Z. angew. Math. Phys. 50 (1999) 258–281
0044-2275/99/020258-24 $ 1.50+0.20/0

Zeitschrift für angewandte Mathematik und Physik ZAMP

Gas-kinetic schemes for the compressible Euler equations: Positivity-preserving analysis

Tao Tang and Kun Xu

Abstract. Numerical schemes based on the collisional BGK model have been developed in recent years. In this paper, we investigate the first-order BGK schemes for the Euler equations. Particular attention is given to finding CFL-like conditions under which the schemes are positivity-preserving (i.e. density and internal energy remain nonnegative). The first-order BGK schemes are linear combinations of collisionless (i.e. kinetic flux-splitting scheme) and collisional approach. We show that the collisionless approach preserves the positivity of density and internal energy under the standard CFL condition. Although the collisionless approach has the positivity-preserving property, it introduces large intrinsic dissipation and heat conductions since the corresponding scheme is based on two half Maxwellians. In order to reduce the viscous error, one obvious method is to use an exact Maxwellian, which leads to the collisional Boltzmann scheme. An CFL-like condition is also found for the collisional approach, which works well for the test problems available in literature. However, by considering a counterexample we find that the collisional approach is not always positivity-preserving. The BGK type schemes are formed by taking the advantages of both approaches, i.e. the less dissipative scheme (collisional) and the more dissipative but positivity-preserving scheme (collisionless).

Mathematics Subject Classification (1991). 65M93, 35L64, 76N10.

Keywords. Euler equations, gas-kinetic schemes, positivity-preserving, Maxwellian distribution, BGK.

1. Introduction

In an important and influential paper, Harten, Lax and van Leer [6] drew a distinction between two numerical approaches to the solution of the Euler equations, namely, Godunov and Boltzmann schemes. Broadly speaking, Godunov scheme is based on the Riemann solutions, and the Boltzmann scheme uses the microscopic particle motion as the basis to construct the scheme, where the macroscopic flow behavior is an average collection of interactions in the microscopic world. Godunov and Boltzmann schemes are based on two different physical considerations. The development of numerical schemes based on a Riemann solver starts from: *Euler* $\longrightarrow$ *Navier-Stokes* $\longrightarrow$ *high-moments equations.* On the other hand, schemes

based on the Boltzmann equation start from: *Rarefied gas flow* $\longrightarrow$ *Navier-Stokes* $\longrightarrow$ *Euler*. Basically, this is also the direct reason why Boltzmann-type schemes always give the entropy satisfying solutions.

Currently, the governing equations in constructing gas-kinetic schemes for the compressible Euler equations can be distinguished mainly in two groups. One of them is based on the Collisionless Boltzmann equation [3, 5, 10, 13, 15], and the other is based on the collisional BGK model [17, 12, 18, 19]. The collisionless Boltzmann solution and the Euler solution are two extreme limits in flow motion. The former describes the gas flow without any particle collisions and the latter includes infinite number of collisions. The transition from one scheme to other can be finished by including more and more particle collisions, where the Navier-Stokes solutions are between them. The BGK scheme is just a scheme to connect collisionless Boltzmann equation and the inviscid Euler equations, which naturally gives the Navier-Stokes solutions.

In computing numerical solutions of the Euler equations, one important requirement is that density and internal energy should remain positive under a Courant-Friedrichs-Lewy (CFL)-like condition. This property is called positivity-preserving property. It is well known that classical approximate Riemann solvers do not satisfy this property [4, 16]. This is a serious drawback when the solution is near vacuum. On the other hand, schemes based on the Boltzmann equation are found to preserve the positivity of density and internal energy under a CFK-like condition [5, 13, 11, 18]. It is quite well understood that the kinetic flux-splitting methods satisfy the positivity-preserving property. In the last few years, numerical schemes based on the collisional BGK model have been developed [12, 18, 19]. It is observed numerically that the BGK schemes also satisfy the positivity-preserving property.

Recently, there have been also interests in developing the positive schemes, see e.g. [9, 7]. In the sense of Liu and Lax [9], a scheme is positive if it can be written in the form

$$U_J^{n+1} = \sum_K C_K^n U_{J+K}^n,$$

so that the coefficient matrices C_K, which themselves depend on all the U_{J+K}^n, have the following properties: (i) Each C_K is symmetric positive definite, (ii) $\sum_K C_K$ is the identity matrix. The positive schemes involve two positive parameters. It is shown that some positive schemes do suffer from negative internal energy for the low density and internal energy test problem (Example 4, [9]). By changing the two parameters, i.e. by adding more dissipation, the positivity-preserving property is recovered for the test problem.

In this paper, we will analyze the positivity-preserving properties for the BGK schemes, in particular the first order BGK type schemes. Our schemes also involve two positive parameters, which combine the so-called collisionless approach (i.e. flux splitting method) and the collisional approach. In order to have a better understanding of the BGK schemes, we begin with the investigations of the col-

lisionless and collisional approaches. In Section 3 we show that the collisionless approach is always positivity-preserving as long as the standard CFL condition is satisfied. In Section 4, we investigate the collisional approach. A practical CFL-like condition is obtained that works well for the test problems in literature. However, unlike the collisionless approach, a rigorous CFL-like condition cannot be obtained for the collisional approach. In fact, counterexamples showing that the collisional approach is not positivity-preserving are found. To take the advantages of the collisionless and collisional approaches, i.e. the positivity-preserving property of the collisionless approach and the less-dissipation property of the collisional approach, the linear combination of the two approaches is used to form the BGK type schemes. In Section 5, it is shown that if the collisionless and collisional schemes are positivity-preserving for a given problem, then the corresponding BGK schemes are also positivity-preserving for the same problem. Furthermore, for problems violating the positivity-preserving property with the collisional approach, the BGK schemes work well if some collisionless contributions are added.

2. Preliminaries

We consider the one dimensional Euler equations of gas dynamics:

$$\begin{cases} \rho_t + m_x = 0, \\ m_t + (mU + p)_x = 0, \\ E_t + (EU + pU)_x = 0, \end{cases} \tag{2.1}$$

where ρ is the density, U the velocity, $m = \rho U$ the momentum, $E = \frac{1}{2}\rho U^2 + \rho e$ the energy per unit mass, e the internal energy, p the pressure. We assume that the gas is a γ-law gas, i.e. $p = (\gamma - 1)\rho e$, $1 \le \gamma \le 3$.

The Boltzmann equation in the 1-D case can be written as [8]

$$f_t + u f_x = Q(f, f), \tag{2.2}$$

where f is the gas-distribution function, u the particle velocity, and $Q(f, f)$ the collision term. The collision term is an integral function which accounts for the binary collisions. In most cases, the collision term can be simplied and the BGK model is the most successful one [1],

$$Q(f, f) = (g - f)/\tau, \tag{2.3}$$

where g is the equilibrium state and τ the collision time. For the Euler equations, the equilibrium state g is a Maxwellian,

$$g = \rho \left(\frac{\lambda}{\pi}\right)^{\frac{K+1}{2}} e^{-\lambda((u-U)^2+\xi^2)}, \tag{2.4}$$

where K is the degree of the internal variable given by

$$K = (3 - \gamma)/(\gamma - 1); \tag{2.5}$$

λ is connected to the gas temperature T.

The connection between the distribution function f and macroscopic flow variables is

$$(\rho, m, E)^T = \int \psi_\alpha f du d\xi, \tag{2.6}$$

where ξ is the internal degree of freedom, such as molecular rotation and vibrations, and

$$\psi_\alpha = (1, u, \frac{1}{2}(u^2 + \xi^2))^T$$

are the moments for density ρ, momentum m and total energy E. The fluxes for the corresponding macroscopic variables are

$$(F_\rho, F_m, F_E)^T = \int u\psi_\alpha f du d\xi. \tag{2.7}$$

The conservation principle for mass, momentum and energy during the course of particle collisions requires $Q(f, f)$ to satisfy the compatibility condition

$$\int Q(f, f)\psi_\alpha du d\xi = 0, \qquad \alpha = 1, 2, 3. \tag{2.8}$$

In the 1st-order BGK scheme [17], the gas distribution function at cell interface is

$$\begin{aligned} f &= (1 - e^{-t/\tau})f_1 + e^{-t/\tau} f_0 \\ &= \beta(t) f_1 + \alpha(t) f_0, \end{aligned} \tag{2.9}$$

where f_1 is the exact Maxwellian located at the cell interface, and f_0 is the initial non-equilibrium two half Maxwellians from the left and right hand side of the numerical boundary. The above solution will go back to collisionless solution in the limit of $\tau \to \infty$, the so-called Kinetic Flux Vector Splitting scheme (KFVS). On the other hand, as τ goes to zero, f will be equal to a Maxwellian $f = f_1$ which is the exact requirement from the Euler equations.

Based on (2.9), we will in this paper consider the positivity-preserving properties for the following three cases:

- 1. Collisionless approach: $\alpha(t) \equiv 1, \beta(t) \equiv 0$;
- 2. Collisional approach: $\alpha(t) \equiv 0, \beta(t) \equiv 1$;
- 3. BGK: $0 \le \alpha(t), \beta(t) \le 1$ satisfying $\alpha(t) + \beta(t) \equiv 1$.

3. Collisionless approach

In this section we consider the kinetic flux-splitting scheme (i.e. collisionless scheme) proposed by Pullin [13] and Deshpande [2, 3]. The scheme uses the fact that the Euler equations (2.1) are the first moments of the Boltzmann equation when the velocity repartition function is Maxwellian. In Section 3.1 we briefly recall the collisionless scheme. In Section 3.2 we prove that the scheme is positivity preserving under the standard CFL condition. A similar result, which is about half of the standard CFL condition, has been given in a recent paper of Estivalezes and Villedieu [5].

3.1. Numerical scheme

In order to derive the collisionless Boltzmann scheme, we first construct the numerical flux which is to use (2.7). We suppose that the initial data $(\rho_0(x), m_0(x), E_0(x))$ are piecewise constant over the cells $C_j = [x_{j-1/2}, x_{j+1/2})$. At each time level, once ρ_j, m_j and E_j are given, the corresponding U_j and λ_j can be obtained by the following formulas:

$$m = \rho U, \qquad E = \frac{1}{2}\rho U^2 + \frac{K+1}{4\lambda}\rho. \tag{3.1}$$

Let

$$g_j(x,t,u,\xi) = \rho_j \left(\frac{\lambda_j}{\pi}\right)^{\frac{K+1}{2}} e^{-\lambda_j((u-U_j)^2+\xi^2)} \tag{3.2}$$

be a Maxwellian distribution in the cell C_j. The corresponding distribution function at cell interface is defined by

$$f_0(x_{j+1/2},t,u,\xi) = \begin{cases} g_j(x,t,u,\xi), & \text{if } u > 0 \\ g_{j+1}(x,t,u,\xi), & \text{if } u < 0. \end{cases} \tag{3.3}$$

Using the formulas (2.7), we obtain the numerical fluxes

$$\begin{pmatrix} F^0_{\rho,j+1/2} \\ F^0_{m,j+1/2} \\ F^0_{E,j+1/2} \end{pmatrix} = \rho_j \begin{pmatrix} \frac{U_j}{2}\mathrm{erfc}(-\sqrt{\lambda_j}U_j) + \frac{1}{2}\frac{e^{-\lambda_j U_j^2}}{\sqrt{\pi\lambda_j}} \\ \left(\frac{U_j^2}{2} + \frac{1}{4\lambda_j}\right)\mathrm{erfc}(-\sqrt{\lambda_j}U_j) + \frac{U_j}{2}\frac{e^{-\lambda_j U_j^2}}{\sqrt{\pi\lambda_j}} \\ \left(\frac{U_j^3}{4} + \frac{K+3}{8\lambda_j}U_j\right)\mathrm{erfc}(-\sqrt{\lambda_j}U_j) + \left(\frac{U_j^2}{4} + \frac{K+2}{8\lambda_j}\right)\frac{e^{-\lambda_j U_j^2}}{\sqrt{\pi\lambda_j}} \end{pmatrix} \tag{3.4}$$

$$+\rho_{j+1} \begin{pmatrix} \frac{U_{j+1}}{2}\mathrm{erfc}(\sqrt{\lambda_{j+1}}U_{j+1}) - \frac{1}{2}\frac{e^{-\lambda_{j+1}U_{j+1}^2}}{\sqrt{\pi\lambda_{j+1}}} \\ \left(\frac{U_{j+1}^2}{2} + \frac{1}{4\lambda_{j+1}}\right)\mathrm{erfc}(\sqrt{\lambda_{j+1}}U_{j+1}) - \frac{U_{j+1}}{2}\frac{e^{-\lambda_{j+1}U_{j+1}^2}}{\sqrt{\pi\lambda_{j+1}}} \\ \left(\frac{U_{j+1}^3}{4} + \frac{K+3}{8\lambda_{j+1}}U_{j+1}\right)\mathrm{erfc}(\sqrt{\lambda_{j+1}}U_{j+1}) - \left(\frac{U_{j+1}^2}{4} + \frac{K+2}{8\lambda_{j+1}}\right)\frac{e^{-\lambda_{j+1}U_{j+1}^2}}{\sqrt{\pi\lambda_{j+1}}} \end{pmatrix},$$

where the complementary error function, which is a special case of the incomplete gamma function, is defined by

$$\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt.$$

Like sine and cosin functions, erfc(x), or its double precision derfc(x), is a given function in FORTRAN. Using the above numerical fluxes, we are able to update ρ_j, m_j, E_j with the standard conservative formulations:

$$\begin{pmatrix} \tilde{\rho}_j \\ \tilde{m}_j \\ \tilde{E}_j \end{pmatrix} = \begin{pmatrix} \rho_j \\ m_j \\ E_j \end{pmatrix} + \sigma \begin{pmatrix} F^0_{\rho,j-1/2} - F^0_{\rho,j+1/2} \\ F^0_{m,j-1/2} - F^0_{m,j+1/2} \\ F^0_{E,j-1/2} - F^0_{E,j+1/2} \end{pmatrix}, \tag{3.5}$$

where

$$\sigma = \frac{\Delta t}{\Delta x},$$

with Δt the stepsize in time, and Δx the mesh size in space. The scheme can be viewed as consisting of the following three steps (although it is not typically implemented this way):

ALGORITHM 0 (Collisionless Approach)

1. Given data $\{\rho_j^n, U_j^n, E_j^n\}$, compute $\{\lambda_j^n\}$ using (3.1).
2. Compute the numerical flux $\{F^0_{\rho,j+1/2}, F^0_{m,j+1/2}, F^0_{E,j+1/2}\}$ using (3.4).
3. Update $\{\rho_j^n, m_j^n, E_j^n\}$ using (3.5). This gives $\{\rho_j^{n+1}, m_j^{n+1}, E_j^{n+1}\}$.

3.2. Positivity-preserving analysis

The numerical schemes (3.5) can be split into two steps. In the first step we consider the case when there is only gas flowing out from the cell C_j. This gives that

$$\begin{pmatrix} \rho_j^* \\ m_j^* \\ E_j^* \end{pmatrix} = \begin{pmatrix} \rho_j \\ m_j \\ E_j \end{pmatrix} + \sigma \begin{pmatrix} \int_{u<0} u g_j du d\xi - \int_{u>0} u g_j du d\xi \\ \int_{u<0} u^2 g_j du d\xi - \int_{u>0} u^2 g_j du d\xi \\ \int_{u<0} \frac{u}{2}(u^2+\xi^2) g_j du d\xi - \int_{u>0} \frac{u}{2}(u^2+\xi^2) g_j du d\xi \end{pmatrix}. \tag{3.6}$$

The second step is to add the correction terms:

$$\begin{pmatrix} \tilde{\rho}_j \\ \tilde{m}_j \\ \tilde{E}_j \end{pmatrix} = \begin{pmatrix} \rho_j^* \\ m_j^* \\ E_j^* \end{pmatrix} + \sigma \begin{pmatrix} \int_{u>0} u g_{j-1} du d\xi - \int_{u<0} u g_{j+1} du d\xi \\ \int_{u>0} u^2 g_{j-1} du d\xi - \int_{u<0} u^2 g_{j+1} du d\xi \\ \int_{u>0} \frac{u}{2}(u^2+\xi^2) g_{j-1} du d\xi - \int_{u<0} \frac{u}{2}(u^2+\xi^2) g_{j+1} du d\xi \end{pmatrix}. \tag{3.7}$$

It can be verified that $(\tilde{\rho}_j, \tilde{m}_j, \tilde{E}_j)$ obtained by (3.5) are exactly the same as those obtained by using (3.7).

Lemma 3.1. *Assume that ρ_j^*, m_j^*, E_j^* be computed by (3.6). If $\rho_j \geq 0$ and $\rho_j E_j \geq \frac{1}{2}(m_j)^2$ for all integers j, then*

$$\rho_j^* \geq 0, \qquad \rho_j^* E_j^* \geq \frac{1}{2}\left(m_j^*\right)^2 \tag{3.8}$$

for all j, provided that the following CFL condition is satisfied:

$$\sigma \leq \frac{1}{\max_j \left(|U_j| + c_j\right)}, \tag{3.9}$$

where $c_j = \sqrt{\gamma/2\lambda_j}$ is the local speed of sound.

Proof. It follows from (3.2) and (3.6) that

$$\begin{aligned}
\rho_j^* &= \rho_j - \sigma\rho_j \left\{ \frac{1}{2} U_j \alpha_j + \beta_j \right\}, \\
m_j^* &= m_j - \sigma\rho_j \left\{ \left(\frac{U_j^2}{2} + \frac{1}{4\lambda_j} \right) \alpha_j + u_j \beta_j \right\}, \\
E_j^* &= E_j - \sigma\rho_j \left\{ \left(\frac{U_j^3}{4} + \frac{K+3}{8\lambda_j} U_j \right) \alpha_j + \left(\frac{U_j^2}{2} + \frac{K+2}{4\lambda_j} \right) \beta_j \right\},
\end{aligned}$$

where

$$\alpha_j = \operatorname{erfc}\left(-\sqrt{\lambda_j} U_j\right) - \operatorname{erfc}\left(\sqrt{\lambda_j} U_j\right); \quad \beta_j = \frac{e^{-\lambda_j U_j^2}}{\sqrt{\pi\lambda_j}}. \tag{3.10}$$

For ease of notations, we drop the subscript j in the remaining of the proof. It follows from (3.10) that

$$0 \leq U\alpha \leq 2|U|, \qquad \beta \leq \frac{1}{\sqrt{\pi\lambda}}.$$

If σ satisfies (3.9), then

$$\rho^* \geq \rho\sigma \left\{ \max_j \left(|U_j| + c_j \right) - \left(|U| + \frac{1}{\sqrt{\pi\lambda}} \right) \right\} \geq 0.$$

Furthermore, we observe that

$$\rho^* E^* - \frac{1}{2}\left(m^*\right)^2 = A\sigma^2 - B\sigma + C,$$

where, by direct calculations

$$A = \left(\frac{K+1}{16\lambda}U^2 - \frac{1}{32\lambda^2}\right)\rho^2\alpha^2 + \frac{K+2}{4\lambda}\rho^2\beta^2 + \frac{2K+3}{8\lambda}U\rho^2\alpha\beta;$$
$$B = \frac{K+1}{4\lambda}\rho^2 U\alpha + \frac{2K+3}{4\lambda}\rho^2\beta;$$
$$C = \rho E - \frac{1}{2}m^2 = \frac{K+1}{4\lambda}\rho^2.$$

The last equation indicates that $C \geq 0$. It follows from Jensen's inequality and the integral formulations (3.6) that $A \geq 0, B \geq 0$. Direct calculation also shows that $B^2 - 4AC \geq 0$. These facts imply that there are two positive roots for the quadratic equation $A\sigma^2 - B\sigma + C = 0$. In order that $\rho^* E^* \geq \frac{1}{2}(m^*)^2$, the σ should satisfy that $\sigma \leq \sigma_1$, here σ_1 is the smaller root of the quadratic equation. Direct calculation gives

$$\sigma_1 = \left(\frac{1}{2}U\alpha + \frac{2K+3}{2K+2}\beta + \frac{1}{K+1}\sqrt{\frac{K+1}{8\lambda}\alpha^2 + \frac{1}{4}\beta^2}\right)^{-1}.$$

Now introduce the following function:

$$F(x,K) = |x| + \sqrt{\frac{K+3}{2K+2}} - \frac{1}{2}x\Big(\operatorname{erfc}(-x) - \operatorname{erfc}(x)\Big) - \frac{2K+3}{2K+2}\frac{e^{-x^2}}{\sqrt{\pi}}$$
$$- \frac{1}{K+1}\sqrt{\frac{K+1}{8}\Big(\operatorname{erfc}(-x) - \operatorname{erfc}(x)\Big)^2 + \frac{e^{-2x^2}}{4\pi}}.$$

It can be shown that $F(x,K)$ is always positive for any $x \in \mathbf{R}$ and for any positive K. This can also be seen from Figure 1 where we have plotted $F(x,K)$ for several values of K. Since $\gamma = (K+3)/(K+1)$, $F(x,K) \geq 0$ indicates that

$$\sigma_1 \geq \frac{1}{|U| + \sqrt{\frac{\gamma}{2\lambda}}}.$$

This completes the proof of this lemma. □

Lemma 3.2. *Assume that $\tilde{\rho}_j, \tilde{m}_j, \tilde{E}_j$ be computed by (3.7). If ρ_j^*, m_j^* and E_j^* used in (3.7) satisfy $\rho_j^* \geq 0$ and $\rho_j^* E_j^* \geq \frac{1}{2}(m_j^*)^2$ for all integers j, then for any choice of $\sigma > 0$ the following positivity-preserving properties are held*

$$\tilde{\rho}_j \geq 0, \qquad \tilde{\rho}_j \tilde{E}_j \geq \frac{1}{2}(\tilde{m}_j)^2 \tag{3.11}$$

for all j.

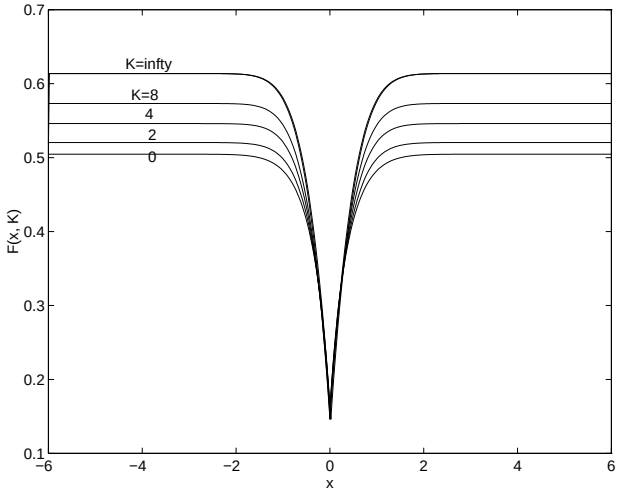

Figure 1.
The function $F(x, K)$, with $K = 0, 2, 4, 8, \infty$.

Proof. It follows from Lemma 3.1 that $\rho_j^* \geq 0$, $\rho_j^* E_j^* \geq \frac{1}{2}\left(m_j^*\right)^2$. It is observed from (3.7) that $\tilde{\rho}_j \geq \rho_j^* \geq 0$. Similar to the proof of Lemma 3.1, we can write $\tilde{\rho}_j \tilde{E}_j - \frac{1}{2}\left(\tilde{m}_j\right)^2$ into the following form:

$$\tilde{\rho}_j \tilde{E}_j - \frac{1}{2}\left(\tilde{m}_j\right)^2 = A\sigma^2 + B\sigma + C,$$

where the coefficients A, B, and C are obtained from (3.7). Using the facts that $\rho_j^* E_j^* \geq \frac{1}{2}\left(m_j^*\right)^2$ and

$$\int_{u>0} \frac{u}{2}(u^2+\xi^2) g_{j-1} du d\xi \geq \int_{u>0} \frac{1}{2} u^3 g_{j-1} du d\xi;$$
$$\int_{u<0} \frac{u}{2}(u^2+\xi^2) g_{j+1} du d\xi \leq \int_{u<0} \frac{1}{2} u^3 g_{j+1} du d\xi,$$

we can show that $A \geq 0, B \geq 0$ and $C \geq 0$. This completes the proof of (3.11). □

Combining Lemmas 3.1 and 3.2, we conclude that the collisionless approach is positivity-preserving as long as the standard CFL condition is satisfied.

Remark 3.1. Lemma 3.2 shows that the CFL condition based on the positivity-preserving analysis for the numerical scheme (3.5) can be determined by analyzing

the simplified scheme (3.6). In other words, the CFL condition is obtained by considering the scheme (3.5) with the following assumption:

$$\rho_{j-1} = 0, \qquad \rho_j > 0, \qquad \rho_{j+1} = 0. \tag{3.12}$$

4. Collisional approach

Although the collisionless Boltzmann scheme has positivity-preserving property, its description of the flow motion is inadquate. As we know, for the compressible Euler equations, the corresponding gas distribution should be Maxwellians. However, the flux function in the collisionless Boltzmann scheme is based on two half Maxwellians at each cell interface. In other words, it introduces intrinsic numerical dissipation and heat conductions. In order to reduce the viscous error, we need in some ways to modify the gas distribution at the cell interface. One of the obvious ways is to use an exact Maxwellian.

4.1. Numerical scheme

We suppose that the initial data $(\rho_0(x), m_0(x), E_0(x))$ are piecewise constant over the cells $C_j = [x_{j-1/2}, x_{j+1/2})$. We further assume that across the cell interface

$$f_1(x_{j+1/2}, t, u, \xi) = \bar{\rho}_{j+1/2} \left(\frac{\bar{\lambda}_{j+1/2}}{\pi}\right)^{\frac{K+1}{2}} e^{-\bar{\lambda}_{j+1/2}\left((u-\bar{U}_{j+1/2})^2+\xi^2\right)}, \tag{4.1}$$

where $\bar{\rho}_{j+1/2}, \bar{\lambda}_{j+1/2}, \bar{U}_{j+1/2}$ are to be determined. In other words, $f_1(x_{j+1/2}, t, u, \xi)$ is a local Maxiwellian distribution. Since mass, momentum and energy are conserved during particle collisions, we have the following condition

$$\int f_1(x_{j+1/2}, t, u, \xi) \begin{pmatrix} 1 \\ u \\ \frac{1}{2}(u^2+\xi^2) \end{pmatrix} du d\xi = \int f_0(x_{j+1/2}, t, u, \xi) \begin{pmatrix} 1 \\ u \\ \frac{1}{2}(u^2+\xi^2) \end{pmatrix} du d\xi, \tag{4.2}$$

where $f_0(x_{j+1/2}, t, u, \xi)$ on the right hand side is given by (3.3). Direct calculation using (4.2) gives

$$\begin{pmatrix} \bar{\rho}_{j+1/2} \\ \bar{\rho}_{j+1/2}\bar{U}_{j+1/2} \\ \frac{1}{2}\bar{\rho}_{j+1/2}\left(\bar{U}^2_{j+1/2} + \frac{K+1}{2\bar{\lambda}_{j+1/2}}\right) \end{pmatrix} = \rho_j \begin{pmatrix} \frac{1}{2}\mathrm{erfc}(-\sqrt{\lambda_j}U_j) \\ \frac{1}{2}U_j\mathrm{erfc}(-\sqrt{\lambda_j}U_j) + \frac{1}{2}\frac{e^{-\lambda_j U_j^2}}{\sqrt{\pi\lambda_j}} \\ \frac{1}{2}\left(\frac{U_j^2}{2} + \frac{K+1}{4\lambda_j}\right)\mathrm{erfc}(-\sqrt{\lambda_j}U_j) + \frac{U_j}{4}\frac{e^{-\lambda_j U_j^2}}{\sqrt{\pi\lambda_j}} \end{pmatrix} \tag{4.3}$$

$$+\rho_{j+1}\begin{pmatrix} \frac{1}{2}\mathrm{erfc}(\sqrt{\lambda_{j+1}}U_{j+1}) \\ \frac{1}{2}U_{j+1}\mathrm{erfc}(\sqrt{\lambda_{j+1}}U_{j+1}) - \frac{1}{2}\frac{e^{-\lambda_{j+1}U_{j+1}^2}}{\sqrt{\pi\lambda_{j+1}}} \\ \frac{1}{2}\left(\frac{U_{j+1}^2}{2}+\frac{K+1}{4\lambda_{j+1}}\right)\mathrm{erfc}(\sqrt{\lambda_{j+1}}U_{j+1}) - \frac{U_{j+1}}{4}\frac{e^{-\lambda_{j+1}U_{j+1}^2}}{\sqrt{\pi\lambda_{j+1}}} \end{pmatrix}$$

The first equation in (4.3) gives $\bar{\rho}_{j+1/2}$, the second equation yields $\bar{U}_{j+1/2}$, and the third one leads to $\bar{\lambda}_{j+1/2}$. Further, using (4.1) and the formulas (2.7), we obtain the numerical fluxes

$$\begin{aligned}\begin{pmatrix} F^1_{\rho,j+1/2} \\ F^1_{m,j+1/2} \\ F^1_{E,j+1/2} \end{pmatrix} &= \int f_1(x_{j+1/2},t,u,\xi)u\begin{pmatrix} 1 \\ u \\ \frac{1}{2}(u^2+\xi^2) \end{pmatrix} dud\xi \\ &= \bar{\rho}_{j+1/2}\begin{pmatrix} \bar{U}_{j+1/2} \\ \bar{U}^2_{j+1/2}+\frac{1}{2\bar{\lambda}_{j+1/2}} \\ \frac{1}{2}\bar{U}^3_{j+1/2}+\frac{K+3}{4\bar{\lambda}_{j+1/2}}\bar{U}_{j+1/2} \end{pmatrix}.\end{aligned} \tag{4.4}$$

Using the above numerical fluxes, we are able to update ρ_j, m_j, E_j with the standard conservative formulations:

$$\begin{pmatrix} \tilde{\rho}_j \\ \tilde{m}_j \\ \tilde{E}_j \end{pmatrix} = \begin{pmatrix} \rho_j \\ m_j \\ E_j \end{pmatrix} + \sigma\begin{pmatrix} F^1_{\rho,j-1/2} - F^1_{\rho,j+1/2} \\ F^1_{m,j-1/2} - F^1_{m,j+1/2} \\ F^1_{E,j-1/2} - F^1_{E,j+1/2} \end{pmatrix}, \tag{4.5}$$

where $F^1_{\rho,j\pm1/2}, F^1_{m,j\pm1/2}, F^1_{E,j\pm1/2}$ are given by (4.4), and $\sigma = \Delta t/\Delta x$. The scheme can be viewed as consisting of the following four steps:

ALGORITHM 1 (Collisional Approach)

1. Given data $\{\rho^n_j, U^n_j, E^n_j\}$, compute $\{\lambda^n_j\}$ using (3.1).
2. Compute $\{\bar{\rho}_{j+1/2}, \bar{U}_{j+1/2}, \bar{\lambda}_{j+1/2}\}$ using (4.3).
3. Compute the numerical flux $\{F^1_{\rho,j+1/2}, F^1_{m,j+1/2}, F^1_{E,j+1/2}\}$ using (4.4).
4. Update $\{\rho^n_j, m^n_j, E^n_j\}$ using (4.5). This gives $\{\rho^{n+1}_j, m^{n+1}_j, E^{n+1}_j\}$.

4.2. Positivity-preserving analysis

If the Algorithm 1 makes sense, then it is required that $\tilde{\rho}_j \geq 0, \tilde{\rho}_j\tilde{E}_j \geq \frac{1}{2}(\tilde{m}_j)^2$. The two requirements will lead to the CFL condition. The collisional scheme is strongly nonlinear, which makes the analysis more difficult. However, it is found for the density $\tilde{\rho}$, the analysis is simple. From (4.3)–(4.5), we obtain

$$\tilde{\rho}_j = \rho_j + \sigma\Big\{ \frac{\rho_{j-1}U_{j-1}}{2}\mathrm{erfc}(-\sqrt{\lambda_{j-1}}U_{j-1}) + \frac{\rho_{j-1}}{2}\frac{e^{-\lambda_{j-1}U_{j-1}^2}}{\sqrt{\pi\lambda_{j-1}}} \tag{4.6}$$
$$+ \frac{\rho_j U_j}{2}\Big(\mathrm{erfc}(\sqrt{\lambda_j U_j}) - \mathrm{erfc}(-\sqrt{\lambda_j U_j})\Big) - \rho_j\frac{e^{-\lambda_j U_j^2}}{\sqrt{\pi\lambda_j}}$$
$$- \frac{\rho_{j+1}U_{j+1}}{2}\mathrm{erfc}(\sqrt{\lambda_{j+1}}U_{j+1}) + \frac{\rho_{j+1}}{2}\frac{e^{-\lambda_{j+1}U_{j+1}^2}}{\sqrt{\pi\lambda_{j+1}}}\Big\}.$$

It can be shown that, for any $U \in \mathbf{R}$ and for any fixed $\lambda > 0$, that

$$U\mathrm{erfc}(-\sqrt{\lambda}U) + \frac{e^{-\lambda U^2}}{\sqrt{\pi\lambda}} \geq 0, \qquad -U\mathrm{erfc}(\sqrt{\lambda}U) + \frac{e^{-\lambda U^2}}{\sqrt{\pi\lambda}} \geq 0. \tag{4.7}$$

This, together with (4.6), yield

$$\tilde{\rho}_j \geq \rho_j + \sigma\left\{\frac{\rho_j U_j}{2}\Big(\mathrm{erfc}(\sqrt{\lambda_j U_j}) - \mathrm{erfc}(-\sqrt{\lambda_j U_j})\Big) - \rho_j\frac{e^{-\lambda_j U_j^2}}{\sqrt{\pi\lambda_j}}\right\}$$
$$\geq \rho_j + \sigma\left\{-\rho_j|U_j| - \rho_j\frac{1}{\sqrt{\pi\lambda_j}}\right\}. \tag{4.8}$$

To ensure that $\tilde{\rho}_j \geq 0$, it is required that

$$\sigma \leq \min_j\left(|U_j| + \frac{1}{\sqrt{\pi\lambda_j}}\right)^{-1}. \tag{4.9}$$

The positivity-preserving analysis for the internal energy is much more complicated than that for the collisionless approach. In fact, an CFL-like condition based on the positivity-preserving analysis seems impossible for the collisional approach. We will consider two special cases in the remaining of the section. The first case uses the assumption (3.12), which is the case that there is only gas flowing out from the cell C_j, but there is no any mass moving into the cell. As pointed out in Remark 3.1, the CFL condition for the collisionless approach is obtained by using (3.12).

4.2.1.Special case I. In order to find a practical CFL number satisfying the positivity-preserving requirements, we consider the extrem case (3.12), i.e.

$$\rho_{j-1} = 0, \qquad \rho_j > 0, \qquad \rho_{j+1} = 0. \tag{4.10}$$

Lemma 4.1. *If (4.10) is satisfied, then*

$$\frac{1}{\bar{\lambda}_{j\pm1/2}} \leq \frac{1}{\lambda_j}. \tag{4.11}$$

Proof. Since $\rho_{j\pm1}=0$, it follows from (4.3) that

$$\bar{\rho}_{j\pm1/2}=\frac{1}{2}\rho_j\mathrm{erfc}(\mp\sqrt{\lambda_j}U_j), \tag{4.12}$$

$$\bar{\rho}_{j\pm1/2}\bar{U}_{j\pm1/2}=\bar{\rho}_{j\pm1/2}U_j\pm\frac{\rho_j}{2}\frac{e^{-\lambda_j U_j^2}}{\sqrt{\pi\lambda_j}}, \tag{4.13}$$

$$\frac{1}{2}\bar{\rho}_{j\pm1/2}\left(\bar{U}_{j\pm1/2}^2+\frac{K+1}{2\bar{\lambda}_{j\pm1/2}}\right)=\frac{1}{2}U_j\left(\bar{\rho}_{j\pm1/2}\bar{U}_{j\pm1/2}\right)+\frac{K+1}{4\lambda_j}\bar{\rho}_{j\pm1/2}. \tag{4.14}$$

Using the above formulas, we can obtain

$$\frac{K+1}{4\bar{\lambda}_{j\pm1/2}}\bar{\rho}_{j\pm1/2}=\frac{K+1}{4\lambda_j}\bar{\rho}_{j\pm1/2}\mp\frac{\rho_j}{4}\frac{e^{-\lambda_j U_j^2}}{\sqrt{\pi\lambda_j}}\bar{U}_{j\pm1/2}. \tag{4.15}$$

It follows from (4.7) that $\bar{U}_{j+1/2}\geq 0$ and $\bar{U}_{j-1/2}\leq 0$. This, together with (4.15), yield (4.11). □

Lemma 4.2. *If (4.10) is satisfied, then*

$$\tilde{\rho}_j\tilde{E}_j-\frac{1}{2}\tilde{m}_j^2\geq-\frac{\rho_j^2}{8\lambda_j^2}\sigma^2-\left(\frac{5(K+1)}{8\lambda_j}\rho_j^2|U_j|+\frac{K+2}{2\lambda_j}\frac{\rho_j^2}{\sqrt{\pi\lambda_j}}\right)\sigma+\frac{K+1}{4\lambda_j}\rho_j^2, \tag{4.16}$$

where $\tilde{\rho}_j,\tilde{m}_j$ and $\tilde{E}_j$ are defined by (4.4) and (4.5).

Proof. It follows from (4.4) and (4.5) that

$$\tilde{\rho}_j\tilde{E}_j-\frac{1}{2}\tilde{m}_j^2=A\sigma^2-B\sigma+C,$$

where

$$\begin{aligned}
A=&\left(\bar{\rho}_{j+1/2}\bar{U}_{j+1/2}-\bar{\rho}_{j-1/2}\bar{U}_{j-1/2}\right)\left(\frac{1}{2}\bar{\rho}_{j+1/2}\bar{U}_{j+1/2}^3-\frac{1}{2}\bar{\rho}_{j-1/2}\bar{U}_{j-1/2}^3\right.\\
&\left.+\frac{K+3}{4\bar{\lambda}_{j+1/2}}\bar{\rho}_{j+1/2}\bar{U}_{j+1/2}-\frac{K+3}{4\bar{\lambda}_{j-1/2}}\bar{\rho}_{j-1/2}\bar{U}_{j-1/2}\right)\\
&-\frac{1}{2}\left(\bar{\rho}_{j+1/2}\bar{U}_{j+1/2}^2-\bar{\rho}_{j-1/2}\bar{U}_{j-1/2}^2+\frac{\bar{\rho}_{j+1/2}}{2\bar{\lambda}_{j+1/2}}-\frac{\bar{\rho}_{j-1/2}}{2\bar{\lambda}_{j-1/2}}\right)^2;\\
B=&\left(\frac{1}{2}\rho_j U_j^2+\frac{K+1}{4\lambda_j}\rho_j\right)\left(\bar{\rho}_{j+1/2}\bar{U}_{j+1/2}\right.\\
&\left.-\bar{\rho}_{j-1/2}\bar{U}_{j-1/2}\right)+\rho_j\left(\frac{1}{2}\bar{\rho}_{j+1/2}\bar{U}_{j+1/2}^3-\frac{1}{2}\bar{\rho}_{j-1/2}\bar{U}_{j-1/2}^3\right.
\end{aligned}$$

$$+\frac{K+3}{4\bar{\lambda}_{j+1/2}}\bar{\rho}_{j+1/2}\bar{U}_{j+1/2}-\frac{K+3}{4\bar{\lambda}_{j-1/2}}\bar{\rho}_{j-1/2}\bar{U}_{j-1/2}\Big)$$
$$-\rho_j U_j\left(\bar{\rho}_{j+1/2}\bar{U}^2_{j+1/2}-\bar{\rho}_{j-1/2}\bar{U}^2_{j-1/2}+\frac{\bar{\rho}_{j+1/2}}{2\bar{\lambda}_{j+1/2}}-\frac{\bar{\rho}_{j-1/2}}{2\bar{\lambda}_{j-1/2}}\right);$$
$$C=\rho_j E_j-\frac{1}{2}m_j^2=\frac{K+1}{4\lambda_j}\rho_j^2.$$

Using the facts that $\bar{U}_{j+1/2}\geq 0, \bar{U}_{j-1/2}\leq 0$, we obtain from the expression for A that

$$A\geq -\frac{1}{2}\left(\frac{\bar{\rho}_{j+1/2}}{2\bar{\lambda}_{j+1/2}}-\frac{\bar{\rho}_{j-1/2}}{2\bar{\lambda}_{j-1/2}}\right)^2\geq -\frac{1}{8}\max\left(\frac{\bar{\rho}^2_{j+1/2}}{\bar{\lambda}^2_{j+1/2}},\ \frac{\bar{\rho}^2_{j-1/2}}{\bar{\lambda}^2_{j-1/2}}\right).$$

It follows from (4.12) that $\bar{\rho}_{j\pm 1/2}\leq \rho_j$. This, together with Lemma 4.1, yield $A\geq -\rho_j^2/8\lambda_j^2$. Further, we split the coefficient B into the following form

$$B=\frac{K+1}{4\lambda_j}\rho_j\left(\bar{\rho}_{j+1/2}\bar{U}_{j+1/2}-\bar{\rho}_{j-1/2}\bar{U}_{j-1/2}\right)+B^+-B^-, \tag{4.17}$$

where

$$B^{\pm}=\frac{1}{2}\rho_j U_j^2\left(\bar{\rho}_{j\pm 1/2}\bar{U}_{j\pm 1/2}\right)+\rho_j\bar{U}_{j\pm 1/2}\left(\frac{1}{2}\bar{\rho}_{j\pm 1/2}\bar{U}^2_{j\pm 1/2}+\frac{K+1}{4\bar{\lambda}_{j\pm 1/2}}\right)$$
$$-\rho_j U_j\bar{U}_{j\pm 1/2}\left(\bar{\rho}_{j\pm 1/2}\bar{U}_{j\pm 1/2}\right)+\frac{1}{2\bar{\lambda}_{j\pm 1/2}}\rho_j\bar{\rho}_{j\pm 1/2}\left(\bar{U}_{j\pm 1/2}-U_j\right).$$

It follows from (4.12)–(4.14) that

$$B^+=\frac{K+1}{4\lambda_j}\rho_j\left(\bar{\rho}_{j+1/2}\bar{U}_{j+1/2}\right)+\frac{\rho_j^2}{4\bar{\lambda}_{j+1/2}}\frac{e^{-\lambda_j U_j^2}}{\sqrt{\pi\lambda_j}}$$
$$+\frac{1}{2}\rho_j U_j\bar{\rho}_{j+1/2}\bar{U}_{j+1/2}(U_j-\bar{U}_{j+1/2}).$$

Using (4.14) gives

$$\rho_j U_j\bar{\rho}_{j+1/2}\bar{U}_{j+1/2}(U_j-\bar{U}_{j+1/2})=\frac{K+1}{4}\left(\frac{1}{\bar{\lambda}_{j+1/2}}-\frac{1}{\lambda_j}\right)\rho_j\bar{\rho}_{j+1/2}U_j$$
$$\leq\frac{K+1}{4\lambda_j}\rho_j\bar{\rho}_{j+1/2}|U_j|,$$

where in the last step we have used the result of Lemma 4.1. Combining the above two results, we obtain

$$B^+\leq\frac{K+1}{4\lambda_j}\rho_j\left(\bar{\rho}_{j+1/2}\bar{U}_{j+1/2}\right)+\frac{\rho_j^2}{4\bar{\lambda}_{j+1/2}}\frac{e^{-\lambda_j U_j^2}}{\sqrt{\pi\lambda_j}}+\frac{K+1}{8\lambda_j}\rho_j\bar{\rho}_{j+1/2}|U_j|.$$

Similarly, we have

$$B^- \geq \frac{K+1}{4\lambda_j}\rho_j\left(\bar{\rho}_{j-1/2}\bar{U}_{j-1/2}\right) - \frac{\rho_j^2}{4\bar{\lambda}_{j-1/2}}\frac{e^{-\lambda_j U_j^2}}{\sqrt{\pi\lambda_j}} - \frac{K+1}{8\lambda_j}\rho_j\bar{\rho}_{j-1/2}|U_j|.$$

It follows from (4.17) and the above two inequalitites that

$$\begin{aligned} B \leq & \frac{K+1}{2\lambda_j}\rho_j\left(\bar{\rho}_{j+1/2}\bar{U}_{j+1/2} - \bar{\rho}_{j-1/2}\bar{U}_{j-1/2}\right) \\ & + \frac{\rho_j^2}{2\lambda_j}\frac{e^{-\lambda_j U_j^2}}{\sqrt{\pi\lambda_j}} + \frac{K+1}{8\lambda_j}\rho_j\left(\bar{\rho}_{j+1/2} + \bar{\rho}_{j-1/2}\right)|U_j|, \end{aligned}$$

where we have used the result of Lemma 4.1. Using (4.12), (4.13) and the fact that $\mathrm{erfc}(-x) + \mathrm{erfc}(x) = 2$, we obtain

$$B \leq \frac{K+1}{2\lambda_j}\rho_j^2\left(|U_j| + \frac{1}{\sqrt{\pi\lambda_j}}\right) + \frac{\rho_j^2}{2\lambda_j}\frac{1}{\sqrt{\pi\lambda_j}} + \frac{K+1}{8\lambda_j}\rho_j^2|U_j|.$$

This completes the proof of this lemma. □

It is easy to show that the positive root of the right hand side polynomial (with respect to σ) of (4.16) satisfies

$$\begin{aligned} \sigma^+ &\geq \left(\frac{5}{2}|U_j| + \frac{2(K+2)}{K+1}\frac{1}{\sqrt{\pi\lambda_j}} + \frac{1}{\sqrt{2(K+1)\lambda_j}}\right)^{-1} \\ &\geq \left(\frac{5}{2}|U_j| + \frac{5}{2}c_j\right)^{-1}, \end{aligned}$$

for all $K \geq 0$, where $c_j = \sqrt{\gamma_j/2\lambda_j}$ is the local speed of sound. In order that the right hand side of (4.16) is positive, we require that

$$\sigma \leq \frac{0.4}{\max_j\left(|U_j| + c_j\right)}. \tag{4.18}$$

Under the above condition, we also obtain from (4.8) that $\tilde{\rho}_j \geq 0$.

We point out that the CFL-like condition (4.18) works well for the available test problems in literature, including those in [5, 9]. For these test problems, we have not found any case with negative internal energy. In particular, the collisional scheme also works for the low density and low internal energy problem proposed in [4].

For the collisionless approach, Lemma 3.2 ensures that a rigourous CFL condition can be obtained by using the assumption (4.10). However, there is no result

similar to Lemma 3.2 that holds for the collisional approach. Is (4.10) again a correct assumption leading to a rigourous CFL condition for the collisional approach?

4.2.2. Special case II.

Physically, the collisional approach is based on the assumption of local equilibrium Maxwellian distribution, in which the viscosity and heat conductivity go to zero automatically. However, in some situations, even for the Euler equations, the gas will not stay on the equilibrium state. The example for this is the discontinuous shocks, where the dissipative terms are crucial to achieve a smooth shock transition by transferring kinetic energy into thermal energy. So, for the strong shock cases, the collisional approach has intrinsic weakness and it will definitely give some mal-behavior due to the incorrect physical assumptions. Let us consider the following example. A stationary shock is located at cell interface $x = 0$ with distributions (ρ_1, U_1, P_1) and (ρ_2, U_2, P_2) on the left and right sides. The upstream and downstream flow conditions are

$$\begin{cases} \rho_1 = 1, \\ U_1 = 1, \qquad x \le 0, \\ P_1 = \frac{1}{\gamma M^2}, \end{cases} \tag{4.19}$$

and

$$\begin{cases} \rho_2 = \rho_1 \frac{\gamma+1}{2} M^2 \Big/ \left(1 + \frac{\gamma-1}{2} M^2\right), \\ U_2 = U_1 \left(\dfrac{\gamma-1}{\gamma+1} + \dfrac{2}{(\gamma+1)M^2}\right), \qquad x > 0, \\ P_2 = P_1 \left(\dfrac{2\gamma}{\gamma+1} M^2 - \dfrac{\gamma-1}{\gamma+1}\right), \end{cases} \tag{4.20}$$

where $\gamma = 1.4$, M is the Mach number. In Fig 2, we plot the internal energy $\tilde{\rho}_0 \tilde{E}_0 - \frac{1}{2}\tilde{m}_0^2$, where $\tilde{\rho}_0, \tilde{E}_0$ and $\tilde{m}_0$ are computed by the collisional scheme Algorithm 1, with $\sigma = 0.2/\max(|U_j| + c_j)$, $0.4/\max(|U_j| + c_j)$, $0.6/\max(|U_j| + c_j)$ and $\sigma = 1/\max(|U_j| + c_j)$.

It is seen from Fig 2 that for $\sigma = 0.4/\max(|U_j| + c_j)$, the internal energy becomes negative if $M \ge 25$. Even with unrealisticly small $\sigma = 0.2/\max(|U_j| + c_j)$, the internal energy becomes negative for larger values of M.

The above analysis suggests that, unlike the collisionless approach, the collisional Boltzmann scheme is not positivity-preserving. Therefore, some modifications are required in order to take the advantage of the less-dissipation property in the collisional scheme.

5. The BGK approach

Collisionless and collisional schemes are two extreme limits in the description of flow motion. In the smooth flow region, collisional approach gives correct flux

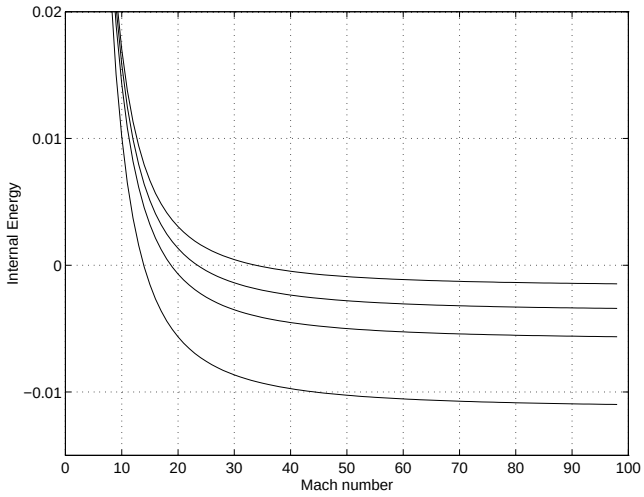

Figure 2.
The internal ebergy: from top to bottom are obtained by using $\sigma = 0.2/\max(|U_j| + c_j)$, $0.4/\max(|U_j| + c_j)$, $0.6/\max(|U_j| + c_j)$ and $\sigma = 1/\max(|U_j| + c_j)$.

functions by its equlibrium Maxwellian. However, in the discontinuous region, the real physical gas distribution function should be a non-Maxwellian and all physical dissipative mechanism are included in the deviation of the distribution function away from the equilibrium state. So, in order to correctly describe both smooth and discontinuous flow motions we need somehow to keep both collisionless and collisional distributions. We hope also by doing this that the positivity-preserving property is recovered. The linear combination of the collisionless and collisional approaches forms the BGK scheme.

5.1. Numerical scheme

We suppose that the initial data $(\rho_0(x), m_0(x), E_0(x))$ are piecewise constant over the cells $C_j = [x_{j-1/2}, x_{j+1/2})$. We assume that across the cell interface we have

$$f_2(x_{j+1/2}, t, u, \xi) = \alpha(t) f_0(x_{j+1/2}, t, u, \xi) + \beta(t) f_1(x_{j+1/2}, t, u, \xi), \tag{5.1}$$

where $\alpha(t) > 0, \beta(t) > 0$ are given coefficients satisfying $\alpha(t) + \beta(t) \equiv 1$, f_0 and f_1 are given by (3.3) and (4.1), respectively. The corresponding numerical fluxes are given by

$$\begin{pmatrix} F^2_{\rho,j+1/2} \\ F^2_{m,j+1/2} \\ F^2_{E,j+1/2} \end{pmatrix} = \alpha(t) \begin{pmatrix} F^0_{\rho,j+1/2} \\ F^0_{m,j+1/2} \\ F^0_{E,j+1/2} \end{pmatrix} + \beta(t) \begin{pmatrix} F^1_{\rho,j+1/2} \\ F^1_{m,j+1/2} \\ F^1_{E,j+1/2} \end{pmatrix}, \tag{5.2}$$

where $F^0_{\rho,j+1/2}, F^0_{m,j+1/2}, F^0_{E,j+1/2}$ are given by (3.4), $F^1_{\rho,j+1/2}, F^1_{m,j+1/2}, F^1_{E,j+1/2}$ are given by (4.4). Using the above numerical fluxes, we update ρ_j, m_j, E_j based on the following conservative formulations:

$$\begin{pmatrix} \tilde{\rho}_j \\ \tilde{m}_j \\ \tilde{E}_j \end{pmatrix} = \begin{pmatrix} \rho_j \\ m_j \\ E_j \end{pmatrix} + \sigma\alpha(t) \begin{pmatrix} F^0_{\rho,j-1/2} - F^0_{\rho,j+1/2} \\ F^0_{m,j-1/2} - F^0_{m,j+1/2} \\ F^0_{E,j-1/2} - F^0_{E,j+1/2} \end{pmatrix} + \sigma\beta(t) \begin{pmatrix} F^1_{\rho,j-1/2} - F^1_{\rho,j+1/2} \\ F^1_{m,j-1/2} - F^1_{m,j+1/2} \\ F^1_{E,j-1/2} - F^1_{E,j+1/2} \end{pmatrix}. \tag{5.3}$$

The scheme can be viewed as consisting of the following four steps:

ALGORITHM 2 (BGK scheme)

1. Given data $\{\rho^n_j, U^n_j, E^n_j\}$, compute $\{\lambda^n_j\}$ using (3.1).
2. Compute the numerical flux $\{F^0_{\rho,j+1/2}, F^0_{m,j+1/2}, F^0_{E,j+1/2}\}$ using (3.4).
3. Compute the numerical flux $\{F^1_{\rho,j+1/2}, F^1_{m,j+1/2}, F^1_{E,j+1/2}\}$ using (4.4).
5. Update $\{\rho^n_j, m^n_j, E^n_j\}$ using (5.3). This gives $\{\rho^{n+1}_j, m^{n+1}_j, E^{n+1}_j\}$.

5.2. Positivity-preserving analysis

Let (ρ^0_j, m^0_j, E^0_j) and (ρ^1_j, m^1_j, E^1_j) be computed by the right hand sides of (3.5) and (4.5), respectively. Namely,

$$\begin{pmatrix} \rho^0_j \\ m^0_j \\ E^0_j \end{pmatrix} = \begin{pmatrix} \rho_j \\ m_j \\ E_j \end{pmatrix} + \sigma \begin{pmatrix} F^0_{\rho,j-1/2} - F^0_{\rho,j+1/2} \\ F^0_{m,j-1/2} - F^0_{m,j+1/2} \\ F^0_{E,j-1/2} - F^0_{E,j+1/2} \end{pmatrix}, \tag{5.4}$$

and

$$\begin{pmatrix} \rho^1_j \\ m^1_j \\ E^1_j \end{pmatrix} = \begin{pmatrix} \rho_j \\ m_j \\ E_j \end{pmatrix} + \sigma \begin{pmatrix} F^1_{\rho,j-1/2} - F^1_{\rho,j+1/2} \\ F^1_{m,j-1/2} - F^1_{m,j+1/2} \\ F^1_{E,j-1/2} - F^1_{E,j+1/2} \end{pmatrix}. \tag{5.5}$$

Since $\alpha(t) + \beta(t) = 1$, the formulas (5.3) can be written in the following form

$$\begin{pmatrix} \tilde{\rho}_j \\ \tilde{m}_j \\ \tilde{E}_j \end{pmatrix} = \alpha(t) \begin{pmatrix} \rho^0_j \\ m^0_j \\ E^0_j \end{pmatrix} + \beta(t) \begin{pmatrix} \rho^1_j \\ m^1_j \\ E^1_j \end{pmatrix}. \tag{5.6}$$

Theorem 5.1. *Assume that under the CFL condition $\sigma \le \tilde{\sigma}$ the following positivity–preserving properties hold*

$$\rho^0_j \ge 0, \qquad \rho^1_j \ge 0, \tag{5.7}$$

$$\rho^0_j E^0_j \ge \frac{1}{2}(m^0_j)^2, \qquad \rho^1_j E^1_j \ge \frac{1}{2}(m^1_j)^2, \tag{5.8}$$

for all j. Then for any $\sigma \le \tilde{\sigma}$, we have

$$\tilde{\rho}_j \ge 0, \qquad \tilde{\rho}_j \tilde{E}_j \ge \frac{1}{2}(\tilde{m}_j)^2, \tag{5.9}$$

for any j, where $(\tilde{\rho}_j, \tilde{m}_j, \tilde{E}_j)$ are given by (5.3).

Proof. It follows from (5.6) that $\tilde{\rho}_j \ge 0$ if (5.8) is satisfied. Moreover, if (5.8) holds, then

$$\begin{aligned}
\tilde{\rho}_j \tilde{E}_j &= \left(\alpha \rho_j^0 + \beta \rho_j^1\right)\left(\alpha E_j^0 + \beta E_j^1\right) \\
&\ge \left(\alpha \sqrt{\rho_j^0 E_j^0} + \beta \sqrt{\rho_j^1 E_j^1}\right)^2 \\
&\ge \frac{1}{2}\left(\alpha m_j^0 + \beta m_j^1\right)^2,
\end{aligned}$$

where in the last step we have used (5.8). This completes the proof of the theorem. □

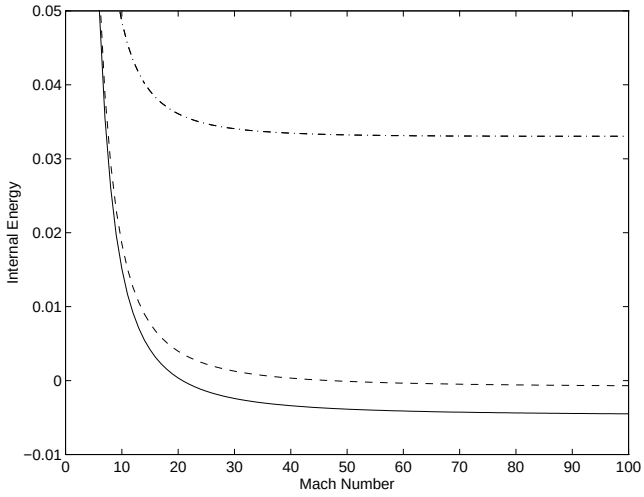

Figure 3.
The internal energy obtained by using the BGK scheme with $\sigma = 0.5/\max(|U_j| + c_j)$: from top to bottom are obtained with $(\alpha, \beta) = (0.1, 0.9), (0.01, 0.99)$, and $(0, 1)$.

The above theorem implies that the CFL condition (based on positivity of the density and the internal energy) for the BGK schemes can be determined by

those for the collisionless and collisional schemes. Since the collisionless approach is positivity-preserving under the standard CFL condition, this theorem suggests that from the positivity-preserving point of view, the BGK scheme is not worse than the collisional scheme. In fact, the numerical experiemnts show that with small contributions from the collisionless part, the difficulties with negative internal energy encounted in the collisional approach can be overcome. To see this, we consider the same example as considered in Section 4.2.2. In Fig 3, we plot the internal energy as defined in Section 4.2.2 by using the scheme (5.6), with $\sigma = 0.5/\max(|U_j| + c_j)$. The top curve is obtained with $(\alpha, \beta) = (0.1, 0.9)$, the middle one is obtained with $(\alpha, \beta) = (0.01, 0.99)$, and the bottom one is obtained with $(\alpha, \beta) = (0, 1)$. It is noted that by adding 1% contribusions from the collisionless approach the trouble with negative internal energy is almost disappeared. If 10% collisionless contribusions are added, then the internal energy is well above the axis of $e = 0$.

It should be pointed out that for all of the three schemes investigated in this work, the density is always positive as long as the requirement (4.9) is satisfied. This requirement is weaker than the standard CFL-condition and is always satisfied. So the main concern is to ensure that the internal energy is non-negative.

With test problems available in literature the most difficult case in keeping positivity of density and internal energy is strong rarefaction waves (see, e.g. [4, 16]). We will test the BGK schemes for two typical examples. In the following computation $\gamma = 1.4$.

Example 1. Low density and internal energy Riemann problem [4] with initial data,

$$(\rho, U, p) = \begin{cases} (1, -2, 0.4) & 0 \le x < 0.5, \\ (1, 2, 0.4) & 0.5 \le x \le 1. \end{cases}$$

The CFL number used in computation is $0.9/\max(|U_j| + c_j)$ and 100 grid points are used. In Fig 4 we plot the logarithm of minimal internal energy

$$e = \min_j \left\{ \rho_j^n E_j^n - \frac{1}{2}\left(m_j^n\right)^2 \right\}.$$

It is observed that for both collisionless and collisional approaches, the internal energy is always positive. By Theorem 5.1, the internal energy is also positive for the corresponding BGK schemes.

Example 2. Vacuum apparition [5] with initial data,

$$(\rho, U, p) = \begin{cases} (1, -5, 0.4) & 0 \le x < 0.5, \\ (1, 5, 0.4) & 0.5 \le x \le 1. \end{cases}$$

The CFL number is $0.9/\max(|U_j| + c_j)$ and 100 grid points are used. We plot the minimal internal energy in Fig 5. Again, it is observed that for both collisionless

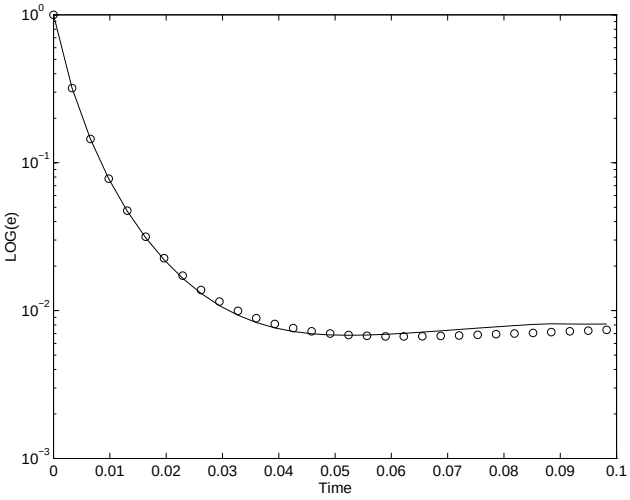

Figure 4.
The history of the minimal internal energy obtained by using the collisionless approach (circle) and the collisional approach (solid) for Example 1.

and collisional approaches the internal energy is always positive. By Theorem 5.1, the internal energy is also positive for the corresponding BGK schemes. We also present numerical results for the density and velocity from the collisional approach at $t = 0.05$ in Figs 6 and 7. As we can see from these figures, even with $\rho \sim 10^{-9}$, the internal energy is still positive and the flows are moving away from each other to form the vaccum at center.

Finally, we point out that the ALGORITHM 2 can also be extended to second order BGK schemes that yield finer resolution, while keeping the positivity-preserving property.

6. Conclusion

In this paper, we have investigated the positivity-preserving property for the first order gas-kinetic schemes, which include collisionless, collisional and BGK type Boltzmann schemes. The positivity-preserving property is a necessary condition for any numerical schemes in order to get physically reasonable solutions. However, it should be pointed out that this property is not sufficient for obtaning a good flow solvers. The positivity property in the gas-kinetic schemes is closely related to the nonlinear coupling in the flow variables in the construction of positive gas distribution functions, and this coupling is hardly satisfied in any other schemes where the flow equations are updated separately, such as Lax-Friedrichs method.

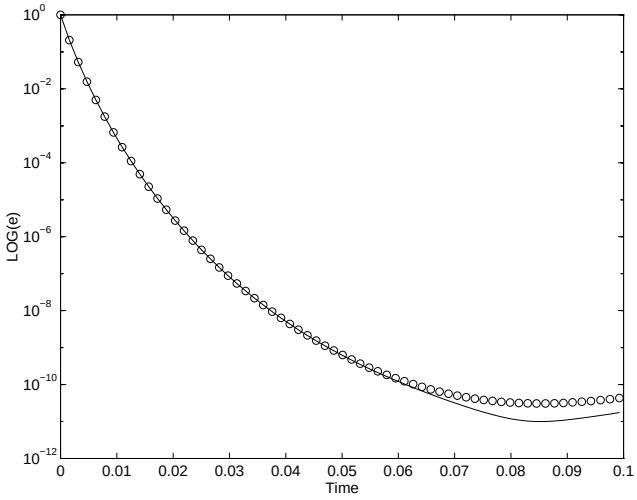

Figure 5.
The history of the minimal internal energy obtained by using the collisionless approach (circle) and the collisional approach (solid) for Example 2.

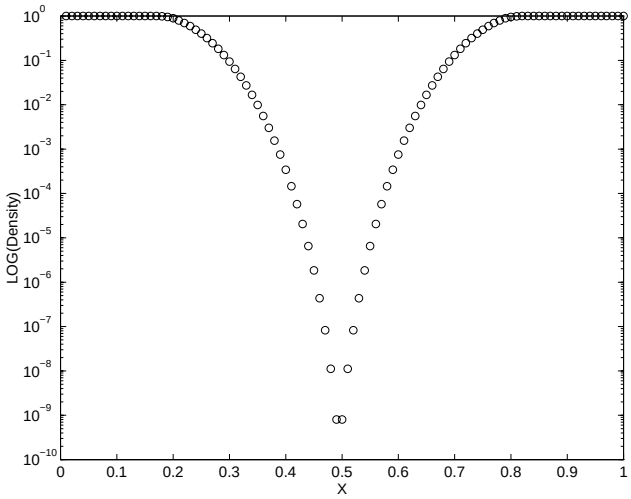

Figure 6.
The density at $t = 0.05$ obtained by using the collisional approach for Example 2.

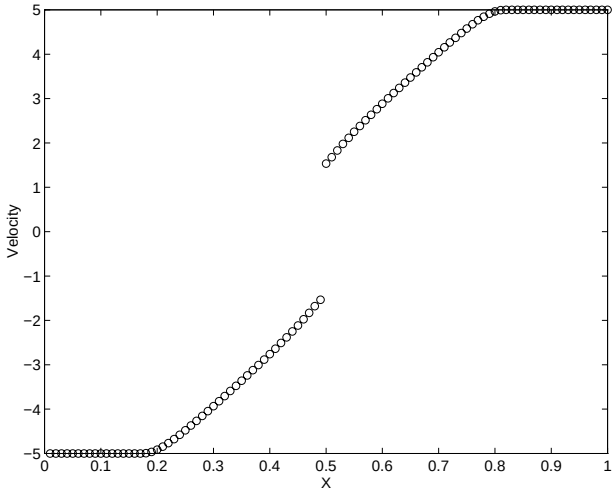

Figure 7.
The velocity at $t = 0.05$ obtained by using the collisional approach.

For these schemes with nonlinear coupling through certain averages, i.e. Roe's scheme, the positivity is also hardly satisfied because the gas evolution is not as simple as upwinding in the scalar equations. Even for the exact Riemann solver, it is also questionable about its solutions in the gas-vaccum interaction cases. The Godunov method and Boltzmann schemes are two main branches in the algorithm development for the compressible flow simulations (not necessarily Euler equations). From the physical point of view, the gas-kinetic schemes give a correct description in a more general flow situations. The commonly anomalous phenomena in the exact and approximate Riemann solvers, such as carbuncle phenomena, post-shock oscillations, odd-even decoupling, negative temperature, can be avoided.

Acknowledgments

This research was done during TT's visit to Hong Kong University of Science and Technology. The hospitality of Professor W.H. Hui and Dr. Xiaoping Wang is gratefully acknowledged.

References

[1] P.L. Bhatnagar, E.P. Gross, and M. Krook, A model for collision processes in gases I: small

amplitude processes in charged and neutral one-component systems, *Phys. Rev.* **94** (1954), 511-525.

[2] S. M. Deshpande, On the Maxwellian Distribution, Symetric Form and Entropy Conservation for the Euler Equations, *Tech. Report* **2583**, NASA Langley, 1986.

[3] S. M. Deshpande, Kinetic Theory Based New Upwind Methods for Inviscid Compressible Flows, *AIAA Paper* **86-0275**, New York, 1986.

[4] B. Enfield, C.D. Munz, P. L. Roe and B. Sjogreen, On Godunov type methods near low density, *J. Comput. Phys.* **92** (1991), 273-295.

[5] J. L. Estivalezes and P. Villedieu, High-order positivity-preserving kinetic schemes for the compressible Euler equations, *SIAM J. Numer. Anal.* **33** (1996), 2050-2067.

[6] A. Harten, P. D. Lax and B. Van Leer, On upstream differencing and Godunov-type schemes for hyperbolic conservation laws, *SIAM Review* **25** (1983), 35-61.

[7] G.-S. Jiang and E. Tadmor, Non-oscillatory central schemes for multidimensional hyperbolic conservation laws, *SIAM J. Sci. Comput.*. **19** (1998), 1892–1917.

[8] M.N. Kogan, *Rarefied Gas Dynamics*, Plenum Press, New York 1969.

[9] X.-D. Liu and P. D. Lax, Positive schemes for solving multi-dimensional hyperbolic sysytems of conservation laws, *CFD J.* **5** (1996), 1–24.

[10] B. Perthame, Second-Order Boltzmann schemes for compressible Euler equation in one and two space dimensions, *SIAM J. Numer. Anal.* **29**(1) (1992).

[11] B. Perthame and C. W. Shu, On positivity preserving finite volume schemes for Euler equations, *Numer. Math.* **73** (1996), 119-130.

[12] K.H. Prendergast and K. Xu, Numerical Hydrodynamics from Gas-Kinetic Theory, *J. of Comput. Phys.* **109** (1993), 53.

[13] D. I. Pullin, Direct simulations methods for compressible inviscid ideal gas-flows, *J. Comput. Phys.* **34** (1980), 231-244.

[14] P.L. Roe, Approximate Riemann solvers, parameter vectors and difference schemes, *J. Comput. Phys.* **43** (1981), 357.

[15] R. Sanders and K. Prendergast, The possible relation of the three-kiloparsec arm to explosions in the galactic nucleus, in *Astrophysical J.* **188** (1974).

[16] G. Toth and D. Odstrcil, Comparison of some flux corrected transport and total variation diminishing numerical schemes for hydrodynamic and magnetohydrodynamic problems, *J. Comput. Phys.* **128** (1996), 82-100.

[17] K. Xu, *Numerical Hydrodynamics from Gas-Kinetic Theory*, Ph.D thesis, Columbia University, 1993.

[18] K. Xu and K. Prendergast, Numerical Navier-Stokes solutions from gas-kinetic theory, *J. Comput. Phys.* **114** (1994), 9-17.

[19] K. Xu, L. Martinelli and A. Jameson, Gas-kinetic finite volume methods, flux-vector splitting and artificial diffusion, *J. Comput. Phys.* **120** (1995), 48-65.

Tao Tang
Department of Mathematics and Statistics
Simon Fraser University
Burnaby, British Columbia
Canada V5A 1S6.

Kun Xu
Department of Mathematics
Hong Kong University of
Science and Technology
Clear Water Bay
Hong Kong

(Received: October 27, 1997)