Int. Conference on Boundary and Interior Layers
BAIL 2006
G. Lube, G. Rapin (Eds)
© University of Göttingen, Germany, 2006

Adaptivity and A Posteriori Error Estimation For DG Methods on Anisotropic Meshes*

Paul Houston[1], Emmanuil H. Georgoulis[2], and Edward Hall[2]

[1]School of Mathematical Sciences, University of Nottingham, University Park, Nottingham NG7 2RD, UK
Paul.Houston@nottingham.ac.uk

[2]Department of Mathematics, University of Leicester, Leicester LE1 7RH, UK
{Emmanuil.Georgoulis,ejch1}@mcs.le.ac.uk

1 Introduction

In many application areas involving the mathematical modeling of convection, diffusion, and reaction processes, diffusion can be small (compared to the convection and the reaction coefficients), degenerate, or even identically equal to zero in subregions of the computational domain. This multi-scale behavior between convection and diffusion creates various challenges in the endeavor of computing accurate numerical approximations to PDE problems of this type in an efficient manner. In particular, computationally demanding features may appear in the analytical solutions of such problems; these include boundary/interior layers or even discontinuities in the subregions where the problem is of hyperbolic type. When structures such as layers or discontinuities are present in the solution, the use of anisotropically refined meshes, which aim to provide the necessary mesh resolution in the directions along these structures, is essential in order to reduce the number of degrees of freedom required to accurately resolve these sharp features.

Discontinuous Galerkin finite element methods (DGFEMs) exhibit attractive properties in the numerical approximation of problems of hyperbolic or nearly–hyperbolic type. Indeed, DGFEMs are, by construction, locally conservative, and moreover exhibit enhanced stability properties in the vicinity of boundary/interior layers and discontinuities present in the analytical solution. Additionally, DGFEMs offer advantages in the context of *hp*-adaptivity, such as increased flexibility in the mesh design (irregular grids are admissible) and the freedom of choosing the elemental polynomial degrees without the need to enforce any conformity requirements. Thereby, the combination of DGFEMs, that produce stable approximations even in unresolved regions of the computational domain, and anisotropic mesh refinement, which aims to provide the desired grid resolution in appropriate spatial directions, is an appealing technique for the numerical approximation of these types of problems.

In this work, we consider the *a priori* and *a posteriori* error analysis of discontinuous Galerkin interior penalty methods for second–order partial differential equations with nonnegative characteristic form on anisotropically refined computational meshes. In particular, we discuss the question of error estimation for linear target functionals, such as the outflow flux and the local average of the solution. The *a priori* error estimation is based on exploiting the analysis developed in the article [9], which assumed that the underlying computational mesh is shape–regular, together with an extension of the techniques developed in [4] which precisely describe the anisotropy of the mesh; for related anisotropic approximation results, we refer to [1, 16, 15, 3], for example. More specifically, we employ tools from tensor analysis, along with local singular-value

*PH was supported by the EPSRC (Grant GR/R76615).

decompositions of the Jacobi matrix of the local elemental mappings, to derive directionally-sensitive bounds for arbitrary polynomial degree approximations, thus generalizing the ideas presented in [4], where only the case of approximation with conforming linear elements was considered. These interpolation error bounds are then employed to derive general anisotropic *a priori* error bounds for the DGFEM approximation of linear functionals of the underlying analytical solution.

Type I *a posteriori* error bounds are derived based on employing the dual weighted residual approach, cf. [2, 10, 13, 14], for example. Based on our *a posteriori* error bound we design and implement the corresponding adaptive algorithm to ensure the reliable and efficient control of the error in the prescribed target functional to within a given tolerance. This involves exploiting both local isotropic and anisotropic mesh refinement. To this end, we develop a new anisotropic mesh refinement strategy, based on choosing the most competitive subdivision of a given element κ from a series of trial (Cartesian) refinements. The superiority of the proposed algorithm in comparison with standard isotropic mesh refinement will be illustrated by a series of numerical experiments.

The paper is structured as follows. In Section 2 we introduce the model problem and formulate its discontinuous Galerkin finite element approximation. Then, in Sections 3, 4, and 5 we develop the *a posteriori* and *a priori* error analyses of the error measured in terms of certain linear target functionals of practical interest. Guided by our *a posteriori* error analysis, in Section 6 we design an adaptive finite element algorithm to guarantee reliable and efficient control of the error in the computed functional to within a fixed user–defined tolerance based on employing a combination of local isotropic and anisotropic mesh refinement. The performance of the resulting refinement strategy is then studied in Section 7 through a series of numerical experiments. Finally, in Section 8 we summarize the work presented in this paper and draw some conclusions.

2 Model problem and discretization

Given that Ω is a bounded Lipschitz domain in $\mathbb{R}^d$, $d \geq 2$, with boundary $\Gamma = \partial\Omega$, we consider the linear second–order partial differential equation

$$\mathcal{L}u \equiv -\nabla \cdot (a\nabla u) + \nabla \cdot (\mathbf{b}u) + cu = f \ , \tag{1}$$

where f is a real–valued function belonging to $L_2(\Omega)$, and the real–valued coefficients a, $\mathbf{b}$, and c satisfy: $a(x) = \{a_{ij}(x)\}_{i,j=1}^d \in L_\infty(\Omega)^{d\times d}_{\rm sym}$, $\mathbf{b}(x) = \{b_i(x)\}_{i=1}^d \in W^{1,\infty}(\Omega)^d$, and $c(x) \in L_\infty(\Omega)$, respectively. We shall suppose throughout that the characteristic form associated with the principal part of the differential operator $\mathcal{L}$ is nonnegative; namely,

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

For simplicity, we shall assume that the entries of the matrix a are piecewise continuous on $\bar{\Omega}$; this hypothesis is sufficiently general to cover most situations of practical relevance. Let $\mathbf{n}(x)$ denote the unit outward normal vector to Γ at $x \in \Gamma$ and define the following subsets of Γ:

$$\Gamma_0 = \left\{ x \in \Gamma : \ \mathbf{n}(x)^\top a(x)\mathbf{n}(x) > 0 \right\} ,$$
$$\Gamma_- = \{ x \in \Gamma\backslash\Gamma_0 : \ \mathbf{b}(x) \cdot \mathbf{n}(x) < 0 \} \ , \quad \Gamma_+ = \{ x \in \Gamma\backslash\Gamma_0 : \ \mathbf{b}(x) \cdot \mathbf{n}(x) \geq 0 \} \ .$$

The sets $\Gamma_\mp$ will be referred to as the inflow and outflow boundary, respectively. With these definitions we have that $\Gamma = \Gamma_0 \cup \Gamma_- \cup \Gamma_+$. We shall further decompose Γ_0 into two connected parts, $\Gamma_{\rm D}$ and $\Gamma_{\rm N}$, and supplement (1) with the following boundary conditions:

$$u = g_{\rm D} \quad \text{on} \quad \Gamma_{\rm D} \cup \Gamma_- \ , \qquad \mathbf{n} \cdot (a\nabla u) = g_{\rm N} \quad \text{on} \quad \Gamma_{\rm N} \ ; \tag{3}$$

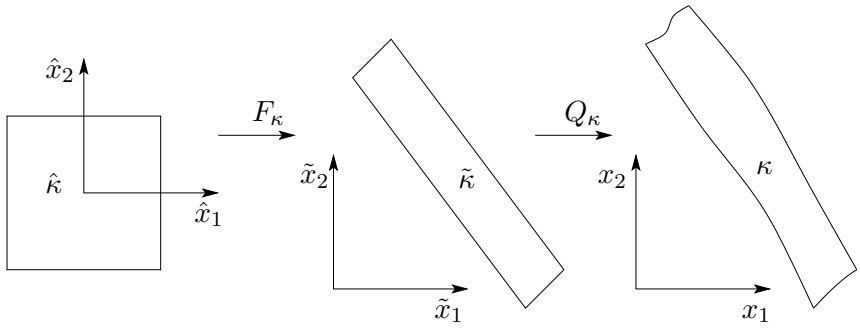

Figure 1: Construction of the element mapping via the composition of an affine mapping F_κ and a C^1–diffeomorphism Q_κ.

here, we adopt the (physically reasonable) hypothesis that $\mathbf{b} \cdot \mathbf{n} \geq 0$ on $\Gamma_{\rm N}$, whenever $\Gamma_{\rm N}$ is nonempty. Additionally, we assume that the following (standard) positivity hypothesis holds: $c(x) + 1/2\, \nabla \cdot \mathbf{b}(x) > 0$ a.e. $x \in \Omega$, and define the positive function c_0 by

$$(c_0(x))^2 = c(x) + \frac{1}{2}\, \nabla \cdot \mathbf{b}(x) \ \text{ a.e. } \ x \in \Omega\,. \tag{4}$$

We note that (1), (3) includes a range of physically relevant problems, such as the mixed boundary value problem for an elliptic equation corresponding to the case when (2) holds with strict inequality, as well as the case of a linear transport problem associated with the choice of $a \equiv 0$ on $\bar{\Omega}$. Our aim here is to develop, in a unified manner, the *a posteriori* and *a priori* error analysis of the interior penalty discontinuous finite element approximation to (1), (3).

2.1 Meshes, finite element spaces, and traces

Let $\mathcal{T}_h = \{\kappa\}$ be a 1–irregular subdivision of the (polygonal) domain Ω into disjoint open element domains κ constructed through the use of the mappings $Q_\kappa \circ F_\kappa$, where $F_\kappa : \hat{\kappa} \to \tilde{\kappa}$ is an affine mapping from the reference element $\hat{\kappa} = (-1,1)^d$ to $\tilde{\kappa}$, and $Q_\kappa : \tilde{\kappa} \to \kappa$ is a C^1–diffeomorphism from $\tilde{\kappa}$ to the physical element κ, cf. [11, 5]; the case when $d = 2$ is depicted in Figure 1. The mapping F_κ defines the size and orientation of the element κ, while Q_κ defines the shape of κ, without any significant rescaling, or indeed change of orientation. With this in mind, we assume that the element mapping Q_κ is close to the identity in the following sense: the Jacobi matrix J_{Q_κ} of Q_κ satisfies

$$C_1^{-1} \leq \|\det J_{Q_\kappa}\|_{L_\infty(\kappa)} \leq C_1, \quad \|J_{Q_\kappa}^{-\top}\|_{L_\infty(\kappa)} \leq C_2, \quad \|J_{Q_\kappa}^{-\top}\|_{L_\infty(\partial\kappa)} \leq C_3 \tag{5}$$

for all κ in $\mathcal{T}_h$ uniformly throughout the mesh for some positive constants C_1, C_2, and C_3. This will be important as our error estimates will be expressed in terms of Sobolev norms over the element domains $\tilde{\kappa}$, in order to ensure that only the scaling and orientation introduced by the affine element maps F_κ are present in the analysis. Writing m_κ, $m_{\tilde{\kappa}}$, and $m_{\hat{\kappa}}$ to denote the d–dimensional (Hausdorff) measure of the elements κ, $\tilde{\kappa}$, and $\hat{\kappa}$, respectively, the above condition (5) implies that there exists a positive constant C_4 such that

$$C_4^{-1} m_{\tilde{\kappa}} \leq m_\kappa \leq C_4 m_{\tilde{\kappa}} \quad \forall \kappa \in \mathcal{T}_h.$$

The above maps are assumed to be constructed in such a manner to ensure that the union of the closure of the disjoint open elements $\kappa \in \mathcal{T}_h$ forms a covering of the closure of Ω, i.e., $\bar{\Omega} = \cup_{\kappa \in \mathcal{T}_h} \bar{\kappa}$. For a function v defined on κ, $\kappa \in \mathcal{T}_h$, we write $\tilde{v} = v \circ Q_\kappa$ and $\hat{v} = \tilde{v} \circ F_\kappa$ to denote the corresponding functions on the elements $\tilde{\kappa}$ and $\hat{\kappa}$, respectively.

Associated with $\mathcal{T}_h$, we write $H^s(\Omega, \mathcal{T}_h)$ to denote the broken Sobolev space of order $s \geq 0$, equipped with the broken Sobolev seminorm $|\cdot|_{s,\mathcal{T}_h}$ and norm $\|\cdot\|_{s,\mathcal{T}_h}$. For $u \in H^1(\Omega, \mathcal{T}_h)$ we define the broken gradient $\nabla_{\mathcal{T}_h} u$ of u by $(\nabla_{\mathcal{T}_h} u)|_\kappa = \nabla(u|_\kappa)$, $\kappa \in \mathcal{T}_h$.

2.2 Interior penalty discontinuous Galerkin method

In this section we introduce the (symmetric) interior penalty discretization of the advection–diffusion–reaction problem (1), (3). To this end, we introduce the following notation. Given a polynomial degree $p \geq 1$ we define the finite element space $S_{h,p}$ as follows

$$S_{h,p} \;=\; \{u \in L_2(\Omega) \,:\, u|_\kappa \circ Q_\kappa \circ F_\kappa \in \mathcal{Q}_p(\kappa);\ \kappa \in \mathcal{T}_h\}\,,$$

where $\mathcal{Q}_p(\kappa) = \operatorname{span}\{\hat{x}^\alpha : 0 \leq \alpha_i \leq p,\ \ 1 \leq i \leq d\}$.

An *interior face* of $\mathcal{T}_h$ is defined as the (non-empty) $(d-1)$–dimensional interior of $\partial\kappa_i \cap \partial\kappa_j$, where κ_i and κ_j are two adjacent elements of $\mathcal{T}_h$, not necessarily matching. A *boundary face* of $\mathcal{T}_h$ is defined as the (non-empty) $(d-1)$–dimensional interior of $\partial\kappa \cap \Gamma$, where κ is a boundary element of $\mathcal{T}_h$. We denote by Γ_{int} the union of all interior faces of $\mathcal{T}_h$. Given a face $f \subset \Gamma_{\text{int}}$, shared by the two elements κ_i and κ_j, where the indices i and j satisfy $i > j$, we write $\mathbf{n}_f$ to denote the (numbering–dependent) unit normal vector which points from κ_i to κ_j; on boundary faces, we set $\mathbf{n}_f = \mathbf{n}$. Further, for $v \in H^1(\Omega, \mathcal{T}_h)$ we define the jump of v across f and the mean value of v on f, respectively, by $[v] = v|_{\partial\kappa_i \cap f} - v|_{\partial\kappa_j \cap f}$ and $\langle v \rangle = 1/2\left(v|_{\partial\kappa_i \cap f} + v|_{\partial\kappa_j \cap f}\right)$. On a boundary face $f \subset \partial\kappa$, we set $[v] = v|_{\partial\kappa \cap f}$ and $\langle v \rangle = v|_{\partial\kappa \cap f}$. Finally, given a function $v \in H^1(\Omega, \mathcal{T}_h)$ and an element $\kappa \in \mathcal{T}_h$, we denote by v_κ^+ (respectively, v_κ^-) the interior (respectively, exterior) trace of v defined on $\partial\kappa$ (respectively, $\partial\kappa \backslash \Gamma$). Since below it will always be clear from the context which element κ in the subdivision $\mathcal{T}_h$ the quantities v_κ^+ and v_κ^- correspond to, for the sake of notational simplicity we shall suppress the letter κ in the subscript and write, respectively, v^+ and v^- instead.

Given that κ is an element in the subdivision $\mathcal{T}_h$, we denote by $\partial\kappa$ the union of $(d-1)$–dimensional open faces of κ. Let $x \in \partial\kappa$ and suppose that $\mathbf{n}_\kappa(x)$ denotes the unit outward normal vector to $\partial\kappa$ at x. We define the inflow and outflow parts of $\partial\kappa$, respectively, by

$$\partial_-\kappa = \{x \in \partial\kappa :\ \mathbf{b}(x) \cdot \mathbf{n}_\kappa(x) < 0\}\,, \quad \partial_+\kappa = \{x \in \partial\kappa :\ \mathbf{b}(x) \cdot \mathbf{n}_\kappa(x) \geq 0\}\,.$$

For simplicity of presentation, we suppose that the entries of the matrix a are constant on each element κ in $\mathcal{T}_h$; i.e., $a \in [S_{h,0}]^{d\times d}_{\text{sym}}$. The extension of our results to general $a \in L_\infty(\Omega)^{d\times d}_{\text{sym}}$ follow analogously based on employing the modified DG method proposed in [7]. In the following, we write $\bar{a} = |\sqrt{a}\,|_2^2$, where $|\cdot|_2$ denotes the matrix norm subordinate to the l_2–vector norm on $\mathbb{R}^d$ and $\bar{a}_\kappa = \bar{a}|_\kappa$.

The DGFEM approximation of (1), (3) is defined as follows: find u_{DG} in $S_{h,p}$ such that

$$B_{\text{DG}}(u_{\text{DG}}, v) = \ell_{\text{DG}}(v) \tag{6}$$

for all $v \in S_{h,p}$. Here, the bilinear form $B_{\text{DG}}(\cdot,\cdot)$ is defined by

$$B_{\text{DG}}(w, v) = B_a(w, v) + B_{\mathbf{b}}(w, v) - B_f(v, w) - B_f(w, v) + B_\vartheta(w, v)\,,$$

where

$$B_a(w, v) = \sum_{\kappa\in\mathcal{T}_h} \int_\kappa a\nabla w \cdot \nabla v \,\mathrm{d}x\,,$$

$$B_{\mathbf{b}}(w, v) = \sum_{\kappa\in\mathcal{T}_h} \left\{ -\int_\kappa (w\,\mathbf{b} \cdot \nabla v - cwv)\,\mathrm{d}x + \int_{\partial_+\kappa} (\mathbf{b} \cdot \mathbf{n}_\kappa)\, w^+ v^+ \,\mathrm{d}s + \int_{\partial_-\kappa\backslash\Gamma} (\mathbf{b} \cdot \mathbf{n}_\kappa)\, w^- v^+ \,\mathrm{d}s \right\},$$

$$B_f(w, v) = \int_{\Gamma_{\text{int}}\cup\Gamma_{\text{D}}} \langle (a\nabla w) \cdot \mathbf{n}_f \rangle [v] \,\mathrm{d}s\,, \qquad B_\vartheta(w, v) = \int_{\Gamma_{\text{int}}\cup\Gamma_{\text{D}}} \vartheta [w][v] \,\mathrm{d}s\,,$$

and the linear functional $\ell_{\text{DG}}(\cdot)$ is given by

$$\begin{aligned} \ell_{\text{DG}}(v) \;=\; & \sum_{\kappa\in\mathcal{T}_h} \left(\int_\kappa fv \,\mathrm{d}x - \int_{\partial_-\kappa\cap(\Gamma_{\text{D}}\cup\Gamma_-)} (\mathbf{b} \cdot \mathbf{n}_\kappa)\, g_{\text{D}}\, v^+ \,\mathrm{d}s \right. \\ & \left. - \int_{\partial\kappa\cap\Gamma_{\text{D}}} g_{\text{D}} ((a\nabla v^+) \cdot \mathbf{n}_\kappa) \,\mathrm{d}s + \int_{\partial\kappa\cap\Gamma_{\text{N}}} g_{\text{N}} v^+ \,\mathrm{d}s + \int_{\partial\kappa\cap\Gamma_{\text{D}}} \vartheta g_{\text{D}} v^+ \,\mathrm{d}s \right). \end{aligned}$$

Here ϑ is called the *discontinuity-penalization* parameter and is defined by

$$\vartheta|_f = \vartheta_f \quad \text{for} \quad f \subset \Gamma_{\text{int}} \cup \Gamma_{\text{D}},$$

where ϑ_f is a nonnegative constant on face f. The precise choice of ϑ_f, which depends on a and the discretization parameters, will be discussed in detail in the next section. We shall adopt the convention that faces $f \subset \Gamma_{\text{int}} \cup \Gamma_{\text{D}}$ with $\vartheta|_f = 0$ are omitted from the integrals appearing in the definition of $B_\vartheta(w,v)$ and $\ell_{\text{DG}}(v)$, although we shall not highlight this explicitly in our notation; the same convention is adopted in the case of integrals where the integrand contains the factor $1/\vartheta$. Thus, in particular, the definition of the DG-norm, cf. (7) below, is meaningful even if $\vartheta|_f$ happens to be equal to zero on certain faces $f \subset \Gamma_{\text{int}} \cup \Gamma_{\text{D}}$, given that such faces are understood to be excluded from the region of integration.

3 Stability analysis

Before embarking on the error analysis of the discontinuous Galerkin method (6), we first derive some preliminary results. Let us first introduce the DG–norm $|||\cdot|||$ by

$$\begin{aligned} |||w|||^2 \;=\; & \sum_{\kappa\in\mathcal{T}_h}\Big(\|\sqrt{a}\,\nabla w\|^2_{L_2(\kappa)} + \|c_0 w\|^2_{L_2(\kappa)} + \frac{1}{2}\|w^+\|^2_{\partial_-\kappa\cap(\Gamma_{\text{D}}\cup\Gamma_-)} + \frac{1}{2}\|w^+ - w^-\|^2_{\partial_-\kappa\setminus\Gamma} \\ & + \frac{1}{2}\|w^+\|^2_{\partial_+\kappa\cap\Gamma}\Big) + \int_{\Gamma_{\text{int}}\cup\Gamma_{\text{D}}} \vartheta[w]^2\,\mathrm{d}s + \int_{\Gamma_{\text{int}}\cup\Gamma_{\text{D}}} \frac{1}{\vartheta}\langle(a\,\nabla w)\cdot\mathbf{n}_f\rangle^2\,\mathrm{d}s\,, \end{aligned} \tag{7}$$

where $\|\cdot\|_\tau$, $\tau\subset\partial\kappa$, denotes the (semi)norm associated with the (semi)inner-product $(v,w)_\tau = \int_\tau |\mathbf{b}\cdot\mathbf{n}_\kappa| vw\,\mathrm{d}s$, and c_0 is as defined in (4).

For a given face $f \subset \Gamma_{\text{int}}\cup\Gamma_{\text{D}}$, such that $f\subset\partial\kappa$, for some $\kappa\in\mathcal{T}_h$, we write $\tilde{f}$ and $\hat{f}$ to denote the respective faces of the mapped elements $\tilde{\kappa}$ and $\hat{\kappa}$, respectively, based on employing the element mappings Q_κ and F_κ. More precisely, we write $\tilde{f} = Q_\kappa^{-1}(f)$ and $\hat{f} = F_\kappa^{-1}(\tilde{f})$. Further, we define m_f, $m_{\tilde{f}}$, and $m_{\hat{f}}$ to denote the $(d-1)$–dimensional measure (volume) of the faces f, $\tilde{f}$, and $\hat{f}$, respectively. In view of (5), we note that there exists a positive constant C_5, such that

$$C_5^{-1} m_{\tilde{f}} \le m_f \le C_5 m_{\tilde{f}} \tag{8}$$

for every face $f\subset\Gamma_{\text{int}}\cup\Gamma_{\text{D}}$. Moreover, the surface Jacobian $S_{f,\tilde{f}}$ arising in the transformation of the face f to $\tilde{f}$ may be uniformly bounded in the following manner

$$\|S_{f,\tilde{f}}\|_{L_\infty(\tilde{f})} \le C_6 \tag{9}$$

for all faces $f\subset\Gamma_{\text{int}}\cup\Gamma_{\text{D}}$, where C_6 is a positive constant.

We now define the function $\mathtt{h}$ in $L_\infty(\Gamma_{\text{int}}\cup\Gamma_{\text{D}})$, as $\mathtt{h}(x) = \min\{m_{\kappa_1}, m_{\kappa_2}\}/m_f$, if x is in the interior of $f = \partial\kappa_1\cap\partial\kappa_2$ for two neighboring elements in the mesh $\mathcal{T}_h$, and $\mathtt{h}(x) = m_\kappa/m_f$, if x is in the interior of $f = \partial\kappa\cap\Gamma_{\text{D}}$. Similarly, we define the function $\mathtt{a}$ in $L_\infty(\Gamma_{\text{int}}\cup\Gamma_{\text{D}})$ by $\mathtt{a}(x) = \min\{\bar{a}_{\kappa_1}, \bar{a}_{\kappa_2}\}$ if x is in the interior of $e = \partial\kappa_1\cap\partial\kappa_2$, and $\mathtt{a}(x) = \bar{a}_\kappa$ if x is in the interior of $\partial\kappa\cap\Gamma_{\text{D}}$. With this notation, we now provide the following coercivity result for the bilinear form $B_{\text{DG}}(\cdot,\cdot)$ over $S_{h,p}\times S_{h,p}$.

Theorem 3.1 *Define the discontinuity-penalization parameter ϑ arising in (6) by*

$$\vartheta|_f \equiv \vartheta_f = C_\vartheta \frac{\mathtt{a}}{\mathtt{h}} \quad \textit{for} \quad f\subset\Gamma_{\text{int}}\cup\Gamma_{\text{D}}, \tag{10}$$

where C_ϑ is a sufficiently large positive constant. Then, there exists a positive constant C, which depends only on the dimension d and the polynomial degree p, such that

$$B_{\text{DG}}(v,v) \ge C|||v|||^2 \quad \forall v\in S_{h,p}.$$

Proof. This result follows by application of the inverse estimate derived in [6], following the general argument presented, for example, in the articles [18, 12]. □

For the proceeding error analysis, we assume that the solution u to the boundary value problem (1), (3) is sufficiently smooth: namely, $u \in H^{3/2+\varepsilon}(\Omega, \mathcal{T}_h)$, $\varepsilon > 0$, and the functions u and $(a\nabla u) \cdot \mathbf{n}_f$ are continuous across each face $f \subset \partial\kappa \backslash \Gamma$ that intersects the subdomain of ellipticity, $\Omega_a = \{x \in \bar{\Omega} : \boldsymbol{\zeta}^\top a(x)\boldsymbol{\zeta} > 0 \ \forall \boldsymbol{\zeta} \in \mathbb{R}^d\}$. If this smoothness requirement is violated, the discretization method has to be modified accordingly, cf. [12]. We note that under these assumptions, the following Galerkin orthogonality property holds:

$$B_{\mathrm{DG}}(u - u_{\mathrm{DG}}, v) = 0 \quad \forall v \in S_{h,p}\,. \tag{11}$$

For simplicity of presentation, it will be assumed in the proceeding analysis, that the velocity vector $\mathbf{b}$ satisfies the following assumption:

$$\mathbf{b} \cdot \nabla_{\mathcal{T}_h} v \in S_{h,p} \quad \forall v \in S_{h,p}\,. \tag{12}$$

To ensure that (1) is then meaningful (i.e., that the characteristic curves of the differential operator $\mathcal{L}$ are correctly defined), we still assume that $\mathbf{b} \in \left[W^1_\infty(\Omega)\right]^d$.

4 Approximation results

In this section we develop the necessary approximation results needed for the forthcoming *a priori* error estimation developed in Section 5. To this end, on the reference element $\hat{\kappa}$, we define $\hat{\Pi}_p$ to denote the orthogonal projector in $L_2(\hat{\kappa})$ onto the space of polynomials $Q_p(\hat{\kappa})$; i.e., given that $\hat{v} \in L_2(\hat{\kappa})$, we define $\hat{\Pi}_p\hat{v}$ by

$$(\hat{v} - \hat{\Pi}_p\hat{v}, \hat{w})_{\hat{\kappa}} = 0 \quad \forall \hat{w} \in Q_p(\hat{\kappa})\,,$$

where $(\cdot,\cdot)_{\hat{\kappa}}$ denotes the $L_2(\hat{\kappa})$ inner product. Similarly, we define the L_2-projection operators $\tilde{\Pi}_p$ and Π_p on $\tilde{\kappa}$ and κ, respectively, by the relations

$$\tilde{\Pi}_p\tilde{v} := (\hat{\Pi}_p(\tilde{v} \circ F_\kappa)) \circ F_\kappa^{-1}, \qquad \Pi_p v := (\tilde{\Pi}_p(v \circ Q_\kappa)) \circ Q_\kappa^{-1},$$

for $\tilde{v} \in L_2(\tilde{\kappa})$ and $v \in L_2(\kappa)$, respectively.

We remark that this choice of projector is essential in the following *a priori* error analysis, in order to ensure that

$$(u - \Pi_p u, \mathbf{b} \cdot \nabla_{\mathcal{T}_h} v) = 0 \tag{13}$$

for all v in $S_{h,p}$. We remark that this same choice of projector is also necessary in the corresponding case when (12) fails to hold; in this situation an equality of the form (13) with $\mathbf{b}$ replaced by a suitable projection of $\mathbf{b}$ is still necessary for the underlying analysis; cf. [5].

We now quote the following approximation result on the reference element $\hat{\kappa}$.

Lemma 4.1 *Let $\hat{\kappa}$ be the reference element $(-1,1)^d$, and let $\hat{f}$ denote one of its faces. Given a function $\hat{v} \in H^k(\hat{\kappa})$, the following error bounds hold:*

$$\|\hat{v} - \hat{\Pi}_p\hat{v}\|_{L_2(\hat{\kappa})} \leq C|\hat{v}|_{H^s(\hat{\kappa})}, \quad 0 \leq s \leq \min(p+1, k), \tag{14}$$

$$\|\hat{v} - \hat{\Pi}_p\hat{v}\|_{L_2(\hat{f})} \leq C|\hat{v}|_{H^s(\hat{\kappa})}, \quad 1 \leq s \leq \min(p+1, k), \tag{15}$$

$$|\hat{v} - \hat{\Pi}_p\hat{v}|_{H^1(\hat{\kappa})} \leq C|\hat{v}|_{H^s(\hat{\kappa})}, \quad 1 \leq s \leq \min(p+1, k), \tag{16}$$

$$|\hat{v} - \hat{\Pi}_p\hat{v}|_{H^1(\hat{f})} \leq C|\hat{v}|_{H^s(\hat{\kappa})}, \quad 2 \leq s \leq \min(p+1, k), \tag{17}$$

where C is a positive constant which depends only on the dimension d and the polynomial order p.

Proof. The proofs of the estimates (14), (15), (16), and (17) are, for example, given, respectively, in [12] (Lemma 3.4 and Lemma 3.6, respectively), and [5] (Lemma 3.7 on p. 48 and Corollary 3.21 on p. 59, respectively). □

Corollary 4.2 *Using the notation of Lemma 4.1, there exists a positive constant C, which depends only on the dimension d and the polynomial order p, such that*

$$\begin{aligned}
\|v - \Pi_p v\|_{L_2(\kappa)} &\leq C|\det(J_{F_\kappa})|^{1/2}\,|\hat{v}|_{H^s(\hat{\kappa})}, && 0 \leq s \leq \min(p+1,k),\\
\|v - \Pi_p v\|_{L_2(f)} &\leq C|m_f|^{1/2}|\hat{v}|_{H^s(\hat{\kappa})}, && 1 \leq s \leq \min(p+1,k),\\
|v - \Pi_p v|_{H^1(\kappa)} &\leq C|\det(J_{F_\kappa})|^{1/2}\,\|J_{F_\kappa}^{-\top}\|_2\,|\hat{v}|_{H^s(\hat{\kappa})}, && 1 \leq s \leq \min(p+1,k),\\
|v - \Pi_p v|_{H^1(f)} &\leq C|m_f|^{1/2}\,\|J_{F_\kappa}^{-\top}\|_2\,|\hat{v}|_{H^s(\hat{\kappa})}, && 2 \leq s \leq \min(p+1,k).
\end{aligned}$$

Proof. The proof of the each inequality stated in the corollary is based on exploiting a standard scaling argument to the respective left–hand sides of the approximation results stated in Lemma 4.1, together with (5), (8), and (9); see [6] for details. □

Finally, it remains to scale the $H^s(\hat{\kappa})$, $s \geq 0$, semi-norm defined on the reference element $\hat{\kappa}$ to $\tilde{\kappa}$ based on employing the affine element transformation F_κ. In order to retain the anisotropic mesh information within the Jacobi matrix J_{F_κ}, we first re-write the square of the $H^s(\hat{\kappa})$ semi-norm of a function $\hat{v}$ in terms of the integral of the square of the Frobenius norm of an sth–order tensor containing the s–order derivatives of $\hat{v}$. With this definition the transformation of the s–order derivatives of $\hat{v}$ defined over $\hat{\kappa}$ may naturally be transformed to derivatives of the (mapped) function $\tilde{v}$ defined over $\tilde{\kappa}$. Indeed, for the case when $s = 2$, this approach is analogous to the technique employed in [4].

To this end, we now introduce the following tensor notation: here, and in the following we use calligraphic letters $\mathcal{A}, \mathcal{B}, \ldots$ to denote Nth–order tensors, where it is understood that a 0th–order tensor is a scalar, a 1st–order tensor is a vector, a 2nd–order tensor is a matrix, and so on. The following discussion regarding tensors is based on the work presented in the article [17].

Firstly, we define a way of multiplying a tensor by a matrix.

Definition 4.3 *The n-mode product of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ by a matrix $U \in \mathbb{R}^{J_n \times I_n}$, denoted by $\mathcal{A} \times_n U$, is an $I_1 \times I_2 \times \ldots \times I_{n-1} \times J_n \times I_{n+1} \times \ldots I_N$-tensor of which the entries are given by*

$$(\mathcal{A} \times_n U)_{i_1 i_2 \ldots i_{n-1} j_n i_{n+1} \ldots i_N} := \sum_{i_n=1}^{I_n} (\mathcal{A})_{i_1 i_2 \ldots i_{n-1} i_n i_{n+1} \ldots i_N} (U)_{j_n i_n}.$$

By considering a vector $\mathbf{v}$ as an $I_n \times 1$ matrix, then an n-mode product of $\mathbf{v}^\top$ and $\mathcal{A}$ can be formed to produce an $I_1 \times I_2 \times \ldots \times I_{n-1} \times 1 \times I_{n+1} \times \ldots \times I_N$-tensor. This tensor could be viewed as an $(N-1)$-tensor, but instead we leave it as an N-tensor in order that we can form other m-mode products without the value of m having to change. However, if we have a $1 \times 1 \times \ldots \times 1$-tensor then we simply view this as a scalar. The n-mode product satisfies the following property: for a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ and the matrices $F \in \mathbb{R}^{J_n \times I_n}$ and $G \in \mathbb{R}^{J_m \times I_m}$, $n \neq m$, we have

$$(\mathcal{A} \times_n F) \times_m G = (\mathcal{A} \times_m G) \times_n F = \mathcal{A} \times_n F \times_m G. \tag{18}$$

The Frobenius-norm, $\|\cdot\|_F$, of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ is defined by

$$\|\mathcal{A}\|_F^2 = \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_N=1}^{I_N} (\mathcal{A})_{i_1 i_2 \cdots i_N}^2.$$

Exploiting the properties of the Frobenius norm and the definition of the n-mode product of a tensor by a matrix, it can be shown that

$$\|\mathcal{A} \times_n Q\|_F = \|\mathcal{A}\|_F, \tag{19}$$

where $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ is an N-tensor and $Q \in \mathbb{R}^{I_n \times I_n}$ is an orthonormal matrix, cf. [6].

In order to rescale $|\hat{v}|_{H^s(\hat{\kappa})}$ to the corresponding quantity on $\tilde{\kappa}$, we first note that

$$|\hat{v}|^2_{H^s(\hat{\kappa})} = \int_{\hat{\kappa}} \|\hat{\mathcal{D}}^s(\hat{v})\|^2_F \, \mathrm{d}\hat{x},$$

where $\hat{\mathcal{D}}^s(\hat{v}) \in \mathbb{R}^{d \times d \times \cdots \times d}$ is the sth–order tensor containing the sth–order derivatives of $\hat{v}$ with respect to the coordinate system $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_d)$, i.e.,

$$(\hat{\mathcal{D}}^s(\hat{v}))_{i_1, i_2, \ldots, i_s} = \frac{\partial^s \hat{v}}{\partial \hat{x}_{i_1} \cdots \partial \hat{x}_{i_s}}, \quad i_k = 1, \ldots, d, \text{ for } k = 1, \ldots, s.$$

Thereby, for $s = 0$, $\hat{\mathcal{D}}^s(\hat{v}) = \hat{v}$, for $s = 1$, $\hat{\mathcal{D}}^s(\hat{v})$ is the gradient vector, and for $s = 2$, $\hat{\mathcal{D}}^s(\hat{v})$ is the Hessian matrix of second–order derivatives. Writing $\tilde{\mathcal{D}}^s(\tilde{v}) \in \mathbb{R}^{d \times d \times \cdots \times d}$ to denote the sth–order tensor containing the sth–order derivatives of $\tilde{v}$ with respect to the coordinate system $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_d)$, we now state the following lemma relating $|\hat{v}|^2_{H^s(\hat{\kappa})}$ to $|\tilde{v}|^2_{H^s(\tilde{\kappa})}$.

Lemma 4.4 *Under the foregoing assumptions, for $\tilde{v} \in H^s(\tilde{\kappa})$, $s \geq 0$, we have that*

$$|\hat{v}|^2_{H^s(\hat{\kappa})} = |\det(J^{-1}_{F_\kappa})| \int_{\tilde{\kappa}} \|\tilde{\mathcal{D}}^s(\tilde{v}) \times_1 J^\top_{F_\kappa} \times_2 J^\top_{F_\kappa} \times_3 \ldots \times_s J^\top_{F_\kappa}\|^2_F \, \mathrm{d}\tilde{x}. \tag{20}$$

Proof. The proof of the lemma follows by repeated application of the chain rule, together with Definition 4.3 and the property (18) above; see [6] for details. □

Remark 4.5 *For the case when $s = 0$, Lemma 4.4 simply states the change of variable formula for the L_2-norm. For $s = 1$ we note that the expression inside the Frobenius norm on the right-hand side of (20) gives rise to the usual change of variables for the gradient operator, namely,*

$$\hat{\mathcal{D}}^s(\hat{v}) \equiv \nabla_{\hat{x}} \hat{v} = \tilde{\mathcal{D}}^s(\tilde{v}) \times_1 J^\top_{F_\kappa} = J^\top_{F_\kappa} \nabla_{\tilde{x}} \tilde{v},$$

where $\nabla_{\hat{x}}$ and $\nabla_{\tilde{x}}$ denote the gradient operator with respect to the coordinate systems $\hat{x}$ and $\tilde{x}$, respectively. Similarly, for $s = 2$, this same expression may be written in the more familiar form

$$H_{\hat{x}}(\hat{v}) = J^\top_{F_\kappa} H_{\tilde{x}}(\tilde{v}) J_{F_\kappa}$$

where $H_{\hat{x}}(\cdot)$ and $H_{\tilde{x}}(\cdot)$ denote the Hessian matrix operators with respect to the coordinate systems $\hat{x}$ and $\tilde{x}$, respectively, cf. [4].

In order to describe the length scales and orientation of the element $\tilde{\kappa}$ we adopt a similar approach to that developed in [4]. Namely, we perform an SVD decomposition of the Jacobi matrix J_{F_κ} of the affine element mapping F_κ. Thereby, we write

$$J_{F_\kappa} = U_\kappa \Sigma_\kappa V^\top_\kappa,$$

where U_κ and V_κ are $d \times d$ orthogonal matrices containing the left and right singular vectors of J_{F_κ}, respectively, and $\Sigma_\kappa = \mathrm{diag}(\sigma_{1,\kappa}, \sigma_{2,\kappa}, \ldots, \sigma_{d,\kappa})$ is a $d \times d$ diagonal matrix containing the singular values $\sigma_{i,\kappa}$, $i = 1, \ldots, d$, of J_{F_κ}. By convention, we assume that $\sigma_{1,\kappa} \geq \sigma_{2,\kappa} \geq \ldots \geq \sigma_{d,\kappa} > 0$. Writing $U_\kappa = (\mathbf{u}_{1,\kappa} \ldots \mathbf{u}_{d,\kappa})$, where $\mathbf{u}_{i,\kappa}$, $i = 1, \ldots, d$, denote the left singular vectors of J_{F_κ}, we note that $\mathbf{u}_{i,\kappa}$, $i = 1, \ldots, d$, give the direction of stretching of the element

κ, while $\sigma_{i,\kappa}$, $i = 1, \ldots, d$, give the stretching lengths in the respective directions. Indeed, for axiparallel meshes, as considered in [5], for example, then $\mathbf{u}_{i,\kappa}$, $i = 1, \ldots, d$ will be parallel to the coordinates axes and $\sigma_{i,\kappa}$, $i = 1, \ldots, d$ will denote the local mesh length within the respective coordinate direction.

With this notation, we make the following observations

$$|\det(J_{F_\kappa})| = \Pi_{i=1}^d \sigma_{i,\kappa}, \quad \|J_{F_\kappa}^{-\top}\|_2 = 1/\sigma_{d,\kappa}, \quad m_f \leq C_7\, \Pi_{i=1}^{d-1} \sigma_{i,\kappa}, \tag{21}$$

where C_7 is a positive constant independent of the element size. Employing (19), we note that

$$\begin{aligned}
&\|\tilde{\mathcal{D}}^s(\tilde{v}) \times_1 J_{F_\kappa}^\top \times_2 J_{F_\kappa}^\top \times_3 \ldots \times_s J_{F_\kappa}^\top\|_F^2 \\
&\quad = \sum_{i_1=1}^{d} \sum_{i_2=1}^{d} \ldots \sum_{i_s=1}^{d} (\sigma_{i_1,\kappa}\sigma_{i_2,\kappa}\ldots\sigma_{i_s,\kappa})^2 (\tilde{\mathcal{D}}^s(\tilde{v}) \times_1 \mathbf{u}_{i_1,\kappa}^\top \times_2 \mathbf{u}_{i_2,\kappa}^\top \times_3 \ldots \times_s \mathbf{u}_{i_s,\kappa}^\top)^2 \\
&\quad \equiv D_{\tilde{\kappa}}^s(\tilde{v}, \Sigma_\kappa, U_\kappa).
\end{aligned} \tag{22}$$

Thereby, exploiting (21) and (22) together with Corollary 4.2, we deduce the following approximation results.

Theorem 4.6 *Using the notation of Lemma 4.1, there exists a positive constant C, which depends only on the dimension d and the polynomial order p, such that*

$$\begin{aligned}
\|v - \Pi_p v\|_{L_2(\kappa)} &\leq C \left[\int_{\tilde{\kappa}} D_{\tilde{\kappa}}^s(\tilde{v}, \Sigma_\kappa, U_\kappa)\, \mathrm{d}\tilde{x}\right]^{1/2}, && 0 \leq s \leq \min(p+1, k), \\
\|v - \Pi_p v\|_{L_2(f)} &\leq C|\sigma_{d,\kappa}|^{-1/2} \left[\int_{\tilde{\kappa}} D_{\tilde{\kappa}}^s(\tilde{v}, \Sigma_\kappa, U_\kappa)\, \mathrm{d}\tilde{x}\right]^{1/2}, && 1 \leq s \leq \min(p+1, k), \\
|v - \Pi_p v|_{H^1(\kappa)} &\leq C|\sigma_{d,\kappa}|^{-1} \left[\int_{\tilde{\kappa}} D_{\tilde{\kappa}}^s(\tilde{v}, \Sigma_\kappa, U_\kappa)\, \mathrm{d}\tilde{x}\right]^{1/2}, && 1 \leq s \leq \min(p+1, k), \\
|v - \Pi_p v|_{H^1(f)} &\leq C \left|\frac{m_f}{m_\kappa}\right|^{1/2} |\sigma_{d,\kappa}|^{-1} \left[\int_{\tilde{\kappa}} D_{\tilde{\kappa}}^s(\tilde{v}, \Sigma_\kappa, U_\kappa)\, \mathrm{d}\tilde{x}\right]^{1/2}, && 2 \leq s \leq \min(p+1, k).
\end{aligned}$$

Remark 4.7 *For the purposes of deriving the forthcoming* a priori *error bound on the error in the computed target functional, cf. Theorem 5.2 below, it is convenient to leave the statement of the fourth approximation result above in terms of m_f and m_κ, rather than in terms of the stretching factors $\sigma_{i,\kappa}$, $i = 1, \ldots, d$, solely, since these quantities naturally arise within the definition of the discontinuity-penalization parameter σ defined in (10).*

In the next section, we consider the *a posteriori* and *a priori* error analysis of the discontinuous Galerkin finite element method (6) in terms of certain linear target functionals of practical interest.

5 *A posteriori* and *a priori* error analysis

Very often in problems of practical importance the quantity of interest is an output or target functional $J(\cdot)$ of the solution. Relevant examples include the lift and drag coefficients for a body immersed into a viscous fluid, the local mean value of the field, or its flux through the outflow boundary of the computational domain. The aim of this section is to develop the *a posteriori* and *a priori* error analysis for general linear target functionals $J(\cdot)$ of the solution; for related work, we refer to [2, 10, 13, 14], for example.

5.1 Type I *a posteriori* error analysis

In this section we consider the derivation of so-called Type I (cf. [13]) or weighted *a posteriori* error bounds. Following the argument presented in [13, 14] we begin our analysis by considering the following *dual* or *adjoint* problem: find $z \in H^2(\Omega, \mathcal{T}_h)$ such that

$$B_{\mathrm{DG}}(w, z) = J(w) \quad \forall w \in H^2(\Omega, \mathcal{T}_h). \tag{23}$$

Let us assume that (23) possesses a unique solution. Clearly, the validity of this assumption depends on the choice of the linear functional under consideration; typical examples covered by our hypothesis are considered in [13].

For a given linear functional $J(\cdot)$ the proceeding *a posteriori* error bound will be expressed in terms of the finite element residual R_{int} defined on $\kappa \in \mathcal{T}_h$ by

$$R_{\mathrm{int}}|_\kappa = (f - \mathcal{L}u_{\mathrm{DG}})|_\kappa \,,$$

which measures the extent to which u_{DG} fails to satisfy the differential equation on the union of the elements κ in the mesh $\mathcal{T}_h$; thus we refer to R_{int} as the *internal residual*. Also, since u_{DG} only satisfies the boundary conditions approximately, the differences $g_{\mathrm{D}} - u_{\mathrm{DG}}$ and $g_{\mathrm{N}} - (a\nabla u_{\mathrm{DG}}) \cdot \mathbf{n}$ are not necessarily zero on $\Gamma_{\mathrm{D}} \cup \Gamma_-$ and Γ_{N}, respectively; thus we define the *boundary residuals* R_{D} and R_{N}, respectively, by

$$R_{\mathrm{D}}|_{\partial\kappa\cap(\Gamma_{\mathrm{D}}\cup\Gamma_-)} = (g_{\mathrm{D}} - u^+_{\mathrm{DG}})|_{\partial\kappa\cap(\Gamma_{\mathrm{D}}\cup\Gamma_-)} \,, \quad R_{\mathrm{N}}|_{\partial\kappa\cap\Gamma_{\mathrm{N}}} = (g_{\mathrm{N}} - (a\nabla u^+_{\mathrm{DG}}) \cdot \mathbf{n})|_{\partial\kappa\cap\Gamma_{\mathrm{N}}}.$$

On choosing $w = u - u_{\mathrm{DG}}$ in (23) and recalling the linearity of $J(\cdot)$ and the Galerkin orthogonality property (11), we deduce the following error representation formula

$$J(u) - J(u_{\mathrm{DG}}) = J(u - u_{\mathrm{DG}}) = B_{\mathrm{DG}}(u - u_{\mathrm{DG}}, z) = B_{\mathrm{DG}}(u - u_{\mathrm{DG}}, z - z_{h,p}) \tag{24}$$

for all $z_{h,p} \in S_{h,p}$, which equates the error in the computed target function $J(\cdot)$ in terms of quantities involving the product of the error $u - u_{\mathrm{DG}}$ in the underlying computed solution u_{DG} and the weighting term (projection error) $z - z_{h,p}$. Using the consistency of the DGFEM, upon application of the divergence theorem, we note that (24) may be expressed in the following form

$$J(u) - J(u_{\mathrm{DG}}) = \mathcal{E}_\Omega(u_{\mathrm{DG}}, z - z_{h,p}) \equiv \sum_{\kappa\in\mathcal{T}_h} \eta_\kappa \,,$$

where

$$\begin{aligned}
\eta_\kappa \;=\; & \int_\kappa R_{\mathrm{int}}(z - z_{h,p})\,\mathrm{d}x - \int_{\partial_-\kappa\cap\Gamma} (\mathbf{b}\cdot\mathbf{n}_\kappa)\, R_{\mathrm{D}}(z - z_{h,p})^+\,\mathrm{d}s \\
& + \int_{\partial_-\kappa\setminus\Gamma} (\mathbf{b}\cdot\mathbf{n}_\kappa)\,[u_{\mathrm{DG}}](z - z_{h,p})^+\,\mathrm{d}s - \int_{\partial\kappa\cap\Gamma_{\mathrm{D}}} R_{\mathrm{D}}((a\nabla(z - z_{h,p})^+)\cdot\mathbf{n}_\kappa)\,\mathrm{d}s \\
& + \int_{\partial\kappa\cap\Gamma_{\mathrm{D}}} \vartheta R_{\mathrm{D}}(z - z_{h,p})^+\,\mathrm{d}s + \int_{\partial\kappa\cap\Gamma_{\mathrm{N}}} R_{\mathrm{N}}(z - z_{h,p})^+\,\mathrm{d}s - \int_{\partial\kappa\setminus\Gamma} \vartheta[u_{\mathrm{DG}}](z - z_{h,p})^+\,\mathrm{d}s \\
& + \frac{1}{2}\int_{\partial\kappa\setminus\Gamma} \left\{\, [u_{\mathrm{DG}}](a\nabla(z - z_{h,p})^+)\cdot\mathbf{n}_\kappa - [(a\nabla u_{\mathrm{DG}})\cdot\mathbf{n}_\kappa](z - z_{h,p})^+ \right\}\mathrm{d}s
\end{aligned} \tag{25}$$

for all $z_{h,p} \in S_{h,p}$. Thereby, on application of the triangle inequality, we deduce the following Type I *a posteriori* error bound.

Proposition 5.1 *Let u and u_{DG} denote the solutions of (1), (3) and (6), respectively, and suppose that the dual solution z is defined by (23). Then, the following Type I* a posteriori *error bound holds:*

$$|J(u) - J(u_{\mathrm{DG}})| \le \mathcal{E}_{|\Omega|}(u_{\mathrm{DG}}, z - z_{h,p}) \equiv \sum_{\kappa\in\mathcal{T}_h} |\eta_\kappa| \,, \tag{26}$$

where η_κ is defined as in (25).

As discussed in [10, 14], the local weighting terms involving the difference between the dual solution z and its projection/interpolant $z_{h,p}$ onto $S_{h,p}$ appearing in the Type I bound (26) provide useful information concerning the global transport of the error. Thereby, we refrain from eliminating the weighting terms involving the (unknown) dual solution z and approximate z numerically; this will be discussed in Section 6.

5.2 *A priori* error bounds

In this section we derive an *a priori* error bound for the interior penalty DGFEM introduced in Section 2.2. To this end, let us now assume that the volume of the elements, denoted by m_κ for each $\kappa \in \mathcal{T}_h$, cf. above, has *bounded local variation*; i.e., there exists a constant $C_8 \geq 1$ such that, for any pair of elements κ and κ' which share a $(d-1)$–dimensional face,

$$C_8^{-1} \leq m_\kappa / m_{\kappa'} \leq C_8 \,. \tag{27}$$

With this hypothesis, we now proceed to prove the main result of this section.

Theorem 5.2 *Let $\Omega \subset \mathbb{R}^d$ be a bounded polyhedral domain, $\mathcal{T}_h = \{\kappa\}$ a subdivision of Ω into elements, constructed as in Section 2.1, such that the elemental volumes satisfy the bounded local variation condition (27). Then, assuming that the foregoing assumptions on the data hold, and $u \in H^k(\Omega, \mathcal{T}_h)$, $k \geq 2$, $z \in H^l(\Omega, \mathcal{T}_h)$, $l \geq 2$, then the solution $u_{\mathrm{DG}} \in S_{h,p}$ of (6) obeys the error bound*

$$\begin{aligned} |J(u) - J(u_{DG})|^2 \quad \leq \quad & C \left(\sum_{\kappa \in \mathcal{T}_h} \left\{ \frac{\alpha}{\sigma_{d,\kappa}^2} + \frac{\beta_2}{\sigma_{d,\kappa}} + (\beta_1 + \gamma_1) \right\} \int_{\tilde{\kappa}} D_{\tilde{\kappa}}^s(\tilde{u}, \Sigma_\kappa, U_\kappa) \, \mathrm{d}\tilde{x} \right) \\ & \times \left(\sum_{\kappa \in \mathcal{T}_h} \left\{ \frac{\alpha}{\sigma_{d,\kappa}^2} + \frac{\beta_2}{\sigma_{d,\kappa}} + (\beta_1 + \gamma_2) \right\} \int_{\tilde{\kappa}} D_{\tilde{\kappa}}^t(\tilde{z}, \Sigma_\kappa, U_\kappa) \, \mathrm{d}\tilde{x} \right), \end{aligned}$$

for $2 \leq s \leq \min(p+1, k)$ and $2 \leq t \leq \min(p+1, l)$, where $\alpha|_\kappa = \bar{a}_{\tilde{\kappa}}$, $\beta_1|_\kappa = \|c + \nabla \cdot \mathbf{b}\|_{L_\infty(\kappa)}$, $\beta_2|_\kappa = \|\mathbf{b}\|_{L_\infty(\kappa)}$, $\gamma_1|_\kappa = \|c/c_0\|_{L_\infty(\kappa)}^2$, $\gamma_2|_\kappa = \|(c + \nabla \cdot \mathbf{b})/c_0\|_{L_\infty(\kappa)}^2$, for all $\kappa \in \mathcal{T}_h$. Here, C is a positive constant depending on the dimension d, the polynomial degree p, and the parameters C_i, $i = 1, \ldots, 8$.

Proof. Decomposing the error $u - u_{\mathrm{DG}}$ as

$$u - u_{\mathrm{DG}} = (u - \Pi_p u) + (\Pi_p u - u_{\mathrm{DG}}) \equiv \eta + \xi \,,$$

where Π_p denotes the L_2–projection operator introduced in Section 4, we note that the error in the target functional $J(\cdot)$ may be expressed as follows:

$$J(u) - J(u_{\mathrm{DG}}) = B_{\mathrm{DG}}(\eta, z - z_{h,p}) + B_{\mathrm{DG}}(\xi, z - z_{h,p}). \tag{28}$$

Defining $z_{h,p} = \Pi_p z$, after a lengthy, but straightforward calculation, the two terms on the right-hand side of (28) may be separately estimated, in order to deduce the following bound:

$$\begin{aligned} & |J(u) - J(u_{DG})|^2 \\ & \leq C \left(\sum_{\kappa \in \mathcal{T}_h} \Big\{ \|\sqrt{a} \nabla \eta\|_{L_2(\kappa)}^2 + (\beta_1 + \gamma_1) \, \|\eta\|_{L_2(\kappa)}^2 + \beta_2 \sigma_{d,\kappa} \|\nabla \eta\|_{L_2(\kappa)}^2 + \|\eta^+\|_{\partial_+ \kappa \cap \Gamma}^2 \right. \\ & \qquad \left. + \|\eta^-\|_{\partial_- \kappa \setminus \Gamma}^2 + \|[\eta]\|_{\partial_- \kappa}^2 + \|\vartheta^{-1/2} \langle a \nabla \eta \rangle\|_{L_2(\partial\kappa \cap (\Gamma_{\mathrm{int}} \cup \Gamma_D))}^2 + \|\vartheta^{1/2} [\eta]\|_{L_2(\partial\kappa \cap (\Gamma_{\mathrm{int}} \cup \Gamma_D))}^2 \Big\} \right) \\ & \quad \times \left(\sum_{\kappa \in \mathcal{T}_h} \Big\{ \|\sqrt{a} \nabla w\|_{L_2(\kappa)}^2 + \Big(\beta_1 + \beta_2 \sigma_{d,\kappa}^{-1} + \gamma_2 \Big) \, \|w\|_{L_2(\kappa)}^2 + \|w^+\|_{\partial_- \kappa}^2 \right. \\ & \qquad \left. + \|\vartheta^{-1/2} \langle a \nabla w \rangle\|_{L_2(\partial\kappa \cap (\Gamma_{\mathrm{int}} \cup \Gamma_D))}^2 + \|\vartheta^{1/2} [w]\|_{L_2(\partial\kappa \cap (\Gamma_{\mathrm{int}} \cup \Gamma_D))}^2 \Big\} \right), \end{aligned}$$

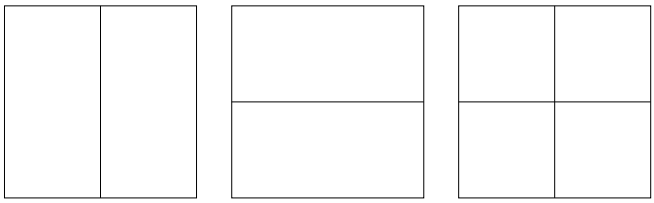

Figure 2: Cartesian refinement in 2D: (a) & (b) Anisotropic refinement; (c) Isotropic refinement.

where $w = z - z_{h,p}$. The statement of the theorem now follows by application of Theorem 4.6, the definition of the discontinuity-penalization parameter ϑ stated in (10), together with the bounded variation of the elemental volumes (27) and the scaling estimates stated in (21). $\square$

Remark 5.3 *The above result represents an extension of the* a priori *error bound derived in the article [9] to the case when general anisotropic computational meshes are employed. We note that although the analysis presented in [9] assumed shape–regular meshes, the explicit dependence of the polynomial degree was retained in the resulting* a priori *error bound; however, following the arguments in [9] an analogous hp–version bound of the form stated in Theorem 5.2 may easily be deduced.*

Remark 5.4 *The* a priori *bound stated in Theorem 5.2 clearly highlights that in order to minimize the error in the computed target functional $J(\cdot)$, the design of an optimal mesh must exploit anisotropic information emanating from both the primal and dual solutions u and z, respectively. Indeed, a mesh solely optimized for u may be completely inappropriate for z, and vice versa, thus there must me a trade-off between aligning the elements with respect to either solution in order to minimize the overall error in $J(\cdot)$.*

6 Adaptive algorithm

For a user-defined tolerance $\texttt{TOL}$, we now consider the problem of designing an appropriate finite element mesh $\mathcal{T}_h$ such that

$$|J(u) - J(u_{\mathrm{DG}})| \leq \texttt{TOL}\,,$$

subject to the constraint that the total number of elements in $\mathcal{T}_h$ is minimized; for simplicity of presentation, in this section we only consider the case when $\Omega \subset \mathbb{R}^2$. Following the discussion presented [13], we exploit the *a posteriori* error bound (26) with z replaced by a discontinuous Galerkin approximation $\hat{z}$ computed on the same mesh $\mathcal{T}_h$ used for the primal solution u_{DG}, but with a higher degree polynomial, i.e., $\hat{z} \in S_{h,\hat{p}}$, $\hat{p} = p + p_{\mathrm{inc}}$; in Section 7, we set $p_{\mathrm{inc}} = 1$, cf. [10, 14]. Thereby, in practice we enforce the stopping criterion

$$\hat{\mathcal{E}}_{|\Omega|} \equiv \mathcal{E}_{|\Omega|}(u_{\mathrm{DG}}, \hat{z} - z_{h,p}) \leq \texttt{TOL}\,. \tag{29}$$

If (29) is not satisfied, then the elements are marked for refinement/derefinement according to the size of the (approximate) error indicators $|\hat{\eta}_\kappa|$; these are defined analogously to $|\eta_\kappa|$ in (25) with z replaced by $\hat{z}$. In Section 7 we use the fixed fraction mesh refinement algorithm, with refinement and derefinement fractions set to 20% and 10%, respectively.

To subdivide the elements which have been flagged for refinement, we employ a simple Cartesian refinement strategy; here, elements may be subdivided either anisotropically or isotropically according to the three refinements (in two–dimensions, i.e., $d = 2$) depicted in Figure 2. In order to determine the optimal refinement, stimulated by the articles [19, 20], we propose the

following strategy based on choosing the most competitive subdivision of κ from a series of trial refinements, whereby an approximate local error indicator on each trial patch is determined. More precisely, given an element κ in the computational mesh $\mathcal{T}_h$ (which has been marked for refinement), we first construct the mesh patches $\mathcal{T}_{h,1}$ and $\mathcal{T}_{h,2}$ based on refining κ anisotropically according to Figures 2(a) & (b), respectively. On each mesh patch, $\mathcal{T}_{h,i}$, $i = 1, 2$, we compute the approximate error estimators

$$\hat{\mathcal{E}}_{\kappa,i}(u_{\mathrm{DG},i}, \hat{z}_i - z_{h,p}) = \sum_{\kappa' \in \mathcal{T}_{h,i}} \eta_{\kappa',i},$$

for $i = 1, 2$, respectively. Here, $u_{\mathrm{DG},i}$, $i = 1, 2$, is the discontinuous Galerkin approximation to (1), (3) computed on the mesh patch $\mathcal{T}_{h,i}$, $i = 1, 2$, respectively, based on enforcing appropriate boundary conditions on $\partial\kappa$ computed from the original discontinuous Galerkin solution u_{DG} on the portion of the boundary $\partial\kappa$ of κ which is interior to the computational domain Ω, i.e., where $\partial\kappa \cap \Gamma = \emptyset$. Similarly, $\hat{z}_i$ denotes the discontinuous Galerkin approximation to z computed on the local mesh patch $\mathcal{T}_{h,i}$, $i = 1, 2$, respectively, with polynomials of degree $\hat{p}$, based on employing suitable boundary conditions on $\partial\kappa \cap \Gamma = \emptyset$ derived from $\hat{z}$. Finally, $\eta_{\kappa',i}$, $i = 1, 2$, is defined in an analogous manner to η_κ, cf. (25) above, with u_{DG} and z replaced by $u_{\mathrm{DG},i}$ and $\hat{z}_i$, respectively.

Given an anisotropy parameter $\theta > 0$, isotropic refinement is selected when

$$\frac{\max_{i=1,2} |\hat{\mathcal{E}}_{\kappa,i}(u_{\mathrm{DG},i}, \hat{z}_i - z_{h,p})|}{\min_{i=1,2} |\hat{\mathcal{E}}_{\kappa,i}(u_{\mathrm{DG},i}, \hat{z}_i - z_{h,p})|} < \theta;$$

otherwise an anisotropic refinement is performed based on which refinement gives rise to the smallest predicted error indicator, i.e., the subdivision for which $|\hat{\mathcal{E}}_{\kappa,i}(u_{\mathrm{DG},i}, \hat{z}_i - z_{h,p})|$, $i = 1, 2$, is minimal. Based on computational experience, we select $\theta = 3$.

7 Numerical experiments

In this section we present a number of experiments to numerically to demonstrate the performance of the anisotropic adaptive algorithm outlined in Section 6.

7.1 Example 1

In this first example we consider a linear advection problem on the rectangular domain $\Omega = (0, 2) \times (0, 1)$, where $a = 0$, $c = 0$, $f = 0$, and

$$\mathbf{b} = \begin{cases} (y, 1 - x)^\top & \text{if } x < 1, \\ (2 - y, x - 1)^\top & \text{if } x \geq 1. \end{cases}$$

On the inflow boundary Γ_-, we select $u(x, y) = 1$ along $y = 0$, $1/8 < x < 3/4$ and $u(x, y) = 0$, elsewhere. Thereby, the analytical is piecewise constant, with discontinuities across the two characteristics emanating from the points $(x, y) = (1/8, 0)$ and $(x, y) = (3/4, 0)$ located on the inflow boundary. Here, we suppose that the aim of the computation is to calculate the value of the (weighted) outflow flux along $x = 2$, $0 \leq y \leq 1$, i.e.,

$$J(u) = \int_0^1 u(2, y)\psi(y)\,\mathrm{d}y.$$

Selecting the weight function ψ as follows:

$$\psi(y) = \mathrm{e}^{(3/8)^{-2} - ((y-5/8)^2 - 3/8)^{-2}},$$

the true value of the functional is given by $J(u) = 0.19280098502579391380$.

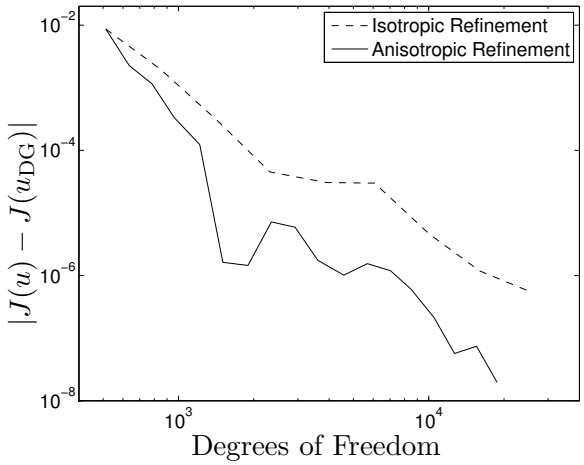

Figure 3: Example 1: Comparison between adaptive isotropic and anisotropic mesh refinement.

In Figure 3 we plot the error in the computed target functional $J(\cdot)$ with $p = 1$, i.e. bilinear elements, using both an isotropic (only) mesh refinement algorithm, together with the anisotropic refinement strategy outlined in Section 6. Here, we clearly observe the superiority of employing anisotropic mesh refinement in comparison with standard isotropic subdivision of the elements. Indeed, the error $|J(u) - J(u_{\mathrm{DG}})|$ computed on the series of anisotropically refined meshes is always less than the corresponding quantity computed on the isotropic grids. Here, we observe that there is an initial transient whereby the error in the computed target functional decays rapidly using the former refinement algorithm, in comparison with the latter, after which the gradient of the two convergence curves become very similar. This type of behavior is indeed expected, since for a fixed order method, i.e. h–version, we can only expect to improve the convergence of the error by a fixed constant, as the mesh is refined. Notwithstanding this, we note that the true error between $J(u)$ and $J(u_{\mathrm{DG}})$ using anisotropic refinement is around an order of magnitude smaller than the corresponding quantity when isotropic refinement is employed alone.

Finally, in Figure 4 we show the meshes generated using the proposed anisotropic mesh adaptation algorithm. Firstly, we note that the grid is primarily concentrated in the vicinity of the discontinuity of the analytical solution u which emanates from the point $(x, y) = (3/4, 0)$ on the inflow boundary; the second discontinuity in u is significantly less refined, as the resolution of this sharp feature in the solution is not important for the accurate computation of the selected target functional, cf. [10], for example. Secondly, we observe that the refinement algorithm has clearly identified the anisotropy in the underlying primal and dual solutions, and refined the mesh accordingly. Indeed, here we observe that in regions where the discontinuities in u are well aligned with the mesh lines of the original background mesh, anisotropic refinement has been employed; in other regions of the computational domain, isotropic refinement has been utilized.

7.2 Example 2

In this second example we consider Poisson's equation on the unit square $\Omega = (0, 1)^2$, where $a = I$, $\mathbf{b} = \mathbf{0}$, $c = 0$, and f is selected so that the analytical solution to (1), (3) is given by

$$u(x, y) = 4y(1 - y)(1 - \mathrm{e}^{-\alpha x} - (1 - \mathrm{e}^{-\alpha})x);$$

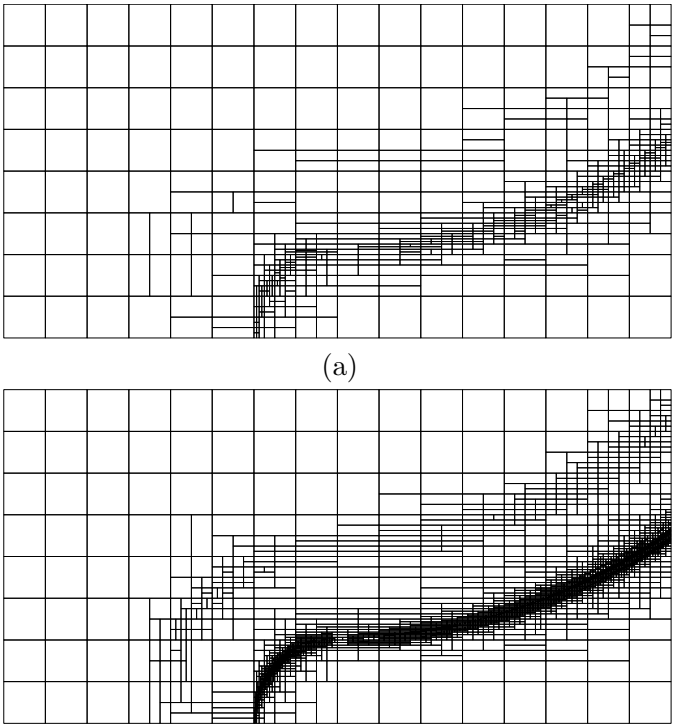

Figure 4: Example 1: Anisotropic meshes after: (a) 8 adaptive refinements, with 731 elements; (b) 15 adaptive refinements, with 3169 elements.

setting $\alpha = 100$ gives rise to a strong boundary layer along the boundary $x = 0$, $0 \leq y \leq 1$, cf. [4]. In this section, we select

$$J(u) = \int_0^1 \frac{\partial u}{\partial \mathbf{n}}(0, y)\psi(y)\,\mathrm{d}y, \qquad \psi = \mathrm{e}^{-10000(y-1/2)^4};$$

thereby, $J(u) = -17.704136538610340970$. For dual consistency we use the following consistent reformulation of $J(\cdot)$:

$$\tilde{J}(u) = \int_0^1 \left(\frac{\partial u}{\partial \mathbf{n}}(0, y) - \vartheta(u(0, y) - g_{\mathrm{D}}) \right) \psi(y)\,\mathrm{d}y,$$

cf. [8], for details.

To demonstrate the versatility of the proposed refinement algorithm, in this section we employ bi-quadratic elements, i.e., $p = 2$. To this end, in Figure 5 we plot the error in the computed target functional $J(\cdot)$ using both isotropic and anisotropic mesh refinement. As in the previous section, we clearly observe the superiority of employing anisotropic mesh refinement in comparison with standard isotropic subdivision of the elements. Indeed, the error $|J(u) - J(u_{\mathrm{DG}})|$ computed on the series of anisotropically refined meshes is always less than the corresponding quantity computed on isotropic grids. As before, we observe that there is an initial transient whereby the error in the computed target functional decays rapidly using the anisotropic refinement algorithm, in comparison with isotropic refinement, after which the gradient of the two convergence curves become very similar. Moreover, after this transient, the true error between $J(u)$ and $J(u_{\mathrm{DG}})$ using anisotropic refinement is over an order of magnitude smaller than the corresponding quantity when isotropic refinement is employed alone.

Finally, in Figure 6 we show the meshes generated using the proposed anisotropic mesh adaptation algorithm. Here, we observe that the boundary layer along $x = 0$, $0 \leq y \leq 1$,

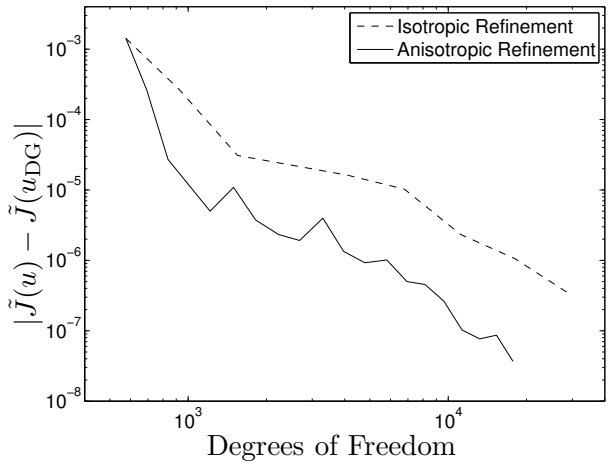

Figure 5: Example 2: Comparison between adaptive isotropic and anisotropic mesh refinement.

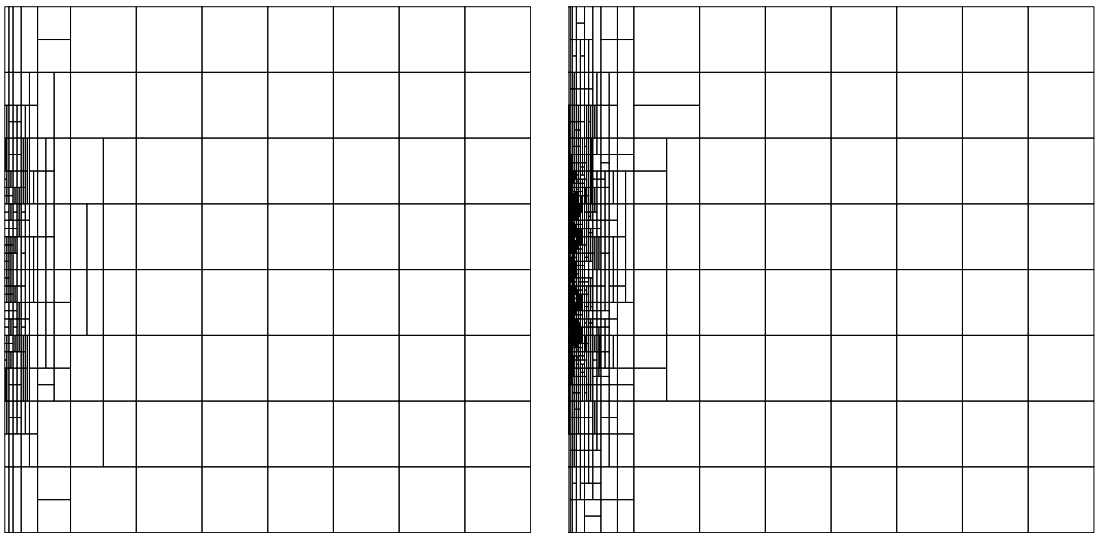

Figure 6: Example 2: Anisotropic meshes after: (a) 8 adaptive refinements, with 298 elements; (b) 15 adaptive refinements, with 1070 elements.

has been significantly refined, as we would expect, with the elements being mostly refined in the direction parallel to the boundary. We note, however, that some isotropic refinement is performed in the region of the boundary layer where the anisotropy of the dual solution z is perpendicular to Γ.

8 Concluding remarks

In this paper, we have been concerned with the *a priori* and *a posteriori* error analyses of the (symmetric) interior penalty discontinuous Galerkin finite element discretization of second–order partial differential equations with nonnegative characteristic form based on employing anisotropically refined computational meshes. We have been particularly interested in the approximation of linear output functionals of the analytical solution. To this end, new sharp directionally-sensitive bounds have been derived for the polynomial approximation on anisotropic elements exploiting the ideas presented in [4]. These bounds were utilized in the proceeding *a priori* error analysis of the approximation error in general linear target functionals of the solution on anisotropic meshes. Moreover, Type I (weighted) *a posteriori* error bounds have been derived and implemented within an adaptive mesh refinement algorithm based on employing a combination of local isotropic and anisotropic mesh refinement. The performance of the resulting refinement strategy was then studied through a series of numerical experiments. In particular, we demonstrated the superiority of the proposed algorithm in comparison with standard isotropic mesh refinement.

References

[1] T. Apel. *Anisotropic finite elements: Local estimates and applications.* Advances in Numerical Mathematics, Teubner, Stuttgart, 1999.

[2] R. Becker and R. Rannacher. Weighted a posteriori error control in FE methods. Technical report. Preprint 1, Interdisziplinäres Zentrum für Wissenschaftliches Rechnen, Universität Heidelberg, Heidelberg, Germany, 1996.

[3] W. Cao. On the error of linear interpolation and the orientation, aspect ratio, and internal angles of a triangle. *SIAM J. Numer. Anal.*, 43(1):19–40, 2005.

[4] L. Formaggia and S. Perotto. New anisotropic a priori error estimates. *Numer. Math.*, 89:641–667, 2001.

[5] E.H. Georgoulis. *hp*–version interior penalty discontinuous Galerkin finite element methods on anisotropic meshes. *Int. J. Numer. Anal. Model.*, 3:52–79, 2006.

[6] E.H. Georgoulis, E. Hall, and P. Houston. Discontinuous Galerkin methods for convection–diffusion–reaction problems on anisotropically refined meshes. In preparation.

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

[8] K. Harriman, D.J. Gavaghan, and E. Süli. The importance of adjoint consistency in the approximation of linear functionals using the discontinuous Galerkin finite element method. Technical Report NA04/18, Oxford University Computing Laboratory, 2004.

[9] K. Harriman, P. Houston, B. Senior, and E. Süli. *hp*–Version discontinuous Galerkin methods with interior penalty for partial differential equations with nonnegative characteristic form. In C.-W. Shu, T. Tang, and S.-Y. Cheng, editors, *Recent Advances in Scientific*

Computing and Partial Differential Equations. Contemporary Mathematics Vol. 330, pages 89–119. AMS, 2003.

[10] R. Hartmann and P. Houston. Adaptive discontinuous Galerkin finite element methods for nonlinear hyperbolic conservation laws. *SIAM J. Sci. Comput.*, 24:979–1004, 2002.

[11] P. Houston, C. Schwab, and E. Süli. Stabilized *hp*–finite element methods for first–order hyperbolic problems. *SIAM J. Numer. Anal.*, 37:1618–1643, 2000.

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

[13] P. Houston and E. Süli. *hp*–Adaptive discontinuous Galerkin finite element methods for hyperbolic problems. *SIAM J. Sci. Comput.*, 23:1225–1251, 2001.

[14] P. Houston and E. Süli. Adaptive finite element approximation of hyperbolic problems. In T. Barth and H. Deconinck, editors, *Error Estimation and Adaptive Discretization Methods in Computational Fluid Dynamics. Lect. Notes Comput. Sci. Engrg.*, volume 25, pages 269–344. Springer, 2002.

[15] W. Huang. Mathematical principles of anisotropic mesh adaptation. *Commun. Comput. Phys.*, 1(2):276–310, 2006.

[16] G. Kunert. *A posteriori error estimation for anisotropic tetrahedral and triangular finite element meshes*. PhD thesis, TU Chemnitz, 1999.

[17] L. De Lathauwer, B. De Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.*, 21:1253–1278, 2000.

[18] S. Prudhomme, F. Pascal, J.T. Oden, and A. Romkes. Review of a priori error estimation for discontinuous Galerkin methods. Technical report, TICAM Report 00–27, Texas Institute for Computational and Applied Mathematics, 2000.

[19] W. Rachowicz, L. Demkowicz, and J.T. Oden. Toward a universal *h–p* adaptive finite element strategy, Part 3. Design of *h–p* meshes. *Comput. Methods Appl. Mech. Engrg.*, 77:181–212, 1989.

[20] R. Schneider and P. Jimack. Toward anisotropic mesh adaptation based upon sensitivity of a posteriori estimates. Technical Report 2005.03, School of Computing, University of Leeds, 2005.