

## ANALYSIS OF AN EMBEDDED DISCONTINUOUS GALERKIN METHOD WITH IMPLICIT-EXPLICIT TIME-MARCHING FOR CONVECTION-DIFFUSION PROBLEMS

GUOSHENG FU AND CHI-WANG SHU

**Abstract.** In this paper, we analyze implicit-explicit (IMEX) Runge-Kutta (RK) time discretization methods for solving linear convection-diffusion equations. The diffusion operator is treated implicitly via the embedded discontinuous Galerkin (EDG) method and the convection operator explicitly via the upwinding discontinuous Galerkin method.

**Key words.** Embedded discontinuous Galerkin method, upwinding discontinuous Galerkin method, implicit-explicit Runge-Kutta time-marching scheme, convection-diffusion equation, stability, error estimate, energy method.

### 1. Introduction

In this paper, we propose and analyze an implicit-explicit embedded discontinuous Galerkin (IMEX-EDG) method for solving the linear convection diffusion equation. We use the IMEX Runge-Kutta time discretization [1] that treats the diffusion term implicitly via the embedded discontinuous Galerkin (EDG) method [7, 6] and the convection term explicitly via the upwinding discontinuous Galerkin method [9]. For a detailed discussion on IMEX RK schemes, see [1, 3, 8] and references therein.

The EDG methods, originally introduced for linear shells in [7], is obtained from hybridizable discontinuous Galerkin (HDG) methods [5] by simply reducing the space of the hybrid (interface) unknowns by requiring them to be continuous across the mesh skeleton. It reduces the globally coupled degrees of freedom (after hybridization) to exactly those for a continuous Galerkin formulation (after static condensation).

Here we consider three specific Runge-Kutta type IMEX schemes given in [1] from first to third order accuracy. Coupling with the EDG (diffusion) and upwinding DG (convection) spatial discretization, we give the stability analysis and error estimates by the energy method. Our work is inspired from [10, 11, 12], where the authors analyzed IMEX time stepping coupled with local discontinuous Galerkin (LDG) methods for linear and nonlinear convection diffusion equations. The only difference of the IMEX-LDG and IMEX-EDG methods is on the discretization of the diffusion operator. While the theoretical results are similar for both spatial approaches, the IMEX-EDG methods is more computationally efficient due to a smaller number of globally coupled degrees of freedom. On a fixed triangular mesh in two dimensions, using polynomials of degree  $k$  approximations, the LDG method results a globally coupled linear system of size  $N_t(k+1)(k+2)/2 \approx N_v(k+1)(k+2)$ , while the EDG method results a globally coupled linear system of size  $N_v + N_e(k-1) \approx N_v(3k-2)$ . Here  $N_v$ ,  $N_e$ , and  $N_t$

---

Received by the editors on November 9, 2016, accepted on January 19, 2017.  
2000 *Mathematics Subject Classification.* 65M12, 65M15, 65M60.

are the numbers of vertices, edges, and triangles. We remark that while we can equally use the HDG methods [5] (with a discontinuous hybrid space) to discretize the diffusion operator, the EDG methods is more efficient in terms of the number of globally coupled degrees of freedom.

The paper is organized as follows. In Section 2 we present the spatial discretization for the model convection diffusion problem and give some preliminary results. Then, in Section 3, we present and analyze the fully discrete schemes with IMEX RK time discretization. Several numerical tests are presented in Section 4 to verify the main results in Section 3. Finally, we conclude in Section 5.

## 2. Semi-discretization with EDG for diffusion and upwinding DG for convection

In this section, we present the spatial discretization for the following linear convection-diffusion problem:

$$(1a) \quad u_t + \nabla \cdot (\boldsymbol{\beta}u) - \nabla \cdot (\epsilon \nabla u) = 0, \quad (\mathbf{x}, t) \in \Omega_T = \Omega \times (0, T],$$

$$(1b) \quad u(\mathbf{x}, 0) = u_0(\mathbf{x}), \quad \mathbf{x} \in \Omega,$$

with a periodic boundary condition. Here  $\Omega \in \mathbb{R}^d$  ( $d = 1, 2, 3$ ) is a bounded rectangular domain,  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_d) \in \mathbb{R}^d$  is a constant velocity field,  $\epsilon$  is the diffusion coefficient, and  $u_0(\mathbf{x})$  is the initial solution.

We use the EDG scheme [7, 6] to discretize the diffusion operator and the upwinding DG scheme [9] to discretize the convection operator. We present properties of these schemes that will be used for the analysis of the fully discrete schemes in Section 3.

We first collect some notation that will be used throughout the paper.

**2.1. Notation and preliminaries.** We denote by  $\|\cdot\|_{H^m(D)}$  the standard  $H^m$ -Sobolev norm on the domain  $D \subset \mathbb{R}^d$ . When  $m = 0$ , we simplify the notation and denote by  $\|\cdot\|_D$  the  $L^2$ -norm on  $D$ .

We denote by  $\mathcal{T}_h := \{K\}$  a quasi-uniform, shape-regular conforming simplicial triangulation of  $\Omega$ , and by  $\mathcal{E}_h$  the mesh skeleton consists the set of facets  $F$  (element nodes in  $1d$ , edges in  $2d$ , and faces in  $3d$ ) of the simplicial elements  $K \in \mathcal{T}_h$ . We denote by  $\partial K$  the element boundary of an element  $K$ .

We denote by  $\text{Volume}(K)$  and  $\text{Volume}(\partial K)$  the volume and surface area of  $K$  in  $3d$ . In  $2d$ ,  $\text{Volume}(K)$  is the area of the triangle  $K$ , and  $\text{Volume}(\partial K)$  is the perimeter length. And in  $1d$ ,  $\text{Volume}(K)$  is the length of the interval  $K$ , and  $\text{Volume}(\partial K)$  is set to be 2. We set  $h_K := \text{diam}(K)$  and  $h := \max_{K \in \mathcal{T}_h} h_K$ .

Associated with the triangulation and mesh skeleton, we define the discontinuous (cell-wise) finite element spaces (on  $\mathcal{T}_h$ ) and continuous (facet-wise) finite element space (on  $\mathcal{E}_h$ ):

$$(2a) \quad \mathbf{R}_h := \{\mathbf{r} \in L^2(\mathcal{T}_h)^d : \mathbf{r}|_K \in \mathcal{P}_{k-1}^d(K), K \in \mathcal{T}_h\},$$

$$(2b) \quad V_h := \{v \in L^2(\mathcal{T}_h) : v|_K \in \mathcal{P}_k(K), K \in \mathcal{T}_h\},$$

$$(2c) \quad M_h := \{\hat{v}_h \in C^0(\mathcal{E}_h) : \hat{v}_h|_F \in \mathcal{P}_k(F), F \in \mathcal{E}_h\},$$

for  $k \geq 1$ . Here  $\mathcal{P}_m(K)$  ( $\mathcal{P}_m^d(K)$ ) stands for the space of scalar (vector) polynomials of degree at most  $m$ . We use the convention that  $\mathcal{P}_m(F)$  is the space of constants

for any  $m \geq 0$  when the facet  $F$  is a node (1d case). Note that we require  $C^0$ -continuity for the approximate trace space  $M_h$  (nodal continuity in 2d, and nodal and edge continuity in 3d).

We recall the following optimal  $hp$  inverse trace inequality from [13] that will be used to define the stabilization parameter of our EDG scheme.

**Lemma 1.** [13, Theorem 5] *For a  $d$ -simplex,  $D$ , the following result holds  $\forall u \in \mathcal{P}_p(D)$*

$$\|u\|_{\partial D} \leq \sqrt{\frac{(p+1)(p+d)}{d} \frac{\text{Volume}(\partial D)}{\text{Volume}(D)}} \|u\|_D.$$

Here, by convention, when  $D = (a, b)$  is an interval,  $\|u\|_{\partial D} = \sqrt{u(a)^2 + u(b)^2}$ .

To simplify notation, we denote the optimal  $hp$  inverse trace constant in Lemma 1 as

$$(3) \quad C_{p,D} := \sqrt{\frac{(p+1)(p+d)}{d} \frac{\text{Volume}(\partial D)}{\text{Volume}(D)}}.$$

Now, we prepare some notation that will be used to define the upwinding DG scheme. We denote by  $\partial K^+ := \{F \in \partial K : \boldsymbol{\beta} \cdot \mathbf{n}_F > 0\}$  the *outflow* boundary, and  $\partial K^- := \{F \in \partial K : \boldsymbol{\beta} \cdot \mathbf{n}_F < 0\}$  the *inflow* boundary. Here  $\mathbf{n}_F$  is the outward normal of  $K$  on  $F$ . For each facet  $F$ , we define the *inflow* and *outflow* elements  $K^-$  and  $K^+$  which share the same facet  $F$  by requiring  $\boldsymbol{\beta} \cdot \mathbf{n}_{K^-} < 0$  and  $\boldsymbol{\beta} \cdot \mathbf{n}_{K^+} > 0$ . When the normal direction of  $F$  is parallel to  $\boldsymbol{\beta}$ , we take  $K^-$  as the element on the left side of the direction  $\boldsymbol{\beta}$ . Along the facet  $F$ , there are two traces for a scalar function  $p$ , denoted by  $p^+ = (p|_{K^+})|_F$  and  $p^- = (p|_{K^-})|_F$ , respectively. We denote the *jump* as  $\llbracket p \rrbracket := p^+ - p^-$ .

We write  $(\eta, \zeta)_{\mathcal{T}_h} := \sum_{K \in \mathcal{T}_h} (\eta, \zeta)_K$ , where  $(\eta, \zeta)_D$  denotes the integral of  $\eta\zeta$  over the domain  $D \subset \mathbb{R}^d$ . We also write  $\langle \eta, \zeta \rangle_{\partial \mathcal{T}_h} := \sum_{K \in \mathcal{T}_h} \langle \eta, \zeta \rangle_{\partial K}$ , where  $\langle \eta, \zeta \rangle_D$  denotes the integral of  $\eta\zeta$  over the domain  $D \subset \mathbb{R}^{d-1}$  (in 1d the integral is just the product of point values) and where  $\partial \mathcal{T}_h := \{\partial K : K \in \mathcal{T}_h\}$ . When vector-valued functions are involved, we use a similar notation.

We denote the velocity magnitude as  $\beta_{\max} = \sqrt{\sum_{i=1}^d \beta_i^2}$ .

**2.2. EDG for diffusion.** In this subsection, we present the EDG discretization for the diffusion operator  $-\nabla \cdot (\epsilon \nabla u)$ . To facilitate our description, we consider the following steady-state diffusion problem with a periodic boundary condition:

$$(4) \quad -\nabla \cdot (\epsilon \nabla u) = f \quad \text{in } \Omega, \quad \int_{\Omega} u = g.$$

Here  $f$  is the source term and  $g$  is a prescribed average of the solution  $u$ .

We start with the following equivalent first-order reformulation of (4):

$$(5a) \quad \epsilon^{-1} \mathbf{q} + \nabla u = 0 \quad \text{in } \Omega,$$

$$(5b) \quad \nabla \cdot \mathbf{q} = f \quad \text{in } \Omega,$$

$$(5c) \quad \int_{\Omega} u = g,$$

with a periodic boundary condition for  $u$ .

The EDG scheme for (5) is defined as follows: find  $(\mathbf{q}_h, u_h, \widehat{u}_h) \in \mathbf{R}_h \times V_h \times M_h$  such that

$$(6a) \quad (\epsilon^{-1} \mathbf{q}_h, \mathbf{r}_h)_{\mathcal{T}_h} - (u_h, \nabla \cdot \mathbf{r}_h)_{\mathcal{T}_h} + \langle \widehat{u}_h, \mathbf{r}_h \cdot \mathbf{n} \rangle_{\partial \mathcal{T}_h} = 0,$$

$$(6b) \quad -(\mathbf{q}_h, \nabla v_h)_{\mathcal{T}_h} + \langle \mathbf{q}_h \cdot \mathbf{n} + \alpha(u_h - \widehat{u}_h), v_h \rangle_{\partial \mathcal{T}_h} = (f, v_h)_{\mathcal{T}_h},$$

$$(6c) \quad -\langle \mathbf{q}_h \cdot \mathbf{n} + \alpha(u_h - \widehat{u}_h), \widehat{v}_h \rangle_{\partial \mathcal{T}_h} = 0,$$

$$(6d) \quad (u_h, 1)_{\mathcal{T}_h} = (g, 1)_{\mathcal{T}_h},$$

for all  $(\mathbf{r}_h, v_h, \widehat{v}_h) \in \mathbf{R}_h \times V_h \times M_h$ , where

$$\alpha \in \{\mu \in L^2(\partial \mathcal{T}_h) : \mu|_F \in \mathcal{P}_0(F), \quad \forall F \in \partial \mathcal{T}_h\}$$

is a (positive, piecewise-constant) stabilization parameter defined on the collection of element boundaries  $\partial \mathcal{T}_h$ . We choose  $\alpha$  to be single-valued on each facet  $F \in \mathcal{T}_h$  based on the sharp  $hp$  inverse trace constant in (3), whose expression restricted on a facet  $F = \partial K^1 \cap \partial K^2$  shared by two elements  $K^1$  and  $K^2$  is given as follows:

$$(6e) \quad \alpha|_F = \epsilon \max\{C_{k,K^1}^2, C_{k,K^2}^2\}.$$

Note that  $\alpha|_F$  scales proportionally to  $\epsilon(k+1)^2/h_K$ , where  $h_K = \min\{h_{K^1}, h_{K^2}\}$ .

The following result states that the above EDG scheme gives  $hp$ -optimal  $L^2$  error estimates. We postpone its proof via an energy argument to the Appendix. Similar analysis was used in [6] in which  $h$ -optimal  $L^2$  estimates were obtained for the EDG scheme for pure diffusion problems.

**Theorem 1.** *There exists a unique solution  $(\mathbf{q}_h, u_h, \widehat{u}_h) \in \mathbf{R}_h \times V_h \times M_h$  to the EDG scheme (6).*

*Moreover, let  $u \in H^{k+1}(\Omega)$  be a smooth function, and  $u_h \in V_h$  be part of the solution to the EDG scheme (6) with right hand sides  $f$  replaced by  $-\nabla \cdot \epsilon \nabla u$  and  $g$  replaced by  $u$ . Then, we have*

$$\|u - u_h\|_0 \leq C \left(\frac{h}{k}\right)^{k+1} \|u\|_{k+1},$$

*with a constant  $C$  only depending on the shape-regularity and quasi-uniformity of the triangulation  $\mathcal{T}_h$ , and independent of the mesh size  $h$  and the polynomial degree  $k$ .*

The above result implies that the elliptic projection defined through the above EDG scheme gives optimal  $L^2$  estimates, we will use this fact to prove the error estimates for the fully-discrete IMEX scheme in Section 3.

To simplify our presentation for the analysis of the fully discrete scheme, we now transform the EDG scheme (6) to its primal form, expressing it in terms of  $u_h \in V_h$  only. To achieve this, we rewrite  $\mathbf{q}_h$  and  $\widehat{u}_h$  as a linear mapping applied to  $u_h$ , which is defined by using equations (6a) and (6c). Thus, for any  $w \in V_h$ , we define  $(Q_w, \widehat{U}_w) \in \mathbf{R}_h \times M_h$  as the solution of

$$(7a) \quad (\epsilon^{-1} Q_w, \mathbf{r}_h)_{\mathcal{T}_h} + \langle \widehat{U}_w, \mathbf{r}_h \cdot \mathbf{n} \rangle_{\partial \mathcal{T}_h} = (w, \nabla \cdot \mathbf{r}_h)_{\mathcal{T}_h},$$

$$(7b) \quad -\langle Q_w \cdot \mathbf{n}, \widehat{v}_h \rangle_{\partial \mathcal{T}_h} + \langle \alpha \widehat{U}_w, \widehat{v}_h \rangle_{\partial \mathcal{T}_h} = \langle \alpha w, \widehat{v}_h \rangle_{\partial \mathcal{T}_h},$$

for all  $(\mathbf{r}_h, \widehat{v}_h) \in \mathbf{R}_h \times M_h$ . A simple energy argument ensures the existence and uniqueness of the solution  $(Q_w, \widehat{U}_w) \in \mathbf{R}_h \times M_h$  for any given  $w \in V_h$ .

Using the above mapping, we define the *primal* EDG scheme for (4) as follows: find  $u_h \in V_h$  with  $(u_h, 1)_{\mathcal{T}_h} = (g, 1)_{\mathcal{T}_h}$  such that, for all  $v_h \in V_h$ ,

$$(8) \quad (\epsilon^{-1}Q_{u_h}, Q_{v_h})_{\mathcal{T}_h} + \langle \alpha(u_h - \widehat{U}_{u_h}), v_h - \widehat{U}_{v_h} \rangle_{\partial\mathcal{T}_h} = (f, v_h)_{\mathcal{T}_h}.$$

We have the following result on the equivalence of the two EDG schemes (6) and (8).

**Lemma 2.** *Let  $u_h \in V_h$  be the solution to the primal EDG scheme (8), then  $(Q_{u_h}, u_h, \widehat{U}_{u_h}) \in \mathbf{R}_h \times V_h \times M_h$  is the unique solution to the EDG scheme (6).*

*Proof.* The fact that  $(Q_{u_h}, u_h, \widehat{U}_{u_h}) \in \mathbf{R}_h \times V_h \times M_h$  satisfies (6a),(6c), (6d) is trivial to verify by definition of the mapping (7). Now, we show  $(Q_{u_h}, u_h, \widehat{U}_{u_h})$  satisfies (6b). Taking  $w = v_h$  in (7a) and choosing test function  $\mathbf{r}_h = Q_{u_h}$ , we get

$$(\epsilon^{-1}Q_{v_h}, Q_{u_h})_{\mathcal{T}_h} + \langle \widehat{U}_{v_h}, Q_{u_h} \cdot \mathbf{n} \rangle_{\partial\mathcal{T}_h} = (v_h, \nabla \cdot Q_{u_h})_{\mathcal{T}_h},$$

taking  $w = u_h$  in (7b) and choosing test function  $\widehat{v}_h = \widehat{U}_{v_h}$ , we get

$$-\langle Q_{u_h} \cdot \mathbf{n}, \widehat{U}_{v_h} \rangle_{\partial\mathcal{T}_h} + \langle \alpha \widehat{U}_{u_h}, \widehat{U}_{v_h} \rangle_{\partial\mathcal{T}_h} = \langle \alpha u_h, \widehat{U}_{v_h} \rangle_{\partial\mathcal{T}_h}.$$

Adding up the above expressions, we get

$$(9) \quad (\epsilon^{-1}Q_{v_h}, Q_{u_h})_{\mathcal{T}_h} - \langle \alpha(u_h - \widehat{U}_{u_h}), \widehat{U}_{v_h} \rangle_{\partial\mathcal{T}_h} = (v_h, \nabla \cdot Q_{u_h})_{\mathcal{T}_h}.$$

Hence

$$\begin{aligned} & - (Q_{u_h}, \nabla v_h)_{\mathcal{T}_h} + \langle Q_{u_h} \cdot \mathbf{n} + \alpha(u_h - \widehat{U}_{u_h}), v_h \rangle_{\partial\mathcal{T}_h} \\ & = (\nabla \cdot Q_{u_h}, v_h)_{\mathcal{T}_h} + \langle \alpha(u_h - \widehat{U}_{u_h}), v_h \rangle_{\partial\mathcal{T}_h} \\ & = (\epsilon^{-1}Q_{v_h}, Q_{u_h})_{\mathcal{T}_h} + \langle \alpha(u_h - \widehat{U}_{u_h}), v_h - \widehat{U}_{v_h} \rangle_{\partial\mathcal{T}_h}, \end{aligned}$$

where we used (9) in the last step. By (8), the above expression is equal to  $(f, v_h)_{\mathcal{T}_h}$ . Hence,  $(Q_{u_h}, u_h, \widehat{U}_{u_h})$  satisfies (6b). This implies that  $(Q_{u_h}, u_h, \widehat{U}_{u_h}) \in \mathbf{R}_h \times V_h \times M_h$  is the unique solution to the EDG scheme (6).  $\square$

We emphasize that the primal form of the EDG scheme (8) is only used for ease of presentation in the analysis of the fully discrete scheme considered in the next section. In the actual implementation, the cell-wise unknowns  $\mathbf{q}_h$  and  $u_h$  can be locally condensed out, and we only need to solve a global linear system for the trace unknown  $\widehat{u}_h$ .

Now, we denote the (symmetric) bilinear form associated with the primal EDG scheme (8) as  $B_d(u_h, v_h)$ , i.e.,

$$(10) \quad B_d(u_h, v_h) := (\epsilon^{-1}Q_{u_h}, Q_{v_h})_{\mathcal{T}_h} + \langle \alpha(u_h - \widehat{U}_{u_h}), v_h - \widehat{U}_{v_h} \rangle_{\partial\mathcal{T}_h}.$$

The following discrete seminorms on  $V_h$  will be useful for our analysis: for  $w_h \in V_h$ , we denote

$$(11a) \quad \|w_h\|_{e, \mathcal{T}_h} = \sqrt{(\epsilon^{-1}Q_{w_h}, Q_{w_h})_{\mathcal{T}_h} + \langle \alpha(w_h - \widehat{U}_{w_h}), w_h - \widehat{U}_{w_h} \rangle_{\partial\mathcal{T}_h}},$$

$$(11b) \quad \|w_h\|_{1, \mathcal{T}_h} = \sqrt{(\nabla w_h, \nabla w_h)_{\mathcal{T}_h} + \sum_{K \in \mathcal{T}_h} \langle \frac{C_{k,K}^2}{2} \llbracket w_h \rrbracket, \llbracket w_h \rrbracket \rangle_{\partial K}}.$$

We have  $B_d(w_h, w_h) = \|w_h\|_{e, \mathcal{T}_h}^2$  for any  $w_h \in V_h$ .

The following result states the control of  $\|w_h\|_{1,\mathcal{T}_h}$  by  $\|w_h\|_{e,\mathcal{T}_h}$ , which plays a key role in obtaining the good stability of the fully discrete IMEX scheme in the next section (compare with [10, Lemma 2.4]).

**Lemma 3.** *For any  $w_h \in V_h$ , we have*

$$\|w_h\|_{1,\mathcal{T}_h} \leq 2\epsilon^{-1/2}\|w_h\|_{e,\mathcal{T}_h}$$

*Proof.* We use equation (7a) (taking  $w = w_h$ ) and the definition of the stability parameter  $\alpha$  in (6e) to prove this result.

Taking  $\mathbf{r}_h = \nabla w_h$  in (7a), and integrating by parts the right hand side of the resulting expression, we have

$$(\epsilon^{-1} Q_{w_h}, \nabla w_h)_{\mathcal{T}_h} + \langle \widehat{U}_{w_h} - w_h, \nabla w_h \cdot \mathbf{n} \rangle_{\partial\mathcal{T}_h} = -(\nabla w_h, \nabla w_h)_{\mathcal{T}_h}.$$

Applying the Cauchy-Schwarz inequality for the left hand side of the above expression, we get

$$\begin{aligned} \|\nabla w_h\|_{\mathcal{T}_h}^2 &\leq \left( \epsilon^{-2} \|Q_{w_h}\|_{\mathcal{T}_h}^2 + \sum_{K \in \mathcal{T}_h} C_{k,K}^2 \|w_h - \widehat{U}_{w_h}\|_{\partial K}^2 \right)^{1/2} \\ &\quad \left( \|\nabla w_h\|_{\mathcal{T}_h}^2 + \sum_{K \in \mathcal{T}_h} C_{k,K}^{-2} \|\nabla w_h\|_{\partial K}^2 \right)^{1/2} \\ &\leq \epsilon^{-1/2} \|w_h\|_{e,\mathcal{T}_h} \left( \|\nabla w_h\|_{\mathcal{T}_h}^2 + \sum_{K \in \mathcal{T}_h} \|\nabla w_h\|_K^2 \right)^{1/2} \\ &\leq \sqrt{2} \epsilon^{-1/2} \|w_h\|_{e,\mathcal{T}_h} \|\nabla w_h\|_{\mathcal{T}_h}, \end{aligned}$$

where, in the second inequality, we used the inverse trace inequality from Lemma 1. Hence,

$$(12) \quad \|\nabla w_h\|_{\mathcal{T}_h}^2 \leq 2\epsilon^{-1} \|w_h\|_{e,\mathcal{T}_h}^2.$$

On the other hand, by the definition of  $\alpha$  in (6e), on any facet  $F = \partial K^1 \cap \partial K^2$ , we have

$$\sum_{i=1}^2 \frac{C_{k,K^i}^2}{2} \|[w_h]\|_F^2 \leq (\epsilon^{-1} \alpha|_F) \|[w_h]\|_F^2 \leq (2\epsilon^{-1} \alpha|_F) \sum_{i=1}^2 \|w_h|_{K^i} - \widehat{U}_{w_h}\|_F^2.$$

Summing the above expressions over all the facets  $F \in \mathcal{E}_h$ , we obtain

$$(13) \quad \sum_{K \in \mathcal{T}_h} \frac{C_{k,K}^2}{2} \|[w_h]\|_{\partial K}^2 = \sum_{F \in \mathcal{E}_h} \sum_{i=1}^2 \frac{C_{k,K^i}^2}{2} \|[w_h]\|_F^2 \\ \leq 2\epsilon^{-1} \langle \alpha(w_h - \widehat{U}_{w_h}), w_h - \widehat{U}_{w_h} \rangle_{\partial\mathcal{T}_h} \leq 2\epsilon^{-1} \|w_h\|_{e,\mathcal{T}_h}^2.$$

Finally, combining the inequalities in (12) and (13), we obtain the desired estimate of Lemma 3.  $\square$

**2.3. Upwinding DG for convection.** In this subsection we present the upwinding DG discretization for the convection operator  $\nabla \cdot (\boldsymbol{\beta} u)$  and states its relevant properties. The associated bilinear form, acting on the space  $V_h$ , is given as follows:

$$(14) \quad B_c(w_h, v_h) := (-\boldsymbol{\beta} w, \nabla v)_{\mathcal{T}_h} + \langle \boldsymbol{\beta} \cdot \mathbf{n} w_h^-, v_h \rangle_{\partial\mathcal{T}_h}.$$

We collect the relevant stability properties of the above bilinear form in the next lemma.

**Lemma 4.** *For any  $w_h, v_h \in V_h$ , we have*

$$\begin{aligned} B_c(w_h, w_h) &= \frac{1}{2} \sum_{F \in \mathcal{E}_h} \|\beta \cdot \mathbf{n}\|^{1/2} \llbracket w_h \rrbracket \|_F^2 \geq 0, \\ B_c(w_h, v_h) &\leq \sqrt{3} \beta_{\max} \|w_h\|_{\mathcal{T}_h} \|v_h\|_{1, \mathcal{T}_h}, \\ B_c(w_h, v_h) &\leq \sqrt{3} \beta_{\max} \|w_h\|_{1, \mathcal{T}_h} \|v_h\|_{\mathcal{T}_h}. \end{aligned}$$

For any  $w \in V_h + H^1(\Omega)$  and  $v_h \in V_h$ , we have

$$B_c(w, v_h) \leq \beta_{\max} (\|w\|_{\mathcal{T}_h}^2 + \sum_{K \in \mathcal{T}_h} \frac{2}{C_{k,K}^2} \|w\|_{\partial K^+}^2)^{1/2} \|v_h\|_{1, \mathcal{T}_h}.$$

Recall that  $\beta_{\max} = \sqrt{\sum_{i=1}^d \beta_i^2}$ .

*Proof.* Taking  $v_h := w_h$  in (14), and integrating by parts the resulting element-wise integrals, we obtain the first equality.

To prove the second inequality of Lemma 4, we note that

$$\begin{aligned} B_c(w_h, v_h) &= -(\beta w_h, \nabla v_h)_{\mathcal{T}_h} + \langle \beta \cdot \mathbf{n} w_h^-, v_h \rangle_{\partial \mathcal{T}_h} \\ &= -(\beta w_h, \nabla v_h)_{\mathcal{T}_h} + \langle \beta \cdot \mathbf{n} w_h^-, (v_h - v_h^-) \rangle_{\partial \mathcal{T}_h} \\ &= -(\beta w_h, \nabla v_h)_{\mathcal{T}_h} + \sum_{K \in \mathcal{T}_h} \langle \beta \cdot \mathbf{n} w_h^-, (v_h^+ - v_h^-) \rangle_{\partial K^-} \\ &= -(\beta w_h, \nabla v_h)_{\mathcal{T}_h} - \sum_{K \in \mathcal{T}_h} \langle \beta \cdot \mathbf{n} w_h^-, (v_h^+ - v_h^-) \rangle_{\partial K^+}, \end{aligned}$$

where we used the identity  $\langle \beta \cdot \mathbf{n} w_h^-, v_h^\pm \rangle_{\partial \mathcal{T}_h} = \sum_{K \in \mathcal{T}_h} \langle \beta \cdot \mathbf{n} w_h^-, v_h^\pm \rangle_{\partial K} = 0$  in the second and fourth equalities. Applying the Cauchy-Schwarz inequality for the above expression, we have

$$\begin{aligned} |B_c(w_h, v_h)| &\leq \beta_{\max} \|w_h\|_{\mathcal{T}_h} \|\nabla v_h\|_{\mathcal{T}_h} \\ &\quad + \beta_{\max} \left( \sum_{K \in \mathcal{T}_h} \frac{2}{C_{k,K}^2} \|w_h^-\|_{\partial K^+}^2 \right)^{1/2} \left( \sum_{K \in \mathcal{T}_h} \frac{C_{k,K}^2}{2} \llbracket v_h \rrbracket \|_{\partial K^+}^2 \right)^{1/2} \\ &\leq \beta_{\max} (\|w_h\|_{\mathcal{T}_h}^2 + \sum_{K \in \mathcal{T}_h} \frac{2}{C_{k,K}^2} \|w_h^-\|_{\partial K^+}^2)^{1/2} \\ &\quad \times (\|\nabla v_h\|_{\mathcal{T}_h}^2 + \sum_{K \in \mathcal{T}_h} \frac{C_{k,K}^2}{2} \llbracket v_h \rrbracket \|_{\partial K^+}^2)^{1/2} \\ &\leq \beta_{\max} \left( \|w_h\|_{\mathcal{T}_h}^2 + \sum_{K \in \mathcal{T}_h} 2 \|w_h\|_K^2 \right)^{1/2} \|v_h\|_{1, \mathcal{T}_h} \\ &= \sqrt{3} \beta_{\max} \|w_h\|_{\mathcal{T}_h} \|v_h\|_{1, \mathcal{T}_h}, \end{aligned}$$

where the third inequality comes from the inverse-trace estimate in Lemma 1.

The third inequality is proven along the same line as that for the second by first rewriting the upwinding DG operator (14) via integration by parts:

$$\begin{aligned} B_c(w_h, v_h) &= -(\boldsymbol{\beta} w_h, \nabla v_h)_{\mathcal{T}_h} + \langle \boldsymbol{\beta} \cdot \mathbf{n} w_h^-, v_h \rangle_{\partial \mathcal{T}_h} \\ &= (\boldsymbol{\beta} \nabla w_h, v_h)_{\mathcal{T}_h} + \langle \boldsymbol{\beta} \cdot \mathbf{n} (w_h^- - w_h), v_h \rangle_{\partial \mathcal{T}_h} \\ &= (\boldsymbol{\beta} \nabla w_h, v_h)_{\mathcal{T}_h} + \sum_{K \in \mathcal{T}_h} \langle \boldsymbol{\beta} \cdot \mathbf{n} (w_h^- - w_h^+), v_h^+ \rangle_{\partial K^-}, \end{aligned}$$

and then applying the Cauchy-Schwarz inequality to the resulting expression.

The fourth inequality follows from the proof of the second one. This completes the proof of Lemma 4.  $\square$

**2.4. The semi-discrete EDG/DG scheme.** Collecting the above two spatial discretizations, we arrive at the following semi-discrete EDG/DG scheme for the convection diffusion equation (1): for  $0 < t \leq T$ , find  $u_h \in V_h$  such that

$$(\partial_t u_h, v_h)_{\mathcal{T}_h} + B_d(u_h, v_h) + B_c(u_h, v_h) = 0$$

for any  $v_h \in V_h$  with initial condition  $u_h(0)$  being the  $L^2$ -projection of the initial data  $u_0$  onto  $V_h$ . The EDG diffusion operator  $B_d(\cdot, \cdot)$  is given in (10), and the DG convection operator  $B_c(\cdot, \cdot)$  is given in (14).

To further simplify our presentation for the analysis of the fully discrete scheme considered in the next section, we define the following linear operators  $\mathbf{B}_d, \mathbf{B}_c : V_h \rightarrow V_h$  associated with the above two bilinear forms:

$$(15a) \quad (\mathbf{B}_d(w_h), v_h)_{V_h \times V_h} := B_d(w_h, v_h), \quad \forall v_h \in V_h,$$

$$(15b) \quad (\mathbf{B}_c(w_h), v_h)_{V_h \times V_h} := B_c(w_h, v_h), \quad \forall v_h \in V_h.$$

Then, the above semi-discrete scheme is equivalent to the following linear operator form: for  $0 < t \leq T$ , find  $u_h \in V_h$  such that

$$(16) \quad \partial_t u_h + \mathbf{B}_d(u_h) + \mathbf{B}_c(u_h) = 0.$$

### 3. The IMEX RK fully discrete schemes

In this section, we apply IMEX Runge-Kutta time stepping for the semi-discrete scheme (16). We treat the diffusive operator implicitly, and the convective operator explicitly. For a detailed introduction to IMEX RK schemes, we refer the readers to [1, 3, 8].

We provide stability and error estimates of the resulting IMEX scheme, obtaining similar results as those in [10, 11, 12], where the authors considered LDG spatial discretization.

We remark that due to the explicit treatment of the DG convective operator, the resulting fully discrete scheme can be implemented with static condensation of the element-wise degrees of freedom ( $\mathbf{R}_h \times V_h$ ) so that the globally coupled degrees of freedom are only those on the mesh skeleton ( $M_h$ ). Moreover, the resulting linear system is symmetric positive definite and of smaller size than the IMEX-LDG scheme [10, 11, 12]. On the other hand, if the DG convective operator were to be treated implicitly, static condensation of the element-wise degrees of freedom ( $\mathbf{R}_h \times V_h$ ) would no longer be possible due to the coupling of neighboring element-wise degrees of freedom.

Let us now introduce a general  $s$ -stage IMEX RK time marching scheme for the semi-discrete equation (16). Denote  $\{t^n = n\tau\}_{n=0}^M$  be the uniform partition of the



time interval  $[0, T]$ , with time step  $\tau$  and  $T = M\tau$ . Given  $u_h^n \in V_h$ , where  $V_h$  is the DG finite element space given in (2b) with a fixed polynomial degree  $k \geq 1$ , we would like to find the numerical solution at the next time level  $t^{n+1} = t^n + \tau$  by applying the following  $s$ -stage IMEX RK time marching scheme:

$$(17a) \quad U_h^{(n,0)} = u_h^n,$$

$$(17b) \quad U_h^{(n,i)} = U_h^{(n,0)} - \tau \sum_{j=1}^i a_{i,j} \mathbf{B}_d(U_h^{(n,j)}) - \tau \sum_{j=1}^i \hat{a}_{i,j} \mathbf{B}_c(U_h^{(n,j-1)}), \quad 1 \leq i \leq s,$$

$$(17c) \quad u_h^{n+1} = U_h^{(n,0)} - \tau \sum_{i=1}^s b_i \mathbf{B}_d(U_h^{(n,i)}) - \tau \sum_{i=1}^{s+1} \hat{b}_i \mathbf{B}_c(U_h^{(n,i-1)}),$$

where  $U_h^{(n,i)} \in V_h$  denotes the intermediate stages for  $i = 1, \dots, s$ . Denote  $A = (a_{i,j})$ ,  $\hat{A} = (\hat{a}_{i,j}) \in \mathbb{R}^{s \times s}$ ,  $\mathbf{b}^\top = [b_1, \dots, b_s]$ ,  $\hat{\mathbf{b}}^\top = [\hat{b}_1, \dots, \hat{b}_{s+1}]$ , then we can express the general  $s$ -stage IMEX RK scheme as the following Butcher tableau

$$(18) \quad \begin{array}{c|cc} \mathbf{c} & A & \hat{A} \\ \hline & \mathbf{b}^\top & \hat{\mathbf{b}}^\top \end{array}$$

Here the vector  $\mathbf{c} \in \mathbb{R}^s$  is the row sum of the matrix  $A$  (or  $\hat{A}$ ), related to the intermediate stage time.

In the above tableau, the pair  $(A|\mathbf{b})$  determines an  $s$ -stage singly diagonally implicit Runge-Kutta (DIRK) method and  $(\hat{A}|\hat{\mathbf{b}})$  defines an  $(s+1)$ -stage ( $s$ -stage if  $\hat{b}_{s+1} = 0$ ) explicit Runge-Kutta method.

We consider three specific IMEX schemes proposed in [1], of order 1, 2, and 3, respectively, which are stiffly accurate meaning that  $a_{sj} = b_j$  and  $\hat{a}_{sj} = \hat{b}_j$  for  $j = 1, \dots, s$ , and  $\hat{b}_{s+1} = 0$ . We have  $u_h^{n+1} = U_h^{(n,s)}$  for these IMEX schemes.

The first-order IMEX scheme, ARS(1,1,1) takes the forward Euler discretization for the explicit part and the backward Euler discretization for the implicit part. The second-order scheme is the L-stable, two-stage, second-order DIRK scheme, ARS(2,2,2). And the third-order scheme is the L-stable, four-stage, third-order DIRK scheme, ARS(4,4,3). Here the first argument in ARS( $\cdot, \cdot, \cdot$ ) is the number of implicit RK stages, the second argument is the number of explicit RK stages, and the last argument is the order of the scheme.

In the following we present these three schemes in the form (18).

First order ARS(1,1,1):

$$(19) \quad \begin{array}{c|cc} 1 & 1 & 1 \\ \hline & 1 & 1 \quad 0 \end{array}$$

Second order ARS(2,2,2):

$$(20) \quad \begin{array}{c|ccc|cc} \gamma & \gamma & 0 & \gamma & 0 \\ 1 & 1-\gamma & \gamma & \delta & 1-\delta \\ \hline & 1-\gamma & \gamma & \delta & 1-\delta & 0 \end{array}$$

where  $\gamma = 1 - \frac{\sqrt{2}}{2}$  and  $\delta = 1 - \frac{1}{2\gamma}$ .

Third order ARS(4,4,3):

$$(21) \quad \begin{array}{c|cccc|cccc} 1/2 & 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 2/3 & 1/6 & 1/2 & 0 & 0 & 11/18 & 1/18 & 0 & 0 \\ 1/2 & -1/2 & 1/2 & 1/2 & 0 & 5/6 & -5/6 & 1/2 & 0 \\ 1 & 3/2 & -3/2 & 1/2 & 1/2 & 1/4 & 7/4 & 3/4 & -7/4 \\ \hline & 3/2 & -3/2 & 1/2 & 1/2 & 1/4 & 7/4 & 3/4 & -7/4 & 0 \end{array}$$

In the rest of this section, we consider the fully discrete IMEX-EDG scheme (17) with spatial discretization given in Section 2 and temporal discretization given by one of the above three stiffly-accurate IMEX RK schemes.

We prove that the resulting IMEX-EDG schemes are  $L^2$ -stable under a time stepping restriction  $\tau \leq C_t \epsilon / \beta_{\max}^2$  with a constant  $C_t$  independent of the mesh size  $h$  and polynomial degree  $k$ . We also derive the corresponding optimal  $L^2$ -error estimates.

Note that similar stability and error estimates were obtained in [10], where their stability  $C_t$  is independent of the mesh size but depends on the polynomial degree. We also note that our first and second order IMEX schemes were the same as the ones considered in [10], but the third order IMEX scheme is different. Therein, the three-stage (with four explicit stages) third-order LIRK3 scheme introduced in [3] was analyzed. LIRK3 has one less implicit stage than ARS(4,4,3), which is potentially computationally more attractive. Our energy argument also applies to LIRK3, but for simplicity of presentation, we leave out the details. We specifically mention that while the energy argument is quite standard for the first- and second-order IMEX schemes, special care is needed to obtain the stability result for the third-order scheme.

Note that [3] also introduced a fourth-order IMEX scheme, namely LIRK4 with five implicit stages and six explicit stages. We are not able to prove similar stability results with the energy argument as for the other lower order IMEX schemes mentioned above. However, our numerical results in the next section indicates similar stability result should still hold.

**3.1. Stability analysis.** We present our main stability result below. The rest of this subsection is devoted to its proof.

**Theorem 2.** *There exists a positive constant  $C_t$ , independent of the mesh size  $h$  and the polynomial degree  $k$ , such that if  $\tau \leq C_t \epsilon / \beta_{\max}^2$ , then the solution of the scheme (17) with ARS(1,1,1), ARS(2,2,2), or ARS(4,4,3) IMEX time-stepping satisfies*

$$\|u_h^{n+1}\|_{\Omega} \leq \|u_h^n\|_{\Omega}, \quad \forall n.$$

We collect some intermediate results that will be used to prove the above theorem.

**Lemma 5.** *We have, for any  $1 \leq i \leq s$ ,*

$$(22a) \quad U_h^{(n,i)} - U_h^{(n,i-1)} = -\tau B_d \left( \sum_{j=1}^i a_{i,j}^1 U_h^{(n,j)} \right) - \tau B_c \left( \sum_{j=1}^i \hat{a}_{i,j}^1 U_h^{(n,j-1)} \right),$$

where

$$a_{i,j}^1 = \begin{cases} a_{1,j} & \text{if } i = 1 \\ a_{i,j} - a_{i-1,j} & \text{if } i > 1 \end{cases}, \quad \hat{a}_{i,j}^1 = \begin{cases} \hat{a}_{1,j} & \text{if } i = 1 \\ \hat{a}_{i,j} - \hat{a}_{i-1,j} & \text{if } i > 1 \end{cases}.$$

Moreover, we have

(22b)

$$\begin{aligned} \|U_h^{(n,i)}\|_\Omega^2 - \|U_h^{(n,i-1)}\|_\Omega^2 &= -\|U_h^{(n,i)} - U_h^{(n,i-1)}\|_\Omega^2 - 2\tau B_d\left(\sum_{j=1}^i a_{i,j}^1 U_h^{(n,j)}, U_h^{(n,i)}\right) \\ &\quad - 2\tau B_c\left(\sum_{j=1}^i \hat{a}_{i,j}^1 U_h^{(n,j-1)}, U_h^{(n,i)}\right), \end{aligned}$$

(22c)

$$\begin{aligned} \|U_h^{(n,i)} - U_h^{(n,i-1)}\|_\Omega^2 &= -\tau B_d\left(\sum_{j=1}^i a_{i,j}^1 U_h^{(n,j)}, U_h^{(n,i)} - U_h^{(n,i-1)}\right) \\ &\quad - \tau B_c\left(\sum_{j=1}^i \hat{a}_{i,j}^1 U_h^{(n,j-1)}, U_h^{(n,i)} - U_h^{(n,i-1)}\right), \end{aligned}$$

*Proof.* The first equation is obtained by dividing the  $i$ th equation of (17b) by the  $(i-1)$ th equation. The second one is obtained by testing the first equation by  $U_h^{(n,i)}$  and the last one is by testing the first equation by  $U_h^{(n,i)} - U_h^{(n,i-1)}$ .  $\square$

The following corollary is a direct consequence of the above result.

**Corollary 1.** *We have*

$$(23a) \quad \|u_h^{n+1}\|_\Omega^2 - \|u_h^n\|_\Omega^2 = \Theta_t + \Theta_d + \Theta_c$$

where

$$(23b) \quad \Theta_t := -\sum_{i=1}^s \|U_h^{(n,i)} - U_h^{(n,i-1)}\|_\Omega^2$$

$$(23c) \quad \Theta_d := -\tau \sum_{i=1}^s B_d\left(\sum_{j=1}^s a_{i,j}^2 U_h^{(n,j)}, U_h^{(n,i)}\right)$$

$$(23d) \quad \Theta_c := -2 \sum_{i=1}^s \tau B_c\left(\sum_{j=1}^i \hat{a}_{i,j}^1 U_h^{(n,j-1)}, U_h^{(n,i)}\right),$$

and  $a_{i,j}^2 := a_{i,j}^1 + a_{j,i}^1$ .

*Proof.* Summing up (22b) over  $i$  from 1 to  $s$ , we obtain

$$\begin{aligned} \|U_h^{(n,s)}\|_\Omega^2 - \|U_h^{(n,0)}\|_\Omega^2 &= -\sum_{i=1}^s \|U_h^{(n,i)} - U_h^{(n,i-1)}\|_\Omega^2 - \tau \sum_{i=1}^s B_d\left(\sum_{j=1}^i a_{i,j}^1 U_h^{(n,j)}, U_h^{(n,i)}\right) \\ &\quad - 2 \sum_{i=1}^s \tau B_c\left(\sum_{j=1}^i \hat{a}_{i,j}^1 U_h^{(n,j-1)}, U_h^{(n,i)}\right). \end{aligned}$$

Using the symmetry of the diffusive bilinear form,  $B_d(w_h, v_h) = B_d(v_h, w_h)$ , and  $u_h^{n+1} = U_h^{(n,s)}$  and  $u_h^n = U_h^{(n,0)}$ , we obtain the desired equality.  $\square$

Denoting the matrix  $A^{(2)} := (a_{i,j}^2) \in \mathbb{R}^{s \times s}$ , we have

$$\begin{aligned}
 A^{(2)} &= 1 && \text{for ARS}(1,1,1), \\
 A^{(2)} &= \begin{pmatrix} 2 - \sqrt{2} & \sqrt{2} - 1 \\ \sqrt{2} - 1 & 2 - \sqrt{2} \end{pmatrix} && \text{for ARS}(2,2,2), \\
 A^{(2)} &= \begin{pmatrix} 1 & -1/3 & -2/3 & 2 \\ -1/3 & 1 & 0 & -2 \\ -2/3 & 0 & 1 & 0 \\ 2 & -2 & 0 & 1 \end{pmatrix} && \text{for ARS}(4,4,3).
 \end{aligned}$$

Note that for ARS(1,1,1) and ARS(2,2,2), the matrix  $A^{(2)}$  has positive eigenvalues, which guarantees the non-positivity of the right hand side (for any time step  $\tau$ ) of the energy identity (23a) in the absence of convection ( $\beta = 0$ ). Hence, unconditional stability in the absence of convection can be obtained. However, the matrix  $A^{(2)}$  for ARS(4,4,3) has one negative eigenvalue, our stability result can not be directly obtained from the above energy identity (23a). Similar difficulty appears for the third-order LIRK3 scheme analyzed in [10]. Indeed, we have to test the equation (22a) with some special test functions with the goal of perturbing the matrix  $A^{(2)}$  so that it become positive definite. We note that, however, we are unable to find suitable test functions for an energy argument to prove the unconditional stability of the fourth-order LIRK4 scheme proposed in [3] for the pure diffusion problem.

The modified energy identity for ARS(4,4,3) is given below.

**Corollary 2.** *For ARS(4,4,3), we have*

$$(24a) \quad \|u_h^{n+1}\|_{\Omega}^2 - \|u_h^n\|_{\Omega}^2 = \tilde{\Theta}_t + \tilde{\Theta}_d + \tilde{\Theta}_c$$

where

(24b)

$$\tilde{\Theta}_t := - \sum_{i=1}^s \|U_h^{(n,i)} - U_h^{(n,i-1)}\|_{\Omega}^2 - 18 \|U_h^{(n,2)} - U_h^{(n,1)}\|_{\Omega}^2$$

(24c)

$$\begin{aligned}
 \tilde{\Theta}_d &:= -\tau \sum_{i=1}^s B_d \left( \sum_{j=1}^s a_{i,j}^2 U_h^{(n,j)}, U_h^{(n,i)} \right) - 18\tau B_d \left( \sum_{j=1}^2 a_{2,j}^1 U_h^{(n,j)}, U_h^{(n,2)} - U_h^{(n,1)} \right) \\
 &= -\tau \sum_{i=1}^s B_d \left( \sum_{j=1}^s \hat{a}_{i,j}^2 U_h^{(n,j)}, U_h^{(n,i)} \right)
 \end{aligned}$$

(24d)

$$\begin{aligned}
 \tilde{\Theta}_c &:= -2 \sum_{i=1}^s \tau B_c \left( \sum_{j=1}^i \hat{a}_{i,j}^1 U_h^{(n,j-1)}, U_h^{(n,i)} \right) - 18\tau B_c \left( \sum_{j=1}^2 \hat{a}_{2,j}^1 U_h^{(n,j-1)}, U_h^{(n,2)} - U_h^{(n,1)} \right) \\
 &= -2 \sum_{i=1}^s \tau B_c \left( \sum_{j=1}^i \hat{a}_{i,j}^1 U_h^{(n,j-1)}, U_h^{(n,i)} \right) - \tau B_c (2U_h^{(n,0)} - U_h^{(n,1)}, U_h^{(n,2)} - U_h^{(n,1)}),
 \end{aligned}$$

and where  $\tilde{A}^{(2)} = (\tilde{a}_{i,j}^2) = \begin{pmatrix} 7 & -47/6 & -2/3 & 2 \\ -47/6 & 10 & 0 & -2 \\ -2/3 & 0 & 1 & 0 \\ 2 & -2 & 0 & 1 \end{pmatrix}$ .

*Proof.* The result is a direct consequence of the equality (22c) and (23a).  $\square$

Note that the matrix  $\tilde{A}^{(2)}$  above is positive definite, with minimal eigenvalue  $\approx 0.043$ .

**Corollary 3.** *We have*

$$\begin{aligned} \Theta_d &= -\tau \left\| \left\| U_h^{(n,1)} \right\| \right\|_{e, \mathcal{T}_h}^2 && \text{for ARS}(1,1,1), \\ \Theta_d &\leq -0.17\tau \sum_{i=1}^2 \left\| \left\| U_h^{(n,i)} \right\| \right\|_{e, \mathcal{T}_h}^2 && \text{for ARS}(2,2,2), \\ \tilde{\Theta}_d &\leq -0.043\tau \sum_{i=1}^4 \left\| \left\| U_h^{(n,i)} \right\| \right\|_{e, \mathcal{T}_h}^2 && \text{for ARS}(4,4,3). \end{aligned}$$

*Proof.* The equality for ARS(1,1,1) is trivial to verify.

To prove the other inequalities, we claim that for any symmetric matrix  $M = (m_{i,j}) \in \mathbb{R}^{s \times s}$  with minimal eigenvalue  $\lambda_{\min}$ , we have

$$\sum_{i=1}^s B_d \left( \sum_{i=1}^s m_{i,j} U_h^{(n,j)}, U_h^{(n,i)} \right) \geq \lambda_{\min} \sum_{i=1}^s B_d(U_h^{(n,i)}, U_h^{(n,i)}) = \lambda_{\min} \sum_{i=1}^s \left\| \left\| U_h^{(n,i)} \right\| \right\|_{e, \mathcal{T}_h}^2.$$

Then, applying this claim, the inequalities are obtained by a simple calculation of the minimal eigenvalues of the related matrices, which is larger than 0.17 for ARS(2,2,2), and larger than 0.043 for ARS(4,4,3).

Let us now prove the claim. Without loss of generality, we assume  $\lambda_{\min} = 0$ . Hence, all the eigenvalues of  $M$  are non-negative. Since  $M$  is symmetric, we have the eigenvalue decomposition  $M = Q\Lambda Q^T$  where  $Q = (q_{i,j}) \in \mathbb{R}^{s \times s}$  is an orthogonal matrix, and  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_s)$  is the diagonal matrix whose diagonal entries are the eigenvalues of  $M$ . Denoting  $V_h^i = \sum_{j=1}^s q_{j,i} U_h^{(n,j)}$ , then we have

$$\sum_{i=1}^s B_d \left( \sum_{i=1}^s m_{i,j} U_h^{(n,j)}, U_h^{(n,i)} \right) = \sum_{k=1}^s B_d(\lambda_k V_h^k, V_h^k) \geq 0$$

This completes the proof.  $\square$

Let us now show that the convective part  $\Theta_c$  in (23d) (and  $\tilde{\Theta}_c$  in (24d)) can be controlled by the energy norm of  $U_h^{(n,i)}$  and  $L^2$ -norm of stage-differences  $U_h^{(n,i)} - U_h^{(n,i-1)}$ .

**Lemma 6.** *There exists a positive constant  $C$ , depending only on the coefficient matrix  $\hat{A}$ , such that*

$$-\sum_{i=1}^s B_c \left( \sum_{j=1}^i \hat{a}_{i,j}^1 U_h^{(n,j-1)}, U_h^{(n,i)} \right) \leq C \frac{\beta_{\max}}{\sqrt{\epsilon}} \sum_{i=1}^s \|U_h^{(n,i)}\|_{e, \mathcal{T}_h} \sum_{j=1}^s \|U_h^{(n,j)} - U_h^{(n,j-1)}\|_{\Omega}$$

Moreover,

$$\begin{aligned} -B_c(2U_h^{(n,0)} - U_h^{(n,1)}, U_h^{(n,2)} - U_h^{(n,1)}) &\leq 2\sqrt{3} \frac{\beta_{\max}}{\sqrt{\epsilon}} \left( \left\| \left\| U_h^{(n,1)} \right\| \right\|_{e, \mathcal{T}_h} + \left\| \left\| U_h^{(n,2)} \right\| \right\|_{e, \mathcal{T}_h} \right) \\ &\quad (2\|U_h^{(n,1)} - U_h^{(n,0)}\|_{\Omega} + \|U_h^{(n,2)} - U_h^{(n,1)}\|_{\Omega}) \end{aligned}$$

*Proof.* We just give the proof for the first inequality since that for the second is similar.

Let  $\hat{a}_{i,j}^2 := \sum_{k=1}^j \hat{a}_{i,k}^1$ . Then, we have  $\hat{a}_{i,i}^2 = \sum_{k=1}^i \hat{a}_{i,k}^1 = c_i - c_{i-1}$  recalling  $c_i = \sum_{k=1}^i \hat{a}_{i,k}^1$  (we set  $c_0 = 0$  for notation consistency). Denote  $\hat{a}_{\max} := \max_{i,j} |\hat{a}_{i,j}^2|$ . We have  $\hat{a}_{\max} = 1$  for ARS(1,1,1) and ARS(2,2,2), and  $\hat{a}_{\max} = 9/4$  for ARS(4,4,3).

Then, we have

$$\begin{aligned} -B_c(\sum_{j=1}^i \hat{a}_{i,j}^1 U_h^{(n,j-1)}, U_h^{(n,i)}) &= -B_c(\sum_{j=1}^i \hat{a}_{i,j}^2 (U_h^{(n,j-1)} - U_h^{(n,j)}), U_h^{(n,i)}) \\ &\quad - \hat{a}_{i,i}^2 B_c(U_h^{(n,i)}, U_h^{(n,i)}) \\ &\leq 2\sqrt{3} \hat{a}_{\max} \frac{\beta_{\max}}{\sqrt{\epsilon}} \sum_{j=1}^i \|U_h^{(n,j)} - U_h^{(n,j-1)}\|_{\Omega} \left\| \|U_h^{(n,i)}\| \right\|_{e, \mathcal{T}_h} \\ &\quad - (c_i - c_{i-1}) B_c(U_h^{(n,i)}, U_h^{(n,i)}), \end{aligned}$$

where the above inequality comes from Lemma 4 and Lemma 3.

Summing up the above inequality over  $i$  from 1 to  $s$ , we get

$$\begin{aligned} -\sum_{i=1}^s B_c(\sum_{j=1}^i \hat{a}_{i,j}^1 U_h^{(n,j-1)}, U_h^{(n,i)}) &\leq 2\sqrt{3} \hat{a}_{\max} \frac{\beta_{\max}}{\sqrt{\epsilon}} \sum_{j=1}^s \|U_h^{(n,j)} - U_h^{(n,j-1)}\|_{\Omega} \sum_{i=1}^s \left\| \|U_h^{(n,i)}\| \right\|_{e, \mathcal{T}_h} \\ &\quad - \sum_{i=1}^s (c_i - c_{i-1}) B_c(U_h^{(n,i)}, U_h^{(n,i)}). \end{aligned}$$

We have  $s = 1$  and  $c_1 = 1$  for the first-order IMEX scheme ARS(1,1,1),  $s = 2$  and  $c_1 = \gamma, c_2 = 1$  for the second-order scheme ARS(2,2,2). For these two schemes,  $c_i$  is increasing, hence  $\sum_{i=1}^s (c_i - c_{i-1}) B_c(U_h^{(n,i)}, U_h^{(n,i)}) \geq 0$  by Lemma 4. So, the inequality in Lemma 6 holds with the constant  $C = 2\sqrt{3}$ .

We have  $s = 4$ , and  $c_1 = 1/2, c_2 = 2/3, c_3 = 1/2, c_4 = 1$  for the third-order IMEX scheme ARS(4,4,3). This implies that

$$\begin{aligned} -\sum_{i=1}^s (c_i - c_{i-1}) B_c(U_h^{(n,i)}, U_h^{(n,i)}) &= -\frac{1}{2} B_c(U_h^{(n,1)}, U_h^{(n,1)}) - \frac{1}{2} B_c(U_h^{(n,4)}, U_h^{(n,4)}) \\ &\quad - \frac{1}{6} (B_c(U_h^{(n,2)}, U_h^{(n,2)}) - B_c(U_h^{(n,3)}, U_h^{(n,3)})) \\ &\leq -\frac{1}{6} (B_c(U_h^{(n,2)}, U_h^{(n,3)}, U_h^{(n,2)}) - B_c(U_h^{(n,3)}, U_h^{(n,3)}, U_h^{(n,2)})) \\ &\leq \frac{1}{6} 2\sqrt{3} \frac{\beta_{\max}}{\sqrt{\epsilon}} \|U_h^{(n,2)} - U_h^{(n,3)}\|_{\Omega} (\left\| \|U_h^{(n,2)}\| \right\|_{e, \mathcal{T}_h} + \left\| \|U_h^{(n,3)}\| \right\|_{e, \mathcal{T}_h}). \end{aligned}$$

Hence, the inequality in Lemma 6 holds with the constant  $C = 2\sqrt{3}(\hat{a}_{\max} + 1/6) = 29\sqrt{3}/6$ .  $\square$

Now, we are ready to prove Theorem 2.

### Proof of Theorem 2.

*Proof.* Combining results in Corollary 1-3 and Lemma 6, we have

$$\begin{aligned} \|u_h^{n+1}\|_{\Omega}^2 - \|u_h^n\|_{\Omega}^2 &\leq -\sum_{i=1}^s \|U_h^{(n,i)} - U_h^{(n,i-1)}\|_{\Omega}^2 - C_1 \tau \sum_{i=1}^s \left\| \|U_h^{(n,i)}\| \right\|_{e, \mathcal{T}_h}^2 \\ &\quad + C_2 \tau \frac{\beta_{\max}}{\sqrt{\epsilon}} \left( \sum_{i=1}^s \|U_h^{(n,i)} - U_h^{(n,i-1)}\|_{\Omega} \right) \left( \sum_{i=1}^s \left\| \|U_h^{(n,i)}\| \right\|_{e, \mathcal{T}_h} \right), \end{aligned}$$

where  $C_1$  and  $C_2$  are two positive constants depending on the IMEX time discretization. We can take  $C_1 = 1$ ,  $C_2 = 2\sqrt{3}$  for ARS(1,1,1),  $C_1 = 0.17$ ,  $C_2 = 2\sqrt{3}$  for ARS(2,2,2), and  $C_1 = 0.043$ ,  $C_2 = 9\sqrt{3}$  for ARS(4,4,3).

Then, applying the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} \|u_h^{n+1}\|_\Omega^2 - \|u_h^n\|_\Omega^2 &\leq -1/2 \sum_{i=1}^s \|U_h^{(n,i)} - U_h^{(n,i-1)}\|_\Omega^2 \\ &\quad + (C_2^2 s^2 \tau^2 \frac{\beta_{\max}^2}{2\epsilon} - C_1 \tau) \sum_{i=1}^s \left\| \|U_h^{(n,i)}\|_{e, \mathcal{T}_h} \right\|^2 \end{aligned}$$

Taking  $\tau \leq \frac{2C_1}{C_2^2 s^2} \epsilon / \beta_{\max}^2$ , we have

$$\|u_h^{n+1}\|_\Omega^2 - \|u_h^n\|_\Omega^2 \leq 0.$$

This completes the proof. □

We remark that the bounding constant in the above proof is not sharp. We will numerically examine the constant in one dimension in the next section.

**3.2. Error estimates.** Now, combining the stability results of Theorem 2 and approximation property of the elliptic projection in Theorem 1, we are ready to obtain optimal  $L^2$ -error estimates for smooth solutions. From now on, we denote  $C > 0$  as a generic constant that is independent of the mesh size  $h$  and time step length  $\tau$ , but depends on the quasi-uniformity and shape-regularity of the mesh, the convection and diffusion coefficients, the final time, and the smoothness of the exact solution. We assume the solution is smooth enough so that the Sobolev norms hidden in the constant  $C$  are bounded. The constant  $C$  may vary in different locations.

Following [14, 10], we introduce stage reference functions (discretize in time only), denoted by  $\{U^{(n,\ell)}\}_{\ell=0}^s$  for  $n = 0, \dots, M$ , associated with the  $s$ -stage IMEX RK time discretization (17). Recall that  $M$  is the number of total steps with  $\tau M = T$ . In detail, for  $0 \leq n \leq M$ ,  $U^{(n,0)} = u(t^n)$  and

(25)

$$U^{(n,i)} = U^{(n,0)} - \tau \sum_{j=1}^i a_{i,j} \nabla \cdot (-\epsilon \nabla U^{(n,j)}) - \tau \sum_{j=1}^i \hat{a}_{i,j} \nabla \cdot (\beta U^{(n,j-1)}), \quad 1 \leq i \leq s$$

We first derive the error bound for the final stage reference solution and the (smooth) exact solution.

**Lemma 7.** *We have*

$$(26) \quad \|U^{(n,s)} - u(t^{n+1})\|_\Omega \leq C\tau^{r+1},$$

with  $r = 1$  for ARS(1,1,1),  $r = 2$  for ARS(2,2,2), and  $r = 3$  for ARS(4,4,3).

*Proof.* By the considered PDE (1) and the definitions of the reference functions (25), it is easy to show that

$$u(t^{n+1}) = U^{(n,0)} - \tau \sum_{j=1}^s a_{s,j} \nabla \cdot (-\epsilon \nabla U^{(n,j)}) - \tau \sum_{j=1}^s \hat{a}_{s,j} \nabla \cdot (\beta U^{(n,j-1)}) + \zeta^n,$$

where  $\zeta^n$  is the local truncation error in each step that satisfies

$$\|U^{(n,s)} - u(t^{n+1})\|_\Omega = \|\zeta^n\|_\Omega \leq C\tau^{r+1}.$$

Similar analysis can be found in [14].  $\square$

We denote  $\pi_V U^{(n,j)} \in V_h$  as the following the elliptic projection:

$$(27) \quad B_d(\pi_V U^{(n,j)}, v_h) = (\nabla \cdot (-\epsilon \nabla U^{(n,j)}), v_h)_{\mathcal{T}_h} \quad \forall v_h \in V_h.$$

Theorem 1 implies the following optimal approximation property:

$$(28) \quad \|\pi_V U^{(n,j)} - U^{(n,j)}\|_{\mathcal{T}_h} + h^{1/2} \|\pi_V U^{(n,j)} - U^{(n,j)}\|_{\partial \mathcal{T}_h} \leq C h^{k+1}.$$

At each stage time, we denote the error between the exact (reference) solution and the numerical solution by  $\varepsilon^{(n,\ell)} = U^{(n,\ell)} - U_h^{(n,\ell)}$ . As the standard treatment in finite element analysis, we would like to divide the error in the form  $\varepsilon^{(n,\ell)} = \xi^{(n,\ell)} - \eta^{(n,\ell)}$ , where

$$(29) \quad \eta^{(n,\ell)} = \pi_V U^{(n,\ell)} - U^{(n,\ell)}, \quad \xi^{(n,\ell)} = \pi_V U^{(n,\ell)} - U_h^{(n,\ell)}.$$

Now, we derive the error equation for  $\xi^{(n,\ell)} \in V_h$ .

**Lemma 8.** *We have, for  $1 \leq i \leq s$ ,*

$$(30) \quad \begin{aligned} \xi^{(n,i)} &= \xi^{(n,0)} - \tau \sum_{j=1}^i a_{i,j} B_d(\xi^{(n,j)}) - \tau \sum_{j=1}^i \hat{a}_{i,j} B_c(\xi^{(n,j-1)}) \\ &\quad + \eta^{(n,s)} - \eta^{(n,0)} + \tau \sum_{j=1}^i \hat{a}_{i,j} (B_c(\pi_V U^{(n,j-1)}) - \nabla \cdot (\beta U^{(n,j-1)})), \end{aligned}$$

*Proof.* The equality is obtained by combining the numerical scheme (17b), the reference solution equation (25) and definition of the elliptic projection (27).  $\square$

Now, we can obtain the energy identity for the error  $\xi$  similar as that in Corollary 1.

**Lemma 9.** *We have*

$$(31a) \quad \|\xi^{(n,s)}\|_{\Omega}^2 - \|\xi^{(n,0)}\|_{\Omega}^2 = \Theta_t^e + \Theta_d^e + \Theta_c^e + \Theta_p^e$$

where

$$(31b) \quad \Theta_t^e := - \sum_{i=1}^s \|\xi^{(n,i)} - \xi^{(n,i-1)}\|_{\Omega}^2$$

$$(31c) \quad \Theta_d^e := - \tau \sum_{i=1}^s B_d(\sum_{j=1}^s a_{i,j}^2 \xi^{(n,j)}, \xi^{(n,i)})$$

$$(31d) \quad \Theta_c^e := - 2 \sum_{i=1}^s \tau B_c(\sum_{j=1}^i \hat{a}_{i,j}^1 \xi^{(n,j-1)}, \xi^{(n,i)}),$$

$$(31e) \quad \Theta_p^e := 2 \sum_{i=1}^s (\eta^{(n,j)} - \eta^{(n,j-1)}, \xi^{(n,i)})_{\Omega} + 2 \sum_{i=1}^s \tau B_c(\sum_{j=1}^i \hat{a}_{i,j}^1 \eta^{(n,j-1)}, \xi^{(n,i)}).$$

Moreover,

$$(32) \quad \Theta_p^e \leq C \tau h^{k+1} \sum_{i=1}^s \|\xi^{(n,i)}\|_{\Omega} + C \tau h^{k+1} \sum_{i=1}^s \left\| \left\| \xi^{(n,i)} \right\| \right\|_{e, \mathcal{T}_h}.$$



*Proof.* The proof for the equality follows from that for Corollary 1, noting that

$$\begin{aligned} B_c(\eta^{(n,j)}, \xi^{(n,i)}) &= B_c(\pi_V U^{(n,j)} - U^{(n,j)}, \xi^{(n,i)}) \\ &= B_c(\pi_V U^{(n,j)}, \xi^{(n,i)}) - (\nabla \cdot (\beta U^{(n,j)}), \xi^{(n,i)})_\Omega. \end{aligned}$$

The estimate (32) is obtained by the Cauchy-Schwarz inequality along with the following approximation estimates

$$\|\eta^{(n,i)} - \eta^{(n,i-1)}\|_\Omega \leq C\tau h^{k+1}, \quad \|\eta^{(n,i)}\|_\Omega + h^{1/2}\|\eta^{(n,i)}\|_{\partial\mathcal{T}_h} \leq C\tau h^{k+1}.$$

□

Following the proof of the stability result in Theorem 2, we are ready to obtain our main result on the  $L^2$ -error estimates.

**Theorem 3.** *Let  $u_h^n$ ,  $n = 1, \dots, M$  be the solution of the fully discrete scheme (17) with  $ARS(1,1,1)$ ,  $ARS(2,2,2)$ , or  $ARS(4,4,3)$  IMEX time stepping and initial condition  $u_h^0$  the  $L^2$ -projection onto  $V_h$  of  $u_0$ .*

*Then, there exists a constant  $C_t$ , independent of the mesh size  $h$  and the polynomial degree  $k$ , such that if  $\tau \leq C_t \epsilon / \beta_{\max}^2$ , the following error estimate holds:*

$$\max_{n=1, \dots, M} \|u_h^n - u(t^n)\|_\Omega \leq C(h^{k+1} + \tau^r),$$

where  $u(t^n)$  is the smooth exact solution at time  $t^n$ , and  $r = 1$  for  $ARS(1,1,1)$ ,  $r = 2$  for  $ARS(2,2,2)$ , and  $r = 3$  for  $ARS(4,4,3)$ .

*Proof.* The proof follows from that for the stability result in Theorem 2. Here we just sketch its main steps.

First, from Lemma 9, we can show (by adapting the proof for Theorem 2) there exist a constant  $C_t > 0$  such that if  $\tau \leq C_t \epsilon / \beta_{\max}^2$ , then

$$\begin{aligned} \|\xi^{(n+1,0)}\|_\Omega^2 - \|\xi^{(n,0)}\|_\Omega^2 &\leq \|\xi^{(n+1,0)}\|_\Omega^2 - \|\xi^{(n,s)}\|_\Omega^2 \\ &\quad - 1/2 \sum_{i=1}^s \|\xi^{(n,i)} - \xi^{(n,i-1)}\|_\Omega^2 - C_1 \tau/2 \sum_{i=1}^s \left\| \xi^{(n,i)} \right\|_{e, \mathcal{T}_h}^2 \\ &\quad + C\tau h^{k+1} \sum_{i=1}^s \|\xi^{(n,i)}\|_\Omega + C\tau h^{k+1} \sum_{i=1}^s \left\| \xi^{(n,i)} \right\|_{e, \mathcal{T}_h} \end{aligned}$$

where  $C_1 > 0$  is a constant only depends on the IMEX RK coefficients. Recall that special test function needs to be used to prove this estimate for the third-order  $ARS(4,4,3)$  scheme.

We have

$$\begin{aligned} \|\xi^{(n+1,0)}\|_\Omega^2 - \|\xi^{(n,s)}\|_\Omega^2 &\leq \|\xi^{(n+1,0)} - \xi^{(n,s)}\|_\Omega (\|\xi^{(n+1,0)} - \xi^{(n,s)}\|_\Omega + 2\|\xi^{(n,s)}\|_\Omega) \\ &\leq C\tau^{r+1}(\tau^{r+1} + \|\xi^{(n,s)}\|_\Omega), \end{aligned}$$

where Lemma 7 is used in the last inequality.

Next, by the triangular inequality  $\|\xi^{(n,\ell)}\|_\Omega \leq \|\xi^{(n,0)}\|_\Omega + \sum_{i=1}^\ell \|\xi^{(n,i)} - \xi^{(n,i-1)}\|_\Omega$ , and the Cauchy-Schwarz inequality, we get

$$\begin{aligned} \|\xi^{(n+1,0)}\|_\Omega^2 - \|\xi^{(n,0)}\|_\Omega^2 &\leq C\tau(h^{k+1} + \tau^r)\|\xi^{(n,0)}\|_\Omega + C\tau(h^{2(k+1)} + \tau^{2r+1}) \\ &\leq \tau\|\xi^{(n,0)}\|_\Omega^2 + C\tau(h^{2(k+1)} + \tau^{2r}). \end{aligned}$$

Hence, for any  $1 \leq n \leq M$ ,

$$\begin{aligned} \|\xi^{(n,0)}\|_{\Omega}^2 &\leq (1 + \tau)^n \|\xi^{(0,0)}\|_{\Omega}^2 + C\tau n (h^{2(k+1)} + \tau^{2r}) \\ &\leq C(\exp(T) + T) h^{2(k+1)} + \tau^{2r}. \end{aligned}$$

Finally, by a triangular inequality, we have, for any  $0 \leq n \leq M$ ,

$$\|u_h^n - u(t^n)\|_{\Omega} \leq \|\xi^{(n,0)}\|_{\Omega} + \|\eta^{(n,0)}\|_{\Omega} \leq C(h^{k+1} + \tau^r).$$

This completes the sketch of the proof. □

#### 4. Numerical experiments

In this section, we first numerically compute the stability constant  $C_t$  for the IMEX EDG schemes (17) presented in the previous section in  $1d$  such that for time step  $\tau \leq C_t\epsilon/\beta_{\max}^2$ , we have the decrease of  $L^2$ -norm in each Runge-Kutta time step. Specifically, we consider the three stiffly-accurate schemes ARS(1,1,1), ARS(2,2,2), and ARS(4,4,3), and the third-order LIRK3 scheme (with tuning variable  $\alpha = -0.35$ ) and fourth-order LIRK4 scheme.

We then perform numerical accuracy tests for the last four IMEX schemes.

**4.1. The stability constant  $C_t$ .** Here we compute the stability constant  $C_t$  in  $1d$  such that for  $\tau \leq C_t\epsilon/\beta_{\max}^2$ , we have the decrease of  $L^2$ -norm from the previous solution  $u_h^n$  to the next solution  $u_h^{n+1}$ . By linearity of the IMEX scheme (17), we can write the scheme as  $u_h^{n+1} = Lu_h^n$  for a square matrix  $L$ . The decrease of the  $L^2$ -norm is then equivalent to the non-negativity of the matrix  $M - L^T ML$ , where  $M$  is the mass matrix for  $V_h$ .

We take the computational domain to be of size  $2\pi$ , use a sequence of uniform mesh with  $2^N$  elements with  $N = 1, 2, \dots, 7$ . We fix the time step  $\tau = 1$ , and vary the diffusion coefficient  $\epsilon = \epsilon_{\alpha} = 0.01 \times 4^{-3+\alpha}$  for  $\alpha = 1, 2, \dots, 7$ . For each value of  $\epsilon_{\alpha}$ , we compute the largest value of  $\beta_{\alpha}$  on the sequences of meshes that makes sure the smallest eigenvalue of  $M - L^T ML$  is not negative. We then obtain

$$C_t = \min_{\alpha \in \{1, \dots, 7\}} \{\tau\beta_{\alpha}^2/\epsilon_{\alpha}\}.$$

The obtained constant  $C_t$ , up to 2 digits accuracy, for various polynomial degree, and IMEX schemes are listed in Table 1. For all the IMEX schemes, it is clear that the polynomial degree  $k$  does not have a significant influence on the constant  $C_t$ . However, we remark again that we are not able to prove this stability result for the fourth-order LIRK4 scheme.

TABLE 1. The stability constant  $C_t$ , up to 2 digits accuracy, that ensure the  $L^2$ -norm decrease with time step  $\tau \leq C_t\epsilon/\beta^2$ .

	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$
ARS(1,1,1)	1.5	1.4	1.3	1.3	1.2	1.2	1.2	1.2
ARS(2,2,2)	0.78	0.71	0.64	0.72	0.61	0.69	0.60	0.60
ARS(4,4,3)	3.7	3.8	3.7	3.8	3.5	3.8	3.6	3.6
LIRK3( $\alpha = -0.35$ )	3.1	3.6	3.7	3.5	3.4	3.1	3.1	3.2
LIRK4	1.3	1.0	0.94	0.80	0.82	0.76	0.73	0.76

TABLE 2. Errors and orders of accuracy for the convection-diffusion problem.

<b>Example 1.</b>		$\epsilon = 1$		$\epsilon = 0.1$		$\epsilon = 0.01$	
scheme	$(nx, ny)$	$L^2$ error	order	$L^2$ error	order	$L^2$ error	order
$\mathcal{P}_1$ ARS(2,2,2)	(10,10)	1.77E-01	-	1.86E-01	-	1.72E-01	-
	(20,20)	4.76E-02	1.90	4.07E-02	2.19	3.66E-02	2.23
	(40,40)	1.21E-02	1.98	9.88E-03	2.04	8.31E-03	2.14
	(80,80)	3.03E-03	1.99	2.45E-03	2.01	1.99E-03	2.06
	(160,160)	7.62E-04	1.99	6.11E-04	2.00	4.89E-04	2.02
$\mathcal{P}_2$ ARS(4,4,3)	(10,10)	3.99E-03	-	1.72E-02	-	2.23E-02	-
	(20,20)	4.18E-04	3.29	2.31E-03	2.90	3.90E-03	2.52
	(40,40)	4.84E-05	3.11	2.86E-04	3.02	5.34E-04	2.87
	(80,80)	5.89E-06	3.04	3.56E-05	3.01	5.56E-05	3.26
	(160,160)	7.36E-07	3.00	4.44E-06	3.00	5.86E-06	3.25
$\mathcal{P}_2$ LIRK3 $\alpha = -0.35$	(10,10)	3.98E-03	-	1.69E-02	-	2.23E-02	-
	(20,20)	4.20E-04	3.25	2.29E-03	2.88	3.89E-03	2.52
	(40,40)	4.87E-05	3.11	2.85E-04	3.01	5.33E-04	2.87
	(80,80)	5.93E-06	3.04	3.55E-05	3.01	5.56E-05	3.26
	(160,160)	7.40E-07	3.00	4.43E-06	3.00	5.86E-06	3.25
$\mathcal{P}_3$ LIRK4	(10,10)	1.64E-04	-	9.65E-04	-	1.08E-03	-
	(20,20)	9.31E-06	4.12	5.57E-05	4.12	5.94E-05	4.19
	(40,40)	5.67E-07	4.04	3.44E-06	4.02	3.54E-06	4.07
	(80,80)	3.52E-08	4.01	2.14E-07	4.00	2.32E-07	3.93
	(160,160)	2.21E-09	3.99	1.34E-08	4.00	1.55E-08	3.91
<b>Example 2.</b>		$\epsilon = 1$		$\epsilon = 0.1$		$\epsilon = 0.01$	
scheme	$(nx, ny)$	$L^2$ error	order	$L^2$ error	order	$L^2$ error	order
$\mathcal{P}_1$ ARS(2,2,2)	(10,10)	1.73E-01	-	1.81E-01	-	1.87E-01	-
	(20,20)	4.62E-02	1.90	4.03E-02	2.17	3.66E-02	2.35
	(40,40)	1.17E-02	1.98	9.79E-03	2.04	8.15E-03	2.17
	(80,80)	2.94E-03	1.99	2.43E-03	2.01	1.95E-03	2.06
	(160,160)	7.39E-04	1.99	6.06E-04	2.00	4.81E-04	2.02
$\mathcal{P}_2$ ARS(4,4,3)	(10,10)	3.86E-03	-	1.58E-02	-	2.45E-02	-
	(20,20)	4.09E-04	3.24	2.20E-03	2.85	3.56E-03	2.78
	(40,40)	4.79E-05	3.11	2.81E-04	2.97	4.18E-04	3.09
	(80,80)	5.86E-06	3.03	3.53E-05	2.99	4.63E-05	3.17
	(160,160)	7.33E-07	3.00	4.43E-06	3.00	5.45E-06	3.09
$\mathcal{P}_2$ LIRK3 $\alpha = -0.35$	(10,10)	3.87E-03	-	1.58E-02	-	2.45E-02	-
	(20,20)	4.10E-04	3.24	2.20E-03	2.85	3.56E-03	2.78
	(40,40)	4.79E-05	3.10	2.81E-04	2.97	4.18E-04	3.09
	(80,80)	5.86E-06	3.03	3.53E-05	2.99	4.63E-05	3.17
	(160,160)	7.33E-07	3.00	4.43E-06	3.00	5.45E-06	3.09
$\mathcal{P}_3$ LIRK4	(10,10)	1.60E-04	-	9.63E-04	-	1.22E-03	-
	(20,20)	9.24E-06	4.11	5.59E-05	4.11	5.99E-05	4.34
	(40,40)	5.65E-07	4.03	3.44E-06	4.02	3.58E-06	4.06
	(80,80)	3.51E-08	4.01	2.14E-07	4.00	2.37E-07	3.92
	(160,160)	2.21E-09	3.99	1.34E-08	4.00	1.56E-08	3.92

**4.2. Accuracy tests in 2d.** Now, we perform accuracy tests in 2d. We consider the same examples used in [12], namely a linear convection-diffusion equation and a nonlinear viscous Burgers equation with smooth exact solutions.

*Example 1. (Linear convection diffusion equation)*

$$\begin{cases} u_t + u_x + u_y - \epsilon(u_{xx} + u_{yy}) = 0, \\ u(x, y, 0) = \sin(x + y), \end{cases}$$

on  $(x, y) \in [-\pi, \pi] \times [-\pi, \pi]$ . The exact solution is

$$u(x, y, t) = \exp(-2\epsilon t) \sin(x + y - 2t).$$

*Example 2. (Viscous Burgers equation)*

$$\begin{cases} u_t + (u^2/2)_x + (u^2/2)_y - \epsilon(u_{xx} + u_{yy}) = f(x, y, t), \\ u(x, y, 0) = \sin(x + y), \end{cases}$$

on  $(x, y) \in [-\pi, \pi] \times [-\pi, \pi]$ , where  $f(x, y, t) = \exp(-4\epsilon t) \sin(2(x + y))$ . The exact solution is

$$u(x, y, t) = \exp(-2\epsilon t) \sin(x + y).$$

We test four IMEX schemes ARS(2,2,2), ARS(4,4,3), LIRK3, and LIRK4. For the second-order ARS(2,2,2) scheme, we take the polynomial degree  $k = 1$ , for the third-order ARS(4,4,3) and LIRK3 schemes, we take  $k = 2$ , and for the fourth-order LIRK4 scheme, we take  $k = 3$ .

In all the numerical tests, we use a sequence of uniform triangular meshes obtained by first obtaining a uniform  $nx \times ny$  rectangular mesh then cutting each rectangle into two triangles in the northwest direction, the final time is  $T = 1$  and the time step  $\tau = 0.1h$ , where  $h = \min\{2\pi/nx, 2\pi/ny\}$ .

In Table 2, we list the  $L^2$ -errors and orders of convergence for the four IMEX EDG schemes for solving the two examples. We can clearly observe optimal orders of convergence. Again, we note that we do not have a convergence proof for the LIRK4 scheme.

## 5. Conclusions

We considered several specific implicit-explicit Runge-Kutta time marching methods for solving linear convection-diffusion problems with periodic boundary conditions. In these methods the diffusion term was treated implicitly with an EDG scheme and the convection term explicitly with an upwinding DG scheme.

We prove stability of the resulting IMEX schemes under the time step restriction  $\tau \leq \tau_0$ , where the constant  $\tau_0$  only depends on the convection and diffusion coefficients, and is independent of the mesh size  $h$  and polynomial degree  $k$ . We also showed optimal error estimates in both space and time, under the same temporal condition  $\tau \leq \tau_0$ . The stability analysis and error estimates can be extended to convection-diffusion problems with a nonlinear convection part as done in [11, 12].

## Acknowledgments

This research is supported by DOE grant DE-FG02-08ER25863 and NSF grant DMS-1418750.

## References

- [1] U. M. Ascher, S. J. Ruuth, and R. J. Spiteri, Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations, *Appl. Numer. Math.*, 25 (1997), pp. 151–167. Special issue on time integration (Amsterdam, 1996).
- [2] I. Babuška and M. Suri, The  $h$ - $p$  version of the finite element method with quasi-uniform meshes, *RAIRO Modél. Math. Anal. Numér.*, 21 (1987), pp. 199–238.
- [3] M. P. Calvo, J. de Frutos, and J. Novo, Linearly implicit Runge-Kutta methods for advection-reaction-diffusion equations, *Appl. Numer. Math.*, 37 (2001), pp. 535–549.
- [4] A. Chernov, Optimal convergence estimates for the trace of the polynomial  $L^2$ -projection operator on a simplex, *Math. Comp.*, 81 (2012), pp. 765–787.
- [5] B. Cockburn, J. Gopalakrishnan, and R. Lazarov, Unified hybridization of discontinuous Galerkin, mixed and continuous Galerkin methods for second order elliptic problems, *SIAM J. Numer. Anal.*, 47 (2009), pp. 1319–1365.
- [6] B. Cockburn, J. Guzmán, S.-C. Soon, and H. K. Stolarski, An analysis of the embedded discontinuous Galerkin method for second-order elliptic problems, *SIAM J. Numer. Anal.*, 47 (2009), pp. 2686–2707.
- [7] S. Güzey, B. Cockburn, and H. K. Stolarski, The embedded discontinuous Galerkin method: application to linear shell problems, *Internat. J. Numer. Methods Engrg.*, 70 (2007), pp. 757–790.
- [8] C. A. Kennedy and M. H. Carpenter, Additive Runge-Kutta schemes for convection-diffusion-reaction equations, *Appl. Numer. Math.*, 44 (2003), pp. 139–181.
- [9] W. H. Reed and T. R. Hill, Triangular mesh methods for the neutron transport equation, Tech. Report LA-UR-73-479, Los Alamos Scientific Laboratory, Los Alamos, NM, (1973).
- [10] H. Wang, C.-W. Shu, and Q. Zhang, Stability and error estimates of local discontinuous Galerkin methods with implicit-explicit time-marching for advection-diffusion problems, *SIAM J. Numer. Anal.*, 53 (2015), pp. 206–227.
- [11] H. Wang, C.-W. Shu, and Q. Zhang, Stability analysis and error estimates of local discontinuous Galerkin methods with implicit-explicit time-marching for nonlinear convection-diffusion problems, *Appl. Math. Comput.*, 272 (2016), pp. 237–258.
- [12] H. Wang, S. Wang, Q. Zhang, and C.-W. Shu, Local discontinuous Galerkin methods with implicit-explicit time-marching for multi-dimensional convection-diffusion problems, *ESAIM Math. Model. Numer. Anal.*, 50 (2016), pp. 1083–1105.
- [13] T. Warburton and J. S. Hesthaven, On the constants in  $hp$ -finite element trace inverse inequalities, *Comput. Methods Appl. Mech. Engrg.*, 192 (2003), pp. 2765–2773.
- [14] Q. Zhang and C.-W. Shu, Error estimates to smooth solution of Runge-Kutta discontinuous Galerkin methods for scalar conservation laws, *SIAM. J. Numer. Anal.*, 42 (2004), pp. 641–666.

## Appendix A

In this appendix, we prove Theorem 1.

We directly prove the  $L^2$ -error estimate in Theorem 1. Uniqueness of the numerical solution is a simple byproduct of the proof of this estimate.

We use a projection-based energy argument similar to the analysis [6], with the only difference being the choice of the projections. With our choice of the projection, optimal  $hp$ -estimate can be derived. We estimate the difference of the numerical solution  $(\mathbf{q}_h, u_h, \widehat{u}_h) \in \mathbf{R}_h \times V_h \times M_h$  of the EDG scheme (6) (with  $f$  replaced by  $-\nabla \cdot (\epsilon \nabla u)$  and  $g$  replaced by  $u$ ) with an appropriate projection of the data, denoted as  $(\mathbf{\Pi}_R(-\epsilon \nabla u), \Pi_V u, \Pi_M u) \in \mathbf{R}_h \times V_h \times M_h$ . Here we take  $\mathbf{\Pi}_R(-\epsilon \nabla u)$  as the  $L^2$ -projection onto  $\mathbf{R}_h$  of  $-\epsilon \nabla u$ , and  $\Pi_V u = u_h^{CG}$ ,  $\Pi_M u = u_h^{CG}|_{\mathcal{E}_h}$ , where  $u_h^{CG}$  is the elliptic projection onto the following continuous Galerkin space  $V_h \cap H^1(\Omega)$ : find  $u_h^{CG} \in V_h \cap H^1(\Omega)$  such that

$$\begin{aligned} (\epsilon \nabla u_h^{CG}, \nabla v_h)_{\mathcal{T}_h} &= (\epsilon \nabla u, \nabla v_h)_{\mathcal{T}_h}, \\ (u_h^{CG}, 1)_{\mathcal{T}_h} &= (u, 1)_{\mathcal{T}_h}. \end{aligned}$$

Note that  $u_h^{CG} \in V_h \cap H^1(\Omega)$  implies that  $\Pi_M u = u_h^{CG}|_{\mathcal{E}_h} \in M_h$ .

To simplify notation, we denote

$$\begin{aligned} \epsilon_q &:= \mathbf{\Pi}_R(-\epsilon \nabla u) - \mathbf{q}_h, & \epsilon_u &:= \Pi_V u - u_h, & \widehat{\epsilon}_u &:= \Pi_M u - \widehat{u}_h, \\ \delta_q &:= \mathbf{\Pi}_R(-\epsilon \nabla u) - (-\epsilon \nabla u), & \delta_u &:= \Pi_V u - u, & \widehat{\delta}_u &:= \Pi_M u - u|_{\mathcal{E}_h}. \end{aligned}$$

Note that  $\delta_u|_{\mathcal{E}_h} = \widehat{\delta}_u$  by the definition of the projections.

It is well-known that the  $L^2$ -projection  $\mathbf{\Pi}_R(-\epsilon \nabla u)$  and the elliptic projection  $\Pi_V u$  have the following optimal  $hp$ -error estimates on quasi-uniform meshes, see [2],

$$(A.1a) \quad \|\delta_q\|_{\Omega} \leq C \epsilon \left(\frac{h}{k}\right)^k \|u\|_{H^{k+1}(\Omega)}$$

$$(A.1b) \quad \|\delta_u\|_{H^j(\Omega)} \leq C \left(\frac{h}{k}\right)^{k+1-j} \|u\|_{H^{k+1}(\Omega)}$$

Moreover, we have the following optimal  $hp$ -error estimate for the trace norm of the  $L^2$ -projection  $\mathbf{\Pi}_R(-\epsilon \nabla u)$ , see [4],

$$(A.1c) \quad \|\delta_q\|_{\partial \mathcal{T}_h} \leq C \epsilon \left(\frac{h}{k}\right)^{k-1/2} \|u\|_{H^{k+1}(\Omega)}.$$

The constant  $C$  in the above estimates only depends on the shape-regularity and quasi-uniformity of the mesh, but indent of polynomial degree  $k$  and mesh size  $h$ .

By the definition of the EDG scheme (6), its consistency (  $(-\epsilon \nabla u, u, u|_{\mathcal{E}_h})$  satisfies equations (6)), and properties of the projections, we have the following set of error equations holds:

$$(A.2a) \quad (\epsilon^{-1} \epsilon_q, \mathbf{r}_h)_{\mathcal{T}_h} - (\epsilon_u, \nabla \cdot \mathbf{r}_h)_{\mathcal{T}_h} + \langle \widehat{\epsilon}_u, \mathbf{r}_h \cdot \mathbf{n} \rangle_{\partial \mathcal{T}_h} = (\nabla \delta_u, \mathbf{r}_h)_{\mathcal{T}_h},$$

$$(A.2b) \quad -(\epsilon_q, \nabla v_h)_{\mathcal{T}_h} + \langle \epsilon_q \cdot \mathbf{n} + \alpha(\epsilon_u - \widehat{\epsilon}_u), v_h \rangle_{\partial \mathcal{T}_h} = \langle \delta_q \cdot \mathbf{n}, v_h \rangle_{\partial \mathcal{T}_h},$$

$$(A.2c) \quad -\langle \epsilon_q \cdot \mathbf{n} + \alpha(\epsilon_u - \widehat{\epsilon}_u), \widehat{v}_h \rangle_{\partial \mathcal{T}_h} = -\langle \delta_q \cdot \mathbf{n}, \widehat{v}_h \rangle_{\partial \mathcal{T}_h},$$

$$(A.2d) \quad (\epsilon_u, 1)_{\mathcal{T}_h} = 0,$$

for all  $(\mathbf{r}_h, v_h, \widehat{v}_h) \in \mathbf{R}_h \times V_h \times M_h$ .

Taking test functions  $(\mathbf{r}_h, v_h, \widehat{v}_h) := (\epsilon_q, \epsilon_u, \widehat{\epsilon}_u)$  in the above equations and summing up, we have

$$(\epsilon^{-1} \epsilon_q, \epsilon_q)_{\mathcal{T}_h} + \langle \alpha(\epsilon_u - \widehat{\epsilon}_u), \epsilon_u - \widehat{\epsilon}_u \rangle_{\partial \mathcal{T}_h} = (\nabla \delta_u, \epsilon_q)_{\mathcal{T}_h} + \langle \delta_q \cdot \mathbf{n}, \epsilon_u - \widehat{\epsilon}_u \rangle_{\partial \mathcal{T}_h}.$$

By the Cauchy-Schwarz inequality, the above right hand side is controlled by

$$(\epsilon \|\nabla \delta_u\|_{\mathcal{T}_h}^2 + \langle \alpha^{-1} \delta_q, \delta_q \rangle_{\partial \mathcal{T}_h})^{1/2} (\epsilon^{-1} \|\epsilon_q\|_{\mathcal{T}_h}^2 + \langle \alpha(\epsilon_u - \widehat{\epsilon}_u), \epsilon_u - \widehat{\epsilon}_u \rangle_{\partial \mathcal{T}_h})^{1/2}$$

Hence,

$$\begin{aligned} (\epsilon^{-1} \epsilon_q, \epsilon_q)_{\mathcal{T}_h} + \langle \alpha(\epsilon_u - \widehat{\epsilon}_u), \epsilon_u - \widehat{\epsilon}_u \rangle_{\partial \mathcal{T}_h} &\leq \epsilon \|\nabla \delta_u\|_{\mathcal{T}_h}^2 + \langle \alpha^{-1} \delta_q, \delta_q \rangle_{\partial \mathcal{T}_h} \\ &\leq C \epsilon \left(1 + \frac{\epsilon k}{h \alpha_{\min}}\right) \left(\frac{h}{k}\right)^{2k} \|u\|_{H^{k+1}(\Omega)}^2, \end{aligned}$$

where we used the  $hp$ -estimates (A.1b) and (A.1c) and  $\alpha_{\min} = \min_{F \in \mathcal{T}_h} \alpha|_F$ . By definition of  $\alpha$  in (6e), we have  $\alpha \geq C \epsilon k^2/h$  with a constant  $C$  depending only on

shape-regularity and quasi-uniformity of the mesh. Hence,

$$(A.3) \quad (\epsilon^{-1} \epsilon_q, \epsilon_q)_{\mathcal{T}_h} + \langle \alpha(\epsilon_u - \widehat{\epsilon}_u), \epsilon_u - \widehat{\epsilon}_u \rangle_{\partial \mathcal{T}_h} \leq C \epsilon \left( \frac{h}{k} \right)^{2k} \|u\|_{H^{k+1}(\Omega)}^2.$$

Recall that our domain is rectangular, hence we have the following  $H^2$ -regularity

$$(A.4) \quad \|\phi\|_{H^2(\Omega)} \leq C \|\epsilon_u\|_{\Omega}$$

where  $\phi \in H^1(\Omega)$  satisfies

$$-\nabla \cdot \nabla \phi = \epsilon_u \text{ in } \Omega, \quad \int_{\Omega} \phi = 0,$$

with a periodic boundary condition. Here  $C$  only depends on the domain  $\Omega$ .

Finally, the  $L^2$ -estimate in Theorem 1 follows from a standard duality argument. Denoting  $\boldsymbol{\psi} := \nabla \phi$  and  $\boldsymbol{\psi}_h = \boldsymbol{\Pi}_R(\nabla \phi)$ , we have

$$\begin{aligned} \|\epsilon_u\|_{\Omega}^2 &= (\epsilon_u, -\nabla \cdot \boldsymbol{\psi})_{\mathcal{T}_h} = (\epsilon_u, -\nabla \cdot (\boldsymbol{\psi} - \boldsymbol{\psi}_h))_{\mathcal{T}_h} - (\epsilon_u, \nabla \cdot \boldsymbol{\psi}_h)_{\mathcal{T}_h} \\ &= \langle \epsilon_u, -(\boldsymbol{\psi} - \boldsymbol{\psi}_h) \cdot \mathbf{n} \rangle_{\partial \mathcal{T}_h} - (\epsilon^{-1} \epsilon_q - \nabla \delta_u, \boldsymbol{\psi}_h)_{\mathcal{T}_h} \\ &\quad - \langle \widehat{\epsilon}_u, \boldsymbol{\psi}_h \cdot \mathbf{n} \rangle_{\partial \mathcal{T}_h} \\ &= \langle \epsilon_u - \widehat{\epsilon}_u, -(\boldsymbol{\psi} - \boldsymbol{\psi}_h) \cdot \mathbf{n} \rangle_{\partial \mathcal{T}_h} - (\epsilon^{-1} \epsilon_q - \nabla \delta_u, \boldsymbol{\psi}_h - \boldsymbol{\psi})_{\mathcal{T}_h} \\ &\quad + (\epsilon^{-1} \epsilon_q - \nabla \delta_u, \nabla \phi)_{\mathcal{T}_h}, \end{aligned}$$

where we used equation (A.2a) with test function  $\boldsymbol{\psi}_h$  in the second equality, and  $\langle \epsilon_u, \boldsymbol{\psi} \cdot \mathbf{n} \rangle_{\partial \mathcal{T}_h} = 0$  and  $\boldsymbol{\psi} = \nabla \phi$  in the third equality. Now, taking  $(v_h, \widehat{v}_h) = (\phi_h, \phi_h|_{\mathcal{E}_h})$  in equations (A.2b) and (A.2c) with  $\phi_h \in V_h \cap H^1(\Omega)$  and summing up the resulting expressions, we have  $(\epsilon_q, \nabla \phi_h)_{\mathcal{T}_h} = 0$ . We also have  $(\delta_u, \nabla \phi_h)_{\mathcal{T}_h} = 0$  by definition of the  $u_h^{CG}$ . Hence, for any  $\phi_h \in V_h \cap H^1(\Omega)$ ,

$$\begin{aligned} \|\epsilon_u\|_{\Omega}^2 &= \langle \epsilon_u - \widehat{\epsilon}_u, -(\boldsymbol{\psi} - \boldsymbol{\psi}_h) \cdot \mathbf{n} \rangle_{\partial \mathcal{T}_h} - (\epsilon^{-1} \epsilon_q - \nabla \delta_u, \boldsymbol{\psi}_h - \boldsymbol{\psi})_{\mathcal{T}_h} \\ &\quad + (\epsilon^{-1} \epsilon_q - \nabla \delta_u, \nabla (\phi - \phi_h))_{\mathcal{T}_h}. \end{aligned}$$

Applying the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} \|\epsilon_u\|_{\Omega}^2 &\leq \left( \langle \frac{\alpha}{\epsilon} (\epsilon_u - \widehat{\epsilon}_u), \epsilon_u - \widehat{\epsilon}_u \rangle_{\partial \mathcal{T}_h} + \epsilon^{-2} \|\epsilon_q\|_{\mathcal{T}_h}^2 + \|\nabla \delta_u\|_{\mathcal{T}_h}^2 \right)^{1/2} \\ &\quad \left( \langle \frac{\epsilon}{\alpha} (\boldsymbol{\psi} - \boldsymbol{\psi}_h), \boldsymbol{\psi} - \boldsymbol{\psi}_h \rangle_{\partial \mathcal{T}_h} + \|\boldsymbol{\psi} - \boldsymbol{\psi}_h\|_{\mathcal{T}_h}^2 + \inf_{\phi_h \in V_h \cap H^1(\Omega)} \|\nabla (\phi - \phi_h)\|_{\mathcal{T}_h}^2 \right)^{1/2} \\ &\leq C \left( \frac{h}{k} \right)^k \|u\|_{H^{k+1}(\Omega)} \frac{h}{k} \|\phi\|_{H^2(\Omega)} \end{aligned}$$

By the  $H^2$ -regularity result (A.4), we get

$$\|\epsilon_u\|_{\Omega} \leq C \left( \frac{h}{k} \right)^{k+1} \|u\|_{H^{k+1}(\Omega)}.$$

This completes the proof of Theorem 1.

Division of Applied Mathematics, Brown University, Providence, RI 02912, U.S.A.

*E-mail:* guosheng\_fu@brown.edu

*URL:* <https://www.brown.edu/academics/applied-mathematics/guosheng-fu>

*E-mail:* shu@dam.brown.edu

*URL:* <http://www.dam.brown.edu/people/shu>