SIAM J. Numer. Anal.
Vol. 34, No. 1, pp. 246–281, February 1997

UNIFORMLY ACCURATE SCHEMES FOR HYPERBOLIC SYSTEMS WITH RELAXATION*

RUSSEL E. CAFLISCH†, SHI JIN‡, AND GIOVANNI RUSSO§

Abstract. We develop high-resolution shock-capturing numerical schemes for hyperbolic systems with relaxation. In such systems the relaxation time may vary from order-1 to much less than unity. When the relaxation time is small, the relaxation term becomes very strong and highly stiff, and underresolved numerical schemes may produce spurious results. Usually one cannot decouple the problem into separate regimes and handle different regimes with different methods. Thus it is important to have a scheme that works uniformly with respect to the relaxation time. Using the Broadwell model of the nonlinear Boltzmann equation we develop a second-order scheme that works effectively, with a fixed spatial and temporal discretization, for all ranges of the mean free path. Formal uniform consistency proof for a first-order scheme and numerical convergence proof for the second-order scheme are also presented. We also make numerical comparisons of the new scheme with some other schemes. This study is motivated by the reentry problem in hypersonic computations.

Key words. hyperbolic systems with relaxation, Broadwell model, stiff source, high-resolution shock-capturing methods

AMS subject classifications. 35L65, 49M25, 34A65, 82B40

PII. S0036142994268090

1. Introduction.

Hyperbolic systems with relaxation are used to describe many physical problems that involve both convection and nonlinear interaction. In the Boltzmann equation from the kinetic theory of rarefied gas dynamics, the collision (relaxation) terms describe the interaction of particles. In viscoelasticity, memory effects are modeled as relaxation. Relaxation occurs in water waves when the gravitational force balances the frictional force of the riverbed. For gas in thermal nonequilibrium the internal state variable satisfies a rate equation that measures a departure of the system from the local equilibrium. Relaxations also occur in other problems ranging from magnetohydrodynamics to traffic flow.

In such systems, when the nonlinear interactions are strong, the relaxation rate is large. In kinetic theory, for example, this occurs when the mean free path between collisions is small (i.e., the Knudsen number is small). Within this regime, which is referred to as the *fluid dynamic limit*, the gas flow is well described by the Euler or Navier–Stokes equations of fluid mechanics, except in shock layers and boundary layers. The characteristic length scale of the kinetic description of the gas is the microscopic, collision distance; in the fluid dynamic limit it is replaced by the macroscopic length scale of fluid dynamics. By analogy with the kinetic theory, we shall refer to the limit of large relaxation rate (or small relaxation time) for a general hyperbolic system with relaxation as the fluid dynamic limit.

The fluid dynamic limit is challenging for numerical methods, because in this regime the relaxation terms become stiff. In particular, a standard numerical scheme

*Received by the editors May 18, 1994; accepted for publication (in revised form) April 5, 1995.
http://www.siam.org/journals/sinum/34-1/26809.html

†Department of Mathematics, University of California, Los Angeles, CA 90024. The research of this author was supported in part by ARO grant DAAL03-91-G-0162.

‡School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332 (jin@math.gatech.edu). The research of this author was supported in part by NSF grant DMS-9404157 and by AFOSR grant F49620-92-J0098 as a visiting member at the Courant Institute.

§Dipartimento di Matematica, Università dell'Aquila, 67010 L'Aquila, Italy.

might fail to give physically correct solutions once the (microscopic) relaxation distance is much smaller than the spatial discretization. Although a full simulation of the relaxation process would require a very fine (and expensive) discretization, it may be possible to accurately compute the solution on a coarser fluid dynamic length scale. The goal of this paper is to present a class of numerical methods that work with uniform accuracy from the rarefied regime to the fluid dynamic limit for the Broadwell model of kinetic theory.

Numerical methods for hyperbolic systems with relaxation terms have attracted a lot of attention in recent years [7], [23], [24], [16], [17]. Studying the numerical behavior for these problems is important not only for the physical applications but also for the development of new numerical methods for conservation laws, such as kinetic schemes [15], [8], [25] and relaxation schemes [18]. Most kinetic or relaxation schemes can be described as fractional step methods, in which the collision step is just a projection of the system into a sort of discrete "local Maxwellian" or local equilibrium. Although the goal of a kinetic scheme or relaxation scheme is different from ours, nevertheless we use them as guidelines for the study of the properties of a numerical scheme near the fluid dynamic regime.

In earlier works on the system of hyperbolic equations with relaxation, the goal was to develop robust numerical schemes that handle the stiffness of the problem effectively and avoid spurious numerical solutions when the grid spacing underresolves the small relaxation time. In regions where the relaxation time is no longer small and the problem becomes nonstiff, however, these schemes usually may not have high-order accuracy uniformly with respect to the wide range of the relaxation time.

Our motivation differs from these earlier approaches in that we seek to develop robust numerical schemes that work *uniformly* for a wide range of relaxation rates. We consider a simpler model of the Boltzmann equation, and we derive a numerical scheme which is of second order uniformly in the mean free path. This is motivated by hypersonic computations of reentry problems. The new scheme we introduce here, called NSPIF, is able to handle all different regimes from the rarefied gas to the fluid limit (the stiff regime) with fixed spatial and temporal grids that are independent of the mean free path. Although we develop our method based on the Broadwell model, this scheme can be applied to a much wider class of hyperbolic systems with relaxation terms and to other discrete velocity kinetic models. In particular, it applies to a class of hyperbolic systems with relaxation characterized by Liu [22] and Chen, Levermore, and Liu [5].

Probably the paper that is closest in spirit to our work is the one by Coron and Perthame [7]. In that paper the authors derive a numerical scheme for solving the BGK model of the Boltzmann equation under a wide range of mean free path. They discretize velocity space and use a splitting scheme. The collision step is treated by a semi-implicit method that guarantees positivity and entropy condition for the time-discrete model. The scheme is first-order accurate in space and time.

Development of numerical methods for the problems considered here is considerably aided by knowledge of the equations for the fluid dynamic limit. In other stiff source problems, such as those arising in reacting flow computations, the corresponding limit may be less well understood, and extra efforts are necessary to develop underresolved numerical methods [6], [21], [9], [23].

The outline of the paper is as follows. In the next section we present the Broadwell model and its fluid dynamic limit. Section 3 deals with the treatment of the convective step. Upwind methods and flux limiters are briefly recalled. Sections 4–7 are devoted to the development and analysis of a second-order scheme (in space and time) called NSP. In section 4, the scheme is derived making use of truncation analysis, and in the following section the linear stability of this scheme is studied. In section 6 the Richardson extrapolation is used for the first time step in order to guarantee second-order accuracy even in presence of an initial layer. The NSP combined with this initial step is called the NSPIF. In section 7 the fluid dynamic limit of the NSPIF is derived. In section 8 we study a first-order splitting scheme (called the SP1). Important properties, such as positivity and entropy inequality, are proved for the scheme. The main result in this section is a uniformly (in the relaxation time) first-order consistency error estimate. Our analysis is based on the assumption of smoothness of the solution. To the best of our knowledge this is the first estimate on numerical methods for hyperbolic systems with stiff relaxation terms, even for smooth solutions. Also, the uniformly first-order accuracy sharpens earlier work in this direction [14]. Estimates on nonsmooth solutions or on higher order methods for such inhomogeneous hyperbolic systems would require more advanced analytic techniques and are beyond the scope of this paper. In the last section we present some numerical results which show that NSPIF overperforms several other approaches and always gives second-order physical solutions over a wide range of the mean free path.

2. The Broadwell model. A simple discrete velocity kinetic model for a gas was introduced by Broadwell [2]. It describes a two-dimensional (2-D) (3-D) gas as composed of particles of only four (six) velocities with a binary collision law and spatial variation in only one direction. When looking for one-dimensional (1-D) solutions of the 2-D gas, the evolution equations of the model are given by

$$\begin{aligned} \partial_t f + \partial_x f &= \frac{1}{\varepsilon}(h^2 - fg), \\ \partial_t h &= -\frac{1}{\varepsilon}(h^2 - fg), \\ \partial_t g - \partial_x g &= \frac{1}{\varepsilon}(h^2 - fg), \end{aligned} \tag{2.1}$$

where ε is the mean free path, f, h, and g denote the mass densities of gas particles with speed $1, 0$, and -1, respectively, in space x and time t. Similar equations are obtained for 1-D solutions of a 3-D gas [3].

The fluid dynamic moment variables are density ρ, momentum m, and velocity u defined by

$$\rho = f + 2h + g, \quad m = f - g, \quad u = \frac{m}{\rho}. \tag{2.2}$$

In addition, define

$$z = f + g. \tag{2.3}$$

Then the Broadwell equations can be rewritten as

$$\partial_t \rho + \partial_x m = 0, \tag{2.4}$$

$$\partial_t m + \partial_x z = 0, \tag{2.5}$$

$$\partial_t z + \partial_x m = \frac{1}{2\varepsilon}(\rho^2 + m^2 - 2\rho z). \tag{2.6}$$

Note that if the fluid variables ρ, m, and z are known then f, g, and h can be recovered from (2.2) and (2.3) as

$$f = \frac{1}{2}(z + m), \quad g = \frac{1}{2}(z - m), \quad h = \frac{1}{2}(\rho - z).$$

A *local Maxwellian* is a density function that satisfies

$$Q(f,h,g) = h^2 - fg = \rho^2 + m^2 - 2\rho z = 0, \tag{2.7}$$

i.e.,

$$z = z_E(\rho, m) \equiv \frac{1}{2\rho}(\rho^2 + m^2) = \frac{1}{2}(\rho + \rho u^2). \tag{2.8}$$

As $\varepsilon \to 0$ equation (2.1) or (2.6) gives the local Maxwellian distribution (2.8). Applying (2.8) in (2.5), one gets the fluid dynamic limit described by the following model Euler equations:

$$\begin{aligned} \partial_t \rho + \partial_x(\rho u) &= 0, \\ \partial_t(\rho u) + \partial_x\left(\frac{1}{2}(\rho + \rho u^2)\right) &= 0. \end{aligned} \tag{2.9}$$

To the next order, a model Navier–Stokes equation can be derived via the Chapman–Enskog expansion [4]. For a description of the Broadwell model and its fluid dynamic limit see, for example, [3] and [13].

Previously the numerical solution of the Broadwell equations has been considered by several authors [10, 12, 1]. In these earlier works the attention was focused on developing methods that work in the rarefied regime $\epsilon = O(1)$. Many of these methods will have problems when $\epsilon \to 0$. Among the problems that arose in the fluid regimes are numerical instability, poor shock, and rarefaction resolutions and even spurious numerical solutions. See the numerical examples in section 9.

The Broadwell equations are a prototypical example for more general hyperbolic systems with relaxations in the sense of Whitham [29] and Liu [22]. These problems can be described mathematically by the system of evolutional equations

$$\partial_t U + \partial_x F(U) = -\frac{1}{\varepsilon} R(U), \qquad U \in \mathcal{R}^N. \tag{2.10}$$

We will call this system the *relaxation system.* Here we use the term relaxation in the sense of Whitham [29] and Liu [22]. The relaxation term is endowed with an $n \times N$ constant matrix $\mathcal{Q}$ with rank $n < N$ such that

$$\mathcal{Q}R(U) = 0 \qquad \text{for all} \quad U. \tag{2.11}$$

This yields n independent conserved quantities $v = \mathcal{Q}U$. In addition we assume that each such v uniquely determines a local equilibrium value $U = \mathcal{E}(v)$ satisfying $R(\mathcal{E}(v)) = 0$ and such that

$$\mathcal{Q}\mathcal{E}(v) = v \qquad \text{for all} \quad v. \tag{2.12}$$

The image of $\mathcal{E}$ then constitutes the manifold of local equilibria of R.

Associated with $\mathcal{Q}$ are n local conservation laws satisfied by every solution of (2.10) that take the form

$$\partial_t(\mathcal{Q}U) + \partial_x(\mathcal{Q}F(U)) = 0. \tag{2.13}$$

These can be closed as a reduced system for $v = \mathcal{Q}U$ if we take the local relaxation approximation

$$U = \mathcal{E}(v), \tag{2.14}$$

$$\partial_t v + \partial_x e(v) = 0, \tag{2.15}$$

where the reduced flux e is defined by

$$e(v) \equiv \mathcal{Q}F(\mathcal{E}(v)). \tag{2.16}$$

A system of conservation laws with relaxation is stiff when ε is small compared with the time scale determined by the characteristic speeds of the system and some appropriate length scales. While we mainly concentrate on the Broadwell equation, the analysis as well as the numerical schemes can certainly be applied to this class of relaxation problems. In fact from time to time we will use the general equation (2.10) to simplify the notation.

3. The discretizations of the convection terms. We introduce the spatial grid points $x_{j+\frac{1}{2}},\ j = \cdots, -1, 0, 1, \cdots$ with uniform mesh spacing $\Delta x = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$ for all j. The time level $t_0, t_1, \cdots$ are also spaced uniformly with space step $\Delta t = t^{n+1} - t^n$ for $n = 0, 1, 2, \cdots$. Here the assumption of a uniform grid is only for simplicity. We use U_j^n to denote the cell average of U in the cell $[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$ at time t^n,

$$U_j^n = \frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} U(t^n, x)\, dx. \tag{3.1}$$

We use the method of lines, in which the time discretization and spatial discretization are taken separately, for the Broadwell equations. In this section we shall discuss the spatial discretization, which concerns the linear convection terms. Note that the linear convection in the Broadwell equation is of hyperbolic type. Thus it is natural to use upwind schemes.

A conservative spatial discretization to the Broadwell equations (2.4)–(2.6) is

$$\partial_t \rho_j + \frac{m_{j+\frac{1}{2}} - m_{j-\frac{1}{2}}}{\Delta x} = 0, \tag{3.2}$$

$$\partial_t m_j + \frac{z_{j+\frac{1}{2}} - z_{j-\frac{1}{2}}}{\Delta x} = 0, \tag{3.3}$$

$$\partial_t z_j + \frac{m_{j+\frac{1}{2}} - m_{j-\frac{1}{2}}}{\Delta x} = \frac{1}{2\varepsilon}(\rho_j^2 + m_j^2 - 2\rho_j z_j). \tag{3.4}$$

Equations (3.2)–(3.4) can be diagonalized into

$$\begin{aligned} \partial_t f_j + \frac{f_{j+\frac{1}{2}} - f_{j-\frac{1}{2}}}{\Delta x} &= \frac{1}{\varepsilon}(h_j^2 - f_j g_j), \\ \partial_t h_j &= -\frac{1}{\varepsilon}(h_j^2 - f_j g_j), \\ \partial_t g_j - \frac{g_{j+\frac{1}{2}} - g_{j-\frac{1}{2}}}{\Delta x} &= \frac{1}{\varepsilon}(h_j^2 - f_j g_j), \end{aligned} \tag{3.5}$$

where f, h, and g are exactly the original density functions for the Broadwell equation. The connection between (3.4) and (3.5) is established through the definition of the fluid moment variables (2.2), (2.3). Since f and g travel along the constant characteristics with speeds 1 and -1, respectively, upwind schemes can be easily applied to them.

3.1. The upwind scheme. The upwind scheme applied to f and g gives

$$f_{j+\frac{1}{2}} = f_j, \quad g_{j+\frac{1}{2}} = g_{j+1} \tag{3.6}$$

while h_j is constant. This implies

$$(z+m)_{j+\frac{1}{2}} = (z+m)_j, \quad (z-m)_{j+\frac{1}{2}} = (z-m)_{j+1}, \tag{3.7}$$

or equivalently in fluid moment variables

$$\begin{aligned} m_{j+\frac{1}{2}} &= \frac{m_{j+1}+m_j}{2} - \frac{z_{j+1}-z_j}{2}, \\ z_{j+\frac{1}{2}} &= \frac{z_{j+1}+z_j}{2} - \frac{m_{j+1}-m_j}{2}. \end{aligned} \tag{3.8}$$

Applying (3.8) in (3.3) gives the semidiscrete upwind scheme for the convection step

$$\begin{aligned} \partial_t \rho_j + \frac{m_{j+1}-m_{j-1}}{2\Delta x} - \frac{z_{j+1}-2z_j+z_{j-1}}{2\Delta x} &= 0, \\ \partial_t m_j + \frac{z_{j+1}-z_{j-1}}{2\Delta x} - \frac{m_{j+1}-2m_j+m_{j-1}}{2\Delta x} &= 0, \\ \partial_t z_j + \frac{m_{j+1}-m_{j-1}}{2\Delta x} - \frac{z_{j+1}-2z_j+z_{j-1}}{2\Delta x} &= \frac{1}{2\varepsilon}(\rho_j^2 + m_j^2 - 2\rho_j z_j). \end{aligned} \tag{3.9}$$

3.2. van Leer's MUSCL scheme. A second-order extension of the upwind scheme is van Leer's MUSCL scheme [28]. While the upwind scheme uses piecewise constant interpolation (3.6), the MUSCL uses piecewise linear interpolation along with slope limiters to eliminate numerical oscillations at discontinuities. Applying MUSCL to the Riemann invariants f and g, respectively, we obtain

$$f_{j+\frac{1}{2}} = f_j + \frac{1}{2}\Delta x\, \sigma_j^+, \quad g_{j+\frac{1}{2}} = g_{j+1} - \frac{1}{2}\Delta x\, \sigma_{j+1}^-. \tag{3.10}$$

Here σ_j^+ and σ_j^- are the slope of f and g on the jth cell, respectively. For $w^+ = f, w^- = g$, $\sigma^\pm$ are defined by [20]

$$\sigma_j^\pm = \frac{1}{\Delta x}(w_{j+1}^\pm - w_j^\pm)\phi(\theta_j^\pm), \quad \theta_j^\pm = \frac{w_j^\pm - w_{j-1}^\pm}{w_{j+1}^\pm - w_j^\pm}, \tag{3.11}$$

and the slope limit function $\phi(\theta)$ by van Leer is

$$\phi(\theta) = \frac{|\theta| + \theta}{1 + |\theta|}. \tag{3.12}$$

With this limiter the MUSCL scheme is total variation diminishing (TVD). Rewriting (3.10) in the fluid moment variables gives

$$m_{j+\frac{1}{2}} = \frac{m_{j+1}+m_j}{2} - \frac{z_{j+1}-z_j}{2} + \frac{\Delta x}{4}(\sigma_j^+ + \sigma_{j+1}^-), \tag{3.13}$$

$$z_{j+\frac{1}{2}} = \frac{z_{j+1}+z_j}{2} - \frac{m_{j+1}-m_j}{2} + \frac{\Delta x}{4}(\sigma_j^+ - \sigma_{j+1}^-). \tag{3.14}$$

Applying (3.13), (3.14) in (3.3) finally gives

$$\partial_t\rho_j + \frac{m_{j+1}-m_{j-1}}{2\Delta x} - \frac{z_{j+1}-2z_j+z_{j-1}}{2\Delta x} + \frac{\Delta x}{4}(\sigma_j^+ + \sigma_{j+1}^- - \sigma_{j-1}^+ - \sigma_j^-) = 0,$$

$$\partial_t m_j + \frac{z_{j+1}-z_{j-1}}{2\Delta x} - \frac{m_{j+1}-2m_j+m_{j-1}}{2\Delta x} + \frac{\Delta x}{4}(\sigma_j^+ - \sigma_{j+1}^- - \sigma_{j-1}^+ + \sigma_j^-) = 0, \tag{3.15}$$

$$\partial_t z_j + \frac{m_{j+1}-m_{j-1}}{2\Delta x} - \frac{z_{j+1}-2z_j+z_{j-1}}{2\Delta x} + \frac{\Delta x}{4}(\sigma_j^+ + \sigma_{j+1}^- - \sigma_{j-1}^+ - \sigma_j^-) = \frac{1}{2\varepsilon}(\rho_j^2 + m_j^2 - 2\rho_j z_j). \tag{3.16}$$

4. A uniformly second-order time discretization.

In this and in the next three sections we develop a second-order scheme, uniformly accurate in the relaxation parameter, and study its properties. Since our goal is to seek a robust scheme that works uniformly for all range of ε, it is essential that the time discretization is stable for every ε. This is especially important near the fluid regime where the mean free path is small and the problem becomes stiff.

Uniformly numerical stability can be achieved with implicit temporal integrators. Since stiffness appears only through the relaxation term, it is natural to keep the convection terms explicit, while the collision step has to be treated implicitly.

Numerical experience for such relaxation problems shows that this problem is not merely a numerical stability problem. Stable numerical discretization may still produce spurious solutions [23], [24], [16]. For example, the Crank–Nicolson scheme

$$z_j^{n+1} = z_j^n + \frac{\Delta t}{2\varepsilon}[(\rho_j^n)^2 + (m_j^n)^2 - \rho_j^n(z_j^n + z_j^{n+1})], \tag{4.1}$$

coupled with the free-streaming convection with $\Delta x = \Delta t$ gives a scheme that is stable independent of ε but does not have the correct fluid limit.

Previous results [16] demonstrate that an effective condition for the correct numerical behavior near the fluid regime is that the numerical scheme should possess the correct fluid limit, in the sense that a discrete analogue of the Chapman–Enskog expansion for the continuous equations remains valid for the numerical discretization, and the resulting numerical fluid limit should be a good discretization for the model Euler limit. A sufficient condition to achieve this is that the collision step always projects the nonequilibrium data into a local Maxwellian at every time step.

We now show how to construct a uniformly second-order scheme. The basic idea is to combine a high-order convection step with an implicit collision step according to the following guidelines:

(i) Truncation error analysis is used to obtain second-order accuracy in the rarefied regime ($\varepsilon = O(1)$).

(ii) The collision step is well posed $\forall\varepsilon > 0$, and it relaxes the system to a local Maxwellian as $\varepsilon \to 0$.

(iii) The scheme should be unconditionally stable in the collision step.

(iv) Check that the limit scheme obtained as $\varepsilon \to 0$ is a second-order scheme for the model Euler equations when $\varepsilon = 0$.

It is possible to show that if condition (ii) is not satisfied, then the scheme may give the wrong behavior in the fluid dynamic limit (this is the case, for example, of the scheme in [10]).

Such a splitting scheme, applied to system (2.10), can be written in the form

$$U_1 = U^n - \alpha \frac{\Delta t}{\varepsilon} R(U_1), \tag{4.2}$$

$$U_2 = U_1 - \tilde{\alpha} \Delta t \mathcal{D} F(U_1), \tag{4.3}$$

$$U_3 = U_2 - \beta \frac{\Delta t}{\varepsilon} R(U_3) - \gamma \frac{\Delta t}{\varepsilon} R(U_1), \tag{4.4}$$

$$U_4 = U_3 - \tilde{\beta} \Delta t \mathcal{D} F(U_3), \tag{4.5}$$

$$U_5 = \xi U^n + \eta U_4, \tag{4.6}$$

$$U^{n+1} = U_5 - \mu \frac{\Delta t}{\varepsilon} R(U^{n+1}), \tag{4.7}$$

where $\mathcal{D}$ denotes a discrete approximation of the convection operator, such as the simple upwind or a second-order (or higher) MUSCL scheme. We assume that $\mathcal{D}$ is at least second-order accurate, so that the splitting scheme will be second-order accurate both in space and time.

The scheme is a fractional step combination of convection and collision steps. The parameters $\alpha, \beta, \gamma, \mu, \tilde{\alpha}, \tilde{\beta}, \xi, \eta$ will be determined by conditions (i)–(iii).

Roughly speaking, this splitting scheme mimics the asymptotic process that leads from the Broadwell equations to the model Euler equations. At $t = t^{(1)}$, the stiff source solver gives $R(U_1) \approx 0$, which is the local Maxwellian. Applying it to the next convection step $t = t^{(2)}$ should give a numerical flux that approximates the flux for the model Euler equation. Similar behavior occurs at $t = t^{(3)}$ and $t^{(4)}$. The last step gives $R(U^{n+1}) \approx 0$, guaranteeing the correct local Maxwellian at $t = t^{n+1}$. The above steps are combined in a second-order way, aiming at a scheme with uniform second-order accuracy.

In order to satisfy condition (i) we apply the scheme to the linear system

$$\partial_t U + AU + BU = 0, \tag{4.8}$$

where A and B are constant matrices. The exact solution of (4.8) at time $t = \Delta t$ is given by

$$U(\Delta t) = e^{-(A+B)\Delta t} U(0).$$

Apply scheme (4.2)–(4.7) to (4.8) and write the difference equation in the compact form

$$U^{n+1} = \mathcal{C} U^n. \tag{4.9}$$

To achieve a second-order accuracy we impose that

$$(\mathcal{C}(\Delta t) - e^{-(A+B)\Delta t}) U(0) = O(\Delta t^3),$$

then the following restrictions on the parameters should be satisfied:

$$
\begin{aligned}
&\xi+\eta=1,\\
&\eta(\tilde\alpha+\tilde\beta)=1,\\
&\eta(\alpha+\beta+\gamma)+\mu(\xi+\eta)=1,\\
&2\eta\tilde\alpha\tilde\beta=1,\\
&2\eta(\alpha\tilde\alpha+\alpha\tilde\beta+\tilde\beta\gamma+\tilde\beta\beta)=1,\\
&2\eta(\tilde\alpha\beta+\mu\tilde\alpha+\mu\tilde\beta)=1,\\
&2\eta(\alpha^2+\alpha\gamma+\alpha\beta+\beta\gamma+\beta^2)+2\mu\eta(\alpha+\beta+\gamma)+2(\xi+\eta)\mu^2=1.
\end{aligned}
\tag{4.10}
$$

This is a system of seven algebraic equations with eight unknowns. The system can be explicitly solved by expressing all the parameters as a function of μ. The solution is given by

$$
\begin{gathered}
\alpha=0,\quad \beta=\frac{2\mu-1}{2(\mu-1)},\quad \gamma=-\frac{2\mu^2-2\mu+1}{2\mu(\mu-1)},\\
\tilde\alpha=\frac{1}{2\mu},\quad \tilde\beta=-\frac{1}{2(\mu-1)},\quad \xi=2\mu^2-2\mu+1,\quad \eta=-2\mu(\mu-1).
\end{gathered}
$$

We shall use this freedom to satisfy the second condition (ii). When applying the scheme to the Broadwell model, the collision step (4.2) becomes

$$
z_1=\frac{z^n+\alpha(\Delta t/2\varepsilon)(\rho_1^2+m_1^2)}{1+\alpha(\Delta t/\varepsilon)\rho_1}.
$$

Therefore, for positive data both numerator and denominator are positive provided $\alpha\ge 0$. The same property holds for step (4.7), provided $\mu\ge 0$.

Step (4.4) becomes

$$
z_3=\frac{z_2+\beta(\Delta t/2\varepsilon)(\rho_2^2+m_2^2)+\gamma(\Delta t/2\varepsilon)(\rho_1^2+m_1^2-2\rho_1 z_1)}{1+\beta(\Delta t/\varepsilon)\rho_2}.
$$

The denominator is positive for positive data provided $\beta\ge 0$. We therefore look for a solution of equations (4.10) that also satisfies the restrictions

$$
\alpha\ge 0,\quad \beta>0,\quad \mu>0. \tag{4.11}
$$

A set of parameters satisfying equations (4.10) and the conditions (4.11) is given by

$$
\xi=\frac{5}{9},\ \eta=\frac{4}{9},\ \alpha=0,\ \mu=\frac{1}{3},\ \beta=\frac{1}{4},\ \gamma=\frac{5}{4},\ \tilde\alpha=\frac{3}{2},\ \tilde\beta=\frac{3}{4}. \tag{4.12}
$$

Linear stability analysis of the scheme is performed in the next section. The fluid dynamical limit of the scheme, which is required to meet condition (iv), will be studied in section 7.

The construction of the time discretization that combines the convection and the relaxation term is similar in spirit to the time discretizations used in [9] and [16]. Both time discretizations of [9] and [16] have negative parameters in the implicit terms which may cause numerical breakdown in the intermediate regime $\Delta t=O(\varepsilon)$.

Finally, our new second-order scheme (abbreviated as NSP) for the Broadwell equation (2.4)–(2.6) is (3.3)–(3.7), where R stands for the source term, $\mathcal{D}F$ is the MUSCL flux (3.13) and (3.14) and the coefficients given in (4.12).

5. Stability analysis. In this section we study the stability of scheme (4.2)–(4.7) applied to (4.8). The main result of this section is summarized in Proposition 5.2, which states that under a certain condition on the explicit step the scheme is unconditionally stable in the implicit step.

We interpret U as a complex function and A, B as complex numbers with positive real part. Equation (4.9) becomes

$$U^{n+1} = P(z_1, z_2)U^n, \tag{5.1}$$

where $z_1 = A\Delta t$, $z_2 = B\Delta t$, and P is a rational function of z_1 and z_2:

$$P = \frac{P_0(z_1) + a(1 - z_1)z_2}{(1 + \mu z_2)(1 + bz_2)}$$

with $P_0(z_1) = 1 - z_1 + z_1^2/2$, $a = \mu - 1/(2 - 2\mu)$, $b = (1 - 2\mu)/(2 - 2\mu)$. The stability region associated to scheme (5.1) is the set $\mathcal{S} \subset C^2 : \mathcal{S} = \{(z_1, z_2) : |P(z_1, z_2)| \leq 1\}$. For $B = 0$ the scheme is a second-order explicit scheme for the equation

$$U_t + AU = 0, \tag{5.2}$$

and for $A = 0$ the scheme is a second-order implicit scheme for the equation

$$U_t + BU = 0. \tag{5.3}$$

We now show that for $A = 0$ the scheme is an A-stable scheme for the equation (5.3), provided $0 < \mu < 1/2$.

It is

$$P_B(z_2) \equiv P(0, z_2) = \frac{1 + az_2}{(1 + \mu z_2)(1 + bz_2)}.$$

The scheme is A-stable if the stability region

$$S_B = \{z_2 \in C : |P(0, z_2)| < 1\}$$

contains the complex half plane $C_+ \equiv \{z \in C : \Re(z) > 0\}$. We express the stability result relative to (5.3) in the following proposition.

PROPOSITION 5.1. *If* $0 < \mu < 1/2$ *then scheme* (4.2)–(4.7) *is an L-stable scheme for equation* (5.3).

Proof. First observe that if $0 < \mu < 1/2$, then $P_B(z_2)$ has two real negative poles; therefore it is analytic in the whole half plane $\Re(z_2) \geq 0$. By the maximum principle of analytic functions, it follows that the maximum of $|P_B(z_2)|$ on any region $\Omega \subseteq C$ lies on the boundary of Ω. Because $\lim_{z_2 \to \infty} P_B(z_2) = 0$, it follows that the maximum of $|P_B(z_2)|$, $z_2 \in C_+$, lies on the imaginary axis. It is therefore enough to prove that $|P_B(iy)| \leq 1 \, \forall y \in R$. It is

$$|P_B(iy)|^2 - 1 = -\frac{\mu^2 b^2 y^4}{(1 + \mu^2 y^2)(1 + b^2 y^2)},$$

and therefore $|P_B(z_2)| \leq 1 \, \forall z_2 \in C_+$. This proves that the scheme is A-stable.

Moreover, it is $\lim_{z_2 \to \infty} P_B(z_2) = 0$, and therefore the scheme is L-stable. □

Now we set $B = 0$ and study the stability properties of the explicit scheme for the equation

$$U_t + AU = 0.$$

We have

$$P(z_1, 0) = P_0(z_1) = 1 - z_1 + \frac{1}{2}z_1^2;$$

therefore there is no dependence on μ. The stability region for z_1 is defined as

$$S_A = \{z_1 \in C : |P(z_1, 0)| \le 1\}.$$

It is convenient to write $z_1 = 1 + \rho e^{i\theta}$. Then it is $P(z_1, 0) = (1 + \rho^2 e^{2i\theta})/2$. The boundary of the stability region is obtained by the equation $|P(z_1, 0)|^2 = 1$, which gives

$$\rho^4 + 2\rho^2 \cos 2\theta - 3 = 0,$$

and therefore the two branches of the boundary are given by

$$\rho_\pm = \pm\sqrt{-\cos 2\theta + \sqrt{\cos^2 2\theta + 3}}, \quad \theta \in [0, \pi).$$

Now we consider the general problem with $z_1 \neq 0, z_2 \neq 0$. We prove the following proposition.

PROPOSITION 5.2. *If* $|z_1 - 1| < 1$ *and* $\Re(z_2) > 0$, *then* $|P(z_1, z_2)| < 1$.

Proof. We have

$$P(z_1, z_2) = \frac{P_0(z_1) + a(1 - z_1)z_2}{(1 + \mu z_2)(1 + bz_2)}.$$

If $0 < \mu < 1/2$, then $a < 0, b > 0$. For any fixed z_1, $P(z_1, z_2)$ is analytic in the half plane $\Re(z_2) \ge 0$, and it vanishes at infinity. Therefore the function has its maximum on the imaginary axis. The square of the modulus of $P(z_1, z_2)$ on the imaginary axis can be expressed in the form

$$|P(z_1, iy)|^2 = \frac{\mathcal{N}}{\mathcal{D}}$$

with

$$\mathcal{N} = 1 + 4a^2\rho^2 y^2 + \rho^4 + 4\rho a y(1 - \rho^2)\sin\theta + 2\rho^2\cos 2\theta, \tag{5.4}$$

$$\mathcal{D} = 4(1 + \mu^2 y^2)(1 + b^2 y^2), \tag{5.5}$$

and z_1 has been written as $z_1 = 1 + \rho e^{i\theta}$. We shall prove that if $\rho < 1$ then $\mathcal{D} - \mathcal{N} > 0 \ \forall y \in R$.

We have

$$\Pi \equiv \mathcal{D} - \mathcal{N} = \chi + \Lambda \sin\theta - 2\rho^2 \cos 2\theta,$$

where $\chi = 3 - \rho^4 + 4a^2(1 - \rho^2)y^2 + 4\mu^2 b^2 y^4$, $\Lambda = 4\rho a y(\rho^2 - 1)$. (We used the identity $\mu^2 + b^2 = a^2$.)

The extrema of Π as functions of θ are obtained by $\partial\Pi/\partial\theta = 0$ and are given by

(i) $\cos\theta = 0$,

(ii) $8\rho^2 \sin\theta + \Lambda = 0$.

We shall consider the two cases separately.

(i) With $\cos\theta = 0$ it is $\Pi = \Pi_\pm \equiv \chi + 2\rho^2 \pm \Lambda$. The minimum is obtained by $\Pi = \Pi_-$ if $y > 0$ and by $\Pi = \Pi_+$ if $y < 0$. Because χ is an even function of y and Λ is an odd function of y, it is enough to consider the case $\Pi = \Pi_-$, $y > 0$. It is

$$\begin{aligned}\Pi_-(\rho, y) &= 3 + 2\rho^2 - \rho^4 - 4\rho(\rho^2 - 1)ay + 4a^2(1 - \rho^2)y^2 + 4\mu^2 b^2 y^4 \\ &> \mathcal{G}(\rho^2, y) + 4\mu^2 b^2 y^4,\end{aligned}$$

where

$$\mathcal{G}(r,y) = 3 + 2r - r^2 - 4(r-1)ay + 4a^2(1-r)y^2.$$

If $0 \le r \le 1$, then $\mathcal{G}(r,y) > 0\,\forall y \in R$. In fact $\mathcal{G}(r,y) = c_0(r) + 4c_1(r)y + 4c_2(r)y^2$ and

$$\frac{\Delta^2}{16} \equiv c_1^2 - c_0c_2 = a^2(1-r)(r^2 - 2r - 2) < 0;$$

therefore $\Pi_- > 0$ and $\mathcal{D} - \mathcal{N} > 0\,\forall \rho \le 1,\ y \in R,\ \cos\theta = 0$.

(ii) For $\sin\theta = -\Lambda/(8\rho^2)$ we have

$$\begin{aligned}\Pi &= \chi - \frac{\Lambda^2}{16\rho^2} - 2\rho^2 \\ &= 3 - 2\rho^2 - \rho^4 + a^2(1-\rho^2)(3+\rho^2)y^2 + 4\mu^2b^2y^4,\end{aligned}$$

which is positive for $\rho < 1$. Therefore $\mathcal{D} - \mathcal{N} > 0\,\forall \rho < 1,\ y \in R,\ \sin\theta = -\Lambda/(8\rho^2)$.

Because Π is positive at the extrema, it is always positive for any $\rho \in [0,1)$, $\theta \in [0,\pi)$, $y \in R$, and therefore if $|z_1 - 1| < 1$ and $\Re(z_2) > 0$ it is $|P(z_1,z_2)| < 1$. □

6. An initial layer fix.

A good feature of the time discretization in [16] is that it has the correct initial layer behavior, since in the first step it projects into the local Maxwellian and in the fluid limit it becomes the second-order TVD Runge–Kutta method for the fluid equation. Our scheme here does not have such a mechanism, because as a result of truncation analysis, $\alpha = 0$ in (4.2), thus the de facto first step (4.4) is not a projection into the local Maxwellian. This will introduce an initial disturbance if the initial layer is not well resolved. Although at later time the scheme has the mechanism to project the data into the local Maxwellian, this initial error will remain at all later times, causing nonconservative numerical solutions, thus degrading the quality of the numerical results. This has been demonstrated in [16]. Hence an extra care must be taken to properly handle the initial layer.

One possibility to overcome this burden is simply to resolve the initial layer by using a time step $\Delta t \ll \varepsilon$ in the first few steps and then to switch to a larger time step. This procedure is very expensive if the problem also contains nonstiff regions where ε is not small and one does not need to resolve the initial layer.

The initial layer analysis performed in [3] indicates that the initial layer projects the initial data into the local Maxwellian. In order to have the correct behavior, the numerical scheme should have the same projection in the first time step. Note that in the first-order splitting scheme (8.2), the first step, due to its fully implicit collision term, always projects the initial data into the local Maxwellian and thus can be used for the first time step along with our new splitting scheme. The first-order accuracy of this initial step will be overcome by a Richardson-type extrapolation on (8.2).

Let us write the exact solution to (2.10) as

$$U(t) = \mathcal{S}(t)U(0), \tag{6.1}$$

where $S(t)$ is the evolutional operator and $U(0)$ is the initial datum. Suppose the difference operator is denoted by $\mathcal{T}(t)$. Then the numerical solution $u(t)$ is given by

$$u(t) = \mathcal{T}(t)U(0). \tag{6.2}$$

A first iteration of the splitting scheme (8.2) can be written as

$$u_1(\Delta t) = \mathcal{T}(\Delta t)U(0), \tag{6.3}$$

while two half-step iterations give

$$u_2(\Delta t) = \mathcal{T}\left(\frac{\Delta t}{2}\right)^2 U(0). \tag{6.4}$$

We define our first time step as

$$u(\Delta t) = 2u_2(\Delta t) - u_1(\Delta t). \tag{6.5}$$

This Richardson extrapolation will give a second-order approximation, as will be demonstrated now.

Since the splitting scheme (8.2) is globally first order (locally second order), one has

$$u_1(\Delta t) - U(\Delta t) = (\mathcal{T}(\Delta t) - \mathcal{S}(\Delta t))U(0) = C\Delta t^2 + O(\Delta t^3), \tag{6.6}$$

where C depends on $\mathcal{T}, \mathcal{S}$, and $U(0)$. Of course this implies

$$u_1\left(\frac{\Delta t}{2}\right) - U\left(\frac{\Delta t}{2}\right) = \left(\mathcal{T}\left(\frac{\Delta t}{2}\right) - \mathcal{S}\left(\frac{\Delta t}{2}\right)\right)U(0) = \frac{1}{4}C\Delta t^2 + O(\Delta t^3). \tag{6.7}$$

Also,

$$\mathcal{T}(\Delta t) = I + O(\Delta t), \tag{6.8}$$

where I is the identity operator. Now,

$$\begin{aligned}
u(\Delta t) &= 2u_2(\Delta t) - u_1(\Delta t)\\
&= 2\mathcal{T}\left(\frac{\Delta t}{2}\right)^2 U(0) - \mathcal{T}(\Delta t)U(0)\\
&= 2\mathcal{T}\left(\frac{\Delta t}{2}\right)\left[\mathcal{S}\left(\frac{\Delta t}{2}\right) + \mathcal{T}\left(\frac{\Delta t}{2}\right) - \mathcal{S}\left(\frac{\Delta t}{2}\right)\right]U(0)\\
&\quad -[\mathcal{S}(\Delta t) + \mathcal{T}(\Delta t) - \mathcal{S}(\Delta t)]U(0).
\end{aligned}$$

Making use of (6.6) one has

$$\begin{aligned}
u(\Delta t) &= 2\mathcal{T}\left(\frac{\Delta t}{2}\right)U\left(\frac{\Delta t}{2}\right) + 2\mathcal{T}\left(\frac{\Delta t}{2}\right)\left[\frac{1}{4}C\Delta t^2 + O(\Delta t^3)\right]\\
&\quad -U(\Delta t) - C\Delta t^2 + O(\Delta t^3)\\
&= 2\left[\mathcal{S}\left(\frac{\Delta t}{2}\right) + \mathcal{T}\left(\frac{\Delta t}{2}\right) - \mathcal{S}\left(\frac{\Delta t}{2}\right)\right]U\left(\frac{\Delta t}{2}\right) - U(\Delta t) - \frac{1}{2}C\Delta t^2 + O(\Delta t^3)\\
&= 2U(\Delta t) + 2\left[\frac{1}{4}C\Delta t^2 + O(\Delta t)^3\right] - U(\Delta t) - \frac{1}{2}C\Delta t^2 + O(\Delta t^3)\\
&= U(\Delta t) + O(\Delta t^3).
\end{aligned}$$

Thus we have a locally third-order, or globally second-order, approximation in this first step. In fact this second-order scheme alone can be applied to system (2.10). However, such a scheme would be slightly less efficient that the one we use, (4.2)–(4.7), since one has to integrate the convection part three times in each time step, while (4.2)–(4.7) only needs two. This Richardson extrapolation (6.5) may also lose positivity of the numerical solution.

The NSP with the first step given by the initial layer fix step described in this section will be called NSPIF.

7. The numerical fluid dynamic limits.

To show that the scheme has the correct fluid dynamic limit as $\epsilon \to 0$ (condition iv in section 4), we assume that the solution is smooth in the sense that all of the spatial derivatives are of $O(1)$. Such an assumption is necessary since the continuous Chapman–Enskog expansion, which leads from the Broadwell equation to the Euler equation, is only justified for smooth solutions [3] or weak shocks [30]. This formal analysis, although not rigorous, does provide a effective tool to understand underresolved (grid spacing does not resolve ϵ) numerical methods for hyperbolic systems with stiff relaxations [19], [17], [16], [18].

Following [19], we separate the grid spacing into three regimes. In the thin regime $\max(\Delta x, \Delta t) = o(\epsilon)$. In the intermediate regime $\max(\Delta x, \Delta t) = O(\epsilon)$. In the coarse regime $\epsilon/\max(\Delta x, \Delta t) \to 0$. In the thin regime the grid space resolves ϵ, thus the truncation error analysis in section 4 is already sufficient to show the second-order accuracy of the scheme (if the MUSCL is applied for the convection term). It is in the intermediate and coarse regimes that the fluid dynamic limit analysis will be applied.

7.1. The coarse regime.

Assume that $\varepsilon/\Delta t \ll 1, \varepsilon/\Delta x \ll 1$. The asymptotic analysis will be performed with fixed Δt and Δx while letting $\varepsilon \to 0$.

We first analyze the limiting behavior of the two spatial discretizations discussed in section 3 and defer the time discretization to later paragraphs. Here we are concerned with space discretization, and therefore we do not perform any discretization in time.

First, for the upwind discretization (3.9) of the Broadwell equation, as $\varepsilon \to 0$ the third equation of (3.9) gives the leading term approximation

$$z = \frac{1}{2\rho}(\rho^2 + m^2) + O(\varepsilon). \tag{7.1}$$

Applying this to the first two equations in (3.9) implies (after ignoring the $O(\varepsilon)$ term)

$$\partial_t \rho_j + \frac{m_{j+1} - m_{j-1}}{2\Delta x} - \frac{\frac{1}{2}(\rho + \rho u^2)_{j+1} - (\rho + \rho u^2)_j + \frac{1}{2}(\rho + \rho u^2)_{j-1}}{2\Delta x} = 0,$$

$$\partial_t m_j + \frac{\frac{1}{2}(\rho + \rho u^2)_{j+1} - \frac{1}{2}(\rho + \rho u^2)_{j-1}}{2\Delta x} - \frac{m_{j+1} - 2m_j + m_{j-1}}{2\Delta x} = 0. \tag{7.2}$$

This gives the limiting scheme of the upwind scheme. It can be easily seen that this is a first-order centered conservative discretization of the model Euler equation (2.9). Thus the upwind scheme has the correct asymptotic limit, which is only first order.

For the MUSCL discretization (3.15), as $\varepsilon \to 0$ the last equation again gives (7.1) which can be applied to the first two equations in (3.15) to yield the limiting scheme of the MUSCL:

$$\partial_t \rho_j + \frac{m_{j+1} - m_{j-1}}{2\Delta x} - \frac{\frac{1}{2}(\rho + \rho u^2)_{j+1} - (\rho + \rho u^2)_j + \frac{1}{2}(\rho + \rho u^2)_{j-1}}{2\Delta x}$$

$$+ \frac{\Delta x}{4}(\sigma_j^{E+} + \sigma_{j+1}^{E-} - \sigma_{j-1}^{E+} - \sigma_j^{E-}) = 0, \tag{7.3}$$

$$\partial_t m_j + \frac{\frac{1}{2}(\rho + \rho u^2)_{j+1} - \frac{1}{2}(\rho + \rho u^2)_{j-1}}{2\Delta x} - \frac{m_{j+1} - 2m_j + m_{j-1}}{2\Delta x}$$

$$+ \frac{\Delta x}{4}(\sigma_j^{E+} - \sigma_{j+1}^{E-} - \sigma_{j-1}^{E+} + \sigma_j^{E-}) = 0, \tag{7.4}$$

where

$$\sigma_j^{E\pm} = \sigma_j^{\pm}|_{z=\frac{1}{2\rho}(\rho^2+m^2)}. \tag{7.5}$$

Note that $\sigma_j^{\pm}$ is defined by ρ, m, and z, and the right-hand side of (7.5) simply means applying the local Maxwellian $z = (\rho^2+m^2)/(2\rho)$ into $\sigma_j^{\pm}$. Define

$$w^{E+} = f|_{z=\frac{1}{2\rho}(\rho^2+m^2)} = \frac{1}{2}(z+m)\Big|_{z=\frac{1}{2\rho}(\rho^2+m^2)} = \frac{1}{2}\left(\frac{1}{2\rho}(\rho^2+m^2)+m\right), \tag{7.6}$$

$$w^{E-} = g|_{z=\frac{1}{2\rho}(\rho^2+m^2)} = \frac{1}{2}(z-m)\Big|_{z=\frac{1}{2\rho}(\rho^2+m^2)} = \frac{1}{2}\left(\frac{1}{2\rho}(\rho^2+m^2)-m\right). \tag{7.7}$$

Then from (3.11),

$$\sigma_j^{E\pm} = \frac{1}{\Delta x}(w_{j+1}^{E\pm} - w_{j-1}^{E\pm})\phi(\theta^{E\pm}), \quad \theta^{E\pm} = \frac{w_j^{E\pm} - w_{j-1}^{E\pm}}{w_{j+1}^{E\pm} - w_j^{E\pm}}. \tag{7.8}$$

One can see that this is a second-order centered conservative discretization of the model equations (2.9). Thus the MUSCL also has the correct fluid limit, which is second order to (2.9).

We now demonstrate that the new temporal splitting scheme (4.2)–(4.7) has the correct fluid limit in the sense to be specified below. We always assume that $R(U)$ is Lipschitz continuous in U and $I+\beta\frac{\Delta t}{\varepsilon}R'$ is invertible for all ε and Δt, where R' is the Jacobian matrix of R. This invertibility is true for general hyperbolic systems with relaxations classified by Liu [22]. In particular, it is easy to check that the collision term of the Broadwell equations satisfies this condition.

By the initial layer fix, we can assume that

$$R(U^n) = O(\varepsilon). \tag{7.9}$$

Now, we want to show that this condition will imply that

$$R(U_1) \approx 0, \quad R(U_3) \approx 0, \quad R(U^{n+1}) \approx 0, \tag{7.10}$$

up to some error terms depending on ε and Δt. First, since $\alpha = 0$, then $U_1 = U^n$ and

$$R(U_1) = R(U^n) = O(\varepsilon). \tag{7.11}$$

Using (4.3),

$$U_2 - U_1 = O(\Delta t). \tag{7.12}$$

By (4.4),

$$\begin{aligned} U_3 - U_2 &= -\beta\frac{\Delta t}{\varepsilon}(R(U_3)-R(U_2)) - \beta\frac{\Delta t}{\varepsilon}R(U_2) + O(\Delta t) \\ &= -\beta\frac{\Delta t}{\varepsilon}R'(U^*)(U_3-U_2) - \beta\frac{\Delta t}{\varepsilon}(R(U_1)+O(\Delta t)) + O(\Delta t) \\ &= -\beta\frac{\Delta t}{\varepsilon}R'(U^*)(U_3-U_2) + O\left(\Delta t + \frac{\Delta t^2}{\varepsilon}\right), \end{aligned} \tag{7.13}$$

where U^* lies between U_2 and U_3. This implies

$$U_3 - U_2 = \left(I + \beta\frac{\Delta t}{\varepsilon}R'(U^*)\right)^{-1} O\left(\Delta t + \frac{\Delta t^2}{\varepsilon}\right) \tag{7.14}$$

$$= O\left(\frac{\varepsilon}{\Delta t}\right) O\left(\Delta t + \frac{\Delta t^2}{\varepsilon}\right) = O(\varepsilon + \Delta t) = O(\Delta t). \tag{7.15}$$

Applying this back to (4.4) then implies

$$R(U_3) = O(\varepsilon). \tag{7.16}$$

Similar arguments also imply

$$R(U_{n+1}) = O(\varepsilon). \tag{7.17}$$

By this argument, for any initial data with the initial layer fixing one always has

$$R(U_3) = O(\varepsilon), \qquad R(U^n) = O(\varepsilon) \quad \text{for all} \quad n \geq 1. \tag{7.18}$$

Thus this scheme projects the numerical data into the local Maxwellian at every time step. This completes the study of the fluid dynamic limit of the new splitting scheme.

Finally, we combine the spatial and temporal discussions above to get the limiting scheme for the fully discrete method that combines the MUSCL and the new splitting scheme. Applying (7.18) into (4.3), (4.5), respectively, after reordering the indices, we get

$$U_1 = U^n - \tilde{\alpha}\Delta t \mathcal{D}F(U^n)|_{R(U^n)=0}, \tag{7.19}$$

$$U_2 = U_1 - \tilde{\beta}\Delta t \mathcal{D}F(U_1)|_{R(U_1)=0}, \tag{7.20}$$

$$U^{n+1} = \xi U^n + \eta U_2 \tag{7.21}$$

by ignoring the $O(\varepsilon)$ terms. In (7.19) and (7.20), $\mathcal{D}F(U)|_{R(U)=0}$ are the limiting spatial discretization given in (7.3) and (7.4) for the model Euler equation (2.9). Clearly (7.21) is a second-order (both spatially and temporally) discretization of the model Euler equation (2.9). This shows that the new splitting scheme combined with MUSCL flux indeed has the correct fluid limit in the coarse regime.

If the upwind scheme is used in the convection term, then the limiting flux $\mathcal{D}F(U)|_{R(U)=0}$ is given in (7.2), which is a first-order spatial discretization of the model Euler equation.

7.2. The intermediate regime. The intermediate regime is defined as $\varepsilon \ll 1, \Delta t \ll 1$ but $\varepsilon/\Delta t = O(1)$. We now show that the splitting scheme (4.2)–(4.7) also has the correct fluid limit in this intermediate regime. The key is to demonstrate that (7.10) are valid also in this regime.

First, since $\alpha = 0$, after the initial layer fix (the assumption on (7.9)), (7.11) is valid. Thus (4.4) implies

$$\begin{aligned} U_3 - U_2 &= -\beta\frac{\Delta t}{\varepsilon}R(U_3) + O(\varepsilon) \\ &= -\beta\frac{\Delta t}{\varepsilon}(R(U_3) - R(U_2)) - \beta\frac{\Delta t}{\varepsilon}R(U_2) + O(\varepsilon) \\ &= -\beta\frac{\Delta t}{\varepsilon}R'(U^*)(U_3 - U_2) - \beta\frac{\Delta t}{\varepsilon}(R(U_1) + O(\Delta t)) + O(\varepsilon) \\ &= -\beta\frac{\Delta t}{\varepsilon}R'(U^*)(U_3 - U_2) + O(\Delta t). \end{aligned}$$

Therefore

$$U_3 - U_2 = \left(1 + \beta \frac{\Delta t}{\varepsilon} R'(U^*)\right)^{-1} O(\Delta t) = O(\Delta t).$$

Applying this to (4.4) gives

$$R(U_3) = O(\Delta t) O\left(\frac{\varepsilon}{\Delta t}\right) = O(\varepsilon).$$

Similarly, one can also show

$$R(U^{n+1}) = O(\varepsilon).$$

Thus, one always has

$$R(U_3) = O(\varepsilon), \qquad R(U^n) = O(\varepsilon) \quad \text{for all } n \geq 1$$

independent of the initial data. So the solution is always a local Maxwellian. Moreover, as $\varepsilon \to 0$ ($\Delta t \to 0$ as well) one gets (2.9). The limiting fully discrete discretization is in fact the same as in (7.19) and (7.20). Thus in the intermediate regime this scheme also has the correct fluid limit.

Remark. In [24] Pember found that in the limit $\epsilon \to 0$ a second-order space discretization may be reduced to first order in a fractional step method. A high-order method needs to incorporate the effect of the source term into the convective flux. Jin [16] also showed such a phenomenon in a Strang's splitting. However, as indicated in [16], a *properly* designed fractional step method may maintain a second-order accuracy even if the convective flux is based *solely* on the Riemann problem of the homogeneous system. The result in this section confirms this for the new splitting scheme NSPIF, and our numerical results in section 9 seem to verify this point.

8. Theoretical results. In this section we shall provide some analytic analyses on a simple splitting scheme (SP1):

$$\tilde{\rho}_j = \rho_j^n, \quad \tilde{m}_j = m_j^n, \quad \tilde{z}_j = z_j^n + \frac{\Delta t}{2\varepsilon}\left((\rho_j^n)^2 + (m_j^n)^2 - 2\rho_j^n \tilde{z}_j\right); \tag{8.1}$$

$$f_j^{n+1} = \tilde{f}_{j-1}^n, \quad h_j^{n+1} = \tilde{h}_j^n, \quad g_j^{n+1} = \tilde{g}_{j+1}. \tag{8.2}$$

This simple first-order method, although significantly simpler compared with NSPIF, does possess some key properties of the NSPIF, such as A-stability, L-stability, and the correct fluid limit as $\epsilon \to 0$, and thus it serves as an analytic model here. Analytic study of the second-order method NSPIF would require more advanced techniques that are out of reach for this work.

In SP1, the collision step is the backward Euler method that has a numerical stability independent of ε. In this step it is very convenient to use the fluid variables. In the convection step, the characteristic method can be applied for the free stream. The overall Courant–Friedrichs–Lewy (CFL) number then will be solely determined by the convection step, which is 1. In (8.2) we denote the intermediate result after the collision step by a tilde. Between the collision and the convection steps the relation (2.2) and (2.3) will be used. This method is going to have the correct fluid limit when $\varepsilon \to 0$ since the first step always yields the correct local Maxwellian $z = (\rho^2 + m^2)/(2\rho)$. Applying it to the next step, one gets a first-order approximation to the model Euler equations. However, the overall accuracy is first order, uniformly in ε, as will be shown later.

In this section we derive several properties of the SP1, including positivity, the entropy inequality, and a bound in the consistency error which is independent on the mean free path. To the best of our knowledge this is the first uniform (in ϵ) result for the numerical discretization of the Broadwell equation.

8.1. Positivity. The continuous Broadwell equations maintain positivity of the densities f, h, and g. This property is important because of the physical meaning of f, h, and g as particle density. Such property is maintained by the discrete scheme.

The velocity densities (f, g, h) are positive if and only if

$$\rho > z > |m|$$

by (2.2) and (2.3). In the collision step $\tilde{\rho} = \rho^n$ and $\tilde{m} = m^n$ are unchanged, while

$$\tilde{z} = (1 + 2\kappa\rho^n)^{-1}(z^n + \kappa((m^n)^2 + (\rho^n)^2)),$$

in which $\kappa = \Delta t/(2\varepsilon)$. Since $\tilde{\rho} = \rho^n > z^n > |m^n| = |\tilde{m}|$, it follows easily that

$$\tilde{\rho} > \tilde{z} > |\tilde{m}|.$$

In fact,

$$\tilde{z} - |\tilde{m}| = \tilde{z} - |m| = \frac{z - |m| + \kappa(\rho - |m|)^2}{1 + 2\kappa\rho} > 0$$

and

$$\tilde{\rho} - \tilde{z} = \rho - \tilde{z} = \frac{\rho - z + \kappa(\rho^2 - m^2)}{1 + 2\kappa\rho} > 0.$$

So $(\tilde{f}, \tilde{g}, \tilde{h})$ is positive. Positivity of $(\tilde{f}, \tilde{g}, \tilde{h})$ clearly implies positivity of f^{n+1}, g^{n+1}, and h^{n+1}, the result of the convective step in (8.2).

8.2. Entropy inequality. For the system of Broadwell equations (with periodic boundary conditions or in unbounded domain) there exists a function that is monotonically increasing and is constant only at equilibrium:

$$\mathcal{H}(t) = \int \hat{H}(x,t)\,dx,$$

with

$$\hat{H}(x,t) = f(x,t)\log f(x,t) + 2h(x,t)\log h(x,t) + g(x,t)\log g(x,t).$$

The proof of this property (the so-called H-theorem) is very simple in this case, and we show it here (in the case of periodic boundary conditions) for completeness.

Making use of the Broadwell equations and of the periodic boundary conditions one has

$$\frac{d\mathcal{H}}{dt} = \int h^2 \left(1 - \frac{fg}{h^2}\right) \log \frac{fg}{h^2}\,dx.$$

The right-hand side is negative, since $(1 - x)\log x \leq 0\ \forall x > 0$, and it is zero if and only if $fg - h^2 \equiv 0$.

The entropy inequality has the meaning of irreversibility in time and shows that the system relaxes to an equilibrium given by a Maxwellian distribution.

We shall prove that such property is maintained by the simple splitting scheme (8.2). This property is relevant, first because it shows that the discrete model relaxes to equilibrium and second because of the connection between the H-function and the entropy of the fluid dynamic limit.

Before proving the H-theorem for the discrete scheme, we shall exploit this connection in detail.

The kinetic model is described by the Broadwell equations, which constitute a 3×3 semilinear system of partial differential equations, while the fluid dynamic limit is described by a 2×2 quasi-linear hyperbolic system of conservation laws. In both cases the entropy inequality has the physical meaning of irreversibility in time. For the quasi-linear system, the entropy principle is used to select the unique weak solution in the presence of shocks.

Now let us consider an $n \times n$ system of conservation laws of the form

$$\frac{\partial U}{\partial t} + \frac{\partial F(U)}{\partial x} = 0. \tag{8.3}$$

An entropy for this system is a convex function $s = s(U)$ which, in case of regular solutions, satisfies an additional conservation law of the form [11]

$$\frac{\partial s(U)}{\partial t} + \frac{\partial q(U)}{\partial x} = 0, \tag{8.4}$$

where $q(U)$ is the entropy flux. By comparing (8.3) and (8.4), the following compatibility condition holds:

$$\frac{\partial q}{\partial u_\beta} = \sum_{\alpha=1}^{n} \frac{\partial s}{\partial u_\alpha} \frac{\partial F^\alpha}{\partial u_\beta}, \quad \beta = 1, \dots, n.$$

We now show that the function $\hat{H}$, when computed at a local Maxwellian (denoted by the suffix M), i.e., if $fg = h^2$, and expressed in terms of the fluid variables ρ and u, is indeed an entropy function for the 2×2 system of fluid dynamic limit.

At equilibrium one has

$$f_M = \frac{\rho}{4}(1+u)^2, \quad g_M = \frac{\rho}{4}(1-u)^2, \quad h_M = \frac{\rho}{4}(1-u^2).$$

Substituting in the expression of $\hat{H}$, after some simplifications one has

$$H_M = \rho(\log \rho + S(u)), \tag{8.5}$$

with

$$S(u) \equiv (1+u)\log(1+u) + (1-u)\log(1-u) - \log 4. \tag{8.6}$$

Let us compute $\partial H_M / \partial t$:

$$\frac{\partial H_M}{\partial t} = \frac{\partial H_M}{\partial \rho}\frac{\partial \rho}{\partial t} + \frac{\partial H_M}{\partial u}\frac{\partial u}{\partial t}.$$

Making use of expression (8.5) and of the equations (2.9) for ρ and u one has

$$\frac{\partial H_M}{\partial t} = -\rho[\log \rho + S(u) + 1]\frac{\partial u}{\partial x} - \left[(\log \rho + S(u) + 1)u + \frac{1}{2}S'(u)(1-u^2)\right]\frac{\partial \rho}{\partial x}.$$

Compatibility with (8.4) requires that

$$\frac{\partial q}{\partial u} = f_1(\rho, u) \equiv \rho[\log \rho + S(u) + 1],$$

$$\frac{\partial q}{\partial \rho} = f_2(\rho, u) \equiv (\log \rho + S(u) + 1)u + \frac{1}{2} S'(u)(1 - u^2).$$

These conditions can be satisfied by a function $q(\rho, u)$ provided

$$\frac{\partial f_1}{\partial \rho} = \frac{\partial f_2}{\partial u}.$$

This last condition becomes

$$1 - \frac{1}{2} S''(u)(1 - u^2) = 0,$$

which is satisfied by $S(u)$, defined in (8.6). This proves that $H_M(\rho, u)$ is indeed an entropy function for the system of equations describing the fluid dynamic limit.

It is evident that a scheme that has the property of preserving the entropy inequality for the Broadwell model for any value of the relaxation parameter ε will also preserve the entropy inequality in the fluid dynamic limit. This property is essential in approximating the correct weak solution of the fluid dynamic equations.

We prove next that the H function given by

$$H = f \log f + 2h \log h + g \log g$$

decreases in the collision step at each grid point j. This result is valid both for periodic boundary conditions and for unbounded domains.

Let $(\tilde{\rho}, \tilde{m}, \tilde{z})$ be the state after the collision step and $(\rho, m, z) = (\tilde{\rho}, \tilde{m}, \tilde{z} + 2\Delta)$ be the state before the collision step, in which

$$\Delta = 2\kappa(\tilde{h}^2 - \tilde{f}\tilde{g}).$$

Since $f = (z + m)/2,\ g = (z - m)/2$, and $h = (\rho - z)/2$, then

$$f = \tilde{f} - \Delta, \quad g = \tilde{g} - \Delta, \quad h = \tilde{h} + \Delta. \tag{8.7}$$

Moreover since (f, g, h) and $(\tilde{f}, \tilde{g}, \tilde{h})$ are all positive, then

$$1 > \max\left(\frac{\Delta}{\tilde{f}}, \frac{\Delta}{\tilde{g}}, -\frac{\Delta}{\tilde{h}}\right).$$

Before the collision the H function is

$$H = f \log f + 2h \log h + g \log g.$$

First rewrite the f term in H as

$$\begin{aligned} f \log f &= (\tilde{f} - \Delta) \log(\tilde{f} - \Delta) \\ &= \tilde{f}\left(1 - \frac{\Delta}{\tilde{f}}\right) \log\left(1 - \frac{\Delta}{\tilde{f}}\right) + (\tilde{f} - \Delta) \log \tilde{f} \\ &\geq -\Delta + (\tilde{f} - \Delta) \log \tilde{f}. \end{aligned}$$

The last bound uses the elementary inequality

$$(1-a)\log(1-a) \geq -a$$

for $1 > a$ in which $a = \Delta/\tilde{f}$. Similarly,

$$g\log g \geq -\Delta + (\tilde{g}-\Delta)\log\tilde{g}, \quad h\log h \geq \Delta + (\tilde{h}+\Delta)\log\tilde{h}. \tag{8.8}$$

Add these together to obtain

$$\begin{aligned} H &\geq (\tilde{f}-\Delta)\log\tilde{f} + 2(\tilde{h}+\Delta)\log\tilde{h} + (\tilde{g}-\Delta)\log\tilde{g} \\ &= \tilde{H} + \Delta\log(\tilde{h}^2/\tilde{f}\tilde{g}) \\ &= \tilde{H} + 2\kappa(\tilde{h}^2 - \tilde{f}\tilde{g})\log(\tilde{h}^2/\tilde{f}\tilde{g}) \\ &\geq \tilde{H}. \end{aligned}$$

Finally, note that the total "entropy"

$$\mathcal{H}^n = \Delta x \sum_j H_j^n$$

is preserved by the convection step of the fractional step method (8.2), since the values of f, g, and h do not change but only move between j points.

The results of these two subsections are summarized by the following proposition.

PROPOSITION 8.1. *Let (f_j^n, g_j^n, h_j^n) be the solution of the fractional step method* (8.2). *Assume that $f_j^0 > 0,\ g_j^0 > 0$, and $h_j^0 > 0$ for all j. It follows that*

(i) $f_j^n > 0, g_j^n > 0, h_j^n > 0$ *for all $n > 0$ and all j;*

(ii) *we define*

$$\mathcal{H}^n = \sum_j (f_j^n \log f_j^n + 2h_j^n \log h_j^n + g_j^n \log g_j^n)$$

and suppose that $\mathcal{H}^0 < \infty$. Then for all $n \geq 0$

$$\mathcal{H}^{n+1} \leq \mathcal{H}^n.$$

8.3. Formal analysis of convergence: Uniform bounds on consistency error. Finally, we present a formal analysis of the splitting method which shows that it is first-order accurate uniformly in ε if the underlying Broadwell solution is smooth. This analysis does not apply for the solution in a boundary layer, initial layer, or shock in which numerical results indicate that a half order of accuracy may be lost. The analysis here is only of the consistency error, which is shown to be of size $O(\Delta t)$ uniformly in ε. Analysis of the stability error has not yet been successful.

Because of the difficulty of the problem, no result is available for higher order schemes. To our knowledge, this is the first analytic result showing a bound on the consistency error, uniformly in the relaxation parameter, for such a problem.

The estimate will be performed by writing the Broadwell equations and the discretized Broadwell equations for the fluid dynamic variables ρ and m and for the difference from equilibrium w, defined by

$$w = z - \frac{\rho^2 + m^2}{2\rho}. \tag{8.9}$$

Define also

$$z_E(\rho, m) = \frac{\rho^2 + m^2}{2\rho}. \tag{8.10}$$

Whereas the equation for z involves a forcing term of size ε^{-1}, in the w equation the factor ε^{-1} appears only in a decay term.

The Broadwell equations (2.1) can be written in terms of ρ, m, and w as

$$\rho_t = -m_x, \tag{8.11}$$

$$m_t = -(w + z_E(\rho, m))_x, \tag{8.12}$$

$$w_t = -\varepsilon^{-1}\rho w - m_x - z_E(\rho, m)_t. \tag{8.13}$$

The equation for w can be written in integral form for each x as

$$w(t) = A(t)w^0 + \int_0^t \frac{A(t)}{A(s)} g(s) ds, \tag{8.14}$$

in which

$$A(t) = \exp\left(-\varepsilon^{-1}\int_0^t \rho(t', x)dt'\right)$$

$$g(t) = -(m_x + z_E(\rho, m)_t).$$

Define the discretized variables $\tilde{\rho}, \tilde{m}$, and $\tilde{w}$ as in (3.2), (3.3) to be the state after a collision step. The discrete Broadwell equations (with $\Delta t = \Delta x$) can then be written with the convection and collision steps combined for each spatial value j as

$$\tilde{\rho}^{n+1} - \tilde{\rho}^n = -\frac{1}{2}D_0\tilde{m}^n + \frac{1}{2}D^2(\tilde{w}^n + \tilde{z}_E^n), \tag{8.15}$$

$$\tilde{m}^{n+1} - \tilde{m}^n = -\frac{1}{2}D_0(\tilde{w}^n + \tilde{z}_E^n) + \frac{1}{2}D^2\tilde{m}^n, \tag{8.16}$$

$$\tilde{w}^{n+1} = \left(1 + \frac{\Delta t}{\varepsilon}\tilde{\rho}^{n+1}\right)^{-1}(\tilde{w}^n + \Delta t\tilde{g}^{n+1}), \tag{8.17}$$

in which

$$\tilde{g}^{n+1} = -\Delta t^{-1}\left(\tilde{z}_E^{n+1} - \tilde{z}_E^n + \frac{1}{2}D_0\tilde{m}^{n+1} - \frac{1}{2}D^2(\tilde{w}^{n+1} + \tilde{z}_E^{n+1})\right).$$

The difference operators D_0 and D^2 are defined as

$$(D_0 f)_j = f_{j+1} - f_{j-1}, \quad (D^2 f)_j = f_{j+1} - 2f_j + f_{j-1}. \tag{8.18}$$

The equation for $\tilde{w}^{n+1}$ can be considered a linear difference equation in $\tilde{w^n}$. The solution can be written as

$$\tilde{w}^n = \tilde{B}^n w^0 + \Delta t \sum_{k=1}^n \frac{\tilde{B}^n}{\tilde{B}^{k-1}}\tilde{g}^k, \tag{8.19}$$

in which

$$\tilde{B}^n = \prod_{j=1}^n \left(1 + \frac{\Delta t}{\varepsilon}\tilde{\rho}^j\right)^{-1}, \quad \tilde{B}^0 = 1. \tag{8.20}$$

Assume that ρ, m, and w are smooth and bounded uniformly in ε (this excludes shocks, boundary layers, and initial layers). Define the continuous solution at discrete times as $(\rho^n, m^n, w^n) = (\rho(n\Delta t), m(n\Delta t), w(n\Delta t))$. This is an abuse of notation, since (ρ^n, m^n, w^n) was used in section 3 to denote the discretized solution after the convection step. Nevertheless, it will be used for simplicity.

The consistency error is defined as the error formed by substituting the continuous solution $(\rho^n, m^n, w^n) = (\rho(n\Delta t), m(n\Delta t), w(n\Delta t))$ into the discrete equations. Define

$$E_1 = (\rho^{n+1} - \rho^n)/\Delta t - \left\{-\frac{1}{2}D_0 m^n + \frac{1}{2}D^2(w^n + z_E^n)\right\} \Big/ \Delta t, \tag{8.21}$$

$$E_2 = (m^{n+1} - m^n)/\Delta t - \left\{-\frac{1}{2}D_0(w^n + z_E^n) + \frac{1}{2}D^2 m^n\right\} \Big/ \Delta t, \tag{8.22}$$

$$E_3 = w^n - \left\{B^n w^0 + \Delta t \sum_{k=1}^{n} \frac{B^n}{B^{k-1}} g^k\right\}, \tag{8.23}$$

in which B^n and g^n are defined as above in terms of the solution (ρ^n, m^n, w^n). Note that the consistency error for the w equations is defined here in terms of the "integrated solution" rather than the finite difference equation. In fact the error due to the implicit collision operator is better behaved over many time steps than over a single step. This is also the reason that the factor $1/\Delta t$ does not appear in the definition of E_3.

The consistency error serves as a forcing term in a finite difference equation for the total error. Define the difference between the numerical solution $(\tilde{\rho}^n, \tilde{m}^n, \tilde{w}^n)$ and the continuous solution $(\rho^n, m^n, w^n) = (\rho, m, w)(n\Delta t)$ as

$$a^n = \tilde{\rho}^n - \rho^n, \quad b^n = \tilde{m}^n - m^n, \quad c^n = \tilde{w}^n - w^n. \tag{8.24}$$

Denote $A(\rho^n, m^n, w^n)$ and $B(\rho^n, m^n, w^n)$ to be the bracketed convection terms in the ρ and m equations (8.21) and (8.22). Also denote $C[\rho, m, w]$ to be the bracketed term in the w equation (8.23). Then the error quantities (a^n, b^n, c^n) satisfy

$$\begin{aligned} a^{n+1} - a^n &= (A(\rho^n + a^n, m^n + b^n, w^n + c^n) - A(\rho^n, m^n, w^n)) + \Delta t E_1, \\ b^{n+1} - b^n &= (B(\rho^n + a^n, m^n + b^n, w^n + c^n) - B(\rho^n, m^n, w^n)) + \Delta t E_2, \\ c^{n+1} - c^n &= (C[\rho + a, m + b, w + c] - C[\rho, m, w]) + E_3. \end{aligned} \tag{8.25}$$

The first term on the right-hand side of these equations is the stability error and is approximately a linear operator in (a, b, c). If the difference scheme is uniformly stable, which we have not proved, then the error will be of the size of the consistency error E_1, E_2, E_3.

Under the assumption that the continuous solution ρ, m, w is smooth and that the density ρ is uniformly bounded above and below, i.e.,

$$\bar{\rho} < \rho < \bar{c}\bar{\rho}, \tag{8.26}$$

we prove the following uniform bound on the consistency error.

PROPOSITION 8.2. *Suppose that ρ, m, w is smooth and that the density ρ satisfies* (8.26) *for some constants $\bar{\rho}$ and $\bar{c}$. Then the consistency error E_1, E_2, E_3, defined by* (8.21), (8.22), *and* (8.23), *for the discretized Broadwell equations* (8.15), (8.16), *and* (8.19) *with $\Delta t = \Delta x$ satisfies*

$$|E_1| + |E_2| + |E_3| \leq c\Delta t \tag{8.27}$$

for some constant c that is independent of ε. In other words, the consistency error is uniformly first order.

The estimates on E_1 and E_2 are straightforward; i.e., for some constant c,

$$|E_1| + |E_2| \le c\Delta t. \tag{8.28}$$

In order to compare the integral (8.14) with the sum (8.23), we first replace the integrand by a piecewise-constant function. Define

$$\bar{A}(s) = A((m-1)\Delta t), \quad \bar{g}(s) = g((m-1)\Delta t) \tag{8.29}$$

for $(m-1)\Delta t \le s < m\Delta t$. For $t = n\Delta t$ define

$$\begin{aligned} \bar{w}(t) &= A(t)w^0 + \int_0^t \frac{A(t)}{\bar{A}(s)}\bar{g}(s)ds \\ &= A(t)w^0 + \Delta t \sum_{k=1}^{n} \frac{A(n\Delta t)}{A((k-1)\Delta t)} g(k\Delta t). \end{aligned} \tag{8.30}$$

Now estimate for $(m-1)\Delta t \le s < m\Delta t$, $m \le n$,

$$|g(s) - \bar{g}(s)| < c\Delta t, \tag{8.31}$$

$$\begin{aligned} \left|\frac{A(t)}{A(s)} - \frac{A(t)}{\bar{A}(s)}\right| &= \left|\frac{A(t)}{A(s)}\right| \cdot \left|1 - \frac{A(s)}{\bar{A}(s)}\right| \\ &\le \exp\left(-\int_s^t \rho(s')ds'/\varepsilon\right) \\ &\quad \left|1 - \exp\left(-\int_{(m-1)\Delta t}^{s} \rho(s')ds'/\varepsilon\right)\right| \\ &\le \min(1, \bar{c}\Delta t\bar{\rho}/\varepsilon)\exp(-(n-m)\Delta t\bar{\rho}/\varepsilon) \end{aligned} \tag{8.32}$$

since $|e^{-\alpha} - 1| \le \min(1, \alpha)$ for any α.

Next estimate the difference between $A(t)$ and B^n. First make a simple general estimate (for $t = n\Delta t$)

$$A(t) \le e^{-t\bar{\rho}/\varepsilon} \le \left(1 + \frac{\Delta t}{\varepsilon}\bar{\rho}\right)^{-n}, \tag{8.33}$$

$$B^n \le \left(1 + \frac{\Delta t}{\varepsilon}\bar{\rho}\right)^{-n}, \tag{8.34}$$

so that

$$|A(t) - B^n| \le 2\left(1 + \frac{\Delta t}{\varepsilon}\bar{\rho}\right)^{-n}. \tag{8.35}$$

If $\Delta t\bar{\rho}/\varepsilon$ is small, a more refined estimate is needed. In this case

$$\log\left(1 + \frac{\Delta t}{\varepsilon}\rho^k\right) = \frac{\Delta t}{\varepsilon}\rho^k + O\left(\frac{\Delta t}{\varepsilon}\bar{\rho}\right)^2. \tag{8.36}$$

Then

$$A(t) - B^n = A(t)\left\{1 - \exp\left(\int_0^t \rho(s)ds/\varepsilon - \sum_{k=1}^{n} \log\left(1 + \frac{\Delta t}{\varepsilon}\rho^k\right)\right)\right\}$$

$$= A(t)\left\{1 - \exp\left(nO\left(\frac{\Delta t}{\varepsilon}\bar{\rho}\right)^2\right)\right\}, \tag{8.37}$$

in which we ignore a term of size $n\,O(\Delta t^2\bar{\rho}/\varepsilon)$ in the exponential.

Now $n(\Delta t\bar{\rho}/\varepsilon)^2 = (t\bar{\rho}/\varepsilon)(\Delta t\bar{\rho}/\varepsilon)$. If this is larger than two, we use the estimate (8.35); if it is less than two, then

$$|A(t) - B^n| \le A(t)O\left(\frac{t\bar{\rho}}{\varepsilon}\left(\frac{\Delta t}{\varepsilon}\bar{\rho}\right)\right). \tag{8.38}$$

Combining these estimates, we obtain the bound

$$|A(t) - B^n| \le \min\left(2, \frac{t\Delta t}{\varepsilon^2}\bar{\rho}^2\right)\left(1 + \frac{\Delta t}{\varepsilon}\bar{\rho}\right)^{-n}. \tag{8.39}$$

Finally, the estimate on $A(t)/A((k-1)\Delta t) - B^n/B^{k-1}$ is found in a similar way to be

$$\left|\frac{A(t)}{A((k-1)\Delta t)} - \frac{B^n}{B^{k-1}}\right| \le \min\left(2, \varepsilon^{-2}(n-k+1)\Delta t^2\bar{\rho}^2\right) \cdot \left(1 + \frac{\Delta t}{\varepsilon}\bar{\rho}\right)^{-(n-k)}. \tag{8.40}$$

Now these basic estimates are combined to estimate E_3. First

$$\begin{aligned} w(t) - \bar{w}(t) &= \int_0^t \left(\frac{A(t)}{A(s)}g(s) - \frac{A(t)}{\bar{A}(s)}\bar{g}(s)\right) ds \\ &\le \int_0^t \left(\left|\frac{A(t)}{A(s)} - \frac{A(t)}{\bar{A}(s)}\right| |\bar{g}(s)| + \left|\frac{A(t)}{A(s)}\right| |g(s) - \bar{g}(s)|\right) ds \\ &\le c\Delta t \sum_{m=1}^{n} \min(1, \bar{c}\Delta t\bar{\rho}\varepsilon^{-1}) \exp(-(n-m)\Delta t\bar{\rho}/\varepsilon) \\ &\quad + c\Delta t \int_0^t e^{-\bar{\rho}(t-s)/\varepsilon} ds \\ &\le c\Delta t \min(1, \Delta t\bar{\rho}\varepsilon^{-1})(1 - e^{-\Delta t\bar{\rho}/\varepsilon})^{-1} + c\varepsilon\Delta t/\bar{\rho} \\ &\le \tilde{c}\Delta t \end{aligned} \tag{8.41}$$

since

$$\max_{y>0}\left(\frac{\min(1, y)}{1 - \exp(-y)}\right) = \frac{e}{e-1}.$$

Next, making use of (8.39),

$$\begin{aligned}|E_3+\bar{w}(t)-w^n| &= \Bigg|(A(t)-B^n)w^0 \\ &\quad + \Delta t\sum_{k=1}^{n}\left(\frac{A(t)}{A((k-1)\Delta t)}\bar{g}(k\Delta t)-\frac{B^n}{B^{k-1}}g^k\right)\Bigg| \\ &\le |w^0|\min\left(2,\frac{t\Delta t}{\varepsilon^2}\bar{\rho}^2\right)\left(1+\frac{\Delta t}{\varepsilon}\bar{\rho}\right)^{-n} \\ &\quad + c\Delta t\sum_{k=1}^{n}\left(1+\frac{\Delta t}{\varepsilon}\bar{\rho}\right)^{-(n-k+1)}\Delta t \\ &\quad + c\Delta t\sum_{k=1}^{n}\min\left(1,\frac{\Delta t^2\bar{\rho}^2}{\varepsilon^2}(n-k+1)\right)\left(1+\frac{\bar{\rho}\Delta t}{\varepsilon}\right)^{-n+k-1}. \qquad (8.42)\end{aligned}$$

Denote the three terms on the right side of (8.42) as D_1, D_2, D_3, respectively.

The first term D_1 is bounded by

$$c\min\left(2,\frac{n\Delta t}{\varepsilon}\bar{\rho}\frac{\Delta t\bar{\rho}}{\varepsilon}\right)\left(1+\frac{\Delta t}{\varepsilon}\bar{\rho}\right)^{-n}\le 2c\Delta t. \qquad (8.43)$$

In fact, if $\Delta t\bar{\rho}/\varepsilon > 1/2$, then

$$D_1 < 2c(3/2)^{-n} < 2c\Delta t \qquad (8.44)$$

while if $\Delta t\bar{\rho}/\varepsilon < 1/2$ then

$$D_1 < c(1/n)(n\Delta t\bar{\rho}/\varepsilon)^2(1+\Delta t\bar{\rho}/\varepsilon)^{-n} < c/n < c\Delta t, \qquad (8.45)$$

since $x^2(1+x/n)^{-n}$ is bounded by 1 uniformly in $n > 2$ and $x \ge 0$.

Now we consider the second term D_2. Let $q=(1+\bar{\rho}\Delta t/\varepsilon)^{-1}$. Then

$$\begin{aligned}D_2 &= c\Delta t^2\sum_{k=1}^{\infty}q^{n-k+1} = c\Delta t^2 q\frac{1-q^n}{1-q} \\ &\le c\Delta t^2\frac{q}{1-q} = \frac{c\Delta t\varepsilon}{\bar{\rho}}. \qquad (8.46)\end{aligned}$$

Finally, consider the third term D_3. If $\Delta t\bar{\rho}/\varepsilon > 2$, then

$$\begin{aligned}D_3 &\le c\Delta t\sum_{k=1}^{n}q^{n-k+1} \\ &= c\Delta tq\frac{1-q^n}{1-q}\le c\Delta t\frac{q}{1-q} \\ &= c\Delta t\frac{\varepsilon}{\bar{\rho}\Delta t}\le\frac{c\Delta t}{2}.\end{aligned}$$

If $\Delta t\bar{\rho}/\varepsilon < 2$, then

$$\begin{aligned}D_3 &\le c\Delta t\left(\frac{\Delta t\bar{\rho}}{\varepsilon}\right)^2\sum_{0}^{\infty}mq^m \\ &= c\Delta t\left(\frac{\Delta t\bar{\rho}}{\varepsilon}\right)^2\frac{q}{(1-q)^2} \\ &= c\Delta t(1+\Delta t\bar{\rho}/\varepsilon)\le 3c\Delta t \qquad (8.47)\end{aligned}$$

since

$$\sum_0^\infty m x^{m-1} = (1-x)^{-2}.$$

This shows that

$$|E_3 + \bar{w}(t) - w^n| \le c\Delta t. \tag{8.48}$$

Combine this with (8.41) to obtain

$$|E_3| \le c\Delta t. \tag{8.49}$$

Together with (8.28), this implies that

$$|E_1| + |E_2| + |E_3| \le c\Delta t. \tag{8.50}$$

This concludes the formal demonstration that consistency error in the discrete method is first-order accurate, uniformly in ε.

9. Numerical results. In this section we present some numerical examples in order to compare some existing schemes with the new splitting scheme NSPIF developed in this article. Rewrite the Broadwell equations (2.4)–(2.6) as

$$\partial_t U + A\partial_x U = Q(U). \tag{9.1}$$

Schemes that will be compared are as follows:

1. NSP (introduced in section 4).
2. NSPIF (introduced in section 6).
3. SP1 (introduced in section 8).
4. SCN, which is free streaming (8.2) for the convection $\partial_t U + A\partial_x U = 0$ followed by a half step Crank–Nicolson method for the collision $\partial_t U = Q(U)$ and then followed by another step of free streaming (8.2) on $\partial_t U + A\partial_x U = 0$. Strang's splitting [27] is used here.
5. SCNvL, which is the same as SCN except that the two free streaming steps on the convection $\partial_t U + A\partial_x U = 0$ must be replaced by van Leer's MUSCL spatially and the second-order TVD Runge–Kutta method [26] temporally. Strang's splitting [27] is used here.
6. SCNvLIF, which is the same as SCNvL except that in the first time step we use the initial layer fix introduced in section 6.
7. SP1vL, which is the same as SP1 except that the free streaming step (8.2) on $\partial_t U + A\partial_x U = 0$ is replaced by MUSCL spatially and the second-order TVD Runge–Kutta method temporally.

Note that the free streaming (8.2) is the characteristic method for the convection and corresponds to the upwind scheme, with the CFL number equaling one. Thus we will always use CFL= 1 for SP1 and SCN. For other methods we will use CFL=0.5, unless otherwise stated.

First we solve the Broadwell equation with the following initial data:

$$\rho = 2, \quad m = 1, \quad z = 1, \qquad \text{for} \quad x < 0.2, \tag{9.2}$$

$$\rho = 1, \quad m = 0.13962, \quad z = 1, \qquad \text{for} \quad x > 0.2. \tag{9.3}$$

We integrate over domain $[-1, 1]$ with reflecting boundary conditions and take $\Delta x = 0.01$ and $\Delta t = O(\Delta x)$. We test six different schemes. The exact solution is obtained using fine grids with $\Delta x = 0.0005$.

First we take $\varepsilon = 1$. This is in the rarefied regime. The numerical solutions of ρ, m, and z are depicted with the "exact" solution in Figure 9.1. In this regime, SP1 yields the best resolution, SCN is very diffusive, SCNvL is slightly better than SCN, and SP1vL, NSP, and NSPIF give comparable results which are more diffusive than SP1 but better than SCN and SCNvL. While SP1vL, NSP, and NSPIF show standard second-order behavior, in the rarefied regime it seems that the free stream is a nice thing to use, for SP1 gives the best result and SCN is better then SCNvL.

Next we take $\varepsilon = 10^{-8}$ and compare the behavior of NSP and NSPIF and show how the initial layer fix works. This is the regime where the mean free path is very small and the limiting Euler equation has a shock wave moving right with a speed $s = 0.86038$ determined by the Rankine–Hugoniot jump condition. Note that the initial datum for z is not a local Maxwellian, which yields an initial layer. The results are displayed in Figure 9.2. The NSP and NSPIF are comparable with respect to the shock; however, without the initial layer fix, the NSP creates a hump near the initial discontinuity $x = 0$, solely caused by the kinetic effect. In the next run we see that for a larger ϵ this hump exists in the exact solution. This shows that NSP is dragged by the kinetic behavior even when ϵ is much smaller and the solution is in the fluid regime.

We then choose $\varepsilon = 0.02$. We still use $\Delta x = 0.02$. This is in the intermediate regime where $\varepsilon, \Delta x$, and Δt are of the same order. The results are depicted in Figure 9.3. The small hump located at $x = 0$ is part of the exact solution. It is due to the fact that the initial condition represents an exact traveling shock for the relaxed system, i.e., for the system with $\varepsilon = 0$, while in this case it is $\varepsilon = 0.02$. It seems that SCNvL, NSP, and NSPIF give comparable and the best results, while SP1vL is slightly more diffusive and SP1 and SCN are much more smeared out and exhibited typical first-order behavior.

We then choose other initial data:

$$\rho = 1, \quad m = 0, \quad z = 1 \qquad \text{for} \quad x < 0.5, \tag{9.4}$$

$$\rho = 0.2, \quad m = 0, \quad z = 1 \qquad \text{for} \quad x > 0.5. \tag{9.5}$$

We integrate over domain $[0, 1]$ with reflecting boundary conditions. We take $\Delta x = 0.01$ and $\varepsilon = 10^{-8}$ so the solution is close to that of the model Euler equation. By solving the model Euler equation one obtains a left-moving rarefaction wave and a right-moving shock wave. The initial data are not in the local Maxwellian. The numerical solutions are plotted in Figure 9.4. In this fluid limit we observe that NSPIF gives the best resolution for both the shock and the rarefaction waves, and NSP is comparable although it smears a tiny bit more than NSPIF. All other schemes smear a lot more. Both SCN and SCNvL give z far from the local Maxwellian, which is not surprising since the Crank–Nicolson does not project to the local Maxwellian if the initial data are not local Maxwellian. This can be fixed rather easily if these schemes are combined with the initial layer fix introduced in section 6. Such a phenomenon also disappears if the initial data are in the local Maxwellian. Very interesting is the result of SCNvL, which gives completely spurious wave structures.

To examine the spurious waves produced by the SCNvL for this problem we depict the results of SCNvL and SCNvLIF for the same problem in Figure 9.5 with CFL = 0.5 and CFL = 0.005, respectively. When CFL = 0.5 both SCNvL and SCNvLIF give spurious but different waves! By taking Δt 100 times smaller such spurious waves disappear and the schemes give qualitatively correct waves, although the waves are smeared more due to the smallness of Δt. This example shows that large temporal spacing in SCNvL may produce numerical results which look completely reasonable

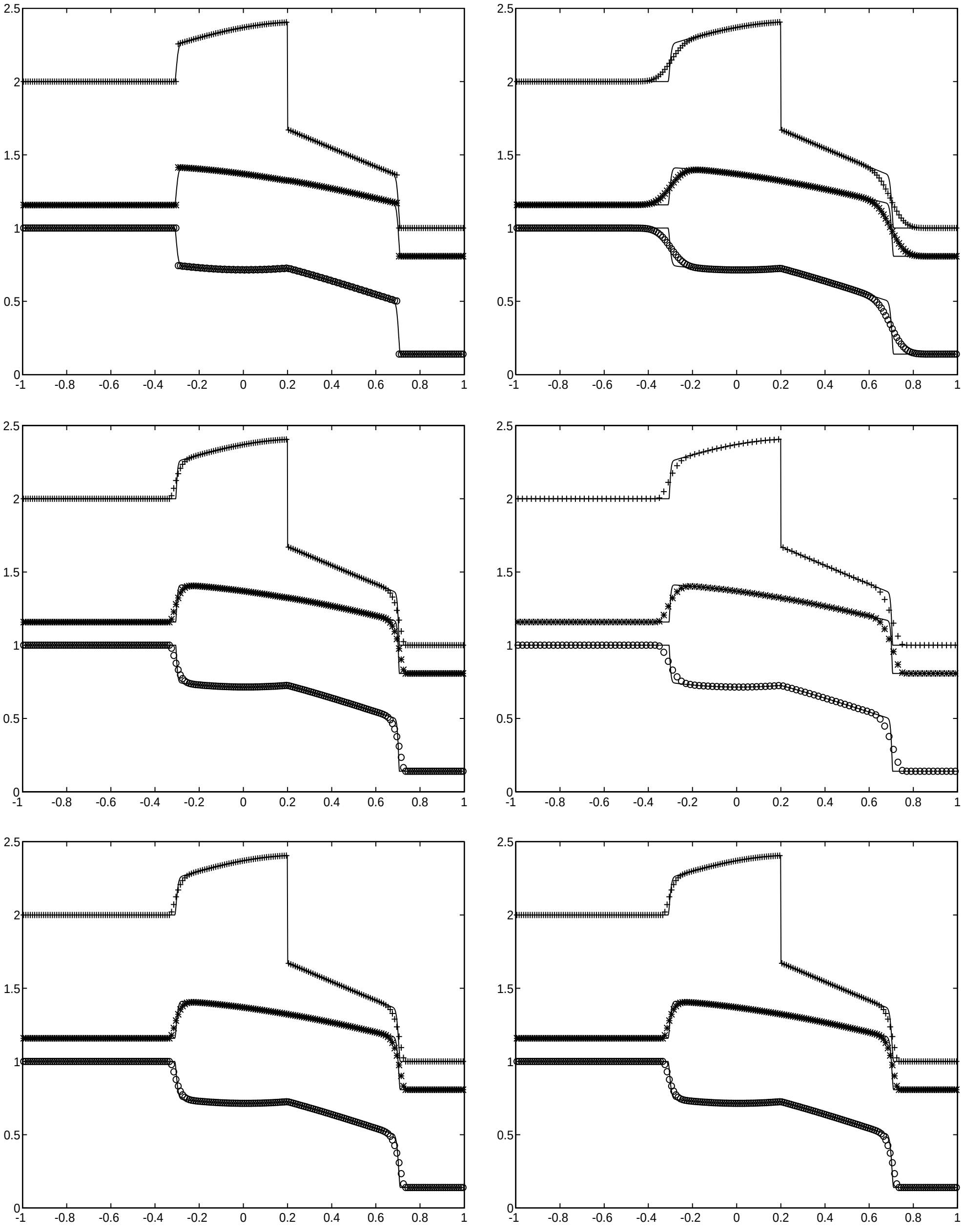

FIG. 9.1. *The numerical solutions of ρ ("+"), m ("o"), and z ("*") at $t = 0.5$ in $x \in (-1, 1)$ for initial data* (9.2) *and* (9.3) *by (from left to right, then top to bottom) SP*1*, SCN, SP*1*vL, SCNvL, NSP, and NSPIF.* $\varepsilon = 1$, $\Delta x = 0.01$. *CFL=* 1 *for SP*1 *and SCN, CFL=* 0.5 *for others.*

but are totally unphysical! The reason is simple: the SCNvL does not have the correct fluid limit in the coarse regime.

In summary, NSPIF seems to produce satisfactory results and to exhibit a typical second-order behavior in all these regimes. Other methods may give unsatisfactory (even wrong) results in one or another regime.

Next we perform the numerical convergence study. We consider an initial value problem with periodic boundary conditions such that the solution is smooth in a time

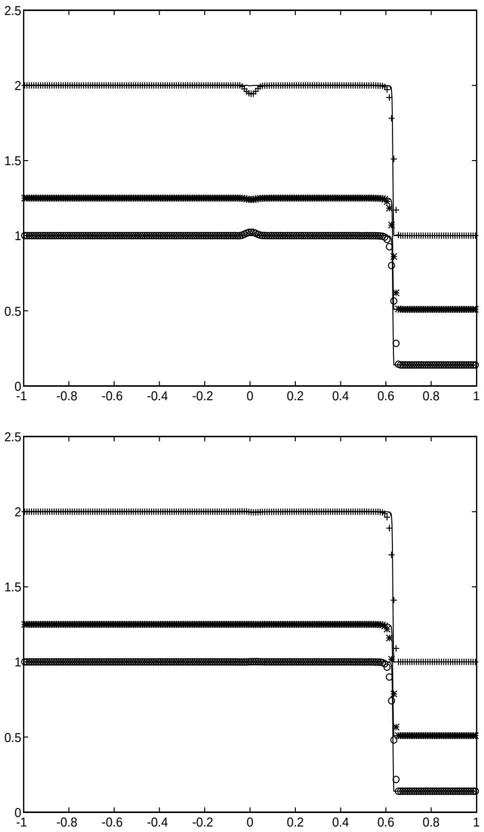

FIG. 9.2. *The numerical solutions of* ρ *(“+”),* m *(“o”), and* z *(“*”) at* $t = 0.5$ *in* $x \in [-1, 1]$ *for initial data* (9.2) *and* (9.3) *by NSP and NSPIF.* $\varepsilon = 10^{-8}$*,* $\Delta x = 0.01$*, and CFL* $= 0.5$*.*

interval $[0, T]$ for any value of the parameter ε. We compute the error at time T by differencing, i.e., by comparing the result obtained with a given grid $(\Delta x, \Delta t)$ with the one obtained with the grid $(\Delta x/2, \Delta t/2)$.

In this section we have proved that the consistency error for SP1 is uniformly first order. By truncation analysis we know that the scheme NSP and NSPIF are second order in both space and time, if $\Delta t \ll \varepsilon$, and they are second order also in the fluid regime (for smooth solutions).

The goal of the test is to perform a numerical study of the convergence rate for a wide range of ε and to check whether the convergence is uniform in ε also in the intermediate regime. The test problem is given by equations (2.1) with periodic boundary conditions: $s(x+L, t) = s(x, t)$ with $s = f, g, h$. The initial data are given by

$$\rho(x, 0) = 1 + a_\rho \sin\frac{2\pi x}{L}, \quad u(x, 0) = \frac{1}{2} + a_u \sin\frac{2\pi x}{L},$$
$$m(x, 0) = \rho(x, 0)u(x, 0), \quad z(x, 0) = z_E(\rho(x, 0)u(x, 0))\theta_M,$$

where θ_M is a real parameter. If $\theta_M = 1$ then the initial condition is a local Maxwellian; otherwise it is not. If $\theta_M \neq 1$, $\varepsilon \ll 1$, there is an initial layer. The system is integrated for $t \in [0, T]$. The values of the parameters used in the computations are

$$L = 20,\ T = 30,\ a_\rho = 0.3,\ a_u = 0.1.$$

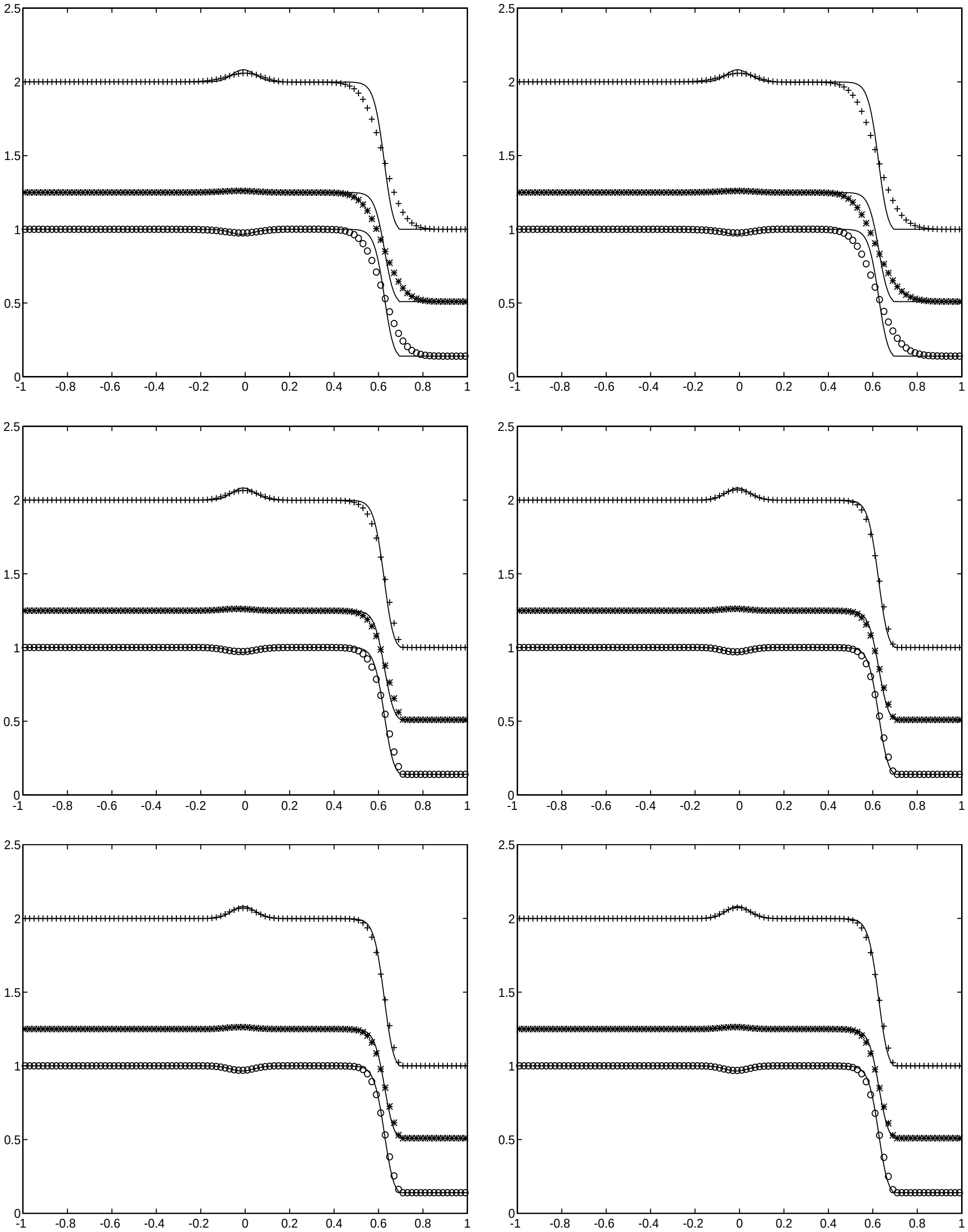

FIG. 9.3. *The numerical solutions of ρ ("+"), m ("o"), and z ("*") at $t = 0.5$ in $x \in [-1, 1]$ for initial data* (9.2) *and* (9.3) *by (from left to right, then top to bottom) SP*1*, SCN, SP*1*vL, SCNvL, NSP, and NSPIF.* $\varepsilon = 0.02$*,* $\Delta x = 0.02$*. CFL* $= 1$ *for SP*1 *and SCN, CFL* $= 0.5$ *for others.*

The values of Δx used in the computations are

$$\Delta x = 0.4, 0.2, 0.1, 0.05, 0.025$$

for the first-order scheme SP1 and

$$\Delta x = 1, 0.5, 0.25, 0.125, 0.0625$$

for the second-order schemes. The time step is chosen in such a way that the CFL condition is satisfied: $\Delta t = \Delta x$ for scheme SP1 and $\Delta t = \Delta x/2$ for the second-order

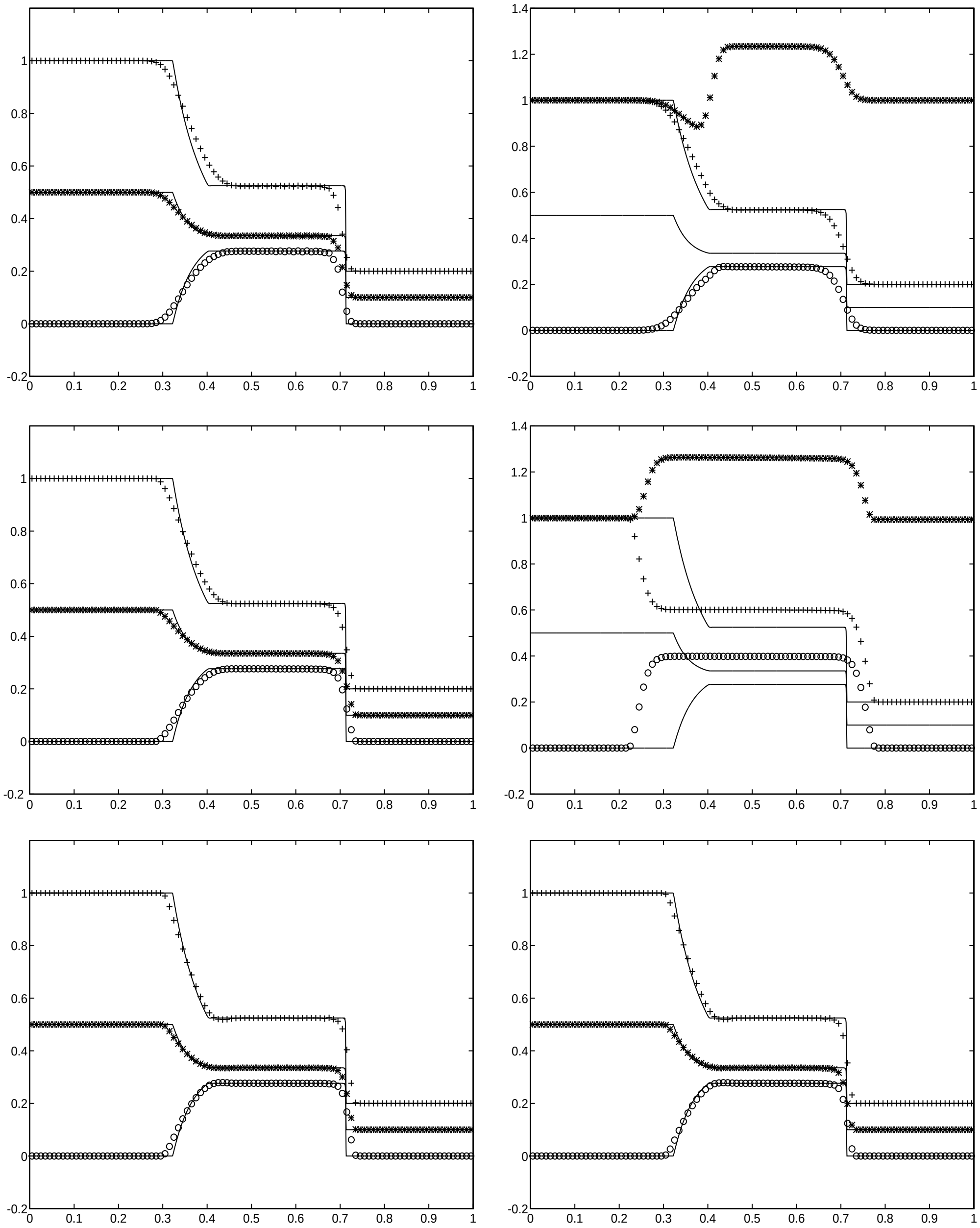

Fig. 9.4. *The numerical solutions of ρ ("+"), m ("o"), and z ("*") at $t = 0.25$ in $x \in [0, 1]$ for initial data* (9.4) *and* (9.5) *by (from left to right, then top to bottom) SP*1*, SCN, SP*1*vL, SCNvL, NSP, NSPIF, and SCNvLIF.* $\varepsilon = 10^{-8}$, $\Delta x = 0.01$. *CFL* = 1 *for SP*1 *and SCN, CFL* = 0.5 *for others.*

schemes. The convergence rate is computed from the error according to the formula

$$\text{convergence rate}_i = \frac{\log(\text{error}_i/\text{error}_{i+1})}{\log(\Delta x_i/\Delta x_{i+1})},$$

where error_i is obtained by comparing the solution obtained using Δx_i with that obtained using Δx_{i+1}. The errors and convergence rate are computed and plotted as function of ε. For each value of ε, five runs have been done for five dif-

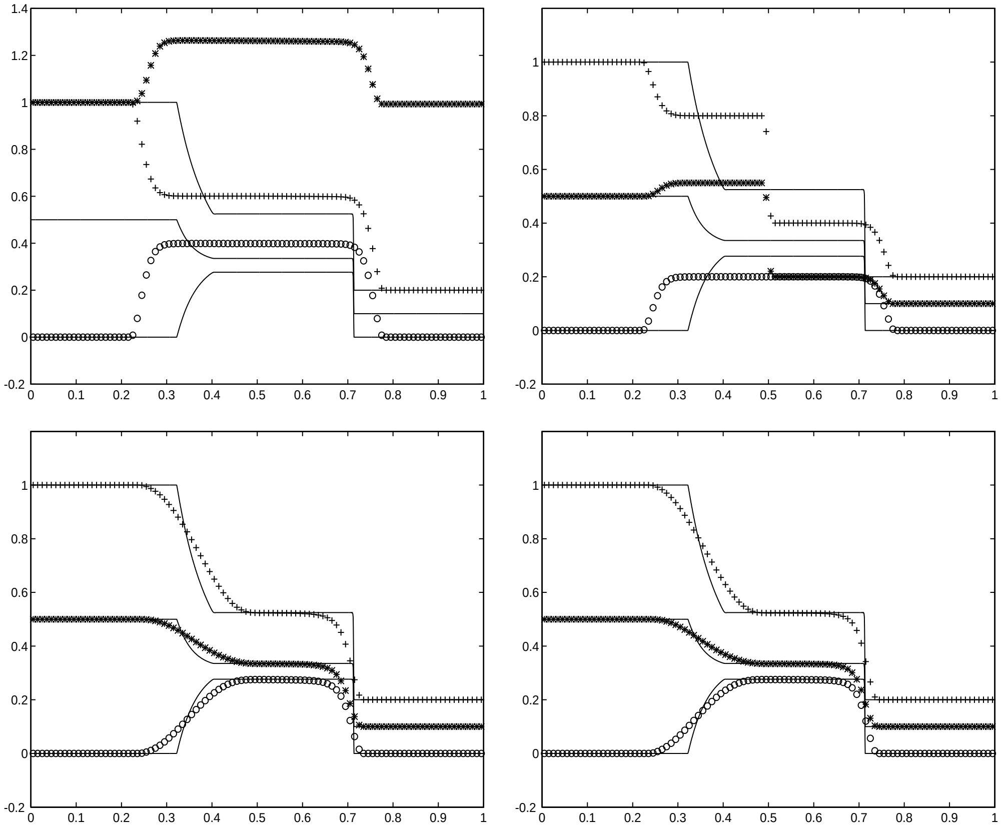

FIG. 9.5. *The numerical solutions of ρ ("+"), m ("o"), and z ("*") at $t = 0.25$ in $x \in [0, 1]$ for initial data* (9.4) *and* (9.5) *by (from left to right, then top to bottom) SCNvL and SCNvLIF for CFL* $= 0.5$ *and SCNvL and SCNvLIF for CFL* $= 0.005$. $\varepsilon = 10^{-8}$, $\Delta x = 0.01$.

ferent values of Δx, resulting in four error curves and three curves of convergence rate.

Several measures of the error have been used, namely, L_1, L_2, and L_∞ relative norm of the error. The different norms give essentially the same results; therefore we shall show only the L_1 norm.

First we consider the simple splitting scheme SP1. In Figure 9.6 we show the relative discrete norm of the error in ρ and in m as a function of ε (left column) and the corresponding convergence rate (right column).

The initial state is a local Maxwellian in the first two cases and it is not in the last. The convergence rate increases when the mesh becomes finer and seems to confirm that the scheme is first order, uniformly in ε, for a fine enough mesh.

Next we consider second-order schemes. In Figure 9.7a–b the result of scheme NSP is shown. The initial condition is a local Maxwellian.

As it is evident from the figures, the scheme is second-order accurate for small and large values of ε, and there is a slight deterioration of the accuracy in the intermediate regime.

Figure 9.7c shows the effect of the initial layer if scheme NSP is used without using Richardson extrapolation for the first step. The accuracy of the scheme of course degrades due to the initial layer. In Figure 9.7d the convergence rate is shown.

The problem of the initial layer can be overcome by using Richardson extrapolation for the first step (scheme NSPIF, Figure 9.7e–f). A similar result is obtained by

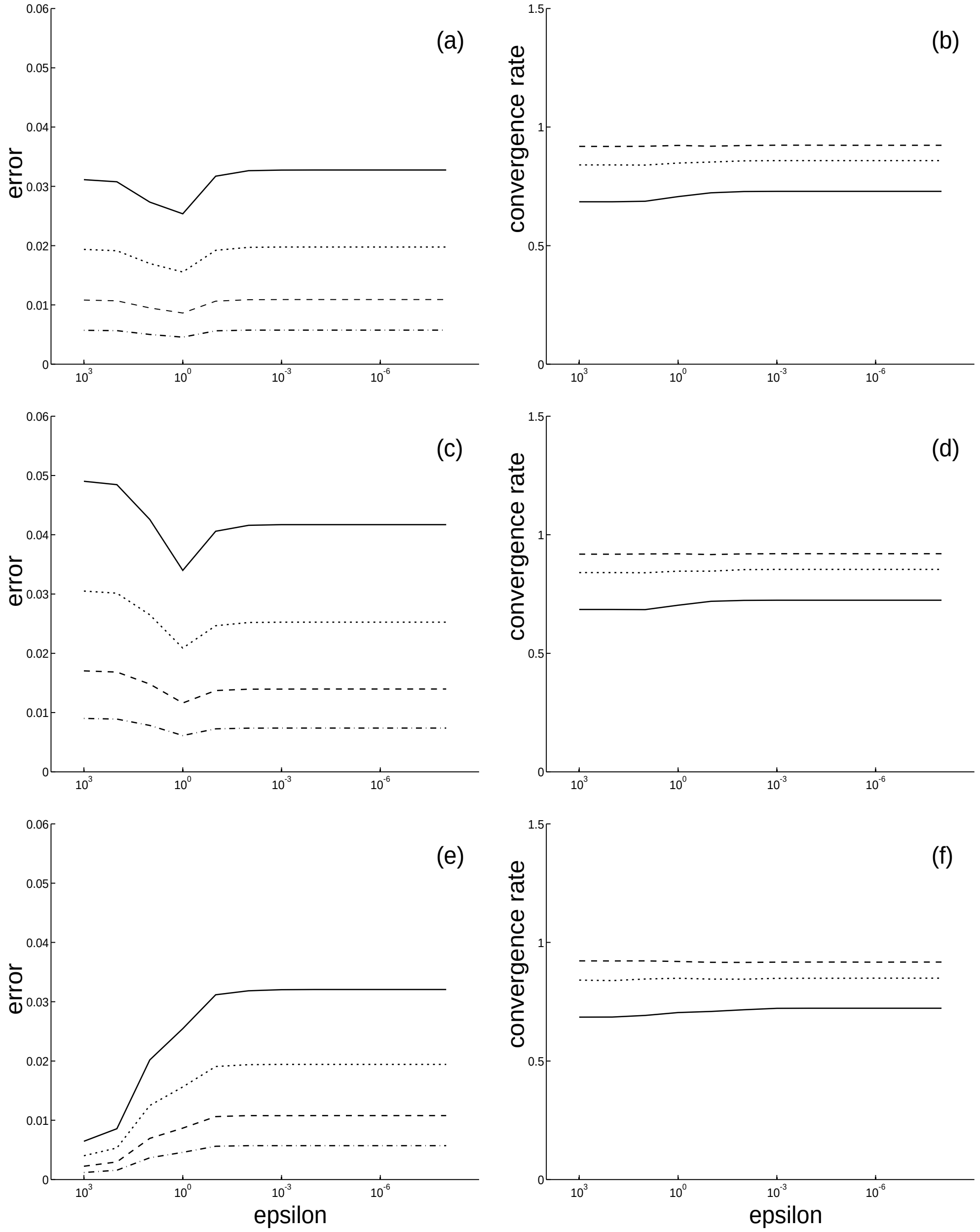

FIG. 9.6. *Uniform convergence of the simple splitting scheme* $SP1$. *Relative error (left) and convergence rate (right) vs* ε *for various values of the grid step size* Δx. *Continuous line: coarsest mesh; dashed line: finest mesh.* (a–b)*:* L_1 *error in* ρ. *Initial condition: local Maxwellian.* (c–d)*:* L_1 *error in* m. *Initial condition: local Maxwellian.* (e–f)*:* L_1 *error in* ρ. *Initial condition: not local Maxwellian.*

using a scheme based entirely on Richardson extrapolation, in which every step is of the form (6.5).

Acknowledgments. S. Jin and G. Russo would like to thank the Mathematics Department at UCLA for its hospitality during their visits there. S. Jin is grateful to Prof. George Papanicolaou for his constant encouragement and support for this

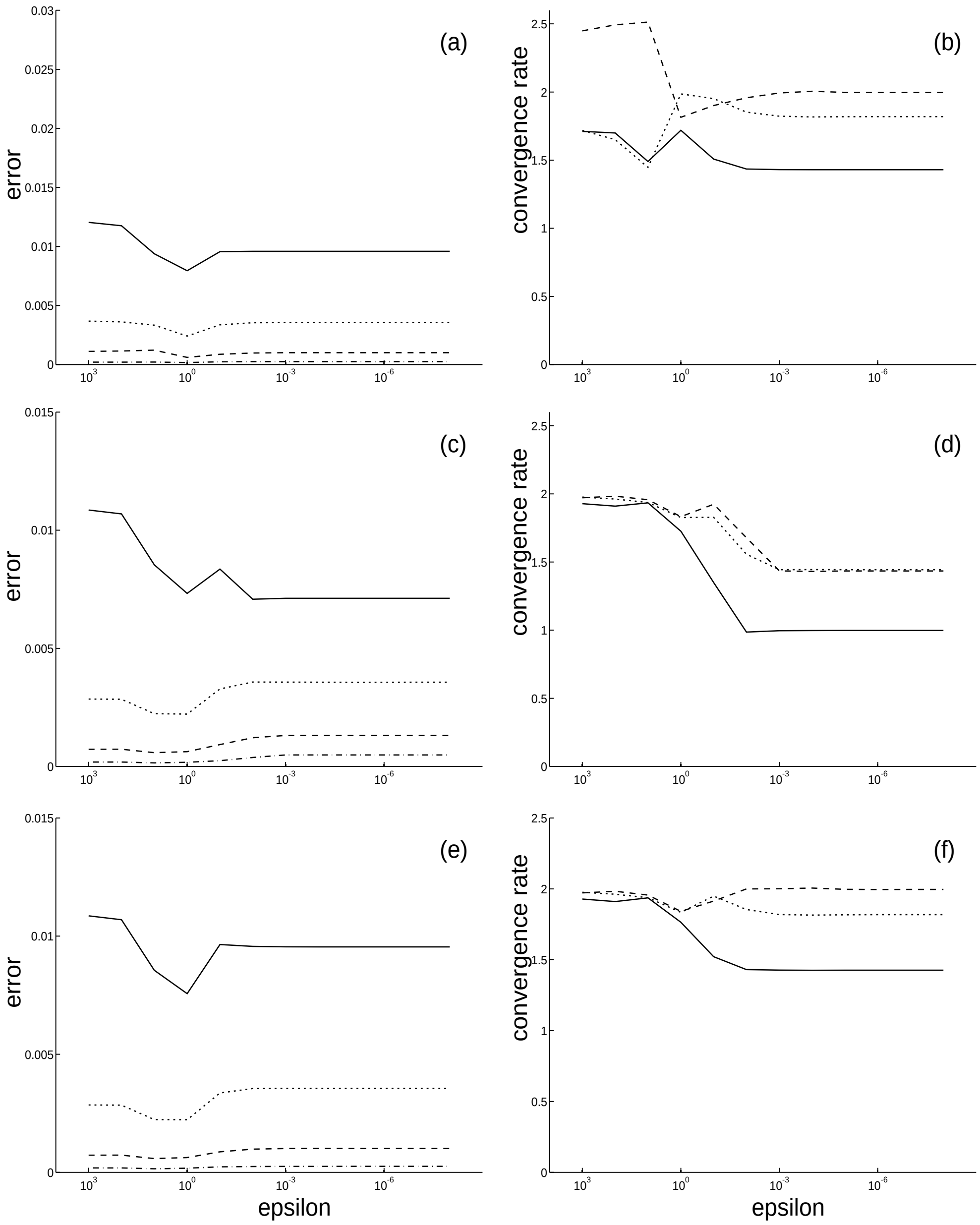

FIG. 9.7. *Relative L^1 norm of the error in density and convergence rate vs ε for various values of the grid step size Δx.* (a) *and* (b)*: Scheme NSP. Initial state: local Maxwellian.* (c) *and* (d)*: Scheme NSP. The initial state is not a local Maxwellian.* (e) *and* (f)*: Scheme NSPIF. Same initial state of* (c–d).

work. We also thank the two unknown referees for their critical remarks on the first draft of the paper.

REFERENCES

[1] A. V. BOBYLEV, E. GABETTA, AND L. PARESCHI, *On a boundary value problem for the plane Broadwell model. Exact solutions and numerical simulation*, Math. Models and Methods in Appl. Sci., 5 (1995), pp. 253–266.

[2] J. E. Broadwell, *Shock structure in a simple discrete velocity gas*, Phys. Fluids, 7 (1964), pp. 1013–1037.

[3] R. E. Caflisch and G. Papanicolau, *The fluid dynamic limit of a nonlinear model Boltzmann equation*, Comm. Pure and Appl. Math., 22 (1979), pp. 586–616.

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

[5] G. Q. Chen, C. D. Levermore, and T. P. Liu, *Hyperbolic conservation laws with stiff relaxation terms and entropy*, Comm. Pure and Appl. Math., 47 (1994), pp. 787–830.

[6] P. Colella, A. Majda, and V. Roytburd, *Theoretical and numerical structure for reacting shock waves*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 1059–1080.

[7] F. Coron and B. Perthame, *Numerical passage from a kinetic to a fluid equation*, SIAM J. Numer. Anal., 28 (1991), pp. 26–42.

[8] S. Deshpande, *A Second Order Accurate, Kinetic-Theory Based Method for Inviscid Compressible Flow*, Tech. paper 2613, NASA, Langley, VA, 1986.

[9] B. Engquist and B. Sjogreen, *Robust Difference Approximations of Stiff Inviscid Detonation Waves*, CAM report 91-03, UCLA, Los Angeles, CA, 1991.

[10] H. Emamirad, *Méthode des pas fractionnaires pour la modèle de Broadwell*, C. R. Acad. Sci. Paris, t.304, Séries II (1987), pp. 487–490.

[11] K. O. Friedrichs and P. D. Lax, *System of conservation laws with a convex extension*, Proc. Nat. Acad. Sci. USA, 68 (1971), pp. 1686–1688.

[12] E. Gabetta and L. Pareschi, *Approximating the Broadwell model in a strip*, Math. Models and Methods in Appl. Sci., 2 (1992), pp. 1–19.

[13] R. Gatignol, *Théorie cinétique des gas à répartition discrète de vitesses*, Lecture Notes in Physics 36, Springer-Verlag, Heidelberg, 1975.

[14] F. Golse, S. Jin, and C. D. Levermore, *The convergence of numerical transfer schemes in diffusive regimes* I: *The discrete-ordinate method*, SIAM. J. Numer. Anal., submitted.

[15] A. Harten, P. D. Lax, and B. van Leer, *On upstream difference and Godunov-type schemes for hyperbolic conservation laws*, SIAM Rev., 25 (1983), pp. 35–61.

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

[17] S. Jin and C. D. Levermore, *Numerical schemes for hyperbolic systems with stiff relaxation terms*, J. Comput. Phys., 126 (1996), pp. 449–467.

[18] S. Jin and Z. P. Xin, *The relaxation schemes for systems of conservation laws in arbitrary space dimensions*, Comm. Pure and Appl. Math., 48 (1995), pp. 235-277.

[19] E. W. Larsen, J. E. Morel, and W. F. Miller, Jr., *Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes*, J. Comput. Phys., 69 (1987), pp. 283–324.

[20] R. J. LeVeque, *Numerical Methods for Conservation Laws*, Birkhauser-Verlag, Basel, 1992.

[21] R. J. LeVeque and H. C. Yee, *A study of numerical methods for hyperbolic conservation laws with stiff source terms*, J. Comput. Phys., 86 (1990), pp. 187–210.

[22] T. P. Liu, *Hyperbolic conservation laws with relaxation*, Comm. Math. Phys., 108 (1987), pp. 153–175.

[23] R. B. Pember, *Numerical methods for hyperbolic conservation laws with stiff relaxation* I. *Spurious solutions*, SIAM J. Appl. Math., 53 (1993), pp. 1293–1330.

[24] R. B. Pember, *Numerical methods for hyperbolic conservation laws with stiff relaxation* II. *Higher order methods*, SIAM J. Sci. Statist. Comp., 14 (1993), pp. 824–859.

[25] B. Perthame, *Second-order Boltzmann schemes for compressible Euler equations in one and two space dimensions*, SIAM J. Numer. Anal., 29 (1992), pp. 1–19.

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

[27] G. Strang, *On the construction and comparison of difference schemes*, SIAM J. Numer. Anal., 5 (1968), pp. 506–517.

[28] B. van Leer, *Toward the ultimate conservative difference scheme* V. *A second-order sequel to Godunov's method*, J. Comput. Phys., 32 (1979), pp. 101–136.

[29] G. B. Whitham, *Linear and Nonlinear Waves*, Wiley, New York, 1974.

[30] Z. P. Xin, *The fluid dynamic limit of the Broadwell model of the nonlinear Boltzmann equation in the presence of shocks*, Comm. Pure and Appl. Math., 44 (1991), pp. 679–713.