

ERROR ESTIMATES OF ASYMPTOTIC-PRESERVING NEURAL NETWORKS IN APPROXIMATING STOCHASTIC LINEARIZED BOLTZMANN EQUATION

JIAYU WAN*, LIU LIU AND ZHENYI ZHU

Abstract. In this paper, we construct an asymptotic-preserving neural network (APNN) (Jin *et al.*, *J. Sci. Comput.* **94** (2023) 57) for the linearized Boltzmann equation with uncertain parameters. Utilizing the micro-macro decomposition, we design the loss function based on the stochastic-Galerkin system conducted from the micro-macro equations. Rigorous analysis is provided to show the capability of the neural network in approximating solutions near the global Maxwellian. By employing hypocoercivity techniques, we demonstrate two key results: (i) the existence of APNN leading to arbitrarily small loss function, and (ii) the convergence of the APNN's approximated solution as the loss tends to zero, with the error exhibiting an exponential decay in time.

Mathematics Subject Classification. 35Q20, 68T07, 82C40, 65F99.

Received March 5, 2025. Accepted November 25, 2025.

1. INTRODUCTION

Kinetic equations have been widely used in many areas such as rarefied gas, plasma physics, astrophysics, semiconductor device modeling, and social and biological sciences [33, 41]. They describe the non-equilibrium dynamics of a system composed of a large number of particles and bridge atomistic and continuum models in the hierarchy of multiscale modeling. The Boltzmann-type equation, as one of the most representative models in kinetic theory, provides a power tool to describe molecular gas dynamics, radiative transfer, plasma physics, and polymer flow [3]. They have significant impacts in designing, optimization, control, and inverse problems. For example, it can be used in the design of semiconductor devices, topology optimization of gas flow channels, or risk management in quantitative finance [10]. Many of these applications often require finding unknown or optimal parameters in the Boltzmann-type equations or mean-field models [2, 6, 7, 9].

In addition, kinetic equations typically involve various sources of uncertainty, such as modeling errors, imprecise measurements, and uncertain initial conditions. In particular, the collision kernel in the Boltzmann equation governs the transition rates during particle collisions. Calculating this collision kernel from first principles is highly complex, and in practice, heuristic approximations or empirical data are often used, inevitably introducing uncertainties. Additionally, uncertainties may stem from inaccurate measurements of initial or boundary conditions, as well as from source terms, further compounding the uncertainties in the model. As a result, addressing

Keywords and phrases. Linearized Boltzmann equation, uncertainty quantification, deep learning, asymptotic-preserving neural networks, hypocoercivity.

The Chinese University of Hong Kong, Hong Kong, P.R. China.

*Corresponding author: jiayuan@cuhk.edu.hk

uncertainty quantification (UQ) becomes essential for evaluating, validating, and improving the underlying models, underscoring our project's significance. For numerical studies of the Boltzmann equation and other kinetic models with or without randomness, we refer readers to works such as [18, 20, 24, 38] and [12, 17, 19, 35, 36]. Among the various numerical approaches, the generalized polynomial chaos (gPC)-based stochastic Galerkin (SG) method and its variations have been widely adopted, demonstrating success in a range of applications [43]. Beyond numerical simulations, theoretical studies have established the stability and convergence of these methods. Spectral convergence for the gPC-based SG method was demonstrated in [16, 28, 29], while [11, 27] introduced a robust framework based on hypocoercivity to perform local sensitivity analysis for a class of multiscale, inhomogeneous kinetic equations with random uncertainties-approximated using the gPC-based SG method. For further reference, we point readers to the recent collection [23] and the survey [37].

Modeling and predicting the evolution of multiscale systems such as the Boltzmann-type equations have always been challenging, which often requires sophisticated knowledge of numerical methods and labor-intensive implementation, in addition to the prohibitive costs due to the well-known curse of dimensionality. The issue of high dimensionality becomes even more overwhelming when uncertainties are considered, making traditional numerical approaches unfavorable. This motivates researchers to develop data-driven models and methods [42] in the recent decade.

Machine Learning, or Deep neural networks (DNNs) in particular, have gained increasing interest in approximating the solutions of partial differential equations (PDE) due to their universal approximation property and ability to handle high-dimensional problems. Many studies have been done to employ DNN for solving deterministic or parameterized PDEs, and have shown remarkable promise in various applications. Among those, *Physics Informed Neural Network* (PINN) [39] is one of the most famous approaches which incorporates the available physical laws and limited data, such as boundary or initial conditions and the source term, to approximate the solution of the underlying PDE. The idea has been successfully applied in the simulation of many forward and inverse problems [8, 31, 32].

However, if one directly applies the standard PINNs when dealing with multiscale kinetic models, it may lead to incorrect inferences and predictions. This is due to the presence of small scales needing to be enforced consistently during the learning process, but a standard PINN formulation only captures the solution at the leading order of the Knudsen number, thereby losing accuracy in the asymptotic limit regimes. To overcome this difficulty, the authors in [25] propose the *Asymptotic-Preserving Neural Networks* (APNNs) to enhance the performance of standard PINN to solve multiscale linear transport equations. The idea of APNN is inspired by the traditional AP schemes [21], which preserve the asymptotic transition from one scale to another at the discrete level and capture the limiting macroscopic behavior of the solution when the scaling parameter approaches zero. Recently, an APNN-based method is performed to solve the linear semiconductor Boltzmann equation with good accuracy [30]. APNN is also applied to study hyperbolic-type linear kinetic equations with multiple scales, for example, see [4].

Although some numerical experiments have been conducted to validate the efficiency and accuracy of APNN-based methods, rigorous analysis of the convergence of these methods is still limited. For literature in this direction, we refer to [1], in which the authors study Boltzmann equations with linear collision kernels and present a formal proof of the convergence of the APNN solutions to the real solutions in the standard L^2 space, as the defined APNN loss approaches zero. In this paper, we carry out rigorous analysis of APNN for linearized Boltzmann equation perturbed around some global Maxwellian. We will perform error estimates for the method and derive some convergence results. Our innovations are twofold. First, we introduce uncertainties into our system and adopt the SG method to show the convergence of the approximated SG coefficients in some weighted normed space, as inspired by Liu and Jin [27]. Second, we conduct hypocoercivity analysis [5, 34] to the APNN system to derive convergence in H^1 space with errors decaying exponentially in time, thus providing a more accurate rate of convergence compared to [1]. Our analysis can be outlined as follows. First, we perform micro-macro decomposition for the linearized Boltzmann equation and define our APNN loss function based on the micro-macro system. We then have two main theorems to prove: the first theorem suggests that there exists neural networks which lead to arbitrarily small APNN loss, and the proof is based on the Universal

Approximation Theorem (UAT) of neural networks. The second theorem states that the errors of the APNN-approximated solutions tend to zero as the APNN loss approaches zero, with an exponential decay in time. The proof is inspired by Briant [5], which constructs a Lyapunov functional that is equivalent to the standard Sobolev norm H^1 . We then estimate the time evolution for each term involved in the functional and keep track of the terms that serve as components of the APNN loss function.

The rest of the paper is organized as follows. In Section 2, we review some important concepts in kinetic theory which serve as the cornerstone for our analysis. This includes a short introduction to linearized Boltzmann equation with uncertainty, a complete list of hypocoercivity assumptions, and a concise review of the stochastic Galerkin method. In Section 3, we derive a coupled system in the micro-macro decomposition framework with a formal analysis of the asymptotic limit. We then define our APNN loss function based on the decomposition and generalize the result to the stochastic Galerkin setting. In Section 4, we prove our main results, which are the two theorems mentioned in the previous paragraph. In Section 5, we conduct numerical experiments and demonstrate that the APNN-based SG method is efficient to solve the linear Boltzmann model with uncertainties. Lastly, we conclude the paper and mention the future work.

2. PRELIMINARIES

2.1. Introduction to Boltzmann equation with uncertainties

Consider the initial value problem for the Boltzmann equation

$$\begin{cases} \partial_t f + \frac{1}{\varepsilon^\alpha} v \cdot \nabla_x f = \frac{1}{\varepsilon^{1+\alpha}} Q(f, f) \\ f(0, x, v, z) = f_I(x, v, z), \end{cases} \quad x \in \mathbb{T}^d, v \in \mathbb{R}^d, z \in I_z \subset \mathbb{R} \quad (1)$$

where $f = f(t, x, v, z)$ represents the particle density distribution in the phase space that depends on time t , particle position x , particle velocity v and a random variable z . The number $d \geq 1$ denotes the dimension of the spatial and velocity spaces, and z is a random variable that lies in domain $I_z \subset \mathbb{R}$ with compact support, which is used to account for the random uncertainties of inputs. For more details on the parameterization of random inputs by a finite-dimensional random vector, we refer the reader to the appendix. The operator Q is quadratic and models the binary collisional interactions between particles. The parameter ε is the dimensionless Knudsen number, the ratio of particle mean free path over the domain size. The choice $\alpha = 1$ is referred to the incompressible Navier–Stokes scaling, while $\alpha = 0$ corresponds to the Euler (acoustic) scaling. Periodic boundary condition for the spatial domain $\Omega = \mathbb{T}^d$ is assumed.

We consider random collision kernels, hence the operator Q is defined by

$$Q(f, g) = \int_{\mathbb{R}^d \times \mathbb{S}^{d-1}} B(|v - v_*|, \cos \theta, z) (f' g'_* + f'_* g' - f g_* - f_* g) dv_* d\sigma \quad (2)$$

where we adopt the notations $f' = f(v')$, $f_* = f(v_*)$, $f'_* = f(v'_*)$ and similar for g . v' and v'_* are the post-collisional velocities of particles depending on the pre-collisional velocities v and v_* , which are given by:

$$v' = \frac{v + v_*}{2} + \frac{|v - v_*|}{2} \sigma, \quad v'_* = \frac{v + v_*}{2} - \frac{|v - v_*|}{2} \sigma$$

$\theta \in [0, \pi]$ is the deviation angle between $v' - v'_*$ and $v - v_*$. The collision kernel $B = B(|v - v_*|, \cos \theta, z)$ is a non-negative function determined by physics, and it is assumed to depend on the random variable $z \in I_z$ in our setting. The Boltzmann collision operator conserves mass, momentum and energy, namely

$$\int_{\mathbb{R}^d} Q(f, f) \phi(v) dv = 0, \quad \phi(v) = 1, v, |v|^2. \quad (3)$$

Moreover, we have the celebrated Boltzmann's H-theorem,

$$\int_{\mathbb{R}^d} Q(f, f) \log(f) \, dv \leq 0, \quad (4)$$

such that

$$\int_{\mathbb{R}^d} Q(f, f) \log(f) \, dv = 0 \Leftrightarrow Q(f, f) = 0 \Leftrightarrow f = \widetilde{M},$$

where \widetilde{M} is the *local* equilibrium state:

$$\widetilde{M} = \frac{\rho}{(2\pi T)^{\frac{d}{2}}} \exp\left(-\frac{|v-u|^2}{2T}\right), \quad (5)$$

with

$$\begin{aligned} \rho(x, t) &= \int_{\mathbb{T}^d \times \mathbb{R}^d} f \, dv, \\ u(x, t) &= \frac{1}{\rho} \int_{\mathbb{T}^d \times \mathbb{R}^d} f v \, dv, \\ T(x, t) &= \frac{1}{d\rho} \int_{\mathbb{T}^d \times \mathbb{R}^d} f |v-u|^2 \, dv, \end{aligned}$$

corresponding to the density, mean velocity and temperature of the gas respectively, which are all determined by the initial datum due to the conservation properties. The *global* equilibrium is the unique stationary solution to (1) and is given by

$$\mathcal{M}(v) = \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{|v|^2}{2}}. \quad (6)$$

We consider a linearization around the *global* equilibrium for the solution:

$$f = \mathcal{M} + \varepsilon M h, \quad (7)$$

where $M = \sqrt{\mathcal{M}}$. Inserting (7) into the Boltzmann equation (1), the fluctuation h satisfies

$$\begin{cases} \partial_t h + \frac{1}{\varepsilon^\alpha} v \cdot \nabla_x h = \frac{1}{\varepsilon^{1+\alpha}} \mathcal{L}(h) + \frac{1}{\varepsilon^\alpha} \mathcal{F}(h, h) \\ h(0, x, v, z) = h_I, \end{cases} \quad (8)$$

where the linearized collision operator \mathcal{L} is defined as

$$\begin{aligned} \mathcal{L}(h) &= M^{-1}(Q(Mh, \mathcal{M}) + Q(\mathcal{M}, Mh)) \\ &= M \int_{\mathbb{R}^d \times \mathbb{S}^{d-1}} B(|v-v_*|, \cos \theta, z) \mathcal{M}_* \left(\frac{h'_*}{M'_*} + \frac{h'}{M'} - \frac{h_*}{M_*} - \frac{h}{M} \right) \, dv_* \, d\sigma, \end{aligned} \quad (9)$$

with the nonlinear operator \mathcal{F} given by

$$\begin{aligned} \mathcal{F}(h, h) &= M^{-1}[Q(Mh, Mh) + Q(Mh, Mh)] \\ &= \int_{\mathbb{R}^d \times \mathbb{S}^{d-1}} B(|v-v_*|, \cos \theta, z) M_* (h'_* h' - h_* h) \, dv_* \, d\sigma. \end{aligned} \quad (10)$$

In this paper, we will focus mainly on the acoustic scaling ($\alpha = 0$) in spatial dimension three ($d = 3$), and we will ignore the influence of the nonlinear operator \mathcal{F} by setting $\mathcal{F} = 0$ [34]. Then (8) becomes

$$\begin{cases} \partial_t h + v \cdot \nabla_x h = \frac{1}{\varepsilon} \mathcal{L}(h) \\ h(0, x, v, z) = h_I. \end{cases} \quad (11)$$

The linear operator \mathcal{L} is self-adjoint on the space L_v^2 , and it can be split as

$$\mathcal{L}(h) = K(h) - \Lambda(h)$$

such that

$$K(h) = L^+(h) - L^*(h), \quad L^*(h) = M[(hM) * \Phi],$$

with

$$L^+(h) = \int_{\mathbb{R}^3 \times \mathbb{S}^2} B(|v - v_*|, \cos \theta, z) (h' M_*' + h_*' M') M_* \, dv_* \, d\sigma.$$

In addition, $\Lambda(h) = \nu(v, z)h$, with the collision frequency

$$\nu(v, z) = \int_{\mathbb{R}^3 \times \mathbb{S}^2} B(|v - v_*|, \cos \theta, z) \mathcal{M}_* \, dv_* \, d\sigma = (\Phi * \mathcal{M})(v). \quad (12)$$

Assumptions on the collision kernel

We will make the same assumptions on the collision kernel as in [11]. In particular, we will consider hard potentials with B satisfying Grad's angular cutoff, that is,

$$\begin{aligned} B(|v - v_*|, \cos \theta, z) &= \Phi(|v - v_*|) b(\cos \theta, z), & \Phi(|v - v_*|) &= C|v - v_*|^\gamma, \quad \gamma \in [0, 1], \quad C > 0, \\ \forall \eta \in [-1, 1], \quad |b(\eta, z)| &\leq C_b, \quad |\partial_\eta b(\eta, z)| \leq \tilde{C}_b, \quad |\partial_z^k b(\eta, z)| \leq C_b^*, \quad \forall 0 \leq k \leq r. \end{aligned} \quad (13)$$

We assume that b is linear in z , namely

$$b(\cos \theta, z) = b_0(\cos \theta) + b_1(\cos \theta)z, \quad |z| \leq C_z. \quad (14)$$

Furthermore, we presume that

$$b_0(\cos \theta) \geq (2^q + 2)|b_1(\cos \theta)|C_z + D(\cos \theta), \quad (15)$$

where $D \in C^1([-1, 1])$ such that $|D(\eta)|, |D'(\eta)| \leq C_D, \forall \eta \in [-1, 1]$. The importance of the linearity assumption (14) will become clear when we introduce the stochastic Galerkin method in Section 2.3. Equation (15) is a technical assumption to ensure that our hypocoercivity analysis can be generalized to the stochastic setting (see Sect. 4 for more details).

2.2. Hypocoercivity assumptions

Our error estimate is based on hypocoercivity analysis, which relies on some assumptions on the collision kernel. In this subsection, we will summarize these assumptions and we comment that the linearized Boltzmann operator \mathcal{L} satisfies all these assumptions [34].

H1: Λ satisfies coercivity conditions. $\mathcal{L} : L_{x,v}^2 = L^2(\mathbb{T}^3 \times \mathbb{R}^3)$ is closed, self-adjoint on L_v^2 and local in x . \mathcal{L} has the form $\mathcal{L} = K - \Lambda$. There is a norm $\|\cdot\|_{\Lambda_v}$ on L_v^2 given by

$$\|h\|_{\Lambda_v} = \left\| h(1 + |v|)^{\frac{\gamma}{2}} \right\|_{L_v^2} \quad (16)$$

such that $\forall h \in L_v^2$, Λ satisfies the coercivity condition:

$$\nu_0^\Lambda \|h\|_{L_v^2}^2 \leq \nu_1^\Lambda \|h\|_{\Lambda_v}^2 \leq \langle \Lambda(h), h \rangle_{L_v^2} \leq \nu_2^\Lambda \|h\|_{\Lambda_v}^2, \quad (17)$$

and $\forall h \in H_v^1$,

$$\langle \nabla_v \Lambda(h), \nabla_v h \rangle_{L_v^2} \geq \nu_3^\Lambda \|\nabla_v h\|_{\Lambda_v}^2 - \nu_4^\Lambda \|h\|_{\Lambda_v}^2, \quad (18)$$

where $(\nu_s^\Lambda)_{0 \leq s \leq 4} > 0$ are constants depending on the operators and the velocity space. We also assume that $\forall h, g \in L_v^2$,

$$\langle \mathcal{L}(h), g \rangle_{L_v^2} \leq C^\mathcal{L} \|h\|_{\Lambda_v}^2 \|g\|_{\Lambda_v}^2. \quad (19)$$

We define a norm $\|\cdot\|_\Lambda$ on $L_{x,v}^2$ by

$$\|\cdot\|_\Lambda^2 = \int_{\mathbb{T}^3} \|\cdot\|_{\Lambda_v}^2 dx. \quad (20)$$

H2: K has a regularizing effect. For any $\delta > 0$, there exists some explicit constant $C(\delta) > 0$ such that $\forall h \in H_v^1$,

$$\langle \nabla_v K(h), \nabla_v h \rangle_{L_v^2} \leq C(\delta) \|h\|_{L_v^2}^2 + \delta \|\nabla_v h\|_{L_v^2}^2. \quad (21)$$

In fact, it is shown in [34] that $\|\nabla_v K(h)\|_{L_v^2}^2 \leq C(\delta) \|h\|_{L_v^2}^2 + \delta \|\nabla_v h\|_{L_v^2}^2$.

H3: \mathcal{L} has a finite dimensional kernel. One has

$$N(\mathcal{L}) = \text{Span}\{\varphi_1, \dots, \varphi_N\},$$

such that $\{\varphi_i\}_{1 \leq i \leq N}$ forms an orthonormal set with $\varphi_i(v) = P_i(v)e^{-\frac{|v|^2}{4}}$, where P_i is a polynomial.

We denote by $\pi_\mathcal{L}$ the orthogonal projection in L_v^2 onto $N(\mathcal{L})$. Then $\forall h \in L_v^2$,

$$\pi_\mathcal{L}(h) = \sum_{i=1}^N \left(\int_{\mathbb{R}^3} h \varphi_i dv \right) \varphi_i. \quad (22)$$

Here \mathcal{L} has the local coercivity property: there exists $\lambda > 0$ such that $\forall h \in L_v^2$,

$$\langle \mathcal{L}(h), h \rangle_{L_v^2} \leq -\lambda \|h^\perp\|_{\Lambda_v}^2, \quad (23)$$

where $h^\perp = h - \pi_\mathcal{L}(h)$ represents the microscopic part of h . We summarize here some key facts about the fluid projection $\pi_\mathcal{L}$, which will be used later.

If $h \in H_{x,v}^1$, write $\pi_\mathcal{L}(h) = \sum h_i \varphi_i$ as in (22), then $\partial_{v_i} \varphi_j$ is still of the form $P(v)e^{-\frac{|v|^2}{4}}$ for a polynomial P , whose norm in L_v^2 is finite. We let $M = \max\{\|\nabla_v \varphi_i\|_{L_v^2}^2\}_{1 \leq i \leq N}$, then $\|\nabla_v \pi_\mathcal{L}(h)\|_{L_{x,v}^2}^2 \leq M \|\pi_\mathcal{L}(h)\|_{L_{x,v}^2}^2$ by Cauchy-Schwarz and the orthonormality of $\{\varphi_i\}$. Similar results can be proved for $\|v \cdot \nabla_v \pi_\mathcal{L}(h)\|_{L_{x,v}^2}^2$ and $\|\nabla_v \pi_\mathcal{L}(v \cdot \mathbf{h})\|_{L_{x,v}^2}^2$ by choosing M to be $\max\{\|v \cdot \nabla_v \varphi_i\|_{L_v^2}^2\}$ and $\max\{\|\nabla_v v_i \varphi_j\|_{L_v^2}^2\}$ respectively. Hence there exists $C_{\pi_1} > 0$ such that $\forall h \in H_{x,v}^1$,

$$\begin{aligned} \|\nabla_v \pi_\mathcal{L}(h)\|_{L_{x,v}^2}^2 &\leq C_{\pi_1} \|\pi_\mathcal{L}(h)\|_{L_{x,v}^2}^2 \\ \|v \cdot \nabla_v \pi_\mathcal{L}(h)\|_{L_{x,v}^2}^2 &\leq C_{\pi_1} \|\pi_\mathcal{L}(h)\|_{L_{x,v}^2}^2 \\ \|\nabla_v \pi_\mathcal{L}(v \cdot \mathbf{h})\|_{L_{x,v}^2}^2 &\leq C_{\pi_1} \|\pi_\mathcal{L}(\mathbf{h})\|_{L_{x,v}^2}^2. \end{aligned} \quad (24)$$

Moreover, setting $M = \max\{\|(1 + |v|)^{\frac{\gamma}{2}} \varphi_i\|_{L_v^2}^2\}$ and $M = \max\{\|(1 + |v|)^{\frac{\gamma}{2}} \nabla_v \varphi_i\|_{L_v^2}^2\}$ respectively, we can prove similar results for $\|\pi_\mathcal{L}(h)\|_\Lambda^2$ and $\|\nabla_v \pi_\mathcal{L}(h)\|_\Lambda^2$. Hence there exists $C_\pi > 0$ such that $\forall h \in H_{x,v}^1$,

$$\begin{aligned} \|\pi_\mathcal{L}(h)\|_\Lambda^2 &\leq C_\pi \|\pi_\mathcal{L}(h)\|_{L_{x,v}^2}^2 \leq C_\pi \|h\|_{L_{x,v}^2}^2 \\ \|\nabla_v \pi_\mathcal{L}(h)\|_\Lambda^2 &\leq C_\pi \|\pi_\mathcal{L}(h)\|_{L_{x,v}^2}^2 \leq C_\pi \|h\|_{L_{x,v}^2}^2. \end{aligned} \quad (25)$$

Inequality (25), combining with **H1**, shows that the Λ -norm and the standard $L_{x,v}^2$ -norm are equivalent on the fluid part of \mathcal{L} . Finally, for all $h \in N(\frac{1}{\varepsilon} \mathcal{L} - v \cdot \nabla_x)^\perp$, the Poincaré inequality on the torus gives that for some $C_p > 0$,

$$\|\pi_\mathcal{L}(h)\|_{L_{x,v}^2}^2 \leq C_p \|\nabla_x \pi_\mathcal{L}(h)\|_{L_{x,v}^2}^2 \leq C_p \|\nabla_x h\|_{L_{x,v}^2}^2. \quad (26)$$

2.3. Stochastic Galerkin (SG) method

In this subsection, we discuss how to manage the random variable z existing in the perturbed Boltzmann equation (11). The main idea involves applying the stochastic Galerkin (SG) method to eliminate the randomness, thereby transforming the equation into a system of equations containing only deterministic coefficients. Throughout the paper, we assume that $z \in I_z \subset \mathbb{R}$ with compact support.

We define the space

$$\mathbb{P}^K := \text{Span}\left\{\phi_i(z) \mid 1 \leq i \leq K\right\}$$

equipped with the inner product with respect to the probability density function $\pi(z)$ in z (which is given *a priori*):

$$\langle f(t, x, v, \cdot), g(t, x, v, \cdot) \rangle_{I_z} = \int_{I_z} f(t, x, v, z) g(t, x, v, z) \pi(z) dz,$$

where $\{\phi_i(z)\}_{i=1}^K$ is an orthonormal gPC basis function, *i.e.*,

$$\int_{I_z} \phi_i(z) \phi_j(z) \pi(z) dz = \delta_{ij}, \quad 1 \leq i, j \leq K. \quad (27)$$

For (11), we employ the SG method and expand $h(t, x, v, z)$ as

$$h(t, x, v, z) \approx h_K(t, x, v, z) := \sum_{i=1}^K h^i(t, x, v) \phi_i(z), \quad (28)$$

with

$$h^i(t, x, v) = \int_{I_z} h(t, x, v, z) \phi_i(z) \pi(z) dz. \quad (29)$$

We perform the Galerkin projection of (11) onto \mathbb{P}^K to obtain a system of equations:

$$\begin{cases} \partial_t h^i + v \cdot \nabla_x h^i = \frac{1}{\varepsilon} \mathcal{L}_i(h_K), \\ h_i(0, x, v) = h_i^I(x, v), \end{cases} \quad (30)$$

for $1 \leq i \leq K$, with periodic boundary conditions and initial data for h_i given by

$$h_i^I(x, v) := \int_{I_z} h_I(x, v, z) \phi_i(z) \pi(z) dz.$$

The linear term $\mathcal{L}_i(h_K)$ is defined by

$$\begin{aligned} \mathcal{L}_i(h_K) &= \langle \mathcal{L}(h_K), \phi_i \rangle_{I_z} = \sum_{k=1}^K \langle \mathcal{L}(h_k \phi_k), \phi_i \rangle_{I_z} \\ &= \sum_{k=1}^K M \int_{\mathbb{R}^3 \times \mathbb{S}^2} S_{ik} \mathcal{M}_* \left(\frac{h_*^{k'}}{M_*'} + \frac{h^{k'}}{M'} - \frac{h_*^k}{M_*} - \frac{h^k}{M} \right) dv_* d\sigma \\ &= \sum_{k=1}^K \mathcal{L}_{ik}(h^k) \end{aligned} \quad (31)$$

with

$$S_{ik}(|v - v_*|, \cos \theta) = \int_{I_z} B(|v - v_*|, \cos \theta, z) \phi_i(z) \phi_k(z) \pi(z) dz. \quad (32)$$

The linearity assumption (14) guarantees that for each fixed i , S_{ik} is nonzero for only three choices of $k : i - 1, i, i + 1$, which will simplify our error analysis in Section 4. We hence define

$$\chi_{ik} = \begin{cases} 0, & S_{ik} = 0 \\ 1, & S_{ik} \neq 0. \end{cases} \quad (33)$$

Similar to [27], we make the technical assumption

$$\|\phi_i\|_{L^\infty} \leq Ci^p, \quad \forall 1 \leq i \leq K,$$

with parameter $p \geq 0$. Let $q > p + 2$, we then define the energy E_t^K by

$$E_t^K(\mathbf{h}) = \sum_{i=1}^K \|i^q h^i(t, \cdot, \cdot)\|_{H_{x,v}^1}^2, \quad (34)$$

where $\mathbf{h} = (h^1, h^2, \dots, h^K)^T$.

3. APNN FRAMEWORK

3.1. Micro-macro decomposition

In this subsection, we derive the micro-macro decomposition for the linearized Boltzmann equation perturbed around the global equilibrium $\mathcal{M}(v)$. As before, we assume acoustic scaling and suppose first that the model is deterministic with spatial dimension $d = 3$. We recall here that the equation is given by

$$\begin{cases} \partial_t h + v \cdot \nabla_x h = \frac{1}{\varepsilon} \mathcal{L}(h), \\ h(0, x, v) = h_I. \end{cases} \quad (35)$$

Denote $M = \sqrt{\mathcal{M}}$. By **H3**, the null space of \mathcal{L} is finite dimensional and it is well known that $N(\mathcal{L})$ has an orthonormal basis given by

$$N(\mathcal{L}) = \text{Span}\{\varphi_0 M, \varphi_1 M, \varphi_2 M, \varphi_3 M, \varphi_4 M\}$$

where

$$\begin{cases} \varphi_0(v) = 1 \\ \varphi_i(v) = v_i, \quad 1 \leq i \leq 3 \\ \varphi_4(v) = \frac{1}{\sqrt{6}}(|v|^2 - 3) \end{cases} \quad (36)$$

with $\langle \varphi_i M, \varphi_j M \rangle_{L_v^2} = \delta_{ij}$. Let h_ε be a solution to (35), we define the fluid quantities associated to h_ε by

$$\begin{cases} \rho_\varepsilon(t, x) = \langle h_\varepsilon, \varphi_0 M \rangle_{L_v^2} \\ u_\varepsilon(t, x) = \langle h_\varepsilon, v M \rangle_{L_v^2} \\ T_\varepsilon(t, x) = \langle h_\varepsilon, \varphi_4 M \rangle_{L_v^2}, \end{cases} \quad (37)$$

which are precisely the coefficients of $\pi_{\mathcal{L}}(h_\varepsilon)$ with respect to the orthonormal basis. We now describe the micro-macro decomposition for (35).

We first decompose h_ε as

$$h_\varepsilon = \pi_{\mathcal{L}}(h_\varepsilon) + \varepsilon g_\varepsilon = \mathbf{m}_\varepsilon^T \boldsymbol{\varphi} M + \varepsilon g_\varepsilon, \quad (38)$$

where $\mathbf{m}_\varepsilon = (\rho_\varepsilon, u_{1,\varepsilon}, u_{2,\varepsilon}, u_{3,\varepsilon}, T_\varepsilon)^T$, $\boldsymbol{\varphi} = (\varphi_0, \varphi_1, \varphi_2, \varphi_3, \varphi_4)^T$. To simplify the notation, we delete the subscript ε in the expression above and let $\tilde{h} = \mathbf{m}^T \boldsymbol{\varphi} M$. Then the decomposition becomes $h = \tilde{h} + \varepsilon g$. We substitute this ansatz into (35) to get

$$\partial_t (\tilde{h} + \varepsilon g) + v \cdot \nabla_x (\tilde{h} + \varepsilon g) = \frac{1}{\varepsilon} \mathcal{L}(\tilde{h} + \varepsilon g)$$

which is equivalent to

$$\partial_t \tilde{h} + \varepsilon \partial_t g + v \cdot \nabla_x \tilde{h} + \varepsilon v \cdot \nabla_x g = \mathcal{L}(g). \quad (39)$$

Taking $\pi_{\mathcal{L}}$ on both sides of (39) gives the macroscopic part of the equation

$$\partial_t \tilde{h} + \pi_{\mathcal{L}}(v \cdot \nabla_x \tilde{h}) + \varepsilon \pi_{\mathcal{L}}(v \cdot \nabla_x g) = 0, \quad (40)$$

where we used $\pi_{\mathcal{L}}(g) = 0$, $\langle \mathcal{L}(f), \varphi_i M \rangle_{L^2_v} = 0$, $\forall f \in L^2_v$, $0 \leq i \leq 4$. Note that (40) is equivalent to

$$\partial_t \mathbf{m} + \nabla_x \cdot \langle v \tilde{h} \boldsymbol{\varphi} M \rangle + \varepsilon \nabla_x \cdot \langle v g \boldsymbol{\varphi} M \rangle = 0, \quad (41)$$

where $\langle f \rangle := \int_{\mathbb{R}^3} f(v) dv$ and $\langle v \tilde{h} \boldsymbol{\varphi} M \rangle$ stands for $(\langle v \tilde{h} \varphi_0 M \rangle, \langle v \tilde{h} \varphi_1 M \rangle, \dots, \langle v \tilde{h} \varphi_4 M \rangle)^T$ (similarly for $\langle v g \boldsymbol{\varphi} M \rangle$). Taking $I - \pi_{\mathcal{L}}$ on both sides of (39) leads to the microscopic part

$$\varepsilon \partial_t g + (I - \pi_{\mathcal{L}})v \cdot \nabla_x \tilde{h} + \varepsilon (I - \pi_{\mathcal{L}})v \cdot \nabla_x g = \mathcal{L}(g). \quad (42)$$

The coupled system (41) and (42) is the kinetic/fluid formulation of (35), and we can easily recover (35) from the system by taking the dot product of (41) with $\boldsymbol{\varphi} M$, and then add the result to (42).

As $\varepsilon \rightarrow 0$, the coupled system becomes

$$\begin{cases} \partial_t \mathbf{m} + \nabla_x \cdot \langle v \tilde{h} \boldsymbol{\varphi} M \rangle = 0, \\ (I - \pi_{\mathcal{L}})v \cdot \nabla_x \tilde{h} = \mathcal{L}(g). \end{cases} \quad (43)$$

Thus $g = \mathcal{L}^{-1}((I - \pi_{\mathcal{L}})v \cdot \nabla_x \tilde{h})$ and (ρ, u, T) satisfies $\partial_t \mathbf{m} + \nabla_x \cdot \langle v \tilde{h} \boldsymbol{\varphi} M \rangle = 0$, which takes the following explicit form

$$\begin{cases} \partial_t \rho + \nabla_x \cdot u = 0 \\ \partial_t u + \nabla_x \left(\rho + \frac{\sqrt{6}}{3} T \right) = 0 \\ \partial_t T + \nabla_x \cdot \left(\frac{\sqrt{6}}{3} u \right) = 0. \end{cases} \quad (44)$$

This is precisely the acoustic equations as described in [14], with a slight change in the constant coefficients due to different choices of the orthonormal basis. Hence our macro-micro formulation correctly captures the fluid limit of (35).

Our goal is to obtain (ρ, u, T, g) using neural networks. We will denote f_{θ} to be the approximation of f using a neural network with parameters θ . In the APNN framework, we will construct networks $(\rho_{\theta}, u_{\theta}, T_{\theta}, g_{\theta})$ to approximate (ρ, u, T, g) , and we refer to [30] for a detailed description of the APNN structure. Notice that the function g contains the velocity variable v , which has an unbounded domain \mathbb{R}^3 . But when using neural networks to approximate functions, we typically consider functions in a bounded domain. So we will restrict our training of g_{θ} over some bounded domain $\mathcal{D} \subset \mathbb{R}^3$. For technical purpose, we then extend g_{θ} over the entire velocity space by setting $g_{\theta} = 0$ for $v \in \mathbb{R}^3 - \mathcal{D}$. Moreover, $\pi_{\mathcal{L}}(g) = 0$ if g is the real solution to (42). This clearly does not necessarily hold for an arbitrary g_{θ} . Hence, we perform a post-processing of our network to replace g_{θ} by $g_{\theta} - \pi_{\mathcal{L}}(g_{\theta})$ to make sure $\pi_{\mathcal{L}}(g_{\theta}) = 0$ and hence $\pi_{\mathcal{L}}(h_{\theta}) = \tilde{h}_{\theta}$, where

$$\begin{cases} \mathbf{m}_{\theta} = (\rho_{\theta}, u_{\theta}, T_{\theta})^T \\ \tilde{h}_{\theta} = \mathbf{m}_{\theta}^T \boldsymbol{\varphi} M \\ h_{\theta} = \tilde{h}_{\theta} + \varepsilon g_{\theta}. \end{cases}$$

The best approximation $(\rho_{\theta}, u_{\theta}, T_{\theta}, g_{\theta})$ is obtained by minimizing the loss function of our neural network, which we now describe. Replacing (ρ, u, T, g) by $(\rho_{\theta}, u_{\theta}, T_{\theta}, g_{\theta})$ in (41) and (42), we get

$$\begin{cases} \partial_t \mathbf{m}_{\theta} + \nabla_x \cdot \langle v \tilde{h}_{\theta} \boldsymbol{\varphi} M \rangle + \varepsilon \nabla_x \cdot \langle v g_{\theta} \boldsymbol{\varphi} M \rangle = \mathbf{d}_{\theta}^1 \\ \varepsilon \partial_t g_{\theta} + (I - \pi_{\mathcal{L}})v \cdot \nabla_x \tilde{h}_{\theta} + \varepsilon (I - \pi_{\mathcal{L}})v \cdot \nabla_x g_{\theta} - \mathcal{L}(g_{\theta}) = d_{\theta}^2. \end{cases} \quad (45)$$

We further define

$$d_\theta^{ini} = h_\theta(0, x, v) - h(0, x, v) \quad (46)$$

$$\begin{aligned} d_\theta^b &= (h_\theta(t, x_1, x_2, \pi, v) - h_\theta(t, x_1, x_2, -\pi, v))^2 + (h_\theta(t, x_1, \pi, x_3, v) - h_\theta(t, x_1, -\pi, x_3, v))^2 \\ &\quad + (h_\theta(t, \pi, x_2, x_3, v) - h_\theta(t, -\pi, x_2, x_3, v))^2 \end{aligned} \quad (47)$$

which correspond to the losses on the initial data and boundary value, respectively. We define the generalization error by

$$\mathcal{R}_\theta^G = \mathcal{R}_\theta^1 + \mathcal{R}_\theta^2 + \mathcal{R}_\theta^{ini} + \mathcal{R}_\theta^b \quad (48)$$

where

$$\begin{cases} \mathcal{R}_\theta^1 = \int_0^T \mathcal{R}_{\theta,t}^1 dt, & \mathcal{R}_{\theta,t}^1 = \int_{\mathbb{T}^3} |d_\theta^1|^2 dx \\ \mathcal{R}_\theta^2 = \int_0^T \mathcal{R}_{\theta,t}^2 dt, & \mathcal{R}_{\theta,t}^2 = \int_{\mathbb{T}^3} \int_{\mathcal{D}} |d_\theta^2|^2 dv dx \\ \mathcal{R}_\theta^{ini} = \int_{\mathbb{T}^3} \int_{\mathcal{D}} |d_\theta^{ini}|^2 dv dx \\ \mathcal{R}_\theta^b = \int_0^T \mathcal{R}_{\theta,t}^b dt, & \mathcal{R}_{\theta,t}^b = \int_{\partial\mathbb{T}^3} \int_{\mathcal{D}} |d_\theta^b|^2 dv d\sigma(x). \end{cases} \quad (49)$$

Notice that as $\epsilon \rightarrow 0$, the residual loss (45) will tend to the residual loss of the fluid limit (43), which clearly demonstrates the AP property of the network.

Since we will eventually prove convergence results in the H^1 space, we need a H^1 -counterpart of the loss function. Hence we further define $d_\theta^{1,\nabla_x} = \nabla_x d_\theta^1$, $d_\theta^{2,\nabla_x} = \nabla_x d_\theta^2$, and let d_θ^{ini,∇_x} , d_θ^{b,∇_x} similarly as d_θ^{ini} , d_θ^b , replacing h_θ, h by $\nabla_x h_\theta, \nabla_x h$. We then define $\mathcal{R}_\theta^{G,\nabla_x}$ similarly as \mathcal{R}_θ^G , replacing each integrand by the corresponding ∇_x -counterpart, and we define $\mathcal{R}_\theta^{G,\nabla_v}$ in the same way. Finally, we let

$$\mathcal{R}_\theta^{G,H^1} = \mathcal{R}_\theta^G + \mathcal{R}_\theta^{G,\nabla_x} + \mathcal{R}_\theta^{G,\nabla_v}, \quad (50)$$

which is the final loss function we are going to minimize during the training process of our networks.

3.2. APNN for the SG system

We now generalize the previous APNN framework to linearized Boltzmann equation with uncertainty. We assume again acoustic scaling and then the equation reads

$$\begin{cases} \partial_t h + v \cdot \nabla_x h = \frac{1}{\epsilon} \mathcal{L}(h), \\ h(0, x, v, z) = h_I. \end{cases} \quad (51)$$

To deal with the random variable z , we follow the stochastic Galerkin method introduced in Section 2.3. Let $h \approx h_K = \sum_{i=1}^K h^i \phi_i(z)$, and we again write $h = \tilde{h} + \epsilon g = \mathbf{m}^T \boldsymbol{\varphi} M + \epsilon g$. Then

$$\begin{aligned} h^i(t, x, v) &= \int_{I_z} h \phi_i(z) \pi(z) dz \\ &= \int_{I_z} (\mathbf{m}^T \boldsymbol{\varphi} M + \epsilon g) \phi_i(z) \pi(z) dz \\ &= \int_{I_z} \mathbf{m}^T \boldsymbol{\varphi} M \phi_i(z) \pi(z) dz + \epsilon \int_{I_z} g \phi_i(z) \pi(z) dz \\ &= (\mathbf{m}^i)^T \boldsymbol{\varphi} M + \epsilon g^i, \end{aligned}$$

where $\mathbf{m}^i = (\rho^i, u^i, T^i)^T$, which can be easily proved to be the coefficients of $\pi_{\mathcal{L}}(h^i)$ with respect to the orthonormal basis $\{\boldsymbol{\varphi} M\}$, due to the fact that $\pi_{\mathcal{L}}$ is interchangeable with $\int_{I_z} \cdot dz$. Hence we also have $g^i \in N(\mathcal{L})^\perp$ for $1 \leq i \leq K$.

As in (30), the stochastic Galerkin system for (51) is given by

$$\partial_t h^i + v \cdot \nabla_x h^i = \frac{1}{\varepsilon} \mathcal{L}_i(h_K), \quad (52)$$

for $1 \leq i \leq K$ where $\mathcal{L}_i(h_K)$ is defined in (31). Substituting $h^i = (\mathbf{m}^i)^T \varphi M + \varepsilon g^i$ into (52) and taking $\pi_{\mathcal{L}}$ and $I - \pi_{\mathcal{L}}$ respectively on both sides of the equation will lead to

$$\begin{cases} \partial_t \mathbf{m}^i + \nabla_x \cdot \langle v \tilde{h}^i \varphi M \rangle + \varepsilon \nabla_x \cdot \langle v g^i \varphi M \rangle = 0, \\ \varepsilon \partial_t g^i + (I - \pi_{\mathcal{L}}) \nabla_x v \cdot \tilde{h}^i + \varepsilon (I - \pi_{\mathcal{L}}) v \cdot \nabla_x g^i = \mathcal{L}_i(g_K), \end{cases} \quad (53)$$

for $1 \leq i \leq K$. This is the macro-micro formulation for (52).

As in the deterministic case, we will construct networks $(\rho_\theta^i, u_\theta^i, T_\theta^i, g_\theta^i)$ to approximate $(\rho_\theta^i, u_\theta^i, T_\theta^i, g_\theta^i)$. The losses $(\mathbf{d}_\theta^{1,i}, d_\theta^{2,i}, d_\theta^{i,i}, d_\theta^{b,i})$ are defined similarly as $(\mathbf{d}_\theta^1, d_\theta^2, d_\theta^i, d_\theta^b)$, replacing $(\rho_\theta, u_\theta, T_\theta, g_\theta)$ by $(\rho_\theta^i, u_\theta^i, T_\theta^i, g_\theta^i)$. Then

$$\begin{cases} \partial_t \mathbf{m}_\theta^i + \nabla_x \cdot \langle v \tilde{h}_\theta^i \varphi M \rangle + \varepsilon \nabla_x \cdot \langle v g_\theta^i \varphi M \rangle = \mathbf{d}_\theta^{1,i}, \\ \varepsilon \partial_t g_\theta^i + (I - \pi_{\mathcal{L}}) v \cdot \nabla_x \tilde{h}_\theta^i + \varepsilon (I - \pi_{\mathcal{L}}) v \cdot \nabla_x g_\theta^i - \mathcal{L}_i(g_{K,\theta}) = d_\theta^{2,i}. \end{cases} \quad (54)$$

The generalization error in the stochastic case is defined by

$$\mathcal{R}_{\theta,sto}^G = \mathcal{R}_{\theta,sto}^1 + \mathcal{R}_{\theta,sto}^2 + \mathcal{R}_{\theta,sto}^i + \mathcal{R}_{\theta,sto}^b, \quad (55)$$

where

$$\mathcal{R}_{\theta,sto}^1 = \sum_{i=1}^K i^{2q} \int_0^T \mathcal{R}_{\theta,t}^{1,i} dt, \quad \mathcal{R}_{\theta,t}^{1,i} = \int_{\mathbb{T}^3} |\mathbf{d}_\theta^{1,i}|^2 dx,$$

and $\mathcal{R}_{\theta,sto}^2, \mathcal{R}_{\theta,sto}^i, \mathcal{R}_{\theta,sto}^b$ are defined similarly. Finally, we define

$$\mathcal{R}_{\theta,sto}^{G,H^1} = \mathcal{R}_{\theta,sto}^G + \mathcal{R}_{\theta,sto}^{G,\nabla_x} + \mathcal{R}_{\theta,sto}^{G,\nabla_v}, \quad (56)$$

where each term is defined analogously to the deterministic case.

4. MAIN RESULTS

In this section, we formally prove our convergence results. Similar results for APNN are obtained in [1]. Compared to [1], our innovations can be summarized as follows. First, our problem contains uncertainties, which are characterized by the random variable z . This stochastic setting will introduce new challenges to our analysis, as we shall see later. Second, our analysis differs significantly from that in [1], in terms of both methodology and implications. In general, we have adopted hypercoercivity methods to derive a more explicit convergence result than [1]. We also mention that our analysis can be easily generalized to other linear models involving Boltzmann-type operators, such as the linear semiconductor Boltzmann model, which will be illustrated by numerical experiments in Section 5.

4.1. Existence of APNN with arbitrarily small loss

Our first result states that there exists a neural network so that the APNN loss $\mathcal{R}_{\theta,sto}^{G,H^1}$ is sufficiently small. Notice that g_θ^i can only approximate g^i in a bounded subset of the velocity space. Hence inspired by Abdo *et al.* [1], we need to make some technical assumptions on g^i outside some bounded subset of \mathbb{R}^3 so that the approximation is valid over the entire velocity space. To be specific, we define the following quantities.

Let

$$c_i = \int_0^T \int_{\mathbb{T}^3} \int_{\mathbb{R}^3 - \mathcal{D}} |g^i|^2 + |v \cdot \nabla_x g^i|^2 dv dx dt.$$

Define $c_i^{\nabla^x}, c_i^{\nabla^v}$ in a similarly way, replacing g^i by $\nabla_x g^i$ and $\nabla_v g^i$ respectively in the definition of c_i . Let

$$c_{ik}^\Lambda = \int_0^T \int_{\mathbb{T}^3} \int_{\mathbb{R}^3 - \mathcal{D}} |\nu_{ik} \nabla_v g^k|^2 + |(\nabla_v \nu_{ik}) g^k|^2 dv dx dt$$

where ν_{ik} is defined similarly as (12), replacing the collision kernel with the proper kernels in the SG setting.

We make the following assumption, which asserts that quantities related to the microscopic part \mathbf{g} outside some sufficiently large \mathcal{D} is negligible.

Assumption 1. *For any $\eta > 0$, there exists a bounded domain $\mathcal{D} \subset \mathbb{R}^3$ large enough, such that $c_{\mathcal{D}} := \sum_{i=1}^K (c_i + c_i^{\nabla^x} + c_i^{\nabla^v} + \sum_{k=1}^K c_{ik}^\Lambda) < \eta$.*

Theorem 1. *Consider the linearized Boltzmann equation with uncertainty (51), and let $(\rho, \mathbf{u}, \theta, \mathbf{g})$ be the solution to the macro-micro formulation (53) such that the microscopic part \mathbf{g} satisfies Assumption 1. Then for any $\delta > 0$, there exists a bounded domain $\mathcal{D} \subset \mathbb{R}^3$ and a network $(\rho_\theta, \mathbf{u}_\theta, \mathbf{T}_\theta, \mathbf{g}_\theta)$ such that*

$$\mathcal{R}_{\theta,sto}^{G,H^1} < \delta.$$

Proof. To prove the statement, we need the following lemma from [30], which is basically the Universal Approximation Theorem for neural networks. \square

Lemma 1. *Let $\Omega \subset \mathbb{R}^N$ be a bounded subset. Suppose $f \in C^2(\Omega)$. Then for any $\eta > 0$, there exists a two-layer neural network f_θ such that,*

$$\|f - f_\theta\|_{W^{2,\infty}(\Omega)} < \eta.$$

Let \mathcal{D} be a bounded domain in \mathbb{R}^3 such that $c_{\mathcal{D}} < d\delta$ for some $d > 0$ to be chosen later. Let $\Omega = [0, T] \times \mathbb{T}^3 \times \mathcal{D}$. Suppose $h^i \in C^2(\Omega)$ for all $1 \leq i \leq K$. Then from Lemma 1, for any $\eta > 0$, there exists a network $(\rho_\theta, \mathbf{u}_\theta, \mathbf{T}_\theta, \mathbf{g}_\theta)$ such that

$$\begin{cases} \|\rho_\theta^i - \rho^i\|_{W^{2,\infty}([0,T] \times \mathbb{T}^3)} < \eta, & \|u_\theta^i - u^i\|_{W^{2,\infty}([0,T] \times \mathbb{T}^3)} < \eta, \\ \|T_\theta^i - T^i\|_{W^{2,\infty}([0,T] \times \mathbb{T}^3)} < \eta, & \|g_\theta^i - g^i\|_{W^{2,\infty}(\Omega)} < \eta. \end{cases} \quad (57)$$

Combining (53) and (54), we have

$$\begin{cases} \partial_t \hat{\mathbf{m}}_\theta^i + \nabla_x \cdot \langle v \hat{h}_\theta^i \varphi M \rangle + \varepsilon \nabla_x \cdot \langle v \hat{g}_\theta^i \varphi M \rangle = -\mathbf{d}_\theta^{1,i} \\ \varepsilon \partial_t \hat{g}_\theta^i + (I - \pi_{\mathcal{L}})v \cdot \nabla_x \hat{h}_\theta^i + \varepsilon (I - \pi_{\mathcal{L}})v \cdot \nabla_x \hat{g}_\theta^i - \mathcal{L}_i(\hat{g}_{K,\theta}) = -d_\theta^{2,i}. \end{cases} \quad (58)$$

For the rest of the proof, we use the notation $\|\cdot\| = \|\cdot\|_{L^2_{[0,T] \times \mathbb{T}^3 \times \mathbb{R}^3}}$, $\|\cdot\|_\Omega = \|\cdot\|_{L^2_\Omega}$ and $\|\cdot\|_{\Omega - \mathcal{D}} = \|\cdot\|_{L^2_{[0,T] \times \mathbb{T}^3 \times (\mathbb{R}^3 - \mathcal{D})}}$. Consider $\mathcal{R}_\theta^{2,i} = \|d_\theta^{2,i}\|_\Omega^2$. From (57),

$$\varepsilon \|\partial_t \hat{g}_\theta^i\|_\Omega^2 < \|\partial_t \hat{g}_\theta^i\|_\Omega^2 \leq |\Omega| \cdot \|\partial_t \hat{g}_\theta^i\|_{L^\infty_\Omega}^2 < |\Omega| \eta^2$$

and

$$\begin{aligned} \|\varepsilon (I - \pi_{\mathcal{L}})v \cdot \nabla_x \hat{g}_\theta^i\|_\Omega^2 &\leq \|(I - \pi_{\mathcal{L}})v \cdot \nabla_x \hat{g}_\theta^i\|_\Omega^2 \leq \|v \cdot \nabla_x \hat{g}_\theta^i\|_\Omega^2 \\ &= \|v \cdot \nabla_x \hat{g}_\theta^i\|_\Omega^2 + \|v \cdot \nabla_x \hat{g}_\theta^i\|_{\Omega - \mathcal{D}}^2 \\ &< C_1 \eta^2 + \|v \cdot \nabla_x g^i\|_{\Omega - \mathcal{D}}^2 \\ &\leq C_1 \eta^2 + d\delta \end{aligned}$$

where we use the fact that g_θ^i vanishes when $v \in \mathbb{R}^3 - \mathcal{D}$ and Assumption 1 in the last two inequalities. Similarly, we can show $\|(I - \pi_{\mathcal{L}})v \cdot \nabla_x \hat{h}_\theta^i\|_\Omega^2 < C_2 \eta^2 + d\delta$.

Finally, since \mathcal{L}_{ik} is a bounded operator on $L^2(\mathbb{R}^3)$, for all $1 \leq i, k \leq K$, we have

$$\|\mathcal{L}_{ik}(\hat{g}_\theta^k)\|_\Omega^2 \leq \|L_{ik}(\hat{g}_\theta^k)\|_\Omega^2 \leq C_3 \|\hat{g}_\theta^k\|_\Omega^2 = C_3 \left(\|\hat{g}_\theta^k\|_\Omega^2 + \|g^k\|_{\Omega-\mathcal{D}}^2 \right) < C_3 \eta^2 + C_3 d \delta.$$

Collecting all the above inequalities, we get $\mathcal{R}_\theta^{2,i} < C\eta^2 + C'd\delta$ for some $C, C' > 0$. Using the same arguments, we can show $\mathcal{R}_\theta^{2,i,\nabla_x} < C\eta^2 + C'd\delta$. For $\mathcal{R}_\theta^{2,i,\nabla_v}$, we will apply similar arguments, with additional observations that

$$\begin{aligned} \|\nabla_v(I - \pi_{\mathcal{L}})v \cdot \nabla_x \hat{g}_\theta^i\|^2 &\leq \|\nabla_x \hat{g}_\theta^i\|^2 + \|v \cdot \nabla_v \nabla_x \hat{g}_\theta^i\|^2 + \|\nabla_v \pi_{\mathcal{L}}(v \cdot \nabla_x \hat{g}_\theta^i)\|^2 \\ &\leq \|\nabla_x \hat{g}_\theta^i\|^2 + \|v \cdot \nabla_v \nabla_x \hat{g}_\theta^i\|^2 + C_{\pi_1} \|v \cdot \nabla_x \hat{g}_\theta^i\|^2 \end{aligned}$$

where we use (24) for the last inequality, and

$$\begin{aligned} \|\nabla_v \mathcal{L}_{ik}(\hat{g}_\theta^k)\|^2 &\leq 2\|\nabla_v K_{ik}(\hat{g}_\theta^k)\|^2 + 2\|\nabla_v \Lambda_{ik}(\hat{g}_\theta^k)\|^2 \\ &\leq 2\eta \|\nabla_v \hat{g}_\theta^k\|^2 + 2C(\eta) \|\hat{g}_\theta^k\|^2 + 4\|(\nabla_v \nu_{ik})\hat{g}_\theta^k\|^2 + 4\|\nu_{ik} \nabla_v \hat{g}_\theta^k\|^2 \end{aligned}$$

where we use (21) and the definition for Λ_{ik} . So we have $\mathcal{R}_{\theta,sto}^2 + \mathcal{R}_{\theta,sto}^{2,\nabla_x} + \mathcal{R}_{\theta,sto}^{2,\nabla_v} < C\eta^2 + C'd\delta$. The error estimates for $\mathcal{R}_{\theta,sto}^1, \mathcal{R}_{\theta,sto}^{ini}, \mathcal{R}_{\theta,sto}^b$ and their ∇_x (or ∇_v)-counterparts are analogous and simpler. So we omit the details here. Finally, setting d and η to be small enough will yield the desired result.

4.2. Convergence of APNN solution

Our next theorem suggests that when the APNN loss $\mathcal{R}_{\theta,sto}^{G,H^1}$ is small enough, $(\rho_\theta, \mathbf{u}_\theta, \mathbf{T}_\theta, \mathbf{g}_\theta)$ will tend towards the true solution $(\rho, \mathbf{u}, \mathbf{T}, \mathbf{g})$ in the weighted H^1 -norm. Again, we need to make some technical assumptions on the microscopic part \mathbf{g} , which is given as follows.

Let

$$\begin{cases} \tilde{R}_{1,t}^i = \int_{\mathbb{T}^3} \int_{\mathbb{R}^3-\mathcal{D}} |\partial_t g^i|^2 + |v \cdot \nabla_x g^i|^2 + |\mathcal{L}_i(g_K)|^2 dv dx, & \tilde{R}_1^i = \int_0^T \tilde{R}_{1,t}^i dt, \\ \tilde{R}_2^i = \int_{\mathbb{T}^3} \int_{\mathbb{R}^3-\mathcal{D}} |g^i(0, x, v)|^2 dv dx, \\ \tilde{R}_{3,t}^i = \int_{\partial\mathbb{T}^3} \int_{\mathbb{R}^3-\mathcal{D}} v |g^i|^2 dv d\sigma(x), & \tilde{R}_3^i = \int_0^T \tilde{R}_{3,t}^i dt. \end{cases} \quad (59)$$

Replacing g by $\nabla_x g$ and $\nabla_v g$ respectively, we can define $\tilde{R}_j^{\nabla_x, i}$ and $\tilde{R}_j^{\nabla_v, i}$ in a similar way for $j = 1, 2, 3$. We let $\tilde{R} = \sum_{i=1}^K \sum_{j=1}^3 (\tilde{R}_j^i + \tilde{R}_j^{\nabla_x, i} + \tilde{R}_j^{\nabla_v, i})$. We make the following assumption on \tilde{R} .

Assumption 2. For any $0 < \varepsilon < 1$, there exists $\mathcal{D} \subset \mathbb{R}^3$ large enough such that $\tilde{R} < \delta(\varepsilon)$ with $\delta(\varepsilon) \rightarrow 0$ as $\varepsilon \rightarrow 0$.

Theorem 2. Let $(\rho_\theta, \mathbf{u}_\theta, \theta_\theta, \mathbf{g}_\theta)$ be approximations to $(\rho, \mathbf{u}, \mathbf{T}, \mathbf{g})$ using a neural network with parameters θ such that \mathbf{g} satisfies Assumption 2, and let $\hat{\mathbf{h}}_\theta = \mathbf{h} - \mathbf{h}_\theta$. Then for any $t \in (0, T]$ and $0 < \varepsilon < 1$,

$$E_t^K(\hat{\mathbf{h}}_\theta) \leq \frac{\tilde{C}(\mathcal{R}_{\theta,sto}^{G,H^1} + \delta(\varepsilon))}{\varepsilon} e^{-\varepsilon\tau t},$$

for some $\tilde{C}, \tau > 0$ independent of ε .

Proof. Combing the two equations in (58) gives

$$\partial_t \hat{h}_\theta^i + v \cdot \nabla_x \hat{h}_\theta^i = \frac{1}{\varepsilon} \mathcal{L}_i(\hat{h}_{K,\theta}) - \mathbf{d}_\theta^{1,i} \cdot \varphi M - \mathbf{d}_\theta^{2,i} \quad (60)$$

where $\hat{h}_\theta^i = h^i - h_\theta^i \ \forall 1 \leq i \leq K$. We let $A_\theta^i = -d_\theta^{1,i} \cdot \varphi M - d_\theta^{2,i}$. Unless specified otherwise, we use the notation $\|\cdot\|^2 = \|\cdot\|_{L_{x,v}^2}^2$ and $\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_{L_{x,v}^2}$ for the rest of the proof. Define

$$\begin{cases} r_{1,t}^i := \int_{\mathbb{T}^3} \int_{\mathbb{R}^3 - \mathcal{D}} |d_\theta^{2,i}|^2 \, dv \, dx \\ r_2^i := \int_{\mathbb{T}^3} \int_{\mathbb{R}^3 - \mathcal{D}} |h^i(0, x, v)|^2 \, dv \, dx \\ r_{3,t}^i := \int_{\partial\mathbb{T}^3} \int_{\mathbb{R}^3 - \mathcal{D}} v |h^i|^2 \, dv \, d\sigma(x). \end{cases}$$

Since g_θ vanishes on $\mathbb{R}^3 - \mathcal{D}$ and the Maxwellian $\mathcal{M}(v)$ is sufficiently small outside some large \mathcal{D} , we can derive that $\sum_{i=1}^K \int_0^T r_{1,t}^i \, dt < \delta(\varepsilon)$ using Assumption 2. Similarly, $\sum_{i=1}^K r_2^i < \delta(\varepsilon)$ and $\sum_{i=1}^K \int_0^T r_{3,t}^i \, dt < \delta(\varepsilon)$. Same arguments also hold for $r_j^{\nabla_x, i}$ and $r_j^{\nabla_v, i}$ with $1 \leq j \leq 3$. We then follow the idea in [5] to construct a Lyapunov functional $\|\cdot\|_{\varepsilon_\perp}^2$ on $H_{x,v}^1$ that is equivalent to the standard Sobolev norm by

$$\|h\|_{\varepsilon_\perp}^2 = a_1 \|h\|^2 + a_2 \|\nabla_x h\|^2 + a_3 \|\nabla_v h^\perp\|^2 + a_4 \varepsilon \langle \nabla_x h, \nabla_v h \rangle \quad \forall h \in H_{x,v}^1 \quad (61)$$

for $a_1, a_2, a_3, a_4 > 0$. □

Lemma 2. *For any fixed choices of $a_1, a_3, a_4 > 0$ and $0 < \varepsilon < 1$, we can choose a_2 to be big enough so that $\|\cdot\|_{\varepsilon_\perp}^2$ is equivalent to the standard Sobolev norm $\|\cdot\|_{H_{x,v}^1}^2$, with equivalence independent of ε .*

Proof of lemma. Using Cauchy–Schwarz, Young’s inequality and the fact that $\varepsilon < 1$, we have

$$\begin{aligned} a_4 \varepsilon \langle \nabla_x h, \nabla_v h \rangle &\leq a_4 \varepsilon \eta \|\nabla_x h\|^2 + \frac{a_4 \varepsilon}{\eta} \|\nabla_v h\|^2 \leq a_4 \eta \|\nabla_x h\|^2 + \frac{a_4}{\eta} \|\nabla_v h\|^2 \\ a_4 \varepsilon \langle \nabla_x h, \nabla_v h \rangle &\geq -a_4 \varepsilon \eta \|\nabla_x h\|^2 - \frac{a_4 \varepsilon}{\eta} \|\nabla_v h\|^2 \geq -a_4 \eta \|\nabla_x h\|^2 - \frac{a_4}{\eta} \|\nabla_v h\|^2. \end{aligned}$$

Write $h = h^\perp + \pi_{\mathcal{L}}(h)$ and using (24), we get

$$\|\nabla_v h\|^2 \leq \|\nabla_v h^\perp\|^2 + C_{\pi_1} \|h\|^2,$$

so

$$\|h\|_{\varepsilon_\perp}^2 \geq \left(a_1 - \frac{a_4 C_{\pi_1}}{\eta} \right) \|h\|^2 + (a_2 - a_4 \eta) \|\nabla_x h\|^2 + \left(a_3 - \frac{a_4}{\eta} \right) \|\nabla_v h^\perp\|^2.$$

Thus, for any fixed choices $a_1, a_3, a_4 > 0$, we can choose η and a_2 big enough, so that $\|h\|_{\varepsilon_\perp}^2 > 0 \ \forall h \neq 0$ and $\|h\|_{\varepsilon_\perp}^2 \geq C_1 \|h\|_\sim^2$ for some $C_1 > 0$ where $\|h\|_\sim^2 = \|h\|^2 + \|\nabla_x h\|^2 + \|\nabla_v h^\perp\|^2$. Similarly, it can be proved that $\|h\|_{\varepsilon_\perp}^2 \leq C_2 \|h\|_\sim^2$ for some $C_2 > 0$, independent of ε . So for any fixed $a_1, a_3, a_4 > 0$ and $0 < \varepsilon < 1$, we can choose a_2 to be big enough such that $\|\cdot\|_{\varepsilon_\perp}^2$ is equivalent to $\|\cdot\|_\sim^2$, with equivalence independent of ε .

Using (24), we can derive

$$\|\nabla_v h\|^2 \leq \|\nabla_v h^\perp\|^2 + C_{\pi_1} \|h\|^2 \leq \|\nabla_v h\|^2 + 2C_{\pi_1} \|h\|^2.$$

Hence $\|\cdot\|_\sim^2$ is equivalent to the standard Sobolev norm $\|\cdot\|_{H_{x,v}^1}^2$, and the result follows.

Similarly, we define

$$E_{\varepsilon_\perp, t}^K(\mathbf{h}) = a_1 \sum_{i=1}^K \|i^q h^i\|^2 + a_2 \sum_{i=1}^K \|i^q \nabla_x h^i\|^2 + a_3 \sum_{i=1}^K \|i^q \nabla_v h^{i,\perp}\|^2 + a_4 \varepsilon \sum_{i=1}^K i^{2q} \langle \nabla_x h^i, \nabla_v h^i \rangle. \quad (62)$$

Then $E_{\varepsilon_\perp, t}^K$ is equivalent to E_t^K for proper choices of $a_1, a_2, a_3, a_4 > 0$. Next, we use four lemmas to bound the time evolution for each term in (62). □

Lemma 3. *We have the estimate*

$$\partial_t \sum_{i=1}^K \left\| i^q \hat{h}_\theta^i \right\|^2 \leq -\frac{\lambda_D}{\varepsilon} \sum_{i=1}^K \left\| i^q \hat{h}_\theta^{i,\perp} \right\|_\Lambda^2 + R_1$$

where

$$R_1 = f_1 \sum_{i=1}^K \left\| i^q A_\theta^i \right\|^2 + \frac{1}{f_1} \sum_{i=1}^K \left\| i^q \hat{h}_\theta^i \right\|^2 + C \sum_{i=1}^K i^{2q} \mathcal{R}_{\theta,t}^{b,i} + \sum_{i=1}^K i^{2q} r_{3,t}^i$$

for $\lambda_D, f_1, C > 0$, all independent of ε with f_1 a free variable to be chosen later.

Proof of lemma. Take the inner product with respect to \hat{h}_θ^i on both sides of (60), we get

$$\frac{1}{2} \partial_t \left\| \hat{h}_\theta^i \right\|^2 = -\langle v \cdot \nabla_x \hat{h}_\theta^i, \hat{h}_\theta^i \rangle + \frac{1}{\varepsilon} \sum_{k=1}^K \langle \mathcal{L}_{ik}(\hat{h}_\theta^k), \hat{h}_\theta^i \rangle + \langle A_\theta^i, \hat{h}_\theta^i \rangle. \quad (63)$$

Notice that since \hat{h}_θ^i is not necessarily periodic, the operator $v \cdot \nabla_x$ is not skew-symmetric. Using integration by part and splitting $\int_{\mathbb{R}^3} = \int_{\mathcal{D}} + \int_{\mathbb{R}^3 - \mathcal{D}}$, we have

$$-\langle v \cdot \nabla_x \hat{h}_\theta^i, \hat{h}_\theta^i \rangle \leq C \mathcal{R}_{\theta,t}^{b,i} + r_{3,t}^i.$$

To bound the term $\frac{1}{\varepsilon} \sum_{k=1}^K \langle \mathcal{L}_{ik}(\hat{h}_\theta^k), \hat{h}_\theta^i \rangle$, we follow exactly the same arguments in [11], which we summarize as follows. Define

$$\text{Term I} = \sum_{i=1}^K i^{2q} \sum_{k=1}^K \langle \mathcal{L}_{ik}(\hat{h}_\theta^k), \hat{h}_\theta^i \rangle. \quad (64)$$

Let $\Theta_i = \frac{\hat{h}_{\theta,*}^{i'}}{M_*'} + \frac{\hat{h}_\theta^{i'}}{M'} - \frac{\hat{h}_{\theta,*}^i}{M_*} - \frac{\hat{h}_\theta^i}{M}$, $\tilde{\Theta}_i = i^q \Theta_i$. Consider the change of variables $(v, v_*) \rightarrow (v', v_*'), (v_*, v)$ respectively. It can be shown that

$$\begin{aligned} \text{Term I} &= \sum_{i,k=1}^K i^{2q} \langle \mathcal{L}_{ik}(\hat{h}_\theta^k), \hat{h}_\theta^i \rangle \\ &= -\frac{1}{4} \int_{\mathbb{T}^3} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3 \times \mathbb{S}^2} \mathcal{M} \mathcal{M}_* \times \text{Term A} \, dv_* \, d\sigma \, dv \, dx, \end{aligned}$$

where

$$\text{Term A} = \sum_{i,k=1}^K \binom{i}{k}^q S_{ik} \tilde{\Theta}_i \tilde{\Theta}_k. \quad (65)$$

Using assumptions (13)–(15) on the collision kernel, it can be proved that

$$\text{Term A} \geq (b_0 - (2^q + 2)|b_1|C_z) \sum_{i=1}^K \tilde{\Theta}_i^2 \geq D(\cos \theta) \sum_{i=1}^K \tilde{\Theta}_i^2,$$

and hence

$$\begin{aligned} \text{Term I} &\leq -\frac{1}{4} \sum_{i=1}^K \int_{\mathbb{T}^3} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3 \times \mathbb{S}^2} \mathcal{M} \mathcal{M}_* \phi(|v - v_*|) D(\cos \theta) \tilde{\Theta}_i^2 \, dv_* \, d\sigma \, dv \, dx \\ &= \sum_{i=1}^K i^{2q} \langle \mathcal{L}^D(\hat{h}_\theta^i), \hat{h}_\theta^i \rangle \leq -\lambda_D \sum_{i=1}^K i^{2q} \left\| \hat{h}_\theta^{i,\perp} \right\|_\Lambda^2, \end{aligned}$$

where

$$\mathcal{L}^D(h) = M \int_{\mathbb{R}^d \times \mathbb{S}^{d-1}} \phi(|v - v_*|) D(\cos \theta) \mathcal{M}_* \left(\frac{h'_*}{M'_*} + \frac{h'}{M'} - \frac{h_*}{M_*} - \frac{h}{M} \right) dv_* d\sigma.$$

We use the Cauchy–Schwarz and Young’s inequality to bound $\langle A_\theta^i, \hat{h}_\theta^i \rangle$ by

$$\langle A_\theta^i, \hat{h}_\theta^i \rangle \leq f_1 \|A_\theta^i\|^2 + \frac{1}{f_1} \|\hat{h}_\theta^i\|^2.$$

Multiply (63) by i^{2q} and sum over all $1 \leq i \leq K$. The result follows by collecting all the inequalities above. \square

Lemma 4. *We have the estimate*

$$\partial_t \sum_{i=1}^K \left\| i^q \nabla_x \hat{h}_\theta^i \right\|^2 \leq -\frac{\lambda_D}{\varepsilon} \sum_{i=1}^K \left\| i^q \nabla_x \hat{h}_\theta^{i,\perp} \right\|_\Lambda^2 + R_2,$$

where

$$R_2 = f_2 \sum_{i=1}^K \left\| i^q \nabla_x A_\theta^i \right\|^2 + \frac{1}{f_2} \sum_{i=1}^K \left\| i^q \nabla_x \hat{h}_\theta^i \right\|^2 + C \sum_{i=1}^K i^{2q} \mathcal{R}_{\theta,t}^{b,i,\nabla_x} + \sum_{i=1}^K i^{2q} r_{3,t}^{\nabla_x,i},$$

for $\lambda_D, f_2, C > 0$, all independent of ε with f_2 a free variable to be chosen later.

Proof of lemma. Take ∇_x on both sides of (60) and then take the inner product with respect to $\nabla_x \hat{h}_\theta^i$, we get

$$\frac{1}{2} \partial_t \left\| \nabla_x \hat{h}_\theta^i \right\|^2 = -\langle v \cdot \nabla_x \nabla_x \hat{h}_\theta^i, \nabla_x \hat{h}_\theta^i \rangle + \frac{1}{\varepsilon} \sum_{k=1}^K \langle \mathcal{L}_{ik}(\nabla_x \hat{h}_\theta^k), \nabla_x \hat{h}_\theta^i \rangle + \langle \nabla_x A_\theta^i, \nabla_x \hat{h}_\theta^i \rangle. \quad (66)$$

Following exactly the same arguments as for $\|\hat{h}_\theta^i\|^2$, we can derive

$$\begin{aligned} -\langle v \cdot \nabla_x \nabla_x \hat{h}_\theta^i, \nabla_x \hat{h}_\theta^i \rangle &\leq C \mathcal{R}_{\theta,t}^{b,i,\nabla_x} + r_{3,t}^{\nabla_x,i} \\ \sum_{i=1}^K i^{2q} \sum_{k=1}^K \langle \mathcal{L}_{ik}(\nabla_x \hat{h}_\theta^k), \nabla_x \hat{h}_\theta^i \rangle &\leq -\lambda_D \sum_{i=1}^K i^{2q} \left\| \nabla_x \hat{h}_\theta^{i,\perp} \right\|_\Lambda^2, \end{aligned}$$

and

$$\langle \nabla_x A_\theta^i, \nabla_x \hat{h}_\theta^i \rangle \leq f_2 \left\| \nabla_x A_\theta^i \right\|^2 + \frac{1}{f_2} \left\| \nabla_x \hat{h}_\theta^i \right\|^2.$$

Hence the result follows by multiplying (66) by i^{2q} and sum over all $1 \leq i \leq K$. \square

Lemma 5. *We have the estimate*

$$\partial_t \sum_{i=1}^K \left\| i^q \nabla_v \hat{h}_\theta^{i,\perp} \right\|^2 \leq \frac{C_1}{\varepsilon} \sum_{i=1}^K \left\| i^q \hat{h}_\theta^{k,\perp} \right\|_\Lambda^2 - \frac{C_2}{\varepsilon} \sum_{i=1}^K \left\| i^q \nabla_v \hat{h}_\theta^{i,\perp} \right\|_\Lambda^2 + \varepsilon C_3 \sum_{i=1}^K \left\| i^q \nabla_x \hat{h}_\theta^i \right\|^2 + R_3,$$

where

$$\begin{aligned} R_3 &= f_3 \tilde{C}_1 \left(\sum_{i=1}^K \left\| i^q A_\theta^i \right\|^2 + \sum_{i=1}^K \left\| i^q \nabla_v A_\theta^i \right\|^2 \right) + \frac{1}{f_3} \tilde{C}_1 \left(\sum_{i=1}^K \left\| i^q \hat{h}_\theta^i \right\|^2 + \sum_{i=1}^K \left\| i^q \nabla_v \hat{h}_\theta^i \right\|^2 \right) \\ &\quad + C'_1 \left(\sum_{i=1}^K i^{2q} \mathcal{R}_{\theta,t}^{b,i} + \sum_{i=1}^K i^{2q} \mathcal{R}_{\theta,t}^{b,i,\nabla_v} \right) + C'_2 \left(\sum_{i=1}^K i^{2q} r_{3,t}^i + \sum_{i=1}^K i^{2q} r_{3,t}^{\nabla_v,i} \right), \end{aligned}$$

for $C_1, C_2, C_3, C'_1, C'_2, \tilde{C}_1, f_3 > 0$, all independent of ε with f_3 a free variable to be chosen later.

Proof of lemma. Take $I - \pi_{\mathcal{L}}$ and ∇_v respectively on both sides of (60), and then take the inner product with respect to $\nabla_v \hat{h}_{\theta}^{i,\perp}$. We get

$$\frac{1}{2} \partial_t \left\| \nabla_v \hat{h}_{\theta}^{i,\perp} \right\|^2 = - \left\langle \nabla_v \left(v \cdot \nabla_x \hat{h}_{\theta}^i \right)^{\perp}, \nabla_v \hat{h}_{\theta}^{i,\perp} \right\rangle + \frac{1}{\varepsilon} \sum_{k=1}^K \left\langle \nabla_v \mathcal{L}_{ik} \left(\hat{h}_{\theta}^{k,\perp} \right), \nabla_v \hat{h}_{\theta}^{i,\perp} \right\rangle + \left\langle \nabla_v A_{\theta}^{i,\perp}, \nabla_v \hat{h}_{\theta}^{i,\perp} \right\rangle, \quad (67)$$

where we used $\mathcal{L}_{ik}(f) \in \mathcal{N}(\mathcal{L}_{ik})^{\perp} \forall f \in L_v^2$, and hence $(\mathcal{L}_{ik}(\hat{h}_{\theta}^k))^{\perp} = \mathcal{L}_{ik}(\hat{h}_{\theta}^k) - \pi_{\mathcal{L}}(\mathcal{L}_{ik}(\hat{h}_{\theta}^k)) = \mathcal{L}_{ik}(\hat{h}_{\theta}^k) = \mathcal{L}_{ik}(\hat{h}_{\theta}^{k,\perp})$.

First, consider the term $\frac{1}{\varepsilon} \sum_{k=1}^K \langle \nabla_v \mathcal{L}_{ik}(\hat{h}_{\theta}^{k,\perp}), \nabla_v \hat{h}_{\theta}^{i,\perp} \rangle$. we multiply the term by i^{2q} and sum over all $0 \leq i \leq K$. Then we use **H1** – **H2** and follow exactly the same arguments in [27] to get

$$\begin{aligned} \frac{1}{\varepsilon} \sum_{i=1}^K i^{2q} \sum_{k=1}^K \left\langle \nabla_v \mathcal{L}_{ik} \left(\hat{h}_{\theta}^{k,\perp} \right), \nabla_v \hat{h}_{\theta}^{i,\perp} \right\rangle &= \frac{1}{\varepsilon} \sum_{i,k=1}^K i^{2q} \left\langle \nabla_v \mathcal{L}_{ik} \left(\hat{h}_{\theta}^{k,\perp} \right), \nabla_v \hat{h}_{\theta}^{i,\perp} \right\rangle \\ &\leq \frac{1}{\varepsilon} \sum_{i,k=1}^K \chi_{ik} \cdot i^{2q} \left(\left(C(\delta) \frac{\nu_1^{\Lambda}}{\nu_0^{\Lambda}} + \nu_4^{\Lambda} \right) \left\| \hat{h}_{\theta}^{k,\perp} \right\|_{\Lambda}^2 + \left(\delta \frac{\nu_1^{\Lambda}}{\nu_0^{\Lambda}} - \nu_3^{\Lambda} \right) \left\| \nabla_v \hat{h}_{\theta}^{i,\perp} \right\|_{\Lambda}^2 \right) \\ &= \frac{1}{\varepsilon} \sum_{i,k=1}^K \chi_{ik} \cdot \frac{i^{2q}}{k^{2q}} \left(C(\delta) \frac{\nu_1^{\Lambda}}{\nu_0^{\Lambda}} + \nu_4^{\Lambda} \right) \left\| k^q \hat{h}_{\theta}^{k,\perp} \right\|_{\Lambda}^2 + \frac{1}{\varepsilon} \sum_{i,k=1}^K \chi_{ik} \cdot \left(\delta \frac{\nu_1^{\Lambda}}{\nu_0^{\Lambda}} - \nu_3^{\Lambda} \right) \left\| i^q \nabla_v \hat{h}_{\theta}^{i,\perp} \right\|_{\Lambda}^2 \\ &\leq \frac{3 \times 4^q}{\varepsilon} \left(C(\delta) \frac{\nu_1^{\Lambda}}{\nu_0^{\Lambda}} + \nu_4^{\Lambda} \right) \sum_{k=1}^K \left\| k^q \hat{h}_{\theta}^{k,\perp} \right\|_{\Lambda}^2 + \frac{3}{\varepsilon} \left(\delta \frac{\nu_1^{\Lambda}}{\nu_0^{\Lambda}} - \nu_3^{\Lambda} \right) \sum_{i=1}^K \left\| i^q \nabla_v \hat{h}_{\theta}^{i,\perp} \right\|_{\Lambda}^2. \end{aligned}$$

For the term $-\langle \nabla_v (v \cdot \nabla_x \hat{h}_{\theta}^i)^{\perp}, \nabla_v \hat{h}_{\theta}^{i,\perp} \rangle$, we expand it as

$$\begin{aligned} - \left\langle \nabla_v \left(v \cdot \nabla_x \hat{h}_{\theta}^i \right)^{\perp}, \nabla_v \hat{h}_{\theta}^{i,\perp} \right\rangle &= - \left\langle \nabla_v \left(v \cdot \nabla_x \hat{h}_{\theta}^i \right), \nabla_v \hat{h}_{\theta}^{i,\perp} \right\rangle + \left\langle \nabla_v \pi_{\mathcal{L}} \left(v \cdot \nabla_x \hat{h}_{\theta}^i \right), \nabla_v \hat{h}_{\theta}^{i,\perp} \right\rangle \\ &= - \left\langle \nabla_x \hat{h}_{\theta}^i, \nabla_v \hat{h}_{\theta}^{i,\perp} \right\rangle - \left\langle v \cdot \nabla_v \nabla_x \hat{h}_{\theta}^i, \nabla_v \hat{h}_{\theta}^{i,\perp} \right\rangle \\ &\quad + \left\langle \nabla_v \pi_{\mathcal{L}} \left(v \cdot \nabla_x \hat{h}_{\theta}^i \right), \nabla_v \hat{h}_{\theta}^{i,\perp} \right\rangle \\ &= - \left\langle \nabla_x \hat{h}_{\theta}^i, \nabla_v \hat{h}_{\theta}^{i,\perp} \right\rangle - \left\langle v \cdot \nabla_v \nabla_x \pi_{\mathcal{L}} \left(\hat{h}_{\theta}^i \right), \nabla_v \hat{h}_{\theta}^{i,\perp} \right\rangle \\ &\quad - \left\langle v \cdot \nabla_x \nabla_v \hat{h}_{\theta}^{i,\perp}, \nabla_v \hat{h}_{\theta}^{i,\perp} \right\rangle + \left\langle \nabla_v \pi_{\mathcal{L}} \left(v \cdot \nabla_x \hat{h}_{\theta}^i \right), \nabla_v \hat{h}_{\theta}^{i,\perp} \right\rangle. \end{aligned}$$

Using integration by part and inequality (24), we have

$$- \left\langle v \cdot \nabla_x \nabla_v \hat{h}_{\theta}^{i,\perp}, \nabla_v \hat{h}_{\theta}^{i,\perp} \right\rangle \leq C_1' \left(\mathcal{R}_{\theta,t}^{b,i} + \mathcal{R}_{\theta,t}^{b,i,\nabla_v} \right) + C_2' \left(r_{3,t}^i + r_{3,t}^{\nabla_v,i} \right).$$

Apply the Cauchy–Schwarz, Young’s inequality and inequality (17), we have

$$\begin{aligned} - \left\langle \nabla_x \hat{h}_{\theta}^i, \nabla_v \hat{h}_{\theta}^{i,\perp} \right\rangle &\leq \eta_1 \left\| \nabla_x \hat{h}_{\theta}^i \right\|^2 + \frac{1}{\eta_1} \left\| \nabla_v \hat{h}_{\theta}^{i,\perp} \right\|^2 \\ &\leq \eta_1 \left\| \nabla_x \hat{h}_{\theta}^i \right\|^2 + \frac{\nu_1^{\Lambda}}{\eta_1 \nu_0^{\Lambda}} \left\| \nabla_v \hat{h}_{\theta}^{i,\perp} \right\|_{\Lambda}^2. \end{aligned}$$

Apply the Cauchy–Schwarz, Young’s inequality, inequality (24) and (17), we can derive

$$\begin{aligned}
-\left\langle v \cdot \nabla_v \nabla_x \pi_{\mathcal{L}}(\hat{h}_\theta^i), \nabla_v \hat{h}_\theta^{i,\perp} \right\rangle &\leq \eta_2 \left\| v \cdot \nabla_v \nabla_x \pi_{\mathcal{L}}(\hat{h}_\theta^i) \right\|^2 + \frac{1}{\eta_2} \left\| \nabla_v \hat{h}_\theta^{i,\perp} \right\|^2 \\
&= \eta_2 \left\| v \cdot \nabla_v \pi_{\mathcal{L}}(\nabla_x \hat{h}_\theta^i) \right\|^2 + \frac{1}{\eta_2} \left\| \nabla_v \hat{h}_\theta^{i,\perp} \right\|^2 \\
&\leq \eta_2 C_{\pi 1} \left\| \pi_{\mathcal{L}}(\nabla_x \hat{h}_\theta^i) \right\|^2 + \frac{1}{\eta_2} \left\| \nabla_v \hat{h}_\theta^{i,\perp} \right\|^2 \\
&\leq \eta_2 C_{\pi 1} \left\| \nabla_x \hat{h}_\theta^i \right\|^2 + \frac{\nu_1^\Lambda}{\eta_2 \nu_0^\Lambda} \left\| \nabla_v \hat{h}_\theta^{i,\perp} \right\|_\Lambda^2
\end{aligned}$$

and

$$\begin{aligned}
\left\langle \nabla_v \pi_{\mathcal{L}}(v \cdot \nabla_x \hat{h}_\theta^i), \nabla_v \hat{h}_\theta^{i,\perp} \right\rangle &\leq \eta_3 \left\| \nabla_v \pi_{\mathcal{L}}(v \cdot \nabla_x \hat{h}_\theta^i) \right\|^2 + \frac{1}{\eta_3} \left\| \nabla_v \hat{h}_\theta^{i,\perp} \right\|^2 \\
&\leq \eta_3 C_{\pi 1} \left\| \pi_{\mathcal{L}}(\nabla_x \hat{h}_\theta^i) \right\|^2 + \frac{1}{\eta_3} \left\| \nabla_v \hat{h}_\theta^{i,\perp} \right\|^2 \\
&\leq \eta_3 C_{\pi 1} \left\| \nabla_x \hat{h}_\theta^i \right\|^2 + \frac{\nu_1^\Lambda}{\eta_3 \nu_0^\Lambda} \left\| \nabla_v \hat{h}_\theta^{i,\perp} \right\|_\Lambda^2.
\end{aligned}$$

Finally, we use the Cauchy–Schwarz, Young’s inequality and (24) to bound $\langle \nabla_v A_\theta^{i,\perp}, \nabla_v \hat{h}_\theta^{i,\perp} \rangle$ by

$$\begin{aligned}
\left\langle \nabla_v A_\theta^{i,\perp}, \nabla_v \hat{h}_\theta^{i,\perp} \right\rangle &\leq f_3 \left\| \nabla_v A_\theta^{i,\perp} \right\|^2 + \frac{1}{f_3} \left\| \nabla_v \hat{h}_\theta^{i,\perp} \right\|^2 \\
&\leq f_3 \tilde{C}_1 \left(\|A_\theta^i\|^2 + \|\nabla_v A_\theta^i\|^2 \right) + \frac{1}{f_3} \tilde{C}_1 \left(\|\hat{h}_\theta^i\|^2 + \|\nabla_v \hat{h}_\theta^i\|^2 \right).
\end{aligned}$$

Multiply both sides of (67) by i^{2q} and sum over all $0 \leq i \leq K$. Then collect all the inequalities above, we get

$$\begin{aligned}
\partial_t \sum_{i=1}^K \left\| i^q \nabla_v \hat{h}_\theta^{i,\perp} \right\|^2 &\leq \frac{3 \times 4^q}{\varepsilon} \left(C(\delta) \frac{\nu_1^\Lambda}{\nu_0^\Lambda} + \nu_4^\Lambda \right) \sum_{i=1}^K \left\| i^q \hat{h}_\theta^{k,\perp} \right\|_\Lambda^2 \\
&\quad + \left(\frac{3}{\varepsilon} \left(\delta \frac{\nu_1^\Lambda}{\nu_0^\Lambda} - \nu_3^\Lambda \right) + \frac{\nu_1^\Lambda}{\eta_1 \nu_0} + \frac{\nu_1^\Lambda}{\eta_2 \nu_0} + \frac{\nu_1^\Lambda}{\eta_3 \nu_0} \right) \sum_{i=1}^K \left\| i^q \nabla_v \hat{h}_\theta^{i,\perp} \right\|_\Lambda^2 \\
&\quad + (\eta_1 + \eta_2 C_{\pi 1} + \eta_3 C_{\pi 1}) \sum_{i=1}^K \left\| i^q \nabla_x \hat{h}_\theta^i \right\|^2 + R_3.
\end{aligned}$$

Let $\eta_1 = \eta_2 = \eta_3 = \varepsilon \hat{\eta}$. We then choose $\hat{\eta}$ to be big enough and δ to be small enough to get the desired result. \square

Lemma 6. *We have the estimate*

$$\partial_t \sum_{i=1}^K i^{2q} \varepsilon \left\langle \nabla_x \hat{h}_\theta^i, \nabla_v \hat{h}_\theta^i \right\rangle \leq -\frac{\varepsilon}{4} \sum_{i=1}^K \left\| i^q \nabla_x \hat{h}_\theta^i \right\|^2 + \frac{D_1 e}{\varepsilon} \sum_{i=1}^K \left\| i^q \nabla_x \hat{h}_\theta^{i,\perp} \right\|_\Lambda^2 + \frac{D_2}{e \varepsilon} \sum_{i=1}^K \left\| i^q \nabla_v \hat{h}_\theta^{i,\perp} \right\|_\Lambda^2 + R_4,$$

where

$$\begin{aligned}
R_4 &= f_4 \sum_{i=1}^K \left\| i^q \nabla_x A_\theta^i \right\|^2 + \frac{1}{f_4} \sum_{i=1}^K \left\| i^q \nabla_v \hat{h}_\theta^i \right\|^2 \\
&\quad + C'_1 \left(\sum_{i=1}^K i^{2q} \mathcal{R}_{\theta,t}^{b,i,\nabla_x} + \sum_{i=1}^K i^{2q} \mathcal{R}_{\theta,t}^{b,i,\nabla_v} \right) + C'_2 \left(\sum_{i=1}^K i^{2q} r_{3,t}^{\nabla_x,i} + \sum_{i=1}^K i^{2q} r_{3,t}^{\nabla_v,i} \right),
\end{aligned}$$

for $D_1, D_2, f_4, e, C'_1, C'_2 > 0$, all independent of ε with $f_4 > 0$ and $e > 4$ two free variables to be chosen later.

Proof of lemma. Take ∇_x on both sides of (60) and then take the inner product with respect to $\nabla_v \hat{h}_\theta^i$, we get

$$\frac{1}{2} \partial_t \langle \nabla_x \hat{h}_\theta^i, \nabla_v \hat{h}_\theta^i \rangle = - \langle \nabla_x (v \cdot \nabla_x \hat{h}_\theta^i), \nabla_v \hat{h}_\theta^i \rangle + \frac{1}{\varepsilon} \sum_{k=1}^K \langle \mathcal{L}_{ik} (\nabla_x \hat{h}_\theta^k), \nabla_v \hat{h}_\theta^i \rangle + \langle \nabla_x A^i, \nabla_v \hat{h}_\theta^i \rangle. \quad (68)$$

Using integration by part, Cauchy–Schwarz and Young’s inequality, one can show that

$$\begin{aligned} - \langle \nabla_x (v \cdot \nabla_x \hat{h}_\theta^i), \nabla_v \hat{h}_\theta^i \rangle &\leq \langle v \cdot \nabla_x \hat{h}_\theta^i, \nabla_x \nabla_v \hat{h}_\theta^i \rangle + C_1 \mathcal{R}_{\theta,t}^{b,i,\nabla_x} + C_2 \mathcal{R}_{\theta,t}^{b,i,\nabla_v} + C'_1 r_{3,t}^{\nabla_x,i} + C'_2 r_{3,t}^{\nabla_v,i} \\ &= - \langle \nabla_v (v \cdot \nabla_x \hat{h}_\theta^i), \nabla_x \hat{h}_\theta^i \rangle + C_1 \mathcal{R}_{\theta,t}^{b,i,\nabla_x} + C_2 \mathcal{R}_{\theta,t}^{b,i,\nabla_v} + C'_1 r_{3,t}^{\nabla_x,i} + C'_2 r_{3,t}^{\nabla_v,i} \\ &= - \left\| \nabla_x \hat{h}_\theta^i \right\|^2 - \langle v \cdot \nabla_v \nabla_x \hat{h}_\theta^i, \nabla_x \hat{h}_\theta^i \rangle + C_1 \mathcal{R}_{\theta,t}^{b,i,\nabla_x} + C_2 \mathcal{R}_{\theta,t}^{b,i,\nabla_v} + C'_1 r_{3,t}^{\nabla_x,i} + C'_2 r_{3,t}^{\nabla_v,i} \\ &\leq - \left\| \nabla_x \hat{h}_\theta^i \right\|^2 + \langle v \cdot \nabla_v \hat{h}_\theta^i, \nabla_x \nabla_x \hat{h}_\theta^i \rangle + 2C_1 \mathcal{R}_{\theta,t}^{b,i,\nabla_x} + 2C_2 \mathcal{R}_{\theta,t}^{b,i,\nabla_v} + 2C'_1 r_{3,t}^{\nabla_x,i} \\ &\quad + 2C'_2 r_{3,t}^{\nabla_v,i}. \end{aligned}$$

Since $\langle v \cdot \nabla_v \hat{h}_\theta^i, \nabla_x \nabla_x \hat{h}_\theta^i \rangle = \langle \nabla_x (v \cdot \nabla_x \hat{h}_\theta^i), \nabla_v \hat{h}_\theta^i \rangle$, we have

$$- \langle \nabla_x (v \cdot \nabla_x \hat{h}_\theta^i), \nabla_v \hat{h}_\theta^i \rangle \leq - \frac{1}{2} \left\| \nabla_x \hat{h}_\theta^i \right\|^2 + C_1 \mathcal{R}_{\theta,t}^{b,i,\nabla_x} + C_2 \mathcal{R}_{\theta,t}^{b,i,\nabla_v} + C'_1 r_{3,t}^{\nabla_x,i} + C'_2 r_{3,t}^{\nabla_v,i}.$$

By (19), for each \mathcal{L}_{ik} , there is a constant $C^{\mathcal{L}_{ik}}$ such that $\langle \mathcal{L}_{ik}(h), g \rangle \leq C^{\mathcal{L}_{ik}} \|h\|_{\Lambda_v} \|g\|_{\Lambda_v} \quad \forall h, g \in L^2_v$. Let $C^{\mathcal{L}} = \max\{C^{\mathcal{L}_{ik}}\}_{0 \leq i, k \leq K}$. Then

$$\begin{aligned} \frac{1}{\varepsilon} \sum_{k=1}^K \langle \mathcal{L}_{ik} (\nabla_x \hat{h}_\theta^k), \nabla_v \hat{h}_\theta^i \rangle &= \frac{1}{\varepsilon} \sum_{k=1}^K \langle \mathcal{L}_{ik} (\nabla_x \hat{h}_\theta^{k,\perp}), \nabla_v \hat{h}_\theta^i \rangle \\ &\leq \frac{1}{\varepsilon} \sum_{k=1}^K C^{\mathcal{L}} \left\| \nabla_x \hat{h}_\theta^{k,\perp} \right\|_{\Lambda} \left\| \nabla_v \hat{h}_\theta^i \right\|_{\Lambda} \cdot \chi_{ik} \\ &\leq \frac{C^{\mathcal{L}} \eta}{\varepsilon} \sum_{k=1}^K \left\| \nabla_x \hat{h}_\theta^{k,\perp} \right\|_{\Lambda}^2 \cdot \chi_{ik} + \frac{C^{\mathcal{L}} (K+1)}{\varepsilon \eta} \left\| \nabla_v \hat{h}_\theta^i \right\|_{\Lambda}^2 \end{aligned}$$

where we apply Young’s inequality in the last inequality above. Using inequality (25) and the Poincare inequality (26), We can further bound $\left\| \nabla_v \hat{h}_\theta^i \right\|_{\Lambda}^2$ by

$$\begin{aligned} \left\| \nabla_v \hat{h}_\theta^i \right\|_{\Lambda}^2 &= \left\| \nabla_v \pi_{\mathcal{L}} (\hat{h}_\theta^i) + \nabla_v \hat{h}_\theta^{i,\perp} \right\|_{\Lambda}^2 \\ &\leq 2 \left\| \nabla_v \pi_{\mathcal{L}} (\hat{h}_\theta^i) \right\|_{\Lambda}^2 + 2 \left\| \nabla_v \hat{h}_\theta^{i,\perp} \right\|_{\Lambda}^2 \\ &\leq 2C_{\pi} \left\| \pi_{\mathcal{L}} (\hat{h}_\theta^i) \right\|_{\Lambda}^2 + 2 \left\| \nabla_v \hat{h}_\theta^{i,\perp} \right\|_{\Lambda}^2 \\ &\leq 2C_{\pi} C_p \left\| \nabla_x \hat{h}_\theta^i \right\|_{\Lambda}^2 + 2 \left\| \nabla_v \hat{h}_\theta^{i,\perp} \right\|_{\Lambda}^2. \end{aligned}$$

Finally, we use Cauchy–Schwarz, Young’s inequality to bound $\langle \nabla_x A_\theta^i, \nabla_v \hat{h}_\theta^i \rangle$ by

$$\langle \nabla_x A_\theta^i, \nabla_v \hat{h}_\theta^i \rangle \leq f_4 \left\| \nabla_x A_\theta^i \right\|^2 + \frac{1}{f_4} \left\| \nabla_v \hat{h}_\theta^i \right\|^2.$$

Set $\eta = \frac{D}{\varepsilon}$ for $D > 0$ and collect all the inequalities above, we get

$$\begin{aligned} \partial_t \langle \nabla_x \hat{h}_\theta^i, \nabla_v \hat{h}_\theta^i \rangle &\leq \left(2C_\pi C_p \frac{C^\mathcal{L}(K+1)}{D} - \frac{1}{2} \right) \left\| \nabla_x \hat{h}_\theta^i \right\|^2 + \frac{C^\mathcal{L} D}{\varepsilon^2} \sum_{k=1}^K \left\| \nabla_x \hat{h}_\theta^{k,\perp} \right\|_\Lambda^2 \cdot \chi_{ik} \\ &\quad + \frac{2C^\mathcal{L}(K+1)}{D} \left\| \nabla_v \hat{h}_\theta^{i,\perp} \right\|_\Lambda^2 + f_4 \left\| \nabla_x A_\theta^i \right\|^2 + \frac{1}{f_4} \left\| \nabla_v \hat{h}_\theta^i \right\|^2 \\ &\quad + C_1 \mathcal{R}_{\theta,t}^{b,i,\nabla_x} + C_2 \mathcal{R}_{\theta,t}^{b,i,\nabla_v} + C'_1 r_{3,t}^{\nabla_x,i} + C'_2 r_{3,t}^{\nabla_v,i}. \end{aligned}$$

We set $D = 2C_\pi C_p C^\mathcal{L}(K+1)e$ for some $e > 4$ to be chosen later. Then we multiply both sides of the inequality above by ε and use the fact that $\varepsilon < 1$ to get

$$\begin{aligned} \partial_t \varepsilon \langle \nabla_x \hat{h}_\theta^i, \nabla_v \hat{h}_\theta^i \rangle &\leq -\frac{\varepsilon}{4} \left\| \nabla_x \hat{h}_\theta^i \right\|^2 + \frac{\tilde{D}_1 e}{\varepsilon} \sum_{k=1}^K \left\| \nabla_x \hat{h}_\theta^{k,\perp} \right\|_\Lambda^2 \cdot \chi_{ik} \\ &\quad + \frac{D_2}{e\varepsilon} \left\| \nabla_v \hat{h}_\theta^{i,\perp} \right\|_\Lambda^2 + f_4 \left\| \nabla_x A_\theta^i \right\|^2 + \frac{1}{f_4} \left\| \nabla_v \hat{h}_\theta^i \right\|^2 \\ &\quad + C_1 \mathcal{R}_{\theta,t}^{b,i,\nabla_x} + C_2 \mathcal{R}_{\theta,t}^{b,i,\nabla_v} + C'_1 r_{3,t}^{\nabla_x,i} + C'_2 r_{3,t}^{\nabla_v,i}. \end{aligned} \tag{69}$$

Finally, multiple both sides of (69) by i^{2q} and sum over all $1 \leq i \leq K$, we get

$$\begin{aligned} \partial_t \sum_{i=1}^K i^{2q} \varepsilon \langle \nabla_x \hat{h}_\theta^i, \nabla_v \hat{h}_\theta^i \rangle &\leq -\frac{\varepsilon}{4} \sum_{i=1}^K \left\| i^q \nabla_x \hat{h}_\theta^i \right\|^2 + \frac{\tilde{D}_1 e}{\varepsilon} \sum_{i,k=1}^K \frac{i^{2q}}{k^{2q}} \left\| k^q \nabla_x \hat{h}_\theta^{k,\perp} \right\|_\Lambda^2 \cdot \chi_{ik} \\ &\quad + \frac{D_2}{e\varepsilon} \sum_{i=1}^K \left\| i^q \nabla_v \hat{h}_\theta^{i,\perp} \right\|_\Lambda^2 + R_4 \\ &\leq -\frac{\varepsilon}{4} \sum_{i=1}^K \left\| i^q \nabla_x \hat{h}_\theta^i \right\|^2 + \frac{D_1 e}{\varepsilon} \sum_{i=1}^K \left\| i^q \nabla_x \hat{h}_\theta^{i,\perp} \right\|_\Lambda^2 \\ &\quad + \frac{D_2}{e\varepsilon} \sum_{i=1}^K \left\| i^q \nabla_v \hat{h}_\theta^{i,\perp} \right\|_\Lambda^2 + R_4, \end{aligned}$$

where $D_1 = 3 \times 4^q \tilde{D}_1$ and the result follows.

Collecting Lemmas 3–6, we can derive that

$$\begin{aligned} \partial_t E_{\varepsilon,\perp,t}^K(\hat{\mathbf{h}}_\theta) &\leq \frac{a_3 C_1 - a_1 \lambda_D}{\varepsilon} \sum_{i=1}^K \left\| i^q \hat{h}_\theta^{i,\perp} \right\|_\Lambda^2 + \frac{a_4 D_1 e - a_2 \lambda_D}{\varepsilon} \sum_{i=1}^K \left\| i^q \nabla_x \hat{h}_\theta^{i,\perp} \right\|_\Lambda^2 \\ &\quad + \frac{a_4 D_2 - a_3 C_2}{\varepsilon} \sum_{i=1}^K \left\| i^q \nabla_v \hat{h}_\theta^{i,\perp} \right\|_\Lambda^2 + \varepsilon \left(a_3 C_3 - \frac{a_4}{4} \right) \sum_{i=1}^K \left\| i^q \nabla_x \hat{h}_\theta^i \right\|^2 + R, \end{aligned} \tag{70}$$

where $R = R_1 + R_2 + R_3 + R_4$.

We first pick a_3, a_4 so that $a_3 C_3 - \frac{a_4}{4} < 0$. Then we pick e big enough so that $\frac{a_4 D_2}{e} - a_3 C_2 < 0$. Then we pick a_1 big enough so that $a_3 C_1 - a_1 \lambda_D < 0$. Finally, we choose a_2 big enough so that $a_4 D_1 e - a_2 \lambda_D < 0$ and

$E_{\varepsilon\perp,t}^K$ is equivalent to E_t^K . Then (70) becomes

$$\begin{aligned}
\partial_t E_{\varepsilon\perp,t}^K(\hat{h}_\theta) &\leq -\frac{K_1}{\varepsilon} \sum_{i=1}^K \left\| i^q \hat{h}_\theta^{i,\perp} \right\|_\Lambda^2 - \frac{K_2}{\varepsilon} \sum_{i=1}^K \left\| i^q \nabla_x \hat{h}_\theta^{i,\perp} \right\|_\Lambda^2 \\
&\quad - \frac{K_3}{\varepsilon} \sum_{i=1}^K \left\| i^q \nabla_v \hat{h}_\theta^{i,\perp} \right\|_\Lambda^2 - \varepsilon K_4 \sum_{i=1}^K \left\| i^q \nabla_x \hat{h}_\theta^i \right\|_\Lambda^2 + R \\
&\leq -\varepsilon \tilde{K} \left(\frac{\sum_{i=1}^K \left\| i^q \hat{h}_\theta^{i,\perp} \right\|_\Lambda^2}{\varepsilon^2} + \frac{\sum_{i=1}^K \left\| i^q \nabla_x \hat{h}_\theta^{i,\perp} \right\|_\Lambda^2}{\varepsilon^2} \right. \\
&\quad \left. + \frac{\sum_{i=1}^K \left\| i^q \nabla_v \hat{h}_\theta^{i,\perp} \right\|_\Lambda^2}{\varepsilon^2} + \sum_{i=1}^K \left\| i^q \nabla_x \hat{h}_\theta^i \right\|_\Lambda^2 \right) + R \\
&\leq -\varepsilon \tilde{K} \left(\sum_{i=1}^K \left\| i^q \hat{h}_\theta^{i,\perp} \right\|_\Lambda^2 + \sum_{i=1}^K \left\| i^q \nabla_x \hat{h}_\theta^{i,\perp} \right\|_\Lambda^2 + \sum_{i=1}^K \left\| i^q \nabla_v \hat{h}_\theta^{i,\perp} \right\|_\Lambda^2 + \sum_{i=1}^K \left\| i^q \nabla_x \hat{h}_\theta^i \right\|_\Lambda^2 \right) + R \\
&\leq -\varepsilon \tilde{K} E_t^K(\hat{h}_\theta) + R.
\end{aligned}$$

The last inequality is based on the following fact: Let

$$\begin{cases} F_1(h) = \|h^\perp\|_\Lambda^2 + \|\nabla_x h^\perp\|_\Lambda^2 + \|\nabla_v h^\perp\|_\Lambda^2 + \|\nabla_x h\|_\Lambda^2 \\ F_2(h) = \|h\|_\Lambda^2 + \|\nabla_x h\|_\Lambda^2 + \|\nabla_v h^\perp\|_\Lambda^2, \end{cases}$$

then use (25) and Poincare inequality (26), we have

$$\|h\|_\Lambda^2 \leq \|h^\perp\|_\Lambda^2 + \|\pi_{\mathcal{L}}(h)\|_\Lambda^2 \leq \|h^\perp\|_\Lambda^2 + C_\pi C_p \|\nabla_x h\|_\Lambda^2.$$

Similarly,

$$\|\nabla_x h\|_\Lambda^2 \leq \|\nabla_x h^\perp\|_\Lambda^2 + C_\pi \|\nabla_x h\|_\Lambda^2,$$

so $F_2(h) \leq C F_1(h)$ for some $C > 0$. On the other hand, using (25) and (17), we get

$$\|\nabla_v h\|_\Lambda^2 \leq \|\nabla_v h^\perp\|_\Lambda^2 + \|\nabla_v \pi_{\mathcal{L}}(h)\|_\Lambda^2 \leq \|\nabla_v h^\perp\|_\Lambda^2 + \frac{C_\pi \nu_1^\Lambda}{\nu_0^\Lambda} \|h\|_\Lambda^2.$$

Hence $\|h\|_{H_\Lambda^1}^2 \leq C' F_2(h)$ for some $C' > 0$. Finally, since Λ -norm controls the standard L^2 norm, we have $\|h\|_{H_{x,v}^1}^2 \leq C'' \|h\|_{H_\Lambda^1}^2$ for some $C'' > 0$. Combining all these inequalities, we get $\|h\|_{H_{x,v}^1}^2 \leq \tilde{C} F_1(h)$ for some $\tilde{C} > 0$.

Finally, we analyze the term R . Set $f_1 = f_2 = f_3 = f_4 = \frac{1}{f\varepsilon}$ for some $f > 0$, we have

$$R \leq \varepsilon f C_1 E_t^K(\hat{h}_\theta) + \frac{C_1}{\varepsilon f} E_t^K(\mathbf{A}_\theta) + C_2 \left(\mathcal{R}_{\theta,t,sto}^b + \mathcal{R}_{\theta,t,sto}^{b,\nabla_x} + \mathcal{R}_{\theta,t,sto}^{b,\nabla_v} \right) + C_3 \left(r_{3,t} + r_{3,t}^{\nabla_x} + r_{3,t}^{\nabla_v} \right),$$

for some $C_1, C_2, C_3 > 0$. We choose f to be small enough so that $fC_1 - \tilde{K} < 0$. Then

$$\partial_t E_{\varepsilon\perp,t}^K(\hat{h}_\theta) \leq -\varepsilon K^* E_t^K(\hat{h}_\theta) + \frac{C}{\varepsilon} \left(E_t^K(\mathbf{A}_\theta) + \mathcal{R}_{\theta,t,sto}^b + \mathcal{R}_{\theta,t,sto}^{b,\nabla_x} + \mathcal{R}_{\theta,t,sto}^{b,\nabla_v} + r_{3,t} + r_{3,t}^{\nabla_x} + r_{3,t}^{\nabla_v} \right).$$

Split $\int_{\mathbb{R}^3} = \int_{\mathcal{D}} + \int_{\mathbb{R}^3 - \mathcal{D}}$ and use Assumption 2, we have

$$\begin{aligned} \|A_\theta^i\|^2 &\leq \left\| \mathbf{d}_\theta^{1,i} \cdot \varphi M \right\|^2 + \left\| d_\theta^{2,i} \right\|^2 \\ &= \mathcal{R}_{\theta,t}^{1,i} + \mathcal{R}_{\theta,t}^{2,i} + r_{1,t}^i. \end{aligned}$$

Similarly, $\|\nabla_x A_\theta^i\|^2 \leq \mathcal{R}_{\theta,t}^{1,i,\nabla_x} + \mathcal{R}_{\theta,t}^{2,i,\nabla_x} + r_{1,t}^{\nabla_x,i}$ and $\|\nabla_v A_\theta^i\|^2 \leq \mathcal{R}_{\theta,t}^{1,i,\nabla_v} + \mathcal{R}_{\theta,t}^{2,i,\nabla_v} + r_{1,t}^{\nabla_v,i}$. So we get

$$\partial_t E_{\varepsilon \perp, t}^K(\hat{\mathbf{h}}_\theta) \leq -\varepsilon K^* E_t^K(\hat{\mathbf{h}}_\theta) + \frac{C}{\varepsilon} \tilde{R}$$

where

$$\begin{aligned} \tilde{R} &= \mathcal{R}_{\theta,t,sto}^1 + \mathcal{R}_{\theta,t,sto}^{1,\nabla_x} + \mathcal{R}_{\theta,t,sto}^{1,\nabla_v} + \mathcal{R}_{\theta,t,sto}^2 + \mathcal{R}_{\theta,t,sto}^{2,\nabla_x} + \mathcal{R}_{\theta,t,sto}^{2,\nabla_v} + \mathcal{R}_{\theta,t,sto}^b \\ &\quad + \mathcal{R}_{\theta,t,sto}^{b,\nabla_x} + \mathcal{R}_{\theta,t,sto}^{b,\nabla_v} + r_{1,t} + r_{1,t}^{\nabla_x} + r_{1,t}^{\nabla_v} + r_{3,t} + r_{3,t}^{\nabla_x} + r_{3,t}^{\nabla_v}. \end{aligned}$$

Thus by Gronwall's inequality, we have

$$\begin{aligned} E_t^K(\hat{\mathbf{h}}_\theta) &\leq \left(\frac{C}{\varepsilon} \int_0^T \tilde{R} dt + E_0^K(\hat{\mathbf{h}}_\theta) \right) e^{-\varepsilon\tau t} \\ &\leq \frac{\tilde{C}(\mathcal{R}_{\theta,sto}^{G,H^1} + \delta(\varepsilon))}{\varepsilon} e^{-\varepsilon\tau t}, \end{aligned}$$

for some $\tilde{C}, \tau > 0$ independent of ε . This completes the proof. \square

Remark 1. Use similar arguments and a Grönwall type inequality, we can also show that $E_t^K(\hat{\mathbf{h}}_\theta) \leq \tilde{C}(\mathcal{R}_{\theta,sto}^{G,H^1} + \delta(\varepsilon))$ for some $\tilde{C} > 0$ independent of ε , hence the error won't explode when $\varepsilon \ll 1$. But in this case we can't derive the exponential decay in time. This is not a real problem since $e^{-\varepsilon\tau t} \rightarrow 1$ as $\varepsilon \rightarrow 0$, hence the exponential decay becomes insignificant when we are in the fluid regime. We comment that in the acoustic scaling, we can't expect an exponential decay with decay rate independent of ε , as suggested in [27].

5. NUMERICAL EXPERIMENTS

5.1. The micro-macro decomposition method

In this section, we will numerically implement the APNN-SG method and demonstrate that it is efficient and can deliver robust results. Instead of studying the full Boltzmann (or its linearized model), we consider the linear semiconductor Boltzmann equation [22] with uncertainties and under the diffusive scaling. As mentioned in the beginning of Section 4, our analysis can be easily generalized to this linear model. The equation of the model is given by

$$\varepsilon \partial_t f + v \cdot \nabla_x f + \nabla_x \phi \cdot \nabla_v f = \frac{1}{\varepsilon} \mathcal{L}(f), \quad (71)$$

where

$$\mathcal{L}(f)(v, z) = \int_{\mathbb{R}^d} \sigma(v, w, z) (M(v)f(w, z) - M(w)f(v, z)) dw.$$

Here t and x dependence in functions are omitted. We employ the micro-macro decomposition method [26] and derive the micro-macro system for the semiconductor Boltzmann equation (71). Assume the ansatz

$$f = \Pi f + \varepsilon g, \quad (72)$$

where $g = g(t, x, v, z)$ and $\Pi f = \langle f \rangle M(v)$, with

$$\langle f \rangle = \int_{\mathbb{R}} f(t, x, v, z) \, dv = \rho(t, x, z).$$

One can derive a coupled system for ρ and g in this micro-macro decomposition framework, then

$$\begin{cases} \partial_t \rho + \nabla_x \cdot \langle vg \rangle + \nabla_x V \cdot \langle \partial_v g \rangle = 0, \\ \varepsilon^2 \partial_t g + v \cdot \nabla_x \rho M(v) + \varepsilon(\mathbb{I} - \Pi)(v \cdot \nabla_x g + \nabla_x \phi \cdot \nabla_v g) - 2v\rho \cdot \nabla_x VM(v) = \mathcal{L}(g). \end{cases} \quad (73)$$

5.2. APNN-SG method

We approximate the solutions ρ , g in (73) by the following ansatzes:

$$\rho_K^{NN}(t, x, z) = \sum_{k=0}^K \hat{\rho}_k^{NN}(t, x) \phi_k(z), \quad g_K^{NN}(t, x, v, z) = \sum_{k=0}^K \hat{g}_k^{NN}(t, x, v) \phi_k(z). \quad (74)$$

To approximate the solution of the stochastic Galerkin system, we represent the expansion coefficients using deep neural networks. Specifically, we employ two distinct sets of $K+1$ neural networks to parameterize the gPC coefficients $\{\hat{\rho}_k(t, x)\}$ and $\{\hat{g}_k(t, x, v)\}$ in the expansions (74). For the density coefficients, a physical constraint is that the leading-order term $\hat{\rho}_0$, must be non-negative. To enforce this property at the network level, we model this coefficient using a neural network $\tilde{\rho}_0^{NN}$ followed by an exponential activation function:

$$\hat{\rho}_0^{NN}(t, x) := \exp(\tilde{\rho}_0^{NN}(t, x; \theta_\rho)), \quad (75)$$

where the inputs to the network are time t and position x . The remaining coefficients, $\hat{\rho}_k^{NN}$ for $k > 0$, are standard fully-connected networks. To ensure this condition is met by construction for all coefficients, we formulate each network approximation, \hat{g}_k^{NN} , using a projection-based approach:

$$\hat{g}_k^{NN}(t, x, v) := \tilde{g}_k^{NN}(t, x, v; \theta_g) - \Pi \tilde{g}_k^{NN}. \quad (76)$$

Here, \tilde{g}_k^{NN} is a standard fully-connected neural network whose inputs include time t , position x , and velocity v . This formulation inherently satisfies the constraint for any trainable network \tilde{g}_k^{NN} , since applying the projection operator Π to \hat{g}_k^{NN} yields: $\Pi \hat{g}_k^{NN} = 0$. This architectural choice correctly embeds the physical constraints of the model directly into our neural network representation.

Conducting the SG method on the micro-macro system (73) and incorporating the initial and boundary conditions, one defines the loss function for the APNN-SG method as the following:

$$\begin{aligned} \mathcal{R}_{\text{Total}} = & \frac{1}{|\mathcal{T} \times \mathcal{D}|} \int_{\mathcal{T}} \int_{\mathcal{D}} \left\| \partial_t \hat{\rho}_\theta^{NN} + \nabla_x \cdot \langle v \hat{g}_\theta^{NN} \rangle + \nabla_x \phi \cdot \langle \nabla_v \hat{g}_\theta^{NN} \rangle \right\|^2 dx dt \\ & + \frac{1}{|\mathcal{T} \times \mathcal{D} \times \Omega|} \int_{\mathcal{T}} \int_{\mathcal{D}} \int_{\Omega} \left\| \varepsilon^2 \partial_t \hat{g}_\theta^{NN} + \varepsilon(\mathbb{I} - \Pi)(v \cdot \nabla_x \hat{g}_\theta^{NN} + \nabla_x \phi \cdot \nabla_v \hat{g}_\theta^{NN}) \right. \\ & \left. - 2v \cdot \nabla_x \phi \hat{\rho}_\theta^{NN} M(v) + v \cdot \nabla_x \hat{\rho}_\theta^{NN} M(v) - \mathbf{L}(\hat{g}_\theta^{NN}) \right\|^2 dv dx dt \\ & + \frac{\lambda_1}{|\mathcal{T} \times \partial \mathcal{D} \times \Omega|} \int_{\mathcal{T}} \int_{\partial \mathcal{D}} \int_{\Omega} \left\| \mathcal{B}(\hat{\rho}_\theta^{NN} M(v) + \varepsilon \hat{g}_\theta^{NN}) - F_B \right\|^2 dv dx dt \\ & + \frac{\lambda_2}{|\mathcal{D} \times \Omega|} \int_{\mathcal{D}} \int_{\Omega} \left\| \mathcal{I}(\hat{\rho}_\theta^{NN} M(v) + \varepsilon \hat{g}_\theta^{NN}) - f_0 \right\|^2 dv dx, \end{aligned} \quad (77)$$

where \mathcal{T} , \mathcal{D} , Ω denote the temporal, spatial and velocity domain respectively, the penalty parameters λ_1 , λ_2 are chosen for good performances. Here $\|\cdot\|$ represents the L^2 norm for vectors, and $\mathbf{L}(\hat{\mathbf{g}}_\theta^{\text{NN}})$ is shown as

$$\mathbf{L}(\hat{\mathbf{g}}_\theta^{\text{NN}}) = M(v) \int_{\mathbb{R}^d} \mathbf{B}(v, w) \hat{\mathbf{g}}_\theta^{\text{NN}}(w) dw - \mathbf{F}(v) \hat{\mathbf{g}}_\theta^{\text{NN}}(v), \quad (78)$$

with matrices $(\mathbf{B}_{ij})_{K \times K}$ and $(\mathbf{F}_{ij})_{K \times K}$ defined by

$$\mathbf{B}_{ij}(v, w) = \int_{I_z} \sigma(v, w, z) \psi_i(z) \psi_j(z) \pi(z) dz, \quad (79)$$

$$\mathbf{F}_{ij}(v) = \int_{I_z} \lambda(v, z) \psi_i(z) \psi_j(z) \pi(z) dz. \quad (80)$$

5.3. Numerical tests

In our numerical experiments, we will show one example to illustrate the robustness and accuracy of our designed APNN-SG method. We employ the APNN-SG for the semiconductor Boltzmann model with uncertainties. We compare the reference solution obtained by traditional numerical methods and APNN-SG methods in different regimes, ranging from the kinetic regime ($\varepsilon \approx O(1)$) to the diffusive regime ($\varepsilon \ll 1$), and observe that the APNN-SG can capture the multiscale nature of the model with uncertainties thanks to the design of the loss function based on the macro-micro decomposition. For the velocity discretization, the integral in v is computed by using the Hermite quadrature rule with N_v quadrature points. We adopt the Hermite polynomials in velocity discretization to provide more accurate velocity derivatives, a technique similar to the moment method [22, 40]. See details in the appendix.

Networks architecture. In our experiments, we use the feed-forward neural network (FNN) with one input layer, one output layer and 4 hidden layers with 128 neurons in each layer, unless otherwise specified. The hyperbolic tangent function (Tanh) is chosen as our activation function.

gPC basis. Since the random variable z is assumed following the uniform distribution, we use the following Legendre polynomials:

$$\begin{cases} \phi_0(z) = 1, \\ \phi_1(z) = \sqrt{3}z, \\ \phi_2(z) = \frac{\sqrt{5}}{2}(3z^2 - 1), \\ \phi_3(z) = \frac{\sqrt{7}}{2}(5z^3 - 3z), \\ \phi_4(z) = \frac{3}{8}(35z^4 - 30z^2 + 3). \end{cases} \quad (81)$$

Training settings. We train the neural network by Adam with Xavier initialization. We set epochs to be 200 000 and the learning rate to be 10^{-4} , and use full batch for most of the following experiments in the numerical experiments unless otherwise specified. All the hyper-parameters are chosen by trial and error.

Loss design. For our experiment, we consider the spatial and temporal domains to be $[0, 1]$ and $[0, 0.1]$ respectively. We choose the collocation points $\{(t_i, x_i, v_i)\}$ for $f(t, x, v)$ in the following way. For spatial points x_i , we select 99 interior points evenly spaced in $[0, 1]$. For temporal points t_i , we select 20 interior points evenly spaced in the range $[0, 0.1]$. We use the tensor product grid for the collocation points. For velocity points v_i , $N_v = 8$ points are generated by the Hermite quadrature rule. The variable z is uniformly defined on the interval $[-1, 1]$, over which the function $\sigma(z)$ is positive. For numerical evaluation, we discretize this domain using 101 uniformly spaced points. The penalty parameters in (77) are set to $(\lambda_1, \lambda_2) = (1, 1)$.

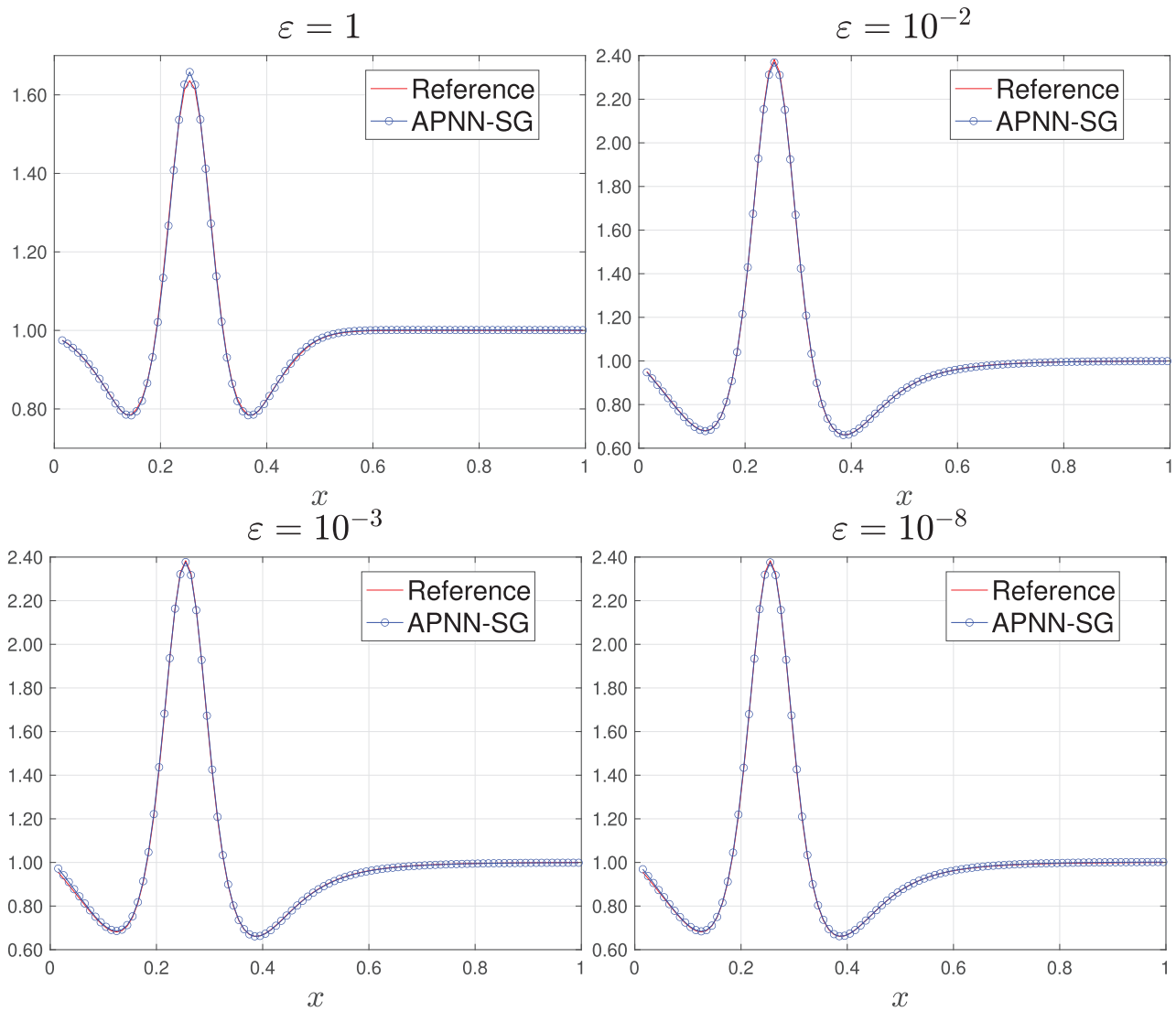


FIGURE 1. Problems I with different ε . Mean of density ρ for APNN-SG and reference solutions at $T = 0.1$.

TABLE 1. Problems I. Relative ℓ^2 error comparisons for the mean of the reference solution and APNNs with different ε at the final time at $T = 0.1$.

| ε | 1 | 10^{-2} | 10^{-3} | 10^{-8} |
|---------------|-----------------------|-----------------------|-----------------------|-----------------------|
| APNN-SG | 3.81×10^{-3} | 2.99×10^{-3} | 4.31×10^{-3} | 3.65×10^{-3} |

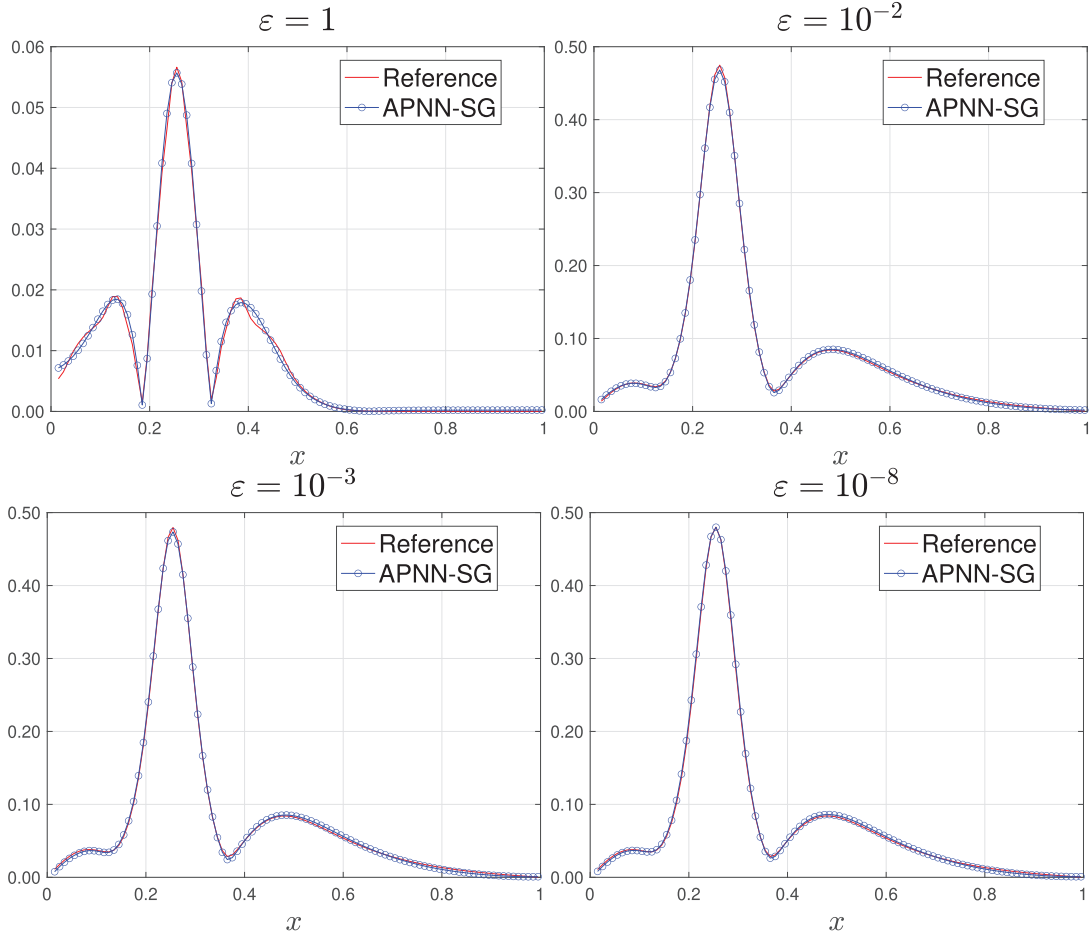


FIGURE 2. Problems I with different ε . Standard deviation of density ρ for APNN-SG and reference solutions at $T = 0.1$.

The reference solutions are obtained by traditional AP scheme [22]. We compute the relative ℓ^2 errors for the mean and standard deviation of ρ obtained from the APNN-SG method and the reference solution, which are defined by

$$\mathcal{E}_{\text{mean}}(t) = \sqrt{\frac{\sum_j |\rho_{\text{mean}}^{\text{NN}}(t, x_j) - \rho_{\text{mean}}^{\text{ref}}(t, x_j)|^2}{\sum_j |\rho_{\text{mean}}^{\text{ref}}(t, x_j)|^2}}, \quad \mathcal{E}_{\text{std}}(t) = \sqrt{\frac{\sum_j |\rho_{\text{std}}^{\text{NN}}(t, x_j) - \rho_{\text{std}}^{\text{ref}}(t, x_j)|^2}{\sum_j |\rho_{\text{std}}^{\text{ref}}(t, x_j)|^2}}. \quad (82)$$

Here the mean and standard deviation of $\rho(t, x, z)$ over a set of M data points $\{z_i\}_{i=1}^M$ are given by

$$\rho_{\text{mean}}(t, x) = \frac{1}{M} \sum_{i=1}^M \rho(t, x, z_i), \quad \rho_{\text{std}}(t, x) = \sqrt{\frac{1}{M} \sum_{i=1}^M (\rho(t, x, z_i) - \rho_{\text{mean}}(t, x))^2}. \quad (83)$$

Uncertainty in the collision kernel. Let $x \in [0, 1]$, the potential $V = e^{-50\exp(1)(1/4-x)^2}$ and the scattering coefficient $\sigma(v, w, z) = 10 + 9z$. In this problem, we assume $z \in [-1, 1]$ following the uniform distribution. Consider the initial data $f(t = 0, x, v, z) = \frac{1}{\sqrt{\pi}} e^{-v^2}$ and incoming boundary conditions in space.

TABLE 2. Problems I. Relative ℓ^2 error comparisons for standard deviation of the reference solution and APNN-SG with different ε at the final time at $T = 0.1$.

| ε | 1 | 10^{-2} | 10^{-3} | 10^{-8} |
|---------------|-----------------------|-----------------------|-----------------------|-----------------------|
| APNN-SG | 4.76×10^{-2} | 1.27×10^{-2} | 1.65×10^{-2} | 1.75×10^{-3} |

We will establish two networks to learn the coefficients $\hat{\rho}_k(t, x)$ and $\hat{g}_k(t, x, v)$ respectively and use the Legendre polynomials as our basis. In numerical experiments, we consider the cases when $K = 4$. We construct five different neural networks to approximate each gPC coefficient in $\hat{\rho}$ and \hat{g} respectively. Figure 1 shows the mean of $\rho(t, \mathbf{x}, z)$ under different ε . To qualitatively demonstrate the performance of the APNN-SG, we report the relative ℓ^2 errors for the test sets generated by APNN-SG in Table 1, where we compute the ℓ^2 errors of the density between APNN-SG solution and the reference solution. Besides the mean of density, we further study the standard deviation. Figure 2 shows the standard deviation of $\rho(t, x, z)$ under different ε . We observe that APNN-SG can capture an accurate solution compared to the reference when ε is a relatively large value, such as $\varepsilon = 1$, and when ε is extremely small, such as $\varepsilon = 10^{-8}$, thanks to the AP property of the loss function in APNN-SG (Tab. 2).

6. CONCLUSION

In this paper, we construct an asymptotic-preserving neural network (APNN) for the linearized Boltzmann equation in the acoustic scaling with uncertainties. We employ the micro-macro decomposition method and build the APNN loss function based on the micro-macro system. Rigorous analysis have been conducted to show the existence of neural networks when the APNN loss goes to zero, and the convergence of the neural network approximated solution when the loss function tends to zero. In numerical examples, we demonstrate that the proposed APNN-SG method works efficiently for the linear semiconductor Boltzmann equation with uncertainties and under the diffusive scaling. In the future work, we will further improve our analysis result by incorporating the Barron-type functions [13] and develop posterior estimates for the neural network approximations, and implement the APNN-based method for higher-dimensional problem and the full Boltzmann equation with uncertainties.

FUNDING

L. Liu acknowledges the support by National Key R&D Program of China (2021YFA1001200), Ministry of Science and Technology in China, General Research Fund (14303022, 14301423 & 14307125) funded by Research Grants Council of Hong Kong.

DATA AVAILABILITY STATEMENT

The research data associated with this article are included in the article.

REFERENCES

- [1] E. Abdo, L. Chai, R. Hu and X. Yang, Error estimates of physics-informed neural networks for approximating Boltzmann equations. Preprint [arXiv:2407.08383](https://arxiv.org/abs/2407.08383) (2024).
- [2] G. Albi, M. Herty and L. Pareschi, Kinetic description of optimal control problems and applications to opinion consensus. *Commun. Math. Sci.* **13** (2015) 1407–1429.
- [3] R. Alonso, Boltzmann-type equations and their applications. Publicações Matemáticas do IMPA. [IMPA Mathematical Publications]. Instituto Nacional de Matemática Pura e Aplicada (IMPA), Rio de Janeiro (2015). 30o Colóquio Brasileiro de Matemática. [30th Brazilian Mathematics Colloquium]
- [4] G. Bertaglia, C. Lu, L. Pareschi and X. Zhu, Asymptotic-preserving neural networks for multiscale hyperbolic models of epidemic spread. *Math. Models Methods Appl. Sci.* **32** (2022) 1949–1985.

- [5] M. Briant, From the Boltzmann equation to the incompressible Navier–Stokes equations on the torus: a quantitative error estimate. *J. Differ. Equ.* **259** (2015) 6072–6141.
- [6] R. Cafilisch, D. Silant'ev and Y. Yang, Adjoint DSMC for nonlinear Boltzmann equation constrained optimization. *J. Comput. Phys.* **439** (2021) 110404.
- [7] M. Caponigro, M. Fornasier, B. Piccoli and E. Trélat, Sparse stabilization and optimal control of the Cucker–Smale model. *Math. Control Relat. Fields* **3** (2013) 447–466.
- [8] Y. Chen, L. Lu, G.E. Karniadakis and L. Dal Negro, Physics-informed neural networks for inverse problems in nano-optics and metamaterials. *Opt. Express* **28** (2020) 11618–11633.
- [9] Y. Cheng, I.M. Gamba and K. Ren, Recovering doping profiles in semiconductor devices with the Boltzmann–Poisson model. *J. Comput. Phys.* **230** (2011) 3391–3412.
- [10] M. Choulli and P. Stefanov, Inverse scattering and inverse boundary value problems for the linear Boltzmann equation. *Comm. Part. Differ. Equ.* **21** (1996) 763–785.
- [11] E.S. Daus, S. Jin and L. Liu, Spectral convergence of the stochastic Galerkin approximation to the Boltzmann equation with multiple scales and large random perturbation in the collision kernel. *Kinetic Related Models* **12** (2019) 909–922.
- [12] G. Dimarco and L. Pareschi, Numerical methods for kinetic equations. *Acta Numer.* **23** (2014) 369–520.
- [13] Y. Gu and M.K. Ng, Deep adaptive basis Galerkin method for high-dimensional evolution equations with oscillatory solutions. *SIAM J. Sci. Comput.* **44** (2022) A3130–A3157.
- [14] Y. Guo, N. Jiang and J. Jang, Acoustic limit for the Boltzmann equation in optimal scaling. *Commun. Pure Appl. Math.* **63** (2010) 337–361.
- [15] B. Hajek, *Random Processes for Engineers*. Cambridge University Press (2015).
- [16] J. Hu and S. Jin, A stochastic Galerkin method for the Boltzmann equation with uncertainty. *J. Comput. Phys.* **315** (2016) 150–168.
- [17] J. Hu and K. Qi, A fast Fourier spectral method for the homogeneous Boltzmann equation with non-cutoff collision kernels. *J. Comput. Phys.* **423** (2020) 109806.
- [18] J. Hu, L. Pareschi and Y. Wang, Uncertainty quantification for the BGK model of the Boltzmann equation using multilevel variance reduced Monte Carlo methods. *SIAM/ASA J. Uncertain. Quantif.* **9** (2021) 650–680.
- [19] J. Hu, K. Qi and T. Yang, A new stability and convergence proof of the Fourier–Galerkin spectral method for the spatially homogeneous Boltzmann equation. *SIAM J. Numer. Anal.* **59** (2021) 613–633.
- [20] S. Jin, Mathematical analysis and numerical methods for multiscale kinetic equations with uncertainties, in *Proceedings of the International Congress of Mathematicians – Rio de Janeiro 2018*. Vol. IV. Invited lectures. World Science Publications, Hackensack, NJ (2018) 3611–3639.
- [21] S. Jin, Asymptotic-preserving schemes for multiscale physical problems. *Acta Numer.* **31** (2022) 415–489.
- [22] S. Jin and L. Pareschi, Discretization of the multiscale semiconductor Boltzmann equation by diffusive relaxation schemes. *J. Comput. Phys.* **161** (2000) 312–330.
- [23] S. Jin and L. Pareschi, editors, *Uncertainty Quantification for Hyperbolic and Kinetic Equations*. *SEMA-SIMAI Springer Series*. Vol. 14. Springer (2017).
- [24] 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** (2015) 35–52.
- [25] S. Jin, Z. Ma and K. Wu, Asymptotic-preserving neural networks for multiscale time-dependent linear transport equations. *J. Sci. Comput.* **94** (2023) 57.
- [26] 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** (2008) 334–368.
- [27] 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. *Multiscale Model. Simul.* **16** (2018) 1085–1114.
- [28] L. Liu and K. Qi, Convergence of the Fourier–Galerkin spectral method for the Boltzmann equation with uncertainties. *Commun. Math. Sci.* **22** (2024) 1897–1925.
- [29] L. Liu and K. Qi, Spectral convergence of a semi-discretized numerical system for the spatially homogeneous Boltzmann equation with uncertainties. *SIAM/ASA J. Uncertain. Quantif.* **12** (2024) 812–841.
- [30] L. Liu, Y. Wang, X. Zhu and Z. Zhu, Asymptotic-preserving neural networks for the semiconductor Boltzmann equation and its application on inverse problems. *J. Comput. Phys.* **523** (2025) 113669.
- [31] Q. Lou, X. Meng and G.E. Karniadakis, Physics-informed neural networks for solving forward and inverse flow problems via the Boltzmann–BGK formulation. *J. Comput. Phys.* **447** (2021) 110676.

- [32] Z. Mao, A.D. Jagtap and G.E. Karniadakis, Physics-informed neural networks for high-speed flows. *Comput. Methods Appl. Mech. Eng.* **360** (2020) 112789.
- [33] P.A. Markowich, C.A. Ringhofer and C. Schmeiser, *Semiconductor Equations*. Springer-Verlag, Vienna (1990).
- [34] C. Mouhot and L. Neumann, Quantitative perturbative study of convergence to equilibrium for collisional kinetic models in the torus. *Nonlinearity* **19** (2006) 969–998.
- [35] C. Mouhot and L. Pareschi, Fast algorithms for computing the Boltzmann collision operator. *Math. Comput.* **75** (2006) 1833–1852.
- [36] C. Mouhot, L. Pareschi and T. Rey, Convolutional decomposition and fast summation methods for discrete-velocity approximations of the Boltzmann equation. *ESAIM: Math. Model. Numer. Anal. (M2AN)* **47** (2013) 1515–1531.
- [37] L. Pareschi, An introduction to uncertainty quantification for kinetic equations and related problems, in *Trails in Kinetic Theory. SEMA SIMAI Springer Series*. Vol. 25. Springer, Cham (2021) 141–181.
- [38] L. Pareschi and M. Zanella, Monte Carlo stochastic Galerkin methods for the Boltzmann equation with uncertainties: space-homogeneous case. *J. Comput. Phys.* **423** (2020) 109822.
- [39] M. Raissi, P. Perdikaris and G.E. Karniadakis, Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **378** (2019) 686–707.
- [40] C. Schmeiser and A. Zwirchmayr, Convergence of moment methods for linear kinetic equations. *SIAM J. Numer. Anal.* **36** (1999) 74–88.
- [41] C. Villani, A review of mathematical topics in collisional kinetic theory, in *Handbook of Mathematical Fluid Mechanics*, edited by S. Friedlander, D. Serre. Vol. I. North-Holland (2002) 71–305.
- [42] E. Weinan, The dawning of a new era in applied mathematics. *Notices Am. Math. Soc.* **68** (2021) 565–571.
- [43] D. Xiu, *Numerical Methods for Stochastic Computations*. Princeton University Press, New Jersey (2010).



Please help to maintain this journal in open access!

This journal is currently published in open access under the Subscribe to Open model (S2O). We are thankful to our subscribers and supporters for making it possible to publish this journal in open access in the current year, free of charge for authors and readers.

Check with your library that it subscribes to the journal, or consider making a personal donation to the S2O programme by contacting subscribers@edpsciences.org.

More information, including a list of supporters and financial transparency reports, is available at <https://edpsciences.org/en/subscribe-to-open-s2o>.

APPENDIX A. VELOCITY DISCRETIZATIONS

For completeness, we briefly mention the velocity discretization similar to what has been studied in [22]. Set $f(t, x, v) = \psi(t, x, v)M(v)$, where $M(v) = \frac{1}{\sqrt{\pi}}e^{-v^2}$, with

$$\psi(t, x, v) = \sum_{k=0}^N \psi_k(t, x) \tilde{H}_k(v), \quad (\text{A.1})$$

being the Hermite expansion. For notation simplicity, we omit the t and x dependence of functions below. Here \tilde{H}_k are the renormalized Hermite polynomials defined as $\tilde{H}_{-1} = 0$, $\tilde{H}_0 = 1/\pi^{1/4}$ and

$$\tilde{H}_{j+1} = v\sqrt{\frac{2}{j+1}}\tilde{H}_j - \sqrt{\frac{j}{j+1}}\tilde{H}_{j-1} \quad \text{for } j \geq 0,$$

satisfying $\partial_v \tilde{H}_j = \sqrt{2j}\tilde{H}_{j-1}$. The inverse Hermite expansion is given by

$$\psi_k = \sum_{j=0}^N \psi(v_j) \tilde{H}_k(v_j) w_j, \quad (\text{A.2})$$

where (v_j, w_j) are the points and corresponding weights of the Gauss–Hermite quadrature rule. Thus the collision operator \mathcal{L} in (71) can be computed by

$$\mathcal{L}(f)(v) = M(v) \sum_{j=0}^N \sigma(v, v_j) \psi(v_j) w_j - \lambda(v)f(v),$$

with $\lambda(v) = \sum_{j=0}^N \sigma(v, v_j) w_j$.

APPENDIX B. PARAMETRIZATION OF RANDOM INPUTS

In this appendix, we discuss how to manage the randomness existing in our model. The key step is to parameterize the random inputs by a finite set of independent random variables to make computational simulations plausible. If the random inputs are already given in the form of finitely many random parameters with a proper probability distribution, *e.g.*, jointly Gaussian, then parametrization is straightforward, *e.g.*, using Cholesky decomposition. However, in many cases, the random inputs are formulated by random processes, which are often characterized by a continuous index $t \in T$. Then we need to apply dimension blackuction techniques to approximate the processes using finitely many random variables. One of the most widely used techniques in this regard is the Karhunen–Loeve (KL) expansion, see [15, 43].

For a random process $\{Y_t(\omega)\}_{t \in T}$ (we use the notation Y_t as an abbreviation) with mean $\mu_Y(t)$ and autocorrelation function $R_Y(t, s) = \mathbb{E}[Y_t Y_s]$, its KL expansion, if exists, is given by

$$Y_t(\omega) = \sum_{i=1}^{\infty} \psi_i(t) Y_i(\omega), \quad (\text{B.1})$$

where the series converges in the mean square sense (m.s.), $\{\psi_i\}$ is an orthonormal family of functions in $L^2(T)$ and Y_i 's are mutually orthogonal, *i.e.*, $\mathbb{E}[Y_i Y_j] = 0 \quad \forall i \neq j$. The existence of KL expansion for Y_t is guaranteed by Mercer's theorem, provided that the random process Y_t is m.s. continuous, *i.e.*, $R_Y(t, s)$ is continuous over $T \times T$.

We now analyze how to derive ψ_i and Y_i if the KL expansion for Y_t exists. The KL expansion for Y_t can be viewed as an analogue of decomposing $f \in L^2(T)$ with respect to an orthonormal family $\{\psi_i\}$, *i.e.*,

$$f = \sum_{i=1}^{\infty} f_i \psi_i,$$

where $f_i = \int_T f \psi_i dt \in \mathbb{R}$ are the Fourier coefficients. In the KL setting, once an orthonormal family $\{\psi_i\}$ is chosen, we can define the “Fourier coefficients” Y_i analogously by

$$Y_i(\omega) = \int_T Y_t(\omega) \psi_i(t) dt. \quad (\text{B.2})$$

The only difference is that the coefficients for the KL expansion are random variables instead of real numbers. The challenge now is to pick $\{\psi_i\}$ properly so that the coefficients $\{Y_i\}$ are mutually orthogonal. This can be accomplished by the following lemma, whose proof is included in [15].

Lemma B.1. *Suppose Y_t is m.s. continuous and (B.1) holds for Y_t with $\{\psi_i\}$ orthonormal and $\{Y_i\}$ not necessarily mutually orthogonal. Then it is a KL expansion (*i.e.*, Y_i 's are mutually orthogonal) if and only if ψ_i 's are eigenfunctions of R_Y :*

$$R_Y(\psi_i) = \lambda_i \psi_i,$$

where R_Y is an operator on $L^2(T)$ given by $R_Y(\psi)(t) = \int_T R_Y(t, s) \psi(s) ds$ for any $\psi(t) \in L^2(T)$. In case (B.1) is a KL expansion, the eigenvalues are given by $\lambda_i = \mathbb{E}[|Y_i|^2]$.

In summary, if Y_t is m.s. continuous, its KL expansion can be established by first solving an eigenvalue problem related to the autocorrelation function $R_Y(t, s)$ to obtain an orthonormal family $\{\psi_i\}$, followed by a computation of the “Fourier coefficients” $Y_i(\omega)$ associated to this orthonormal family. Once the KL expansion is established, the analysis of Y_t can be naturally transformed into the analysis of the coefficients Y_i .

For practical purposes, we need to truncate the series appearing in (B.1) to obtain a finite dimensional parametrization of the random process, *i.e.*,

$$Y_t(\omega) \approx \sum_{i=1}^d \psi_i(t) Y_i(\omega), \quad d \geq 1. \quad (\text{B.3})$$

In most situations, the eigenvalues λ_i as appeablack in Lemma B.1 will decay as i increases. Hence we can choose the truncation order d based on the decay rate of the eigenvalues. For more details, we refer the readers to [43]. Once (B.3) is established, we can represent the random process Y_t by finitely many orthogonal random variables Y_i as we desire. Note that in general Y_i 's are not mutually independent, unless additional assumptions on Y_t are made, *e.g.*, Y_t is a Gaussian process with zero mean. We will not pursue further in this direction and shall be content with finite representation of Y_t by orthogonal random variables. Some remarks are in order.

Remark B.1. The condition of m.s. continuity to guarantee the existence of a KL expansion is not very restrictive. Many random processes we use for modeling, *e.g.*, Brownian motion and Poisson process, satisfy this property.

Remark B.2. We assume the distribution of the random process Y_t is prescribed. Hence, we can derive the probability distributions of Y_i using (B.2). This is especially straightforward and useful when Y_t is a Gaussian process.

Remark B.3. The truncated KL expansion (B.3) identifies the “most accurate” d -dimensional approximation of Y_t in the sense that it minimizes $\mathbb{E}[\|Y_t - Z_t\|^2]$ over all d -dimensional random processes Z_t . A random process Z_t is said to be d -dimensional if it has the form $Z_t(\omega) = \sum_{i=1}^d \phi_i(t) Z_i(\omega)$ for any d random variables Z_1, \dots, Z_d and functions ϕ_1, \dots, ϕ_d .