MULTISCALE MODEL. SIMUL.
Vol. 17, No. 4, pp. 1123–1146

A SECOND-ORDER ASYMPTOTIC-PRESERVING AND POSITIVITY-PRESERVING EXPONENTIAL RUNGE–KUTTA METHOD FOR A CLASS OF STIFF KINETIC EQUATIONS[*]

JINGWEI HU[†] AND RUIWEN SHU[‡]

Abstract. We introduce a second-order time discretization method for stiff kinetic equations. The method is asymptotic-preserving—can capture the Euler limit without numerically resolving the small Knudsen number—and positivity-preserving—can preserve the non-negativity of the solution which is a probability density function for arbitrary Knudsen numbers. The method is based on a new formulation of the exponential Runge–Kutta method and can be applied to a large class of stiff kinetic equations including the Bhatnagar–Gross–Krook equation (relaxation type), the Fokker–Planck equation (diffusion type), and even the full Boltzmann equation (nonlinear integral type). Furthermore, we show that when coupled with suitable spatial discretizations the fully discrete scheme satisfies an entropy-decay property. Various numerical results are provided to demonstrate the theoretical properties of the method.

Key words. stiff kinetic equation, exponential Runge–Kutta method, asymptotic-preserving, positivity-preserving, entropy-decay

AMS subject classifications. 82C40, 65L04, 35Q31, 35Q84, 35Q20, 65L06, 65F60

DOI. 10.1137/18M1226774

1. Introduction.

Kinetic equations describe the nonequilibrium dynamics of a gas or system comprised of a large number of particles. In multiscale modeling hierarchy, they serve as a bridge that connects microscopic Newtonian mechanics and macroscopic continuum mechanics. In this paper, we are concerned with the following class of kinetic equations:

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

where $f=f(t,x,v)$ is the one-particle probability density function (PDF) of time t, position x, and particle velocity v (d_x and d_v are the dimensions of x and v, respectively). $\mathcal{Q}$ is the collision operator that acts only in the velocity space and models the interactions between particles. Examples of $\mathcal{Q}$ include: the Boltzmann collision operator (a nonlinear integral operator) [6], the Bhatnagar–Gross–Krook (BGK) operator (a relaxation type operator) [3], the kinetic Fokker–Planck operator (a diffusion type operator) [26], among others. Finally, ε is the Knudsen number defined as the ratio of the mean free path and typical length scale. The magnitude of ε indicates the degree of rarefaction of the system. When ε is small, collisions happen very frequently so that the system is close to the fluid regime. In fact, one can derive the compressible Euler equations from (1.1) as the leading-order asymptotics by sending $\varepsilon\to 0$.

When ε is small, numerically solving (1.1) is challenging due to the stiff collision term on the right-hand side. Any explicit time discretization would suffer from severe

[*] Received by the editors November 14, 2018; accepted for publication (in revised form) July 26, 2019; published electronically October 11, 2019.
https://doi.org/10.1137/18M1226774

Funding: This 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.

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

[‡] Department of Mathematics, University of Maryland, College Park, MD 20742 (rshu@cscamm.umd.edu).

stability constraint (time step Δt has to be $O(\varepsilon)$). As such, schemes that can remove this constraint are highly desirable. The so-called asymptotic-preserving (AP) scheme [18] is exactly designed for this kind of problem: it solves the kinetic equation without resolving small scales (Δt can be chosen independent of ε), yet when $\varepsilon \to 0$ while keeping Δt fixed, it automatically becomes a macroscopic fluid solver, i.e., a consistent discretization to the limiting Euler equations (see [19, 14] for a comprehensive review of AP schemes).

The AP property is certainly a desired feature when handling multiscale kinetic equations, especially in the near fluid regime. However, most of AP schemes require some implicit treatment or reformulation of the equation such that the positivity of the solution is lost during the construction. This is unphysical since f is a PDF, and sometimes even causes the simulation to break down. The design of high order (at least second order) schemes that are both AP and positivity-preserving turns out to be highly nontrivial and needs to be handled in a problem-dependent basis. Recently, we developed a family of second-order AP and positivity-preserving schemes for the stiff BGK equation [15]. The method is based on the implicit-explicit (IMEX) Runge–Kutta framework plus a key correction step utilizing the special structure of the BGK operator. It also works for some hyperbolic systems but is limited to relaxation type operators.

In this paper, we propose a more general time discretization method based on a new exponential Runge–Kutta formulation that can be applied to a large class of stiff kinetic equations including the BGK, the Fokker–Planck, and even the full Boltzmann equations. To summarize, our method possesses the following features:

- The scheme is second-order accurate in the kinetic regime $\varepsilon = O(1)$;
- The scheme is AP: For fixed Δt, when $\varepsilon \to 0$, it reduces to a second-order scheme for the limiting Euler equations (in fact, the limiting scheme can be made as the optimal second-order strong-stability-preserving (SSP) Runge–Kutta method, i.e., the Heun's method [10, 9]);
- The scheme is positivity-preserving for any $\varepsilon \geq 0$: If $f^n \geq 0$, then $f^{n+1} \geq 0$;
- The time step of the scheme is only constrained by the transport part and can be chosen the same as in the forward Euler method;
- The scheme satisfies an entropy-decay property when coupled with suitable spatial discretizations.

The rest of this paper is organized as follows. In section 2, we construct the numerical method for the general kinetic equation (1.1) without specifying the collision operator. The emphasis is to make the method second order and positivity preserving. In section 3, we consider the application of the method to specific kinetic equations and discuss its AP property. In section 4, we address the issue of solving the homogeneous equation ((1.1) without transport term) which is an important building block of the proposed method. A comparison with existing similar methods is given in section 5. In section 6, we prove the entropy-decay property of the method when coupled with suitable spatial discretizations. Some remarks regarding the spatial and velocity domain discretizations are given in section 7. Numerical examples are presented in section 8. The paper is concluded in section 9.

2. A new exponential Runge-Kutta method for general stiff kinetic equations.

We now present the procedure to construct the new exponential Runge–Kutta method. Since the method is quite general and can be applied to a large class of kinetic equations, we will start with (1.1) without specifying the collision operator and derive a scheme that is both second-order accurate and positivity preserving.

Then in section 3, we will consider specific collision operators and discuss the AP property of the scheme as this latter part is problem dependent.

To begin with, let us introduce the following notation: For the autonomous ODE

$$\frac{\mathrm{d}}{\mathrm{d}t}f = A(f), \quad f|_{t=t_0} = g, \tag{2.1}$$

where A is an operator, being either linear or nonlinear, we use $\varphi_A^s g$, $s \geq 0$ to represent its exact solution at $t = t_0 + s$. In the following, whenever it is clear from the context, the subscript A is dropped for notational simplicity.

We now consider an ODE resulting from the semidiscretization of (1.1) (only space x is discretized while time t and velocity v are left continuous):

$$\frac{\mathrm{d}}{\mathrm{d}t}f = \mathcal{T}(f) + \frac{1}{\varepsilon}\mathcal{Q}(f). \tag{2.2}$$

Here $\mathcal{T}(f)$ is a discretized operator for the transport term $-v \cdot \nabla_x f$ and $\mathcal{Q}(f)$ is the collision operator which may take various forms depending on the application. We assume the operators $\mathcal{T}(f)$ and $\mathcal{Q}(f)$ are positivity preserving. To be precise,

- for $\mathcal{T}(f)$, we assume
  (2.3)
  $f \geq 0 \implies f + a\Delta t\,\mathcal{T}(f) \geq 0 \quad \forall$ constant a such that (s.t.) $0 \leq a\Delta t \leq \Delta t_{\mathrm{FE}}$,

  where Δt_{FE} is the maximum time step allowance such that the forward Euler method is positivity preserving;
- for $\mathcal{Q}(f)$, we assume the solution to the homogeneous equation

  $$\frac{\mathrm{d}}{\mathrm{d}t}f = \mathcal{Q}(f) \tag{2.4}$$

  satisfies $f \geq 0\ \forall\ t \geq t_0$ if the initial data $f|_{t=t_0} = g \geq 0$. In other words,

  $$g \geq 0 \implies \varphi^s g \geq 0 \quad \forall \text{ constant } s \geq 0. \tag{2.5}$$

Remark 2.1. The condition (2.3) can be easily satisfied if a positivity-preserving spatial discretization is used, as was done in [15]. The condition (2.5) is a theoretical property that holds for any kinetic equations.

We are ready to construct the numerical method for (2.2). We propose an exponential Runge–Kutta scheme of the following form:

$$\begin{aligned} f^{(0)} &= \varphi^{a_0\Delta t/\varepsilon} f^n, \\ f^{(1)} &= \varphi^{a_1\Delta t/\varepsilon}\left(f^{(0)} + b_1\Delta t\mathcal{T}(f^{(0)})\right), \\ f^{(2)} &= f^{(1)} + b_2\Delta t\mathcal{T}(f^{(1)}), \\ f^{n+1} &= \varphi^{a_2\Delta t/\varepsilon}\left[w f^{(2)} + (1-w)\varphi^{(1-a_2)\Delta t/\varepsilon} f^n\right], \end{aligned} \tag{2.6}$$

where the constants a_0, a_1, a_2, b_1, b_2, and w are to be determined. Note that in the above, $\varphi^{a_0\Delta t/\varepsilon}$ denotes the solution map by solving the homogeneous equation (2.4) with time step $a_0\Delta t/\varepsilon$, or equivalently, $\varphi^{a_0\Delta t/\varepsilon} f^n$ is the exact solution to (2.4) at time $t^n + a\Delta t/\varepsilon$ with initial condition f^n. Other terms of the similar form in (2.6) are understood analogously. Of course this step relies on the specific collision operator $\mathcal{Q}$, and we will get back to this issue in section 4. For the time being, we assume that this solution map is known.

With the previous assumptions on $\mathcal{T}$ and $\mathcal{Q}$, it is easy to see the following.

PROPOSITION 2.2. *The scheme* (2.6) *is positivity-preserving, i.e., if $f^n \geq 0$, then $f^{n+1} \geq 0$ provided*

$$a_0, a_1, b_1, b_2 \geq 0, \quad 0 \leq a_2, w \leq 1, \tag{2.7}$$

under the Courant–Friedrichs–Lewy (CFL) condition

$$\Delta t \leq \frac{\Delta t_{FE}}{\max(b_1, b_2)}, \tag{2.8}$$

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

We next derive the conditions for (2.6) to be second order in the kinetic regime. Without loss of generality, we assume $\varepsilon = 1$.

First of all, given the solution $f^n = f(t_n)$, if we Taylor expand the exact solution of (2.2) at t_{n+1} around t_n, we have

$$\begin{aligned} f^{n+1}_{\text{exact}} &= f^n + \Delta t(\partial_t f)^n + \frac{1}{2}\Delta t^2(\partial_{tt} f)^n + O(\Delta t^3) \\ &= f^n + \Delta t[\mathcal{T}(f^n) + \mathcal{Q}(f^n)] \\ &\quad + \frac{1}{2}\Delta t^2[\mathcal{T}'(f^n)\mathcal{T}(f^n) + \mathcal{T}'(f^n)\mathcal{Q}(f^n) + \mathcal{Q}'(f^n)\mathcal{T}(f^n) + \mathcal{Q}'(f^n)\mathcal{Q}(f^n)] + O(\Delta t^3), \end{aligned} \tag{2.9}$$

where $\mathcal{Q}'$, $\mathcal{T}'$ are the Fréchet derivative of $\mathcal{Q}$ and $\mathcal{T}$, given by

$$\mathcal{Q}'(f)g = \lim_{\delta \to 0} \frac{\mathcal{Q}(f + \delta g) - \mathcal{Q}(f)}{\delta}, \tag{2.10}$$

and similarly for $\mathcal{T}$.

On the other hand, given $f^n = f(t_n)$, if we Taylor expand the exact solution of (2.4) at $t_n + \Delta t$ around t_n, we have

$$\begin{aligned} \varphi^{\Delta t} f^n &= f^n + \Delta t(\partial_t f)^n + \frac{1}{2}\Delta t^2(\partial_{tt} f)^n + O(\Delta t^3) \\ &= f^n + \Delta t\mathcal{Q}(f^n) + \frac{1}{2}\Delta t^2\mathcal{Q}'(f^n)\mathcal{Q}(f^n) + O(\Delta t^3). \end{aligned} \tag{2.11}$$

Using this in the first equation of (2.6), we have

$$f^{(0)} = f^n + a_0\Delta t\mathcal{Q}(f^n) + \frac{1}{2}a_0^2\Delta t^2\mathcal{Q}'(f^n)\mathcal{Q}(f^n) + O(\Delta t^3). \tag{2.12}$$

Continuing the Taylor expansion of $f^{(1)}$, $f^{(2)}$, and f^{n+1} in (2.6), we have

$$\begin{aligned} f^{(1)} &= (f^{(0)} + b_1\Delta t\mathcal{T}(f^{(0)})) + a_1\Delta t\mathcal{Q}(f^{(0)} + b_1\Delta t\mathcal{T}(f^{(0)})) \\ &\quad + \frac{1}{2}a_1^2\Delta t^2\mathcal{Q}'(f^{(0)} + b_1\Delta t\mathcal{T}(f^{(0)}))\mathcal{Q}(f^{(0)} + b_1\Delta t\mathcal{T}(f^{(0)})) + O(\Delta t^3) \\ &= f^n + \Delta t((a_0 + a_1)\mathcal{Q}(f^n) + b_1\mathcal{T}(f^n)) \\ &\quad + \Delta t^2\left(b_1 a_0\mathcal{T}'(f^n)\mathcal{Q}(f^n) + a_1 b_1\mathcal{Q}'(f^n)\mathcal{T}(f^n) + \frac{1}{2}(a_0 + a_1)^2\mathcal{Q}'(f^n)\mathcal{Q}(f^n)\right) \\ &\quad + O(\Delta t^3). \end{aligned} \tag{2.13}$$

$$
\begin{aligned}
f^{(2)} &= f^{(1)} + b_2\Delta t\mathcal{T}(f^{(1)}) \\
&= f^n + \Delta t((a_0+a_1)\mathcal{Q}(f^n) + b_1\mathcal{T}(f^n)) \\
&\quad + \Delta t^2\left(b_1a_0\mathcal{T}'(f^n)\mathcal{Q}(f^n) + a_1b_1\mathcal{Q}'(f^n)\mathcal{T}(f^n) + \frac{1}{2}(a_0+a_1)^2\mathcal{Q}'(f^n)\mathcal{Q}(f^n)\right) \\
&\quad + b_2\Delta t\mathcal{T}(f^n) + b_2\Delta t^2\mathcal{T}'(f^n)((a_0+a_1)\mathcal{Q}(f^n) + b_1\mathcal{T}(f^n)) + O(\Delta t^3) \\
&= f^n + \Delta t((a_0+a_1)\mathcal{Q}(f^n) + (b_1+b_2)\mathcal{T}(f^n)) \\
&\quad + \Delta t^2\Big(b_1b_2\mathcal{T}'(f^n)\mathcal{T}(f^n) + (b_1a_0+b_2a_0+b_2a_1)\mathcal{T}'(f^n)\mathcal{Q}(f^n) \\
&\quad + a_1b_1\mathcal{Q}'(f^n)\mathcal{T}(f^n) + \frac{1}{2}(a_0+a_1)^2\mathcal{Q}'(f^n)\mathcal{Q}(f^n)\Big) + O(\Delta t^3).
\end{aligned}
\tag{2.14}
$$

$$
\begin{aligned}
&wf^{(2)} + (1-w)\varphi^{(1-a_2)\Delta t}f^n = w\Big[f^n + \Delta t((a_0+a_1)\mathcal{Q}(f^n) + (b_1+b_2)\mathcal{T}(f^n)) \\
&\quad + \Delta t^2\Big(b_1b_2\mathcal{T}'(f^n)\mathcal{T}(f^n) + (b_2a_1+b_1a_0+b_2a_0)\mathcal{T}'(f^n)\mathcal{Q}(f^n) \\
&\quad + a_1b_1\mathcal{Q}'(f^n)\mathcal{T}(f^n) + \frac{1}{2}(a_0+a_1)^2\mathcal{Q}'(f^n)\mathcal{Q}(f^n)\Big)\Big] \\
&\quad + (1-w)\left[f^n + (1-a_2)\Delta t\mathcal{Q}(f^n) + \frac{1}{2}(1-a_2)^2\Delta t^2\mathcal{Q}'(f^n)\mathcal{Q}(f^n)\right] + O(\Delta t^3) \\
&= f^n + \Delta t[(w(a_0+a_1) + (1-w)(1-a_2))\mathcal{Q}(f^n) + w(b_1+b_2)\mathcal{T}(f^n)] \\
&\quad + \Delta t^2\Big[wb_1b_2\mathcal{T}'(f^n)\mathcal{T}(f^n) + w(b_2a_1+b_1a_0+b_2a_0)\mathcal{T}'(f^n)\mathcal{Q}(f^n) + wa_1b_1\mathcal{Q}'(f^n)\mathcal{T}(f^n) \\
&\quad + \frac{1}{2}(w(a_0+a_1)^2 + (1-w)(1-a_2)^2)\mathcal{Q}'(f^n)\mathcal{Q}(f^n)\Big] + O(\Delta t^3).
\end{aligned}
\tag{2.15}
$$

Finally,

$$
\begin{aligned}
f^{n+1} &= f^n + \Delta t[(w(a_0+a_1) + (1-w)(1-a_2))\mathcal{Q}(f^n) + w(b_1+b_2)\mathcal{T}(f^n)] \\
&\quad + \Delta t^2\Big[wb_1b_2\mathcal{T}'(f^n)\mathcal{T}(f^n) + w(b_2a_1+b_1a_0+b_2a_0)\mathcal{T}'(f^n)\mathcal{Q}(f^n) \\
&\quad + wa_1b_1\mathcal{Q}'(f^n)\mathcal{T}(f^n) + \frac{1}{2}(w(a_0+a_1)^2 + (1-w)(1-a_2)^2)\mathcal{Q}'(f^n)\mathcal{Q}(f^n)\Big] \\
&\quad + a_2\Delta t[\mathcal{Q}(f^n) + \Delta t\mathcal{Q}'(f^n)((w(a_0+a_1) + (1-w)(1-a_2))\mathcal{Q}(f^n) \\
&\quad + w(b_1+b_2)\mathcal{T}(f^n))] + \frac{1}{2}a_2^2\Delta t^2\mathcal{Q}'(f^n)\mathcal{Q}(f^n) + O(\Delta t^3) \\
&= f^n + \Delta t[(w(a_0+a_1+a_2) + (1-w))\mathcal{Q}(f^n) + w(b_1+b_2)\mathcal{T}(f^n)] \\
&\quad + \Delta t^2\Big[wb_1b_2\mathcal{T}'(f^n)\mathcal{T}(f^n) + w(b_2a_1+b_1a_0+b_2a_0)\mathcal{T}'(f^n)\mathcal{Q}(f^n) \\
&\quad + w(a_1b_1+a_2b_2+a_2b_1)\mathcal{Q}'(f^n)\mathcal{T}(f^n) + \frac{1}{2}(w(a_0+a_1+a_2)^2 \\
&\quad + (1-w))\mathcal{Q}'(f^n)\mathcal{Q}(f^n)\Big] + O(\Delta t^3).
\end{aligned}
\tag{2.16}
$$

Comparing (2.9) and (2.16), we arrive at the following order conditions:

$$\begin{aligned}
&w(a_0+a_1+a_2)+(1-w)=1; \quad w(b_1+b_2)=1; \quad wb_1b_2=\frac{1}{2};\\
&w(b_2a_1+b_1a_0+b_2a_0)=\frac{1}{2}; \quad w(a_1b_1+a_2b_2+a_2b_1)=\frac{1}{2};\\
&w(a_0+a_1+a_2)^2+(1-w)=1.
\end{aligned} \tag{2.17}$$

Further simplification yields the following.

PROPOSITION 2.3. *The scheme* (2.6) *is second-order accurate for* $\varepsilon = O(1)$ *provided*

$$a_0+a_1+a_2=1, \tag{2.18}$$

$$w(b_1+b_2)=1, \tag{2.19}$$

$$wb_1b_2=\frac{1}{2}, \tag{2.20}$$

$$w(b_2a_1+(b_1+b_2)a_0)=\frac{1}{2}. \tag{2.21}$$

Combining the positivity conditions and order conditions found in Propositions 2.2 and 2.3, one can obtain a second-order positivity-preserving scheme for (2.2). To find a set of parameters satisfying these conditions, first notice that (2.19) and (2.20) imply b_1, b_2 are the solutions of the quadratic equation

$$b^2-\frac{1}{w}b+\frac{1}{2w}=0, \tag{2.22}$$

whose solutions are given by

$$b_{1,2}=\frac{1}{1\pm\sqrt{1-2w}} \quad \text{for } 0<w\leq\frac{1}{2}. \tag{2.23}$$

In order to obtain the best CFL condition (minimize $\max(b_1,b_2)$ in (2.8)), we choose

$$w=\frac{1}{2}, \quad b_1=b_2=1; \tag{2.24}$$

hence the CFL condition (2.8) is the same as the forward Euler method. Then (2.18) and (2.21) reduce to

$$a_0+a_1+a_2=1, \quad a_0=a_2. \tag{2.25}$$

To insure positivity, we only need additionally $a_0, a_1 \geq 0$, $0 \leq a_2 \leq 1$ (see (2.7)). However, to obtain a good AP property, we require

$$a_0, a_1 > 0, \quad 0 < a_2 < 1. \tag{2.26}$$

This will be further elaborated in section 3. One choice of a_0, a_1, a_2 is

$$a_0=a_1=a_2=\frac{1}{3}. \tag{2.27}$$

Remark 2.4. For $a_0, a_1 > 0$ and $0 < a_2 < 1$, (2.6) would require 4 times evaluation of the operator φ^s in each time step. However, similar to the Strang splitting, one can combine the operator $\varphi^{a_2\Delta t/\varepsilon}$ in the last stage of the nth step with the operator $\varphi^{a_0\Delta t/\varepsilon}$ in the first stage of the $(n+1)$th step, so that effectively one only needs 3 times of such evaluations in each time step.

Remark 2.5. Note that in (2.6) (with the choice (2.24)–(2.26)), if one sets $\mathcal{T} = 0$, then the scheme becomes $f^{n+1} = \varphi^{\Delta t/\varepsilon} f^n$, which is the exact solution to the homogeneous equation $\partial_t f = \frac{1}{\varepsilon}\mathcal{Q}(f)$. On the other hand, if one sets $\mathcal{Q} = 0$, then the scheme just becomes Heun's method applied to the purely transport equation $\partial_t f = \mathcal{T}(f)$, which is optimal among all second-order explicit SSP Runge–Kutta schemes of two stages [10, 9].

3. Application to specific kinetic equations and AP property. By now, we have obtained a second-order positivity-preserving scheme ((2.6) with coefficients satisfying (2.24)–(2.26)) for the general stiff kinetic equation (2.2). In this section, we apply the scheme to some specific kinetic equations and discuss its AP property.

We will consider (2.2) with the following collision operators:

- The BGK operator [3], a simple relaxation type operator used to mimic the complicated Boltzmann collision operator:

  $$\mathcal{Q}(f) = \eta(\mathcal{M}[f] - f), \tag{3.1}$$

  where $\mathcal{M}[f]$ is the Maxwellian defined by

  $$\mathcal{M}[f] = \frac{\rho}{(2\pi T)^{\frac{d_v}{2}}} \exp\left(-\frac{|v-u|^2}{2T}\right), \tag{3.2}$$

  with the density ρ, bulk velocity u, and temperature T given by the moments of f,

  $$\rho = \int_{\mathbb{R}^{d_v}} f \,\mathrm{d}v, \quad u = \frac{1}{\rho}\int_{\mathbb{R}^{d_v}} f \,\mathrm{d}v, \quad T = \frac{1}{d_v \rho}\int_{\mathbb{R}^{d_v}} f|v-u|^2 \,\mathrm{d}v, \tag{3.3}$$

  and η is some positive function depending only on ρ and T.
- The ellipsoidal statistics-BGK (ES-BGK) operator [13], a generalized BGK model used to fit realistic values of the transport coefficients:

  $$\mathcal{Q}(f) = \eta(\mathcal{G}[f] - f), \tag{3.4}$$

  where $\mathcal{G}[f]$ is a Gaussian function defined by

  $$\mathcal{G}[f] = \frac{\rho}{\sqrt{\det(2\pi \bar{T})}} \exp\left(-\frac{1}{2}(v-u)^T \bar{T}^{-1}(v-u)\right), \tag{3.5}$$

  with ρ, u, and T given in (3.3) and

  $$\bar{T} = (1-\nu)TI + \nu\Theta, \quad \Theta = \frac{1}{\rho}\int_{\mathbb{R}^{d_v}} f(v-u)\otimes(v-u)\,\mathrm{d}v, \tag{3.6}$$

  where $-\frac{1}{2} \le \nu < 1$ is a parameter and I is the identity matrix. η is again some positive function of ρ and T.
- The Boltzmann collision operator [6], a fundamental equation in kinetic theory describing the binary collisions in a rarefied gas:

  $$\mathcal{Q}(f) = \int_{\mathbb{R}^{d_v}} \int_{S^{d_v-1}} B(v-v_*, \sigma)[f(v')f(v'_*) - f(v)f(v_*)] \,\mathrm{d}\sigma \,\mathrm{d}v_*, \tag{3.7}$$

  where v' and v'_* (postcollisional velocities) are defined in terms of v and v_* (precollisional velocities) as

  $$v' = \frac{v+v_*}{2} + \frac{|v-v_*|}{2}\sigma, \quad v'_* = \frac{v+v_*}{2} - \frac{|v-v_*|}{2}\sigma, \tag{3.8}$$

  with σ being a vector varying on the unit sphere S^{d_v-1}. B is the collision kernel characterizing the scattering rate and is a non-negative function.

- The kinetic Fokker–Planck operator [26], a kinetic model describing the drift and diffusion effects of particles:

$$\mathcal{Q}(f) = \nabla_v \cdot \left(\mathcal{M}[f] \nabla_v \frac{f}{\mathcal{M}[f]} \right), \tag{3.9}$$

  where $\mathcal{M}[f]$ is the same as in the BGK model. Using the definition (3.2), (3.9) can be written equivalently as

$$\mathcal{Q}(f) = \nabla_v \cdot \left(\nabla_v f + \frac{(v-u)}{T} f \right), \tag{3.10}$$

  with u and T given by (3.3). This is the more commonly seen drift-diffusion type equation in the literature.

All of the above collision operators $\mathcal{Q}$ satisfy the following properties that can be found in many standard textbooks [6, 26] with perhaps the ES-BGK operator as an exception whose proof is given in [1].

- Conservation of mass, momentum, and energy:

$$\langle \mathcal{Q}(f)\phi \rangle = 0, \quad \langle \cdot \phi \rangle := \int_{\mathbb{R}^{d_v}} \cdot \phi \, \mathrm{d}v, \quad \phi(v) = \left(1, v, \frac{|v|^2}{2}\right)^T \tag{3.11}$$

  for any function f.
  This implies that $\varphi^s g$, the solution to the homogeneous equation (2.4) at $t = t_0 + s$ with initial data $f|_{t=t_0} = g$, satisfies the conservation property

$$\langle (\varphi^s g)\phi \rangle = \langle g\phi \rangle \quad \forall s \geq 0. \tag{3.12}$$

- Decay of entropy

$$\int_{\mathbb{R}^{d_v}} \mathcal{Q}(f) \log f \, \mathrm{d}v \leq 0, \tag{3.13}$$

  and

$$\int_{\mathbb{R}^{d_v}} \mathcal{Q}(f) \log f \, \mathrm{d}v = 0 \Longleftrightarrow \mathcal{Q}(f) = 0 \ \Longleftrightarrow f = \mathcal{M}[f], \tag{3.14}$$

  where $\mathcal{M}[f]$ is the Maxwellian defined in (3.2).
  This implies that $\varphi^s g$, the solution to the homogeneous equation (2.4) at $t = t_0 + s$ with initial data $f|_{t=t_0} = g$, has the long time behavior

$$\lim_{s \to \infty} \varphi^s g = \mathcal{M}[g], \tag{3.15}$$

  i.e., $\varphi^s g$ approaches the Maxwellian determined by the moments of the initial condition.

Using these properties, it is easy to (formally) show that the spatially inhomogeneous equation (1.1) has the compressible Euler equations as the leading-order asymptotics when $\varepsilon \to 0$. Indeed, taking the moments $\langle \cdot \phi \rangle$ on both sides of (1.1), one obtains

$$\partial_t \langle f\phi \rangle + \nabla_x \cdot \langle f v \phi \rangle = 0, \tag{3.16}$$

by the conservation property of $\mathcal{Q}$. On the other hand, when $\varepsilon \to 0$, (1.1) formally implies $\mathcal{Q}(f) \to 0$; hence $f \to \mathcal{M}[f]$. Substituting $f = \mathcal{M}[f] := \mathcal{M}[U]$ into (3.16) yields

$$\partial_t U + \nabla_x \cdot \langle \mathcal{M}[U] v \phi \rangle = 0, \tag{3.17}$$

where we used the vector U to denote the first d_v+2 moments of f: $U=(\rho,\rho u,E)^T$ with $E=\frac{1}{2}\rho u^2+\frac{d_v}{2}\rho T$ being the total energy. The closed system (3.17) is nothing but the compressible Euler equations

$$\begin{cases} \partial_t\rho+\nabla_x\cdot(\rho u)=0, \\ \partial_t(\rho u)+\nabla_x\cdot(\rho u\otimes u+pI)=0, \\ \partial_t E+\nabla_x\cdot((E+p)u)=0, \end{cases} \tag{3.18}$$

where $p=\rho T$ is the pressure.

We now prove the AP property of the proposed scheme. Note that this proof is only formal as the rigorous transition from the Boltzmann equation to the compressible Euler equations even at the continuous level is an open problem.

PROPOSITION 3.1. *The scheme* (2.6) *(with coefficients satisfying* (2.24)–(2.26)*) applied to the stiff kinetic equation* (2.2) *with the collision operator* $\mathcal{Q}$ *being the BGK operator* (3.1)*, the ES-BGK operator* (3.4)*, the Boltzmann collision operator* (3.7)*, or the kinetic Fokker–Planck operator* (3.9) *is asymptotic-preserving, i.e., for any initial data and fixed* Δt*, in the limit* $\varepsilon\to 0$*,* (2.6) *becomes a second-order Heun's method applied to the limiting Euler system* (3.18)*. Furthermore,*

$$\lim_{\varepsilon\to 0} f^{n+1}=\mathcal{M}[U^{n+1}], \tag{3.19}$$

i.e., after each time step, f^{n+1} *is driven to its corresponding Maxwellian.*

Proof. First of all, taking the moments $\langle\,\cdot\,\phi\rangle$ on (2.6) and using (3.12), one obtains

$$\begin{aligned} &U^{(0)}=U^n, \\ &U^{(1)}=U^{(0)}+b_1\Delta t\langle\mathcal{T}(f^{(0)})\phi\rangle, \\ &U^{(2)}=U^{(1)}+b_2\Delta t\langle\mathcal{T}(f^{(1)})\phi\rangle, \\ &U^{n+1}=wU^{(2)}+(1-w)U^n. \end{aligned} \tag{3.20}$$

On the other hand, for $a_0,a_1,a_2>0$, using (3.15), it can be seen from (2.6) that as $\varepsilon\to 0$, $f^{(0)}$, $f^{(1)}$, and f^{n+1} are driven to their corresponding Maxwellian:

$$\begin{aligned} &f^{(0)}\to\mathcal{M}[U^n]=\mathcal{M}[U^{(0)}], \\ &f^{(1)}\to\mathcal{M}[U^{(0)}+b_1\Delta t\langle\mathcal{T}(f^{(0)})\phi\rangle]=\mathcal{M}[U^{(1)}], \\ &f^{n+1}\to\mathcal{M}[wU^{(2)}+(1-w)U^n]=\mathcal{M}[U^{n+1}]. \end{aligned} \tag{3.21}$$

Finally, substituting $f^{(0)}$ and $f^{(1)}$ into (3.20), one has

$$\begin{aligned} &U^{(1)}=U^n+b_1\Delta t\langle\mathcal{T}(M[U^n])\phi\rangle, \\ &U^{(2)}=U^{(1)}+b_2\Delta t\langle\mathcal{T}(M[U^{(1)}])\phi\rangle, \\ &U^{n+1}=wU^{(2)}+(1-w)U^n. \end{aligned} \tag{3.22}$$

With the coefficients (2.24) and $\mathcal{T}$ a discretized operator for $-v\cdot\nabla_x$, this is just a kinetic scheme for the limiting Euler equations (3.17) using Heun's method for time discretization. □

Remark 3.2. Note that the requirement for nonzero a_1, a_2, a_3 plays an important role here. In order for the scheme to have a nice AP property (working for any initial data, driving f to the corresponding Maxwellian after each time step, the limiting scheme maintains second-order accuracy, etc.), we need all these coefficients to be nondegenerate. See also the discussion in section 5.

4. Solving the homogeneous equation.

A key assumption we made in section 2 is that the solution to the homogeneous equation (2.4), or equivalently, the solution map φ^s, can be found exactly. Since we are content with the second-order scheme, this can be relaxed by finding an approximate solution, or an approximate map Φ^s, such that it is

- at least second-order accurate in time, i.e.,

$$\varphi^s g \approx \Phi^s g + O(s^3) \quad \text{for } s \sim O(\Delta t); \tag{4.1}$$

- positivity preserving, i.e.,

$$g \geq 0 \implies \Phi^s g \geq 0 \quad \forall \text{ constant } s \geq 0; \tag{4.2}$$

- AP, for which to hold we need Φ^s satisfy the same long time behavior as φ^s, i.e.,

$$\lim_{s\to\infty} \Phi^s g = \mathcal{M}[g]. \tag{4.3}$$

In the following, we will provide the strategy to construct the exact map φ^s or the approximate map Φ^s for all the kinetic equations discussed in section 3. Then using φ^s or Φ^s as a building block in (2.6), the whole scheme is completed.

4.1. The BGK equation.

For the homogeneous BGK equation

$$\partial_t f = \mathcal{Q}(f) = \eta(\mathcal{M}[f] - f), \quad f|_{t=t_0} = g, \tag{4.4}$$

since $\mathcal{Q}$ conserves mass, momentum, and energy, $\mathcal{M}[f] = \mathcal{M}[g]$ does not change with time, neither does η. Hence the solution at $t = t_0 + s$ can be found exactly:

$$\varphi^s g = e^{-\eta s} g + (1 - e^{-\eta s})\mathcal{M}[g]. \tag{4.5}$$

Remark 4.1. Instead of solving the BGK equation exactly, one can also find an approximate solution at $t = t_0 + s$ using the scheme

$$\begin{aligned} f^{(1)} &= g + s\mathcal{Q}(f^{(1)}), \\ f^1 &= f^{(1)} - \frac{1}{2}s^2\mathcal{Q}'(f^{(1)})\mathcal{Q}(f^1), \end{aligned} \tag{4.6}$$

and define the approximate map Φ^s as

$$\Phi^s g = f^1. \tag{4.7}$$

Note that the map Φ^s such defined is second-order accurate, positivity preserving, and AP. Using this Φ^s in (2.6) would give an IMEX Runge–Kutta type scheme, similar to our previous work [15].

4.2. The ES-BGK equation.

For the homogeneous ES-BGK equation

$$\partial_t f = \mathcal{Q}(f) = \eta(\mathcal{G}[f] - f), \quad f|_{t=t_0} = g, \tag{4.8}$$

since $\mathcal{Q}$ conserves mass, momentum, and energy, ρ, u, T do not change with time, neither does η. Taking the moment $\langle \cdot \frac{1}{\rho}(v-u)\otimes(v-u)\rangle$ on both sides of (4.8) gives

$$\partial_t\Theta = \eta\left(\frac{1}{\rho}\langle(v-u)\otimes(v-u)\mathcal{G}[f]\rangle - \Theta\right) = \eta(\bar{T} - \Theta) = \eta(1-\nu)(TI - \Theta), \tag{4.9}$$

whose solution is given by

$$\Theta(t_0+s) = e^{-\eta(1-\nu)s}\Theta(t_0) + (1-e^{-\eta(1-\nu)s})TI. \tag{4.10}$$

Hence

$$\bar{T}(t_0+s) = \nu e^{-\eta(1-\nu)s}\Theta(t_0) + (1-\nu e^{-\eta(1-\nu)s})TI. \tag{4.11}$$

On the other hand, (4.8) can be integrated to yield

$$\varphi^s g = f(t_0+s) = e^{-\eta s}g + \int_{t_0}^{t_0+s} \eta e^{-\eta(t_0+s-\tau)}\mathcal{G}[f(\tau)]\ \mathrm{d}\tau, \tag{4.12}$$

where $\mathcal{G}[f(\tau)]$ only depends on $\rho, u, \bar{T}(\tau)$. Rather than solving (4.12) exactly, we propose to use a quadrature to approximate the integral part. We adopt the two-point Gauss–Lobatto quadrature; that is,

$$\int_{t_0}^{t_0+s} \eta e^{-\eta(t_0+s-\tau)}\psi(\tau)\ \mathrm{d}\tau \approx w_1\psi(t_0) + w_2\psi(t_0+s), \tag{4.13}$$

where the weights w_1, w_2 are determined by requiring this approximation to be exact for $\psi(\tau) = 1, \tau$. A simple calculation gives

$$w_1 = \frac{1-e^{-\eta s}}{\eta s} - e^{-\eta s}, \quad w_2 = 1 - \frac{1-e^{-\eta s}}{\eta s}. \tag{4.14}$$

The quadrature in (4.13) has an error $O(s^3)$ for general functions.

Therefore, we approximate the solution in (4.12) as

$$\Phi^s g = e^{-\eta s}g + \left(\frac{1-e^{-\eta s}}{\eta s} - e^{-\eta s}\right)\mathcal{G}[\rho, u, \bar{T}(t_0)] + \left(1 - \frac{1-e^{-\eta s}}{\eta s}\right)\mathcal{G}[\rho, u, \bar{T}(t_0+s)], \tag{4.15}$$

with $\bar{T}(t_0+s)$ given by (4.11). This approximate map is positivity preserving since (4.15) is a convex combination of positive functions. It is AP since $s \to \infty$ one has $\bar{T}(t_0+s) \to TI$, thus $\mathcal{G}[\rho, u, \bar{T}(t_0+s)] \to \mathcal{M}[g]$; also three weights in (4.15) converge to 0, 0, 1, respectively, hence $\Phi^s g \to \mathcal{M}[g]$.

4.3. The Boltzmann equation.

For the homogeneous Boltzmann equation

$$\partial_t f = \mathcal{Q}(f) = \int_{\mathbb{R}^{d_v}}\int_{S^{d_v-1}} B(v-v_*, \sigma)[f(v')f(v'_*) - f(v)f(v_*)]\ \mathrm{d}\sigma\ \mathrm{d}v_*, \quad f|_{t=t_0} = g, \tag{4.16}$$

we adopt the exponential Runge–Kutta method introduced in [7] to find an approximate solution. Since $\mathcal{Q}$ conserves mass, momentum, and energy, $\mathcal{M}[f] = \mathcal{M}[g]$ does not change with time. Thus we can rewrite (4.16) as

$$\partial_t((f-\mathcal{M})e^{\mu t}) = (P(f) - \mu\mathcal{M})e^{\mu t}, \tag{4.17}$$

where $P(f) := \mathcal{Q}(f) + \mu f$, $\mu > 0$ being a constant, large enough so that $P(f) \geq 0$ (a simple choice is $\mu = \sup_v \int_{\mathbb{R}^{d_v}}\int_{S^{d_v-1}} B(v-v_*, \sigma)f(v_*)\ \mathrm{d}\sigma\ \mathrm{d}v_*$). Then, by applying the midpoint method to (4.17), one obtains a second-order scheme

$$\begin{aligned} (f^{(1)} - \mathcal{M})e^{\frac{\lambda}{2}} &= (g-\mathcal{M}) + \frac{\lambda}{2}\left(\frac{P(g)}{\mu} - \mathcal{M}\right), \\ (f^1 - \mathcal{M})e^{\lambda} &= (g-\mathcal{M}) + \lambda e^{\frac{\lambda}{2}}\left(\frac{P(f^{(1)})}{\mu} - \mathcal{M}\right), \end{aligned} \tag{4.18}$$

with $\lambda = \mu s$, which simplifies to

$$\begin{aligned} f^{(1)} &= e^{-\frac{\lambda}{2}} g + \left(1 - e^{-\frac{\lambda}{2}} - \frac{\lambda}{2} e^{-\frac{\lambda}{2}}\right)\mathcal{M} + \frac{\lambda}{2} e^{-\frac{\lambda}{2}} \frac{P(g)}{\mu}, \\ f^1 &= e^{-\lambda} g + \left(1 - e^{-\lambda} - \lambda e^{-\frac{\lambda}{2}}\right)\mathcal{M} + \lambda e^{-\frac{\lambda}{2}} \frac{P(f^{(1)})}{\mu}. \end{aligned} \tag{4.19}$$

Therefore, we choose f^1 to be the approximate solution at $t = t_0 + s$, i.e.,

$$\Phi^s g = f^1. \tag{4.20}$$

This approximate map is positivity preserving since both $f^{(1)}$ and f^1 are convex combinations of positive functions. It is AP since $s \to \infty$ implies $\lambda \to \infty$, thus $f^1 \to \mathcal{M}$.

Remark 4.2. As an alternative to the above described method, one can also use the so-called Wild sum expansion of the Boltzmann collision operator, which leads to another class of exponential method for the homogeneous Boltzmann equation possessing the desired properties; see, for instance [8, 23].

Remark 4.3. Here we did not address the issue of velocity domain discretization. To get a fully discrete scheme, one also needs an efficient and positivity-preserving solver for the Boltzmann collision operator (to evaluate the term $P(f)$ in the scheme). Available choices are the direct simulation Monte Carlo (DSMC) method [4], the discrete velocity method [22], or the recently proposed entropic Fourier method [5].

4.4. The kinetic Fokker–Planck equation. For the homogeneous kinetic Fokker–Planck equation

$$\partial_t f = \mathcal{Q}(f) = \nabla_v \cdot \left(\mathcal{M}[f] \nabla_v \frac{f}{\mathcal{M}[f]}\right), \quad f|_{t=t_0} = g, \tag{4.21}$$

since $\mathcal{Q}$ conserves mass, momentum, and energy, $\mathcal{M}[f] = \mathcal{M}[g]$ does not change with time. We adopt the approximation proposed in [20] to discretize $\mathcal{Q}$. Define $\tilde{f} = \frac{f}{\sqrt{\mathcal{M}}}$; then $\tilde{f}$ solves

$$\partial_t \tilde{f} = \tilde{\mathcal{Q}}(\tilde{f}) := \frac{1}{\sqrt{\mathcal{M}}} \nabla_v \cdot \left(\mathcal{M} \nabla_v \frac{\tilde{f}}{\sqrt{\mathcal{M}}}\right), \quad \tilde{f}|_{t=t_0} = \tilde{g} := \frac{g}{\sqrt{\mathcal{M}}}. \tag{4.22}$$

Hence

$$\varphi^s g = \sqrt{\mathcal{M}} \varphi^s_{\tilde{\mathcal{Q}}} \tilde{g}. \tag{4.23}$$

Now it suffices to approximate $\varphi^s_{\tilde{\mathcal{Q}}} \tilde{g}$. To do so, we first discretize the velocity and then use the matrix exponential to solve the resulting ODE system. Specifically, we truncate the velocity domain (consider $d_v = 1$ for simplicity) into a large enough interval $[-|v|_{\max}, |v|_{\max}]$ and discretize it into N_v grid points with $v_i = -|v|_{\max} + (i - 1/2)\Delta v$, $i = 1, \dots, N_v$, $\Delta v = 2|v|_{\max}/N_v$. Then the operator $\tilde{\mathcal{Q}}$ can be approximated by a tridiagonal symmetric matrix $\tilde{\mathcal{Q}}^h$ with the entries given by

$$\begin{aligned} \tilde{\mathcal{Q}}^h_{i,i} &= -\frac{1}{\Delta v^2} \frac{\sqrt{\mathcal{M}_{i-1}} + \sqrt{\mathcal{M}_{i+1}}}{\sqrt{\mathcal{M}_i}}, \\ \tilde{\mathcal{Q}}^h_{i,i-1} &= \tilde{\mathcal{Q}}^h_{i,i+1} = \frac{1}{\Delta v^2}, \end{aligned} \tag{4.24}$$

where $\mathcal{M}_i = \mathcal{M}(v_i)$. Define the vector $\tilde{g}^h = (\tilde{g}_1, \ldots, \tilde{g}_{N_v})^T$ with $\tilde{g}_i = \tilde{g}(v_i)$; then we approximate $\varphi^s g$ by

$$(\Phi^s g)_i = \sqrt{\mathcal{M}_i}\left(\exp(s\tilde{\mathcal{Q}}^h)\tilde{g}^h\right)_i, \tag{4.25}$$

where $\exp(s\tilde{\mathcal{Q}}^h)$ is the matrix exponential and can be computed very accurately by existing matrix exponential algorithms (we assume there is no error occurring at this step). The approximate map Φ^s is accurate in time as time evolution is solved exactly. It is positivity preserving since the off-diagonal entries of $\tilde{\mathcal{Q}}^h$ are non-negative. It is AP since $s \to \infty$ implies $\sqrt{\mathcal{M}_i}\left(\exp(s\tilde{\mathcal{Q}}^h)\tilde{g}^h\right)_i \to \mathcal{M}_i$. To see this, note that the discretization (4.24) for (4.22) is equivalent to the following:

$$\partial_t f_i = \frac{F_{i+1/2} - F_{i-1/2}}{\Delta v}, \quad F_{i+1/2} := \frac{\sqrt{\mathcal{M}_i \mathcal{M}_{i+1}}}{\Delta v}\left(\frac{f_{i+1}}{\mathcal{M}_{i+1}} - \frac{f_i}{\mathcal{M}_i}\right). \tag{4.26}$$

Define the discrete relative entropy as

$$H = \sum_i f_i \log \frac{f_i}{\mathcal{M}_i}\Delta v; \tag{4.27}$$

then

$$\begin{aligned} \partial_t H &= \sum_i \partial_t f_i \left(\log \frac{f_i}{\mathcal{M}_i} + 1\right)\Delta v = \sum_i (F_{i+1/2} - F_{i-1/2})\left(\log \frac{f_i}{\mathcal{M}_i} + 1\right) \\ &= -\sum_i F_{i+1/2}\left(\log \frac{f_{i+1}}{\mathcal{M}_{i+1}} - \log \frac{f_i}{\mathcal{M}_i}\right) \\ &= -\sum_i \frac{\sqrt{\mathcal{M}_i \mathcal{M}_{i+1}}}{\Delta v}\left(\frac{f_{i+1}}{\mathcal{M}_{i+1}} - \frac{f_i}{\mathcal{M}_i}\right)\left(\log \frac{f_{i+1}}{\mathcal{M}_{i+1}} - \log \frac{f_i}{\mathcal{M}_i}\right) \le 0, \end{aligned} \tag{4.28}$$

and the equality holds if and only if $f_i/\mathcal{M}_i$ is independent of i. This implies $f_i = \mathcal{M}_i$ by conservation.

Remark 4.4. Although we only discussed the kinetic Fokker–Planck operator in this paper, it is clear from the above analysis that for any Fokker–Planck operator, as long as its steady state $\mathcal{M}$ is explicitly known so that it can be written in the form

$$\partial_t f = \nabla_v \cdot \left(\mathcal{M}\nabla_v \frac{f}{\mathcal{M}}\right) \quad \text{and} \tag{4.29}$$

it can be treated in a similar fashion using matrix exponential.

5. Comparison with existing methods.

Searching the literature, there have been several methods available to solve the stiff kinetic equation (1.1) or equations of a similar structure. Therefore, we devote this section to a careful comparison of our method with some of the existing methods. For a general discussion on exponential integrators, the readers are referred to the review article [12].

- The following two existing second-order methods for (2.2) are special cases of (2.6):
  1. If one considers the second-order Strang splitting

$$\varphi^{\Delta t}_{\mathcal{T}+\mathcal{Q}} = \varphi^{\Delta t/2}_{\mathcal{Q}} \varphi^{\Delta t}_{\mathcal{T}} \varphi^{\Delta t/2}_{\mathcal{Q}} + O(\Delta t^3), \tag{5.1}$$

and discretizes $\varphi_{\mathcal{T}}^{\Delta t}$ by Heun's method, then one arrives at (2.6) with

$$a_0 = a_2 = \frac{1}{2}, \quad a_1 = 0, \quad b_1 = b_2 = 1, \quad w = \frac{1}{2}. \tag{5.2}$$

2. For the case $\mathcal{Q}(f) = -\mu f$ with $\mu > 0$ a constant, [16] rewrites (2.2) as

$$\partial_t(f e^{\frac{\mu}{\varepsilon} t}) = \mathcal{T}(f) e^{\frac{\mu}{\varepsilon} t} \tag{5.3}$$

and applies Heun's method to (5.3) directly. Then one arrives at (2.6) with

$$a_0 = a_2 = 0, \quad a_1 = 1, \quad b_1 = b_2 = 1, \quad w = \frac{1}{2}. \tag{5.4}$$

These two methods would suffer from order degeneracy in the fluid regime when applied to (1.1). In fact, in the first method $a_1 = 0$, and thus $f^{(1)}$ is not at local Maxwellian. Therefore, the flux term $b_2 \Delta t \langle \mathcal{T}(f^{(1)})\phi \rangle$ in (3.20) only approximates the flux in the limiting system up to first-order accuracy, which makes the limiting scheme first order. This order degeneracy of the Strang spliting was discovered already in an early work [17]. Similarly in the second method $a_0 = a_2 = 0$, and thus $f^{(0)} = f^n$ is not at local Maxwellian, which means the flux term $b_1 \Delta t \langle \mathcal{T}(f^{(0)})\phi \rangle$ in (3.20) is only first-order accurate in the limiting scheme. Moreover, even one starts with a consistent initial data, i.e., $f^n = \mathcal{M}[f^n]$, this method will not drive f^{n+1} to the local Maxwellian since $a_2 = 0$. Hence this error will pollute the solution as well in the next time step.

For the second method, [16] showed that the limiting scheme is second order with consistent initial data, in the case of $\mathcal{Q}(f) = -\mu f$ and $\mathcal{T}$ satisfying a maximum principle. Their proof is based on the following fact: If f is at local equilibrium (say $f - f^{\text{eq}} = O(\varepsilon)$), then $f + \Delta t \mathcal{T}(f)$ is also at local equilibrium $((f + \Delta t \mathcal{T}(f)) - (f^{\text{eq}} + \Delta t \mathcal{T}(f^{\text{eq}})) = O(\varepsilon))$. This is clearly not the case for (1.1), since generally speaking $f - \Delta t v \cdot \nabla_x f$ is $O(\Delta t)$ away from its local Maxwellian, even if f itself is at local Maxwellian.

- In [7], an exponential Runge–Kutta method was proposed for the homogeneous Boltzmann equation. This method is high order, AP, and positivity preserving. But it is extended to the nonhomogeneous equation (1.1) based on the Strang splitting, hence suffering from the order degeneracy as mentioned above.
- A nonsplitting version of the exponential Runge–Kutta method was proposed in [21] by applying an explicit Runge–Kutta scheme to a reformulated spatially inhomogeneous Boltzmann equation. There are two types of schemes proposed. One uses the time varying Maxwellian (called "ExpRK-V" in the paper) which cannot guarantee the positivity of f except for the density ρ. The other one is based on a fixed Maxwellian (called "ExpRK-F" in the paper) and can preserve the positivity of f provided a separate fluid equation is solved simultaneously and the underlying Runge–Kutta scheme satisfies certain conditions. However, the existence of such schemes (second or third order) that satisfy these conditions as well as AP remains to be discovered. Indeed, the second-order midpoint method and third-order Heun's method cannot satisfy these conditions, unlike what was claimed in [21].

To summarize, by a careful choice of the coefficients (2.24)–(2.26), our scheme (2.6) is different from any existing exponential Runge–Kutta type methods. It is second-order accurate, positivity preserving, and AP (capturing the Euler limit with second-order accuracy for any initial data).

6. Entropy-decay property.

In this section, we discuss the entropy-decay property of our scheme. First of all, we recall the following well-known result in kinetic theory. For the kinetic equation (1.1) with the collision operator being the BGK operator (3.1), the ES-BGK operator (3.4), the Boltzmann collision operator (3.7), or the kinetic Fokker–Planck operator (3.9), one has

$$\frac{\mathrm{d}}{\mathrm{d}t}\iint_{\mathbb{R}^{d_x}\times\mathbb{R}^{d_v}} f\log f \,\mathrm{d}v\,\mathrm{d}x \le 0 \tag{6.1}$$

under a periodic boundary condition in x. This is the famous H-theorem that says that the total entropy of the system is always nonincreasing.

We would like to show that our scheme (2.6) coupled with the first-order upwind discretization for the transport term and the homogeneous solvers discussed in section 4 satisfies a discrete entropy-decay property (a discrete analog of (6.1)). In order to do so, we assume the velocity space is *continuous*, in particular, this means the Fokker–Planck operator is not discretized and the solution to its homogeneous equation can be found analytically.

For simplicity, we consider (1.1) with $d_x = 1$, $d_v \ge 1$ (i.e., $x \in \mathbb{R}$, $v = (v_1, \dots) \in \mathbb{R}^{d_v}$):

$$\partial_t f + v_1 \partial_x f = \frac{1}{\varepsilon}\mathcal{Q}(f). \tag{6.2}$$

We truncate the velocity domain to a large enough box $D_v = [-|v|_{\max}, |v|_{\max}]^{d_v}$ and discretize the transport term by the upwind method (j is the spatial index):

$$(v_1\partial_x f)_j = \chi_{v_1\ge 0} v_1 \frac{f_j - f_{j-1}}{\Delta x} + \chi_{v_1<0} v_1 \frac{f_{j+1}-f_j}{\Delta x}. \tag{6.3}$$

Define the discrete entropy as

$$\mathcal{S}[f] := \Delta x \sum_j S[f_j], \quad S[f_j] := \int_{D_v} f_j \log f_j \,\mathrm{d}v; \tag{6.4}$$

then we claim that (2.6) satisfies a discrete entropy-decay property,

$$\mathcal{S}[f^{n+1}] \le \mathcal{S}[f^n], \tag{6.5}$$

provided the error coming from the velocity domain truncation is negligible (this assumption is reasonable since the distribution function often decays exponentially at large velocity).

To prove (6.5), we need two building blocks; one is the exponential step decays entropy, i.e., for either the exact φ^s or approximate Φ^s, one has

$$\mathcal{S}[\varphi^s g] \le \mathcal{S}[g] \ \text{ or } \ \mathcal{S}[\Phi^s g] \le \mathcal{S}[g] \quad \forall \text{ constant } s \ge 0; \tag{6.6}$$

the other is the transport step decays entropy, i.e., for step of the form $g = f + a\Delta t\mathcal{T}(f)$, one has

$$\mathcal{S}[g] \le \mathcal{S}[f], \quad \text{under the CFL condition } \Delta t \le \frac{\Delta x}{a|v|_{\max}}. \tag{6.7}$$

We now prove (6.6) and (6.7), respectively.

- For the BGK and Fokker–Planck operators, we have the exact φ^s; hence $\mathcal{S}[\varphi^s g] \leq \mathcal{S}[g]$ follows directly from the analytical result (3.13).
  For the ES-BGK operator, note that (4.15) is a convex combination of g, $\mathcal{G}[g]$ and $\mathcal{G}[\varphi^s g]$. One has $\mathcal{S}[\mathcal{G}[g]] \leq S[g]$ from [1]; hence $\mathcal{S}[\mathcal{G}[\varphi^s g]] \leq \mathcal{S}[\varphi^s g] \leq \mathcal{S}[g]$ (the second inequality comes from the analytical result (3.13)). Therefore, $\mathcal{S}[\Phi^s g] \leq \mathcal{S}[g]$ follows from the convexity of $\mathcal{S}$.
  For the Boltzmann operator, note that in the approximation (4.19), $f^{(1)}$ is a convex combination of g, $\mathcal{M}$ and $P(g)/\mu$, and f^1 is a convex combination of g, $\mathcal{M}$ and $P(f^{(1)})/\mu$. In [25], it is proved that $\mathcal{S}[P(f)/\mu] \leq \mathcal{S}[f]$ for Maxwell molecules (see Corollary 4.3 on page 825 of [25]). Therefore, by the convexity of $\mathcal{S}$ and $\mathcal{S}[\mathcal{M}[g]] \leq \mathcal{S}[g]$, one has $\mathcal{S}[f^{(1)}] \leq \mathcal{S}[g]$, hence $\mathcal{S}[f^1] \leq \mathcal{S}[g]$. Therefore, $\mathcal{S}[\Phi^s g] \leq \mathcal{S}[g]$.
- The transport step $g = f + a\Delta t \mathcal{T}(f)$ with (6.3) plugged in reads

$$\begin{aligned} g_j &= f_j - a\Delta t \left(\chi_{v_1 \geq 0} v_1 \frac{f_j - f_{j-1}}{\Delta x} + \chi_{v_1 < 0} v_1 \frac{f_{j+1} - f_j}{\Delta x} \right) \\ &= \left(1 - a\frac{|v_1|\Delta t}{\Delta x}\right) f_j + a\frac{|v_1|\Delta t}{\Delta x} \left(\chi_{v_1 \geq 0} f_{j-1} + \chi_{v_1 < 0} f_{j+1}\right). \end{aligned} \tag{6.8}$$

  Hence the right-hand side is a convex combination of f_j and $\chi_{v_1 \geq 0} f_{j-1} + \chi_{v_1<0} f_{j+1}$ under the CFL condition $\Delta t \leq \frac{\Delta x}{a|v|_{\max}}$. Then using the convexity of function $f \log f$, one has

$$\begin{aligned} S[g_j] \leq & \int_{D_v} \left(1 - a\frac{|v_1|\Delta t}{\Delta x}\right) f_j \log f_j \, \mathrm{d}v \\ & + \int_{D_v} a\frac{|v_1|\Delta t}{\Delta x} \left(\chi_{v_1 \geq 0} f_{j-1} + \chi_{v_1 < 0} f_{j+1}\right) \log \left(\chi_{v_1 \geq 0} f_{j-1} + \chi_{v_1 < 0} f_{j+1}\right) \, \mathrm{d}v \\ = & S[f_j] - a\frac{\Delta t}{\Delta x} \left(F_{j+1/2} - F_{j-1/2}\right), \end{aligned} \tag{6.9}$$

  where

$$F_{j+1/2} := \int_{D_v} |v_1| \left(\chi_{v_1 \geq 0} f_j \log f_j - \chi_{v_1 < 0} f_{j+1} \log f_{j+1}\right) \, \mathrm{d}v \tag{6.10}$$

  is the discrete entropy flux. Summing over j in (6.9) and assuming the periodic boundary condition in x, one obtains

$$\mathcal{S}[g] \leq \mathcal{S}[f]. \tag{6.11}$$

Now applying the previous two results in (2.6), we have

$$\mathcal{S}[f^{(2)}] \leq \mathcal{S}[f^{(1)}] \leq \mathcal{S}[f^{(0)}] \leq \mathcal{S}[f^n]; \tag{6.12}$$

hence

$$\mathcal{S}[f^{n+1}] \leq w\mathcal{S}[f^{(2)}] + (1-w)S[f^n] \leq \mathcal{S}[f^n]. \tag{6.13}$$

The assertion is proved.

7. A remark on spatial and velocity discretizations.

Most of the spatial and velocity discretizations follow our previous paper [15]; namely, we use a finite volume method for the x-variable and finite difference method for the v-variable.

For the transport term, we adopt the fifth-order finite volume WENO method [24] with a bound-preserving limiter [27, 28] to insure the positivity. Since the treatment of this part is standard and has been described in [15], we omit the detail.

For the collision term, special care needs to be paid when switching between the finite volume and finite difference framework. We briefly describe the procedure in the following. For convenience, we regard v as continuous and omit it in the discussion.

Let $I_j = [x_{j-1/2}, x_{j+1/2}]$ be the jth spatial cell and $\{x_{j,l}\}$ $(l = 1, 2, 3)$ denote the three Gauss–Legendre quadrature points in this cell and $\{w_l\}$ be the corresponding quadrature weights. For a fixed v, suppose we are given the cell average $f_j \geq 0$ in I_j, we would like to construct a polynomial $f_j(x)$ of degree four such that

- $f_j(x)$ is a fifth-order accurate approximation to $f(x)$ in I_j with f_j being its cell average, i.e.,

$$\frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} f_j(x)\,\mathrm{d}x = f_j. \tag{7.1}$$

- $f_j(x)$ is non-negative at the Gauss quadrature points, i.e.,

$$f_{j,l} := f_j(x_{j,l}) \geq 0, \quad l = 1, 2, 3. \tag{7.2}$$

The construction of such a polynomial can be done similarly as described in section 3.2.2 of our previous paper [15]. Provided with $f_j(x)$, it is easy to see (7.1) reduces to

$$\sum_{l=1}^{3} w_l f_{j,l} = f_j, \tag{7.3}$$

since the three-point Gauss–Legendre quadrature is exact for polynomials with degree no more than five.

Then we approximate the jth cell average of $\varphi^s f$ by

$$(\varphi^s f)_j = \sum_{l=1}^{3} w_l \varphi^s f_{j,l}. \tag{7.4}$$

This approximation is fifth-order accurate in x since the reconstruction $f_{j,l}$ is. It is also conservative, since

$$\langle(\varphi^s f)_j \phi\rangle = \sum_{l=1}^{3} w_l \langle \varphi^s f_{j,l}\phi\rangle = \sum_{l=1}^{3} w_l \langle f_{j,l}\phi\rangle = \left\langle \sum_{l=1}^{3} w_l f_{j,l}\phi \right\rangle = \langle f_j \phi\rangle, \tag{7.5}$$

where we used (3.12) in the second equality and (7.3) in the last one.

For the mixed regime problem where $\varepsilon = \varepsilon(x)$, one needs to compute $\varphi^{s(x)} f$ with $s(x)$ a given function depending on x. To do this, we use the same reconstruction $f_{j,l}$ and approximate $\varphi^{s(x)} f$ by

$$(\varphi^{s(x)} f)_j = \sum_{l=1}^{3} w_l \varphi^{s(x_{j,l})} f_{j,l}, \tag{7.6}$$

which is still fifth-order accurate and conservative.

8. Numerical examples. In this section we demonstrate numerically the properties of the proposed scheme (2.6) with coefficients (2.24) and (2.27). We will mainly use the 1D1V ($d_x = d_v = 1$) BGK and Fokker–Planck equations as prototype examples since the main purpose of this work is to develop a generic time integrator that can be potentially applied to a large class of equations rather than to study a particular kinetic equation.

Unless otherwise specified, we consider the spatial domain $x \in [0, 2]$ with periodic boundary condition (except the test in section 8.2, where the Dirichlet boundary condition is assumed) and a large enough velocity domain $v \in [-|v|_{\max}, |v|_{\max}]$ with $|v|_{\max} = 15$. The x-space is discretized into N_x cells with $\Delta x = 2/N_x$ and cell center $x_j = (j - 1/2)\Delta x$, $j = 1, \dots, N_x$. The v-space is discretized into N_v grid points with $\Delta v = 2|v|_{\max}/N_v$ and $v_i = -|v|_{\max} + (i - 1/2)\Delta v$, $i = 1, \dots, N_v$. $N_v = 150$ is used so that the discretization error in v is much smaller than that in x and t.

To compute the matrix exponential (4.25) resulting from the discretization of the Fokker–Planck operator, we used the code by Guttel [11] for the test in section 8.1, and the MATLAB function "expm" for other tests.

8.1. Accuracy test. We first verify the second-order accuracy of the scheme. We consider inconsistent initial data

$$f(0, x, v) = 0.5M_{\rho,u,T} + 0.3M_{\rho,-0.5u,T}, \tag{8.1}$$

with

$$\rho = 1 + 0.2\sin(\pi x), \quad u = 1, \quad T = \frac{1}{1 + 0.2\sin(\pi x)}, \tag{8.2}$$

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 ($\varepsilon = 10^{-10}$). We choose different Δx and set $\Delta t = 0.5\Delta x/|v|_{\max}$. This CFL number is not small enough to guarantee the positivity which is pretty restrictive due to the spatial discretization. We will consider the positivity-preserving property in the following test. For the same reason, the positivity-preserving limiter is 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 size Δx:

$$\text{error}_{\Delta t,\Delta x} := \|f_{\Delta t,\Delta x} - f_{\Delta t/2,\Delta x/2}\|_{L^2_{x,v}}. \tag{8.3}$$

The results are shown in Tables 1 and 2. For the Fokker–Planck equation, due to the second-order discretization error in the velocity space, one has to choose a larger N_v in order to see the temporal error. In all these results, the spatial error dominates for small N_x, and the temporal error dominates for large N_x. One can clearly see that in both the kinetic regime $\varepsilon = O(1)$ and the fluid regime $\varepsilon \ll 1$, the scheme is second order. Note that there is some extent of order reduction in the intermediate regime $\varepsilon = O(\Delta t)$. The uniform accuracy of the AP scheme is an open problem, and we do not attempt to address this issue in the current work.

We also perform a similar test for the ES-BGK equation to validate the approximation (4.15) proposed in section 4.2. Note that for this equation, one needs to consider $d_v > 1$ as the ES-BGK operator reduces to the BGK one when $d_v = 1$. Therefore, we consider the ES-BGK equation with $d_x = 1$ and $d_v = 2$. The parameter ν is taken as $-1/2$. We take $N_v = 150$ in each velocity dimension, and other discretization is the same as before. The initial data is

$$f(0, x, v) = 0.5M_{\rho,u,T} + 0.3M_{\rho,-0.5u,T}, \tag{8.4}$$

with

$$\rho = 1 + 0.2\sin(\pi x), \quad u = (1,0), \quad T = \frac{1}{1+0.2\sin(\pi x)}. \tag{8.5}$$

The results are similar to the previous two tests, and are shown in Table 3.

Table 1
Accuracy test of the scheme for the BGK equation.

	$\varepsilon = 1e+00$	$\varepsilon = 1e-02$	$\varepsilon = 1e-04$	$\varepsilon = 1e-06$	$\varepsilon = 1e-08$	$\varepsilon = 1e-10$
Nx=10	5.60e-04	4.64e-04	4.67e-04	4.67e-04	4.67e-04	4.67e-04
Nx=20	5.91e-05	3.93e-05	3.65e-05	3.65e-05	3.65e-05	3.65e-05
Order	3.25	3.56	3.68	3.68	3.68	3.68
Nx=40	4.33e-06	2.83e-06	4.46e-06	2.46e-06	2.46e-06	2.46e-06
Order	3.77	3.80	3.03	3.89	3.89	3.89
Nx=80	2.11e-07	2.86e-07	5.24e-06	1.10e-07	1.10e-07	1.10e-07
Order	4.36	3.31	-0.23	4.49	4.49	4.49
Nx=160	1.27e-08	6.24e-08	3.25e-06	6.29e-09	6.29e-09	6.29e-09
Order	4.05	2.19	0.69	4.12	4.12	4.12
Nx=320	2.89e-09	1.55e-08	1.23e-06	1.45e-09	1.45e-09	1.45e-09
Order	2.14	2.01	1.40	2.11	2.11	2.11
Nx=640	7.30e-10	3.88e-09	3.74e-07	3.68e-10	3.68e-10	3.68e-10
Order	1.99	2.00	1.72	1.98	1.98	1.98
Nx=1280	1.83e-10	9.71e-10	1.03e-07	2.82e-10	9.20e-11	9.20e-11
Order	2.00	2.00	1.86	0.38	2.00	2.00

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

$$f(0,x,v) = M_{\rho,u,T}, \tag{8.6}$$

with

$$(\rho, u, T) = \begin{cases} (1,0,1), & 0 \le x \le 1, \\ (0.125, 0, 0.25), & 1 < x \le 2. \end{cases} \tag{8.7}$$

With the positivity-preserving limiter, the CFL condition of our scheme is $\Delta t \le \frac{1}{12}\frac{\Delta x}{|v|_{\max}}$ (note that 1/12 comes from the spatial discretization and the forward Euler method also has the same constraint). We choose $\Delta t = \frac{1}{24}\frac{\Delta x}{|v|_{\max}}$ and $N_x = 80$.

For the BGK equation, no negative cells are detected in the simulation. For the Fokker–Planck equation, one technical issue is that we are not aware of any algorithms that can guarantee the numerically computed matrix exponential is positive if the exact matrix exponential is. To demonstrate that no negative values are caused by our time discretization, we use "expm" function in MATLAB to compute the matrix exponential and set the negative entries of the resulting matrix to zero. With this modification, no negative cells are detected in the simulation.

As a comparison, we solve the same equations with the same initial data and spatial/velocity discretization, but using the ARS(2,2,2) scheme in time [2], which is a standard second-order accurate IMEX scheme without positivity-preserving property. The number of negative cells during the simulation is tracked. The result for the BGK equation is already included in the previous paper [15] and is omitted here. The result for the Fokker–Planck equation is shown in Figure 1. Here to make the comparison fair, when we compute $(I - s\tilde{Q}^h)^{-1}g^h$ (an operator needs to be evaluated in the IMEX

Table 2
Accuracy test of the scheme for the Fokker–Planck equation. Here $N_v = 600$.

	$\varepsilon = 1e+00$	$\varepsilon = 1e-01$	$\varepsilon = 1e-02$	$\varepsilon = 1e-03$	$\varepsilon = 1e-04$	$\varepsilon = 1e-05$	$\varepsilon = 1e-06$	$\varepsilon = 1e-07$
Nx=10	5.30e-04	4.64e-04	4.62e-04	4.66e-04	4.66e-04	4.66e-04	4.66e-04	4.66e-04
Nx=20	5.50e-05	4.32e-05	3.93e-05	4.63e-05	3.65e-05	3.65e-05	3.65e-05	3.65e-05
Order	3.27	3.42	3.56	3.33	3.68	3.68	3.68	3.68
Nx=40	3.89e-06	2.82e-06	3.42e-06	1.29e-05	2.54e-06	2.46e-06	2.46e-06	2.46e-06
Order	3.82	3.94	3.52	1.85	3.85	3.89	3.89	3.89
Nx=80	1.80e-07	1.29e-07	5.47e-07	4.23e-06	2.16e-06	1.10e-07	1.10e-07	1.10e-07
Order	4.43	4.45	2.64	1.61	0.23	4.49	4.49	4.49
Nx=160	1.13e-08	9.34e-09	1.35e-07	1.25e-06	2.97e-06	1.16e-08	6.30e-09	6.29e-09
Order	3.99	3.79	2.02	1.76	-0.46	3.25	4.12	4.12
Nx=320	2.64e-09	2.07e-09	3.56e-08	3.53e-07	1.80e-06	1.08e-07	1.50e-09	1.45e-09
Order	2.10	2.17	1.92	1.82	0.72	-3.22	2.07	2.11
Nx=640	6.66e-10	5.77e-10	9.93e-09	1.01e-07	7.15e-07	3.91e-07	3.62e-10	3.68e-10
Order	1.98	1.85	1.84	1.80	1.33	-1.86	2.05	1.98

TABLE 3
Accuracy test of the scheme for the ES-BGK equation.

	$\varepsilon = 1e+00$	$\varepsilon = 1e-02$	$\varepsilon = 1e-04$	$\varepsilon = 1e-06$	$\varepsilon = 1e-08$	$\varepsilon = 1e-10$
Nx=10	1.56e-04	1.52e-04	1.57e-04	1.57e-04	1.57e-04	1.57e-04
Nx=20	1.50e-05	9.69e-06	8.86e-06	8.83e-06	8.83e-06	8.83e-06
Order	3.38	3.97	4.15	4.16	4.16	4.16
Nx=40	8.05e-07	6.10e-07	1.38e-06	3.88e-07	3.88e-07	3.88e-07
Order	4.22	3.99	2.68	4.51	4.51	4.51
Nx=80	2.81e-08	9.30e-08	1.97e-06	9.04e-09	9.04e-09	9.04e-09
Order	4.84	2.71	-0.51	5.42	5.42	5.42
Nx=160	2.35e-09	2.26e-08	1.21e-06	5.33e-10	5.10e-10	5.10e-10
Order	3.58	2.04	0.71	4.08	4.15	4.15
Nx=320	5.65e-10	5.67e-09	4.48e-07	1.87e-10	1.34e-10	1.34e-10
Order	2.06	2.00	1.43	1.51	1.93	1.93

scheme), we first compute the matrix $(I - s\tilde{\mathcal{Q}}^h)^{-1}$ which is not necessarily positive at the numerical level, and then set the negative entries to zero in this matrix. This is to make sure that no negative values are generated due to the failure of positivity-preserving in the matrix inversion. In Figure 1 one can still see a lot of negative cells in the fluid regime.

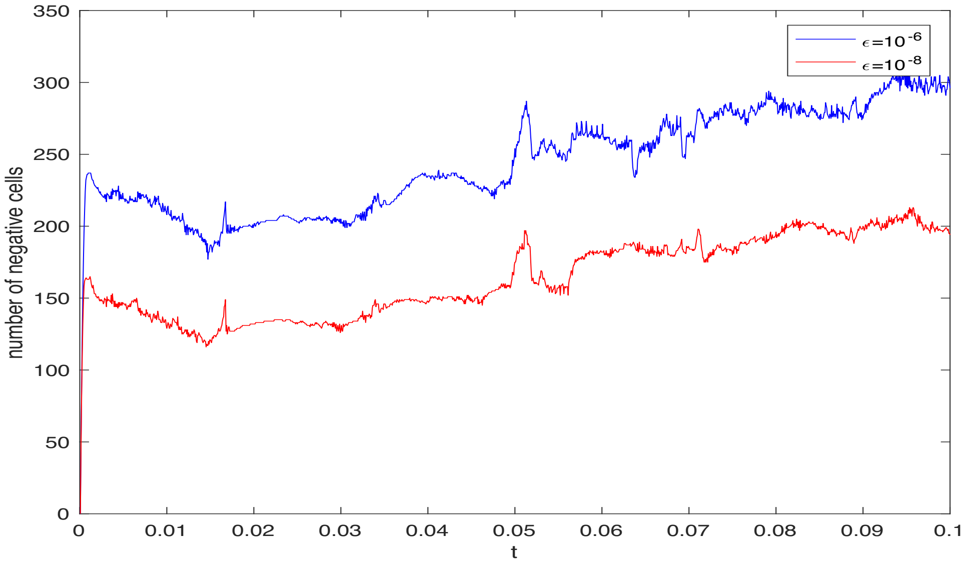

FIG. 1. *Total number of negative cells for the ARS$(2,2,2)$ scheme applied to the Fokker–Planck equation during time evolution. Blue line: $\varepsilon = 10^{-6}$. Red line: $\varepsilon = 10^{-8}$.*

8.3. AP property.

Finally, to illustrate the AP property, we use the proposed scheme to solve the BGK and Fokker–Planck equations in a mixed regime (ε is a function of x so that in part of the domain the problem is in kinetic regime and while in other part it is in fluid regime). We take the same initial data as in (8.1)–(8.2) and $N_x = 40$.

For the BGK equation, we consider $\varepsilon = \varepsilon(x)$ as follows:

$$\varepsilon(x) = \varepsilon_0 + (\tanh(1 - 11(x-1)) + \tanh(1 + 11(x-1))), \quad \varepsilon_0 = 10^{-5}. \tag{8.8}$$

We compare the macroscopic quantities at time $t = 0.5$ with a reference solution computed by Heun's method with $N_x = 80$. Note that for our scheme, $\Delta t = \frac{1}{24}\frac{\Delta x}{|v|_{\max}} \approx 7\times 10^{-5}$; while for the (explicit) Heun's method, $\Delta t = \frac{1}{240}\frac{\Delta x}{|v|_{\max}} \approx 7\times 10^{-6}$ which needs to resolve ε. One can see a good agreement with the reference solution in Figure 2.

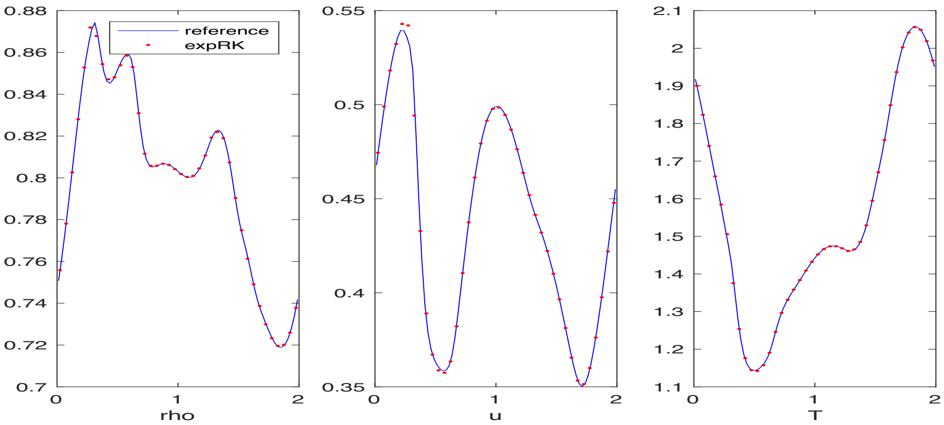

FIG. 2. *The BGK equation in a mixed regime. Left to right: density ρ, velocity u, and temperature T. Solid line: reference solution computed by the explicit Heun's method. Dots: solution computed by the proposed scheme.*

For the Fokker–Planck equation, we consider the following $\varepsilon(x)$:

$$\varepsilon(x) = \varepsilon_0 + (\tanh(1-11(x-1)) + \tanh(1+11(x-1))), \quad \varepsilon_0 = 5\times 10^{-4}. \tag{8.9}$$

The numerical parameters are chosen the same as the BGK case, except in the reference solution we need $\Delta t = \frac{1}{540}\frac{\Delta x}{|v|_{\max}} \approx 3\times 10^{-6}$ in order to satisfy the explicit parabolic CFL condition. The result is shown in Figure 3 and again with good agreement.

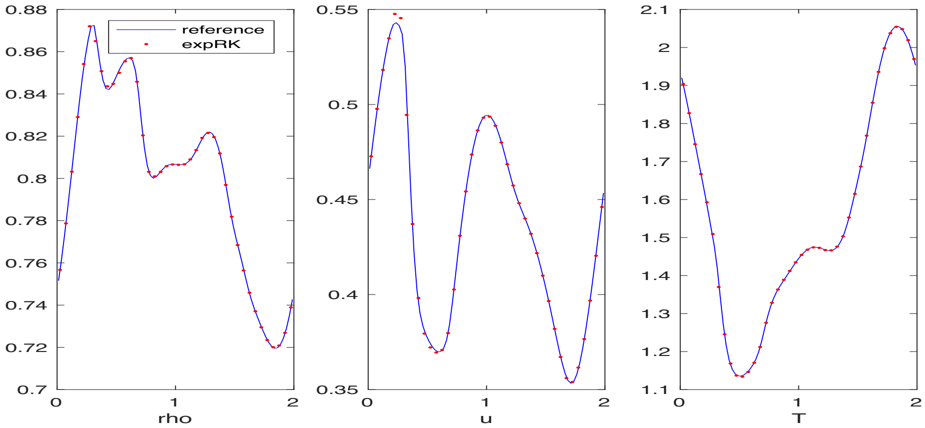

FIG. 3. *The Fokker–Planck equation in a mixed regime. Left to right: density ρ, velocity u, and temperature T. Solid line: reference solution computed by the explicit Heun's method. Dots: solution computed by the proposed scheme.*

9. Conclusion. We introduced a new exponential Runge–Kutta time discretization method (2.6) for a class of stiff kinetic equations (1.1). The main contribution is a careful blend of the transport term and the collision term so that the method is second order, AP, and positivity preserving. With suitably chosen homogeneous solvers, the method can be applied to the relaxation type equation (BGK and ES-BGK equations), the diffusion type equation (kinetic Fokker–Planck equation), and the full Boltzmann equation. Further, we showed that the method satisfies an entropy-decay property when coupled with upwind discretization for the transport term. A series of numerical examples were presented to demonstrate the properties of the proposed method.

Acknowledgement. The first author would like to thank Prof. Lili Ju for helpful discussion on exponential methods.

REFERENCES

[1] P. Andries, P. L. Tallec, J.-P. Perlat, and B. Perthame, *The Gaussian-BGK model of Boltzmann equation with small Prandtl number*, Eur. J. Mech. B Fluids, 19 (2000), pp. 813–830.

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

[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 (1954), pp. 511–525.

[4] G. A. Bird, *Molecular Gas Dynamics and the Direct Simulation of Gas Flows*, Clarendon Press, Oxford, 1994.

[5] Z. Cai, Y. Fan, and L. Ying, *An entropic Fourier method for the Boltzmann equation*, SIAM J. Sci. Comput., 40 (2018), pp. A2858–A2882.

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

[7] G. Dimarco and L. Pareschi, *Exponential Runge-Kutta methods for stiff kinetic equations*, SIAM J. Numer. Anal., 49 (2011), pp. 2057–2077.

[8] E. Gabetta, L. Pareschi, and G. Toscani, *Relaxation schemes for nonlinear kinetic equations*, SIAM J. Numer. Anal., 34 (1997), pp. 2168–2194.

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

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

[11] S. Güttel, *Rational Krylov methods for operator functions*, PhD thesis, Institut für Numerische Mathematik und Optimierung, Technische Universität Bergakademie Freiberg, 2010.

[12] M. Hochbruck and A. Ostermann, *Exponential integrators*, Acta Numer., 19 (2010), pp. 209–286.

[13] L. Holway, *Kinetic theory of shock structure using an ellipsoidal distribution function*, in Proceedings of the 4th International Symposium on Rarefied Gas Dynamics, vol. I, Academic Press, New York, 1966, pp. 193–215.

[14] J. Hu, S. Jin, and Q. Li, *Asymptotic-preserving schemes for multiscale hyperbolic and kinetic equations*, in Handbook of Numerical Methods for Hyperbolic Problems: Applied and Modern Issues, R. Abgrall and C.-W. Shu, eds., North-Holland, Amsterdam, 2017, pp. 103–129.

[15] J. Hu, R. Shu, and X. Zhang, *Asymptotic-preserving and positivity-preserving implicit-explicit schemes for the stiff BGK equation*, SIAM J. Numer. Anal., 56 (2018), pp. 942–973.

[16] J. Huang and C.-W. Shu, *Bound-preserving modified exponential Runge-Kutta discontinuous Galerkin methods for scalar hyperbolic equations with stiff source terms*, J. Comput. Phys., 361 (2018), pp. 111–135.

[17] S. Jin, *Runge-Kutta methods for hyperbolic conservation laws with stiff relaxation terms*, J. Comput. Phys., 122 (1995), pp. 51–67.

[18] S. Jin, *Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations*, SIAM J. Sci. Comput., 21 (1999), pp. 441–454.

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

[20] S. Jin and B. Yan, *A class of asymptotic-preserving schemes for the Fokker-Planck-Landau equation*, J. Comput. Phys., 230 (2011), pp. 6420–6437.

[21] Q. Li and L. Pareschi, *Exponential Runge-Kutta for the inhomogeneous Boltzmann equations with high order of accuracy*, J. Comput. Phys., 259 (2014), pp. 402–420.

[22] C. Mouhot, L. Pareschi, and T. Rey, *Convolutive decomposition and fast summation methods for discrete-velocity approximations of the Boltzmann equation*, ESAIM: Math. Model. Numer. Anal., 47 (2013), pp. 1515–1531.

[23] L. Pareschi and G. Russo, *Time relaxed Monte Carlo methods for the Boltzmann equation*, SIAM J. Sci. Comput., 23 (2001), pp. 1253–1273.

[24] C.-W. Shu, *Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws*, in Advanced Numerical Approximation of Nonlinear Hyperbolic Equations, Springer, New York, 1998, pp. 325–432.

[25] C. Villani, *Fisher information estimates for Boltzmann's collision operator*, J. Math. Pure Appl., 77 (1998), pp. 821–837.

[26] C. Villani, *A review of mathematical topics in collisional kinetic theory*, in Handbook of Mathematical Fluid Mechanics, S. Friedlander and D. Serre, eds., vol. I, North-Holland, Amsterdam, 2002, pp. 71–305.

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

[28] 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, The Royal Society, London, 2011, rspa20110153.