arXiv:1812.08320v1 [math.NA] 14 Dec 2018

STABILITY ANALYSIS OF QUADRATURE-BASED MOMENT METHODS FOR KINETIC EQUATIONS*

QIAN HUANG†, SHUIQING LI‡, AND WEN-AN YONG§

Abstract. In this paper, we present a systematic stability analysis of the quadrature-based moment method (QBMM) for the one-dimensional Boltzmann equation with BGK or Shakhov models. As reported in recent literature, the method has revealed its potential for modeling non-equilibrium flows, while a thorough theoretical analysis is largely missing but desirable. We show that the method can yield non-hyperbolic moment systems if the distribution function is approximated by a linear combination of δ-functions. On the other hand, if the δ-functions are replaced by their Gaussian approximations with a common variance, we prove that the moment systems are strictly hyperbolic and preserve the dissipation property (or H-theorem) of the kinetic equation. In the proof we also determine the equilibrium manifold that lies on the boundary of the state space. The proofs are quite technical and involve detailed analyses of the characteristic polynomials of the coefficient matrices.

Key words. quadrature based moment methods, Boltzmann equation, structural stability condition, hyperbolicity, BGK and Shakhov models

AMS subject classifications. 35Q79, 76P05, 82-08

1. Introduction.

Kinetic theories pioneered by L. Boltzmann arise in a variety of fields beyond the classical rarefield gas dynamics, ranging from multiphase flows [14, 19], aerosol dynamics in atmospheric environments [10, 19], and active matter physics [13], to galactic dynamics in the universe [29]. In the kinetic framework [16], various physical systems are described with a distribution function f which depends on the spatial and other problem-specific microscopic variables and its time evolution is governed by kinetic equations like the Boltzmann equation. Although the kinetic equations have solid physical ground, they are computationally costly and therefore not directly usable in engineering applications.

Because of the above reason, various simplifications or approximations of the kinetic equations have been proposed, including the BGK model [1], discrete velocity models [11, 21], and moment closure systems [12, 17, 19]. All these approximations have their advantages and disadvantages. This work is concerned with moment closure systems, in which the governing equations of several moments of the distribution function are derived from the kinetic equation and an additional procedure must be accompanied to close the moment system [19]. The resultant moment systems consist usually of first-order partial differential equations (PDEs).

To correctly model the observability of physical processes, the derived system of PDEs should be well-posed (or hyperbolic for first-order systems). For instance, the well-known Grad's closure method yields non-hyperbolic PDEs and produces unphysical results [12, 22]. Its hyperbolic regularization has attracted much attention [2, 3, 17, 20, 26]. A recent work is [4] where the authors introduced a framework to construct hyperbolic moment closure systems.

*Submitted to the editors DATE.

Funding: This work was funded by China Postdoctoral Science Foundation (under contract no. 043201001) and National Natural Science Foundation of China (no. 51725601 and 11471185).

†Zhou Pei-Yuan Center for Applied Mathematics, Tsinghua University, Beijing 100084, China (hqqh91@qq.com, https://www.researchgate.net/profile/Qian_Huang34).

‡Department of Energy and Power Engineering, Tsinghua University, Beijing 100084, China (lishuiqing@tsinghua.edu.cn, http://www.thu-lishuiqing.org).

§Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China (wayong@tsinghua.edu.cn, https://www.researchgate.net/profile/Wen-An_Yong).

Furthermore, the moment closure systems derived from the kinetic equations should preserve the key physical properties of the original kinetic equations. For the Boltzmann equation, one of the key properties is the celebrated H-theorem characterizing the dissipation property of the mesoscopic system under consideration [16]. In this regard, a paradigm is the widely used BGK model that not only simplifies the collision term in the Boltzmann equation, but also inherits the key conservation and dissipation properties thereof [16]. At this point, an immediate question is how to manifest the H-theorem in such moment systems.

It turns out that the structural stability condition proposed in [27] for hyperbolic relaxation systems is a proper counterpart of the H-theorem for the kinetic equation. Indeed, this condition has been tacitly respected by many well-developed physical theories [28]. Recently, it was shown in [7] to be satisfied by the hyperbolic regularization models derived in [2, 3, 4]. In contrast, the Biot/squirt (BISQ) model for wave propagation in saturated porous media violates this condition and thus allows exponentially exploding asymptotic solutions [18]. On the other hand, this condition also implies that the resultant moment system is compatible with the classical theories [27]. The implication is important because the lower-order moments are usually associated with the macroscopic parameters of the system [16]. Therefore, we believe that the structural stability condition is a proper criterion to evaluate the moment closure systems.

The objective of this paper is to investigate whether or not the quadrature-based moment method (QBMM, [19]) yields hyperbolic PDEs which satisfy the structural stability condition above. In QBMM, the distribution function f is approximated with a linear combination of N ($N \geq 1$) δ-functions with unknown centers or their Gaussian approximations with unknown variance and centers (named QMOM or EQMOM, respectively) [19]. QBMM has become an effective and popular method in simulating the evolution of fine particulate matter, where the distribution function is independent of the particle velocity and the resultant governing equation is termed population balance equation [19, 23, 30]. However, the QMOM-derived moment system of the Boltzmann equation leads to unphysical shocks in the numerical solution of Riemann problems [9], which is confirmed by our own numerical results (see the Supplementary Material). Thus it is appealing to find the cause for the irregular behaviors and the aforementioned criteria are expected to be useful in clarifying such issues.

This paper deals only with the spatial one-dimensional (1-D) Boltzmann equation with hypothetical collisions (BGK or Shakhov type), just to figure out a road map for further investigations of general cases. We show that the QMOM-derived moment system is not strongly hyperbolic for any number N of nodes, while the Gaussian EQMOM produces strictly hyperbolic moment systems when the variance is positive. For the latter, we further determine their equilibrium manifolds and verify the structural stability condition. The proofs are quite technical and purely analytic. They involve detailed analyses of characteristic polynomials of the coefficient matrices.

Let us remark that for $N = 2$, the hyperbolicity of moment systems has been studied in [6] for 1-D QMOM and in [5] for 1-D Gaussian-EQMOM. The proofs rely on direct calculations of the eigenvalues of the coefficient matrix of the moment systems [6, 5] and does not seem generalizable to N-node systems. Thus new techniques are needed to handle the general cases. Moreover, the stability of EQMOM has not been analyzed in the existing literature. Given our positive results, EQMOM reveals its potential in solving a wider range of kinetic equations.

The paper is organized as follows. Section 2 presents a brief introduction on QBMM (QMOM and EQMOM) and states our main results. Section 3 is devoted to

a proof of non-hyperbolicity of QMOM for N-node systems. In Section 4, we verify the structural stability condition for the EQMOM with N nodes. In particular, the hyperbolicity is demonstrated in Subsection 4.2, the equilibrium states are determined in Subsection 4.3, and the dissipation property is shown in Subsections 4.4 and 4.5. Finally, we conclude our paper in Section 5.

2. Preliminaries. For simplicity, we only consider a hypothetical 1-D ideal gas with the probability density function $f = f(t, x, \xi)$ of time $t \in \mathbb{R}^+$, spatial position $x \in \mathbb{R}$ and velocity $\xi \in \mathbb{R}$. The temporal evolution of f is governed by the Boltzmann equation [16]:

$$\frac{\partial f}{\partial t} + \xi \frac{\partial f}{\partial x} = Q(f). \tag{2.1}$$

Here the volumetric force is neglected and the right-hand side $Q(f)$ represents the collisions. As a standard assumption [16], $Q = Q(f)$ has only 1, ξ and ξ^2 as locally conserved quantities:

$$\int_{\mathbb{R}} Q(f)\phi(\xi)d\xi = 0, \qquad \phi(\xi) = 1,\ \xi,\ \xi^2, \tag{2.2}$$

and vanishes at a local equilibrium distribution

$$f_{eq} = f_{eq}(t, x, \xi) = \frac{\rho}{(2\pi\theta)^{1/2}} \exp\left(-\frac{(\xi - U)^2}{2\theta}\right), \tag{2.3}$$

where ρ, U and θ are the density, velocity and temperature of the gas, respectively. They are the classical macroscopic parameters related to f as

$$\rho = \int_{\mathbb{R}} f d\xi, \quad \rho U = \int_{\mathbb{R}} \xi f d\xi, \quad \rho\theta = \int_{\mathbb{R}} (\xi - U)^2 f d\xi, \tag{2.4}$$

In this paper, we mainly consider the BGK model [1], where

$$Q = Q_{BGK}(f) = \nu(f_{eq} - f). \tag{2.5}$$

Here ν is the collision frequency. This simple model has been widely used since it preserves several key properties of the kinetic equation, including (2.2) and the H-theorem. Because the BGK model results in the Prantle number $Pr = 1$, inconsistent with most realistic cases [16], the Shakhov model was proposed [25]:

$$Q_S(f) = \nu(f_S - f). \tag{2.6}$$

Here an alternative equilibrium distribution f_S is assumed:

$$f_S = f_{eq} \times \left(1 + \frac{(1 - Pr)q(\xi - U)}{3\rho\theta^2}\left(\frac{(\xi - U)^2}{\theta} - 3\right)\right) \tag{2.7}$$

with q the heat flux defined as $q = \int_{\mathbb{R}} \frac{1}{2}(\xi - U)^3 f d\xi$.

Denote by $M_j(t, x) = \int_{\mathbb{R}} \xi^j f d\xi$ the jth velocity-moment of f. From (2.4) we see that

$$M_0 = \rho,\ M_1 = \rho U,\ M_2 = \rho(U^2 + \theta),\ M_3 = \rho(U^3 + 3U\theta) + 2q. \tag{2.8}$$

The evolution equation for M_j can be derived from the Boltzmann equation (2.1) with the BGK collision (2.5):

$$\partial_t M_j + \partial_x M_{j+1} = \nu \left[\rho \Delta_j(U,\theta) - M_j\right]. \tag{2.9}$$

Here $\Delta_j(U,\theta)$ denotes the jth moment of the normalized Gaussian distribution

$$\delta_\theta(\xi;U) = \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{(\xi-U)^2}{2\theta}\right).$$

Notice that $\Delta_0(U,\theta) = 1$ and $\Delta_1(U,\theta) = U$.

There are infinitely many equations in (2.9). The first N equations for moments $M_0, \ldots, M_{N-1}$ are not closed, because the M_{N-1}-equation contains the term $\partial_x M_N$. Hence a closure method is needed.

In the rest of this section, we introduce the QBMM methods, the structural stability condition for hyperbolic relaxation systems, and our main results of this paper.

2.1. Quadrature-based moment methods. In QBMM, the lower-order moments determine the weights and nodes of the quadrature for the integration $\int f(\xi)g(\xi)d\xi$. Then the unclosed term can be expressed in terms of the lower-order moments and thereby the closure is done [19].

2.1.1. Quadrature method of moment (QMOM). In QMOM, the distribution function f is assumed to be a sum of N Dirac delta functions

$$f(\xi) = \sum_{i=1}^{N} w_i \delta(\xi - u_i). \tag{2.10}$$

In order to determine the weights w_i and nodes u_i, the first $2N$ lower-order moments $M_0, \ldots, M_{2N-1}$ are employed:

$$M_j = \sum_{i=1}^{N} w_i u_i^j \quad \text{for } j = 0, \ldots, 2N-1. \tag{2.11}$$

These non-linear algebraic equations can be solved to obtain w_i and u_i as in [19]. Then the next moment M_{2N} can be found as

$$\bar{M}_{2N} = \sum_{i=1}^{N} w_i u_i^{2N}. \tag{2.12}$$

Namely, w_i and u_i are functions of $M_1, \ldots, M_{2N-1}$, and so is $\bar{M}_{2N}$. In this way, we obtain the following system of PDEs:

$$\partial_t M + A(M)\partial_x M = \nu S(M). \tag{2.13}$$

Here $M = (M_0, \ldots, M_{2N-1})^T \in \mathbb{R}^{2N}$, $S(M) = \rho(\Delta_0(U,\theta), \ldots, \Delta_{2N-1}(U,\theta))^T - M$, and

$$A(M) = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ a_0 & a_1 & \cdots & a_{2N-2} & a_{2N-1} \end{bmatrix}, \tag{2.14}$$

with $a_j = \frac{\partial \bar{M}_{2N}}{\partial M_j}$ for $0 \le j \le 2N-1$.

2.1.2. Extended-QMOM (EQMOM). In order to improve QMOM [5], the delta function in (2.10) is replaced with its Gaussian approximation

$$\delta_{\sigma^2}(\xi;u)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(\xi-u)^2}{2\sigma^2}\right),$$

that is,

$$f(\xi)=\sum_{i=1}^{N}w_i\delta_{\sigma^2}(\xi;u_i). \tag{2.15}$$

Set $W=(w_1,u_1,\dots,w_N,u_N,\sigma^2)^T\in\mathbb{R}^{2N+1}$ and $M=(M_0,\dots,M_{2N})^T\in\mathbb{R}^{2N+1}$. They are related with the map $M=\mathcal{M}(W)$:

$$M_j=\sum_{i=1}^{N}w_i\Delta_j(u_i,\sigma^2)\quad\text{for } j=0,\dots,2N, \tag{2.16}$$

defined for $W\in\Omega_W=\Omega_W^{open}\cup\Omega_W^{eq}$, where

$$\Omega_W^{open}=\{W:\ w_i>0;\ \sigma^2>0;\ \forall\, i\neq j,\ u_i\neq u_j\}, \tag{2.17a}$$

$$\Omega_W^{eq}=\{W:\ w_i>0;\ \sigma^2>0;\ u_1=u_2=\cdots=u_N\}. \tag{2.17b}$$

Remark that $\Delta_j(u_i,\sigma^2)$ is exactly the same as that in (2.9). It is shown in Appendix A that the map $M=\mathcal{M}(W)$ is one-to-one for $W\in\Omega_W^{open}$. Therefore, W can be uniquely solved from (2.16) for $M\in\mathcal{M}(\Omega_W^{open})$. In this way, the next moment M_{2N+1} is a function of the lower-order moments $M\in\mathcal{M}(\Omega_W^{open})$:

$$\bar{M}_{2N+1}=\sum_{i=1}^{N}w_i\Delta_{2N+1}(u_i,\sigma^2). \tag{2.18}$$

Therefore, the following moment system is derived:

$$\partial_t M+A(M)\partial_x M=\nu S(M) \tag{2.19}$$

for $M\in\mathcal{M}(\Omega_W^{open})$. Here $S(M)=\rho\left(\Delta_0(U,\theta),\dots,\Delta_{2N}(U,\theta)\right)^T-M$ and

$$A(M)=\begin{bmatrix}0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1\\ a_0 & a_1 & \cdots & a_{2N-1} & a_{2N}\end{bmatrix} \tag{2.20}$$

with $a_j=\frac{\partial\bar{M}_{2N+1}}{\partial M_j}$ for $0\le j\le 2N$.

For such systems, the moment set $\mathcal{M}(\Omega_W)$ and its closure $\overline{\mathcal{M}(\Omega_W)}$ have been extensively studied as a realizability issue in the literature [5, 19, 23, 24]. A further discussion on this issue is beyond the scope of this paper.

2.2. Structural stability condition. The both QMOM and EQMOM moment systems consist of first-order PDEs derived from the Boltzmann equation. To clarify whether or not these systems inherits the H-theorem characterizing the dissipation property of the Boltzmann equation, we recall the structural stability condition proposed in [27] for systems of D-dimensional PDEs:

$$\frac{\partial M}{\partial t} + \sum_{d=1}^{D} A_d(M) \frac{\partial M}{\partial x_d} = S(M). \tag{2.21}$$

Here M is the unknown n-vector valued function, $A_d = A_d(M)$ is the dth $n \times n$ coefficient matrix, and the source term $S = S(M)$ is a given n-vector valued function of $M \in \mathbb{G} \subset \mathbb{R}^n$. As in [27], we assume that the equilibrium manifold $\mathcal{E} = \{M \in \mathbb{G} \mid S(M) = 0\}$ is not empty and denote the Jacobian matrix of $S(M)$ as $S_M(M)$. The stability condition reads as

(i) There exist an invertible $n \times n$ matrix $P(M)$ and an invertible $r \times r$ $(0 < r \leq n)$ matrix $\hat{T}(M)$ such that

$$P(M) S_M(M) = \begin{bmatrix} 0 & 0 \\ 0 & \hat{T}(M) \end{bmatrix} P(M), \quad \forall M \in \mathcal{E};$$

(ii) There exists a positive definite symmetric matrix $A_0(M)$ such that

$$A_0(M) A_d(M) = A_d^T(M) A_0(M) \quad \text{for any } M \in \mathbb{G} \text{ and } d = 1, \ldots, D;$$

(iii) The spatial derivative parts and the source are coupled as

$$A_0(M) S_M(M) + S_M^T(M) A_0(M) \leq -P^T(M) \begin{bmatrix} 0 & 0 \\ 0 & I_r \end{bmatrix} P(M), \quad \forall M \in \mathcal{E}.$$

Here I_r is the unit matrix of order r.

As shown in [28], this set of conditions has been tacitly respected by many well-developed physical theories. Condition (i) is classical for initial value problems of the system of ordinary differential equations (ODE, spatially homogeneous systems), while (ii) means the symmetrizable hyperbolicity of the PDE system. Condition (iii) characterizes a kind of coupling between the ODE and PDE parts. Recently, this structural stability condition is shown in [7] to be proper for certain moment closure systems. On the other hand, this set of conditions implies the existence and stability of the zero relaxation limit of the corresponding initial value problems [27]. Thanks to these, we believe that the structural stability condition is essential for a reasonable moment closure system.

2.3. Main results. For the moment systems derived above, we will establish the following facts as the main result of this paper,

THEOREM 2.1 (Non-hyperbolicity of QMOM). *The QMOM-derived moment system (2.13) is not strongly hyperbolic.*

THEOREM 2.2 (Stability of EQMOM). *The EQMOM-derived moment system (2.19) satisfies the structural stability condition for $M \in \mathcal{M}(\Omega_W)$.*

A proof of Theorem 2.1 will be presented in the next section. In Section 4, Theorem 2.2 is divided as Theorem 4.4 (hyperbolicity of EQMOM), Theorem 4.11 (equilibrium state), Theorem 4.12 (BGK model) and Theorem 4.14 (Shakhov model), which will be proved in Subsections 4.2 to 4.5, respectively.

3. Non-hyperbolicity of QMOM. This section is devoted to a proof of Theorem 2.1 for the QMOM-derived moment system (2.13) with $N \geq 2$. We should mention that this theorem has been proved in [6] but only for $N = 2$. For our purpose, we need to consider the $2N \times 2N$ coefficient matrix $A = A(M)$ in (2.14).

Proof of Theorem 2.1. Let λ be an eigenvalue of A and $\mathbf{v} = (v_1, \ldots, v_{2N})^T$ the corresponding right eigenvector. A direct calculation indicates that

$$v_k = \lambda v_{k-1} = \lambda^{k-1} v_1 \quad \text{for } k = 2, \ldots, 2N, \tag{3.1a}$$

$$\sum_{k=1}^{2N} a_{k-1} v_k = \lambda v_{2N} = \lambda^{2N} v_1. \tag{3.1b}$$

Then we have $\mathbf{v} = v_1(1, \lambda, \ldots, \lambda^{2N-1})^T$ and thereby $v_1 \neq 0$. This shows that the geometric multiplicity of each eigenvalue is 1.

On the other hand, we see from (3.1b) that the characteristic polynomial of A is

$$c(\lambda) = \lambda^{2N} - a_{2N-1}\lambda^{2N-1} - \cdots - a_1\lambda - a_0. \tag{3.2}$$

Note that $(a_0, a_1, \ldots, a_{2N-1}) = \left(\frac{\partial \bar{M}_{2N}}{\partial M_0}, \frac{\partial \bar{M}_{2N}}{\partial M_1}, \ldots, \frac{\partial \bar{M}_{2N}}{\partial M_{2N-1}}\right) = \frac{\partial \bar{M}_{2N}}{\partial M}$ with $\bar{M}_{2N}$ defined in (2.12) and $M = (M_0, \ldots, M_{2N-1})^T$. Writing $W = (w_1, u_1, \ldots, w_N, u_N)^T \in \mathbb{R}^{2N}$, we have

$$(a_0, a_1, ..., a_{2N-1})\frac{\partial M}{\partial W} = \left(\frac{\partial \bar{M}_{2N}}{\partial M}\right)\left(\frac{\partial M}{\partial W}\right) = \frac{\partial \bar{M}_{2N}}{\partial W}. \tag{3.3}$$

In addition, it follows from (2.11) that the $2N \times 2N$ Jacobian matrix $\partial M/\partial W$ is

$$\frac{\partial M}{\partial W} = \begin{bmatrix} 1 & 0 & \cdots & 1 & 0 \\ u_1 & w_1 & \cdots & u_N & w_N \\ \vdots & \vdots & & \vdots & \vdots \\ u_1^j & j w_1 u_1^{j-1} & \cdots & u_N^j & j w_N u_N^{j-1} \\ \vdots & \vdots & & \vdots & \vdots \\ u_1^{2N-1} & (2N-1) w_1 u_1^{2N-2} & \cdots & u_N^{2N-1} & (2N-1) w_N u_N^{2N-2} \end{bmatrix}$$

and from (2.12) that

$$\frac{\partial \bar{M}_{2N}}{\partial W} = \left(u_1^{2N}, 2N w_1 u_1^{2N-1}, \ldots, u_N^{2N}, 2N w_N u_N^{2N-1}\right).$$

Substituting the last two relations into (3.3), we obtain

$$\begin{aligned} u_k^{2N} - a_{2N-1} u_k^{2N-1} - \cdots - a_1 u_k - a_0 = 0, \\ 2N u_k^{2N-1} - (2N-1) a_{2N-1} u_k^{2N-2} - \cdots - a_1 = 0 \end{aligned}$$

for $k = 1, \ldots, N$. These mean that $c(u_k) = 0$ and $\left.\frac{dc(\lambda)}{d\lambda}\right|_{\lambda = u_k} = 0$ for $k = 1, \ldots, N$. Since $c = c(\lambda)$ is a monic polynomial of order $2N$, there must be

$$c(\lambda) = (\lambda - u_1)^2 \cdots (\lambda - u_N)^2. \tag{3.4}$$

As a result, the eigenvalues of A are $u_1, u_2, \ldots, u_N$ and each of them has the algebraic multiplicity 2 and the geometric multiplicity 1. In view of its Jordan canonical form, the coefficient matrix A is similar to

$$\begin{bmatrix} u_1 & 1 & & & \\ 0 & u_1 & & & \\ & & \ddots & & \\ & & & u_N & 1 \\ & & & 0 & u_N \end{bmatrix}. \tag{3.5}$$

Hence the moment closure system (2.13) is not strongly hyperbolic. ∎

4. Stability of EQMOM. We prove Theorem 2.2 in this section. In particular, Subsection 4.2 is devoted to Condition (ii), while Conditions (i) and (iii) are verified in Subsections 4.4 and 4.5 for both the BGK and Shakhov collision models.

4.1. Preliminaries. Recall that in Subsection 2.1, we use the notation

$$\Delta_j = \Delta_j(u, \sigma^2) = \int_{\mathbb{R}} \xi^j \delta_{\sigma^2}(\xi; u) d\xi$$

for the jth moment of the Gaussian distribution $\delta_{\sigma^2} = \delta_{\sigma^2}(\xi; u) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\xi-u)^2}{2\sigma^2}\right)$. A direct calculation shows $\Delta_0(u, \sigma^2) = 1$ and $\Delta_1(u, \sigma^2) = u$. Moreover, we can show with Lemma 4.1(a) below that $\Delta_j(u, \sigma^2)$ is a bivariate polynomial of u and σ^2.

LEMMA 4.1.

(a) $\Delta_j(u, \sigma^2) = u\Delta_{j-1}(u, \sigma^2) + (j-1)\sigma^2\Delta_{j-2}(u, \sigma^2)$ *for* $j \geq 2$,

(b) $\Delta_j(u, \sigma^2) = \sum_{k=0}^{\infty} \left(\frac{\sigma^2}{2}\right)^k \frac{(u^j)^{(2k)}}{k!}$ *(this is a finite sum),*

(c) $\frac{\partial \Delta_j(u, \sigma^2)}{\partial u} = j\Delta_{j-1}(u, \sigma^2)$ *for* $j \geq 1$,

(d) $\frac{\partial \Delta_j(u, \sigma^2)}{\partial \sigma^2} = \frac{j(j-1)}{2}\Delta_{j-2}(u, \sigma^2)$ *for* $j \geq 2$.

Proof. (a): Note that $d\delta_{\sigma^2}/d\xi = -(\xi - u)\delta_{\sigma^2}/\sigma^2$. Then for $j \geq 2$ we have

$$\begin{aligned} \Delta_j &= \int_{\mathbb{R}} (\xi - u + u)\xi^{j-1}\delta_{\sigma^2} d\xi = u\Delta_{j-1} + \int_{\mathbb{R}} (\xi - u)\xi^{j-1}\delta_{\sigma^2} d\xi \\ &= u\Delta_{j-1} - \sigma^2 \int_{\mathbb{R}} \xi^{j-1} \frac{d\delta_{\sigma^2}}{d\xi} d\xi = u\Delta_{j-1} + (j-1)\sigma^2\Delta_{j-2}. \end{aligned}$$

This, together with $\Delta_0(u, \sigma^2) = 1$ and $\Delta_1(u, \sigma^2) = u$, indicates that $\Delta_j = \Delta_j(u, \sigma^2)$ is a polynomial of both u and σ^2.

(b): This can be proven by induction on j. It obviously holds for $\Delta_0 = 1$ and $\Delta_1 = u$. Suppose it is true for $j-1$ and j. Then for $j+1$ it follows from (a) that

$$\begin{aligned} \Delta_{j+1} &= u\Delta_j + j\sigma^2\Delta_{j-1} = u\sum_{k=0}^{\infty} \left(\frac{\sigma^2}{2}\right)^k \frac{(u^j)^{(2k)}}{k!} + \sigma^2 \sum_{k=0}^{\infty} \left(\frac{\sigma^2}{2}\right)^k \frac{(u^j)^{(2k+1)}}{k!} \\ &= u^{j+1} + \sum_{k=1}^{\infty} \left(\frac{\sigma^2}{2}\right)^k \frac{u(u^j)^{(2k)} + 2k(u^j)^{(2k-1)}}{k!} = \sum_{k=0}^{\infty} \left(\frac{\sigma^2}{2}\right)^k \frac{(u^{j+1})^{(2k)}}{k!}. \end{aligned}$$

Hence the proof is complete.

(c & d): These two follow immediately from (b):

$$\frac{\partial \Delta_j}{\partial u} = \sum_{k=0}^{\infty} \left(\frac{\sigma^2}{2}\right)^k \frac{(u^j)^{(2k+1)}}{k!} = \sum_{k=0}^{\infty} \left(\frac{\sigma^2}{2}\right)^k \frac{j(u^{j-1})^{(2k)}}{k!} = j\Delta_{j-1};$$
$$\frac{\partial \Delta_j}{\partial \sigma^2} = \frac{1}{2}\sum_{k=1}^{\infty} k \left(\frac{\sigma^2}{2}\right)^{k-1} \frac{(u^j)^{(2k)}}{k!} = \frac{1}{2}\sum_{k=0}^{\infty} \left(\frac{\sigma^2}{2}\right)^k \frac{(u^j)^{(2k+2)}}{k!}$$
$$= \frac{1}{2}\sum_{k=0}^{\infty} \left(\frac{\sigma^2}{2}\right)^k \frac{j(j-1)(u^{j-2})^{(2k)}}{k!} = \frac{j(j-1)}{2}\Delta_{j-2}.$$

∎

Remark 4.2. Lemma 4.1(b) is obviously equivalent to

$$\Delta_j(u,\sigma^2) = \sum_{k=0}^{[j/2]} \frac{j!}{k!(j-2k)!} \left(\frac{\sigma^2}{2}\right)^k u^{j-2k}, \tag{4.1}$$

which was established in [19]. But the former is more convenient for our later use.

Inspired by Lemma 4.1(b), we introduce a family of linear operators $\mathcal{D}_\vartheta$, parameterized with $\vartheta \in \mathbb{R}$, acting on the polynomial algebra $\mathbb{R}[u]$. For $f \in \mathbb{R}[u]$, $\mathcal{D}_\vartheta f$ is defined as

$$\mathcal{D}_\vartheta f = \sum_{k=0}^{\infty} \left(\frac{\vartheta}{2}\right)^k \frac{f^{(2k)}}{k!}, \tag{4.2}$$

which is a finite sum. Obviously, $\mathcal{D}_0$ is an identical operator, $\mathcal{D}_{\sigma^2} f$ is a polynomial of u and σ^2, and $\mathcal{D}_{\sigma^2} u^j = \Delta_j(u,\sigma^2)$. Further useful properties of $\mathcal{D}_\vartheta$ are

LEMMA 4.3.

(a) *(composition)* $\mathcal{D}_\alpha \circ \mathcal{D}_\vartheta = \mathcal{D}_{\alpha+\vartheta}$,

(b) $\mathcal{D}_\vartheta$ *is invertible and* $\mathcal{D}_\vartheta^{-1} = \mathcal{D}_{-\vartheta}$,

(c) $\dfrac{\partial}{\partial u}\mathcal{D}_\vartheta f(u) = \mathcal{D}_\vartheta f'(u), \quad \dfrac{\partial}{\partial \vartheta}\mathcal{D}_\vartheta f(u) = \dfrac{1}{2}\mathcal{D}_\vartheta f''(u)$,

(d) $\mathcal{D}_\vartheta(uf) = u\mathcal{D}_\vartheta f + \vartheta \mathcal{D}_\vartheta f'$,

(e) *If* $\mathcal{D}_\vartheta f(u_0) = 0$, *then* $\mathcal{D}_\vartheta(uf)|_{u=u_0} = \vartheta \mathcal{D}_\vartheta f'(u_0)$.

Proof. (a): For the composition, we deduce from the definition that

$$(\mathcal{D}_\alpha \circ \mathcal{D}_\vartheta) f = \sum_{k=0}^{\infty} \left(\frac{\alpha}{2}\right)^k \frac{1}{k!} \sum_{l=0}^{\infty} \left(\frac{\vartheta}{2}\right)^l \frac{f^{(2k+2l)}}{l!}$$
$$= \sum_{p=0}^{\infty} \frac{f^{(2p)}}{p!} \left[\sum_{l=0}^{p} \frac{p!}{l!(p-l)!} \left(\frac{\alpha}{2}\right)^{p-l} \left(\frac{\vartheta}{2}\right)^l\right] = \sum_{p=0}^{\infty} \frac{f^{(2p)}}{p!} \left(\frac{\alpha+\vartheta}{2}\right)^p = \mathcal{D}_{\alpha+\vartheta} f.$$

(b) follows immediately from (a) and $\mathcal{D}_0 = id$.

For (c), the first one is obvious, while the second can be shown as Lemma 4.1(d).

(d): By using $(uf)^{(2k)} = uf^{(2k)} + 2kf^{(2k-1)}$, this can be proved as Lemma 4.1(b). Then (e) follows immediately from (d). ∎

4.2. Hyperbolicity of EQMOM. In this section we prove that the EQMOM-derived moment system (2.19) for the 1-D Boltzmann equation is strictly hyperbolic, which will be shown to be sufficient for the structural stability condition (ii). The conclusion can be stated as

THEOREM 4.4. *For $M \in \mathcal{M}(\Omega_W)$, the $(2N+1)\times(2N+1)$ coefficient matrix $A = A(M)$ in (2.20) has $(2N+1)$ distinct real eigenvalues. Namely, the EQMOM-derived moment system (2.19) is strictly hyperbolic.*

We should mention that this theorem was already established in [5] for $N = 2$ (the two-node system) but the proof does not seem to work for $N > 2$.

Our proof of this theorem needs some preparations. First of all, the characteristic polynomial of A in (2.20) reads as

$$c(u;W) = u^{2N+1} - a_{2N}u^{2N} - \cdots - a_1 u - a_0. \tag{4.3}$$

Here the coefficient $a_j = a_j(W) = \frac{\partial \bar{M}_{2N+1}}{\partial \bar{M}_j}$ $(j = 0, 1, \ldots, 2N)$, with $\bar{M}_{2N+1}$ defined in (2.18), is a function of W. To show that $c(u;W)$, as a polynomial of u, has $(2N+1)$ distinct real roots for $M \in \mathcal{M}(\Omega_W)$, we introduce an auxiliary function

$$g(u;W) = \mathcal{D}_{\sigma^2} c(u;W) = \sum_{k=0}^{\infty} \left(\frac{\sigma^2}{2}\right)^k \frac{\partial_u^{2k} c(u;W)}{k!}. \tag{4.4}$$

By Lemma 4.3(b), we have

$$c(u;W) = \mathcal{D}_{-\sigma^2} g(u;W) = \sum_{k=0}^{\infty} \left(-\frac{\sigma^2}{2}\right)^k \frac{\partial_u^{2k} g(u;W)}{k!}. \tag{4.5}$$

Set $a_{2N+1} = -1$. Then $c(u;W)$ can be rewritten as $-\sum_{j=0}^{2N+1} a_j u^j$ and from the linearity of $\mathcal{D}_{\sigma^2}$ it follows that

$$g(u;W) = -\sum_{j=0}^{2N+1} a_j \mathcal{D}_{\sigma^2} u^j = -\sum_{j=0}^{2N+1} a_j \Delta_j(u,\sigma^2). \tag{4.6}$$

Moreover, from (4.3) and (4.4) we see that $g(u;W)$ is a u-polynomial of degree $(2N+1)$:

$$g(u;W) = -\sum_{j=0}^{2N+1} g_j u^j \tag{4.7}$$

with $g_{2N+1} = a_{2N+1} = -1$. Further relations between the coefficients of $g(u;W)$ and $c(u;W)$ are

$$a_j = \sum_{k=0}^{N-[j/2]} g_{j+2k} \frac{(j+2k)!}{j!k!} \left(-\frac{\sigma^2}{2}\right)^k, \quad j = 0, 1, \ldots, 2N+1. \tag{4.8}$$

This can be shown as

$$\begin{aligned} c(u;W) &= \sum_{k=0}^{N} \frac{\partial_u^{2k} g(u;W)}{k!} \left(-\frac{\sigma^2}{2}\right)^k = -\sum_{k=0}^{N} \sum_{j=0}^{2N+1-2k} \frac{(j+2k)!}{j!k!} \left(-\frac{\sigma^2}{2}\right)^k g_{j+2k} u^j \\ &= -\sum_{j=0}^{2N+1} \left[\sum_{k=0}^{N-[j/2]} \frac{(j+2k)!}{j!k!} \left(-\frac{\sigma^2}{2}\right)^k g_{j+2k} \right] u^j. \end{aligned}$$

Furthermore, $g = g(u) = g(u; W)$ has the following elegant expression.

LEMMA 4.5.

$$g(u; W) = (u - u_1)^2 \cdots (u - u_N)^2 (u - \tilde{U}),$$

where $u_1, \ldots, u_N$ are the nodes solved from (2.16), and

$$\tilde{U} = \tilde{U}(W) = \frac{\sum_{i=1}^{N} w_i u_i \prod_{1 \le j \le N, j \ne i} (u_j - u_i)^2}{\sum_{i=1}^{N} w_i \prod_{1 \le j \le N, j \ne i} (u_j - u_i)^2}$$

for $W \in \Omega_W^{open}$.

Remark 4.6. This lemma shows that for $W \in \Omega_W^{open}$, $\tilde{U}$ is a convex combination of the u_i's. Moreover, for $W = (w_1, U, w_2, U, \ldots, w_N, U, \sigma^2) \in \Omega_W^{eq}$ and any sequence $\{W_k\} \subset \Omega_W^{open}$ approaching W, $\tilde{U}(W_k)$ converges to U. Because of this, for $W \in \Omega_W^{eq}$ we define $\tilde{U}(W) = U$ $(= M_1/M_0)$ and thereby $g(u; W) = (u - U)^{2N+1}$.

Proof. By Lemma 4.1(c&d), the Jacobian matrix of the map $M = \mathcal{M}(W)$ defined in (2.16) is

(4.9)
$$\begin{bmatrix} \Delta_0(u_1) & 0 & \cdots & \Delta_0(u_N) & 0 & 0 \\ \Delta_1(u_1) & w_1\Delta_0(u_1) & \cdots & \Delta_1(u_N) & w_N\Delta_1(u_N) & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ \Delta_j(u_1) & jw_1\Delta_{j-1}(u_1) & \cdots & \Delta_j(u_N) & jw_N\Delta_{j-1}(u_N) & \binom{j}{2} M_{j-2} \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ \Delta_{2N}(u_1) & 2Nw_1\Delta_{2N-1}(u_1) & \cdots & \Delta_{2N}(u_N) & 2Nw_N\Delta_{2N-1}(u_N) & \binom{2N}{2} M_{2N-2} \end{bmatrix}.$$

Note that the dependence of Δ_j on σ^2 has been omitted here for clarity. Moreover, from (2.18), $\partial \bar{M}_{2N+1}/\partial W$ reads as

$$\left(\Delta_{2N+1}(u_1), (2N+1)w_1\Delta_{2N}(u_1), \ldots, \Delta_{2N+1}(u_N), (2N+1)w_N\Delta_{2N}(u_N), \binom{2N+1}{2} M_{2N-1} \right).$$

Then from the simple relation

$$(a_0, a_1, ..., a_{2N}) \frac{\partial \mathcal{M}}{\partial W} = \left(\frac{\partial \bar{M}_{2N+1}}{\partial M} \right) \left(\frac{\partial \mathcal{M}}{\partial W} \right) = \frac{\partial \bar{M}_{2N+1}}{\partial W}$$

we obtain

$$a_0\Delta_0(u_k) + a_1\Delta_1(u_k) + \cdots + a_{2N}\Delta_{2N}(u_k) = \Delta_{2N+1}(u_k), \tag{4.10a}$$

$$w_k a_1\Delta_0(u_k) + \cdots + 2Nw_k a_{2N}\Delta_{2N-1}(u_k) = (2N+1)w_k\Delta_{2N}(u_k), \tag{4.10b}$$

$$\binom{2}{2} a_2 M_0 + \cdots + \binom{2N}{2} a_{2N} M_{2N-2} = \binom{2N+1}{2} M_{2N-1}. \tag{4.10c}$$

(4.10a) and (4.10b), together with (4.6) and Lemma 4.1(c), imply that $g(u_k) = 0$ and $\frac{dg(u)}{du}\Big|_{u=u_k} = 0$ for $k = 1, \ldots, N$. Thus we see the expected expression of $g(u; W)$ from (4.7) with $\tilde{U}$ to be determined.

Next, we use (4.10c) to determine $\tilde{U}$. Recall that $a_{2N+1} = -1$. We use (2.16) to rewrite (4.10c) as

$$0 = \sum_{j=2}^{2N+1} \binom{j}{2} a_j M_{j-2} = \sum_{i=1}^{N} w_i \sum_{j=2}^{2N+1} \binom{j}{2} a_j \Delta_{j-2}(u_i, \sigma^2). \tag{4.11}$$

On the other hand, we deduce from (4.6) and (4.7) that

$$-\frac{1}{2} g''(u) = \sum_{j=2}^{2N+1} \binom{j}{2} g_j u^{j-2} = \sum_{j=2}^{2N+1} \binom{j}{2} a_j \Delta_{j-2}(u, \sigma^2).$$

Then we see from (4.11) that

$$\sum_{i=1}^{N} w_i \sum_{j=2}^{2N+1} \binom{j}{2} g_j u_i^{j-2} = 0. \tag{4.12}$$

Now we define $\tilde{g}(u) = (u-u_1)^2 \cdots (u-u_N)^2 = -\sum_{j=0}^{2N} \tilde{g}_j u^j$ with $\tilde{g}_{2N} = -1$. Then $g(u) = (u - \tilde{U})\tilde{g}(u)$ and the coefficients are related with

$$g_j = \tilde{g}_{j-1} - \tilde{U}\tilde{g}_j$$

for $0 \le j \le 2N+1$ ($\tilde{g}_{-1} = \tilde{g}_{2N+1} = 0$). Substituting this relation into (4.12), we obtain

$$\left[\sum_{j=2}^{2N} \binom{j}{2} \tilde{g}_j M^*_{j-2} \right] \tilde{U} = \sum_{j=2}^{2N+1} \binom{j}{2} \tilde{g}_{j-1} M^*_{j-2},$$

where $M^*_j = \sum_{i=1}^{N} w_i u_i^j$. It remains to show

$$\sum_{j=2}^{2N} \binom{j}{2} \tilde{g}_j M^*_{j-2} = -\sum_{i=1}^{N} w_i \prod_{1 \le k \le N, k \ne i} (u_k - u_i)^2,$$
$$\sum_{j=2}^{2N+1} \binom{j}{2} \tilde{g}_{j-1} M^*_{j-2} = -\sum_{i=1}^{N} w_i u_i \prod_{1 \le k \le N, k \ne i} (u_k - u_i)^2.$$

These two follow from the obvious relations

$$-\sum_{j=2}^{2N} \binom{j}{2} \tilde{g}_j u_i^{j-2} = \frac{1}{2}\tilde{g}''(u_i) = \prod_{1 \le k \le N, k \ne i} (u_k - u_i)^2,$$
$$-\sum_{j=2}^{2N+1} \binom{j}{2} \tilde{g}_{j-1} u_i^{j-2} = \frac{1}{2}\left(u\tilde{g}\right)''(u_i) = u_i \prod_{1 \le k \le N, k \ne i} (u_k - u_i)^2$$

for any $1 \le i \le N$. This completes the proof. ∎

Remark 4.7. Lemma 4.5 indicates that the coefficients g_j of $g(u; W)$ in (4.7) are independent of σ^2. From (4.5) we see that $c(u; W)$ is a bivariate polynomial of u and σ^2, the coefficients a_j of $c(u; W)$ are polynomials of σ^2, and $c(u; W) = g(u)$ for $\sigma^2 = 0$. Furthermore, the jth derivative $c^{(j)}(u; W)$ of $c(u; W)$ with respect to u can be viewed as a perturbation of $g^{(j)}(u)$ with the single parameter $\sigma^2 \ge 0$ for $0 \le j \le 2N+1$.

By Lemma 4.5, $g(u)$ has $(2N+1)$ (= the degree of g) real roots (including multiplicity). This fact can be further generalized as follows.

LEMMA 4.8. *For any $0 \le j \le 2N$, $g^{(j)}(u)$ has $(2N+1-j)$ real roots (including multiplicity). Hence any local minimum (maximum) value of $g^{(j)}(u)$ is non-positive (non-negative).*

Proof. We prove by induction on j. As discussed above, the conclusion holds for $j=0$. Namely, g has $(2N+1)$ roots. Suppose it holds for $0,\dots,j$. Then we have $g^{(j)}(u) = C(u-\tilde{u}_1)^{k_1}\cdots(u-\tilde{u}_m)^{k_m}$, where $m \ge 1$, $\tilde{u}_1 < \tilde{u}_2 < \cdots < \tilde{u}_m$, $k_i \ge 1$ and $k_1+\cdots+k_m = 2N+1-j$. Thus $(u-\tilde{u}_i)^{k_i-1}$ is a factor of $g^{(j+1)}(u)$ for any $1 \le i \le m$. Besides, Rolle's theorem implies the existence of at least one root of $g^{(j+1)}(u)$ in each open interval $(\tilde{u}_i, \tilde{u}_{i+1})$ for $1 \le i \le m-1$. Therefore, the number of roots of $g^{(j+1)}$ is no less than

$$(k_1-1)+\cdots+(k_m-1)+(m-1) = (k_1+\cdots+k_m)-1 = 2N-j.$$

Since $g^{(j+1)}$ is of degree $(2N-j)$, it must have $(2N-j)$ roots (including multiplicity). This also indicates that $g^{(j)}$ has only one extreme point in each open interval above. Hence any local minimum (maximum) value of $g^{(j)}(u)$ is non-positive (non-negative). ∎

With the preparations above, we are in a position to prove Theorem 4.4.

Proof of Theorem 4.4. We will prove the following stronger statement: for $0 \le j \le 2N+1$, $c^{(2N+1-j)}(u;W)$ has j distinct roots for any $W = (w_1,u_1,\dots,w_N,u_N,\sigma^2) \in \Omega_W$ with $\sigma^2 > 0$. This will be done with induction on j. For $j=0$, 1, the statement is obvious because $c^{(2N+1)}(u;W) = (2N+1)!$ and $c^{(2N)}(u;W)$ is of degree 1.

Suppose the conclusion holds for $j \le k (\le 2N+1)$. From Remark 4.7 we know that $c^{(2N+1-k)}(u;W)$ is a bivariate polynomial of u and σ^2 on $\mathbb{R}\times[0,\infty)$. Denote $u^*(\sigma^2) \in \mathbb{R}$ to be one root of $c^{(2N+1-k)}(u;W)$. Thus $u^*(\sigma^2)$ is an extreme point of $c^{(2N-k)}(u;W)$ and $u^*(0)$ is a root of $g^{(2N+1-k)}(u)$. Moreover, $u^*(\sigma^2)$ is continuous on $\sigma^2 \in [0,\infty)$ and differentiable on $(0,\infty)$ because the roots are distinct [15].

Next we consider the extreme values of $c^{(2N-k)}(u;W)$ at $u = u^*(\sigma^2)$. Since $c^{(2N-k)}(u;W)$ is a polynomial of u and σ^2, the composite $h_k(\sigma^2) := c^{(2N-k)}(u^*(\sigma^2);W)$ is continuous on $[0,\infty)$ and differentiable on $(0,\infty)$. And $h_k(0)$ is the extreme value of $g^{(2N-k)}(u)$. According to Lemma 4.8, $h_k(0) \ge 0$ if it is a local maximum and $h_k(0) \le 0$ if it is a local minimum. For $\sigma^2 > 0$, because $c^{(2N+1-k)}(u^*(\sigma^2);W) = 0$, the derivative of $h_k(\sigma^2)$ reads as

$$\begin{aligned}
\frac{\partial h_k(\sigma^2)}{\partial \sigma^2} &= \frac{\partial}{\partial \sigma^2} c^{(2N-k)}(u^*(\sigma^2);W) = \frac{\partial}{\partial \sigma^2}\sum_{l=0}^{\infty}\left(-\frac{\sigma^2}{2}\right)^l \frac{g^{(2l+2N-k)}(u^*(\sigma^2))}{l!} \\
&= -\frac{1}{2}\sum_{l=1}^{\infty}\left(-\frac{\sigma^2}{2}\right)^{l-1}\frac{g^{(2l+2N-k)}(u^*(\sigma^2))}{(l-1)!} + c^{(2N+1-k)}(u^*(\sigma^2);W)\frac{\partial u^*(\sigma^2)}{\partial \sigma^2} \\
&= -\frac{1}{2}c^{(2N+2-k)}(u^*(\sigma^2);W).
\end{aligned}$$

Thus, if $c^{(2N-k+2)}(u^*(\sigma^2);W) < 0$ (that is, $u^*(\sigma^2)$ is a local maximum point of $c^{(2N-k)}(u;W)$), then the local maximum value $h_k(\sigma^2)$ strictly increases on $\sigma^2 \in (0,\infty)$. Since $h_k(0) \ge 0$ and $h_k(\sigma^2)$ is continuous at $\sigma^2 = 0$, we conclude that $h_k(\sigma^2) > 0$ for all $\sigma^2 > 0$. Similarly, if $c^{(2N-k+2)}(u^*(\sigma^2);W) > 0$, we have $h_k(\sigma^2) < 0$ for all $\sigma^2 > 0$.

In summary, the above arguments show that each local maximum value of the $(k+1)$th oder polynomial $c^{(2N-k)}(u;W)$ is positive and each local minimum value

is negative. On the other hand, by the induction assumption $c^{(2N+1-k)}(u;W)$ has k distinct real roots, which are naturally extreme points of $c^{(2N-k)}(u;W)$ for $\sigma^2>0$. Therefore, $c^{(2N-k)}(u;W)$ has $(k-1)$ distinct real roots among the extreme points. Moreover, the induction assumption implies that $c^{(2N-k)}(u;W)$ has one root larger and another one less than all the extreme points. Thus, for each $\sigma^2>0$, $c^{(2N-k)}(u;W)$ has $(k+1)$ distinct real roots. By the induction principle this completes the proof. □

Remark 4.9. By Theorem 4.4, the coefficient matrix $A=A(M)$ of the 1-D moment system (2.19) has $n=(2N+1)$ distinct real eigenvalues λ_i $(1\le i\le n)$ for $\sigma^2>0$. Denote by r_i the corresponding left eigenvectors. Set $L=(r_1^T,\dots,r_n^T)^T$. It is clear that $A_0(M)=L^T\Lambda L$ with Λ an arbitrary positive diagonal matrix is a symmetrizer in the structural stability condition (ii). As a matter of fact, it is straightforward to show that such a symmetrizer can only be of the form $L^T\Lambda L$.

4.3. Equilibrium state. As stated in Subsection 2.2, (i) and (iii) of the structural stability condition should be examined on the equilibrium manifold $\mathcal{E}$ where $S(M(W))=0$. In this section we determine the equilibrium manifold.

For the BGK model, $S(M(W))=0$ is equivalent to

$$\sum_{i=1}^{N} w_i\Delta_j(u_i,\sigma^2)=M_j=\rho\Delta_j(U,\theta)\quad\text{for } j=0,\dots,2N \tag{4.13}$$

(see Subsection 2.1.2). Thus, the equilibrium state $W=(w_1,u_1,\dots,w_N,u_N,\sigma^2)^T$ is determined by the three macroscopic parameters ρ, U and θ. And we need to find W from (4.13) for $1\le i\le N$.

For this purpose, we recall from Subsection 4.1 that $\Delta_0(u,\sigma^2)=1$, $\Delta_1(u,\sigma^2)=u$ and $\Delta_2(u,\sigma^2)=u^2+\sigma^2$. Thus, for $j=0,1,2$, (4.13) is just

$$\sum_{i=1}^{N} w_i=\rho,\quad \sum_{i=1}^{N} w_iu_i=\rho U,\quad \sum_{i=1}^{N} w_iu_i^2=\rho U^2+\rho(\theta-\sigma^2).$$

Then we deduce from the inequality $\left(\sum_{i=1}^N w_i\right)\left(\sum_{i=1}^N w_iu_i^2\right)\ge\left(\sum_{i=1}^N w_iu_i\right)^2$ that

$$\sigma^2\le\theta\quad\text{and } \sigma^2=\theta \text{ if and only if all the } u_i\text{'s are equal.} \tag{4.14}$$

For further discussions, we need the following fact.

Proposition 4.10.

$$M_j^*:=\sum_{i=1}^{N} w_iu_i^j=\sum_{k=0}^{[j/2]}\frac{j!}{k!(j-2k)!}\left(-\frac{\sigma^2}{2}\right)^k M_{j-2k}.$$

Proof. Recall that $\Delta_j(u,\sigma^2)=\mathcal{D}_{\sigma^2}u^j$. From Lemma 4.3(b) and Lemma 4.1(c) we deduce that

$$u^j=\sum_{k=0}^{\infty}\frac{\partial_u^{2k}\Delta_j(u,\sigma^2)}{k!}\left(-\frac{\sigma^2}{2}\right)^k=\sum_{k=0}^{[j/2]}\frac{j!}{k!(j-2k)!}\left(-\frac{\sigma^2}{2}\right)^k\Delta_{j-2k}(u,\sigma^2).$$

Then taking the weighted summation $\sum_{i=1}^N w_i$ and using (2.16) give the proposition. □

Next we define $\zeta^2 = \theta - \sigma^2 \geq 0$ and show that (4.13) is equivalent to

$$\sum_{i=1}^{N} w_i u_i^j = \rho \Delta_j(U, \zeta^2) \quad \text{for } j = 0, \ldots, 2N. \tag{4.15}$$

Indeed, if (4.13) holds (i.e. $M(W) \in \mathcal{E}$), the last proposition implies that

$$\begin{aligned}\sum_i w_i u_i^j = M_j^* &= \sum_{k=0}^{[j/2]} \left(-\frac{\sigma^2}{2}\right)^k \frac{j!}{k!(j-2k)!} \left[\rho \Delta_{j-2k}(U, \theta)\right] \\ &= \rho \sum_{k=0}^{\infty} \left(-\frac{\sigma^2}{2}\right)^k \left. \frac{\partial_u^{2k} \Delta_j(u, \theta)}{k!} \right|_{u=U} = \rho \mathcal{D}_{-\sigma^2} \mathcal{D}_\theta U^j = \rho \mathcal{D}_{\theta - \sigma^2} U^j.\end{aligned}$$

Here the expression $\mathcal{D}_\vartheta f(U)$ denotes $\mathcal{D}_\vartheta f(u)|_{u=U}$ for arbitrary polynomial f and the last step is due to Lemma 4.3(a). This is just (4.15). The deduction of (4.13) from (4.15) is similar.

Now we are in a position to state the central result of this section.

THEOREM 4.11. *The equilibrium state belongs to Ω_W^{eq}, that is,*

$$u_1 = \cdots = u_N = U, \ \sigma^2 = \theta, \ \text{and} \ \sum_{i=1}^{N} w_i = \rho.$$

Hence, at equilibrium $\bar{M}_{2N+1} = \rho \Delta_{2N+1}(U, \theta)$.

Proof. Thanks to (4.14), it suffices to show that $\zeta^2 := \theta - \sigma^2 = 0$. Otherwise, the u_i's must take N' different values ($1 < N' \leq N$). Then, by redefining w_i, the summation $\sum_{i=1}^{N} w_i u_i^j$ in the left-hand side of (4.15) is reduced to $\sum_{k=1}^{N'} w_k u_k^j$ where the u_k's are distinct ($1 \leq k \leq N'$). Thus, we may as well assume that all the u_i's are distinct and $\zeta^2 > 0$. Then we will derive a contradiction in three steps, where the abbreviation

$$\mathcal{D}_\vartheta f(U) \equiv \mathcal{D}_\vartheta f(u)|_{u=U}$$

will be frequently used.

Step I. Because $\Delta_j(u, \zeta^2) = \mathcal{D}_{\zeta^2} u^j$, the first N equations ($j = 0, \ldots, N-1$) in (4.15) can be rewritten as a system of linear algebraic equations:

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ u_1 & u_2 & \cdots & u_N \\ \vdots & \vdots & & \vdots \\ u_1^{N-1} & u_2^{N-1} & \cdots & u_N^{N-1} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix} = \rho \begin{bmatrix} \mathcal{D}_{\zeta^2}(1) \\ \mathcal{D}_{\zeta^2}(U^1) \\ \vdots \\ \mathcal{D}_{\zeta^2}(U^{N-1}) \end{bmatrix}.$$

Since all the u_i's are distinct, this gives a unique $(w_1, \ldots, w_N)$ in terms of $(u_1, \ldots, u_N)$, ρ, U and ζ^2. We claim that for $1 \leq i \leq N$,

$$w_i \prod_{1 \leq k \leq N, k \neq i} (u_i - u_k) = \rho \mathcal{D}_{\zeta^2} \left(\prod_{1 \leq k \leq N, k \neq i} (U - u_k) \right). \tag{4.16}$$

To see this, we use the uniqueness and only need to show that the w_i's solve the system of equations above. Indeed, thanks to the Lagrange interpolating polynomial

$$\sum_{i=1}^{N} \frac{\prod_{1 \leq k \leq N, k \neq i} (u - u_k)}{\prod_{1 \leq k \leq N, k \neq i} (u_i - u_k)} u_i^j = u^j \quad \text{for } 0 \leq j \leq N-1$$

and the linearity of the operator $\mathcal{D}_{\zeta^2}$,(4.16) implies that for $0 \le j \le N-1$,

$$\sum_{i=1}^{N} w_i u_i^j = \rho \mathcal{D}_{\zeta^2} \left(\sum_{i=1}^{N} \frac{\prod_{1\le k\le N, k\neq i}(U-u_k)}{\prod_{1\le k\le N, k\neq i}(u_i-u_k)} u_i^j \right) = \rho \mathcal{D}_{\zeta^2}(U^j).$$

Namely, the w_i's defined in (4.16) solve the system of linear algebraic equations above.

Step II. With the w_i's defined in (4.16), we turn to the next N equations ($j = N, \dots, 2N-1$) in (4.15) to solve u_i:

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ u_1 & u_2 & \cdots & u_N \\ \vdots & \vdots & & \vdots \\ u_1^{N-1} & u_2^{N-1} & \cdots & u_N^{N-1} \end{bmatrix} \begin{bmatrix} w_1 u_1^N \\ w_2 u_2^N \\ \vdots \\ w_N u_N^N \end{bmatrix} = \rho \begin{bmatrix} \mathcal{D}_{\zeta^2}(U^N) \\ \mathcal{D}_{\zeta^2}(U^{N+1}) \\ \vdots \\ \mathcal{D}_{\zeta^2}(U^{2N-1}) \end{bmatrix}.$$

Again, the solution $w_i u_i^N$ is unique. As in **Step I**, we can show that

$$w_i u_i^N \prod_{1\le k\le N, k\neq i} (u_i - u_k) = \rho \mathcal{D}_{\zeta^2} \left(U^N \prod_{1\le k\le N, k\neq i} (U-u_k) \right) \tag{4.17}$$

for $1 \le i \le N$.

Substituting (4.16) into (4.17), we obtain

$$u_i^N \mathcal{D}_{\zeta^2} \left(\prod_{1\le k\le N, k\neq i} (U-u_k) \right) = \mathcal{D}_{\zeta^2} \left(U^N \prod_{1\le k\le N, k\neq i} (U-u_k) \right) \tag{4.18}$$

for $1 \le i \le N$. By the linearity of $\mathcal{D}_{\zeta^2}$, (4.18) is equivalent to

$$\mathcal{D}_{\zeta^2} \left((U^{N-1} + u_i U^{N-2} + \cdots + u_i^{N-1}) \prod_{k=1}^{N} (U-u_k) \right) = 0,$$

which can be rewritten as

$$\begin{bmatrix} 1 & u_1 & \cdots & u_1^{N-1} \\ \vdots & \vdots & & \vdots \\ 1 & u_N & \cdots & u_N^{N-1} \end{bmatrix} \begin{bmatrix} \mathcal{D}_{\zeta^2}(U^{N-1}F) \\ \vdots \\ \mathcal{D}_{\zeta^2}(F) \end{bmatrix} = 0$$

with $F = F(U) = \prod_{k=1}^{N}(U-u_k)$. Since all the u_i's are distinct, this says

$$\mathcal{D}_{\zeta^2}(F) = \mathcal{D}_{\zeta^2}(UF) = \cdots = \mathcal{D}_{\zeta^2}(U^{N-1}F) = 0. \tag{4.19}$$

Having this, in Lemma 4.3(e) we take $u_0 = U$ and $f(u) = u^j F(u)$ ($0 \le j \le N-2$) and deduce from (4.19) that

$$0 = \mathcal{D}_{\zeta^2}(U^{j+1}F) = \zeta^2 \mathcal{D}_{\zeta^2}\left((U^jF)'\right) = \zeta^2 \mathcal{D}_{\zeta^2}\left(U^jF'\right).$$

Hence $\mathcal{D}_{\zeta^2}(U^jF') = 0$ for $0 \le j \le N-2$. This procedure can be repeated for the derivative of $F'(u)$ to yield $\mathcal{D}_{\zeta^2}(U^jF'') = 0$ for $0 \le j \le N-3$. Moreover, we have

$$\mathcal{D}_{\zeta^2}\left(U^jF^{(k)}\right) = 0 \tag{4.20}$$

for $0 \le k \le N-1$ and $0 \le j \le N-1-k$.

Step III. In this step, we use (4.17) and the Lagrange interpolating polynomial

$$\sum_{i=1}^{N} \frac{\prod_{1\le k\le N, k\neq i}(u-u_k)}{\prod_{1\le k\le N, k\neq i}(u_i-u_k)} u_i^N = u^N - \prod_{k=1}^{N}(u-u_k)$$

to deduce that

$$\sum_{i=1}^{N} w_i u_i^{2N} = \rho\mathcal{D}_{\zeta^2}\left(U^N \sum_{i=1}^{N} \frac{\prod_{1\le k\le N, k\neq i}(U-u_k)}{\prod_{1\le k\le N, k\neq i}(u_i-u_k)} u_i^N\right) = \rho\mathcal{D}_{\zeta^2}\left(U^N(U^N-F)\right).$$

Thus, the last equation in (4.15) is equivalent to $\mathcal{D}_{\zeta^2}(U^N(U^N-F)) = \mathcal{D}_{\zeta^2}(U^{2N})$ or

$$\mathcal{D}_{\zeta^2}\left(U^N F\right) = 0.$$

Then we use Lemma 4.3(d) and (4.20) to see that

$$0 = \mathcal{D}_{\zeta^2}(U^N F) = \zeta^2 \mathcal{D}_{\zeta^2}(U^{N-1}F') = \cdots = \zeta^{2N}\mathcal{D}_{\zeta^2}(F^{(N)}) = N!\cdot\zeta^{2N},$$

which implies that $\zeta^2 = 0$. This contradicts the assumption that all the u_i's are distinct and $\zeta^2 > 0$. Hence the proof is complete. ∎

4.4. BGK model. In this subsection we show that the EQMOM moment system (2.19) with BGK source term

$$S(M) = \rho\left(\Delta_0(U,\theta),\ldots,\Delta_{2N}(U,\theta)\right)^T - M$$

satisfies the structural stability condition (i)–(iii). Indeed, (ii) has been verified in Remark 4.9.

To see Condition (i), we compute the Jacobian matrix of $S = S(M)$. Notice that the first three components of S vanish identically and ρ, U, θ depend only on M_0, M_1 and M_2. Then the Jacobian matrix can be written as

$$S_M(M) := \frac{\partial S}{\partial M} = \begin{bmatrix} 0_{3\times 3} & \\ \hat{S}_M & -I_{2N-2} \end{bmatrix}, \tag{4.21}$$

where $\hat{S}_M$ is a $(2N-2)\times 3$ matrix with

$$\left(\hat{S}_M\right)_{i-2,\ j+1} = \chi_i^j := \partial(\rho\Delta_i(U,\theta))/\partial M_j \tag{4.22}$$

for $3 \le i \le 2N$ and $j = 0,1,2$. Now we take

$$P = \begin{bmatrix} I_3 & \\ -\hat{S}_M & I_{2N-2} \end{bmatrix} \tag{4.23}$$

and see that $PS_M = \begin{bmatrix} 0_{3\times 3} & \\ & -I_{2N-2} \end{bmatrix} P$, which justifies Condition (i).

The rest of this subsection is to show Condition (iii). To this end, we need to choose the symmetrizer $A_0 = A_0(M)$. As pointed out in Remark 4.9, such a symmetrizer A_0 can only be of the form $L^T\Lambda L$ with Λ a diagonal positive definite matrix to be determined.

Firstly, we specify the matrix $L = (r_1^T, \ldots, r_{2N+1}^T)^T$ with r_i a left eigenvector of the coefficient matrix $A = A(M)$ corresponding to the eigenvalues λ_i for $1 \le i \le 2N+1$. Let $r_i = \left(r_i^{(1)}, \ldots, r_i^{(2N+1)}\right)$. From $r_i A = \lambda_i r_i$ we have

$$r_i^{(j)} + a_j r_i^{(2N+1)} = \lambda_i r_i^{(j+1)} \quad \text{for } 0 \le j \le 2N.$$

Here we have assumed $r_i^{(0)} = 0$ for simplicity. From the last equation we see that $r_i^{(2N+1)} \neq 0$; otherwise the eigenvector $r_i = 0$. Thus we may as well assume $r_i^{(2N+1)} = 1$. Recall that $a_{2N+1} = -1$. Then we can easily obtain

$$r_i^{(j)} = -\sum_{k=j}^{2N+1} a_k \lambda_i^{k-j}$$

for $0 \le j \le 2N$. Therefore, we have

$$L = \begin{bmatrix} \lambda_1^{2N} & \lambda_1^{2N-1} & \lambda_1^{2N-2} & \cdots & 1 \\ \lambda_2^{2N} & \lambda_2^{2N-1} & \lambda_2^{2N-2} & \cdots & 1 \\ \lambda_3^{2N} & \lambda_3^{2N-1} & \lambda_3^{2N-2} & \cdots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ \lambda_{2N+1}^{2N} & \lambda_{2N+1}^{2N-1} & \lambda_{2N+1}^{2N-2} & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 & & & & \\ -a_{2N} & 1 & & & \\ -a_{2N-1} & -a_{2N} & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ -a_1 & -a_2 & -a_3 & \cdots & 1 \end{bmatrix}. \tag{4.24}$$

With this L, we can state our main result of this subsection.

THEOREM 4.12. *For the EQMOM moment system (2.19), the inequality in the structural stability condition (iii) holds with $A_0 = L^T L$ and P defined in (4.23).*

Proof. According to Theorem 2.1 in [27], it suffices to show that at equilibrium states M,

$$K(M) := P^{-T} A_0 P^{-1} = (LP^{-1})^T (LP^{-1})$$

is of the block-diagonal form $\mathrm{diag}(K_1, K_2)$, in which K_1 and K_2 are 3×3 and $(2N-2) \times (2N-2)$ matrices, respectively. Namely, the first three columns of LP^{-1} are orthogonal to its other columns. In what follows all the states M are in equilibrium.

To show the orthogonality, we compute the $(2N+1)$-matrix $LP^{-1} := (b_{il})$. From (4.23) we see that

$$P^{-1} = \begin{bmatrix} I_3 & \\ \hat{S}_M & I_{2N-2} \end{bmatrix}. \tag{4.25}$$

This, together with (4.24), gives

$$b_{il} = \begin{cases} -\displaystyle\sum_{j=0}^{2N} \chi_j^{l-1} \sum_{k=j+1}^{2N+1} a_k \lambda_i^{k-j-1} & \text{for } 1 \le l \le 3, \\ -\displaystyle\sum_{k=l}^{2N+1} a_k \lambda_i^{k-l} & \text{for } 4 \le l \le 2N+1. \end{cases}$$

The expression above indicates that the last $(2N-2)$ columns of LP^{-1} are linear combinations of $(\lambda_1^\beta, \ldots, \lambda_{2N+1}^\beta)^T \in \mathbb{R}^{2N+1}$ for $0 \le \beta \le 2N-3$. Thus it reduces to show that

$$\sum_{i=1}^{2N+1} b_{il} \lambda_i^\beta = 0 \quad \text{for } l = 1, 2, 3 \text{ and } 0 \le \beta \le 2N-3.$$

Set $p_k = \sum_{i=1}^{2N+1} \lambda_i^k$. By using the above expression of b_{il} for $1 \le l \le 3$, the last equation is equivalent to

$$\sum_{j=0}^{2N} \chi_j^l \sum_{k=j+1}^{2N+1} a_k p_{k-j-1+\beta} = \sum_{j=-\beta}^{2N-\beta} \chi_{j+\beta}^l \sum_{k=\beta}^{2N-j} a_{j+k+1} p_k = 0 \tag{4.26}$$

for $l = 0, 1, 2$ and $0 \le \beta \le 2N-3$.

To prove (4.26), it suffices to show that

$$\sum_{j=-\beta}^{2N-\beta} \mathcal{H}_{j+\beta} \sum_{k=\beta}^{2N-j} a_{j+k+1} p_k = 0, \quad \text{for } 0 \le \beta \le 2N-3, \tag{4.27}$$

where $\mathcal{H}_j$ can be replaced by any of $\Delta_j = \Delta_j(U,\theta)$, $\partial_U \Delta_j$ and $\partial_\theta \Delta_j$. Indeed, (4.26) follows immediately from (4.27) and (4.22) which says

$$\chi_j^l = \frac{\partial}{\partial M_l}(\rho \Delta_j(U,\theta)) = \left(\frac{\partial \rho}{\partial M_l}\right) \Delta_j + \left(\rho \frac{\partial U}{\partial M_l}\right) \partial_U \Delta_j + \left(\rho \frac{\partial \theta}{\partial M_l}\right) \partial_\theta \Delta_j$$

for $0 \le j \le 2N$ and $l = 0, 1, 2$.

Before proceeding, two tools are needed. The first one is Newton's power sum formulas for p_k [8]:

$$\sum_{k=0}^{2N+1} a_k p_{k-j-1} = \sum_{k=-1-j}^{2N-j} a_{j+k+1} p_k = 0 \quad \text{for } j \le -2, \tag{4.28a}$$

$$(2N-j) a_{j+1} + \sum_{k=1}^{2N-j} a_{j+k+1} p_k = 0 \quad \text{for } -1 \le j \le 2N. \tag{4.28b}$$

The second tool is the following relation

$$\mathcal{D}_\theta \left(u^k c^{(j)}\right)\Big|_{u=U} = 0 \tag{4.29}$$

for $0 \le k \le 2N$ and $0 \le j \le 2N-k$, where $c^{(j)}$ denotes the jth derivative of the characteristic polynomial $c = c(u; W)$ with respect to u. This relation can be proved as below. Lemma 4.5 tells $g(u) = (u-U)^{2N+1}$ in equilibrium and therefore $g^{(j)}(U) = 0$ for $0 \le j \le 2N$. From Lemma 4.3(c) we see that $g^{(j)}(u) = \mathcal{D}_\theta c^{(j)}$ and thereby $\mathcal{D}_\theta c^{(j)}\big|_{u=U} = 0$ for $0 \le j \le 2N$. This is just the case for $k = 0$ in (4.29). Then using Lemma 4.3(d) we have

$$\mathcal{D}_\theta \left(u c^{(j)}\right)\Big|_{u=U} = U\ \mathcal{D}_\theta c^{(j)}\Big|_{u=U} + \theta\ \mathcal{D}_\theta c^{(j+1)}\Big|_{u=U} = 0$$

for $j = 0, \ldots, 2N-1$, which validates the case for $k = 1$. This procedure can be repeated to show (4.29) for other $k \le 2N$.

With these preparations, we only need to prove (4.27) for the following two cases.

Case I: $\beta = 0$. Noting that $p_0 = 2N+1$, we deduce from (4.28b) that

$$\sum_{k=0}^{2N-j} a_{k+j+1} p_k = (j+1) a_{j+1}.$$

Thus (4.27) in this case is equivalent to

$$\sum_{j=0}^{2N}(j+1)a_{j+1}\mathcal{H}_j = 0.$$

When taking $\mathcal{H}_j$ to be Δ_j, $\partial_U\Delta_j$ or $\partial_\theta\Delta_j$, the left-hand side of the last equation is equivalent to $\mathcal{D}_\theta c'|_{u=U}$, $\mathcal{D}_\theta c''|_{u=U}$ or $\mathcal{D}_\theta c'''|_{u=U}$, respectively. They are all equal to zero due to (4.29) and hence (4.27) with $\beta = 0$ is proved.

Case II: $\beta \geq 1$. As in Case I, we first simplify the coefficients $\sum_{k=\beta}^{2N-j} a_{j+k+1}p_k$ of $\mathcal{H}_{j+\beta}$ in (4.27) by using Newton's power sum formulas (4.28a) and (4.28b). They can be rewritten as

$$\sum_{k=-1-j}^{\beta-1} a_{j+k+1}p_k + \sum_{k=\beta}^{2N-j} a_{j+k+1}p_k = 0 \quad \text{for } j \leq -2,$$
$$\left((2N-j)a_{j+1} + \sum_{k=1}^{\beta-1} a_{j+k+1}p_k\right) + \sum_{k=\beta}^{2N-j} a_{j+k+1}p_k = 0 \quad \text{for } j \geq -1.$$

With these two relations, (4.27) is equivalent to

$$\begin{aligned} 0 &= \sum_{j=-\beta}^{2N-\beta} \mathcal{H}_{j+\beta} \sum_{k=\max\{1,-1-j\}}^{\beta-1} a_{j+k+1}p_k + \sum_{j=-1}^{2N-\beta}(2N-j)\mathcal{H}_{j+\beta}a_{j+1} \\ &= \sum_{k=1}^{\beta-1} p_k \sum_{j=-1-k}^{2N-\beta} \mathcal{H}_{j+\beta}a_{j+k+1} + \sum_{j=-1}^{2N-\beta}(2N-j)\mathcal{H}_{j+\beta}a_{j+1} \end{aligned} \tag{4.30}$$

for $1 \leq \beta \leq 2N-3$.

(4.30) can be further simplified by using the following relations

$$\sum_{j=-1-k}^{2N-k} \mathcal{H}_{j+\beta}a_{j+k+1} = 0 \quad \text{for } 1 \leq k \leq \beta-1.$$

Indeed, replacing $\mathcal{H}_j$ by Δ_j, $\partial_U\Delta_j$ or $\partial_\theta\Delta_j$, the sum is just $\mathcal{D}_\theta\left(u^{\beta-1-k}c\right)\big|_{u=U}$, $\mathcal{D}_\theta\left(u^{\beta-1-k}c'\right)\big|_{u=U}$ or $\mathcal{D}_\theta\left(u^{\beta-1-k}c''\right)\big|_{u=U}$, respectively. They are all equal to zero due to (4.29) and the fact that $0 \leq \beta-1-k \leq \beta-2 \leq 2N-5$. With the last relation, the first term in the right-hand side of (4.30) is reduced to

$$\begin{aligned} \sum_{k=1}^{\beta-1} p_k \sum_{j=-1-k}^{2N-\beta} \mathcal{H}_{j+\beta}a_{j+k+1} &= -\sum_{k=1}^{\beta-1} p_k \sum_{j=2N-\beta+1}^{2N-k} \mathcal{H}_{j+\beta}a_{j+k+1} \\ &= -\sum_{j=2N-\beta+1}^{2N-1} \mathcal{H}_{j+\beta} \sum_{k=1}^{2N-j} a_{j+k+1}p_k = \sum_{j=2N-\beta+1}^{2N-1} \mathcal{H}_{j+\beta}\left[(2N-j)a_{j+1}\right]. \end{aligned}$$

The last step resorts again to (4.28b) for $j \geq 2N+1-\beta \geq 4$. With this, (4.30) is equivalent to

$$\sum_{j=-1}^{2N-1}(2N-j)\mathcal{H}_{j+\beta}a_{j+1} = 0 \quad \text{for } 1 \leq \beta \leq 2N-3. \tag{4.31}$$

This is our final task.

In (4.31), we take $\mathcal{H}_j$ to be Δ_j, $\partial_U\Delta_j$ or $\partial_\theta\Delta_j$ and arrive at

$$\mathcal{D}_\theta\left(u^\beta c' - (2N+1)u^{\beta-1}c\right)^{(k)}\Big|_{u=U} = 0$$

for $1 \le \beta \le 2N-3$. Here $k = 0, 1, 2$ correspond to $\mathcal{H}_j = \Delta_j$, $\partial_U\Delta_j$ or $\partial_\theta\Delta_j$, respectively. The last relations hold due to (4.29) and the linearity of the operator $\mathcal{D}_\theta$. Hence the orthogonality is validated and the proof is completed. ∎

Remark 4.13. It is worth pointing out that the structural stability condition still holds if the collision frequency $\nu = \nu(M)$ in the BGK model depends on M, because in equilibrium $S = S(M) = 0$ and thus $\partial_M(\nu S) = \nu S_M(M) + S\partial_M\nu = \nu S_M(M)$. Hence all the analyses above are valid.

4.5. Shakhov model. This subsection is devoted to the EQMOM moment system of the 1-D Boltzmann equation with Shakhov source term (2.6). We first introduce the notation

$$\Delta_j^S = \Delta_j^S(U,\theta,q) = \frac{1}{\rho}\int_{\mathbb{R}} \xi^j f_S d\xi$$

with the equilibrium distribution $f_S = f_S(t,x,\xi)$ defined in (2.7). In this situation the moment system has the form (2.19) but the source term is different:

$$S^{Sh} = S^{Sh}(M) = \rho(\Delta_0^S, \Delta_1^S, \dots, \Delta_{2N}^S)^T - M.$$

We need to investigate whether this source term satisfies the structural stability condition (i) & (iii). For this purpose, some basic properties of Δ_j^S are required. A direct calculation shows that $\Delta_0^S = \Delta_0 = 1$, $\Delta_1^S = \Delta_1 = U$ and $\Delta_2^S = \Delta_2 = U^2 + \theta$. Moreover, for $j \ge 3$ we have

$$\begin{aligned}\rho\Delta_j^S - \rho\Delta_j &= \frac{q(1-Pr)}{3\theta^2}\int_{\mathbb{R}} \xi^j(\xi-U)\left(\frac{(\xi-U)^2}{\theta} - 3\right) f_{eq} d\xi \\ &= \binom{j}{3}(1-Pr)(2q)\Delta_{j-3} = \binom{j}{3}(1-Pr)(M_3 - \rho\Delta_3)\Delta_{j-3}.\end{aligned} \tag{4.32}$$

Here Lemma 4.1(a) is used for the integration and the last step is due to the definition of q.

As in Subsection 4.3, the equilibrium state W needs to be determined. From $S^{Sh}(M) = 0$ we see that $\rho\Delta_j^S = M_j$ for $0 \le j \le 2N$. With (4.32), the equation $\rho\Delta_3^S = M_3$ clearly implies that $M_3 = \rho\Delta_3$ for $Pr \ne 1$. Therefore, we have $\rho\Delta_j^S = \rho\Delta_j$ for any $0 \le j \le 2N$ and the equilibrium manifold $\mathcal{E}$ is determined by $M_j = \rho\Delta_j$ for any $0 \le j \le 2N$. This is exactly the same as that of the BGK model, which has already been determined in Theorem 4.11 to be $W \in \Omega_W^{eq}$.

At equilibrium, the Jacobian matrix of S^{Sh} can be computed with (4.32):

$$S_M^{Sh} := \frac{\partial S^{Sh}}{\partial M}\Big|_{S^{Sh}(M)=0} = \left(I_{2N+1} - (1-Pr)\sum_{i=3}^{2N}\binom{i}{3}\Delta_{i-3}E_{(i+1),4}\right) S_M, \tag{4.33}$$

where S_M is the Jacobian matrix (4.21) for the BGK model and the $(2N+1)$-matrix $E_{ij} = (e_{ij})$ with $e_{ij} = 1$ and all the other entities being zero. S_M^{Sh} is diagonalizable by an invertible matrix P^S such that $P^S S_M^{Sh} = -\operatorname{diag}(0,0,0,Pr,1,\dots,1)P^S$, and

$$\left(P^S\right)^{-1} = P^{-1} + \sum_{i=4}^{2N}\binom{i}{3}\Delta_{i-3}E_{(i+1),4}, \tag{4.34}$$

where P^{-1} is defined in (4.25). Hence the structural stability condition (i) is justified.

For Condition (iii), we take the same symmetrizer $A_0 = L^T L$ as that for the BGK model. This is reasonable since the equilibrium state is the same. It then suffices to show that the first three columns of $L\left(P^S\right)^{-1}$ are orthogonal to its other columns in equilbrium. From (4.34) we see that the only difference between $L\left(P^S\right)^{-1}$ and LP^{-1} is the fourth column. For $L\left(P^S\right)^{-1}$, its fourth column is a linear combination of the last $(2N-2)$ columns of L, while the last $(2N-2)$ columns of LP^{-1} are exactly those of L. Since the first three columns of LP^{-1} are orthogonal to its other columns, the fourth column of $L\left(P^S\right)^{-1}$ is also orthogonal to its first three columns. This has validated Condition (iii). In this way, we have the main result of this subsection:

THEOREM 4.14. *For the 1-D Boltzmann equation with the Shakhov model, the EQMOM moment system satisfies the structural stability condition.*

5. Conclusions. This paper presents a rigorous stability analysis of the quadrature based moment methods (QBMM) for the Boltzmann equation. To figure out a road map for more general cases, only the spatial one-dimensional (1-D) Boltzmann equation with hypothetical collisions (BGK or Shakhov type) is considered here. In the QBMM, the distribution function f is approximated with a linear combination of N ($N \geq 1$) δ-functions with unknown centers or their Gaussian approximations with unknown variance and centers (named QMOM or EQMOM, respectively). For QMOM, we show purely analytically that the resulting moment systems of first-order PDEs are not strongly hyperbolic for any N. Furthermore, we prove that the moment systems produced by the Gaussian EQMOM are strictly hyperbolic, when the variance is positive, and preserve the dissipation property of the kinetic equation. As a step in the proof, we also determine the equilibrium manifold that lies on the boundary of the state space for the parameters (w_i, u_i, σ^2) $(1 \leq i \leq N)$. These conclusions explain why the EQMOM gives reasonable numerical results while QMOM does not.

The proofs are quite technical and involve detailed analyses of the characteristic polynomial of the coefficient matrices. They offer a guideline to investigate the multidimensional cases with multiple nodes, which is underway.

Appendix A. Injectivity of EQMOM. In this appendix we show

PROPOSITION A.1. *For EQMOM, the map $M = \mathcal{M}(W)$ in (2.16) is injective for $W \in \Omega_W^{open}$ defined in (2.17a).*

Proof. It suffices to demonstrate that the Jacobian matrix $\frac{\partial \mathcal{M}}{\partial W}$ in (4.9) is invertible for $W \in \Omega_W^{open}$. In fact, we can show that

$$\det\left(\frac{\partial \mathcal{M}}{\partial W}\right) = \left(\prod_{i=1}^{N} w_i\right) \cdot \left(\sum_{i=1}^{N} w_i \prod_{\substack{j=1 \\ j\neq i}}^{N} (u_i - u_j)^2\right) \cdot \prod_{1\leq i<j\leq N} (u_i - u_j)^4$$

for multiple nodes $N \geq 2$. To this end, we set $\mathcal{F}(u) = \left(\Delta_0(u,\sigma^2), \Delta_1(u,\sigma^2), \cdots, \Delta_{2N}(u,\sigma^2)\right)^T$ and see from (4.9) that

$$\begin{aligned}
\det\left(\frac{\partial \mathcal{M}}{\partial W}\right) &= \left(\prod_{i=1}^{N} w_i\right) \cdot \det\left(\mathcal{F}(u_1), \mathcal{F}'(u_1), \cdots, \mathcal{F}(u_N), \mathcal{F}'(u_N), \frac{1}{2}\sum_{i=1}^{N} w_i \mathcal{F}''(u_i)\right) \\
&= \left(\prod_{i=1}^{N} w_i\right) \cdot \sum_{i=1}^{N} w_i \det\left(\mathcal{F}(u_1), \mathcal{F}'(u_1), \cdots, \mathcal{F}(u_N), \mathcal{F}'(u_N), \frac{1}{2}\mathcal{F}''(u_i)\right).
\end{aligned}$$

Denote

$$f_N(u_1,u_2,\dots,u_N;\sigma)=\det\left(\mathcal{F}(u_1),\mathcal{F}'(u_1),\cdots,\mathcal{F}(u_N),\mathcal{F}'(u_N),\frac{1}{2}\mathcal{F}''(u_1)\right).$$

And we see that for each $1\le i\le N$,

$$\det\left(\mathcal{F}(u_1),\mathcal{F}'(u_1),\cdots,\mathcal{F}(u_N),\mathcal{F}'(u_N),\frac{1}{2}\mathcal{F}''(u_i)\right)=f_N(u_i,u_2,\dots,u_{i-1},u_1,u_{i+1},\dots,u_N;\sigma),$$

and thereby

$$\det\left(\frac{\partial\mathcal{M}}{\partial W}\right)=\left(\prod_{i=1}^N w_i\right)\cdot\sum_{i=1}^N w_i f_N(u_i,u_2,\dots,u_{i-1},u_1,u_{i+1},\dots,u_N;\sigma).$$

Thus, it remains to show

$$f_N(u_1,u_2,\dots,u_N;\sigma)=C(N)\prod_{j=2}^N(u_j-u_1)^2\cdot\prod_{1\le i<j\le N}(u_i-u_j)^4 \tag{A.1}$$

and

$$C(N)=1\quad\text{for } N\ge 2. \tag{A.2}$$

Note that $f_N(u_1,\dots,u_N;\sigma)$ is a homogeneous polynomial of $u_1,\dots,u_N,\sigma$ with degree $2(N^2-1)$. This can be seen from the definition of determinant and the fact that $\Delta_j(u,\sigma^2)$ is a homogeneous polynomial of u and σ with degree j (see Lemma 4.1(a)). On the other hand, the right-hand side of (A.1) is also a homogeneous polynomial of $u_1,\dots,u_N,\sigma$ with degree $2(N^2-1)$. Thus, to prove (A.1), we need to show that $(u_j-u_1)^6$ and $(u_j-u_i)^4$ are factors of $f(u_1,\dots,u_N;\sigma)$ for any $2\le j\ne i\le N$.

From the definition of $f_N=f_N(u_1,\dots,u_N;\sigma)$, it is not difficult to compute that for $j\ne 1$,

$$\begin{aligned}
\partial_{u_j}f_N&=\det\left(\mathcal{F}(u_1),\mathcal{F}'(u_1),\cdots,\mathcal{F}(u_j),\mathcal{F}''(u_j),\cdots,\frac{1}{2}\mathcal{F}''(u_1)\right),\\
\partial_{u_j}^2f_N&=\det\left(\cdots,\mathcal{F}'(u_j),\mathcal{F}''(u_j),\cdots\right)+\det\left(\cdots,\mathcal{F}(u_j),\mathcal{F}'''(u_j),\cdots\right),\\
\partial_{u_j}^3f_N&=\det\left(\cdots,\mathcal{F}(u_j),\mathcal{F}^{(4)}(u_j),\cdots\right)+2\det\left(\cdots,\mathcal{F}'(u_j),\mathcal{F}'''(u_j),\cdots\right),\\
\partial_{u_j}^4f_N&=\det\left(\cdots,\mathcal{F}(u_j),\mathcal{F}^{(5)}(u_j),\cdots\right)+3\det\left(\cdots,\mathcal{F}'(u_j),\mathcal{F}^{(4)}(u_j),\cdots\right)\\
&\quad+2\det\left(\cdots,\mathcal{F}''(u_j),\mathcal{F}'''(u_j),\cdots\right),\\
\partial_{u_j}^5f_N&=\det\left(\cdots,\mathcal{F}(u_j),\mathcal{F}^{(6)}(u_j),\cdots\right)+4\det\left(\cdots,\mathcal{F}'(u_j),\mathcal{F}^{(5)}(u_j),\cdots\right)\\
&\quad+5\det\left(\cdots,\mathcal{F}''(u_j),\mathcal{F}^{(4)}(u_j),\cdots\right).
\end{aligned}$$

Thus it follows that for $j\ne 1$,

$$f_N|_{u_j=u_i}=0,\quad \partial_{u_j}f_N\big|_{u_j=u_i}=0,\quad \partial_{u_j}^2f_N\Big|_{u_j=u_i}=0,\quad \partial_{u_j}^3f_N\Big|_{u_j=u_i}=0$$

for any $1\le i\ne j\le N$ and that

$$\partial_{u_j}^4f_N\Big|_{u_j=u_1}=0,\quad \partial_{u_j}^5f_N\Big|_{u_j=u_1}=0.$$

This justifies (A.1) and $C(N)$ is a constant.

Then all we need is to prove (A.2). A direct calculation for $N = 2$ indicates that $f_2(u_1, u_2; \sigma) = (u_1 - u_2)^4$ and thus $C(2) = 1$. For $N > 2$, we deduce from (A.1) that the leading coefficient of u_N (with degree $(4N - 2)$) is

$$C(N) \prod_{j=2}^{N-1} (u_j - u_1)^2 \cdot \prod_{1 \le i < j \le N-1} (u_i - u_j)^4$$

On the other hand, the determinant definition of $f_N(u_1, \ldots, u_N; \sigma)$ implies that the leading term of u_N (with degree $(4N - 2)$) is included in the following part:

$$f_{N-1}(u_1, \ldots, u_{N-1}; \sigma) \times \det \begin{bmatrix} \Delta_{2N-1}(u_N) & (2N-1)\Delta_{2N-2}(u_N) \\ \Delta_{2N}(u_N) & 2N\Delta_{2N-1}(u_N) \end{bmatrix}.$$

Thus, using (A.1), the leading coefficient of u_N is

$$f_{N-1}(u_1, \ldots, u_{N-1}; \sigma) = C(N-1) \prod_{j=2}^{N-1} (u_j - u_1)^2 \cdot \prod_{1 \le i < j \le N-1} (u_i - u_j)^4.$$

By equating the leading coefficients of u_N in the above two expressions, we see immediately that $C(N) = C(N - 1)$ for $N > 2$. Since $C(2) = 1$, this justifies (A.2) and hence completes the proof. ∎

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), pp. 511–525.

[2] Z. Cai, Y. Fan, and R. Li, *Globally hyperbolic regularization of Grad's moment system*, Commun. Pur. Appl. Math., 67 (2013), pp. 464–518.

[3] Z. Cai, Y. Fan, and R. Li, *Globally hyperbolic regularization of Grad's moment system in one dimensional space*, Commun. Math. Sci., 11 (2013), pp. 547–571.

[4] Z. Cai, Y. Fan, and R. Li, *A framework on moment model reduction for kinetic equation*, SIAM J. Appl. Math., 75 (2015), pp. 2001–2023.

[5] C. Chalons, R. Fox, F. Laurent, M. Massot, and A. Vié, *Multivariate Gaussian extended quadrature method of moments for turbulent disperse multiphase flow*, Multiscale Model. Simul., 15 (2017), pp. 1553–1583.

[6] C. Chalons, D. Kah, and M. Massot, *Beyond pressureless gas dynamics: Quadrature-based velocity moment models*, Commun. Math. Sci., 10 (2012), pp. 1241–1272.

[7] Y. Di, Y. Fan, R. Li, and L. Zheng, *Linear stability of hyperbolic moment models for Boltzmann equation*, Numer. Math. Theor. Meth. Appl., 10 (2017), pp. 255–277.

[8] J. A. Eidswick, *A proof of Newton's power sum formulas*, Am. Math. Mon., 75 (1968), pp. 396–397.

[9] R. O. Fox, *A quadrature-based third-order moment method for dilute gas-particle flows*, J. Comput. Phys., 227 (2008), pp. 6313–6350.

[10] S. K. Friedlander, *Smoke, Dust, and Haze: Fundamentals of Aerosol Dynamics*, Oxford University Press, Oxford, 2nd ed., 2000.

[11] R. Gatignol, *Théorie Cinétique de Gaz à Répartition Discrète de Vitesses*, Springer, New York, 1975.

[12] H. Grad, *On the kinetic theory of rarefield gases*, Comm. Pure Appl. Math., 2 (1949), pp. 331–407.

[13] M. Herty, A. Tosin, G. Visconti, and M. Zanella, *Hybrid stochastic kinetic description of two-dimensional traffic dynamics*, SIAM J. Appl. Math., 78 (2018), pp. 2737–2762.

[14] Q. Huang, S. Li, G. Li, and Q. Yao, *Mechanisms on the size partitioning of sodium in particulate matter from pulverized coal combustion*, Combust. Flame, 182 (2017), pp. 313–323.

[15] T. Kato, *Perturbation Theory for Linear Operators*, Springer, New York, 2nd ed., 1980.
[16] G. M. Kremer, *An Introduction to the Boltzmann Equation and Transport Processes in Gases*, Springer, New York, 2010.
[17] C. Levermore, *Moment closure hierarchies for kinetic theories*, J. Stat. Phys., 83 (1996), pp. 1021–1065.
[18] J. Liu and W.-A. Yong, *Stability analysis of the Biot/Squirt models for wave propagation in saturated porous media*, Geophys. J. Int., 204 (2016), pp. 535–543.
[19] D. L. Marchisio and R. O. Fox, *Computational Models for Polydisperse Particulate and Multiphase Systems*, Cambridge University Press, Cambridge, 2013.
[20] J. McDonald and M. Torrilhon, *Affordable robust moment closures for cfd based on the maximum-entropy hierarchy*, J. Comput. Phys., 251 (2013), pp. 500–523.
[21] L. Mieussens, *Discrete velocity model and implicit scheme for the BGK equation of rarefield gas dynamics*, Math. Mod. Meth. Appl. S, 10 (2000), pp. 1121–1149.
[22] I. Müller and T. Ruggeri, *Rational Extended Thermodynamics*, Springer, New York, 2nd ed., 1998.
[23] T. T. Nguyen, F. Laurent, R. O. Fox, and M. Massot, *Solution of population balance equations in applications with fine particles: Mathematical modeling and numerical schemes*, J. Comput. Phys., 325 (2016), pp. 129–156.
[24] M. Pigou, J. Morchain, P. Fede, M. Penet, and G. Laronze, *New developments of the extended quadrature method of moments to solve population balance equations*, J. Comput. Phys., 365 (2018), pp. 243–268.
[25] E. M. Shakhov, *Generalization of the Krook kinetic relaxation equation*, Fluid Dyn., 3 (1968), pp. 95–96.
[26] H. Struchtrup and M. Torrilhon, *Regularization of grad's 13 moment equaitons: Derivation and linear analysis*, Phys. Fluids, 15 (2003), pp. 2668–2680.
[27] W.-A. Yong, *Singular perturbations of fisrt-order hyperbolic systems with stiff source terms*, J. Differ. Equations, 155 (1999), pp. 89–132.
[28] W.-A. Yong, *An interesting class of partial differential equations*, J. Math. Phys., 49 (2008), 033503.
[29] K. Yoshikawa, N. Yoshida, and M. Umemura, *Direct integration of the collisionless Boltzmann equation in six-dimensional phase space: Self-gravitating systems*, Astrophys. J., 762 (2013), 116.
[30] C. Yuan and R. Fox, *Conditional quadrature method of moments for kinetic equations*, J. Comput. Phys., 230 (2011), pp. 8216–8246.