Missouri University of Science and Technology
Scholars' Mine

Mathematics and Statistics Faculty Research & Creative Works

Mathematics and Statistics

01 Oct 2019

# An Optimal EDG Method for Distributed Control of Convection Diffusion PDEs

X. Zhang

Y. Zhang

John R. Singler
*Missouri University of Science and Technology*, singlerj@mst.edu

Follow this and additional works at: https://scholarsmine.mst.edu/math_stat_facwork

Part of the Mathematics Commons, and the Statistics and Probability Commons

## Recommended Citation

X. Zhang et al., "An Optimal EDG Method for Distributed Control of Convection Diffusion PDEs," *International Journal of Numerical Analysis and Modeling*, vol. 16, no. 4, pp. 519-542, University of Alberta, Oct 2019.

This Article - Journal is brought to you for free and open access by Scholars' Mine. It has been accepted for inclusion in Mathematics and Statistics Faculty Research & Creative Works by an authorized administrator of Scholars' Mine. This work is protected by U. S. Copyright Law. Unauthorized use including reproduction for redistribution requires the permission of the copyright holder. For more information, please contact scholarsmine@mst.edu.

INTERNATIONAL JOURNAL OF
NUMERICAL ANALYSIS AND MODELING
Volume 16, Number 4, Pages 519–542

© 2019 Institute for Scientific
Computing and Information

# AN OPTIMAL EDG METHOD FOR DISTRIBUTED CONTROL OF CONVECTION DIFFUSION PDES

XIAO ZHANG, YANGWEN ZHANG, AND JOHN R. SINGLER

**Abstract.** We propose an embedded discontinuous Galerkin (EDG) method to approximate the solution of a distributed control problem governed by convection diffusion PDEs, and obtain optimal a priori error estimates for the state, dual state, their fluxes, and the control. Moreover, we prove the optimize-then-discretize (OD) and discrtize-then-optimize (DO) approaches coincide. Numerical results confirm our theoretical results.

**Key words.** Distributed optimal control, convection diffusion, embedded discontinuous Galerkin method, error analysis, optimize-then-discretize, discrtize-then-optimize.

## 1. Introduction

We study the following distributed optimal control problem:

$$\min J(u) = \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)} + \frac{\gamma}{2}\|u\|^2_{L^2(\Omega)}, \quad \gamma > 0, \tag{1}$$

subject to

$$\begin{aligned} -\Delta y + \boldsymbol{\beta} \cdot \nabla y &= f + u \quad \text{in } \Omega, \\ y &= g \quad\quad\quad \text{on } \partial\Omega, \end{aligned} \tag{2}$$

where $\Omega \subset \mathbb{R}^d$ $(d \geq 2)$ is a Lipschitz polyhedral domain with boundary $\Gamma = \partial\Omega$, $f \in L^2(\Omega)$, $g \in C^0(\partial\Omega)$, and the vector field $\boldsymbol{\beta}$ satisfies

$$\nabla \cdot \boldsymbol{\beta} \leq 0. \tag{3}$$

Optimal control problems for convection diffusion equations have been extensively studied using many different finite element methods, such as standard finite elements [11–13], mixed finite elements [13, 35, 39], discontinuous Galerkin (DG) methods [16, 21, 33, 34, 36, 40, 41] and hybrid discontinuous Galerkin (HDG) methods [17, 18]. HDG methods were first introduced by Cockburn et al. in [4] for second order elliptic problems, and they have subsequently been applied to many other problems [2, 3, 5, 7, 8, 23–26, 32]. HDG methods keep the advantages of DG methods, but have a lower number of globally coupled degrees of freedom compared to mixed methods and DG methods. However, the degrees of freedom for HDG methods is still larger compared to standard finite element methods. Embedded discontinuous Galerkin (EDG) methods were first proposed in [15], and then analyzed in [6]. EDG methods are obtained from the HDG methods by forcing the numerical trace space to be continuous. This simple change significantly reduces the number of degrees of freedom and make EDG methods competitive for flow problems [27] and many other applications [9, 10, 19, 27, 29].

In [38], we utilized an EDG method for a distributed optimal control problem for the Poisson equation. We obtained optimal convergence rates for the state,

Received by the editors December 29, 2017.
1991 *Mathematics Subject Classification.* 65N30, 49M25.

dual state and the control, but *suboptimal* convergence rates for their fluxes. This suboptimal flux convergence rate for the Poisson equation is a limitation of the EDG method with equal order polynomial degrees for all variables [6]. However, Zhang, Xie, and Zhang recently proposed a new EDG method and proved optimal convergence rates for all variables for the Poisson equation [37]. This new EDG method is obtained by simply using a lower degree finite element space for the flux. In this work, we use this new EDG method to approximate the solution of the above convection diffusion distributed optimal control problem, and in Section 3 we prove optimal convergence rates for all variables.

There are two main approaches to compute the numerical solution of PDE constrained optimal control problems: the optimize-then-discretize (OD) and discretize-then-optimize (DO) approaches. In the OD approach, one first derives the first-order necessary optimality conditions, then discretizes the optimality system, and then solves the resulting discrete system by utilizing efficient iterative solvers [31]. In the DO approach, one first discretizes the PDE optimization problem to obtain a finite dimensional optimization problem, which is then solved by existing optimization algorithms, such as [1,28]. The discretization methods for which these two approaches coincide are called *commutative*. Intuitively, the DO approach is more straightforward in practice; however, not all discretization schemes are commutative. In the non-commutative case, the DO approach may result in badly behaved numerical results; see, e.g., [20, 22]. Therefore, devising commutative numerical methods is very important. In Section 2, we prove the EDG method studied here is commutative for the convection diffusion distributed control problem. Moreover, we provide numerical examples to confirm our theoretical results in Section 4.

## 2. EDG scheme for the optimal control problem

**2.1. Notation.** Throughout the paper we adopt the standard notation $W^{m,p}(\Omega)$ for Sobolev spaces on $\Omega$ with norm $\|\cdot\|_{m,p,\Omega}$ and seminorm $|\cdot|_{m,p,\Omega}$ . We denote $W^{m,2}(\Omega)$ by $H^m(\Omega)$ with norm $\|\cdot\|_{m,\Omega}$ and seminorm $|\cdot|_{m,\Omega}$. Specifically, $H_0^1(\Omega) = \{v \in H^1(\Omega) : v = 0 \text{ on } \partial\Omega\}$. We denote the $L^2$-inner products on $L^2(\Omega)$ and $L^2(\Gamma)$ by

$$
\begin{aligned}
(v, w) &= \int_\Omega vw \quad \forall v, w \in L^2(\Omega),\\
\langle v, w\rangle &= \int_\Gamma vw \quad \forall v, w \in L^2(\Gamma).
\end{aligned}
$$

Define the space $H(\text{div}, \Omega)$ as

$$H(\text{div}, \Omega) = \{\boldsymbol{v} \in [L^2(\Omega)]^d, \nabla \cdot \boldsymbol{v} \in L^2(\Omega)\}.$$

Let $\mathcal{T}_h$ be a collection of disjoint elements that partition $\Omega$. We denote by $\partial\mathcal{T}_h$ the set $\{\partial K : K \in \mathcal{T}_h\}$. For an element $K$ of the collection $\mathcal{T}_h$, let $e = \partial K \cap \Gamma$ denote the boundary face of $K$ if the $d-1$ Lebesgue measure of $e$ is non-zero. For two elements $K^+$ and $K^-$ of the collection $\mathcal{T}_h$, let $e = \partial K^+ \cap \partial K^-$ denote the interior face between $K^+$ and $K^-$ if the $d-1$ Lebesgue measure of $e$ is non-zero. Let $\varepsilon_h^o$ and $\varepsilon_h^\partial$ denote the set of interior and boundary faces, respectively. We denote by $\varepsilon_h$ the union of $\varepsilon_h^o$ and $\varepsilon_h^\partial$. We finally introduce

$$(w, v)_{\mathcal{T}_h} = \sum_{K\in\mathcal{T}_h} (w, v)_K, \qquad\qquad \langle \zeta, \rho\rangle_{\partial\mathcal{T}_h} = \sum_{K\in\mathcal{T}_h} \langle \zeta, \rho\rangle_{\partial K}\,.$$

Let $\mathcal{P}^k(D)$ denote the set of polynomials of degree at most $k \geq 0$ on a domain $D$. We introduce the discontinuous finite element spaces

$$\boldsymbol{V}_h := \{\boldsymbol{v} \in [L^2(\Omega)]^d : \boldsymbol{v}|_K \in [\mathcal{P}^k(K)]^d, \forall K \in \mathcal{T}_h\}, \tag{4}$$

$$W_h := \{w \in L^2(\Omega) : w|_K \in \mathcal{P}^{k+1}(K), \forall K \in \mathcal{T}_h\}, \tag{5}$$

$$M_h := \{\mu \in L^2(\varepsilon_h) : \mu|_e \in \mathcal{P}^{k+1}(e), \forall e \in \varepsilon_h\}. \tag{6}$$

Define $M_h(o)$ and $M_h(\partial)$ in the same way as $M_h$, but with $\varepsilon_h^o$ and $\varepsilon_h^\partial$ replacing $\varepsilon_h$. Note that $M_h$ consists of functions which are continuous inside the faces (or edges) $e \in \varepsilon_h$ and discontinuous at their borders. In addition, for any function $w \in W_h$ we use $\nabla w$ to denote the piecewise gradient on each element $K \in \mathcal{T}_h$. A similar convention applies to the divergence $\nabla \cdot \boldsymbol{r}$ for all $\boldsymbol{r} \in \boldsymbol{V}_h$.

For EDG methods, we only change the space of numerical traces $M_h$, which is discontinuous, into a continuous space $\widetilde{M}_h$ as follows:

$$\widetilde{M}_h := M_h \cap \mathcal{C}^0(\varepsilon_h). \tag{7}$$

The spaces $\widetilde{M}_h(o)$ and $\widetilde{M}_h(\partial)$ are defined in the same way as $M_h(o)$ and $M_h(\partial)$.

Recall we assume the Dirichlet boundary data $g$ is continuous. Let $\mathcal{I}_h$ be an interpolation operator, so that $\mathcal{I}_h g$ is a continuous interpolation of $g$ on $\varepsilon_h^\partial$.

Again, in most of the EDG works in the literature the polynomial degree is equal for the three spaces $\boldsymbol{V}_h$, $W_h$, and $\widetilde{M}_h$. We lower the polynomial degree for the flux space $\boldsymbol{V}_h$ as in the recent work [37].

### 2.2. Optimize-then-Discretize.

First, we consider the optimize-then-discretize (OD) approach: we use the EDG method to discretize the optimality system for the convection diffusion control problem.

It is well known that the optimal control problem (1)-(2) is equivalent to the optimality system

$$-\Delta y + \boldsymbol{\beta} \cdot \nabla y = f + u \qquad \text{in } \Omega, \tag{8a}$$

$$y = g \qquad \text{on } \partial\Omega, \tag{8b}$$

$$-\Delta z - \nabla \cdot (\boldsymbol{\beta} z) = y - y_d \qquad \text{in } \Omega, \tag{8c}$$

$$z = 0 \qquad \text{on } \partial\Omega, \tag{8d}$$

$$z + \gamma u = 0 \qquad \text{in } \Omega. \tag{8e}$$

For $\boldsymbol{q} = -\nabla y$ and $\boldsymbol{p} = -\nabla z$, the mixed weak form of the optimality system (8a)-(8e) is given by

$$(\boldsymbol{q}, \boldsymbol{r}) - (y, \nabla \cdot \boldsymbol{r}) + \langle y, \boldsymbol{r} \cdot \boldsymbol{n} \rangle = 0, \tag{9a}$$

$$(\nabla \cdot (\boldsymbol{q} + \boldsymbol{\beta} y), w) - (y \nabla \cdot \boldsymbol{\beta}, w) = (f + u, w), \tag{9b}$$

$$(\boldsymbol{p}, \boldsymbol{r}) - (z, \nabla \cdot \boldsymbol{r}) = 0, \tag{9c}$$

$$(\nabla \cdot (\boldsymbol{p} - \boldsymbol{\beta} z), w) = (y - y_d, w), \tag{9d}$$

$$(z + \gamma u, v) = 0, \tag{9e}$$

for all $(\boldsymbol{r}, w, v) \in H(\text{div}, \Omega) \times L^2(\Omega) \times L^2(\Omega)$.

To approximate the solution of this system, the EDG method seeks approximate fluxes $\boldsymbol{q}_h, \boldsymbol{p}_h \in \boldsymbol{V}_h$, states $y_h, z_h \in W_h$, interior element boundary traces $\widehat{y}_h^o, \widehat{z}_h^o \in$

$\widetilde{M}_h(o)$, and control $u_h \in W_h$ satisfying

$$(\boldsymbol{q}_h, \boldsymbol{r}_1)_{\mathcal{T}_h} - (y_h, \nabla \cdot \boldsymbol{r}_1)_{\mathcal{T}_h} + \langle \widehat{y}_h^o, \boldsymbol{r}_1 \cdot \boldsymbol{n} \rangle_{\partial \mathcal{T}_h \backslash \varepsilon_h^\partial} = -\langle \mathcal{I}_h g, \boldsymbol{r}_1 \cdot \boldsymbol{n} \rangle_{\varepsilon_h^\partial}, \tag{10a}$$

$$\begin{aligned} &-(\boldsymbol{q}_h + \boldsymbol{\beta} y_h, \nabla w_1)_{\mathcal{T}_h} - (y_h \nabla \cdot \boldsymbol{\beta}, w_1)_{\mathcal{T}_h} + \langle \widehat{\boldsymbol{q}}_h \cdot \boldsymbol{n}, w_1 \rangle_{\partial \mathcal{T}_h} \\ &+\langle \boldsymbol{\beta} \cdot \boldsymbol{n} \widehat{y}_h^o, w_1 \rangle_{\partial \mathcal{T}_h \backslash \varepsilon_h^\partial} - (u_h, w_1)_{\mathcal{T}_h} = -\langle \boldsymbol{\beta} \cdot \boldsymbol{n} \mathcal{I}_h g, w_1 \rangle_{\varepsilon_h^\partial} + (f, w_1)_{\mathcal{T}_h}, \end{aligned} \tag{10b}$$

for all $(\boldsymbol{r}_1, w_1) \in \boldsymbol{V}_h \times W_h$,

$$(\boldsymbol{p}_h, \boldsymbol{r}_2)_{\mathcal{T}_h} - (z_h, \nabla \cdot \boldsymbol{r}_2)_{\mathcal{T}_h} + \langle \widehat{z}_h^o, \boldsymbol{r}_2 \cdot \boldsymbol{n} \rangle_{\partial \mathcal{T}_h \backslash \varepsilon_h^\partial} = 0, \tag{10c}$$

$$\begin{aligned} &-(\boldsymbol{p}_h - \boldsymbol{\beta} z_h, \nabla w_2)_{\mathcal{T}_h} + \langle \widehat{\boldsymbol{p}}_h \cdot \boldsymbol{n}, w_2 \rangle_{\partial \mathcal{T}_h} \\ &-\langle \boldsymbol{\beta} \cdot \boldsymbol{n} \widehat{z}_h^o, w_2 \rangle_{\partial \mathcal{T}_h \backslash \varepsilon_h^\partial} - (y_h, w_2)_{\mathcal{T}_h} = -(y_d, w_2)_{\mathcal{T}_h}, \end{aligned} \tag{10d}$$

for all $(\boldsymbol{r}_2, w_2) \in \boldsymbol{V}_h \times W_h$,

$$\langle \widehat{\boldsymbol{q}}_h \cdot \boldsymbol{n} + \boldsymbol{\beta} \cdot \boldsymbol{n} \widehat{y}_h^o, \mu_1 \rangle_{\partial \mathcal{T}_h \backslash \varepsilon_h^\partial} = 0, \tag{10e}$$

$$\langle \widehat{\boldsymbol{p}}_h \cdot \boldsymbol{n} - \boldsymbol{\beta} \cdot \boldsymbol{n} \widehat{z}_h^o, \mu_2 \rangle_{\partial \mathcal{T}_h \backslash \varepsilon_h^\partial} = 0, \tag{10f}$$

for all $\mu_1, \mu_2 \in \widetilde{M}_h(o)$, and the optimality condition

$$(z_h + \gamma u_h, w_3)_{\mathcal{T}_h} = 0, \tag{10g}$$

for all $w_3 \in W_h$.

The numerical traces on $\partial \mathcal{T}_h$ are defined as

$$\widehat{\boldsymbol{q}}_h \cdot \boldsymbol{n} = \boldsymbol{q}_h \cdot \boldsymbol{n} + h^{-1}(y_h - \widehat{y}_h^o) + \tau_1 (y_h - \widehat{y}_h^o) \qquad \text{on } \partial \mathcal{T}_h \backslash \varepsilon_h^\partial, \tag{10h}$$

$$\widehat{\boldsymbol{q}}_h \cdot \boldsymbol{n} = \boldsymbol{q}_h \cdot \boldsymbol{n} + h^{-1}(y_h - \mathcal{I}_h g) + \tau_1 (y_h - \mathcal{I}_h g) \quad \text{on } \varepsilon_h^\partial, \tag{10i}$$

$$\widehat{\boldsymbol{p}}_h \cdot \boldsymbol{n} = \boldsymbol{p}_h \cdot \boldsymbol{n} + h^{-1}(z_h - \widehat{z}_h^o) + \tau_2 (z_h - \widehat{z}_h^o) \qquad \text{on } \partial \mathcal{T}_h \backslash \varepsilon_h^\partial, \tag{10j}$$

$$\widehat{\boldsymbol{p}}_h \cdot \boldsymbol{n} = \boldsymbol{p}_h \cdot \boldsymbol{n} + h^{-1} z_h + \tau_2 z_h \qquad \text{on } \varepsilon_h^\partial, \tag{10k}$$

where $\tau_1$ and $\tau_2$ are positive stabilization functions defined on $\partial \mathcal{T}_h$. We show below that the OD and DO approaches coincide if $\tau_2 = \tau_1 - \boldsymbol{\beta} \cdot \boldsymbol{n}$. The implementation of the OD approach is very similar to the HDG method in [18], and hence is omitted here.

### 2.3. Discretize-then-Optimize.

Now we derive the optimality conditions for the discretize-then-optimize (DO) approach when the optimal control problem is discretized by the EDG method. Therefore, we solve

$$\min_{u_h \in W_h} \frac{1}{2} \| y_h - y_d \|_{\mathcal{T}_h}^2 + \frac{\gamma}{2} \| u_h \|_{\mathcal{T}_h}^2, \qquad \gamma > 0, \tag{11}$$

subject to the discrete state equations

$$\begin{aligned} &(\boldsymbol{q}_h, \boldsymbol{r}_1)_{\mathcal{T}_h} - (y_h, \nabla \cdot \boldsymbol{r}_1)_{\mathcal{T}_h} + \langle \widehat{y}_h^o, \boldsymbol{r}_1 \cdot \boldsymbol{n} \rangle_{\partial \mathcal{T}_h \backslash \varepsilon_h^\partial} = -\langle \mathcal{I}_h g, \boldsymbol{r}_1 \cdot \boldsymbol{n} \rangle_{\varepsilon_h^\partial}, \\ &\quad -(\boldsymbol{q}_h + \boldsymbol{\beta} y_h, \nabla w_1)_{\mathcal{T}_h} - (y_h \nabla \cdot \boldsymbol{\beta}, w_1)_{\mathcal{T}_h} + \langle \boldsymbol{q}_h \cdot \boldsymbol{n}, w_1 \rangle_{\partial \mathcal{T}_h} \\ &\quad +\langle (h^{-1} + \tau_1) y_h, w_1 \rangle_{\partial \mathcal{T}_h} + \langle \boldsymbol{\beta} \cdot \boldsymbol{n} - (h^{-1} + \tau_1) y_h^o, w_1 \rangle_{\partial \mathcal{T}_h \backslash \varepsilon_h^\partial} \end{aligned} \tag{12}$$

$$-(u_h, w_1)_{\mathcal{T}_h} = -\langle \boldsymbol{\beta} \cdot \boldsymbol{n} - (h^{-1} + \tau_1) \mathcal{I}_h g, w_1 \rangle_{\partial \mathcal{T}_h \backslash \varepsilon_h^\partial} + (f, w_1)_{\mathcal{T}_h}, \tag{13}$$

$$\langle \boldsymbol{q}_h \cdot \boldsymbol{n} + (h^{-1} + \tau_1)(y_h - \widehat{y}_h^o), \mu_1 \rangle_{\partial \mathcal{T}_h \backslash \varepsilon_h^\partial} = 0, \tag{14}$$

for any $(\boldsymbol{r}_1, w_1, \mu_1) \in \boldsymbol{V}_h \times W_h \times \widetilde{M}_h(o)$.

The discretized Lagrangian functional is defined by

$$
\begin{aligned}
\mathcal{L}_h(\boldsymbol{q}_h, y_h, \widehat{y}_h^o; \boldsymbol{p}_h, z_h, \widehat{z}_h^o) &= \frac{1}{2}\|y_h - y_d\|_{\mathcal{T}_h}^2 + \frac{\gamma}{2}\|u_h\|_{\mathcal{T}_h}^2 \\
&+ (\boldsymbol{q}_h, \boldsymbol{p}_h)_{\mathcal{T}_h} - (y_h, \nabla\cdot\boldsymbol{p}_h)_{\mathcal{T}_h} + \langle \widehat{y}_h^o, \boldsymbol{p}_h\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial} + \langle \mathcal{I}_h g, \boldsymbol{p}_h\cdot\boldsymbol{n}\rangle_{\varepsilon_h^\partial} \\
&+ (\boldsymbol{q}_h + \boldsymbol{\beta} y_h, \nabla z_h)_{\mathcal{T}_h} + (y_h \nabla\cdot\boldsymbol{\beta}, z_h)_{\mathcal{T}_h} - \langle \boldsymbol{q}_h\cdot\boldsymbol{n}, z_h\rangle_{\partial\mathcal{T}_h} \\
&- \langle (h^{-1}+\tau_1) y_h, z_h\rangle_{\partial\mathcal{T}_h} - \langle (\boldsymbol{\beta}\cdot\boldsymbol{n} - h^{-1} - \tau_1)\widehat{y}_h^o, z_h\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial} \\
&+ (u_h, z_h)_{\mathcal{T}_h} - \langle (\boldsymbol{\beta}\cdot\boldsymbol{n} - h^{-1} - \tau_1)\mathcal{I}_h g, z_h\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial} + (f, z_h)_{\mathcal{T}_h} \\
&+ \langle \boldsymbol{q}_h\cdot\boldsymbol{n} + (h^{-1}+\tau_1)(y_h - \widehat{y}_h^o), \widehat{z}_h^o\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial}.
\end{aligned}
\tag{15}
$$

Since the constraint PDE is linear and the cost functional is convex, the necessary and sufficient optimality conditions can be obtained by setting the partial Fréchet-derivatives of (15) with respect to the flux $\boldsymbol{q}_h$, state $y_h$, numerical trace $\widehat{y}_h^o$ and control $u_h$ equal to zero. Thus, we obtain the system consisting of the discrete adjoint equations

$$
\begin{aligned}
\frac{\partial\mathcal{L}_h}{\partial\boldsymbol{q}_h}\boldsymbol{r}_2 &= (\boldsymbol{p}_h, \boldsymbol{r}_2)_{\mathcal{T}_h} + (\nabla z_h, \boldsymbol{r}_2)_{\mathcal{T}_h} - \langle z_h, \boldsymbol{r}_2\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h} + \langle \widehat{z}_h^o, \boldsymbol{r}_2\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial} \\
&= (\boldsymbol{p}_h, \boldsymbol{r}_2)_{\mathcal{T}_h} - (z_h, \nabla\cdot\boldsymbol{r}_2)_{\mathcal{T}_h} + \langle \widehat{z}_h^o, \boldsymbol{r}_2\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial} = 0, \\
\frac{\partial\mathcal{L}_h}{\partial y_h}w_2 &= -(\nabla\cdot\boldsymbol{p}_h, w_2)_{\mathcal{T}_h} + (\boldsymbol{\beta}\nabla z_h, w_2)_{\mathcal{T}_h} + (z_h\nabla\cdot\boldsymbol{\beta}, w_2)_{\mathcal{T}_h} \\
&\quad - \langle (h^{-1}+\tau_1) z_h, w_2\rangle_{\partial\mathcal{T}_h} + \langle (h^{-1}+\tau_1)\widehat{z}_h^o, w_2\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial} + (y_h - y_d, w_2)_{\mathcal{T}_h} \\
&= (\boldsymbol{p}_h - \boldsymbol{\beta} z_h, \nabla w_2)_{\mathcal{T}_h} - \langle \boldsymbol{p}_h\cdot\boldsymbol{n} + (h^{-1} + \tau_1 - \boldsymbol{\beta}\cdot\boldsymbol{n}) z_h, w_2\rangle_{\partial\mathcal{T}_h} \\
&\quad + \langle (h^{-1}+\tau_1)\widehat{z}_h^o, w_2\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial} + (y_h - y_d, w_2)_{\mathcal{T}_h} = 0, \\
\frac{\partial\mathcal{L}_h}{\partial\widehat{y}_h^o}\mu_2 &= \langle \boldsymbol{p}_h\cdot\boldsymbol{n} - (\boldsymbol{\beta}\cdot\boldsymbol{n} - h^{-1} - \tau_1) z_h - (h^{-1}+\tau_1)\widehat{z}_h^o, \mu_2\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial} = 0,
\end{aligned}
$$

Furthermore, we obtain the same optimality condition (10g) as in the OD approach.

$$
\frac{\partial\mathcal{L}_h}{\partial u_h}w_3 = (\gamma u_h + z_h, w_3)_{\mathcal{T}_h} = 0.
$$

In the OD approach, if the stabilization functions $\tau_1$ and $\tau_2$ satisfy

$$
\tau_2 = \tau_1 - \boldsymbol{\beta}\cdot\boldsymbol{n},
\tag{17}
$$

then by comparing the above discrete adjoint equations with (10) we obtain identical discrete systems; therefore, the two approaches coincide in this case, i.e., OD = DO.

### 2.4. Implementation of DO.

In the DO approach, we need to deal with a large optimization problem (11) and (12)-(14) since the EDG method generates three variables: the flux $\boldsymbol{q}_h$, the scalar variable $y_h$, and the numerical trace $\widehat{y}_h$. Fortunately, we can reduce the large scale problem into a smaller problem using the local solver for the EDG method.

#### 2.4.1. Matrix equations.

Assume $\boldsymbol{V}_h = \operatorname{span}\{\boldsymbol{\varphi}_i\}_{i=1}^{N_1}$, $W_h = \operatorname{span}\{\phi_i\}_{i=1}^{N_2}$, and $\widetilde{M}_h(o) = \operatorname{span}\{\psi_i\}_{i=1}^{N_3}$. Then

$$
\boldsymbol{q}_h = \sum_{j=1}^{N_1}\alpha_j\boldsymbol{\varphi}_j, \quad y_h = \sum_{j=1}^{N_2}\beta_j\phi_j, \quad \widehat{y}_h^o = \sum_{j=1}^{N_3}\gamma_j\psi_j, \quad u_h = \sum_{j=1}^{N_2}\zeta_j\phi_j.
\tag{18}
$$

Substitute (18) into (11)-(14) to give the following finite dimensional optimization problem:

$$\min_{\boldsymbol{\zeta}\in\mathbb{R}^{N_2}} \frac{1}{2}\boldsymbol{\beta}^T A_6 \boldsymbol{\beta} - b_1^T\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\zeta}^T A_6 \boldsymbol{\zeta} \tag{19a}$$

subject to

$$\begin{bmatrix} A_1 & -A_2 & A_3 & 0 \\ A_2^T & A_4 & A_5 & -A_6 \\ A_3^T & A_7 & -A_8 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \\ \boldsymbol{\gamma} \\ \boldsymbol{\zeta} \end{bmatrix} = \begin{bmatrix} -b_2 \\ b_3 - b_4 \\ 0 \end{bmatrix}, \tag{19b}$$

where $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\zeta}$ are the coefficient vectors for $\boldsymbol{q}_h, y_h, \widehat{y}_h^o, u_h$, respectively, and

$$\begin{gathered} A_1 = [(\boldsymbol{\varphi}_j, \boldsymbol{\varphi}_i)_{\mathcal{T}_h}], \quad A_2 = [(\phi_j, \nabla\cdot\boldsymbol{\varphi_i})_{\mathcal{T}_h}], \quad A_3 = [\langle \psi_j, \boldsymbol{\varphi}_i\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial}], \\ A_4 = -[(\phi_j, \nabla\cdot(\boldsymbol{\beta}\cdot\phi_i))_{\mathcal{T}_h}] + [\langle (h^{-1}+\tau_1)\phi_j, \phi_i\rangle_{\partial\mathcal{T}_h}], \\ A_5 = [\langle (\boldsymbol{\beta}\cdot\boldsymbol{n} - h^{-1} - \tau_1)\psi_j, \phi_i\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial}], \quad A_6 = [(\phi_j, \phi_i)_{\mathcal{T}_h}], \\ A_7 = [\langle (h^{-1}+\tau_1)\phi_j, \psi_i\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial}], \quad A_8 = [\langle (h^{-1}+\tau_1)\psi_j, \psi_i\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial}], \\ b_1 = [(y_d, \phi_i)_{\mathcal{T}_h}], \quad b_2 = [\langle \mathcal{I}_h g, \boldsymbol{r}_1\cdot\boldsymbol{n}\rangle_{\varepsilon_h^\partial}], \quad b_3 = [(f, \phi_i)_{\mathcal{T}_h}], \\ b_4 = [\langle (\boldsymbol{\beta}\cdot\boldsymbol{n} - h^{-1} - \tau_1) g, \phi_i\rangle_{\varepsilon_h^\partial}]. \end{gathered}$$

Due to the discontinuous nature of the approximation spaces $\boldsymbol{V_h}$ and $W_h$, the first two equations of (19b) can be used to eliminate both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in an element-by-element fashion. As a consequence, we can write system (19b) as

$$\begin{cases} \boldsymbol{\alpha} = G_1\boldsymbol{\gamma} + G_2\boldsymbol{\zeta} + H_1, \\ \boldsymbol{\beta} = G_3\boldsymbol{\gamma} + G_4\boldsymbol{\zeta} + H_2, \\ G_5\boldsymbol{\gamma} + G_6\boldsymbol{\zeta} = H_3. \end{cases} \tag{20}$$

We provide details on the element-by-element construction of the coefficient matrices $G_1, \ldots, G_6$ and $H_1, H_2, H_3$ in the appendix.

Substituting (20) into (19) gives the reduced optimization problem

$$\min_{\boldsymbol{\zeta}\in\mathbb{R}^{N_2}} \frac{1}{2} \begin{bmatrix} \boldsymbol{\gamma}^T & \boldsymbol{\zeta}^T \end{bmatrix} \begin{bmatrix} B_1 & B_2 \\ B_3 & B_4 \end{bmatrix} \begin{bmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\zeta} \end{bmatrix} + \begin{bmatrix} b_5^T & b_6^T \end{bmatrix} \begin{bmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\zeta} \end{bmatrix}, \tag{21a}$$

subject to

$$\begin{bmatrix} G_5 & G_6 \end{bmatrix} \begin{bmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\zeta} \end{bmatrix} = H_3, \tag{21b}$$

where

$$\begin{gathered} B_1 = G_3^T A_6 G_3, \quad B_2 = G_3^T A_6 G_4, \quad B_3 = G_4^T A_6 G_3, \quad B_4 = G_4^T A_6 G_4 + A_6, \\ b_5 = G_3^T(A_6 H_2 - b_1), \quad b_6 = G_4^T(A_6 H_2 - b_1). \end{gathered}$$

**Remark 1.** In the DO approach, we need to solve the optimization problem (21); there are many existing optimization algorithms [14] that can efficiently solve this problem.

## 3. Error Analysis

Next, we provide a convergence analysis of the above EDG method for the optimal control problem. Throughout this section, we assume $\boldsymbol{\beta} \in [W^{1,\infty}(\Omega)]^d$, $\Omega$ is a bounded convex polyhedral domain, the solution is smooth enough, and $h \leq 1$.

**3.1. Main result.** For our theoretical results, we require the stabilization functions $\tau_1$ and $\tau_2$ are chosen to satisfy

**(A1):** $\tau_2 = \tau_1 - \boldsymbol{\beta} \cdot \boldsymbol{n}$.
**(A2):** For any $K \in \mathcal{T}_h$, $\min{(\tau_1 - \frac{1}{2}\boldsymbol{\beta} \cdot \boldsymbol{n})}|_{\partial K} > 0$.

We note that **(A1)** and **(A2)** imply

$$\min{(\tau_2 + \frac{1}{2}\boldsymbol{\beta} \cdot \boldsymbol{n})}|_{\partial K} > 0 \quad \text{for any } K \in \mathcal{T}_h. \tag{22}$$

Furthermore, **(A1)** implies the OD and DO approaches yield equivalent results; therefore, all of our convergence analysis is for the OD approach.

**Theorem 1.** We have

$$\begin{aligned}
\|\boldsymbol{q} - \boldsymbol{q}_h\|_{\mathcal{T}_h} &\lesssim h^{k+1}(|\boldsymbol{q}|_{k+1} + |y|_{k+2} + |\boldsymbol{p}|_{k+1} + |z|_{k+2}),\\
\|\boldsymbol{p} - \boldsymbol{p}_h\|_{\mathcal{T}_h} &\lesssim h^{k+1}(|\boldsymbol{q}|_{k+1} + |y|_{k+2} + |\boldsymbol{p}|_{k+1} + |z|_{k+2}),\\
\|y - y_h\|_{\mathcal{T}_h} &\lesssim h^{k+2}(|\boldsymbol{q}|_{k+1} + |y|_{k+2} + |\boldsymbol{p}|_{k+1} + |z|_{k+2}),\\
\|z - z_h\|_{\mathcal{T}_h} &\lesssim h^{k+2}(|\boldsymbol{q}|_{k+1} + |y|_{k+2} + |\boldsymbol{p}|_{k+1} + |z|_{k+2}),\\
\|u - u_h\|_{\mathcal{T}_h} &\lesssim h^{k+2}(|\boldsymbol{q}|_{k+1} + |y|_{k+2} + |\boldsymbol{p}|_{k+1} + |z|_{k+2}).
\end{aligned}$$

**3.2. Preliminary material.** Next, we introduce the standard $L^2$-orthogonal projection operators $\boldsymbol{\Pi}_V$ and $\Pi_W$ as follows:

$$(\boldsymbol{\Pi}_V \boldsymbol{q}, \boldsymbol{r})_K = (\boldsymbol{q}, \boldsymbol{r})_K \quad \forall \boldsymbol{r} \in [\mathcal{P}_k(K)]^d, \tag{23a}$$

$$(\Pi_W y, w)_K = (y, w)_K \quad \forall w \in \mathcal{P}_{k+1}(K). \tag{23b}$$

We use the following well-known bounds:

$$\|\boldsymbol{q} - \boldsymbol{\Pi}_V \boldsymbol{q}\|_{\mathcal{T}_h} \leq Ch^{k+1} \|\boldsymbol{q}\|_{k+1,\Omega}, \ \|y - \Pi_W y\|_{\mathcal{T}_h} \leq Ch^{k+2} \|y\|_{k+2,\Omega}, \tag{24a}$$

$$\|y - \Pi_W y\|_{\partial \mathcal{T}_h} \leq Ch^{k+\frac{3}{2}} \|y\|_{k+2,\Omega}, \ \|\boldsymbol{q} - \boldsymbol{\Pi}_V \boldsymbol{q}\|_{\partial \mathcal{T}_h} \leq Ch^{k+\frac{1}{2}} \|\boldsymbol{q}\|_{k+1,\Omega}, \tag{24b}$$

$$\|y - \mathcal{I}_h y\|_{\partial \mathcal{T}_h} \leq Ch^{k+\frac{3}{2}} \|y\|_{k+2,\Omega}, \ \|w\|_{\partial \mathcal{T}_h} \leq Ch^{-\frac{1}{2}} \|w\|_{\mathcal{T}_h}, \forall w \in W_h, \tag{24c}$$

where $\mathcal{I}_h$ is the continuous interpolation operator introduced earlier.

We define the following EDG operators $\mathscr{B}_1$ and $\mathscr{B}_2$:

$$\begin{aligned}
&\mathscr{B}_1(\boldsymbol{q}_h, y_h, \widehat{y}_h^o; \boldsymbol{r}_1, w_1, \mu_1)\\
&\quad= (\boldsymbol{q}_h, \boldsymbol{r}_1)_{\mathcal{T}_h} - (y_h, \nabla\cdot\boldsymbol{r}_1)_{\mathcal{T}_h} + \langle \widehat{y}_h^o, \boldsymbol{r}_1\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial}\\
&\qquad - (\boldsymbol{q}_h + \boldsymbol{\beta} y_h, \nabla w_1)_{\mathcal{T}_h} - (\nabla\cdot\boldsymbol{\beta} y_h, w_1)_{\mathcal{T}_h}\\
&\qquad + \langle \boldsymbol{q}_h\cdot\boldsymbol{n} + (h^{-1}+\tau_1) y_h, w_1\rangle_{\partial\mathcal{T}_h} + \langle (\boldsymbol{\beta}\cdot\boldsymbol{n} - h^{-1} - \tau_1)\widehat{y}_h^o, w_1\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial}\\
&\qquad - \langle \boldsymbol{q}_h\cdot\boldsymbol{n} + \boldsymbol{\beta}\cdot\boldsymbol{n}\widehat{y}_h^o + (h^{-1}+\tau_1)(y_h - \widehat{y}_h^o), \mu_1\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial},
\end{aligned}\tag{25}$$

$$\begin{aligned}
&\mathscr{B}_2(\boldsymbol{p}_h, z_h, \widehat{z}_h^o; \boldsymbol{r}_2, w_2, \mu_2)\\
&\quad= (\boldsymbol{p}_h, \boldsymbol{r}_2)_{\mathcal{T}_h} - (z_h, \nabla\cdot\boldsymbol{r}_2)_{\mathcal{T}_h} + \langle \widehat{z}_h^o, \boldsymbol{r}_2\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial} - (\boldsymbol{p}_h - \boldsymbol{\beta} z_h, \nabla w_2)_{\mathcal{T}_h}\\
&\qquad + \langle \boldsymbol{p}_h\cdot\boldsymbol{n} + (h^{-1}+\tau_2) z_h, w_2\rangle_{\partial\mathcal{T}_h} - \langle (\boldsymbol{\beta}\cdot\boldsymbol{n} + h^{-1} + \tau_2)\widehat{z}_h^o, w_2\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial}\\
&\qquad - \langle \boldsymbol{p}_h\cdot\boldsymbol{n} - \boldsymbol{\beta}\cdot\boldsymbol{n}\widehat{z}_h^o + (h^{-1}+\tau_2)(z_h - \widehat{z}_h^o), \mu_2\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial}.
\end{aligned}\tag{26}$$

By the definition of $\mathscr{B}_1$ and $\mathscr{B}_2$, we can rewrite the EDG formulation of the optimality system (10) as follows: find $(\boldsymbol{q}_h, \boldsymbol{p}_h, y_h, z_h, u_h, \widehat{y}_h^o, \widehat{z}_h^o) \in \boldsymbol{V}_h\times\boldsymbol{V}_h\times W_h\times W_h\times W_h\times \widetilde{M}_h(o)\times\widetilde{M}_h(o)$ such that

$$\begin{aligned}
\mathscr{B}_1(\boldsymbol{q}_h, y_h, \widehat{y}_h^o; \boldsymbol{r}_1, w_1, \mu_1) &= (f + u_h, w_1)_{\mathcal{T}_h}\\
&\qquad - \langle \mathcal{I}_h g, (\boldsymbol{\beta}\cdot\boldsymbol{n} - \tau_1 - h^{-1}) w_1 + \boldsymbol{r}_1\cdot\boldsymbol{n}\rangle_{\varepsilon_h^\partial},
\end{aligned}\tag{27a}$$

$$\mathscr{B}_2(\boldsymbol{p}_h, z_h, \widehat{z}_h^o; \boldsymbol{r}_2, w_2, \mu_2) = (y_h - y_d, w_2)_{\mathcal{T}_h},\tag{27b}$$

$$(z_h + \gamma u_h, w_3)_{\mathcal{T}_h} = 0,\tag{27c}$$

for all $(\boldsymbol{r}_1, \boldsymbol{r}_2, w_1, w_2, w_3, \mu_1, \mu_2) \in \boldsymbol{V}_h\times\boldsymbol{V}_h\times W_h\times W_h\times W_h\times\widetilde{M}_h(o)\times\widetilde{M}_h(o)$.

Next, we present two fundamental properties of the operators $\mathscr{B}_1$ and $\mathscr{B}_2$, and show the EDG equations (27) have a unique solution. The proofs of these results are similar to proofs in [17, 18] and are omitted. We note that condition **(A1)** is used in the proof of Lemma 2, which is fundamental to the error analysis in this work. Furthermore, **(A1)** and **(A2)** are used in the proof of Proposition 1.

**Lemma 1.** For any $(\boldsymbol{v}_h, w_h, \mu_h) \in \boldsymbol{V}_h\times W_h\times\widetilde{M}_h$, we have

$$\begin{aligned}
&\mathscr{B}_1(\boldsymbol{v}_h, w_h, \mu_h; \boldsymbol{v}_h, w_h, \mu_h)\\
&\quad= (\boldsymbol{v}_h, \boldsymbol{v}_h)_{\mathcal{T}_h} + \langle (h^{-1} + \tau_1 - \frac{1}{2}\boldsymbol{\beta}\cdot\boldsymbol{n})(w_h - \mu_h), w_h - \mu_h\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial}\\
&\qquad - \frac{1}{2}(\nabla\cdot\boldsymbol{\beta} w_h, w_h)_{\mathcal{T}_h} + \langle (h^{-1} + \tau_1 - \frac{1}{2}\boldsymbol{\beta}\cdot\boldsymbol{n}) w_h, w_h\rangle_{\varepsilon_h^\partial},\\
&\mathscr{B}_2(\boldsymbol{v}_h, w_h, \mu_h; \boldsymbol{v}_h, w_h, \mu_h)\\
&\quad= (\boldsymbol{v}_h, \boldsymbol{v}_h)_{\mathcal{T}_h} + \langle (h^{-1} + \tau_2 + \frac{1}{2}\boldsymbol{\beta}\cdot\boldsymbol{n})(w_h - \mu_h), w_h - \mu_h\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial}\\
&\qquad - \frac{1}{2}(\nabla\cdot\boldsymbol{\beta} w_h, w_h)_{\mathcal{T}_h} + \langle (h^{-1} + \tau_2 + \frac{1}{2}\boldsymbol{\beta}\cdot\boldsymbol{n}) w_h, w_h\rangle_{\varepsilon_h^\partial}.
\end{aligned}$$

**Lemma 2.** The EDG operators satisfy

$$\mathscr{B}_1(\boldsymbol{q}_h, y_h, \widehat{y}_h^o; \boldsymbol{p}_h, -z_h, -\widehat{z}_h^o) + \mathscr{B}_2(\boldsymbol{p}_h, z_h, \widehat{z}_h^o; -\boldsymbol{q}_h, y_h, \widehat{y}_h^o) = 0.$$

**Proposition 1.** There exists a unique solution of the EDG equations (27).

**3.3. Proof of Main Result.** To prove the convergence result, we split the proof into eight steps. We first consider the following auxiliary problem: find

$$(\boldsymbol{q}_h(u), \boldsymbol{p}_h(u), y_h(u), z_h(u), \widehat{y}_h^o(u), \widehat{z}_h^o(u)) \in \boldsymbol{V}_h \times \boldsymbol{V}_h \times W_h \times W_h \times \widetilde{M}_h(o) \times \widetilde{M}_h(o)$$

such that

$$\begin{aligned} \mathscr{B}_1(\boldsymbol{q}_h(u), y_h(u), \widehat{y}_h^o(u); \boldsymbol{r}_1, w_1, \mu_1) &= (f + u, w_1)_{\mathcal{T}_h} \\ &\quad - \langle \mathcal{I}_h g, (\boldsymbol{\beta} \cdot \boldsymbol{n} - \tau_1 - h^{-1}) w_1 + \boldsymbol{r}_1 \cdot \boldsymbol{n} \rangle_{\varepsilon_h^\partial}, \end{aligned} \tag{28a}$$

$$\mathscr{B}_2(\boldsymbol{p}_h(u), z_h(u), \widehat{z}_h^o(u); \boldsymbol{r}_2, w_2, \mu_2) = (y_h(u) - y_d, w_2)_{\mathcal{T}_h}, \tag{28b}$$

for all $(\boldsymbol{r}_1, \boldsymbol{r}_2, w_1, w_2, \mu_1, \mu_2) \in \boldsymbol{V}_h \times \boldsymbol{V}_h \times W_h \times W_h \times \widetilde{M}_h(o) \times \widetilde{M}_h(o)$.

In Steps 1-3, we focus on the primary variables, i.e., the state $y$ and the flux $\boldsymbol{q}$, and we use the following notation:

$$\begin{aligned} \delta^{\boldsymbol{q}} &= \boldsymbol{q} - \boldsymbol{\Pi}_V \boldsymbol{q}, & \varepsilon_h^{\boldsymbol{q}} &= \boldsymbol{\Pi}_V \boldsymbol{q} - \boldsymbol{q}_h(u), \\ \delta^y &= y - \Pi_W y, & \varepsilon_h^y &= \Pi_W y - y_h(u), \\ \delta^{\widehat{y}} &= y - \mathcal{I}_h y, & \varepsilon_h^{\widehat{y}} &= \mathcal{I}_h y - \widehat{y}_h(u), \\ \widehat{\boldsymbol{\delta}}_1 &= \delta^{\boldsymbol{q}} \cdot \boldsymbol{n} + \boldsymbol{\beta} \cdot \boldsymbol{n} \delta^{\widehat{y}} + (\tau_1 + h^{-1})(\delta^y - \delta^{\widehat{y}}), \end{aligned} \tag{29}$$

where $\widehat{y}_h(u) = \widehat{y}_h^o(u)$ on $\varepsilon_h^o$ and $\widehat{y}_h(u) = \mathcal{I}_h g$ on $\varepsilon_h^\partial$, which implies $\varepsilon_h^{\widehat{y}} = 0$ on $\varepsilon_h^\partial$.

#### 3.3.1. Step 1: The error equation for part 1 of the auxiliary problem (28a).

**Lemma 3.** We have the following error equation

$$\begin{aligned} \mathscr{B}_1(\varepsilon_h^{\boldsymbol{q}}, \varepsilon_h^y, \varepsilon_h^{\widehat{y}}; \boldsymbol{r}_1, w_1, \mu_1) &= -\langle \delta^{\widehat{y}}, \boldsymbol{r}_1 \cdot \boldsymbol{n} \rangle_{\partial \mathcal{T}_h} + (\boldsymbol{\beta} \delta^y, \nabla w_1)_{\mathcal{T}_h} + (\nabla \cdot \boldsymbol{\beta} \delta^y, w_1)_{\mathcal{T}_h} \\ &\quad - \langle \widehat{\boldsymbol{\delta}}_1, w_1 \rangle_{\partial \mathcal{T}_h} + \langle \widehat{\boldsymbol{\delta}}_1, \mu_1 \rangle_{\partial \mathcal{T}_h \backslash \varepsilon_h^\partial}. \end{aligned} \tag{30}$$

*Proof.* By definition of the operator $\mathscr{B}_1$ in (25), we have

$$\begin{aligned} &\mathscr{B}_1(\boldsymbol{\Pi}_V \boldsymbol{q}, \Pi_W y, \mathcal{I}_h y; \boldsymbol{r}_1, w_1, \mu_1) \\ &\quad = (\boldsymbol{\Pi}_V \boldsymbol{q}, \boldsymbol{r}_1)_{\mathcal{T}_h} - (\Pi_W y, \nabla \cdot \boldsymbol{r}_1)_{\mathcal{T}_h} + \langle \mathcal{I}_h y, \boldsymbol{r}_1 \cdot \boldsymbol{n} \rangle_{\partial \mathcal{T}_h \backslash \varepsilon_h^\partial} \\ &\quad\quad - (\boldsymbol{\Pi}_V \boldsymbol{q} + \boldsymbol{\beta} \Pi_W y, \nabla w_1)_{\mathcal{T}_h} - (\nabla \cdot \boldsymbol{\beta} \Pi y, w_1)_{\mathcal{T}_h} \\ &\quad\quad + \langle \boldsymbol{\Pi}_V \boldsymbol{q} \cdot \boldsymbol{n} + (\tau_1 + h^{-1}) \Pi_W y, w_1 \rangle_{\partial \mathcal{T}_h} + \langle (\boldsymbol{\beta} \cdot \boldsymbol{n} - \tau_1 - h^{-1}) \mathcal{I}_h y, w_1 \rangle_{\partial \mathcal{T}_h \backslash \varepsilon_h^\partial} \\ &\quad\quad - \langle \boldsymbol{\Pi}_V \boldsymbol{q} \cdot \boldsymbol{n} + \boldsymbol{\beta} \cdot \boldsymbol{n} \mathcal{I}_h y + (\tau_1 + h^{-1})(\Pi_W y - \mathcal{I}_h y), \mu_1 \rangle_{\partial \mathcal{T}_h \backslash \varepsilon_h^\partial}. \end{aligned}$$

Using properties of the $L^2$-orthogonal projection operators (23) gives

$$\begin{aligned} &\mathscr{B}_1(\boldsymbol{\Pi}_V \boldsymbol{q}, \Pi_W y, \mathcal{I}_h y; \boldsymbol{r}_1, w_1, \mu_1) \\ &\quad = (\boldsymbol{q}, \boldsymbol{r}_1)_{\mathcal{T}_h} - (y, \nabla \cdot \boldsymbol{r}_1)_{\mathcal{T}_h} + \langle y, \boldsymbol{r}_1 \cdot \boldsymbol{n} \rangle_{\partial \mathcal{T}_h \backslash \varepsilon_h^\partial} - \langle \delta^{\widehat{y}}, \boldsymbol{r}_1 \cdot \boldsymbol{n} \rangle_{\partial \mathcal{T}_h \backslash \varepsilon_h^\partial} \\ &\quad\quad - (\boldsymbol{q} + \boldsymbol{\beta} y, \nabla w_1)_{\mathcal{T}_h} + (\boldsymbol{\beta} \delta^y, \nabla w_1)_{\mathcal{T}_h} - (\nabla \cdot \boldsymbol{\beta} y, w_1)_{\mathcal{T}_h} + (\nabla \cdot \boldsymbol{\beta} \delta^y, w_1)_{\mathcal{T}_h} \\ &\quad\quad + \langle \boldsymbol{q} \cdot \boldsymbol{n} + (\tau_1 + h^{-1}) y, w_1 \rangle_{\partial \mathcal{T}_h} - \langle \delta^{\boldsymbol{q}} \cdot \boldsymbol{n} + (\tau_1 + h^{-1}) \delta^y, w_1 \rangle_{\partial \mathcal{T}_h} \\ &\quad\quad + \langle (\boldsymbol{\beta} \cdot \boldsymbol{n} - \tau_1 - h^{-1}) y, w_1 \rangle_{\partial \mathcal{T}_h \backslash \varepsilon_h^\partial} - \langle (\boldsymbol{\beta} \cdot \boldsymbol{n} - \tau_1 - h^{-1}) \delta^{\widehat{y}}, w_1 \rangle_{\partial \mathcal{T}_h \backslash \varepsilon_h^\partial} \\ &\quad\quad - \langle \boldsymbol{q} \cdot \boldsymbol{n} + \boldsymbol{\beta} \cdot \boldsymbol{n} y, \mu_1 \rangle_{\partial \mathcal{T}_h \backslash \varepsilon_h^\partial} \\ &\quad\quad + \langle \delta^{\boldsymbol{q}} \cdot \boldsymbol{n} + \boldsymbol{\beta} \cdot \boldsymbol{n} \delta^{\widehat{y}} + (\tau_1 + h^{-1})(\delta^y - \delta^{\widehat{y}}), \mu_1 \rangle_{\partial \mathcal{T}_h \backslash \varepsilon_h^\partial}. \end{aligned}$$

Note that the exact solution $\boldsymbol{q}$ and $y$ satisfies

$$
\begin{aligned}
(\boldsymbol{q},\boldsymbol{r}_1)_{\mathcal{T}_h}-(y,\nabla\cdot\boldsymbol{r}_1)_{\mathcal{T}_h}+\langle y,\boldsymbol{r}_1\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial}&=-\langle g,\boldsymbol{r}_1\cdot\boldsymbol{n}\rangle_{\varepsilon_h^\partial},\\
-(\boldsymbol{q}+\boldsymbol{\beta}y,\nabla w_1)_{\mathcal{T}_h}-(\nabla\cdot\boldsymbol{\beta}y,w_1)_{\mathcal{T}_h}+\langle\boldsymbol{q}\cdot\boldsymbol{n}+\boldsymbol{\beta}\cdot\boldsymbol{n}y,w_1\rangle_{\partial\mathcal{T}_h}&=(f+u,w_1)_{\mathcal{T}_h},\\
-\langle\boldsymbol{q}\cdot\boldsymbol{n}+\boldsymbol{\beta}\cdot\boldsymbol{n}y,\mu_1\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial}&=0,
\end{aligned}
$$

for all $(\boldsymbol{r}_1,w_1,\mu_1)\in\boldsymbol{V}_h\times W_h\times\widetilde{M}_h(o)$. Therefore, we have

$$
\begin{aligned}
&\mathscr{B}_1(\boldsymbol{\Pi}_V\boldsymbol{q},\Pi_W y,I_h y;\boldsymbol{r}_1,w_1,\mu_1)\\
&\quad=-\langle g,\boldsymbol{r}_1\cdot\boldsymbol{n}\rangle_{\varepsilon_h^\partial}-\langle\delta^{\widehat{y}},\boldsymbol{r}_1\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial}+(\boldsymbol{\beta}\delta^y,\nabla w_1)_{\mathcal{T}_h}\\
&\qquad+(\nabla\cdot\boldsymbol{\beta}\delta^y,w_1)_{\mathcal{T}_h}+(f+u,w_1)_{\mathcal{T}_h}-\langle\delta^{\boldsymbol{q}}\cdot\boldsymbol{n},w_1\rangle_{\partial\mathcal{T}_h}\\
&\qquad-\langle(\boldsymbol{\beta}\cdot\boldsymbol{n}-\tau_1-h^{-1})y,w_1\rangle_{\varepsilon_h^\partial}-\langle(\boldsymbol{\beta}\cdot\boldsymbol{n}-\tau_1-h^{-1})\delta^{\widehat{y}},w_1\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial}\\
&\qquad+\langle\delta^{\boldsymbol{q}}\cdot\boldsymbol{n}+\boldsymbol{\beta}\cdot\boldsymbol{n}\delta^{\widehat{y}}+(\tau_1+h^{-1})(\delta^y-\delta^{\widehat{y}}),\mu_1\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial}.
\end{aligned}
$$

Finally, subtracting (28a) from the above equation completes the proof. □

**3.3.2. Step 2: Estimate for $\varepsilon_h^q$ by an energy argument.** First, we give an auxiliary result that is very similar to a result from [30]. The proof is also very similar, and is omitted.

**Lemma 4.** We have

$$
\|\nabla\varepsilon_h^y\|_{\mathcal{T}_h}\lesssim\|\varepsilon_h^{\boldsymbol{q}}\|_{\mathcal{T}_h}+h^{-\frac{1}{2}}\|\varepsilon_h^y-\varepsilon_h^{\widehat{y}}\|_{\partial\mathcal{T}_h}. \tag{31}
$$

**Lemma 5.** We have

$$
\|\varepsilon_h^{\boldsymbol{q}}\|_{\mathcal{T}_h}+h^{-\frac{1}{2}}\|\varepsilon_h^y-\varepsilon_h^{\widehat{y}}\|_{\partial\mathcal{T}_h}\lesssim h^{k+1}(|\boldsymbol{q}|_{k+1}+|y|_{k+2}). \tag{32}
$$

*Proof.* Taking $(\boldsymbol{r}_1,w_1,\mu_1)=(\varepsilon_h^{\boldsymbol{q}},\varepsilon_h^y,\varepsilon_h^{\widehat{y}})$ in (30) in Lemma 3 gives

$$
\begin{aligned}
\mathscr{B}_1(\varepsilon_h^{\boldsymbol{q}},\varepsilon_h^y,\varepsilon_h^{\widehat{y}};\varepsilon_h^{\boldsymbol{q}},\varepsilon_h^y,\varepsilon_h^{\widehat{y}})&=-\langle\delta^{\widehat{y}},\varepsilon_h^{\boldsymbol{q}}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h}+(\boldsymbol{\beta}\delta^y,\nabla\varepsilon_h^y)_{\mathcal{T}_h}\\
&\quad+(\nabla\cdot\boldsymbol{\beta}\delta^y,w_1)_{\mathcal{T}_h}-\langle\widehat{\boldsymbol{\delta}}_1,\varepsilon_h^y-\varepsilon_h^{\widehat{y}}\rangle_{\partial\mathcal{T}_h}\\
&=:T_1+T_2+T_3+T_4,
\end{aligned}
$$

where we used $\varepsilon_h^{\widehat{y}}=0$ on $\varepsilon_h^\partial$. We estimate $T_i$, for $i=1,2,3,4$, as follows. First,

$$
T_1\leq Ch^{-1}\|\delta^{\widehat{y}}\|_{\partial\mathcal{T}_h}^2+\frac{1}{4}\|\varepsilon_h^{\boldsymbol{q}}\|_{\mathcal{T}_h}^2,
$$

where we used trace and inverse inequalities. For the second term $T_2$, by Lemma 4, we have

$$
T_2\leq C\|\delta^y\|_{\mathcal{T}_h}^2+\frac{1}{4}\|\varepsilon_h^{\boldsymbol{q}}\|_{\mathcal{T}_h}^2+\frac{1}{4h}\|\varepsilon_h^y-\varepsilon_h^{\widehat{y}}\|_{\partial\mathcal{T}_h}^2.
$$

For the third term $T_3$, we have

$$
T_3\leq C\|\delta^y\|_{\mathcal{T}_h}^2+\frac{1}{2}\|(-\nabla\cdot\boldsymbol{\beta})^{\frac{1}{2}}\varepsilon_h^y\|_{\mathcal{T}_h}^2.
$$

For the last term $T_4$,

$$
T_4\leq Ch\|\widehat{\delta}_1\|_{\partial\mathcal{T}_h}^2+\frac{1}{4h}\|\varepsilon_h^y-\varepsilon_h^{\widehat{y}}\|_{\partial\mathcal{T}_h}.
$$

Sum all the estimates for $\{T_i\}_{i=1}^4$ to obtain

$$\begin{aligned}\|\varepsilon_h^{\boldsymbol{q}}\|_{\mathcal{T}_h}^2 + h^{-1}\|\varepsilon_h^y - \varepsilon_h^{\widehat{y}}\|_{\partial\mathcal{T}_h}^2 &\lesssim h^{-1}\|\delta^{\widehat{y}}\|_{\partial\mathcal{T}_h}^2 + \|\delta^y\|_{\mathcal{T}_h}^2 + h\|\widehat{\delta}_1\|_{\partial\mathcal{T}_h}^2\\ &\lesssim h^{2(k+1)}(|\boldsymbol{q}|_{k+1}^2 + |y|_{k+2}^2).\end{aligned}$$

□

### 3.3.3. Step 3: Estimate for $\varepsilon_h^y$ by a duality argument.

Next, we introduce the dual problem for any given $\Theta$ in $L^2(\Omega)$:

$$\begin{aligned}\boldsymbol{\Phi} + \nabla\Psi &= 0 &&\text{in } \Omega,\\ \nabla\cdot(\boldsymbol{\Phi} - \boldsymbol{\beta}\Psi) &= \Theta &&\text{in } \Omega,\\ \Psi &= 0 &&\text{on } \partial\Omega.\end{aligned}\tag{33}$$

Since the domain $\Omega$ is convex, we have the following regularity estimate

$$\|\boldsymbol{\Phi}\|_{1,\Omega} + \|\Psi\|_{2,\Omega} \le C_{\rm reg}\,\|\Theta\|_{\Omega}\,.\tag{34}$$

We use the following notation below:

$$\delta^{\boldsymbol{\Phi}} = \boldsymbol{\Phi} - \boldsymbol{\Pi}_V\boldsymbol{\Phi}, \quad \delta^{\Psi} = \Psi - \Pi_W\Psi, \quad \delta^{\widehat{\Psi}} = \Psi - \mathcal{I}_h\Psi.\tag{35}$$

**Lemma 6.** We have

$$\|\varepsilon_h^y\|_{\mathcal{T}_h} \lesssim h^{k+2}(|\boldsymbol{q}|_{k+1} + |y|_{k+2}).\tag{36}$$

*Proof.* First we take $(\boldsymbol{r}_1, w_1, \mu_1) = (\boldsymbol{\Pi}_V\boldsymbol{\Phi}, -\Pi_W\Psi, -\mathcal{I}_h\Psi)$ in equation (30) to get

$$\begin{aligned}&\mathscr{B}_1(\varepsilon_h^{\boldsymbol{q}}, \varepsilon_h^y, \varepsilon_h^{\widehat{y}}; \boldsymbol{\Pi}_V\boldsymbol{\Phi}, -\Pi_W\Psi, -\mathcal{I}_h\Psi)\\ &\qquad= (\varepsilon_h^{\boldsymbol{q}}, \boldsymbol{\Pi}_V\boldsymbol{\Phi})_{\mathcal{T}_h} - (\varepsilon_h^y, \nabla\cdot\boldsymbol{\Pi}_V\boldsymbol{\Phi})_{\mathcal{T}_h} + \langle\varepsilon_h^{\widehat{y}}, \boldsymbol{\Pi}_V\boldsymbol{\Phi}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial}\\ &\qquad\quad+ (\varepsilon_h^{\boldsymbol{q}} + \boldsymbol{\beta}\varepsilon_h^y, \nabla\Pi_W\Psi)_{\mathcal{T}_h} + (\nabla\cdot\boldsymbol{\beta}\varepsilon_h^y, \Pi_W\Psi)_{\mathcal{T}_h}\\ &\qquad\quad- \langle\varepsilon_h^{\boldsymbol{q}}\cdot\boldsymbol{n} + (h^{-1}+\tau_1)\varepsilon_h^y, \Pi_W\Psi\rangle_{\partial\mathcal{T}_h}\\ &\qquad\quad- \langle(\boldsymbol{\beta}\cdot\boldsymbol{n} - h^{-1} - \tau_1)\varepsilon_h^{\widehat{y}}, \Pi_W\Psi\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial}\\ &\qquad\quad+ \langle\varepsilon_h^{\boldsymbol{q}}\cdot\boldsymbol{n} + \boldsymbol{\beta}\cdot\boldsymbol{n}\varepsilon_h^{\widehat{y}} + (h^{-1}+\tau_1)(\varepsilon_h^y - \varepsilon_h^{\widehat{y}}), \mathcal{I}_h\Psi\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial}.\end{aligned}$$

Moreover, we have

$$\begin{aligned}-(\varepsilon_h^y, \nabla\cdot\boldsymbol{\Pi}_V\boldsymbol{\Phi})_{\partial\mathcal{T}_h} &= (\nabla\varepsilon_h^y, \boldsymbol{\Phi})_{\mathcal{T}_h} - \langle\varepsilon_h^y, \boldsymbol{\Pi}_V\Phi\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h}\\ &= -(\varepsilon_h^y, \nabla\cdot\boldsymbol{\Phi})_{\mathcal{T}_h} + \langle\varepsilon_h^y, \delta^{\boldsymbol{\Phi}}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h},\\ (\varepsilon_h^{\boldsymbol{q}}, \nabla\Pi_W\Psi)_{\mathcal{T}_h} &= -(\nabla\cdot\varepsilon_h^{\boldsymbol{q}}, \Psi)_{\mathcal{T}_h} + \langle\varepsilon_h^{\boldsymbol{q}}\cdot\boldsymbol{n}, \Pi_W\Psi\rangle_{\partial\mathcal{T}_h}\\ &= (\varepsilon_h^{\boldsymbol{q}}, \nabla\Psi)_{\mathcal{T}_h} - \langle\varepsilon_h^{\boldsymbol{q}}\cdot\boldsymbol{n}, \delta^{\Psi}\rangle_{\partial\mathcal{T}_h},\\ (\boldsymbol{\beta}\varepsilon_h^y, \nabla\Pi_W\Psi)_{\mathcal{T}_h} + (\nabla\cdot\boldsymbol{\beta}\varepsilon_h^y, \Pi_W\Psi)_{\mathcal{T}_h} &= (\varepsilon_h^y, \nabla\cdot(\boldsymbol{\beta}\Pi_W\Psi))_{\mathcal{T}_h}\\ &= (\varepsilon_h^y, \nabla\cdot(\boldsymbol{\beta}\Psi))_{\mathcal{T}_h} - (\varepsilon_h^y, \nabla\cdot(\boldsymbol{\beta}\delta^{\Psi}))_{\mathcal{T}_h}\\ &= (\varepsilon_h^y, \nabla\cdot(\boldsymbol{\beta}\Psi))_{\mathcal{T}_h} + (\boldsymbol{\beta}\cdot(\nabla\varepsilon_h^y), \delta^{\Psi})_{\mathcal{T}_h}\\ &\quad\;- \langle\boldsymbol{\beta}\cdot\boldsymbol{n}\varepsilon_h^y, \delta^{\Psi}\rangle_{\partial\mathcal{T}_h}.\end{aligned}$$

Together with the dual problem (33), using $\Theta = -\varepsilon_h^{y}$, we have

$$\begin{aligned}
&\mathscr{B}(\varepsilon_h^{\boldsymbol{q}}, \varepsilon_h^{y}, \varepsilon_h^{\widehat{y}}; \boldsymbol{\Pi}_V \boldsymbol{\Phi}, -\Pi_W \Psi, -\mathcal{I}_h \Psi)\\
&\qquad = (\varepsilon_h^{\boldsymbol{q}}, \boldsymbol{\Phi})_{\mathcal{T}_h} - (\varepsilon_h^{y}, \nabla\cdot\boldsymbol{\Phi})_{\mathcal{T}_h} + \langle \varepsilon_h^{y} - \varepsilon_h^{\widehat{y}}, \delta^{\boldsymbol{\Phi}}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h}\\
&\qquad\quad + (\varepsilon_h^{\boldsymbol{q}}, \nabla\Psi)_{\mathcal{T}_h} - \langle \varepsilon_h^{\boldsymbol{q}}\cdot\boldsymbol{n}, \delta^{\Psi}\rangle_{\partial\mathcal{T}_h} + (\varepsilon_h^{y}, \nabla\cdot(\boldsymbol{\beta}\Psi))_{\mathcal{T}_h}\\
&\qquad\quad + (\boldsymbol{\beta}\cdot(\nabla\varepsilon_h^{y}), \delta^{\Psi})_{\mathcal{T}_h} - \langle \boldsymbol{\beta}\cdot\boldsymbol{n}\varepsilon_h^{y}, \delta^{\Psi}\rangle_{\partial\mathcal{T}_h}\\
&\qquad\quad - \langle (\tau_1 + h^{-1})(\varepsilon_h^{y} - \varepsilon_h^{\widehat{y}}) + \boldsymbol{\beta}\cdot\boldsymbol{n}\varepsilon_h^{\widehat{y}}, \Pi_W\Psi\rangle_{\partial\mathcal{T}_h}\\
&\qquad\quad + \langle \varepsilon_h^{\boldsymbol{q}}\cdot\boldsymbol{n} + (\tau_1 + h^{-1})(\varepsilon_h^{y} - \varepsilon_h^{\widehat{y}}) + \boldsymbol{\beta}\cdot\boldsymbol{n}\varepsilon_h^{\widehat{y}}, \mathcal{I}_h\Psi\rangle_{\partial\mathcal{T}_h}\\
&\qquad = (\varepsilon_h^{y}, \varepsilon_h^{y})_{\mathcal{T}_h} + \langle \varepsilon_h^{y} - \varepsilon_h^{\widehat{y}}, \delta^{\boldsymbol{\Phi}}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h} - \langle \varepsilon_h^{\boldsymbol{q}}\cdot\boldsymbol{n}, \delta^{\widehat{\Psi}}\rangle_{\partial\mathcal{T}_h}\\
&\qquad\quad + (\boldsymbol{\beta}\cdot\nabla\varepsilon_h^{y}, \delta^{\Psi})_{\mathcal{T}_h} - \langle \boldsymbol{\beta}\cdot\boldsymbol{n}(\varepsilon_h^{y} - \varepsilon_h^{\widehat{y}}), \delta^{\Psi}\rangle_{\partial\mathcal{T}_h}\\
&\qquad\quad + \langle (\tau_1 + h^{-1})(\varepsilon_h^{y} - \varepsilon_h^{\widehat{y}}), \delta^{\Psi} - \delta^{\widehat{\Psi}}\rangle_{\partial\mathcal{T}_h}.
\end{aligned}$$

Here, we used that $\langle \varepsilon_h^{\widehat{y}}, \boldsymbol{\Phi}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^{\partial}} = 0$, $\Psi = \varepsilon_h^{\widehat{y}} = 0$ on $\varepsilon_h^{\partial}$, and

$$\langle \boldsymbol{\beta}\cdot\boldsymbol{n}\varepsilon_h^{\widehat{y}}, \delta^{\widehat{\Psi}}\rangle_{\partial\mathcal{T}_h} = 0,$$

since $\varepsilon_h^{\widehat{y}}$ is single-valued on interior faces and $\varepsilon_h^{\widehat{y}} = 0$ on boundary faces. On the other hand, from equation (30),

$$\begin{aligned}
&\mathscr{B}(\varepsilon_h^{\boldsymbol{q}}, \varepsilon_h^{y}, \varepsilon_h^{\widehat{y}}; \boldsymbol{\Pi}_V \boldsymbol{\Phi}, -\Pi_W \Psi, -\mathcal{I}_h \Psi)\\
&\qquad = -\langle \delta^{\widehat{y}}, \boldsymbol{\Pi}_V\boldsymbol{\Phi}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h} - (\boldsymbol{\beta}\delta^{y}, \nabla\Pi_W\Psi)_{\mathcal{T}_h} - (\nabla\cdot\boldsymbol{\beta}\delta^{y}, \Pi_W\Psi)_{\mathcal{T}_h}\\
&\qquad\quad + \langle \widehat{\boldsymbol{\delta}}_1, \Pi_W\Psi - \mathcal{I}_h\Psi\rangle_{\partial\mathcal{T}_h}.
\end{aligned}$$

Comparing the two equations above, we have

$$\begin{aligned}
\|\varepsilon_h^{y}\|_{\mathcal{T}_h}^2 &= \langle \delta^{\widehat{y}}, \delta^{\boldsymbol{\Phi}}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h} - \langle \widehat{\boldsymbol{\delta}}_1, \delta^{\Psi} - \delta^{\widehat{\Psi}}\rangle_{\partial\mathcal{T}_h} - (\boldsymbol{\beta}\delta^{y}, \nabla\Pi_W\Psi)_{\mathcal{T}_h}\\
&\quad - (\nabla\cdot\boldsymbol{\beta}\delta^{y}, \Pi_W\Psi)_{\mathcal{T}_h} - \langle \varepsilon_h^{y} - \varepsilon_h^{\widehat{y}}, \delta^{\boldsymbol{\Phi}}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h} + \langle \varepsilon_h^{\boldsymbol{q}}\cdot\boldsymbol{n}, \delta^{\widehat{\Psi}}\rangle_{\partial\mathcal{T}_h}\\
&\quad - \langle (\tau_1 + h^{-1})(\varepsilon_h^{y} - \varepsilon_h^{\widehat{y}}), \delta^{\Psi} - \delta^{\widehat{\Psi}}\rangle_{\partial\mathcal{T}_h} + (\boldsymbol{\beta}\cdot(\nabla\varepsilon_h^{y}), \delta^{\Psi})_{\mathcal{T}_h}\\
&\quad - \langle \boldsymbol{\beta}\cdot\boldsymbol{n}(\varepsilon_h^{y} - \varepsilon_h^{\widehat{y}}), \delta^{\Psi}\rangle_{\partial\mathcal{T}_h}\\
&=: \sum_{i=1}^{9} T_i.
\end{aligned}$$

We estimate each term separately. For the first term,

$$T_1 \le \|\delta^{\widehat{y}}\|_{\partial\mathcal{T}_h}\|\delta^{\boldsymbol{\Phi}}\|_{\partial\mathcal{T}_h} \lesssim h^{\frac{1}{2}}\|\delta^{\widehat{y}}\|_{\partial\mathcal{T}_h}\|\boldsymbol{\Phi}\|_{1,\Omega} \lesssim h^{\frac{1}{2}}\|\delta^{\widehat{y}}\|_{\partial\mathcal{T}_h}\|\varepsilon_h^{y}\|_{\Omega}.$$

For the second term,

$$T_2 \lesssim h^{\frac{3}{2}}\|\widehat{\delta}_1\|_{\partial\mathcal{T}_h}\|\Psi\|_{2,\Omega} \lesssim h^{\frac{3}{2}}\|\widehat{\delta}_1\|_{\partial\mathcal{T}_h}\|\varepsilon_h^{y}\|_{\mathcal{T}_h}.$$

For the third term $T_3$,

$$\begin{aligned}
T_3 &\le \|\boldsymbol{\beta}\|_{0,\infty,\Omega}\|\delta^{y}\|_{\mathcal{T}_h}(\|\nabla\delta^{\Psi}\|_{\mathcal{T}_h} + \|\nabla\Psi\|_{\Omega})\\
&\lesssim \|\delta^{y}\|_{\mathcal{T}_h}(\|\Psi\|_{2,\Omega} + \|\Psi\|_{1,\Omega})\\
&\lesssim \|\delta^{y}\|_{\mathcal{T}_h}\|\varepsilon_h^{y}\|_{\mathcal{T}_h}.
\end{aligned}$$

For $T_4$,

$$T_4 \lesssim \|\boldsymbol{\beta}\|_{1,\infty,\Omega}\|\delta^y\|_{\mathcal{T}_h}\|\Pi_W\Psi\|_{\mathcal{T}_h} \lesssim \|\delta^y\|_{\mathcal{T}_h}\|\varepsilon_h^y\|_{\mathcal{T}_h}.$$

For $T_5$,

$$\begin{aligned} T_5 &\le \|\varepsilon_h^y - \varepsilon_h^{\widehat{y}}\|_{\partial\mathcal{T}_h}\|\delta^{\boldsymbol{\Phi}}\|_{\partial\mathcal{T}_h} \\ &\lesssim h^{\frac{1}{2}}\|\varepsilon_h^y - \varepsilon_h^{\widehat{y}}\|_{\partial\mathcal{T}_h}\|\boldsymbol{\Phi}\|_{1,\Omega} \\ &\lesssim h^{\frac{1}{2}}\|\varepsilon_h^y - \varepsilon_h^{\widehat{y}}\|_{\partial\mathcal{T}_h}\|\varepsilon_h^y\|_{\mathcal{T}_h}. \end{aligned}$$

For $T_6$, $T_7$, and $T_9$, following the same idea for $T_5$, we have

$$\begin{aligned} T_6 &\lesssim h\|\varepsilon_h^{\boldsymbol{q}}\|_{\mathcal{T}_h}\|\varepsilon_h^y\|_{\mathcal{T}_h}, \\ T_7 &\lesssim h^{\frac{1}{2}}\|\varepsilon_h^y - \varepsilon_h^{\widehat{y}}\|_{\partial\mathcal{T}_h}\|\varepsilon_h^y\|_{\mathcal{T}_h}, \\ T_9 &\lesssim \|\boldsymbol{\beta}\|_{0,\infty,\Omega}h^{\frac{1}{2}}\|\varepsilon_h^y - \varepsilon_h^{\widehat{y}}\|_{\partial\mathcal{T}_h}\|\varepsilon_h^y\|_{\mathcal{T}_h}. \end{aligned}$$

And by Lemma 4, we have

$$\begin{aligned} T_8 &\lesssim \|\boldsymbol{\beta}\|_{0,\infty,\Omega}h\|\nabla\varepsilon_h^y\|_{\mathcal{T}_h}\|\Psi\|_1 \\ &\lesssim h(\|\varepsilon_h^{\boldsymbol{q}}\|_{\mathcal{T}_h} + h^{-\frac{1}{2}}\|\varepsilon_h^y - \varepsilon_h^{\widehat{y}}\|_{\mathcal{T}_h})\|\varepsilon_h^y\|_{\mathcal{T}_h}. \end{aligned}$$

Therefore, summing the estimates and using the bounds (24) and Lemma 5 gives the result. □

The triangle inequality yields optimal convergence rates for $\|\boldsymbol{q} - \boldsymbol{q}_h(u)\|_{\mathcal{T}_h}$ and $\|y - y_h(u)\|_{\mathcal{T}_h}$:

**Lemma 7.** We have

$$\|\boldsymbol{q} - \boldsymbol{q}_h(u)\|_{\mathcal{T}_h} \le \|\delta^{\boldsymbol{q}}\|_{\mathcal{T}_h} + \|\varepsilon_h^{\boldsymbol{q}}\|_{\mathcal{T}_h} \lesssim h^{k+1}(|\boldsymbol{q}|_{k+1} + |y|_{k+2}), \tag{37a}$$

$$\|y - y_h(u)\|_{\mathcal{T}_h} \le \|\delta^y\|_{\mathcal{T}_h} + \|\varepsilon_h^y\|_{\mathcal{T}_h} \lesssim h^{k+2}(|\boldsymbol{q}|_{k+1} + |y|_{k+2}). \tag{37b}$$

**3.3.4. Step 4: The error equation for part 2 of the auxiliary problem (28b).** Next, we bound the error between the solution of the dual convection diffusion equation (8c)-(8d) for $z$ and the auxiliary HDG equation (28b).

First, we define

$$\begin{aligned} &\delta^{\boldsymbol{p}} = \boldsymbol{p} - \boldsymbol{\Pi}_V\boldsymbol{p}, && \varepsilon_h^{\boldsymbol{p}} = \boldsymbol{\Pi}_V\boldsymbol{p} - \boldsymbol{p}_h(u), \\ &\delta^z = z - \Pi_W z, && \varepsilon_h^z = \Pi_W z - z_h(u), \\ &\delta^{\widehat{z}} = z - \mathcal{I}_h z, && \varepsilon_h^{\widehat{z}} = \mathcal{I}_h z - \widehat{z}_h(u), \\ &\widehat{\boldsymbol{\delta}}_2 = \delta^{\boldsymbol{p}}\cdot\boldsymbol{n} - \boldsymbol{\beta}\cdot\boldsymbol{n}\delta^{\widehat{z}} + (\tau_2 + h^{-1})(\delta^z - \delta^{\widehat{z}}), \end{aligned} \tag{38}$$

where $\widehat{z}_h(u) = \widehat{z}_h^o(u)$ on $\varepsilon_h^o$ and $\widehat{z}_h(u) = 0$ on $\varepsilon_h^\partial$. This gives $\varepsilon_h^{\widehat{z}} = 0$ on $\varepsilon_h^\partial$.

Following the same idea with Lemma 3, we have the following error equation:

**Lemma 8.** We have

$$\begin{aligned} &\mathscr{B}_2(\varepsilon_h^{\boldsymbol{p}}, \varepsilon_h^z, \varepsilon_h^{\widehat{z}}; \boldsymbol{r}_2, w_2, \mu_2) \\ &\qquad = -\langle\delta^{\widehat{z}}, \boldsymbol{r}_2\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h} - (\boldsymbol{\beta}\delta^z, \nabla w_2)_{\mathcal{T}_h} \\ &\qquad\quad - \langle\widehat{\boldsymbol{\delta}}_2, w_2\rangle_{\partial\mathcal{T}_h} + \langle\widehat{\boldsymbol{\delta}}_2, \mu_2\rangle_{\partial\mathcal{T}_h\backslash\varepsilon_h^\partial} + (y_h(u) - y, w_2)_{\mathcal{T}_h}. \end{aligned} \tag{39}$$

**3.3.5. Step 5: Estimates for $\varepsilon_h^p$ and $\varepsilon_h^z$ by an energy and duality argument.** First, it is easy to see that Lemma 4 still holds for $\varepsilon_h^z$, $\varepsilon_h^{\widehat{z}}$, and $\varepsilon_h^{\boldsymbol{p}}$.

**Lemma 9.** We have

$$\|\nabla\varepsilon_h^z\|_{\mathcal{T}_h} \le C(\|\varepsilon_h^{\boldsymbol{q}}\|_{\mathcal{T}_h} + h^{-\frac{1}{2}}\|\varepsilon_h^z - \varepsilon_h^{\widehat{z}}\|_{\partial\mathcal{T}_h}). \tag{40}$$

Also, to estimate $\varepsilon_h^{\boldsymbol{p}}$ we need the following discrete Poincaré inequality that is very similar to a result from [30]. The proof is essentially the same, and is omitted.

**Lemma 10.** We have

$$\|\varepsilon_h^z\|_{\mathcal{T}_h} \le C(\|\nabla\varepsilon_h^z\|_{\mathcal{T}_h} + h^{-\frac{1}{2}}\|\varepsilon_h^z - \varepsilon_h^{\widehat{z}}\|_{\partial\mathcal{T}_h}). \tag{41}$$

**Lemma 11.** We have

$$\|\varepsilon_h^{\boldsymbol{p}}\|_{\mathcal{T}_h} + h^{-\frac{1}{2}}\|\varepsilon_h^z - \varepsilon_h^{\widehat{z}}\|_{\partial\mathcal{T}_h} \lesssim h^{k+1}(|\boldsymbol{q}|_{k+1} + |y|_{k+2} + |\boldsymbol{p}|_{k+1} + |z|_{k+2}), \tag{42}$$

$$\|\varepsilon_h^z\|_{\mathcal{T}_h} \lesssim h^{k+1}(|\boldsymbol{q}|_{k+1} + |y|_{k+2} + |\boldsymbol{p}|_{k+1} + |z|_{k+2}). \tag{43}$$

*Proof.* Since $\varepsilon_h^{\widehat{z}} = 0$ on $\varepsilon_h^\partial$, the energy identity for $\mathscr{B}_2$ in Lemma 1 gives

$$\begin{aligned}
&\mathscr{B}_2(\varepsilon_h^{\boldsymbol{p}}, \varepsilon_h^z, \varepsilon_h^{\widehat{z}}, \varepsilon_h^{\boldsymbol{p}}, \varepsilon_h^z, \varepsilon_h^{\widehat{z}})\\
&\quad = \|\varepsilon_h^{\boldsymbol{p}}\|_{\mathcal{T}_h}^2 + \|(h^{-1} + \tau_2 + \frac{1}{2}\boldsymbol{\beta}\cdot\boldsymbol{n})^{\frac{1}{2}}(\varepsilon_h^z - \varepsilon_h^{\widehat{z}})\|_{\partial\mathcal{T}_h}^2 + \frac{1}{2}\|(-\nabla\cdot\boldsymbol{\beta})^{\frac{1}{2}}\varepsilon_h^z\|_{\mathcal{T}_h}^2.
\end{aligned}$$

Take $(\boldsymbol{r}_2, w_2, \mu_2) = (\varepsilon_h^{\boldsymbol{p}}, \varepsilon_h^z, \varepsilon_h^{\widehat{z}})$ in the error equation (39) to obtain

$$\begin{aligned}
\|\varepsilon_h^{\boldsymbol{p}}\|_{\mathcal{T}_h}^2 &+ \|(h^{-1} + \tau_2 + \frac{1}{2}\boldsymbol{\beta}\cdot\boldsymbol{n})^{\frac{1}{2}}(\varepsilon_h^z - \varepsilon_h^{\widehat{z}})\|_{\partial\mathcal{T}_h}^2 + \frac{1}{2}\|(-\nabla\cdot\boldsymbol{\beta})^{\frac{1}{2}}\varepsilon_h^z\|_{\mathcal{T}_h}^2\\
&= -\langle\delta^{\widehat{z}}, \boldsymbol{\varepsilon}_h^{\boldsymbol{p}}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h} - (\boldsymbol{\beta}\delta^z, \nabla\varepsilon_h^z)_{\mathcal{T}_h}\\
&\quad - \langle\widehat{\boldsymbol{\delta}}_2, \varepsilon_h^z - \varepsilon_h^{\widehat{z}}\rangle_{\partial\mathcal{T}_h} + (y_h(u) - y, \varepsilon_h^z)_{\mathcal{T}_h}\\
&=: T_1 + T_2 + T_3 + T_4.
\end{aligned}$$

By the same argument as in the proof of Lemma 5, apply (40) and (41) to get

$$\begin{aligned}
T_1 &\lesssim h^{-\frac{1}{2}}\|\delta^{\widehat{z}}\|_{\partial\mathcal{T}_h}\|\varepsilon_h^{\boldsymbol{p}}\|_{\mathcal{T}_h},\\
T_2 &\lesssim \|\boldsymbol{\beta}\|_{0,\infty,\Omega}\|\delta^z\|_{\mathcal{T}_h}\|\nabla\varepsilon_h^z\|_{\mathcal{T}_h}\\
&\lesssim \|\boldsymbol{\beta}\|_{0,\infty,\Omega}\|\delta^z\|_{\mathcal{T}_h}(\|\varepsilon_h^{\boldsymbol{p}}\|_{\mathcal{T}_h} + h^{-\frac{1}{2}}\|\varepsilon_h^z - \varepsilon_h^{\widehat{z}}\|_{\partial\mathcal{T}_h}),\\
T_3 &\lesssim h^{\frac{1}{2}}\|\widehat{\boldsymbol{\delta}}_2\|_{\partial\mathcal{T}_h}h^{-\frac{1}{2}}\|\varepsilon_h^z - \varepsilon_h^{\widehat{z}}\|_{\partial\mathcal{T}_h},\\
T_4 &\lesssim \|y - y_h(u)\|_{\mathcal{T}_h}\|\varepsilon_h^z\|_{\mathcal{T}_h}\\
&\lesssim \|y - y_h(u)\|_{\mathcal{T}_h}(\|\varepsilon_h^{\boldsymbol{p}}\|_{\mathcal{T}_h} + h^{-\frac{1}{2}}\|\varepsilon_h^z - \varepsilon_h^{\widehat{z}}\|_{\partial\mathcal{T}_h}).
\end{aligned}$$

Finally, applying (24) and Lemma 7 yields (42). Together with (42) and (41), we can obtain (43). □

**3.3.6. Step 6: Estimate for $\varepsilon_h^z$ by a duality argument.** Next, we introduce the dual problem for any given $\Theta$ in $L^2(\Omega)$:

$$\begin{aligned}
\boldsymbol{\Phi} + \nabla\Psi &= 0 &&\text{in } \Omega,\\
\nabla\cdot\boldsymbol{\Phi} - \boldsymbol{\beta}\cdot\nabla\Psi &= \Theta &&\text{in } \Omega,\\
\Psi &= 0 &&\text{on } \partial\Omega.
\end{aligned} \tag{44}$$

Since the domain $\Omega$ is convex, we have the following regularity estimate

$$\|\boldsymbol{\Phi}\|_{1,\Omega} + \|\Psi\|_{2,\Omega} \le C_{\text{reg}}\|\Theta\|_{\Omega}. \tag{45}$$

**Lemma 12.** We have

$$\|\varepsilon_h^z\|_{\mathcal{T}_h} \lesssim h^{k+2}(|\boldsymbol{q}|_{k+1} + |y|_{k+2} + |\boldsymbol{p}|_{k+1} + |z|_{k+2}). \tag{46a}$$

*Proof.* Consider the dual problem (44), and let $\Theta = \varepsilon_h^z$. Take $(\boldsymbol{r}_2, w_2, \mu_2) = (\boldsymbol{\Pi}_V \boldsymbol{\Phi}, -\Pi_W \Psi, -I_h\Psi)$ in (39) in Lemma 8. Since $\Psi = 0$ on $\varepsilon_h^\partial$ we have

$$\begin{aligned}
&\mathscr{B}_2(\varepsilon_h^{\boldsymbol{p}}, \varepsilon_h^z, \varepsilon_h^{\widehat{z}}; \boldsymbol{\Pi}_V\boldsymbol{\Phi}, -\Pi_W\Psi, -I_h\Psi)\\
&\quad = (\varepsilon_h^{\boldsymbol{p}}, \boldsymbol{\Pi}_V\boldsymbol{\Phi})_{\mathcal{T}_h} - (\varepsilon_h^z, \nabla\cdot\boldsymbol{\Pi}_V\boldsymbol{\Phi})_{\mathcal{T}_h} + \langle \varepsilon_h^{\widehat{z}}, \boldsymbol{\Pi}_V\boldsymbol{\Phi}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h}\\
&\qquad + (\varepsilon_h^{\boldsymbol{p}} - \boldsymbol{\beta}\varepsilon_h^z, \nabla\Pi_W\Psi)_{\mathcal{T}_h} - \langle \varepsilon_h^{\boldsymbol{p}}\cdot\boldsymbol{n} - \boldsymbol{\beta}\cdot\boldsymbol{n}\varepsilon_h^{\widehat{z}} + \tau_2(\varepsilon_h^z - \varepsilon_h^{\widehat{z}}), \Pi_W\Psi - \mathcal{I}_h\Psi\rangle_{\partial\mathcal{T}_h}.
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
-(\varepsilon_h^z, \nabla\cdot\boldsymbol{\Pi}_V\boldsymbol{\Phi})_{\partial\mathcal{T}_h} &= (\nabla\varepsilon_h^z, \boldsymbol{\Phi})_{\mathcal{T}_h} - \langle\varepsilon_h^z, \boldsymbol{\Pi}_V\boldsymbol{\Phi}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h}\\
&= -(\varepsilon_h^z, \nabla\cdot\boldsymbol{\Phi})_{\mathcal{T}_h} + \langle\varepsilon_h^z, \delta^{\boldsymbol{\Phi}}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h},
\end{aligned}$$

$$\begin{aligned}
(\varepsilon_h^{\boldsymbol{p}}, \nabla\Pi_W\Psi)_{\mathcal{T}_h} &= -(\nabla\cdot\varepsilon_h^{\boldsymbol{p}}, \Psi)_{\mathcal{T}_h} + \langle\varepsilon_h^{\boldsymbol{p}}\cdot\boldsymbol{n}, \Pi_W\Psi\rangle_{\partial\mathcal{T}_h}\\
&= (\varepsilon_h^{\boldsymbol{p}}, \nabla\Psi)_{\mathcal{T}_h} - \langle\varepsilon_h^{\boldsymbol{p}}\cdot\boldsymbol{n}, \delta^\Psi\rangle_{\partial\mathcal{T}_h},
\end{aligned}$$

$$\begin{aligned}
-(\boldsymbol{\beta}\varepsilon_h^z, \nabla\Pi_W\Psi)_{\mathcal{T}_h} &= -(\boldsymbol{\beta}\varepsilon_h^z, \nabla\delta^\Psi)_{\mathcal{T}_h} + (\boldsymbol{\beta}\varepsilon_h^z, \nabla\Psi)_{\mathcal{T}_h}\\
&= -\langle\boldsymbol{\beta}\cdot\boldsymbol{n}\varepsilon_h^z, \delta^\Psi\rangle_{\partial\mathcal{T}_h} + (\nabla\cdot\boldsymbol{\beta}\varepsilon_h^z, \delta^\Psi)_{\mathcal{T}_h}\\
&\quad + (\boldsymbol{\beta}\cdot\nabla\varepsilon_h^z, \delta^\Psi)_{\mathcal{T}_h} + (\boldsymbol{\beta}\varepsilon_h^z, \nabla\Psi)_{\mathcal{T}_h}.
\end{aligned}$$

Then we have

$$\begin{aligned}
&\mathscr{B}_2(\varepsilon_h^{\boldsymbol{p}}, \varepsilon_h^z, \varepsilon_h^{\widehat{z}}; \boldsymbol{\Pi}_V\boldsymbol{\Phi}, -\Pi_W\Psi, -I_h\Psi)\\
&\quad = (\varepsilon_h^{\boldsymbol{p}}, \boldsymbol{\Phi})_{\mathcal{T}_h} - (\varepsilon_h^z, \nabla\cdot\boldsymbol{\Phi})_{\mathcal{T}_h} + \langle\varepsilon_h^z, \delta^{\boldsymbol{\Phi}}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h} + \langle\varepsilon_h^{\widehat{z}}, \boldsymbol{\Pi}_V\boldsymbol{\Phi}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h}\\
&\qquad + (\varepsilon_h^{\boldsymbol{p}}, \nabla\Psi)_{\mathcal{T}_h} - \langle\varepsilon_h^{\boldsymbol{p}}\cdot\boldsymbol{n}, \delta^{\widehat{\Psi}}\rangle_{\partial\mathcal{T}_h} - \langle\boldsymbol{\beta}\cdot\boldsymbol{n}\varepsilon_h^z, \delta^\Psi\rangle_{\partial\mathcal{T}_h}\\
&\qquad + (\nabla\cdot\boldsymbol{\beta}\varepsilon_h^z, \delta^\Psi)_{\mathcal{T}_h} + (\boldsymbol{\beta}\cdot\nabla\varepsilon_h^z, \delta^\Psi)_{\mathcal{T}_h} + (\varepsilon_h^z, \boldsymbol{\beta}\cdot\nabla\Psi)_{\mathcal{T}_h}\\
&\qquad + \langle\boldsymbol{\beta}\cdot\boldsymbol{n}\varepsilon_h^{\widehat{z}}, \delta^\Psi\rangle_{\partial\mathcal{T}_h} + \langle\tau_2(\varepsilon_h^z - \varepsilon_h^{\widehat{z}}), \delta^\Psi - \delta^{\widehat{\Psi}}\rangle_{\partial\mathcal{T}_h}\\
&\quad = (\varepsilon_h^z, \varepsilon_h^z)_{\mathcal{T}_h} + \langle\varepsilon_h^z - \varepsilon_h^{\widehat{z}}, \delta^{\boldsymbol{\Phi}}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h} - \langle\varepsilon_h^{\boldsymbol{p}}\cdot\boldsymbol{n}, \delta^{\widehat{\Psi}}\rangle_{\partial\mathcal{T}_h} + (\nabla\cdot\boldsymbol{\beta}\varepsilon_h^z, \delta^\Psi)_{\mathcal{T}_h}\\
&\qquad + (\boldsymbol{\beta}\cdot\nabla\varepsilon_h^z, \delta^\Psi)_{\mathcal{T}_h} - \langle\boldsymbol{\beta}\cdot\boldsymbol{n}(\varepsilon_h^z - \varepsilon_h^{\widehat{z}}), \delta^\Psi\rangle_{\partial\mathcal{T}_h}\\
&\qquad + \langle(\tau_2 + h^{-1})(\varepsilon_h^z - \varepsilon_h^{\widehat{z}}), \delta^\Psi - \delta^{\widehat{\Psi}}\rangle_{\partial\mathcal{T}_h}.
\end{aligned}$$

Here, we used $\langle\varepsilon_h^{\widehat{z}}, \boldsymbol{\Phi}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h} = 0$, which holds since $\varepsilon_h^{\widehat{z}}$ is single-valued function on interior edges and $\varepsilon_h^{\widehat{z}} = 0$ on $\varepsilon_h^\partial$. We also used $\langle\boldsymbol{\beta}\cdot\boldsymbol{n}\varepsilon_h^{\widehat{z}}, \delta^{\widehat{\Psi}}\rangle_{\partial\mathcal{T}_h} = 0$, which is derived similarly.

On the other hand, by Lemma 8

$$\begin{aligned}
&\mathscr{B}_2(\varepsilon_h^{\boldsymbol{p}}, \varepsilon_h^z, \varepsilon_h^{\widehat{z}}; \boldsymbol{\Pi}_V\boldsymbol{\Phi}, -\Pi_V\Psi, -\mathcal{I}_h\Psi)\\
&\quad = -\langle\delta^{\widehat{z}}, \boldsymbol{\Pi}_V\boldsymbol{\Phi}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h} + (\boldsymbol{\beta}\delta^z, \nabla\Pi_V\Psi)_{\mathcal{T}_h}\\
&\qquad + \langle\widehat{\boldsymbol{\delta}}_2, \Pi_V\Psi - \mathcal{I}_h\Psi\rangle_{\partial\mathcal{T}_h} - (y_h(u) - y, \Pi_V\Psi)_{\mathcal{T}_h}.
\end{aligned} \tag{47}$$

Comparing the above two equalities gives

$$\begin{aligned}
\|\varepsilon_h^z\|_{\mathcal{T}_h}^2 &= -\langle \varepsilon_h^z - \varepsilon_h^{\widehat{z}}, \delta^{\boldsymbol{\Phi}}\cdot\boldsymbol{n} + (\tau_2 + h^{-1})(\delta^{\Psi} - \delta^{\widehat{\Psi}}) - \boldsymbol{\beta}\cdot\boldsymbol{n}\delta^{\Psi}\rangle_{\partial\mathcal{T}_h}\\
&\quad + \langle \varepsilon_h^{\boldsymbol{p}}\cdot\boldsymbol{n}, \delta^{\widehat{\Psi}}\rangle_{\partial\mathcal{T}_h} + (\boldsymbol{\beta}\cdot\nabla\varepsilon_h^z, \delta^{\Psi})_{\mathcal{T}_h} + (\nabla\cdot\boldsymbol{\beta}\varepsilon_h^z, \delta^{\Psi})_{\mathcal{T}_h}\\
&\quad - \langle \delta^{\widehat{z}}, \delta^{\boldsymbol{\Phi}}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h} + (\boldsymbol{\beta}\delta^z, \nabla\Pi_V\Psi)_{\mathcal{T}_h}\\
&\quad + \langle \widehat{\boldsymbol{\delta}}_2, \Pi_V\Psi - \mathcal{I}_h\Psi\rangle_{\partial\mathcal{T}_h} - (y_h(u) - y, \Pi_V\Psi)_{\mathcal{T}_h}\\
&=: \sum_{i=1}^{8} R_i.
\end{aligned}$$

For the terms $R_1$-$R_4$, Lemma 11 gives

$$\begin{aligned}
R_1 &= -\langle \varepsilon_h^z - \varepsilon_h^{\widehat{z}}, \delta^{\boldsymbol{\Phi}}\cdot\boldsymbol{n} - \boldsymbol{\beta}\cdot\boldsymbol{n}\delta^{\Psi} + (\tau_2 + h^{-1})(\delta^{\Psi} - \delta^{\widehat{\Psi}})\rangle_{\partial\mathcal{T}_h}\\
&\lesssim h^{\frac{1}{2}}\|(\tau_2 + h^{-1} + \boldsymbol{\beta}\cdot\boldsymbol{n})^{\frac{1}{2}}(\varepsilon_h^z - \varepsilon_h^{\widehat{z}})\|_{\partial\mathcal{T}_h}(\|\boldsymbol{\Phi}\|_{1,\Omega} + \|\Psi\|_{1,\Omega})\\
&\lesssim h^{\frac{1}{2}}\|(\tau_2 + h^{-1} + \boldsymbol{\beta}\cdot\boldsymbol{n})^{\frac{1}{2}}(\varepsilon_h^z - \varepsilon_h^{\widehat{z}})\|_{\partial\mathcal{T}_h}\|\varepsilon_h^z\|_{\mathcal{T}_h},\\
R_2 &\lesssim h^{\frac{3}{2}}\|\varepsilon_h^{\boldsymbol{p}}\|_{\partial\mathcal{T}_h}\|\Psi\|_{2,\Omega} \lesssim h^{\frac{3}{2}}\|\varepsilon_h^{\boldsymbol{p}}\|_{\partial\mathcal{T}_h}\|\varepsilon_h^z\|_{\mathcal{T}_h},\\
R_3 &\lesssim \|\boldsymbol{\beta}\|_{0,\infty,\Omega}h\|\nabla\varepsilon_h^z\|_{\mathcal{T}_h}\|\Psi\|_{1,\Omega},\\
R_4 &\lesssim h\|(-\nabla\cdot\boldsymbol{\beta})^{\frac{1}{2}}\varepsilon_h^z\|_{\mathcal{T}_h}\|\Psi\|_{1,\Omega} \lesssim h\|(-\nabla\cdot\boldsymbol{\beta})^{\frac{1}{2}}\varepsilon_h^z\|_{\mathcal{T}_h}\|\varepsilon_h^z\|_{\mathcal{T}_h}.
\end{aligned}$$

For $R_5$, we have

$$R_5 \lesssim h^{\frac{1}{2}}\|\delta^{\widehat{z}}\|_{\partial\mathcal{T}_h}\|\varepsilon_h^z\|_{\mathcal{T}_h}.$$

For the terms $R_6$ and $R_8$, we use the triangle inequality, the regularity estimate (34), and the assumption $h \leq 1$ to give

$$\begin{aligned}
R_6 &\lesssim \|\boldsymbol{\beta}\|_{0,\infty,\Omega}\|\delta^z\|_{\mathcal{T}_h}(\|\nabla\delta^{\Psi}\|_{\mathcal{T}_h} + \|\Psi\|_{\mathcal{T}_h}) \lesssim \|\boldsymbol{\beta}\|_{0,\infty,\Omega}\|\delta^z\|_{\mathcal{T}_h}\|\varepsilon_h^z\|_{\mathcal{T}_h},\\
R_8 &\lesssim \|y_h(u) - y\|_{\mathcal{T}_h}\|\varepsilon_h^z\|_{\mathcal{T}_h}.
\end{aligned}$$

For the term $R_7$,

$$\begin{aligned}
R_7 &\lesssim h^{\frac{3}{2}}\|\delta^{\boldsymbol{p}}\cdot\boldsymbol{n} + (\tau_1 + h^{-1})(\delta^z - \delta^{\widehat{z}})\|_{\partial\mathcal{T}_h}\|\Psi\|_{2,\Omega}\\
&\lesssim h^{\frac{3}{2}}(\|\delta^{\boldsymbol{p}}\|_{\partial\mathcal{T}_h} + \|\delta^z\|_{\mathcal{T}_h} + \|\delta^{\widehat{z}}\|_{\partial\mathcal{T}_h})\|\varepsilon_h^z\|_{\mathcal{T}_h}.
\end{aligned}$$

Summing $R_1$ to $R_8$, together with (24), (40), (42), and (43) gives

$$\|\varepsilon_h^z\|_{\mathcal{T}_h} \lesssim h^{k+2}(|\boldsymbol{q}|_{k+1} + |y|_{k+2} + |\boldsymbol{p}|_{k+1} + |z|_{k+2}).$$

□

The triangle inequality gives optimal convergence rates for $\|\boldsymbol{p} - \boldsymbol{p}_h(u)\|_{\mathcal{T}_h}$ and $\|z - z_h(u)\|_{\mathcal{T}_h}$:

**Lemma 13.**

$$\begin{aligned}
\|\boldsymbol{p} - \boldsymbol{p}_h(u)\|_{\mathcal{T}_h} &\leq \|\delta^{\boldsymbol{p}}\|_{\mathcal{T}_h} + \|\varepsilon_h^{\boldsymbol{p}}\|_{\mathcal{T}_h}\\
&\lesssim h^{k+1}(|\boldsymbol{q}|_{k+1} + |y|_{k+2} + |\boldsymbol{p}|_{k+1} + |z|_{k+2}), \qquad (48a)\\
\|z - z_h(u)\|_{\mathcal{T}_h} &\leq \|\delta^z\|_{\mathcal{T}_h} + \|\varepsilon_h^z\|_{\mathcal{T}_h}\\
&\lesssim h^{k+2}(|\boldsymbol{q}|_{k+1} + |y|_{k+2} + |\boldsymbol{p}|_{k+1} + |z|_{k+2}). \qquad (48b)
\end{aligned}$$

**3.3.7. Step 7: Estimates for $\|u-u_h\|_{\mathcal{T}_h}$, $\|y-y_h\|_{\mathcal{T}_h}$, and $\|z-z_h\|_{\mathcal{T}_h}$.** Next, we bound the error between the solutions of the auxiliary problem and the EDG discretization of the optimality system (27). We use these error bounds and the error bounds in Lemma 7 and Lemma 13 to obtain the main result.

The proofs in Steps 7 and 8 are similar to proofs in our earlier work [18]; we include the proofs here to make the final steps self-contained.

For the remaining steps, we denote

$$\begin{aligned}\zeta_{\boldsymbol{q}} &= \boldsymbol{q}_h(u)-\boldsymbol{q}_h, \quad \zeta_y = y_h(u)-y_h, \quad \zeta_{\widehat{y}} = \widehat{y}_h^o(u)-\widehat{y}_h^o,\\ \zeta_{\boldsymbol{p}} &= \boldsymbol{p}_h(u)-\boldsymbol{p}_h, \quad \zeta_z = z_h(u)-z_h, \quad \zeta_{\widehat{z}} = \widehat{z}_h^o(u)-\widehat{z}_h^o.\end{aligned}$$

Subtracting the auxiliary problem and the EDG problem gives the following error equations

$$\mathscr{B}_1(\zeta_{\boldsymbol{q}},\zeta_y,\zeta_{\widehat{y}};\boldsymbol{r}_1,w_1,\mu_1) = (u-u_h,w_1)_{\mathcal{T}_h}, \tag{49a}$$

$$\mathscr{B}_2(\zeta_{\boldsymbol{p}},\zeta_z,\zeta_{\widehat{z}};\boldsymbol{r}_2,w_2,\mu_2) = -(\zeta_y,w_2)_{\mathcal{T}_h}. \tag{49b}$$

**Lemma 14.** We have

$$\begin{aligned}&\gamma\|u-u_h\|_{\mathcal{T}_h}^2 + \|y_h(u)-y_h\|_{\mathcal{T}_h}^2\\ &\qquad = (z_h+\gamma u_h, u-u_h)_{\mathcal{T}_h} - (z_h(u)+\gamma u, u-u_h)_{\mathcal{T}_h}.\end{aligned} \tag{50}$$

*Proof.* First, we have

$$\begin{aligned}&(z_h+\gamma u_h, u-u_h)_{\mathcal{T}_h} - (z_h(u)+\gamma u, u-u_h)_{\mathcal{T}_h}\\ &\qquad = -(\zeta_z, u-u_h)_{\mathcal{T}_h} + \gamma\|u-u_h\|_{\mathcal{T}_h}^2.\end{aligned}$$

Next, Lemma 2 gives

$$\mathscr{B}_1(\zeta_{\boldsymbol{q}},\zeta_y,\zeta_{\widehat{y}};\zeta_{\boldsymbol{p}},-\zeta_z,-\zeta_{\widehat{z}}) + \mathscr{B}_2(\zeta_{\boldsymbol{p}},\zeta_z,\zeta_{\widehat{z}};-\zeta_{\boldsymbol{q}},\zeta_y,\zeta_{\widehat{y}}) = 0.$$

On the other hand, using the definition of $\mathscr{B}_1$ and $\mathscr{B}_2$ gives

$$\begin{aligned}&\mathscr{B}_1(\zeta_{\boldsymbol{q}},\zeta_y,\zeta_{\widehat{y}};\zeta_{\boldsymbol{p}},-\zeta_z,-\zeta_{\widehat{z}}) + \mathscr{B}_2(\zeta_{\boldsymbol{p}},\zeta_z,\zeta_{\widehat{z}};-\zeta_{\boldsymbol{q}},\zeta_y,\zeta_{\widehat{y}})\\ &\qquad = -(u-u_h,\zeta_z)_{\mathcal{T}_h} - \|\zeta_y\|_{\mathcal{T}_h}^2.\end{aligned}$$

Comparing the above two equalities gives

$$-(u-u_h,\zeta_z)_{\mathcal{T}_h} = \|\zeta_y\|_{\mathcal{T}_h}^2.$$

This completes the proof. □

**Theorem 2.** We have

$$\|u-u_h\|_{\mathcal{T}_h} \lesssim h^{k+2}(|\boldsymbol{q}|_{k+1}+|y|_{k+2}+|\boldsymbol{p}|_{k+1}+|z|_{k+2}), \tag{51a}$$

$$\|y-y_h\|_{\mathcal{T}_h} \lesssim h^{k+2}(|\boldsymbol{q}|_{k+1}+|y|_{k+2}+|\boldsymbol{p}|_{k+1}+|z|_{k+2}), \tag{51b}$$

$$\|z-z_h\|_{\mathcal{T}_h} \lesssim h^{k+2}(|\boldsymbol{q}|_{k+1}+|y|_{k+2}+|\boldsymbol{p}|_{k+1}+|z|_{k+2}). \tag{51c}$$

*Proof.* Recalling the continuous and discretized optimality conditions (8e) and (27c) gives

$$\begin{aligned}\gamma\|u-u_h\|_{\mathcal{T}_h}^2 &+ \|\zeta_y\|_{\mathcal{T}_h}^2\\ &= (z_h+\gamma u_h, u-u_h)_{\mathcal{T}_h} - (z_h(u)+\gamma u, u-u_h)_{\mathcal{T}_h}\\ &= -(z_h(u)-z, u-u_h)_{\mathcal{T}_h}\\ &\le \|z_h(u)-z\|_{\mathcal{T}_h}\|u-u_h\|_{\mathcal{T}_h}\\ &\le \frac{1}{2\gamma}\|z_h(u)-z\|_{\mathcal{T}_h}^2 + \frac{\gamma}{2}\|u-u_h\|_{\mathcal{T}_h}^2.\end{aligned}$$

By Lemma 13, we have

$$\|u-u_h\|_{\mathcal{T}_h} + \|\zeta_y\|_{\mathcal{T}_h} \lesssim h^{k+2}(|\boldsymbol{q}|_{k+1}+|y|_{k+2}+|\boldsymbol{p}|_{k+1}+|z|_{k+2}). \tag{52}$$

Then, by the triangle inequality and Lemma 7 we obtain

$$\|y-y_h\|_{\mathcal{T}_h} \lesssim h^{k+2}(|\boldsymbol{q}|_{k+1}+|y|_{k+2}+|\boldsymbol{p}|_{k+1}+|z|_{k+2}).$$

Finally, since $z=\gamma u$ and $z_h=\gamma u_h$ we have

$$\|z-z_h\|_{\mathcal{T}_h} \lesssim h^{k+2}(|\boldsymbol{q}|_{k+1}+|y|_{k+2}+|\boldsymbol{p}|_{k+1}+|z|_{k+2}).$$

□

### 3.3.8. Step 8: Estimate for $\|q-q_h\|_{\mathcal{T}_h}$ and $\|p-p_h\|_{\mathcal{T}_h}$.

**Lemma 15.** We have

$$\|\zeta_{\boldsymbol{q}}\|_{\mathcal{T}_h} \lesssim h^{k+2}(|\boldsymbol{q}|_{k+1}+|y|_{k+2}+|\boldsymbol{p}|_{k+1}+|z|_{k+2}), \tag{53a}$$

$$\|\zeta_{\boldsymbol{p}}\|_{\mathcal{T}_h} \lesssim h^{k+2}(|\boldsymbol{q}|_{k+1}+|y|_{k+2}+|\boldsymbol{p}|_{k+1}+|z|_{k+2}). \tag{53b}$$

*Proof.* By Lemma 1, the error equation (49a), and the estimate (52) we have

$$\begin{aligned}\|\zeta_{\boldsymbol{q}}\|_{\mathcal{T}_h}^2 &\lesssim \mathscr{B}_1(\zeta_{\boldsymbol{q}},\zeta_y,\zeta_{\widehat{y}};\zeta_{\boldsymbol{q}},\zeta_y,\zeta_{\widehat{y}})\\ &= (u-u_h,\zeta_y)_{\mathcal{T}_h}\\ &\le \|u-u_h\|_{\mathcal{T}_h}\|\zeta_y\|_{\mathcal{T}_h}\\ &\lesssim h^{2k+4}(|\boldsymbol{q}|_{k+1}+|y|_{k+2}+|\boldsymbol{p}|_{k+1}+|z|_{k+2})^2.\end{aligned}$$

Similarly, by Lemma 1, the error equation (49b), Lemma 13, and Theorem 2 we have

$$\begin{aligned}\|\zeta_{\boldsymbol{p}}\|_{\mathcal{T}_h}^2 &\lesssim \mathscr{B}_2(\zeta_{\boldsymbol{p}},\zeta_z,\zeta_{\widehat{z}};\zeta_{\boldsymbol{p}},\zeta_z,\zeta_{\widehat{z}})\\ &= -(\zeta_y,\zeta_z)_{\mathcal{T}_h}\\ &\le \|\zeta_y\|_{\mathcal{T}_h}\|\zeta_z\|_{\mathcal{T}_h}\\ &\le \|\zeta_y\|_{\mathcal{T}_h}(\|z_h(u)-z\|_{\mathcal{T}_h}+\|z-z_h\|_{\mathcal{T}_h})\\ &\lesssim h^{2k+4}(|\boldsymbol{q}|_{k+1}+|y|_{k+2}+|\boldsymbol{p}|_{k+1}+|z|_{k+2})^2.\end{aligned}$$

□

The above lemma along with the triangle inequality, Lemma 7, and Lemma 13 complete the proof of the main result:

Table 1. Example 1: Errors for the state $y$, adjoint state $z$, and the fluxes $\boldsymbol{q}$ and $\boldsymbol{p}$ when $k = 0$ with the OD approach.

| $h/\sqrt{2}$ | 1/8 | 1/16 | 1/32 | 1/64 | 1/128 |
|---|---|---|---|---|---|
| $\|\boldsymbol{q}-\boldsymbol{q}_h\|_{0,\Omega}$ | 2.8775E-01 | 1.4501E-01 | 7.2649E-02 | 3.6342E-02 | 1.8173E-02 |
| order | - | 0.98861 | 0.99716 | 0.99929 | 0.99982 |
| $\|\boldsymbol{p}-\boldsymbol{p}_h\|_{0,\Omega}$ | 2.1036E-01 | 1.0341E-01 | 5.1480E-02 | 2.5712E-02 | 1.2852E-02 |
| order | - | 1.0244 | 1.0063 | 1.0016 | 1.0004 |
| $\|y-y_h\|_{0,\Omega}$ | 1.1842E-02 | 3.2095E-03 | 8.4824E-04 | 2.1887E-04 | 5.5641E-05 |
| order | - | 1.8834 | 1.9198 | 1.9544 | 1.9759 |
| $\|z-z_h\|_{0,\Omega}$ | 1.8304E-02 | 5.3420E-03 | 1.4422E-03 | 3.7460E-04 | 9.5451E-05 |
| order | - | 1.7767 | 1.8891 | 1.9449 | 1.9725 |

**Theorem 3.** We have

$$\|\boldsymbol{q}-\boldsymbol{q}_h\|_{\mathcal{T}_h} \lesssim h^{k+1}(|\boldsymbol{q}|_{k+1}+|y|_{k+2}+|\boldsymbol{p}|_{k+1}+|z|_{k+2}), \tag{54a}$$

$$\|\boldsymbol{p}-\boldsymbol{p}_h\|_{\mathcal{T}_h} \lesssim h^{k+1}(|\boldsymbol{q}|_{k+1}+|y|_{k+2}+|\boldsymbol{p}|_{k+1}+|z|_{k+2}). \tag{54b}$$

## 4. Numerical Experiments

In this section, we present three numerical examples to confirm our theoretical results. We consider the problems on a square domain $\Omega = [0,1]\times[0,1] \subset \mathbb{R}^2$. For the first two examples, we take $\gamma = 1$, $\tau_1 = 1$, $\boldsymbol{\beta} = [x_2, x_1]$, and the exact state $y(x_1,x_2) = \sin(\pi x_1)$. We used the optimize-then-discretize (OD) approach in Example 1 and the discretize-then-optimize (DO) approach in Example 2. In the third example, we take $\gamma = 1$, $\tau_1 = 5$, $\boldsymbol{\beta} = [\cos(x_1)\exp(x_2), x_1\cos(x_2)]$, and the same exact state $y(x_1,x_2) = \sin(\pi x_1)$. In these examples, the data $f$, $g$, and $y_d$ is generated from the optimality system (8) after we specified the exact dual state $z(x_1,x_2) = \sin(\pi x_1)\sin(\pi x_2)$.

Numerical results for $k = 0$ and $k = 1$ for the two approaches are shown in Table 1–Table 4 for the first two examples. The observed convergence rates and numerical results exactly match the theoretical results.

**Example 1.** For the OD approach, we set the stabilization parameter $\tau_2$ using **(A1)**; hence, conditions **(A1)-(A2)** are satisfied. We obtain optimal convergence rates for all variables for $k = 0$ and $k = 1$ in Table 1 and Table 2, respectively. This matches our theoretical results.

**Example 2.** For the DO approach, we used the same data as in Example 1. From the tables we can see that the numerical results are exactly the same with the OD approach, which confirms our theoretical results.

**Example 3.** In this example, we give a brief comparison of the EDG method with an HDG method. Specifically, we use the HDG method for the optimal control of convection diffusion PDEs from [18]. This HDG method uses discontinuous polynomials of equal degree for all variables. There is no doubt that the degrees of freedom for the EDG method is much smaller than the HDG method if we use the same polynomial degree for the numerical trace in both methods. However, in this case the convergence rates of the HDG method (with an element-by-element postprocessing for the state variables) are one order higher than the EDG method.

Table 2. Example 1: Errors for the state $y$, adjoint state $z$, and the fluxes $\boldsymbol{q}$ and $\boldsymbol{p}$ when $k = 1$ with the OD approach.

| $h/\sqrt{2}$ | 1/8 | 1/16 | 1/32 | 1/64 | 1/128 |
|---|---|---|---|---|---|
| $\|\boldsymbol{q} - \boldsymbol{q}_h\|_{0,\Omega}$ | 1.8365E-02 | 4.9165E-03 | 1.2726E-03 | 3.2189E-04 | 8.0742E-05 |
| order | - | 1.9012 | 1.9498 | 1.9831 | 1.9952 |
| $\|\boldsymbol{p} - \boldsymbol{p}_h\|_{0,\Omega}$ | 1.6649E-02 | 5.6050E-03 | 1.5952E-03 | 4.1463E-04 | 1.0475E-04 |
| order | - | 1.5707 | 1.8129 | 1.9439 | 1.9848 |
| $\|y - y_h\|_{0,\Omega}$ | 1.3524E-03 | 1.8347E-04 | 2.3956E-05 | 3.0691E-06 | 3.8882E-07 |
| order | - | 2.8819 | 2.9371 | 2.9645 | 2.9807 |
| $\|z - z_h\|_{0,\Omega}$ | 3.2125E-03 | 4.2489E-04 | 5.4721E-05 | 6.9745E-06 | 8.8190E-07 |
| order | - | 2.9186 | 2.9569 | 2.9719 | 2.9834 |

Table 3. Example 2: Errors for the state $y$, adjoint state $z$, and the fluxes $\boldsymbol{q}$ and $\boldsymbol{p}$ when $k = 0$ with the DO approach.

| $h/\sqrt{2}$ | 1/8 | 1/16 | 1/32 | 1/64 | 1/128 |
|---|---|---|---|---|---|
| $\|\boldsymbol{q} - \boldsymbol{q}_h\|_{0,\Omega}$ | 2.8775E-01 | 1.4501E-01 | 7.2649E-02 | 3.6342E-02 | 1.8173E-02 |
| order | - | 0.98861 | 0.99716 | 0.99929 | 0.99982 |
| $\|\boldsymbol{p} - \boldsymbol{p}_h\|_{0,\Omega}$ | 2.1036E-01 | 1.0341E-01 | 5.1480E-02 | 2.5712E-02 | 1.2852E-02 |
| order | - | 1.0244 | 1.0063 | 1.0016 | 1.0004 |
| $\|y - y_h\|_{0,\Omega}$ | 1.1842E-02 | 3.2095E-03 | 8.4824E-04 | 2.1887E-04 | 5.5641E-05 |
| order | - | 1.8834 | 1.9198 | 1.9544 | 1.9759 |
| $\|z - z_h\|_{0,\Omega}$ | 1.8304E-02 | 5.3420E-03 | 1.4422E-03 | 3.7460E-04 | 9.5451E-05 |
| order | - | 1.7767 | 1.8891 | 1.9449 | 1.9725 |

Table 4. Example 2: Errors for the state $y$, adjoint state $z$, and the fluxes $\boldsymbol{q}$ and $\boldsymbol{p}$ when $k = 1$ with the DO approach.

| $h/\sqrt{2}$ | 1/8 | 1/16 | 1/32 | 1/64 | 1/128 |
|---|---|---|---|---|---|
| $\|\boldsymbol{q} - \boldsymbol{q}_h\|_{0,\Omega}$ | 1.8365E-02 | 4.9165E-03 | 1.2726E-03 | 3.2189E-04 | 8.0742E-05 |
| order | - | 1.9012 | 1.9498 | 1.9831 | 1.9952 |
| $\|\boldsymbol{p} - \boldsymbol{p}_h\|_{0,\Omega}$ | 1.6649E-02 | 5.6050E-03 | 1.5952E-03 | 4.1463E-04 | 1.0475E-04 |
| order | - | 1.5707 | 1.8129 | 1.9439 | 1.9848 |
| $\|y - y_h\|_{0,\Omega}$ | 1.3524E-03 | 1.8347E-04 | 2.3956E-05 | 3.0691E-06 | 3.8882E-07 |
| order | - | 2.8819 | 2.9371 | 2.9645 | 2.9807 |
| $\|z - z_h\|_{0,\Omega}$ | 3.2125E-03 | 4.2489E-04 | 5.4721E-05 | 6.9745E-06 | 8.8190E-07 |
| order | - | 2.9186 | 2.9569 | 2.9719 | 2.9834 |

Hence, to make a more fair comparison, for the numerical traces we take discontinuous piecewise linear basis functions for the HDG method and continuous piecewise quadratic basis functions for the EDG method; in this case, the convergence rates for all variables are the same for both methods (using postprocessing for the HDG method). From Table 5 and Table 6, we can see that the EDG method is competitive both in terms of accuracy and globally coupled degrees of freedom.

Table 5. Example 3: Errors for the state $y$, adjoint state $z$, and the fluxes $\boldsymbol{q}$ and $\boldsymbol{p}$ for the EDG method with continuous piecewise quadratic basis functions for the numerical trace. Here, DoF is the number of globally coupled degrees of freedom.

| $h/\sqrt{2}$ | 1/8 | 1/16 | 1/32 | 1/64 | 1/128 |
|---|---|---|---|---|---|
| $\|\boldsymbol{q}-\boldsymbol{q}_h\|_{0,\Omega}$ | 1.8432E-02 | 4.9222E-03 | 1.2730E-03 | 3.2192E-04 | 8.0743E-05 |
| order | - | 1.9048 | 1.9511 | 1.9835 | 1.9953 |
| $\|\boldsymbol{p}-\boldsymbol{p}_h\|_{0,\Omega}$ | 1.6809E-02 | 5.6229E-03 | 1.5966E-03 | 4.1473E-04 | 1.0476E-04 |
| order | - | 1.5798 | 1.8163 | 1.9448 | 1.9851 |
| $\|y-y_h\|_{0,\Omega}$ | 1.3561E-03 | 1.8358E-04 | 2.3959E-05 | 3.0692E-06 | 3.8882E-07 |
| order | - | 2.8851 | 2.9378 | 2.9646 | 2.9807 |
| $\|z-z_h\|_{0,\Omega}$ | 3.2125E-03 | 4.2475E-04 | 5.4714E-05 | 6.9743E-06 | 8.8190E-07 |
| order | - | 2.9190 | 2.9566 | 2.9718 | 2.9834 |
| DoF | 226 | 962 | 3970 | 16130 | 65026 |

Table 6. Example 3: Errors for the state $y$, adjoint state $z$, and the fluxes $\boldsymbol{q}$ and $\boldsymbol{p}$ for the HDG method (with postprocessing) from [18] with discontinuous piecewise linear basis functions for the numerical trace. Here, DoF is the number of globally coupled degrees of freedom, and the superscript $\star$ denotes the postprocessed approximations.

| $h/\sqrt{2}$ | 1/8 | 1/16 | 1/32 | 1/64 | 1/128 |
|---|---|---|---|---|---|
| $\|\boldsymbol{q}-\boldsymbol{q}_h\|_{0,\Omega}$ | 1.8427E-02 | 4.7138E-03 | 1.1891E-03 | 2.9831E-04 | 7.4684E-05 |
| order | - | 1.9668 | 1.9870 | 1.9950 | 1.9979 |
| $\|\boldsymbol{p}-\boldsymbol{p}_h\|_{0,\Omega}$ | 3.5193E-02 | 8.9732E-03 | 2.2614E-03 | 5.6736E-04 | 1.4208E-04 |
| order | - | 1.9716 | 1.9884 | 1.9949 | 1.9976 |
| $\|y-y_h\|_{0,\Omega}$ | 1.2751E-02 | 3.2022E-03 | 8.0021E-04 | 1.9989E-04 | 4.9944E-05 |
| order | - | 1.9935 | 2.0006 | 2.0012 | 2.0008 |
| $\|z-z_h\|_{0,\Omega}$ | 2.3555E-02 | 5.9284E-03 | 1.4837E-03 | 3.7092E-04 | 9.2716E-05 |
| order | - | 1.9903 | 1.9984 | 2.0000 | 2.0002 |
| $\|y-y_h^\star\|_{0,\Omega}$ | 8.2590E-04 | 1.0219E-04 | 1.2658E-05 | 1.5731E-06 | 1.9600E-07 |
| order | - | 3.0147 | 3.0132 | 3.0083 | 3.0047 |
| $\|z-z_h^\star\|_{0,\Omega}$ | 1.3013E-03 | 1.6247E-04 | 2.0306E-05 | 2.5383E-06 | 3.1729E-07 |
| order | - | 3.0017 | 3.0002 | 3.0000 | 3.0000 |
| DoF | 352 | 1472 | 6016 | 24320 | 97792 |

## 5. Conclusions

We considered a recently proposed EDG method to approximate the solution of an optimal distributed control problem for an elliptic convection diffusion equation. We showed the optimize-then-discretize and discretize-then-optimize approaches coincide, and proved optimal a priori error estimates for the control, state, dual state, and their fluxes. EDG methods are known to be competitive for convection dominated problems; therefore, this new EDG method has potential for optimal control problems involving such PDEs.

## Acknowledgements

X. Zhang thanks Missouri University of Science and Technology for hosting him as a visiting scholar; some of this work was completed during his research visit. Y. Zhang and J. Singler were supported in part by National Science Foundation grant DMS-1217122. Y. Zhang and J. Singler thank the IMA for funding research visits, during which some of this work was completed.

## References

[1] JT Betts and I Kolmanovsky. Practical methods for optimal control using nonlinear programming. Applied Mechanics Reviews, 55:B68, 2002.

[2] Aycil Cesmelioglu, Bernardo Cockburn, and Weifeng Qiu. Analysis of a hybridizable discontinuous Galerkin method for the steady-state incompressible Navier-Stokes equations. Math. Comp., 86(306):1643–1670, 2017.

[3] Yanlai Chen, Bernardo Cockburn, and Bo Dong. Superconvergent HDG methods for linear, stationary, third-order equations in one-space dimension. Math. Comp., 85(302):2715–2742, 2016.

[4] Bernardo Cockburn, Jayadeep Gopalakrishnan, and Raytcho Lazarov. Unified hybridization of discontinuous Galerkin, mixed, and continuous Galerkin methods for second order elliptic problems. SIAM J. Numer. Anal., 47(2):1319–1365, 2009.

[5] Bernardo Cockburn, Jayadeep Gopalakrishnan, Ngoc Cuong Nguyen, Jaume Peraire, and Francisco-Javier Sayas. Analysis of HDG methods for Stokes flow. Math. Comp., 80(274):723–760, 2011.

[6] Bernardo Cockburn, Johnny Guzmán, See-Chew Soon, and Henry K. Stolarski. An analysis of the embedded discontinuous Galerkin method for second-order elliptic problems. SIAM J. Numer. Anal., 47(4):2686–2707, 2009.

[7] Bernardo Cockburn and Kassem Mustapha. A hybridizable discontinuous Galerkin method for fractional diffusion problems. Numer. Math., 130(2):293–314, 2015.

[8] Bernardo Cockburn and Jiguang Shen. A hybridizable discontinuous Galerkin method for the $p$-Laplacian. SIAM J. Sci. Comput., 38(1):A545–A566, 2016.

[9] P. Fernandez, N.C. Nguyen, X. Roca, and J. Peraire. Implicit large-eddy simulation of compressible flows using the interior embedded discontinuous Galerkin method. In 54th AIAA Aerospace Sciences Meeting, AIAA SciTech Forum, AIAA 2016-1332, 2016.

[10] Gousheng Fu and Chi-Wang Shu. Analysis of an embedded discontinuous Galerkin method with implicit-explicit time-marching for convection-diffusion problems. International Journal of Numerical Analysis & Modeling, 14(4):477–499, 2017.

[11] Hongfei Fu. A characteristic finite element method for optimal control problems governed by convection-diffusion equations. J. Comput. Appl. Math., 235(3):825–836, 2010.

[12] Hongfei Fu and Hongxing Rui. A priori error estimates for optimal control problems governed by transient advection-diffusion equations. J. Sci. Comput., 38(3):290–315, 2009.

[13] Hongfei Fu and Hongxing Rui. A characteristic-mixed finite element method for time-dependent convection-diffusion optimal control problem. Appl. Math. Comput., 218(7):3430–3440, 2011.

[14] Philip E. Gill, Walter Murray, and Michael A. Saunders. SNOPT: an SQP algorithm for large-scale constrained optimization. SIAM Rev., 47(1):99–131, 2005.

[15] S. Güzey, B. Cockburn, and H. K. Stolarski. The embedded discontinuous Galerkin method: application to linear shell problems. Internat. J. Numer. Methods Engrg., 70(7):757–790, 2007.

[16] Matthias Heinkenschloss and Dmitriy Leykekhman. Local error estimates for SUPG solutions of advection-dominated elliptic linear-quadratic optimal control problems. SIAM J. Numer. Anal., 47(6):4607–4638, 2010.

[17] Weiwei Hu, Jiguang Shen, John R. Singler, Yangwen Zhang, and Xiabo Zheng. A superconvergent HDG method for distributed control of convection diffusion PDEs. Journal of Scientific Computing. To appear.

[18] Weiwei Hu, Jiguang Shen, John R. Singler, Yangwen Zhang, and Xiabo Zheng. An HDG method for distributed control of convection diffusion PDEs. Journal of Computational and Applied Mathematics. To appear.

[19] D. S. Kamenetskiy. On the relation of the embedded discontinuous Galerkin method to the stabilized residual-based finite element methods. Appl. Numer. Math., 108:271–285, 2016.

[20] Dmitriy Leykekhman. Investigation of commutative properties of discontinuous Galerkin methods in PDE constrained optimal control problems. Journal of Scientific Computing, 53(3):483–511, 2012.

[21] Dmitriy Leykekhman and Matthias Heinkenschloss. Local error analysis of discontinuous Galerkin methods for advection-dominated elliptic linear-quadratic optimal control problems. SIAM J. Numer. Anal., 50(4):2012–2038, 2012.

[22] Jun Liu and Zhu Wang. Non-commutative discretize-then-optimize algorithms for elliptic PDE-constrained optimal control problems. arXiv preprint arXiv:1706.07652.

[23] Kassem Mustapha, Maher Nour, and Bernardo Cockburn. Convergence and superconvergence analyses of HDG methods for time fractional diffusion problems. Adv. Comput. Math., 42(2):377–393, 2016.

[24] N. C. Nguyen, J. Peraire, and B. Cockburn. An implicit high-order hybridizable discontinuous Galerkin method for linear convection-diffusion equations. J. Comput. Phys., 228(9):3232–3254, 2009.

[25] N. C. Nguyen, J. Peraire, and B. Cockburn. An implicit high-order hybridizable discontinuous Galerkin method for nonlinear convection-diffusion equations. J. Comput. Phys., 228(23):8841–8855, 2009.

[26] N. C. Nguyen, J. Peraire, and B. Cockburn. A hybridizable discontinuous Galerkin method for Stokes flow. Comput. Methods Appl. Mech. Engrg., 199(9-12):582–597, 2010.

[27] N. C. Nguyen, J. Peraire, and B. Cockburn. A class of embedded discontinuous Galerkin methods for computational fluid dynamics. J. Comput. Phys., 302:674–692, 2015.

[28] Jorge Nocedal and Stephen J Wright. Sequential quadratic programming. Springer, 2006.

[29] Jaime Peraire, NC Nguyen, and Bernardo Cockburn. An embedded discontinuous Galerkin method for the compressible Euler and Navier-Stokes equations. In Proceedings of the 20th AIAA Computational Fluid Dynamics Conference, AIAA 2011-3228, 2011.

[30] Weifeng Qiu and Ke Shi. An HDG method for convection diffusion equation. J. Sci. Comput., 66(1):346–357, 2016.

[31] Yousef Saad. Iterative methods for sparse linear systems. Society for Industrial and Applied Mathematics, Philadelphia, PA, second edition, 2003.

[32] M. Stanglmeier, N. C. Nguyen, J. Peraire, and B. Cockburn. An explicit hybridizable discontinuous Galerkin method for the acoustic wave equation. Comput. Methods Appl. Mech. Engrg., 300:748–769, 2016.

[33] Tongjun Sun. Discontinuous Galerkin finite element method with interior penalties for convection diffusion optimal control problem. Int. J. Numer. Anal. Model., 7(1):87–107, 2010.

[34] Chunguang Xiong and Yuan Li. Error analysis for optimal control problem governed by convection diffusion equations: DG method. J. Comput. Appl. Math., 235(10):3163–3177, 2011.

[35] Ningning Yan and Zhaojie Zhou. A RT mixed FEM/DG scheme for optimal control governed by convection diffusion equations. J. Sci. Comput., 41(2):273–299, 2009.

[36] Hamdullah Yücel, Martin Stoll, and Peter Benner. A discontinuous Galerkin method for optimal control problems governed by a system of convection-diffusion PDEs with nonlinear reaction terms. Comput. Math. Appl., 70(10):2414–2431, 2015.

[37] X. Zhang, X. Xie, and S. Zhang. An optimal embedded discontinuous Galerkin method for second-order elliptic problems. Computational Methods in Applied Mathematics. To appear.

[38] X. Zhang, Y. Zhang, and John R. Singler. An EDG method for distributed optimal control of elliptic PDEs. Advances in Applied Mathematics and Mechanics. To appear.

[39] Zhaojie Zhou, Fengxin Chen, and Huanzhen Chen. Characteristic mixed finite element approximation of transient convection diffusion optimal control problems. Math. Comput. Simulation, 82(11):2109–2128, 2012.

[40] Zhaojie Zhou and Ningning Yan. The local discontinuous Galerkin method for optimal control problem governed by convection diffusion equations. Int. J. Numer. Anal. Model., 7(4):681–699, 2010.

[41] Zhaojie Zhou, Xiaoming Yu, and Ningning Yan. Local discontinuous Galerkin approximation of convection-dominated diffusion optimal control problems with control constraints. Numer. Methods Partial Differential Equations, 30(1):339–360, 2014.

## Appendix

By simple algebraic operations in equation (19b), we obtain the following formulas for $G_1$, $G_2$, $G_3$, $G_4$, $H_1$, and $H_2$ in (20):

$$\begin{aligned}
G_1 &= -A_1^{-1}A_2(A_4 + A_2^T A_1^{-1} A_2)^{-1}(A_5 - A_2^T A_1^{-1} A_3) - A_1^{-1} A_3,\\
G_2 &= A_1^{-1}A_2(A_4 + A_2^T A_1^{-1} A_2)^{-1} A_6,\\
G_3 &= -(A_4 + A_2^T A_1^{-1} A_2)^{-1}(A_5 - A_2^T A_1^{-1} A_3),\\
G_4 &= (A_4 + A_2^T A_1^{-1} A_2)^{-1} A_6,\\
H_1 &= A_1^{-1}A_2(A_4 + A_2^T A_1^{-1} A_2)^{-1}(b_3 - b_4 + A_2^T A_1^{-1} b_2) - A_1^{-1} b_2,\\
H_2 &= (A_4 + A_2^T A_1^{-1} A_2)^{-1}(b_3 - b_4 + A_2^T A_1^{-1} b_2).
\end{aligned}$$

In general, forming these quantities is impractical; however, for the EDG method described in this work these matrices can be easily computed. We briefly sketch this process below.

Since the spaces $\boldsymbol{V}_h$ and $W_h$ consist of discontinuous polynomials, some of the system matrices are block diagonal and each block is small and symmetric positive definite (SSPD). The inverse of such a matrix is another matrix of the same type, and the inverse is easily computed by inverting each small block. Furthermore, the inverse of each small block can be computed in parallel.

It can be checked that $A_1$ is a SSPD block diagonal matrix, and therefore $A_1^{-1}$ is easily computed and is also a SSPD block diagonal matrix. Therefore, $G_1$, $G_2$, $G_3$, $G_4$, $H_1$, and $H_2$ are easily computed since $A_4 + A_2^T A_1^{-1} A_2$ is also a SSPD block diagonal matrix. Also, once these quantities are computed, $G_5$, $G_6$, and $H_3$ in (20) are also easy to compute using (19b).

College of Mathematics, Sichuan University, China
*E-mail*: zhangxiaofem@163.com

Department of Mathematics and Statistics, Missouri University of Science and Technology, Rolla, MO 65409, USA
*E-mail*: ywzfg4@mst.edu

Department of Mathematics and Statistics, Missouri University of Science and Technology, Rolla, MO 65409, USA
*E-mail*: singlerj@mst.edu