UNCERTAINTY QUANTIFICATION AND SENSITIVITY ANALYSIS FOR MULTISCALE KINETIC EQUATIONS WITH RANDOM INPUTS

by

Ruiwen Shu

A thesis submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Department of Mathematics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2018

Date of final oral examination: 04/26/2018

The dissertation is approved by the following members of the Final Oral Committee:

Shi Jin, Professor, Mathematics
Qin Li, Assistant Professor, Mathematics
Hung V. Tran, Assistant Professor, Mathematics
Carl Sovinec, Professor, Engineering Physics

ACKNOWLEDGMENTS

Firstly, I would like to thank my advisor, Prof. Shi Jin, for helping me with great patience in the past four years. He told me how to identify important questions in applied math, and how to attack those questions with persistence. Both ingredients will be indispensable in my future academic career.

Next, I would like to thank my family, especially my wife, for their thorough support. She really likes to discuss with me about my research, which gives me many nice ideas, due to her splendid intuition of mathematical models.

Finally, I would like to thank Prof. Jingwei Hu, Prof. Qin Li, Prof. Jian-Guo Liu, and all fellow students in Prof. Jin's group, for their sincere support and enlightening discussion with me.

TABLE OF CONTENTS

Page

Appendix

Page

LIST OF TABLES

LIST OF FIGURES

Appendix
Figure Page

ABSTRACT

This thesis gives an overview of the current results on uncertainty quantification and sensitivity analysis for multiscale kinetic equations with random inputs, with an emphasis on the author's contribution to this field.

In the first part of this thesis we consider a kinetic-fluid model for disperse two-phase flows with uncertainty in the fine particle regime. We propose a stochastic asymptotic-preserving (s-AP) scheme in the generalized polynomial chaos stochastic Galerkin (gPC-sG) framework, which allows the efficient computation of the problem in both kinetic and hydrodynamic regimes. The s-AP property is proved by deriving the equilibrium of the gPC version of the Fokker-Planck operator. The coefficient matrices that arise in a Helmholtz equation and a Poisson equation, essential ingredients of the algorithms, are proved to be positive definite under reasonable and mild assumptions. The computation of the gPC version of a translation operator that arises in the inversion of the Fokker-Planck operator is accelerated by a spectrally accurate splitting method. Numerical examples illustrate the s-AP property and the efficiency of the gPC-sG method in various asymptotic regimes.

In the second part of this thesis we consider the same kinetic-fluid model with random initial inputs in the light particle regime. Using energy estimates, we prove the uniform regularity in the random space of the model for random initial data near the global equilibrium in some suitable Sobolev spaces, with the randomness in the initial particle distribution and fluid velocity. By hypocoercivity arguments, we prove that the energy decays exponentially in time, which means that the long time behavior of the solution is insensitive to such

randomness in the initial data. Then we consider the gPC-sG method for the same model. For initial data near the global equilibrium and smooth enough in the physical and random spaces, we prove that the gPC-sG method has spectral accuracy, uniformly in time and the Knudsen number, and the error decays exponentially in time.

In the third part of this thesis we propose a stochastic Galerkin method using sparse wavelet bases for the Boltzmann equation with multi-dimensional random inputs. The method uses locally supported piecewise polynomials as an orthonormal basis of the random space. By a sparse approach, only a moderate number of basis functions is required to achieve good accuracy in multi-dimensional random spaces. We discover a sparse structure of a set of basis-related coefficients, which allows us to accelerate the computation of the collision operator. Regularity of the solution of the Boltzmann equation in the random space and an accuracy result of the stochastic Galerkin method are proved in multi-dimensional cases. The efficiency of the method is illustrated by numerical examples with uncertainties from the initial data, boundary data and collision kernel.

In the fourth part of this thesis we explore the possibility of using Generalized polynomial chaos (gPC) for uncertainty quantification in hyperbolic problems. GPC has been extensively used in uncertainty quantification problems to handle random variables. For gPC to be valid, one requires high regularity on the random space that hyperbolic type problems usually cannot provide, and thus it is believed to behave poorly in those systems. We provide a counter-argument, and show that despite the solution profile itself develops singularities in the random space, which prevents the use of gPC, the physical quantities such as shock emergence time, shock location, and shock width are all smooth functions of random variables in the initial data: with proper shifting, the solution's polynomial interpolation approximates with high accuracy. The studies were inspired by the stability results from hyperbolic systems. We use the Burgers' equation as an example for thorough analysis, and the analysis could be extended to general conservation laws with convex fluxes.

Chapter 1

Introduction

1.1 Kinetic equations

Kinetic equations are widely used in rarefied gas dynamics, plasma physics, biology, etc. They describe the behavior of particles in the mesoscopic scale, for which the number of particles is too large for particle models to be computationally affordable, and the material is far from local equilibrium so that hydrodynamic models are not accurate enough. Kinetic equations usually take the form

$$\partial_t f + v \cdot \nabla_x f + F \cdot \nabla_v f = \mathcal{Q}(f). \tag{1.1}$$

Here $t \in \mathbb{R}_+$, $x \in \Omega \subset \mathbb{R}^{d_x}$, $v \in \mathbb{R}^{d_v}$ are the time, space and velocity variables, respectively. $f = f(t, x, v)$ is the particle distribution function defined on the phase space, which represents the particle density at time t, space position x, and with velocity v. $F = F(x, t)$ is force on the particles, which may be external forces (e.g., gravity) or self-consistence forces (e.g., electric field generated by the particles themselves, in case the particles are charged). $\mathcal{Q}(f)$ is the collision operator, which describes the collisions between the particles and some background (for which $\mathcal{Q}$ is linear, usually written as $L(f)$), or between two particles (for which $\mathcal{Q}$ is quadratic, usually written as $Q(f, f)$). Typical examples of kinetic equations include

- The Boltzmann equation [18]

$$\partial_t f + v \cdot \nabla_x f = Q_B(f, f), \tag{1.2}$$

$$Q_B(f, f) = \int_{\mathbb{R}^{d_v}} \int_{\mathbb{S}^{d_v - 1}} B(v, v_*, \sigma) \left[f(v') f(v'_*) - f(v) f(v_*) \right] \, \mathrm{d}\sigma \, \mathrm{d}v_*. \tag{1.3}$$

f is the density of some uncharged particles. $Q_B(f,f)$ is the Boltzmann collision operator, describing the binary collisions between particles. v, v_* are the velocities of the two particles before collision, while v', v'_* are those after collision, given by

$$\begin{cases} v' = \dfrac{v+v_*}{2} + \dfrac{|v-v_*|}{2}\sigma, \\ v'_* = \dfrac{v+v_*}{2} - \dfrac{|v-v_*|}{2}\sigma, \end{cases} \tag{1.4}$$

where σ is the angle between relative pre-collision and post-collision outgoing velocities. $B(v, v_*, \sigma)$ is the cross section, which describes the frequency of collisions between a pair of particles with velocities v, v_* and with collision angle σ. A commonly used model for the collision kernel is the variable hard sphere (VHS) model [9], which takes the form

$$B = b|v - v_*|^\lambda, \tag{1.5}$$

where b and λ are some constants whose values are usually determined by matching with the experimental data to reproduce the correct transport coefficients such as the viscosity.

- The Fokker-Planck-Landau (FPL) equation (or Vlasov-Poisson-Landau equation) [65, 99]

$$\partial_t f + v \cdot \nabla_x f + E(t,x) \cdot \nabla_v f = Q_L(f,f), \tag{1.6}$$

f is the density of electrons in a plasma. $E(t,x)$ is the self-consistent electric field given by

$$E(t,x) = -\nabla_x \phi(t,x), \tag{1.7}$$

and $\phi(t,x)$ is a self-consistent electrostatic potential function satisfying the Poisson equation

$$\Delta_x \phi(t,x) = \mu(x) - \int_{\mathbb{R}^{d_v}} f(t,x,v)\,\mathrm{d}v, \tag{1.8}$$

where $\mu(x)$ is a neutralizing background satisfying

$$\int_{\mathbb{R}^{d_x}} \mu(x)\,\mathrm{d}x = \int_{\mathbb{R}^{d_x}} \int_{\mathbb{R}^{d_v}} f(t,x,v)\,\mathrm{d}v\,\mathrm{d}x. \tag{1.9}$$

$Q_L(f,f)$ on the right-hand side of (1.6) is the FPL collision operator (or Landau collision operator) that models binary interactions among particles:

$$Q_L(f,f)(v) = \nabla_v \cdot \int_{\mathbb{R}^{d_v}} A(v-v_*)[f(v_*)\nabla_v f(v) - f(v)\nabla_{v_*} f(v_*)]\,\mathrm{d}v_*. \tag{1.10}$$

Here $A(w)$ is a semi-positive definite matrix defined by

$$A(w) = \Psi(w)\left(I - \frac{w\otimes w}{|w|^2}\right), \tag{1.11}$$

where I is the identity matrix. For inverse-power law potentials, $\Psi(w) = |w|^{\gamma+2}$ with $-3 \le \gamma \le 1$. The case $\gamma = -3$ corresponds to the Coulomb interaction which is of primary importance in plasma applications.

- Kinetic-fluid two-phase flow models. One example is the following system [40, 41]:

$$\begin{cases} \partial_t f + v\cdot\nabla_x f - \nabla_x\Phi\cdot\nabla_v f = \mathcal{L}_u f, \\ \partial_t u + \nabla_x\cdot(u\otimes u) + \nabla_x p - \dfrac{1}{Re}\Delta_x u = \kappa\displaystyle\int (v-u) f\,\mathrm{d}v, \\ \nabla_x\cdot u = 0, \end{cases} \tag{1.12}$$

  where $\mathcal{L}_u f$ is the Fokker-Planck (FP) operator

$$\mathcal{L}_u f = \nabla_v\cdot((v-u)f + \nabla_v f). \tag{1.13}$$

  This model describes disperse two-phase flows, in which the primary phase is fluid, whose velocity field u satisfies the incompressible Navier-Stokes equations, while the secondary phase is particles, whose distribution function f satisfies a kinetic equation. Here Re is the Reynolds number, κ is a coupling constant describing the mass ratio of the two phases, and Φ is a external potential (e.g., gravity). The drag force between the two phases is assumed to satisfy the Stokes law, i.e., proportional to their relative velocity.

We remark that the general form (1.1) does not include all kinetic equations: for example, aggregation models in biology and social science [47, 80] often have collision operators which are nonlocal in x; models from quantum mechanics [28] may involve collisions that are not binary; etc. In this paper we will only focus on kinetic equations in the form (1.1).

1.1.1 Basic properties of kinetic equations

We briefly summarize some basic properties of kinetic equations. First of all, most collision operators satisfy conservation properties, which is a direct consequence of the physical conservation laws of particle collisions. Such conservation property takes the form

$$\langle \mathcal{Q}(f)\phi(v)\rangle = 0, \quad \forall f = f(v), \tag{1.14}$$

for some given function $\phi(v)$, called a collision invariant. Here $\langle\cdot\rangle$ denotes the integral in v. For example, the Fokker-Planck operator $\mathcal{L}_u f$ in (1.13) satisfies the mass conservation property, which corresponds to $\phi(v) = 1$; the Boltzmann and FPL collision operators satisfy the mass, momentum and energy conservation, which correspond to $\phi(v) = 1, v, |v|^2$, respectively. As a result of (1.14), if one multiplies (1.2) by $\phi(v)$ and integrate in v, one obtains

$$\partial_t \langle f\phi(v)\rangle + \nabla_x \cdot \langle fv\phi(v)\rangle = 0, \tag{1.15}$$

which means that the quantity $\langle f\phi(v)\rangle$ (as a function of x) is transported by the flux $\langle fv\phi(v)\rangle$. In the cases $\phi(v) = 1, v, |v|^2$, this quantity is nothing but the macroscopic quantities (density, momentum, energy) at position x. If one further integrates (1.15) in x and adds proper boundary conditions (e.g., periodic boundary), one concludes that the total amount of the macroscopic quantities $\int \langle f\phi(v)\rangle \,\mathrm{d}x$ do not change in time.

Another important property of collision operators is the entropy dissipation property. For the Boltzmann and FPL collision operators, it takes the form

$$\langle Q_B(f, f)\ln f\rangle \le 0, \quad \forall f > 0, \tag{1.16}$$

which is called the Boltzmann H-theorem. We remark that the Fokker-Planck operator also satisfies a similar entropy dissipation property, see [17]. If one multiplies (1.2) by $(\ln f + 1)$ and integrate in v, one obtains

$$\partial_t \langle f\ln f\rangle + \nabla_x \cdot \langle vf\ln f\rangle \le 0, \tag{1.17}$$

which means that the entropy $\langle f\ln f\rangle$ is transported by the entropy flux $\langle vf\ln f\rangle$ and may decrease due to collision. If one further integrates (1.17) in x and adds proper boundary

conditions, one concludes that the total entropy $\int \langle f \ln f \rangle$ is non-increasing in time. This is nothing but the Second Law of Thermodynamics. We remark that the same conclusion holds for the FPL collision operator.

For the FPL equation (1.6) and the two-phase flow model (1.12), one can use the same manipulation to obtain similar conservation properties and entropy dissipation properties. We omit the details here, and refer to Section 3.1 for those properties of (1.12).

1.1.2 Hydrodynamic limits

After doing nondimensionalization of the kinetic equation (1.1), one often obtains equations with dimensionless parameters. Usually one would like to take a specific scaling and reduce to the case where there is only one dimensionless parameter, the Knudsen number ϵ, which describes the strength of collision: smaller ϵ means collisions are more frequent. In the case of the Boltzmann equation, ϵ is defined as the ratio between the mean free path and the typical length scale. Two typical scalings are the hyperbolic scaling and diffusive scaling.

The hyperbolic scaling assumes that the collision is strong and keeps the time scale at $O(1)$. For (1.2), the hyperbolic scaling, usually called the compressible Euler scaling [8], takes the form

$$\partial_t f + v \cdot \nabla_x f = \frac{1}{\epsilon} Q_B(f, f). \tag{1.18}$$

For (1.12), a similar scaling is the fine particle regime [41], which takes the form

$$\begin{cases} \partial_t f + v \cdot \nabla_x f - \nabla_x \Phi \cdot \nabla_v f = \dfrac{1}{\epsilon} \mathcal{L}_u f, \\ \partial_t u + \nabla_x \cdot (u \otimes u) + \nabla_x p - \dfrac{1}{Re} \Delta_x u = \dfrac{1}{\epsilon} \kappa \displaystyle\int (v - u) f \, \mathrm{d}v, \\ \nabla_x \cdot u = 0. \end{cases} \tag{1.19}$$

The diffusive scaling assumes that the collision is strong and takes the long time scale $O(\frac{1}{\epsilon})$. For (1.2), the diffusive scaling, usually called the incompressible Navier-Stokes scaling [8], takes the form

$$\epsilon \partial_t f + v \cdot \nabla_x f = \frac{1}{\epsilon} Q_B(f, f). \tag{1.20}$$

For (1.12), a similar scaling is the light particle regime [40], which takes the form

$$\begin{cases} \partial_t f + \dfrac{1}{\epsilon} v \cdot \nabla_x f - \nabla_x \Phi \cdot \nabla_v f = \dfrac{1}{\epsilon^2} \mathcal{L}_{\epsilon u} f, \\ \partial_t u + \nabla_x \cdot (u \otimes u) + \nabla_x p - \dfrac{1}{Re} \Delta_x u = \dfrac{1}{\epsilon} \kappa \displaystyle\int (v - \epsilon u) f \, \mathrm{d}v, \\ \nabla_x \cdot u = 0. \end{cases} \tag{1.21}$$

In the regime where collision is very strong, i.e., as $\epsilon \to 0$, f will achieve local equilibrium, i.e., $\mathcal{Q}(f) \approx 0$, and the behavior of f can be effectively approximated by macroscopic hydrodynamic equations. This is called the hydrodynamic limit of kinetic equations. One can formally derive this limit by the Chapman-Enskog expansion.

For (1.18), let $\epsilon \to 0$, one formally has $Q_B(f, f) = 0$, which implies

$$f = M(v)_{(\rho,u,T)} = \frac{\rho}{(2\pi T)^{d_v/2}} e^{-\frac{|v-u|^2}{2T}}, \tag{1.22}$$

which is called the Maxwellian. ρ, u and T are the density, bulk velocity and temperature, given by

$$\rho(t,x) = \int f \, \mathrm{d}v, \quad u(t,x) = \frac{1}{\rho} \int f v \, \mathrm{d}v, \quad T(t,x) = \frac{1}{d_v \rho} \int f |v-u|^2 \, \mathrm{d}v. \tag{1.23}$$

Notice that ρ, u, T are determined explicitly by the moments $\int f \phi(v) \, \mathrm{d}v$, $\phi = 1, v, |v|^2$ against the collision invariants of Q_B. Then, taking moments of (1.18) against the collision invariants and expressing $f = M(v)_{(\rho,u,T)}$ in terms of the macroscopic quantities, one obtains the compressible Euler equations

$$\begin{cases} \rho_t + \nabla_x \cdot (\rho u) = 0, \\ (\rho u)_t + \nabla_x \cdot (\rho u \otimes u + pI) = 0, \\ E_t + \nabla_x \cdot ((E+p)u) = 0, \end{cases} \tag{1.24}$$

for the macroscopic quantities, where $E = \frac{1}{2}\rho u^2 + \frac{d_v}{2}\rho T$ is the total energy, and $p = \rho T$ is the pressure. See [8] for more details about this limit.

For (1.19), we first define the moments

$$n(t,x) = \int f(t,x,v) \, \mathrm{d}v, \quad J(t,x) = \int v f(t,x,v) \, \mathrm{d}v, \quad P(t,x) = \int v \otimes v f(t,x,v) \, \mathrm{d}v. \tag{1.25}$$

Let $\epsilon \to 0$, one formally has $\mathcal{L}_u f = 0$, which implies

$$f(t,x,v) = n(t,x)M_u(v), \tag{1.26}$$

where

$$M_u(v) = \frac{1}{(2\pi)^{d_v/2}} \exp(-\frac{|v-u(t,x)|^2}{2}), \tag{1.27}$$

is the local Maxwellian. This implies

$$J = nu, \quad P = nu \otimes u + nI, \tag{1.28}$$

where I is the identity matrix. Integrating the first equation of (1.19) in v against $1, v$, one has

$$\partial_t n + \nabla_x \cdot J = 0, \tag{1.29}$$

$$\partial_t J + \nabla_x \cdot P = -\frac{1}{\epsilon} \int (v-u) f \, \mathrm{d}v + n\nabla_x \Phi. \tag{1.30}$$

Then multiplying (1.30) by κ and adding it to the second equation of (1.19), using (1.28), one gets

$$\begin{cases} \partial_t n + \nabla_x \cdot (nu) = 0, \\ \partial_t((1+\kappa n)u) + \nabla_x((1+\kappa n)u \otimes u) + \nabla_x(p + \kappa n) + \kappa n \nabla_x \Phi = \dfrac{1}{Re} \Delta_x u, \end{cases} \tag{1.31}$$

which is the incompressible Navier-Stokes equations with variable density $(1+\kappa n)$. See [41] for more details about this limit.

For (1.20), one can derive the hydrodynamic limit as the incompressible Navier-Stokes-Fourier equations for the macroscopic quantities. For (1.21), one can derive the hydrodynamic limit as the incompressible Navier-Stokes equations for u and a convection-diffusion equation for the particle density n. We refer to [8] and [40] respectively for details.

Rigorous justification of hydrodynamic limits have been done for many kinetic equations in the last two decades, based on entropy dissipation estimates for weak solutions [37, 40, 41] for example, or energy estimates for strong solutions in the perturbative regime [29]. A further discussion on this topic is out of the scope of this paper.

1.1.3 Asymptotic-preserving (AP) schemes

Due to the stiffness of the collision term, a direct numerical discretization of the multi-scale kinetic equations with a small Knudsen number ϵ usually requires stability constraint $\Delta t = O(\epsilon)$, which is prohibitively restrictive if ϵ is very small. Asymptotic-preserving (AP) schemes [52] are those numerical schemes that are stable even if $\Delta t >> \epsilon$, and as $\epsilon \to 0$, the scheme automatically becomes a consistent discretization of the limiting hydrodynamic equations. With the AP property, one can use a fixed set of numerical parameters (Δt, Δx, Δv) to solve effectively a stiff kinetic equation with all sizes of ϵ, ranging from $O(1)$ to very small, because as $\epsilon \to 0$, the scheme automatically captures the hydrodynamic limit of the kinetic equation. In other words, the numerical parameters can be chosen *independent of* ϵ.

We first take the Boltzmann equation in the hyperbolic scaling (1.18) as an example. If one could afford to treat the collision operator implicitly, then the forward-backward Euler scheme

$$\frac{f^{n+1} - f^n}{\Delta t} + v \cdot \nabla_x f^n = \frac{1}{\epsilon} Q_B(f^{n+1}, f^{n+1}), \tag{1.32}$$

is AP. In fact, as $\epsilon \to 0$, one formally has $Q_B(f^{n+1}, f^{n+1}) = 0$, which implies f^{n+1} is a Maxwellian for any n. Then taking moments against $1, v, |v|^2$ one obtains a kinetic scheme for the limiting Euler equations with forward Euler time discretization. However, treating the Boltzmann collision operator implicitly requires nonlinear iterations, which can be very expensive due to the complexity of this operator. [30] proposed an AP scheme to avoid this implicit treatment, based on the idea of penalization. They use

$$\frac{f^{n+1} - f^n}{\Delta t} + v \cdot \nabla_x f^n = \frac{1}{\epsilon}(Q_B(f^n, f^n) - \beta P(f^n) + \beta P(f^{n+1})), \tag{1.33}$$

where $P(f) = M[f] - f$ is the BGK operator, $M[f]$ being the Maxwellian with the same moments of f against $1, v, |v|^2$. In this way they can make sure f^{n+1} is driven to the Maxwellian, without treating Q_B implicitly. Similar idea is used in [59] to deal with the FPL equation, using a Fokker-Planck operator as penalization operator. Another method to avoid the implicit treatment of Q_B is the exponential Runge-Kutta method [70], for which we omit the details here.

For the two-phase flow model in the fine particle regime (1.19), an AP scheme is proposed in [42], based on a combination of the projection method for the incompressible Navier-Stokes equation and an implicit treatment for the stiff Fokker-Planck operator.

AP schemes for equations with diffusive scalings usually require a reformulation of the equation before doing numerical discretizations. Two popular approaches are the even-odd decomposition and the micro-macro decomposition, see [56] and [67] respectively as examples.

Once a first order AP scheme is obtained, one can use implicit-explicit (IMEX) Runge-Kutta or multi-step methods [86] to extend to higher order AP schemes.

1.2 Uncertainty quantification (UQ) for kinetic equations with random inputs

Most previous works on kinetic equations only consider the deterministic case. However, there are many sources of *uncertainty* in kinetic equations:

- The initial data and boundary data often come from experiments, thus have measurement error.
- The parameters inside the model may have uncertainty, since many models are empirical, and the parameters are obtained by matching with experimental data. For example, the variable hard sphere model (1.5) is an empirical model with two parameters b and λ. Such parameter has uncertainty due to experimental error.

To provide reliable predictions and a guidance to improve the model, it is imperative to incorporate these uncertainties into the system, and quantify these uncertainties by numerically solving the resulting system with uncertain inputs.

To model the uncertainties, we introduce a random parameter z lying in a random space I_z with probability distribution $\pi(z)\,\mathrm{d}z$. The uncertainties are modeled by letting the random quantities depend on z: for example, a random initial data is given by $f_{in} = f_{in}(x, v, z)$, and a random collision kernel [50] may appear as $B = b(z)|v - v_*|^\lambda$, etc. As a result, the solution

also depends on z. Then, to quantify the effect of the uncertainties on the solution, one needs to solve the same kinetic equation with an extra random parameter z.

We summarize some popular numerical methods for uncertainty quantification (UQ) [34, 45, 75, 102, 103]: the first one is Monte-Carlo (MC) methods [83], which take random samples in I_z, solve the deterministic problem on these samples, and then get the statistical moments by taking the average on these samples. MC methods are half-order accurate for any dimensional random spaces, and thus they are not accurate enough for low dimensional random spaces, but very efficient for high-dimensional random spaces. The second method is stochastic collocation (sC) methods [4, 7, 84, 104], which take sample points on a well-designed grid (quadrature points, sparse grids, or by some optimization procedure), compute the deterministic solutions on the samples, and then reconstruct the solution in the whole random domain by some interpolation rules. SC methods can achieve good accuracy in low dimensional random spaces, but the efficiency drops as the dimension becomes high. The third method is stochastic Galerkin (sG) methods [7, 5, 105], which takes an orthonormal basis in the random domain, approximate the functions by a truncated sG series, i.e., the generalized Fourier series with respect to this basis, and then obtain a deterministic system of equations on the sG coefficients via the Galerkin projection. SG methods are as accurate as sC methods for low dimensional random spaces, and behave better than sC for moderately high dimensional random spaces if one wants to achieve high accuracy [7]. In this paper we will focus on sG methods.

1.2.1 Stochastic Galerkin (sG) methods

The sG method takes a set of orthonormal basis functions $\{\phi_k(z)\}_{k=1}^{\infty}$ with respect to the probability measure $\pi(z)\,\mathrm{d}z$, and expands any function depending on z into an sG series. For example, the solution $f = f(t, x, v, z)$ is expanded as

$$f(t, x, v, z) = \sum_{k=1}^{\infty} f_k(t, x, v)\phi_k(z), \tag{1.34}$$

where $f_k(t,x,v) = \int f(t,x,v,z)\phi(z)\pi(z)\,\mathrm{d}z$ are the sG coefficients, which *no longer depends on* z. Then one approximates f by truncating this series:

$$f(t,x,v,z) \approx f^K(t,x,v,z) := \sum_{k=1}^{K} f_k(t,x,v)\phi_k(z). \tag{1.35}$$

Using this ansatz in the kinetic equation and conducting the Galerkin projection, one obtains a *deterministic* system for the coefficients $f_1, \ldots, f_K$. Then one can design numerical schemes to solve this deterministic system. Finally, the statistical information of the solution can be obtained from the coefficients. For example, if the first basis function is a constant function $\phi_1(z) = 1$, then the mean and variance can be computed by

$$\mathbb{E}(f) = f_1, \quad \mathrm{var}(f) = \sum_{k=2}^{K} f_k^2. \tag{1.36}$$

A popular choice of the basis functions is the generalized polynomial chaos (gPC) basis [105]. When the random space I_z is one-dimensional, the gPC basis is the set of polynomials defined on I_z, orthonormal with respect to the given probability measure $\pi(z)\,\mathrm{d}z$, with ϕ_k being a polynomial of degree $k-1$. GPC basis functions for higher dimensional random spaces are the tensor product basis functions. The biggest advantage of the gPC basis is that it can achieve *spectral accuracy* for low dimensional random spaces, if the function to approximate is *smooth*. When the dimension becomes moderately high, one can still truncate the gPC basis in a proper way (total degree, hyperbolic cross, etc.) [7] in order to achieve good accuracy with a moderately large number of basis functions.

1.2.2 Stochastic asymptotic-preserving (s-AP) schemes

For gPC-sG methods for multiscale kinetic equations with random inputs, a desirable property is the stochastic asymptotic-preserving (s-AP) [58], which means that as the Knudsen number $\epsilon \to 0$, the gPC-sG method for the kinetic equation automatically becomes a consistent gPC-sG approximation for the limiting hydrodynamic system. Similar to the deterministic AP property, the s-AP property enables one to use a fixed set of numerical parameters, including the number of basis functions K, to effectively solve the multiscale

kinetic equation with uncertainty for all sizes of ϵ, i.e., K can be chosen *independent of* ϵ. Apart from the linear transport equation considered in [58], s-AP schemes for other kinetic equations have been developed in the recent two years [54, 57, 110, 55].

Since the gPC coefficients $\{f_k\}_{k=1}^K$ satisfy a system of deterministic kinetic equations, the key to check the s-AP property is to derive (at least formally) a hydrodynamic limit for this system. To do this, one needs to know the *null space* of the gPC version of the stiff collision operator, i.e., what one can conclude for $\{f_k\}$ from

$$\int \mathcal{Q}(f^K)\phi_k(z)\pi(z)\,\mathrm{d}z = 0, \quad k = 1, \ldots, K. \tag{1.37}$$

If (1.37) implies that $\{f_k\}$ is some equilibrium form determined by some moments, then (in the cases of hyperbolic scalings) one can take the moments of the kinetic system to obtain a hydrodynamic system for the moments of $\{f_k\}$. For most of the known cases, this system is indeed a consistent gPC-sG approximation of the hydrodynamic limit of the original equation. After deriving the hydrodynamic limit of the kinetic gPC system, one needs to design a (deterministic) AP scheme to capture this limit. This can be done by mimicking the AP scheme for the original deterministic kinetic equation.

In the case where $\mathcal{Q}(f) = \sigma(z)L(f)$ where L is a deterministic linear operator and $\sigma(z) > 0$ is given, (1.37) implies

$$\sum_{j=1}^{K} \left(\int \sigma(z)\phi_j(z)\phi_k(z)\pi(z)\,\mathrm{d}z \right) L(f_j) = 0, \quad k = 1, \ldots, K. \tag{1.38}$$

It is easy to show that the matrix whose (k, j) term is $\left(\int \sigma(z)\phi_j(z)\phi_k(z)\pi(z)\,\mathrm{d}z\right)$ is a symmetric positive-definite matrix, thus one concludes $L(f_k) = 0$, $k = 1, \ldots, K$, which simply means f_k is in the null space of the deterministic operator L, for each k. This is indeed the cases in [58].

The case of the Fokker-Planck operator (1.13) is a little harder. In the author's work with S. Jin [57], we show that when $\mathcal{Q}(f) = \mathcal{L}_u(f)$ where $u = u(z)$ is given, (1.37) implies

$$\vec{f}(v) = \vec{\mathcal{T}}_{\vec{u}}(M_0\vec{C}), \tag{1.39}$$

where we use the vector notation $\vec{f} = (f_1, \dots, f_K)^T$, $M_0 = e^{-|v|^2/2}$ is the Maxwellian centered at 0, and $\vec{C}$ is any constant vector. The vector-valued translation operator $\vec{\mathcal{T}}_{\vec{u}}$ is defined by

$$\vec{\mathcal{T}}_{\vec{u}}(\vec{f}) = \exp(-A^{(1)}\partial_{v_1} - A^{(2)}\partial_{v_2})(\vec{f}), \quad A^{(i)}_{jk} = \sum_{j,k=1}^{K} \int u^{(i)}(z)\phi_j(z)\phi_k(z)\,\mathrm{d}z, \tag{1.40}$$

in the case where v is two-dimensional, $u^{(i)}$ standing for the i-th component of u.

We remark that the problem of designing an s-AP scheme for the Boltzmann equation and the FPL equation is still open. One difficulty is that the implication of (1.37) for the Boltzmann or FPL collision operator is unknown. Another difficulty is that (in the hyperbolic scaling) a direct gPC approximation of the limiting Euler equations may fail to be hyperbolic [23].

1.3 Sensitivity analysis for kinetic equations with random inputs

For kinetic equations with uncertainty, in order to understand the sensitivity of the solution on the random inputs, one needs to estimate $\partial_z f$. In order to guarantee the spectral accuracy of gPC based methods, including sG and sC methods, it is necessary to give estimates for $\partial_z^k f$, $k = 1, 2, \dots$, which provides the spectral accuracy of the polynomial approximations in the random space.

1.3.1 Energy/hypocoercivity estimates

In the recent two years, energy/hypocoercivity estimates have been adopted to estimate $\partial_z^k f$ for kinetic equations with uncertainty, for both linear [53, 71, 54, 73] and nonlinear [60, 95, 74] kinetic equations. To illustrate the idea, we first consider a linear kinetic equation with uncertainty

$$\partial_t f + v \cdot \nabla_x f = \sigma(z) L(f), \tag{1.41}$$

where L is a deterministic linear collision operator and $\sigma(z) \geq \sigma_0 > 0$. Multiplying by f and integrating in x, v, adopting a proper boundary condition (e.g., periodic boundary), one

obtains

$$\frac{1}{2}\partial_t \|f\|_{L^2}^2 = \sigma(z) \int \int L(f) f \, \mathrm{d}v \, \mathrm{d}x. \tag{1.42}$$

Many linear kinetic operators have a finite-dimensional null space (which is the local equilibrium), with a spectral gap estimates

$$\int L(f) f \, \mathrm{d}v \leq -c \|f^{\perp}\|_{L^2}^2, \tag{1.43}$$

where $f^{\perp}$ is the projection of f onto the orthogonal complement of the null space of L. This implies

$$\frac{1}{2}\partial_t \|f\|_{L^2}^2 \leq -c\sigma_0 \|f^{\perp}\|_{L^2}^2. \tag{1.44}$$

This estimate means the L^2 norm of f is non-increasing. However, in order to obtain the *exponential decay in time* of f, which is an essential long-time behavior, one has to find a dissipation term for the macroscopic quantities, i.e., the projection onto the null space of L. For many collision operators, hypocoercivity theory [100] uses a carefully designed Lyapunov functional $|||\cdot|||$, equivalent to the H^1 norm, such that there is indeed an estimate

$$\frac{1}{2}\partial_t |||f|||^2 \leq -c_1 |||f|||^2, \tag{1.45}$$

which gives the exponential decay of f.

In order to estimate the z-derivatives of f, we take ∂_z^k on (1.41) and get

$$\partial_t \partial_z^k f + v \cdot \nabla_x \partial_z^k f = \sigma(z) L(\partial_z^k f) + \mathcal{S}_k, \tag{1.46}$$

where the source term

$$\mathcal{S}_k = \sum_{j=1}^{k} \binom{k}{j} \partial_z^j \sigma L(\partial_z^{k-j} f), \tag{1.47}$$

only involves the z-derivatives of f of lower order. Conducting energy estimate for $\partial_z^k f$, this source term can be controlled by adding suitable multiples of the energy estimates of $\partial_z^j f$, $j = 1, \ldots, k-1$, see [53]. Then, combined with a hypocoercivity estimate, one can show the exponential decay of f.

For kinetic equations with multiscale features (small Knudsen number ϵ), one can prove uniform-in-ϵ energy estimates for $\partial_z^k f$, if the size of the initial data is suitable with respect to

ϵ ($O(1)$ for some cases and $O(\epsilon)$ for other cases). Most of the scaling issues can be handled in the same way as the deterministic case. We refer to [74] for more details on the scaling issues for a large class of collisional kinetic equations.

For nonlinear kinetic equations, the same method can be applied to prove the exponential decay of solution as well as its z-derivatives, if the initial data is close enough to the global equilibrium [60, 95, 74]. In this case, for any nonlinear term $\Gamma(f, f)$, as long as there is an estimate

$$\|\Gamma(f, g)\| \leq C\|f\|\|g\|, \tag{1.48}$$

in a suitable space (e.g., some Sobolev space), one will be able to absorb this term by the dissipation from the linear terms, under the small data assumption. The author's work with S. Jin [95] uses this method to prove the uniform-in-ϵ random space regularity of the two-phase flow model in the light particle regime (1.21) under smallness conditions, and exponential decay of $\partial_z^k f$ is proved via hypocoercivity.

1.3.2 Spectral accuracy of the gPC-sG method

The spectral accuracy of gPC-sC methods follows directly from the random space regularity. However, for gPC-sG methods, apart from the interpolation error, there is also an error from the Galerkin projection, thus require further analysis to justify its accuracy.

The spectral accuracy of the gPC-sG method for the linear equation (1.41) is proved in [53], in which the linear equation satisfied by the error $f - f^K$ is analyzed. To deal with nonlinear equations, it is desirable to have a small initial data condition *independent of* the numerical parameter K, which means that the accuracy results are true for a fixed set of initial data, for all K. If one directly uses (1.48) to estimate the terms $\int \Gamma(f^K, f^K)\phi_k(z)\,\mathrm{d}z$, $k = 1, \ldots, K$, then one will end up with the terms

$$C|S_{ijk}|\|f_i\|\|f_j\|, \quad i, j, k = 1, \ldots, K, \tag{1.49}$$

which are a large number (K^3) of terms involving the gPC coefficients $f_1, \dots, f_K$. Here S_{ijk} is the basis related triple product coefficient

$$S_{ijk} = \int \phi_i \phi_j \phi_k \pi(z) \, \mathrm{d}z. \tag{1.50}$$

This will give an estimate which requires the smallness condition depending on K.

To overcome this difficulty, the author's work with S. Jin [95] introduced a weighted sum of the Sobolev norm of the gPC coefficients to control the nonlinear terms with an estimate independent of K. Instead of estimating a naive summation of the norms of f_k, we estimate

$$\sum_{k=1}^{K} \|k^q f_k\|^2, \tag{1.51}$$

where we assume that the basis functions satisfy a technical condition $|S_{ijk}| \leq k^p$, and the constant q satisfies $q > p + 2$. This extra weight enables us to combine the terms (1.49) together as part of a convergent series, and obtain an estimate independent of K. Using this technique, we prove the uniform-in-(t, ϵ) spectral accuracy of the gPC-sG method for the two-phase flow model in the light particle regime (1.21) under smallness conditions independent of K. Later this technique is adopted by [74] to prove the spectral accuracy of the gPC-sG methods for a class of collisional kinetic equations.

1.3.3 Random space regularity for the Vlasov-Poisson equation

For the Vlasov-Poisson equation

$$\partial_t f + v \cdot \nabla_x f + E \cdot \nabla_v f = 0, \quad E = F[f] = -\nabla W * \rho, \quad \rho = \int f \, \mathrm{d}v, \tag{1.52}$$

where W is a given interacting potential and $x \in \mathbb{T}^d$, there does not hold any energy dissipation estimate due to the *time-reversibility*. However, if the initial data is close enough to some specific spatial homogeneous equilibrium and has enough regularity, [82] showed that the long time behavior of the solution exhibits the *Landau damping* behavior [64], i.e., the electric field E decays exponentially in time, and the particle distribution f is well approximated by the free transport of some fixed profile. Then, it is natural to ask whether the long time behavior of the Landau damping solution is sensitive to initial perturbations.

In the author's work with S. Jin [94], we proved a first result in this direction: there exists initial data for the Vlasov-Poisson equation which depending on z smoothly, such that this random space regularity persists for all time. To be precise, $\partial_z^k E$ decays exponentially in time, and $\partial_z^k f$ converges exponentially fast to the free transport of some fixed profile. Our result is based on an earlier deterministic Landau damping result [16]. Currently we are working on the random space regularity for all initial data near a linearly-stable spatial homogeneous equilibrium, based on [82].

1.4 Other related topics

1.4.1 UQ for kinetic equations with high-dimensional random inputs

Uncertainty quantification for kinetic equations with high-dimensional random inputs is hard due to the fact that deterministic kinetic equations are already high-dimensional (3d in x and 3d in v). In the author's work with J. Hu and S. Jin [93], we did a first attempt in this direction. For the Boltzmann equation with multi-dimensional random inputs, we adopt the sparse wavelet basis [101], which uses locally supported piecewise polynomials and a Smolyak type construction. Using this basis, we conduct the sG method for the Boltzmann equation, and accelerate the computation of the collision part by observing a sparsity structure in the basis related coefficients S_{ijk} defined in (1.50), which is a consequence of the local support of the basis functions. Numerical experiments suggest that this method has convergence order higher than the Monte Carlo method, and is effective in at least 6 dimensional random space.

1.4.2 UQ for hyperbolic equations with discontinuous solutions

The spectral accuracy of gPC based UQ methods relies on the random space regularity, which may fail in some cases, for example, nonlinear hyperbolic equations and kinetic equations with hyperbolic scaling and small Knudsen number. It is very challenging to design numerical methods which have good accuracy for such problems. In the author's work with Q. Li and J.-G. Liu [69], we did a first attempt in this direction. We consider the Burgers

equation

$$\partial_t u + \partial_x(\frac{1}{2}u^2) = 0, \tag{1.53}$$

with random inputs, and we assume the initial data satisfies a few conditions so that only one shock is developed for each value of z. In this special case, we proved that the shock location depends smoothly on z, if the initial data does. The main idea of the proof is to analyze the ODE satisfied by the upper and lower boundary of the shock. This result enables us to use gPC based interpolation methods to obtain the shock location as well as the solution away from the shock with high accuracy.

1.5 Outline of this thesis

In the rest of this thesis, we go through some of the author's results: Chapter 2 is the author's work with S. Jin [57] on the s-AP method for the two-phase flow model in the fine particle regime (1.19); Chapter 3 is our work [95] on the random space regularity and spectral accuracy of the gPC-sG method for the two-phase flow model in the light particle regime (1.21); Chapter 4 is the author's work with J. Hu and S. Jin [93] on the sparse wavelet based sG method for the Boltzmann equation with multi-dimensional random inputs; Chapter 5 is the author's work with Q. Li and J.-G. Liu [69] on the polynomial interpolations for the Burgers equation with random inputs. The thesis is concluded in Chapter 6.

Chapter 2

S-AP method for the two-phase flow model in the fine particle regime

In this chapter we go through the author's work with S. Jin [57] on the s-AP method for the two-phase flow model in the fine particle regime (1.19).

2.1 Introduction

We are concerned with kinetic-fluid models for disperse two-phase flows. Such models arise naturally in the study of mixture of a continuum of fluid, such as gas and liquids, and small particles, such as droplets and suspension of solids. The fluid phase is described by hydrodynamic equations, such as the Euler equations or the Navier-Stokes equations, while the particle phase is described by a kinetic equation. The application of kinetic-fluid models includes the dynamic of sprays [27, 85, 51], granular flows [3, 36, 26], and combustion theory [38, 88], to name a few.

We assume the following physical assumptions:

1. The primary phase is liquid or dilute gas, and therefore modeled by the incompressible Navier-Stokes equations.

2. The secondary phase is small particles (or droplets, bubbles), scattered in the fluid, and it is modeled by a kinetic equation.

3. The interaction between the two phases is assumed to be the Stokes drag force, i.e., a particle is subject to a force proportional to the relative velocity between it and the fluid.

4. The particles are assumed to be subject to the Brownian motions.

There are two scalings that are physically important: one is the light particle regime [40], which assumes:

1. The velocity of the fluid is small compared to the typical molecular velocity of the particles.

2. The particles are light, and thus its effect on the fluid is small.

3. The relaxation time is much smaller than the typical time scale.

Another one is the fine particle regime [41], which assumes:

1. The particle size is small compared to the typical length scale.

2. The density of the fluid and particles are of the same order.

3. The relaxation time is much smaller than the typical time scale.

In this chapter we focus on the fine particle regime with two-dimensional physical space. The model is given by (1.19) where the Fokker-Planck operator $\mathcal{L}_f f$ is defined by (1.13). $x = (x_1, x_2) \in \Omega \subset \mathbb{R}^2$ is the space variable, and $v = (v_1, v_2) \in \mathbb{R}^2$ is the velocity variable. $f = f(t, x, v)$ is the density function of the particles. $u = u(t, x) = (u^{(1)}(t, x), u^{(2)}(t, x))$ is the velocity field of the fluid. $\Phi = \Phi(x)$ is an external force field. The first equation of (1.19) describes the motion of particles. The two terms in the FP operator comes from the drag force from the fluid and the effect of Brownian motions, respectively. The $\nabla_x \Phi$ term is the effect of the external force field on the particles. The second and third equations of (1.19) are the standard incompressible Navier-Stokes equations for the fluid, with the right-hand-side term describing the force coming from the particle. $\kappa > 0$ is the coupling constant depending

on the typical mass ratio between the particles and the fluid, and Re is the Reynolds number. ϵ is the Knudsen number given by $\epsilon = \frac{2\rho_P a^2}{9\mu}$, where μ is the dynamic viscosity of the fluid, a the typical radius of the particles, and ρ_P the density of the particles.

The hydrodynamic limit of this model was first investigated by Goudon et al. [40, 41] in the light particle regime and the fine particle regime, respectively. In [42] Goudon et al. proposed an Asymptotic-preserving (AP) [52] scheme for the two-phase flow system (1.19), which are efficient for both the cases of small and large ϵ. The main idea of this work is to incorporate the evolution of the moments of the particles into the projection method [20, 98] for the NS system. The possibly stiff (when ϵ is small) FP operator is treated fully implicitly, with a well-balanced spatial discretization proposed by Jin and Yan [59]. The second order time discretization is given by the backward difference.

The paper [42] only concerns with the case where all the physical quantities and parameters are deterministic. However, there are many sources of uncertainties in this model. For example, the initial data of f and u come from experimental measurements, hence may have measurement errors. If one adopts the Maxwellian boundary condition for f with the accommodation coefficient, or the no-slip boundary condition for u against a wall with a non-zero velocity, then these boundary data will contain parameters, which come from direct measurements or matching with experimental data. Such parameters will also give rise to uncertainties. Furthermore, the parameters ϵ, κ, Re and the external field Φ come from measurements and have uncertainties. In [] we proposed an s-AP scheme for the two-phase flow model (1.19). In order to simplify the presentation and emphasize the main idea, we only consider the case of uncertain initial data. Uncertainties from other terms can be treated similarly in the sG framework, see [110].

Compared with the deterministic problem in [42], there are several new difficulties to overcome. First, the formal proof of the s-AP property is less obvious, due to the vector form of the scheme. The way to overcome this difficulty is already mentioned in Section 1.2.2. Second, one needs to show that the resulted Helmholtz and Poisson systems, essential ingredients of the s-AP schemes, are well-defined systems. Indeed these properties, which are

based on the positive-definiteness of the coefficient matrices in these systems, will be proven under reasonable and mild assumptions. Thirdly, to treat $\vec{\mathcal{L}}$ implicitly, which is needed for good numerical stability property, it is necessary to compute the translation operator $\vec{\mathcal{T}}$, which is very expensive if computed directly. We accelerate this computation dramatically by using a spectrally accurate splitting together with the Fast Fourier Transform. Then the problem is reduced to the inversion of $\mathcal{L}_0$, which can be easily and efficiently computed by the method for the deterministic Fokker-Planck system in [59]. As a result, we just need to diagonalize *two* matrices of size K in each time step, while the direct method needs to diagonalize N_v^2 such matrices, where K is the number of basis functions and N_v the number of mesh points in the one-dimensional velocity space.

This chapter is organized as follows: in Section 2.2 we briefly review the deterministic problem; in Section 2.3 we propose the gPC approximation to the problem with uncertain initial data, and prove the s-AP property of the gPC-sG method; in Section 2.4 we give the fully discrete s-AP scheme, prove the positive-definiteness of two coefficient matrices, and give a spectrally accurate splitting method for the translation operator $\vec{\mathcal{T}}$; in Section 2.5 we demonstrate the s-AP property and the efficiency of the gPC-sG method by some numerical examples.

2.2 The Deterministic Problem

In this section we briefly review the results for the deterministic problem. The hydrodynamic limit of the system (1.19) was proved in [41]. Since we already explained the formal derivation of this limit in Section 1.1.2, here we only give a derivation of the null space of $\mathcal{L}_u$.

First note that

$$\mathcal{L}_u f = \nabla_v \cdot \left(M_u \nabla_v \left(\frac{f}{M_u} \right) \right), \tag{2.1}$$

with the local Maxwellian M_u defined by (1.27). This implies that $\mathcal{L}_u$ is a self-adjoint operator on $L_u^2(\mathbb{R}^2)$, defined as the weighted L^2 space with weight M_u^{-1}. Denote the inner

product on $L^2_u(\mathbb{R}^2)$ by $\langle f, g\rangle = \int fgM_0^{-1}\,\mathrm{d}v$. Then one has the coercivity estimate

$$\langle f, \mathcal{L}_u f\rangle = -\int M_u \left|\nabla_v\left(\frac{f}{M_u}\right)\right|^2 \mathrm{d}v \le 0, \tag{2.2}$$

which implies that the null space of $\mathcal{L}_u$ is one dimensional, spanned by M_u.

2.3 The gPC Approximation to the Problem with Uncertainty

We consider the system (1.19) with the random variable z as introduced in Section 1.2. We assume that the only random input is the initial data, i.e., f_{in} and u_{in} depend on z

2.3.1 Equations for the gPC coefficients

We adopt the gPC-sG approximation for (1.19) (as introduced in Section 1.2.1) and obtain the following deterministic system for the gPC coefficients:

$$\begin{cases} \partial_t f_j + v\cdot\nabla_x f_j - \nabla_x\Phi\cdot\nabla_v f_j = \dfrac{1}{\epsilon}(\mathcal{L}_u f^K)_j, \\ \partial_t u_j + \nabla_x\cdot(u^K\otimes u^K)_j + \nabla_x p_j - \dfrac{1}{Re}\Delta_x u_j = \dfrac{1}{\epsilon}\kappa\displaystyle\int((v-u^K)f^K)_j\,\mathrm{d}v, \\ \nabla_x\cdot u_j = 0, \end{cases} \tag{2.3}$$

where the sub-index $j = 1,\dots,K$.

With a chosen K, for a function $g = g(z) \approx g^K$, we introduce the notation

$$\begin{aligned} &\vec{g} = (g_1,\dots,g_K)^T, \\ &A(g) \text{ is a size } K+1 \text{ matrix defined by} A(g)_{ij} = \sum_{k=1}^{K} S_{ijk}g_k, \end{aligned} \tag{2.4}$$

where S_{ijk} is defined by (1.50). Then the system of gPC coefficients can be written as

$$\begin{cases} \partial_t\vec{f} + v\cdot\nabla_x\vec{f} - \nabla_x\Phi\cdot\nabla_v\vec{f} = \dfrac{1}{\epsilon}\vec{\mathcal{L}}_{\vec{u}}\vec{f}, \\ \partial_t\vec{u}^{(i)} + \partial_{x_1}(A^{(i)}\vec{u}^{(1)}) + \partial_{x_2}(A^{(i)}\vec{u}^{(2)}) + \partial_{x_i}\vec{p} - \dfrac{1}{Re}\Delta_x\vec{u}^{(i)} = \dfrac{1}{\epsilon}\kappa\displaystyle\int(v_i - A^{(i)})\vec{f}\,\mathrm{d}v, \quad i = 1,2, \\ \nabla_x\cdot\vec{u} = 0, \end{cases} \tag{2.5}$$

where $\vec{\mathcal{L}}$ is the gPC version of the Fokker-Planck operator

$$\vec{\mathcal{L}}_{\vec{u}}(\vec{f}) = \Delta_v \vec{f} + \partial_{v_1}(v_1\vec{f}) + \partial_{v_2}(v_2\vec{f}) - \partial_{v_1}(A^{(1)}\vec{f}) - \partial_{v_2}(A^{(2)}\vec{f}), \tag{2.6}$$

and $A^{(1)}, A^{(2)}$ are given by

$$A^{(1)} = A(u^{(1)}), \quad A^{(2)} = A(u^{(2)}), \tag{2.7}$$

and a term like $A^{(i)}\vec{u}^{(1)}$ or $A^{(i)}\vec{f}$ is interpreted as a matrix-vector multiplication.

2.3.2 The structure and coercivity of $\vec{\mathcal{L}}_{\vec{u}}$

We will analyze the structure of $\vec{\mathcal{L}}_{\vec{u}}$ and give a coercivity result as a generalization of (2.2). We begin with introducing the gPC version of the translation operator $\vec{\mathcal{T}}_u$ on $L^2(\mathbb{R}^2, \mathbb{R}^{K+1})$ by

$$\vec{\mathcal{T}}_{\vec{u}}(\vec{f}) = \exp(-A^{(1)}\partial_{v_1} - A^{(2)}\partial_{v_2})(\vec{f}). \tag{2.8}$$

In other words, $\vec{\mathcal{T}}_{\vec{u}}(\vec{f})$ is the solution at $s = 1$ of the hyperbolic system

$$\partial_s\vec{g}(s,v) + A^{(1)}\partial_{v_1}\vec{g} + A^{(2)}\partial_{v_2}\vec{g} = 0, \quad \vec{g}(0,\cdot) = \vec{f}(\cdot). \tag{2.9}$$

Note that this system is hyperbolic because $A^{(1)}, A^{(2)}$ are symmetric. It is easy to see that $\vec{\mathcal{T}}_{\vec{u}}$ is a unitary operator, i.e., $\int \vec{\mathcal{T}}_{\vec{u}}\vec{f}\cdot\vec{\mathcal{T}}_{\vec{u}}\vec{g}\,\mathrm{d}v = \int \vec{f}\cdot\vec{g}\,\mathrm{d}v$. We first prove the following lemma:

Lemma 2.1.

$$\vec{\mathcal{T}}_{\vec{u}}[\partial_{v_1}(v_1\vec{f}) + \partial_{v_2}(v_2\vec{f})] = \partial_{v_1}(v_1\vec{\mathcal{T}}_{\vec{u}}\vec{f}) + \partial_{v_2}(v_2\vec{\mathcal{T}}_{\vec{u}}\vec{f}) - (A^{(1)}\partial_{v_1} + A^{(2)}\partial_{v_2})\vec{\mathcal{T}}_{\vec{u}}\vec{f}. \tag{2.10}$$

Proof. Let $\vec{g}(s,v)$ be the solution of the hyperbolic system (2.9). Then

$$\begin{aligned}
&\partial_s[\partial_{v_1}(v_1\vec{g})] \\
=&\partial_{v_1}[v_1(-A^{(1)}\partial_{v_1} - A^{(2)}\partial_{v_2})\vec{g}] \\
=&\partial_{v_1}[(-A^{(1)}\partial_{v_1} - A^{(2)}\partial_{v_2})(v_1\vec{g}) + A^{(1)}\vec{g}] \\
=&(-A^{(1)}\partial_{v_1} - A^{(2)}\partial_{v_2})\partial_{v_1}(v_1\vec{g}) + A^{(1)}\partial_{v_1}\vec{g}
\end{aligned} \tag{2.11}$$

Writing a similar expression for the second component of v and adding together, one gets

$$\begin{aligned}&\partial_s[\partial_{v_1}(v_1\vec{g})+\partial_{v_2}(v_2\vec{g})]\\=&(-A^{(1)}\partial_{v_1}-A^{(2)}\partial_{v_2})[\partial_{v_1}(v_1\vec{g})+\partial_{v_2}(v_2\vec{g})]+(A^{(1)}\partial_{v_1}+A^{(2)}\partial_{v_2})\vec{g}\end{aligned}\tag{2.12}$$

Then

$$\begin{aligned}&\partial_s[\partial_{v_1}(v_1\vec{g})+\partial_{v_2}(v_2\vec{g})-s(A^{(1)}\partial_{v_1}+A^{(2)}\partial_{v_2})\vec{g}]\\=&(-A^{(1)}\partial_{v_1}-A^{(2)}\partial_{v_2})[\partial_{v_1}(v_1\vec{g})+\partial_{v_2}(v_2\vec{g})]-s(A^{(1)}\partial_{v_1}+A^{(2)}\partial_{v_2})\partial_s\vec{g}\\=&(-A^{(1)}\partial_{v_1}-A^{(2)}\partial_{v_2})[\partial_{v_1}(v_1\vec{g})+\partial_{v_2}(v_2\vec{g})-s(A^{(1)}\partial_{v_1}+A^{(2)}\partial_{v_2})\vec{g}]\end{aligned}\tag{2.13}$$

Since the function $\partial_{v_1}(v_1\vec{g})+\partial_{v_2}(v_2\vec{g})-s(A^{(1)}\partial_{v_1}+A^{(2)}\partial_{v_2})\vec{g}$ evaluated at $s=0$ is $\partial_{v_1}(v_1\vec{f})+\partial_{v_2}(v_2\vec{f})$, and it satisfies the hyperbolic system (2.9), we get the conclusion by the definition of $\vec{\mathcal{T}}_{\vec{u}}$ (RHS of (2.10) is just this function evaluated at $s=1$). □

Theorem 2.3.1.

$$\vec{\mathcal{L}}_{\vec{u}}=\vec{\mathcal{T}}_{\vec{u}}\vec{\mathcal{L}}_{\vec{0}}\vec{\mathcal{T}}_{\vec{u}}^{-1}.\tag{2.14}$$

Proof. By Lemma 3.1 and the fact that the Laplacian Δ_v commutes with $\vec{\mathcal{T}}_{\vec{u}}$, one gets

$$\vec{\mathcal{T}}_{\vec{u}}\vec{\mathcal{L}}_{\vec{0}}=\vec{\mathcal{L}}_{\vec{u}}\vec{\mathcal{T}}_{\vec{u}},\tag{2.15}$$

which is the desired conclusion.

□

Define

$$\langle\vec{f},\vec{g}\rangle_{\vec{u}}=\int(\vec{\mathcal{T}}_{\vec{u}}^{-1}\vec{f})\cdot(\vec{\mathcal{T}}_{\vec{u}}^{-1}\vec{g})M_0^{-1}\,\mathrm{d}v,\tag{2.16}$$

where $M_0=\frac{1}{2\pi}\exp(-\frac{|v|^2}{2})$. It is clear that $\langle\vec{f},\vec{f}\rangle_{\vec{u}}$ is positive if $\vec{f}\neq\vec{0}$, and thus $\langle\cdot,\cdot\rangle_{\vec{u}}$ is an inner product on $L^2_{\vec{u}}(\mathbb{R}^2,\mathbb{R}^{K+1})$, defined as the subspace of $L^2(\mathbb{R}^2,\mathbb{R}^{K+1})$ consisting of functions $\vec{f}$ with $\langle\vec{f},\vec{f}\rangle_{\vec{u}}<\infty$. This space is the vector analog of $L^2_u(\mathbb{R}^2)$ defined in Section 2.2.

From

$$\langle\vec{f},\vec{\mathcal{L}}_{\vec{u}}\vec{g}\rangle_{\vec{u}}=\int\vec{\mathcal{T}}_{\vec{u}}^{-1}\vec{f}\cdot(\vec{\mathcal{T}}_{\vec{u}}^{-1}\vec{\mathcal{L}}_{\vec{u}}\vec{g})M_0^{-1}\,\mathrm{d}v=\int\vec{\mathcal{T}}_{\vec{u}}^{-1}\vec{f}\cdot(\vec{\mathcal{L}}_{\vec{0}}\vec{\mathcal{T}}_{\vec{u}}^{-1}\vec{g})M_0^{-1}\,\mathrm{d}v,\tag{2.17}$$

and the self-adjointness of $\mathcal{L}_0$, it is clear that $\vec{\mathcal{L}}_{\vec{u}}$ is self-adjoint on $L^2_{\vec{u}}(\mathbb{R}^2, \mathbb{R}^{K+1})$. Also, it follows from (2.2) that

$$\langle \vec{f}, \vec{\mathcal{L}}_{\vec{u}} \vec{f} \rangle_{\vec{u}} = \int \vec{\mathcal{T}}_{\vec{u}}^{-1} \vec{f} \cdot (\vec{\mathcal{L}}_{\vec{0}} \vec{\mathcal{T}}_{\vec{u}}^{-1} \vec{f}) M_0^{-1} \, \mathrm{d}v = - \int M_0 \left| \nabla_v \left(\frac{\vec{\mathcal{T}}_{\vec{u}}^{-1} \vec{f}}{M_0} \right) \right|^2 \mathrm{d}v \leq 0, \tag{2.18}$$

which is a vector analog of the coercivity estimate (2.2). From (2.18) it is easy to prove the following theorem:

Theorem 2.3.2. *The null space of $\vec{\mathcal{L}}_{\vec{u}}$ is given by*

$$\vec{M}(v) = \vec{\mathcal{T}}_{\vec{u}}(M_0 \vec{C}), \tag{2.19}$$

where $M_0 = \frac{1}{2\pi} \exp(-\frac{|v|^2}{2})$, and $\vec{C}$ is any constant vector.

2.3.3 The hydrodynamic limit of the gPC system

Based on Theorem 2.3.2, we formally derive the hydrodynamic limit of the gPC system (2.5), and show that the limit is the gPC approximation to the limiting Navier-Stokes system (1.31), which means that the gPC approximation (2.5) is s-AP.

Define the moments of $\vec{f}$ by

$$\vec{n} = \int \vec{f} \, \mathrm{d}v, \quad \vec{J} = \int v \vec{f} \, \mathrm{d}v, \quad \vec{P} = \int v \otimes v \vec{f} \, \mathrm{d}v. \tag{2.20}$$

As $\epsilon \to 0$, from (2.5) one formally has $\vec{\mathcal{L}}_{\vec{u}} \vec{f} = 0$, which implies

$$\vec{f}(v) = \vec{M}(v) = \vec{\mathcal{T}}_{\vec{u}}(M_0 \vec{C}), \tag{2.21}$$

by Theorem 2.3.2. Substituting into (2.20), one gets

$$\begin{aligned} \vec{n} &= \int \vec{M}(v) \, \mathrm{d}v = \vec{C}, \\ \vec{J}^{(i)} &= \int v_i \vec{M}(v) \, \mathrm{d}v = A^{(i)} \vec{n}, \quad i = 1, 2, \\ \vec{P}^{(ij)} &= \int v_i v_j \vec{M}(v) \, \mathrm{d}v = \delta_{ij} \vec{n} + \frac{1}{2} (A^{(i)} A^{(j)} + A^{(j)} A^{(i)}) \vec{n}, \quad i, j = 1, 2, \end{aligned} \tag{2.22}$$

which are consistent with the expression of J, P in the deterministic case (see (1.28)). Integrating the first equation of (2.5) against $1, v$ one gets

$$\begin{cases} \partial_t \vec{n} + \nabla_x \cdot \vec{J} = 0, \\ \partial_t \vec{J}^{(i)} + \nabla_x \cdot \vec{P}^{(i)} - \vec{n} \nabla_{x_i} \Phi = -\dfrac{1}{\epsilon} \displaystyle\int (v_i - A^{(i)}) \vec{f} \, \mathrm{d}v, \quad i = 1, 2. \end{cases} \tag{2.23}$$

Substituting (2.22) into the above equations and doing proper linear combinations, one has

$$\begin{cases} \partial_t \vec{n} + \nabla_x \cdot (A\vec{n}) = 0, \\ \partial_t (\vec{u}^{(i)} + \kappa A^{(i)} \vec{n}) + \nabla_x \cdot [A\vec{u}^{(i)} + \dfrac{\kappa}{2} (A^{(i)} A + A A^{(i)}) \vec{n}] \\ \qquad + \partial_{x_i} (\vec{p} + \kappa \vec{n}) + \kappa \vec{n} \partial_{x_i} \Phi = \dfrac{1}{Re} \Delta_x \vec{u}^{(i)}, \quad i = 1, 2. \end{cases} \tag{2.24}$$

Note if one inserts the gPC ansatz $f = f^K = \sum_{k=1}^K f_k \phi_k$, $u = u^K = \sum_{k=1}^K u_k \phi_k$ with definitions in (2.20) into the Navier-Stokes system (1.31) and conducts the Galerkin projection, one gets exactly (2.24). Thus the s-AP property of (2.5) is justified.

2.4 The Fully Discrete s-AP Scheme for the System with Uncertainty

For simplicity of notation, in this section, an expression of the form $c + B$ where c is a scalar and B is a matrix is interpreted as $cI + B$, where I is the identity matrix.

In order to get a first order AP scheme for the system with uncertainty, we follow the steps of the deterministic AP scheme [42]. Starting from (2.5), we first give an outline of the scheme, and then provide justifications in the subsequent sections.

STEP 1: Integrating the first equation of (2.5) over v,

$$\frac{1}{\Delta t} (\vec{n}^{k+1} - \vec{n}^k) = - \int v \cdot \nabla_x \vec{f}^k \, \mathrm{d}v. \tag{2.25}$$

Then one gets $\vec{n}^{k+1}$.

STEP 2: We adopt the idea of the projection method for the Navier-Stokes equations. We first solve the second equation of (2.5) with the pressure term being $\nabla_x \vec{p}^k$, and then solve for the increment of $\vec{p}$. Multiplying the first equation of (2.5) and integrating over v,

combining with the second equation of (2.5), one gets

$$\frac{1}{\Delta t}(\vec{J}^* - \vec{J}^k) = -\int v \otimes v \cdot \nabla_x \vec{f}^k \, \mathrm{d}v - \vec{n}^k \nabla_x \Phi - \frac{1-\alpha}{\epsilon}[\vec{J}^* - A(\vec{n}^{k+1})\vec{u}^*], \tag{2.26}$$

$$\frac{1}{\Delta t}(\vec{u}^* - \vec{u}^k) + \nabla_x \vec{p}^k - \Delta_x \vec{u}^* = -\nabla_x \cdot [A(\vec{u}^k) \otimes \vec{u}^k] + \frac{1-\alpha}{\epsilon}\kappa[\vec{J}^* - A(\vec{n}^{k+1})\vec{u}^*], \tag{2.27}$$

where $\alpha \in (0,1)$ is a fixed parameter to be chosen. Only a part of the stiff term $\frac{1-\alpha}{\epsilon}(\vec{J}^* - A(\vec{n}^{k+1})\vec{u}^*)$ is contained in this step, and another part is contained in STEP 3. The purpose of this is to make sure that as $\epsilon \to 0$, the equation $\vec{J} = A(\vec{n})\vec{u}$ holds at each step. Eliminating $\vec{J}^*$ one gets,

$$\begin{aligned}(B_1 - \frac{1}{Re}\Delta_x)\vec{u}^* = &\frac{\vec{u}^k}{\Delta t} - \nabla_x \cdot [A(\vec{u}^k) \otimes \vec{u}^k] \\ &- \nabla_x \vec{p}^k + \frac{(1-\alpha)\kappa}{\epsilon + (1-\alpha)\Delta t}(\vec{J}^k - \Delta t \int v \otimes v \nabla_x \vec{f}^k \, \mathrm{d}v - \Delta t \vec{n}^k \nabla_x \Phi),\end{aligned} \tag{2.28}$$

where

$$B_1 = \frac{1}{\Delta t} + \frac{1-\alpha}{\epsilon + (1-\alpha)\Delta t}\kappa A(\vec{n}^{k+1}). \tag{2.29}$$

This is a system of the Helmholtz equation, whose coefficient matrix B_1 is symmetric positive definite under reasonable and mild assumptions on $\vec{n}^{k+1}$ (given in Section 2.4.1). Solving it by the conjugate gradient method to get $\vec{u}^*$, and then substituting the result of $\vec{u}^*$ into (2.26), one gets $\vec{J}^*$.

STEP 3: To solve for the increment of $\vec{p}$, make $\vec{u}^{k+1}$ divergence-free, and include the other part of the stiff term $\frac{\alpha}{\epsilon}(\vec{J}^{**} - A(\vec{n}^{k+1})\vec{u}^{k+1})$, one has

$$\begin{aligned}&\frac{1}{\Delta t}(\vec{J}^{**} - \vec{J}^*) = -\frac{\alpha}{\epsilon}[\vec{J}^{**} - A(\vec{n}^{k+1})\vec{u}^{k+1}], \\ &\frac{1}{\Delta t}(\vec{u}^{k+1} - \vec{u}^*) + \nabla_x(\vec{p}^{k+1} - \vec{p}^k) = \frac{\alpha}{\epsilon}\kappa[\vec{J}^{**} - A(\vec{n}^{k+1})\vec{u}^{k+1}].\end{aligned} \tag{2.30}$$

Eliminate J^{**},

$$\begin{aligned}&\vec{u}^{k+1} + \left[\frac{1}{\Delta t} + \frac{\alpha}{\epsilon}(1 + \kappa A(\vec{n}^{k+1}))\right]^{-1} (\frac{1}{\Delta t} + \frac{\alpha}{\epsilon})\Delta t \nabla_x(\vec{p}^{k+1} - \vec{p}^k) \\ &= \left[\frac{1}{\Delta t} + \frac{\alpha}{\epsilon}(1 + \kappa A(\vec{n}^{k+1}))\right]^{-1} \left((\frac{1}{\Delta t} + \frac{\alpha}{\epsilon})\vec{u}^* + \frac{\alpha}{\epsilon}\kappa\vec{J}^*\right).\end{aligned} \tag{2.31}$$

To solve for $\vec{u}^{k+1}$, first take divergence:

$$\nabla_x \cdot (B_2 \nabla_x (p^{k+1} - p^k)) = \frac{1}{\Delta t} \nabla_x \cdot \left[\left[\frac{1}{\Delta t} + \frac{\alpha}{\epsilon}(1 + \kappa A(\vec{n}^{k+1})) \right]^{-1} \left((\frac{1}{\Delta t} + \frac{\alpha}{\epsilon})\vec{u}^* + \frac{\alpha}{\epsilon}\kappa \vec{J}^* \right) \right], \tag{2.32}$$

with

$$B_2 = \left[\frac{1}{\Delta t} + \frac{\alpha}{\epsilon}(1 + \kappa A(\vec{n}^{k+1})) \right]^{-1} (\frac{1}{\Delta t} + \frac{\alpha}{\epsilon}). \tag{2.33}$$

This is a system of the variable coefficient Poisson equation for $(\vec{p}^{k+1} - \vec{p}^{k})$, whose coefficient matrix is symmetric positive definite under reasonable assumptions on $\vec{n}^{k+1}$ (given in Section 2.4.1). Solving it with the Neumann boundary condition

$$\left. \frac{\partial (\vec{p}^{k+1} - \vec{p}^{k})}{\partial \nu} \right|_{\partial \Omega} = 0, \tag{2.34}$$

by the conjugate gradient method to get $\vec{p}^{k+1}$, and then substituting back to (2.31), one gets $\vec{u}^{k+1}$.

STEP 4: To get $\vec{f}^{k+1}$ we solve the first equation of (2.5) with the stiff term $\vec{\mathcal{L}}_{\vec{u}} \vec{f}$ treated implicitly, using the $\vec{u}^{k+1}$ obtained above:

$$\frac{\vec{f}^{k+1} - \vec{f}^{k}}{\Delta t} + v \cdot \nabla_x \vec{f}^{k} - \nabla_x \Phi \cdot \nabla_v \vec{f}^{k} = \frac{1}{\epsilon} \vec{\mathcal{L}}_{\vec{u}^{k+1}} \vec{f}^{k+1}. \tag{2.35}$$

Using the relation $\vec{\mathcal{L}}_{\vec{u}} = \vec{\mathcal{T}}_{\vec{u}} \vec{\mathcal{L}}_{\vec{0}} \vec{\mathcal{T}}_{\vec{u}}^{-1}$ (where we omit the index $k+1$ on $\vec{u}$), one gets

$$\vec{f}^{k+1} = \epsilon \vec{\mathcal{T}}_{\vec{u}} (\frac{\epsilon}{\Delta t} - \vec{\mathcal{L}}_{\vec{0}})^{-1} \vec{\mathcal{T}}_{\vec{u}}^{-1} \left(\frac{\vec{f}^{k}}{\Delta t} - v \cdot \nabla_x \vec{f}^{k} + \nabla_x \Phi \cdot \nabla_v \vec{f}^{k} \right). \tag{2.36}$$

The computation of the operator $\vec{\mathcal{T}}_{\vec{u}}$ and $\vec{\mathcal{T}}_{\vec{u}}^{-1}$ is given in Section 2.4.2, and the inversion of $(\frac{\epsilon}{\Delta t} - \vec{\mathcal{L}}_{\vec{0}})$ can be done by the same method as the deterministic case [59] (see the Appendix A for details), since $\vec{\mathcal{L}}_{\vec{0}}$ is the operator $\mathcal{L}_0$ acting on each gPC mode. Thus one can get $\vec{f}^{k+1}$, and get $\vec{J}^{k+1}$ by taking moments of $\vec{f}^{k+1}$.

The second order AP scheme can be derived in the same way as the deterministic case, see [42] for details.

2.4.1 On the Helmholtz equation and the Poisson equation

In STEP 2 and STEP 3 one needs to solve a system of Helmholtz equations and a system of Poisson equations respectively. We give the proof of positive definiteness of the coefficient matrices of these equations under reasonable assumptions.

In STEP 2 the system of Helmholtz equations for $\vec{u}^*$ has the form (we omit the time step indexes)

$$(B_1 - \frac{1}{Re}\Delta_x)\vec{u}(x) = \vec{g}(x), \tag{2.37}$$

where B_1 is defined in (2.29). Define the LHS of (2.37) as $\mathcal{B}_1(\vec{u})$, and then $\mathcal{B}_1$ is an (unbounded) operator on the Hilbert space $L^2(\Omega, \mathbb{R}^{K+1})$. $\mathcal{B}_1$ is symmetric since

$$\begin{aligned}
&< \mathcal{B}_1(\vec{u}), \vec{w} > \\
&= \int \left[\sum_{i,j,k=0}^{K} \frac{1-\alpha}{\epsilon + (1-\alpha)\Delta t} S_{ijk} n_i(x) u_j(x) + \sum_{k=0}^{K} (\frac{1}{\Delta t} - \frac{1}{Re}\Delta_x u_k(x)) \right] w_k(x)\,\mathrm{d}x \\
&= \int \sum_{k=0}^{K} \left(\frac{1}{\Delta t} u_k(x) w_k(x) + \frac{1}{Re} \nabla_x u_k(x) \cdot \nabla_x w_k(x) \right) \mathrm{d}x \\
&\qquad + \frac{1-\alpha}{\epsilon + (1-\alpha)\Delta t} \int \int n^K(x,z) u^K(x,z) w^K(x,z) \pi(z)\,\mathrm{d}z\,\mathrm{d}x
\end{aligned} \tag{2.38}$$

for $\vec{u}, \vec{w} \in H_0^1(\Omega, \mathbb{R}^{K+1})$.

By Poincaré's Inequality,

$$\int |\nabla u_k(x)|^2\,\mathrm{d}x \geq C_1 \int |u_k(x)|^2\,\mathrm{d}x, \tag{2.39}$$

for some positive constant C_1 depending on Ω, the x-domain. Then, under the assumption that

$$n^K(x,z) > -\delta_1, \tag{2.40}$$

where

$$\delta_1 = \frac{(1 + C_1 \Delta t/Re)(\epsilon + (1-\alpha)\Delta t)}{(1-\alpha)\Delta t} > 1, \tag{2.41}$$

one can see that $\mathcal{B}_1$ is positive definite. The assumption (2.40) is reasonable since $n^K(x,z)$ is an approximation of $n(x,z) \geq 0$ with a spectral accuracy (see Remark 2.2 for more details).

In STEP 3 the system of Poisson equations for $\vec{u}^{**}$ has the form

$$\nabla \cdot (B_2 \nabla \vec{u}(x)) = \vec{g}(x), \tag{2.42}$$

where B_2 is defined in (2.33). Define the LHS as $\mathcal{B}_2(\vec{u})$, and then $\mathcal{B}_2$ is an (unbounded) operator on the Hilbert space $L^2(\Omega, \mathbb{R}^{K+1})$. $\mathcal{B}$ is symmetric since

$$\begin{aligned} & < \mathcal{B}_2(\vec{u}), \vec{w} > \\ = & \int \sum_{j,k=0}^{K} (\nabla_x \cdot (B_{2,jk}(x) \nabla u_j(x)) w_k(x) \, \mathrm{d}x \\ = & -\int \sum_{j,k=0}^{K} B_{2,jk}(x) \nabla_x u_j(x) \cdot \nabla_x w_k(x) \, \mathrm{d}x \end{aligned}$$

for $\vec{u}, \vec{w} \in H_0^1(\Omega, \mathbb{R}^{K+1})$. Similar to the case of $\mathcal{B}_1$, one can show that the matrix $\left[\frac{1}{\Delta t} + \frac{\alpha}{\epsilon}(1 + \kappa A(\vec{n}))\right]$ is positive definite if

$$n^K(x, z) > -\delta_2, \tag{2.43}$$

where

$$\delta_2 = \frac{\epsilon + \alpha \Delta t}{\alpha \kappa \Delta t} > \frac{1}{\kappa}. \tag{2.44}$$

This assumption is also reasonable due to the positivity of the exact solution $n(x, z)$ (see Remark 2.2 for more details). Then one can deduce that the inverse matrix in the definition of B_2 exists, and B_2 is also positive definite. Therefore the operator $\mathcal{B}_2$ is negative definite.

Remark 2.2. *In general the numerical solution $n^K(x, z)$ is not necessarily positive everywhere. However, since the exact solution $n(x, z)$ is positive everywhere, and the gPC approximation is spectrally accurate (based on the smoothness of $n(x, z)$), it is expected that (2.40) and (2.43) are easily satisfied if K is reasonably large.*

The loss of positivity is a pitfall of any spectral method, with no exception to the gPC method. We did not encounter any problem in this paper. If positivity is necessary (for example, if one considers a more complicated model where n enters into the definition of drag coefficients), one might need to use some positivity-preserving techniques, such as the one developed in [109], to fix the problem. This will be studied in a subsequent work.

2.4.2 Computing $\vec{\mathcal{T}}$ numerically

In STEP 4 it is required to compute the operator $\vec{\mathcal{T}}_{\vec{u}} = \exp(-A^{(1)}\partial_{v_1} - A^{(2)}\partial_{v_2})$. If one computes it directly by the discrete Fourier transform, then one needs to multiply $\exp(-iA^{(1)}\xi_1 - iA^{(2)}\xi_2)$ on each Fourier mode (ξ_1, ξ_2). Since the matrices $A^{(1)}$ and $A^{(2)}$ do not commute in general, in order to compute these matrix exponentials, one has to diagonalize the matrix $(A^{(1)}\xi_1 + A^{(2)}\xi_2)$ for each (ξ_1, ξ_2). Thus one needs N_v^2 times of diagonalization of matrices of size $K+1$, where N_v is the number of mesh points in one dimension of the velocity space. This can be extremely expensive if N_v and K are large. Here we introduce an operator splitting

$$\exp(-A^{(1)}\partial_{v_1} - A^{(2)}\partial_{v_2})\vec{f} \approx \exp(-A^{(1)}\partial_{v_1})\exp(-A^{(2)}\partial_{v_2})\vec{f}. \tag{2.45}$$

In general, on the Fourier mode (ξ_1, ξ_2), the difference between the matrices $\exp(-iA^{(1)}\xi_1 - iA^{(2)}\xi_2)$ and $\exp(-iA^{(1)}\xi_1)\exp(-iA^{(2)}\xi_2)$ can be of order $O(1)$ since the commutator of $A^{(1)}$ and $A^{(2)}$ can be of order $O(1)$. However, we will prove that under smoothness assumptions, the splitting (2.45) has spectral accuracy and show that it saves the computational cost of $\vec{\mathcal{T}}_{\vec{u}}$ dramatically. By saying that an approximation has *spectral accuracy*, or an error is *spectrally small*, we mean that the error of the approximation in L^2 norm is less than $\frac{C_m}{K^m}$, for any $m \geq 1$. Our accuracy result is

Theorem 2.4.1. *Assume that $u(z) = u^K$ is sufficiently smooth in z, $f(v,z) = f^K$ is sufficiently smooth in v, z. Then the splitting (2.45) has spectral accuracy, namely,*

$$\exp(-A^{(1)}\partial_{v_1} - A^{(2)}\partial_{v_2})\vec{f} - \exp(-A^{(1)}\partial_{v_1})\exp(-A^{(2)}\partial_{v_2})\vec{f} = O(\frac{1}{K^m}), \quad \forall m \geq 1, \tag{2.46}$$

where m depends on the regularity of u and f.

To prove the theorem, we need the following lemma concerning the spectral accuracy of the gPC approximation of the linear transport equation:

Lemma 2.3. *Consider the equation*

$$\partial_s g + u \cdot \nabla_v g = 0, \quad g|_{s=0} = g^0, \tag{2.47}$$

where $g = g(s, v, z)$, $u(z) = u^K(z)$, and its gPC approximation

$$\partial_s \tilde{g} + A(u) \cdot \nabla_v \tilde{g} = 0, \quad \tilde{g}|_{s=0} = (g^0)^K, \tag{2.48}$$

where the matrix $A(u)$ acts on $\tilde{g}$ by matrix multiplication as if $\tilde{g}$ is written in the vector form $\vec{g}$ (see (2.4)). Suppose u is sufficiently smooth in z and g^0 is sufficiently smooth in v, z. Then $\tilde{g}|_{s=1}$ as an approximation of $g|_{s=1}$ has spectral accuracy.

Proof. Denote P_K as the projection onto the gPC approximation space. Then $P_K g$ satisfies

$$\partial_s P_K g + P_K(u \cdot \nabla_v P_K g) + P_K(u \cdot \nabla_v (I - P_K)g) = 0. \tag{2.49}$$

By the definition of $A(u)$ in (2.4), one can see that

$$A(u) \cdot \nabla_v \tilde{g} = P_K(u \cdot \nabla_v \tilde{g}). \tag{2.50}$$

Thus

$$\partial_s \tilde{g} + P_K(u \cdot \nabla_v \tilde{g}) = 0. \tag{2.51}$$

Take the difference between (2.49) and (2.51) one gets

$$\partial_s (P_K g - \tilde{g}) + A(u) \cdot \nabla_v (P_K g - \tilde{g}) + P_K(u \cdot \nabla_v (I - P_K)g) = 0. \tag{2.52}$$

Since $P_K g = \tilde{g} = (g^0)^K$ at $s = 0$, one solves the above equation as

$$(P_K g - \tilde{g})|_{s=1} = -\int_0^1 \exp(-(1-\tau)A(u) \cdot \nabla_v) P_K(u \cdot \nabla_v (I - P_K)g(\tau, v, z)) \, \mathrm{d}\tau. \tag{2.53}$$

Since $g(s, v, z) = g^0(v - su(z), z)$ and g^0, u are sufficiently smooth, the higher order v, z-derivatives of g are bounded in $L^\infty_s([0,1], L^2_{v,z})$. Thus $\nabla_v (I - P_K)g$ is spectrally small in the same norm, and so is $P_K(u \cdot \nabla_v (I - P_K)g)$. Since $\exp(-(1-\tau)A(u) \cdot \nabla_v)$ is unitary, one concludes that $(P_K g - \tilde{g})|_{s=1}$ is spectrally small. Then one can get the conclusion since $(P_K g - g)|_{s=1}$ is spectrally small due to the smoothness of $g|_{s=1}$. □

Proof of the Theorem. One notes that the previous lemma implies that

$$\exp(-A^{(1)}\partial_{v_1} - A^{(2)}\partial_{v_2})\vec{f} - f^K(v - u(z), z) \tag{2.54}$$

is spectrally small, where the first term is interpreted as a function of v, z. Also, by taking $u = (0, u^{(2)})$, the lemma implies that

$$\exp(-A^{(2)}\partial_{v_2})\vec{f} - f^K(v_1, v_2 - u^{(2)}(z), z) \tag{2.55}$$

is spectrally small. Then, by taking $u = (u^{(1)}, 0)$, the lemma implies that

$$\exp(-A^{(1)}\partial_{v_1})\exp(-A^{(2)}\partial_{v_2})\vec{f} - f^K(v_1 - u^{(1)}(z), v_2 - u^{(2)}(z), z) \tag{2.56}$$

is spectrally small. Then combining (2.54) and (2.56) the theorem is proved. □

Thus if one computes $\vec{\mathcal{T}}_{\vec{u}}$ by the RHS of (2.45), then at most a spectral error is introduced. The RHS of (2.45) can be computed by using the discrete Fourier transform. For example, the operator $\exp(-A^{(1)}\partial_{v_1})$ is the multiplication of $\exp(-iA^{(1)}\xi_1)$ on the Fourier mode ξ_1. By diagonalizing the matrix $A^{(1)}$, all such matrix exponentials are easily computed. Thus, with this splitting, *only two* (instead of N_v^2 as in the direct method) matrix diagonalizations are required.

2.5 Numerical Results

In the numerical tests in this section, the x-domain is $[0,1] \times [0,1]$, and the no-slip boundary condition for u is taken. We take the parameters $\kappa = 2$, $Re = 1000$, and $\Phi(x) = x_2$ the gravity field. The random domain is taken as $I_z = [-1, 1]$ with the uniform distribution, except for the last example, where $I_z = (-\infty, \infty)$ with the normal distribution $\mathcal{N}(0, 1)$. For the stochastic Galerkin method, we take

$$K = 7, \quad N_x = 128, \quad N_v = 32, \quad R_v = 7, \quad \Delta t = 1/2560. \tag{2.57}$$

(Notice that the notation is a little different from []: here K is the number of basis functions, while in [] $(K + 1)$ is the number of basis functions.) The mesh sizes are given by

$$\Delta x = \frac{1}{N_x}, \quad \Delta v = \frac{2R_v}{N_v}. \tag{2.58}$$

The Δt is about $0.18\frac{\Delta x}{R_v}$, which satisfies the CFL condition for the kinetic flux term $v \cdot \nabla_x f$. For the problems in Section 2.5.2, the total time t is taken as 2000 time steps, which is about $t = 0.39$.

The mesh points of the x-domain and the v-domain are defined by

$$\begin{aligned} x_{i,j} &= \left((i+\frac{1}{2})\Delta x, (j+\frac{1}{2})\Delta x \right), \quad i,j = 0,\dots,N_x - 1 \\ v_{i,j} &= \left(-R_v + (i+\frac{1}{2})\Delta v, -R_v + (j+\frac{1}{2})\Delta v \right), \quad i,j = 0,\dots,N_v - 1 \end{aligned} \tag{2.59}$$

The kinetic flux term $v \cdot \nabla_x f$ and the forcing term $\nabla_x \Phi \cdot \nabla_v f$ are numerically approximated by the second order upwind scheme with the minmod slope limiter. Other flux terms are approximated by the centered difference scheme.

Given a function $g(z) = \sum_{k=1}^{K} g_k \phi_k(z)$, the expectation value is given by g_0 and the standard deviation is given by $\sqrt{\sum_{k=2}^{K} g_k^2}$.

Given the bulk density $n(z) = \sum_{k=1}^{K} n_k \Phi_k(z)$ and the momentum $J(z) = \sum_{k=1}^{K} J_k \Phi_k(z)$ of the particles, the bulk velocity of the particles $u^p(z) \approx \sum_{k=1}^{K} u_k^p \Phi_k(z)$ is computed by solving u_k^p from the linear equations

$$A(n)\vec{u^p} = \vec{J}, \tag{2.60}$$

where $A(n)$ is defined in (2.4). It is easy to check that the matrix $A(n)$ is symmetric positive-definite if $n(z)$ is everywhere positive. Thus $\vec{u^p}$ can be solved from this set of linear equations. In all the numerical experiments we conducted in this paper, $A(n)$ is always invertible. However, the gPC method does not guarantee positivity. To deal with the positivity issue, see discussions in Remark 2.2.

2.5.1 The s-AP property

To verify the s-AP property of the stochastic Galerkin method, we take initial fluid velocity

$$\begin{aligned} u = (x_2 - 0.5, -(x_1 - 0.5)) \cdot 1000(1 + 0.2z) \cdot [(x_1 - 0.5)^2 + (x_2 - 0.5)^2] \\ \cdot \exp[-100((x_1 - 0.5)^2 + (x_2 - 0.5)^2)], \end{aligned} \tag{2.61}$$

and particle distribution as the Maxwellian with density

$$n = 0.5 - 0.4 \arctan[10(x_1 - 0.5)]/(\pi/2), \tag{2.62}$$

and bulk velocity $\frac{J}{n} = -u$. Notice that this is different from the equilibrium, which is given by a Maxwellian with bulk velocity u. We observe the time evolution of the difference between $\vec{f}$ and the corresponding equilibrium $\vec{M} = \vec{\mathcal{T}}_{\vec{u}}(M_0\vec{n})$ with $\vec{n} = \int \vec{f}\,\mathrm{d}v$ for different values of ϵ (see Section 2.3.3 for related definitions). The difference is measured in $L^\infty_x(L^2_{v,z})$. The result is shown in Figure 1. One can clearly see that $\vec{f}$ is driven to $\vec{M}$ as t increases, until $\|\vec{f} - \vec{M}\| = O(\epsilon)$.

2.5.2 Problems with random initial data

Example 1: random initial fluid velocity.

We take the initial fluid velocity the same as (2.61) and the particle distribution as the equilibrium with density as in (2.62) and bulk velocity the same as the fluid velocity. With this initial data, the fluid near the center of the domain is rotating clockwisely, and the particle density has the tendency of falling to the left bottom due to gravity, as well as the tendency of diffusion due to the gradient of the density.

The expectation and standard deviation of the fluid velocity u and the macroscopic quantities of the particles are shown in Figures 2, 3, 4 (for $\epsilon = 1, 0.01, 10^{-8}$ respectively).

One can clearly see that for $\epsilon = 1$, the fluid velocity and the particle bulk velocity differ much. From the expectations of n and u^p one can clearly see that the particles are falling down (second component of u^p) and there is a strong diffusion of the particles due to the gradient of the initial density (first component of u^p). The randomness on the initial bulk velocity and fluid velocity have little effect on the particle density and bulk velocity afterwards.

For $\epsilon = 0.01$ the fluid velocity and the particle bulk velocity are very close but not the same; for $\epsilon = 10^{-8}$ they look the same. Also, in these two cases one can clearly see that the fluid together with the particles are rotating clockwisely, while for $\epsilon = 0.01$ there is still some visible kinetic effect which is smoothing, compared to the $\epsilon = 10^{-8}$ case. The s-AP property

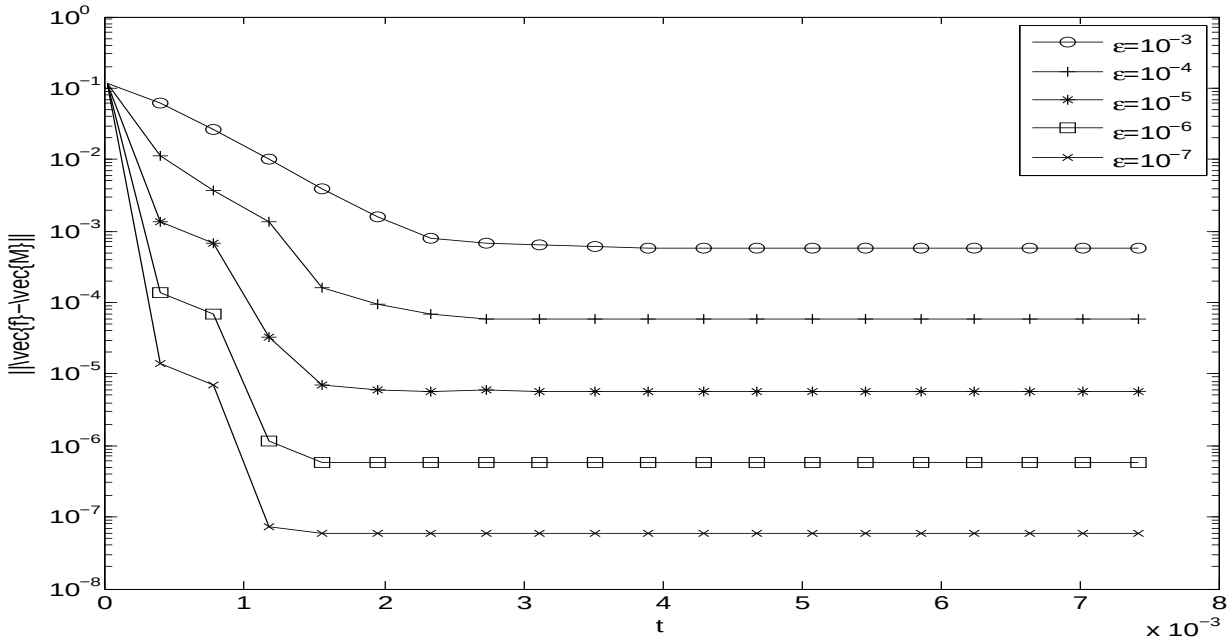

Figure 2.1 The s-AP property: time evolution of $\|\vec{f} - \vec{M}\|$ measured in $L_x^\infty(L_{v,z}^2)$, $\epsilon = 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}$.

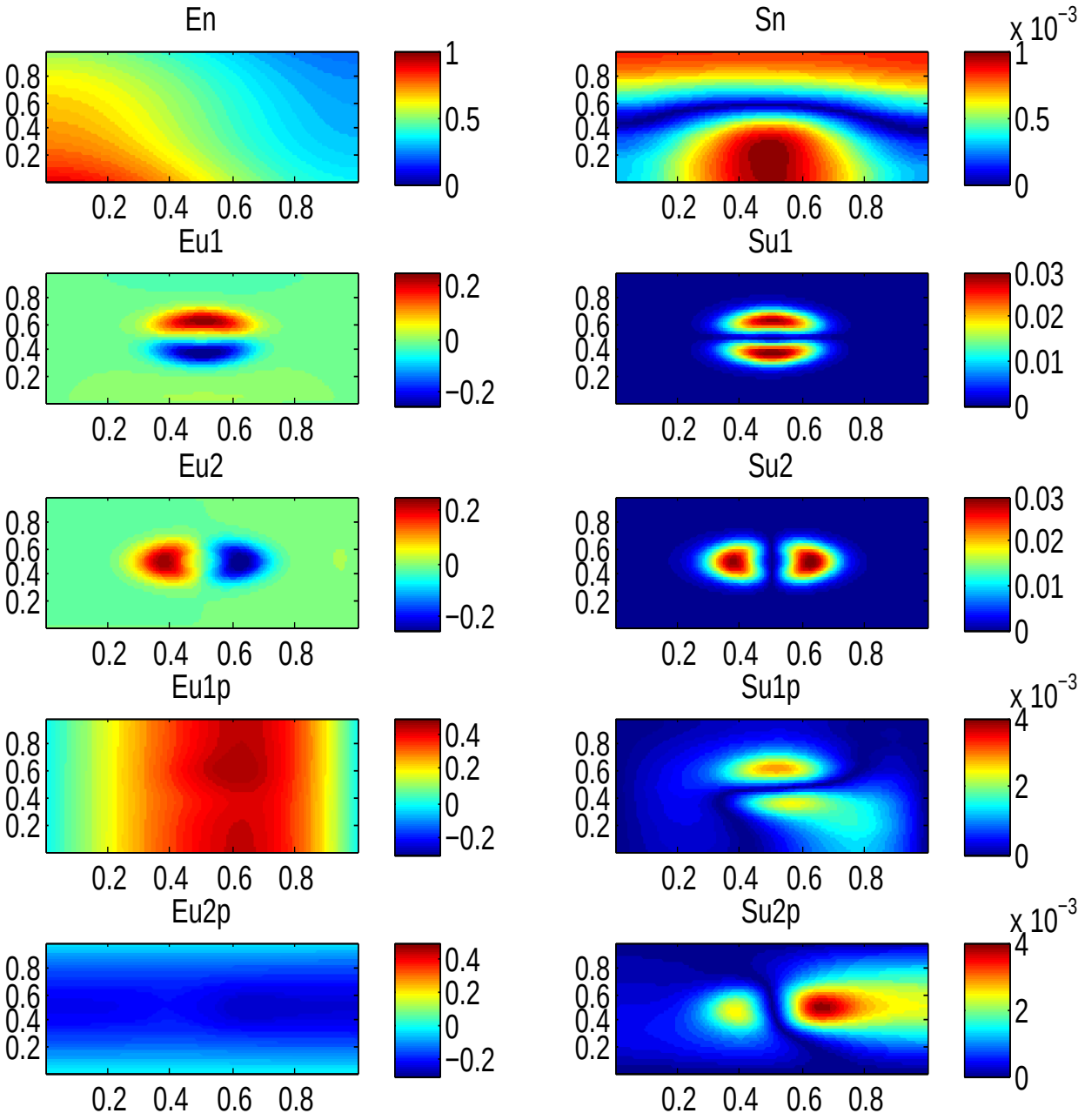

Figure 2.2 Example 1: $\epsilon = 1$. Left column: expectation. Right column: standard deviation. Rows: particle density n, fluid velocity u (two components), particle bulk velocity $\frac{J}{n}$ (two components).

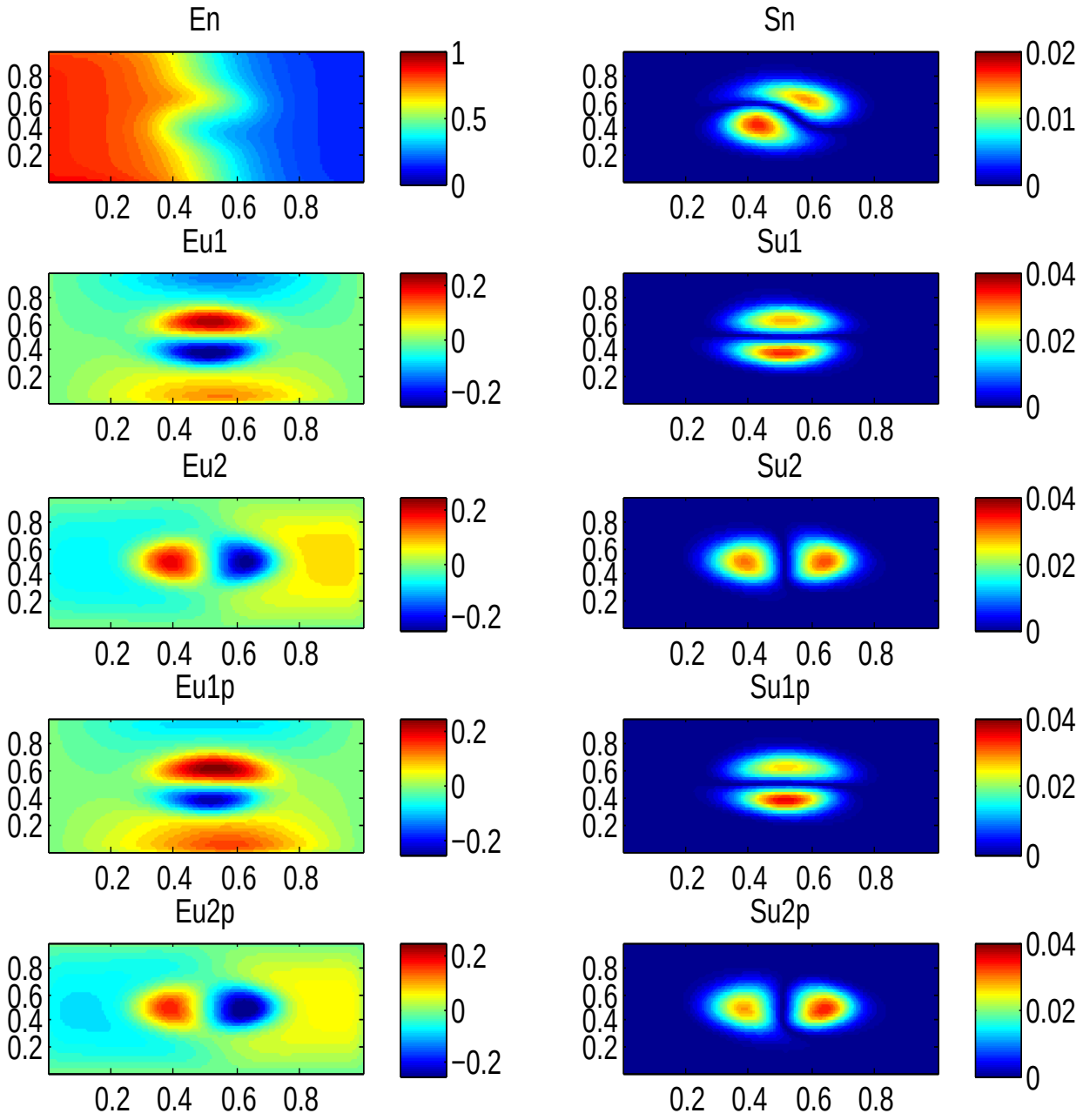

Figure 2.3 Example 1: $\epsilon = 0.01$. Left column: expectation. Right column: standard deviation. Rows: particle density n, fluid velocity u (two components), particle bulk velocity $\frac{J}{n}$ (two components).

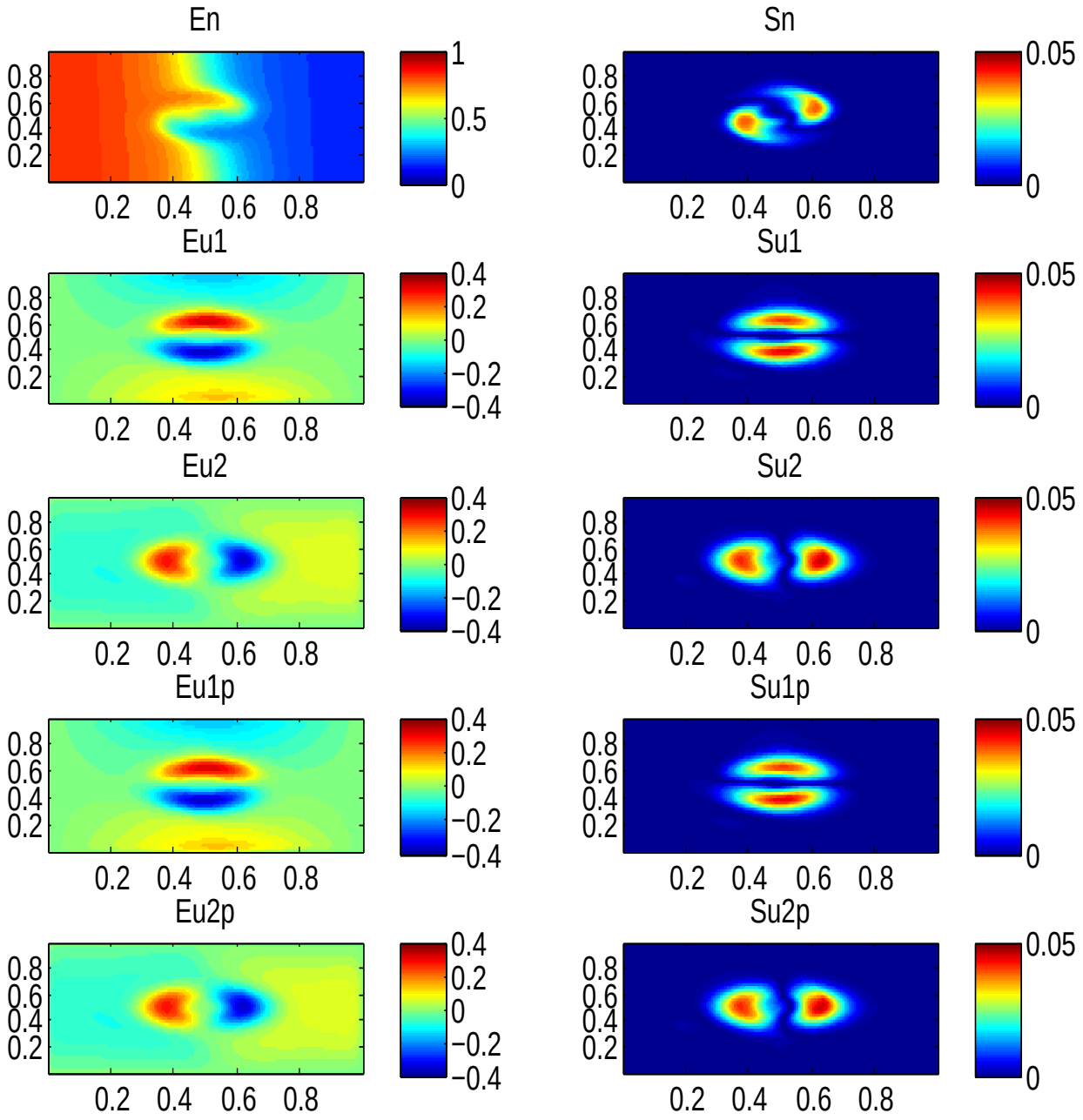

Figure 2.4 Example 1: $\epsilon = 10^{-8}$. Left column: expectation. Right column: standard deviation. Rows: particle density n, fluid velocity u (two components), particle bulk velocity $\frac{J}{n}$ (two components).

is verified because (1) smaller ϵ makes the fluid velocity and the particle bulk velocity closer, (2) the solution with $\epsilon = 0.01$ is close to the solution with $\epsilon = 10^{-8}$, which can be viewed as the solution of the limiting system since $\epsilon << \Delta t, \Delta x$.

One can also see that in all three cases, the fluid velocity is sensitive to the initial uncertainty, while for the particle density and bulk velocity, they are insensitive to the initial uncertainty for $\epsilon = 1$ case, and sensitive for the $\epsilon = 0.01, 10^{-8}$ cases. The reason for the insensitiveness in the kinetic regime is that the interaction between the particles and the fluid is weak, and the kinetic diffusion effect and the gravity dominate the motion of the particles.

We also compare the solution by the stochastic Galerkin method with the solution by a stochastic collocation (sC) method with 10 Gauss-Legendre collocation points. The results are compared at the 1d slice $x_1 = 0.55$ and shown in Figure 5.

One can see that the expectations computed by the two methods agree very well, while some standard deviations have some discrepancy due to their small magnitude.

To compare the efficiency of the sG method and the sC method, since the exact solution is expected to be smooth, one expects that the sG method with degree K polynomials will give the same accuracy as the Gauss-Legendre sC method with K collocation points in each random direction. Since the sC method is non-intrusive, one expects that for 1d random space sC is more efficient than sG. However, if one considers a d-dimensional random space, then sG requires $\binom{K+d}{d}$ basis functions, while sC with tensor grids requires K^d collocation points. From this one can see that sG can be much more efficient than sC if d is large. However, to seriously compare the efficiency of the two methods in very high dimension, one needs to utilize sparse grids in both methods. This is out of the scope of the paper.

Example 2: random initial particle density.

We take initial fluid velocity as zero and particle distribution as the equilibrium with density

$$n = 0.5 - 0.4 \arctan[10(x_1 - 0.5 + 0.1z + 0.2x_2)]/(\pi/2), \tag{2.63}$$

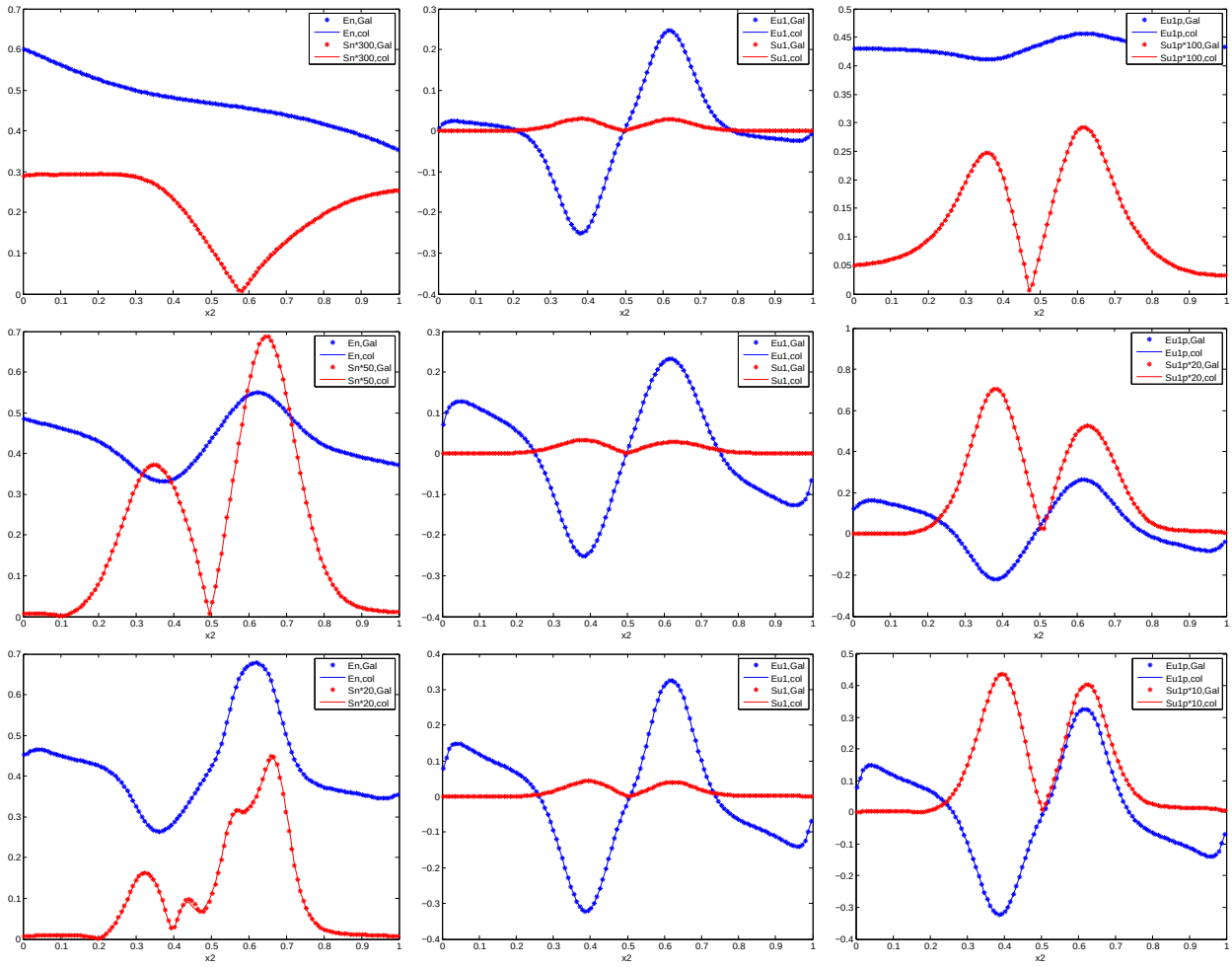

Figure 2.5 Example 1: at $x_1 = 0.55$. Upper: $\epsilon = 1$; middle: $\epsilon = 0.01$; lower: $\epsilon = 10^{-8}$. Curve: collocation; asterisks: Galerkin. Blue: expectation; red: standard deviation. From left to right: particle density n, fluid velocity u (first component), particle bulk velocity $\frac{J}{n}$ (first component). Some standard deviations are multiplied by constants to make them easy to observe, as noted on the figure.

and bulk velocity as zero. The particles still tend to fall to left bottom, and diffuse, and as a result, the fluid will be forced to rotate counter-clockwisely.

The expectation and standard deviation of the fluid velocity and the macroscopic quantities of the particles are shown in Figures 6, 7, 8 (for $\epsilon = 1, 0.01, 10^{-8}$ respectively).

For $\epsilon = 1$ one can see the expected behavior of the particles, as well as the rotation of the fluid. Notice that the fluid velocity is much smaller than the particle bulk velocity, and their shapes do not look similar. This is because for $\epsilon = 1$ the interaction between the fluid and the particles is weak. Also one can see the randomness on the initial density propagates into the particle bulk velocity and the fluid velocity.

As ϵ getting smaller, one can clearly see that the fluid velocity and the particle bulk velocity get closer, and they are rotating counter-clockwisely. Also, the solution with $\epsilon = 0.01$ is close to the solution with $\epsilon = 10^{-8}$. This verifies the s-AP property.

In all three cases, the particle density and bulk velocity are sensitive to the initial uncertainty, while the fluid velocity becomes sensitive only when ϵ gets small ($\epsilon = 0.01, 10^{-8}$). This is because there is no uncertainty in the initial fluid velocity, and only when ϵ is small, the interaction between the particles and the fluid is strong enough to propagate a significant amount of the initial uncertainty from the particles to the fluid.

Example 3: random initial particle density with the normal distribution.

In this example, the random variable $\tilde{z}$ obeys the normal distribution $\mathcal{N}(0, 1)$. We take initial fluid velocity as zero and particle distribution as the equilibrium with density

$$n = 0.5 - 0.2 \arctan[10(x_1 - 0.5 + 0.1\tilde{z} + 0.2x_2)]/(\pi/2), \tag{2.64}$$

and bulk velocity as zero. This initial data is similar to Example 2, so one expects similar physical behaviors. The expectation and standard deviation of the fluid velocity and the macroscopic quantities of the particles are shown in Figures 9, 10, 11 (for $\epsilon = 1, 0.01, 10^{-8}$ respectively). One can see that the result is similar to Example 2. One difference is that the standard deviations of Example 3 are smoother than the corresponding quantities in

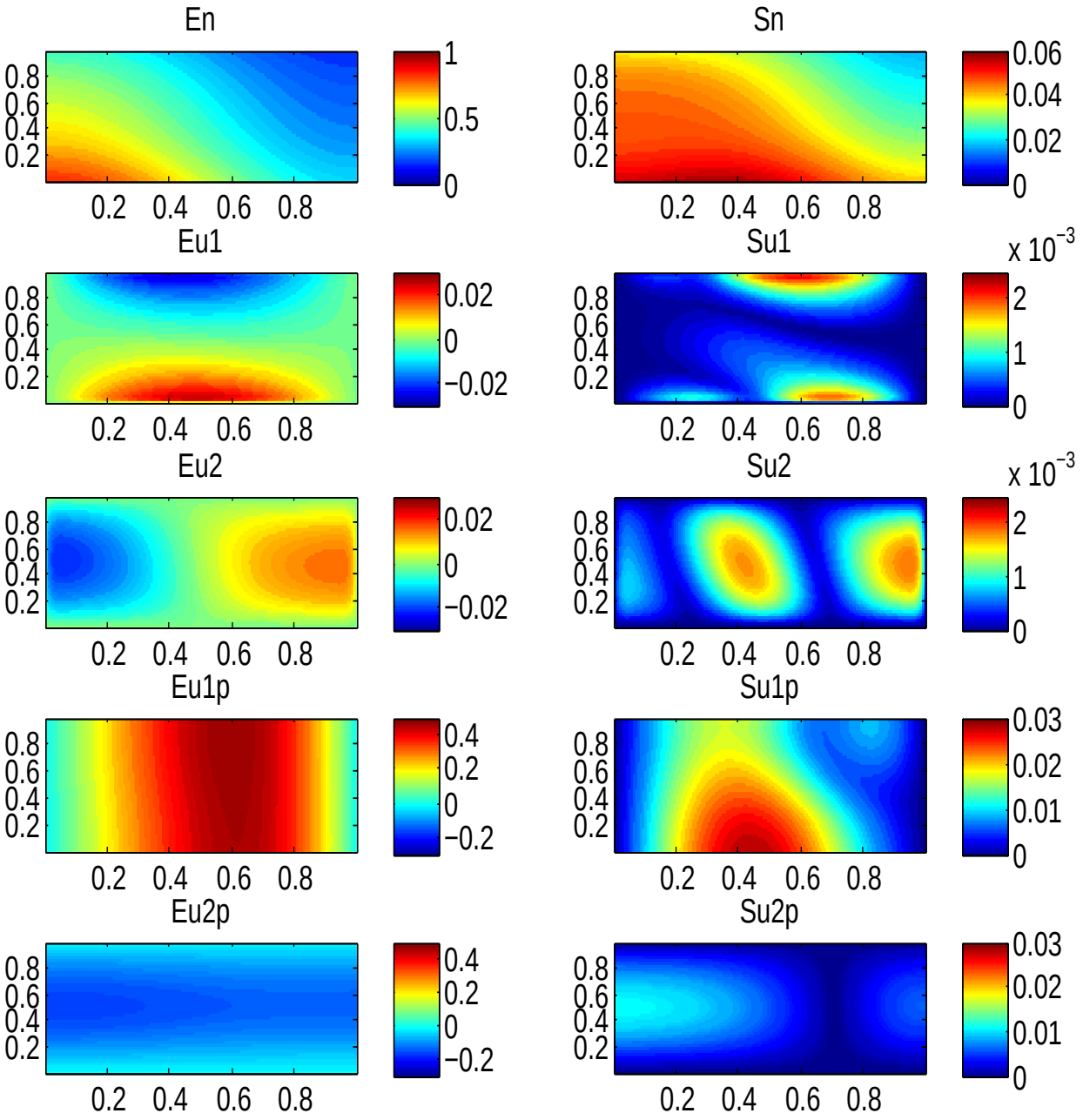

Figure 2.6 Example 2: $\epsilon = 1$. Left column: expectation. Right column: standard deviation. Rows: particle density n, fluid velocity u (two components), particle bulk velocity $\frac{J}{n}$ (two components).

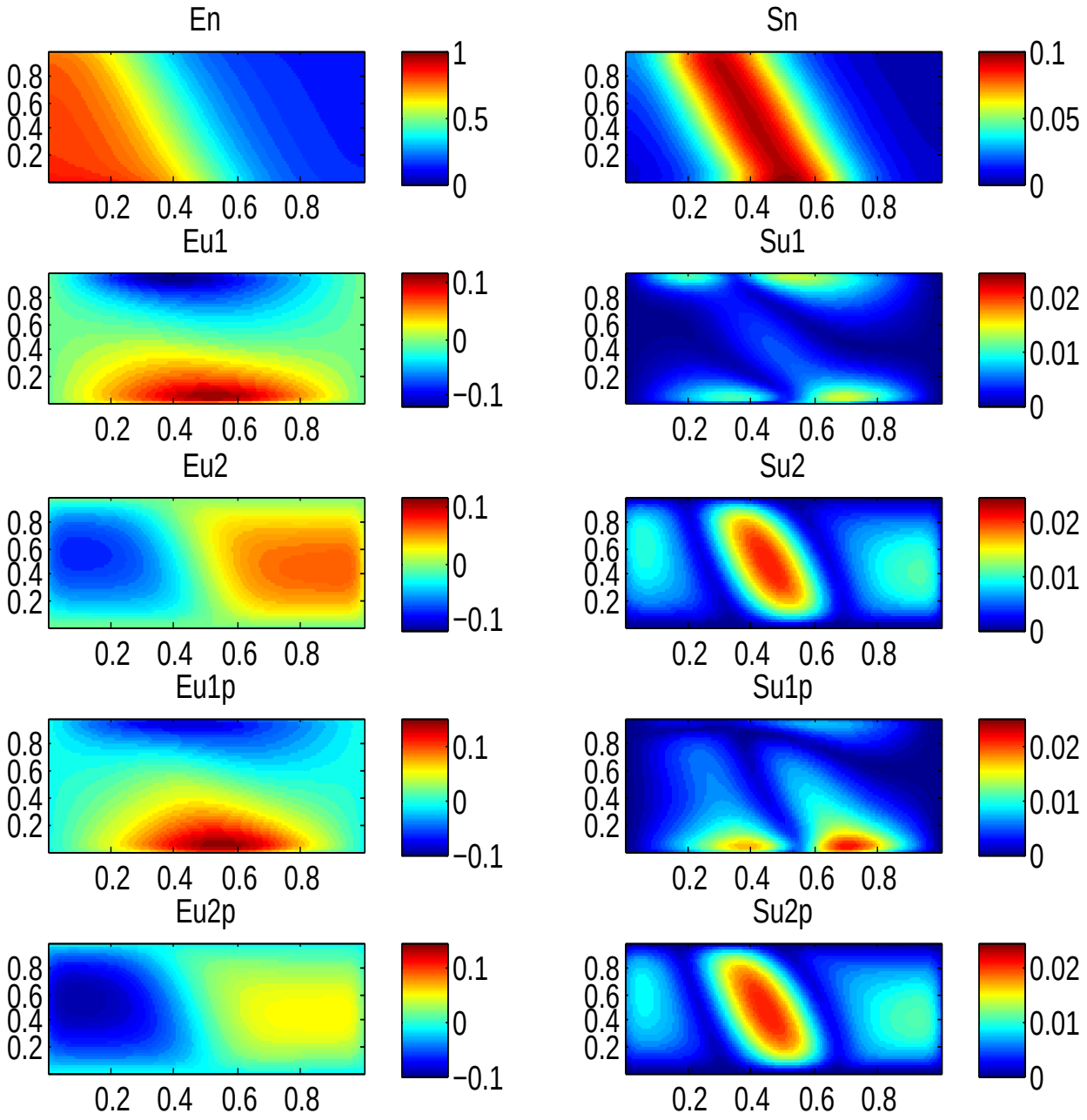

Figure 2.7 Example 2: $\epsilon = 0.01$. Left column: expectation. Right column: standard deviation. Rows: particle density n, fluid velocity u (two components), particle bulk velocity $\frac{J}{n}$ (two components).

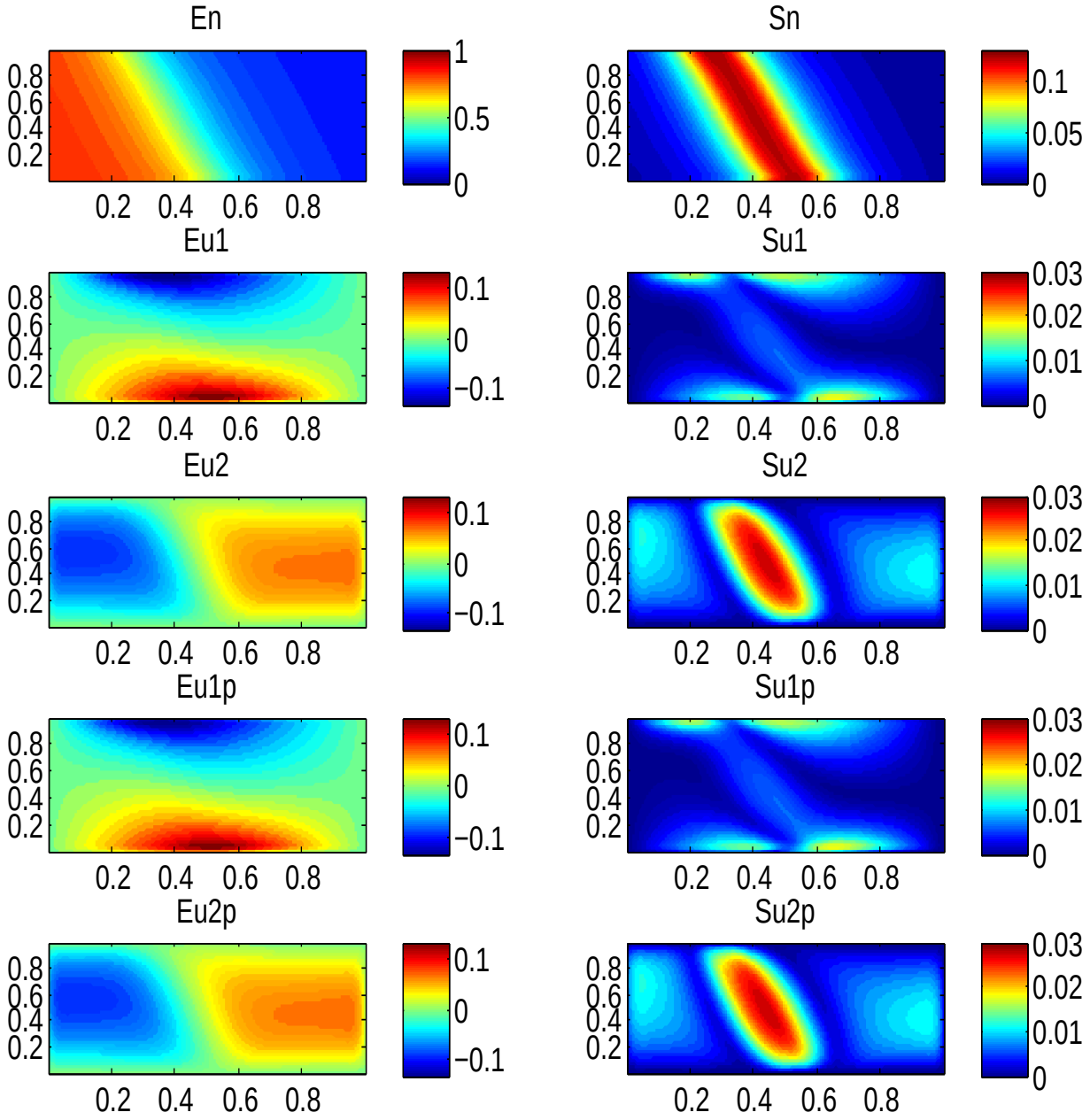

Figure 2.8 Example 2: $\epsilon = 10^{-8}$. Left column: expectation. Right column: standard deviation. Rows: particle density n, fluid velocity u (two components), particle bulk velocity $\frac{J}{n}$ (two components).

Example 2. This is because the Gaussian random variable $\tilde{z}$ ranges in $(-\infty, \infty)$, but the uniform random variable z ranges in $[-1, 1]$. In the initial data the random variables appear as the position of the sharp gradient of n. Therefore with a larger range of $\tilde{z}$, the uncertainty is less concentrated in the x-space.

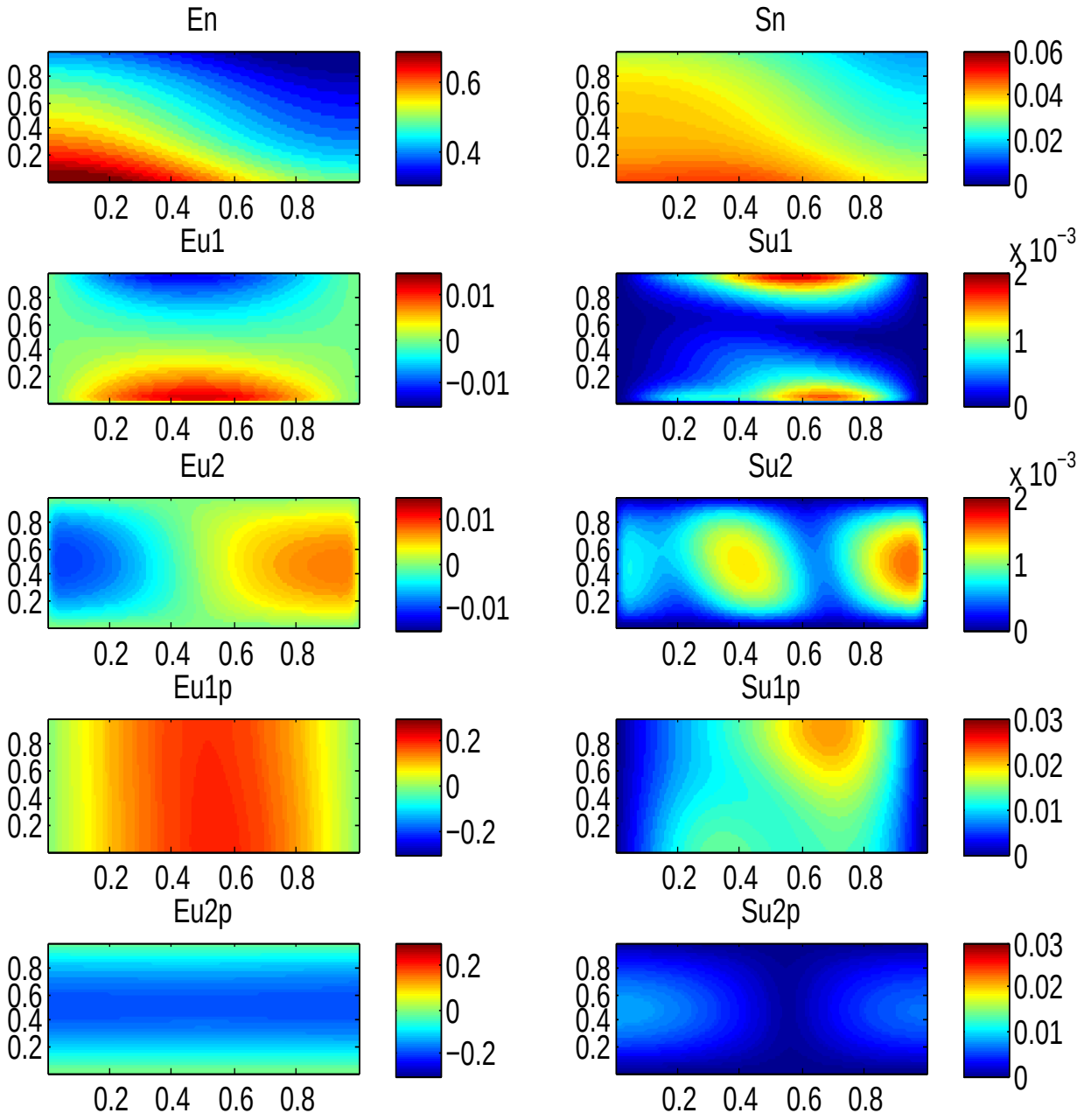

Figure 2.9 Example 3: $\epsilon = 1$. Left column: expectation. Right column: standard deviation. Rows: particle density n, fluid velocity u (two components), particle bulk velocity $\frac{J}{n}$ (two components).

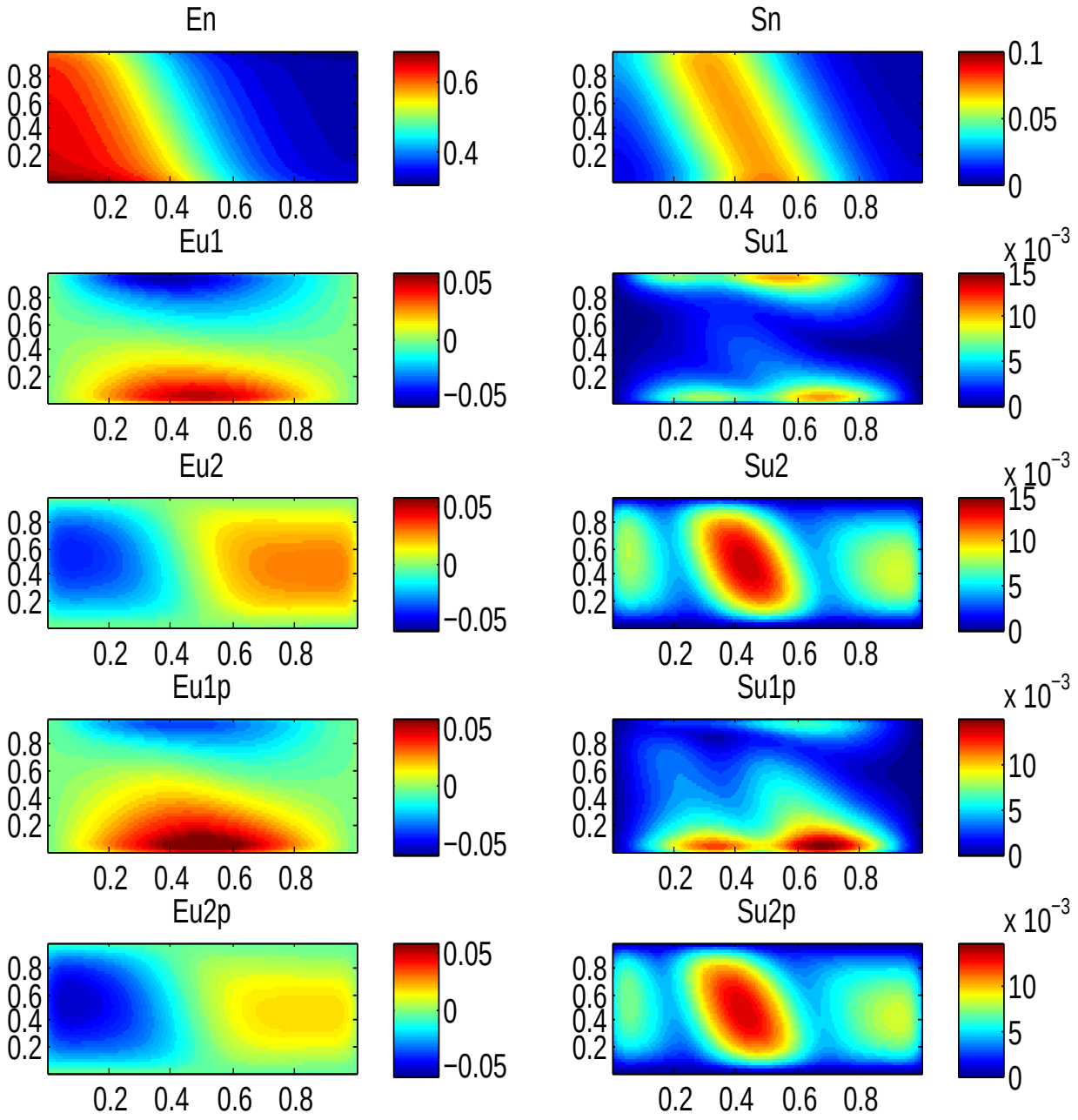

Figure 2.10 Example 3: $\epsilon = 0.01$. Left column: expectation. Right column: standard deviation. Rows: particle density n, fluid velocity u (two components), particle bulk velocity $\frac{J}{n}$ (two components).

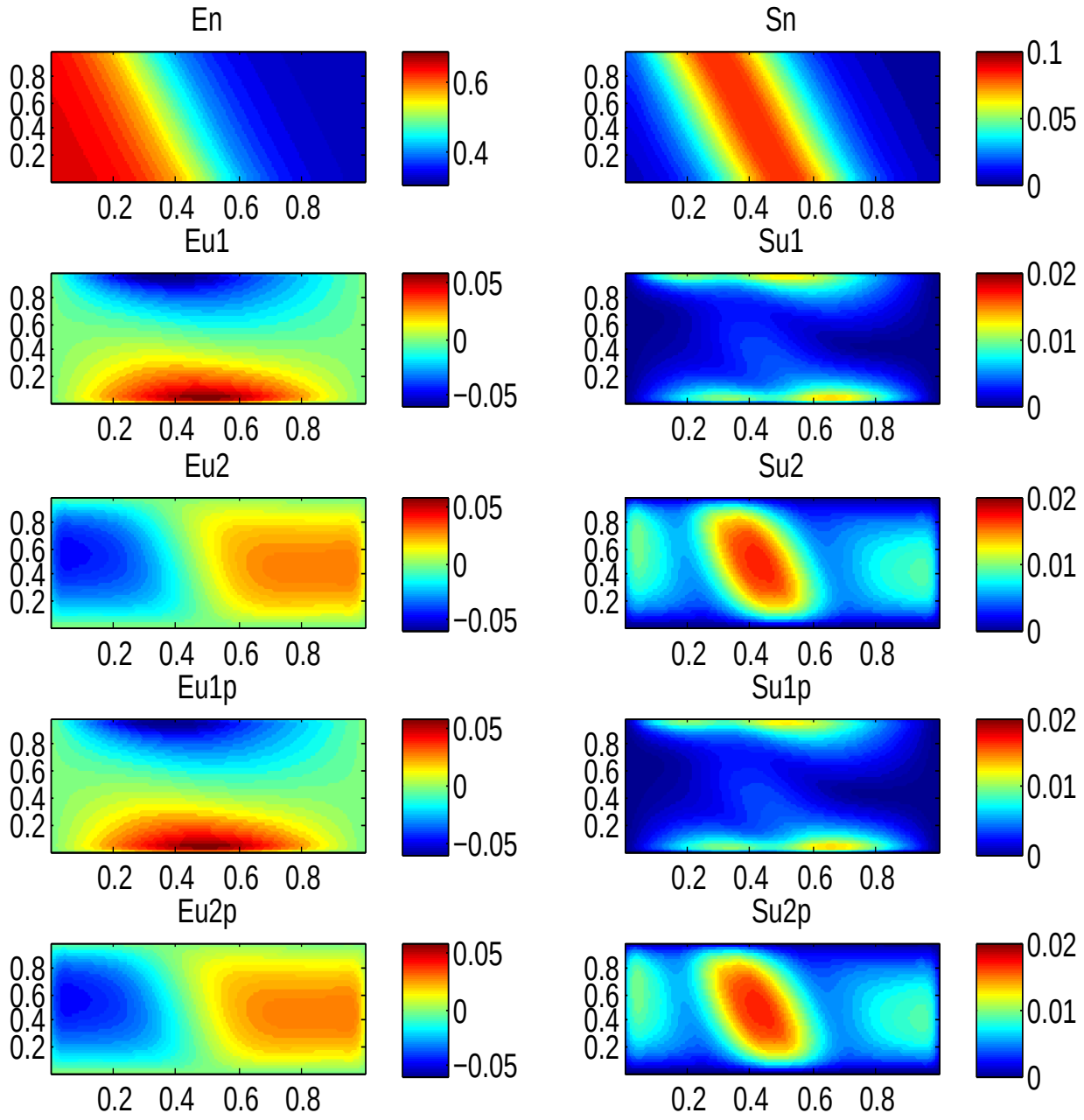

Figure 2.11 Example 3: $\epsilon = 10^{-8}$. Left column: expectation. Right column: standard deviation. Rows: particle density n, fluid velocity u (two components), particle bulk velocity $\frac{J}{n}$ (two components).

Chapter 3

Random space regularity and spectral accuracy of the gPC-sG method for the two-phase flow model in the light particle regime

In this chapter we go through the author's work with S. Jin on the random space regularity and spectral accuracy of the gPC-sG method for the two-phase flow model in the light particle regime (1.21).

3.1 Introduction

For simplicity the space is taken as $\mathbb{T}^3 = [-\pi, \pi]^3$ with periodic boundary condition. To be consistent with the notations in [95], we rewrite the model as

$$\begin{cases} u_t + u \cdot \nabla_x u + \nabla_x p - \Delta_x u = \dfrac{1}{\epsilon} \displaystyle\int (v - \epsilon u) F \, \mathrm{d}v, \\ \nabla_x \cdot u = 0, \\ F_t + \dfrac{1}{\epsilon} v \cdot \nabla_x F = \dfrac{1}{\epsilon^2} \nabla_v \cdot (\nabla_v F + (v - \epsilon u) F), \end{cases} \tag{3.1}$$

with initial data

$$u|_{t=0} = u_0, \quad \nabla_x \cdot u_0 = 0, \quad F|_{t=0} = F_0, \tag{3.2}$$

where $t \in \mathbb{R}^+$ is the time variable, $x \in \mathbb{T}^3$ is the space variable, and $v \in \mathbb{R}^3$ is the velocity variable. $u = u(t, x)$ is the velocity field of the fluid, and $F = F(t, x, v)$ is the distribution function of the particles. ϵ is the Knudsen number, which satisfies $0 < \epsilon \leq 1$. $\epsilon = O(1)$ corresponds to the kinetic regime, while $\epsilon \to 0$ corresponds to the fluid regime.

This system satisfies the following conservation properties:

$$\begin{aligned}
&\text{Mass conservation: } \frac{\mathrm{d}}{\mathrm{d}t}\int\int F\,\mathrm{d}v\,\mathrm{d}x = 0,\\
&\text{Momentum conservation: } \frac{\mathrm{d}}{\mathrm{d}t}\left(\int u\,\mathrm{d}x + \epsilon\int\int vF\,\mathrm{d}v\,\mathrm{d}x\right) = 0,\\
&\text{Energy/Entropy dissipation: } \frac{\mathrm{d}}{\mathrm{d}t}\left(\int\frac{|u|^2}{2}\,\mathrm{d}x + \int\int(F\ln F + \frac{|v|^2}{2}F)\,\mathrm{d}v\,\mathrm{d}x\right)\\
&\quad + \frac{1}{\epsilon^2}\int\int\frac{|(\epsilon u - v)F - \nabla_v F|^2}{F}\,\mathrm{d}v\,\mathrm{d}x + \int|\nabla_x u|^2\,\mathrm{d}x = 0.
\end{aligned} \tag{3.3}$$

As $\epsilon \to 0$, it is shown in [40] that (3.1) has a hydrodynamic limit

$$\begin{cases}
u_t + u\cdot\nabla_x u + \nabla_x p - \Delta_x u = 0,\\
\nabla_x\cdot u = 0,\\
\partial_t\rho + \nabla_x\cdot(u\rho - \nabla_x\rho) = 0,
\end{cases} \tag{3.4}$$

with $\rho(x) = \int F(x,v)\,\mathrm{d}v$ being the particle density, which is self-consistent Navier-Stokes equations for u, and a convection-diffusion equation for ρ with drift velocity u.

Goudon et al. [39] proved the first existence result of (3.1), in the case of kinetic regime ($\epsilon = O(1)$) and initial data near the global equilibrium, which means that F is close enough to the global Maxwellian

$$\mu(v) = \frac{1}{(2\pi)^{3/2}|\mathbb{T}^3|}e^{-|v|^2/2}, \tag{3.5}$$

and u is close to 0, in some suitable Sobolev spaces. In fact their method also works for small ϵ. They first write

$$F = \mu + \sqrt{\mu}f. \tag{3.6}$$

Then (3.1) becomes the following system for (u, f):

$$\begin{cases}
u_t + u\cdot\nabla_x u + \nabla_x p - \Delta_x u + u + \displaystyle\int\sqrt{\mu}uf\,\mathrm{d}v - \frac{1}{\epsilon}\int v\sqrt{\mu}f\,\mathrm{d}v = 0,\\
\nabla_x\cdot u = 0,\\
f_t + \dfrac{1}{\epsilon}v\cdot\nabla_x f + \dfrac{1}{\epsilon}(\nabla_v - \dfrac{v}{2})\cdot(uf) - \dfrac{1}{\epsilon}u\cdot v\sqrt{\mu} = \dfrac{1}{\epsilon^2}(\dfrac{-|v|^2}{4} + \dfrac{3}{2} + \Delta_v)f,
\end{cases} \tag{3.7}$$

with initial data

$$u|_{t=0} = u_0, \quad f|_{t=0} = f_0. \tag{3.8}$$

They assume that (u_0, f_0), the perturbation of initial data, satisfies the conditions

$$\int u_0 \,\mathrm{d}x + \int\int v\sqrt{\mu} f_0 \,\mathrm{d}v \,\mathrm{d}x = 0, \quad \nabla_x \cdot u_0 = 0, \tag{3.9}$$

$$\int\int \sqrt{\mu} f_0 \,\mathrm{d}v \,\mathrm{d}x = 0, \tag{3.10}$$

which mean that the perturbation does not affect the total momentum and mass, and the perturbation of the fluid velocity is divergence-free. Then, combining with a relation for the mean fluid velocity

$$\bar{u}(t) = \frac{1}{|\mathbb{T}^3|}\int u(t,x)\,\mathrm{d}x, \tag{3.11}$$

$$\bar{u}_t + 2\bar{u} + \frac{1}{|\mathbb{T}^3|}\int\int \sqrt{\mu}(uf)\,\mathrm{d}v\,\mathrm{d}x = 0, \tag{3.12}$$

which is a consequence of (3.9), using energy estimates, they proved the decay of an energy functional, defined as the summation of some suitable Sobolev norms, under the assumption that it is small enough initially. Then, by using hypocoercivity arguments, they proved that the L^2 norms of u and f decay exponentially in time, under some smoothness assumptions.

As introduced in Section 1.2 and 1.3, it is important to consider the system (3.7) with an extra random parameter z, and give estimates for the z-derivatives of the solution. Also, it is desirable to prove the uniform-in-(t, ϵ) spectral accuracy of the gPC-sG method for (3.7). We will assume the random space I_z is one-dimensional, for simplicity of notation. Our results can be extend to the case multi-dimensional random spaces. See more discussions at the end of Section 3.2. We also assume that the only random input is the initial data.

In this chapter, we first analyze the z-regularity of (3.7) for random initial data near the global equilibrium in some suitable Sobolev spaces (with derivatives with respect to x and z). We use energy estimates and hypocoercivity arguments similar to [39] on the z-derivatives of u and f. Our result implies that for near equilibrium initial data with regular dependence on x and z, the solution depends regularly on z for all time, and is insensitive to random perturbations on the initial data for large time. Then for the sG method, we consider the most popular choice of basis functions, the generalized polynomial chaos (gPC) [105], i.e.,

the orthonormal polynomials with respect to $\pi(z)\,\mathrm{d}z$. We write the equations for the gPC coefficients and do energy estimates. As mentioned in Section 1.3.2, using a weighted sum of the Sobolev norm of the gPC coefficients (Lemma 3.10), we manage to make this estimate independent of K, the number of basis functions. Finally we write the equations for the error of the gPC-sG method and do energy and hypocoercivity estimates. Our result implies that if the random initial data (u_0, f_0) is small enough in some suitable Sobolev spaces, then the gPC-sG method has spectral accuracy, uniformly in time and ϵ, and captures the exponential decay in time of the exact solution. An important feature of our results is that all the constants involved are *independent of* ϵ.

This chapter is organized as follows: in Section 3.2, we introduce some notations and state the main results; in Section 3.3 we prove the energy estimates for the z-derivatives of u and f; in Section 3.4 we use hypocoercivity arguments to prove the exponential decay of these derivatives; in Section 3.5 we prove the spectral accuracy of the sG method.

3.2 Notations and statements of main results

3.2.1 Notations

Due to the extra variable z (compared to [39]), our notation is slightly different from that in [39]. All the norms or inner products with a single bound (like $|\cdot|, \langle\cdot,\cdot\rangle$) are only integrated in x, v and pointwise in z (thus the result is a function in z). All the norms or inner products with a double bound (like $\|\cdot\|, \langle\langle\cdot,\cdot\rangle\rangle$) are integrated in all variables, thus the result is a number.

Let $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ be a multi-index. Then define

$$\partial^\alpha = \partial_{x_1}^{\alpha_1}\partial_{x_2}^{\alpha_2}\partial_{x_3}^{\alpha_3}. \tag{3.13}$$

The z-derivative of order γ of a function f is denoted by

$$f^\gamma = \partial_z^\gamma f. \tag{3.14}$$

There will not be any Sobelev norm with v-derivatives, so we do not give a short notation for them.

For function $u = u(x)$, $f = f(x, v)$, define the Sobolev norm (with x-derivatives)

$$\|u\|_s^2 = \sum_{|\alpha| \le s} \|\partial^\alpha u\|_{L_x^2}^2, \quad \|f\|_s^2 = \sum_{|\alpha| \le s} \|\partial^\alpha f\|_{L_{x,v}^2}^2. \tag{3.15}$$

In particular, $\|u\|_0$ denote the L_x^2 norm of u. For function $u = u(x, z)$, $f = f(x, v, z)$, define the sum of Sobolev norms

$$|u|_{s,r}^2 = \sum_{|\gamma| \le r} \|u^\gamma(\cdot, z)\|_s^2, \quad |f|_{s,r}^2 = \sum_{|\gamma| \le r} \|f^\gamma(\cdot, \cdot, z)\|_s^2, \tag{3.16}$$

where $|u|_{s,r}$ and $|f|_{s,r}$ are functions of z. Then define the expected value of the total Sobolev norm by

$$\|u\|_{s,r}^2 = \int |u|_{s,r}^2 \pi(z)\,\mathrm{d}z, \quad \|f\|_{s,r}^2 = \int |f|_{s,r}^2 \pi(z)\,\mathrm{d}z. \tag{3.17}$$

For function $\bar{u} = \bar{u}(z)$, define the sum of derivatives and the Sobolev norm by

$$|\bar{u}|_r^2 = \sum_{|\gamma| \le r} |\bar{u}^\gamma|^2, \quad \|\bar{u}\|_r^2 = \int |\bar{u}|_r^2 \pi(z)\,\mathrm{d}z. \tag{3.18}$$

In all these notations, the sub-index r is omitted when $r = 0$.

The L^2 inner product of functions defined on x-space of x, v-space will be denoted by $\langle \cdot, \cdot \rangle$, i.e.,

$$\langle f, g \rangle = \int f g\,\mathrm{d}x, \quad \text{or} \quad \langle f, g \rangle = \int\int f g\,\mathrm{d}v\,\mathrm{d}x. \tag{3.19}$$

In case the inputs also depend on z, $\langle f, g \rangle$ only integrates in x or (x, v), and the result is a function in z. For example,

$$\langle f, g \rangle(z) = \int f(x, z) g(x, z)\,\mathrm{d}x. \tag{3.20}$$

Then we introduce the inner products related to the hypocoercivity arguments. Define

$$\mathcal{K} = \nabla_v + \frac{v}{2}, \quad \mathcal{P} = v \cdot \nabla_x, \quad \mathcal{S}_i = [\mathcal{K}_i, \mathcal{P}] = \mathcal{K}_i \mathcal{P} - \mathcal{P} \mathcal{K}_i = \partial_{x_i}, \quad \mathcal{K}^* = -\nabla_v + \frac{v}{2}, \tag{3.21}$$

where $\mathcal{K}^*$ is the adjoint operator of $\mathcal{K}$, in the sense that $\langle \mathcal{K} f, g \rangle = \langle f, \mathcal{K}^* \cdot g \rangle$, where f has one component and g has three components .

For functions $f = f(x, v)$, $g = g(x, v)$, define

$$\begin{aligned}(f, g) &= 2\langle \mathcal{K}f, \mathcal{K}g\rangle + \epsilon\langle \mathcal{K}f, \mathcal{S}g\rangle + \epsilon\langle \mathcal{S}f, \mathcal{K}g\rangle + \epsilon^2\langle \mathcal{S}f, \mathcal{S}g\rangle,\\ [f, g] &= \langle \mathcal{K}f, \mathcal{K}g\rangle + \epsilon^2\langle \mathcal{S}f, \mathcal{S}g\rangle + \langle \mathcal{K}^2 f, \mathcal{K}^2 g\rangle + \epsilon^2\langle \mathcal{K}\mathcal{S}f, \mathcal{K}\mathcal{S}g\rangle,\end{aligned} \tag{3.22}$$

where we denote $\langle \mathcal{K}\mathcal{S}f, \mathcal{K}\mathcal{S}g\rangle := \sum_{i,j=1}^{3}\langle \mathcal{K}_i\mathcal{S}_j f, \mathcal{K}_i\mathcal{S}_j g\rangle$.

For functions $f = f(x, v, z)$, $g = g(x, v, z)$, define

$$(f, g)_{s,r} = \sum_{|\gamma|\le r}\sum_{|\alpha|\le s}(\partial^\alpha f^\gamma(\cdot, \cdot, z), \partial^\alpha g^\gamma(\cdot, \cdot, z)), \tag{3.23}$$

where $(f, g)_{s,r}$ is a function of z. Similarly define $[f, g]_{s,r}$.

Then we introduce the inner product in the (x, v, z) space:

$$\langle\langle f, g\rangle\rangle = \int \langle f, g\rangle\pi(z)\,\mathrm{d}z, \tag{3.24}$$

and similarly define $((f, g))$, $((f, g))_{s,r}$, $[[f, g]]$, $[[f, g]]_{s,r}$ as the corresponding inner products integrated in z. We also define the following norms in the (x, v, z) space:

$$\begin{aligned}\|u\|_{W^{s,\infty}} &= \max_{|\alpha|\le s}\|\partial^\alpha u\|_{L^\infty_{x,z}},\\ \|f\|_{W^{s,\infty}} &= \max_{|\alpha|\le s}\|\partial^\alpha f\|_{L^\infty_{x,z}(L^2_v)}.\end{aligned} \tag{3.25}$$

3.2.2 Regularity in the random space

Now we focus on the system (3.7) with the random variable z. In all of our results, the constants involved are *independent of* ϵ.

Both results in this subsection can be viewed as generalization of those in [39]. Our first main result is the following energy estimate assuming near equilibrium initial data:

Theorem 3.2.1. *Assume (u, f) solves (3.7) with initial data verifying (3.9). Fix a point z. Define the energy*

$$E(t; z) = E_{s,r}(t; z) = |u|^2_{H^{s,r}} + |f|^2_{s,r} + |\bar{u}|^2_r, \tag{3.26}$$

with integers $s \ge 2$ and $r \ge 0$. Then there exists a constant $c_1 = c_1(s, r) > 0$, such that $E(0; z) \le c_1$ implies that $E(t; z)$ is non-increasing in t.

This theorem is proved by an energy estimate on $\partial^\alpha f^\gamma$. This theorem means that for initial data near the global equilibrium, in the sense that $E(0;z)$ is small, the solution depends regularly in z for all time and all ϵ, and the z-derivatives are bounded uniformly in t and ϵ.

From now on we will omit the dependence on z of E, in case there is no confusion.

Next, by a standard hypocoercivity argument, we strengthen the above theorem into the following one:

Theorem 3.2.2. *Assume (u,f) solves (3.7) with initial data verifying (3.9) and (3.10). There exists a constant $c_1'(s,r)$ such that, if we assume $s \geq 0$, $E_{s+3,r}(0) \leq c_1'(s,r)$, and that $C^h_{s,r} = (f,f)_{s,r}|_{t=0}$ (defined by (3.23)) is finite, then there exists a constant $\lambda > 0$ such that*

$$E_{s,r}(t) \leq C(E_{s,r}(0) + C^h_{s,r})e^{-\lambda t}, \tag{3.27}$$

where $C = C(s,r)$.

This theorem implies that as long as the random perturbation (u_0, f_0) on the initial data is small in suitable Sobolev spaces and has vanishing total mass and momentum, the long-time behavior of the solution is not sensitive to the random initial data. The smallness condition is independent of ϵ.

3.2.3 Error estimate for the gPC-sG method

The gPC-sG method (as introduced in Section 1.2.1) for (3.7) reads

$$\begin{cases} \partial_t u_k + (u\cdot\nabla_x u)_k + \nabla_x p_k - \Delta_x u_k + u_k + \displaystyle\int \sqrt{\mu}(uf)_k \,\mathrm{d}v - \int v\sqrt{\mu} f_k \,\mathrm{d}v = 0, \\ \nabla_x \cdot u_k = 0, \\ \partial_t f_k + v\cdot\nabla_x f_k + (\nabla_v - \dfrac{v}{2})\cdot(uf)_k - u_k \cdot v\sqrt{\mu} = (\dfrac{-|v|^2}{4} + \dfrac{3}{2} + \Delta_v) f_k, \end{cases} \tag{3.28}$$

with initial data

$$u_k|_{t=0} = (u_0)_k = \int u_0 \phi_k(z)\pi(z)\,\mathrm{d}z, \quad f_k|_{t=0} = (f_0)_k. \tag{3.29}$$

Here the gPC coefficient of a product is given by

$$(uw)_k = \sum_{i,j=1}^{K} S_{ijk} u_i w_j, \tag{3.30}$$

where S_{ijk} as defined in (1.50).

The goal is to show that under smallness assumptions on the initial data, the gPC-sG method (3.28) has uniform-in-ϵ spectral accuracy for all K. We start from an energy estimate for (3.28). Although being similar to the original system (3.7), it indeed requires some new idea to obtain an estimate *independent of* K, i.e., the smallness requirement on the initial data is independent of K. The difficulty comes from the K^2 nonlinear terms appeared in the gPC product (3.30).

To overcome this difficulty, we introduce the technical condition (3.31), and introduce the weighted Sobolev norm $\sum_{k=1}^{K} \|k^q u_k\|_s^2$ (see the theorem below for detail). This idea originates from the analog between the gPC series and Fourier series. If one takes a function $\Phi = \Phi(y)$ defined on $y \in [-1, 1]$, then $\|\Phi\|_{H^r}^2 = \sum_k |\hat{\Phi}_k (1 + |k|^2)^{r/2}|^2$ where $\hat{\Phi}_k$ denotes the k-th Fourier coefficient. Therefore our weighted Sobolev norm is almost a Sobolev norm in both x and z spaces. With this viewpoint, it is natural to expect a nonlinear estimate with this norm (Lemma 3.10, the key nonlinear estimate), being *independent of* K, similar to the nonlinear estimate $\|uw\|_{H^s} \le C\|u\|_{H^s}\|w\|_{H^s}$ in the x-space. With the aid of this new technique, we prove

Theorem 3.2.3. *Assume the technical condition*

$$\|\phi_k\|_{L^\infty} \le C k^p, \quad \forall k, \tag{3.31}$$

with a parameter $p > 0$. Let $q > p + 2$ and $s \ge 2$. Let (u_k, f_k), $k = 1, \dots, K$, solve (3.28) with initial data verifying (3.9), and define the energy E^K by

$$E^K(t) = E^K_{s,q}(t) = \sum_{k=1}^{K} (\|k^q u_k\|_s^2 + \|k^q f_k\|_s^2 + |k^q \bar{u}_k|^2). \tag{3.32}$$

Then there exists a constant $c_2 = c_2(s, q) > 0$, independent of K, such that $E^K(0) \le c_2$ implies that $E^K(t)$ is decreasing in t.

Next we give a sufficient condition on the initial data, under which the assumption $E^K(0) \le c_2$ in Theorem 3.2.3 holds:

Proposition 3.1. *With the same assumptions as Theorem 3.2.3, the condition* $E^K_{s,q}(0) \le c_2(s,q)$ *holds if* $\|E_{s,r}(0)\|_{L^1_z} \le C c_2(s,q)$ *with* $r > q + \frac{1}{2}$, *and* $C = C(s,q,r)$.

Theorem 3.2.3 is proved by the same type of energy estimate as Theorem 3.2.1, with the aid of the nonlinear estimate Lemma 3.10. Notice that c_2 being independent of K is important, because it implies that the condition $E^K(0) \le c_2$ is in fact, in view of Proposition 3.1, a consequence of a smoothness condition on (u_0, f_0), for all K. This means for such initial data, the gPC-sG method is stable *for all* K.

We remark that (3.31) holds for gPC basis with respect to a large class of probability measures supported on a finite interval. To be precise, we have

Proposition 3.2. *Suppose* $I_z = [-R, R]$, $R < +\infty$ *with* $\pi(z)$ *satisfying* $1/\pi(z) \in L^{p_1}$ *for some* $p_1 > 0$. *Then (3.31) holds with* $p = 1 + 1/p_1$.

This proposition gives (3.31) for the uniform distribution on $[-1, 1]$ (with normalized Legendre polynomials as gPC basis), the distribution $\pi(z) = \frac{2}{\pi\sqrt{1-z^2}}$ on $[-1, 1]$ (with normalized Chebyshev polynomials as gPC basis), and all piecewise polynomial probability distributions on a finite interval with isolated zeros. More details about (3.31) can be found in Section 3.5.

Finally, by a combination of the above results, we obtain the spectral accuracy of the gPC-sG method, uniformly in t and ϵ, with a small initial data assumption on (u_0, f_0), independent of K and ϵ:

Theorem 3.2.4. *Assume (3.31) holds. Let* (u_k, f_k), $k = 1, \ldots, K$, *solve (3.28) with initial data verifying (3.9)(3.10). There exists a constant* $c_1''(s,r)$ *such that the following holds: Assume* $s \ge 0$, $r > p + \frac{5}{2}$, $\|E_{s+3,r}(0)\|_{L^\infty_z} \le c_1''(s,r)$, *and* $C^h_{s,r}$ *is finite. Then* E^e, *the energy of the gPC approximation error, defined by*

$$E^e = \|u^e\|_s^2 + \|f^e\|_s^2 + \|\bar{u}^e\|^2, \quad u^e = u - u^K, \quad f^e = f - f^K, \tag{3.33}$$

satisfies

$$E^e \leq \frac{C}{K^{2r}}, \tag{3.34}$$

for all time, i.e., the gPC-sG method has r-th order accuracy uniformly in time.

This theorem is proved by an energy estimate in the (x, v, z) space on (u^e, f^e) with the aid of the previous theorems.

Finally we prove that the error also decays exponentially in time, by a hypocoercivity argument:

Theorem 3.2.5. *Assume (3.31) holds. Let (u_k, f_k), $k = 1, \ldots, K$, solves (3.28) with initial data verifying (3.9)(3.10). There exists a constant $c_2''(s, r)$ such that the following holds: Assume $s \geq 0$, $r > p + \frac{5}{2}$, $\|E_{s+6,r}(0)\|_{L_z^\infty} \leq c_2''(s, r)$, and $C^h_{s+3,r}$ is finite. Then there exists a constant $\lambda^e > 0$ such that*

$$E^e \leq \frac{C}{K^{r-p-1/2}} e^{-\lambda^e t}. \tag{3.35}$$

These theorems imply that for random initial data near the global equilibrium, in the sense that (u_0, f_0) is small in some suitable Sobolev spaces, the gPC-sG method has spectral accuracy, uniformly in time and ϵ, and it captures the long-time behavior of (3.7) with random initial data.

Remark 3.3. *In cases where the random space I_z has dimension $d > 1$, the proof of Theorem 3.2.1 and Theorem 3.2.2 stays valid, but the results of other theorems may deteriorate due to:*

1. *The spectral accuracy of gPC approximation deteriorates. To be precise, suppose one takes the multi-dimensional gPC basis as the tensor product of one-dimensional ones, then the approximation error becomes $\frac{C}{K^{r/d}}$, where K is the number of basis functions.*
2. *The constant p in (3.31) will become pd in the case of tensor product basis.*

To investigate how d affects the estimate for the gPC-sG method is left as our future work.

3.3 Basic energy estimate: proof of Theorem 3.2.1

We first state some lemmas on nonlinear estimates. Denote the space of functions with finite $\|\cdot\|_s$ norm as

$$H^s = \{u(x) : \|u\|_s < \infty\}, \quad \tilde{H}^s = \{f(x,v) : \|f\|_s < \infty\}. \tag{3.36}$$

The following lemma is from [39]:

Lemma 3.4. *Let $u = u(x) \in H^s, w = w(x) \in H^s, f = f(x,v) \in \tilde{H}^s$. Then for $s > 3/2$,*

$$\|uw\|_s \le C\|u\|_s\|w\|_s, \tag{3.37}$$

$$\|uf\|_s \le C\|u\|_s\|f\|_s, \tag{3.38}$$

where $C = C(s)$.

It follows that

Lemma 3.5. *Let $u = u(x,z) \in W_z^{r,\infty}(H^s), w = w(x,z) \in W_z^{r,\infty}(H^s), f = f(x,v,z) \in W_z^{r,\infty}(\tilde{H}^s)$. Let $|\gamma| \le r$. Then for $s > 3/2$ and all z,*

$$|(uw)^\gamma|_s \le C|u|_{s,r}|w|_{s,r}, \tag{3.39}$$

$$|(uf)^\gamma|_s \le C|u|_{s,r}|f|_{s,r}, \tag{3.40}$$

where $C = C(s,r)$.

Proof. By the Leibniz rule,

$$(uw)^\gamma = \sum_{\beta=0}^{\gamma} \binom{\gamma}{\beta} u^\beta w^{\gamma-\beta}. \tag{3.41}$$

Then

$$|(uw)^\gamma|_s \le \sum_{\beta=0}^{\gamma} \binom{\gamma}{\beta} |u^\beta w^{\gamma-\beta}|_s \le C(s) \sum_{\beta=0}^{\gamma} \binom{\gamma}{\beta} |u^\beta|_s |w^{\gamma-\beta}|_s \le C(s,r)|u|_{s,r}|w|_{s,r}, \tag{3.42}$$

where the second inequality uses (3.37). This finishes the proof of (3.39). The proof of (3.40) is similar, in view of (3.38). □

And then a bilinear version follows:

Lemma 3.6. *Let* $u = u(x,z) \in W_z^{r,\infty}(H^s), w = w(x,z) \in W_z^{r,\infty}(H^s), y = y(x,z) \in W_z^{r,\infty}(H^s), f = f(x,v,z) \in W_z^{r,\infty}(\tilde{H}^s), g = g(x,v,z) \in W_z^{r,\infty}(\tilde{H}^s)$. *Let* $|\gamma| \le r$, $|\alpha| \le s$. *Then for* $s > 3/2$ *and all* z,

$$|\langle \partial^\alpha (uw)^\gamma, y^\gamma \rangle| \le C(\delta, s, r)|u|_{s,r}^2 |w|_{s,r}^2 + \delta |y|_{0,r}^2, \tag{3.43}$$

$$|\langle \partial^\alpha (uf)^\gamma, g^\gamma \rangle| \le C(\delta, s, r)|u|_{s,r}^2 |f|_{s,r}^2 + \delta |g|_{0,r}^2, \tag{3.44}$$

where δ *is any positive number.*

Proof. To prove (3.43),

$$\begin{aligned} ||\langle \partial^\alpha (uw)^\gamma, y^\gamma \rangle| &\le \frac{1}{4\delta} |\partial^\alpha (uw)^\gamma|_{L^2}^2 + \delta |y^\gamma|_{L^2}^2 \le \frac{1}{4\delta} |(uw)^\gamma|_s^2 + \delta |y|_{0,r}^2 \\ &\le C(\delta, s, r)|u|_{s,r}^2 |w|_{s,r}^2 + \delta |y|_{0,r}^2, \end{aligned} \tag{3.45}$$

where the first inequality uses Young's inequality, and the last inequality uses (3.39). The proof of (3.44) is similar. □

Proof of Theorem 3.2.1. Taking z-derivative of order γ and x-derivative of order α of (3.7), and taking z-derivative of order γ of (3.12) gives

$$\begin{aligned} &\partial_t \partial^\alpha u^\gamma + \partial^\alpha (u \cdot \nabla_x u)^\gamma + \nabla_x \partial^\alpha p^\gamma \underline{- \Delta_x \partial^\alpha u^\gamma} \underbrace{+ \partial^\alpha u^\gamma} + \int \sqrt{\mu} \partial^\alpha (uf)^\gamma \, \mathrm{d}v \underbrace{- \frac{1}{\epsilon} \int v \sqrt{\mu} \partial^\alpha f^\gamma \, \mathrm{d}v} = 0, \\ &\nabla_x \cdot \partial^\alpha u^\gamma = 0, \\ &\partial_t \partial^\alpha f^\gamma + \frac{1}{\epsilon} v \cdot \nabla_x \partial^\alpha f^\gamma + \frac{1}{\epsilon} (\nabla_v - \frac{v}{2}) \cdot \partial^\alpha (uf)^\gamma \underbrace{- \frac{1}{\epsilon} \partial^\alpha u^\gamma \cdot v \sqrt{\mu}} = \underbrace{\frac{1}{\epsilon^2} (\frac{-|v|^2}{4} + \frac{3}{2} + \Delta_v) \partial^\alpha f^\gamma}, \\ &\partial_t \bar{u}^\gamma \underline{+ 2\bar{u}^\gamma} + \frac{1}{|\mathbb{T}^3|} \int \int \sqrt{\mu} (uf)^\gamma \, \mathrm{d}v \, \mathrm{d}x = 0. \end{aligned} \tag{3.46}$$

Now do L^2 estimate on each equation above (except the second one), i.e., multiply the first equation by $\partial^\alpha u^\gamma$ and integrate in x; multiply the third equation by $\partial^\alpha f^\gamma$ and integrate in

(v,x); multiply the fourth equation by $\bar{u}^\gamma$. And then add the results together and sum over $|\gamma|\le r, |\alpha|\le s$. Then one gets the following equation (at each z):

$$\frac{1}{2}\partial_t E + G + B = 0, \tag{3.47}$$

where the energy E is given by (3.26). The good terms G are given by

$$G = \underline{G_1} + \underbrace{G_2} = \sum_{|\gamma|\le s} G_{1,\gamma} + \sum_{|\gamma|\le s} G_{2,\gamma}, \tag{3.48}$$

with

$$\begin{aligned} G_{1,\gamma} &= |\nabla_x u^\gamma|_s^2 + 2|\bar{u}^\gamma|^2 \ge C|u^\gamma|_{s+1}^2, \\ G_{2,\gamma} &= \left| u^\gamma\sqrt{\mu} - \frac{1}{\epsilon}\nabla_v f^\gamma - \frac{1}{\epsilon}\frac{v}{2} f^\gamma \right|_s^2, \end{aligned} \tag{3.49}$$

where the above inequality is by the Poincare-Wirtinger inequality. G_1 and G_2 come from the underlined terms and the underbraced terms in (3.46), respectively. To verify the G_2 term, we provide the following calculation:

$$\begin{aligned} &\langle \partial^\alpha u^\gamma, \partial^\alpha u^\gamma\rangle - \frac{1}{\epsilon}\langle v\sqrt{\mu}\partial^\alpha u^\gamma, \partial^\alpha f^\gamma\rangle - \frac{1}{\epsilon}\langle \partial^\alpha u^\gamma \cdot v\sqrt{\mu}, \partial^\alpha f^\gamma\rangle - \frac{1}{\epsilon^2}\langle (\frac{-|v|^2}{4} + \frac{3}{2} + \Delta_v)\partial^\alpha f^\gamma, \partial^\alpha f^\gamma\rangle \\ =&\langle \partial^\alpha u^\gamma\sqrt{\mu}, \partial^\alpha u^\gamma\sqrt{\mu}\rangle - 2\frac{1}{\epsilon}\langle \partial^\alpha u^\gamma\sqrt{\mu}, \frac{v}{2}\partial^\alpha f^\gamma\rangle - 2\frac{1}{\epsilon}\langle \partial^\alpha u^\gamma\sqrt{\mu}, \nabla_v\partial^\alpha f^\gamma\rangle \\ &+ \frac{1}{\epsilon^2}\langle \nabla_v\partial^\alpha f^\gamma + \frac{v}{2}\partial^\alpha f^\gamma, \nabla_v\partial^\alpha f^\gamma + \frac{v}{2}\partial^\alpha f^\gamma\rangle \\ =&\langle A_1, A_1\rangle - 2\langle A_1, A_3\rangle - 2\langle A_1, A_2\rangle + \langle A_2 + A_3, A_2 + A_3\rangle \\ =&|A_1 - A_2 - A_3|_0^2 = \left| \partial^\alpha \left(u^\gamma\sqrt{\mu} - \frac{1}{\epsilon}\nabla_v f^\gamma - \frac{1}{\epsilon}\frac{v}{2} f^\gamma \right) \right|_0^2, \end{aligned} \tag{3.50}$$

where we used integration by parts in v, $\nabla_v\sqrt{\mu} = -\frac{v}{2}\sqrt{\mu}$, and the notations

$$A_1 = \partial^\alpha u^\gamma\sqrt{\mu}, \quad A_2 = \frac{1}{\epsilon}\nabla_v\partial^\alpha f^\gamma, \quad A_3 = \frac{1}{\epsilon}\frac{v}{2}\partial^\alpha f^\gamma. \tag{3.51}$$

The notation $|\cdot|_0$ is interpreted by (3.16) with $s = r = 0$, i.e., taking $L^2_{x,v}$ norm for a fixed z.

The bad terms B are given by

$$B = B_1 + B_2 + B_3 = \sum_{|\gamma|\le r, |\alpha|\le s} B_{1,\alpha,\gamma} + \sum_{|\gamma|\le r, |\alpha|\le s} B_{2,\alpha,\gamma} + \sum_{|\gamma|\le r} B_{3,\gamma}, \tag{3.52}$$

with

$$\begin{aligned} B_{1,\alpha,\gamma} &= \langle \partial^\alpha (u\cdot\nabla_x u)^\gamma, \partial^\alpha u^\gamma\rangle, \\ B_{2,\alpha,\gamma} &= \left\langle \partial^\alpha (uf)^\gamma, \partial^\alpha \left[u^\gamma\sqrt{\mu} - \frac{1}{\epsilon}\nabla_v f^\gamma - \frac{1}{\epsilon}\frac{v}{2} f^\gamma \right] \right\rangle, \\ B_{3,\gamma} &= \frac{1}{|\mathbb{T}^3|}\langle (uf)^\gamma, \bar{u}^\gamma\sqrt{\mu}\rangle, \end{aligned} \tag{3.53}$$

coming from the nonlinear terms.

By using Lemma 3.6, the bad terms are controlled by

$$\begin{aligned} |B_{1,\alpha,\gamma}| &\le C(\delta)|u|^2_{s+1,r}|u|^2_{s,r} + \delta|u|^2_{s,r} \le C(\delta)EG_1 + \delta G_1, \\ |B_{2,\alpha,\gamma}| &\le C(\delta)|u|^2_{s,r}|f|^2_{s,r} + \delta\left| u^\gamma\sqrt{\mu} - \frac{1}{\epsilon}\nabla_v f^\gamma - \frac{1}{\epsilon}\frac{v}{2} f^\gamma \right|_s \le C(\delta)EG_1 + \delta G_2, \\ |B_{3,\gamma}| &\le C(\delta)|u|^2_{s,r}|f|^2_{s,r} + \delta|\bar{u}^\gamma|^2 \le C(\delta)EG_1 + \delta G_1. \end{aligned} \tag{3.54}$$

In conclusion, we have the energy estimate

$$\frac{1}{2}\partial_t E \le -(1 - C(\delta)E - C\delta)G. \tag{3.55}$$

Take $\delta = \frac{1}{4C}$ where C is the constant in (3.55), and $c_1(s,r) = \frac{1}{4C(\delta)}$. Then we will show that $E(t) \le c_1$ for all t. In fact, let

$$T^* = \sup\{\tilde{T} \ge 0 : \sup_{0\le t<\tilde{T}} E(t) \le c_1\}. \tag{3.56}$$

Then it follows that $E(t) \le c_1$ for $0 \le t \le T^*$. Then by our choice of δ and c_1,

$$1 - C(\delta)E - C\delta \ge 1 - \frac{1}{4} - \frac{1}{4} = \frac{1}{2}, \tag{3.57}$$

and therefore (3.55) implies

$$\partial_t E + G \le 0, \tag{3.58}$$

for $0 \le t \le T^*$. This prevents T^* from being finite. Thus we proved $E(t) \le c_1$ for all t, and as a result, (3.58) holds for all t. Thus $E(t)$ is decreasing in t. □

3.4 Hypocoercivity estimates: proof of Theorem 3.2.2

We will use the following lemma, which is Proposition 4.2 in [39]:

Lemma 3.7. *There exisits a constant $C > 0$ such that for $f(x,v) \in L^2_{x,v}$ orthogonal to $\sqrt{\mu}$, one has*

$$\|f\|^2_{L^2} \leq C(\|\mathcal{K}f\|^2_{L^2} + \|\mathcal{S}f\|^2_{L^2}). \tag{3.59}$$

We begin by proving the following lemma, which is indeed a modification of part of the proof of Proposition 4.1 in [39]:

Lemma 3.8. *For f and g orthogonal to $\sqrt{\mu}$,*

$$|(u \cdot \mathcal{K}^* f, g)| \leq C\frac{1}{\epsilon}\|u\|_{H^3}([f,f] + [g,g]). \tag{3.60}$$

Proof. Using the commutator relation

$$\mathcal{K}(u \cdot \mathcal{K}^* f) = (u \cdot \mathcal{K}^*)\mathcal{K}f + uf, \tag{3.61}$$

i.e.,

$$\mathcal{K}_i\left(\sum_{j=1}^3 u_j \mathcal{K}_j^* f\right) = \sum_{j=1}^3 u_j \mathcal{K}_j^* \mathcal{K}_i f + u_i f, \tag{3.62}$$

one gets

$$\begin{aligned}
(u \cdot \mathcal{K}^* f, g) =& 2\langle \mathcal{K}(u \cdot \mathcal{K}^* f), \mathcal{K}g\rangle + \epsilon\langle \mathcal{K}(u \cdot \mathcal{K}^* f), \mathcal{S}g\rangle + \epsilon\langle \mathcal{S}(u \cdot \mathcal{K}^* f), \mathcal{K}g\rangle + \epsilon^2\langle \mathcal{S}(u \cdot \mathcal{K}^* f), \mathcal{S}g\rangle \\
=& 2\langle u\mathcal{K}f, \mathcal{K}^2 g\rangle + 2\langle uf, \mathcal{K}g\rangle + \epsilon\langle u\mathcal{K}f, \mathcal{K}\mathcal{S}g\rangle + \epsilon\langle uf, \mathcal{S}g\rangle \\
&+ \epsilon\langle \mathcal{S}(uf), \mathcal{K}^2 g\rangle + \epsilon^2\langle \mathcal{S}(uf), \mathcal{S}\mathcal{K}g\rangle \\
=& 2\langle u\mathcal{K}f, \mathcal{K}^2 g\rangle + 2\langle uf, \mathcal{K}g\rangle + \epsilon\langle u\mathcal{K}f, \mathcal{K}\mathcal{S}g\rangle + \epsilon\langle uf, \mathcal{S}g\rangle \\
&+ \epsilon\langle (\mathcal{S}u)f, \mathcal{K}^2 g\rangle + \epsilon\langle u(\mathcal{S}f), \mathcal{K}^2 g\rangle + \epsilon^2\langle (\mathcal{S}u)f, \mathcal{S}\mathcal{K}g\rangle + \epsilon^2\langle u(\mathcal{S}f), \mathcal{K}\mathcal{S}g\rangle.
\end{aligned} \tag{3.63}$$

where the term $\langle \mathcal{S}(uf), \mathcal{S}\mathcal{K}g\rangle := \sum_{i,j=1}^3 \langle \mathcal{S}_i(u_j f), \mathcal{S}_i\mathcal{K}_j g\rangle$. Now use the Cauchy-Schwarz inequality, Lemma 3.7, and the Sobolev inequality

$$\|u\|_{L^\infty} + \|\nabla_x u\|_{L^\infty} \leq C\|u\|_{H^3}, \tag{3.64}$$

on each term. We provide the details for two of them and omit the others:

$$\begin{aligned}\langle uf, \mathcal{K}g\rangle \leq& \|u\|_{L^\infty}\|f\|_{L^2}\|\mathcal{K}g\|_{L^2} \leq C\|u\|_{L^\infty}(\|\mathcal{K}f\|_{L^2}+\|\mathcal{S}f\|_{L^2})\|\mathcal{K}g\|_{L^2}\\ \leq& C\|u\|_{L^\infty}(\|\mathcal{K}f\|_{L^2}^2+\|\mathcal{K}g\|_{L^2}^2+\epsilon\|\mathcal{S}f\|_{L^2}^2+\frac{1}{\epsilon}\|\mathcal{K}g\|_{L^2}^2),\\ \epsilon\langle uf, \mathcal{S}g\rangle \leq& \epsilon\|u\|_{L^\infty}\|f\|_{L^2}\|\mathcal{S}g\|_{L^2} \leq C\epsilon\|u\|_{L^\infty}(\|\mathcal{K}f\|_{L^2}+\|\mathcal{S}f\|_{L^2})\|\mathcal{S}g\|_{L^2}\\ \leq& C\epsilon\|u\|_{L^\infty}(\|\mathcal{K}f\|_{L^2}^2+\|\mathcal{S}g\|_{L^2}^2+\|\mathcal{S}f\|_{L^2}^2+\|\mathcal{S}g\|_{L^2}^2).\end{aligned} \tag{3.65}$$

Then one gets the conclusion, in view of the definition of $[\cdot,\cdot]$. □

Now we prove the following lemma, which is an analog to Proposition 4.1 of [39]:

Lemma 3.9. *Let the assumptions of Theorem 3.2.2 be fulfilled. Then there exists a constant $c_1'(s,r) \leq c_1(s+3,r)$ such that, if we assume that $E_{s+3,r}(0) \leq c_1'(s,r)$ is small enough, then there exists a constant $\lambda_1 > 0$ such that (at each z)*

$$\partial_t(f,f)_{s,r} + \lambda_1\frac{1}{\epsilon^2}[f,f]_{s,r} \leq C(\lambda_1)(|u|_{s,r}^2 + |\nabla_x u|_{s,r}^2 + \frac{1}{\epsilon^2}|\mathcal{K}f|_{s,r}^2). \tag{3.66}$$

Proof. One can write the evolution equation of $\partial^\alpha f^\gamma$ as

$$\partial_t\partial^\alpha f^\gamma + \frac{1}{\epsilon}\mathcal{P}\partial^\alpha f^\gamma + \frac{1}{\epsilon^2}(\mathcal{K}^*\cdot\mathcal{K})\partial^\alpha f^\gamma = \frac{1}{\epsilon}\partial^\alpha u^\gamma\cdot v\sqrt{\mu} + \frac{1}{\epsilon}\sum_{0\leq\eta\leq\alpha}\sum_{0\leq\beta\leq\gamma}\binom{\gamma}{\beta}\binom{\alpha}{\eta}\partial^\eta u^\beta\cdot\mathcal{K}^*\partial^{\alpha-\eta}f^{\gamma-\beta}. \tag{3.67}$$

We will take the $(\cdot,\cdot)$ inner product of (3.67) with $\partial^\alpha f^\gamma$. For the linear terms, by the same argument as the proof of Proposition 4.1 of [39], one gets

$$\begin{aligned}\frac{1}{\epsilon}(\mathcal{P}\partial^\alpha f^\gamma, \partial^\alpha f^\gamma) =& 2\frac{1}{\epsilon}\langle\mathcal{S}\partial^\alpha f^\gamma, \mathcal{K}\partial^\alpha f^\gamma\rangle + |\mathcal{S}\partial^\alpha f^\gamma|_0^2 \geq \frac{3}{4}|\mathcal{S}\partial^\alpha f^\gamma|_0^2 - 4\frac{1}{\epsilon^2}|\mathcal{K}\partial^\alpha f^\gamma|_0^2,\\ \frac{1}{\epsilon^2}(\mathcal{K}^*\cdot\mathcal{K}\partial^\alpha f^\gamma, \partial^\alpha f^\gamma) =& 2\frac{1}{\epsilon^2}|\mathcal{K}\partial^\alpha f^\gamma|_0^2 + 2\frac{1}{\epsilon^2}|\mathcal{K}^2\partial^\alpha f^\gamma|_0^2 + |\mathcal{S}\mathcal{K}\partial^\alpha f^\gamma|_0^2\\ &+ \frac{1}{\epsilon}\langle\mathcal{K}\partial^\alpha f^\gamma, \mathcal{S}\partial^\alpha f^\gamma\rangle + 2\frac{1}{\epsilon}\langle\mathcal{K}^2\partial^\alpha f^\gamma, \mathcal{S}\mathcal{K}\partial^\alpha f^\gamma\rangle\\ \geq& \frac{3}{2}\frac{1}{\epsilon^2}|\mathcal{K}\partial^\alpha f^\gamma|_0^2 + \frac{1}{2}\frac{1}{\epsilon^2}|\mathcal{K}^2\partial^\alpha f^\gamma|_0^2 + \frac{1}{3}|\mathcal{S}\mathcal{K}\partial^\alpha f^\gamma|_0^2 - \frac{1}{2}|\mathcal{S}\partial^\alpha f^\gamma|_0^2,\\ \frac{1}{\epsilon}|(\partial^\alpha u^\gamma\cdot v\sqrt{\mu}, \partial^\alpha f^\gamma)| =& |2\frac{1}{\epsilon}\langle\mathcal{K}(\partial^\alpha u^\gamma\cdot v\sqrt{\mu}), \mathcal{K}\partial^\alpha f^\gamma\rangle + \langle\mathcal{K}(\partial^\alpha u^\gamma\cdot v\sqrt{\mu}), \mathcal{S}\partial^\alpha f^\gamma\rangle\\ &+ \langle\mathcal{S}(\partial^\alpha u^\gamma\cdot v\sqrt{\mu}), \mathcal{K}\partial^\alpha f^\gamma\rangle + \epsilon\langle\mathcal{S}(\partial^\alpha u^\gamma\cdot v\sqrt{\mu}), \mathcal{S}\partial^\alpha f^\gamma\rangle|\\ \leq& \delta(\frac{1}{\epsilon^2}|\mathcal{K}\partial^\alpha f^\gamma|_0^2 + |\mathcal{S}\partial^\alpha f^\gamma|_0^2) + C(\delta)(|u|_{s,r}^2 + |\nabla_x u|_{s,r}^2).\end{aligned} \tag{3.68}$$

The notation $|\cdot|_0$ is interpreted by (3.16) with $s = r = 0$, i.e., taking $L^2_{x,v}$ norm for a fixed z. For the nonlinear term (the summation), we apply Lemma 3.8 and get

$$\begin{aligned}
&\frac{1}{\epsilon}|(\partial^\eta u^\beta \cdot \mathcal{K}^* \partial^{\alpha-\eta} f^{\gamma-\beta}, \partial^\alpha f^\gamma)| \\
\leq & C\frac{1}{\epsilon^2}|\partial^\eta u^\beta|_{3,0}([\partial^{\alpha-\eta} f^{\gamma-\beta}, \partial^{\alpha-\eta} f^{\gamma-\beta}] + [\partial^\alpha f^\gamma, \partial^\alpha f^\gamma]) \\
\leq & C\frac{1}{\epsilon^2}|u|_{s+3,r}[f,f]_{s,r},
\end{aligned} \tag{3.69}$$

where we used the fact that the x and z derivatives commute with the operators $\mathcal{K}$ and $\mathcal{S}$. With these estimates, we get

$$\begin{aligned}
&\frac{1}{2}\partial_t(\partial^\alpha f^\gamma, \partial^\alpha f^\gamma) + \frac{1}{2}\frac{1}{\epsilon^2}|\mathcal{K}^2\partial^\alpha f^\gamma|_0^2 + \frac{1}{3}|\mathcal{S}\mathcal{K}\partial^\alpha f^\gamma|_0^2 + \frac{1}{4}|\mathcal{S}\partial^\alpha f^\gamma|_0^2 - \frac{5}{2}\frac{1}{\epsilon^2}|\mathcal{K}\partial^\alpha f^\gamma|_0^2 \\
&\quad \leq \delta(\frac{1}{\epsilon^2}|\mathcal{K}\partial^\alpha f^\gamma|_0^2 + |\mathcal{S}\partial^\alpha f^\gamma|_0^2) + C(\delta)(|u|_{s,r}^2 + |\nabla_x u|_{s,r}^2) + C\frac{1}{\epsilon^2}|u|_{s+3,r}[f,f]_{s,r}.
\end{aligned} \tag{3.70}$$

Then we choose $\delta = 1/8$ to absorb the term $|\mathcal{S}\partial^\alpha f^\gamma|_0^2$ on the RHS by the same term on the LHS. Summing over α, γ, we get

$$\partial_t(f,f)_{s,r} + (\frac{1}{8} - C_1|u|_{s+3,r})\frac{1}{\epsilon^2}[f,f]_{s,r} \leq C_2(|u|_{s,r}^2 + |\nabla_x u|_{s,r}^2 + \frac{1}{\epsilon^2}|\mathcal{K}f|_{s,r}^2), \tag{3.71}$$

where $C_1 = NC$, $C_2 = \max\{3, NC(\delta)\}$, N being the number of possible pairs (α, γ).

Thus if one chooses $c_1' = \min\{c_1(s+3, r), \frac{1}{16C_1}\}$, then by Theorem 3.2.1, $E_{s+3,r}(t)$ is decreasing, so $E_{s+3,r}(t) \leq c_1'$ for all t. Thus $|u|_{s+3,r} \leq E_{s+3,r} \leq c_1'$ for all t, and one gets the conclusion, with $\lambda_1 = 1/16$.

□

Proof of Theorem 3.2.2. To obtain the energy decay estimate, we write

$$\begin{aligned}
G &= |\nabla_x u|_{s,r}^2 + 2|\bar{u}|_r^2 + |u\sqrt{\mu} - \frac{1}{\epsilon}\mathcal{K}f|_{s,r}^2 \\
&\geq |\bar{u}|_r^2 + 2\lambda_2|u|_{s+1,r}^2 + |u\sqrt{\mu} - \frac{1}{\epsilon}\mathcal{K}f|_{s,r}^2 \\
&\geq |\bar{u}|_r^2 + \lambda_2|u|_{s+1,r}^2 + \frac{1}{2}|u\sqrt{\mu} - \frac{1}{\epsilon}\mathcal{K}f|_{s,r}^2 + \lambda_3\frac{1}{\epsilon^2}|\mathcal{K}f|_{s,r}^2,
\end{aligned} \tag{3.72}$$

where $\lambda_3 = \min\{\frac{\lambda_2}{2}, \frac{1}{4}\}$. The first inequality is by the Poincare-Wirtinger inequality. The second inequality is because

$$\begin{aligned} |\frac{1}{\epsilon}\mathcal{K}f|_{s,r}^2 &= |(\frac{1}{\epsilon}\mathcal{K}f - u\sqrt{\mu}) + u\sqrt{\mu}|_{s,r}^2 \leq 2(|\frac{1}{\epsilon}\mathcal{K}f - u\sqrt{\mu}|_{s,r}^2 + |u\sqrt{\mu}|_{s,r}^2) \\ &= 2(|\frac{1}{\epsilon}\mathcal{K}f - u\sqrt{\mu}|_{s,r}^2 + |u|_{s,r}^2). \end{aligned} \tag{3.73}$$

Thus, by adding to (3.58) some positive constant λ_4 (to be chosen) times (3.66), we have

$$\partial_t \tilde{E} + \tilde{G} \leq \lambda_4 \tilde{B}, \tag{3.74}$$

where

$$\tilde{E} = E + \lambda_4 (f, f)_{s,r}, \quad \tilde{G} = G + \lambda_4 \lambda_1 \frac{1}{\epsilon^2}[f, f]_{s,r}, \tag{3.75}$$

$$\tilde{B} = C(\lambda_1)(|u|_{s,r}^2 + |\nabla_x u|_{s,r}^2 + \frac{1}{\epsilon^2}|\mathcal{K}f|_{s,r}^2). \tag{3.76}$$

It is clear from (3.72) that $\tilde{B} \leq CG \leq C\tilde{G}$. Thus by choosing $\lambda_4 = \min\{\frac{1}{2C}, 1\}$, C being the previous constant, we get

$$\partial_t \tilde{E} + \frac{1}{2}\tilde{G} \leq 0. \tag{3.77}$$

Notice that Lemma 3.7 implies that

$$|f|_{s,r}^2 \leq C(|\mathcal{K}f|_{s,r}^2 + |\mathcal{S}f|_{s,r}^2), \tag{3.78}$$

and by definition one also has

$$(f, f)_{s,r} \leq C(|\mathcal{K}f|_{s,r}^2 + |\mathcal{S}f|_{s,r}^2) \leq C\frac{1}{\epsilon^2}(f, f)_{s,r}. \tag{3.79}$$

Thus

$$\tilde{E} \leq C(G + |f|_{s,r}^2) + \lambda_4((f, f))_{s,r} \leq C(G + |\mathcal{K}f|_{s,r}^2 + |\mathcal{S}f|_{s,r}^2) \leq C\tilde{G}. \tag{3.80}$$

This together with (3.77) implies

$$\tilde{E}(t) \leq \tilde{E}(0)e^{-\lambda t}, \tag{3.81}$$

where $\lambda = \frac{1}{2C}$, C being the constant in (3.80).

Finally, the proof of Theorem 3.2.2 is finished by noticing that

$$E(t) \leq \tilde{E}(t) \leq \tilde{E}(0)e^{-\lambda t} \leq (E(0) + C^h)e^{-\lambda t}. \tag{3.82}$$

□

3.5 Proof of spectral accuracy of the gPC-sG approximation

In order to prove the accuracy of the gPC-sG method, we first prove Theorem 3.2.3, which is an energy estimate for the gPC coefficients (u_k, f_k).

3.5.1 Estimate of the gPC coefficients: proof of Theorem 3.2.3

In this section, all the norms and inner products acting on ϕ_k are taken on the random space I_z, and with respect to the measure $\pi(z)\,\mathrm{d}z$ if not stated otherwise. In order to prove the estimate for the gPC coefficients, we need the extra assumption (3.31) on the basis functions.

We mention some special cases where (3.31) is satisfied [97]. For the case $I_z = [-1, 1]$ with uniform distribution, ϕ_k are the normalized Legendre polynomials, and (3.31) holds with $p = 1/2$. For the case $I_z = [-1, 1]$ with the distribution $\pi(z) = \frac{2}{\pi\sqrt{1-z^2}}$, ϕ_k are the normalized Chebyshev polynomials, and (3.31) holds with $p = 0$.

Now we prove Proposition 3.2, which guarantees (3.31) for gPC basis with respect to a large class of probability measures supported on a finite interval.

Proof of Proposition 3.2. First, if $\{\phi_k\}$ is the gPC basis for the probability measure $\pi(z)\,\mathrm{d}z$ on $[-R, R]$, then $\{\phi_k(R\cdot)\}$ is the gPC basis for the probability measure $R\pi(Rz)\,\mathrm{d}z$ on $[-1, 1]$. Therefore we can assume $R = 1$ without loss of generality.

Let $\Phi(z)$ be any degree k polynomial on $I_z = [-1, 1]$ with $\int_{I_z} \Phi(z)^2 \frac{1}{2}\,\mathrm{d}z = 1$, i.e., has norm 1 in the L^2 space with uniform distribution $\frac{1}{2}\,\mathrm{d}z$. Then one can expand it into series of normalized Legendre polynomials $\{\psi_j\}$:

$$\Phi(z) = \sum_{j=1}^{k+1} \Phi_j \psi_j(z), \quad \sum_{j=1}^{k+1} \Phi_j^2 = 1. \tag{3.83}$$

Then it follows that

$$|\Phi(z)| \le \left(\sum_{j=1}^{k+1} \Phi_j^2\right)^{1/2} \left(\sum_{j=1}^{k+1} \psi_j(z)^2\right)^{1/2} \le Ck, \tag{3.84}$$

by the fact that $\{\psi_j\}$ satisfies (3.31) with $p = 1/2$. Thus $\|\Phi\|_{L^\infty} \le Ck$.

Take $\Phi = \frac{\phi_k}{\|\phi_k\|_{L^2(1/2\,\mathrm{d}z)}}$, we obtain

$$\|\phi_k\|_{L^\infty} \le Ck\|\phi_k\|_{L^2(1/2\,\mathrm{d}z)}. \tag{3.85}$$

Next writing $p_2 = p_1 + 1 > 1$, $p_2' = p_2/(p_2-1) = 1 + 1/p_1$, we estimate

$$\begin{aligned}
\|\phi_k\|_{L^2(1/2\,\mathrm{d}z)} &= \left(\int \phi_k(z)^2 \frac{1}{2\pi(z)}\pi(z)\,\mathrm{d}z\right)^{1/2} \\
&\le \left(\int |\phi_k(z)|^{2p_2'}\pi(z)\,\mathrm{d}z\right)^{1/(2p_2')} \left(\int \frac{1}{(2\pi(z))^{p_2}}\pi(z)\,\mathrm{d}z\right)^{1/(2p_2)} \\
&\le C\|\phi_k\|_{L^\infty}^{(p_2'-1)/p_2'}\|\phi_k\|_{L^2(\pi(z)\,\mathrm{d}z)}^{1/p_2'} \\
&= C\|\phi_k\|_{L^\infty}^{(p_2'-1)/p_2'},
\end{aligned} \tag{3.86}$$

where we use the assumption that $\int \pi(z)^{1-p_2}\,\mathrm{d}z < \infty$ in the second inequality, and $\|\phi_k\|_{L^2(\pi(z)\,\mathrm{d}z)} = 1$ in the last equality. Combining with (3.85) we conclude the proof. □

This proposition gives (3.31) for a large class of probability measures on a finite interval. For example, if $\pi(z)$ is continuous and has only finite number of zeros, with $\pi(z - z_0) \ge c|z - z_0|^{p_3}$ for some $p_3 > 0$, $c > 0$ near any zero $z = z_0$, then the condition of Lemma 3.2 is satisfied with any $p_1 < 1/p_3$. This already includes all piecewise polynomial weights with separated zeros.

It follows from (3.31) that

$$|S_{ijk}| \le Ci^p, \tag{3.87}$$

since

$$|S_{ijk}| \le \|\phi_i\|_{L^\infty}\langle|\phi_j|, |\phi_k|\rangle \le \|\phi_i\|_{L^\infty}\|\phi_j\|_{L^2}\|\phi_k\|_{L^2} \le Ci^p. \tag{3.88}$$

Also, note that ϕ_k is a polynomial of degree $k-1$, orthogonal to all lower order polynomials. If $(i-1) + (j-1) < k-1$, then $S_{ijk} = 0$. Thus S_{ijk} may be nonzero only when the triangle inequality

$$i + j \ge k + 1, \tag{3.89}$$

holds.

Note that due to the symmetry in i, j, k of S_{ijk}, (3.87) and (3.89) also hold if i, j, k are permuted.

Then we have the following lemma, which is the key nonlinear estimate:

Lemma 3.10. *Assume condition (3.87). Let $q > p + 2$. Let $s > \frac{3}{2}$, α be a multi-index with $|\alpha| \le s$. Let $u_k = u_k(x) \in H^s, w_k = w_k(x) \in H^s, y_k = y_k(x) \in L^2, f_k = f_k(x, v) \in \tilde{H}^2, g_k = g_k(x, v) \in L^2$. Then*

$$\left| \sum_{k=1}^{K} k^{2q} \langle \partial^\alpha (uw)_k, y_k \rangle \right| \le C(\delta) \sum_{i=1}^{K} \|i^q u_i\|_s^2 \sum_{j=1}^{K} \|j^q w_j\|_s^2 + \delta \sum_{k=1}^{K} \|k^q y_k\|_0^2, \tag{3.90}$$

$$\left| \sum_{k=1}^{K} k^{2q} \langle \partial^\alpha (uf)_k, g_k \rangle \right| \le C(\delta) \sum_{i=1}^{K} \|i^q u_i\|_s^2 \sum_{j=1}^{K} \|j^q f_j\|_s^2 + \delta \sum_{k=1}^{K} \|k^q g_k\|_0^2, \tag{3.91}$$

where the constants are independent of K, and δ is any positive constant.

Proof. We focus on the proof of the first inequality, and the second one is similar (just use (3.38) instead of (3.37)). Note (by (3.37))

$$k^{2q} \|S_{ijk} \partial^\alpha (u_i w_j)\|_0 \le C k^{2q} |S_{ijk}| \|u_i\|_s \|w_j\|_s = C \frac{k^{2q}}{i^q j^q} |S_{ijk}| \cdot \|i^q u_i\|_s \cdot \|j^q w_j\|_s. \tag{3.92}$$

First we consider the case $i \ge j$. Then since

$$i^q j^q \ge \frac{1}{C} (\frac{k+1}{2})^q |S_{ijk}| j^{q-p}, \tag{3.93}$$

by (3.87) and (3.89), we conclude that

$$\frac{k^{2q}}{i^q j^q} |S_{ijk}| \le C k^q j^{p-q}. \tag{3.94}$$

Thus if we write the $(uw)_k$ on the LHS of (3.90) as a summation in i, j by (3.30), the $i \geq j$ terms can be estimated by

$$\begin{aligned}
&\left| \sum_{k=1}^{K} k^{2q} \sum_{i,j=1;\, i\geq j}^{K} \chi_{ijk} S_{ijk} \langle \partial^\alpha (u_i w_j), y_k \rangle \right| \\
\leq & \sum_{i,j,k=1;\, i\geq j}^{K} k^{2q} \| S_{ijk} \partial^\alpha (u_i w_j) \|_0 \cdot \|y_k\|_0 \cdot \chi_{ijk} \\
\leq & C \sum_{i,j,k=1;\, i\geq j}^{K} j^{p-q} \cdot \|i^q u_i\|_s \cdot \|j^q w_j\|_s \cdot \|k^q y_k\|_0 \cdot \chi_{ijk} \\
\leq & C \sum_{i,j,k=1}^{K} j^{p-q} \cdot \|i^q u_i\|_s \cdot \|j^q w_j\|_s \cdot \|k^q y_k\|_0 \cdot \chi_{ijk} \\
\leq & C(\delta) \sum_{i,j,k=1}^{K} j^{p-q} \cdot \|i^q u_i\|_s^2 \cdot \|j^q w_j\|_s^2 \cdot \chi_{ijk} + \delta \sum_{i,j,k=1}^{K} j^{p-q} \|k^q u_k\|_0^2 \cdot \chi_{ijk} \\
= & C(\delta) I + \delta I I,
\end{aligned} \tag{3.95}$$

where the second inequality uses (3.94), and χ_{ijk} is the indicator function of the set of indexes (i, j, k) for which $S_{ijk} \neq 0$.

Now we claim that

$$I \leq 2 \sum_{i=1}^{K} \|i^q u_i\|_s^2 \cdot \sum_{j=1}^{K} \|j^q w_j\|_s^2. \tag{3.96}$$

In fact, fix i, then one can write

$$I = \sum_{i=1}^{K} \|i^q u_i\|_s^2 I_i, \quad I_i = \sum_{j,k=1}^{K} j^{p-q} \cdot \|j^q w_j\|_s^2 \chi_{ijk}. \tag{3.97}$$

Notice that $\chi_{ijk} = 1$ implies that $i - j + 1 \leq k \leq i + j - 1$, by (3.89). Thus in the last summation, there is at most $2j$ terms corresponding to a fixed j. Thus

$$I_i \leq 2 \sum_{j=1}^{K} j^{p-q+1} \|j^q w_j\|_s^2 \leq 2 \sum_{j=1}^{K} \|j^q w_j\|_s^2, \tag{3.98}$$

if $q \geq p + 1$. This proves (3.96).

II is controlled by

$$II \leq 2 \sum_{j=1}^{K} j^{p-q+1} \sum_{k=1}^{K} \|k^q y_k\|_0^2, \tag{3.99}$$

since for each fixed (j,k) there is at most $2j$ choices for i. Thus if $q > p+2$, one has

$$II \le C\sum_{k=1}^{K} \|k^q y_k\|_0^2, \quad C = 2\sum_{j=1}^{\infty} j^{p-q+1} \le 2(1+(p-q+2)^{-1}). \tag{3.100}$$

Thus we conclude that the $i \ge j$ terms can be controlled by the RHS of (3.90) (with δ replaced by $C\delta$).

For the terms of the LHS of (3.90) with $i \le j$, we exchange the indexes i and j, and get the LHS of (3.95) with u and w exchanged. Thus one proceeds as before and get the same conclusion, since the RHS of (3.90) is invariant if u and w are exchanged. □

Remark 3.11. *The weight k^q appeared in the above lemma is essential. Suppose one uses a summation $\sum_{k=1}^{K}\langle\partial^\alpha(uw)_k, y_k\rangle$, then one ends up with the estimate*

$$\begin{aligned}\left|\sum_{k=1}^{K}\langle\partial^\alpha(uw)_k, y_k\rangle\right| &= \left|\sum_{i,j,k=1}^{K} S_{ijk}\langle\partial^\alpha(u_i w_j), y_k\rangle\right| \\ &\le \sum_{i,j,k=1}^{K} \min(i,j,k)^p [C(\delta)\|u_i\|_s^2\|w_j\|_s^2 + \delta\|y_k\|_0^2] \\ &\le C(\delta)C_1(K)\sum_{i=1}^{K}\|u_i\|_s^2\sum_{j=1}^{K}\|w_j\|_s^2 + \delta C_2(K)\sum_{k=1}^{K}\|y_k\|_0^2,\end{aligned} \tag{3.101}$$

where $C_1(K) = \sum_{k=1}^{K} k^p = O(K^{p+1})$, $\quad C_2(K) = K\sum_{i=1}^{K} i^p = O(K^{p+2})$. Thus in this way one gets an estimate with the coefficient depending on K. If one uses this estimate to prove an analog of Theorem 3.2.3, then one will get a constant c_2 depending on K.

In view of Proposition 3.1, c_2 being independent of K implies that the conclusion of Theorem 3.2.3 holds if the initial data satisfies a smoothness condition independent of K. If c_2 depends on K, then the initial data needs to satisfy a K-dependent condition to make the conclusion of Theorem 3.2.3 true. This is not good, since it is desirable that the gPC-sG method is stable for a class of initial data, for all K.

Remark 3.12. *For gPC basis with respect to a probability measure supported on $\mathbb{R}$, for example, the Gaussian distribution, numerical evidence suggests that (3.87) may fail. In this case one might require a weaker condition, for example, (3.87) with k^p replaced by λ^k for*

some $\lambda > 1$, and prove results similar to Lemma 3.10 with different weights. This is out of the scope of this paper.

Due to the similarity of Lemma 3.6 and Lemma 3.10, it is straightforward to modify the proof of Theorem 3.2.1 into a proof of Theorem 3.2.3:

Proof of Theorem 3.2.3. We take ∂^α on the first and third equations of (3.28), and do L^2 estimates on them as well as on the fourth equation, and then sum over k and α with the k-th equation multiplied by k^{2q}. Then we get

$$\frac{1}{2}\partial_t E^K + G^K + B^K = 0, \tag{3.102}$$

where

$$\begin{aligned}
&E^K(t) = \sum_{k=1}^{K}(\|k^q u_k\|_s^2 + \|k^q f_k\|_s^2 + |k^q \bar{u}_k|^2),\\
&G^K = G_1^K + G_2^K = \sum_{k=1}^{K}(\|\nabla_x k^q u_k\|_s^2 + 2|k^q \bar{u}_k|^2) + \sum_{k=1}^{K}\left\|k^q\left(u_k\sqrt{\mu} - \frac{1}{\epsilon}\nabla_v f_k - \frac{1}{\epsilon}\frac{v}{2}f_k\right)\right\|_s^2,\\
&B^K = B_1^K + B_2^K + B_3^K = \sum_{|\alpha|\le s} B_{1,\alpha}^K + \sum_{|\alpha|\le s} B_{2,\alpha}^K + B_3^K,
\end{aligned} \tag{3.103}$$

with

$$\begin{aligned}
B_{1,\alpha}^K &= \sum_{k=1}^{K} k^{2q}\langle \partial^\alpha (u\cdot\nabla_x u)_k, \partial^\alpha u_k\rangle,\\
B_{2,\alpha}^K &= \sum_{k=1}^{K} k^{2q}\left\langle \partial^\alpha (uf)_k, \partial^\alpha\left[u_k\sqrt{\mu} - \frac{1}{\epsilon}\nabla_v f_k - \frac{1}{\epsilon}\frac{v}{2}f_k\right]\right\rangle,\\
B_3^K &= \frac{1}{|\mathbb{T}^3|}\sum_{k=1}^{K} k^{2q}\langle (uf)_k, \bar{u}_k\sqrt{\mu}\rangle.
\end{aligned} \tag{3.104}$$

Now apply Lemma 3.10 to get

$$
\begin{aligned}
|B^K_{1,\alpha}| &\le C(\delta)\sum_{k=1}^K \|k^q u_k\|^2_{s+1}\sum_{k=1}^K \|k^q u_k\|^2_s + \delta\sum_{k=1}^K \|k^q u_k\|^2_{s+1} \le C(\delta)E^K G^K_1 + \delta G^K_1,\\
|B^K_{2,\alpha}| &\le C(\delta)\sum_{k=1}^K \|k^q u_k\|^2_s\sum_{k=1}^K \|k^q f_k\|^2_s + \delta G^K_2 \le C(\delta)E^K G^K_1 + \delta G^K_2,\\
|B^K_3| &\le C(\delta)\sum_{k=1}^K \|k^q u_k\|^2_s\sum_{k=1}^K \|k^q f_k\|^2_s + \sum_{k=1}^K \delta|k^q \bar{u}_k|^2 \le C(\delta)E^K G^K_1 + \delta G^K_1.
\end{aligned}
\tag{3.105}
$$

And then one concludes

$$
\frac{1}{2}\partial_t E^K \le -(1 - C(\delta)E^K - C\delta)G^K. \tag{3.106}
$$

Assuming $\delta = \frac{1}{4C}$ where C is the constant in (3.106), and $c_2(s,r) = \frac{1}{4C(\delta)}$, then by the same argument as in the proof of Theorem 3.2.1, if $E^K(0) \le c_2(s,r)$, then one has

$$
\partial_t E^K + G^K \le 0, \tag{3.107}
$$

and E^K is non-increasing.

□

Proof of Proposition 3.1. Note that $(u_0)_k$ is the k-th gPC coefficient of the initial data u_0, and thus satisfies the spectral accuracy estimate

$$
|(u_0)_k(x)| \le C\frac{\|u_0(x,\cdot)\|_{H^r_z}}{k^r}, \tag{3.108}
$$

at each fixed x. By integrating (3.108) in x and replacing u by $\partial^\alpha u$ and summing over α, one gets

$$
\|k^q(u_k)_0\|_s \le Ck^{q-r}\|u_0\|_{s,r}. \tag{3.109}
$$

Thus if $r > q + \frac{1}{2}$, one has

$$
\sum_{k=1}^K \|k^q(u_k)_0\|^2_s \le C\|u_0\|^2_{s,r}. \tag{3.110}
$$

Similar estimate holds for f and $\bar{u}$. Thus one has

$$
E^K_{s,q}(0) \le C\|E_{s,r}(0)\|_{L^1_z}, \tag{3.111}
$$

and the proof is finished. □

3.5.2 Accuracy analysis: proof of Theorem 3.2.4

Recall the reconstructed gPC solution

$$u^K(x,z) = \sum_{k=1}^{K} u_k(x)\phi_k(z). \tag{3.112}$$

Then at a fixed x point one has

$$\|u^K(x,\cdot)\|_{L_z^2}^2 = \sum_{k=1}^{K} |u_k(x)|^2 \le \sum_{k=1}^{K} |k^q u_k(x)|^2. \tag{3.113}$$

Thus

$$\|u^K\|_0^2 \le E_{0,q}^K. \tag{3.114}$$

for any $q \ge 0$.

Furthermore, with the assumption (3.31), one has the estimate

$$\|u^K(x)\|_{L_z^\infty}^2 \le C\left(\sum_{k=1}^{K} |u_k(x)|k^p\right)^2 \le C\left(\sum_{k=1}^{K} |k^q u_k(x)|^2\right)\left(\sum_{k=1}^{K} k^{2(p-q)}\right) \le C\left(\sum_{k=1}^{K} |k^q u_k(x)|^2\right), \tag{3.115}$$

since $q > p+2$. Thus

$$\|u^K\|_{L_z^\infty(L_x^2)}^2 \le \|u^K\|_{L_x^2(L_z^\infty)}^2 \le C E_{0,q}^K. \tag{3.116}$$

Similar estimates hold for f and $\bar{u}$ and their x derivatives.

Proof of Theorem 3.2.4. The gPC coefficients of the mean fluid velocity satisfies

$$\partial_t \bar{u}_k + 2\bar{u}_k + C\int\int \sqrt{\mu}(uf)_k \,\mathrm{d}v\,\mathrm{d}x = 0. \tag{3.117}$$

Denote the projection operator onto the span of $\{\phi_k\}_{k=1}^K$ by P_K. Multiplying (3.28) and (3.117) by $\phi_k(z)$ and summing in k, one gets the equations for (u^K, f^K)

$$\begin{aligned}
&\partial_t u^K + P_K(u^K\cdot\nabla_x u^K) + \nabla_x p^K - \Delta_x u^K + u^K + \int \sqrt{\mu} P_K(u^K f^K)\,\mathrm{d}v - \frac{1}{\epsilon}\int v\sqrt{\mu} f^K \,\mathrm{d}v = 0,\\
&\nabla_x \cdot u^K = 0,\\
&\partial_t f^K + \frac{1}{\epsilon} v\cdot\nabla_x f^K + \frac{1}{\epsilon}(\nabla_v - \frac{v}{2})\cdot P_K(u^K f^K) - \frac{1}{\epsilon} u^K\cdot v\sqrt{\mu} = \frac{1}{\epsilon^2}(\frac{-|v|^2}{4} + \frac{3}{2} + \Delta_v) f^K,\\
&\partial_t \bar{u}^K + 2\bar{u}^K + \frac{1}{|\mathbb{T}^3|}\int\int \sqrt{\mu} P_K(u^K f^K)\,\mathrm{d}v\,\mathrm{d}x = 0.
\end{aligned} \tag{3.118}$$

Then subtracting from (3.7) and (3.12), one gets

$$\begin{aligned}
&\partial_t u^e + [(I-P_K)(u\cdot\nabla_x u) + P_K(u^e\cdot\nabla_x u + u^K\cdot\nabla_x u^e)] + \nabla_x p^e - \Delta_x u^e + u^e \\
&\qquad + \int \sqrt{\mu}[(I-P_K)(uf) + P_K(u^e f + u^K f^e)]\,\mathrm{d}v - \frac{1}{\epsilon}\int v\sqrt{\mu} f^e\,\mathrm{d}v = 0, \\
&\nabla_x\cdot u^e = 0, \\
&\partial_t f^e + \frac{1}{\epsilon} v\cdot\nabla_x f^e + \frac{1}{\epsilon}(\nabla_v - \frac{v}{2})\cdot[(I-P_K)(uf) + P_K(u^e f + u^K f^e)] \\
&\qquad - \frac{1}{\epsilon} u^e\cdot v\sqrt{\mu} = \frac{1}{\epsilon^2}(\frac{-|v|^2}{4} + \frac{3}{2} + \Delta_v) f^e, \\
&\partial_t \bar{u}^e + 2\bar{u}^e + \frac{1}{|\mathbb{T}^3|}\int\int \sqrt{\mu}[(I-P_K)(uf) + P_K(u^e f + u^K f^e)]\,\mathrm{d}v\,\mathrm{d}x = 0,
\end{aligned} \tag{3.119}$$

where (u^e, f^e) is the approximation error

$$u^e = u - u^K, \quad f^e = f - f^K. \tag{3.120}$$

Notice that (3.119) is linear in (u^e, f^e).

Now take ∂^α on the first and third equations of (3.119), and do L^2 estimates on the first, third, fourth equations in (x,z), (x,v,z), z, respectively. First notice that P_K commutes with x-derivatives, and has operator norm 1 on L^2_z. Thus one has

$$|\langle\langle \partial^\alpha P_K(u^e\cdot\nabla_x u + u^K\cdot\nabla_x u^e), \partial^\alpha u^e\rangle\rangle| \le C(\|u\|_{W^{s+1,\infty}} + \|u^K\|_{W^{s,\infty}})\|u^e\|^2_{s+1}, \tag{3.121}$$

where the W norms mean the Sobolev norms with power index ∞, as defined in (3.25), and the sub-index $r=0$ is omitted. By estimating the terms $P_K(u^e f + u^K f^e)$ in the same manner, i.e.,

$$|\langle\langle \partial^\alpha P_K(u^e f + u^K f^e), \partial^\alpha u^e\rangle\rangle| \le C(\|f\|_{W^{s,\infty}} + \|u^K\|_{W^{s,\infty}})(\|u^e\|^2_s + \|f^e\|^2_s), \tag{3.122}$$

one gets the energy estimate

$$\frac{1}{2}\partial_t E^e \le -(\frac{2}{3} - CH)G^e + CS, \tag{3.123}$$

where

$$\begin{aligned} E^e =& \|u^e\|_s^2 + \|f^e\|_s^2 + \|\bar{u}^e\|^2, \\ G^e =& \|\nabla_x u^e\|_s^2 + 2\|\bar{u}^e\|^2 + \left\| u^e\sqrt{\mu} - \frac{1}{\epsilon}\nabla_v f^e - \frac{1}{\epsilon}\frac{v}{2} f^e \right\|_s^2, \\ S =& (\|(I-P_K)(u\cdot\nabla_x u)\|_s^2 + \|(I-P_K)(uf)\|_s^2), \\ H =& \|u\|_{W^{s+1,\infty}} + \|u^K\|_{W^{s,\infty}} + \|f\|_{W^{s,\infty}}. \end{aligned} \tag{3.124}$$

First notice that by Sobolev embedding,

$$\|u\|_{W^{s+1,\infty}} \le C\|u\|_{L_z^\infty(H_x^{s+3})}, \quad \|f\|_{W^{s,\infty}} \le C\|f\|_{L_z^\infty(H_x^{s+2}(L_v^2))}, \tag{3.125}$$

and by (3.116)

$$\|u^K\|_{W^{s,\infty}}^2 \le C E_{s+2,q}^K. \tag{3.126}$$

Thus H can be controlled by

$$H \le C(\|E_{s+3,0}\|_{L_z^\infty} + E_{s+2,q}^K)^{1/2}. \tag{3.127}$$

In view of Lemma 3.1, for $r > p + \frac{5}{2}$ one has

$$H \le C\|E_{s+3,r}\|_{L_z^\infty}^{1/2}, \tag{3.128}$$

which implies that

$$CH \le \frac{1}{6}, \tag{3.129}$$

in (3.123) for all time if $\|E_{s+3,r}(0)\|_{L_z^\infty} \le c_1''(s,r) \le \min\{\frac{1}{4C}, c_1(s,r), c_2(s,q)\}$, in view of Theorem 3.2.1 and Theorem 3.2.3.

To estimate the source term S, notice that at each fixed x, v,

$$\|(I-P_K)\partial^\alpha(uf)(x,v)\|_{L_z^2} \le C\frac{\|\partial^\alpha(uf)(x,v)\|_{H_z^r}}{K^r}. \tag{3.130}$$

Integrating in x, v and summing over α,

$$\|(I-P_K)(uf)\|_s \le C\frac{\|uf\|_{s,r}}{K^r}. \tag{3.131}$$

Notice that at each z,

$$|uf|_{s,r} \leq \max_{|\alpha|\leq s,\,|\gamma|\leq r} \|\partial^\alpha u^\gamma\|_{L^\infty_x} |f|_{s,r} \leq C|u|_{s+2,r}|f|_{s,r}. \tag{3.132}$$

Thus

$$\|uf\|_{s,r} \leq \||uf|_{s,r}\|_{L^\infty_z} \leq C\||u|_{s+2,r}\|_{L^\infty_z}\||f|_{s,r}\|_{L^\infty_z} \leq C\|E_{s+2,r}\|^{1/2}_{L^\infty_z}\|E_{s,r}\|^{1/2}_{L^\infty_z}. \tag{3.133}$$

Then by Theorem 3.2.1 and Theorem 3.2.2 (suppress the dependence on C^h), taking $c_1'' \leq c_1'(s,r)$,

$$E_{s+2,r}(t) \leq C, \quad E_{s,r}(t) \leq Ce^{-\lambda t}. \tag{3.134}$$

Thus we finally get

$$\|(I-P_K)(uf)\|_s \leq \frac{Ce^{-\frac{\lambda}{2}t}}{K^r}. \tag{3.135}$$

The term $\|(I-P_K)(u\cdot\nabla_x u)\|_s$ can be estimated similarly, by using $|u\cdot\nabla_x u|_{s,r} \leq C|u|_{s+3,r}|u|_{s,r}$, and we get

$$S \leq \frac{Ce^{-\lambda t}}{K^{2r}}. \tag{3.136}$$

In conclusion, we have the estimate

$$\partial_t E^e + G^e \leq \frac{C}{K^{2r}} e^{-\lambda t}. \tag{3.137}$$

Finally, combining (3.123), (3.129) and (3.136), noticing that $\int_0^\infty e^{-2\lambda t}\,\mathrm{d}t$ converges, one concludes that $E^e \leq \frac{C}{K^{2r}}$ uniformly in time and ϵ.

□

3.5.3 Hypocoercivity estimates for the error: proof of Theorem 3.2.5

Proof of Theorem 3.2.5. In order to get a hypocoercivity estimate for (u^e, f^e), we write the equation of $\partial^\alpha f^e$ as

$$\begin{aligned}\partial_t\partial^\alpha f^e + \frac{1}{\epsilon}\mathcal{P}\partial^\alpha f^e + \frac{1}{\epsilon^2}\mathcal{K}^*\mathcal{K}\partial^\alpha f^e &= \frac{1}{\epsilon}\partial^\alpha u^e\cdot v\sqrt{\mu} \\ &+ \frac{1}{\epsilon}[(I-P_K)\partial^\alpha(u\cdot\mathcal{K}^*f) + P_K\partial^\alpha(u^e\cdot\mathcal{K}^*f) + P_K\partial^\alpha(u^K\cdot\mathcal{K}^*f^e)].\end{aligned} \tag{3.138}$$

and then do energy estimate in (x, v, z). The linear terms can be handled in the same way as Lemma 3.8. The first nonlinear term is estimated by

$$\left|\frac{1}{\epsilon}((\,(I-P_K)\partial^\alpha(u\cdot\mathcal{K}^*f),\partial^\alpha f^e))\right| \le \frac{C}{K^r}\frac{1}{\epsilon^2}\|u\|_{L^\infty_z(H^{s+3,r})}([[f,f]]_{s,r}+[[f^e,f^e]]_s). \tag{3.139}$$

In fact, by modifying the proof of Lemma 3.8, one can get an expression like (3.63):

$$((\,(I-P_K)\partial^\alpha(u\cdot\mathcal{K}^*f),\partial^\alpha f^e)) = 2\langle\langle(I-P_K)\partial^\alpha(u\mathcal{K}f),\mathcal{K}^2\partial^\alpha f^e\rangle\rangle + \text{similar terms}. \tag{3.140}$$

The first term in (3.140) is estimated by

$$\begin{aligned}
&|\langle\langle(I-P_K)\partial^\alpha(u\mathcal{K}f),\mathcal{K}^2\partial^\alpha f^e\rangle\rangle| \\
\le&\|(I-P_K)\partial^\alpha(u\mathcal{K}f)\|_0\|\mathcal{K}^2\partial^\alpha f^e\|_0 \le \frac{C}{K^r}\|\partial^\alpha(u\cdot\mathcal{K}f)\|_{0,r}\|\mathcal{K}^2 f^e\|_s \\
\le&\frac{C}{K^r}\max_{|\gamma|\le r,|\beta|\le s}\|\partial^\beta u^\gamma\|_{L^\infty}\|\mathcal{K}f\|_{s,r}\|\mathcal{K}^2 f^e\|_s \\
\le&\frac{C}{K^r}\|u\|_{L^\infty_z(H^{s+3,r})}(\|\mathcal{K}f\|^2_{s,r}+\|\mathcal{K}^2 f^e\|^2_s),
\end{aligned} \tag{3.141}$$

and other terms in (3.140) can be estimated similarly.

The second nonlinear term in (3.138) is estimated by Lemma 3.8 as follows:

$$\begin{aligned}
\left|\frac{1}{\epsilon}((P_K\partial^\alpha(u^e\cdot\mathcal{K}^*f),\partial^\alpha f^e))\right| &\le \left|\frac{1}{\epsilon}((\partial^\alpha(u^e\cdot\mathcal{K}^*f),\partial^\alpha f^e))\right| \\
&\le C\frac{1}{\epsilon^2}\max_{|\beta|\le s}\|\partial^\beta u^e\|_{L^\infty}(C(\delta)[[f,f]]_s+\delta[[f^e,f^e]]_s).
\end{aligned} \tag{3.142}$$

The third nonlinear term is estimated by Lemma 3.8 as follows:

$$\left|\frac{1}{\epsilon}((P_K\partial^\alpha(u^K\cdot\mathcal{K}^*f^e),\partial^\alpha f^e))\right| \le \left|\frac{1}{\epsilon}((\partial^\alpha(u^K\cdot\mathcal{K}^*f^e),\partial^\alpha f^e))\right| \le C\frac{1}{\epsilon^2}\|u^K\|_{L^\infty_z(H^{s+3})}[[f^e,f^e]]_s. \tag{3.143}$$

Now by assumption, $\|E_{s+3,r}(t)\|_{L^\infty_z}$ is small enough at $t=0$ (which implies that they are small enough for all time, by Theorem 3.2.1). Similar result holds for $E^K_{s+3,q}\le C\|E_{s+3,r}\|_{L^\infty_z}$, by Theorem 3.2.3. As a result, $\|u\|_{L^\infty_z(H^{s+3,r})}$ and $\|u^K\|_{L^\infty_z(H^{s+3})}$ are small enough, see (3.116) for the latter.

To bound the term $\max_{|\beta|\le s}\|\partial^\beta u^e\|_{L^\infty}$ appeared in (3.142), we estimate

$$\|u^e\|_{L^\infty_z} = \left\|\sum_{k=1}^K(u^e)_k\phi_k(z)\right\|_{L^\infty_z} \le C\left(\sum_{k=1}^K|(u^e)_k|^2\right)^{1/2}\left(\sum_{k=1}^K k^{2p}\right)^{1/2} \le C\|u^e\|_{L^2_z}K^{p+1/2}, \tag{3.144}$$

at any fixed x. Then taking L^∞ in x we obtain

$$\|u^e\|_{L^\infty} \le CK^{p+1/2}\|u^e\|_{L^\infty_x(L^2_z)} \le CK^{p+1/2}\|u^e\|_{L^2_z(L^\infty_x)} \le CK^{p+1/2}\|u^e\|_{L^2_z(H^2_x)}. \tag{3.145}$$

By Theorem 3.2.4, $\|u^e\|_{L^2_z(H^{s+3})}$ is bounded by $\frac{C}{K^r}$. Doing the same estimate for the x-derivatives of u^e, we obtain

$$\max_{|\beta|\le s}\|\partial^\beta u^e\|_{L^\infty} \le \frac{C}{K^{r-p-1/2}}. \tag{3.146}$$

Then by choosing δ in (3.142) small enough, all the $[[f^e, f^e]]_s$ terms from the nonlinear terms can be absorbed by the corresponding term from the linear terms, and then one concludes the estimate

$$\partial_t((f^e, f^e))_s + \lambda_1^e\frac{1}{\epsilon^2}[[f^e, f^e]]_s \le C(\lambda_1^e)(\|u^e\|_s^2 + \|\nabla_x u^e\|_s^2 + \frac{1}{\epsilon^2}\|\mathcal{K}f^e\|_s^2) + \frac{C}{K^{r-p-1/2}}\frac{1}{\epsilon^2}[[f, f]]_{s,r}. \tag{3.147}$$

Finally, similar to the proof of Theorem 3.2.2, by taking a suitable linear combination of (3.147)(3.137) and (3.77) integrated in z (where the appearance of (3.77) is to control the term $[[f, f]]_{s,r}$ in (3.147)), we get

$$\partial_t\tilde{E}^e + \frac{1}{2}\tilde{G}^e \le \lambda_4^e\frac{C}{K^{r-p-1/2}}\frac{1}{\epsilon^2}[[f, f]]_{s,r} + \frac{C}{K^r}e^{-\lambda t}, \tag{3.148}$$

where

$$\tilde{E}^e = E^e + \lambda_4^e((f^e, f^e))_s + \frac{1}{K^{r-p-1/2}}\lambda_5^e\|\tilde{E}\|_{L^1_z}, \tag{3.149}$$

and

$$\tilde{G}^e = G^e + \lambda_4^e\lambda_1^e\frac{1}{\epsilon^2}[[f^e, f^e]]_s + \frac{1}{2K^{r-p-1/2}}\lambda_5^e\|\tilde{G}\|_{L^1_z}. \tag{3.150}$$

The choice of λ_4^e is in the same way as the choice of λ_4. To choose λ_5^e, one wants the $\tilde{G}$ term to control the first RHS term in (3.148), and thus choose

$$\lambda_5^e = 4\frac{C\lambda_4^e}{\lambda_4\lambda_1}, \tag{3.151}$$

where the C is the first constant in (3.148). Then

$$\partial_t\tilde{E}^e + \frac{1}{4}\tilde{G}^e \le \frac{C}{K^r}e^{-\lambda t}. \tag{3.152}$$

Then since $\tilde{E}^e \leq C\tilde{G}^e$ (which can be proved similarly as the proof of $\tilde{E} \leq C\tilde{G}$, see (3.80)), and $\tilde{E}^e(0) \leq \frac{C}{K^{r-p-1/2}}$, one concludes that

$$\tilde{E}^e \leq \frac{C}{K^{r-p-1/2}} e^{-\lambda^e t}, \tag{3.153}$$

where $\lambda^e = \min\{\lambda, \frac{1}{4C}\} - \delta$ for some $\delta > 0$ small enough, in view of the lemma below. □

Lemma 3.13. *Let $\Phi = \Phi(t)$ satisfy*

$$\frac{\mathrm{d}\Phi}{\mathrm{d}t} + a_1\Phi \leq a_2 e^{-a_3 t}. \tag{3.154}$$

Then

$$\Phi(t) \leq e^{-at}(\Phi(0) + a_2 C(\delta)), \tag{3.155}$$

with $a = \min\{a_1, a_3\} - \delta$, δ being any positive constant.

Proof.

$$\frac{\mathrm{d}}{\mathrm{d}t}(e^{a_1 t}\Phi) \leq a_2 e^{(a_1 - a_3)t}, \tag{3.156}$$

$$e^{a_1 t}\Phi \leq \Phi(0) + \int_0^t a_2 e^{(a_1 - a_3)s}\,\mathrm{d}s, \tag{3.157}$$

$$\Phi(t) \leq e^{-a_1 t}\Phi(0) + a_2 \frac{e^{-a_3 t} - e^{-a_1 t}}{a_1 - a_3} = e^{-a_1 t}\Phi(0) + a_2 t e^{-\xi t}, \tag{3.158}$$

for some ξ between a_1 and a_3, by the mean value theorem. Then the conclusion follows since

$$te^{-\xi t} \leq e^{-at}(te^{-\delta t}) \leq C(\delta)e^{-at}, \tag{3.159}$$

where $C(\delta) = (\delta e)^{-1}$. □

Chapter 4

A sparse wavelet based sG method for the Boltzmann equation with multi-dimensional random inputs

In this chapter we go through the author's work with J. Hu and S. Jin on the sparse wavelet based sG method for the Boltzmann equation (1.2) with multi-dimensional random inputs. In order to emphasize the vector notations, in this chapter we will use bold letters to denote vectors.

4.1 Introduction

The Boltzmann equation plays an essential role in kinetic theory [18]. It describes the time evolution of the density distribution of dilute gases, where fluid dynamics equations, such as the Euler equations and the Navier-Stokes equations, fail to provide reliable information. It is an indispensable tool in fields concerning non-equilibrium statistical mechanics, such as rarefied gas dynamics and astronautical engineering.

As introduced in Section 1.2, it is important to consider the Boltzmann equation with random inputs. Hu and Jin [50] gave a first numerical method to solve the Boltzmann equation with uncertainty by a gPC-sG method. By a singular value decomposition on a set of basis related coefficients, together with the fast spectral method for the Boltzmann collision operator proposed by [81], the computational cost of the collision operator is decreased dramatically. However, their work focuses on low dimensional random spaces, and a direct extension of their method to multi-dimensional random spaces will suffer from the curse of dimensionality, which means K, the total number of basis functions, will grow like $K =$

$\binom{K_1+d}{K_1}$, where K_1 is the number of basis in one dimension, and d is the dimension of the random space. This cost is not affordable if both K_1 and d are large. Monte-Carlo methods are feasible, but a halfth order convergence rate can be unsatisfactory in many applications. Therefore it is desirable to have an efficient and accurate method to solve the Boltzmann equation with multi-dimensional random inputs.

In this work, we adopt a sparse approach [15, 32] for the stochastic Galerkin method to circumvent the curse of dimensionality. The idea of sparse approaches traces back to Smolyak [96]. In recent years, sparse approaches have become a major way to break the curse of dimensionality in various contexts, for example in Galerkin finite element methods [15, 107, 92], finite difference methods [43, 44], high-dimensional stochastic differential equations [104, 84] and uncertainty quantification [89, 76]. The sparse approach we adopt was first proposed by Schwab et al. [90] for transport-dominated diffusion problems, and then applied to discontinuous Galerkin methods for elliptic equations by Wang et al. [101] and transport equations by Guo and Cheng [46]. Simply speaking, we start from a hierarchical basis in one dimension. To construct the sparse wavelet basis in multi-dimension, we take the tensor basis and discard those basis functions that are in deep levels in most dimensions. In this way only a small number of basis functions are kept, yet it can be proved that the accuracy is still as good as the corresponding tensor basis, if the function to approximate is smooth enough. With a hierarchical basis with N levels and piecewise polynomials of degree at most m, our method can achieve an accuracy of $O(N^{d-1}2^{-N(m+1)})$ with number of basis $K = O((m+1)^d 2^N N^{d-1})$ for d-dimensional random spaces. This accuracy is $O(K^{-(m+1)}(\log K)^{(m+2)(d-1)})$ in terms of K. It is algebraically accurate, but as d increases, the accuracy deteriorates very slowly. Furthermore, we discover a sparse structure of a set of basis related coefficients, S_{ijk}, which greatly reduces the cost of the expensive collision operator evaluation.

The rest of the chapter is organized as follows: in Section 4.2 we introduce the Boltzmann equation with uncertainty and the framework of stochastic Galerkin (sG) method; in Section 4.3 we introduce our sparse method with multi-wavelet functions; in Section 4.4 we give an

estimate of the sparsity of the coefficients S_{ijk}; in Section 4.5 we prove the random space regularity of the solution of the Boltzmann equation with uncertainty, as well as the accuracy of the sG method with sparse wavelet basis; in Section 4.6 we give some numerical results.

4.2 The Boltzmann equation with uncertainty

To make the notation consistent with [93], we rewrite the classical (deterministic) Boltzmann equation in its dimensionless form as

$$\partial_t f + \mathbf{v} \cdot \nabla_{\mathbf{x}} f = \frac{1}{\mathrm{Kn}} Q(f, f), \tag{4.1}$$

where $f = f(t, \mathbf{x}, \mathbf{v})$ is the density distribution function of a dilute gas at time $t \in \mathbb{R}^+$, position $\mathbf{x} \in \Omega \subset \mathbb{R}^{d_x}$, and with particle velocity $\mathbf{v} \in \mathbb{R}^{d_v}$. Kn is the Knudsen number, a dimensionless number defined as the ratio of the mean free path and a typical length scale, such as the size of the spatial domain. The collision operator $Q(f, f)$ is defined by (1.2) (where we used Q_B instead of the notation Q here). The basic properties of the Boltzmann equation, as well as its hydrodynamic limit as $\mathrm{Kn} \to 0$, are already mentioned in Section 1.1.

The initial condition of the Boltzmann equation is given by

$$f(0, \mathbf{x}, \mathbf{v}) = f^0(\mathbf{x}, \mathbf{v}), \tag{4.2}$$

and a boundary condition is needed if the spatial domain Ω is a proper subset of $\mathbb{R}^{d_x}$. We adopt the Maxwell boundary condition, which takes the form

$$f(t, \mathbf{x}, \mathbf{v}) = g(t, \mathbf{x}, \mathbf{v}), \quad \mathbf{x} \in \partial\Omega, \quad \mathbf{v} \cdot \mathbf{n} > 0, \tag{4.3}$$

with

$$\begin{aligned} g(t, \mathbf{x}, \mathbf{v}) =& (1 - \alpha) f(t, \mathbf{x}, \mathbf{v} - 2(\mathbf{v} \cdot \mathbf{n})\mathbf{n}) \\ &+ \frac{\alpha}{(2\pi)^{(d_v-1)/2} T_w(\mathbf{x})^{(d_v+1)/2}} e^{-\frac{|\mathbf{v}|^2}{2T_w(\mathbf{x})}} \int_{\mathbf{v}\cdot\mathbf{n}<0} f(t, \mathbf{x}, \mathbf{v}) |\mathbf{v} \cdot \mathbf{n}| \, \mathrm{d}\mathbf{v}, \end{aligned} \tag{4.4}$$

where T_w is the temperature of the wall, and $\mathbf{n}$ is the inner normal unit vector of the wall. The first term is the specular reflective part, and the second term is the diffusive part. α is

the accommodation coefficient. $\alpha = 1$ implies purely diffusive boundary, while $\alpha = 0$ implies purely reflective boundary. For simplicity we only consider the case where the wall is static.

As mentioned before, there are many sources of uncertainties in the Boltzmann equation, such as the initial data, boundary data, and collision kernel. To quantify these uncertainties we introduce the Boltzmann equation with uncertainty

$$\begin{cases} \partial_t f(t,\mathbf{x},\mathbf{v},\mathbf{z}) + \mathbf{v}\cdot\nabla_{\mathbf{x}} f(t,\mathbf{x},\mathbf{v},\mathbf{z}) = \dfrac{1}{\mathrm{Kn}} Q_{\mathbf{z}}(f,f), \quad t\in\mathbb{R}_+,\ \mathbf{x}\in\Omega\subset\mathbb{R}^{d_x},\ \mathbf{v}\in\mathbb{R}^{d_v},\ \mathbf{z}\in I_{\mathbf{z}}\subset\mathbb{R}^d, \\ f(0,\mathbf{x},\mathbf{v},\mathbf{z}) = f^0(\mathbf{x},\mathbf{v},\mathbf{z}), \quad \mathbf{x}\in\Omega,\ \mathbf{v}\in\mathbb{R}^{d_v},\ \mathbf{z}\in I_{\mathbf{z}}, \\ f(t,\mathbf{x},\mathbf{v},\mathbf{z}) = g(t,\mathbf{x},\mathbf{v},\mathbf{z}), \quad t\in\mathbb{R}_+,\ \mathbf{x}\in\partial\Omega,\ \mathbf{v}\in\mathbb{R}^{d_v},\ \mathbf{z}\in I_{\mathbf{z}}. \end{cases} \tag{4.5}$$

Here $\mathbf{z}\in I_{\mathbf{z}}$ is a d-dimensional random vector with probability distribution $\pi(\mathbf{z})$ characterizing the uncertainty in the system. We assume that the collision kernel has the form

$$B(\mathbf{v},\mathbf{v}_*,\sigma,\mathbf{z}) = b(\mathbf{z})B_0(\mathbf{v},\mathbf{v}_*,\sigma),$$

which means that $Q_{\mathbf{z}}$ can be written as

$$Q_{\mathbf{z}}(f,f) = b(\mathbf{z})Q(f,f).$$

The Maxwell boundary data $g(t,\mathbf{x},\mathbf{v},\mathbf{z})$ is given by

$$\begin{aligned} g(t,\mathbf{x},\mathbf{v},\mathbf{z}) =& (1-\alpha(\mathbf{z}))f(t,\mathbf{x},\mathbf{v}-2(\mathbf{v}\cdot\mathbf{n})\mathbf{n},\mathbf{z}) \\ &+ \frac{\alpha(\mathbf{z})}{(2\pi)^{(d_v-1)/2}T_w(\mathbf{x},\mathbf{z})^{(d_v+1)/2}} e^{-\frac{|\mathbf{v}|^2}{2T_w(\mathbf{x},\mathbf{z})}} \int_{\mathbf{v}\cdot\mathbf{n}<0} f(t,\mathbf{x},\mathbf{v},\mathbf{z})|\mathbf{v}\cdot\mathbf{n}|\,\mathrm{d}\mathbf{v}. \end{aligned} \tag{4.6}$$

To solve the stochastic system (4.5), Hu and Jin [50] proposed a stochastic Galerkin (sG) method. As introduced in Section 1.2, the sG method for the Boltzmann equation reads

$$\partial_t f_k(t,\mathbf{x},\mathbf{v}) + \mathbf{v}\cdot\nabla_{\mathbf{x}} f_k(t,\mathbf{x},\mathbf{v}) = Q_k(f^K,f^K), \tag{4.7}$$

$$f_k(0,\mathbf{x},\mathbf{v}) = f_k^0(\mathbf{x},\mathbf{v}), \tag{4.8}$$

$$Q_k(f^K,f^K) = \sum_{i,j=1}^{K} S_{ijk}Q(f_i,f_j), \tag{4.9}$$

where S_{ijk} is defined by (1.50). The boundary condition is given by

$$\begin{aligned} g_k = &\sum_{j=1}^{K} \int_{I_{\mathbf{z}}} (1-\alpha(\mathbf{z}))\Phi_k(\mathbf{z})\Phi_j(\mathbf{z})\pi(\mathbf{z})\,\mathrm{d}\mathbf{z} f_j(t,\mathbf{x},\mathbf{v}-2(\mathbf{v}\cdot\mathbf{n})\mathbf{n}) \\ &+ \sum_{j=1}^{K} D_{kj}(\mathbf{x},\mathbf{v}) \int_{\mathbf{v}\cdot\mathbf{n}<0} f_j(t,\mathbf{x},\mathbf{v},\mathbf{z})|\mathbf{v}\cdot\mathbf{n}|\,\mathrm{d}\mathbf{v}, \end{aligned} \tag{4.10}$$

where

$$D_{kj}(\mathbf{x},\mathbf{v}) = \int_{I_{\mathbf{z}}} \frac{\alpha(\mathbf{z})}{(2\pi)^{(d_v-1)/2} T_w(\mathbf{x},\mathbf{z})^{(d_v+1)/2}} e^{-\frac{|\mathbf{v}|^2}{2T_w(\mathbf{x},\mathbf{z})}} \Phi_k(\mathbf{z})\Phi_j(\mathbf{z})\pi(\mathbf{z})\,\mathrm{d}\mathbf{z} \tag{4.11}$$

is a matrix that is time independent hence can be pre-computed.

This gPC-sG method works well for low dimensional random inputs, but for high dimensional ones, it might require a very large number of basis functions (K large) to approximate f to a given accuracy. If one takes K_1 basis functions in each dimension of a d-dimensional random space, then a direct extension of the gPC-sG method will require $K = \binom{K_1+d}{K_1}$ basis functions, which is prohibitively expensive if both K_1 and d are large. Furthermore, since the computation of Q_k typically requires $O(K^2)$ times evaluation of the deterministic collision operator, one has to choose a relatively small K in order to afford the computation. Also, [50] uses the singular value decomposition of a size K matrix as pre-computation for the collision operator, which reduces the computational cost by one order of magnitude, but this pre-computation can be prohibitively expensive if K is large. In the following sections we propose a stochastic Galerkin method with sparse wavelet basis functions, which requires much fewer basis functions for multi-dimensional random spaces.

4.3 A sparse approach with multi-wavelet basis functions

4.3.1 The sparse wavelet basis construction

For simplicity we restrict to the case $I_{\mathbf{z}} = [-1,1]^d$, and $\pi(\mathbf{z}) = \frac{1}{2^d}$ is the uniform distribution. We follow the notation by Guo and Cheng [46]. We start by constructing a hierarchical decomposition of the space consisting of piecewise polynomials of degree at most m. Let $P^m(a,b)$ be the space of polynomials of degree at most m on the interval (a,b), and for every

$N \geq 0$,

$$V_N^m = \{\phi : \phi \in P^m(-1 + 2^{-N+1}j, -1 + 2^{-N+1}(j+1)), j = 0, 1, \ldots, 2^N - 1\}. \tag{4.12}$$

Then define the wavelet space $W_N^m, N = 1, 2, \ldots$ as the orthogonal complement of V_{N-1}^m inside V_N^m. For convenience we define $W_0^m = V_0^m$. Then one obtains the hierarchical decomposition $V_N^m = \oplus_{0 \leq j \leq N} W_j^m$.

Then a standard sparse trick can be applied. For simplicity we introduce the following vector notations:

If $\mathbf{i} = (i_1, \ldots, i_d), \quad \mathbf{j} = (j_1, \ldots, j_d)$ then

$$\mathbf{i} \leq \mathbf{j} \text{ means } i_1 \leq j_1, \ldots, i_d \leq j_d,$$

$$\binom{\mathbf{j}}{\mathbf{i}} := \binom{j_1}{i_1} \times \cdots \times \binom{j_d}{i_d},$$

$\mathbf{1_m}$ is the vector with 1 at m-th component and 0 elsewhere,

$$|\mathbf{i}|_\infty = \max_{1 \leq m \leq d} \{|i_m|\}, \quad |\mathbf{i}|_1 = |i_1| + \cdots + |i_d|.$$

Define the d-fold tensor product of V_N^m by

$$\mathbf{V}_{N,\mathbf{z}}^m = V_{N,z_1}^m \times \cdots \times V_{N,z_d}^m. \tag{4.13}$$

Similarly define the d-fold tensor product of $W_{\mathbf{j}}^m$ by

$$\mathbf{W}_{\mathbf{j},\mathbf{z}}^m = W_{j_1,z_1}^m \times \cdots \times W_{j_d,z_d}^m. \tag{4.14}$$

Then

$$\mathbf{V}_{N,\mathbf{z}}^m = \oplus_{0 \leq |\mathbf{j}|_\infty \leq N} \mathbf{W}_{\mathbf{j},\mathbf{z}}^m.$$

The sparse trick is to replace the l^∞ norm on $\mathbf{j}$ by the l^1 norm. In this way we define the sparse wavelet space

$$\hat{\mathbf{V}}_{N,\mathbf{z}}^m = \oplus_{0 \leq |\mathbf{j}|_1 \leq N} \mathbf{W}_{\mathbf{j},\mathbf{z}}^m. \tag{4.15}$$

From now on we will omit the subscript $\mathbf{z}$ for these spaces.

4.3.2 Construction of the basis functions

We adopt the basis functions of W_j^m constructed by Alpert [2]. The basis functions of W_j^m are denoted by $\psi_{j,l}^{m'}$, $m' = 0, 1, \ldots, m$, $l = 0, 1, \ldots, 2^{j-1} - 1$ for $j \geq 1$ and $l = 0$ for $j = 0$. $\psi_{0,0}^{m'}$ are the orthonormal Legendre polynomials of degree m' on $[-1, 1]$, and $\psi_{1,0}^{m'}$ are piecewise polynomials on $[-1, 0]$ and $[0, 1]$ that are orthogonal to those Legendre polynomials, which can be constructed by a procedure similar to the Gram-Schmidt orthogonalization. Other $\psi_{j,l}^{m'}$ are defined by dilation and translation of $\psi_{1,0}^{m'}$:

$$\psi_{j,l}^{m'}(y) = 2^{(j-1)/2}\psi_{1,0}^{m'}(2^{j-1}y + 2^{j-1} - 1 - 2l), \quad j = 2, 3, \ldots, \quad l = 0, 1, \ldots, 2^{j-1} - 1,$$

which has support on the interval $[-1 + 2^{2-j}l, -1 + 2^{2-j}(l + 1)]$.

The basis functions of $\mathbf{W}_{\mathbf{j}}^m$ are tensor products of the one dimensional basis functions:

$$\psi_{\mathbf{j},\mathbf{l}}^{\mathbf{m}'}(\mathbf{z}) = \psi_{j_1,l_1}^{m_1'}(z_1) \times \cdots \times \psi_{j_d,l_d}^{m_d'}(z_d), \quad 0 \leq |\mathbf{m}'|_\infty \leq m, 0 \leq l_1 \leq 2^{j_1-1} - 1, \ldots, 0 \leq l_d \leq 2^{j_d-1} - 1,$$

and the basis functions of $\hat{\mathbf{V}}_N^m$ consist of all the above functions for $0 \leq |\mathbf{j}|_1 \leq N$. By re-ordering the basis functions for $\hat{\mathbf{V}}_N^m$ we make them $\Phi_1(\mathbf{z}), \ldots, \Phi_K(\mathbf{z})$, where $K = K(m, N, d)$ is the total number of basis functions. It is proved in Lemma 2.3 of [101] that

$$K = O((m + 1)^d 2^N N^{d-1}). \tag{4.16}$$

4.4 Estimate of the Sparsity of S_{ijk}

Recall the triple product tensor S_{ijk} defined in (1.50). Due to the local support of the sparse wavelet basis functions Φ_k, this tensor is sparse, especially when N and d are large. Due to this sparsity, when one computes $Q_k = \sum_{i,j=1}^K S_{ijk} Q(f_i, f_j)$, one only needs to compute those $Q(f_i, f_j)$ where there is at least one k with $S_{ijk} \neq 0$. Now we prove some results on its sparsity. We focus on the dependence on N, so every $O(\cdot)$ notation means multiplication by a constant that may depend on d.

Recall that when one takes the sparse wavelet space $\hat{\mathbf{V}}_N^m$, the basis functions are

$$\begin{aligned} \psi_{\mathbf{j},\mathbf{l}}^{\mathbf{m}'}(\mathbf{z}) = \psi_{j_1,l_1}^{m_1'}(z_1) \times \cdots \times \psi_{j_d,l_d}^{m_d'}(z_d), \quad 0 \leq |\mathbf{m}'|_\infty \leq m, \\ 0 \leq l_1 \leq 2^{j_1-1} - 1, \ldots, 0 \leq l_d \leq 2^{j_d-1} - 1, \quad |\mathbf{j}|_1 \leq N. \end{aligned} \tag{4.17}$$

The function $\psi_{j,l}^{m'}(z)$ is supported on the interval $[-1+2^{2-j}l, -1+2^{2-j}(l+1)]$ for $j \geq 1$. Since this support is independent of m', we omit the m' index in the following consideration. If $\psi_{\mathbf{j}^1,\mathbf{l}^1}$ and $\psi_{\mathbf{j}^2,\mathbf{l}^2}$ have non-intersecting supports, then

$$\int_{I_{\mathbf{z}}} b(\mathbf{z})\psi_{\mathbf{j}^1,\mathbf{l}^1}(\mathbf{z})\psi_{\mathbf{j}^2,\mathbf{l}^2}(\mathbf{z})\psi_{\mathbf{j}^3,\mathbf{l}^3}(\mathbf{z})\pi(\mathbf{z})\,\mathrm{d}\mathbf{z} = 0, \quad \forall \mathbf{j}_3, \mathbf{l}_3.$$

Recall that the number of basis functions, in $\hat{\mathbf{V}}_N^m$, which includes those $\psi_{\mathbf{j},\mathbf{l}}$ with $|\mathbf{j}|_1 \leq N$ and $0 \leq l_1 \leq 2^{j_1-1}-1, \ldots, 0 \leq l_d \leq 2^{j_d-1}-1$, is $O((m+1)^d 2^N N^{d-1})$. Thus the number of the pairs of such functions is $O((m+1)^{2d}2^{2N}N^{2d-2})$. Now we state our result:

Theorem 4.4.1. *The pairs of basis functions of $\hat{\mathbf{V}}_N^m$ with intersecting supports have a total number at most $O((m+1)^{2d}2^{2N}N^{d+1})$.*

Proof. The number of $\phi_{j,l}$ for a fixed j is $(m+1)2^{j-1}$ for $j \geq 1$, and $m+1$ if $j = 0$. Thus it is less than or equal to $(m+1)2^j$ for all j. For fixed j^1, j^2, suppose $j^1 \geq j^2$, then ϕ_{j^1,l^1} and ϕ_{j^2,l^2} have intersecting supports if and only if the support of ϕ_{j^1,l^1} is a subinterval of the support of ϕ_{j^2,l^2}. For every l^1, there is one and only one such l^2. Thus the number of pairs l^1, l^2 such that ϕ_{j^1,l^1} and ϕ_{j^2,l^2} have intersecting supports is 2^{j^1}, which is $2^{\max\{j^1,j^2\}}$ in general.

Thus the desired number is

$$S = (m+1)^{2d} \sum_{0 \leq |\mathbf{j}^1|_1 \leq N, 0 \leq |\mathbf{j}^2|_1 \leq N} 2^{\max\{j_1^1,j_1^2\}+\cdots+\max\{j_d^1,j_d^2\}}. \tag{4.18}$$

Let $\mathbf{k}^1 = \max\{\mathbf{j}^1, \mathbf{j}^2\}$, where the maximum acts on each component of vectors. Similarly let $\mathbf{k}^2 = \min\{\mathbf{j}^1, \mathbf{j}^2\}$. Then $|\mathbf{k}^1 + \mathbf{k}^2|_1 = |\mathbf{j}^1 + \mathbf{j}^2|_1 = |\mathbf{j}^1|_1 + |\mathbf{j}^2|_1 \leq 2N$, and for each fixed $\mathbf{k}^1, \mathbf{k}^2$, there are at most 2^d pairs of $\mathbf{j}^1, \mathbf{j}^2$ satisfying the conditions $\mathbf{k}^1 = \max\{\mathbf{j}^1, \mathbf{j}^2\}$ and $\mathbf{k}^2 = \min\{\mathbf{j}^1, \mathbf{j}^2\}$. Thus

$$\begin{aligned}
S &\leq C(d)(m+1)^{2d} \sum_{0 \leq |\mathbf{k}^1|_1 + |\mathbf{k}^2|_1 \leq 2N} 2^{|\mathbf{k}^1|_1} \\
&= C(d)(m+1)^{2d} \sum_{k=0}^{2N} 2^k \binom{k+d-1}{d-1} \sum_{l=0}^{2N-k} \binom{l+d-1}{d-1} \\
&\leq C(d)(m+1)^{2d} N \sum_{k=0}^{2N} 2^k (k+1)^{d-1}(2N-k+1)^{d-1}.
\end{aligned}$$

The first equality is because there are $\binom{k+d-1}{d-1}$ choices of $\mathbf{k}^1$ with $|\mathbf{k}^1|_1 = k$, and similarly for $\mathbf{k}^2$. The second inequality is because $\binom{k+d-1}{d-1} = \frac{k+1}{1}\frac{k+2}{2}\cdots\frac{k+d-1}{d-1} \leq (k+1)^{d-1}$, and taking the largest term in the l summation.

Then by taking derivative with respect of k, it is easy to see that the previous summation is optimized at $k_{max} = 2N - O(d)$. Thus

$$\begin{aligned} S &\leq C(d)(m+1)^{2d}N^2 2^{k_{max}}(k_{max}+1)^{d-1}(2N-k_{max}+1)^{d-1} \\ &\leq C(d)(m+1)^{2d}2^{2N}N^{d+1}, \end{aligned}$$

which finishes the proof. □

Remark 4.1. *When $d \geq 4$, one has $2^{2N}N^{2d-2} > 2^{2N}N^{d+1}$, thus in this case the number of $Q(f_i, f_j)$ needed to be computed is much less than the total number of pairs of f_i, f_j. And the bigger d is, the more saving one will gain.*

Numerically, we observe this sparsity result even in the cases $d = 2, 3$ (see Section 4.6.1), and for a fixed d, the percentage of $Q(f_i, f_j)$ needed to be computed decreases exponentially as N increases, which is better than what one expects from the above theorem (where the percentage is $O(\frac{1}{N^{d-3}})$). This suggests that the above theorem is not sharp.

4.5 Regularity and accuracy

In this section, we prove the regularity of the solution to the Boltzmann equation in the random space, and the accuracy of the stochastic Galerkin method using sparse wavelet basis. These are straightforward multi-dimensional extensions of the corresponding results in [50]. We assume that the random collision kernel depends linearly on $\mathbf{z}$. This is a reasonable assumption because when one uses the Karhunen-Loeve expansion to approximate a random field, the resulting dependence on $\mathbf{z}$ is linear.

We consider the spatially homogeneous Boltzmann equation

$$\frac{\partial f}{\partial t} = Q(f, f), \tag{4.19}$$

subject to random initial data and random collision kernel

$$f(0,\mathbf{v},\mathbf{z}) = f^0(\mathbf{v},\mathbf{z}), \quad B = B(\mathbf{v},\mathbf{v}_*,\sigma,\mathbf{z}), \quad \mathbf{z}\in I_{\mathbf{z}}.$$

4.5.1 Regularity in the random space for the Boltzmann equation

We define the norms and operators:

$$\|f(t,\cdot,\mathbf{z})\|_{L^p_{\mathbf{v}}} = \left(\int_{\mathbb{R}^{d_v}} |f(t,\mathbf{v},\mathbf{z})|^p \,\mathrm{d}\mathbf{v}\right)^{1/p}, \quad \|f(t,\mathbf{v},\cdot)\|_{L^2_{\mathbf{z}}} = \left(\int_{I_{\mathbf{z}}} f(t,\mathbf{v},\mathbf{z})^2\pi(\mathbf{z})\,\mathrm{d}\mathbf{z}\right)^{1/2},$$

$$|||f(t,\cdot,\cdot)|||_k = \sup_{\mathbf{z}\in I_{\mathbf{z}}}\left(\sum_{|\mathbf{l}|=0}^{k} \|\partial^{\mathbf{l}}_{\mathbf{z}} f(t,\mathbf{v},\mathbf{z})\|^2_{L^2_{\mathbf{v}}}\right)^{1/2},$$

$$Q(g,h)(\mathbf{v}) = \int_{\mathbb{R}^{d_v}}\int_{\mathbb{S}^{d_v-1}} B(\mathbf{v},\mathbf{v}_*,\sigma,\mathbf{z})\left[g(\mathbf{v}')h(\mathbf{v}'_*) - g(\mathbf{v})h(\mathbf{v}_*)\right]\,\mathrm{d}\sigma\,\mathrm{d}\mathbf{v}_*,$$

$$Q_{1,j}(g,h)(\mathbf{v}) = \int_{\mathbb{R}^{d_v}}\int_{\mathbb{S}^{d_v-1}} \partial_{z_j} B(\mathbf{v},\mathbf{v}_*,\sigma,\mathbf{z})\left[g(\mathbf{v}')h(\mathbf{v}'_*) - g(\mathbf{v})h(\mathbf{v}_*)\right]\,\mathrm{d}\sigma\,\mathrm{d}\mathbf{v}_*.$$

We first state the following estimates of $Q(g,h)$ and $Q_{1,j}(g,h)$, which are standard results proved in [72, 12] and its extension to the uncertain case is straightforward:

Lemma 4.2. *Assume the collision kernel B depends on $\mathbf{z}$ linearly, B and $\partial_{\mathbf{z}}B$ are locally integrable and bounded in $\mathbf{z}$. If $g,h\in L^1_{\mathbf{v}}\cap L^2_{\mathbf{v}}$, then*

$$\|Q(g,h)\|_{L^2_{\mathbf{v}}},\ \|Q_{1,j}(g,h)\|_{L^2_{\mathbf{v}}} \le C_B\|g\|_{L^1_{\mathbf{v}}}\|h\|_{L^2_{\mathbf{v}}}, \tag{4.20}$$

$$\|Q(g,h)\|_{L^2_{\mathbf{v}}},\ \|Q_{1,j}(g,h)\|_{L^2_{\mathbf{v}}} \le C_B\|g\|_{L^2_{\mathbf{v}}}\|h\|_{L^2_{\mathbf{v}}}, \tag{4.21}$$

where the constant $C_B>0$ depends only on B and $\partial_{z_j}B, j=1,\dots,d$.

Now we state our estimate on $|||f|||_k$.

Theorem 4.5.1. *Assume that B satisfies the assumption in Lemma 4.2, and $\sup_{\mathbf{z}\in I_{\mathbf{z}}}\|f^0\|_{L^1_{\mathbf{v}}} \le M$, $|||f^0|||_k<\infty$ for some integer $k\ge 0$. Then there exists a constant $C_k>0$, depending only on C_B, M, T, and $|||f^0|||_k$ such that*

$$|||f|||_k \le C_k, \qquad \text{for any} \quad t\in[0,T]. \tag{4.22}$$

The proof of the theorem is provided in the Appendix B.

4.5.2 Accuracy analysis

In this subsection, we will prove the convergence rate of the stochastic Galerkin method using the previously established regularity. As in section 4.5.1, we will still restrict to the spatially homogeneous equation (4.19).

We use the sparse wavelet space $\hat{\mathbf{V}}_N^m$ with parameters m, N. For this space, the number of basis functions $K = O((m+1)^d 2^N N^{d-1})$.

Define the space $\mathcal{H}^m(I_{\mathbf{z}})$ by

$$\|f\|_{\mathcal{H}^m(I_{\mathbf{z}})} = \max \sum_{0\le m_{i_1},\dots,m_{i_r}\le m} \sum_{0\le m_{j_1},\dots,m_{j_{d-r}}\le 1} \|\partial_{z_{i_1}}^{m_{i_1}} \cdots \partial_{z_{i_r}}^{m_{i_r}} \partial_{z_{j_1}}^{m_{j_1}} \cdots \partial_{z_{j_{d-r}}}^{m_{j_{d-r}}} f\|_{L^2(I_{\mathbf{z}})},$$

where the maximum is taken over all non-empty subsets $\{i_1, \dots, i_r\} \subset \{1, \dots, d\}$, and $\{j_1, \dots, j_{d-r}\}$ is the complement of $\{i_1, \dots, i_r\}$. Using the orthonormal basis $\{\Phi_k(z)\}$, the solution f to (4.19) can be represented as

$$f(t, \mathbf{v}, \mathbf{z}) = \sum_{k=1}^{\infty} f_k(t, \mathbf{v}) \Phi_k(\mathbf{z}), \quad \text{where} \quad f_k(t, \mathbf{v}) = \int_{I_{\mathbf{z}}} f(t, \mathbf{v}, \mathbf{z}) \Phi_k(\mathbf{z}) \pi(\mathbf{z}) \, \mathrm{d}\mathbf{z} \,. \tag{4.23}$$

Let P_K be the projection operator defined as

$$P_K f(t, \mathbf{v}, \mathbf{z}) = \sum_{k=1}^{K} f_k(t, \mathbf{v}) \Phi_k(\mathbf{z}).$$

Then one has the following projection error estimate (Theorem 5.1 in [90]):

Lemma 4.3. *For any $f \in \mathcal{H}^{m+1}(I_{\mathbf{z}})$, $N \ge 1$, we have*

$$\|P_K f - f\|_{L^2(I_{\mathbf{z}})} \le C(m, d) N^{d-1} \, 2^{-N(m+1)} \|f\|_{\mathcal{H}^{m+1}(I_{\mathbf{z}})}. \tag{4.24}$$

This lemma implies that the projection error

$$\|P_K f - f\|_{L^2(I_{\mathbf{z}})} \le C(m, d) K^{-(m+1)} (\log K)^{(m+2)(d-1)} \|f\|_{\mathcal{H}^{m+1}(I_{\mathbf{z}})}. \tag{4.25}$$

Define the norms

$$\|f(t, \cdot, \cdot)\|_{L^2_{\mathbf{v},\mathbf{z}}} = \left(\int_{I_{\mathbf{z}}} \int_{\mathbb{R}^d} f(t, \mathbf{v}, \mathbf{z})^2 \, \mathrm{d}\mathbf{v} \pi(\mathbf{z}) \, \mathrm{d}\mathbf{z} \right)^{1/2}, \tag{4.26}$$

then we have the following:

Lemma 4.4. *Assume* $\mathbf{z}$ *obeys the uniform distribution, i.e.,* $\mathbf{z} \in I_{\mathbf{z}} = [-1,1]^d$ *and* $\pi(\mathbf{z}) = 1/2^d$*. If* $|||f^0|||_{d(m+1)}$ *is bounded, then*

$$\|P_K f - f\|_{L^2_{\mathbf{v},z}} \leq C(m,d) K^{-(m+1)} (\log K)^{(m+2)(d-1)}, \tag{4.27}$$

where $C(m,d)$ *is a constant depending on* m *and* d*.*

Given the gPC approximation of f:

$$f^K(t,\mathbf{v},\mathbf{z}) = \sum_{k=1}^{K} \hat{f}_k(t,\mathbf{x},\mathbf{v})\Phi_k(\mathbf{z}), \tag{4.28}$$

we now define the error function

$$e^K(t,\mathbf{v},\mathbf{z}) = P_K f(t,\mathbf{v},\mathbf{z}) - f^K(t,\mathbf{v},\mathbf{z}) := \sum_{k=1}^{K} e_k(t,\mathbf{v})\Phi_k(\mathbf{z}),$$

where $e_k = \hat{f}_k - f_k$. Then we have

Theorem 4.5.2. *Assume the random variable* z *and initial data* f^0 *satisfy the assumption in Lemma 4.3, and the gPC approximation* f^K *is uniformly bounded in* K*, then*

$$\|f - f^K\|_{L^2_{\mathbf{v},\mathbf{z}}} \leq C(t)\left\{C(m,d)K^{-(m+1)}(\log K)^{(m+2)(d-1)} + \|e^K(0)\|_{L^2_{\mathbf{v},\mathbf{z}}}\right\}.$$

The proof of Lemma 4.4 and Theorem 4.5.2 can be proved in the same way as Section 4.2 in Hu and Jin [50], in view of Lemma 4.3. We omit the details.

Remark 4.5. *In general, wavelet bases are used for functions with low regularity. Here we briefly explain the reason why we use them for smooth functions. For low dimensional random spaces (*$d \leq 4$*), by choosing a large* m *(i.e.,* $m \geq 2$*) one can obtain a good accuracy order (almost* $(m+1)$*-th order) with the wavelet basis. However, due to the factor* $(m+1)^d$ *in the number of basis functions* K *(see (4.16)),* m *cannot be large for higher dimensional random spaces (*$d \geq 5$*). Thus for such random spaces one has to sacrifice the accuracy order a little and take* $m = 0, 1$ *in order to make the number of basis functions* K *affordable.*

4.6 Numerical results

In this section we give some numerical results of the stochastic Galerkin method with sparse technique. We first demonstrate the efficiency of the sparse wavelet basis, and then show its application to the Boltzmann equation with uncertainty.

The random space is taken as $[0,1]^d$ with the uniform distribution. For the Boltzmann equation with uncertainty, the physical space is taken as $[0,1]$, and the velocity space is truncated as $[-R_v, R_v]^2$. The physical space is discretized into N_x grid points

$$x_i = (i + \frac{1}{2})\Delta x, \quad i = 0, 1, \ldots, N_x - 1, \tag{4.29}$$

where $\Delta_x = \frac{1}{N_x}$. The velocity space is discretized into N_v grid points in each dimension:

$$v_{i,j} = (-R_v + (i + \frac{1}{2})\Delta v, -R_v + (j + \frac{1}{2})\Delta v), \quad i, j = 0, 1, \ldots, N_v - 1, \tag{4.30}$$

where $\Delta v = \frac{2R_v}{N_v}$.

The flux term $\mathbf{v} \cdot \nabla_{\mathbf{x}} f_k$ in (4.8) is discretized by the second order upwind scheme with the minmod slope limiter. The collision operator is computed by the fast spectral method [81]. The time discretization is given by the second order Runge-Kutta scheme.

4.6.1 The sparse wavelet basis

- **Number of basis functions**

  We first give a comparison of number of basis functions between our sparse wavelet function space $\hat{\mathbf{V}}_N^m$ and the tensor basis $\mathbf{V}_N^m$. The result is shown in Table 1. It is clear that the sparse technique saves a great number of basis functions, especially in multi-dimensional random spaces.

- **Efficiency of the sparse wavelet function space**

  We give a comparison of the L^2 approximation error of $\hat{\mathbf{V}}_N^m$ and $\mathbf{V}_N^m$. For each random dimension $d = 2, 3, 4$ we pick a smooth test function as follows:

$$f(\mathbf{z}) = \frac{1}{2\pi\mathcal{K}(\mathbf{z})^2} \exp\left(-\frac{1}{2\mathcal{K}(\mathbf{z})}\right)\left(2\mathcal{K}(\mathbf{z}) - 1 + \frac{1 - \mathcal{K}(\mathbf{z})}{2\mathcal{K}(\mathbf{z})}\right), \tag{4.31}$$

(a) $m = 0$

	$N = 3$	$N = 4$	$N = 5$
$d = 1$	8,8	16,16	32,32
$d = 2$	20,64	48,256	112,1024
$d = 3$	38,512	104,4096	272,32768
$d = 4$	63,4096	192,65536	552,1048576

(b) $m = 1$

	$N = 3$	$N = 4$	$N = 5$
$d = 1$	16,16	32,32	64,64
$d = 2$	80,256	192,1024	448,4096
$d = 3$	304,4096	832,32768	2176,262144

Table 4.1 Comparison of number of basis functions: m is the maximal degree of polynomials. d is the dimension; in each cell, the left number is the number of basis of functions of $\hat{\mathbf{V}}_N^m$; the right number is the number of basis of functions of $\mathbf{V}_N^m$.

where

$$\begin{aligned}
\mathcal{K}_{d=2}(\mathbf{z}) &= 1 - 0.5(0.5 + 0.1\sin(z_1) + 0.1\sin(2z_2)),\\
\mathcal{K}_{d=3}(\mathbf{z}) &= 1 - 0.5(0.5 + 0.1\sin(z_1) + 0.1\sin(2z_2) + 0.1\cos(z_3)),\\
\mathcal{K}_{d=4}(\mathbf{z}) &= 1 - 0.5(0.5 + 0.1\sin(z_1) + 0.1\sin(2z_2) + 0.1\cos(z_3) + 0.1\cos(2z_4)).
\end{aligned} \tag{4.32}$$

We use the function spaces $\hat{\mathbf{V}}_N^m$ and $\mathbf{V}_N^m$ with different m, N values to approximate these functions, and compute their relative L^2 error $\frac{\|f - P_K f\|_{L^2}}{\|f\|_{L^2}}$, where P_K is the projection operator onto the corresponding function space. The result is shown in Figure 1. It can be seen that the sparse wavelet method performs much better than the tensor method.

Sparsity of S_{ijk} We give a test of the sparsity of the tensor S_{ijk}, as well as the number of $Q(f_i, f_j)$ needed to compute. We take a random collision kernel $b(\mathbf{z}) = 1 + 0.2z_1$. For simplicity we only show the results with $m = 0$, since the sparsity of S_{ijk} with larger m is similar. The result is shown in Figure 2. One can clearly see an exponential decay of the percentage of nonzeros in S_{ijk}, as well as the percentage of $Q(f_i, f_j)$ needed to compute, as N or d increase. This is even better than what we have proved.

To further demonstrate the sparsity of S_{ijk} we give a graph of nonzero elements of S_{ijk} for $m = 0, N = 4, d = 3$, shown in Figure 3. The points in the first graph represent nonzero elements in S_{ijk}. The second graph is the projection of the first graph onto i, j coordinates, and the points in it represent those $Q(f_i, f_j)$ needed to compute.

4.6.2 Application to the Boltzmann equation with uncertainty

In this subsection, the velocity space is assumed to be two-dimensional and its discretization is always given by $N_v = 32$. The time discretization is given by 0.8 times the CFL condition for spatial inhomogeneous problems.

- **Accuracy of the approximation of the collision operator**

  We first check the accuracy of the collision operator $Q(f, f)$ computed by the sparse stochastic Galerkin method. The function f is given by the Bobylev-Krook-Wu [10, 63]

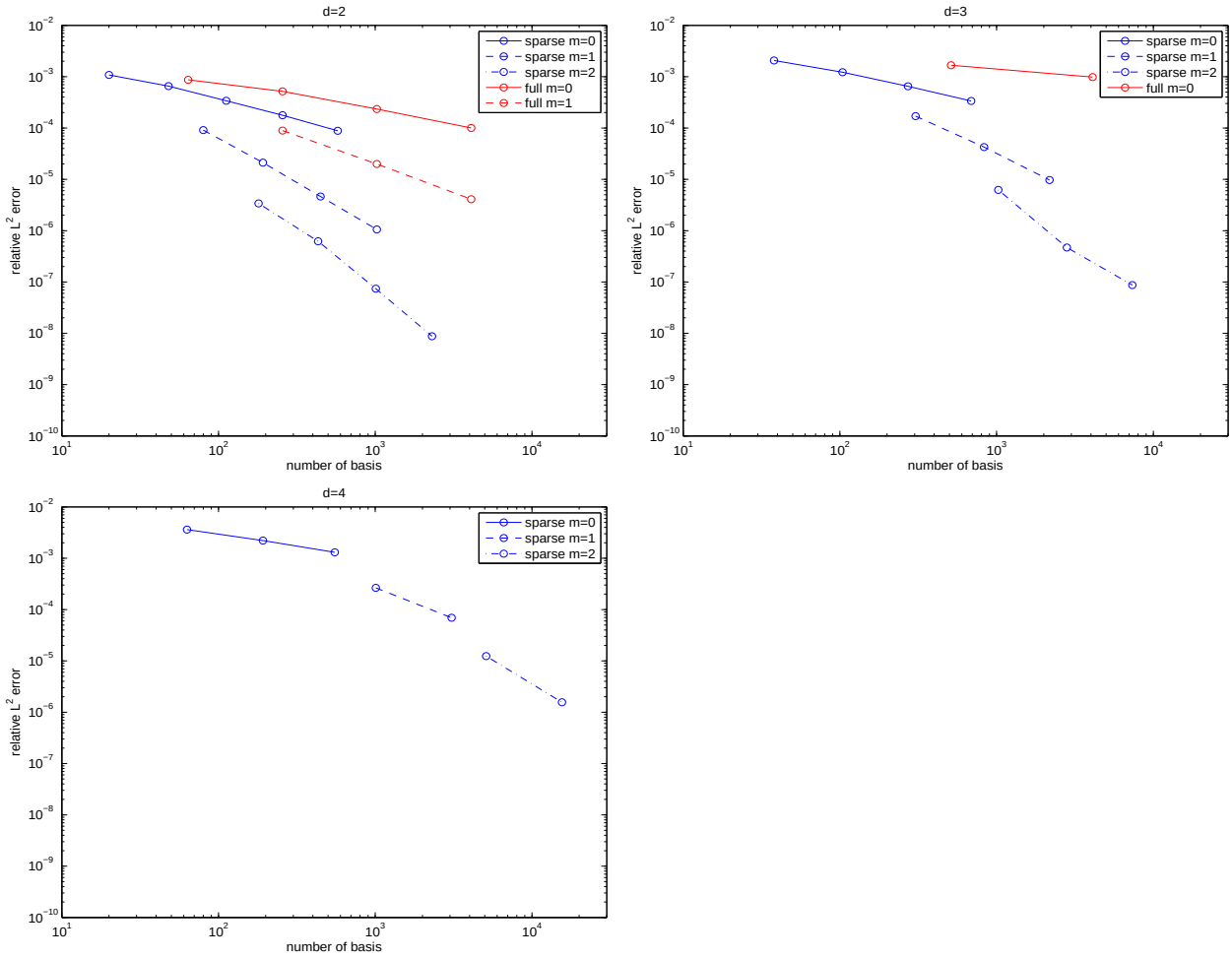

Figure 4.1 Comparison of approximation error of both sparse basis and full tensor basis for $d = 2, 3, 4$. For $d = 4$ we do not give the result by tensor basis because the number of basis functions is too large.

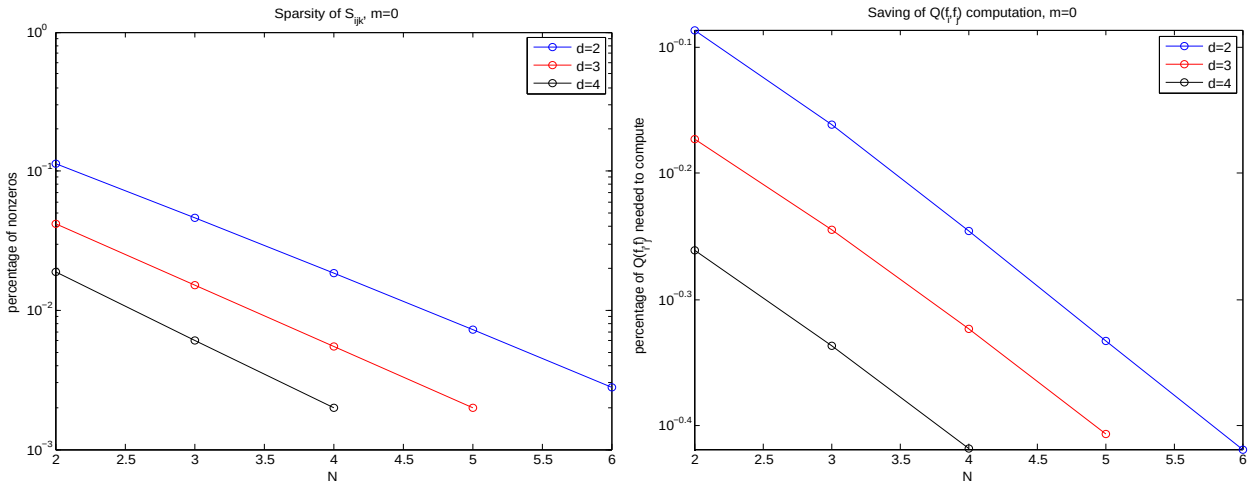

Figure 4.2 Sparsity of S_{ijk} and the number of $Q(f_i, f_j)$ needed to compute, $d = 2, 3, 4$, $m = 0$.

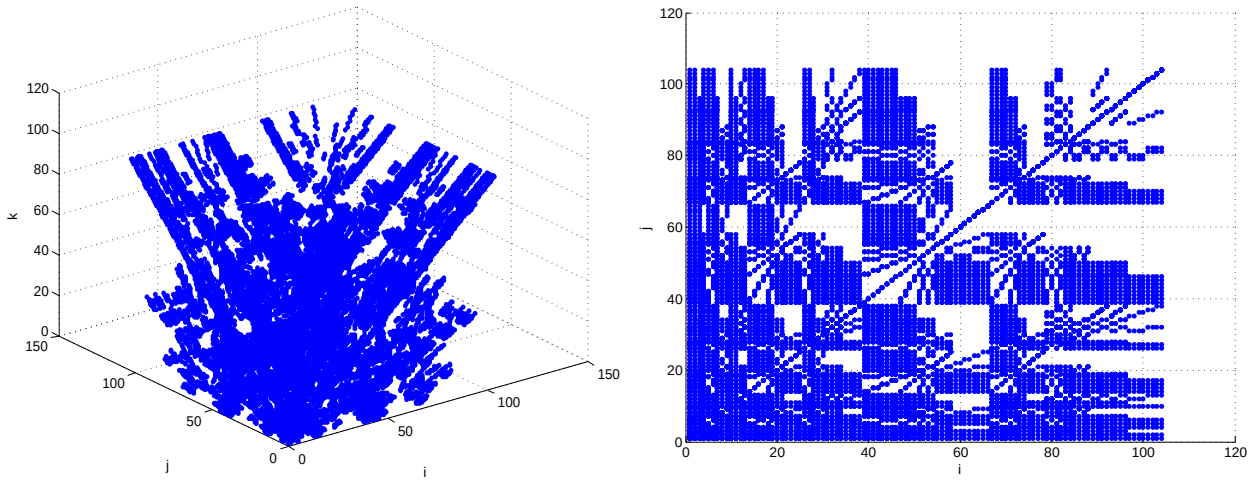

Figure 4.3 Demonstration of sparsity of S_{ijk}: $m = 0, N = 4, d = 3$. Left: blue points represent non-zeros terms of S_{ijk}. Right: blue points represent a pair (i, j) with $S_{ijk} \neq 0$ for some k.

solution with uncertainty:

$$f(\mathbf{v},\mathbf{z}) = \frac{1}{2\pi\mathcal{K}(\mathbf{z})^2}\exp\left(-\frac{|\mathbf{v}|^2}{2\mathcal{K}(\mathbf{z})}\right)\left(2\mathcal{K}(\mathbf{z}) - 1 + \frac{1-\mathcal{K}(\mathbf{z})}{2\mathcal{K}(\mathbf{z})}\mathbf{v}^2\right), \tag{4.33}$$

where

$$\begin{aligned}
\mathcal{K}_{d=2}(\mathbf{z}) &= 1 - 0.5(0.5 + 0.1\sin(z_1) + 0.1\sin(2z_2)),\\
\mathcal{K}_{d=3}(\mathbf{z}) &= 1 - 0.5(0.5 + 0.1\sin(z_1) + 0.1\sin(2z_2) + 0.1\cos(z_3)),\\
\mathcal{K}_{d=4}(\mathbf{z}) &= 1 - 0.5(0.5 + 0.1\sin(z_1) + 0.1\sin(2z_2) + 0.1\cos(z_3) + 0.1\cos(2z_4)).
\end{aligned} \tag{4.34}$$

For this f, $Q(f,f)$ with collision kernel $B = \frac{1}{2\pi}$ is given explicitly by

$$\begin{aligned}
Q(f,f)(\mathbf{v},\mathbf{z}) = &\left(\left(-\frac{2}{\mathcal{K}(\mathbf{z})} + \frac{|\mathbf{v}|^2}{2\mathcal{K}(\mathbf{z})^2}\right) f\right.\\
&\left. + \frac{1}{2\pi\mathcal{K}(\mathbf{z})^2}\exp\left(-\frac{|\mathbf{v}|^2}{2\mathcal{K}(\mathbf{z})}\right)\left(2 - \frac{1}{2\mathcal{K}(\mathbf{z})^2}|\mathbf{v}|^2\right)\right)\frac{1-\mathcal{K}(\mathbf{z})}{8}.
\end{aligned} \tag{4.35}$$

The numerical solution is given by

$$\tilde{Q}(f,f)(\mathbf{v},\mathbf{z}) = \sum_{k=0}^{K} Q_k(\mathbf{v})\Phi_k(\mathbf{z}), \quad \text{where } Q_k(\mathbf{v}) = \sum_{i,j=0}^{K} S_{ijk}Q(f_i,f_j)(\mathbf{v}).$$

We compare the relative L^2 error for $d = 2,3,4$ and sparse basis $\hat{\mathbf{V}}_N^m$ with different m, N. The result is shown in Figure 4. One can clearly see the error is a little worse than $O(K^{-(m+1)})$, and it becomes a little worse as d increases. This is caused by the $\log K$ factor in the error estimate.

- **The homogeneous Boltzmann equation with uncertainty on the collision kernel**

  We solve the homogeneous Boltzmann equation with deterministic initial data and a random collision kernel. We take the dimension of the random space $d = 2,3$, and the collision kernels are

$$\begin{aligned}
b(\mathbf{z}) &= 1 + 0.2z_1 + 0.1z_2, \quad d = 2,\\
b(\mathbf{z}) &= 1 + 0.2z_1 + 0.1z_2 + 0.07z_3, \quad d = 3.
\end{aligned} \tag{4.36}$$

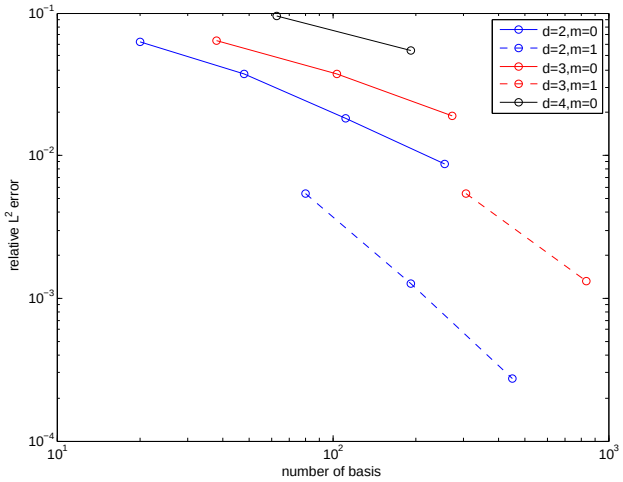

Figure 4.4 Accuracy of the approximation of the collision operator for $d = 2, 3, 4$.

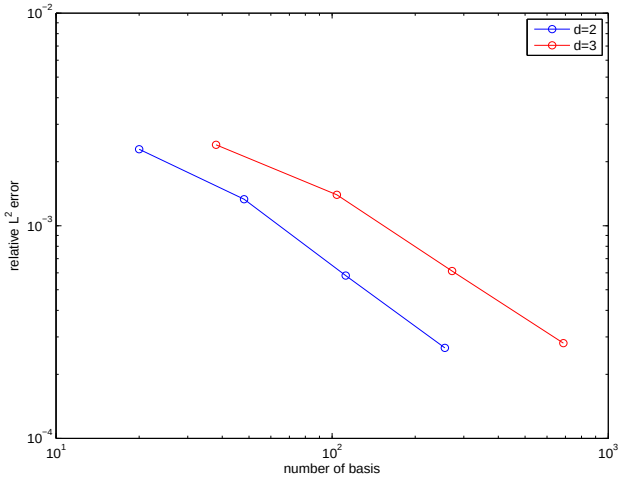

Figure 4.5 The homogeneous Boltzmann equation with a random collision kernel: accuracy result. $m = 0$, $\Delta t = 0.01$, $t = 1$.

The initial data is the BKW solution

$$f_0(\mathbf{v}, \mathbf{z}) = \frac{1}{\pi} \exp(-|\mathbf{v}|^2) \frac{|\mathbf{v}|^2}{2}, \tag{4.37}$$

and the exact solution is given by

$$f(t, \mathbf{v}, \mathbf{z}) = \frac{1}{2\pi\mathcal{K}^2} \exp\left(-\frac{|\mathbf{v}|^2}{2\mathcal{K}}\right) \left(2\mathcal{K} - 1 + \frac{1-\mathcal{K}}{2\mathcal{K}}|\mathbf{v}|^2\right), \tag{4.38}$$

with

$$\mathcal{K}(t, \mathbf{z}) = 1 - \exp(-b(\mathbf{z})t/8)/2. \tag{4.39}$$

We solve this equation by the sparse sG method with $m = 0$, time step $\Delta t = 0.01$ and final time $t = 1$, and check the relative L^2 error with the exact solution. The result is shown in Figure 5. The phenomenon is similar to the previous accuracy test.

- **The Boltzmann equation with random initial data**

We test our method on the (inhomogeneous) Boltzmann equation with uncertainty. The random space is 4-dimensional. We take the x-domain to be $[0, 1]$ with the periodic

boundary condition. We use the following random initial data to mimic the Karhunen-Loeve expansion

$$\begin{cases} \rho_0 = \dfrac{1}{3}\left(2+\sin(2\pi x)+\sin(4\pi x)z_1/2+\sin(6\pi x)z_2/4+\sin(8\pi x)z_3/6+\sin(10\pi x)z_4/7\right), \\ \mathbf{u}_0 = (0.2, 0), \\ T_0 = \dfrac{1}{4}\left(3+\cos(2\pi x)+\cos(4\pi x)z_1/2+\cos(6\pi x)z_2/4+\cos(8\pi x)z_3/6+\cos(10\pi x)z_4/7\right), \\ f = \dfrac{\rho_0}{4\pi T_0}\left(\exp(-\dfrac{|\mathbf{v}-\mathbf{u}_0|^2}{2T_0})+\exp(-\dfrac{|\mathbf{v}+\mathbf{u}_0|^2}{2T_0})\right). \end{cases} \tag{4.40}$$

The x-domain is discretized into $N_x = 50$ mesh points, and we compare the solution by the sparse stochastic Galerkin method with $m = 0, N = 3$ and a stochastic collocation method with full tensor basis in random space at time $t = 0.1$. The collocation method is implemented by solving the deterministic problem at points of the form $\mathbf{z} = (z_1, \ldots, z_d)$ where each z_i is one of the $M_z = 8$ Gauss-Legendre quadrature points (thus one needs to solve M_z^d deterministic problems). And then the mean and standard deviation are computed by numerical quadrature. The comparison result is shown in Figure 6. We see the results by the two methods agree well.

- **The Boltzmann equation with randomness on initial data, boundary data, and collision kernel**

  We finally solve the inhomogeneous Boltzmann equation with uncertainty on initial data, boundary data, and collision kernel. The random domain is taken to be 6-dimensional. We take the initial data to be the equilibrium with

$$\rho(x,\mathbf{z}) = 1, \quad \mathbf{u}(x,\mathbf{z}) = 0, \quad T = 1+0.5(1+0.2z_2)\exp(-100(1+0.1z_3)(x-0.4-0.01z_1)^2), \tag{4.41}$$

  and the boundary data is given by the Maxwellian boundary condition with random parameters

$$T_w = 1+0.2z_4, \quad \alpha = 0.5+0.3z_5. \tag{4.42}$$

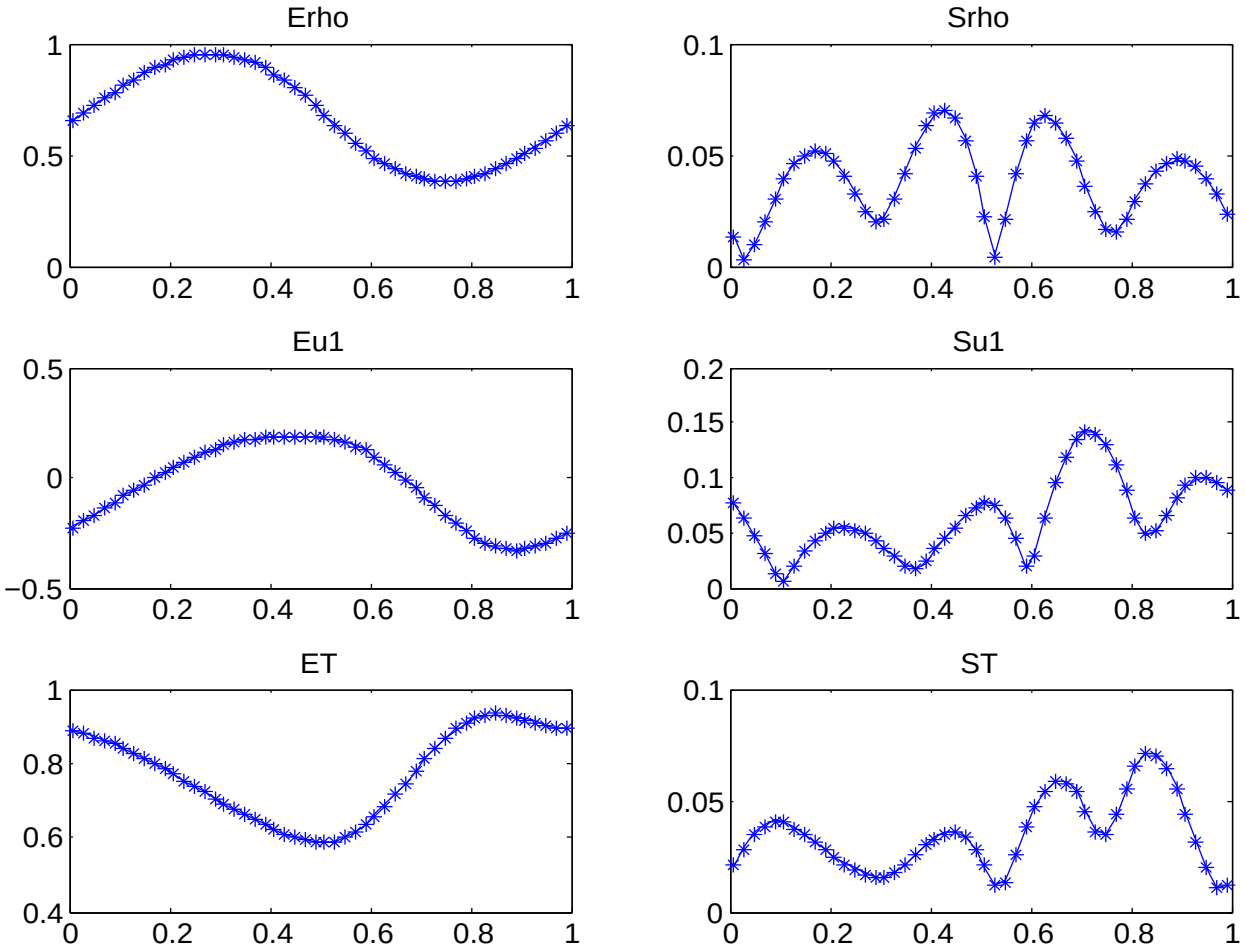

Figure 4.6 The Boltzmann equation with random initial data. $N_x = 50$, $t = 0.1$. Curve: collocation with $M_z = 8$; asterisks: Galerkin with $m = 0, N = 3$. Left column: mean of density, first component of bulk velocity, and temperature. Right column: standard deviation of density, first component of bulk velocity, and temperature.

The collision kernel is given by

$$b(\mathbf{z}) = 1 + 0.2z_6. \tag{4.43}$$

The spatial discretization is given by $N_x = 100$ to better capture the details near the boundary. We compare the result by the stochastic Galerkin method with sparse technique with the stochastic collocation method with full grid at time $t = 0.04$. The Galerkin method has parameters $m = 0, N = 3$, and the collocation method is as described in the previous numerical result with $M_z = 4$ collocation points in each dimension. The result is shown in Figure 7. One can see that the two results agree well.

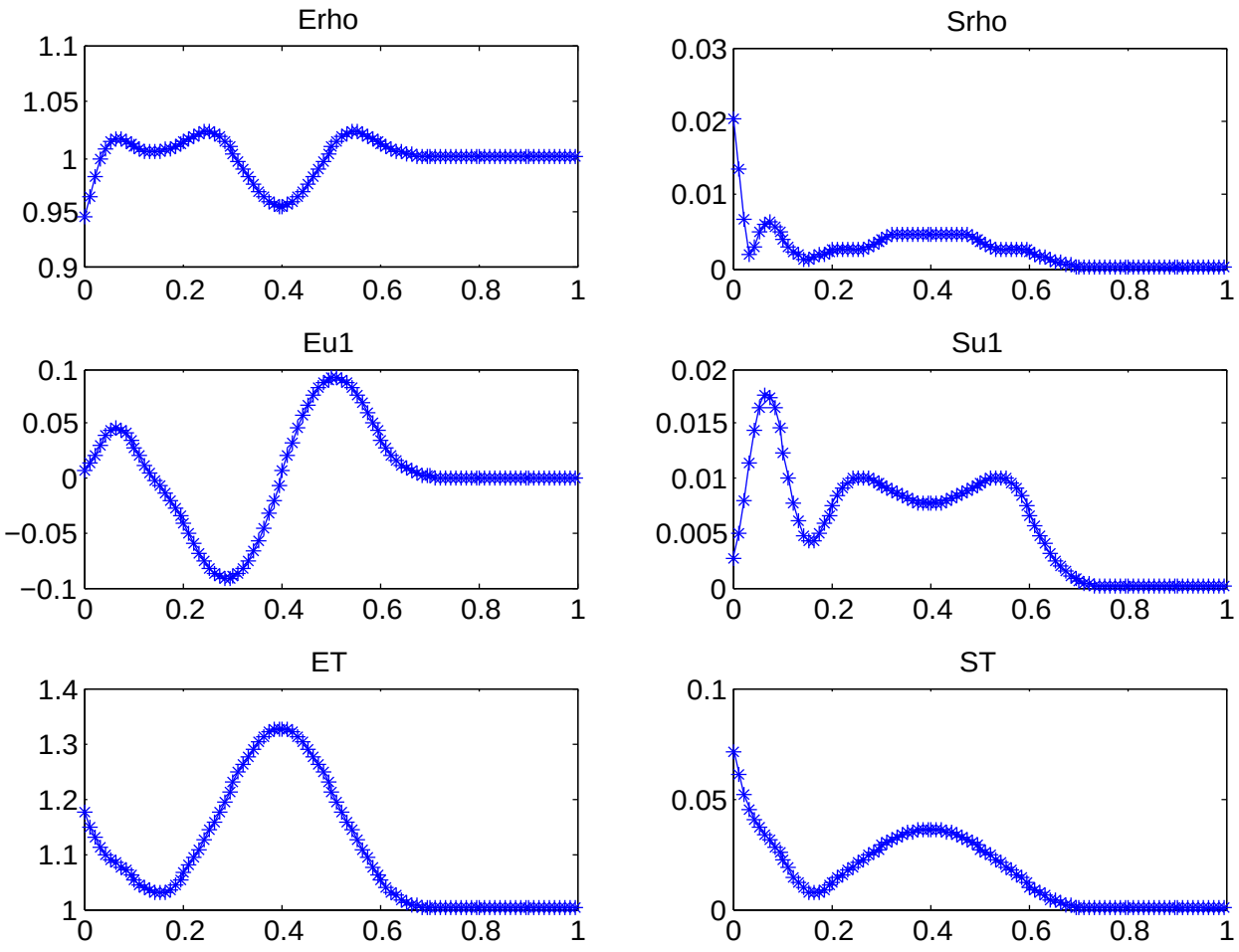

Figure 4.7 The Boltzmann equation with randomness on initial data, boundary data, and collision kernel ($d = 6$). $N_x = 100$, $t = 0.04$. Curve: collocation with $M_z = 4$; asterisks: Galerkin with $m = 0, N = 3$. Left column: mean of density, first component of bulk velocity, and temperature. Right column: standard deviation of density, first component of bulk velocity, and temperature.

Chapter 5

Polynomial interpolations for the Burgers equation with random inputs

In this chapter we go through the author's work with Q. Li and J.-G. Liu on the polynomial interpolations for the Burgers equation with random inputs.

5.1 Introduction

In this chapter we address one simple problem: can we use polynomial interpolation idea to compute the Burgers' equation, or general hyperbolic conservation laws with random initial data?

Polynomial interpolation type algorithms have been extensively used under the uncertainty quantification (UQ) framework. Given a PDE with some unknown coefficients, for the computation, one needs to make an ansatz for the solution. More specifically, we are given a PDE in a general form:

$$\mathcal{L}(x;z)u = 0\,,$$

with $\mathcal{L}$ being some operator on x and z being the random variables, and we assume that the solution u enjoys a certain form:

$$u(x;z) \in \mathrm{Span}\{\phi_i(x)\} \otimes \mathrm{Span}\{\psi_j(z)\}, \tag{5.1}$$

where ϕ_i and ψ_j are basis functions prepared on x and z directions respectively.

The so-termed generalized polynomial chaos (gPC) method and related ideas regard the solution well approximated by polynomials of random variables in the random space,

meaning $\psi_j(z)$ are simply set as polynomials of z. The generalized polynomial chaos-stochastic Galerkin method (gPC-SG) plugs the ansatz back into the equation and seeks for the governing equations for the projection coefficients, while the generalized polynomial chaos-stochastic collocation method (gPC-SC) computes the original equation multiple times on collocation grid points, and then interpolates on z for every x. The latter method is regarded non-intrusive since the forward PDE solver with deterministic coefficients could be recycled and run multiple times before assembled for the interpolation step, while the former one, on the opposite, is intrusive, and a different code is called for. Compared with the traditional Monte Carlo method, the advantage is obvious: the two polynomial based methods enjoy much faster convergence: ideally, the convergence is of spectral type (r^N) with N being the number of modes used in the ansatz and $r < 1$, instead of $1/\sqrt{N}$ provided by the Monte Carlo method.

Since the introduction of gPC based methods [35, 104, 106], they were widely used to deal with many equations under the UQ framework, and the polynomial ansatz was made for many equations with unknown coefficients. Many aspects centered around the methods have drawn lots of attention: this includes the challenges induced from the high dimensionality, which was partially addressed by the sparse grid idea [84], and then low rank structure or sparsity concept was also explored to reduce the computation [91, 19, 21, 6, 49, 48].

One challenge, however, is still left unaddressed: using polynomial basis functions lies in the spectral method framework. It brings the possibility of obtaining spectral accuracy, but it also naturally inherits the strong assumptions: for the spectral accuracy to be achieved, the solution has to be very smooth in the random space. The order of the regularity determines the order of the convergence rate: analyticity leads to exponential convergence but irregular solution leads to Gibbs phenomena, see reviews [45, 103]. The smoothness of the solution, however, needs to be studied case by case. For elliptic and parabolic equation, it was very well-studied. In [5, 4] the authors proved the analyticity in the random space that justifies the polynomial expansion for elliptic equations, and similar analysis was conducted for parabolic equations in [108]. Extensions to kinetic equations were performed in [50, 54, 71]. All

such analysis that justifies the validity of gPC type methods concentrates on showing the regularity of the solution in the random space, and it is typically done for equations that are essentially "good" – the solutions are known to be smooth. For systems that are known to be "bad", gPC seems to be useless [13]. If we take hyperbolic conservation laws, or particularly, the Burgers' equation, as a toy model for example: these equations are nonlinear, and even with C_∞ initial data, the solutions generically develop singularities, meaning that the accuracy deteriorates as the Gibbs phenomenon emerges.

We would like to better understand the break down of gPC on the conservation laws. More specifically, we study the regularity of the solution to the Burgers' equation with random initial data, and later extend the idea to hyperbolic equations with convex flux terms. We are going to show that although the solution itself, as a function of space, is not regular in the random space, the physical quantities such as shock locations, shock widths, and shock heights all smoothly depend on random variables. This means that the gPC method can still be used for these equations, but has to be used to deal with physical quantities mainly. We will prove this argument, and our numerical experiments also support it.

The idea was initially inspired by the rich studies on L_1 contraction [11, 31], L_2 contraction [61, 14], and other stability related properties of the hyperbolic conservation laws [77]. According to the studies, for the Burgers' equation, the initial random perturbation will be unseen in L_2 in long time, if the solution gets shifted by a right amount. In certain cases, such shifts can even be computed explicitly [66, 62]. This indicates the low rank structure of the solution space in the random space upon the correct "shifting". This property is simply not used in the brute-force application of gPC, but should be. Technically, such "shifting" is related to physical quantities such as shock location and shock emerging time. To explicitly trace their evolution, we employ the hodograph transformation method that flips the coordinate and the solution, representing x as a function of u. Such method reduces the nonlinear conservation law problems to linear problems, and the dynamics of physical quantities can be computed through a set of ODEs, which are much easier to handle. We note that the

hodograph method, aside from its application on most 1D evolutional problems, could also be used to treat 2D stationary isentropic gas dynamics, and we leave that for future research.

We should emphasize that the main goal of this chapter is not only to justify the usage of gPC on the Burgers' equation. In fact, except some very special cases such as WENO type methods, almost all numerical methods require certain regularity of the solution, and they simply cannot be applied to hyperbolic conservation laws at the emergence of the shock. We do want to mention that there have been efforts made on finding the correct quantification of the regularity. In [78, 79], the authors proved the wellposedness of entropy solution when randomness is present in initial data and flux, and L_1 contraction is used for estimating the error from interpolation method. In [87], the authors introduced entropic variable and expand the solution as polynomials of the new variable with the understanding that it is smoother when represented by the new variable. Other approaches include [1, 33] where authors either employed the so-termed truncate-encode framework, or in [22] where kinetic formulation is utilized. In this chapter, we aim at providing a different perspective to look at the solution space, and we emphasize on the physical quantities instead of the solution profile. This allows us to single out the quantities that are smooth and can be computed using standard numerical methods.

5.2 Burgers' equation – deterministic case

In this section, we collect some understanding of the shock formation of the Burgers' equation, as a preparation for introducing random variables later.

Consider the inviscid Burgers' equation with known fixed initial data at two infinite ends:

$$\begin{cases} \partial_t u + \partial_x(\frac{1}{2}u^2) = 0, \\ \lim_{x\to\pm\infty} u_{\mathrm{i}}(x) = \mp 1\,. \end{cases} \tag{5.2}$$

Here u_{i} is the initial data. For simplicity, we assume [68]

Assumption 5.2.1. *on the initial data:*

- u_i *monotonically decreases in* x*: i.e.* $u_i'(x) < 0$ *for all* x*.*

- *u_i has a unique inflection point (x^*, u^*), meaning $u_i(x^*) = u_i^*$ and $u_i''(x^*) = 0$.*
- $u_i'''(x^*) > 0$.

With the standard conservation law technique, the solution keeps being constants along characteristics:

$$u(t, x(t)) = u_{\mathrm{i}}(x_0)\,, \quad \text{on} \quad x(t) = x_0 + u_{\mathrm{i}}(x_0)t\,, \tag{5.3}$$

and shocks form when two characteristics collide, meaning:

$$x(t) = x_1 + u_{\mathrm{i}}(x_1)t = x_2 + u_{\mathrm{i}}(x_2)t \quad \Rightarrow \quad t^* = -\frac{x_2 - x_1}{u_{\mathrm{i}}(x_2) - u_{\mathrm{i}}(x_1)}\,.$$

The earliest time for the shock to form, i.e. the smallest possible t^* is achieved by the smallest $u_{\mathrm{i}}'(x)$ (or biggest in absolute value), which is given by the inflection point (x^*, u^*).

5.2.1 Reformulation of the Burgers' equation

The monotonicity of u on x for all time ensures that one could flip $x - u$ coordinate to $u - x$ coordinate and write down the inverse function. We reformulate the equation in this section and write x as a function of u. Let $x = x(t, u)$ be the inverse function of u, then the domain is $[0, \infty) \times (-1, 1)$.

- **Before the formation of the shock**

  Here, since $x(u)$ is the coordinate that sits on u-level set, it propagates with speed u, meaning before t^* when no shocks have been formed, the equation writes:

$$\begin{cases} \partial_t x(t, u) = u\,, \quad u \in (-1, 1), \\ x_{\mathrm{i}}(u) = x(t = 0, u). \end{cases} \tag{5.4}$$

  Take ∂_u of (5.4), we also have:

$$\partial_t \partial_u x(t, u) = 1\,, \quad \Rightarrow \quad \partial_u x(t, u) = x_{\mathrm{i}}'(u) + t\,. \tag{5.5}$$

  According to Assumption 5.2.1, we have, before t^*:

- $\partial_u x(t,u) \leq 0$ for all time, meaning $x_{\mathrm{i}}'(u) + t \leq 0$ (this also means that $t^* = \min(-x_{\mathrm{i}}'(u))$ is the earliest time for a shock emergence, which is achieved at $-x_{\mathrm{i}}'(u^*)$);
- x_{i} has one unique inflection point at (u^*, x^*);
- $(x_{\mathrm{i}})'''(u^*) < 0$.

- **After the formation of the shock**

  However, at $t^* = -x_{\mathrm{i}}'(u^*)$, a shock forms. Just as the strong solution to (5.2) breaks down, equation (5.4) no longer correctly characterizes the solution behavior. The weak solution to (5.2) needs to satisfy the entropy condition, and thus was shown that the shock propagates with the average speed of the two sides. Denote u_1 and u_2 the top and the bottom of the shock, the shock speed is governed by the Rankine-Hugoniot jump condition

$$s = \frac{u_1(t)^2/2 - u_2(t)^2/2}{u_1(t) - u_2(t)} = \frac{u_1(t) + u_2(t)}{2}\,. \tag{5.6}$$

  Such adjust is reflected on the equation for $x(u)$ as well. A shock on $u(t,x)$ plane translates to a "flat" region on $x(t,u)$ plane. When regarded as a function of u, x stays as a constant between u_1 and u_2. While $u_1(t)$ expands to the right and $u_2(t)$ expands to the left as time propagates, the constant changes with speed s in (5.6), meaning the "flat" region moves horizontally with speed s. More specifically, if we denote x^c the height of the flat region, then:

$$\frac{\mathrm{d}}{\mathrm{d}t}x^c = \frac{u_1(t) + u_2(t)}{2}\,, \quad \text{with} \quad x^c(t^*) = x^*\,. \tag{5.7}$$

  We now study the ODE system for $u_{1,2}(t)$.

  In the neighborhood of u_1, the flat region propagates with speed s, while x propagates with speed u_1, and thus in δt,

$$\delta x = (u_1 - s)\delta t = \frac{u_1(t) - u_2(t)}{2}\delta t,$$

is absorbed into the shock. The corresponding δu is:

$$\delta u = -\partial_x u(t,x)|_{u=u_1} \frac{u_1(t)-u_2(t)}{2}\delta t = (-\partial_u x(t,u_1))^{-1}\frac{u_2(t)-u_1(t)}{2}\delta t\,. \tag{5.8}$$

Take (5.5) into account one has:

$$\frac{\delta u}{\delta t} = (-x_{\mathrm{i}}'(u_1)-t)^{-1}\frac{u_2(t)-u_1(t)}{2}\,. \tag{5.9}$$

Perform the same analysis on u_2 and denote $f(u) = -x_{\mathrm{i}}'(u)$, we derive:

$$\begin{aligned}\frac{\mathrm{d}u_1}{\mathrm{d}t} &= F_1(u_1,u_2) = \frac{1}{2}(u_1-u_2)(f(u_1)-t)^{-1},\\ \frac{\mathrm{d}u_2}{\mathrm{d}t} &= F_2(u_1,u_2) = -\frac{1}{2}(u_1-u_2)(f(u_2)-t)^{-1},\end{aligned} \tag{5.10}$$

with the initial condition

$$u_1(t^*) = u_2(t^*) = u^*\,. \tag{5.11}$$

Here we use $F_{1,2}$ to denote the forcing terms for $u_{1/2}$ respectively, and f purely depends on the initial data.

Note that by (5.5) and the monotonicity of $\partial_u x$, we see $-f(u_{1,2})+t = x_{\mathrm{i}}'(u_{1,2})+t \le 0$. Combined with (5.10), it is shown that $u_1(t)$ monotonically increases in time and $u_2(t)$ monotonically decreases in time. Thus

$$u_1 \ge u^*, \quad u_2 \le u^*, \quad -x_{\mathrm{i}}'(u_{1,2}) - t \ge 0\,. \tag{5.12}$$

- **Summary** To summarize the reformulation, the Burgers' equation, when writes on $x(u)$ plane, becomes:

$$\begin{cases} t < t^* = -x_{\mathrm{i}}'(u^*): & \text{Equation (5.4)} \\ t > t^*: & \begin{cases}\text{Equation (5.4)} \quad \text{with} \quad u \in (-1,u_2)\cup(u_1,1) \\ \text{Equation (5.7)} \quad \text{with} \quad u \in (u_2,u_1)\end{cases}\end{cases}, \tag{5.13}$$

where u_1 and u_2 are the shock locations satisfying the ODE system (5.10).

5.2.2 Shock behavior in small time

To understand the shock behavior is the key for unraveling the random variable dependence. We thus carefully study (5.10) here. To start, we first assume $u^* = 0$ and $t^* = 0$, then since $t^* = -x_{\mathrm{i}}'(u^*)$ and $u_1(t^*) = u_2(t^*) = u^*$, one has:

$$u_1(0) = u_2(0) = 0\,, \quad \text{and} \quad f(0) = 0\,. \tag{5.14}$$

Physically it means the flat region start forming right at the initial time $t = 0$, at location $u = 0$.

Considering the properties of x_{i}, we summarize the properties of f:

- $f(u) \geq 0$;
- $f'(u) = 0$ at only one point u^*;
- $f'(u^* < 0) < 0$ and $f'(u^* > 0) > 0$ with $f''(u^*) > 0$.

And when the conditions in (5.14) hold, locally around $u = 0$, $f(u)$ behaves like a quadratic function

$$f(u) \sim au^2\,, \quad \text{with} \quad a = \frac{1}{2}f''(0) > 0\,.$$

Note that the wellposedness of the ODE system is not guaranteed. In fact, the two forcing terms F_1 and F_2 in (5.10) are Lipschitz continuous on u_1 and u_2 if $f(u_{1,2}) - t$ are away from 0. In fact, we show in the lemma below that once $f(u_{1,2}) - t > 0$, it keeps being that way:

Lemma 5.1. *Assume $u_{1,2}(t)$ solves (5.10) with $f(u_{1,2}(t_1)) - t_1 > 0$ for some t_1. Then there exists $c > 0$ such that $f(u_{1,2}(t)) - t > c$ for all $t > t_1$.*

Proof.

$$\frac{\mathrm{d}}{\mathrm{d}t}(f(u_1) - t) = \frac{1}{2}(u_1 - u_2)f'(u_1)(f(u_1) - t)^{-1} - 1\,. \tag{5.15}$$

Since u_1 is increasing, u_2 is decreasing, and $f'(u) > 0$ is increasing for $u > u^*$, one has

$$\frac{1}{2}(u_1 - u_2)f'(u_1) \geq [\frac{1}{2}(u_1 - u_2)f'(u_1)]|_{t=t_1} := c_1 > 0, \quad \forall t \geq t_1\,. \tag{5.16}$$

According to the ODE (5.15),

$$f(u_1) - t \geq \min\{c_1/2, (f(u_1) - t)_{t=t_1}\} := c > 0, \quad \forall t \geq t_1 \,. \tag{5.17}$$

The proof for $f(u_2) - t$ is similar. □

However, at $t = 0$, $f(u_{1,2}) - t = f(u^*) - t^* = 0$ and that $F_{1,2}$ are infinitely big, and special treatment is thus needed. Below we focus on the small time behavior and show the existence and the uniqueness when the Lipschitz continuity of $F_{1,2}$ breaks down.

Remark 5.2. *In fact, In small time, u is small too. If we simply set $f(u) = au^2$, the equation* (5.10) *becomes:*

$$\begin{aligned} \frac{\mathrm{d}u_1}{\mathrm{d}t} &= \frac{1}{2}(u_1 - u_2)(au_1^2 - t)^{-1}, \\ \frac{\mathrm{d}u_2}{\mathrm{d}t} &= -\frac{1}{2}(u_1 - u_2)(au_2^2 - t)^{-1}, \end{aligned} \tag{5.18}$$

and one has the explicit solution:

$$u_1 = -u_2 = (3a^{-1}t)^{1/2} \,. \tag{5.19}$$

For general $f(u)$, we need to sandwich it with two quadratic functions bounding from above and below.

We first state our theorem:

Theorem 5.2.1. *Under the above assumptions on f, assume $u_{1,2}(t)$ solve the ODE system* (5.10) *with initial condition* (5.14) *and satisfy $u_1 > 0$, $u_2 < 0$, $f(u_{1,2}) - t > 0$ for $t > 0$. Then, for any $\epsilon > 0$, there holds*

$$(3(a+\epsilon)^{-1}t)^{1/2} \leq u_1 \leq (3(a-\epsilon)^{-1}t)^{1/2}, \quad -(3(a-\epsilon)^{-1}t)^{1/2} \leq u_2 \leq -(3(a+\epsilon)^{-1}t)^{1/2}, \tag{5.20}$$

for t small enough.

Before proving the theorem, we first study the following ODE:

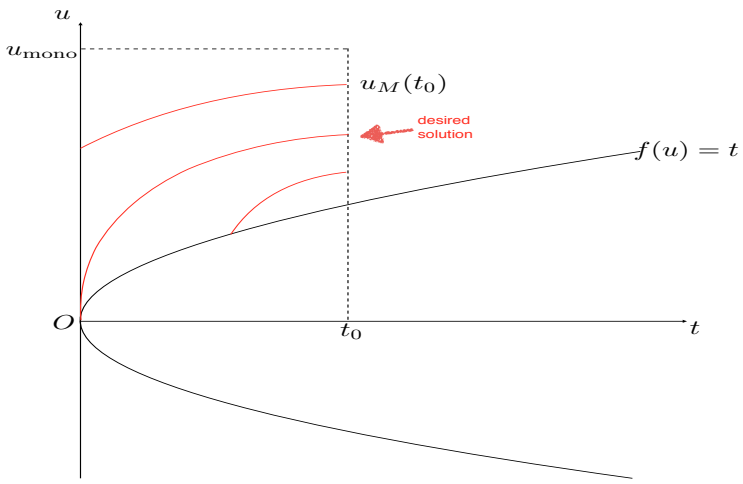

Figure 5.1 Proof of Lemma 5.3

Lemma 5.3. *The ODE*

$$\frac{\mathrm{d}u}{\mathrm{d}t} = u(f(u) - t)^{-1}, \tag{5.21}$$

has a unique solution which satisfies the initial condition $u(0) = 0$ *and the constraint* $u(t) > 0, f(u) > t, \ \forall t > 0$.

Proof. The wellposedness is not obvious only if the solution touches $f(u) = t$ line, making the forcing term blowing up. So to show the wellposedness we fix an arbitrary $t_0 > 0$ and an arbitrary u_0 with $f(u_0) > t_0$, and compute the equation backwards in time. Denote the solution $u(t; u_0)$. We can show if u_0 is set too large, the solution $u(0; u_0) > 0$, while if u_0 is too small the solution $u(t; u_0)$ will hit the $f(u) = t$ for positive t. There is a unique u_0 such that solving the equation backwards give a solution that passes through $(0, 0)$. More specifically we make the following three statements (see Figure 5.1 for demonstration):

1 The solution, computing backwards in time, monotonically depends on u_0, meaning, there exists a small enough positive number u_{mono} such that:

$$u(t; u_0) - u(t; u_1) \geq u_0 - u_1 \geq 0\,, \quad \forall t < t_0, \tag{5.22}$$

for all $u_{0,1} < u_{\text{mono}}$ and $f(u_1) - t_0 > 0$. To show this, we take the derivative with respect to u of the forcing term $u(f(u) - t)^{-1}$.

$$\frac{\mathrm{d}}{\mathrm{d}u}[u(f(u) - t)^{-1}] = (f(u) - t)^{-2}(f(u) - t - uf'(u)) \,. \tag{5.23}$$

With expansion $f(u) = au^2 + O(u^3)$, $uf'(u) = u(2au + O(u^2)) = 2au^2 + O(u^3)$, it is seen that this quantity is negative, for u is small enough, meaning smaller u leads to bigger slope, and (5.22) is shown.

2 If t_0 is small enough, then there exists $u_M = u_M(t_0) > 0$ with $u_M < u_{\text{mono}}$ such that computing (5.21) backwards in time with initial data $u(t_0) = u_M$ gives a solution entirely above the $f(u) = t$ line, and thus the forcing term in (5.21) is always well-defined and is Lipschitz, up to $t = 0$. The solution therefore exists and is unique in $t \in [0, t_0]$, with $u(t = 0; u_0) > 0$, and it continuously depends on u_0.

To show this, we take a positive number $u_{M1} < u_{\text{mono}}/2$ and evolve (5.21) forward with $u(0) = u_{M1}$. Then for t_0 small enough, there holds $u(t_0) < u_{\text{mono}}$. This t_0 together with $u_M = u(t_0)$ are what we desired.

3 If u_0 is chosen to be too close to $f(u) = t$ line, then the slope is given by a large number, meaning computing (5.21) backwards in time gives blow-up solution with u hitting $f(u) = t$ for some $t < t_0$;

Items 1 and 2 combined means with small enough t_0 and u_0, $u(t = 0; u_0)$ monotonically depends on u_0 and approaches 0. Item 3 guarantees that 0 is reachable since too small of the u_0 hits the $f(u) = t$ line. □

Lemma 5.4. *If $f(u) = f(-u)$ is a symmetric function then $(u, -u)$ solves* (5.10) *if u solves* (5.21)*.*

We omit the proof. It is easily obtained by symmetry.

Lemma 5.5. *Let (u_1, u_2) solves* (5.10) *with initial condition* (5.14)*, and $(v, -v)$ solves* (5.10) *with initial condition* (5.14) *where f is replaced by an even function g ($g(u) = g(-u)$). If*

$f(u) < g(u)$ for all u, then $u_1 \geq v, u_2 \leq -v$ for t small enough. If $f(u) > g(u)$ for all u, then $u_1 \leq v, u_2 \geq -v$ for t small enough.

Proof. We prove the first statement by contradiction. Suppose it is not true, then there exists a $t_0 > 0$ small enough such that $u_1(t_0) < v(t_0)$ or $u_2(t_0) > -v(t_0)$. Without loss of generality, we assume the former case, and that $-u_2(t_0) \geq u_1(t_0)$. From the previous lemma, there exists a g-solution $(v_1, -v_1)$ with $v_1(t_0) > u_1(t_0)$, and this solution hits $g(v_1) - t = 0$ line before $t = 0$, meaning there is $t_1 > 0$ such that:

$$f\left(v_1 - \int_{t_1}^{t_0} F_1^g(v_1, -v_1)\,\mathrm{d}s\right) = t_1\,.$$

Here F_1^g is the forcing term for v_1 defined by g.

We claim that there is no $t \leq t_0$ such that $u_1(t) \geq v_1(t), u_2(t) \leq -v_1(t)$. In fact, this is not true at $t = t_0$. Suppose t_2 is the largest time such that this holds, then without loss of generality, we assume that $u_1(t_2) = v_1(t_2)$. Then we have:

$$\begin{aligned} F_1^f(u_1(t_2), u_2(t_2)) &= \frac{1}{2}(u_1(t_2) - u_2(t_2))(f(u_1(t_2)) - t_2)^{-1} \geq \frac{1}{2}(v_1(t_2) + v_1(t_2))(f(v_1(t_2)) - t_2)^{-1} \\ &> \frac{1}{2}(v_1(t_2) + v_1(t_2))(g(v_1(t_2)) - t_2)^{-1} = F_1^g(v_1(t_2), -v_1(t_2)). \end{aligned}$$

This contradicts the choice of t_2. See Figure 5.2 (left) for an illustration.

This claim contradicts the fact that (u_1, u_2) can be continued to the time t_1, since at this time, $f(u_1) - t < g(u_1) - t \leq g(v_1) - t = 0$ or $f(u_2) - t < g(u_2) - t \leq g(-v_1) - t = 0$. Thus the first statement is proved. See Figure 5.2 (right) for an illustration. The second statement can be proved similarly. □

We now turn to show Theorem 5.2.1.

Proof. Locally at $t = 0$, $f(u) \sim au^2$ and thus we approximate (5.10) by:

$$\begin{aligned} \frac{\mathrm{d}u_1}{\mathrm{d}t} &= \frac{1}{2}(u_1 - u_2)(au_1^2 - t)^{-1}, \\ \frac{\mathrm{d}u_2}{\mathrm{d}t} &= -\frac{1}{2}(u_1 - u_2)(au_2^2 - t)^{-1}. \end{aligned} \tag{5.24}$$

Then one can see that

$$u_1 = -u_2 = (3a^{-1}t)^{1/2}\,. \tag{5.25}$$

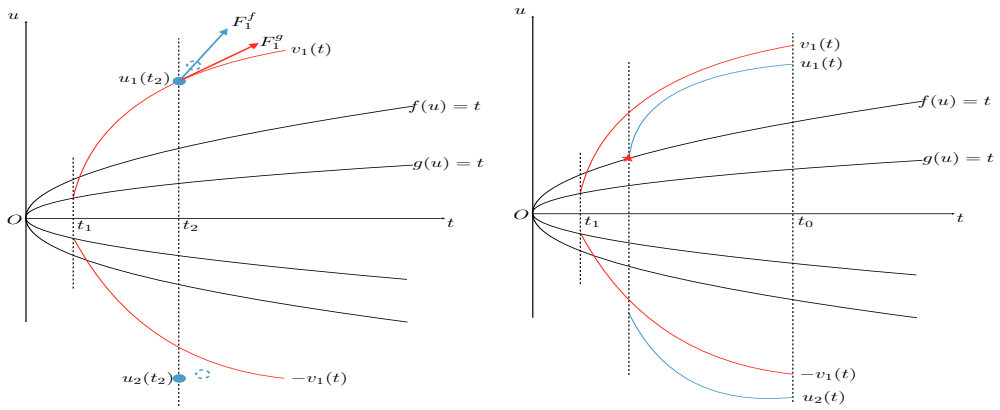

Figure 5.2 Proof of Lemma 5.5. Left: obtaining a contradiction at time t_2. The dashed circle is the approximate positions of $u_{1,2}$ at time slightly larger than t_2, which contradicts the choice of t_2. Right: obtaining the final contradiction. u_1 or u_2 must touches the curve $f(u) = t$ at some time larger than t_1 (the star in the picture).

solves (5.24)(5.14). To estimate the solution of (5.10) near $t = 0$, we assume that $\epsilon < a$, and then there exists δ such that

$$|f(u, z) - a(z)u^2| < \epsilon u^2, \quad \forall |u| < \delta, \forall z. \tag{5.26}$$

Then by Lemma 5.5 it is clear that for sufficiently small t, i.e., t such that $u_1^- < \delta, -u_2^- < \delta$,

$$u_1^+ \leq u_1 \leq u_1^-, u_2^- \leq u_2 \leq u_2^+, \tag{5.27}$$

where $u_i^\pm$ is the solution of (5.24) with a replaced by $a \pm \epsilon$, i.e., (5.25) with this replacement.

□

Remark 5.6. *We note that the initial condition imposed in* (5.14) *is not the reason for having singularities. If we do not shift the emergence time and location of the shock, the singularity will simply appears at* $(t = t^*, u_1 = u^*, u_2 = u^*)$*, with* $f(u^*) = t^*$*. We imposed the condition* (5.14) *not to induce new singularities, but to move the singularity from* (t^*, u^*, u^*) *to* $(0, 0, 0)$*.*

5.3 Random variable dependence

With the shock behavior in hand, we are ready to study the random variable's influence. As stated in Introduction, we are interested mainly in showing the regularity of physical quantities, namely if the initial data $f(u)$ depends smoothly on z, the random variable, then both u_1 and u_2 that define the shock height, and x^c that determines the shock location depend smoothly on z also. In the end we will show:

Theorem 5.3.1. *Assume* u *solves* (5.2)*, with smooth initial data satisfies the convexity conditions. We assume that* $t^*(z)$ *is bounded in* z*, and that there exists* $\delta > 0$ *such that* $-1+\delta \leq u^*(z) \leq 1-\delta$ *for all* z*. Then for* (t, z) *with* $t - t^*(z)$ *small enough, the* z*-derivatives of the physical quantities are bounded:*

$$\partial_z^k u_{1,2} = P_k(D^k u^*, D^k t^*) + \mathcal{O}((t - t^*)^{1/2}), \quad \partial_z^k x^c = P_k(D^k x^*, D^k t^*) + \mathcal{O}((t - t^*)^{3/2}),$$

where P_k *is a polynomial, and* $D^k = (\partial_z^0, \partial_z^1, \ldots, \partial_z^k)$*.*

For the sake of conciseness, in this chapter we only study regularity in one random variable, and will leave higher dimensional case to future studies. As a preparation we first study the case where the initial conditions in (5.14) are satisfied, and then perform the shifting in both t and u for the proof of the theorem above.

Proposition 5.7. *Consider* (5.10) *with initial condition* (5.14). *Suppose the initial profile represented by* $f(u)$ *has smooth* z*-dependence where* z *is the random variable, i.e.,* $f(u;z) \in C^k(u,z)$. *Then for* t *small enough, the* z*-derivatives of* u_1, u_2 *satisfy the estimate*

$$\partial_z u_{1,2} = \mathcal{O}(t^{1/2})\,.$$

Proof. We take z-derivatives of the system (5.10):

$$\begin{aligned}
\frac{\mathrm{d}\partial_z u_1}{\mathrm{d}t} &= \frac{1}{2}(\partial_z u_1 - \partial_z u_2)(f(u_1)-t)^{-1} - \frac{1}{2}(u_1-u_2)(f(u_1)-t)^{-2}(f'(u_1)\partial_z u_1 + \partial_z f(u_1)),\\
\frac{\mathrm{d}\partial_z u_2}{\mathrm{d}t} &= -\frac{1}{2}(\partial_z u_1 - \partial_z u_2)(f(u_2)-t)^{-1} + \frac{1}{2}(u_1-u_2)(f(u_2)-t)^{-2}(f'(u_2)\partial_z u_2 + \partial_z f(u_2)).
\end{aligned}$$

or in a compact form:

$$\begin{aligned}
\frac{\mathrm{d}\partial_z u_1}{\mathrm{d}t} &= A_{11}\partial_z u_1 + A_{12}\partial_z u_2 + S_1,\\
\frac{\mathrm{d}\partial_z u_2}{\mathrm{d}t} &= A_{21}\partial_z u_1 + A_{22}\partial_z u_2 + S_2,
\end{aligned} \tag{5.28}$$

with initial data

$$\partial_z u_1(0) = \partial_z u_2(0) = 0.$$

By Theorem 5.2.1, $u_{1,2}(t) \approx \pm(3a^{-1}t)^{1/2}$. Thus we have the estimates:

$$A_{11} = \frac{1}{2}(f(u_1)-t)^{-1} - \frac{1}{2}(u_1-u_2)(f(u_1)-t)^{-2}f'(u_1) \approx -\frac{5}{4}t^{-1}\,,$$

and

$$A_{12} = -\frac{1}{2}(f(u_1)-t)^{-1} \approx -\frac{1}{4}t^{-1}\,.$$

Similarly

$$A_{21} \approx -\frac{1}{4}t^{-1}, \quad A_{22} \approx -\frac{5}{4}t^{-1}\,.$$

Here the approximation $\approx$ sign means: if $A(t) \approx B(t)$ then $\lim_{t\to 0+} \frac{A(t)}{B(t)} = 1$. For the source terms, one uses $f(u;z) \sim a(z)u^2$ to obtain $\partial_z f \sim \partial_z a u^2$. Therefore:

$$S_1 = -\frac{1}{2}(u_1 - u_2)(f(u_1) - t)^{-2}\partial_z f(u_1) = O(t^{-1/2}), \quad S_2 = O(t^{-1/2})\,. \tag{5.29}$$

With these estimates, we perform L^2 type estimate for (5.28). We multiply (5.28) on both sides with $\partial_z u_{1,2}$ and incorporate the estimates from above. For a small positive ϵ, and small t, one has:

$$\begin{aligned}
\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}(\partial_z u_1)^2 &\le -(\frac{5}{4} - \epsilon)t^{-1}(\partial_z u_1)^2 + (\frac{1}{4} + \epsilon)t^{-1}|\partial_z u_1 \partial_z u_2| + Ct^{-1/2}|\partial_z u_1|\,,\\
\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}(\partial_z u_2)^2 &\le -(\frac{5}{4} - \epsilon)t^{-1}(\partial_z u_2)^2 + (\frac{1}{4} + \epsilon)t^{-1}|\partial_z u_1 \partial_z u_2| + Ct^{-1/2}|\partial_z u_2|\,.
\end{aligned}$$

Add the two inequalities and use the fact that

$$|\partial_z u_1 \partial_z u_2| \le \frac{1}{2}((\partial_z u_1)^2 + (\partial_z u_2)^2)\,, \quad \text{and} \quad t^{-1/2}|\partial_z u_1| \le \epsilon_1 t^{-1}(\partial_z u_1)^2 + \frac{1}{4\epsilon_1}, \quad \forall \epsilon_1 > 0\,,$$

one gets:

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}((\partial_z u_1)^2 + (\partial_z u_2)^2) \le -\left(\frac{5}{4} - \epsilon - (\frac{1}{4} + \epsilon) - C\epsilon_1\right) t^{-1}((\partial_z u_1)^2 + (\partial_z u_2)^2) + \frac{C}{2\epsilon_1}\,.$$

Choosing $\epsilon = 1/4, \quad \epsilon_1 = 1/(2C)$, one gets

$$\frac{\mathrm{d}}{\mathrm{d}t}((\partial_z u_1)^2 + (\partial_z u_2)^2) \le 2C^2\,,$$

which finishes the proof. □

To extend to higher derivatives, one conducts induction and the computation is more involved, but the conclusion still holds true.

Proposition 5.8. *Consider* (5.10) *with initial condition* (5.14). *With the same conditions as in Proposition 5.7, the higher z-derivatives of* u_1, u_2 *still satisfy:*

$$\partial_z^k u_{1,2} = \mathcal{O}(t^{1/2})\,. \tag{5.30}$$

Proof. We show this by induction. Assume $\partial_z^j u_{1,2} = \mathcal{O}(t^{1/2})$ holds true for all $j < k$, and we show it for k-th derivative.

To start, we first take k-th z-derivative of the ODE system (5.10). It gives:

$$\begin{aligned} \frac{\mathrm{d}\partial_z^k u_1}{\mathrm{d}t} &= A_{11}\partial_z^k u_1 + A_{12}\partial_z^k u_2 + S_1^k, \\ \frac{\mathrm{d}\partial_z^k u_2}{\mathrm{d}t} &= A_{21}\partial_z^k u_1 + A_{22}\partial_z^k u_2 + S_2^k, \end{aligned} \tag{5.31}$$

where A_{mn} are the same as defined in (5.28) $(m, n = 1, 2)$, and thus share the estimations, but the source term S_1^k becomes much more complicated: it is a summation of terms of the form

$$c\partial_z^{r_1}(u_1 - u_2)(f(u_1) - t)^{-(1+r_2)} \prod_{j=1}^{r_2} \partial_z^{r_{3,j}} \partial_u^{r_{4,j}} f(u_1) \prod_{l=1}^{L} \partial_z^{r_{5,l}} u_1 \,, \tag{5.32}$$

where the indices satisfy the relation

$$L = \sum_j r_{4,j} \,,$$

and every $r_{5,l}$ is at most $k-1$. We notice that

$$f(u_1) \sim u_1^2 = \mathcal{O}(t), \quad f'(u_1) \sim u = \mathcal{O}(t^{1/2}), \quad \text{and} \quad \partial_u^r f(u_1) = \mathcal{O}(1),\ r \geq 2\,.$$

As a result, $\partial_u^r f(u_1) \lesssim \mathcal{O}(t^{1-r/2}),\ r \geq 0$. This also holds with f replaced by its z-derivatives. Thus, with the assumptions that $\partial_z^l u_1 = \mathcal{O}(t^{1/2})$ for $l \leq k-1$, the order of the term (5.32) is (in term of the power of t)

$$\frac{1}{2} - (1 + r_2) + \sum_j (1 - \frac{r_{4,j}}{2}) + \frac{L}{2} = -\frac{1}{2}\,.$$

We analyze S_2^k in the same way, and obtain the same L^2 estimate which finished the proof. □

Proposition 5.9. *With the same assumptions as in Theorem 5.8, one has*

$$\partial_z^k x^c = \partial_z^k x^*(z) + \mathcal{O}(t^{3/2})\,.$$

Proof. It follows easily from checking (5.7) which holds true on $t > t^* = 0$ with $x^c(0) = x^*$. Take its z derivative, we have:

$$\frac{\mathrm{d}}{\mathrm{d}t}\partial_z^k x^c = \frac{\partial_z^k u_1 + \partial_z^k u_2}{2} = \mathcal{O}(t^{1/2})\,,$$

which leads to the conclusion. □

With all these preparations for equations with special initial data, we are ready to perform shifting for proving Theorem 5.3.1.

Proof of Theorem 5.3.1. This theorem deals with the equation with initial conditions not satisfying (5.14), and thus both u^* and t^* vary with respect to z. We first shift the system pointwisely in z by $t^*(z)$. Define:

$$\tilde{u}_{1,2}(t,z) = u_{1,2}(t - t^*(z), z)\,, \tag{5.33}$$

then $\tilde{u}_{1,2}$ satisfy the ODE system (5.10) with initial condition:

$$\tilde{f}(u,z) = f(u,z) - t^*(z)u\,.$$

Correspondingly, the physical quantities need adjustments: the shock location x^c gets shifted in time, but u^* is kept the same, meaning:

$$x^c(t,z) = \tilde{x}^c(t^*(z) + t, z)\,, \quad u^*(z) = \tilde{u}^*(z)\,, \quad \text{and} \quad \tilde{t}^*(z) = 0\,. \tag{5.34}$$

We then shift the systems by u^*. Define:

$$\bar{u}_{1,2}(t,z) = \tilde{u}_{1,2}(t,z) - u^*(z)\,, \tag{5.35}$$

then $\bar{u}$ satisfy (5.10) with initial condition:

$$\bar{f}(u,z) = \tilde{f}(u + u^*, z)\,.$$

Correspondingly, the physical quantities are adjusted:

$$\bar{x}^c(z) = \tilde{x}^c(z)\,, \quad \bar{u}^* = \tilde{u}^* - u^* = 0\,, \quad \text{and} \quad \bar{t}^*(z) = \tilde{t}^*(z) = 0\,. \tag{5.36}$$

Given the assumption that there exists $\delta > 0$ such that $-1 + \delta \le u^*(z) \le 1 - \delta$ for all z, we have $\bar{f}$ well-defined for $u \in [-\delta, \delta]$.

With these shiftings, $\bar{u}$ satisfies the conditions in Proposition 5.8 and 5.9 and thus:

$$\partial_z^k \bar{u}_{1,2} = \mathcal{O}(t^{1/2})\,, \quad \text{and} \quad \partial_z^k \bar{x}^c = \mathcal{O}(t^{3/2})\,.$$

Then according to (5.35) and (5.36):

$$\partial_z^k \tilde{u}_{1,2} = \partial_z^k \bar{u}_{1,2} + \partial_z^k u^* = \partial_z^k u^* + \mathcal{O}(t^{1/2}), \quad \text{and} \quad \partial_z^k \tilde{x}^c = \partial_z^k \bar{x}^c = \mathcal{O}(t^{3/2}).$$

Plugging it back in (5.33) and (5.34), one gets:

$$\partial_z^k u_{1,2} = \partial_z^k \tilde{u}_{1,2}(t - t^*(z), z), \quad \text{and} \quad \partial_z^k x^c = \partial_z^k \tilde{x}^c(t - t^*(z)),$$

which concludes the theorem. □

5.4 Numerical methods

According to Theorem 5.3.1, the physical quantities of the Burger's equation enjoy high order regularity in the random space. This suggests that polynomial interpolation is still feasible if done properly. We propose two interpolation methods in this section, both of which are based on the regularities proved before.

To proceed, we denote $\{z_j\}_{j=1}^N$ a set of quadrature points. Following the gPC idea, in the offline stage, we first compute all solutions on these quadrature points and obtain $u(t, x, z_j)$ for all $j = 1, \cdots, N$. The goal is to utilize these solutions to find a good approximation to $u(t, x, z_0)$ for arbitrary given z_0 in the online stage through interpolation.

5.4.1 Method 1: the x-transformed interpolation

In the first method we propose, we shift all solutions in x by x^c, the shock location, so that solutions with different z develop shocks at the same location. Let

$$\tilde{u}(t, x, z) = u(t, x + x^c(t, z), z), \tag{5.37}$$

be the shifted solution, then its polynomial interpolation is defined as:

$$\tilde{u}_N(t, x, z) = \sum_{j=1}^{N} l_j(z) \tilde{u}(t, x, z_j), \tag{5.38}$$

where $l_j(z)$ is the Lagrange polynomial associated with z_j. From Proposition 5.9, $x^c \in C^k(z)$, and at the same time, it is a standard practice that $u(t, x, z)$ smoothly depends on

z away from the shock too. We show below that $\tilde{u}(t,x,z) \in C^k(z)$, permitting an accurate polynomial interpolation:

Proposition 5.10. *The shifted $\tilde{u}$ (defined in* (5.37)*) has continuous dependence on z away from the shock. To be precise,*

$$|\partial_z^k \tilde{u}(t,x_0,z)| \le \frac{C}{|x_0|}. \tag{5.39}$$

Proof. We only show the boundedness of $\partial_z \tilde{u}$ and that of higher order derivatives are similar. For all $x_0 \neq 0$, we take z-derivative of (5.37):

$$\begin{aligned}
\partial_z \tilde{u}(t,x_0,z) &= \partial_z u(t, x_0 + x^c(t,z), z) + \partial_x u(t, x_0 + x^c(t,z), z)\partial_z x^c(t,z) \\
&= \partial_x u(t, x_0 + x^c(t,z), z)\left[\partial_z x(t,u_0,z) + \partial_z x^c(t,z)\right] \\
&= \partial_x u(t, x_0 + x^c(t,z), z)\left[\partial_z (x_{\mathrm{i}}(u_0,z) + u_0 t) + \partial_z x^c(t,z)\right] \\
&= \partial_x u(t, x_0 + x^c(t,z), z)\left[\partial_z x_{\mathrm{i}}(u_0,z) + \partial_z x^c(t,z)\right] ,
\end{aligned} \tag{5.40}$$

where $u_0 = u(t, x_0 + x^c(t,z), z)$, and the third line comes from solving (5.4). We then claim that

$$\left| \frac{1}{\partial_x u(t, x_0 + x^c(t,z), z)} \right| = |\partial_u x(t, u_0, z)| \ge \frac{|x_0|}{2} \tag{5.41}$$

In fact, suppose $x_0 > 0$, then

$$u(t, x_0 + x^c(t,z), z) - u(t, x^c(t,z), z) = \int_{x^c(t,z)}^{x_0 + x^c(t,z)} \partial_x u(t,y,z)\,\mathrm{d}y \le x_0 \partial_x u(t, x_0 + x^c(t,z), z)\,,$$

where the inequality is because $u(t,x,z)$ is convex in x for $x > x^c(t,z)$, and thus $\partial_x u(t,y,z) \le \partial_x u(t, x_0 + x^c(t,z), z)$. Taking absolute value, noticing that $\partial_x u < 0$, we get

$$\begin{aligned}
|x_0 \partial_x u(t, x_0 + x^c(t,z), z)| \le& |u(t, x_0 + x^c(t,z), z) - u(t, x^c(t,z), z)| \\
\le& |u(t, x_0 + x^c(t,z), z)| + |u(t, x^c(t,z), z)| \le 2,
\end{aligned}$$

and the claim is proved. The case $x_0 < 0$ is similar.

Combining (5.40) and (5.41), we get the constant bound:

$$|\partial_z \tilde{u}(t,x_0,z)| \le \frac{2}{|x_0|}\left[|\partial_z x_i(u_0,z)| + |\partial_z x^c(t,z)|\right], \tag{5.42}$$

away from the shock. □

Note that by the definition of (5.37), the solution gets shifted in x by x^c, the shock location. It implicitly requires that at time t, the solutions for all z have the shock developed already. Therefore this method only works for t big enough so that $t > t^*(z), \forall z$. For smaller t, we use Method 2.

5.4.2 Method 2: the (x,t)-transformed interpolation

The second method is relatively more subtle. We perform shifting in both x and t. We first interpolate $t^*(z)$ by:

$$t^*_N(z) = \sum_{j=1}^{N} l_j(z) t^*(z_j)\,,$$

and then define:

$$\bar{u}(t,x,z) = u(t + t^*_N(z), x + x^c(t + t^*_N(z), z), z),$$

and interpolate it using:

$$\bar{u}_N(t,x,z) = \sum_{j=1}^{N} l_j(z)\bar{u}(t,x,z_j)\,.$$

For example to obtain $u(t,x,z_0)$, we have:

$$u(t,x,z_0) = \sum_{j=1}^{N} l_j(z_0) u(t + t^*_N(z_j), x + x^c(t + t^*_N(z_j), z_j), z_j)\,.$$

5.4.3 Comments on numerical error

In practice one is given profile of u for all collocation points z_j, and needs to find, for all these collocation points, the evaluation of x^c and t^*, with which interpolates for their values for arbitrary z_0. Two parts of error incur: finding the evaluation of x^c and t^* numerically for all z_j, and use these values for interpolating their values at z_0.

To numerically evaluating x^c and t^*, we simply take numerical differentiation. For obtaining x^c, we take

$$k^* \approx \text{argmax}_k |u(t, x_{k+1}, z) - u(t, x_{k-1}, z)|\,,$$

and we call $x^c = x_k$. Here we do introduce numerical error of $\mathcal{O}(\Delta x)$. t^* is the first time for the shock to appear. We thus define

$$t^* \approx \text{argmin}_t[\max_k \frac{|u(t, x_{k+1}, z) - u(t, x_{k-1}, z)|}{2\Delta x} > C] \,,$$

with C set as a big threshold. Naturally it gives $\mathcal{O}(\Delta t)$ error.

These errors would propagate into numerical interpolation. Take interpolating x^c for example. Since the interpolation formula writes:

$$x^c(z) \sim \sum_j (\sum_k x^c(z_k)\phi_j(z_k)w_k)\phi_j(z) \,,$$

where ϕ_j is a set of orthonormal polynomials (with respect to some probability measure), and (z_k, w_k) is the corresponding Gauss quadrature location and weight, perturbing every $x^c(z_k)$ by $\mathcal{O}(\Delta x)$ means that $x^c(z)$ is perturbed by at most

$$\mathcal{O}(\Delta x) \sum_{j,k} |\phi_j(z_k)|w_k|\phi_j(z)| \leq \mathcal{O}(\Delta x) \sum_j |\phi_j(z)| \,.$$

Suppose one uses Chebyshev polynomials $\phi_j(z) \leq 1$, then the error would be at most $\mathcal{O}(N\Delta x)$.

5.5 Numerical results

In this section we demonstrate the efficiency of the method numerically. We solve the Burgers equation (5.2) with initial data

$$u_{\text{i}}(x) = v(x) - 0.2(v(x) + 0.5)(1 - v(x)^2)z, \quad \text{with} \quad v(x) = \frac{1 - e^{x-3z^3}}{1 + e^{x-3z^3}} \,, \tag{5.43}$$

and the random variable $z \in [-1, 1]$. This initial data, for every z, satisfies the assumptions listed below (5.2). We truncate the x-domain to $[-R, R]$ with $R = 15$ and discretize the domain by $N_x = 1500$ grid points. The equation is computed using 5th order finite difference WENO scheme with the 3rd order SSP Runge-Kutta in time. We choose to use very accurate solver in x and t space to minimize the error induced by the discretization in physical domain.

5.5.1 Accuracy of the interpolation methods

On the random space, we sample $N_z = 10$ Chebyshev quadrature points

$$z_j = \cos\left(\frac{2j-1}{2N_z}\pi\right), \quad j = 1, \ldots, N_z\,, \tag{5.44}$$

and we use the two methods proposed in the Section 5.4 for the interpolation. We plot the solution at $z_0 = 0.234$ at time $T = 2.2$. In Figure 5.3, we plot the numerical result using method 1. It is obvious that the x-transformation put all the shocks together, and the interpolation error is small (less than 10^{-3}) apart from the a few grid points near the shock. The results using method 2 achieve similar accuracy and are shown in Figure 5.4.

For comparison, we give the result of a direct polynomial interpolation in z. The numerical result is shown in Figure 5.5. One can clearly see that the two solutions do not agree well, and the error is large (larger than 10^{-2}) inside the interval $[-2.5, 2.5]$. This is precisely the region where most of the shocks are located at (as seen in top left of Figure 5.3).

5.5.2 Regularity of u_1, u_2, x^c

We also show the regularity of u_1, u_2, x^c. To do this, we sample $N_z = 50$ Chebyshev quadrature points

$$z_j = \cos\left(\frac{2j-1}{2N_z}\pi\right), \quad j = 1, \ldots, N_z\,, \tag{5.45}$$

on the random space and then solve the Burgers equation up to $T = 4$. Numerically we find x^c using the method proposed in Section 5.4.3, and we assign u_1 and u_2 the values of u taken in a small neighborhood of x^c (extending to the two sides by 2 grid points). In Figure 5.6 we plot $u_{1,2}\,, x^c$ as functions of both time and the random variable z. They clearly have the regular dependence. We have to note the zigzags at the boundary comes from the $\mathcal{O}(\Delta x)$ error from numerically finding x^c. The empty area for small t means that shocks have not been developed yet.

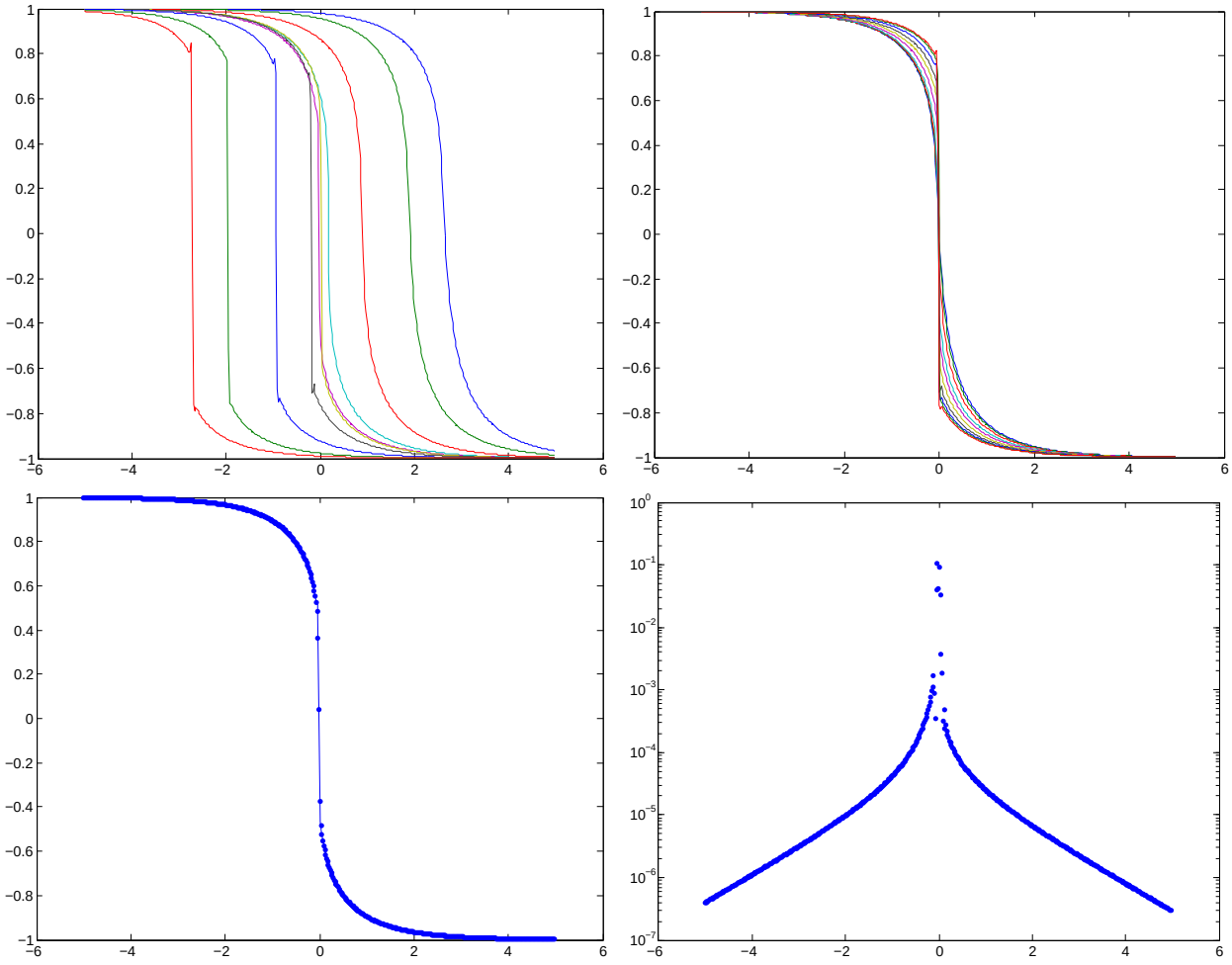

Figure 5.3 Top left: solutions at all $\{z_j\}$; top right: transformed solutions by method 1; bottom left: compare the interpolation solution (by method 1, dots) with the numerical scheme solution (line) at z_0; bottom right: error (difference between the two solutions given in bottom left).

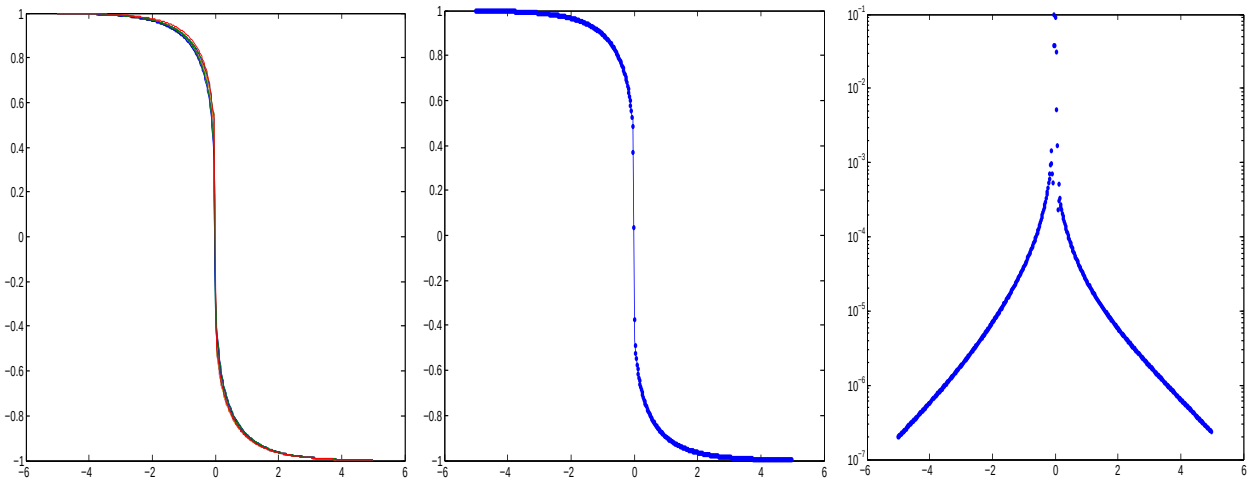

Figure 5.4 Left: transformed solutions by method 2; middle: compare the interpolation solution (by method 2, dots) with the numerical scheme solution (line) at z_0; right: error (difference between the two solutions given in bottom left).

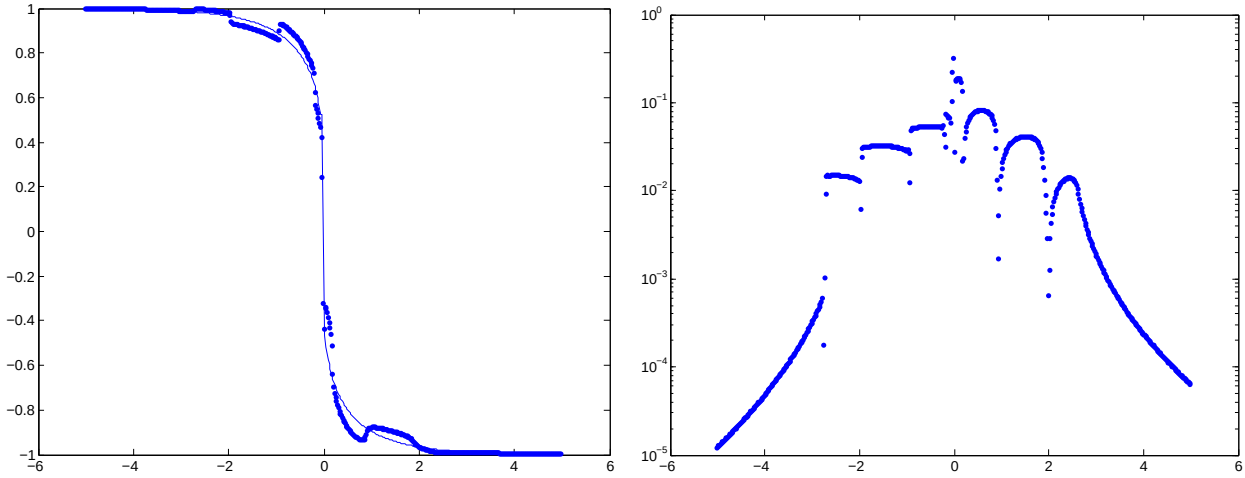

Figure 5.5 Left: compare the interpolation solution (direct, dots) with the numerical scheme solution (line) at z_0; right: error (difference between the two solutions given in bottom left).

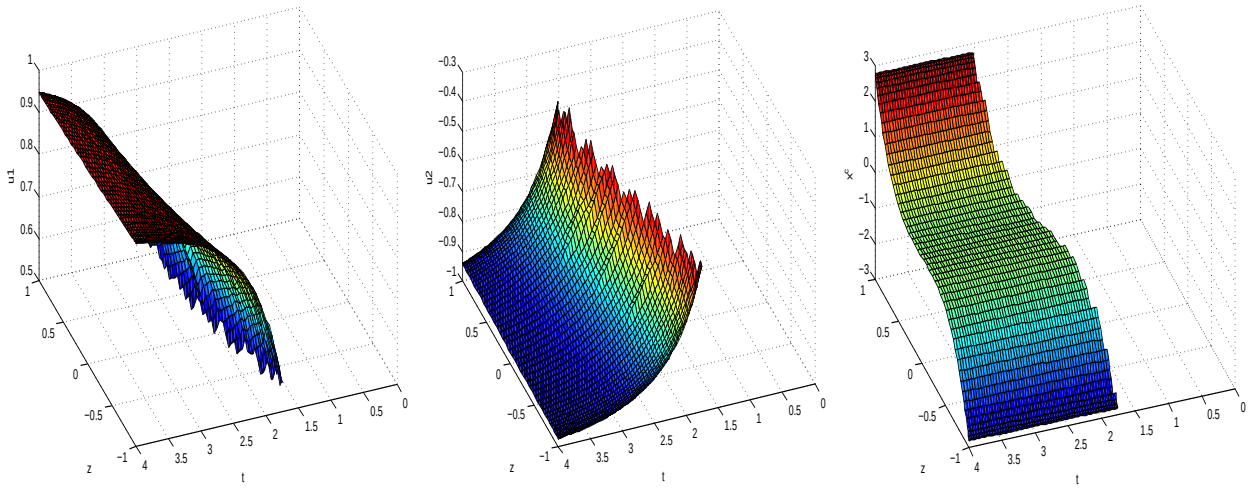

Figure 5.6 Left to right: u_1, u_2, x^c as functions of time and the random variable.

5.6 General convex scalar conservation laws

All the analysis can be extended to study general scalar conservation laws with convex flux term. The equation writes:

$$\begin{cases} \partial_t u + \partial_x F(u) = 0, \\ \lim_{x\to\pm\infty} u_{\mathrm{i}}(x) = \mp 1\,. \end{cases} \tag{5.46}$$

where the flux function F is smooth and strictly convex. As done for the Burgers equation, we assume that the initial data is decreasing, and therefore the inverse function $x(u)$ is well-defined on $(-1,1)$.

We once again flip $x-u$ space to $u-x$ space, and derive the equation for $x(u)$ as a function of u.

5.6.1 Reformulation of the equation

- **Before shocking emergence**

  The solution keeps being constant along characteristic lines. This naturally gives the equation of x as a function of u:

$$\begin{cases} \partial_t x(t,u) = F'(u)\,, \quad u \in (-1,1), \\ x_{\mathrm{i}}(u) = x(t=0,u). \end{cases} \tag{5.47}$$

  Since F is convex, F' is increasing, and thus has a smooth inverse denoted by G, then:

$$G(F'(u)) = F'(G(u)) = u\,. \tag{5.48}$$

  Plugging it back into (5.47) and denoting $y(t,u) = x(t,G(u))$, we have:

$$\begin{cases} \partial_t y(t,u) = \partial_t x(t,G(u)) = u\,, \quad u \in (-1,1), \\ y_{\mathrm{i}}(u) = x(t=0,G(u)) = x_{\mathrm{i}}(G(u)). \end{cases}$$

  As was done for the Burgers' equation, we take one more derivative on u and obtain:

$$\begin{cases} \partial_t \partial_u y(t,u) = 1\,, \quad u \in (-1,1)\,, \\ y'_{\mathrm{i}}(u) = x'_{\mathrm{i}}(G(u))G'(u)\,. \end{cases}$$

Therefore

$$\partial_u y = y_{\mathrm{i}}'(u) + t\,,$$

and equivalently, the earliest shock appears at $t^* = -\min y_{\mathrm{i}}'(u)$ and we assume there is one and only one and set it as:

$$t^* = -\min y_{\mathrm{i}}'(u) = -y_{\mathrm{i}}'(F'(u^*))\,.$$

- **After the emergence of the shock**

  Once the shock appears, on $u - x$ plane, a "flat" region appears. We denote u_1 and u_2 the top and the bottom of the shock point, then on the $u - x$ plane, in (u_2, u_1) the solution is a constant. Away from such "flat" region, we still use (5.47), but the shock itself moves horizontally with speed:

$$s = \frac{F(u_1) - F(u_2)}{u_1 - u_2}\,.$$

  and thus if we call the shock location, or the value for the flat region x^c, it satisfies:

$$\frac{\mathrm{d}}{\mathrm{d}t}x^c = \frac{F(u_1) - F(u_2)}{u_1 - u_2}\,, \quad \text{with} \quad x^c(t^*) = x^*\,. \tag{5.49}$$

  With the same derivation as in Section 5.2.1, one has:

$$\begin{aligned}
\frac{\mathrm{d}u_1}{\mathrm{d}t} &= F_1(u_1, u_2) = \left(F'(u_1) - \frac{F(u_1) - F(u_2)}{u_1 - u_2}\right)(f(u_1) - F''(u_1)t)^{-1}, \\
\frac{\mathrm{d}u_2}{\mathrm{d}t} &= F_2(u_1, u_2) = -\left(\frac{F(u_1) - F(u_2)}{u_1 - u_2} - F'(u_2)\right)(f(u_2) - F''(u_2)t)^{-1},
\end{aligned} \tag{5.50}$$

  with initial condition $u_1(t^*) = u_2(t^*) = u^*$. Here we denoted $f(u) = -x_{\mathrm{i}}'(u)$. Considering F' is an increasing function, we see that

$$\frac{\mathrm{d}u_1}{\mathrm{d}t} > 0 > \frac{\mathrm{d}u_2}{\mathrm{d}t}\,.$$

- **Summary**

To summarize the reformulation, in the general convex flux case, when writes on $x(u)$ plane, x satisfies equation

$$\begin{cases} t < t^* = -y_{\mathrm{i}}'(u^*): & \text{Equation (5.47)} \quad \text{with} \\ t > t^*: & \begin{cases} \text{Equation (5.47)} \quad \text{with} \quad u \in (-1, u_2) \cup (u_1, 1) \\ \text{Equation (5.49)} \quad \text{with} \quad u \in (u_2, u_1) \end{cases} \end{cases} \tag{5.51}$$

with u_1 and u_2 being the shock locations satisfying the ODE system (5.50).

5.6.2 Shock behavior in small time

We assume $t^* = 0$, $u^* = 0$. Then one has $u_1 > 0$, $u_2 < 0$ for $t > 0$. To study the shock time behavior of equation (5.50), we analyze the two forces. Near $u_{1,2} = 0$ we can approximate one term:

$$\begin{aligned} & F'(u_1) - \frac{F(u_1) - F(u_2)}{u_1 - u_2} \\ =& F'(0) + F''(0)u_1 - \frac{F'(0)(u_1 - u_2) + \frac{1}{2}F''(0)(u_1^2 - u_2^2))}{u_1 - u_2} + \mathcal{O}(u_1^2, u_2^2, u_1 u_2) \\ =& \frac{1}{2}F''(0)(u_1 - u_2) + \mathcal{O}(u_1^2, u_2^2, u_1 u_2). \end{aligned} \tag{5.52}$$

Since $-y_{\mathrm{i}}'$ achieves its maximum at $F'(u^*) = F'(0)$, we have $f(G(u))G'(u)$ achieves the minimum at $u = F'(u^*) = F'(0)$. Furthermore, since $t^* = 0$, one has

$$0 = t^* = f(G(F'(u^*))) = f(u^*) = f(0)\,.$$

Thus one has the expansion

$$f(G(u))G'(u) = a_1(u - F'(0))^2 + \mathcal{O}((u - F'(0))^3), \quad a_1 = \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}u^2} f(G(u))G'(u)|_{u=F'(0)} > 0\,.$$

Then, near $u = 0$,

$$\begin{aligned} f(u) =& \frac{f(G(F'(u)))G'(F'(u))}{G'(F'(u))} \\ =& \frac{a(F'(u) - F'(0))^2 + \mathcal{O}((F'(u) - F'(0))^3)}{G'(F'(0)) + \mathcal{O}(u)} \\ =& au^2 + \mathcal{O}(u^3)\,, \end{aligned}$$

where $a = \frac{a_1 F''(0)^2}{G'(F'(0))}$ is a positive number.

Therefore, if we replace the terms in (5.50) by their first order approximations, we get

$$\begin{aligned} \frac{\mathrm{d}u_1}{\mathrm{d}t} &= \frac{1}{2}F''(0)(u_1-u_2)(au_1^2 - F''(0)t)^{-1}, \\ \frac{\mathrm{d}u_2}{\mathrm{d}t} &= -\frac{1}{2}F''(0)(u_1-u_2)(au_2^2 - F''(0)t)^{-1}. \end{aligned} \tag{5.53}$$

It is a scaled version of (5.18) and has the explicit solution

$$u_1 = -u_2 = (ct)^{1/2}, \quad c = \frac{3F''(0)}{a}\,. \tag{5.54}$$

Away from the leading order when $f(u)$ no longer gets approximated by au^2, one needs to perform similar analysis as is done in Theorem 5.2.1 by sandwiching two sides with two quadratic functions. We omit the proof but simply conjecture that Theorem 5.2.1 still holds true for equation (5.50) in the general conservation law setting, with $3a^{-1}$ replaced by c.

5.6.3 Regularities in the random space

We study the solution's regularity in the random space in this section. It is the counterpart of Section 5.3. Due to the complexity of the formula, we only present the first derivative in z of (5.50). Higher derivatives are all similar.

Taking the first derivation of (5.50), we get:

$$\begin{aligned} \frac{\mathrm{d}\partial_z u_1}{\mathrm{d}t} =& \left[F_1''\partial_z u_1 - \frac{F_1'\partial_z u_1 - F_2'\partial_z u_2}{u_1-u_2} + \frac{(F_1-F_2)(\partial_z u_1 - \partial_z u_2)}{(u_1-u_2)^2}\right](f_1 - F_1''t)^{-1} \\ &- \left[F_1' - \frac{F_1-F_2}{u_1-u_2}\right](f_1-F_1''t)^{-2}(f_1'\partial_z u_1 - F_1'''\partial_z u_1 t + \partial_z f_1), \\ \frac{\mathrm{d}\partial_z u_2}{\mathrm{d}t} =& -\left[\frac{F_1'\partial_z u_1 - F_2'\partial_z u_2}{u_1-u_2} - \frac{(F_1-F_2)(\partial_z u_1 - \partial_z u_2)}{(u_1-u_2)^2} - F_1''\partial_z u_1\right](f_2 - F_2''t)^{-1} \\ &+ \left[\frac{F_1-F_2}{u_1-u_2} - F_2'\right](f_2-F_2''t)^{-2}(f_2'\partial_z u_2 - F_2'''\partial_z u_2 t + \partial_z f_2), \end{aligned} \tag{5.55}$$

where we have used $\gamma_{1,2}$ to denote $\gamma(u_1)$ or $\gamma(u_2)$ respectively for all quantities. In a compact form, it writes:

$$\begin{aligned} \frac{\mathrm{d}\partial_z u_1}{\mathrm{d}t} &= A_{11}\partial_z u_1 + A_{12}\partial_z u_2 + S_1, \\ \frac{\mathrm{d}\partial_z u_2}{\mathrm{d}t} &= A_{21}\partial_z u_1 + A_{22}\partial_z u_2 + S_2, \end{aligned}$$

with initial data

$$\partial_z u_1(0) = \partial_z u_2(0) = 0\,.$$

Here:

$$\begin{aligned}
A_{11} &= \left[F_1'' - \frac{F_1'}{u_1-u_2} + \frac{F_1-F_2}{(u_1-u_2)^2}\right](f_1 - F_1''t)^{-1} - \left[F_1' - \frac{F_1-F_2}{u_1-u_2}\right](f_1-F_1''t)^{-2}(f_1'-F_1'''t)\,,\\
A_{12} &= \left[\frac{F_2'}{u_1-u_2} - \frac{F_1-F_2}{(u_1-u_2)^2}\right](f_1-F_1''t)^{-1}\,,\\
S_1 &= -\left[F_1' - \frac{F_1-F_2}{u_1-u_2}\right](f_1-F_1''t)^{-2}\partial_z f_1\,.
\end{aligned}$$

To analyze the term A_{11}, we note that

$$\begin{aligned}
&F_1'' \approx F''(0)\,;\\
&-\frac{F_1'}{u_1-u_2} + \frac{F_1-F_2}{(u_1-u_2)^2} = \frac{1}{u_1-u_2}\left[-F'(u_1) + \frac{F(u_1)-F(u_2)}{u_1-u_2}\right] \approx -\frac{1}{2}F''(0)\,;\\
&f(u_1) - F''(u_1)t \approx (ac - F''(0))t = 2F''(0)t\,;\\
&f'(u_1) - F'''(u_1)t \approx 2a(ct)^{1/2},.
\end{aligned} \tag{5.56}$$

Plugging them back in A_{11}, we have

$$A_{11} \approx \frac{1}{2}F''(0)\cdot(2F''(0)t)^{-1} - 2actF''(0)(2F''(0)t)^{-2} = -\frac{5}{4}t^{-1}\,,$$

Similarly one has:

$$A_{22} \approx -\frac{1}{4}t^{-1}\,, \quad \text{and} \quad S_1 = \mathcal{O}(t^{-1/2})\,.$$

All together,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[(\partial_z u)^2 + (\partial_z u_2)^2\right] \le C,$$

and $H_1(\,\mathrm{d}z)$ norm grows at most algebraically.

Chapter 6

Conclusion

Kinetic equations with random inputs have been studied extensively during the recent three years. On the numerical aspect, stochastic asymptotic-preserving (s-AP) schemes have been designed for many multiscale kinetic equations, including the author's work with S. Jin [57] on a kinetic-fluid two-phase flow model. On the theoretical aspect, random space regularity has been proved for many multiscale kinetic equations using energy/hypocoercivity estimates, including the author's work with S. Jin [95] on the two-phase flow model. Spectral accuracy of generalized polynomial chaos based stochastic Galerkin (gPC-sG) methods are also proved, for which the author's work with S. Jin [95] is the first result for nonlinear kinetic equations. The random space regularity of the Vlasov-Poisson equation, which is time-reversible, is also studied by the author's work with S. Jin [94]. The author also contributes to some other related topics, including gPC-sG methods for kinetic equations with multi-dimensional random inputs [93], and spectrally accurate interpolation methods for the Burgers' equation with random inputs [69].

Several questions are still open: on the numerical aspect,

1. How to design s-AP schemes for kinetic equations with nonlinear collision operators? [95] deals with the Fokker-Planck operator $\mathcal{L}_u$, which is only weakly nonlinear, in the sense that it is linear for a fixed u. For equations with quadratic collision operators, like the Boltzmann or FPL equations, s-AP schemes have not been developed yet.

2. How to design accurate UQ methods for hyperbolic equations with discontinuous solutions? This is closely related to the previous question, since the hydrodynamic limits

of many kinetic equations are hyperbolic. One result in this direction is [87], which uses the entropy variables to ensure the hyperbolicity of the gPC system. However, their method suffers from a costly optimization procedure. Another result is [69], in which we interpolate the shock location, but this only works for very special solutions.

3. How to design efficient UQ methods for kinetic equations with high-dimensional random inputs? In [93] we can deal with moderately high-dimensional random spaces, but for really high dimensions (like 20 or more), there are currently no available UQ methods other than Monte Carlo.

On the theoretical aspect,

1. Is it possible to prove random space regularity for nonlinear kinetic equations with large initial data? Deterministic results for existence and long-time behavior are available for some special cases (e.g., [24, 25]). However, it is not clear whether one can generalize such results to prove random space regularity.

2. Is there any kinetic equation such that the solution with random inputs behaves worse than the deterministic solution? All current theoretical results for kinetic equations with random inputs suggest that a deterministic long time decay can be generalized into the random case. It would be interesting if one can find a counterexample to this.

Appendix A: Inversion of the deterministic operator $\mathcal{L}_u$

Here, for readers' convenience, we give a review of the method of inversion of the deterministic operator $\mathcal{L}_u$ proposed in [59]. $\mathcal{L}_u$ can be written as

$$\mathcal{L}_u f = \sqrt{M_u}\tilde{\mathcal{L}}_u h, \tag{A.1}$$

where

$$h = \frac{f}{\sqrt{M_u}}, \quad \tilde{\mathcal{L}}_u h = \frac{1}{\sqrt{M_u}}\nabla_v \cdot \left(M_u \nabla_v \left(\frac{h}{\sqrt{M_u}} \right) \right), \tag{A.2}$$

and M_u as defined in (1.27). The key observation is that the operator $\tilde{\mathcal{L}}_u$ is symmetric in L_v^2. This symmetry can be preserved by the spatial discretization

$$(\tilde{L}_u h)_{i,j} = \frac{1}{\Delta v^2}\left(h_{i,j+1} + h_{i,j-1} + h_{i+1,j} + h_{i-1,j} - \frac{\sqrt{M_{i,j+1}} + \sqrt{M_{i,j-1}} + \sqrt{M_{i+1,j}} + \sqrt{M_{i-1,j}}}{\sqrt{M_{i,j}}} h_{i,j} \right), \tag{A.3}$$

where the subindex u of M_u is omitted. Note that the discrete operator satisfies the well-balanced property $\tilde{L}_u(\sqrt{M_u}) = 0$. If one uses this spatial discretization to solve an equation

$$(a - \mathcal{L}_u) f = g, \tag{A.4}$$

where $a > 0$ is a constant, then the resulting system of linear equations is symmetric positive definite, and one can solve it by the Conjugate Gradient method.

Appendix B: Proof of Theorem 4.5.1

Proof. First, from the conservation property of Q, one has

$$\|f(t,\cdot,\mathbf{z})\|_{L^1_{\mathbf{v}}} = \|f^0(\cdot,\mathbf{z})\|_{L^1_{\mathbf{v}}} \le M\,.$$

Then we use mathematical induction on k. For $k=0$, multiplying (4.19) by f and integrating on $\mathbf{v}$, by the Cauchy-Schwarz inequality and (4.20), one obtains

$$\frac{1}{2}\partial_t \int_{\mathbb{R}^d} f^2 \,\mathrm{d}\mathbf{v} = \int_{\mathbb{R}^d} fQ(f,f)\,\mathrm{d}\mathbf{v} \le \|f\|_{L^2_{\mathbf{v}}}\|Q(f,f)\|_{L^2_{\mathbf{v}}} \le C_B\|f\|_{L^1_{\mathbf{v}}}\|f\|^2_{L^2_{\mathbf{v}}} \le C_B M\|f\|^2_{L^2_{\mathbf{v}}}\,.$$

Now Gronwall's inequality implies that there is a positive constant C_0 such that (4.22) is true for $k=0$.

Now for some $k \ge 0$ assume (4.22) holds. Take any multi-index $\mathbf{j}$ with $|\mathbf{j}|_1 = k+1$. Taking $\mathbf{j}$-th derivative of z on (4.19) gives

$$\partial_t \partial^{\mathbf{j}}_{\mathbf{z}} f = \sum_{\mathbf{l}=\mathbf{0}}^{\mathbf{j}} \binom{\mathbf{j}}{\mathbf{l}} Q(\partial^{\mathbf{l}}_{\mathbf{z}} f, \partial^{\mathbf{j}-\mathbf{l}}_{\mathbf{z}} f) + \sum_{m=1}^{d} j_m \sum_{\mathbf{l}=\mathbf{0}}^{\mathbf{j}-\mathbf{1_m}} \binom{\mathbf{j}-\mathbf{1_m}}{\mathbf{l}} Q_{1,m}(\partial^{\mathbf{l}}_{\mathbf{z}} f, \partial^{\mathbf{j}-\mathbf{1_m}-\mathbf{l}}_{\mathbf{z}} f), \tag{A.1}$$

where we used the bilinearity of the collision operator and the assumption that B is linear in $\mathbf{z}$.

Multiplying (A.1) by $\partial^{\mathbf{j}}_{\mathbf{z}} f$ and integrating over $\mathbf{v}$ yields

$$\begin{aligned}
&\frac{1}{2}\partial_t \int_{\mathbb{R}^d} (\partial^{\mathbf{j}}_{\mathbf{z}} f)^2 \,\mathrm{d}\mathbf{v} \\
&\le \sum_{\mathbf{l}=\mathbf{0}}^{\mathbf{j}} \binom{\mathbf{j}}{\mathbf{l}} \|\partial^{\mathbf{j}}_{\mathbf{z}} f\|_{L^2_{\mathbf{v}}} \|Q(\partial^{\mathbf{l}}_{\mathbf{z}} f, \partial^{\mathbf{j}-\mathbf{l}}_{\mathbf{z}} f)\|_{L^2_{\mathbf{v}}} + \sum_{m=1}^{d} j_m \sum_{\mathbf{l}=\mathbf{0}}^{\mathbf{j}-\mathbf{1_m}} \binom{\mathbf{j}-\mathbf{1_m}}{\mathbf{l}} \|\partial^{\mathbf{j}}_{\mathbf{z}} f\|_{L^2_{\mathbf{v}}} \|Q_{1,m}(\partial^{\mathbf{l}}_{\mathbf{z}} f, \partial^{\mathbf{j}-\mathbf{1_m}-\mathbf{l}}_{\mathbf{z}} f)\|_{L^2_{\mathbf{v}}} \\
&\le \sum_{\mathbf{l}=\mathbf{0}}^{\mathbf{j}} \binom{\mathbf{j}}{\mathbf{l}} C_B \|\partial^{\mathbf{j}}_{\mathbf{z}} f\|_{L^2_{\mathbf{v}}} \|\partial^{\mathbf{l}}_{\mathbf{z}} f\|_{L^2_{\mathbf{v}}} \|\partial^{\mathbf{j}-\mathbf{l}}_{\mathbf{z}} f\|_{L^2_{\mathbf{v}}} + \sum_{m=1}^{d} j_m \sum_{\mathbf{l}=\mathbf{0}}^{\mathbf{j}-\mathbf{1_m}} \binom{\mathbf{j}-\mathbf{1_m}}{\mathbf{l}} C_B \|\partial^{\mathbf{j}}_{\mathbf{z}} f\|_{L^2_{\mathbf{v}}} \|\partial^{\mathbf{l}}_{\mathbf{z}} f\|_{L^2_{\mathbf{v}}} \|\partial^{\mathbf{j}-\mathbf{1_m}-\mathbf{l}}_{\mathbf{z}} f\|_{L^2_{\mathbf{v}}} \\
&\le C_B C_k^2 \|\partial^{\mathbf{j}}_{\mathbf{z}} f\|_{L^2_{\mathbf{v}}} \sum_{\mathbf{0}\le\mathbf{l}\le\mathbf{j},\mathbf{l}\ne\mathbf{0},\mathbf{j}} \binom{\mathbf{j}}{\mathbf{l}} + 2C_B C_0 \|\partial^{\mathbf{j}}_{\mathbf{z}} f\|^2_{L^2_{\mathbf{v}}} + C_B C_k^2 \|\partial^{\mathbf{j}}_{\mathbf{z}} f\|_{L^2_{\mathbf{v}}} \sum_{m=1}^{d} j_m \sum_{\mathbf{l}=\mathbf{0}}^{\mathbf{j}-\mathbf{1_m}} \binom{\mathbf{j}-\mathbf{1_m}}{\mathbf{l}} \\
&= (2^{k+1}-2) C_B C_k^2 \|\partial^{\mathbf{j}}_{\mathbf{z}} f\|_{L^2_{\mathbf{v}}} + 2C_B C_0 \|\partial^{\mathbf{j}}_{\mathbf{z}} f\|^2_{L^2_{\mathbf{v}}} + 2^k (k+1) C_B C_k^2 \|\partial^{\mathbf{j}}_{\mathbf{z}} f\|_{L^2_{\mathbf{v}}}.
\end{aligned} \tag{A.2}$$

In the first inequality we used the Cauchy-Schwarz inequality, and in the second inequality we used (4.21). In the third inequality the induction assumption is used for the second sum,

since the indexes $\mathbf{l}$ and $\mathbf{j}-\mathbf{1_m}-\mathbf{l}$ appeared there have order less than or equal to k. Every term in the first sum can be treated similarly except terms corresponding to the cases of $\mathbf{l}=\mathbf{0}$ and $\mathbf{l}=\mathbf{j}$, which are treated separately. In the final equality, we used the identity $\sum_{l=0}^{L}\binom{L}{l}=(1+1)^L=2^L$.

Then we apply Gronwall's inequality to (A.2) and get the control

$$\sup_{\mathbf{z}\in I_{\mathbf{z}}}\left(\|\partial_{\mathbf{z}}^{\mathbf{j}}f(t,\mathbf{v},\mathbf{z})\|_{L^2_{\mathbf{v}}}^2\right)^{1/2}\leq C_{k+1},$$

with a positive constant C_{k+1}. Sum over all $\mathbf{j}$ with $|\mathbf{j}|_1=k+1$ we get (4.22) for $k+1$. This completes the mathematical induction and the proof. □

LIST OF REFERENCES

[1] R. Abgrall, P. M. Congedo, and G. Geraci. A one-time truncate and encode multiresolution stochastic framework. *J. Comput. Phys.*, 257:19–56, 2014.

[2] B. Alpert. A class of bases in L^2 for the sparse representation of integral operators. *SIAM Journal on Mathematical Analysis*, 24(1):246–262, 1993.

[3] M. J. Andrews and P. J. O'Rourke. The multiphase particle-in-cell (mp-pic) method for dense particulate flows. *International Journal of Multiphase Flow*, 22(2):379–402, 1996.

[4] I. Babuska, F. Nobile, and R. Tempone. A stochastic collocation method for elliptic partial differential equations with random input data. *SIAM Journal on Numerical Analysis*, 45(3):1005–1034, 2007.

[5] I. Babuska, R. Tempone, and G. E. Zouraris. Galerkin finite element approximations of stochastic elliptic partial differential equations. *SIAM Journal on Numerical Analysis*, 42(2):800–825, 2004.

[6] M. Bachmayr, A. Cohen, R. DeVore, and G. Migliorati. Sparse polynomial approximation of parametric elliptic pdes. part ii: lognormal coefficients. *ESAIM: M2AN*, 51(1):341–363, 2017.

[7] J. Back, F. Nobile, L. Tamellini, and R. Tempone. Stochastic spectral galerkin and collocation methods for pdes with random coefficients: a numerical comparison. In E. M. Rønquist J. S. Hesthaven, editor, *Spectral and High Order Methods for Partial Differential Equations*. Springer-Verlag Berlin Heidelberg, 2011.

[8] C. Bardos, F. Golse, and D. Levermore. Fluid dynamic limits of kinetic equations. I. Formal derivations. *J. Stat. Phys.*, 63:323–344, 1991.

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

[10] A. V. Bobylev. One class of invariant solutions of the Boltzmann equation. *Akademiia Nauk SSSR, Doklady*, 231:571–574, 1976.

[11] F. Bolley, Y. Brenier, and G. Loeper. Contractive metrics for scalar conservation laws. *Journal of Hyperbolic Differential Equations*, 02(01):91–107, 2005.

[12] F. Bouchut and L. Desvillettes. A proof of the smoothing properties of the positive part of Boltzmann's kernel. *Revista Matemática Iberoamericana*, 14:47–61, 1998.

[13] M. Branicki and A. J. Majda. Fundamental limitations of polynomial chaos for uncertainty quantification in systems with intermittent instabilities. *Communications in Mathematical Sciences*, 11(1):55 – 103, 2013.

[14] Y. Brenier. l_2 formulation of multidimensional scalar conservation laws. *Archive for Rational Mechanics and Analysis*, 193(1):1–19, Jul 2009.

[15] H.-J. Bungartz and M. Griebel. Sparse grids. *Acta Numerica*, 13:147–269, 2004.

[16] E. Caglioti and C. Maffei. Time asymptotics for solutions of Vlasov-Poisson equation in a circle. *J. Stat. Phys.*, 92:301–323, 1998.

[17] J. A. Carrillo and G. Toscani. Exponential convergence toward equilibrium for homogeneous Fokker–Planck-type equations. *Mathematical Methods in the Applied Sciences*, 21:1269–1286, 1998.

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

[19] A. Chkifa, A. Cohen, and C. Schwab. Breaking the curse of dimensionality in sparse polynomial approximation of parametric PDEs. *Journal de Mathématiques Pures et Appliquées*, April 2014.

[20] A. Chorin. The numerical solution of the Navier-Stokes equations for an incompressible fluid. *Bull. Am. Math. Soc.*, 73:928–931, 1967.

[21] A. Cohen, R. DeVore, and C. Schwab. Convergence rates of best N-term Galerkin approximations for a class of elliptic sPDEs. *Foundations of Computational Mathematics*, 10(6):615–646, 2010.

[22] B. Despres and B. Perthame. Uncertainty propagation: Intrusive kinetic formulations of scalar conservation laws. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):980–1013, 2016.

[23] B. Despres, G. Poette, and D. Lucor. Robust uncertainty propagation in systems of conservation laws with the entropy closure method. In *Uncertainty Quantification in Computational Fluid Dynamics*. Springer, 2013.

[24] L. Desvillettes and C. Villani. On the trend to global equilibrium in spatially inhomogeneous entropy-dissipating systems: The linear Fokker-Planck equation. *Communications on Pure and Applied Mathematics*, 54(1):1–42, 2001.

[25] L. Desvillettes and C. Villani. On the trend to global equilibrium for spatially inhomogeneous kinetic systems: the Boltzmann equation. *Inventiones mathematicae*, 159(2):245–316, 2005.

[26] J. Ding and D. Gidaspow. A bubbling fluidization model using kinetic theory of granular flow. *AlChE journal*, 36(4):523–538, 1990.

[27] J. K. Dukowicz. A particle-fluid numerical model for the liquid sprays. *Journal of Computational Physics*, 35(2):229–253, 1980.

[28] U. Eckern. Relaxation processes in a condensed Bose gas. *J. Low Temp. Phys.*, 54:333–359, 1984.

[29] R. Esposito, Y. Guo, C. Kim, and M. Marra. Stationary solutions to the Boltzmann equation in the hydrodynamic limit. *Annals of PDE*, 4(1):1–119, 2018.

[30] F. Filbet and S. Jin. A class of asymptotic-preserving schemes for kinetic equations and related problems with stiff sources. *Journal of Computational Physics*, 229:7625–7648, 2010.

[31] H. Freistuhler and D. Serre. l_1 stability of shock waves in scalar viscous conservation laws. *Communications on Pure and Applied Mathematics*, 51(3):291–301, 1998.

[32] J. Garcke and M. Griebel. *Sparse Grids and Applications*. Springer, 2013.

[33] G. Geraci, P. M. Congedo, R. Abgrall, and G. Iaccarino. A novel weakly-intrusive nonlinear multiresolution framework for uncertainty quantification in hyperbolic partial differential equations. *Journal of Scientific Computing*, 66(1):358–405, Jan 2016.

[34] R. G. Ghanem and P. D. Spanos. *Stochastic Finite Elements: A Spectral Approach*. Springer-Verlag, New York, 1991.

[35] Roger G. Ghanem and Robert M. Kruger. Numerical solution of spectral stochastic finite element systems. *Computer Methods in Applied Mechanics and Engineering*, 129(3):289 – 303, 1996.

[36] D. Gidaspow, R. Bezburuah, and J. Ding. *Hydrodynamics of circulating fluidized beds: kinetic theory approach*. Illinois Institute of Technology, Chicago, IL, USA. Department of Chemical Engineering, 1991.

[37] F. Golse and L. Saint-Raymond. The Navier-Stokes limit of the Boltzmann equation for bounded collision kernels. *Invent. Math.*, 155:81–161, 2004.

[38] A. D. Gosman and E. Loannides. Aspects of computer simulation of liquid-fueled combustors. *Journal of Energy*, 7(6):482–490, 1983.

[39] T. Goudon, L. He, A. Moussa, and P. Zhang. The Navier-Stokes-Vlasov-Fokker-Planck system near equilibrium. *SIAM J. Math. Anal.*, 42(5):2177–2202, 2010.

[40] T. Goudon, P.-E. Jabin, and A. Vasseur. Hydrodynamic limit for the Vlasov-Navier-Stokes equations. I. Light particles regime. *Indiana University Mathematics Journal*, 2004.

[41] T. Goudon, P.-E. Jabin, and A. Vasseur. Hydrodynamic limit for the Vlasov-Navier-Stokes equations. II. Fine particles regime. *Indiana University Mathematics Journal*, 2004.

[42] T. Goudon, S. Jin, J.-G. Liu, and B. Yan. Asymptotic-preserving schemes for kinetic-fluid modeling of disperse two-phase flows. *Journal of Computational Physics*, 246:145–164, 2013.

[43] M. Griebel. Adaptive sparse grid multilevel methods for elliptic PDEs based on finite differences. *Computing*, 61(2):151–179, 1998.

[44] M. Griebel and G. Zumbusch. Adaptive sparse grids for hyperbolic conservation laws. In *Hyperbolic Problems: Theory, Numerics, Applications*, pages 411–422. Springer, 1999.

[45] Max D. Gunzburger, Clayton G. Webster, and Guannan Zhang. Stochastic finite element methods for partial differential equations with random input data. *Acta Numer.*, 23:521–650, 2014.

[46] W. Guo and Y. Cheng. A sparse grid discontinuous Galerkin method for high-dimensional transport equations and its application to kinetic simulations. SIAM Journal on Scientific Computing, accepted.

[47] S.-Y. Ha and E. Tadmor. From particle to kinetic and hydrodynamic descriptions of flocking. *Kinetic and Related Models*, 1(3):415–435, 2008.

[48] Y. T. Hou, Q. Li, and P. Zhang. Exploring the locally low dimensional structure in solving random elliptic PDEs. *SIAM Multiscale Model. Simul.*, accepted, 2016.

[49] Y. T. Hou, Q. Li, and P. Zhang. A sparse decomposition of low rank symmetric positive semi-definite matrices. *SIAM Multiscale Model. Simul.*, accepted, 2016.

[50] J. Hu and S. Jin. A stochastic Galerkin method for the Boltzmann equation with uncertainty. *Journal of Computational Physics*, 315:150–168, 2016.

[51] X. Jiang, G. A. Siamas, K. Jagus, and et al. Physical modelling and advanced simulations of gas–liquid two-phase jet flows in atomization and sprays. *Progress in Energy and Combustion Science*, 36(2):131–167, 2010.

[52] S. Jin. Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations. *SIAM Journal on Scientific Computing*, 21(2):441–454, 1999.

[53] S. Jin, J.-G. Liu, and 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. *Research in Math. Sci.*, 4:15, 2017.

[54] S. Jin and L. Liu. An asymptotic-preserving stochastic Galerkin method for the semiconductor Boltzmann equation with random inputs and diffusive scalings. *SIAM Multiscale Modeling and Simulation*, 15:157–183, 2017.

[55] S. Jin and H. Lu. An asymptotic-preserving stochastic Galerkin method for the radiative heat transfer equations with random inputs and diffusive scalings. *J. Comput. Phys.*, 334:182–206, 2017.

[56] S. Jin and L. Pareschi. Discretization of the multiscale semiconductor Boltzmann equation by diffusive relaxation schemes. *J. Comput. Phys.*, 161:312–330, 2000.

[57] S. Jin and R. Shu. A stochastic asymptotic-preserving scheme for a kinetic-fluid model for disperse two-phase flows with uncertainty. *J. Comput. Phys.*, 335:905–924, 2017.

[58] S. Jin, D. Xiu, and X. Zhu. Asymptotic-preserving methods for hyperbolic and transport equations with random inputs and diffusive scalings. *J. Comput. Phys.*, 289:35–52, 2015.

[59] S. Jin and B. Yan. A class of asymptotic-preserving schemes for the Fokker-Planck-Landau equation. *Journal of Computational Physics*, 230:6420–6437, 2011.

[60] S. Jin and Y. Zhu. Hypocoercivity and uniform regularity for the Vlasov-Poisson-Fokker-Planck system with uncertainty and multiple scales. *SIAM J. Math. Anal.*, to appear.

[61] M.-J. Kang and A. F. Vasseur. Criteria on contractions for entropic discontinuities of systems of conservation laws. *Archive for Rational Mechanics and Analysis*, 222(1):343–391, Oct 2016.

[62] M.-J. Kang and A. F. Vasseur. l_2-contraction for shock waves of scalar viscous conservation laws. *Annales de l'Institut Henri Poincare (C) Non Linear Analysis*, 34:139 – 156, 2017.

[63] M. Krook and T. T. Wu. Formation of Maxwellian tails. *Physics of Fluids*, 20:1589–1595, 1977.

[64] L. D. Landau. On the vibration of the electronic plasma. *J. Phys. USSR*, 10:25, 1946.

[65] L.D. Landau. The transport equation in the case of the Coulomb interaction. In *Collected Papers of L.D. Landau*, pages 163–170. Pergamon press, 1981.

[66] N. Leger. l_2 stability estimates for shock solutions of scalar conservation laws using the relative entropy method. *Archive for Rational Mechanics and Analysis*, 199(3):761–778, Mar 2011.

[67] M. Lemou and L. Mieussens. A new asymptotic preserving scheme based on micro-macro formulation for linear kinetic equations in the diffusion limit. *SIAM J. Sci. Comput.*, 31(1):334–368, 2008.

[68] D. Li and W. Yu. On the necessity of the solvable conditions of the typical boundary value problems for quasilinear hyperbolic systems. *Communications in Partial Differential Equations*, 6(11):1225–1234, 1981.

[69] Q. Li, J.-G. Liu, and R. Shu. Polynomial interpolation of Burgers' equation with randomness. *submitted*.

[70] Q. Li and L. Pareschi. Exponential Runge-Kutta for the inhomogeneous Boltzmann equations with high order of accuracy. *J. Comput. Phys.*, 259:402–420, 2014.

[71] Q. Li and L. Wang. Uniform regularity for linear kinetic equations with random input based on hypocoercivity. *SIAM/ASA Journal on Uncertainty Quantification*, to appear.

[72] P.-L. Lions. Compactness in Boltzmann's equation via Fourier integral operators and applications. I, II. *Journal of Mathematics of Kyoto University*, 34(2):391–427,429–461, 1994.

[73] L. Liu. Uniform spectral convergence of the stochastic Galerkin method for the linear semiconductor Boltzmann equation with random inputs and diffusive scalings. *Kinetic and Related Models-AIMS*, to appear.

[74] L. Liu and S. Jin. Hypocoercivity based sensitivity analysis and spectral convergence of the stochastic Galerkin approximation to collisional kinetic equations with multiple scales and random inputs. *SIAM Multiscale Model. Simul.*, to appear.

[75] O. P. Le Maître and O. M. Knio. *Spectral Methods for Uncertainty Quantification, Scientific Computation, with Applications to Computational Fluid Dynamics*. Springer, New York, 2010.

[76] O. P. Le Maître, H. N. Najm, R. G. Ghanem, and O. M. Knio. Multi-resolution analysis of Wiener-type uncertainty propagation schemes. *Journal of Computational Physics*, 197(2):502–531, 2004.

[77] A. Majda. *Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables*, volume 53 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1984.

[78] S. Mishra, N. H. Risebro, C. Schwab, and S. Tokareva. Numerical solution of scalar conservation laws with random flux functions. *SIAM/ASA Journal on Uncertainty Quantification*, 4:552–591, 2016.

[79] S. Mishra and Ch. Schwab. Sparse tensor multi-level monte carlo finite volume methods for hyperbolic conservation laws with random initial data. *Math. Comp.*, 81:1979–2018, 2012.

[80] S. Motsch and E. Tadmor. Heterophilious dynamics enhances consensus. *SIAM Review*, 56(4):577–621, 2014.

[81] C. Mouhot and L. Pareschi. Fast algorithms for computing the Boltzmann collision operator. *Mathematics of Computation*, 75:1833–1852, 2006.

[82] C. Mouhot and C. Villani. On Landau damping. *Acta Math.*, 207:29–201, 2011.

[83] H. Niederreiter, P. Hellekalek, G. Larcher, and P. Zinterhof. *Monte Carlo and Quasi-Monte Carlo Methods 1996*. Springer-Verlag, 1998.

[84] F. Nobile, R. Tempone, and C. G. Webster. A sparse grid stochastic collocation method for partial differential equations with random input data. *SIAM J. Numer. Anal.*, 46(5):2309–2345, 2008.

[85] P. J. O'Rourke. *Collective drop effects on vaporizing liquid sprays*. Los Alamos National Lab., NM (USA), 1981.

[86] L. Pareschi and G. Russo. Implicit-explicit Runge-Kutta schemes and applications to hyperbolic systems with relaxation. *J. Sci. Comput.*, 25:129–155, 2005.

[87] G. Poette, B. Despres, and D. Lucor. Uncertainty quantification for systems of conservation laws. *Journal of Computational Physics*, 228(7):2443 – 2467, 2009.

[88] S. C. Saxena. Devolatilization and combustion characteristics of coal particles. *Progress in Energy and Combustion Science*, 16(1):55–94, 1990.

[89] D. Schiavazzi, A. Doostan, and G. Iaccarino. Sparse multiresolution stochastic approximation for uncertainty quantification. *Recent Advances in Scientific Computing and Applications*, 586:295, 2013.

[90] C. Schwab, E. Süli, and R. A. Todor. Sparse finite element approximation of high-dimensional transport-dominated diffusion problems. *ESAIM: Mathematical Modelling and Numerical Analysis*, 42(5):777–819, 2008.

[91] C. Schwab and R.-A. Todor. Sparse finite elements for elliptic problems with stochastic loading. *Numerische Mathematik*, 95(4):707–734, 2003.

[92] J. Shen and H. Yu. Efficient spectral sparse grid methods and applications to high-dimensional elliptic problems. *SIAM Journal on Scientific Computing*, 32(6):3228–3250, 2010.

[93] R. Shu, J. Hu, and S. Jin. A stochastic Galerkin method for the Boltzmann equation with multi-dimensional random inputs using sparse wavelet bases. *Numer. Math. Theor. Meth. Appl.*, 10:465–488, 2017.

[94] R. Shu and S. Jin. A study of Landau damping with random initial inputs. *submitted*.

[95] R. Shu and S. 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.*, accepted.

[96] S. Smolyak. Quadrature and interpolation formulas for tensor products of certain classes of functions. *Doklady Akademii Nauk SSSR*, 4:240–243, 1963.

[97] G. Szegő. *Orthogonal Polynomials*. American Mathematical Society, 1939.

[98] R. Temam. Sur l'approximation de la solution des equations de Navier–stokes par la méthode des pas fractionnarires ii. *Arch. Ration. Mech. Anal.*, 33:377–385, 1969.

[99] C. Villani. A review of mathematical topics in collisional kinetic theory. In *Handbook of Mathematical Fluid Dynamics*, volume 1, pages 71–305. Amsterdam: North-Holland, 2002.

[100] C. Villani. Hypocoercivity. arXiv:math/0609050, 2006.

[101] Z. Wang, Q. Tang, W. Guo, and Y. Cheng. Sparse grid discontinuous Galerkin methods for high-dimensional elliptic equations. Journal of Computational Physics, accepted.

[102] D. Xiu. Fast numerical methods for stochastic computations: a review. *Communications in Computational Physics*, 5(2-4):242–272, 2009.

[103] D. Xiu. *Numerical methods for stochastic computations: A spectral method approach*. Princeton University Press, Princeton, NJ, 2010.

[104] D. Xiu and J. S. Hesthaven. High-order collocation methods for differential equations with random inputs. *SIAM J. Sci. Comput.*, 27(3):1118–1139 (electronic), 2005.

[105] D. Xiu and G. E. Karniadakis. The Wiener-Askey polynomial chaos for stochastic differential equations. *SIAM J. Sci. Comput.*, 24(2):619–644 (electronic), 2002.

[106] D. Xiu and G. E. Karniadakis. Modeling uncertainty in flow simulations via generalized polynomial chaos. *Journal of Computational Physics*, 187(1):137 – 167, 2003.

[107] C. Zenger. Sparse grids. In *Parallel Algorithms for Partial Differential Equations, Proceedings of the Sixth GAMM-Seminar*, volume 31. 1990.

[108] G. Zhang and M. Gunzburger. Error analysis of a stochastic collocation method for parabolic partial differential equations with random input data. *SIAM J. Numer. Anal.*, 50(4):1922–1940, 2012.

[109] X. Zhang and C.-W. Shu. On positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations on rectangular meshes. *Journal of Computational Physics*, 229(23):8918–8934, 2010.

[110] Y. Zhu and S. Jin. The Vlasov-Poisson-Fokker-Planck system with uncertainty and a one-dimensional asymptotic-preserving method. *SIAM Multiscale Model. Simul.*, 15:1502–1529, 2017.