory since it requires *a priori* knowledge of the interface in question. For simple problems the interface is easy to locate, but it can become difficult whenever nonlinear models with solution-dependent diffusivity are used, or in the presence of free-boundary problems.

The aim of the present work is to design a DG method that can handle internal layers resulting from locally vanishing diffusion in an automated fashion. The key ingredient is the use of weighted instead of arithmetic averages in the design of certain terms in the DG method. The idea of utilizing weighted averages stems from the mortar finite-element method originally proposed by Nitsche [17, 18]. This method imposes weakly the continuity of fluxes between different regions. Various authors have highlighted the possibility of using an average with weights that differ from one half; see [21, 14, 12, 13] where several mortaring techniques are presented to match conforming finite elements on possibly nonconforming computational meshes. The weighted averages are introduced as a generalization of the standard average and the analysis is carried out in the general framework. However, the cited works do not consider any connection between the weights and the coefficients of the problem. This dependency was investigated recently in [3] for isotropic advection–diffusion problems, using a weighted interior penalty technique with mortars which, when applied elementwise, yields a DG method. It was shown in [3] that a specific choice of weights improves the stability of the numerical scheme in the locally vanishing diffusivity limit. The reason why weighted averages are needed to properly handle internal layers is rooted in the dissipative structure of the underlying Friedrichs's system. The design of the corresponding DG bilinear form, where dissipation at the discrete level is enforced by a consistency term involving averages, has been recently proposed in [8] for the general case, and in [6] for advection–diffusion equations in the locally vanishing diffusivity limit.

In the present work, we extend the DG method implicitly derived in [3] for isotropic diffusivity to anisotropic problems. This task is not as simple as it may appear on first sight since the presence of internal layers now depends on the spectral structure of the diffusion tensor on both sides of each mesh interface. The spectral structure also raises the question of the appropriate choice of the penalty term at each mesh interface. The analysis presented below will tackle these issues.

We design and analyze one specific DG method with weighted averages, namely the Symmetric Weighted Interior Penalty (SWIP) method, obtained by modifying the well-known (Symmetric) Interior Penalty (IP) method [2, 1]. Many other well-known DG methods, including the Local Discontinuous Galerkin method [4] and the Nonsymmetric Interior Penalty Galerkin method [20], can also be modified to fit the present scope; for brevity, these developments are omitted herein.

This paper is organized as follows: Section 2 presents the setting under scrutiny and formulates the SWIP method, while Section 3 contains the error analysis in the natural energy norm for the problem. The error estimate is robust, with respect to both locally vanishing and anisotropic diffusivity. Section 4 is concerned with the error analysis of the advective derivative. The estimate is again robust with respect to locally vanishing diffusivity, but the constant can in some cases depend on local anisotropies. Numerical results are presented in Section 5 and illustrate the benefits of using weighted interior penalties to approximate advection–diffusion equations with locally vanishing and anisotropic diffusivity. Finally, Section 6 contains some concluding remarks.

2 The SWIP method

Let Ω be a domain in $\mathbb{R}^d$ with boundary $\partial\Omega$ in space dimension $d \in \{2,3\}$. We consider the following advection-diffusion equation with homogeneous Dirichlet boundary conditions:

$$\begin{cases} -\nabla\cdot(K\nabla u) + \beta\cdot\nabla u + \mu u = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{1}$$

Here $\mu \in L^\infty(\Omega)$, $\beta \in [W^{1,\infty}(\Omega)]^d$, the diffusion tensor K is a symmetric, positive definite field in $[L^\infty(\Omega)]^{d,d}$ and $f \in L^2(\Omega)$. The regularity assumption on β can be relaxed, but is sufficient for the present purpose. The weak formulation of (1) consists of finding $u \in H^1_0(\Omega)$ such that

$$(K\nabla u, \nabla v)_{0,\Omega} + (\beta\cdot\nabla u, v)_{0,\Omega} + (\mu u, v)_{0,\Omega} = (f, v)_{0,\Omega} \quad \forall v \in H^1_0(\Omega) \tag{2}$$

where $(\cdot,\cdot)_{0,\Omega}$ denotes the L^2-scalar product on Ω. Henceforth, we assume that

$$\mu - \tfrac{1}{2}\nabla\cdot\beta \geq \mu_0 > 0 \qquad \text{a.e in } \Omega. \tag{3}$$

Furthermore, we assume that the smallest eigenvalue of K is bounded from below by a positive (but possibly very small) constant. Then, owing to the Lax–Milgram Lemma, (2) is well–posed.

Let $\{\mathcal{T}_h\}_{h>0}$ be a shape-regular family of affine triangulations of the domain Ω. The meshes $\mathcal{T}_h$ may possess hanging nodes. For simplicity we assume that the meshes cover Ω exactly, i.e., Ω is a polyhedron. A generic element in $\mathcal{T}_h$ is denoted by T, h_T denotes the diameter of T and n_T its outward unit normal. Set $h = \max_{T\in\mathcal{T}_h} h_T$. We assume without loss of generality that $h \leq 1$. Let $p \geq 1$. We define the classical DG approximation space

$$V_h = \{v_h \in L^2(\Omega); \forall T \in \mathcal{T}_h, v_h|_T \in \mathbb{P}_p\} \tag{4}$$

where $\mathbb{P}_p$ is the set of polynomials of total degree less than or equal to p. Henceforth, we assume that the diffusivity tensor K is piecewise constant on $\mathcal{T}_h$. This assumption, which is reasonable in the context of groundwater flow models, can be generalized by assuming a smooth enough behaviour of K inside each mesh element. For the sake of simplicity, these technicalities are avoided.

We say that F is an interior face of the mesh if there are $T^-(F)$ and $T^+(F)$ in $\mathcal{T}_h$ such that $F = T^-(F) \cap T^+(F)$. We set $\mathcal{T}(F) = \{T^-(F), T^+(F)\}$ and let n_F be the unit normal vector to F pointing from $T^-(F)$ towards $T^+(F)$. The analysis hereafter does not depend on the arbitrariness of this choice. Similarly, we say that F is a boundary face of the mesh if there is $T(F) \in \mathcal{T}_h$ such that $F = T(F) \cap \partial\Omega$. We set $\mathcal{T}(F) = \{T(F)\}$ and let n_F coincide with the outward normal to $\partial\Omega$. All the interior (resp., boundary) faces of the mesh are collected into the set $\mathcal{F}_h^i$ (resp., $\mathcal{F}_h^{\partial\Omega}$) and we let $\mathcal{F}_h = \mathcal{F}_h^i \cup \mathcal{F}_h^{\partial\Omega}$. Henceforth, we shall often deal with functions that are double-valued on $\mathcal{F}_h^i$ and single-valued on $\mathcal{F}_h^{\partial\Omega}$. This is the case, for instance, of functions in V_h. On interior faces, when the two branches of the function in question, say v, are associated with restrictions to the neighboring elements $T^{\mp}(F)$, these branches are denoted by $v^{\mp}$ and the jump of v across F is defined as

$$[\![v]\!]_F = v^- - v^+. \tag{5}$$

On a boundary face $F \in \mathcal{F}^{\partial\Omega}$, we set $[\![v]\!]_F = v|_F$. Furthermore, on an interior face $F \in \mathcal{F}_h^i$, we define the standard (arithmetic) average as $\{v\}_F = \frac{1}{2}(v^- + v^+)$. The subscript F in the above jumps and averages is omitted if there is no ambiguity.

The L^2-scalar product and its associated norm on a region $R \subset \Omega$ are indicated by the subscript $0, R$. For $s \geq 1$, a norm (seminorm) with the subscript s, R designates the usual norm (seminorm) in $H^s(R)$. When the region R is the boundary of a mesh element ∂T and the arguments in the scalar product or the norm are double-valued functions, it is implicitly assumed that the value considered is that of the branch associated with the restriction to T. For $s \geq 1$, $H^s(\mathcal{T}_h)$ denotes the usual broken Sobolev space on $\mathcal{T}_h$ and for $v \in H^1(\mathcal{T}_h)$, $\nabla_h v$ denotes the piecewise gradient of v, that is, $\nabla_h v \in [L^2(\Omega)]^d$ and for all $T \in \mathcal{T}_h$, $(\nabla_h v)|_T = \nabla(v|_T)$. It is also convenient to set $V(h) = H^2(\mathcal{T}_h) + V_h$.

The formulation of the SWIP method requires two parameters. As in the formulation of the usual IP method we introduce a single- and scalar-valued function γ defined on $\mathcal{F}_h$. The purpose of this function is to penalize jumps across interior faces and values at boundary faces. Additionally, we define a scalar- and double-valued function ω on $\mathcal{F}_h^i$. This function, which is not present in the usual IP method, is used to evaluate weighted averages of diffusive fluxes. On an interior face $F \in \mathcal{F}_h^i$, the values taken by the two branches of ω are denoted by $(\omega|_F)^{\mp}$, or simply $\omega^{\mp}$ if there is no ambiguity.

Henceforth, it is assumed that for all $F \in \mathcal{F}_h^i$,

$$\omega^- + \omega^+ = 1. \tag{6}$$

For $v \in V(h)$, we define the weighted average of the diffusive flux $K\nabla_h v$ on an interior face $F \in \mathcal{F}_h^i$ as

$$\{K\nabla_h v\}_\omega = \omega^-(K\nabla_h v)^- + \omega^+(K\nabla_h v)^+. \tag{7}$$

For convenience, we extend the above definitions to boundary faces as follows: on $F \in \mathcal{F}_h^{\partial\Omega}$, ω is single-valued and equal to 1, and we set $\{K\nabla v\}_\omega = K\nabla v$.

The SWIP bilinear form $B_h(\cdot,\cdot)$ is defined on $V(h) \times V(h)$ as follows

$$\begin{aligned} B_h(v,w) &= (K\nabla_h v, \nabla_h w)_{0,\Omega} + ((\mu - \nabla\cdot\beta)v, w)_{0,\Omega} - (v, \beta\cdot\nabla_h w)_{0,\Omega} \\ &\quad + \sum_{F\in\mathcal{F}_h} \big((\gamma[\![v]\!], [\![w]\!])_{0,F} - (n_F^t\{K\nabla_h v\}_\omega, [\![w]\!])_{0,F} - (n_F^t\{K\nabla_h w\}_\omega, [\![v]\!])_{0,F} \big) \\ &\quad + \sum_{F\in\mathcal{F}_h^i} (\beta\cdot n_F\{v\}, [\![w]\!])_{0,F} + \sum_{F\in\mathcal{F}_h^{\partial\Omega}} \tfrac{1}{2}(\beta\cdot n_F v, w)_{0,F}. \end{aligned} \tag{8}$$

The SWIP bilinear form can equivalently be expressed, after integrating the advective derivative by parts, as

$$\begin{aligned} B_h(v,w) &= (K\nabla_h v, \nabla_h w)_{0,\Omega} + (\mu v, w)_{0,\Omega} + (\beta\cdot\nabla_h v, w)_{0,\Omega} \\ &\quad + \sum_{F\in\mathcal{F}_h} \big((\gamma[\![v]\!], [\![w]\!])_{0,F} - (n_F^t\{K\nabla_h v\}_\omega, [\![w]\!])_{0,F} - (n_F^t\{K\nabla_h w\}_\omega, [\![v]\!])_{0,F} \big) \\ &\quad - \sum_{F\in\mathcal{F}_h^i} (\beta\cdot n_F\{w\}, [\![v]\!])_{0,F} - \sum_{F\in\mathcal{F}_h^{\partial\Omega}} \tfrac{1}{2}(\beta\cdot n_F v, w)_{0,F}. \end{aligned} \tag{9}$$

Both (8) and (9) will be used in the analysis. The discrete problem consists of finding $u_h \in V_h$ such that

$$B_h(u_h, v_h) = (f, v_h)_{0,\Omega} \qquad \forall v_h \in V_h. \tag{10}$$

The penalty parameter γ is defined as

$$\forall F \in \mathcal{F}_h, \qquad \gamma = \alpha\frac{\gamma_K}{h_F} + \gamma_\beta, \tag{11}$$

where α is a positive scalar (α can also vary from face to face) and where

$$\forall F \in \mathcal{F}_h^i, \qquad \gamma_K = (\omega^-)^2\delta_{Kn}^- + (\omega^+)^2\delta_{Kn}^+ \tag{12}$$

$$\forall F \in \mathcal{F}_h^{\partial\Omega}, \qquad \gamma_K = \delta_{Kn}, \tag{13}$$

$$\forall F \in \mathcal{F}_h, \qquad \gamma_\beta = \tfrac{1}{2}|\beta\cdot n_F|, \tag{14}$$

with $\delta_{Kn}^\mp = n_F^t K^\mp n_F$ if $F \in \mathcal{F}_h^i$ and $\delta_{Kn} = n_F^t K n_F$ if $F \in \mathcal{F}_h^{\partial\Omega}$. Note that the choice for γ_β amounts to the usual upwind scheme to stabilize the advective derivative.

For the error analysis in the energy norm (see Section 3), no other assumption than (6) is made for the weights. In particular, it is possible to choose $\omega^{\mp} = \frac{1}{2}$, in which case the SWIP bilinear form B_h reduces to the standard IP bilinear form with the penalty parameter scaling as the standard average of the diffusion in the normal direction; this method has been analyzed in [11]. Note also that the choice made in [15] for the penalty parameter is different since it involves the maximum eigenvalue of K.

For the error analysis in the advective derivative (see Section 4), a specific choice of the weights differing from $\omega^{\mp} = \frac{1}{2}$ has to be made to yield robust error estimates with respect to the diffusivity. Specifically, we shall set

$$\omega^- = \frac{\delta_{Kn}^+}{\delta_{Kn}^+ + \delta_{Kn}^-}, \qquad \omega^+ = \frac{\delta_{Kn}^-}{\delta_{Kn}^+ + \delta_{Kn}^-}, \tag{15}$$

and thus

$$\forall F \in \mathcal{F}_h^i, \qquad \gamma_K = \frac{\delta_{Kn}^+ \delta_{Kn}^-}{\delta_{Kn}^+ + \delta_{Kn}^-}. \tag{16}$$

Note that with this choice $\gamma_K = \omega^- \delta_{Kn}^- = \omega^+ \delta_{Kn}^+$, and that $2\gamma_K$ is the harmonic average of the normal component of the diffusion tensor across the interface. Observe also that $\gamma_K \leq \inf(\delta_{Kn}^-, \delta_{Kn}^+)$, a point that becomes important to ensure even the consistency of the method when the diffusivity is actually allowed to vanish locally, see [6].

3 Error analysis in the energy norm

The goal of this section is to establish an error estimate for the SWIP method in the energy norm, the estimate being robust with respect to both locally vanishing and anisotropic diffusion. The analysis is performed by establishing coercivity, consistency and continuity properties for the SWIP bilinear form in the spirit of Strang's Second Lemma [7].

Without loss of generality, we assume that the problem data have been normalized so that $\|\beta\|_{[W^{1,\infty}(\Omega)]^d}$ is of order unity. We also assume that $\|\mu\|_{L^\infty(\Omega)} \leq 1$ since we are not interested in strong reaction regimes. In the sequel, the symbol $\lesssim$ indicates an inequality involving a positive constant C independent of the size of the mesh family and of the diffusion tensor. The constant may depend on the advection field β, the reaction term μ, and the shape-regularity parameter of the mesh family. In the analysis we will make use of the following inverse trace and inverse inequalities: For all $T \in \mathcal{T}_h$ and for all $v_h \in V_h$,

$$\|v_h\|_{0,\partial T} \lesssim h_T^{-\frac{1}{2}} \|v_h\|_{0,T}, \tag{17}$$

$$\|\nabla_h v_h\|_{0,T} \lesssim h_T^{-1} \|v_h\|_{0,T}, \tag{18}$$

which result from the shape regularity of the mesh family $\{\mathcal{T}_h\}_{h>0}$.

For a function $v \in V(h)$, we consider the following jump seminorms

$$|[\![v]\!]|_\sigma^2 = \sum_{F\in\mathcal{F}_h} |[\![v]\!]|_{\sigma,F}^2, \qquad |[\![v]\!]|_{\sigma,F}^2 = (\sigma[\![v]\!],[\![v]\!])_{0,F}, \tag{19}$$

with $\sigma := \gamma_\beta$, $\sigma := \gamma_K$ or $\sigma := \gamma$. The natural energy norm with which to equip $V(h)$ is

$$\|v\|_{h,B} = \|v\|_{0,\Omega} + \|\kappa\nabla_h v\|_{0,\Omega} + |[\![v]\!]|_\gamma \tag{20}$$

where κ denotes the (unique) symmetric positive definite tensor-valued field such that $\kappa^2 = K$ a.e. in Ω.

Lemma 3.1. (Coercivity) *Assume that α in* (11) *is large enough. Then, the bilinear form B_h is $\|\cdot\|_{h,B}$-coercive, i.e., for all $v_h \in V_h$,*

$$B_h(v_h,v_h) \gtrsim \|v_h\|_{h,B}^2. \tag{21}$$

Proof. Let $v_h \in V_h$. Taking $v = w = v_h$ in (8) yields

$$\begin{aligned} B_h(v_h,v_h) &= \|\kappa\nabla_h v_h\|_{0,\Omega}^2 + (\mu v_h,v_h)_{0,\Omega} - ((\nabla\cdot\beta)v_h,v_h)_{0,\Omega} - (v_h,\beta\cdot\nabla_h v_h)_{0,\Omega} \\ &\quad + |[\![v_h]\!]|_\gamma^2 - \sum_{F\in\mathcal{F}_h} 2(n_F^t\{K\nabla v_h\}_\omega,[\![v_h]\!])_{0,F} \\ &\quad + \sum_{F\in\mathcal{F}_h^i} (\beta\cdot n_F\{v_h\},[\![v_h]\!])_{0,F} + \sum_{F\in\mathcal{F}_h^{\partial\Omega}} \tfrac{1}{2}(\beta\cdot n_F v_h,v_h)_{0,F}. \end{aligned} \tag{22}$$

Integrating by parts the fourth term on the right hand side of (22) and owing to hypothesis (3), we obtain

$$\begin{aligned} &(\mu v_h,v_h)_{0,\Omega} - ((\nabla\cdot\beta)v_h,v_h)_{0,\Omega} - (v_h,\beta\cdot\nabla_h v_h)_{0,\Omega} + \sum_{F\in\mathcal{F}_h^i} (\beta\cdot n_F\{v_h\},[\![v_h]\!])_{0,F} \\ &+ \sum_{F\in\mathcal{F}_h^{\partial\Omega}} \tfrac{1}{2}(\beta\cdot n_F v_h,v_h)_{0,F} = ((\mu - \tfrac{1}{2}\nabla\cdot\beta)v_h,v_h)_{0,\Omega} \gtrsim \|v_h\|_{0,\Omega}^2. \end{aligned} \tag{23}$$

Consider now the sixth term in the right-hand side of (22). Let $F \in \mathcal{F}_h$. First, observe that owing to Young's inequality

$$\begin{aligned} |2(n_F^t\omega^\mp(K\nabla_h v_h)^\mp,[\![v_h]\!])_{0,F}| &= |2((\kappa\nabla_h v_h)^\mp,\omega^\mp\kappa^\mp n_F[\![v_h]\!])_{0,F}| \\ &\le h_F\alpha_0\|(\kappa\nabla_h v_h)^\mp\|_{0,F}^2 + \frac{1}{\alpha_0}\left(\frac{(\omega^\mp)^2\delta_{Kn}^\mp}{h_F}[\![v_h]\!],[\![v_h]\!]\right)_{0,F} \end{aligned}$$

where $\alpha_0 > 0$ can be chosen as small as needed. Using the trace inverse inequality (17) and the definition of γ_K (12)-(13) yields

$$|2(n_F^t\{K\nabla_h v_h\}_\omega,[\![v_h]\!])_{0,F}| \lesssim \alpha_0\|\kappa\nabla_h v_h\|_{0,\mathcal{T}(F)}^2 + \frac{1}{\alpha_0 h_F}|[\![v_h]\!]|_{\gamma_K,F}^2.$$

Choosing α in (11) to be large enough yields

$$\|\kappa\nabla_h v_h\|_{0,\Omega}^2 + |[\![v_h]\!]|_\gamma^2 - \sum_{F\in\mathcal{F}_h} 2(n_F^t\{K\nabla_h v_h\}_\omega, [\![v_h]\!])_{0,F} \gtrsim \|\kappa\nabla_h v_h\|_{0,\Omega}^2 + |[\![v_h]\!]|_\gamma^2. \tag{24}$$

Combining (24) with (23) we obtain (21). □

Lemma 3.2. (Consistency) *Let u solve* (2) *and let u_h solve* (10). *Assume that $u \in H^2(\mathcal{T}_h)$. Then*

$$\forall v_h \in V_h, \qquad B_h(u-u_h, v_h) = 0 \tag{25}$$

Proof. Let $v_h \in V_h$. Since u is continuous by assumption and vanishes on $\partial\Omega$, using (9) yields

$$B_h(u, v_h) = (K\nabla u, \nabla_h v_h)_{0,\Omega} + (\mu u, v_h)_{0,\Omega} + (\beta\cdot\nabla u, v_h)_{0,\Omega} \\ - \sum_{F\in\mathcal{F}_h} (n_F^t\{K\nabla u\}_\omega, [\![v_h]\!])_{0,F}.$$

Using the fact that $n_F^t K\nabla u$ is continuous on interior faces yields $n_F^t\{K\nabla u\}_\omega = (\omega^- + \omega^+)n_F^t K\nabla u = n_F^t K\nabla u$ owing to (6). Hence, integrating by parts leads to

$$(K\nabla u, \nabla_h v_h)_{0,\Omega} - \sum_{F\in\mathcal{F}_h} (n_F^t\{K\nabla u\}_\omega, [\![v_h]\!])_{0,F} = -\sum_{T\in\mathcal{T}_h} (\nabla\cdot(K\nabla u), v_h)_{0,T}.$$

As a result,

$$B_h(u, v_h) = \sum_{T\in\mathcal{T}_h} (-\nabla\cdot(K\nabla u) + \beta\cdot\nabla u + \mu u, v_h)_{0,T} = (f, v_h)_{0,\Omega} = B_h(u_h, v_h),$$

yielding (25). □

We now establish a continuity property for the SWIP bilinear form B_h. To this purpose, we introduce on $V(h)$ the norm

$$\|v\|_{h,\frac{1}{2}} = \|v\|_{h,B} + \left(\sum_{T\in\mathcal{T}_h} \|v\|_{0,\partial T}^2\right)^{\frac{1}{2}} + \left(\sum_{F\in\mathcal{F}_h} h_F \|\kappa\nabla_h v\|_{0,F}^2\right)^{\frac{1}{2}}. \tag{26}$$

Let $V_h^\perp = \{v \in V(h), \forall v_h \in V_h, (v, v_h)_{0,\Omega} = 0\}$.

Lemma 3.3. (Continuity) *The following holds:*

$$\forall (v, w_h) \in V_h^\perp \times V_h, \qquad |B_h(v, w_h)| \lesssim \|v\|_{h,\frac{1}{2}} \|w_h\|_{h,B}. \tag{27}$$

Proof. Let $(v, w_h) \in V_h^{\perp} \times V_h$. The first two terms in (8) are easily bounded as

$$|(K\nabla_h v, \nabla_h w_h)_{0,\Omega}| + |((\mu - \nabla\cdot\beta)v, w_h)_{0,\Omega}| \lesssim \|v\|_{h,B}\|w_h\|_{h,B}.$$

To bound the third term, let $\overline{\beta}$ be the piecewise constant, vector-valued field equal to the mean value of β on each $T \in \mathcal{T}_h$. Then,

$$\begin{aligned}(v, \beta\cdot\nabla_h w_h)_{0,\Omega} &= (v, \overline{\beta}\cdot\nabla_h w_h)_{0,\Omega} + (v, (\beta - \overline{\beta})\cdot\nabla_h w_h)_{0,\Omega} \\ &= (v, (\beta - \overline{\beta})\cdot\nabla_h w_h)_{0,\Omega}\end{aligned}$$

since $\overline{\beta}\cdot\nabla_h w_h \in V_h$ and $v \in V_h^{\perp}$. Moreover, since $\beta \in [W^{1,\infty}(\Omega)]^d$,

$$\forall T \in \mathcal{T}_h, \qquad \|\beta - \overline{\beta}\|_{[L^\infty(T)]^d} \lesssim h_T$$

so that the inverse inequality (18) yields

$$|(v, \beta\cdot\nabla_h w_h)_{0,\Omega}| \lesssim \|v\|_{0,\Omega}\|w_h\|_{0,\Omega} \leq \|v\|_{h,B}\|w_h\|_{h,B}.$$

Furthermore, proceeding as in the proof of Lemma 3.1 yields, for all $F \in \mathcal{F}_h$,

$$|(n_F^t\{K\nabla_h v\}_\omega, [\![w_h]\!])_{0,F}| \lesssim h_F^{\frac{1}{2}}\|\kappa\nabla_h v\|_{0,F}\, h_F^{-\frac{1}{2}}|[\![w_h]\!]|_{\gamma_K,F}$$

and

$$|(n_F^t\{K\nabla_h w_h\}_\omega, [\![v]\!])_{0,F}| \lesssim h_F^{-\frac{1}{2}}|[\![v]\!]|_{\gamma_K,F}\|\kappa\nabla_h w_h\|_{0,T(F)}$$

so that

$$\sum_{F\in\mathcal{F}_h}\left(|(n_F^t\{K\nabla v\}_\omega, [\![w_h]\!])_{0,F}| + |(n_F^t\{K\nabla w_h\}_\omega, [\![v]\!])_{0,F}|\right) \lesssim \|v\|_{h,\frac{1}{2}}\|w_h\|_{h,B}.$$

For the remaining terms, we obtain

$$\begin{aligned}&\sum_{F\in\mathcal{F}_h}|(\gamma[\![v]\!], [\![w_h]\!])_{0,F}| + \sum_{F\in\mathcal{F}_h^i}|(\beta\cdot n_F\{v\}, [\![w_h]\!])_{0,F}| + \sum_{F\in\mathcal{F}_h^{\partial\Omega}}|\tfrac{1}{2}(\beta\cdot n_F v, w_h)_{0,F}| \\ &\qquad\lesssim |[\![v]\!]|_\gamma|[\![w_h]\!]|_\gamma + \sum_{F\in\mathcal{F}_h^i}\|\{v\}\|_{0,F}|[\![w_h]\!]|_{\gamma_\beta,F} \leq \|v\|_{h,\frac{1}{2}}\|w_h\|_{h,B}.\end{aligned}$$

This completes the proof since $\|\cdot\|_{h,B} \leq \|\cdot\|_{h,\frac{1}{2}}$. □

Theorem 3.1. *Let $\Pi_h u$ be the L^2-projection of u onto V_h. Then,*

$$\|u - u_h\|_{h,B} \lesssim \|u - \Pi_h u\|_{h,\frac{1}{2}}. \tag{28}$$

Proof. Owing to Lemmata 3.1, 3.2 and 3.3,

$$\|u_h - \Pi_h u\|_{h,B} \lesssim \frac{B_h(u_h - \Pi_h u, u_h - \Pi_h u)}{\|u_h - \Pi_h u\|_{h,B}} \lesssim \frac{B_h(u - \Pi_h u, u_h - \Pi_h u)}{\|u_h - \Pi_h u\|_{h,B}}$$
$$\lesssim \|u - \Pi_h u\|_{h,\frac{1}{2}}. \tag{29}$$

We complete the proof by applying the triangle inequality and using the fact that $\|\cdot\|_{h,B} \leq \|\cdot\|_{h,\frac{1}{2}}$. □

Corollary 3.1. *Set $\lambda_{M,K} := \max(1, \lambda_K)$, where λ_K indicates the maximum eigenvalue of K on Ω. Then, if the exact solution u is in $H^{p+1}(\mathcal{T}_h)$,*

$$\|u - u_h\|_{h,B} \lesssim \lambda_{M,K}^{\frac{1}{2}} h^p \|u\|_{H^{p+1}(\mathcal{T}_h)}. \tag{30}$$

Proof. Use Theorem 3.1 and standard approximation properties for the L^2-orthogonal projector Π_h. □

We now prove that when the domain Ω has elliptic regularity and the diffusion is not too small, the error estimate in the L^2-norm can be improved by using the Aubin-Nitsche duality argument. To this purpose, we introduce the following dual problem: seek $\psi \in H_0^1(\Omega)$ such that

$$(K\nabla v, \nabla\psi)_{0,\Omega} + (\beta\cdot\nabla v, \psi)_{0,\Omega} + (\mu v, \psi)_{0,\Omega} = (v, u - u_h)_{0,\Omega} \quad \forall v \in H_0^1(\Omega). \tag{31}$$

We assume that elliptic regularity holds in the broken H^2-norm, namely that

$$\|\psi\|_{H^2(\mathcal{T}_h)} \lesssim \lambda_{m,K}^{-1} \|u - u_h\|_{0,\Omega} \tag{32}$$

where $\lambda_{m,K}$ denotes the lowest eigenvalue of K on Ω.

Theorem 3.2. *In the above framework,*

$$\|u - u_h\|_{0,\Omega} \leq \frac{\lambda_{M,K}^{\frac{1}{2}}}{\lambda_{m,K}} h \left(\|u - u_h\|_{h,B} + \inf_{w_h \in V_h} \|u - w_h\|_{h,B_+} \right) \tag{33}$$

where for all $v \in V(h)$,

$$\|v\|_{h,B_+} = \|v\|_{h,B} + \left(\sum_{T\in\mathcal{T}_h} h_T^2 \|\nabla_h v\|_{0,T}^2 \right)^{\frac{1}{2}} + \left(\sum_{F\in\mathcal{F}_h} h_F \|\kappa\nabla_h v\|_{0,F}^2 \right)^{\frac{1}{2}}. \tag{34}$$

Proof. Step (i): observe that for all $v \in V(h)$, using (8) yields

$$B_h(v,\psi) = (K\nabla_h v, \nabla\psi)_{0,\Omega} + ((\mu - \nabla\cdot\beta)v, \psi)_{0,\Omega} - (v, \beta\cdot\nabla\psi)_{0,\Omega}$$
$$- \sum_{F\in\mathcal{F}_h} (n_F^t\{K\nabla\psi\}_\omega, [\![v]\!])_{0,F} = \sum_{T\in\mathcal{T}_h} (v, -\nabla\cdot(K\nabla\psi) - \beta\cdot\nabla\psi + (\mu - \nabla\cdot\beta)\psi)_{0,T}$$
$$= (v, u - u_h)_{0,\Omega}. \tag{35}$$

Step (ii): define on $V(h)$ the norm

$$\|v\|_{h,1} = \|v\|_{h,\frac{1}{2}} + \left(\sum_{T\in\mathcal{T}_h} h_T^{-2} \|v\|_{0,T}^2 \right)^{\frac{1}{2}} \tag{36}$$

and let us prove that for all $(v,w) \in V(h) \times V(h)$,

$$|B_h(v,w)| \lesssim \|v\|_{h,B_+} \|w\|_{h,1}. \tag{37}$$

Indeed, indicating by T_i, $1 \le i \le 8$ the eight terms on the right-hand side of (9), and proceeding as in the proof of Lemma 3.3, it is clear that $\sum_{i\neq 3} |T_i| \lesssim \|v\|_{h,B_+} \|w\|_{h,\frac{1}{2}}$. Moreover,

$$\begin{aligned} |T_3| = |(\beta{\cdot}\nabla_h v, w)_{0,\Omega}| &\lesssim \sum_{T\in\mathcal{T}_h} \|\nabla_h v\|_{0,T} \|w\|_{0,T} \\ &= \sum_{T\in\mathcal{T}_h} h_T \|\nabla_h v_h\|_{0,T} h_T^{-1} \|w\|_{0,T} \le \|v\|_{h,B_+} \|w\|_{h,1}. \end{aligned}$$

Hence, (37) holds.

Step (iii): taking $v = u - u_h$ in (35), applying Lemma 3.2 and using (37) yields for all $\psi_h \in V_h$,

$$\|u - u_h\|_{0,\Omega}^2 = B_h(u-u_h, \psi) = B_h(u-u_h, \psi - \psi_h) \lesssim \|u-u_h\|_{h,B_+} \|\psi - \psi_h\|_{h,1}.$$

Using standard interpolation results leads to

$$\inf_{\psi_h\in V_h} \|\psi - \psi_h\|_{h,1} \lesssim \lambda_{M,K}^{\frac{1}{2}} h \|\psi\|_{H^2(\mathcal{T}_h)},$$

and taking into account (32) yields

$$\|u - u_h\|_{0,\Omega} \lesssim \frac{\lambda_{M,K}^{\frac{1}{2}}}{\lambda_{m,K}} h \|u - u_h\|_{h,B_+}. \tag{38}$$

Using the inverse inequalities (17) and (18), we infer that for all $v_h \in V_h$,

$$\|v_h\|_{h,B_+} \lesssim \|v_h\|_{h,B} + \|v_h\|_{0,\Omega} + \|\kappa \nabla_h v_h\|_{0,\Omega} \lesssim \|v_h\|_{h,B}. \tag{39}$$

Applying the triangle inequality together with (39) leads to

$$\begin{aligned} \|u - u_h\|_{h,B_+} &\le \|u - w_h\|_{h,B_+} + \|u_h - w_h\|_{h,B_+} \\ &\lesssim \|u - w_h\|_{h,B_+} + \|u_h - w_h\|_{h,B} \\ &\lesssim \|u - w_h\|_{h,B_+} + \|u - u_h\|_{h,B}. \end{aligned} \tag{40}$$

where w_h is arbitrary in V_h. Substituting (40) into (38) yields (33). □

Corollary 3.2. *If the exact solution u is in $H^{p+1}(\mathcal{T}_h)$, then*

$$\|u - u_h\|_{0,\Omega} \lesssim \frac{\lambda_{M,K}}{\lambda_{m,K}} h^{p+1} \|u\|_{H^{p+1}(\mathcal{T}_h)}. \tag{41}$$

Proof. Use Theorem 3.2, Corollary 3.1 and standard approximation properties of V_h. □

4 Error analysis for the advective derivative

For vanishing diffusion it is no longer possible to control the advective derivative by means of Theorem 3.1. The goal of this section is to obtain a control of the error in the advective derivative that is robust with respect to the diffusivity. We will see that this goal can be achieved in the isotropic case. Moreover, in the anisotropic case we establish an estimate that is fully independent of the diffusivity in the advection-dominant regime.

We introduce the following norm on $V(h)$,

$$\|v\|_{h,B\beta} = \|v\|_{h,B} + \|v\|_{h,\beta} \tag{42}$$

where

$$\|v\|_{h,\beta} = \left(\sum_{T\in\mathcal{T}_h} h_T \|\beta\cdot\nabla_h v\|_{0,T}^2 \right)^{\frac{1}{2}} . \tag{43}$$

The aim of this section is to obtain a convergence result in the $\|\cdot\|_{h,\beta}$-norm. To this purpose, the first step is to derive a stability property for the SWIP bilinear form B_h in this norm.

Lemma 4.1. (Stability) *Define*

$$\forall T \in \mathcal{T}_h, \qquad \Delta_{K,T} = \begin{cases} 1 & \textit{if } \|\beta\|_{[L^\infty(T)]^d} \gtrsim \frac{\lambda_{M,T}}{h_T} \\ \frac{\lambda_{M,T}}{\lambda_{m,T}} & \textit{otherwise} \end{cases} \tag{44}$$

where $\lambda_{M,T}$ and $\lambda_{m,T}$ are respectively the maximum and the minimum eigenvalue of $K|_T$. Set $\Delta_K = \max_{T\in\mathcal{T}_h} \Delta_{K,T}$. Then,

$$\inf_{v_h\in V_h\setminus\{0\}} \sup_{w_h\in V_h\setminus\{0\}} \frac{B_h(v_h,w_h)}{\|v_h\|_{h,B\beta}\|w_h\|_{h,B\beta}} \gtrsim \Delta_K^{-1}. \tag{45}$$

Remark 4.1. *We stress the fact that the inf-sup condition is robust in the isotropic case and in the anisotropic case for dominant advection. Note also that the anisotropies are local to the mesh element, i.e., ratios of eigenvalues between adjacent elements are not considered. To achieve this result, the key point (see the control of $|[\![\pi_h]\!]|_{\gamma_K}^2$ in the proof below) is that the choice* (15) *for the weights yields $\gamma_K \le \inf(\delta_{Kn}^-, \delta_{Kn}^+)$.*

Proof. Step (i): let $v_h \in V_h$ and set $\mathbf{S} = \sup_{w_h\in V_h\setminus\{0\}} \frac{B_h(v_h,w_h)}{\|w_h\|_{h,B\beta}}$. Owing to Lemma 3.1, we infer that

$$\|v_h\|_{h,B}^2 \lesssim \mathbf{S}\|v_h\|_{h,B\beta}, \tag{46}$$

so it only remains to control the advective derivative.

Step (ii): let $\pi_h \in V_h$ be such that for all $T \in \mathcal{T}_h$ $\pi_h|_T = h_T\overline{\beta}\cdot\nabla_h v_h$ where $\overline{\beta}$ is defined in the proof of Lemma 3.3. Let us prove that

$$\|\pi_h\|_{h,B\beta} \lesssim \Delta_K^{\frac{1}{2}}\|v_h\|_{h,B\beta}. \tag{47}$$

The inverse inequality (18) and the regularity of β yield for all $T \in \mathcal{T}_h$,

$$\|\pi_h\|_{0,T} \lesssim h_T\|\beta\cdot\nabla_h v_h\|_{0,T} + h_T\|v_h\|_{0,T}, \tag{48}$$

while the inverse inequality (17) yields for all $F \in \mathcal{F}_h$

$$|[\![\pi_h]\!]|^2_{\gamma_\beta,F} \lesssim \sum_{T\in\mathcal{T}(F)} \|\pi_h\|^2_{0,\partial T} \lesssim \sum_{T\in\mathcal{T}(F)} \left(h_T\|\beta\cdot\nabla_h v_h\|^2_{0,T} + h_T\|v_h\|^2_{0,T}\right).$$

Hence,

$$\|\pi_h\|_{0,\Omega} + |[\![\pi_h]\!]|_{\gamma_\beta} \lesssim \|v_h\|_{h,B\beta}.$$

Let us estimate $h_F^{-\frac{1}{2}}|[\![\pi_h]\!]|_{\gamma_K,F}$ for all $F \in \mathcal{F}_h$. Observe first that $\gamma_K = \omega^{\mp}\delta^{\mp}_{Kn} \leq \delta^{\mp}_{Kn}$ if $F \in \mathcal{F}^i_h$ and $\gamma_K = \delta_{Kn}$ if $F \in \mathcal{F}^{\partial\Omega}_h$. Hence, if there is $T \in \mathcal{T}_h(F)$ such that $\|\beta\|_{[L^\infty(T)]^d} \gtrsim \frac{\lambda_{M,T}}{h_T}$, then

$$h_F^{-1}|[\![\pi_h]\!]|^2_{\gamma_K,F} \leq h_F^{-1}\lambda_{M,T}\|[\![\pi_h]\!]\|^2_{0,F} \leq \sum_{T\in\mathcal{T}(F)} \left(h_T\|\beta\cdot\nabla_h v_h\|^2_{0,T} + h_T\|v_h\|^2_{0,T}\right).$$

Otherwise, for all $F \in \mathcal{F}^i_h$,

$$\begin{aligned} h_F^{-1}\gamma_K[\![\pi_h]\!]^2 &\lesssim h_F\gamma_K\left(((\overline{\beta}\cdot\nabla_h v_h)^-)^2 + ((\overline{\beta}\cdot\nabla_h v_h)^+)^2\right) \\ &\lesssim h_F\left(\delta^-_{K,n}((\overline{\beta}\cdot\nabla_h v_h)^-)^2 + \delta^+_{K,n}((\overline{\beta}\cdot\nabla_h v_h)^+)^2\right), \end{aligned}$$

and similarly for $F \in \mathcal{F}^{\partial\Omega}_h$. Hence, using the trace inverse inequality (17),

$$h_F^{-1}|[\![\pi_h]\!]|^2_{\gamma_K,F} \lesssim \sum_{T\in\mathcal{T}(F)} \lambda_{M,T}\|\nabla_h v_h\|^2_{0,T} \lesssim \sum_{T\in\mathcal{T}(F)} \frac{\lambda_{M,T}}{\lambda_{m,T}}\|\kappa\nabla_h v_h\|^2_{0,T}$$

Hence, $|[\![\pi_h]\!]|_\gamma \lesssim \Delta_K^{\frac{1}{2}}\|v_h\|_{h,B\beta}$. Furthermore, since κ is piecewise constant,

$$\|\kappa\nabla_h\pi_h\|_{0,T} = h_T\|\overline{\beta}\cdot\nabla_h(\kappa\nabla_h v_h)\|_{0,T} \lesssim \|\kappa\nabla_h v_h\|_{0,T},$$

implying that $\|\kappa\nabla_h\pi_h\|_{0,\Omega} \lesssim \|v_h\|_{h,\beta}$. Finally, the advective derivative of π_h is controlled by

$$\|\pi_h\|^2_{h,\beta} \lesssim \sum_{T\in\mathcal{T}_h} h_T^{-1}\|\pi_h\|^2_{0,T} \lesssim \|v_h\|^2_{h,B\beta},$$

owing to (48). This proves (47).
Step (iii): we can now examine the term $\|v_h\|_{h,\beta}^2$ by making use of (9):

$$\begin{aligned}\|v_h\|_{h,\beta}^2 &= B_h(v_h,\pi_h) - (K\nabla_h v_h, \nabla_h \pi_h)_{0,\Omega} - (\mu v_h, \pi_h)_{0,\Omega} \\ &\quad + \sum_{T\in\mathcal{T}_h} (\beta\cdot\nabla_h v_h, h_T\beta\cdot\nabla_h v_h - \pi_h)_{0,T} + \sum_{F\in\mathcal{F}_h^i} (\beta\cdot n_F\{\pi_h\}, [\![v_h]\!])_{0,F} \\ &\quad + \sum_{F\in\mathcal{F}_h^{\partial\Omega}} \tfrac{1}{2}(\beta\cdot n_F v_h, \pi_h)_{0,F} - \sum_{F\in\mathcal{F}_h} (\gamma[\![v_h]\!], [\![\pi_h]\!])_{0,F} \\ &\quad + \sum_{F\in\mathcal{F}_h} \big((n_F^t\{K\nabla_h v_h\}_\omega, [\![\pi_h]\!])_{0,F} + (n_F^t\{K\nabla_h \pi_h\}_\omega, [\![v_h]\!])_{0,F}\big) \\ &= B_h(v_h,\pi_h) + T_1 + T_2 + T_3 + T_4 + T_5 + T_6 + T_7 + T_8.\end{aligned}$$

We observe that

$$|B_h(v_h,\pi_h)| \leq \mathbf{S}\|\pi_h\|_{h,B\beta} \leq \mathbf{S}\Delta_K^{\frac{1}{2}}\|v_h\|_{h,B\beta}.$$

It is also clear that

$$|T_1| + |T_2| + |T_6| + |T_7| + |T_8| \lesssim \|v_h\|_{h,B}\|\pi_h\|_{h,B} \lesssim \mathbf{S}^{\frac{1}{2}}\Delta_K^{\frac{1}{2}}\|v_h\|_{h,B\beta}^{\frac{3}{2}}.$$

Furthermore, using the inverse inequality (17) together with (48) yields

$$\begin{aligned}|T_4| + |T_5| \lesssim |[\![v_h]\!]|_{\gamma\beta} \left(\sum_{T\in\mathcal{T}_h} \|\pi_h\|_{0,\partial T}^2\right)^{\frac{1}{2}} &\lesssim |[\![v_h]\!]|_{\gamma\beta} \left(\sum_{T\in\mathcal{T}_h} h_T^{-1}\|\pi_h\|_{0,T}^2\right)^{\frac{1}{2}} \\ &\lesssim \|v_h\|_{h,B}\|v_h\|_{h,B\beta} \lesssim \mathbf{S}^{\frac{1}{2}}\|v_h\|_{h,B\beta}^{\frac{3}{2}}.\end{aligned}$$

Finally,

$$\begin{aligned}|T_3| &\leq \sum_{T\in\mathcal{T}_h} h_T|(\beta\cdot\nabla_h v_h, (\beta-\overline{\beta})\cdot\nabla_h v_h)_{0,T}| \lesssim \sum_{T\in\mathcal{T}_h} h_T^2\|\beta\cdot\nabla_h v_h\|_{0,T}\|\nabla_h v_h\|_{0,T} \\ &\lesssim \sum_{T\in\mathcal{T}_h} h_T\|\beta\cdot\nabla_h v_h\|_{0,T}\|v_h\|_{0,T} \lesssim \|v_h\|_{h,B\beta}\|v_h\|_{0,\Omega} \lesssim \mathbf{S}^{\frac{1}{2}}\|v_h\|_{h,B\beta}^{\frac{3}{2}}.\end{aligned}$$

Hence,

$$\begin{aligned}\|v_h\|_{h,B\beta}^2 &\lesssim \|v_h\|_{h,B}^2 + \sum_{T\in\mathcal{T}_h} h_T\|\beta\cdot\nabla_h v_h\|_{0,T}^2 \\ &\lesssim \mathbf{S}\|v_h\|_{h,B\beta} + \mathbf{S}\Delta_K^{\frac{1}{2}}\|v_h\|_{h,B\beta} + \mathbf{S}^{\frac{1}{2}}\Delta_K^{\frac{1}{2}}\|v_h\|_{h,B\beta}^{\frac{3}{2}} + \mathbf{S}^{\frac{1}{2}}\|v_h\|_{h,B\beta}^{\frac{3}{2}}.\end{aligned}$$

Applying Young's inequality yields $\|v_h\|_{h,B\beta} \lesssim \mathbf{S}\Delta_K$, and thus (45). □

Proceeding as above, the following result is readily inferred:

Theorem 4.1. *In the above framework,*

$$\|u - u_h\|_{h,B\beta} \lesssim \Delta_K \inf_{v_h \in V_h} \|u - v_h\|_{h,\frac{1}{2}\beta}, \tag{49}$$

where, for all $v \in V(h)$,

$$\|v\|_{h,\frac{1}{2}\beta} = \|v\|_{h,\beta} + \left(\sum_{T \in \mathcal{T}_h} \|v\|^2_{0,\partial T} \right)^{\frac{1}{2}} + \left(\sum_{T \in \mathcal{T}_h} h_T \|\kappa \nabla_h v\|^2_{0,\partial T} \right)^{\frac{1}{2}}. \tag{50}$$

In particular, if the exact solution u is in $H^{p+1}(\mathcal{T}_h)$ and $\|\beta\|_{[L^\infty(T)]^d} \gtrsim \frac{\lambda_{M,T}}{h_T}$ $\forall T \in \mathcal{T}_h$, then

$$\|u - u_h\|^2_{h,\beta} \lesssim h^{p+\frac{1}{2}} \|u\|_{H^{p+1}(\mathcal{T}_h)}. \tag{51}$$

5 Numerical tests

5.1 A test case with discontinuous coefficients

To verify the convergence of the SWIP method and to make quantitative comparisons between this and other IP methods, we consider the test problem proposed in [3], featuring discontinuous coefficients and where the exact solution is known analytically. We split the domain $\Omega = [0,1] \times [0,1]$ into two subregions: $\Omega_1 = [0, \frac{1}{2}] \times [0,1]$, $\Omega_2 = [\frac{1}{2}, 1] \times [0,1]$. The diffusivity tensor K is constant within each subregion, and defined as

$$K(x,y) = \begin{pmatrix} \epsilon(x) & 0 \\ 0 & 1.0 \end{pmatrix}$$

where $\epsilon(x)$ is a discontinuous function across the interface $x = \frac{1}{2}$. Indicating with the subscript 1 (resp. 2) the restriction to the subdomain Ω_1 (resp. Ω_2), we will consider different values of ϵ_1, while ϵ_2 is set equal to 1. Setting $\beta = (1,0)^t$, $\mu = 0$ and $f = 0$, the exact solution is independent of the y-coordinate, and is exponential with respect to the x-coordinate. The following conditions must be satisfied at the interface between the two subdomains:

$$\lim_{x \to \frac{1}{2}^-} u(x,y) = \lim_{x \to \frac{1}{2}^+} u(x,y), \text{ and } \lim_{x \to \frac{1}{2}^-} -\epsilon_1 \partial_x u(x,y) = \lim_{x \to \frac{1}{2}^+} -\partial_x u(x,y).$$

Setting $u(0,y) = 1$, $u(1,y) = 0$ and applying the matching conditions, we obtain the value of the exact solution at the interface:

$$u\left(\tfrac{1}{2}, y\right) = \frac{\exp(\frac{1}{2\epsilon_1})}{1 - \exp(\frac{1}{2\epsilon_1})} \left(\frac{\exp(\frac{1}{2\epsilon_1})}{1 - \exp(\frac{1}{2\epsilon_1})} + \frac{1}{1 - \exp(\frac{1}{2})} \right)^{-1}.$$

As a result, the exact solution in each subdomain can be expressed as

$$u_1(x,y) = \frac{u(\frac{1}{2},y) - \exp(\frac{1}{2\epsilon_1}) + (1 - u(\frac{1}{2},y))\exp(\frac{x}{\epsilon_1})}{1 - \exp(\frac{1}{2\epsilon_1})},$$

$$u_2(x,y) = \frac{-\exp(\frac{1}{2})u(\frac{1}{2},y) + u(\frac{1}{2},y)\exp(x - \frac{1}{2})}{1 - \exp(\frac{1}{2})}.$$

5.2 Accuracy of the SWIP method

Just like any discontinuous Galerkin method based on interior penalties, the precise setting of the SWIP method depends on the definition of the penalty parameter α which has to be large enough in order to obtain a well posed discrete formulation. In the following numerical tests we have used $\alpha = 1.0$ for $\mathbb{P}_1$ elements and $\alpha = 4.0$ for $\mathbb{P}_2$.

We apply the SWIP method to the test case presented in the previous section. In order to assess the accuracy of the SWIP method with respect to the mesh-size h, we consider a family of uniform triangulations $\{\mathcal{T}_h\}_{h>0}$ which are conforming with respect to the interface between Ω_1 and Ω_2. These triangulations are obtained starting from a uniform partition of $\partial\Omega$ in sub-intervals of length $h = 0.1$, $h = 0.05$, $h = 0.025$ and $h = 0.0125$ respectively. The numerical results obtaied with $\epsilon_1 = 0.1$ are found in Tables 1 and 2. Since the exact solution u is sufficiently smooth locally, and the computational mesh is conforming with respect to the interface where $\epsilon(x)$ is discontinuous, we expect that our method satisfies the order of convergence in the norms $\|\cdot\|_{h,B}$, $\|\cdot\|_{h,\beta}$ and $\|\cdot\|_{0,\Omega}$ provided by the theory (see (30), (51) and (41)). These properties are clearly verified by the numerical experiments, where the order of convergence is computed with respect to the last two rows of each table.

5.3 SWIP versus IP

We compare the performance of the SWIP method with respect to two IP methods differing in their choice of the penalty parameter. The first method, indicated by IP-A, corresponds to the SWIP method with weights $\omega^{\mp} = \frac{1}{2}$. The penalty parameter γ_K is thus the arithmetic average of the diffusion in the direction normal to the face. This method was analyzed in [11]. The second method (IP-B), proposed in [15], differs

Table 1: Convergence rates of the SWIP method ($p = 1$)

h	$\|u - u_h\|_{h,B}$	$\|u - u_h\|_{h,\beta}$	$\|u - u_h\|_{0,\Omega}$
0.1000	1.62e-01	1.49e-01	6.94e-03
0.0500	7.96e-02	5.45e-02	2.11e-03
0.0250	3.67e-02	1.87e-02	4.80e-04
0.0125	1.70e-02	6.37e-03	1.21e-04
order	1.11	1.55	1.98

from IP-A in the choice of penalty parameter: here γ_K is the arithmetic average of the maximum eigenvalue of K on the triangles sharing the face F.

We consider the test case proposed in Section 5.1 on a uniform triangulation $\mathcal{T}_h$ characterized by $h = 0.05$. The quantitative analysis is based on the norms $\|\cdot\|_{h,B}$, $\|\cdot\|_{h,\beta}$, $\|\cdot\|_{0,\Omega}$ and the indicator

$$M = \max(|\max_{\Omega}(u_h) - \max_{\Omega}(u)|, |\min_{\Omega}(u_h) - \min_{\Omega}(u)|) \tag{52}$$

which quantifies overshoots and undershoots of the calculated solution. The numerical results are found in Tables 3 and 4 and in Figure 1. These results show that the SWIP scheme performs better than the considered IP methods, particularly when the computational mesh is not completely adequate to capture the singularities of the exact solution. This is evident in the case of $\epsilon_1 = 5$e-3 where the weights permit sharper discontinuities in the calculated solution, leading to smaller oscillations in the internal layer. Indeed, the indicator M shows that the maximal and the minimal values of the exact solution are closely respected. This is not the case for the other IP methods, where the solution is forced to be almost continuous. As can be observed in Figure 1, this limitation promotes the instability of the approximate solution in the neighborhood of the internal layer. The spurious oscillations generated in this case lead to an overshoot of about 40%.

The robustness of the SWIP method with respect to standard IP schemes is also confirmed by further numerical tests concerning vanishing values of ϵ_1. In Figure 2 we see that as the diffusivity decreases the difference between the IP methods and the SWIP method augments. Comparing the error measures, the SWIP method performs favourably with respect to the IP methods as the internal layer becomes sharper. These observations are confirmed by figure 3, where we compare SWIP and IP methods in the case $\epsilon_1 = 1$e-6. The solution computed by the SWIP method is very close to the exact solution, whereas the IP-A and IP-B methods become unstable.

5.4 A test case with genuine anisotropic properties

To conclude the sequence of numerical tests, we consider a test case with genuine anisotropic properties. Because of the complexity of the problem, it is not possible

Table 2: Convergence rates of the SWIP method ($p = 2$)

h	$\|u - u_h\|_{h,B}$	$\|u - u_h\|_{h,\beta}$	$\|u - u_h\|_{0,\Omega}$
0.1000	2.31e-02	2.15e-02	6.80e-04
0.0500	4.63e-03	3.31e-03	4.29e-05
0.0250	1.17e-03	5.93e-04	5.20e-06
0.0125	2.95e-04	1.05e-04	6.41e-07
order	1.99	2.49	3.02

Figure 1: Graphical comparison between the methods SWIP and IP-A. The test case with $\epsilon_1 = 5$e-2 is reported on the left while the case with $\epsilon_1 = 5$e-3 is on the right. In both cases $\epsilon_2 = 1$. Each column shows the one-dimensional exact solution $u(x)$ of the test problem (top) and the numerical approximation u_h obtained with the methods SWIP (center) and IP-A (bottom), by means of piecewise-linear elements ($p = 1$). The case IP-B has been omitted since it is qualitatively equivalent to IP-A.

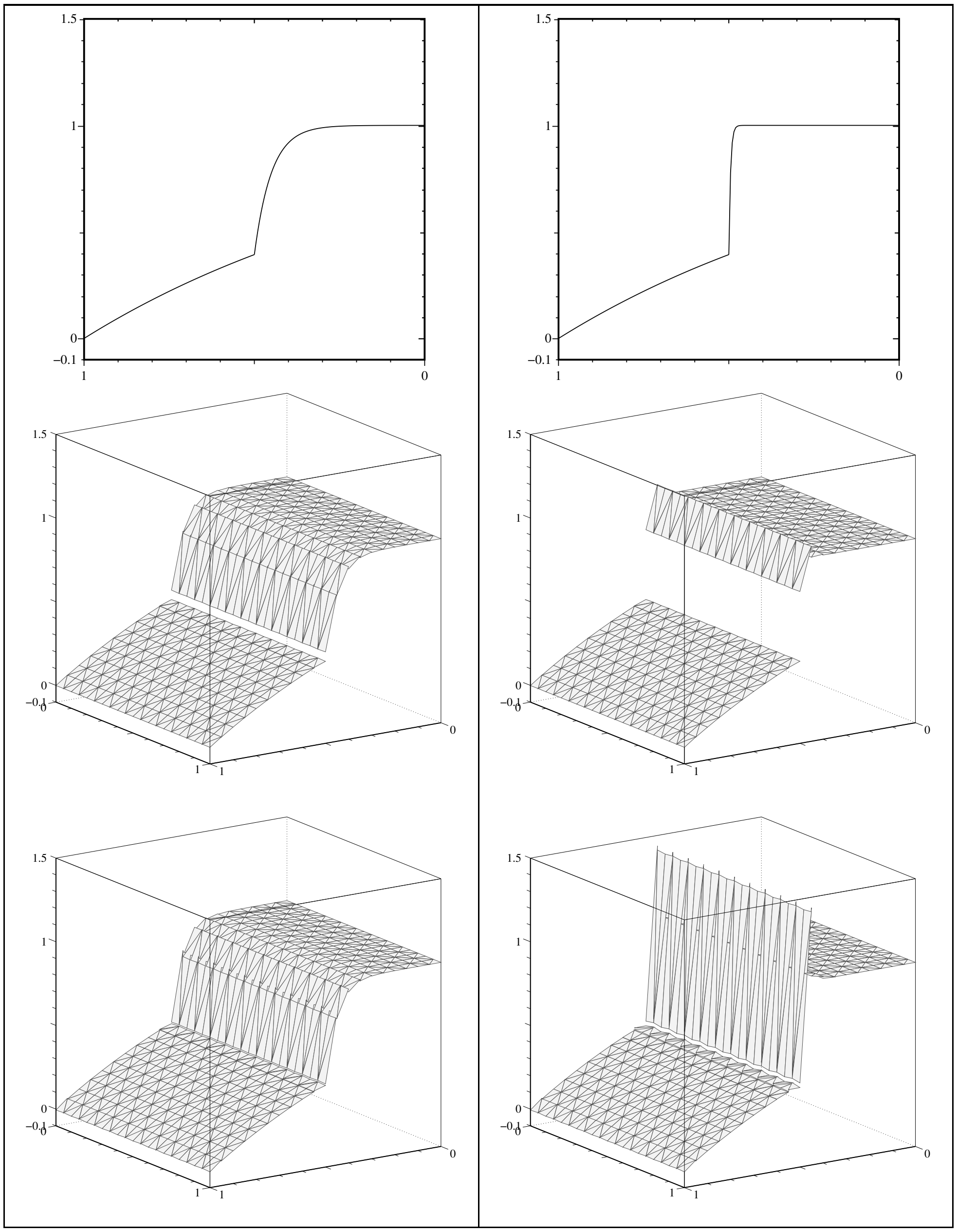


Figure 2: The norm $\|\cdot\|_{0,\Omega}$ and the indicator (52) (denoted with *M*) are plotted for the values $\epsilon_1 = 2^{-i},\ i = 0, \ldots, 16$. The methods SWIP, IP-A and IP-B are compared with respect to these indicators.

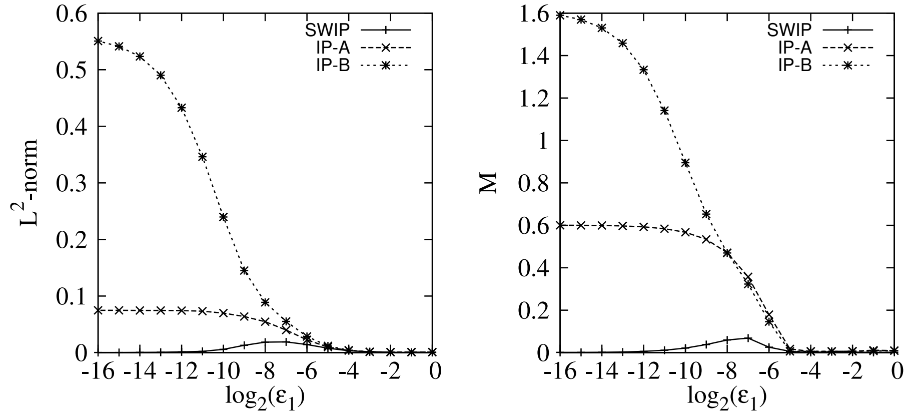
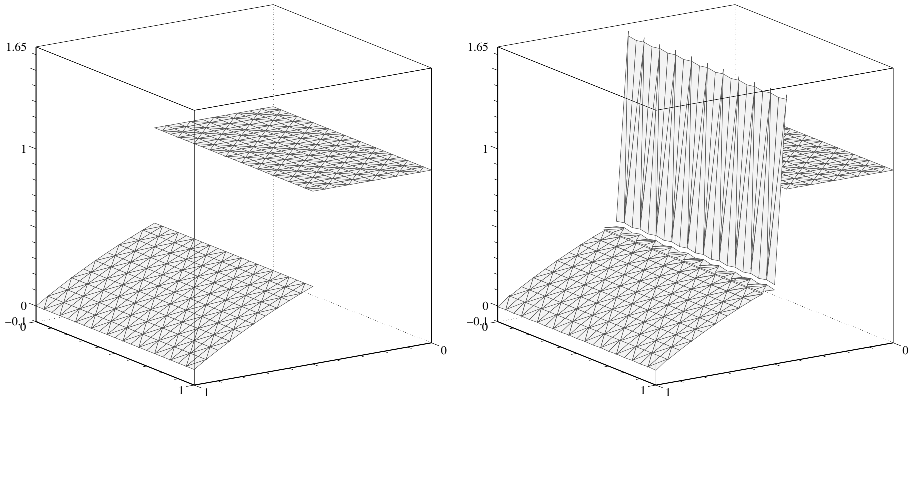


Figure 3: Graphical comparison between the methods SWIP (left) and IP-A (right) in the case $\epsilon_1 = $ 1e-6.



to compute analytically the exact solution. Consequently, the comparison between the SWIP and the IP methods will only be qualitative.

We consider the unit square $\Omega = [0,1] \times [0,1]$ split into four subdomains: $\Omega_1 = [0, \frac{1}{2}] \times [0, \frac{1}{2}]$, $\Omega_2 = [\frac{1}{2}, 1] \times [0, \frac{1}{2}]$, $\Omega_3 = [\frac{1}{2}, 1] \times [\frac{1}{2}, 1]$ and $\Omega_4 = [0, \frac{1}{2}] \times [\frac{1}{2}, 1]$. The diffusion tensor K takes different values in each subregion:

$$K(x,y) = \begin{pmatrix} 1\mathrm{e}{-6} & 0 \\ 0 & 1.0 \end{pmatrix} \text{ for } (x,y) \in \Omega_1,\ \Omega_3,$$
$$K(x,y) = \begin{pmatrix} 1.0 & 0 \\ 0 & 1\mathrm{e}{-6} \end{pmatrix} \text{ for } (x,y) \in \Omega_2,\ \Omega_4.$$

For the advection term we consider a solenoidal field $\beta = (\beta_x, \beta_y)^t$ with $\beta_x = -10(2y-1)(1-(2x-1)^2)$ and $\beta_y = -40y(2x-1)(y-1)$. We note that the field is oriented along the normal of the interfaces $x = \frac{1}{2}$ and $y = \frac{1}{2}$ where $K(x,y)$ is discontinuous, in the direction of increasing diffusion. Examining the variations along a radius originated in the center of Ω, the forcing term $f(x,y) = 10^{-2} \exp(-(\sqrt{(x-0.5)^2 + (y-0.5)^2} - 0.35)^2/0.005)$ is a Gaussian hill with center at $r = 0.35$. Finally, we choose $\mu = 1$. For the simulations, we consider a uniform mesh characterized by $h = 0.025$. This mesh is conforming with respect to the discontinuity of K.

A qualitative representation of the data is found in Figure 4. Similarly to what happens for the test case described in Section 5.1, the orientation of the advection field at the interfaces where K is discontinuous in combination with a change from a dominant advective to a dominant diffusive regime, induces quite steep internal layers.

In Figure 5 we compare the solutions obtained with the SWIP and the IP methods. The contour plots of the numerical solutions confirm that the methods at hand behave differently in the neighborhood of the interfaces where the tensor K is discontinuous. We observe that the SWIP scheme approximates the internal layers by means

Table 3: The accuracy of the SWIP and the IP methods: $\epsilon_1 = $ 5e-2

method	$\|\|u - u_h\|\|_{h,B}$	$\|\|u - u_h\|\|_{h,\beta}$	$\|\|u - u_h\|\|_{0,\Omega}$	M
SWIP	1.583e-01	1.505e-01	4.586e-03	9.555e-04
IP-A	1.483e-01	1.403e-01	5.153e-03	5.882e-03
IP-B	1.338e-01	1.378e-01	5.903e-03	5.882e-03

Table 4: The accuracy of the SWIP and the IP methods: $\epsilon_1 = $ 5e-3

method	$\|\|u - u_h\|\|_{h,B}$	$\|\|u - u_h\|\|_{h,\beta}$	$\|\|u - u_h\|\|_{0,\Omega}$	M
SWIP	4.917e-01	1.280	1.474e-02	6.594e-02
IP-A	5.886e-01	1.303	4.973e-02	4.373e-01
IP-B	6.625e-01	1.634	7.553e-02	4.173e-01

Figure 4: The test case with genuine anisotropic properties. At the top, an illustration of the domain and its subregions together with a synoptic description of the diffusivity tensor. The advection field β and the forcing term f are shown bottom left and right respectively.

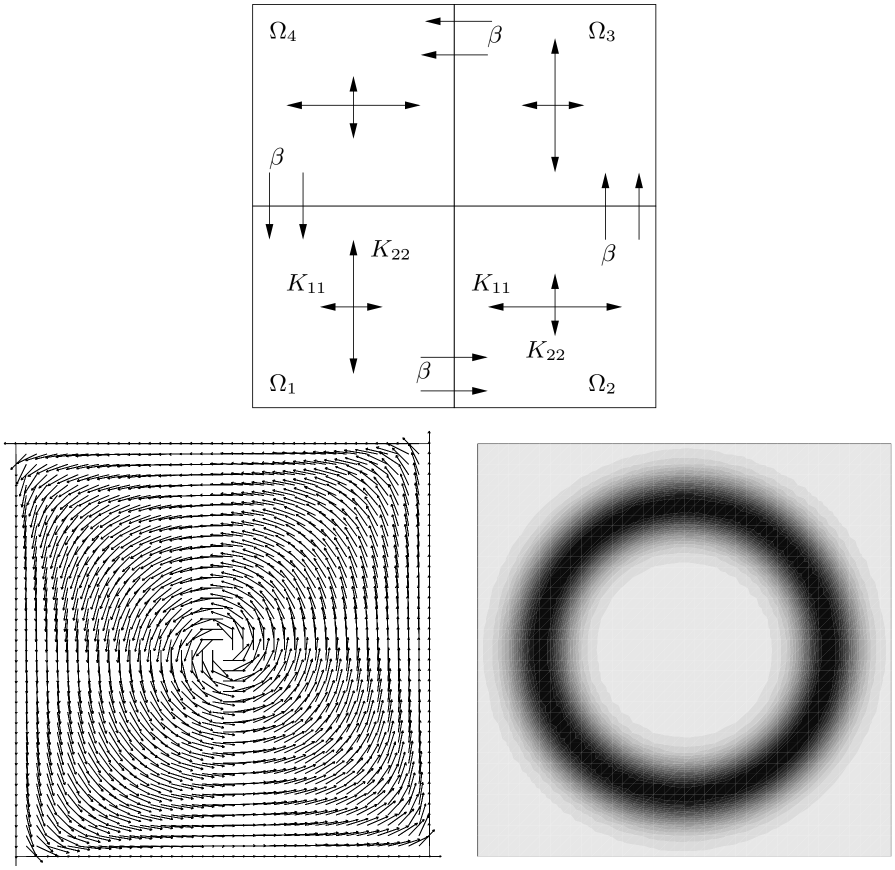
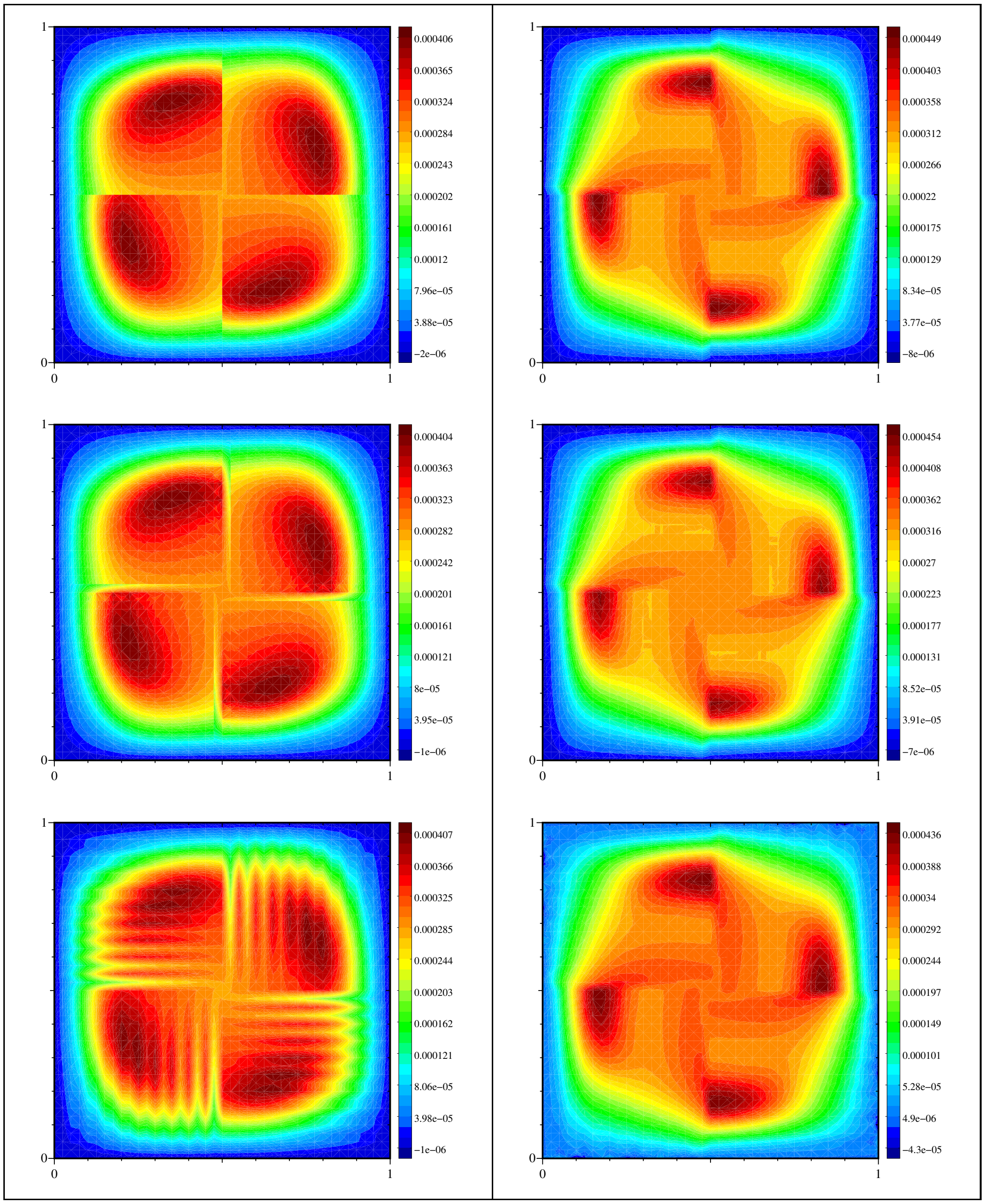


Figure 5: The test cases of section 5.4. The rotating field is counterclockwise on the left (see figure 4) and clockwise on the right. The solution obtained by the SWIP scheme is reported on the top while the ones relative to the interior penalty methods IP-A and IP-B are depicted below.



of jumps, while the IP schemes attempt to recover a numerical solution which is almost continuous. Since the computational mesh is insufficiently refined, the scheme IP-A generates some slight undershoots near the interfaces where K is discontinuous. For the IP-B method the oscillations generated by the approximation of the internal layer are much more evident and propagate quite far from the interfaces. This behavior can be explained by observing that this type of penalty does not distinguish between the principal directions of diffusion. Consequently, an excessive penalty is applied along the direction of low diffusivity.

To strengthen these conclusions, we also consider a numerical test where the advection field is the opposite of the one reported in Figure 4, i.e. it rotates clockwise. Following this advection field along the interfaces between subregions, the diffusivity decreases. These conditions lead to an exact solution which is continuous in the neighborhood of the interfaces. In this case, we expect that the SWIP, the IP-A and IP-B methods behave similarly. Indeed, this is confirmed by the numerical results reported in Figure 5, on the right hand side. Although the SWIP method enforces the continuity between elements in a weaker way with respect to IP-A and IP-B, it provides a solution that is comparable with the others.

6 Concluding remarks

The SWIP method analyzed in this paper is a DG method with weighted averages designed to approximate satisfactorily advection-diffusion equations with anisotropic and possibly locally vanishing diffusivity. A thorough a priori analysis has been carried out, yielding robust and optimal error estimates that have been supported by numerical evidence. The SWIP method is an interesting alternative to other IP methods in the presence of internal layers caused by locally vanishing diffusivity, since these are approximated more sharply without increasing the computational complexity.

References

[1] D.N. Arnold. An interior penalty finite element method with discontinuous elements. *SIAM J. Numer. Anal.*, 19:742–760, 1982.

[2] G.A. Baker. Finite element methods for elliptic equations using nonconforming elements. *Math. Comp.*, 31(137):45–59, 1977.

[3] E. Burman and P. Zunino. A domain decomposition method based on weighted interior penalties for advection-diffusion-reaction problems. *SIAM Journal on Numerical Analysis*, 44(4):1612–1638, 2006.

[4] B. Cockburn and C.W. Shu. The local discontinuous Galerkin method for time-dependent convection-diffusion systems. *SIAM J. Numer. Anal.*, 35:2440–2463, 1998.

[5] J.-P. Croisille, A. Ern, T. Lelièvre, and J. Proft. Analysis and simulation of a coupled hyperbolic/parabolic model problem. *J. Numer. Math.*, 13(2):81–103, 2005.

[6] D.A. Di Pietro, A. Ern, and J.-L. Guermond. Discontinuous Galerkin methods for anisotropic diffusion with advection. *SIAM J. Numer. Anal.*, 2006. submitted.

[7] A. Ern and J.-L. Guermond. *Theory and Practice of Finite Elements*, volume 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, NY, 2004.

[8] A. Ern and J.-L. Guermond. Discontinuous Galerkin methods for Friedrichs' systems. I. General theory. *SIAM J. Numer. Anal.*, 44(2):753–778, 2006.

[9] A. Ern and J. Proft. Multi-algorithmic methods for coupled hyperbolic-parabolic problems. *Int. J. Numer. Anal. Model.*, 1(3):94–114, 2006.

[10] F. Gastaldi and A. Quarteroni. On the coupling of hyperbolic and parabolic systems: analytical and numerical approach. *Appl. Numer. Math.*, 6(1-2):3–31, 1989/90. Spectral multi-domain methods (Paris, 1988).

[11] E. H. Georgoulis and A. Lasis. A note on the design of hp-version interior penalty Galerkin finite element methods *IMA J. Numer. Anal.*, 26(2):381–390, 2006.

[12] B. Heinrich and S. Nicaise. The Nitsche mortar finite-element method for transmission problems with singularities. *IMA J. Numer. Anal.*, 23(2):331–358, 2003.

[13] B. Heinrich and K. Pietsch. Nitsche type mortaring for some elliptic problem with corner singularities. *Computing*, 68(3):217–238, 2002.

[14] B. Heinrich and K. Pönitz. Nitsche type mortaring for singularly perturbed reaction-diffusion problems. *Computing*, 75(4):257–279, 2005.

[15] P. Houston, Ch. Schwab, and E. Süli. Discontinuous hp-finite element methods for advection-diffusion-reaction problems. *SIAM J. Numer. Anal.*, 39(6):2133–2163, 2002.

[16] P. Lesaint and P.-A. Raviart. On a finite element method for solving the neutron transport equation. In C. de Boors, editor, *Mathematical aspects of Finite Elements in Partial Differential Equations*, pages 89–123. Academic Press, 1974.

[17] J. Nitsche. Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind. *Abh. Math. Sem. Univ. Hamburg*, 36:9–15, 1971. Collection of articles dedicated to Lothar Collatz on his sixtieth birthday.

[18] J. Nitsche. On Dirichlet problems using subspaces with nearly zero boundary conditions. In *The mathematical foundations of the finite element method with applications to partial differential equations (Proc. Sympos., Univ. Maryland, Baltimore, Md., 1972)*, pages 603–627. Academic Press, New York, 1972.

[19] W.H. Reed and T.R. Hill. Triangular mesh methods for the neutron transport equation. Technical Report LA-UR-73-479, Los Alamos Scientific Laboratory, Los Alamos, NM, 1973.

[20] B. Rivière, M. Wheeler, and V. Girault. Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems. Part I. *Comput. Geosci.*, 3:337–360, 1999.

[21] R. Stenberg. Mortaring by a method of J.A. Nitsche. In Idelsohn S.R., Oñate E., and Dvorkin E.N., editors, *Computational Mechanics: New trends and applications*, Barcelona, Spain, 1998.