Journal of Scientific Computing _####################_
https://doi.org/10.1007/s10915-020-01197-x

Optimal Error Estimates of the Discontinuous Galerkin Method with Upwind-Biased Fluxes for 2D Linear Variable Coefficients Hyperbolic Equations

Minghui Liu[1] · Boying Wu[2] · Xiong Meng[2]

Received: 9 January 2020 / Revised: 3 March 2020 / Accepted: 14 March 2020

Abstract

In this paper, we consider the discontinuous Galerkin method with upwind-biased numerical fluxes for two-dimensional linear hyperbolic equations with degenerate variable coefficients on Cartesian meshes. The L^2-stability is guaranteed by the numerical viscosity of the upwind-biased fluxes, and the adjustable numerical viscosity is useful in resolving waves and is beneficial for long time simulations. To derive optimal error estimates, a new projection is introduced and analyzed, which is the tensor product of the corresponding one-dimensional piecewise global projection for each variable. The analysis of uniqueness and optimal interpolation properties of the proposed projection is subtle, as the projection requires different collocations for the projection errors involving the volume integral, the boundary integral and the boundary points. By combining the optimal interpolation estimates and a sharp bound for the projection errors, optimal error estimates are obtained. Numerical experiments are shown to confirm the validity of the theoretical results.

Keywords Discontinuous Galerkin method · Upwind-biased fluxes · 2D hyperbolic equation · Error estimates · Projections

Mathematics Subject Classification 65M60 · 65M15

The research of Boying Wu was supported by NSFC Grants U1637208 and 71773024. The research of Xiong Meng was supported by NSFC Grant 11971132.

✉ Xiong Meng
xiongmeng@hit.edu.cn

Minghui Liu
lmh@hit.edu.cn

Boying Wu
mathwby@hit.edu.cn

1 School of Mathematics, Harbin Institute of Technology, Harbin 150001, Heilongjiang, China

2 School of Mathematics and Institute for Advanced Study in Mathematics, Harbin Institute of Technology, Harbin 150001, Heilongjiang, China

1 Introduction

In this paper, we study optimal error estimates for the discontinuous Galerkin (DG) methods with upwind-biased numerical fluxes for two-dimensional linear hyperbolic equations with degenerate variable coefficients

$$u_t + (a(x, y)u)_x + (b(x, y)u)_y = 0, \qquad (x, y, t) \in \Omega \times (0, T], \tag{1.1a}$$

$$u(x, y, 0) = u_0(x, y), \qquad (x, y) \in \Omega, \tag{1.1b}$$

where $a(x, y)$ and $b(x, y)$ are given smooth functions that have turning points on a bounded rectangular domain in $\mathbb{R}^2$, and $u_0(x, y)$ is a smooth initial condition. The periodic boundary conditions are mainly discussed, and for the Dirichlet boundary condition case, we refer to [15, Sect. 3.5]. By constructing a special piecewise *global* projection and establishing the optimal interpolation properties as well as a sharp bound for projection error terms, we are able to derive optimal error estimates for the DG methods with upwind-biased fluxes on Cartesian meshes.

The DG method is a class of nonconforming finite element methods, designed mainly to capture shocks without nonphysical oscillations and to achieve a uniform high order accuracy for smooth solutions. Proposed by Reed and Hill [21] for solving a linear steady-state hyperbolic equation, the DG methods were developed by Cockburn and Shu [6,9,10,12] for solving nonlinear time-dependent conservation laws. Since the basis functions can be completely discontinuous at element interfaces, the DG method provides more flexibility for h-p adaptivity. Due to its excellent features for computing both smooth and discontinuous solutions, the DG method was generalized to lots of different partial differential equations (PDEs), such as diffusion equations and high order wave equations, for which the local DG (LDG) method [11] and the ultra weak DG method [5] are proposed. For recent development and applications of DG methods, we refer to the survey papers [8,22].

Traditionally, purely upwind fluxes are chosen in the DG scheme for hyperbolic equations. However, in order to better resolve discontinuities and capture the wave for long time integrations, the upwind-biased flux in possession of adjustable numerical viscosities can be helpful. Specifically, in order to simulate shocks, a small amount of numerical dissipation that is lower than that of an upwind flux can be considered, and this is achieved by choosing suitable weights in the generalized local Lax–Friedrichs flux in [17]. On the other hand, for smooth solutions of hyperbolic equations, a numerical flux with negligible numerical dissipation can be chosen which will produce a smaller magnitude of the error (especially for even polynomial degrees) [15,20]. For linearized Korteweg–de Vries (KdV) equations, by choosing a downwind-biased flux for the convection term, a nearly energy conserving LDG scheme [16] shows a better result for long time simulations, when compared with the standard upwind flux. In addition, an energy conserving DG scheme is proposed and analyzed with central fluxes for generalized KdV equations in [1], and a special global projection is constructed.

First proposed in [20], the idea of the upwind-biased flux has shown its flexibility and advantages for solving different types of PDEs. In [18], Liu and Ploymaklam consider the LDG method for Burgers–Poisson equations, in which weighted numerical fluxes are used for the diffusion term. In [4], Cheng et al. adopt upwind-biased and generalized alternating numerical fluxes for solving linear convection-diffusion equations; for fully discretized analysis, please refer to [23]. In addition to optimal error estimates for these generalized numerical fluxes, superconvergence of the DG and LDG methods have been studied for linear hyperbolic equations in [3,13] and convection–diffusion equations in [19]. We would like

Journal: 10915 Article No.: 1197 TYPESET DISK LE CP Disp.:2020/3/19 Pages: 19 Layout: Small

to point out that a main technical issue related to the analysis of upwind-biased fluxes is the coupling feature of the projection, as it uses information from both sides of cell interfaces due to a weight for the flux. Therefore, in contrast to an explicit formula for local projections for purely upwind fluxes, a linear algebraic system of equations needs to be solved when upwind-biased fluxes are considered and the unknowns of the global projection in the discrete L^2 norm should be uniformly bounded. Moreover, for linear hyperbolic equations with degenerate coefficients, if we simply use the approach as that for the linear equations, the resulting matrix may be singular and thus existence of the designed projection cannot be obtained. To solve this problem, we proposed in [15] a piecewise *global* projection by imposing an additional exact collocation condition at one of the boundary point, at which the value of $f'(u)$ is of the mesh size. Consequently, the whole region can be divided into three parts connected by a varying sign element on which a local Gauss–Radau (GR) projection is defined. By requiring some suitable collocations of points at which $f'(u)$ does not change sign, we obtain two matrices that are diagonally dominant, indicating that the resulting two matrices are always invertible and thus uniqueness as well as optimal interpolation properties can be proved.

As a continued work of [15,20], we consider in this paper the optimal error analysis of DG methods with upwind-biased fluxes for 2D hyperbolic equations with degenerate coefficients on Cartesian meshes. To this end, we first define a new projection which is a tensor product of the 1D piecewise *global* projection in [15]. However, the projection is not easy to analyze, as it involves different collocations for the volume integral, the boundary integral and boundary points of different cells. Noting that this projection cannot completely eliminate the contribution for projection errors, a sharp estimate for the projection errors is derived, which is based on a global inequality rather than a local equality as that in [4,7].

The rest of this paper is organized as follows. In Sect. 2, we present the DG scheme with upwind-biased fluxes for 2D linear hyperbolic equations with degenerate variable coefficients and show L^2 stability. In Sect. 3, we begin by presenting some notation and recalling some preliminaries for the 1D piecewise *global* projection in Sect. 3.1. In Sect. 3.2, we define a new piecewise *global* projection and show existence and optimal approximation properties. A sharp bound of the projection error is shown in Sect. 3.3. The optimal error estimates are given in Sect. 3.4. In Sect. 4, numerical experiments are given to confirm theoretical results. Some concluding remarks are given in Sect. 5.

2 The DG Method

In this section, we define the DG scheme and show the L^2 stability.

2.1 The DG Scheme

Prior to giving the definition of the DG scheme, let us first present some notation. For any positive integer r, let $\mathbb{Z}_r = \{1, \ldots, r\}$ and denote by $\Omega_h = \{K \triangleq I_i \times J_j\}$ a Cartesian mesh of Ω, where K are shape regular rectangular elements and $I_i = (x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}})$, $J_j = (y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}})$ with $i \in \mathbb{Z}_{N_1}$ and $j \in \mathbb{Z}_{N_2}$. The cell center is (x_i, y_j), where $x_i = \frac{1}{2}\big(x_{i-\frac{1}{2}} + x_{i+\frac{1}{2}}\big)$, $y_j = \frac{1}{2}\big(y_{j-\frac{1}{2}} + y_{j+\frac{1}{2}}\big)$. We set $\partial\Omega_h = \{\partial K : K \in \Omega_h\}$ being a collection of cell boundaries. Moreover, we denote $h_x = \max_{i\in\mathbb{Z}_{N_1}} h_i^x$, $h_y = \max_{j\in\mathbb{Z}_{N_2}} h_j^y$ with $h_i^x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$, $h_j^y = y_{j+\frac{1}{2}} - y_{j-\frac{1}{2}}$, and $h = \max(h_x, h_y)$. Associated with the mesh Ω_h, the

Journal: 10915 Article No.: 1197 TYPESET DISK LE CP Disp.:2020/3/19 Pages: 19 Layout: Small

finite element space is

$$V_h = \{v \in L^2(\Omega) : v|_K \in Q^k(K) \quad \forall K \in \Omega_h\},$$

where $Q^k(K)$ is the space of the tensor product of polynomials of degree at most k for each variable on K.

Since functions in V_h can be discontinuous across element boundaries, for $y \in J_j$ and $j \in \mathbb{Z}_{N_2}$, we use $v^-_{i+\frac{1}{2},y}$ and $v^+_{i+\frac{1}{2},y}$ to denote the traces evaluated from the left element $I_i \times J_j$ and the right element $I_{i+1} \times J_j$; the jump and the average of v are denoted by $[\![v]\!]_{i+\frac{1}{2},y} = v^+_{i+\frac{1}{2},y} - v^-_{i+\frac{1}{2},y}$ and $\{\!\{v\}\!\}_{i+\frac{1}{2},y} = \frac{1}{2}(v^-_{i+\frac{1}{2},y} + v^+_{i+\frac{1}{2},y})$. Analogously, $v^-_{x,j+\frac{1}{2}}$, $v^+_{x,j+\frac{1}{2}}$, $[\![v]\!]_{x,j+\frac{1}{2}}$ and $\{\!\{v\}\!\}_{x,j+\frac{1}{2}}$ can be well defined on horizontal edges when $x \in I_i$ and $i \in \mathbb{Z}_{N_1}$.

As usual, we adopt $W^{\ell,p}(D)$ to represent the standard Sobolev space on D equipped with the norm $\|\cdot\|_{\ell,p,D}$ with $\ell \ge 0$, $p = 2, \infty$, and $D = K, \Omega$ etc. The subscripts D, ℓ will be omitted when $D = \Omega$ or $\ell = 0$, and $W^{\ell,p}(D) = H^\ell(D)$ when $p = 2$. Similarly, the boundary L^2 norm is $\|v\|_{\partial\Omega_h} = \left(\sum_{K\in\Omega_h} \|v\|^2_{\partial K}\right)^{\frac{1}{2}}$ with $\|v\|^2_{\partial K} = \int_{J_j} [(v^+_{i-\frac{1}{2},y})^2 + (v^-_{i+\frac{1}{2},y})^2]\mathrm{d}y + \int_{I_i} [(v^+_{x,j-\frac{1}{2}})^2 + (v^-_{x,j+\frac{1}{2}})^2]\mathrm{d}x$.

We are now ready to present the DG scheme for (1.1). For all $t \in (0, T]$, find $u_h(t) \in V_h$ such that

$$\begin{aligned}
&\int_K u_{ht} v_h \mathrm{d}x\mathrm{d}y - \int_K a u_h (v_h)_x \mathrm{d}x\mathrm{d}y + \int_{J_j} (a\hat{u}_h v_h^-)_{i+\frac{1}{2},y}\mathrm{d}y - \int_{J_j} (a\hat{u}_h v_h^+)_{i-\frac{1}{2},y}\mathrm{d}y \\
&- \int_K b u_h (v_h)_y \mathrm{d}x\mathrm{d}y + \int_{I_i} (b\hat{u}_h v_h^-)_{x,j+\frac{1}{2}}\mathrm{d}x - \int_{I_i} (b\hat{u}_h v_h^+)_{x,j-\frac{1}{2}}\mathrm{d}x
\end{aligned} \tag{2.1}$$

holds for all $v_h \in V_h$ and $K \in \Omega_h$. Instead of using purely upwind fluxes for the hat terms, here we consider a more generalized upwind-biased fluxes in the form

$$\hat{u}_h = \begin{cases} u_h^{(\theta_1)} & \text{if } a(x_{i+\frac{1}{2}}, y_j) \ge 0, \\ u_h^{(\tilde{\theta}_1)} & \text{if } a(x_{i+\frac{1}{2}}, y_j) < 0, \end{cases} \quad \text{at } (x_{i+\frac{1}{2}}, y), \tag{2.2a}$$

$$\hat{u}_h = \begin{cases} u_h^{(\theta_2)} & \text{if } b(x_i, y_{j+\frac{1}{2}}) \ge 0, \\ u_h^{(\tilde{\theta}_2)} & \text{if } b(x_i, y_{j+\frac{1}{2}}) < 0, \end{cases} \quad \text{at } (x, y_{j+\frac{1}{2}}). \tag{2.2b}$$

Here and in what follows, $w^{(\theta_1)}_{i+\frac{1}{2},y} = \theta_1 w^-_{i+\frac{1}{2},y} + \tilde{\theta}_1 w^+_{i+\frac{1}{2},y}$, $w^{(\theta_2)}_{x,j+\frac{1}{2}} = \theta_2 w^-_{x,j+\frac{1}{2}} + \tilde{\theta}_2 w^+_{x,j+\frac{1}{2}}$ and $\theta_s > \frac{1}{2}$ are the weights in the upwind-biased fluxes with $\tilde{\theta}_s = 1 - \theta_s$ for $s = 1, 2$. For the numerical initial discretization, we can simply take the L^2 projection of u_0. This completes the definition of the DG scheme.

For notational convenience, we would like to use the DG spatial discretization operators in the form

$$\mathcal{H}^x_K(w, v) = \int_K w v_x \mathrm{d}x\mathrm{d}y - \int_{J_j} (\hat{w}v^-)_{i+\frac{1}{2},y}\mathrm{d}y + \int_{J_j} (\hat{w}v^+)_{i-\frac{1}{2},y}\mathrm{d}y, \tag{2.3a}$$

$$\mathcal{H}^y_K(w, v) = \int_K w v_y \mathrm{d}x\mathrm{d}y - \int_{I_i} (\hat{w}v^-)_{x,j+\frac{1}{2}}\mathrm{d}x + \int_{I_i} (\hat{w}v^+)_{x,j-\frac{1}{2}}\mathrm{d}x, \tag{2.3b}$$

and the removal of the subscript K indicates the summation of all $K \in \Omega_h$.

Journal: **10915** Article No.: **1197** ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:**2020/3/19** Pages: **19** Layout: **Small**

2.2 Stability

The DG scheme (2.1) with the upwind-biased fluxes (2.2) satisfies the following L^2 stability.

Proposition 2.1 *The solution of the DG scheme* (2.1) *with the fluxes* (2.2) *satisfies*

$$\|u_h(t)\| \le C\|u_h(0)\|, \quad \forall t > 0,$$

where C is a positive constant depending on a_x and b_y.

Proof Taking $v_h = u_h$ in (2.1) and summing over all K, we get

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u_h\|^2 = \mathcal{H}^x(au_h, v_h) + \mathcal{H}^y(bu_h, v_h). \tag{2.4}$$

It follows from integration by parts and a local linearization $a_{i+\frac{1}{2},y} = \left(a_{i+\frac{1}{2},y} - a_{i+\frac{1}{2},j}\right) + a_{i+\frac{1}{2},j}$ that

$$\begin{aligned}
\mathcal{H}^x(au_h, u_h) = &\sum_{K\in\Omega_h}\int_K -\frac{a_x}{2}u_h^2\mathrm{d}x\mathrm{d}y \\
&- \left(\theta_1 - \frac{1}{2}\right)\sum_{j=1}^{N_2}\int_{J_j}\sum_{i=1}^{N_1}|a_{i+\frac{1}{2},j}|[\![u_h]\!]^2_{i+\frac{1}{2},y}\mathrm{d}y \\
&+ \sum_{j=1}^{N_2}\int_{J_j}\sum_{i=1}^{N_1}\left(a_{i+\frac{1}{2},y} - a_{i+\frac{1}{2},j}\right)\left(\hat{u}_h - \{\!\{u_h\}\!\}\right)_{i+\frac{1}{2},y}[\![u_h]\!]_{i+\frac{1}{2},y}\mathrm{d}y \\
\le\ & C\|u_h\|^2 + Ch\|u_h\|^2_{\partial\Omega_h} \\
\le\ & C\|u_h\|^2,
\end{aligned}$$

since $\theta_1 > \frac{1}{2}$ and $\left|a_{i+\frac{1}{2},y} - a_{i+\frac{1}{2},j}\right| \le Ch$, and we have also used the fact that $\sum_{K\in\Omega_h} w = \sum_{j=1}^{N_2}\sum_{i=1}^{N_1} w$ implied by the structure of Cartesian meshes and the inverse property (ii). Analogously, for $\mathcal{H}^y(bu_h, u_h)$, we have

$$\mathcal{H}^y(bu_h, v_h) \le C\|u_h\|^2.$$

A substitution of the above two inequalities into (2.4) together with the Gronwall's inequality leads to the L^2 stability. This finishes the proof of Proposition 2.1. □

3 Optimal Error Estimates

3.1 Preliminaries

3.1.1 A Special Projection in 1D

Basically, for optimal error estimates of the DG methods with upwind-biased fluxes solving linear hyperbolic equations with variable coefficients, the design of special projection is mainly divided into two cases. The first case is that the derivatives of flux functions a, b do not change signs over Ω; for such a case, one can simply employ the local linearization approach for a, b at each element and take the projection proposed in [20]. Note that in [20] a

Journal: 10915 Article No.: 1197 TYPESET DISK LE CP Disp.:2020/3/19 Pages: 19 Layout: Small

global linear system of size $N_1N_2 \times N_1N_2$ needs to be solved, which, however, can be treated as the tensor product of two matrices of size N_1 and N_2, respectively. For the second case when a, b do change signs over Ω, the situation is totally different, for which we should be careful to rearrange different collocation conditions. A successful treatment proposed in [15] is to split the whole projection into a piecewise *global* projection via replacing a collocation condition by a decoupling condition for a sign varying element (connecting cell).

To be more specific, let us recall the definition of the special piecewise *global* projection in 1D. Thus, (1.1) reduces to

$$u_t + (c(x)u)_x = 0.$$

According to the sign variation of $c(x)$ together with an assumption that $f'(u) = c(x)$ has only two zeros, we follow [15] and denote

$$\beta = \{j |\ c(x_{j-\frac{1}{2}}) < 0 \text{ and } c(x_{j+\frac{1}{2}}) \geq 0,\ \forall j \in \mathbb{Z}_N\}, \tag{3.1a}$$

$$\gamma = \{j |\ c(x_{j-\frac{1}{2}}) > 0 \text{ and } c(x_{j+\frac{1}{2}}) \leq 0,\ \forall j \in \mathbb{Z}_N\}, \tag{3.1b}$$

and

$$\mathbb{b}^+ = \{\beta, \ldots, \gamma - 1\}, \quad \mathbb{b}^- = \{\gamma + 1, \ldots, \beta - 1\} \tag{3.1c}$$

no matter whether γ is greater than β or not. Note that $\mathbb{Z}_N \backslash \{\mathbb{b}^+ \cup \mathbb{b}^-\} = \gamma$, allowing us to define an additional decoupling condition for the element I_γ in (3.2b) below. The piecewise *global* projection $\mathcal{P}_h^\theta u$ is defined as a delicate collocation at different points with the purpose of obtaining matrices that are always diagonally dominant, resulting in the uniqueness and existence of the projection. It reads

$$\int_{I_i} (\mathcal{P}_h^\theta u)\varphi \mathrm{d}x = \int_{I_i} u\varphi \mathrm{d}x \qquad \forall \varphi \in P^{k-1}(I_i), \quad i \in \mathbb{Z}_N, \tag{3.2a}$$

$$(\mathcal{P}_h^\theta u)^-_{i+\frac{1}{2}} = u^-_{i+\frac{1}{2}} \qquad \text{at } x_{i+\frac{1}{2}}, \qquad i = \gamma, \tag{3.2b}$$

$$\widehat{(\mathcal{P}_h^\theta u)}_{i+\frac{1}{2}} = \hat{u}_{i+\frac{1}{2}} \qquad \text{at } x_{i+\frac{1}{2}}, \qquad i \in \mathbb{b}^+, \tag{3.2c}$$

$$\widehat{(\mathcal{P}_h^\theta u)}_{i-\frac{1}{2}} = \hat{u}_{i-\frac{1}{2}} \qquad \text{at } x_{i-\frac{1}{2}}, \qquad i \in \mathbb{b}^-, \tag{3.2d}$$

where $\hat{w} = \theta w^- + \widetilde{\theta} w^+$ for $c(x_{i+\frac{1}{2}}) \geq 0$ and $\hat{w} = \widetilde{\theta} w^- + \theta w^+$ for $c(x_{i+\frac{1}{2}}) < 0$ with $\theta > \frac{1}{2}$ and $\widetilde{\theta} = 1 - \theta$. We can see that the projection is doubly defined at $x_{\gamma+\frac{1}{2}}$ without any collocation at $x_{\beta-\frac{1}{2}}$ (namely $(u - \mathcal{P}_h^\theta u)_{\beta-\frac{1}{2}} \neq 0$), which will give us a local GR projection on I_γ, entailing that $\mathcal{P}_h^\theta u$ can be decoupled starting from this element. For more details, see [15, Lemma 3.1 and Remark 3.1].

3.1.2 Inverse Properties in 2D

For any function $v \in V_h$, the following inverse inequalities hold [2]:

$$\text{(i)} \|\nabla v\| \leq Ch^{-1}\|v\|, \ \text{(ii)} \|v\|_{\partial\Omega_h} \leq Ch^{-\frac{1}{2}}\|v\|, \ \text{(iii)} \|v\|_\infty \leq Ch^{-1}\|v\|, \tag{3.3}$$

where $\|\nabla v\| = \left(\|v_x\|^2 + \|v_y\|^2\right)^{\frac{1}{2}}$ and the bounding constant C is independent of h.

Journal: 10915 Article No.: 1197 TYPESET DISK LE CP Disp.:2020/3/19 Pages: 19 Layout: Small

3.2 A Special Piecewise Global Projection in 2D

We are now ready to define a new projection for the 2D case. For simplicity, we consider the univariate case of (1.1), namely $a(x, y) = a(x)$ and $b(x, y) = b(y)$; definition of the projection for the multivariate case of (1.1) is more involved, as sign variations will be quite complicated. Analogous to $\beta, \gamma, \mathbb{b}^+, \mathbb{b}^-$ in (3.1a)–(3.1c) for 1D, we can define $\beta_1, \gamma_1, \beta_2, \gamma_2$, and further $\mathbb{b}_1^+, \mathbb{b}_1^-, \mathbb{b}_2^+, \mathbb{b}_2^-$. Next, for $u \in W^{1,\infty}(\Omega_h)$, the new projection, denoted by $\Pi_h^{\theta_1,\theta_2} u$, is defined to be the tensor product of the corresponding 1D projection. That is,

$$\Pi_h^{\theta_1,\theta_2} u = \mathcal{P}_{h_x}^{\theta_1} \otimes \mathcal{P}_{h_y}^{\theta_2} u, \tag{3.4}$$

where the subscripts x and y denote the 1D projection is used as given in (3.2). Taking into account collocations at different boundary points, the projection $\Pi_h^{\theta_1,\theta_2} u$ is a polynomial in V_h satisfying the following four groups of identities, i.e., the volume integrals

$$\int_K \Pi_h^{\theta_1,\theta_2} u(x, y) v_h(x, y) \mathrm{d}x\mathrm{d}y = \int_K u(x, y) v_h(x, y) \mathrm{d}x\mathrm{d}y, \tag{3.5a}$$

the collocations for vertical boundary integrals

$$\int_{J_j} (\Pi_h^{\theta_1,\theta_2} u)^-_{i+\frac12,y} (v_h)^-_{i+\frac12,y} \mathrm{d}y = \int_{J_j} u^-_{i+\frac12,y} (v_h)^-_{i+\frac12,y} \mathrm{d}y \quad i = \gamma_1, \tag{3.5b}$$

$$\int_{J_j} (\Pi_h^{\theta_1,\theta_2} u)^{(\theta_1)}_{i+\frac12,y} (v_h)^-_{i+\frac12,y} \mathrm{d}y = \int_{J_j} u^{(\theta_1)}_{i+\frac12,y} (v_h)^-_{i+\frac12,y} \mathrm{d}y \quad i \in \mathbb{b}_1^+, \tag{3.5c}$$

$$\int_{J_j} (\Pi_h^{\theta_1,\theta_2} u)^{(\widetilde{\theta}_1)}_{i-\frac12,y} (v_h)^+_{i-\frac12,y} \mathrm{d}y = \int_{J_j} u^{(\widetilde{\theta}_1)}_{i-\frac12,y} (v_h)^+_{i-\frac12,y} \mathrm{d}y \quad i \in \mathbb{b}_1^-, \tag{3.5d}$$

the collocations for horizontal boundary integrals

$$\int_{I_i} (\Pi_h^{\theta_1,\theta_2} u)^-_{x,j+\frac12} (v_h)^-_{x,j+\frac12} \mathrm{d}x = \int_{I_i} u^-_{x,j+\frac12} (v_h)^-_{x,j+\frac12} \mathrm{d}x \quad j = \gamma_2, \tag{3.5e}$$

$$\int_{I_i} (\Pi_h^{\theta_1,\theta_2} u)^{(\theta_2)}_{x,j+\frac12} (v_h)^-_{x,j+\frac12} \mathrm{d}x = \int_{I_i} u^{(\theta_2)}_{x,j+\frac12} (v_h)^-_{x,j+\frac12} \mathrm{d}x \quad j \in \mathbb{b}_2^+, \tag{3.5f}$$

$$\int_{I_i} (\Pi_h^{\theta_1,\theta_2} u)^{(\widetilde{\theta}_2)}_{x,j-\frac12} (v_h)^+_{x,j-\frac12} \mathrm{d}x = \int_{I_i} u^{(\widetilde{\theta}_2)}_{x,j-\frac12} (v_h)^+_{x,j-\frac12} \mathrm{d}x \quad j \in \mathbb{b}_2^-, \tag{3.5g}$$

which hold for all $v_h \in Q^{k-1}(K)$ and $K \in \Omega_h$, and the collocations for boundary points

$$(\Pi_h^{\theta_1,\theta_2} u)^{-,-}_{i+\frac12,j+\frac12} = u^{-,-}_{i+\frac12,j+\frac12} \qquad (i, j) = (\gamma_1, \gamma_2), \tag{3.5h}$$

$$(\Pi_h^{\theta_1,\theta_2} u)^{(\theta_1),-}_{i+\frac12,j+\frac12} = u^{(\theta_1),-}_{i+\frac12,j+\frac12} \qquad (i, j) \in (\mathbb{b}_1^+, \gamma_2), \tag{3.5i}$$

$$(\Pi_h^{\theta_1,\theta_2} u)^{(\widetilde{\theta}_1),-}_{i-\frac12,j+\frac12} = u^{(\widetilde{\theta}_1),-}_{i-\frac12,j+\frac12} \qquad (i, j) \in (\mathbb{b}_1^-, \gamma_2), \tag{3.5j}$$

$$(\Pi_h^{\theta_1,\theta_2} u)^{-,(\theta_2)}_{i+\frac12,j+\frac12} = u^{-,(\theta_2)}_{i+\frac12,j+\frac12} \qquad (i, j) \in (\gamma_1, \mathbb{b}_2^+), \tag{3.5k}$$

$$(\Pi_h^{\theta_1,\theta_2} u)^{-,(\widetilde{\theta}_2)}_{i+\frac12,j-\frac12} = u^{-,(\widetilde{\theta}_2)}_{i+\frac12,j-\frac12} \qquad (i, j) \in (\gamma_1, \mathbb{b}_2^-), \tag{3.5l}$$

$$(\Pi_h^{\theta_1,\theta_2} u)^{(\theta_1,\theta_2)}_{i+\frac12,j+\frac12} = u^{(\theta_1,\theta_2)}_{i+\frac12,j+\frac12} \qquad (i, j) \in (\mathbb{b}_1^+, \mathbb{b}_2^+), \tag{3.5m}$$

$$(\Pi_h^{\theta_1,\theta_2} u)^{(\theta_1,\tilde{\theta}_2)}_{i+\frac12,j-\frac12} = u^{(\theta_1,\tilde{\theta}_2)}_{i+\frac12,j-\frac12} \qquad (i, j) \in (\mathbb{b}_1^+, \mathbb{b}_2^-), \tag{3.5n}$$

Journal: 10915 Article No.: 1197 TYPESET DISK LE CP Disp.:2020/3/19 Pages: 19 Layout: Small

$$(\Pi_h^{\theta_1,\theta_2}u)_{i-\frac{1}{2},j+\frac{1}{2}}^{(\widetilde{\theta}_1,\theta_2)} = u_{i-\frac{1}{2},j+\frac{1}{2}}^{(\widetilde{\theta}_1,\theta_2)} \qquad (i,j) \in (\mathbb{b}_1^-, \mathbb{b}_2^+), \tag{3.5o}$$

$$(\Pi_h^{\theta_1,\theta_2}u)_{i-\frac{1}{2},j-\frac{1}{2}}^{(\widetilde{\theta}_1,\widetilde{\theta}_2)} = u_{i-\frac{1}{2},j-\frac{1}{2}}^{(\widetilde{\theta}_1,\widetilde{\theta}_2)} \qquad (i,j) \in (\mathbb{b}_1^-, \mathbb{b}_2^-). \tag{3.5p}$$

Here and below,

$$w_{i+\frac{1}{2},j+\frac{1}{2}}^{(\theta_1,\theta_2)} = \theta_1\theta_2 w_{i+\frac{1}{2},j+\frac{1}{2}}^{-,-} + \theta_1\widetilde{\theta}_2 w_{i+\frac{1}{2},j+\frac{1}{2}}^{-,+} + \widetilde{\theta}_1\theta_2 w_{i+\frac{1}{2},j+\frac{1}{2}}^{+,-} + \widetilde{\theta}_1\widetilde{\theta}_2 w_{i+\frac{1}{2},j+\frac{1}{2}}^{+,+}.$$

In order to show uniqueness, existence and optimal approximation properties of the projection $\Pi_h^{\theta_1,\theta_2}u$, we need to recall the definition Π_h^- as defined in [7,20]. Specifically, for $u \in W^{1,\infty}(\Omega_h)$, the projection $\Pi_h^- u$ is a unique polynomial in V_h such that

$$\int_K \Pi_h^- u(x,y)v_h(x,y)\mathrm{d}x\mathrm{d}y = \int_K u(x,y)v_h(x,y)\mathrm{d}x\mathrm{d}y, \tag{3.6a}$$

$$\int_{J_j} (\Pi_h^- u)_{i+\frac{1}{2},y}^-(v_h)_{i+\frac{1}{2},y}^-\mathrm{d}y = \int_{J_j} u_{i+\frac{1}{2},y}^-(v_h)_{i+\frac{1}{2},y}^-\mathrm{d}y, \tag{3.6b}$$

$$\int_{I_i} (\Pi_h^- u)_{x,j+\frac{1}{2}}^-(v_h)_{x,j+\frac{1}{2}}^-\mathrm{d}x = \int_{I_i} u_{x,j+\frac{1}{2}}^-(v_h)_{x,j+\frac{1}{2}}^-\mathrm{d}x, \tag{3.6c}$$

$$(\Pi_h^- u)_{i+\frac{1}{2},j+\frac{1}{2}}^{-,-} = u_{i+\frac{1}{2},j+\frac{1}{2}}^{-,-} \tag{3.6d}$$

hold for all $v_h \in Q^{k-1}(K)$ and $K \in \Omega_h$. Clearly, $\Pi_h^- u$ is locally defined and satisfies the optimal approximation property [2,7]:

$$\|u - \Pi_h^- u\| + h^{\frac{1}{2}}\|u - \Pi_h^- u\|_{\partial\Omega_h} + h\|u - \Pi_h^- u\|_\infty \le Ch^{k+1}\|u\|_{k+1}, \tag{3.7}$$

where C is independent of h.

Existence and optimal approximation properties of the piecewise *global* projection $\Pi_h^{\theta_1,\theta_2}$ are established in the following lemma.

Lemma 3.1 *There exists a unique $\Pi_h^{\theta_1,\theta_2}$ satisfying (3.5a)–(3.5p). Moreover, assume that u is sufficiently smooth, i.e. $u \in H^{k+1}(\Omega_h)$, and periodic. Then, there holds the optimal approximation property:*

$$\|u - \Pi_h^{\theta_1,\theta_2}u\| + h^{\frac{1}{2}}\|u - \Pi_h^{\theta_1,\theta_2}u\|_{\partial\Omega_h} \le Ch^{k+1}\|u\|_{k+1}, \tag{3.8}$$

where $\|u\|_{k+1} = \left(\sum_{K\in\Omega_h}\|u\|_{k+1,K}^2\right)^{\frac{1}{2}}$ is the broken Sobolev $k+1$ norm of u and C is independent of the mesh size h.

Proof Denote $\Pi_h^{\theta_1,\theta_2}u - u = \Pi_h^{\theta_1,\theta_2}u - \Pi_h^- u + \Pi_h^- u - u \triangleq E + \psi$ with $E = \Pi_h^{\theta_1,\theta_2}u - \Pi_h^- u \in V_h$ and $\psi = \Pi_h^- u - u$. Since $\Pi_h^- u$ defined in (3.6) has already known, if we can prove the existence and uniqueness of E, then $\Pi_h^{\theta_1,\theta_2}u = E + \Pi_h^- u$ will be unique. By the definitions of $\Pi_h^{\theta_1,\theta_2}$ and Π_h^-, E satisfies the following identities

$$\int_K Ev_h\mathrm{d}x\mathrm{d}y = 0, \tag{3.9a}$$

$$\int_{J_j} E_{i+\frac{1}{2},y}^-(v_h)_{i+\frac{1}{2},y}^-\mathrm{d}y = 0 \qquad i = \gamma_1, \tag{3.9b}$$

Journal: 10915 Article No.: 1197 TYPESET DISK LE CP Disp.:2020/3/19 Pages: 19 Layout: Small

$$\int_{J_j} E^{(\theta_1)}_{i+\frac12,y}(v_h)^-_{i+\frac12,y}\mathrm{d}y = -\widetilde{\theta}_1 \int_{J_j} \psi^+_{i+\frac12,y}(v_h)^-_{i+\frac12,y}\mathrm{d}y \quad i \in \mathbb{b}_1^+, \tag{3.9c}$$

$$\int_{J_j} E^{(\widetilde{\theta}_1)}_{i-\frac12,y}(v_h)^+_{i-\frac12,y}\mathrm{d}y = -\theta_1 \int_{J_j} \psi^+_{i-\frac12,y}(v_h)^+_{i-\frac12,y}\mathrm{d}y \quad i \in \mathbb{b}_1^-, \tag{3.9d}$$

$$\int_{I_i} E^-_{x,j+\frac12}(v_h)^-_{x,j+\frac12}\mathrm{d}x = 0 \qquad j = \gamma_2, \tag{3.9e}$$

$$\int_{I_i} E^{(\theta_2)}_{x,j+\frac12}(v_h)^-_{x,j+\frac12}\mathrm{d}x = -\widetilde{\theta}_2 \int_{I_i} \psi^+_{x,j+\frac12}(v_h)^-_{x,j+\frac12}\mathrm{d}x \quad j \in \mathbb{b}_2^+, \tag{3.9f}$$

$$\int_{I_i} E^{(\widetilde{\theta}_2)}_{x,j-\frac12}(v_h)^+_{x,j-\frac12}\mathrm{d}x = -\theta_2 \int_{I_i} \psi^+_{x,j-\frac12}(v_h)^+_{x,j-\frac12}\mathrm{d}x \quad j \in \mathbb{b}_2^-, \tag{3.9g}$$

$$E^{-,-}_{i+\frac12,j+\frac12} = 0 \qquad (i,j) = (\gamma_1,\gamma_2), \tag{3.9h}$$

$$E^{(\theta_1),-}_{i+\frac12,j+\frac12} = -\widetilde{\theta}_1 \psi^{+,-}_{i+\frac12,j+\frac12} \qquad (i,j) = (\mathbb{b}_1^+,\gamma_2), \tag{3.9i}$$

$$E^{(\widetilde{\theta}_1),-}_{i-\frac12,j+\frac12} = -\theta_1 \psi^{+,-}_{i-\frac12,j+\frac12} \qquad (i,j) = (\mathbb{b}_1^-,\gamma_2), \tag{3.9j}$$

$$E^{-,(\theta_2)}_{i+\frac12,j+\frac12} = -\widetilde{\theta}_2 \psi^{-,+}_{i+\frac12,j+\frac12} \qquad (i,j) = (\gamma_1,\mathbb{b}_2^+), \tag{3.9k}$$

$$E^{-,(\widetilde{\theta}_2)}_{i+\frac12,j-\frac12} = -\theta_2 \psi^{-,+}_{i+\frac12,j-\frac12} \qquad (i,j) = (\gamma_1,\mathbb{b}_2^-), \tag{3.9l}$$

$$E^{(\theta_1,\theta_2)}_{i+\frac12,j+\frac12} = -\psi^{(\theta_1,\theta_2)}_{i+\frac12,j+\frac12} \qquad (i,j) = (\mathbb{b}_1^+,\mathbb{b}_2^+), \tag{3.9m}$$

$$E^{(\theta_1,\tilde{\theta}_2)}_{i+\frac12,j-\frac12} = -\psi^{(\theta_1,\tilde{\theta}_2)}_{i+\frac12,j-\frac12} \qquad (i,j) = (\mathbb{b}_1^+,\mathbb{b}_2^-), \tag{3.9n}$$

$$E^{(\widetilde{\theta}_1,\theta_2)}_{i-\frac12,j+\frac12} = -\psi^{(\widetilde{\theta}_1,\theta_2)}_{i-\frac12,j+\frac12} \qquad (i,j) = (\mathbb{b}_1^-,\mathbb{b}_2^+), \tag{3.9o}$$

$$E^{(\widetilde{\theta}_1,\widetilde{\theta}_2)}_{i-\frac12,j-\frac12} = -\psi^{(\widetilde{\theta}_1,\widetilde{\theta}_2)}_{i-\frac12,j-\frac12} \qquad (i,j) = (\mathbb{b}_1^-,\mathbb{b}_2^-), \tag{3.9p}$$

which hold for all $v_h \in Q^{k-1}(K)$ and $K \in \Omega_h$.

Since $E \in V_h$, we can express the restriction of E to $K = I_i \times J_j$ in terms of the orthogonal Legendre basis functions, i.e.,

$$E|_K \triangleq E_K(x,y) = \sum_{\ell_1=0}^{k}\sum_{\ell_2=0}^{k} \alpha^{\ell_1,\ell_2}_{i,j} P_{i,\ell_1}(x)P_{j,\ell_2}(y) = \sum_{\ell_1=0}^{k}\sum_{\ell_2=0}^{k} \alpha^{\ell_1,\ell_2}_{i,j} P_{\ell_1}(\hat{x})P_{\ell_2}(\hat{y}),$$

where $P_{\ell_1}(\hat{x})$ is the ℓ_1th order Legendre polynomial on the reference element $[-1,1]$ with $\hat{x} = \frac{2(x-x_i)}{h_i^x}$; likewise for $P_{\ell_2}(\hat{y})$.

Below we will finish the proof of Lemma 3.1 with the following five steps.

Step 1 It follows from (3.9a) and the orthogonality property of the Legendre polynomials that

$$\begin{aligned} E_K(x,y) &= \sum_{\ell_2=0}^{k-1} \alpha^{k,\ell_2}_{i,j} P_k(\hat{x})P_{\ell_2}(\hat{y}) + \sum_{\ell_1=0}^{k-1} \alpha^{\ell_1,k}_{i,j} P_{\ell_1}(\hat{x})P_k(\hat{y}) + \alpha^{k,k}_{i,j} P_k(\hat{x})P_k(\hat{y}) \\ &\triangleq W_1 + W_2 + W_3, \end{aligned} \tag{3.10}$$

since $\alpha^{\ell_1,\ell_2}_{i,j} = 0$ for $\ell_1, \ell_2 = 0,\ldots,k-1$, $i \in \mathbb{Z}_{N_1}$ and $j \in \mathbb{Z}_{N_2}$.

Journal: 10915 Article No.: 1197 TYPESET DISK LE CP Disp.:2020/3/19 Pages: 19 Layout: Small

Step 2 Estimate to W_1. Taking $v_h = P_{\ell_2}(\hat{y})$ in (3.9b)–(3.9d) with $\ell_2 = 0, \ldots, k-1$ and using the orthogonality property of Legendre polynomials, we obtain consecutively

$$\alpha_{i,j}^{k,\ell_2} = 0 \qquad i = \gamma_1, \tag{3.11a}$$

$$\theta_1 \alpha_{i,j}^{k,\ell_2} + \widetilde{\theta}_1 (-1)^k \alpha_{i+1,j}^{k,\ell_2} = \widetilde{\theta}_1 g_{i+1,j}^{k,\ell_2} \qquad i \in \mathbb{b}_1^+, \tag{3.11b}$$

$$\tilde{\theta}_1 \alpha_{i-1,j}^{k,\ell_2} + \theta_1 (-1)^k \alpha_{i,j}^{k,\ell_2} = \theta_1 g_{i,j}^{k,\ell_2} \qquad i \in \mathbb{b}_1^-, \tag{3.11c}$$

for $\ell_2 = 0, \ldots, k-1$, $j \in \mathbb{Z}_{N_2}$, where $g_{i+1,j}^{k,\ell_2} = -\frac{2\ell_2+1}{2} \int_{-1}^{1} \psi_{i+\frac{1}{2},y}^{+} P_{\ell_2}(\hat{y}) \mathrm{d}\hat{y}$ with $y = y_j + \frac{h_j^y}{2}\hat{y}$. Next, a combination of (3.11a) with (3.11b) and (3.11c), respectively, gives us, for $\ell_2 = 0, \ldots, k-1$, $j \in \mathbb{Z}_{N_2}$, the linear systems of equations

$$A_{\mathbb{b}_1^+} \alpha_{\mathbb{b}_1^+,j}^{k,\ell_2} = \widetilde{\theta}_1 g_{\mathbb{b}_1^+,j}^{k,\ell_2}, \tag{3.12a}$$

$$A_{\mathbb{b}_1^-} \alpha_{\mathbb{b}_1^-,j}^{k,\ell_2} = \theta_1 g_{\mathbb{b}_1^-,j}^{k,\ell_2}, \tag{3.12b}$$

where the vectors $\alpha_{\mathbb{b}_1^+,j}^{k,\ell_2} = (\alpha_{\beta_1,j}^{k,\ell_2}, \ldots, \alpha_{\gamma_1-1,j}^{k,\ell_2})^{\mathrm{T}}$, $\alpha_{\mathbb{b}_1^-,j}^{k,\ell_2} = (\alpha_{\gamma_1+1,j}^{k,\ell_2}, \ldots, \alpha_{\beta_1-1,j}^{k,\ell_2})^{\mathrm{T}}$, $g_{\mathbb{b}_1^+,j}^{k,\ell_2} = (g_{\beta_1+1,j}^{k,\ell_2}, \ldots, g_{\gamma_1,j}^{k,\ell_2})^{\mathrm{T}}$, $g_{\mathbb{b}_1^-,j}^{k,\ell_2} = (g_{\gamma_1+1,j}^{k,\ell_2}, \ldots, g_{\beta_1-1,j}^{k,\ell_2})^{\mathrm{T}}$, and the diagonally dominant matrices

$$A_{\mathbb{b}_1^+} = \begin{pmatrix} \theta_1 & \widetilde{\theta}_1(-1)^k & & \\ & \ddots & \ddots & \\ & & \theta_1 & \widetilde{\theta}_1(-1)^k \\ & & & \theta_1 \end{pmatrix}, \quad A_{\mathbb{b}_1^-} = \begin{pmatrix} \theta_1(-1)^k & & & \\ \widetilde{\theta}_1 & \theta_1(-1)^k & & \\ & \ddots & \ddots & \\ & & \widetilde{\theta}_1 & \theta_1(-1)^k \end{pmatrix}. \tag{3.13}$$

Obviously, by (2.2a) with $\theta_1 > \frac{1}{2}$, the determinants of $A_{\mathbb{b}_1^+}$ and $A_{\mathbb{b}_1^-}$ are not zero. Thus, $\alpha_{i,j}^{k,\ell_2}$ exists uniquely for $\ell_2 = 0, \ldots, k-1$, $j \in \mathbb{Z}_{N_2}$ and $i \in \mathbb{Z}_{N_1}$.

Step 3 Estimate to W_2. Analogously, taking $v_h = P_{\ell_1}(\hat{x})$ in (3.9e)–(3.9g) with $\ell_1 = 0, \ldots, k-1$ and using the orthogonality property of Legendre polynomials, we obtain consecutively

$$\alpha_{i,j}^{\ell_1,k} = 0 \qquad j = \gamma_2, \tag{3.14a}$$

$$\theta_2 \alpha_{i,j}^{\ell_1,k} + \tilde{\theta}_2 (-1)^k \alpha_{i,j+1}^{\ell_1,k} = \widetilde{\theta}_2 g_{i,j+1}^{\ell_1,k} \qquad j \in \mathbb{b}_2^+, \tag{3.14b}$$

$$\tilde{\theta}_2 \alpha_{i,j-1}^{\ell_1,k} + \theta_2 (-1)^k \alpha_{i,j}^{\ell_1,k} = \theta_2 g_{i,j}^{\ell_1,k} \qquad j \in \mathbb{b}_2^-, \tag{3.14c}$$

for $\ell_1 = 0, \ldots, k-1$, $i \in \mathbb{Z}_{N_1}$, where $g_{i,j+1}^{\ell_1,k} = -\frac{2\ell_1+1}{2} \int_{-1}^{1} \psi_{x,j+\frac{1}{2}}^{+} P_{\ell_1}(\hat{x}) \mathrm{d}\hat{x}$ with $x = x_i + \frac{h_i^x}{2}\hat{x}$. Next, a combination of (3.14a) with (3.14b) and (3.14c), respectively, gives us, for $\ell_1 = 0, \ldots, k-1$, $i \in \mathbb{Z}_{N_1}$, the linear systems of equations

$$A_{\mathbb{b}_2^+} \alpha_{i,\mathbb{b}_2^+}^{\ell_1,k} = \widetilde{\theta}_2 g_{i,\mathbb{b}_2^+}^{\ell_1,k}, \tag{3.15a}$$

$$A_{\mathbb{b}_2^-} \alpha_{i,\mathbb{b}_2^-}^{\ell_1,k} = \theta_2 g_{i,\mathbb{b}_2^-}^{\ell_1,k}, \tag{3.15b}$$

Journal: 10915 Article No.: 1197 TYPESET DISK LE CP Disp.:2020/3/19 Pages: 19 Layout: Small

where the vectors $\alpha_{i,\mathbb{b}_2^+}^{\ell_1,k} = (\alpha_{i,\beta_2}^{\ell_1,k},\ldots,\alpha_{i,\gamma_2-1}^{\ell_1,k})^{\mathrm{T}}$, $\alpha_{i,\mathbb{b}_2^-}^{\ell_1,k} = (\alpha_{i,\gamma_2+1}^{\ell_1,k},\ldots,\alpha_{i,\beta_2-1}^{\ell_1,k})^{\mathrm{T}}$, $g_{i,\mathbb{b}_2^+}^{\ell_1,k} = (g_{i,\beta_2+1}^{\ell_1,k},\ldots,g_{i,\gamma_2}^{\ell_1,k})^{\mathrm{T}}$, $g_{i,\mathbb{b}_2^-}^{\ell_1,k} = (g_{i,\gamma_2+1}^{\ell_1,k},\ldots,g_{i,\beta_2-1}^{\ell_1,k})^{\mathrm{T}}$, and the diagonally dominant matrices

$$A_{\mathbb{b}_2^+} = \begin{pmatrix} \theta_2 & \widetilde{\theta}_2(-1)^k & & \\ & \ddots & \ddots & \\ & & \theta_2 & \widetilde{\theta}_2(-1)^k \\ & & & \theta_2 \end{pmatrix}, \quad A_{\mathbb{b}_2^-} = \begin{pmatrix} \theta_2(-1)^k & & & \\ \widetilde{\theta}_2 & \theta_2(-1)^k & & \\ & \ddots & \ddots & \\ & & \widetilde{\theta}_2 & \theta_2(-1)^k \end{pmatrix}. \tag{3.16}$$

Obviously, by (2.2b) with $\theta_2 > \frac{1}{2}$, the determinants of $A_{\mathbb{b}_2^+}$ and $A_{\mathbb{b}_2^-}$ are not zero. Thus, $\alpha_{i,j}^{\ell_1,k}$ exists uniquely for $\ell_1 = 0,\ldots,k-1$, $i \in \mathbb{Z}_{N_1}$ and $j \in \mathbb{Z}_{N_2}$.

Step 4 Estimate to W_3. By the exact collocation at $(x_{\gamma_1+\frac{1}{2}}, y_{\gamma_2+\frac{1}{2}})$ in (3.9h), we have that

$$\alpha_{\gamma_1,\gamma_2}^{k,k} = -\sum_{\ell_2=0}^{k-1}\alpha_{\gamma_1,\gamma_2}^{k,\ell_2} - \sum_{\ell_1=0}^{k-1}\alpha_{\gamma_1,\gamma_2}^{\ell_1,k} = 0, \tag{3.17a}$$

since, by (3.11a) and (3.14a), $\alpha_{\gamma_1,j}^{k,\ell_2} = 0$ for $\ell_2 = 0,\ldots,k-1$, $j \in \mathbb{Z}_{N_2}$ and $\alpha_{i,\gamma_2}^{\ell_1,k} = 0$ for $\ell_1 = 0,\ldots,k-1$, $i \in \mathbb{Z}_{N_1}$.

Then, the conditions (3.9i) and (3.9j) imply that

$$\theta_1\alpha_{i,\gamma_2}^{k,k} + \widetilde{\theta}_1(-1)^k\alpha_{i+1,\gamma_2}^{k,k} = g_{i+1,\gamma_2}^{k,k} \qquad i \in \mathbb{b}_1^+, \tag{3.17b}$$

$$\widetilde{\theta}_1\alpha_{i-1,\gamma_2}^{k,k} + \theta_1(-1)^k\alpha_{i,\gamma_2}^{k,k} = g_{i,\gamma_2}^{k,k} \qquad i \in \mathbb{b}_1^-, \tag{3.17c}$$

where

$$\begin{aligned} g_{i+1,\gamma_2}^{k,k} &= -\widetilde{\theta}_1\psi_{i+\frac{1}{2},\gamma_2+\frac{1}{2}}^{+,-} - \theta_1\left(\sum_{\ell_2=0}^{k-1}\alpha_{i,\gamma_2}^{k,\ell_2} + \sum_{\ell_1=0}^{k-1}\alpha_{i,\gamma_2}^{\ell_1,k}\right) \\ &\quad - \widetilde{\theta}_1\left(\sum_{\ell_2=0}^{k-1}\alpha_{i+1,\gamma_2}^{k,\ell_2}(-1)^k + \sum_{\ell_1=0}^{k-1}\alpha_{i+1,\gamma_2}^{\ell_1,k}(-1)^{\ell_1}\right) \qquad i \in \mathbb{b}_1^+, \\ g_{i,\gamma_2}^{k,k} &= -\theta_1\psi_{i-\frac{1}{2},\gamma_2+\frac{1}{2}}^{+,-} - \theta_1\left(\sum_{\ell_2=0}^{k-1}\alpha_{i-1,\gamma_2}^{k,\ell_2} + \sum_{\ell_1=0}^{k-1}\alpha_{i-1,\gamma_2}^{\ell_1,k}\right) \\ &\quad - \widetilde{\theta}_1\left(\sum_{\ell_2=0}^{k-1}\alpha_{i,\gamma_2}^{k,\ell_2}(-1)^k + \sum_{\ell_1=0}^{k-1}\alpha_{i,\gamma_2}^{\ell_1,k}(-1)^{\ell_1}\right) \qquad i \in \mathbb{b}_1^-. \end{aligned}$$

Inserting (3.17a) into (3.17b) and (3.17c), we obtain, for $j = \gamma_2$, the linear systems of equations

$$A_{\mathbb{b}_1^+}\alpha_{\mathbb{b}_1^+,\gamma_2} = g_{\mathbb{b}_1^+,\gamma_2}, \tag{3.18a}$$

$$A_{\mathbb{b}_1^-}\alpha_{\mathbb{b}_1^-,\gamma_2} = g_{\mathbb{b}_1^-,\gamma_2}, \tag{3.18b}$$

where $\alpha_{\mathbb{b}_1^+,\gamma_2} = (\alpha_{\beta_1,\gamma_2}^{k,k},\ldots,\alpha_{\gamma_1-1,\gamma_2}^{k,k})^{\mathrm{T}}$, $\alpha_{\mathbb{b}_1^-,\gamma_2} = (\alpha_{\gamma_1+1,\gamma_2}^{k,k},\ldots,\alpha_{\beta_1-1,\gamma_2}^{k,k})^{\mathrm{T}}$, $g_{\mathbb{b}_1^+,\gamma_2} = (g_{\beta_1+1,\gamma_2}^{k,k},\ldots,g_{\gamma_1,\gamma_2}^{k,k})^{\mathrm{T}}$, $g_{\mathbb{b}_1^-,\gamma_2} = (g_{\gamma_1+1,\gamma_2}^{k,k},\ldots,g_{\beta_1-1,\gamma_2}^{k,k})^{\mathrm{T}}$, and the diagonally dominant matrices $A_{\mathbb{b}_1^+}$, $A_{\mathbb{b}_1^-}$ have been given in (3.13). Therefore, $\alpha_{i,\gamma_2}^{k,k}$ exists uniquely for $i \in \mathbb{Z}_{N_1}$.

Journal: 10915 Article No.: 1197 TYPESET DISK LE CP Disp.:2020/3/19 Pages: 19 Layout: Small

Similarly, the conditions (3.9k) and (3.9l) together with (3.17a) produce, for $i = \gamma_1$, the linear systems of equations

$$A_{\mathbb{b}_2^+}\alpha_{\gamma_1,\mathbb{b}_2^+} = g_{\gamma_1,\mathbb{b}_2^+}, \tag{3.19a}$$

$$A_{\mathbb{b}_2^-}\alpha_{\gamma_1,\mathbb{b}_2^-} = g_{\gamma_1,\mathbb{b}_2^-}, \tag{3.19b}$$

where $\alpha_{\gamma_1,\mathbb{b}_2^+} = (\alpha^{k,k}_{\gamma_1,\beta_2},\ldots,\alpha^{k,k}_{\gamma_1,\gamma_2-1})^{\mathrm{T}}$, $\alpha_{\gamma_1,\mathbb{b}_2^-} = (\alpha^{k,k}_{\gamma_1,\gamma_2+1},\ldots,\alpha^{k,k}_{\gamma_1,\beta_2-1})^{\mathrm{T}}$, $g_{\gamma_1,\mathbb{b}_2^+} = (g^{k,k}_{\gamma_1,\beta_2+1},\ldots,g^{k,k}_{\gamma_1,\gamma_2})^{\mathrm{T}}$, $g_{\gamma_1,\mathbb{b}_2^-} = (g^{k,k}_{\gamma_1,\gamma_2+1},\ldots,g^{k,k}_{\gamma_1,\beta_2-1})^{\mathrm{T}}$, and $A_{\mathbb{b}_2^+}$, $A_{\mathbb{b}_2^-}$ have been given in (3.16). Therefore, $\alpha^{k,k}_{\gamma_1,j}$ exists uniquely for $j \in \mathbb{Z}_{N_2}$.

In what follows, we shall deal with some more complicated terms involving two weights in (3.9m)–(3.9p). Since the analysis to (3.9m)–(3.9p) are similar, we only take (3.9m) as an example. After rearranging terms, the condition (3.9m) yields, for $i \in \mathbb{b}_1^+$ and $j \in \mathbb{b}_2^+$, that

$$\theta_1\theta_2\alpha^{k,k}_{i,j} + \theta_1\widetilde{\theta}_2(-1)^k\alpha^{k,k}_{i,j+1} + \widetilde{\theta}_1\theta_2(-1)^k\alpha^{k,k}_{i+1,j} + \widetilde{\theta}_1\widetilde{\theta}_2\alpha^{k,k}_{i+1,j+1} = g^{k,k}_{i,j}, \tag{3.20}$$

where

$$\begin{aligned} g^{k,k}_{i,j} = &-\psi^{(\theta_1,\theta_2)}_{i+\frac{1}{2},j+\frac{1}{2}} - \theta_1\theta_2\left(\sum_{\ell_2=0}^{k-1}\alpha^{k,\ell_2}_{i,j} + \sum_{\ell_1=0}^{k-1}\alpha^{\ell_1,k}_{i,j}\right)\\ &- \theta_1\widetilde{\theta}_2\left(\sum_{\ell_2=0}^{k-1}\alpha^{k,\ell_2}_{i,j+1}(-1)^{\ell_2} + \sum_{\ell_1=0}^{k-1}\alpha^{\ell_1,k}_{i,j+1}(-1)^k\right)\\ &- \widetilde{\theta}_1\theta_2\left(\sum_{\ell_2=0}^{k-1}\alpha^{k,\ell_2}_{i+1,j}(-1)^k + \sum_{\ell_1=0}^{k-1}\alpha^{\ell_1,k}_{i+1,j}(-1)^{\ell_1}\right)\\ &- \widetilde{\theta}_1\widetilde{\theta}_2\left(\sum_{\ell_2=0}^{k-1}\alpha^{k,\ell_2}_{i+1,j+1}(-1)^{k+\ell_2} + \sum_{\ell_1=0}^{k-1}\alpha^{\ell_1,k}_{i+1,j+1}(-1)^{k+\ell_1}\right) \end{aligned} \tag{3.21}$$

is known. If we now denote

$$\alpha_{\mathbb{b}_1^+,\mathbb{b}_2^+} = (\alpha^{k,k}_{\beta_1,\beta_2},\ldots,\alpha^{k,k}_{\beta_1,\gamma_2-1},\ldots,\alpha^{k,k}_{\gamma_1-1,\beta_2},\ldots,\alpha^{k,k}_{\gamma_1-1,\gamma_2-1})^{\mathrm{T}},$$

$$g_{\mathbb{b}_1^+,\mathbb{b}_2^+} = (g^{k,k}_{\beta_1,\beta_2},\ldots,g^{k,k}_{\beta_1,\gamma_2-1},\ldots,g^{k,k}_{\gamma_1-1,\beta_2},\ldots,g^{k,k}_{\gamma_1-1,\gamma_2-1})^{\mathrm{T}},$$

then (3.20) can be rewritten as

$$A_{\mathbb{b}_1^+} \otimes A_{\mathbb{b}_2^+}\, \alpha_{\mathbb{b}_1^+,\mathbb{b}_2^+} = g_{\mathbb{b}_1^+,\mathbb{b}_2^+}, \tag{3.22}$$

where $A_{\mathbb{b}_1^+}$, $A_{\mathbb{b}_2^+}$ have been defined in (3.13) and (3.16), and $\otimes$ is the Kronecker product of two matrices. Since $A_{\mathbb{b}_1^+}$ and $A_{\mathbb{b}_2^+}$ are invertible, we can deduce from

$$(A_{\mathbb{b}_1^+} \otimes A_{\mathbb{b}_2^+})^{-1} = A^{-1}_{\mathbb{b}_1^+} \otimes A^{-1}_{\mathbb{b}_2^+}$$

that $A_{\mathbb{b}_1^+} \otimes A_{\mathbb{b}_2^+}$ is also invertible. Therefore, $\alpha^{k,k}_{i,j}$ exists uniquely for $i \in \mathbb{b}_1^+$, $j \in \mathbb{b}_2^+$. Applying the same arguments as that for (3.9m) to (3.9n)–(3.9p), we conclude that $\alpha^{k,k}_{i,j}$ exists uniquely for $i \in \mathbb{b}_1^+$, $j \in \mathbb{b}_2^-$ and $i \in \mathbb{b}_1^-$, $j \in \mathbb{b}_2^+ \cup \mathbb{b}_2^-$.

Till now, we have proved that $\alpha^{\ell_1,\ell_2}_{i,j}$ can be solved for $\ell_1, \ell_2 = 0,\ldots,k$ and $i \in \mathbb{Z}_{N_1}$, $j \in \mathbb{Z}_{N_2}$; then $E_K(x,y)$ and thus $\Pi_h^{\theta_1,\theta_2}u$ is uniquely determined on each element $K \in \Omega_h$.

Journal: 10915 Article No.: 1197 TYPESET DISK LE CP Disp.:2020/3/19 Pages: 19 Layout: Small

Step 5 Optimal approximation property. The optimal approximation property of $\Pi_h^{\theta_1,\theta_2}$ can be derived from that of E, and, by (3.10), we need only to consider the bounds for $\alpha_{i,j}^{k,\ell_2}$, $\alpha_{i,j}^{\ell_1,k}$ and $\alpha_{i,j}^{k,k}$ with $\ell_1, \ell_2 = 0, \ldots, k-1$, $i \in \mathbb{Z}_{N_1}$, $j \in \mathbb{Z}_{N_2}$.

Firstly, we estimate the coefficients in W_1, i.e., $\alpha_{i,j}^{k,\ell_2}$. To do that, we solve (3.12a) and use the special form of $A_{\mathbb{b}_1^+}^{-1}$ in [15, Appendix A] (which is an upper triangular matrix) to get

$$\begin{aligned}\|\alpha_{\mathbb{b}_1^+,j}^{k,\ell_2}\|_2^2 &\le \widetilde{\theta}_1^2 \|A_{\mathbb{b}_1^+}^{-1}\|_2^2 \, \|g_{\mathbb{b}_1^+,j}^{k,\ell_2}\|_2^2 \le \widetilde{\theta}_1^2 \|A_{\mathbb{b}_1^+}^{-1}\|_1 \, \|A_{\mathbb{b}_1^+}^{-1}\|_\infty \, \|g_{\mathbb{b}_1^+,j}^{k,\ell_2}\|_2^2 \\ &\le \frac{q_1^2}{(1-|q_1|)^2} \, \|g_{\mathbb{b}_1^+,j}^{k,\ell_2}\|_2^2, \end{aligned} \tag{3.23a}$$

where $q_1 = -\frac{\widetilde{\theta}_1(-1)^k}{\theta_1}$ with $|q_1| < 1$, and $\|\cdot\|_p$ denotes the ℓ^p norm for a vector or matrix with $p = 1, 2, \infty$. Moreover, it follows from the Cauchy–Schwarz inequality and the change of variables that

$$\|g_{\mathbb{b}_1^+,j}^{k,\ell_2}\|_2^2 \le Ch^{-1} \sum_{i \in \mathbb{b}_1^+} \int_{J_j} (\psi_{i+\frac{1}{2},y}^+)^2 \mathrm{d}y \le Ch^{-1} \sum_{i \in \mathbb{b}_1^+} \|\psi\|_{\partial K_R}^2, \tag{3.23b}$$

with $K_R = I_{i+1} \times J_j$. A combination of (3.23a) and (3.23b) gives us

$$\|\alpha_{\mathbb{b}_1^+,j}^{k,\ell_2}\|_2^2 \le Ch^{-1} \sum_{i \in \mathbb{b}_1^+} \|\psi\|_{\partial K_R}^2. \tag{3.24a}$$

Analogously, for (3.12b), we have

$$\|\alpha_{\mathbb{b}_1^-,j}^{k,\ell_2}\|_2^2 \le \frac{1}{(1-|q_2|)^2} \, \|g_{\mathbb{b}_1^-,j}^{k,\ell_2}\|_2^2 \le Ch^{-1} \sum_{i \in \mathbb{b}_1^-} \|\psi\|_{\partial K}^2, \tag{3.24b}$$

where $q_2 = -\frac{\widetilde{\theta}_2(-1)^k}{\theta_2}$ with $|q_2| < 1$. If we now denote $\alpha_j^{k,\ell_2} = ((\alpha_{\mathbb{b}_1^+,j}^{k,\ell_2})^{\mathrm{T}}, 0, (\alpha_{\mathbb{b}_1^-,j}^{k,\ell_2})^{\mathrm{T}})^{\mathrm{T}}$ with $j \in \mathbb{Z}_{N_2}$, $\ell_2 = 0, \ldots, k-1$, we arrive at

$$\|\alpha_j^{k,\ell_2}\|_2^2 = \|\alpha_{\mathbb{b}_1^+,j}^{k,\ell_2}\|_2^2 + \|\alpha_{\mathbb{b}_1^-,j}^{k,\ell_2}\|_2^2 \le Ch^{-1} \sum_{i=1}^{N_1} \|\psi\|_{\partial K}^2. \tag{3.25}$$

Secondly, performing the same procedure as that in deriving (3.25) to (3.15), we obtain the bound for the coefficients in W_2, namely $\alpha_{i,j}^{\ell_1,k}$. It reads

$$\|\alpha_i^{\ell_1,k}\|_2^2 = \|\alpha_{i,\mathbb{b}_2^+}^{\ell_1,k}\|_2^2 + \|\alpha_{i,\mathbb{b}_2^-}^{\ell_1,k}\|_2^2 \le Ch^{-1} \sum_{j=1}^{N_2} \|\psi\|_{\partial K}^2, \tag{3.26}$$

where $\alpha_i^{\ell_1,k} = ((\alpha_{i,\mathbb{b}_2^+}^{\ell_1,k})^{\mathrm{T}}, 0, (\alpha_{i,\mathbb{b}_2^-}^{\ell_1,k})^{\mathrm{T}})^{\mathrm{T}}$ with $i \in \mathbb{Z}_{N_1}$, $\ell_1 = 0, \ldots, k-1$.

Thirdly, let us consider estimates to the coefficients in W_3, i.e., $\alpha_{i,j}^{k,k}$. By an argument similar to that in the proof of (3.24a), we deduce from (3.22) that

$$\begin{aligned} &\|\alpha_{\mathbb{b}_1^+,\mathbb{b}_2^+}\|_2^2 \\ &\quad \le \|(A_{\mathbb{b}_1^+} \otimes A_{\mathbb{b}_2^+})^{-1}\|_2^2 \, \|g_{\mathbb{b}_1^+,\mathbb{b}_2^+}\|_2^2 \\ &\quad \le \|A_{\mathbb{b}_1^+}^{-1} \otimes A_{\mathbb{b}_2^+}^{-1}\|_1 \, \|A_{\mathbb{b}_1^+}^{-1} \otimes A_{\mathbb{b}_2^+}^{-1}\|_\infty \, \|g_{\mathbb{b}_1^+,\mathbb{b}_2^+}\|_2^2 \end{aligned}$$

$$
\begin{aligned}
&\leq \frac{Cq_1^2}{(1-|q_1|)^2(1-|q_2|)^2}\left(\sum_{K\in\Omega_h}\|\psi\|_{\infty,K}^2+\sum_{\ell_2=0}^{k-1}\sum_{j=1}^{N_2}\|\alpha_j^{k,\ell_2}\|_2^2+\sum_{\ell_1=0}^{k-1}\sum_{i=1}^{N_1}\|\alpha_i^{\ell_1,k}\|_2^2\right)\\
&\leq C\sum_{K\in\Omega_h}\left(h^{2k}\|u\|_{k+1,K}^2+h^{-1}\|\psi\|_{\partial K}^2\right)=C(h^{2k}\|u\|_{k+1}^2+h^{-1}\|\psi\|_{\partial\Omega_h}^2)\\
&\leq Ch^{2k}, \qquad (3.27)
\end{aligned}
$$

where in the second line we have used the fact that $(A_{\mathbb{b}_1^{+}}\otimes A_{\mathbb{b}_2^{+}})^{-1}=A_{\mathbb{b}_1^{+}}^{-1}\otimes A_{\mathbb{b}_2^{+}}^{-1}$ as well as the Hölder's inequality for the matrix norm, in the third line we have utilized $\left(\sum_{\ell_2=0}^{k-1}\alpha_{i,j}^{k,\ell_2}\right)^2\leq k\sum_{\ell_2=0}^{k-1}\left(\alpha_{i,j}^{k,\ell_2}\right)^2$ and have substituted (3.25), (3.26) into (3.21), in the fourth line we have employed the property $\|\psi\|_{\infty,K}\leq Ch^k\|u\|_{k+1,K}$ implied by the Sobolev inequality, the Bramble–Hilbert lemma and scaling arguments in [2, Corollary 4.4.7], and in the last line we have taken into account the approximation result in (3.7). Similar bounds for $\|\alpha_{\mathbb{b}_1^{+},\mathbb{b}_2^{-}}\|_2^2$, $\|\alpha_{\mathbb{b}_1^{-},\mathbb{b}_2^{+}}\|_2^2$, and $\|\alpha_{\mathbb{b}_1^{-},\mathbb{b}_2^{-}}\|_2^2$ can also be shown.

Finally, we are now ready to present the optimal approximation property for $\Pi_h^{\theta_1,\theta_2}$. Collecting (3.25)–(3.27) into (3.10), we have

$$
\begin{aligned}
\|E\|^2 &\leq Ch^2\left(\sum_{K\in\Omega_h}\sum_{\ell_2=0}^{k-1}(\alpha_{i,j}^{k,\ell_2})^2+\sum_{K\in\Omega_h}\sum_{\ell_1=0}^{k-1}(\alpha_{i,j}^{\ell_1,k})^2+h^{2k}\right)\\
&\leq Ch^2\left(h^{-1}\|\psi\|_{\partial\Omega_h}^2+h^{2k}\right)\\
&\leq Ch^{2k+2},
\end{aligned}
$$

where we have also used the interpolation error estimate (3.7). This, together with the triangle inequality, leads to the desired result (3.8). Also, the boundary norm estimate can be derived by the inverse property (ii). The proof of Lemma 3.1 is complete. □

Remark 3.1 For the special case that $a(x)$ keeps its sign and $b(y)$ changes its sign on I, we can modify the projection to be the tensor product of $P_h^{\star}$ in [20, Lemma 2.6] and $\mathcal{P}_h$ in [15, Lemma 3.1], and similar conclusions as that in Lemma 3.1 can be obtained. □

3.3 A Sharp Bound for Projection Error Terms

Due to the lack of degrees of freedom in defining projections, the projection error terms cannot be eliminated. However, the following sharp bound of the projection $\Pi_h^{\theta_1,\theta_2}$ helps to recover the order for the leading term of the projection error. Denote by a_L a piecewise constant with $a_L|_{I_i}=a(x_i)\triangleq a_i$; likewise for b_L.

Lemma 3.2 *Assume that $u\in H^{k+2}(\Omega)$ and $v_h\in V_h$. Then we have*

$$
\left|\mathcal{H}^x(a_L(u-\Pi_h^{\theta_1,\theta_2}u),v_h)+\mathcal{H}^y(b_L(u-\Pi_h^{\theta_1,\theta_2}u),v_h)\right|\leq Ch^{k+1}\|u\|_{k+2}\|v_h\|, \qquad (3.28)
$$

where C is independent of h.

Proof The proof is similar to that in [4] in which linear convection–diffusion equations with alternating fluxes are considered. We only point out the main differences. Without loss of generality, in what follows we only concentrate on the bound for $\mathcal{H}^x(a_L(u-\Pi_h^{\theta_1,\theta_2}u),v_h)$. In contrast to a local identity in [7, Lemma 3.6], here we have a global inequality

$$
\mathcal{H}^x(a_L(w-\Pi_h^{\theta_1,\theta_2}w),v_h)\leq Ch^{k+\frac{3}{2}}\|v_h\| \quad \forall w\in P^{k+1}(\Omega_h),\ v_h\in Q^k(\Omega_h), \qquad (3.29)
$$

Journal: 10915 Article No.: 1197 TYPESET DISK LE CP Disp.:2020/3/19 Pages: 19 Layout: Small

since the one-dimensional projection $\mathcal{P}_h^{\theta}u$ does not enforce any collocation condition at the point $x_{\beta-\frac{1}{2}}$. Noting that $\mathcal{H}^x(a_L(w-\Pi_h^{\theta_1,\theta_2}w), v_h) = 0$ for $w \in P^k(K)$, as $\Pi_h^{\theta_1,\theta_2}$ is a polynomial preserving operator, to prove (3.29) we need only to consider $w|_K = x^{k+1}$ and $w|_K = y^{k+1}$. Specifically, for $w|_K = x^{k+1}$, since $\Pi_h^{\theta_1,\theta_2}$ reduces to a one-dimensional projection $\mathcal{P}_{h_x}^{\theta_1}$ for the univariate function $w = x^{k+1}$ and, by (3.5c)–(3.5d), $(w-\Pi_h^{\theta_1,\theta_2}w)_{\beta_1-\frac{1}{2},y} = (w-\mathcal{P}_{h_x}^{\theta_1}w)_{\beta_1-\frac{1}{2}} \neq 0$, we conclude that

$$\mathcal{H}^x_{I_{\beta_1}\times J_j}(a_L(w-\Pi_h^{\theta_1,\theta_2}w), v_h) = a_{\beta_1}(w-\mathcal{P}_{h_x}^{\theta_1}w)^{(\theta_1)}_{\beta_1-\frac{1}{2}}\int_{J_j}(v_h)^+_{\beta_1-\frac{1}{2},y}\mathrm{d}y,$$

$$\mathcal{H}^x_{I_{\beta_1-1}\times J_j}(a_L(w-\Pi_h^{\theta_1,\theta_2}w), v_h) = -a_{\beta_1-1}(w-\mathcal{P}_{h_x}^{\theta_1}w)^{(\theta_1)}_{\beta_1-\frac{1}{2}}\int_{J_j}(v_h)^-_{\beta_1-\frac{1}{2},y}\mathrm{d}y,$$

and for other elements, i.e. $\forall K \in \Omega_h\backslash\{(I_{\beta_1}\cup I_{\beta_1-1})\times J_j\}$,

$$\mathcal{H}^x_K(a_L(w-\Pi_h^{\theta_1,\theta_2}w), v_h) = 0.$$

In addition, for $w|_K = y^{k+1}$, after using integration by parts

$$\mathcal{H}^x_K(a_L(w-\Pi_h^{\theta_1,\theta_2}w), v_h) = 0, \quad \forall K \in \Omega_h.$$

For more details, see [4, Appendix A]. Therefore, summing over all K, we obtain for $w \in P^{k+1}(K)$

$$\begin{aligned}\mathcal{H}^x(a_L(w-\Pi_h^{\theta_1,\theta_2}w), v_h) &= \mathcal{H}^x_{(I_{\beta_1}\cup I_{\beta_1-1})\times J_j}(a_L(w-\Pi_h^{\theta_1,\theta_2}w), v_h)\\ &\le Chh^{k+\frac{1}{2}}h^{\frac{1}{2}}\|v_h\|_{\partial\Omega_h}\\ &\le Ch^{k+\frac{3}{2}}\|v_h\|,\end{aligned}$$

where $C = C(\|w\|_{k+1})$ with $\|w\|_{k+1}$ being the broken Sobolev norm of w and in the second step we have used the approximation result for $\mathcal{P}_{h_x}^{\theta_1}$, the Cauchy–Schwarz inequality and the fact that $|a_{\beta_1}| + |a_{\beta_1-1}| \le Ch$, and in the last step we have employed the inverse property (ii).

Next, we use the inverse inequalities (i) and (ii) in combination with the optimal approximation property for $\Pi_h^{\theta_1,\theta_2}$ with $k=0$ in (3.8) to get

$$\left|\mathcal{H}^x(a_L(u-\Pi_h^{\theta_1,\theta_2}u), v_h)\right| \le C\|u\|_1\|v_h\|. \tag{3.30}$$

Consequently,

$$\begin{aligned}&\left|\mathcal{H}^x(a_L(u-\Pi_h^{\theta_1,\theta_2}u), v_h)\right|\\ &\le \left|\mathcal{H}^x(a_L((u-w)-\Pi_h^{\theta_1,\theta_2}(u-w)), v_h)\right| + \left|\mathcal{H}^x(a_L(w-\Pi_h^{\theta_1,\theta_2}w), v_h)\right|\\ &\le C\inf_{w\in P^{k+1}(\Omega_h)}\|u-w\|_1\|v_h\| + Ch^{k+\frac{3}{2}}\|v_h\|\\ &\le Ch^{k+1}\|u\|_{k+2}\|v_h\|,\end{aligned}$$

where we in the first step we have added and subtracted $\mathcal{H}^x(a_L(w-\Pi_h^{\theta_1,\theta_2}w), v_h)$ for all $w \in P^{k+1}(\Omega_h)$, and in the second step we have taken into account (3.30) and (3.29), and in the last step we have employed the standard approximation theory. This finishes the proof of Lemma 3.2. □

3.4 Optimal Error Estimates

Let us now show our main result regarding the optimal error estimates. Denote $e = u - u_h = u - \Pi_h^{\theta_1,\theta_2} u + \Pi_h^{\theta_1,\theta_2} u - u_h \triangleq \eta + \xi$ with $\xi \in V_h$.

Theorem 3.1 (Error estimate) *Assume that $u \in H^{k+2}(\Omega)$, $u_t \in H^{k+1}(\Omega)$. Let u_h be the numerical solution of the DG scheme (2.1) with upwind-biased numerical fluxes (2.2a), (2.2b). For any regular mesh, if the discontinuous finite element space V_h of degree k is used, then there holds the error estimate*

$$\|u(t) - u_h(t)\| \le Ch^{k+1}, \quad \forall t \in (0, T], \tag{3.31}$$

where C is independent of the mesh size h.

Proof By Galerkin orthogonality and using the DG operator in (2.3), we have the cell error equation

$$\int_K e_t v_h \mathrm{d}x\mathrm{d}y = \mathcal{H}_K^x(ae, v_h) + \mathcal{H}_K^y(be, v_h)$$

for any $v_h \in V_h$ and $K \in \Omega_h$. Taking $v_h = \xi$ and summing over all K, we get

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\xi\|^2 + \int_{\Omega_h} \eta_t \xi \mathrm{d}x\mathrm{d}y = \mathcal{H}^x(a\xi, \xi) + \mathcal{H}^y(b\xi, \xi) + \mathcal{H}^x(a\eta, \xi) + \mathcal{H}^y(b\eta, \xi). \tag{3.32}$$

Using the same arguments as that in the proof of the stability property in Proposition 2.1, we have that

$$\mathcal{H}^x(a\xi, \xi) + \mathcal{H}^y(b\xi, \xi) \le C\|\xi\|^2, \tag{3.33a}$$

since $\theta_1, \theta_2 > \frac{1}{2}$.

Let us now consider the estimate to $\mathcal{H}^x(a\eta, \xi) + \mathcal{H}^y(b\eta, \xi)$. Using a local linearization for $a(x) = a(x) - a_L + a_L$ and $b(y) = b(y) - b_L + b_L$, we obtain

$$\begin{aligned}
&\mathcal{H}^x(a\eta, \xi) + \mathcal{H}^y(b\eta, \xi) \\
&\quad = \mathcal{H}^x((a - a_L)\eta, \xi) + \mathcal{H}^y((b - b_L)\eta, \xi) + \mathcal{H}^x(a_L\eta, \xi) + \mathcal{H}^y(b_L\eta, \xi) \\
&\quad \le Ch\left(\|\eta\|(\|\xi_x\| + \|\xi_y\|) + \|\eta\|_{\partial\Omega_h}\|\xi\|_{\partial\Omega_h}\right) + Ch^{k+1}\|\xi\| \\
&\quad \le C\left(\|\eta\| + h^{\frac{1}{2}}\|\eta\|_{\partial\Omega_h}\right)\|\xi\| + Ch^{k+1}\|\xi\| \\
&\quad \le Ch^{k+1}\|\xi\|,
\end{aligned} \tag{3.33b}$$

where we have also used the inverse inequalities (i), (ii), the sharp bound in Lemma 3.2 and the optimal approximation property for $\Pi_h^{\theta_1,\theta_2}$ in (3.8).

Collecting (3.33a) and (3.33b) into (3.32) together with the fact that the projection $\Pi_h^{\theta_1,\theta_2}$ is linear and independent of t, namely $\|\eta_t\| \le Ch^{k+1}\|u_t\|_{k+1}$, we have

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\xi\|^2 \le C\|\xi\|^2 + Ch^{2k+2},$$

where we have also used the Cauchy–Schwarz inequality and Young's inequality. Since the numerical initial condition is taken as an L^2 projection of u_0, then a simple application of Gronwall's inequality and the triangle inequality gives us (3.31). This completes the proof of Theorem 3.1. □

Remark 3.2 For the case of $a(x)$ or $b(y)$ having more zeros, we can use the same approach as that in [15, Lemma 3.3] to construct a special projection for 2D equations. The optimal

Journal: 10915 Article No.: 1197 TYPESET DISK LE CP Disp.:2020/3/19 Pages: 19 Layout: Small

Table 1 The errors $\|u - u_h\|$ and orders for Example 4.1 using Q^k polynomials with different (θ_1, θ_2) on a random mesh of $N_1 \times N_2$ cells. $T = 1$. $CFL = 0.1$

$N_1 \times N_2$	$(\theta_1, \theta_2) = (0.7, 0.7)$		$(\theta_1, \theta_2) = (0.7, 1.5)$		$(\theta_1, \theta_2) = (1.5, 1.5)$	
	L^2 error	Order	L^2 error	Order	L^2 error	Order
Q^1						
10 × 10	3.71E−02	–	2.45E−02	–	1.90E−02	–
20 × 20	1.09E−02	1.87	7.50E−02	1.81	4.68E−03	2.14
40 × 40	2.95E−03	1.94	2.14E−03	1.86	1.20E−03	2.02
80 × 80	7.55E−04	2.03	5.60E−04	2.00	2.94E−04	2.10
160 × 160	1.89E−04	2.01	1.47E−04	1.99	7.27E−05	2.03
Q^2						
10 × 10	1.05E−03	–	1.75E−03	–	2.05E−03	–
20 × 20	1.30E−04	3.20	2.03E−04	3.29	2.69E−04	3.10
40 × 40	1.71E−05	3.02	2.66E−05	3.03	3.39E−05	3.08
80 × 80	2.00E−06	3.20	3.28E−06	3.12	4.15E−06	3.14
160 × 160	2.61E−07	3.03	4.28E−07	3.04	5.32E−07	3.06

approximation result as well as a sharp bound for projection error terms will also be obtained. The optimal error estimates will still hold. Details are omitted to save space. □

4 Numerical Experiments

In this section, a numerical example is given to demonstrate the sharpness of optimal error estimates in Theorem 3.1. To reduce time errors, the five stage fourth order strong stability preserving Runge–Kutta discretizations [14] are employed and $\Delta t = CFL\, h_{\min}$. A nonuniform mesh is used, which is a 10% random perturbation of the uniform mesh. Periodic boundary conditions are considered.

Example 4.1

$$\begin{aligned} &u_t + (a(x,y)u)_x + (b(x,y)u)_y = g(x,y,t), && (x,y,t) \in [0,2\pi]^2 \times (0,T], \\ &u(x,y,0) = u_0(x,y), && (x,y) \in [0,2\pi]^2, \end{aligned} \tag{4.1}$$

where $a(x, y) = \sin(x + y)$, $b(x, y) = \cos(x + y)$, $g(x, y, t)$ is chosen such that the exact solution of (4.1) is

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

Different combinations of the weights (θ_1, θ_2) are taken, and the results for the L^2 errors are given in Table 1, from which we can observe the expected optimal $(k+1)$th order. Moreover, for the fixed mesh, it seems that for even (odd) values of k, smaller (bigger) weights would lead to a better approximation with a smaller magnitude of the error. This may come from the different dispersive and diffusive errors of the DG scheme with upwind-biased fluxes.

Journal: 10915 Article No.: 1197 TYPESET DISK LE CP Disp.:2020/3/19 Pages: 19 Layout: Small

5 Concluding Remarks

In this paper, we analyze the DG scheme with upwind-biased fluxes for two-dimensional linear hyperbolic equations with variable coefficients on Cartesian meshes. By constructing a special piecewise *global* projection, we derive the existence and optimal approximation property of the projection. The main technicality is an elaborate treatment for the boundary collocation terms, for which couplings from different directions should be clarified and estimated. Moreover, due to the tensor product structure of the mesh and basis functions, a sharp bound for the leading error of projection error terms is shown. Therefore, optimal error estimates are obtained. Numerical experiments are presented to verify the theoretical results. Extensions to multivariate linear variable coefficient equations and the 2D nonlinear equations are challenging, which constitute of our future work.

Acknowledgements The authors thank referees for their valuable suggestions that result in the improvement of the paper.

References

1. Bona, J.L., Chen, H., Karakashian, O., Xing, Y.: Conservative, discontinuous Galerkin-methods for the generalized Korteweg–de Vries equation. Math. Comp. **82**(283), 1401–1432 (2013). https://doi.org/10.1090/S0025-5718-2013-02661-0
2. Brenner, S., Scott, R.: The Mathematical Theory of Finite Element Methods, vol. 15. Springer, Berlin (2008)
3. Cao, W., Li, D., Yang, Y., Zhang, Z.: Superconvergence of discontinuous Galerkin methods based on upwind-biased fluxes for 1D linear hyperbolic equations. ESAIM Math. Model. Numer. Anal. **51**(2), 467–486 (2017). https://doi.org/10.1051/m2an/2016026
4. Cheng, Y., Meng, X., Zhang, Q.: Application of generalized Gauss-Radau projections for the local discontinuous Galerkin method for linear convection-diffusion equations. Math. Comp. **86**(305), 1233–1267 (2017). https://doi.org/10.1090/mcom/3141
5. Cheng, Y., Shu, C.W.: A discontinuous Galerkin finite element method for time dependent partial differential equations with higher order derivatives. Math. Comp. **77**(262), 699–730 (2008). https://doi.org/10.1090/S0025-5718-07-02045-5
6. Cockburn, B., Hou, S., Shu, C.W.: The Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws IV. The multidimensional case. Math. Comp. **54**(190), 545–581 (1990). https://doi.org/10.2307/2008501
7. Cockburn, B., Kanschat, G., Perugia, I., Schötzau, D.: Superconvergence of the local discontinuous Galerkin method for elliptic problems on Cartesian grids. SIAM J. Numer. Anal. **39**(1), 411–435 (2001). https://doi.org/10.1137/S0036142900371544
8. Cockburn, B., Karniadakis, G.E., Shu, C.W.: Discontinuous Galerkin Methods: Theory, Computation and Applications, vol. 11. Springer, Berlin (2012)
9. Cockburn, B., Lin, S.Y., Shu, C.W.: TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws III. One-dimensional systems. J. Comput. Phys. **84**(1), 90–113 (1989). https://doi.org/10.1016/0021-9991(89)90183-6
10. Cockburn, B., Shu, C.W.: TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws II. General framework. Math. Comp. **52**(186), 411–435 (1989). https://doi.org/10.2307/2008474
11. Cockburn, B., Shu, C.W.: The local discontinuous Galerkin method for time-dependent convection–diffusion systems. SIAM J. Numer. Anal. **35**(6), 2440–2463 (1998)
12. Cockburn, B., Shu, C.W.: The Runge–Kutta discontinuous Galerkin method for conservation laws. V. Multidimensional systems. J. Comput. Phys. **141**(2), 199–224 (1998). https://doi.org/10.1006/jcph.1998.5892
13. Frean, D., Ryan, J.: Superconvergence and the numerical flux: A study using the upwind-biased flux in discontinuous Galerkin methods. Commun. Appl. Math. Comput. (to appear). https://doi.org/10.1007/s42967-019-00049-2

14. Gottlieb, S., Ketcheson, D.I., Shu, C.W.: High order strong stability preserving time discretizations. J. Sci. Comput. **38**(3), 251–289 (2009). https://doi.org/10.1007/s10915-008-9239-z
15. Li, J., Zhang, D., Meng, X., Wu, B.: Analysis of discontinuous Galerkin methods with upwind-biased fluxes for one dimensional linear hyperbolic equations with degenerate variable coefficients. J. Sci. Comput. **78**(3), 1305–1328 (2019). https://doi.org/10.1007/s10915-018-0831-6
16. Li, J., Zhang, D., Meng, X., Wu, B.: Analysis of local discontinuous Galerkin methods with generalized numerical fluxes for linearized KdV equations. Math. Comput. (submitted) 2
17. Li, J., Zhang, D., Meng, X., Wu, B., Zhang, Q.: Discontinuous Galerkin methods for nonlinear scalar conservation laws: generalized local Lax–Friedrichs numerical fluxes. SIAM J. Numer. Anal. **58**(1), 1–20 (2020). https://doi.org/10.1137/19M1243798
18. Liu, H., Ploymaklam, N.: A local discontinuous Galerkin method for the Burgers–Poisson equation. Numer. Math. **129**(2), 321–351 (2015). https://doi.org/10.1007/s00211-014-0641-1
19. Liu, X., Zhang, D., Meng, X., Wu, B.: Superconvergence of local discontinuous Galerkin methods with generalized alternating fluxes for 1D linear convection-diffusion equations. Sci. China Math. (to appear)
20. Meng, X., Shu, C.W., Wu, B.: Optimal error estimates for discontinuous Galerkin methods based on upwind-biased fluxes for linear hyperbolic equations. Math. Comput. **85**(299), 1225–1261 (2016). https://doi.org/10.1090/mcom/3022
21. Reed, W.H., Hill, T.R.: Triangular mesh methods for the neutron transport equation. Technical Report LA-UR-73-479, Los Alamos Scientific Laboratory, Los Alamos, NM (1973)
22. Shu, C.W.: Discontinuous Galerkin methods for time-dependent convection dominated problems: basics, recent developments and comparison with other methods. In: Building Bridges: Connections and Challenges in Modern Approaches to Numerical Partial Differential Equations. Springer, pp. 371–399 (2016)
23. Wang, H., Zhang, Q., Shu, C.W.: Implicit explicit local discontinuous Galerkin methods with generalized alternating numerical fluxes for convection–diffusion problems. J. Sci. Comput. **81**(3), 2080–2114 (2019). https://doi.org/10.1007/s10915-019-01072-4

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Journal: **10915** Article No.: **1197** TYPESET DISK LE CP Disp.:**2020/3/19** Pages: **19** Layout: **Small**