Commun. Comput. Phys.
doi: 10.4208/cicp.OA-2016-0053

Vol. **21**, No. 4, pp. 981-1011
April 2017

Numerical Study of Partially Conservative Moment Equations in Kinetic Theory

Julian Koellermeier[1,*] and Manuel Torrilhon[1]

[1] *Department of Mathematics, RWTH Aachen University, Schinkelstraße 2, 52062 Aachen, Germany.*

Received 13 April 2016; Accepted (in revised version) 21 September 2016

Abstract. Moment models are often used for the solution of kinetic equations such as the Boltzmann equation. Unfortunately, standard models like Grad's equations are not hyperbolic and can lead to nonphysical solutions. Newly derived moment models like the Hyperbolic Moment Equations and the Quadrature-Based Moment Equations yield globally hyperbolic equations but are given in partially conservative form that cannot be written as a conservative system.

In this paper we investigate the applicability of different dedicated numerical schemes to solve the partially conservative model equations. Caused by the non-conservative type of equation we obtain differences in the numerical solutions, but due to the structure of the moment systems we show that these effects are very small for standard simulation cases. After successful identification of useful numerical settings we show a convergence study for a shock tube problem and compare the results to a discrete velocity solution. The results are in good agreement with the reference solution and we see convergence considering an increasing number of moments.

AMS subject classifications: 35Q20, 35Q35, 35L02

Key words: Kinetic theory, moment method, hyperbolicity, non-conservative numerics.

1 Introduction

Numerical simulations of applications involving rarefied gases require the solution of the Boltzmann transport equation (BTE), especially if moderate to large Knudsen numbers are considered and particle methods like Direct Simulation Monte Carlo (see [3]) on the one hand are still too costly whereas standard fluid dynamics equations on the other hand are not accurate enough to describe possible non-equilibrium effects like the Knudsen paradox (see [20]).

*Corresponding author. *Email addresses:* koellermeier@mathcces.rwth-aachen.de (J. Koellermeier), mt@mathcces.rwth-aachen.de (M. Torrilhon)

 ©2017 Global-Science Press

https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press

Moment methods are one possibility to efficiently solve the BTE as relatively few variables (usually the so-called *moments*) are used to describe the flow solution. The first example is Grad's method (see [8]) which uses an expansion of the unknown distribution function around local equilibrium in Hermite polynomials and results in a system of equations for the coefficients of this expansion, also called *moment system*. However, it has been shown in [4] that the resulting PDE system of Grad's method looses hyperbolicity for larger deviations from equilibrium, resulting in nonphysical infinite or imaginary propagation wave speeds. Hyperbolicity is a necessary condition for stable solutions of a moment system (see also [24]).

The maximum entropy approach proposed by Levermore (see e.g. [15]) has good approximation properties and is hyperbolic by construction, but unfortunately does not have an explicit flux formulation. Instead it requires a costly optimization which makes this method difficult to apply for a large number of moments. An ansatz based on multivariate Pearson-IV-Distributions was proposed by Torrilhon in [22] but is difficult to generalize to the case with more higher order moments.

Following the work in [4] and [10], a new framework for the derivation of hyperbolic moment models has been derived in [7]. The proposed operator projection framework can be seen as a generalization of Grad's method with additional modification of the equations such that the resulting moment system is globally hyperbolic. Two examples are the hyperbolic moment equations (HME, first derived in [4]) and the quadrature-based moment equations (QBME, first derived in [11]). Both methods yield very similar systems of equations as exemplified for a simple test case in [12]. In comparison to Grad's standard model, only some of the higher order equations are changed. Both models exhibit the same physical adaptivity as can be found either in the original model by Grad or due to the transformation of the Boltzmann equation by Kauf in [9].

Despite the beneficial hyperbolicity, a major drawback of the novel equations is the lack of a conservative form. Due to small changes in the last equations, the new HME and QBME systems are in so-called *partially conservative* form, which means that only a subset of the whole system can be written in conservative form. One or two of the last equations are not in conservative form which makes the application of standard finite volume schemes impossible. In this paper we therefore apply non-conservative numerical schemes to solve the system, following the theory in [16].

Various analytical properties of HME and QBME have been extensively studied in the papers mentioned before. The purpose of this paper is the investigation and justification of different numerical schemes to solve the partially conservative PDE system. It is important to study the performance of different schemes first, before addressing model accuracy itself because the solution of non-conservative systems is much more difficult than in the standard conservative case and different schemes can converge to different numerical solutions. It is our aim to investigate these problems and eventually identify the best scheme. Afterwards, we want to use this scheme to get more numerical results for the new moment systems. Consequently, the tests with different schemes will be followed by a study of the model performance with respect to accuracy and convergence

https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press

towards a reference solution.

The rest of this paper is organized as follows: In Section 2 we recall the Boltzmann transport equation and explain the necessary steps of a moment method to derive a closed system of equations describing rarefied gas flows using Grad's method as one example. The loss of hyperbolicity is described in Section 3 along with the derivation of the new hyperbolic moment models. The 1D examples for Grad, HME and QBME are given explicitly for an arbitrary number of equations along with explicit eigenvalues and eigenvectors for HME and QBME. The important question of suitable numerical methods is considered by Section 4, in which we describe different numerical algorithms that can be used to solve the non-conservative PDE system. Results are given in Section 5 for a shock tube test case identifying suitable non-conservative numerical schemes. After that we compare the new models and address their accuracy with respect to a discrete velocity solution by a convergence study. The paper ends with concluding remarks.

2 Moment methods for the Boltzmann equation

In this section, we introduce our notation and the Boltzmann equation before we explain the moment method by recalling the derivation of Grad's moment system.

2.1 Boltzmann transport equation

In kinetic theory the motion of particles is described by the mass density function $f(t,x,c)$ and its evolution is given according to the Boltzmann transport equation. As our kinetic equation we consider a 1D version of the Boltzmann equation

$$\frac{\partial}{\partial t}f(t,x,c)+c\frac{\partial}{\partial x}f(t,x,c)=S(f), \tag{2.1}$$

where we have $x\in\mathbb{R}$ and $c\in\mathbb{R}$ denoting position and microscopic velocity, respectively. Throughout the paper we will only consider the 1D case. The collision operator $S(f)$ on the right hand side of Eq. (2.1) can be modelled in various ways. For the tests conducted in Section 5 we consider a BGK collision operator (see [2])

$$S(f)=-\frac{1}{\tau}(f-f_M), \tag{2.2}$$

where $\tau\in\mathbb{R}$ denotes the relaxation time and $f_M(t,x,c)$ is the local Maxwellian

$$f_M(t,x,c)=\frac{\rho(t,x)}{\sqrt{2\pi\theta(t,x)}}\exp\left(-\frac{(c-u(t,x))^2}{2\theta(t,x)}\right). \tag{2.3}$$

More advanced models for the collision operator $S(f)$ can be found in [20].

https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press

From the distribution function $f(t,x,c)$ we can readily extract the values of the macroscopic quantities density $\rho(t,x)$, velocity $u(t,x)$ and temperature $\theta(t,x)$ by integration of $f(t,x,c)$ over the velocity space.

$$\rho(t,x)=\int_{\mathbb{R}} f(t,x,c)\ dc, \tag{2.4}$$

$$\rho(t,x)u(t,x)=\int_{\mathbb{R}} cf(t,x,c)\ dc, \tag{2.5}$$

$$\rho(t,x)\theta(t,x)=\int_{\mathbb{R}} |c-u|^2 f(t,x,c)\ dc. \tag{2.6}$$

In order to derive a system of equations for these macroscopic variables, the multiplication of the Boltzmann equation (2.1) with monomials $\left(1,c,c^2\right)$ followed by integration over the velocity space yields the well-known conservation laws of mass, momentum and energy, see e.g. [23].

2.2 Moment method

Rarefied flows often contain so-called non-equilibrium effects which cannot be described by basic models like the Euler equations. The aim of the moment method is to derive more advanced equations that describe the fluid flow more accurately. Moment methods date back to the first approach by Grad in 1949 (see [8]), who expanded the distribution function in a series of Hermite basis functions around local equilibrium. The expansion can be written in the following way (see also [7]):

$$f(t,x,c)=\sum_{\alpha\in\mathbb{N}} f_\alpha(t,x)\mathcal{H}_\alpha^{[u(t,x),\theta(t,x)]}(c), \tag{2.7}$$

where $f_\alpha(t,x)$, $\alpha\in\mathbb{N}$, are the expansion coefficients and the weighted Hermite polynomials $\mathcal{H}_\alpha^{[u,\theta]}$ are defined as

$$\mathcal{H}_\alpha^{[u,\theta]}(c)=(-1)^\alpha \frac{\mathrm{d}^\alpha}{\mathrm{d}c^\alpha}\omega^{[u,\theta]}(c),\quad \alpha\geq 0,\quad \omega^{[u,\theta]}(c)=\frac{1}{\sqrt{2\pi\theta}}\exp\left(-\frac{(c-u)^2}{2\theta}\right). \tag{2.8}$$

Using orthogonality of the weighted Hermite polynomials with respect to the scalar product $<f,g>=\int f\cdot g\cdot\omega^{[u,\theta]}(c)^{-1}dc$, we get the following constraints:

$$f_0=\rho,\quad f_1=f_2=0. \tag{2.9}$$

After substitution of expansion (2.7) into the Boltzmann equation (2.1) the application of suitable derivative and recurrence relations yields explicit formulas for the time derivative and the spatial term in (2.1) (see [7] for details on the derivation).

Grad's moment system can be readily derived matching the coefficients of the basis functions or projecting the equation onto Hermite test functions via integration over velocity space.

https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press

According to [7] the infinite expansion (2.7) leads to the following system:

$$\begin{aligned}&\frac{\partial f_\alpha}{\partial t}+u\frac{\partial f_\alpha}{\partial x}+\theta\frac{\partial f_{\alpha-1}}{\partial x}+(\alpha+1)\frac{\partial f_{\alpha+1}}{\partial x}\\&\quad+f_{\alpha-1}\frac{\partial u}{\partial t}+(uf_{\alpha-1}+\theta f_{\alpha-2}+(\alpha+1)f_\alpha)\frac{\partial u}{\partial x}\\&\quad+\frac{f_{\alpha-2}}{2}\frac{\partial\theta}{\partial t}+\frac{1}{2}(uf_{\alpha-2}+\theta f_{\alpha-3}+(\alpha+1)f_{\alpha-1})\frac{\partial\theta}{\partial x}=S_\alpha,\quad \alpha\geq 3,\end{aligned}\tag{2.10}$$

where the term S_α can be derived by the same expansion of the substituted collision term $S(f)$. Due to the constraints (2.9) we let the unknown vector be $\boldsymbol{w}=(\rho,u,\theta,f_3,f_4,\cdots)$ and write (2.10) as follows:

$$\boldsymbol{D}\frac{\partial\boldsymbol{w}}{\partial t}+\boldsymbol{M}\boldsymbol{D}\frac{\partial\boldsymbol{w}}{\partial x}=\boldsymbol{S},\tag{2.11}$$

where the infinite matrices $\boldsymbol{M}$ and $\boldsymbol{D}$ can be computed using the derivative and recurrence relations of the Hermite polynomials and the vector $\boldsymbol{S}=(S_\alpha)_{\alpha\in\mathbb{N}}$ contains the corresponding right hand side terms in ascending order of the index α.

System (2.11) yields an infinite number of equations. In order to end up with a finite system, the expansion is cut off after an index $M\in\mathbb{N}$, leaving only $M+1$ equations. In Grad's approach, the term f_{M+1} appearing in the last equation is set to zero to close the system. The result can be written with new finite matrices and variable vector as

$$\boldsymbol{D}_M\frac{\partial\boldsymbol{w}_M}{\partial t}+(\boldsymbol{M}\boldsymbol{D})_M\frac{\partial\boldsymbol{w}_M}{\partial x}=\boldsymbol{S}_M,\tag{2.12}$$

where the matrices $\boldsymbol{D}_M$ and $(\boldsymbol{M}\boldsymbol{D})_M$ can be computed either directly from (2.10) or with the help of a projection procedure explained in the next section.

Remark 2.1 (Transformed velocity space)**.** Note that expansion (2.7) together with the definition of the weighted basis functions (2.8) is equivalent to the use of a shifted version of the Boltzmann transport equation and suitable expansion as used in [11]. In the first case, the velocity variable inside the weighted Hermite polynomial is transformed according to $c\mapsto\frac{c-u(t,x)}{\theta(t,x)}$, whereas in the latter case the equation itself is transformed and the basis function is then simply defined in the transformed velocity variable. The reason for the transformed velocity space in both cases is a more efficient discretization of the transformed velocity space that automatically adapts to the current flow situation. This allows to exhibit physical adaptivity obviously at the expense of a more difficult system of equations. Without the transformed velocity space it would be much more difficult to capture flow behavior in situations with varying and large velocities or temperatures.

3 Hyperbolic moment equations

Grad's system as derived in Section 2 is not hyperbolic. We will point out consequences thereof and present the operator projection framework first developed in [7] in order to

https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press

overcome this and to derive new hyperbolic equations. We will describe two of the new models and their analytical properties.

3.1 Hyperbolicity and its importance

According to [4] we define hyperbolicity as follows:

Definition 3.1 (Hyperbolicity)**.** A system of PDEs in 1D for $\boldsymbol{w}\in\mathbb{R}^{M+1}$ and $\boldsymbol{A}(\boldsymbol{w}),\boldsymbol{B}(\boldsymbol{w})\in\mathbb{R}^{(M+1)\times(M+1)}$, e.g.

$$\boldsymbol{B}(\boldsymbol{w})\frac{\partial}{\partial t}\boldsymbol{w}+\boldsymbol{A}(\boldsymbol{w})\frac{\partial}{\partial x}\boldsymbol{w}=0, \tag{3.1}$$

is *globally hyperbolic* if $\boldsymbol{B}(\boldsymbol{w})$ is invertible and the eigenvalues of the diagonalizable matrix

$$\boldsymbol{A}_{sys}(\boldsymbol{w})=(\boldsymbol{B}(\boldsymbol{w}))^{-1}\boldsymbol{A}(\boldsymbol{w}) \tag{3.2}$$

are real for every $\boldsymbol{w}\in\mathbb{R}^{n+1}$.

Hyperbolic systems can be diagonalized by means of a proper eigenvalue decomposition and thus describe physical processes in which waves propagate with certain real-valued propagation speeds that are the eigenvalues of the system matrix. As we want our models to accurately capture physical phenomena, it is crucial to obtain hyperbolic systems. In case hyperbolicity is lost, waves can propagate with either infinite or even imaginary wave speeds, which is clearly nonphysical.

Because of Definition 3.1 we see that also conditionally hyperbolic systems can occur. Such systems behave well when inside their hyperbolicity region, e.g. for a specific range of values $\boldsymbol{w}\in\mathbb{R}^{M+1}$, but as soon as hyperbolicity is lost, numerical simulations will show nonphysical oscillations possibly causing a breakdown of the simulation and at least leading to inaccurate results.

3.2 Hyperbolicity loss of Grad's equations

For Grad's system with $M+1$ equations, we can analytically derive the system matrix $\boldsymbol{A}_{\text{Grad}}=\boldsymbol{D}_M^{-1}(\boldsymbol{M}\boldsymbol{D})_M$, such that

$$\frac{\partial\boldsymbol{w}_M}{\partial t}+\boldsymbol{A}_{\text{Grad}}\frac{\partial\boldsymbol{w}_M}{\partial x}=\boldsymbol{S}_M, \tag{3.3}$$

and we note that $\boldsymbol{D}_M^{-1}\boldsymbol{S}_M=\boldsymbol{S}_M$ due to the special structure of $\boldsymbol{S}_M$ acting only on higher order coefficients.

https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press

In order to later ease comparison to our new model equations, we recall from [4] that the matrix $\boldsymbol{A}_{\text{Grad}}$ is given by

$$\begin{pmatrix}
u & \rho & 0 & \dots & & & & \dots & 0 \\
\frac{\theta}{\rho} & u & 1 & 0 & \dots & & & & \vdots \\
0 & 2\theta & u & \frac{6}{\rho} & 0 & \dots & & & \\
0 & 4f_3 & \frac{\rho\theta}{2} & u & 4 & 0 & \dots & & \vdots \\
-\frac{\theta f_3}{\rho} & 5f_4 & \frac{3f_3}{2} & \theta & u & 5 & 0 & \dots & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
-\frac{\theta f_{M-2}}{\rho} & Mf_{M-1} & \frac{(M-2)f_{M-2}+\theta f_{M-4}}{2} & -\frac{3f_{M-3}}{\rho} & 0 & \dots & 0 & \theta & u & M \\
-\frac{\theta f_{M-1}}{\rho} & (M+1)f_M & \frac{(M-1)f_{M-1}+\theta f_{M-3}}{2} & -\frac{3f_{M-2}}{\rho} & 0 & \dots & \dots & 0 & \theta & u
\end{pmatrix}. \tag{3.4}$$

The characteristic polynomial χ_M of the system matrix $\boldsymbol{A}_{\text{Grad}} \in \mathbb{R}^{(M+1)\times(M+1)}$ can be derived analytically and reads as follows:

$$\chi_M(\lambda) = \theta^{\frac{M+1}{2}} \text{He}_{M+1}\left(\frac{\lambda-u}{\sqrt{\theta}}\right) - \frac{(M+1)!}{2\rho}\left(\left((\lambda-u)^2-\theta\right)f_{M-1}+2(\lambda-u)f_M\right), \tag{3.5}$$

where He_{M+1} is the Hermite polynomial of degree $M+1$. The characteristic polynomial of Grad's model thus depends on the two highest coefficients f_{M-1} and f_M. The roots of χ_M can be shown to have the following form

$$\lambda_j = u + \sqrt{\theta} c_j(\theta, f_{M-1}, f_M), \quad j = 1, \cdots, M+1, \tag{3.6}$$

with $c_j = c_j(\theta, f_{M-1}, f_M)$ depending non-linearly on the higher order coefficients and θ. As the roots of the characteristic polynomial are the eigenvalues of the system matrix $\boldsymbol{A}_{\text{Grad}}$, those eigenvalues can become imaginary for certain values of f_{M-1} and f_M. Already for the popular five moment case $M=4$, the model is only hyperbolic in a small range of values around equilibrium, which is denoted by $f_3 = f_4 = 0$. Even for moderate non-equilibrium or relatively small values of f_{M-1} and f_M, the model is not hyperbolic and shows nonphysical behavior as shown in [4]. This includes test cases with large Mach numbers or strong shocks as exemplified in Section 5.1 by a shock tube test.

3.3 Operator projection framework

The operator projection framework first developed in [7] is a successful way to derive new hyperbolic model equations. With few inputs and an algorithmic fashion it is possible to analytically derive explicit expressions of a new moment system and prove hyperbolicity. For the theoretical background we refer to the original paper, here we only want to give a short summary of the method and focus on the models afterwards.

https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press

Starting from the infinite moment system (2.11), a new system of equations can be obtained by subsequent *projection* of the equations onto a subspace of the polynomial ansatz space. In the end this reduces to the projection of the occurring matrices $\boldsymbol{D}$ and $\boldsymbol{M}$ in (2.11) using a projection operator $\boldsymbol{P}\in\mathbb{R}^{(M+1)\times\infty}$ as follows:

$$(\boldsymbol{D})_M\frac{\partial \boldsymbol{w}_M}{\partial t}+(\boldsymbol{M})_M(\boldsymbol{D})_M\frac{\partial \boldsymbol{w}_M}{\partial x}=(\boldsymbol{S})_M, \tag{3.7}$$

with finite-dimensional projected variables and matrices

$$\begin{aligned}\boldsymbol{w}_M&=\boldsymbol{P}\boldsymbol{w},\\ (\boldsymbol{D})_M&=\boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^T,\\ (\boldsymbol{M})_M&=\boldsymbol{P}\boldsymbol{M}\boldsymbol{P}^T,\\ (\boldsymbol{S})_M&=\boldsymbol{P}\boldsymbol{S}.\end{aligned}$$

Provided invertibility of the matrix $(\boldsymbol{D})_M$, the system can thus be written in the following form:

$$\frac{\partial \boldsymbol{w}_M}{\partial t}+(\boldsymbol{D})_M^{-1}(\boldsymbol{M})_M(\boldsymbol{D})_M\frac{\partial \boldsymbol{w}_M}{\partial x}=(\boldsymbol{D})_M^{-1}\boldsymbol{S}_M, \tag{3.8}$$

where global hyperbolicity can be directly obtained if $(\boldsymbol{M})_M$ has only real eigenvalues.

The important difference of the operator projection approach compared to Grad's approach is the fact that in front of the spatial derivative both matrices $\boldsymbol{M}$ and $\boldsymbol{D}$ are projected individually before multiplication. In the case of Grad's equations, the product $\boldsymbol{MD}$ is projected by cutting off higher order equations.

3.4 Model equations

The projection operator approach leads to new moment models which can be shown to be globally hyperbolic. We will now give some examples that differ in the explicit form of the projection operator and the way it is applied in detail.

3.4.1 Hyperbolic moment equations

The hyperbolic moment equations (HME) first derived in [4] using a special modification of the system matrix can be derived by operator projection when choosing the following projection operator:

$$\boldsymbol{P}_{HME}=(\boldsymbol{I}_{M+1},\boldsymbol{0})\in\mathbb{R}^{(M+1)\times\infty}, \tag{3.9}$$

corresponding to a simple cut-off of higher order terms in the equations.

The resulting system of equations (3.7) can be written in the following form:

$$\frac{\partial \boldsymbol{w}_M}{\partial t}+\boldsymbol{A}_{HME}\frac{\partial \boldsymbol{w}_M}{\partial x}=\boldsymbol{S}_M, \tag{3.10}$$

https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press

where $\boldsymbol{A}_{HME}$ differs from $\boldsymbol{A}_{Grad}$ in (3.4) only in two entries in the last equation. Those entries are

$$\boldsymbol{A}_{HME}(M+1,2)=0, \tag{3.11}$$

$$\boldsymbol{A}_{HME}(M+1,3)=\frac{f_{M-3}\theta}{2}, \tag{3.12}$$

whereas all other entries remain unchanged.

The separate application of the projection operator furthermore leads to a different characteristic polynomial of the system matrix $\boldsymbol{A}_{HME}$ which reads

$$\chi_M(\lambda)=\theta^{\frac{M+1}{2}}\mathrm{He}_{M+1}\left(\frac{\lambda-u}{\sqrt{\theta}}\right), \tag{3.13}$$

i.e. exactly the same polynomial as (3.5) but without the dependence on the highest coefficients. The (real) eigenvalues are thus the shifted and scaled roots of the Hermite polynomial He_{M+1}

$$\lambda_j=u+\sqrt{\theta}c_j,\quad j=1,\cdots,M+1, \tag{3.14}$$

where the c_j are now the Hermite roots such that $\mathrm{He}_{M+1}(c_j)=0$ and the c_j do not depend on the higher coefficients any more (compare (3.6)). Obviously, the eigenvalues adaptively adjust according to the current velocity and temperature. This nonlinearity allows to capture flow phenomena accurately as mentioned in Remark 2.1.

The corresponding entries of the eigenvectors v_j can be derived as

$$\begin{aligned}
v_{j,1}&=\rho\mathrm{He}_0(c_j),\\
v_{j,2}&=\theta^{1/2}\mathrm{He}_1(c_j),\\
v_{j,3}&=\theta\mathrm{He}_2(c_j),\\
v_{j,4}&=\frac{1}{3!}\rho\theta^{3/2}\mathrm{He}_3(c_j),\\
v_{j,5}&=\frac{1}{4!}\rho\theta^{4/2}\mathrm{He}_4(c_j)-f_3\sqrt{\theta}c_j,\\
v_{j,6}&=\frac{1}{5!}\rho\theta^{5/2}\mathrm{He}_5(c_j)+f_3\theta\frac{1-c_j^2}{2}-f_4\sqrt{\theta}c_j,\\
&\vdots\\
v_{j,M+1}&=\frac{1}{M!}\rho\theta^{M/2}\mathrm{He}_M(c_j)+f_{M-2}\theta\frac{1-c_j^2}{2}-f_{M-1}\sqrt{\theta}c_j.
\end{aligned}$$

Remark 3.1 (Conservation property of HME)**.** Due to the projection, the last equation of the system is changed in order to gain hyperbolicity of the system. As mass, momentum and energy conservation should not be changed, the system only makes sense for $M>2$. However, since with increasing M more equations are exactly recovered and only equations for higher order coefficients are changed, the use of this method can be justified.

https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press

3.4.2 Quadrature-based moment equations

The so-called quadrature-based moment equations (QBME) were first derived in [11] following a substitution of exact integration by Gauss-Hermite quadrature formulas. In [7] it was shown that the system can easily be formulated within the operator projection framework as well by interpreting the quadrature formulas as a special case of a projection.

The easiest way to derive the QBME system by operator projection is to redefine the matrix $(\boldsymbol{D})_M$ in the general operator projection framework to be

$$(\boldsymbol{D})_M = \boldsymbol{P}_{QBME}\boldsymbol{D}\boldsymbol{P}^T, \tag{3.15}$$

which means that a different matrix is used for the projection of the matrix $\boldsymbol{D}$ whereas all other projections are kept the same as before.

The new projection matrix $\boldsymbol{P}_{QBME}$ is very similar to the previous $\boldsymbol{P}_{HME}$ and only varies in one entry because we have

$$\boldsymbol{P}_{QBME} = \boldsymbol{P}_{HME} - \frac{M+1}{\theta}\boldsymbol{E}_{M-1,M-3}, \tag{3.16}$$

where $\boldsymbol{E}_{i,j}$ is a matrix where only the i,j-entry is set to one and all others entries are set to zero.

Using (3.15) and (3.16), the resulting finite system of equations is the quadrature-based moment system and can be written in the following way similar to the HME system

$$\frac{\partial \boldsymbol{w}_M}{\partial t} + \boldsymbol{A}_{QBME}\frac{\partial \boldsymbol{w}_M}{\partial x} = \boldsymbol{S}_M, \tag{3.17}$$

where $\boldsymbol{A}_{QBME}$ also differs from $\boldsymbol{A}_{Grad}$ only in few entries, here in the last two equations, while all other entries remain unchanged. The affected entries are

$$\boldsymbol{A}_{QBME}(M,3) = \frac{(M-2)f_{M-2}+\theta f_{M-4}}{2} - \frac{M(M+1)}{2}\frac{f_M}{\theta}, \tag{3.18}$$

$$\boldsymbol{A}_{QBME}(M+1,3) = \frac{f_{M-3}\theta}{2}, \tag{3.19}$$

$$\boldsymbol{A}_{QBME}(M+1,4) = -\frac{3f_{M-2}}{\rho} + \frac{3(M+1)f_M}{\rho\theta}. \tag{3.20}$$

Due to the operator projection framework, we can calculate the characteristic polynomial of the system matrix $\boldsymbol{A}_{QBME}$ which is exactly the same as for the HME equations, see (3.13). As before, the eigenvalues are the shifted and scaled roots of the Hermite polynomial He_{M+1}, see (3.14).

A new result is the analytical computation of the eigenvectors of $\boldsymbol{A}_{QBME}$ which are the same as in the case of $\boldsymbol{A}_{HME}$ except for the last entry that is given by

$$v_{j,M+1} = \frac{1}{M!}\rho\theta^{M/2}\mathrm{He}_M(c_j) + f_{M-2}\theta\frac{1-c_j^2}{2} - f_{M-1}\sqrt{\theta}c_j + (M+1)f_M\frac{c_j^2-1}{2}.$$

https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press

Remark 3.2 (Other projection possibilities)**.** The derivation of the QBME system (3.17) can also be done using the same projection operator for every matrix appearing in the general framework according to [7] but this requires rewriting the equation in a more complicated form first. For details we refer to the derivation of the transformed equation in [10].

Remark 3.3 (Conservation property of QBME)**.** Due to the different projection procedure, the last two equations of system (3.17) are changed in order to gain hyperbolicity of the system. As mass, momentum and energy conservation should not be changed, the system only makes sense for $M>3$. Similar to HME, since with increasing M more equations are exactly recovered and only equations for higher order coefficients are changed, the use of this method can be justified.

3.4.3 Change of variables

An interesting point has first been reported in [12] and is related to a change of variables of the respective systems of equations.

Rewriting Grad's system (3.3) in terms of the *convective moments* m_j with

$$m_j = m_j(t,x,c) := \int_{-\infty}^{\infty} f(t,x,c)\, c^j dc \tag{3.21}$$

and choosing $\boldsymbol{m}_M = (m_0,\cdots,m_M)^T \in \mathbb{R}^{M+1}$ results in a new system of equations that describes the same physical behavior but with different variables. Assuming the new variables $\boldsymbol{m}_M$ can be computed from $\boldsymbol{w}_M$ with Jacobian $\frac{\partial \boldsymbol{m}_M}{\partial \boldsymbol{w}_M} = \boldsymbol{B} \in \mathbb{R}^{(M+1)\times(M+1)}$, which is a lower triangular matrix depending on the entries of $\boldsymbol{w}_M$, the change of variables can be performed as follows:

$$\begin{aligned} & \frac{\partial \boldsymbol{w}_M}{\partial t} + \boldsymbol{A}\frac{\partial \boldsymbol{w}_M}{\partial x} = -\boldsymbol{P}\boldsymbol{w}_M \\ \Leftrightarrow \quad & \frac{\partial \boldsymbol{m}_M}{\partial t} + \boldsymbol{B}\boldsymbol{A}\boldsymbol{B}^{-1}\frac{\partial \boldsymbol{m}_M}{\partial x} = -\boldsymbol{P}\boldsymbol{m}_M, \end{aligned}$$

where we used $\boldsymbol{B}\boldsymbol{P}\boldsymbol{w}_M = \boldsymbol{P}\boldsymbol{m}_M$ due to the variable transformation and the special form of the right hand side collision operator $\boldsymbol{P}$ which is a diagonal matrix for the BGK case (2.2).

The resulting system thus yields a new transformed system matrix $\widetilde{\boldsymbol{A}} = \boldsymbol{B}\boldsymbol{A}\boldsymbol{B}^{-1}$ with the same eigenvalues. For Grad's system (3.3), we obtain a so-called companion matrix, see [18]. It has the following structure:

$$\partial_t \boldsymbol{m}_M + \widetilde{A}_{\text{Grad}}\, \partial_x \boldsymbol{m}_M = -\boldsymbol{P}\boldsymbol{m}_M, \quad \widetilde{A}_{\text{Grad}} = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ \widetilde{a}_{M+1} & \cdots & \cdots & \widetilde{a}_1 \end{pmatrix}, \tag{3.22}$$

https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press

with last row coefficients $\widetilde{a}_i$, $i=1,\cdots,M$, depending non-linearly on the variables. However, it is easy to write down the last row at equilibrium, i.e. $\boldsymbol{w}_M^{eq}=(\rho,u,\theta,0,\cdots,0)^T\in\mathbb{R}^{M+1}$ as

$$\widetilde{a}_i^{eq}=-He_i\left(\frac{u}{\sqrt{\theta}}\right)\left(-\sqrt{\theta}\right)^i\prod_{j=1}^{i}\frac{M+2-j}{j}. \tag{3.23}$$

When performing the same variable transformation for the HME system (3.10), the resulting system matrix is exactly Grad's matrix evaluated at equilibrium, i.e. we get

$$\partial_t\boldsymbol{m}_M+\widetilde{\boldsymbol{A}}_{\mathrm{HME}}\,\partial_x\boldsymbol{m}_M=-\boldsymbol{P}\boldsymbol{m}_M,\quad \widetilde{\boldsymbol{A}}_{\mathrm{HME}}(\boldsymbol{m}_M)=\widetilde{\boldsymbol{A}}_{\mathrm{Grad}}\left(\boldsymbol{m}_M^{eq}\right), \tag{3.24}$$

which is particularly interesting as the equilibrium state of Grad's system is obviously inside the hyperbolicity region and immediately proves hyperbolicity for HME. Furthermore, the system matrix of the transformed system does not depend on any of the higher order coefficients.

For the QBME system, the resulting convective system is different. In fact it can be shown that the system is a linear deviation of Grad's system around equilibrium

$$\widetilde{\boldsymbol{A}}_{\mathrm{QBME}}(\boldsymbol{m}_M)=\widetilde{\boldsymbol{A}}_{\mathrm{Grad}}\left(\boldsymbol{m}_M^{eq}\right)+\widetilde{\boldsymbol{M}}\cdot f_M, \tag{3.25}$$

where $\widetilde{\boldsymbol{M}}\in\mathbb{R}^{(M+1)\times(M+1)}$ is the deviation matrix depending only on equilibrium values ρ,u,θ and f_M is the highest order expansion coefficient (see also [12]).

Because of the linear deviation from Grad's equations in equilibrium compared to the direct evaluation of Grad's matrix at equilibrium for the HME case, we expect the QBME model to give different results than HME while keeping global hyperbolicity.

4 Non-conservative numerics

Despite all beneficial properties of the new model equations HME and QBME, both hyperbolic models are modifications of the original Grad system and can no longer guarantee conservation form of all equations involved. In fact, the systems are in partially conservative form meaning that only a subset of the equations can be written in conservative form. We want to describe numerical schemes that allow for an accurate solution while keeping the conservation property for the conserved quantities.

We first neglect the right-hand collision operator to focus on the left-hand side of the equation. A conservative PDE system of $n\in\mathbb{N}$ equations can be written as

$$\frac{\partial\boldsymbol{u}}{\partial t}+\frac{\partial\boldsymbol{F}(\boldsymbol{u})}{\partial x}=\boldsymbol{0}, \tag{4.1}$$

with *flux function* $F:\mathbb{R}^n\to\mathbb{R}^n$ that has a Jacobian $\frac{\partial\boldsymbol{F}(\boldsymbol{u})}{\partial\boldsymbol{u}}=\boldsymbol{A}(\boldsymbol{u})\in\mathbb{R}^{n\times n}$. The Jacobian allows to write the system in non-conservative form as

$$\frac{\partial\boldsymbol{u}}{\partial t}+\boldsymbol{A}(\boldsymbol{u})\frac{\partial\boldsymbol{u}}{\partial x}=\boldsymbol{0}. \tag{4.2}$$

https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press

For conservative systems like (4.1) it can be shown that besides boundary effects the integral of the variables over a domain remains constant over time, i.e.

$$\frac{\partial}{\partial t}\int \boldsymbol{u}(x,t)dt = \mathbf{0}. \tag{4.3}$$

It is always possible to write a conservative system in non-conservative form. The other way is only possible if there exists a flux function $\boldsymbol{F}$ such that the system matrix is the Jacobian of this flux function.

For Grad's system (2.12), there exists a flux function and the system can be written in conservative form, see e.g. [18].

However, as the hyperbolic equations include changes in the last equations it can be shown that it is impossible to derive a general flux function for the new models. Instead, we can only derive flux functions for the first unchanged equations, which are equal to the components of Grad's flux function. The respective variables are still conserved and only the variables corresponding to the last equations are not conserved. We call this a *partially conservative* PDE system. As stated in Remarks 3.1 and 3.3, it is important to always have the physical quantities mass, momentum and energy conserved which requires to take into account a minimum number of equations for both the HME and the QBME case.

In case of HME, only the last equation is changed, see (3.11), and therefore not in conservative form. For QBME on the other hand, the two last equations are changed, see (3.18). So QBME leads to two equations in non-conservative form. Both are examples of partially conservative systems.

It is important to note that one or two equations in non-conservative form do not necessarily undermine the physical relevance of the whole equation system. Depending on the total number of equations, only less important moments of higher order coefficients in the expansion are not conserved exactly. One can always simply take more equations into account to allow for better conservation properties of the equation systems. In cases including a right-hand side collision operator, however, higher order moments are not conserved anyway as briefly explained in Section 4.6.

In the following section we describe different numerical solution schemes to solve the partially conservative PDE systems and especially focus on the difficulties related to the missing flux formulation.

4.1 Generalized Roe matrix for non-conservative systems

A standard conservative system like (4.1) can be solved using a finite volume scheme of the form

$$\boldsymbol{u}_i^{n+1} = \boldsymbol{u}_i^n - \frac{\Delta t}{\Delta x}\left(\boldsymbol{F}^*_{i+\frac{1}{2}} - \boldsymbol{F}^*_{i-\frac{1}{2}}\right), \tag{4.4}$$

where the numerical flux function $\boldsymbol{F}^*_{i+\frac{1}{2}}$ is evaluated at the respective cell boundary. The conservation property can be directly carried over from the analytical case, such that

https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press

the integral over the numerical solution $\boldsymbol{u}^n$ remains constant. Schemes of this type are also relatively easy to extend to 2D and unstructured grids as the fluxes can be directly evaluated at cell boundaries which are faces in 2D. Examples for popular numerical flux functions in finite volume schemes are the well-known Lax-Friedrichs or Lax-Wendroff flux functions, see e.g. [14].

In case of a non-conservative system according to (4.2), there may not exist any flux function and if we desire to use a similar formulation as in the standard finite volume case (4.4), we need to find a clever way to evaluate the system matrix at the cell boundaries. This should be done in a way such that the scheme recovers conservation properties if the matrix $\boldsymbol{A}(\boldsymbol{u})$ is a flux Jacobian. In order for this to hold, standard methods use a Roe matrix $\boldsymbol{A}_{\text{Roe}}$ to generalize the notation as follows:

$$\boldsymbol{A}_{Roe}(\boldsymbol{u}_L,\boldsymbol{u}_R)(\boldsymbol{u}_R-\boldsymbol{u}_L)=\boldsymbol{F}(\boldsymbol{u}_R)-\boldsymbol{F}(\boldsymbol{u}_L), \tag{4.5}$$

where $\boldsymbol{u}_L$ and $\boldsymbol{u}_R$ are the states at the left or right boundary of the corresponding cells and $\boldsymbol{F}$ is the flux function for $\frac{\partial \boldsymbol{F}(\boldsymbol{u})}{\partial \boldsymbol{u}}=\boldsymbol{A}(\boldsymbol{u})$. However, the construction of an analytical Roe matrix $\boldsymbol{A}_{\text{Roe}}$ can be difficult if not impossible for certain physical systems as reported also in [14].

In the non-conservative case, the notation of a Roe matrix is extended to a generalized Roe matrix $\boldsymbol{A}_{\boldsymbol{\psi}}$ that depends on the left and right states as well as on a connecting path $\boldsymbol{\psi}$ between both states (see [16] for more details about the definition of a proper path). For now, we can think of the simple linear path $\boldsymbol{\psi}(s,\boldsymbol{u}_L,\boldsymbol{u}_R)=\boldsymbol{u}_L+s\cdot(\boldsymbol{u}_R-\boldsymbol{u}_L)$, $s\in[0,1]$, connecting both states linearly. The generalized Roe matrix is then given according to [16] as

$$\boldsymbol{A}_{\boldsymbol{\psi}}(\boldsymbol{u}_L,\boldsymbol{u}_R)(\boldsymbol{u}_R-\boldsymbol{u}_L)=\int_0^1 \boldsymbol{A}(\boldsymbol{\psi}(s,\boldsymbol{u}_L,\boldsymbol{u}_R))\frac{\partial \boldsymbol{\psi}}{\partial s}ds. \tag{4.6}$$

The computation of the generalized Roe matrix in (4.6) is obviously not simple at all as the analytical integration on the right hand side may be very difficult or fail. Following [6], we will exemplify the computation using the linear path from above. Due to the simple choice of $\boldsymbol{\psi}$ and the derivative $\frac{\partial \boldsymbol{\psi}}{\partial s}=\boldsymbol{u}_R-\boldsymbol{u}_L$ the definition of the generalized Roe matrix simplifies to

$$\boldsymbol{A}_{\boldsymbol{\psi}}(\boldsymbol{u}_L,\boldsymbol{u}_R)=\int_0^1 \boldsymbol{A}(\boldsymbol{\psi}(s,\boldsymbol{u}_L,\boldsymbol{u}_R))ds. \tag{4.7}$$

A simple suggestion was made in [6] to substitute the exact integration by a quadrature rule of sufficient accuracy. The generalized Roe matrix is thus computed numerically using

$$\boldsymbol{A}_{\boldsymbol{\psi}}(\boldsymbol{u}_L,\boldsymbol{u}_R)\approx\sum_{j=1}^{P}\omega_j\boldsymbol{A}\left(\boldsymbol{\psi}\left(s_j,\boldsymbol{u}_L,\boldsymbol{u}_R\right)\right), \tag{4.8}$$

for a quadrature formula with $P\in\mathbb{N}$ quadrature points $s_j\in[0,1]$ and quadrature weights $\omega_j\in\mathbb{R}$, $j=1,\cdots,P$. The application of a quadrature formula heavily simplifies the computation and still leads to good results, as demonstrated in [6]. However, it is important

https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press

to use enough quadrature points such that the approximation error does not spoil the accuracy of the whole method. It is advised to use Gauss-Legendre quadrature rules that result in a relatively high accuracy together with relatively few quadrature points and thus few potentially costly evaluations of the system matrix per computation.

The linear path is obviously not the only option for a path connecting both states. Other possibilities include polynomial paths $\boldsymbol{\psi}_-^N$ and $\boldsymbol{\psi}_+^N$

$$\boldsymbol{\psi}_-^N(s,\boldsymbol{u}_L,\boldsymbol{u}_R)=\boldsymbol{u}_L+s^N\cdot(\boldsymbol{u}_R-\boldsymbol{u}_L), \tag{4.9}$$

$$\boldsymbol{\psi}_+^N(s,\boldsymbol{u}_L,\boldsymbol{u}_R)=\boldsymbol{u}_R+(s-1)^N\cdot(\boldsymbol{u}_L-\boldsymbol{u}_R), \tag{4.10}$$

where the path variable accelerates along the path way.

Theoretical results (see e.g. [16]) show that the choice of the path can be very important and a precise numerical integration for the generalized Roe matrix is crucial for an accurate scheme. As reported by many other authors, however, the influence of the choice of the path on the solution vanishes in many application cases provided the quadrature is chosen accurately enough (see [17] for more details). This is a result we also obtained in our first simulations and the reason why we conducted all our subsequent tests using the most simple linear path $\boldsymbol{\psi}_-^1$. Other paths require more quadrature points as they depend on the path variable s in a non-linear way. Too few quadrature points can then lead to a loss of the conservation property due to wrong propagation speeds especially for strong shocks.

As the non-conservative products added to the hyperbolic moment models (see (3.11) and (3.18)) depend at most only linearly on the higher order coefficients f_M, f_{M-1} representing the non-conservative equations, only $P=3$ quadrature points of a Gauss-Legendre quadrature rule are sufficient to guarantee accuracy of the scheme and further refinement does not lead to improved results. In the results in Section 5 we omit numerical results for this to focus on the comparison of the different numerical schemes. All of the simulations there are using the linear path with three Gauss-Legendre quadrature points.

4.2 Castro scheme

One way to use the definition of the generalized Roe matrix in (4.8) is the first order upwind type scheme by Castro [16]

$$\boldsymbol{u}_i^{n+1}=\boldsymbol{u}_i^n-\frac{\Delta t}{\Delta x}\left(\boldsymbol{A}^-_{i+\frac{1}{2}}\left(\boldsymbol{u}_{i+1}^n-\boldsymbol{u}_i^n\right)+\boldsymbol{A}^+_{i-\frac{1}{2}}\left(\boldsymbol{u}_i^n-\boldsymbol{u}_{i-1}^n\right)\right), \tag{4.11}$$

where the left and right going contributions $\boldsymbol{A}^-_{i+\frac{1}{2}}$ and $\boldsymbol{A}^+_{i-\frac{1}{2}}$ are computed via an eigenvalue decomposition of the generalized Roe matrix as follows

$$\boldsymbol{A}^\pm_{i+\frac{1}{2}}=\boldsymbol{A}_{\boldsymbol{\psi}}(\boldsymbol{u}_i,\boldsymbol{u}_{i+1})^\pm=\boldsymbol{R}_{\boldsymbol{\psi}}\cdot\boldsymbol{\Lambda}^\pm_{\boldsymbol{\psi}}\cdot\boldsymbol{R}^{-1}_{\boldsymbol{\psi}}. \tag{4.12}$$

https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press

The scheme is of upwind type as it uses eigenvalue information of the generalized Roe matrix. It is therefore expected to accurately capture flow properties and show less diffusion than central schemes. On the other hand, the eigenvalue and eigenvector computations require more runtime. Note that it is not possible to use the analytical eigenvalues of the system matrix as the generalized Roe matrix is a linear combination of evaluations of this system matrix at values along the specific path chosen due to the quadrature rule. The only exception is the case of $P=1$ quadrature point, but this might not be accurate enough.

A benefit of the method is that it can readily be implemented for 2D unstructured grids using straightforward modifications.

4.3 PRICE-C scheme

A different family of schemes follows a centered discretization, i.e. without the use of eigenvalue and eigenvector information. For the conservative case a well-known centered method is the FORCE method (first order centered). The scheme can be written in terms of the fluxes similar to (4.4)

$$\boldsymbol{u}_i^{n+1} = \boldsymbol{u}_i^n - \frac{\Delta t}{\Delta x}\left(\boldsymbol{F}_{i+\frac{1}{2}}^{\mathrm{FORCE}} - \boldsymbol{F}_{i-\frac{1}{2}}^{\mathrm{FORCE}}\right) \tag{4.13}$$

with force flux $\boldsymbol{F}_{i+\frac{1}{2}}^{\mathrm{FORCE}}$. We refer to [21] for more detailed information about the definition and derivation of the FORCE scheme. It is known that the FORCE scheme is the optimal centered scheme that is a convex average of the Lax-Friedrichs and the Lax-Wendroff scheme and it is the least dissipative of all three-point centered methods that are monotone and stable under a suitable CFL condition (see [6]). We would therefore like to recover the properties of the FORCE scheme in case our system can be written in conservative form.

The PRICE-C scheme that was published by Canestrelli et al. in [6] has been derived based upon the FORCE scheme and the recent version of the scheme also reduces to the FORCE scheme in the conservative non-linear case. The first order version of the scheme is similar to the Castro scheme above and can be written as

$$\boldsymbol{u}_i^{n+1} = \boldsymbol{u}_i^n - \frac{\Delta t}{\Delta x}\left(\boldsymbol{A}_{i+\frac{1}{2}}^{-}\left(\boldsymbol{u}_{i+1}^n - \boldsymbol{u}_i^n\right) + \boldsymbol{A}_{i-\frac{1}{2}}^{+}\left(\boldsymbol{u}_i^n - \boldsymbol{u}_{i-1}^n\right)\right), \tag{4.14}$$

where the matrices $\boldsymbol{A}_{i+\frac{1}{2}}^{-}$ and $\boldsymbol{A}_{i-\frac{1}{2}}^{+}$ are now computed according to the following formulas

$$\boldsymbol{A}_{i+\frac{1}{2}}^{-} = \frac{1}{4}\left(2\boldsymbol{A}_{\psi}\left(\boldsymbol{u}_i^n, \boldsymbol{u}_{i+1}^n\right) - \frac{\Delta x}{\Delta t}\boldsymbol{I} - \frac{\Delta t}{\Delta x}\left(\boldsymbol{A}_{\psi}\left(\boldsymbol{u}_i^n, \boldsymbol{u}_{i+1}^n\right)\right)^2\right), \tag{4.15}$$

$$\boldsymbol{A}_{i-\frac{1}{2}}^{+} = \frac{1}{4}\left(2\boldsymbol{A}_{\psi}\left(\boldsymbol{u}_{i-1}^n, \boldsymbol{u}_i^n\right) - \frac{\Delta x}{\Delta t}\boldsymbol{I} - \frac{\Delta t}{\Delta x}\left(\boldsymbol{A}_{\psi}\left(\boldsymbol{u}_{i-1}^n, \boldsymbol{u}_i^n\right)\right)^2\right). \tag{4.16}$$

https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press

The definitions of the matrices include a diffusion term $\frac{\Delta x}{\Delta t}\boldsymbol{I}$ that stabilizes the method and the respective generalized Roe matrix occurs in both formulas. It is important to note that each matrix is only a function of the two neighbouring values $\boldsymbol{u}_i^n, \boldsymbol{u}_{i+1}^n$ or $\boldsymbol{u}_{i-1}^n, \boldsymbol{u}_i^n$, respectively. Other versions of the scheme like PRICE-R do not have this property which makes generalization to 2D or higher order versions much more difficult, see e.g. [5].

In contrast to the Castro scheme which uses an eigenvalue decomposition (4.12), the PRICE-C scheme requires less computations per cell due to the central origin of the scheme. However, the added diffusion will definitely smooth out shock structures of the hyperbolic system.

A 2D extension of the scheme is again straightforward and can be done analogously to any standard finite volume scheme.

4.4 UPRICE scheme

The centered scheme PRICE-C can be augmented with eigenvalue information to obtain the UPRICE-Cδ method by Stecca, see [19]. The computation of the matrices then reads

$$\boldsymbol{A}^-_{i+\frac{1}{2}} = \frac{1}{2}\left(\boldsymbol{A}_{\boldsymbol{\psi}}\left(\boldsymbol{u}_i^n, \boldsymbol{u}_{i+1}^n\right) - \frac{|s_1^-||s_N^-|}{|s_1^-|+|s_N^-|}\boldsymbol{I} - \frac{1}{|s_1^-|+|s_N^-|}\left(\boldsymbol{A}_{\boldsymbol{\psi}}\left(\boldsymbol{u}_i^n, \boldsymbol{u}_{i+1}^n\right)\right)^2\right), \tag{4.17}$$

$$\boldsymbol{A}^+_{i-\frac{1}{2}} = \frac{1}{2}\left(\boldsymbol{A}_{\boldsymbol{\psi}}\left(\boldsymbol{u}_{i-1}^n, \boldsymbol{u}_i^n\right) - \frac{|s_1^+||s_N^+|}{|s_1^+|+|s_N^+|}\boldsymbol{I} - \frac{1}{|s_1^+|+|s_N^+|}\left(\boldsymbol{A}_{\boldsymbol{\psi}}\left(\boldsymbol{u}_{i-1}^n, \boldsymbol{u}_i^n\right)\right)^2\right), \tag{4.18}$$

where $s_N^\pm$ and $s_1^\pm$ are the fastest and slowest travelling wave speeds at the interfaces of the respective cells, see [19] for details. The computation of the wave speeds is relatively cheap as only eigenvalues are needed and the eigenvalues of the system matrix at the neighbouring cells corresponding to $\boldsymbol{u}_{i-1}^n$, $\boldsymbol{u}_i^n$ and $\boldsymbol{u}_{i+1}^n$ can all be obtained analytically for the models HME and QBME using the given formula (3.14). The resulting numerical method is expected to give less spurious results than PRICE-C and reduce diffusion due to the additional eigenvalue information.

Remark 4.1 (Higher order extension)**.** All three schemes presented in this sections so far are of first order in space and time. Due to the definition of the schemes it is possible to extend them to higher order. This first requires a reconstruction in space, for example using a WENO scheme and afterwards the *arbitrary order* (ADER, see [5]) approach allows to expand the unknowns in time as well and obtain a space-time reconstruction of the following form:

$$\boldsymbol{u}_i(x,t) = \boldsymbol{u}(x_i,t^n) + (x-x_i)\frac{\partial \boldsymbol{u}}{\partial x} + (t-t^n)\frac{\partial \boldsymbol{u}}{\partial t} + \cdots, \tag{4.19}$$

where the spatial derivative is computed from the reconstruction in space and the unknown time (and possible higher order or mixed) derivatives are computed using the original PDE (4.2), for example

$$\frac{\partial \boldsymbol{u}}{\partial t} = -\boldsymbol{A}(\boldsymbol{u})\frac{\partial \boldsymbol{u}}{\partial x}. \tag{4.20}$$

https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press

With the help of the space-time reconstruction, the integration of the PDE over a space-time volume and computation of the occurring integrals using appropriate Gaussian quadrature rules then leads to a higher order scheme of arbitrary order. The procedure is straightforward but computationally very expensive especially in 2D as more mixed derivatives are present in the higher order expansion (4.19). In this work we restrict ourselves to first order versions of the aforementioned schemes.

4.5 Wave propagation scheme

The wave propagation scheme by LeVeque [13] can also be used to solve non-conservative systems. However, it is not derived using the generalized Roe matrix and uses a special averaging of the adjacent cell values to evaluate the system matrix at the cell boundary. In simple cases this is equivalent to a generalized Roe matrix using only one quadrature point for the approximate integration. The Roe matrix is then given as

$$\boldsymbol{A}_{\boldsymbol{\psi}}(\boldsymbol{u}_L,\boldsymbol{u}_R)=\boldsymbol{A}\left(\frac{1}{2}(\boldsymbol{u}_L+\boldsymbol{u}_R)\right). \tag{4.21}$$

Nevertheless, the scheme can also be extended in a straightforward way by substituting the formula of the generalized Roe matrix. In our computations, we used this extension to allow for a better comparison with the other schemes.

Following the original notation the scheme can be written as follows:

$$\boldsymbol{u}_i^{n+1}=\boldsymbol{u}_i^n-\frac{\Delta t}{\Delta x}\left(\boldsymbol{A}^+\boldsymbol{\Delta u}_i+\boldsymbol{A}^-\boldsymbol{\Delta u}_{i+1}\right)-\frac{\Delta t}{\Delta x}\left(\widetilde{\boldsymbol{F}}_{i+1}-\widetilde{\boldsymbol{F}}_i\right), \tag{4.22}$$

with fluctuations $\boldsymbol{A}^{\pm}\boldsymbol{\Delta u}_i$ that are computed just like the corresponding terms in the Castro scheme (4.12) and second order corrections

$$\widetilde{\boldsymbol{F}}_i=\frac{1}{2}\sum_p|\lambda_i^p|\left(1-\frac{\Delta t}{\Delta x}|\lambda_i^p|\right)\widetilde{\boldsymbol{W}}_i^p \tag{4.23}$$

that use eigenvalues λ_i^p and limited versions of the waves $\boldsymbol{W}_i^p$ that were obtained by the eigenvalue decomposition of the system matrix. For more information we refer to [13]. Without the second order correction terms, the scheme gives exactly the same results as the Castro scheme.

The wave propagation scheme is almost second order as some of the leading second order error terms are removed, but formally only first order can be shown. It is an upwind scheme similar to the Castro scheme (4.11) that is very accurate and resolves shock structures particularly well. On the other hand, it is not extendable to higher order in this form, as the second order corrections are specially tailored to cancel the leading error terms. Typically, the wave propagation scheme is only applicable for equidistant and structured grids. Due to these limitations we will only consider the scheme as a reference for some of the other methods which we also plan to extend to 2D unstructured grids in subsequent work.

https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press

4.6 Source term discretization

Taking into account a right-hand side collision operator, the partially conservative PDE system has the following form:

$$\frac{\partial \boldsymbol{u}}{\partial t}+\boldsymbol{A}(\boldsymbol{u})\frac{\partial \boldsymbol{u}}{\partial x}=\boldsymbol{S}(\boldsymbol{u}) \tag{4.24}$$

with a collision operator $\boldsymbol{S}\in\mathbb{R}^n$ acting only on the higher order equations, because we have $S_j=0$ for $j=0,1,2$. Consequently, the collision term on the right hand side is zero for the first three equations resulting in conservation of mass, momentum and energy. For all other equations we obtain a relaxation term using the BGK model (2.2), as exemplified in the test case in Section 5.1 and in equation (5.3). As all our numerical methods will be first order in time, we do not apply a splitting scheme but simply evaluate the source term at the center of the grid cell and add the term to the residual on the right hand side of the update formula for the next time step.

In the partially convective system only the last one or two equations cannot be written in conservative form. However, the right hand side collision operator for those equations yields a relaxation of the corresponding moments anyway, such that even in the case of a conservative system (e.g. for Grad's model) the quantities would not be conserved as a consequence of the collision operator for finite Knudsen number.

5 Results

In this section we compare different numerical schemes for the solution of the partially conservative PDE systems derived in Section 3. We first consider the accuracy of the schemes, the different sets of variables and then compare the model equations themselves before performing a convergence study with respect to increasing M.

5.1 Shock tube test case

In order to test the equation systems (3.3), (3.10) and (3.17), we consider the following non-dimensionalized shock tube problem (see [4]):

$$\partial_t \boldsymbol{u}_M+\boldsymbol{A}\partial_x \boldsymbol{u}_M=-\frac{1}{\tau}\boldsymbol{P}\boldsymbol{u}_M, \tag{5.1}$$

$$\boldsymbol{u}_M(0,x)=\begin{cases}\boldsymbol{u}_M^L & \text{if } x<0,\\ \boldsymbol{u}_M^R & \text{if } x>0,\end{cases} \tag{5.2}$$

where the collisions are modelled using a BGK operator with non-linear relaxation time $\tau=\frac{\mathrm{Kn}}{\rho}$ that leads to the following form of the matrix $\boldsymbol{P}$ after the projections:

$$\boldsymbol{P}=\mathrm{diag}\,(0,0,0,1,\cdots,1)\in\mathbb{R}^{(M+1)\times(M+1)} \tag{5.3}$$

https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press

for variable vector $\boldsymbol{u}_M = (\rho,u,\theta,f_3,\cdots,f_M)$, $M \geq 4$. We will consider different Knudsen numbers $\mathrm{Kn}_1 = 0.05$ representing a small Knudsen number, $\mathrm{Kn}_2 = 0.5$ for a relatively large Knudsen number and $\mathrm{Kn}_3 = \infty$ leading to vanishing right hand side and strongest shocks.

According to the tests by Cai et al. (see [4]), the left and right states are chosen as

$$\boldsymbol{u}_M^L = (7,0,1,0,\cdots,0)^T, \qquad \boldsymbol{u}_M^R = (1,0,1,0,\cdots,0)^T, \tag{5.4}$$

and for a stronger shock test case we consider the following initial data:

$$\boldsymbol{u}_M^L = (30,0,1,0,\cdots,0)^T, \qquad \boldsymbol{u}_M^R = (1,0,1,0,\cdots,0)^T, \tag{5.5}$$

both corresponding to a jump in density at the discontinuity at $x = 0$. We therefore expect a rarefaction wave travelling to the left, a contact discontinuity travelling to the right and a faster shock wave also travelling to the right in front of the contact discontinuity.

The results show part of the computational domain $[-2,2]$ discretized using 4000 cells at end time $t_{\mathrm{END}} = 0.3$ using a constant $\Delta t = 0.0001$ which corresponds to a CFL number of approximately 0.45. As all schemes are formally first order, no limiter is used except for the wave propagation scheme, where the waves are limited using a minmod limiter to cancel out higher order error terms in smooth areas.

5.2 Variables

The numerical solution of a non-conservative system depends on the set of chosen variables. Following the explanation in Section 3.4.3 we consider three different choices

(1) primitive variables: $\boldsymbol{w}_M = (\rho,u,\theta,f_3,f_4,\cdots,f_M)$,

(2) partially conservative variables: $\widetilde{\boldsymbol{w}}_M = (\rho,\rho u,\rho(u^2+\theta),f_3,f_4,\cdots,f_M)$,

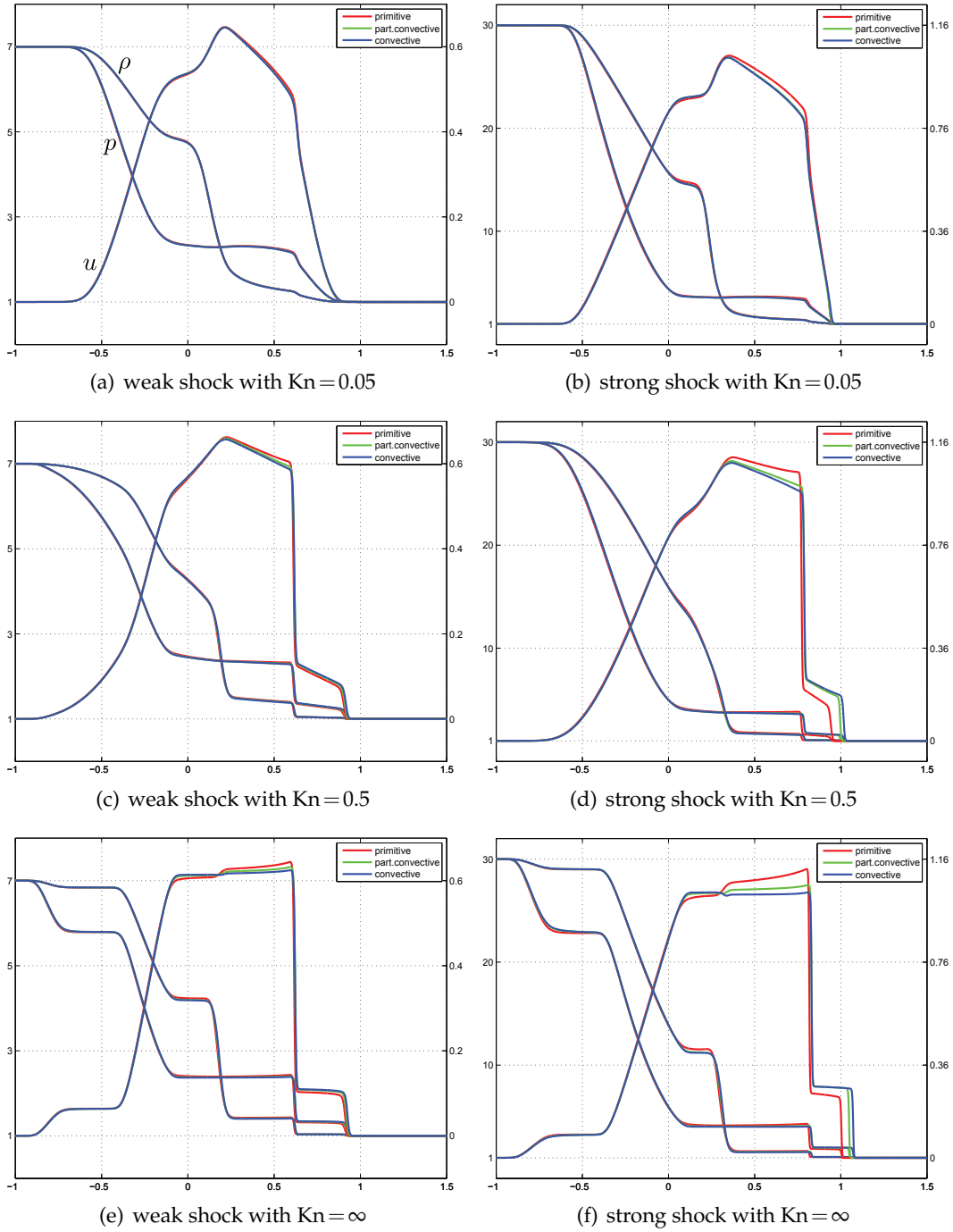
(3) conservative variables: $\boldsymbol{m}_M = (m_0,\cdots,m_M)$,

where the use of conservative variables yields exact conservation for all equations in conservation form including mass, momentum and energy conservation. Partially conservative variables only result in the conservation of the macroscopic quantities mass, momentum and energy. Primitive variables are simple, but have no conservation properties at all, except for the first variable ρ which coincides with the first conservative variable and for which the integral is obviously conserved.

In Fig. 1, the results for the three different sets of variables are shown for all combinations of weak shocks (5.4) and strong shocks (5.5) as well as the three Knudsen numbers $\mathrm{Kn}_1 = 0.05$, $\mathrm{Kn}_2 = 0.5$ and $\mathrm{Kn}_3 = \infty$. In order to focus on the variable sets themselves, we used the QBME model and the PRICE scheme in all tests of Fig. 1.

Starting from small Knudsen numbers and a weak shock in Fig. 1(a) we observe almost no difference between the three variable sets. However, differences become more obvious with increasing shock strength and Knudsen number. The results for the primitive variables first differ from the other two for the velocity variable u directly in front of

https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press



(a) weak shock with Kn = 0.05 (b) strong shock with Kn = 0.05

(c) weak shock with Kn = 0.5 (d) strong shock with Kn = 0.5

(e) weak shock with Kn = ∞ (f) strong shock with Kn = ∞

Figure 1: Comparison of variable sets for primitive, partially conservative and conservative variables obtained with PRICE scheme and QBME moment model. The left y-axis is for ρ and p, the right y-axis is for u.

https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press

the contact discontinuity. Further increase in the Knudsen number also reveals some deviations in the density ρ at the plateau between shock discontinuity and rarefaction wave. The extreme cases of strong shocks with moderate and large Knudsen numbers $\mathrm{Kn}_2 = 0.5$ and $\mathrm{Kn}_3 = \infty$ show a difference in propagation speeds of the subshock travelling right at $x \approx 1$, mainly caused by wrong velocity values behind the shock wave.

In the limit of very small Knudsen numbers, the Euler equations are recovered and it is known that primitive variables will not result in accurate shock propagation speeds. We also observe this in tests with smaller Knudsen numbers $\mathrm{Kn} \ll 0.01$. Especially the velocity maximum is different for primitive and conservative variables. However, there is no difference in the results for conservative and partially conservative variables, as all higher-order coefficients tend to zero in the limit of small Knudsen number.

On the other hand, the conservative variables can obtain large values as they have a higher order of magnitude, even more for larger systems. This can lead to numerical problems appearing especially for very large Knudsen numbers. Note that the values of the coefficients f_α decrease for larger α as opposed to the values of the conservative variables m_α which increase. This is one reason why we do not use conservative variables for tests including more equations. For all our test cases, we thus do not observe any numerical problems related to this issue.

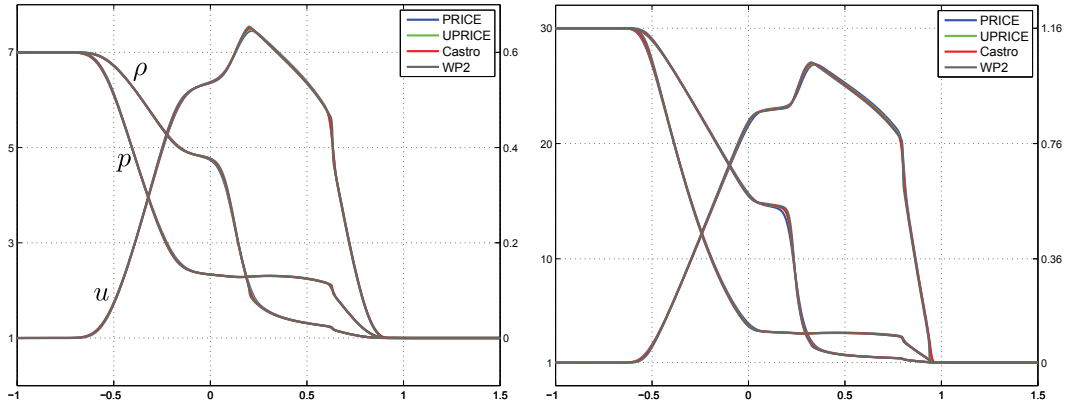
As all three choices of variables can be seen as part of the model choice, it is not clear which one is the best. Intuitively, we trust the convective variables as they are the physically most relevant. One example is the convective variable *energy* which should be conserved and is physically relevant. Furthermore, Grad's system written in convective variables results a balance law form with well defined fluxes. The primitive variables on the other hand are somewhat artificial in that sense and also give results that differ from the other two quite a bit in the regions mentioned before. The first test case shows that it is in fact important to use the right set of variables. The partially conservative variables nevertheless give relatively good results compared to the primitive variables and only slightly differ from the conservative variables in the extreme cases. However, for $\mathrm{Kn}_1 = 0.05$ and for weak shocks with $\mathrm{Kn} < \infty$, primitive variables are sufficient.

Obviously, partially conservative and conservative variables need more runtime for certain schemes, as the eigenvectors of the generalized Roe matrix are not given explicitly. Also the transformation of the system of equations needs to be performed. However, in case of conservative variables an efficient way is to directly use the explicit form of the system matrix described in Section 3.4.3 for HME and QBME.

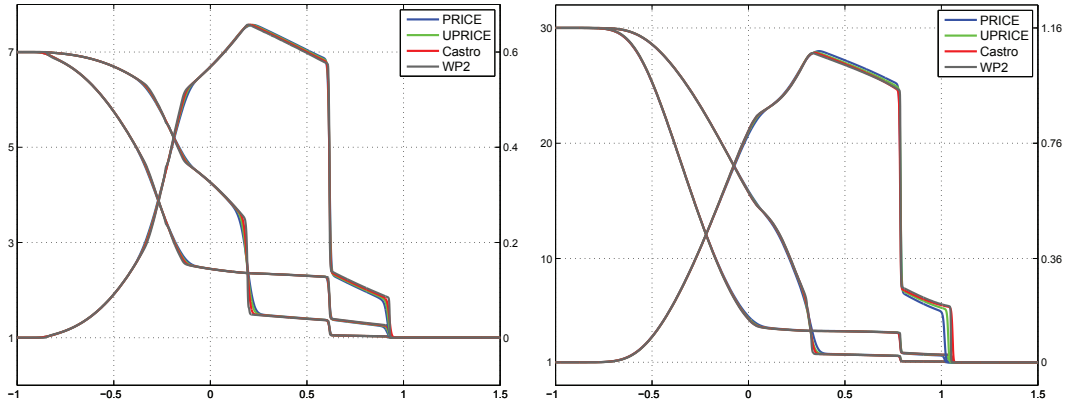
5.3 Non-conservative scheme

We now want to consider the different numerical schemes for the solution of the partially conservative system as mentioned in Section 4. We will compare the centered PRICE scheme [6], the augmented UPRICE scheme [19], the upwind type scheme by Castro [16] as well as the almost second order wave propagation scheme by LeVeque [13]. Fig. 2 shows the results again for all combinations of weak and strong shocks as well as the

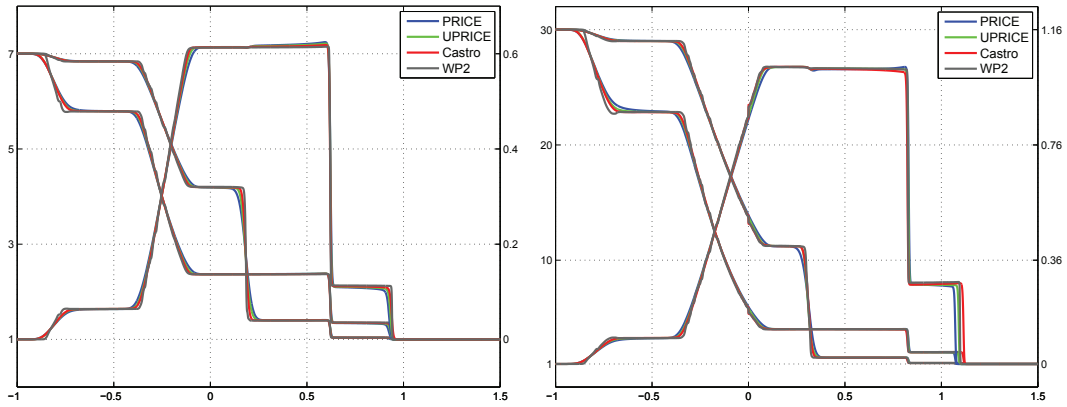
https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press



(a) weak shock with Kn = 0.05 (b) strong shock with Kn = 0.05



(c) weak shock with Kn = 0.5 (d) strong shock with Kn = 0.5



(e) weak shock with Kn = ∞ (f) strong shock with Kn = ∞

Figure 2: Numerical scheme comparison for PRICE, UPRICE, Castro and wave propagation scheme, all in conservative variables and for QBME moment model. The left y-axis is for ρ and p, the right y-axis is for u.

https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press

three Knudsen numbers.

For the small Knudsen number $\mathrm{Kn}_1 = 0.05$, there is almost no difference between the solutions, as the right hand side collision operator smoothes out the shock structure and the discretization of the transport term has less influence on the solution. However, for the larger Knudsen number $\mathrm{Kn}_2 = 0.5$, slight differences start to appear. First of all, we obtain different diffusive behavior close to the shocks, best visible at the contact discontinuity in the middle. We clearly observe that the PRICE scheme shows most diffusivity followed by UPRICE, Castro and the wave propagation scheme. This is an expected result as the central discretization together with the diffusion term in the PRICE scheme, especially in (4.15), adds a lot of diffusion. On the other hand, UPRICE uses eigenvalue information and Castro's scheme is a pure upwind scheme that naturally results in less diffusion and sharper profiles. The higher order scheme by LeVeque provides the most accurate solution close to the shocks as not only the diffusion is small but also most of the second order error terms are removed.

In front of the shock wave, we see a smaller subshock which is positioned slightly different for the four methods. This is because of small differences in the velocity prior to the shock that are again caused by diffusivity. Again, the wave propagation scheme results in the sharpest shock profile, also visible for the strong shock. However, all schemes still give good approximations and the other shock propagation speeds seem to coincide almost perfectly.

Using $\mathrm{Kn}_3 = \infty$, we obtain Figs. 2(e) and 2(f). Here we see an amplification of the differences already observed at smaller Knudsen numbers. The wave propagation scheme resolves the shocks very well and the other schemes follow as described above. Still the propagation speeds are very similar, only for the strong shock case in front of the shock wave we see some differences. Furthermore, we see small oscillations appearing for the wave propagation scheme. We attribute this to the limiting process in the presence of the strong shocks. A closer look at the higher order moments reveals that these oscillations only occur where the higher order moments have strong shocks.

It is also important to note that we see no differences in the values at the plateaus between the shocks, compared to the different variable sets in Fig. 1.

In order to see whether the different shock positions in Fig. 2(d) and Fig. 2(f) are due to different artificial viscosity of the numerical scheme or numerical accuracy, we repeated test case 2(d) with different numerical grids and constant CFL number. The results are shown in Fig. 3. Starting from a very coarse grid in Fig. 3(a), we see relatively large differences between the four numerical schemes. These differences can be observe mainly as different viscosity close to the shocks, especially in the vicinity of the fastest shock wave. In addition, the maximum velocity values are not the same. As the grid is refined, the differences decrease quickly. The left part of the domain $x < 0$ is already converged for the grid with $\Delta x = 0.001$ as can be seen in Fig. 3(c) and the right part of the domain $x > 0$ with smaller density shows only little difference between the numerical schemes as shown in Fig. 3(d). We thus conclude that all numerical methods apparently seem to converge to one solution in this test case. The impact of the non-conservative

https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press

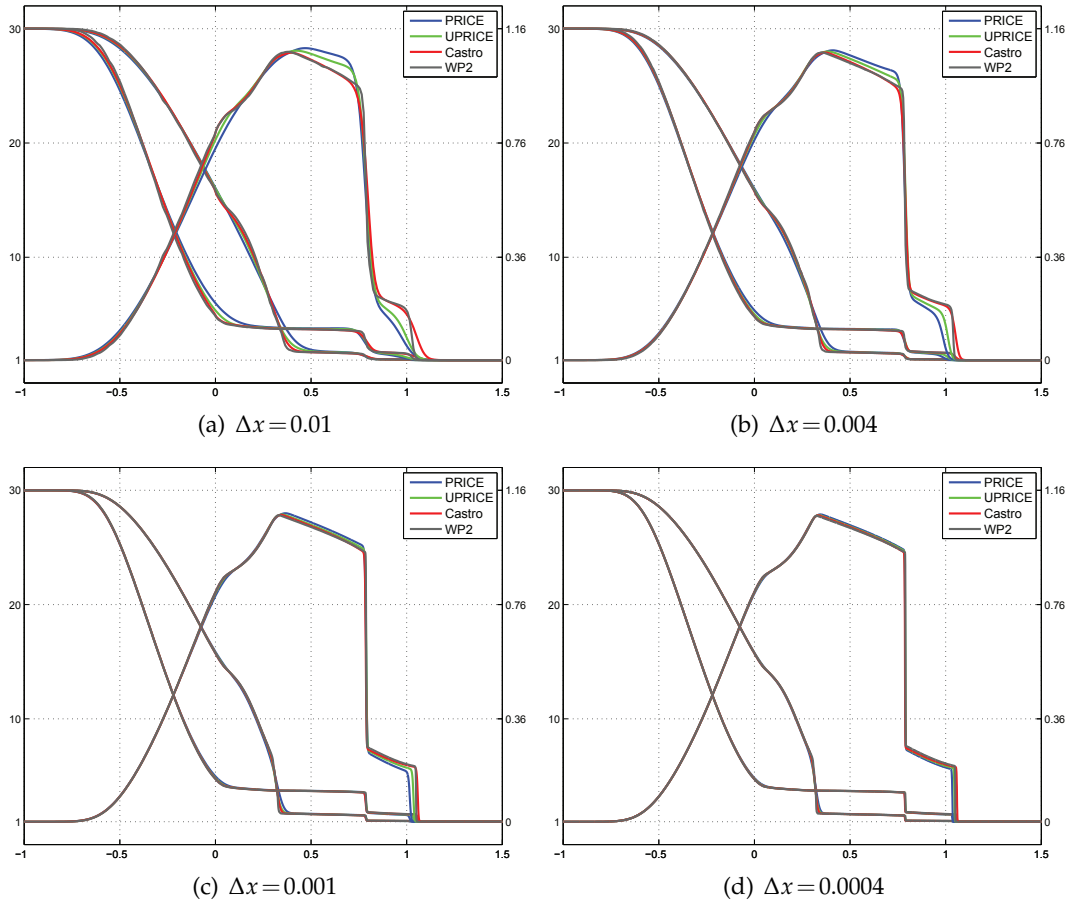


Figure 3: Grid comparison for PRICE, UPRICE, Castro and wave propagation scheme, all in conservative variables and for QBME moment model. Strong shock with $\mathrm{Kn} = 0.5$. The left y-axis is for ρ and p, the right y-axis is for u.

products in the last equations seems to be very small and the difference of the numerical schemes can be neglected if very fine grids are used. This is an important observation, as non-conservative methods do not always converge to the same solution.

Note that we only show results for different grids using the QBME moment model because this model contains the most non-conservative terms and in fact has two non-conservative equations. The influence of the non-conservative terms is thus even smaller for the HME model system.

5.4 Models

After successful comparison of the different numerical schemes, we can now compare the different moment models Grad, HME and QBME mentioned in Section 3. Following the first results of the HME model in [4], we consider the weak shock with Knudsen

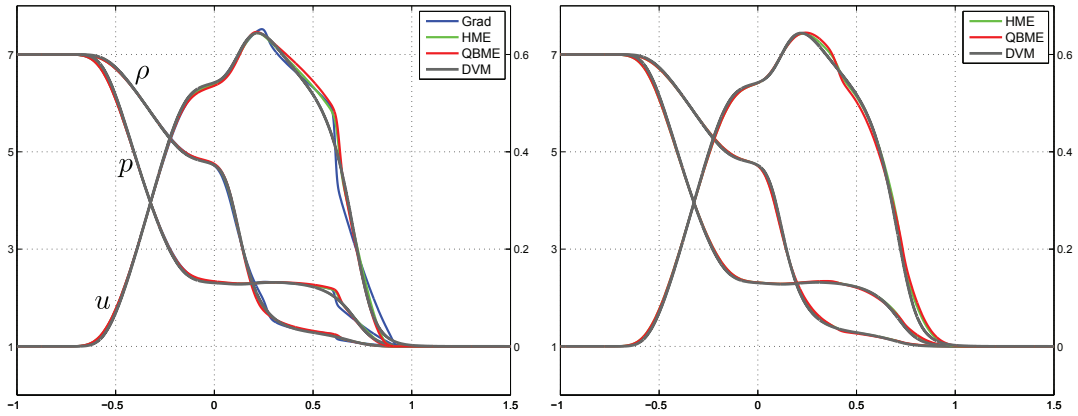
https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press

numbers $\mathrm{Kn}_1 = 0.05$ and $\mathrm{Kn}_2 = 0.5$ as these test cases correspond to small and moderate non-equilibrium and can be seen as the most relevant ones with respect to actual experiments. For simplicity, we use the PRICE scheme and primitive variables. As we have seen in the previous test cases, the choice of variables and the numerical scheme is not crucial when a weak shock with small to moderate Knudsen number is considered. Furthermore, runtime is significantly reduced for this choice. For comparison we use a discrete velocity method (DVM) solution on a very fine velocity grid which can thus be seen as a reference solution. Note that the CFL number changes with increasing number M as the characteristic velocities become larger as well (compare (3.14)). However, the results are very similar for a constant CFL number and for simplicity, we show the results obtained with constant Δt.

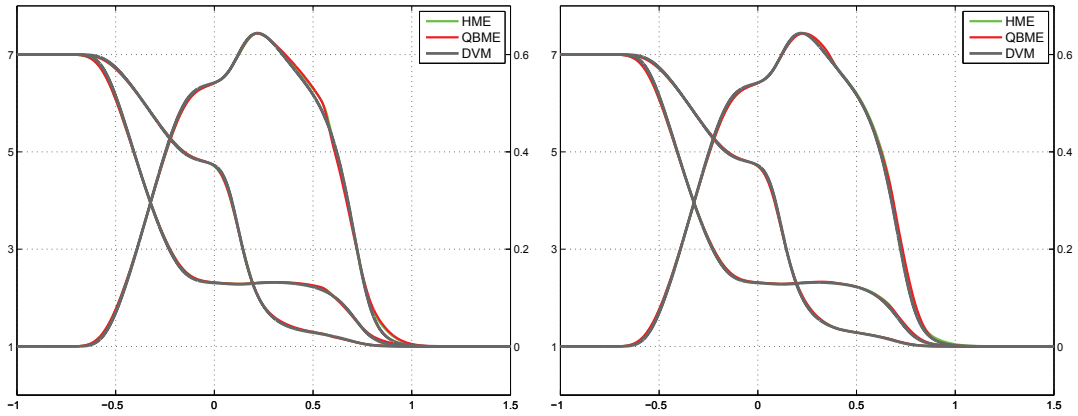
In Fig. 4, the results for a weak shock with the small Knudsen number $\mathrm{Kn}_1 = 0.05$ are shown. The number of equations ranges from $M = 4$ to $M = 9$ and we compare all applicable moment models. Already for $M = 4$ we see a good agreement with the discrete velocity reference solution, especially for $x < 0$. This is similar to previous tests. Note again that the differences in u seem to be larger due to the different scale, but actually they are similar compared to ρ and p. An interesting observation is that the new hyperbolic moment models HME and QBME seem to be as accurate as Grad's method, with all relative errors being around 1%. However, Grad's method breaks down for $M > 4$ and thus we can only compare the solutions of the hyperbolic models. With increasing M, we see even smaller differences to the DVM solution in ρ and p and also the velocity u becomes much more accurate. For the cases with larger M, there is no visual difference to the DVM reference solution, meaning that both moment models converged to the exact solution.

Fig. 5 on the other hand shows results for the larger Knudsen number $\mathrm{Kn}_2 = 0.5$. First of all, Grad's method does not yield any solution as the equations are unstable in this region due to loss of hyperbolicity. Furthermore, we directly see larger differences between HME, QBME and DVM as expected because of the stronger non-equilibrium, best visible for $x > 0$. For $M = 4$, the HME model seems to approximate the maximum velocity value much better but gives worse results in front of the strongest subshock compared to the QBME model. However, the differences between the two new hyperbolic models HME and QBME quickly become smaller with increasing M. Simultaneously, the solutions are approaching the DVM reference better and better. An interesting point is the apparent decoupling of the solutions for even and odd M. This has been observed in [1], too, and is due to the different propagation speeds of the respective models. As the roots of the M-th Hermite polynomial lie between the roots of the $M+1$-st Hermite polynomial, the same holds for the wave speeds as computed in (3.14). Overall, we can still see the solution slowly converging to the DVM reference. Note that the DVM simulation is usually much more costly as more variables are used (in this case there are 400 discrete velocities in the range $[-5,5]$ for every grid cell). On the other hand, moment models like HME or QBME only use $M+1$ variables at every spatial cell to represent the solution. This is a significant model order reduction and the accuracy of the models is still very good.

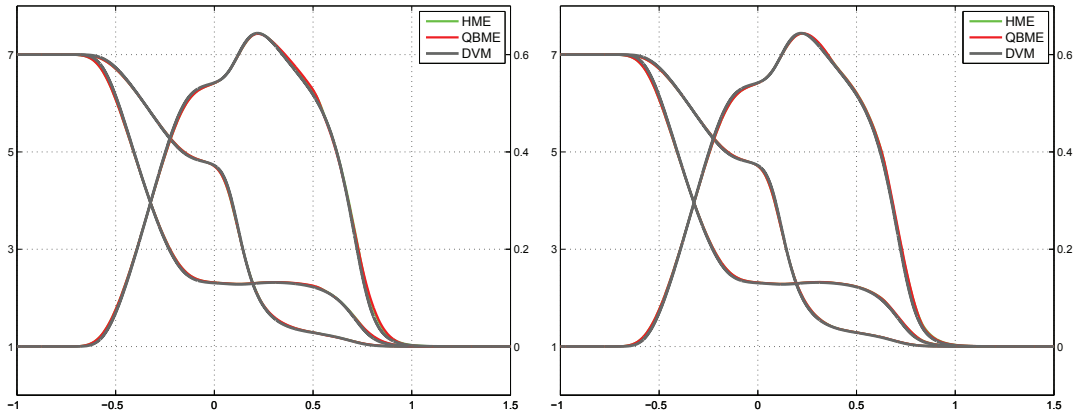
https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press



(a) $M=4$ (b) $M=5$



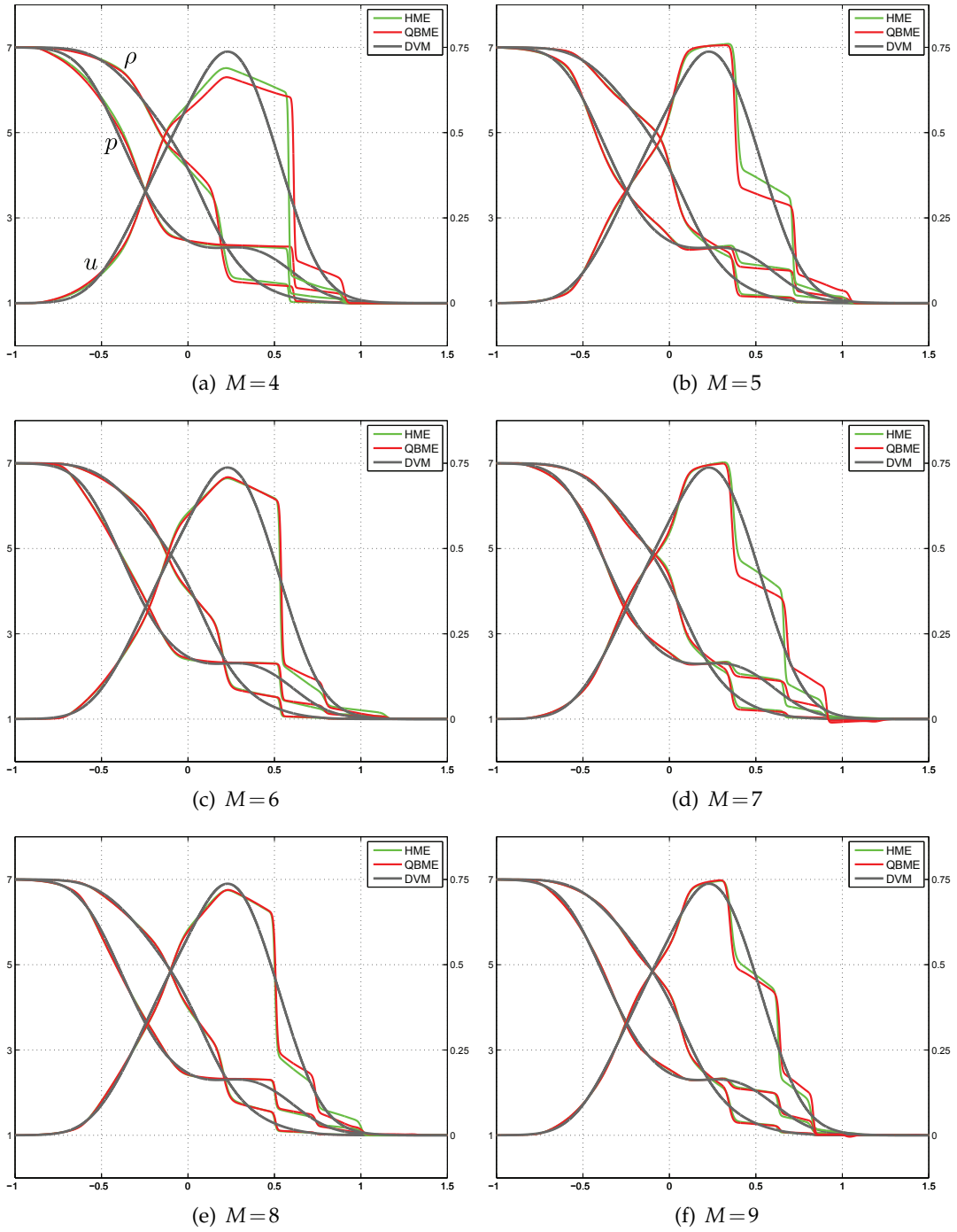
(c) $M=6$ (d) $M=7$



(e) $M=8$ (f) $M=9$

Figure 4: Moment model comparison for Grad, HME, QBME and DVM reference solution, $\mathrm{Kn}=0.05$, all obtained using primitive variables and PRICE scheme. The left y-axis is for ρ and p, the right y-axis is for u.

https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press



(a) $M=4$ (b) $M=5$

(c) $M=6$ (d) $M=7$

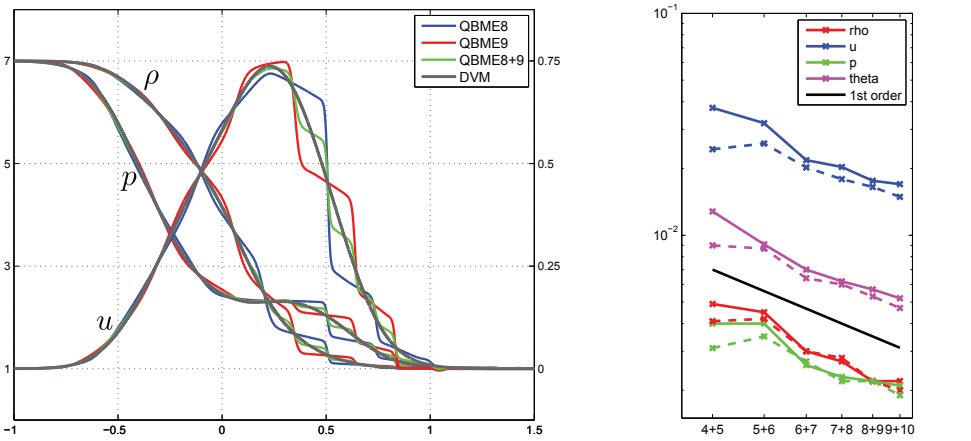
(e) $M=8$ (f) $M=9$

Figure 5: Moment model comparison for Grad, HME, QBME and DVM reference solution, $\mathrm{Kn}=0.5$, all obtained using primitive variables and PRICE scheme. The left y-axis is for ρ and p, the right y-axis is for u.

https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press

5.5 Numerical convergence study

Besides the plots for varying M, we also want to give a plot of the error showing the convergence of the moment models to the DVM reference solution. Here we only consider the case $\mathrm{Kn}_2 = 0.5$. Because of the differences in even and odd M, we first average two subsequent solutions for M and $M+1$, see also [1]. The effect of this averaging is shown in Fig. 6(a). We clearly see the green average approximating the DVM solution much better than both QBME models with different M. This is most obvious for the velocity u, but also holds for the other variables.



(a) Averaged solution of $M=8$ and $M=9$

(b) Convergence of QBME (solid), HME (dashed)

Figure 6: Averaged solution and convergence of averaged solution for $\mathrm{Kn}=0.5$, all results using primitive variables and PRICE scheme.

The relative L^1 errors on the displayed domain $[-1,1.5]$ between the two hyperbolic moment models HME and QBME and the reference DVM model are shown in Fig. 6(b). We see a slightly smaller error for the HME method especially for small M. This can be explained by the fact that the QBME model changes the two last equations which influences the accuracy more if less equations are used. However, the methods show almost the same error for larger M and converge with the same rate. More importantly, all variables ρ, u, p and θ show the same convergence rate with largest relative errors for velocity u due to the different scale. The results stay the same if we use other sets of variables or other numerical methods for this test case. Summarized we obtain linear convergence of both moment systems for increasing number of equations.

6 Conclusion

This paper presented a numerical study of partially conservative moment models using different numerical schemes. It can be seen as an investigation of proper numerical

https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press

schemes for the solution of the newly developed hyperbolic moment models.

After a brief introduction we recalled the derivation of the HME and QBME models as an alternative to the standard Grad model. We gave explicit expressions for the 1D equations together with eigenvalues and eigenvectors for the first time. Additionally we showed new results regarding the change of variables leading to a simple explicit form of the system matrix.

A detailed description of different numerical schemes for the solution of the non-conservative PDE systems was followed by shock tube tests focussing on different aspects of the schemes like the computation of the generalized Roe matrix, the choice of variables and the schemes themselves. As expected we obtained different results for strong shocks or other extreme cases, but for our standard test cases the non-conservative product terms did not cause any problems and all variables and schemes resulted in the same solution, because of the special structure of the new model equations.

Lastly, the accuracy of the new model equations was addressed by comparison with a discrete velocity solution. We saw a very good agreement with the reference solution and increased accuracy for more and more moments also demonstrated by a convergence study.

Further work will include multi-dimensional settings, higher order numerical schemes and test cases that benefit from the rotational invariance of the moment models.

Acknowledgments

This work has been carried out with the support of the German National Academic Foundation.

References

[1] J. D. Au, M. Torrilhon, and W. Weiss. The shocktube-experiment in extended thermodynamics. *Phys. Fluids*, 13(8):2423–2432, 2001.

[2] 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(3):511–525, 1954.

[3] G. A. Bird. *Molecular Gas Dynamics and the Direct Simulation of Gas Flows*. Oxford: Clarendon Press, 1994.

[4] Z. Cai, Y. Fan, and R. Li. Globally hyperbolic regularization of Grad's moment system in one dimensional space. *Comm. Math. Sci.*, 11(2):547–571, 2013.

[5] A. Canestrelli, M. Dumbser, A. Siviglia, and E.F. Toro. Well-balanced high-order centered schemes on unstructured meshes for shallow water equations with fixed and mobile bed. *Adv. Water Resour.*, 33:291–303, 2010.

[6] A. Canestrelli, A. Siviglia, M. Dumbser, and E.F. Toro. Well-balanced high-order centred schemes for non-conservative hyperbolic systems. applications to shallow water equations with fixed and mobile bed. *Adv. Water Resour.*, 32(6):834–844, 2009.

https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press

[7] Y. Fan, J. Koellermeier, J. Li, R. Li, and M. Torrilhon. Model reduction of kinetic equations by operator projection. *J. Stat. Phys.*, 162(2):457–486, 2016.
[8] H. Grad. On the kinetic theory of rarefied gases. *Comm. Pure Appl. Math.*, 2(4):331–407, 1949.
[9] P. Kauf. *Multi-Scale Approximation Models for the Boltzmann Equation*. PhD thesis, ETH Zrich, 2011.
[10] J. Koellermeier. Hyperbolic approximation of kinetic equations using quadrature-based projection methods. Master's thesis, RWTH Aachen University, 2013.
[11] J. Koellermeier, R. Schaerer, and M. Torrilhon. A framework for hyperbolic approximation of kinetic equations using quadrature-based projection methods. *Kinet. Relat. Mod.*, 7(3):531–549, 2014.
[12] J. Koellermeier and M. Torrilhon. On new hyperbolic moment models for the boltzmann equation. In *Conference Proceedings of the YIC GACM 2015, Publication Server of RWTH Aachen University*, 2015.
[13] R. J. LeVeque. Wave propagation algorithms for multidimensional hyperbolic systems. *J. Comp. Phys.*, 131:327–353, 1997.
[14] R.J. LeVeque. *Finite Volume Methods for Hyperbolic Problems*. Cambridge, 2002.
[15] C. D. Levermore. Moment closure hierarchies for kinetic theories. *J. Stat. Phys.*, 83(5–6):1021–1065, 1996.
[16] C. Pars and M. Castro. On the well-balance property of roe's method for nonconservative hyperbolic systems. applications to shallow-water systems. *Math. Model. Anal.*, 38(5):821–852, 2004.
[17] S. Rhebergen, O. Bokhove, and J.J.W. van der Vegt. Discontinuous galerkin finite element methods for hyperbolic nonconservative partial differential equations. *J. Comp. Phys.*, 227:1887–1922, 2008.
[18] R.P. Schärer and M. Torrilhon. On singular closures for the 5-moment system in kinetic gas theory. *Commun. Comput. Phys.*, 17(2):371–400, 2015.
[19] G. Stecca. *Numerical modelling of gravel-bed river morphodynamics*. PhD thesis, Universita' Degli Studi di Trento, 2012.
[20] H. Struchtrup. *Macroscopic Transport Equations for Rarefied Gas Flows*. Springer, 2005.
[21] E. F. Toro and S. J. Billett. Centred tvd schemes for hyperbolic conservation laws. *IMA J. Num. Anal.*, 20:47–79, 2000.
[22] M. Torrilhon. Hyperbolic moment equations in kinetic gas theory based on multi-variate Pearson-IV-distributions. *Commun. Comput. Phys.*, 7(4):639–673, 2010.
[23] F. M. White. *Fluid Mechanics*. WCB/McGraw-Hill, 1999.
[24] W.-A. Yong. Singular perturbations of first-order hyperbolic systems with stiff source terms. *J. Diff. Equations*, 155:89–132, 1999.

https://doi.org/10.4208/cicp.OA-2016-0053 Published online by Cambridge University Press