

## SOLVING STATIONARY NONLINEAR FOKKER–PLANCK EQUATIONS VIA SAMPLING\*

LEI LI<sup>†</sup>, YIJIA TANG<sup>‡</sup>, AND JINGTONG ZHANG<sup>§</sup>

**Abstract.** Solving the stationary nonlinear Fokker–Planck equations is important in applications and examples include the Poisson–Boltzmann equation and the two-layer neural networks. Making use of the connection between the interacting particle systems and the nonlinear Fokker–Planck equations, we propose solving the stationary solution by sampling from the  $N$ -body Gibbs distribution. This avoids simulation of the  $N$ -body system for a long time and the requirement of uniform propagation of chaos from direct simulation of the particle systems. While the sampling strategy could be used for any given temperature, we establish the convergence of the Gibbs measure to the stationary solution when the interaction kernel is bounded (not necessarily continuous) and the temperature is high enough. Numerical experiments are performed for the Poisson–Boltzmann equation and the two-layer neural networks to validate the method and the theory.

**Key words.** McKean–Vlasov equation, Gibbs distribution, mean-field limit, random batch Monte Carlo method

**MSC codes.** 65C05, 65C40, 65N75

**DOI.** 10.1137/23M1616054

**1. Introduction.** The nonlinear Fokker–Planck equations describe the evolution of laws for the McKean–Vlasov processes [47, 21], which often describe the evolution of distributions under the self-consistent mean field [31]. A famous example is the Poisson–Nernst–Planck (PNP) model, which describes the motion of charged chemical species in a fluid medium and is important in electrochemistry [6, 19]. In this paper, our focus is on solving the stationary solutions of the nonlinear Fokker–Planck equations. For a single species, the dynamic nonlinear Fokker–Planck equation may be written as

$$(1.1) \quad \partial_t \rho = \nabla \cdot (\rho (\nabla U + \nabla (W * \rho))) + \beta^{-1} \Delta \rho,$$

and thus the stationary equation we are concerned with is given by

$$(1.2) \quad \nabla \cdot (\rho (\nabla U + \nabla (W * \rho))) + \beta^{-1} \Delta \rho = 0.$$

Here,  $U : \mathbb{R}^d \rightarrow \mathbb{R}$  is some external potential field,  $W : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$  is a symmetric (i.e.,  $W(x, y) = W(y, x)$ ) interacting potential,  $\rho$  is the species distribution,  $\beta$  is the inverse temperature. Throughout this paper, the operation  $*$  is defined as

\*Received by the editors November 13, 2023; accepted for publication (in revised form) August 27, 2024; published electronically February 4, 2025.

<https://doi.org/10.1137/23M1616054>

**Funding:** This work was financially supported by the National Key R&D Program of China, Projects 2021YFA1002800 and 2020YFA0712000. The work of the first author was partially supported by NSFC 12371400 and 12031013, Shanghai Municipal Science and Technology Major Project 2021SHZDZX0102, Shanghai Science and Technology Commission (grants 21JC1403700, 21JC1402900), and the Strategic Priority Research Program of Chinese Academy of Sciences, grant XDA25010403. The work of the second author was funded by the Priority Programme SPP2298 “Theoretical Foundations of Deep Learning” by the Deutsche Forschungsgemeinschaft (DFG).

<sup>†</sup>School of Mathematical Sciences, Institute of Natural Sciences, MOE-LSC, Shanghai Jiao Tong University, Shanghai, 200240, China, and Shanghai AI Laboratory, Shanghai, 200030, China (leili2010@sjtu.edu.cn).

<sup>‡</sup>Scientific Computing Center, Karlsruhe Institute of Technology, Karlsruhe, 76131, Germany (yijia.tang@kit.edu).

<sup>§</sup>School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai, 200240, China (sffred@sjtu.edu.cn).

$$(1.3) \quad W * \rho := \int_{\mathbb{R}^d} W(x, y) \rho(y) dy$$

with the standard convolution as a special case when  $W(x, y) = W(x - y)$ . The multispecies case can be similarly written out. For example, if there are two species, the stationary equations could be written as

$$(1.4) \quad \begin{aligned} \nabla \cdot (\rho_1 (\nabla U_1 + \nabla(W_1 * \rho_1) + \nabla(W_c * \rho_2))) + \beta^{-1} \Delta \rho_1 &= 0, \\ \nabla \cdot (\rho_2 (\nabla U_2 + \nabla(W_2 * \rho_2) + \nabla(W_c * \rho_1))) + \beta^{-1} \Delta \rho_2 &= 0. \end{aligned}$$

Here,  $U_i$ 's are the external potentials for the two species,  $W_i$ 's are the interaction potentials within the species, and  $W_c$  is the cross species interaction,  $W_i$ 's and  $W_c$  are assumed to be symmetric. See section 4 for the details.

The nonlinear Fokker–Planck equations are naturally associated with the interacting particle systems at the microscopic level, which are ubiquitous in applications such as molecular dynamics [20], flocking and swarming [11, 13], chemotaxis [29], clustering [28], and consensus [50]. The models consisting of exchangeable particles (one species) may be described by the first order interacting particle systems, i.e., for  $i = 1, \dots, N$ ,

$$(1.5) \quad dX_i = -\nabla U(X_i) dt - \frac{1}{N-1} \sum_{j=1, j \neq i}^N \nabla_1 W(X_i, X_j) dt + \sqrt{\frac{2}{\beta}} dB_i.$$

$\nabla_1 W(x, y)$  means the derivative of  $W$  w.r.t. the first argument  $x$ . The particle systems for multispecies could be similarly written out; see section 4 for details. Here, we will call the label  $X_i$  the “position”. The concrete significance may be other quantities (for example,  $X_i$  represents the opinion in the opinion dynamics and may represent the velocity in the homogeneous Landau equation). The processes  $B_i \in \mathbb{R}^d$  are  $N$  independent standard Brownian motions. The initial data  $X_i^0$ 's are i.i.d. sampled from some distributions. There are many models that are described by the second order interacting particle systems if the “velocity” is also considered (see section 2.4). In some applications, the first order systems arise as the zero-inertia/overdamped limit of the second order Langevin systems. As remarked in section 2.4, when we consider the stationary distributions, the second order systems are not very special. In this work, we will consider first order systems as the primary examples.

The factor  $1/(N-1)$  indicates that we focus on the regime (by choosing suitable units for physical quantities) where the interaction energy is comparable to the external energy, and the total mass of the particles is of order 1. This scaling is crucial for the connection between the particle system (1.5) and the nonlinear Fokker–Planck equation (1.1). In this regime, as  $N \rightarrow \infty$ , one particle is expected to feel a mean field given by  $-\nabla U - \nabla(W * \rho)$ , and the motions of two given particles move independently. The phenomenon that the motions of certain particles tend to be independent copies under the mean field is known as the “propagation of chaos”. The term “propagation” means that the chaotic configuration is kept from the initial time  $t = 0$ . This is the classical mean field limit, and this regime may be called the mean field regime. The nonlinear Fokker–Planck equations (1.1) thus describe the mean field limits of the interacting particle systems. The rigorous justification of the mean field limit and the propagation of chaos could date back to Kac and McKean [38, 47]; see [15, 55] for some classical works and see [8] for a review. Currently, this is still a popular research topic, especially for singular potentials [54, 33, 5]. Most of these results are for finite time  $T$ . In various practical applications, such as the Poisson–Boltzmann

(PB) equation [23, 9] and the two-layer neural network [48], we are more interested in the stationary solution of the nonlinear Fokker–Planck equations. In fact, the PB equation describes the stationary form of the PNP model. Classical discretization using finite difference or finite element may not be convenient for irregular domain and suffer from curse of dimensionality. A popular approach is to simulate the corresponding interacting particle systems and approximate the solution of the nonlinear PB equation by particle densities [40]. This requires the simulation for sufficiently long time to reach equilibrium. Moreover, theoretically, this needs the uniform propagation of chaos, which needs a lot of additional requirements [45, 26, 25]. These results often need the potentials to be convex in some sense [26, 45, 46] or to be considered on torus [25] and thus are limited in applications.

We are then motivated to adopt an alternative approach, where we refrain from directly simulating the  $N$ -particle systems, and instead to do sampling from the  $N$ -body Gibbs distribution of the particle systems. Then the samples can be regarded as the approximation to the stationary solutions. Hence, the main idea of this paper is solving the stationary solution by sampling from the  $N$ -body Gibbs distribution. In this sense, the requirement of the uniform-in-time of propagation of chaos has been changed to the requirement of the ergodicity of the sampling methods, which is relatively easier to obtain as one can design the sampling dynamics with mixing properties. Moreover, the justification of the convergence from the stationary Gibbs distribution to the nonlinear Fokker–Planck equations may avoid the strong requirements of the uniform propagation of chaos and the result here is valid in  $\mathbb{R}^d$ . There are some existing results on the convergence of the  $N$ -body Gibbs measure to the corresponding limit. Arous and Brunaud [1] and Arous and Zeitouni [2] used the relative entropy to study the convergence of the Gibbs distribution in the mean-field limit. The proof is based on the Laplace asymptotics and requires the continuity and boundedness of the free energy functional. Moreover, they assumed the interaction kernel to be continuous and required information on the second variation of the free energy functional. Provided that the interaction kernel  $W$  is bounded and continuous, using the standard large deviation principle, one can also establish the convergence of the empirical measure to the minimizer of mean field free energy without the second variation of the free energy, but there is not explicit convergence rate given (see, for example, the discussion in [7]). When  $W$  is not continuous, the large deviation argument fails.

In this paper, similar to [1, 2] and inspired by the recent works for propagation of chaos [32, 34, 54], we will use the tool of relative entropy to gauge the discrepancy between the joint law and the tensor product of  $N$  copies of the stationary solution. Based on the approach in [32, 34, 54] and a symmetrization technique, we prove that if  $\beta$  is small enough (temperature is large enough), both the first marginal distribution and empirical measure for the  $N$ -body Gibbs measure converge to the unique stationary distribution of the nonlinear Fokker–Planck equation as  $N \rightarrow \infty$ . Our proof does not rely on the continuity of  $W$  and we need only boundedness of  $W$ . More importantly, the dependence on the interaction kernel and the temperature in the bound of the rescaled relative entropy can be made explicit and thus an explicit rate of convergence can be given. In addition, we also establish the results for multiple species cases. The results can thus be used to justify the validity of solving the stationary nonlinear Fokker–Planck equations using sampling algorithms. We remark that the method, in principle, can be used for any temperature though which solution the  $N$ -body Gibbs distribution converges to is unclear when there is phase transition. Moreover, the sampling method does not have to be exact and we actually allow some

bias that is within the tolerance. In the numerical examples in this paper, we will use the Random Batch Monte Carlo method (RBMC) [43], which has been shown to be efficient in practice with some bias introduced by the random batch strategy.

The rest of this paper is organized as follows. Section 2 is devoted to the introduction of the basic properties of the  $N$ -body Gibbs measure for the particle system (1.5) and the stationary solution to the nonlinear Fokker–Planck equation (1.1). In section 3, we show the convergence of the  $N$ -body Gibbs measure to the stationary solution of the nonlinear Fokker–Planck equation for one species, which illustrates our main methodology and lays the foundation of sampling for solving the nonlinear Fokker–Planck equations. In section 4, we consider the multispecies cases. The convergence is proved for the two-species case as the typical example. In section 5, we perform numerical experiments for two typical applications, namely, the PB equations and the two-layer neural network, where sampling from the  $N$ -body Gibbs distribution is performed using RBMC. Conclusions follow in section 6.

## 2. Setup and properties of the Gibbs measure and stationary solutions.

In this section, we first collect some basic facts and properties for the Gibbs measure of the interacting particle systems in section 2.1 and the stationary solution of the nonlinear Fokker–Planck equation (1.1), namely (1.2) in section 2.2. Then, the idea of sampling to solve the stationary nonlinear Fokker–Planck equations is explained in section 2.3. We perform a discussion on the second order systems in section 2.4. For the clarity of the presentation, we focus only on the one species case here. The multispecies case will be studied in section 4.

**2.1. The Gibbs measure of the interacting particle system.** The interacting particle system (1.5) is associated with the following energy functional:

$$(2.1) \quad \begin{aligned} E_N(x_1, \dots, x_N) &= \sum_{i=1}^N U(x_i) + \frac{1}{(N-1)} \sum_{1 \leq i < j \leq N} W(x_i, x_j) \\ &= \sum_{i=1}^N U(x_i) + \frac{1}{2(N-1)} \sum_{i, j: i \neq j} W(x_i, x_j). \end{aligned}$$

Note that we do not call this “Hamiltonian” as called in other literature because this is a functional of the spatial variables only without the conjugate variables (i.e., the momentum). Then, (1.5) can be written as

$$(2.2) \quad dX_i = -\nabla_{X_i} E_N dt + \sqrt{\frac{2}{\beta}} dB_i,$$

or, in terms of  $X := (X_1, \dots, X_N)$  and  $\mathbf{B} := (B_1, \dots, B_N)$ ,

$$(2.3) \quad dX = -\nabla_X E_N dt + \sqrt{2/\beta} d\mathbf{B}.$$

*Remark 2.1.* If we think that each particle has weight  $1/N$ , as considered in [7] the energy may be given by

$$(2.4) \quad \tilde{E}_N(x_1, \dots, x_N) = \frac{1}{N} \sum_{i=1}^N U(x_i) + \frac{1}{N(N-1)} \sum_{1 \leq i < j \leq N} W(x_i, x_j).$$

With this scaling, the (scaled) inverse temperature  $\tilde{\beta}$  should be like  $\tilde{\beta} = N\beta$  to be consistent with (2.2) (the time should be rescaled as well). Another different scaling has been used in [39], where the energy used is roughly  $NE_N = N^2 \tilde{E}_N$ .

The joint distribution of  $(X_1, \dots, X_N)$  is known to satisfy the following  $N$ -body Fokker–Planck equation [4]:

$$(2.5) \quad \partial_t \rho_N = \nabla_x \cdot (\nabla_x E_N \rho_N) + \beta^{-1} \Delta_x \rho_N,$$

where the derivative is taken with respect to  $\mathbf{x} := (x_1, \dots, x_N)$ . This linear Fokker–Planck is also connected to the following free energy for the  $N$ -body system:

$$(2.6) \quad \mathcal{F}_N(\rho_N) := \int_{\mathbb{R}^{Nd}} E_N \rho_N dx + \beta^{-1} \int_{\mathbb{R}^{Nd}} \rho_N \log \rho_N dx.$$

Physically, this is the total energy plus the entropy. Equation (2.5) is the gradient flow of (2.6) under the Wasserstein distance. We recall that the  $W_2$  gradient flow of the free energy functional  $F$  is given by [37]:

$$(2.7) \quad \partial_t \rho = \nabla \cdot \left( \rho \nabla \frac{\delta F}{\delta \rho} \right).$$

Based on this, the following is standard.

LEMMA 2.2 (Gibbs distribution). *Suppose that  $\exp(-\beta E_N) \in L^1(\mathbb{R}^{Nd})$ . Then, the  $N$ -body system (1.5) has a unique invariant probability measure, which is the  $N$ -body Gibbs measure given by*

$$(2.8) \quad \begin{aligned} \rho_N(\mathbf{x}) &= \bar{Z}_N^{-1} \exp(-\beta E_N) \\ &= \bar{Z}_N^{-1} \exp \left( -\beta \left( \sum_i U(x_i) + \frac{1}{2(N-1)} \sum_{i,j:i \neq j} W(x_i, x_j) \right) \right), \end{aligned}$$

where  $\bar{Z}_N$  is the normalizing factor such that the integral of  $\rho_N$  is 1. This is the unique minimizer of the free energy (2.6).

It is clear that any (local) minimizer of  $\mathcal{F}_N$  is a stationary solution of (2.5) (but not vice versa). It is easy to verify that (2.8) is a stationary solution of (2.5). Any stationary distribution  $\tilde{\rho}_N$  must have a full support in  $\mathbb{R}^{Nd}$  and be absolutely continuous with respect to the Lebesgue measure [4, Corollary 3.5.8]. The uniqueness of the stationary solution follows then from [4, Theorem 4.1.11]. This then implies that the free energy has a unique minimizer, which is clear because the free energy functional  $\mathcal{F}_N$  is convex with respect to the linear structure.

## 2.2. The stationary solution of the nonlinear Fokker–Planck equation.

In this subsection, we investigate the stationary solution of the nonlinear Fokker–Planck equations. Similar to the  $N$ -particle case, the nonlinear Fokker–Planck equation (1.1) is associated with the mean-field free energy

$$(2.9) \quad \mathcal{F}(\rho) := \int_{\mathbb{R}^d} U \rho dx + \frac{1}{2} \int_{\mathbb{R}^d} \rho(W * \rho) dx + \beta^{-1} \int_{\mathbb{R}^d} \rho \log \rho dx.$$

In fact, it is the gradient flow of  $\mathcal{F}$  under the  $W_2$  distance. As is well-known, the mean-field free energy may exhibit phase transitions and could have multiple local minimizers, and the global minimizers may not be unique either [12].

For any local minimizer, taking variation subject to  $\rho \geq 0$  and  $\int \rho dx = 1$ , one has

$$(2.10) \quad \frac{\delta \mathcal{F}}{\delta \rho} = U + W * \rho + \beta^{-1}(\log \rho + 1) = \text{const}$$

on the support of  $\rho$ . One can further verify that the support of the minimizer is full (see the appendix of [48]). These results can also be derived formally using the KKT conditions [10]. Hence, any minimizers of  $\mathcal{F}$  satisfies the following relation:

$$(2.11) \quad \rho = Z^{-1} \exp(-\beta(U + W * \rho)),$$

where  $Z$  is the normalizing constant, and can be verified to be a stationary solution of (1.1) and thus a solution to (1.2).

**2.3. Sampling to solve the stationary nonlinear Fokker–Planck equation.** We aim to connect the Gibbs measure (2.8) with the solution to (1.2) so that we can do sampling from (2.8) to solve the stationary Fokker–Planck equation. Our goal in this work is then to identify the convergence of (2.8) to one of the stationary solution to (1.2) in the form of (2.11) in a certain sense.

To make the meaning precise, consider the  $k$ -marginal of  $\rho_N$  defined as

$$(2.12) \quad \rho_{N,k}(x_1, \dots, x_k) = \int_{\mathbb{R}^{(N-k)d}} \rho_N(x_1, \dots, x_k, dx_{k+1}, \dots, dx_N)$$

and the empirical distribution

$$(2.13) \quad \mu_N := \frac{1}{N} \sum_{i=1}^N \delta(\cdot - X_i), \quad (X_1, X_2, \dots, X_N) \sim \rho_N.$$

We aim to show that  $\rho_{N,1}$  and  $\mu_N$  converge to  $\rho$  in a certain sense. As can be imagined, when there is a phase transition so that the local minimizers of  $\mathcal{F}$  are not unique, while the  $N$ -body system has a unique Gibbs measure, there could be intrinsic difficulty. Our result is that when the temperature is large enough, the stationary solution of the form (2.11) is unique and the convergence can be quantified. The approach in section 3 is to use the relative entropy to do the quantitative estimate.

In particular, using the relative entropy, one can conclude that  $\rho_{N,1}$  converges to  $\rho$  in total variation norm with rate  $1/\sqrt{N}$  and  $\mu_N$  converges to  $\rho$  in  $H^{-\alpha}$ ,  $\alpha > d/2$  in expectation. See sections 3 and 4 for details. In practical applications, we combine these two to obtain the following algorithm.

Clearly,  $\bar{\mu}_N$  is the average of  $N_s$  empirical measures. One may do some postprocessing to make this distribution more smooth. Clearly,

$$\|\bar{\mu}_N - \rho\|_{H^{-\alpha}} \leq \frac{1}{N_s} \sum_{m=1}^{N_s} \|\mu_N^{(m)} - \rho\|_{H^{-\alpha}}.$$

---

**Algorithm 2.1.** Sampling to solve the stationary nonlinear Fokker–Planck equation.

---

- 1: Set  $N \geq 2$  and  $N_s \geq 1$ .
- 2: Draw  $N_s$  samples from  $\rho_N$  using certain sampling methods (e.g., some Markov chain Monte Carlo methods like Metropolis–Hastings (MH) or RBMC).
- 3: Collect all  $X_i$ 's from all the  $N_s$  samples to obtain  $N \times N_s$  points in  $\mathbb{R}^d$ .
- 4: Use the  $NN_s$  points to form a density approximation of  $\rho$  using

$$(2.14) \quad \bar{\mu}_N = \frac{1}{NN_s} \sum_{j=1}^{NN_s} \delta(\cdot - X_j).$$


---

Here,  $\mu_N^{(m)}$  means the empirical measure for the  $m$ th sample. Since the typical rate of convergence for  $\mu_N$  to  $\rho$  is  $N^{-1/2}$ , this is also  $N^{-1/2}$ . In applications, sometimes we care about the convergence of some quantity of the form  $\int \varphi(\cdot) d\bar{\mu}_N$ . This is again the weak convergence and we expect that the rate is again  $N^{-1/2}$  like the law of large numbers.

We also remark that the algorithm and the proof later in sections 3 and 4 can be applied to cases when  $x$  is in bounded domains (like torus and bounded domain with reflection boundary condition). See section 5.1 for one example.

**2.4. Discussion on the second order interacting particle systems.** If the “velocity” is considered, the models for the interacting particle systems are described by the second order systems. For example, for the Langevin system, the particle system may be given by the following for  $i = 1, \dots, N$ :

$$(2.15) \quad \begin{aligned} dX_i &= V_i dt, \\ dV_i &= -\gamma V_i dt - \nabla U(X_i) dt - \frac{1}{N-1} \sum_{j=1, j \neq i}^N \nabla_1 W(X_i, X_j) dt + \sqrt{\frac{2\gamma}{\beta}} dB_i. \end{aligned}$$

This particle system is also associated with the energy functional (2.1). In fact, it can be written as

$$(2.16) \quad \begin{aligned} dX &= V dt, \\ dV &= -\gamma V dt - \nabla_X E_N dt + \sqrt{\frac{2\gamma}{\beta}} d\mathbf{B}. \end{aligned}$$

*Remark 2.3.* If we consider the rescaled energy functional in Remark 2.1, and each particle has mass  $1/N$ , one may obtain

$$\begin{aligned} dX &= V dt, \\ \frac{1}{N} dV &= -\tilde{\gamma} V dt - \frac{1}{N} \nabla_X E_N dt + \sqrt{\frac{2\tilde{\gamma}}{\beta}} d\mathbf{B}. \end{aligned}$$

Now, we need  $\tilde{\beta} = N\beta$  and  $\tilde{\gamma}N = \gamma$  to be consistent with (2.15). The regime considered is the one where the total mass is 1, the temperature is low, and the friction coefficient is also small.

Using (2.16), it is not difficult to verify that the invariant measure is given by

$$\pi(dx, dv) \propto \exp\left(-\beta\left(E_N + \frac{|v|^2}{2}\right)\right) dx dv.$$

The Gibbs measure for the second order systems is just the Gibbs measure of the first order system tensored by the Maxwellian distribution in the velocity space. From the sampling viewpoint, we only have to focus on the Gibbs measure of the first order systems.

Similarly, the second order nonlinear Fokker–Planck equation may be written as

$$(2.17) \quad \partial_t f = -\nabla_x \cdot (vf) + \nabla_v \cdot (f(\gamma v + \nabla U + \nabla(W *_x f))) + \beta^{-1} \gamma \Delta_v f,$$

where  $*_x$  means the convolution in the spatial variable, i.e.,

$$(W *_x f)(x) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} W(x, y) f(y, v) dv dy$$

for distribution  $f$  in the phase space. One can find that if  $\rho(x)$  is a solution to (1.2), then  $\rho(x)e^{-\beta|v|^2/2}$  is a stationary solution to (2.17). In fact, there is a one-to-one correspondence between invariant measures of the first and second order Fokker–Planck equations (1.1) and (2.17). Hence, the (non)-uniqueness and phase transition properties of the stationary solutions of (2.17) can be deduced from the results of (1.1) [16]. Furthermore, the convergence of (2.17) towards an invariant measure is also studied using free-energy approach in [17]. Hence, if we choose to solve stationary equations in the form of (2.17) by sampling, we do not have to distinguish the first order and second order systems.

In some applications, there may be coupling between the space and velocity variables. Then, the invariant measure is no longer of the product form. Nevertheless, one may view  $(X_i, V_i)$  as a new particle, and the convergence of the Gibbs measure to the solution of the stationary nonlinear Fokker–Planck equation can be similarly studied as in section 3 or 4.

**3. Convergence of the stationary solutions for one species.** In this section, we study the convergence of the Gibbs measure directly using the relative entropy as mentioned in the introduction [32, 34, 44]. This may allow weaker assumptions on  $W$ . Compared to many results in literature [32, 34, 44, 57], our result is valid in  $\mathbb{R}^d$ .

To start with, we recall some basic notations. Let  $\mathcal{P}(E)$  denote the set of probability measures on a generic Polish space  $E$ . The relative entropy (KL divergence) between two probability measures  $\mu$  and  $\nu$  on Polish space  $E$  is defined as

$$\mathcal{H}(\mu | \nu) = \begin{cases} \int_E \log \frac{d\mu}{d\nu} d\mu & \text{if } \mu \ll \nu, \\ \infty & \text{else,} \end{cases}$$

where  $\frac{d\mu}{d\nu}$  denotes the Radon–Nikodym derivative of  $\mu$  with respect to  $\nu$ . Let  $\mathcal{P}_{\text{sym}}(E^N)$  denote the set of symmetric probability measures on the product space  $E^N$ , that is, for any Borel set  $A = A_1 \times \cdots \times A_N \subset E^N$  and any permutation  $\sigma$  of  $\{1, \dots, N\}$ , the probability measure  $\mu_N$  satisfies  $\mu_N(A) = \mu_N(\sigma(A))$ , where  $\sigma(A) = A_{\sigma(1)} \times \cdots \times A_{\sigma(N)}$ . For two probability measures  $\mu, \nu \in \mathcal{P}(E^k)$ , the scaled (normalized) relative entropy is given by

$$(3.1) \quad \mathcal{H}_k(\mu | \nu) = \frac{1}{k} \mathcal{H}(\mu | \nu).$$

The scaled relative entropy satisfies the following useful monotonicity property [27, 34].

LEMMA 3.1. *For any probability density  $\rho_N \in \mathcal{P}_{\text{sym}}(E^N)$  and  $\rho \in \mathcal{P}(E)$ . Then,*

$$\mathcal{H}_k(\rho_{N,k} | \rho^{\otimes k}) \leq \mathcal{H}_N(\rho_N | \rho^{\otimes N}), \quad 1 \leq k \leq N.$$

Here  $\rho_{N,k} \in \mathcal{P}(E^k)$  is the  $k$ -marginal of  $\rho_N$  defined as

$$\rho_{N,k}(x_1, \dots, x_k) = \int_{E^{(N-k)}} \rho_N(x_1, \dots, x_k, dx_{k+1}, \dots, dx_N),$$

while  $\rho^{\otimes k} \in \mathcal{P}(E^k)$  is given by  $\rho^{\otimes k}(x_1, \dots, x_k) = \prod_{i=1}^k \rho(x_i)$ .

Next, we state our main result.

THEOREM 3.2. *Suppose that the interacting kernel  $W \in L^\infty(\mathbb{R}^d, \mathbb{R}^d)$ , and assume  $U$  is an external potential with  $\exp(-\beta E_N) \in L^1(\mathbb{R}^{Nd})$  so that the  $N$ -body Gibbs*

distribution  $\rho_N$  in (2.8) is well-defined. Then, if  $\beta < (2\sqrt{2}e\|W\|_\infty)^{-1}$ , for a minimizer  $\rho$  of the free energy  $\mathcal{F}$ , there exists  $C$  independent of  $N$  and  $N_0 > 0$ , s.t.

$$(3.2) \quad \mathcal{H}_N(\rho_N | \rho^{\otimes N}) \leq \frac{C}{N} \quad \forall N > N_0.$$

*Remark 3.3.* By going over the proof of Theorem 3.2, the dependence of  $C$  on the interaction kernel and the temperature can be made explicit. For example, we may choose  $c \sim \frac{2}{(1+\theta)}$ , where  $\theta = 2\sqrt{2}e\|W\|_\infty\beta$ , then

$$C = -\frac{\log(1-c\theta)}{(c-1)} \sim \frac{1+\theta}{1-\theta} \log \frac{1+\theta}{1-\theta}.$$

This result indicates that  $\mathcal{H}_N(\rho_N | \rho^{\otimes N}) \rightarrow 0$  as  $N \rightarrow \infty$  under the assumptions of Theorem 3.2. Due to the indistinguishability assumption on the particles, their joint law is invariant under relabelling of the particles. That is,  $\rho_N$  together with  $\rho_{N,k}$  are symmetric. It follows from Lemma 3.1 that  $\rho_{N,k} \rightharpoonup \rho^{\otimes k}$ , which is indeed the propagation of chaos at equilibrium. Specifically, the first marginal of the invariant Gibbs measure (the law of a single particle) weakly converges to the limit law, i.e.,  $\rho_{N,1} \rightharpoonup \rho$  as  $N \rightarrow \infty$ .

As a direct consequence of Theorem 3.2, the minimizer of the mean-field free energy  $\mathcal{F}$  is unique for the  $\beta$  considered here. In the case where there is phase separation for the limit mean-field equation, so the stationary solutions are not unique at low temperatures, which stationary solution the  $N$ -body distribution would converge to is an interesting question (recall that the  $N$ -body Gibbs distribution is unique). Sampling from the  $N$ -body distribution can be viewed as a method for solving this solution. The selection principle for the stationary solution by this strategy is left for future study and it may be related to the large deviation principle as for the Hamilton–Jacobi equation in [22].

Our approach to prove this result has three steps. First, calculate the relative entropy between  $\rho_N$  and  $\rho^{\otimes N}$  using the characterizations of invariant  $N$ -body Gibbs measure and the stationary solution to nonlinear Fokker–Planck equation. Then control the relative entropy by resorting to the Fenchel–Young’s inequality. Finally, a uniform in  $N$  bound is established for the exponential random fluctuation for the interactions. The key observation here is to introduce a symmetrized version of the fluctuation for the interaction. Compared to the studies of Gibbs measure using the large deviation principle [7] or the Laplace method [2], we need the boundedness of the interaction kernel, but we do not require it to be continuous.

While in the time-dependent studies [44, 34], a common approach is to bound the derivative of the relative entropy, and then Grönwall inequality gives  $\mathcal{H}_N(\rho_N | \rho^{\otimes N})(t) \leq e^{C_T t} (\frac{1}{N} + \mathcal{H}_N(\rho_N | \rho^{\otimes N})(0))$  for finite time horizon  $(0, T]$ . To further obtain uniform-in-time propagation of chaos, the uniform logarithmic Sobolev inequality [26, 25] and the convexity of the potentials [45, 46] are required. For example, in [26, Theorem 9], the uniform logarithmic Sobolev inequality is proved with assumptions including the one-particle conditional log-Sobolev inequality (e.g. when  $U$  is superconvex), and most importantly, a translation of Zegarlinski’s uniqueness condition stating that the product of the Lipschitz spectral gap of one particle and the  $L_\infty$  norm of the Hessian  $\nabla_{x,y}^2 W$  is smaller than 1. Meanwhile, in [46, Theorem 5.1] when there is no confinement potential ( $U = 0$ ), the interaction potential  $W$  is assumed to be symmetric and uniformly convex. While in [45],  $W$  is even, convex, with polynomial growth and  $U$  is uniformly convex. We do not consider the uniform propagation of chaos, and

instead focus on the stationary case and give a convergence of the invariant Gibbs measure to the stationary solution directly. Compared to these results, we need only  $\|W\|_\infty < \infty$  and no other regularity assumptions on  $W$ . It would be an interesting question whether one could combine the convergence of the stationary distribution and finite time propagation of chaos to get some uniform-in-time convergence results under weaker assumptions compared to the uniform-in-time propagation of chaos.

As mentioned above, to prove Theorem 3.2, we need some auxiliary lemmas. The first is a type of Fenchel–Young’s inequality and has been written out explicitly in [34, Lemma 1].

LEMMA 3.4. *For any two probability densities  $\rho$  and  $\tilde{\rho}$  on  $E$  and some test function  $\Phi \in L^1(\rho)$ , one has that  $\forall \eta > 0$ ,*

$$\int_E \Phi \rho \, dx \leq \frac{1}{\eta} \left( \mathcal{H}(\rho | \tilde{\rho}) + \log \int_E e^{\eta \Phi} \tilde{\rho} \, dx \right).$$

Next, we need the following lemma.

LEMMA 3.5. *Consider the random variable*

$$(3.3) \quad B = (W * \rho)(X_1) - W(X_1, X_2) + (W * \rho)(X_2) - \int_{\mathbb{R}^d} \rho(W * \rho) \, dx,$$

where  $X_1, X_2$  are i.i.d. from some probability distribution  $\rho$ . If there exists  $\alpha > 0$  such that  $\lim_{p \rightarrow \infty} \alpha^{-p} \|B\|_{L^p(\rho)}^p \leq 1$ , then for  $\eta < \sqrt{2}/(e\alpha)$ , one has

$$(3.4) \quad \limsup_{N \rightarrow \infty} \int_{\mathbb{R}^{Nd}} \exp \left( \eta \sum_{i=1}^N \left( W * \rho(x_i) - \frac{1}{2(N-1)} \sum_{j:j \neq i} W(x_i, x_j) - \frac{1}{2} \int_{\mathbb{R}^d} \rho(W * \rho) \, dx \right) \right) \rho^{\otimes N} dx_1 \cdots dx_N < \infty.$$

Lemma 3.5 is reminiscent of the following “large deviation estimate” by Jabin and Wang [34, Theorem 4] using combinatoric techniques.

LEMMA 3.6. *Consider  $\rho \in \mathcal{P}(E)$  and  $\phi(x, y)$  satisfies  $\gamma := C(\sup_{p \geq 1} \|\sup_y |\phi(\cdot, y)|\|_{L^p(\rho)}/p)^2 < 1$  with a universal constant  $C$ . Assume that  $\phi$  satisfies the following cancellations:*

$$\int_E \phi(x, y) \rho(x) \, dx = 0 \quad \forall y, \quad \int_E \phi(x, y) \rho(y) \, dy = 0 \quad \forall x.$$

Then

$$\sup_{N \geq 2} \int_{E^N} \exp \left( \frac{1}{N} \sum_{i,j=1}^N \phi(x_i, x_j) \right) \rho^{\otimes N} dx \leq \frac{2}{1-\gamma} < \infty.$$

It is tempting to apply Lemma 3.6 with

$$\phi(x, y) = (W * \rho)(x) - \frac{1}{2} W(x, y) - \frac{1}{2} \int_{\mathbb{R}^d} \rho(W * \rho) \, dx.$$

It is easily found that  $\int_{\mathbb{R}^d} \phi(x, y) \rho(x) \, dx = \frac{1}{2} \int_{\mathbb{R}^d} \rho(W * \rho) \, dx - \frac{1}{2} (W * \rho)(y)$  and  $\int_{\mathbb{R}^d} \phi(x, y) \rho(y) \, dy = \frac{1}{2} (W * \rho)(x) - \frac{1}{2} \int_{\mathbb{R}^d} \rho(W * \rho) \, dx$  do not cancel out.

Our key observation in this work is that if we introduce a symmetric version of the above function

$$(3.5) \quad \phi(x, y) = \frac{1}{2}(W * \rho)(x) + \frac{1}{2}(W * \rho)(y) - \frac{1}{2}W(x, y) - \frac{1}{2} \int_{\mathbb{R}^d} \rho(W * \rho) dx,$$

then  $\phi$  satisfies the two cancellation conditions in Lemma 3.6, and Lemma 3.5 can be established. Here, with this observation, we would like to accommodate the martingale proof in [44, Lemma 4.3] to the new case here. We remark that the results in [44, Lemma 4.3] are weaker than Lemma 3.6 but the conditions needed could be explicit in the proof and might be generalized to other cases in the future.

*Proof of Lemma 3.5.* Let  $X_1, \dots, X_N$  be i.i.d. samples from  $\rho$ . Let  $\mathcal{F}_k = \sigma(X_1, \dots, X_k)$ .

Define

$$D_k = \sum_{j=1}^{k-1} \left( (W * \rho)(X_j) - W(X_j, X_k) \right) + (k-1)(W * \rho)(X_k) - (k-1) \int_{\mathbb{R}^d} \rho(W * \rho) dx.$$

It can be verified that  $\mathbb{E}(D_k | \mathcal{F}_{k-1}) = 0$  and

$$\sum_{k=1}^N D_k = (N-1) \sum_{k=1}^N (W * \rho)(X_k) - \sum_{k,j:k>j} W(X_j, X_k) - \frac{N(N-1)}{2} \int_{\mathbb{R}^d} \rho(W * \rho) dx.$$

The estimate (3.4) then becomes  $\mathbb{E} \exp\left(\frac{\eta}{N-1} \sum_{k=1}^N D_k\right)$ .

Applying the Marcinkiewicz–Zygmund type inequality (see [51, Theorem 2.1]), one has

$$\left\| \sum_{k=1}^N D_k \right\|_{L^p}^2 \leq (p-1) \sum_{k=1}^N \|D_k\|_{L^p}^2, \quad p \geq 2.$$

Define

$$B_j^k = (W * \rho)(X_j) - W(X_j, X_k) + (W * \rho)(X_k) - \int_{\mathbb{R}^d} \rho(W * \rho) dx.$$

Since  $\mathbb{E}(B_j^k | X_k) = 0$ ,  $D_k = \sum_{j=1}^{k-1} B_j^k$ , one can apply the Marcinkiewicz–Zygmund type inequality again to obtain that

$$\|D_k\|_{L^p}^2 \leq (p-1) \sum_{j=1}^{k-1} \|B_j^k\|_{L^p}^2, \quad p \geq 2.$$

Consequently,

$$\left\| \sum_{k=1}^N D_k \right\|_{L^p} \leq (p-1) \sqrt{\sum_{k=1}^N \sum_{j=1}^{k-1} \|B_j^k\|_{L^p}^2} = (p-1) \sqrt{\frac{N(N-1)}{2}} \|B\|_{L^p}, \quad p \geq 2.$$

Using this estimate and  $\mathbb{E} \sum_{k=1}^N D_k = 0$ , one has

$$\mathbb{E} \exp\left(\frac{\eta}{N-1} \sum_{k=1}^N D_k\right) \leq 1 + \sum_{p \geq 2} \frac{\eta^p (p-1)^p \|B\|_{L^p}^p}{p!} \left(\frac{N}{2(N-1)}\right)^{\frac{p}{2}}.$$

Hence, it follows from the Stirling's formula that

$$\mathbb{E} \exp \left( \frac{\eta}{N-1} \sum_{k=1}^N D_k \right) \leq 1 + \sum_{p=2}^{\infty} \frac{1}{\sqrt{2\pi p}} \left( \frac{\eta e(p-1) \|B\|_{L^p}}{p} \sqrt{\frac{N}{2(N-1)}} \right)^p < \infty,$$

when  $\eta < \sqrt{2}/(e\alpha)$  for  $N$  large enough. Specifically,  $\alpha$  can be chosen to be  $\|B\|_{L^\infty}$  if  $B$  is bounded.  $\square$

Now, we prove Theorem 3.2.

*Proof of Theorem 3.2.* Recall the stationary solution  $\rho$  to the nonlinear Fokker-Planck equation in (2.11). One has

$$\rho^{\otimes N} = \prod_{i=1}^N \rho(x_i) = Z^{-N} \exp \left( -\beta \left( \sum_{i=1}^N U(x_i) + \sum_{i=1}^N (W * \rho)(x_i) \right) \right).$$

According to Lemma 2.2, the  $N$ -body Gibbs measure can be rewritten as

$$\rho_N = Z_N^{-1} \exp \left( -\beta \left( \sum_i U(x_i) + \frac{1}{2(N-1)} \sum_{i \neq j} W(x_i, x_j) + \frac{N}{2} \int_{\mathbb{R}^d} \rho(W * \rho) dx \right) \right),$$

where  $Z_N = \bar{Z}_N \exp(-\frac{N\beta}{2} \int_{\mathbb{R}^d} \rho(W * \rho) dx)$ . Direct computation reveals that

$$\mathcal{H}(\rho^{\otimes N} | \rho_N) = \int_{\mathbb{R}^{dN}} \rho^{\otimes N} \log \frac{\rho^{\otimes N}}{\rho_N} dx = \log \left( \frac{Z^{-N}}{Z_N^{-1}} \right).$$

Then, the nonnegativity of the relative entropy yields the relation

$$(3.6) \quad -N \log Z + \log Z_N \geq 0.$$

On the other hand,

$$\begin{aligned} \mathcal{H}(\rho_N | \rho^{\otimes N}) &= \int_{\mathbb{R}^{dN}} \rho_N \log \rho_N dx - \int_{\mathbb{R}^{dN}} \rho_N \log \rho^{\otimes N} dx \\ &= \int_{\mathbb{R}^{dN}} \rho_N \left( -\log Z_N - \beta \left( \sum_{i=1}^N U(x_i) + \frac{1}{2(N-1)} \sum_{i \neq j} W(x_i, x_j) + \frac{N}{2} \int_{\mathbb{R}^d} \rho(W * \rho) \right) \right) dx \\ &\quad - \int_{\mathbb{R}^{dN}} \rho_N \left( -N \log Z - \beta \left( \sum_{i=1}^N U(x_i) + \sum_{i=1}^N (W * \rho)(x_i) \right) \right) dx \\ &= N \log Z - \log Z_N + \beta \int_{\mathbb{R}^{dN}} \rho_N \Phi dx \leq \beta \int_{\mathbb{R}^{dN}} \rho_N \Phi dx, \end{aligned}$$

where

$$(3.7) \quad \Phi(x_1, \dots, x_N) = -\frac{1}{2(N-1)} \sum_{i, j: i \neq j} W(x_i, x_j) + \sum_{i=1}^N (W * \rho)(x_i) - \frac{N}{2} \int_{\mathbb{R}^d} \rho(W * \rho) dx.$$

Obviously, we do not know the properties of  $\rho_N$  and would much prefer having expectations with respect to the tensorized law  $\rho^{\otimes N}$ . Applying Lemma 3.4 with  $\eta = c\beta$ , one has

$$(3.8) \quad \int_{\mathbb{R}^{dN}} \Phi \rho_N dx \leq \frac{1}{c\beta} \left( \mathcal{H}(\rho_N | \rho^{\otimes N}) + \log \int_{\mathbb{R}^{dN}} \exp(c\beta \Phi) \rho^{\otimes N} dx \right).$$

Hence, for  $c > 1$ , we can get

$$\mathcal{H}(\rho_N | \rho^{\otimes N}) \leq \frac{1}{c-1} \log \int_{\mathbb{R}^{dN}} \exp(c\beta\Phi)\rho^{\otimes N} dx.$$

Applying Lemma 3.5, since  $\|B\|_{L^\infty} \leq 4\|W\|_{L^\infty} < \infty$  (note that  $B$  is the random variable defined in (3.3)), we find that when

$$\beta < \frac{1}{2\sqrt{2}e\|W\|_\infty},$$

there exists  $c > 1$  and  $N_0 > 0$  such that

$$\mathcal{H}(\rho_N | \rho^{\otimes N}) \leq \frac{C}{c-1} \quad \forall N > N_0. \quad \square$$

Like in [32, 44], our result also needs the assumption that the interaction kernel  $W$  is bounded. More specifically, the proof of Lemma 3.5 requires  $\lim_{p \rightarrow \infty} \|B\|_{L^p}$  to be almost bounded, which is not true for general singular kernels.

Besides, we can obtain the convergence in the total variation norm or the Wasserstein metric under the same condition. The results follow from the monotonicity property, the Csiszár–Kullback–Pinsker inequality [56] for the TV norm and the Talagrand transportation inequality for the Wasserstein distance [3].

**COROLLARY 3.7.** *For  $\beta < (2e\sqrt{2}\|W\|_\infty)^{-1}$  and  $N$  large enough, for any fixed  $k$ ,*

$$\|\rho_{N,k} - \rho^{\otimes k}\|_{TV} \leq \sqrt{2\mathcal{H}(\rho_{N,k} | \rho^{\otimes k})} \leq \sqrt{\frac{2k}{N}\mathcal{H}(\rho_N | \rho^{\otimes N})} \lesssim \sqrt{\frac{k}{N}}.$$

*A similar result holds for the Wasserstein distance*

$$W_1(\rho_{N,1}, \rho) \lesssim \sqrt{\frac{1}{N}}.$$

The following estimate concerns the convergence from the empirical measure  $\mu_N = \frac{1}{N} \sum_{i=1}^N \delta(x - X_i)$  of particle system (1.5) at steady state to  $\rho$  in  $H^{-\alpha}$ , for  $\alpha > d/2$ . Together with the tightness, this can imply the weak convergence of the empirical measures, justifying using sampling to compute the stationary nonlinear Fokker–Planck equations.

**COROLLARY 3.8.** *For each  $\alpha > d/2$ , there exists constant  $C$  depending only on  $\alpha$  such that*

$$\mathbb{E}\|\mu_N - \rho\|_{H^{-\alpha}}^2 \leq \frac{C}{N}$$

*for  $\beta$  small enough and  $N$  large enough.*

*Proof.* The proof is the whole space version of [57, Lemma 2.6].

Since the Dirac measure belongs to  $H^{-\alpha}(\mathbb{R}^d)$  for  $\alpha > d/2$ ,  $\mu_N - \rho \in H^{-\alpha}(\mathbb{R}^d)$ . Denote  $\Psi = \|\mu_N - \rho\|_{H^{-\alpha}}^2$ , then

$$\Psi = \int_{\mathbb{R}^d} (1 + |\xi|^2)^{-\alpha} |(\widehat{\mu_N - \rho})(\xi)|^2 d\xi.$$

Here,  $\hat{f}$  represents the Fourier transform of  $f$ . Since  $\Psi < \infty$ , applying Lemma 3.4 with  $\eta \leftarrow \eta N$ ,

$$\mathbb{E}\Psi = \int_{\mathbb{R}^{dN}} \Psi \rho_N dx \leq \frac{1}{\eta} \left( \mathcal{H}_N(\rho_N | \rho^{\otimes N}) + \frac{1}{N} \log \int_{\mathbb{R}^{dN}} \exp(N\eta\Psi)\rho^{\otimes N} dx \right).$$

Let  $C_\alpha = \int_{\mathbb{R}^d} (1 + |\xi|^2)^{-\alpha} d\xi$ , then the Jensen's inequality gives

$$\exp(N\eta\Psi) \leq \int_{\mathbb{R}^d} \frac{(1 + |\xi|^2)^{-\alpha}}{C_\alpha} \exp\left(N\eta C_\alpha |\widehat{(\mu_N - \rho)}(\xi)|^2\right) d\xi.$$

Define

$$\phi(x, y) = \left( e^{i\xi \cdot x} - \int_{\mathbb{R}^d} e^{i\xi \cdot x} \rho(x) dx \right) \left( e^{-i\xi \cdot y} - \int_{\mathbb{R}^d} e^{-i\xi \cdot y} \rho(y) dy \right).$$

Since  $\phi$  satisfies the two cancellations and

$$|\widehat{(\mu_N - \rho)}(\xi)|^2 = \frac{1}{N^2} \sum_{i,j} \operatorname{Re} [\phi(X_i, X_j)],$$

Lemma 3.6 implies that for  $\eta$  small enough (depending on  $\alpha$  and  $W$ ),  $\exists C$  depends on  $\alpha$ , s.t.

$$\sup_{\xi} \sup_{N \geq 2} \int_{\mathbb{R}^{dN}} \exp(N\eta C_\alpha \langle \phi, \mu_N \otimes \mu_N \rangle) \rho^{\otimes N} dx \leq C.$$

Together with Theorem 3.2, one then concludes that

$$\mathbb{E} \|\mu_N - \rho\|_{H^{-\alpha}}^2 \leq \frac{C}{N}. \quad \square$$

Again, we do not need  $W$  and  $U$  be to continuous any more and we have the rate of convergence. The rate  $1/\sqrt{N}$  is widely considered to be optimal for the convergence of empirical measures as it corresponds to the stochastic fluctuations.

As a next corollary, we consider the fluctuation measure

$$\mathcal{G}_N = \sqrt{N}(\mu_N - \rho)$$

around the mean field limit. Corollary 3.8 suggests

$$(3.9) \quad \limsup_{N \rightarrow \infty} \mathbb{E} \|\mathcal{G}_N\|_{H^{-\alpha}}^2 \leq C \text{ for } \alpha > d/2,$$

when  $\beta$  is small enough. Here, we consider the interaction fluctuation where  $W$  does not necessarily belong to  $H^\alpha$  with  $\alpha > d/2$ .

**COROLLARY 3.9.** *For  $\beta < (2e\sqrt{2}\|W\|_\infty)^{-1}$  there exists  $N_0 > 0$  such that*

$$(3.10) \quad \sup_{N > N_0} \mathbb{E} \langle W(x, y), \mathcal{G}_N^{\otimes 2} \rangle < \infty.$$

*Proof.* It can be computed that

$$\langle W(x, y), \mathcal{G}_N^{\otimes 2} \rangle = \frac{1}{N} \sum_{i,j=1}^N W(X_i, X_j) - 2 \sum_{i=1}^N (W * \rho)(X_i) + N \int_{\mathbb{R}^d} \rho(W * \rho) dx.$$

This is essentially  $-2\Phi$  with  $\Phi$  given in (3.7). The only difference is that the self-interactions  $W(X_i, X_i)$  are also considered and there is  $1/N$  here while  $1/(N-1)$  there. The difference is clearly tiny. Using the same proof for Theorem 3.2, the result follows.  $\square$

In [14], they also tried to study the fluctuation of the Gibbs measure. There, they assumed (3.10), while we can see that (3.10) actually holds when the interaction kernel is bounded (not necessarily continuous).

**4. Convergence for the multispecies case.** In this section, we consider the multispecies case, which is actually common in applications. For example, in electrochemistry, we often want to find the distributions of different charged chemical species in a fluid medium. To make the presentation clean, we consider only two species, while the general case can be treated similarly. [18] also studied the propagation of chaos and invariant measures for a two-species case of McKean–Vlasov diffusions, but there is no quantitative convergence towards stationary solution.

**4.1. The problem setup.** We will assume that there are  $N$  particles for each species. The particles for the first species are labelled by  $X_i$ ,  $1 \leq i \leq N$  while the particles for the second species are labelled by  $Y_j$ ,  $1 \leq j \leq N$ . The equations of the motion using the overdamped Langevin may be given by

$$(4.1) \quad \begin{aligned} dX_i = & - \left( \nabla U_1(X_i) + \frac{1}{N-1} \sum_{k=1, k \neq i}^N \nabla_i W_1(X_i, X_k) + \frac{1}{N} \sum_{j=1}^N \nabla_i W_c(X_i, Y_j) \right) dt \\ & + \sqrt{\frac{2}{\beta}} dB_{1i}(t), \quad i = 1, \dots, N, \\ dY_j = & - \left( \nabla U_2(Y_j) + \frac{1}{N-1} \sum_{\ell=1, \ell \neq j}^N \nabla_j W_2(Y_j, Y_\ell) + \frac{1}{N} \sum_{i=1}^N \nabla_j W_c(Y_j, X_i) \right) dt \\ & + \sqrt{\frac{2}{\beta}} dB_{2j}(t), \quad j = 1, \dots, N. \end{aligned}$$

Here,  $U_1, U_2$  are the external potentials for the two species,  $W_1$  is the interaction potential within the first species,  $W_2$  is the interaction potential within the second species, and  $W_c$  is the cross species interaction. The Brownian motions  $B_{1,i}, B_{2,j}$  are all independent. For the charged chemical species, one may have  $W_1(x, y) = z_1^2 \Phi(x - y)$ ,  $W_2(x, y) = z_2^2 \Phi(x - y)$ , and  $W_c(x, y) = z_1 z_2 \Phi(x - y)$ , where  $z_i$ 's are the valence of the ions and  $\Phi$  is the Coulomb potential. Typically,  $z_1 = 1$  and  $z_2 = -1$ . We assume that all the interaction potentials are symmetric while the asymmetric case can be similarly treated. Then, one has

$$\nabla_x W_c(x, y) = -\nabla_x W_c(y, x).$$

It is not hard to find that the corresponding nonlinear Fokker–Planck equations are given by

$$(4.2) \quad \begin{aligned} \partial_t \rho_1 &= \nabla_x \cdot (\rho_1 (\nabla_x U_1 + \nabla_x W_1 * \rho_1 + \nabla_x W_c * \rho_2)) + \beta^{-1} \Delta_x \rho_1, \\ \partial_t \rho_2 &= \nabla_y \cdot (\rho_2 (\nabla_y U_2 + \nabla_y W_2 * \rho_2 + \nabla_y W_c * \rho_1)) + \beta^{-1} \Delta_y \rho_2. \end{aligned}$$

The stationary equations are thus (1.4).

The energy functional for the system (4.1) is

$$(4.3) \quad \begin{aligned} E_N(x, y) &= \sum_{i=1}^N U_1(x_i) + \sum_{i=1}^N U_2(y_i) + \frac{1}{2(N-1)} \sum_{i,j=1, i \neq j}^N W_1(x_i, x_j) \\ &+ \frac{1}{2(N-1)} \sum_{i,j=1, i \neq j}^N W_2(y_i, y_j) + \frac{1}{N} \sum_{i,j=1}^N W_c(x_i, y_j), \end{aligned}$$

where  $\mathbf{x} = (x_1, \dots, x_N)$  and  $\mathbf{y} = (y_1, \dots, y_N)$ . Similar to the one species case, the equations (4.1) are actually

$$(4.4) \quad d \begin{pmatrix} X \\ Y \end{pmatrix} = -\nabla_{(\mathbf{x}, \mathbf{y})} E_N(X, Y) dt + \sqrt{\frac{2}{\beta}} d\mathbf{B},$$

so that the Gibbs distribution is given by

$$(4.5) \quad \rho_N(\mathbf{x}, \mathbf{y}) = \bar{Z}_N^{-1} \exp(-\beta E_N(\mathbf{x}, \mathbf{y})).$$

The nonlinear Fokker–Planck equation has the free energy

$$(4.6) \quad \begin{aligned} \mathcal{F}(\rho_1, \rho_2) &= \int_{\mathbb{R}^d} U_1 \rho_1 dx + \int_{\mathbb{R}^d} U_2 \rho_2 dy \\ &+ \frac{1}{2} \int_{\mathbb{R}^d \times \mathbb{R}^d} W_1(x, x') \rho_1(x) \rho_1(x') dx dx' + \frac{1}{2} \int_{\mathbb{R}^d \times \mathbb{R}^d} W_2(y, y') \rho_2(y) \rho_2(y') dy dy' \\ &+ \int_{\mathbb{R}^d \times \mathbb{R}^d} W_c(x, y) \rho_1(x) \rho_2(y) dx dy + \beta^{-1} \int_{\mathbb{R}^d} \rho_1 \log \rho_1 dx + \beta^{-1} \int_{\mathbb{R}^d} \rho_2 \log \rho_2 dy. \end{aligned}$$

The equation can be viewed as the  $W_2$  gradient flow in the space  $\mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d)$ , namely

$$\partial_t \rho_1 = \nabla_x \cdot \left( \rho_1 \nabla_x \frac{\delta \mathcal{F}}{\delta \rho_1} \right), \quad \partial_t \rho_2 = \nabla_y \cdot \left( \rho_2 \nabla_y \frac{\delta \mathcal{F}}{\delta \rho_2} \right).$$

By a similar argument, a stationary pair  $(\rho_1, \rho_2)$  satisfies

$$(4.7) \quad \begin{aligned} \rho_1(x) &= Z_1^{-1} \exp \left( -\beta (U_1(x) + (W_1 * \rho_1)(x) + (W_c * \rho_2)(x)) \right), \\ \rho_2(y) &= Z_2^{-1} \exp \left( -\beta (U_2(y) + (W_2 * \rho_2)(y) + (W_c * \rho_1)(y)) \right), \end{aligned}$$

where  $Z_i$ 's are normalizing constants.

Our approach to investigate the convergence is based on the observation that the Gibbs distribution is symmetric about  $(X_i, Y_i)$  and thus we will look at the convergence to the joint distribution  $\rho_1(x) \rho_2(y) dx dy \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ . Using this idea, one may view  $P_i \equiv (X_i, Y_i)$  as a grand particle. The interaction energy  $\frac{1}{N} W_c(X_i, Y_i)$  can then be treated as the self-interaction. This can be formally reduced to the one species case. This viewpoint, however, cannot identify the independence between  $x$  and  $y$  in the mean field limit explicitly.

*Remark 4.1.* This coupling strategy can be extended to multispecies system with different interactions and species with different number of particles as long as they scale simultaneously in the mean field limit. Namely, if the particle number of the  $m$ th species is  $N_m = a_m N$  with fixed  $a_m \in \mathbb{N}^+$  for  $m = 1, \dots, M$ , then we can collect  $a_m$  particles from the  $m$ th species, and consider the  $\sum_{m=1}^M a_m$  particles as a whole. Therefore, one can still pass the  $N \rightarrow \infty$  limit with suitably defined interactions among the  $N$  grand particles. See section 5.1 for an example. However, if the numbers of different species tend to infinity at different rates, our approach is no longer applicable.

**4.2. The convergence result and the sketch of the proof.** We will use the treatment in the above section to obtain the quantitative convergence. In particular, we have the following claim.

THEOREM 4.2. *Suppose that the interaction potentials are bounded and*

$$(4.8) \quad \beta 2\sqrt{2}e(2\|W_c\|_\infty + \max(\|W_1\|_\infty, \|W_2\|_\infty)) < 1.$$

Let  $(\rho_1, \rho_2)$  be a stationary pair and define  $\rho_1\rho_2(dx, dy) := \rho_1(x)\rho_2(y)dxdy$ . Then for  $N$  large enough, one has

$$(4.9) \quad \mathcal{H}_N(\rho_N | (\rho_1\rho_2)^{\otimes N}) \leq \frac{C}{N}.$$

The proof is similar to the proof of Theorem 3.2. Here, we list out the key ingredients and omit the details. By noting that

$$(4.10) \quad (\rho_1\rho_2)^{\otimes N} = (Z_1Z_2)^{-N} \exp \left( -\beta \left( \sum_{i=1}^N (U_1(x_i) + (W_1 * \rho_1)(x_i) + (W_c * \rho_2)(x_i)) + \sum_{j=1}^N (U_2(y_j) + (W_2 * \rho_2)(y_j) + (W_c * \rho_1)(y_j)) \right) \right),$$

we rewrite the  $\rho_N$  in (4.5) as

$$(4.11) \quad \rho_N = Z_N^{-1} \exp \left( -\beta \left( E_N(x, y) + \frac{N}{2} \int_{\mathbb{R}^d} W_1 * \rho_1(x)\rho_1(x) dx + \frac{N}{2} \int_{\mathbb{R}^d} W_2 * \rho_2(y)\rho_2(y) dy + N \int_{\mathbb{R}^d} W_c * \rho_2(x)\rho_1(x) dx \right) \right).$$

The difference between  $\bar{Z}_N$  and  $Z_N$  is a constant depending on  $N$  but not on  $x, y$ .

It is similarly computed that

$$\mathcal{H}((\rho_1\rho_2)^{\otimes N} | \rho_N) = \log \left( \frac{(Z_1Z_2)^{-N}}{Z_N^{-1}} \right) \geq 0,$$

and

$$\mathcal{H}(\rho_N | (\rho_1\rho_2)^{\otimes N}) = \log \left( \frac{Z_N^{-1}}{(Z_1Z_2)^{-N}} \right) + \beta \int_{\mathbb{R}^{Nd} \times \mathbb{R}^{Nd}} (\Phi_1(x) + \Phi_2(y) + \Phi_c(x, y))\rho_N(dx, dy),$$

where

$$\Phi_1(x) = \sum_{i=1}^N (W_1 * \rho_1)(x_i) - \frac{1}{2(N-1)} \sum_{i \neq j} W_1(x_i, x_j) - \frac{N}{2} \int_{\mathbb{R}^d} \rho_1(W_1 * \rho_1) dx,$$

$$\Phi_2(y) = \sum_{j=1}^N (W_2 * \rho_2)(y_j) - \frac{1}{2(N-1)} \sum_{i \neq j} W_2(y_i, y_j) - \frac{N}{2} \int_{\mathbb{R}^d} \rho_2(W_2 * \rho_2) dy,$$

$$\Phi_c(x, y) = \sum_{i=1}^N ((W_c * \rho_2)(x_i) + (W_c * \rho_1)(y_i)) - \frac{1}{N} \sum_{i, j} W_c(x_i, y_j) - N \int_{\mathbb{R}^d} \rho_1(W_c * \rho_2) dx.$$

The treatment of  $\Phi_1$  and  $\Phi_2$  are the same as in Lemma 3.5. In particular, using Lemma 3.4, one has

$$\begin{aligned} \beta \int_{\mathbb{R}^{Nd} \times \mathbb{R}^{Nd}} (\Phi_1(x) + \Phi_2(y)) \rho_N(dx, dy) &\leq \frac{1}{c} \mathcal{H}(\rho_N | (\rho_1 \rho_2)^{\otimes N}) \\ &+ \frac{1}{c} \log \left( \int \exp(c\beta(\Phi_1 + \Phi_2)) (\rho_1 \rho_2)^{\otimes N}(dx, dy) \right). \end{aligned}$$

The integral of the exponential can be split into two integrals. Hence, the conditions are similar for this to be finite.

The cross term  $\Phi_c$  is much simpler. In fact, we do not need the symmetrization technique in (3.5). Set

$$\phi_c(x, y) = (W_c * \rho_2)(x) + (W_c * \rho_1)(y) - W_c(x, y) - \int_{\mathbb{R}^d} \rho_1(W_c * \rho_2) dx$$

such that  $\sum_{i,j} \phi_c(x_i, y_j) = N\Phi_c$ . One finds that  $\phi_c$  satisfies the required cancellation conditions. Repeating the proof of Lemma 3.5 with

$$B_j^k = \phi_c(X_j, Y_k) + \phi_c(X_k, Y_j),$$

one finds the following fact.

LEMMA 4.3. *Suppose that  $W_c \in L^\infty$  and for  $\eta < 1/(4\sqrt{2}e\|W_c\|_\infty)$ , one has*

$$\sup_{N \geq 2} \int_{\mathbb{R}^{Nd} \times \mathbb{R}^{Nd}} \exp \left( \eta \frac{1}{N} \sum_{i,j} \phi_c(x_i, y_j) \right) (\rho_1 \rho_2)^{\otimes N}(dx, dy) < \infty.$$

Here, the limsup is replaced by  $\sup_{N \geq 2}$  because there is no discrepancy between  $N - 1$  and  $N$ . Moreover, the coefficient is now  $4\sqrt{2}$  instead of  $2\sqrt{2}$  because in construction of the martingale difference  $B_j^k$ , there are more terms.

The condition (4.8) is required because we use

$$\begin{aligned} \beta \int_{\mathbb{R}^{Nd} \times \mathbb{R}^{Nd}} \Phi_c(x, y) \rho_N(dx, dy) &\leq \frac{1}{c'} \mathcal{H}(\rho_N | (\rho_1 \rho_2)^{\otimes N}) \\ &+ \frac{1}{c'} \log \left( \int \exp(c'\beta\Phi_c) (\rho_1 \rho_2)^{\otimes N} \right) \end{aligned}$$

and need  $1/c + 1/c' < 1$ .

**5. Numerical experiments.** According to Theorems 3.2 and 4.2 and Corollary 3.8, both the one marginal of the  $N$ -body Gibbs measure and the empirical measure constructed from the samples converge to the solution of the stationary Fokker–Planck equation. This then justifies our motivation to use sampling methods to solve the stationary nonlinear Fokker–Planck equations. Upon adequately sampling from the  $N$ -body Gibbs measure (2.8), one can obtain a good approximation to the solution of the stationary Fokker–Planck equation.

To sample from the Gibbs distribution like (2.8), standard Markov Chain Monte Carlo method, such as the Langevin Monte Carlo [52] or the direct Metropolis–Hastings (MH) algorithm [49] are not convenient as  $W$  could be singular and the complexity for moving one particle is  $\mathcal{O}(N)$ . The Random Batch Monte Carlo (RBMC) method proposed in [43] can overcome these two difficulties, where the singularity is

treated by a kernel splitting strategy while the complexity is reduced to  $\mathcal{O}(1)$  by a random batch strategy [35].

By using the sampling methods, the requirement of the uniform-in-time propagation of chaos for the direct simulation has been changed to the requirement of the mixing property of the designed Markov chain Monte Carlo. We would like to emphasize that we have assumed there is one sampling method that can sample from the  $N$ -body Gibbs distribution, which may have a bias within the tolerance. In the numerical experiments here, we have chosen the RBMC for which the rigorous proof of the ergodicity is not known yet. We would like to point out that RBMC is just an example of the sampling methods, so the study of the RBMC is not a central issue of the strategy proposed in this paper. The RBMC has turned out to be efficient in practice if the designed Langevin dynamics have good mixing property, and this is why we chose it. About the random batch strategy, we remark that the extra noise could introduce large error by changing the critical temperature for systems with phase transition [24]. Nevertheless, the ergodicity and the error of some Langevin methods with random batches have been proved in related papers [42, 41, 30].

Let us briefly describe RBMC here. We decompose the singular interacting potential  $W$  into  $W = W^{(1)} + W^{(2)}$ , where  $W^{(2)}$  is singular and of short range while  $W^{(1)}$  is the remaining part which is smooth and often of long range. Then,  $W^{(1)}$  is used in the Langevin dynamics step, while  $W^{(2)}$  is used in the Metropolis rejection step. Regarding the complexity, consider only the interactions within a randomly selected mini batch [35], the Langevin step is  $\mathcal{O}(1)$ . Besides, Metropolis rejection step is also practically  $\mathcal{O}(1)$  since  $W^{(2)}$  is local. The RBMC method is summarized in Algorithm 5.1 and the details can be found in [43].

In this section, we will first consider the Poisson–Boltzmann (PB) equation where two types of ions are considered. Note that the interaction kernel is not bounded so

---

**Algorithm 5.1.** Random Batch Monte Carlo method.

---

- 1: Split  $W = W^{(1)} + W^{(2)}$ . Generate  $N$  initial particles  $X_i, i = 1, \dots, N$ . Choose  $m \geq 1$ , batch size  $p > 1$ , number of iterations in the burn in phase and sampling phase  $N_b, N_s$ .
- 2: **for**  $n = 1 : N_b + N_s$  **do**
- 3: Randomly select a particle  $X_i$  with uniform probability.
- 4: **for**  $k = 1 : m$  **do**
- 5: Choose  $\xi_k$  (random set not containing  $i$ ),  $z_k \sim N(0, I_d)$ ,  $\tau_k > 0$ .

$$X_i = X_i - \left( \nabla U(X_i) + \frac{1}{p-1} \sum_{j \in \xi_k} \nabla_1 W^{(1)}(X_i, X_j) \right) \tau_k + \sqrt{\frac{2\tau_k}{\beta}} z_k.$$

- 6: **end for**
- 7: Compute the acceptance ratio using cell list or other data structures:

$$a(X_i^*, X_i) = \min \left\{ 1, \exp \left( -\beta \sum_{j:j \neq i} \left( W^{(2)}(X_i^*, X_j) - W^{(2)}(X_i, X_j) \right) \right) \right\}.$$

Accept  $X_i^*$  with probability  $a(X_i^*, X_i)$ .

- 8: **end for**
-

the theory will not apply but the numerical results still agree with the theory. As a second example, we look at using sampling for training the two-layer neural networks.

**5.1. The Poisson–Boltzmann equation.** The PB equation [23, 9] is a classical implicit solvent model describing the electric potential at equilibrium, widely used in electrochemistry. Consider a charged colloid  $\Omega$  immersed in some symmetric monovalent electrolyte. Then the ions (two types with valences  $z_+ = 1, z_- = -1$ ) will concentrate close to the surface of  $\Omega$  and form a screening layer. At equilibrium, the charge distribution outside the charged surface is given by the Boltzmann distribution, i.e.,

$$\rho_+ \propto \exp(-\beta\phi), \quad \rho_- \propto \exp(\beta\phi).$$

Here  $\phi$  denotes the electric potential. The concentration of both anions and cations are  $\rho_\infty$  in the far field. The potential then satisfies

$$-\varepsilon\Delta\phi = \begin{cases} \rho_\infty(e^{-\beta\phi} - e^{\beta\phi}), & x \in \Omega^c, \\ \rho_f, & x \in \Omega. \end{cases}$$

Here  $\varepsilon$  is the dielectric constant. The equation in  $\Omega^c$  is the PB equation.

To do simulation, we truncate the external solution into bounded domain  $D_L = B(0, L) \setminus \Omega$  and prescribe a Neumann boundary condition

$$-\frac{\partial\phi}{\partial n} = 0, \quad x \in \partial B(0, L).$$

It has been shown that this approximation problem converges to the PB equation for  $L$  large enough [40].

Unlike the direction simulation approach in [40], we solve the PB equation by sampling method here. Consider the corresponding particle system where the interaction potential between the particles is the Coulomb potential. Given the free charge  $\rho_f$ , we adjust the total charge of the cations  $Q_+$  so that the total charge of the anions satisfy

$$Q_+ - Q_- + \int_{\Omega} \rho_f dx = 0.$$

We choose a numerical charge unit  $q$  such that the numbers of numerical cations and anions are given, respectively, by

$$N_+ = \left\lceil \frac{Q_+}{z_+q} \right\rceil, \quad N_- = \left\lceil \frac{Q_-}{|z_-|q} \right\rceil.$$

Note that here  $N_+ \neq N_-$  but if we decrease  $q$ ,  $N_+$  and  $N_-$  will increase proportionally so that the proof in section 4 can be applied similarly. Let  $W$  be the fundamental solution to

$$-\varepsilon\Delta W = \delta.$$

Then the external potential is given by  $U = \rho_f * W$  and the Gibbs distribution can be written as

$$(5.1) \quad \rho_N \propto \exp\left(-\beta\left(\sum_i z_i U(x_i) + \frac{1}{2} \sum_{i \neq j} z_i z_j q W(x_i - x_j)\right)\right) 1_{\{x_i \in D_L, 1 \leq i \leq N\}},$$

where  $z_i$  is the valence of  $i$ th particle,  $N = N_+ + N_-$ , the cations and anions are numbered together. Note that the Gibbs free energy used is scaled by  $1/q$  so that the mean field limit exists, consistent with Remark 2.1. In fact, if we consider the corresponding overdamped Langevin equation, one has

$$dX_i = -z_i \left( \nabla U(x_i) + \sum_{j \neq i} z_j q \nabla W(x_i - x_j) \right) dt + \sqrt{2\beta^{-1}} dB_i + dR_i,$$

where  $dR_i$  is the reflection at the boundary (see [40]). The formal mean field limit is

$$\begin{aligned} \partial_t \rho_+ &= \nabla \cdot (\rho_+ (\nabla U + \nabla W * \rho_+ - \nabla W * \rho_-)) + \beta^{-1} \Delta \rho_+, \quad x \in D_L, \\ \partial_t \rho_- &= -\nabla \cdot (\rho_- (\nabla U + \nabla W * \rho_+ - \nabla W * \rho_-)) + \beta^{-1} \Delta \rho_-, \quad x \in D_L. \end{aligned}$$

This is the system of PNP equations, with the equilibrium to be the PB equation.

For the one-dimensional (1D) case,  $W(x) = (2\epsilon)^{-1}|x|$ , and  $\nabla W$  is bounded, the theory above can be applied. In fact, though the domain here is bounded, the proof can be performed without difficulty. For the three-dimensional (3D) case,  $W(x) = \frac{1}{4\pi\epsilon|x|}$ , the theory above does not apply. If we use this potential directly in the Gibbs distribution, it is not integrable due to the singularity between one anion particle and one cation particle. One technique to avoid this attraction is adding some hard sphere repulsive potential such as the Lennard-Jones potential [36] given by

$$(5.2) \quad U_{LJ}(r) = 4\epsilon \left[ \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^6 \right],$$

where  $\epsilon$  is the depth of the potential well (i.e., the strength of Lennard-Jones potential),  $\sigma$  is the finite distance at which the interparticle potential is zero. The parameters  $\epsilon, \sigma$  vanish in the mean field limit so that the effect of the Lennard-Jones does not appear in the PB equation. Another approach is to take a cutoff of  $W$  that vanishes as  $N \rightarrow \infty$ . For example,

$$\bar{W}(r) = \begin{cases} \frac{1}{8\pi\epsilon r_N} \left( 3 - \frac{r^2}{r_N^2} \right), & 0 < r < r_N, \\ \frac{1}{4\pi\epsilon r}, & r > r_N, \end{cases}$$

where  $r_N = N^{-\gamma(d)}$  for some  $\gamma(d) > 0$ . Clearly,  $|\bar{W}| \leq \frac{3}{8\pi\epsilon r_N} = \frac{3}{8\pi\epsilon} N^\gamma$ .  $\bar{W}$  converges to  $W$  as  $N \rightarrow \infty$ . Either with the hard sphere potential or the cutoff, the Gibbs measure in three dimensions then can be normalized to a probability measure. Though our theory does not apply due to the singularity, we can still test the performance numerically nevertheless.

For the 1D case, the particles can be imagined as charged sheets in three dimensions and the Green's function does not blow up at  $x = 0$ . there is no need to do splitting in the RBMC. That means we run a particle using the random minibatch version of a Langevin equation and then accept the proposal totally. This is, in fact, a modified version of the stochastic gradient Langevin dynamics (SGLD) [58], but we are allowed to move several particles only in each iteration. When the particles fall out of the domain considered, we do simple reflection.

In Figure 1(a), we present the sampling results for solving the 1D PB equation. The particles are all put in  $(1, L)$  with  $L = 15$  in the experiment and a free charge  $Q_f$  is put at  $x = 0$  (for the 1D case, the location of the free charge in  $(-\infty, 1)$  is not

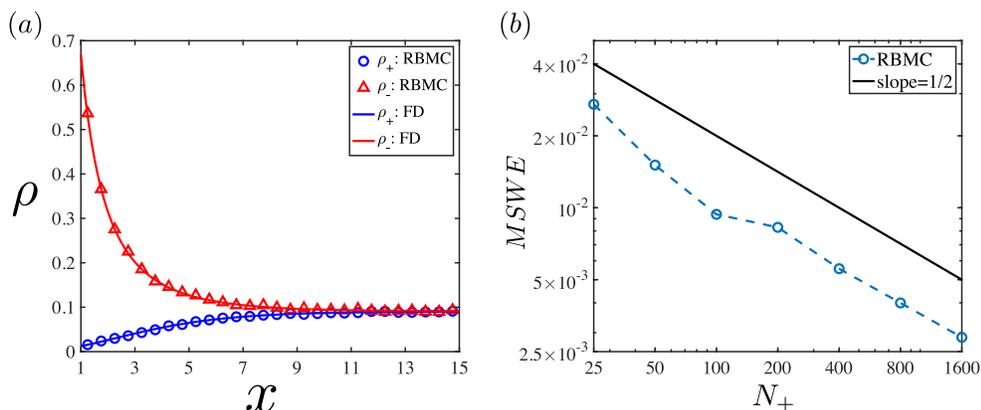


FIG. 1. Solving the 1D PB equation using sampling. (a) The density distribution obtained by RBMC sampling method versus finite difference method. The parameters are detailed in the context. (b) The mean square weak error versus particle number for positive charge (number for negative charge is scaled proportionally). The 1/2 rate is observed.

important). Here,  $\varepsilon = 1$ ,  $Q_f = 1$ ,  $Q_+ = 1$ ,  $Q_- = 2$ , and  $N_+ = 800$  so that  $q = 0.00125$ . We move one particle each iteration in RBMC with the batch size is  $p = 2$ , step size  $\tau = 0.01$ , steps per iteration  $m = 20$ , and the number of iterations in the burn in phase and sampling phase are  $N_b = 5e6$  and  $N_s = 1e6$ .

In Figure 1(b), we show the relative weak error of the sampling method versus the particle number of cations  $N_+$  used in the Gibbs distribution. In one experiment, for each given  $N_+$ , we collect samples every 500 iterations in the sampling phase (i.e.,  $N_{tot} = NN_s/500$  particles) to approximate  $\mathbb{E} \int \varphi d\mu_N$  for the empirical measure  $\mu_N$  and compute

$$(5.3) \quad \text{err} := \left| \int \varphi(x)\rho(x) dx - \frac{1}{N_{tot}} \sum_{i=1}^{N_{tot}} \varphi(x_i) \right| / \int \varphi(x)\rho(x) dx.$$

The reference solution  $\rho$  is computed by the finite difference method as explained in [40]. The test function used is  $\varphi(x) = x^2$  and the errors for the positive and negative particles are computed separately and we repeat the experiments for  $M = 32$  times to get the mean square weak error (MSWE)

$$(5.4) \quad \text{MSWE} = \sqrt{\frac{\sum \text{err}_+^2 + \sum \text{err}_-^2}{2M}}.$$

The results are shown for  $N_+ = 25, 50, 100, 200, 400, 800, 1600$ . Again,  $N_b = 5e6$  and  $N_s = 1e6$ , and  $m = N/15$  in RBMC. Clearly, the convergence rate is 1/2 which is consistent with the result  $\|\mu_N - \rho\|_{H^{-\alpha}} \sim N^{-1/2}$ .

For the 3D case, to run the RBMC algorithm, we let  $r = |x|$  and split the singular Coulomb potential as

$$W^{(1)}(r) = \begin{cases} \frac{1}{4\pi\varepsilon} \left[ -\frac{1}{r_c^2}(r - r_c) + \frac{1}{r_c} \right], & 0 < r \leq r_c, \\ \frac{1}{4\pi\varepsilon} \frac{1}{r}, & r > r_c, \end{cases}$$

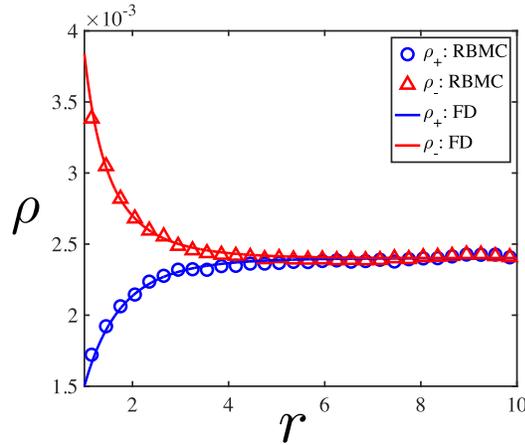


FIG. 2. Solving 3D PB equation by RBMC sampling. The Lennard-Jones potential is used to overcome the singularity of 3D Coulomb potential.

and

$$W^{(2)}(r) = \begin{cases} \frac{1}{4\pi\epsilon} \left[ \frac{1}{r} + \frac{1}{r_c^2}(r - r_c) - \frac{1}{r_c} \right], & 0 < r \leq r_c, \\ 0, & r > r_c. \end{cases}$$

In Figure 2, we present the simulation results for  $\epsilon = 0.01$ ,  $Q_f = 10\epsilon$  centered at the origin,  $Q_+ = 10$ . Here,  $\rho_{\pm}(r)$  are the 3D densities such that  $\int \rho_{\pm}(r)4\pi r^2 dr = Q_{\pm}$ . The cell is  $B(0, 1)$  and  $Q_f$  is put at the origin. The domain is truncated at  $L = 10$ . In the RBMC sampling method, we use  $N_+ = 100$  particles for cations so that  $q = 0.1$ . The cut-off radius of the Coulomb potential is chosen as  $r_c = 0.1$ , while the zero and strength of Lennard-Jones potential  $U_{LJ}$  is  $\sigma = r_c = 0.1$  and  $\epsilon = 0.01$ , respectively. We move one particle each step with batch size  $p = 2$ , steps per iteration  $m = 9$ , step size  $\tau = 0.01$ . Also, when particles fall out of the domain, a simple reflection is performed. The numbers of iterations in the burn-in phase and the sampling phase are  $N_b = 5 \times 10^5$  and  $N_s = 5 \times 10^7$ . To compute the density distributions at equilibrium, we collect the samplers every 1000 iterations in the sampling phase. Hence, the samplers used for cations and anions are  $N_s N_+ / 1000 = 5 \times 10^6$  and  $N_s N_- / 1000 = 5.05 \times 10^6$ , respectively. Clearly, we can see from Figure 2 that the RBMC sampling result matches well with the reference solution, given by the finite difference method as mentioned in [40].

**5.2. Two-layer neural networks.** Following [48, 53], we consider in this subsection the supervised learning task to fit data of the form  $\{(x, y)\}$  using a two-layer neural network of the following form:

$$(5.5) \quad \hat{y}(x; \theta) = \frac{1}{N} \sum_{i=1}^N \sigma_*(x; \theta_i).$$

Here,  $\hat{y}(x; \theta)$  represents the output of the network for input  $x$ , and  $\sigma_*(x; \theta_i)$  is a neuron with parameter  $\theta_i$ . Often

$$(5.6) \quad \sigma_*(x, \theta) = c\sigma(w^T x + b), \quad \theta = (c, w, b),$$

and  $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_N)$ . Hence, the model (5.5) is a two-layer neural network with the hidden layer to be  $x \mapsto w^T x + b$ , and the output layer is  $(\sigma_i) \mapsto \frac{1}{N} \sum_i c_i \sigma_i$  after the activation. The mean square loss function is given by

$$(5.7) \quad R(\boldsymbol{\theta}) = \frac{1}{2} \mathbb{E}_{x,y} |\hat{y}(x; \boldsymbol{\theta}) - y|^2.$$

If we are given  $P$  training data  $\{(x_i, y_i)\}$ , then

$$(5.8) \quad \frac{1}{2} \mathbb{E}_{x,y} |\hat{y}(x; \boldsymbol{\theta}) - y|^2 = \frac{1}{2P} \sum_{j=1}^P |\hat{y}(x_j; \boldsymbol{\theta}) - y_j|^2.$$

If we have the task for fitting the distribution of the data, the expectation is then taken over the population distribution, which is the law of  $(x, y)$  as  $P \rightarrow \infty$ . We will use  $\mathbb{E}$  to mean the expectation over  $(x, y)$ , either over the training data or over the population distribution.

An important observation in [48, 53] is that the loss function can also be written as

$$(5.9) \quad R(\boldsymbol{\theta}) = \frac{1}{2N^2} \sum_{i,i'} W(\theta_i, \theta_{i'}) + \frac{1}{N} \sum_{i=1}^N U(\theta_i) + \frac{1}{2} \mathbb{E}|y|^2 =: E(\mu_N) + C,$$

where

$$(5.10) \quad E(\mu) := \int U \mu(d\theta) + \frac{1}{2} \int W(\theta, \theta') \mu(d\theta) \mu(d\theta'), \quad \mu_N = \frac{1}{N} \sum_{i=1}^N \delta(\cdot - \theta_i),$$

and  $U$  and  $W$  are given, respectively, by

$$(5.11) \quad U(\theta) = -\mathbb{E}[y \sigma_*(x, \theta)], \quad W(\theta, \theta') = \mathbb{E}[\sigma_*(x; \theta) \sigma_*(x; \theta')].$$

This indicates that the loss function is, in fact, an energy functional of the empirical measure of the parameters of the neurons.

The noisy stochastic gradient descent (SGD) algorithm is often used to train the neuron networks [48]

$$(5.12) \quad \theta_i^{k+1} = \theta_i^k - \lambda s_k \theta_i^k + s_k (y_k - \hat{y}_k) \nabla_{\theta_i} \sigma_*(x_k; \theta_i^k) + \sqrt{2s_k / \beta} z_i^k,$$

where  $z_i^k \sim \mathcal{N}(0, I_D)$  where  $D$  is the dimension of  $\theta_i$ ,  $\hat{y}_k = \hat{y}(x; \boldsymbol{\theta}^k)$ . Here,  $\lambda \theta_i^k$  comes from an  $\ell^2$ -regularization [48]. Note that the coefficient used here is slightly different from there, now  $s_k$  directly represents the learning rate. This update rule is clearly an approximation of

$$\theta_i^{k+1} = \theta_i^k - \lambda s_k \theta_i^k - s_k \nabla_{\theta_i} (NR(\boldsymbol{\theta}^k)) + \sqrt{2s_k / \beta} z_i^k,$$

where only one sample  $(x_k, y_k)$  is used to replace the expectation over the distribution of the data in  $R(\boldsymbol{\theta})$ , and this is the mini-batch idea in SGD. Of course, several samples could be used. Note that the rescaled loss (or energy functional)  $NR(\boldsymbol{\theta}^k)$  is used consistent with Remark 2.1. This is the discretization of

$$d\theta_i = -\lambda \theta_i - \nabla_{\theta_i} (NR(\boldsymbol{\theta})) dt + \sqrt{2\beta^{-1}} dB_i.$$

Clearly, the mean field limit of this interacting particle system would be the following nonlinear Fokker–Planck equation:

$$(5.13) \quad \partial_t \rho = \nabla_\theta \cdot (\rho \nabla_\theta \Psi_\lambda) + \beta^{-1} \Delta_\theta \rho,$$

where

$$\Psi_\lambda = U(\theta) + \int W(\theta, \theta') \rho(d\theta') + \frac{\lambda}{2} |\theta|^2 = \frac{\delta}{\delta \rho} \left( E(\rho) + \lambda \int \frac{|\theta|^2}{2} \rho(d\theta) \right).$$

This mean field limit has been justified in a probabilistic sense in [48]. Clearly, for the  $N$ -neuron system, the stationary distribution is

$$(5.14) \quad \rho_N \propto \exp \left( -\beta \left( \sum_i \left( U(\theta_i) + \frac{\lambda}{2} |\theta_i|^2 \right) + \frac{1}{N} \sum_{i,j} W(\theta_i, \theta_j) \right) \right).$$

As proved in [48], the stationary solution of the nonlinear Fokker–Planck equation (5.13) is close to the global minimizer of the two-layer neural networks with infinite width, and has nearly zero training loss.

We thus conclude with the following corollary.

**COROLLARY 5.1.** *Suppose that  $W$  is bounded under suitable choices of activation functions and the domains for  $\theta$ . When  $\beta$  is not very big, as the width of neural network  $N \rightarrow \infty$ , the empirical measure  $\mu_N$  converges to the stationary solution  $\rho$  in  $H^{-s}$  for  $s > D/2$  almost surely. Moreover, the neural network converges to the nearly optimal predictor  $\int \sigma_*(x, \theta) \rho(d\theta)$  for all  $x$ .*

The second part is actually a consequence of Corollary 3.9. In practice, training a very wide network requires more memory. Moreover, it is challenging to reach a true “steady state” during training. The above result then suggests that we may do sampling from the  $N$ -body Gibbs measure to approximate the nearly optimal predictor.

Let us consider a typical regression task. Here, measurement of the output value is taken after each input, with measurement errors following a Gaussian distribution. The goal is to predict new outputs for a new input. Consider the following data which are generated by

$$(5.15) \quad y_i = \sin(3x) + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(0, 0.2), \quad x \sim U[0, 1].$$

We use the two-layer neural network model above to fit the data to approximate the model  $y = f(x) = \sin(3x)$ . Here, we use the sigmoid function  $\sigma(x) = \frac{1}{1+e^{-x}}$  in (5.6). The network width is chosen as  $N = 64$ .

The noisy SGD update (5.12) is applied with  $\lambda = 0$  for optimization. The time step (or learning rate)  $s_k = 10$  and  $\beta = 2000$ . For sampling, we apply RBMC without splitting. Hence, the algorithm becomes almost the same with SGLD or the noisy SGD, with the difference that we are allowed to update a few neurons in each iteration. The stepsize is the same as above. In the RBMC, the warm-up phase and sampling phase are  $N_b = 10000$  and  $N_s = 20000$ . In other words, we take  $NN_s = 1.28 \times 10^6$  samples for the empirical distribution of  $\theta$ . The steady-state distribution is reconstructed using the empirical distribution:

$$(5.16) \quad \tilde{y}(x; \boldsymbol{\theta}) = \mathbb{E}[\sigma_*(x; \boldsymbol{\theta})] \approx \frac{1}{M} \sum_{i=1}^M \sigma_*(x; \theta_i).$$

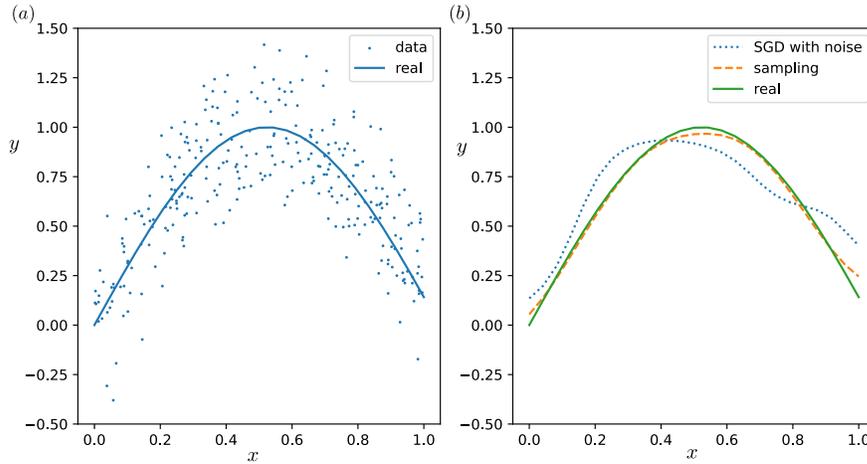


FIG. 3. Training the two-layer neural networks by sampling.

TABLE 1  
Neural network errors.

	$\overline{\text{err}}_{sgd}$	$\text{err}_s$
Training set	0.0478	0.0409
Test set	0.0490	0.0420

Figure 3 illustrates the numerical results. The result by sampling (red dashed line) is closer to the true curve  $y = f(x)$  compared to the SGD training result (blue dotted line). It exhibits better training and generalization errors. This is clearly expected as we used more samples for  $\theta$  and the results by sampling can be thought as the average of  $N_s$  networks with width  $N = 64$ .

Table 1 shows the errors of the neural network. The “ $\overline{\text{err}}_{sgd}$ ” refers to the average of  $N_s$  runs of the SGD mean square errors defined in (5.8), while “ $\text{err}_s$ ” corresponds to the mean square error (5.8) of the empirical distribution. Clearly, the error of the predictor constructed by sampling is smaller.

**6. Conclusions.** Compared to some direct simulation methods like in [40], we have proposed in this work to solve the stationary nonlinear Fokker–Planck equation by sampling from the N-body Gibbs measure, which could be acceptable if some efficient sampling methods based on random batch strategy (e.g., the RBMC) are used. On the theoretical side, we have established the convergence in the mean field limit from the N-body Gibbs measure to the solution of the nonlinear Fokker–Planck equation using relative entropy mainly following the strategy in [34, 44] with a symmetrization technique for the interaction terms. Compared to earlier results using relative entropy based on Laplace asymptotics [1, 2], we do not need the continuity of the interaction kernel and continuity of the free energy functional. On the numerical side, the method has been mainly tested on the Poisson–Boltzmann equation and satisfactory results have been obtained.

For future work, extension to more singular kernels (like the 3D Coulomb potential) could be considered. How to compute a particular stationary solution when the uniqueness is not available is also an important question, which is related to the interesting phase transition problems.

## REFERENCES

- [1] G. B. AROUS AND M. BRUNAUD, *Méthode de laplace: étude variationnelle des fluctuations de diffusions de type "champ moyen"*, Stochastics Stochastics Rep., 31 (1990), pp. 79–144, <https://doi.org/10.1080/03610919008833649>.
- [2] G. BEN AROUS AND O. ZEITOUNI, *Increasing propagation of chaos for mean field models*, Ann. Inst. Henri Poincaré Probab. Stat., 35 (1999), pp. 85–102, [https://doi.org/10.1016/S0246-0203\(99\)80006-5](https://doi.org/10.1016/S0246-0203(99)80006-5).
- [3] S. G. BOBKOV AND F. GÖTZE, *Exponential integrability and transportation cost related to logarithmic Sobolev inequalities*, J. Funct. Anal., 163 (1999), pp. 1–28, <https://doi.org/10.1006/jfan.1998.3326>.
- [4] V. I. BOGACHEV, N. V. KRYLOV, M. RÖCKNER, AND S. V. SHAPOSHNIKOV, *Fokker–Planck–Kolmogorov Equations*, Math. Surveys Monogr. 207, American Mathematical Society, Providence, RI, 2022.
- [5] D. BRESCH, P.-E. JABIN, AND Z. WANG, *On mean-field limits and quantitative estimates with a large class of singular kernels: Application to the Patlak–Keller–Segel model*, C. R. Math. Acad. Sci. Paris, 357 (2019), pp. 708–720, <https://doi.org/10.1016/j.crma.2019.09.007>.
- [6] T. R. BRUMLEVE AND R. P. BUCK, *Numerical solution of the Nernst-Planck and Poisson equation system with applications to membrane electrochemistry and solid state physics*, J. Electroanal. Chem., 90 (1978), pp. 1–31, [https://doi.org/10.1016/S0022-0728\(78\)80137-5](https://doi.org/10.1016/S0022-0728(78)80137-5).
- [7] D. CHAFAÏ, N. GOZLAN, AND P.-A. ZITT, *First-order global asymptotics for confined particles with singular pair repulsion*, Ann. Appl. Probab., 24 (2014), pp. 2371–2413, <https://doi.org/10.1214/13-AAP980>.
- [8] L.-P. CHAINTRON AND A. DIEZ, *Propagation of chaos: A review of models, methods and applications. I. Models and methods*, Kinet. Relat. Models, 15 (2022), pp. 895–1015, <https://doi.org/10.3934/krm.2022017>.
- [9] D. L. CHAPMAN, *A contribution to the theory of electrocapillarity*, Lond. Edinb. Dubl. Phil. Mag., 25 (1913), pp. 475–481, <https://doi.org/10.1080/14786440408634187>.
- [10] J. CHEN AND L. LI, *Fluctuation Suppression and Enhancement in Interacting Particle Systems*, preprint, <https://arxiv.org/abs/2204.07757>, 2022.
- [11] F. CUCKER AND S. SMALE, *Emergent behavior in flocks*, IEEE Trans. Automat. Control, 52 (2007), pp. 852–862, <https://doi.org/10.1109/TAC.2007.895842>.
- [12] D. A. DAWSON, *Critical dynamics and fluctuations for a mean-field model of cooperative behavior*, J. Stat. Phys., 31 (1983), pp. 29–85, <https://doi.org/10.1007/BF01010922>.
- [13] P. DEGOND, J.-G. LIU, AND R. L. PEGO, *Coagulation–fragmentation model for animal group-size statistics*, J. Nonlinear Sci., 27 (2017), pp. 379–424, <https://doi.org/10.1007/s00332-016-9336-3>.
- [14] M. G. DELGADINO, R. S. GVALANI, AND G. A. PAVLIOTIS, *On the diffusive-mean field limit for weakly interacting diffusions exhibiting phase transitions*, Arch. Ration. Mech. Anal., 241 (2021), pp. 91–148, <https://doi.org/10.1007/s00205-021-01648-1>.
- [15] R. L. DOBRUSHIN, *Vlasov equations*, Funct. Anal. Appl., 13 (1979), pp. 115–123, <https://doi.org/10.1007/BF01077243>.
- [16] M. H. DUONG AND J. TUGAUT, *Stationary solutions of the Vlasov-Fokker-Planck equation: Existence, characterization and phase-transition*, Appl. Math. Lett., 52 (2016), pp. 38–45, <https://doi.org/10.1016/j.aml.2015.08.003>.
- [17] M. H. DUONG AND J. TUGAUT, *The Vlasov-Fokker-Planck equation in non-convex landscapes: Convergence to equilibrium*, Electron. Commun. Probab., 23 (2018), pp. 1–10, <https://doi.org/10.1214/18-ECP116>.
- [18] M. H. DUONG AND J. TUGAUT, *Coupled McKean-Vlasov diffusions: Wellposedness, propagation of chaos and invariant measures*, Stochastics, 92 (2020), pp. 900–943, <https://doi.org/10.1080/17442508.2019.1677663>.
- [19] A. FLAVELL, M. MACHEN, B. EISENBERG, J. KABRE, C. LIU, AND X. LI, *A conservative finite difference scheme for Poisson–Nernst–Planck equations*, J. Comput. Electron., 13 (2014), pp. 235–249, <https://doi.org/10.1007/s10825-013-0506-3>.
- [20] D. FRENKEL AND B. SMIT, *Understanding Molecular Simulation: From Algorithms to Applications*, Vol. 1, Elsevier, 2001.
- [21] T. FUNAKI, *A certain class of diffusion processes associated with nonlinear parabolic equations*, Z. Wahrscheinlichkeitstheorie verw. Gebiete, 67 (1984), pp. 331–348, <https://doi.org/10.1007/BF00535008>.
- [22] Y. GAO AND J.-G. LIU, *A selection principle for weak KAM solutions via Freidlin–Wentzell large deviation principle of invariant measures*, SIAM J. Math. Anal., 55 (2023), pp. 6457–6495, <https://doi.org/10.1137/22M1519717>.
- [23] G. GOUY, *Constitution of the electric charge at the surface of an electrolyte*, J. Phys., 9 (1910), pp. 457–468, <https://doi.org/10.1051/jphysap:019100090045700>.

- [24] A. GUILLIN, P. LE BRIS, AND P. MONMARCHÉ, *Some Remarks on the Effect of the Random Batch Method on Phase Transition*, preprint, <https://arxiv.org/abs/2308.01624>, 2023.
- [25] A. GUILLIN, P. LE BRIS, AND P. MONMARCHÉ, *Uniform in time propagation of chaos for the 2D vortex model and other singular stochastic systems*, J. Eur. Math. Soc., (2024), in press, <https://doi.org/10.4171/jems/1413>.
- [26] A. GUILLIN, W. LIU, L. WU, AND C. ZHANG, *Uniform Poincaré and logarithmic Sobolev inequalities for mean field particles systems*, Ann. Appl. Probab., 32 (2022), pp. 1590–1614, <https://doi.org/10.1214/21-AAP1707>.
- [27] M. HAURAY AND S. MISCHLER, *On Kac's chaos and related problems*, J. Funct. Anal., 266 (2014), pp. 6055–6157, <https://doi.org/10.1016/j.jfa.2014.02.030>.
- [28] P. W. HOLLAND, K. B. LASKEY, AND S. LEINHARDT, *Stochastic blockmodels: First steps*, Soc. Networks, 5 (1983), pp. 109–137, [https://doi.org/10.1016/0378-8733\(83\)90021-7](https://doi.org/10.1016/0378-8733(83)90021-7).
- [29] D. HORSTMANN, *From 1970 until present: the Keller-Segel model in chemotaxis and its consequences*. I, Jahresber. Deutsch. Math.-Verein., 105 (2003), pp. 103–165, <https://files-www.mis.mpg.de/mpi-typo3/preprints/2003/preprint2003.3.pdf>.
- [30] Z. HUANG, S. JIN, AND L. LI, *Mean field error estimate of the random batch method for large interacting particle system*, ESAIM: M2AN, 59 (2025), pp. 265–289, <https://doi.org/10.1051/m2an/2024071>.
- [31] P.-E. JABIN, *A review of the mean field limits for Vlasov equations*, Kinet. Relat. Models, 7 (2014), pp. 661–711, <https://doi.org/10.3934/krm.2014.7.661>.
- [32] P.-E. JABIN AND Z. WANG, *Mean field limit and propagation of chaos for Vlasov systems with bounded forces*, J. Funct. Anal., 271 (2016), pp. 3588–3627, <https://doi.org/10.1016/j.jfa.2016.09.014>.
- [33] P.-E. JABIN AND Z. WANG, *Mean field limit for stochastic particle systems*, in Active Particles, Vol. 1, Springer, 2017, pp. 379–402.
- [34] P.-E. JABIN AND Z. WANG, *Quantitative estimates of propagation of chaos for stochastic systems with  $W^{-1,\infty}$  kernels*, Invent. Math., 214 (2018), pp. 523–591, <https://doi.org/10.1007/s00222-018-0808-y>.
- [35] S. JIN, L. LI, AND J.-G. LIU, *Random batch methods (RBM) for interacting particle systems*, J. Comput. Phys., 400 (2020), 108877, <https://doi.org/10.1016/j.jcp.2019.108877>.
- [36] J. E. JONES, *On the determination of molecular fields. –II. From the equation of state of a gas*, in Proceedings of the Royal Society of London Series A, Containing Papers of a Mathematical and Physical Character, 106 (2024), pp. 463–477.
- [37] R. JORDAN, D. KINDERLEHRER, AND F. OTTO, *The variational formulation of the Fokker–Planck equation*, SIAM J. Math. Anal., 29 (1998), pp. 1–17, <https://doi.org/10.1137/S0036141096303359>.
- [38] M. KAC, *Foundations of kinetic theory*, in Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Vol. 3, University of California Press, Berkeley and Los Angeles, CA, 1956, pp. 171–197.
- [39] T. LEBLÉ AND S. SERFATY, *Large deviation principle for empirical fields of Log and Riesz gases*, Invent. Math., 210 (2017), pp. 645–757, <https://doi.org/10.1007/s00222-017-0738-0>.
- [40] L. LI, J.-G. LIU, AND Y. TANG, *Some random batch particle methods for the Poisson-Nernst-Planck and Poisson-Boltzmann equations*, Commun. Comput. Phys., 32 (2022), pp. 41–82, <https://doi.org/10.4208/cicp.OA-2021-0159>.
- [41] L. LI, J.-G. LIU, AND Y. WANG, *Geometric ergodicity of SGLD via reflection coupling*, Stoch. Dyn., 24 (2024), 2450035, <https://doi.org/10.1142/S0219493724500357>.
- [42] L. LI AND Y. WANG, *A Sharp Uniform-in-Time Error Estimate for Stochastic Gradient Langevin Dynamics*, preprint, <https://arxiv.org/abs/2207.09304>, 2022.
- [43] L. LI, Z. XU, AND Y. ZHAO, *A random-batch Monte Carlo method for many-body systems with singular kernels*, SIAM J. Sci. Comput., 42 (2020), pp. A1486–A1509, <https://doi.org/10.1137/19M1302077>.
- [44] T. S. LIM, Y. LU, AND J. H. NOLEN, *Quantitative propagation of chaos in a bimolecular chemical reaction-diffusion model*, SIAM J. Math. Anal., 52 (2020), pp. 2098–2133, <https://doi.org/10.1137/19M1287687>.
- [45] F. MALRIEU, *Logarithmic Sobolev inequalities for some nonlinear PDE's*, Stochastic Process. Appl., 95 (2001), pp. 109–132, [https://doi.org/10.1016/S0304-4149\(01\)00095-3](https://doi.org/10.1016/S0304-4149(01)00095-3).
- [46] F. MALRIEU, *Convergence to equilibrium for granular media equations and their Euler schemes*, Ann. Appl. Probab., 13 (2003), pp. 540–560, <https://doi.org/10.1214/aoap/1050689593>.
- [47] H. P. MCKEAN, *Propagation of chaos for a class of non-linear parabolic equations*, in Stochastic Differential Equations (Lecture Series in Differential Equations, Session 7), Catholic University, 1967, pp. 41–57.

- [48] S. MEI, A. MONTANARI, AND P.-M. NGUYEN, *A mean field view of the landscape of two-layer neural networks*, Proc. Natl. Acad. Sci., 115 (2018), pp. E7665–E7671, [10.1073/pnas.1806579115](https://doi.org/10.1073/pnas.1806579115).
- [49] N. METROPOLIS, A. W. ROSENBLUTH, M. N. ROSENBLUTH, A. H. TELLER, AND E. TELLER, *Equation of state calculations by fast computing machines*, J. Chem. Phys., 21 (1953), pp. 1087–1092, <https://doi.org/10.1063/1.1699114>.
- [50] S. MOTSCH AND E. TADMOR, *Heterophilious dynamics enhances consensus*, SIAM Rev., 56 (2014), pp. 577–621, <https://doi.org/10.1137/120901866>.
- [51] E. RIO, *Moment inequalities for sums of dependent random variables under projective conditions*, J. Theoret. Probab., 22 (2009), pp. 146–163, <https://doi.org/10.1007/s10959-008-0155-9>.
- [52] G. O. ROBERTS AND R. L. TWEEDIE, *Exponential convergence of Langevin distributions and their discrete approximations*, Bernoulli, 2 (1996), pp. 341–363, <https://doi.org/10.2307/3318418>.
- [53] G. M. ROTSKOFF AND E. VANDEN-EIJNDEN, *Trainability and accuracy of neural networks: An interacting particle system approach*, Comm. Pure Appl. Math., 75 (2022), pp. 1889–1935, <https://doi.org/10.1002/cpa.22074>.
- [54] S. SERFATY, *Mean field limit for Coulomb-type flows*, Duke Math. J., 169 (2020), pp. 2887–2935, <https://doi.org/10.1215/00127094-2020-0019>.
- [55] A. S. SZNITMAN, *Topics in propagation of chaos*, Ecole d’Ete de Probabilites de Saint-Flour XIX - 1989, 1464 (1989), pp. 165–251.
- [56] C. VILLANI, *Optimal Transport: Old and New*, Grundlehren Math. Wiss. 338, Springer-Verlag, Berlin, 2009.
- [57] Z. WANG, X. ZHAO, AND R. ZHU, *Gaussian fluctuations for interacting particle systems with singular kernels*, Arch. Ration. Mech. Anal., 247 (2023), 101, <https://doi.org/10.1007/s00205-023-01932-2>.
- [58] M. WELLING AND Y. W. TEH, *Bayesian learning via stochastic gradient Langevin dynamics*, in Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011, pp. 681–688.