

Interior over-penalized enriched Galerkin methods for second order elliptic equations

Jeonghun J. Lee¹, Omar Ghattas^{1,1}

^a*Department of Mathematics, Baylor University, Waco, Texas, USA*

^b*Department of Mechanical Engineering, The University of Texas at Austin, Austin, Texas, USA*

^c*The Oden Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, Texas, USA*

Abstract

In this paper we propose a variant of enriched Galerkin methods for second order elliptic equations with over-penalization of interior jump terms. The bilinear form with interior over-penalization gives a non-standard norm which is different from the discrete energy norm in the classical discontinuous Galerkin methods. Nonetheless we prove that optimal a priori error estimates with the standard discrete energy norm can be obtained by combining a priori and a posteriori error analysis techniques. We also show that the interior over-penalization is advantageous for constructing preconditioners robust to mesh refinement by analyzing spectral equivalence of bilinear forms. Numerical results are included to illustrate the convergence and preconditioning results.

Keywords: discontinuous Galerkin methods, error analysis, preconditioning
2000 MSC: 68Q25, 68R10, 68U05

1. Introduction

It is well known that numerical fluxes of Galerkin methods with Lagrange finite elements (CG) do not satisfy local mass conservation. Since local mass conservation is an important physical principle in numerical simulations, numerical methods providing locally mass conservative numerical flux with/without postprocessing have been intensively studied. These include post-processing of the CG methods [1, 2, 3], the mixed method [4, 5, 6, 7], and a class of discontinuous Galerkin (DG) methods [8, 9, 1, 1, 1]), to name a few.

Enriched Galerkin (EG) methods [1] are proposed to achieve locally mass conservative numerical methods with low computational costs. Compared to classical DG methods for the primal formulation, EG methods have much fewer degrees of freedom and provide a simple local post-processing for reconstruction of locally mass conservative numerical fluxes. For efficient solvers, block diagonal preconditioners based on algebraic multigrid methods have been proposed as efficient preconditioners for EG methods [1]. However, the preconditioning

analysis in this work relies on the full elliptic regularity assumption which requires restrictive conditions on the domain geometry and material coefficients. In fact, numerical experiments show that the block preconditioners in [1] are not scalable with mesh refinement when the boundary condition is not a pure Neumann boundary condition.

The purpose of this paper is to develop new EG methods that retain the advantages of local mass conservation while also providing more robust fast solvers that do not need restrictive assumptions on domain geometry and boundary conditions. The key in the development of these new EG methods is to replace the original bilinear form with a new bilinear form that has an over-penalized interior jump term. As such, we call them interior over-penalized enriched Galerkin methods (IOP-EG). The over-penalization parameters depend on the mesh sizes, so the over-penalized jump term and other terms in the new bilinear form do not have the same scaling property. Because of these incompatible scaling properties, the standard techniques for error analysis are not available for the optimal error estimates of the new EG methods.

The first main result in this paper is an error analysis with optimal error estimates. To overcome the difficulty caused by the bilinear forms with different scalings, we use a medius analysis idea utilizing a posteriori error estimate results for our a priori error estimates. Through this idea we can obtain optimal error estimates for IOP-EG methods. The second main result is the construction of scalable preconditioners for the IOP-EG methods. More precisely, we propose an abstract form of preconditioners based on appropriate mesh-dependent norms and carry out spectral equivalence analysis for the abstract preconditioners following the framework in [1]. It turns out that over-penalization is crucial for construction of scalable preconditioners via an analysis of spectral equivalence.

The paper is organized as follows. In Section 2 we introduce notation and define the IOP-EG methods. In Section 3, we prove optimal error estimates of the IOP-EG methods with a minimal regularity assumption using a medius error analysis. Finally, numerical results illustrating our theoretical results and concluding remarks will be given in Sections 5 and 6.

2. Preliminaries

2.1. Notations

For a set $D \subset \mathbb{R}^n$ we use $L^2(D)$ to denote the space of square integrable functions on D . For a finite-dimensional vector space \mathbb{X} , $L^2(D; \mathbb{X})$ is the space of \mathbb{X} -valued square-integrable functions on D whose inner product is naturally defined with the inner product on \mathbb{X} and the inner product on $L^2(D)$.

Let Ω be a bounded polygonal/polyhedral domain in \mathbb{R}^n with $n = 2$ or 3 . For a nonnegative real number s , $H^s(\Omega)$ denotes the standard Sobolev spaces based on the L^2 norm (see [1] for details). We use \mathcal{T}_h to denote a shape-regular triangulation of Ω where h is the maximum diameter of triangles or tetrahedra

in \mathcal{T}_h . We also use \mathcal{E}_h to denote the set of facets in \mathcal{T}_h , i.e., the set of $(n-1)$ -dimensional simplices in the triangulation determined by \mathcal{T}_h . In particular, \mathcal{E}_h^0 and \mathcal{E}_h^∂ are the sets of interior and boundary facets of \mathcal{T}_h , respectively.

For $e \in \mathcal{E}_h$ and functions $f, g \in L^2(e)$ we define $\langle f, g \rangle_e = \int_e f \cdot g \, ds$. For a set Γ which is a union of facets in \mathcal{E}_h , we define

$$\langle f, g \rangle_\Gamma = \sum_{e \in \Gamma} \langle f, g \rangle_e.$$

This notation is naturally extended to $\langle \mathbf{f}, \mathbf{g} \rangle_\Gamma$ for vector-valued functions $\mathbf{f}, \mathbf{g} \in L^2(e; \mathbb{R}^n)$ with the natural inner product on \mathbb{R}^n .

For an integer $k \geq 0$ and for each $T \in \mathcal{T}_h$, $\mathcal{P}_k(T)$ is the space of polynomials of degree $\leq k$ on T , and $\mathcal{P}_k(\mathcal{T}_h)$ denotes the space

$$\mathcal{P}_k(\mathcal{T}_h) = \begin{cases} \{q \in H^1(\Omega) : q|_T \in \mathcal{P}_k(T), T \in \mathcal{T}_h\} & \text{if } k \geq 1 \\ \{q \in L^2(\Omega) : q|_T \in \mathcal{P}_k(T), T \in \mathcal{T}_h\} & \text{if } k = 0 \end{cases}.$$

We use $X \lesssim Y$ to denote an inequality $X \leq CY$ with a constant C which depends only on the shape regularity of \mathcal{T}_h and polynomial degree k , and $X \sim Y$ stands for $X \lesssim Y$ and $Y \lesssim X$.

2.2. Enriched Galerkin methods with interior over-penalization

In this subsection we introduce the interior over-penalized enriched Galerkin methods.

For $V_h^c = \mathcal{P}_k(\mathcal{T}_h)$ with $k \geq 1$ and $V_h^0 = \mathcal{P}_0(\mathcal{T}_h)$, we define V_h as the linear space

$$V_h = \{v^c + v^0 : v^c \in V_h^c, v^0 \in V_h^0\}.$$

A function $v \in V_h$ with $v = v^c + v^0$ is determined by the standard degrees of freedom of V_h^c and V_h^0 . However, the degrees of freedom of V_h is not the union of the two sets of degrees of freedom, of V_h^c and of V_h^0 . In fact, a piecewise polynomial function $v \in V_h$ may have more than one expression as a sum of elements in V_h^c and V_h^0 . If we use $v = v^c + v^0$ with $v^c \in V_h^c$, $v^0 \in V_h^0$ to denote an element in V_h , the zero function can have infinitely many expressions by choosing $v^c = -v^0 = C$ for any constant C . However, such non-unique expressions are unique up to constant addition. To see it, assume that $v = v_1^c + v_1^0 = v_2^c + v_2^0$. Then $v_1^c - v_2^c = v_2^0 - v_1^0 \in V_h^c \cap V_h^0$, so v_1^c and v_2^c are the same up to constant addition. So are v_1^0 and v_2^0 .

We use \mathbb{V}_h to denote the space $V_h^c \times V_h^0$. Note that one element in $\phi \in \mathbb{V}_h$ uniquely determines $\phi_c \in V_h^c$ and $\phi_0 \in V_h^0$. Let \mathbb{V}_h^{ker} be the one dimensional subspace $\mathbb{V}_h^{ker} := \{\phi \in \mathbb{V}_h : \phi_c = -\phi_0\}$ in \mathbb{V}_h . If we denote the quotient space $\mathbb{V}_h / \mathbb{V}_h^{ker}$ by \mathbb{V}_h^\perp , there is a one-to-one correspondence between V_h and \mathbb{V}_h^\perp .

To introduce the bilinear form for the enriched Galerkin method, we define the jump and average operators of functions which have well-defined traces on edges/faces in the context of classical discontinuous Galerkin methods. For

$$e \in \mathcal{E}_h^\partial$$

$$[[q]]|_e = q|_e \mathbf{n}, \quad \{\!\!\{q\}\!\!\}|_e = q|_e, \quad [[\mathbf{q}]]|_e = \mathbf{q}|_e \cdot \mathbf{n}, \quad \{\!\!\{\mathbf{q}\}\!\!\}|_e = \mathbf{q}|_e$$

where q and \mathbf{q} are \mathbb{R} - and \mathbb{R}^n -valued functions such that their traces on e are well-defined, and \mathbf{n} is the outward unit normal vector field on e . For $e \in \mathcal{E}_h^0$, let T_+ and T_- be the two elements sharing e as $e = \partial T_+ \cap \partial T_-$. If q is a scalar function on $T_+ \cup T_-$, then we use q^\pm and ∇q^\pm to denote the restrictions of q and ∇q on T_+ and T_- . We use \mathbf{n}^\pm to denote the unit outward normal vector fields of T_\pm and we again omit the restriction $|_e$ if it is clear in context. Then the jumps and the averages of q and \mathbf{q} on $e \in \mathcal{E}_h^0$ are defined by

$$\begin{aligned} [[q]]|_e &= q^+ \mathbf{n}^+ + q^- \mathbf{n}^-, & \{\!\!\{q\}\!\!\}|_e &= \frac{1}{2}(q^+ + q^-), \\ [[\mathbf{q}]]|_e &= \mathbf{q}^+ \cdot \mathbf{n}^+ + \mathbf{q}^- \cdot \mathbf{n}^-, & \{\!\!\{\mathbf{q}\}\!\!\}|_e &= \frac{1}{2}(\mathbf{q}^+ + \mathbf{q}^-). \end{aligned}$$

If $e \in \mathcal{E}_h$ is clear in context, we will use $[[q]]$ instead of $[[q]]|_e$ for the jump of q on $e \in \mathcal{E}_h$. The same simplification will apply to other quantities.

We consider a model second order elliptic equation

$$-\Delta u = f \quad \text{in } \Omega, \tag{1}$$

$$u|_{\Gamma_D} = u_D \text{ on } \Gamma_D, \quad \nabla u \cdot \mathbf{n}|_{\Gamma_N} = u_N \text{ on } \Gamma_N \tag{2}$$

where Γ_D, Γ_N are disjoint open subsets of $\partial\Omega$, the boundary of Ω , such that $\overline{\Gamma_D} \cup \overline{\Gamma_N} = \partial\Omega$.

Let $\gamma > 0$ be a sufficiently large positive constant. We use \mathcal{E}_h^D to denote the set $\{e \in \mathcal{E}_h^\partial : e \subset \Gamma_D\}$. For $V := H^1(\Omega)$, let $a : V \times V_h \rightarrow \mathbb{R}$ be

$$a(v, w) = (\nabla v, \nabla w) + \langle \gamma h_e^{-1} v, w \rangle_{\mathcal{E}_h^D} \tag{3}$$

where

$$\langle \gamma h_e^{-1} v, w \rangle_{\mathcal{E}_h^D} := \sum_{e \in \mathcal{E}_h^D} \langle \gamma h_e^{-1} v, w \rangle_e$$

with h_e , the diameter of $e \in \mathcal{E}_h$. We also define the bilinear form a_h on $V_h \times V_h$ as

$$\begin{aligned} a_h(v, w) &= (\nabla v, \nabla w) - \left(\langle \{\!\!\{\nabla v\}\!\!\}, [[w]] \rangle_{\mathcal{E}_h^0 \cup \mathcal{E}_h^D} + \langle [[v]], \{\!\!\{\nabla w\}\!\!\} \rangle_{\mathcal{E}_h^0 \cup \mathcal{E}_h^D} \right) \\ &\quad + \langle \gamma h_e^{-1-\alpha} [[v]], [[w]] \rangle_{\mathcal{E}_h^0} + \langle \gamma h_e^{-1} [[v]], [[w]] \rangle_{\mathcal{E}_h^D} \end{aligned} \tag{4}$$

for $\alpha \geq 0$. Here ∇ for $v \in V_h$ is the element-wise gradient operator. If $\alpha = 0$, this is the bilinear form of the symmetric interior penalty discontinuous Galerkin method (SIPG), which is used in the enriched Galerkin (EG) method in [1]. In this paper we are interested in the cases $\alpha \geq 1$ such that the interior penalization terms are overly penalized. We will call these methods interior over-penalized

enriched Galerkin methods (IOP-EG). For example, if $\alpha = 1$, then a_h is

$$\begin{aligned} a_h(v, w) = & (\nabla v, \nabla w) - \left(\langle \llbracket \nabla v \rrbracket, \llbracket w \rrbracket \rangle_{\mathcal{E}_h^0 \cup \mathcal{E}_h^D} + \langle \llbracket v \rrbracket, \llbracket \nabla w \rrbracket \rangle_{\mathcal{E}_h^0 \cup \mathcal{E}_h^D} \right) \\ & + \langle \gamma h_e^{-2} \llbracket v \rrbracket, \llbracket w \rrbracket \rangle_{\mathcal{E}_h^0} + \langle \gamma h_e^{-1} \llbracket v \rrbracket, \llbracket w \rrbracket \rangle_{\mathcal{E}_h^D}. \end{aligned}$$

In IOP-EG methods we seek $u_h \in V_h$ such that

$$a_h(u_h, v) = \mathcal{F}(v) \quad \forall v \in V_h \quad (5)$$

for a_h in (4) with

$$\mathcal{F}(v) := (f, v) + \langle u_N, v \rangle_{\Gamma_N} + \langle u_D, \mathbf{n} \cdot \nabla v \rangle_{\Gamma_D} + \langle \gamma h_e^{-1} u_D, v \rangle_{\Gamma_D}. \quad (6)$$

To show the stability of (5), let us define a norm on V_h associated with a_h as

$$\|v\|_{a_h}^2 := (\nabla v, \nabla v) + \langle \gamma h_e^{-1-\alpha} \llbracket v \rrbracket, \llbracket v \rrbracket \rangle_{\mathcal{E}_h^0} + \langle \gamma h_e^{-1} \llbracket v \rrbracket, \llbracket v \rrbracket \rangle_{\mathcal{E}_h^D}. \quad (7)$$

For later use, we also define

$$\|v\|_{a_h,0}^2 := (\nabla v, \nabla v) + \langle \gamma h_e^{-1-\alpha} \llbracket v \rrbracket, \llbracket v \rrbracket \rangle_{\mathcal{E}_h^0}, \quad \|v\|_{a_h,\partial}^2 := \langle \gamma h_e^{-1} \llbracket v \rrbracket, \llbracket v \rrbracket \rangle_{\mathcal{E}_h^D}. \quad (8)$$

By the inverse trace inequality and the standard argument in DG methods, it is not difficult to show

$$|a_h(v, w)| \lesssim \|v\|_{a_h} \|w\|_{a_h}$$

for $v, w \in V_h$. We can obtain that

$$|\langle \llbracket \nabla v \rrbracket, \llbracket v \rrbracket \rangle_{\mathcal{E}_h^0 \cup \mathcal{E}_h^D}| \leq \frac{1}{2} (\nabla v, \nabla v)^{\frac{1}{2}} \langle \gamma h_e^{-1} \llbracket v \rrbracket, \llbracket v \rrbracket \rangle_{\mathcal{E}_h^0 \cup \mathcal{E}_h^D}^{\frac{1}{2}} \quad (9)$$

for sufficiently large γ . Since $h \leq 1$, if γ is sufficiently large, then we can use (9) to derive a coercivity inequality

$$C_{\text{co}} \|v\|_{a_h}^2 \leq a_h(v, v) \quad (10)$$

with $C_{\text{co}} > 0$ independent of $0 < h \leq 1$. Therefore (5) is stable.

3. The a priori error estimates

In this section we prove the a priori error analysis of the IOP-EG methods. For the analysis we adopt the idea of [1] utilizing estimates of the a posteriori error analysis. As a consequence, we prove optimal a priori error estimates with minimal regularity assumption of exact solutions.

We recall efficiency estimates of the a posteriori error analysis in Nitsche's method [1, 1]. We take into account only the data oscillation of f in the discussion below for simplicity. For detailed discussion with data oscillations of

boundary conditions, we refer to [1]. For $v \in V_h^c$ the local efficiency estimates

$$h_T^2 \|f + \Delta v\|_{0,T}^2 \lesssim (\nabla(u-v), \nabla(u-v))_{0,T}^2 + (\text{osc}_T(f))^2, \quad (11)$$

$$h_e \|[\![\nabla v]\!]\|_{0,e}^2 \lesssim \sum_{T \in \mathcal{T}_h, e \subset T} ((\nabla(u-v), \nabla(u-v))_{0,T}^2 + (\text{osc}_T(f))^2) \quad (12)$$

are proved in [1] with

$$\text{osc}_T(f) := h_T \|f - f_h\|_{0,T}, \quad \text{osc}(f) := \left(\sum_{T \in \mathcal{T}_h} (\text{osc}_T(f))^2 \right)^{\frac{1}{2}}$$

where f_h is the L^2 projection of f into V_h^c .

Here we give some definitions and results which are necessary in our a priori error analysis. We will use V_h^{00} to denote the space of mean-value zero piecewise constant functions, i.e.,

$$V_h^{00} := \{v \in V_h^0 : \int_{\Omega} v \, dx = 0\}.$$

Suppose that D is a union of (closed) simplices in \mathcal{T}_h and $\text{int } D$ is the interior of D . A discrete seminorm $|v|_{H_h^1(D)}$ for element-wise H^1 function v is defined as

$$|v|_{H_h^1(D)}^2 = \|\nabla v\|_{L^2(D)}^2 + \sum_{e \in \mathcal{E}_h^0, e \subset \text{int } D} h_e^{-1} \|[\![v]\!]\|_{0,e}^2. \quad (13)$$

If $D = \overline{\Omega}$, we simply use H_h^1 instead of $H_h^1(\overline{\Omega})$.

For a vertex z in the triangulation \mathcal{T}_h , $T \in \mathcal{T}_h$, $e \in \mathcal{E}_h$, we define

$$M_z = \bigcup_{z \in T} \{T \in \mathcal{T}_h : z \in T\}, \quad M_T = \bigcup_{z \in T} M_z, \quad M_e = \bigcup_{T \in \mathcal{T}_h, e \subset \partial T} M_T. \quad (14)$$

Geometrically, these are the union of simplices which become proper neighborhoods whose interiors contain z , T , and e , respectively. For $v \in V_h^0$ it is known that there is a linear interpolation operator $I_h : V_h^0 \rightarrow V_h^c$ such that,

$$\|v - I_h v\|_{L^2(T)} \lesssim h_T |v|_{H_h^1(M_T)} \quad \forall T \in \mathcal{T}_h, \quad (15)$$

$$\|v - I_h v\|_{L^2(e)} \lesssim h_e^{\frac{1}{2}} |v|_{H_h^1(M_e)} \quad \forall e \in \mathcal{E}_h^0 \cup \mathcal{E}_h^D, \quad (16)$$

$$\|\nabla I_h v\|_{L^2(T)} \lesssim |v|_{H_h^1(M_T)} \quad \forall T \in \mathcal{T}_h \quad (17)$$

hold, where the implicit constants in these estimates are independent of the mesh sizes (see [2, Theorem 3.1]).

We define \mathcal{T}_h^D and $\mathcal{E}_h(\mathcal{T}_h^D)$ as

$$\mathcal{T}_h^D := \{T \in \mathcal{T}_h : \partial T \cap \partial\Omega \subset \Gamma_D\}, \quad (18)$$

$$\mathcal{E}_h(\mathcal{T}_h^D) := \{e \in \mathcal{E}_h : e \subset \partial T, \text{ for some } T \in \mathcal{T}_h^D\}. \quad (19)$$

Lemma 1. *For $v^0 \in V_h^{00}$ it holds that*

$$\|v^0\|_{a_h, \partial} \leq C \max_{e \in \mathcal{E}_h^D} \{h_e^{-\frac{1}{2}}\} \max_{e \in \mathcal{E}_h^0} \{h_e^{\alpha/2}\} \|v^0\|_{a_h, 0} \quad (20)$$

with C depending on the shape regularity, the constants of the trace theorem and the discrete Poincaré inequality.

Proof. Suppose that $T \in \mathcal{T}_h^D$ and $e \in \mathcal{E}_h^D$ with $e \subset \partial T$, are fixed. If we set $z \in V_h^{00}$ as $z = h_e^{-1/2} v^0$, then, by (16),

$$\begin{aligned} \langle h_e^{-1} v^0, v^0 \rangle_e &= \|z\|_{L^2(e)}^2 \leq 2 \|I_h z\|_{L^2(e)}^2 + C h_e |z|_{H_h^1(M_e)}^2 \\ &\leq 2 h_e^{-1} \|I_h v^0\|_{L^2(e)}^2 + C |v^0|_{H_h^1(M_e)}^2. \end{aligned}$$

Applying a similar argument for every such pair (T, e) , $T \in \mathcal{T}_h^D$,

$$\|v^0\|_{a_h, \partial}^2 \leq 2 \max_{e \in \mathcal{E}_h^D} \{h_e^{-1}\} \gamma \|I_h v^0\|_{L^2(\Gamma_D)}^2 + C \gamma |v^0|_{H_h^1}^2. \quad (21)$$

By the trace theorem and the triangle inequality, we have

$$\begin{aligned} \|I_h v^0\|_{L^2(\Gamma_D)} &\lesssim \|I_h v^0\|_{L^2(\Omega)} + \|\nabla I_h v^0\|_{L^2(\Omega)} \\ &\leq \|I_h v^0 - v^0\|_{L^2(\Omega)} + \|v^0\|_{L^2(\Omega)} + \|\nabla I_h v^0\|_{L^2(\Omega)}. \end{aligned}$$

By a discrete Poincaré inequality (cf. [2, Remark 1.1]), $\|v^0\|_{L^2(\Omega)} \lesssim |v^0|_{H_h^1}$ because $v^0 \in V_h^{00}$. Applying this discrete Poincaré inequality, (15), and (17), to the above inequality, we get

$$\|I_h v^0\|_{L^2(\Gamma_D)} \lesssim |v^0|_{H_h^1}.$$

From this, (21), and $|v^0|_{H_h^1} \leq \max_{e \in \mathcal{E}_h^0} \gamma^{-1/2} h_e^{\alpha/2} \|v^0\|_{a_h, 0}$, we can obtain

$$\|v^0\|_{a_h, \partial} \leq C \max_{e \in \mathcal{E}_h^D} \{h_e^{-\frac{1}{2}}\} \max_{e \in \mathcal{E}_h^0} \{h_e^{\alpha/2}\} \|v^0\|_{a_h, 0}.$$

□

In the theorem below, we assume that $\max_{e \in \mathcal{E}_h^D} \{h_e^{-\frac{1}{2}}\} \max_{e \in \mathcal{E}_h^0} \{h_e^{\alpha/2}\}$ is bounded by $C h^{(\alpha-1)/2}$ with $C > 0$ independent of the mesh sizes. This condition is fulfilled, for example, if \mathcal{T}_h is quasi-uniform. Under this assumption, we prove the a priori error estimates for solutions $u \in H^s(\Omega)$, $s \geq 1$, of (1)–(2).

Theorem 1. Suppose that u and u_h are the solutions of (1)–(2) and (5). Assume that $\alpha \geq 1$ and

$$\max_{e \in \mathcal{E}_h^D} \{h_e^{-\frac{1}{2}}\} \max_{e \in \mathcal{E}_h^0} \{h_e^{\alpha/2}\} \leq Ch^{(\alpha-1)/2}$$

with $C > 0$ independent of the mesh sizes. Then,

$$\|u - u_h\|_{a_h} \leq C \left(\inf_{v \in V_h^c} \|u - v\|_1 + \text{osc}(f) \right) \quad (22)$$

with $C > 0$ independent of the mesh sizes. In addition, the interior jump terms satisfy

$$|u_h^0|_{H_h^1} \lesssim h^\alpha \left(\inf_{v \in V_h^c} \|u - v\|_1 + \text{osc}(f) \right). \quad (23)$$

We remark that (22) can be replaced by $\|u - u_h\|_{a_h} \leq Ch^s \|u\|_{s+1}$, a more common form of error estimate, if $1 \leq s \leq k+1$ because $\|f\|_{s-1} \leq C \|u\|_{s+1}$.

Proof. We define $\Pi_h^c : H^1(\Omega) \rightarrow V_h^c$ by the linear map seeking $w_h^c \in V_h^c$ such that

$$a(w_h^c, v) = \mathcal{F}(v) \quad \forall v \in V_h^c. \quad (24)$$

Further, we define $\Pi_h^0 : (V_h^{00})^* \rightarrow V_h^{00}$ by the linear map seeking the solution $w_h^0 \in V_h^{00}$ of

$$a_h(w_h^0, v) = a_h(\Pi_h^c u, v) - \mathcal{F}(v) \quad \forall v \in V_h^{00}. \quad (25)$$

We first claim that for $v^0 \in V_h^{00}$

$$|a_h(\Pi_h^0 u, v^0)| \lesssim \left(\inf_{w \in V_h^c} \|u - w\|_1 + \text{osc}(f) \right) |v^0|_{H_h^1}. \quad (26)$$

To prove it, we use (25), the Galerkin orthogonality, and the integration by parts to get

$$\begin{aligned} a_h(\Pi_h^0 u, v^0) &= a_h(\Pi_h^c u, v^0 - I_h v^0) - \mathcal{F}(v^0 - I_h v^0) \\ &= - \sum_{T \in \mathcal{T}_h} (f + \nabla \cdot (\nabla \Pi_h^c u), v^0 - I_h v^0) \\ &\quad - \sum_{e \in \mathcal{E}_h^0 \cup \mathcal{E}_h^D} \langle \llbracket \nabla \Pi_h^c u \rrbracket, v^0 - I_h v^0 \rangle_e. \end{aligned}$$

By (15) and (16), we have

$$\begin{aligned}
(f + \Delta \Pi_h^c u, v^0 - I_h v^0)_T &= \|f + \Delta \Pi_h^c u\|_{0,T} \|v^0 - I_h v^0\|_{0,T} \\
&\lesssim h_T \|f + \Delta \Pi_h^c u\|_{0,T} |v^0|_{H_h^1(M_T)}, \\
\langle \llbracket \nabla \Pi_h^c u \rrbracket, v^0 - I_h v^0 \rangle_e &\leq \|\llbracket \nabla \Pi_h^c u \rrbracket\|_{0,e} \|v^0 - I_h v^0\|_{0,e} \\
&\lesssim h_e^{\frac{1}{2}} \|\llbracket \nabla \Pi_h^c u \rrbracket\|_{0,e} |v^0|_{H_h^1(M_e)}.
\end{aligned}$$

The elements in $\{M_T\}$ and $\{M_e\}$ overlap only finitely many times with a uniform finite number, so the Cauchy–Schwarz inequality and the a posteriori error estimates (11)–(12) yield (26).

Note that

$$|v^0|_{H_h^1}^2 \lesssim h^\alpha \|v^0\|_{a_h,0}^2 \lesssim h^\alpha \|v^0\|_{a_h}^2. \quad (27)$$

By this and by taking $v^0 = \Pi_h^0 u$ in (26),

$$h^{-\alpha} |\Pi_h^0 u|_{H_h^1}^2 \lesssim \|\Pi_h^0 u\|_{a_h,0}^2 \leq \|\Pi_h^0 u\|_{a_h}^2 \lesssim |\Pi_h^0 u|_{H_h^1} \left(\inf_{w \in V_h^c} \|u - w\|_1 + \text{osc}(f) \right),$$

so we have

$$|\Pi_h^0 u|_{H_h^1} \lesssim h^\alpha \left(\inf_{w \in V_h^c} \|u - w\|_1 + \text{osc}(f) \right), \quad (28)$$

$$\|\Pi_h^0 u\|_{a_h} \lesssim h^{\alpha/2} \left(\inf_{w \in V_h^c} \|u - w\|_1 + \text{osc}(f) \right). \quad (29)$$

We now define $\Pi_h : H^1(\Omega) \rightarrow V_h$ as $\Pi_h = \Pi_h^c + \Pi_h^0$. Then

$$\begin{aligned}
&a_h(u_h - \Pi_h u, u_h - \Pi_h u) \\
&= \mathcal{F}(u_h - \Pi_h u) - a_h(\Pi_h u, u_h - \Pi_h u) \\
&= \mathcal{F}(u_h^c - \Pi_h^c u) + \mathcal{F}(u_h^0 - \Pi_h^0 u) - a_h(\Pi_h^c u, u_h^c - \Pi_h^c u) \\
&\quad - a_h(\Pi_h^0 u, u_h^c - \Pi_h^c u) - a_h(\Pi_h^c u, u_h^0 - \Pi_h^0 u) - a_h(\Pi_h^0 u, u_h^0 - \Pi_h^0 u).
\end{aligned}$$

Since

$$\begin{aligned}
\mathcal{F}(u_h^c - \Pi_h^c u) - a_h(\Pi_h^c u, u_h^c - \Pi_h^c u) &= 0, \\
\mathcal{F}(u_h^0 - \Pi_h^0 u) - a_h(\Pi_h^c u, u_h^0 - \Pi_h^0 u) &= -a_h(\Pi_h^0 u, u_h^0 - \Pi_h^0 u),
\end{aligned}$$

by (24) and (25), we have

$$\begin{aligned}
& |a_h(u_h - \Pi_h u, u_h - \Pi_h u)| \\
&= |a_h(\Pi_h^0 u, (\Pi_h^c u - u_h^c)) + 2a_h(\Pi_h^0 u, (\Pi_h^0 u - u_h^0))| \\
&= |a_h(\Pi_h^0 u, \Pi_h u - u_h) + a_h(\Pi_h^0 u, \Pi_h^0 u - u_h^0)| \\
&\lesssim \|\Pi_h^0 u\|_{a_h} \|\Pi_h u - u_h\|_{a_h} + \|\Pi_h^0 u\|_{a_h} \|\Pi_h^0 u - u_h^0\|_{a_h}.
\end{aligned} \tag{30}$$

By Lemma 1, $\|\Pi_h^0 u - u_h^0\|_{a_h, \partial} \lesssim h^{(\alpha-1)/2} \|\Pi_h^0 u - u_h^0\|_{a_h, 0}$. Since

$$\|\Pi_h^0 u - u_h^0\|_{a_h, 0} \leq \|\Pi_h u - u_h\|_{a_h, 0} \leq \|\Pi_h u - u_h\|_{a_h},$$

we have

$$\|\Pi_h^0 u - u_h^0\|_{a_h} \leq \|\Pi_h^0 u - u_h^0\|_{a_h, 0} + \|\Pi_h^0 u - u_h^0\|_{a_h, \partial} \lesssim \|\Pi_h u - u_h\|_{a_h},$$

therefore we can obtain

$$\|u_h - \Pi_h u\|_{a_h} \lesssim \|\Pi_h^0 u\|_{a_h}$$

from (30). If we use (29), then one can obtain with the coercivity of a_h that

$$\|u_h - \Pi_h u\|_{a_h} \lesssim h^{\alpha/2} \left(\inf_{v \in V_h^c} \|u - v\|_1 + \text{osc}(f) \right).$$

In particular, by $|u_h^0 - \Pi_h^0 u|_{H_h^1} \lesssim h^{\alpha/2} \|u_h^0 - \Pi_h^0 u\|_{a_h, 0} \lesssim h^{\alpha/2} \|u_h - \Pi_h u\|_{a_h, 0}$, we obtain

$$|u_h^0 - \Pi_h^0 u|_{H_h^1} \lesssim h^{\alpha} \left(\inf_{v \in V_h^c} \|u - v\|_1 + \text{osc}(f) \right),$$

therefore (23) follows from this, (28), and the triangle inequality.

To complete the proof we estimate $\|u - \Pi_h u\|_{a_h}$. By the triangle inequality and (29),

$$\|u - \Pi_h u\|_{a_h} \leq \|u - \Pi_h^c u\|_{a_h} + \|\Pi_h^0 u\|_{a_h} \lesssim \inf_{w \in V_h^c} \|u - w\|_1 + \text{osc}(f)$$

where we used the estimate $\|u - \Pi_h^c u\|_{a_h} \lesssim \inf_{w \in V_h^c} \|u - w\|_1 + \text{osc}(f)$ which holds because $\Pi_h^c u$ is the solution of Nitsche's method of the problem (24). \square

Recall that the flux is $\mathbf{z} := -\nabla u$ by Darcy's law. We now discuss recovery of locally mass conservative numerical flux \mathbf{z}_h via local post-processing and prove an estimate of $\mathbf{z} - \mathbf{z}_h$. The key result is that the estimate is robust for the over-penalization.

We use $H(\text{div}; \Omega)$ to denote the subspace of $L^2(\Omega; \mathbb{R}^n)$ such that the function and its divergence are square-integrable.

Definition 1. We say $\mathbf{z}_h \in H(\text{div}; \Omega)$ is locally mass conservative if

$$(\text{div } \mathbf{z}_h, 1_T) = (f, 1_T)$$

for any $T \in \mathcal{T}_h$ where 1_T is the indicator function on T .

For construction of locally mass conservative numerical flux we use the Raviart–Thomas–Nédélec element of order k , N_h , which has local shape functions

$$\mathcal{P}_{k-1}(T; \mathbb{R}^n) + \mathbf{x}\mathcal{P}_{k-1}(T)$$

where \mathbf{x} is the vector-valued polynomial $\mathbf{x} = (x \ y)^T$ if $n = 2$ and $\mathbf{x} = (x \ y \ z)^T$ if $n = 3$.

Theorem 2. If we define $\mathbf{z}_h \in N_h$ as

$$\begin{aligned} (\mathbf{z}_h, \mathbf{q})_T &= (-\nabla u_h, \mathbf{q})_T, & \mathbf{q} &\in \mathcal{P}_{k-2}(T), \\ \langle \mathbf{z}_h \cdot \mathbf{n}, q \rangle_e &= \langle -\{\!\!\{ \nabla u_h \}\!\!\} \cdot \mathbf{n} + \gamma h_e^{-1-\alpha} \llbracket u_h \rrbracket \cdot \mathbf{n}, q \rangle_e, & q &\in \mathcal{P}_{k-1}(e), e \in \mathcal{E}_h^0, \\ \langle \mathbf{z}_h \cdot \mathbf{n}, q \rangle_e &= \langle u_N, q \rangle_e, & q &\in \mathcal{P}_{k-1}(e), e \notin \mathcal{E}_h^0 \cup \mathcal{E}_h^D, \\ \langle \mathbf{z}_h \cdot \mathbf{n}, q \rangle_e &= \langle -\nabla u_h \cdot \mathbf{n} + \gamma h_e^{-1} (u_h - u_D), q \rangle_e, & q &\in \mathcal{P}_{k-1}(e), e \in \mathcal{E}_h^D, \end{aligned}$$

then

$$\|\mathbf{z}_h + \nabla u_h\|_0 \lesssim \inf_{w \in V_h^c} \|u - w\|_1 + \text{osc}(f)$$

with an implicit constant independent of h .

Proof. By definition \mathbf{z}_h is in $H(\text{div}; \Omega)$ and

$$\begin{aligned} (\text{div } \mathbf{z}_h, 1_T) &= \sum_{e \subset \partial T, e \in \mathcal{E}_h^D} \langle -\nabla u_h \cdot \mathbf{n} + \gamma h_e^{-1} (u_h - u_D), 1_T \rangle_e \\ &\quad + \sum_{e \subset \partial T, e \in \mathcal{E}_h^0} \langle -\{\!\!\{ \nabla u_h \}\!\!\} + \gamma h_e^{-1-\alpha} \llbracket u_h \rrbracket \cdot \mathbf{n}, 1_T \rangle_e \\ &\quad + \sum_{e \subset \partial T, e \notin \mathcal{E}_h^0 \cup \mathcal{E}_h^D} \langle u_N, 1_T \rangle_e. \end{aligned}$$

If we take $v \in V_h$ in (5) as the indicator function 1_T and use the definition of a_h in (4), then

$$(\text{div } \mathbf{z}_h, 1_T) = (f, 1_T),$$

so \mathbf{z}_h is locally mass conservative.

From the definition of \mathbf{z}_h , $\mathbf{z}_h + \nabla u_h$ satisfies

$$(\mathbf{z}_h + \nabla u_h, \mathbf{q})_T = 0, \quad \mathbf{q} \in \mathcal{P}_{k-2}(T),$$

and

$$\langle (\mathbf{z}_h + \nabla u_h) \cdot \mathbf{n}, q \rangle_e = \begin{cases} \langle -\frac{1}{2} \llbracket \nabla u_h \rrbracket + \gamma h_e^{-1-\alpha} \llbracket u_h \rrbracket \cdot \mathbf{n}, q \rangle_e, & e \in \mathcal{E}_h^0 \\ 0, & e \notin \mathcal{E}_h^0 \cup \mathcal{E}_h^D \\ \langle \gamma h_e^{-1} (u_h - u_D), q \rangle_e, & e \in \mathcal{E}_h^D \end{cases} \quad (31)$$

for $q \in \mathcal{P}_{k-1}(e)$. For $e \in \mathcal{E}_h^0$, the standard scaling argument gives

$$\|\mathbf{z}_h + \nabla u_h\|_{0,T} \lesssim \sum_{e \subset \partial T} h_e^{\frac{1}{2}} \|(\mathbf{z}_h + \nabla u_h) \cdot \mathbf{n}\|_{0,e}.$$

We now estimate the above term for three cases of $e \in \mathcal{E}_h$. If $e \notin \mathcal{E}_h^0 \cup \mathcal{E}_h^D$, then there is nothing to estimate by (31). If $e \in \mathcal{E}_h^0$, then

$$\begin{aligned} \|\mathbf{z}_h + \nabla u_h\|_{0,T} &\lesssim \sum_{e \subset \partial T} h_e^{\frac{1}{2}} \left\| -\frac{1}{2} \llbracket \nabla u_h \rrbracket + \gamma h_e^{-1-\alpha} \llbracket u_h \rrbracket \cdot \mathbf{n} \right\|_{0,e} \\ &\lesssim \sum_{e \subset \partial T} h_e^{\frac{1}{2}} \left(\|\llbracket \nabla u_h \rrbracket\|_{0,e} + \gamma h_e^{-1-\alpha} \|\llbracket u_h \rrbracket\|_{0,e} \right), \end{aligned}$$

so we need to show

$$\begin{aligned} &\left(\sum_{e \in \mathcal{E}_h^0} h_e \|\llbracket \nabla u_h \rrbracket\|_{0,e}^2 \right)^{\frac{1}{2}} + \left(\sum_{e \in \mathcal{E}_h^0} h_e^{-1-2\alpha} \|\llbracket u_h \rrbracket\|_{0,e}^2 \right)^{\frac{1}{2}} \\ &\lesssim \inf_{v \in V_h^c} \|u - v\|_1 + \text{osc}(f). \end{aligned} \quad (32)$$

The first term in this inequality is easily obtained by (12). For the second term note that the quasiuniformity assumption implies

$$\sum_{e \in \mathcal{E}_h^0} h_e^{-1-2\alpha} \|\llbracket u_h \rrbracket\|_{0,e}^2 \lesssim h^{-2\alpha} |u_h|_{H_h^1}^2 \lesssim \left(\inf_{v \in V_h^c} \|u - v\|_1 + \text{osc}(f) \right)^2$$

by (23). For $e \in \mathcal{E}_h^D$, the scaling argument gives

$$\|\mathbf{z}_h + \nabla u_h\|_{0,T} \lesssim \sum_{e \subset \partial T} h_e^{\frac{1}{2}} \|\gamma h_e^{-1} (u_h - u_D)\|_{0,e} \lesssim \|u - u_h\|_{a_h}.$$

Therefore, the conclusion follows. \square

4. Preconditioning

In this section we propose an abstract form of block diagonal preconditioners. Through the operator preconditioning approach, we show that the preconditioners are spectrally equivalent to the matrix given by (5) independent of mesh sizes.

We define a norm $\|\cdot\|$ on \mathbb{V}_h as

$$\begin{aligned} \| (v^c, v^0) \|^2 &:= a_h(v^c, v^c) + a_h(v^0, v^0) \\ &= (\nabla v^c, \nabla v^c) + \langle \gamma h_e^{-1} v^c, v^c \rangle_{\mathcal{E}_h^D} \\ &\quad + \langle \gamma h_e^{-1-\alpha} \llbracket v^0 \rrbracket, \llbracket v^0 \rrbracket \rangle_{\mathcal{E}_h^0} + \langle \gamma h_e^{-1} v^0, v^0 \rangle_{\mathcal{E}_h^D} \end{aligned} \quad (33)$$

for $v^c \in V_h^c$, $v^0 \in V_h^0$. Note that this is a norm on \mathbb{V}_h , not only on V_h . For an element $v \in V_h$ the decomposition $v = v^c + v^0$ is not unique, so $\|v\|$ is not uniquely defined. However, we will use $\|v^c\|$ and $\|v^0\|$ for $v^c \in V_h^c$ and $v^0 \in V_h^0$ instead of $\|(v^c, 0)\|$ and $\|(0, v^0)\|$ for convenience even though these quantities are not norms.

Recall that \mathbb{V}_h^\perp and V_h have a one-to-one correspondence, so we will use v to denote an element in V_h and its corresponding element in \mathbb{V}_h^\perp . Let $\mathcal{A}_h : \mathbb{V}_h^\perp \rightarrow (\mathbb{V}_h^\perp)^*$ be the linear operator defined by $\langle \mathcal{A}_h v, w \rangle_{\langle \mathbb{V}_h, \mathbb{V}_h^* \rangle} = a_h(v, w)$. Following the operator preconditioning approach [1], we first show that \mathcal{A}_h is a bounded isomorphism from \mathbb{V}_h^\perp to $(\mathbb{V}_h^\perp)^*$ with the norm $\|\cdot\|$.

For preconditioning we consider the linear operator $\tilde{\mathcal{A}}_h : \mathbb{V}_h \rightarrow \mathbb{V}_h^*$ defined by the bilinear form

$$a_h(v^c, v^c) + a_h(v^0, v^0),$$

and show that $\tilde{\mathcal{A}}_h$ is spectrally equivalent to \mathcal{A}_h on \mathbb{V}_h^\perp with $\|\cdot\|$. As a consequence, we can expect that $\mathcal{B}_h = \tilde{\mathcal{A}}_h^{-1}$ is a good preconditioner of \mathcal{A}_h .

It is obvious that \mathcal{A}_h gives a bounded bilinear form on $\mathbb{V}_h^\perp \times \mathbb{V}_h^\perp$ with $\|\cdot\|$ norm. To show that $\tilde{\mathcal{A}}_h$ is spectrally equivalent to \mathcal{A}_h on \mathbb{V}_h^\perp , we prove

$$\inf_{v=v^c+v^0 \in V_h} \sup_{w=w^c+w^0 \in V_h} \frac{a_h(v, w)}{\| (v^c, v^0) \| \| (w^c, w^0) \|} \geq C > 0.$$

We remark that the coercivity of $a_h(\cdot, \cdot)$ for $\|\cdot\|_{a_h}$ does not imply this condition because $\|v\|_{a_h} \lesssim \| (v^c, v^0) \|$ for $v = v^c + v^0$ but the inequality of the other direction is not true in general.

Definition 2. For $T \in \mathcal{T}_h$, we define $N(T)$ as the number of elements adjacent to T , i.e., $N(T) = |\{T' \in \mathcal{T}_h : \partial T' \cap \partial T \neq \emptyset, T' \neq T\}|$. For $v^0 \in V_h^0$, $T_0 \in \mathcal{T}_h$, and the elements $\{T_i\}_{i=1, \dots, N(T_0)}$ adjacent to T_0 , let p_i be the value of $v^0|_{T_i}$,

$i = 0, \dots, N(T_0)$. We define $\phi_{v^0, T_0} \in V_h^0$ as

$$\phi_{v^0, T_0}|_T = \begin{cases} \frac{1}{N(T_0)+1} \sum_{i=0}^{N(T_0)} (p_i - p_0) & \text{if } T = T_0 \\ 0 & \text{otherwise} \end{cases}. \quad (34)$$

Lemma 2. Suppose that $\alpha \geq 1$. Given $v^0 \in V_h^{00}$, define \bar{v}^0 as $\bar{v}^0 := \sum_{T \in \mathcal{T}_h^D} \phi_{v^0, T}$. Then

$$\|\bar{v}^0\|_{a_h, \partial} \leq \|\bar{v}^0\| \leq C_0 \|v^0\|_{a_h, 0}. \quad (35)$$

with $C_0 > 0$ independent of h . As a corollary, $\tilde{v}^0 := v^0 + \bar{v}^0$ satisfies

$$\|\tilde{v}^0\|_{a_h, 0} \leq (1 + C_0) \|v^0\|_{a_h, 0}, \quad (36)$$

$$\|\tilde{v}^0\|_{a_h, \partial} \lesssim \|v^0\|_{a_h, 0} \quad (37)$$

with an implicit constant depending on the shape regularity of \mathcal{T}_h , the constants of the trace theorem and discrete Poincaré inequality.

Proof. Let T_0 be an element in \mathcal{T}_h^D . For simplicity of presentation we assume that T_0 has only one facet in $\partial T_0 \cap \Gamma_D$ but extension of the arguments below to more general cases is straightforward.

Assume that $e_0 = \partial T_0 \cap \Gamma_D$, and $\{T_i\}_{i=1}^{N(T_0)}$ are the elements adjacent to T_0 such that $e_i = \partial T_i \cap \partial T_0$. Denoting the value of v^0 on T_i by p_i , $i = 0, \dots, N(T_0)$, $\llbracket v^0 \rrbracket$ on each e_i is $|p_i - p_0|$. By the definition of ϕ_{v^0, T_0}

$$|\llbracket \phi_{v^0, T_0} \rrbracket_{e_j}| = \left| \frac{1}{N(T_0)+1} \sum_{i=1}^{N(T_0)} (p_i - p_0) \right| \leq \frac{1}{N(T_0)+1} \sum_{i=1}^{N(T_0)} |p_i - p_0|$$

for $0 \leq j \leq N(T_0)$, so the Cauchy-Schwarz inequality and the definition of $\|\cdot\|$ give

$$\begin{aligned} \|\phi_{v^0, T_0}\|^2 &= \sum_{e \subset \partial T_0, e \in \mathcal{E}_h^0} \gamma h_e^{-1-\alpha} \|\llbracket \phi_{v^0, T_0} \rrbracket\|_{0,e}^2 + \gamma h_{e_0}^{-1} \|\phi(v^0, T_0)\|_{0,e_0}^2 \\ &\lesssim \gamma \sum_{i=1}^{N(T_0)} h_{e_i}^{-1-\alpha} \|\llbracket v^0 \rrbracket\|_{e_i}^2, \\ \|\phi_{v^0, T_0}\|_{a_h, \partial}^2 &= \gamma h_{e_0}^{-1} \|\phi(v^0, T_0)\|_{0,e_0}^2 \lesssim h^\alpha \sum_{i=1}^{N(T_0)} \gamma h_{e_i}^{-1-\alpha} \|\llbracket v^0 \rrbracket\|_{0,e_i}^2. \end{aligned}$$

By the definition of \bar{v}^0 , $\bar{v}^0|_T = \phi(v^0, T)$ for all $T \in \mathcal{T}_h^D$ and $\bar{v}^0|_T = 0$ for $T \notin \mathcal{T}_h^D$. Since each facet is shared at most by two elements, the triangle inequality with

the above argument gives

$$\|\bar{v}^0\|^2 \lesssim \|v^0\|_{a_h,0}^2, \quad (38)$$

which also implies $\|\bar{v}^0\|_{a_h,\partial}^2 \lesssim \|v^0\|_{a_h,0}^2$. Then, (36) follows by the triangle inequality $\|\tilde{v}^0\|_{a_h,0} \leq \|v^0\|_{a_h,0} + \|\bar{v}^0\|_{a_h,0}$ and the inequality $\|\bar{v}^0\|_{a_h,0} \leq \|\bar{v}^0\|$. (37) follows by $\|\tilde{v}^0\|_{a_h,\partial} \leq \|v^0\|_{a_h,\partial} + \|\bar{v}^0\|_{a_h,\partial}$, (20) and (38). \square

Theorem 3. *Assume that the assumptions in Lemma 2 hold. If $\alpha \geq 1$, then there exists $C > 0$ such that*

$$\inf_{v=v^c+v^0 \in V_h} \sup_{w=w^c+w^0 \in V_h} \frac{a_h(v, w)}{\|(v^c, v^0)\| \|(w^c, w^0)\|} \geq C > 0. \quad (39)$$

Proof. For $v \in V_h$ let $v = v^c + v^0$ for $v^c \in V_h^c$ and $v^0 \in V_h^{00}$. Suppose that \bar{v}^0 and \tilde{v}^0 are defined as in Lemma 2. By (35), (36), (37), and (20) there exists C_1 independent of h satisfying

$$\|\bar{v}^0\|_{a_h,\partial}, \|\tilde{v}^0\|_{a_h,0}, \|\tilde{v}^0\|_{a_h,\partial}, \|v^0\|_{a_h,\partial} \leq C_1 \|v^0\|_{a_h,0}. \quad (40)$$

We take

$$w := w^c + w^0, \quad w^c := v^c, \quad w^0 := \delta \tilde{v}^0 = v^0 + \delta(v^0 + \bar{v}^0)$$

with $\delta > 0$ which will be determined later. From the definition (33) and (35), it holds that

$$\|(w^c, w^0)\|^2 = \|(v^c, v^0 + \delta \tilde{v}^0)\|^2 = \|v^c\|^2 + \|v + \delta \tilde{v}^0\|^2 \leq C \|(v^c, v^0)\|^2 \quad (41)$$

with C which is uniformly bounded for $\delta \leq 1$.

From the definition of w and the coercivity of a_h

$$\begin{aligned}
a_h(v, w) &= a_h(v, v) + \delta a_h(v, \tilde{v}^0) \\
&\geq C_{\text{co}} \left(\|v^c\|_{a_h,0}^2 + \|v^0\|_{a_h,0}^2 + \|v^0 + v^c\|_{a_h,\partial}^2 \right) \\
&\quad - \delta \langle \llbracket \nabla v^c \rrbracket, \llbracket \tilde{v}^0 \rrbracket \rangle_{\mathcal{E}_h^0 \cup \mathcal{E}_h^D} + \delta \langle \gamma h_e^{-1-\alpha} \llbracket v^0 \rrbracket, \llbracket \tilde{v}^0 \rrbracket \rangle_{\mathcal{E}_h^0} \\
&\quad + \delta \langle \gamma h_e^{-1} (v^0 + v^c), \tilde{v}^0 \rangle_{\mathcal{E}_h^D} \\
&= C_{\text{co}} \| (v^c, v^0) \|^2 + 2C_{\text{co}} \langle \gamma h_e^{-1} v^c, v^0 \rangle_{\mathcal{E}_h^D} \\
&\quad - \delta \langle \llbracket \nabla v^c \rrbracket, \llbracket \tilde{v}^0 \rrbracket \rangle_{\mathcal{E}_h^0 \cup \mathcal{E}_h^D} + \delta \langle \gamma h_e^{-1-\alpha} \llbracket v^0 \rrbracket, \llbracket \tilde{v}^0 \rrbracket \rangle_{\mathcal{E}_h^0} \\
&\quad + \delta \langle \gamma h_e^{-1} v^0, \tilde{v}^0 \rangle_{\mathcal{E}_h^D} + \delta \langle \gamma h_e^{-1} v^c, v^0 + \bar{v}^0 \rangle_{\mathcal{E}_h^D} \quad (\text{because } \tilde{v}^0 = v^0 + \bar{v}^0) \\
&= C_{\text{co}} \| (v^c, v^0) \|^2 + (2C_{\text{co}} + \delta) \langle \gamma h_e^{-1} v^c, v^0 \rangle_{\mathcal{E}_h^D} + \delta \langle \gamma h_e^{-1} v^c, \bar{v}^0 \rangle_{\mathcal{E}_h^D} \\
&\quad - \delta \langle \llbracket \nabla v^c \rrbracket, \llbracket \tilde{v}^0 \rrbracket \rangle_{\mathcal{E}_h^0 \cup \mathcal{E}_h^D} + \delta \langle \gamma h_e^{-1-\alpha} \llbracket v^0 \rrbracket, \llbracket \tilde{v}^0 \rrbracket \rangle_{\mathcal{E}_h^0} + \delta \langle \gamma h_e^{-1} v^0, \bar{v}^0 \rangle_{\mathcal{E}_h^D} \\
&=: C_{\text{co}} \| (v^c, v^0) \|^2 + (2C_{\text{co}} + \delta) I_1 + \delta (I_2 + I_3 + I_4 + I_5)
\end{aligned}$$

with

$$\begin{aligned}
I_1 &:= \langle \gamma h_e^{-1} v^c, v^0 \rangle_{\mathcal{E}_h^D} & I_2 &:= \langle \gamma h_e^{-1} v^c, \bar{v}^0 \rangle_{\mathcal{E}_h^D} & I_3 &:= \langle \llbracket \nabla v^c \rrbracket, \llbracket \tilde{v}^0 \rrbracket \rangle_{\mathcal{E}_h^0 \cup \mathcal{E}_h^D}, \\
I_4 &:= \langle \gamma h_e^{-1-\alpha} \llbracket v^0 \rrbracket, \llbracket \tilde{v}^0 \rrbracket \rangle_{\mathcal{E}_h^0} & I_5 &:= \langle \gamma h_e^{-1} v^0, \bar{v}^0 \rangle_{\mathcal{E}_h^D}.
\end{aligned}$$

For simplicity we introduce additional notations

$$X_{c,\partial} = \|v^c\|_{a_h,\partial}, \quad X_{c,0} = \|v^c\|_{a_h,0}, \quad X_{0,\partial} = \|v^0\|_{a_h,\partial}, \quad X_{0,0} = \|v^0\|_{a_h,0}.$$

By the Cauchy–Schwarz inequality

$$\begin{aligned}
|I_1| &\leq X_{c,\partial} X_{0,\partial} \leq C_1 X_{c,\partial} X_{0,0}, \quad (\text{by (40)}) \\
|I_2| &\leq X_{c,\partial} \|\bar{v}^0\|_{a_h,\partial} \leq C_1 X_{c,\partial} X_{0,0}, \quad (\text{by (40)}) \\
|I_3| &\leq C_2 X_{c,\partial} (X_{0,\partial} + X_{0,0}) \leq C_2 (1 + C_1) X_{c,\partial} X_{0,0}, \quad (\text{by (9) and (40)}) \\
|I_4| &\leq C_2 X_{0,0}^2. \quad (\text{by (40)})
\end{aligned}$$

Moreover,

$$I_5 = X_{0,\partial}^2 + \langle \gamma h_e^{-1} v^0, \bar{v}^0 \rangle_{\mathcal{E}_h^D} \geq X_{0,\partial}^2 - C_1 X_{0,\partial} X_{0,0}. \quad (\text{by (40)})$$

If we use these inequalities to the previous form of $a_h(v, w)$, then

$$\begin{aligned}
a_h(v, w) &\geq C_{\text{co}} (X_{c,\partial}^2 + X_{c,0}^2 + X_{0,\partial}^2 + X_{0,0}^2 - 2X_{c,\partial} X_{0,\partial}) \\
&\quad - \delta ((3C_1 + C_1 C_2) X_{c,\partial} X_{0,0} + C_2 X_{0,0}^2) + \delta X_{0,0}^2
\end{aligned}$$

By Young's inequality we can have

$$\begin{aligned} a_h(v, w) \geq & C_{\text{co}}(X_{c,\partial}^2 + X_{c,0}^2 + X_{0,\partial}^2 + X_{0,0}^2 - 2X_{c,\partial}X_{0,\partial}) \\ & - \delta^2 C_3 X_{c,\partial}^2 - \frac{C_{\text{co}}}{4} X_{0,0}^2 - \delta C_2 X_{0,0}^2 + \delta X_{0,\partial}^2 \end{aligned}$$

with $C_3 > 0$ depending on C_1 , C_2 , and C_{co} . If we use

$$2X_{c,\partial}X_{0,\partial} \leq \frac{1}{1 + \frac{\delta}{2}} X_{c,\partial}^2 + \left(1 + \frac{\delta}{2}\right) X_{0,\partial}^2,$$

we can obtain

$$a_h(v, w) \geq \left(C_{\text{co}} \frac{\delta}{2 + \delta} - 9\delta^2 C_1^2\right) X_{c,\partial}^2 + C_{\text{co}} X_{c,0}^2 + \frac{C_{\text{co}}\delta}{2} X_{0,\partial}^2 + \frac{3C_{\text{co}}}{4} X_{0,0}^2.$$

If we choose δ sufficiently small, then

$$a_h(v, w) \geq C \|(v^c, v^0)\|^2.$$

The conclusion follows by combining it with (41). \square

To construct preconditioners recall that $\tilde{\mathcal{A}}_h$ is given by the bilinear form

$$a_h(v^c, w^c) + a_h(v^0, w^0), \quad v, w \in V_h, \quad (42)$$

and its matrix form is

$$\begin{pmatrix} \mathbb{M}_c & 0 \\ 0 & \mathbb{M}_0 \end{pmatrix} \quad (43)$$

where \mathbb{M}_c and \mathbb{M}_0 are matrices obtained from $a_h(v^c, w^c)$ and $a_h(v^0, w^0)$, respectively.

To construct preconditioners in practice, we use algebraic multigrid methods to obtain an approximate inverse of this block diagonal matrix. It is known that algebraic multigrid methods give good preconditioners for \mathbb{M}_c . Since \mathbb{M}_0 is a weakly diagonally dominant matrix [1], standard algebraic multigrid methods give an efficient preconditioner for this as well. Therefore it is feasible to use a preconditioner of the form

$$\begin{pmatrix} \text{AMG}(\mathbb{M}_c) & 0 \\ 0 & \text{AMG}(\mathbb{M}_0) \end{pmatrix}, \quad (44)$$

where $\text{AMG}(\mathbb{M})$ is a preconditioner of \mathbb{M} constructed by algebraic multigrid methods. We will see in the next section that this form of preconditioner gives robust numerical results.

5. Numerical results

In this section we present results of numerical experiments. In all numerical experiments, $\Omega = [0, 1] \times [0, 1] \subset \mathbb{R}^2$ and meshes consist of triangles that are bisections of $N \times N$ subsquares of Ω ($N = 4, 8, 16, 32, 64, 128$). All numerical experiments are implemented with Firedrake [2].

κ_0	k	h_{\max}	$\ u - u_h\ _0$		$\ u - u_h\ _{a_h}$		$\ \mathbf{z} - \mathbf{z}_h\ _{\underline{\kappa}-1}$		$\ P_0(f - \operatorname{div} \mathbf{z}_h)\ _0$	
			error	rate	error	rate	error	rate	error	rate
1	1	1/4	1.7308e-02	—	2.3161e-01	—	1.8076e-01	—	6.7771e-06	—
		1/8	5.1275e-03	1.76	1.1741e-01	0.98	8.2400e-02	1.13	2.5351e-07	4.74
		1/16	1.3562e-03	1.92	5.8387e-02	1.01	3.7214e-02	1.15	9.4939e-10	8.06
		1/32	3.4623e-04	1.97	2.9030e-02	1.01	1.7672e-02	1.07	1.7626e-11	5.75
		1/64	8.7325e-05	1.99	1.4464e-02	1.01	8.6656e-03	1.03	7.8165e-13	4.50
		1/128	2.1919e-05	1.99	7.2184e-03	1.00	4.3051e-03	1.01	2.6110e-12	-1.74
	2	1/4	9.4928e-04	—	3.2761e-02	—	2.2220e-02	—	3.4991e-06	—
		1/8	1.2497e-04	2.93	8.3853e-03	1.97	5.7538e-03	1.95	7.1763e-08	5.61
		1/16	1.6095e-05	2.96	2.1117e-03	1.99	1.4424e-03	2.00	7.7163e-09	3.22
		1/32	2.0441e-06	2.98	5.2921e-04	2.00	3.5990e-04	2.00	7.6639e-09	0.01
		1/64	2.5760e-07	2.99	1.3242e-04	2.00	8.9842e-05	2.00	2.1702e-12	11.79
		1/128	3.2332e-08	2.99	3.3119e-05	2.00	2.2443e-05	2.00	5.6520e-12	-1.38
10	1	1/4	1.7137e-02	—	2.4350e-01	—	1.4273e+00	—	9.5339e-06	—
		1/8	5.0961e-03	1.75	1.1878e-01	1.04	7.1024e-01	1.01	2.1197e-06	2.17
		1/16	1.3477e-03	1.92	5.8490e-02	1.02	3.4157e-01	1.06	2.6746e-09	9.63
		1/32	3.4400e-04	1.97	2.9036e-02	1.01	1.6712e-01	1.03	6.7721e-11	5.30
		1/64	8.6762e-05	1.99	1.4465e-02	1.01	8.2866e-02	1.01	3.5851e-12	4.24
		1/128	2.1779e-05	1.99	7.2184e-03	1.00	4.1319e-02	1.00	1.7368e-11	-2.28
	2	1/4	9.2460e-04	—	3.4138e-02	—	1.8100e-01	—	1.3883e-04	—
		1/8	1.2355e-04	2.90	8.5794e-03	1.99	4.6441e-02	1.96	1.8428e-07	9.56
		1/16	1.6029e-05	2.95	2.1321e-03	2.01	1.1725e-02	1.99	4.1800e-09	5.46
		1/32	2.0410e-06	2.97	5.3134e-04	2.00	2.9323e-03	2.00	3.8990e-11	6.74
		1/64	2.5744e-07	2.99	1.3266e-04	2.00	7.3234e-04	2.00	1.4930e-10	-1.94
		1/128	3.2324e-08	2.99	3.3146e-05	2.00	1.8294e-04	2.00	4.1893e-11	1.83

Table 1: Convergence with IOP-EG ($\alpha = 1$)

κ_0	k	h_{\max}	$\ u - u_h\ _0$		$\ u - u_h\ _{a_h}$		$\ \mathbf{z} - \mathbf{z}_h\ _{\underline{\kappa}-1}$		$\ P_0(f - \operatorname{div} \mathbf{z}_h)\ _0$	
			error	rate	error	rate	error	rate	error	rate
1	1	1/4	1.7401e-02	—	2.3188e-01	—	1.8320e-01	—	6.8532e-06	—
		1/8	5.1355e-03	1.76	1.1747e-01	0.98	8.2788e-02	1.15	1.8947e-07	5.18
		1/16	1.3566e-03	1.92	5.8393e-02	1.01	3.7252e-02	1.15	7.5461e-10	7.97
		1/32	3.4625e-04	1.97	2.9030e-02	1.01	1.7675e-02	1.08	1.6568e-10	2.19
		1/64	8.7325e-05	1.99	1.4464e-02	1.01	8.6658e-03	1.03	8.2004e-13	7.66
		1/128	2.1919e-05	1.99	7.2184e-03	1.00	4.3051e-03	1.01	3.2178e-12	-1.97
	2	1/4	9.5448e-04	—	3.2725e-02	—	2.2288e-02	—	8.6195e-06	—
		1/8	1.2558e-04	2.93	8.3760e-03	1.97	5.7739e-03	1.95	2.4908e-07	5.11
		1/16	1.6143e-05	2.96	2.1101e-03	1.99	1.4448e-03	2.00	7.6211e-09	5.03
		1/32	2.0474e-06	2.98	5.2900e-04	2.00	3.6016e-04	2.00	4.2024e-11	7.50
		1/64	2.5781e-07	2.99	1.3240e-04	2.00	8.9872e-05	2.00	1.7440e-12	4.59
		1/128	3.2346e-08	2.99	3.3115e-05	2.00	2.2446e-05	2.00	6.0409e-12	-1.79
10	1	1/4	1.7247e-02	—	2.4384e-01	—	1.4472e+00	—	1.0088e-05	—
		1/8	5.1065e-03	1.76	1.1880e-01	1.04	7.1414e-01	1.02	4.3373e-07	4.54
		1/16	1.3482e-03	1.92	5.8491e-02	1.02	3.4198e-01	1.06	2.0379e-08	4.41
		1/32	3.4403e-04	1.97	2.9036e-02	1.01	1.6715e-01	1.03	1.7823e-10	6.84
		1/64	8.6763e-05	1.99	1.4465e-02	1.01	8.2869e-02	1.01	1.4241e-11	3.65
		1/128	2.1779e-05	1.99	7.2184e-03	1.00	4.1319e-02	1.00	1.2476e-11	0.19
	2	1/4	9.3036e-04	—	3.4126e-02	—	1.8096e-01	—	4.6005e-05	—
		1/8	1.2421e-04	2.90	8.5715e-03	1.99	4.6526e-02	1.96	4.2893e-07	6.74
		1/16	1.6080e-05	2.95	2.1306e-03	2.01	1.1734e-02	1.99	6.3984e-09	6.07
		1/32	2.0444e-06	2.98	5.3112e-04	2.00	2.9330e-03	2.00	1.2150e-10	5.72
		1/64	2.5766e-07	2.99	1.3263e-04	2.00	7.3240e-04	2.00	4.0834e-11	1.57
		1/128	3.2338e-08	2.99	3.3142e-05	2.00	1.8295e-04	2.00	4.3021e-11	-0.08

Table 2: Convergence with IOP-EG ($\alpha = 2$)

In numerical experiments of this section, we consider

$$-\nabla \cdot (\underline{\kappa} \nabla u) = f \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega \quad (45)$$

k	κ_0	N				
		8	16	32	64	128
1	1	27	36	48	64	86
	2	29	37	50	67	91
	4	29	40	54	73	99
	8	30	43	59	81	110
	10	30	43	59	83	115
2	1	38	45	56	73	95
	2	40	47	59	76	101
	4	42	51	63	83	110
	8	45	55	70	93	124
	10	46	57	72	97	129

Table 3: Number of iterations for $\alpha = 0$

k	κ_0	N				
		8	16	32	64	128
1	1	17	18	19	19	18
	2	18	19	19	19	19
	4	19	20	20	19	19
	8	20	21	21	21	20
	10	20	21	21	21	21
2	1	31	32	31	30	28
	2	32	33	33	32	30
	4	33	34	34	32	32
	8	34	35	35	34	33
	10	35	36	35	34	33

Table 4: Number of iterations for $\alpha = 1$

with a tensor

$$\underline{\kappa} = \begin{pmatrix} \kappa_0 & 0 \\ 0 & 1 \end{pmatrix} \quad (46)$$

where $\kappa_0 > 0$ is a constant. Let $\underline{\kappa}_n := \underline{\kappa}n \cdot n$ on edge/face e with a unit normal vector n on e . Then, a modification of (4) for (45) is

$$\begin{aligned} a_h(v, w) = & (\underline{\kappa} \nabla v, \nabla w) - \left(\langle \llbracket \underline{\kappa} \nabla v \rrbracket, \llbracket w \rrbracket \rangle_{\mathcal{E}_h^0 \cup \mathcal{E}_h^D} + \langle \llbracket v \rrbracket, \llbracket \underline{\kappa} \nabla w \rrbracket \rangle_{\mathcal{E}_h^0 \cup \mathcal{E}_h^D} \right) \\ & + \langle \gamma \underline{\kappa}_n h_e^{-1-\alpha} \llbracket v \rrbracket, \llbracket w \rrbracket \rangle_{\mathcal{E}_h^0} + \langle \gamma \underline{\kappa}_n h_e^{-1} \llbracket v \rrbracket, \llbracket w \rrbracket \rangle_{\mathcal{E}_h^D}, \end{aligned} \quad (47)$$

and flux reconstruction formula is

$$\begin{aligned} (\mathbf{z}_h, \mathbf{q})_T &= (-\underline{\kappa} \nabla u_h, \mathbf{q})_T, & \mathbf{q} &\in \mathcal{P}_{k-2}(T), \\ \langle \mathbf{z}_h \cdot \mathbf{n}, q \rangle_e &= \langle -\llbracket \underline{\kappa} \nabla u_h \rrbracket \cdot \mathbf{n} + \gamma \underline{\kappa}_n h_e^{-1-\alpha} \llbracket u_h \rrbracket \cdot \mathbf{n}, q \rangle_e, & q &\in \mathcal{P}_{k-1}(e), e \in \mathcal{E}_h^0, \\ \langle \mathbf{z}_h \cdot \mathbf{n}, q \rangle_e &= \langle u_N, q \rangle_e, & q &\in \mathcal{P}_{k-1}(e), e \notin \mathcal{E}_h^0 \cup \mathcal{E}_h^D, \\ \langle \mathbf{z}_h \cdot \mathbf{n}, q \rangle_e &= \langle -\underline{\kappa} \nabla u_h \cdot \mathbf{n} + \gamma \underline{\kappa}_n h_e^{-1} (u_h - u_D), q \rangle_e, & q &\in \mathcal{P}_{k-1}(e), e \in \mathcal{E}_h^D. \end{aligned}$$

For simplicity of presentation we only showed a priori error analysis in Section 3 for $\kappa_0 = 1$. The error analysis can be extended to (45) with $\underline{\kappa}$ weighted flux and penalization terms. In contrast, for the preconditioning discussed in Section 4, a constant which is related to the anisotropy of $\underline{\kappa}$, is involved in the operator preconditioning analysis. Thus, it does not seem to be straightforward to get an analytic proof that abstract preconditioners of the form (42) are spectrally equivalent to the operator given by (47) with equivalence constants independent

k	κ_0	N				
		8	16	32	64	128
1	1	15	15	14	14	13
	2	15	15	14	14	14
	4	16	15	15	14	14
	8	16	15	15	15	14
	10	16	16	15	15	14
2	1	28	29	29	27	25
	2	29	30	29	29	28
	4	31	31	30	29	29
	8	31	31	31	30	29
	10	32	31	31	30	29

Table 5: Number of iterations for $\alpha = 2$

k	κ_0	N				
		8	16	32	64	128
1	1	15	17	19	22	25
	2	16	17	20	23	26
	4	16	18	20	23	28
	8	17	19	21	25	31
	10	17	19	22	26	32
2	1	29	31	31	32	35
	2	30	32	33	34	37
	4	31	33	34	35	38
	8	32	34	35	37	41
	10	32	34	35	37	41

Table 6: Number of iterations for large $\gamma (= 200)$ on \mathcal{E}_h^0

of the anisotropy constant. Nonetheless, we present numerical results of convergence and preconditioners for anisotropic coefficients in this section because the numerical test results below show that IOP-EG methods with preconditioner (42) are robust for the anisotropy of $\boldsymbol{\kappa}$ (see Tables 4-5). As seen in Table 3, preconditioned EG methods show worse performances if $\boldsymbol{\kappa}$ is more anisotropic (i.e., κ_0 is larger), so this anisotropy robustness is another advantage of IOP-EG methods.

In the first set of numerical experiments we present convergence rates of errors of various versions of IOP-EG methods with $\alpha = 1, 2$, $\kappa_0 = 1, 10$ in (47) and the manufactured solution $u = x(1-x)\sin(\pi y)$. Recall that a larger α implies a stronger interior over-penalization. From the results in Tables 1-2, one can see that convergence rates are optimal for the L^2 error $\|u - u_h\|_0$, the a_h -norm $\|u - u_h\|_{a_h}$ endowed by (7) with the bilinear form in (47), the $\boldsymbol{\kappa}^{-1}$ -weighted L^2 flux error $\|\mathbf{z} - \mathbf{z}_h\|_{\boldsymbol{\kappa}^{-1}} := (\boldsymbol{\kappa}^{-1}(\mathbf{z} - \mathbf{z}_h), \mathbf{z} - \mathbf{z}_h)^{1/2}$. The errors of local mass reach the level of machine precision zero quickly as mesh is refined. As shown in the proof of Theorem 2, exact local mass conservation theoretically holds for all meshes. However, the error $\|P_0(f - \operatorname{div} \mathbf{z}_h)\|_0$ is not machine precision zero for coarse meshes in our experiments because the errors of numerical quadrature of manufactured solution $u = x(1-x)\sin(\pi y)$ are involved in flux reconstruction.

In the second set of experiments we present performance of preconditioners of the form (44) as a function of mesh refinement and anisotropy of the permeability tensor. We set $\kappa_0 = 1, 2, 4, 8, 10$ to test the proposed numerical methods with preconditioners for anisotropic tensors. In the results of experiments, k is the polynomial degree of V_h^c , κ_0 is the coefficient in (46) and $\gamma = 10$.

Numerical results for iterative solvers with $\alpha = 1, 2$ are presented in Tables 4-5. More specifically, we present the number of iterations of a preconditioned

k	κ_0	N				
		8	16	32	64	128
1	1	21	23	26	29	32
	2	22	25	27	31	34
	4	23	27	30	33	37
	8	24	29	32	36	41
	10	24	29	34	38	42
	1	34	35	37	39	41
2	2	35	37	39	40	42
	4	36	39	40	42	45
	8	38	41	44	46	48
	10	39	42	45	47	51

Table 7: Number of iterations for $\alpha = 0.5$

MinRes method with the block diagonal preconditioner of the form (44), and iteration stops when relative error becomes smaller than 10^{-12} of the initial error or when the number of iterations is more than 10^4 .

We use the MinRes method instead of the conjugate gradient method for guaranteed convergence because the matrix is not symmetric positive definite in either EG or IOP-EG methods. We use the hypre library (cf. [2]) as an algebraic multigrid preconditioner for the blocks in (44). The corresponding preconditioner is constructed as in (42) with a_h in (47).

We also tested performance of preconditioners for the original EG method ($\alpha = 0$) and two IOP-EG methods with $\alpha = 1, 2$. The results are given in Tables 3–5. In Table 3, the number of iterations for the EG method clearly increases for mesh refinement in all cases. Moreover, the number of iterations increases if κ is more anisotropic, and the increment is not negligible for the two cases $\kappa_0 = 1$ and $\kappa_0 = 10$ in the finest mesh ($N = 128$). In contrast, the IOP-EG methods with $\alpha = 1, 2$ perform much better than the ones of the original EG method in all cases. As can be seen in Tables 4–5, the number of iterations is very robust for mesh refinement. For κ anisotropy, the number of iterations increases as κ becomes more anisotropic. Nevertheless, the IOP-EG methods perform much better than the original EG methods in all cases.

Finally, we present two additional preconditioning experiment results. The purpose of the first additional experiment is to verify that the mesh-dependent over-penalization is necessary for robust preconditioning for mesh refinement. For this we set $\alpha = 0$ and use 10 times stronger interior penalization parameter γ on the interior edges/faces than the penalization parameter on the boundary edges/faces. Specifically, we set γ on the interior edges/faces as $\gamma_{\text{int}} = 100$ and set γ on the boundary edges/faces as $\gamma_{\partial} = 10$. In Table 6 one can see that the number of iterations is smaller than the ones in Table 3, so the interior over-penalization with a constant factor clearly improves preconditioning performances. However, the results in Table 6 also show that the number of iterations still considerably increases for mesh refinement. The purpose of the second additional experiment is to obtain numerical evidence of how large $\alpha > 0$ need to be for robust preconditioning while our analysis showed that $\alpha \geq 1$ is sufficient if $\kappa_0 = 1$. For this, we ran preconditioning experiments with different α values and the results are given in Figure 1. The results show that the number of iterations increases if $\alpha < 0.9$ while for mesh refinement and anisotropy

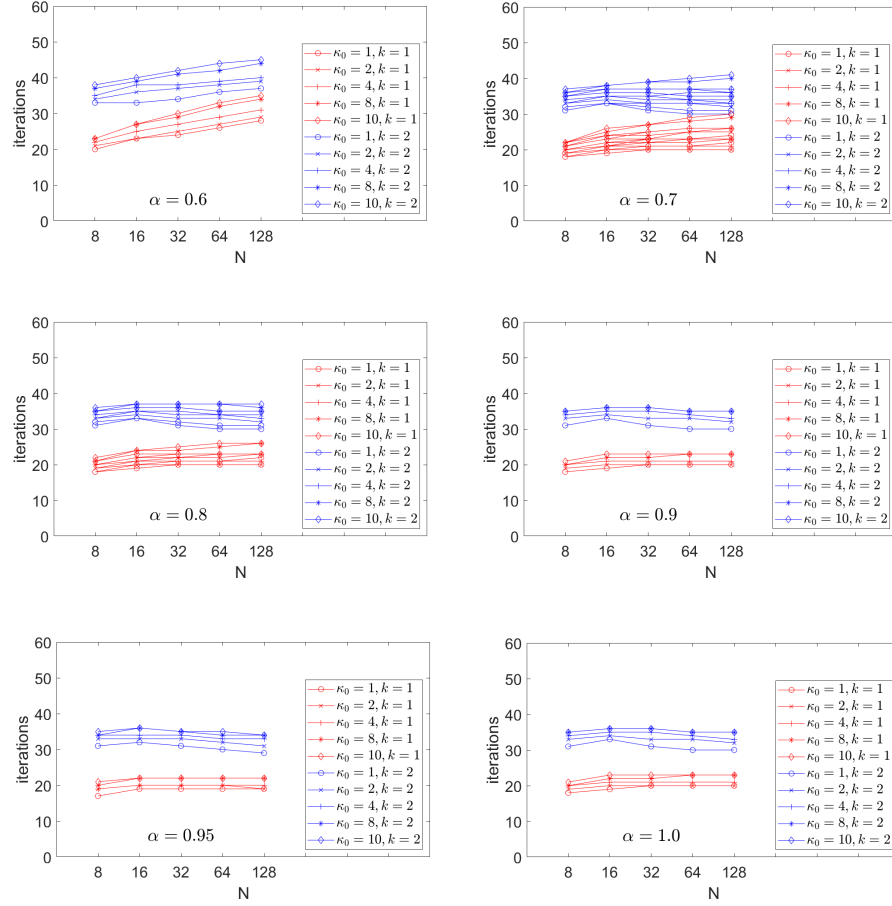


Figure 1: Comparison of iteration numbers for different α and κ_0 values

of κ , the iteration numbers seem to be nearly stable for mesh refinement and anisotropy if $\alpha \geq 0.9$. The numerical results show that while $\alpha = 1$ may not be a sharp threshold value for preconditioning robustness we can observe that α should be sufficiently close to 1.

Acknowledgement

Omar Ghattas gratefully acknowledge support by Department of Energy, Office of Advanced Scientific Computing Research (award number DE-SC0019303). Jeonghun J. Lee is partially supported by University Research Committee grant of Baylor University and by the National Science Foundation under grant number DMS-2110781.

6. Conclusion

In this paper we propose interior over-penalized enriched Galerkin (IOP-EG) methods for second order elliptic equations. From theoretical point of view, a new medius error analysis and a new spectral equivalence analysis, are developed for optimal convergence of errors and for construction of robust iterative solvers. In numerical experiment results comparing preconditioned IOP-EG and EG methods, we observe that preconditioned IOP-EG methods show very robust iterative solver performances for mesh refinement and for the anisotropy of permeability coefficients. In conclusion, the IOP-EG methods can be a good replacement of the original EG methods providing parameter-robust scalable iterative solvers.

References

- [1] B. Cockburn, J. Gopalakrishnan, H. Wang, Locally conservative fluxes for the continuous Galerkin method, *SIAM J. Numer. Anal.* 45 (4) (2007) 1742–1776. doi:10.1137/060666305.
URL <http://dx.doi.org/10.1137/060666305>
- [2] S. Chippada, C. N. Dawson, M. L. Martínez, M. F. Wheeler, A projection method for constructing a mass conservative velocity field, *Comput. Methods Appl. Mech. Engrg.* 157 (1-2) (1998) 1–10. doi:10.1016/S0045-7825(98)80001-7.
URL [http://dx.doi.org/10.1016/S0045-7825\(98\)80001-7](http://dx.doi.org/10.1016/S0045-7825(98)80001-7)
- [3] T. J. R. Hughes, G. Engel, L. Mazzei, M. G. Larson, The continuous Galerkin method is locally conservative, *J. Comput. Phys.* 163 (2) (2000) 467–488. doi:10.1006/jcph.2000.6577.
URL <http://dx.doi.org/10.1006/jcph.2000.6577>
- [4] F. Brezzi, J. J. Douglas, L. D. Marini, Two families of mixed finite elements for second order elliptic problems, *Numer. Math.* 47 (2) (1985) 217–235. doi:10.1007/BF01389710.
URL <http://dx.doi.org/10.1007/BF01389710>
- [5] P.-A. Raviart, J. M. Thomas, A mixed finite element method for 2nd order elliptic problems, in: *Mathematical aspects of finite element methods* (Proc. Conf., Consiglio Naz. delle Ricerche (C.N.R.), Rome, 1975), Springer, Berlin, 1977, pp. 292–315. *Lecture Notes in Math.*, Vol. 606.
- [6] J.-C. Nédélec, Mixed finite elements in \mathbf{R}^3 , *Numer. Math.* 35 (3) (1980) 315–341. doi:10.1007/BF01396415.
URL <http://dx.doi.org/10.1007/BF01396415>
- [7] J.-C. Nédélec, A new family of mixed finite elements in \mathbf{R}^3 , *Numer. Math.* 50 (1) (1986) 57–81. doi:10.1007/BF01389668.
URL <http://dx.doi.org/10.1007/BF01389668>

- [8] D. N. Arnold, An interior penalty finite element method with discontinuous elements, *SIAM J. Numer. Anal.* 19 (4) (1982) 742–760. doi:10.1137/0719052.
URL <http://dx.doi.org/10.1137/0719052>
- [9] D. N. Arnold, F. Brezzi, B. Cockburn, L. D. Marini, Unified analysis of discontinuous Galerkin methods for elliptic problems, *SIAM J. Numer. Anal.* 39 (5) (2001/02) 1749–1779. doi:10.1137/S0036142901384162.
URL <http://dx.doi.org/10.1137/S0036142901384162>
- [10] M. F. Wheeler, A priori L_2 error estimates for Galerkin approximations to parabolic partial differential equations, *SIAM J. Numer. Anal.* 10 (1973) 723–759.
- [11] B. Cockburn, J. Gopalakrishnan, R. Lazarov, Unified hybridization of discontinuous Galerkin, mixed, and continuous Galerkin methods for second order elliptic problems, *SIAM J. Numer. Anal.* 47 (2) (2009) 1319–1365. doi:10.1137/070706616.
URL <http://dx.doi.org/10.1137/070706616>
- [12] J. Wang, X. Ye, A weak Galerkin finite element method for second-order elliptic problems, *J. Comput. Appl. Math.* 241 (2013) 103–115. doi:10.1016/j.cam.2012.10.003.
URL <https://doi.org/10.1016/j.cam.2012.10.003>
- [13] S. Sun, J. Liu, A locally conservative finite element method based on piecewise constant enrichment of the continuous Galerkin method, *SIAM J. Sci. Comput.* 31 (4) (2009) 2528–2548. doi:10.1137/080722953.
URL <http://dx.doi.org/10.1137/080722953>
- [14] S. Lee, Y.-J. Lee, M. F. Wheeler, A locally conservative enriched Galerkin approximation and efficient solver for elliptic and parabolic problems, *SIAM J. Sci. Comput.* 38 (3) (2016) A1404–A1429. doi:10.1137/15M1041109.
URL <http://dx.doi.org/10.1137/15M1041109>
- [15] K.-A. Mardal, R. Winther, Preconditioning discretizations of systems of partial differential equations, *Numer. Linear Algebra Appl.* 18 (1) (2011) 1–40. doi:10.1002/nla.716.
URL <http://dx.doi.org/10.1002/nla.716>
- [16] L. C. Evans, Partial differential equations, Vol. 19 of Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, 1998.
- [17] T. Gudi, A new error analysis for discontinuous finite element methods for linear elliptic problems, *Math. Comp.* 79 (272) (2010) 2169–2189. doi:10.1090/S0025-5718-10-02360-4.
URL <http://dx.doi.org/10.1090/S0025-5718-10-02360-4>

- [18] J. Nitsche, über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind, Abh. Math. Sem. Univ. Hamburg 36 (1971) 9–15, collection of articles dedicated to Lothar Collatz on his sixtieth birthday. doi:10.1007/BF02995904.
URL <https://doi.org/10.1007/BF02995904>
- [19] N. Lüthen, M. Juntunen, R. Stenberg, An improved a priori error analysis of Nitsche’s method for Robin boundary conditions, Numer. Math. 138 (4) (2018) 1011–1026. doi:10.1007/s00211-017-0927-1.
URL <https://doi.org/10.1007/s00211-017-0927-1>
- [20] A. Buffa, C. Ortner, Compact embeddings of broken Sobolev spaces and applications, IMA J. Numer. Anal. 29 (4) (2009) 827–855. doi:10.1093/imanum/drn038.
URL <http://dx.doi.org/10.1093/imanum/drn038>
- [21] S. C. Brenner, Poincaré-Friedrichs inequalities for piecewise H^1 functions, SIAM J. Numer. Anal. 41 (1) (2003) 306–324. doi:10.1137/S0036142902401311.
URL <http://dx.doi.org/10.1137/S0036142902401311>
- [22] F. Rathgeber, D. A. Ham, L. Mitchell, M. Lange, F. Luporini, A. T. T. McRae, G.-T. Bercea, G. R. Markall, P. H. J. Kelly, Firedrake: automating the finite element method by composing abstractions, ACM Trans. Math. Software 43 (3) (2017) Art. 24, 27. doi:10.1145/2998441.
URL <https://doi.org/10.1145/2998441>
- [23] R. D. Falgout, U. M. Yang, Computational Science—ICCS 2002: International Conference Amsterdam, The Netherlands, April 21–24, 2002 Proceedings, Part III (2002) 632–641doi:10.1007/3-540-47789-6_66.
URL http://dx.doi.org/10.1007/3-540-47789-6_66