A note on the design of \(hp\)-version interior penalty discontinuous Galerkin finite element methods for degenerate problems

Emmanuil H. Georgoulis \(^a\) Andris Lasis \(^b\)

\(^a\)Department of Mathematics, University of Leicester, University Road, Leicester LE1 7RH, Emmanuil.Georgoulis@mcs.le.ac.uk

\(^b\)Oxford University Computing Laboratory, Wolfson Building, Parks Road, Oxford OX1 3QD, Andris.Lasis@comlab.ox.ac.uk

We consider a variant of the \(hp\)-version interior penalty discontinuous Galerkin finite element method (IP-DGFE M) for second order problems of degenerate type. We do not assume uniform ellipticity of the diffusion tensor. Moreover, diffusion tensors of arbitrary form are covered in the theory presented. A new, refined recipe for the choice of the discontinuity-penalisation parameter (that is present in the formulation of the IP-DGFE M) is given. Making use of the recently introduced augmented Sobolev space framework, we prove an \(hp\)-optimal error bound in the energy norm and an \(h\)-optimal and slightly \(p\)-suboptimal (by only half an order of \(p\)) bound in the \(L^2\) norm, provided that the solution belongs to an augmented Sobolev space.

Subject classifications: AMS(MOS): 65N12, 65N15, 65N30

Key words and phrases: \(hp\)-finite element methods, discontinuous Galerkin methods, degenerate problems

Oxford University Computing Laboratory
Numerical Analysis Group
Wolfson Building
Parks Road
Oxford, England OX1 3QD

June, 2005
## Contents

1 Introduction .................................................. 3

2 Model problem and preliminaries .................. 4

3 Discontinuous Galerkin finite element methods ... 5

4 Error analysis ................................................. 7

5 Concluding remarks ........................................ 11
1 Introduction

The numerical solution of boundary value problems containing second order equations of degenerate type is of interest, as in many models diffusion can be small, degenerate, or even identically equal to zero in subregions of the domain under consideration. Owing to stability properties and flexibility in the choice of discretisation in different parts of the differential operator, discontinuous Galerkin finite element methods have been proposed for the numerical solution of such equations [9, 6] (see the important volume [4] and the references therein for the history and the development of discontinuous Galerkin methods). However, the error analysis presented in [9, 6] required that diffusion coefficients are of a particular form in the regions of degeneracy, excluding many equations of degenerate type.

Here, we present and analyse a variant of the the interior-penalty discontinuous Galerkin method for general positive semidefinite diffusion tensors. The method proposed in this work emerges from an inconsistent broken weak formulation. Hence, Strang’s Second Lemma is employed for the error analysis. We also show that this modification delivers an $hp$-optimal error bound in the energy norm and and $h$-optimal and slightly $p$-suboptimal (by only half an order of $p$) bound in the $L^2$ norm when the solution admits sufficient regularity in the context of augmented Sobolev spaces, presented in [7]. We note that, to the best of our knowledge, the bounds in the $L^2$ norm derived here are the sharpest known in the literature. A significant advantage of this approach is the weak regularity assumption imposed on the analytical solution of the problem in the definition of the method. Indeed, the global regularity of solutions to problems with degenerate diffusion is, in general, low due to the lack of ellipticity in subregions of the domain of definition.

The present analysis is inspired by the analysis by Perugia and Schötzau [11] in the case of the local DG method for linear elliptic problems, where an inconsistent weak formulation of the problem is derived (through the use of lifting operators) and Strang’s Second Lemma is employed for the error analysis.

For simplicity, we do not include convection and reaction terms in the differential operator; the corresponding $hp$-DGFEM for the solution of equations with nonnegative characteristic form can be defined in a straightforward fashion, analogously to [9].

The paper is structured as follows. In Section 2 we state the model problem and, in Section 3, we present the broken weak formulation that leads to the IP-DGFEM. The error analysis of the $hp$-version IP-DGFEM is discussed in Section 4: we prove an $hp$-optimal $a$ priori error bound in the energy norm and an $h$-optimal and slightly $p$-suboptimal (by only half an order of $p$) bound in the $L^2$ norm, using the projection error bounds introduced in [7]. Section 5 contains some final comments on our results in this work.
2 Model problem and preliminaries

For $\omega \subset \mathbb{R}^2$, let $H^s(\omega)$, $s \geq 0$, signify the usual Sobolev space equipped with the norm $\| \cdot \|_{s,\omega}$ and seminorm $|\cdot|_{s,\omega}$; we also write $L^2(\omega) \equiv H^0(\omega)$.

Let $\Omega$ be a bounded open convex polygonal domain in $\mathbb{R}^2$, and $a = \{a_{ij}\}_{i,j=1}^2$ be a symmetric positive semidefinite tensor whose entries $a_{ij}$ are bounded, piecewise continuously differentiable real-valued functions defined on $\Omega$, with

$$\zeta^\top a(x)\zeta > 0 \quad \forall \zeta \in \mathbb{R}^2, \quad \text{a.e. } x \in \bar{\Omega}. \quad (2.1)$$

Note that we do not assume uniform ellipticity of the form $\zeta^\top a(x)\zeta \geq \alpha$, for some positive real number $\alpha$. We consider the linear problem in the form

$$-\nabla \cdot (a \nabla u) = f \quad \text{in } \Omega, \quad u = g_D \quad \text{on } \partial \Omega, \quad (2.2)$$

with $f \in L^2(\Omega)$ and $g_D \in H^{1/2}(\partial \Omega)$. We assume that $a$ is such that the boundary value problem has a unique solution. Equations of such type include the Grušin type equations where the diffusion tensor is of the form $a = \text{diag}(1, \lambda^2(x))$, $(x, y) \in \Omega$, for some continuously differentiable function of one variable $\lambda$ which is non-negative for a.e. $x \in \mathbb{R}$ such that $(x, y) \in \mathbb{R}^2$ (the analysis presented in [6] includes this case), and the equation of M. V. Keldyš where the diffusion tensor is of the form $a = \text{diag}(1, y^r)$ for some $r > 0$ (see, e.g., [10]), which is not covered by the analysis in [6].

Let $T_h$ be a subdivision of $\Omega$ into disjoint open elements $\kappa$ such that each side of $\kappa$ has at most one hanging node. We let $h_\kappa := \text{diam}(\kappa)$ and $h := \max_{\kappa \in T_h} h_\kappa$. We assume that the family of subdivisions $\{T_h\}$ is shape-regular (see, e.g., page 124 in [3]), constructed via mappings $Q_\kappa \circ F_\kappa$, where $F_\kappa : \tilde{\kappa} := (-1, 1)^2 \to \kappa$ is an affine mapping of the form $F_\kappa(x) := A_\kappa x + b_\kappa$, with $A_\kappa := (h_\kappa^2/2) \text{id}_2$, where $h_\kappa$ the length of the edges of $\tilde{\kappa}$, and $\text{id}_2$ the $2 \times 2$ identity matrix, $b_\kappa$ a two-component real-valued vector, and $Q_\kappa : \tilde{\kappa} \to \kappa$ is a $C^1$-diffeomorphism, such that the Jacobian $J_{Q_\kappa}$ of $Q_\kappa$ satisfies

$$C_1^{-1} \leq \det J_{Q_\kappa} \leq C_1, \quad \|(J_{Q_\kappa})_{ij}\|_{L^\infty(\kappa)} \leq C_2, \quad i, j = 1, 2, \quad \text{for all } \kappa \in T_h,$$

uniformly throughout the mesh for some positive constants $C_1, C_2$. The above maps are assumed to be constructed so as to ensure that the union of the closures of the elements $\kappa \in T_h$ forms a covering of the closure of $\Omega$, i.e., $\bar{\Omega} = \cup_{\kappa \in T_h} \bar{\kappa}$.

Heuristically, we can say that the affine mapping $F_\kappa$ defines the size of the element $\kappa$ and the diffeomorphism $Q_\kappa$ defines the “shape”. We refer to [7] for a more detailed discussion of these ideas.

We assign to the subdivision $T_h$ the broken Sobolev space of composite order $s := \{s_\kappa : \kappa \in T_h\}$,

$$H^s(\Omega, T_h) := \{ u \in L^2(\Omega) : u|_\kappa \in H^{s_\kappa}(\kappa) \text{ for all } \kappa \in T_h \},$$

equipped with the standard broken Sobolev norm. When $s_\kappa = s$ for all $\kappa \in T_h$, we write $H^s(\Omega, T_h)$. 

We shall also make use of the notion of augmented Sobolev spaces, introduced in [7]. In that paper, augmented Sobolev spaces were shown to be a suitable function space framework for the derivation of the first \(h^p\)-optimal energy norms error bounds for the interior penalty discontinuous Galerkin finite element method.

**Definition 2.1** Let \( s \) be a nonnegative integer and \( \kappa \) an open (curvilinear) quadrilateral domain with diameter \( h_\kappa := \text{diam}(\kappa) \), constructed as the image of an axiparallel rectangle \( \tilde{\kappa} \) through a \( C^{s+1} \)-diffeomorphism \( Q_\kappa \), as indicated above. We define the augmented Sobolev space of order \( s \) on \( \kappa \) by

\[
\tilde{H}^s(\kappa) := \{ (u \circ Q_\kappa) \in H^s(\tilde{\kappa}) : (\alpha, \beta) \in I, \ \partial_\alpha^\alpha \partial_\beta^\beta (u \circ Q_\kappa) \in L^2(\tilde{\kappa}) \},
\]

where \( I := \{ (\alpha, \beta) \in \mathbb{N}_0^2 : \alpha + \beta = s + 1, \text{ excluding } (s + 1, 0) \text{ and } (0, s + 1) \} \), with associated norm \( \| \cdot \|_{s, \kappa} :\)

\[
\| u \|_{s, \kappa} := \left( \| u \circ Q_\kappa \|_{s, \kappa}^2 + \sum_{(\alpha, \beta) \in I} h_\kappa^{2s} \| \partial_\alpha^\alpha \partial_\beta^\beta (u \circ Q_\kappa) \|_{0, \kappa}^2 \right)^{1/2}.
\]

Finally, for a nonnegative integer \( p \), we denote by \( Q_p(\tilde{\kappa}) \), the set of all tensor-product polynomials on \( \kappa \) of degree \( p \) in each coordinate direction. To each \( \kappa \in \mathcal{T}_h \) we assign nonnegative integers \( p_\kappa \) (the local polynomial degree), collect the \( p_\kappa \) and \( F_\kappa \) into vectors \( \mathbf{p} = \{ p_\kappa : \kappa \in \mathcal{T}_h \} \) and \( \mathbf{F} = \{ Q_\kappa \circ F_\kappa : \kappa \in \mathcal{T}_h \} \), respectively, and consider the finite element space

\[
S^p \equiv S^p(\Omega, \mathcal{T}_h, \mathbf{F}) := \{ v \in L^2(\Omega) : v|_\kappa \circ Q_\kappa \circ F_\kappa \in Q_{p_\kappa}(\tilde{\kappa}), \kappa \in \mathcal{T}_h \}.
\]

We shall assume throughout that the meshsize vector \( \mathbf{h} := \{ h_\kappa : \kappa \in \mathcal{T}_h \} \) and polynomial degree vector \( \mathbf{p} \), with \( p_\kappa \geq 1 \) for each \( \kappa \in \mathcal{T}_h \), have bounded local variation, i.e., there exist constants \( \eta, \rho \geq 1 \), independent of \( \mathbf{h} \) and \( \mathbf{p} \), such that, for any pair of elements \( \kappa \) and \( \kappa' \) in \( \mathcal{T}_h \) which share a side,

\[
\eta^{-1} \leq h_\kappa / h_{\kappa'} \leq \eta \quad \text{and} \quad \rho^{-1} \leq p_\kappa / p_{\kappa'} \leq \rho.
\]

## 3 Discontinuous Galerkin finite element methods

Let \( \mathcal{E} \) denote the set of all open sides of the subdivision \( \mathcal{T}_h \), containing the smallest common interfaces \( e \) of neighbouring elements, and define \( \Gamma := \{ x \in \Omega : x \in e, \ e \in \mathcal{E} \} \). We denote by \( \mathcal{E}_{\text{int}} \) the set of all sides in \( \mathcal{E} \) that are contained in the interior of \( \Omega \), and let \( \Gamma_{\text{int}} := \{ x \in \Omega : x \in e \text{ for some } e \in \mathcal{E}_{\text{int}} \} \). Further, we denote by \( \mathcal{E}_{\partial} \) the set of all boundary sides. Given that \( e \in \mathcal{E}_{\text{int}} \), there exist positive integers \( i, j \) such that \( i > j \) and \( \kappa_i \) and \( \kappa_j \) share the side \( e \); we define the jump of \( v \in H^s(\Omega, \mathcal{T}_h) \), \( s_\kappa > 1/2 \), \( \kappa \in \mathcal{T}_h \), across \( e \) and the mean value of \( v \) on \( e \) by

\[
[u]_e := v|_{\partial\kappa_i \cap e} - v|_{\partial\kappa_j \cap e} \quad \text{and} \quad \langle v \rangle_0 := \frac{1}{2} (v|_{\partial\kappa_i \cap e} + v|_{\partial\kappa_j \cap e}).
\]
respectively, with $\partial \kappa$ denoting the union of all open sides of the element $\kappa$. With each side $e$ we associate the unit normal vector $\nu$ pointing from the element $\kappa_i$ to $\kappa_j$ when $i > j$; when the side belongs to $\mathcal{E}_0$, we choose $\nu$ to be the unit outward normal vector.

In what follows, we shall write $\bar{a} := \sqrt{a}_{\|2}$, with $\sqrt{a}$ denoting the positive semidefinite square root of the symmetric matrix $a$ and $\| \cdot \|_2$ denoting the matrix-2-norm; we also write $\bar{a}_{\kappa} := \bar{a}|_{\kappa}$ and denote by $\alpha_{\kappa}$ the arithmetic mean of $\bar{a}_{\kappa'}$ over the elements $\kappa'$ (including $\kappa$) that share a side with $\kappa$.

For $\theta \in [-1, 1]$, we introduce the bilinear form

$$B(w, v) := \sum_{\kappa \in T_h} \int_\kappa a \nabla w \cdot \nabla v \, dx + \int_{\partial\Omega} \sigma_{\kappa} w v \, ds + \int_{\Gamma_{\text{int}}} \sigma[w][v] \, ds$$

$$+ \int_{\partial\Omega} \{\theta (a \nabla v \cdot \nu) - (a \nabla w \cdot \nu)\} w \, ds$$

$$+ \int_{\Gamma_{\text{int}}} \{\theta (a \nabla v \cdot \nu)[w] - (a \nabla w \cdot \nu)[v]\} \, ds$$

and the linear functional

$$l(v) := \int_\Omega f v \, dx + \int_{\partial\Omega} \theta (a \nabla v \cdot \nu) g_D \, ds + \int_{\partial\Omega} \sigma g_D v \, ds,$$

where $\sigma : \Gamma \to \mathbb{R}$ is the discontinuity-penalisation function and is defined by

$$|\sigma|_e = C_\sigma (|\sqrt{a} \nu|_{2}^{2} p^2 / h_e) \quad \text{for} \quad e \in \mathcal{E},$$

where $C_\sigma$ is a positive constant whose size will be determined later, and $h_e$ is the diameter of a side $e \in \mathcal{E}$. When $e \in \mathcal{E}_0$, the contribution from outside $\Omega$ in the definition of $\sigma_e$ is set to 0. The subscript $e$ in this definition will be suppressed when no confusion is likely to occur. The choice of $\theta = -1$ leads to a symmetric bilinear form $B(\cdot, \cdot)$; we call the associated method the Symmetric Interior Penalty, or SIP, method. For $\theta = 1$ the bilinear form $B(\cdot, \cdot)$ is nonsymmetric, but it is coercive for any $\sigma > 0$; we call the corresponding method the Nonsymmetric Interior Penalty, or NIP, method. The choice of $\theta = 0$ leads to a nonsymmetric bilinear form $B(\cdot, \cdot)$; we call the associated method the Incomplete Interior Penalty, or IIP, method.

Then, the broken weak formulation of the problem (2.2) reads as follows:

$$\text{find } u \in \mathcal{X} \text{ such that } B(u, v) = l(v) \quad \forall v \in H^2(\Omega, T_h),$$

where by $\mathcal{X}$ we denote the function space

$$\mathcal{X} = \{ w \in H^2(\Omega, T_h) : w, (a \nabla w) \cdot \nu \text{ are continuous across each } e \in \mathcal{E}_{\text{int}} \}.$$

The well-posedness of this problem has been extensively discussed (see, e.g., [9]); the coercivity of the bilinear form $B$ is guaranteed by choosing the constant $C_\sigma$ large enough [1].
In addition to (3.1) and (3.2), we shall define the bilinear form $B_{in}(\cdot, \cdot)$ and the linear functional $\ell_{in}(\cdot)$ by

$$B_{in}(w, v) := \sum_{\kappa \in T_h} a \nabla w \cdot \nabla v \ dx + \int_{\partial\Omega} \sigma w v \ ds + \int_{\Gamma_{int}} \sigma[w][v] \ ds$$

$$+ \int_{\partial\Omega} \left\{ \theta(\sqrt{a}P_0(\sqrt{a}d) \cdot \nu)w - (\sqrt{a}P_0(\sqrt{a}d) \cdot \nu)v \right\} \ ds$$

$$+ \int_{\Gamma_{int}} \left\{ \theta(\sqrt{a}P_0(\sqrt{a}d) \cdot \nu)[w] - (\sqrt{a}P_0(\sqrt{a}d) \cdot \nu)[v] \right\} \ ds$$

and

$$\ell_{in}(v) := \int_{\Omega} f v \ dx + \int_{\partial\Omega} \theta(\sqrt{a}P_0(\sqrt{a}d) \cdot \nu) g_D \ ds + \int_{\partial\Omega} \sigma g_D v \ ds,$$

with $P_0$ denoting the $L^2$ (orthogonal) projection operator onto $[S^p]$. The modified broken weak formulation of the problem (2.2) reads as follows:

$$\text{find } u \in H^1(\Omega) \text{ such that } B_{in}(u, v) = \ell_{in}(v) \quad \forall \ v \in H^1(\Omega, T_h).$$

We note that this weak formulation is inconsistent with (2.2). On the other hand, we observe that weaker regularity assumptions on the analytical solution to (2.2) are required here: we shall merely assume that $u \in H^1(\Omega)$ (as opposed to $u \in X$ in the original error analysis of DGFEM [6]), since no additional regularity of $u$ is required to ensure the existence of traces of $\sqrt{a}d u$ in the definition of the method. The $hp$-DGFEM approximation of the problem (2.2) reads as follows:

$$\text{find } u_{DG} \in S^p \text{ such that } B_{in}(u_{DG}, v) = \ell_{in}(v) \quad \forall \ v \in S^p.$$ 

Note that this formulation is equivalent to the original broken weak formulation (3.4) for discrete trial and test functions, when the diffusion tensor is elementwise constant.

**Remark 3.1** In practice, the projections can be evaluated via an elemental mass matrix computation. Hence, the amount of additional calculations required by this approach during the assembly of the stiffness matrix is marginal.

### 4 Error analysis

We equip $H^1(\Omega, T_h)$ with the norm $\| \cdot \|_{1,h}$ defined by

$$\| w \|_{1,h} := \left( \sum_{\kappa \in T_h} \| \sqrt{a} \nabla w \|_{0,\kappa}^2 + \int_{\partial\Omega} \sigma w^2 \ ds + \int_{\Gamma_{int}} \sigma[w]^2 \ ds \right)^{1/2},$$

with $\sigma$ as in (3.3). The bilinear form $B_{in}(\cdot, \cdot)$ has the following continuity and coercivity properties.
Lemma 4.1 There exist positive constant $C_1$ and $C_2$, independent of $u$ and the discretisation parameters, such that the bilinear form $B_{in}(\cdot, \cdot)$ is continuous and coercive in the sense that

$$|B_{in}(v, w)| \leq C_1 \|v\|_{1,h} \|w\|_{1,h} \quad \text{for all } v, w \in H^1(\Omega, T_h),$$

$$B_{in}(w, w) \geq C_2 \|w\|_{1,h}^2 \quad \text{for all } w \in H^1(\Omega, T_h).$$

Proof. Following the proof of Lemma 4.3 in [9], the first three terms on the right-hand side of (3.5) can be bounded as follows:

$$\sum_{\kappa \in T_h} \int_{\kappa} a \nabla w \cdot \nabla v \, dx + \int_{\partial \Omega} \sigma w v \, ds + \int_{\Gamma_{int}} \sigma[w]v \, ds \leq \|w\|_{1,h} \|v\|_{1,h}.$$

Also, using the symmetry of $\sqrt{a}$, we have

$$\left| \int_{\Gamma_{int}} \theta \langle \sqrt{a} \Pi(\sqrt{a} \nabla v) \cdot \nu \rangle[w] \, ds \right| \leq |\theta| \sum_{e \in E_{int}} \int_e \left| \Pi(\sqrt{a} \nabla v) \right|_{\kappa} \cdot \left( \sqrt{a} \nu \right) [w] \, ds.$$

Using the discrete and continuous versions of Cauchy-Schwarz inequality, and multiplying and dividing by $\sqrt{h_{\kappa}}/p_{\kappa}$ and $\sqrt{h_{\kappa'}}/p_{\kappa'}$ the first and the second terms on the right-hand side of (4.2), respectively, we obtain

$$\left| \int_{\Gamma_{int}} \theta \langle \sqrt{a} \Pi(\sqrt{a} \nabla v) \cdot \nu \rangle[w] \, ds \right| \leq |\theta| \left( \sum_{e \in E_{int}} \frac{h_{\kappa}}{p_{\kappa}} \left\| \Pi(\sqrt{a} \nabla v) \right\|_{0,e}^2 + \frac{h_{\kappa'}}{p_{\kappa'}} \left\| \Pi(\sqrt{a} \nabla v) \right\|_{0,e}^2 \right)^{1/2} \times \left( \sum_{e \in E_{int}} \int_e \left( \frac{p_{\kappa}}{\bar{h}} \right)^2 \left( \sqrt{a} \nu \right)_2/w \, ds \right)^{1/2} \leq C|\theta| \left( \sum_{\kappa \in T_h} \left\| \Pi(\sqrt{a} \nabla v) \right\|_{0,\kappa}^2 \right)^{1/2} \|w\|_{1,h} \leq C_1 \|v\|_{1,h} \|w\|_{1,h}.$$

where in the last two steps we have used the standard inverse inequality and the stability of the $L^2$ projection operator, respectively. Applying analogous arguments to the remaining terms on the right-hand side of (3.5), we obtain the result on continuity. The coercivity follows by a standard argument, with the constant $C_\sigma$ in (3.3) chosen large enough.

The next result (Strang’s Second Lemma) gives us the abstract error bound that can be used for the analysis of the inconsistent formulation (cf. Theorem 4.2.2 [3]).
Let \( a \) and \( \sigma \) be as above and assume that the analytical solution \( u \in H^1(\Omega) \). Then, we have
\[
\|u - u_{DG}\|_{1,h} \leq \left(1 + \frac{C_1}{C_2}\right) \inf_{v \in S^p} \|u - v\|_{1,h} + \frac{1}{C_2} \sup_{w \in S^p} \frac{|R(u, w)|}{\|w\|_{1,h}},
\]
(4.3)
where \( R(u, w) = B_{u,v}(u, w) - \ell_{in}(w) \) is the residual.

We continue by estimating the residual.

**Lemma 4.3** Let \( \{T_h\} \) be a family of shape-regular subdivisions of \( \Omega \), and let \( u \) be the analytical solution of (2.2). Assume further that \( (\sqrt{a} \nabla u)|_k \in [H^{k_\kappa}(\kappa)]^2 \), for all \( k \in T_h \). Then, for all \( w \in S^p \),
\[
|R(u, w)| \leq C \left( \sum_{k \in T_h} \frac{h_{2k_\kappa}}{p_{2k_\kappa+1}} \|\sqrt{a} \nabla u\|_{k_\kappa,\kappa}^2 \right)^{1/2} \|w\|_{1,h},
\]
(4.4)
with \( 0 \leq s_\kappa \leq \min \{p_\kappa, k_\kappa\} \), \( p_\kappa \geq 1 \); the constant \( C \) is independent of \( u \), the data \( a, f \) and \( g_0 \), and the discretisation parameters.

**Proof.** A straightforward calculation yields
\[
R(u, w) = \int_{\Gamma_{int}} \langle (a \nabla u - \sqrt{a} \Pi(\sqrt{a} \nabla u)) \cdot \nu \rangle [w] \, ds + \int_{\partial_\Omega} (a \nabla u - \sqrt{a} \Pi(\sqrt{a} \nabla u)) \cdot \nu w \, ds.
\]
Then, by analogous arguments to the ones in the proof of Lemma 4.1, we deduce that
\[
|R(u, w)| \leq \left( \sum_{k \in T_h} \frac{h_{2k_\kappa}}{p_{2k_\kappa+1}} \|\sqrt{a} \nabla u - \Pi(\sqrt{a} \nabla u)\|_{0,0}^2 \right)^{1/2} \|w\|_{1,h}
\]
\[
\leq C \left( \sum_{k \in T_h} \frac{h_{2k_\kappa}}{p_{2k_\kappa+1}} \|\sqrt{a} \nabla u\|_{k_\kappa,\kappa}^2 \right)^{1/2} \|w\|_{1,h},
\]
where in the last step we made use of the \( L^2 \) projection approximation bound from Lemma 3.6 in [9].

We are now ready to present the approximation error bound in the energy norm.

**Theorem 4.4** Let \( \Omega \subset \mathbb{R}^2 \) be a bounded (curvilinear) polygonal domain. Suppose that \( \{T_h\} \) is a family of shape-regular subdivisions of \( \Omega \), and \( p \) is a polynomial degree vector of bounded local variation. Assume further that \( u|_k \in H^{k_\kappa+1}(\kappa) \) and that \( (\sqrt{a} \nabla u)|_k \in [H^{k_\kappa}(\kappa)]^2 \) for some Sobolev index \( k_\kappa \geq 1 \) and each \( k \in T_h \). Let the discontinuity-penalisation function \( \sigma : \Gamma \to \mathbb{R} \) be defined as in (3.3). Then, the following energy error bound holds:
\[
\|u - u_{DG}\|_{1,h} \leq C \left( \sum_{k \in T_h} \frac{h_{2k_\kappa}}{p_{2k_\kappa}^2} \left( \alpha_\kappa^2 \|u\|_{k_\kappa+1,\kappa}^2 + p_\kappa^{-1} \|\sqrt{a} \nabla u\|_{k_\kappa,\kappa}^2 \right) \right)^{1/2}
\]
(4.5)
where \( s_\kappa = \min \{p_\kappa, k_\kappa\} \), and the constant \( C \) is independent of \( u, p, \) and \( h_\kappa \), but depends on \( k = \max_{k \in T_h} k_\kappa \), the parameter \( \rho \) in (2.5), the constant \( C_\sigma \), and on \( \Omega \).
Proof. Using (4.3), with \( v = \Pi_{H^1} u \) such that \((\Pi_{H^1} u)|_\kappa = \pi^h_p(u|_\kappa)\) where \(\pi^h_p\) denotes the \(H^1\) projection operator (see Definition 6.10 in [7] for details), along with the approximation error bounds from Theorems 6.11 and 6.12 in [7], we deduce that

\[
\inf_{v \in S^h} \| u - v \|_{1,h} \leq \| u - \Pi_{H^1} u \|_{1,h} \leq C \left( \sum_{\kappa \in T_h} \alpha^2 \frac{h^2_{2\kappa}}{p^{2\kappa}} \| u \|_{k_{\kappa+1,\kappa}} \right)^{1/2},
\]

where \( s_\kappa = \min \{ p_\kappa, k_\kappa \} \). The result follows by combining the above with Lemma 4.3. \( \square \)

Provided that sufficient regularity is assumed on the solution to the boundary value problem (2.2), we now show that the symmetric version of the IP-DGFEM (with \( \theta = -1 \)) is \( h \)-optimal and slightly \( p \)-suboptimal when the discretisation error is measured in the \( L^2 \) norm.

**Theorem 4.5** Consider the setting of Theorem 4.4, and let \( \theta = -1 \). Assume, in addition, that the coefficient \( a \) (being the same as in (2.2)) and the domain \( \Omega \) of the boundary value problem

\[
-\nabla \cdot (a \nabla z) = g \quad \text{in} \quad \Omega, \quad z = 0 \quad \text{on} \quad \partial \Omega,
\]

are such that \( z \in H^2(\Omega), \quad (\sqrt{a} \nabla z) \in H^1(\Omega), \) and for any \( g \in L^2(\Omega) \) we have \( \| z \|_{2,\Omega} \leq C \| g \|_{0,\Omega} \), with the positive constant \( C \) depending only on the domain \( \Omega \) (c.f., Remark 4.6). Then, the following error bound holds:

\[
\| u - u_{DG} \|_{0,\Omega} \leq C h \left( \sum_{\kappa \in T_h} \frac{\alpha^2}{p^{2\kappa}} \| u \|_{k_{\kappa+1,\kappa}} \right)^{1/2},
\]

where \( s_\kappa = \min \{ p_\kappa, k_\kappa \} \), \( C_a = \max_{\kappa \in T_h} \{ \alpha_\kappa, \max_{\{i,j\} \in \{1,2\}} \| \nabla (\sqrt{a})_{ij} \|_{\infty,\kappa} \} \), and \( C \) is independent of \( u, p_\kappa \) and \( h_\kappa \), but depends on \( k = \max_{\kappa \in T_h} k_\kappa \), the parameter \( \rho \) in (2.5), the constant \( C_\sigma \), and on \( \Omega \).

**Proof.** The proof follows the same steps as the one of Theorem 3.3 in [11]. We let \( g = u - u_{DG} \); then, for the solution \( z \) of the problem (4.7), we have that \( \| u - u_{DG} \|^2_{0,\Omega} = \ell_{in}(u - u_{DG}) = B_{in}(z, u - u_{DG}) - R(z, u - u_{DG}) \). In addition, for any \( \tilde{z} \in S^p \) we have \( B_{in}(\tilde{z}, u - u_{DG}) = B_{in}(u, \tilde{z}) - B_{in}(u_{DG}, \tilde{z}) = B_{in}(u, \tilde{z}) - \ell_{in}(\tilde{z}) = R(u, \tilde{z}) = -R(u, z - \tilde{z}) \). This implies that

\[
\| u - u_{DG} \|^2_{0,\Omega} = B_{in}(z - \tilde{z}, u - u_{DG}) - R(u, z - \tilde{z}) - R(z, u - u_{DG}).
\]

For the first term on the right-hand side, we use the continuity result from Lemma 4.1; for the second term we use (4.4) with \( k_\kappa = 1 \); and for the third term we use (4.4) with \( s_\kappa = \min \{ p_\kappa, k_\kappa \} \), to obtain

\[
\| u - u_{DG} \|^2_{0,\Omega} \leq C \left( \sum_{\kappa \in T_h} \frac{h^2_{2\kappa}}{p^{2\kappa}} \| u \|_{k_{\kappa+1,\kappa}} \right)^{1/2} \| z - \tilde{z} \|_{1,h}.
\]

(4.9)
where we have used the (easy to prove) bound $\|\sqrt{a} \nabla z\|_{1,\kappa} \leq C_a \|z\|_{2,\kappa}$. Choosing $\tilde{z} = \Pi_{H^1} z$, and working as in (4.6) with $k_\kappa = 1$, we have that

$$
\|z - \tilde{z}\|_{1,h} \leq C \sum_{\kappa \in T_h} \frac{\alpha_\kappa h_\kappa}{p_\kappa^{1/2}} \|z\|_{2,\kappa} \leq C \max_{\kappa \in T_h} \left( \frac{\alpha_\kappa h_\kappa}{p_\kappa^{1/2}} \right) \|z\|_{2,\Omega} \\
\leq C C_a \max_{\kappa \in T_h} \left( \frac{h_\kappa}{p_\kappa^{1/2}} \right) \|u - u_{DG}\|_{0,\Omega}.
$$

(4.10)

Applying this along with (4.5), we obtain the result. \qed

**Remark 4.6** The assumptions of Theorem 4.5, regarding the regularity of the dual problem cannot be removed, due to the lack of uniform ellipticity of the differential operator. In the case whereby the problem is uniformly elliptic and is defined on a bounded convex polygonal domain, with square-integrable right-hand side and homogeneous Dirichlet boundary condition, the solution is $H^2$ regular; the same is true in $n$ dimensions (see [8]). A similar result holds for mixed homogeneous Dirichlet – homogeneous Neumann boundary conditions, provided that the internal angles of the polyhedron are sufficiently small (cf. Chapter 8 in [5]). The error bound (4.8) still holds in this, more general, case.

For the sake of brevity, we only considered augmented Sobolev space regularity for the solution $u$ (i.e., $u|_\kappa \in H^{k_\kappa+1}(\kappa)$, $\kappa \in T_h$) in Theorem 4.4 and in Theorem 4.5. When the solution $u$ belongs merely to a (standard) Sobolev space (i.e., $u|_\kappa \in H^{k_\kappa+1}(\kappa)$, $\kappa \in T_h$), completely analogous error analysis holds. However, the resulting error bounds, in this case, will still be optimal with respect to the meshsize $h$, but suboptimal with respect to the polynomial degree $p$ (by half an order of $p$ for the energy norm bound, and by a whole order of $p$ for the $L^2$ error norm for the SIP DGFEM; c.f., [7]).

Moreover, when the solution is analytic in an open neighbourhood of $\Omega$, it is possible to prove exponential convergence estimates for the $p$-version DGFEM considered here. The method of proof is completely analogous to the proof presented, e.g., in [6].

Also, the error bounds presented in this section hold in the case of quadrilateral meshes. The analysis in the case of triangular meshes proceeds along the same lines, but instead of the $H^1$ projector a standard $H^1$-stable projection operator (as, for example, Babuška-Suri projector [2]) is used. The energy norm error bound, in this case, is half-an-order suboptimal in the polynomial degree, and in the case of the SIP DGFEM the error bound in the $L^2$ norm will be $p$-suboptimal by a single power of $p$.

## 5 Concluding remarks

In this paper we presented a variant of the IP-DGFEM using an inconsistent formulation of the method, that enables us to prove $hp$-optimal bounds for degenerate problems. For the sake of simplicity, we confined ourselves to the 2-dimensional case; the error bounds presented here can be easily restored in the general $n$-dimensional case by performing a standard tensor-product argument on the projection error bounds. The bounds obtained
are also valid in the case of mixed Dirichlet–Neumann boundary conditions; the analysis in these cases proceeds along the same lines and, for the sake of brevity, is not considered here for brevity.

Acknowledgment

We would like to express our sincere gratitude to Professor Endre Süli (University of Oxford) for his valuable comments on an earlier version of this note.

References


