Journal of Computational Physics 397 (2019) 108838

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

On stochastic Galerkin approximation of the nonlinear Boltzmann equation with uncertainty in the fluid regime ☆

Jingwei Hu [a], Shi Jin [b,*], Ruiwen Shu [c]

[a] Department of Mathematics, Purdue University, West Lafayette, IN 47907, USA
[b] School of Mathematical Sciences, Institute of Natural Sciences, MOE-LSC, Shanghai Jiao Tong University, Shanghai, 200240, PR China
[c] Department of Mathematics, University of Maryland, College Park, MD 20742, USA

ARTICLE INFO

Article history:
Received 29 January 2019
Received in revised form 14 July 2019
Accepted 16 July 2019
Available online 31 July 2019

Keywords:
Boltzmann equation with uncertainty
Stochastic Galerkin
Asymptotic-preserving schemes
Fluid dynamic limit

ABSTRACT

The Boltzmann equation may contain uncertainties in initial/boundary data or collision kernel. To study the impact of these uncertainties, a stochastic Galerkin (sG) method was proposed in [18] and studied in the kinetic regime. When the system is close to the fluid regime (the Knudsen number is small), the method would become prohibitively expensive due to the stiff collision term. In this work, we develop efficient sG methods for the Boltzmann equation that work for a wide range of Knudsen numbers, and investigate, in particular, their behavior in the fluid regime.

1. Introduction

Kinetic equations describe the non-equilibrium dynamics of a large number of particles from a statistical point of view [11]. Their applications range from rarefied gas dynamics, plasma physics to biological and social sciences. The Boltzmann equation, which takes into account particle transport and binary collisions, is the most fundamental kinetic equation [8]. Let $f(t,x,v) \geq 0$ be a phase space distribution function of time t, position x, and particle velocity v, then the equation reads:

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f = \frac{1}{\varepsilon} Q(f,f), \quad x \in \Omega \subset \mathbb{R}^d, \quad v \in \mathbb{R}^d, \tag{1.1}$$

where $Q(f,f)$, called the collision operator, is a high-dimensional, nonlinear, nonlocal integral operator (form to be presented later). ε is the Knudsen number, defined as the ratio of the mean free path and the typical length scale. Simply speaking, ε measures the degree of rarifiedness of the particle system. Depending on the application, ε may be large: $\varepsilon = O(1)$ (kinetic regime, system far from equilibrium), small: $\varepsilon \ll 1$ (fluid regime, system close to equilibrium), or even varies across different scales (multiscale phenomena).

In the past decades, there have been numerous research studies on equation (1.1) both theoretically (cf. [10,41]) and numerically (cf. [9,13]). However, all these work dealt with the deterministic problem, namely, the initial/boundary conditions and parameters appear in the collision operator are completely certain. Recently, there has been a significant interest to

☆ J.H. was partially supported by NSF grant DMS-1620250 and NSF CAREER grant DMS-1654152. S.J. was partially supported by the NSFC grants No. 31571071 and No. 11871297.

* Corresponding author.
E-mail addresses: jingweihu@purdue.edu (J. Hu), shijin-m@sjtu.edu.cn (S. Jin), rshu@cscamm.umd.edu (R. Shu).

https://doi.org/10.1016/j.jcp.2019.07.037
0021-9991/

study the impact of random inputs to the kinetic equations, see [26,24,23,25], and in particular, [18,33] for the nonlinear Boltzmann equation. The rationale of these studies is twofold: 1) As a reliable equation bridging mesoscopic and macroscopic descriptions, the initial/boundary conditions of the Boltzmann equation are usually given in terms of macroscopic quantities like density, temperature, pressure, etc., whose measurement is based on experiments and inevitably contain uncertainties — in this aspect, the problem is similar to those investigated extensively in computational fluid dynamics [34,6,35]; 2) In the equation itself, there is a key term called collision kernel (or cross section) which characterizes the law of interaction between particles. Calculating its form from first principles is fairly complicated, and even impossible most of the time, thus in practice, empirical kernels containing adjustable parameters are often used with the aim to reproduce viscosity and diffusion coefficients [7,17,5,30]. This, on the other hand, brings a source of uncertainty to the system because a minor perturbation in the parameters may affect the solution significantly.

To quantify the aforementioned uncertainties, the work in [18] took a generalized polynomial chaos based stochastic Galerkin (gPC-sG) approximation. The gPC-sG approach, essentially a spectral method in the random domain, has been successfully applied to many science and engineering problems, see for instance [16,44,35]. Due to the nonlinear, nonlocal nature of the Boltzmann equation, the main focus of [18] is to develop a fast algorithm to evaluate the collision operator under the Galerkin projection, while time and spatial derivatives are discretized using simple explicit schemes. This works well in the kinetic regime, and the time step constraint for numerical stability is not as severe as in those problems with possibly small mean free time (the case of fluid dynamic regime). Unfortunately, the method will become extremely expensive, if not impossible, in the fluid regime, because the stiffness induced by the collision operator would require $\Delta t = O(\varepsilon)$. To overcome this bottleneck, the goal of this paper is to design efficient time discretizations that are uniformly stable for a wide range of Knudsen numbers (i.e., Δt can be chosen independently of ε). We mention that the construction of such schemes is not new for deterministic kinetic/hyperbolic equations involving small parameters, and they fall under the umbrella of the so-called Asymptotic-Preserving (AP) scheme, whose main feature is that it becomes a consistent discretization to the limiting equation as $\varepsilon \to 0$ with time step fixed, see [22] and references therein. In the UQ framework, the issue was addressed only for transport equations with diffusive scaling [26,23,27], or kinetic equations with incompressible Navier-Stokes limit [25], and no work has been done for nonlinear kinetic equations such as the Boltzmann equation in the compressible Euler regime. On the analysis side, hypocoercivity theory based sensitivity and long-time behavior of uncertain nonlinear kinetic equations and their stochastic approximations were studied in [38,29,33,12].

We now summarize our contributions of this work. Following the AP scheme for the deterministic Boltzmann equation [15], we construct a direct analogue in the gPC-sG setting. This approach requires the computation of the Maxwellian (equilibrium of the collision operator) which is expensive and may break down when density or temperature becomes negative. Inspired by [14], we then develop an alternative approach which uses a pseudo Maxwellian constructed easily from macroscopic quantities rather than the true Maxwellian. Yet the choice of pseudo Maxwellian is not unique and how to choose an appropriate one is not known a priori. Finally, we propose a new, novel method based on simple operator splitting with a linear penalization. It does not need the evaluation of any macroscopic quantities nor Maxwellian, and is shown to be robust for both continuous and discontinuous problems. The three methods will be shown to be AP, and be compared numerically and their pros and cons will be studied. Apart from the numerical contribution, we also investigate theoretically the hyperbolicity of the gPC-sG scheme for the Boltzmann equation when ε goes to zero. It is known that the sG method applied to the hyperbolic systems such as the compressible Euler equations may lead to an enlarged system which is not necessarily hyperbolic [37]. Here, from a kinetic point of view, we present an asymptotic analysis by assuming the random perturbation is *small*, and show that up to linear order, the limiting scheme is *weakly hyperbolic*. The limiting behavior for general random perturbation remains an open problem.

Finally, we mention that the gPC-sG framework adopted in this work, like the spectral method, suffers from the Gibb's phenomenon when the solution becomes discontinuous. Furthermore, it does not preserve the positivity of positive physical quantities. These issues have been studied in some UQ literatures in other applications, for examples [43,42,20,31,32,1,4,2,39], and remain to be addressed for the Boltzmann equation.

The rest of this paper is organized as follows. In the next section we briefly summarize the properties of the Boltzmann equation and review the gPC-sG method in [18]. We also study the hyperbolicity of the gPC-sG system under certain assumptions. Section 3 describes in detail the construction of three AP gPC-sG schemes for the Boltzmann equation with uncertainty. Section 4 presents various numerical examples for the proposed schemes, where AP properties, kinetic, fluid and mixed regimes are carefully examined and different numerical schemes are compared. The paper is concluded in section 5.

2. The Boltzmann equation with uncertainty and its gPC-sG approximation

In order to set up the basis for the construction of AP schemes, in this section we briefly summarize the properties of the Boltzmann equation and review its gPC-sG approximation. For simplicity, we will only consider the uncertainty coming from the initial data and the one-dimensional random variable. For a discussion of more general case, we refer to paper [18].

2.1. Basics of the Boltzmann equation

When the random inputs appear in the initial condition, the Boltzmann equation reads exactly the same as in (1.1) except that the function $f = f(t, x, v, z)$ depends on an extra random variable z. The collision operator also takes the same form as in the deterministic case (note that $\mathcal{Q}$ acts only in the velocity space, so t, x, and z are suppressed from the functions below):

$$\mathcal{Q}(f, g)(v) = \int_{\mathbb{R}^d} \int_{S^{d-1}} B(v - v_*, \sigma) \left[f(v')g(v'_*) - f(v)g(v_*) \right] \mathrm{d}\sigma \, \mathrm{d}v_*. \tag{2.1}$$

Here (v, v_*) and (v', v'_*) are the velocity pairs before and after a collision, during which the momentum and energy are conserved; hence (v', v'_*) can be represented in terms of (v, v_*) as

$$\begin{cases} v' = \dfrac{v + v_*}{2} + \dfrac{|v - v_*|}{2}\sigma, \\ v'_* = \dfrac{v + v_*}{2} - \dfrac{|v - v_*|}{2}\sigma, \end{cases} \tag{2.2}$$

with the parameter σ varying on the unit sphere S^{d-1}. The collision kernel $B(v - v_*, \sigma)$ is a non-negative function depending only on $|v - v_*|$ and cosine of the deviation angle $\frac{\sigma \cdot (v - v_*)}{|v - v_*|}$. For numerical purpose, the variable hard sphere (VHS) model [7] is commonly used:

$$B(|v - v_*|, \cos\theta) = b_\lambda |v - v_*|^\lambda, \quad -d < \lambda \leq 1, \tag{2.3}$$

where b_λ is a positive constant, $\lambda > 0$ corresponds to the hard potentials, and $\lambda < 0$ to the soft potentials.

$\mathcal{Q}(f, f)$ conserves mass, momentum, and energy:

$$\int_{\mathbb{R}^d} \mathcal{Q}(f, f)\phi(v) \, \mathrm{d}v = 0, \quad \phi(v) = 1, v, \frac{|v|^2}{2}. \tag{2.4}$$

Also, it satisfies the celebrated Boltzmann's H-theorem:

$$\int_{\mathbb{R}^d} \mathcal{Q}(f, f) \ln f \, \mathrm{d}v \leq 0. \tag{2.5}$$

Moreover,

$$\int_{\mathbb{R}^d} \mathcal{Q}(f, f) \ln f \, \mathrm{d}v = 0 \Longleftrightarrow \mathcal{Q}(f, f) = 0 \Longleftrightarrow f = M, \tag{2.6}$$

where M is the local equilibrium, also called *Maxwellian*, given by

$$M(v)_{(\rho,u,T)} = \frac{\rho}{(2\pi T)^{\frac{d}{2}}} e^{-\frac{|v-u|^2}{2T}}, \tag{2.7}$$

and ρ, u, T are, respectively, the density, bulk velocity, and temperature defined as

$$\rho = \int_{\mathbb{R}^d} f(v) \, \mathrm{d}v, \quad u = \frac{1}{\rho} \int_{\mathbb{R}^d} f(v) v \, \mathrm{d}v, \quad T = \frac{1}{d\rho} \int_{\mathbb{R}^d} f(v) |v - u|^2 \, \mathrm{d}v. \tag{2.8}$$

When $\varepsilon \to 0$ in (1.1), formally one has $f \to M$. Replacing f with M and taking the moments $\int \cdot \phi(v) \, \mathrm{d}v$ on both sides of the equation, one can derive the compressible Euler equations:

$$\begin{cases} \dfrac{\partial \rho}{\partial t} + \nabla_x \cdot (\rho u) = 0, \\ \dfrac{\partial \rho u}{\partial t} + \nabla_x \cdot (\rho u \otimes u + pI) = 0, \\ \dfrac{\partial E}{\partial t} + \nabla_x \cdot ((E + p)u) = 0, \end{cases} \tag{2.9}$$

where $E = \rho \frac{d}{2} T + \frac{1}{2} \rho u^2$ is the total energy, and $p = \rho T$ is the pressure. Note that this limiting system also contains uncertainty since the macroscopic quantities ρ, u and T depend on z.

2.2. The gPC-sG approximation

In the gPC-sG method, one seeks a solution in the following form

$$f(t,x,v,z) \approx \sum_{k=0}^{K} f_k(t,x,v)\Phi_k(z) := f^K(t,x,v,z), \tag{2.10}$$

where $\{\Phi_k(z)\}$ are orthonormal polynomials that form the gPC basis, and satisfy

$$\int_{I_z} \Phi_k(z)\Phi_j(z)\pi(z)\,\mathrm{d}z = \delta_{kj}, \quad 0 \le k, j \le K, \tag{2.11}$$

where $\pi(z)$ is the probability distribution function of z and δ_{kj} is the Kronecker delta function. Inserting (2.10) into (1.1) and performing a Galerkin projection, one obtains

$$\frac{\partial f_k}{\partial t} + v\cdot\nabla_x f_k = \frac{1}{\varepsilon}\mathcal{Q}_k, \quad 0 \le k \le K, \tag{2.12}$$

where

$$\mathcal{Q}_k(t,x,v) := \int_{I_z} \mathcal{Q}(f^K, f^K)(t,x,v,z)\Phi_k(z)\pi(z)\,\mathrm{d}z. \tag{2.13}$$

This, together with the initial condition

$$f_k(0,x,v) = \int_{I_z} f^0(x,v,z)\Phi_k(z)\pi(z)\,\mathrm{d}z, \tag{2.14}$$

constitutes the complete stochastic Galerkin system.

Given f_k, the statistical information, e.g., the mean and variance of the solution can be approximated as

$$\mathbb{E}[f] \approx f_0, \quad \mathrm{Var}[f] \approx \sum_{k=1}^{K} f_k^2. \tag{2.15}$$

2.3. The equilibrium of $\mathcal{Q}_k$

In this subsection, we study the equilibrium of the numerical collision operator $\mathcal{Q}_k$ in order to understand the behavior of the sG system (2.12). We can find the equilibrium by assuming that the random perturbation is small. For more general random perturbation the local equilibrium of $\mathcal{Q}_k$ is still unknown.

We assume $f_0 \ge 0$ and it is larger than other gPC coefficients (i.e., the uncertainty is small):

$$f^K(v,z) = f_0(v) + \omega\sum_{k=1}^{K} f_k(v)\Phi_k(z), \quad 0 < \omega \ll 1. \tag{2.16}$$

Suppose

$$\mathcal{Q}_k(v) = \int_{I_z} \mathcal{Q}(f^K, f^K)(v,z)\Phi_k(z)\pi(z)\,\mathrm{d}z = 0, \quad \text{for } 0 \le k \le K, \tag{2.17}$$

we derive a form of f^K up to $O(\omega^2)$ in the following.

Substituting (2.16) into (2.17) and using the orthogonality of $\{\Phi_k(z)\}$, one has

$$\mathcal{Q}(f_0,f_0)\delta_{0k} + \omega\sum_{i=1}^{K}(\mathcal{Q}(f_0,f_i) + \mathcal{Q}(f_i,f_0))\delta_{ik} + O(\omega^2) = 0, \quad 0 \le k \le K. \tag{2.18}$$

Balancing $O(1)$ and $O(\omega)$ terms require

$$\mathcal{Q}(f_0,f_0) = 0, \tag{2.19}$$

and

$$\mathcal{Q}(f_0,f_k) + \mathcal{Q}(f_k,f_0) = 0, \quad 1 \le k \le K. \tag{2.20}$$

(2.19) implies that f_0 should take the form of a Maxwellian:

$$f_0(v) := M_0(v) = \frac{\rho_0}{(2\pi T_0)^{\frac{d}{2}}} e^{-\frac{|v-u_0|^2}{2T_0}}. \tag{2.21}$$

Then the left hand side of (2.20) is just the linearized collision operator around M_0, whose kernel is given by

$$f_k(v) = M_0(v)(a_k + b_k \cdot v + c_k|v|^2), \quad 1 \le k \le K, \tag{2.22}$$

where a_k, b_k, and c_k are some constants (or constant vector) independent of ω. Therefore, we have

$$f^K(v,z) = M_0(v)\left[1 + \omega \sum_{k=1}^{K} (a_k + b_k \cdot v + c_k|v|^2)\Phi_k(z)\right] + O(\omega^2). \tag{2.23}$$

Furthermore, we define

$$\int_{\mathbb{R}^d} f^K \,\mathrm{d}v = \rho(z), \quad \int_{\mathbb{R}^d} f^K v \,\mathrm{d}v = \rho(z)u(z), \quad \int_{\mathbb{R}^d} f^K \frac{1}{2}|v|^2 \,\mathrm{d}v = \frac{1}{2}\rho(z)u(z)^2 + \frac{d}{2}\rho(z)T(z). \tag{2.24}$$

If we assume $\rho(z)$, $u(z)$, $T(z)$ have the following expansion:

$$\rho(z) = \tilde{\rho}_0 + \omega \sum_{k=1}^{K} \rho_k \Phi_k(z), \quad u(z) = \tilde{u}_0 + \omega \sum_{k=1}^{K} u_k \Phi_k(z), \quad T(z) = \tilde{T}_0 + \omega \sum_{k=1}^{K} T_k \Phi_k(z), \tag{2.25}$$

by matching the leading order terms of (2.24), it is not difficult to see that

$$\tilde{\rho}_0 = \rho_0, \quad \tilde{u}_0 = u_0, \quad \tilde{T}_0 = T_0. \tag{2.26}$$

If we further match the $O(\omega)$ terms and use the orthogonality of $\{\Phi_k(z)\}$, we can obtain

$$a_k = \frac{1}{\rho_0}\rho_k - \frac{u_0}{T_0} \cdot u_k + \left(\frac{u_0^2}{2T_0^2} - \frac{d}{2T_0}\right) T_k, \quad b_k = \frac{u_k}{T_0} - \frac{u_0}{T_0^2} T_k, \quad c_k = \frac{1}{2T_0^2} T_k. \tag{2.27}$$

On the other hand, consider a Maxwellian function

$$M(v,z) = \frac{\rho(z)}{(2\pi T(z))^{\frac{d}{2}}} e^{-\frac{|v-u(z)|^2}{2T(z)}}, \tag{2.28}$$

with $\rho(z)$, $u(z)$, $T(z)$ given by (2.25). Expanding $M(v,z)$ around ρ_0, T_0, u_0 yields

$$\begin{aligned} M(v,z) =& M_0(v) + \omega \left(\frac{\partial M}{\partial \rho}|_{\rho_0}\right)\left(\sum_{k=1}^{K} \rho_k \Phi_k(z)\right) + \omega \left(\frac{\partial M}{\partial u}|_{u_0}\right)\left(\sum_{k=1}^{K} u_k \Phi_k(z)\right) \\ &+ \omega \left(\frac{\partial M}{\partial T}|_{\rho_0}\right)\left(\sum_{k=1}^{K} T_k \Phi_k(z)\right) + O(\omega^2) \\ =& M_0(v)\left[1 + \omega \sum_{k=1}^{K}\left(\frac{1}{\rho_0}\rho_k + \left(\frac{v-u_0}{T_0}\right)\cdot u_k + \left(\frac{(v-u_0)^2}{2T_0^2} - \frac{d}{2T_0}\right) T_k\right)\Phi_k(z)\right] + O(\omega^2). \end{aligned} \tag{2.29}$$

Comparing (2.23) and (2.29) (and taking into account (2.27)), we see that

$$f^K(v,z) = M(v,z) + O(\omega^2). \tag{2.30}$$

2.4. The hyperbolicity

In this subsection, we study the hyperbolicity of the gPC-sG system (2.12) when $\varepsilon \to 0$. Assume $d = 1$ for simplicity. The discussion in the previous subsection implies that as $\varepsilon \to 0$,

$$f_0 = M_0(v), \quad f_k = \omega M_0(v)\left(\frac{1}{\rho_0}\rho_k + \left(\frac{v-u_0}{T_0}\right)u_k + \left(\frac{(v-u_0)^2}{2T_0^2} - \frac{1}{2T_0}\right)T_k\right). \tag{2.31}$$

Substituting f_k $(0 \le k \le K)$ into (2.12) and taking the moments $\int \cdot (1, v, v^2/2)^T \,\mathrm{d}v$, we get (after lengthy calculation)

$$\frac{\partial U}{\partial t}+\frac{\partial F(U)}{\partial x}=0, \tag{2.32}$$

with

$$\begin{aligned} &U=(U_1,U_2,U_3,\ldots,U_{3k+1},U_{3k+2},U_{3k+3},\ldots,U_{3K+1},U_{3K+2},U_{3K+3}),\\ &F(U)=(F_1,F_2,F_3,\ldots,F_{3k+1},F_{3k+2},F_{3k+3},\ldots,F_{3K+1},F_{3K+2},F_{3K+3}), \end{aligned} \tag{2.33}$$

where

$$U_1=\rho_0,\quad U_2=\rho_0u_0,\quad U_3=\frac{1}{2}\rho_0u_0^2+\frac{1}{2}\rho_0T_0, \tag{2.34}$$

$$F_1=U_2,\quad F_2=2U_3,\quad F_3=\frac{3U_2U_3}{U_1}-\frac{U_2^3}{U_1^2}, \tag{2.35}$$

and for $1<k\le K$,

$$\begin{aligned} &U_{3k+1}=\omega\rho_k,\quad U_{3k+2}=\omega(\rho_ku_0+\rho_0u_k),\quad U_{3k+3}=\omega\left(\frac{1}{2}\rho_ku_0^2+\frac{1}{2}\rho_kT_0+\rho_0u_0u_k+\frac{1}{2}\rho_0T_k\right),\\ &F_{3k+1}=U_{3k+2},\quad F_{3k+2}=2U_{3k+3},\\ &F_{3k+3}=\left(\frac{2U_2^3}{U_1^3}-\frac{3U_2U_3}{U_1^2}\right)U_{3k+1}+\left(\frac{3U_3}{U_1}-\frac{3U_2^2}{U_1^2}\right)U_{3k+2}+\frac{3U_2}{U_1}U_{3k+3}. \end{aligned} \tag{2.36}$$

Therefore, the Jacobian matrix $F'(U)$ (size $(3K+3)\times(3K+3)$) looks like

$$\begin{pmatrix} A & 0 & 0 & \cdots & 0\\ B & A & 0 & \cdots & 0\\ B & 0 & A & \cdots & 0\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ B & 0 & 0 & \cdots & A \end{pmatrix},$$

with

$$A=\begin{pmatrix} 0 & 1 & 0\\ 0 & 0 & 2\\ a & b & c \end{pmatrix},\quad B=\begin{pmatrix} 0 & 0 & 0\\ 0 & 0 & 0\\ d & e & f \end{pmatrix},$$

and

$$\begin{aligned} &a=\frac{2U_2^3}{U_1^3}-\frac{3U_2U_3}{U_1^2},\quad b=\frac{3U_3}{U_1}-\frac{3U_2^2}{U_1^2},\quad c=\frac{3U_2}{U_1},\\ &d=\left(-\frac{6U_2^3}{U_1^4}+\frac{6U_2U_3}{U_1^3}\right)U_{3k+1}+\left(-\frac{3U_3}{U_1^2}+\frac{6U_2^2}{U_1^3}\right)U_{3k+2}-\frac{3U_2}{U_1^2}U_{3k+3},\\ &e=\left(\frac{6U_2^2}{U_1^3}-\frac{3U_3}{U_1^2}\right)U_{3k+1}-\frac{6U_2}{U_1^2}U_{3k+2}+\frac{3}{U_1}U_{3k+3},\\ &f=-\frac{3U_2}{U_1^2}U_{3k+1}+\frac{3}{U_1}U_{3k+2}. \end{aligned} \tag{2.37}$$

The eigenvalues of the first block are

$$\lambda_1=u_0-\sqrt{3T_0},\quad \lambda_2=u_0,\quad \lambda_3=u_0+\sqrt{3T_0}. \tag{2.38}$$

One can show by math induction that the eigenvalues of the whole matrix $F'(U)$ are

$$\lambda_1=u_0-\sqrt{3T_0},\quad \lambda_2=u_0,\quad \lambda_3=u_0+\sqrt{3T_0}, \tag{2.39}$$

each with multiplicity $(K+1)$.

For each eigenvalue λ_i, $i=1,2,3$, there are K eigenvectors and they are of the form:

$$(\mathbf{0}_{3k},\xi_i,\mathbf{0}_{3(K-k)}),\quad k=1,\ldots,K, \tag{2.40}$$

with $\mathbf{0}_k$ denoting the zero vector of length k, and

$$\xi_1 = \left(\frac{2U_1^2}{(U_2 - \sqrt{6U_1U_3 - 3U_2^2})^2}, \frac{2U_1}{U_2 - \sqrt{6U_1U_3 - 3U_2^2}}, 1 \right) = \left(\frac{2}{(u_0 - \sqrt{3T_0})^2}, \frac{2}{u_0 - \sqrt{3T_0}}, 1 \right),$$
$$\xi_2 = \left(\frac{2U_1^2}{U_2^2}, \frac{2U_1}{U_2}, 1 \right) = \left(\frac{2}{u_0^2}, \frac{2}{u_0}, 1 \right), \tag{2.41}$$
$$\xi_3 = \left(\frac{2U_1^2}{(U_2 + \sqrt{6U_1U_3 - 3U_2^2})^2}, \frac{2U_1}{U_2 + \sqrt{6U_1U_3 - 3U_2^2}}, 1 \right) = \left(\frac{2}{(u_0 + \sqrt{3T_0})^2}, \frac{2}{u_0 + \sqrt{3T_0}}, 1 \right).$$

The above calculation shows that the system is weakly hyperbolic (the set of eigenvectors is not complete, since the multiplicity is $K+1$, and the number of eigenvectors is K). We also point out that this is only true for an asymptotic expansion on $O(\omega)$ up to first order. It remains to understand the hyperbolicity, or the opposite, of the limiting gPC-sG system in the non-perturbative setting, or for more general random perturbation. For nonlinear Fokker-Planck equations leading to isentropic Euler equations in the fluid limit, counter examples which show loss of hyperbolicity were given in [28].

3. AP gPC-sG schemes for the Boltzmann equation with uncertainty

Clearly the main difficulty of solving the system (2.12) is to compute the Galerkin projected collision operator $\mathcal{Q}_k$. A fast algorithm was proposed in [18] which significantly reduces the complexity compared to a direct evaluation. Nevertheless, it is still the most expensive part of the whole computation and should be kept as minimal as possible. When ε is small, as mentioned in the introduction, an explicit scheme would require $\Delta t = O(\varepsilon)$ which is prohibitively expensive because of the excessive evaluation of $\mathcal{Q}_k$. An implicit scheme allows larger Δt, but inverting a nonlinear integral operator is computationally daunting. With these in mind, in this section we propose a series of AP schemes (each has its own pros and cons) for the gPC-sG system (2.12).

3.1. AP schemes for the deterministic Boltzmann equation

To illustrate the idea, we start with the deterministic equation (1.1). We will consider three different schemes and the third one is new to the best of our knowledge. The designing principles are: 1) the scheme is free of Newton type nonlinear algebraic solvers; 2) when $\varepsilon \to 0$, it should automatically become a consistent discretization of the limiting Euler system, with a stability condition independent of ε (in other words, allowing Δt to be independent of ε); 3) the generalization to the uncertain case should be straightforward.

3.1.1. The scheme with Maxwellian penalization

To remove the stiffness in the collision term, [15] introduces a BGK penalty method. The idea is to penalize $\mathcal{Q}(f,f)$ by the BGK operator $P(f) = M - f$ which can be inverted easily, and to treat $\mathcal{Q}(f,f)$ explicitly. A first-order scheme reads as

$$\frac{f^{n+1} - f^n}{\Delta t} + v \cdot \nabla_x f^n = \frac{\mathcal{Q}(f^n, f^n) - \lambda^n (M^n - f^n)}{\varepsilon} + \frac{\lambda^n (M^{n+1} - f^{n+1})}{\varepsilon}, \tag{3.1}$$

where λ is some constant that can be made time/spatially dependent. (3.1) appears to be implicit at first sight as it contains M^{n+1}. But note that $\mathcal{Q}(f,f)$ conserves mass, momentum and energy, so does $P(f)$. Thus taking the moments $\int \cdot \phi(v)\,\mathrm{d}v$, $\phi(v) := (1, v, v^2/2)^T$ on both sides of the scheme yields

$$\frac{U^{n+1} - U^n}{\Delta t} + \nabla_x \cdot \int v\phi(v) f^n \,\mathrm{d}v = 0, \tag{3.2}$$

where $U := (\rho, m, E)^T$ is a vector of mass, momentum, and total energy. From U^{n+1}, one can easily define M^{n+1} through (2.7). Plugging it back to (3.1), f^{n+1} can be explicitly determined.

To formally prove the AP property, we assume $0 < \varepsilon \ll \Delta t \ll 1$, $0 < \lambda = O(1)$, and all functions are smooth in the following. (3.1) can be rearranged as

$$f^{n+1} - M^{n+1} = \frac{\varepsilon + D^n \Delta t}{\varepsilon + \lambda^n \Delta t}(f^n - M^n) - \frac{\varepsilon \Delta t}{\varepsilon + \lambda^n \Delta t}\left[v \cdot \nabla_x f^n + \frac{M^{n+1} - M^n}{\Delta t} \right], \tag{3.3}$$

where

$$D^n := \frac{\mathcal{Q}(f^n, f^n) - \mathcal{Q}(M^n, M^n)}{f^n - M^n} + \lambda^n. \tag{3.4}$$

Suppose one can choose λ^n such that

$$\sup_v |D^n| < \alpha\lambda^n, \quad \forall n \tag{3.5}$$

for some constant $0 < \alpha < 1$, then

$$|f^{n+1} - M^{n+1}| \le \alpha |f^n - M^n| + O(\varepsilon). \tag{3.6}$$

Iteratively, this gives

$$|f^n - M^n| \le \alpha^n |f^0 - M^0| + O(\varepsilon). \tag{3.7}$$

Therefore, for arbitrary initial data f^0, there holds

$$f^n - M^n = O(\varepsilon), \tag{3.8}$$

if n is sufficiently large.

Substituting this into (3.2), we get

$$\frac{U^{n+1} - U^n}{\Delta t} + \nabla_x \cdot \int v\phi(v) M^n \,\mathrm{d}v + O(\varepsilon) = 0, \tag{3.9}$$

which shows that the scheme is AP, i.e. as $\varepsilon \to 0$, it becomes a first-order time discretization of the Euler system.

Remark 3.1. It is not easy to show the existence of λ that satisfies the condition (3.5) for general collision operators. For the BGK operator $\mathcal{Q}(f, f) = M - f$, it just requires $\lambda > \frac{1}{2}$. For the Boltzmann operator, numerical experiments suggest that $\lambda = \sup_v \mathcal{Q}^-(f)$ usually gives the desired convergence ($\mathcal{Q}(f, f) := \mathcal{Q}^+(f, f) - f\mathcal{Q}^-(f)$).

3.1.2. The scheme with pseudo Maxwellian penalization

The previous scheme relies on the evaluation of the Maxwellian at every spatial point and time step. Sometimes, it may be difficult/expensive to compute the true Maxwellian (especially so when one considers uncertainty in the gPC-sG setting, see next section). The scheme proposed in [14] provides an alternative. The only difference is to replace M in (3.1) with a pseudo Maxwellian M_p, and the minimum requirement on M_p is that it has the same first $d+2$ moments as f: $\int M_p \phi(v)\,\mathrm{d}v = \int f\phi(v)\,\mathrm{d}v = U$. The scheme then reads:

$$\frac{f^{n+1} - f^n}{\Delta t} + v \cdot \nabla_x f^n = \frac{\mathcal{Q}(f^n, f^n) - \lambda^n (M_p^n - f^n)}{\varepsilon} + \frac{\lambda^n (M_p^{n+1} - f^{n+1})}{\varepsilon}. \tag{3.10}$$

By construction, the right-hand side of (3.10) is still conservative, so (3.2) remains valid.

To show the AP property of this scheme, we rewrite (3.10) as

$$\begin{aligned} f^{n+1} - M^{n+1} =& \frac{\varepsilon + D^n \Delta t}{\varepsilon + \lambda^n \Delta t}(f^n - M^n) - \frac{\varepsilon \Delta t}{\varepsilon + \lambda^n \Delta t}(v \cdot \nabla_x f^n) \\ & - (M^{n+1} - M^n) + \frac{\lambda^n \Delta t}{\varepsilon + \lambda^n \Delta t}(M_p^{n+1} - M_p^n), \end{aligned} \tag{3.11}$$

where D^n is given by (3.4). Note that the last two terms in the above equation are of $O(\Delta t)$. Then by a similar argument as in the previous section, one can derive that for arbitrary initial data f^0, after a few time steps (see [14]),

$$f^n - M^n = O(\Delta t). \tag{3.12}$$

Substituting this into (3.10) and taking the moments, we have

$$\frac{U^{n+1} - U^n}{\Delta t} + \nabla_x \cdot \int v\phi(v) M^n \,\mathrm{d}v + O(\Delta t) = 0, \tag{3.13}$$

so this scheme is again AP although the local truncation error is $O(\Delta t)$, larger than $O(\varepsilon)$ in the scheme using the true Maxwellian penalization (if ε is much smaller than Δt).

3.1.3. The scheme without Maxwellian: a linear penalty

The previous scheme with pseudo Maxwellian works in our formal analysis. However, there are infinitely many ways to construct M_p, and we are lack of theoretical foundation to make a good choice. To avoid such ambiguity, a natural question to ask is: is it possible to design a scheme without Maxwellian? The answer is yes, and it can be realized by a simple splitting on (3.1). That is, we split the transport part and the collision part, and then do a linear penalization on the second step:

$$\begin{cases} \dfrac{f^* - f^n}{\Delta t} + v \cdot \nabla_x f^n = 0, \\ \dfrac{f^{n+1} - f^*}{\Delta t} = \dfrac{\mathcal{Q}(f^*, f^*) + \lambda^* f^*}{\varepsilon} - \dfrac{\lambda^* f^{n+1}}{\varepsilon}. \end{cases} \tag{3.14}$$

A key fact about this scheme is that the collision step is conservative, so $U^{n+1} = U^*$, hence $M^{n+1} = M^*$ (i.e., we don't need the Maxwellian at all in this step).

We now formally prove the AP property. Some manipulation on (3.14) yields

$$\begin{aligned} f^{n+1} - M^{n+1} =& \frac{\varepsilon + D^* \Delta t}{\varepsilon + \lambda^* \Delta t}(f^n - M^n) - \frac{\varepsilon \Delta t}{\varepsilon + \lambda^* \Delta t}\left[v \cdot \nabla_x f^n + \frac{M^{n+1} - M^n}{\Delta t} \right] \\ &+ \frac{D^* \Delta t}{\varepsilon + \lambda^* \Delta t}[(f^* - f^n) - (M^* - M^n)], \end{aligned} \tag{3.15}$$

where D^* is defined analogously as in (3.4). Note that the last term in the above equation is $O(\Delta t)$ due to the consistency error. Therefore, for arbitrary initial data f^0, after a few time steps,

$$f^n - M^n = O(\Delta t), \tag{3.16}$$

if (3.5) is satisfied. On the other hand, if we add the two equations in (3.14) together and take the moments $\int \cdot \phi(v)\, \mathrm{d}v$, we again get (3.2), which upon substituting (3.16) becomes

$$\frac{U^{n+1} - U^n}{\Delta t} + \nabla_x \cdot \int v \phi(v) M^n \, \mathrm{d}v + O(\Delta t) = 0. \tag{3.17}$$

Thus a similar AP property is obtained as the scheme using the pseudo Maxwellian.

Remark 3.2. To get a better idea of how $O(\Delta t)$ term looks like in (3.17), noting that from scheme (3.14), we have (for simplicity we assume $\mathcal{Q}(f, f) = M - f$ and λ is a constant)

$$M^{n+1} + (\lambda - 1)(f^n - \Delta t v \cdot \nabla_x f^n) - \lambda f^{n+1} = O(\varepsilon). \tag{3.18}$$

Expanding f^{n+1} and M^{n+1} around time step n in Taylor series yields

$$M^n - f^n + \Delta t \left(\frac{\partial M^n}{\partial t} - (\lambda - 1) v \cdot \nabla_x f^n - \lambda \frac{\partial f^n}{\partial t} \right) + O(\Delta t^2) + O(\varepsilon) = 0. \tag{3.19}$$

The second and third f^n in above equation can be replaced by M^n due to (3.16), hence

$$f^n = M^n - \Delta t (\lambda - 1) \left(\frac{\partial M^n}{\partial t} + v \cdot \nabla_x M^n \right) + O(\Delta t^2) + O(\varepsilon). \tag{3.20}$$

Plugging this into (3.2), we get

$$\frac{U^{n+1} - U^n}{\Delta t} + \nabla_x \cdot \int v \phi(v) M^n \, \mathrm{d}v = \Delta t (\lambda - 1) \nabla_x \cdot \int v \phi(v) \left(\frac{\partial M^n}{\partial t} + v \cdot \nabla_x M^n \right) \mathrm{d}v + O(\Delta t^2) + O(\varepsilon). \tag{3.21}$$

The term $\int v \phi(v) \left(\frac{\partial M^n}{\partial t} + v \cdot \nabla_x M^n \right) \mathrm{d}v$ can be integrated out analytically thanks to the special form of the Maxwellian (in fact, it is nothing but the dissipation term that appears in the derivation of the Navier-Stokes limit of the BGK equation ([3]). For instance, in the one-dimensional case, (3.21) reads as

$$\frac{U^{n+1} - U^n}{\Delta t} + \left(\int v \phi(v) M^n \, \mathrm{d}v \right)_x = \Delta t (\lambda - 1) \begin{pmatrix} 0 \\ 0 \\ \left(\frac{3}{2} \rho T \right) T_x \end{pmatrix}_x + O(\Delta t^2) + O(\varepsilon). \tag{3.22}$$

Therefore, $O(\Delta t)$ term in (3.17) behaves like a diffusion operator if $\lambda > 1$ (recall that the AP condition is $\lambda > 1/2$). Compared to the scheme with Maxwellian penalization, we expect more numerical diffusion in this one (which is actually observed in our numerical results).

Performing a similar analysis on the scheme with the pseudo Maxwellian penalization, equation (3.13) with $O(\Delta t)$ term explicitly written down is

$$\frac{U^{n+1}-U^n}{\Delta t}+\nabla_x\cdot\int v\phi(v)M^n\,\mathrm{d}v=\Delta t\lambda\nabla_x\cdot\frac{\partial}{\partial t}\int v\phi(v)(f^n-M_p^n)\,\mathrm{d}v+O(\Delta t^2)+O(\varepsilon). \tag{3.23}$$

This may suggest that M_p is better chosen to also satisfy $\int v\phi(v)M_p\,\mathrm{d}v=\int v\phi(v)f\,\mathrm{d}v=0$. Namely, M_p matches f to one higher moment than a classical Maxwellian.

Remark 3.3. One may ask that can we do the linear penalty on the original unsplitted equation? That is,

$$\frac{f^{n+1}-f^n}{\Delta t}+v\cdot\nabla_x f^n=\frac{Q(f^n,f^n)+\lambda^n f^n}{\varepsilon}-\frac{\lambda^n f^{n+1}}{\varepsilon}, \tag{3.24}$$

for proper λ^n. This scheme does the job to remove the stiffness. However, because it is non-conservative on the right hand side, when taking the moments of (3.24), one arrives at

$$U^{n+1}=U^n-\frac{\varepsilon\Delta t}{\varepsilon+\lambda^n\Delta t}\nabla_x\cdot\int v\phi(v)f^n\,\mathrm{d}v. \tag{3.25}$$

As $\varepsilon\to 0$, this gives $U^{n+1}=U^n$. Thus the scheme is not AP! See numerical experiments in [15].

3.2. AP gPC-sG schemes for the uncertain Boltzmann equation

We now investigate the possibility of generalizing the previous deterministic AP schemes to the gPC-sG system (2.12).

3.2.1. The scheme with Maxwellian penalization

Starting from (2.12), our first scheme is a natural generalization of the deterministic AP scheme (3.1):

$$\frac{f_k^{n+1}-f_k^n}{\Delta t}+v\cdot\nabla_x f_k^n=\frac{Q_k^n-\lambda^n(M_k^n-f_k^n)}{\varepsilon}+\frac{\lambda^n(M_k^{n+1}-f_k^{n+1})}{\varepsilon}, \tag{3.26}$$

where f_k, M_k denote the gPC coefficients of $f(t,x,v,z)$ and $M(t,x,v,z)$ respectively. For simplicity, we choose the same λ for all $0\le k\le K$ (its choice will be discussed later). Although this scheme looks formally the same as before, an essential difference arises when it comes to the evaluation of M_k^{n+1}. Indeed, we take the following procedure:

1) Take the moments $\int\cdot\,\phi(v)\,\mathrm{d}v$ on both sides of (3.26):

$$\frac{U_k^{n+1}-U_k^n}{\Delta t}+\nabla_x\cdot\int v\phi(v)f_k^n\,\mathrm{d}v=0, \tag{3.27}$$

which gives $U_k^{n+1}=(\rho_k^{n+1},m_k^{n+1},E_k^{n+1})^T$, the gPC coefficients of ρ^{n+1}, m^{n+1} and E^{n+1}.

2) From ρ_k^{n+1}, reconstruct ρ^{n+1} by

$$\rho^{n+1}(x,z_j)=\sum_{k=0}^{K}\rho_k^{n+1}(x)\Phi_k(z_j), \tag{3.28}$$

where $\{z_j\}_{j=1}^{N_z}\in I_z$ are the Gauss quadrature points chosen according to probability $\pi(z)$. $m^{n+1}(x,z_j)$ and $E^{n+1}(x,z_j)$ are reconstructed similarly.

3) For each x and z_j, compute $u^{n+1}(x,z_j)$, $T^{n+1}(x,z_j)$ directly from $\rho^{n+1}(x,z_j)$, $m^{n+1}(x,z_j)$ and $E^{n+1}(x,z_j)$ using relation $m=\rho u$, $E=\rho\frac{d}{2}T+\frac{1}{2}\rho u^2$. Then construct $M^{n+1}(x,v,z_j)$ through (2.7) using $\rho^{n+1}(x,z_j)$, $u^{n+1}(x,z_j)$ and $T^{n+1}(x,z_j)$.

4) Decompose $M^{n+1}(x,v,z_j)$ to get $M_k^{n+1}(x,v)$ via

$$M_k^{n+1}(x,v)=\sum_{j=1}^{N_z}w_jM^{n+1}(x,v,z_j)\Phi_k(z_j), \tag{3.29}$$

where $\{w_j\}$ are the corresponding quadrature weights of $\{z_j\}$.

To show the AP property of this scheme, we need to show two properties. First, the uniform (in ε) stability, and second, the consistency to the Euler limit as $\varepsilon\to 0$.

The uniform stability was not rigorously proved even in the deterministic case [15], although it was shown for simplified models and verified numerically for the Boltzmann equation. Since the time discretization and penalization remain the same as that in [15], we expect the same uniform stability for the gPC-sG method proposed here, and will verify this numerically later.

To show the consistency to the Euler limit as $\varepsilon \to 0$, first we need to show that

$$f_k^n - M_k^n = O(\varepsilon). \tag{3.30}$$

For the Boltzmann collision operator this is difficult, even in the deterministic setting (see [15] and Remark 3.1). However, if the collision operator is the BGK operator, this can be formally shown, in the same way as outlined in section 3.1.1. Our numerical experiments in section 4 also verify that, after a few time steps, one has (3.30). Next, one needs to show that, when (3.30) is satisfied, the scheme becomes a consistent discretization of the limiting equation, as $\varepsilon \to 0$. Here, the limiting scheme is the gPC-sG approximation of the Euler equations (2.9). To this aim, we first write down the Euler equations (2.9) in vector form:

$$\frac{\partial U}{\partial t} + \nabla_x \cdot F(U) = 0, \tag{3.31}$$

where $F(U(t,x,z))$ is the flux. Then the gPC-sG for it is

$$\frac{\partial U_k}{\partial t} + \nabla_x \cdot \int_{I_z} F(U)\Phi_k(z)\pi(z)\,\mathrm{d}z = 0. \tag{3.32}$$

When (3.30) holds, substituting it to (3.27), one obtains

$$\frac{U_k^{n+1} - U_k^n}{\Delta t} + \nabla_x \cdot \int v\phi(v) M_k^n \,\mathrm{d}v + O(\varepsilon) = 0. \tag{3.33}$$

We just need to show that the flux in (3.33) is consistent to the flux in (3.32). Note that, by (3.29),

$$\int v\phi(v) M_k \,\mathrm{d}v = \int v\phi(v) \sum_{j=1}^{N_z} w_j M(x,v,z_j)\Phi_k(z_j)\,\mathrm{d}v \tag{3.34}$$

$$= \sum_{j=1}^{N_z} w_j \int v\phi(v) M(x,v,z_j)\mathrm{d}v\, \Phi_k(z_j) \tag{3.35}$$

$$= \sum_{j=1}^{N_z} w_j F(U(x,z_j))\Phi_k(z_j). \tag{3.36}$$

The last term in above equation is clearly the Gaussian quadrature rule for the flux in (3.32), thus (3.33), as $\varepsilon \to 0$, becomes a scheme consistent (modulus numerical error of the Gaussian quadrature rule) to the gPC-sG of the Euler equations. This, combined with the uniform stability, implies that our scheme is AP.

3.2.2. The scheme with pseudo Maxwellian penalization

Although the scheme with true Maxwellian penalization may give better AP property, computing M in the uncertain case can be very time-consuming or even impossible (as we will show later in numerical examples). Our second scheme is a generalization of the deterministic AP scheme (3.10),

$$\frac{f_k^{n+1} - f_k^n}{\Delta t} + v \cdot \nabla_x f_k^n = \frac{\mathcal{Q}_k^n - \lambda^n (M_{pk}^n - f_k^n)}{\varepsilon} + \frac{\lambda^n (M_{pk}^{n+1} - f_k^{n+1})}{\varepsilon}. \tag{3.37}$$

Recall that all we need for M_{pk} is that it has the same first $d+2$ moments as f_k: $\int M_{pk}\phi(v)\,\mathrm{d}v = \int f_k \phi(v)\,\mathrm{d}v = U_k$. We next give a simple way to construct M_{pk}^{n+1} given the moments U_k^{n+1} of f_k^{n+1} (superscript $n+1$ is omitted below for brevity and we take two-dimensional velocity space as an example).

For $0 \le k \le K$, define M_{pk} as

$$M_{pk}(x,v) = \rho_k(x)G_1(v) + m_k^{(1)}(x)G_2(v) + m_k^{(2)}(x)G_3(v) + E_k(x)G_4(v), \tag{3.38}$$

where ρ_k, $m_k^{(1)}$, $m_k^{(2)}$ and E_k are the four moments of f_k, and functions G_i satisfy

$$\int G_i(v)(1, v^{(1)}, v^{(2)}, v^2/2)^T \,\mathrm{d}v = e_i^T, \tag{3.39}$$

with e_i being the four-vector with 1 on the ith entry and zero elsewhere. We then assume G_i is a linear combination of four simple Maxwellians (each of the form (2.7)):

$$G_i(v) = b_{i1}M_{(1,(0,0),1)} + b_{i2}M_{(1,(1,0),1)} + b_{i3}M_{(1,(-1,0),1)} + b_{i4}M_{(1,(0,1),1)}. \tag{3.40}$$

It is not difficult to verify that in order to satisfy (3.39), the coefficients b_{ij} should take the value

$$(b_{ij})_{4\times 4} = \begin{pmatrix} 3 & -1 & -1 & 0 \\ 0 & \frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & -\frac{1}{2} & -\frac{1}{2} & 1 \\ -2 & 1 & 1 & 0 \end{pmatrix}. \tag{3.41}$$

The AP property of this scheme can be similarly argued as the previous subsection, except that (3.30) should be changed to

$$f_k^n - M_k^n = O(\Delta t), \tag{3.42}$$

as was discussed in section 3.1.2. We omit the details.

3.2.3. The linear penalty scheme without Maxwellian

As mentioned previously, the way of constructing pseudo Maxwellian M_{pk} is not unique and it would be nice for our scheme to be independent of this choice. Our third scheme is based on the deterministic AP scheme (3.14), whose gPC-sG counterpart is:

$$\begin{cases} \dfrac{f_k^* - f_k^n}{\Delta t} + v\cdot\nabla_x f_k^n = 0, \\ \dfrac{f_k^{n+1} - f_k^*}{\Delta t} = \dfrac{\mathcal{Q}_k^* + \lambda^* f_k^*}{\varepsilon} - \dfrac{\lambda^* f_k^{n+1}}{\varepsilon}. \end{cases} \tag{3.43}$$

The advantage of this scheme in the uncertain case should be evident by now. The AP property is similar to the scheme using the pseudo Maxwellian penalization, and we omit the details. Most importantly, it does not need the evaluation of any macroscopic quantities or Maxwellian, hence is much easier to implement!

3.2.4. The choice of λ

A remaining question is how to choose an appropriate λ in the above three AP gPC-sG schemes. Inspired by the deterministic case (see Remark 3.1), we consider the loss term of the collision operator under the Galerkin projection:

$$\mathcal{Q}_k^- = \sum_{i=0}^{K} a_{ki}(v) f_i(v), \tag{3.44}$$

where

$$a_{ki} := \sum_{j=0}^{K} S_{kij} \int_{\mathbb{R}^d}\int_{S^{d-1}} B(v - v_*, \sigma) f_j(v_*) \mathrm{d}\sigma \mathrm{d}v_*, \quad S_{kij} := \int_{I_z} \Phi_k(z)\Phi_i(z)\Phi_j(z)\pi(z)\,\mathrm{d}z.$$

Therefore, we take λ to be the largest eigenvalue of matrix $A = (a_{ki})$:

$$\lambda = \max|\mathrm{eig}(\sup_v A)|. \tag{3.45}$$

3.3. Discussion of the three AP gPC-sG schemes and their second-order extension

Based on our numerical experiments, the three AP gPC-sG schemes all perform well for continuous problems. The situation can be subtle for discontinuous problems (see section 4.5). First of all, the scheme with the Maxwellian penalization may break down since the oscillations in Gibb's phenomenon can cause negative density or temperature. As a result, the true Maxwellian is not well defined. In this aspect, the schemes using the pseudo Maxwellian or without Maxwellian are more robust. Second and surprisingly, the scheme with the pseudo Maxwellian penalization works for discontinuous data but appears to require much more restrictive CFL condition in practice. This is probably due to the particular choice of the pseudo Maxwellian.

The three AP schemes presented above are all first-order accurate in time. Following the deterministic case [15], the scheme with the Maxwellian penalization admits a straightforward second-order extension (provided M_k can be successfully constructed) via the so-called IMEX scheme. The scheme with the pseudo Maxwellian penalization can be extended to

second-order via backward differentiation formula (BDF) [19] but may suffer from divergence for certain initial data. Furthermore, it may also be subject to the same issue as its first-order counterpart for discontinuous data. The linear penalty scheme without the Maxwellian, albeit simple and robust for both continuous and discontinuous data, does not have an immediate second-order extension. Strang splitting seems to be a natural choice, but as pointed out in [21], it will degenerate to first order as $\varepsilon \to 0$. In this paper, we restrict to first-order (in time) schemes, and leave their second-order extension as future work.

4. Numerical examples

In this section, we present several numerical examples to illustrate the performance of the proposed AP gPC-sG schemes. We assume (unless otherwise specified):

- The random variable z obeys a uniform distribution on $[-1, 1]$, thus the Legendre polynomials are used as the gPC basis, and the Gauss-Legendre quadrature are adopted whenever numerical integration is needed.
- The spatial variable $x \in [0, 1]$ and periodic boundary condition is used except for the shock tube tests (section 4.5). The second-order MUSCL scheme [40] with minmod slope limiter is employed for spatial discretization.
- The velocity variable $v \in [-L_v, L_v]^2$ with $L_v = 8.4$. The fast spectral method proposed in [18] is applied to evaluate the collision operator $\mathcal{Q}_k$. For simplicity, we always use 32 points in each velocity dimension.
- The time step Δt is chosen based on the CFL condition from the transport part: $\Delta t = n_{\text{CFL}} \frac{\Delta x}{L_v}$, and $n_{\text{CFL}} < 1$ is the CFL number.
- The mean and standard deviation of the macroscopic quantities ρ, u and T are computed from their gPC coefficients ρ_k, u_k, T_k similarly as in (2.15), and these coefficients are obtained from f_k via (see [18] for details):
  - compute ρ_k, m_k, and E_k by direct integration

$$\rho_k = \int f_k \, \mathrm{d}v, \quad m_k = \int f_k v \, \mathrm{d}v, \quad E_k = \frac{1}{2} \int f_k |v|^2 \, \mathrm{d}v;$$

  - compute ρ_k^{-1} from

$$\rho \rho^{-1} = 1 \Rightarrow \left(\sum_{i=0}^{K} \rho_i \Phi_i \right) \left(\sum_{j=0}^{K} \rho_j^{-1} \Phi_j \right) = 1,$$

    which upon projection amounts to solving the linear system

$$\sum_{j=0}^{K} a_{kj} \rho_j^{-1} = \delta_{0k}$$

    with

$$a_{kj} = \sum_{i=0}^{K} \rho_i S_{kij}, \quad S_{kij} = \int_{I_z} \Phi_k(z) \Phi_i(z) \Phi_j(z) \pi(z) \, \mathrm{d}z;$$

  - compute u_k and T_k via

$$u_k^{(l)} = \sum_{i,j=0}^{K} m_i^{(l)} \rho_j^{-1} S_{kij}, \quad T_k = \sum_{i,j=0}^{K} \left(E_i \rho_j^{-1} - \frac{1}{2} \sum_{l=1}^{2} u_i^{(l)} u_j^{(l)} \right) S_{kij},$$

    where superscripts l denote the lth component of a 2-dimensional vector.

In the following, the AP gPC-sG scheme with the Maxwellian penalization introduced in section 3.2.1 is abbreviated as APGalwM; the AP gPC-sG scheme with the pseudo Maxwellian penalization in section 3.2.2 is abbreviated as APGalwPM; and the linear penalty AP gPC-sG scheme without Maxwellian in section 3.2.3 is abbreviated as APGalwoM.

Finally, in all numerical tests we consider $O(1)$ randomness in the initial data. This is in contrast to the theoretical discussion in Sections 2.3 and 2.4, where small randomness is assumed.

4.1. The AP property

We first numerically verify the AP property of the three schemes: namely, we take a very small $\varepsilon = 10^{-6}$, and check the distance between f and the corresponding Maxwellian. The following metric is used:

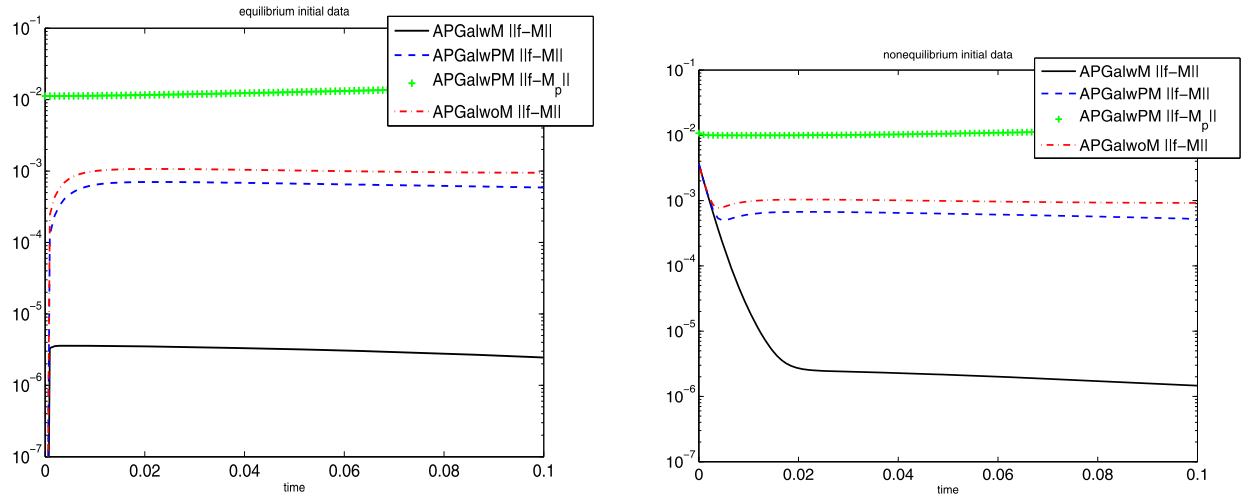

Fig. 1. $\|f - M\|_{L^2}$ versus time for different AP schemes. $\varepsilon = 1\text{e}{-6}$, $K = 7$, $N_x = 100$, $\Delta t = 9.5\text{e}{-4}$ ($n_{\text{CFL}} = 0.8$). Left: equilibrium initial data. Right: non-equilibrium initial data.

$$\|f(x, v, z)\|_{L^2} := \left(\iiint f(x, v, z)^2 \pi(z)\, \mathrm{d}x\, \mathrm{d}v\, \mathrm{d}z \right)^{\frac{1}{2}} \approx \left(\sum_{x,v,k} f_k^2(x, v) \Delta v^2 \Delta x \right)^{\frac{1}{2}}. \tag{4.1}$$

The random initial data is given by

$$\rho^0(x, z) = \frac{2 + \sin(2\pi x) + \frac{1}{2}\sin(4\pi x) z}{3}, \quad u^0 = (0.2, 0), \quad T^0(x, z) = \frac{3 + \cos(2\pi x) + \frac{1}{2}\cos(4\pi x) z}{4}, \tag{4.2}$$

and we consider both equilibrium and non-equilibrium initial distribution:

$$\text{(I)} \quad f^0(x, v, z) = \frac{\rho^0}{2\pi T^0} e^{-\frac{|v-u^0|^2}{2T^0}}, \tag{4.3}$$

$$\text{(II)} \quad f^0(x, v, z) = \frac{\rho^0}{4\pi T^0} \left(e^{-\frac{|v-u^0|^2}{2T^0}} + e^{-\frac{|v+u^0|^2}{2T^0}} \right). \tag{4.4}$$

The results are shown in Fig. 1. As expected, the distance between f and M is about $O(\varepsilon)$ for APGalwM, whereas the distance is about $O(\Delta t)$ for both APGalwPM and APGalwoM (this error will decrease as Δt decreases). To see that f is indeed driven to the true Maxwellian rather than the pseudo Maxwellian, we also plot $\|f - M_p\|$ for APGalwPM in the same figure.

4.2. The kinetic regime

We next take $\varepsilon = 1$ and consider initial data (4.2) (4.4). An explicit second-order (in both time and space) collocation scheme (20 Gauss-Legendre points are used in random space) is taken as a reference solution. In fact, this problem can be solved efficiently by the explicit scheme as the collision term is not stiff. Our purpose here is just to test the consistency of the proposed schemes in the kinetic regime. Results are shown in Fig. 2. The solutions of AP schemes all agree well with the reference solution.

4.3. The fluid regime

We now take $\varepsilon = 1\text{e}{-6}$, and check the accuracy of the schemes in the fluid regime. The initial data is given by (4.2) (4.4). This is a regime such that the explicit computation would be extremely expensive. Therefore, we compare with a second-order kinetic collocation scheme (a solver for the Euler system). The results are shown in Fig. 3 where we again observe good agreement. The solution by APGalwM is slightly better than APGalwPM and APGalwoM, mainly due to the $O(\Delta t)$ AP error in the latter two schemes.

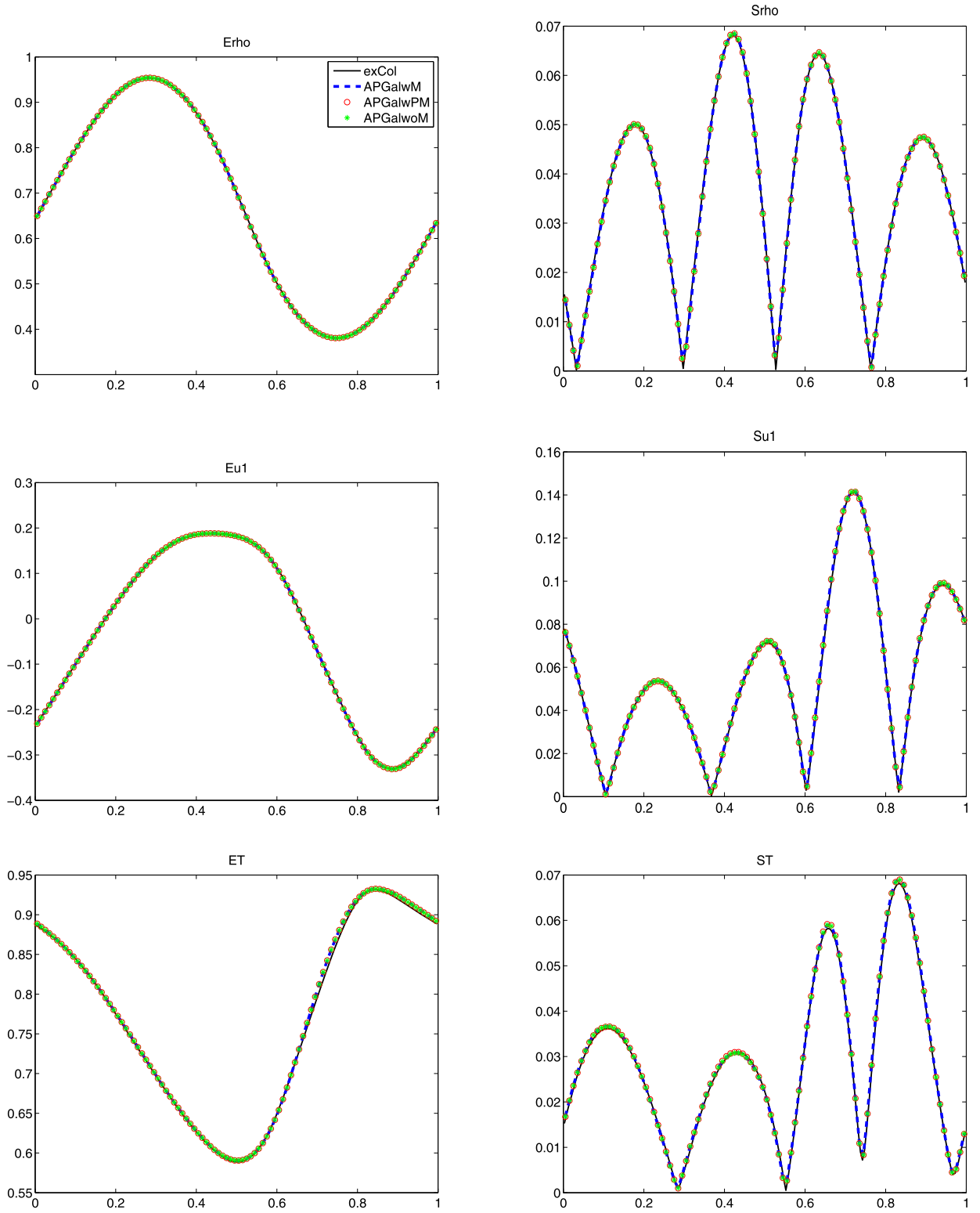

Fig. 2. Kinetic regime $\varepsilon = 1$. Solutions at $t = 0.1$. Left column: mean of ρ, u and T. Right column: standard deviation of ρ, u and T. Solid line: second-order explicit collocation with $N_z = 20$, $N_x = 200$, $\Delta t = 4.8\text{e}{-4}$ ($n_{\text{CFL}} = 0.8$). Dashed line: APGalwM. Circle: APGalwPM. Star: APGalwoM. For all three AP schemes $K = 7$, $N_x = 100$, $\Delta t = 9.5\text{e}{-4}$ ($n_{\text{CFL}} = 0.8$).

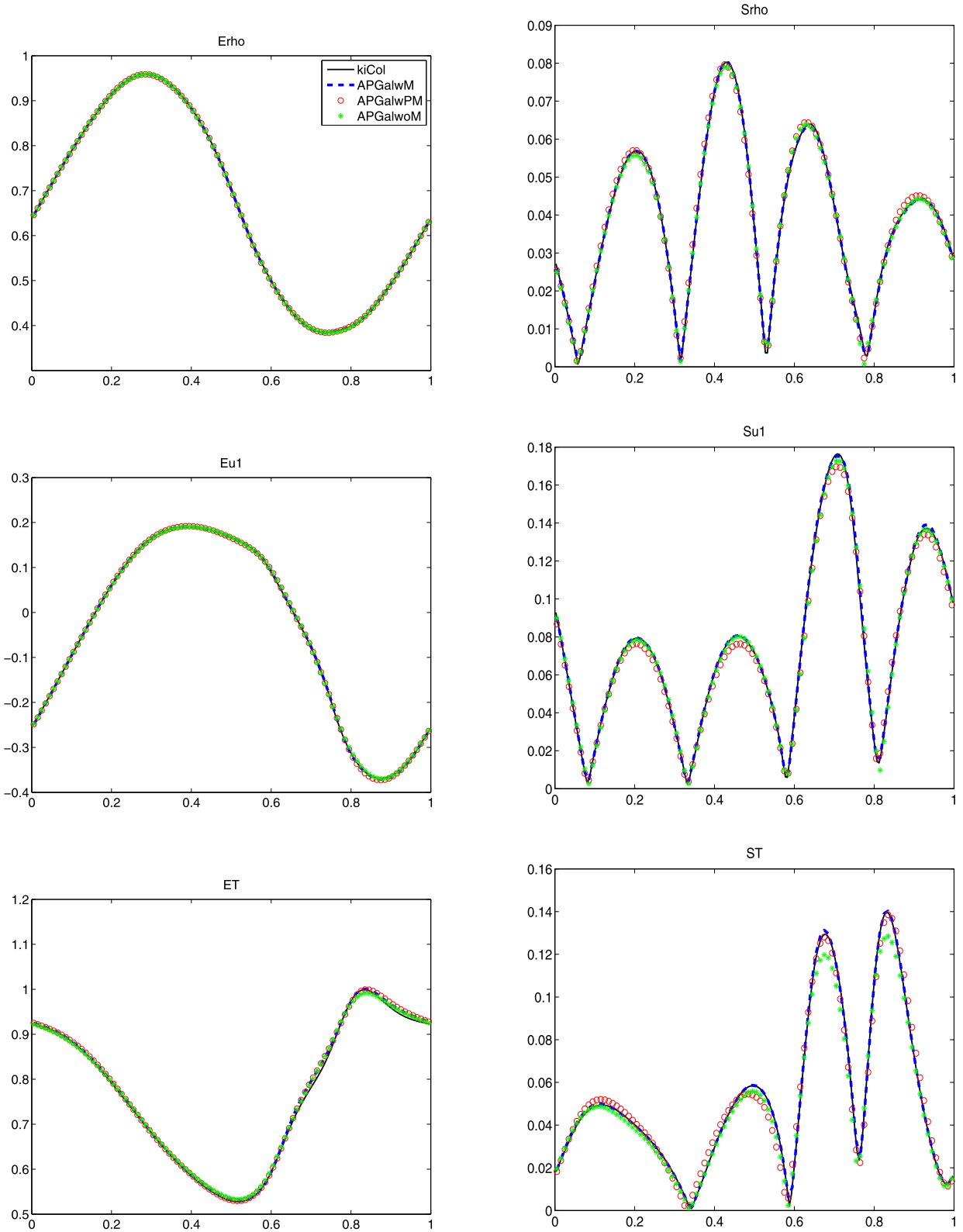

Fig. 3. Fluid regime $\varepsilon = 1e{-}6$. Solutions at $t = 0.1$. Left column: mean of ρ, u and T. Right column: standard deviation of ρ, u and T. Solid line: second-order kinetic collocation with $N_z = 20$, $N_x = 200$, $\Delta t = 4.8e{-}4$ ($n_{\mathrm{CFL}} = 0.8$). Dashed line: APGalwM. Circle: APGalwPM. Star: APGalwoM. For all three AP schemes $K = 7$, $N_x = 100$, $\Delta t = 9.5e{-}4$ ($n_{\mathrm{CFL}} = 0.8$).

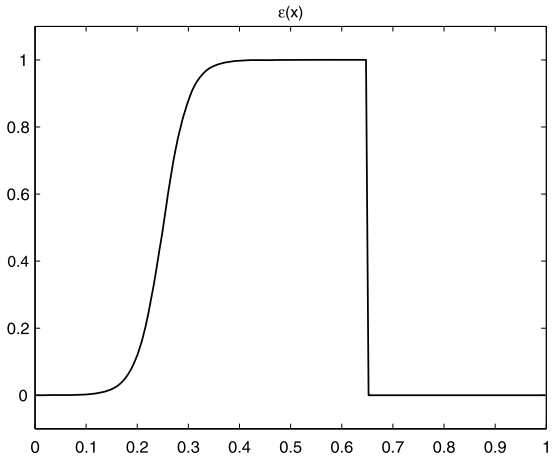

Fig. 4. A spatially varying Knudsen number $\varepsilon(x)$.

4.4. A mixed regime

In this example, we assume ε is varying in space:

$$\varepsilon(x) = \begin{cases} 1\text{e}{-4} + \dfrac{1}{2}(\tanh(25 - 20x) + \tanh(-5 + 20x)), & x \le 0.65, \\ 1\text{e}{-4}, & x > 0.65. \end{cases} \tag{4.5}$$

The profile of ε is depicted in Fig. 4, whose values range from 1e−4 to 1. The initial data is given by (4.2) (4.4). We compare with the explicit second-order collocation scheme. In order to get stability, explicit scheme requires Δt to be at least 1e−4 while AP schemes are only subject to CFL condition. The results are given in Fig. 5. For this mixed regime problem, APGalwM works well; APGalwPM seems to perform a little better than APGalwoM.

4.5. Shock tube problems

So far we have restricted to continuous initial data. In this last numerical test, we examine the performance of the three AP schemes for discontinuous initial data. The test we take is the classical shock tube problem. Two kinds of initial uncertainties are considered: case (I), uncertainty in the state variable; case (II), uncertainty in the location of the interface.

$$\text{(I)} \quad \begin{cases} \rho_l = 1 + 0.2\left(\dfrac{z+1}{2}\right), \quad u_l = (0,0), \quad T_l = 1, \quad x \le 0.5, \\ \rho_r = 0.125, \quad u_r = (0,0), \quad T_r = 0.25, \quad x > 0.5. \end{cases} \tag{4.6}$$

$$\text{(II)} \quad \begin{cases} \rho_l = 1, \quad u_l = (0,0), \quad T_l = 1, \quad x \le 0.5 + 0.05z, \\ \rho_r = 0.125, \quad u_r = (0,0), \quad T_r = 0.25, \quad x > 0.5 + 0.05z. \end{cases} \tag{4.7}$$

We mention that these two types of initial condition are very typical and often used for testing gPC-sG codes for uncertainty quantification of the Euler equations [37,36]. In particular, case (II) is harder in the sense that the initial condition is discontinuous in the z direction which causes a severe Gibb's phenomenon. As a result, the gPC approximated density or temperature will assume negative values immediately, and the conventional gPC-sG method fails even at the first step ([37])!

Now we tackle this problem from a kinetic point of view. When ε is small, we expect that the solution of AP schemes for the Boltzmann equation will be close to that of the Euler equations. Hence we take $\varepsilon = 1\text{e}{-6}$, and compare our AP schemes with a second-order kinetic collocation scheme for the Euler system. What we observe is the following: For case (I) (see Fig. 6), both APGalwM and APGalwoM perform well (the latter is a bit more smearing). APGalwPM, however, turns out to require a much smaller Δt to get a stable result for fixed Δx (n_{CFL} has to be about 0.1). For case (II), APGalwM completely breaks down. The reason is similar to that mentioned earlier: this method relies on the evaluation of macroscopic quantities, once density or temperature becomes negative at some point, the Maxwellian cannot be constructed. APGalwPM still needs a much smaller Δt for stability, thus we omit its results. APGalwoM performs reasonably well under a coarse mesh and the result can be improved by using a finer mesh (see Fig. 7).

Remark 4.1. For discontinuous data, it seems that the choice of pseudo Maxwellian can be quite sensitive. This strange behavior of APGalwPM deserves further investigation.

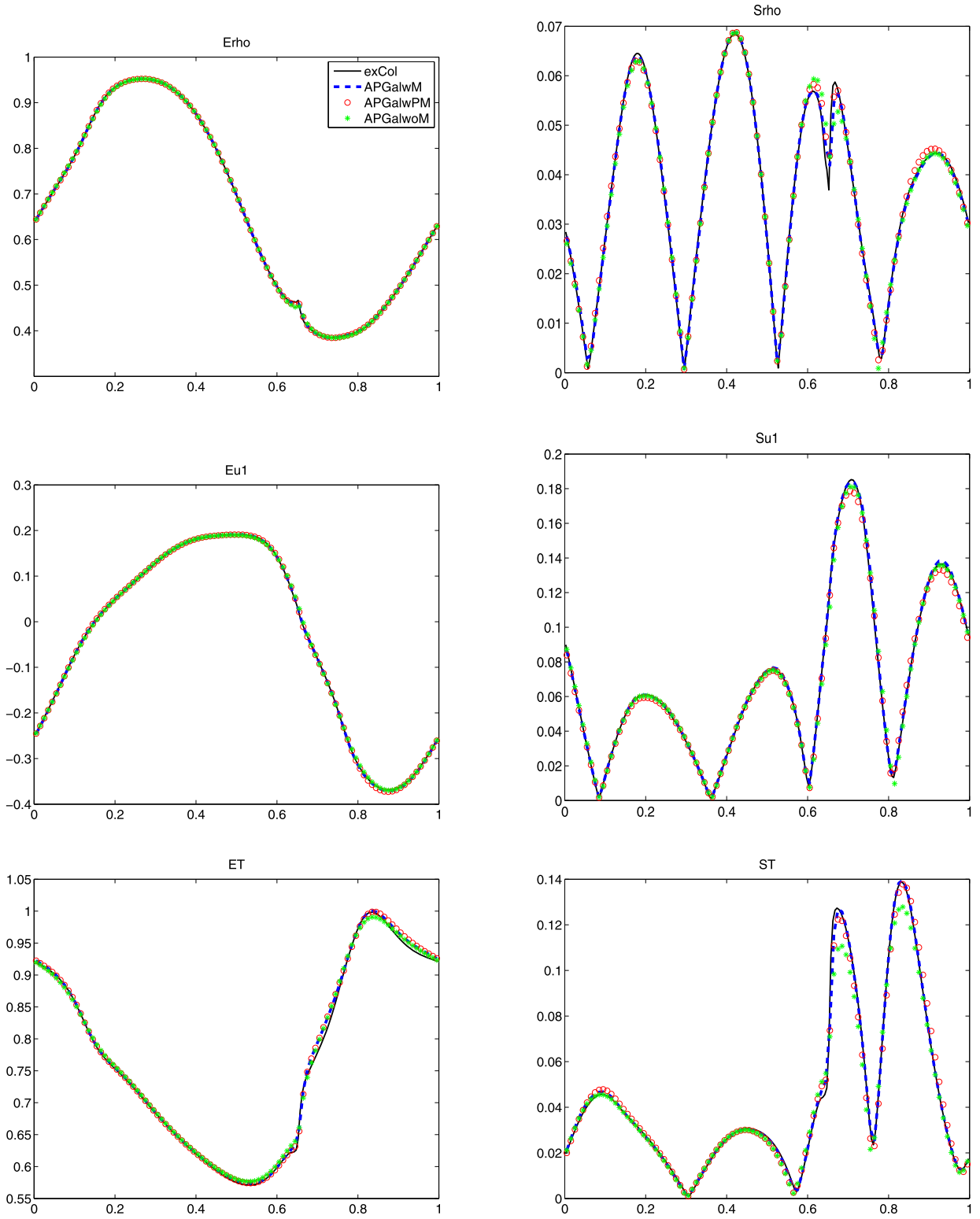

Fig. 5. Mixed regime. Solutions at $t = 0.1$. Left column: mean of ρ, u and T. Right column: standard deviation of ρ, u and T. Solid line: second-order explicit collocation with $N_z = 20$, $N_x = 200$, $\Delta t = 2\text{e}{-5}$. Dashed line: APGalwM. Circle: APGalwPM. Star: APGalwoM. For all three AP schemes $K = 7$, $N_x = 100$, $\Delta t = 9.5\text{e}{-4}$ ($n_{\text{CFL}} = 0.8$).

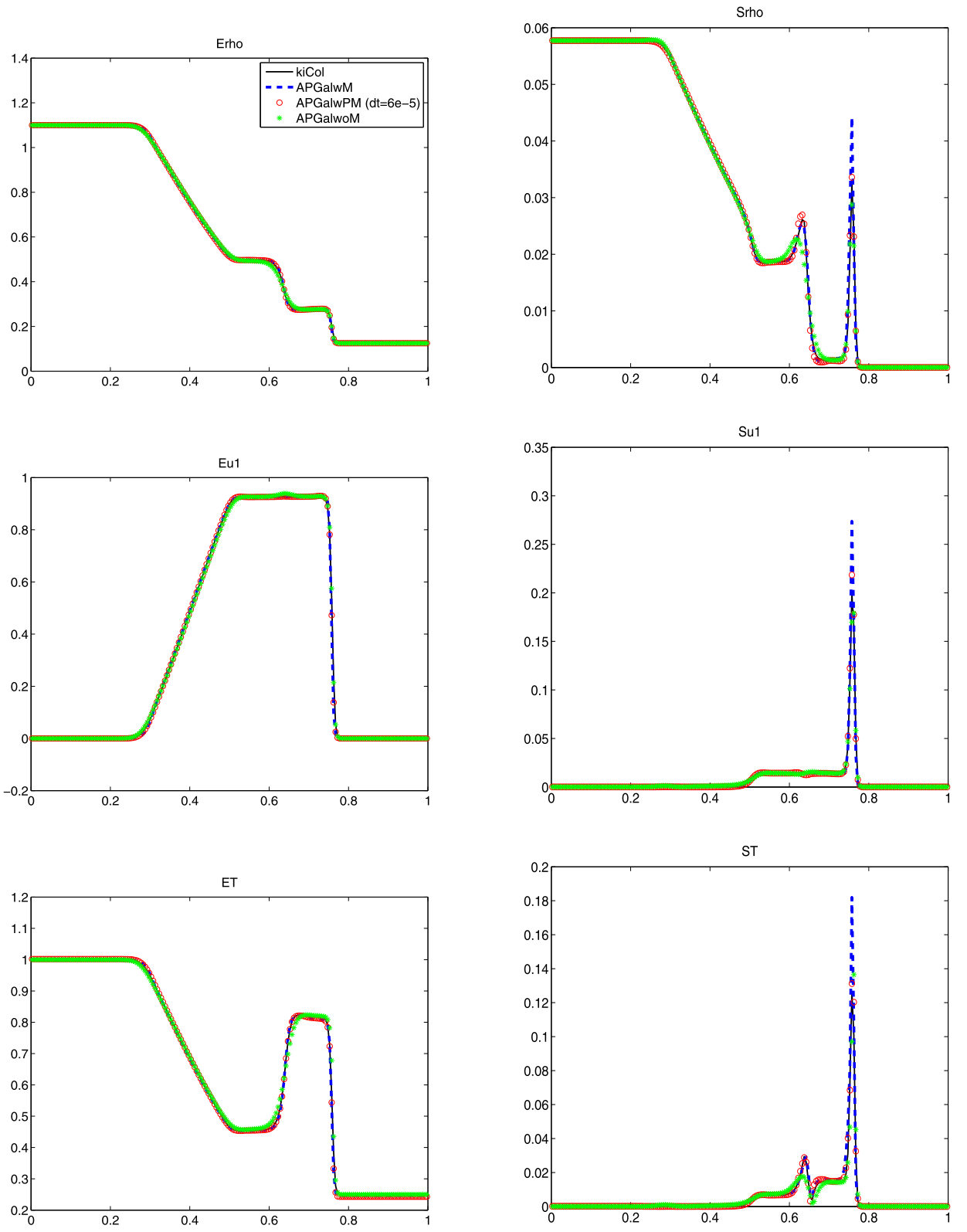

Fig. 6. Shock tube problem. Case (I). $\varepsilon = 1\text{e}{-6}$. Solutions at $t = 0.15$. Left column: mean of ρ, u and T. Right column: standard deviation of ρ, u and T. Solid line: second-order kinetic collocation with $N_z = 20$, $N_x = 200$, $\Delta t = 4.8\text{e}{-4}$ ($n_{\text{CFL}} = 0.8$). Dashed line: APGalwM. Circle: APGalwPM. Star: APGalwoM. For all three AP schemes $K = 7$, $N_x = 200$. $\Delta t = 4.8\text{e}{-4}$ ($n_{\text{CFL}} = 0.8$) for APGalwM and APGalwoM. $\Delta t = 6\text{e}{-05}$ ($n_{\text{CFL}} = 0.1$) for APGalwPM.

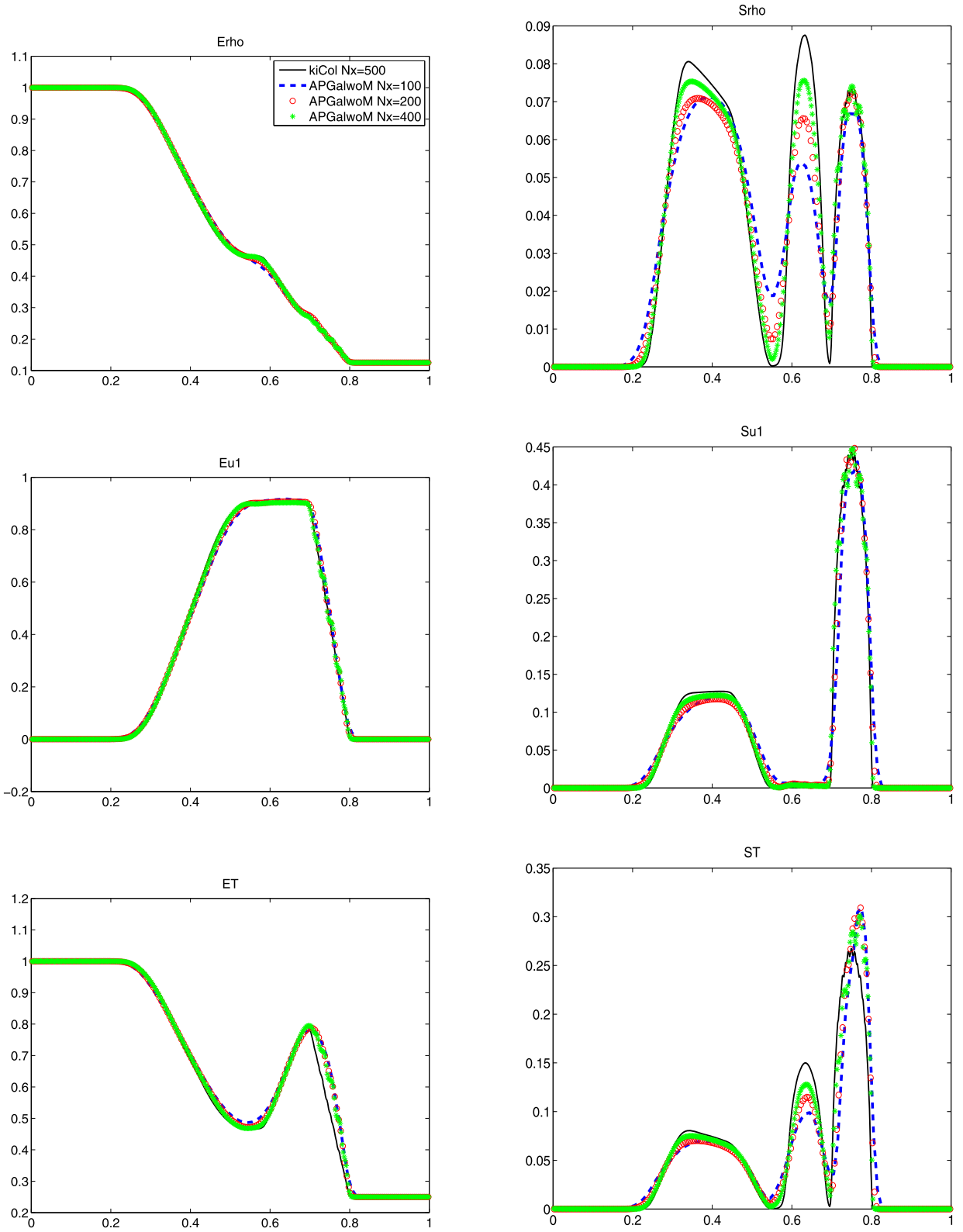

Fig. 7. Shock tube problem. Case (II). $\varepsilon = 1e{-}6$. Solutions at $t = 0.15$. Left column: mean of ρ, u and T. Right column: standard deviation of ρ, u and T. Solid line: second-order kinetic collocation with $N_z = 20$, $N_x = 500$, $\Delta t = 1.9e{-}4$ ($n_{\mathrm{CFL}} = 0.8$). APGalwoM with $K = 7$ and various N_x (fixed $n_{\mathrm{CFL}} = 0.8$).

Remark 4.2. When the global gPC basis is used for discontinuous data, the oscillations are inevitable and we do not attempt to address this issue in the current work. What we try to convey is that unlike the conventional gPC-sG (for the Euler equations) and APGalwM which relies crucially on the positivity of macroscopic quantities, the scheme APGalwoM only computes time evolution of the distribution function f_k itself and seems to be a more robust approach when dealing with problems such as case (II).

5. Conclusion

In this paper, we develop efficient stochastic Galerkin methods for the randomly uncertain nonlinear Boltzmann equation near the fluid dynamic regime. These schemes have the merit of being stochastic asymptotic-preserving, namely they allow Knudsen number independent polynomial chaos orders and time steps and yet, although having implicit collision terms, can be inverted quite easily. When the randomness is small, we show that the leading order of its fluid dynamical limit is weakly hyperbolic. Numerical examples were given which demonstrate the robustness of the proposed schemes, particularly in the compressible Euler regimes.

It remains an open question whether the fluid dynamic limit of the schemes gives rise to a hyperbolic discretization of the compressible Euler equations with uncertainty, for general random inputs, and to develop efficient schemes for high dimensional random inputs.

References

[1] R. Abgrall, P.M. Congedo, A semi-intrusive deterministic approach to uncertainty quantification in non-linear fluid flow problems, J. Comput. Phys. 235 (2013) 828–845.
[2] R. Abgrall, P.M. Congedo, G. Geraci, A one-time truncate and encode multiresolution stochastic framework, J. Comput. Phys. 257 (part A) (2014) 19–56.
[3] C. Bardos, F. Golse, D. Levermore, Fluid dynamic limits of kinetic equations. I. Formal derivations, J. Stat. Phys. 63 (1991) 323–344.
[4] T. Barth, Non-intrusive uncertainty propagation with error bounds for conservation laws containing discontinuities, in: Uncertainty Quantification in Computational Fluid Dynamics, in: Lect. Notes Comput. Sci. Eng., vol. 92, Springer, Heidelberg, 2013, pp. 1–57.
[5] P.R. Berman, J.E.M. Haverkort, J.P. Woerdman, Collision kernels and transport coefficients, Phys. Rev. A 34 (1986) 4647–4656.
[6] H. Bijl, D. Lucor, S. Mishra, C. Schwab (Eds.), Uncertainty Quantification in Computational Fluid Dynamics, Springer, 2013.
[7] G.A. Bird, Molecular Gas Dynamics and the Direct Simulation of Gas Flows, Clarendon Press, Oxford, 1994.
[8] C. Cercignani, The Boltzmann Equation and Its Applications, Springer-Verlag, New York, 1988.
[9] C. Cercignani, Rarefied Gas Dynamics: From Basic Concepts to Actual Calculations, Cambridge University Press, Cambridge, 2000.
[10] C. Cercignani, R. Illner, M. Pulvirenti, The Mathematical Theory of Dilute Gases, Springer-Verlag, 1994.
[11] S. Chapman, T.G. Cowling, The Mathematical Theory of Non-Uniform Gases, third edition, Cambridge University Press, Cambridge, 1991.
[12] Esther S. Daus, Shi Jin, Liu Liu, Spectral convergence of the stochastic Galerkin approximation to the Boltzmann equation with multiple scales and large random perturbation in the collision kernel, Kinet. Relat. Models 12 (2019) 909–922.
[13] G. Dimarco, L. Pareschi, Numerical methods for kinetic equations, Acta Numer. 23 (2014) 369–520.
[14] F. Filbet, J. Hu, S. Jin, A numerical scheme for the quantum Boltzmann equation with stiff collision terms, ESAIM: Math. Model. Numer. Anal. 46 (2012) 443–463.
[15] F. Filbet, S. Jin, A class of asymptotic-preserving schemes for kinetic equations and related problems with stiff sources, J. Comput. Phys. 229 (2010) 7625–7648.
[16] R.G. Ghanem, P.D. Spanos, Stochastic Finite Elements: A Spectral Approach, Springer-Verlag, New York, 1991.
[17] J. Hirschfelder, R. Bird, E. Spotz, The transport properties for non-polar gases, J. Chem. Phys. 16 (1948) 968–981.
[18] J. Hu, S. Jin, A stochastic Galerkin method for the Boltzmann equation with uncertainty, J. Comput. Phys. 315 (2016) 150–168.
[19] J. Hu, S. Jin, B. Yan, A numerical scheme for the quantum Fokker-Planck-Landau equation efficient in the fluid regime, Commun. Comput. Phys. 12 (2012) 1541–1561.
[20] J. Jakeman, R. Archibald, D. Xiu, Characterization of discontinuities in high-dimensional stochastic problems on adaptive sparse grids, J. Comput. Phys. 230 (2011) 3977–3997.
[21] S. Jin, Runge-Kutta methods for hyperbolic conservation laws with stiff relaxation terms, J. Comput. Phys. 122 (1995) 51–67.
[22] S. Jin, Asymptotic preserving (AP) schemes for multiscale kinetic and hyperbolic equations: a review, Riv. Mat. Univ. Parma 3 (2012) 177–216.
[23] S. Jin, J.-G. Liu, Z. Ma, Uniform spectral convergence of the stochastic Galerkin method for the linear transport equations with random inputs in diffusive regime and a micro-macro decomposition based asymptotic preserving method, Res. Math. Sci. 4 (15) (2017).
[24] S. Jin, L. Liu, An asymptotic-preserving stochastic Galerkin method for the semiconductor Boltzmann equation with random inputs and diffusive scalings, SIAM Multiscale Model. Simul. 15 (2017) 157–183.
[25] S. Jin, R. Shu, A stochastic asymptotic-preserving scheme for a kinetic-fluid model for disperse two-phase flows with uncertainty, J. Comput. Phys. 335 (2017) 905–924.
[26] S. Jin, D. Xiu, X. Zhu, Asymptotic-preserving methods for hyperbolic and transport equations with random inputs and diffusive scalings, J. Comput. Phys. 289 (2015) 35–52.
[27] Shi Jin, Hanqing Lu, Lorenzo Pareschi, Efficient stochastic asymptotic-preserving implicit-explicit methods for transport equations with diffusive scalings and random inputs, SIAM J. Sci. Comput. 40 (2) (2018) A671–A696.
[28] Shi Jin, Ruiwen Shu, A study of hyperbolicity of kinetic stochastic Galerkin system for the isentropic Euler equations with uncertainty, Chin. Ann. Math., Ser. B (2019), in press.
[29] Shi Jin, Yuhua Zhu, Hypocoercivity and uniform regularity for the Vlasov–Poisson–Fokker–Planck system with uncertainty and multiple scales, SIAM J. Math. Anal. 50 (2) (2018) 1790–1816.
[30] K. Koura, H. Matsumoto, Variable soft sphere molecular model for inverse-power-law or Lennard-Jones potential, Phys. Fluids A 3 (1991) 2459–2465.
[31] O. Le Maitre, O. Knio, H. Najm, R. Ghanem, Uncertainty propagation using Wiener-Haar expansions, J. Comput. Phys. 197 (2004) 28–57.
[32] O. Le Maitre, H. Najm, R. Ghanem, O. Knio, Multi-resolution analysis of Wiener-type uncertainty propagation schemes, J. Comput. Phys. 197 (2004) 502–531.
[33] Liu Liu, Shi Jin, Hypocoercivity based sensitivity analysis and spectral convergence of the stochastic Galerkin approximation to collisional kinetic equations with multiple scales and random inputs, Multiscale Model. Simul. 16 (3) (2018) 1085–1114.
[34] O.P. Le Maitre, O.M. Knio, Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics, Springer, 2010.

[35] M. Per Pettersson, G. Iaccarino, J. Nordstrom, Polynomial Chaos Methods for Hyperbolic Partial Differential Equations, Springer, 2015.
[36] P. Pettersson, G. Iaccarino, J. Nordstrom, A stochastic Galerkin method for the Euler equations with Roe variable transformation, J. Comput. Phys. 257 (2014) 481–500.
[37] G. Poette, B. Despres, D. Lucor, Uncertainty quantification for systems of conservation laws, J. Comput. Phys. 228 (2009) 2443–2467.
[38] Ruiwen Shu, Shi Jin, Uniform regularity in the random space and spectral accuracy of the stochastic Galerkin method for a kinetic-fluid two-phase flow model with random initial inputs in the light particle regime, ESAIM: Math. Model. Numer. Anal. 52 (5) (2018) 1651–1678.
[39] J. Tryoen, O. Le Maitre, M. Ndjinga, A. Ern, Intrusive Galerkin methods with upwinding for uncertain nonlinear hyperbolic systems, J. Comput. Phys. 229 (18) (2010) 6485–6511.
[40] B. van Leer, Towards the ultimate conservative difference scheme V. A second order sequel to Godunov's method, J. Comput. Phys. 32 (1979) 101–136.
[41] C. Villani, A review of mathematical topics in collisional kinetic theory, in: S. Friedlander, D. Serre (Eds.), Handbook of Mathematical Fluid Mechanics, vol. I, North-Holland, 2002, pp. 71–305.
[42] X. Wan, G.E. Karniadakis, An adaptive multi-element generalized polynomial chaos method for stochastic differential equations, J. Comput. Phys. 209 (2) (2005) 617–642.
[43] X. Wan, G.E. Karniadakis, Multi-element generalized polynomial chaos for arbitrary probability measures, SIAM J. Sci. Comput. 28 (2006) 901–928.
[44] D. Xiu, Numerical Methods for Stochastic Computations, Princeton University Press, New Jersey, 2010.