TOWARDS MORE THEORETICALLY-GROUNDED PARTICLE OPTIMIZATION SAMPLING FOR DEEP LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Many deep-learning based methods such as Bayesian deep learning (DL) and deep reinforcement learning (RL) have heavily relied on the ability of a model being able to efficiently explore via Bayesian sampling. Particle-optimization sampling (POS) is a recently developed technique to generate high-quality samples from a target distribution by iteratively updating a set of interactive particles, with a representative algorithm the Stein variational gradient descent (SVGD). Though obtaining significant empirical success, the non-asymptotic convergence behavior of SVGD remains unknown. In this paper, we generalize POS to a stochasticity setting by injecting random noise in particle updates, called stochastic particle-optimization sampling (SPOS). Notably, for the first time, we develop non-asymptotic convergence theory for the SPOS framework, characterizing convergence of a sample approximation w.r.t. the number of particles and iterations under both convex- and nonconvex-energy-function settings. Interestingly, we provide theoretical understanding of a pitfall of SVGD that can be avoided in the proposed SPOS framework, i.e., particles tend to collapse to a local mode in SVGD under some particular conditions. Our theory is based on the analysis of nonlinear stochastic differential equations, which serves as an extension and a complementary development to the asymptotic convergence theory for SVGD such as (Liu, 2017). With such theoretical guarantees, SPOS can be safely and effectively applied on both Bayesian DL and deep RL tasks. Extensive results demonstrate the effectiveness of our proposed framework.

1 INTRODUCTION

Recent development of deep-learning techniques have required the ability of an algorithm to efficiently explore some particular space (e.g., parameter space) via Bayesian sampling, due to high model complexity of modern deep models. Meanwhile, recent years have seen the development of scalable sampling methods such as stochastic gradient MCMC (SG-MCMC) (Welling & Teh, 2011; Chen et al., 2014; Ding et al., 2014; Chen et al., 2015) and Stein variational gradient descent (SVGD) (Liu & Wang, 2016) to facilitate big-data analysis. SG-MCMC is a family of scalable Bayesian sampling algorithms relying on Itô diffusions, linear stochastic differential equations (SDE) with appropriately designed coefficients such that the corresponding stationary distributions match a target distribution. One potential issue is that samples might be highly correlated partially due to the nature of Markov chains, leading to low sample efficiency, an undesirable property of SG-MCMC.

SVGD, on the other hand, belongs to the family of particle-based sampling methods, which optimize a set of interacting particles to minimize some distance metric between the target distribution and the distribution induced by the particles. By optimizing the distance measure, one maintains an optimal set of particles at each time. Recent development of SVGD has shown that the underlying mathematical principle is based on a family of nonlinear SDEs, in the sense that coefficients of the SDE depend on the current density of the particles. Though achieving numerous practical successes (Liu & Wang, 2016; Feng et al., 2017; Liu et al., 2017; Haarnoja et al., 2017; Liu & Zhu, 2018), little theory has been developed to understand the convergence property of the algorithm. A recent theoretical development has interpreted SVGD as a special type of gradient flow in the space of probability measures, and developed theory to disclose its asymptotic convergence behavior (Liu, 2017).
Recently, Chen et al. (2018) unified SG-MCMC and SVGD by proposing a particle-optimization sampling framework that interprets both as Wasserstein gradient flows (WGFs). Generally speaking, a WGF is a partial differential equation (PDE) defined on the space of probability measures, describing the evolution of a density function over time. In (Chen et al., 2018), the authors define a WGF by combining the corresponding Fokker-Planck equations for both SG-MCMC and SVGD, and solve it with deterministic particle approximations. However, due to the diffusion nature, deterministic-particle approximation leads to a hard-to-control error, challenging for theoretical analysis.

Based on the unified framework in (Chen et al., 2018), we propose to solve WGFs with stochastic particle approximation, leading to stochastic particle-optimization sampling (SPOS). The idea is instead of solving the WGF with an uncontrollable deterministic approximation for a diffusion term, we solve it stochastically by injecting random noise to the particle-update equations. Remarkably, for the first time, we develop nonasymptotic convergence theory for the family of SPOS algorithms, considering both convex- and nonconvex-energy functions. Different from existing theory for SG-MCMC based algorithms (Teo et al., 2016; Vollmer et al., 2016; Chen et al., 2015; Raginsky et al., 2017; Zhang et al., 2017; Xu et al., 2018), our development relies on the theory of nonlinear SDEs, which is more involved and less explored in literature. Particularly, we adopt tools from granular media equations (Malrieu, 2003; Cattiaux et al., 2008) to develop nonasymptotic error bounds in terms of the 1-Wasserstein distance. Please refer to Section M in the Supplementary Material (SM) for detailed distinctions of our work to existing work. Within our theoretical framework, we provide a formal theoretical understanding of a pitfall of SVGD, i.e., particles tend to collapse to one point under some particular conditions; whereas this can be avoided in the proposed SPOS framework due to the injected random noise. Our theory is general enough to be applied for various deep-learning tasks, including Bayesian deep learning and Bayesian exploration in reinforcement learning. Our extensive experimental results well suggest advantages of our framework compared to existing methods.

2 Preliminaries

This section introduces necessary preliminaries, along with notations used in this paper. For the sake of clarity, through out the paper, we use bold letters to denote variables in continuous-time diffusions and model definitions, e.g., \( \theta_t \) in (1) defined below (indexed by “time” \( t \)). By contrast, normal unbold letters are used to denote parameters in algorithms (discrete solutions of continuous-time diffusions), e.g., \( \theta_k^{(i)} \) in (1) below (indexed by “iteration” \( k \)). For conciseness, all the proofs as well as some extra experimental results are presented in the SM. Discussion on the complexity of our algorithm is also included in Section L of the SM.

2.1 Stochastic Gradient MCMC

In Bayesian sampling, we aim to generate random samples from a posterior distribution \( p(\theta | \mathcal{X}) \propto p(\mathcal{X} | \theta) p(\theta) \), where \( \theta \in \mathbb{R}^d \) represents the model parameter with a prior distribution \( p(\theta) \), and \( \mathcal{X} \triangleq \{ \mathbf{x}_1 \}_{q=1}^N \) represents the observed data with likelihood \( p(\mathcal{X} | \theta) = \prod_q p(\mathbf{x}_q | \theta) \). Define the potential energy as: \( U(\theta) \triangleq -\log p(\mathcal{X} | \theta) - \log p(\theta) = -\sum_{q=1}^N \left( \log p(\mathbf{x}_q | \theta) + \frac{1}{\beta} \log p(\theta) \right) \triangleq \sum_{q=1}^N U_q(\theta) \).

SG-MCMC algorithms belong to diffusion-based sampling methods, where a continuous-time diffusion process is designed such that its stationary distribution matches the target posterior distribution. The diffusion process is driven by a specific stochastic differential equation (SDE). For example, in stochastic gradient Langevin dynamic (SGLD), the SDE endows the following form:

\[
d\theta_t = -\beta^{-1} F(\theta_t) dt + \sqrt{2/\beta} d\mathcal{W}_t, \tag{1}
\]

where \( F(\theta) \triangleq \nabla U(\theta) = \sum_{q=1}^N \nabla_x U_q(\theta) \triangleq \sum_{q=1}^N F_q(\theta) \); \( t \) is the time index; \( \beta > 0 \) is the temperature parameter; and \( \mathcal{W}_t \in \mathbb{R}^d \) is a d-dimensional Brownian motion. More instances of SDEs corresponding to other SG-MCMC algorithms can be defined by specifying different forms of \( F \) and potentially other diffusion coefficients. We focus on SGLD and (1) in this paper, and refer interested readers to Ma et al. (2015) for more detailed description of general SG-MCMC algorithms. Denote the probability density function of \( \theta_t \) in (1) as \( \nu_t \), and \( a \cdot b \triangleq a^\top b \) for two vectors \( a \) and \( b \). It is known that \( \nu_t \) is characterized by the following Fokker-Planck (FP) equation (Risken, 1989):

\[
\partial_t \nu_t = \nabla_x \cdot (\beta^{-1} \nu_t F(\theta) + \beta^{-1} \nabla_x \nu_t), \tag{2}
\]

where the stationary distribution \( \nu_\infty \) equals to our target distribution \( p(\theta | \mathcal{X}) \) according to Chiang & Hwang (1987). SGLD generates samples from \( p(\theta | \mathcal{X}) \) by numerically solving the SDE (1). For
scalloplity, it replaces $F(\theta_t)$ in each iteration with an unbiased evaluation by randomly sampling a subset of $\mathcal{X}$, i.e. $F(\theta_t)$ is approximated by: $G_k \overset{d}{=} \frac{N}{M} \sum_{q \in \mathcal{I}_k} F_q(\theta_t)$, where $\mathcal{I}_k$ is a random subset of $[1, 2, \ldots, N]$ with size $B_k$ in each iteration. Based on the above settings, SGLD uses the Euler method with stepsize $h_k$ to numerically solve (1) and obtains the update equation: $\theta_{k+1} = \theta_k - \beta^{-1} G_k h_k + \sqrt{2\beta^{-1}} h_k \xi_k, \xi_k \sim \mathcal{N}(0, \mathbf{I})$.

### 2.2 Stein variational gradient descent

Different from SG-MCMC, SVGD is a deterministic particle-optimization algorithm that is able to generate samples from a target distribution. In the algorithm, a set of particles interact with each other, driving them to high density locations in the parameter space while keeping them far away from each other with repulsive force. The update equations of the particles follow the fastest descent direction of the KL-divergence between current empirical distribution of the particles and the target distribution, on an RKHS induced by a kernel function $\kappa(\cdot, \cdot)$ (Liu & Wang, 2016). Formally, Liu & Wang (2016) derived the following updating rules for the particles $\{\theta_k^{(i)}\}_{i=1}^M$ at the $k$-th iteration with stepsize $h_k$ and $C_k^{(i)} \overset{d}{=} \frac{N}{M} \sum_{q \in \mathcal{I}_k} F_q(\theta_k^{(i)})$:

$$\theta_{k+1}^{(i)} = \theta_k^{(i)} + h_k \sum_{j=1}^M \left[ \kappa(\theta_k^{(j)}, \theta_k^{(i)}) C_k^{(i)} + \nabla \theta_k^{(i)} \kappa(\theta_k^{(j)}, \theta_k^{(i)}) \right], \quad \forall i$$

(3)

where the first term in the bracket encourages particles to locate on high density modes, and the second term serves as repulsive force that pushes away different particles. Different from SG-MCMC, only particles at the current iteration, $\{\theta_k^{(i)}\}_{i=1}^M$, are used to approximate the target distribution.

### 2.3 Particle-optimization based sampling methods

SG-MCMC and SVGD, though look closely related, behave very differently in terms of algorithms, e.g., stochastic and noninteractive versus deterministic and interactive particle updates. Recently, Chen et al. (2018) proposed a deterministic particle-optimization framework that unifies both SG-MCMC and SVGD. Specifically, the authors viewed both SG-MCMC and SVGD as Wasserstein gradient flows (WGFs) on the space of probabilistic measures, and derived several deterministic particle-optimization techniques for particle evolutions, like what SVGD does. For SG-MCMC, the FP equation (2) for SGLD is a special type of WGFs. Together with an interpretation of SVGD as a special case of the Vlasov equation in nonlinear PDE literature, Chen et al. (2018) proposed a general form of PDE to characterize the evolution of the density for the model parameter $\theta$, denoted as $\nu_t$ at time $t$ with $\nu_\infty$ matching our target (posterior) distribution, i.e.,

$$\partial_t \nu_t = \nabla_\theta \cdot (\nu_t \beta^{-1} F(\theta) + \nu_t (\kappa \ast \nu_t(\theta)) + \beta^{-1} \nabla_\theta \nu_t),$$

(4)

where $\kappa$ is a function controlling the interaction of particles in the PDE system. For example, in SVGD, Chen et al. (2018) showed that $\kappa$ and $\kappa \ast \nu_t(\theta)$ endow the following forms:

$$\kappa(\theta, \theta') \overset{d}{=} F(\theta') \kappa(\theta', \theta) - \nabla_\theta \kappa(\theta', \theta) \text{ and } \kappa \ast \nu_t(\theta) \overset{d}{=} \int \kappa(\theta, \theta') \nu_t(\theta') d\theta'$$

(5)

where $\kappa(\cdot, \cdot)$ is a kernel function such as the RBF kernel. In the following, we introduce a new unary function $K(\theta) = \exp(-\frac{||\theta||^2}{2\beta^2})$, thus $\kappa(\theta, \theta')$ can be rewritten as $\kappa(\theta, \theta') = K(\theta - \theta')$. Hence, (4) with $\kappa$ defined in (5) can be equivalently rewritten as:

$$\partial_t \nu_t = \nabla_\theta \cdot (\nu_t \beta^{-1} F(\theta) + \nu_t (E_{Y \sim \nu_t} K(\theta - Y) F(Y) - \nabla K \ast \nu_t(\theta)) + \beta^{-1} \nabla_\theta \nu_t),$$

(6)

where $Y$ is a random sample from $\nu_t$ independent of $\theta$. Importantly, Proposition 1 (Chen et al. 2018) The stationary distribution of (6) equals to our target distribution, which means $\nu_\infty(\theta) = \tilde{p}(\theta | X)$.

Chen et al. (2018) proposed to solve (6) numerically with deterministic particle-optimization algorithms such as the blob method. Specifically, the continuous density $\nu_t$ is approximated by a set of $M$ particles $\{\theta_k^{(i)}\}_{i=1}^M$ that evolve over time $t$, i.e. $\nu_t \approx \frac{1}{M} \sum_{i=1}^M \delta_{\theta_k^{(i)}}(\theta)$, where $\delta_{\theta_k^{(i)}}(\theta) = 1$ if $\theta = \theta_k^{(i)}$ and 0 otherwise. Note $\nabla_\theta \nu_t$ in (6) is no longer a valid definition when adopting particle approximation for $\nu_t$. Consequently, $\nabla q_{\theta_k} \nu_t$ needs nontrivial approximations, e.g., by discrete gradient flows or blob methods proposed in Chen et al. (2018). We omit the details here for simplicity.
3 Stochastic Particle-Optimization Sampling Algorithms

The deterministic particle-approximation methods proposed by Chen et al. (2013) introduce approximation errors for \( \nabla \theta \nu_t \) that are hard to control analytically. To overcome this problem, we propose to solve (4) stochastically to replace the \( \nabla \theta \nu_t \) term with a Brownian motion. Specifically, first note that the term \( \beta^{-1} \nabla \theta \cdot \nabla \theta \nu_t \) is contributed from Brownian motion, i.e., solving the SDE, \( d\theta_t = \sqrt{2\beta^{-1}}dW_t \), is equivalent to solving the corresponding FP equation: \( \partial \theta_t = \beta^{-1} \nabla \theta \cdot \nabla \theta \nu_t \). Consequently, we decompose RHS of (4) into two parts: \( F_1 \triangleq \nabla \theta \cdot (\nu_t \beta^{-1} F(\theta_t) + (K \ast \nu_t) \nu_t) \) and \( F_2 \triangleq \beta^{-1} \nabla \theta \cdot \nabla \theta \nu_t \). Our idea is to solve \( F_1 \) deterministically under a PDE setting, and solve \( F_2 \) stochastically based on its corresponding SDE. When adopting particle approximation for the density \( \nu_t \), both solutions of \( F_1 \) and \( F_2 \) are represented in terms of particles \( \{\theta_t^{(i)}\} \). Thus we can combine the solutions from the two parts directly to approximate the original exact solution of (4).

Similar to the results of SVGD in Section 3.3 in (Li 2017), we first formally show in Theorem 2 that when approximating \( \nu_t \) with particles, i.e., \( \nu_t \approx \frac{1}{M} \sum_{i=1}^{M} \delta_{\theta_t^{(i)}}(\theta) \), the PDE can be transformed into a system of deterministic differential equations with interacting particles.

**Theorem 2** When approximating \( \nu_t \) in (4) with particles \( \{\theta_t^{(i)}\} \), the PDE \( \partial \nu_t = F_1 \) reduces to the following system of differential equations describing evolutions of the particles over time: \( \forall i \)

\[
d\theta_t^{(i)} = -\beta^{-1} F(\theta_t^{(i)})dt - \frac{1}{M} \sum_{j=1}^{M} K(\theta_t^{(i)} - \theta_t^{(j)})F(\theta_t^{(j)})dt + \frac{1}{M} \sum_{j=1}^{M} \nabla K(\theta_t^{(i)} - \theta_t^{(j)})dt \tag{7}
\]

On the other hand, by solving \( \partial \nu_t = F_2 \) stochastically with its equivalent SDE counterpart, we arrive at the following differential equation system, describing evolutions of the particles \( \{\theta_t^{(i)}\} \) over time \( t \): \( \forall i \)

\[
d\theta_t^{(i)} = \left( -\beta^{-1} F(\theta_t^{(i)}) - \frac{1}{M} \sum_{j=1}^{M} K(\theta_t^{(i)} - \theta_t^{(j)})F(\theta_t^{(j)}) + \frac{1}{M} \sum_{j=1}^{M} \nabla K(\theta_t^{(i)} - \theta_t^{(j)}) \right) dt + \sqrt{2\beta^{-1}}dW_t^{(i)} \tag{8}
\]

Our intuition is that if the particle evolution (8) can be solved exactly, the solution of (4) \( \nu_t \) will be well-approximated by the particles \( \{\theta_t^{(i)}\} \). In our theory, we show this intuition is actually true. In practice, however, solving (8) is typically infeasible, and thus numerical methods are adopted. Furthermore, in the case of big data, following SG-MCMC, \( F(\theta_k^{(i)}) \) is typically replaced by a stochastic version \( G_k^{(i)} \triangleq \frac{N}{B_k} \sum_{q \in \mathcal{I}_k} F_q(\theta_k^{(i)}) \) evaluated with a minibatch of data of size \( B_k \) for computational feasibility. Based on the Euler method (Chen et al. 2015) with a stepsize \( h_k \), (8) leads to the following updates for the particles at the \( k \)-th iteration

\[
\theta_{k+1}^{(i)} = \theta_k^{(i)} - h_k(\beta^{-1} G_k^{(i)} - \frac{h_k}{M} \sum_{j=1}^{M} K(\theta_k^{(i)} - \theta_k^{(j)})G_j^{(j)} + \frac{h_k}{M} \sum_{j=1}^{M} \nabla K(\theta_k^{(i)} - \theta_k^{(j)}) + \sqrt{2\beta^{-1}h_k} \xi_k^{(i)}, \xi_k^{(i)} \sim \mathcal{N}(0, I) \quad \forall i \). \tag{9}
\]

We called the algorithm with particle update equations (9) stochastic particle-optimization sampling (described in Algorithm 3), in the sense that particles are optimized stochastically with extra random Gaussian noise. Intuitively, the added Gaussian noise would enhance the ability of the algorithm to jump out of local modes, leading to better ergodic properties compared to standard SVGD. This serves as one of our motivations to generalize SVGD to SPOS. To illustrate the advantage of introducing the noise term, we compare SPOS and SVGD on sampling a difficult multi-mode distribution, with the density function given in Section A of the SM. The particles are initialized on a local mode close to zero. Figure 1 plots the final locations of the particles along with the true density, which shows that particles in SPOS are able to reach different modes, while they are all trapped at one mode in SVGD. This pitfall of SVGD will be studied formally in Section 4.4.

![Figure 1: Comparison of SPOS (left) and SVGD (right) on a multi-mode distribution. The circles with different colors are the resulting 100 particles, which are able to spread over all modes for SPOS.](image-url)
4 NON-ASYMPTOTIC CONVERGENCE ANALYSIS: THE CONVEX CASE

In this section, we develop non-asymptotic convergence theory for the proposed SPOS when the energy function $U(\theta)$ is convex. The nonconvex case is discussed in Section 5. We prove non-asymptotic convergence rates for SPOS algorithm under the 1-Wasserstein metric $W_1$, a special case of p-Wasserstein metric defined as $W_p(\mu, \nu) = \left( \int_{\mathbb{R}^d \times \mathbb{R}^d} \|X_\mu - X_\nu\|^p d\zeta(X_\mu, X_\nu) \right)^{1/p}$, where $\zeta(\mu, \nu)$ is the joint distribution on $\mathbb{R}^d \times \mathbb{R}^d$ with marginal distribution $\mu$ and $\nu$. Note that SPOS reduces to SVGD when $\beta \to \infty$, thus our theory also sheds light on the convergence behavior of SVGD, where non-asymptotic theory is currently missing, despite the asymptotic theory developed recently (Liu 2017; Lu et al. 2018). It is worth noting that part of our proofs are generalization of techniques for analyzing granular media equations in (Malrieu 2003; Cattiaux et al. 2008).

4.1 Basic setup and assumptions

Due to the exchangeability of the particle system $\{\theta_t^{(i)}\}^M_{i=1}$ in (8), if we initialize all the particles $\theta_t^{(i)}$ with the same distribution $\rho_0$, they would endow the same distribution for each time $t$. We denote the distribution of each $\theta_t^{(i)}$ as $\rho_t$. Similar arguments hold for the particle system $\{\theta_k^{(i)}\}^M_{i=1}$ in (9), and thus we denote the distribution of each $\theta_k^{(i)}$ as $\mu_k$. To this end, our analysis aims at bounding $W_1(\mu_T, \nu_\infty)$ since $\nu_\infty$ equals to our target distribution $\nu(\theta, \mathcal{X})$ according to Proposition 3. Before proceeding to our theoretical results, we first present the following basic assumptions.

**Assumption 1** Assume $F$ and $K$ satisfy the following conditions:

- There exist positive $m_F$ and $m_K$, such that $\langle F(\theta) - F(\theta'), \theta - \theta' \rangle \geq m_F \|\theta - \theta'\|^2$ and $\langle \nabla K(\theta) - \nabla K(\theta'), \theta - \theta' \rangle \leq -m_K \|\theta - \theta'\|^2$.
- $F$ is bounded by $H_F$ and $L_F$-Lipschitz continuous i.e., $\|F(\theta')\| \leq H_F$ and $\|F(\theta_1) - F(\theta_2)\| \leq L_F \|\theta_1 - \theta_2\|$, $K$ is $L_K$-Lipschitz continuous.
- $K$ is an even function, i.e., $K(-\theta) = K(\theta)$.

Note the first bullet indicates $U$ to be a convex function and $W$ to be a concave function. For an RBF kernel, the later could be achieved by setting the bandwidth large enough and only considering the concave region for simplicity. This assumption is used for revealing some undesired property of SVGD developed below. We do not need such an assumption when analyzing under a nonconvex energy function $U$ in Section 5.

The high-level idea of bounding $W_1(\mu_T, \nu_\infty)$ in this section is to decompose it as follows:

$$ W_1(\mu_T, \nu_\infty) \leq W_1(\mu_T, \rho_{\sum_{k=0}^{T-1} h_k}) + W_1(\rho_{\sum_{k=0}^{T-1} h_k}, \nu_{\sum_{k=0}^{T-1} h_k}) + W_1(\nu_{\sum_{k=0}^{T-1} h_k}, \nu_\infty). \quad (10) $$

4.2 BOUNDS WITH STOCHASTIC PARTICLE APPROXIMATION

We firstly bound $W_1(\rho_{\sum_{k=0}^{T-1} h_k}, \nu_{\sum_{k=0}^{T-1} h_k})$ and $W_1(\nu_{\sum_{k=0}^{T-1} h_k}, \nu_\infty)$ with the following theorems.

**Theorem 3** Under Assumption 2 and let $\rho_0 = \nu_0$, there exist some positive constants $c_1$ and $c_2$ independent of $(M, d)$ and satisfying $c_2 < \beta^{-1}$ such that

$$ W_1(\rho_t, \nu_t) \leq c_1 (\beta^{-1} - c_2)^{-1} M^{-1/2}, \forall t. \quad (11) $$

**Remark 1** According to Theorem 3, we can bound the $W_1(\rho_{\sum_{k=0}^{T-1} h_k}, \nu_{\sum_{k=0}^{T-1} h_k})$ term as $W_1(\rho_{\sum_{k=0}^{T-1} h_k}, \nu_{\sum_{k=0}^{T-1} h_k}) \leq \frac{c_1}{\sqrt{M} (\beta^{-1} - c_2)}$. Furthermore, by letting $t \to \infty$, we have $W_1(\nu_\infty, \nu_\infty) \leq \frac{c_1}{\sqrt{M} (\beta^{-1} - c_2)}$, which is an important intermediate result to prove the following theorem.

---

**Algorithm 1** Stochastic Particle-Optimization Sampling

**Input:** Initial particles $\{\theta_0^{(i)}\}^M_{i=1}$ with $\theta_0^{(i)} \in \mathbb{R}^d$, step size $h_k$, batch size $B_k$

1: for iteration $k=0,1,\ldots,T$ do
2: Update $\theta_k^{(i)}$ with (7) for $\forall i$.
3: end for

**Output:** $\{\theta_T^{(i)}\}^M_{i=1}$
Then the EPD of SPOS is bounded as: \( EPD(\nu_0) \leq c_3 e^{-2\lambda t}, \) where \( \lambda_1 = \beta^4 m_F - 3H_F L_W - 2L_F \) and \( c_3 \) is some positive constant independent of \((M, d).\) Hence, the EPD for SVGD can be obtained by taking the repulsive force between particles, e.g., when particles are initialized with the same value \( i \rightarrow \infty \) discussed in Section 4.4.

4.3 Bounds with a Numerical Solution

To bound the \( W_1(\mu_T, \rho_{\sum_{k=0}^{T-1} h_k}) \) term, we adopt techniques from \cite{Raginsky2017, Xu2018} on analyzing the behaviors of SGLD, and derive the following results for our SPOS algorithm.

Theorem 5 Under Assumptions 7, for a fixed step size \( h_k = h \) that is small enough, the corresponding \( W_1(\mu_T, \rho_{Th}) \) is bounded as
\[
W_1(\mu_T, \rho_{Th}) \leq c_4 M d^\frac{3}{2} (c_5 \beta^3 B^{-1} + c_6 h) \frac{1}{2} T \frac{1}{2} h^2
\]
where \( B \) is the fixed size of the minibatch in each iteration and \( (c_4, c_5, c_6) \) are some positive constants.

The dependence of \( T \) in the above bound makes the bound relatively loose. Fortunately, we can improve the bound to make it independent of \( T \) by considering a decreasing-stepsize SPOS algorithm, stated in Theorem 6.

Theorem 6 Under Assumptions 7, for a decreasing step size \( h_k = h_0/(k+1) \), and let the minibatch size in each iteration \( k \) be \( B_k = B_0 + \log(k+1)/(100/99) \), the corresponding \( W_1(\mu_T, \rho_{\sum_{k=0}^{T-1} h_k}) \) term is bounded, for some \( \beta \) small enough, as
\[
W_1(\mu_T, \rho_{\sum_{k=0}^{T-1} h_k}) \leq c_4 \beta^3 M d^\frac{3}{2} (c_7 h_0^3 + c_8 \beta^3 h_0/B_0 + c_9 h_0^2 \beta^2)^{1/2},
\]
where \( B_0 \) is the initial minibatch size, and \( (c_4, c_7, c_8, c_9) \) are some positive constants.

4.4 A Pitfall of SVGD

Based on the above analysis, we now formally show a pitfall of SVGD, i.e., particles in SVGD tend to collapse to a local mode under some particular conditions. Inspired by the work on analyzing the granular media equations by \cite{Malrieu2003, Cattiaux2008}, we measure this by calculating the expected distance between particles, called expected particle distance (EPD). Firstly, we bound the EPD for the proposed SPOS algorithm in Theorem 7.

Theorem 7 Under Assumption 7, further assuming every \( \{\theta_{t}^{(i)}\} \) with the same initial probability law \( \rho_0 \) and \( \Gamma \),\( \|\theta_0 - \theta'_0\|^2 < \infty \). Choose a \( \beta \) such that \( \lambda = \frac{\beta^4 m_F}{2} + m_F - H_F L_K > 0. \) Then the EPD of SPOS is bounded as: \( EPD \leq \sqrt{\sum_{i,j} M E\|\theta_{t}^{(i)} - \theta_{t}^{(j)}\|^2} \leq C_1 e^{-2\lambda t} + 4\sqrt{\frac{d M}{\lambda}}, \)
where \( C_1 = M(M-1)\Gamma - 4\sqrt{d^3 M/\lambda}. \)

Remark 2 There are two interesting cases: i) When \( C_1 > 0, \) the EPD would decrease to the bound \( 4\sqrt{d^3 M/\lambda} \) along time \( t. \) This represents the phenomenon of an attraction force between particles; ii) When \( C_1 < 0, \) the EPD would increase to the same bound, which represents the phenomenon of a repulsive force between particles, e.g., when particles are initialized with the same value \( (\Gamma = 0), \) they would be pushed away from each other until the EPD increases to the aforementioned bound.

Intuitively, the EPD for SVGD can be obtained by taking the \( \beta \rightarrow \infty \) limit. Corollary 8 formally characterizes the particle-degeneracy phenomenon of SVGD, which has been empirically studied in \cite{Zhuo2018}.

Under review as a conference paper at ICLR 2019
Corollary 8 Under the same conditions of Theorem 7 the EPD in SVGD is bounded as: \( \text{EPD} \triangleq \sqrt{\sum_{i,j} M \left| \theta_{i}^{(1)} - \theta_{j}^{(1)} \right|^2} \leq C_0 e^{-2\lambda t} \), where \( C_0 = \sqrt{\sum_{i,j} M \left| \theta_{0}^{(1)} - \theta_{0}^{(1)} \right|^2} \) and \( \lambda = m_{K} - H_{F}L_{K} \).

Remark 3 We would like to emphasize two points: 1) In the case of \( \lambda \geq 0 \), Corollary 3 indicates that particles in SVGD would collapse to a point when \( t \to \infty \). In practice, we usually find that particles are trapped in a local mode instead of collapsing. This is due to two reasons: i) numerical errors inject noise into the particles; ii) some particles are out of the concave region of \( K \) stated in Assumption 7 in SVGD, which is required for the theory to hold. All these make the empirical EPD not exactly the same as the true particle distance. 2) Corollary 8 also applies when the energy function is nonconvex. Our proof in the SM considers the nonconvex case as well. Consequently, this serves as a strong theoretical motivation to apply SPOS instead of SVGD in deep learning.

5 Non-Asymptotic Convergence Analysis: The Nonconvex Case

Since the non-convex case is much more complicated than the convex case, we reply on different assumptions and adopt another distance metric, denoted as \( \mathcal{B} \), to characterize the convergence behavior of SPOS under the non-convex case. Specifically, \( \mathcal{B}(\mu, \nu) \) is defined as \( \mathcal{B}(\mu, \nu) \triangleq \left| \mathbb{E}_{\theta \sim \mu}[f(\theta)] - \mathbb{E}_{\theta \sim \nu}[f(\theta)] \right| \) for a known \( L_{F} \)-continuous function \( f \) satisfying Assumption 2 below. Note such metric has also been adopted in (Vollmer et al., 2016; Chen et al., 2015). Our analysis considers \( (T, M, h) \) as variables in \( \mathcal{B} \). In addition, we use \( \{\hat{\theta}_{k}^{(1)}, \ldots, \hat{\theta}_{k}^{(M)}\} \) to denote the particles when full gradients are adopted at \( \hat{\theta}_{k} \). The distribution of the particles is denoted as \( \hat{\mu}_{k} \).

Our high-level idea of bounding \( \mathcal{B}(\mu_{T}, \nu_{\infty}) \) is to decompose it as follows:

\[
\mathcal{B}(\mu_{T}, \nu_{\infty}) \leq \mathcal{B}(\mu_{T}, \tilde{\mu}_{T}) + \mathcal{B}(\tilde{\mu}_{T}, \hat{\mu}_{\infty}) = \mathcal{B}(\tilde{\mu}_{T}, \hat{\mu}_{\infty}) \leq \mathcal{B}(\mu_{T}, \tilde{\mu}_{T}) + \mathcal{B}(\tilde{\mu}_{T}, \hat{\mu}_{\infty}) \leq \mathcal{B}(\mu_{T}, \tilde{\mu}_{T}) + \mathcal{B}(\tilde{\mu}_{T}, \hat{\mu}_{\infty}) (15)
\]

Our second idea is to concatenate the particles at each time into a single vector representation, i.e. defining the new parameter at time \( \Theta_{t} \) as \( \Theta_{t} \triangleq \{\theta_{1}^{(1)}, \ldots, \theta_{t}^{(M)}\} \in \mathbb{R}^{Md} \). Consequently, the nonlinear SDE system (8) can be turned into a linear SDE, which means \( \Theta_{t} \) is driven by the following linear SDE:

\[
d\Theta_{t} = -F_{\Theta}(\Theta_{t})dt + \sqrt{2\beta^{-1}}dW_{t}^{(Md)}, (16)
\]

where \( F_{\Theta}(\Theta_{t}) \triangleq [\beta^{-1}F(\theta_{t}^{(1)}) - \frac{1}{M} \sum_{j=1}^{M} \nabla K(\theta_{t}^{(1)} - \theta_{t}^{(j)}) + \frac{1}{M} \sum_{j=1}^{M} K(\theta_{t}^{(1)} - \theta_{t}^{(j)}) + \frac{1}{M} \sum_{j=1}^{M} K(\theta_{t}^{(M)} - \theta_{t}^{(j)}) F(\theta_{t}^{(j)})] \) is a vector function \( \mathbb{R}^{Md} \to \mathbb{R}^{Md} \) and \( W_{t}^{(Md)} \) is Brownian motion of dimension \( Md \). Similarly, we can define \( \Theta_{t} \triangleq [\theta_{1}^{(1)}, \ldots, \theta_{t}^{(M)}] \in \mathbb{R}^{Md} \) for the full-gradient case. Hence, it can be seen that through such a decomposition in (15), the bound related to a nonlinear SDE system (8) reduces to that of a linear SDE. The second term \( \mathcal{B}(\tilde{\mu}_{T}, \tilde{\mu}_{\infty}) \) reflects the geometric ergodicity of a linear dynamic system with a numerical method. It is known that even if a dynamic system has an exponential convergence rate to its equilibrium, its corresponding numerical method might not. Our bound for \( \mathcal{B}(\tilde{\mu}_{T}, \tilde{\mu}_{\infty}) \) is essentially a specification of the result of Mattingly et al. (2002), which has also been applied by Xu et al. (2018). The third term \( \mathcal{B}(\hat{\mu}_{\infty}, \nu_{\infty}) \) reflects the numerical error of a linear SDE, which has been studied in related literature such as (Chen et al., 2015). To this end, we adopt standard assumptions used in the analysis of SDEs (Vollmer et al., 2016; Chen et al., 2015), rephrased in Assumption 2.

Assumption 2 (Assumption in (Vollmer et al., 2016; Chen et al., 2015)) For the linear SDE (16) and a Lipschitz function \( f \), let \( \psi \) be the solution functional of the Poisson equation: \( \hat{\psi}(\Theta_{k}) = \frac{1}{M} \sum_{j=1}^{M} f(\theta_{k}^{(j)}) - \mathbb{E}_{\theta \sim \mu(\Theta_{k})}[f(\theta)] \), where \( \hat{\psi} \) denotes the infinite generator of the linear SDE (16). Assume \( \psi \) and its up to 4th-order derivatives, \( D^{k}\psi \), are bounded by a function \( V \), i.e., \( \|D^{k}\psi\| \leq H_{k}V^{p} \) for \( k = 0, 1, 2, 3, 4, H_{k}, p_{k} \geq 0 \). Furthermore, the expectation of \( V \) on \( \{\Theta_{t}\} \) is bounded: \( \sup_{t} \mathbb{E}V_{p}(\Theta_{t}) < \infty \), and \( V \) is smooth such that \( \sup_{t} \mathbb{E}V_{p}(s\Theta_{t}) \leq H_{k}V^{p}(s\Theta_{t}) \leq H_{k}V^{p}(\Theta_{t}) \) for all \( s \leq \min\{2p_{k}\} \).

Assumption 3 i) \( F \), \( K \) and \( V \) are \( L_{F} \), \( L_{K} \) and \( L_{V} \) Lipschitz; ii) \( F \) satisfies the dissipative property, i.e., \( (F(\theta,\Theta_{t}) \geq m \|\theta\|^{2} - b \) for some \( m, b > 0 \); iii) Remark 7 applies to the nonconvex setting, i.e. \( \sup_{t} \|\theta_{t}\|_{L_{F}} \leq \|\Theta_{0}\|_{L_{F}} \).
Remark 4 Assumption 2 is necessary to control the gap between a numerical solution and the exact solution of an SDE. Specifically, it is used to bound the $\tilde{B}(\hat{\mu}_\infty, \rho_\infty)$ term and the $\tilde{B}(\mu_T, \hat{\mu}_T)$ term above. Purely relying on the dissipative assumption in Assumption 3 as in non-convex optimization with SG-MCMC (Raginsky et al., 2017; Xu et al., 2018) would induce a bound increasing linearly w.r.t. time $t$. Thus it is not suitable for our goal. Finally, iii) in Assumption 3 is a mild and reasonable condition because we expect particles to be able to approximate all distributions equally well in the asymptotic limit of $t \to \infty$ by ergodicity due to the injected noise. How to remove/replace this assumption is an interesting future work.

Based on the assumptions above, the bounds for $\tilde{B}(\mu_T, \hat{\mu}_T)$ and $\tilde{B}(\hat{\mu}_\infty, \rho_\infty)$ are summarized below.

Theorem 9 Under Assumption 2,3 if we set the stepsize $h_k = h$, we can have the following results:

\[
\tilde{B}(\mu_T, \hat{\mu}_\infty) \leq C_2\varsigma^{-M^2/2}(1 + \varsigma e^{m_\theta h}) \exp \left(-2m_\theta T \sigma \sigma M d / \log(\varsigma)\right) \tag{17}
\]

\[
\tilde{B}(\hat{\mu}_\infty, \rho_\infty) \leq C_3 h / \beta, \tag{18}
\]

where $\varsigma = 2L_\theta (Mb\beta + m_\theta \beta + Md)/m_\theta$, $L_\theta = \sqrt{2}\beta^{-1}L_F + l'$, $m_\theta = \beta^{-1}m - m'$, and $(\sigma, C_2, C_3, l, m)$ are some positive constants independent of $(T, M, h)$ and $\sigma \in (0, 1)$

Remark 5 It is seen that in order to make $\tilde{B}(\mu_T, \hat{\mu}_\infty)$ term asymptotically decrease to zero, the number of running iteration $T$ should increase at a rate faster enough to compensate the effect of increasing $M$. We believe there is room for improving this bound, which is an interesting future work.

Next we bound the $\tilde{B}(\mu_T, \hat{\mu}_T)$ term related to stochastic gradients. By adapting results from linear SDE (Xu et al., 2018), $\tilde{B}(\mu_T, \hat{\mu}_T)$ can be bounded with Theorem 10.

Theorem 10 Under Assumptions 2,3 if we set $B_k = B$ and $h_k = h$, $\tilde{B}(\mu_T, \hat{\mu}_T)$ is bounded as

\[
\tilde{B}(\mu_T, \hat{\mu}_T) \leq C_5 T h (L_\theta \Gamma' + MC_4) \left(\frac{(6 + 2\Gamma')\beta}{BM}\right)^{1/2}.
\]

where $\Gamma' = 2(1 + 1/m_\theta)(Mb + 2M^2C_4^2 + Md/\beta)$ and $(C_4, C_5)$ is some positive constant independent of $(T, M, h)$

Finally, by combining the results from Theorem 9,10 and iii) in Assumption 3, we arrive at a bound for $\tilde{B}(\mu_T, \nu_\infty)$, summarized in Theorem 11.

Theorem 11 Under Assumptions 2,3 there exist some positive constants $(C_2, C_3, C_4, C_5, C_6)$ such that:

\[
\tilde{B}(\mu_T, \nu_\infty) \leq C_2\varsigma^{-M^2/2}(1 + \varsigma e^{m_\theta h}) \exp \left(-2m_\theta T \sigma \sigma M d / \log(\varsigma)\right) + C_3 h / \beta + C_5 T h (L_\theta \Gamma' + MC_4) \left(\frac{(6 + 2\Gamma')\beta}{BM}\right)^{1/2} + C_6 M^{-1/2},
\]

where $\sigma$, $\varsigma$ and $\Gamma'$ are the same as those in Theorem 9,10

6 Experiments

In this section, we illustrate the effectiveness of our proposed Bayesian sampling framework on several deep-learning models, including Bayesian learning of deep neural network and Bayesian exploration in deep reinforcement learning. We start with a simple illustrative example.

6.1 Bounds illustration with a simple Gaussian example

We follow Chen et al. (2015) and consider a standard Gaussian model where $x_i \sim \mathcal{N}(\theta, 1), \theta \sim \mathcal{N}(0, 1)$. 1000 data samples

Figure 2: The estimation errors with different particles.
\{x_i\} are generated, and every minibatch in the stochastic gradient is of size 10. The test function is defined as \( f(\theta) \equiv \theta^2 \), with explicit expression for the posterior average. To evaluate the expectations in the bias and MSE, we average over 200 runs with random initializations. The estimation errors are plotted in Figure 2. It is seen from the figure that at the beginning, the errors for the ones with less particles decrease faster than those with more particles. This is reflected in the bounds in Theorem 15 which are dominated by the bound in Theorem 3 (indicating larger \( M \) results in larger errors at the beginning). When more running time/iterations are given, the increase of error in Theorem 5 by increasing \( M \) is essentially canceled out the exponentially-decay term in Theorem 4. According to Theorem 3, the error would eventually decreasing with increasing number of particles.

6.2 Bayesian Neural Networks for Regression

Table 1: Averaged predictions with standard deviations in terms of RMSE and log-likelihood on test sets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Test RMSE</th>
<th>Test Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boston_Housing</td>
<td>2.937 ± 0.099</td>
<td>-2.054 ± 0.029</td>
</tr>
<tr>
<td>Concrete</td>
<td>5.324 ± 0.104</td>
<td>-3.082 ± 0.018</td>
</tr>
<tr>
<td>Energy</td>
<td>1.374 ± 0.045</td>
<td>-1.152 ± 0.071</td>
</tr>
<tr>
<td>Kin8nm (0.4)</td>
<td>0.090 ± 0.001</td>
<td>0.984 ± 0.008</td>
</tr>
<tr>
<td>Naval (0.4)</td>
<td>0.006 ± 0.000</td>
<td>0.093 ± 0.013</td>
</tr>
<tr>
<td>CCPP</td>
<td>4.033 ± 0.033</td>
<td>-2.815 ± 0.008</td>
</tr>
<tr>
<td>Winequality</td>
<td>0.609 ± 0.010</td>
<td>-0.925 ± 0.014</td>
</tr>
<tr>
<td>Yacht (0.4)</td>
<td>0.864 ± 0.052</td>
<td>-1.225 ± 0.042</td>
</tr>
<tr>
<td>Protein</td>
<td>4.606 ± 0.013</td>
<td>-2.947 ± 0.003</td>
</tr>
<tr>
<td>YearPredict</td>
<td>8.684 ± 0.820</td>
<td>-3.580 ± 0.216</td>
</tr>
</tbody>
</table>

We conduct experiments for Bayesian learning of DNNs, where we Bayesian DNNs are used to model weight uncertainty of neural networks, an important topic that has been well explored (Hernández-Lobato & Adams, 2015). Blundell et al. (2015) and Li et al. (2016) studied Stein variational inference and its application to neural networks. We assign simple isotropic Gaussian priors to the weights, and perform posterior sampling with different methods. For all methods, we use a RBF kernel \( k(\theta, \theta') = \exp(-\|\theta - \theta'\|_2^2/\eta^2) \), with the bandwidth set to \( \eta = \text{med}^2/\log M \). Here \( \text{med} \) is the median of the pairwise distance between particles. We use a single-layer BNN for regression tasks. Following Li et al. (2015), 10 UCI public datasets are considered: 100 hidden units for 2 large datasets (Protein and YearPredict), and 50 hidden units for the other 8 small datasets. Following Zhang et al. (2016), we repeat the experiments 20 times with batchsize 100 for all datasets except for Protein and YearPredict, which we repeat 5 times and once with batchsize 1000. The datasets are randomly split into 90% training and 10% testing. We adopt the root mean squared error (RMSE) and test log-likelihood as the evaluation criteria. The experimental results are shown in Table I from where we can see the proposed SPOS outperforms other methods, achieving state-of-the-art results.

6.3 Bayesian Exploration in Deep Reinforcement Learning

Following Liu et al. (2017), Zhang et al. (2018a), we define policies in RL with Bayesian neural networks. This naturally introduces uncertainty into action selections under a specific state-action pair, rendering Bayesian explorations to make policy learning more effective.

Specifically, denote the policy as \( \pi(a|s) \) parameterized by \( \theta \) with prior distribution \( p(\theta) \), where \( a \) represent the action variable, and \( s \) the state variable. According to Liu et al. (2017), learning the optimal policy corresponds to calculating the following posterior distribution for \( \theta \):

\[
q(\theta) \propto \exp \left( \frac{1}{\alpha} J(\theta) \right) p(\theta),
\]

where \( J(\theta) \) denotes the expected cumulative reward under the policy with parameter \( \theta \) and \( \alpha \) a hyperparameter. Consequently, \( \theta \) could be updated by drawing samples from (19) with the proposed SPOS. We denote this method as SPOS-PG. In addition, when drawing samples with SVGD, the resulting algorithm is called Stein variational policy gradient (SVPG) (Liu et al., 2017). Note in implementation, the term \( J(\theta) \) can be approximated with REINFORCE (Williams, 1992) or advantage actor critic (Schulman et al., 2015), which we will investigate in our experiments.

We follow the same setting as in Liu et al. (2017), and conduct experiments on three classical continuous control tasks are considered: Cartpole Swing-Up, Double Pendulum, and Cartpole. Specifically, the policy is parameterized as a two-layer (25-10 hidden units) neural network with tanh as the activation function. The maximal length of horizon is set to 500. We use a sample size of 10000 for policy gradient estimation, and \( M = 16, \alpha = 10 \). For the simplest task, Cartpole, all
agents are trained for 100 episodes. For the other two complex tasks, all agents are trained up to 1000 episodes. The average reward versus number of episodes are plotted in Figure 3. It is observed that our proposed SPOS-PG obtains much larger average rewards as well as smaller variance compared to SVPG, though the convergence behaviors are similar in the simplest task Cartpole.

7 Conclusion

Motivated by the need of effective and efficient Bayesian sampling techniques in modern deep learning, we propose a probability approach for particle-optimization-based sampling that unifies SG-MCMC and SVGD. Notably, for the first time, by analyzing the corresponding nonlinear SDE, we develop non-asymptotic convergence theory for the proposed SPOS framework, a missing yet important theoretical result since the development of SVGD. Within our theoretical framework, a pitfall of SVGD, which has been studied empirically (Zhuo et al., 2018), is formally analyzed. Our theory also indicates the convergence of SPOS to the true posterior distribution in the asymptotic limit of infinite particles and iterations under appropriate conditions. Our theory is of great practice value, as for the first time it provides nonasymptotic theoretical guarantees for the recently proposed particle-optimization-based algorithms such as the SVGD, whose advantages have also been extensively evaluated by experiments on Bayesian learning of DNNs and Bayesian exploration of DRL. There are a number of interesting future works. For example, one might explore more recently developed techniques such as (Cheng et al., 2018) to improve the convergence bound; one can also adopt the SPOS framework for non-convex optimization like what SG-MCMC is used for, and develop corresponding theory to study the convergence property of the algorithm to the global optimum.
REFERENCES


A  DENSITY FUNCTION OF THE MULTI-MODE DISTRIBUTION IN SECTION 5

The negative log-density function of the multi-mode distribution in Section 5 is defined as:

\[ U(\theta) \triangleq e^{\frac{3}{4} \theta^2 - \frac{3}{2} \sum_{i=1}^{10} c_i \sin(\frac{1}{4} \pi i(\theta + 4) )} \]

where \( c = (-0.47, -0.83, -0.71, -0.02, 0.24, 0.01, 0.27, -0.37, 0.87, -0.37) \) is a vector, \( c_i \) is the \( i \)-th element of \( c \).

B  GRONWALL LEMMA

The Gronwall Lemma plays an important role in parts of our proofs, which is stated in Lemma 12.

Lemma 12 (Gronwall Lemma) Let \( f \) and \( g \) functions. Specifically, let the function \( u \) with compact support in \( \Omega \subset \mathbb{R}^d \) be in a form of \( \partial_\theta u(\theta) \), that is, Lebesgue integrable on any closed bounded set \( \mathcal{F} \subset \omega \). Then \( u \) is differentiable in \( \Omega \), written as \( \phi = \partial_\theta u \), if for any infinitely-differentiable function \( v \) with compact support in \( \Omega \), we have

\[ \int_{\Omega} g(\theta) \partial_\theta u(\theta) d\theta = - \int_{\Omega} \phi(\theta) u(\theta) d\theta . \]

Proof The proof relies on further expansions on the definition of generalized derivative on specific functions. Specifically, let the function \( g \) in Definition 1 be in a form of \( g \triangleq Gf \) for two functions \( G \) and \( f \). The generalized derivative of \((Gf)(\theta,t)\), written as \( \partial_\theta(Gf)(\theta,t) \), satisfies

\[ \int \partial_\theta(Gf)(\theta,t)u(\theta)d\theta = - \int Gf(\theta,t)\partial_\theta u(\theta)d\theta \]  \hspace{1cm} (20)

for all differentiable function \( u(\cdot) \).

C  PROOF OF THEOREM 2

To prove Theorem 2, we rely on the definition of generalized derivative in Definition 1.

Definition 1 (Generalized Derivative) Let \( g \) and \( \phi \) be locally integrable functions on an open set \( \Omega \subset \mathbb{R}^d \), that is, Lebesgue integrable on any closed bounded set \( \mathcal{F} \subset \omega \). Then \( \phi \) is the generalized derivative of \( g \) with respect to \( \theta \) on \( \Omega \), written as \( \phi = \partial_\theta g \), if for any infinitely-differentiable function \( u \) with compact support in \( \Omega \), we have

\[ \int_{\Omega} g(\theta) \partial_\theta u(\theta) d\theta = - \int_{\Omega} \phi(\theta) u(\theta) d\theta . \]

Proof The proof relies on further expansions on the definition of generalized derivative on specific functions. Specifically, let the function \( g \) in Definition 1 be in a form of \( g \triangleq Gf \) for two functions \( G \) and \( f \). The generalized derivative of \((Gf)(\theta,t)\), written as \( \partial_\theta(Gf)(\theta,t) \), satisfies

\[ \int \partial_\theta(Gf)(\theta,t)u(\theta)d\theta = - \int Gf(\theta,t)\partial_\theta u(\theta)d\theta \]  \hspace{1cm} (20)

for all differentiable function \( u(\cdot) \).

In Theorem 2 we want to prove a particle representation of the following PDE:

\[ \partial_t \mu_t = F_1 = - \nabla_{\theta} \cdot ((\mu_t F(\theta_t)) + (K * \mu_t)) \triangleq - \partial_\theta(Gf)(\theta,t) \]

\[ \Rightarrow \int \partial_t \mu_t u(\theta)d\theta = - \int \partial_\theta(Gf)(\theta,t)u(\theta)d\theta , \]

where \( f(\theta,t) = \mu_t \). Consequently, we have

\[ \int \partial_t f(\theta,t)u(\theta)d\theta = - \int \partial_\theta(Gf)(\theta,t)u(\theta)d\theta \]  \hspace{1cm} (21)

By applying (20) in (21), we have

\[ \int \partial_t f(\theta,t)u(\theta)d\theta = - \int \partial_\theta(Gf)(\theta,t)u(\theta)d\theta = \int Gf(\theta,t)\partial_\theta u(\theta)d\theta . \]
Since \( f d\theta \triangleq \mu(d\theta, t) \), we have

\[
\int \partial_t \mu(d\theta, t)u(\theta) = \int G \cdot \mu(d\theta, t) \partial_\theta u(\theta)
\]

\[
\Rightarrow \frac{d}{dt} \int \mu(d\theta, t)u(\theta) = \int G \cdot \mu(d\theta, t) \partial_\theta u(\theta)
\]

\[
\Rightarrow \frac{d}{dt} E_{\mu(t)}[u(\theta)] = E_{\mu(t)}[G \cdot \partial_\theta u(\theta)].
\]

(22)

In particle approximation, we have \( \mu(t) = \frac{1}{M} \sum_{i=1}^{M} \delta_{(\theta_i(t))} \). For each particle, let \( u(\theta) = \theta \), (22) reduces to the following equation:

\[
d\theta_t^{(i)} = G(\theta_t^{(i)}) dt.
\]

This completes the proof. 

D PROOF OF THEOREM 3

Note that one challenge in our analysis compared with the analysis for linear SDEs, such as those for SG-MCMC [Vollmer et al. (2016); Chen et al. (2015)], is how to bound the gap between the original nonlinear PDE \( (\ref{eq:nonlinear_pde}) \) and the reduced nonlinear SDE \( (\ref{eq:reduced_sde}) \). Based on the techniques on analyzing granular media equations in [Malrieu (2003); Cattiaux et al. (2008); Durmus et al. (2018)], we introduce a nonlinear SDE as an element in-between \( (\ref{eq:nonlinear_pde}) \) and \( (\ref{eq:reduced_sde}) \) like:

\[
\begin{align*}
\begin{cases}
  d\theta_t = -\beta^{-1} F(\theta_t) - E_{Y \sim \nu_t} K(\theta_t - Y) F(Y) + \nabla K \ast \nu_t(\theta_t) + \sqrt{2\beta^{-1}} d\tilde{W}_t \\
  \mathcal{L}(\theta_t) = \nu_t d\theta_t
\end{cases}
\end{align*}
\]

(23)

where \( \mathcal{L}(\theta_t) \) denotes the probability law of \( \theta_t \), \( \tilde{W}_t \in \mathbb{R}^d \) is a \( d \)-dimensional Brownian motion and \( Y \) is a random variable independent of \( \theta_t \) and just used here for the sake of clarity. In order to match the SDE system \( (\ref{eq:reduced_sde}) \) of the particles \( \{\theta_t^{(i)}\}_{i=1}^{M} \), we duplicate \( M \) times each, endowing with an exact solution \( \theta_t^{(i)} \) indexed by \( i \). The distribution of each particle \( \{\theta_t^{(i)}\}_{i=1}^{M} \) is \( \nu_t \) and the corresponding \( \tilde{W}_t^{(i)} \) can be set exactly the same as the \( \tilde{W}_t^{(i)} \) in \( (\ref{eq:reduced_sde}) \):

\[
\begin{align*}
\begin{cases}
  d\tilde{\theta}_t^{(i)} = -\beta^{-1} F(\tilde{\theta}_t^{(i)}) - E_{Y_i \sim \nu_t} K(\tilde{\theta}_t^{(i)} - Y_i) F(Y_i) + \nabla K \ast \nu_t(\tilde{\theta}_t^{(i)}) + \sqrt{2\beta^{-1}} d\tilde{W}_t^{(i)} \\
  \mathcal{L}(\tilde{\theta}_t^{(i)}) = \nu_t d\theta_t
\end{cases}
\end{align*}
\]

(24)

where \( Y_i \) is a random variable independent of \( \theta_t^{(i)} \) and just used here for the convenience of the proof.

Proof [Proof of Theorem 3] Firstly we have

\[
\begin{align*}
&\frac{d}{dt} \left( \theta_t^{(i)} - \bar{\theta}_t^{(i)} \right) = -\beta^{-1} \left( F(\theta_t^{(i)}) - F(\bar{\theta}_t^{(i)}) \right) dt + \frac{1}{M} \sum_{j=1}^{M} \left[ \nabla K(\theta_t^{(i)} - \theta_t^{(j)}) - \nabla K \ast \nu_t(\bar{\theta}_t^{(i)}) \right] dt \\
&\quad - \frac{1}{M} \sum_{j=1}^{M} \left( F(\theta_t^{(i)}) W(\theta_t^{(i)} - \theta_t^{(j)}) - E_{Y_j \sim \nu_t} F(Y_j) W(\bar{\theta}_t^{(i)} - Y_j) \right) dt \\
\Rightarrow &\frac{d}{dt} \left( \sum_{i=1}^{M} \| \theta_t^{(i)} - \bar{\theta}_t^{(i)} \|^2 \right) = \frac{2}{M} \sum_{i,j=1}^{M} (A_{ij}(t) + B_{ij}(t) + C_{ij}(t) + F_{ij}(t) + G_{ij}(t) + H_{ij}(t)) dt
\end{align*}
\]

(25)

where \( A_{ij}(t) = -\beta^{-1} \left( F(\theta_t^{(i)}) - F(\theta_t^{(j)}) \right) \cdot (\theta_t^{(i)} - \theta_t^{(j)}) \)

\[
\begin{align*}
B_{ij}(t) = \left( \nabla K(\theta_t^{(i)} - \theta_t^{(j)}) - \nabla K(\bar{\theta}_t^{(i)} - \bar{\theta}_t^{(j)}) \right) \cdot (\theta_t^{(i)} - \theta_t^{(j)}) 
\end{align*}
\]
Under review as a conference paper at ICLR 2019

\[ C_{ij}(t) = \left( \nabla K(\theta_t^{(i)} - \bar{\theta}_t^{(j)}) - \nabla K * \nu_t(\bar{\theta}_t^{(i)}) \right) \cdot \left( \theta_t^{(i)} - \bar{\theta}_t^{(i)} \right) \]

\[ F_{ij}(t) = - \left( F(\theta_t^{(j)}) K(\theta_t^{(i)} - \bar{\theta}_t^{(j)}) - F(\bar{\theta}_t^{(j)}) K(\theta_t^{(i)} - \bar{\theta}_t^{(j)}) \right) \cdot \left( \theta_t^{(i)} - \bar{\theta}_t^{(i)} \right) \]

\[ G_{ij}(t) = - \left( F(\theta_t^{(j)}) K(\theta_t^{(i)} - \bar{\theta}_t^{(j)}) - F(\bar{\theta}_t^{(j)}) K(\bar{\theta}_t^{(i)} - \bar{\theta}_t^{(j)}) \right) \cdot \left( \theta_t^{(i)} - \bar{\theta}_t^{(i)} \right) \]

\[ H_{ij}(t) = - \left( F(\theta_t^{(j)}) K(\theta_t^{(i)} - \bar{\theta}_t^{(j)}) - E_{Y_j \sim \nu_t} F(Y_j) K(\theta_t^{(i)} - Y_j) \right) \cdot \left( \theta_t^{(i)} - \bar{\theta}_t^{(i)} \right) \]

For the \( A_{ij}(t) \) term, according to the \( i \) in Assumption [1] for \( F \), we have

\[
\sum_{ij} A_{ij}(t) = - \sum_{ij} \beta^{-1} \left( F(\theta_t^{(i)}) - F(\bar{\theta}_t^{(i)}) \right) \cdot \left( \theta_t^{(i)} - \bar{\theta}_t^{(i)} \right)
\]

\[
\leq -\beta^{-1} m_F M \sum_i \| \theta_t^{(i)} - \bar{\theta}_t^{(i)} \|^2
\]

For the \( B_{ij}(t) \) term, applying the concave condition for \( K \) and the oddness of \( \nabla K \) in Assumption [1] we have

\[
\sum_{ij} B_{ij}(t) = \sum_{ij} \left( \nabla K(\theta_t^{(i)} - \theta_t^{(j)}) - \nabla K(\bar{\theta}_t^{(i)} - \bar{\theta}_t^{(j)}) \right) \cdot \left( \theta_t^{(i)} - \bar{\theta}_t^{(i)} \right)
\]

\[
= \frac{1}{2} \sum_{ij} \left( \nabla K(\theta_t^{(i)} - \theta_t^{(j)}) - \nabla K(\theta_t^{(i)} - \bar{\theta}_t^{(j)}) \right) \cdot \left( \theta_t^{(i)} - \bar{\theta}_t^{(i)} - (\theta_t^{(j)} - \bar{\theta}_t^{(j)}) \right)
\]

\[
\leq -\frac{1}{2} m_K \sum_{ij} \| \theta_t^{(i)} - \bar{\theta}_t^{(i)} - (\theta_t^{(j)} - \bar{\theta}_t^{(j)}) \|^2 \leq 0
\]

For the \( C_{ij}(t) \) term, we have

\[
\mathbb{E} \sum_j C_{ij}(t) \leq \left( \mathbb{E} \| \theta_t^{(i)} - \bar{\theta}_t^{(i)} \|^2 \right)^{1/2} \left( \mathbb{E} \left\| \sum_j \left( \nabla K(\theta_t^{(i)} - \bar{\theta}_t^{(j)}) - \nabla K * \nu_t(\bar{\theta}_t^{(i)}) \right) \right\|^2 \right)^{1/2}
\]

\[
= \left( \mathbb{E} \| \theta_t^{(i)} - \bar{\theta}_t^{(i)} \|^2 \right)^{1/2} \left( \mathbb{E} \left( \sum_j \mathbb{E} \left( \nabla K(\theta_t^{(i)} - \bar{\theta}_t^{(j)}) - \nabla K * \nu_t(\bar{\theta}_t^{(i)}) \right)^2 \right) \right)^{1/2}
\]

\[
\leq H_{\nabla K} \sqrt{2M} \left( \mathbb{E} \| \theta_t^{(i)} - \bar{\theta}_t^{(i)} \|^2 \right)^{1/2}
\]

where the (1) is obtained by applying the Cauchy-Schwarz inequality and (2) by the fact that \( \mathbb{E} \left( K(\theta_t^{(i)} - \bar{\theta}_t^{(j)}) - K * \nu_t(\bar{\theta}_t^{(i)}) \right) = 0 \). We can tune the bandwidth of the RBF kernel to make \( \nabla K \leq H_{\nabla K} \), which is omitted in the Assumption due to the space limit. Hence (3) is obtain by the boundedness of \( \nabla K(\theta) \).

Similarly, since \( K \leq 1 \), we have the following result for \( H_{ij}(t) \) term,

\[
\mathbb{E} \sum_j H_{ij}(t)
\]

\[
\leq \left( \mathbb{E} \| \theta_t^{(i)} - \bar{\theta}_t^{(i)} \|^2 \right)^{1/2} \left( \mathbb{E} \left( \sum_j \left( F(\theta_t^{(j)}) K(\theta_t^{(i)} - \bar{\theta}_t^{(j)}) - E_{Y_j \sim \nu_t} F(Y_j) K(\theta_t^{(i)} - Y_j) \right) \right)^2 \right)^{1/2}
\]

\[
= \left( \mathbb{E} \| \theta_t^{(i)} - \bar{\theta}_t^{(i)} \|^2 \right)^{1/2} \left( \sum_j \mathbb{E} \left( F(\theta_t^{(j)}) K(\theta_t^{(i)} - \bar{\theta}_t^{(j)}) - E_{Y_j \sim \nu_t} F(Y_j) K(\theta_t^{(i)} - Y_j) \right)^2 \right)^{1/2}
\]
We denote \( \gamma^2 = \sqrt{2M} \left( E \left\| \theta_t^{(i)} - \bar{\theta}_t^{(i)} \right\|^2 \right)^{1/2} \)

For the \( F_{ij}(t) \) and \( G_{ij}(t) \) terms, we have:

\[
\sum_{ij} F_{ij}(t) = - \sum_{ij} \left( F(\theta_t^{(j)})K(\theta_t^{(i)} - \theta_t^{(j)}) - F(\bar{\theta}_t^{(j)})K(\bar{\theta}_t^{(i)} - \bar{\theta}_t^{(j)}) \right) \cdot \left( \theta_t^{(i)} - \bar{\theta}_t^{(i)} \right)
\]
\[
\leq \sum_{ij} L_F \left\| \theta_t^{(j)} - \bar{\theta}_t^{(j)} \right\| \left\| \theta_t^{(i)} - \bar{\theta}_t^{(i)} \right\|
\]
\[
\leq 2LFM \sum_i \left\| \theta_t^{(i)} - \bar{\theta}_t^{(i)} \right\|^2 .
\]

\[
\sum_{ij} G_{ij}(t) = \sum_{ij} \left( F(\bar{\theta}_t^{(j)})K(\theta_t^{(i)} - \theta_t^{(j)}) - F(\bar{\theta}_t^{(j)})K(\bar{\theta}_t^{(i)} - \bar{\theta}_t^{(j)}) \right) \cdot \left( \theta_t^{(i)} - \bar{\theta}_t^{(i)} \right)
\]
\[
\leq H_F L_K \sum_{ij} \left\| \theta_t^{(i)} - \bar{\theta}_t^{(i)} - (\theta_t^{(j)} - \bar{\theta}_t^{(j)}) \right\| \left\| \theta_t^{(i)} - \bar{\theta}_t^{(i)} \right\|
\]
\[
\leq 3H_F L_K M \sum_i \left\| \theta_t^{(i)} - \bar{\theta}_t^{(i)} \right\|^2.
\]

We denote \( \gamma_i(t) \equiv E \left\| \theta_t^{(i)} - \bar{\theta}_t^{(i)} \right\|^2 \). Due to the exchangeability of the particles, \( \gamma_i(t) \) are the same for all the particles. Then according to (25), we have:

\[
\gamma'(t) \leq -2\lambda_1 \gamma(t) + \frac{H_F K \sqrt{2} + H_F \sqrt{2}}{\sqrt{M}} \sqrt{\gamma(t)}.
\]

where \( \lambda_1 = \beta^{-1} m_F - 3H_F L_K - 2L_F \)

\[
\Rightarrow \left( \sqrt{\gamma(t)} - \frac{(H_F K + H_F) / \sqrt{2}}{\sqrt{M} (\beta^{-1} - 3H_F L_K - 2L_F)} \right) \leq -\lambda_1 \left( \sqrt{\gamma(t)} - \frac{(H_F K + H_F) / \sqrt{2}}{\sqrt{M} (\beta^{-1} - 3H_F L_K - 2L_F)} \right)
\]

Note that \( \theta_t^{(i)} \) and \( \bar{\theta}_t^{(i)} \) are initialized with the same initial distribution \( \mu_0 = \nu_0 \) and we can also set \( \theta_0^{(i)} \) to be independent of \( \bar{\theta}_0^{(i)} \), we can have \( \gamma(0) = 0 \). Then according to the Gronwall Lemma, we have:

\[
\sqrt{\gamma(t)} \leq \frac{(H_F K + H_F) / \sqrt{2}}{\sqrt{M} (\beta^{-1} - 3H_F L_K - 2L_F)}
\]

Hence, there exist some positive constant \( (c_1, c_2) \) such that:

\[
W_1(\rho_t, \nu_t) \leq W_2(\rho_t, \nu_t)
\]
\[
\leq \sqrt{E} \left\| \theta_t^{(i)} - \bar{\theta}_t^{(i)} \right\|^2 \leq \frac{c_1}{\sqrt{M} (\beta^{-1} - c_2)},
\]

where (2) holds due to the relationship between \( W_1 \) and \( W_2 \) metric, \( \text{Givens & Shortt (1984)} \), (2) due to the definition of \( W_2 \) and (3) due to the result from the previous proof.
E Proof of Theorem 4

Proof [Proof of Theorem 4] Firstly, what we aim at is $W_1(\nu_t, \nu_\infty) \leq c_3 \exp(-2\lambda_1 t)$ in this theorem. According to the relationship between $W_1$ and $W_2$ metric [Givens & Shortt (1984)], once we bound $W_2(\nu_t, \nu_\infty)$ as $W_2(\nu_t, \nu_\infty) \leq c_3 \exp(-2\lambda_1 t)$, we will finish our proof.

Next, look at the equation (5):
If we set the initial distribution of each particle to be $\nu_0$, which means $\rho_0 = \mathcal{L}(\theta_0^{(i)}) = \nu_0$, we will derive $M$ particles denoted as $\{\theta_{t,1}^{(i)}\}_{i=1}^M$. We denote the distribution of each $\theta_{t,1}^{(i)}$ at $t$ as $\rho_{t,1}$.
If we set the initial distribution of each particle to be $\nu_\infty$, which means $\rho_0 = \mathcal{L}(\theta_0^{(i)}) = \nu_\infty$, we will derive $M$ particles denoted as $\{\theta_{t,2}^{(i)}\}_{i=1}^M$. We denote the distribution of each $\theta_{t,2}^{(i)}$ at $t$ as $\rho_{t,2}$.

Since we need to bound $W_2(\nu_t, \nu_\infty)$, we make the following decomposition:

$$W_2(\nu_t, \nu_\infty) \leq W_2(\nu_t, \rho_{t,1}) + W_2(\rho_{t,1}, \rho_{t,2}) + W_2(\rho_{t,2}, \nu_\infty).$$

(27)

Note that $\rho_{0,1} = \nu_0$ and $\rho_{0,2} = \nu_\infty$. Then, according to (26), we have

$$W_2(\nu_t, \rho_{t,1}) \leq \frac{c_1}{\sqrt{M(\beta^{-1} - c_2)}} \text{ and } W_2(\rho_{t,2}, \nu_\infty) \leq \frac{c_1}{\sqrt{M(\beta^{-1} - c_2)}}$$

Now we need to focus on the term $W_2(\rho_{t,1}, \rho_{t,2})$. Since $W_2(\rho_{t,1}, \rho_{t,2}) \leq \mathbb{E}\left(\left\|\theta_{t,1}^{(i)} - \theta_{t,2}^{(i)}\right\|^2\right) \triangleq r(t)$, we will derive a bound for $\mathbb{E}\left(\left\|\theta_{t,1}^{(i)} - \theta_{t,2}^{(i)}\right\|^2\right)$ in the following. We have

$$d \left(\theta_{t,1}^{(i)} - \theta_{t,2}^{(i)}\right) = -\beta^{-1} \left(F(\theta_{t,1}^{(i)}) - F(\theta_{t,2}^{(i)})\right) dt$$

$$+ \frac{1}{M} \sum_{j}^{M} \left[\nabla K(\theta_{t,1}^{(i)} - \theta_{t,1}^{(j)}) - \nabla K(\theta_{t,2}^{(i)} - \theta_{t,2}^{(j)})\right] dt$$

$$- \frac{1}{M} \sum_{j}^{M} \left(F(\theta_{t,1}^{(j)})K(\theta_{t,1}^{(i)} - \theta_{t,1}^{(j)}) - F(\theta_{t,2}^{(j)})K(\theta_{t,2}^{(i)} - \theta_{t,2}^{(j)})\right) dt$$

$$\Rightarrow d \left(\sum_{i}^{M} \left\|\theta_{t,1}^{(i)} - \theta_{t,2}^{(i)}\right\|^2\right) = -\frac{2}{M} \sum_{i,j}^{M} \left(\xi_{ij}^{1}(t) + \xi_{ij}^{2}(t) + \xi_{ij}^{3}(t) + \xi_{ij}^{4}(t)\right) dt$$

where $\xi_{ij}^{1}(t) = -\beta^{-1} \left(F(\theta_{t,1}^{(i)}) - F(\theta_{t,2}^{(i)})\right) \cdot \left(\theta_{t,1}^{(i)} - \theta_{t,2}^{(i)}\right)$

$$\xi_{ij}^{2}(t) = -\beta^{-1} \left(F(\theta_{t,1}^{(j)})K(\theta_{t,1}^{(i)} - \theta_{t,1}^{(j)}) - F(\theta_{t,2}^{(j)})K(\theta_{t,2}^{(i)} - \theta_{t,2}^{(j)})\right) \cdot \left(\theta_{t,1}^{(i)} - \theta_{t,2}^{(i)}\right)$$

$$\xi_{ij}^{3}(t) = -\beta^{-1} \left(F(\theta_{t,1}^{(j)})K(\theta_{t,1}^{(i)} - \theta_{t,1}^{(j)}) - F(\theta_{t,2}^{(j)})K(\theta_{t,2}^{(i)} - \theta_{t,2}^{(j)}\right) \cdot \left(\theta_{t,1}^{(i)} - \theta_{t,2}^{(i)}\right)$$

$$\xi_{ij}^{4}(t) = -\beta^{-1} \left(F(\theta_{t,1}^{(j)})K(\theta_{t,1}^{(i)} - \theta_{t,1}^{(j)}) - F(\theta_{t,2}^{(j)})K(\theta_{t,2}^{(i)} - \theta_{t,2}^{(j)}\right) \cdot \left(\theta_{t,1}^{(i)} - \theta_{t,2}^{(i)}\right)$$

For the $\xi_{ij}^{1}(t)$ terms, according to the $i$ in Assumption 4 for $F$, we have

$$\sum_{ij} \xi_{ij}^{1}(t) = -\sum_{ij} \beta^{-1} \left(F(\theta_{t,1}^{(i)}) - F(\theta_{t,2}^{(i)})\right) \cdot \left(\theta_{t,1}^{(i)} - \theta_{t,2}^{(i)}\right)$$

$$\leq -\beta^{-1} m_F M \sum_{i} \left\|\theta_{t,1}^{(i)} - \theta_{t,2}^{(i)}\right\|^2.$$
For the $\xi^2_{ij}(t)$ term, applying the concave condition for $K$ and the oddness of $\nabla K$ in Assumption 1, we have

\[
\sum_{ij} \xi^2_{ij}(t) = \frac{1}{2} \sum_{ij} \left( \nabla K(\theta_{ij}^{(i)} - \theta_{ij}^{(j)}) - \nabla K(\theta_{ij}^{(i)} - \theta_{ij}^{(j)}) \right) \cdot \left( \theta_{ij}^{(i)} - \theta_{ij}^{(j)} \right)
= \frac{1}{2} \sum_{ij} \left( \nabla K(\theta_{ij}^{(i)} - \theta_{ij}^{(j)}) - \nabla K(\theta_{ij}^{(i)} - \theta_{ij}^{(j)}) \right) \cdot \left( \theta_{ij}^{(i)} - \theta_{ij}^{(j)} - (\theta_{ij}^{(j)} - \theta_{ij}^{(j)}) \right)
\leq -\frac{1}{2} M \sum_{ij} \| \theta_{ij}^{(i)} - \theta_{ij}^{(j)} \| \leq 0.
\]

For the $\xi^3_{ij}(t)$ terms, after applying the $L_F$-Lipschitz property for $F$ and $K \leq 1$, we have

\[
\sum_{ij} \xi^3_{ij}(t) = -\sum_{ij} \left( F(\theta_{ij}^{(j)}) K(\theta_{ij}^{(i)} - \theta_{ij}^{(j)}) - F(\theta_{ij}^{(j)}) K(\theta_{ij}^{(i)} - \theta_{ij}^{(j)}) \right) \cdot \left( \theta_{ij}^{(i)} - \theta_{ij}^{(j)} \right)
\leq \sum_{ij} L_F \left\| \theta_{ij}^{(i)} - \theta_{ij}^{(j)} \right\| \leq 2L_F M \sum_{ij} \left\| \theta_{ij}^{(i)} - \theta_{ij}^{(j)} \right\|^2.
\]

For the $\xi^4_{ij}(t)$ terms:

\[
\sum_{ij} \xi^4_{ij}(t) = -\sum_{ij} \left( F(\theta_{ij}^{(j)}) K(\theta_{ij}^{(i)} - \theta_{ij}^{(j)}) - F(\theta_{ij}^{(j)}) K(\theta_{ij}^{(i)} - \theta_{ij}^{(j)}) \right) \cdot \left( \theta_{ij}^{(i)} - \theta_{ij}^{(j)} \right)
\leq \sum_{ij} H_F L_K \left\| \theta_{ij}^{(i)} - \theta_{ij}^{(j)} \right\| \leq 3H_F L_K M \sum_{ij} \left\| \theta_{ij}^{(i)} - \theta_{ij}^{(j)} \right\|^2.
\]

Now we have

\[
r'(t) \leq -2(\beta^{-1} m_F - 3H_F L_K - 2L_F) r(t).
\]

According to the Gronwall lemma,

\[
r(t) \leq r(0) e^{-2\lambda_1 t},
\]

where $\lambda_1 = \beta^{-1} m_F - 3H_F L_K - 2L_F$.

Consequently, there exists some positive constant $c_3$ such that

\[
W_2(\rho_{t,1}, \rho_{t,2}) \leq c_3 e^{-2\lambda_1 t}.
\]

Then we have

\[
W_2(\nu_t, \nu_\infty) \leq c_3 e^{-2\lambda_1 t} + \frac{c_1}{\sqrt{M}(\beta^{-1} - c_2)} + \frac{c_1}{\sqrt{M}(\beta^{-1} - c_2)}.
\]

However, it worth noting that $\nu_t$ is the solution of (6) which has nothing to do with the number of particles, $M$. Then let $M \to \infty$, we can derive that $W_2(\nu_t, \nu_\infty) \leq c_3 e^{-2\lambda_1 t}$. Now we finish our proof.
F Proof of Theorem 5

To bound the $\mathcal{W}_1(\mu, \rho_{\sum_{k=1}^{T-1} h_k})$ term, note the original SDE driving the particles $\{\theta^{(i)}_t\}$ in (8) is a nonlinear SDE, which is hard to deal with. Fortunately, (8) can be turned into a linear SDE by concatenating the particles at each time into a single vector representation, i.e., by defining the new parameter at time $t$ as $\Theta_t = [\theta^{(1)}_t, \ldots, \theta^{(M)}_t] \in \mathbb{R}^{Md}$. Consequently, $\Theta_t$ is driven by the following linear SDE:

$$d\Theta_t = -F_\Theta(\Theta_t)dt + \sqrt{2\beta^{-1}}d\mathcal{W}_t(\Theta),$$

(28)

where $F_\Theta(\Theta_t) \triangleq [\beta^{-1}1F(\theta^{(1)}_t) - \frac{1}{M} \sum_{j=1}^{M} \nabla K(\theta^{(1)}_t - \theta^{(j)}_t) + \frac{1}{M} \sum_{j=1}^{M} K(\theta^{(1)}_t - \theta^{(j)}_t)]$ is a vector function $\mathbb{R}^{Md} \rightarrow \mathbb{R}^{Md}$, and $\mathcal{W}_t(\Theta)$ is Brownian motion of dimension $Md$.

Now we define the $F(q_\phi)(\Theta_t) \triangleq [\beta^{-1}1F(\theta^{(1)}_t) - \frac{1}{M} \sum_{j=1}^{M} \nabla K(\theta^{(1)}_t - \theta^{(j)}_t) + \frac{1}{M} \sum_{j=1}^{M} K(\theta^{(1)}_t - \theta^{(j)}_t)]F(\theta^{(j)}_t), \ldots, \beta^{-1}1F(\theta^{(M)}_t) - \frac{1}{M} \sum_{j=1}^{M} \nabla K(\theta^{(M)}_t - \theta^{(j)}_t) + \frac{1}{M} \sum_{j=1}^{M} K(\theta^{(M)}_t - \theta^{(j)}_t)]$. We can verify that $F_\Theta(\Theta_t) = \sum_{q=1}^{N} F(q_\phi)(\Theta_t)$.

Then we define $\theta_k \triangleq [\theta^{(1)}_k, \ldots, \theta^{(M)}_k]$ and $G_{\Theta_k} \triangleq \frac{N}{\sqrt{K}} \sum_{q \in \mathcal{I}_k} F(q_\phi)(\Theta_k)$. We can verify that the following result holds:

$$\Theta_{k+1} = \Theta_k - \beta^{-1}G_{\Theta_k}h_k + \sqrt{2\beta^{-1}}h_k \xi_k,$$

(29)

where $\xi_k \sim \mathcal{N}(0, I_{Md})$. Now we reach the conclusion that $\Theta_k$ of (29) is accutally the numerical solution of the SDE (28) via stochastic gradients.

We denote the distribution of $\Theta_k$ as $\mu_k^{\Theta}$ and the distribution of $\Theta_t$ as $\mu_t^{\Theta}$. Before proceeding to our theoretical results, we need to present the following Lemmas which is very important in our proof.

Lemma 13 $\mathcal{W}_1(\mu_k, \rho_t) \leq \frac{1}{\sqrt{M}} \mathcal{W}_1(\mu_k^{\Theta}, \rho_t^{\Theta})$

Proof [Proof of Lemma 13] Let us recall the definition of $W_1$ metric and its Kantorovich-Rubinstein duality [Villani, 2008], i.e., $\mathcal{W}_1(\mu, \nu) \triangleq \sup_{g \in \Pi(\mu, \nu)} |\mathbb{E}_{\theta \sim \mu}[g(\theta)] - \mathbb{E}_{\theta \sim \nu}[g(\theta)]|$. We can prove the fact that if $g(\theta): \mathbb{R}^d \rightarrow \mathbb{R}$ is a $L_g$-Lipschitz function in $\mathbb{R}^d$, the $g_\Theta(\Theta)$, defined as $g_\Theta(\Theta) = \frac{1}{\sqrt{M}} \sum_{i=1}^{M} g(\theta^{(i)}_t)$, is a $L_g$-Lipschitz function in $\mathbb{R}^{Md}$, where $\Theta \triangleq [\theta^{(1)}, \ldots, \theta^{(M)}]$

$$||g_\Theta(\Theta_1) - g_\Theta(\Theta_2)|| \leq \frac{1}{\sqrt{M}} \sum_{i=1}^{M} ||g(\theta^{(i)}_1) - g(\theta^{(i)}_2)|| \leq \frac{L_g}{\sqrt{M}} \sum_{i=1}^{M} ||\theta^{(i)}_1 - \theta^{(i)}_2||$$

$$\leq \frac{L_g}{\sqrt{M}} \sqrt{\sum_{i=1}^{M} ||\theta^{(i)}_1 - \theta^{(i)}_2||^2} = L_g ||\Theta_1 - \Theta_2||$$

Then we have:

$$\frac{1}{\sqrt{M}} \sum_{i=1}^{M} |\mathbb{E}_{\theta_k^{(i)} \sim \mu_k^{\Theta}}[g(\theta^{(i)}_k)] - \mathbb{E}_{\theta_i^{(i)} \sim \rho_k}[g(\theta^{(i)}_k)]| \leq \frac{1}{\sqrt{M}} \sum_{i=1}^{M} (|\mathbb{E}_{\theta_k^{(i)} \sim \mu_k^{\Theta}}[g(\theta^{(i)}_k)] - \mathbb{E}_{\theta_i^{(i)} \sim \rho_k}[g(\theta^{(i)}_k)]|)

= \frac{1}{\sqrt{M}} |\mathbb{E}_{\Theta_k \sim \mu_k^{\Theta}}[g_\Theta(\Theta_k)] - \mathbb{E}_{\Theta_i \sim \rho_k}[g_\Theta(\Theta_i)]|$$

The (1) holds since $\mathbb{E}_{\theta_k^{(i)} \sim \mu_k^{\Theta}}[g(\theta^{(i)}_k)] = \cdots = \mathbb{E}_{\theta^{(M)} \sim \mu_k^{\Theta}}[g(\theta^{(M)}_k)]$ for all the particles $\theta^{(i)}_k$ and $\mathbb{E}_{\theta_k^{(i)} \sim \rho_k}[g(\theta^{(i)}_k)] = \cdots = \mathbb{E}_{\theta^{(M)} \sim \rho_k}[g(\theta^{(M)}_k)]$ for all the particles $\theta^{(i)}_k$. Then according to the
definition of \( \mathcal{W}_1 \) metric, we derive that

\[
\mathcal{W}_1(\mu_k, \rho_t) = \sup_{\|g\|_{lip} \leq 1} \frac{1}{M} \sum_{i=1}^{M} \left| E_{\theta_k \sim \mu_k}[g(\theta_k)] - E_{\theta_t \sim \rho_t}[g(\theta_t)] \right|
\]

\[
= \frac{1}{\sqrt{M}} \sup_{\|g\|_{lip} \leq 1} \left| E_{\theta_k \sim \mu_k}[g(\Theta_k)] - E_{\theta_t \sim \rho_t}[g(\Theta_t)] \right| = \frac{1}{\sqrt{M}} \sup_{\|g\|_{lip} \leq 1} \left| E_{\theta_k \sim \mu_k}[g(\Theta_k)] - E_{\theta_t \sim \rho_t}[g(\Theta_t)] \right|
\]

\[
\leq \frac{1}{\sqrt{M}} \mathcal{W}_1(\mu_k, \rho_t)
\]

**Lemma 14** Assuming \( F(0) = 0 \). If \( F \) in \( \mathcal{Q} \) is Lipschitz with constant \( L_F \), and satisfies the dissipative property that \( \langle F(\theta), \theta \rangle \geq m_F \|\theta\|^2 - b \). Then \( F_\phi \) in \( \mathcal{Q} \) is Lipschitz-continuous with constant \( \sqrt{2} \beta^{-1} L_F + l' \) and satisfies \( \langle F_\phi(\Theta), \Theta \rangle \geq (\beta^{-1} m_F - m') \|\Theta\|^2 - \beta^{-1} M b \), where \( l' \) and \( m' \) are some positive constants.

**Proof** [Proof of Lemma 14]

\[
\|F_\phi(\Theta_1) - F_\phi(\Theta_2)\| = \sqrt{\sum_{i=1}^{M} (\|\omega^1_i\| + \|\omega^2_i\| + \|\omega^3_i\|)^2} \leq \sqrt{\sum_{i=1}^{M} (\|\omega^1_i\| + \|\omega^2_i\| + \|\omega^3_i\|)^2}
\]

where

\[
\|\omega^1_i\| = \|\beta^{-1} F(\theta^{(i)}_1) - \beta^{-1} F(\theta^{(i)}_2)\|
\]

\[
\leq \beta^{-1} L_F \|\theta^{(i)}_1 - \theta^{(i)}_2\|
\]

\[
\|\omega^2_i\| = \|\frac{1}{M} \left( \sum_{j} K(\theta^{(i)}_1 - \theta^{(j)}_1) F(\theta^{(j)}_1) - \sum_{j} K(\theta^{(i)}_2 - \theta^{(j)}_2) F(\theta^{(j)}_2) \right)\|
\]

\[
\leq \|\frac{1}{M} \left( \sum_{j} K(\theta^{(i)}_1 - \theta^{(j)}_1) F(\theta^{(j)}_1) - \sum_{j} K(\theta^{(i)}_2 - \theta^{(j)}_2) F(\theta^{(j)}_2) \right)\|
\]

\[
+ \|\frac{1}{M} \left( \sum_{j} K(\theta^{(i)}_2 - \theta^{(j)}_2) F(\theta^{(j)}_1) - \sum_{j} K(\theta^{(i)}_2 - \theta^{(j)}_2) F(\theta^{(j)}_2) \right)\|
\]

\[
\leq (2L_K H_F + L_F) \|\theta^{(i)}_1 - \theta^{(i)}_2\|
\]

\[
\|\omega^3_i\| = \|\frac{1}{M} \left( \sum_{j} \nabla K(\theta^{(i)}_1 - \theta^{(j)}_1) - \sum_{j} \nabla K(\theta^{(i)}_2 - \theta^{(j)}_2) \right)\|
\]

\[
\leq L_{\nabla K} \|\theta^{(i)}_1 - \theta^{(i)}_2\|
\]

It is easy to verify that there exits some positive constant \( l' \) such that

\[
\|F_\phi(\Theta_1) - F_\phi(\Theta_2)\| = \sqrt{\sum_{i=1}^{M} (\|\omega^1_i\| + \|\omega^2_i\| + \|\omega^3_i\|)^2}
\]

*This is reasonable, as \( F \) in our setting corresponds to an unnormalized log-posterior, which can be shifted so that \( F(0) = 0 \) for a specific problem.*
Assumption 1 recovers the dissipative assumption as

\[ \langle F_\theta(\Theta), \Theta \rangle \]

As a result, we can derive the following result:

Next, we have

\[ (F_\theta(\Theta), \Theta) \]

Notice:

\[ \sum_{i}^{M} \beta^{-1} F(\theta^{(i)}) \theta^{(i)} \geq \beta^{-1} m_F \sum_{i}^{M} \| \theta^{(i)} \|^2 - \beta^{-1} \beta^{-1} M_b \]

Since we have assumed \( F(0) = 0 \), we have:

\[ \sum_{i}^{M} \frac{1}{M} \sum_{j}^{M} K(\theta^{(i)} - \theta^{(j)}) F(\theta^{(j)}) \theta^{(i)} \geq - \frac{1}{M} \sum_{i}^{M} \sum_{j}^{M} L_F \| \theta^{(i)} \| \| \theta^{(j)} \| \]

\[ \geq -2L_F \sum_{i=1}^{M} \| \theta^{(i)} \|^2 = -2L_F \| \Theta \|^2 \]

Since \( \nabla K \) is an odd function, we have:

\[ \sum_{i}^{M} \frac{1}{M} \sum_{j}^{M} \nabla K(\theta^{(i)} - \theta^{(j)}) \theta^{(i)} \geq - \frac{1}{M} \sum_{i}^{M} \sum_{j}^{M} L_{\nabla K} \| \theta^{(i)} - \theta^{(j)} \| \| \theta^{(i)} \| \]

\[ \geq -3L_{\nabla K} \sum_{i}^{M} \| \theta^{(i)} \|^2 = -3L_{\nabla K} \| \Theta \|^2 \]

As a result, we can derive the following result:

\[ \langle F_\theta(\Theta), \Theta \rangle \geq (\beta^{-1} m - 2L_F - 3L_{\nabla K}) \| \Theta \|^2 - \beta^{-1} M_b \]

Now it is ready to prove Theorem 5. It is worth noting that after assuming \( F(0) = 0 \), the first bullet in Assumption 1 recovers the dissipative assumption as \( \langle F(\theta), \Theta \rangle \geq m_F \| \Theta \|^2 \).

Proof: Next we use Lemma C.5 in [Xu et al. (2018)] to verify that \( F_\theta \) satisfies the assumptions in [Raginsky et al. (2017)] by setting \( \delta = \frac{2a}{B} \), a positive constant and \( B \) the size of the random set \( \mathcal{I} \). Let \( \mu^\Theta \) := \( \mathcal{L}(\Theta) \) and \( \rho^\Theta := \mathcal{L}(\Theta) \). Now we can borrow the result of Lemma 3.6 in [Raginsky et al. (2017)]. The relative entropy \( D_{KL}(\mu^\Theta \| \rho^\Theta) \) satisfies:

\[ D_{KL}(\mu^\Theta \| \rho^\Theta) \leq (A_0 \beta a / B + A_1 h) kh \]
Applying the Lemma 13, we have

\[
\beta \quad I \quad \text{where} 
\]

where

\[
A_0 = \left(2 \left( \frac{L_F}{\beta} + l' \right)^2 \left( a_2 + 2 \left( 1 \vee \frac{1}{\beta - 1} m_F - m' \right) (2a_1^2 + \frac{Md}{\beta}) \right) + a_1^2 \right)
\]

\[
A_1 = 6 \left( \frac{L_F}{\beta} + l' \right)^2 (\beta A_0 + Md)
\]

and \(a_1, a_2\) are some positive constants. When the \(\beta\) is small enough, there exist some positive constants \(a_3, a_4\) such that

\[
A_0 \leq a_3 \frac{Md}{\beta^3}, A_1 \leq a_4 \frac{Md}{\beta^4}
\]

Similar to the proof of Lemma \[4\] it is easy to verify that there exists some positive constant \(a_5\) such that \(\langle F_{\Theta}(\Theta_1) - F_{\Theta}(\Theta_2), \Theta_1 - \Theta_2 \rangle \geq (\beta^{-1} m_F - a_5) \| \Theta_1 - \Theta_2 \|^2\). Notice, when \(\beta\) is small enough, (28) satisfies the conditions of Proposition 4.2 in Cattiaux et al. (2008). Hence, there exits some positive constant \(C\) such that \(W_1(\mu^\Theta_k, \rho_{kh}) \leq C \sqrt{D_{KL}(\mu^\Theta_k \| \rho_{kh})}\).

According to Corollary 4 and Lemma 8 in Bolley & Villani (2005), we can derive an explicit expression for \(C\):

\[
C \leq a_6 \beta^{-1} Md
\]

when \(\beta\) is small enough and \(a_6\) is some positive constant.

Applying the Lemma \[13\] we have

\[
W_1(\mu_k, \rho_{kh}) \leq \frac{1}{\sqrt{M}} W_1(\mu^\Theta_k, \rho_{kh}) \leq a_6 M d^2 \beta^{-3} (a_3 a' \beta^{-1} B^{-1} + a_4 h)^{\frac{1}{2}} k^\frac{1}{2} h^{\frac{1}{2}}
\]

Let \(k = T\) and we can finish the proof. \(\blacksquare\)

G Proof of Theorem 6

**Proof** Our proof is based on the proof of Lemma 3.6 in Raginsky et al. (2017) with some modifications. Firstly, adopt the same notations in the Section \[6\] and we get the following update:

\[
\Theta_{k+1} = \Theta_k - \beta^{-1} G_{\Theta}^{\Theta_k, h_k} \Xi_k + \sqrt{2 \beta^{-1} h_k} \Xi_k,
\]

where \(\Xi_k \sim \mathcal{N}(0, \text{Id} \times Md)\) and \(h_k = \frac{h}{k + 1}\). We assume \(\mathbb{E}(G_{\Theta}^{\Theta}) = F_{\Theta}(\Theta_k), \forall \Theta \in \mathbb{R}^{Md}\), which is a general assumption due to the way that we choose the minibatch \(\mathcal{T}_k\). We need to define \(q(t)\), which will be used in the following proof:

\[
q(t) = \{k \in \mathbb{R} \mid \sum_{i=0}^{k-1} h_i \leq t < \sum_{i=0}^{k} h_i \}
\]

Furthermore, we define \(\sum_{i=0}^{k-1} h_i \equiv 0\) and \(\sum_{i=0}^{0} h_i \equiv h_0\) here just used for the convenience of statement in the following.

Now we focus on the following continuous-time interpolation of \(\Theta_k\):

\[
\Theta(t) = \Theta_0 - \int_0^t \frac{G_{\Theta}^{\Theta(s)}}{\mathcal{T}(s)} \left( \Theta \left( \sum_{i=0}^{q(s)-1} h_i \right) \right) ds + \sqrt{\frac{2}{\beta}} \int_0^t \mathcal{W}_s^{(Md)}
\]

where \(\mathcal{T}(s) \equiv \mathcal{T}_k\) for \(t \in [\sum_{i=0}^{k-1} h_i, \sum_{i=0}^k h_i)\). \(G_{\Theta}^{\Theta(s)}(\Theta) \equiv \frac{N}{N(s)} \sum_{q \in \mathcal{T}(s)} \sum_{r \in \mathcal{T}(s)} F_{\Theta}^{(Md)}(\Theta)\) and \(B(s)\) is the size of the minibatch \(\mathcal{T}(s)\). And for each \(k, \Theta \left( \sum_{i=0}^{k-1} h_i \right)\) and \(\Theta_k\) have the same probability law \(\rho_{kh}\). Since \(\Theta(t)\) is not a Markov process, we define the following Itô process which has the same one-time marginals as \(\Theta(t)\):

\[
\Lambda(t) = \Theta_0 - \int_0^t G \left( \Lambda(s) \right) ds + \sqrt{\frac{2}{\beta}} \int_0^t \mathcal{W}_s^{(Md)}
\]
Thus, we can use Lemma 3.1 and 3.2 in Raginsky et al. (2017), and Lemma C.5 in Xu et al. (2018) to derive a similar result for the relative entropy of \( P^t_{\Lambda} \) and \( P^t_{\Theta} \):

\[
\text{Lemma 3.6 in Raginsky et al. (2017), we can derive a similar result for the relative entropy of } P^t_{\Lambda} \text{ and } P^t_{\Theta}:
\]

\[
D_{KL}(P^t_{\Lambda} \| P^t_{\Theta}) = -\int dP^t_{\Lambda} \log \frac{dP^t_{\Lambda}}{dP^t_{\Theta}}
\]

\[
= \frac{\beta}{4} \int_0^t \mathbb{E}[F_{\Theta} (\Lambda(s)) - G_\Lambda (\Lambda(s))] \|2^2 ds
\]

\[
= \frac{\beta}{4} \int_0^t \mathbb{E}[F_{\Theta} (\Theta(s)) - G_\Theta (\Theta(s))] \|2^2 ds
\]

The last line follows that \( \mathcal{L} (\Theta(s)) = \mathcal{L} (\Lambda(s)) \), \( \forall s \).

In the following proof, we will let \( t = \sum_{i=0}^{k-1} h_i \) for some \( k \in \mathbb{R} \). Now we can use the martingale property of Itô integral to derive:

\[
D_{KL}(P^t_{\Lambda} \| P^t_{\Theta}) = -\int dP^t_{\Lambda} \log \frac{dP^t_{\Lambda}}{dP^t_{\Theta}}
\]

\[
= \frac{\beta}{4} \int_0^t \mathbb{E}[F_{\Theta} (\Theta(s)) - G_\Theta (\Theta(s))] \|2^2 ds
\]

\[
= \frac{\beta}{4} \int_0^t \mathbb{E}[F_{\Theta} (\Theta(s)) - G_\Theta (\Theta(s))] \|2^2 ds
\]

\[
\leq \frac{\beta}{2} \sum_{j=0}^{k-1} \sum_{i=0}^{j-1} h_i \mathbb{E}[F_{\Theta} (\Theta(s)) - F_{\Theta} (\sum_{i=0}^{q(s)-1} h_i)] \|2^2 ds
\]

\[
+ \frac{\beta}{2} \sum_{j=0}^{k-1} \sum_{i=0}^{j-1} h_i \mathbb{E}[F_{\Theta} (\sum_{i=0}^{q(s)-1} h_i)) - G_{\Theta} (\sum_{i=0}^{q(s)-1} h_i)] \|2^2 ds
\]

\[
\leq \frac{\beta L^2_{\Theta}}{2} \sum_{j=0}^{k-1} \sum_{i=0}^{j-1} h_i \mathbb{E}[\Theta(s) - \Theta(\sum_{i=0}^{q(s)-1} h_i)] \|2^2 ds
\]

\[
+ \frac{\beta}{2} \sum_{j=0}^{k-1} \sum_{i=0}^{j-1} h_i \mathbb{E}[F_{\Theta} (\sum_{i=0}^{q(s)-1} h_i)) - G_{\Theta} (\sum_{i=0}^{q(s)-1} h_i)] \|2^2 ds
\]

For the first part (31), we consider some \( s \in \bigcup_{i=0}^{j-1} [h_i, h_i + h_{i+1}] \). The following equation holds:

\[
\Theta(s) - \Theta(\sum_{i=0}^{j-1} h_i) = -(s - \sum_{i=0}^{j-1} h_i) G_{\Theta}^s + \frac{1}{\beta} \mathbb{W}_s^{(Md)} - \mathbb{W}_s^{(Md)}
\]

\[
= -(s - \sum_{i=0}^{j-1} h_i) (F_{\Theta} (\Theta(s)) - G_{\Theta}^s)
\]

\[
+ \frac{1}{\beta} \mathbb{W}_s^{(Md)} - \mathbb{W}_s^{(Md)}
\]

Thus, we can use Lemma 3.1 and 3.2 in Raginsky et al. (2017), and Lemma C.5 in Xu et al. (2018) to get the following result:

\[
\mathbb{E}[(\Theta(s) - \Theta(\sum_{i=0}^{j-1} h_i))^2] \leq 3 \frac{h_0^2}{(j+1)^2} \mathbb{E}[G_{\Theta}^s]^2 + 3 \frac{h_0^2}{(j+1)^2} \mathbb{E}[F_{\Theta} (\Theta(s)) - G_{\Theta}^s]^2 + \frac{6h_0Md}{\beta(j+1)}
\]

\[
\leq 12 \frac{h_0^2}{(j+1)^2} \max_{0 \leq j \leq k-1} (L_{\Theta}^2 \mathbb{E}[\Theta]_2^2 + 2b_1) + \frac{6h_0Md}{\beta(j+1)}
\]
where \( b_1 \) is some positive constant.

Consequently, the first part above (31) can be bounded as:

\[
\frac{\beta L_F^2}{2} \sum_{j=0}^{k-1} \left[ \sum_{i=0}^{q(s)-1} h_i \right] \| \Theta(s) - \Theta( \sum_{i=0}^{q(s)-1} h_i ) \|_2^2 ds
\]

\[
\leq \frac{\beta L_F^2}{2} \sum_{j=0}^{k-1} \left[ 12 \frac{h_0^3}{(j+1)^3} \max_{0 \leq j \leq k-1} \left( L_F^2 \| \Theta_j \| ^2 + b_1 \right) + \frac{6h_0^2Md}{\beta(j+1)^2} \right]
\]

\[
\leq \pi^2 \frac{\beta L_F^2}{2} h_0^3 \max_{0 \leq j \leq k-1} \left( L_F^2 \| \Theta_j \| ^2 + b_1 \right) + \frac{\pi^2 L_F^2 h_0^2 Md}{2},
\]

where the last line follows from the fact that

\[
\sum_{j=0}^{k-1} \frac{1}{(j+1)^3} \leq \sum_{j=0}^{k-1} \frac{1}{(j+1)^2} \leq \sum_{j=0}^{\infty} \frac{1}{(j+1)^2} = \frac{\pi^2}{6}.
\]

According to Lemma C.5 in Xu et al. (2018), the second part (32) can be bounded as follows:

\[
\frac{\beta}{2} \sum_{j=0}^{k-1} \left[ \sum_{i=0}^{q(s)-1} h_i \right] \mathbb{E} \left[ \| F_{\Theta}(\Theta( \sum_{i=0}^{q(s)-1} h_i )) - G_{2(s)}^\Theta \right] \|_2^2 ds
\]

\[
= \beta h_0 \mathbb{E} \left[ \| F_{\Theta}(\Theta) - G_{2(s)}^\Theta \|_2^2 \right]
\]

\[
\leq \beta h_0 \max_{0 \leq j \leq k-1} \left( L_F^2 \| \Theta_j \| ^2 + b_1 \right) \left( 4 \frac{B_0}{B_0 + \sum_{j=1}^{k-1} (j+1)(B_0 + \log \frac{\Theta}{\Theta}) (j+1)) \right)
\]

\[
\leq \beta h_0 \max_{0 \leq j \leq k-1} \left( L_F^2 \| \Theta_j \| ^2 + b_1 \right) \left( 4 \frac{B_0}{B_0 + \sum_{j=1}^{k-1} (j+1)(\log \frac{\Theta}{\Theta}) (j+1)) \right)
\]

\[
\leq (b_2 + \frac{4}{B_0}) \beta h_0 \max_{0 \leq j \leq k-1} \left( L_F^2 \| \Theta_j \| ^2 + b_1 \right),
\]

where the last line follows from the fact that when \( r > 1 \),

\[
\sum_{j=1}^{k-1} \frac{4}{(j+1) \log^r (j+1)} \leq \sum_{j=1}^{\infty} \frac{4}{(j+1) \log^r (j+1)} \leq \frac{4 \log^{1-r} 2}{r-1}.
\]

Denote \( \mu_k^\Theta := \mathcal{L}(\Theta_k) \) and \( \rho_k^\Theta := \mathcal{L}(\Theta_k) \). Due to the data-processing inequality for the relative entropy, we have

\[
D_{KL}(\mu_k^\Theta \| \rho_k^\Theta) \leq D_{KL}(\mathbb{P}_{\Theta}^{\sum_{i=0}^{k-1} h_i} \| \mathbb{P}_{\Theta}^{\sum_{i=0}^{k-1} h_i})
\]

\[
\leq \pi^2 \beta L_F^2 h_0^3 \max_{0 \leq j \leq k-1} \left( L_F^2 \| \Theta_j \| ^2 + b_1 \right) + \frac{\pi^2 L_F^2 h_0^2 Md}{2}
\]

\[
+ (b_2 + \frac{4}{B_0}) \beta h_0 \max_{0 \leq j \leq k-1} \left( L_F^2 \| \Theta_j \| ^2 + b_1 \right)
\]

\[
\leq (\pi^2 \beta L_F^2 h_0^3 + b_2 \beta h_0 + \frac{4}{B_0} \beta h_0) \max_{0 \leq j \leq k-1} \left( L_F^2 \| \Theta_j \| ^2 + b_1 \right) + \frac{\pi^2 L_F^2 h_0^2 Md}{2}.
\]

Lemma 3.2 in Raginsky et al. (2017) has provided a uniform bound to \( \max_{0 \leq j \leq k-1} \left( L_F^2 \| \Theta_j \| ^2 + b_1 \right) \). Hence we can tell that \( D_{KL}(\mathbb{P}_{\Theta}^{\sum_{i=0}^{k-1} h_i} \| \mathbb{P}_{\Theta}^{\sum_{i=0}^{k-1} h_i}) \) would not increase w.r.t. \( k \). This is a nice
property that the fixed-step-size SPOS does not endow. Since $L_{F\theta} = \sqrt{2}\beta^{-1}L_F + \nu$, it is easy to verify that when $\beta$ is small enough, there exists some positive constants $b_3, b_4, b_5$ and $b_6$ such that:

$$D_{KL}(\mu_k^\theta \| \rho_{\sum_{i=1}^{k-1} h_i}) \leq (\frac{\pi^2}{\beta}L_{F\theta}^2h_0^3 + b_2\beta h_0 + \frac{4\beta h_0}{B_0}) \max_{0 \leq s \leq K-1} (L_{F\theta}^2 \mathbb{E}\|\Theta_j\|^2) + b_1) + \frac{\pi^2 L_{F\theta}^2 h_0^2Md}{2}.$$ 

$$\leq (b_3 h_0^3 + \frac{b_4 \beta^3 h_0}{B_0} + b_5 h_0^2 \beta^2) \frac{Md}{\beta^4}.$$ 

Similar to the proof of Theorem 5, we can bound the $W_1(\mu_k^\theta \| \rho_{\sum_{i=1}^{k-1} h_i})$ term with Corollary 4, Lemma 8 in Bolley & Villani (2005) and Proposition 4.2 in Cattiaux et al. (2008). Specifically, when $\beta$ is small enough, there exist some positive constant $a_0$ such that:

$$W_1(\mu_k^\theta \| \rho_{\sum_{i=1}^{k-1} h_i}) \leq a_0(Md/\beta) \sqrt{D_{KL}(\mu_k^\theta \| \rho_{\sum_{i=1}^{k-1} h_i})} \leq a_0 \beta^{-3} Md^\frac{3}{2} (b_3 h_0^3 + \frac{b_4 \beta^3 h_0}{B_0} + b_5 h_0^2 \beta^2)^\frac{1}{2}.$$ 

According to Lemma 13 we have:

$$W_1(\mu_k, \rho_{kh}) \leq \frac{1}{\sqrt{M}} W_1(\mu_k^\theta \| \rho_{\sum_{i=1}^{k-1} h_i}) = a_0 \beta^{-3} Md^\frac{3}{2} (b_3 h_0^3 + \frac{b_4 \beta^3 h_0}{B_0} + b_5 h_0^2 \beta^2)^\frac{1}{2}.$$ 

Let $k = T$ and we can finish the proof.

**H DETAILED STATEMENTS OF NONASYMPTOTIC CONVERGENCE UNDER THE CONVEX CASE**

We give detailed statements of non-asymptotic convergence of SPOS under the convex case, which can be proved by directly combining results from Theorem 15 and Theorem 16 thus they are omitted for simplicity. We consider a fixed-stepsize and a decreasing-stepsize cases in Theorem 15 and Theorem 16 respectively.

**Theorem 15 (Fixed Stepsize)** Under Assumption 7 and $F(0) = 0$, if we set $h_k = h_0$ and $B_k = B_0$, we can bound the $W_1(\mu_T, \nu_\infty)$ as

$$W_1(\mu_T, \nu_\infty) \leq \frac{c_1}{\sqrt{M}(\beta^{-1} - c_2)} + c_3 \exp\left\{-2 \left(\beta^{-1} m_F - 3H_F L_K - 2L_F \right) Th\right\} \ + c_3 \beta^{-3} \mathbb{E} \sum_{k=0}^{T-1} \left\{ c_4 \beta^2 B^{-1} + c_5 h_k \right\} \frac{3}{2} T^\frac{1}{2} h^\frac{1}{2}. $$

where $(c_1, c_2, c_3, c_4, c_5, c_6, \beta)$ are some positive constants such that $\frac{1}{\beta} > c_2$ and $\frac{m_F}{\beta} > 3H_F L_K - 2L_F$.

**Theorem 16 (Decreasing Stepsize)** Under Assumption 7 and $F(0) = 0$, if we set $h_k = h_0/(k+1)$ and $B_k = B_0 + \log(k+1)^{100/99}$, we can bound the $W_1(\mu_T, \nu_\infty)$ as

$$W_1(\mu_T, \nu_\infty) \leq \frac{c_1}{\sqrt{M}(\beta^{-1} - c_2)} + c_3 \exp\left\{-2 \left(\beta^{-1} m_F - 3H_F L_K - 2L_F \right) \sum_{k=0}^{T-1} h_k\right\} \ + c_3 \beta^{-3} Md^\frac{3}{2} (c_6 h_0^3 + \frac{c_7 \beta^3 h_0}{B_0} + c_8 h_0^2 \beta^2)^\frac{1}{2}. $$

where $(c_1, c_2, c_3, c_6, c_7, c_8, \beta)$ are some positive constants such that $\frac{1}{\beta} > c_2$ and $\frac{m_F}{\beta} > 3H_F L_K - 2L_F$, as
I Proof of Theorem 7 and Corollary 8

Proof [Proof of Theorem 7]
\[
\begin{align*}
\text{d} \left( \theta^{(i)}_t - \theta^{(j)}_t \right) &= -\beta^{-1} \left( F(\theta^{(i)}_t) - F(\theta^{(j)}_t) \right) \text{d}t \\
+ & \frac{1}{M} \sum_{q} \left[ \nabla K(\theta^{(i)}_t - \theta^{(q)}_t) - \nabla K(\theta^{(j)}_t - \theta^{(q)}_t) \right] \text{d}t \\
- & \frac{1}{M} \sum_{q} \left( F(\theta^{(q)}_t) K(\theta^{(i)}_t - \theta^{(q)}_t) - F(\theta^{(q)}_t) K(\theta^{(j)}_t - \theta^{(q)}_t) \right) \text{d}t \\
+ & \sqrt{\frac{\gamma}{\beta}} (dW_t^{(i)} - dW_t^{(j)}) \\
= & \text{d} \left( \mathbb{E} \sum_{ij} \left\| \theta^{(i)}_t - \theta^{(j)}_t \right\|^2 \right) = -\mathbb{E} 2 \sum_{ij} \beta^{-1} \left( F(\theta^{(i)}_t) - F(\theta^{(j)}_t) \right) \left( \theta^{(i)}_t - \theta^{(j)}_t \right) \text{d}t \\
+ & \mathbb{E} \sum_{ij} \frac{2}{M} \sum_{q} \left[ \nabla K(\theta^{(i)}_t - \theta^{(q)}_t) - \nabla K(\theta^{(j)}_t - \theta^{(q)}_t) \right] \left( \theta^{(i)}_t - \theta^{(j)}_t \right) \text{d}t \\
- & \mathbb{E} \sum_{ij} \frac{2}{M} \sum_{q} \left( F(\theta^{(q)}_t) K(\theta^{(i)}_t - \theta^{(q)}_t) - F(\theta^{(q)}_t) K(\theta^{(j)}_t - \theta^{(q)}_t) \right) \left( \theta^{(i)}_t - \theta^{(j)}_t \right) \text{d}t \\
+ & \mathbb{E} 2 \sum_{ij} \sqrt{\frac{2}{\beta}} \left( dW_t^{(i)} - dW_t^{(j)} \right) \left( \theta^{(i)}_t - \theta^{(j)}_t \right) \\
\leq & -2\beta^{-1} m_F \mathbb{E} \sum_{ij} \left\| \theta^{(i)}_t - \theta^{(j)}_t \right\|^2 \text{d}t - 2m_K \mathbb{E} \sum_{ij} \left\| \theta^{(i)}_t - \theta^{(j)}_t \right\|^2 \text{d}t \\
+ & 2H_F K \mathbb{E} \sum_{ij} \left\| \theta^{(i)}_t - \theta^{(j)}_t \right\|^2 \text{d}t \\
+ & 2 \sqrt{\frac{2}{\beta}} \left( \mathbb{E} \sum_{ij} (dW_t^{(i)} - dW_t^{(j)}) \right)^{1/2} \left( \mathbb{E} \sum_{ij} \left\| \theta^{(i)}_t - \theta^{(j)}_t \right\|^2 \right)^{1/2}.
\end{align*}
\]

Denote \( z(t) = \mathbb{E} \sum_{ij} \left\| \theta^{(i)}_t - \theta^{(j)}_t \right\|^2 \). We have
\[
z(t)' \leq -(2\beta^{-1} m_F + 2m_K - 2H_F L_K) z(t) + 4M \sqrt{\frac{d}{\beta} z(t)} (33)
\]

We can finish our proof by applying Gronwall Lemma on (33). □

Proof [Proof for Corollary 8]
\[
\begin{align*}
\text{d} \left( \theta^{(i)}_t - \theta^{(j)}_t \right) &= \\
+ & \frac{1}{M} \sum_{q} \left[ \nabla K(\theta^{(i)}_t - \theta^{(q)}_t) - \nabla K(\theta^{(j)}_t - \theta^{(q)}_t) \right] \text{d}t \\
- & \frac{1}{M} \sum_{q} \left( F(\theta^{(q)}_t) K(\theta^{(i)}_t - \theta^{(q)}_t) - F(\theta^{(q)}_t) K(\theta^{(j)}_t - \theta^{(q)}_t) \right) \text{d}t \\
\end{align*}
\]
\[
\Rightarrow d \left( E \sum_{ij}^M \left\| \theta_{ij}^{(i)} - \theta_{ij}^{(j)} \right\|^2 \right) = \\
+ E \sum_{ij}^M 2 \sum_\mathcal{M} \left[ \nabla K(\theta_{ij}^{(i)} - \theta_{ij}^{(q)}) - \nabla K(\theta_{ij}^{(j)} - \theta_{ij}^{(q)}) \right] \left( \theta_{ij}^{(i)} - \theta_{ij}^{(j)} \right) dt \\
- E \sum_{ij}^M 2 \sum_\mathcal{M} \left( F(\theta_{ij}^{(q)})K(\theta_{ij}^{(i)} - \theta_{ij}^{(q)}) - F(\theta_{ij}^{(q)})K(\theta_{ij}^{(j)} - \theta_{ij}^{(q)}) \right) \left( \theta_{ij}^{(i)} - \theta_{ij}^{(j)} \right) dt \\
\leq -2m_K E \sum_{ij}^M \left\| \theta_{ij}^{(i)} - \theta_{ij}^{(j)} \right\|^2 dt + 2H_F L_K E \sum_{ij}^M \left\| \theta_{ij}^{(i)} - \theta_{ij}^{(j)} \right\|^2 dt
\]

Denote \( z(t) = E \sum_{ij}^M \left\| \theta_{ij}^{(i)} - \theta_{ij}^{(j)} \right\|^2 \). We have

\[
z(t)' \leq -(2m_K - 2H_F L_K)z(t) \tag{34}
\]

We can finish our proof by applying Gronwall Lemma on (34).  

\[\blacksquare\]

J  PROOF OF THEOREM 9

Proof  [Proof of Theorem 9] Our conclusion for \( \tilde{B}(\hat{\mu}_T, \hat{\mu}_\infty) \) is essentially a specification of the result in Mattingly et al. (2002), which has also been applied in Xu et al. (2018).

Specifically, we rely on the following lemma, which is essentially Theorem 7.3 in Mattingly et al. (2002) and Lemma C.3 in Xu et al. (2018), considering the SDE (16):

\[d\Theta_t = -F_{\Theta}(\Theta_t)dt + \sqrt{2\beta} \frac{1}{M}dW_t^{(M,d)}\]

As mentioned in Section 5, we firstly denote the distribution of \( \Theta_t \) as \( \rho_{\Theta}^t \). Then we define the \( \hat{\Theta}_k \triangleq [\hat{\theta}_{1k}, \cdots, \hat{\theta}_{MK}] \in \mathbb{R}^{M,d} \), which is actually the numerical solution of (16) using full gradient with Euler method. And we denote the distribution of \( \hat{\Theta}_k \) as \( \hat{\rho}_k \).

Lemma 17  Let \( F_{\Theta} \) be Lipschitz-continuous with constant \( L_\Theta \), and satisfy the dissipative property that \( (F_{\Theta}(\Theta), \Theta) \geq m_\Theta \| \Theta \|^2 - b_\Theta \). Define \( V_{\Theta}(\Theta) = C_0 + L_\Theta/2 \| \Theta \|^2 \). The Euler method for (16) has a unique invariant measure \( \hat{\mu}_\infty^\Theta \), and for all test function \( f_\Theta \) such that \( |f_\Theta| \leq V_{\Theta}(\Theta) \), we have

\[
\left| E[f_\Theta(\hat{\Theta}_k))] - E_{\Theta_j \sim \hat{\mu}_\infty^\Theta} [f(\hat{\Theta}_{\infty})] \right| \leq C K \rho^{-M/2} (1 + \kappa e^{m_\Theta h}) \exp \left( \frac{-2m_\Theta kh \rho^{M/2}}{\log(\kappa)} \right),
\]

where \( \rho \in (0, 1) \), \( C > 0 \) are positive constants, and \( k = 2L_\Theta(b_\Theta \beta + m_\Theta \beta + Md)/m_\Theta \).

Now we adopt the \( f_\Theta : \mathbb{R}^{M,d} \rightarrow \mathbb{R} \) defined as \( f_\Theta(\Theta) = \frac{1}{M} \sum_{i=1}^M f(\theta^{(i)}) \), where \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) is a \( L_f \)-Lipschitz function and satisfies our Assumption 2 and \( \Theta \triangleq [\theta^{(1)}, \cdots, \theta^{(M)}] \). Similar to the proof of Lemma 12, we can find that \( f_\Theta : \mathbb{R}^{M,d} \) is a \( L_f/\sqrt{M} \)-Lipschitz function. Furthermore, according to the Lemma 14, it is easily check that \( F_{\Theta} \) is \( L_\Theta \)-Lipschitz where \( L_\Theta = \sqrt{2} \beta^{-1} L_F + l' \).

Hence, when the \( \beta \) is small enough, we have \( L_f/\sqrt{M} \leq \sqrt{2} \beta^{-1} L_F + l' \). Then we can set the \( C_0 \) large enough to force \( f_\Theta \) to satisfy the condition in Lemma 17 that \( |f_\Theta| \leq V_{\Theta}(\Theta) \). Then, according to the exchangeability of the particle system \( \{\hat{\theta}_{ik}\} \) and Lemma 14, we can bound the \( \hat{B}(\hat{\mu}_T, \hat{\mu}_\infty) \) as

\[
\hat{B}(\hat{\mu}_T, \hat{\mu}_\infty) \leq \left| E[f_\Theta(\hat{\Theta}_T))] - E_{\Theta_j \sim \hat{\mu}_\infty^\Theta} [f(\hat{\Theta}_{\infty})] \right| \leq C_2 \sigma^{-M/2}(1 + \varsigma e^{m_\Theta h}) \exp \left( \frac{-2m_\Theta Th \sigma^{M/2}}{\log(\varsigma)} \right)
\]
where \( \varsigma = 2L_\Theta(Mb + m_\varphi \beta + Md)/m_\varphi \), \( L_\Theta = \sqrt{2} \beta^{-1} L_F + l' \), \( m_\varphi = \beta^{-1} m - m' \), and \( (\sigma, C_2, l', m') \) are some positive constants independent of \((T, M, h)\) and \( \sigma \in (0, 1) \).

To prove the bound for \( \tilde{B}(\hat{\mu}_\infty, \rho_\infty) \). Since \( \hat{\Theta}_k = (\hat{\theta}_k^{(1)}, \ldots, \hat{\theta}_k^{(M)}) \) can be considered as a solution to the SDE \( \{16\} \), standard results from linear FP equation can be applied. Specifically, for the \( \tilde{B}(\hat{\mu}_\infty, \rho_\infty) \) term, we rely on the following lemma adapted from Lemma C.4 in Xu et al. (2018), and is essentially the work of Chen et al. (2015) when taking \( K \to \infty \).

**Lemma 18** Under the same assumption as in Lemma [17] for the Lipschitz-continuous function \( f(\Theta) = \frac{1}{N} \sum_i f(\theta_i) \) mentioned above, the following bound is satisfied for some positive constant \( C \):

\[
\left| \frac{1}{K} \sum_{k=1}^{K-1} E[f(\hat{\Theta}_k)] - E[f(\hat{\Theta}_\infty)] \right| \leq C \left( \frac{h}{\beta} + \frac{1}{K \beta} \right).
\]

The uniqueness of invariant measure from Lemma [17] implies the numerical solution \( \hat{\Theta}_k \) to be ergodic. Then similar to the proof of Lemma 4.2 in Xu et al. (2018), we consider the case where \( K \to \infty \). Take average over the \( \{\hat{\Theta}_k\}_{k=0}^K \), we have

\[
E_{\hat{\Theta}_\infty \sim \hat{\mu}_\infty} [f(\hat{\Theta}_\infty)] = \lim_{K \to \infty} \frac{1}{K} \sum_{k=1}^{K-1} E[f(\hat{\Theta}_k)]
\]

Now according to the exchangeability of the particle system \( \{\hat{\theta}_k^{(i)}\} \) and \( \{\hat{\theta}_k^{(i)}\} \), we can bound the \( \tilde{B}(\hat{\mu}_\infty, \rho_\infty) \) as:

\[
\tilde{B}(\hat{\mu}_\infty, \rho_\infty) \leq \left| E_{\hat{\Theta}_\infty \sim \hat{\mu}_\infty} [f(\hat{\Theta}_\infty)] - E_{\hat{\Theta}_\infty \sim \hat{\mu}_\infty} [f(\hat{\Theta}_\infty)] \right| \leq C_3 h / \beta
\]

where \( C_3 \) are some positive constant.

**K PROOF OF THEOREM [10]**

**Proof** [Proof of Theorem [10]] Adopting the same notation used in the proof of the Theorem [5] we define \( \Theta_k \triangleq [\hat{\theta}_k^{(1)}, \ldots, \hat{\theta}_k^{(M)}] \) and \( G^\Theta_{I_k} \triangleq \frac{N}{B_\varepsilon} \sum_i \mathbb{1}_{I_k} f(\theta_i) \Theta_k \). We denote the distribution of \( \Theta_k \) as \( \mu^\Theta_k \).

\[
\Theta_{k+1} = \Theta_k - \beta^{-1} G^\Theta_{I_k} h_k + \sqrt{2\beta^{-1} h_k} \varepsilon_k.
\]

We firstly give a bound to \( \mathcal{W}_2(\mu^\Theta_k, \bar{\mu}^\Theta_k) \) (the definition of \( \bar{\mu}^\Theta_k \) has been mention in the last section). According to the proof of the Lemma 4.4 in Xu et al. (2018)

\[
\mathcal{W}_2(\mu^\Theta_k, \bar{\mu}^\Theta_k) \leq kh(L_\Theta \Gamma' + MC_4)((6 + 2\Gamma') \beta/B)^{1/2}
\]

where \( \Gamma' = 2(1 + 1/m_\varphi)(Mb + 2M^2C_2^2 + Md/\beta) \) and \( C_4 \) is some positive constant independent of \((T, M, h)\). Note the fact that \( \mathcal{W}_1(\mu^\Theta_k, \bar{\mu}^\Theta_k) \leq \mathcal{W}_2(\mu^\Theta_k, \bar{\mu}^\Theta_k) \) and \( \mathcal{W}_1(\mu_k, \bar{\mu}_k) \leq \frac{1}{\sqrt{M}} \mathcal{W}_1(\mu^\Theta_k, \bar{\mu}^\Theta_k) \) (see the proof of Lemma [13] similar result holds here), We get

\[
\mathcal{W}_1(\mu_T, \bar{\mu}_T) \leq Th(L_\Theta \Gamma' + MC_4)((6 + 2\Gamma') \beta/(BM))^{1/2}.
\]

Let us compare the definition of \( \mathcal{W}_1(\mu, \nu) \) and \( \tilde{B}(\mu, \nu) \)

\[
\mathcal{W}_1(\mu, \nu) \triangleq \sup_{\|g\|_{L_1} \leq 1} |E_{\Theta \sim \mu}[g(\Theta)] - E_{\Theta \sim \nu}[g(\Theta)]|
\]

\[
\tilde{B}(\mu, \nu) \triangleq |E_{\Theta \sim \mu}[f(\Theta)] - E_{\Theta \sim \nu}[f(\Theta)]|
\]

and we can derive the result that \( \tilde{B}(\mu_T, \bar{\mu}_T) \leq L_1 \mathcal{W}_1(\mu_T, \bar{\mu}_T) \). Now we finish our proof.
L DISCUSSION ON THE COMPLEXITY OF OUR METHOD

The complexity of an algorithm mainly refers to its time complexity (corresponding to the number of iterations in our method i.e. $T$) and space complexity (corresponding to the number of particles used in our method i.e. $M$). Hence the complexity of our method can be well explored with our work, since our non-asymptotic convergence theory was developed w.r.t. the number of particles $i.e. M$ and iterations $i.e. T$. Their relationship (tradeoff) was even discussed further in Experiment 6.1. Moreover, comparing \[9\] with \[3\], one can easily find that our space complexity is exactly the same as SVGD and our computational time in each iteration is almost the same as SVGD with an extra addition operation. However, it is worth noting that our method has much better performance in practice and no "pitfall" verified by both our theory and our experiments.

M COMPARISON WITH RELATED WORK

Firstly, our proposed framework SPOS is different from the recently proposed particle-optimization sampling framework (Chen et al., 2018) in the sense that we solve the nonlinear PDE \[10\] stochastically. For example, they solve the equation in \[6\] \[\partial_t \nu_t = \beta^{-1} \nabla \theta \cdot \nabla \theta \nu_t \] approximately using blob method adopted from (Carrillo et al., 2017). Secondly, our method is also distinguishable to existing work on granular media equations such as (Durmus et al., 2018). Their work about the granular media equations focuses on the following PDE:

\[\partial_t \nu_t = \nabla \theta \cdot \left( \nu_t \beta^{-1} F(\theta) + \nu_t (\nabla K \ast \nu_t) + \beta^{-1} \nabla \theta \nu_t \right), \quad \text{(35)}\]

whereas our framework focuses on the following one:

\[\partial_t \nu_t = \nabla \theta \cdot \left( \nu_t \beta^{-1} F(\theta) + \nu_t \left( E_{Y \sim \mu_{\theta}} K(\theta - Y) F(Y) - \nabla K \ast \nu_t(\theta) \right) + \beta^{-1} \nabla \theta \nu_t \right). \quad \text{(36)}\]

The extra term \(\nu_t \left( E_{Y \sim \mu_{\theta}} K(\theta - Y) F(Y) \right)\) in our framework makes the analysis much more challenging. The main differences are summarized below:

- Formulations are different. The extra term \(E_{Y \sim \mu_{\theta}} K(\theta - Y) F(Y)\) cannot be combined with the \(F(\theta)\) term in their (35). This is because function \(F(\theta)\) itself is a function independent of \(t\); while \(E_{Y \sim \mu_{\theta}} K(\theta - Y) F(Y)\) depends on both \(\theta\) and \(t\). This makes our problem much more difficult.

- Assumptions are different. For example, the analysis on granular media equations in (Cattiaux et al., 2008) requires that \(F\) satisfies a special condition \(C(\alpha, \sigma)\), which is a strong condition impractical to be satisfied in our case; And (Durmus et al., 2018) adopts different assumptions from ours with a different goal.

- For the Euler integrator, (Durmus et al., 2018) does not consider an Euler solution. Furthermore, our sampling method needs "stochastic gradient" i.e. \(G^{(i)}_k \triangleq \frac{N}{N_k} \sum_{q \in I_k} \nabla_{\theta} F_q(\theta^{(i)}_k)\) in (9) for computational feasibility, which is quite different from the former work on particle-SDE such as (Malrieu 2003; Cattiaux et al. 2008). Few of the former work on particle-SDE considered the stochastic gradient issue.

To sum up, the main purpose of our paper is to provide a non-asymptotic analysis of our method instead of improving the former work on a certain type of PDE. This is also the reason why we said that part of our proof techniques are based on those for analyzing granular media equations.

N EXTRA EXPERIMENTS

N.1 TOY EXPERIMENTS

We compare the proposed SPOS with other popular methods such as SVGD and standard SGLD on four multi-mode toy examples. We aim to sample from four unnormalized 2D densities \(p(z)/\exp(U(z))\), with the functional form provided in (Rezende & Mohamed, 2015). We optimize/sample 50 and 2000 particles to approximate the target distributions. The results are illustrated in Figure 4 and Figure 5 respectively.
N.1.1 BAYESIAN NEURAL NETWORKS FOR MNIST CLASSIFICATION

We perform the classification tasks on the standard MNIST dataset. A two-layer MLP 784-X-X-10 with ReLU activation function is used, with X being the number of hidden units for each layer. The training epoch is set to 100. The test errors are reported in Table 2. Surprisingly, the proposed SPOS outperforms other algorithms such as SVGD at a significant level, though it is just a simple modification of SVGD by adding in random Gaussian noise. This is partly due to the fact that our SPOS algorithm can jump out of local modes efficiently, as explained in Section 4.4.
Table 2: Classification error of FNN on MNIST.

<table>
<thead>
<tr>
<th>Method</th>
<th>Test Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>400-400</td>
</tr>
<tr>
<td>SPOS</td>
<td>1.32%</td>
</tr>
<tr>
<td>SVGD</td>
<td>1.56%</td>
</tr>
<tr>
<td>SGLD</td>
<td>1.64%</td>
</tr>
<tr>
<td>RMSprop</td>
<td>1.59%</td>
</tr>
<tr>
<td>RMSspectral</td>
<td>1.65%</td>
</tr>
<tr>
<td>SGD</td>
<td>1.72%</td>
</tr>
<tr>
<td>BPB, Gaussian</td>
<td>1.82%</td>
</tr>
<tr>
<td>SGD, dropout</td>
<td>1.51%</td>
</tr>
</tbody>
</table>