A second-order asymptotic-preserving and positivity-preserving discontinuous Galerkin scheme for the Kerr-Debye model

Juntao Huang[1] and Chi-Wang Shu[2]

Abstract

In this paper, we develop a second-order asymptotic-preserving and positivity-preserving discontinuous Galerkin (DG) scheme for the Kerr-Debye model. By using the energy estimate and Taylor expansion first introduced by Zhang and Shu in [46], the asymptotic-preserving property of the semi-discrete DG methods is proved rigorously. In addition, we propose a class of unconditional positivity-preserving implicit-explicit (IMEX) Runge-Kutta methods for the system of ordinary differential equations arising from the semi-discretization of the Kerr-Debye model. The new IMEX Runge-Kutta methods are based on the modification of the strong-stability-preserving (SSP) implicit Runge-Kutta method and have second-order accuracy. The numerical results validate our analysis.

Keywords: Discontinuous Galerkin; positivity-preserving; asymptotic-preserving; stiff systems; Runge-Kutta methods; implicit-explicit; Kerr-Debye model

[1]Zhou Pei-Yuan Center for Applied Mathematics, Tsinghua University, Beijing 100084, China. E-mail: huangjt13@mails.tsinghua.edu.cn

[2]Division of Applied Mathematics, Brown University, Providence, RI 02912, USA. E-mail: shu@dam.brown.edu. Research supported by DOE grant DE-FG02-08ER25863 and NSF grant DMS-1418750.

1 Introduction

The propagation of electromagnetic waves in medium is described by the system of Maxwell equations. The constitutive equations characterize different property of the medium. A model for the nonlinear optical phenomena in isotropic crystal is the Kerr-Debye model. In this paper, we focus on the non-dimensional Kerr-Debye model in 1D case:

$$\partial_t D + \partial_x H = 0, \tag{1.1a}$$

$$\partial_t H + \partial_x (\frac{D}{1+\chi}) = 0, \tag{1.1b}$$

$$\partial_t \chi = \frac{1}{\varepsilon}\left(\frac{D^2}{(1+\chi)^2} - \chi\right). \tag{1.1c}$$

Here $D = D(x,t)$ is the electric displacement, $H = H(x,t)$ is the magnetic field, $\chi = \chi(x,t)$ is the third-order nonlinear susceptibility of the medium, and the positive constant ε is the non-dimensional relaxation time. For the physical background and the derivation of the model, we refer the readers to [49] and [1].

Let p denote the reciprocal function of $q(e) = e + e^3$. Formally when ε tends to 0, $\frac{D^2}{(1+\chi)^2} - \chi$ converges to 0, which is equivalent to $\chi = (p(D))^2$. Thus the limiting equation of (1.1) behaves as

$$\partial_t D + \partial_x H = 0, \tag{1.2a}$$

$$\partial_t H + \partial_x p(D) = 0, \tag{1.2b}$$

which is called the Kerr model.

First we make a brief review on the results of this model in the aspect of analysis. There have been many works discussing the validity of the approximation for general hyperbolic systems with stiff relaxation terms, see e.g. [35, 8, 45]. For this specific model, the convergence of smooth solutions for (1.1) to those of (1.2) was proved in [23, 7]. The shocks for which there exists a Kerr-Debye profile were characterized in [1].

The main difficulty in designing numerical schemes for (1.1) is the approximation of the source term when ε is very small. Asymptotic-preserving (AP) methods are proposed for

this kind of problems and have been intensively studied in different settings [31, 21, 19]. The basic idea of the AP schemes is to develop numerical schemes that preserve the asymptotic limits in the discrete setting. We refer the readers to [32] for a review of the subject. Below, to the best of our knowledge, we briefly review some existing works on the AP schemes with discontinuous Galerkin (DG) spatial discretization [13, 14]. With the technique of modified equation and asymptotic analysis, Lowrie and Morel showed the AP property of a semi-discrete DG method with piecewise-linear elements for solving linear hyperbolic systems with linear relaxation in [37]. Recently, in [18], Dumsber et al. proposed a class of finite volume schemes of arbitrary high order accuracy in time and space for hyperbolic systems of balance laws with stiff source terms by combining a high order WENO reconstruction and a space-time DG scheme, and obtained good numerical results for the stiff scalar model equation in [34] and the relaxation system in [33]. There are also many works on applying the DG methods with asymptotic property to various problems, including radiation hydrodynamics [36] and extended hydrodynamics [26, 43, 44]. More recently, Jang et al. developed a family of high order asymptotic preserving DG schemes for some discrete-velocity kinetic equations under a diffusive scaling in [29], and obtained uniform stability as well as error estimates in the linear case in [28]. We remark that rigorous proofs for the AP property are relatively rare in spite of much computational effort and many applications.

Another issue is the time integration method. To allow the step size in time much larger than ε, implicit time discretization techniques have to be used for the stiff source term. A popular class of time discretization technique is the implicit-explicit (IMEX) Runge-Kutta (RK) schemes [2, 39]. In [40], the IMEX RK method was applied to hyperbolic systems with relaxation by treating convection terms explicitly and stiff source terms implicitly. Many splitting Runge-Kutta methods could also be written as the form of IMEX RK schemes [30, 6].

For the Kerr-Debye model (1.1), an important property is that if χ is initially positive, then it remains positive for $t > 0$. The violation of this positivity-preserving property may

result in unphysical numerical solutions. There are some works on positivity-preserving property of time discretization methods but a restrictive time step of order ε is required to maintain the desired property for most of them [25, 15]. Below we comment on several works on the discretization of stiff ordinary differential equations (ODEs) which preserves the positivity of solutions unconditionally. The well-known Patankar-trick was developed to treat stiff source terms in geobiochemical models [41], and was further improved and applied in different areas [4, 5, 16, 3]. Recently, Chertock et al. developed a class of second-order semi-implicit time integration methods with steady state and sign preserving property for systems of ODEs with stiff terms [9], and successfully applied it to the shallow water equations with stiff friction term [10].

In this article, we aim to construct DG schemes with proper time discretization for the Kerr-Debye system (1.1). The schemes have several properties: (i) it has a stability constraint independent of the small parameter ε; (ii) When ε tends to 0, the scheme is consistent with the Kerr model (1.2); (iii) The positivity of χ should be preserved unconditionally. To be more precise, we rigorously prove the AP property of the semi-discrete DG scheme in the limit of $\varepsilon \to 0$. Due to the nonlinearity of the relaxation term, the semi-discrete DG scheme for (1.1) with $\varepsilon \to 0$ does not solve the degenerate equations (1.2) exactly but could be taken as a small "perturbation" of the semi-discrete DG scheme for (1.2). By using energy estimate and Taylor expansion first introduced in [46, 47] and similar idea in [27], an *a priori* error estimate is obtained and thus the consistency is proved. As to the time integration, it is well-known that the Euler backward method enjoys a nice unconditional bound-preserving property for a class of stiff ODEs with some requirements (see Proposition 4.1). However, due to the nonexistence of higher order strong-stability-preserving (SSP) implicit Runge-Kutta scheme [22], higher order schemes can not be constructed by a simple convex combination of Euler backward methods. Therefore, inspired by [9], we introduce a correction step to the Euler backward method, and develop a class of second-order modified IMEX RK methods which can preserve the positivity of χ unconditionally.

The paper is organized as follows. In section 2, we introduce some basic notations and define the semi-discrete DG method. In section 3, we prove the asymptotic-preserving property of semi-discrete DG schemes. In section 4, the new IMEX RK methods are presented. Numerical results are reported in section 5. Some concluding remarks are given in section 6.

2 Semi-discrete DG method

We first discretize the Kerr-Debye model (1.1) in space following [12]. Denote the computational domain by $I \subset \mathbb{R}$. For each partition of the interval I, $\{x_{j+\frac{1}{2}}\}_{j=0}^{N}$, we set $I_j = (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$ for $j = 1, \cdots, N$. For simplicity, we assume the partition is uniform with mesh size h. We define the finite element space:

$$V_h^k = \{v \in L^1(I) : v|_{I_j} \in \mathbb{P}^k(I_j), \quad j = 1, \cdots, N\}, \tag{2.3}$$

where $\mathbb{P}^k(I_j)$ denotes the space of polynomials in I_j of degree at most $k \geq 0$.

We are now ready to define the semi-discrete DG method for (1.1). Look for $D_h(\cdot, t)$, $H_h(\cdot, t)$, $\chi_h(\cdot, t) \in V_h^k$, such that $\forall\ \phi_h,\ \psi_h,\ \varphi_h \in V_h^k$ and $\forall\ j$,

$$\int_{I_j} \partial_t D_h \phi_h - \int_{I_j} H_h \partial_x \phi_h + \hat{H}_{j+\frac{1}{2}}(D_h, H_h, \chi_h)(\phi_h)^-_{j+\frac{1}{2}} - \hat{H}_{j-\frac{1}{2}}(D_h, H_h, \chi_h)(\phi_h)^+_{j-\frac{1}{2}} = 0, \tag{2.4a}$$

$$\int_{I_j} \partial_t H_h \psi_h - \int_{I_j} \frac{D_h}{1+\chi_h} \partial_x \psi_h + \widehat{(\frac{D}{1+\chi})}_{j+\frac{1}{2}} (D_h, H_h, \chi_h)(\psi_h)^-_{j+\frac{1}{2}} - \widehat{(\frac{D}{1+\chi})}_{j-\frac{1}{2}} (D_h, H_h, \chi_h)(\psi_h)^+_{j-\frac{1}{2}} = 0, \tag{2.4b}$$

$$\int_{I_j} \partial_t \chi_h \varphi_h = \frac{1}{\varepsilon} \int_{I_j} ((\frac{D_h}{1+\chi_h})^2 - \chi_h)\varphi_h. \tag{2.4c}$$

Here $\hat{H}_{j+\frac{1}{2}}$ and $\widehat{(\frac{D}{1+\chi})}_{j+\frac{1}{2}}$ are numerical fluxes. Throughout this paper, we will use the global Lax-Friedrichs flux:

$$\hat{H}_{j+\frac{1}{2}} = \frac{1}{2}\left(((H_h)^-_{j+\frac{1}{2}} + (H_h)^+_{j+\frac{1}{2}}) - \alpha((D_h)^+_{j+\frac{1}{2}} - (D_h)^-_{j+\frac{1}{2}})\right), \tag{2.5a}$$

$$\widehat{(\frac{D}{1+\chi})}_{j+\frac{1}{2}} = \frac{1}{2}\left(((\frac{D_h}{1+\chi_h})^-_{j+\frac{1}{2}} + (\frac{D_h}{1+\chi_h})^+_{j+\frac{1}{2}}) - \alpha((H_h)^+_{j+\frac{1}{2}} - (H_h)^-_{j+\frac{1}{2}})\right), \tag{2.5b}$$

with α the maximum of the absolute value of all the eigenvalues of the Jacobian matrix for (1.1), i.e., $\alpha = \max_\chi \frac{1}{\sqrt{1+\chi}}$.

In the computation, we use the $(k+1)$-points Gauss-Legendre quadrature rules to approximate the integral in the source term of (2.4c), and (2.4c) is replaced by:

$$\int_{I_j} \partial_t \chi_h \varphi_h = \frac{1}{\varepsilon} h \sum_{\beta=0}^{k} \hat{\omega}_\beta ((\frac{D_h(\hat{x}_{j,\beta})}{1+\chi_h(\hat{x}_{j,\beta})})^2 - \chi_h(\hat{x}_{j,\beta}))\varphi_h(\hat{x}_{j,\beta}), \tag{2.6}$$

where $\hat{\omega}_\beta$ and $\hat{x}_{j,\beta}$ denote the quadrature weights and quadrature points in the cell I_j.

3 Asymptotic-preserving property of semi-discrete DG schemes

In this section, we will rigorously prove the asymptotic-preserving property of semi-discrete DG schemes. To be more specific, we would like to prove that, as ε tends to 0, the semi-discrete methods (2.4a)-(2.4b)-(2.6) become a consistent discretization of the limiting equation (1.2). Throughout this section, we do not pay attention to boundary conditions: thus the solution is considered either periodic or compactly supported.

To explicitly indicate how the solutions depend on the small parameter ε, we denote the exact solutions to the Kerr-Debye model (1.1) by $D^\varepsilon = D^\varepsilon(x,t)$, $H^\varepsilon = H^\varepsilon(x,t)$ and $\chi^\varepsilon = \chi^\varepsilon(x,t)$. We denote the solutions to the semi-discrete DG method with source term approximated by quadrature rules (2.4a)-(2.4b)-(2.6) by D_h^ε, H_h^ε and χ_h^ε. As usual, we take the initial value of D_h^ε, H_h^ε and χ_h^ε to be the L^2-projection of $D^\varepsilon(\cdot, t=0)$, $H^\varepsilon(\cdot, t=0)$ and $\chi^\varepsilon(\cdot, t=0)$, see e.g. [46].

By letting ε formally tend to 0 in (2.6), one can obtain the equilibrium set for the semi-discrete DG schemes:

$$\sum_{\beta=0}^{k} \hat{\omega}_\beta ((\frac{D_h^0(\hat{x}_{j,\beta})}{1+\chi_h^0(\hat{x}_{j,\beta})})^2 - \chi_h^0(\hat{x}_{j,\beta}))\varphi_h(\hat{x}_{j,\beta}) = 0, \tag{3.7}$$

which is equivalent to

$$\chi_h^0 - p^2(D_h^0) = 0, \tag{3.8}$$

at all $(k+1)$ Gauss-Legendre quadrature points in all cells. For notational convenience, we define an interpolation operator π_h: for any piecewise continuous function u, $\pi_h u$ is defined as the unique function in V_h^k which satisfies $\pi_h u = u$ at all $(k+1)$ Gauss-Legendre quadrature points in all cells. Then (3.8) could be written compactly as

$$\chi_h^0(\cdot,t) = \pi_h(p^2(D_h^0(\cdot,t))). \tag{3.9}$$

Now as $\varepsilon \to 0$, the semi-discrete DG method (2.4a)-(2.4b)-(2.6) formally becomes

$$\int_{I_j} \partial_t D_h^0 \phi_h - \int_{I_j} H_h^0 \partial_x \phi_h + \hat{H}_{j+\frac{1}{2}}(D_h^0, H_h^0, \pi_h(p^2(D_h^0)))(\phi_h)^-_{j+\frac{1}{2}} - \hat{H}_{j-\frac{1}{2}}(D_h^0, H_h^0, \pi_h(p^2(D_h^0)))(\phi_h)^+_{j-\frac{1}{2}} = 0, \tag{3.10a}$$

$$\int_{I_j} \partial_t H_h^0 \psi_h - \int_{I_j} \frac{D_h^0}{1+\pi_h(p^2(D_h^0))} \partial_x \psi_h + \widehat{(\frac{D}{1+\chi})}_{j+\frac{1}{2}}(D_h^0, H_h^0, \pi_h(p^2(D_h^0)))(\psi_h)^-_{j+\frac{1}{2}} - \widehat{(\frac{D}{1+\chi})}_{j-\frac{1}{2}}(D_h^0, H_h^0, \pi_h(p^2(D_h^0)))(\psi_h)^+_{j-\frac{1}{2}} = 0. \tag{3.10b}$$

for all ϕ_h, $\psi_h \in V_h^k$ and j.

We remark that if the relaxation source term is linear, i.e., p^2 is a linear function, then $\pi_h(p^2(D_h^0)) \equiv p^2(D_h^0)$ and thus (3.10) is exactly the semi-discrete DG scheme for (1.2). Due to the nonlinearity of p^2, (3.10) does not solve (1.2) exactly. However, it could be taken as a small "perturbation" of the semi-discrete DG methods for (1.2) because the interpolation operator π_h preserves piecewise polynomials of degree $\leq k$. Therefore, this problem is much similar to the error estimate of semi-discrete DG methods with quadrature rules in [27]. Our main idea also originates from [27].

In the following part, we will estimate the error between (D_h^0, H_h^0) (the solutions to the degenerate semi-discrete method (3.10)) and (D^0, H^0) (the exact solutions to the degenerate equations (1.2)). We present the main theorem and lemmas here. Some technical details will be left in the appendix.

Theorem 3.1 (Asymptotic-preserving property of semi-discrete DG schemes)**.** *Let* (D^0, H^0) *be the exact solutions to the degenerate equations* (1.2) *which are both bounded and sufficiently*

smooth. Let (D_h^0, H_h^0) *be the solutions to the degenerate semi-discrete DG scheme* (3.10) *with piecewise polynomials of degree* $k \geq 3$*, and denote the corresponding numerical error by* $e = (D^0 - D_h^0, H^0 - H_h^0)$*. Then for small enough* h*, there holds the following error estimates:*

$$\max_{0\leq t\leq T} \|e(t,\cdot)\|_{L^2(I)} \leq Ch^k. \tag{3.11}$$

Here the positive constant C *is independent of* h *and the approximation solution* (D_h^0, H_h^0)*.*

Remark 3.1. *We point out that the error estimate of* $O(h^k)$ *is not optimal. Moreover, we would like to mention that our proof does not work for* $k = 1, 2$*. Such assumptions are purely needed for the a priori assumption. In practice, it does not seem necessary. We will report convergence of order two with piecewise linear finite element space in the numerical experiments.*

Remark 3.2. *Our proof does not rely on the specific form of the equation. It could be extended to general systems of hyperbolic conservation laws with relaxation in multi-dimensional cases.*

Remark 3.3. *The above process of deducing the limiting scheme* (3.10) *is only formal and not rigorous. The rigorous proof for the reasonability of the limit is highly nontrivial and beyond the scope of this work.*

Before starting to prove the main results for error estimates, we present some interpolation inequalities for the projections. The usual notation of norms and seminorms in Sobolev spaces will be used, see e.g. [46]. For vectors and matrices, we use the 2-norm. For the L^2 projection $\mathbb{P}_h$ and the interpolation operator π_h mentioned above, it is easy to show the following lemma (cf. [11]):

Lemma 3.1 (Interpolation inequalities)**.** *There exists a constant* C*, which does not depend on* h*, such that, for any piecewise smooth function* u*, we have*

$$\|u - \pi_h u\| + h^{\frac{1}{2}} \|u - \pi_h u\|_\Gamma \leq C \left\|\partial_x^{k+1} u\right\| h^{k+1}, \tag{3.12}$$

$$\|u - \pi_h u\|_\infty \le C \left\|\partial_x^{k+1} u\right\|_\infty h^{k+1}. \tag{3.13}$$

The same inequalities hold for π_h replaced by $\mathbb{P}_h$.

We also present some inverse properties of the finite element space V_h^k that will be used in our analysis. For more details, we refer the reader to [11].

Lemma 3.2 (Inverse inequalities)**.** *There exists a constant $C > 0$, independent of h, such that for any $v_h \in V_h^k$,*

$$\|\partial_x v_h\| \le C h^{-1} \|v_h\|, \tag{3.14a}$$

$$\|v_h\|_\Gamma \le C h^{-1/2} \|v_h\|, \tag{3.14b}$$

$$\|v_h\|_\infty \le C h^{-1/2} \|v_h\|. \tag{3.14c}$$

First, we introduce some notations and put the problem in a general form. We use the notation $u = (D, H)^\intercal$ and $v = \chi$. Now the Kerr model (1.2) is a system for the unknown function u. The physical flux of the first two equations of (1.1) is $f(u, v) = (f_1(u, v), f_2(u, v))^\intercal = (H, \frac{D}{1+\chi})^\intercal$. Denote

$$H(p, q, r) = \sum_j \int_{I_j} \partial_x r^\intercal f(p, q) + \sum_j [r]^\intercal_{j+\frac{1}{2}} \hat{f}_{j+\frac{1}{2}}(p, q) \tag{3.15}$$

Here p and r a vector-valued function with two components, q is a scalar-valued function and $\hat{f}_{j+\frac{1}{2}} = (\hat{f}_{j+\frac{1}{2},1}, \hat{f}_{j+\frac{1}{2},2})^\intercal$ is the Lax-Friedrichs flux (2.5):

$$\hat{f}_{j+\frac{1}{2},1} = \frac{1}{2}\left((f_1(u, v)^-_{j+\frac{1}{2}} + f_1(u, v)^+_{j+\frac{1}{2}}) - \alpha((u_1)^+_{j+\frac{1}{2}} - (u_1)^-_{j+\frac{1}{2}})\right), \tag{3.16a}$$

$$\hat{f}_{j+\frac{1}{2},2} = \frac{1}{2}\left((f_2(u, v)^-_{j+\frac{1}{2}} + f_2(u, v)^+_{j+\frac{1}{2}}) - \alpha((u_2)^+_{j+\frac{1}{2}} - (u_2)^-_{j+\frac{1}{2}})\right), \tag{3.16b}$$

with α the maximum of the absolute value of eigenvalue of the Jacobian matrix of (1.1). We denote the equilibrium set $\chi - (p(D))^2 = 0$ by $v = g(u)$. With periodic or zero boundary conditions, making a summation of (3.10) over j obtains

$$\sum_j \int_{I_j} \phi_h^\intercal \partial_t u_h = H(u_h, \pi_h(g(u_h)), \phi_h). \tag{3.17}$$

The exact smooth solution to the degenerate PDE (1.2) satisfies

$$\sum_j \int_{I_j} \phi_h^\intercal \partial_t u = H(u, g(u), \phi_h). \tag{3.18}$$

Here and in what follows, we omit the superscript 0 in (3.10) for notational convenience.

We would like to estimate the error $e = u - u_h$. As is customary in error analysis of finite element methods, we denote $\xi = \mathbb{P}_h u - u_h$ and $\eta = \mathbb{P}_h u - u$ with $\mathbb{P}_h$ the usual L^2-projection [46]. Subtracting (3.18) from (3.17), we obtain the energy equality:

$$\sum_j \int_{I_j} \phi_h^\intercal \partial_t \xi = H(u, g(u), \phi_h) - H(u_h, \pi_h(g(u_h)), \phi_h). \tag{3.19}$$

Due to the symmetrization theory [24], one can seek a symmetric positive definite matrix $Q = Q(u)$ such that $H = H(u) := Q(u) f'_u$ is symmetric where $f'_u \equiv \frac{\partial f(u,g(u))}{\partial u}$ denote the Jacobian matrix for (1.2). Following [38], we further define a piecewise constant matrix $Q_c = Q(u_c)$ with u_c denoting the evaluation of the exact solution at the element central points. We take $\phi_h = Q_c \xi$ in (3.19) and split the RHS as follows:

$$\begin{aligned}
\sum_j \int_{I_j} (Q_c\xi)^\intercal \partial_t \xi =& H(u, g(u), Q_c\xi) - H(u_h, \pi_h(g(u_h)), Q_c\xi), \\
=& \left(H(u, g(u), Q_c\xi) - H(u_h, g(u_h), Q_c\xi)\right) \\
&+ \left(H(u_h, g(u_h), Q_c\xi) - H(u_h, \pi_h(g(u_h)), Q_c\xi)\right), \\
=& \left(H(u, g(u), Q_c\xi) - H(u_h, g(u_h), Q_c\xi)\right) \\
&+ E(u_h, g(u_h), Q_c\xi), \\
=& \left(H(u, g(u), Q_c\xi) - H(u_h, g(u_h), Q_c\xi)\right) \\
&+ E(u, g(u), Q_c\xi) + \left(E(u_h, g(u_h), Q_c\xi) - E(u, g(u), Q_c\xi)\right), \\
\equiv& \mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3.
\end{aligned}$$

Here $E(p, q, r) := H(p, q, r) - H(p, \pi_h q, r)$.

Before going to the details of the estimate of each part, we would like to make some further assumptions on the physical flux f and the matrix Q following [38]. Each component of $f(p)$ and $f'_u(p)$ is bounded for all p. The matrix $Q(p)$ is symmetric positive definite uniformly in

the sense that there exist two constants C_1 and C_2 such that $0 < C_1 \leq \|Q(p)\| \leq C_2$ for all p. Furthermore, we assume that $f'_u(p)$ and $Q(p)$ are globally Lipschitz continuous. These assumptions are reasonable with a cut-off modification of f and Q, due to the boundedness of the exact solutions. We refer the readers to [46] and [38] for more details.

The first term $\mathcal{T}_1$ only involves the error of the semi-discrete DG method for (1.2) and has been estimated in [38]. The authors introduced a definition called "the generalized-E flux" for systems [38]. We remark that if the Lax-Friedrichs flux (3.16) is used, then the reduced flux is

$$\hat{f}_{j+\frac{1}{2},1} = \frac{1}{2}\left((f_1(u,g(u))^-_{j+\frac{1}{2}} + f_1(u,g(u))^+_{j+\frac{1}{2}}) - \alpha((u_1)^+_{j+\frac{1}{2}} - (u_1)^-_{j+\frac{1}{2}})\right), \tag{3.20a}$$

$$\hat{f}_{j+\frac{1}{2},2} = \frac{1}{2}\left((f_2(u,g(u))^-_{j+\frac{1}{2}} + f_2(u,g(u))^+_{j+\frac{1}{2}}) - \alpha((u_2)^+_{j+\frac{1}{2}} - (u_2)^-_{j+\frac{1}{2}})\right), \tag{3.20b}$$

It is also a Lax-Friedrichs flux for the degenerate systems (1.2), due to the fact that the maximum of the absolute value of eigenvalue of (1.1) is $\frac{1}{\sqrt{1+\chi}}$, which is greater than $\frac{1}{\sqrt{1+3\chi}}$, the maximum of (1.2). Hence, it satisfies the definition of the generalized-E flux in [38]. We list the results in the following lemma and refer the readers to section 5.3 in [38]:

Lemma 3.3 (Estimate of $\mathcal{T}_1$)**.** *There exist a constant $C > 0$, independent of h, such that*

$$\mathcal{T}_1 = H(u,g(u),Q_c\xi) - H(u_h,g(u_h),Q_c\xi) \leq C(1 + h^{-1}\|e\|_\infty)(\|\xi\|^2 + h^{2k+1}). \tag{3.21}$$

Lemma 3.4 (Estimate of $\mathcal{T}_2$)**.** *There exist a constant $C > 0$, independent of h, such that*

$$|\mathcal{T}_2| \leq C(h^{2k} + \|\xi\|^2). \tag{3.22}$$

Proof. We expand the second term

$$\begin{aligned}
\mathcal{T}_2 =& E(u,g(u),Q_c\xi),\\
=& H(u,g(u),Q_c\xi) - H(u,\pi_h(g(u)),Q_c\xi),\\
=& \sum_j \int_{I_j} (Q_c\xi)^{\intercal}_x (f(u,g(u)) - f(u,\pi_h(g(u))))\\
&+ \sum_j (Q_c\xi)^{\intercal}_{j+\frac{1}{2}} (\hat{f}_{j+\frac{1}{2}}(u,g(u)) - \hat{f}_{j+\frac{1}{2}}(u,\pi_h(g(u))))
\end{aligned}$$

to have the following estimate

$$\begin{aligned}|\mathcal{T}_2| \leq & C \left\|\partial_x \xi\right\| \left\|g(u) - \pi_h(g(u))\right\| + C \left\|\xi\right\|_\Gamma \left\|g(u) - \pi_h(g(u))\right\|_\Gamma, \\ \leq & C h^k \left\|\xi\right\| \leq C(h^{2k} + \left\|\xi\right\|^2).\end{aligned}$$

Here we have used Cauchy's inequality, the interpolation inequality (3.12) and the inverse inequality (3.14a)-(3.14b). □

For the estimate of the third term $\mathcal{T}_3$, we present the results below and leave the technical proof in the appendix.

Lemma 3.5 (Estimate of $\mathcal{T}_3$)**.** *There exist a constant $C > 0$, independent of h, such that*

$$|\mathcal{T}_3| \leq C(1 + h^{-2} \left\|e\right\|_\infty)(h^{2k} + \left\|\xi\right\|^2). \tag{3.23}$$

We are now ready to prove our main theorem 3.1. Following [46], we first make an *a priori* assumption that, for small enough h, there holds the inequality

$$\left\|e(\cdot, t)\right\| \leq h^{5/2}, \tag{3.24}$$

for $0 \leq t \leq T$. From Lemma 3.3, Lemma 3.4 and Lemma 3.5, an estimate is obtained based on (3.19):

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} \left\|Q_c^{\frac{1}{2}} \xi\right\|^2 = & \frac{\mathrm{d}}{\mathrm{d}t} \sum_j \int_{I_j} \xi^\intercal Q_c \xi, \\ = & 2 \sum_j \int_{I_j} (Q_c \xi)^\intercal \partial_t \xi + \sum_j \int_{I_j} \xi^\intercal \partial_t Q_c \xi, \\ \leq & C(1 + h^{-1} \left\|e\right\|_\infty)(\left\|\xi\right\|^2 + h^{2k+1}) + C(\left\|\xi\right\|^2 + h^{2k}) \\ & + C(1 + h^{-2} \left\|e\right\|_\infty)(\left\|\xi\right\|^2 + h^{2k}) + C \left\|\xi\right\|^2, \\ \leq & C(1 + h^{-2} \left\|e\right\|_\infty)(\left\|Q_c^{\frac{1}{2}} \xi\right\|^2 + h^{2k}) \leq C(\left\|Q_c^{\frac{1}{2}} \xi\right\|^2 + h^{2k}).\end{aligned}$$

Here we have used the uniform equivalence among the norms $\|\cdot\|$ and $\left\|Q_c^{\frac{1}{2}} \cdot\right\|$, due to the uniform boundedness of Q_c. Thus again by the uniform equivalence of the two norms, we finally reach the conclusion of the theorem:

$$\max_{0 \leq t \leq T} \left\|e(t, \cdot)\right\|_{L^2(I)} \leq C h^k, \tag{3.25}$$

To complete proof of this theorem, the *a priori* assumption (3.24) needs to be justified. The details are omitted here and we refer the readers to [46] and [27].

4 Positivity-preserving IMEX RK methods

First, we choose an appropriate set of basis functions for the finite element space V_h^k and write down the ODEs for the coefficients of the basis. If we use the Lagrange polynomials at Gauss-Legendre points as basis of V_h^k, i.e., for $x \in I_j$,

$$D_h(x,t) = \sum_{\beta=0}^{k} D_j^{(\beta)}(t) l_j^{(\beta)}(x),$$
$$H_h(x,t) = \sum_{\beta=0}^{k} H_j^{(\beta)}(t) l_j^{(\beta)}(x),$$
$$\chi_h(x,t) = \sum_{\beta=0}^{k} \chi_j^{(\beta)}(t) l_j^{(\beta)}(x),$$

where the Lagrange polynomials satisfy $l_j^{(\alpha)}(\hat{x}_{j,\beta}) = \delta_{\alpha\beta}$ for $\alpha, \beta = 0, \cdots, k$ and $\hat{x}_{j,\beta}$ denotes the quadrature points in I_j, then (2.6) could be rewritten as

$$\partial_t \chi_j^{(\beta)}(t) = \frac{1}{\varepsilon} \left((\frac{D_j^{(\beta)}(t)}{1+\chi_j^{(\beta)}(t)})^2 - \chi_j^{(\beta)}(t) \right), \tag{4.26}$$

for $\beta = 0, \cdots, k$ and any j. Choosing this set of basis preserves the original structure of the PDE and thus all the coefficients $\chi_j^{(\beta)}(t)$ will stay positive as long as the initial value is positive. In the following part, we focus on developing a class of positivity-preserving implicit-explicit (IMEX) Runge-Kutta (RK) methods for the ODEs of the form (4.26).

4.1 Scalar ODE

To start with the simple case, we first discuss the initial value problem for scalar ODE of the form:

$$\frac{\mathrm{d}u}{\mathrm{d}t} = L(u), \quad t > 0 \tag{4.27}$$

Here L is a stiff term which satisfies:

($\mathcal{A}1$) $L = L(u)$ is decreasing w.r.t. u;

($\mathcal{A}2$) There exists one and only one real number u^* satisfying $L(u^*) = 0$.

Under these assumptions, the exact solution of (4.27) enjoys the following property:

(i) if $u(0) < u^*$, then $u(t) < u^*$ for $t > 0$ and u is increasing w.r.t $t > 0$;

(ii) if $u(0) = u^*$, then $u(t) \equiv u^*$;

(iii) if $u(0) > u^*$, then $u(t) > u^*$ for $t > 0$ and u is decreasing w.r.t $t > 0$.

In the special case of $u^* \geq 0$, the solution $u(t)$ stays positive as long as the initial value is positive.

Thanks to the decreasing property of L, one can easily show that the numerical solution of (4.27) with the Euler backward method shares the same property:

Proposition 4.1. *Suppose that L satisfies the assumptions ($\mathcal{A}1$) and ($\mathcal{A}2$). Then the numerical solution of* (4.27) *with the Euler backward method*

$$u_{n+1} = u_n + kL(u_{n+1}) \tag{4.28}$$

has the property: for any step size $k > 0$,

(i) if $u_n < u^*$, then $u_n < u_{n+1} < u^*$;

(ii) if $u_n = u^*$, then $u_{n+1} = u^*$;

(iii) if $u_n > u^*$, then $u^* < u_{n+1} < u_n$.

A natural way of generalizing this property to higher order methods is to write the diagonally implicit Runge-Kutta method into a convex combination of the Euler backward method [22]:

$u^{(0)} = u_n$,

$$u^{(i)} = \sum_{j=0}^{i-1} \alpha_{i,j} u^{(j)} + k\beta_i L(u^{(i)}), \quad \alpha_{i,j} \geq 0, \quad \sum_{j=0}^{i-1} \alpha_{i,j} = 1, \quad \beta_i \geq 0, \quad i = 1, \cdots, m, \tag{4.29}$$
$$u_{n+1} = u^{(m)},$$

with $k > 0$ the step size. With non-negativity of the coefficients $\alpha_{i,k}$ and β_i, the implicit Runge-Kutta schemes (4.29) shares the same bound-preserving property in Proposition 4.1. We remark that here the restrictions on the coefficients are stronger than those for the strong-stability-preserving (SSP) implicit Runge-Kutta methods discussed in [22]. The SSP RK method does not require $\beta_i \geq 0$ by using a trick of solving the negative-in-time version of the conservation law. Unfortunately, the existence of (4.29) of order higher than 1 is completely ruled out even if the non-negativity of β_i is not required (cf. Proposition 6.2 in [22]).

We have to try another approach to construct high order bound-preserving implicit Runge-Kutta methods. To begin with, we analyze the Euler backward method (4.28) by using Taylor expansion:

$$u_{n+1} = u_n + kL(u_n) + k^2 L(u_n)L'(u_n) + O(k^3). \tag{4.30}$$

Clearly, it is only first-order accurate. Inspired by [9], we add one stage and compensate some second-order term after the Euler backward to enforce it to be second-order:

$$u^{(1)} = u_n + kL(u^{(1)}), \tag{4.31a}$$
$$u_{n+1} = u^{(1)} - \frac{1}{2}k^2 L(u_{n+1})L'(u^{(1)}). \tag{4.31b}$$

With the aid of decreasing property of L, it is easy to show that the second stage also enjoys the bound-preserving property in Proposition 4.1. The result is summarized in the following:

Proposition 4.2. *Suppose that L satisfies the assumptions ($\mathcal{A}1$) and ($\mathcal{A}2$). Then the numerical solution of* (4.27) *with modified implicit Runge-Kutta method* (4.31) *has the property:*

(1) It is second-order accurate;

(2) It is bound-preserving: for any step size $k > 0$,

(i) if $u_n < u^*$, then $u_n < u^{(1)} < u_{n+1} < u^*$;

(ii) if $u_n = u^*$, then $u^{(1)} = u_{n+1} = u^*$;

(iii) if $u_n > u^*$, then $u^* < u_{n+1} < u^{(1)} < u_n$.

Remark 4.1. *We can also do the modification based on the implicit Runge-Kutta method with more stages. Here the modified two stage implicit RK method is presented:*

$$u^{(1)} = u_n + ka_{11}L(u^{(1)}), \tag{4.32a}$$

$$u^{(2)} = u_n + k(a_{21}L(u^{(1)}) + a_{22}L(u^{(2)})), \tag{4.32b}$$

$$u_{n+1} = u^{(2)} - ck^2L(u_{n+1})L'(u^{(2)}), \tag{4.32c}$$

with the parameters satisfying

$$a_{21} + a_{22} = 1, \quad a_{11} \geq a_{21} \geq 0, \quad a_{22} \geq 0, \quad c = (a_{21}a_{11} + a_{22}(a_{21} + a_{22})) - \frac{1}{2} > 0. \tag{4.33}$$

It is also second-order accurate and has the bound-preserving property as in Proposition 4.2. However, it seems difficult to extend this idea to bound-preserving methods of higher order.

4.2 Coupling with non-stiff parts

In our semi-discrete scheme (2.4a)-(2.4b)-(2.6), there are also the non-stiff ODEs for D_h and H_h. Therefore, we need to generalize the modified implicit RK solver (4.31) or (4.32) to solve the systems of ODEs of the form:

$$\frac{\mathrm{d}u}{\mathrm{d}t} = f(u, v), \tag{4.34a}$$

$$\frac{\mathrm{d}v}{\mathrm{d}t} = g(u, v) = \frac{1}{\varepsilon}(N(u, v) - v), \tag{4.34b}$$

where f is a non-stiff term and g is a stiff term. For the simplicity of presentation, we start from the simple case in which u and v are both scalar-valued functions and will make a comment on the vector-valued case later. We make some assumptions on $N = N(u, v)$:

($\mathcal{B}1$) $N(u,v) \geq 0$ for any u and v;

($\mathcal{B}2$) $N = N(u,v)$ is decreasing w.r.t. v.

Under these assumptions, the exact solution of (4.34) has the property: $v(t)$ remains positive for $t > 0$ if $v(0)$ is positive.

Now we use the IMEX RK method [39] to solve (4.34), and try to keep the positivity-preserving property:

$$\begin{aligned}
u^{(1)} &= u_n, \\
v^{(1)} &= v_n + k\tilde{a}_{11} g(u^{(1)}, v^{(1)}), \\
u^{(2)} &= u_n + k a_{21} f(u^{(1)}, v^{(1)}), \\
v^{(2)} &= v_n + k\tilde{a}_{21} g(u^{(1)}, v^{(1)}) + k\tilde{a}_{22} g(u^{(2)}, v^{(2)}), \\
u_{n+1} &= u_n + k b_1 f(u^{(1)}, v^{(1)}) + k b_2 f(u^{(2)}, v^{(2)}), \\
v_{n+1} &= v_n + k\tilde{b}_1 g(u^{(1)}, v^{(1)}) + k\tilde{b}_2 g(u^{(2)}, v^{(2)}).
\end{aligned}$$

By using Taylor expansion, it is easy to obtain

$$u_{n+1} = u_n + k(b_1 + b_2) f + k^2 (b_2 a_{21} f f'_u + (b_1 \tilde{a}_{11} + b_2(\tilde{a}_{21} + \tilde{a}_{22})) g f'_v) + O(k^3),$$

and

$$v_{n+1} = v_n + k(\tilde{b}_1 + \tilde{b}_2) g + k^2 (\tilde{b}_2 a_{21} f g'_u + (\tilde{b}_1 \tilde{a}_{11} + \tilde{b}_2(\tilde{a}_{21} + \tilde{a}_{22})) g g'_v) + O(k^3)$$

where the arguments in f, g and their derivatives are all (u_n, v_n) and omitted.

Also, by using Taylor expansion on the exact solutions, we have

$$\begin{aligned}
u(t+k) &= u + k u' + \frac{1}{2} k^2 u'' + O(k^3) \\
&= u + k f + \frac{1}{2} k^2 (f f'_u + g f'_v) + O(k^3)
\end{aligned}$$

and

$$v(t+k) = v + k v' + \frac{1}{2} k^2 v'' + O(k^3)$$

$$= v + kg + \frac{1}{2}k^2(fg'_u + gg'_v) + O(k^3).$$

Now we make a summary on the restrictions on the coefficients a_{21}, $\tilde{a}_{11}$, $\tilde{a}_{21}$, $\tilde{a}_{22}$, b_1, b_2, $\tilde{b}_1$, $\tilde{b}_2$:

- (first-order accuracy for u)

$$b_1 + b_2 = 1. \tag{4.35}$$

- (first-order accuracy for v)

$$\tilde{b}_1 + \tilde{b}_2 = 1. \tag{4.36}$$

- (second-order accuracy for u)

$$b_2 a_{21} = \frac{1}{2}, \tag{4.37a}$$

$$b_1\tilde{a}_{11} + b_2(\tilde{a}_{21} + \tilde{a}_{22}) = \frac{1}{2}. \tag{4.37b}$$

- (second-order accuracy for v)

$$\tilde{b}_2 a_{21} = \frac{1}{2}, \tag{4.38a}$$

$$\tilde{b}_1\tilde{a}_{11} + \tilde{b}_2(\tilde{a}_{21} + \tilde{a}_{22}) = \frac{1}{2}. \tag{4.38b}$$

- (positivity-preserving property for v)

$$\tilde{a}_{11} \geq \tilde{a}_{21} \geq 0, \quad \tilde{a}_{22} \geq 0, \quad \tilde{a}_{21} = \tilde{b}_1 \quad \tilde{a}_{22} = \tilde{b}_2. \tag{4.39}$$

However, from the last section, we know that these restrictions (4.35)-(4.39) could not be satisfied simultaneously even if we are only limited to the stiff parts.

Now we try to drop some restrictions and make up some additional terms in the final stage to meet these abandoned conditions as we have done in the last section. The second-order condition (4.37) for u must be satisfied because compensating the derivatives terms on f will make variables in all cells couple with each other and it would cost too much to solve a large algebraic system. The restriction (4.38a) must be satisfied because we do not impose

any conditions on the sign of the non-stiff part f and thus compensating the term involving fg_u of undetermined sign will lose the positivity-preserving property. Based on the above analysis, we try to relax the restrictions (4.35) and (4.38b). In the spirit of the last section, we modify the final stage to

$$\tilde{u}_{n+1} = u_{n+1} + k(1 - b_1 - b_2)f(u_n, v_n), \tag{4.40}$$

$$\tilde{v}_{n+1} = v_{n+1} - ck^2 g(\tilde{u}_{n+1}, \tilde{v}_{n+1})g'_v(u^{(2)}, v^{(2)}), \tag{4.41}$$

with $c = \tilde{b}_1\tilde{a}_{11} + \tilde{b}_2(\tilde{a}_{21} + \tilde{a}_{22}) - \frac{1}{2} > 0$. With the aid of the assumptions ($\mathcal{B}1$) and ($\mathcal{B}2$), it is easy to show the positivity of $\tilde{v}_{n+1}$ in (4.41). Under these restrictions, the parameters are not unique and here we take one set of parameters: $a_{21} = 2$, $\tilde{a}_{11} = \frac{3}{4}$, $\tilde{a}_{21} = \frac{3}{4}$, $\tilde{a}_{22} = \frac{1}{4}$, $b_1 = \frac{1}{3}$, $b_2 = \frac{1}{4}$, $\tilde{b}_1 = \frac{3}{4}$, $\tilde{b}_2 = \frac{1}{4}$, $c = \frac{5}{16}$.

Now we make a summary of our modified IMEX RK method for (4.34):

$$u^{(1)} = u_n, \tag{4.42a}$$

$$v^{(1)} = v_n + k\tilde{a}_{11}g(u^{(1)}, v^{(1)}), \tag{4.42b}$$

$$u^{(2)} = u_n + ka_{21}f(u^{(1)}, v^{(1)}), \tag{4.42c}$$

$$v^{(2)} = v_n + k\tilde{a}_{21}g(u^{(1)}, v^{(1)}) + k\tilde{a}_{22}g(u^{(2)}, v^{(2)}), \tag{4.42d}$$

$$\hat{u}_n = u_n + kb_1 f(u^{(1)}, v^{(1)}) + kb_2 f(u^{(2)}, v^{(2)}), \tag{4.42e}$$

$$\hat{v}_n = v^{(2)}, \tag{4.42f}$$

$$u_{n+1} = \hat{u}_n + k(1 - b_1 - b_2)f(u_n, v_n), \tag{4.42g}$$

$$v_{n+1} = \hat{v}_n - ck^2 g(u_{n+1}, v_{n+1})g'_v(\hat{u}_n, \hat{v}_n) \tag{4.42h}$$

with $a_{21} = 2$, $\tilde{a}_{11} = \frac{3}{4}$, $\tilde{a}_{21} = \frac{3}{4}$, $\tilde{a}_{22} = \frac{1}{4}$, $b_1 = \frac{1}{3}$, $b_2 = \frac{1}{4}$, $c = \frac{5}{16}$.

Proposition 4.3. *Suppose that* (4.34) *satisfies the assumptions* ($\mathcal{B}1$) *and* ($\mathcal{B}2$). *Then the numerical solutions of* (4.34) *with the modified IMEX RK method* (4.42) *have the following properties:*

(1) It is second-order accurate;

(2) It is positivity-preserving: for any $k > 0$, if $v_n > 0$, then $v^{(1)}$, $v^{(2)}$, $\hat{v}_n$ and $v_{n+1} > 0$.

Remark 4.2. *Although we have only discussed the case in which u and v are scalar-valued functions, the scheme (4.42) could be also applied to systems as long as the equations for v are not coupled with each other. To be more specific, we assume that the system of ODEs is in the form:*

$$\begin{aligned} \frac{\mathrm{d}u}{\mathrm{d}t} &= f(u, v), \\ \frac{\mathrm{d}v}{\mathrm{d}t} &= g(u, v) = \frac{1}{\varepsilon}(N(u, v) - v), \end{aligned}$$

with $u = (u_1, u_2, \cdots, u_m)^\intercal$ and $v = (v_1, v_2, \cdots, v_n)^\intercal$. The function $N = (N_1, N_2, \cdots, N_n)^\intercal$ should satisfy the property that N_j only depends on u and v_j for $1 \le j \le n$, and

(i) $N_j \ge 0$;

(ii) N_j is decreasing w.r.t. v_j.

Remark 4.3. *In [9], the authors developed a class of second-order semi-implicit RK methods for the systems of ODEs of the form:*

$$u' = f(u, t) + G(u, t)u, \tag{4.43}$$

where u is an unknown vector function, f is a non-stiff term and $G(u,t)u$ is a stiff term with G a diagonal nonpositive definite matrix. The main idea is to do Taylor expansion and modify the final stage of the classical semi-implicit RK method which is at most first-order accurate and thus obtain a second-order scheme. Our idea in constructing the IMEX RK scheme is similar to theirs. Our building block is the Euler backward method while their building block is the semi-implicit RK method.

At last, we discuss the asymptotic preserving property of our modified IMEX RK solver (4.42). Under the assumptions ($\mathcal{B}1$) and ($\mathcal{B}2$), it is easy to show that, for any u, there exists a unique v such that $N(u, v) - v = 0$. We denote this equilibrium point by $v = e(u)$. Then

as $\varepsilon \to 0$, (4.34) formally degenerates to

$$\frac{\mathrm{d}u}{\mathrm{d}t} = f(u, e(u)). \tag{4.44a}$$

Following [20], we show the consistency as ε tends to 0 under some stability assumptions on the numerical solutions to (4.42). To explicitly indicate how the numerical solutions depend on ε, we add ε in the superscript. The results are listed in the following and the proof is left in the appendix.

Proposition 4.4 (consistency in the limit of $\varepsilon \to 0$ for a fixed k)**.** *Suppose that f and g in* (4.34) *are sufficiently smooth. Fix the time step $k > 0$, a final time $T > 0$, and set $N_T = [T/k]$. Assume that the numerical solutions $(u_n^\varepsilon, v_n^\varepsilon)_{0\le n\le N_T}$ given by* (4.42) *is such that for all $0 \le n \le N_T$, $(u_n^\varepsilon, \varepsilon v_n^\varepsilon)_{\varepsilon>0}$ is bounded with respect to $\varepsilon > 0$. The initial data $(u_0^\varepsilon, v_0^\varepsilon) \to (w_0, v_0)$ as $\varepsilon \to 0$ and $v_0^\varepsilon \ge 0$ for $\varepsilon > 0$. Then there exist sequences w_n, $w_n^{(1)}$, $w_n^{(2)}$ and $\hat{w}_n$, such that for $0 \le n \le N_T$, $u_n^\varepsilon \to w_n$, and for $0 \le n \le N_T - 1$, $u_n^{(1),\varepsilon} \to w_n^{(1)}$, $u_n^{(2),\varepsilon} \to w_n^{(2)}$ and $\hat{u}_n^\varepsilon \to \hat{w}_n$ as $\varepsilon \to 0$, and they satisfy the following scheme which is a consistent and first-order approximation of* (4.44)*:*

$$\begin{aligned}
w_n^{(1)} &= w_n,\\
w_n^{(2)} &= w_n + ka_{21} f(w_n^{(1)}, e(w_n^{(1)})),\\
\hat{w}_n &= w_n + kb_1 f(w_n^{(1)}, e(w_n^{(1)})) + kb_2 f(w_n^{(2)}, e(w_n^{(2)})),\\
w_{n+1} &= w_n + k(1 - b_1 - b_2) f(w_n, e(w_n))
\end{aligned}$$

for $1 \le n \le N_T - 1$. And for $n = 0$,

$$\begin{aligned}
w_0^{(1)} &= w_0,\\
w_0^{(2)} &= w_0 + ka_{21} f(w_0^{(1)}, e(w_0^{(1)})),\\
\hat{w}_0 &= w_0 + kb_1 f(w_0^{(1)}, e(w_0^{(1)})) + kb_2 f(w_0^{(2)}, e(w_0^{(2)})),\\
w_1 &= w_0 + k(1 - b_1 - b_2) f(w_0, v_0).
\end{aligned}$$

If we further assume that the initial value is consistent, i.e., $v_0 = e(w_0)$, then the limiting scheme is a second-order approximation of (4.44).

To conclude this section, we would like to mention that using this modified IMEX RK solver to solve the semi-discrete DG schemes for (1.1), it can be guaranteed that the values of the numerical solutions χ_h at all Gauss-Legendre quadrature points are positive. There is no need to use the limiter introduced by Zhang and Shu in [48].

5 Numerical examples

In this section, we perform several numerical examples to validate the accuracy of our schemes in different cases.

5.1 IMEX RK methods

We first test the accuracy of our modified IMEX RK ODE solver (4.42).

Example 5.1.1 (Accuracy test for the ODE solver)**.** Consider the following system of ODEs which satisfy the assumption ($\mathcal{B}1$) and ($\mathcal{B}2$):

$$\frac{\mathrm{d}u}{\mathrm{d}t} = -u^2 - v, \tag{5.45a}$$

$$\frac{\mathrm{d}v}{\mathrm{d}t} = \frac{1}{\varepsilon}(\frac{u^2}{v^2} - v), \tag{5.45b}$$

with two sets of initial values:

(i) (without initial layer)

$$u(0) = 1, \quad v(0) = 1; \tag{5.46}$$

(ii) (with initial layer)

$$u(0) = 2, \quad v(0) = 1. \tag{5.47}$$

We compute the numerical solutions of (5.45) with our solver (4.42) up to time $T = 1$ with $\varepsilon = 1 \times 10^2, 1 \times 10^{-2}, 1 \times 10^{-6}$. Denote the numerical solutions by u^{num} and v^{num}, and the reference solutions by u^{ref} and v^{ref}. Here the "reference solutions" to (5.45) are computed by the classical fourth-order Runge-Kutta method with small enough step size k.

Presented in Table 1 and Table 2 are the errors between numerical solutions and reference solutions with initial values (5.46) and (5.47). We observe that in both cases the convergence orders are two when $k \ll \varepsilon$. When $\varepsilon \ll k$, the convergence orders for u and v are both two if there is no initial layer. If there exists an initial layer, then the convergence orders for u and v both degenerate to one. These numerical results validate our analysis in Proposition 4.4. Moreover, in Table 1, we can see the deterioration of the accuracy in the intermediate region where $k = O(\varepsilon)$, as is already observed for many IMEX RK solvers (see e.g. [39]).

Step size k	$\lvert u^{\text{exa}}(T) - u^{\text{ref}}(T)\rvert$					
	$\varepsilon = 1 \times 10^{2}$	order	$\varepsilon = 1 \times 10^{-2}$	order	$\varepsilon = 1 \times 10^{-6}$	order
1/20	1.86e-03	-	5.46e-04	-	1.56e-04	-
1/40	4.49e-04	2.05	2.70e-04	1.01	3.98e-05	1.97
1/80	1.11e-04	2.02	6.08e-06	5.48	9.97e-06	2.00
1/160	2.74e-05	2.01	1.19e-04	-4.29	2.43e-06	2.04
1/320	6.84e-06	2.01	8.32e-05	0.52	5.36e-07	2.18
1/640	1.71e-06	2.00	3.29e-05	1.34	6.05e-08	3.15
1/1280	4.26e-07	2.00	1.01e-05	1.70	5.81e-08	0.06
1/2560	1.06e-07	2.00	2.79e-06	1.86	8.73e-08	-0.59
Step size k	$\lvert v^{\text{exa}}(T) - v^{\text{ref}}(T)\rvert$					
	$\varepsilon = 1 \times 10^{2}$	order	$\varepsilon = 1 \times 10^{-2}$	order	$\varepsilon = 1 \times 10^{-6}$	order
1/20	1.84e-05	-	2.19e-03	-	1.91e-04	-
1/40	4.45e-06	2.05	1.84e-03	0.25	4.87e-05	1.97
1/80	1.10e-06	2.02	1.28e-03	0.52	1.21e-05	2.01
1/160	2.72e-07	2.01	6.72e-04	0.93	2.86e-06	2.09
1/320	6.77e-08	2.01	2.51e-04	1.42	5.27e-07	2.44
1/640	1.69e-08	2.00	7.40e-05	1.76	5.67e-08	3.22
1/1280	4.22e-09	2.00	1.97e-05	1.91	2.03e-07	-1.84
1/2560	1.05e-09	2.00	5.03e-06	1.97	2.39e-07	-0.24

Table 1: Example 5.1.1: Errors between numerical solutions and reference solutions of u and v at time $T = 1$ for (5.45) with the initial value (5.46) (without the initial layer).

5.2 The Kerr-Debye model

In this part, we will demonstrate the performance of the proposed schemes by applying them to several numerical examples for the Kerr-Debye model. Since our ODE solver (4.42) is second-order accurate, the finite element space of piecewise linear polynomials is used here. The CFL number is taken to be 0.1, unless otherwise stated.

Step size k	$\left\|u^{\mathrm{exa}}(T)-u^{\mathrm{ref}}(T)\right\|$					
	$\varepsilon=1\times10^{2}$	order	$\varepsilon=1\times10^{-2}$	order	$\varepsilon=1\times10^{-6}$	order
1/20	2.11e-03	-	2.58e-04	-	1.31e-03	-
1/40	4.73e-04	2.16	1.33e-04	0.95	6.27e-04	1.07
1/80	1.12e-04	2.08	2.26e-04	-0.76	2.98e-04	1.07
1/160	2.73e-05	2.04	2.42e-04	-0.10	1.44e-04	1.05
1/320	6.73e-06	2.02	1.38e-04	0.81	7.09e-05	1.03
1/640	1.67e-06	2.01	5.15e-05	1.42	3.50e-05	1.02
1/1280	4.16e-07	2.00	1.55e-05	1.73	1.74e-05	1.01
1/2560	1.04e-07	2.00	4.24e-06	1.87	8.60e-06	1.01
Step size k	$\left\|v^{\mathrm{exa}}(T)-v^{\mathrm{ref}}(T)\right\|$					
	$\varepsilon=1\times10^{2}$	order	$\varepsilon=1\times10^{-2}$	order	$\varepsilon=1\times10^{-6}$	order
1/20	1.79e-05	-	1.74e-03	-	1.33e-03	-
1/40	3.95e-06	2.18	1.84e-03	-0.09	6.37e-04	1.07
1/80	9.29e-07	2.09	1.47e-03	0.33	3.03e-04	1.07
1/160	2.25e-07	2.04	8.33e-04	0.82	1.47e-04	1.05
1/320	5.54e-08	2.02	3.26e-04	1.36	7.19e-05	1.03
1/640	1.37e-08	2.01	9.86e-05	1.72	3.55e-05	1.02
1/1280	3.42e-09	2.01	2.66e-05	1.89	1.75e-05	1.02
1/2560	8.54e-10	2.00	6.84e-06	1.96	8.57e-06	1.03

Table 2: Example 5.1.1: Errors between numerical solutions and reference solutions of u and v at time $T=1$ for (5.45) with the initial value (5.47) (with the initial layer).

We first test the accuracy of our schemes for smooth solutions in the case of $h \ll \varepsilon$ and $\varepsilon \ll h$.

Example 5.2.1 (Smooth solutions, $h \ll \varepsilon$)**.** We choose the relaxation shock profiles of the form [1]:

$$D(x,t) = d(\frac{x-\sigma t}{\varepsilon}), \tag{5.48a}$$

$$H(x,t) = h(\frac{x-\sigma t}{\varepsilon}), \tag{5.48b}$$

$$\chi(x,t) = \Upsilon(\frac{x-\sigma t}{\varepsilon}), \tag{5.48c}$$

and

$$D_\pm = d(\pm\infty),$$

$$H_\pm = h(\pm\infty),$$

$$\chi_\pm = \Upsilon(\pm\infty),$$

with σ a constant. The profile is determined by the solution of an ODE [1] for d, h and Υ which we solve by the classical fourth-order Runge-Kutta method with small enough time step.

In this numerical test, we set $D_- = 2.4$, $D_+ = 1.5$, $H_- = 1$ and $\varepsilon = 1$. The four parameters D_-, D_+, H_- and ε can uniquely determine the other parameters H_-, $\chi_\pm$, σ as well as the solutions d, h and Υ (cf. [1]). The computational domain is taken to be the interval $[-10, 10]$. We denote the cell number by N and the mesh size in space by h. We compute the solutions up to $T = 1$. The profiles of numerical solutions and reference solutions are presented in Figure 1, which stay very smooth. The errors are listed in Table 3. Here we take the maximum value of three L^1-errors and L^∞-errors of D, H and χ. It is clearly observed that the designed second-order accuracy is achieved.

Example 5.2.2 (Smooth solutions without initial layer, $\varepsilon \ll h$)**.** We take a consistent initial value for (1.1):

$$D(x,0) = (\sin^4(\pi x)(1+\sin^4(\pi x))^2 + \varepsilon)^{1/2}, \tag{5.49a}$$

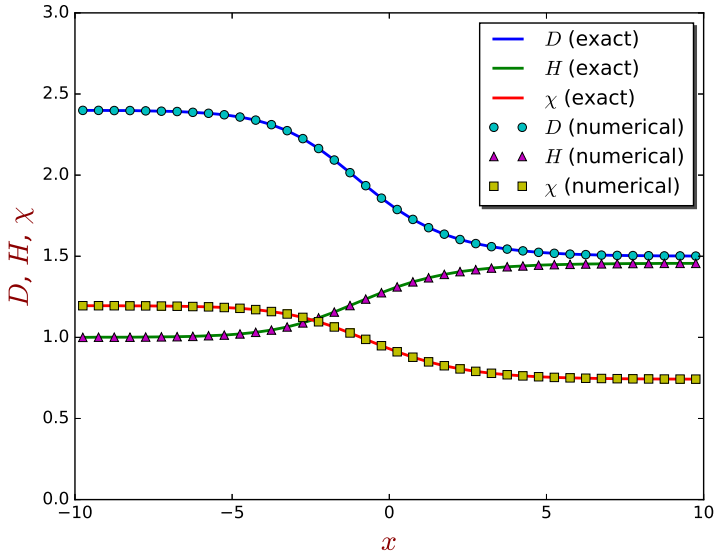


Figure 1: Example 5.2.1: Profile of solutions for the Kerr-Debye model at $T = 1$ with DG method and IMEX RK method. $\varepsilon = 1$ and $N = 20$. Solid: exact solutions; symbols: numerical solutions (cell averages).

N	L^1-error	order	L^∞-error	order
20	1.05e-02	-	1.74e-03	-
40	2.62e-03	2.01	4.17e-04	2.06
80	6.51e-04	2.01	1.01e-04	2.05
160	1.62e-04	2.00	2.50e-05	2.02
320	4.05e-05	2.00	6.19e-06	2.01

Table 3: Example 5.2.1: Error table for the Kerr-Debye model at $T = 1$ with DG method and IMEX RK method. $\varepsilon = 1$.

$$H(x,0) = \sin(\pi x), \tag{5.49b}$$

$$\chi(x,0) = \sin^4(\pi x). \tag{5.49c}$$

The computational domain is $[-1,1]$ with periodic boundary conditions. We compute the solutions up to $T = 0.3$ at which the solutions remain smooth. The reference solutions are computed by the spectral method with a fine enough mesh. In Table 4, we show the errors and orders of accuracy with $\varepsilon = 1 \times 10^{-6}$. A uniform second-order accuracy is observed.

N	L^1-error	order	L^∞-error	order
20	2.00e-02	-	6.24e-02	-
40	4.78e-03	2.06	2.12e-02	1.56
80	1.19e-03	2.01	6.17e-03	1.78
160	2.94e-04	2.01	1.67e-03	1.89
320	7.67e-05	1.94	4.50e-04	1.89
640	1.97e-05	1.96	1.15e-04	1.97

Table 4: Example 5.2.2: Error table for the Kerr-Debye model at $T = 0.3$ with consistent initial value (5.49). $\varepsilon = 1 \times 10^{-6}$.

Example 5.2.3 (Smooth solutions with initial layer, $\varepsilon \ll h$)**.** In this example, we take a non-consistent initial value:

$$D(x,0) = \sin(\pi x),$$

$$H(x,0) = \sin(\pi x),$$

$$\chi(x,0) = \sin^4(\pi x),$$

As before, the reference solutions are also computed by the spectral method. The errors and orders of accuracy are reported in Table 5 with $\varepsilon = 1 \times 10^{-6}$. Similar to the numerical example for the ODE with initial layer, only first-order accuracy is observed.

Next, a numerical example with discontinuous solutions is chosen to validate the performance of our scheme in capturing shocks.

Example 5.2.4 (Discontinuous solutions)**.** We take the discontinuous solutions of the form [1]:

$$D(x,t) = D_- \quad \text{if} \quad \xi < 0, \quad 0 \quad \text{otherwise}, \tag{5.51a}$$

$$H(x,t) = H_- \quad \text{if} \quad \xi < 0, \quad H_+ \quad \text{otherwise}, \tag{5.51b}$$

$$\chi(x,t) = \chi_- \quad \text{if} \quad \xi < 0, \quad \chi_- e^{\xi/\sigma} \quad \text{otherwise}, \tag{5.51c}$$

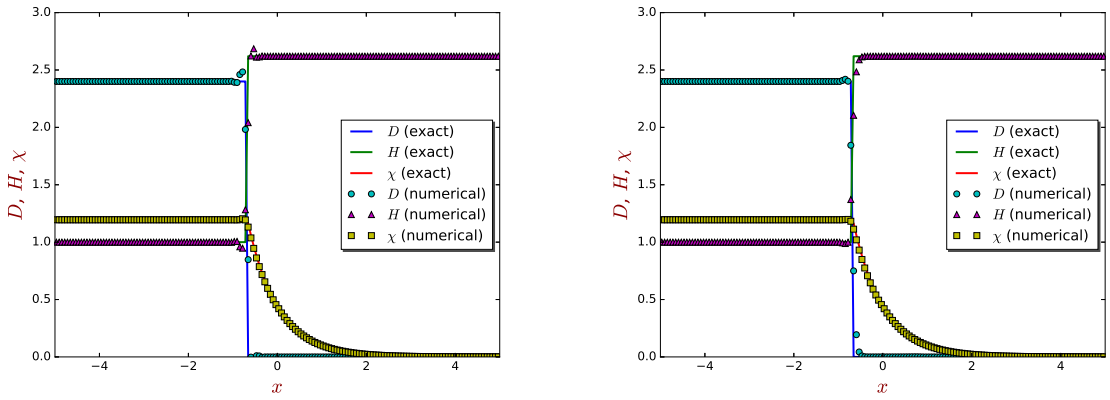
with $\xi \equiv (x - \sigma t)/\varepsilon$ and σ a constant. In the computation, we take $D_- = 2.4$, $H_- = 1$, $\varepsilon = 1$. These three parameters could uniquely determine H_+, χ_- and σ (cf. [1]). We compare the numerical behaviors without and with the TVB limiter [42] in Figure 2(a) and Figure 2(b), respectively. In both cases, the shock is captured well. However, there exists slight oscillation near the shock without the TVB limiter. We remark that our scheme is designed only to preserve the positivity of the solution, not to enforce non-oscillatory performance, similar to the philosophy of the maximum-principle-satisfying schemes for scalar conservation laws in [48].

6 Concluding remarks

In this paper, we develop a second-order asymptotic-preserving and positivity-preserving discontinuous Galerkin (DG) scheme for the Kerr-Debye model. We prove the asymptotic-preserving property of the semi-discrete DG methods rigorously. The main techniques are the energy estimate and Taylor expansion first introduced by Zhang and Shu in [46] and the idea is similar to that in the error estimate for DG methods with quadrature rules in [27]. For the time discretization, we propose a class of unconditional positivity-preserving implicit-explicit (IMEX) Runge-Kutta (RK) methods for a system of ODEs arising from the semi-discretization of the model. Inspired by [9], the new IMEX RK methods are based

N	L^1-error	order	L^∞-error	order
20	1.16e-02	-	3.75e-02	-
40	3.74e-03	1.64	1.15e-02	1.71
80	1.37e-03	1.45	3.56e-03	1.69
160	5.76e-04	1.25	1.36e-03	1.39
320	2.68e-04	1.10	5.79e-04	1.24
640	1.31e-04	1.04	2.63e-04	1.14

Table 5: Example 5.2.3: Error table for the Kerr-Debye model at $T = 0.3$ with a non-consistent initial value (5.50). $\varepsilon = 1 \times 10^{-6}$.



(a) (b)

Figure 2: Example 5.2.4: Profile of solutions for Kerr-Debye model at $T = 1$ with DG method and IMEX RK method. $\varepsilon = 1$ and $N = 160$. Solid: exact solutions; symbols: numerical solutions (cell averages). Left: no TVB limiter; Right: TVB limiter with TVB constant 0.1.

on the modification of the strong-stability-preserving (SSP) implicit RK method and have second-order accuracy. When ε tends to 0, the ODE solver is consistent with the degenerate ODE. Numerical experiments validate our analysis.

We also mention several drawbacks of this work. The AP property of the semi-discrete DG scheme is proved rigorously but the limit of $\varepsilon \to 0$ is only formal. The validity of this limit needs to be verified. Moreover, we do not analyze the AP property of the full discretization scheme but only use an AP semi-discrete scheme coupled with an AP ODE solver and numerically obtain good results. This issue need to be investigated in details. For the time discretization, the modified positivity-preserving IMEX RK solver is only second-order accurate. It seems difficult to extend this methodology to higher order. Moreover, the solver degenerates to first-order when $\varepsilon \ll k$, which is similar to Strang's splitting method [30]. New and powerful ideas need to be introduced to construct higher order and uniformly accurate ODE solver with positivity-preserving property. These issues constitute our ongoing work.

A The proof of Lemma 3.5

We expand the third term as follows:

$$\begin{aligned}\mathcal{T}_3 =& E(u_h, g(u_h), Q_c\xi) - E(u, g(u), Q_c\xi),\\ =& \sum_j \int_{I_j} (Q_c\xi)^\intercal_x (f(u_h, g(u_h)) - f(u_h, \pi_h(g(u_h))) - f(u, g(u)) + f(u, \pi_h(g(u))))\\ &+ \sum_j [Q_c\xi]^\intercal_{j+\frac{1}{2}} (\hat{f}_{j+\frac{1}{2}}(u_h, g(u_h)) - \hat{f}_{j+\frac{1}{2}}(u_h, \pi_h(g(u_h))) - \hat{f}_{j+\frac{1}{2}}(u, g(u)) + \hat{f}_{j+\frac{1}{2}}(u, \pi_h(g(u)))),\\ \equiv& W_1 + W_2,\end{aligned}$$

where W_1 denotes the integral term and W_2 denotes the interface term.

A.1 Estimate of W_1

By doing the Taylor expansion at $(u, g(u))$ for the function $f(p, q)$

$$f(p, q) = f + f'_u(p - u) + f'_v(q - g(u)) + O((p - u)^2 + (q - g(u))^2),$$

we have

$$\begin{aligned}& f(u_h, g(u_h)) - f(u_h, \pi_h(g(u_h))) - f(u, g(u)) + f(u, \pi_h(g(u)))\\ =& f'_u(u_h - u) + f'_v(g(u_h) - g(u)) - f'_u(u_h - u) - f'_v(\pi_h(g(u_h)) - g(u)) + f'_v(\pi_h(g(u)) - g(u)) + \text{H.O.T.}\\ =& f'_v(g(u_h) - g(u) - \pi_h(g(u_h) - g(u))) + \text{H.O.T.}\end{aligned}$$

with the second order term

$$\text{H.O.T.} = O((u_h - u)^2 + (g(u_h) - g(u))^2 + (\pi_h(g(u_h)) - g(u))^2 + (\pi_h(g(u)) - g(u))^2).$$

Note that here u is a vector and $u^2 \equiv u^\intercal u$ for notation convenience.

We further perform a Taylor expansion at u for the function g:

$$g(u_h) - g(u) = g'_u\eta - g'_u\xi + O(e^2), \tag{A.52}$$

and thus obtain

$$f(u_h, g(u_h)) - f(u_h, \pi_h(g(u_h))) - f(u, g(u)) + f(u, \pi_h(g(u)))$$

$$=f_v'(g_u'\eta-\pi_h(g_u'\eta))-f_v'(g_u'\xi-\pi_h(g_u'\xi))+f_v'(O(e^2)-\pi_h(O(e^2)))+\text{H.O.T.}$$

Now we have

$$\begin{aligned}W_1=&\sum_j\int_{I_j}(Q_c\xi)_x^{\intercal}f_v'(g_u'\eta-\pi_h(g_u'\eta))-\sum_j\int_{I_j}(Q_c\xi)_x^{\intercal}f_v'(g_u'\xi-\pi_h(g_u'\xi))\\&+\sum_j\int_{I_j}(Q_c\xi)_x^{\intercal}f_v'(O(e^2)-\pi_h(O(e^2)))+\sum_j\int_{I_j}(Q_c\xi)_x^{\intercal}\text{H.O.T.},\\\equiv&S_1+S_2+S_3+S_4.\end{aligned}$$

In the next, by using Lemma 3.1 and Lemma 3.2, we estimate S_1, S_2, S_3 and S_4 one by one:

$$\begin{aligned}|S_1|\leq&C\left\|\xi_x\right\|\left\|g_u'\eta-\pi_h(g_u'\eta)\right\|,\\\leq&Ch^{-1}\left\|\xi\right\|h^{k+1}\left\|\partial_x^{k+1}(g_u'\eta)\right\|,\\=&Ch^k\left\|\xi\right\|(\left\|\eta\right\|+\left\|\partial_x\eta\right\|+\cdots+\left\|\partial_x^{k+1}\eta\right\|),\\\leq&C(h^{2k}+\left\|\xi\right\|^2).\end{aligned}$$

$$\begin{aligned}|S_2|\leq&C\left\|\xi_x\right\|\left\|g_u'\xi-\pi_h(g_u'\xi)\right\|,\\\leq&Ch^{-1}\left\|\xi\right\|h^{k+1}\left\|\partial_x^{k+1}(g_u'\xi)\right\|,\\=&Ch^k\left\|\xi\right\|(\left\|\xi\right\|+\left\|\partial_x\xi\right\|+\cdots+\left\|\partial_x^{k+1}\xi\right\|),\\\leq&C\left\|\xi\right\|^2.\end{aligned}$$

$$|S_3|\leq C\left\|\xi_x\right\|\left\|e\right\|_\infty^2\leq Ch^{-3/2}\left\|e\right\|_\infty(\left\|\xi\right\|^2+h^{2k+2}).$$

$$\begin{aligned}|S_4|\leq&C\left\|\xi_x\right\|(\left\|e^2\right\|+\left\|(\pi_h(g(u_h))-g(u))^2\right\|+\left\|(\pi_h(g(u))-g(u))^2\right\|),\\\leq&Ch^{-1}\left\|\xi\right\|(\left\|e\right\|_\infty\left\|e\right\|+\left\|\pi_h(g(u_h))-g(u)\right\|_\infty^2+h^{2k+2}),\\\leq&Ch^{-1}\left\|\xi\right\|(\left\|e\right\|_\infty\left\|e\right\|+\left\|\pi_h(g(u_h)-g(u))\right\|_\infty^2+h^{2k+2}),\\\leq&Ch^{-1}\left\|\xi\right\|(\left\|e\right\|_\infty\left\|e\right\|+\left\|g(u_h)-g(u)\right\|_\infty^2+h^{2k+2}),\\\leq&Ch^{-1}\left\|\xi\right\|(\left\|e\right\|_\infty\left\|e\right\|+\left\|e\right\|_\infty^2+h^{2k+2}),\\\leq&Ch^{-3/2}\left\|e\right\|_\infty(\left\|\xi\right\|^2+h^{2k+2}).\end{aligned}$$

In the estimate of S_3 and S_4, we have used the inequality:

$$\|\pi_h u\|_\infty \leq C \|u\|_\infty . \tag{A.53}$$

A.2 Estimate of W_2

Similar to the estimate of W_1, we perform a Taylor expansion at $(u, u, g(u), g(u))$ up to second order for the numerical flux $\hat{f}_{j+\frac{1}{2}}(p^-, p^+, q^-, q^+)$. Note that we use the Lax-Friedrichs flux and thus it is smooth enough for our Taylor expansion. Since the norm of v^- or v^+ on the edge Γ will be controlled by $\|v\|_\Gamma$, for notational convenience, in the following estimate, we will not distinguish v^- or v^+ and write it as v in a uniform way. We have the estimate for W_2:

$$\begin{aligned}
|W_2| \leq & C \sum_j |[Q_c\xi]|^{\mathsf{T}}_{j+\frac{1}{2}} |g'_u\eta - \pi_h(g'_u\eta)|_{j+\frac{1}{2}} + C \sum_j |[Q_c\xi]|^{\mathsf{T}}_{j+\frac{1}{2}} |g'_u\xi - \pi_h(g'_u\xi)|_{j+\frac{1}{2}} \\
& + \sum_j |[Q_c\xi]|^{\mathsf{T}}_{j+\frac{1}{2}} \left|O(e^2) - \pi_h(O(e^2))\right|_{j+\frac{1}{2}} + \sum_j |[Q_c\xi]|^{\mathsf{T}}_{j+\frac{1}{2}} |\text{H.O.T}|_{j+\frac{1}{2}} , \\
\equiv & S_1 + S_2 + S_3 + S_4.
\end{aligned}$$

In the following, we estimate the four terms one by one:

$$\begin{aligned}
|S_1| \leq & C \|\xi\|_\Gamma \|g'_u\eta - \pi_h(g'_u\eta)\|_\Gamma , \\
\leq & Ch^{-\frac{1}{2}} \|\xi\| h^{k+\frac{1}{2}}, \\
\leq & C(\|\xi\|^2 + h^{2k}),
\end{aligned}$$

here we use the multiplicative trace inequality (cf. Lemma 3.1 in [17]):

$$\|v\|^2_\Gamma \leq C(\|v\| \|\partial_x v\| + h^{-1} \|v\|^2) \tag{A.54}$$

for $v \in H^1$. By the same approach, we obtain the estimate for S_2, S_3 and S_4:

$$\begin{aligned}
|S_2| \leq & C \|\xi\|_\Gamma \|g'_u\xi - \pi_h(g'_u\xi)\|_\Gamma \leq C \|\xi\|^2 . \\
|S_3| \leq & C \|\xi\|_\Gamma h^{-1} \|e\|^2_\infty \leq Ch^{-2} \|e\|_\infty (\|\xi\|^2 + h^{2k+2}). \\
|S_4| \leq & Ch^{-2} \|e\|_\infty (\|\xi\|^2 + h^{2k+2}).
\end{aligned}$$

Finally, we collect the above estimate about W_1 and W_2 to complete the proof of Lemma 3.5.

B Proof of Proposition 4.4

The proof of Proposition 4.4 is similar to that of Proposition 4 in [20]. First, we present two trivial lemmas and their proofs are omitted.

Lemma B.1. *Suppose that the function $N = N(u, v)$ satisfies assumptions ($\mathcal{B}1$) and ($\mathcal{B}2$). Assume the two sequences $(u^\varepsilon)_{\varepsilon>0}$ and $(v^\varepsilon)_{\varepsilon>0}$ satisfy that $(u^\varepsilon)_{\varepsilon>0}$ and $(N(u^\varepsilon, v^\varepsilon) - v^\varepsilon)_{\varepsilon>0}$ are both bounded. Then $(v^\varepsilon)_{\varepsilon>0}$ is also bounded.*

Lemma B.2. *Suppose that the function $N = N(u, v)$ satisfies assumptions ($\mathcal{B}1$) and ($\mathcal{B}2$). Assume the two sequences $(u^\varepsilon)_{\varepsilon>0}$ and $(v^\varepsilon)_{\varepsilon>0}$ satisfy*

$$N(u^\varepsilon, v^\varepsilon) - v^\varepsilon \to 0, \quad as \quad \varepsilon \to 0.$$

Then we have

$$v^\varepsilon - e(u^\varepsilon) \to 0, \quad as \quad \varepsilon \to 0.$$

To explicitly indicate how the solutions depend on the small parameter ε, we rewrite the method (4.42) in the following form with superscript ε:

$$\begin{aligned}
u_n^{(1),\varepsilon} &= u_n^\varepsilon, \\
v_n^{(1),\varepsilon} &= v_n^\varepsilon + k\tilde{a}_{11} g(u_n^{(1),\varepsilon}, v_n^{(1),\varepsilon}), \\
u_n^{(2),\varepsilon} &= u_n^\varepsilon + k a_{21} f(u_n^{(1),\varepsilon}, v_n^{(1),\varepsilon}), \\
v_n^{(2),\varepsilon} &= v_n^\varepsilon + k\tilde{a}_{21} g(u_n^{(1),\varepsilon}, v_n^{(1),\varepsilon}) + k\tilde{a}_{22} g(u_n^{(2),\varepsilon}, v^{(2),\varepsilon}), \\
\hat{u}_n^\varepsilon &= u_n^\varepsilon + k b_1 f(u_n^{(1),\varepsilon}, v_n^{(1),\varepsilon}) + k b_2 f(u_n^{(2),\varepsilon}, v_n^{(2),\varepsilon}), \\
\hat{v}_n^\varepsilon &= v_n^{(2),\varepsilon}, \\
u_{n+1}^\varepsilon &= \hat{u}_n^\varepsilon + k(1 - b_1 - b_2) f(u_n^\varepsilon, v_n^\varepsilon), \\
v_{n+1} &= \hat{v}_n - ck^2 g(u_{n+1}^\varepsilon, v_{n+1}^\varepsilon) g_v'(\hat{u}_n^\varepsilon, \hat{v}_n^\varepsilon).
\end{aligned}$$

Proof of Proposition 4.4. First, we remark that all the variables are well-defined, and $v_n^{(1),\varepsilon}$, $v_n^{(2),\varepsilon}$, $\hat{v}_n^{\varepsilon}$ are all non-negative for any $\varepsilon > 0$ and n, provided that $v_0^{\varepsilon} \geq 0$ for any $\varepsilon > 0$. Also, from the boundedness of (u_n^{ε}), one could choose a subsequence $u_n^{\varepsilon} \to w_n$ as $\varepsilon \to 0$ for $0 \leq n \leq N_T$. In the following the proof is divided into several parts for clean presentation:

1. First, it is trivial that $(u_n^{(1),\varepsilon})_{\varepsilon>0}$ is bounded for $0 \leq n \leq N_T - 1$ because of the first stage $u_n^{(1),\varepsilon} = u_n^{\varepsilon}$.

2. Next, we prove that $(\varepsilon v_n^{(1),\varepsilon})_{\varepsilon>0}$ is bounded for all n by contradiction. If this conclusion does not hold, as $\varepsilon v_n^{(1),\varepsilon} \geq 0$ thus has lower bound, then $\forall\ M > 0$, $\exists\ \varepsilon_0 > 0$ and $n_0 > 0$ such that $\varepsilon_0 v_{n_0}^{(1),\varepsilon_0} > \max(M, 1)$. Hence from the second stage, we have

$$\begin{aligned} \varepsilon_0 v_{n_0}^{\varepsilon_0} &= \varepsilon_0 v_{n_0}^{(1),\varepsilon_0} - k\tilde{a}_{11}(N(u_{n_0}^{(1),\varepsilon_0}, v_{n_0}^{(1),\varepsilon_0}) - v_{n_0}^{(1),\varepsilon_0}), \\ &\geq M - k\tilde{a}_{11}(N(u_{n_0}^{(1),\varepsilon_0}, 1) - M), \end{aligned}$$

   which is in contradiction with the boundedness of $(\varepsilon v_n^{\varepsilon})$ and $(u_n^{(1),\varepsilon})$.

3. Also from the second stage

$$\varepsilon v_n^{(1),\varepsilon} = \varepsilon v_n^{\varepsilon} + k\tilde{a}_{11}(N(u_n^{(1),\varepsilon}, v_n^{(1),\varepsilon}) - v_n^{(1),\varepsilon}),$$

   we have $(N(u_n^{(1),\varepsilon}, v_n^{(1),\varepsilon}) - v_n^{(1),\varepsilon})_{\varepsilon>0}$ is bounded. By using Lemma B.1, we have $(v_n^{(1),\varepsilon})_{\varepsilon>0}$ is bounded, and thus $\varepsilon v_n^{(1),\varepsilon} \to 0$ as $\varepsilon \to 0$. Then $N(u_n^{(1),\varepsilon}, v_n^{(1),\varepsilon}) - v_n^{(1),\varepsilon} \to 0$, and it immediately follows that $v_n^{(1),\varepsilon} - e(u_n^{(1),\varepsilon}) \to 0$ by using Lemma B.2.

4. Thanks to the smoothness of f, we can deduce that $(u_n^{(2),\varepsilon})$ is bounded from the third stage.

5. With a similar approach, we could prove $(\varepsilon v_n^{(2),\varepsilon})_{\varepsilon>0}$ is bounded. And it follows that $(N(u_n^{(2),\varepsilon}, v_n^{(2),\varepsilon}) - v_n^{(2),\varepsilon})$ is bounded, $v_n^{(2),\varepsilon}$ is bounded, $\varepsilon v_n^{(2),\varepsilon} \to 0$. Then $(N(u_n^{(2),\varepsilon}, v_n^{(2),\varepsilon}) - v_n^{(2),\varepsilon}) \to 0$, $v_n^{(2),\varepsilon} - e(u_n^{(2),\varepsilon}) \to 0$. Now one can show that $(\hat{u}_n^{\varepsilon})$ and $(\hat{v}_n^{\varepsilon})$ are bounded, because of the smoothness of f.

6. At last, let $\varepsilon \to 0$ in the final stage,

$$\varepsilon^2 v_{n+1}^{\varepsilon} + c(N(u_{n+1}^{\varepsilon}, v_{n+1}^{\varepsilon}) - v_{n+1}^{\varepsilon})(N_v'(u_n^{(2),\varepsilon}, v_n^{(2),\varepsilon}) - 1)k^2 = \varepsilon^2 \hat{v}_n^{\varepsilon}.$$

Noticing that $N_v'(u_n^{(2),\varepsilon}, v_n^{(2),\varepsilon}) - 1 \leq -1$, we have $(N(u_{n+1}^{\varepsilon}, v_{n+1}^{\varepsilon}) - v_{n+1}^{\varepsilon}) \to 0$. It follows that $v_n^{\varepsilon} - e(u_n^{\varepsilon}) \to 0$ for $1 \leq n \leq N_T$.

7. From the third stage, and $v_n^{(1),\varepsilon} - e(u_n^{(1),\varepsilon}) \to 0$, we know that as $\varepsilon \to 0$, there exists a limit of $u_n^{(2),\varepsilon}$ denoted by $w_n^{(2)}$, which satisfies

$$w_n^{(2)} = w_n + ka_{21} f(w_n^{(1)}, e(w_n^{(1)})).$$

Also in the fifth stage, $\hat{w}_n$ is the limit of $\hat{u}_n^{\varepsilon}$ and satisfies

$$\hat{w}_n = w_n + kb_1 f(w_n^{(1)}, e(w_n^{(1)})) + kb_2 f(w_n^{(2)}, e(w_n^{(2)})).$$

In the final stage, $v_n^{\varepsilon} - e(u_n^{\varepsilon}) \to 0$ holds for $1 \leq n \leq N_T$. Hence, we get

$$w_{n+1} = w_n + k(1 - b_1 - b_2) f(w_n, e(w_n)),$$

for $1 \leq n \leq N_T - 1$, and

$$w_1 = w_0 + k(1 - b_1 - b_2) f(w_0, v_0).$$

8. Collecting the above proof, we have, for $1 \leq n \leq N_T - 1$,

$$\begin{aligned}
w_n^{(1)} &= w_n, \\
w_n^{(2)} &= w_n + ka_{21} f(w_n^{(1)}, e(w_n^{(1)})), \\
\hat{w}_n &= w_n + kb_1 f(w_n^{(1)}, e(w_n^{(1)})) + kb_2 f(w_n^{(2)}, e(w_n^{(2)})), \\
w_{n+1} &= w_n + k(1 - b_1 - b_2) f(w_n, e(w_n)),
\end{aligned}$$

and for $n = 0$,

$$w_0^{(1)} = w_0,$$

$$
\begin{aligned}
w_0^{(2)} &= w_0 + k a_{21} f(w_0^{(1)}, e(w_0^{(1)})), \\
\hat{w}_0 &= w_0 + k b_1 f(w_0^{(1)}, e(w_0^{(1)})) + k b_2 f(w_0^{(2)}, e(w_0^{(2)})), \\
w_1 &= w_0 + k(1 - b_1 - b_2) f(w_0, v_0).
\end{aligned}
$$

Since w_n, $w_n^{(1)}$, $w_n^{(2)}$ are uniquely determined, all the sequences (u_n^{ε}), $(u_n^{(1),\varepsilon})$ and $(u_n^{(2),\varepsilon})$ converge.

□

References

[1] D. Aregba-Driollet and B. Hanouzet. Kerr-Debye relaxation shock profiles for Kerr equations. *Communications in Mathematical Sciences*, 9(1):1–31, 2011.

[2] U. M. Ascher, S. J. Ruuth, and R. J. Spiteri. Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations. *Applied Numerical Mathematics*, 25(2):151–167, 1997.

[3] J. Bruggeman, H. Burchard, B. W. Kooi, and B. Sommeijer. A second-order, unconditionally positive, mass-conserving integration scheme for biochemical systems. *Applied Numerical Mathematics*, 57(1):36–58, 2007.

[4] H. Burchard, E. Deleersnijder, and A. Meister. A high-order conservative Patankar-type discretisation for stiff systems of production–destruction equations. *Applied Numerical Mathematics*, 47(1):1–30, 2003.

[5] H. Burchard, E. Deleersnijder, and A. Meister. Application of modified Patankar schemes to stiff biogeochemical models for the water column. *Ocean Dynamics*, 55(3-4):326–337, 2005.

[6] R. E. Caflisch, S. Jin, and G. Russo. Uniformly accurate schemes for hyperbolic systems with relaxation. *SIAM Journal on Numerical Analysis*, 34(1):246–281, 1997.

[7] G. Carbou and B. Hanouzet. Relaxation approximation of the Kerr model for the three-dimensional initial-boundary value problem. *Journal of Hyperbolic Differential Equations*, 6(03):577–614, 2009.

[8] G.-Q. Chen, C. D. Levermore, and T.-P. Liu. Hyperbolic conservation laws with stiff relaxation terms and entropy. *Communications on Pure and Applied Mathematics*, 47(6):787–830, 1994.

[9] A. Chertock, S. Cui, A. Kurganov, and T. Wu. Steady state and sign preserving semi-implicit Runge–Kutta methods for ODEs with stiff damping term. *SIAM Journal on Numerical Analysis*, 53(4):2008–2029, 2015.

[10] A. Chertock, S. Cui, A. Kurganov, and T. Wu. Well-balanced positivity preserving central-upwind scheme for the shallow water system with friction terms. *International Journal for Numerical Methods in Fluids*, 78(6):355–383, 2015.

[11] P. G. Ciarlet. The Finite Element Method for Elliptic Problems. 1978.

[12] B. Cockburn, S. Hou, and C.-W. Shu. The Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. IV. The multidimensional case. *Mathematics of Computation*, 54(190):545–581, 1990.

[13] B. Cockburn and C.-W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. II. General framework. *Mathematics of Computation*, 52(186):411–435, 1989.

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

[15] G. Dimarco and L. Pareschi. Asymptotic preserving implicit-explicit Runge–Kutta methods for nonlinear kinetic equations. *SIAM Journal on Numerical Analysis*, 51(2):1064–1087, 2013.

[16] D. T. Dimitrov and H. V. Kojouharov. Positive and elementary stable nonstandard numerical methods with applications to predator–prey models. *Journal of Computational and Applied Mathematics*, 189(1):98–108, 2006.

[17] V. Dolejší, M. Feistauer, and C. Schwab. A finite volume discontinuous Galerkin scheme for nonlinear convection–diffusion problems. *Calcolo*, 39(1):1–40, 2002.

[18] M. Dumbser, C. Enaux, and E. F. Toro. Finite volume schemes of very high order of accuracy for stiff hyperbolic balance laws. *Journal of Computational Physics*, 227(8):3971–4001, 2008.

[19] F. Filbet and S. Jin. A class of asymptotic-preserving schemes for kinetic equations and related problems with stiff sources. *Journal of Computational Physics*, 229(20):7625–7648, 2010.

[20] F. Filbet and L. Rodrigues. Asymptotically stable particle-in-cell methods for the Vlasov–Poisson system with a strong external magnetic field. *SIAM Journal on Numerical Analysis*, 54(2):1120–1146, 2016.

[21] L. Gosse and G. Toscani. An asymptotic-preserving well-balanced scheme for the hyperbolic heat equations. *Comptes Rendus Mathematique*, 334(4):337–342, 2002.

[22] S. Gottlieb, C.-W. Shu, and E. Tadmor. Strong stability-preserving high-order time discretization methods. *SIAM Review*, 43(1):89–112, 2001.

[23] B. Hanouzet and P. Huynh. Approximation par relaxation d'un système de Maxwell non linéaire. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, 330(3):193–198, 2000.

[24] A. Harten. On the symmetric form of systems of conservation laws with entropy. *Journal of Computational Physics*, 49(1):151–164, 1983.

[25] I. Higueras and T. Roldán. Positivity-preserving and entropy-decaying IMEX methods. In *Ninth International Conference Zaragoza-Pau on Applied Mathematics and Statistics. Monogr. Semin. Mat. Garcia Galdeano*, volume 33, pages 129–136, 2006.

[26] J. A. Hittinger, Y. Suzuki, and B. Van Leer. Investigation of the discontinuous Galerkin method for first-order PDE approaches to CFD. *AIAA Paper*, 4989, 2005.

[27] J. Huang and C.-W. Shu. Error estimates to smooth solutions of semi-discrete discontinuous Galerkin methods with quadrature rules for scalar conservation laws. *Numerical Methods for Partial Differential Equations*. In press.

[28] J. Jang, F. Li, J.-M. Qiu, and T. Xiong. Analysis of asymptotic preserving DG-IMEX schemes for linear kinetic transport equations in a diffusive scaling. *SIAM Journal on Numerical Analysis*, 52(4):2048–2072, 2014.

[29] J. Jang, F. Li, J.-M. Qiu, and T. Xiong. High order asymptotic preserving DG-IMEX schemes for discrete-velocity kinetic equations in a diffusive scaling. *Journal of Computational Physics*, 281:199–224, 2015.

[30] S. Jin. Runge-Kutta methods for hyperbolic conservation laws with stiff relaxation terms. *Journal of Computational Physics*, 122(1):51–67, 1995.

[31] S. Jin. Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations. *SIAM Journal on Scientific Computing*, 21(2):441–454, 1999.

[32] S. Jin. Asymptotic preserving (AP) schemes for multiscale kinetic and hyperbolic equations: a review. *Lecture Notes for Summer School on "Methods and Models of Kinetic Theory" (M&MKT), Porto Ercole (Grosseto, Italy)*, pages 177–216, 2010.

[33] S. Jin and Z. Xin. The relaxation schemes for systems of conservation laws in arbitrary space dimensions. *Communications on Pure and Applied Mathematics*, 48(3):235–276, 1995.

[34] R. J. LeVeque and H. C. Yee. A study of numerical methods for hyperbolic conservation laws with stiff source terms. *Journal of Computational Physics*, 86(1):187–210, 1990.

[35] T.-P. Liu. Hyperbolic conservation laws with relaxation. *Communications in Mathematical Physics*, 108(1):153–175, 1987.

[36] R. B. Lowrie and J. E. Morel. Discontinuous Galerkin for hyperbolic systems with stiff relaxation. In *Discontinuous Galerkin Methods*, pages 385–390. Springer, 2000.

[37] R. B. Lowrie and J. E. Morel. Methods for hyperbolic systems with stiff relaxation. *International Journal for Numerical Methods in Fluids*, 40(3-4):413–423, 2002.

[38] J. Luo, C.-W. Shu, and Q. Zhang. A priori error estimates to smooth solutions of the third order Runge-Kutta discontinuous Galerkin method for symmetrizable systems of conservation laws. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(4):991–1018, 2015.

[39] L. Pareschi and G. Russo. Implicit-explicit Runge-Kutta schemes for stiff systems of differential equations. *Recent Trends in Numerical Analysis*, 3:269–289, 2000.

[40] L. Pareschi and G. Russo. Implicit-explicit Runge-Kutta schemes and applications to hyperbolic systems with relaxation. *Journal of Scientific Computing*, 25(1-2):129–155, 2005.

[41] S. V. Patankar. *Numerical Heat Transfer and Fluid Flow*. McGraw-Hill, 1980.

[42] C.-W. Shu. TVB uniformly high-order schemes for conservation laws. *Mathematics of Computation*, 49(179):105–121, 1987.

[43] Y. Suzuki. *Discontinuous Galerkin methods for Extended Hydrodynamics*. ProQuest, 2008.

[44] Y. Suzuki and B. van Leer. A discontinuous Galerkin method with Hancock-type time integration for hyperbolic systems with stiff relaxation source terms. In *Computational Fluid Dynamics 2006*, pages 59–64. Springer, 2009.

[45] W.-A. Yong. Singular perturbations of first-order hyperbolic systems with stiff source terms. *Journal of Differential Equations*, 155(1):89–132, 1999.

[46] Q. Zhang and C.-W. Shu. Error estimates to smooth solutions of Runge–Kutta discontinuous Galerkin methods for scalar conservation laws. *SIAM Journal on Numerical Analysis*, 42(2):641–666, 2004.

[47] Q. Zhang and C.-W. Shu. Error estimates to smooth solutions of Runge-Kutta discontinuous Galerkin method for symmetrizable systems of conservation laws. *SIAM Journal on Numerical Analysis*, 44(4):1703–1720, 2006.

[48] X. Zhang and C.-W. Shu. On maximum-principle-satisfying high order schemes for scalar conservation laws. *Journal of Computational Physics*, 229(9):3091–3120, 2010.

[49] R. W. Ziolkowski and J. B. Judkins. Full-wave vector Maxwell equation modeling of the self-focusing of ultrashort optical pulses in a nonlinear Kerr medium exhibiting a finite response time. *Journal of the Optical Society of America B*, 10(2):186–198, 1993.