Journal of Computational Physics 229 (2010) 7625–7648

Contents lists available at ScienceDirect

Journal of Computational Physics

journal homepage: www.elsevier.com/locate/jcp

A class of asymptotic-preserving schemes for kinetic equations and related problems with stiff sources ☆

Francis Filbet [a,*], Shi Jin [b]

[a] Université de Lyon, Université Lyon I, CNRS UMR 5208, Institut Camille Jordan 43, Boulevard du 11 Novembre 1918, 69622 Villeurbanne Cedex, France
[b] Department of Mathematics, University of Wisconsin Madison, WI 53706, USA

ARTICLE INFO

Article history:
Received 8 May 2009
Received in revised form 8 June 2010
Accepted 12 June 2010
Available online 26 June 2010

Keywords:
Boltzmann equation
Asymptotic-preserving scheme
Stiff source terms

ABSTRACT

In this paper, we propose a general time-discrete framework to design asymptotic-preserving schemes for initial value problem of the Boltzmann kinetic and related equations. Numerically solving these equations are challenging due to the nonlinear stiff collision (source) terms induced by small mean free or relaxation time. We propose to penalize the nonlinear collision term by a BGK-type relaxation term, which can be solved explicitly even if discretized implicitly in time. Moreover, the BGK-type relaxation operator helps to drive the density distribution toward the local Maxwellian, thus naturally imposes an asymptotic-preserving scheme in the Euler limit. The scheme so designed does not need any nonlinear iterative solver or the use of Wild Sum. It is uniformly stable in terms of the (possibly small) Knudsen number, and can capture the macroscopic fluid dynamic (Euler) limit even if the small scale determined by the Knudsen number is not numerically resolved. It is also consistent to the compressible Navier–Stokes equations if the viscosity and heat conductivity are numerically resolved. The method is applicable to many other related problems, such as hyperbolic systems with stiff relaxation, and high order parabolic equations.

1. Introduction

The Boltzmann equation describes the time evolution of the density distribution of a dilute gas of particles when the only interactions taken into account are binary elastic collisions. For space variable $x \in \Omega \subset \mathbb{R}^{d_x}$, particle velocity $v \in \mathbb{R}^{d_v}$ ($d_v \geqslant 2$), the Boltzmann equation reads:

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f = \frac{1}{\varepsilon} Q(f), \tag{1.1}$$

where $f := f(t,x,v)$ is the time-dependent particles distribution function in the phase space. Here for simplicity, we do not study the case of Maxwell diffusion boundary condition for which boundary layers may be generated, but only consider specular or periodic boundary condition in space. The parameter $\varepsilon > 0$ is the dimensionless Knudsen number defined as the ratio of the mean free path over a typical length scale such as the size of the spatial domain, which measures the rarefiedness of the gas. The Boltzmann collision operator Q is a quadratic operator,

☆ F. Filbet is partially supported by the French ANR project "Jeunes Chercheurs" Méthodes Numériques pour les Équations Cinétiques (MNEC) and by the European Research Council ERC Starting Grant 2009, project 239983-*NuSiKiMo*. S. Jin was partially supported by NSF Grant No. DMS-0608720, NSF FRG Grant DMS-0757285, and a Van Vleck Distinguished Research Prize from University of Wisconsin-Madison.

* Corresponding author. Tel.: +33 4 72 44 62 62; fax: +33 4 72 43 16 87.
E-mail addresses: filbet@math.univ-lyon1.fr (F. Filbet), jin@math.wisc.edu (S. Jin).

0021-9991/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2010.06.017

$$\mathcal{Q}(f)(v) = \int_{\mathbb{R}^{d_v}} \int_{\mathbb{S}^{d_v-1}} B(|v - v_\star|, \cos\theta)(f'_\star f' - f_\star f) d\sigma\, dv_\star, \tag{1.2}$$

where we used the shorthanded notation $f = f(v)$, $f_\star = f(v_\star)$, $f' = f(v')$, $f'_\star = f(v'_\star)$. The velocities of the colliding pairs $(v, v_\star)$ and $(v', v'_\star)$ are related by

$$\begin{cases} v' = v - \frac{1}{2}((v - v_\star) - |v - v_\star|\sigma), \\ v'_\star = v - \frac{1}{2}((v - v_\star) + |v - v_\star|\sigma), \end{cases}$$

with $\sigma \in \mathbb{S}^{d_v-1}$. The collision kernel B is a non-negative function which by physical arguments of invariance only depends on $|v - v_\star|$ and $\cos\theta = u \cdot \sigma$ (where $u = (v - v_\star)/|v - v_\star|$ is the normalized relative velocity). In this work we assume that B is locally integrable and we will simply take

$$B(|u|, \cos\theta) = C_\alpha |u|^\alpha \tag{1.3}$$

for some $\alpha \in [0,1]$ and a constant $C_\alpha > 0$.

Boltzmann's collision operator has the fundamental properties of conserving mass, momentum and energy: at the formal level

$$\int_{\mathbb{R}^{d_v}} \mathcal{Q}(f)\phi(v) dv = 0, \quad \text{for } \phi(v) = 1, v, |v|^2 \tag{1.4}$$

and it satisfies the well-known Boltzmann's H theorem

$$-\frac{d}{dt}\int_{\mathbb{R}^{d_v}} f \log f dv = -\int_{\mathbb{R}^{d_v}} \mathcal{Q}(f)\log(f) dv \geqslant 0.$$

The functional $-\int f \log f$ is the *entropy* of the solution. Boltzmann's H theorem implies that any equilibrium distribution function, *i.e.*, any function which is a maximum of the entropy, has the form of a local Maxwellian distribution

$$\mathcal{M}_{\rho,u,T}(v) = \frac{\rho}{(2\pi T)^{d_v/2}} \exp\left(-\frac{|u - v|^2}{2T}\right),$$

where ρ, u, T are the *density*, *macroscopic velocity* and *temperature* of the gas, defined by

$$\rho = \int_{\mathbb{R}^{d_v}} f(v) dv = \int_{\mathbb{R}^{d_v}} \mathcal{M}_{\rho,u,T}(v), \quad u = \frac{1}{\rho}\int_{\mathbb{R}^{d_v}} v f(v) dv = \frac{1}{\rho}\int_{\mathbb{R}^{d_v}} v\mathcal{M}_{\rho,u,T}(v) dv, \tag{1.5}$$

$$T = \frac{1}{d_v \rho}\int_{\mathbb{R}^{d_v}} |u - v|^2 f(v) dv = \frac{1}{d_v \rho}\int_{\mathbb{R}^{d_v}} |u - v|^2 \mathcal{M}_{\rho,u,T}(v) dv. \tag{1.6}$$

Therefore, when the Knudsen number $\varepsilon > 0$ becomes very small, the macroscopic model, which describes the evolution of averaged quantities such as the density ρ, momentum ρu and temperature T of the gas, by fluid dynamics equations, namely, the compressible Euler or Navier–Stokes equations, become adequate [1,5]. More specifically, as $\varepsilon \to 0$, the distribution function will converge to a local Maxwellian $\mathcal{M}$, and system (1.2) becomes a closed system for the $2 + d_v$ moments. The conserved quantities satisfy the classical Euler equations of gas dynamics for a mono-atomic gas:

$$\begin{cases} \frac{\partial \rho}{\partial t} + \nabla_x \cdot \rho u = 0, \\ \frac{\partial \rho u}{\partial t} + \nabla_x \cdot (\rho u \otimes u + p\mathrm{I}) = 0, \\ \frac{\partial E}{\partial t} + \nabla_x \cdot ((E + p)u) = 0, \end{cases} \tag{1.7}$$

where p is the pressure, E represents the total energy

$$E = \frac{1}{2}\rho u^2 + \frac{d_v}{2}\rho T,$$

and I is the identity matrix. These equations constitute a system of $2 + d_v$ equations in $3 + d_v$ unknowns. The pressure is related to the internal energy by the constitutive relation for a polytropic gas

$$p = (\gamma - 1)\left(E - \frac{1}{2}\rho|u|^2\right),$$

where the polytropic constant $\gamma = (d_v + 2)/d_v$ represents the ratio between specific heat at constant pressure and at constant volume, thus yielding $p = \rho T$. For small but non zero values of the Knudsen number ε, the evolution equation for the moments can be derived by the so-called Chapman–Enskog expansion [10], applied to the Boltzmann equation. This approach gives the Navier–Stokes equations as a second order approximation with respect to ε to the solution of the Boltzmann equation:

$$\begin{cases} \frac{\partial \rho_\varepsilon}{\partial t} + \nabla_x \cdot \rho_\varepsilon u_\varepsilon = 0, \\ \frac{\partial \rho_\varepsilon u_\varepsilon}{\partial t} + \nabla_x \cdot (\rho_\varepsilon u_\varepsilon \otimes u_\varepsilon + p_\varepsilon \mathrm{I}) = \varepsilon \nabla_x \cdot [\mu_\varepsilon \sigma(u_\varepsilon)], \\ \frac{\partial E_\varepsilon}{\partial t} + \nabla_x \cdot (E_\varepsilon + p_\varepsilon) u_\varepsilon) = \varepsilon \nabla_x \cdot \left(\mu_\varepsilon \sigma(u_\varepsilon) u + \kappa_\varepsilon \nabla_x T_\varepsilon\right). \end{cases} \tag{1.8}$$

In these equations $\sigma(u)$ denotes the strain-rate tensor given by

$$\sigma(u) = \nabla_x u + (\nabla_x u)^T - \frac{2}{d_v} \nabla_x \cdot u \mathrm{I},$$

while the viscosity $\mu_\varepsilon = \mu(T_\varepsilon)$ and the thermal conductivity $\kappa_\varepsilon = \kappa(T_\varepsilon)$ are defined according to the linearized Boltzmann operator with respect to the local Maxwellian [1].

The connection between kinetic and macroscopic fluid dynamics results from two properties of the collision operator [1,5]:

(i) conservation properties and an entropy relation that imply that the equilibria are Maxwellian distributions for the zeroth order limit;
(ii) the derivative of $\mathcal{Q}(f)$ satisfies a formal Fredholm alternative with a kernel related to the conservation properties of (i).

Past progress on developing robust numerical schemes for kinetic equations that also work in the fluid regimes has been guided by the fluid dynamic limit, in the framework of *asymptotic-preserving* (AP) scheme. As summarized by Jin [35], a scheme for the kinetic equation is AP if

- it preserves the discrete analogy of the Chapman–Enskog expansion, namely, it is a suitable scheme for the kinetic equation, yet, when holding the mesh size and time step fixed and letting the Knudsen number go to zero, the scheme becomes a suitable scheme for the limiting Euler equations,
- implicit collision terms can be implemented explicitly, or at least more efficiently than using the Newton type solvers for nonlinear algebraic systems.

Comparing with a multi-physics domain decomposition type method [6,18,20,33,46,56], the AP schemes avoid the coupling of physical equations of different scales where the coupling conditions are difficult to obtain, and interface locations hard to determine. The AP schemes are based on solving one equation – the kinetic equation, and they become robust macroscopic (fluid) solvers *automatically* when the Knudsen number goes to zero. A generic way to prove that an AP scheme implies a numerical convergence uniformly in the Knudsen number was given by Golse–Jin–Levermore for the linear discrete-ordinate transport equation in the diffusion regime [31]. This result can be extended to essentially all AP schemes, although the specific proof is problem dependent. We refer to AP schemes for kinetic equations in the fluid dynamic or diffusive regimes [2,7,14,32,40–42,44,45,47–49]. The AP framework has also been extended in [15,16] for the study of the quasi-neutral limit of Euler–Poisson and Vlasov–Poisson systems, and in [19,21,34] for all-speed (Mach number) fluid equations bridging the passage from compressible flows to the incompressible flows. One should note that under-resolved computation may not yield accurate or even physically correct approximations in areas with sharp transitions, such as shock and boundary layers. In these areas one may want to use resolved calculations. The AP schemes allow one to use suitable mesh size and time step at needed domains with one first-principle equation, thus is especially suitable for problems with localized sharp transitions where macroscopic simulation is necessary.

To satisfy the first condition for AP, the scheme must be driven to the local Maxwellian when $\varepsilon \to 0$. Let $t^n (n = 0, 1, 2, \ldots)$ be the discrete time, and $U^n = U(t^n)$ for a general quantity U. Then an AP scheme requires that, for $\Delta t \gg \varepsilon$,

$$f^n - \mathcal{M}^n = O(\varepsilon), \quad n \geqslant 1 \tag{1.9}$$

for *any* initial data f^0. Namely, the numerical solution projects *any* data into the local Maxwellian, with an accuracy of $O(\varepsilon)$, in *one step*. This can usually be achieved by a backward Euler or any L-stable ODE solvers for the collision term [36]. Such a scheme requires an implicit collision term to guarantee a uniform stability in time. However, how to invert such an implicit, yet non-local and nonlinear, collision operator is a delicate numerical issue. Namely, it is hard to realize the second condition for AP schemes. One solution was offered by Gabetta et al. [28]. They first penalize $\mathcal{Q}$ by a linear function λf, and then absorb the linearly stiff part into the time variable to remove the stiffness. The remaining implicit nonlinear collision term is approximated by finite terms in the Wild Sum, with the infinite sum replaced by the local Maxwellian. This yields a uniformly stable AP scheme for the collision term, capturing the Euler limit when $\varepsilon \to 0$. Such a time-relaxed method was also used to develop AP Monte Carlo method, see [8,51].

When the collision operator $\mathcal{Q}$ is the BGK collision operator

$$\mathcal{Q}_{BGK} = \mathcal{M} - f, \tag{1.10}$$

it is well known that even an implicit collision term can be solved explicitly, using the property that $\mathcal{Q}$ preserves mass, momentum and energy [14]. *Our new idea* in this paper is to utilize this property, and penalize the Boltzmann collision operator $\mathcal{Q}$ by the BGK operator:

$$\mathcal{Q} = [\mathcal{Q} - \lambda(\mathcal{M} - f)] + \lambda[\mathcal{M} - f], \tag{1.11}$$

where λ is the spectral radius of the linearized collision operator of $\mathcal{Q}$ around the local Maxwellian $\mathcal{M}$.

Now the first term on the right hand side of (1.11) is either not stiff, or less stiff and less dissipative compared to the second term, thus it can be discretized *explicitly*, so as to avoid inverting the nonlinear operator $\mathcal{Q}$. The second term on the right hand side of (1.11) is stiff or dissipative, thus will be treated implicitly. As mentioned earlier, the implicit BGK operator can be inverted explicitly. Therefore we arrive at a scheme which is uniformly stable in ε, with an implicit source term that can be solved explicitly. In other words, in terms of handling the stiffness, the general Boltzmann collision operator can be handled as easily as the much simpler BGK operator, thus we significantly simplify an implicit Boltzmann solver!

A related problem is hyperbolic systems with relaxations. Such systems arise in reacting gases, shallow water equations, discrete-velocity kinetic models, etc. [57], and have been mathematically studied extensively in recent years (see for example [4,12,43,50]). A prototype example is the following 2×2 nonlinear hyperbolic system with relaxation:

$$\begin{cases} \frac{\partial u}{\partial t} + f_1(u, v)_x = 0, \\ \frac{\partial v}{\partial t} + f_2(u, v)_x = \frac{1}{\varepsilon} R(u, v). \end{cases} \tag{1.12}$$

The relaxation term $R : \mathbb{R}^2 \mapsto \mathbb{R}$ is dissipative in the sense of [12]:

$$\partial_v R \leqslant 0. \tag{1.13}$$

It possesses a unique local equilibrium, namely, $R(u, v) = 0$ implies $v = g(u)$. At the local equilibrium, one has the macroscopic system

$$u_t + f_1(u, g(u))_x = 0.$$

This system can be derived by sending $\varepsilon \to 0$ in (1.12), the so-called zero relaxation limit [12]. This limit is analogous to the passage from kinetic equations to their fluid limit, and in the last decade the development in these two areas – both analytic studies and numerical approximations – have strongly intervened. The numerical methods for such systems are similar to those developed for the Boltzmann equations, especially for discrete-velocity kinetic models [7,36]. The guiding principle for the AP schemes is the same for both classes of problems, and in this paper we will study both applications whenever appropriate.

Let V^n and U^n be the time-discrete approximations to v and u respectively in (1.12). A classical AP scheme requires that, for $\Delta t \gg \varepsilon$,

$$V^n - g(U^n) = O(\varepsilon), \quad n \geqslant 1 \tag{1.14}$$

for *any* initial data V^0. Namely, the numerical solution projects *any* data V into the local equilibrium $V = f(U)$, with an accuracy of $O(\varepsilon)$, in *one step*. This is the analogy of (1.9). Our new method is not necessarily AP in the classical sense of (1.14). Nevertheless, we can show that, for any ε, and $\Delta t \gg \varepsilon$, there exists an $N_\varepsilon \geqslant 1$, such that

$$V^n - g(U^n) = O(\varepsilon), \quad n \geqslant N_\varepsilon \tag{1.15}$$

for *any* initial data V^0. Namely, the numerical solution projects the solution into the local equilibrium after the initial transient time, for *any* initial data. This is a slightly weaker condition than (1.14), but is enough to guarantee the desired numerical performance as good as the classical AP schemes.

Although a linear penalty (by removing $\mathcal{M}$ on the right hand side of (1.11)) can also remove the stiffness, we can show that, when applied to the relaxation system (1.12), it only has the following property:

$$V^n - g(U^n) = O(\Delta t), \quad n \geqslant N \tag{1.16}$$

for *any* initial data V^0, when $\Delta t \gg \varepsilon$. Since in the fluid regime, we really want to take $\Delta t \gg \varepsilon$, schemes with a weak AP property (1.16) is much less accurate than our scheme which has the property (1.15). The BGK operator that we use in (1.11) helps to drive f into $\mathcal{M}$ (or V into $g(u)$) more effectively than a linear damping $-\lambda f$, thus preserves the Euler limit more accurately. Moreover, if $v^n - g(U^n) = O(\varepsilon)$ (well-prepared initial data), then our method implies that $v^{n+1} - g(U^{n+1}) = O(\varepsilon)$, while the linear penalty method always yields $v^{n+1} - g(U^{n+1}) = O(\Delta t)$ even for well-prepared initial data.

For the Boltzmann equation, although we cannot analytically prove an analogy of (1.15) for $f - \mathcal{M}$, our numerical examples show that this is true. We can prove, however, that if the initial data are well prepared,

$$f^n - \mathcal{M}^n = O(\varepsilon) \quad \text{for some} \quad n = N \geqslant 0.$$

then the scheme captures the correct Euler limit for later time $n > N$. Moreover, for suitably small time step, our method is also consistent to the Navier–Stokes Eq. (1.8) for $\varepsilon \ll 1$.

Our method is partly motivated by the work of Haack et al. [34], where by subtracting the leading linear part of the pressure in the compressible Euler equations with a low Mach number, the nonlinear stiffness in the pressure term due to the low Mach number is removed and an AP scheme was proposed for the compressible Euler or Navier–Stokes equations that capture the incompressible Euler or Navier–Stokes limit when the Mach number goes to zero. In terms of removing the stiffness of nonlinear parabolic equations Smereka used the idea of adding and subtracting a linear elliptic operator. However his approach was not aimed at achieving the AP property.

Our method is not restricted to the Boltzmann equation. It applies to general nonlinear hyperbolic systems with stiff nonlinear relaxation terms [12,13,36,39], as will be shown in Section 3, and higher-order parabolic equations (see Section 6). Indeed, it applies to any *stiff source term that admits a stable local equilibrium*.

We will present and study this framework for stiff ODEs (Section 2), nonlinear hyperbolic system with relaxation (Section 3), and the Boltzmann equation (Section 4). We present different numerical tests on the Boltzmann equation in Section 5 to illustrate the efficiency of the present method. In particular, we will include a multi-scale problem where the Knudsen number ε depends on the space variable and takes different values ranging from 10^{-4} (hydrodynamic regime) to *1* (kinetic regime). Finally, in Section 6, we design a scheme for the nonlinear Fokker–Planck equations for which the asymptotic-preserving scheme can be used to remove the CFL constraint of a parabolic equation. We conclude the paper in Section 7.

2. Asymptotic-preserving (AP) stiff ODE solvers

We first present our method for stiff ordinary differential equations. Let us consider a Hilbert space H and the following nonlinear autonomous ordinary differential system

$$\begin{cases} \frac{df_\varepsilon}{dt}(t) = \frac{\mathcal{Q}(f_\varepsilon)}{\varepsilon}, \quad t \geqslant 0, \\ f_\varepsilon(0) = f_0 \in H, \end{cases} \tag{2.1}$$

where the source term $\mathcal{Q}(f)$ satisfies the following properties:

- there exists a unique stationary solution $\mathcal{M}$ to (2.1), namely, $\mathcal{Q}(\mathcal{M}) = 0$;
- the solution to (2.1) converges to the steady state $\mathcal{M}$ when time goes to infinity, and the spectrum of $\nabla\mathcal{Q}(f) \subset \mathbb{C}^- = \{z | z \in \mathbb{C}^-, \mathrm{Re}(z) < 0\}$,

$$0 < \alpha_m \leqslant \|\nabla\mathcal{Q}(f)\| \leqslant \alpha_M, \quad \forall f \in H \setminus \{0\}. \tag{2.2}$$

where $\nabla\mathcal{Q}(f)$ denotes the Frechet derivative of $\mathcal{Q}$.

Remark 2.1. The second hypothesis above is certainly not the most general, but is convenient for our purpose. The lower bound implies that the solution converges to the steady state $\mathcal{M}$, while the upper bound is a sufficient condition for existence and uniqueness of a global solution.

When ε becomes small, the differential Eq. (2.1) becomes stiff and explicit schemes are subject to severe stability constraints. Of course, implicit schemes allow larger time step, but new difficulty arises in seeking the numerical solution of a fully nonlinear problem at each time step. Here we want to combine both advantages of implicit and explicit schemes: large time step for stiff problems and low computational complexity of the numerical solution at each time step.

Two classical procedures handle the aforementioned difficulties well. One is to linearize the unknown $\mathcal{Q}(f^{n+1})$ at time step t^{n+1} around f at the previous time step f^n:

$$\mathcal{Q}(f^{n+1}) \approx \mathcal{Q}(f^n) + \nabla\mathcal{Q}(f^n)(f^{n+1} - f^n). \tag{2.3}$$

This yields a problem that only needs to solve a linear system with coefficient matrices depending on $\nabla\mathcal{Q}(f^n)$ [58]. This approach gives a uniformly stable time discretization without nonlinear solvers. The second approach, introduced in [28], takes

$$\mathcal{Q}(f) = [\mathcal{Q}(f) - \mu f] + \mu f. \tag{2.4}$$

In [42], the second μf term in absorbed into the time derivative, which removes the stiffness, and then $Q(f)$ is approximated by the Wild Sum which is truncated at finite terms with the remaining infinite series replaced by the local Maxwellian in order to become AP. If one is just interested in removing the stiffness, one can just approximate the right hand side of (2.4) by

$$[\mathcal{Q}(f^n) - \mu f^n] + \mu f^{n+1}.$$

For sufficiently large μ, this yields a scheme with stability independent of ε, yet can be solved explicitly. However, a disadvantage of the linear penalty method, as well as method (2.3), is that the operators on the right hand size do not preserve exactly the mass, momentum and total energy as the BGK operator does.

As will be shown in Section 3, these two classical approaches project the data into the local equilibrium in the sense of (1.16).

We propose to split the source term of (2.1) as the sum of a stiff-dissipative part and a non-(or less) stiff and non-dissipative part as

$$\frac{\mathcal{Q}(f)}{\varepsilon} = \underbrace{\frac{\mathcal{Q}(f) - P(f)}{\varepsilon}}_{\text{less stiff part}} + \underbrace{\frac{P(f)}{\varepsilon}}_{\text{stiff, dissipative part}}, \tag{2.5}$$

where $P(f)$ is a *well balanced*, i.e. preserving the steady state, $P(\mathcal{M}) = 0$, linear operator and is asymptotically close to the source term $\mathcal{Q}(f)$. For instance, performing a simple Taylor expansion, we get

$$\mathcal{Q}(f) = \mathcal{Q}(\mathcal{M}) + \nabla\mathcal{Q}(\mathcal{M})(f-\mathcal{M}) + O(\|f-\mathcal{M}\|_H^2)$$

and we may choose

$$P(f) := \nabla\mathcal{Q}(\mathcal{M})(f-\mathcal{M}).$$

Since it is not always possible to compute exactly $\nabla\mathcal{Q}(\mathcal{M})$, we may simply choose

$$P(f) := \beta(\mathcal{M}-f),$$

where β is an upper bound of $\|\nabla\mathcal{Q}(\mathcal{M})\|$ or some approximation of it such as $[\mathcal{Q}(f)-\mathcal{Q}(\mathcal{M})]/(f-\mathcal{M})$.

In the following we propose a discretization to (2.5) based on IMEX schemes.

We simply apply a first order implicit-explicit (IMEX) scheme for the time discretization of (2.1):

$$\frac{f^{n+1}-f^n}{\Delta t} = \frac{\mathcal{Q}(f^n)-P(f^n)}{\varepsilon} + \frac{P(f^{n+1})}{\varepsilon}, \tag{2.6}$$

or

$$f^{n+1} = [\varepsilon \mathrm{I} - \Delta t\nabla\mathcal{Q}(\mathcal{M})]^{-1}[\varepsilon f^n + \Delta t(\mathcal{Q}(f^n)-P(f^n)) - \Delta t\nabla\mathcal{Q}(\mathcal{M})\mathcal{M}].$$

This method is easy to implement, since f^{n+1} is linear in the right hand side of (2.6). For linear problems, we have the following result:

Theorem 2.2. *Consider the differential system (2.1) with $\mathcal{Q}(f) = -\lambda f$, where Re $(\lambda) > 0$. Set $P(f) := -\nu\lambda f$ with $\nu \geqslant 0$. Then, the scheme (2.6) is A-stable and L-stable for $\nu > 1/2$.*

Proof. For linear systems with $\mathcal{Q}(f) = -\lambda f$, the scheme simple reads

$$f^{n+1} = \frac{\varepsilon + (\nu-1)\lambda\Delta t}{\varepsilon + \nu\lambda\Delta t} f^n = \left(1 - \frac{\lambda\Delta t}{\varepsilon + \nu\lambda\Delta t}\right) f^n.$$

Observe that $\nu = 0$ gives the explicit Euler scheme, which is stable only for $\Delta t \leqslant \varepsilon/\lambda$, whereas for $0 \leqslant \nu \leqslant 1$, it yields the so-called θ-scheme, which is A-stable for $\nu > 1/2$. For $\nu = 1$ it corresponds to the A-stable implicit Euler scheme. Moreover,

$$\|f^{n+1}\|_H \leqslant \left|1 - \frac{\lambda\Delta t}{\varepsilon + \nu\lambda\Delta t}\right| \|f^n\|_H \sim \left(1 - \frac{1}{\nu}\right)\|f^n\|_H \quad \text{for } \varepsilon \sim 0 \quad \text{or } \lambda\Delta t \gg 1,$$

where $|1 - \frac{1}{\nu}| < 1$ for $\nu > 1/2$. This is also the condition for the L-stability [30]. Clearly $\lambda \sim 1$ gives the fastest convergence to the equilibrium. □

Concerning nonlinear problems, we observe that the scheme (2.6) is not AP in the sense of (1.9). However, we can prove that it is AP in the sense of (1.15).

Theorem 2.3. *Assume that the operator $\mathcal{Q}$ satisfies (2.2) and*

$$\frac{\mathcal{Q}(f^n)-\mathcal{Q}(\mathcal{M})}{f^n-\mathcal{M}} < 0. \tag{2.7}$$

Assume that $\Delta t \gg \varepsilon$. Then, for β sufficiently large, there exists an $0 < r < 1$, independent of Δt and ε, such that

$$|f^n - \mathcal{M}| \leqslant r^n |f^0 - \mathcal{M}|.$$

Consequently scheme (2.6) is AP in the sense of (1.15).

Proof. We choose $\beta > 0$ such that

$$\beta > \frac{1}{2}\sup_{f\in H}\left|\frac{\mathcal{Q}(f)-\mathcal{Q}(\mathcal{M})}{f-\mathcal{M}}\right| = \frac{1}{2}\alpha_M.$$

Scheme (2.6) can be written as

$$\frac{[f^{n+1}-\mathcal{M}]-[f^n-\mathcal{M}]}{\Delta t} = \frac{1}{\varepsilon}\left[\frac{\mathcal{Q}(f^n)-\mathcal{Q}(\mathcal{M})}{f^n-\mathcal{M}} + \beta\right](f^n-\mathcal{M}) - \frac{\beta[f^{n+1}-\mathcal{M}]}{\varepsilon}.$$

This gives

$$\left(1 + \frac{\beta\Delta t}{\varepsilon}\right)[f^{n+1}-\mathcal{M}] = \left(1 + \frac{\Delta t}{\varepsilon}\mathcal{D}^n\right)[f^n-\mathcal{M}],$$

where $\mathcal{D}^n$ is given by

$$\mathcal{D}^n = \frac{\mathcal{Q}(f^n) - \mathcal{Q}(\mathcal{M})}{f^n - \mathcal{M}} + \beta.$$

Clearly, under the assumption (2.7),

$$|\mathcal{D}^n| \leqslant \beta - \frac{1}{2}\alpha_M.$$

Thus,

$$f^{n+1} - \mathcal{M} = \frac{\varepsilon + \Delta t \mathcal{D}^n}{\varepsilon + \Delta t \beta}[f^n - \mathcal{M}^n].$$

For $\Delta t \gg \varepsilon$,

$$r = \sup_{\varepsilon,n,\Delta t}\left|\frac{\varepsilon + \Delta t \mathcal{D}^n}{\varepsilon + \Delta t \beta}\right| \sim \sup_n \frac{\beta - \frac{1}{2}\alpha_M}{\beta} < 1$$

hence (2.8) implies

$$|f^{n+1} - \mathcal{M}| \leqslant r|f^n - \mathcal{M}|.$$

From here it is simple to see that

$$|f^n - \mathcal{M}| \leqslant r^n|f^0 - \mathcal{M}|.$$

So for any $\varepsilon > 0$, and any initial data, there exists an $N_\varepsilon \geqslant 1$ such that when $n \geqslant N_\varepsilon, f^n - \mathcal{M} = O(\varepsilon)$. This is the AP property defined in (1.15). □

To improve the numerical accuracy, second order schemes are sometimes more desirable. Thus, we propose the following second order IMEX extension. Assume that an approximate solution f^n is known at time t^n, we compute a first approximation at time t^* using a first order IMEX scheme and next apply the trapezoidal rule and the mid-point formula. The scheme reads

$$\begin{cases} 2\frac{f^\star - f^n}{\Delta t} = \frac{\mathcal{Q}(f^n) - P(f^n)}{\varepsilon} + \frac{P(f^\star)}{\varepsilon}, \\ \frac{f^{n+1} - f^n}{\Delta t} = \frac{\mathcal{Q}(f^\star) - P(f^\star)}{\varepsilon} + \frac{P(f^n) + P(f^{n+1})}{2\varepsilon}. \end{cases} \tag{2.8}$$

For $Q = -\lambda f$, $P = -v\lambda f$, (2.8) gives

$$f^{n+1} = \frac{1 + \frac{\Delta t}{\varepsilon}\lambda(v-1) + \frac{1}{4}\left(\frac{\Delta t}{\varepsilon}\right)^2\lambda^2(v^2 - 4v + 2)}{\left(1 + \frac{\Delta t}{2\varepsilon}v\lambda\right)^2} f^n.$$

For $\Delta t \gg \varepsilon$ this gives

$$f^{n+1} \sim \frac{v^2 - 4v + 2}{v^2} f^n.$$

Note that

$$r = \left|\frac{v^2 - 4v + 2}{v^2}\right| < 1 \quad \text{if } v > \frac{1}{2},$$

thus the second order IMEX scheme has the same AP property as the first order scheme (2.6). Moreover, we can prove a theorem similar to Theorem 2.3 for (2.8) but the details are omitted here.

To illustrate the efficiency of (2.6) and (2.8) in various situations, we consider a simple linear problem with different scales for which only some components rapidly converge to a steady state whereas the remaining part oscillates. We solve

$$\mathcal{Q}(f) = Af, \tag{2.9}$$

where

$$A = \begin{pmatrix} -1000 & 1 & 0 \\ -1 & -1000 & 0 \\ 0 & 0 & i \end{pmatrix} \tag{2.10}$$

for which the eigenvalues are $\mathrm{Sp}(A) = \{-1000 + i, -1000 - i, i\}$. The first block represents the fast scales whereas the last one is the oscillating part. Indeed, the first components go to zero exponentially fast whereas the third one oscillates with respect to time with a period of 2π. We want to solve accurately the oscillating part with a large time step without resolving the small scales. Then, we apply the first order (2.6) and second order (2.8) schemes by choosing

$$P(f) = \nu A f,$$

with $\nu \geqslant 0$. Here we take a large time step $\Delta t = 0.3$ and $\nu = 2$, which means that $P(f)$ has the same structure of $\mathcal{Q}(f)$ but the eigenvalues are over estimated. Thus, fast scales are under-resolved whereas this times step is a good discretization of the third oscillating component. Therefore, an efficient AP scheme would give an accurate behavior of the slow oscillating scale with large time step with respect to the fast scale.

In Fig. 1, we present the real part of the numerical solution to the differential system (2.9), (2.10) corresponding to the initial datum $f(0) = (2,1,1)$ on the time interval $[0,15]$. We compare the numerical solution obtained with our first (2.6) and second (2.8) order AP schemes using a large time step ($\Delta t = 0.3$) and the one obtained with a first and second order explicit Runge–Kutta scheme using a small time step ($\Delta t = 0.0001$) for which the numerical solution is stable.

It clearly appears in Fig. 1 (1) that the time step is too large to give accurate results for the first order scheme (2.6): the solution is stable but the oscillation of the third component is damped for this time step which is too large. This approximation is compared with the one obtained with a first order explicit Euler using a times step 300 times smaller. Thus, the first order AP scheme gives a numerical solution which is stable for large time step but the accuracy is not satisfying.

Therefore, we also compare the numerical solution of the second order scheme (2.8) with the one obtained using a second order explicit Runge–Kutta scheme corresponding to $\nu = 0$ with a time step three hundred times smaller. In Fig. 1, we observe the stability and good accuracy of the second order scheme (2.8). Let us emphasize that for the same time step, the numerical solution given by an explicit Runge–Kutta scheme blows-up (hence the result is not reported here)!

3. Hyperbolic systems with relaxations

In this section, we propose and study the method for hyperbolic system with (stiff) relaxations. We propose the following temporal approximation to (1.12):

$$\begin{cases} \frac{U^{n+1}-U^n}{\Delta t} + f_1(U^n, V^n)_x = 0, \\ \frac{V^{n+1}-V^n}{\Delta t} + f_2(U^n, V^n)_x = \frac{1}{\varepsilon}\left[R(U^n, V^n) + \beta(V^n - g(U^n))\right] - \frac{\beta}{\varepsilon}\left[V^{n+1} - g(U^{n+1})\right]. \end{cases} \tag{3.1}$$

Assume all functions are smooth. Some simple mathematical manipulations on (3.1) give

$$V^{n+1} - g(U^{n+1}) = -\left[f_2(U^n, V^n)_x + (g(U^{n+1}) - g(U^n))/\Delta t\right] \frac{\varepsilon \Delta t}{\varepsilon + \beta \Delta t} + \frac{1 + \frac{\Delta t}{\varepsilon}\left[\beta + \frac{R(U^n, V^n)}{V^n - g(U^n)}\right]}{1 + \beta \frac{\Delta t}{\varepsilon}} (V^n - g(U^n)). \tag{3.2}$$

Note that

$$\frac{R(U^n, V^n)}{V^n - g(U^n)} = \frac{R(U^n, V^n) - R(U^n, g(U^n))}{V^n - g(U^n)} = \partial_v R(U^n, W^n) < 0 \quad \text{for some } W^n,$$

thus if

$$\beta > \frac{1}{2} \sup |\partial_v R|,$$

there exists a constant C, and $0 < r < 1$ such that

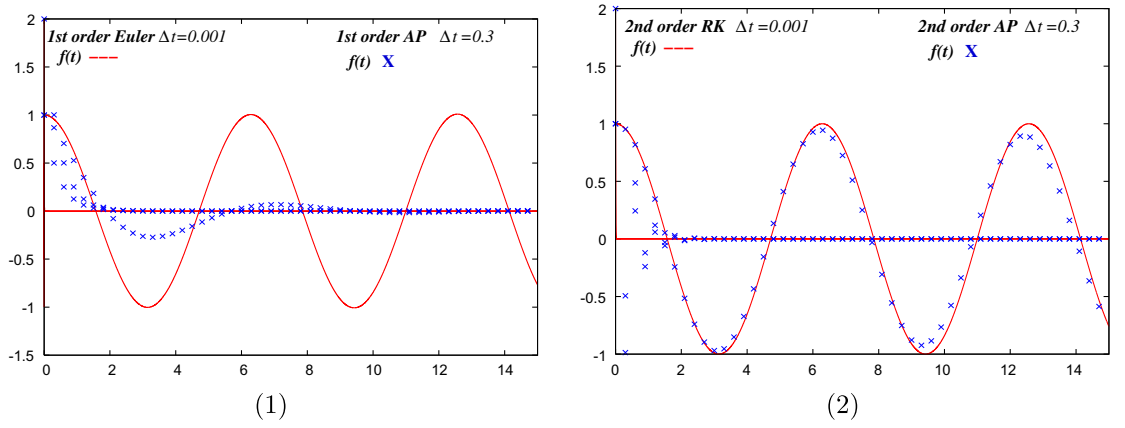


Fig. 1. Comparison of the time evolution of the numerical approximation to the differential system (2.9), (2.10) with $f(0) = (2,1,1)$. (1) first and (2) second order asymptotic-preserving and explicit Runge–Kutta schemes.

$$|V^{n+1} - g(U^{n+1})| \leqslant C\frac{\varepsilon\Delta t}{\varepsilon + \beta\Delta t} + r|V^n - g(U^n)|.$$

From here it is easy to see that

$$|V^n - g(U^n)| \leqslant \frac{C}{1-r}\frac{\varepsilon\Delta t}{\varepsilon + \beta\Delta t} + r^n|V^0 - g(U^0)|.$$

This clearly gives

$$|V^n - g(U^n)| \leqslant \frac{C}{(1-r)\beta}\varepsilon + r^n|V^0 - g(U^0)| \tag{3.3}$$

in which the first term on the right hand side is $O(\varepsilon)$ independent of Δt. For any $\varepsilon \ll 1$, there exists an $N_\varepsilon \geqslant 1$ such that

$$r^n|V^0 - g(U^0)| \leqslant \varepsilon,$$

therefore (3.3) implies the desired AP property (1.15).

Next we consider the linear penalty method (2.4):

$$\begin{cases} \frac{U^{n+1}-U^n}{\Delta t} + f_1(U^n, V^n)_x = 0, \\ \frac{V^{n+1}-V^n}{\Delta t} + f_2(U^n, V^n)_x = \frac{1}{\varepsilon}[R(U^n, V^n) + \beta V^n] - \frac{\beta}{\varepsilon}V^{n+1}. \end{cases} \tag{3.4}$$

A simple mathematical manipulation on (3.4) gives

$$V^{n+1} - g(U^{n+1}) = -f_2(U^n, V^n)_x\frac{\varepsilon\Delta t}{\varepsilon + \beta\Delta t} - \left[g(U^{n+1}) - g(U^n)\right] + \frac{1 + \frac{\Delta t}{\varepsilon}\left[\mu + \frac{R(U^n,V^n)}{V^n - g(U^n)}\right]}{1 + \mu\frac{\Delta t}{\varepsilon}}(V^n - g(U^n)). \tag{3.5}$$

The first two terms on the right hand side of (3.5) can only be bounded by $C(\varepsilon + \Delta t)$, while the third term, under the condition

$$\mu > \frac{1}{2}\sup|\partial_v R|,$$

is similar to the second term on the right hand side of (3.2). In conclusion, corresponding to (3.3), here we can only obtain

$$|V^n - g(U^n)| \leqslant C(\varepsilon + \Delta t) + r^n|V^0 - g(U^0)|, \tag{3.6}$$

which, if $\Delta t \gg \varepsilon$, gives only (1.16).

Another observation is the following. From (3.3), one sees that for prepared initial data

$$V^0 = g(U^0) + O(\varepsilon) \tag{3.7}$$

(3.3) implies that

$$V^n = g(U^n) + O(\varepsilon), \quad \text{for any } n \geqslant 1.$$

Namely, if the data are within $O(\varepsilon)$ of the local equilibrium, they remain so for all future times. However, for the linear penalty method, even if the initial data are well prepared as in (3.7), from (3.6) one sees that

$$V^1 = g(U^1) + O(\varepsilon + \Delta t),$$

so the deviation from the local equilibrium at later times is always of $O(\Delta t)$ rather than $O(\varepsilon)$. A similar analysis on method (2.3) gives a result as in (3.6). We omit the details here.

Now to illustrate the efficiency of our approach, we present numerical simulations on (1.12). We simply consider

$$\begin{cases} \frac{\partial u}{\partial t} + \frac{\partial v}{\partial x} = 0, \\ \frac{\partial v}{\partial t} + a\frac{\partial u}{\partial x} = \frac{g(v)}{\varepsilon}(f(u) - v), \\ u(t=0,x) = 1 + 0.9\sin(\pi x), \quad v(t=0,x) = \cos(\pi x), \quad x \in (-1,1), \end{cases} \tag{3.8}$$

with $g(v) = 1 + |v|^4$, $f(u) = u^2/2$ and $a = \sup_u|f'(u)|^2$. In Fig. 2, we represent the approximation of the solution at time $t = 0.1$ for different values of $\varepsilon = 10^{-1}$; 10^{-2} and 10^{-7} obtained with our AP scheme (3.1) and the linear penalty method (3.4). The number of points in space is $n_x = 800$ and $\Delta t = 0.0006$. Clearly, for the same time step, our scheme gives the correct behavior which corresponds to the well known solution to the Burgers equations when ε tends to zero, whereas the linear penalty method gives a stable approximation which is not accurate. Of course, when Δt becomes smaller and ε is fixed, the linear penalty method is accurate. Here, the initial data is far from the equilibrium hence the linear penalty method is not appropriate since it does not have any mechanism of projection to the steady state when ε is small.

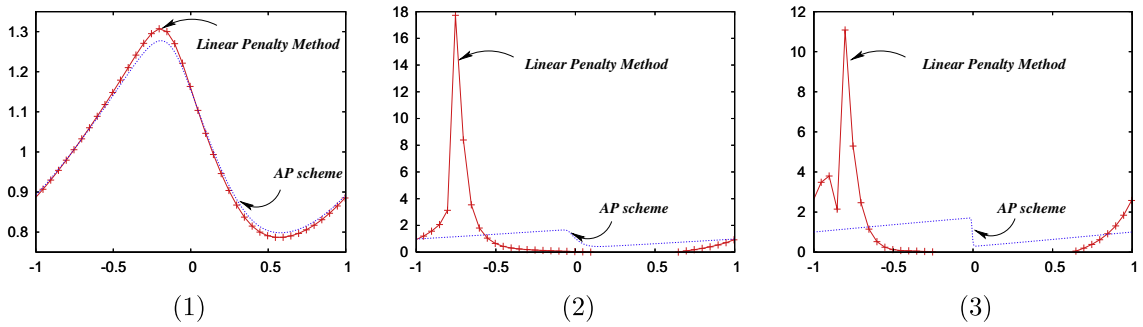


Fig. 2. Approximation of the solution to (3.8) obtained from our AP scheme (3.1) and the linear penalty method (3.4) for different values of the Knudsen number $\varepsilon = 10^{-1}$, 10^{-2} and 10^{-7}.

4. The Boltzmann equation

We now extend the method to the Boltzmann Eq. (1.1). To this aim, we rewrite the Boltzmann Eq. (1.1) in the following form

$$\begin{cases} \frac{\partial f}{\partial t} + v\nabla_x f = \frac{Q(f)-P(f)}{\varepsilon} + \frac{P(f)}{\varepsilon}, \quad x \in \Omega \subset \mathbb{R}^{d_x}, \ v \in \mathbb{R}^{d_v}, \\ f(0,x,v) = f_0(x,v), \quad x \in \Omega, \ v \in \mathbb{R}^{d_v}, \end{cases} \tag{4.1}$$

where the operator P is a "well-balanced relaxation approximation" of $Q(f)$, which means that it satisfies the following (balance law)

$$\int_{\mathbb{R}^{d_v}} P(f)\phi(v)dv = 0, \quad \phi(v) = 1, v, |v|^2,$$

and preserves the steady state *i.e.* $P(\mathcal{M}_{\rho,u,T}) = 0$ where $\mathcal{M}_{\rho,u,T}$ is the Maxwellian distribution associated to ρ, u and T given by (1.5). Moreover, it is a relaxation operator in velocity

$$P(f) = \beta\big[\mathcal{M}_{\rho,u,T}(v) - f(v)\big]. \tag{4.2}$$

4.1. Choice of the free parameter β

For instance, $P(f)$ can be computed from an expansion of the Boltzmann operator with respect to $\mathcal{M}_{\rho,u,T}$:

$$Q(f) \simeq Q(\mathcal{M}_{\rho,u,T}) + \nabla Q(\mathcal{M}_{\rho,u,T})\big[\mathcal{M}_{\rho,u,T} - f\big].$$

Thus, we choose $\beta > 0$ as an upper bound of the operator $\nabla Q(\mathcal{M}_{\rho,u,T})$. Other choices of β are also possible, for example

$$\beta = \sup\left|\frac{Q(f) - Q(\mathcal{M})}{f - \mathcal{M}}\right| = \sup\left|\frac{Q(f)}{f - \mathcal{M}}\right|,$$

or, at time t^n,

$$\beta^n = \sup\left|\frac{Q(f^n) - Q(f^{n-1})}{f^n - f^{n-1}}\right|.$$

Then $P(f)$ given by (4.2) is just the BGK collisional operator [3].

One can also choose β such that the operator $P(f)$ gives the same viscosity (of order to ε) as $Q(f)$ when applying a Chapman–Enskog expansion.

4.2. Discretization to the Boltzmann equation

Since the convection term in (4.1) is not stiff, we will treat it explicitly. The source terms on the right hand side of (4.1) will be handled using the ODE solver in the previous section. For example, if the first order scheme (2.6) is used, then we have

$$\begin{cases} \frac{f^{n+1}-f^n}{\Delta t} + v \cdot \nabla_x f^n = \frac{Q(f^n)-P(f^n)}{\varepsilon} + \frac{P(f^{n+1})}{\varepsilon}, \\ f^0(x,v) = f_0(x,v). \end{cases} \tag{4.3}$$

Using the relaxation structure of $P(f)$ given in (4.2), it can be written as

$$f^{n+1} = \frac{\varepsilon}{\varepsilon + \beta^{n+1}\Delta t}[f^n - \Delta t\, v\nabla_x f^n] + \Delta t \frac{\mathcal{Q}(f^n) - P(f^n)}{\varepsilon + \beta^{n+1}\Delta t} + \frac{\beta^{n+1}\Delta t}{\varepsilon + \beta^{n+1}\Delta t}\mathcal{M}^{n+1}, \tag{4.4}$$

where $\beta^{n+1} = \beta(\rho^{n+1}, T^{n+1})$ and $\mathcal{M}^{n+1}$ is the Maxwellian distribution computed from f^{n+1}.

Although (4.4) appears nonlinearly implicit, it can be computed explicitly. Specifically, upon multiplying (4.4) by $\phi(v)$ defined in (1.4), and use the conservation property of $\mathcal{Q}$ and P and the definition of $\mathcal{M}$ in (1.5), we define the macroscopic quantity U by $U := (\rho, \rho u, T)$ computed from f and get [14,53]

$$U^{n+1} = \frac{\varepsilon}{\varepsilon + \beta^{n+1}\Delta t}\int \phi(v)(f^n - \Delta t\, v\cdot\nabla_x f^n)dv + \frac{\beta^{n+1}\Delta t}{\varepsilon + \beta^{n+1}\Delta t}U^{n+1},$$

or simply

$$U^{n+1} = \int \phi(v)(f^n - \Delta t\, v\cdot\nabla_x f^n)dv.$$

Thus U^{n+1} can be obtained explicitly, which defines $\mathcal{M}^{n+1}$. Now f^{n+1} can be obtained from (4.4) explicitly. In summary, although (4.3) is nonlinearly implicit, it can be solved *explicitly*, thus satisfies the second condition of an AP scheme.

Clearly, scheme (4.3) satisfies the following properties.

Proposition 4.1. *Consider the numerical solution given by (4.3) . Then,*

(i) *If $\varepsilon \to 0$ and $f^n = \mathcal{M}^n + O(\varepsilon)$, then $f^{n+1} = \mathcal{M}^{n+1} + O(\varepsilon)$. Thus, when $\varepsilon \to 0$, the (moments of the) scheme becomes a consistent discretization of the Euler system (1.7).*

(ii) *Assume $\varepsilon \ll 1$ and $f^n = \mathcal{M}^n + \varepsilon g^n$. If there exists a constant $C > 0$ such that*

$$\left\|\frac{g^{n+1} - g^n}{\Delta t}\right\| + \|\nabla_x(vg^n)\| + \|g^n\| \leqslant C, \tag{4.5}$$

and

$$\|U^n\| + \left\|\frac{U^{n+1} - U^n}{\Delta t}\right\| \leqslant C, \tag{4.6}$$

then the scheme (4.3) asymptotically becomes a first order in time approximation of the compressible Navier–Stokes (1.8).

Proof. We easily first check that for $\varepsilon \to 0$ and $f^n = \mathcal{M}^n + O(\varepsilon)$, we get $f^{n+1} = \mathcal{M}^{n+1} + O(\varepsilon)$. Therefore, we multiply (4.3) by $(1, v, -v^2/2)$ and integrate with respect to v, which yields that U^n is given by a time explicit scheme of the Euler system (1.7).

Now let us prove (ii). We apply the classical Chapman–Enskog expansion:

$$f^n = \mathcal{M}^n + \varepsilon g^n \tag{4.7}$$

and integrate (4.3) with respect to $v \in \mathbb{R}^{d_v}$. By using the conservation properties of the Boltzmann operator (1.4) and of the well-balanced approximation $P(f)$,

$$\frac{U^{n+1} - U^n}{\Delta t} + \nabla_v \cdot \int_{\mathbb{R}^{d_v}} \begin{pmatrix} 1 \\ v \\ \frac{|v|^2}{2} \end{pmatrix} v(\mathcal{M}^n + \varepsilon g^n)dv = 0. \tag{4.8}$$

For $\varepsilon g = 0$, this is the compressible Euler equations (1.7). Thus, a consistent approximation of the compressible Navier–Stokes is directly related to a consistent approximation of g^n. Inserting decomposition (4.7) into the scheme (4.3) gives

$$\frac{\mathcal{M}^{n+1} - \mathcal{M}^n}{\Delta t} + v\nabla_x\mathcal{M}^n + \varepsilon\left(\frac{g^{n+1} - g^n}{\Delta t} + v\nabla_x g^n\right) = \frac{\mathcal{Q}(\mathcal{M}^n + \varepsilon g^n)}{\varepsilon} - \left[\beta^n g^n - \beta^{n+1}g^{n+1}\right],$$

Since $\mathcal{Q}$ is bilinear and $\mathcal{Q}(\mathcal{M}) = 0$, one has

$$\mathcal{Q}(\mathcal{M} + \varepsilon g) = \mathcal{Q}(\mathcal{M}) + \varepsilon\mathcal{L}_{\mathcal{M}}(g) + \varepsilon^2\mathcal{Q}(g),$$

where $\mathcal{L}_{\mathcal{M}}$ is the linearized collision operator with respect to $\mathcal{M}$. Thus, we get

$$\frac{\mathcal{M}^{n+1} - \mathcal{M}^n}{\Delta t} - \left[\beta^n g^n - \beta^{n+1}g^{n+1}\right] + \varepsilon\left[\frac{g^{n+1} - g^n}{\Delta t} + v\nabla_x g^n - \mathcal{Q}(g^n)\right] = \mathcal{L}_{\mathcal{M}}(g^n) - v\nabla_x\mathcal{M}^n, \tag{4.9}$$

It is well known that $\mathcal{L}_{\mathcal{M}}$ is a non-positive self-adjoint operator on $L^2_{\mathcal{M}}$ defined by the set

$$L^2_{\mathcal{M}} := \{\varphi : \quad \varphi\mathcal{M}^{-1/2} \in L^2(\mathbb{R}^{d_v})\}$$

and that its kernel is $\mathcal{N}(\mathcal{L}_{\mathcal{M}}) = \mathrm{Span}\{\mathcal{M}, v\mathcal{M}, |v|^2\mathcal{M}\}$. Let $\Pi_{\mathcal{M}}$ be the orthogonal projection in $L^2_{\mathcal{M}}$ onto $\mathcal{N}(\mathcal{L}_{\mathcal{M}})$. After easy computations in the orthogonal basis, one finds that [5]

$$\Pi_{\mathcal{M}}(\psi) = \frac{\mathcal{M}}{\rho}\left[m_0 + \frac{v-u}{T}m_1 + \left(\frac{|v-u|^2}{2T} - \frac{d_v}{2}\right)m_2\right],$$

where

$$m_0 = \int_{\mathbb{R}^{d_v}} \psi dv, \quad m_1 = \int_{\mathbb{R}^{d_v}} (v-u)\psi dv, \quad m_2 = \int_{\mathbb{R}^{d_v}} \left(\frac{|v-u|^2}{2T} - \frac{d_v}{2}\right)\psi dv.$$

It is easy to verify that $\Pi_{\mathcal{M}^n}(\mathcal{M}^n) = \mathcal{M}^n$ and

$$\Pi_{\mathcal{M}^n}(g^n) = \Pi_{\mathcal{M}^n}(g^{n+1}) = \Pi_{\mathcal{M}^n}(\mathcal{Q}(g^n)) = \Pi_{\mathcal{M}^n}(\mathcal{L}_{\mathcal{M}^n}(g^n)) = 0.$$

Then applying the orthogonal projection $\mathrm{I} - \Pi_{\mathcal{M}^n}$ to (4.9), it yields

$$(\mathrm{I}-\Pi_{\mathcal{M}^n})\left(\frac{\mathcal{M}^{n+1}-\mathcal{M}^n}{\Delta t}\right) - \left(\beta^n g^n - \beta^{n+1}g^{n+1}\right) + \varepsilon\left[\frac{g^{n+1}-g^n}{\Delta t} + (\mathrm{I}-\Pi_{\mathcal{M}^n})(v\nabla_x g^n) - \mathcal{Q}(g^n)\right] = \mathcal{L}_{\mathcal{M}}(g^n) - (\mathrm{I}-\Pi_{\mathcal{M}^n})(v\nabla_x \mathcal{M}^n).$$

Using the assumption (4.5) we get that the term

$$\varepsilon\left[\frac{g^{n+1}-g^n}{\Delta t} + (\mathrm{I}-\Pi_{\mathcal{M}^n})(v\nabla_x g^n) - \mathcal{Q}(g^n)\right]$$

is of order ε. Then, it remains to estimate the terms $\beta^{n+1}g^{n+1} - \beta^n g^n$ and

$$(\mathrm{I}-\Pi_{\mathcal{M}^n})\left(\frac{\mathcal{M}^{n+1}-\mathcal{M}^n}{\Delta t}\right).$$

First, we have

$$\beta^{n+1}g^{n+1} - \beta^n g^n = \beta^{n+1}(g^{n+1}-g^n) + (\beta^{n+1}-\beta^n)g^n.$$

Under the assumption (4.5) and (4.6), and since β^n only depends on U^n, we easily get

$$\beta^{n+1}g^{n+1} - \beta^n g^n = O(\Delta t). \tag{4.10}$$

Next using a Taylor expansion we find that

$$\mathcal{M}^{n+1} = \mathcal{M}^n\left[1 + \frac{\rho^{n+1}-\rho^n}{\rho^n} + \frac{v-u^n}{T^n}(u^{n+1}-u^n) + \left(\frac{|v-u^n|^2}{2T^n} - \frac{d}{2}\right)\frac{T^{n+1}-T^n}{T^n}\right] + O(\Delta t^2)$$

and by definition of $\Pi_{\mathcal{M}}$

$$\begin{aligned}\Pi_{\mathcal{M}^n}(\mathcal{M}^{n+1}) = {}& \mathcal{M}^n\left(1 + \frac{\rho^{n+1}-\rho^n}{\rho^n} + \frac{v-u^n}{T^n}(u^{n+1}-u^n) + \left(\frac{|v-u^n|^2}{2T^n} - \frac{d}{2}\right)\frac{T^{n+1}-T^n}{T^n}\right)\\ &+ \mathcal{M}^n\left(\frac{|v-u^n|^2}{2T^n} - \frac{d}{2}\right)\left[\frac{T^{n+1}-T^n}{\rho^n T^n}(\rho^{n+1}-\rho^n) + \frac{\rho^{n+1}}{d\rho^n T^n}(u^{n+1}-u^n)^2\right]\\ &+ \mathcal{M}^n\frac{v-u^n}{T^n}\frac{\rho^{n+1}-\rho^n}{\rho^n}(u^{n+1}-u^n) + O(\Delta t^2).\end{aligned}$$

Thus, under assumption (4.6), we have

$$(\mathrm{I}-\Pi_{\mathcal{M}^n})\left(\frac{\mathcal{M}^{n+1}-\mathcal{M}^n}{\Delta t}\right) = O(\Delta t). \tag{4.11}$$

Gathering (4.10) and (4.11), the residual distribution function is given by

$$g^n = \mathcal{L}^{-1}_{\mathcal{M}^n}((\mathrm{I}-\Pi_{\mathcal{M}^n})(v\cdot\nabla_x\mathcal{M}^n)) + O(\varepsilon) + O(\Delta t).$$

Now, substituting this latter expression in (4.8), we get

$$\frac{U^{n+1}-U^n}{\Delta t} + \nabla_x\cdot F(U) = -\varepsilon\nabla_x\cdot\int_{\mathbb{R}^{d_v}}\begin{pmatrix} v \\ v\otimes v \\ v\frac{|v|^2}{2}\end{pmatrix}\mathcal{L}^{-1}_{\mathcal{M}^n}((\mathrm{Id}-\Pi_{\mathcal{M}^n})(v\cdot\nabla_x\mathcal{M}^n))dv \tag{4.12}$$

$$+ O(\varepsilon\Delta t + \varepsilon^2), \tag{4.13}$$

where

$$F(U) = \begin{pmatrix} \rho u \\ \rho u \otimes u + p\mathrm{I} \\ (E+p)u \end{pmatrix}.$$

To complete the proof, it remains to compute the term in $O(\varepsilon)$. After some computations, we first get

$$(\mathrm{I} - \Pi_{\mathcal{M}^n})(v \cdot \nabla_x \mathcal{M}^n) = \left[B\left(\nabla u + (\nabla u)^T - \frac{d}{2} \nabla \cdot u \mathrm{I} \right) + A \frac{\nabla T}{\sqrt{T}} \right] \mathcal{M}(v),$$

with

$$A = \left(\frac{|v-u|^2}{2T} - \frac{d+2}{2} \right) \frac{v-u}{\sqrt{T}}, \quad B = \frac{1}{2} \left(\frac{(v-u) \otimes (v-u)}{2T} - \frac{|v-u|^2}{dT} \mathrm{I} \right).$$

Therefore, it yields

$$\mathcal{L}^{-1}_{\mathcal{M}^n}((\mathrm{I} - \Pi_{\mathcal{M}^n})(v \cdot \nabla_x \mathcal{M}^n)) = \mathcal{L}^{-1}_{\mathcal{M}^n}(B\mathcal{M}) \left(\nabla u + (\nabla u)^T - \frac{d}{2} \nabla \cdot u \mathrm{I} \right) + \mathcal{L}^{-1}_{\mathcal{M}^n}(A\mathcal{M}) \frac{\nabla T}{\sqrt{T}}.$$

Substituting this expression in (4.8), we get a consistent time discretization scheme to the compressible Navier–Stokes system where the term of order of ε is given by

$$\varepsilon \nabla_x \cdot \begin{pmatrix} 0 \\ \mu_\varepsilon \sigma(u_\varepsilon) \\ \mu_\varepsilon \sigma(u_\varepsilon) u + \kappa_\varepsilon \nabla_x T_\varepsilon \end{pmatrix},$$

with

$$\sigma(u) = \nabla_x u + (\nabla_x u)^T - \frac{2}{d_v} \nabla_x \cdot u \mathrm{I},$$

while the viscosity $\mu_\varepsilon = \mu(T_\varepsilon)$ and the thermal conductivity $\kappa_\varepsilon = \kappa(T_\varepsilon)$ are defined according to the linearized Boltzmann operator with respect to the local Maxwellian [1]. □

At this stage, let us address several comments concerning Proposition 4.1.

- Note that the assumption (4.5) is very difficult to prove for our scheme. However, a similar assumption is done in [2].
- We only prove theoretical results when the initial data is close enough to the local Maxwellian. However, it is expected that, as it is shown in Section 3, after the initial transient time the solution is only $O(\varepsilon)$ distance from the local equilibrium. While this has not been proven for the Boltzmann equation (since it does not have a property similar to (1.13), our numerical results in Section 4 strongly suggest so. A rigorous proof remains an open question.
- Under-resolved computations using AP schemes can only capture the solutions of the Euler equations. To capture the Navier–Stokes approximation that has $O(\varepsilon)$ viscosity and heat conductivity, one needs the mesh size and $c\Delta t$ to be $o(\varepsilon)$ (c is a characteristic speed). Thus conclusion (ii) in the above proposition shows that the scheme is *consistent* to the Navier–Stokes equations provided that the viscous terms are resolved. In other words, one cannot expect to capture the Navier–Stokes solution with under-resolved ($\frac{\Delta x}{c}, \Delta t \gg \varepsilon$) mesh sizes and time steps. On the other hand, if one has to resolve the viscous term using $\frac{\Delta x}{c}, \Delta t = o(\varepsilon)$ it will be more efficient to directly solve the Boltzmann equation directly. Thus we do not advocate an AP scheme for the compressible Navier–Stokes limit. Nevertheless, the result of Proposition 4.1 (ii) is still analytically interesting. If one directly compares the error of numerical solutions f with the solution of the Boltzmann equation by, say a first order method, one usually arrives at an error of $O(\Delta t/\varepsilon)$ (see a related study in [31]), but if compared with the solutions of the Navier–Stokes equation, which are moments of f, (4.12) shows that the error is of order $O(\Delta t + \varepsilon \Delta t)$. Here $O(\Delta t)$ comes from the Euler time discretization of U_t.

4.3. Second order IMEX scheme for the Boltzmann equation

In the following section, which is devoted to numerical simulations to the Boltzmann equation, we also have implemented a second order IMEX scheme:

$$\begin{cases} f^\star = \frac{\varepsilon}{\varepsilon + \beta^\star \Delta t} [f^n - \Delta t v \nabla_x f^n] + \Delta t \frac{Q(f^n) - P(f^n)}{\varepsilon + \beta^\star \Delta t} + \frac{\beta^\star \Delta t}{\varepsilon + \beta^\star \Delta t} \mathcal{M}^\star, \\ f^{n+1} = \frac{\varepsilon}{\varepsilon + \beta^{n+1} \Delta t/2} [f^n - \Delta t v \nabla_x f^\star] + \Delta t \frac{Q(f^\star) - P(f^\star)}{\varepsilon + \beta^{n+1} \Delta t/2} + \frac{\Delta t}{2\varepsilon + \beta^{n+1} \Delta t} \left(\beta^{n+1} \mathcal{M}^{n+1} + \beta^n (\mathcal{M}^n - f^n) \right), \end{cases} \tag{4.14}$$

where $\beta^\star = \beta(\rho^\star, T^\star)$ and $\mathcal{M}^\star$ is the Maxwellian distribution computed from $f^\star$.

4.4. Space discretization

On the one hand, the approximation of the Boltzmann operator is performed by a fast spectral Fourier-Galerkin method already proposed in [27]. On the other hand, the approximation in space is achived using a second order finite volume scheme. Let $(x_{i+1/2})_{i\in I}$ a set of points of the space domain and I a bounded set of integers, hence for $\Delta x = x_{i+1/2} - x_{i-1/2}$

$$\int_{x_{i-1/2}}^{x_{i+1/2}} v_x \frac{\partial f}{\partial x} dx = \frac{\mathcal{F}_{i+1/2} - \mathcal{F}_{i-1/2}}{\Delta x},$$

where $\mathcal{F}_{i+1/2} = v_x^+ f^l_{i+1/2} - v_x^- f^r_{i+1/2}$ and $v_x^+ = \max(v_x, 0)$, $v_x^- = \max(-v_x, 0)$,

$$f^l_{i+1/2} = f_i + \frac{\delta f_{i+1/2}}{2}, \quad f^r_{i+1/2} = f_{i+1} - \frac{\delta f_{i+1/2}}{2}.$$

and δf represents a slope with a slope limiter (see for instance [43]).

5. Numerical tests

In this section we perform several numerical simulations for the Boltzmann equation in different asymptotic regimes in order to check the performance (in stability and accuracy) of our methods. We have implemented the first order (2.6) and second order (2.8) scheme for the approximation of the Boltzmann equation. Here, the Boltzmann collision operator is discretized by a deterministic method [22–25,27,52], which gives a spectrally accurate approximation. A classical second order finite volume scheme with slope limiters is applied for the transport operator as sdescribed in Section 4.4.

For all numerical simulations, we have considered Maxwellian molecules, that is $\alpha = 0$ in (1.3). Hence, we take $\beta = 2\pi\rho$ such that both operators $P(f)$ and the full Boltzmann operator $Q(f)$ have the same loss term corresponding to the dissipative part.

5.1. Approximation of smooth solutions

This test is used to evaluate the order of accuracy of our new methods. More precisely, we want to show that our methods (4.4) and (4.14) are uniformly accurate with respect to the parameter $\varepsilon > 0$. We consider the Boltzmann Eq. (1.1) in $1d_x \times 2d_v$. We take a smooth initial data

$$f_0(x, v) = \frac{\rho_0(x)}{2\pi T_0(x)} \exp\left(-\frac{|v|^2}{2T_0(x)}\right), \quad (x, v) \in [-L, L] \times \mathbb{R}^2,$$

with $\rho_0(x) = (11 - 9\tanh(x))/10$, $T_0(x) = (3 - \tanh(x))/4$, $L = 1$ and assume specular reflection boundary conditions in x. Numerical solutions are computed from different phase space meshes: the number of point in space is $n_x = 50, 100, 200, \ldots, 1600$ and the number of points in velocity is n_v^2 with $n_v = 8, \ldots, 64$ (for which the spectral accuracy is achieved), the time step is computed such that the CFL condition for the transport is satisfied $\Delta t \leqslant \Delta x / v_{\max}$, where Δx is the space step and $v_{\max} = 7$ is the truncation of the velocity domain. Then different values of ε are considered starting from the fully kinetic regime $\varepsilon = 1$, up to the fluid limit $\varepsilon = 10^{-5}$ corresponding to the solution of the Euler system (1.7). The final time is $T_{\max} = 1$ such that the solution is smooth for the different regimes.

An estimation of the relative error in L^p norm is given by

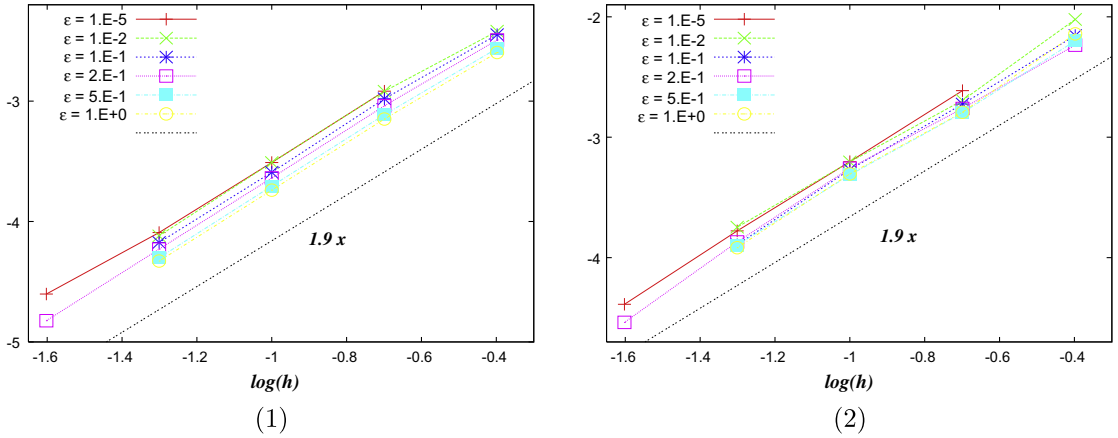


Fig. 3. The L^1 and L^∞ errors of the second order method (4.14) for different values of the Knudsen number $\varepsilon = 10^{-5}, \ldots, 1$.

$$e_{2h} = \max_{t\in(0,T)} \left(\frac{\|f_h(t) - f_{2h}(t)\|_p}{\|f_0\|_p} \right), \quad 1 \leqslant p \leqslant +\infty,$$

where f_h represents the approximation computed from a grid of order h. The numerical scheme is said to be kth order if $e_{2h} \leqslant Ch^k$, for all $0 < h \ll 1$.

In Fig. 3, the L^1 and L^∞ errors of the second order method (4.14) are presented. They show a uniformly second order convergence rate (an estimation of the slope is 1.9) in space and time (the velocity discretization is spectrally accuracy in v thus does not contribute much to the errors). The time step is not constrained by the value of ε, showing a uniform stability in time.

5.2. The Riemann problem

This test deals with the numerical solution to the $1d_x \times 2d_v$ Boltzmann equation for Maxwellian molecules ($\gamma = 0$). We present numerical simulations for one dimensional Riemann problem and compute an approximation for different Knudsen numbers, from rarefied regime to the fluid regime.

Here, the initial data corresponding to the Boltzmann equations are given by the Maxwellian distributions computed from the following macroscopic quantities

$$\begin{cases} (\rho_l, u_l, T_l) = (1, 0, 1), & \text{if } 0 \leqslant x \leqslant 0.5, \\ (\rho_r, u_r, T_r) = (0.125, 0, 0.25), & \text{if} 0.5 < x \leqslant 1. \end{cases}$$

We perform several computations for $\varepsilon = 1, 10^{-1}, 10^{-2}, \ldots, 10^{-4}$. In Fig. 4, we only show the results obtained in the kinetic regime (10^{-2}) using a spectral scheme for the discretization of the collision operator [27] (with $n_v = 32^2$ and a truncation of the velocity domain $v_{\max} = 7$) and second order explicit Runge–Kutta and second order method (4.14) for the time discretization with a time step $\Delta t = 0.005$ satisfying the CFL condition for the transport part (with $n_x = 100$). For such a value of ε, the problem is not stiff and this test is only performed to compare the accuracy of our second order scheme (4.14) with the classical (second order) Runge–Kutta method. We present several snapshots of the density, mean velocity, temperature and heat flux

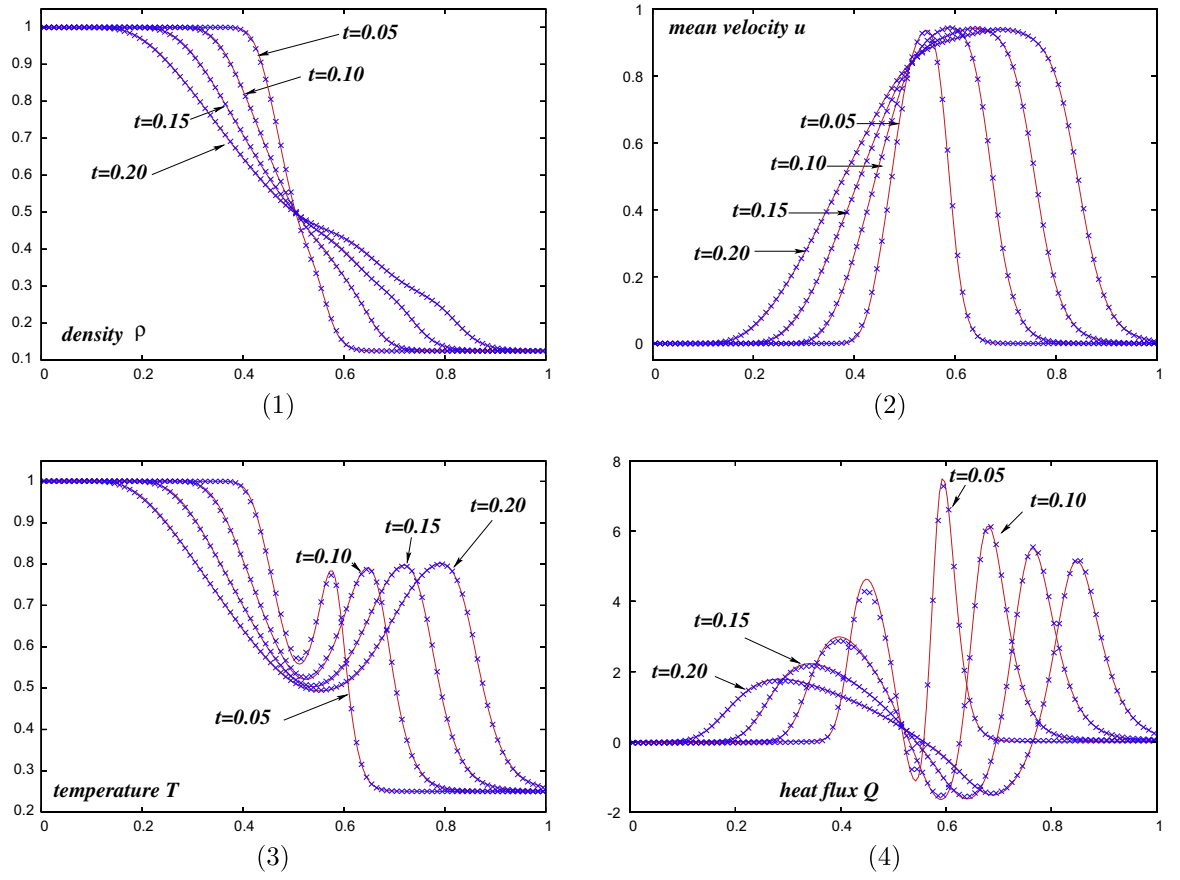


Fig. 4. Riemann problem ($\varepsilon = 10^{-2}$), crosses (x) represent the numerical solution obtained with our second order method (4.14) and lines with the explicit Runge–Kutta method: evolution of (1) the density ρ, (2) mean velocity u, (3) temperature T and (4) heat flux $\mathbb{Q}$ at time $t = 0.05$, 0.1, 0.15 and 0.2.

$$\mathbb{Q}(t,x) := \frac{1}{\varepsilon}\int_{\mathbb{R}^{d_v}} (v-u_\varepsilon)|v-u_\varepsilon|^2 f_\varepsilon(t,x,v)dv$$

at different time $t = 0.10$ and 0.20. Both results agree well with only $n_x = 100$ in the space domain and $n_v = 32$ for the velocity space. Thus, in the kinetic regime our second order method (4.14) gives the same accuracy as a second order fully explicit scheme without any additional computational effort.

Now, we investigate the cases of small values of ε for which an explicit scheme requires the time step to be of order $O(\varepsilon)$. In order to evaluate the accuracy of our method (4.14) in the Navier–Stokes regime (for small $\varepsilon \ll 1$ but not negligible), we compared the numerical solution for $\varepsilon = 10^{-3}$ with one obtained with a small time step $\Delta t = O(\varepsilon)$ (for which the computation is still feasible). Note that a direct comparison with the numerical solution to the compressible Navier–Stokes system (1.8) is difficult since the viscosity $\mu_\varepsilon = \mu(T_\varepsilon)$ and the thermal conductivity $\kappa_\varepsilon = \kappa(T_\varepsilon)$ are not explicitly known. Therefore, in Fig. 5, we report the numerical results for $\varepsilon = 10^{-3}$ and propose a comparison between the numerical solution obtained with the scheme (4.14) and the one obtained with a second order explicit Runge–Kutta method. In this case, the behavior of macroscopic quantities (density, mean velocity, temperature and heat flux) agree very well even if the time step is at least ten times larger with our method (4.4) or (4.14).

Then in Fig. 6, we compare the numerical solution of the Boltzmann Eq. (1.1) with the numerical solution to the compressible Navier–Stokes system derived from the BGK model since the viscosity and heat conductivity are in that case explicitly known [2]. To approximate the compressible Navier–Stokes system, we apply a second order Lax–Friedrich scheme using a large number of points ($n_x = 1000$) whereas we only used $n_x = 100$, and 200 points in space and $n_v^2 = 32^2$ points in velocity for the approximation of the kinetic Eq. (1.1). In this problem, the density, mean velocity and temperature are relatively close to the one obtained with the approximation of the Navier–Stokes system. Even the qualitative behavior of the heat flux agrees well with the heat flux corresponding to the compressible Navier–Stokes system $\kappa_\varepsilon \nabla_x T_\varepsilon$, with $\kappa_\varepsilon = \rho_\varepsilon T_\varepsilon$ (see Fig. 6), yet some differences can be observed, which means that the use of BGK models to derive macroscopic models has a strong influence on the heat flux.

Finally in Fig. 7, we present a comparison to the numerical solution obtained with our AP scheme for a very small value of $\varepsilon = 10^{-8}$ with the numerical solution to the Euler system. The agreement on the density, mean velocity and temperature is very satisfying with only $n_x = 100$ in the space domain for the solution to the kinetic model.

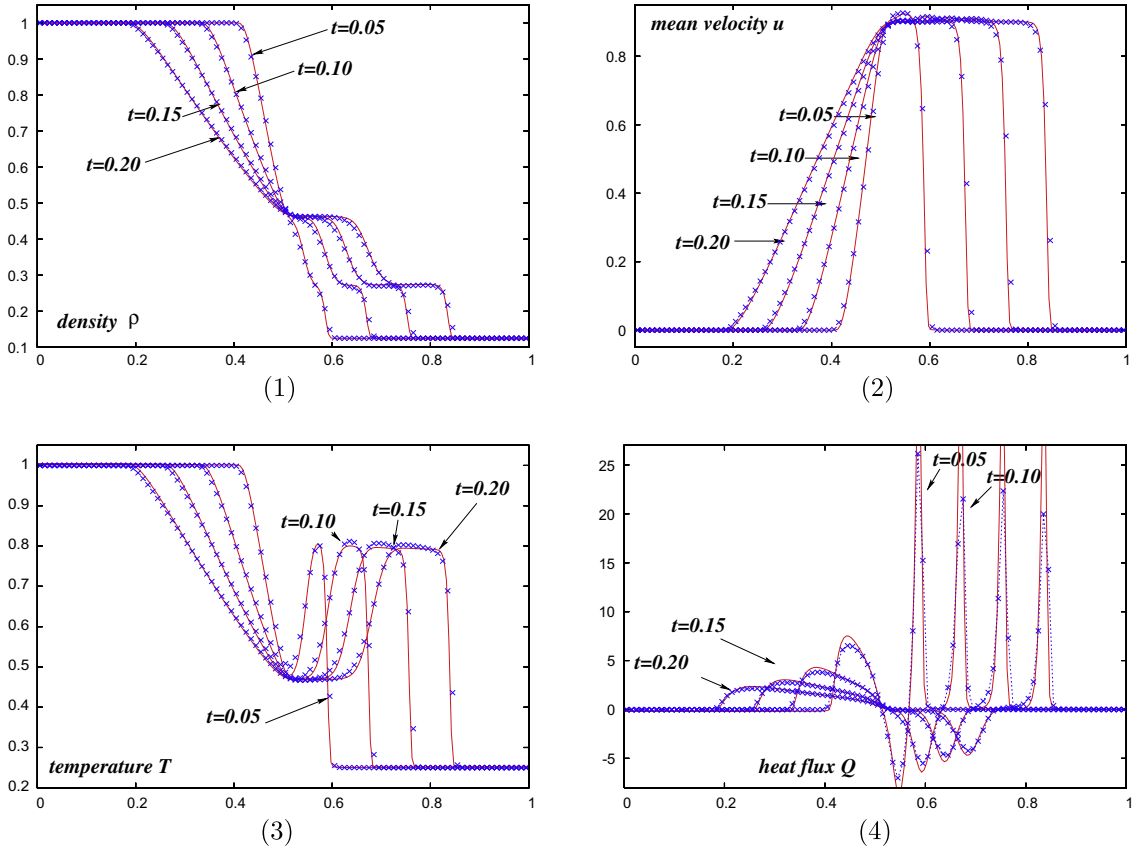


Fig. 5. Riemann problem ($\varepsilon = 10^{-3}$), crosses (x) represent the numerical solution obtained with our second order method (4.14) and lines with the explicit Runge–Kutta method: evolution of (1) the density ρ, (2) mean velocity u, (3) temperature T and (4) heat flux $\mathbb{Q}$ at time $t = 0.05$, 0.1, 0.15 and 0.2.

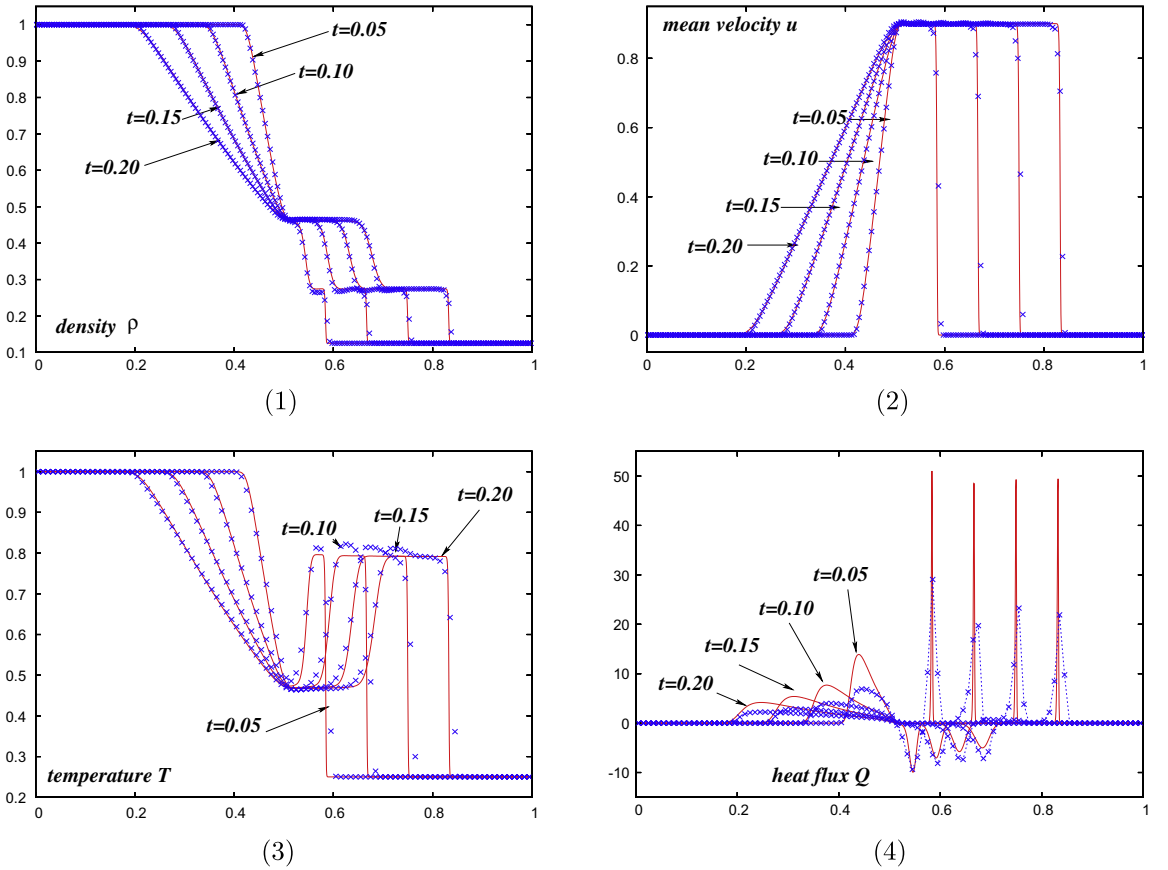


Fig. 6. Riemann problem ($\varepsilon = 10^{-4}$), comparison between the numerical solution to the Boltzmann equation with our second order method (4.14) represented with crosses (x) and the numerical solution to the compressible Navier–Stokes system (lines): evolution of (1) the density ρ, (2) mean velocity u, (3) temperature T and (4) heat flux $\mathbb{Q}$ at time $t = 0.05$, 0.1, 0.15 and 0.2.

5.3. A problem with mixing regimes

Now we consider the Boltzmann Eq. (1.1) with the Knudsen number $\varepsilon > 0$ depending on the space variable in a wide range of mixing scales.

This kind of problem was already studied by several authors for the BGK model [20] or the radiative transfer equation [42]. In this problem, $\varepsilon : \mathbb{R} \mapsto \mathbb{R}^+$ is given by

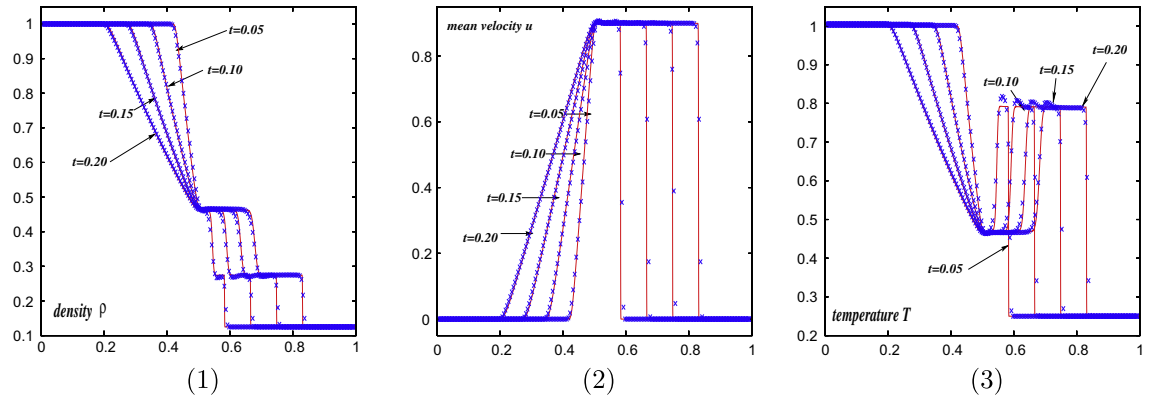
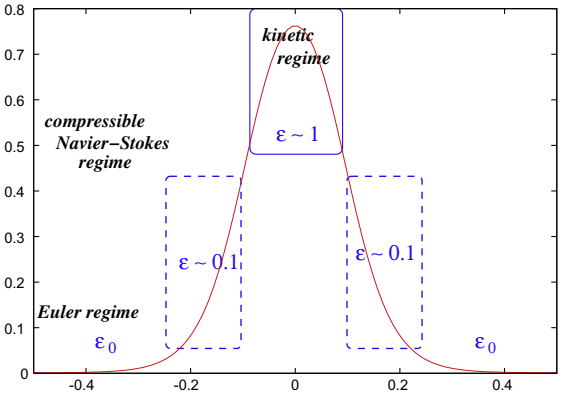


Fig. 7. Riemann problem ($\varepsilon = 10^{-8}$), comparison with the solution to the Euler system: evolution of (1) the density ρ, (2) mean velocity u, and (3) temperature T at time $t = 0.05$, 0.1, 0.15 and 0.2.

$$\varepsilon(x) = \varepsilon_0 + \frac{1}{2}[\tanh(1 - 11x) + \tanh(1 + 11x)],$$

which varies smoothly from ε_0 to $O(1)$.



This numerical test is difficult because different scales are involved. It requires a good accuracy of the numerical scheme for all range of ε. In order to focus on the multi-scale nature we only consider periodic boundary conditions, even if the method has also been used with specular reflection in space. Furthermore, to increase the difficulty we consider an initial data which is far from the local equilibrium of the collision operator:

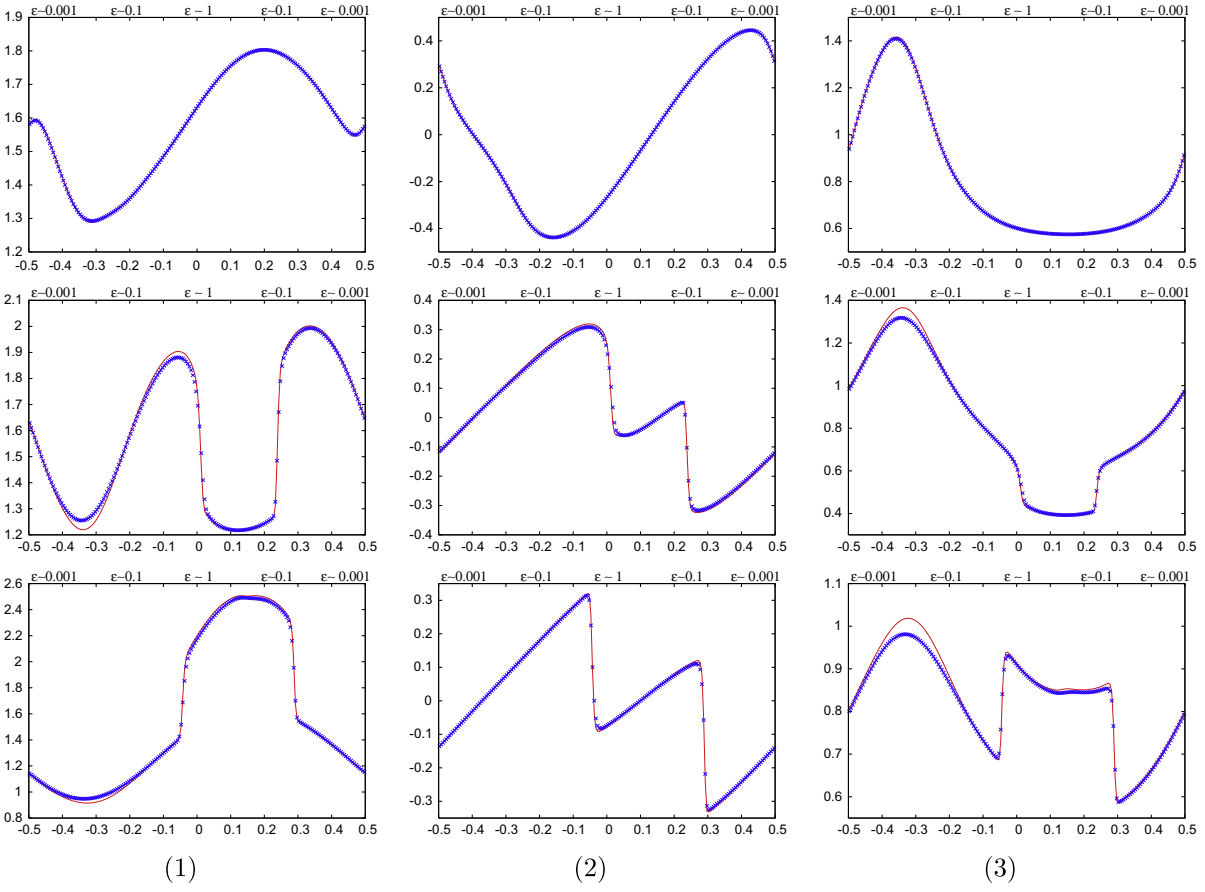


Fig. 8. Mixing regime problem ($\varepsilon_0 = 10^{-3}$), comparison of the numerical solution to the Boltzmann equation with the second order method (4.14) represented with crosses (x) with the numerical solution obtained with the explicit Runge–Kutta method using a small time step (line): evolution of (1) the density ρ, (2) mean velocity u, (3) temperature T at time $t = 0.25$, 0.5 and 0.75.

$$f_0(x,v) = \frac{\rho_0}{2}\left[\exp\left(-\frac{|v-u_0|^2}{T}\right) + \exp\left(-\frac{|v+u_0|^2}{T_0}\right)\right], \quad x \in [-L,L],\ v \in \mathbb{R}^2,$$

with $u_0 = (3/4, -3/4)$,

$$\rho_0(x) = \frac{2+\sin(kx)}{2}, \quad T_0(x) = \frac{5+2\cos(kx)}{20},$$

where $k = \pi/L$ and $L = 1/2$.

Here we cannot compare the numerical solution with the one obtained by a macroscopic model. From the numerical simulations, we observe that the solution is smooth during a short time and some discontinuities are formed in the region where the Knudsen number ε is very small and then propagate into the physical domain.

On the one hand, we only take $\varepsilon_0 = 10^{-3}$ in order to propose a comparison of numerical solutions computed with a second order method using a time step $\Delta t = 0.001$ (such that the CFL condition for the transport part is satisfied) and the one by the second order explicit Runge–Kutta method with a smaller time step $\Delta t = 0.0001$) to get stability. The number of points in space is $n_x = 200$ and in velocity is $n_v^2 = 32^2$. Clearly, in Fig. 8, the results are in good agreement even if our new method does not solve accurately small time scales when the solution is far from the local equilibrium. Moreover in Fig. 9, we present numerical results with only $n_x = 50$ and $n_x = 200$, and $n_v^2 = 32^2$ to show the performance of the method with a small number of discretization points in space. With $n_x = 50$ points the qualitative behavior of the macroscopic quantities (ρ, u, T) is fairly good.

On the other hand, we have performed different numerical results when $\varepsilon_0 = 10^{-4}$, then the variations of ε starts from 10^{-4} to *1* in the space domain. In that case, the computational time of a fully explicit scheme would be more than one hundred times larger than the one required for the asymptotic-preserving scheme (4.14). We observe that discontinuities appear

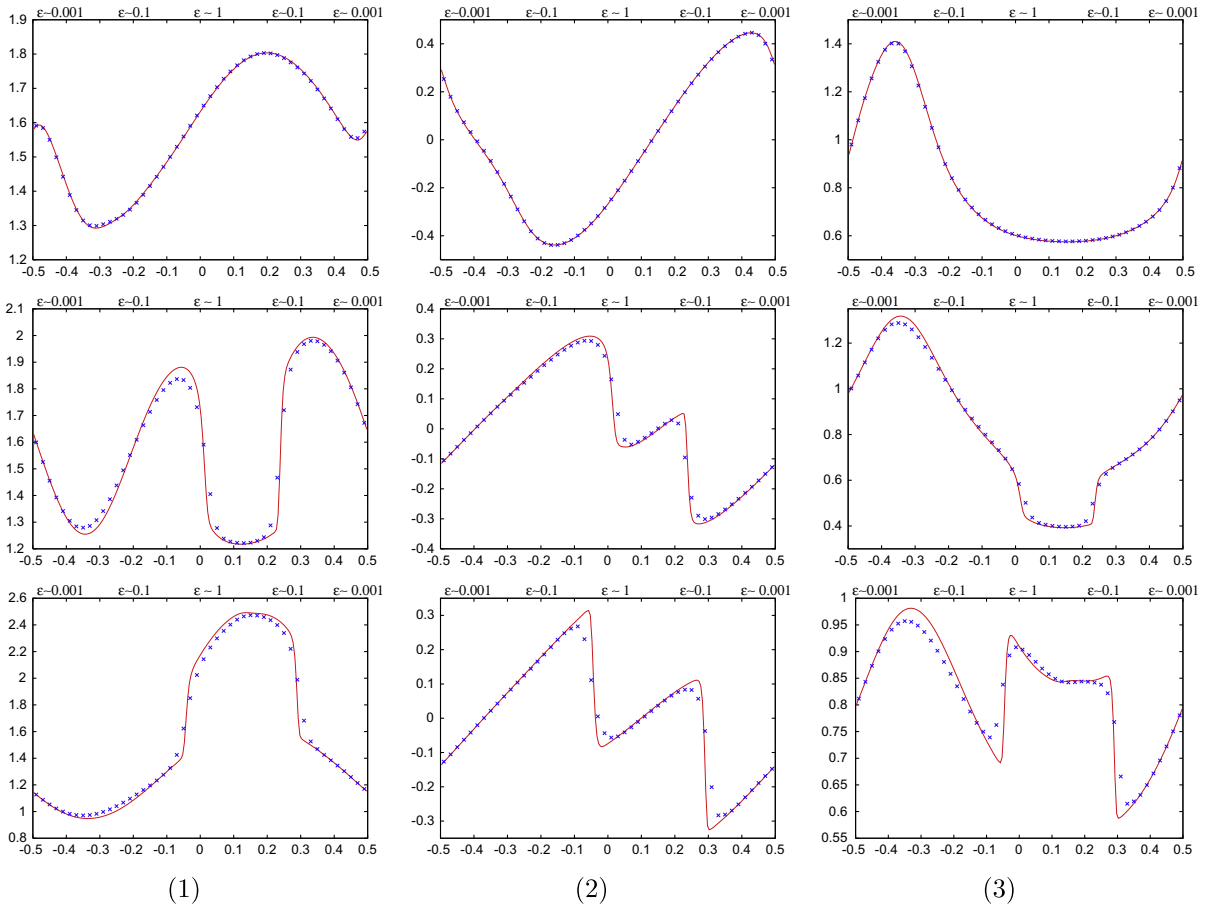


Fig. 9. Mixing regime problem ($\varepsilon_0 = 10^{-3}$), comparison of the numerical solution to the Boltzmann equation obtained with the AP scheme (4.14) using $n_x = 50$ (crosses x) and $n_x = 200$ points (line): evolution of (1) the density ρ, (2) mean velocity u, (3) temperature T at time $t = 0.25$, 0.5 and 0.75.

on the density, mean velocity and temperature and then propagate accurately into the domain. The shock speed is roughly the same for the different numerical resolutions. Therefore, this method gives a very good compromise between accuracy and stability for the different regimes. Numerical results are not plotted since they are relatively close to the ones presented in Figs. 8 and 9.

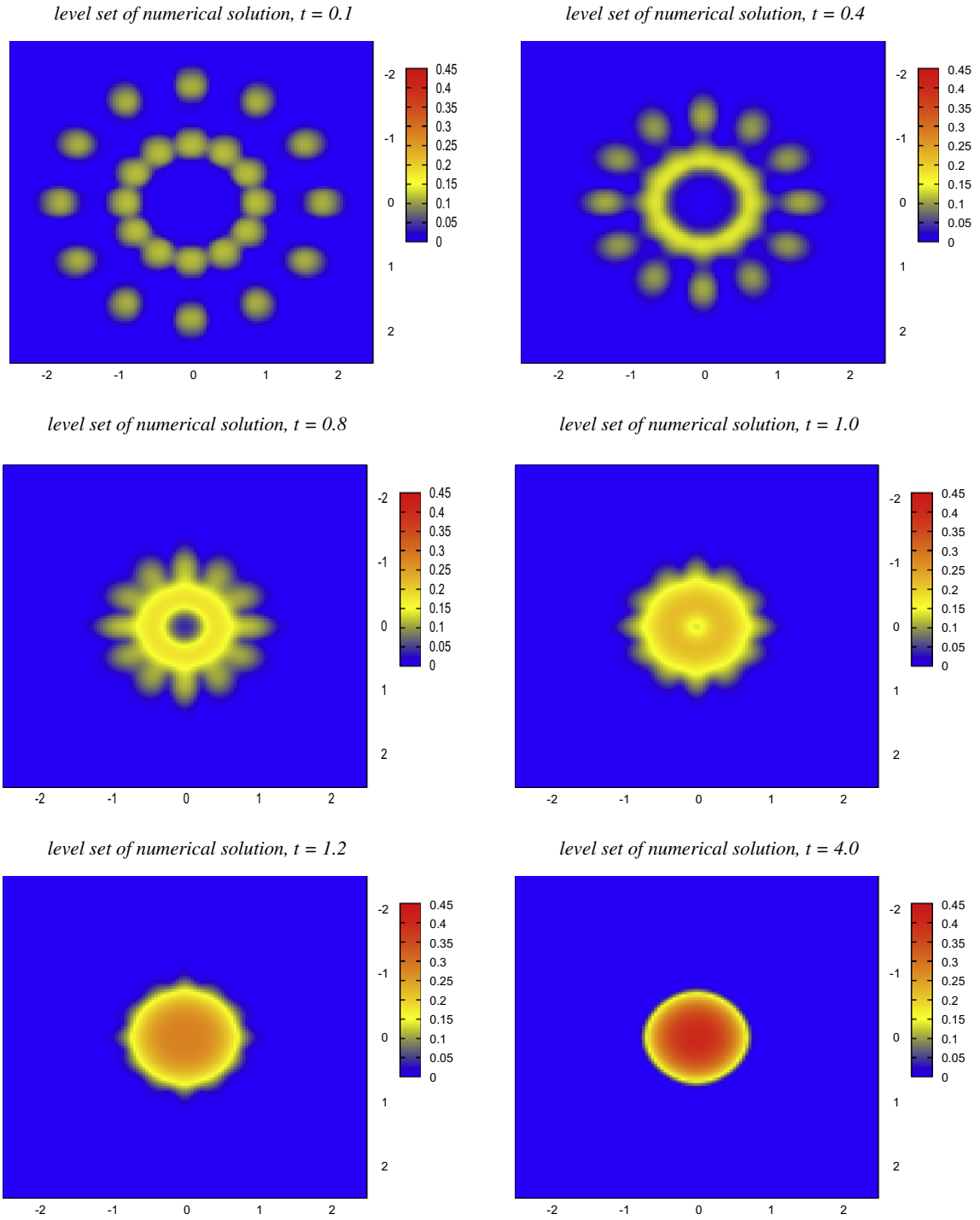


Fig. 10. Nonlinear Fokker–Planck solution: convergence toward equilibrium (Barenblatt–Pattle distribution) obtained with the first order method (2.6) using $n_x = 100$ at time $t = 0.1$, 0.4, 0.8, 1.0, 1.2 and 4 with a large time step.

6. Other applications: a nonlinear diffusion equation

In this section, we want to illustrate the efficiency of the asymptotic preserving scheme to treat high order differential operators. Such a scheme has already been applied to Willmore flow, a fourth order differential operator [54]. Here, we consider the flow of gas in a two dimensional porous medium with initial density $g_0(v) \geqslant 0$. The distribution function $g(t,v)$ then satisfies the nonlinear degenerate parabolic equation

$$\begin{cases} \frac{\partial g}{\partial t} = \Delta_v g^m, \quad v \in \mathbb{R}^{d_v}, \\ g(t=0,v) = g_0(v), \quad v \in \mathbb{R}^{d_v}, \end{cases} \tag{6.1}$$

where $m > 1$ is a physical constant. Assuming that

$$\int_{\mathbb{R}^2} (1+|v|^2) g_0(v) dv < +\infty,$$

Carrillo and Toscani [9] proved that $g(t,v)$ behaves asymptotically in a self-similar way like the Barenblatt–Pattle solution, as $t \to +\infty$. More precisely, it is easy to see that if we consider the change of variables

$$g(t,v) = \frac{1}{s(t)} f\left(\log(s(t)), \frac{v}{s(t)}\right), \tag{6.2}$$

where $s(t) := \sqrt{1+2t}$, the new distribution function f is solution to

$$\frac{\partial f}{\partial t} = \nabla_v \cdot (vf + \nabla_v f^m),$$

and converges to the Barenblatt–Pattle distribution

$$\mathcal{M}(v) = \left(C - \frac{m-1}{2m}|v|^2\right)_+^{1/(m-1)},$$

where C is uniquely determined and depends on the initial mass g_0 but not on the "details" of the initial data.

Instead of working on (6.1) directly, we will study the asymptotic decay towards its equilibrium. The key argument on the proof of Carrillo and Toscani is the control of the entropy functional

$$H(f) = \int_{\mathbb{R}^2} \left[|v|^2 f(t,v) + \frac{m}{m-1} f^m(t,v)\right] dv,$$

which satisfies

$$\frac{dH(f)}{dt} = -\int_{\mathbb{R}^2} f(t,v) \left| v + \frac{m}{m-1} \nabla f^{m-1} \right|^2 dv \leqslant 0$$

or the control of the relative entropy $H(f|\mathcal{M}) = H(f) - H(\mathcal{M})$ with respect to the steady state $\mathcal{M}$.

Numerical discretization of this problem leads to the following difficulty: explicit schemes are constrained by a CFL condition $\Delta t \simeq \Delta v^2$ whereas implicit schemes require the numerical resolution of a nonlinear problem at each time step (with a local constraint on the time step). We refer to [11,26] for a fully implicit approximation preserving steady states for nonlinear Fokker–Planck type equations.

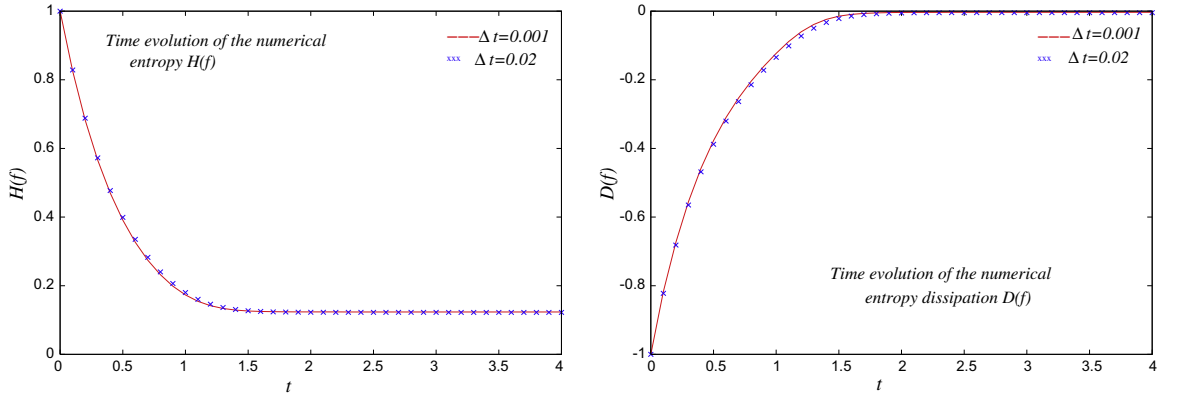


Fig. 11. Nonlinear Fokker–Planck solution: convergence toward equilibrium (Barenblatt–Pattle distribution) obtained with the first order method (2.6) using $n_x = 100$ with $\Delta t = 0.02$ and 0.001.

Here we do not focus on the velocity discretization, but only want to apply our splitting operator technique to remove this severe constraint on the time step. Here the parameter ε does not represent a physical time scale but is only related to the velocity space discretization Δv. Therefore, we set $Q(f) = \nabla_v \cdot (vf + \nabla_v f^m)$ and $P(f) = \nabla Q(\mathcal{M})(f - \mathcal{M})$, which leads to the following decomposition

$$\frac{\partial f}{\partial t} = \underbrace{\nabla_v \cdot \left(v\mathcal{M} + \nabla_v(f^m - m\mathcal{M}^{m-1}(f - \mathcal{M}))\right)}_{\text{non dissipative part}} + \underbrace{\nabla_v \cdot \left(v(f - \mathcal{M}) + m\nabla_v\left(\mathcal{M}^{m-1}(f - \mathcal{M})\right)\right)}_{\text{stiff, dissipative linear part}}.$$

Then we apply a simple IMEX scheme which only requires the numerical resolution of a linear system at each time step.

We choose $m = 3$ and a discontinuous initial datum far from the equilibrium

$$f_0(v) = \sum_{l\in\{1,2\}} \sum_{k\in\{0,\dots,n-1\}} \frac{1}{10}\mathbf{1}_{\mathcal{B}(0,r_0)}(v - v_{k,l}),$$

where $n = 12$, $r_0 = 1/4$ and $v_{k,l} = le^{i\theta_k}$, with $\theta_k = 2k\pi/n$, $k = 0, \dots, n-1$. We use a standard velocity discretization in the velocity space based on an upwind finite volume approximation for the transport term and a center difference for the diffusive part. We take $n_v^2 = 120^2$ in velocity and a time step $\Delta t = 0.02$ which is much larger than the time step satisfying a classical CFL condition for this problem $\Delta t \simeq O(\Delta v^2)$. The numerical scheme (2.6) is still stable and the numerical solution preserves nonegativity at each time step (see Fig. 10)! For large time, the solution converges to an approximation of the steady state even if the present scheme is not exactly well balanced (it does not preserve exactly the steady state). Moreover, to get a better idea on the behavior of the numerical solution, we plot the evolution of the entropy and its dissipation for different time steps (see Fig. 11). More surprisingly, the numerical entropy is decreasing and the dissipation converges towards zero when time goes to infinity.

7. Conclusion

We have proposed a new class of numerical schemes for physical problems with multiple time and spatial scales described by a stiff nonlinear source term. A prototype equation of this type is the Boltzmann equation for rarefied gas. When the Knudsen number is small, the stiff collision term of the Boltzmann equation drives the density distribution to the local Maxwellian, thus the macroscopic quantities such as mass, velocity and temperature evolve according to fluid dynamic equations such as the Euler or Navier–Stokes equations. Asymptotic-preserving (AP) schemes for kinetic equations have been successful since they capture the fluid dynamic behavior even without numerically resolving the small Knudsen number. However, the AP schemes need to treat the stiff collision terms implicitly, thus it yields a complicated numerical algebraic problem due to the nonlinearity and non-locality of the collision term. In this paper, we propose to augment the nonlinear Boltzmann collision operator by a much simpler BGK collision operator, and apply an implicit scheme only on the BGK operator which can be handled much more easily. For hyperbolic systems with relaxations We show that this method is AP in the Euler regime, after the initial transient time, and is also consistent to the Navier–Stokes approximations for suitably small time steps and mesh sizes. Numerical examples, including those with mixing scales and non-local Maxwellian initial data, demonstrate the AP property as well as uniform convergence (in the Knudsen number) of this method.

This method can be extended to a wide class of PDEs (or ODEs) with stiff source terms that admit a stable and unique local equilibrium. One example is the hyperbolic system with relaxations which are studied in this paper. We also use the nonlinear Fokker–Planck equation as an example to illustrate this point, and will pursue more applications in the future.

It is worth to mention that the present method is essentially based on a decomposition of the nonlinear operator as the sum of a linear and dissipative part and a nonlinear part (2.5). Therefore, it does not need a specific velocity and space discretization and can be easily applied to different stochastic and deterministic schemes. Moreover, based on this decomposition, other schemes can be constructed [17].

In this paper we do not mention the numerical treatment of boundary conditions although the method naturally applies to periodic and specular reflection boundary conditions. However, for physical boundary conditions, as Maxwell diffusive conditions, boundary layers will be generated where the solution is very far away from Gaussian distributions [55]. Therefore, adequate space and time discretizations deserve attention and will constitute a very interesting problem that we would like to deal with. We refer to [29] for a numerical treatment of boundary conditions for the Boltzmann equation using deterministic method [29], where the influence of boundary conditions is studied far away the boundary. This work was not aimed at the AP property. There have been very few studies on AP schemes for boundary value problems, except those on linear transport equation in the diffusive regimes with Dirichlet boundary conditions [31,37,38]. This is an important subject for future research.

Acknowledgments

F. Filbet thanks Ph. Laurençot, and L. Pareschi for interesting discussions on the topic.

References

[1] C. Bardos, F. Golse, D. Levermore, Fluid dynamic limits of kinetic equations. I. Formal derivations, J. Statist. Phys. 63 (1991) 323–344.
[2] M. Bennoune, M. Lemou, L. Mieussens, Uniformly stable numerical schemes for the Boltzmann equation preserving the compressible Navier Stokes asymptotics, J. Comput. Phys. 227 (2008) 3781–3803.
[3] P.L. Bhatnagar, E.P. Gross, K. Krook, A model for collision processes in gases, Phys. Rev. 94 (1954) 511–524.
[4] S. Bianchini, Hyperbolic limit of the Jin–Xin relaxation model, Comm. Pure Appl. Math. 47 (1994) 787–830.
[5] F. Bouchut, F. Golse, M. Pulvirenti, Kinetic Equations and Asymptotic Theory, Gauthiers-Villars, 2000.
[6] J.-F. Bourgat, P. Le Tallec, B. Perthame, Y. Qiu, Coupling Boltzmann and Euler equations without overlapping, in Domain decomposition methods in science and engineering (Como, 1992), Contemp. Math., 157, American Mathematical Society, Providence, RI, 1994. pp. 377–398.
[7] R. Caflish, S. Jin, G. Russo, Uniformly accurate schemes for hyperbolic systems with relaxation, SIAM J. Numer. Anal. 34 (1997) 246–281.
[8] R.E. Caflisch, L. Pareschi, An implicit Monte Carlo method for rarefied gas dynamics. I: The space homogeneous case, J. Comput. Phys. 154 (1999) 90–116.
[9] J.A. Carrillo, G. Toscani, Asymptotic L^1-decay of solutions of the porous medium equation to self-similarity, Indiana Univ. Math. J. 49 (2000) 113–142.
[10] C. Cercignani, The Boltzmann Equation and its Applications, Springer, 1998.
[11] C. Chainais-Hillairet, F. Filbet, Asymptotic behavior of a finite volume scheme for the transient drift-diffusion model, IMA J. Numer. Anal. 27 (2007) 689–716.
[12] G.Q. Chen, T.P. Liu, C.D. Levermore, Hyperbolic conservation laws with stiff relaxation terms and entropy, Comm. Pure Appl. Math. 47 (6) (1994) 787–830.
[13] F. Coquel, B. Perthame, Relaxation of energy and approximate Riemann solvers for general pressure laws in fluid dynamics, SIAM J. Numer. Anal. 35 (6) (1998) 2223–2249 (English summary).
[14] F. Coron, B. Perthame, Numerical passage from kinetic to fluid equations, SIAM J. Numer. Anal. 28 (1991) 26–42.
[15] P. Crispel, P. Degond, M.-H. Vignal, An asymptotically preserving scheme for the two-fluid Euler–Poisson model in the quasi-neutral limit, J. Comput. Phys. 223 (2007) 208–234.
[16] P. Degond, F. Deluzet, L. Navoret, An asymptotically stable Particle-in-Cell (PIC) scheme for collisionless plasma simulations near quasineutrality, C. R. Acad. Sci. Paris Sér. I Math. 343 (2006) 613–618.
[17] P. Degond, G. Dimarco, L. Pareschi, in press.
[18] P. Degond, S. Jin, A smooth transition model between kinetic and diffusion equations, SIAM J. Numer. Anal. 41 (6) (2005) 2671–2687.
[19] P. Degond, S. Jin, J.-G. Liu, Mach-number uniform asymptotic-preserving gauge schemes for compressible flows, Bull. Inst. Math. Acad. Sin. (N.S.) 2 (2007) 851–892.
[20] P. Degond, S. Jin, L. Mieussens, A smooth transition model between kinetic and hydrodynamic equations, J. Comput. Phys. 207 (2005) 665–694.
[21] P. Degond, M. Tang, All speed scheme for the low mach number limit of the isentropic Euler equation, Commun. Comput. Phys., preprint.
[22] R. Duclous, B. Dubroca, F. Filbet, V. Tikhonchuk, High order resolution of the Maxwell–Fokker–Planck–Landau model intended for ICF applications, J. Comput. Phys 228 (2009) 5072–5100.
[23] F. Filbet, L. Pareschi, A numerical method for the accurate solution of the Fokker–Planck–Landau equation in the non homogeneous case, J. Comput. Phys. 179 (2002) 1–26.
[24] F. Filbet, G. Russo, High order numerical methods for the space non-homogeneous Boltzmann equation, J. Comput. Phys. 186 (2003) 457–480.
[25] F. Filbet, L. Pareschi, G. Toscani, Accurate numerical methods for the collisional motion of (heated) granular flows, J. Comput. Phys. 202 (2005) 216–235.
[26] F. Filbet, A finite volume scheme for the Patlak–Keller–Segel chemotaxis model, Numerische Mathematik 104 (2006) 457–488.
[27] F. Filbet, C. Mouhot, L. Pareschi, Solving the Boltzmann equation in $N \log_2 N$, SIAM J. Sci. Comput. 28 (2006) 1029–1053.
[28] E. Gabetta, L. Pareschi, G. Toscani, Relaxation schemes for nonlinear kinetic equations, SIAM J. Numer. Anal. 34 (1997) 2168–2194.
[29] I. Gamba, S.H. Tharskabhushanam, Shock and boundary structure formation by spectral-Lagrangian methods for the inhomogeneous Boltzmann transport equation, J. Comp. Math. (2009).
[30] C.W. Gear, Numerical Initial Value Problems in Ordinary Differential Equations, Prentice Hall, 1971.
[31] F. Golse, S. Jin, C.D. Levermore, The convergence of numerical transfer schemes in diffusive regimes I: The discrete–ordinate method, SIAM J. Numer. Anal. 36 (1999) 1333–1369.
[32] L. Gosse, G. Toscani, An asymptotic-preserving well-balanced scheme for the hyperbolic heat equations, C. R. Math. Acad. Sci. Paris 334 (2002) 337–342.
[33] M. Günther, P. Le Tallec, J.-P. Perlat, J. Struckmeier, Numerical modeling of gas flows in the transition between rarefied and continuum regimes. Numerical flow simulation I, (Marseille, 1997), Notes Numer. Fluid Mech., 66, Vieweg, Braunschweig, 1998. pp. 222–241.
[34] J. Haack, S. Jin, J.-G. Liu, An all-speed asymptotic-preserving schemes for compressible flows, in press.
[35] S. Jin, Effcient asymptotic-preserving (AP) schemes for some multiscale kinetic equations, SIAM J. Sci. Comput. 21 (1999) 441–454.
[36] S. Jin, Runge–Kutta methods for hyperbolic conservation laws with stiff relaxation terms, J. Comput. Phys. 122 (1995) 51–67.
[37] S. Jin, C.D. Levermore, The discrete–ordinate method in diffusive regime, Transp. Theory Stat. Phys. 20 (1991) 413–439.
[38] S. Jin, C.D. Levermore, Fully-discrete numerical transfer in diffusive regimes, Transp. Theory Stat. Phys. 22 (1993) 739–791.
[39] S. Jin, C.D. Levermore, Numerical schemes for hyperbolic conservation laws with stiff relaxation terms, J. Comput. Phys. 126 (2) (1996) 449–467.
[40] S. Jin, L. Pareschi, Discretization of the multiscale semiconductor Boltzmann equation by diffusive relaxation schemes, J. Comput. Phys. 161 (2000) 312–330.
[41] S. Jin, L. Pareschi, G. Toscani, Diffusive relaxation schemes for discrete-velocity kinetic equations, SIAM J. Num. Anal. 35 (1998) 2405–2439.
[42] S. Jin, L. Pareschi, G. Toscani, Uniformly accurate diffusive relaxation schemes for multiscale transport equations, SIAM J. Numer. Anal. 38 (2000) 913–936.
[43] S. Jin, Z.P. Xin, The relaxation schemes for systems of conservation laws in arbitrary space dimensions, Comm. Pure Appl. Math. 48 (1995) 235–276.
[44] A. Klar, An asymptotic-induced scheme for nonstationary transport equations in the diffusive limit, SIAM J. Numer. Anal. 35 (1998) 1073–1094.
[45] A. Klar, An asymptotic preserving numerical scheme for kinetic equations in the low Mach number limit, SIAM J. Numer. Anal. 36 (1999) 1507–1527.
[46] A. Klar, H. Neunzert, J. Struckmeier, Transition from kinetic theory to macroscopic fluid equations: a problem for domain decomposition and a source for new algorithm, Transp. Theory Stat. Phys. 29 (2000) 93–106.
[47] E.W. Larsen, J.E. Morel, W.F. Miller Jr., Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes, J. Comp. Phys. 69 (1987) 283–324.
[48] E. Larsen, J.E. Morel, Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes II, J. Comp. Phys. 83 (1989) 212–236.
[49] M. Lemou, L. Mieussens, A new asymptotic preserving scheme based on micro-macro formulation for linear kinetic equations in the diffusion limit, SIAM J. Sci. Comput. 31 (1) (2008) 334–368.
[50] R. Natalini, Recent results on hyperbolic relaxation problems. Analysis of systems of conservation laws (Aachen, 1997), Chapman and Hall/CRC Monogr. Surv. Pure Appl. Math., 99, Chapman & Hall/CRC, Boca Raton, FL, 1999. pp. 128–198.
[51] L. Pareschi, G. Russo, Time relaxed Monte Carlo methods for the Boltzmann equation, SIAM J. Sci. Comput. 23 (2001) 1253–1273.
[52] L. Pareschi, G. Russo, Numerical solution of the Boltzmann equation I: Spectrally accurate approximation of the collision operator, SIAM J. Numer. Anal. 37 (2000) 1217–1245.
[53] G. Puppo, S. Pieraccini, Implicit–explicit schemes for BGK kinetic equations, J. Sci. Comput. 32 (2007) 1–28.

[54] P. Smereka, Semi-implicit level set methods for curvature and surface diffusion motion, J. Sci. Comput. 19 (2003) 439–456 (Special issue in honor of the 60th birthday of Stanley Osher).
[55] Y. Sone, Kinetic Theory and Fluid Dynamics, Birkhauser, Boston, 2002.
[56] P. Le Tallec, F. Mallinger, Coupling Boltzmann and Navier–Stokes equations by half fluxes, J. Comput. Phys. 136 (1997) 51–67.
[57] G.B. Whitham, Linear and Nonlinear Waves, Wiley-Interscience, 1974.
[58] H.C. Yee, A Class of High-Resolution Explicit and Implicit Shock-Capturing Methods, Von Karman Institute for Fluid Dynamics, Lecture Series, 1989.