Improved Error estimates of hybridizable interior penalty methods using a variable penalty for highly anisotropic diffusion problems

Grégory Etangsale, Marwan Fahs, Vincent Fontaine, Nalitiana Rajaonison

Error estimates of hybridizable interior penalty methods using a variable penalty for highly anisotropic diffusion problems

Grégory Etangsale[a], Marwan Fahs[b], Vincent Fontaine[a,*], Nalitiana Rajaonison[a]

[a]*Department of Building and Environmental Sciences, University of La Réunion, France*
[b]*Université de Strasbourg, CNRS, ENGEES, LHYGES UMR 7517, F-67000 Strasbourg, France*

Abstract

In this paper, we derive improved *a priori* error estimates for families of hybridizable interior penalty discontinuous Galerkin (H-IP) methods using a variable penalty for second-order elliptic problems. The strategy is to use a penalization function of the form $O(1/h^{1+\delta})$, where h denotes the mesh size and δ is a user-dependent parameter. We then quantify its direct impact on the convergence analysis, namely, the (strong) consistency, discrete coercivity and boundedness (with h^{δ}-dependency), and we derive updated error estimates for both discrete energy- and L^2-norms. All theoretical results are supported by numerical evidence.

Keywords: Hybridizable discontinuous Galerkin, interior penalty methods, variable-penalty technique, convergence analysis, updated *a priori* error estimates
2020 MSC: 65N12, 65N15, 65N30, 65N38

1. Introduction

Hybridizable discontinuous Galerkin (HDG) methods were first introduced in the last decade by Cockburn *et al.* [1] (see, e.g., [2]) and have since received extensive attention from the research community. They are popular and very efficient numerical approaches for solving a large class of partial differential equations (see, e.g., [3, 4, 5, 6, 7] for a historical perspective). Indeed, they inherit attractive features from both (i) discontinuous Galerkin (DG) methods such as local conservation, *hp*-adaptivity and high-order polynomial approximation [8] and (ii) standard conforming Galerkin (CG) methods such as the Schur complement strategy [9]. One undeniable additional benefit of the HDG methods is their superconvergence property, obtained through the application of a local postprocessing technique on each element of the mesh [4]. In the hybrid formalism, additional unknowns are introduced along the mesh skeleton corresponding to discrete trace approximations. Thanks to the specific localization of its additional degrees of freedom (dofs), interior variables can be eliminated in favor of its Lagrange multipliers by only static condensation [10]. The resulting matrix system is significantly smaller and sparser than those associated with CG or DG methods for any given mesh and polynomial degree [9]. Several HDG formulations have been derived in the literature and can be classified into two main categories. The first is based on a primal form of the continuous problem, such as the class of interior penalty (IP) methods [11], whereas the second relies on a dual (often called mixed) form, such as local discontinuous Galerkin (LDG) methods [1, 4, 12]. In the latter formulation, the flux variable is introduced as an additional unknown of the problem.

Our focus is on families of hybridizable interior penalty discontinuous Galerkin (H-IP) methods [13]. They are hybridized counterparts of the well-known interior penalty DG (IPDG) methods [14, 15, 16] and have been analyzed until quite recently by several authors [11, 6]. Specifically, in our exposition, we considered the incomplete, non-symmetric and symmetric schemes denoted by H-IIP, H-NIP and H-SIP, respectively. The main difference between these schemes concerns the role of the *symmetrization* term in the discrete bilinear form [15]. Fabien *et al.* recently

*Corresponding author
Email address: vincent.fontaine@univ-reunion (Vincent Fontaine)

analyzed these schemes using a stabilization function of the form $O(1/h)$ for solving second-order elliptic problems [11]. The authors conclude that H-IP methods inherit similar convergence properties to their IPDG equivalents. Notably, they theoretically establish (i) optimal energy error estimates, and because of the lack of symmetry of the associated discrete operator, (ii) only suboptimal L^2-norm error estimates for H-IIP and H-NIP schemes. In addition, they numerically conclude that the L^2orders of convergence of both non-symmetric variants are suboptimal for only even polynomial degrees and are optimal otherwise. Similar conclusions have also been suggested by Oikawa for second-order elliptic problems [5].

To restore optimal L^2-error estimates for the nonsymmetric IPDG method, Rivière *et al.* suggest using a sort of superpenalty on the jumps [17, 18]. In the present paper, we explore a similar idea in the general context of H-IP methods by using a variable penalty function of the form $\tau := O(1/h^{1+\delta})$, where $\delta \in \mathbb{R}$. Here, we analyze the direct impact of the parameter δ on *a priori* error estimates in different norms. First, we propose a convergence analysis by investigating three key properties: (strong) consistency, discrete coercivity and boundedness. One remarkable feature of this strategy is the h^{δ}-dependency of the coercivity condition and the continuity (or boundedness) constant C_{bnd}, which consequently impacts the error estimates. Improved error estimates are then derived in the spirit of the second Strang lemma [16], and we first prove that the order of convergence in the natural energy-norm is linear, δ-dependent, and optimal when $\delta \geq 0$ for any scheme. Then, by using a duality argument, i.e., the so-called Aubin–Nitsche technique, we also prove that the optimal convergence is theoretically reached as soon as $\delta \geq 0$ for the H-SIP scheme only, and when $\delta \geq 2$ for both non-symmetric variants, i.e., H-NIP and H-IIP schemes. We recover some well-known theoretical error estimates proposed in the literature for both the natural energy- and L^2-norms in the particular case of $\delta = 0$.

The rest of the material is organized as follows: Section 2 describes the model problem, mesh notation and assumptions, and recalls some definitions and useful (trace) inequalities, while Section 3 derives the discrete H-IP formulation and discusses its stability properties. In Section 4, optimal error estimates are provided for both the energy- and L^2-norms by using a standard duality argument. Section 5 concerns the numerical experiments that validate our theoretical results. We briefly end with some remarks and perspectives.

2. Some preliminaries

2.1. The model problem

Let Ω be a bounded (polyhedron) domain in $\mathbb{R}^d$ with Lipschitz boundary $\partial\Omega$ in spatial dimension $d \geq 2$. For clarity, we consider the anisotropic diffusion problem with homogeneous Dirichlet boundary conditions:

$$-\nabla \cdot (\boldsymbol{\kappa} \nabla u) = f \quad \text{in } \Omega \quad \text{and} \quad u = 0 \quad \text{on } \partial\Omega, \tag{1}$$

where $\boldsymbol{\kappa} \in [L^{\infty}(\Omega)]^{d\times d}$ is a bounded, symmetric, uniformly positive-definite matrix-valued function and $f \in L^2(\Omega)$ is a forcing term. Thus, the weak formulation of problem (1) is to find $u \in H_0^1(\Omega)$ such that

$$\int_{\Omega} \boldsymbol{\kappa} \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v \mathrm{d}x = \int_{\Omega} f v \mathrm{d}x \quad \forall v \in H_0^1(\Omega). \tag{2}$$

It is well known that under elliptic regularity assumptions, the variational problem (2) is well posed.

2.2. Mesh notation and assumptions

Let h be a positive parameter; we assume without loss of generality that $h \leq 1$. We denote by $\{\mathcal{T}_h\}_{h>0}$ a family of affine triangulations of the domain Ω, where h stands for the largest diameter: $h_E := \text{diam}(E)$. We also assume that $\mathcal{T}_h$ is *quasi-uniform*, meaning that for all $E \in \mathcal{T}_h$, there exists $0 < \rho_0 \leq 1$ independent of h such that $\rho_0 h \leq h_E \leq h$. Following our notation, the generic term *interface* indicates a $(d-1)$-dimensional geometric object, i.e., an edge, if $d = 2$ and a face if $d = 3$. Thus, we denote by $\mathcal{F}_h^{\text{i}}$ the set of interior interfaces; i.e., $F \in \mathcal{F}_h^{\text{i}}$ if there exist E_1 and E_2 in $\mathcal{T}_h$ such that $F := \partial E_1 \cap \partial E_2$. The set of boundary interfaces is denoted by $\mathcal{F}_h^{\text{b}}$; i.e., $F \in \mathcal{F}_h^{\text{b}}$ if there exists E in $\mathcal{T}_h$ such that $F := \partial E \cap \partial\Omega$. The set of all interfaces is often called the mesh skeleton and is denoted by $\mathcal{F}_h$, i.e., $\mathcal{F}_h := \mathcal{F}_h^{\text{i}} \cup \mathcal{F}_h^{\text{b}}$. We denote by $\partial\mathcal{T}_h := \{\cup \partial E, \forall E \in \mathcal{T}_h\}$, the collection of interfaces of all mesh elements. Let X be a mesh element or an interface; we then denote by $|X|$ a positive d- or $(d-1)$-dimensional Lebesgue measure of X,

respectively. Moreover, for any mesh element $E \in \mathcal{T}_h$, we denote by $\mathcal{F}_E := \{F \in \mathcal{F}_h \,:\, F \subset \partial E\}$ the set of interfaces composing the boundary of E; we define $\eta_E := \mathrm{card}(\mathcal{F}_E)$ and $\eta_0 := \max_{\forall E \in \mathcal{T}_h} (\eta_E)$.

2.3. Broken polynomial spaces

For any polyhedral domain $D \subset \mathbb{R}^d$ with $\partial D \subset \mathbb{R}^{d-1}$, we denote by $(\cdot,\cdot)_{0,D}$ (resp., $\langle\cdot,\cdot\rangle_{0,\partial D}$) the L^2-inner product in $L^2(D)$ (resp., $L^2(\partial D)$) equipped with its natural norm $\|\cdot\|_{0,D}$ (resp., $\|\cdot\|_{0,\partial D}$). Let us now introduce some compact notation associated with the discrete L^2-inner scalar product:

$$(\cdot,\cdot)_{0,\mathcal{T}_h} := \sum_{E\in\mathcal{T}_h} (\cdot,\cdot)_{0,E} \quad \text{and} \quad \langle\cdot,\cdot\rangle_{0,\partial\mathcal{T}_h} := \sum_{E\in\mathcal{T}_h} \langle\cdot,\cdot\rangle_{0,\partial E}. \tag{3}$$

We denote by $\|\cdot\|_{0,\mathcal{T}_h}$ and $\|\cdot\|_{0,\partial\mathcal{T}_h}$ the corresponding norms. Similarly, we denote by $H^s(D)$ the usual Hilbert space of index s on D equipped with its natural norm $\|\cdot\|_{s,D}$ and seminorm $|\cdot|_{s,D}$, respectively. If $s = 0$, then we set $H^0(D) = L^2(D)$. We denote by $H^s(\mathcal{T}_h)$ the usual broken Sobolev space and by $\boldsymbol{\nabla}_h$ the broken gradient operator acting on $H^s(\mathcal{T}_h)$ with $s \geq 1$. We assume an extended regularity requirement of the exact solution u of the weak problem (2), i.e., $u \in H^s_0(\Omega) \cap H^2(\mathcal{T}_h)$ with $s > 3/2$. We also introduce the additional unknown $\hat{u} \in L^2(\mathcal{F}_h)$ corresponding to the trace of u on the skeleton of the mesh. Let us now introduce the composite variable $\boldsymbol{u} := (u, \hat{u})$, which belongs to the continuous approximation space $\boldsymbol{V} := H^s_0(\Omega) \cap H^2(\mathcal{T}_h) \times L^2(\mathcal{F}_h)$; i.e., $\boldsymbol{u} \in \boldsymbol{V}$. As usual in HDG methods, we consider broken Sobolev spaces:

$$\mathbb{P}_k(\mathcal{T}_h) := \{v_h \in L^2(\mathcal{T}_h) \,:\, v_h|_E \in \mathbb{P}_k(E),\ \forall E \in \mathcal{T}_h\}, \tag{4}$$

and similarly for $\mathbb{P}_k(\mathcal{F}_h)$. Here, $\mathbb{P}_k(X)$ denotes the space of polynomials of at least degree k on X, where X corresponds to a generic element of $\mathcal{T}_h$ or $\mathcal{F}_h$, respectively. For H-IP discretization, two types of discrete variables are necessary to approximate the weak solution u of problem (2). First, the discrete variable $u_h \in V_h$ is defined within each mesh element, and its trace $\hat{u}_h \in \hat{V}_h$ is defined on the mesh skeleton with respect to the imposed homogeneous Dirichlet boundary conditions. Thus, we set $V_h := \mathbb{P}_k(\mathcal{T}_h)$ and $\hat{V}_h := \mathbb{P}^0_k(\mathcal{F}_h)$, where

$$\mathbb{P}^0_k(\mathcal{F}_h) := \{\hat{v}_h \in \mathbb{P}_k(\mathcal{F}_h) \,:\, \hat{v}_h|_F = 0,\ \forall F \in \mathcal{F}_h^{\mathrm{b}}\}. \tag{5}$$

Throughout the manuscript, we use the following compact notation: Let $\boldsymbol{V}_h := V_h \times \hat{V}_h$ denote the composite approximation space and a generic element of $\boldsymbol{V}_h$ be denoted by $\boldsymbol{v}_h := (v_h, \hat{v}_h)$. For all $E \in \mathcal{T}_h$ and $F \in \mathcal{F}_E$, we define the jump of $\boldsymbol{v}_h \in \boldsymbol{V}_h$ across F as $[\![\boldsymbol{v}_h]\!]_{E,F} := (v_h|_E - \hat{v}_h|_F)\mathbf{n}_F$, where $\mathbf{n}_F$ denotes the unit normal vector to F pointing out of E. When confusion cannot arise, we omit the subscripts E and F from the definition, and we simply write $[\![\boldsymbol{v}_h]\!] := (v_h - \hat{v}_h)\mathbf{n}$. Finally, we introduce the space $\boldsymbol{V}(h) := \boldsymbol{V} + \boldsymbol{V}_h$ to analyze the boundedness of the discrete bilinear form.

2.4. Useful inequalities

We recall here some useful inequalities that will be used extensively later on (see, e.g., [19, 16, 15]). For clarity, C denotes a generic constant that is independent of h, h_E and $\boldsymbol{\kappa}$ in the rest of the manuscript. Owing to the shape regularity of $\mathcal{T}_h$, we now introduce multiplicative trace inequalities. Let $E \in \mathcal{T}_h$ and $F \in \mathcal{F}_E$. For all $v \in H^2(E)$, there exists a positive constant C_{M} independent of h_E, v and E such that

$$\|v\|^2_{0,F} \leq C_{\mathrm{M}}(\|v\|_{0,E}|v|_{1,E} + h_E^{-1}\|v\|^2_{0,E}), \tag{6a}$$

$$\|\boldsymbol{\nabla}_h v\|^2_{0,F} \leq C_{\mathrm{M}}(|v|_{1,E}|v|_{2,E} + h_E^{-1}|v|^2_{1,E}). \tag{6b}$$

On broken polynomial spaces $v_h \in V_h$, we obtain the discrete and inverse trace inequalities, respectively:

$$\|v_h\|_{0,F} \leq C_{\mathrm{tr}} h_E^{-1/2}\|v_h\|_{0,E}, \tag{7a}$$

$$\|\boldsymbol{\nabla}_h v_h\|_{0,E} \leq C_{\mathrm{inv}} h_E^{-1}\|v_h\|_{0,E}, \tag{7b}$$

where C_{tr} and C_{inv} are positive constants independent of h_E.

Remark 2.1. *Following Rivière [15] (see Section 2.1.3, p.24), one can obtain an exact expression of the constant C_{tr} used in the discrete trace inequality (7a) for a d-simplex mesh element:*

$$C_{\mathrm{tr}} := \sqrt{\frac{(k+1)(k+d)}{d}}, \tag{8}$$

where k denotes the polynomial degree of V_h and d denotes the spatial dimension. This expression is particularly important in our analysis since it will be used later in the definition of the penalty parameter.

We are now in a position to introduce the energy-norm used in the stability analysis and error estimations [13, 10]. For any given composite function $\boldsymbol{v}_h \in \boldsymbol{V}_h$, we consider the jump seminorm:

$$|\boldsymbol{v}_h|_\gamma^2 := \sum_{E\in\mathcal{T}_h} |\boldsymbol{v}_h|_{\gamma,\partial E}^2 \quad \text{with} \quad |\boldsymbol{v}_h|_{\gamma,\partial E}^2 := \sum_{F\in\mathcal{F}_E} \|\gamma_F^{1/2}[\![\boldsymbol{v}_h]\!]\|_{0,F}^2, \tag{9}$$

where $\gamma_F \geq 0$ is an arbitrary positive constant associated with $F \in \mathcal{F}_E$. The natural energy-norm equipping the discrete approximation space $\boldsymbol{V}_h$ is given by

$$\|\boldsymbol{v}_h\|_*^2 := \|\boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}_h v_h\|_{0,\mathcal{T}_h}^2 + |\boldsymbol{v}_h|_\gamma^2, \tag{10}$$

which clearly depends on $\boldsymbol{\kappa}$.

3. Hybridizable interior penalty methods

The discrete H-IP problem is to find $\boldsymbol{u}_h \in \boldsymbol{V}_h$ such that

$$\mathcal{B}_h^{(\epsilon)}(\boldsymbol{u}_h, \boldsymbol{v}_h) = l(\boldsymbol{v}_h), \quad \forall \boldsymbol{v}_h \in \boldsymbol{V}_h, \tag{11}$$

where $l(\boldsymbol{v}_h) := (f, v_h)_{0,\mathcal{T}_h}$ and the bilinear form $\mathcal{B}_h^{(\epsilon)} : \boldsymbol{V}_h \times \boldsymbol{V}_h \to \mathbb{R}$ is given by

$$\begin{aligned}\mathcal{B}_h^{(\epsilon)}(\boldsymbol{u}_h, \boldsymbol{v}_h) :=& (\boldsymbol{\kappa}\boldsymbol{\nabla}_h u_h, \boldsymbol{\nabla}_h v_h)_{0,\mathcal{T}_h} - \langle \boldsymbol{\kappa}\boldsymbol{\nabla}_h u_h, [\![\boldsymbol{v}_h]\!]\rangle_{0,\partial\mathcal{T}_h} \\ &- \epsilon\langle \boldsymbol{\kappa}\boldsymbol{\nabla}_h v_h, [\![\boldsymbol{u}_h]\!]\rangle_{0,\partial\mathcal{T}_h} + \langle \tau[\![\boldsymbol{u}_h]\!], [\![\boldsymbol{v}_h]\!]\rangle_{0,\partial\mathcal{T}_h},\end{aligned} \tag{12}$$

where $\epsilon \in \{0, \pm 1\}$. The second, third and fourth terms on the right-hand side of (12) are called the consistency, symmetry and penalty terms, respectively. The discrete bilinear operator $\mathcal{B}_h^{(\epsilon)}$ is symmetric iff $\epsilon = 1$ and is nonsymmetric otherwise. We obtain the symmetric scheme (H-SIP) if $\epsilon = 1$, the incomplete scheme (H-IIP) if $\epsilon = 0$ and the nonsymmetric scheme (H-NIP) if $\epsilon = -1$. For all $E \in \mathcal{T}_h$ and $F \in \mathcal{F}_E$, the penalty term is chosen as follows:

$$\tau_F := \frac{\gamma_0 C_{\mathrm{tr}}^2 \kappa_F}{h_F^{1+\delta}} \quad \text{with} \quad \delta \in \mathbb{R}, \tag{13}$$

where γ_0 is a user-dependent parameter, C_{tr} is given by (8) and results from the discrete trace inequality (7a), h_F is a local length scale associated with the interface F, and $\kappa_F := \boldsymbol{n}_F\boldsymbol{\kappa}_E\boldsymbol{n}_F$ denotes the normal diffusivity. We then assume that the quantity h_F satisfies the following *equivalence condition*, where for all $E \in \mathcal{T}_h$ and $F \in \mathcal{F}_E$, there exist positive constants ρ_1 and ρ_2 independent of h_E such that

$$\rho_1 h_E \leq h_F \leq \rho_2 h_E. \tag{14}$$

Remark 3.1. *Different choices of the local length scale h_F have been suggested in the literature, i.e., $h_F := \mathrm{diam}(F)$, $h_F := h_E$ (the diameter of E), $h_F := |F|$ (the Lebesgue measure of F) and $h_F := |E|/|F|$ (the Hausdorff measure of F) (see, e.g., [16]). For simplicity, we assume that $\boldsymbol{\kappa}$ is approximated by piecewise constants on the mesh element $\mathcal{T}_h$; i.e., $\boldsymbol{\kappa}_{|E} \in \mathbb{R}^{d\times d}$ for all $E \in \mathcal{T}_h$.*

Lemma 3.1 (Consistency). *Let $\boldsymbol{u} = (u, \hat{u}) \in \boldsymbol{V}$, where $u \in H^s(\Omega)$ is a solution of the weak problem (2) with $s > 3/2$. Then, the following holds:*

$$\mathcal{B}_h^{(\epsilon)}(\boldsymbol{u}, \boldsymbol{v}_h) = l(\boldsymbol{v}_h), \quad \forall \boldsymbol{v}_h \in \boldsymbol{V}_h. \tag{15}$$

Proof. The regularity of the weak solution implies that the quantities u and $\boldsymbol{\kappa}\boldsymbol{\nabla}_h u \cdot \boldsymbol{n}$ are single-valued fields on the mesh skeleton; i.e., $[\![\![\boldsymbol{u}]\!]\!] = 0$ for all $E \in \mathcal{T}_h$ and $F \in \mathcal{F}_E$, and $[\![\boldsymbol{\kappa}\boldsymbol{\nabla}_h u]\!] = 0$ for all $F \in \mathcal{F}_h^{\mathrm{i}}$, where $[\![\cdot]\!]$ denotes the standard jump operator as used in the DG method [8]. After integrating by parts, the bilinear form $\mathcal{B}_h^{(\epsilon)}$ yields

$$\begin{aligned}\mathcal{B}_h^{(\epsilon)}(\boldsymbol{u}, \boldsymbol{v}_h) =& (\boldsymbol{\kappa}\boldsymbol{\nabla}_h u, \boldsymbol{\nabla}_h v_h)_{0,\mathcal{T}_h} - \langle \boldsymbol{\kappa}\boldsymbol{\nabla}_h u, [\![\![\boldsymbol{v}_h]\!]\!]\rangle_{0,\partial\mathcal{T}_h}, \\ =& \sum_{E\in\mathcal{T}_h} (\boldsymbol{\nabla}_h \cdot (-\boldsymbol{\kappa}\boldsymbol{\nabla}_h u), v_h)_{0,E} + \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} \langle \underbrace{[\![\boldsymbol{\kappa}\boldsymbol{\nabla}_h u]\!]}_{=0}, \hat{v}_h\rangle_{0,F} = \sum_{E\in\mathcal{T}_h} (f, \boldsymbol{v}_h)_{0,E} \quad \forall \boldsymbol{v}_h \in \boldsymbol{V}_h,\end{aligned}$$

since $\hat{v}_h$ vanishes on the boundary skeleton $\mathcal{F}_h^{\mathrm{b}}$. This completes the proof. □

A straightforward consequence of the consistency property is the Galerkin orthogonality.

Proposition 3.1 (Galerkin orthogonality). *Let $\boldsymbol{u} = (u, \hat{u}) \in \boldsymbol{V}$, where $u \in H^s(\Omega)$ a solution of the weak problem (2) with $s > 3/2$. We denote by $\boldsymbol{u}_h \in \boldsymbol{V}_h$ the approximate solution of the discrete problem (11). Then,*

$$\mathcal{B}_h^{(\epsilon)}(\boldsymbol{u} - \boldsymbol{u}_h, \boldsymbol{v}_h) = 0 \quad \forall \boldsymbol{v}_h \in \boldsymbol{V}_h. \tag{16}$$

Proof. Subtracting (15) and (11) yields the assertion. □

3.1. Coercivity and well-posedness

The next step is to prove the key property, i.e., the discrete coercivity of the bilinear form $\mathcal{B}_h^{(\epsilon)}$, to ensure the well-posedness of the discrete problem (11). To this end, we first need to establish an upper bound of the consistency term using the jump seminorm $|\cdot|_\tau$.

Lemma 3.2 (Bound on consistency term). *Let $(\boldsymbol{w}_h, \boldsymbol{v}_h) \in \boldsymbol{V}_h \times \boldsymbol{V}_h$; then, there exists a constant $C_\delta > 0$ such that*

$$\left|\langle \boldsymbol{\kappa}\boldsymbol{\nabla}_h w_h, [\![\![\boldsymbol{v}_h]\!]\!]\rangle_{0,\partial\mathcal{T}_h}\right| \leq C_\delta^{1/2} \|\boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}_h w_h\|_{0,\mathcal{T}_h} |\boldsymbol{v}_h|_\tau, \tag{17}$$

where $C_\delta := C_0 h^\delta$ and $C_0 := C\eta_0\gamma_0^{-1}$ is a positive constant independent of h.

Proof. The decomposition of the consistency term yields

$$\langle \boldsymbol{\kappa}\boldsymbol{\nabla}_h w_h, [\![\![\boldsymbol{v}_h]\!]\!]\rangle_{0,\partial\mathcal{T}_h} = \sum_{E\in\mathcal{T}_h} \sum_{F\in\mathcal{F}_E} \langle \boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}_h w_h, \boldsymbol{\kappa}^{1/2}[\![\![\boldsymbol{v}_h]\!]\!]\rangle_{0,F}. \tag{18}$$

For any $F \in \mathcal{F}_h$, successively applying the Cauchy–Schwarz and the discrete trace inequalities (7a), using the definition (13) and the *equivalence condition* (14), we infer that

$$\begin{aligned}\left|\langle \boldsymbol{\kappa}\boldsymbol{\nabla}_h w_h, [\![\![\boldsymbol{v}_h]\!]\!]\rangle_{0,F}\right| \leq& \left[\frac{h_F^{1+\delta}}{\gamma_0 C_{\mathrm{tr}}^2}\right]^{1/2} (\|\boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}_h w_h\|_{0,F})(\tau_F^{1/2}\|[\![\![\boldsymbol{v}_h]\!]\!]\|_{0,F}), \\ \leq& \left[\frac{C h_E^\delta}{\gamma_0}\right]^{1/2} \|\boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}_h w_h\|_{0,E} |\boldsymbol{v}_h|_{\tau,F}.\end{aligned}$$

By summing over all interfaces $F \in \mathcal{F}_E$ and then over all mesh elements $E \in \mathcal{T}_h$ and by using the quasi-uniformity property of the mesh $\mathcal{T}_h$, we obtain the assertion

$$\left|\langle \boldsymbol{\kappa}\boldsymbol{\nabla}_h w_h, [\![\![\boldsymbol{v}_h]\!]\!]\rangle_{0,\partial\mathcal{T}_h}\right| \leq \left[\frac{C\eta_0}{\gamma_0} h^\delta\right]^{1/2} \|\boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}_h w_h\|_{0,\mathcal{T}_h} |\boldsymbol{v}_h|_\tau, \tag{19}$$

which concludes the proof. □

Lemma 3.3 (Coercivity)**.** *For a penalty parameter γ_0 that is large enough—i.e., $\gamma_0 > 4C\eta_0 h^\delta$—the discrete bilinear form $\mathcal{B}_h^{(\epsilon)}$ is $\boldsymbol{V}_h$-coercive with respect to the energy-norm $\|\cdot\|_*$; i.e.,*

$$\mathcal{B}_h^{(\epsilon)}(\boldsymbol{v}_h, \boldsymbol{v}_h) \geq \frac{1}{2}\|\boldsymbol{v}_h\|_*^2, \tag{20}$$

for all $\boldsymbol{v}_h \in \boldsymbol{V}_h$ and for any value of the parameter ϵ.

Proof. Setting $\boldsymbol{u}_h = \boldsymbol{v}_h$ in the definition of the bilinear form (12), we obtain

$$\mathcal{B}_h^{(\epsilon)}(\boldsymbol{v}_h, \boldsymbol{v}_h) = \|\boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}_h v_h\|_{0,\mathcal{T}_h}^2 - (1+\epsilon)\langle \boldsymbol{\kappa}\boldsymbol{\nabla}_h v_h, [\![v_h]\!]\rangle_{0,\partial\mathcal{T}_h} + |\boldsymbol{v}_h|_\tau^2. \tag{21}$$

Thus, owing to Lemmata 3.2 and using Young's inequality, for any $\zeta > 0$, there exists a constant $C_\zeta^{(\epsilon)} > 0$ such that

$$\mathcal{B}_h^{(\epsilon)}(\boldsymbol{v}_h, \boldsymbol{v}_h) \geq \left[1 - \frac{1+\epsilon}{2}\frac{C_\delta}{\zeta}\right]\|\boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}_h v_h\|_{0,\mathcal{T}_h}^2 + \left[1 - \frac{1+\epsilon}{2}\zeta\right]|\boldsymbol{v}_h|_\tau^2 \geq C_\zeta^{(\epsilon)}\|\boldsymbol{v}_h\|_*^2,$$

where $C_\zeta^{(\epsilon)}$ is given by

$$C_\zeta^{(\epsilon)} := 1 - \frac{1+\epsilon}{2}\max(C_\delta/\zeta, \zeta).$$

We now select γ_0 such that $C_\delta < \zeta^2$; i.e., $\gamma_0 > \zeta^{-2} C\eta_0 h^\delta$. Setting $\zeta = 1/2$, we easily bound $C_{1/2}^{(\epsilon)} \geq 1/2$ for any value of the parameter ϵ, thus completing the proof. □

Remark 3.2. *Note the h^δ-dependency of the coercivity condition. A straightforward consequence of the consistency and coercivity requirements via the Lax–Milgram Theorem is the well-posedness of the weak problem (11); i.e., the existence and uniqueness of the discrete solution $\boldsymbol{u}_h \in \boldsymbol{V}_h$ are ensured.*

3.2. Boundedness

We now assume that the discrete bilinear form $\mathcal{B}_h^{(\epsilon)}$ can be extended to $\boldsymbol{V}(h)\times\boldsymbol{V}(h)$, and we assert the boundedness of the product space . To this end, we introduce the enriched energy-norm on $\boldsymbol{V}(h)$ denoted by $|\!|\!|\cdot|\!|\!|$ (which is also a natural norm on $\boldsymbol{V}_h$) to bound the (normal) derivative terms [10]:

$$|\!|\!|\boldsymbol{v}|\!|\!|^2 := \|\boldsymbol{v}\|_*^2 + \sum_{E\in\mathcal{T}_h} h_E\|\boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}_h v\|_{0,\partial E}^2, \quad \forall \boldsymbol{v} \in \boldsymbol{V}(h). \tag{22}$$

Lemma 3.4 (Equivalency of $\|\cdot\|_*$- and $|\!|\!|\cdot|\!|\!|$-norms)**.** *For all $\boldsymbol{v} \in \boldsymbol{V}(h)$, the norms $\|\boldsymbol{v}\|_*$ and $|\!|\!|\boldsymbol{v}|\!|\!|$ are equivalent; i.e., there exists a constant $\rho > 0$ such that*

$$\rho^{-1}|\!|\!|\boldsymbol{v}|\!|\!| \leq \|\boldsymbol{v}\|_* \leq |\!|\!|\boldsymbol{v}|\!|\!|, \tag{23}$$

where $\rho := (1 + \eta_0 C_{\mathrm{tr}}^2)^{\frac{1}{2}}$ depends only on the element shape.

Proof. Following the definition (22), we notice that $\|\boldsymbol{v}\|_* \leq |\!|\!|\boldsymbol{v}|\!|\!|$. We now can easily bound the difference of both norms by using the discrete trace inequality (7a)

$$|\!|\!|\boldsymbol{v}|\!|\!|^2 - \|\boldsymbol{v}\|_*^2 \leq \eta_0 C_{\mathrm{tr}}^2\|\boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}_h v\|_{0,\mathcal{T}_h}^2 \leq \eta_0 C_{\mathrm{tr}}^2\|\boldsymbol{v}\|_*^2, \tag{24}$$

which yields the assertion. □

Lemma 3.5 (Boundedness with h^δ-dependency)**.** *For all $(\boldsymbol{w}, \boldsymbol{v}) \in \boldsymbol{V}(h)\times\boldsymbol{V}(h)$, there exists a constant $C_{\mathrm{bnd}} > 0$ such that*

$$\mathcal{B}_h^{(\epsilon)}(\boldsymbol{w}, \boldsymbol{v}) \leq C_{\mathrm{bnd}}|\!|\!|\boldsymbol{w}|\!|\!|\cdot|\!|\!|\boldsymbol{v}|\!|\!|, \tag{25}$$

where $C_{\mathrm{bnd}} := 2 + C_1 h^\delta$ and $C_1 := (\gamma_0 C_{\mathrm{tr}}^2)^{-1}$ is a positive constant independent of h.

Proof. Following the definition of the bilinear form (12), we deduce that

$$\begin{aligned}\left|\mathcal{B}_h^{(\epsilon)}(\boldsymbol{w},\boldsymbol{v})\right| &\leq \left|(\boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}_h w, \boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}_h v)_{0,\mathcal{T}_h} + \langle \tau^{1/2}[\![[\boldsymbol{w}]]\!], \tau^{1/2}[\![[\boldsymbol{v}]]\!]\rangle_{0,\partial\mathcal{T}_h}\right| + \\ &\quad \left|\langle \boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}_h w, \boldsymbol{\kappa}^{1/2}[\![[\boldsymbol{v}]]\!]\rangle_{0,\partial\mathcal{T}_h}\right| + |\epsilon|\left|\langle \boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}_h v, \boldsymbol{\kappa}^{1/2}[\![[\boldsymbol{w}]]\!]\rangle_{0,\partial\mathcal{T}_h}\right| \\ &\leq |\mathcal{T}_1 + \mathcal{T}_2| + |\mathcal{T}_3| + |\epsilon|\,|\mathcal{T}_4|\,.\end{aligned}$$

Applying the Cauchy–Schwarz inequality, the first two terms can be bounded as follows:

$$|\mathcal{T}_1 + \mathcal{T}_2| \leq [\|\boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}_h w\|_{0,\mathcal{T}_h}^2 + |\boldsymbol{w}|_\tau^2]^{1/2}[\|\boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}_h v\|_{0,\mathcal{T}_h}^2 + |\boldsymbol{v}|_\tau^2]^{1/2} = \|\boldsymbol{w}\|_*\|\boldsymbol{v}\|_*. \tag{26}$$

Proceeding as in the proof of Lemmata 3.2, the third and fourth terms can also be bounded as follows:

$$|\mathcal{T}_3| \leq \left[C_1 h^\delta \sum_{E\in\mathcal{T}_h} h_E\|\boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}_h w\|_{0,\partial E}^2\right]^{1/2}\|\boldsymbol{v}\|_*, \tag{27a}$$

$$|\mathcal{T}_4| \leq \left[C_1 h^\delta \sum_{E\in\mathcal{T}_h} h_E\|\boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}_h v\|_{0,\partial E}^2\right]^{1/2}\|\boldsymbol{w}\|_*, \tag{27b}$$

where $C_1 := (\gamma_0 C_{\mathrm{tr}}^2)^{-1}$. By combining these estimates via the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}\left|\mathcal{B}_h^{(\epsilon)}(\boldsymbol{w},\boldsymbol{v})\right| &\leq \left[(1+|\epsilon|)\|\boldsymbol{w}\|_*^2 + C_1 h^\delta \sum_{E\in\mathcal{T}_h} h_E\|\boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}_h w\|_{0,\partial E}^2\right]^{1/2}\left[2\|\boldsymbol{v}\|_*^2 + |\epsilon|\,C_1 h^\delta \sum_{E\in\mathcal{T}_h} h_E\|\boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}_h v\|_{0,\partial E}^2\right]^{1/2}, \\ &\leq \max(2, C_1 h^\delta)|\!|\!|\boldsymbol{w}|\!|\!|\boldsymbol{v}|\!|\!|,\end{aligned}$$

which yields the assertion. □

Remark 3.3. *Let us emphasize that* $C_{\mathrm{bnd}} \leq Ch^r$*, where* $r = \min(0,\delta)$ *and* $C := 2\max(2, C_1)$ *and is a positive constant independent of* h*.*

4. A priori error analysis

We now derive *a priori* error estimates in both the discrete $\|\cdot\|_*$- and $\|\cdot\|_{0,\mathcal{T}_h}$-norms to show the accuracy of the H-IP method. To this end, we first recall some definitions such as the continuous interpolant and derive standard interpolation estimates that will be used extensively in the rest of the document (for more details, we refer the reader to [16, 19]). Let us introduce π_h^{i} and π_h^{b}, the standard L^2-orthogonal projectors on the discrete approximation spaces V_h and $\hat{V}_h$, respectively. Then, if $\phi \in H^s(\Omega)$ with $s \geq 2$, the standard interpolation estimate is written as

$$|\phi - \pi_h^{\mathrm{i}}\phi|_{q,\mathcal{T}_h} \leq Ch^{\mu-q}|\phi|_{\mu,\mathcal{T}_h}, \quad \forall q \in \{0,\ldots,s-1\}, \tag{28a}$$

$$\left[\sum_{E\in\mathcal{T}_h} h_E^\alpha\|\boldsymbol{\nabla}_h(\phi - \pi_h^{\mathrm{i}}\phi)\|_{0,\partial E}^2\right]^{1/2} \leq Ch^{\mu+\frac{\alpha-3}{2}}|\phi|_{\mu,\mathcal{T}_h}, \tag{28b}$$

where $\mu := \min(k+1, s)$ and k denote the polynomial degrees of approximation spaces V_h and $\hat{V}_h$, respectively.

Lemma 4.1 (Optimal error estimates)**.** *Let* $\boldsymbol{u} := (u,\hat{u}) \in H^s(\mathcal{T}_h)\times L^2(\mathcal{F}_h)$*, where* u *is the weak solution of* (2) *and* $s > 3/2$*. We denote by* $\pi_h\boldsymbol{u} := (\pi_h^{\mathrm{i}}u, \pi_h^{\mathrm{b}}\hat{u})$ *the continuous interpolant of the composite variable* $\boldsymbol{u}$*, which is contained in* $\boldsymbol{V}_h$*; i.e.,* $\pi_h\boldsymbol{u} \in \boldsymbol{V}_h$*. Then,*

$$\|\boldsymbol{u} - \pi_h\boldsymbol{u}\|_* \leq |\!|\!|\boldsymbol{u} - \pi_h\boldsymbol{u}|\!|\!| \leq C_{\boldsymbol{\kappa}} h^{\mu_0}|u|_{\mu,\mathcal{T}_h}, \tag{29}$$

where $\mu_0 := \min(k, s-1)$ *and* $C_{\boldsymbol{\kappa}} := C\|\boldsymbol{\kappa}^{1/2}\|_{\infty,\Omega}$.

Proof. Successively using the definition of the $|||\cdot|||$-norm (22), the Cauchy–Schwarz inequality, and the interpolation estimates (28) yields

$$\begin{aligned}\|\boldsymbol{u}-\pi_h\boldsymbol{u}\|_*^2 \overset{(23)}{\leq} |||\boldsymbol{u}-\pi_h\boldsymbol{u}|||^2 &\overset{(22)}{=} \|\boldsymbol{u}-\pi_h\boldsymbol{u}\|_*^2 + \sum_{E\in\mathcal{T}_h} h_E\|\boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}_h(u-\pi_h^{\mathrm{i}}u)\|_{0,\partial E}^2,\\ &\overset{(28)}{\leq} \|\boldsymbol{\kappa}^{1/2}\|_{\infty,\Omega}^2 \sum_{E\in\mathcal{T}_h}(|u-\pi_h^{\mathrm{i}}u|_{1,E}^2 + h_E\|\boldsymbol{\nabla}_h(u-\pi_h^{\mathrm{i}}u)\|_{0,\partial E}^2),\\ &\leq C^2\|\boldsymbol{\kappa}^{1/2}\|_{\infty,\Omega}^2 h^{2\mu-2}|u|_{\mu,\mathcal{T}_h}^2,\end{aligned}$$

which concludes the proof. □

4.1. Energy-norm error estimates

We now derive an error estimation of the discrete composite variable $\boldsymbol{u}_h$ in the natural $\|\cdot\|_*$-norm.

Theorem 4.1 ($\|\cdot\|_*$-norm estimate and optimal convergence rate)**.** *Let $\boldsymbol{u} := (u,\hat{u}) \in H^s(\Omega)\times L^2(\mathcal{F}_h)$, where u is a solution of (2) with $s > 3/2$. We denote by $\boldsymbol{u}_h \in \boldsymbol{V}_h$ the approximate solution of the discrete problem (11). Then, for any value of the parameter δ, the following estimate holds:*

$$\|\boldsymbol{u}-\boldsymbol{u}_h\|_* \leq |||\boldsymbol{u}-\boldsymbol{u}_h||| \leq C_{\boldsymbol{\kappa}} h^{\mu_0+r}|u|_{\mu,\mathcal{T}_h}, \tag{30}$$

where $\mu_0 := \min(k, s-1)$, $r := \min(0,\delta)$, and $C_{\boldsymbol{\kappa}} := C\|\boldsymbol{\kappa}^{1/2}\|_{\infty,\Omega}$.

Proof. We decompose this quantity as $\boldsymbol{u}-\boldsymbol{u}_h = \boldsymbol{u}-\pi_h\boldsymbol{u}+\pi_h\boldsymbol{u}-\boldsymbol{u}_h$. By using the triangle inequality, we easily infer that

$$|||\boldsymbol{u}-\boldsymbol{u}_h||| \leq |||\boldsymbol{u}-\pi_h\boldsymbol{u}||| + |||\pi_h\boldsymbol{u}-\boldsymbol{u}_h|||. \tag{31}$$

Only an upper bound on the last term of (31) remains to be established. Successively using the coercivity, energy-norm equivalency, Galerkin orthogonality, and boundedness, we deduce that

$$\begin{aligned}\frac{1}{2\rho^2}|||\pi_h\boldsymbol{u}-\boldsymbol{u}_h|||^2 \overset{(23)}{\leq} \frac{1}{2}\|\pi_h\boldsymbol{u}-\boldsymbol{u}_h\|_*^2 &\overset{(20)}{\leq} \mathcal{B}_h^{(\epsilon)}(\pi_h\boldsymbol{u}-\boldsymbol{u}_h,\pi_h\boldsymbol{u}-\boldsymbol{u}_h),\\ &\overset{(16)}{\leq} \mathcal{B}_h^{(\epsilon)}(\pi_h\boldsymbol{u}-\boldsymbol{u},\pi_h\boldsymbol{u}-\boldsymbol{u}_h) \overset{(25)}{\leq} C_{\mathrm{bnd}}|||\boldsymbol{u}-\pi_h\boldsymbol{u}|||\,|||\pi_h\boldsymbol{u}-\boldsymbol{u}_h|||,\end{aligned}$$

and then we insert $|||\pi_h\boldsymbol{u}-\boldsymbol{u}_h||| \leq 2\rho^2 C_{\mathrm{bnd}}|||\boldsymbol{u}-\pi_h\boldsymbol{u}|||$ into (31) to obtain

$$\|\boldsymbol{u}-\boldsymbol{u}_h\|_* \overset{(23)}{\leq} |||\boldsymbol{u}-\boldsymbol{u}_h||| \leq (1+2\rho^2 C_{\mathrm{bnd}})|||\boldsymbol{u}-\pi_h\boldsymbol{u}|||.$$

Proceeding as in Remark 3.3, we can conclude that there exists a positive constant C such that $1+2\rho^2 C_{\mathrm{bnd}} \leq Ch^r$, which yields the assertion. □

Corollary 4.1 (Estimate for strong-regularity solutions)**.** *Assume that $s \geq k+1$ with $u \in H_0^{k+1}(\mathcal{T}_h)$ and $\delta \in \mathbb{R}$. Then, we have the estimate*

$$\|\boldsymbol{u}-\boldsymbol{u}_h\|_* \leq C_{\mathrm{u},\boldsymbol{\kappa}} h^{k+r}, \tag{32}$$

where $r := \min(0,\delta)$, $C_{\mathrm{u},\boldsymbol{\kappa}} := C_{\boldsymbol{\kappa}}|u|_{k+1,\mathcal{T}_h}$ and $C_{\boldsymbol{\kappa}} := C\|\boldsymbol{\kappa}^{1/2}\|_{\infty,\Omega}$.

Proof. (Evident) □

Remark 4.1. *Following Di Pietro and Ern, since C in Theorem 4.1 is independent of $\boldsymbol{\kappa}$, the discrete method is said to be robust with respect to the anisotropy and heterogeneity of the diffusion tensor. The given estimate (32) indicates that the order of convergence in the $\|\cdot\|_*$-norm, or equivalently, $|||\cdot|||$-norm, is linear and δ-dependent, i.e., suboptimal if $\delta < 0$ and optimal otherwise.*

4.2. L^2-norm error estimate

Using a standard Aubin–Nitsche duality argument, we now derive an improved L^2-error estimate of the H-IP method in terms of the parameter δ. To this end, we define an auxiliary function ψ as the solution of the adjoint problem:

$$-\boldsymbol{\nabla}\cdot(\boldsymbol{\kappa}\boldsymbol{\nabla}\psi) = u - u_h \quad \text{in } \Omega, \quad \text{and} \quad \psi = 0 \quad \text{on } \partial\Omega. \tag{33}$$

By assuming elliptic regularity, the following estimate holds:

$$\|\psi\|_{2,\Omega} \leq C_{\boldsymbol{\kappa}}\|u - u_h\|_{0,\Omega}, \tag{34}$$

where $C_{\boldsymbol{\kappa}}$ depends on the shape regularity (i.e., the convexity) of Ω and the distribution of $\boldsymbol{\kappa}$ inside it [20]. The weak adjoint problem is to find $\psi \in H^2(\Omega)\cap H_0^1(\Omega)$ such that

$$(\boldsymbol{\kappa}\boldsymbol{\nabla}_h\psi, \boldsymbol{\nabla}_h v)_{0,\mathcal{T}_h} - \langle \boldsymbol{\kappa}\boldsymbol{\nabla}_h\psi\cdot\boldsymbol{n}, v\rangle_{0,\partial\mathcal{T}_h} = (u-u_h, v)_{0,\mathcal{T}_h}, \quad \forall v \in H_0^1(\Omega). \tag{35}$$

By setting $v = u - u_h$ in (35), we obtain

$$\|u-u_h\|_{0,\mathcal{T}_h}^2 = (\boldsymbol{\kappa}\boldsymbol{\nabla}_h\psi, \boldsymbol{\nabla}_h(u-u_h))_{0,\mathcal{T}_h} - \langle \boldsymbol{\kappa}\boldsymbol{\nabla}_h\psi, (u-u_h)\boldsymbol{n}\rangle_{0,\partial\mathcal{T}_h}. \tag{36}$$

Let us now introduce the composite error variable $\boldsymbol{e}_h^u := \boldsymbol{u} - \boldsymbol{u}_h = (e_h^u, \hat{e}_h^u)$. From the regularity of the variables $\hat{u}$, $\hat{u}_h$ and ψ, we deduce that $\langle \boldsymbol{\kappa}\boldsymbol{\nabla}_h\psi, (\hat{u}-\hat{u}_h)\boldsymbol{n}\rangle_{0,\partial\mathcal{T}_h} = 0$. By embedding this condition in (36), we obtain an equivalent reformulation of the weak adjoint problem in terms of the discrete bilinear operator $\mathcal{B}_h^{(\epsilon)}$:

$$\|e_h^u\|_{0,\mathcal{T}_h}^2 = (\boldsymbol{\kappa}\boldsymbol{\nabla}\psi, \boldsymbol{\nabla}e_h^u)_{0,\mathcal{T}_h} - \langle \boldsymbol{\kappa}\boldsymbol{\nabla}\psi, [\![\boldsymbol{e}_h^u]\!]\rangle_{0,\partial\mathcal{T}_h} = \mathcal{B}_h^{(\epsilon)}(\boldsymbol{\psi}, \boldsymbol{e}_h^u), \tag{37}$$

where $\boldsymbol{\psi} := (\psi, \hat{\psi})$. Following the definition of the bilinear form $\mathcal{B}_h^{(\epsilon)}$ (12) and using the Galerkin orthogonality $\mathcal{B}_h^{(\epsilon)}(\boldsymbol{e}_h^u, \pi_h\boldsymbol{\psi}) = 0$, since $\pi_h\boldsymbol{\psi} \in \boldsymbol{V}_h$ (see Proposition 3.1), we easily infer

$$\mathcal{B}_h^{(\epsilon)}(\boldsymbol{\psi}, \boldsymbol{e}_h^u) = \mathcal{B}_h^{(\epsilon)}(\boldsymbol{e}_h^u, \boldsymbol{e}_\pi^\psi) - (1-\epsilon)\langle \boldsymbol{\kappa}\boldsymbol{\nabla}\psi, [\![\boldsymbol{e}_h^u]\!]\rangle_{0,\partial\mathcal{T}_h} := \mathcal{T}_1^{(\epsilon)} - (1-\epsilon)\mathcal{T}_2, \tag{38}$$

where $\boldsymbol{e}_\pi^\psi := \boldsymbol{\psi} - \pi_h\boldsymbol{\psi}$. We will now determine an upper bound of the quantity $\|e_h^u\|_{0,\mathcal{T}_h}^2$. Owing to Lemmas 3.5 and 4.1 and using the regularity assumption $\psi \in H^2(\Omega)$, we can bound the first term $\mathcal{T}_1$:

$$\left|\mathcal{T}_1^{(\epsilon)}\right| \leq C_{\mathrm{bnd}}|||\boldsymbol{e}_\pi^\psi|||\,|||\boldsymbol{e}_h^u||| \leq C_{\boldsymbol{\kappa}}C_{\mathrm{bnd}}h\|\psi\|_{2,\Omega}|||\boldsymbol{e}_h^u|||. \tag{39}$$

Using the trace inequality $\|\boldsymbol{\nabla}_h\psi\|_{0,\partial\mathcal{T}_h} \leq Ch^{-1/2}\|\psi\|_{2,\Omega}$ [19], the second term $\mathcal{T}_2$ can be bounded as follows:

$$|\mathcal{T}_2| \leq C_{\boldsymbol{\kappa}}h^{\frac{1+\delta}{2}}\|\boldsymbol{\nabla}_h\psi\|_{0,\partial\mathcal{T}_h}|\boldsymbol{e}_h^u|_\tau \leq C_{\boldsymbol{\kappa}}h^{\frac{\delta}{2}}\|\psi\|_{2,\Omega}|||\boldsymbol{e}_h^u|||. \tag{40}$$

Combining (39) and (40), we obtain the estimate

$$\|u-u_h\|_{0,\mathcal{T}_h} \leq C_{\boldsymbol{\kappa}}(C_{\mathrm{bnd}}h + (1-\epsilon)h^{\frac{\delta}{2}})|||\boldsymbol{e}_h^u|||, \tag{41}$$

and we can assert the theorem below.

Theorem 4.2 (L^2-norm estimate)**.** *Let $\boldsymbol{u} := (u, \hat{u}) \in H^s(\Omega)\times L^2(\mathcal{F}_h)$, where u is a solution of (2) with $s > 3/2$. We denote by $\boldsymbol{u}_h \in \boldsymbol{V}_h$ the approximate solution of the discrete problem (11). Then, for any value of the parameters δ and $\epsilon \in \{0, \pm 1\}$, the following estimate holds for the H-IP method:*

$$\|u-u_h\|_{0,\mathcal{T}_h} \leq C_{\boldsymbol{\kappa}}h^{\mu_0+s_\delta^{(\epsilon)}}|u|_{s,\mathcal{T}_h}, \tag{42}$$

where $C_{\boldsymbol{\kappa}} := C\|\boldsymbol{\kappa}^{1/2}\|_{\infty,\Omega}$, $\mu_0 := \min(k, s-1)$, and the parameter $s_\delta^{(\epsilon)}$ is only dependent on ϵ and δ and is given by

$$s_\delta^{(\epsilon)} := \begin{cases} \min(1, 1+2\delta) & \text{if} \quad \epsilon = 1, \\ \min(1, \delta/2) & \text{if} \quad \epsilon \neq 1 \quad \text{and} \quad \delta \geq 0, \\ \min(1+2\delta, 3\delta/2) & \text{if} \quad \epsilon \neq 1 \quad \text{and} \quad \delta < 0. \end{cases} \tag{43}$$

Proof. The estimate (42) using (43) follows after some algebraic manipulations from the previous equation (41), the definition of C_{bnd} given in Lemma 3.5 and the optimal error estimate given in Lemma 4.1. □

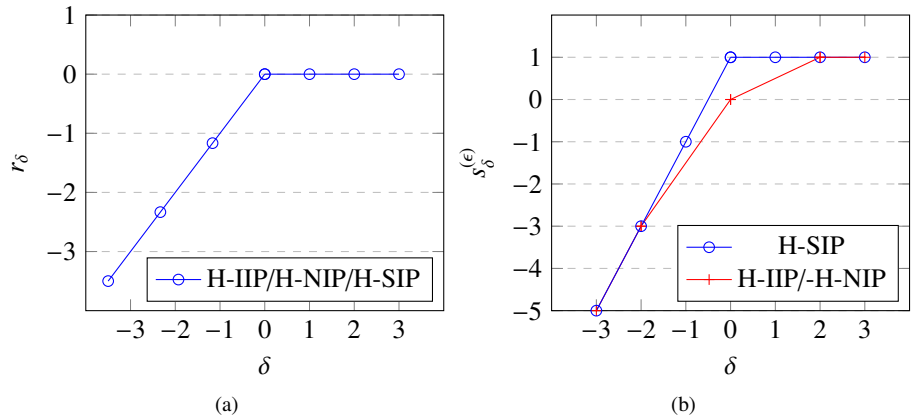


Figure 1: Representation of the quantities r_δ and $s_\delta^{(\epsilon)}$ vs. δ given in Theorems 4.1 and 4.2, respectively.

Remark 4.2. *The authors are certain that the estimates given in Theorems 4.1 and 4.2 have already been established in the literature, but we have not been able to find them.*

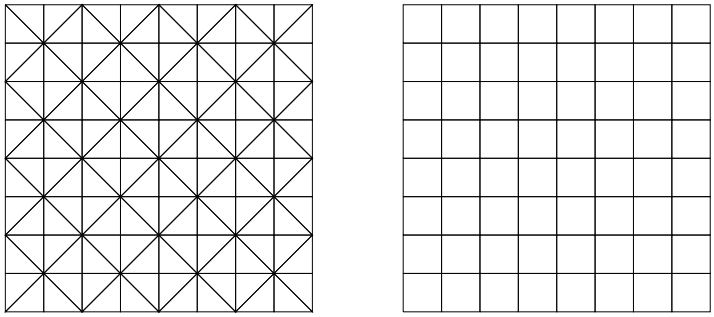
5. Numerical experiments

In the previous sections, we built families of hybridizable interior penalty methods based on an adaptive definition of the penalty parameter that depends on several coefficients. This section highlights the benefit these methods provide in the approximation of diffusion problems with anisotropic and/or discontinuous coefficients and in the validation of a priori error estimates. In the rest of the document, we assume that the local length scale h_F in (13) is chosen to be equal to the diameter of the associated element, i.e., $h_F := h_E$, for all $E \in \mathcal{T}_h$ and for all $F \in \mathcal{F}_E$. All numerical experiments are performed using the high-performance finite element library NGSolve [21]. Then, the physical domain is taken to be a unit square—i.e., $\Omega := [0,1]^2 \subset \mathbb{R}^2$—and the right-hand-side f is chosen such that the given exact solution u respecting the homogeneous boundary conditions is verified. We use a sequence of subdivisions $\mathcal{T}_h$, where regular triangles or squares form each partition (see, e.g., Figure 2). Standard h- and k-refinement strategies are used to compute the numerical errors and estimated convergence rates (ECRs). To pursue our quantitative analysis, we first measure the impact of the parameter δ on the a posteriori error estimates. Second, we point out the crucial role of the factor κ_n arising in (12) for the robustness of the H-IP methods when the medium becomes highly anisotropic and/or discontinuous. Finally, we complete our experiments by pointing out some unexpected benefits of the value of γ_0 for the ECRs of the H-SIP scheme.

5.1. Test A: Influence of the parameter δ

We consider the following test case, which was previously proposed in Fabien *et al.* [11]: the diffusion tensor is homogeneous and isotropic—$\boldsymbol{\kappa} = \boldsymbol{I_2}$ (identity matrix)—and the exact smooth solution is given by $u(x,y) = xy(1-x)(1-y)\exp(-x^2-y^2)$. Then, for all $E \in \mathcal{T}_h$ and for all $F \in \mathcal{F}_E$, we assume that the penalty parameter has the following simplified form:

$$\tau_F := \frac{\tau_0}{h_E^{1+\delta}}, \tag{44}$$



(a) (b)

Figure 2: Uniform triangular (a) and square (b) meshes with $h = 1/8$, respectively.

where $\tau_0 > 0$ is a positive constant chosen to be large enough in accordance with Lemma 3.3. The objective here is to measure the impact of the parameter δ on the ECRs in both the L^2 and energy-norms. A history of convergence is shown in Figures 3 ($|\!|\!| \cdot |\!|\!|$-norm) and 4 ($\| \cdot \|_{0,\mathcal{T}_h}$-norm) for uniform triangular meshes and for polynomial degrees $k \in \{1, \cdots, 3\}$, and Table 1 summarizes our numerical observations.

		H-IIP			H-NIP			H-SIP	
Degree	Norm	$\delta = -1$	$\delta = 0$	$\delta \geq 2$	$\delta = -1$	$\delta = 0$	$\delta \geq 2$	$\delta = -1$	$\delta = 0$
$k = 1$	$\| \cdot \|_{0,\mathcal{T}_h}$	1.0	2.0	–	1.0	2.0	–	1.0	2.0
	$\|\!\| \cdot \|\!\|$	1.0	1.0	–	1.0	1.0	–	1.0	1.0
odd k	$\| \cdot \|_{0,\mathcal{T}_h}$	k	$k+1$	–	$k+1$	–	–	$k+1$	–
	$\|\!\| \cdot \|\!\|$	$k-1$	k	–	k	–	–	k	–
even k	$\| \cdot \|_{0,\mathcal{T}_h}$	$k-1$	k	$k+1$	k	k	$k+1$	$k+1$	–
	$\|\!\| \cdot \|\!\|$	$k-1$	k	k	k	k	k	k	–

Table 1: Test A: a summary of the ECRs in the L^2- and energy-norm of H-IP methods in terms of the parameter δ and the polynomial parity k.

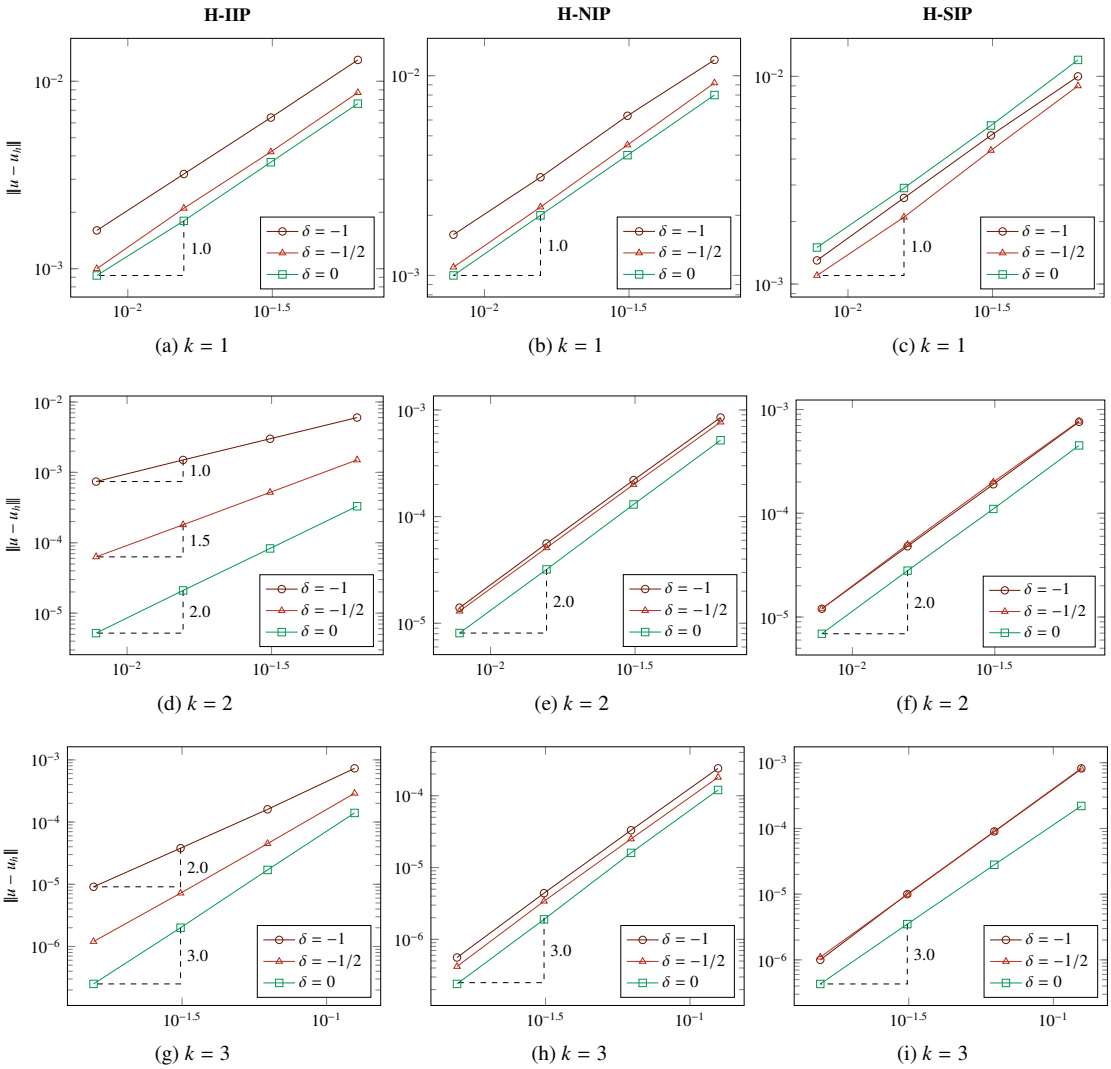
As expected, these observations are in agreement with theoretical estimates and underline that the stabilization parameter δ influences the convergence rate. In particular, we recover some well-known estimates if $\delta = 0$. First, we notice that the convergence of the H-IP method in the energy-norm is linearly δ-dependent if $\delta \leq 0$ and optimal if $\delta \geq 0$, which is in accordance with Lemma 4.1 (see Figure 3). A brief analysis of the convergence in the L^2-norm indicates that both the H-IIP and H-NIP schemes behave differently from the H-SIP scheme. Nonsymmetric variants are strongly influenced by the polynomial parity of k and by the penalty parameter δ. We observe that the convergence rate increases linearly and optimally if $\delta \geq 0$ for odd k and $\delta \geq 2$ for even k. In this last case, let us point out that the optimal convergence is nearly reached once $\delta \geq 1$. As expected, the symmetric scheme converges optimally when $\delta \geq 0$. These results agree with the theoretical results established in Theorem 4.2.

5.2. *Test B: Influence of the parameter κ_F*

In the second experiment, we analyze the behavior of the discretization method in the context of genuine anisotropic and heterogeneous properties. Then, the unit square Ω is split into four subdomains $\Omega_1 = [0, 1/2]^2$, $\Omega_2 = [1/2, 1] \times [0, 1/2]$, $\Omega_3 = [1/2, 1]^2$ and $\Omega_4 = [0, 1/2] \times [1/2, 1]$, such that $\Omega := \cup_{i=1}^4 \Omega_i$. The exact solution on the whole domain Ω is given by $u(x, y) = \sin(\pi x) \sin(\pi y)$, and the diffusivity tensor takes different values in each subregion:

$$\boldsymbol{\kappa} = \begin{bmatrix} 1 & 0 \\ 0 & \lambda \end{bmatrix} \quad \text{for } (x, y) \in \Omega_1,\, \Omega_3, \quad \text{and} \quad \boldsymbol{\kappa} = \begin{bmatrix} 1/\lambda & 0 \\ 0 & 1 \end{bmatrix} \quad \text{for } (x, y) \in \Omega_2,\, \Omega_4, \tag{45}$$

where the parameter $\lambda > 0$ simultaneously controls both the anisotropy and the medium heterogeneity. Here, we focus on the influence of the parameter κ_F on the robustness of the discretization method in the context of highly anisotropic and heterogeneous coefficients, and we choose $\lambda = 10^{-3}$. In this context, the anisotropy and heterogeneity

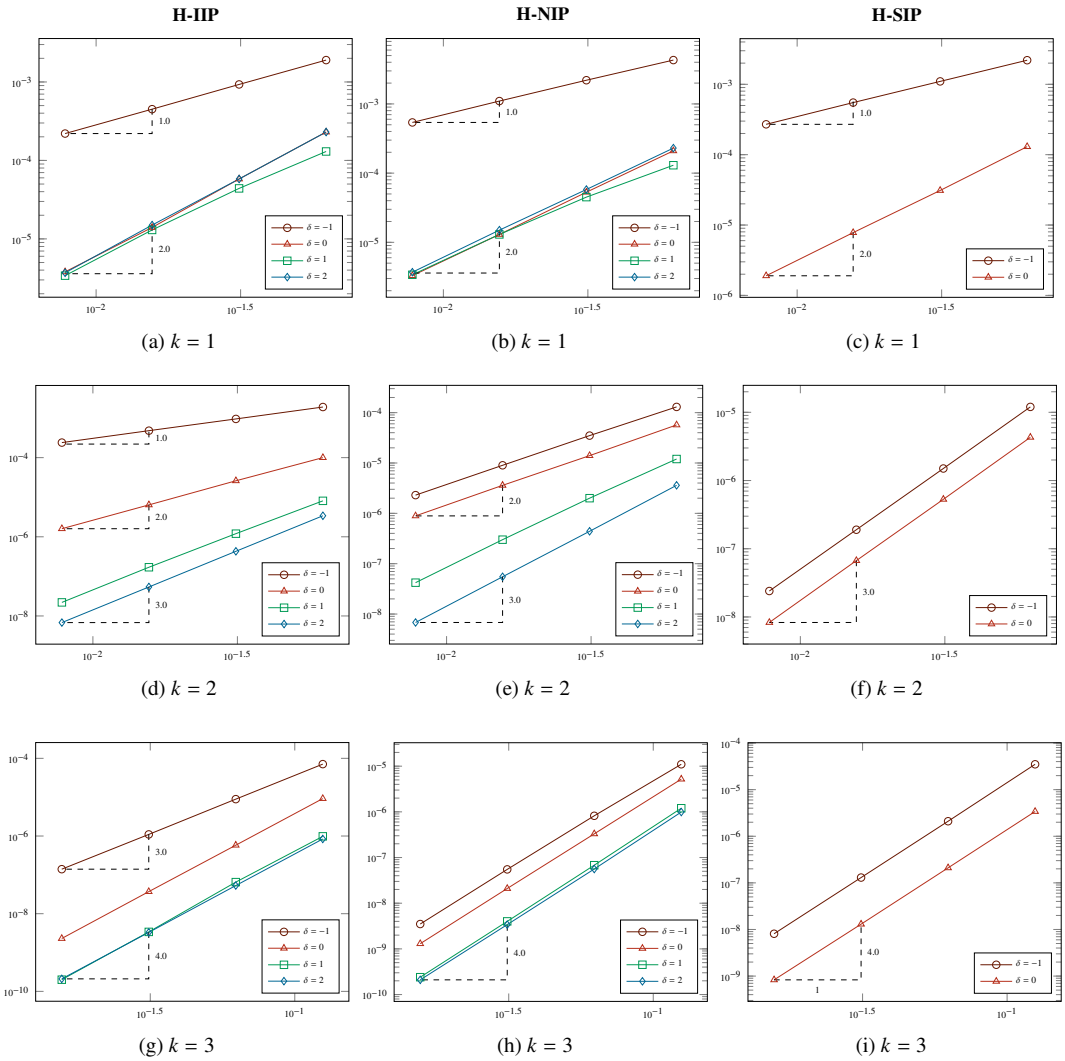


(a) $k = 1$ (b) $k = 1$ (c) $k = 1$

(d) $k = 2$ (e) $k = 2$ (f) $k = 2$

(g) $k = 3$ (h) $k = 3$ (i) $k = 3$

Figure 3: Test A: history of convergence of the H-IP methods with $-1 \leq \delta \leq 0$: $\|u - u_h\|$ vs. h for the three H-IP variants and various polynomial degrees ($1 \leq k \leq 3$) on uniform triangular meshes.



(a) $k = 1$ (b) $k = 1$ (c) $k = 1$

(d) $k = 2$ (e) $k = 2$ (f) $k = 2$

(g) $k = 3$ (h) $k = 3$ (i) $k = 3$

Figure 4: Test A: history of convergence of the H-IP methods with $-1 \leq \delta \leq 2$: $\|u - u_h\|_{0,\mathcal{T}_h}$ vs. h for the three H-IP variants and various polynomial degrees ($1 \leq k \leq 3$) on uniform triangular meshes.

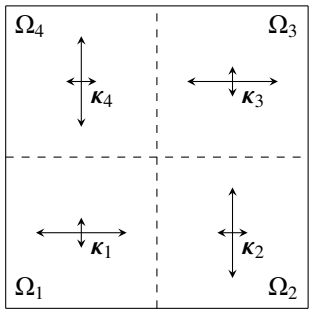


Figure 5: Description of test case B with genuine anisotropic and heterogeneous properties.

ratios are approximately 10^3 and 10^6, respectively. For the simulations, we consider a conforming triangular mesh ($h = 1/32$) respecting the discontinuities of $\boldsymbol{\kappa}$, we use piecewise linear approximations of the discrete variable u_h, and we set $\delta = 0$ in the definition of the penalty parameter (13). Here, the comparisons are only graphical (Figure 6). We depict the discrete solutions u_h obtained successively using $\kappa_F := 1$ (Case 1) and $\kappa_F := \boldsymbol{n}_F\boldsymbol{\kappa}_E\boldsymbol{n}_F$ (Case 2) for all variations of $\epsilon \in \{0, \pm 1\}$. In the first situation (Figures 6-a, 6-b and 6-c), the discrete solutions exhibit spurious oscillations and erratic behavior, thus violating the discrete maximum principle (see, e.g., Table 2). This can be easily explained by observing that the first formulation does not distinguish between the principal directions of the diffusivity tensor. Consequently, a misestimated penalty is applied in directions of low or high diffusivity. In the second situation (Figures 6-d, 6-e and 6-f), the jumps in diffusivity are better captured at the interfaces of discontinuities, and the discrete solutions are significantly more robust, i.e., exhibit less erratic behavior.

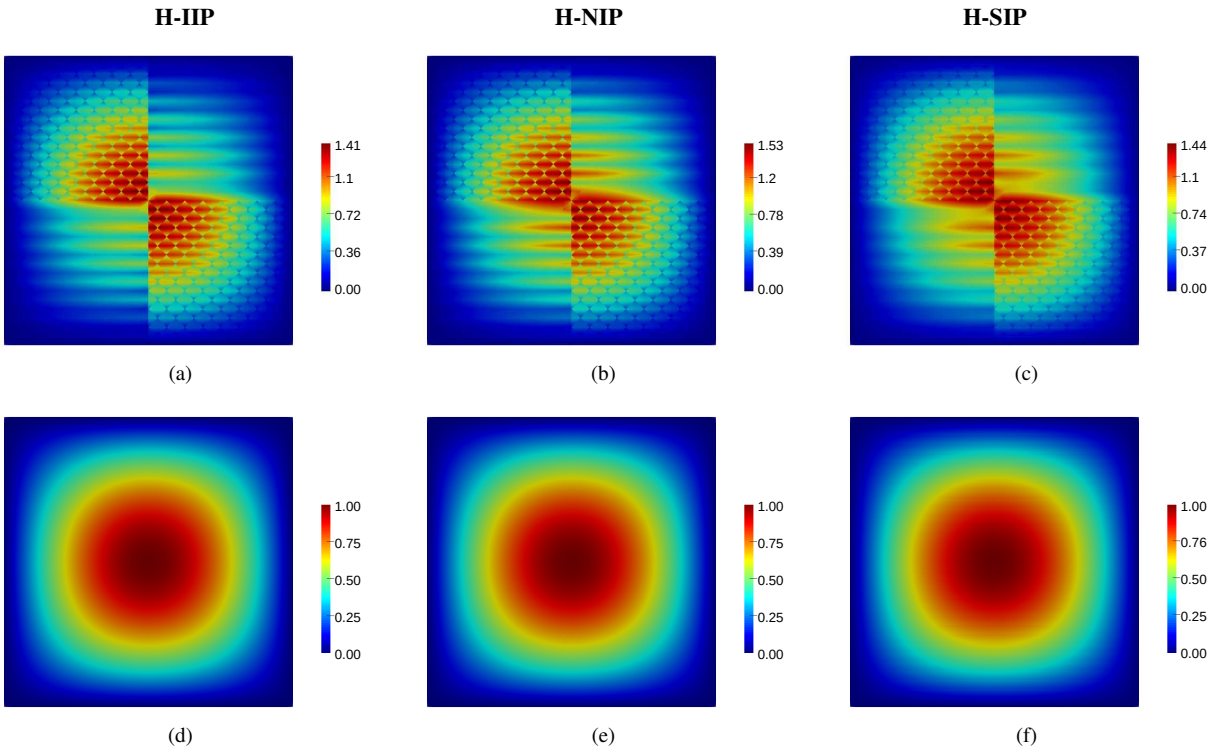


Figure 6: Test B: representation of the discrete solution u_h obtained by the H-IIP, H-NIP and H-SIP schemes, respectively, on the structured triangular mesh ($h = 1/32$). In the top images, the parameter κ_F in (13) is chosen as $\kappa_F := 1$, and in the bottom images, $\kappa_F := \boldsymbol{n}_F\boldsymbol{\kappa}_E\boldsymbol{n}_F$.

	H-IIP		H-NIP		H-SIP	
	Case 1	Case 2	Case 1	Case 2	Case 1	Case 2
$\min(u_h)$	$1.54e-03$	$2.14e-03$	$2.79e-03$	$2.09e-03$	$2.68e-03$	$2.12e-03$
$\max(u_h)$	$1.25e+00$	$9.97e-01$	$1.33e+00$	$9.97e-01$	$1.30e+00$	$9.97e-01$
$\|u-u_h\|_{0,\mathcal{T}_h}$	$1.31e-01$	$4.33e-04$	$1.39e-01$	$5.43e-04$	$1.21e-01$	$1.96e-03$

Table 2: Test B: comparison of H-IP methods using a piecewise linear approximation ($u_h \in \mathbb{P}_1(\mathcal{T}_h)$) and two distinct definitions of the coefficient κ_F for highly anisotropic and heterogeneous media ($\lambda = 10^{-3}$). In Case 1, $\kappa_F := 1$, and in Case 2, $\kappa_F := \boldsymbol{n}_F\boldsymbol{\kappa}_E\boldsymbol{n}_F$.

5.3. Test C: Influence of the parameter γ_0

To conclude the sequence of numerical tests, we analyze the influence of the parameter γ_0 on the convergence of the H-SIP method for $\boldsymbol{\kappa}$-orthogonal grids only. For simplicity, we consider the same test case as Test B, (5.2), and we set two values of the parameter λ: (i) $\lambda = 1$ for a homogeneous and isotropic media and (ii) $\lambda = 0.1$ for a

heterogeneous and anisotropic media. We plot the computed L^2-error of the H-SIP method for a wide range of values of the parameter γ_0—i.e., $1 \leq \gamma_0 \leq 6$—using a uniform square mesh ($h = 1/32$). The analysis is done for polynomial degrees $1 \leq k \leq 4$, but the results are presented for $k = 1, 2$ only. Analyzing Figure 7, we observe that there exists an optimal value of the parameter $\gamma_0 := \gamma_{\text{opt}}$ that minimizes the L^2-error of the scheme. In the context of $\boldsymbol{\kappa}$-orthogonal grids, this optimal value ($\gamma_{\text{opt}} = 2$) is insensitive to the mesh form, the mesh size h, the polynomial degree k, and the heterogeneity and/or anisotropy of the media λ. A history of the convergence of the H-SIP method using $\gamma_{\text{opt}} = 2$ is then given in Table 7, and we note the surprising superconvergence of u_h $(k + 2)$ in the discrete L^2-norm obtained without any postprocessing. We emphasize that the *superconvergence* property is not achieved for any triangular mesh or any value of the parameter $\epsilon \neq 1$, even using the optimal parameter γ_{opt} in (13).

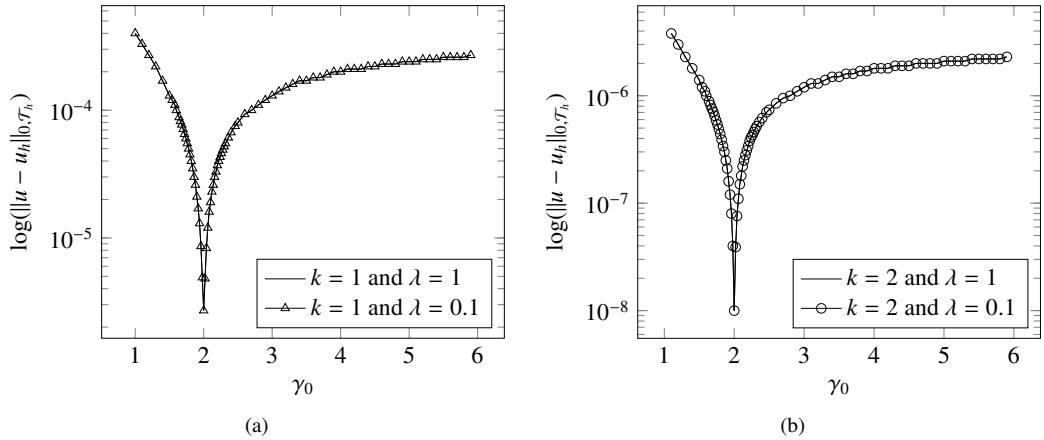


Figure 7: Test C: the L^2-error of the H-SIP method vs. γ_0 for a uniform square mesh using piecewise linear (a) and quadratic (b) approximations.

	H-SIP ($k = 1$)				H-SIP ($k = 2$)			
	$\lambda = 1$		$\lambda = 0.1$		$\lambda = 1$		$\lambda = 0.1$	
h^{-1}	$\|u - u_h\|_{0,\mathcal{T}_h}$	ECR	$\|u - u_h\|_{0,\mathcal{T}_h}$	ECR	$\|u - u_h\|_{0,\mathcal{T}_h}$	ECR	$\|u - u_h\|_{0,\mathcal{T}_h}$	ECR
8	$1.7e-04$	–	$1.7e-04$	–	$2.6e-06$	–	$2.6e-06$	–
16	$2.1e-05$	3.00	$2.1e-05$	3.00	$1.6e-07$	3.99	$1.6e-07$	3.99
32	$2.7e-06$	3.00	$2.7e-06$	3.00	$1.0e-08$	4.00	$1.0e-08$	4.00
64	$3.4e-07$	3.00	$3.4e-07$	3.00	$6.4e-10$	4.00	$6.4e-10$	4.00

Table 3: Test C: history of the convergence $\|u - u_h\|_{0,\mathcal{T}_h}$ of the H-SIP method using the optimal parameter γ_{opt} on uniform square meshes

6. Conclusion

We derive improved *a priori* error estimates of families of hybridizable interior penalty discontinuous Galerkin methods using a variable penalty to solve highly anisotropic diffusion problems. The convergence analysis highlights the h^{δ}-dependency of the coercivity condition and the boundedness requirement that strongly impacts the derived error estimates in terms of both energy- and L^2-norms. The optimal convergence of the energy-norm is proven for any penalty parameter $\delta \geq 0$ and $\epsilon \in \{0, \pm 1\}$. The situation is somewhat different in L^2, and distinctive features can be found between the three schemes. Indeed, the symmetric method theoretically converges optimally if $\delta \geq 0$, and non-symmetric variants converge only if $\delta \geq 2$ independently of the polynomial parity. All of these estimates are corroborated by numerical evidence. Notably, the superconvergence of the H-SIP scheme is achieved for $\boldsymbol{\kappa}$-orthogonal grids without any postprocessing but only if an appropriate γ_0 is selected.

Acknowledgments

By convention, the names of the authors are listed in alphabetical order. The corresponding author is grateful to Sander Rhebergen for his invitation to the Department of Applied Mathematics at the University of Waterloo (UW) in April 2019. Thank you very much for introducing me to HDG methods and to the subtle mechanisms of the NGSolve library. Special thanks to UW for the kind hospitality. He also would like to thank Béatrice Rivière at Rice University for her insightful suggestions and remarks concerning *a priori* error estimates. Our fruitful discussions of hybridizable interior penalty methods using superpenalties were the source of inspiration of the present work.

[1] B. Cockburn, J. Gopalakrishnan, R. Lazarov, Unified hybridization of discontinuous Galerkin, mixed, and continuous Galerkin methods for second order elliptic problems, SIAM Journal on Numerical Analysis 47 (2) (2009) 1319–1365.

[2] H. Egger, J. Schöberl, A mixed-hybrid-discontinuous Galerkin finite element method for convection-diffusion problems, IMA J. Numer. Anal 30 (2009) 1–2.

[3] B. Cockburn, B. Dong, J. Guzmán, M. Restelli, R. Sacco, A hybridizable discontinuous Galerkin method for steady-state convection-diffusion-reaction problems, SIAM Journal on Scientific Computing 31 (5) (2009) 3827–3846.

[4] N. C. Nguyen, J. Peraire, B. Cockburn, An implicit high-order hybridizable discontinuous Galerkin method for linear convection-diffusion equations, Journal of Computational Physics 228 (9) (2009) 3232–3254.

[5] I. Oikawa, HDG METHODS FOR SECOND-ORDER ELLIPTIC PROBLEMS (Numerical Analysis: New Developments for Elucidating Interdisciplinary Problems II), RIMS Kokyuroku 2037 (2017) 61–74.

[6] K. L. Kirk, S. Rhebergen, Analysis of a Pressure-Robust Hybridized Discontinuous Galerkin Method for the Stationary Navier–Stokes Equations, Journal of Scientific Computing 81 (2) (2019) 881–897.

[7] M. S. Fabien, M. Knepley, B. Riviere, A high order hybridizable discontinuous Galerkin method for incompressible miscible displacement in heterogeneous media, Results in Applied Mathematics (2020) 100089doi:\let\@tempa\bibinfo@X@doihttps://doi.org/10.1016/j.rinam.2019.100089.

[8] D. N. Arnold, F. Brezzi, B. Cockburn, L. D. Marini, Unified analysis of discontinuous Galerkin methods for elliptic problems, SIAM journal on numerical analysis 39 (5) (2002) 1749–1779.

[9] R. M. Kirby, S. J. Sherwin, B. Cockburn, To CG or to HDG: a comparative study, Journal of Scientific Computing 51 (1) (2012) 183–212.

[10] C. Lehrenfeld, Hybrid discontinuous Galerkin methods for solving incompressible flow problems, Rheinisch-Westfalischen Technischen Hochschule Aachen (2010) 111.

[11] M. S. Fabien, M. G. Knepley, B. M. Riviere, Families of Interior Penalty Hybridizable discontinuous Galerkin methods for second order elliptic problems, Journal of Numerical Mathematics (0), doi:\let\@tempa\bibinfo@X@doihttps://doi.org/10.1515/jnma-2019-0027.

[12] L. Dijoux, V. Fontaine, T. A. Mara, A projective hybridizable discontinuous Galerkin mixed method for second-order diffusion problems, Applied Mathematical Modelling 75 (2019) 663–677, ISSN 0307-904X, doi:\let\@tempa\bibinfo@X@doihttps://doi.org/10.1016/j.apm.2019.05.054.

[13] G. N. Wells, Analysis of an interface stabilized finite element method: the advection-diffusion-reaction equation, SIAM Journal on Numerical Analysis 49 (1) (2011) 87–109.

[14] D. N. Arnold, An interior penalty finite element method with discontinuous elements, SIAM journal on numerical analysis 19 (4) (1982) 742–760.

[15] B. Riviere, Discontinuous Galerkin methods for solving elliptic and parabolic equations: theory and implementation, SIAM, 2008.

[16] D. A. Di Pietro, A. Ern, Mathematical aspects of discontinuous Galerkin methods, vol. 69, Springer Science & Business Media, 2011.

[17] B. Rivière, M. Wheeler, V. Girault, Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems, Computational Geosciences 3 (3) (1998) 337–360, doi:\let\@tempa\bibinfo@X@doihttps://doi.org/10.1023/A:1011591328604.

[18] J. Guzmán, B. Rivière, Sub-optimal Convergence of Non-symmetric Discontinuous Galerkin Methods for Odd Polynomial Approximations, Journal of Scientific Computing 40 (1) (2009) 273–280.

[19] P. G. Ciarlet, Basic error estimates for elliptic problems, Handbook of Numerical Analysis 2 (1991) 17–351.

[20] A. Ern, A. F. Stephansen, P. Zunino, A discontinuous Galerkin method with weighted averages for advection–diffusion equations with locally small and anisotropic diffusivity, IMA Journal of Numerical Analysis 29 (2) (2009) 235–256.

[21] J. Schöberl, C++ 11 implementation of finite elements in NGSolve, Institute for Analysis and Scientific Computing, Vienna University of Technology .