On Generalized Gauss–Radau Projections and Optimal Error Estimates of Upwind-Biased DG Methods for the Linear Advection Equation on Special Simplex Meshes

On generalized Gauss–Radau projections and optimal error estimates of upwind-biased DG methods for the linear advection equation on special simplex meshes

Zheng Sun* Yulong Xing†

Abstract: Generalized Gauss–Radau (GGR) projections are global projection operators that are widely used for the error analysis of discontinuous Galerkin (DG) methods with generalized numerical fluxes. In previous work, GGR projections were constructed for Cartesian meshes and analyzed through an algebraic approach. In this paper, we first present an alternative energy approach for analyzing the one-dimensional GGR projection, which does not require assembling and explicitly solving a global system over the entire computational domain as that in the algebraic approach. We then generalize this energy argument to construct a global projection operator on special simplex meshes in multidimensions satisfying the so-called flow condition. With this projection, optimal error estimates are proved for upwind-biased DG methods for the linear advection equation on these meshes, which generalizes the error analysis for the purely upwind case in [9] in a time-dependent setting.

1 Introduction

In this paper, we study an energy-based method for the construction and analysis of global projection operators and use them to analyze optimal error estimates of the upwind-biased discontinuous Galerkin (DG) methods for linear advection equations on special simplex meshes in multidimensions satisfying the so-called flow condition. This energy approach is based by the techniques developed in [36]. In contrast to the algebraic-type argument in previous work, it avoids the assembly and solution of a global system in the analysis and can be easily extended to unstructured meshes in multidimensions. To understand the applicability of the method, we first revisit the work of [26, 6] and use this energy approach to reproduce existing results on generalized Gauss–Radau (GGR) projections in one dimension. Then we generalize our argument, without much complication, to construct a global projection operator on multidimensional simplex meshes satisfying the flow condition. This global

2020 *Mathematics Subject Classification.* Primary 65M15, 65M60.

Key words and phrases. discontinuous Galerkin methods, generalized Gauss–Radau projections, upwind-biased fluxes, optimal error estimates, global projections.

*Department of Mathematics, The University of Alabama, Tuscaloosa, AL 35487, USA. E-mail: zsun30@ua.edu. The work of this author is partially supported by the NSF grant DMS-2208391.

†Department of Mathematics, The Ohio State University, Columbus, OH 43210, USA. E-mail: xing.205@osu.edu. The work of this author is partially supported by the NSF grant DMS-1753581.

projection is a generalization of the local projection in [10] and can be used to acquire optimal error estimates of the upwind-biased DG method for the linear advection equation on these special meshes. Despite our analysis concerns a time-dependent problem rather than a steady state problem, the optimal error estimate in this paper is essentially a generalization of the results in [9] from the purely upwind case to the upwind-biased case.

The DG methods are a class of finite element methods using discontinuous piecewise polynomial spaces. They were first introduced by Reed and Hill in [28] for solving the transport equation and were then further developed in the past decades for different applications [1, 15, 30, 14]. The DG methods come with many advantages and have now become one of the main-stream numerical methods for solving partial differential equations arising from science and engineering.

For the DG methods, the so-called numerical fluxes play a central role in the algorithm design and have a crucial effect on the stability and accuracy of the schemes. In the earlier literature, classical numerical fluxes, such as the upwind fluxes (or more generally, the monotone fluxes) for hyperbolic equations and the alternating fluxes for equations with high-order derivatives, are usually considered. Recently, there is a rising interest in analyzing DG schemes with generalized numerical fluxes, such as the upwind-biased fluxes [26, 19, 22], the generalized alternating fluxes [6, 7, 44, 41], the generalized Lax–Friedrichs fluxes [21], the $\alpha\beta$-fluxes [5, 18, 36], etc. These numerical fluxes are perturbed from the classical numerical fluxes with some adjustable parameters. The motivation for using the generalized fluxes is mainly in two folds. Firstly, the parameters in the numerical fluxes may relate to the jump dissipation in the stability estimates. One can make the numerical scheme more stable or less dissipative by adjusting the parameters. In some cases, this will also improve the accuracy of the numerical methods. Secondly, for some complex systems, the classical numerical fluxes, such as the upwind fluxes, may not be easily determined. The generalized fluxes will provide more flexibility in the algorithm design.

The error estimates of DG methods with generalized fluxes can be more involved than the classical methods. It is known that the essential ingredient for proving error estimates of the DG methods is to construct appropriate projection operators, see for example, [13, 12, 16, 23, 25, 35]. For the classical cases, these projections are typically locally-defined. Their well-definedness and approximation properties can usually be proved by looking into the solution of a local system on a single element. For example, the (locally-defined) Gauss–Radau (GR) projection [13, 4] has been used for proving the optimal error estimates of the upwind DG method and the local DG methods with alternating fluxes [30]. However, with generalized fluxes, the required projection operator for optimal error estimates can be global, coupling all mesh cells on the entire computational domain.

An important global projection for error analysis of the DG methods is the GGR projection, which recovers the GR projection in the special case. The GGR projection is introduced by Meng et al. in [26] for optimal error estimates of the upwind-biased DG methods for the linear advection equation. Their analysis is based on an algebraic approach motivated by an earlier work by Bona et al. [2]. The key is to look into the difference between the GGR projection and the GR projection, which is denoted by δ. The well-definedness and approximation property of the GGR projection can be implied by those of δ and the GR projection. To prove the properties of δ, a global linear system is assembled and solved to obtain the explicit formula of δ. The GGR projection on two-dimensional (2D) Cartesian meshes has

also been studied in [26] following the similar idea.

Beyond the work of [26], the approximation estimate of the GGR projection is improved by Cheng et al. in [6] and is used for the optimal error estimate of the local DG method with generalized alternating fluxes for the convection-diffusion equations. After that, the GGR projection along with its variants has been used for the optimal error estimates of the DG method with upwind-biased fluxes for the linear advection equation with degenerate variable coefficients [19, 22], with generalized local Lax–Friedrichs fluxes for 1D nonlinear scalar conservation laws [21], with generalized numerical fluxes for the 1D nonlinear convection-diffusion systems [42], with generalized numerical fluxes for the linearized KdV equations [20], with generalized numerical fluxes for stochastic Maxwell equations with additive noise [32], with generalized alternating fluxes for 2D nonlinear Schrödinger equations [41], etc. The fully discrete error estimates using the GGR projection can also be found in the literature. See, for example, [38, 37, 40]. We remark that due to the construction of the GGR projection, these error estimates are mostly for Cartesian meshes in one and two dimensions.

Besides the GGR projection, recently in [36], Sun and Xing introduced another global projection to prove the optimal error estimates of DG methods with generalized numerical fluxes for wave equations on unstructured simplex meshes. In special cases, this global projection retrieves the locally-defined HDG projection in [12]. The key step in constructing this global projection is again to consider its difference δ from the HDG projection in [12]. However, instead of considering the algebraic system satisfied by δ, the authors used an energy argument for the estimates: appropriate bilinear forms are constructed from the conditions satisfied by δ and then the desired estimates can be deduced from the weak coercivity of the bilinear form. This global projection is also used for the error analysis of the DG methods for stochastic Maxwell equations with multiplicative noise in a recent work [31]. We remark that the energy argument in [36] is different from the construction of elliptic projections. Although they share similarities in terms of both using the coercivity of certain bilinear form, the required coercivity in [36] is much weaker (usually only for the jump seminorm) and is used to analyze the difference term δ — the argument still relies on the existence and approximation properties of a local projection.

So far, we have seen two ways of extending a local projection to a global projection. See Table 1.1. Their common argument is to consider the difference, δ, between the global and the local projections. But then the analysis of δ proceeds differently: one is an algebraic approach in the analysis of the GGR projection [26, 6], the other is an energy approach in the analysis of the global projection in [36]. This paper is an effort to gain an improved understanding of the energy approach for analyzing global projections. We wonder whether it can be used to reproduce the existing results proved through the algebraic approach and whether it can be used to construct new projections that could be less easy to handle by the algebraic approach.

To this end, we start by reproducing existing one-dimensional (1D) results in [26, 6, 19] in a different way. This part of the analysis is given in Section 2, in which we use the energy approach to analyze the 1D GGR projection for the optimal error estimates of the upwind-biased DG method. In Section 3, we study the generalization of the 1D GGR projection to special simplex meshes in multidimensions, which leads to a novel global projection that extends the local projection in [10, 9]. In [9], the authors studied the upwind DG scheme for the steady state transport equation on special simplex meshes satisfying the so-called

flow condition, which requires each mesh cell to have a unique outflow face that is contained in an inflow face of the neighboring cells. See (3.3) and note the meshes can possibly be unstructured. They used the local projection introduced in [10] to prove optimal error estimates of the scheme. In this paper, we consider the time-dependent linear advection equation and construct a global projection that generalizes the local projection in [10]. The main idea is to use the weak coercivity of the DG discretization of the advection operator to analyze the difference term δ. Note this is different from that in [36], where the argument essentially relies on the bilinear form associated with the wave equation. With this novel projection operator, we are able to extend the optimal error estimates of the purely upwind DG schemes in [9] to the upwind-biased DG schemes on these special meshes.

Compared with the algebraic approach in the analysis of GGR projections in [26, 6], the energy argument in [36] and this paper has the following advantages: firstly, the argument is insensitive to the spatial dimension and one can prove the two- and three- dimensional cases in one shot; secondly, since no matrix assembly is needed in the energy approach, the argument can be easily used to construct global projections on unstructured meshes. However, we remark that with the energy approach, one may encounter difficulty in constructing global projections with certain superconvergence properties. Hence it may not substitute the algebraic approach in some cases. For example, we are not able to prove the properties of the 2D GGR projection on Cartesian meshes with the energy approach. See Subsection 3.4 for further discussions.

The rest of the paper is organized as the following. In Section 2, we revisit the optimal error estimates of the upwind-biased DG method for the linear advection equation in one dimension. In particular, we use the energy approach to prove the well-definedness and approximation property of the 1D GGR projection. See Subsection 2.3. In Section 3, we extend the 1D GGR projection to 2D and 3D simplex meshes satisfying the flow condition and apply it to prove the optimal error estimates of the upwind-biased DG method for linear advection equations on these meshes. In Section 4, numerical tests are presented to validate the error estimates. Finally, conclusions are given in Section 5.

Context	Meshes	Local projections	Global projections	Argument
Advection	1D	GR [4, Corollary 3.13]	GGR [26, Lemma 2.6], [6, Lemma 3.2]	Algebraic
			GGR Lemma 2.1	Energy
	2D Cartesian	GR [13, Lemma 3.2]	GGR [26, Lemma 3.3], [6, Lemma 3.3]	Algebraic
	Mulit-D simplex*	[10, Lemma 3.1]	Lemma 3.4	Energy
Wave	1D	[5, Lemma 2.4]	[36, Lemma 2.1]	**
	Mulit-D simplex	HDG [12, Theorem 2.1]	[36, Lemma 3.1]	Energy

* Flow condition is required.

** Constructed with linear combinations of GGR projections. Not built from scratch.

Table 1.1: Local projections and their extensions as global projections.

2 One-dimensional case

In this section, we study the optimal error estimates of the 1D linear advection equation

$$u_t + u_x = 0, \quad u = u(x,t), \quad (x,t) \in \Omega \times (0,T), \quad \Omega = (0,1) \subseteq \mathbb{R} \tag{2.1}$$

along with the 1D GGR projection. Both the periodic boundary condition $u(0,t) = u(1,t)$ and the inflow boundary condition $u(0,t) = g(t)$ are considered.

2.1 Notations

Let $\mathcal{T} = \{I_j\}_{j=1}^N$ be a partition of the computational domain Ω, where the mesh cell $I_j = (x_{j-1/2}, x_{j+1/2})$ has the length $h_j = x_{j+1/2} - x_{j-1/2}$ and $h = \max_{1\le j\le N} h_j$. The finite element space of the DG method is chosen as

$$V_h = \{v \in L^2(\Omega) : v|_{I_j} \in \mathcal{P}_k(I_j), \forall j = 1, \cdots, N\}. \tag{2.2}$$

Here $\mathcal{P}_k(I_j)$ is the space spanned by polynomials on I_j of degree less than or equal to k. Note that functions in V_h can be double-valued at cell interfaces. We denote by $v^{\pm}_{j+1/2} = \lim_{\varepsilon\to 0^{\pm}} v(x_{j+1/2} + \varepsilon)$ the left and right limits of v at $x_{j+1/2}$. The notations

$$[v]_{j+\frac{1}{2}} = v^+_{j+\frac{1}{2}} - v^-_{j+\frac{1}{2}} \quad \text{and} \quad \{v\}^{(\theta)}_{j+\frac{1}{2}} = (\theta v)^-_{j+\frac{1}{2}} + (\tilde{\theta} v)^+_{j+\frac{1}{2}}, \quad \text{with } \tilde{\theta} = 1 - \theta \tag{2.3}$$

are used for the jump and the weighted average of v across $x_{j+1/2}$, respectively. Here $\theta = \{\theta_{j+1/2}\}_{j=1}^N$ is a given set of parameters that may vary with j. Given a function v, we use the following convention for its trace outside of the domain at $x_{N+1/2}$: when the periodic boundary condition is considered, we have $v^+_{N+1/2} = v^+_{1/2}$; when the inflow boundary condition is considered, we have $v^+_{N+1/2} = 0$. We also use

$$\begin{aligned} (w,v)_{I_j} &= \int_{I_j} wv \mathrm{d}x, \quad (w,v)_{\mathcal{T}_h} = \sum_{j=1}^N (w,v)_{I_j}, \\ \|v\|_{L^2(I_j)} &= \sqrt{(v,v)_{I_j}}, \quad \|v\|_{L^2(\mathcal{T}_h)} = \sqrt{(v,v)_{\mathcal{T}_h}}, \end{aligned} \tag{2.4}$$

for the inner products and norms. Let $\mathcal{E}_h^+ = \{x_{j+1/2}\}_{j=1}^N$. For a function w that is single-valued on $\mathcal{E}_h^+$, we define

$$\|w\|_{L^2(\mathcal{E}_h^+)} = \sqrt{\sum_{j=1}^N \left| w_{j+\frac{1}{2}} \right|^2}. \tag{2.5}$$

For a function v that is double-valued along $\mathcal{E}_h^+$, we define

$$\|v\|_{L^2(\mathcal{E}_h^+)} = \sqrt{\frac{1}{2}\left(\|v^+\|^2_{L^2(\mathcal{E}_h^+)} + \|v^-\|^2_{L^2(\mathcal{E}_h^+)} \right)}. \tag{2.6}$$

Note that the left end $x_{1/2}$ is excluded from $\mathcal{E}_h^+$ and $\|\cdot\|_{L^2(\mathcal{E}_h^+)}$.

Moreover, we use the standard notation $H^\ell(I_j)$ to represent the Sobolev space on I_j with the seminorm $|v|_{H^\ell(I_j)} = \|\partial_x^\ell v\|_{L^2(I_j)}$ and the norm $\|v\|_{H^\ell(I_j)} = \sqrt{\sum_{i=0}^\ell |v|^2_{H^i(I_j)}}$, where $\ell \ge 0$ is an integer. We denote by

$$H^\ell(\mathcal{T}_h) = \{v \in L^2(\Omega) : v|_{I_j} \in H^\ell(I_j), \forall j = 1, \cdots, N\} \tag{2.7}$$

the broken Sobolev space with the seminorm $|v|_{H^\ell(\mathcal{T}_h)} = \sqrt{\sum_{j=1}^N |v|^2_{H^\ell(I_j)}}$ and the norm $\|v\|_{H^\ell(\mathcal{T}_h)} = \sqrt{\sum_{j=1}^N \|v\|^2_{H^\ell(I_j)}}$.

2.2 Upwind-biased DG scheme and its error estimate

The upwind-biased DG method for (2.1) is formulated as the following: Find $u_h \in V_h$ such that

$$((u_h)_t, v)_{I_j} - (u_h, v_x)_{I_j} + \widehat{u}_{h,j+\frac{1}{2}} v^-_{j+\frac{1}{2}} - \widehat{u}_{h,j-\frac{1}{2}} v^+_{j-\frac{1}{2}} = 0, \quad \forall v \in V_h, \quad \forall j = 1, \cdots, N, \quad (2.8)$$

where $\widehat{u}_h$ is the so-called upwind-biased numerical flux. To be more specific, we take

$$\widehat{u}_{h,j+\frac{1}{2}} = \begin{cases} \{u_h\}^{(\theta)}_{j+\frac{1}{2}}, & j = 1, \cdots, N \\ \{u_h\}^{(\theta)}_{N+\frac{1}{2}}, & j = 0 \end{cases} \quad (2.9)$$

for the periodic boundary condition, and

$$\widehat{u}_{h,j+\frac{1}{2}} = \begin{cases} \{u_h\}^{(\theta)}_{j+\frac{1}{2}}, & j = 1, \cdots, N-1 \\ g, & j = 0 \\ (u_h)^-_{N+\frac{1}{2}}, & j = N \end{cases} \quad (2.10)$$

for the inflow boundary condition [26, (2.3)-(2.4)]. Recall that $\theta = \{\theta_{j+1/2}\}_{j=1}^N$ contains parameters that may vary with the grid points. Here and in what follows, we assume there are positive constants μ_* and μ^* such that

$$0 < \mu_* \le \theta_{j+\frac{1}{2}} - \frac{1}{2} \le \mu^* < +\infty, \quad \forall j = 1, \cdots, N. \quad (2.11)$$

Note according to our definition, we have $\theta_{N+1/2} = 1$ for the inflow boundary condition. In the special case that $\theta_{j+1/2} \equiv 1$ for all j, it retrieves the standard purely upwind fluxes.

After summing over all mesh cells, the scheme (2.8) can be written in the global form

$$((u_h)_t, v)_{\mathcal{T}_h} = \mathcal{H}(u_h, v) + \mathcal{G}(v), \quad \forall v \in V_h, \quad (2.12)$$

where

$$\mathcal{H}(u_h, v) = (u_h, v_x)_{\mathcal{T}_h} + \sum_{j=1}^N \{u_h\}^{(\theta)}_{j+\frac{1}{2}} [v]_{j+\frac{1}{2}} \quad \text{and} \quad \mathcal{G}(v) = \begin{cases} 0 & \text{for periodic b.c.} \\ g v^+_{1/2} & \text{for inflow b.c.} \end{cases}. \quad (2.13)$$

The bilinear form $\mathcal{H}(\cdot,\cdot)$ is seminegative, in the sense that [6, 43]

$$\mathcal{H}(v, v) = -\sum_{j=1}^N \left(\theta_{j+\frac{1}{2}} - \frac{1}{2}\right) [v]^2_{j+\frac{1}{2}} - \frac{\chi}{2}\left(v^+_{\frac{1}{2}}\right)^2 \le -\mu_* \|v\|^2_{L^2(\mathcal{E}_h^+)} \le 0, \quad \forall v \in V_h. \quad (2.14)$$

Here

$$\chi = \begin{cases} 0 & \text{for periodic b.c.} \\ 1 & \text{for inflow b.c.} \end{cases}. \quad (2.15)$$

Also note that with the inflow boundary condition, we have $\theta_{N+1/2} = 1$ and $v^+_{N+1/2} = 0$ at $x_{N+1/2}$, which yields $\left(\theta_{N+1/2} - 1/2\right)[v]^2_{N+1/2} = \frac{1}{2}(v^-_{N+1/2})^2$, included in $\mathcal{H}(v, v)$ in (2.14).

The key to the error analysis of the upwind-biased DG method is to construct the so-called GGR projection. See Lemma 2.1. This lemma was proved using an algebraic approach in [26, 6]. In Subsection 2.3, we will provide an alternative proof based on an energy approach.

6

Lemma 2.1 (GGR projection)**.** *Suppose the flux parameter θ satisfies the assumption* (2.11)*. Then for a sufficiently smooth function u, there exists a uniquely defined $\Pi_\theta u$, such that*

$$(\Pi_\theta u, v)_{I_j} = (u, v)_{I_j}, \quad \forall v \in \mathcal{P}_{k-1}(I_j), \quad \forall j = 1, \cdots, N, \tag{2.16a}$$

$$\{\Pi_\theta u\}^{(\theta)}_{j+\frac{1}{2}} = \{u\}^{(\theta)}_{j+\frac{1}{2}}, \quad \forall j = 1, \cdots, N. \tag{2.16b}$$

Furthermore, it satisfies the following approximation property

$$\|u - \Pi_\theta u\|_{L^2(\mathcal{T}_h)} + h^{\frac{1}{2}} \|u - \Pi_\theta u\|_{L^2(\mathcal{E}_h^+)} \le C_\theta h^{k+1} |u|_{H^{k+1}(\mathcal{T}_h)}, \tag{2.17}$$

where $C_\theta = C\left(1 + (\mu^* + 1/2)\mu_*^{-1}\right)\left(1 + (\mu^* + 1/2)\right)$, and C is a constant that may depend on k, but is independent of μ^*, μ_* and h.

With the above projection, one can derive the error estimate of the semidiscrete upwind-biased DG method. See Theorem 2.2. Its proof can be found in [26] and is also given below for completeness.

Theorem 2.2. *Consider either the periodic boundary condition or the inflow boundary condition. Suppose the exact solution of* (2.1) *is sufficiently smooth, with uniformly bounded derivatives $\|u\|_{H^{k+1}(\mathcal{T}_h)}$ and $\|u_t\|_{H^{k+1}(\mathcal{T}_h)}$ in time. Suppose θ satisfies* (2.11)*. Then the upwind-biased DG scheme for* (2.1) *admits the following error estimate.*

$$\|u - u_h\|_{L^2(\mathcal{T}_h)}\Big|_{t=T} \le \|u - u_h\|_{L^2(\mathcal{T}_h)}\Big|_{t=0} + C_{\theta,u}(1 + T)h^{k+1}, \tag{2.18}$$

where $C_{\theta,u}$ depends on C_θ in Lemma 2.1, $\|u\|_{H^{k+1}(\mathcal{T}_h)}$, and $\|u_t\|_{H^{k+1}(\mathcal{T}_h)}$, but is independent of h.

Proof. Let $e = u - u_h$, $\eta = u - \Pi_\theta u$ and $\xi = u_h - \Pi_\theta u$. Note the exact solution u admits the variational equation

$$(u_t, v)_{\mathcal{T}_h} = \mathcal{H}(u, v) + \mathcal{G}(v), \quad \forall v \in V_h. \tag{2.19}$$

After subtracting (2.12) from (2.19), we have

$$(e_t, v)_{\mathcal{T}_h} = \mathcal{H}(e, v), \quad \forall v \in V_h. \tag{2.20}$$

Note that $e = \eta - \xi$ and $\mathcal{H}(\eta, v) = 0$ according to the construction of Π_θ. We can split the terms to obtain

$$(\xi_t, v)_{\mathcal{T}_h} = \mathcal{H}(\xi, v) + (\eta_t, v)_{\mathcal{T}_h}, \quad \forall v \in V_h. \tag{2.21}$$

Take $v = \xi$. Recalling the seminegativity of $\mathcal{H}(\cdot, \cdot)$ and applying Cauchy–Schwarz inequality yield

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\xi\|^2_{L^2(\mathcal{T}_h)} = (\xi_t, \xi)_{\mathcal{T}_h} = \mathcal{H}(\xi, \xi) + (\eta_t, \xi)_{\mathcal{T}_h} \le \|\eta_t\|_{L^2(\mathcal{T}_h)} \|\xi\|_{L^2(\mathcal{T}_h)}. \tag{2.22}$$

After simplification, one can obtain $\frac{\mathrm{d}}{\mathrm{d}t}\|\xi\|_{L^2(\mathcal{T}_h)} \le \|\eta_t\|_{L^2(\mathcal{T}_h)}$, which gives

$$\|\xi(\cdot, T)\|_{L^2(\mathcal{T}_h)} \le \|\xi(\cdot, 0)\|_{L^2(\mathcal{T}_h)} + T \sup_{0 \le t \le T} \|\eta_t(\cdot, t)\|_{L^2(\mathcal{T}_h)}. \tag{2.23}$$

The proof can be completed after applying the triangle inequality $\|e\|_{L^2(\mathcal{T}_h)} \le \|\eta\|_{L^2(\mathcal{T}_h)} + \|\xi\|_{L^2(\mathcal{T}_h)}$ and the approximation estimate of Π_θ for $\|\eta\|_{L^2(\mathcal{T}_h)}$ and $\|\eta_t\|_{L^2(\mathcal{T}_h)}$ in Lemma 2.1. □

7

In this section, we assumed $\theta_{j+1/2} - 1/2 \geq \mu_* > 0$ to be uniformly away from 0 by a positive constant μ_* for the optimal convergence. In the case that $\mu_* = C_0 h^\omega$ is very close to 0, where $\omega > 0$ is a constant, one can prove a suboptimal convergence rate for the corresponding upwind-biased DG schemes.

Theorem 2.3. *Under the setting of Theorem 2.2, if $\theta_{j+1/2} - 1/2 \geq \mu_* = C_0 h^\omega$ with $\omega > 0$, then we have*

$$\|u - u_h\|_{L^2(\mathcal{T}_h)}\Big|_{t=T} \leq \|u - u_h\|_{L^2(\mathcal{T}_h)}\Big|_{t=0} + C_{\theta,u}(1+T)h^{k+\max(1-\omega,0)}, \tag{2.24}$$

where $C_{\theta,u}$ depends on C_0, μ^*, $\|u\|_{H^{k+1}(\mathcal{T}_h)}$, and $\|u_t\|_{H^{k+1}(\mathcal{T}_h)}$, but is independent of h.

Proof. Here we give a very sketched proof. When $\omega \geq 1$, one can use the standard L^2 projection with the argument in the proof of [25, Theorem 2.2] to show the kth order convergence rate. When $0 < \omega \leq 1$, by following the proof in Section 2.3, one can see that Lemma 2.1 still holds while $C_\theta \leq C h^{-\omega}$. This gives us

$$\|u - \Pi_\theta u\|_{L^2(\mathcal{T}_h)} + h^{\frac{1}{2}} \|u - \Pi_\theta u\|_{L^2(\mathcal{E}_h^+)} \leq C h^{k+1-\omega} |u|_{H^{k+1}(\mathcal{T}_h)}, \tag{2.25}$$

where C depends on C_0 and μ^*. Using the approximation estimate (2.25) in the proof of Theorem 2.2, we obtain the $(k+1-\omega)$th order convergence rate. □

Remark 2.4. *Through the numerical tests in Example 4.1, we can see that the error estimates in Theorem 2.3 are sharp in general. However, on uniform meshes with an even polynomial order k, one may observe the optimal $(k+1)$th order convergence rate. This relates to the fact that the DG methods with central fluxes ($\theta = 1/2$) are optimal. We refer to [25] for details.*

2.3 An energy-based proof of Lemma 2.1

In this section, we provide proof of Lemma 2.1 based on the energy approach.

Proof. Note the case $\theta \equiv 1$ retrieves the classical GR projection Π_1. The operator is locally-defined through the relationships

$$(\Pi_1 u, v)_{I_j} = (u, v)_{I_j}, \quad \forall v \in \mathcal{P}_{k-1}(I_j), \quad \forall j = 1, \cdots, N, \tag{2.26a}$$

$$(\Pi_1 u)^-_{j+\frac{1}{2}} = u^-_{j+\frac{1}{2}}, \qquad \forall j = 1, \cdots, N. \tag{2.26b}$$

This projection is well-defined and its approximation property (2.17) is well-understood [4, Corollary 3.13]. We observe that Lemma 2.1 holds for $\theta \equiv 1$ and will use perturbation analysis to prove the general case.

Let us define

$$\delta := (\Pi_\theta - \Pi_1)u. \tag{2.27}$$

Note $\delta \in V_h$. By subtracting (2.26) from (2.16), it can be seen that δ satisfies the following equations

$$(\delta, v)_{I_j} = 0, \qquad \forall v \in \mathcal{P}_{k-1}(I_j), \quad \forall j = 1, \cdots, N, \tag{2.28a}$$

$$\{\delta\}^{(\theta)}_{j+\frac{1}{2}} = \bar{\eta}_{j+\frac{1}{2}}, \quad \forall j = 1, \cdots, N, \tag{2.28b}$$

where

$$\bar{\eta}_{j+\frac{1}{2}} = \{u - \Pi_1 u\}^{(\theta)}_{j+\frac{1}{2}} = \tilde{\theta}_{j+\frac{1}{2}} \left(u - \Pi_1 u\right)^{+}_{j+\frac{1}{2}}. \tag{2.29}$$

We claim that (which will be proved later in Lemma 2.6): if (2.28) has a solution, then the solution admits the estimate

$$\|\delta\|_{L^2(\mathcal{T}_h)} + h^{\frac{1}{2}} \|\delta\|_{L^2\left(\mathcal{E}_h^+\right)} \leq \hat{C}_\theta h^{\frac{1}{2}} \|\bar{\eta}\|_{L^2\left(\mathcal{E}_h^+\right)}, \quad \text{with } \hat{C}_\theta = C\left(1 + \left(\mu^* + \frac{1}{2}\right)\mu_*^{-1}\right). \tag{2.30}$$

With this estimate, we can show that (2.28) has a unique solution as follows. Indeed, when $\bar{\eta} = 0$, we know that $\delta = 0$ is a solution to (2.28). Furthermore, $\delta = 0$ has to be the only solution because (2.30) implies $\|\delta\|_{L^2(\mathcal{T}_h)} + h^{\frac{1}{2}} \|\delta\|_{L^2\left(\mathcal{E}_h^+\right)} \leq 0$. Therefore, when the system (2.28) is homogeneous, with $\bar{\eta} = 0$, it has a unique solution $\delta = 0$. Recall that $A\boldsymbol{x} = \boldsymbol{0}$ has a unique solution $\boldsymbol{x} = \boldsymbol{0}$ implies that the solution to $A\boldsymbol{x} = \boldsymbol{b}$, if it exists, is unique. Hence we prove the uniqueness of the solution to (2.28) also for $\bar{\eta} \neq 0$. Moreover, note that (2.28) is a linear, square, and finite-dimensional system, the uniqueness of the solution also implies the existence of the solution. Hence (2.28) is unisolvent.

For the uniquely defined δ, we once again look into the estimate (2.30). Also note that $\bar{\eta} = \{u - \Pi_1 u\}^{(\theta)}$, and the error term $u - \Pi_1 u$ satisfies the estimate (2.17) with $\theta = 1$, which leads to

$$\|\bar{\eta}\|_{L^2\left(\mathcal{E}_h^+\right)} \leq C\left(\mu^* + \frac{1}{2}\right) h^{k+\frac{1}{2}} |u|_{H^{k+1}(\mathcal{T}_h)}. \tag{2.31}$$

Therefore, substituting (2.31) into (2.30), we have the estimate of the difference term:

$$\|\delta\|_{L^2(\mathcal{T}_h)} + h^{\frac{1}{2}} \|\delta\|_{L^2\left(\mathcal{E}_h^+\right)} \leq C_\theta h^{k+1} |u|_{H^{k+1}(\mathcal{T}_h)}. \tag{2.32}$$

Recall that $\Pi_\theta u = \Pi_1 u + \delta$. Hence $\Pi_\theta u$ is also uniquely determined. Its approximation estimate (2.17) is based on that of δ and $\Pi_1 u$, and can be obtained after applying the triangle inequality

$$\begin{aligned} &\|u - \Pi_\theta u\|_{L^2(\mathcal{T}_h)} + h^{\frac{1}{2}} \|u - \Pi_\theta u\|_{L^2\left(\mathcal{E}_h^+\right)} \\ \leq &\|u - \Pi_1 u\|_{L^2(\mathcal{T}_h)} + h^{\frac{1}{2}} \|u - \Pi_1 u\|_{L^2\left(\mathcal{E}_h^+\right)} + \|\delta\|_{L^2(\mathcal{T}_h)} + h^{\frac{1}{2}} \|\delta\|_{L^2\left(\mathcal{E}_h^+\right)} \\ \leq &C_\theta h^{k+1} |u|_{H^{k+1}(\mathcal{T}_h)}. \end{aligned} \tag{2.33}$$

□

It now suffices to prove (2.30), which is obtained from Proposition 2.5 and Lemma 2.6.

Proposition 2.5. *Given a real number z, there is a unique function $Z \in \mathcal{P}_k(I_j)$ such that*

$$(Z, v)_{I_j} = 0, \quad \forall v \in \mathcal{P}_{k-1}(I_j), \tag{2.34a}$$

$$Z^-_{j+\frac{1}{2}} = z. \tag{2.34b}$$

Moreover, we have

$$\|Z\|_{L^2(I_j)} \leq C h_j^{\frac{1}{2}} |z|, \tag{2.35}$$

where C is a constant only dependent on k.

Proof. Firstly, we make the following assumption which will be proved in the next paragraph: if Z is a solution to (2.34), then it satisfies the estimate (2.35). With this assumption, we can prove that (2.34) has a unique solution: When $z = 0$, we know that the system (2.34) has a solution $Z = 0$, which is indeed the only solution due to the estimate $\|Z\|_{L^2(I_j)} \leq Ch_j^{1/2}|z| = 0$. Hence the solution to (2.34) is unique when $z = 0$. By the linearity of the equation system, the solution to (2.34) with $z \neq 0$, if it exists, is also unique. This proves the uniqueness of the solution to (2.34). Furthermore, note that (2.34) is a linear, square, and finite-dimensional system of Z. The uniqueness of the solution to (2.34) also implies the existence of the solution. Hence (2.34) is unisolvent.

Now we prove the estimate (2.35). Let us denote by $\hat{I} = [-1, 1]$. We can write (2.34a) as $Z(\cdot) \in \mathcal{P}_{k-1}^{\perp}(I_j)$ and hence $Z(\cdot h_j/2 + x_j) \in \mathcal{P}_{k-1}^{\perp}(\hat{I})$. Here x_j is the midpoint of I_j. Furthermore, by changing the variable, it yields

$$\|Z(\cdot)\|_{L^2(I_j)}^2 = \frac{h_j}{2}\|Z\left(\cdot h_j/2 + x_j\right)\|_{L^2(\hat{I})}^2. \tag{2.36}$$

Note that $|||v||| := |v(1)|$ is a norm on $P_{k-1}^{\perp}(\hat{I})$.[1] Using the norm equivalence in the finite-dimensional space, we have

$$\frac{h_j}{2}\|Z\left(\cdot h_j/2 + x_j\right)\|_{L^2(\hat{I})}^2 \leq Ch_j|||Z\left(\cdot h_j/2 + x_j\right)|||^2 = Ch_j\left|Z_{j+\frac{1}{2}}^-\right|^2 = Ch_j|z|^2. \tag{2.37}$$

The proof of (2.35) is completed after combining (2.36) and (2.37). □

Lemma 2.6. *Let δ be the solution of* (2.28). *Then δ is well-defined and satisfies* (2.30).

Proof. We divide the proof of this lemma into three steps.

Step 1: Estimate of $\|[\delta]\|_{L^2(\mathcal{E}_h^+)}$. We take $v = \delta_x$ in (2.28a), multiply (2.28b) with $[\delta]$, add the two equations and sum over all j. It then yields

$$\sum_{j=1}^{N}\left((\delta, \delta_x)_j + \{\delta\}_{j+\frac{1}{2}}^{(\theta)}[\delta]_{j+\frac{1}{2}}\right) = \sum_{j=1}^{N}\bar{\eta}_{j+\frac{1}{2}}[\delta]_{j+\frac{1}{2}}.$$

Note that the left side assembles the bilinear form $\mathcal{H}(\delta, \delta)$. According to (2.14), we have

$$\mu_*\|[\delta]\|_{L^2(\mathcal{E}_h^+)}^2 \leq \sum_{j=1}^{N}\left(\theta_{j+\frac{1}{2}} - \frac{1}{2}\right)[\delta]_{j+\frac{1}{2}}^2 + \frac{\chi}{2}\left(\delta_{\frac{1}{2}}^+\right)^2 = |\mathcal{H}(\delta, \delta)| = \left|\sum_{j=1}^{N}\bar{\eta}_{j+\frac{1}{2}}[\delta]_{j+\frac{1}{2}}\right|. \tag{2.38}$$

One can then apply Cauchy–Schwarz inequality on the right side to obtain

$$\mu_*\|[\delta]\|_{L^2(\mathcal{E}_h^+)}^2 \leq \|\bar{\eta}\|_{L^2(\mathcal{E}_h^+)}\|[\delta]\|_{L^2(\mathcal{E}_h^+)}, \tag{2.39}$$

which gives

$$\|[\delta]\|_{L^2(\mathcal{E}_h^+)} \leq \mu_*^{-1}\|\bar{\eta}\|_{L^2(\mathcal{E}_h^+)}. \tag{2.40}$$

[1]Since $|||\cdot|||$ is already a seminorm, it suffices to show $|v(1)| = 0$ implies $v \equiv 0$, $\forall v \in P_{k-1}^{\perp}(\hat{I})$. Indeed, note that $P_{k-1}^{\perp}(\hat{I}) = \{al_k(x)|a \in \mathbb{R}\}$, where $l_k(x)$ is the kth-order Legendre polynomial on $\hat{I}$. For $v = al_k(x)$, since $l_k(1) \neq 0$, one can see that $v(1) = 0$ implies $a = 0$ and hence $v \equiv 0$.

Step 2: Estimate of $\|\delta\|_{L^2(\mathcal{E}_h^+)}$. Add $\delta^-_{j+1/2} - \{\delta\}^{(\theta)}_{j+1/2}$ on both sides of (2.28b). It gives

$$\delta^-_{j+\frac{1}{2}} = \bar{\eta}_{j+\frac{1}{2}} - \tilde{\theta}_{j+\frac{1}{2}}[\delta]_{j+\frac{1}{2}}. \tag{2.41}$$

Hence using the triangle inequality and the estimate (2.40), we have

$$\begin{aligned} \|\delta^-\|_{L^2(\mathcal{E}_h^+)} \leq & \|\bar{\eta}\|_{L^2(\mathcal{E}_h^+)} + \left(\sup_{1\leq j\leq N}\left|\tilde{\theta}_{j+\frac{1}{2}}\right|\right)\|[\delta]\|_{L^2(\mathcal{E}_h^+)} \\ \leq & \left(1 + \sup_{1\leq j\leq N}\left|\tilde{\theta}_{j+\frac{1}{2}}\right|\mu_*^{-1}\right)\|\bar{\eta}\|_{L^2(\mathcal{E}_h^+)} \leq \hat{C}_\theta\|\bar{\eta}\|_{L^2(\mathcal{E}_h^+)}. \end{aligned} \tag{2.42}$$

Similarly, we can add $\delta^+_{j+1/2} - \{\delta\}^{(\theta)}_{j+1/2}$ on both sides of (2.28b) to obtain

$$\delta^+_{j+\frac{1}{2}} = \bar{\eta}_{j+\frac{1}{2}} + \theta_{j+\frac{1}{2}}[\delta]_{j+\frac{1}{2}}. \tag{2.43}$$

Following the derivation in (2.42) yields a similar estimate $\|\delta^+\|_{L^2(\mathcal{E}_h^+)} \leq \hat{C}_\theta\|\bar{\eta}\|_{L^2(\mathcal{E}_h^+)}$. Therefore, we have

$$\|\delta\|_{L^2(\mathcal{E}_h^+)} = \sqrt{\frac{1}{2}\left(\|\delta^+\|^2_{L^2(\mathcal{E}_h^+)} + \|\delta^-\|^2_{L^2(\mathcal{E}_h^+)}\right)} \leq \hat{C}_\theta\|\bar{\eta}\|_{L^2(\mathcal{E}_h^+)}. \tag{2.44}$$

Step 3: Estimate of $\|\delta\|_{L^2(\mathcal{T}_h)}$. Note that $\delta|_{I_j}$ satisfies (2.34) with $z = \delta^-_{j+1/2}$. Therefore, Proposition 2.5 implies

$$\|\delta\|_{L^2(I_j)} \leq Ch_j^{\frac{1}{2}}\left|\delta^-_{j+\frac{1}{2}}\right|. \tag{2.45}$$

After taking the square, summing over all j, and applying the estimate (2.42), one can obtain

$$\|\delta\|^2_{L^2(\mathcal{T}_h)} = \sum_{j=1}^N \|\delta\|^2_{L^2(I_j)} \leq C\sum_{j=1}^N h_j\left|\delta^-_{j+\frac{1}{2}}\right|^2 \leq Ch\|\delta^-\|^2_{L^2(\mathcal{E}_h^+)} \leq \hat{C}_\theta^2 h\|\bar{\eta}\|^2_{L^2(\mathcal{E}_h^+)}. \tag{2.46}$$

Finally, the proof of (2.30) can be completed after combining (2.44) and (2.46). □

Remark 2.7. *In the analysis of [26] and [6], the authors provide an algebraic proof of Lemma 2.6. The proof uses the fact that the linear system of* (2.28) *under a given basis can be assembled and solved explicitly. For periodic boundary condition, this methodology has been applied to construct the GGR projection with one of the following settings:*

1. $\theta_{j+1/2} \equiv \theta_{1/2}$ *is constant;*
2. *There exists an index* j_* *such that* $\theta_{j_*+1/2} = 1$ *(or* -1 *for* $u_t = u_x$*).*

The associated matrix is circulant for the first case and is triangular (after permutation) for the second case, which can be both inverted analytically in a neat form. For general θ*, the algebraic approach to construct the corresponding GGR projection would still work, but one has to deal with the complication of inverting the bidiagonal matrix with a periodic boundary. While the energy-based analysis in Subsection 2.3 does not rely on these specialties of* θ *and the case with general* θ *can be covered. See Subsection 2.4 for further discussions.*

2.4 A more general projection

The energy approach can be used to analyze the following projection operator, for which the flux coefficient θ may vary at different mesh cells. This projection can be used to prove the optimal error estimates of the upwind-biased DG method for the linear advection equation with degenerate variable coefficients $u_t + a(x)u_x = 0$ and the DG methods with generalized local Lax–Friedrichs fluxes for nonlinear conservation laws $u_t + f(u)_x = 0$. See Remark 2.10.

Lemma 2.8. *Given $\theta = \{\theta_j\}_{j=1}^N$ such that*

$$0 < \mu_* \le \left|\theta_j - \frac{1}{2}\right| \le \mu^* < +\infty, \quad \forall j = 1, \cdots, N, \tag{2.47}$$

there exists a uniquely defined $\Pi_\theta u$ satisfying

$$(\Pi_\theta u, v)_{I_j} = (u, v)_{I_j}, \quad \forall v \in \mathcal{P}_{k-1}(I_j), \quad \forall j = 1, \cdots, N, \tag{2.48a}$$

$$\{\Pi_\theta u\}^{(\theta_j)}_{j+\frac{1}{2}} = \{u\}^{(\theta_j)}_{j+\frac{1}{2}}, \quad \text{if } \theta_j > \frac{1}{2}, \qquad \forall j = 1, \cdots, N, \tag{2.48b}$$

$$\{\Pi_\theta u\}^{(\theta_j)}_{j-\frac{1}{2}} = \{u\}^{(\theta_j)}_{j-\frac{1}{2}}, \quad \text{if } \theta_j < \frac{1}{2}, \qquad \forall j = 1, \cdots, N. \tag{2.48c}$$

Furthermore, we have

$$\|u - \Pi_\theta u\|_{L^2(\mathcal{T}_h)} + h^{\frac{1}{2}} \|u - \Pi_\theta u\|_{L^2(\mathcal{E}_h^+)} \le C_\theta h^{k+1} |u|_{H^{k+1}(\mathcal{T}_h)}, \tag{2.49}$$

where $C_\theta = C\left(1 + (\mu^* + 1/2)\,\mu_*^{-1}\right)\left(1 + (\mu^* + 1/2)\,\mu_*^{-1/2}\right)(1 + (\mu^* + 1/2))$, and C is a constant dependent on k, but is independent of μ^*, μ_* and h.

An energy-based proof of Lemma 2.8 is given in Appendix A.

Remark 2.9. *With $\theta_j = \theta_{j+1/2} > 1/2$, (2.48) retrieves the GGR projection in Lemma 2.1.*

Remark 2.10. *Suppose the sequence $\theta = \{\theta_{j+1/2}\}_{j=1}^N$ changes sign only twice. We have*

$$\begin{cases} \theta_{j+\frac{1}{2}} > \frac{1}{2}, & \text{if } \beta \le j \le \gamma - 1 \\ \theta_{j+\frac{1}{2}} < \frac{1}{2}, & \text{otherwise} \end{cases}. \tag{2.50}$$

The choice

$$\theta_j = \begin{cases} 1 & \text{if } j = \gamma \\ \theta_{j+\frac{1}{2}} & \text{if } \beta \le j \le \gamma - 1 \\ \theta_{j-\frac{1}{2}} & \text{otherwise} \end{cases} \tag{2.51}$$

yields the projection

$$(\Pi_\theta u, v)_{I_j} = (u, v)_{I_j}, \quad \forall v \in \mathcal{P}_{k-1}(I_j), \quad \forall j = 1, \cdots N, \tag{2.52a}$$

$$(\Pi_\theta u)^-_{j+\frac{1}{2}} = u^-_{j+\frac{1}{2}}, \quad \text{if } j = \gamma, \tag{2.52b}$$

$$\{\Pi_\theta u\}^{(\theta)}_{j+\frac{1}{2}} = \{u\}^{(\theta)}_{j+\frac{1}{2}}, \quad \text{if } \beta \le j \le \gamma - 1, \tag{2.52c}$$

$$\{\Pi_\theta u\}^{(\theta)}_{j-\frac{1}{2}} = \{u\}^{(\theta)}_{j-\frac{1}{2}}, \quad \text{otherwise}, \tag{2.52d}$$

which retrieves the projection constructed in [19, Lemma 3.1] for optimal error estimates of the upwind-biased DG methods for the linear advection equation with degenerate variable coefficients. Moreover, let us assume $\lambda > |\nu|$. If we change the parametrization in (2.50) *as*

$$\begin{cases} \theta_{j+\frac{1}{2}} := \frac{1}{2} + \left(\lambda_{j+\frac{1}{2}} + \nu_{j+\frac{1}{2}}\right) > \frac{1}{2}, & \text{if } \beta \le j \le \gamma - 1 \\ \theta_{j+\frac{1}{2}} := \frac{1}{2} - \left(\lambda_{j+\frac{1}{2}} - \nu_{j+\frac{1}{2}}\right) < \frac{1}{2}, & \text{otherwise} \end{cases}, \tag{2.53}$$

then (2.52) *will retrieve the piecewise global projection in [21, Lemma 3.2] that is used for optimal error estimates of the DG methods for nonlinear conservation laws with generalized local Lax–Friedrichs fluxes.*

3 Multi-dimensional case

In this section, we consider the linear advection equation in multidimensions,

$$u_t + \partial_{\boldsymbol{\beta}} u = 0, \quad u = u(\boldsymbol{x}, t), \quad (\boldsymbol{x}, t) \in \Omega \times (0, T). \tag{3.1}$$

Here $\partial_{\boldsymbol{\beta}} = \boldsymbol{\beta} \cdot \nabla$ and $\boldsymbol{\beta}$ is a non-zero constant vector. We assume $\Omega \subseteq \mathbb{R}^d$, $d = 2, 3$. To avoid unnecessary technicality, let us only consider the periodic boundary condition and hence assume Ω is a rectangular domain in 2D or a cuboid domain in 3D, although the inflow boundary condition with a convex polygonal domain can be analyzed along similar lines.

3.1 Notations

3.1.1 Mesh partition

Let $\mathcal{T}_h = \{K\}$ be a partition of the domain Ω with simplices. Given a simplex K and a face $e \in \partial K$, we use $\boldsymbol{n}_{e_K}$ to represent the outward unit vector along e with respect to K. The subscripts of $\boldsymbol{n}$ may be omitted when it does not cause confusion. Let h_K be the diameter of K and $h = \max_{K \in \mathcal{T}_h} h_K$. We assume $\mathcal{T}_h$ to be shape-regular. In other words, there exists a positive constant $\sigma > 0$, such that

$$h_K / \rho_K \le \sigma, \quad \forall K \in \mathcal{T}_h, \tag{3.2}$$

where ρ_K is the diameter of the inscribed ball of K. In addition to the shape-regularity assumption, we also assume $\mathcal{T}_h$ satisfies the following flow condition [9]:

(A1) Each simplex K has a unique outflow face with respect to $\boldsymbol{\beta}$, denoted by e_K^+. (3.3a)

(A2) Each interior face e_K^+ is included in an inflow face with respect to $\boldsymbol{\beta}$ of another simplex. (3.3b)

Here we say e is an outflow (inflow) face with respect to $\boldsymbol{\beta}$ if $\boldsymbol{\beta} \cdot \boldsymbol{n}_{e_K} > (<)\ 0$. The set of all outflow faces is denoted by $\mathcal{E}_h^+ := \cup_{K \in \mathcal{T}_h} \{e_K^+\}$. Note that hanging nodes are allowed if they do not appear on the outflow face of a simplex. Further characterizations on meshes satisfying the flow condition (3.3), including their construction on general polygonal domains in any dimensions, can be found in [9].

Remark 3.1. *Typically, the flow condition* (3.3) *may imply a strong assumption that many faces in the mesh partition have to be parallel to the flow direction* $\boldsymbol{\beta}$*, so that the upwind-biased DG scheme can be written in the form of Proposition 3.3 and the number of cell-interface terms in the error estimates can be reduced. For optimal error estimates, the flow condition* (3.3) *can be relaxed. It can allow more than one outflow face by either having faces to be "almost parallel" to* $\boldsymbol{\beta}$ *or requiring the number of those outflow faces to be appropriately bounded. We refer to [11] for details. More generally, when the flow condition is not satisfied, we may observe* $(k+1/2)$*th order convergence rate for some numerical tests, see [27] for an example.*

In addition, we note that the flow condition (3.3a) together with the shape-regularity condition (3.2) implies the transversality condition on $\mathcal{E}_h^+$ (but not on all edges of $\mathcal{T}_h$). See Lemma 3.2, whose proof is given in Appendix B.

Lemma 3.2 (Transversality condition on $\mathcal{E}_h^+$)**.** *For* $d = 2, 3$*, there exists a positive constant* γ*, which depends on the shape-regularity constant* σ*, such that*

$$\boldsymbol{\beta} \cdot \boldsymbol{n}_{e_K^+} \geq |\boldsymbol{\beta}|\gamma > 0, \quad \forall K \in \mathcal{T}_h. \tag{3.4}$$

3.1.2 Finite element space, inner products, and norms

The finite element space of DG discretization is taken as

$$V_h = \{v \in L^2(\Omega) : v|_K \in \mathcal{P}_k(K)\}, \tag{3.5}$$

where $\mathcal{P}_k(K)$ is the linear span of polynomials on K of degree less than or equal to k. Along a face e, we denote by $v^{\pm} = \lim_{\varepsilon \to 0^{\pm}} v(x + \varepsilon\boldsymbol{\beta})$. As those in the 1D case, we use

$$[v] = v^+ - v^- \quad \text{and} \quad \{v\}^{(\theta)} = (\theta v)^- + \left(\tilde{\theta} v\right)^+, \quad \text{with } \tilde{\theta} = 1 - \theta, \tag{3.6}$$

for the jump and weighted average of v across a face, respectively. Let us denote by $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}/|\boldsymbol{\beta}|$ the unit vector with the same direction as $\boldsymbol{\beta}$. Let w and v be single-valued functions defined along the element edges in (3.7), and be functions in V_h in (3.8) and (3.9). The following notations will be used in our analysis.

$$\langle w, v\rangle_e = \int_e wv \mathrm{d}l, \qquad \|v\|_{L^2(e)} = \sqrt{\langle v, v\rangle_e}, \qquad \|v\|_{\hat{\boldsymbol{\beta}}, L^2(e_K^+)} = \sqrt{\left\langle \hat{\boldsymbol{\beta}} \cdot \boldsymbol{n} v, v \right\rangle_{e_K^+}}, \tag{3.7}$$

$$(w, v)_K = \int_K wv \mathrm{d}\boldsymbol{x}, \qquad \|v\|_{L^2(K)} = \sqrt{(v, v)_K}, \qquad \|v\|_{\hat{\boldsymbol{\beta}}, L^2(K)} = \sqrt{\left(\hat{\boldsymbol{\beta}} \cdot \boldsymbol{n}_{e_K^+} v, v\right)_K}, \tag{3.8}$$

$$(w, v)_{\mathcal{T}_h} = \sum_{K \in \mathcal{T}_h} (w, v)_K, \ \|v\|_{L^2(\mathcal{T}_h)} = \sqrt{\sum_{K \in \mathcal{T}_h} \|v\|^2_{L^2(K)}}, \ \|v\|_{\hat{\boldsymbol{\beta}}, L^2(\mathcal{T}_h)} = \sqrt{\sum_{K \in \mathcal{T}_h} \|v\|^2_{\hat{\boldsymbol{\beta}}, L^2(K)}}. \tag{3.9}$$

14

Furthermore, for a single-valued function w and a double-valued function v along the outflow edges, let us define

$$\|w\|_{L^2(\mathcal{E}_h^+)} = \sqrt{\sum_{K\in\mathcal{T}_h} \|w\|^2_{L^2(e_K^+)}}, \qquad \|v\|_{L^2(\mathcal{E}_h^+)} = \sqrt{\frac{1}{2}\left(\|v^+\|^2_{L^2(\mathcal{E}_h^+)} + \|v^-\|^2_{L^2(\mathcal{E}_h^+)}\right)}, \tag{3.10}$$

$$\|w\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)} = \sqrt{\sum_{K\in\mathcal{T}_h} \|w\|^2_{\hat{\boldsymbol{\beta}},L^2(e_K^+)}}, \ \|v\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)} = \sqrt{\frac{1}{2}\left(\|v^+\|^2_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)} + \|v^-\|^2_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)}\right)}. \tag{3.11}$$

Note that due to Lemma 3.2, $\|\cdot\|_{L^2(\mathcal{E}_h^+)}$ and $\|\cdot\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)}$ are equivalent, and $\|\cdot\|_{L^2(\mathcal{T}_h)}$ and $\|\cdot\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{T}_h)}$ are equivalent, upto a constant dependent on γ (and hence σ).

As before, letting $\ell \geq 0$ be an integer, we use the standard notation $H^\ell(K)$ to represent the Sobolev space on K with the seminorm $|\cdot|_{H^\ell(K)}$ and the norm $\|\cdot\|_{H^\ell(K)}$. We denote by

$$H^\ell(\mathcal{T}_h) = \{v \in L^2(\Omega) : v|_K \in H^\ell(K), \forall K \in \mathcal{T}_h\} \tag{3.12}$$

the broken Sobolev space with the seminorm $|v|_{H^\ell(\mathcal{T}_h)} = \sqrt{\sum_{K\in\mathcal{T}_h} |v|^2_{H^\ell(K)}}$ and the norm $\|v\|_{H^\ell(\mathcal{T}_h)} = \sqrt{\sum_{K\in\mathcal{T}_h} \|v\|^2_{H^\ell(K)}}$.

3.2 Upwind-biased DG scheme and its error estimate

The upwind-biased DG scheme for (3.1) is defined as the following: Find $u_h \in V_h$, such that

$$((u_h)_t, v)_K - (u_h, \partial_{\boldsymbol{\beta}} v)_K + \sum_{e_K\in\partial K} \left\langle \{u_h\}^{(\theta)}, \boldsymbol{\beta}\cdot\boldsymbol{n} v\right\rangle_{e_K} = 0, \quad \forall v \in V_h. \tag{3.13}$$

Proposition 3.3. *Under the flow condition* (3.3a)*, the DG scheme* (3.13) *can be equivalently written as*

$$((u_h)_t, v)_{\mathcal{T}_h} = \mathcal{H}(u_h, v; \boldsymbol{\beta}), \quad \forall v \in V_h, \tag{3.14}$$

where

$$\mathcal{H}(u_h, v; \boldsymbol{\beta}) = (u_h, \partial_{\boldsymbol{\beta}} v)_{\mathcal{T}_h} + \sum_{K\in\mathcal{T}_h} \left\langle \{u_h\}^{(\theta)}, \boldsymbol{\beta}\cdot\boldsymbol{n}\,[v]\right\rangle_{e_K^+}. \tag{3.15}$$

Moreover, we have

$$\mathcal{H}(v, v; \boldsymbol{\beta}) = -|\boldsymbol{\beta}| \sum_{K\in\mathcal{T}_h} \left(\left(\theta_{e_K^+} - \frac{1}{2}\right) \|[v]\|^2_{\hat{\boldsymbol{\beta}},L^2(e_K^+)}\right), \quad \forall v \in V_h. \tag{3.16}$$

Proof. (3.14) can be proved by taking the summation of (3.13) over all mesh cells, combining the integrals along cell interfaces for adjacent elements, and finally noting that $\boldsymbol{\beta}\cdot\boldsymbol{n} = 0$ if the edge is not an inflow or outflow edge for any K. (3.16) can be verified through a similar argument as the proof of (2.14). □

15

As one can see from (3.14) and (3.15), one only needs to specify θ along $\mathcal{E}_h^+$. Here we make the following assumption.

$$0 \le \mu_* \le \theta_{e_K^+} - \frac{1}{2} \le \mu^* < +\infty, \quad \forall K \in \mathcal{T}_h. \tag{3.17}$$

In the following lemma, we define a global projection associated with the special simplex mesh, whose proof is based on an energy approach and is postponed to Section 3.3.

Lemma 3.4. *Suppose $\mathcal{T}_h$ is a shape-regular mesh* (3.2) *satisfying the flow condition* (3.3) *and the flux parameter $\{\theta_{e_K^+}\}_{K\in\mathcal{T}_h}$ satisfies* (3.17)*. Then for any sufficiently smooth function u, there exists a unique $\Pi_\theta u$ such that*

$$(\Pi_\theta u, v)_K = (u, v)_K, \qquad \forall v \in \mathcal{P}_{k-1}(K), \quad \forall K \in \mathcal{T}_h, \tag{3.18a}$$

$$\left\langle \{\Pi_\theta u\}^{(\theta)}, w \right\rangle_{e_K^+} = \left\langle \{u\}^{(\theta)}, w \right\rangle_{e_K^+}, \quad \forall w \in \mathcal{P}_k\left(e_K^+\right), \quad \forall K \in \mathcal{T}_h. \tag{3.18b}$$

Furthermore, we have

$$\|u - \Pi_\theta u\|_{L^2(\mathcal{T}_h)} + h^{\frac{1}{2}} \|u - \Pi_\theta u\|_{L^2\left(\mathcal{E}_h^+\right)} \le C_\theta h^{k+1} |u|_{H^{k+1}(\mathcal{T}_h)}, \tag{3.19}$$

where $C_\theta = C(1 + (\mu^* + 1/2)\mu_*^{-1})\,(1 + (\mu^* + 1/2))$ and C is a constant dependent on k and σ, but is independent of μ^*, μ_* and h.

With the projection in Lemma 3.4, we are able to prove the optimal error estimate of (3.13), as outlined in the theorem below. The proof is omitted here, since it is the same as that of the 1D case, except for replacing $\mathcal{H}(\cdot,\cdot)$ with $\mathcal{H}(\cdot,\cdot;\boldsymbol{\beta})$.

Theorem 3.5. *Suppose the exact solution of* (3.1) *is sufficiently smooth, with uniformly bounded $\|u\|_{H^{k+1}(\mathcal{T}_h)}$ and $\|u_t\|_{H^{k+1}(\mathcal{T}_h)}$. For $\mathcal{T}_h$ and θ satisfying conditions in Lemma 3.4, the upwind-biased DG scheme* (3.13) *for* (3.1) *admits the following error estimate*

$$\|u - u_h\|_{L^2(\mathcal{T}_h)}\Big|_{t=T} \le \|u - u_h\|_{L^2(\mathcal{T}_h)}\Big|_{t=0} + C_{\theta,u}(1+T)h^{k+1}, \tag{3.20}$$

where $C_{\theta,u}$ depends on C_θ in Lemma 3.4, $\|u\|_{H^{k+1}(\mathcal{T}_h)}$, and $\|u_t\|_{H^{k+1}(\mathcal{T}_h)}$, but is independent of h.

Remark 3.6. *In general, when $\theta_{j+1/2} - 1/2 \ge \mu_* = C_0 h^\omega$ with $\omega > 0$, we expect similar suboptimal convergence as that in the 1D case (see Theorem 2.3). A numerical test with P^1 elements on unstructured meshes is given in Table 4.7 of Example 4.3.*

Remark 3.7. *The projection in Lemma 3.4 can be considered as a multidimensional extension of those in Lemmas 2.1 and 2.8. Indeed, it is written in a closer format as that in Lemma 2.8. The main complication in defining the 1D projection in Lemma 2.8 is to specify whether the outflow edge should be $x_{j-1/2}$ or $x_{j+1/2}$. While this complication has been automatically taken care of in the multidimensional case with the notation of e_K^+.*

16

Remark 3.8. *Here although we focus on the case with constant coefficients, we expect similar optimal error estimates can be obtained for the case with variable coefficients. In [11], Cockburn et al. relaxed the mesh condition in* (3.3) *and proved optimal error estimates of the purely upwind DG methods for the steady state transport equation with variable coefficients. The analysis utilizes the local projection in [9, 10] corresponding to* $\theta = 1$ *in Lemma 3.4. We expect that optimal error estimates can be extended to the variable coefficient case by following similar argument in [11] and replacing the local projection by the global projection in Lemma 3.4.*

3.3 Proof of Lemma 3.4

Note that $\theta \equiv 1$ retrieves a local projection operator. It is well-defined and its approximation property has been shown in [9, Lemma 2.1] and [10, Proposition 2.1].[2]

Lemma 3.9. *Lemma 3.4 holds for* $\theta \equiv 1$.

The proof of Lemma 3.9 is based on a multi-dimensional version of Proposition 2.5, which is stated in Lemma 3.10. The proof of Lemma 3.10 can be found in [10, Lemma 3.1].

Lemma 3.10. *Given a face* e *of the simplex* K *and a function* $z \in L^2(e)$*, there is a unique function* $Z \in \mathcal{P}_k(K)$ *such that*

$$(Z, v)_K = 0, \qquad \forall v \in \mathcal{P}_{k-1}(K), \tag{3.21a}$$
$$\langle Z, w\rangle_e = \langle z, w\rangle_e, \quad \forall w \in \mathcal{P}_k(e). \tag{3.21b}$$

Moreover,

$$\|Z\|_{L^2(K)} \le C h_K^{\frac{1}{2}} \|z\|_{L^2(e)}, \tag{3.22}$$

where C *depends solely on the polynomial degree* k *and the shape regularity constant* σ.

With a well-defined local projection Π_1 and the estimate with trace (3.22). We can use an energy argument to prove Lemma 3.4. The proof is very similar to that of the 1D result in Subsection 2.3.

Proof of Lemma 3.4. Let $\delta := (\Pi_\theta - \Pi_1)u$. Set $\theta \equiv 1$ in (3.18) and subtract the resulted equation from (3.18) with a general θ. Then it yields

$$(\delta, v)_K = 0, \qquad \forall v \in \mathcal{P}_{k-1}(K), \quad \forall K \in \mathcal{T}_h, \tag{3.23a}$$
$$\left\langle \{\delta\}^{(\theta)}, w\right\rangle_{e_K^+} = \langle \bar{\eta}, w\rangle_{e_K^+}, \quad \forall w \in \mathcal{P}_k\left(e_K^+\right), \quad \forall K \in \mathcal{T}_h. \tag{3.23b}$$

Here $\bar{\eta} = \{u - \Pi_1 u\}^{(\theta)}$. As that in the 1D case, the key is to show that: if δ solves (3.23), then

$$\|\delta\|_{L^2(\mathcal{T}_h)} + h^{\frac{1}{2}} \|\delta\|_{L^2\left(\mathcal{E}_h^+\right)} \le \hat{C}_\theta h^{\frac{1}{2}} \|\bar{\eta}\|_{L^2\left(\mathcal{E}_h^+\right)}, \quad \text{with } \hat{C}_\theta = C\left(1 + \left(\mu^* + \frac{1}{2}\right)\mu_*^{-1}\right). \tag{3.24}$$

[2] In the papers by Cockburn et al., the estimate of $\|u - \Pi_1 u\|_{L^2(K)}$ is proved. The estimate of the trace $\|u - \Pi_1 u\|_{L^2\left(e_K^+\right)}$ can be obtained after applying the inverse trace inequality.

Recall the transversality condition on outflow edges in Lemma 3.2. Since $0 < \gamma \le \hat{\boldsymbol{\beta}} \cdot \boldsymbol{n} \le 1$, $\|\cdot\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{T}_h)}$ and $\|\cdot\|_{L^2(\mathcal{T}_h)}$ are equivalent and $\|\cdot\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)}$ and $\|\cdot\|_{L^2(\mathcal{E}_h^+)}$ are equivalent, upto a positive constant dependent on γ (and hence σ). Therefore, it suffices to show that

$$\|\delta\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{T}_h)} + h^{\frac{1}{2}}\|\delta\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)} \le \hat{C}_\theta h^{\frac{1}{2}}\|\bar{\eta}\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)}, \tag{3.25}$$

which is proved through the following three steps.

Step 1: Estimate of $\|[\delta]\|_{L^2(\mathcal{E}_h^+)}$. Under the assumption (3.3b), we have

$$[\delta]\big|_{e_K^+} \in \mathcal{P}_k\left(e_K^+\right). \tag{3.26}$$

Hence we can take $v = \partial_{\hat{\boldsymbol{\beta}}}\delta := \hat{\boldsymbol{\beta}} \cdot \nabla\delta$ and $w = [\delta]\,\hat{\boldsymbol{\beta}} \cdot \boldsymbol{n}$ in (3.23). After summing over all mesh cells, it then yields

$$\left(\delta, \partial_{\hat{\boldsymbol{\beta}}}\delta\right)_{\mathcal{T}_h} + \sum_{K\in\mathcal{T}_h}\left\langle \{\delta\}^{(\theta)}, [\delta]\,\hat{\boldsymbol{\beta}} \cdot \boldsymbol{n}\right\rangle_{e_K^+} = \sum_{K\in\mathcal{T}_h}\left\langle \bar{\eta}, [\delta]\,\hat{\boldsymbol{\beta}} \cdot \boldsymbol{n}\right\rangle_{e_K^+}. \tag{3.27}$$

Note the left hand side is simply $\mathcal{H}(\delta,\delta;\boldsymbol{\beta})/|\boldsymbol{\beta}|$. Taking the absolute value on both sides and applying (3.16) and (3.17) to the left side, it yields

$$\mu_*\|[\delta]\|^2_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)} \le \sum_{K\in\mathcal{T}_h}\left(\left(\theta_{e_K^+} - \frac{1}{2}\right)\|[\delta]\|^2_{\hat{\boldsymbol{\beta}},L^2(e_K^+)}\right) = \frac{|\mathcal{H}(\delta,\delta;\boldsymbol{\beta})|}{|\boldsymbol{\beta}|} = \left|\sum_{K\in\mathcal{T}_h}\left\langle \bar{\eta}, [\delta]\,\hat{\boldsymbol{\beta}} \cdot \boldsymbol{n}\right\rangle_{e_K^+}\right|. \tag{3.28}$$

We then apply the Cauchy–Schwarz inequality to the right side to get

$$\mu_*\|[\delta]\|^2_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)} \le \|\bar{\eta}\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)}\|[\delta]\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)}, \tag{3.29}$$

which gives

$$\|[\delta]\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)} \le \mu_*^{-1}\|\bar{\eta}\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)}. \tag{3.30}$$

Step 2: Estimate of $\|\delta\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)}$. One can deduce from (3.23b) that

$$\left\langle \delta^-, w\right\rangle_{e_K^+} = \left\langle \bar{\eta} - \tilde{\theta}\,[\delta], w\right\rangle_{e_K^+}, \quad \forall w \in \mathcal{P}_k\left(e_K^+\right). \tag{3.31}$$

Take $w = \hat{\boldsymbol{\beta}} \cdot \boldsymbol{n}\delta^-$, sum over all elements K, and then apply the Cauchy–Schwarz inequality. It yields

$$\begin{aligned}\|\delta^-\|^2_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)} &= \sum_{K\in\mathcal{T}_h}\left\langle \delta^-, \hat{\boldsymbol{\beta}} \cdot \boldsymbol{n}\delta^-\right\rangle_{e_K^+} = \sum_{K\in\mathcal{T}_h}\left\langle \bar{\eta} - \tilde{\theta}\,[\delta], \hat{\boldsymbol{\beta}} \cdot \boldsymbol{n}\delta^-\right\rangle_{e_K^+} \\ &\le \|\bar{\eta} - \tilde{\theta}\,[\delta]\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)}\|\delta^-\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)}.\end{aligned} \tag{3.32}$$

We then divide by $\|\delta^-\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)}$ on both sides, apply the triangle inequality, and recall the estimate of $\|[\delta]\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)}$ in (3.30). It gives

$$\begin{aligned}\|\delta^-\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)} &\le \|\bar{\eta}\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)} + \left(\sup_{K\in\mathcal{T}_h}\left|\tilde{\theta}_{e_K^+}\right|\right)\|[\delta]\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)} \\ &\le \left(1 + \sup_{K\in\mathcal{T}_h}\left|\tilde{\theta}_{e_K^+}\right|\mu_*^{-1}\right)\|\bar{\eta}\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)} \\ &\le \hat{C}_\theta\|\bar{\eta}\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)}.\end{aligned} \tag{3.33}$$

Similarly, with

$$\left\langle \delta^+, w \right\rangle_{e_K^+} = \left\langle \bar{\eta} + \theta \left[\delta\right], w \right\rangle_{e_K^+}, \quad \forall w \in \mathcal{P}_k\left(e_K^+\right), \tag{3.34}$$

we can use a similar argument to obtain $\|\delta^+\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)} \leq \hat{C}_\theta \|\bar{\eta}\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)}$. Therefore we have

$$\|\delta\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)} = \sqrt{\frac{1}{2}\left(\|\delta^+\|^2_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)} + \|\delta^-\|^2_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)}\right)} \leq \hat{C}_\theta \|\bar{\eta}\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{E}_h^+)}. \tag{3.35}$$

Step 3: Estimate of $\|\delta\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{T}_h)}$. Applying Lemma 3.10 with

$$e = e_K^+, \quad Z = \delta\sqrt{\hat{\boldsymbol{\beta}} \cdot \boldsymbol{n}_{e_K^+}}, \quad \text{and} \quad z = \delta^-\sqrt{\hat{\boldsymbol{\beta}} \cdot \boldsymbol{n}_{e_K^+}}, \tag{3.36}$$

one can obtain

$$\|\delta\|_{\hat{\boldsymbol{\beta}},L^2(K)} \leq C h_K^{\frac{1}{2}} \|\delta^-\|_{\hat{\boldsymbol{\beta}},L^2\left(e_K^+\right)}, \tag{3.37}$$

which gives

$$\|\delta\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{T}_h)} = \sqrt{\sum_{K\in\mathcal{T}_h} \|\delta\|^2_{\hat{\boldsymbol{\beta}},L^2(K)}} \leq \sqrt{\sum_{K\in\mathcal{T}_h} C h_K \|\delta^-\|^2_{\hat{\boldsymbol{\beta}},L^2\left(e_K^+\right)}} \leq C h^{\frac{1}{2}} \|\delta^-\|_{\hat{\boldsymbol{\beta}},L^2\left(\mathcal{E}_h^+\right)}. \tag{3.38}$$

Then we use the estimate (3.33) to obtain

$$\|\delta\|_{\hat{\boldsymbol{\beta}},L^2(\mathcal{T}_h)} \leq \hat{C}_\theta h^{\frac{1}{2}} \|\bar{\eta}\|_{\hat{\boldsymbol{\beta}},L^2\left(\mathcal{E}_h^+\right)}. \tag{3.39}$$

We can combine (3.35) and (3.39) to obtain (3.25) and hence (3.24).

Finally, to prove Lemma 3.4, we can use the estimate in (3.39) to show that the solution to (3.23) is unique. Furthermore, through a simple dimension count in Proposition C.1, one can see that (3.23) is a square system, for which the uniqueness of the solution implies the existence of the solution. Hence δ is uniquely solvable. Furthermore, noting that $\bar{\eta} = \{u - \Pi_1 u\}^{(\theta)}$, Lemma 3.9 implies

$$\|\bar{\eta}\|_{L^2(\mathcal{E}_h^+)} \leq C\left(\mu^* + \frac{1}{2}\right)\|u - \Pi_1 u\|_{L^2(\mathcal{E}_h^+)} \leq C\left(\mu^* + \frac{1}{2}\right) h^{k+\frac{1}{2}} |u|_{H^{k+1}(\mathcal{T}_h)}. \tag{3.40}$$

Together with (3.24), it gives

$$\|\delta\|_{L^2(\mathcal{T}_h)} + h^{\frac{1}{2}}\|\delta\|_{L^2\left(\mathcal{E}_h^+\right)} \leq\ C_\theta h^{k+1} |u|_{H^{k+1}(\mathcal{T}_h)}. \tag{3.41}$$

As a result, $\Pi_\theta u = \Pi_1 u + \delta$ is well-defined, with $\Pi_1 u$ admitting the approximation property in Lemma 3.9 and δ admitting the estimate (3.41). The approximation estimate (3.19) can be obtained after applying the triangle inequality

$$\begin{aligned} &\|u - \Pi_\theta u\|_{L^2(\mathcal{T}_h)} + h^{\frac{1}{2}}\|u - \Pi_\theta u\|_{L^2\left(\mathcal{E}_h^+\right)} \\ \leq\ &\|u - \Pi_1 u\|_{L^2(\mathcal{T}_h)} + h^{\frac{1}{2}}\|u - \Pi_1 u\|_{L^2\left(\mathcal{E}_h^+\right)} + \|\delta\|_{L^2(\mathcal{T}_h)} + h^{\frac{1}{2}}\|\delta\|_{L^2\left(\mathcal{E}_h^+\right)} \\ \leq\ &C_\theta h^{k+1} |u|_{H^{k+1}(\mathcal{T}_h)}. \end{aligned} \tag{3.42}$$

□

3.4 Difficulties on extension to 2D Cartesian meshes

In this subsection, we briefly comment on the difficulties of using the energy approach to construct global projection operators for optimal error estimates of the upwind-biased DG methods on 2D Cartesian meshes with either Q^k or P^k elements. Here we only present some native attempts based on the techniques in the previous subsection, and comment on the potential difficulties. There may be ways to circumvent these difficulties and we will leave them to future investigations.

For simplicity, we assume $\boldsymbol{\beta} = (1,1)^T$ and the equation (3.1) becomes

$$u_t + u_x + u_y = 0, \qquad u = u(x,y,t). \tag{3.43}$$

The mesh partition is given by $\Omega = \cup_{i,j}\{K_{ij}\}$, where $K_{ij} = I_i \times J_j = (x_{i-1/2}, x_{i+1/2}) \times (y_{j-1/2}, y_{j+1/2})$. The finite element space is set as

$$V_h = \{v \in L^2(\Omega) : v|_{K_{ij}} \in \mathcal{Z}_k\left(K_{ij}\right), \forall i,j\}. \tag{3.44}$$

Here $\mathcal{Z}_k\left(K_{ij}\right) = \mathcal{P}_k\left(K_{ij}\right)$ for P^k elements and $\mathcal{Z}_k\left(K_{ij}\right) = \mathcal{Q}_k\left(K_{ij}\right)$ for Q^k elements. $\mathcal{Q}_k\left(K_{ij}\right)$ is the space spanned by polynomials on K_{ij} of degree less than or equal to k in each variable. In below, $\theta_1 > 1/2$ and $\theta_2 > 1/2$ are given constant parameters.

3.4.1 Q^k elements

The optimal error estimates of upwind-biased DG methods on 2D Cartesian meshes were proved in [26] using the 2D GGR projection [26, 6]. The 2D GGR projection $\Pi_{\theta_1,\theta_2} := \Pi_{\theta_1} \otimes \Pi_{\theta_2}$ is defined as the tensor product of the one-dimensional projections. To be more specific, for any u, we want to find $\Pi_{\theta_1,\theta_2} u \in V_h$ such that

$$\int_{K_{ij}} \left(\Pi_{\theta_1,\theta_2} u\right) v \mathrm{d}x\mathrm{d}y = \int_{K_{ij}} uv \mathrm{d}x\mathrm{d}y, \qquad \forall v \in \mathcal{Q}_{k-1}(K_{ij}), \tag{3.45a}$$

$$\int_{J_j} \{\Pi_{\theta_1,\theta_2} u\}^{(\theta_1,y)}_{i+\frac{1}{2},y} v \mathrm{d}y = \int_{J_j} \{u\}^{(\theta_1,y)}_{i+\frac{1}{2},y} v \mathrm{d}y, \qquad \forall v \in \mathcal{P}_{k-1}(J_j), \tag{3.45b}$$

$$\int_{I_i} \{\Pi_{\theta_1,\theta_2} u\}^{(x,\theta_2)}_{x,j+\frac{1}{2}} v \mathrm{d}x = \int_{I_i} \{u\}^{(x,\theta_2)}_{x,j+\frac{1}{2}} v \mathrm{d}x, \qquad \forall v \in \mathcal{P}_{k-1}(I_i), \tag{3.45c}$$

$$\{\Pi_{\theta_1,\theta_2} u\}^{(\theta_1,\theta_2)}_{i+\frac{1}{2},j+\frac{1}{2}} = \{u\}^{(\theta_1,\theta_2)}_{i+\frac{1}{2},j+\frac{1}{2}}. \tag{3.45d}$$

Here we have

$$\{w\}^{(\theta_1,y)}_{i+\frac{1}{2},y} = \theta_1 w\left(x^-_{i+\frac{1}{2}}, y\right) + \tilde{\theta}_1 w\left(x^+_{i+\frac{1}{2}}, y\right), \tag{3.46a}$$

$$\{w\}^{(x,\theta_2)}_{x,j+\frac{1}{2}} = \theta_2 w\left(x, y^-_{j+\frac{1}{2}}\right) + \tilde{\theta}_2 w\left(x, y^+_{j+\frac{1}{2}}\right), \tag{3.46b}$$

$$\begin{aligned}\{w\}^{(\theta_1,\theta_2)}_{i+\frac{1}{2},j+\frac{1}{2}} = {} & \theta_1\theta_2 w\left(x^-_{i+\frac{1}{2}}, y^-_{j+\frac{1}{2}}\right) + \theta_1\tilde{\theta}_2 w\left(x^-_{i+\frac{1}{2}}, y^+_{j+\frac{1}{2}}\right) \\ & + \tilde{\theta}_1\theta_2 w\left(x^+_{i+\frac{1}{2}}, y^-_{j+\frac{1}{2}}\right) + \tilde{\theta}_1\tilde{\theta}_2 w\left(x^+_{i+\frac{1}{2}}, y^+_{j+\frac{1}{2}}\right),\end{aligned} \tag{3.46c}$$

and $\tilde{\theta}_l = 1 - \theta_l$ for $l = 1, 2$.

To study the projection Π_{θ_1,θ_2} by the energy approach, we want to construct the bilinear form associated with the linear advection

$$\begin{aligned}\mathcal{H}(w,v) = &\sum_{i,j}\left(\int_{K_{ij}} w\left(a_1 v_x + a_2 v_y\right)\mathrm{d}x\mathrm{d}y\right.\\ &-a_1\int_{J_j}\{w\}^{(\theta_1,y)}_{i+\frac{1}{2},y} v\left(x^-_{i+\frac{1}{2}},y\right)\mathrm{d}y + a_1\int_{J_j}\{w\}^{(\theta_1,y)}_{i-\frac{1}{2},y} v\left(x^+_{i-\frac{1}{2}},y\right)\mathrm{d}y\\ &\left.-a_2\int_{I_i}\{w\}^{(x,\theta_2)}_{x,j+\frac{1}{2}} v\left(x,y^-_{j+\frac{1}{2}}\right)\mathrm{d}x + a_2\int_{I_i}\{w\}^{(x,\theta_2)}_{x,j-\frac{1}{2}} v\left(x,y^+_{j-\frac{1}{2}}\right)\mathrm{d}x\right).\end{aligned} \tag{3.47}$$

Here $w, v \in V_h$ and a_1 and a_2 are some constants that can be chosen in the analysis. However, note that for $v \in \mathcal{Q}_k(K_{ij})$, we may have $v_x, v_y \notin \mathcal{Q}_{k-1}(K_{ij})$. Hence the term $\int_{K_{ij}} w\left(a_1 v_x + a_2 v_y\right)\mathrm{d}x\mathrm{d}y$ in (3.47) may not be directly constructed from (3.45a). This hinders the analysis of the 2D GGR projection on Cartesian meshes by the energy approach. Further investigation is needed to overcome this difficulty.

3.4.2 P^k elements

In [24], Liu et al. proved the optimal error estimates of the upwind DG method with P^k elements for the linear advection equation on 2D Cartesian meshes. The main ingredient of the proof is to construct the special local projection [24, Lemma 2.1]. However, there seems to be very limited results on extending their optimal error estimates to the upwind-biased case, and the exact form of the required projection may not even be clear. A tentative attempt is to generalize the local projection [24, Lemma 2.1] as

$$\int_{K_{ij}} \left(\Pi_{\theta_1,\theta_2} u\right)\mathrm{d}x\mathrm{d}y = \int_{K_{ij}} u\mathrm{d}x\mathrm{d}y, \tag{3.48a}$$

$$\mathcal{L}_{ij}(\Pi_{\theta_1,\theta_2} u, v) = \mathcal{L}_{ij}(u,v), \quad \forall v \in \mathcal{P}_k(K_{ij}), \tag{3.48b}$$

where

$$\begin{aligned}\mathcal{L}_{ij}(w,v) = &\int_{K_{ij}} w\left(v_x + v_y\right)\mathrm{d}x\mathrm{d}y - \int_{J_j}\{w\}^{(\theta_1,y)}_{i+\frac{1}{2},y}\left(v\left(x^-_{i+\frac{1}{2}},y\right) - v\left(x^+_{i-\frac{1}{2}},y\right)\right)\mathrm{d}y\\ &-\int_{I_i}\{w\}^{(x,\theta_2)}_{x,j+\frac{1}{2}}\left(v\left(x,y^-_{j+\frac{1}{2}}\right)\mathrm{d}x - v\left(x,y^+_{j-\frac{1}{2}}\right)\right)\mathrm{d}x,\end{aligned} \tag{3.49}$$

and $\{w\}^{(\theta_1,y)}_{i+1/2,y}$ and $\{w\}^{(x,\theta_2)}_{x,j+1/2}$ are defined in (3.46a) and (3.46b), respectively. When $\theta_1 = \theta_2 = 1$, this retrieves the local projection in [24, Lemma 2.1].

The structure of the projection (3.48) is very different from those of Lemmas 2.1 and 3.4, and it is not easy to derive the bilinear form (3.47) from (3.48). Although the projection (3.48) naturally induces the bilinear form $\mathcal{L}(w,v) = \sum_{ij}\mathcal{L}_{ij}(w,v)$, it seems to be difficult to use $\mathcal{L}(\delta,\delta)$ to control $[\delta]$, and $\mathcal{L}(\delta,\delta)$ may not be used in replace of the bilinear form $\mathcal{H}(\delta,\delta)$ in (3.47).

21

4 Numerical Tests

4.1 1D Tests

The detailed numerical verification of Theorem 2.2 can be found in [26]. In this section, we examine Theorem 2.3 and test the 1D upwind-biased DG methods using polynomials of degrees $k = 1, 2$ and $\theta = 1/2 + h^\omega$ with various values of ω .

Example 4.1. In this test, we solve (2.1) with the initial condition $u(x, 0) = \sin x$ on the domain $\Omega = (0, 2\pi)$ coupled with the periodic boundary condition. The exact solution is $u(x, t) = \sin(x - t)$. The second-order Runge–Kutta method is used for the $k = 1$ case and the third-order Runge–Kutta method is used for the $k = 2$ case. We set $\Delta t = 0.05h$ and use very fine spatial meshes for a clean convergence rate. ω is set as $0.5, 0.75, 1, 2$. We have also tested other values of ω, but the results are very similar and are hence omitted.

In Table 4.1, uniform meshes with N cells are used for computation. In Table 4.2, the meshes are nonuniform and the cell length alternates between $h = 2\pi/N \cdot 4/3$ and $h/2$. Except for P^2 elements on uniform mesh, for which the optimal third-order convergence rate is observed [25], we observe the $(k + \max(1 - \omega, 0))$th order convergence rate in all other cases, which matches the results in Theorem 2.3.

		$\omega = 0.5$		$\omega = 0.75$		$\omega = 1$		$\omega = 2$	
	N	L^2 error	Order	L^2 error	Order	L^2 error	Order	L^2 error	Order
P^1	640	4.068E-05	-	1.263E-04	-	2.864E-04	-	4.883E-04	-
	1280	1.439E-05	1.50	5.388E-05	1.23	1.435E-04	1.00	2.456E-04	0.99
	2560	5.093E-06	1.50	2.284E-05	1.24	7.181E-05	1.00	1.231E-04	1.00
	5120	1.803E-06	1.50	9.643E-06	1.24	3.593E-05	1.00	6.162E-05	1.00
P^2	640	4.765E-09	-	4.725E-09	-	4.739E-09	-	4.756E-09	-
	1280	5.915E-10	3.01	5.923E-10	3.00	5.943E-10	3.00	5.946E-10	3.00
	2560	7.382E-11	3.00	7.413E-11	3.00	7.430E-11	3.00	7.434E-11	3.00
	5120	9.232E-12	3.00	9.275E-12	3.00	9.289E-12	3.00	9.292E-12	3.00

Table 4.1: The L^2 error and the convergence order of upwind-biased DG methods on 1D uniform mesh with N mesh cells. $\theta = 1/2 + h^\omega$.

4.2 2D Tests

In this section, we test the 2D upwind-biased DG methods with polynomials of degrees $k = 1, 2$ and various of parameters $\theta = 0.75$ (under-upwinding), $\theta = 1$ (upwinding), and $\theta = 2$ (over-upwinding). The spatial domain is set as $\Omega = [0, 1] \times [0, 1]$. For periodic boundary conditions, we use the fourth-order Runge–Kutta method for time-marching. The resulted fully discrete scheme is stable under the usual CFL condition $\Delta t \leq Ch$ [33, 34, 39]. For the inflow boundary condition, the fourth-order Lax–Wendroff method is adopted for time discretization to avoid the possible order reduction due to the inflow boundary condition [17].

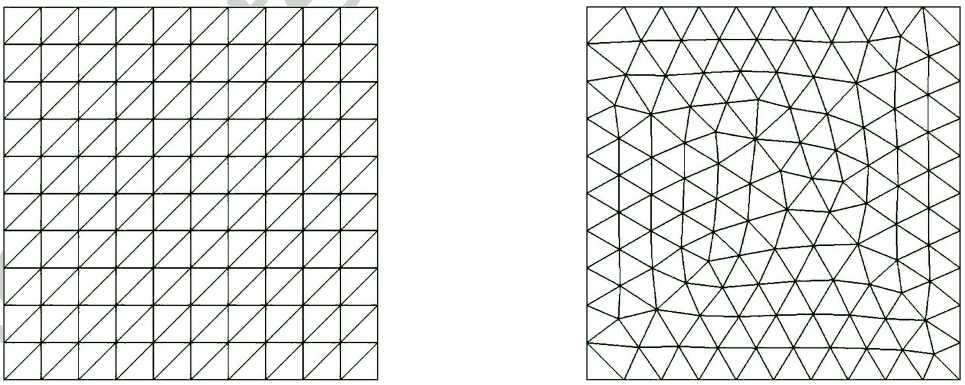
22

		$\omega = 0.5$		$\omega = 0.75$		$\omega = 1$		$\omega = 2$	
	N	L^2 error	Order	L^2 error	Order	L^2 error	Order	L^2 error	Order
P^1	5120	2.462E-06	-	1.442E-05	-	4.719E-05	-	6.852E-05	-
	10240	8.706E-07	1.50	6.098E-06	1.24	2.360E-05	1.00	3.427E-05	1.00
	20480	3.079E-07	1.50	2.571E-06	1.25	1.180E-05	1.00	1.714E-05	1.00
	40960	1.089E-07	1.50	1.080E-06	1.25	5.886E-06	1.00	8.551E-06	1.00
P^2	640	3.936E-08	-	1.322E-07	-	3.419E-07	-	6.147E-07	-
	1280	6.836E-09	2.53	2.799E-08	2.24	8.571E-08	2.00	1.546E-07	1.99
	2560	1.200E-09	2.51	5.911E-09	2.24	2.146E-08	2.00	3.876E-08	2.00
	5120	2.115E-10	2.50	1.244E-09	2.25	5.360E-09	2.00	9.689E-09	2.00

Table 4.2: The L^2 error and the convergence order of upwind-biased DG methods on nonuniform meshes in 1D with the cell length alternating between $h = 2\pi/N \cdot 4/3$ and $h/2$. $\theta = 1/2 + h^\omega$.

Example 4.2. We consider the linear advection equation with $\boldsymbol{\beta} = (1, 1)^T$. The initial condition is set as $u(x, y, 0) = \sin(2\pi(x + y))$. We consider both the periodic boundary condition and the inflow boundary condition. For both cases, the exact solution is given by $u(x, y, t) = \sin(2\pi(x + y - 2t))$ and the final time is set as $T = 0.2$. We take $\Delta t = 0.01/N$ to reduce the temporal error, although a larger time step size can be used in practice.

We use structured triangular meshes in this numerical example. These meshes are generated by splitting the uniform Cartesian meshes by connecting the lower-left and the upper-right nodes in each square. See Figure 4.1(a). This uniform mesh satisfies the presumed flow condition with respect to $\boldsymbol{\beta} = (1, 1)^T$. The numerical results with the periodic and inflow boundary conditions are given in Tables 4.3 and 4.4, respectively. The optimal convergence rates are observed, as that has been proved in Theorem 3.5.



(a) Structured mesh. (b) Unstructured mesh.

Figure 4.1: Meshes for the accuracy test in Examples 4.2 and 4.3 with $N = 10$.

Example 4.3. In this numerical test, we repeat Example 4.2 on unstructured meshes, which are generated with Netgen [29] by specifying the mesh parameters. For example, the mesh with the maximal mesh size $1/N = 1/10$ admitting the periodic boundary condition is

23

		$\theta = 0.75$		$\theta = 1$		$\theta = 2$	
	N	L^2 error	Order	L^2 error	Order	L^2 error	Order
P^1	10	4.819e-02	-	3.820e-02	-	3.592e-02	-
	20	1.152e-02	2.06	9.439e-03	2.02	8.243e-03	2.12
	40	2.802e-03	2.04	2.349e-03	2.01	2.023e-03	2.03
	80	6.945e-04	2.01	5.864e-04	2.00	5.036e-04	2.01
P^2	10	3.854e-03	-	3.333e-03	-	5.982e-03	-
	20	4.615e-04	3.06	4.256e-04	2.97	9.441e-04	2.66
	40	5.654e-05	3.03	5.331e-05	3.00	1.312e-04	2.85
	80	7.039e-06	3.01	6.670e-06	3.00	1.714e-05	2.94

Table 4.3: The L^2 error and the convergence order of upwind-biased DG methods on structured meshes using periodic boundary conditions. The mesh is generated by subdividing a square mesh with $N \times N$ elements.

		$\theta = 0.75$		$\theta = 1$		$\theta = 2$	
	N	L^2 error	Order	L^2 error	Order	L^2 error	Order
P^1	10	4.717e-02	-	3.847e-02	-	3.774e-02	-
	20	1.150e-02	2.04	9.485e-03	2.02	8.420e-03	2.16
	40	2.802e-03	2.04	2.353e-03	2.01	2.044e-03	2.04
	80	6.945e-04	2.01	5.867e-04	2.00	5.061e-04	2.01
P^2	10	3.784e-03	-	3.351e-03	-	5.514e-03	-
	20	4.568e-04	3.05	4.246e-04	2.98	8.637e-04	2.67
	40	5.608e-05	3.03	5.318e-05	3.00	1.214e-04	2.83
	80	6.979e-06	3.01	6.654e-06	3.00	1.597e-05	2.93

Table 4.4: The L^2 error and the convergence order of upwind-biased DG methods on structured meshes using inflow boundary conditions. The mesh is generated by subdividing a square mesh with $N \times N$ elements.

depicted in Figure 4.1(b). Although the meshes do not satisfy the flow condition, we still observe optimal convergence rates. See Tables 4.5 and 4.6.

We have also used the meshes to test the upwind-biased DG methods with $\theta = 1/2 + (1/N)^\omega$ and $\omega = 0.25, 0.5, 0.75, 1, 2$ using P^1 elements. The results are documented in Table 4.7. Degenerated convergence rates are observed as those in the 1D case. By comparing Tables 4.5 and 4.7, it is clear that the degeneracy should be attributed to the vanishing values of $\theta - 1/2$.

Example 4.4. This example is modified from the numerical test in [9]. We consider the linear advection equation with $\boldsymbol{\beta} = (1, 0)^T$. The periodic boundary condition is imposed at $x = 0$ and $x = 1$. Again, we take the initial data to be $u(x, y, 0) = \sin(2\pi(x + y))$ and the corresponding exact solution is $u(x, y, t) = \sin(2\pi(x + y - t))$. We compute to $T = 0.2$ with the time step size $\Delta t = 0.01/N_y$, where N_y is the number of horizontal strips in the mesh partition.

To construct the spatial mesh, we start with a uniform mesh of size $1/N_y$ in Figure 4.2(a). Then we perturb the interior nodes randomly by at most $2/(5N_y)$ along the x direction. See

24

		$\theta = 0.75$		$\theta = 1$		$\theta = 2$	
	N	L^2 error	Order	L^2 error	Order	L^2 error	Order
P^1	10	2.577e-02	-	1.892e-02	-	1.800e-02	-
	20	6.527e-03	1.98	4.508e-03	2.07	3.984e-03	2.18
	40	1.690e-03	1.95	1.143e-03	1.98	9.991e-04	2.00
	80	4.200e-04	2.01	2.813e-04	2.02	2.452e-04	2.03
P^2	10	1.274e-03	-	1.236e-03	-	1.830e-03	-
	20	1.401e-04	3.18	1.417e-04	3.12	2.219e-04	3.04
	40	1.868e-05	2.91	1.909e-05	2.89	3.014e-05	2.88
	80	2.275e-06	3.04	2.346e-06	3.02	3.734e-06	3.01

Table 4.5: The L^2 error and the convergence order of upwind-biased DG methods on unstructured meshes with the mesh parameter $1/N$ using periodic boundary conditions.

		$\theta = 0.75$		$\theta = 1$		$\theta = 2$	
	N	L^2 error	Order	L^2 error	Order	L^2 error	Order
P^1	10	2.341e-02	-	1.742e-02	-	1.592e-02	-
	20	6.234e-03	1.91	4.375e-03	1.99	3.849e-03	2.05
	40	1.643e-03	1.92	1.118e-03	1.97	9.679e-04	1.99
	80	4.183e-04	1.97	2.808e-04	1.99	2.428e-04	2.00
P^2	10	1.103e-03	-	1.088e-03	-	1.643e-03	-
	20	1.321e-04	3.06	1.351e-04	3.01	2.086e-04	2.98
	40	1.743e-05	2.92	1.807e-05	2.90	2.848e-05	2.87
	80	2.177e-06	3.00	2.262e-06	3.00	3.595e-06	2.99

Table 4.6: The L^2 error and the convergence order of upwind-biased DG methods on unstructured meshes with the mesh parameter $1/N$ using inflow boundary conditions.

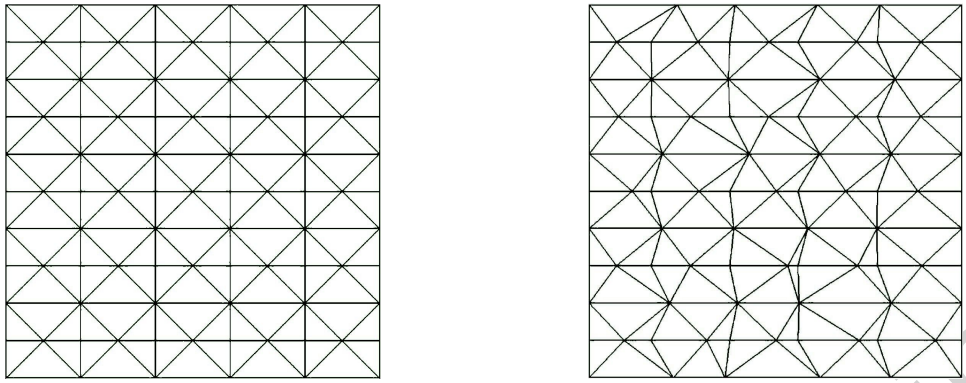
	$\omega = 0.25$		$\omega = 0.5$		$\omega = 0.75$		$\omega = 1$		$\omega = 2$	
N	L^2 error	Order	L^2 error	Order	L^2 error	Order	L^2 error	Order	L^2 error	Order
10	1.841e-02	-	2.260e-02	-	3.193e-02	-	4.534e-02	-	8.807e-02	-
20	4.596e-03	2.00	7.087e-03	1.67	1.264e-02	1.34	2.033e-02	1.16	3.880e-02	1.18
40	1.255e-03	1.87	2.507e-03	1.50	5.458e-03	1.21	9.801e-03	1.05	1.857e-02	1.06
80	3.411e-04	1.88	8.760e-04	1.52	2.302e-03	1.25	4.775e-03	1.04	9.222e-03	1.01
160	9.948e-05	1.78	3.299e-04	1.41	1.056e-03	1.12	2.491e-03	0.94	4.625e-03	1.00
320	2.827e-05	1.82	1.149e-04	1.52	4.353e-04	1.28	1.171e-03	1.09	2.210e-03	1.07

Table 4.7: The L^2 error and the convergence order of upwind-biased DG methods with P^1 elements on unstructured meshes with the mesh parameter $1/N$ using periodic boundary conditions. $\theta = 1/2 + (1/N)^\omega$.

Figure 4.2(b). The resulting mesh is no longer uniform but still satisfies the flow condition with respect to $\boldsymbol{\beta} = (1, 0)^T$. The numerical results are given in Table 4.8. We observe the optimal convergence rates for all cases, as that has been proved in Theorem 3.5.

Example 4.5. This example is modified from the numerical test in [27], which showed that the convergence rate of $k + 1/2$ is sharp for the upwind DG method for linear transport

25



(a) Uniform unperturbed mesh. (b) Nonuniform perturbed mesh.

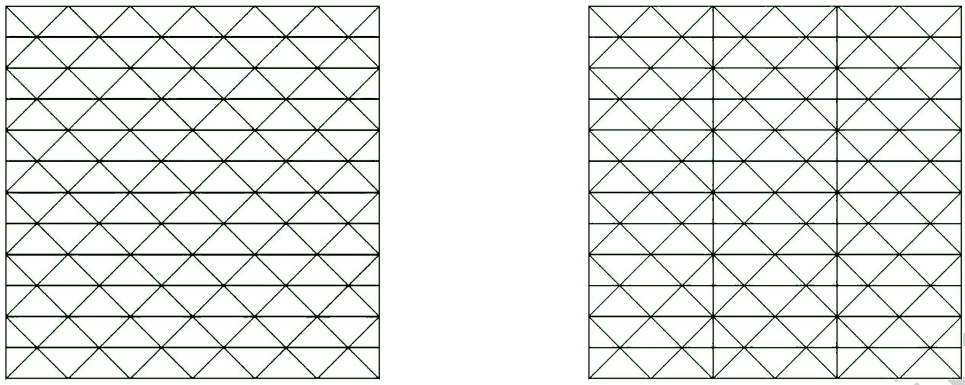
Figure 4.2: Meshes satisfying the flow condition with respect to $\boldsymbol{\beta} = (1, 0)^T$: Uniform mesh with $N_y = 10$ and its nonuniform perturbation. Figure 4.2(b) is used for the accuracy test in Example 4.4.

		$\theta = 0.75$		$\theta = 1$		$\theta = 2$	
	N_y	L^2 error	Order	L^2 error	Order	L^2 error	Order
P^1	10	4.534e-02	-	4.263e-02	-	4.863e-02	-
	20	1.141e-02	1.99	1.073e-02	1.99	1.235e-02	1.98
	40	2.935e-03	1.96	2.743e-03	1.97	3.203e-03	1.95
	80	7.367e-04	1.99	6.856e-04	2.00	8.031e-04	2.00
P^2	10	3.990e-03	-	3.969e-03	-	5.100e-03	-
	20	5.683e-04	2.81	5.722e-04	2.79	7.575e-04	2.75
	40	7.127e-05	3.00	7.204e-05	2.99	9.814e-05	2.95
	80	9.011e-06	2.98	9.112e-06	2.98	1.250e-05	2.97

Table 4.8: The L^2 error and the convergence order of Example 4.4 with meshes having a similar structure in Figure 4.2(b).

over generic triangular meshes. We consider the linear advection equation with $\boldsymbol{\beta} = (0, 1)^T$. The periodic boundary condition is imposed at $y = 0$ and $y = 1$. As before, the initial data is set as $u(x, y, 0) = \sin(2\pi(x + y))$ and the corresponding exact solution is $u(x, y, t) = \sin(2\pi(x + y - t))$. We compute to $T = 0.2$ with $\Delta t = 0.05/N_y$.

To construct the spatial mesh, we start with a uniform mesh of size $1/N_y$ in Figure 4.3(a). Then we add vertical edges to divide the mesh into m vertical strips. When $m = \mathcal{O}(h^{-0.75})$, the reduced convergence rate of $k + 1/2$ is observed with this mesh for the test problem in [27]. The numerical test is given in Table 4.9. We also observe an order degeneration in the convergence rates. This does not contradict our analysis since the flow condition is not satisfied for this set of meshes.



(a) Undivided mesh. (b) Divided mesh with $m = 3$.

Figure 4.3: Mesh structure for the accuracy test in Example 4.5 with $N_y = 12$: Undivided mesh and divided mesh with $m = 3$.

			$\theta = 0.75$		$\theta = 1$		$\theta = 2$	
	N_y	m	L^2 error	Order	L^2 error	Order	L^2 error	Order
P^1	32	8	8.713e-03	-	7.336e-03	-	6.108e-03	-
	128	21	1.047e-03	1.53	7.840e-04	1.61	5.028e-04	1.80
	512	64	1.433e-04	1.43	1.024e-04	1.47	5.724e-05	1.57
P^2	32	8	1.608e-04	-	1.906e-04	-	3.134e-04	-
	128	21	3.446e-06	2.77	4.346e-06	2.73	7.877e-06	2.66
	512	64	9.958e-08	2.56	1.265e-07	2.55	2.197e-07	2.58

Table 4.9: The L^2 error and the convergence order of Example 4.5 using meshes with a similar structure as that in Figure 4.3(b).

5 Conclusions

In this paper, we study the global projection operators using the energy approach developed in [36]. Firstly, we revisit the 1D GGR projection along with optimal error estimates of the upwind-biased DG method for the 1D linear advection equation. In particular, an energy approach is proposed to prove the well-definedness and the approximation property of the 1D GGR projection. Then we extend the argument to multidimensions, which leads to a novel global projection operator on 2D and 3D simplex meshes satisfying the so-called flow condition. This global projection generalizes the local projection in [9] and is used to prove the optimal error estimates of the upwind-biased DG methods for linear advection on these special meshes.

Funding The work of Z. Sun is partially supported by the NSF grant DMS-2208391. The work of Y. Xing is partially supported by the NSF grant DMS-1753581.

Data Availability Datasets generated during the current study are available from the corresponding author upon reasonable request.

Declaration

Conflict of Interest The authors declare that they have no conflict of interest.

A Proof of Lemma 2.8

Before starting, we first state the following proposition, which can be deduced from Proposition 2.5 through symmetry.

Proposition A.1. *Proposition 2.5 holds after replacing* (2.34b) *with* $Z^+_{j-1/2} = z$.

The rest of the section is dedicated to the proof of Lemma 2.8.

Proof of Lemma 2.8. By default, we have $1 \le j \le N$ within the proof. From Propositions 2.5 and A.1, it can be seen that the following local projection is well-defined

$$(\Pi u, v)_{I_j} = (u, v)_{I_j}, \quad \forall v \in \mathcal{P}_{k-1}(I_j), \quad \forall j = 1, \cdots, N, \tag{A.1a}$$

$$(\Pi u)^-_{j+\frac{1}{2}} = u^-_{j+\frac{1}{2}}, \qquad \text{if } \theta_j > \frac{1}{2}, \tag{A.1b}$$

$$(\Pi u)^+_{j-\frac{1}{2}} = u^+_{j-\frac{1}{2}}, \qquad \text{if } \theta_j < \frac{1}{2}. \tag{A.1c}$$

Indeed, it can be equivalently written as

$$\Pi u = \begin{cases} \Pi_1 u, & \text{if } \theta_j > \frac{1}{2}, \\ \Pi_0 u, & \text{if } \theta_j < \frac{1}{2}. \end{cases} \tag{A.2}$$

Here Π_1 and Π_0 correspond to (2.48) with $\theta \equiv 1$ and $\theta \equiv 0$, respectively. All three projections, Π_1, Π_0 and Π, satisfy (2.49) with $C_\theta = C$ independent of μ^* and μ_*.

We denote by $\delta = (\Pi_\theta - \Pi)u$. Subtracting (A.1) from (2.48), one can see that the difference δ satisfies the following equations.

$$(\delta, v)_{I_j} = 0, \qquad v \in \mathcal{P}_{k-1}(I_j), \quad \forall j = 1, \cdots, N, \tag{A.3a}$$

$$\{\delta\}^{(\theta_j)}_{j+\frac{1}{2}} = \bar{\eta}_{j+\frac{1}{2}}, \quad \text{if } \theta_j > \frac{1}{2}, \tag{A.3b}$$

$$\{\delta\}^{(\theta_j)}_{j-\frac{1}{2}} = \bar{\zeta}_{j-\frac{1}{2}}, \quad \text{if } \theta_j < \frac{1}{2}. \tag{A.3c}$$

Here $\bar{\eta}_{j+1/2} = \{u - \Pi_1 u\}^{(\theta_j)}_{j+1/2}$ and $\bar{\zeta}_{j-1/2} = \{u - \Pi_0 u\}^{(\theta_j)}_{j-1/2}$. Using the same argument as that in the proof of Lemma 2.1, it suffices to show the solution to (A.3) satisfies

$$\|\delta\|_{L^2(\mathcal{T}_h)} + h^{\frac{1}{2}} \|\delta\|_{L^2(\mathcal{E}_h^+)} \le \hat{C}_\theta h^{\frac{1}{2}} \left(\|\bar{\eta}\|_{L^2(\mathcal{E}_h^+)} + \|\bar{\zeta}\|_{L^2(\mathcal{E}_h^+)} \right) \tag{A.4}$$

with

$$\hat{C}_\theta = C \left(1 + (\mu^* + 1/2)\, \mu_*^{-1}\right) \left(1 + (\mu^* + 1/2)\, \mu_*^{-1/2}\right) \tag{A.5}$$

to complete the proof. We now proceed to prove (A.4).

To facilitate the discussion, we introduce the index sets

$$J^{-,+}=\left\{j:\theta_j<\frac{1}{2}<\theta_{j+1}\right\},\qquad J^{+,-}=\left\{j:\theta_j>\frac{1}{2}>\theta_{j+1}\right\};\tag{A.6a}$$

$$J^{-,-}=\left\{j:\theta_j<\frac{1}{2},\theta_{j+1}<\frac{1}{2}\right\},\quad J^{+,+}=\left\{j:\theta_j>\frac{1}{2},\theta_{j+1}>\frac{1}{2}\right\},\tag{A.6b}$$

and $\mathcal{E}_h^{-,+}$, $\mathcal{E}_h^{+,-}$, $\mathcal{E}_h^{-,-}$ and $\mathcal{E}_h^{+,+}$ for corresponding sets of $\{x_{j+1/2}\}$. It can be seen that we are imposing one condition on $\mathcal{E}_h^{-,-}$ and $\mathcal{E}_h^{+,+}$, two conditions on $\mathcal{E}_h^{+,-}$, and no condition on $\mathcal{E}_h^{-,+}$, for each mesh point.

Step 1: *Estimate of* $\|\delta\|_{L^2(\mathcal{E}_h^{+,-})}$. Note that for $x_{j+\frac{1}{2}}\in\mathcal{E}_h^{+,-}$, we have

$$\theta_j\delta^-_{j+\frac{1}{2}}+\tilde{\theta}_j\delta^+_{j+\frac{1}{2}}=\bar{\eta}_{j+\frac{1}{2}},\tag{A.7a}$$

$$\theta_{j+1}\delta^-_{j+\frac{1}{2}}+\tilde{\theta}_{j+1}\delta^+_{j+\frac{1}{2}}=\bar{\zeta}_{j+\frac{1}{2}}.\tag{A.7b}$$

Since $\theta\neq\tilde{\theta}$, we can solve the equation system (A.7) to get

$$\delta^-_{j+\frac{1}{2}}=\frac{\tilde{\theta}_{j+1}\bar{\eta}_{j+\frac{1}{2}}-\tilde{\theta}_j\bar{\zeta}_{j+\frac{1}{2}}}{\tilde{\theta}_{j+1}-\tilde{\theta}_j}\quad\text{and}\quad\delta^+_{j+\frac{1}{2}}=\frac{\theta_{j+1}\bar{\eta}_{j+\frac{1}{2}}-\theta_j\bar{\zeta}_{j+\frac{1}{2}}}{\theta_{j+1}-\theta_j}.\tag{A.8}$$

Recall our assumption $0<\mu_*\le|\theta_j-1/2|\le\mu^*<+\infty$ and let us define

$$\kappa=\left(\mu^*+\frac{1}{2}\right)\mu_*^{-1}.\tag{A.9}$$

Then it can be estimated that

$$\|\delta\|_{L^2(\mathcal{E}_h^{+,-})}\le C\kappa\left(\|\bar{\eta}\|_{L^2(\mathcal{E}_h^{+,-})}+\|\bar{\zeta}\|_{L^2(\mathcal{E}_h^{+,-})}\right).\tag{A.10}$$

Step 2: *Estimate of* $\|\delta\|_{L^2(\mathcal{E}_h^{-,+})}$ *and* $\|[\delta]\|_{L^2(\mathcal{E}_h^{+,+}\cup\mathcal{E}_h^{-,-})}$. From (A.3a), it can be seen that

$$\begin{aligned}
0=&\sum_{j:\theta_j>\frac{1}{2}}(\delta,\delta_x)_{I_j}-\sum_{j:\theta_j<\frac{1}{2}}(\delta,\delta_x)_{I_j}\\
=&\frac{1}{2}\sum_{j:\theta_j>\frac{1}{2}}\left(\left|\delta^-_{j+\frac{1}{2}}\right|^2-\left|\delta^+_{j-\frac{1}{2}}\right|^2\right)-\frac{1}{2}\sum_{j:\theta_j<\frac{1}{2}}\left(\left|\delta^-_{j+\frac{1}{2}}\right|^2-\left|\delta^+_{j-\frac{1}{2}}\right|^2\right)\\
=&\frac{1}{2}\sum_{j\in J^{+,+}}\left(\left|\delta^-_{j+\frac{1}{2}}\right|^2-\left|\delta^+_{j+\frac{1}{2}}\right|^2\right)-\frac{1}{2}\sum_{j\in J^{-,-}}\left(\left|\delta^-_{j+\frac{1}{2}}\right|^2-\left|\delta^+_{j+\frac{1}{2}}\right|^2\right)\\
&-\frac{1}{2}\sum_{j\in J^{-,+}}\left(\left|\delta^-_{j+\frac{1}{2}}\right|^2+\left|\delta^+_{j+\frac{1}{2}}\right|^2\right)+\frac{1}{2}\sum_{j\in J^{+,-}}\left(\left|\delta^-_{j+\frac{1}{2}}\right|^2+\left|\delta^+_{j+\frac{1}{2}}\right|^2\right)\\
=&-\sum_{j\in J^{+,+}}\{\delta\}^{(1/2)}_{j+\frac{1}{2}}[\delta]_{j+\frac{1}{2}}+\sum_{j\in J^{-,-}}\{\delta\}^{(1/2)}_{j+\frac{1}{2}}[\delta]_{j+\frac{1}{2}}-\|\delta\|^2_{L^2(\mathcal{E}_h^{-,+})}+\|\delta\|^2_{L^2(\mathcal{E}_h^{+,-})}.
\end{aligned}\tag{A.11}$$

Here we have used the identity $|\delta^+|^2 - |\delta^-|^2 = 2\{\delta\}^{(1/2)}\,[\delta]$. With (A.3b) and (A.3c), it can be shown that

$$\sum_{j\in J^{+,+}} \{\delta\}^{(\theta_j)}_{j+\frac{1}{2}}\,[\delta]_{j+\frac{1}{2}} - \sum_{j\in J^{-,-}} \{\delta\}^{(\theta_j)}_{j+\frac{1}{2}}\,[\delta]_{j+\frac{1}{2}} = \sum_{j\in J^{+,+}} \bar{\eta}_{j+\frac{1}{2}}\,[\delta]_{j+\frac{1}{2}} - \sum_{j\in J^{-,-}} \bar{\zeta}_{j+\frac{1}{2}}\,[\delta]_{j+\frac{1}{2}}\,. \tag{A.12}$$

Note that $\{\delta\}^{(\theta_j)}_{j+\frac{1}{2}} = \{\delta\}^{(1/2)}_{j+\frac{1}{2}} - (\theta_j - 1/2)\,[\delta]_{j+\frac{1}{2}}$. Combining (A.11) and (A.12) yields

$$\sum_{j\in J^{+,+}\cup J^{-,-}} \left|\theta_j - \frac{1}{2}\right| [\delta]^2_{j+\frac{1}{2}} + \|\delta\|^2_{L^2(\mathcal{E}_h^{-,+})} - \|\delta\|^2_{L^2(\mathcal{E}_h^{+,-})} = \sum_{j\in J^{-,-}} \bar{\zeta}_{j+\frac{1}{2}}\,[\delta]_{j+\frac{1}{2}} - \sum_{j\in J^{+,+}} \bar{\eta}_{j+\frac{1}{2}}\,[\delta]_{j+\frac{1}{2}} \tag{A.13}$$

Using the assumption $|\theta_j - 1/2| \geq \mu_*$ and the Cauchy–Schwartz inequality on the right side, we can obtain

$$\begin{aligned} &\mu_* \|\,[\delta]\,\|^2_{L^2(\mathcal{E}_h^{+,+}\cup\mathcal{E}_h^{-,-})} + \|\delta\|^2_{L^2(\mathcal{E}_h^{-,+})} - \|\delta\|^2_{L^2(\mathcal{E}_h^{+,-})} \\ &\leq \|\bar{\eta}\|_{L^2(\mathcal{E}_h^{+,+})} \|\,[\delta]\,\|_{L^2(\mathcal{E}_h^{+,+})} + \|\bar{\zeta}\|_{L^2(\mathcal{E}_h^{-,-})} \|\,[\delta]\,\|_{L^2(\mathcal{E}_h^{-,-})}. \end{aligned} \tag{A.14}$$

Using the inequality $ab \leq (a^2+b^2)/2$ on the right side, we can simplify the inequality as

$$\frac{\mu_*}{2} \|\,[\delta]\,\|^2_{L^2(\mathcal{E}_h^{+,+}\cup\mathcal{E}_h^{-,-})} + \|\delta\|^2_{L^2(\mathcal{E}_h^{-,+})} \leq (2\mu_*)^{-1} \left(\|\bar{\eta}\|^2_{L^2(\mathcal{E}_h^{+,+})} + \|\bar{\zeta}\|^2_{L^2(\mathcal{E}_h^{-,-})}\right) + \|\delta\|^2_{L^2(\mathcal{E}_h^{+,-})}. \tag{A.15}$$

Taking the square root and using the inequality $(|a|+|b|)/\sqrt{2} \leq \sqrt{a^2+b^2} \leq |a|+|b|$ yield

$$\mu_*^{\frac{1}{2}} \|\,[\delta]\,\|_{L^2(\mathcal{E}_h^{+,+}\cup\mathcal{E}_h^{-,-})} + \|\delta\|_{L^2(\mathcal{E}_h^{-,+})} \leq C\mu_*^{-\frac{1}{2}} \left(\|\bar{\eta}\|_{L^2(\mathcal{E}_h^{+,+})} + \|\bar{\zeta}\|_{L^2(\mathcal{E}_h^{-,-})}\right) + C\|\delta\|_{L^2(\mathcal{E}_h^{+,-})}. \tag{A.16}$$

Note that $(\mu_*)^{-\frac{1}{2}} \leq 1/2 + 1/(2\mu_*) \leq C\,(1+\kappa)$. Combining with (A.10), it can be shown that

$$\begin{aligned} &\mu_*^{\frac{1}{2}} \|\,[\delta]\,\|_{L^2(\mathcal{E}_h^{+,+}\cup\mathcal{E}_h^{-,-})} + \|\delta\|_{L^2(\mathcal{E}_h^{-,+})} \\ &\leq C(1+\kappa) \left(\|\bar{\eta}\|_{L^2(\mathcal{E}_h^{+,+}\cup\mathcal{E}_h^{+,-})} + \|\bar{\zeta}\|_{L^2(\mathcal{E}_h^{-,-}\cup\mathcal{E}_h^{+,-})}\right) \\ &\leq C(1+\kappa) \left(\|\bar{\eta}\|_{L^2(\mathcal{E}_h^{+})} + \|\bar{\zeta}\|_{L^2(\mathcal{E}_h^{+})}\right). \end{aligned} \tag{A.17}$$

Step 3: *Estimate of* $\|\delta\|_{L^2(\mathcal{E}_h^{+})}$. From (A.3b) and (A.3c), we have

$$\delta^-_{j+\frac{1}{2}} = \bar{\eta}_{j+\frac{1}{2}} - \tilde{\theta}_j\,[\delta]_{j+\frac{1}{2}}\,, \quad \forall j \in \mathcal{E}_h^{+,+}. \tag{A.18a}$$

$$\delta^+_{j+\frac{1}{2}} = \bar{\zeta}_{j+\frac{1}{2}} + \theta_j\,[\delta]_{j+\frac{1}{2}}\,, \quad \forall j \in \mathcal{E}_h^{-,-}. \tag{A.18b}$$

With the triangle inequality, the estimate (A.17), and the fact $|\theta_j|, |\tilde{\theta}_j| \leq \mu^* + 1/2 = \kappa\mu_*$, it can be seen that

$$\|\delta\|_{L^2(\mathcal{E}_h^{+,+}\cup\mathcal{E}_h^{-,-})} \leq \left(1 + C(1+\kappa)\kappa\mu_*^{\frac{1}{2}}\right) \left(\|\bar{\eta}\|_{L^2(\mathcal{E}_h^{+})} + \|\bar{\zeta}\|_{L^2(\mathcal{E}_h^{+})}\right). \tag{A.19}$$

Therefore, with (A.10), (A.17), and (A.19), we have

$$\|\delta\|_{L^2(\mathcal{E}_h^{+})} \leq \left(1 + C(1+\kappa)\left(1 + \kappa\mu_*^{\frac{1}{2}}\right)\right) \left(\|\bar{\eta}\|_{L^2(\mathcal{E}_h^{+})} + \|\bar{\zeta}\|_{L^2(\mathcal{E}_h^{+})}\right) \leq \hat{C}_\theta \left(\|\bar{\eta}\|_{L^2(\mathcal{E}_h^{+})} + \|\bar{\zeta}\|_{L^2(\mathcal{E}_h^{+})}\right). \tag{A.20}$$

30

Step 4: *Estimate of* $\|\delta\|_{L^2(\mathcal{T}_h)}$. We can deduce from Propositions 2.5 and A.1 that

$$\|\delta\|_{L^2(I_j)} \le C h_j^{\frac{1}{2}} \left|\delta^-_{j+\frac{1}{2}}\right|, \quad \forall j : \theta_j > \frac{1}{2}. \tag{A.21a}$$

$$\|\delta\|_{L^2(I_j)} \le C h_j^{\frac{1}{2}} \left|\delta^+_{j-\frac{1}{2}}\right|, \quad \forall j : \theta_j < \frac{1}{2}. \tag{A.21b}$$

Taking the square and summing over all mesh cells yields

$$\|\delta\|^2_{L^2(\mathcal{T}_h)} = \sum_{j=1}^{N} \|\delta\|^2_{L^2(I_j)} \le Ch \|\delta\|^2_{L^2(\mathcal{E}_h^+)}. \tag{A.22}$$

Apply the estimate in (A.20) and take the square root. One can obtain

$$\|\delta\|_{L^2(\mathcal{T}_h)} \le \hat{C}_\theta h^{\frac{1}{2}} \left(\|\bar{\eta}\|_{L^2(\mathcal{E}_h^+)} + \|\bar{\zeta}\|_{L^2(\mathcal{E}_h^+)} \right). \tag{A.23}$$

Finally, the estimate (A.4) can be obtained by combining (A.20) and (A.23). □

B Proof of Lemma 3.2

Proof. Two-dimensional case: It is known that the shape-regularity condition (3.2) is equivalent to the following minimal angle condition in 2D (also known as the Zlámal's condition [8, Exercise 3.1.3]):

$$\text{There exists a constant } \alpha_0 > 0, \text{ such that } \alpha_K \ge \alpha_0 \text{ for all } K \in \mathcal{T}_h, \tag{B.1}$$

where α_K is the minimum angle of K.

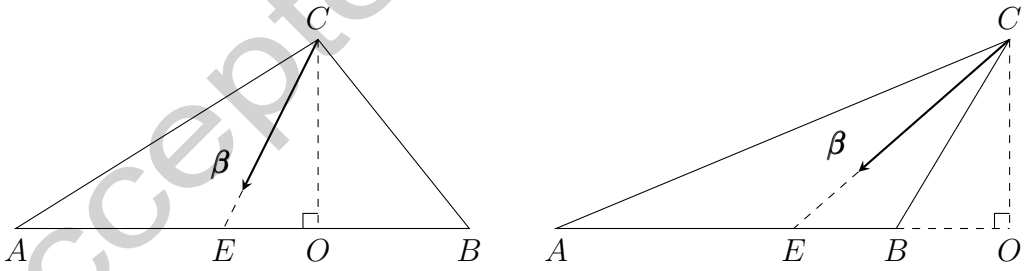


Figure B.4: Triangular elements for the proof of Lemma 3.2 in 2D.

Now we consider the triangular elements $K = \triangle ABC$ in Figure B.4. Let $e_K^+ = AB$ be the outflow edge. Suppose $\boldsymbol{\beta}$ starts at C. Due to the flow condition (3.3a), we must have the extension of $\boldsymbol{\beta}$ intersect the line segment AB at some point E. Furthermore, we set O to be the foot of the altitude from C to AB. Then we will have either $|OA| \ge |OE|$ or $|OB| \ge |OE|$. Without loss of generality, we assume $|OA| \ge |OE|$, which implies $\angle OCE \le \angle OCA$. As a result, we have

$$\boldsymbol{\beta} \cdot \boldsymbol{n}_{e_K^+} = |\boldsymbol{\beta}| \cos \angle OCE \ge |\boldsymbol{\beta}| \cos \angle OCA = |\boldsymbol{\beta}| \sin \angle OAC \ge |\boldsymbol{\beta}| \sin \alpha_0 > 0. \tag{B.2}$$

31

Here we have used the Zlámal's condition (B.1). Hence the transversality condition (3.4) holds with $\gamma = \sin\alpha_0$.

Three-dimensional case: In 3D, it is proven in [3] that the shape-regularity condition (3.2) is equivalent to the following minimal angle condition.

There exists a constant $\alpha_0 > 0$, such that for any simplex $K \in \mathcal{T}_h$, any dihedral angle α, and any solid angle α of K, we have $\alpha \geq \alpha_0$. (B.3)

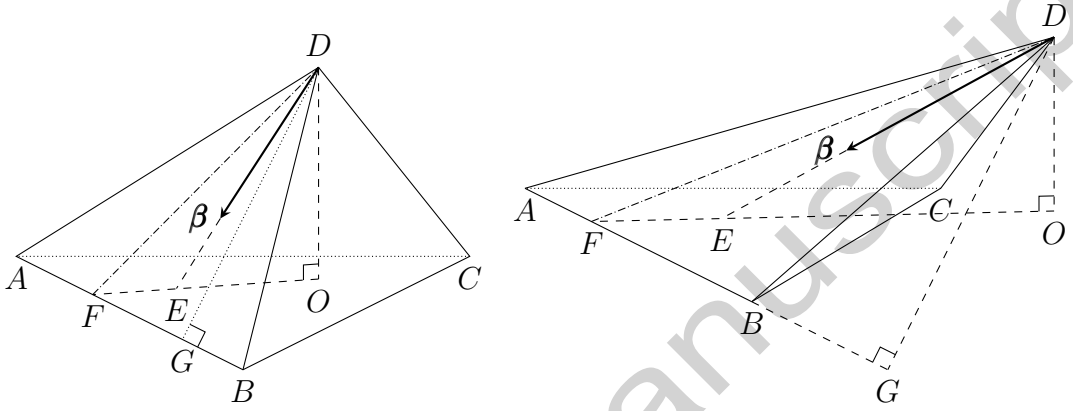


Figure B.5: Tetrahedral elements for the proof of Lemma 3.2 in 3D.

Now let us consider the tetrahedrons $K = ABCD$ in Figure B.5. We assume $e_K^+ = \triangle ABC$ to be the outflow face. Suppose $\boldsymbol{\beta}$ starts at D. Due to the flow condition (3.3a), to ensure a unique outflow face, we need that the extension of $\boldsymbol{\beta}$ intersect $\triangle ABC$ within the triangle at some point E. Furthermore, we set O to be the foot of the altitude from D to $\triangle ABC$. We connect OE and extend it until it intersects the edge of $\triangle ABC$ at some point F (so that $|OF| \geq |OE|$). Then we must have

$$\cos\angle ODE \geq \cos\angle ODF = \sin\angle OFD. \tag{B.4}$$

Without loss of generality, we assume that F is on the edge AB. Note we have either $|OA| \geq |OF|$ or $|OB| \geq |OF|$. We only consider the case $|OA| \geq |OF|$ and the other case can be proved similarly. When $|OA| \geq |OF|$, we have

$$\sin\angle OFD \geq \sin\angle OAD = \frac{|OD|}{|DA|}. \tag{B.5}$$

Then we set G to be the foot of the altitude from D to AB on $\triangle ABD$. In can be seen that

$$\frac{|OD|}{|DA|} = \frac{|OD|}{|DG|}\cdot\frac{|DG|}{|DA|} = \sin\angle OGD\cdot\sin\angle GAD \geq \sin^2\alpha_0. \tag{B.6}$$

Here we have used the fact that $\angle OGD$ is the dihedral angle between the plane ABC and the plane ABD and $\angle GAD$ is a solid angle in $\triangle ABD$, which are both greater than or equal

to α_0 according to the minimal angle condition (B.3). Combining (B.4), (B.5), and (B.6), we get

$$\boldsymbol{\beta} \cdot \boldsymbol{n}_{e_K^+} = |\boldsymbol{\beta}| \cos \angle ODE \geq |\boldsymbol{\beta}| \sin^2 \alpha_0 > 0. \tag{B.7}$$

Hence the transversality condition (3.4) holds with $\gamma = \sin^2 \alpha_0$.

□

C Dimension count in the proof of Lemma 3.4

Proposition C.1. *The finite-dimensional linear system determined by* (3.23) *is square.*

Proof. On each mesh cell $K \in \mathcal{T}_h$, the degrees of freedom of the unknown δ is $\dim(\mathcal{P}_k(K))$, the number of equations associated with (3.23a) is $\dim(\mathcal{P}_{k-1}(K))$, and the number of equations associated with (3.23b) is $\dim(\mathcal{P}_k(e_K^+))$. Since

$$\dim\left(\mathcal{P}_k(K)\right) = \binom{k+d}{d}, \ \dim\left(\mathcal{P}_{k-1}(K)\right) = \binom{k-1+d}{d}, \ \dim\left(\mathcal{P}_k\left(e_K^+\right)\right) = \binom{k+d-1}{d-1}, \tag{C.8}$$

and

$$\binom{k+d}{d} = \binom{k-1+d}{d} + \binom{k+d-1}{d-1}, \tag{C.9}$$

we know that $\dim(\mathcal{P}_k(K)) = \dim(\mathcal{P}_{k-1}(K)) + \dim(\mathcal{P}_k(e_K^+))$ — the degrees of freedom equals to the number of equations on each mesh cell. As a result, the global system (3.23) is square with $|\mathcal{T}_h|\binom{k+d}{d}$ unknowns, where $|\mathcal{T}_h|$ is the number of mesh cells. □

References

[1] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini. Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM Journal on Numerical Analysis*, 39(5):1749–1779, 2002.

[2] J. Bona, H. Chen, O. Karakashian, and Y. Xing. Conservative, discontinuous Galerkin–methods for the generalized Korteweg–de Vries equation. *Mathematics of Computation*, 82(283):1401–1432, 2013.

[3] J. Brandts, S. Korotov, and M. Křížek. On the equivalence of regularity criteria for triangular and tetrahedral finite element partitions. *Computers & Mathematics with Applications*, 55(10):2227–2233, 2008.

[4] P. Castillo, B. Cockburn, D. Schötzau, and C. Schwab. Optimal a priori error estimates for the *hp*-version of the local discontinuous Galerkin method for convection–diffusion problems. *Mathematics of computation*, 71(238):455–478, 2002.

[5] Y. Cheng, C.-S. Chou, F. Li, and Y. Xing. L^2 stable discontinuous Galerkin methods for one-dimensional two-way wave equations. *Mathematics of Computation*, 86(303):121–155, 2017.

33

[6] Y. Cheng, X. Meng, and Q. Zhang. Application of generalized Gauss–Radau projections for the local discontinuous Galerkin method for linear convection-diffusion equations. *Mathematics of Computation*, 86(305):1233–1267, 2017.

[7] Y. Cheng and Q. Zhang. Local analysis of the local discontinuous Galerkin method with generalized alternating numerical flux for one-dimensional singularly perturbed problem. *Journal of Scientific Computing*, 72(2):792–819, 2017.

[8] P. G. Ciarlet. *The finite element method for elliptic problems*. North Holland, 1978.

[9] B. Cockburn, B. Dong, and J. Guzmán. Optimal convergence of the original DG method for the transport-reaction equation on special meshes. *SIAM Journal on Numerical Analysis*, 46(3):1250–1265, 2008.

[10] B. Cockburn, B. Dong, and J. Guzmán. A superconvergent LDG-hybridizable Galerkin method for second-order elliptic problems. *Mathematics of Computation*, 77(264):1887–1916, 2008.

[11] B. Cockburn, B. Dong, J. Guzmán, and J. Qian. Optimal convergence of the original DG method on special meshes for variable transport velocity. *SIAM Journal on Numerical Analysis*, 48(1):133–146, 2010.

[12] B. Cockburn, J. Gopalakrishnan, and F.-J. Sayas. A projection-based error analysis of HDG methods. *Mathematics of Computation*, 79(271):1351–1367, 2010.

[13] B. Cockburn, G. Kanschat, I. Perugia, and D. Schötzau. Superconvergence of the local discontinuous Galerkin method for elliptic problems on Cartesian grids. *SIAM Journal on Numerical Analysis*, 39(1):264–285, 2001.

[14] B. Cockburn, G. E. Karniadakis, and C.-W. Shu. *Discontinuous Galerkin methods: theory, computation and applications*, volume 11. Springer Science & Business Media, 2012.

[15] B. Cockburn and C.-W. Shu. Runge–Kutta discontinuous Galerkin methods for convection-dominated problems. *Journal of Scientific Computing*, 16(3):173–261, 2001.

[16] B. Dong. Optimally convergent hdg method for third-order Korteweg–de Vries type equations. *Journal of Scientific Computing*, 73(2):712–735, 2017.

[17] G. Fu and C.-W. Shu. Optimal energy-conserving discontinuous Galerkin methods for linear symmetric hyperbolic systems. *Journal of Computational Physics*, 394(1):329–363, 2019.

[18] P. Fu, Y. Cheng, F. Li, and Y. Xu. Discontinuous Galerkin methods with optimal L^2 accuracy for one dimensional linear PDEs with high order spatial derivatives. *Journal of Scientific Computing*, 78(2):816–863, 2019.

[19] J. Li, D. Zhang, X. Meng, and B. Wu. Analysis of discontinuous Galerkin methods with upwind-biased fluxes for one dimensional linear hyperbolic equations with degenerate variable coefficients. *Journal of Scientific Computing*, 78(3):1305–1328, 2019.

34

[20] J. Li, D. Zhang, X. Meng, and B. Wu. Analysis of local discontinuous Galerkin methods with generalized numerical fluxes for linearized KdV equations. *Mathematics of Computation*, 89(325):2085–2111, 2020.

[21] J. Li, D. Zhang, X. Meng, B. Wu, and Q. Zhang. Discontinuous Galerkin methods for nonlinear scalar conservation laws: Generalized local Lax–Friedrichs numerical fluxes. *SIAM Journal on Numerical Analysis*, 58(1):1–20, 2020.

[22] M. Liu, B. Wu, and X. Meng. Optimal error estimates of the discontinuous Galerkin method with upwind-biased fluxes for 2d linear variable coefficients hyperbolic equations. *Journal of Scientific Computing*, 83(1):1–19, 2020.

[23] Y. Liu, C.-W. Shu, and M. Zhang. Optimal error estimates of the semidiscrete central discontinuous Galerkin methods for linear hyperbolic equations. *SIAM Journal on Numerical Analysis*, 56(1):520–541, 2018.

[24] Y. Liu, C.-W. Shu, and M. Zhang. Optimal error estimates of the semidiscrete discontinuous Galerkin methods for two dimensional hyperbolic equations on Cartesian meshes using P^k elements. *ESAIM: Mathematical Modelling and Numerical Analysis*, 54:705–726, 2020.

[25] Y. Liu, C.-W. Shu, and M. Zhang. Sub-optimal convergence of discontinuous Galerkin methods with central fluxes for linear hyperbolic equations with even degree polynomial approximations. *Journal of Computational Mathematics*, 39(4):518–537, 2021.

[26] X. Meng, C.-W. Shu, and B. Wu. Optimal error estimates for discontinuous Galerkin methods based on upwind-biased fluxes for linear hyperbolic equations. *Mathematics of Computation*, 85(299):1225–1261, 2016.

[27] T. E. Peterson. A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation. *SIAM Journal on Numerical Analysis*, 28(1):133–140, 1991.

[28] W. H. Reed and T. Hill. Triangular mesh methods for the neutron transport equation. Technical report, Los Alamos Scientific Lab., N. Mex. (USA), 1973.

[29] J. Schöberl. NETGEN an advancing front 2D/3D-mesh generator based on abstract rules. *Computing and Visualization in Science*, 1(1):41–52, 1997.

[30] C.-W. Shu. Discontinuous Galerkin methods: general approach and stability. *Numerical Solutions of Partial Differential Equations*, 201, 2009.

[31] J. Sun, C.-W. Shu, and Y. Xing. Discontinuous Galerkin methods for stochastic Maxwell equations with multiplicative noise. *ESAIM: Mathematical Modelling and Numerical Analysis*, 2022. `https://doi.org/10.1051/m2an/2022084`.

[32] J. Sun, C.-W. Shu, and Y. Xing. Multi-symplectic discontinuous Galerkin methods for the stochastic Maxwell equations with additive noise. *Journal of Computational Physics*, 461:111199, 2022.

35

[33] Z. Sun and C.-W. Shu. Stability of the fourth order Runge–Kutta method for time-dependent partial differential equations. *Annals of Mathematical Sciences and Applications*, 2(2):255–284, 2017.

[34] Z. Sun and C.-W. Shu. Strong stability of explicit Runge–Kutta time discretizations. *SIAM Journal on Numerical Analysis*, 57(3):1158–1182, 2019.

[35] Z. Sun and C.-W. Shu. Error analysis of Runge–Kutta discontinuous Galerkin methods for linear time-dependent partial differential equations. *arXiv preprint arXiv:2001.00971*, 2020.

[36] Z. Sun and Y. Xing. Optimal error estimates of discontinuous Galerkin methods with generalized fluxes for wave equations on unstructured meshes. *Mathematics of Computation*, 90(330):1741–1772, 2021.

[37] H. Wang, Q. Zhang, and C.-W. Shu. Implicit–explicit local discontinuous Galerkin methods with generalized alternating numerical fluxes for convection–diffusion problems. *Journal of Scientific Computing*, 81(3):2080–2114, 2019.

[38] Y. Xu, C.-W. Shu, and Q. Zhang. Error estimate of the fourth-order Runge–Kutta discontinuous Galerkin methods for linear hyperbolic equations. *SIAM Journal on Numerical Analysis*, 58(5):2885–2914, 2020.

[39] Y. Xu, Q. Zhang, C.-W. Shu, and H. Wang. The L^2-norm stability analysis of Runge–Kutta discontinuous Galerkin methods for linear hyperbolic equations. *SIAM Journal on Numerical Analysis*, 57(4):1574–1601, 2019.

[40] Y. Xu, D. Zhao, and Q. Zhang. Local error estimates for Runge–Kutta discontinuous Galerkin methods with upwind-biased numerical fluxes for a linear hyperbolic equation in one-dimension with discontinuous initial data. *Journal of Scientific Computing*, 91(1):1–30, 2022.

[41] H. Zhang, B. Wu, and X. Meng. A local discontinuous Galerkin method with generalized alternating fluxes for 2D nonlinear Schrödinger equations. *Communications on Applied Mathematics and Computation*, 4(1):84–107, 2022.

[42] H. Zhang, B. Wu, and X. Meng. Analysis of the local discontinuous galerkin method with generalized fluxes for one-dimensional nonlinear convection-diffusion systems. *Science China Mathematics*, 2022, 65. `https://doi.org/10.1007/s11425-022-2035-y`.

[43] Q. Zhang and C.-W. Shu. Stability analysis and a priori error estimates of the third order explicit Runge–Kutta discontinuous Galerkin method for scalar conservation laws. *SIAM Journal on Numerical Analysis*, 48(3):1038–1063, 2010.

[44] D. Zhao and Q. Zhang. Local discontinuous Galerkin methods with generalized alternating numerical fluxes for two-dimensional linear Sobolev equation. *Journal of Scientific Computing*, 78(3):1660–1690, 2019.

36