Commun. Comput. Phys.
doi: 10.4208/cicp.OA-2020-0050

Vol. **29**, No. 1, pp. 1-56
January 2021

High Order Conservative Semi-Lagrangian Scheme for the BGK Model of the Boltzmann Equation

Sebastiano Boscarino[1,*], Seung-Yeon Cho[1], Giovanni Russo[1] and Seok-Bae Yun[2]

[1] *Department of Mathematics and Computer Science, University of Catania, 95125 Catania, Italy.*
[2] *Department of Mathematics, Sungkyunkwan University, Suwon 440-746, Republic of Korea.*

Received 17 March 2020; Accepted (in revised version) 15 July 2020

Abstract. In this paper, we present a conservative semi-Lagrangian finite-difference scheme for the BGK model. Classical semi-Lagrangian finite difference schemes, coupled with an L-stable treatment of the collision term, allow large time steps, for all the range of Knudsen number [17,27,30]. Unfortunately, however, such schemes are not conservative. Lack of conservation is analyzed in detail, and two main sources are identified as its cause. First, when using classical continuous Maxwellian, conservation error is negligible only if velocity space is resolved with sufficiently large number of grid points. However, for a small number of grid points in velocity space such error is not negligible, because the parameters of the Maxwellian do not coincide with the discrete moments. Secondly, the non-linear reconstruction used to prevent oscillations destroys the translation invariance which is at the basis of the conservation properties of the scheme. As a consequence the schemes show a wrong shock speed in the limit of small Knudsen number. To treat the first problem and ensure machine precision conservation of mass, momentum and energy with a relatively small number of velocity grid points, we replace the continuous Maxwellian with the discrete Maxwellian introduced in [22]. The second problem is treated by implementing a conservative correction procedure based on the flux difference form as in [26]. In this way we can construct conservative semi-Lagrangian schemes which are Asymptotic Preserving (AP) for the underlying Euler limit, as the Knudsen number vanishes. The effectiveness of the proposed scheme is demonstrated by extensive numerical tests.

AMS subject classifications: 65L06, 65M25, 76P05

Key words: BGK model, Boltzmann equation, semi-Lagrangian scheme, conservative correction, discrete Maxwellian.

[*] Corresponding author. *Email addresses:* boscarino@dmi.unict.it (S. Boscarino), chosy89@skku.edu (S.-Y. Cho), russo@dmi.unict.it (G. Russo), sbyun01@skku.edu (S.-B. Yun)

1 Introduction

The dynamics of a non-ionized dilute gas at mesoscopic level is described by the celebrated Boltzmann equation [9]. The development of efficient numerical methods for its solution, however, constitutes a formidable challenge, due, among others, to the high dimensionality of the problem, the complicated structure of the collision operator, the need to preserve the collision invariants at a discrete level, and the stiffness issue arising when the Knudsen number is very small.

In view of this situation, Bhatnagar, Gross and Krook, in 1954, suggested a relaxation model of the Boltzmann equation, which now goes by the name of the BGK model [5]. This approximation preserves several important qualitative features of the original Boltzmann equation, such as conservation of mass, momentum and energy, H-theorem and relaxation to equilibrium, and is now widely used as a simplified alternative to the Boltzmann equation because it is much less expensive to treat at a numerical level.

Initial value problem for the BGK model on a periodic domain reads

$$\begin{aligned} &\frac{\partial f}{\partial t}+v\cdot\nabla_x f=\frac{1}{\kappa\tau_0}(\mathcal{M}(f)-f),\\ &f(x,v,0)=f_0(x,v). \end{aligned} \tag{1.1}$$

The velocity distribution function $f(x,v,t)$ represents the mass density of particles at point $(x,v)\in\mathbb{R}^d\times\mathbb{R}^d$ in phase space, at time $t>0$. The quantity $\tau=\kappa\tau_0$ represent the relaxation time. Here κ is the Knudsen number, defined as a ratio between the mean free path and a macroscopic characteristic length of the physical system. We assume it may change by several orders of magnitude, and in particular it may become extremely small. The time τ_0 expresses the dependence of the relaxation time on the deviation of temperature and density from the reference one. We assume such dependence is not very strong, and for simplicity we consider τ_0 to be constant in our treatment and analysis. By suitable non-dimensionalization of the problem we shall omit to write the term τ_0. The local Maxwellian $\mathcal{M}(f)$ is given by

$$\mathcal{M}(f)(x,v,t):=\frac{\rho(x,t)}{\sqrt{(2\pi T(x,t))^d}}\exp\left(-\frac{|v-U(x,t)|^2}{2T}\right),$$

where the macroscopic fields of local density $\rho(x,t)\in\mathbb{R}^+$, bulk velocity $U(x,t)\in\mathbb{R}^d$ and local temperature $T(x,t)\in\mathbb{R}^+$ are defined through the following relation:

$$(\rho(x,t),\rho(x,t)U(x,t),E(x,t))^T=\langle f\phi(v)\rangle, \tag{1.2}$$

where

$$\phi(v)=\left(1,v,\frac{1}{2}|v|^2\right)^T,\quad \text{and}\quad \langle g\rangle\equiv\int_{\mathbb{R}^d}g(v)dv.$$

The physical quantity $E(x,t)$ is the total energy density per unit volume, and it is related to the temperature $T(x,t)$ by the following relation:

$$E(x,t)=\frac{d}{2}\rho(x,t)T(x,t)+\frac{1}{2}\rho(x,t)|U(x,t)|^2.$$

The BGK model (1.1) satisfies the main properties of the Boltzmann equation such as conservation of mass, momentum and energy:

$$\langle\mathcal{M}(f)\phi(v)\rangle=\langle f\phi(v)\rangle,$$

as well as entropy dissipation:

$$\int_{\mathbb{R}^d}(\mathcal{M}(f)-f)\ln f dv\le 0.$$

Note that the equilibrium state clearly is the local Maxwellian determined by f. Indeed the collision operator vanished for $f=\mathcal{M}(f)$. Therefore, the BGK model gives the correct Euler limit as $\kappa\to 0$, i.e., the moments of solution to (1.1), in the limit of vanishing Knudsen number, satisfy the macroscopic compressible Euler equations for a monatomic gas [4,6]:

$$\begin{aligned}&\partial_t\rho+\nabla\cdot(\rho U)=0,\\&\partial_t(\rho U)+\nabla\cdot(\rho U\otimes U+pI)=0,\\&\partial_t E+\nabla\cdot((E+p)U)=0,\end{aligned}\tag{1.3}$$

with pressure p given by the constitutive relation to close the system (1.3) $p=\rho T$.

Navier-Stokes equations can be derived by the Chapman-Enskog equation (see for example [10]), by inserting a formal expansion of the distribution function f in terms of the Knudsen number. To zero-th order one obtains compressible Euler's equations, while to first order in κ one derives the Navier-Stokes equations associated to the BGK model.

We mention that such Navier-Stokes limit is slightly inconsistent with the one obtained from the Boltzmann equation, in that the Prandtl number $\mathrm{Pr}=c_p\mu/k$ (c_p is the specific heat at constant pressure, μ is the viscosity and k the thermal conductivity) derived from the BGK model is numerically different from the value computed using the Boltzmann equation. Several techniques have been proposed to overcome this drawback, the most widely adopted being the so-called Ellipsoidal BGK (ES-BGK), see [1,2,20]. A semi-Lagrangian method for the ES-BGK model has recently been proposed and analyzed in [29].

There have been several efficient numerical methods for the BGK model of the Boltzmann equation. In [32], the authors propose a high order conservative A-stable scheme which performs well for both fluid and rarefied regime. The procedure requires to update the microscopic distribution function coupled with the update of the macroscopic conservative variables. One can also find efficient numerical methods for BGK model in [15, 16]. Among various numerical approaches, in this paper, we aim to analyze the lack of conservation of classical semi-Lagrangian schemes for the BGK model. Then, we propose an alternative technique which gives high order conservative semi-Lagrangian (SL) finite-difference schemes for the BGK model.

The loss of conservation has been an important issue in the field of the method of characteristics, and it can be more relevant if one solves equations with variable coefficients [13]. In kinetic theory, conservative semi-Lagrangian methods have recently attracted a lot of attention, especially in the context of the Vlasov-Poisson model (see [12, 14]).

General procedures have been developed for the construction of conservative SL schemes, as in [25], however such procedures are often restricted to treat one dimensional problems.

SL schemes for BGK models have recently received increasing interest [17,27–30] since the SL treatment avoids the classical CFL stability restriction. Furthermore, the implicit treatment of the collision term, which can be easily computed, allows the methods to capture the underlying fluid dynamic limit.

Unfortunately, however, classical SL schemes do not necessarily conserve the total mass, momentum and energy, and the error may become more relevant as the Knudsen number gets smaller [17].

We identify the cause of lack of conservation in the use of continuous Maxwellian in the collision term, and in the non-linear weights adopted in the high order non-oscillatory reconstruction, and propose a remedy based on the use of a discrete Maxwellian (as in [22]) and on a conservative correction to fully restore the conservation properties of the schemes, such as the one adopted in [26] in the case of the Vlasov-Poisson equation.

The paper is organized as follows. Section 2 is devoted to first order schemes. It is shown that the conservation error depends sensitively on the number of velocity grid points, and the cause is identified in the use of a continuous Maxwellian in a discrete scheme. We prove that the SL schemes can be made conservative within round-off errors by adopting a discrete Maxwellian in place of the classical continuous one.

Section 3 considers high order SL schemes, which exhibit lack of conservation even with the use of the discrete Maxwellian in the collision term. A conservative correction is then adopted, which restores exact conservation of the methods (within round-off). Section 4 is devoted to linear stability analysis, to explain the stability limitations introduced by the conservative correction. The extension of the methods to the two dimensional case is described in Section 5. In Section 6 we present several numerical tests, which confirm the expected accuracy and conservation properties of the proposed schemes, and provide numerical evidence of the AP property of the scheme towards the underlying fluid dynamic limit as the Knudsen number vanishes. At the end of this paper, we draw some conclusions.

2 First order Semi-Lagrangian schemes

We start from the basic first order semi-Lagrangian scheme [28], and gradually build up to derive our conservative high order semi-Lagrangian scheme (see Section 3).

2.1 First order SL scheme

We start from the characteristic formulation of (1.1) :

$$\frac{df}{dt}=\frac{1}{\kappa}(\mathcal{M}(f)-f), \quad \frac{dx}{dt}=v, \tag{2.1}$$

subject to the initial data: $f(x,v,0)=f_0(x,v)$.

We consider one dimensional problem in space and velocity, and we divide the spatial and velocity domain into uniform grids with mesh spacing Δx and Δv, respectively. We also use uniform time step Δt. Given a computational domain, $[x_{\min},x_{\max}]\times[v_{\min},v_{\max}]\times[0,t_f]$, we denote the grid points by

$$\begin{aligned} &x_i=x_{\min}+\left(i-\frac{1}{2}\right)\Delta x, && i=1,\cdots,N_x,\\ &v_j=v_{\min}+j\Delta v, && j=0,\cdots,N_v,\\ &t^n=n\Delta t, && n=0,\cdots,N_t, \end{aligned}$$

where N_x, N_v+1 and N_t are the number of grid nodes in space, velocity and time, respectively, so that $x_{\max}=x_{\min}+N_x\Delta x$, $v_{\max}=v_{\min}+N_v\Delta v$ and $t_f=N_t\Delta t$.

Let $f_{i,j}^n$ denote a discrete approximation of $f(x_i,v_j,t^n)$ and $\phi(v_j)=\left(1,v_j,\frac{v_j^2}{2}\right)^T$. Applying first order semi-Lagrangian implicit Euler (IE-SL) scheme to (2.1), we get

$$f_{i,j}^{n+1}=\tilde{f}_{ij}^n+\frac{\Delta t}{\kappa}\left(\mathcal{M}(f_{i,j}^{n+1})-f_{i,j}^{n+1}\right), \tag{2.2}$$

where $\mathcal{M}(f_{i,j}^n)\equiv(\mathcal{M}[f^n])_{i,j}$, and $\tilde{f}_{ij}^n$ is an approximation of $f(x_i-v_j\Delta t,v_j,n\Delta t)$ obtained by a suitable interpolation from $\{f_{i,j}^n\}$. Note that linear reconstruction will be sufficient for first order SL scheme, while a higher order non-oscillatory reconstruction is necessary for high order accuracy. The Maxwellian $\mathcal{M}(f_{i,j}^{n+1})$ is given by

$$\mathcal{M}(f_{i,j}^{n+1})=\frac{\rho_i^{n+1}}{\sqrt{2\pi T_i^{n+1}}}\exp\left(-\frac{|v_j-U_i^{n+1}|^2}{2T_i^{n+1}}\right),$$

where discrete macroscopic moments are constructed from f^{n+1} as follows:

$$\begin{pmatrix}\rho_i^{n+1}\\ \rho_i^{n+1}U_i^{n+1}\\ E_i^{n+1}\end{pmatrix}=\sum_{j=0}^{N_v}f_{i,j}^{n+1}\phi(v_j)\Delta v,$$

which is equivalent to using midpoint rule in the computation of the moments, Eq. (1.2).

We now employ a technique which enables us to explicitly solve the implicit scheme (2.2). The idea behind it is that the Maxwellian and the distribution function at time t^{n+1}

have the same moments. Such a technique has been adopted several times by various authors. For example, Pieraccini and Puppo adopted it in [24] in the context of Eulerian schemes, Santagati used it independently in his PhD thesis [30] in the context of semi-Lagrangian schemes, and Coron and Perthame [11] adopted a similar idea using a splitting strategy, in which a convection step is followed by a relaxation step, during which the local Maxwellian does not change.

We multiply both sides in (2.2) by $\phi(v_j)$, sum over j, and use the property that the moments of $\mathcal{M}(f_{i,j}^{n+1})-f_{i,j}^{n+1}$ up to second order vanish, to obtain

$$\sum_{j=0}^{N_v} f_{i,j}^{n+1}\phi(v_j)\Delta v = \sum_{j=0}^{N_v} \tilde{f}_{ij}^{n}\phi(v_j)\Delta v.$$

This gives

$$\begin{pmatrix} \rho_i^{n+1} \\ U_i^{n+1} \\ E_i^{n+1} \end{pmatrix} = \begin{pmatrix} \tilde{\rho}_i^{n} \\ \tilde{U}_i^{n} \\ \tilde{E}_i^{n} \end{pmatrix},$$

with

$$\begin{pmatrix} \tilde{\rho}_i^{n} \\ \tilde{\rho}_i^{n}\tilde{U}_i^{n} \\ \tilde{E}_i^{n} \end{pmatrix} = \sum_{j=0}^{N_v} \tilde{f}_{ij}^{n}\phi(v_j)\Delta v, \quad \tilde{E}_i^n = \frac{1}{2}\tilde{\rho}_i^n|\tilde{U}_i^n|^2 + \frac{1}{2}\tilde{\rho}_i^n\tilde{T}_i^n.$$

Therefore, we can legitimately replace $\mathcal{M}(f_{i,j}^{n+1})$ with $\mathcal{M}(\tilde{f}_{i,j}^{n})$, so that the scheme becomes

$$f_{i,j}^{n+1} = \tilde{f}_{ij}^{n} + \frac{\Delta t}{\kappa}\left(\mathcal{M}(\tilde{f}_{ij}^{n}) - f_{i,j}^{n+1}\right),$$

which gives

$$f_{i,j}^{n+1} = \frac{\kappa \tilde{f}_{ij}^{n} + \Delta t \mathcal{M}(\tilde{f}_{ij}^{n})}{\kappa + \Delta t}. \tag{2.3}$$

This approach has been also fruitfully used, for example, in [17,27,28].

Summarizing, we have the following procedure (see Fig. 1):

1. Use linear interpolation to obtain $\tilde{f}_{ij}^{n}$ from $\{f_{i,j}^{n}\}$.

2. Compute $\mathcal{M}(\tilde{f}_{ij}^{n})$ from $\{\tilde{f}_{ij}^{n}\}$ by using the macroscopic moments, i.e. $(\tilde{\rho}_i^n, \tilde{U}_i^n, \tilde{E}_i^n)^T$.

3. Compute numerical solution using (2.3).

We apply the scheme to the propagation of a single shock, where we can compare the numerical solution to the exact one, and therefore accurately check the conservation properties of the scheme.

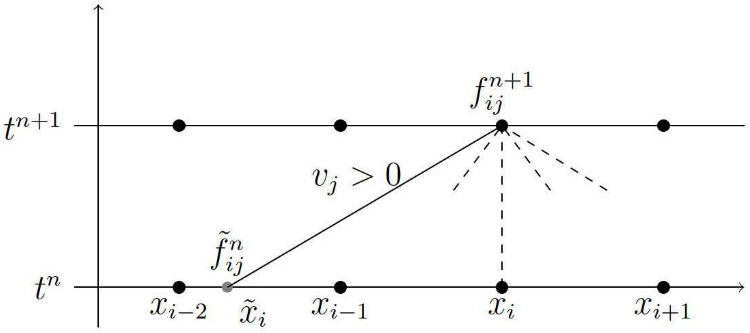


Figure 1: Representation of the implicit first order scheme.

2.2 Test 1. Check exact conservation

The aim of this test is to check conservation of the conservative semi-Lagrangian scheme to machine precision. We apply IE-SL scheme (2.3) to Eq. (1.1) with f_0 given by the Maxwellian w.r.t macroscopic quantities

$$(\rho_0,u_0,p_0)=\begin{cases} \left(\frac{(\gamma+1)M^2}{(\gamma-1)M^2+2},\frac{2\sqrt{\gamma}(M^2-1)}{(\gamma+1)M},1+\frac{2\gamma(M^2-1)}{(\gamma+1)}\right), & \text{for } x\leq 0.5,\\ (1,0,1), & \text{for } x>0.5. \end{cases}$$

We take the Knudsen number $\kappa=10^{-6}$, the polytropic constant $\gamma=3$ (corresponding to a polytropic gas with one degree of freedom per gas molecule) and Mach number $M=2$.

To prevent the solution from reaching the boundary, final time is taken $t_f=0.4$. We used free flow boundary conditions and performed the computation on $(x,v)\in[0,5]\times[-20,20]$.

The results are summarized in Table 1, where the conservation errors are reported for various values of N_x and N_v.

From the results we can make the following observations:

1. Table 1 shows that the first order IE-SL scheme with enough points in velocity space maintains conservation within machine precision, independently of the number of grid point in space;

2. the same scheme with smaller number of points in velocity produces non-negligible conservation errors.

This numerical evidence suggests that the convection part is conservative, while errors in conservation are a consequence of the numerical approximation of the relaxation term. The lack of conservation is indeed due to the use of a continuous Maxwellian on a discrete scheme in velocity: the parameters of the continuous Maxwellian do not coincide with the discrete moments, they are just approximated by them with spectral accuracy

Table 1: Test 1. CFL $=4$, $\kappa=10^{-6}$. Conservation errors of discrete moments in the relative L_1 norm for first order scheme applied to single shock with velocity domain $[-20,20]$.

	IE-SL-Linear-CM, $N_x=100$			IE-SL-Linear-CM, $N_x=200$		
N_v	Mass	Momentum	Energy	Mass	Momentum	Energy
30	3.63e-04	0.0012	0.0021	9.10e-04	0.0030	0.0051
40	5.54e-08	3.26e-07	6.03e-07	1.15e-07	6.43e-07	1.25e-06
50	8.55e-13	7.81e-12	1.43e-11	1.78e-12	1.54e-11	2.97e-11
60	3.55e-14	4.96e-14	3.89e-14	7.45e-14	8.24e-14	7.23e-14
90	3.24e-14	4.82e-14	3.77e-14	7.16e-14	7.32e-14	7.45e-14

when the integrals are replaced by a summation. The spectral accuracy of the quadrature explains, for example, the dramatic drop of the conservation error when the number of points in velocity is increased from 40 to 50.

2.3 Classical SL scheme with the discrete Maxwellian

In this section, we replace the continuous Maxwellian with the discrete Maxwellian to resolve the problem of strong dependence of the conservation error on the number of velocity grids.

2.3.1 Discrete Maxwellian

We start by describing the discrete Maxwellian introduced in [22]. In that work, the author proved that a discrete entropy minimization problem has a unique solution called the discrete Maxwellian ($d\mathcal{M}$). More precisely a consequence of his Theorem 3.1 in [22] is that for any discrete distribution function $\{f_j\}$, with discrete moments $\mathbf{m}\in\mathbb{R}^{2+d}$, $\mathbf{m}=\sum_j f_j\phi(v_j)(\Delta v)^d$, there is a unique discrete distribution $d\mathcal{M}(v_j)$ that minimizes the discrete entropy $H[g]=\sum_j g_j\log(g_j)(\Delta v)^d$, subject to the condition that its moments are $\mathbf{m}$, and that such a discrete Maxwellian can be expressed as

$$d\mathcal{M}(x,v_j,t):=\exp\left(\mathbf{a}\cdot\phi(v_j)\right),$$

with a suitable vector $\mathbf{a}\in\mathbb{R}^{2+d}$.

For $d=1$, the vector $\mathbf{a}(x,t)$ is determined by solving the following non-linear system:

$$\sum_{j=0}^{N_v} f(x,v_j,t)\phi(v_j)\Delta v=\sum_{j=0}^{N_v}\exp\left(\mathbf{a}(x,t)\cdot\phi(v_j)\right)\phi(v_j)\Delta v.$$

In practice, employing a Newton algorithm, we find $a(x,t)$ such that

$$\max_{1\leq\ell\leq 3}\left|\sum_{j=0}^{N_v}\left(f(x,v_j,t)-d\mathcal{M}(x,v_j,t)\right)\phi_\ell(v_j)\Delta v\right|<tol \tag{2.4}$$

for arbitrary small tolerance(tol). Throughout this paper, we take *tol* to be the order of 10^{-14}. Here, we denote the ℓth component of $\phi(v_j)$ by $\phi_\ell(v_j)$, $\ell=1,2,3$. With the use of discrete Maxwellian in (2.3),

$$f_{i,j}^{n+1}=\frac{\kappa\tilde{f}_{ij}^n+\Delta t d\mathcal{M}(\tilde{f}_{ij}^n)}{\kappa+\Delta t}, \tag{2.5}$$

and it is possible to prove the following estimate on the conservation error (see Appendix A):

$$\max_{1\le\ell\le3}\left|\sum_{i=1}^{N_x}\sum_{j=0}^{N_v}\left(f_{i,j}^{Nt}-f_{i,j}^0\right)\phi_\ell(v_j)\Delta v\Delta x\right|\le\frac{N_t\Delta t}{\kappa+\Delta t}(x_{\max}-x_{\min})tol.$$

On the other hand, we recall that, in discrete velocity models, we need to take the velocity domain sufficiently large to secure correct profile of macroscopic moments, especially when there is a large space variation of mean velocity U and temperature T.

Therefore, it is necessary to balance the size of the velocity domain needed for the accurate computation of macroscopic fields, and the efficient choice of smallest possible number of grids to guarantee the efficient performance of the scheme (see Test 1 in Section 6.)

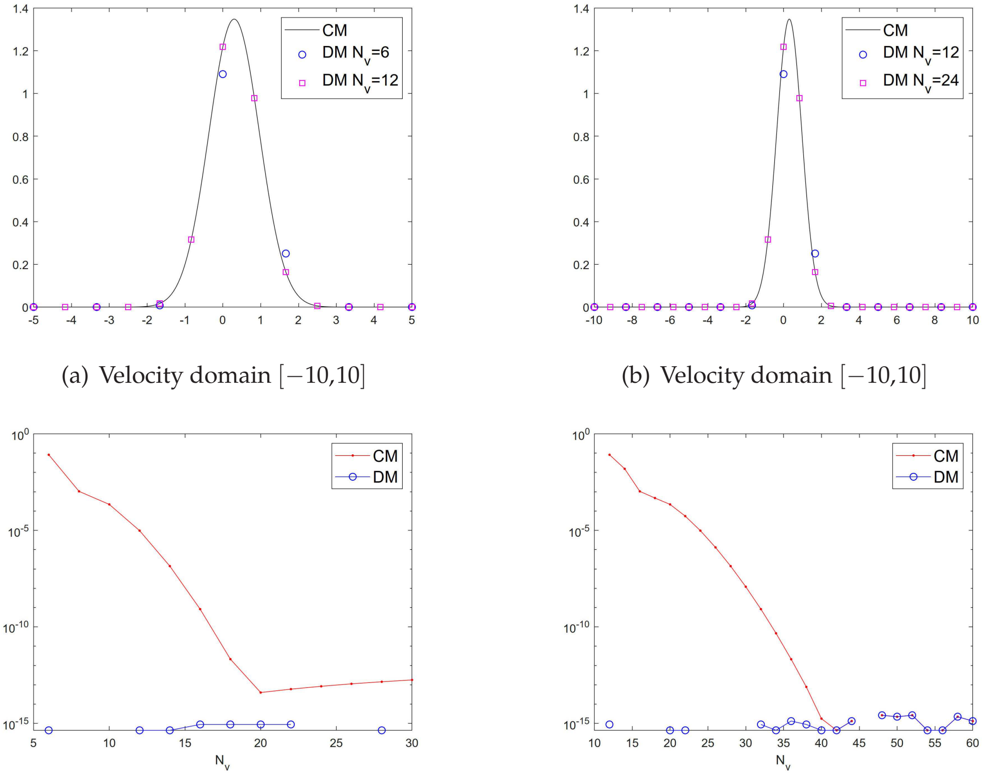
Such issues of the optimal choice of the grid points in velocity space are not considered here and will be left to future investigation.

Here we provide a simple example which demonstrates the usefulness of discrete Maxwellian. Let us consider a set of reference macroscopic moments $(\rho_{ref},u_{ref},E_{ref})=(2.25,0.3,0.6)$. In Fig. 2, we compare continuous and discrete Maxwellians for different velocity domains with various number of velocity grid points N_v. Figs. 2(a)-2(b) show that continuous and discrete Maxwellians get closer as N_v increases. With just 12 grid points there is a good agreement when looking at the pictures.

In Figs. 2(c)-2(d), we compare the computation of the density obtained respectively from the continuous and discrete Maxwellian constructed from the reference macroscopic moments. As expected, the error in the moments computed from the discrete Maxwellian is of the order of the round off error, while the error computed using the continuous Maxwellian decreases spectrally as the number of grid points in velocity increases. The blue circles that do not appear in panels (c) and (d) mean that the error in density is less that 10^{-16}. If the support of the grid in velocity is not sufficiently large, a small error remains even for large values of N_v.

3 High order schemes and conservative correction

Several techniques can be adopted to obtain high order accuracy and to ensure the shock capturing properties near the fluid regime, avoiding spurious oscillations. Here we consider two of the schemes adopted in [17], namely third order schemes obtained by com-



(a) Velocity domain $[-10,10]$

(b) Velocity domain $[-10,10]$

(c) The difference of ρ_{ref} and ρ for velocity domain $[-5,5]$

(d) The difference of ρ_{ref} and ρ for velocity domain $[-10,10]$

Figure 2: Comparison of the continuous and discrete Maxwellians.

bining high order methods in time (RK3 and BDF3) with a high order non-oscillatory spatial interpolation technique that we call generalized WENO (G-WENO) [8].

We repeat the same moving shock test using third order schemes, and the results are summarized in Table 2 for SL schemes using Runge-Kutta time advancement (RK3-W35), and in Table 3 for the BDF-based SL schemes (BDF3-W35). Fully resolved high order schemes both in space and velocity produce finite conservation error, which is much larger than the conservation error of the first order scheme, shown in Table 1.

This indicates that there are cases where high order schemes may show even bigger conservation errors compared to those obtained by the first order scheme.

The main qualitative difference between first order and high order methods is that the former uses a fixed stencil for the linear interpolation at the feet of the characteristics, while high order non-oscillatory reconstructions such as G-WENO use a weighted sum of reconstructions on different stencils, the weight depending on the local regularity properties of the function to be reconstructed. As a result, in the first order SL scheme the interpolation weights are the same for all intervals, whereas in high order SL schemes,

Table 2: Test 1. CFL$=2$, $\kappa=10^{-6}$. Conservation errors of discrete moments in the relative L_1 norm for high order schemes applied to single shock problem with velocity domain $[-20,20]$.

	Classical RK3-W35-CM			Classical RK3-W35-DM		
(N_x,N_v)	Mass	Momentum	Energy	Mass	Momentum	Energy
(100,42)	1.28e-03	1.25e-02	1.40e-01	1.22e-03	1.29e-02	1.47e-02
(100,50)	1.06e-03	1.31e-02	1.47e-02	1.06e-03	1.36e-02	1.47e-02
(100,60)	1.43e-03	1.26e-02	1.49e-02	1.43e-03	1.26e-02	1.49e-02
(100,90)	1.35e-03	1.28e-02	1.48e-02	1.35e-03	1.28e-02	1.48e-02
(200,42)	1.54e-03	1.30e-02	1.45e-02	1.48e-03	1.33e-02	1.52e-02
(200,50)	1.30e-03	1.35e-02	1.51e-02	1.30e-03	1.35e-02	1.51e-02
(200,60)	1.68e-03	1.30e-02	1.53e-02	1.68e-03	1.30e-02	1.53e-02
(200,90)	1.60e-03	1.32e-02	1.53e-02	1.60e-03	1.32e-02	1.53e-02
(400,42)	1.68e-03	1.32e-02	1.47e-02	1.61e-03	1.35e-02	1.54e-02
(400,50)	1.42e-03	1.36e-02	1.53e-02	1.42e-03	1.36e-02	1.53e-02
(400,60)	1.80e-03	1.32e-02	1.55e-02	1.80e-03	1.32e-02	1.55e-02
(400,90)	1.73e-03	1.34e-02	1.54e-02	1.73e-03	1.34e-02	1.54e-02
(800,60)	1.86e-03	1.33e-02	1.55e-02	1.86e-03	1.33e-02	1.55e-02
(800,90)	1.80e-03	1.34e-02	1.55e-02	1.80e-03	1.34e-02	1.55e-02

Table 3: CFL$=2$, $\kappa=10^{-6}$. Conservation errors of discrete moments in the relative L_1 norm for high order schemes applied to single shock problem with velocity domain $[-20,20]$.

	Classical BDF3-W35-CM			Classical BDF3-W35-DM		
(N_x,N_v)	Mass	Momentum	Energy	Mass	Momentum	Energy
(100,42)	1.73e-03	1.03e-02	1.39e-02	1.72e-03	1.03e-02	1.39e-02
(100,50)	1.58e-03	1.04e-02	1.38e-02	1.58e-03	1.04e-02	1.38e-02
(100,60)	1.73e-03	1.00e-02	1.38e-02	1.73e-03	1.00e-02	1.38e-02
(100,90)	1.75e-03	1.03e-02	1.40e-02	1.75e-03	1.03e-02	1.40e-02
(200,42)	2.02e-03	1.10e-02	1.46e-02	2.01e-03	1.10e-02	1.46e-02
(200,50)	1.88e-03	1.11e-02	1.45e-02	1.88e-03	1.11e-02	1.45e-02
(200,60)	2.01e-03	1.07e-02	1.44e-02	2.01e-03	1.07e-02	1.44e-02
(200,90)	2.03e-03	1.10e-02	1.46e-02	2.03e-03	1.10e-02	1.46e-02
(400,42)	2.18e-03	1.14e-02	1.49e-02	2.18e-03	1.14e-02	1.49e-02
(400,50)	2.05e-03	1.15e-02	1.48e-02	2.05e-03	1.15e-02	1.48e-02
(400,60)	2.16e-03	1.11e-02	1.47e-02	2.16e-03	1.11e-02	1.47e-02
(400,90)	2.19e-03	1.14e-02	1.49e-02	2.19e-03	1.14e-02	1.49e-02
(800,60)	2.24e-03	1.13e-02	1.49e-02	2.24e-03	1.13e-02	1.49e-02
(800,90)	2.27e-03	1.16e-02	1.51e-02	2.27e-03	1.16e-02	1.51e-02

due to the nonlinearity of the non-oscillatory reconstruction, the interpolation weights are not the same for all intervals, thus destroying the translation invariance which is at the basis of the conservation property of the schemes.

3.1 Conservative correction and discrete Maxwellian

In Subsection 2.3, we achieved a machine precision conservation error for first order scheme by implementing the discrete Maxwellian in place of the continuous one. This remedy, however, is not sufficient in high order implementations, as was indicated in Tables 2 and 3.

To overcome this, we modify the scheme (2.5) using the conservative correction procedure based on a flux difference form [23,26] to derive our main scheme.

For clarity of exposition, we start by describing the procedure in the case of first order schemes, although its real benefit appears in its application to high order methods.

The conservative method can be viewed as a predictor-corrector method. It is based on a SL non-conservative prediction, and a conservative correction.

Referring to Fig. 3, the first order scheme with conservative correction works as follows:

1. using (2.5), predict $f_{i,j}^{(1)}$ from $\{f_{i,j}^n\}$ at time t^{n+1};

2. reconstruct $\widehat{F}_{i+\frac{1}{2},j}^{(1)}$ and $\widehat{F}_{i-\frac{1}{2},j}^{(1)}$ from $\{v_j f_{i,j}^{(1)}\}$, by using a suitable high order reconstruction (see Section 3.2);

3. compute the convective term $f{*}_{i,j}^{n+1}$ by the conservative scheme

$$f{*}_{i,j}^{n+1} = f_{i,j}^n - \frac{\Delta t}{\Delta x}\left(\widehat{F}_{i+\frac{1}{2},j}^{(1)} - \widehat{F}_{i-\frac{1}{2},j}^{(1)}\right);$$

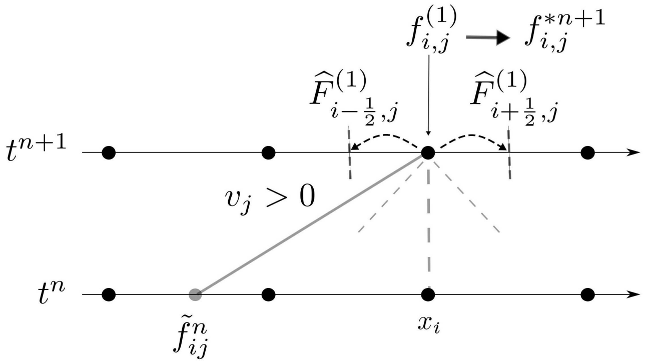

$f_{i,j}^{(1)} \longrightarrow f_{i,j}^{*n+1}$
$\widehat{F}_{i-\frac{1}{2},j}^{(1)}$
$\widehat{F}_{i+\frac{1}{2},j}^{(1)}$
t^{n+1}
$v_j > 0$
t^n
$\tilde{f}_{ij}^n$
x_i

Figure 3: Representation of first order scheme with conservative correction.

4. compute the discrete Maxwellian $d\mathcal{M}^{*n+1}_{i,j}$ from $f^{*n+1}_{i,j}$;

5. update the solution $f^{n+1}_{i,j}$ using

$$f^{n+1}_{i,j} = f^{*n+1}_{i,j} + \frac{\Delta t}{\kappa}(d\mathcal{M}^{*n+1}_{i,j} - f^{n+1}_{i,j}). \tag{3.1}$$

Here $\widehat{F}$ is an accurate reconstruction of the flux vf in the sense of conservative finite difference [31]. We only present the formulation in 1D. Extension to more dimensions can be obtained performing a dimension by dimension 1D reconstruction of the fluxes, as explained in Section 4.

Remark 3.1. The conservative correction imposes severe stability restriction on the CFL number for the C-SL schemes (for a theoretical investigation see also [26]). An accurate analysis for high order Runge-Kutta or BDF C-SL schemes will be given in Section 4.

Remark 3.2. Analysis of semi-Lagrangian schemes for the BGK equation is very challenging. Refs. [28] and [29] deal with this problem for just first order accurate schemes, however the proof that the schemes are convergent uniformly in the Knudsen number is still open to date, as far as we know. In Appendix E, we prove the consistency of both first order schemes appearing in Eqs. (2.3) and (3.1) to the compressible Euler equations in the limit of vanishing Knudsen number. The nonlinear stability analysis and the uniform convergence of the schemes are beyond the scope of the present paper.

3.2 Spatial discretization

We restrict ourselves to 1D case and adopt a uniform grid $\Delta x := x_{i+1} - x_i$.

Flux computation at the feet of the characteristics

We use the Generalized WENO reconstruction (G-WENO) introduced in [8] for non-oscillatory high-order reconstruction of $\tilde{f}^n_{ij}$. The main advantage of such a reconstruction is its use of polynomial weights, which provide a general framework to implement WENO interpolation on any points in a cell. See Appendix B for details.

In our C-SL scheme, we need an accurate approximation of the convection term: $v\partial_x f$. For this, we set $F(f) := vf$, and look for a function $\widehat{F}$ such that

$$F_i = \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \widehat{F} dx, \tag{3.2}$$

where $F_i = F(f(x_i,v,t))$. Then we can compute the convection term using the following relation:

$$\partial_x F_i = \frac{1}{\Delta x}\left(\widehat{F}(x_{i+\frac{1}{2}}) - \widehat{F}(x_{i-\frac{1}{2}})\right).$$

To compute $\widehat{F}(x_{i\pm\frac{1}{2}})$, we use the classical WENO reconstruction in [31] to guarantee non-oscillatory high-order approximation of $\widehat{F}_{i\pm\frac{1}{2}}$. In this reconstruction, we actually find a piecewise polynomial function that interpolates $\{F_i\}_{i=1,\cdots,N_x}$. Since those polynomials contain discontinuity at cell boundaries $x_{i\pm\frac{1}{2}}$, it is necessary to pick the correct direction where information comes from. For this reason, upwinding is introduced by flux splitting:

$$F=F^{+}+F^{-},$$

where

$$F^{+}(f)=\begin{cases} vf, & v>0, \\ 0, & \text{otherwise,} \end{cases} \qquad F^{-}(f)=\begin{cases} 0, & v>0, \\ vf, & \text{otherwise,} \end{cases}$$

so that (3.2) can be rewritten as $F_i=F_i^{+}+F_i^{-}$, where

$$F^{\pm}(x)=\frac{1}{\Delta x}\int_{x-\Delta x/2}^{x+\Delta x/2}\widehat{F}^{\pm}(\xi)d\xi.$$

The half fluxes $\widehat{F}^{\pm}(x)$ are obtained by piecewise polynomial reconstruction:

$$\widehat{F}^{\pm}(x)=\sum_i \chi_i(x)\widehat{F}_i^{\pm}(x),$$

where $\chi_i(x)$ denotes the characteristic function of interval $[x_{i-1/2},x_{i+1/2}]$.

Then, by standard WENO process [31], we reconstruct $\widehat{F}_i^{\pm}(x)$ from $\{F_i^{\pm}\}$. Finally, our numerical flux is obtained as follows:

$$\widehat{F}_{i+\frac{1}{2}}=\widehat{F}_i^{+}(x_{i+1/2})+\widehat{F}_{i+1}^{-}(x_{i+1/2}).$$

3.3 Time discretization

High order discretization in time can be obtained by Runge-Kutta methods (RK) or backward differentiation formulas (BDF) [19]. For the sake of simplicity, we again consider the one-dimensional problem in space and velocity with uniform grid in time.

Runge-Kutta methods

Our system (2.1) becomes *stiff* as $\kappa\to 0$. To overcome this difficulty, we need stable schemes. In view of this, *L-stable diagonally implicit Runge-Kutta* (DIRK) methods provide a balanced performance between stability and efficiency [18].

DIRK methods can be represented using the Butcher's table

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array}$$

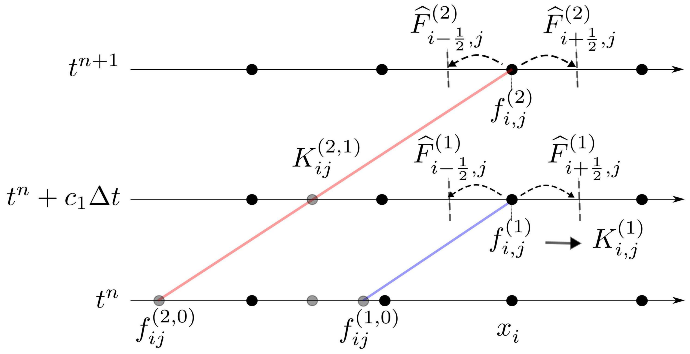

t^{n+1}
$t^n + c_1\Delta t$
t^n
$\widehat{F}^{(2)}_{i-\frac{1}{2},j}$
$\widehat{F}^{(2)}_{i+\frac{1}{2},j}$
$f^{(2)}_{i,j}$
$K^{(2,1)}_{ij}$
$\widehat{F}^{(1)}_{i-\frac{1}{2},j}$
$\widehat{F}^{(1)}_{i+\frac{1}{2},j}$
$f^{(1)}_{i,j} \to K^{(1)}_{i,j}$
$f^{(2,0)}_{ij}$
$f^{(1,0)}_{ij}$
x_i

Figure 4: Representation of DIRK2 scheme with conservative correction.

where $A=(a_{kl})$ is a $s\times s$ lower triangle matrix and $c=(c_1,\cdots,c_s)^T$ and $b=(b_1,\cdots,b_s)^T$ are coefficients vectors [19].

In order to guarantee L-stability, here we make use of *stiffly accurate* schemes (SA), i.e. schemes for which the last row of matrix A is equal to the vector of weights: $a_{s,j}=b_j$, $j=1,\cdots,s$. This will ensure that the absolute stability function vanishes at infinity. As a consequence, an A -stable scheme which is SA is also L-stable [18].

Now, we illustrate our L-stable DIRK schemes to approximate the characteristic system (2.1) coupled with the conservative correction and the discrete Maxwellian. Let us consider the backward characteristic curve to (2.1), corresponding to stage k, which passes through the location x_i with velocity v_j at time $t^n+c_k\Delta t$. We hereafter call it k-th characteristic for each $k=1,\cdots,s$. For example, in Fig. 4 the blue and red lines respectively stand for 1-st and 2-nd characteristics associated to DIRK2 method. In the following, $f_{ij}^{(k,\ell)}$, $\ell=0,\cdots,s$, denotes the ℓ-th stage value computed along the k-th characteristic corresponding to each x_i and v_j (see Fig. 4). For example, in the case of $\ell=0$, $f_{ij}^{(k,0)}$ is the approximation of $f(x_i-c_k\Delta t v_j,v_j,t^n)$ reconstructed from $\{f_{i,j}^n\}$. The k-the stage RK flux $K_{i,j}^{(k)}$ is defined by

$$K_{i,j}^{(k)}=\frac{1}{\kappa}\left(d\mathcal{M}_{i,j}^{(k)}-f_{i,j}^{(k)}\right),\quad k=1,\cdots,s.$$

Denoting the k-th stage characteristic foot of the ℓ-th characteristic by

$$x_{ij}^{(\ell,k)}:=x_i-(c_\ell-c_k)v_j\Delta t,\quad \ell=k+1,\cdots,s,$$

we define the k-th stage RK flux in $x_{ij}^{(\ell,k)}$ by $K_{ij}^{(\ell,k)}$ which is computable from $\{K_{i,j}^{(k)}\}$.

3.3.1 Algorithm DIRK

- *Non-conservative step*
  For $k=1,\cdots,s$

  1. Compute $f_{ij}^{(k,0)}$ in $x_{ij}^{(k,0)}:=x_i-c_kv_j\Delta t$ along the k-th characteristic by interpolation from $\{f^n_{i,j}\}$ with a suitable *generalized* WENO reconstruction in [8].
  2. Compute:

  $$f_{i,j}^{(k)}=f_{ij}^{(k,0)}+\Delta t\sum_{\ell=1}^{k-1}a_{k\ell}K_{ij}^{(k,\ell)}+\frac{\Delta t}{\kappa}a_{kk}\left(d\mathcal{M}_{i,j}^{(k)}-f_{i,j}^{(k)}\right),$$

  where $d\mathcal{M}_{i,j}^{(k)}$ is computed imposing, within some tolerance, that

  $$\sum_j\phi_j d\mathcal{M}_{i,j}^{(k)}\Delta v=\sum_j\phi_j\big(f_{ij}^{(k,0)}+\Delta t\sum_{\ell=1}^{k-1}a_{k\ell}K_{ij}^{(k,\ell)}\big)\Delta v,$$

  for $\phi_j=1,v_j,v_j^2/2$.
  3. Compute:

  $$K_{i,j}^{(k)}=\frac{1}{\kappa}\left(d\mathcal{M}_{i,j}^{(k)}-f_{i,j}^{(k)}\right).$$

  4. Compute the RK flux $K_{ij}^{(\ell,k)}$ in $x_{ij}^{(\ell,k)}:=x_i-(c_\ell-c_k)v_j\Delta t$ with $\ell=k+1,\cdots,s$ along the ℓ-th characteristic by interpolation from $\{K_{i,j}^{(k)}\}$ with a suitable *generalized* WENO reconstruction in [8].
  5. Reconstruct $\widehat{F}_{i+1/2,j}^{(k)}$ from $\{v_jf_{i,j}^{(k)}\}$ using WENO reconstruction [31] within a *finite difference formulation* (fd).

  end

- *Conservative correction step*

  1. Compute the conservative convection:

  $$f_{i,j}^{*}=f_{i,j}^{n}-\frac{\Delta t}{\Delta x}\sum_{\ell=1}^{s}b_\ell\left(\widehat{F}_{i+1/2,j}^{(\ell)}-\widehat{F}_{i-1/2,j}^{(\ell)}\right).$$

  2. Compute conservative solution:

  $$f_{i,j}^{n+1}=f_{i,j}^{*}+\Delta t\sum_{\ell=1}^{s-1}b_\ell K_{i,j}^{(\ell)}+\frac{\Delta t}{\kappa}b_s\left(d\mathcal{M}_{i,j}^{(*)}-f_{i,j}^{n+1}\right), \tag{3.3}$$

where $d\mathcal{M}_{i,j}^{(*)}$ is computed imposing, within some tolerance, that

$$\sum_j \phi_j d\mathcal{M}_{i,j}^{(*)} \Delta v = \sum_j \phi_j f_{i,j}^* \Delta v, \quad \phi_j = 1, v_j, v_j^2/2.$$

In Eq. (3.3), the moments of $K_{i,j}^{(\ell)}$, $\ell = 1,\cdots,s-1$, vanish:

$$\sum_{j=0}^{N_v+1} K_{i,j}^{(\ell)} \phi_j \Delta v = \frac{1}{\kappa} \sum_{j=0}^{N_v+1} \left(d\mathcal{M}_{i,j}^{(\ell)} - f_{i,j}^{(\ell)} \right) \phi_j \Delta v = 0,$$

because for each node x_i the ℓ-th stage values of Maxwellians and distribution functions have the same moments. A schematic representation of DIRK2 is illustrated in Fig. 4.

BDF methods

Another time discretization we use for the stable approximation of stiff problems (2.1) is the backward differentiation formula (BDF) (see [18]) whose general form is given by

$$BDF: y^{n+1} = \sum_{k=1}^{s} a_k y^{n+1-k} + \beta_s \Delta t g(y^{n+1}, t_{n+1})$$

with $\beta_s \neq 0$. For our work, we use BDF2 and BDF3:

$$\begin{aligned} &\text{BDF2:} \quad y^{n+1} = \frac{4}{3} y^n - \frac{1}{3} y^{n-1} + \frac{2}{3} \Delta t g(y^{n+1}, t_{n+1}), \\ &\text{BDF3:} \quad y^{n+1} = \frac{18}{11} y^n - \frac{9}{11} y^{n-1} + \frac{2}{11} y^{n-2} + \frac{6}{11} \Delta t g(y^{n+1}, t_{n+1}). \end{aligned}$$

BDF schemes have some advantages over DIRK since a smaller number of numerical evaluation of the discrete Maxwellian and fluxes are needed and fewer interpolations are required. For BDF2 and BDF3, there is only one stage in which we have to compute the discrete Maxwellian and fluxes while two and three stages are required for DIRK2 and DIRK3 schemes respectively. Moreover, BDF2 and BDF3 schemes require two and three steps for interpolations whereas DIRK2 and DIRK3 schemes require three and six steps respectively. The price to pay is that BDF has more severe stability restriction than DIRK (see Section 5).

3.3.2 Algorithm BDF

Let a_k, and β_s be the coefficients of a BDF method of order s. Given a discrete approximation $\{f_{ij}^n\}$ of the distribution function at time t_n, $\{f_{ij}^{n+1}\}$ is computed by the following steps

- *Non-conservative step*

1. For $k=1,\cdots,s$, interpolate $f_{ij}^{n,k}=f(x_i-kv_j\Delta t,v_j,t^{n+1-k})$ in $x_{ij}^k:=x_i-kv_j\Delta t$ from $\{f_{i,j}^{n+1-k}\}$ with a suitable *generalized* WENO reconstruction in [8].
2. Compute $f_{i,j}^*=\sum_{k=1}^s a_k f_{ij}^{n,k}$ and

$$f_{i,j}^{(1)}=f_{i,j}^*+\beta_s\frac{\Delta t}{\kappa}\left(d\mathcal{M}_{i,j}^{(1)}-f_{i,j}^{(1)}\right),$$

where $d\mathcal{M}_{i,j}^{(1)}$ is computed imposing, within some tolerance, that

$$\sum_j \phi_j d\mathcal{M}_{i,j}^{(1)}\Delta v=\sum_j \phi_j f_{i,j}^*\Delta v,\quad \phi_j=1,v_j,v_j^2/2. \tag{3.4}$$

- *Conservative correction step*

1. Reconstruct $\widehat{F}_{i+1/2,j}^{(1)}$ from $\{v_j f_{i,j}^{(1)}\}$ using WENO reconstruction in the framework of the *conservative finite difference formulation* (fd) [31].
2. Conservative convection:

$$f_{i,j}^{**}=\sum_{k=1}^s a_k f_{i,j}^{n+1-k}-\beta_s\frac{\Delta t}{\Delta x}\left(\widehat{F}_{i+1/2,j}^{(1)}-\widehat{F}_{i-1/2,j}^{(1)}\right).$$

3. Compute conservative solution:

$$f_{i,j}^{n+1}=f_{i,j}^{**}+\beta_s\frac{\Delta t}{\kappa}\left(d\mathcal{M}_{i,j}^{n+1}-f_{i,j}^{n+1}\right).$$

Note that $d\mathcal{M}_{i,j}^{n+1}=d\mathcal{M}(f_{i,j}^{**})$ as in (3.4).

A schematic representation of BDF2 is illustrated in Fig. 5.

The next section is devoted to the stability analysis of RK and BDF schemes applied to the linear advection equation.

4 Linear stability analysis

In this section we perform the stability analysis of conservative semi-Lagrangian scheme for the 1D advection equation. Following [26] we consider the linear transport equation

$$u_t+vu_x=0,\quad u(x,0)=u_0(x),\quad v\in\mathbb{R}. \tag{4.1}$$

For simplicity, we assume a periodic boundary condition and $x\in[-\pi,\pi]$.

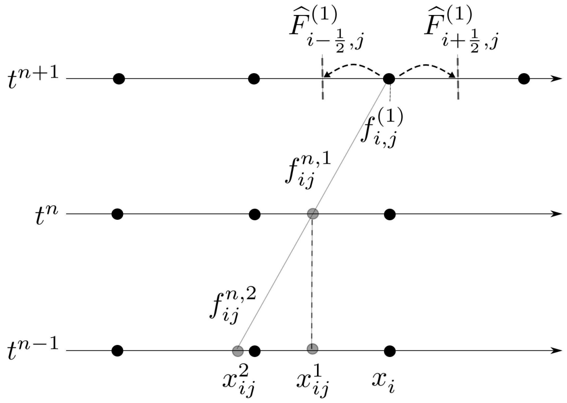


Figure 5: Representation of BDF2 scheme with conservative correction. Black circles: grid nodes, grey circles: points where interpolation is needed.

4.1 DIRK methods

Algorithm 3.3.1 applied to (4.1) gives

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x}\sum_{\ell=1}^{s} b_\ell v\left(\widehat{u}^{(\ell)}_{j+1/2} - \widehat{u}^{(\ell)}_{j-1/2}\right), \tag{4.2}$$

where $\widehat{u}^{(\ell)}_{j+1/2}$ are obtained from the stage values $u_j^{(\ell)} = u^n(x_j - vc_\ell\Delta t)$ by reconstruction, and $u^n(x)$ denotes a suitable interpolation from $\{u_j^n\}$.

4.1.1 Fourier interpolation

We look for the evolution of a Fourier mode of the form

$$u_j^n = \rho^n e^{ikj\Delta x} = \rho^n e^{ij\xi}, \quad \xi = k\Delta x, \quad i = \sqrt{-1}.$$

In the analysis we first consider Fourier interpolation, so

$$u^n(x) = \rho^n e^{ikx} = \rho^n e^{i\xi x/\Delta x}, \quad \xi \in [-\pi,\pi], \tag{4.3}$$

where $\rho^n = \rho^n(\xi)$ is the amplification factor associated with mode ξ. Plugging such *ansatz* into the stage values, we get

$$u_j^{(\ell)} = \rho^n \exp\left(i\xi(x_j - v\Delta t c_\ell)/\Delta x\right) = \rho^n e^{ij\xi} e^{-ic_\ell a\xi},$$

where $a = v\Delta t/\Delta x$ denotes the CFL number. In [31], the relation between $u(x)$ and $\widehat{u}$ is given by

$$\frac{\widehat{u}(x+\Delta x/2) - \widehat{u}(x-\Delta x/2)}{\Delta x} = \frac{\partial u}{\partial x}(x).$$

Using this relation and (4.3) one has

$$\widehat{u}^n(x) = \frac{u^n(x)}{\operatorname{sinc}(\xi/2)}, \tag{4.4}$$

where $\operatorname{sinc}(x) = \sin(x)/x$.

Making use (4.4) and (4.3) in Eq. (4.2), one obtains the following formula for the amplification factor:

$$\rho(\xi) = 1 - i\xi a \sum_{\ell=1}^{s} b_\ell \exp(-ic_\ell a\xi). \tag{4.5}$$

The scheme is stable if $|\rho(\xi)| \le 1$ for all $\xi \in [-\pi,\pi]$.

Such stability problem is closely related to the linear stability of the quadrature formula when applied to the approximation of the integral form of a scalar linear ODE,

$$y' = \lambda y, \quad y(0) = 1, \quad \forall \lambda \in \mathbb{C}.$$

In fact, the solution after one step of this ODE is: $y(\Delta t) = e^{\lambda \Delta t} = e^z$, where $z := \Delta t \lambda$. Such solution is stable iff $\mathcal{R}(z) \le 0$, i.e. if $\mathcal{R}(\lambda) \le 0$. Considering the following identity

$$e^z = 1 + z\int_0^1 e^{cz}dc$$

and approximating the integral by a quadrature formula with nodes c_ℓ and weights b_ℓ, one obtains the approximation of the exact solution after one step:

$$R(z) = 1 + z\sum_{\ell=1}^{s} b_\ell e^{c_\ell z}, \tag{4.6}$$

with which the stability region can be drawn by the set $\{z \in \mathbb{C} : |R(z)| \le 1\}$. Comparing Eq. (4.5) with (4.6), ones has $\rho(\xi) = R(-ia\xi)$ with $\xi \in [\pi,\pi]$. Thus the stability of a quadrature formula in a conservative semi-Lagrangian scheme for a linear advection equation is closely related to the stability on the imaginary axis. Then in order to guarantee stability we look of the largest interval $I^* = [-y^*, y^*]$ of the imaginary axis such that $|R(iy)| \le 1$ $\forall y \in I^*$. Note that the bound $a^* = y^*/\pi$ quantifies the maximum CFL number for the semi-Lagrangian scheme that guarantees stability.

Now in order to maximize the stability interval on imaginary axis, we construct quadrature formulas that allow a wide stability region. Let us consider the expression $R(iy)$ and write it in the form

$$R(iy) = 1 + iy(C_s(y) + iS_s(y)) = 1 - yS_s(y) + iyC_s(y),$$

where

$$C_s(y) = \sum_{\ell=1}^{s} b_\ell \cos(c_\ell y), \quad S_s(y) = \sum_{\ell=1}^{s} b_\ell \sin(c_\ell y).$$

The stability condition therefore becomes

$$|R(iy)|^2 = 1-2yS_s(y)+y^2(C_s^2(y)+S_s^2(y)) \leq 1.$$

Such condition can be written in the form

$$yF_s(y) \geq 0, \quad \text{where} \quad F_s(y) := S_s(y) - \frac{1}{2}\left(C_s^2(y)+S_s^2(y)\right). \tag{4.7}$$

Then the problem to find quadrature formulas with the widest stability region is connected to determine the coefficients $\mathbf{b}=(b_1,\cdots,b_s)$ and $\mathbf{c}=(c_1,\cdots,c_s)$ so that the interval in which (4.7) is satisfied is the widest. The analysis of quadrature formulas with even s and symmetric distribution of nodes around the middle of the interval is performed in [26].

Here we numerically compute nodes and weights for a particular class of third order DIRK schemes that satisfy the simplification conditions

$$\sum_{j=1}^{s} a_{ij} = c_i, \quad i=1,\cdots,s$$

and which is stiffly accurate, i.e. for which the last row of the A-matrix coincides with the weights, $a_{s,j}=b_j$, $j=1,\cdots,s$. This constraint is imposed in order to have L-stable schemes, in view of the AP property in the fluid dynamic regime. Such schemes have the following structure:

$$\begin{array}{c|ccc} c_1 & c_1 & 0 & 0 \\ c_2 & c_2-\gamma_2 & \gamma_2 & 0 \\ 1 & b_1 & b_2 & b_3 \\ \hline & b_1 & b_2 & b_3 \end{array} \tag{4.8}$$

The coefficients of the scheme are determined taking into account the following requirements:

- the scheme has to be at least third order accurate;
- the scheme has to be A-stable (and therefore L-stable, because it is Stiffly Accurate, (SA) i.e. $a_{si}=b_i$ for $i=1,2,3$, see [18]);
- nodes and weight are selected in such a way that condition (4.7) is satisfied for a wide region.

Order conditions for scheme (4.8), up to third order accuracy, are:

$$\sum_{i=1}^{s} b_i = 1, \quad \sum_{i=1}^{s} b_i c_i = 1/2, \quad \sum_{i=1}^{s} b_i c_i^2 = 1/3, \quad \sum_{i,j=1}^{s} b_i a_{ij} c_j = 1/6. \tag{4.9}$$

Solving these equations allows to express four parameters of the scheme as a function of c_1 and c_2:

$$b_2 = \frac{3c_1-1}{6(c_2-c_1)(c_2-1)}, \quad b_3 = \frac{6c_1c_2-3c_1-3c_2+2}{6(c_2-1)(c_1-1)}, \quad \gamma_2 = \frac{6c_1^2c_2-4c_1c_2-c_1+c_2}{2(3c_1-1)(c_1-1)} \tag{4.10}$$

and $b_1 = 1 - b_2 - b_3$. This leaves two free parameters, which are chosen according to the two additional conditions.

In order to impose A-stability, from [18], we recall the following result.

An implicit R-K method is A-stable iff

1. the stability function $R(z) = P(z)/Q(z)$ is analytic in $\mathbb{C}$ for $Re(z) < 0$;

2. the method is I-stable, i.e. $|R(iy)| \leq 1$ for all $y \in \mathbb{R}$ (stability on the imaginary axis).

The I-stability is equivalent to the request that the polynomial

$$E(y) = |Q(iy)|^2 - |P(iy)|^2 = \sum_{j=0}^{s} E_{2j} y^{2j} \tag{4.11}$$

satisfies $E(y) \geq 0$ for all $y \in \mathbb{R}$ and $i = \sqrt{-1}$.

Performing a detailed calculation (reported in Appendix C), the condition for I-stability (4.11) and (4.7) becomes: either

$$c_1 < 1/3, \quad c_2 > 1 \tag{4.12}$$

or

$$c_1 > 1/3, \quad c_2 < 1.$$

The latter has to be excluded since it implies $b_3 < 0$ and condition (1) for the analyticity of the function $R(z)$ above is not satisfied.

Then DIRK scheme (4.8) is A-stable and by the SA property, it's also L-stable.

Remark 4.1. If we look for a third order Singly DIRK scheme i.e. with $\gamma_1 = \gamma_2 = \gamma_3 = \gamma$, as in [3], then there are no free parameters, and one obtains $\gamma \simeq 0.4358665215$ and $\delta = \frac{3}{2}\gamma^2 - 5\gamma + \frac{5}{4} \simeq 644363171$. This scheme is A-stable and L-stable, but the weights and the nodes do not satisfy condition (4.7) for any $y > 0$, i.e. scheme (4.2) is not stable.

This remark suggests to look for DIRK methods as (4.8) such that $\gamma_1 = \gamma_3 = \gamma$, i.e.

$$\begin{array}{c|ccc} \gamma & \gamma & 0 & 0 \\ c_2 & c_2 - \gamma_2 & \gamma_2 & 0 \\ 1 & 1 - b_2 - \gamma & b_2 & \gamma \\ \hline & 1 - b_2 - \gamma & b_2 & \gamma \end{array} \tag{4.13}$$

From (4.9), we have four equations with five unknowns b_2, c_2 γ_2, γ and b_1 and from (4.10), with $c_1 = b_3 = \gamma$, we compute b_2, c_2 and γ_2 as functions of γ and $b_1 = 1 - b_2 - \gamma$.

Performing a detailed calculation (reported in Appendix C), we require to choose γ in the union of the following intervals

$$]1 - \sqrt{2}/2, 1/3[, \quad]1 + \sqrt{2}/2, +\infty[. \tag{4.14}$$

Note that the second interval can not be accepted because this implies values of γ larger than $1 + \sqrt{2}/2 \approx 1.70710\cdots$, and this is in contradiction with the hypothesis (4.12).

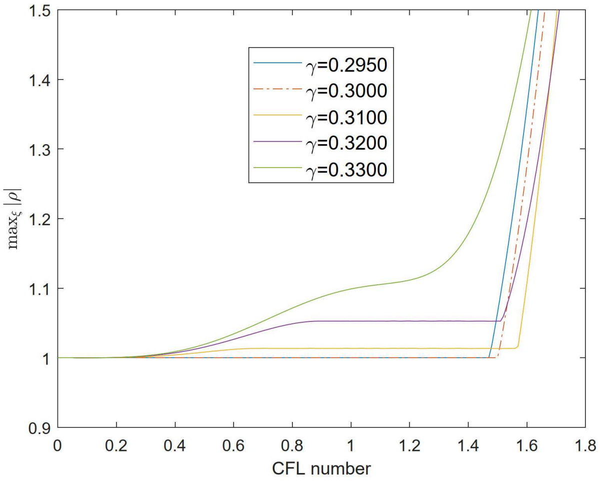


Figure 6: Optimal choice of γ for scheme (4.13).

Numerical experiments show that the optimal value of γ in the first interval is approximately $\gamma=0.3$ (see Fig. 6). Then for this choice of γ, the coefficients of scheme (4.13) are: $\gamma=0.3$, $\gamma_2=13/3$, $b_2=-3/710$ and $c_2=8/3$. This scheme is stable under the condition (4.7) for $y\leq y^*=4.715426442$ with $a^*\approx 1.5$ and is also L-stable.

For the numerical experiments, we use the following two types of DIRK methods. The first is a second order DIRK scheme (DIRK2) [3]

$$\text{DIRK2} = \begin{array}{c|cc} \alpha & \alpha & 0 \\ 1 & 1\text{-}\alpha & \alpha \\ \hline & 1\text{-}\alpha & \alpha \end{array}$$

where $\alpha=1-1/\sqrt{2}$. This scheme is stable under condition (4.7) for $y\leq y^*=4.586275880$ with $a^*\approx 1.46$ and is also L-stable. The second one is the third order DIRK scheme (DIRK3) (4.13).

4.2 BDF schemes

Now apply k-th order BDF schemes to system (4.1), and we get

$$u_j^{n+1}=\sum_{\ell=1}^{k} a_\ell u_j^{n-\ell+1}-\beta_k v\frac{\Delta t}{\Delta x}\left(\widehat{u}_{j+1/2}^{n+1}-\widehat{u}_{j-1/2}^{n+1}\right), \tag{4.15}$$

and by (4.4)

$$\widehat{u}_{j+1/2}^{n+1}=\frac{\tilde{u}^{n+1}(x_{j+1/2})}{\text{sinc}(\xi/2)}, \tag{4.16}$$

with

$$\tilde{u}^{n+1}(x_j)=\sum_{\ell=1}^{k} a_\ell u^n(x_j-v\ell\Delta t). \tag{4.17}$$

Then we look for the evolution of the Fourier mode identified by the parameter $\xi\in[-\pi,\pi]$. We set $u^n(x)=\rho^n e^{ikx}=\rho^n e^{i\xi x/\Delta x}$ so that $u^n(x_j)=u_j^n=\rho^n e^{ij\xi}$. Then (4.17) becomes

$$\tilde{u}^{n+1}(x_j)=\sum_{\ell=1}^{k} a_\ell \rho^{n-\ell+1} e^{i\xi(x_j-\ell v\Delta t)/\Delta x},$$

and (4.16) becomes

$$\widehat{u}^{n+1}(x_{j+1/2})=\left(\sum_{\ell=1}^{k} a_\ell \rho^{n-\ell+1} e^{ij\xi} e^{-i\xi a} e^{i\xi/2}\right)/\operatorname{sinc}(\xi/2).$$

After some algebraic manipulation, we obtain for (4.15):

$$\rho^{n+1}=\sum_{\ell=1}^{k} a_\ell \rho^{n-\ell+1}\left(1-\beta_k a e^{-i\xi\ell a} i\xi\right), \tag{4.18}$$

where $a=v\Delta t/\Delta x$. Then the characteristic polynomial associated to (4.18) is:

$$p(\rho)=\rho^k-\sum_{\ell=1}^{k} a_\ell \rho^{k-\ell}\left(1-\beta_k a e^{-i\xi\ell a} i\xi\right).$$

Now again we compute the maximum a^* such that

$$\max_{\xi\in[-\pi,\pi]}|\rho(a,\xi)|\leq 1, \quad \forall a\in[0,a^*]. \tag{4.19}$$

Here $\rho(a,\xi)$ represents the largest root in absolute value of the polynomial $p(\rho)$. In particular, we consider the two BDF schemes BDF2 and BDF3 with $k=2$ and $k=3$, respectively. We compute numerically (4.19) and we get for BDF2 $a^*\approx 0.5678$, while BDF3 is unstable for each $a>0$.

This analysis confirms that the conservative correction imposes stability restriction on the CFL number a^* for the BDF methods.

We conclude that C-SL schemes based on RK framework have better stability properties that those based on BDF when applied to linear advection equation (4.1).

The result of the linear stability analysis performed on the advection equation is that there are CFL restrictions imposed by the conservative reconstruction. We observe that in practice CFL number higher that those predicted theoretically can be adopted, because of the stabilising effect of the collision term.

4.3 Collision stabilization

We analyse the stabilisation introduced by a dissipative term on the right hand side by considering the model equation

$$u_t+vu_x=-\mu u. \tag{4.20}$$

We limit to simplest case corresponding to implicit Euler scheme, so $s=1$, $b_1=c_1=1$.

First compute the non-conservative step:

$$u_j^{(1)}=u^n(x_j-v\Delta t)-\mu\Delta t u_j^{(1)},$$

where $u^n(x)$ denotes a suitable reconstruction of the solution at time $n\Delta t$. This gives

$$u_j^{(1)}=\frac{u^n(x_j-v\Delta t)}{1+\mu\Delta t}.$$

Now, computing the conservative correction step related to (4.20), we obtain

$$u_j^*=u_j^n-\left(\frac{1}{1+\mu\Delta t}\right)v\frac{\Delta t}{\Delta x}\left(\widehat{u}_{j+1/2}^{(1)}-\widehat{u}_{j-1/2}^{(1)}\right). \tag{4.21}$$

Finally we compute the conservative solution

$$u_j^{n+1}=u_j^*-\mu\Delta t u_j^{n+1}.$$

Plugging (4.21) into the previous formula we get

$$u_j^{n+1}=\frac{(1+\mu\Delta t)u_j^n-\frac{\Delta t}{\Delta x}v\left(\widehat{u}_{j+1/2}^{(1)}-\widehat{u}_{j-1/2}^{(1)}\right)}{(1+\mu\Delta t)^2}.$$

We now repeat the Fourier stability analysis introduced in Subsection 4.1.1. Making use of (4.4) and (4.3) the expression of the amplification factor becomes

$$\rho(a\xi,\mu\Delta t)=\frac{1+\mu\Delta t-ia\xi\exp(-ia\xi)}{(1+\mu\Delta t)^2}, \tag{4.22}$$

where $a=v\Delta t/\Delta x$ denotes the Courant number and $\xi\in[-\pi,\pi)$ denotes the angle corresponding to a particular Fourier mode. The stability region is obtained by imposing that $|\rho(\xi)|\le 1$. For $\mu=0$ it is

$$|\rho|^2=1+y^2-2y\sin y,$$

where $y=a\xi$. Now $|\rho^2|\le 1\Leftrightarrow y(y-2\sin y)\le 0$ therefore the scheme is stable for $y\in[0,y^*]$, with $y^*\approx 1.895494267033981$, corresponding to a maximum allowed Courant number $a^*=y^*/\pi\approx 0.603354564401614$. When $\mu>0$ the stability region is obtained by imposing that the amplification factor in (4.22) is bounded by 1:

$$\max_{-\pi\le\xi<\pi}|\rho^2(a\xi,\mu\Delta t)|\le 1. \tag{4.23}$$

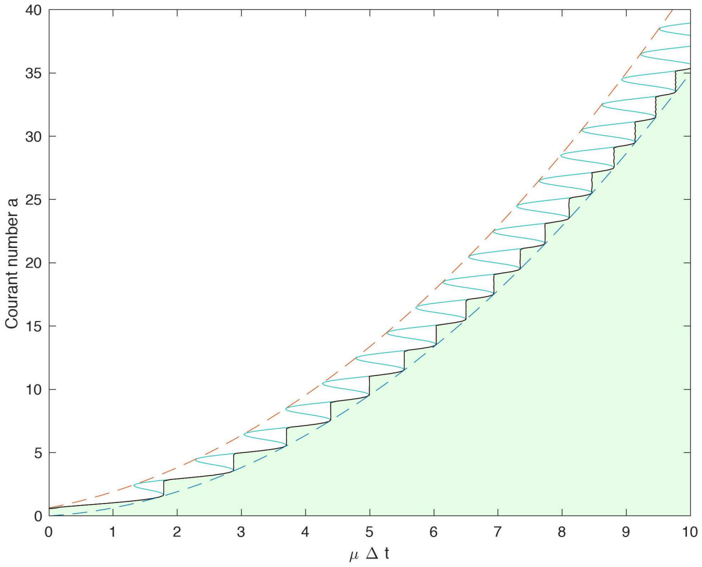


Figure 7: Contour plot for the magnitude of amplification factor ρ in (4.22). The zigzag light blue curve represents the line for which $|\rho^2(a\pi,\mu\Delta t)|=1$, the light green region is the one for which condition (4.23) is satisfied. Note that the zigzag curve is between red and blue dashed lines corresponding, respectively, to $\pi a=(1+\mu\Delta t)^2+(1+\mu\Delta t)$ and $\pi a=(1+\mu\Delta t)^2-(1+\mu\Delta t)$.

The stability region in the $\mu\Delta t$-a plane is the shaded green one illustrated in Fig. 7. In particular, if $a<(\mu\Delta t)^2+\mu\Delta t$ then the scheme is stable. From the analysis it follows that for large values of $\mu\Delta t$, the maximum allowed CFL number increases proportionally to $(\mu\Delta t)^2$. A full stability analysis of the scheme for the BGK equation has not been performed and will be subject of future investigation.

As a final remark, it appears from the analysis that for small values of μ, corresponding to larger values of the Knudsen number, the stabilising effect of implicit treatment of the collision term is less pronounced, therefore the stability limitation of the scheme become more severe, so the gain with respect to the use of a conservative Eulerian scheme is not as strong as in the case of small Knudsen number. On the other hand, we expect that in the rarefied regime conservation properties are not so crucial, so classical semi-Lagrangian schemes would be very effective. Furthermore, if the solution is smooth, then classical semi-Lagrangian schemes have very good conservation properties. However, in the rarefied regime the validity of the BGK model itself becomes questionable.

5 Extension of the scheme to two space dimensions

The extension to the two space dimensions is straightforward, since we just need to perform a dimension by dimension 1D interpolation on characteristic foots and 1D reconstruction of fluxes. For the description, let us denote by $i\equiv(i_1,i_2)$ the index for the space variable and the size of mesh is taken by $(\Delta x,\Delta y)$. Similarly, we use the index $j\equiv(j_1,\cdots,j_{d_v})$ for velocity variable, where d_v is the dimension of the velocity domain, and assume a uniform Cartesian mesh in velocity, with the same mesh spacing Δv in all directions. Using a similar notation as in the 1D case, we summarize the s-order DIRK based methods for two space dimensions as follows:

Algorithm DIRK for two space dimension

For $k=1,\cdots,s$.

- *Non-conservative step*

  1. Reconstruct the solution $f_{ij}^{(k,0)}$ on $x_{ij}^{(k,0)}:=(x_{i_1}-c_kv_{j_1}\Delta t,y_{i_2}-c_kv_{j_2}\Delta t)$ along the k-th characteristic with a suitable *generalized* WENO reconstruction in [8] as follows:
     – for each i_2, approximate $f(x_{i_1}-c_kv_{j_1}\Delta t,y_{i_2},v_j,t^n)$ along the x-axis by interpolation from $\{f_{i,j}^n\}$;
     – using the values obtained on $(x_{i_1}-c_kv_{j_1}\Delta t,y_{i_2})$, interpolate $f_{ij}^{(k,0)}$ at the point $x_{ij}^{(k,0)}$ along the y-axis.
  2. Compute:
     $$f_{i,j}^{(k)}=f_{ij}^{(k,0)}+\Delta t\sum_{\ell=1}^{k-1}a_{k\ell}K_{ij}^{(k,\ell)}+\frac{\Delta t}{\kappa}a_{kk}\left(d\mathcal{M}_{i,j}^{(k)}-f_{i,j}^{(k)}\right),$$
     where $d\mathcal{M}_{i,j}^{(k)}$ is computed imposing, within some tolerance, that
     $$\sum_j\phi_j d\mathcal{M}_{i,j}^{(k)}(\Delta v)^{d_v}=\sum_j\phi_j(f_{ij}^{(k,0)}+\Delta t\sum_{\ell=1}^{k-1}a_{k\ell}K_{ij}^{(k,\ell)})(\Delta v)^{d_v},$$
     for $\phi_j=1,v_j,|v_j|^2/2$.
  3. Store the quantity:
     $$K_{i,j}^{(k)}=\frac{1}{\kappa}\left(d\mathcal{M}_{i,j}^{(k)}-f_{i,j}^{(k)}\right)$$
     that was used to compute $f_{i,j}^{(k)}$.
  4. For each $\ell=k+1,\cdots,s$, compute the RK flux $K_{ij}^{(\ell,k)}$ in $x_{ij}^{(\ell,k)}:=(x_{i_1}-(c_\ell-c_k)v_{j_1}\Delta t,x_{i_2}-(c_\ell-c_k)v_{j_2}\Delta t)$ along the ℓ-th characteristic with a suitable *generalized* WENO reconstruction in [8] as follows:

– for each i_2, approximate the RK flux at $(x_{i_1}-(c_\ell-c_k)v_{j_1}\Delta t, y_{i_2})$ along the x-axis by interpolation from $\{K_{i,j}^{(k)}\}$;
– using the values obtained on $(x_{i_1}-(c_\ell-c_k)v_{j_1}\Delta t, y_{i_2})$, interpolate the RK flux at the point $x_{ij}^{(\ell,k)}$ along the y-axis.

5. Reconstruct $\widehat{F}_{i_1+1/2,i_2,j}^{(k)}$ along x-axis and $\widehat{F}_{i_1,i_2+1/2,j}^{(k)}$ along y-axis from $\{v_j f_{i,j}^{(k)}\}$, respectively, using 1D WENO reconstruction [31] within a *finite difference formulation* (fd).

- *Conservative correction step*

1. Compute the conservative convection:

$$f_{i,j}^{*}=f_{i,j}^{n}-\frac{\Delta t}{\Delta x}\sum_{\ell=1}^{s}b_\ell\left(\widehat{F}_{i_1+1/2,i_2,j}^{(\ell)}-\widehat{F}_{i_1-1/2,i_2,j}^{(\ell)}\right)$$
$$-\frac{\Delta t}{\Delta y}\sum_{\ell=1}^{s}b_\ell\left(\widehat{F}_{i_1,i_2+1/2,j}^{(\ell)}-\widehat{F}_{i,i_2-1/2,j}^{(\ell)}\right).$$

2. Compute conservative solution:

$$f_{i,j}^{n+1}=f_{i,j}^{*}+\Delta t\sum_{\ell=1}^{s-1}b_\ell K_{i,j}^{(\ell)}+\frac{\Delta t}{\kappa}b_s\left(d\mathcal{M}_{i,j}^{(*)}-f_{i,j}^{n+1}\right),$$

where $d\mathcal{M}_{i,j}^{(*)}$ is computed imposing, within some tolerance, that

$$\sum_j \phi_j d\mathcal{M}_{i,j}^{(*)}(\Delta v)^{d_v}=\sum_j \phi_j f_{i,j}^{*}(\Delta v)^{d_v}, \quad \phi_j=1,v_j,|v_j|^2/2.$$

As in the 1D case, the terms $K_{i,j}^{(\ell)}$ vanish when we take moments with respect to ϕ_j. A schematic representation of DIRK2 based method for two space dimension is described in Fig. 8. For brevity, we here omit the descriptions for BDF based methods for two space dimensions.

6 Numerical experiments

In this section, we present several tests to verify some properties of the proposed schemes.

In test 1, we compute the conservation error of the schemes. In test 2, we check the correct order of accuracy for smooth solutions for various values of the Knudsen number κ. In test 3, we check the relaxation to equilibrium for small Knudsen number. Test 4 is devoted to verify the shock capturing capability in the Euler limit. Test 5 explores the

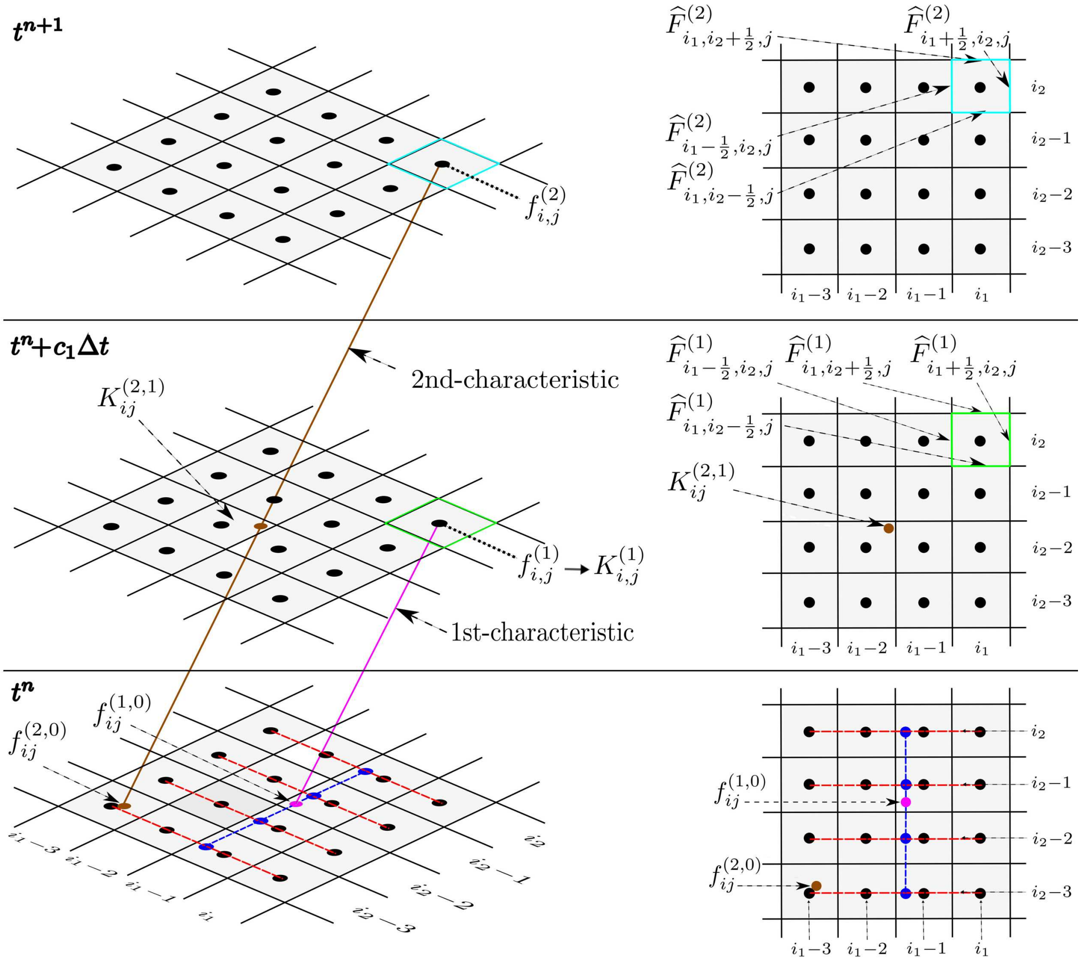

t^{n+1}
$f^{(2)}_{i,j}$
$\widehat{F}^{(2)}_{i_1,i_2+\frac{1}{2},j}$
$\widehat{F}^{(2)}_{i_1+\frac{1}{2},i_2,j}$
$\widehat{F}^{(2)}_{i_1-\frac{1}{2},i_2,j}$
$\widehat{F}^{(2)}_{i_1,i_2-\frac{1}{2},j}$
$t^n+c_1\Delta t$
$K^{(2,1)}_{ij}$
2nd-characteristic
$f^{(1)}_{i,j} \to K^{(1)}_{i,j}$
1st-characteristic
$\widehat{F}^{(1)}_{i_1-\frac{1}{2},i_2,j}$
$\widehat{F}^{(1)}_{i_1,i_2+\frac{1}{2},j}$
$\widehat{F}^{(1)}_{i_1+\frac{1}{2},i_2,j}$
$\widehat{F}^{(1)}_{i_1,i_2-\frac{1}{2},j}$
t^n
$f^{(1,0)}_{ij}$
$f^{(2,0)}_{ij}$
i_1-3
i_1-2
i_1-1
i_1
i_2
i_2-1
i_2-2
i_2-3

Figure 8: Representation of DIRK2 scheme for two space dimension with conservative correction. Given velocity v_j, to interpolate $f^{(1,0)}_{ij}$ on a characteristic foot $x^{(1,0)}_{ij}$ (magenta circle), interpolate first the solutions on blue points on $x=x_{i_1}-c_k v_{j_1}\Delta t$ for $y=y_{i_2-3}$, y_{i_2-2}, y_{i_2-1}, y_{i_2} along x-direction (red line), then use them in the interpolation for $f^{(1,0)}_{ij}$ along y-direction (blue line). The interpolations for $f^{(2,0)}_{ij}$ and $K^{(2,1)}_{ij}$ can be done similarly. The green and cyan colored squares are RK fluxes at time level $t^n+c_1\Delta t$ and t^{n+1} which are computable along x and y directions, respectively.

use of large CFL numbers for small Knudsen numbers. Finally, test 6 deals with a two dimensional problem with small Knudsen number. For the time step, we use the relation:

$$\Delta t = \text{CFL}\times \Delta x/|v_{\max}|. \tag{6.1}$$

For space and velocity grids, we discretize $\Delta v := (v_{\max}-v_{\min})/N_v$ and $\Delta x := (x_{\max}-x_{\min})/N_x$. To distinguish the proposed conservative schemes from the non-conservative SL schemes [17], we denote each scheme as follows:

Scheme name	Conservative	ODE solver	Reconstruction	Maxwellian
RK2-W23-DM	YES	DIRK2	WENO 2-3	Discrete
RK3-W35-DM	YES	DIRK3	WENO 3-5	Discrete
RK2-W23-CM	YES	DIRK2	WENO 2-3	Continuous
RK3-W35-CM	YES	DIRK3	WENO 3-5	Continuous
RK2-W23	NO	DIRK2	WENO 2-3	Continuous
RK3-W35	NO	DIRK3	WENO 3-5	Continuous

A similar notation is used for the schemes based on BDF time integrator.

In particular for small Knudsen number, RK based schemes work well compared to BDF based schemes.

6.1 Test 1. Check exact conservation

We consider the same single shock test adopted in Section 2.2, and apply the various schemes based on the conservative correction. The results are summarized in Tables 4 and 5.

When using with the continuous Maxwellian with the conservative correction, a negligible conservation error can be achieved, but only using a large enough number of velocity grid points. In contrast, conservation error can be suppressed to a negligible level with relatively small number of velocity grid points when the discrete Maxwellian is used (See Tables 4 and 5). We present the conservation error estimates in Appendix D.

It appears that the combined use of discrete Maxwellian and conservative correction provides a scheme which maintains conservation within round-off error. In particular, the use of discrete Maxwellian allows to maintain conservation with a small number of

Table 4: RK-based schemes, CFL=2. Conservation error of discrete moments in the relative L_1 norm for Test 1. Comparison between continuous Maxwellian (CM) and discrete Maxwellian (DM). We do not report the result for $(N_x,N_v)=(100,30)$ with RK3-W35-CM because solutions are destroyed due to the negativity of solutions.

	RK3-W35-CM			RK3-W35-DM		
(N_x,N_v)	Mass	Momentum	Energy	Mass	Momentum	Energy
(100,30)	·	·	·	1.28e-12	1.37e-12	1.52e-14
(100,40)	3.74e-07	1.51e-05	4.07e-06	5.47e-15	7.52e-14	8.59e-14
(100,50)	5.80e-12	3.58e-10	9.66e-11	5.55e-14	4.04e-13	5.22e-13
(100,60)	1.70e-13	1.45e-13	3.25e-13	1.92e-13	1.45e-13	3.43e-13
(100,90)	1.47e-13	8.85e-14	1.85e-13	1.30e-13	8.51e-14	2.27e-13
(200,40)	7.70e-07	4.37e-06	8.38e-06	1.28e-14	2.39e-15	7.45e-14
(200,50)	1.21e-11	1.03e-10	1.99e-10	2.13e-13	3.04e-13	2.59e-13
(200,60)	3.63e-13	1.74e-14	1.28e-13	3.73e-13	8.07e-14	1.79e-13
(200,90)	2.85e-13	1.47e-13	1.51e-13	2.83e-13	1.62e-13	1.75e-13

Table 5: BDF-based schemes, CFL$=2$. Conservation error of discrete moments in the relative L_1 norm for Test 1. Comparison between continuous one (CM) and discrete Maxwellian (DM). Here, we also do not report the result for $(N_x,N_v)=(100,30)$ with BDF3-W35-CM because solutions are destroyed due to the negativity of solutions.

	BDF3-W35-CM			BDF3-W35-DM		
(N_x,N_v)	Mass	Momentum	Energy	Mass	Momentum	Energy
(100,30)	·	·	·	4.96e-14	1.78e-13	2.74e-13
(100,40)	2.13e-07	6.30e-07	2.31e-06	3.55e-15	6.50e-15	3.40e-15
(100,50)	3.33e-12	1.50e-11	5.47e-11	6.08e-14	6.09e-14	3.43e-14
(100,60)	1.05e-13	1.95e-14	1.61e-14	1.06e-13	3.42e-15	6.49e-15
(100,90)	8.12e-14	1.16e-14	1.17e-14	8.39e-14	3.25e-14	1.54e-14
(200,40)	4.31e-07	2.19e-06	4.68e-06	1.49e-14	3.08e-14	3.80e-14
(200,50)	6.78e-12	5.22e-11	1.11e-10	1.60e-13	8.65e-14	1.51e-14
(200,60)	1.80e-13	4.75e-14	3.68e-14	1.82e-13	5.61e-14	4.48e-14
(200,90)	1.34e-13	8.03e-14	3.77e-14	1.36e-13	8.99e-14	4.42e-14

velocity nodes, which is particularly useful when adopting the method to capture the fluid dynamic limit for small Knudsen number.

6.2 Test 2. Accuracy

This test is proposed in [17] to check the accuracy of the scheme. The initial condition for the distribution function is the Maxwellian

$$f_0(x,v)=\frac{\rho_0}{\sqrt{2\pi T_0}}\exp\left(-\frac{|v-u_0(x)|^2}{2T_0}\right),$$

where initial velocity profile is given by

$$u_0(x)=0.1\exp\left(-(10x-1)^2\right)-2\exp\left(-(10x+3)^2\right).$$

Initial density and temperature are uniform, with constant value $\rho_0(x)=1$ and $T_0(x)=1$. We use periodic boundary condition in space. The computation is performed on $(x,v)\in[-1,1]\times[-10,10]$.

Since a shock appears at approximately $t=0.35$, we integrate up to time $t_f=0.32$ when the solution is still smooth even in the limit of vanishing Knudsen number. To check the convergence rate, we take $N_x=160$, 320, 640, 1280, 2560, and 5120 uniform grid points in x direction, and $N_v=20$ uniform grid points in v direction.

Concerning the conclusion of Section 4, here we set CFL$=2$ for RK based schemes, while CFL$=0.5$ for BDF based schemes in all range of κ.

The time step is computed using (6.1). The relative L^1 norm is used to check the accuracy. Here we expect the accuracy of the schemes to be between the order of accuracy of time discretization and spatial reconstruction.

Table 6: Test 2: Convergence rate for second and third order RK and BDF schemes. Final time $t_f = 0.32$.

Test 2 RK2-W23-DM Mass, CFL$=$2								
	$\kappa = 10^{-6}$		$\kappa = 10^{-4}$		$\kappa = 10^{-2}$		$\kappa = 10^{-0}$	
N_x	error	rate	error	rate	error	rate	error	rate
160-320	1.01e-03	1.74	9.80e-04	1.79	1.79e-04	2.17	7.13e-04	1.69
320-640	3.02e-04	1.78	2.84e-04	1.86	3.97e-05	2.06	2.21e-04	1.84
640-1280	8.83e-05	2.17	7.82e-05	2.29	9.54e-06	2.01	6.19e-05	2.12
1280-2560	1.96e-05	2.38	1.60e-05	2.37	2.37e-06	1.99	1.42e-05	2.53
2560-5120	3.76e-06		3.09e-06		5.95e-07		2.47e-06	

Test 2 BDF2-W23-DM Mass, CFL$=$0.5								
	$\kappa = 10^{-6}$		$\kappa = 10^{-4}$		$\kappa = 10^{-2}$		$\kappa = 10^{-0}$	
N_x	error	rate	error	rate	error	rate	error	rate
160-320	1.01e-03	1.74	9.80e-04	1.78	1.77e-04	2.17	7.09e-04	1.68
320-640	3.03e-04	1.77	2.85e-04	1.86	3.93e-05	2.05	2.21e-04	1.84
640-1280	8.86e-05	2.18	7.84e-05	2.29	9.46e-06	2.01	6.19e-05	2.13
1280-2560	1.96e-05	2.38	1.60e-05	2.37	2.35e-06	1.99	1.42e-05	2.53
2560-5120	3.75e-06		3.09e-06		5.93e-07		2.45e-06	

Test 2 RK3-W35-DM Mass, CFL$=$2								
	$\kappa = 10^{-6}$		$\kappa = 10^{-4}$		$\kappa = 10^{-2}$		$\kappa = 10^{-0}$	
N_x	error	rate	error	rate	error	rate	error	rate
160-320	5.74e-05	3.31	5.07e-05	3.47	2.28e-06	4.34	1.28e-05	4.74
320-640	5.77e-06	4.23	4.58e-06	4.39	1.12e-07	3.59	4.80e-07	4.88
640-1280	3.08e-07	4.61	2.19e-07	4.43	9.31e-09	3.09	1.63e-08	4.66
1280-2560	1.26e-08	4.28	1.02e-08	3.58	1.09e-09	2.98	6.43e-10	4.06
2560-5120	6.49e-10		8.50e-10		1.38e-10		3.84e-11	

Test 2 BDF3-W35-DM Mass, CFL$=$0.5								
	$\kappa = 10^{-6}$		$\kappa = 10^{-4}$		$\kappa = 10^{-2}$		$\kappa = 10^{-0}$	
N_x	error	rate	error	rate	error	rate	error	rate
160-320	5.59e-05	3.30	4.93e-05	3.47	1.93e-06	4.99	1.26e-05	4.81
320-640	5.69e-06	4.28	4.44e-06	4.47	6.07e-08	5.31	4.47e-07	5.18
640-1280	2.93e-07	4.77	2.01e-07	4.90	1.53e-09	4.61	1.24e-08	5.00
1280-2560	1.07e-08	4.96	6.72e-09	4.98	6.27e-11	2.38	3.85e-10	2.95
2560-5120	3.45e-10		2.13e-10		1.20e-11		5.00e-11	

In Table 6, the results show that the desired accuracy is obtained for each scheme. For small time step space error appears to be dominant, and this explains the order of accuracy higher than expected from the order of the RK or BDF schemes. Also, some order reductions are observed in intermediate regimes.

6.3 Test 3. Relaxation to a local Maxwellian

The aim of this test is to check relaxation to local equilibrium of the numerical solution obtained by the C-SL scheme, as the Knudsen number becomes smaller. To this purpose we take a similar test as in [7]. We check numerically that $\|f-\mathcal{M}\|_1=\mathcal{O}(\kappa)$ for different values of $\kappa=10^{-4},10^{-5},10^{-6},10^{-7}$.

We take the following non-equilibrium initial data

$$f_0(x,v)=\frac{1}{2}\frac{\rho_0(x)}{\sqrt{2\pi T_0(x)}}\left(\exp\left(-\frac{(v-u_0(x))^2}{2T_0(x)}\right)+\exp\left(-\frac{(v+u_0(x))^2}{2T_0(x)}\right)\right),$$

where initial density, velocity and temperature are given by

$$\rho_0(x)=\frac{2+\sin 2\pi x}{3},\quad u_0(x)=\frac{\cos 2\pi x}{5},\quad T_0(x)=\frac{3+\cos 2\pi x}{4}.$$

We use periodic boundary condition. The computation is performed on $x\in[-1,1]$, $v\in[-8,8]$. The final time is taken 0.02. We implemented RK3-W35-CM and RK3-W35-DM with $N_x=100$ and CFL$=1$.

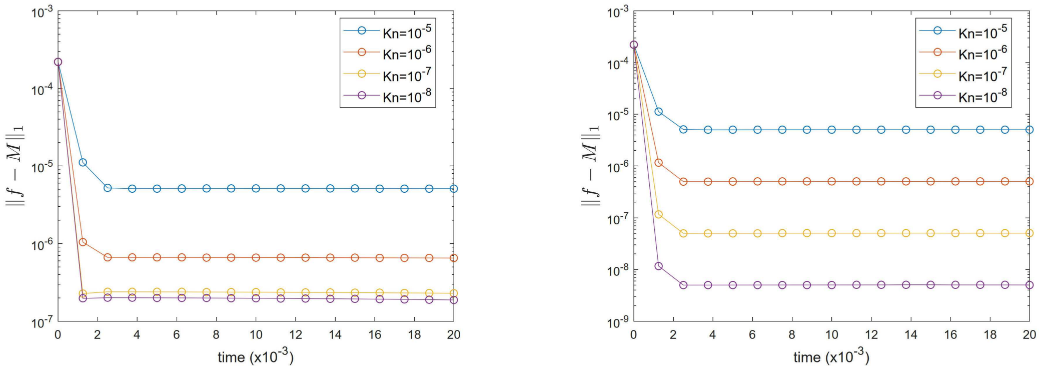
In Figs. 9-10, we show the time evolution of $\|f-\mathcal{M}\|_1$ for our C-SL scheme for different values of κ and different values for the number of grid points in velocity space, i.e., $N_v=20,32$.

From the figures it appears that the norm of the difference between f and the Maxwellian is roughly proportional to the Knudsen number κ, as expected. However, if we use RK3-W35-CM with a continuous Maxwellian, such a norm depends also on the number of velocity grid points, as appears in Fig. 9(a), where with $N_v=20$ the difference does not decrease significantly when going from $\kappa=10^{-7}$ to $\kappa=10^{-8}$. On the contrary, when using RK3-W35-DM with a discrete Maxwellian, the discrepancy between f and the Maxwellian only depends on the Knudsen number: $\|f-\mathcal{M}\|_1=\mathcal{O}(\kappa)$. We obtain similar results for BDF based methods, and we omit to report them in the paper.

The proposed C-SL scheme (3.1) is an asymptotic preserving (AP) scheme for the kinetic equation (1.1), that is, it becomes a consistent scheme for the underlying hydrodynamic limit. Note that in a recent review on AP schemes for kinetic and hyperbolic equations [21], a necessary condition to be AP for a scheme for BGK model (1.1) is that the solution f^n must be driven to the local equilibrium $\mathcal{M}^n$ when $\kappa\to 0$

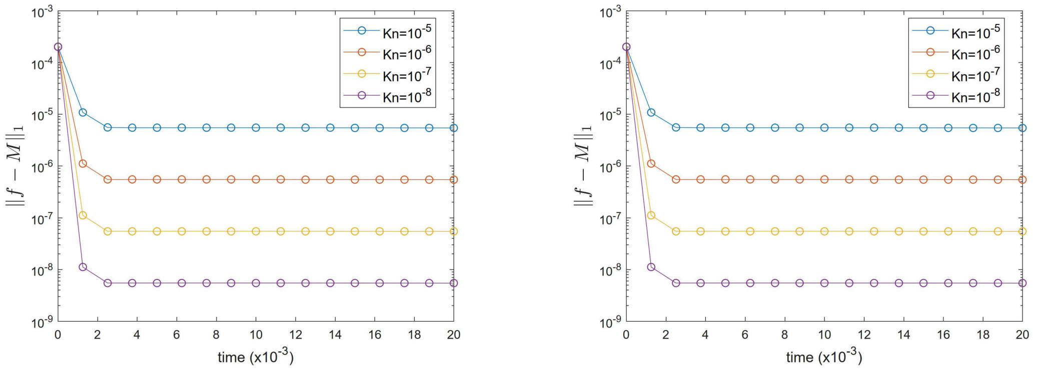
$$f^n-\mathcal{M}(f^n)=\mathcal{O}(\kappa),\quad \text{for } n\geq 1$$

for any initial data f^0, namely, the numerical solution projects any data into the local equilibrium $\mathcal{M}^n$, with a discrepancy of $\mathcal{O}(\kappa)$, in one step. Such AP schemes are referred to as *strongly* AP. In these tests we show that as the Knudsen number vanishes, the distribution function approaches a local Maxwellian, and the numerical solution of the moments agree with the solution of the Euler equation. A detailed consistency analysis of both conservative and non-conservative schemes, for smooth solutions, is presented in the following proposition (see Appendix E for its proof):



(a) RK3-W35-CM, $\|f-CM\|_1$ (b) RK3-W35-DM, $\|f-d\mathcal{M}\|_1$

Figure 9: Time evolution of $\|f-M\|_1$ for high order methods for the BGK model. $N_v=20$. When using the continuous Maxwellian (left panel), the discrepancy between the distribution function and the Maxwellian saturates for small values of the Knudsen number, while the method based on the discrete Maxwellian (right panel) shows the expected behaviour $\|f-M\|_1=\mathcal{O}(\kappa)$.



(a) RK3-W35-CM, $\|f-CM\|_1$ (b) RK3-W35-DM, $\|f-d\mathcal{M}\|_1$

Figure 10: Time evolution of $\|f-M\|_1$ for high order methods for the BGK model. $N_v=32$. For large enough number of velocity grid points the discrepancy between the function and the Maxwellian appears to be proportional to κ, up to $\kappa=10^{-8}$, for both schemes.

Proposition 6.1. Let $f^n(x,v)$ be a solution with time discretization. Assume there exist positive constants C_1,C_2 independent of n such that

$$\sum_{0\le\alpha\le2}\sup_{x,v}|\partial_x^\alpha f^n(x,v)|(1+|v|^6)<C_1, \tag{6.2}$$

and

$$\int_{\mathbb{R}} f^n(x,v)dv>C_2>0. \tag{6.3}$$

Then, the first order SL scheme (2.3) and C-SL scheme (3.1) are first order approximations in time of the compressible Euler system.

6.4 Test 4. Shock capturing capability

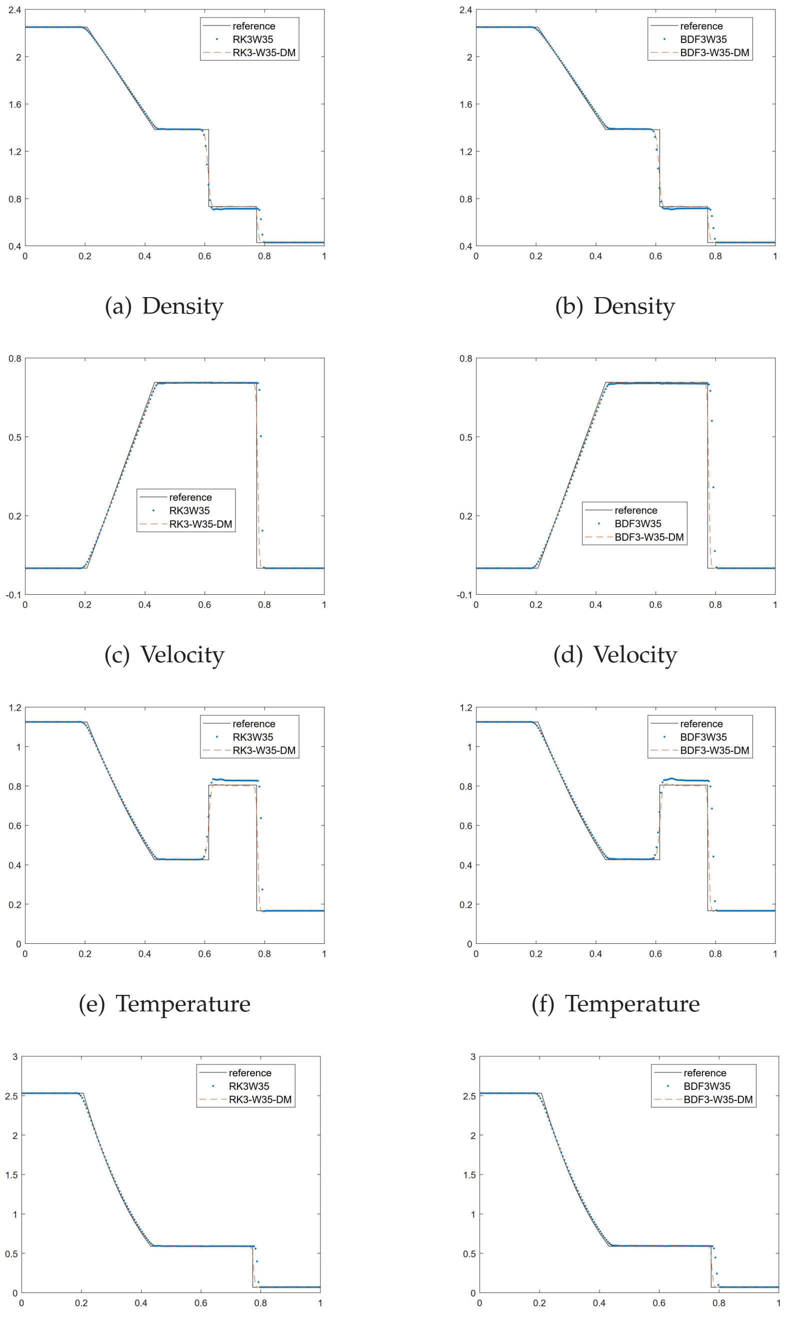
The aim of the test is to check the shock capturing capability of the schemes, and to compare conservative and non-conservative schemes in presence of shocks, when very small (under-resolved) Knudsen number are adopted. To this purpose we consider some classical Riemann problem. To observe the Euler limit, we take $\kappa = 10^{-6}$. Moreover, to see the influence of the conservative correction, we compare our scheme with the standard non-conservative semi-Lagrangian scheme. Initial condition is given by the Maxwellian computed from

$$(\rho_0, u_0, p_0) = \begin{cases} (2.25, 0, 1.125), & \text{for } x \leq 0.5, \\ (3/7, 0, 1/6), & \text{for } x > 0.5. \end{cases}$$

We use free-flow boundary condition. Computations are performed on $x \in [0,1]$, $v \in [-10,10]$ up to final time $t_f = 0.16$. With small Knudsen number, both conservative and non-conservative schemes are stable and they enable us to use CFL$=2$. We take $N_x = 200$ for both schemes. For the proposed schemes, RK3-W35-DM and BDF3-W35-DM, we take $N_v = 30$. For non-conservative schemes, RK3W35 and BDF3W35, we take $N_v = 60$. The larger number of velocity grid points ensures that the conservation error due to the use of continuous Maxwellian is negligible with respect to the one due to lack of conservation of the transport term. The results are shown in Fig. 11. It appears that conservative schemes capture shocks correctly and are in perfect agreement with the exact reference solution.

6.5 Test 5. Long time behaviour

Here we provide a numerical evidence to support the stabilizing effect of the collision operator on numerical schemes for the solution of the BGK equations. Several schemes are used and compared, namely: three semi-Lagrangian Runge-Kutta schemes, RK3W35, RK3-W35-CM and RK3-W35-DM, and an Eulerian Runge-Kutta scheme IMEX3-W35-CM taken from [24], in which the convective terms are treated explicitly in Eulerian framework, while the collision term is treated implicitly, through the use of an IMEX-Runge-Kutta scheme of order 3. We verify that both our RK based semi-Lagrangian conservative schemes and the IMEX based Eulerian scheme remain stable for CFL numbers much larger than 1 when the Knudsen number is sufficiently small, even in presence of shocks. We consider the same initial data as in Test 2 with a small Knudsen number $\kappa = 10^{-6}$ (see Fig. 12). We compare macroscopic variables obtained by the various methods with a reference solution of the Euler equations computed by a WENO3 shock capturing scheme with RK4 time integrator. The final time is $t_f = 10$. For the first four schemes, we take $N_x = 200$, and two values of CFL numbers, CFL$=1$ (left column) and CFL$=5$ (right column). The reference solution is computed with $N_x = 5000$ grid points in space, CFL$=0.2$, and WENO3. Also, to emphasize the efficiency of the proposed conservative method, we



(a) Density

(b) Density

(c) Velocity

(d) Velocity

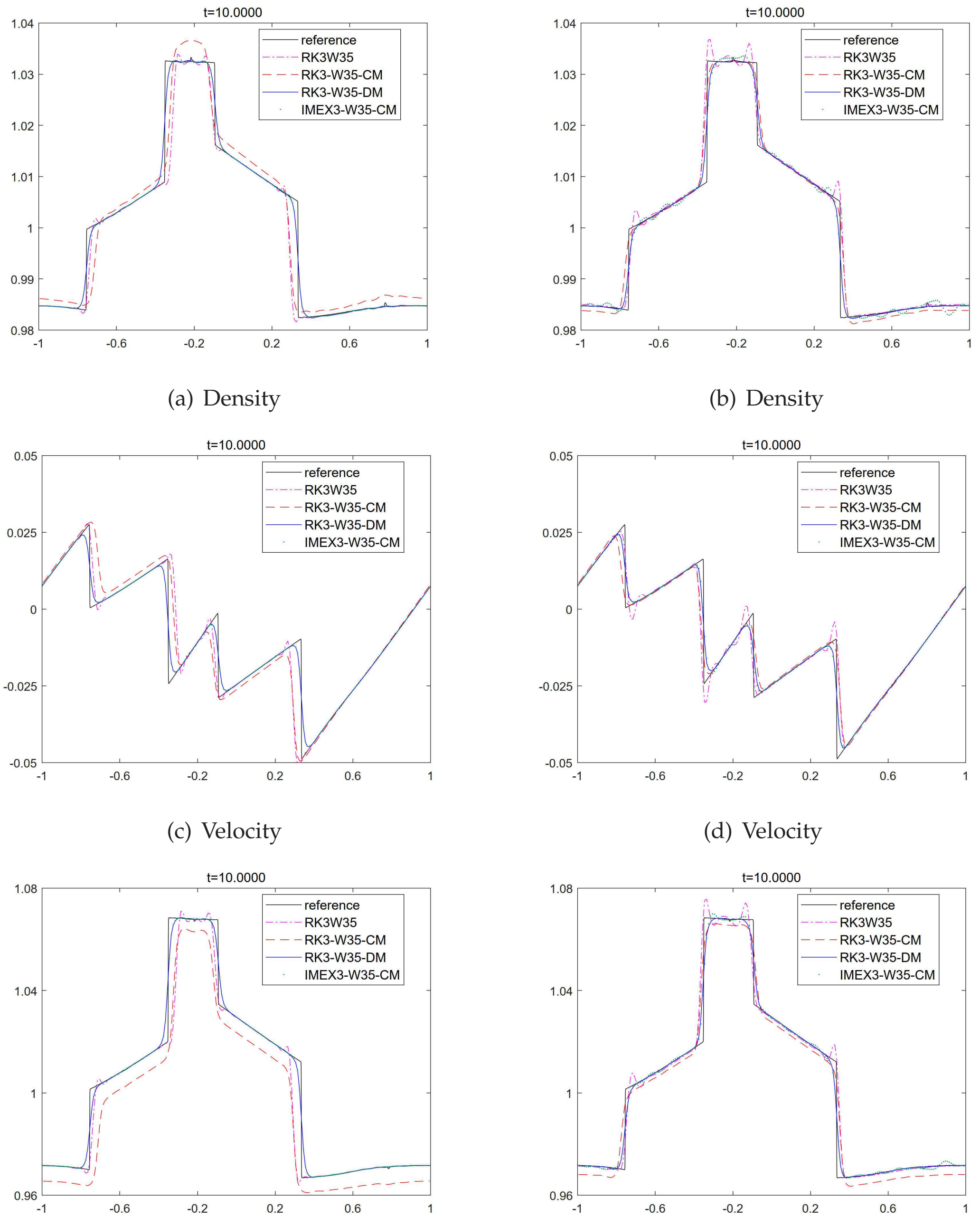
(e) Temperature

(f) Temperature

(g) Pressure

(h) Pressure

Figure 11: Test 4. Riemann problem in 1D space and velocity with $\kappa = 10^{-6}$. RK based schemes (left column) and BDF based schemes (right column). From top to bottom: Density, Velocity, Temperature and Pressure. Blue point: standard SL schemes, red dashed line: new conservative schemes, black solid line: exact solution.



(a) Density

(b) Density

(c) Velocity

(d) Velocity

(e) Temperature

(f) Temperature

Figure 12: Test 5. Long time shock test in 1D space and velocity with $\kappa = 10^{-6}$. Black solid lines: reference solution, magenta dash-dotted lines: RK3W35, red dash-dashed lines: RK3-W35-CM, blue solid lines: RK3-W35-DM, green dots: IMEX3-W35-CM. We used CFL$=1$ for left column, and CFL$=5$ for right one.

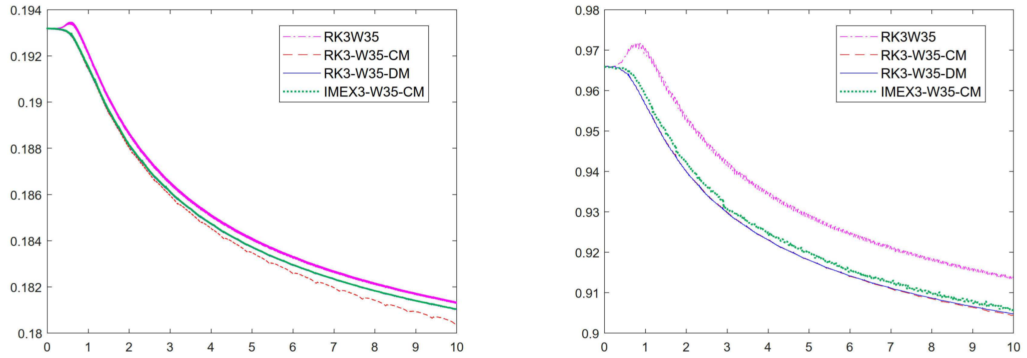


Figure 13: Test 5. The value of fluid dynamic CFL number:$=\max_i\left(|u_i|+\sqrt{3T_i}\right)\frac{\Delta t}{\Delta x}$ up to final time $t_f=10$. Magenta dash-dotted lines: RK3W35, red dash-dashed lines: RK3-W35-CM, blue solid lines: RK3-W35-DM, green dots: IMEX3-W35-CM. We used CFL$=1$ for left column, and CFL$=5$ for right one.

take different number of velocity grid points: $N_v=60$ for RK3W35 and IMEX3-W35, and $N_v=20$ for RK3-W35-CM and RK3-W35-DM methods.

In Fig. 12, one can observe that both non-conservative schemes and RK3-W35-CM give inaccurate solutions regardless of the CFL numbers, because of lack of conservation. In case of RK3W35, although it is computed with enough velocity grids, severe oscillations are observed near discontinuity, and the shock positions are located inaccurately. The RK3-W35-CM scheme gives non-oscillatory solutions, however, the profile of moments are far from the reference solution. On the contrary, RK3-W35-DM scheme gives non-oscillatory solutions and capture shock position exactly even with $N_v=20$. These observations imply that oscillations in the classical schemes come from non-conservative reconstructions such as GWENO, and conservation is an essential property for solutions to be consistent with Euler's equations in presence of shocks. The Eulerian based scheme (dots) is able to correctly capture shocks, and is stable even for large CFL numbers, however in this case its solution becomes oscillatory. We also note that the fluid dynamic CFL number, defined as

$$CFL_F\equiv\max_i\left(|u_i|+\sqrt{3T_i}\right)\frac{\Delta t}{\Delta x},$$

does not exceed 1 for whole simulation time. (See Fig. 13.)

6.6 Test 6. Mixed regime

In this section, we consider a mixed regime 1d space and 2d velocity problem in [7]. We set initial data by

$$f_0(x,v_1,v_2)=\frac{\rho_0(x)}{4\pi T_0(x)}\left(e^{-\frac{|v-u_0(x)|^2}{2T_0(x)}}+e^{-\frac{|v+u_0(x)|^2}{2T_0(x)}}\right),$$

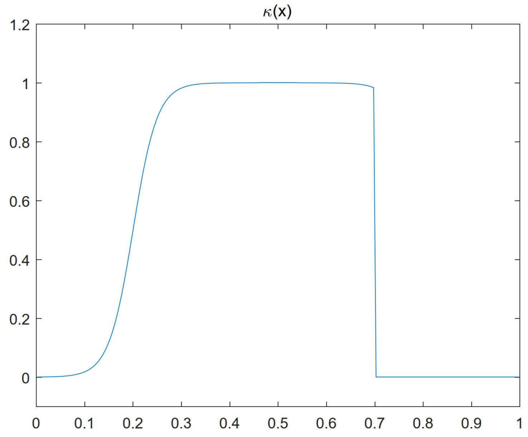


Figure 14: The values of $\kappa(x)$ given in (6.4).

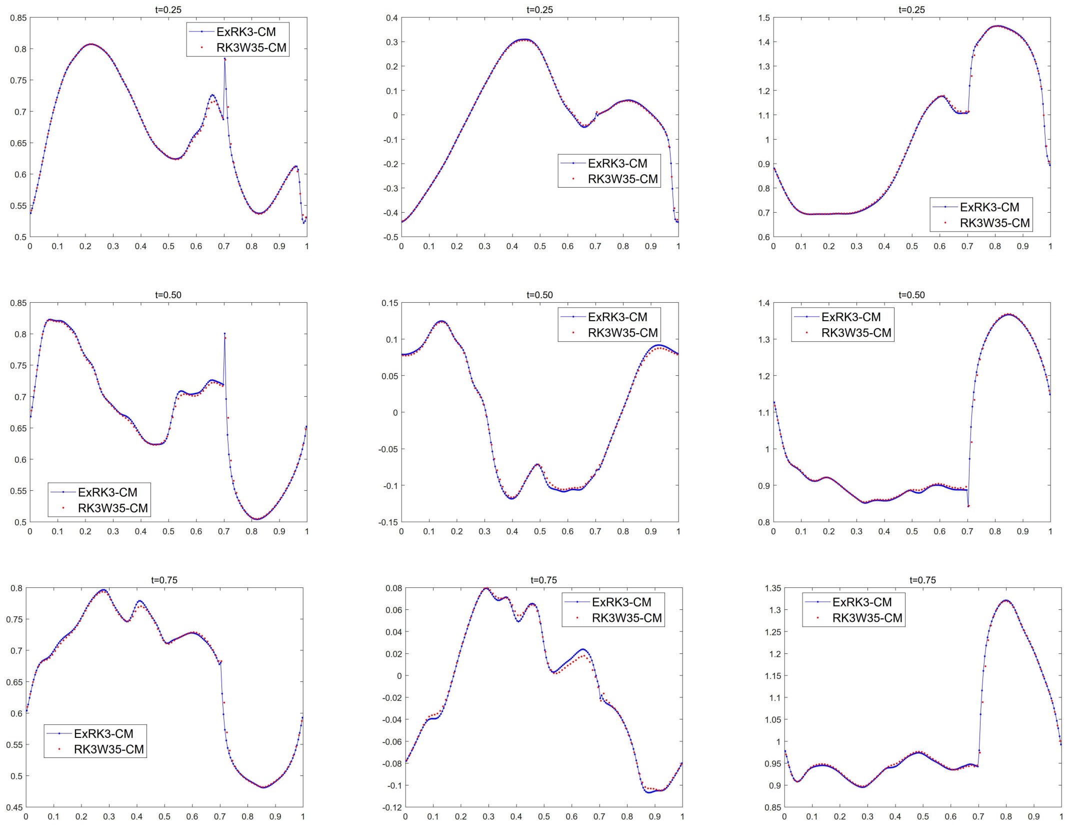
where

$$\rho_0(x)=\frac{2+\sin(2\pi x)}{3},\quad u_0(x)=\begin{pmatrix}\cos(2\pi x)\\ 0\end{pmatrix},\quad T_0(x)=\frac{3+\cos(2\pi x)}{4}.$$

The Knudsen number is given by

$$\kappa(x)=\begin{cases}\kappa_0+0.5(\tanh(16-20x)+\tanh(-4+20x)), & x<0.7,\\ \kappa_0, & x>0.7,\end{cases}\tag{6.4}$$

with $\kappa_0=5\times10^{-4}$. The values of $\kappa(x)$ are plotted in Fig. 14. Therefore, in this problem we are treating Knudsen numbers which range from 5×10^{-4} to 1. For physical space, we consider periodic boundary conditions, while velocity domain is truncated by $[-8,8]$. For comparison, we compute a reference solution using an second order explicit RK method in the Eulerian framework. Due to the stability restriction of the explicit RK scheme, we use a time step $\Delta t=1.25\times10^{-4}$ corresponding to CFL$=0.2$ for $N_x=200$. For numerical solutions, we implement the RK2-W23-CM method, and take a relatively large CFL$=2$ and less grid points $N_x=100$. For both methods, we take the same number of velocity grids $N_v=32^2$, which is enough to guarantee good conservation properties even with the use of a continuous Maxwellian. Fig. 15 shows that RK3-W35-CM scheme gives results similar to the reference solutions even with coarse grids and larger time step. Since our result is obtained by solving only BGK model, it is different from the literature [7]. In order to obtain more accurate solutions for this problem, one may apply our technique to another approximation model of the Boltzmann equation such as ES-BGK model [20].



(a) Density (b) Velocity (c) Temperature

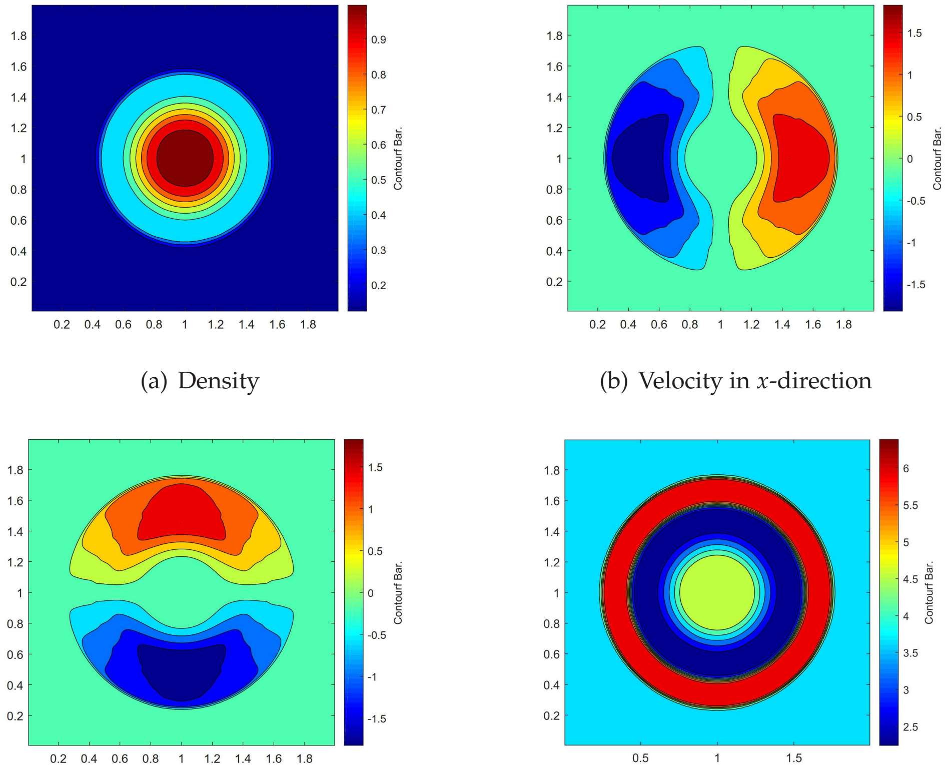
Figure 15: Test 6. Macroscopic variables at different times. From top to bottom, we plot results on the space domain $[0,1]$ at time $t=0.25$, $t=0.5$ and $t=0.75$.

6.7 Test 7. 2D shock problem

Here we test a 2D shock problem in [15]. The initial condition is the local Maxwellian with parameters

$$(\rho_0,u_{10},u_{20},T_0)=\begin{cases} (1,0,0,5), & \text{for } (x-1)^2+(y-1)^2\le(0.2)^2,\\ (0.125,0,0,4), & \text{for } (x-1)^2+(y-1)^2>(0.2)^2, \end{cases}$$

where x,y denotes the Cartesian coordinates. Here we compare two-dimensional versions of RK2W23 and RK2-W23-DM. We use $\kappa=10^{-4}$ and perform the computation on $(x,v)\in[0,2]^2\times[-15,15]^2$ upto final time $t_f=0.07$. Here we take a uniform mesh with 200×200 grid points in physical space. Time steps are taken to satisfy CFL$=v_{\max}\frac{\Delta t}{\Delta x}=v_{\max}\frac{\Delta t}{\Delta y}=1$ and CFL$=4$ where $v_{\max}=15$. In the velocity discretizations, we use $(60+1)\times(60+1)$ points for RK2W23, and $(30+1)\times(30+1)$ points for RK2-W23-DM.



(a) Density

(b) Velocity in x-direction

(c) Velocity in y-direction

(d) Temperature

Figure 16: Test 6. Shock test in 2D space and velocity. Macroscopic quantities obtained by RK2-W23-DM with CFL=4.

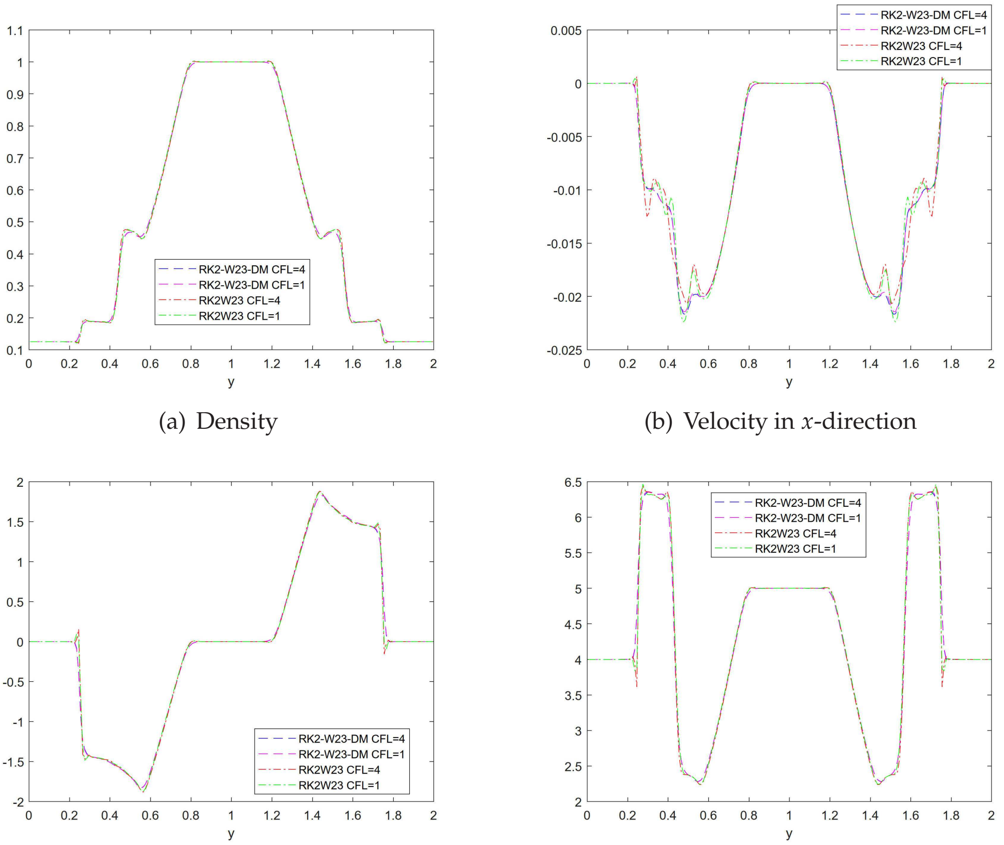
In Figs. 16 and 17, our conservative scheme shows good agreement with the result in [15] even for CFL=4. Furthermore, our solution shows much smaller oscillations near shocks compared to the one obtained by scheme RK2W23.

7 Conclusions

In this paper we present high-order conservative semi-Lagrangian schemes for the numerical solution of the BGK model of the Boltzmann equation.

Lack of conservation in standard semi-Lagrangian schemes is shown to be due to nonlinear weights in the non-oscillatory reconstruction of the distribution function at the foot of the characteristic, and to the use of a continuous Maxwellian for the computation of the moments.

Conservation properties are restored at a discrete level by making use of a discrete Maxwellian in the collision operator, and by a conservative correction of the advection



(a) Density

(b) Velocity in x-direction

(c) Velocity in y-direction

(d) Temperature

Figure 17: Test 6. Macroscopic quantities obtained by RK2W23 and RK2-W23-DM at $x=x_{100}$. Blue and magenta dashed lines: new conservative schemes, red and green dash-dotted lines: standard SL schemes.

term. Exact conservation can be reached up to round-off errors. Together with L-stable treatment of the collisions, exact conservation allows the construction of schemes which become consistent *shock-capturing* schemes for the underlying Euler limit, as the Knudsen number κ vanishes (AP property), even when using a relatively small number of grid points in velocity.

High order accuracy in space is obtained by a modification of the classical WENO reconstruction, while high order accuracy in time is obtained by either Runge-Kutta of multistep time discretization.

The method is described and implemented both in the one and two dimensions, both in space and velocity.

The conservation properties, and the consequent AP property, have been proven mathematically and verified in several numerical tests. A drawback of the conservative correction procedure is the limitation it imposes on the stability of the schemes. A stability analysis has been performed to understand the reason of such limitation. It is observed that Runge-Kutta based schemes have a wider stability region than multistep-bases ones, with a net improvement over Eulerian based schemes for all Knudsen numbers. Stability restrictions become less severe for small Knudsen numbers, as supported by the stability analysis at the end of Section 4, making the schemes competitive in such regimes.

As a work in progress, we are developing a new conservative semi-Lagrangian scheme that does not suffer from such a CFL limitation, which will be the subject of a forthcoming paper.

Acknowledgments

G. Russo is partially supported by ITN-ETN Horizon 2020 Project ModCompShock, Modeling and Computation on Shocks and Interfaces, Project Reference 642768. S. Boscarino is supported by the University of Catania "Piano triennale della Ricerca 2016/2018, Linea di intervento 2". S.-B. Yun was supported by Samsung Science and Technology Foundation under Project Number SSTF-BA1801-02.

Appendices

A Proof of the estimate on the conservation error for IE-SL scheme with the discrete Maxwellian

Let us check in what sense the scheme (2.5) has a better conservative nature compared to that of (2.3). For this, we first rewrite (2.5) as

$$f_{i,j}^{n+1}-\tilde{f}_{ij}^{n}=\frac{\Delta t}{\kappa+\Delta t}\left(d\mathcal{M}(\tilde{f}_{ij}^{n})-\tilde{f}_{ij}^{n}\right), \tag{A.1}$$

where $\tilde{f}_{ij}^{n}=\theta_j f_{i^*+1,j}^{n}+(1-\theta_j)f_{i^*,j}^{n}$ with $i^*=\lfloor i-v_j\Delta t/\Delta x\rfloor$, $\theta_j=(\tilde{x}_{ij}-x_{i^*})/\Delta x$ and $\tilde{x}_{ij}=x_i-v_j\Delta t$.

Since θ_j does not depend on i, we find that the following telescoping cancellation holds so that $f_{i,j}^{n}$ and $\tilde{f}_{ij}^{n}$ share the first moment:

$$\sum_{i=1}^{N_x}\tilde{f}_{ij}^{n}=\sum_{i=1}^{N_x}\left(\theta_j f_{i^*+1,j}^{n}+(1-\theta_j)f_{i^*,j}^{n}\right)=\sum_{i=1}^{N_x}f_{i,j}^{n}. \tag{A.2}$$

Multiplying (A.1) by $\phi(v_j)=(1;v_j;v_j^2/2)$, taking summation on i,j and inserting (A.2), one gets

$$\sum_{i=1}^{N_x}\sum_{j=0}^{N_v}\left(f_{i,j}^{n+1}-f_{i,j}^{n}\right)\phi(v_j)\Delta v\Delta x=\frac{\Delta t}{\kappa+\Delta t}\sum_{i=1}^{N_x}\sum_{j=0}^{N_v}\left(d\mathcal{M}(\tilde{f}_{ij}^{n})-\tilde{f}_{ij}^{n}\right)\phi(v_j)\Delta v\Delta x.$$

Summing further in time step, we have

$$
\begin{aligned}
&\sum_{i=1}^{N_x}\sum_{j=0}^{N_v}\left(f_{i,j}^{Nt}-f_{i,j}^{0}\right)\phi(v_j)\Delta v\Delta x\\
=&\frac{\Delta t}{\kappa+\Delta t}\sum_{k=0}^{N_t-1}\sum_{i=1}^{N_x}\sum_{j=0}^{N_v}\left(d\mathcal{M}(\tilde{f}_{ij}^{k})-\tilde{f}_{ij}^{k}\right)\phi(v_j)\Delta v\Delta x,
\end{aligned}
$$

then, denoting the ℓth component of $\phi(v_j)$ by $\phi_\ell(v_j)$, ℓ=1,2,3, and using a variant of (2.4):

$$
\max_{1\le\ell\le3}\left|\sum_{j=0}^{N_v}\left[d\mathcal{M}(\tilde{f}_{ij}^{k})-\tilde{f}_{ij}^{k}\right]\phi_\ell(v_j)\Delta v\right|<tol,
$$

we obtain the following estimate:

$$
\begin{aligned}
&\max_{1\le\ell\le3}\left|\sum_{i=1}^{N_x}\sum_{j=0}^{N_v}\left(f_{i,j}^{Nt}-f_{i,j}^{0}\right)\phi_\ell(v_j)\Delta v\Delta x\right|\\
\le&\frac{\Delta t}{\kappa+\Delta t}\sum_{k=0}^{N_t-1}\sum_{i=1}^{N_x}\max_{1\le\ell\le3}\left|\sum_{j=0}^{N_v}\left(d\mathcal{M}(\tilde{f}_{ij}^{k})-\tilde{f}_{ij}^{k}\right)\phi_\ell(v_j)\Delta v\right|\Delta x\\
\le&\frac{N_t\Delta t}{\kappa+\Delta t}(x_{\max}-x_{\min})tol.
\end{aligned}
$$

This estimate tells us that error is stacked in each time step by tol. So, for small values of κ, the total conservation error in the end essentially depends on $N_t\times tol$ uniformly in κ. Therefore, tol should be taken small enough to attain a machine precision conservation error.

B General framework of G-WENO interpolation

In this section, we illustrate the G-WENO interpolation of degree $2n-1$. Let $U=\{u_j\},j\in I$ be a set of given values of a function u on a space grid x_j, $j\in I$.

We start with the Lagrange polynomial $Q(x)$ built on the stencil $S=\{x_{j-n+1},\cdots,x_{j+n}\}$:

$$
Q(x)=\sum_{k=1}^{n}C_k(x)P_k(x),
$$

where the "linear weights" $C_k(x)$ are polynomials of degree $n-1$ and P_k are polynomials of degree n interpolating U on the stencil $S_k=\{x_{j-n+k},\cdots,x_{j+k}\}$, $k=1,\cdots,n$. The linear weights $C_k(x)$ $(k=1,\cdots,n)$ are determined to satisfy the following two properties [8]:

1. $C_k(x_i)=0$ for $x_i\in S-S_k$.

2. $\sum_k C_k(x_i) = 1$ for $x_i \in S$.

To guarantee non-oscillatory property, we replace the linear weights $C_k(x)$ by the non-linear weights $\omega_k(x)$:

$$\omega_k(x) = \frac{\alpha_k(x)}{\sum_l \alpha^l(x)},$$

where $\alpha_k(x)$ is defined by

$$\alpha_k(x) = \frac{C_k(x)}{(\beta_k + \epsilon)^2}, \tag{B.1}$$

with the choice of $\epsilon = 10^{-6}$. The smoothness indicators β_k in (B.1) is defined by

$$\beta_k = \sum_{l=1}^{n} \int_{x_j}^{x_{j+1}} \Delta x^{2l-1} (P_k^{(l)})^2 dx.$$

The nonlinear weights $\omega_k(x)$ are designed to put more weights on the smooth part of u and less weights on the discontinuous part of u.

Finally, the G-WENO reconstruction of the values $U = \{u_j\}_{j \in I}$ reads

$$I[U](x) = \sum_{k=1}^{n} \omega_k(x) P_k(x).$$

In the following, we explicitly construct the G-WENO interpolations of degree 3 and 5.

B.1 G-WENO of degree 3 (WENO23)

The G-WENO interpolation of degree 3 can be represented with two degree 2 polynomials P_L and P_R built respectively on stencils $\{x_{j-1}, x_j, x_{j+1}\}$ and $\{x_j, x_{j+1}, x_{j+2}\}$:

$$P(x) = \omega_L P_L(x) + \omega_R P_R(x),$$

where the non-linear weights ω_L and ω_R are given by

$$\omega_\ell = \frac{\alpha_\ell}{\sum_\ell \alpha_\ell}, \quad \alpha_\ell = \frac{C_\ell}{(\epsilon + \beta_\ell)^2}, \quad \ell = L, R,$$

with

$$C_L = \frac{x_{j+2} - x}{3\Delta x}, \quad C_R = \frac{x - x_{j-1}}{3\Delta x},$$

and

$$\begin{aligned}
\beta_L &= \frac{13}{12} u_{j-1}^2 + \frac{16}{3} u_j^2 + \frac{25}{12} u_{j+1}^2 - \frac{13}{3} u_{j-1} u_j + \frac{13}{6} u_{j-1} u_{j+1} - \frac{19}{3} u_j u_{j+1}, \\
\beta_R &= \frac{13}{12} u_j^2 + \frac{16}{3} u_{j+1}^2 + \frac{25}{12} u_{j+2}^2 - \frac{13}{3} u_j u_{j+1} + \frac{13}{6} u_j u_{j+2} - \frac{19}{3} u_{j+1} u_{j+2}.
\end{aligned}$$

B.2 G-WENO of degree 5 (WENO35)

For the G-WENO interpolation of degree 5, we use degree 3 polynomials P_L, P_C and P_R built respectively on stencils $\{x_{j-2},x_{j-1},x_j,x_{j+1}\}$, $\{x_{j-1},x_j,x_{j+1},x_{j+2}\}$ and $\{x_j,x_{j+1},x_{j+2},x_{j+3}\}$:

$$P(x)=\omega_L P_L(x)+\omega_C P_C(x)+\omega_R P_R(x),$$

where the non-linear weights ω_L, ω_C and ω_R are given by

$$\omega_\ell=\frac{\alpha_\ell}{\sum_\ell \alpha_\ell},\quad \alpha_\ell=\frac{C_\ell}{(\epsilon+\beta_\ell)^2},\quad \ell=L,C,R,$$

with

$$C_L=\frac{(x-x_{j+2})(x-x_{j+3})}{20\Delta x^2},\quad C_C=-\frac{(x-x_{j-2})(x-x_{j+3})}{10\Delta x^2},\quad C_R=\frac{(x-x_{j-2})(x-x_{j-1})}{20\Delta x^2},$$

and

$$\begin{aligned}
\beta_L=&\frac{407}{90}u_{j+1}^2+\frac{721}{30}u_j^2+\frac{248}{15}u_{j-1}^2+\frac{61}{45}u_{j-2}^2-\frac{1193}{60}u_{j+1}u_{j-2}+\frac{439}{30}u_{j+1}u_{j-1}\\
&-\frac{683}{180}u_{j+1}u_{j-2}-\frac{2309}{60}u_ju_{j-1}+\frac{309}{30}u_ju_{j-2}-\frac{553}{60}u_{j-1}u_{j-2},\\
\beta_C=&\frac{61}{45}u_{j-1}^2+\frac{331}{30}u_j^2+\frac{331}{30}u_{j+1}^2+\frac{61}{45}u_{j+2}^2-\frac{141}{20}u_{j-1}u_j+\frac{179}{30}u_{j-1}u_{j+1}\\
&-\frac{293}{180}u_{j-1}u_{j+2}-\frac{1259}{60}u_ju_{j+1}+\frac{179}{30}u_ju_{j+2}-\frac{141}{20}u_{j+1}u_{j+2},\\
\beta_R=&\frac{407}{90}u_j^2+\frac{721}{30}u_{j+1}^2+\frac{248}{15}u_{j+2}^2+\frac{61}{45}u_{j+3}^2-\frac{1193}{60}u_ju_{j+3}+\frac{439}{30}u_ju_{j+2}\\
&-\frac{683}{180}u_ju_{j+3}-\frac{2309}{60}u_{j+1}u_{j+2}+\frac{309}{30}u_{j+1}u_{j+3}-\frac{553}{60}u_{j+2}u_{j+3}.
\end{aligned}$$

C Details on the stability analysis

Here we find conditions such that Eq. (4.11) is satisfied. In order to make clear the calculation in the method (4.8) we take $c_1=\gamma_1$ and $b_3=\gamma_1$ then we get the stability function (see [18]). Note that it is sufficient to have all $E_{2j}\geq 0$ for the I-stability. These are the conditions that we actually use, in order to simplify the analysis.

We consider

$$R(z)=\frac{P(z)}{Q(z)}=\frac{p_0+p_1z+p_2z^2+p_3z^3}{q_0-q_1z+q_2z^2-q_3z^3} \tag{C.1}$$

with the following quantities:

$$p_0=1,\quad p_1=\left(\frac{q_0}{1!}-\frac{q_1}{0!}\right),\quad p_2=\left(\frac{q_0}{2!}-\frac{q_1}{1!}+\frac{q_2}{0!}\right),\quad p_3=\left(\frac{q_0}{3!}-\frac{q_1}{2!}+\frac{q_2}{1!}-\frac{q_3}{0!}\right) \tag{C.2}$$

and

$$q_0=1,\quad q_1=\gamma_1+\gamma_2+\gamma_3,\quad q_2=\gamma_1\gamma_2+\gamma_1\gamma_3+\gamma_2\gamma_3,\quad q_3=\gamma_1\gamma_2\gamma_3, \tag{C.3}$$

and from (4.11) we have

$$\begin{aligned}E_2&=(q_1^2-p_1^2)-2(q_2q_0-p_2p_0)\geq 0,\\ E_4&=(q_2^2-p_2^2)-2(q_3q_1-p_3p_1)\geq 0,\\ E_6&=q_3^2-p_3^2\geq 0.\end{aligned}$$

By (C.3) it follows $E_2=0$, and by SA we have $R(\infty)=0$, and from (C.1) we get $p_3=0$ and then $E_6=q_3^2\geq 0$. Now we compute E_4, and by (C.2)-(C.3) we get $E_4=(q_2^2-p_2^2)-2q_3q_1$. From $p_3=0$, it follows

$$q_3=\frac{1}{6}-\frac{q_1}{2}+q_2$$

and substituting this in E_4 we obtain $E_4=8q_1-12q_2-3\geq 0$.

Now if we substitute the quantities (C.3) in E_4, and using (4.10) we get a function that depends on γ_1 and c_2

$$-\frac{S}{(3\gamma_1-1)(\gamma_1-1)^2(c_2-1)}\geq 0,$$

with

$$\begin{aligned}S=&(108c_2^2-72c_2+18)\gamma_1^4+(-144c_2^2+105c_2-33)\gamma_1^3\\ &+(84c_2^2-69c_2+24)\gamma_1^2+(-24c_2^2+21c_2-7)\gamma_1+3c_2^2-3c_2+1.\end{aligned}$$

The function S is always positive for $\gamma_1=c_2\geq 0$, then we have that $E_4\geq 0$, if

$$\gamma_1\leq 1/3,\quad c_2\geq 1,\quad \text{or}\quad \gamma_1\geq 1/3,\quad c_2\leq 1.$$

Now in order to justify the requirement to choose γ in the intervals (4.14) we consider again $E_4=8q_1-12q_2-3\geq 0$.

In (4.13) with $\gamma_1=\gamma_3=\gamma$, we compute from (4.10) b_2, c_2 and γ_2 as functions of γ:

$$b_2=-\frac{3}{4}\frac{(2\gamma^2-4\gamma+1)^2}{3\gamma^3-9\gamma^2+6\gamma-1},\quad c2=\frac{1}{3}\frac{(6\gamma^2-9\gamma+2}{(2\gamma^2-4\gamma+1)},\quad \gamma_2=\frac{1}{3}\frac{(6\gamma^2-6\gamma+1}{(2\gamma^2-4\gamma+1)} \tag{C.4}$$

and $b_1=1-b_2-\gamma$. Furthermore it follows: $q_1=(2\gamma+\gamma_2)$ and $q_2=(2\gamma_2\gamma+\gamma^2)$ and substituting this values in E_4 we get

$$\gamma_2\geq\frac{3-16\gamma+12\gamma^2}{8-24\gamma}. \tag{C.5}$$

Now substituting γ_2 from (C.4) in (C.5) and solving this inequality for γ we get (4.14).

D Conservation error estimates for discrete moments

In this section, we carry out some elementary conservation error estimate for each of the schemes derived so far. For simplicity we denote macroscopic moments $m_i^n := (\rho_i^n, \rho_i^n U_i^n, E_i^n)^T$ and tolerance tol.

Now, we give discrete conservation error estimates with high order C-SL schemes with the discrete Maxwellian.

Proposition D.1. In the periodic boundary condition, conservation error estimates for the DIRK scheme of order $s=1,2,3$ in mass, momentum and energy are given by

$$\left\| \sum_{i=1}^{N_x} (m_i^{N_t} - m_i^0)\Delta x \right\|_\infty \leq \left(\sum_{k=1}^{s-1} |b_k| + |b_s| \right) \frac{N_t \Delta t}{\kappa + b_s \Delta t} (x_{\max} - x_{\min}) tol,$$

where b_k, $k=1,\cdots,s$ are determined for each $s=1,2,3$.

Proof. The DIRK scheme of order s is given by

$$f_{i,j}^{n+1} = f{*}_{i,j}^{n+1} + \frac{\Delta t}{\kappa} \sum_{k=1}^{s-1} b_k \left(d\mathcal{M}_{i,j}^{(k),n+1} - f_{i,j}^{(k),n+1} \right) + \frac{\Delta t}{\kappa} b_s \left(d\mathcal{M}{*}_{i,j}^{n+1} - f_{i,j}^{n+1} \right), \tag{D.1}$$

where $f{*}_{i,j}^{n+1} = f_{i,j}^n - \sum_{k=1}^s b_k \frac{\Delta t}{\Delta x} \big(\widehat{F}_{i+\frac{1}{2},j}^{(k),n} - \widehat{F}_{i-\frac{1}{2},j}^{(k),n} \big)$ and $d\mathcal{M}{*}_{i,j}^{n+1}$ is the discrete Maxwellian computed from $f{*}_{i,j}^{n+1}$. For $k=1,\cdots,s-1$, we denote by $d\mathcal{M}_{i,j}^{(k),n+1}$ the discrete Maxwellian computed from the k-th stage values $f_{i,j}^{(k),n+1}$ computed with classical schemes. From (D.1), we have

$$\begin{aligned}
&\left(1 + b_s \frac{\Delta t}{\kappa} \right) \sum_{i=1}^{N_x} \sum_{j=0}^{N_v} \left(f_{i,j}^{n+1} - f{*}_{i,j}^{n+1} \right) \phi(v_j) \Delta v \Delta x \\
=& \sum_{i=1}^{N_x} \sum_{j=0}^{N_v} \left[\frac{\Delta t}{\kappa} \sum_{k=1}^{s-1} b_k \left(d\mathcal{M}_{i,j}^{(k),n+1} - f_{i,j}^{(k),n+1} \right) + \frac{\Delta t}{\kappa} b_s \left(d\mathcal{M}{*}_{i,j}^{n+1} - f{*}_{i,j}^{n+1} \right) \right] \phi(v_j) \Delta v \Delta x.
\end{aligned} \tag{D.2}$$

Using

$$\begin{aligned}
&\max_i \left\| \sum_{j=0}^{N_v} \left[f_{i,j}^{(k),n+1} - d\mathcal{M}_{i,j}^{(k),n+1} \right] \phi(v_j) \Delta v \right\|_\infty \leq tol, \\
&\max_i \left\| \sum_{j=0}^{N_v} \left[f{*}_{i,j}^{n+1} - d\mathcal{M}{*}_{i,j}^{n+1} \right] \phi(v_j) \Delta v \right\|_\infty \leq tol,
\end{aligned}$$

we can get from (D.2)

$$\begin{aligned}
&\left\|\left(1+b_s\frac{\Delta t}{\kappa}\right)\sum_{i=1}^{N_x}\sum_{j=0}^{N_v}\left(f_{i,j}^{n+1}-f^{*}{}_{i,j}^{n+1}\right)\phi(v_j)\Delta v\Delta x\right\|_\infty\\
\le&\frac{\Delta t}{\kappa}\sum_{i=1}^{N_x}\left(\sum_{k=1}^{s-1}|b_k|+|b_s|\right)tol\Delta x\\
=&\left(\sum_{k=1}^{s-1}|b_k|+|b_s|\right)\frac{\Delta t}{\kappa}(x_{\max}-x_{\min})tol.
\end{aligned}$$

For DIRK schemes of order s in Section 4.1, we have $b_s>0$, which implies

$$\left\|\sum_{i=1}^{N_x}\left(f_{i,j}^{n+1}-f^{*}{}_{i,j}^{n+1}\right)\phi(v_j)\Delta v\Delta x\right\|_\infty\le\left(\sum_{k=1}^{s-1}|b_k|+|b_s|\right)\frac{\Delta t}{\kappa+b_s\Delta t}(x_{\max}-x_{\min})tol.$$

Moreover, the periodic boundary condition gives

$$\sum_{i=1}^{N_x}\sum_{j=0}^{N_v}f^{*}{}_{i,j}^{n+1}\phi(v_j)\Delta v\Delta x=\sum_{i=1}^{N_x}\sum_{j=0}^{N_v}f_{i,j}^{n}\phi(v_j)\Delta v\Delta x,$$

and hence

$$\left\|\sum_{i=1}^{N_x}(m_i^{n+1}-m_i^n)\Delta x\right\|_\infty\le\left(\sum_{k=1}^{s-1}|b_k|+|b_s|\right)\frac{\Delta t}{\kappa+b_s\Delta t}(x_{\max}-x_{\min})tol.$$

Finally, we can conclude that

$$\begin{aligned}
\left\|\sum_{i=1}^{N_x}(m_i^{N_t}-m_i^0)\Delta x\right\|_\infty\le&\sum_{i=1}^{N_x}\sum_{r=1}^{N_t}\left\|(m_i^r-m_i^{r-1})\Delta x\right\|_\infty\\
\le&\left(\sum_{k=1}^{s-1}|b_k|+|b_s|\right)\frac{N_t\Delta t}{\kappa+b_s\Delta t}(x_{\max}-x_{\min})tol.
\end{aligned}$$

This completes the proof. □

Analogue results hold for the BDF methods, which can be derived by similar (but more tedious) calculations. We present it without proof.

Proposition D.2. In the periodic boundary condition, conservation error estimates for the BDF scheme of order $s=2,3$ in mass, momentum and energy are given by

$$\begin{aligned}
&\left\|\sum_{i=1}^{N_x}(m_i^{N_t}-m_i^0)\Delta x\right\|_\infty\\
\le&\gamma_s\left(\frac{(N_t-s)\beta_s\Delta t}{\kappa+\beta_s\Delta t}+\left(\sum_{k=1}^{s-1}|b_k|+|b_s|\right)\frac{s\Delta t}{\kappa+b_s\Delta t}\right)(x_{\max}-x_{\min})tol,
\end{aligned}$$

where b_k, $k=1,\cdots,s$ and β_s are determined for each $s=2,3$ and $\gamma_2=\frac{3}{2}$ and $\gamma_3=\frac{146}{11}$.

E Consistency of C-SL method to the Euler equation in the limit of small Knudsen number

Here, we prove Proposition 6.1.

Proof. **Step 1: SL scheme**: Consider the characteristic formation of (1.1):

$$\begin{aligned}&\frac{df}{ds}=\frac{1}{\kappa}(\mathcal{M}(f)-f), \quad f(x(t),v(t),t)=f(x,v,t),\\&\frac{dx}{ds}=v, \quad x(t)=x, \quad \frac{dv}{ds}=0, \quad v(t)=v.\end{aligned}$$

Applying the implicit Euler method to this, we obtain

$$\frac{f^{n+1}-\tilde{f}^n}{\Delta t}=\frac{1}{\kappa}\Big(\mathcal{M}(f^{n+1})-f^{n+1}\Big), \tag{E.1}$$

where $\tilde{f}^n:=f^n(x-v\Delta t,v)$. Now, take moments of (E.1) with respect to $\phi(v)$

$$\int_{\mathbb{R}}\phi(v)f^{n+1}dv=\int_{\mathbb{R}}\phi(v)\tilde{f}^n dv, \tag{E.2}$$

which implies that

$$\mathcal{M}(f^{n+1})=\mathcal{M}(\tilde{f}^n) \tag{E.3}$$

because both Maxwellians have the same moments. Then, the classical SL scheme is explicitly represented by

$$f^{n+1}=\frac{\kappa}{\kappa+\Delta t}\tilde{f}^n+\frac{\Delta t}{\kappa+\Delta t}\mathcal{M}(\tilde{f}^n), \tag{E.4}$$

and hence

$$f^{n+1}=\mathcal{M}(\tilde{f}^n)+\mathcal{O}(\kappa),$$

for any $n\geq 0$ regardless of the initial data. Furthermore, the relation (E.3) implies

$$f^{n+1}=\mathcal{M}(f^{n+1})+\mathcal{O}(\kappa). \tag{E.5}$$

Therefore, as $\kappa\to 0$, we get from (E.5) that

$$f^n=\mathcal{M}(f^n), \quad \text{for } n\geq 1. \tag{E.6}$$

For $n\geq 1$, we substitute this into (E.2)

$$\int_{\mathbb{R}}\phi(v)\mathcal{M}(f^{n+1})dv=\int_{\mathbb{R}}\phi(v)f^n(x-v\Delta t,v)dv=\int_{\mathbb{R}}\phi(v)\mathcal{M}(f^n)(x-v\Delta t,v)dv,$$

and expand the r.h.s. in Taylor series to get,

$$\begin{aligned}\int_{\mathbb{R}} \phi(v)\mathcal{M}(f^{n+1})dv &= \int_{\mathbb{R}} \phi(v)\mathcal{M}(f^n)(x-v\Delta t,v)dv \\ &= \int_{\mathbb{R}} \phi(v)\Big[\mathcal{M}(f^n)(x,v) - v\Delta t\partial_x \mathcal{M}(f^n)(x,v) \\ &\qquad + \int_x^{x-v\Delta t} \partial_x^2 f^n(y,v)(x-v\Delta t-y)dy\Big]dv \\ &=: \int_{\mathbb{R}} \phi(v)\left[\mathcal{M}(f^n)(x,v) - v\Delta t\partial_x\mathcal{M}(f^n)(x,v)\right]dv + R. \end{aligned} \tag{E.7}$$

From our assumption (6.2), the remainder term R is estimated as

$$\begin{aligned}|R| &\le \int_{\mathbb{R}} (1+|v|^2)\left|\int_x^{x-v\Delta t} \partial_x^2 f^n(y,v)(x-v\Delta t-y)dy\right|dv \\ &\le \sup_{x,v}\left|\partial_x^2 f^n(x,v)(1+|v|^2)^3\right| \int_{\mathbb{R}} \frac{1}{(1+|v|^2)^2}\left|\int_x^{x-v\Delta t}(x-v\Delta t-y)dy\right|dv \\ &\le \sup_{x,v}\left|\partial_x^2 f^n(x,v)(1+|v|^2)^3\right| \int_{\mathbb{R}} \frac{v^2(\Delta t)^2}{2}\frac{1}{(1+|v|^2)^2}dydv \\ &= \mathcal{O}(\Delta t^2). \end{aligned} \tag{E.8}$$

Hence, we have from (E.7) and (E.8) that

$$\frac{m[f^{n+1}]-m[f^n]}{\Delta t} + \partial_x \begin{pmatrix} \rho^n U^n \\ \rho^n|U^n|^2 + \rho^n T^n \\ (E^n + \rho^n T^n)U^n \end{pmatrix} = \mathcal{O}(\Delta t), \tag{E.9}$$

where $m[f^{n+1}] := (\rho^{n+1}, \rho^{n+1}U^{n+1}, E^{n+1})^\top$.

Step 2. C-SL scheme: We apply the implicit Euler method to (1.1):

$$\frac{f^{n+1}-f^n}{\Delta t} = -v\partial_x f^{n+1} + \frac{1}{\kappa}\left(\mathcal{M}(f^{n+1}) - f^{n+1}\right). \tag{E.10}$$

Then, we compute the convection term $v\partial_x f^{n+1}$ is using the classical SL scheme (E.4):

$$f^{(1),n+1} = \frac{\kappa}{\kappa+\Delta t}\tilde{f}^n + \frac{\Delta t}{\kappa+\Delta t}\mathcal{M}(\tilde{f}^n). \tag{E.11}$$

If there is no confusion, we hereafter omit the time index for $f^{(1),n+1}$. Now, we replace $v\partial_x f^{n+1}$ in (E.10) with $v\partial_x f^{(1)}$:

$$\frac{f^{n+1}-f^n}{\Delta t} = -v\partial_x f^{(1)} + \frac{1}{\kappa}\left(\mathcal{M}(f^{n+1}) - f^{n+1}\right). \tag{E.12}$$

Integrating this with respect to $\phi(v)dv$, we obtain

$$\int_{\mathbb{R}}\phi(v)f^{n+1}dv=\int_{\mathbb{R}}\phi(v)(f^n-v\Delta t\partial_x f^{(1)})dv, \tag{E.13}$$

and hence

$$\mathcal{M}(f^{n+1})=\mathcal{M}(f^n-v\Delta t\partial_x f^{(1)}). \tag{E.14}$$

To sum up, we can rewrite our C-SL scheme (E.12) as

$$f^{n+1}=\frac{\kappa}{\kappa+\Delta t}(f^n-v\Delta t\partial_x f^{(1)})+\frac{\Delta t}{\kappa+\Delta t}\mathcal{M}(f^n-v\Delta t\partial_x f^{(1)}). \tag{E.15}$$

Note that (E.14) and (E.15) imply

$$f^{n+1}=\mathcal{M}(f^{n+1})+\mathcal{O}(\kappa), \tag{E.16}$$

for any $n\geq 0$ regardless of initial data. Therefore, as $\kappa\to 0$, we can deduce from (E.6) and (E.16) that

$$f^{(1),n+1}=\mathcal{M}(f^{(1),n+1}),\quad f^{n+1}=\mathcal{M}(f^{n+1}),\quad n\geq 0.$$

For $n\geq 1$, substituting these identities into (E.13) yields

$$\int_{\mathbb{R}}\phi(v)\mathcal{M}(f^{n+1})dv=\int_{\mathbb{R}}\phi(v)\left(\mathcal{M}(f^n)-v\Delta t\partial_x\mathcal{M}(f^{(1)})\right)dv,$$

which leads to

$$\frac{m[f^{n+1}]-m[f^n]}{\Delta t}+\partial_x\begin{pmatrix}\rho^{(1)}U^{(1)}\\ \rho^{(1)}|U^{(1)}|^2+\rho^{(1)}T^{(1)}\\ (E^{(1)}+\rho^{(1)}T^{(1)})U^{(1)}\end{pmatrix}=0,$$

where

$$\begin{pmatrix}\rho^{(1)}\\ \rho^{(1)}U^{(1)}\\ \tilde{E}^{(1)}\end{pmatrix}(x)=\int_{\mathbb{R}}\phi(v)f^{(1)}(x,v)dv,\quad T^{(1)}(x)=\frac{1}{\rho^{(1)}}\int_{\mathbb{R}}|v-U^n(x,v)|^2f^{(1)}(x,v)dv.$$

Now, we will show that

$$\partial_x\begin{pmatrix}\rho^{(1)}U^{(1)}\\ \rho^{(1)}|U^{(1)}|^2+\rho^{(1)}T^{(1)}\\ (E^{(1)}+\rho^{(1)}T^{(1)})U^{(1)}\end{pmatrix}=\partial_x\begin{pmatrix}\rho^nU^n\\ \rho^n|U^n|^2+\rho^nT^n\\ (E^n+\rho^nT^n)U^n\end{pmatrix}+\mathcal{O}(\Delta t). \tag{E.17}$$

For this, we recall (E.7):

$$
\begin{aligned}
&\int_{\mathbb{R}} \phi(v) f^{(1)} dv \\
&= \int_{\mathbb{R}} \phi(v) \left[\mathcal{M}(f^n)(x,v) - v\Delta t \partial_x \mathcal{M}(f^n)(x,v) + \int_x^{x-v\Delta t} \partial_x^2 f^n(y,v)(x-v\Delta t - y) dy \right] dv \\
&=: \int_{\mathbb{R}} \phi(v) \left[\mathcal{M}(f^n)(x,v) - v\Delta t \partial_x \mathcal{M}(f^n)(x,v) \right] dv + R.
\end{aligned}
$$

Here the remainder term R satisfies

$$
\partial_x R = \partial_x \left\{ \int_{\mathbb{R}} \phi(v) \left[\int_x^{x-v\Delta t} \partial_x^2 f^n(y,v)(x-v\Delta t - y) dy \right] dv \right\} = \int_{\mathbb{R}} \phi(v) \partial_x^2 f^n(x,v) v\Delta t dv,
$$

which, in view of the assumption (6.2) gives

$$
\begin{aligned}
\partial_x \begin{pmatrix} \rho^{(1)} \\ \rho^{(1)} U^{(1)} \\ E^{(1)} \end{pmatrix} &= \partial_x \left\{ \int_{\mathbb{R}} \phi(v) \left[\mathcal{M}(f^n)(x,v) - v\Delta t \partial_x \mathcal{M}(f^n)(x,v) \right] dv + R \right\} \\
&= \int_{\mathbb{R}} \phi(v) \left[\partial_x \mathcal{M}(f^n)(x,v) - v\Delta t \partial_x^2 \mathcal{M}(f^n)(x,v) \right] dv + \partial_x R \\
&= \partial_x \begin{pmatrix} \rho^n \\ \rho^n U^n \\ E^n \end{pmatrix} + \mathcal{O}(\Delta t). \qquad \text{(E.18)}
\end{aligned}
$$

Therefore, the first row in (E.17) follows from the second row in (E.18). Owing to the following identity

$$
\rho^{(1)} |U^{(1)}|^2 + \rho^{(1)} T^{(1)} = 2E^{(1)},
$$

the second row in (E.17) also holds due to the third row in (E.18). Then, it remains to show that

$$
\partial_x \left\{ (E^{(1)} + \rho^{(1)} T^{(1)}) U^{(1)} \right\} = \partial_x \left\{ (E^n + \rho^n T^n) U^n \right\} + \mathcal{O}(\Delta t).
$$

For this, we need to show the following identities:

$$
U^{(1)} = U^n + \mathcal{O}(\Delta t), \quad \partial_x U^{(1)} = \partial_x U^n + \mathcal{O}(\Delta t).
$$

Note that assumptions (6.2) and (6.3) together with estimates (E.9) and (E.18) implies that

$$
U^{(1)} - U^n = \frac{(\rho^{(1)} U^{(1)} - \rho^n U^n) + (\rho^n U^n - \rho^{(1)} U^n)}{\rho^{(1)}} = \mathcal{O}(\Delta t),
$$

and

$$\begin{aligned}\partial_x\{U^{(1)}-U^n\}=&\frac{\partial_x\left\{\rho^{(1)}U^{(1)}-\rho^nU^n\right\}}{\rho^{(1)}}-\frac{\left(\rho^{(1)}U^{(1)}-\rho^nU^n\right)\partial_x\rho^{(1)}}{(\rho^{(1)})^2}\\&+\frac{\left(\partial_x\{\rho^n-\rho^{(1)}\}U^n+\{\rho^n-\rho^{(1)}\}\partial_xU^n\right)\rho^{(1)}-\left(\rho^nU^n-\rho^{(1)}U^n\right)\partial_x\rho^{(1)}}{(\rho^{(1)})^2}\\=&\mathcal{O}(\Delta t).\end{aligned}$$

Consequently, our claim (E.17) holds and it gives the desired result:

$$\frac{m[f^{(n+1)}]-m[f^n]}{\Delta t}+\partial_x\begin{pmatrix}\rho^nU^n\\ \rho^n|U^n|^2+\rho^nT^n\\ (E^n+\rho^nT^n)U^n\end{pmatrix}=\mathcal{O}(\Delta t).$$

This completes the proof. □

In the proof, it is worth noting that both the classical SL scheme (E.11) and the C-SL scheme (E.15) preserve macroscopic moments. For simplicity, consider the periodic boundary condition $\mathbb{T}:=\mathbb{R}/\mathbb{Z}$ on the physical space. The classical SL scheme (E.11) satisfies

$$\begin{aligned}\int_{\mathbb{T}}\int_{\mathbb{R}}\phi(v)f^{(1)}dvdx&=\int_{\mathbb{R}}\int_{\mathbb{T}}\phi(v)f^n(x-v\Delta t,v)dxdv\\&=\int_{\mathbb{R}}\int_{\mathbb{T}}\phi(\tilde{v})f^n(\tilde{x},\tilde{v})\left|\frac{\partial(x,v)}{\partial(\tilde{x},\tilde{v})}\right|d\tilde{x}d\tilde{v},\end{aligned}$$

where we use the change of variable $\tilde{x}=x-v\Delta t$, $\tilde{v}=v$. Using $\left|\frac{\partial(x,v)}{\partial(\tilde{x},\tilde{v})}\right|=1$, we obtain

$$\int_{\mathbb{T}}\int_{\mathbb{R}}\phi(v)f^{(1)}dvdx=\int_{\mathbb{T}}\int_{\mathbb{R}}\phi(v)f^ndvdx.$$

Also, in case of the C-SL scheme (E.15), it satisfies (E.13) and hence

$$\int_{\mathbb{T}}\int_{\mathbb{R}}\phi(v)f^{n+1}dvdx=\int_{\mathbb{T}}\int_{\mathbb{R}}\phi(v)f^ndvdx-\int_{\mathbb{T}}\int_{\mathbb{R}}\phi(v)v\Delta t\partial_xf^{(1)}dvdx.$$

Using (E.11), the second term can be written as

$$\int_{\mathbb{T}}\int_{\mathbb{R}}\phi(v)v\Delta t\partial_xf^{(1)}dvdx=\int_{\mathbb{T}}\int_{\mathbb{R}}\phi(v)v\Delta t\partial_x\left(\frac{\kappa}{\kappa+\Delta t}\tilde{f}^n+\frac{\Delta t}{\kappa+\Delta t}\mathcal{M}(\tilde{f}^n)\right)dvdx,$$

which vanishes due to the periodicity of the physical domain. Therefore, we can conclude that

$$\int_{\mathbb{T}}\int_{\mathbb{R}}\phi(v)f^{n+1}dvdx=\int_{\mathbb{T}}\int_{\mathbb{R}}\phi(v)f^ndvdx.$$

Here we only consider time discretization for SL and C-SL schemes. Note that conservation is also valid for the fully discretized C-SL schemes (See Appendix D).

References

[1] P. Andries, J.-F. Bourgat, P. Le Tallec and B. Perthame, Numerical comparison between the Boltzmann and ES-BGK models for rarefied gases, Comput. Methods Appl. Mech. Engrg., 191 (2002), 3369-3390.

[2] P. Andries, P. Le 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), 813-830.

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

[4] C. Bardos, F. Golse and D. Levermore, Fluid dynamic limits of kinetic equations. I. Formal derivations, J. Stat. Phys., 63 (1991), 323-344.

[5] 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), 511.

[6] F. Bouchut, F. Golse and M. Pulvirenti, Kinetic equations and asymptotic theory, Gauthier-Villars, Paris, 2000.

[7] B. Yan and S. Jin, A successive penalty-based asymptotic-preserving scheme for kinetic equations, SIAM J. Sci. Comput., 35 (2013), A150-A172.

[8] E. Carlini, R. Ferretti and G. Russo, A weighted essentially nonoscillatory, large time-step scheme for Hamilton–Jacobi equations, SIAM J. Sci. Comput., 27 (2005), 1071-1091.

[9] C. Cercignani, The Boltzmann equation and its applications, Springer, New York, 1988.

[10] S. Chapman and T. G. Cowling, The mathematical theory of non-uniform gases, Cambridge Univ. Press, England, 1970.

[11] F. Coron and B. Perthame, Numerical passage from kinetic to fluid equations, SIAM J. Numer. Anal., 28 (1991), 26-42.

[12] N. Crouseilles and M. Mehrenberger and E. Sonnendrcker, Conservative semi-Lagrangian schemes for Vlasov equations, J. Comput. Phys., 229 (2010), 1927-1953.

[13] J. Douglas Jr., C. S. Huang and F. Pereira, The modified method of characteristics with adjusted advection, Numer. Math., 83 (1999), 353-369.

[14] F. Filbet, E. Sonnendrücker and P. Bertrand, Conservative numerical schemes for the Vlasov equation, J. Comput. Phys., 172 (2001), 166-187.

[15] G. Dimarco and R. Loubere, Towards an ultra efficient kinetic scheme. Part II: The high order case, J. Comput. Phys., 255 (2013), 699-719.

[16] G. Dimarco, R. Loubere and J. Narski, Towards an ultra efficient kinetic scheme. Part III: High-performance-computing, J. Comput. Phys., 284 (2015), 22-39.

[17] M. Groppi, G. Russo and G. Stracquadanio, High order semi-Lagrangian methods for the BGK equation, Commun. Math. Sci., 14 (2016), 389-414.

[18] G. Wanner and E. Hairer, Solving ordinary differential equations II, Springer Berlin Heidelberg, 1996.

[19] E. Hairer, S. P. Nørsett and G. Wanner, Solving ordinary differential equations I: Nonstiff Problem, Springer Series in Computational Mathematics, 1993.

[20] L. H. Holway, Kinetic theory of shock structure using and ellipsoidal distribution function in rarefied gas dynamics, Proceedings of the Fourth International Symposium, 1 (1964), 193-215.

[21] S. Jin, Asymptotic preserving (AP) schemes for multiscale kinetic and hyperbolic equations, Rivista di Matematica della Universita di Parma, 2010.

[22] L. Mieussens, Discrete velocity model and implicit scheme for the BGK equation of rarefied

gas dynamics, Math. Models Methods Appl. Sci., 10 (2000), 1121-1149.

[23] R. M. Pidatella, G. Puppo, G. Russo and P. Santagati, Semiconservative finite volume schemes for conservation laws, J. Sci. Comput., 41 (2019), B576-B600.

[24] S. Pieraccini and G. Puppo, Implicit-explicit schemes for BGK kinetic equations, J. Sci. Comput., 32 (2007), 1-28.

[25] J. M. Qiu and C. W. Shu, Conservative high order semi-Lagrangian finite difference WENO methods for advection in incompressible flow, J. Sci. Comput., 230 (2011), 863-889.

[26] G. Russo, J. Qiu and X. Tao, Conservative multi-dimensional semi-Lagrangian finite difference scheme: stability and applications to the kinetic and fluid simulations, J. Sci. Comput., 79 (2019), 1-30.

[27] G. Russo and P. Santagati, A new class of large time step methods for the BGK models of the Boltzmann equation, preprint, 2011.

[28] G. Russo, P. Santagati and S.-B. Yun, Convergence of a semi-Lagrangian scheme for the BGK model of the Boltzmann equation, SIAM J. Numer. Anal., 50 (2012), 1111-1135.

[29] G. Russo and S.-B. Yun, Convergence of a semi-Lagrangian scheme for the ellipsoidal BGK model of the Boltzmann equation, SIAM J. Numer. Anal., 56 (2018), 3580-3610.

[30] P. Santagati, High order semi-Lagrangian methods for the BGK model of the Boltzmann equation, Universita'degli Studi di Catania, 2008.

[31] C. W. Shu, Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws. In Advanced numerical approximation of nonlinear hyperbolic equations (Cetraro, 1997), volume 1697 of Lecture Notes in Math, Springer, Berlin, 325-432, 1998

[32] K. Xu and J. C. Huang, A unified gas-kinetic scheme for continuum and rarefied flows, J. Comput. Phys., 229 (2010), 7747-7764.