An a posteriori estimator of eigenvalue/eigenvector error for penalty-type discontinuous Galerkin methods

Stefano Giani[a], Luka Grubišić[b], Harri Hakula[c], Jeffrey S. Ovall[d]

[a] *Durham University, School of Engineering and Computing Sciences, South Road, Durham DH1 3LE, United Kingdom*
[b] *University of Zagreb, Department of Mathematics, Bijenička 30, 10000 Zagreb, Croatia*
[c] *Department of Mathematics and Systems Analysis, Aalto University, Finland*
[d] *Portland State University, Fariborz Maseeh Department of Mathematics and Statistics, 315 Neuberger Hall, Portland, OR 97201, USA*

Abstract

We provide an abstract framework for analyzing discretization error for eigenvalue problems discretized by discontinuous Galerkin methods such as the local discontinuous Galerkin method and symmetric interior penalty discontinuous Galerkin method. The analysis applies to clusters of eigenvalues that may include degenerate eigenvalues. We use asymptotic perturbation theory for linear operators to analyze the dependence of eigenvalues and eigenspaces on the penalty parameter. We first formulate the DG method in the framework of quadratic forms and construct a companion infinite dimensional eigenvalue problem. With the use of the companion problem, the eigenvalue/vector error is estimated as a sum of two components. The first component can be viewed as a "non-conformity" error that we argue can be neglected in practical estimates by properly choosing the penalty parameter. The second component is estimated a posteriori using auxiliary subspace techniques, and this constitutes the practical estimate.

Keywords: eigenvalue problem, finite element method, a posteriori error estimates, discontinuous Galerkin method
2000 MSC: Primary: 65N30, Secondary: 65N25, 65N15

1. Introduction

We present an a posteriori error analysis for penalty type Discontinuous Galerkin (DG) methods. All such numerical methods have in common the presence of a penalty term that ensures the stability of the methods. The function of the penalty term is to control the magnitude of the jumps of the discontinuous solution across the faces of the mesh [35, 37, 5]. An example of penalty term is the bilinear form $J(\cdot,\cdot)$ defined below. The strength of the penalization delivered by penalty terms can be adjusted using a parameter denoted in this work with τ. In general, there is not a unique way to choose the value of τ, but the analysis can prescribe how an appropriate value can be chosen to ensure stability; see, for example, Proposition 3.6, which shows that the penalty term τ has to be sufficiently large for the symmetric interior penalty method. When penalty terms are applied to

Email addresses: stefano.giani@durham.ac.uk (Stefano Giani), luka.grubisic@math.hr (Luka Grubišić), harri.hakula@aalto.fi (Harri Hakula), jovall@pdx.edu (Jeffrey S. Ovall)

faces of the mesh along the boundary of the domain, their function becomes to enforce boundary conditions weakly. This is natural in the context of DG methods, but this practice can be traced back to continuous Galerkin methods [36].

Motivated by the approach from [14], we introduce a notion of the companion discontinuous Galerkin forms. We present estimates for both multiple and clustered eigenvalues and also provide estimates for the associated invariant subspaces. In contrast to [14], our approach is based on the theory of the monotone convergence for quadratic forms from [22, 31]. This theory has been adapted to the application in numerical analysis in [16]. In the present paper we apply the abstract results from [16] to split the approximation error for the eigenvalues and spectral projections into the nonconformity estimate and the a posteriori computable error estimator.

Although our analysis is based on the abstract operator theory from [16] and our results directly include more general second order differential operators in the div-grad form, we will concentrate our presentation on the Laplace eigenvalue problem as a prototypical elliptic eigenvalue problem.

Let us make this claim more plausible. Assume we are given a positive-definite family of forms

$$\mathcal{B}_\tau(u,v) = B(u,v) + \tau J(u,v), \qquad u,v \in V \subset \mathcal{H} \ . \tag{1}$$

It is assumed that $\mathcal{B}_\tau$ are closed and densely defined in $\mathcal{H}$ and that the form J is positive semidefinite and bounded on V. An example of $\mathcal{B}_\tau$ which satisfies these requirements is immediately provided by the Symmetric Interior Penalty Discontinuous Galerkin (SIPDG) [5] and Local Discontinuous Galerkin (LDG) [23] discretizations of the Laplace operator in a bounded polygonal domain Ω.

To be precise, let $\mathcal{T}$ be a triangulation of Ω and let the functions $\underline{p}$ and h be the degree distribution function and the element diameter function, respectively. Here and in what follows, we will consider the piecewise polynomial space $S_{\underline{p}}(\mathcal{T})$, which is defined by requiring that each $u \in S_{\underline{p}}(\mathcal{T})$ is such that each restriction $u|_K$ is a polynomial of degree at most $\underline{p}(K)$ for each element $K \in \mathcal{T}$. We now define the discontinuous Galerkin space $V = S_{\underline{p}}(\mathcal{T}) + H^1_0(\Omega)$ and recall the notation $\{\!\{\cdot\}\!\}$ and $[\![\cdot]\!]$ for standard jump operators and the appropriate lifting operator $\mathcal{L}$ for the discrete gradients. We now introduce the bilinear forms

$$B(u,v) = \sum_{K\in\mathcal{T}} \int_K (\nabla u - \mathcal{L}([\![u]\!])) \cdot (\nabla v - \mathcal{L}([\![v]\!])) \ ,$$

$$J(u,v) = \sum_{F\in\mathcal{F}(\mathcal{T})} \frac{p_F^2}{h_F} \int_F [\![u]\!] \cdot [\![v]\!] \ .$$

It is a standard result on the local discontinuous Galerkin method that the forms B and J satisfy the requirements of (1), cf. [5, 23] and Section 3. We call this $\mathcal{B}_\tau$ the companion form of the Laplace eigenvalue problem

$$-\Delta u = \lambda u \text{ in } \Omega \quad , \quad u = 0 \text{ on } \partial\Omega \ . \tag{2}$$

This is further justified if we observe that the variational formulation of (2)

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \lambda(u,v) \qquad \forall\, v \in H^1_0(\Omega) \ , \tag{3}$$

where $(\cdot,\cdot)$ is the L^2 inner-product and $\|\cdot\|$ the L^2 norm, can be obtained by restricting the form $\mathcal{B}_\tau$ to the subspace $H^1_0(\Omega) \subset V$. On the other hand the discrete eigenvalue problem

$$\sum_{K\in\mathcal{T}} \int_K (\nabla u - \mathcal{L}([\![u]\!])) \cdot (\nabla v - \mathcal{L}([\![v]\!])) + \tau \sum_{F\in\mathcal{F}(\mathcal{T})} \frac{p_F^2}{h_F} \int_F [\![u]\!] \cdot [\![v]\!] = \hat{\lambda}_\tau(u,v), \ \ \forall\, v \in S_{\underline{p}}(\mathcal{T}) \tag{4}$$

is obtained by restricting the form $\mathcal{B}_\tau$ to $S_{\underline{p}}(\mathcal{T}) \subset V$. This construction provides a framework for the simultaneous abstract analysis of both the continuous as well as the discrete eigenvalue problems. We first establish an estimate of the error in the projection of the companion form from V to $H_0^1(\Omega)$ and call this the nonconformity error and denote it by $\mathcal{R}_{nc}$. We then proceed and estimate the error in the projection of V to $S_{\underline{p}}(\mathcal{T})$ and call this the a posteriori error and denote it by $\mathcal{R}_{ap}(\tau)$. For the a posteriori error component we will also present a computable error estimator using the technique of hierarchical bases. Finally, both estimators are combined in to provide an estimate of the approximation error. In particular,

$$\sum_{i=1}^{m} \frac{|\hat{\lambda}_{\tau,i} - \lambda_i|}{\lambda_i} \leq \frac{C_1}{\tau - 1}\mathcal{R}_{nc} + C_2\mathcal{R}_{ap}(\tau)$$

is an example of a type of estimates—for the cluster of lowermost eigenvalues $\lambda_1 \leq \cdots \leq \lambda_m < \lambda_{m+1}$ of (3)—which we will present in Section 4. Further, we will show that by an appropriate choice of τ we may make the a posteriori error the dominant part of the error.

In Section 2 we present main abstract results in the context of discontinuous Galerkin methods and introduce the notation. In Section 3 we review basic facts on discontinuous penalty type Galerkin methods and establish results which guarantee that the concrete formulations satisfy the requirements of the abstract theory. In Section 4 we construct a computable a posteriori error estimator and establish its basic properties. Extensive numerical results will be presented in the subsequent publication.

2. Abstract variational source and eigenvalue problems

Given an open, bounded domain $\Omega \subset \mathbb{R}^d$, let $V = H_0^1(\Omega) + S$, where $S \subset L^2(\Omega)$ is finite dimensional, $S \not\subset H_0^1(\Omega)$ and $H_0^1(\Omega) \cap S \neq \{0\}$. Let $B, J : V \times V \to \mathbb{R}$ be symmetric bilinear forms on V satisfying:

1. For all $u, v \in H_0^1(\Omega)$, $B(u, v) = \int_\Omega \nabla u \cdot \nabla v \, dx$.
2. $J(v, v) \geq 0$ for all $v \in V$, and $\mathrm{Ker}(J) = \{v \in V : \ J(v, v) = 0\} = H_0^1(\Omega)$.
3. If $v \in V$ and $B(v, v) + J(v, v) = 0$ then $v = 0$.

For $\tau \geq 1$, we define

$$\mathcal{B}(u, v) = \mathcal{B}_\tau(u, v) = B(u, v) + \tau J(u, v) \ .$$

In later sections, S will be a space of discontinuous, piecewise polynomials defined on a simple partition of the domain Ω, and J will be used to penalize discontinuities of functions in V, as is typical in penalty-type discontinuous Galerkin (DG) discretizations [5].

We see that $\mathcal{B}$ is an inner-product on V when $\tau \geq 1$, and we denote its induced norm by $\mathcal{B}$ as $\|\cdot\|_{\mathcal{B}}$. We further decompose S as

$$S = (H_0^1(\Omega) \cap S) \oplus_{\mathcal{B}} R \ ,$$

so R is the $\mathcal{B}$-orthogonal complement of $H_0^1(\Omega) \cap S$ in S. We now have $V = H_0^1(\Omega) \oplus R$; here we are not asserting any sort of orthogonality between $H_0^1(\Omega)$ and R, just that $H_0^1(\Omega) \cap R = \{0\}$.

Suppose that $\{v_n\}$, where $v_n = w_n + r_n$ with $w_n \in H_0^1(\Omega)$ and $r_n \in R$, is a Cauchy sequence with respect to $\|\cdot\|_{\mathcal{B}}$. Because $H_0^1(\Omega)$ is closed with respect to $\|\cdot\|_{\mathcal{B}}$, R is finite dimensional, and

$H_0^1(\Omega) \cap R = \{0\}$, there is a constant $\gamma \in [0,1)$ for which $\mathcal{B}(w,r) \le \gamma \|w\|_{\mathcal{B}} \|r\|_{\mathcal{B}}$ for all $w \in H_0^1(\Omega)$ and $r \in R$. This strong Cauchy inequality (cf. [13, 28]), together with Young's inequality, implies that

$$\begin{aligned}\|v_n - v_m\|_{\mathcal{B}}^2 &= \|w_n - w_m\|_{\mathcal{B}}^2 + \|r_n - r_m\|_{\mathcal{B}}^2 - 2\mathcal{B}(w_n - w_m, r_n - r_m) \\ &\ge \|w_n - w_m\|_{\mathcal{B}}^2 + \|r_n - r_m\|_{\mathcal{B}}^2 - 2\gamma \|w_n - w_m\|_{\mathcal{B}} \|r_n - r_m\|_{\mathcal{B}} \\ &\ge (1-\gamma^2) \max(\|w_n - w_m\|_{\mathcal{B}}^2 \,,\, \|r_n - r_m\|_{\mathcal{B}}^2) \,.\end{aligned}$$

So we see that both $\{w_n\}$ and $\{r_n\}$ are Cauchy sequences with respect to $\|\cdot\|_{\mathcal{B}}$. But $H_0^1(\Omega)$ and R are Banach spaces in this norm, so $w_n \to w \in H_0^1(\Omega)$ and $r_n \to r \in R$ in $\|\cdot\|_{\mathcal{B}}$. Therefore, we deduce that V is closed with respect to $\|\cdot\|_{\mathcal{B}}$. Therefore, $\mathcal{B}$ is a closed form on V.

Remark 2.1. The strong Cauchy inequality is an assertion that there is a positive angle between the subspaces $H_0^1(\Omega)$ and R with respect to the inner-product $\mathcal{B}$. This is equivalent to asserting that the oblique projector that maps a vector in V to its component in R is a bounded operator with respect to $\|\cdot\|_{\mathcal{B}}$.

Let $Q \subset V$ be any finite dimensional subspace of V. We consider the following source problems: given $f \in L^2(\Omega)$, find $u(f) \in H_0^1(\Omega)$, $u_\tau(f) \in V$ and $q(f) \in Q$ such that

$$\mathcal{B}(u(f), v) = \mathcal{B}_\tau(u(f), v) = (f, v) \text{ for all } v \in H_0^1(\Omega) \,, \tag{5}$$

$$\mathcal{B}_\tau(u_\tau(f), v) = (f, v) \text{ for all } v \in V \,, \tag{6}$$

$$\mathcal{B}_\tau(q_\tau(f), v) = (f, v) \text{ for all } v \in Q \,. \tag{7}$$

As a general notational convention we will use capital letters to denote subspaces of V and lowercase letters to denote the Galerkin projections of the source problem associated to that subspace. In particular, we will use $q_\tau(f)$ or $\hat{q}_\tau(f)$ or $\hat{q}_{m,\tau}(f)$ to denote Galerkin projections from (7) for generic finite dimensional subspaces Q, $\hat{Q}$ or $\hat{Q}_m$. We will freely modify the notation for any particular finite dimensional subspace following this general rule where appropriate. The first usage of this type of notation is in (16).

Here and below, we use (v, w) to denote the $L^2(\Omega)$ inner-product of $v, w \in L^2(\Omega)$. We also use $|\!|\!| \cdot |\!|\!|_\tau$ to indicate $\|\cdot\|_{\mathcal{B}_\tau}$, in order to emphasize the τ-dependence of the norm, noting that

$$H_0^1(\Omega) = \{v \in V \;:\; \lim_{\tau \to \infty} |\!|\!| v |\!|\!|_\tau < \infty\} \,. \tag{8}$$

The following result appeared as [16, Theorem 4.3].

Theorem 2.2. *For $f \in \mathrm{Ker}(J)$ and $\tau > 1$,*

$$\frac{|\!|\!| u(f) - u_2(f) |\!|\!|_1^2}{\tau - 1} \le |\!|\!| u(f) - u_\tau(f) |\!|\!|_\tau^2 \le \frac{|\!|\!| u(f) - u_1(f) |\!|\!|_1^2}{\mathfrak{j}_{LBB}(\tau - 1)} \,,$$

where $\mathfrak{j}_{LBB} > 0$ is the inf-sup (Ladyzhenskaya-Babuška-Brezzi) constant,

$$\mathfrak{j}_{LBB} = \inf_{q \in V/\,\mathrm{Ker}(J)} \sup_{\psi \in V} \frac{|J(q,\psi)|}{|\!|\!| q |\!|\!|_1 \, |\!|\!| \psi |\!|\!|_1} \,. \tag{9}$$

The associated eigenvalue problems are: Find $(\lambda, \psi) \in \mathbb{R} \times (H_0^1(\Omega) \setminus \{0\})$ and $(\lambda_\tau, \psi_\tau) \in \mathbb{R} \times (V \setminus \{0\})$ such that

$$B(\psi, v) = \lambda(\psi, v) \text{ for all } v \in H_0^1(\Omega) \ , \tag{10}$$

$$\mathcal{B}_\tau(\psi_\tau, v) = \lambda_\tau(\psi_\tau, v) \text{ for all } v \in V \ . \tag{11}$$

These variational eigenvalue problems are attained by discrete sequences of eigenvalues, $0 < \lambda_1 < \lambda_2 \leq \lambda_3 \leq \cdots$ and $0 < \lambda_{\tau,1} \leq \lambda_{\tau,2} \leq \lambda_{\tau,3} \leq \cdots$. One may choose an associated $L^2(\Omega)$-orthonormal Riesz basis of eigenvectors in both cases:

$$B(\psi_j, v) = \lambda_j(\psi_j, v) \text{ for all } v \in H_0^1(\Omega) \ , \ (\psi_i, \psi_j) = \delta_{ij} \ , \ \overline{\text{span}\{\psi_j\}} = L^2(\Omega) \ . \tag{12}$$

$$\mathcal{B}_\tau(\psi_{\tau,j}, v) = \lambda_{\tau,j}(\psi_{\tau,j}, v) \text{ for all } v \in V \ , \ (\psi_{\tau,i}, \psi_{\tau,j}) = \delta_{ij} \ , \ \overline{\text{span}\{\psi_{\tau,j}\}} = L^2(\Omega) \ . \tag{13}$$

For a bounded interval I, we denote by $E(I)$ the $L^2(\Omega)$-orthogonal projector onto the B-invariant subspace associated with the eigenvalues of B in I; $E_\tau(I)$ is the natural analogue for $\mathcal{B}_\tau$. It follows from Kato's monotone convergence theorem [22] that

$$\lim_{\tau \to \infty} \|E(I) - E_\tau(I)\| = 0 \ ,$$

where $\|\cdot\|$ is the operator norm on the space of bounded operators on $L^2(\Omega)$. Furthermore, we have monotone convergence of eigenvalues, including multiplicities. In other words, for a given m such that $\lambda_m < \lambda_{m+1}$ we have

$$\lambda_{\tau,j} \leq \lambda_{\tau',j} \leq \lambda_j \text{ for } \tau' \geq \tau \ , \text{ and } \lim_{\tau \to \infty} \lambda_{\tau,j} = \lambda_j \text{ for } 1 \leq j \leq m \ .$$

Following [38], for a bounded linear operator A on $L^2(\Omega)$ and a complete orthonormal system $\{e_k : k \in \mathbb{N}\}$ in $L^2(\Omega)$, we define

$$\|A\|_{HS}^2 = \sum_{k=1}^{\infty} \|Ae_k\|^2 \ , \tag{14}$$

where $\|v\|$ is the standard norm on $L^2(\Omega)$. If this quantity is finite, A is called a Hilbert-Schmidt operator, and $\|A\|_{HS}$ is its Hilbert-Schmidt norm. It is shown in [38, Lemma 6.58] that $\|A\|_{HS}$ is independent of the particular choice of complete orthonormal system, and that

$$\|A\| \leq \|A\|_{HS} \ , \tag{15}$$

where $\|A\|$ is the operator norm for bounded linear operators on $L^2(\Omega)$, as before.

Given an L^2-orthonormal set of vectors $\{\hat{\psi}_1, \cdots, \hat{\psi}_m\} \subset V$, whose span we denote by $\hat{S}_m$, we define the corresponding *approximation defects* by

$$\eta_{\tau,j}(\hat{S}_m) = \max_{\substack{\mathcal{S} \subset \hat{S}_m \\ \dim \mathcal{S} = m-j-1}} \min_{f \in \mathcal{S}} \frac{|\!|\!| u_\tau(f) - \hat{s}_{m,\tau}(f) |\!|\!|_\tau}{|\!|\!| u_\tau(f) |\!|\!|_\tau} \quad , \quad 1 \leq j \leq m \ . \tag{16}$$

Here we have tacitly assumed that $\hat{u}_\tau(f)$ is defined by (7) for the subspace $Q = \hat{S}_m$. In what follows we set $I = [0, D]$, where $\lambda_{\tau,m} < D < \lambda_{\tau,m+1}$ for all $\tau > 1$. With this we recall the main error estimates from [7, 17, 16].

Theorem 2.3. *Suppose $m \in \mathbb{N}$ is such that $\lambda_{\tau,m} < \lambda_{\tau,m+1}$, and let $\hat{\lambda}_{\tau,j} = \|\hat{\psi}_j\|_\tau^2$ denote the Ritz values associated with $\hat{S}_m$. If $\hat{S}_m$ is such that*

$$\frac{\eta_{\tau,m}(\hat{S}_m)}{1-\eta_{\tau,m}(\hat{S}_m)} < \frac{\lambda_{\tau,m+1}-\hat{\lambda}_{\tau,m}}{\lambda_{\tau,m+1}+\hat{\lambda}_{\tau,m}} ,$$

then

$$\begin{aligned} \frac{\hat{\lambda}_{\tau,1}}{4\hat{\lambda}_{\tau,m}} \sum_{i=1}^{m} \frac{\|u_\tau(\hat{\psi}_i) - \hat{\lambda}_{\tau,i}^{-1}\hat{\psi}_i\|_\tau^2}{\|u_\tau(\hat{\psi}_i)\|_\tau^2} \leq \sum_{i=1}^{m} \frac{|\lambda_{\tau,i}-\hat{\lambda}_{\tau,i}|}{\hat{\lambda}_{\tau,i}} \leq C_{\tau,m} \sum_{i=1}^{m} \frac{\|u_\tau(\hat{\psi}_i) - \hat{s}_{m,\tau}(\hat{\psi}_i)\|_\tau^2}{\|u_\tau(\hat{\psi}_i)\|_\tau^2} \\ \|E_\tau(I) - \hat{E}_m\|_{HS} \leq C_{\tau,m} \sqrt{\sum_{i=1}^{m} \frac{\|u_\tau(\hat{\psi}_i) - \hat{s}_{m,\tau}(\hat{\psi}_i)\|_\tau^2}{\|u_\tau(\hat{\psi}_i)\|_\tau^2}} . \end{aligned} \tag{17}$$

The constant $C_{\tau,m}$ depends solely on the reciprocal relative distance to the unwanted component of the spectrum (e.g. $\frac{\lambda_{\tau,m+1}+\lambda_{\tau,m}}{\lambda_{\tau,m+1}-\lambda_{\tau,m}}$), $\hat{E}_m$ is the L^2-orthogonal projection onto $\hat{S}_m$.

Remark 2.4. Although we have emphasized the τ-dependence of $C_{\tau,m}$ by including it in the subscript, this quantity can be bounded independently of τ when τ is sufficiently large. This is due to the monotone convergence of the τ-dependent eigenvalues $\lambda_{\tau,j}$ to the actual eigenvalues λ_j.

Following [16], we now formulate a result that combines Theorems 2.2 and 2.3 for the particular choice of the subspace $\hat{S}_m$. Namely, we choose that $\hat{S}_m$ is the eigenspace of B associated to the eigenvalues $\lambda_1 \leq \cdots \leq \lambda_m$.

Proposition 2.5. *Let m be such that $\lambda_m < \lambda_{m+1}$, then we have*

$$\frac{c_m^{\text{low}}}{\tau-1} \sum_{i=1}^{m} \|u_2(\psi_i) - \lambda_i^{-1}\psi_i\|_1^2 \leq \sum_{i=1}^{m} \frac{\lambda_i - \lambda_{\tau,i}}{\lambda_i} \leq \frac{C_m^{\text{high}}}{\tau-1} \sum_{i=1}^{m} \|u_1(\psi_i) - \lambda_i^{-1}\psi_i\|_1^2 \tag{18}$$

$$\|E(I) - E_\tau(I)\|_{HS} \leq \sqrt{\frac{C_m^{\text{high}}}{\tau-1} \sum_{i=1}^{m} \|u_1(\psi_i) - \lambda_i^{-1}\psi_i\|_1^2}. \tag{19}$$

Proof. Note that $J(\psi_i, \psi_i) = 0$ for each $i = 1, \cdots, m$ and that $\hat{u}_\tau(\psi_i) = u(\psi_i) = \lambda_i^{-1}\psi_i$. Here we have used $\hat{u}_\tau(f)$ as defined by (7) for the subspace $Q = S_m$, the eigenspace associated to the m lowermost eigenvalues of B. The results now follow directly from Theorems 2.2 and 2.3. The constant c_m^{low} depends on the quotient λ_1/λ_m, whereas the constant C_m^{high} depends on $\mathfrak{j}_{LBB}$ and the reciprocal of the relative spectral gap measure $\frac{\lambda_{m+1}-\lambda_m}{\lambda_{m+1}+\lambda_m}$. □

Remark 2.6. Note that, in (17) and (19), we can take either the operator norm or the Hilbert-Schmidt norm to measure the difference between orthogonal projections. The inequality for the operator norm is an overestimate in comparison to the Hilbert-Schmidt norm version of the result as can be seen in (15). Optimal estimates (independent of the size of the cluster of eigenvalues) of the operator norm of the projection difference can be obtained by solving a small optimization problem as has been done in [17, 7]. An alternative approach to obtaining estimates for approximating clustered eigenvalues can be found in [10]. There the authors use a different measure of the error and obtain estimates that are independent of the size of the cluster.

3. LDG and SIPDG within the abstract framework

We now introduce the notation necessary for defining discontinuous Galerkin methods, and discussing various lifting operators that will be used in what follows. We assume that the computational domain Ω can be partitioned into a shape-regular mesh $\mathcal{T}$, *i.e.* there exists a constant $C_{\rm reg}$ such that for any element K

$$h_K \leq C_{\rm reg}\ \rho_K\ , \tag{20}$$

where h_K is the diameter of the element K, and ρ_K is the diameter of the largest inscribed ball in K. For the analysis we will assume that the elements are affine quadrilaterals ($d = 2$) or hexahedra ($d = 3$). Associated with each K is an affine bijection $T_K : \widehat{K} \to K$ from the reference element $\widehat{K} = [0,1]^d$. This map induces a bijection between the faces F of K and the faces $\widehat{F}$ of $\widehat{K}$. We allow one-irregular meshes (cf. [26]).

The element diameters are collected in the mesh-size vector $\underline{h} = \{\, h_K \,:\, K \in \mathcal{T} \,\}$, and h denotes the maximum of all h_K in the mesh. We refer to F as an interior mesh face of $\mathcal{T}$ if $F = \partial K \cap \partial K'$ for two neighbouring elements $K, K' \in \mathcal{T}$ whose intersection has a positive surface measure. The set of all interior mesh faces is denoted by $\mathcal{F}_I(\mathcal{T})$. Analogously, if the intersection $F = \partial K \cap \partial\Omega$ of the boundary of an element $K \in \mathcal{T}$ and $\partial\Omega$ is of positive surface measure, we refer to F as a boundary mesh face of $\mathcal{T}$. The set of all boundary mesh faces of $\mathcal{T}$ is denoted by $\mathcal{F}_B(\mathcal{T})$ and we set $\mathcal{F}(\mathcal{T}) = \mathcal{F}_I(\mathcal{T}) \cup \mathcal{F}_B(\mathcal{T})$. The diameter of a face F is denoted by h_F.

Let an interior face $F \in \mathcal{F}_I(\mathcal{T})$ be shared by two neighbouring elements K and K'. We define the average and jump associated with F of a scalar-valued piecewise smooth function v, by

$$\{\!\{v\}\!\} = \frac{1}{2}(v|_F + v'|_F), \qquad [\![v]\!] = v|_F\ \underline{n}_K + v'|_F\ \underline{n}_{K'}\ ,$$

where $v|_F$ and $v'|_F$ are the traces of v on F taken from K and K', and $\underline{n}_K$ and $\underline{n}_{K'}$ denote the unit outward normal vectors on the boundary of elements K and K', respectively. Similarly, if $\underline{q}$ is piecewise smooth vector field, its average and (normal) jump across F are given by

$$\{\!\{\underline{q}\}\!\} = \frac{1}{2}\left(\underline{q}|_F + \underline{q}'|_F\right), \qquad [\![\underline{q}]\!] = \underline{q}|_F \cdot \underline{n}_K + \underline{q}'|_F \cdot \underline{n}_{K'}.$$

On a boundary face $F \in \mathcal{F}_B(\mathcal{T})$, we accordingly set $\{\!\{\underline{q}\}\!\} = \underline{q}$ and $[\![v]\!] = v\underline{n}$, with $\underline{n}$ denoting the unit outward normal vector on $\partial\Omega$. The other trace operators will not be used on boundary faces and are thereby left undefined. We note that the jump operator $[\![\cdot]\!]$ changes scalars to vectors and vectors to scalars.

Given an element $K \in \mathcal{T}$ and an integer $p \geq 0$, we define the local polynomial space

$$\mathcal{Q}_p(K) = \{\, v \,:\, K \to \mathbb{R} \,:\, v \circ T_K \in \mathcal{Q}_p(\widehat{K}) \,\}, \tag{21}$$

with $\mathcal{Q}_p(\widehat{K})$ denoting the set of tensor product polynomials on the reference element $\widehat{K}$ of degree less than or equal to p in each coordinate direction on $\widehat{K}$. In addition, if $F \in \mathcal{F}(K)$ is a face of K and $\widehat{F}$ the corresponding face on the reference element $\widehat{K}$, we define

$$\mathcal{Q}_p(F) = \{\, v \,:\, F \to \mathbb{R} \,:\, v \circ T_K|_F \in \mathcal{Q}_p(\widehat{F}) \,\}, \tag{22}$$

where $\mathcal{Q}_p(\widehat{F})$ denotes the set of tensor product polynomials on $\widehat{F}$ of degree less than or equal to p in each coordinate direction on $\widehat{F}$. Now, given a polynomial degree vector $\underline{p}$ on $\mathcal{T}$, we define the corresponding *hp*-DG finite element space by

$$S_{\underline{p}}(\mathcal{T}) = \{\, v \in L^2(\Omega) \,:\, v|_K \in \mathcal{Q}_{p_K}(K),\ K \in \mathcal{T} \,\}. \tag{23}$$

We take p to be the minimum of all p_K in the mesh. If F is a boundary face with adjacent element K, we take $p_F = p_K$. Otherwise, we take $p_F = \max\{p_K, p_{K'}\}$, where K, K' are the two elements sharing the face F.

3.1. Lifting operators

Let Γ denote the union of the boundaries of the elements K in $\mathcal{T}$, which we refer to as the mesh skeleton. We let $T(\Gamma) := \Pi_{K\in\mathcal{T}} L^2(\partial K)$ be the product space of functions that are double-valued on $\Gamma^0 := \Gamma \setminus \partial\Omega$ and single-valued on $\partial\Omega$. The space $L^2(\Gamma)$ is defined as the subspace of $T(\Gamma)$ consisting of functions for which the values on the joint faces between two adjacent elements coincide. This space is endowed with the product norm for $T(\Gamma)$, and $L^2(\Gamma^0)$ is taken to be its restriction on Γ^0. Given these function spaces defined on the skeleton Γ, we further define lifting operators to the space $[S_{\underline{p}}(\mathcal{T})]^d$, following [5]:

Definition 3.1 (Lifting operators)**.** We define four lifting operators,

$$r, r_F, \mathcal{L} : [L^2(\Gamma)]^d \to [S_{\underline{p}}(\mathcal{T})]^d \quad , \quad l : L^2(\Gamma^0) \to [S_{\underline{p}}(\mathcal{T})]^d \ ,$$

given by

$$\int_\Omega r(\varphi)\cdot\tau\, dx = -\int_\Gamma \varphi\cdot\{\!\{\tau\}\!\}\, ds, \quad \forall\tau\in[S_{\underline{p}}(\mathcal{T})]^d \ , \tag{24}$$

$$\int_\Omega r_F(\varphi)\cdot\tau\, dx = -\int_F \varphi\cdot\{\!\{\tau\}\!\}\, ds, \quad \forall\tau\in[S_{\underline{p}}(\mathcal{T})]^d \ \text{ for each face } F \ , \tag{25}$$

$$\int_\Omega l(q)\cdot\tau\, dx = -\int_{\Gamma^0} q[\![\tau]\!]\, ds, \quad \forall\tau\in[S_{\underline{p}}(\mathcal{T})]^d \ , \tag{26}$$

$$\mathcal{L}(\varphi) = -r(\varphi) + l(\underline{\beta}\cdot\varphi) \ , \tag{27}$$

where $\underline{\beta}\in[L^2(\Gamma^0)]^d$ is a vector-valued function that is constant on each face. Definition (27) comes from [23], which is slightly different from the one in [5] used to define the LDG [23] method. A final lifting operator $\mathcal{R} : V \to [S_{\underline{p}}(\mathcal{T})]^d$ that is useful in analysing the SIPDG [5] method is given by $\mathcal{R}(v) = -r([\![v]\!])$, so

$$\int_\Omega \mathcal{R}(v)\cdot\tau\, dx = \sum_{F\in\mathcal{F}(\mathcal{T})}\int_F [\![v]\!]\cdot\{\!\{\tau\}\!\}\, ds \ , \quad \forall\tau\in[S_{\underline{p}}(\mathcal{T})]^d \ . \tag{28}$$

This $\mathcal{R}$ differs by sign from what is given in [5].

Lemma 3.2. *For any function $u\in H_0^1(\Omega)$, any face $F\in\mathcal{F}(\mathcal{T})$ and any piecewise constant vector $\underline{\beta}$ defined on the skeleton, it holds that*

$$[\![u]\!] = r_F([\![u]\!]) = r([\![u]\!]) = \mathcal{R}(u) = l(\underline{\beta}\cdot[\![u]\!]) = \mathcal{L}([\![u]\!]) = 0 \ .$$

Proof. For $u\in H_0^1(\Omega)$, we clearly have $[\![u]\!] = 0 \in [L^2(\Gamma)]^d$, so the fact that $r(u)$ and $r_F(u)$ are both defined in terms the L^2-projection on $[S_{\underline{p}}(\mathcal{T})]^d$ with zero right-hand side guarantees that $r([\![u]\!]) = r_F([\![u]\!]) = \mathcal{R}(u) = 0$. For any piecewise-constant vector $\underline{\beta}$ defined on the skeleton, we have $\underline{\beta}\cdot[\![u]\!] = 0$ on Γ^0, so it follows that $l(\underline{\beta}\cdot[\![u]\!]) = 0$ and $\mathcal{L}([\![u]\!]) = 0$ as well. □

3.2. Local discontinuous Galerkin method

The Local DG (LDG) method (cf. [5, 23]) can be formulated in terms of the following symmetric bilinear forms on $V = S_{\underline{p}}(\mathcal{T}) + H_0^1(\Omega)$,

$$B(u,v) = \sum_{K\in\mathcal{T}} \int_K (\nabla u - \mathcal{L}([\![u]\!])) \cdot (\nabla v - \mathcal{L}([\![v]\!])) \ , \tag{29}$$

$$J(u,v) = \sum_{F\in\mathcal{F}(\mathcal{T})} \frac{p_F^2}{h_F} \int_F [\![u]\!] \cdot [\![v]\!] \ . \tag{30}$$

We take $\mathcal{B}_\tau = B + \tau J$ as in our abstract formulation. Combining Lemma 3.2 and a Poincaré inequality, we see that condition (1) from Section 2 is satisfied. It is also clear that condition (2) holds, so it remains to verify condition (3), and that $B(v,v) < \infty$ and $J(v,v) < \infty$ for $v \in V$. To do so, we will use the energy norm defined as

$$\|u\|_{\mathrm{E},\tau}^2 = \sum_{K\in\mathcal{T}} \|\nabla u\|_{L^2(K)}^2 + \tau \sum_{F\in\mathcal{F}(\mathcal{T})} \frac{p_F^2}{h_F} \|[\![u]\!]\|_{L^2(F)}^2 \ , \tag{31}$$

on V. It is shown in [23, Proposition 3.1] that, for any $\tau > 0$,

$$\mathcal{B}_\tau(w,v) \le C_{\mathrm{cont}} \|w\|_{\mathrm{E},\tau} \|v\|_{\mathrm{E},\tau} \ \forall w, v \in V \ , \tag{32}$$

where $C_{\mathrm{cont}} > 0$ depends only on τ, $\delta = \| \underline{\beta} \|_{L^\infty(\Gamma^0)}$ and the shape-regularity of the mesh. Now suppose that $\mathcal{B}_\tau(v,v) = 0$. This implies that $J(v,v) = 0$, so $v \in H_0^1(\Omega)$, and we have $B_\tau(v,v) = \|\nabla v\|_{L^2(\Omega)}^2 = 0$. Therefore, $v = 0$, and we have verified (3).

To take advantage of the abstract results of Section 2, we must establish (9) in this context. Specifically, we wish to show that $\mathfrak{j}_{LDG}$ given by

$$\mathfrak{j}_{LDG} = \inf_{q\in V/H_0^1(\Omega)} \sup_{\psi\in V} \frac{|J(q,\psi)|}{|\!|\!| q |\!|\!|_1 |\!|\!| \psi |\!|\!|_1} \tag{33}$$

is bounded away from 0 independent of $\underline{h}$ and $\underline{p}$.

Proposition 3.3. *There is a constant $c_{LDG} > 0$, independent of $\underline{h}$ and $\underline{p}$ (and τ), such that, for all $v \in V$,*

$$c_{LDG} \inf_{w\in H_0^1(\Omega)} |\!|\!| v - w |\!|\!|_1^2 \le J(v,v) \ .$$

Proof. As discussed in Section 2, we can decompose V as $V = H_0^1(\Omega) \oplus R$, where $R \subset S_{\underline{p}}(\mathcal{T})$. Writing $v = v_1 + v_2$, where $v_1 \in H_0^1(\Omega)$ and $v_2 \in R$, we clearly have

$$\inf_{w\in H_0^1(\Omega)} |\!|\!| v - w |\!|\!|_1^2 = \inf_{w\in H_0^1(\Omega)} |\!|\!| v_2 - w |\!|\!|_1^2 \ .$$

In [34] (2D) and [33] (3D), an averaging operator $\mathcal{I} : S_{\underline{p}}(\mathcal{T}) \to H_0^1(\Omega)$ is constructed for which there exists a constant c, independent of $\underline{h}$ and $\underline{p}$, such that, for all $v_2 \in S_{\underline{p}}(\mathcal{T})$,

$$\sum_{K\in\mathcal{T}} \|\nabla(v_2 - \mathcal{I}v_2)\|_{L^2(K)}^2 \le c\, J(v_2, v_2) \ .$$

From this, it is clear that $|\!|\!| v_2 - \mathcal{I}v_2 |\!|\!|_1^2 \le C_{\mathrm{cont}}(c+1) J(v_2,v_2) = c_{LDG}^{-1} J(v,v)$, which completes the proof. □

We can now state the main result of this section.

Proposition 3.4. *It holds that* $\mathfrak{j}_{LDG} \geq c_{LDG}$*, where* c_{LDG} *is the constant in Proposition 3.3.*

Proof. Each member of $V/H_0^1(\Omega)$ has the form $r - \mathcal{I}r$ for some $r \in R$, so

$$\mathfrak{j}_{LDG} \geq \inf_{r\in R} \sup_{\psi\in V} \frac{|J(r-\mathcal{I}r,\psi)|}{|\!|\!| r-\mathcal{I}r |\!|\!|_1 |\!|\!| \psi |\!|\!|_1} \geq \inf_{r\in R} \frac{|J(r-\mathcal{I}r, r-\mathcal{I}r)|}{|\!|\!| r-\mathcal{I}r |\!|\!|_1^2} = \inf_{r\in R} \frac{|J(r,r)|}{|\!|\!| r-\mathcal{I}r |\!|\!|_1^2} \geq c_{LDG} \ .$$

□

3.3. Symmetric interior penalty discontinuous Galerkin method

Taking J as in the LDG case (30), the symmetric interior penalty DG (SIPDG) method can be formulated in terms of the bilinear form $A_\tau : V \times V \to \mathbb{R}$

$$A_\tau(u,v) = \sum_{K\in\mathcal{T}} \int_K \nabla u \cdot \nabla v \, dx - \sum_{K\in\mathcal{T}} \int_K \mathcal{R}(u)\cdot\nabla v + \mathcal{R}(v)\cdot\nabla u \, dx + \tau J(u,v) \ . \tag{34}$$

We note that, when $u, v \in S_{\underline{p}}(\mathcal{T})$, we have

$$\sum_{K\in\mathcal{T}} \int_K \mathcal{R}(u)\cdot\nabla v + \mathcal{R}(v)\cdot\nabla u \, dx = \sum_{F\in\mathcal{F}(\mathcal{T})} \int_F \Big(\{\!\{\nabla u\}\!\} \cdot [\![v]\!] + \{\!\{\nabla v\}\!\} \cdot [\![u]\!] \Big) \, ds \ , \tag{35}$$

which provides a common alternative bilinear form for SIPDG.

Lemma 3.5. *There is a constant* $C_{\mathcal{R}}$*, independent of* $\underline{h}$ *and* $\underline{p}$ *such that, for any* $u \in V$*,*

$$\|\mathcal{R}(u)\|_{L^2(\Omega)} \leq C_{\mathcal{R}} \sqrt{J(u,u)} \ .$$

Proof. Letting $\Pi : [L^2(\Omega)]^d \to [S_{\underline{p}}(\mathcal{T})]^d$ be the L^2-projection, we have

$$\begin{aligned} \|\mathcal{R}(u)\|_{L^2(\Omega)} &= \sup_{z\in[L^2(\Omega)]^d} \frac{\int_\Omega \mathcal{R}(u)\cdot z \, dx}{\|z\|_{L^2(\Omega)}} = \sup_{z\in[L^2(\Omega)]^d} \frac{\int_\Omega \mathcal{R}(u)\cdot \Pi z \, dx}{\|z\|_{L^2(\Omega)}} \\ &= \sup_{z\in[L^2(\Omega)]^2} \frac{\int_\Gamma [\![u]\!]\cdot\{\!\{\Pi z\}\!\} \, ds}{\|z\|_{L^2(\Omega)}} \leq \sup_{z\in[L^2(\Omega)]^2} \frac{\|h^{-1/2}p[\![u]\!]\|_{L^2(\Gamma)} \|h^{1/2}p^{-1}\{\!\{\Pi z\}\!\}\|_{L^2(\Gamma)}}{\|z\|_{L^2(\Omega)}} \ . \end{aligned}$$

Note that $\|h^{-1/2}p[\![u]\!]\|_{L^2(\Gamma)} = \sqrt{J(u,u)}$, so it remains to bound $\|h^{1/2}p^{-1}\{\!\{\Pi z\}\!\}\|_{L^2(\Gamma)}$ in terms of $\|z\|_{L^2(\Omega)}$. The key to this is a trace inequality (cf. [24, Equation 4.6]),

$$\|q\|_{0,\partial K}^2 \leq C_{\mathrm{inv}} \frac{p_K^2}{h_K} \|q\|_{0,K}^2 \ , \quad \forall q \in S_{\underline{p}}(K) \ .$$

The constant C_{inv} depends only on the shape-regularity of the mesh. We have

$$\begin{aligned} \|h^{1/2}p^{-1}\{\!\{\Pi z\}\!\}\|_{L^2(\Gamma)}^2 &= \sum_{F\in\mathcal{F}(\mathcal{T})} \frac{h_F}{p_F^2} \int_F \{\!\{\Pi z\}\!\}\cdot\{\!\{\Pi z\}\!\} \, ds \leq \frac{1}{2} \sum_{K\in\mathcal{T}} \frac{h_K}{p_K^2} \|\Pi z\|_{L^2(\partial K)}^2 \\ &\leq \frac{C_{\mathrm{inv}}}{2} \sum_{K\in\mathcal{T}} \|\Pi z\|_{L^2(\partial K)}^2 \leq (C_{\mathrm{inv}}/2)\|z\|_{L^2(\Omega)}^2 \end{aligned}$$

Combining this with the results above completes the proof. □

This result naturally leads to a coercivity result for A_τ provided τ is sufficiently large. We recall the definition of the energy norm (31) for generic τ.

Proposition 3.6. *There are constants $0 < c_A < 1$ and $\tau_0 > 1$ such that $A_\tau(v,v) \geq c_A \|v\|_{E,1}^2$ for all $v \in V$ when $\tau \geq \tau_0$.*

Proof. Using Lemma 3.5 and Young's inequality, we see that, for any $0 < s < 1$,

$$\begin{aligned} A_\tau(v,v) &\geq (1-s) \sum_{K \in \mathcal{T}} \|\nabla v\|_{L^2(K)}^2 - s^{-1} \|\mathcal{R}(v)\|_{L^2(\Omega)}^2 + \tau J(v,v) \\ &\geq (1-s) \sum_{K \in \mathcal{T}} \|\nabla v\|_{L^2(K)}^2 + (\tau - C_{\mathcal{R}}^2/s) J(v,v) \ . \end{aligned}$$

At this stage, it is clear that choosing s and τ appropriately completes the proof. For example, choosing $s = 1/2$ and $\tau_0 = 2C_{\mathcal{R}}^2 + 1/2$ yields the coercivity bound with $c_A = 1/2$. More generally, any $\tau_0 > C_{\mathcal{R}}^2$ will yield coercivity with some $c_A > 0$, and it is clear that τ_0 has only to do with the shape-regularity of the mesh. □

The continuity of A_τ with respect to $\|\cdot\|_{E,1}$ is easy to prove because of Lemma 3.5, so we state the result without proof.

Proposition 3.7. *For $\tau \geq 0$, $A_\tau(u,v) \leq \max\{2, C_{\mathcal{R}}^2 + \tau\} \|u\|_{E,1} \|v\|_{E,1}$.*

We now define B and $\mathcal{B}_\tau$ for SIPDG by

$$B(u,v) = A_{\tau_0 - 1}(u,v) \quad , \quad \mathcal{B}_\tau(u,v) = B(u,v) + \tau J(u,v) \ , \tag{36}$$

where $\tau_0 > 1$ is a penalty parameter guaranteed by Lemma 3.6 to make A_{τ_0} coercive with respect to $\|\cdot\|_{E,1}$. The discussion above makes it clear that properties (1)-(3) of Section 2 are satisfied. As in the previous section, we define the relevant inf-sup constant

$$\mathfrak{j}_{SIP} = \inf_{q \in V/H_0^1(\Omega)} \sup_{\psi \in V} \frac{|J(q,\psi)|}{|\!|\!| q |\!|\!|_1 |\!|\!| \psi |\!|\!|_1} \ . \tag{37}$$

Using essentially the same argument as in Proposition 3.4, we obtain the analogous result for SIPDG,

Proposition 3.8. *There is a constant $c_{SIP} > 0$, independent of $\underline{h}$ and $\underline{p}$, such that $\mathfrak{j}_{SIP} \geq c_{SIP}$.*

4. An a posteriori error estimator

Having in mind the DG space $S_{\underline{p}}(\mathcal{T})$ as motivation, we introduce the one-parameter family of finite dimensional spaces S_ν, $\nu > 0$, satisfying the basic assumptions at the beginning of Section 2. Think of increasing ν as corresponding to *hp*-refinement of the DG space. We take $V_\nu = S_\nu + H_0^1(\Omega)$, and $B_\nu, J_\nu : V_\nu \times V_\nu \to \mathbb{R}$ to satisfy conditions (1)–(3), so that $\mathcal{B}_{\nu,1}$ is an inner-product on V_ν, where $\mathcal{B}_{\nu,\tau} = B_\nu + \tau J_\nu$. We make the further assumption, as in (9), that there is a constant $\mathfrak{j}_{LBB} > 0$, independent of ν, such that

$$\mathfrak{j}_{LBB} = \inf_{q \in V_\nu / \operatorname{Ker}(J_\nu)} \sup_{\psi \in V_\nu} \frac{|J_\nu(q,\psi)|}{|\!|\!| q |\!|\!|_{\nu,1} |\!|\!| \psi |\!|\!|_{\nu,1}} \ . \tag{38}$$

As shown in Section 3, both LDG and SIPDG fit within this abstract framework. What was implicit in the notation and results of Section 3, namely the dependence of the bilinear forms and norms on the discretization parameters $\underline{h}$ and $\underline{p}$, is here made explicit by use of the parameter ν as an index.

We recall the notational convention for the Galerkin approximation (7) from the finite dimensional space $\hat{S}_\nu \subset V_\nu$ for the data $f \in L^2(\Omega)$ as the function $\hat{s}_{\nu,\tau}(f) \in \hat{S}_\nu$ that satisfies

$$\mathcal{B}_{\nu,\tau}(\hat{s}_{\nu,\tau}(f), v) = (f, v) \text{ for all } v \in \hat{S}_\nu \ . \tag{39}$$

Given an L^2-orthonormal set of vectors $\{\hat{\psi}_{\nu,1}, \cdots, \hat{\psi}_{\nu,m}\} \subset V_\nu$, whose span we denote by $\hat{S}_{\nu,m}$, we also recall the definition of the approximation defects (16)

$$\eta_{\nu,\tau,j}(\hat{S}_{\nu,m}) = \max_{\substack{\mathcal{S} \subset \hat{S}_{\nu,m} \\ \dim \mathcal{S} = m-i-1}} \min_{f \in \mathcal{S}} \frac{|\!|\!| u_{\nu,\tau}(f) - \hat{s}_{\nu,m,\tau}(f) |\!|\!|_{\nu,\tau}}{|\!|\!| u_{\nu,\tau}(f) |\!|\!|_{\nu,\tau}} \quad , \quad 1 \le j \le m \ . \tag{40}$$

We might naturally think of $\hat{S}_{\nu,m}$ coming from a computed approximation of S_m in the discrete space S_ν, but this understanding is not necessary for the results below. We now express both eigenvalue error and invariant subspace projection error in terms of a "non-conformity" component, $\mathcal{R}_{nc}$, that will not be factored into practical computations, and an "a posteriori" component, $\mathcal{R}_{ap}$, for which a posteriori error estimates will be computed and local indicators used to drive an adaptive algorithm. Let $(\lambda_i, \psi_i) \in \mathbb{R} \times H_0^1(\Omega)$ be eigenpairs of B as in (12), and let S_m be the span of $\{\psi_1, \ldots, \psi_m\}$. We define the corresponding non-conformity error by

$$\mathcal{R}_{nc}(S_m, \nu) = \sum_{i=1}^{m} |\!|\!| u_{\nu,1}(\psi_i) - \lambda_i^{-1} \psi_i |\!|\!|_{\nu,1}^2 \ . \tag{41}$$

The a posteriori component of the error is given by

$$\mathcal{R}_{ap}(\hat{S}_{\nu,m}, \nu, \tau) = \sum_{i=1}^{m} \frac{|\!|\!| u_{\nu,\tau}(\hat{\psi}_{\nu,i}) - \hat{s}_{\nu,\tau}(\hat{\psi}_{\nu,i}) |\!|\!|_{\nu,\tau}^2}{|\!|\!| u_{\nu,\tau}(\hat{\psi}_{\nu,i}) |\!|\!|_{\nu,\tau}^2} \ . \tag{42}$$

Proposition 4.1. *In the context of the previous paragraph, if both $\eta_{\nu,\tau,j}(\hat{S}_{\nu,m})$ and $\eta_{\nu,1,j}(S_m)$ satisfy the conditions of Theorem 2.3, and $\hat{\lambda}_{\nu,\tau,i} = |\!|\!| \hat{\psi}_{\nu,i} |\!|\!|_{\nu,\tau}^2$, then*

$$\sum_{i=1}^{m} \frac{|\hat{\lambda}_{\nu,\tau,i} - \lambda_i|}{\lambda_i} \le \frac{C_m^{high}}{\tau - 1} \mathcal{R}_{nc}(S_m, \nu) + G_m \mathcal{R}_{ap}(\hat{S}_{\nu,m}, \nu, \tau) \ , \tag{43}$$

$$\|E(I) - \hat{E}_{\nu,\tau}(I)\|_{HS} \le \sqrt{\frac{C_m^{high}}{\tau - 1} \mathcal{R}_{nc}(S_m, \nu)} + G_m \sqrt{\mathcal{R}_{ap}(\hat{S}_{\nu,m}, \nu, \tau)} \ . \tag{44}$$

Here $\hat{E}_{\nu,\tau}(I)$ denotes the L^2-orthogonal projection onto the subspace $\hat{S}_{\nu,m}$.

Proof. We start from the simple inequality

$$\begin{aligned} \frac{|\hat{\lambda}_{\nu,\tau,i} - \lambda_i|}{\lambda_i} &\le \frac{\lambda_i - \lambda_{\nu,\tau,i}}{\lambda_i} + \frac{|\hat{\lambda}_{\nu,\tau,i} - \lambda_{\nu,\tau,i}|}{\lambda_i} \\ &\le \frac{\lambda_i - \lambda_{\nu,\tau,i}}{\lambda_i} + \frac{|\hat{\lambda}_{\nu,\tau,i} - \lambda_{\nu,\tau,i}|}{\lambda_{\nu,\tau,i}} \ , \end{aligned}$$

which holds for any $i = 1, .., m$. The proof now follows the reasoning of [7]. We obtain from Theorem 2.3 the estimate

$$\sum_{i=1}^{m} \frac{\hat{\lambda}_{\nu,\tau,i} - \lambda_{\nu,\tau,i}}{\lambda_i} \leq G_m \sum_{i=1}^{m} \frac{\|\!| u_{\nu,\tau}(\hat{\psi}_{\nu,i}) - \hat{s}_{\nu,\tau}(\hat{\psi}_{\nu,i}) |\!\|^2_{\nu,\tau}}{\|\!| u_{\nu,\tau}(\hat{\psi}_{\nu,i}) |\!\|^2_{\nu,\tau}} \ ,$$

where the constant G_m is obtained by modifying the constant $C_{\tau,m}$ to account for the different relative measure of the eigenvalue error. As discussed in Remark 2.4, $C_{\tau,m}$ can be bounded independently of τ, so we are not being remiss by excluding τ from the subscript of G_m. The sum $\sum_{i=1}^{m}(\lambda_i - \lambda_{\nu,\tau,i})/\lambda_i$ is bounded as in Proposition 2.5. The bound (44) follows by a similar argument. □

Remark 4.2. To relate the estimates from the preceding proposition to other approaches in the literature recall Remark 2.6. Using the approach from [7, 17] we may obtain similar estimates for the difference in projections measured in other unitarily invariant norms—in particular, we may do so for the operator norm. The practical estimates presented in Section 5 are of hierarchical type, which allows for the computation of estimates in such norms, as was shown in [7, 17]. We have opted for the Hibert-Schmidt norm in the present work because it provides the cleanest statements of such estimates within our framework. For an alternative approache to optimality in error estimation for eigenvalue clusters, we again refer to [10].

4.1. Controlling the non-conformity error

For the following we assume the regularity estimate, as in [3, Property 1], that

$$\|\!| u_{\nu,\tau}(f) |\!\|_{\nu,\tau} = \sup_{v \in V_\nu} \frac{B_{\nu,\tau}(u_{\nu,\tau}(f), v)}{\|\!| v |\!\|_{\nu,\tau}} \leq C\|f\|, \qquad f \in L^2(\Omega) \ ,$$

where C is independent of ν and τ. A trivial, though pessimistic, estimate of the non-conformity error is given by

$$\mathcal{R}_{nc}(S_m, \nu) = \sum_{i=1}^{m} \|\!| u_{\nu,1}(\psi_i) - \lambda_i^{-1}\psi_i |\!\|^2_{\nu,1} \leq 2mC^2 + 2\sum_{i=1}^{m} \lambda_i^{-2} \ .$$

Even with this crude estimate, we see that $\mathcal{R}_{ap}(\hat{S}_{\nu,m}, \nu, \tau)$ is readily made the dominant term in the error bounds (43) and (44) by choosing τ sufficiently large.

If the spaces V_ν and forms $\mathcal{B}_{\nu,1}$ are such that the forms $\mathcal{B}_{\nu,1}$ are monotone increasing in ν and

$$H_0^1(\Omega) = \left\{ v \in \cup_\nu V_\nu \ : \ \lim_{\nu\to\infty} \|\!| v |\!\|_{\nu,1} < \infty \right\} \ , \tag{45}$$

we have the stronger result,

$$\lim_{\nu\to\infty} \|\!| u_{\nu,1}(\psi_i) - \lambda_i^{-1}\psi_i |\!\|_{\nu,1} = 0 \ , \tag{46}$$

which implies that the non-conformity error decays as ν increases. The limit (46) follows from the monotone convergence theorem for forms [32]—for its use in the context of residual error estimates,

see [16, Theorem 2.1]. Such a family of spaces is generated by pure p-refinement of $S_{\underline{p}}(\mathcal{T})$ for a fixed triangulation $\mathcal{T}$, if $\mathcal{B}_{\nu,1}$ incorporates the penalty term J in (30). More specifically, we have

$$H_0^1(\Omega) = \left\{ v \in \cup_{\underline{p}} \left(S_{\underline{p}}(\mathcal{T}) + H_0^1(\Omega) \right) \ : \lim_{\nu \to \infty} |\!|\!| v |\!|\!|_{\nu,1} < \infty \right\} ,$$

where $S_\nu = S_{\underline{p}_\nu}(\mathcal{T})$ and $\#\nu = \min\{\underline{p}_\nu\}$. This suggests that, for discretizations such as LDG and SIPDG, a modest choice of τ is likely to make $\mathcal{R}_{ap}$ the dominant contributor to eigenvalue and eigenvector error bounds.

5. An auxiliary subspace error estimator

We now turn to providing a computable estimate of $\mathcal{R}_{ap}(\hat{S}_m, \nu, \tau)$. At this stage, we will no longer consider the approximation space S_ν in its most abstract form, but will focus on the hp-spaces $S_\nu = S_{\underline{p}}(\mathcal{T})$ that motivated our abstract development. The primary reason for being more specific here is that the design of a posteriori estimates that are both efficient and reliable requires more specificity in practice. As such, we replace the subscript ν with hp to reflect this shift. We wish to obtain a practical a posteriori estimate of

$$\begin{aligned} \mathcal{R}_{ap}(\hat{S}_m, hp, \tau) &= \sum_{i=1}^{m} \frac{|\!|\!| u_{hp,\tau}(\hat{\psi}_{\tau,i}) - \hat{s}_{hp,\tau}(\hat{\psi}_{\tau,i}) |\!|\!|_{\nu,\tau}^2}{|\!|\!| u_{hp,\tau}(\hat{\psi}_{\tau,i}) |\!|\!|_{\nu,\tau}^2} \\ &= \sum_{i=1}^{m} \frac{|\!|\!| u_{hp,\tau}(\hat{\psi}_{\tau,i}) - \hat{s}_{hp,\tau}(\hat{\psi}_{\tau,i}) |\!|\!|_{\nu,\tau}^2}{|\!|\!| u_{hp,\tau}(\hat{\psi}_{\tau,i}) - \hat{s}_{hp,\tau}(\hat{\psi}_{\tau,i}) |\!|\!|_{\nu,\tau}^2 + |\!|\!| \hat{s}_{hp,\tau}(\hat{\psi}_{\tau,i}) |\!|\!|_{\nu,\tau}^2} , \end{aligned} \tag{47}$$

so we see that it is sufficient to estimate errors of the form $|\!|\!| u_{hp,\tau}(f) - \hat{s}_{hp,\tau}(f) |\!|\!|_{\nu,\tau}$, where $f \in L^2(\Omega)$, and $u_{hp,\tau}(f) \in V_{hp} = H_0^1(\Omega) + S_{\underline{p}}(\mathcal{T})$ and $\hat{s}_{hp,\tau}(f) \in \hat{S}_m \subset S_{\underline{p}}(\mathcal{T})$ satisfy (6) and (7).

We will employ a hierarchical basis type error estimator, which is novel in the DG setting, but is a well-known approach for continuous Galerkin discretizations (cf. [6, 2, 29]), and has been used successfully for eigenvalue problems (cf. [17, 7]). Recall that the restriction of a function in $S_{\underline{p}}(\mathcal{T})$ to an element K is in the space of tensor-product polynomials $\mathcal{Q}_{p_K}(K)$ (21). We decompose a p-enrichment of this local space hierarchically,

$$\mathcal{Q}_{p_K+1}(K) = \mathcal{Q}_{p_K}(K) \oplus \mathcal{E}_{p_K+1}(K) , \tag{48}$$

and define an auxiliary space in which we will approximate the error $u_{hp,\tau}(f) - \hat{s}_{hp,m,\tau}(f)$ as a function,

$$\mathcal{E}_{\underline{p}}(\mathcal{T}) = \{ \, v \in L^2(\Omega) \, : \, v|_K \in \mathcal{E}_{p_K+1}(K), \ K \in \mathcal{T} \, \} . \tag{49}$$

We remark that the definition of $\mathcal{E}_{p_K+1}(K)$ via the direct-sum (48) leaves some ambiguity, and we now describe how to make a well-defined choice. For a given $p > 1$, a hierarchical basis of $\mathcal{Q}_p(K)$ is typically built up recursively, beginning with a basis for $\mathcal{Q}_1(K)$, extending it to a basis of $\mathcal{Q}_2(K)$ by adding appropriate functions from $\mathcal{Q}_2(K) \setminus \mathcal{Q}_1(K)$, and continuing in this fashion until a basis for $\mathcal{Q}_p(K)$ is obtained. A number of popular strategies for building such hierarchical bases exist for simplices and tensorial elements, with the construction for the latter being based on simple 1D hierarchies. The particular strategy used will not factor into our discussion, and we merely point interested readers to [25, 27, 9, 8, 1, 26] for discussion of several of them. Once a particular strategy

has been chosen, $\mathcal{E}_{p_K+1}(K)$ is spanned by those basis functions for $\mathcal{Q}_{p_K+1}(K)$ that are not part of the basis for $\mathcal{Q}_{p_K}(K)$.

Our hierarchical error estimate is based on the function $\varepsilon_{hp,\tau}(f) \in \mathcal{E}_{\underline{p}}(\mathcal{T})$ satisfying

$$\mathcal{B}_{hp,\tau}(\varepsilon_{hp,\tau}(f), v) = (f, v) - \mathcal{B}_{hp,\tau}(\hat{s}_{hp,m,\tau}(f), v) \quad \text{for all } v \in \mathcal{E}_{\underline{p}}(\mathcal{T}) \ . \tag{50}$$

It is convenient to introduce the space $V'_{hp} = V_{hp} + \mathcal{E}_{\underline{p}}(\mathcal{T})$ and the function $u'_{hp,\tau}(f) \in V'_{hp}$ satisfying

$$\mathcal{B}_{hp,\tau}(u'_{hp,\tau}(f), v) = (f, v) \quad \text{for all } v \in V'_{hp} \ . \tag{51}$$

We emphasize that the bilinear form is not modified in either case just because we use polynomials of one degree higher. The constants we obtained in Section 3 need only be modestly adjusted, in a p-independent way, to accommodate the use of $\mathcal{B}_{hp,\tau}$ on $S_{\underline{p}}(\mathcal{T}) \oplus \mathcal{E}_{\underline{p}}(\mathcal{T})$.

The definitions (50) and (51) make it clear that $\varepsilon_{hp,\tau}(f)$ is the $\mathcal{B}_{hp,\tau}$-orthogonal projection of $u'_{hp,\tau}(f) - \hat{s}_{hp,\tau}(f)$ on $\mathcal{E}_{\underline{p}}(\mathcal{T})$, so immediately have the lower bound

$$|\!|\!|\varepsilon_{hp,\tau}(f)|\!|\!|_{hp,\tau} \le |\!|\!|u'_{hp,\tau}(f) - \hat{s}_{hp,\tau}(f)|\!|\!|_{hp,\tau} \ . \tag{52}$$

Using that $S_{\underline{p}}(\mathcal{T}) \subset V_{hp} \subset V'_{hp}$, together with Galerkin orthogonality, we have the Pythagorean identity

$$|\!|\!|u'_{hp,\tau}(f) - \hat{s}_{hp,\tau}(f)|\!|\!|^2_{hp,\tau} = |\!|\!|u'_{hp,\tau}(f) - u_{hp,\tau}(f)|\!|\!|^2_{hp,\tau} + |\!|\!|u_{hp,\tau}(f) - \hat{s}_{hp,\tau}(f)|\!|\!|^2_{hp,\tau} \ . \tag{53}$$

We make the following saturation assumption: there is a constant $0 < q < 1$, independent of h and p, such that

$$|\!|\!|u'_{hp,\tau}(f) - u_{hp,\tau}(f)|\!|\!|_{hp,\tau} \le q|\!|\!|u'_{hp,\tau}(f) - \hat{s}_{hp,\tau}(f)|\!|\!|_{hp,\tau} \ . \tag{54}$$

Under this assumption, we have a lower bound on the actual quantity of interest,

$$|\!|\!|\varepsilon_{hp,\tau}(f)|\!|\!|_{hp,\tau} \le (1-q^2)^{-1/2}|\!|\!|u_{hp,\tau}(f) - \hat{s}_{hp,\tau}(f)|\!|\!|_{hp,\tau} \ . \tag{55}$$

At this stage, we have the following, computable, estimate of $\mathcal{R}_{ap}$

$$\mathcal{R}_{ap}(\hat{S}_m, hp, \tau) \approx \hat{\mathcal{R}}_{ap}(\hat{S}_m, hp, \tau) = \sum_{i=1}^{m} \frac{|\!|\!|\varepsilon_{hp,\tau}(\hat{\psi}_{\tau,i})|\!|\!|^2_{hp,\tau}}{|\!|\!|\varepsilon_{hp,\tau}(\hat{\psi}_{\tau,i})|\!|\!|^2_{hp,\tau} + |\!|\!|\hat{s}_{hp,\tau}(\hat{\psi}_{\tau,i})|\!|\!|^2_{hp,\tau}} \ . \tag{56}$$

The computation of the terms in $\hat{\mathcal{R}}_{ap}(\hat{S}_m, hp, \tau)$ deserves further comment. If the vectors $\{\hat{\psi}_{\tau,i} : 1 \le i \le m\}$ have been provided without any further information about their origin, the computation of $\hat{\mathcal{R}}_{ap}(\hat{S}_m, hp, \tau)$ requires the solution of m source problems in $S_{\underline{p}}(\mathcal{T})$ to obtain the functions $\hat{s}_{hp,\tau}(\hat{\psi}_{\tau,i})$, and m additional source problems in $\mathcal{E}_{\underline{p}}(\mathcal{T})$ to obtain the approximate error functions $\varepsilon_{hp,\tau}(\hat{\psi}_{\tau,i})$. The latter collection of problems is unavoidable for hierarchical-type error estimates, but the nature of the space $\mathcal{E}_{\underline{p}}(\mathcal{T})$ is such that these problems are not as expensive as one might think. The computation of the $\hat{s}_{hp,\tau}$ is greatly simplified if the approximate eigenvectors $\hat{\psi}_{\tau,i}$ have been obtained as via a discrete eigenvalue problem posed in $S_{\underline{p}}(\mathcal{T})$, as will typically be the case. Specifically, if $0 < \hat{\lambda}_{hp,\tau,1} \le \hat{\lambda}_{hp,\tau,2} \le \cdots \le \hat{\lambda}_{hp,\tau,m}$ are the smallest m eigenvalues of the discrete problem,

$$\mathcal{B}_{hp,\tau}(\hat{\psi}_{hp,\tau}, v) = \hat{\lambda}_{hp,\tau}(\hat{\psi}_{hp,\tau}, v) \qquad \forall v \in S_{\underline{p}}(\mathcal{T}) \ , \tag{57}$$

and $\{\hat{\psi}_{hp,\tau,i} : 1 \leq i \leq m\}$ are corresponding L^2-orthonormal eigenvectors, with $\hat{S}_m = \hat{S}_{hp,m}$ being their span, we have $\hat{s}_{hp,\tau}(\hat{\psi}_{hp,\tau,i}) = \hat{\lambda}^{-1}_{hp,\tau,i}\hat{\psi}_{hp,\tau,i}$, and (56) simplifies to

$$\hat{\mathcal{R}}_{ap}(\hat{S}_m, hp, \tau) = \sum_{i=1}^{m} \frac{|||\varepsilon_{hp,\tau}(\hat{\psi}_{\tau,i})|||^2_{hp,\tau}}{|||\varepsilon_{hp,\tau}(\hat{\psi}_{\tau,i})|||^2_{hp,\tau} + \hat{\lambda}^{-2}_{hp,\tau,i}} \ . \tag{58}$$

Remark 5.1. It may seem odd that, in a paper about eigenvalue and eigenvector error estimation, the first true mention of how such approximations might be computed comes near the end of the paper. This was done to emphasize the fact that the results up to this point have not required that the approximations are obtained by (exactly) solving (57). In particular, inexact solves of (57) are typically expected. In practice, we advocate using $\hat{\lambda}^{-1}_{hp,\tau,i}\hat{\psi}_{hp,\tau,i}$ instead of $\hat{s}_{hp,\tau}(\hat{\psi}_{hp,\tau,i})$, as well as (58), when these quantities are obtained from inexact solves of (57).

We have yet to establish that $|||\varepsilon_{hp,\tau}(f)|||_{hp,\tau}$ can be used to bound $|||u_{hp,\tau}(f) - \hat{s}_{hp,\tau}(f)|||_{hp,\tau}$ from above. Since $|||u_{hp,\tau}(f) - \hat{s}_{hp,\tau}(f)|||_{hp,\tau} \leq |||u'_{hp,\tau}(f) - \hat{s}_{hp,\tau}(f)|||_{hp,\tau}$, if we show that $|||\varepsilon_{hp,\tau}(f)|||_{hp,\tau}$ can be used to bound $|||u'_{hp,\tau}(f) - \hat{s}_{hp,\tau}(f)|||_{hp,\tau}$ from above, then we have the bound we actually seek. In order to do this, we employ the standard approach used in the continuous finite element setting (cf. [6, 2, 29]), which is based on a strong Cauchy inequality and a (second) saturation assumption. The strong Cauchy inequality states that there is a constant, $0 < \gamma < 1$, such that

$$\mathcal{B}_{hp,\tau}(v, w) \leq \gamma |||v|||_{hp,\tau} |||w|||_{hp,\tau} \qquad \forall v \in S_{\underline{p}}(\mathcal{T}) \ , \ \forall w \in \mathcal{E}_{\underline{p}}(\mathcal{T}) \ . \tag{59}$$

The saturation assumption states that $u'_{hp,\tau}(f)$ is better approximated in $S_{\underline{p}}(\mathcal{T}) \oplus \mathcal{E}_{\underline{p}}(\mathcal{T})$ than it is in $S_{\underline{p}}(\mathcal{T})$, which makes intuitive sense because $S_{\underline{p}}(\mathcal{T}) \oplus \mathcal{E}_{\underline{p}}(\mathcal{T})$ locally contains polynomials of one degree higher than $S_{\underline{p}}(\mathcal{T})$. More formally, this saturation assumption states that there is a constant, $0 < \mu < 1$, such that

$$\inf_{v \in S_{\underline{p}}(\mathcal{T}) \oplus \mathcal{E}_{\underline{p}}(\mathcal{T})} |||u'_{hp,\tau}(f) - v|||_{hp,\tau} \leq \mu |||u'_{hp,\tau}(f) - \hat{s}_{hp,\tau}(f)|||_{hp,\tau} \ . \tag{60}$$

With these ingredients, one may take the argument of [6, Theorem 1] essentially verbatim to obtain

$$|||u_{hp,\tau}(f) - \hat{s}_{hp,\tau}(f)|||^2_{hp,\tau} \leq |||u'_{hp,\tau}(f) - \hat{s}_{hp,\tau}(f)|||^2_{hp,\tau} \leq \frac{|||\varepsilon_{hp,\tau}(\hat{\psi}_{\tau,i})|||^2_{hp,\tau}}{(1-\gamma^2)(1-\mu^2)} \ . \tag{61}$$

Based on the discussion above, we obtain our final result, which compares the ideal quantity $\mathcal{R}_{ap}(\hat{S}_m, hp, \tau)$ to the computable one $\hat{\mathcal{R}}_{ap}(\hat{S}_m, hp, \tau)$.

Proposition 5.2. *Under the saturation assumptions* (54) *and* (60), $\hat{\mathcal{R}}_{ap}(\hat{S}_m, hp, \tau)$ *provides an approximation of* $\mathcal{R}_{ap}(\hat{S}_m, hp, \tau)$ *that is both efficient and reliable. More specifically,*

$$c\,\hat{\mathcal{R}}_{ap}(\hat{S}_m, hp, \tau) \leq \mathcal{R}_{ap}(\hat{S}_m, hp, \tau) \leq c^{-1}\hat{\mathcal{R}}_{ap}(\hat{S}_m, hp, \tau) \ ,$$

where $c = (1-\gamma^2)(1-\mu^2)(1-q^2)$ *and* γ *is the optimal constant in the strong Cauchy inequality* (59).

Proof. We bound the terms in $\mathcal{R}_{ap}(\hat{S}_m, hp, \tau)$ by their corresponding terms in $\hat{\mathcal{R}}_{ap}(\hat{S}_m, hp, \tau)$. To save space, we use the shorthand $e_{hp,\tau}(\hat{\psi}_{\tau,i}) = u_{hp,\tau}(\hat{\psi}_{\tau,i}) - \hat{s}_{hp,\tau}(\hat{\psi}_{\tau,i})$. Combining (55) and (61),

we have the upper- and lower-bounds,

$$\frac{|||e_{hp,\tau}(\hat{\psi}_{\tau,i})|||^2_{hp,\tau}}{|||e_{hp,\tau}(\hat{\psi}_{\tau,i})|||^2_{hp,\tau} + |||\hat{s}_{hp,\tau}(\hat{\psi}_{\tau,i})|||^2_{hp,\tau}} \leq c^{-1} \frac{|||\varepsilon_{hp,\tau}(\hat{\psi}_{\tau,i})|||^2_{hp,\tau}}{|||\varepsilon_{hp,\tau}(\hat{\psi}_{\tau,i})|||^2_{hp,\tau} + |||\hat{s}_{hp,\tau}(\hat{\psi}_{\tau,i})|||^2_{hp,\tau}} ,$$

$$\frac{|||e_{hp,\tau}(\hat{\psi}_{\tau,i})|||^2_{hp,\tau}}{|||e_{hp,\tau}(\hat{\psi}_{\tau,i})|||^2_{hp,\tau} + |||\hat{s}_{hp,\tau}(\hat{\psi}_{\tau,i})|||^2_{hp,\tau}} \geq c \frac{|||\varepsilon_{hp,\tau}(\hat{\psi}_{\tau,i})|||^2_{hp,\tau}}{|||\varepsilon_{hp,\tau}(\hat{\psi}_{\tau,i})|||^2_{hp,\tau} + |||\hat{s}_{hp,\tau}(\hat{\psi}_{\tau,i})|||^2_{hp,\tau}} ,$$

where $c = (1-\gamma^2)(1-\mu^2)(1-q^2)$. Summing these inequalities completes the proof. □

Remark 5.3. Since the introduction of hierarchical type error estimators in the 1980s, saturation assumptions of the form (60) have been typical in their reliability analysis for continuous Galerkin approximations in the energy norm setting (cf. [6]). Although the saturation assumption can be removed if one is willing to incorporate an additional "oscillation term" in the upper-bound, as was done in [17, 7, 19, 21], we thought that type of reliability analysis would unnecessarily add to the technical burden on readers, so we have not pursued it here. For additional insights on, and justification of, the use of saturation assumptions in finite element analysis, we refer interested readers to [12, 11]. References [17, 7] above are specifically related to eigenvalue problems, and provide several numerical examples that justify the use of hierarchical error estimators in this context for continuous Galerkin discretizations.

Acknowledgements. The authors very gratefully acknowledge the Mathematisches Forschungsinstitut Oberwolfach for hosting them for the Research-In-Pairs program "High-Order Finite Element Methods Elliptic Eigenvalue Problems". It was during this stay that the groundwork was laid for this and several other projects. The work of J.S. Ovall was also partially supported by the National Science Foundation grant DMS-1414365. L. Grubišić has in part been supported by the Croatian Science Foundation grant HRZZ grant number 9345.

References

[1] S. Adjerid, M. Aiffa, and J. E. Flaherty. Hierarchical finite element bases for triangular and tetrahedral elements. *Comput. Methods Appl. Mech. Engrg.*, 190(22-23):2925 – 2941, 2001.

[2] M. Ainsworth and J. T. Oden. *A posteriori error estimation in finite element analysis*. Pure and Applied Mathematics (New York). Wiley-Interscience [John Wiley & Sons], New York, 2000.

[3] P. F. Antonietti, A. Buffa, and I. Perugia. Discontinuous Galerkin approximation of the Laplace eigenproblem. *Comput. Methods Appl. Mech. Engrg.*, 195(25-28):3483–3503, 2006.

[4] M. G. Armentano and R. G. Durán. Asymptotic lower bounds for eigenvalues by nonconforming finite element methods. *Electron. Trans. Numer. Anal.*, 17:93–101 (electronic), 2004.

[5] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini. Unified analysis of discontinuous galerkin methods for elliptic problems. *SIAM J. Numer. Anal*, pages 1749–1779, 2001.

[6] R. E. Bank. Hierarchical bases and the finite element method. In *Acta numerica, 1996*, volume 5 of *Acta Numer.*, pages 1–43. Cambridge Univ. Press, Cambridge, 1996.

[7] R. E. Bank, L. Grubišić, and J. S. Ovall. A framework for robust eigenvalue and eigenvector error estimation and Ritz value convergence enhancement. *Appl. Numer. Math.*, 66:1–29, 2013.

[8] S. Beuchler and J. Schöberl. New shape functions for triangular p-FEM using integrated Jacobi polynomials. *Numer. Math.*, 103(3):339–366, 2006.

[9] P. Carnevali, R. B. Morris, Y. Tsuji, and G. Taylor. New basis functions and computational procedures for p-version finite element analysis. *International Journal for Numerical Methods in Engineering*, 36(22):3759–3779, 1993.

[10] A. Bonito and A. Demlow. Convergence and optimality of higher-order adaptive finite element methods for eigenvalue clusters. *SIAM J. Numer. Anal.*, 54(4), 2379–2388, 2016.

[11] Carstensen, C., Gallistl, D., Gedicke, J.: Justification of the saturation assumption. Numer. Math. **134**(1), 1–25 (2016).

[12] Dörfler, W., Nochetto, R.H.: Small data oscillation implies the saturation assumption. Numer. Math. **91**(1), 1–12 (2002)

[13] V. Eijkhout and P. Vassilevski. The role of the strengthened Cauchy-Buniakowskiĭ-Schwarz inequality in multilevel methods. *SIAM Rev.*, 33(3):405–419, 1991.

[14] D. Gallistl. Adaptive nonconforming finite element approximation of eigenvalue clusters. *Comput. Methods Appl. Math.*, 14(4):509–535, 2014.

[15] S. Giani and L. Grubišić and J.S. Ovall. Benchmark results for testing adaptive finite element eigenvalue procedures. *Appl. Numer. Math.*, 62(2):121–140, 2012.

[16] L. Grubišić. Relative convergence estimates for the spectral asymptotic in the large coupling limit. *Integral Equations Operator Theory*, 65(1):51–81, 2009.

[17] L. Grubišić and J. S. Ovall. On estimators for eigenvalue/eigenvector approximations. *Math. Comp.*, 78:739–770, 2009.

[18] H. Hakula and T. Tuominen. Mathematica implementation of the high order finite element method applied to eigenproblems. *Computing*, 95(1, suppl.):S277–S301, 2013.

[19] H. Hakula and M. Neilan and J.S. Ovall. A Posteriori Estimates Using Auxiliary Subspace Techniques. *Journal of Scientific Computing*, (in press): 1–31, 2017. url="http://dx.doi.org/10.1007/s10915-016-0352-0"

[20] J. S. Hesthaven and T. Warburton. *Nodal discontinuous Galerkin methods*, volume 54 of *Texts in Applied Mathematics*. Springer, New York, 2008. Algorithms, analysis, and applications.

[21] M. Holst, J. S. Ovall, and R. Szypowski. An efficient, reliable and robust error estimator for elliptic problems in R^3. *Applied Numerical Mathematics*, 61(5):675 – 695, 2011.

[22] T. Kato. *Perturbation theory for linear operators*. Classics in Mathematics. Springer-Verlag, Berlin, 1995. Reprint of the 1980 edition.

[23] I. Perugia and D. Schötzau. An *hp*-analysis of the local discontinuous galerkin method for diffusion problems. *J. Sci. Comput.*, 17(1-4):561–571, 2002.

[24] I. Perugia and D. Schötzau. The hp-local Discontinuous Galerkin Method for Low-Frequency Time-Harmonic Maxwell Equations. *Mathematics of Computation*, 72(243):1179–1214, 2003.

[25] C. Schwab. *p- and hp-finite element methods*. Numerical Mathematics and Scientific Computation. The Clarendon Press, Oxford University Press, New York, 1998. Theory and applications in solid and fluid mechanics.

[26] P. Šolín, K. Segeth, and I. Doležel. *Higher-order finite element methods*. Studies in Advanced Mathematics. Chapman & Hall/CRC, Boca Raton, FL, 2004. With 1 CD-ROM (Windows, Macintosh, UNIX and LINUX).

[27] B. Szabó and I. Babuška. *Finite element analysis*. A Wiley-Interscience Publication. John Wiley & Sons Inc., New York, 1991.

[28] D. B. Szyld. The many proofs of an identity on the norm of oblique projections. *Numer. Algorithms*, 42(3-4):309–323, 2006.

[29] R. Verfürth. *A posteriori error estimation techniques for finite element methods*. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, 2013.

[30] T. Warburton and M. Embree. The role of the penalty in the local discontinuous Galerkin method for Maxwell's eigenvalue problem. *Comput. Methods Appl. Mech. Engrg.*, 195(25-28):3205–3223, 2006.

[31] J. Weidmann. Stetige Abhängigkeit der Eigenwerte und Eigenfunktionen elliptischer Differentialoperatoren vom Gebiet. *Math. Scand.*, 54(1):51–69, 1984.

[32] J. Weidmann. Stetige Abhängigkeit der Eigenwerte und Eigenfunktionen elliptischer Differentialoperatoren vom Gebiet. *Math. Scand.*, 54(1):51–69, 1984.

[33] L. Zhu, S. Giani, P. Houston, and D. Schötzau. Energy norm a posteriori error estimation for hp-adaptive discontinuous galerkin methods for elliptic problems in three dimensions. *Mathematical Models and Methods in Applied Sciences (M3AS)*, 21(2):267–306, 2011.

[34] L. Zhu and D. Schötzau. A robust *a posteriori* error estimate for *hp*-adaptive DG methods for convection-diffusion equations. *IMA J. Numer. Anal.*, 31(3):971–1005, 2011.

[35] D. N. Arnold. An interior penalty finite element method with discontinuous elements. *SIAM J. Numer. Anal.*, 19(4):742–760, 1982.

[36] I. Babuška. The finite element method with penalty. *Mathematics of Computation*, 27(122):221–228, 1973.

[37] B. Cockburn and C.-W. Shu The local discontinuous Galerkin method for time-dependent convection-diffusion systems. *SIAM J. Numer. Anal.*, 35(6):2440–2463, 1998.

[38] R. A. Adams and J. J. F. Fournier. *Sobolev spaces*. Pure and applied Mathematics. Academic Press, 2003. Second edition.