Asymptotic-preserving and positivity-preserving implicit-explicit schemes for the stiff BGK equation*

Jingwei Hu† Ruiwen Shu‡ and Xiangxiong Zhang§

August 21, 2017

Abstract

We develop a family of second-order implicit-explicit (IMEX) schemes for the stiff BGK kinetic equation. The method is asymptotic-preserving (can capture the Euler limit without numerically resolving the small Knudsen number) as well as positivity-preserving — a feature that is not possessed by any of the existing second or high order IMEX schemes. The method is based on the usual IMEX Runge-Kutta framework plus a key correction step utilizing the special structure of the BGK operator. Formal analysis is presented to demonstrate the property of the method and is supported by various numerical results. Moreover, we show that the method satisfies an entropy-decay property when coupled with suitable spatial discretizations. Additionally, we discuss the generalization of the method to some hyperbolic relaxation system and provide a strategy to extend the method to third order.

Key words. Stiff kinetic equation, BGK model, compressible Euler equations, implicit-explicit Runge-Kutta (IMEX-RK) scheme, asymptotic-preserving scheme, positivity-preserving scheme.

AMS subject classifications. 82C40, 65L04, 35Q31, 65L06.

1 Introduction

Kinetic equations describe the non-equilibrium dynamics of a gas or any system comprised of a large number of particles. Compared to macroscopic fluid/continuum equations, they provide information at the mesoscopic scale using a probability density function (PDF). Kinetic equations often contain complicated integral operators modeling particle collisions (for example, the Boltzmann equation [7, 29]). To simplify the analysis and computation, the so-called Bhatnagar-Gross-Krook (BGK) model [3], or its variants, has been widely used in many disciplines of science and engineering (cf. [8, 21, 24]). After nondimensionalization, the equation reads

$$\partial_t f + v \cdot \nabla_x f = \frac{1}{\varepsilon} Q(f), \quad t \geq 0, \quad v \in \mathbb{R}^{d_v}, \quad x \in \Omega \subset \mathbb{R}^{d_x}, \tag{1.1}$$

*J. Hu and R. Shu's research was supported by NSF grant DMS-1620250 and NSF CAREER grant DMS-1654152. Support from DMS-1107291: RNMS KI-Net is also gratefully acknowledged. X. Zhang's research was supported by NSF grant DMS-1522593.

†Department of Mathematics, Purdue University, West Lafayette, IN 47907, USA (jingweihu@purdue.edu).

‡Department of Mathematics, University of Wisconsin-Madison, Madison, WI 53706, USA (rshu2@wisc.edu).

§Department of Mathematics, Purdue University, West Lafayette, IN 47907, USA (zhan1966@purdue.edu).

where $f = f(t, x, v)$ is the one-particle PDF (t is time, x is space, and v is velocity). ε is the Knudsen number which is the ratio of the mean free path and typical length scale. The collision operator Q is a relaxation type:

$$Q(f) = \tau_f(M[f] - f), \tag{1.2}$$

here M is the Maxwellian, or local equilibrium, defined as

$$M[f] = \frac{\rho}{(2\pi T)^{\frac{d_v}{2}}} \exp\left(-\frac{|v-u|^2}{2T}\right), \tag{1.3}$$

where ρ, u and T are density, bulk velocity, and temperature given by the moments of f:

$$\rho = \int_{\mathbb{R}^{d_v}} f \, \mathrm{d}v, \quad u = \frac{1}{\rho}\int_{\mathbb{R}^{d_v}} fv \, \mathrm{d}v, \quad T = \frac{1}{d_v\rho}\int_{\mathbb{R}^{d_v}} f|v-u|^2 \, \mathrm{d}v. \tag{1.4}$$

Finally τ_f is some positive function that depends only on the macroscopic quantities such as ρ and T.

It can be easily shown that the BGK operator (1.2) satisfies similar properties as the full Boltzmann collision operator:

- conservation:
$$\int_{\mathbb{R}^{d_v}} Q(f)\phi(v) \, \mathrm{d}v = 0, \quad \phi(v) = (1, v, |v|^2/2)^T; \tag{1.5}$$
- H-theorem:
$$\int_{\mathbb{R}^{d_v}} Q(f) \ln f \, \mathrm{d}v \le 0. \tag{1.6}$$

Moreover, using the Chapman-Enskog expansion, one can derive the compressible Euler equations as the leading order asymptotics of the BGK model [2]. A simple way to see this is to let $\varepsilon \to 0$ in (1.1), then formally $f \to M[f]$. On the other hand, taking the moments $\langle \cdot \phi \rangle := \int_{\mathbb{R}^{d_v}} \cdot \phi(v) \, \mathrm{d}v$ on both sides of (1.1), one obtains (using (1.5)):

$$\partial_t \langle f\phi \rangle + \nabla_x \cdot \langle fv\phi \rangle = 0. \tag{1.7}$$

Replacing f by $M[f]$ in (1.7) thus yields the compressible Euler equations:

$$\begin{cases} \partial_t \rho + \nabla_x \cdot (\rho u) = 0, \\ \partial_t (\rho u) + \nabla_x \cdot (\rho u \otimes u + pI) = 0, \\ \partial_t E + \nabla_x \cdot ((E+p)u) = 0, \end{cases} \tag{1.8}$$

where $p = \rho T$ is pressure and $E = \frac{d_v}{2}\rho T + \frac{1}{2}\rho u^2$ is total energy.

When ε is small (the system is close to the Euler limit), the right hand side of (1.1) presents strong stiffness. Hence explicit numerical schemes would impose very restrictive time step, i.e., Δt has to be $O(\varepsilon)$. To remove this constraint, implicit-explicit (IMEX) Runge-Kutta (RK) schemes are natural and popular high order methods, in which the stiff collision part is solved implicitly and the non-stiff convection part is treated explicitly [26, 11] (for IMEX-RK schemes applied to other problems, see, e.g., [1, 22, 25, 4]). As a result, the time step can be chosen independently of ε and is determined by the non-stiff part only. Furthermore, it can be shown that (see [11] for details) for fixed Δt and suitable initial conditions, as $\varepsilon \to 0$, the numerical

scheme becomes an explicit RK scheme applied to the limiting Euler equations, i.e., asymptotic-preserving (AP) [20, 18].

AP property is a desired property for handling multiscale kinetic equations, for it guarantees to capture the correct fluid limit without resolving ε. Nevertheless, the implicit treatment of the collision term would usually cause the numerical solution to lose positivity, which is unphysical since f is a PDF. We point out that the first-order IMEX scheme is an exception whose positivity can be easily achieved. Indeed, applying a forward-backward Euler scheme to (1.1) gives

$$\frac{f^{n+1}-f^n}{\Delta t}+v\cdot\nabla_x f^n=\frac{\tau_{f^{n+1}}}{\varepsilon}(M[f^{n+1}]-f^{n+1}), \tag{1.9}$$

which is equivalent to

$$f^{n+1}=\frac{\varepsilon}{\varepsilon+\Delta t\tau_{f^{n+1}}}(f^n-\Delta t v\cdot\nabla_x f^n)+\frac{\Delta t\tau_{f^{n+1}}}{\varepsilon+\Delta t\tau_{f^{n+1}}}M[f^{n+1}]. \tag{1.10}$$

Therefore, if f^n is non-negative, f^{n+1} is non-negative provided a positivity-preserving spatial discretization is used for the convection term. The situation becomes, however, highly non-trivial for the method beyond first order. The positivity of the IMEX-RK schemes is closely related to the monotonicity property (also known as strong stability [13]) of the method. In [17, 16], it was found that for the Broadwell model (a hyperbolic relaxation system, see Section 4), in order to preserve monotonicity or positivity, a sufficient condition requires the time step to be proportional to ε. This suggests that it may be very difficult to achieve the AP property, which requires Δt to be independent of ε, and positivity simultaneously. Another evidence is, even for the spatially homogeneous problem (no convection term in (1.1) and the IMEX scheme reduces to a fully implicit one), the construction of implicit positive RK scheme is still not straightforward. In fact, as proved in [14], there does not exist unconditionally strong stability preserving (SSP) implicit RK schemes of order higher than one.

Recently, a class of second-order semi-implicit RK schemes was proposed for the ODEs with stiff damping term [9]. The method is based on the modification of the explicit SSP RK schemes and is shown to be well-balanced as well as sign-preserving. Later, a second-order AP discontinuous Galerkin scheme was introduced in [19] for the Kerr-Debye model (a special relaxation system). The method is based on the modification of an IMEX-RK scheme and can preserve the positivity of one component of the solution vector. Inspired by these work, we propose to add a correction step to the standard IMEX-RK scheme. Due to the special structure of the BGK operator, this step can maintain both positivity and AP property. To insure second-order accuracy and overall positivity of the scheme, new conditions including both equalities and inequalities are derived for the RK coefficients. We then construct two IMEX-RK schemes fulfilling these conditions, one of type A and one of type ARS (two commonly used forms of IMEX-RK schemes, see Section 2.2 for definitions).

To summarize, we develop a new IMEX time discretization method for the BGK equation (1.1) that has the following feature:

- the scheme is second-order accurate for $\varepsilon=O(1)$;
- the scheme is AP: for fixed Δt, as $\varepsilon\to 0$, it reduces to a second-order scheme for the limiting Euler system (1.8);

- the scheme is positivity-preserving: if $f^n \geq 0$, then $f^{n+1} \geq 0$.

Note that the AP property implies that the time step is independent of ε. In fact, the CFL condition for the new method can be made comparable to that of the first-order scheme (1.9). We also provide a strategy to extend the method to third order. Furthermore, we show that the method satisfies an entropy-decay property when coupled with suitable spatial discretizations, and that it is possible to generalize it to some hyperbolic relaxation system which demands positivity.

The rest of this paper is organized as follows. In Section 2, we introduce a general problem and present the procedure to construct the new IMEX schemes, where the main focus is to achieve second-order accuracy as well as positivity. In Section 3, we apply the new method to the BGK equation and show that it is AP and entropy-decaying. To insure the fully discretized scheme is positivity-preserving and AP, special attention needs to be paid for spatial and velocity domain discretizations. These are described in Section 3.3. In Section 4, we briefly discuss the generalization of the method to the hyperbolic relaxation system. In Section 5, we perform several tests for the BGK equation and demonstrate numerically the properties of the proposed method. The paper is concluded in Section 6. Extension of the method to third order is provided in Appendix.

2 New IMEX-RK schemes

We now present the procedure of constructing the new IMEX schemes that are both AP and positivity-preserving. Although we mainly consider the BGK equation (1.1), the framework is quite general and can be applied to other problems that share a similar structure. Therefore, we will start with a general setting and derive conditions for the RK coefficients to insure accuracy and positivity, and will get back to the BGK model in Section 3 when discussing the AP property as this latter part is problem dependent.

2.1 A general problem and basic assumptions

Consider an ODE of the form:

$$\frac{\mathrm{d}}{\mathrm{d}t} f = \mathcal{T}(f) + \frac{1}{\varepsilon}\mathcal{Q}(f), \tag{2.1}$$

where $f = f(t)$ lies in some function space, $\mathcal{T}$ and $\mathcal{Q}$ are some operators. The equation (2.1) may arise from semi-discretizations of time-dependent PDEs by the method of lines.

We assume the terms $\mathcal{T}(f)$ and $\mathcal{Q}(f)$ are positivity-preserving. To be precise, we assume

$$f \geq 0 \implies f + a\Delta t \mathcal{T}(f) \geq 0, \quad \forall\, 0 \leq a\Delta t \leq \mathcal{C}, \tag{2.2}$$

where $\mathcal{C}$ is the Courant-Friedrichs-Lewy (CFL) type constraint for positivity. If $\mathcal{T} = \mathcal{T}_{\Delta x}$ is a discretized transport operator, then $\mathcal{C} = \Delta t_{\mathrm{FE}}$ with Δt_{FE} being the maximum time step allowance such that the forward Euler scheme is positivity-preserving. For operator $\mathcal{Q}$, we assume

$$g \geq 0, \quad (I - b\mathcal{Q})f = g \implies f \geq 0, \quad \forall\, b \geq 0. \tag{2.3}$$

We also assume a similar property for $\mathcal{Q}'(g)\mathcal{Q}(f)$ and $\mathcal{Q}'(f)\mathcal{Q}(f)$:

$$g,\ h \geq 0, \quad (I + b\mathcal{Q}'(g)\mathcal{Q})f = h \implies f \geq 0, \quad \forall\, b \geq 0, \tag{2.4}$$

$$h \geq 0, \quad (I + b\mathcal{Q}'(f)\mathcal{Q})f = h \implies f \geq 0, \quad \forall\, b \geq 0, \tag{2.5}$$

where $\mathcal{Q}'(g)$ is the Fréchet derivative of $\mathcal{Q}$ at g, given by

$$\mathcal{Q}'(g)f = \lim_{\delta \to 0} \frac{\mathcal{Q}(g + \delta f) - \mathcal{Q}(g)}{\delta}. \tag{2.6}$$

2.2 The standard IMEX-RK scheme

The standard IMEX-RK scheme applied to equation (2.1) reads [25]:

$$\begin{aligned} f^{(i)} &= f^n + \Delta t \sum_{j=1}^{i-1} \tilde{a}_{ij} \mathcal{T}(f^{(j)}) + \Delta t \sum_{j=1}^{i} a_{ij} \frac{1}{\varepsilon} \mathcal{Q}(f^{(j)}), \quad i = 1, \ldots, \nu, \\ f^{n+1} &= f^n + \Delta t \sum_{i=1}^{\nu} \tilde{w}_i \mathcal{T}(f^{(i)}) + \Delta t \sum_{i=1}^{\nu} w_i \frac{1}{\varepsilon} \mathcal{Q}(f^{(i)}). \end{aligned} \tag{2.7}$$

Here $\tilde{A} = (\tilde{a}_{ij})$, $\tilde{a}_{ij} = 0$ for $j \geq i$ and $A = (a_{ij})$, $a_{ij} = 0$ for $j > i$ are $\nu \times \nu$ matrices. Along with the vectors $\tilde{\mathbf{w}} = (\tilde{w}_1, \ldots, \tilde{w}_\nu)^T$, $\mathbf{w} = (w_1, \ldots, w_\nu)^T$, they can be represented by a double Butcher tableau:

$$\begin{array}{c|c} \tilde{\mathbf{c}} & \tilde{A} \\ \hline & \tilde{\mathbf{w}}^T \end{array} \qquad \begin{array}{c|c} \mathbf{c} & A \\ \hline & \mathbf{w}^T \end{array} \tag{2.8}$$

where the vectors $\tilde{\mathbf{c}} = (\tilde{c}_1, \ldots, \tilde{c}_\nu)^T$, $\mathbf{c} = (c_1, \ldots, c_\nu)^T$ are defined as

$$\tilde{c}_i = \sum_{j=1}^{i-1} \tilde{a}_{ij}, \quad c_i = \sum_{j=1}^{i} a_{ij}. \tag{2.9}$$

The tableau (2.8) must satisfy certain order conditions [15, 25]. According to the structure of matrix A in the implicit tableau, one usually classifies the IMEX schemes into following categories [4, 11]:

- **Type A**: if the matrix A is invertible.
- **Type CK**: if the matrix A can be written as

$$\begin{pmatrix} 0 & 0 \\ \mathbf{a} & \hat{A} \end{pmatrix}, \tag{2.10}$$

  and the submatrix $\hat{A} \in \mathbb{R}^{(\nu-1)\times(\nu-1)}$ is invertible; in particular, if the vector $\mathbf{a} = 0$, $w_1 = 0$, the scheme is of type ARS.
- If $a_{\nu i} = w_i$, $\tilde{a}_{\nu i} = \tilde{w}_i$, $i = 1, \ldots, \nu$, i.e., $f^{n+1} = f^{(\nu)}$, the scheme is said to be **globally stiffly accurate (GSA)**.

2.3 The new IMEX-RK scheme with correction

We now propose to add a correction step to the standard IMEX scheme (2.7):

$$f^{(i)} = f^n + \Delta t \sum_{j=1}^{i-1} \tilde{a}_{ij} \mathcal{T}(f^{(j)}) + \Delta t \sum_{j=1}^{i} a_{ij} \frac{1}{\varepsilon} \mathcal{Q}(f^{(j)}), \quad i = 1, \dots, \nu, \tag{2.11}$$

$$\tilde{f}^{n+1} = f^n + \Delta t \sum_{i=1}^{\nu} \tilde{w}_i \mathcal{T}(f^{(i)}) + \Delta t \sum_{i=1}^{\nu} w_i \frac{1}{\varepsilon} \mathcal{Q}(f^{(i)}), \tag{2.12}$$

$$f^{n+1} = \tilde{f}^{n+1} - \alpha \Delta t^2 \frac{1}{\varepsilon^2} \mathcal{Q}'(f^*) \mathcal{Q}(f^{n+1}), \tag{2.13}$$

where f^* can be chosen as f^n, $f^{(i)}$, $\tilde{f}^{n+1}$ or f^{n+1}, as long as it is a first-order approximation to f^n: $f^* = f^n + O(\Delta t)$. The coefficients $a_{ij}, \tilde{a}_{ij}, w_i, \tilde{w}_i$, and α remain to be determined.

2.4 Second-order accuracy

Due to the extra correction step (2.13), the standard order conditions for the IMEX-RK schemes need to be modified. In this subsection, we analyze the order conditions of (2.11)-(2.13), up to second order, in the regime $\varepsilon = O(1)$. Without loss of generality, we assume $\varepsilon = 1$.

First, (2.11) gives

$$f^{(i)} = f^n + \Delta t \tilde{c}_i \mathcal{T}(f^n) + \Delta t c_i \mathcal{Q}(f^n) + O(\Delta t^2), \tag{2.14}$$

where we used $f^{(j)} = f^n + O(\Delta t)$ and (2.9). Substituting it into (2.12) yields

$$\begin{aligned}
\tilde{f}^{n+1} &= f^n + \Delta t \sum_{i=1}^{\nu} \tilde{w}_i \mathcal{T}(f^n + \Delta t \tilde{c}_i \mathcal{T}(f^n) + \Delta t c_i \mathcal{Q}(f^n)) \\
&\quad + \Delta t \sum_{i=1}^{\nu} w_i \mathcal{Q}(f^n + \Delta t \tilde{c}_i \mathcal{T}(f^n) + \Delta t c_i \mathcal{Q}(f^n)) + O(\Delta t^3) \\
&= f^n + \Delta t \sum_{i=1}^{\nu} \tilde{w}_i [\mathcal{T}(f^n) + \mathcal{T}'(f^n)(\Delta t \tilde{c}_i \mathcal{T}(f^n) + \Delta t c_i \mathcal{Q}(f^n))] \\
&\quad + \Delta t \sum_{i=1}^{\nu} w_i [\mathcal{Q}(f^n) + \mathcal{Q}'(f^n)(\Delta t \tilde{c}_i \mathcal{T}(f^n) + \Delta t c_i \mathcal{Q}(f^n))] + O(\Delta t^3) \\
&= f^n + \Delta t \left[(\sum_{i=1}^{\nu} \tilde{w}_i) \mathcal{T}(f^n) + (\sum_{i=1}^{\nu} w_i) \mathcal{Q}(f^n) \right] + \Delta t^2 \left[(\sum_{i=1}^{\nu} \tilde{w}_i \tilde{c}_i) \mathcal{T}'(f^n) \mathcal{T}(f^n) \right. \\
&\quad \left. + (\sum_{i=1}^{\nu} \tilde{w}_i c_i) \mathcal{T}'(f^n) \mathcal{Q}(f^n) + (\sum_{i=1}^{\nu} w_i \tilde{c}_i) \mathcal{Q}'(f^n) \mathcal{T}(f^n) + (\sum_{i=1}^{\nu} w_i c_i) \mathcal{Q}'(f^n) \mathcal{Q}(f^n) \right] + O(\Delta t^3),
\end{aligned} \tag{2.15}$$

where $\mathcal{T}', \mathcal{Q}'$ are the Fréchet derivatives of $\mathcal{T}$ and $\mathcal{Q}$. The last step (2.13) implies

$$f^{n+1} = \tilde{f}^{n+1} - \alpha \Delta t^2 \mathcal{Q}'(f^n) \mathcal{Q}(f^n) + O(\Delta t^3). \tag{2.16}$$

Combining (2.15) and (2.16), we have

$$f^{n+1} = f^n + \Delta t \left[(\sum_{i=1}^{\nu} \tilde{w}_i)\mathcal{T}(f^n) + (\sum_{i=1}^{\nu} w_i)\mathcal{Q}(f^n) \right] + \Delta t^2 \left[(\sum_{i=1}^{\nu} \tilde{w}_i \tilde{c}_i)\mathcal{T}'(f^n)\mathcal{T}(f^n) \right.$$
$$\left. + (\sum_{i=1}^{\nu} \tilde{w}_i c_i)\mathcal{T}'(f^n)\mathcal{Q}(f^n) + (\sum_{i=1}^{\nu} w_i \tilde{c}_i)\mathcal{Q}'(f^n)\mathcal{T}(f^n) + (\sum_{i=1}^{\nu} w_i c_i - \alpha)\mathcal{Q}'(f^n)\mathcal{Q}(f^n) \right] + O(\Delta t^3). \tag{2.17}$$

On the other hand, if we Taylor expand the exact solution of (2.1) around time t^n, we have

$$\begin{aligned} f^{n+1}_{\text{exact}} = f^n + \Delta t[\mathcal{T}(f^n) + \mathcal{Q}(f^n)] + \frac{1}{2}\Delta t^2 [\mathcal{T}'(f^n)\mathcal{T}(f^n) + \mathcal{T}'(f^n)\mathcal{Q}(f^n) \\ + \mathcal{Q}'(f^n)\mathcal{T}(f^n) + \mathcal{Q}'(f^n)\mathcal{Q}(f^n)] + O(\Delta t^3). \end{aligned} \tag{2.18}$$

Comparing (2.17) with (2.18), we obtain the following order conditions:

$$\begin{aligned} &\sum_{i=1}^{\nu} \tilde{w}_i = \sum_{i=1}^{\nu} w_i = 1, \\ &\sum_{i=1}^{\nu} \tilde{w}_i \tilde{c}_i = \sum_{i=1}^{\nu} \tilde{w}_i c_i = \sum_{i=1}^{\nu} w_i \tilde{c}_i = \sum_{i=1}^{\nu} w_i c_i - \alpha = \frac{1}{2}. \end{aligned} \tag{2.19}$$

Note that compared to the standard IMEX-RK order conditions [25], the only difference is the term containing α.

2.5 Positivity-preserving property

In this subsection, we analyze the positivity-preserving property of the IMEX-RK scheme (2.11)-(2.13). To this end, we assume $f^n \geq 0$, and derive conditions to insure $f^{(i)}$, $\tilde{f}^{n+1}$ and f^{n+1} all non-negative.

First of all, we observe that if f^n, $f^{(i)}$, $\tilde{f}^{n+1}$ are all non-negative, then the last step (2.13) preserves positivity of the solution provided $\alpha \geq 0$. Indeed, (2.13) can be written as

$$\left(I + \alpha \Delta t^2 \frac{1}{\varepsilon^2} \mathcal{Q}'(f^*)\mathcal{Q} \right) f^{n+1} = \tilde{f}^{n+1}, \tag{2.20}$$

then $f^{n+1} \geq 0$ follows directly from assumption (2.4) if $f^* = f^n, f^{(i)}, \tilde{f}^{n+1}$, and assumption (2.5) if $f^* = f^{n+1}$.

Next, we concentrate on the first two steps (2.11)-(2.12). To simplify the derivation, we assume the IMEX-RK scheme is GSA, that is, $\tilde{f}^{n+1} = f^{(\nu)}$, and consider type A and type ARS schemes, respectively.

2.5.1 Type A and GSA schemes

From (2.11), we know

$$\frac{1}{\varepsilon}\mathcal{Q}(f^{(i)}) = \frac{1}{a_{ii}} \left(\frac{f^{(i)} - f^n}{\Delta t} - \sum_{j=1}^{i-1} \tilde{a}_{ij} \mathcal{T}(f^{(j)}) - \sum_{j=1}^{i-1} a_{ij} \frac{1}{\varepsilon} \mathcal{Q}(f^{(j)}) \right), \quad i = 1, \ldots, \nu. \tag{2.21}$$

Using this relation recursively, we obtain

$$\frac{1}{\varepsilon}\mathcal{Q}(f^{(i)}) = \frac{1}{\Delta t}\sum_{j=1}^{i} b_{ij}(f^{(j)} - f^n) + \sum_{j=1}^{i-1}\tilde{b}_{ij}\mathcal{T}(f^{(j)}), \tag{2.22}$$

where

$$b_{ii} := \frac{1}{a_{ii}}, \quad b_{ij} := -\frac{1}{a_{ii}}\sum_{l=j}^{i-1} a_{il}b_{lj}, \quad \tilde{b}_{ij} := \frac{1}{a_{ii}}(-\tilde{a}_{ij} - \sum_{l=j+1}^{i-1} a_{il}\tilde{b}_{lj}). \tag{2.23}$$

Then (2.11) can be rewritten as

$$\begin{aligned} f^{(i)} &= f^n + \Delta t\sum_{j=1}^{i-1}\tilde{a}_{ij}\mathcal{T}(f^{(j)}) + \Delta t\sum_{j=1}^{i-1} a_{ij}\left[\frac{1}{\Delta t}\sum_{l=1}^{j} b_{jl}(f^{(l)} - f^n) + \sum_{l=1}^{j-1}\tilde{b}_{jl}\mathcal{T}(f^{(l)})\right] + \Delta t a_{ii}\frac{1}{\varepsilon}\mathcal{Q}(f^{(i)}) \\ &= (1 - \sum_{j=1}^{i-1}\sum_{l=j}^{i-1} a_{il}b_{lj})f^n + \sum_{j=1}^{i-1}\left[(\sum_{l=j}^{i-1} a_{il}b_{lj})f^{(j)} + \Delta t(\tilde{a}_{ij} + \sum_{l=j+1}^{i-1} a_{il}\tilde{b}_{lj})\mathcal{T}(f^{(j)})\right] + \Delta t a_{ii}\frac{1}{\varepsilon}\mathcal{Q}(f^{(i)}), \\ &= c_{i0}f^n + \sum_{j=1}^{i-1}\left[c_{ij}f^{(j)} + \Delta t\tilde{c}_{ij}\mathcal{T}(f^{(j)})\right] + \Delta t a_{ii}\frac{1}{\varepsilon}\mathcal{Q}(f^{(i)}), \end{aligned} \tag{2.24}$$

where

$$c_{i0} := 1 - \sum_{j=1}^{i-1}\sum_{l=j}^{i-1} a_{il}b_{lj}, \quad c_{ij} := \sum_{l=j}^{i-1} a_{il}b_{lj}, \quad \tilde{c}_{ij} := \tilde{a}_{ij} + \sum_{l=j+1}^{i-1} a_{il}\tilde{b}_{lj}. \tag{2.25}$$

Thus

$$\left(I - \Delta t a_{ii}\frac{1}{\varepsilon}\mathcal{Q}\right) f^{(i)} = c_{i0}f^n + \sum_{j=1}^{i-1}\left[c_{ij}f^{(j)} + \Delta t\tilde{c}_{ij}\mathcal{T}(f^{(j)})\right]. \tag{2.26}$$

Therefore, to make $f^{(i)} \geq 0$, using assumptions (2.2) and (2.3), it suffices to have

$$\begin{aligned} &a_{ii} > 0, \quad c_{i0} \geq 0, \quad i = 1, \ldots, \nu, \\ &c_{ij} \geq 0, \quad \tilde{c}_{ij} \geq 0, \quad i = 2, \ldots, \nu, \quad j = 1, \ldots, i-1, \end{aligned} \tag{2.27}$$

and the CFL condition is given by

$$\Delta t \leq c_{\text{sch}}\mathcal{C}, \tag{2.28}$$

where c_{sch} is the extra factor from the scheme, defined as

$$c_{\text{sch}} = \min_{\substack{i=2,\ldots,\nu \\ j=1,\ldots,i-1}} \left\{\frac{c_{ij}}{\tilde{c}_{ij}}\right\}, \tag{2.29}$$

and the ratio is understood as infinite if the denominator is zero.

Remark 2.1. *Requiring $a_{ii} > 0$ rather than $a_{ii} \geq 0$ is to make sure the diagonal matrix A in the implicit tableau (2.8) is invertible so the scheme is of type A.*

Remark 2.2. *Note that $c_{i0} + \sum_{j=1}^{i-1} c_{ij} = 1$. Therefore, written in (2.24), the explicit part of the scheme is a convex combination of forward Euler steps, which is the so-called Shu-Osher form [28]. This enables us to derive some nice properties of the scheme that rely on convexity such as entropy decay, see Section 3.2.*

Remark 2.3. *If $\mathcal{T} = \mathcal{T}_{\Delta x}$ is a discretized transport operator, the constraint $\tilde{c}_{ij} \geq 0$ in (2.27) can be removed by using downwinding [13]. This allows more freedom in choosing coefficients and would possibly yield a better CFL condition. For simplicity, we do not consider this situation in the current work.*

We now write down explicitly the above positivity conditions for $\nu = 3$ (the minimum stage required for RK coefficients to exist). First, the double Butcher tableau (2.8) looks like

$$\begin{array}{c|ccc} 0 & 0 & 0 \\ \tilde{a}_{21} & 0 & 0 \\ \tilde{a}_{31} & \tilde{a}_{32} & 0 \\ \hline \tilde{a}_{31} & \tilde{a}_{32} & 0 \end{array} \qquad \begin{array}{c|ccc} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & a_{33} \\ \hline a_{31} & a_{32} & a_{33} \end{array} \tag{2.30}$$

where the vectors $\mathbf{c}$ and $\tilde{\mathbf{c}}$ are omitted. Then the positivity conditions (2.27) reduce to

- for $i = 1$,

$$a_{11} > 0, \quad c_{10} = 1 \geq 0, \tag{2.31}$$

- for $i = 2$,

$$\begin{aligned} &a_{22} > 0, \quad c_{20} = 1 - \frac{a_{21}}{a_{11}} \geq 0, \\ &c_{21} = \frac{a_{21}}{a_{11}} \geq 0, \quad \tilde{c}_{21} = \tilde{a}_{21} \geq 0, \end{aligned} \tag{2.32}$$

- for $i = 3$,

$$\begin{aligned} &a_{33} > 0, \quad c_{30} = 1 - \frac{a_{31}}{a_{11}} + \frac{a_{32}a_{21}}{a_{22}a_{11}} - \frac{a_{32}}{a_{22}} \geq 0, \\ &c_{31} = \frac{a_{31}}{a_{11}} - \frac{a_{32}a_{21}}{a_{22}a_{11}} \geq 0, \quad c_{32} = \frac{a_{32}}{a_{22}} \geq 0, \quad \tilde{c}_{31} = \tilde{a}_{31} - \frac{a_{32}\tilde{a}_{21}}{a_{22}} \geq 0, \quad \tilde{c}_{32} = \tilde{a}_{32} \geq 0. \end{aligned} \tag{2.33}$$

2.5.2 Type ARS and GSA schemes

The analysis for type ARS schemes is similar. Note that since $a_{11} = 0$, $f^{(1)} = f^n$.

First we recursively derive

$$\frac{1}{\varepsilon}\mathcal{Q}(f^{(i)}) = \frac{1}{\Delta t}\sum_{j=2}^{i} b_{ij}(f^{(j)} - f^n) + \sum_{j=1}^{i-1} \tilde{b}_{ij}\mathcal{T}(f^{(j)}), \quad i = 2, \ldots, \nu, \tag{2.34}$$

where

$$b_{ii} := \frac{1}{a_{ii}}, \quad b_{ij} := -\frac{1}{a_{ii}}\sum_{l=j}^{i-1} a_{il}b_{lj}, \quad \tilde{b}_{ij} := \frac{1}{a_{ii}}(-\tilde{a}_{ij} - \sum_{l=j+1}^{i-1} a_{il}\tilde{b}_{lj}). \tag{2.35}$$

Then (2.11) can be rewritten as

$$f^{(i)} = [c_{i0}f^n + \Delta t\tilde{c}_{i0}\mathcal{T}(f^n)] + \sum_{j=2}^{i-1}\left[c_{ij}f^{(j)} + \Delta t\tilde{c}_{ij}\mathcal{T}(f^{(j)})\right] + \Delta t a_{ii}\frac{1}{\varepsilon}\mathcal{Q}(f^{(i)}), \tag{2.36}$$

where

$$c_{i0} := 1 - \sum_{j=2}^{i-1}\sum_{l=j}^{i-1} a_{il}b_{lj}, \quad \tilde{c}_{i0} := \tilde{a}_{i1} + \sum_{j=2}^{i-1} a_{ij}\tilde{b}_{j1}, \quad c_{ij} := \sum_{l=j}^{i-1} a_{il}b_{lj}, \quad \tilde{c}_{ij} = \tilde{a}_{ij} + \sum_{l=j+1}^{i-1} a_{il}\tilde{b}_{lj}. \tag{2.37}$$

Therefore, to make $f^{(i)} \geq 0$, using assumptions (2.2) and (2.3), it suffices to have

$$\begin{aligned} &a_{ii} > 0, \quad c_{i0} \geq 0, \quad \tilde{c}_{i0} \geq 0, \quad i = 2, \ldots, \nu, \\ &c_{ij} \geq 0, \quad \tilde{c}_{ij} \geq 0, \quad i = 3, \ldots, \nu, \quad j = 2, \ldots, i-1, \end{aligned} \tag{2.38}$$

and the CFL condition is given by

$$\Delta t \leq c_{\text{sch}}\mathcal{C}, \tag{2.39}$$

where

$$c_{\text{sch}} = \min\left\{ \min_{i=2,\ldots,\nu} \frac{c_{i0}}{\tilde{c}_{i0}}, \min_{\substack{i=3,\ldots,\nu \\ j=2,\ldots,i-1}} \frac{c_{ij}}{\tilde{c}_{ij}} \right\}, \tag{2.40}$$

and the ratio is understood as infinite if the denominator is zero. Note that similar considerations as pointed out in Remarks 2.1-2.3 apply here as well.

We now write down explicitly the above positivity conditions for $\nu = 4$ (the minimum stage required for RK coefficients to exist). First, the double Butcher tableau (2.8) looks like

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 \\ \tilde{a}_{21} & 0 & 0 & 0 \\ \tilde{a}_{31} & \tilde{a}_{32} & 0 & 0 \\ \tilde{a}_{41} & \tilde{a}_{42} & \tilde{a}_{43} & 0 \\ \hline \tilde{a}_{41} & \tilde{a}_{42} & \tilde{a}_{43} & 0 \end{array} \qquad \begin{array}{c|cccc} 0 & 0 & 0 & 0 \\ 0 & a_{22} & 0 & 0 \\ 0 & a_{32} & a_{33} & 0 \\ 0 & a_{42} & a_{43} & a_{44} \\ \hline 0 & a_{42} & a_{43} & a_{44} \end{array} \tag{2.41}$$

where the vectors $\mathbf{c}$ and $\tilde{\mathbf{c}}$ are omitted. Then the positivity conditions (2.38) reduce to

- for $i = 2$,

$$a_{22} > 0, \quad c_{20} = 1 \geq 0, \quad \tilde{c}_{20} = \tilde{a}_{21} \geq 0, \tag{2.42}$$

- for $i = 3$,

$$\begin{aligned} &a_{33} > 0, \quad c_{30} = 1 - \frac{a_{32}}{a_{22}} \geq 0, \quad \tilde{c}_{30} = \tilde{a}_{31} - \frac{a_{32}\tilde{a}_{21}}{a_{22}} \geq 0, \\ &c_{32} = \frac{a_{32}}{a_{22}} \geq 0, \quad \tilde{c}_{32} = \tilde{a}_{32} \geq 0, \end{aligned} \tag{2.43}$$

- for $i = 4$,

$$\begin{aligned} &a_{44} > 0, \quad c_{40} = 1 - \frac{a_{42}}{a_{22}} + \frac{a_{43}a_{32}}{a_{33}a_{22}} - \frac{a_{43}}{a_{33}} \geq 0, \quad \tilde{c}_{40} = \tilde{a}_{41} - \frac{a_{42}\tilde{a}_{21}}{a_{22}} - \frac{a_{43}\tilde{a}_{31}}{a_{33}} + \frac{a_{43}a_{32}\tilde{a}_{21}}{a_{33}a_{22}} \geq 0, \\ &c_{42} = \frac{a_{42}}{a_{22}} - \frac{a_{43}a_{32}}{a_{33}a_{22}} \geq 0, \quad c_{43} = \frac{a_{43}}{a_{33}} \geq 0, \quad \tilde{c}_{42} = \tilde{a}_{42} - \frac{a_{43}\tilde{a}_{32}}{a_{33}} \geq 0, \quad \tilde{c}_{43} = \tilde{a}_{43} \geq 0. \end{aligned} \tag{2.44}$$

2.6 Combining order conditions and positivity conditions

Combining the results from Sections 2.4 and 2.5, we conclude that as long as one can find the RK coefficients such that they satisfy the order conditions (2.19), positivity conditions (2.27) (resp. (2.38)), and $\alpha \geq 0$, the resulting scheme (2.11)-(2.13) would be both second-order accurate and positivity-preserving. It turns out that such sets of coefficients are very easy to find. Below we give two IMEX schemes, one of type A and GSA with $\nu = 3$ and one of type ARS and GSA with $\nu = 4$. These coefficients are searched to yield a relatively large CFL constant c_{sch}, but we do not claim their optimality.

2.6.1 A second-order positivity-preserving type A and GSA scheme

A type A and GSA scheme of form (2.30):

$$\begin{aligned}
&\tilde{a}_{21} = 0.7369502715,\\
&\tilde{a}_{31} = 0.3215281691, \quad \tilde{a}_{32} = 0.6784718309,\\
&a_{11} = 0.6286351712,\\
&a_{21} = 0.2431004655, \quad a_{22} = 0.1959392570,\\
&a_{31} = 0.4803651051, \quad a_{32} = 0.0746432814, \quad a_{33} = 0.4449916135.
\end{aligned}$$

α in the correction step (2.13) and the CFL constant (2.29) are given by

$$\alpha = 0.2797373792, \quad c_{\text{sch}} = 0.5247457524.$$

2.6.2 A second-order positivity-preserving type ARS and GSA scheme

A type ARS and GSA scheme of form (2.41):

$$\begin{aligned}
&\tilde{a}_{21} = 0,\\
&\tilde{a}_{31} = 1.0, \quad \tilde{a}_{32} = 0,\\
&\tilde{a}_{41} = 0.5, \quad \tilde{a}_{42} = 0, \quad \tilde{a}_{43} = 0.5,\\
&a_{22} = 1.6,\\
&a_{32} = 0.3, \quad a_{33} = 0.7,\\
&a_{42} = 0.5, \quad a_{43} = 0.3, \quad a_{44} = 0.2.
\end{aligned}$$

α in the correction step (2.13) and the CFL constant (2.40) are given by

$$\alpha = 0.8, \quad c_{\text{sch}} = 0.8125.$$

Remark 2.4. *For simplicity, we only give examples for second-order method. Following a similar procedure in Section 2.4, it is not difficult to derive order conditions for third-order method (see Appendix). This, combined with the positivity conditions in Section 2.5, would yield a third-order positivity-preserving scheme.*

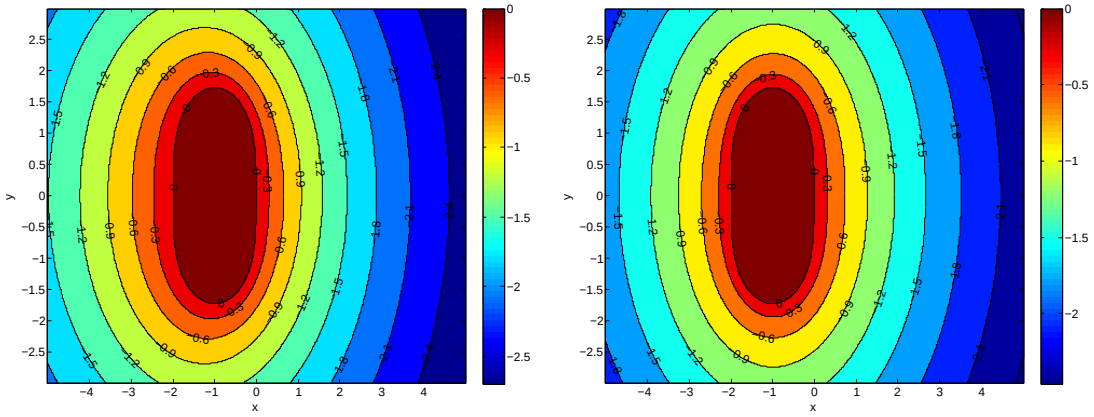

Figure 1: Contours of the stability region $\mathcal{S}$ (2.47) at some fixed values of z_2: Left: the type A scheme given in Section 2.6.1; Right: the type ARS scheme given in Section 2.6.2.

2.7 Absolute stability

In this subsection, we analyze the absolute stability of the proposed IMEX scheme. We consider the linear ODE

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \lambda_1 f + \lambda_2 f, \quad \lambda_1 \in \mathbb{C},\, \lambda_2 < 0, \tag{2.45}$$

and solve it by scheme (2.11)-(2.13), i.e.,

$$\begin{aligned} f^{(i)} &= f^n + \Delta t \sum_{j=1}^{i-1} \tilde{a}_{ij}\lambda_1 f^{(j)} + \Delta t \sum_{j=1}^{i} a_{ij}\lambda_2 f^{(j)}, \quad i = 1, \dots, \nu, \\ \tilde{f}^{n+1} &= f^n + \Delta t \sum_{i=1}^{\nu} \tilde{w}_i \lambda_1 f^{(i)} + \Delta t \sum_{i=1}^{\nu} w_i \lambda_2 f^{(i)}, \\ f^{n+1} &= \tilde{f}^{n+1} - \alpha \Delta t^2 \lambda_2^2 f^{n+1}. \end{aligned} \tag{2.46}$$

Define $z_i = \lambda_i \Delta t$, $i = 1, 2$, then one can write $f^{n+1} = P(z_1, z_2) f^n$, where $P(z_1, z_2)$ is the amplification factor of the scheme. The absolute stability region of the scheme is defined as [23]:

$$\mathcal{S} = \{(z_1, z_2) : |P(z_1, z_2)| \le 1\}. \tag{2.47}$$

In Figure 1, we plot the stability regions of the two schemes given in Section 2.6. Here we write $z_1 = x + iy$, and plot the level sets $\mathcal{S} \cap \{z_2 = C\}$ for different values of C in the x, y plane. One can clearly see that for both schemes, as z_2 becomes more negative, the level sets are strictly increasing. This suggests that the schemes are linearly stable, if the stability condition for the explicit part (corresponding to the case $z_2 = 0$) is satisfied.

3 Application to the BGK equation

We now apply the previously derived general framework to the BGK equation (1.1). The convection operator $-v\cdot\nabla_x$ and the collision operator Q correspond, respectively, to the operators $\mathcal{T}$ and $\mathcal{Q}$ in the general setting (2.1). We have the following:

Proposition 3.1. *The operators $\mathcal{T}(f) = -v \cdot \nabla_x f$ and $\mathcal{Q}(f) = \tau_f(M[f] - f)$ satisfy the assumptions (2.2)-(2.5).*

Proof. First of all, the operator $\mathcal{T}(f)$ can satisfy the assumption (2.2) if a positivity-preserving spatial discretization is used (see Section 3.3).

To verify (2.3), note that

$$(I - b\mathcal{Q})f = g \iff f - b\tau_f(M[f] - f) = g \iff f = \frac{b\tau_f M[f] + g}{1 + b\tau_f}. \tag{3.1}$$

Taking the moments $\langle \cdot \phi \rangle$ on both sides of the middle equation gives $\langle f\phi \rangle = \langle g\phi \rangle$ since $\langle f\phi \rangle = \langle M[f]\phi \rangle$. Thus if $g \geq 0$, then $\tau_g > 0$, so $\tau_f = \tau_g > 0$. Then the last equation of (3.1) implies $f \geq 0$ for any constant $b \geq 0$.

We now compute $\mathcal{Q}'(g)\mathcal{Q}(f)$:

$$\mathcal{Q}'(g)\mathcal{Q}(f) = \lim_{\delta \to 0} \frac{\mathcal{Q}(g + \delta\mathcal{Q}(f)) - \mathcal{Q}(g)}{\delta}. \tag{3.2}$$

Since $\langle (g + \delta\mathcal{Q}(f))\phi \rangle = \langle (g + \delta\tau_f(M[f] - f))\phi \rangle = \langle g\phi \rangle$, so

$$\mathcal{Q}(g + \delta\mathcal{Q}(f)) - \mathcal{Q}(g) = \tau_g(M[g] - g - \delta\mathcal{Q}(f)) - \tau_g(M[g] - g) = -\tau_g\delta\mathcal{Q}(f). \tag{3.3}$$

Hence

$$\mathcal{Q}'(g)\mathcal{Q}(f) = -\tau_g\mathcal{Q}(f). \tag{3.4}$$

Then

$$(I + b\mathcal{Q}'(g)\mathcal{Q})f = h \iff (I - b\tau_g\mathcal{Q})f = h. \tag{3.5}$$

If $g \geq 0$, then $\tau_g > 0$. Thus (2.4) follows from (2.3). To verify (2.5), note that

$$(I + b\mathcal{Q}'(f)\mathcal{Q})f = h \iff (I - b\tau_f\mathcal{Q})f = h, \tag{3.6}$$

from which we know $\langle f\phi \rangle = \langle h\phi \rangle$. If $h \geq 0$, then $\tau_f = \tau_h > 0$. Thus (2.5) follows again from (2.3). □

Therefore, applying the scheme (2.11)-(2.13) to the BGK equation, we get a second-order, positivity-preserving method:

$$\begin{aligned} f^{(i)} &= f^n - \Delta t \sum_{j=1}^{i-1} \tilde{a}_{ij} v \cdot \nabla_x f^{(j)} + \Delta t \sum_{j=1}^{i} a_{ij} \frac{\tau_{f^{(j)}}}{\varepsilon}(M[f^{(j)}] - f^{(j)}), \quad i = 1, \ldots, \nu, \\ f^{n+1} &= f^{(\nu)} + \alpha \Delta t^2 \frac{\tau_{f^*}}{\varepsilon^2}(M[f^{n+1}] - f^{n+1}), \end{aligned} \tag{3.7}$$

where f^* can be taken as f^n, any $f^{(i)}$ or f^{n+1}, and the coefficients $\tilde{a}_{ij}$, a_{ij}, α and the CFL constant c_{sch} are given in Section 2.6. Note that we have restricted to GSA schemes to get positivity, so there is no middle step $\tilde{f}^{n+1}$. Furthermore, due to the special structure (3.4) of the BGK operator, the implementation of the correction step is just as easy as solving the collision operator implicitly.

Remark 3.2. *The scheme (3.7) appears implicit since at every stage i one needs to compute $\tau_{f^{(i)}}, M[f^{(i)}]$ first in order to evaluate $f^{(i)}$ (also for the last step). This can be achieved by taking the moments $\langle \cdot \phi \rangle$ on both sides of the scheme:*

$$\begin{aligned} &\langle f^{(i)}\phi\rangle = \langle f^n\phi\rangle - \Delta t \sum_{j=1}^{i-1} \tilde{a}_{ij} \nabla_x \cdot \langle f^{(j)} v\phi\rangle, \quad i = 1, \dots, \nu, \\ &\langle f^{n+1}\phi\rangle = \langle f^{(\nu)}\phi\rangle. \end{aligned} \tag{3.8}$$

Hence one can obtain the macroscopic quantities ρ, u, T at stage i first, which will define $\tau_{f^{(i)}}$ and $M[f^{(i)}]$ (the last step is treated similarly). This idea has been used in several papers to solve the BGK equation implicitly [10, 26, 12, 11].

3.1 Asymptotic-preserving (AP) property

There remains to prove the scheme (3.7) is AP. To this end, we discuss type A schemes and type ARS schemes separately.

Proposition 3.3. *If the IMEX scheme (3.7) is of type A and GSA, it is AP: for fixed Δt, in the limit $\varepsilon \to 0$, the scheme becomes a second-order explicit RK scheme applied to the limiting Euler system (1.8).*

Proof. We rewrite the first ν steps of (3.7) using vector notations:

$$\mathbf{F} = f^n \mathbf{e} - \Delta t \tilde{A} v \cdot \nabla_x \mathbf{F} + \Delta t A \frac{\tau}{\varepsilon}(M[\mathbf{F}] - \mathbf{F}), \tag{3.9}$$

where $\mathbf{F} := (f^{(1)}, \dots, f^{(\nu)})^T$, $\mathbf{e} := (1, \dots, 1)^T$, $M[\mathbf{F}] := (M[f^{(1)}], \dots, M[f^{(\nu)}])^T$, and $\tau := \mathrm{diag}(\tau_{f^{(1)}}, \dots, \tau_{f^{(\nu)}})$. Now fixing Δt, formally passing the limit $\varepsilon \to 0$ in (3.9), one has $\Delta t A \tau (M[\mathbf{F}] - \mathbf{F}) \to 0$. This implies $\mathbf{F} \to M[\mathbf{F}]$ since both A and τ are invertible (the scheme is of type A and positivity-preserving). Replacing $\mathbf{F}$ by $M[\mathbf{F}]$ in the moment system (3.8), we obtain

$$\begin{aligned} &U^{(i)} = U^n - \Delta t \sum_{j=1}^{i-1} \tilde{a}_{ij} \nabla_x \cdot \langle M[f^{(j)}] v\phi\rangle, \quad i = 1, \dots, \nu, \\ &U^{n+1} = U^{(\nu)}, \end{aligned} \tag{3.10}$$

where $U := (\rho, \rho u, E)^T$. This is a second-order explicit RK scheme applied to the compressible Euler system (1.8). □

Proposition 3.4. *If the IMEX scheme (3.7) is of type ARS and GSA, it is AP: for fixed Δt and consistent initial data $f^0 = M[f^0]$, in the limit $\varepsilon \to 0$, the scheme becomes a second-order explicit RK scheme applied to the limiting Euler system (1.8). If the initial data is inconsistent, the limiting scheme will degenerate to first order.*

Proof. For the ARS scheme, $f^{(1)} = f^n$ and $\mathbf{a} = 0$. Rewrite $\mathbf{F} = (f^{(1)}, \hat{\mathbf{F}})$, $\mathbf{e} = (1, \hat{\mathbf{e}})$, $M[\mathbf{F}] = (M[f^{(1)}], M[\hat{\mathbf{F}}])$, $\hat{\tau} := \mathrm{diag}(\tau_{f^{(2)}}, \dots, \tau_{f^{(\nu)}})$, then (3.9) becomes

$$\hat{\mathbf{F}} = f^n \hat{\mathbf{e}} - \Delta t \tilde{\mathbf{a}} v \cdot \nabla_x f^n - \Delta t \hat{\tilde{A}} v \cdot \nabla_x \hat{\mathbf{F}} + \Delta t \hat{A} \frac{\hat{\tau}}{\varepsilon}(M[\hat{\mathbf{F}}] - \hat{\mathbf{F}}), \tag{3.11}$$

where we have used a similar notation for matrix $\tilde{A}$ as that in (2.10):

$$\begin{pmatrix} 0 & 0 \\ \tilde{\mathbf{a}} & \hat{\tilde{A}} \end{pmatrix}. \tag{3.12}$$

Now fix Δt, let $\varepsilon \to 0$, one has $\Delta t \hat{A}\hat{\tau}(M[\hat{\mathbf{F}}] - \hat{\mathbf{F}}) \to 0$. So $\hat{\mathbf{F}} \to M[\hat{\mathbf{F}}]$ since both $\hat{A}$ and $\hat{\tau}$ are invertible (the scheme is of type CK and positivity-preserving). Replacing $\hat{\mathbf{F}}$ by $M[\hat{\mathbf{F}}]$ in the moment system (3.8), we have

$$\begin{aligned} &U^{(i)} = U^n - \Delta t \tilde{a}_{i1} \nabla_x \cdot \langle f^n v\phi \rangle - \Delta t \sum_{j=2}^{i-1} \tilde{a}_{ij} \nabla_x \cdot \langle M[f^{(j)}] v\phi \rangle, \quad i = 2, \dots, \nu, \\ &U^{n+1} = U^{(\nu)}, \end{aligned} \tag{3.13}$$

which is a second-order explicit RK scheme applied to the compressible Euler system (1.8) if $f^n = M[f^n]$. On the other hand, the last step of (3.7) implies $f^{n+1} \to M[f^{n+1}]$ as $\varepsilon \to 0$. Therefore, as long as the initial data is consistent $f^0 = M[f^0]$, the scheme is second order. Otherwise, the initial data will bring an $O(\Delta t)$ error and the scheme is reduced to first order. □

3.2 Entropy-decay property

It can be shown that the second-order scheme (3.7) satisfies an entropy-decay property if the simple first-order upwind scheme is used for spatial derivative.

Consider the following 1D BGK equation for simplicity:

$$\partial_t f + v \partial_x f = \frac{1}{\varepsilon}(M[f] - f), \tag{3.14}$$

for which we have the entropy inequality

$$\frac{\mathrm{d}}{\mathrm{d}t} \iint f \log f \, \mathrm{d}v \, \mathrm{d}x \le 0. \tag{3.15}$$

Now assume that the velocity domain is truncated to a large enough symmetric interval $[-|v|_{\max}, |v|_{\max}]$ and the convection term $v\partial_x f$ is discretized by the first-order upwind scheme

$$(v\partial_x f)_k = \chi_{v \ge 0} v \frac{f_k - f_{k-1}}{\Delta x} + \chi_{v<0} v \frac{f_{k+1} - f_k}{\Delta x}, \tag{3.16}$$

together with the periodic or compactly supported boundary condition in x. Then we claim that the scheme (3.7) satisfies a discrete entropy inequality:

$$S[f^{n+1}] \le S[f^n], \tag{3.17}$$

where the entropy S is defined as

$$S[f] = \Delta x \sum_k \int s[f_k] \, \mathrm{d}v, \quad \text{with} \quad s[f_k] := f_k \log f_k. \tag{3.18}$$

We prove it for type A and GSA schemes. Type ARS and GSA schemes can be treated similarly.

First applying (3.16) in (2.24) gives

$$\begin{aligned} f_k^{(i)} = c_{i0} f_k^n + \sum_{j=1}^{i-1} \left[c_{ij} f_k^{(j)} - \frac{v\Delta t}{\Delta x} \tilde{c}_{ij} [\chi_{v \ge 0}(f_k^{(j)} - f_{k-1}^{(j)}) + \chi_{v<0}(f_{k+1}^{(j)} - f_k^{(j)})] \right] \\ + \Delta t a_{ii} \frac{1}{\varepsilon}(M[f_k^{(i)}] - f_k^{(i)}), \end{aligned} \tag{3.19}$$

and the CFL condition (2.28) becomes

$$\Delta t \le \min_{i,j} \left\{ \frac{c_{ij}}{\tilde{c}_{ij}} \right\} \frac{\Delta x}{|v|_{\max}}. \tag{3.20}$$

Note that (3.19) can be written equivalently as

$$f_k^{(i)*} = c_{i0} f_k^n + \sum_{j=1}^{i-1} \left[\left(c_{ij} - \tilde{c}_{ij} \frac{|v|\Delta t}{\Delta x} \right) f_k^{(j)} + \tilde{c}_{ij} \frac{|v|\Delta t}{\Delta x} [\chi_{v\ge 0} f_{k-1}^{(j)} + \chi_{v<0} f_{k+1}^{(j)}] \right], \tag{3.21}$$

$$f_k^{(i)} = \left(1 + \frac{\Delta t}{\varepsilon} a_{ii} \right)^{-1} \left(f_k^{(i)*} + \frac{\Delta t}{\varepsilon} a_{ii} M[f_k^{(i)}] \right). \tag{3.22}$$

Recall that

$$a_{ii} > 0, \quad c_{i0} \ge 0, \quad c_{ij} \ge 0, \quad \tilde{c}_{ij} \ge 0, \quad c_{i0} + \sum_{j=1}^{i-1} c_{ij} = 1, \tag{3.23}$$

hence the right hand side of (3.21) is a convex combination of f_k^n, $f_k^{(j)}$, and $[\chi_{v\ge 0} f_{k-1}^{(j)} + \chi_{v<0} f_{k+1}^{(j)}]$, provided the CFL condition is satisfied. Since $s[f_k]$ is a convex function for $f_k > 0$, by Jensen's inequality, (3.21) after integration in v and summing over in k yields

$$S[f^{(i)*}] \le c_{i0} S[f^n] + \sum_{j=1}^{i-1} \left[\left(c_{ij} - \tilde{c}_{ij} \frac{|v|\Delta t}{\Delta x} \right) S[f^{(j)}] + \tilde{c}_{ij} \frac{|v|\Delta t}{\Delta x} \Delta x \sum_k \int s[\chi_{v\ge 0} f_{k-1}^{(j)} + \chi_{v<0} f_{k+1}^{(j)}] \, \mathrm{d}v \right]. \tag{3.24}$$

Notice that

$$\begin{aligned} \Delta x \sum_k \int s[\chi_{v\ge 0} f_{k-1}^{(j)} + \chi_{v<0} f_{k+1}^{(j)}] \, \mathrm{d}v &= \Delta x \sum_k \left[\int_{v\ge 0} f_{k-1}^{(j)} \log f_{k-1}^{(j)} \, \mathrm{d}v + \int_{v<0} f_{k+1}^{(j)} \log f_{k+1}^{(j)} \, \mathrm{d}v \right] \\ &= \Delta x \sum_k \left[\int_{v\ge 0} f_k^{(j)} \log f_k^{(j)} \, \mathrm{d}v + \int_{v<0} f_k^{(j)} \log f_k^{(j)} \, \mathrm{d}v \right] = S[f^{(j)}], \end{aligned} \tag{3.25}$$

thus

$$S[f^{(i)*}] \le c_{i0} S[f^n] + \sum_{j=1}^{i-1} c_{ij} S[f^{(j)}]. \tag{3.26}$$

On the other hand, using the fact that[1]

$$S[M[f^{(i)}]] \le S[f^{(i)}], \tag{3.27}$$

from (3.22), which is also a convex combination, one has

$$S[f^{(i)}] \le \left(1 + \frac{\Delta t}{\varepsilon} a_{ii} \right)^{-1} \left(S[f^{(i)*}] + \frac{\Delta t}{\varepsilon} a_{ii} S[M[f^{(i)}]] \right) \le \left(1 + \frac{\Delta t}{\varepsilon} a_{ii} \right)^{-1} \left(S[f^{(i)*}] + \frac{\Delta t}{\varepsilon} a_{ii} S[f^{(i)}] \right), \tag{3.28}$$

which implies

$$S[f^{(i)}] \le S[f^{(i)*}]. \tag{3.29}$$

[1] An easy way to show this is: $\int M \log M \, \mathrm{d}v - \int f \log f \, \mathrm{d}v = \int f \log \frac{M}{f} \, \mathrm{d}v = \int f [\log \frac{M}{f} - \frac{M}{f} + 1] \, \mathrm{d}v \le 0$, where we used the fact that f and M have the same moments $\langle f\phi \rangle = \langle M\phi \rangle$, and the inequality $\log x \le x - 1$ for $x > 0$.

Therefore,

$$S[f^{(i)}] \le c_{i0} S[f^n] + \sum_{j=1}^{i-1} c_{ij} S[f^{(j)}], \tag{3.30}$$

from which it follows easily that $S[f^{(\nu)}] \le S[f^n]$. Finally, the last step of (3.7) has the same structure as (3.22), thus it can be shown in the same way that $S[f^{n+1}] \le S[f^{(\nu)}]$. Altogether, we have proved $S[f^{n+1}] \le S[f^n]$.

Remark 3.5. *To the best of our knowledge, there is no second or high order fully discretized scheme that satisfies an entropy-decay property for the BGK equation. The purpose of this subsection is to illustrate that one can readily obtain a fully discretized second-order entropy-decay scheme for the BGK equation provided a second-order entropy-decay spatial discretization exists since the proposed time discretization is based on convex combinations.*

3.3 Spatial and velocity domain discretizations

In this subsection, we describe in detail how to obtain a fully discretized scheme for the BGK equation. We emphasize that it is not straightforward to apply the established techniques. Special care needs to be given for both spatial and velocity domain discretizations in order to maintain the properties (positivity and AP) of the semi-discretized scheme.

First of all, to preserve the positivity of the solution, a positivity-preserving spatial discretization must be used for the convection term. One can use a high order accurate discontinuous Galerkin or finite volume scheme with a high order accurate bound-preserving limiter by Zhang and Shu in [31, 33]. Here we choose to use a finite volume method for x-variable and a finite difference method for v-variable.

Consider solving the 1D BGK equation (3.14) with a possibly x-dependent Knudsen number $\varepsilon(x)$ (this is usually the case when handling a multiscale problem). We propose to conduct the temporal discretization first and then the spatial and velocity discretizations. For simplicity, we use the first-order IMEX scheme as an illustration (the high order IMEX can be implemented in a similar fashion), which can be performed in three steps:

$$\frac{f^* - f^n}{\Delta t} + v\partial_x f^n = 0, \tag{3.31a}$$

$$U^{n+1} = \langle f^* \phi \rangle, \quad M^{n+1} = M[U^{n+1}], \tag{3.31b}$$

$$f^{n+1} = \frac{1}{1 + \Delta t/\varepsilon(x)} f^* + \frac{\Delta t/\varepsilon(x)}{1 + \Delta t/\varepsilon(x)} M^{n+1}, \tag{3.31c}$$

where the middle step is to take the moments of f^* to get macroscopic quantities $U = (\rho, m, E)$ which will define ρ, u, T, hence $M[U]$ accordingly. Now define the grid points in x as $x_{j+\frac{1}{2}} = (j + \frac{1}{2})\Delta x$. After integration of the above scheme in x over the interval $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$ at the grid point $v = v_k$, we obtain

$$\frac{f^*_{j,k} - f^n_{j,k}}{\Delta t} + \frac{\hat{F}^n_{j+1/2,k} - \hat{F}^n_{j-1/2,k}}{\Delta x} = 0, \tag{3.32a}$$

$$U^{n+1} = \langle f^* \phi \rangle, \quad M^{n+1} = M[U^{n+1}], \tag{3.32b}$$

$$f^{n+1}_{j,k} = \frac{1}{\Delta x} \int_{I_j} \left[\frac{1}{1 + \Delta t/\varepsilon(x)} f^*_k(x) + \frac{\Delta t/\varepsilon(x)}{1 + \Delta t/\varepsilon(x)} M^{n+1}_k(x) \right] \mathrm{d}x, \tag{3.32c}$$

where $f_{j,k}$ denotes the cell average of f on the interval I_j at k-th velocity grid point, $\hat{F}_{j+1/2,k}$ is the numerical flux approximating $v_k f(t,x,v_k)$ at $x = x_{j+\frac{1}{2}}$, and $f_k^*(x)$ and $M_k^{n+1}(x)$ are high order accurate reconstruction polynomials (reconstructed by the cell averages $\{f_{j,k}^*\}_{j=1}^{N_x}$ and $\{M_{j,k}^{n+1}\}_{j=1}^{N_x}$) approximating the functions $f^*(\cdot, v_k)$ and $M^{n+1}(\cdot, v_k)$ respectively.

In the following, we explain the details of the scheme (3.32) step by step.

3.3.1 Handling the convection term

First we discuss how to enforce the non-negativity of $f_{j,k}^*$ in (3.32a). We omit the index k for convenience. Given the cell averages f_j^n, we use the fifth-order finite volume WENO reconstruction [27] to construct fifth-order accurate approximations $f_{j+\frac{1}{2}}^+$ and $f_{j+\frac{1}{2}}^-$ to the point value f at $x = x_{j+\frac{1}{2}}$ and $t = t^n$. Notice that $f_{j+\frac{1}{2}}^{\pm}$ might be negative. There exists a degree four polynomial $p_j(x)$ on the j-th cell, which is a fifth-order approximation to f on the cell, and satisfies the property that the cell average of $p_j(x)$ is exactly f_j^n, and $p_j(x_{j-1/2}) = f_{j-1/2}^+$, $p_j(x_{j+1/2}) = f_{j+1/2}^-$. For instance, such a polynomial can be obtained by interpolation, even though the construction of this polynomial is not needed in the implementation. Then the four-point Gauss-Lobatto quadrature $f_j^n = \sum_{l=1}^4 p_j(x_{j,l})\omega_l$ is exact, where $\{x_{j,1} = x_{j-1/2}, x_{j,2}, x_{j,3}, x_{j,4} = x_{j+1/2}\}$ are the quadrature points, and $\{w_l\}$ are the corresponding quadrature weights on the interval $[-\frac{1}{2}, \frac{1}{2}]$ such that $\sum_{l=1}^4 w_l = 1$. Next by the simplified bound-preserving limiter for finite volume methods described in [33], we modify $p_j(x)$ into

$$\tilde{p}_j(x) = \theta_j(p_j(x) - f_j^n) + f_j^n, \quad \theta_j = \min\left\{\left|\frac{f_j^n}{m_j - f_j^n}\right|, 1\right\}, \quad m_j = \min\{p_j(x_{j-1/2}), p_j(x_{j+1/2}), \xi_j\}, \tag{3.33a}$$

with

$$\xi_j = \frac{p_j(x_{j,2})\omega_2 + p_j(x_{j,3})\omega_3}{\omega_2 + \omega_3} = \frac{f_j^n - f_{j-1/2}^+\omega_1 - f_{j+1/2}^-\omega_4}{\omega_2 + \omega_3}. \tag{3.33b}$$

The limiter (3.33) guarantees that $\tilde{f}_{j+\frac{1}{2}}^- = \tilde{p}_j(x_{j+\frac{1}{2}}) \geq 0$, $\tilde{f}_{j-\frac{1}{2}}^+ = \tilde{p}_j(x_{j-\frac{1}{2}}) \geq 0$ and $\tilde{\xi}_j = \frac{f_j^n - \tilde{f}_{j-1/2}^+\omega_1 - \tilde{f}_{j+1/2}^-\omega_4}{\omega_2+\omega_3} \geq 0$. Moreover, the quadrature $f_j^n = \sum_{l=1}^4 \tilde{p}_j(x_{j,l})\omega_l$ is still exact and $\tilde{f}_{j+\frac{1}{2}}^{\pm}$ are still fifth-order accurate approximations to the the point value of f at $x = x_{j+\frac{1}{2}}$, see [31, 33, 30]. Since we only need $\tilde{f}_{j+\frac{1}{2}}^-$ and $\tilde{f}_{j-\frac{1}{2}}^+$, the limiter (3.33) is equivalent to the following implementation without using $p_j(x)$:

$$\tilde{f}_{j+\frac{1}{2}}^- = \theta_j(f_{j+\frac{1}{2}}^- - f_j^n) + f_j^n, \quad \tilde{f}_{j-\frac{1}{2}}^+ = \theta_j(f_{j-\frac{1}{2}}^+ - f_j^n) + f_j^n, \quad \theta_j = \min\left\{\left|\frac{f_j^n}{m_j - f_j^n}\right|, 1\right\}, \tag{3.34a}$$

$$m_j = \min\{f_{j-1/2}^+, f_{j+1/2}^-, \xi_j\}, \quad \xi_j = \frac{f_j^n - f_{j-1/2}^+\omega_1 - f_{j+1/2}^-\omega_4}{\omega_2 + \omega_3}. \tag{3.34b}$$

Then we define the upwind flux as

$$\hat{F}_{j+1/2}^n = \begin{cases} v_k \tilde{f}_{j+\frac{1}{2}}^-, & \text{if } v_k \geq 0, \\ v_k \tilde{f}_{j+\frac{1}{2}}^+, & \text{if } v_k < 0. \end{cases} \tag{3.35}$$

To see the positivity of f_j^* in (3.32a) using (3.35), we only discuss the case $v_k \geq 0$ with the other case being similar. We have

$$\begin{aligned} f_j^* &= [\tilde{p}_j(x_{j-1/2})\omega_1 + \tilde{p}_j(x_{j+1/2})\omega_4 + \tilde{\xi}_j(\omega_2+\omega_3)] - \frac{v_k\Delta t}{\Delta x}(\tilde{p}_j(x_{j+1/2}) - \tilde{p}_{j-1}(x_{j-1/2})) \\ &= \tilde{p}_j(x_{j-1/2})\omega_1 + \tilde{p}_j(x_{j+1/2})(\omega_4 - \frac{v_k\Delta t}{\Delta x}) + \tilde{\xi}_j(\omega_2+\omega_3) + \frac{v_k\Delta t}{\Delta x}\tilde{p}_{j-1}(x_{j-1/2}), \end{aligned} \tag{3.36}$$

which implies the positivity of f_j^* since it is a convex combination of non-negative quantities under the CFL condition $\frac{v_k\Delta t}{\Delta x} \leq \omega_4 = \frac{1}{12}$.

3.3.2 Handling the collision term

Now we describe how to compute $M^{n+1} = M[U^{n+1}]$ under the finite volume discretization in x. For convenience, we regard v as a continuous variable and omit the superscript $n+1$.

Let U_j be the moments of $f_j^*(v) \geq 0$ on the j-th cell, then U_j belongs to a convex set of admissible states with positive density and temperature:

$$G = \left\{ (\rho, m, E)^T : \rho > 0, \quad E - \frac{1}{2}\frac{m^2}{\rho} > 0 \right\}. \tag{3.37}$$

Let $\{\tilde{x}_{j,l}\}$ $(l = 1,2,3)$ denote the three-point Gauss-Legendre quadrature on the j-th cell $[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$ and $\{\tilde{w}_l\}$ $(l = 1,2,3)$ be the corresponding quadrature weights on the interval $[-\frac{1}{2}, \frac{1}{2}]$, which is exact for integrating polynomials of degree five. Given cell averages of macroscopic quantities $U_j \in G$, we would like to reconstruct fifth-order approximations to $U(x)$ at $x = \tilde{x}_{j,l}$, denoted as $U_{j,l}, l = 1,2,3$. Moreover, we need them to be positive so that $M[U_{j,l}]$ can be well-defined; and conservative so that the final scheme is AP. Namely, we need

$$U_{j,l} \in G \quad \text{and} \quad \sum_{l=1}^{3} \tilde{w}_l U_{j,l} = U_j. \tag{3.38}$$

Such a reconstruction can be done in the following way. First, we construct a polynomial $U_j(x)$ of degree four, which is a fifth-order accurate approximation to $U(x)$ on the interval I_j with U_j as its cell average. There are many ways to construct such a polynomial, e.g., we can first reconstruct two cell ends values by the WENO method then construct a Hermite type reconstruction polynomial using these two point values and three averages U_{j-1}, U_j, U_{j+1}, see [31]. Thus $\frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} U_j(x)\,\mathrm{d}x = U_j$. Second, we apply the simple positivity-preserving limiter in [32, 30] to $U_j(x)$ to obtain a modified polynomial $\tilde{U}_j(x)$ such that $\tilde{U}_j(\tilde{x}_{j,l}) \in G$ and the cell average of $\tilde{U}_j(x)$ is still U_j. Finally, we set $U_{j,l} = \tilde{U}_j(\tilde{x}_{j,l})$, and we have

$$\sum_{l=1}^{3} \tilde{w}_l U_{j,l} = \sum_{l=1}^{3} \tilde{w}_l \tilde{U}_j(\tilde{x}_{j,l}) = \frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \tilde{U}_j(x)\,\mathrm{d}x = U_j. \tag{3.39}$$

Then $M[U_{j,l}], l = 1,2,3$ are well-defined and we set

$$M_j = \sum_{l=1}^{3} \tilde{w}_l M[U_{j,l}]. \tag{3.40}$$

This method is fifth-order in x, since the reconstruction is fifth-order and the positivity-preserving limiter does not affect the accuracy. Also, this method is conservative:

$$\langle M_j\phi\rangle = \sum_{l=1}^{3} \tilde{w}_l \langle M[U_{j,l}]\phi\rangle = \sum_{l=1}^{3} \tilde{w}_l U_{j,l} = U_j = \langle f_j^*\phi\rangle, \tag{3.41}$$

which is the key to obtain AP property.

3.3.3 Handling the variable $\varepsilon(x)$

In the last step (3.32c) we need to compute an integral on I_j, which can be approximated by the Gauss-Legendre quadrature:

$$\begin{aligned}&\int_{I_j}\left[\frac{1}{1+\Delta t/\varepsilon(x)}f_k^*(x)+\frac{\Delta t/\varepsilon(x)}{1+\Delta t/\varepsilon(x)}M_k^{n+1}(x)\right]\mathrm{d}x\\ \approx&\sum_{l=1}^{3}\tilde{w}_l\left[\frac{1}{1+\Delta t/\varepsilon(\tilde{x}_{j,l})}f_k^*(\tilde{x}_{j,l})+\frac{\Delta t/\varepsilon(\tilde{x}_{j,l})}{1+\Delta t/\varepsilon(\tilde{x}_{j,l})}M_k^{n+1}(\tilde{x}_{j,l})\right].\end{aligned}\tag{3.42}$$

Thus we only need the approximation of the functions $f_k^*(x)$ and $M_k^{n+1}(x)$ at the quadrature points $\{\tilde{x}_{j,l}\}$ $(l=1,2,3)$. The values for M can be read directly from the previous step. The construction of f can be done in the same way as we construct $U_{j,l}\in G$ in the previous section, with the convex set G replaced by the set $\{f: f\geq 0\}$.

3.3.4 AP property of the fully discretized scheme

Now we show that the fully discretized scheme (3.32) is AP. As $\varepsilon\to 0$, step (3.32c) implies

$$f_{j,k}^{n+1}=\sum_{l=1}^{3}\tilde{w}_l M_k^{n+1}(\tilde{x}_{j,l})=M_{j,k}^{n+1}.\tag{3.43}$$

Hence after one time step, the solution is projected to the local Maxwellian. For $n\geq 1$, replacing $f_{j,k}^n$ with $M_{j,k}^n$ in (3.32a) and taking the moments gives

$$\frac{\langle f_{j,\cdot}^*\phi\rangle-\langle M_{j,\cdot}^n\phi\rangle}{\Delta t}+\left\langle\frac{\hat{M}_{j+1/2,\cdot}^n-\hat{M}_{j-1/2,\cdot}^n}{\Delta x}\phi\right\rangle=0,\tag{3.44}$$

where $\hat{M}_{j+1/2,k}$ is the numerical flux approximating $v_kM(x,v_k)$ at $x=x_{j+1/2}$. Finally, using (3.41), we have

$$\frac{\langle M_{j,\cdot}^{n+1}\phi\rangle-\langle M_{j,\cdot}^n\phi\rangle}{\Delta t}+\left\langle\frac{\hat{M}_{j+1/2,\cdot}^n-\hat{M}_{j-1/2,\cdot}^n}{\Delta x}\phi\right\rangle=0.\tag{3.45}$$

This is a fully discretized kinetic scheme for the limiting Euler equations. Thus the scheme (3.32) is AP.

4 Generalization to the hyperbolic relaxation system

The general framework presented in this paper can also be generalized to other problems that have a similar structure, for instance, the hyperbolic relaxation system. We give one example here.

The Broadwell model [5] is a simple discrete velocity kinetic model:

$$\begin{cases}\partial_t f_+ + \partial_x f_+ = \dfrac{1}{\varepsilon}(f_0^2-f_+f_-),\\ \partial_t f_0 = -\dfrac{1}{\varepsilon}(f_0^2-f_+f_-),\\ \partial_t f_- - \partial_x f_- = \dfrac{1}{\varepsilon}(f_0^2-f_+f_-),\end{cases}\tag{4.1}$$

where ε is the mean free path, f_+, f_0, and f_- denote the mass densities of particles with speed 1, 0, and -1, respectively. The model can be written equivalently in terms of moment variables:

$$\begin{cases} \partial_t \rho + \partial_x m = 0, \\ \partial_t m + \partial_x z = 0, \\ \partial_t z + \partial_x m = \dfrac{1}{2\varepsilon}(\rho^2 + m^2 - 2\rho z), \end{cases} \tag{4.2}$$

where $\rho := f_+ + 2f_0 + f_-$, $m := f_+ - f_-$, and $z := f_+ + f_-$. From (4.2), it is clear that when $\varepsilon \to 0$, $z \to \frac{\rho^2+m^2}{2\rho}$. This, substituted into the first two equations, yields a closed hyperbolic system, an analog of the Euler limit:

$$\begin{cases} \partial_t \rho + \partial_x m = 0, \\ \partial_t m + \partial_x \left(\dfrac{\rho^2 + m^2}{2\rho} \right) = 0. \end{cases} \tag{4.3}$$

Similarly as the BGK model, it would be desirable to have a high order scheme for (4.1) that is AP (can capture the limit (4.3) without resolving ε) as well as maintains the positivity of the solution (f_+, f_0, and f_- need to be non-negative by their physical meaning). We mention that [6] proposed a second-order AP scheme for the Broadwell model but it is not positivity-preserving.

We now define $f = (f_+, f_0, f_-)^T$, $\mathcal{T}(f) = (-\partial_x f_+, 0, \partial_x f_-)^T$, and $\mathcal{Q}(f) = (f_0^2 - f_+ f_-, -(f_0^2 - f_+ f_-), f_0^2 - f_+ f_-)^T$. Then (4.1) falls into the general form (2.1). Define the matrix P as

$$\begin{pmatrix} 1 & 2 & 1 \\ 1 & 0 & -1 \\ 1 & 0 & 1 \end{pmatrix}, \tag{4.4}$$

then $Pf = (\rho, m, z)^T$, and $P\mathcal{Q}(f) = (0, 0, \frac{1}{2}(\rho^2 + m^2 - 2\rho z))^T$.

In order to apply the general framework, we need to verify the operators $\mathcal{T}$ and $\mathcal{Q}$ satisfy the assumptions given in Section 2.1. The transport operator $\mathcal{T}$ can definitely satisfy the positivity condition (2.2) provided a positivity-preserving spatial discretization is used. To analyze the positivity conditions for $\mathcal{Q}$, first notice that $(I - b\mathcal{Q})f = g$, upon multiplication of P on both sides from the left, implies

$$\begin{aligned} &\rho_f = \rho_g, \\ &m_f = m_g, \\ &z_f - \frac{b}{2}(\rho_f^2 + m_f^2 - 2\rho_f z_f) = z_g, \end{aligned} \tag{4.5}$$

from which one has

$$z_f = \left(\frac{b}{2}(\rho_f^2 + m_f^2) + z_g \right) / (1 + b\rho_f). \tag{4.6}$$

If $g \geq 0$, or equivalently, $\rho_g \geq z_g \geq |m_g|$, then, to check $f \geq 0$ for any $b \geq 0$, it suffices to check $\rho_f \geq z_f$ and $z_f \geq |m_f|$, which follow from

$$\rho_f - z_f = \frac{\frac{b}{2}(\rho_f^2 - m_f^2) + \rho_f - z_g}{1 + b\rho_f} = \frac{\frac{b}{2}(\rho_g^2 - m_g^2) + \rho_g - z_g}{1 + b\rho_g} \geq 0, \tag{4.7}$$

$$z_f - |m_f| = \frac{\frac{b}{2}(\rho_f - |m_f|)^2 + z_g - |m_f|}{1 + b\rho_f} = \frac{\frac{b}{2}(\rho_g - |m_g|)^2 + z_g - |m_g|}{1 + b\rho_g} \geq 0. \tag{4.8}$$

This proves (2.3). To show (2.4), notice that

$$\mathcal{Q}'(g)\mathcal{Q}(f) = -\rho_g \mathcal{Q}(f), \tag{4.9}$$

and (2.4) follows from (2.3) since $\rho_g \geq 0$. Finally, for (2.5),

$$(I + b\mathcal{Q}'(f)\mathcal{Q})f = h \iff (I - b\rho_f \mathcal{Q})f = h, \tag{4.10}$$

which upon multiplication of P on the left gives $\rho_f = \rho_h$. If $h \geq 0$, $\rho_f = \rho_h \geq 0$. Then (2.5) follows again from (2.3).

Therefore, the scheme (2.11)-(2.13) can be applied to the Broadwell model, resulting in a second-order, positivity-preserving scheme. A similar AP property as for the BGK equation can be proved straightforwardly using the (ρ, m, z) formulation (4.2). We omit the detail.

Finally, we briefly outline how to prove the entropy-decay property of the scheme when using the upwind spatial discretization. The entropy for the Broadwell model is defined by

$$S[f] = \Delta x \sum_k [f_{+,k} \log f_{+,k} + 2f_{0,k} \log f_{0,k} + f_{-,k} \log f_{-,k}], \tag{4.11}$$

where k is the spatial index. We show that $S[f^{n+1}] \leq S[f^n]$.

First, the transport part can be done in the same way as (3.26). For the collision part,

$$f^{(i)} = f^{(i)*} + \Delta t a_{ii} \frac{1}{\varepsilon} \mathcal{Q}(f^{(i)}), \tag{4.12}$$

the entropy inequality for this step, namely, $S[f^{(i)}] \leq S[f^{(i)*}]$, was proved in [6]. As for the last step

$$f^{n+1} = f^{(\nu)} + \alpha \Delta t^2 \frac{1}{\varepsilon^2} \rho_{f^*} \mathcal{Q}(f^{n+1}), \tag{4.13}$$

if $f^* = f^n$ or $f^{(i)}$, ρ_{f^*} is a known non-negative constant, and the proof for (4.12) implies $S[f^{n+1}] \leq S[f^{(\nu)}]$; if $f^* = f^{n+1}$, one first takes the moment of (4.13) (i.e., multiply P on both sides from the left) and gets

$$\rho_{f^{n+1}} = \rho_{f^{(\nu)}} \geq 0, \tag{4.14}$$

and then can obtain the same conclusion.

5 Numerical results

In this section we demonstrate numerically the properties of the proposed IMEX schemes. We will solve the 1D BGK equation (3.14) in $x \in [0, 2]$ with periodic boundary condition (except the test in Section 5.2, where the Dirichlet boundary condition is assumed), and in a large enough velocity domain $v \in [-|v|_{\max}, |v|_{\max}]$. The x-space is discretized into N_x cells with $\Delta x = \frac{2}{N_x}$. The v-space is discretized into N_v grid points with $\Delta v = \frac{2|v|_{\max}}{N_v}$. We fix the parameters $N_v = 150$ and $|v|_{\max} = 15$ such that the discretization error in v is much smaller than that in space and time. We will test the two IMEX schemes given in Section 2.6. For brevity, in the following we refer the scheme in Section 2.6.1 as scheme A, and the scheme in Section 2.6.2 as scheme ARS.

5.1 Accuracy test

We first verify the second-order accuracy of the proposed schemes. We expect that 1) in the kinetic regime $\varepsilon = O(1)$, both scheme A and scheme ARS are second-order accurate; 2) in the fluid regime $\varepsilon \ll 1$, for consistent initial data, both schemes exhibit second-order accuracy; for inconsistent initial data, scheme A is still second order while scheme ARS will degrade to first order (see Propositions 3.3 and 3.4).

We first consider inconsistent initial data

$$f(0,x,v) = 0.5M_{\rho,u,T} + 0.3M_{\rho,-0.5u,T}, \tag{5.1}$$

with

$$\rho = 1 + 0.2\sin(\pi x), \quad u = 1, \quad T = \frac{1}{1+0.2\sin(\pi x)}, \tag{5.2}$$

and compute the solution to time $t = 0.1$. We choose different values of ε, ranging from the kinetic regime ($\varepsilon = 1$) to the fluid regime ($\varepsilon = 1e-10$). We choose different Δx and set $\Delta t = 0.5\frac{\Delta t}{|v|_{\max}}$, i.e., fix the CFL number as 0.5, which guarantees both schemes are stable. (This CFL number is not small enough to guarantee positivity. We will consider the positivity-preserving property in the following test. For the same reason, the positivity-preserving limiters are turned off here.) Since the exact solution is not available, the numerical solution on a finer mesh $\Delta x/2$ is used as a reference solution to compute the error for the solution on the mesh of size Δx:

$$\text{error}_{\Delta t,\Delta x} := \|f_{\Delta t,\Delta x} - f_{\Delta t/2,\Delta x/2}\|_{L^2_{x,v}}. \tag{5.3}$$

The results are shown in Tables 1 and 2. In all the results, the spatial error dominates for small N_x, and the time error dominates for large N_x. One can clearly see that in the kinetic regime ($\varepsilon = 1, 1e-2$), both schemes are second order; in the fluid regime ($\varepsilon = 1e-8, 1e-10$), the scheme A is second order and the scheme ARS is first order, as expected.

We also solve the equation in the intermediate and fluid regimes with a consistent initial data

$$f(0,x,v) = M_{\rho,u,T}, \tag{5.4}$$

where ρ, u and T are the same as in (5.2). The results are shown in Tables 3 and 4. It is clear that in the fluid regime both schemes remain second-order accuracy.

Note that there is always some extent of order reduction in the intermediate regime $\varepsilon = O(\Delta t)$. The uniform accuracy of IMEX schemes is an open problem and we do not attempt to address this issue in the current work.

5.2 Positivity-preserving property

We now illustrate the positivity-preserving property of the scheme. Consider the initial data

$$f(0,x,v) = M_{\rho,u,T}, \tag{5.5}$$

with

$$(\rho,u,T) = \begin{cases} (1,0,1), & 0 \le x \le 1, \\ (0.125,0,0.25), & 1 < x \le 2. \end{cases} \tag{5.6}$$

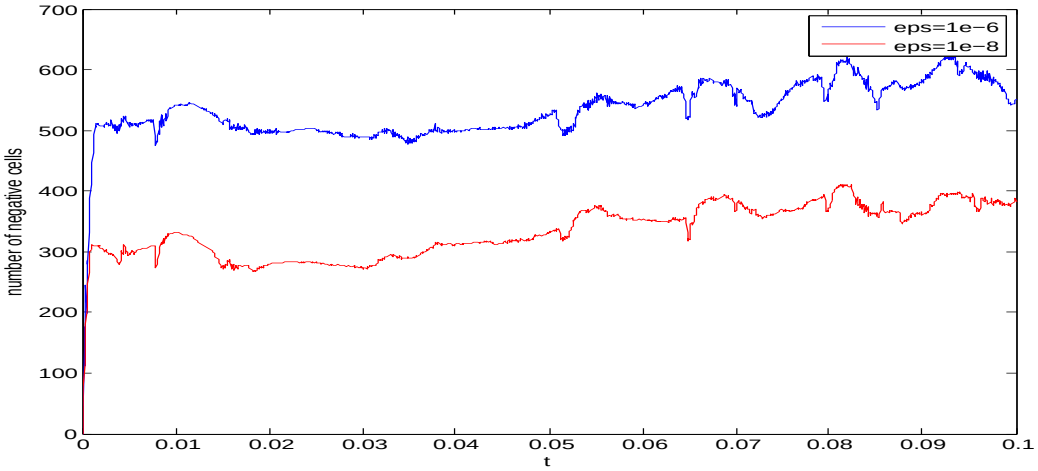

Figure 2: Total number of negative cells for the ARS(2,2,2) scheme during time evolution. Blue line: $\varepsilon = 1e-6$; Red line: $\varepsilon = 1e-8$.

With the positivity-preserving limiters, the CFL coefficient of the spatial discretization is $1/12$, that is, the constant $\mathcal{C}$ in (2.28) and (2.39) is $\frac{1}{12}\frac{\Delta x}{|v|_{\max}}$. In view of both time and spatial discretizations, we choose the time step as $\Delta t = \frac{1}{24}\frac{\Delta x}{|v|_{\max}}$ to satisfy the positivity CFL condition. We take $N_x = 80$.

The numerical solutions computed by both scheme A and scheme ARS exhibit no negative cell averages and are omitted here. As a comparison, we solve the same equation with the same initial data and spatial discretization, but using the ARS(2,2,2) scheme in time [1], which is a standard second-order accurate IMEX scheme with no positivity-preserving property. The number of negative cells (out of $80 \times 150 = 12000$ cells) is tracked and reported in Figure 2. One can see that a significant number of cell averages become negative in the fluid regime, if the time discretization is not positivity-preserving.

5.3 AP property

Finally, to illustrate the AP property, we solve the BGK equation in a mixed regime. We take $\varepsilon = \varepsilon(x)$ as follows:

$$\varepsilon(x) = \varepsilon_0 + (\tanh(1 - 11(x-1)) + \tanh(1 + 11(x-1))), \quad \varepsilon_0 = 1e-5, \tag{5.7}$$

as shown in Figure 3. The ε is chosen such that in the middle part of the domain, the problem is in the kinetic regime ($\varepsilon(x) = O(1)$); while in the left and right parts, the problem is in the fluid regime ($\varepsilon \approx 1e-5$). To handle this multiscale problem, one can use the domain decomposition approach, i.e., solve the BGK equation in the kinetic regime and the Euler equations in the fluid regime. But identifying the interface and coupling conditions between two regimes is a challenging task. An alternative approach is to solve the BGK equation exclusively in the entire domain. But to insure stability, an explicit scheme would require the time step to resolve the smallest value of ε which is extremely expensive. This is where the AP scheme shows its power:

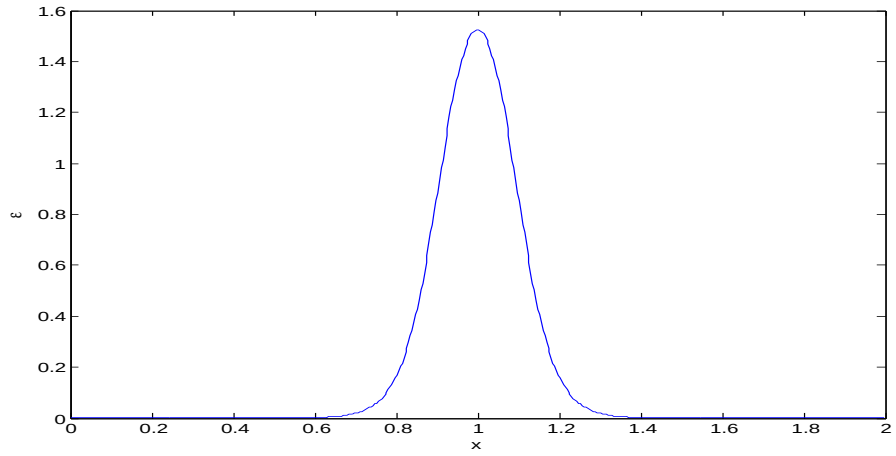

Figure 3: Profile of $\varepsilon(x)$ in a mixed regime problem.

it is a consistent scheme to the kinetic equation when $\varepsilon = O(1)$, and will automatically become a consistent scheme for the fluid equation when $\varepsilon \to 0$.

We take the same initial data as in (5.1)-(5.2) and solve the problem using scheme A and scheme ARS with $N_x = 40$. We compare the macroscopic quantities at time $t = 0.5$ with a reference solution computed by the explicit second-order SSP RK scheme [28] with $N_x = 80$. Note that for AP schemes, $\Delta t = \frac{1}{24}\frac{\Delta x}{|v|_{\max}} \approx 7e-5$; while for the explicit SSP scheme, $\Delta t = \frac{1}{240}\frac{\Delta x}{|v|_{\max}} \approx 7e-6$ which needs to resolve ε. One can see that the solutions of AP schemes agree well with the reference solution in Figure 4.

6 Conclusion

We have introduced a family of second-order IMEX schemes for the BGK equation. The method is asymptotic-preserving: it reduces to a second-order explicit RK scheme for the compressible Euler equations as the Knudsen number $\varepsilon \to 0$. Meanwhile, the method is positivity-preserving, provided the time step satisfies a CFL condition independent of ε. The method also satisfies an entropy-decay property when coupled with proper spatial discretizations. The key idea is to add a correction step to the conventional IMEX RK schemes. Due to the special structure of the BGK operator, this step maintains both positivity and AP property, and is very easy to implement. We considered two types of commonly used IMEX RK schemes (one of type A and one of type ARS) and constructed two examples, one of each type respectively. We investigated, both analytically and numerically, the properties of the proposed schemes. Furthermore, we showed that it is possible to generalize the method to some hyperbolic relaxation system such as the Broadwell model which demands positivity, and provided a strategy to extend the method to third order. Some future work include the construction of high-order asymptotic-preserving and positivity-preserving schemes for other kinetic models, for example, the Fokker-Planck equation, the full Boltzmann equation, etc.

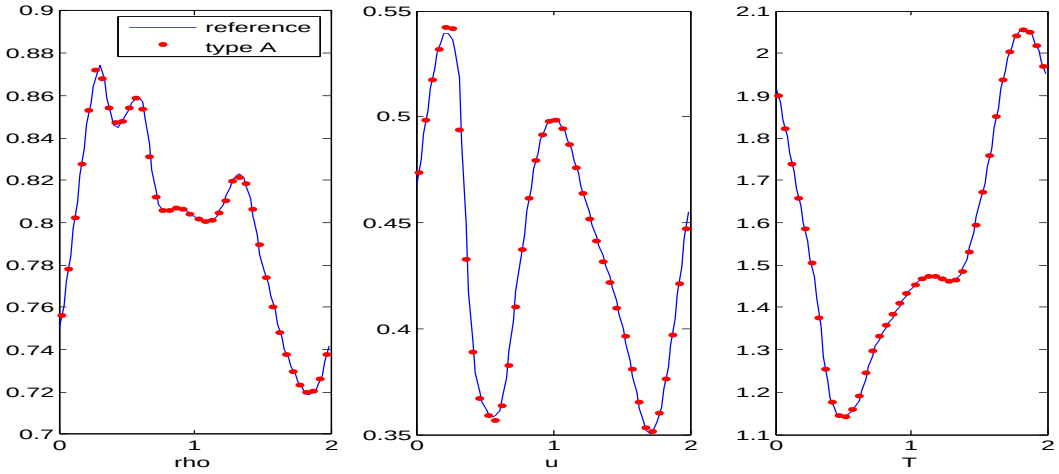

Figure 4: Mixed regime problem. Left to right: density ρ, velocity u and temperature T. Solid line: reference solution computed by the second-order SSP RK scheme. Dots: solution computed by scheme A. The result of scheme ARS is omitted since it is indistinguishable from that of scheme A in the picture.

Appendix: Extension to third order

In this Appendix, we briefly present the strategy to extend the proposed method to third order.

To this end, we need to derive order conditions of the scheme (2.11)-(2.13) up to third order. We consider the cases that $f^* = f^n$, $\tilde{f}^{n+1}$ or f^{n+1}.

Substituting (2.14) into (2.11), one obtains

$$\begin{aligned}
f^{(i)} &= f^n + \Delta t \sum_{j=1}^{i-1} \tilde{a}_{ij} \mathcal{T}(f^n + \Delta t \tilde{c}_j \mathcal{T}(f^n) + \Delta t c_j \mathcal{Q}(f^n)) \\
&\quad + \Delta t \sum_{j=1}^{i} a_{ij} \mathcal{Q}(f^n + \Delta t \tilde{c}_j \mathcal{T}(f^n) + \Delta t c_j \mathcal{Q}(f^n)) + O(\Delta t^3) \\
&= f^n + \Delta t \sum_{j=1}^{i-1} \tilde{a}_{ij} [\mathcal{T}(f^n) + \Delta t \mathcal{T}'(f^n)(\tilde{c}_j \mathcal{T}(f^n) + c_j \mathcal{Q}(f^n))] \\
&\quad + \Delta t \sum_{j=1}^{i} a_{ij} [\mathcal{Q}(f^n) + \Delta t \mathcal{Q}'(f^n)(\tilde{c}_j \mathcal{T}(f^n) + c_j \mathcal{Q}(f^n))] + O(\Delta t^3) \\
&= f^n + \Delta t [\tilde{c}_i \mathcal{T}(f^n) + c_i \mathcal{Q}(f^n)] + \Delta t^2 \left[\sum_{j=1}^{i-1} \tilde{a}_{ij} \mathcal{T}'(f^n)(\tilde{c}_j \mathcal{T}(f^n) + c_j \mathcal{Q}(f^n)) \right. \\
&\quad \left. + \sum_{j=1}^{i} a_{ij} \mathcal{Q}'(f^n)(\tilde{c}_j \mathcal{T}(f^n) + c_j \mathcal{Q}(f^n)) \right] + O(\Delta t^3).
\end{aligned} \tag{6.1}$$

Substituting it into (2.12) yields

$$
\begin{aligned}
\tilde{f}^{n+1} &= f^n + \Delta t \sum_{i=1}^{\nu} \tilde{w}_i \mathcal{T} \left\{ f^n + \Delta t[\tilde{c}_i \mathcal{T}(f^n) + c_i \mathcal{Q}(f^n)] + \Delta t^2 \left[\sum_{j=1}^{i-1} \tilde{a}_{ij} \mathcal{T}'(f^n)(\tilde{c}_j \mathcal{T}(f^n) + c_j \mathcal{Q}(f^n)) \right.\right. \\
&\quad \left.\left. + \sum_{j=1}^{i} a_{ij} \mathcal{Q}'(f^n)(\tilde{c}_j \mathcal{T}(f^n) + c_j \mathcal{Q}(f^n)) \right] \right\} \\
&\quad + \Delta t \sum_{i=1}^{\nu} w_i \mathcal{Q} \left\{ f^n + \Delta t[\tilde{c}_i \mathcal{T}(f^n) + c_i \mathcal{Q}(f^n)] + \Delta t^2 \left[\sum_{j=1}^{i-1} \tilde{a}_{ij} \mathcal{T}'(f^n)(\tilde{c}_j \mathcal{T}(f^n) + c_j \mathcal{Q}(f^n)) \right.\right. \\
&\quad \left.\left. + \sum_{j=1}^{i} a_{ij} \mathcal{Q}'(f^n)(\tilde{c}_j \mathcal{T}(f^n) + c_j \mathcal{Q}(f^n)) \right] \right\} + O(\Delta t^4) \\
&= f^n + \Delta t \left[(\sum_{i=1}^{\nu} \tilde{w}_i) \mathcal{T}(f^n) + (\sum_{i=1}^{\nu} w_i) \mathcal{Q}(f^n) \right] + \Delta t^2 \left[(\sum_{i=1}^{\nu} \tilde{w}_i \tilde{c}_i) \mathcal{T}'(f^n) \mathcal{T}(f^n) + (\sum_{i=1}^{\nu} \tilde{w}_i c_i) \mathcal{T}'(f^n) \mathcal{Q}(f^n) \right. \\
&\quad \left. + (\sum_{i=1}^{\nu} w_i \tilde{c}_i) \mathcal{Q}'(f^n) \mathcal{T}(f^n) + (\sum_{i=1}^{\nu} w_i c_i) \mathcal{Q}'(f^n) \mathcal{Q}(f^n) \right] \\
&\quad + \Delta t^3 \left\{ \sum_{i=1}^{\nu} \sum_{j=1}^{i-1} [\tilde{w}_i \tilde{a}_{ij} \tilde{c}_j \mathcal{T}'(f^n) \mathcal{T}'(f^n) \mathcal{T}(f^n) + \tilde{w}_i \tilde{a}_{ij} c_j \mathcal{T}'(f^n) \mathcal{T}'(f^n) \mathcal{Q}(f^n)] \right. \\
&\quad + \sum_{i=1}^{\nu} \sum_{j=1}^{i} [\tilde{w}_i a_{ij} \tilde{c}_j \mathcal{T}'(f^n) \mathcal{Q}'(f^n) \mathcal{T}(f^n) + \tilde{w}_i a_{ij} c_j \mathcal{T}'(f^n) \mathcal{Q}'(f^n) \mathcal{Q}(f^n)] \\
&\quad + \frac{1}{2} \sum_{i=1}^{\nu} [\tilde{w}_i \tilde{c}_i \tilde{c}_i \mathcal{T}''(f^n)(\mathcal{T}(f^n), \mathcal{T}(f^n)) + 2\tilde{w}_i \tilde{c}_i c_i \mathcal{T}''(f^n)(\mathcal{T}(f^n), \mathcal{Q}(f^n)) + \tilde{w}_i c_i c_i \mathcal{T}''(f^n)(\mathcal{Q}(f^n), \mathcal{Q}(f^n))] \\
&\quad + \sum_{i=1}^{\nu} \sum_{j=1}^{i-1} [w_i \tilde{a}_{ij} \tilde{c}_j \mathcal{Q}'(f^n) \mathcal{T}'(f^n) \mathcal{T}(f^n) + w_i \tilde{a}_{ij} c_j \mathcal{Q}'(f^n) \mathcal{T}'(f^n) \mathcal{Q}(f^n)] \\
&\quad + \sum_{i=1}^{\nu} \sum_{j=1}^{i} [w_i a_{ij} \tilde{c}_j \mathcal{Q}'(f^n) \mathcal{Q}'(f^n) \mathcal{T}(f^n) + w_i a_{ij} c_j \mathcal{Q}'(f^n) \mathcal{Q}'(f^n) \mathcal{Q}(f^n)] \\
&\quad \left. + \frac{1}{2} \sum_{i=1}^{\nu} [w_i \tilde{c}_i \tilde{c}_i \mathcal{Q}''(f^n)(\mathcal{T}(f^n), \mathcal{T}(f^n)) + 2w_i \tilde{c}_i c_i \mathcal{Q}''(f^n)(\mathcal{T}(f^n), \mathcal{Q}(f^n)) + w_i c_i c_i \mathcal{Q}''(f^n)(\mathcal{Q}(f^n), \mathcal{Q}(f^n))] \right\} \\
&\quad + O(\Delta t^4),
\end{aligned}
\tag{6.2}
$$

where the second-order Fréchet derivative is given by

$$
\mathcal{Q}''(g)(f_1, f_2) = \lim_{\delta_1, \delta_2 \to 0} \frac{\mathcal{Q}(g + \delta_1 f_1 + \delta_2 f_2) - \mathcal{Q}(g + \delta_1 f_1) - \mathcal{Q}(g + \delta_2 f_2) + \mathcal{Q}(g)}{\delta_1 \delta_2}, \tag{6.3}
$$

which is a symmetric bilinear operator.

In the case $f^* = f^n$, (2.13) gives (using the first order conditions $\sum_{i=1}^{\nu} \tilde{w}_i = \sum_{i=1}^{\nu} w_i = 1$)

$$
\begin{aligned}
f^{n+1} &= \tilde{f}^{n+1} - \alpha \Delta t^2 \mathcal{Q}'(f^n) \mathcal{Q}(f^n + \Delta t(\mathcal{T}(f^n) + \mathcal{Q}(f^n))) + O(\Delta t^4) \\
&= \tilde{f}^{n+1} - \alpha \Delta t^2 \mathcal{Q}'(f^n) \mathcal{Q}(f^n) - \alpha \Delta t^3 [\mathcal{Q}'(f^n) \mathcal{Q}'(f^n) \mathcal{T}(f^n) + \mathcal{Q}'(f^n) \mathcal{Q}'(f^n) \mathcal{Q}(f^n)] + O(\Delta t^4),
\end{aligned}
\tag{6.4}
$$

while in the case $f^* = \tilde{f}^{n+1}$ or f^{n+1},

$$\begin{aligned} f^{n+1} &= \tilde{f}^{n+1} - \alpha\Delta t^2 \mathcal{Q}'(f^n + \Delta t(\mathcal{T}(f^n) + \mathcal{Q}(f^n)))\mathcal{Q}(f^n + \Delta t(\mathcal{T}(f^n) + \mathcal{Q}(f^n))) + O(\Delta t^4) \\ &= \tilde{f}^{n+1} - \alpha\Delta t^2 \mathcal{Q}'(f^n)\mathcal{Q}(f^n) - \alpha\Delta t^3[\mathcal{Q}''(f^n)(\mathcal{T}(f^n), \mathcal{Q}(f^n)) + \mathcal{Q}''(f^n)(\mathcal{Q}(f^n), \mathcal{Q}(f^n)) \\ &\quad + \mathcal{Q}'(f^n)\mathcal{Q}'(f^n)\mathcal{T}(f^n) + \mathcal{Q}'(f^n)\mathcal{Q}'(f^n)\mathcal{Q}(f^n)] + O(\Delta t^4). \end{aligned} \tag{6.5}$$

On the other hand, if we Taylor expand the exact solution of (2.1) around time t^n, we have

$$\begin{aligned} f^{n+1}_{\text{exact}} &= f^n + \Delta t[\mathcal{T}(f^n) + \mathcal{Q}(f^n)] + \frac{1}{2}\Delta t^2[\mathcal{T}'(f^n)\mathcal{T}(f^n) + \mathcal{T}'(f^n)\mathcal{Q}(f^n) + \mathcal{Q}'(f^n)\mathcal{T}(f^n) + \mathcal{Q}'(f^n)\mathcal{Q}(f^n)] \\ &\quad + \frac{1}{6}\Delta t^3[\mathcal{T}''(f^n)(\mathcal{T}(f^n), \mathcal{T}(f^n)) + 2\mathcal{T}''(f^n)(\mathcal{Q}(f^n), \mathcal{T}(f^n)) + \mathcal{T}''(f^n)(\mathcal{Q}(f^n), \mathcal{Q}(f^n)) \\ &\quad + \mathcal{Q}''(f^n)(\mathcal{T}(f^n), \mathcal{T}(f^n)) + 2\mathcal{Q}''(f^n)(\mathcal{Q}(f^n), \mathcal{T}(f^n)) + \mathcal{Q}''(f^n)(\mathcal{Q}(f^n), \mathcal{Q}(f^n)) \\ &\quad + (\mathcal{T} + \mathcal{Q})'(f^n)(\mathcal{T} + \mathcal{Q})'(f^n)(\mathcal{T} + \mathcal{Q})(f^n)] + O(\Delta t^4). \end{aligned} \tag{6.6}$$

Comparing (6.6) with (6.4) or (6.5), we obtain the following order conditions:

$$\begin{aligned} &\sum_{i,j} \tilde{w}_i \tilde{a}_{ij} \tilde{c}_j = \sum_{i,j} \tilde{w}_i \tilde{a}_{ij} c_j = \sum_{i,j} \tilde{w}_i a_{ij} \tilde{c}_j = \sum_{i,j} \tilde{w}_i a_{ij} c_j \\ =&\sum_{i,j} w_i \tilde{a}_{ij} \tilde{c}_j = \sum_{i,j} w_i \tilde{a}_{ij} c_j = \sum_{i,j} w_i a_{ij} \tilde{c}_j - \alpha = \sum_{i,j} w_i a_{ij} c_j - \alpha = \frac{1}{6}, \\ &\sum_{i} \tilde{w}_i \tilde{c}_i \tilde{c}_i = \sum_{i} \tilde{w}_i \tilde{c}_i c_i = \sum_{i} \tilde{w}_i c_i c_i \\ =&\sum_{i} w_i \tilde{c}_i \tilde{c}_i = \sum_{i} w_i \tilde{c}_i c_i = \sum_{i} w_i c_i c_i = \frac{1}{3}, \end{aligned} \tag{6.7}$$

in the case $f^* = f^n$, and

$$\begin{aligned} &\sum_{i,j} \tilde{w}_i \tilde{a}_{ij} \tilde{c}_j = \sum_{i,j} \tilde{w}_i \tilde{a}_{ij} c_j = \sum_{i,j} \tilde{w}_i a_{ij} \tilde{c}_j = \sum_{i,j} \tilde{w}_i a_{ij} c_j \\ =&\sum_{i,j} w_i \tilde{a}_{ij} \tilde{c}_j = \sum_{i,j} w_i \tilde{a}_{ij} c_j = \sum_{i,j} w_i a_{ij} \tilde{c}_j - \alpha = \sum_{i,j} w_i a_{ij} c_j - \alpha = \frac{1}{6}, \\ &\sum_{i} \tilde{w}_i \tilde{c}_i \tilde{c}_i = \sum_{i} \tilde{w}_i \tilde{c}_i c_i = \sum_{i} \tilde{w}_i c_i c_i \\ =&\sum_{i} w_i \tilde{c}_i \tilde{c}_i = \sum_{i} w_i \tilde{c}_i c_i - \alpha = \sum_{i} w_i c_i c_i - 2\alpha = \frac{1}{3}, \end{aligned} \tag{6.8}$$

in the case $f^* = \tilde{f}^{n+1}$ or f^{n+1}.

Note that compared to the standard IMEX-RK (third) order conditions [25], the only difference is the terms containing α.

Therefore, in order to get a third-order positivity-preserving scheme, one only needs to find RK coefficients in (2.11)-(2.13) such that they satisfy the order conditions (2.19) and (6.7) (resp. (6.8)) as well as the positivity conditions derived in Section 2.5 ($\alpha \geq 0$ and (2.27) for type A and GSA schemes or (2.38) for type ARS and GSA schemes). This can be done via a computer program.

	Nx=10	Nx=20	order	Nx=40	order	Nx=80	order	Nx=160	order	Nx=320	order	Nx=640	order	Nx=1280	order
ϵ=1e+00	5.60e-04	5.91e-05	3.25	4.33e-06	3.77	2.11e-07	4.36	1.29e-08	4.03	2.94e-09	2.13	7.42e-10	1.99	1.86e-10	2.00
ϵ=1e-02	4.67e-04	4.63e-05	3.33	7.11e-06	2.70	1.67e-06	2.09	4.22e-07	1.99	1.06e-07	1.99	2.67e-08	1.99	6.69e-09	2.00
ϵ=1e-04	4.67e-04	3.62e-05	3.69	3.31e-06	3.45	2.92e-06	0.18	3.03e-06	-0.05	2.79e-06	0.12	1.52e-06	0.88	5.46e-07	1.47
ϵ=1e-06	4.67e-04	3.65e-05	3.68	2.46e-06	3.89	1.09e-07	4.49	6.58e-09	4.06	4.71e-09	0.48	8.30e-09	-0.82	1.44e-08	-0.80
ϵ=1e-08	4.67e-04	3.65e-05	3.68	2.46e-06	3.89	1.10e-07	4.49	6.28e-09	4.13	1.45e-09	2.11	3.67e-10	1.98	9.20e-11	2.00
ϵ=1e-10	4.67e-04	3.65e-05	3.68	2.46e-06	3.89	1.10e-07	4.49	6.28e-09	4.13	1.45e-09	2.11	3.68e-10	1.98	9.20e-11	2.00

Table 1: Accuracy test. Scheme A. Inconsistent initial data.

	Nx=10	Nx=20	order	Nx=40	order	Nx=80	order	Nx=160	order	Nx=320	order	Nx=640	order	Nx=1280	order
ϵ=1e+00	5.60e-04	5.91e-05	3.25	4.33e-06	3.77	2.12e-07	4.36	1.22e-08	4.11	2.71e-09	2.17	6.83e-10	1.99	1.71e-10	2.00
ϵ=1e-02	5.02e-04	9.82e-05	2.35	2.89e-05	1.76	8.14e-06	1.83	2.17e-06	1.91	5.59e-07	1.95	1.42e-07	1.98	3.58e-08	1.99
ϵ=1e-04	4.70e-04	3.71e-05	3.66	4.82e-06	2.94	2.35e-06	1.04	2.00e-06	0.23	2.94e-06	-0.56	2.99e-06	-0.02	1.76e-06	0.76
ϵ=1e-06	4.70e-04	3.71e-05	3.66	4.79e-06	2.95	2.21e-06	1.12	1.12e-06	0.99	5.58e-07	1.00	2.79e-07	1.00	1.40e-07	1.00
ϵ=1e-08	4.70e-04	3.71e-05	3.66	4.79e-06	2.95	2.21e-06	1.12	1.12e-06	0.99	5.58e-07	1.00	2.79e-07	1.00	1.40e-07	1.00
ϵ=1e-10	4.70e-04	3.71e-05	3.66	4.79e-06	2.95	2.21e-06	1.12	1.12e-06	0.99	5.58e-07	1.00	2.79e-07	1.00	1.40e-07	1.00

Table 2: Accuracy test. Scheme ARS. Inconsistent initial data.

	Nx=10	Nx=20	order	Nx=40	order	Nx=80	order	Nx=160	order	Nx=320	order	Nx=640	order	Nx=1280	order
ϵ=1e-04	1.04e-03	1.01e-04	3.38	8.05e-06	3.64	4.17e-06	0.95	4.76e-06	-0.19	4.46e-06	0.10	2.40e-06	0.89	8.54e-07	1.49
ϵ=1e-06	1.05e-03	1.01e-04	3.37	7.64e-06	3.73	4.79e-07	4.00	1.83e-08	4.71	6.16e-09	1.58	1.11e-08	-0.85	1.94e-08	-0.80
ϵ=1e-08	1.05e-03	1.01e-04	3.37	7.64e-06	3.73	4.79e-07	3.99	1.82e-08	4.72	1.52e-09	3.58	4.03e-10	1.92	1.03e-10	1.97
ϵ=1e-10	1.05e-03	1.01e-04	3.37	7.64e-06	3.73	4.79e-07	3.99	1.82e-08	4.72	1.52e-09	3.58	4.03e-10	1.92	1.02e-10	1.98

Table 3: Accuracy test. Scheme A. Consistent initial data.

	Nx=10	Nx=20	order	Nx=40	order	Nx=80	order	Nx=160	order	Nx=320	order	Nx=640	order	Nx=1280	order
ϵ=1e-04	1.04e-03	1.01e-04	3.37	7.62e-06	3.73	1.24e-06	2.62	2.65e-06	-1.09	4.51e-06	-0.77	4.56e-06	-0.02	2.67e-06	0.78
ϵ=1e-06	1.05e-03	1.01e-04	3.37	7.64e-06	3.73	4.79e-07	3.99	1.82e-08	4.72	1.60e-09	3.50	9.94e-10	0.69	1.67e-09	-0.75
ϵ=1e-08	1.05e-03	1.01e-04	3.37	7.64e-06	3.73	4.79e-07	3.99	1.82e-08	4.72	1.52e-09	3.58	4.03e-10	1.92	1.02e-10	1.97
ϵ=1e-10	1.05e-03	1.01e-04	3.37	7.64e-06	3.73	4.79e-07	3.99	1.82e-08	4.72	1.52e-09	3.58	4.03e-10	1.92	1.02e-10	1.98

Table 4: Accuracy test. Scheme ARS. Consistent initial data.

References

[1] U. Ascher, S. Ruuth, and R. Spiteri. Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations. *Appl. Numer. Math.*, 25:151–167, 1997.

[2] C. Bardos, F. Golse, and D. Levermore. Fluid dynamic limits of kinetic equations. I. Formal derivations. *J. Stat. Phys.*, 63:323–344, 1991.

[3] P. L. Bhatnagar, E. P. Gross, and M. Krook. A model for collision processes in gases. I. Small amplitude processes in charged and neutral one-component systems. *Phys. Rev.*, 94:511–525, 1954.

[4] S. Boscarino, L. Pareschi, and G. Russo. Implicit-explicit Runge-Kutta schemes for hyperbolic systems and kinetic equations in the diffusion limit. *SIAM J. Sci. Comput.*, 35:A22–A51, 2013.

[5] J. Broadwell. Shock structure in a simple discrete velocity gas. *Phys. Fluids*, 7:1013–1037, 1964.

[6] R. E. Caflisch, S. Jin, and G. Russo. Uniformly accurate schemes for hyperbolic systems with relaxation. *SIAM J. Numer. Anal.*, 34:246–281, 1997.

[7] C. Cercignani. *The Boltzmann Equation and Its Applications*. Springer-Verlag, New York, 1988.

[8] C. Cercignani. *Rarefied Gas Dynamics: From Basic Concepts to Actual Calculations*. Cambridge University Press, Cambridge, 2000.

[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 J. Numer. Anal.*, 53:2008–2029, 2015.

[10] F. Coron and B. Perthame. Numerical passage from kinetic to fluid equations. *SIAM J. Numer. Anal.*, 28:26–42, 1991.

[11] G. Dimarco and L. Pareschi. Asymptotic preserving implicit-explicit Runge-Kutta methods for nonlinear kinetic equations. *SIAM J. Numer. Anal.*, 51:1064–1087, 2013.

[12] F. Filbet and S. Jin. A class of asymptotic-preserving schemes for kinetic equations and related problems with stiff sources. *J. Comput. Phys.*, 229:7625–7648, 2010.

[13] S. Gottlieb, D. Ketcheson, and C.-W. Shu. *Strong Stability Preserving Runge-Kutta and Multistep Time Discretizations*. World Scientific, 2011.

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

[15] E. Hairer. Order conditions for numerical methods for partitioned ordinary differential equations. *Numer. Math.*, 36:431–445, 1981.

[16] I. Higueras. Strong stability for additive Runge-Kutta methods. *SIAM J. Numer. Anal.*, 44:1735–1758, 2006.

[17] I. Higueras and T. Roldan. Positivity-preserving and entropy-decaying IMEX methods. *Monografias del Seminario Matematico Garcia de Galdeano*, 33:129–136, 2006.

[18] J. Hu, S. Jin, and Q. Li. Asymptotic-preserving schemes for multiscale hyperbolic and kinetic equations. In R. Abgrall and C.-W. Shu, editors, *Handbook of Numerical Methods for Hyperbolic Problems*, chapter 5, pages 103–129. North-Holland, 2017.

[19] J. Huang and C.-W. Shu. A second-order asymptotic-preserving and positivity-preserving discontinuous Galerkin scheme for the Kerr-Debye model. *Math. Models Methods Appl. Sci.*, 27:549–579, 2017.

[20] S. Jin. Asymptotic preserving (AP) schemes for multiscale kinetic and hyperbolic equations: a review. *Riv. Mat. Univ. Parma*, 3:177–216, 2012.

[21] A. Jüngel. *Transport Equations for Semiconductors*, volume 773 of *Lecture Notes in Physics*. Springer, Berlin, 2009.

[22] C. Kennedy and M. Carpenter. Additive Runge-Kutta schemes for convection-diffusion-reaction equations. *Appl. Numer. Math.*, 44:139–181, 2003.

[23] R. J. LeVeque. *Finite Difference Methods for Ordinary and Partial Differential Equations*. SIAM, Philadelphia, PA, 2007.

[24] G. Naldi, L. Pareschi, and G. Toscani, editors. *Mathematical Modeling of Collective Behavior in Socio-Economic and Life Sciences*. Birkhäuser Basel, 2010.

[25] L. Pareschi and G. Russo. Implicit-Explicit Runge-Kutta methods and applications to hyperbolic systems with relaxation. *J. Sci. Comput.*, 25:129–155, 2005.

[26] S. Pieraccini and G. Puppo. Implicit-Explicit schemes for BGK kinetic equations. *J. Sci. Comput.*, 1:1–28, 2007.

[27] C.-W. Shu. Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws. In *Advanced numerical approximation of nonlinear hyperbolic equations*, pages 325–432. Springer, 1998.

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

[29] C. Villani. A review of mathematical topics in collisional kinetic theory. In S. Friedlander and D. Serre, editors, *Handbook of Mathematical Fluid Mechanics*, volume I, pages 71–305. North-Holland, 2002.

[30] X. Zhang. On positivity-preserving high order discontinuous Galerkin schemes for compressible Navier–Stokes equations. *J. Comput. Phys.*, 328:301–343, 2017.

[31] X. Zhang and C.-W. Shu. On maximum-principle-satisfying high order schemes for scalar conservation laws. *J. Comput. Phys.*, 229:3091–3120, 2010.

[32] X. Zhang and C.-W. Shu. On positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations on rectangular meshes. *J. Comput. Phys.*, 229:8918–8934, 2010.

[33] X. Zhang and C.-W. Shu. Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: survey and new developments. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, page rspa20110153. The Royal Society, 2011.