ESAIM: M2AN 49 (2015) 991–1018
DOI: 10.1051/m2an/2014063

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

A PRIORI ERROR ESTIMATES TO SMOOTH SOLUTIONS OF THE THIRD ORDER RUNGE–KUTTA DISCONTINUOUS GALERKIN METHOD FOR SYMMETRIZABLE SYSTEMS OF CONSERVATION LAWS *, **

Juan Luo[1], Chi-Wang Shu[2] and Qiang Zhang[1]

Abstract. In this paper we present an *a priori* error estimate of the Runge–Kutta discontinuous Galerkin method for solving symmetrizable conservation laws, where the time is discretized with the third order explicit total variation diminishing Runge–Kutta method and the finite element space is made up of piecewise polynomials of degree $k \geq 2$. Quasi-optimal error estimate is obtained by energy techniques, for the so-called generalized E-fluxes under the standard temporal-spatial CFL condition $\tau \leq \gamma h$, where h is the element length and τ is time step, and γ is a positive constant independent of h and τ. Optimal estimates are also considered when the upwind numerical flux is used.

Mathematics Subject Classification. 65M60, 65M12.

Received August 22, 2013. Revised November 2, 2014.
Published online June 19, 2015.

1. Introduction

In this paper we would like to continue the works in [29–31] and present error estimates of the Runge–Kutta discontinuous Galerkin (RKDG) method for smooth solutions of symmetrizable systems of conservation laws. For simplicity of presentation, we consider the model equation in the spatial domain $I = (0, 1)$ and the time interval $[0, T]$,

$$\boldsymbol{u}_t + \boldsymbol{f}(\boldsymbol{u})_x = 0, \quad (x, t) \in I \times (0, T], \tag{1.1}$$

with the initial solution $\boldsymbol{u}_0(x)$. Here $\boldsymbol{u}(x, t) : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}^m$ is the unknown vector-valued solution, and $\boldsymbol{f}(\boldsymbol{u}) : \mathbb{R}^m \to \mathbb{R}^m$ is the given smooth flux function. We will, however, point out similarities and differences when the analysis is generalized to multiple space dimensions. Furthermore, we do not pay much attention to boundary conditions in this paper; hence the solution is considered to be periodic or compactly-supported. Generic solutions to (1.1) will contain discontinuities, which are much more difficult to analyze, see [4, 32] for

Keywords and phrases. Discontinuous Galerkin method, Runge–Kutta method, error estimates, symmetrizable system of conservation laws, energy analysis.

* *Research of Luo and Zhang is supported by NSFC Grant 11271187, 11071116 and 10931004.*

** *Research of Shu is supported by NSF Grants DMS-1112700 and DMS-1418750, and DOE Grant DE-FG02-08ER25863.*

[1] Department of Mathematics, Nanjing University, Nanjing, 210093, Jiangsu Province, P.R. China. luojuan.nikki@163.com. qzh@nju.edu.cn.

[2] Division of Applied Mathematics, Brown University, Providence, RI 02912, USA. shu@dam.brown.edu.

some preliminary results in this direction. In this paper we consider only smooth solutions, therefore, we assume the initial solution $\boldsymbol{u}_0(x)$ is smooth and we consider only a short time interval before discontinuities develop.

The symmetrizable system of conservation laws has a wide background. Well-known examples are the shallow water wave equations and the Euler equations of compressible gas dynamics. As is well-known, a conservation law system (1.1) is symmetrizable if and only if it has a convex entropy function [14]. Due to the symmetrization theory [14, 25], one can seek a mapping $\boldsymbol{u}(\boldsymbol{v})\colon \mathbb{R}^m \to \mathbb{R}^m$ applied to (1.1), so that when transformed, $\boldsymbol{u}'_{\boldsymbol{v}}\boldsymbol{v}_t + \boldsymbol{f}'_{\boldsymbol{v}}\boldsymbol{v}_x = 0$, the Jacobian matrix $\boldsymbol{u}'_{\boldsymbol{v}} = (\partial \boldsymbol{u}_i/\partial \boldsymbol{v}_j)_{i=1,\dots,m}^{j=1,\dots,m}$ is symmetric positive definite and the Jacobian matrix $\boldsymbol{f}'_{\boldsymbol{v}} = \boldsymbol{f}'_{\boldsymbol{u}}\boldsymbol{u}'_{\boldsymbol{v}}$ is also symmetric.

In this paper we consider the numerical method to solve directly (1.1) by using the RKDG method. The first version of discontinuous Galerkin (DG) method was introduced in 1973 by Reed and Hill [23] for the steady linear neutron transport. It was later developed into RKDG methods by Cockburn *et al.* [5–7, 10, 12] for nonlinear hyperbolic conservation laws, which use a DG discretization in space and combine it with an explicit total variation diminishing Runge–Kutta (TVDRK) time-marching algorithm [26]. Later, this method was developed to solve equations with higher order derivatives. It is well-known that the DG method has strong stability and optimal accuracy to capture discontinuous jumps and/or sharp transient layers, and it combines the advantages of finite element and finite volume methods. For a fairly complete set of references on this methods, we refer to the review papers [9, 11].

However, up to now there has been relatively few work on stability analysis and error estimates for the fully discrete RKDG methods with the explicit TVDRK time marching to solve (1.1). The method of line version (continuous in time) of the DG scheme for linear equations has been considered in [8, 18, 20], and has been proved to maintain good L^2-norm stability and optimal error estimates. For nonlinear equations, there exists the well-known local entropy inequality [17] for the semi-discrete DG scheme, as well as for the fully discrete DG scheme with some special time-discretizations such as the backward Euler and Crank–Nicolson algorithms. Recently, RKDG method with explicit time-discretization for nonlinear conservation laws has been analyzed in [2, 29, 31], where the (quasi)-optimal *a priori* error estimates are obtained for the second order and the third order explicit TVDRK time discretization, respectively, for scalar equations.

As to symmetrizable systems, error estimates are more difficult to obtain. In [30], error estimates are obtained for RKDG methods with a second order TVDRK time discretization, under a restrictive time step constraint in which $\tau = o(h)$. This restriction is not surprising, as RKDG methods with a second order TVDRK time discretization is linearly unconditionally unstable under the regular time step restriction $\tau \le ch$ for any fixed constant $c > 0$, when the polynomial degree $k \ge 2$ [9]. However, the time step restriction $\tau = o(h)$ is not realistic in actual computation, and we would like to use third order TVDRK time discretization under regular CFL time step restriction. As far as the authors know, there are still no error estimates for the symmetrizable system, when explicit Runge–Kutta time discretization is used under regular CFL conditions. The RKDG3 method uses the third order explicit TVDRK time-marching combined with piecewise polynomials of arbitrary degree, and is more popular because in practice it provides better linear stability and higher order accuracy in time. However, the techniques used in [30] to obtain error estimates for RKDG2 under the more restrictive time step constraints are not applicable to RKDG3 methods under regular CFL conditions. In a recent work [31], a new technique which explores the specific dissipation natures of third order TVDRK time discretization is developed to analyze the scalar equations. The extension of this technique to symmetrizable systems is the purpose of this paper. This extension is non-trivial, as we must carefully study the numerical fluxes (referred to as the generalized E-flux) in the system case and their influence towards the errors. We can no longer use monotonicity of the fluxes in the scalar case. Our result is a quasi-optimal or optimal error estimate depending on the numerical flux, similar to that in [31] for the scalar case, under the standard temporal-spatial CFL condition $\tau \le \gamma h$. Here h and τ are the element length and time step, respectively, and the positive constant γ is independent of h and τ.

The main line of analysis in this paper follows that in [30, 31], by using energy analysis as the main tool. It is rather straightforward to generalize the error estimates in [31] from the scalar case to *symmetric* systems. However, as we will see later in this paper, it is significantly more difficult to carry out the above error estimates

for the symmetrizable systems. In this development, we need to pay more attention to several issues owing to the symmetrizable theory, for instance, the suitable norm with local rotational matrix in each element, and Lipschitz continuity of the rotational matrix and the Jacobian matrix of the flux. In this process, many issues about vector-valued functions need to be addressed.

The essential difficulty in this paper is how to describe the numerical viscosity resulting from the DG spatial discretization. We have considered the RKDG2 method in [30], where the explicit second order TVDRK time-marching is used, for the symmetrizable system, and a description on this issue has been given there. However, the stability mechanism in the RKDG3 method is completely different from that of the RKDG2 method, where the additional stability provided by the explicit third order TVDRK time-marching plays an important role. As we have done in Section 3.3 of [31], we need to use the Lipschitz continuity of the numerical viscosity quantity to maintain the structure of this stability term in the time direction. This property is not needed in the analysis for the RKDG method with the second order time-marching, as in [29,30]. This new property demands that we have to modify the analysis in a suitable way for the symmetrizable system.

To this end, we firstly try to propose a careful classification of the necessary properties for the numerical fluxes used in the symmetrizable system case. This kind of numerical flux is defined as the *generalized E-flux* in this paper, where we relax the demand than that in [30]. For example, we only need two inequalities related to the two states under consideration, and we add the freedom on the rotation position and the adjusting matrix. Thus this new definition is more easily verified than that in [30]. Then we establish an important matrix termed *generalized numerical viscosity matrix*, to describe the total numerical stability coming from the jumps at every element interface. In order to ensure the Lipschitz continuity of this matrix, the definition in this paper is different from that in [30]. We carry out this idea through the generalized Newton difference quotients. Furthermore, we present three typical assumptions on the generalized numerical viscosity matrix, which are enough to obtain good error estimates. The details are given in Section 2.2.

An outline of this paper is as follows. In Section 2 we present the RKDG3 scheme for the symmetrizable system of conservation laws. The so-called *generalized E-flux* and the *generalized numerical viscosity matrix* are presented. Some assumptions on the numerical viscosity matrix and the smoothness of the exact solution are also given here, which yield the main conclusion about the quasi-optimal error estimate. In the remaining part of this paper, we would like to present the detailed proof to this main conclusion. In Section 3 we obtain the error representation and the corresponding error equations, and in Section 4 we provide some elementary discussions on the error functional in detail. In Section 5 we carry out the energy analysis and complete the error estimate, with some technical proofs left to the appendix. Finally, a concluding remark is given in Section 6.

2. RKDG3 scheme and the main conclusion

In this section we present the detailed implementation of the RKDG3 method, following the notations in [30,31]. The generalized numerical viscosity matrix is defined for the so-called generalized E-flux. Finally, we present the main conclusion on the quasi-optimal error estimates in general, and on the optimal error estimate for the upwind numerical flux.

2.1. RKDG3 scheme

Let $\mathcal{J}_h = \{I_j = (x_{j-1/2}, x_{j+1/2})\}_{j=1}^N$ be a partition of $I = (0,1)$, with each element length being $h_j = x_{j+1/2} - x_{j-1/2}$. The maximum length of this mesh is denoted by $h = \max_j h_j$. In this paper we assume the partition is quasi-uniform, namely, there exists a positive constant ν such that $h \le \nu h_j$ for all $j = 1, 2, \ldots, N$, as h goes to zero. The discontinuous finite element space is defined as

$$\mathbb{V}_h = \{\, \boldsymbol{v} \in [L^2(0,1)]^m : \boldsymbol{v}|_{I_j} \in [P^k(I_j)]^m, j = 1, \ldots, N \,\}, \tag{2.1}$$

where $P^k(I_j)$ denotes the space of polynomials in I_j of degree at most k. Note that the function $\boldsymbol{p} \in \mathbb{V}_h$ is allowed to have discontinuities across element interfaces. Two limits from the left- and the right-directions are

denoted by $\boldsymbol{p}^-$ and $\boldsymbol{p}^+$, respectively. Furthermore, the jump and the mean, respectively, are denoted by

$$[\![\boldsymbol{p}]\!] = \boldsymbol{p}^+ - \boldsymbol{p}^-, \quad \text{and} \quad \{\!\{\boldsymbol{p}\}\!\} = \frac{1}{2}(\boldsymbol{p}^+ + \boldsymbol{p}^-). \tag{2.2}$$

We discretize the time interval $[0, T]$ with the time step τ, which could actually change from step to step but is taken as a constant in this paper for simplicity. In the RKDG3 method, we would like to seek successively the numerical solution, denoted by $\boldsymbol{u}_h^n(x) = \boldsymbol{u}_h(x, n\tau)$, in the discontinuous finite element space.

The initial solution $\boldsymbol{u}_h^0$ is taken as the approximation of $\boldsymbol{u}_0(x)$, for instance, the standard L^2-projection $\mathbb{P}_h\boldsymbol{u}_0(x)$. It is defined as the unique function in $\mathbb{V}_h$ such that

$$(\mathbb{P}_h\boldsymbol{u}_0(x) - \boldsymbol{u}_0(x), \boldsymbol{v}_h) = 0 \quad \forall \boldsymbol{v}_h \in \mathbb{V}_h. \tag{2.3}$$

Here $(\boldsymbol{p}, \boldsymbol{q}) = \int_I \boldsymbol{p}^\top \boldsymbol{q}\,\mathrm{d}x$ is the inner product as usual in the space $[L^2(0,1)]^m$. Note that this projection is locally defined on each element and hence we will also refer to it as the local L^2-projection later.

For each $n \geq 0$, the numerical solution of the RKDG3 method from the time $n\tau$ to the next time $(n+1)\tau$ is defined as follows: find $\boldsymbol{u}_h^{n,1}, \boldsymbol{u}_h^{n,2}$ and $\boldsymbol{u}_h^{n+1}$ in the finite element space $\mathbb{V}_h$, such that for any $\boldsymbol{v}_h \equiv \boldsymbol{v}_h(x) \in \mathbb{V}_h$ there hold

$$(\boldsymbol{u}_h^{n,1}, \boldsymbol{v}_h) = (\boldsymbol{u}_h^n, \boldsymbol{v}_h) + \tau\mathcal{H}(\boldsymbol{u}_h^n, \boldsymbol{v}_h), \tag{2.4a}$$

$$(\boldsymbol{u}_h^{n,2}, \boldsymbol{v}_h) = \frac{3}{4}(\boldsymbol{u}_h^n, \boldsymbol{v}_h) + \frac{1}{4}(\boldsymbol{u}_h^{n,1}, \boldsymbol{v}_h) + \frac{\tau}{4}\mathcal{H}(\boldsymbol{u}_h^{n,1}, \boldsymbol{v}_h), \tag{2.4b}$$

$$(\boldsymbol{u}_h^{n+1}, \boldsymbol{v}_h) = \frac{1}{3}(\boldsymbol{u}_h^n, \boldsymbol{v}_h) + \frac{2}{3}(\boldsymbol{u}_h^{n,2}, \boldsymbol{v}_h) + \frac{2\tau}{3}\mathcal{H}(\boldsymbol{u}_h^{n,2}, \boldsymbol{v}_h). \tag{2.4c}$$

Here $\mathcal{H}(\boldsymbol{p}, \boldsymbol{q})$ is the DG spatial discretization, expressed compactly in the form

$$\mathcal{H}(\boldsymbol{p}, \boldsymbol{q}) = \sum_{1\leq j\leq N} \int_{I_j} \boldsymbol{q}_x^\top \boldsymbol{f}(\boldsymbol{p})\,\mathrm{d}x + \sum_{1\leq j\leq N} [\![\boldsymbol{q}]\!]^\top_{j+\frac{1}{2}} \hat{\boldsymbol{f}}(\boldsymbol{p}^-, \boldsymbol{p}^+)_{j+\frac{1}{2}}, \quad \forall \boldsymbol{p}, \boldsymbol{q}, \tag{2.5}$$

since the numerical solution is periodic or compactly-supported.

In (2.5), $\hat{\boldsymbol{f}}(\boldsymbol{p}^-, \boldsymbol{p}^+)$ is the numerical flux to ensure the good stability and high-order accuracy. In general, it depends on the two vectors along the left and right directions at the element boundary point. The well-known examples include the global (local) Lax–Friedrichs flux, the Roe linearization flux function [24] with or without Harten's entropy fix [15], and those fluxes constructed by the flux vector splitting [27]. They can be looked upon as exact or approximate Riemann solvers. For more details, we refer to [27, 30].

To ensure numerical stability, the time step should satisfy a temporal-spatial restriction. The detailed CFL condition will be given later; see (5.1) and (5.25). We have now completed the definition of the RKDG3 method.

2.2. Numerical flux

In this subsection we would like to recall some issues for the numerical fluxes used in practice. Then an important matrix will be defined to describe the numerical viscosity for symmetrizable systems.

2.2.1. Some important matrices in the symmetrizable theory

It follows from the symmetrizable theory that $\boldsymbol{f}'_{\boldsymbol{u}}$ has a strong relationship with an important symmetric matrix, since

$$\boldsymbol{Q}^{1/2}\boldsymbol{f}'_{\boldsymbol{u}}\boldsymbol{Q}^{-1/2} = \boldsymbol{Q}^{1/2}\boldsymbol{f}'_{\boldsymbol{v}}\boldsymbol{Q}^{1/2} \equiv \boldsymbol{K}, \tag{2.6}$$

where $\boldsymbol{Q} \equiv \boldsymbol{Q}(\boldsymbol{u}) = \boldsymbol{v}'_{\boldsymbol{u}}$ is the Jacobian matrix of the transformation $\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{u})$. Note that $\boldsymbol{K}$ is a symmetric matrix with the same spectrum as $\boldsymbol{f}'_{\boldsymbol{u}}$. Thus there exists the eigenvector decomposition $\boldsymbol{K} = \boldsymbol{X}^{-1}\tilde{\boldsymbol{K}}\boldsymbol{X}$, where $\tilde{\boldsymbol{K}} = \mathrm{diag}\{\lambda_i\}_{i=1}^m$. Let $\tilde{\boldsymbol{K}}^+ = \mathrm{diag}\{\max(\lambda_i, 0)\}_{i=1}^m$ and $\tilde{\boldsymbol{K}}^- = \mathrm{diag}\{\min(\lambda_i, 0)\}_{i=1}^m$. The positive and negative

parts of $\boldsymbol{K}$ are then defined as $\boldsymbol{K}^{\pm} = \boldsymbol{X}^{-1}\tilde{\boldsymbol{K}}^{\pm}\boldsymbol{X}$. The standard absolute value matrix of $\boldsymbol{K}$ is defined as $|\boldsymbol{K}| = \boldsymbol{K}^{+} - \boldsymbol{K}^{-}$.

In this paper, we will use another important matrix and its generalized absolute value matrix, respectively,

$$\boldsymbol{H} = \boldsymbol{v}_{\boldsymbol{u}}'\boldsymbol{f}_{\boldsymbol{u}}' = \boldsymbol{Q}^{1/2}\boldsymbol{K}\boldsymbol{Q}^{1/2}, \quad \langle \boldsymbol{H} \rangle = \boldsymbol{Q}^{1/2}|\boldsymbol{K}|\boldsymbol{Q}^{1/2}. \tag{2.7}$$

Note that $\langle \boldsymbol{H} \rangle$ is a symmetric positive semidefinite matrix.

Furthermore, in this paper we will use $\|\cdot\|_{\mathrm{M}}$ to represent the length of a vector, or the spectrum norm of a matrix, respectively. Namely, $\|\boldsymbol{p}\|_{\mathrm{M}} = (\sum_{i=1}^{m} p_i^2)^{1/2}$ for any vector $\boldsymbol{p} = (p_1, \ldots, p_m)^{\top}$, and $\|\boldsymbol{C}\|_{\mathrm{M}} = \max_{\|\boldsymbol{p}\|_{\mathrm{M}}=1} \|\boldsymbol{C}\boldsymbol{p}\|_{\mathrm{M}}$ for any matrix $\boldsymbol{C}$. If the matrix $\boldsymbol{C}$ is symmetric, then $\|\boldsymbol{C}\|_{\mathrm{M}}$ is equal to the spectral radius of this matrix, denoted by $\rho(\boldsymbol{C})$. We will also use the following inequalities

$$|\boldsymbol{p}^{\top}\boldsymbol{C}\boldsymbol{q}| \le \|\boldsymbol{C}\|_{\mathrm{M}}\|\boldsymbol{p}\|_{\mathrm{M}}\|\boldsymbol{q}\|_{\mathrm{M}}, \quad \left|\boldsymbol{p}^{\top}\boldsymbol{C}\boldsymbol{q}\right| \le \left(\boldsymbol{p}^{\top}|\boldsymbol{C}|\boldsymbol{p}\right)^{1/2}\left(\boldsymbol{q}^{\top}|\boldsymbol{C}|\boldsymbol{q}\right)^{1/2}, \tag{2.8}$$

for any matrix $\boldsymbol{C}$ and any vectors $\boldsymbol{p}$ and $\boldsymbol{q}$. Note that $\boldsymbol{C} = |\boldsymbol{C}|$ if $\boldsymbol{C}$ is positive semidefinite. Both inequalities in (2.8) are named Cauchy–Schwarz inequalities in this paper.

2.2.2. Generalized numerical fluxes

Now we would like to recall some elementary properties of numerical fluxes for symmetrizable systems.

In general, the numerical flux $\hat{\boldsymbol{f}}(\boldsymbol{a}, \boldsymbol{b})$ is locally Lipschitz continuous with respect to each argument, and consistent with the true flux $\boldsymbol{f}(\boldsymbol{p})$, namely, $\hat{\boldsymbol{f}}(\boldsymbol{p}, \boldsymbol{p}) = \boldsymbol{f}(\boldsymbol{p})$. Furthermore, successful fluxes used in practice should satisfy certain conditions to ensure stability and convergence of RKDG schemes.

In this paper we would like to present an abstract framework to describe these conditions, suitable for further analysis. This abstract framework is naturally defined for the scalar case (when $m = 1$) in terms of monotonicity, where the numerical flux $\hat{f}(p^-, p^+)$ is assumed to be a nondecreasing function of its first argument and a nonincreasing function of its second argument. Such fluxes are called monotone fluxes. Following [22], a more general class of fluxes are termed as entropy fluxes (E-fluxes), namely, there always holds for any q between p^- and p^+, that

$$[\![p]\!](f(q) - \hat{f}(p^-, p^+)) \ge 0. \tag{2.9}$$

However, this description is not trivial to extend to the system case.

For symmetric systems of conservation laws, an important concept along this line is the so-called generalized E-flux, proposed in [16]. That is to say that

$$[\![\boldsymbol{p}]\!]^{\top}\{\boldsymbol{f}(\boldsymbol{r}) - \hat{\boldsymbol{f}}(\boldsymbol{p}^-, \boldsymbol{p}^+)\} \ge 0, \tag{2.10}$$

for any $\boldsymbol{r}$ standing on the straight-line segment with two endpoints $\boldsymbol{p}^{\pm}$. However, this condition can not describe desirable numerical fluxes when solving symmetrizable (but not symmetric) systems of conservation laws.

Thanks to (2.6), we would like in this paper to propose an extension of the generalized E-fluxes to symmetrizable systems, by the help of the local rotation matrix $\boldsymbol{Q}$.

Definition 2.1. The numerical flux $\hat{\boldsymbol{f}}(\boldsymbol{p}^-, \boldsymbol{p}^+)$ is called a generalized E-flux, if we have a rotation position $\boldsymbol{s}_{\kappa}$ and a positive semidefinite matrix $\boldsymbol{Y}_{\kappa}$, such that

$$[\![\boldsymbol{p}]\!]^{\top}\boldsymbol{Q}(\boldsymbol{s}_{\kappa})\{\boldsymbol{f}(\boldsymbol{r}_{\kappa}) - \hat{\boldsymbol{f}}(\boldsymbol{p}^-, \boldsymbol{p}^+)\} + [\![\boldsymbol{p}]\!]^{\top}\boldsymbol{Y}_{\kappa}[\![\boldsymbol{p}]\!] \ge 0, \quad \kappa = 1, 2, \tag{2.11}$$

for $\boldsymbol{r}_1 = \boldsymbol{p}^-$ and $\boldsymbol{r}_2 = \boldsymbol{p}^+$. Here $\boldsymbol{s}_{\kappa} = \boldsymbol{s}_{\kappa}(\boldsymbol{p}^-, \boldsymbol{p}^+)$ lies in the standard super-rectangle with two vertices $\boldsymbol{p}^{\pm}$. Furthermore, each element in $\boldsymbol{Y}_{\kappa} = \boldsymbol{Y}_{\kappa}(\boldsymbol{p}^-, \boldsymbol{p}^+)$ has the order $\mathcal{O}(\|[\![\boldsymbol{p}]\!]\|_{\mathrm{M}})$, with the common bound depending only on the local Lipschitz constant of $\boldsymbol{Q}$ in the above super-rectangle.

Note that the adjusting matrix $\boldsymbol{Y}_\kappa$ is introduced to overcome the trouble in [30] on how to seek a reasonable rotation position $\boldsymbol{s}_\kappa$. It is easy to seek a good rotation position for the scalar case and for symmetric systems; in such cases we can take $\boldsymbol{Y}_\kappa = 0$. In general, with the help of the new freedom $\boldsymbol{Y}_\kappa$, we can take the rotation position $\boldsymbol{s}_\kappa$ as an arbitrary point along the straight line between $\boldsymbol{p}^-$ and $\boldsymbol{p}^+$. Hence, this definition enlarges the class of generalized E-fluxes, given in [16,30].

Many numerical fluxes can be verified easily to be generalized E-fluxes. A detailed example will be given in the appendix, for the (global/local) Lax–Friedrichs flux. More examples can be found in [21].

2.2.3. Generalized numerical viscosity matrix

It is well-known that the numerical viscosity in the semi-discrete DG method comes from the square of the jumps at element boundary points. Below we would like to define an important matrix for the generalized E-flux, to describe the strength of the numerical stability in space. Similar quantity has been defined and analyzed for the scalar case; we refer to [29].

To extend to the system case, we would like to use the following notation. Let $\boldsymbol{g} = (g_1, g_2, \ldots, g_m)^\top$ be an m-dimensional vector-valued function with respect to the m-dimensional variable. For any given two vectors $\boldsymbol{a} = (a_1, a_2, \ldots, a_m)^\top$ and $\boldsymbol{b} = (b_1, b_2, \ldots, b_m)^\top$, define $\boldsymbol{a}^{(0)} = \boldsymbol{a}$, $\boldsymbol{a}^{(m)} = \boldsymbol{b}$, and

$$\boldsymbol{a}^{(j)} = (b_1, \ldots, b_{j-1}, b_j, a_{j+1}, \ldots, a_m)^\top, \quad j = 1, 2, \ldots, m-1. \tag{2.12}$$

Note that $\boldsymbol{g}(\boldsymbol{a}) - \boldsymbol{g}(\boldsymbol{b}) = \sum_{j=1}^m [\boldsymbol{g}(\boldsymbol{a}^{(j-1)}) - \boldsymbol{g}(\boldsymbol{a}^{(j)})]$. The generalized Newton difference quotient is an m-dimensional matrix, denoted by $\mathcal{D}\boldsymbol{g}[\boldsymbol{a}, \boldsymbol{b}]$, with the element in the (i, j)-entry being

$$\Big(\mathcal{D}\boldsymbol{g}[\boldsymbol{a}, \boldsymbol{b}]\Big)_{ij} = \frac{\boldsymbol{g}_i(\boldsymbol{a}^{(j-1)}) - \boldsymbol{g}_i(\boldsymbol{a}^{(j)})}{a_j - b_j}. \tag{2.13}$$

If the denominator is equal to zero, the term should be understood as the limit when the denominator goes to zero.

Definition 2.2. Let $\hat{\boldsymbol{f}}(\boldsymbol{p}^-, \boldsymbol{p}^+)$ be a generalized E-flux, locally Lipschitz continuous and consistent with $\boldsymbol{f}(\boldsymbol{u})$. The generalized numerical viscosity matrix $\mathcal{A}(\hat{\boldsymbol{f}}; \boldsymbol{p})$ is defined at each element boundary point, in the form

$$\begin{aligned} \mathcal{A}(\hat{\boldsymbol{f}}; \boldsymbol{p}) &= \frac{1}{2}\mathcal{A}_1(\hat{\boldsymbol{f}}; \boldsymbol{p}) + \frac{1}{2}\mathcal{A}_2(\hat{\boldsymbol{f}}; \boldsymbol{p}) \\ &\equiv \frac{1}{2}\boldsymbol{Q}(\boldsymbol{s}_1)\mathcal{D}\hat{\boldsymbol{f}}^{(1)}[\boldsymbol{p}^-, \boldsymbol{p}^+] - \frac{1}{2}\boldsymbol{Q}(\boldsymbol{s}_2)\mathcal{D}\hat{\boldsymbol{f}}^{(2)}[\boldsymbol{p}^-, \boldsymbol{p}^+], \end{aligned} \tag{2.14}$$

where $\hat{\boldsymbol{f}}^{(1)}(\boldsymbol{r}) = \hat{\boldsymbol{f}}(\boldsymbol{r}, \boldsymbol{p}^+)$ and $\hat{\boldsymbol{f}}^{(2)}(\boldsymbol{r}) = \hat{\boldsymbol{f}}(\boldsymbol{p}^-, \boldsymbol{r})$. Here $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ are the two rotation positions, stated in Definition 2.1.

We would like to mention that the above definition confirms the result for the linear constant system, with $\boldsymbol{f}(\boldsymbol{u}) = \boldsymbol{A}\boldsymbol{u}$ and $\boldsymbol{A}$ being a constant matrix. In this case, all upwind numerical fluxes used in practice are equal to $\hat{\boldsymbol{f}}(\boldsymbol{p}^-, \boldsymbol{p}^+) = \boldsymbol{A}^+\boldsymbol{p}^- + \boldsymbol{A}^-\boldsymbol{p}^+$. Due to the symmetrizable theory, there exists a symmetric positive definite matrix $\boldsymbol{A}_0$ such that $\boldsymbol{A}\boldsymbol{A}_0$ is symmetric. A tedious stability analysis [21] for the semi-discrete DG method will give the total numerical viscosity,

$$\frac{1}{2}\sum_{1\le j\le N}\Big[[\![\boldsymbol{u}_h]\!]^\top\langle\boldsymbol{A}_0^{-1}\boldsymbol{A}\rangle[\![\boldsymbol{u}_h]\!]\Big]_{j+\frac{1}{2}}. \tag{2.15}$$

The generalized numerical viscosity matrix given by (2.14) is the same as the above involved matrix, namely, $\frac{1}{2}\langle\boldsymbol{A}_0^{-1}\boldsymbol{A}\rangle$. From this viewpoint, Definition 2.2 is nicer than that in [30].

Through observation on many numerical fluxes, we would like in this paper to make some elementary assumptions on the generalized numerical viscosity matrix, for the convenience of analysis. They read

A1 The matrix $\mathcal{A}(\hat{\boldsymbol{f}};\boldsymbol{p})$ is nearly positive definite, in the sense

$$[\![\boldsymbol{p}]\!]^\top \mathcal{A}(\hat{\boldsymbol{f}};\boldsymbol{p})[\![\boldsymbol{p}]\!] \geq -L_{\rm np}\|[\![\boldsymbol{p}]\!]\|_{\rm M}^3, \quad \forall \boldsymbol{p}. \tag{2.16a}$$

A2 The matrix $\mathcal{A}(\hat{\boldsymbol{f}};\boldsymbol{p})$ is Lipschitz continuous, in the sense,

$$\|\mathcal{A}(\hat{\boldsymbol{f}};\boldsymbol{p}) - \mathcal{A}(\hat{\boldsymbol{f}};\{\!\{\boldsymbol{p}\}\!\})\|_{\rm M} \leq L_{\rm Lip}\|[\![\boldsymbol{p}]\!]\|_{\rm M}, \quad \forall \boldsymbol{p}. \tag{2.16b}$$

A3 The matrix $\mathcal{A}(\hat{\boldsymbol{f}};\boldsymbol{p})$ is upper consistent with $\frac{1}{2}\langle \boldsymbol{H}(\{\!\{\boldsymbol{p}\}\!\})\rangle$, in the sense: there exists a positive semidefinite matrix $\mho = \mho(\boldsymbol{f},\hat{\boldsymbol{f}};\boldsymbol{p})$, such that

$$\left|[\![\boldsymbol{p}]\!]^\top \left(\frac{1}{2}\langle \boldsymbol{H}(\{\!\{\boldsymbol{p}\}\!\})\rangle + \mho - \mathcal{A}(\hat{\boldsymbol{f}};\boldsymbol{p})\right)[\![\boldsymbol{p}]\!]\right| \leq L_{\rm uc}\|[\![\boldsymbol{p}]\!]\|_{\rm M}^3, \quad \forall \boldsymbol{p}. \tag{2.16c}$$

Note that the above three bounding constants are all independent of $\|[\![\boldsymbol{p}]\!]\|_{\rm M}$, but may depend on the local Lipschitz constants of the true flux $\boldsymbol{f}(\boldsymbol{u})$ and/or the numerical flux $\hat{\boldsymbol{f}}(\boldsymbol{p}^-,\boldsymbol{p}^+)$.

The assumption A1 is provided directly from Definition 2.1. Actually, it follows from (2.11) and (2.14) that

$$\begin{aligned}[\![\boldsymbol{p}]\!]^\top \mathcal{A}(\hat{\boldsymbol{f}};\boldsymbol{p})[\![\boldsymbol{p}]\!] &= \frac{1}{2}[\![\boldsymbol{p}]\!]^\top \boldsymbol{Q}(\boldsymbol{s}_1)\{\boldsymbol{f}(\boldsymbol{p}^+) - \hat{\boldsymbol{f}}(\boldsymbol{p}^-,\boldsymbol{p}^+)\} + \frac{1}{2}[\![\boldsymbol{p}]\!]^\top \boldsymbol{Q}(\boldsymbol{s}_2)\{\boldsymbol{f}(\boldsymbol{p}^-) - \hat{\boldsymbol{f}}(\boldsymbol{p}^-,\boldsymbol{p}^+)\} \\ &\geq -[\![\boldsymbol{p}]\!]^\top \boldsymbol{Y}_1[\![\boldsymbol{p}]\!] - [\![\boldsymbol{p}]\!]^\top \boldsymbol{Y}_2[\![\boldsymbol{p}]\!],\end{aligned}$$

where $\|\boldsymbol{Y}_\kappa\|_{\rm M} \leq L_\kappa\|[\![\boldsymbol{p}]\!]\|_{\rm M}$ for $\kappa = 1,2$. The remaining two assumptions can be verified easily for many numerical fluxes. An example is given in the appendix, for the local/global Lax–Friedrich flux (A.1).

It is worthy to point out that the generalized numerical viscosity matrix satisfies the kernel property

$$\mathcal{A}_\kappa(\hat{\boldsymbol{f}};\boldsymbol{p})[\![\boldsymbol{p}]\!] = \boldsymbol{Q}(\boldsymbol{s}_\kappa)\Big[\boldsymbol{f}(\boldsymbol{r}_\kappa) - \hat{\boldsymbol{f}}(\boldsymbol{p}^-,\boldsymbol{p}^+)\Big], \tag{2.17}$$

for $\kappa = 1,2$, where $\boldsymbol{r}_1 = \boldsymbol{p}^-$ and $\boldsymbol{r}_2 = \boldsymbol{p}^+$. This property will play an important role in our analysis, as in [30], in order to express explicitly the numerical viscosity provided by the numerical flux. Detailed discussion can be found in section 4.4 and in section A.4 on the proof of Lemma 5.4.

Remark 2.3. In the above discussion, we do not pay attention to whether the generalized numerical viscosity matrix $\mathcal{A}(\hat{\boldsymbol{f}};\boldsymbol{p})$ is symmetric or not.

Remark 2.4. Assumption A3 can be verified easily for the upwind numerical fluxes (namely, one-sided numerical fluxes for each characteristic variable after suitable local characteristic decomposition, see Sect. 5.6 for more details), for example, the upwind numerical flux for a linear flux and the Steger–Warming flux [27] for Euler equations. In such cases, we can take $\mho = 0$, and consequently there holds $\mathcal{A}(\hat{\boldsymbol{f}};\boldsymbol{p})$ tends to $\frac{1}{2}\langle \boldsymbol{H}(\{\!\{\boldsymbol{p}\}\!\})\rangle$ as $[\![\boldsymbol{p}]\!]$ goes to zero.

2.3. The main conclusion

In this paper, some standard norms will be used. Let $\boldsymbol{p}$ be a vector-valued function and/or matrix-valued function. The $[L^2(I)]^m$ norm and the infinity norm, respectively, are denoted by

$$\|\boldsymbol{p}\| = \left(\int_I \|\boldsymbol{p}(x)\|_{\rm M}^2\,{\rm d}x\right)^{1/2}, \quad \|\boldsymbol{p}\|_\infty = \operatorname{ess}\sup_{x\in(0,1)} \|\boldsymbol{p}(x)\|_{\rm M}. \tag{2.18}$$

We also use $\|\cdot\|_s$ to denote the standard norm in the Sobolev space $[H^s(I)]^m$, in which the function and its derivatives up to sth order are all in $[L^2(I)]^m$.

In order to obtain the *a priori* error estimates to the RKDG3 method, we would like in this paper to assume that the exact solution $\boldsymbol{u}(x,t)$ to problem (1.1) is sufficiently smooth. Namely,

H1 Both $\|\boldsymbol{u}\|_{k+1}(t)$ and $\|\boldsymbol{u}_{tttt}\|(t)$ are bounded uniformly by a constant for any time $t \in [0,T]$. Furthermore, the exact solution is continuous and bounded. For the purpose of optimal error estimate, we assume that $\boldsymbol{u}$ has a higher order smoothness, *i.e.*, $\|\boldsymbol{u}\|_{k+2}(t)$ is also bounded uniformly for any time $t \in [0,T]$.

H2 Each component of $\boldsymbol{f}(\boldsymbol{p})$ and $\boldsymbol{f}'_{\boldsymbol{u}}(\boldsymbol{p})$ is bounded for all $\boldsymbol{p} \in \mathbb{R}^m$. Moreover, $\boldsymbol{f}'_{\boldsymbol{u}}(\boldsymbol{p})$ is Lipschitz continuous with the bounding constant $C_\star^{\mathrm{Flux}}$.

H3 The rotation matrix $\boldsymbol{Q}(\boldsymbol{p})$ is symmetric positive definite uniformly, namely, there exist two constants $\gamma_\star$ and $\gamma^\star$, such that $0 < \gamma_\star \le \|\boldsymbol{Q}(\boldsymbol{p})\|_{\mathrm{M}} \le \gamma^\star$ holds for all $\boldsymbol{p} \in \mathbb{R}^m$. Furthermore, $\boldsymbol{Q}(\boldsymbol{p})$ is Lipschitz continuous with the bounding constant $C_\star^{\mathrm{Rot}}$.

As a consequence, the bounding constants in the assumptions A1–A3 can be simplified by three common constants, respectively.

A remark is given here. Due to the boundedness of the exact solution, the last two assumptions H2 and H3 are reasonable with a cut-off modification on $\boldsymbol{f}(\boldsymbol{u})$ and $\boldsymbol{u}(\boldsymbol{v})$. We refer to [29] for more details.

Now we present the main conclusion in the following theorem.

Theorem 2.5. *Let $\boldsymbol{u}_h$ be the numerical solution of the RKDG3 scheme (2.4), using piecewise polynomials with arbitrary degree $k \ge 2$, defined on any quasi-uniform triangulations of $I = (0,1)$. The numerical flux is a generalized E-flux associated with the generalized numerical viscosity matrix satisfying assumptions A1–A3. Let $\boldsymbol{u}$ be the exact solution of problem (1.1), which satisfies the above smoothness assumptions H1–H3. Then we have the following error estimate*

$$\max_{n\tau \le T} \|\boldsymbol{u}(x,t^n) - \boldsymbol{u}_h^n\| \le C(h^{k+\sigma} + \tau^3), \tag{2.19}$$

under a standard CFL condition $\tau \le \gamma h$ with a suitably fixed constant $\gamma > 0$, where the bounding constant $C > 0$ is independent of h and τ. Here $\sigma = \frac{1}{2}$ for generalized E-fluxes; and $\sigma = 1$ for upwind numerical fluxes.

The proof is technical and long, which will be given in several steps in the subsequent sections. To show the main ideas clearly, we focus our attention on the quasi-optimal error estimate ($\sigma = \frac{1}{2}$) for generalized E-fluxes. The optimal error estimate for upwind numerical fluxes can be proved in a similar line; a sketch will be given in Section 5.6.

3. The error equations

To obtain the error estimate, we have to first establish the error equations. This process is similar to that in [30,31].

3.1. Reference stage solutions

Following [31], three reference functions are defined as the local time discretization of the third order explicit TVDRK algorithm for the exact solution of the conservation law (1.1). Let $\boldsymbol{u}^{(0)} = \boldsymbol{u}$, and

$$\boldsymbol{u}^{(1)} = \boldsymbol{u}^{(0)} - \tau \boldsymbol{f}(\boldsymbol{u}^{(0)})_x, \quad \boldsymbol{u}^{(2)} = \frac{3}{4}\boldsymbol{u}^{(0)} + \frac{1}{4}\boldsymbol{u}^{(1)} - \frac{1}{4}\tau \boldsymbol{f}(\boldsymbol{u}^{(1)})_x. \tag{3.1}$$

Owing to the Sobolev embedding theory [1], it follows from the smoothness assumptions H1–H3 that the above reference values are all continuous and bounded in the whole domain $I \times [0,T]$.

A simple manipulation yields the local truncation error in time, denoted by $\mathcal{E}(x,t)$, such that

$$\boldsymbol{u}^{(0)}(x,t+\tau) = \frac{1}{3}\boldsymbol{u}^{(0)}(x,t) + \frac{2}{3}\boldsymbol{u}^{(2)}(x,t) - \frac{2}{3}\tau[\boldsymbol{f}(\boldsymbol{u}^{(2)}(x,t))]_x + \mathcal{E}(x,t). \tag{3.2}$$

There holds uniformly $\|\mathcal{E}(x,t)\| = \mathcal{O}(\tau^4)$ for any time, due to assumption H1. Similar discussion can be found in [31] for the scalar case.

Define $\boldsymbol{u}^{n,\ell} = \boldsymbol{u}^{(\ell)}(x,t^n)$ to be the reference stage solution. Here and below we may drop the index ℓ if $\ell = 0$. For convenience of notations, we denote three differences

$$\mathbb{E}_1^n \boldsymbol{u}^n = \boldsymbol{u}^{n,1} - \boldsymbol{u}^n,\ \mathbb{E}_2^n \boldsymbol{u}^n = 4\boldsymbol{u}^{n,2} - \boldsymbol{u}^{n,1} - 3\boldsymbol{u}^n,\ \mathbb{E}_3^n \boldsymbol{u}^n = \frac{1}{2}(3\boldsymbol{u}^{n+1} - 2\boldsymbol{u}^{n,2} - \boldsymbol{u}^n), \tag{3.3}$$

to describe the evolution of the solution at each time stage.

Multiply the test function $\boldsymbol{v}_h \in \mathbb{V}_h$ on both sides of equations (3.1) and (3.2), respectively. Let $t = t^n$ and then integrate them in each element. Due to the consistency of the numerical flux, this process yields a set of equalities for $\ell = 0, 1, 2$,

$$(\mathbb{E}_{\ell+1}^n \boldsymbol{u}^n, \boldsymbol{v}_h) = \tau \mathcal{H}^{n,\ell}(\boldsymbol{u}^{n,\ell}, \boldsymbol{v}_h) + (\mathcal{E}^{n,\ell}, \boldsymbol{v}_h), \quad \forall \boldsymbol{v}_h \in \mathbb{V}_h, \tag{3.4}$$

which are almost the same as the RKDG3 scheme (2.4). In this paper we would like to denote $\mathcal{E}^{n,2} = \mathcal{E}(x,t^n)$, and denote $\mathcal{E}^n = \mathcal{E}^{n,1} = 0$ just for simplicity.

3.2. Projection properties

Define the stage error $\boldsymbol{e}^{n,\ell} = \boldsymbol{u}^{n,\ell} - \boldsymbol{u}_h^{n,\ell}$. As the usual treatment in a finite element analysis, we divide the stage error in the form

$$\boldsymbol{e}^{n,\ell} = \left[\pi_h \boldsymbol{u}^{n,\ell} - \boldsymbol{u}_h^{n,\ell}\right] - \left[\pi_h \boldsymbol{u}^{n,\ell} - \boldsymbol{u}^{n,\ell}\right] \equiv \boldsymbol{\xi}^{n,\ell} - \boldsymbol{\eta}^{n,\ell}, \tag{3.5}$$

where π_h is a suitable projection. The projection error $\boldsymbol{\eta}^{n,\ell}$ can be estimated easily.

It is enough to take $\pi_h = \mathbb{P}_h$ as the standard L^2-projection (refer to (2.3)), in order to obtain the quasi-optimal error estimate. Since $\boldsymbol{u}(x,t)$ is smooth enough, a standard scaling argument [3, 20] yields that the projection error $\boldsymbol{\eta}^{n,\ell}$ satisfies

$$\|\boldsymbol{\eta}^{n,\ell}\| + h\|\boldsymbol{\eta}_x^{n,\ell}\| + h^{1/2}\|\boldsymbol{\eta}^{n,\ell}\|_{\Gamma_h} \le C_1 h^{k+1}, \quad \forall\, n\colon n\tau \le T, \ell = 0,1,2. \tag{3.6a}$$

Here Γ_h is the union of all element interfaces, and, for any function $\boldsymbol{p}$, we denote

$$\|\boldsymbol{p}\|_{\Gamma_h} = \left(\sum_{1\le j\le N} \frac{1}{2}\|\boldsymbol{p}_{j+\frac{1}{2}}^+\|_{\mathrm{M}}^2 + \frac{1}{2}\|\boldsymbol{p}_{j+\frac{1}{2}}^-\|_{\mathrm{M}}^2 \right)^{1/2}.$$

It follows from the interpolation theory [3] that

$$\|\boldsymbol{\eta}^{n,\ell}\|_\infty \le C_2 h^{k+\frac{1}{2}}, \quad \forall\, n\colon n\tau \le T, \ell = 0,1,2. \tag{3.6b}$$

Since the projection $\mathbb{P}_h$ is linear and independent of the time, $\mathbb{E}_\ell^n \boldsymbol{\eta}^n$ is also the projection error of $\mathbb{E}_\ell^n \boldsymbol{u}^n$ under the same projection. Thus we also have the estimates

$$\|\mathbb{E}_\ell^n \boldsymbol{\eta}^n\| + h^{1/2}\|\mathbb{E}_\ell^n \boldsymbol{\eta}^n\|_{\Gamma_h} \le C_3 h^{k+1}\tau, \quad \forall\, n\colon n\tau \le T, \ell = 1,2,3. \tag{3.6c}$$

Note that the above bounding constants, C_1, C_2 and C_3, depend solely on the smoothness of the exact solution.

3.3. The error equations

The remaining work in this paper is to estimate the errors in the finite element space, namely $\boldsymbol{\xi}^{n,\ell} = \pi_h \boldsymbol{e}^{n,\ell} \in \mathbb{V}_h$. To this end, we need to set up the error equations as follows. Subtracting (3.4) from (2.4) gives the error equations for $\ell = 0, 1, 2$,

$$\begin{aligned} &(\mathbb{E}_{\ell+1}^n \boldsymbol{\xi}^n, \boldsymbol{v}_h) = \tau \mathcal{K}^{n,\ell}(\boldsymbol{v}_h) \\ \equiv\ &(\mathbb{E}_{\ell+1}^n \boldsymbol{\eta}^n + \mathcal{E}^{n,\ell}, \boldsymbol{v}_h) + \tau \mathcal{H}(\boldsymbol{u}^{n,\ell}, \boldsymbol{v}_h) - \tau \mathcal{H}(\boldsymbol{u}_h^{n,\ell}, \boldsymbol{v}_h), \quad \forall \boldsymbol{v}_h \in \mathbb{V}_h. \end{aligned} \tag{3.7}$$

Here $\mathcal{K}^{n,\ell}(v_h)$ is called the error functional at each time stage. To state it more clearly, we would like to separate it into four parts with different meanings, through a simple re-arranging process and using consistency of the numerical flux. Namely,

$$\mathcal{K}^{n,\ell}(v_h) \equiv \mathcal{L}^{n,\ell}(\boldsymbol{e}^{n,\ell}, \boldsymbol{v}_h) + \mathcal{N}^{n,\ell}(\boldsymbol{v}_h) + \mathcal{V}^{n,\ell}(\boldsymbol{v}_h) + \mathcal{T}^{n,\ell}(\boldsymbol{v}_h). \tag{3.8}$$

The terms on the right-hand side are named the linear part, the nonlinear part, the viscosity part, and the time-marching part, respectively, which are defined below.

The linear part is defined as a bilinear functional with respect to $\boldsymbol{w}$ and $\boldsymbol{v}_h$, with the given Jacobian matrix $\boldsymbol{f}'_{\boldsymbol{u}}(\boldsymbol{u}^{n,\ell})$. It reads in the form

$$\begin{aligned}\mathcal{L}^{n,\ell}(\boldsymbol{w}, \boldsymbol{v}_h) &= \mathcal{L}_{\text{main}}(\boldsymbol{f}'_{\boldsymbol{u}}(\boldsymbol{u}^{n,\ell}); \boldsymbol{w}, \boldsymbol{v}_h) \\ &= \sum_{1\le j\le N} \left[\llbracket \boldsymbol{v}_h \rrbracket^\top \boldsymbol{f}'_{\boldsymbol{u}}(\boldsymbol{u}^{n,\ell}) \{\!\{\boldsymbol{w}\}\!\}\right]_{j+\frac{1}{2}} + \int_I (\boldsymbol{v}_h)_x^\top \boldsymbol{f}'_{\boldsymbol{u}}(\boldsymbol{u}^{n,\ell}) \boldsymbol{w}\, \mathrm{d}x. \end{aligned} \tag{3.9a}$$

Here and below the integration on I means the sum of those integrations on every elements I_j. The remaining three parts are linear functionals with respect to $\boldsymbol{v}_h$. The nonlinear part is given in the form

$$\begin{aligned}\mathcal{N}^{n,\ell}(\boldsymbol{v}_h) = &\sum_{1\le j\le N} \llbracket \boldsymbol{v}_h \rrbracket^\top_{j+\frac{1}{2}} \left[\boldsymbol{f}(\boldsymbol{u}^{n,\ell}) - \{\!\{\boldsymbol{f}(\boldsymbol{u}_h^{n,\ell})\}\!\} - \boldsymbol{f}'_{\boldsymbol{u}}(\boldsymbol{u}^{n,\ell}) \{\!\{\boldsymbol{e}^{n,\ell}\}\!\}\right]_{j+\frac{1}{2}} \\ &+ \int_I (\boldsymbol{v}_h)_x^\top \left[\boldsymbol{f}(\boldsymbol{u}^{n,\ell}) - \boldsymbol{f}(\boldsymbol{u}_h^{n,\ell}) - \boldsymbol{f}'_{\boldsymbol{u}}(\boldsymbol{u}^{n,\ell}) \boldsymbol{e}^{n,\ell}\right] \mathrm{d}x. \end{aligned} \tag{3.9b}$$

If $\boldsymbol{f}(\boldsymbol{u})$ is a linear flux, the nonlinear part disappears. Further, the viscosity part solely depends on the numerical solutions, defined as

$$\mathcal{V}^{n,\ell}(\boldsymbol{v}_h) = \sum_{1\le j\le N} \llbracket \boldsymbol{v}_h \rrbracket^\top_{j+\frac{1}{2}} \left(\{\!\{\boldsymbol{f}(\boldsymbol{u}_h^{n,\ell})\}\!\} - \hat{\boldsymbol{f}}(\boldsymbol{u}_h^{n,\ell})\right)_{j+\frac{1}{2}}. \tag{3.9c}$$

The time-marching part is defined as

$$\mathcal{T}^{n,\ell}(\boldsymbol{v}_h) = \tau^{-1}(\boldsymbol{\zeta}^{n,\ell}, \boldsymbol{v}_h), \tag{3.9d}$$

where $\boldsymbol{\zeta}^{n,\ell} = \mathbb{E}^n_{\ell+1}\boldsymbol{\eta}^n + \mathcal{E}^{n,\ell}$, representing the evolution of the projection error and the local truncation error in time.

Remark 3.1. In the splitting of the error (3.8) above, we have introduced the reference vector at each element boundary point, which depends on the specific projection being used. For the local L^2-projection $\mathbb{P}_h$, the reference vector is taken as the simple average $\{\!\{\boldsymbol{u}_h^{n,\ell}\}\!\}$. This corresponds to three terms, namely, $\{\!\{\boldsymbol{w}\}\!\}$ in (3.9a), and $\{\!\{\boldsymbol{f}(\boldsymbol{u}_h^{n,\ell})\}\!\}$ in both (3.9b) and (3.9c).

To obtain the optimal error estimate for the upwind numerical flux, we will use the local Gauss–Radau projection. In this case, we would need to introduce a different reference vector at each element boundary point, and make corresponding modifications; see Section 5.6 for more details.

4. Elementary estimates

In this section we would like to set up some basic discussions on the error functional. For notational convenience, in this section we will drop the super-index, n and ℓ, for the reference stage solutions, the numerical solutions and the operators.

4.1. Preliminaries

4.1.1. Notations

In this paper we will use notations C, K, ε to denote generic positive constants independent of h, τ and n. Here ε is a small positive constant, and K depends solely on the inverse constants, to be specified in Section 4.1.3 below. To emphasize the nonlinearity of the flux $\boldsymbol{f}(\boldsymbol{u})$ and the transformation $\boldsymbol{u}(\boldsymbol{v})$, we would like to use $C_\star$ to denote a nonnegative constant which vanishes, *i.e.* $C_\star = 0$ for a linear flux $\boldsymbol{f}(\boldsymbol{u}) = \boldsymbol{A}\boldsymbol{u}$. Furthermore, for any function $\boldsymbol{p}$, we would like to use a macro notation

$$C(\boldsymbol{p}) = C + C_\star h^{-1}\|\boldsymbol{p}\|_\infty + C_\star h^{-2}\|\boldsymbol{p}\|_\infty^2. \tag{4.1}$$

If the bounding constant depends on ε, we will denote it by $C(\varepsilon; \boldsymbol{p})$. Note that the above notations may have a different value in each occurrence.

In this paper we use the notations $\boldsymbol{u}_c$ and $\boldsymbol{u}_b$ to denote the evaluation of the reference solutions at the element center points and the element boundary points, respectively. Recalling $\boldsymbol{Q} = \boldsymbol{v}'_{\boldsymbol{u}}$, we define the piecewise constant rotation matrix as

$$\boldsymbol{Q}_c = \boldsymbol{Q}(\boldsymbol{u}_c), \tag{4.2}$$

which is equal to the evaluation of the matrix at the center point in each element. Also we denote $\boldsymbol{Q}_b = \boldsymbol{Q}(\boldsymbol{u}_b)$ at every element boundary point.

4.1.2. Local focus shifting

Let $\boldsymbol{C}$ be one of the matrix-valued functions $\boldsymbol{f}'_{\boldsymbol{u}}$, $\boldsymbol{Q}$, $\boldsymbol{Q}^{1/2}$, $\boldsymbol{H}$ and $\langle \boldsymbol{H} \rangle$. Their focus shifting (*i.e.* change of the vector at which this matrix-valued function is evaluated) causes many complexity in the following analysis. Due to the Lipschitz continuity and the famous Wielandt–Hoffman Theorem [13], we have

$$\|\boldsymbol{C}(\boldsymbol{a}) - \boldsymbol{C}(\boldsymbol{b})\|_{\mathrm{M}} \le C_\star \|\boldsymbol{a} - \boldsymbol{b}\|_{\mathrm{M}}, \tag{4.3}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are two considered focuses. Furthermore, we will use the inequalities $\|\boldsymbol{f}(\boldsymbol{a}) - \boldsymbol{f}(\boldsymbol{b})\|_{\mathrm{M}} \le C\|\boldsymbol{a} - \boldsymbol{b}\|_{\mathrm{M}}$ and

$$\|\boldsymbol{f}(\boldsymbol{a}) - \boldsymbol{f}(\boldsymbol{b}) - \boldsymbol{f}'_{\boldsymbol{u}}(\boldsymbol{b})(\boldsymbol{a} - \boldsymbol{b})\|_{\mathrm{M}} \le \frac{1}{2} C_\star^{\mathrm{Flux}} \|\boldsymbol{a} - \boldsymbol{b}\|_{\mathrm{M}}^2, \tag{4.4}$$

in the local linearization process. We refer to [19] for more details.

The focuses considered in this paper will be taken from the reference solutions $\boldsymbol{u}_c$, $\boldsymbol{u}_b$ and $\boldsymbol{u}(x)$, or the numerical solutions $\{\!\{\boldsymbol{u}_h\}\!\}$ and $\boldsymbol{s}_i = \boldsymbol{s}_i(\boldsymbol{u}_h)$ in the definition of generalized E-flux, in the same cell or in adjacent cells, at the same time stage level, or at different time stage levels in the same time-marching step. In all these cases, we have

$$\|\boldsymbol{a} - \boldsymbol{b}\|_{\mathrm{M}} \le C(h + \tau) + \|[\![\boldsymbol{e}]\!]\|_{\mathrm{M}} \le C(h + \tau) + \|\boldsymbol{e}\|_\infty, \tag{4.5}$$

since the reference solution $\boldsymbol{u}$ is Lipschitz continuous and thus

$$[\![\boldsymbol{u}_h]\!] = [\![\boldsymbol{\eta}]\!] - [\![\boldsymbol{\xi}]\!] = -[\![\boldsymbol{e}]\!], \tag{4.6}$$

at each element boundary point. Note that (4.5) simply gives a very crude bound. In many cases, we do not have all the terms on the right-hand side of this inequality.

4.1.3. The inverse properties

Some inverse properties will be used in this paper. For any function $\boldsymbol{v}_h \in \mathbb{V}_h$, there exists a positive constant μ, independent of $\boldsymbol{v}_h$ and h, such that

$$\text{(i)}\ \|(\boldsymbol{v}_h)_x\| \le \mu h^{-1}\|\boldsymbol{v}_h\|; \quad \text{(ii)}\ \|\boldsymbol{v}_h\|_{\Gamma_h} \le \mu^{1/2} h^{-1/2}\|\boldsymbol{v}_h\|; \quad \text{(iii)}\ \|\boldsymbol{v}_h\|_\infty \le \mu h^{-1/2}\|\boldsymbol{v}_h\|.$$

For more details of these inverse properties, we refer the reader to [3].

To cope with the local moving of focus near the element boundary, we will use the following inequalities

$$\|\boldsymbol{Q}_b^{1/2}\boldsymbol{v}_h\|_{\Gamma_h} \le K\mu^{1/2}\Big[h^{-1/2}\|\boldsymbol{Q}_c^{1/2}\boldsymbol{v}_h\| + C_\star h^{1/2}\|\boldsymbol{v}_h\|\Big], \tag{4.7a}$$

$$\|\boldsymbol{Q}^{1/2}(\{\!\{\boldsymbol{u}_h\}\!\})\boldsymbol{v}_h\|_{\Gamma_h}^2 \le K\|\boldsymbol{Q}_b^{1/2}\boldsymbol{v}_h\|_{\Gamma_h}^2 + C_\star h^{-1}\|\boldsymbol{e}\|_\infty\|\boldsymbol{v}_h\|^2, \tag{4.7b}$$

for any $\boldsymbol{v}_h \in \mathbb{V}_h$. These can be obtained from the application of focus shifting and the elementary inverse properties.

4.2. The linear part

Starting from this subsection we will discuss separately the four parts in the error functional. Let $S_{\max} = \max \varrho(\boldsymbol{f}'_{\boldsymbol{u}}(\boldsymbol{u}))$, representing the maximum of the flow speed; here $\varrho(\boldsymbol{f}'_{\boldsymbol{u}})$ is the spectral radius, and the maximum is taken over all vectors in $\mathbb{R}^m$.

Lemma 4.1. *The linear part is a bounded bilinear functional in $\mathbb{V}_h \times \mathbb{V}_h$. Namely, there exist bounding constants K and $C_\star$ independent of h and τ, such that*

$$|\mathcal{L}(\boldsymbol{w}_h, \boldsymbol{Q}_c\boldsymbol{v}_h)| \le KS_{\max}\mu h^{-1}\|\boldsymbol{Q}_c^{1/2}\boldsymbol{w}_h\|\|\boldsymbol{Q}_c^{1/2}\boldsymbol{v}_h\| + C_\star\|\boldsymbol{w}_h\|\|\boldsymbol{v}_h\| \tag{4.8}$$

holds for any $\boldsymbol{w}_h, \boldsymbol{v}_h \in \mathbb{V}_h$.

Proof. For any point x in the interior of each element, we move the focus of the rotation matrix $\boldsymbol{Q}$ from $\boldsymbol{u}_c$ to $\boldsymbol{u}(x)$. Then from (3.9a) we get a splitting $\mathcal{L}(\boldsymbol{w}_h, \boldsymbol{Q}_c\boldsymbol{v}_h) = \mathcal{L}^{(1)}(\boldsymbol{w}_h, \boldsymbol{v}_h) + \mathcal{L}^{(2)}(\boldsymbol{w}_h, \boldsymbol{v}_h)$, where

$$\mathcal{L}^{(1)}(\boldsymbol{w}_h, \boldsymbol{v}_h) = \int_I (\boldsymbol{v}_h)_x^\top \boldsymbol{H}(\boldsymbol{u})\boldsymbol{w}_h \,\mathrm{d}x + \sum_{1\le j\le N}\Big[\![\boldsymbol{v}_h]\!]^\top \boldsymbol{H}(\boldsymbol{u}_b)\{\!\{\boldsymbol{w}_h\}\!\}\Big]_{j+\frac12},$$

$$\mathcal{L}^{(2)}(\boldsymbol{w}_h, \boldsymbol{v}_h) = \int_I (\boldsymbol{v}_h)_x^\top \boldsymbol{M}(\boldsymbol{u})\boldsymbol{f}'_{\boldsymbol{u}}(\boldsymbol{u})\boldsymbol{w}_h \,\mathrm{d}x + \sum_{1\le j\le N}\Big[\![\boldsymbol{M}(\boldsymbol{u}_b)\boldsymbol{v}_h]\!]^\top \boldsymbol{f}'_{\boldsymbol{u}}(\boldsymbol{u}_b)\{\!\{\boldsymbol{w}_h\}\!\}\Big]_{j+\frac12},$$

where $\boldsymbol{M}(\boldsymbol{u}) = \boldsymbol{Q}(\boldsymbol{u}_c) - \boldsymbol{Q}(\boldsymbol{u})$. Note that $\boldsymbol{M}(\boldsymbol{u})$ may be discontinuous at the element interface, and hence $\boldsymbol{M}(\boldsymbol{u}_b)$ has two limits from different directions.

Now we consider the first term $\mathcal{L}^{(1)}(\boldsymbol{w}_h, \boldsymbol{v}_h)$. The two terms involved here are denoted by $\mathcal{L}^{(1)}_{\mathrm{int}}$ and $\mathcal{L}^{(1)}_{\mathrm{bry}}$, respectively. Recall that $\boldsymbol{H}(\boldsymbol{u}) = \boldsymbol{Q}^{1/2}(\boldsymbol{u})\boldsymbol{K}(\boldsymbol{u})\boldsymbol{Q}^{1/2}(\boldsymbol{u})$, and $\boldsymbol{K}(\boldsymbol{u})$ is a symmetric matrix satisfying $\|\boldsymbol{K}(\boldsymbol{u})\|_{\mathrm{M}} = \rho(\boldsymbol{f}'_{\boldsymbol{u}}(\boldsymbol{u}))$. Noting (2.8), a simple application of the inverse property (i) yields that

$$\begin{aligned}
|\mathcal{L}^{(1)}_{\mathrm{int}}| &\le \Big|\int_I (\boldsymbol{v}_h)_x \boldsymbol{H}(\boldsymbol{u}_c)\boldsymbol{w}_h \,\mathrm{d}x\Big| + \Big|\int_I (\boldsymbol{v}_h)_x\Big[\boldsymbol{H}(\boldsymbol{u}) - \boldsymbol{H}(\boldsymbol{u}_c)\Big]\boldsymbol{w}_h \,\mathrm{d}x\Big| \\
&\le \|\boldsymbol{K}(\boldsymbol{u}_c)\|_\infty\|(\boldsymbol{Q}_c^{1/2}\boldsymbol{v}_h)_x\|\|\boldsymbol{Q}_c^{1/2}\boldsymbol{w}_h\| + \|\boldsymbol{H}(\boldsymbol{u}) - \boldsymbol{H}(\boldsymbol{u}_c)\|_\infty\|(\boldsymbol{v}_h)_x\|\|\boldsymbol{w}_h\| \\
&\le S_{\max}\mu h^{-1}\|\boldsymbol{Q}_c^{1/2}\boldsymbol{w}_h\|\|\boldsymbol{Q}_c^{1/2}\boldsymbol{v}_h\| + C_\star\|\boldsymbol{v}_h\|\|\boldsymbol{w}_h\|,
\end{aligned} \tag{4.10}$$

since $\|\boldsymbol{H}(\boldsymbol{u}) - \boldsymbol{H}(\boldsymbol{u}_c)\|_{\mathrm{M}} \le C_\star h$ holds for all x, due to (4.3) and the smoothness assumption H1 on $\boldsymbol{u}$. Also we can use (4.7a) and the boundedness of $\boldsymbol{Q}_c$ (see assumption H3), to get

$$\begin{aligned}
|\mathcal{L}^{(1)}_{\mathrm{bry}}| &\le S_{\max}\|\boldsymbol{Q}_b^{1/2}[\![\boldsymbol{v}_h]\!]\|_{\Gamma_h}\|\boldsymbol{Q}_b^{1/2}\{\!\{\boldsymbol{w}_h\}\!\}\|_{\Gamma_h} \\
&\le KS_{\max}\mu h^{-1}\|\boldsymbol{Q}_c^{1/2}\boldsymbol{w}_h\|\|\boldsymbol{Q}_c^{1/2}\boldsymbol{v}_h\| + C_\star\|\boldsymbol{v}_h\|\|\boldsymbol{w}_h\|.
\end{aligned} \tag{4.11}$$

Since $\|\boldsymbol{M}(\boldsymbol{u})\|_{\mathrm{M}} = \mathcal{O}(h)$ also holds for all x (as we have discussed on the focus shifting), using Cauchy–Schwarz inequality, as well as the inverse properties (i) and (ii), we achieve

$$|\mathcal{L}^{(2)}(\boldsymbol{w}_h, \boldsymbol{v}_h)| \le C_\star\|\boldsymbol{w}_h\|\|\boldsymbol{v}_h\|. \tag{4.12}$$

Hence, collecting up the above inequalities completes the proof of this lemma. □

In this paper the following conclusion plays an important role. The two inequalities reflect the approximate skew-symmetric and negative semidefinite properties, respectively.

Lemma 4.2. *For any functions $\boldsymbol{w}_h$ and $\boldsymbol{v}_h$ in $\mathbb{V}_h$, there holds*

$$|\mathcal{L}(\boldsymbol{w}_h, \boldsymbol{Q}_c\boldsymbol{v}_h) + \mathcal{L}(\boldsymbol{v}_h, \boldsymbol{Q}_c\boldsymbol{w}_h)| \le C_\star \|\boldsymbol{w}_h\| \|\boldsymbol{v}_h\|. \tag{4.13}$$

As a corollary, we have $|\mathcal{L}(\boldsymbol{v}_h, \boldsymbol{Q}_c\boldsymbol{v}_h)| \le C_\star \|\boldsymbol{v}_h\|^2$.

Proof. We just need to estimate $\mathcal{L}^{(1)}(\boldsymbol{w}_h, \boldsymbol{v}_h)$ again, which is defined in the proof of Lemma 4.1. By $\boldsymbol{H}_{\rm prj}(\boldsymbol{u})$ we denote the piecewise linear interpolation of $\boldsymbol{H}(\boldsymbol{u})$. Since this matrix is symmetric, we can have

$$(\boldsymbol{w}_h)_x^\top \boldsymbol{H}_{\rm prj}\boldsymbol{v}_h + (\boldsymbol{v}_h)_x^\top \boldsymbol{H}_{\rm prj}\boldsymbol{w}_h = \Big[\boldsymbol{w}_h^\top \boldsymbol{H}_{\rm prj}\boldsymbol{v}_h\Big]_x - \boldsymbol{w}_h^\top (\boldsymbol{H}_{\rm prj})_x \boldsymbol{v}_h, \tag{4.14}$$

where $(\boldsymbol{H}_{\rm prj})_x|_{I_j} = [\boldsymbol{H}(\boldsymbol{u}_{j+\frac{1}{2}}) - \boldsymbol{H}(\boldsymbol{u}_{j-\frac{1}{2}})]/[x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}]$. As we have discussed in Section 4.1.2, we know that $\|(\boldsymbol{H}_{\rm prj})_x\|_{\rm M}$ is bounded for all x.

Since the solution is periodic or compactly-supported, we can solve the integration in each element and obtain

$$\mathcal{L}^{(1)}(\boldsymbol{w}_h, \boldsymbol{v}_h) + \mathcal{L}^{(1)}(\boldsymbol{v}_h, \boldsymbol{w}_h) = \Xi_{\rm bry} + \Xi_{\rm int}, \tag{4.15}$$

where

$$\Xi_{\rm bry} = \sum_{1\le j\le N} \Big[\llbracket \boldsymbol{w}_h \rrbracket^\top \boldsymbol{H}(\boldsymbol{u}_b) \{\!\{\boldsymbol{v}_h\}\!\} + \llbracket \boldsymbol{v}_h \rrbracket^\top \boldsymbol{H}(\boldsymbol{u}_b) \{\!\{\boldsymbol{w}_h\}\!\} - \llbracket (\boldsymbol{w}_h)^\top \boldsymbol{H}(\boldsymbol{u}_b)\boldsymbol{v}_h \rrbracket\Big]_{j+\frac{1}{2}},$$

$$\Xi_{\rm int} = \int_I \Big\{(\boldsymbol{w}_h)_x^\top \Big[\boldsymbol{H} - \boldsymbol{H}_{\rm prj}\Big]\boldsymbol{v}_h + (\boldsymbol{v}_h)_x^\top \Big[\boldsymbol{H} - \boldsymbol{H}_{\rm prj}\Big]\boldsymbol{w}_h - \boldsymbol{w}_h^\top (\boldsymbol{H}_{\rm prj})_x \boldsymbol{v}_h\Big\}\,{\rm d}x.$$

It is easy to see that $\Xi_{\rm bry} = 0$, since the term in the bracket is equal to zero. Further, we have $|\Xi_{\rm int}| \le C_\star \|\boldsymbol{w}_h\| \|\boldsymbol{v}_h\|$ by the inverse property (i), since the former two matrices involved here are of the size $\mathcal{O}(h)$. Finally, by collecting up the above analysis and (4.12) we complete the proof of this lemma. □

In this paper we would like to use the following compact notations to denote the collection of jumps on the element boundary, like

$$\llbracket \boldsymbol{p} \rrbracket^\top \langle \boldsymbol{H}(\boldsymbol{u}_b) \rangle \llbracket \boldsymbol{q} \rrbracket = \sum_{1\le j\le N} \Big[\llbracket \boldsymbol{p} \rrbracket^\top \langle \boldsymbol{H}(\boldsymbol{u}_b) \rangle \llbracket \boldsymbol{q} \rrbracket\Big]_{j+\frac{1}{2}}, \quad \forall \boldsymbol{p}, \forall \boldsymbol{q}. \tag{4.17}$$

This kind of compact notations will also be used to the matrices $\mathcal{A}, \mathcal{S}$, etc.

Lemma 4.3. *There exists a constant C independent of h and τ, such that*

$$|\mathcal{L}(\boldsymbol{\eta}, \boldsymbol{Q}_c\boldsymbol{v}_h)| \le \varepsilon \llbracket \boldsymbol{v}_h \rrbracket^\top \langle \boldsymbol{H}(\boldsymbol{u}_b) \rangle \llbracket \boldsymbol{v}_h \rrbracket + \varepsilon \|\boldsymbol{v}_h\|^2 + C\varepsilon^{-1} h^{2k+1}, \quad \forall\, \boldsymbol{v}_h \in \mathbb{V}_h. \tag{4.18}$$

Here ε is any given positive constant.

Proof. Similar to what we have done in Lemma 4.1, we move the focus of the rotation matrix in definition (3.9a). This yields $\mathcal{L}(\boldsymbol{\eta}, \boldsymbol{Q}_c\boldsymbol{v}_h) = \mathcal{L}^{(3)}(\boldsymbol{v}_h) + \mathcal{L}^{(4)}(\boldsymbol{v}_h)$, where

$$\mathcal{L}^{(3)}(\boldsymbol{v}_h) = \sum_{1\le j\le N} \Big[\llbracket \boldsymbol{v}_h \rrbracket^\top \boldsymbol{H}(\boldsymbol{u}_b) \{\!\{\boldsymbol{\eta}\}\!\}\Big]_{j+\frac{1}{2}},$$

$$\mathcal{L}^{(4)}(\boldsymbol{v}_h) = \int_I (\boldsymbol{v}_h)_x^\top \boldsymbol{Q}_c \Big[\boldsymbol{f}'_{\boldsymbol{u}}(\boldsymbol{u}) - \boldsymbol{f}'_{\boldsymbol{u}}(\boldsymbol{u}_c)\Big]\boldsymbol{\eta}\,{\rm d}x + \sum_{1\le j\le N} \Big[\llbracket \boldsymbol{M}_b \boldsymbol{v}_h \rrbracket^\top \boldsymbol{f}'_{\boldsymbol{u}}(\boldsymbol{u}_b) \{\!\{\boldsymbol{\eta}\}\!\}\Big]_{j+\frac{1}{2}}.$$

The first term is bounded by Cauchy–Schwarz inequality (2.8) and Young's inequality. The projection property (3.6a) yields that

$$|\mathcal{L}^{(3)}(\boldsymbol{v}_h)| \leq \varepsilon[\![\boldsymbol{v}_h]\!]^\top \langle \boldsymbol{H}(\boldsymbol{u}_b)\rangle [\![\boldsymbol{v}_h]\!] + C\varepsilon^{-1}h^{2k+1}. \tag{4.20}$$

Since $\|\boldsymbol{f}'_{\boldsymbol{u}}(\boldsymbol{u}) - \boldsymbol{f}'_{\boldsymbol{u}}(\boldsymbol{u}_c)\|_{\mathrm{M}} = \mathcal{O}(h)$ holds for all x, as we have discussed in Section 4.1.2, we can bound the second term in the form

$$|\mathcal{L}^{(4)}(\boldsymbol{v}_h)| \leq C\Big[\|\boldsymbol{v}_h\|^2 + \|\boldsymbol{\eta}\|^2 + h\|\boldsymbol{\eta}\|^2_{\Gamma_h}\Big] \leq \varepsilon\|\boldsymbol{v}_h\|^2 + C\varepsilon^{-1}h^{2k+1}, \tag{4.21}$$

using the inverse properties (i) and (ii), as well as the projection property (3.6a). This completes the proof of this lemma. □

Remark 4.4. In the above discussion, the average $\{\!\{\boldsymbol{\eta}\}\!\}$ comes from the setting of the reference vector and the local L^2-projection. When the local Gauss–Radau projection is used to obtain optimal error estimate for upwind numerical fluxes, this term will be changed into the new setting along the upwind direction. See Section 5.6 for details.

4.3. The nonlinear part

Now we turn to show that the nonlinear part does not cause serious trouble in the error estimate. The conclusion is stated in the following lemma.

Lemma 4.5. *There exists a constant $C_\star \geq 0$ independent of h and τ, such that*

$$|\mathcal{N}(\boldsymbol{Q}_c\boldsymbol{v}_h)| \leq C_\star\|\boldsymbol{v}_h\|^2 + C_\star h^{-2}\|\boldsymbol{e}\|^2_\infty\Big[\|\boldsymbol{\xi}\|^2 + h^{2k+2}\Big], \quad \forall\, \boldsymbol{v}_h \in \mathbb{V}_h. \tag{4.22}$$

Proof. Noticing (4.4), we can bound the nonlinear term $\mathcal{N}(\boldsymbol{Q}_c\boldsymbol{v}_h)$ in the form

$$\begin{aligned}
|\mathcal{N}(\boldsymbol{Q}_c\boldsymbol{v}_h)| &\leq C_\star \sum_{1\leq j\leq N}\left[\Big(\|[\![\boldsymbol{v}_h]\!]\|_{\mathrm{M}}\|\{\!\{\boldsymbol{e}\}\!\}\|^2_{\mathrm{M}}\Big)_{j+\frac{1}{2}} + \int_{I_j}\|(\boldsymbol{v}_h)_x\|_{\mathrm{M}}\|\boldsymbol{e}\|^2_{\mathrm{M}}\,\mathrm{d}x\right]\\
&\leq C_\star\|\boldsymbol{e}\|_\infty\Big[\|[\![\boldsymbol{v}_h]\!]\|_{\Gamma_h}\|[\![\boldsymbol{e}]\!]\|_{\Gamma_h} + \|(\boldsymbol{v}_h)_x\|\|\boldsymbol{e}\|\Big]\\
&\leq C_\star\|\boldsymbol{e}\|_\infty\Big[\mu^{1/2}h^{-1/2}\|\boldsymbol{v}_h\|\Big[\mu^{1/2}h^{-1/2}\|\boldsymbol{\xi}\| + \|\boldsymbol{\eta}\|_{\Gamma_h}\Big] + \mu h^{-1}\|\boldsymbol{v}_h\|(\|\boldsymbol{\xi}\| + \|\boldsymbol{\eta}\|)\Big]\\
&\leq C_\star h^{-1}\|\boldsymbol{e}\|_\infty\|\boldsymbol{v}_h\|\Big[\|\boldsymbol{\xi}\| + h^{k+1}\Big].
\end{aligned}$$

Here we have used the inverse properties (i) and (ii), together with the projection property (3.6a). This and Young's inequality complete the proof of this lemma. □

4.4. The numerical viscosity part

In this subsection we consider the numerical viscosity part, which is an important feature of the DG method.

4.4.1. A few propositions

The generalized numerical viscosity matrix $\mathcal{A}(\hat{\boldsymbol{f}};\boldsymbol{u}_h)$ is not guaranteed to be positive semidefinite, causing some inconvenience to the analysis. However, it follows from assumptions A1 and A2 that there exists a positive semidefinite matrix $\mathcal{M}(\hat{\boldsymbol{f}};\boldsymbol{p})$ such that

$$\mathcal{S}(\hat{\boldsymbol{f}};\boldsymbol{u}_h) = \mathcal{A}(\hat{\boldsymbol{f}};\boldsymbol{u}_h) + \mathcal{M}(\hat{\boldsymbol{f}};\boldsymbol{u}_h) \tag{4.23}$$

is a symmetric positive semidefinite matrix, and

$$\|\mathcal{M}(\hat{\boldsymbol{f}};\boldsymbol{u}_h)\|_{\mathrm{M}} \leq C_\star\|[\![\boldsymbol{u}_h]\!]\|_{\mathrm{M}}. \tag{4.24}$$

The total numerical viscosity in the DG spatial discretization can be shown explicitly in the form

$$[\![\boldsymbol{u}_h]\!]^\top \mathcal{S}[\![\boldsymbol{u}_h]\!] = \sum_{1\le j\le N} \Big[[\![\boldsymbol{u}_h]\!]^\top \mathcal{S}(\hat{\boldsymbol{f}};\boldsymbol{u}_h)[\![\boldsymbol{u}_h]\!]\Big]_{j+\frac{1}{2}}. \tag{4.25}$$

Note that this treatment is not essential, but only for the analysis. For example, the second Cauchy–Schwarz inequality in (2.8) holds for $\mathcal{S}(\hat{\boldsymbol{f}};\boldsymbol{u}_h)$, but not for $\mathcal{A}(\hat{\boldsymbol{f}};\boldsymbol{u}_h)$.

Associated with this issue, we will encounter another expression $[\![\boldsymbol{u}_h]\!]^\top\langle \boldsymbol{H}(\boldsymbol{u}_b)\rangle[\![\boldsymbol{u}_h]\!]$. Below we would like to present some useful propositions between them.

Proposition 4.6. *There exists a bounding constant $C_\star$ such that*

$$\frac{1}{2}[\![\boldsymbol{u}_h]\!]^\top\langle \boldsymbol{H}(\boldsymbol{u}_b)\rangle[\![\boldsymbol{u}_h]\!] \le [\![\boldsymbol{u}_h]\!]^\top \mathcal{S}[\![\boldsymbol{u}_h]\!] + C(\boldsymbol{e})(\|\boldsymbol{\xi}\|^2 + h^{2k+2}).$$

Note that the notation $C_\star$ is included in $C(\boldsymbol{e})$; see (4.1).

Proof. Noticing (4.23), a simple manipulation yields

$$\frac{1}{2}[\![\boldsymbol{u}_h]\!]^\top\langle \boldsymbol{H}(\boldsymbol{u}_b)\rangle[\![\boldsymbol{u}_h]\!] + [\![\boldsymbol{u}_h]\!]^\top \mho[\![\boldsymbol{u}_h]\!] = [\![\boldsymbol{u}_h]\!]^\top \mathcal{S}[\![\boldsymbol{u}_h]\!] + \Lambda_3 + \Lambda_4, \tag{4.26}$$

where $\mho$ is the symmetric positive semidefinite matrix given in assumption A3, and

$$\Lambda_3 = \frac{1}{2}[\![\boldsymbol{u}_h]\!]^\top\Big(\langle \boldsymbol{H}(\boldsymbol{u}_b)\rangle - \langle \boldsymbol{H}(\{\!\{\boldsymbol{u}_h\}\!\})\rangle - \mathcal{M}(\hat{\boldsymbol{f}};\boldsymbol{u}_h)\Big)[\![\boldsymbol{u}_h]\!],$$
$$\Lambda_4 = [\![\boldsymbol{u}_h]\!]^\top\Big(\frac{1}{2}\langle \boldsymbol{H}(\{\!\{\boldsymbol{u}_h\}\!\})\rangle + \mho - \mathcal{A}(\hat{\boldsymbol{f}},\boldsymbol{u}_h)\Big)[\![\boldsymbol{u}_h]\!].$$

We can bound easily $|\Lambda_3|$ by using (4.3), (4.5) and (4.24). The bound of $|\Lambda_4|$ is given by assumption A3, (4.6) and the inverse property (ii). Together with (4.6) again, we have completed the proof of this proposition. □

Proposition 4.7. *Let ε be any given positive constant. There holds*

$$\frac{1}{2}[\![\boldsymbol{\xi}]\!]^\top\langle \boldsymbol{H}(\boldsymbol{u}_b)\rangle[\![\boldsymbol{\xi}]\!] \le (1+\varepsilon)[\![\boldsymbol{u}_h]\!]^\top \mathcal{S}[\![\boldsymbol{u}_h]\!] + C(\varepsilon;\boldsymbol{e})(\|\boldsymbol{\xi}\|^2 + h^{2k+1}). \tag{4.27}$$

Proof. Since $\langle \boldsymbol{H}\rangle$ is a symmetric positive semidefinite matrix, we can use Cauchy–Schwarz inequality and Young's inequality to get

$$\begin{aligned}[\![\boldsymbol{\xi}]\!]^\top\langle \boldsymbol{H}(\boldsymbol{u}_b)\rangle[\![\boldsymbol{\xi}]\!] &= [\![\boldsymbol{\eta}-\boldsymbol{u}_h]\!]^\top\langle \boldsymbol{H}(\boldsymbol{u}_b)\rangle[\![\boldsymbol{\eta}-\boldsymbol{u}_h]\!]\\ &\le (1+\varepsilon)[\![\boldsymbol{u}_h]\!]^\top\langle \boldsymbol{H}(\boldsymbol{u}_b)\rangle[\![\boldsymbol{u}_h]\!] + (1+\varepsilon^{-1})[\![\boldsymbol{\eta}]\!]^\top\langle \boldsymbol{H}(\boldsymbol{u}_b)\rangle[\![\boldsymbol{\eta}]\!].\end{aligned} \tag{4.28}$$

Then we can finish the proof of this proposition by using the projection property of the finite element space, together with Proposition 4.6. □

4.4.2. Boundedness

Let $F_{\max} = \max \|\boldsymbol{Q}^{-1/2}(\boldsymbol{s})\mathcal{A}(\hat{\boldsymbol{f}};\boldsymbol{s},\boldsymbol{s})\boldsymbol{Q}^{-1/2}(\boldsymbol{s})\|_{\mathrm{M}}$ represent the maximum strength of the numerical viscosity. Here the maximum is taken over all vectors in $\mathbb{R}^m$. Due to the smoothness assumptions H1–H3, we know that $F_{\max}$ is a finite number.

The next lemma will show a rough amplification in each time-marching, for any functions in the finite element space.

Lemma 4.8. *Let $\varepsilon \le 1$ be any positive constant. For any $\boldsymbol{v}_h \in \mathbb{V}_h$, there holds*

$$|\mathcal{V}(\boldsymbol{Q}_c\boldsymbol{v}_h)| \le \varepsilon F_{\max}\mu h^{-1}\|\boldsymbol{Q}_c^{1/2}\boldsymbol{v}_h\|^2 + K\varepsilon^{-1}[\![\boldsymbol{u}_h]\!]^\top \mathcal{S}[\![\boldsymbol{u}_h]\!] + C_\star h^{-1}\|\boldsymbol{e}\|_\infty\Big[\|\boldsymbol{\xi}\|^2 + \|\boldsymbol{v}_h\|^2 + h^{2k+2}\Big]. \tag{4.29}$$

Proof. Recall that $\boldsymbol{r}_1 = \boldsymbol{u}_h^+$ and $\boldsymbol{r}_2 = \boldsymbol{u}_h^-$, with the local rotation positions $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$. It follows from the definition (2.14) that $\mathcal{V}(\boldsymbol{Q}_c\boldsymbol{v}_h) = \mathcal{V}^{(1)}(\boldsymbol{v}_h) + \mathcal{V}^{(2)}(\boldsymbol{v}_h)$, where

$$\mathcal{V}^{(1)}(\boldsymbol{v}_h) = [\![\boldsymbol{v}_h]\!]^\top \mathcal{A}(\hat{\boldsymbol{f}}; \boldsymbol{u}_h)[\![\boldsymbol{u}_h]\!], \tag{4.30a}$$

$$\mathcal{V}^{(2)}(\boldsymbol{v}_h) = -\frac{1}{2}\sum_{\kappa=1,2}\sum_{1\le j\le N}\Big[[\![\boldsymbol{M}_\kappa\boldsymbol{v}_h]\!]^\top(\boldsymbol{f}(\boldsymbol{r}_\kappa) - \hat{\boldsymbol{f}}(\boldsymbol{u}_h))\Big]_{j+\frac{1}{2}}. \tag{4.30b}$$

Here $\boldsymbol{M}_\kappa = \boldsymbol{Q}(\boldsymbol{u}_c) - \boldsymbol{Q}(\boldsymbol{s}_\kappa)$ describes the focus shifting of the local rotation matrix. Below we will estimate these terms one by one.

Noticing (4.23), an application of Cauchy–Schwarz inequality to the first term yields the estimate

$$\begin{aligned}|\mathcal{V}^{(1)}(\boldsymbol{v}_h)| &\le |[\![\boldsymbol{v}_h]\!]^\top \mathcal{S}[\![\boldsymbol{u}_h]\!]| + |[\![\boldsymbol{v}_h]\!]^\top \mathcal{M}[\![\boldsymbol{u}_h]\!]| \\ &\le K\varepsilon^{-1}[\![\boldsymbol{u}_h]\!]^\top \mathcal{S}[\![\boldsymbol{u}_h]\!] + \varepsilon[\![\boldsymbol{v}_h]\!]^\top \mathcal{S}[\![\boldsymbol{v}_h]\!] + |[\![\boldsymbol{v}_h]\!]^\top \mathcal{M}[\![\boldsymbol{u}_h]\!]|.\end{aligned} \tag{4.31}$$

The second term is equal to the sum of $\mathcal{V}^{(1a)} = \varepsilon[\![\boldsymbol{v}_h]\!]^\top \mathcal{A}(\hat{\boldsymbol{f}}; \{\!\{\boldsymbol{u}_h\}\!\})[\![\boldsymbol{v}_h]\!]$ and

$$\mathcal{V}^{(1b)} = \varepsilon[\![\boldsymbol{v}_h]\!]^\top\Big[\mathcal{A}(\hat{\boldsymbol{f}}; \boldsymbol{u}_h) - \mathcal{A}(\hat{\boldsymbol{f}}; \{\!\{\boldsymbol{u}_h\}\!\}) + \mathcal{M}(\hat{\boldsymbol{f}}; \boldsymbol{u}_h)\Big][\![\boldsymbol{v}_h]\!], \tag{4.32}$$

which are bounded by the inverse property (4.7b), and by assumption A2 and (4.24), respectively. This gives

$$\varepsilon[\![\boldsymbol{v}_h]\!]^\top \mathcal{S}[\![\boldsymbol{v}_h]\!] \le \varepsilon K F_{\max}\mu h^{-1}\|\boldsymbol{Q}_c^{1/2}\boldsymbol{v}_h\|^2 + C_\star h^{-1}\|\boldsymbol{e}\|_\infty\|\boldsymbol{v}_h\|^2, \tag{4.33}$$

since $\varepsilon \le 1$. The last term can be bounded by using (4.24) and (4.6), which reads

$$|[\![\boldsymbol{v}_h]\!]^\top \mathcal{M}[\![\boldsymbol{u}_h]\!]| \le C_\star h^{-1}\|\boldsymbol{e}\|_\infty\Big[\|\boldsymbol{v}_h\|^2 + \|\boldsymbol{\xi}\|^2 + h^{2k+2}\Big]. \tag{4.34}$$

As we have shown in Section 4.1.2, we know that both $\|\boldsymbol{M}_\kappa\|_{\mathrm{M}} \le C_\star(h + \|\boldsymbol{e}\|_\infty)$ and $\|\boldsymbol{f}(\boldsymbol{u}_h^+) - \hat{\boldsymbol{f}}(\boldsymbol{u}_h^-, \boldsymbol{u}_h^+)\|_{\mathrm{M}} \le C\|[\![\boldsymbol{u}_h]\!]\|_{\mathrm{M}}$ hold at every element boundary point. The inverse property (ii) and the projection property (3.6a) yield

$$|\mathcal{V}^{(2)}(\boldsymbol{v}_h)| \le C_\star h^{-1}\|\boldsymbol{e}\|_\infty(\|\boldsymbol{\xi}\| + h^{k+1})\|\boldsymbol{v}_h\|. \tag{4.35}$$

Finally, summing up the above estimates completes the proof of this lemma. For convenience we use a new notation ε instead of εK in the final conclusion. □

Furthermore, for a special test function $\boldsymbol{v}_h = \boldsymbol{\xi}$, we can show explicitly the total numerical viscosity in the DG discretization. It is stated in the following lemma.

Lemma 4.9. *For any given positive constant $\varepsilon < 1$, there holds*

$$\mathcal{V}(\boldsymbol{Q}_c\boldsymbol{\xi}) \le -(1-\varepsilon)[\![\boldsymbol{u}_h]\!]^\top \mathcal{S}[\![\boldsymbol{u}_h]\!] + C(\varepsilon; \boldsymbol{e})\Big[\|\boldsymbol{\xi}\|^2 + h^{2k+1}\Big]. \tag{4.36}$$

Proof. At this time, we still use the splitting (4.30), and need to pay more attention to the first term. It reads

$$\mathcal{V}^{(1)}(\boldsymbol{\xi}) = [\![\boldsymbol{\xi}]\!]^\top \mathcal{A}[\![\boldsymbol{u}_h]\!] = [\![\boldsymbol{\eta}]\!]^\top \mathcal{S}[\![\boldsymbol{u}_h]\!] - [\![\boldsymbol{\xi}]\!]^\top \mathcal{M}[\![\boldsymbol{u}_h]\!] - [\![\boldsymbol{u}_h]\!]^\top \mathcal{S}[\![\boldsymbol{u}_h]\!], \tag{4.37}$$

following (4.23) and (4.6). The first two terms on the right-hand side are denoted by Λ_1 and Λ_2, respectively.

The definition (2.14) implies that every component in $\mathcal{A}$ are bounded uniformly, owing to assumptions H2 and H3. Hence we have $\|\mathcal{S}\|_{\mathrm{M}} \leq C + C_\star\|\boldsymbol{e}\|_\infty$, due to (4.24) and (4.6). Employing Young's inequality, we get

$$|\Lambda_1| \leq \varepsilon[\![\boldsymbol{u}_h]\!]^\top \mathcal{S}[\![\boldsymbol{u}_h]\!] + \frac{1}{4\varepsilon}[\![\boldsymbol{\eta}]\!]^\top \mathcal{S}[\![\boldsymbol{\eta}]\!] \leq \varepsilon[\![\boldsymbol{u}_h]\!]^\top \mathcal{S}[\![\boldsymbol{u}_h]\!] + \varepsilon^{-1}\Big[C + C_\star\|\boldsymbol{e}\|_\infty\Big]\|[\![\boldsymbol{\eta}]\!]\|^2_{\Gamma_h}.$$

Using (4.24) and (4.6), as well as the inverse property (ii), we have

$$|\Lambda_2| \leq \|\mathcal{M}\|_\infty \|[\![\boldsymbol{u}_h]\!]\|_{\Gamma_h}\|[\![\boldsymbol{\eta}]\!]\|_{\Gamma_h} \leq C_\star h^{-1}\|\boldsymbol{e}\|_\infty\Big[\|\boldsymbol{\xi}\|^2 + h\|[\![\boldsymbol{\eta}]\!]\|^2_{\Gamma_h}\Big].$$

Substituting the above inequalities into (4.37) yields

$$\mathcal{V}^{(1)}(\boldsymbol{\xi}) \leq -(1-\varepsilon)[\![\boldsymbol{u}_h]\!]^\top \mathcal{S}[\![\boldsymbol{u}_h]\!] + C(\varepsilon;\boldsymbol{e})\Big[\|\boldsymbol{\xi}\|^2 + h^{2k+2}\Big], \tag{4.38}$$

where the projection property (3.6a) has been used.

The next term $\mathcal{V}^{(2)}(\boldsymbol{\xi})$ can be bounded by (4.35). This completes the proof of this lemma. □

Remark 4.10. Actually the numerical viscosity of the DG method can be shown by another process. For example, if we use the reference vector along the upwind direction at each element boundary point, the numerical viscosity can be shown explicitly by the quantity $[\![\boldsymbol{u}_h]\!]^\top\langle \boldsymbol{H}(\boldsymbol{u}_b)\rangle[\![\boldsymbol{u}_h]\!]$. When the local Gauss–Radau projection is used, the numerical viscosity is shown to come mainly from the linear part. This complex line of analysis is only necessary to obtain the optimal error estimate, when the upwind numerical flux is used. In this case, the new viscosity part will basically disappear. See section 5.6.

4.5. The time-marching part

By using the approximation property (3.6c), and the estimate of the local truncation error in time, we can easily get the following conclusion.

Lemma 4.11. *There exists a constant $C > 0$ independent of h and τ, such that*

$$|\mathcal{T}(\boldsymbol{Q}_c\boldsymbol{v}_h)| \leq C\Big[\|\boldsymbol{v}_h\|^2 + h^{2k+2} + \tau^6\Big], \quad \forall v_h \in \mathbb{V}_h. \tag{4.39}$$

Proof. The proof is a simple application of Young's inequality, so it is omitted. □

5. Proof of the theorem

In this section we come back to prove Theorem 2.5, under a standard CFL condition

$$\max\{S_{\max}, F_{\max}\}\mu\tau h^{-1} \leq \lambda_{\max}, \tag{5.1}$$

where $S_{\max}$ and $F_{\max}$ have been defined in the previous section. Note that $\lambda_{\max} \leq 1$ is a suitable CFL number, independent of h and τ.

5.1. Three useful differences

The following energy analysis depends strongly on three important differences, which have been presented in [31] for the scalar case. They read

$$\mathbb{G}_1^n\boldsymbol{\xi}^n = \boldsymbol{\xi}^{n,1} - \boldsymbol{\xi}^n,\ \mathbb{G}_2^n\boldsymbol{\xi}^n = 2\boldsymbol{\xi}^{n,2} - \boldsymbol{\xi}^{n,1} - \boldsymbol{\xi}^n,\ \mathbb{G}_3^n\boldsymbol{\xi}^n = \boldsymbol{\xi}^{n+1} - 2\boldsymbol{\xi}^{n,2} + \boldsymbol{\xi}^n. \tag{5.2}$$

It is worthy to point out that the group $\{\mathbb{E}_\ell^n\boldsymbol{\xi}^n\}_{\ell=1,2,3}$ and the group $\{\mathbb{G}_\ell^n\boldsymbol{\xi}^n\}_{\ell=1,2,3}$ can be linearly expressed by each other.

It is convenient to understand $\mathbb{G}_\ell^n$ as an operator acting on different objects defined on time stage levels, for example, on matrix-valued functions, or on linear functionals. If the object is the error in the finite element space, we would omit the object and use a simplified notation $\mathbb{G}_\ell^n = \mathbb{G}_\ell^n \boldsymbol{\xi}^n$, in this section and in the appendix.

The following conclusion shows that the above differences $\mathbb{G}_\ell^n$ have a strong relationship to approximations of time derivatives.

Lemma 5.1. *For the RKDG3 method (2.4) and $\ell = 0, 1, 2$, we have*

$$(\mathbb{G}_{\ell+1}^n, \boldsymbol{v}_h) = \frac{\tau}{\ell+1} \mathbb{G}_\ell^n \mathcal{K}^n(\boldsymbol{v}_h), \quad \boldsymbol{v}_h \in \mathbb{V}_h. \tag{5.3}$$

Here $\mathbb{G}_0^n$ is just the identity operator, namely $\mathbb{G}_0^n \mathcal{K}^n = \mathcal{K}^n$.

The proof of this lemma is trivial by the suitable linear combinations of those error equations in (3.7). The detailed process is omitted; we refer to [31]. Based on this lemma, we can achieve easily a crude estimate for each stage in L^2-norm within one time step. The proof of the following lemma is given in the appendix.

Lemma 5.2. *Let n be any time level and $\ell = 0, 1, 2$. Assume $\|\boldsymbol{e}^{n,\kappa}\|_\infty \le h$ holds for any $\kappa \le \ell$. Then there holds*

$$\|\boldsymbol{\xi}^{n,\ell+1}\|^2 \le C \sum_{0 \le \kappa \le \ell} \|\boldsymbol{\xi}^{n,\kappa}\|^2 + C h^{2k+2} + C\tau^8, \tag{5.4}$$

under the CFL condition (5.1), where the bounding constant C is independent of n, h and τ. Here $\boldsymbol{\xi}^{n,3} = \boldsymbol{\xi}^{n+1}$ for notational convenience.

5.2. The energy equation

To carry out the energy analysis, we need to establish the energy equation. To this end, we take three test functions $\boldsymbol{Q}_c^n \boldsymbol{\xi}^n$, $\boldsymbol{Q}_c^{n,1} \boldsymbol{\xi}^{n,1}$ and $\boldsymbol{Q}_c^{n,2} \boldsymbol{\xi}^{n,2}$ in (3.7), for $\ell = 0, 1, 2$, respectively. Recall that the piecewise constant rotation matrix is given as

$$\boldsymbol{Q}_c^{n,\ell} = \boldsymbol{Q}(\boldsymbol{u}_c^{n,\ell}), \quad \ell = 0, 1, 2. \tag{5.5}$$

A simple manipulation yields the energy equation in the form

$$3\|(\boldsymbol{Q}_c^{n+1})^{1/2} \boldsymbol{\xi}^{n+1}\|^2 - 3\|(\boldsymbol{Q}_c^n)^{1/2} \boldsymbol{\xi}^n\|^2 = \mathit{\Pi}_1 + \mathit{\Pi}_2 + \mathit{\Pi}_3, \tag{5.6}$$

which are used to reflect the DG discretization, the TVDRK time-marching, and the variation of the rotation matrix in time, respectively. In details, they read

$$\mathit{\Pi}_1 = \tau[\mathcal{K}^n(\boldsymbol{Q}_c^n \boldsymbol{\xi}^n) + \mathcal{K}^{n,1}(\boldsymbol{Q}_c^{n,1} \boldsymbol{\xi}^{n,1}) + 4\mathcal{K}^{n,2}(\boldsymbol{Q}_c^{n,2} \boldsymbol{\xi}^{n,2})], \tag{5.7a}$$

$$\mathit{\Pi}_2 = (\mathbb{G}_2^n, \boldsymbol{Q}_c^n \mathbb{G}_2^n) + 3(\mathbb{G}_1^n, \boldsymbol{Q}_c^n \mathbb{G}_3^n) + 3(\mathbb{G}_2^n, \boldsymbol{Q}_c^n \mathbb{G}_3^n) + 3(\mathbb{G}_3^n, \boldsymbol{Q}_c^n \mathbb{G}_3^n), \tag{5.7b}$$

$$\mathit{\Pi}_3 = 3(\boldsymbol{\xi}^{n+1}, \widetilde{\boldsymbol{Q}}_c^{n,3} \boldsymbol{\xi}^{n+1}) - (\mathbb{E}^{n,2} \boldsymbol{\xi}^n, \widetilde{\boldsymbol{Q}}_c^{n,1} \boldsymbol{\xi}^{n,1}) - 4(\mathbb{E}^{n,3} \boldsymbol{\xi}^n, \widetilde{\boldsymbol{Q}}_c^{n,2} \boldsymbol{\xi}^{n,2}). \tag{5.7c}$$

Here $\widetilde{\boldsymbol{Q}}_c^{n,\ell} = \boldsymbol{Q}_c^{n,\ell} - \boldsymbol{Q}_c^n$ for $\ell = 1, 2, 3$, reflect the focus shifting, with the notations $\boldsymbol{Q}_c^{n,3} = \boldsymbol{Q}_c^{n+1}$ and $\boldsymbol{u}_c^{n,3} = \boldsymbol{u}_c^{n+1}$ for convenience.

Below we will estimate the above three terms one by one. For notational convenience, we would like to use a macro notation

$$\mathit{\Psi}_{\kappa_1,\kappa_2}^n = \sum_{0 \le \ell \le \kappa_1} C(\boldsymbol{e}^{n,\ell}) \left[\sum_{0 \le \ell \le \kappa_2} \|\boldsymbol{\xi}^{n,\ell}\|^2 + h^{2k+1} + \tau^6 \right], \tag{5.8}$$

where κ_1 and κ_2 are either 2 or 3, and the notation $C(\boldsymbol{e}^{n,\ell})$ has been given in (4.1).

5.3. Estimates to the first term and the third term

The estimate to Π_1 is straightforward, since the test function and the error functional are taken at the same time stage level. Namely, we use Lemmas 4.2 and 4.3 for the linear part, Lemma 4.5 for the nonlinear part, Lemma 4.9 for the viscosity part, and Lemma 4.11 for the time-marching part. All the test functions are taken as $\boldsymbol{v}_h = \boldsymbol{\xi}^{n,\ell}$. Then we use Proposition 4.7 and obtain

$$\tau\mathcal{K}^{n,\ell}(\boldsymbol{Q}_c^{n,\ell}\boldsymbol{\xi}^{n,\ell}) \le -(1-\varepsilon)[\![\boldsymbol{u}_h^{n,\ell}]\!]^\top \mathcal{S}^{n,\ell}[\![\boldsymbol{u}_h^{n,\ell}]\!]\tau + C(\varepsilon;\boldsymbol{e}^{n,\ell})\Big[\|\boldsymbol{\xi}^{n,\ell}\|^2 + h^{2k+1} + \tau^6\Big]\tau,$$

for $\ell = 0, 1, 2$. By taking ε small enough, we achieve the estimate

$$\Pi_1 \le -\frac{3}{4}\sum_{\ell=0,1,2}[\![\boldsymbol{u}_h^{n,\ell}]\!]^\top \mathcal{S}^{n,\ell}[\![\boldsymbol{u}_h^{n,\ell}]\!]\tau + \Psi_{2,2}^n\tau. \tag{5.9}$$

Note that the last term gives a very crude bound.

The estimate to Π_2 is more complex, so we present the detailed discussion in the next subsection. As for the third term, a simple application of Cauchy–Schwarz inequality yields

$$\Pi_3 \le C\Big[\|\boldsymbol{\xi}^n\|^2 + \|\boldsymbol{\xi}^{n,1}\|^2 + \|\boldsymbol{\xi}^{n,2}\|^2 + \|\boldsymbol{\xi}^{n+1}\|^2\Big]\tau. \tag{5.10}$$

since $\|\tilde{\boldsymbol{Q}}_c\|_{\mathrm{M}} = \|\boldsymbol{Q}_c^{n,\ell} - \boldsymbol{Q}_c^n\|_{\mathrm{M}} = \mathcal{O}(\tau)$ holds for all x, as we have discussed before.

5.4. Estimate to the second term

The four terms involved in Π_2 are denoted by $\Pi_2^{(\kappa)}$ for $\kappa = 1, 2, 3, 4$. The estimates to them are a little long and technical, depending strongly on the three differences (5.2) and the relationships among them.

One main trick used here is that we consider the first two terms $\Pi_2^{(1)}$ and $\Pi_2^{(2)}$ together. By taking the test functions $\boldsymbol{Q}_c^n\mathbb{G}_2^n$ and $\boldsymbol{Q}_c^n\mathbb{G}_1^n$, respectively, in Lemma 5.1 for $\ell = 1, 2$, we have

$$\begin{aligned}\Pi_2^{(1)} + \Pi_2^{(2)} =& -(\mathbb{G}_2^n, \boldsymbol{Q}_c^n\mathbb{G}_2^n) + 2(\mathbb{G}_2^n, \boldsymbol{Q}_c^n\mathbb{G}_2^n) + 3(\mathbb{G}_3^n, \boldsymbol{Q}_c^n\mathbb{G}_1^n)\\ =& -\|\mathbb{G}_2^n\|_n^2 + \tau\mathbb{G}_1^n\mathcal{K}^n(\boldsymbol{Q}_c^n\mathbb{G}_2^n) + \tau\mathbb{G}_2^n\mathcal{K}^n(\boldsymbol{Q}_c^n\mathbb{G}_1^n)\\ =& -\|\mathbb{G}_2^n\|_n^2 + \Upsilon_1 + \Upsilon_2 + \Upsilon_3 + \Upsilon_4.\end{aligned} \tag{5.11}$$

Note that the new norm $\|\boldsymbol{w}_h\|_n = \|(\boldsymbol{Q}_c^n)^{1/2}\boldsymbol{w}_h\|$ is used for convenience. Here the last four terms represent respectively the linear part, the nonlinear part, the numerical viscosity part and the time-marching part. For example,

$$\Upsilon_1 = \tau\mathbb{G}_1^n\mathcal{L}^n(\boldsymbol{Q}_c^n\mathbb{G}_2^n) + \tau\mathbb{G}_2^n\mathcal{L}^n(\boldsymbol{Q}_c^n\mathbb{G}_1^n). \tag{5.12}$$

We will estimate them separately below.

The term Υ_1 can be bounded along the same line as that for the linear part at the given time stage. However, we need an additional treatment to deal with the variation of the Jacobian matrix $\boldsymbol{f}_{\boldsymbol{u}}'(\boldsymbol{u}^{n,\ell})$ and the local setting of the local rotation matrix $\boldsymbol{Q}_c^{n,\ell}$. This again involves many applications of local shifting of the focus. A key point in the technique is to fully use the approximate skew-symmetric property (see Lem. 4.2). The complete proof is given in the appendix, and we just present the conclusion here.

Lemma 5.3. *There exists a bounding constant C, such that*

$$|\Upsilon_1| \le \frac{1}{8}\sum_{\ell=0,1}[\![\boldsymbol{u}_h^{n,\ell}]\!]^\top \mathcal{S}^{n,\ell}[\![\boldsymbol{u}_h^{n,\ell}]\!]\tau + C\Psi_{2,2}^n\tau. \tag{5.13}$$

Each term in Υ_2 is bounded easily by using Lemma 4.5. Together with the triangle inequalities to the norms $\|\mathbb{G}_1^n\|$ and $\|\mathbb{G}_2^n\|$, we have

$$|\Upsilon_2| \le C\Psi_{2,2}^n\tau. \tag{5.14}$$

There are some minor troubles to estimate Υ_3, since the generalized viscosity matrix $\mathcal{A}^{n,\ell}$ involves different time stages. Thanks to assumption A2, we have the following result to dig out the complete information about $\|\mathbb{G}_2^n\|_n$. The proof is again postponed to the appendix.

Lemma 5.4. *There exist bounding constants K and C, such that*

$$|\Upsilon_3| \le K\lambda_{\max}\|\mathbb{G}_2^n\|_n^2 + \frac{1}{8}\sum_{\ell=0,1}[\![\boldsymbol{u}_h^{n,\ell}]\!]^\top \mathcal{S}^{n,\ell}[\![\boldsymbol{u}_h^{n,\ell}]\!]\tau + C\Psi_{2,2}^n\tau. \tag{5.15}$$

The last term Υ_4 is easily estimated from Lemma 4.9. Together with the projection property (3.6c) and the triangle inequality, we have

$$|\Upsilon_4| \le C\Psi_{2,2}^n\tau. \tag{5.16}$$

Next we turn to estimate separately the remaining terms, $\Pi_2^{(3)}$ and $\Pi_2^{(4)}$. It follows from Lemma 5.1, with different test functions $\boldsymbol{Q}_c^n\mathbb{G}_2^n$ and $\boldsymbol{Q}_c^n\mathbb{G}_3^n$, respectively, for $\ell = 2$, that

$$\Pi_2^{(3)} = 3(\mathbb{G}_3^n, \boldsymbol{Q}_c^n\mathbb{G}_2^n) = \tau\mathbb{G}_2^n\mathcal{K}^n(\boldsymbol{Q}_c^n\mathbb{G}_2^n), \quad \Pi_2^{(4)} = 3\|\mathbb{G}_3^n\|_n^2 = \tau\mathbb{G}_2^n\mathcal{K}^n(\boldsymbol{Q}_c^n\mathbb{G}_3^n). \tag{5.17}$$

Both estimates depend on the boundedness of $\mathbb{G}_2^n\mathcal{K}^n(\boldsymbol{Q}_c^n\boldsymbol{v}_h)$ for any $\boldsymbol{v}_h \in \mathbb{V}_h$. This property can be achieved along almost the same line as we have done in section 4. However, in the discussion on the linear part, we will also encounter the trouble on how to dig out the complete information about $\|\mathbb{G}_2^n\|_n$. The final conclusion is stated in the following lemma.

Lemma 5.5. *There holds for any $\boldsymbol{v}_h \in \mathbb{V}_h$, that*

$$\begin{aligned}|\tau\mathbb{G}_2^n\mathcal{K}^n(\boldsymbol{Q}_c^n\boldsymbol{v}_h)| \le{}& \varepsilon\|\boldsymbol{v}_h\|_n^2 + \sum_{\ell=0,1,2} C_\star h^{-1}\|\boldsymbol{e}^{n,\ell}\|_\infty\|\boldsymbol{v}_h\|^2\tau + C\Psi_{2,2}^n\tau \\ &+ K\varepsilon^{-1}\lambda_{\max}\left[\|\mathbb{G}_2^n\|_n^2 + \sum_{\ell=0,1,2}[\![\boldsymbol{u}_h^{n,\ell}]\!]^\top\mathcal{S}^{n,\ell}[\![\boldsymbol{u}_h^{n,\ell}]\!]\tau\right],\end{aligned} \tag{5.18}$$

where ε is any small positive constant.

We would like to postpone the proof to the appendix. Noticing (5.17), we take $\boldsymbol{v}_h = \mathbb{G}_2^n$ and $\varepsilon = 1/4$ in Lemma 5.5. Therefore,

$$|\Pi_2^{(3)}| \le C\Psi_{2,2}^n\tau + \left[\frac{1}{4} + K\lambda_{\max}\right]\|\mathbb{G}_2^n\|_n^2 + K\lambda_{\max}\sum_{\ell=0,1,2}[\![\boldsymbol{u}_h^{n,\ell}]\!]^\top\mathcal{S}^{n,\ell}[\![\boldsymbol{u}_h^{n,\ell}]\!]\tau. \tag{5.19}$$

Here the triangle inequality to $\|\mathbb{G}_2^n\|$ is used. The estimate to $\Pi_2^{(4)}$ can be achieved by using the same lemma with $\boldsymbol{v}_h = \mathbb{G}_3^n$, and small enough parameter ε. Thus we can get from this process that

$$|\Pi_2^{(4)}| \le C\Psi_{2,3}^n\tau + K\lambda_{\max}\left[\|\mathbb{G}_2^n\|_n^2 + \sum_{\ell=0,1,2}[\![\boldsymbol{u}_h^{n,\ell}]\!]^\top\mathcal{S}^{n,\ell}[\![\boldsymbol{u}_h^{n,\ell}]\!]\tau\right]. \tag{5.20}$$

Here we use the bigger bound $\Psi_{2,3}^n$, since $\mathbb{G}_3^n$ involves the information about $\boldsymbol{\xi}^{n+1}$.

Finally, we collect all of the above estimates, namely, identity (5.11), inequalities (5.13)–(5.16), (5.19) and (5.20). This gives an estimate

$$\Pi_2 \leq -\left[\frac{3}{4} - K_1\lambda_{\max}\right] |\!|\!| \mathbb{G}_2^n |\!|\!|_n^2 + \left[\frac{1}{4} + K_2\lambda_{\max}\right] \sum_{\ell=0,1,2} [\![\boldsymbol{u}_h^{n,\ell}]\!]^\top \mathcal{S}^{n,\ell} [\![\boldsymbol{u}_h^{n,\ell}]\!] \tau + C\Psi_{2,3}^n \tau, \tag{5.21}$$

where K_1 and K_2 solely depend on the inverse constant μ. We have now estimated every term on the right-hand side of the energy equation (5.6).

5.5. The energy estimate

In this subsection we would like to complete the proof of Theorem 2.5. To deal with the nonlinearity of the flux function $\boldsymbol{f}(\boldsymbol{u})$, we would like first make the *a priori* assumption that, for all n, if $n\tau < T$, we have

$$\|\boldsymbol{e}^{n,\ell}\|_\infty \leq h, \quad \ell = 0, 1, 2, \tag{5.22}$$

which holds for h small enough. We note that, for the linear flux $\boldsymbol{f}(\boldsymbol{u}) = \boldsymbol{A}\boldsymbol{u}$ with the constant matrix $\boldsymbol{A}$, this *a priori* assumption is not necessary. Later, we will verify the correctness of this *a priori* assumption for the piecewise polynomials of degree $k \geq 2$.

The assumption (5.22) implies that $C(\boldsymbol{e}^{n,\ell}) \leq C$ for any n satisfying $n\tau < T$ and $\ell = 0, 1, 2$, where the bounding constant C is independent of n, h and τ. Substitute (5.9), (5.21) and (5.10) into the energy identity (5.6). By employing Lemma 5.2, we finally obtain for any n satisfying $n\tau < T$ that

$$3\|(\boldsymbol{Q}_c^{n+1})^{1/2}\boldsymbol{\xi}^{n+1}\|^2 - 3\|(\boldsymbol{Q}_c^n)^{1/2}\boldsymbol{\xi}^n\|^2 + \Theta_{\mathrm{RKDG}}^n \leq C\|\boldsymbol{\xi}^n\|^2\tau + Ch^{2k+1}\tau + C\tau^7, \tag{5.23}$$

where the bounding constant C is independent of n, h and τ, and

$$\Theta_{\mathrm{RKDG}}^n = \left(\frac{1}{2} - K_1\lambda_{\max}\right) |\!|\!| \mathbb{G}_2^n |\!|\!|_n^2 + \left(\frac{1}{2} - K_2\lambda_{\max}\right) \sum_{\ell=0,1,2} [\![\boldsymbol{u}_h^{n,\ell}]\!]^\top \mathcal{S}^{n,\ell} [\![\boldsymbol{u}_h^{n,\ell}]\!] \tau, \tag{5.24}$$

with the constants K_1 and K_2 defined in (5.21). Suppose that the CFL condition (5.1) holds under the restriction

$$\lambda_{\max} \leq \min((2K_1)^{-1}, (2K_2)^{-1}), \tag{5.25}$$

there holds $\Theta_{\mathrm{RKDG}}^n \geq 0$ to reflect the numerical stability of the RKDG3 method. Summing up the estimate (5.23) from 0 to n yields

$$\|\boldsymbol{\xi}^{n+1}\|^2 \leq C \sum_{n'=0}^{n} \|\boldsymbol{\xi}^{n'}\|^2 \tau + C\Big[\|\boldsymbol{\xi}^0\|^2 + h^{2k+1} + \tau^6\Big], \quad \forall n : n\tau < T, \tag{5.26}$$

where the bounding constant $C > 0$ is independent of n, h and τ. Here we have used the uniform equivalence among the norms $\|\cdot\|$ and $\|(\boldsymbol{Q}_c^n)^{1/2}\cdot\| \equiv |\!|\!|\cdot|\!|\!|_n$ for any n, due to the uniform boundedness of $\boldsymbol{Q}_c^n$; see assumption H3.

It follows from the setting of the initial solution that $\boldsymbol{\xi}^0 = 0$. An application of the discrete Gronwall lemma yields the error estimate for the fully-discrete DG scheme with TVDRK3 time-marching, in the form

$$\|\boldsymbol{\xi}^{n+1}\|^2 \leq Ch^{2k+1} + C\tau^6, \quad \forall n : n\tau < T, \tag{5.27}$$

where the bounding constant C is independent of n, h and τ. This inequality together with the projection property (3.6a) yields the conclusion of this theorem.

To complete the proof of this theorem for the generalized E-flux, we need to verify the *a priori* assumption (5.22), by using an induction process, as well as the inverse property (iii) and the projection property (3.6b).

Obviously there holds $\|\boldsymbol{e}^0\|_\infty \le h$, owing to the setting of the initial solution. An application of Lemma 5.2 yields $\|\boldsymbol{\xi}^{0,1}\| \le Ch^{k+1}$. Thus it follows that

$$\|\boldsymbol{e}^{0,1}\|_\infty \le \mu h^{-\frac{1}{2}}\|\boldsymbol{\xi}^{0,1}\| + \|\boldsymbol{\eta}^{0,1}\|_\infty \le Ch^{k+\frac{1}{2}} \le h, \tag{5.28}$$

if h is small enough, since $k \ge 2$. Repeating this discussion shows $\|\boldsymbol{e}^{0,2}\|_\infty \le h$. Next, we will prove the *a priori* assumption (5.22) by induction. Assume it holds on all time level from 0 to n. We can achieve the inequality (5.27) along the above analysis line, which follows

$$\|\boldsymbol{e}^{n+1}\|_\infty \le C\mu h^{-1/2}\|\boldsymbol{\xi}^{n+1}\| + \|\boldsymbol{\eta}^{n+1}\|_\infty \le Ch^k + h^{-1/2}\tau^3 \le h, \tag{5.29}$$

if h is small enough. As we have done for the initial time level, we can employ Lemma 5.2 to get $\|\boldsymbol{e}^{n+1,1}\|_\infty \le h$ and $\|\boldsymbol{e}^{n+1,2}\|_\infty \le h$. Thus the assumption (5.22) is also true for $n+1$, and we complete the proof of the theorem.

Remark 5.6. The main technique used in this paper is the energy analysis. The energy technique does not require a uniform mesh and can be easily generalized to arbitrary triangulation in multi-dimensions and for linear equations with variable coefficients, as well as to some types of non-periodic boundary conditions. We have discussed in [28] the inflow boundary condition for the scalar hyperbolic equation.

Remark 5.7. For linear conservation laws, the error estimates hold for any piecewise polynomials, if the exact solution is smooth enough. However, for nonlinear conservation laws, the error estimates only hold for piecewise polynomials with degree $k \ge d/2+1$, where d is the spatial dimension. This restriction is solely used to ensure the a priori assumption (5.22). As a result, in this paper we do not consider the piecewise linear polynomials, which are rarely used together with the third order TVDRK time-marching.

5.6. Optimal error estimate for upwind fluxes

In this subsection, we would like to upgrade the error estimate to be optimal for the upwind numerical fluxes, as we have stated in Theorem 2.5. Since almost the same analysis line is used here, we would like in this section to only point out the new techniques and main differences to the above.

To obtain the optimal error estimate, we would need to use some standard tricks in the DG analysis, as we have done in [30, 31]. These consist of two main ingredients, both of which are carried out by the help of eigenvector decomposition. Let $\lambda_i(\boldsymbol{u})$ be the eigenvalues of the Jacobian matrix $\boldsymbol{f}'_{\boldsymbol{u}}(\boldsymbol{u})$, which has the left eigenvector $\boldsymbol{l}_i(\boldsymbol{u})$ and the right eigenvector $\boldsymbol{r}_i(\boldsymbol{u})$. It is assumed that $\lambda_i(\boldsymbol{u}), \boldsymbol{l}_i(\boldsymbol{u})$ and $\boldsymbol{r}_i(\boldsymbol{u})$ are smooth mappings which for strictly hyperbolic system follows from the regularity assumption of $\boldsymbol{f}$. Furthermore, we assume for any space point $x \in I$ that $\lambda_i(\boldsymbol{u}(x,t))$ does not change its sign for any $t \in [0,T]$.

The first one is the local Gauss–Radau projection, denoted by $\mathbb{Q}_h\boldsymbol{u}^{n,\ell}$, instead of the local L^2-projection, where $\boldsymbol{u}^{n,\ell}$ is the reference stage solution as we have defined in section 3.1. In each element I_j, this process involves three steps.

1. Transform $\boldsymbol{u}^{n,\ell}$ to the characteristics fields. This is achieved by left multiplying this by the matrix whose rows are the left eigenvectors of $\boldsymbol{f}'_{\boldsymbol{u}}(\boldsymbol{u}_c^n)$.
2. Apply the scalar Gauss–Radau projection to each of the components of the transformed vector-valued function, which depends on the sign of the corresponding eigenvalue of $\boldsymbol{f}'_{\boldsymbol{u}}(\boldsymbol{u}_c^n)$. If the eigenvalue is not smaller than zero, we demand the exact collocation of the corresponding component on the right boundary point. Otherwise, we demand the exact collocation on the left boundary point.
3. The result is transformed back to the original space by left multiplying the matrix whose columns are the right eigenvectors $\boldsymbol{r}_i(\boldsymbol{u}_c^n)$.

For this projection, the projection properties given in Section 3.2 still hold. The advantage of this projection is that we can correct the order reduction in the element boundary.

The other one is the upwind setting of the reference vector for the numerical solution at each element boundary point, denoted by $(\boldsymbol{u}_h^{n,\ell})^{\mathrm{upw}}$, which is the so-called upwind direction we have referred to before. This treatment

is similar as the above and use the eigenvector decomposition of the Jacobian matrix $\boldsymbol{f}'_{\boldsymbol{u}}(\boldsymbol{u}_b^n)$. At the element boundary point $x_{j+1/2}$, this process involves three steps.

1. Transform the two limits of $\boldsymbol{u}_h^{n,\ell}$ to the characteristics fields of $\boldsymbol{f}'_{\boldsymbol{u}}(\boldsymbol{u}_b^n)$.
2. Apply the scalar upwind setting to each of the components of the transformed vector, which depends on the sign of the corresponding eigenvalue of $\boldsymbol{f}'_{\boldsymbol{u}}(\boldsymbol{u}_b^n)$. If the eigenvalue is not smaller than zero, we take the left limit. Otherwise, we take the right limit.
3. The result is transformed back to the original space.

Then we change the definition of the linear part (3.9a), where $\{\!\{\boldsymbol{w}\}\!\}$ is replaced by $\boldsymbol{w}^{\mathrm{upw}}$. It follows from the new linear part that the total numerical viscosity is shown explicitly in the term $[\![\boldsymbol{u}_h^{n,\ell}]\!]^\top \langle \boldsymbol{H}(\boldsymbol{u}_b^{n,\ell})\rangle [\![\boldsymbol{u}_h^{n,\ell}]\!]$.

Also we change the definitions of the nonlinear part (3.9b) and the viscosity part (3.9c), where $\{\!\{\boldsymbol{f}(\boldsymbol{u}_h^{n,\ell})\}\!\}$ is replaced by $\boldsymbol{f}((\boldsymbol{u}_h^{n,\ell})^{\mathrm{upw}})$. Note that we do not need to use the generalized numerical matrix. In this case, the new viscosity part is very small. The corresponding result is modified from the conclusion in Lemma 4.8, where only the last term on the right-hand side is left. It is worthy to point out that this new viscosity part is equal to zero for the special cases that each eigenvalue keeps the sign in the considered range for the exact solution and the numerical solution. If the sign changes locally, there must exist locally a zero eigenvalue. We refer to [31] for the detailed discussions for the scalar case. This is the key ingredient of the upwind numerical flux.

6. Concluding remarks

In this paper we present quasi-optimal or optimal error estimates to the RKDG scheme with the explicit third order TVDRK time discretization, to solve a symmetrizable system of conservation laws with sufficiently smooth solution. Symmetrizable but not symmetric systems have several difficult issues. Although many numerical fluxes have been used successively in practical computation, the theoretical framework for error analysis has not been clear. In this paper we attempt to set an abstract framework to handle these difficulties, and to establish the so-called generalized numerical viscosity matrix. Another difficulty is related to the local variation of the numerical viscosity, which can be handled by the definition of the local rotation matrix. We remark that this local rotation matrix causes complexity in the analysis. The abstract framework established in this paper includes many commonly used numerical fluxes and is flexible to use for our analysis.

In this paper the exact solution of the conservation laws is assumed to be sufficiently smooth. However, generic solutions of nonlinear conservation laws contain discontinuities. In future work, we will consider piecewise smooth solutions and attempt to analyze the error in smooth regions away from discontinuities, extending the preliminary results for one-dimensional linear equations in [32].

Appendix

A.1. Verification for the Lax–Friedrichs flux

In this subsection we would like to show that the (global/local) Lax–Friedrichs flux is a generalized E-flux, and the corresponding generalized numerical viscosity matrix satisfies assumptions A2 and A3. The (global/local) Lax–Friedrichs flux is defined as follows:

$$\hat{\boldsymbol{f}}^{\mathrm{LF}}(\boldsymbol{p}^-,\boldsymbol{p}^+)=\frac{1}{2}\Big[\boldsymbol{f}(\boldsymbol{p}^-)+\boldsymbol{f}(\boldsymbol{p}^+)-\alpha[\![\boldsymbol{p}]\!]\Big],\quad \text{with } \alpha=\max_{\boldsymbol{p}}\varrho(\boldsymbol{f}'_{\boldsymbol{u}}(\boldsymbol{p})), \tag{A.1}$$

where the maximum is taken locally or globally.

Let $\boldsymbol{s}_1$ be any point in the super-rectangle with the vertexes $\boldsymbol{p}^{\pm}$, and define

$$\boldsymbol{Y}_1^{\mathrm{LF}}=\frac{1}{2}\int_0^1\Big(\boldsymbol{Q}(\boldsymbol{s})-\boldsymbol{Q}(\boldsymbol{s}_1)\Big)(\boldsymbol{f}'_{\boldsymbol{u}}(\boldsymbol{s})+\alpha\boldsymbol{I})\mathrm{d}s,$$

where $\boldsymbol{I}$ is the identity matrix, and $\boldsymbol{s} = \boldsymbol{p}^- + s[\![\boldsymbol{p}]\!]$. Every element in $\boldsymbol{Y}_1^{\mathrm{LF}}$ is bounded by $\mathcal{O}(\|[\![\boldsymbol{p}]\!]\|_{\mathrm{M}})$, since $\boldsymbol{v}_{\boldsymbol{u}}'$ is Lipschitz continuous and $\|\boldsymbol{s} - \boldsymbol{s}_1\|_{\mathrm{M}} \le \|[\![\boldsymbol{p}]\!]\|_{\mathrm{M}}$. A simple manipulation yields

$$[\![\boldsymbol{p}]\!]^\top \boldsymbol{v}_{\boldsymbol{u}}'(\boldsymbol{s}_1)(\boldsymbol{f}(\boldsymbol{p}^+) - \hat{\boldsymbol{f}}^{\mathrm{LF}}(\boldsymbol{p}^-, \boldsymbol{p}^+)) + [\![\boldsymbol{p}]\!]^\top \boldsymbol{Y}_1^{\mathrm{LF}}[\![\boldsymbol{p}]\!] = [\![\boldsymbol{p}]\!]^\top \widetilde{\boldsymbol{Y}}_1^{\mathrm{LF}}[\![\boldsymbol{p}]\!] \ge 0, \tag{A.2}$$

where

$$\widetilde{\boldsymbol{Y}}_1^{\mathrm{LF}} = \frac{1}{2}\int_0^1 \boldsymbol{Q}(\boldsymbol{s})(\boldsymbol{f}_{\boldsymbol{u}}'(\boldsymbol{s}) + \alpha \boldsymbol{I})\mathrm{d}s,$$

is a symmetric positive semidefinite matrix, following (2.6) and (A.1). Similar discussion can be done for $\kappa = 2$. So the Lax–Friedrichs flux is a generalized E-flux.

Next we would like to verify assumption A2. We still consider $\kappa = 1$ as an example. In fact, it follows from the definition of the Lax–Friedrichs flux that

$$\left(\mathcal{D}\hat{\boldsymbol{f}}^{(1)}[\boldsymbol{p}^-, \boldsymbol{p}^+]\right)_{ij} = \frac{1}{2}\left[\frac{\partial f_i}{\partial p_j}(\tilde{\boldsymbol{p}}_{ij}) + \alpha\right], \quad i, j = 1, 2, \ldots, m, \tag{A.3}$$

where $\tilde{\boldsymbol{p}}_{ij}$ is a mean point lying near $\{\!\{\boldsymbol{p}\}\!\}$. Since $\boldsymbol{s}_1 = \boldsymbol{s}_1(\boldsymbol{p}^-, \boldsymbol{p}^+)$ also lies near $\{\!\{\boldsymbol{p}\}\!\}$, it follows from the Lipschitz continuity of $\boldsymbol{f}_{\boldsymbol{u}}'$ and $\boldsymbol{Q}$, that $\mathcal{A}_1(\hat{\boldsymbol{f}}; \boldsymbol{p})$ is Lipschitz continuous. Similar discussion for $\mathcal{A}_2(\hat{\boldsymbol{f}}; \boldsymbol{p})$ implies assumption A2.

To verify assumption A3, we take $\mho = \frac{1}{2}\boldsymbol{Q}^{1/2}[\alpha \boldsymbol{I} - \boldsymbol{K}(\{\!\{\boldsymbol{p}\}\!\})]\boldsymbol{Q}^{1/2}$, which is a symmetric positive semidefinite matrix. Noticing (2.17), a simple manipulation yields

$$\text{LHS of (2.16c)} = \left| -\frac{1}{2}\sum_{\kappa=1,2} [\![\boldsymbol{p}]\!]^\top \Big[\boldsymbol{Q}(\boldsymbol{s}_\kappa) - \boldsymbol{Q}(\{\!\{\boldsymbol{p}\}\!\})\Big]\Big[\boldsymbol{f}(\boldsymbol{r}_\kappa) - \hat{\boldsymbol{f}}^{\mathrm{LF}}(\boldsymbol{p}^-, \boldsymbol{p}^+)\Big] \right|. \tag{A.4}$$

Noticing the Lipschitz continuity of $\boldsymbol{Q}$ and $\hat{\boldsymbol{f}}^{\mathrm{LF}}$, we can complete the verification of assumption A3.

A.2. Proof of Lemma 5.2

We can obtain this lemma by an induction process, to bound $\|\mathbb{G}_{\ell+1}^n\|$ by $\|\mathbb{G}_\ell^n\|$ and the others. This can be established by Lemma 5.1, with the test function $\boldsymbol{v}_h = \boldsymbol{Q}_c^n \mathbb{G}_{\ell+1}^n$.

The key here is to estimate the error functional $\mathbb{G}_\ell^n \mathcal{K}(\boldsymbol{Q}_c^n \boldsymbol{v}_h)$ for any test function in the finite element space. Similar analysis will been given in the proof of Lemma 5.5, so the detail is omitted here. The only difference is that after we bound the numerical viscosity part by Lemma 4.8, we amplify deeply those jumps by the inverse property (ii). Namely, for any $\boldsymbol{v}_h \in \mathbb{V}_h$,

$$\begin{aligned} |\mathcal{V}(\boldsymbol{u}_h; \boldsymbol{Q}_c^n \boldsymbol{v}_h)| \le\ & K F_{\max} \mu h^{-1}\Big[\varepsilon \|(\boldsymbol{Q}_c^n)^{1/2}\boldsymbol{v}_h\|^2 + \varepsilon^{-1}\|(\boldsymbol{Q}_c^n)^{1/2}\boldsymbol{\xi}\|^2\Big] \\ & + C_\star h^{-1}\|\boldsymbol{e}\|_\infty \Big[\|\boldsymbol{\xi}\|^2 + \|\boldsymbol{v}_h\|^2 + h^{2k+2}\Big]. \end{aligned} \tag{A.5}$$

Here ε is any given small positive constant.

Finally, we just need to use the CFL condition (5.1) and the assumption to simplify the obtained conclusion. This will yield the conclusion of this lemma. □

A.3. Proof of Lemma 5.3

We can prove this lemma from the properties of the linear part and the equivalent expression $\Upsilon_1 = \Phi_{\mathrm{DG}} + \Phi_{\mathrm{prj}} + \Phi_{\mathrm{RK}}$, where

$$\begin{aligned} \Phi_{\mathrm{DG}} &= \tau \mathcal{L}_{\mathrm{main}}(\boldsymbol{f}_{\boldsymbol{u}}'(\boldsymbol{u}^n); \mathbb{G}_1^n, \boldsymbol{Q}_c^n \mathbb{G}_2^n) + \tau \mathcal{L}_{\mathrm{main}}(\boldsymbol{f}_{\boldsymbol{u}}'(\boldsymbol{u}^n); \mathbb{G}_2^n, \boldsymbol{Q}_c^n \mathbb{G}_1^n), \\ \Phi_{\mathrm{prj}} &= -\tau \mathcal{L}_{\mathrm{main}}(\boldsymbol{f}_{\boldsymbol{u}}'(\boldsymbol{u}^n); \mathbb{G}_1^n \boldsymbol{\eta}^n, \boldsymbol{Q}_c^n \mathbb{G}_2^n) - \tau \mathcal{L}_{\mathrm{main}}(\boldsymbol{f}_{\boldsymbol{u}}'(\boldsymbol{u}^n); \mathbb{G}_2^n \boldsymbol{\eta}^n, \boldsymbol{Q}_c^n \mathbb{G}_1^n), \\ \Phi_{\mathrm{RK}} &= \tau \mathcal{L}_{\mathrm{main}}(\boldsymbol{N}^{n,1}; \boldsymbol{e}^{n,1}, \boldsymbol{Q}_c^n \mathbb{G}_2^n - \boldsymbol{Q}_c^n \mathbb{G}_1^n) + 2\tau \mathcal{L}_{\mathrm{main}}(\boldsymbol{N}^{n,2}; \boldsymbol{e}^{n,2}, \boldsymbol{Q}_c^n \mathbb{G}_1^n). \end{aligned}$$

Here $\boldsymbol{N}^{n,\ell} = \boldsymbol{f}'_{\boldsymbol{u}}(\boldsymbol{u}^{n,\ell}) - \boldsymbol{f}'_{\boldsymbol{u}}(\boldsymbol{u}^n)$ for $\ell = 1, 2$. Below we will estimate them one by one. It follows from Lemma 4.2 that

$$|\Phi_{\rm DG}| \le C_\star \|\mathbb{G}_1^n\| \|\mathbb{G}_2^n\| \tau \le C_\star \Big[\|\mathbb{G}_1^n\|^2 + \|\mathbb{G}_2^n\|^2 \Big] \tau. \tag{A.6}$$

This is the key step in the proof. Each term in $\Phi_{\rm prj}$ can be bounded by Lemma 4.3 and Proposition 4.7. That is to say

$$\begin{aligned} |\Phi_{\rm prj}| &\le \varepsilon \sum_{\ell=1,2} [\![\mathbb{G}_\ell^n]\!]^\top \langle \boldsymbol{H}(\boldsymbol{u}_b^n) \rangle [\![\mathbb{G}_\ell^n]\!] \tau + \varepsilon \sum_{\ell=1,2} \|\mathbb{G}_\ell^n\|^2 \tau + C h^{2k+1} \tau \\ &\le K\varepsilon(1+\varepsilon) \sum_{\ell=0,1} [\![\boldsymbol{u}_h^{n,\ell}]\!]^\top \mathcal{S}^{n,\ell} [\![\boldsymbol{u}_h^{n,\ell}]\!] \tau + \sum_{\ell=0,1} C(\varepsilon; \boldsymbol{e}^{n,\ell}) (\|\boldsymbol{\xi}^{n,\ell}\|^2 + h^{2k+1}) \tau, \end{aligned} \tag{A.7}$$

where the triangle inequalities are used for $\|\mathbb{G}_1^n\|$ and $\|\mathbb{G}_2^n\|$. Since $\|\boldsymbol{N}^{n,\ell}\|_{\rm M} = \mathcal{O}(\tau) = \mathcal{O}(h)$ holds for all x, along almost the same line as that in Lemmas 4.1 and 4.3, we have

$$|\Phi_{\rm RK}| \le C_\star \Big[\|\boldsymbol{\xi}^n\|^2 + \|\boldsymbol{\xi}^{n,1}\|^2 + \|\boldsymbol{\xi}^{n,2}\|^2 + \|\boldsymbol{\xi}^{n+1}\|^2 + h^{2k+2} + \tau^6 \Big] \tau. \tag{A.8}$$

Finally we collect up the above estimates and take small enough ε to complete the proof of this lemma. □

A.4. Proof of Lemma 5.4

Recalling the definition of generalized E-fluxes, we need to move the focus of $\boldsymbol{Q}$ from the element center $\boldsymbol{u}_c^n$ to the rotation position $\boldsymbol{s}_\kappa^{n,\ell}$, for $\kappa = 1, 2$. After this is done, the term Υ_3 is changed to $\tilde{\Upsilon}_3$. The difference between them is easy to bound. Since at every element boundary point there holds $\|\boldsymbol{Q}(\boldsymbol{u}_c^n) - \boldsymbol{Q}(\boldsymbol{s}_\kappa^{n,\ell})\|_{\rm M} \le C[h + \tau + \|\boldsymbol{e}^{n,\ell}\|_\infty]$, as stated in Section 4.1.2, the inverse properties (i) and (ii) imply that

$$|\Upsilon_3 - \tilde{\Upsilon}_3| \le C \Psi_{2,2}^n \tau, \tag{A.9}$$

due to $\tau = \mathcal{O}(h)$, where $\Psi_{2,2}^n$ has been given in (5.8). Hence we only need to estimate the new term $\tilde{\Upsilon}_3$, which has the decomposition $\tilde{\Upsilon}_3 = \tilde{\Upsilon}_3(\boldsymbol{\eta}) - \tilde{\Upsilon}_3(\boldsymbol{u}_h)$. Here

$$\tilde{\Upsilon}_3(\boldsymbol{q}) = \tau [\![\mathbb{G}_2^n \boldsymbol{q}^n]\!]^\top \mathbb{G}_1^n (\mathcal{A}^n [\![\boldsymbol{u}_h^n]\!]) + \tau [\![\mathbb{G}_1^n \boldsymbol{q}^n]\!]^\top \mathbb{G}_2^n (\mathcal{A}^n [\![\boldsymbol{u}_h^n]\!]). \tag{A.10}$$

Below we will estimate them separately.

The typical term in $\tilde{\Upsilon}_3(\boldsymbol{\eta})$ is $[\![\mathbb{G}_\kappa \boldsymbol{\eta}]\!]^\top \mathcal{A}^{n,\ell} [\![\boldsymbol{u}_h^{n,\ell}]\!]$, where $\kappa = 1, 2$ and $\ell = 0, 1, 2$. Recall that all the elements in $\mathcal{A}^{n,\ell} \equiv \mathcal{A}(\hat{\boldsymbol{f}}; \boldsymbol{u}_h^{n,\ell})$ are bounded uniformly. Using the Cauchy–Schwarz inequality and the inverse property (ii), it is easy to see that

$$|\tau [\![\mathbb{G}_\kappa \boldsymbol{\eta}]\!]^\top \mathcal{A}^{n,\ell} [\![\boldsymbol{u}_h^{n,\ell}]\!]| \le C \|\mathbb{G}_\kappa^n \boldsymbol{\eta}^n\|_{I_h} \Big[\|\boldsymbol{\xi}^{n,\ell}\|_{I_h} + \|\boldsymbol{\eta}^{n,\ell}\|_{I_h} \Big] \tau \le C \Big[\|\boldsymbol{\xi}^{n,\ell}\| + h^{2k+2} \Big] \tau,$$

due to $\tau = \mathcal{O}(h)$, where the projection properties (3.6a) and (3.6c) are used also. Hence we will have

$$|\tilde{\Upsilon}_3(\boldsymbol{\eta})| \le C \left[\sum_{\ell=0,1,2} \|\boldsymbol{\xi}^{n,\ell}\|^2 + h^{2k+2} \right] \tau \le C \Psi_{2,2}^n \tau. \tag{A.11}$$

In order to estimate sharply the term $\tilde{\Upsilon}_3(\boldsymbol{u}_h)$, we have to dig out the complete information about $\mathbb{G}_2$, and cope with the difficulties resulting from the changing of $\mathcal{A}^{n,\ell}$ at different time stages. Define $\mathcal{B}^{n,\ell} = \mathcal{A}^{n,\ell} - \mathcal{A}^n$. A simple manipulation yields that

$$\tilde{\Upsilon}_3(\boldsymbol{u}_h) = -\tilde{\Upsilon}_3^{(1)} + \tilde{\Upsilon}_3^{(2)} + \tilde{\Upsilon}_3^{(3)},$$

where

$$\tilde{\Upsilon}_3^{(1)} = \tau[\![\boldsymbol{u}_h^n]\!]^\top \mathcal{A}^n [\![\mathbb{G}_2^n \boldsymbol{u}_h^n]\!] + \tau[\![\mathbb{G}_2^n \boldsymbol{u}_h^n]\!]^\top \mathcal{A}^n [\![\boldsymbol{u}_h^n]\!], \tag{A.12a}$$

$$\tilde{\Upsilon}_3^{(2)} = \tau[\![\mathbb{G}_2^n \boldsymbol{u}_h^n]\!]^\top \mathcal{A}^{n,1} [\![\boldsymbol{u}_h^{n,1}]\!] + \tau[\![\boldsymbol{u}_h^{n,1}]\!]^\top \mathcal{A}^{n,1} [\![\mathbb{G}_2^n \boldsymbol{u}_h^n]\!], \tag{A.12b}$$

$$\tilde{\Upsilon}_3^{(3)} = 2\tau[\![\mathbb{G}_1^n \boldsymbol{u}_h^n]\!]^\top \mathcal{B}^{n,2} [\![\boldsymbol{u}_h^{n,2}]\!] - \tau[\![\mathbb{G}_1^n \boldsymbol{u}_h^n]\!]^\top \mathcal{B}^{n,1} [\![\boldsymbol{u}_h^{n,1}]\!] - \tau[\![\boldsymbol{u}_h^{n,1}]\!]^\top \mathcal{B}^{n,1} [\![\mathbb{G}_2^n \boldsymbol{u}_h^n]\!]. \tag{A.12c}$$

The first two terms include every term in the same sup-index between the matrices at least once. Here we do not use the symmetric property of $\mathcal{A}^{n,\ell}$.

Since $\mathcal{A}^{n,\ell} = \mathcal{S}^{n,\ell} - \mathcal{M}^{n,\ell}$, an application of Young's inequality yields

$$\begin{aligned} |\tilde{\Upsilon}_3^{(1)}| &\le \varepsilon [\![\boldsymbol{u}_h^n]\!]^\top \mathcal{S}^n [\![\boldsymbol{u}_h^n]\!] \tau + K\varepsilon^{-1} [\![\mathbb{G}_2^n]\!]^\top \mathcal{S}^n [\![\mathbb{G}_2^n]\!] \tau + C\Psi_{2,2}^n \tau \\ &\le \varepsilon [\![\boldsymbol{u}_h^n]\!]^\top \mathcal{S}^n [\![\boldsymbol{u}_h^n]\!] \tau + K\varepsilon^{-1} \mu F_{\max} \tau h^{-1} \|\mathbb{G}_2^n\|_n^2 + C\Psi_{2,2}^n \tau. \end{aligned} \tag{A.13}$$

The above analysis is almost the same as that in Lemma 4.8. The second term $\tilde{\Upsilon}_3^{(2)}$ can be bounded similarly. It follows from assumption A2 and the continuity of the reference stage solutions that

$$\|\mathcal{B}^{n,\ell}\|_{\mathrm{M}} \le C_\star \|(\boldsymbol{u}_h^{n,\ell} - \boldsymbol{u}_h^n)^{\pm}\|_{\mathrm{M}} \le C_\star (\|\boldsymbol{e}^{n,\ell}\|_\infty + \|\boldsymbol{e}^n\|_\infty + \tau), \quad \ell = 1, 2.$$

Then the inverse property (ii) implies $|\tilde{\Upsilon}_3^{(3)}| \le C\Psi_{2,2}^n \tau$.

Finally, collecting up the above conclusions completes the proof of this lemma. □

A.5. Proof of Lemma 5.5

The proof is almost the same as that for the error functional with respect to the single DG discretization (see Sect. 4). The key here is the estimate to the linear part, where the information about $\mathbb{G}_2^n \equiv \mathbb{G}_2^n \boldsymbol{\xi}^n$ will be found explicitly.

Along the same line as in Lemma 5.3, we fix the flow speed to be the same and rewrite the linear part in the equivalent form $\mathbb{G}_2^n \mathcal{L}^n(\boldsymbol{Q}^n \boldsymbol{v}_h) = \Omega_1 - \Omega_2 + \Omega_3$, where

$$\Omega_1 = \mathcal{L}_{\mathrm{main}}(\boldsymbol{f}_{\boldsymbol{u}}'(\boldsymbol{u}^n); \mathbb{G}_2^n, \boldsymbol{Q}_c^n \boldsymbol{v}_h), \tag{A.14a}$$

$$\Omega_2 = \mathcal{L}_{\mathrm{main}}(\boldsymbol{f}_{\boldsymbol{u}}'(\boldsymbol{u}^n); \mathbb{G}_2^n \boldsymbol{\eta}^n, \boldsymbol{Q}_c^n \boldsymbol{v}_h), \tag{A.14b}$$

$$\Omega_3 = 2\mathcal{L}_{\mathrm{main}}(\boldsymbol{N}^{n,2}; \boldsymbol{e}^{n,2}, \boldsymbol{Q}_c^n \boldsymbol{v}_h) - \mathcal{L}_{\mathrm{main}}(\boldsymbol{N}^{n,1}; \boldsymbol{e}^{n,1}, \boldsymbol{Q}_c^n \boldsymbol{v}_h). \tag{A.14c}$$

Here $\boldsymbol{N}^{n,\ell} = \boldsymbol{f}_{\boldsymbol{u}}'(\boldsymbol{u}^{n,\ell}) - \boldsymbol{f}_{\boldsymbol{u}}'(\boldsymbol{u}^n)$ for $\ell = 1, 2$, as stated in Section A.3. There holds $\|\boldsymbol{N}^{n,\ell}\|_\infty = \mathcal{O}(\tau)$.

Using Lemmas 4.1, it follows from Cauchy–Schwarz inequality and Young's inequality, that

$$|\tau\Omega_1| \le K\varepsilon^{-1} \lambda_{\max}^2 \|\mathbb{G}_2^n\|_n^2 + \varepsilon \|\boldsymbol{v}_h\|_n^2 + C_\star \Big[\|\mathbb{G}_2^n\|^2 + \|\boldsymbol{v}_h\|^2 \Big] \tau. \tag{A.15}$$

Along the same line, we can get $|\tau\Omega_2| \le \varepsilon \|\boldsymbol{v}_h\|_n^2 + C_\star \|\boldsymbol{v}_h\|^2 \tau + Ch^{2k+2}\tau^2$. Furthermore, we can estimate Ω_3 along the same line as that for Lemmas 4.1 and 4.3. Using the inverse properties (i) and (ii), together with approximation property (3.6a), we have $|\tau\Omega_3| \le C_\star (\|\boldsymbol{v}_h\|^2 + \|\boldsymbol{\xi}^{n,1}\|^2 + \|\boldsymbol{\xi}^{n,2}\|^2 + h^{2k+2})\tau$. Finally, the sum of the three inequalities above gives the result to the linear part. Namely,

$$\tau \mathbb{G}_2^n \mathcal{L}^n(\boldsymbol{Q}_c^n \boldsymbol{v}_h) \le \varepsilon \|\boldsymbol{v}_h\|_n^2 + K\varepsilon^{-1} \lambda_{\max}^2 \|\mathbb{G}_2^n\|_n^2 + C_\star \|\boldsymbol{v}_h\|^2 \tau + C\Psi_{2,2}^n \tau. \tag{A.16}$$

The next three parts are easy to estimate. Lemma 4.5 gives the estimate to the nonlinear part in the form

$$\tau \mathbb{G}_2^n \mathcal{N}^n(\boldsymbol{Q}_c^n \boldsymbol{v}_h) \le C\|\boldsymbol{v}_h\|^2 \tau + C\Psi_{2,2}^n \tau. \tag{A.17}$$

In order to estimate the viscosity part, the focus of the local rotation matrix $\boldsymbol{Q}$ will be moved from $\boldsymbol{u}_c^n$ to $\boldsymbol{u}_b^{n,\ell}$, and then be moved back. Using mainly Lemma 4.8, we have

$$\tau\mathbb{G}_2^n\mathcal{V}^n(\boldsymbol{Q}_c^n\boldsymbol{v}_h) \le \varepsilon\|\boldsymbol{v}_h\|_n^2 + C\varepsilon^{-1}F_{\max}\mu\tau h^{-1}\sum_{\ell=0,1,2}\tau[\![\boldsymbol{u}_h^{n,\ell}]\!]^\top\mathcal{S}^{n,\ell}[\![\boldsymbol{u}_h^{n,\ell}]\!] + \sum_{\ell=0,1,2}C_\star h^{-1}\|\boldsymbol{e}^{n,\ell}\|_\infty\left[\sum_{\ell=0,1,2}\|\boldsymbol{\xi}^{n,\ell}\|^2 + \|\boldsymbol{v}_h\|^2 + h^{2k+2}\right]\tau. \tag{A.18}$$

Note that we have used a new parameter ε. It is easy to bound the time-marching part, in the form

$$\tau\mathbb{G}_2^n\mathcal{T}^n(\boldsymbol{Q}_c^n\boldsymbol{v}_h) \le \varepsilon\|\boldsymbol{v}_h\|^2\tau + C(h^{2k+2}+\tau^6)\tau. \tag{A.19}$$

Here we have also lost a factor τ, since $\tau < 1$ as we have assumed.

Finally, we collect up the above four estimates and complete the proof of this lemma, since $\lambda_{\max} \le 1$. □

References

[1] R.A. Adams, Sobolev Spaces. Academic Press, New York (1975).

[2] E. Burman, A. Ern and M.A. Fernandez, Explicit Runge–Kutta schemes and finite elements with symmetric stabilization for first-order linear PDE systems. *SIAM. J. Numer. Anal.* **48** (2010) 2019–2042.

[3] P.G. Ciarlet, Finite Element Method for Elliptic Problems. North–Holland, Amsterdam (1978).

[4] B. Cockburn and J. Guzmán, Error estimates for the Runge–Kutta discontinuous Galerkin method for the transport equation with discontinuous initial data. *SIAM. J. Numer. Anal.* **46** (2008) 1364–1398.

[5] B. Cockburn and C.-W. Shu, TVB Runge–Kutta local projection discontinuous Galerkin finite element method for scalar conservation laws II: General framework. *Math. Comput.* **52** (1989) 411–435.

[6] B. Cockburn and C.-W. Shu, The Runge–Kutta local projection P^1-discontinuous Galerkin method for scalar conservation laws. *RAIRO Modél. Math. Anal. Numér.* **25** (1991) 337–361.

[7] B. Cockburn and C.-W. Shu, The Runge–Kutta discontinuous Galerkin finite element method for conservation laws V: Multidimensional systems *J. Comput. Phys.* **141** (1998a) 199–224.

[8] B. Cockburn and C.-W. Shu, The local discontinuous Galerkin method for time-dependent convection-diffusion systems *SIAM. J. Numer. Anal.* **35** (1998b) 2440–2463.

[9] B. Cockburn and C.-W. Shu, Runge–Kutta discontinuous Galerkin methods for convection-dominated problems *J. Sci. Comput.* **16** (2001) 173–261.

[10] B. Cockburn, S. Hou, and C.-W. Shu, TVB Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws IV: The multidimensional case. *Math. Comp.* **54** (1990) 545–581.

[11] B. Cockburn, C. Johnson, C.-W. Shu and E. Tadmor, An introduction to the discontinuous Galerkin method for convection-dominated problems, in *Advanced Numerical Approximation of Nonlinear Hyperbolic Equations*, edited by Quarteroni. Vol. 1697 of *Lect. Notes Math.* Springer, Berlin (1998) 151–268.

[12] B. Cockburn, S.-Y. Lin, and C.-W. Shu, TVB Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws III: One dimensional systems. *J. Comput. Phys.* **84** (1989) 90–113.

[13] G.H. Golub and C.F. Van Loan, Matrix Computations. Posts and Telecom Press (2011).

[14] A. Harten, On the symmetric form of systems of conservation laws with entropy. *J. Comput. Phys.* **49** (1983a) 151–164.

[15] A. Harten, High resolution schemes for hyperbolic conservation laws. *J. Comput. Phys.* **49** (1983b) 357–393.

[16] S.-M. Hou and X.-D. Liu, Solutions of multidimensional hyperbolic systems of conservation laws by square entropy condition satisfying discontinuous Galerkin method. *J. Sci. Comput.* **31** (2007) 127–151.

[17] G.-S. Jiang and C.-W. Shu, On cell entropy inequality for discontinuous Galerkin methods. *Math. Comp.* **62** (1994) 531–538.

[18] C. Johnson and J. Pitkäranta, An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation. *Math. Comp.* **46** (1986) 1–26.

[19] R. Kress, Numerical analysis. Springer-Verlag (1998).

[20] P. Lesaint and P.A. Raviart, On a finite element method for solving the neutron transport equation, in *Mathematical Aspects of Finite Elements in Partial Differential Equations*, edited by C. de Boor. Academic Press, New York (1974) 89–145.

[21] J. Luo, *A priori error estimates to Runge–Kutta discontinuous Galerkin finite element method for symmetrizable system of conservation laws with sufficiently smooth solutions*. Ph.D. thesis, Nanjing University (2013).

[22] S. Osher, Riemann solvers, the entropy condition, and difference approximations. *SIAM. J. Numer. Anal.* **21** (1984) 217–235.

[23] W.H. Reed and T.R. Hill, *Triangular mesh methods for the neutron transport equation*. Los Alamos Scientific Laboratory report LA-UR-73-479, Los Alamos, NM (1973).

[24] P.L. Roe, Approximate Riemann solvers, parameter vectors, and difference schemes. *J. Comput. Phys.* **43** (1981) 357–372.

[25] C.W. Schulz-Rinne, Classification of the Riemann problem for two-dimensional gas dynamics. *SIAM J. Math. Anal.* **24** (1993) 76–88.

[26] C.-W. Shu and S. Osher, Efficient implementation of essentially non-oscillatory shock capturing schemes. *J. Comput. Phys.* **77** (1988) 439–471.

[27] E.F. Toro, Riemann solvers and numerical methods for fluid dynamics: a practical introduction. Springer (2009).

[28] Q. Zhang, Third order explicit Runge–Kutta discontinuous Galerkin method for linear conservation law with inflow boundary condition. *J. Sci. Comput.* **46** (2010) 294–313.

[29] Q. Zhang and C.-W. Shu, Error estimates to smooth solution of Runge–Kutta discontinuous Galerkin methods for scalar conservation laws. *SIAM. J. Numer. Anal.* **42** (2004) 641–666.

[30] Q. Zhang and C.-W. Shu, Error estimates to smooth solution of Runge–Kutta discontinuous Galerkin methods for symmetrizable system of conservation laws. *SIAM. J. Numer. Anal.* **44** (2006) 1702–1720.

[31] Q. Zhang and C.-W. Shu, Stability analysis and a priori error estimates to the third order explicit Runge–Kutta discontinuous Galerkin Method for scalar conservation laws. *SIAM. J. Numer. Anal.* **48** (2010) 1038–1063.

[32] Q. Zhang and C.-W. Shu, Error estimates for the third order explicit Runge–Kutta discontinuous Galerkin method for a linear hyperbolic equation in one-dimension with discontinuous initial data. *Numer. Math.* **126** (2014) 703–740.