Stochastic Particle-Optimization Sampling and the Non-Asymptotic Convergence Theory

Jianyi Zhang[1] **Ruiyi Zhang**[1] **Lawrence Carin**[1] **Changyou Chen**[2] ✉
[1]Duke University [2]University at Buffalo, SUNY

Abstract

Particle-optimization-based sampling (POS) is a recently developed effective sampling technique that interactively updates a set of particles to approximate a target distribution. A representative algorithm is the Stein variational gradient descent (SVGD). We prove, under certain conditions, SVGD experiences a theoretical pitfall, *i.e.*, particles tend to collapse. As a remedy, we generalize POS to a stochastic setting by injecting random noise into particle updates, thus termed stochastic particle-optimization sampling (SPOS). Notably, for the first time, we develop *non-asymptotic convergence theory* for the SPOS framework (related to SVGD), characterizing algorithm convergence in terms of the 1-Wasserstein distance w.r.t. the numbers of particles and iterations. Somewhat surprisingly, with the same number of updates (not too large) for each particle, our theory suggests adopting more particles does not necessarily lead to a better approximation of a target distribution, due to limited computational budget and numerical errors. This phenomenon is also observed in SVGD and verified via a synthetic experiment. Extensive experimental results verify our theory and demonstrate the effectiveness of our proposed framework.

1 Introduction

Recently there has been extensive development of scalable Bayesian sampling algorithms, such as stochastic gradient MCMC (SG-MCMC) [Welling and Teh, 2011, Chen et al., 2014, Ding et al., 2014, Chen et al., 2015] and Stein variational gradient descent (SVGD) [Liu and Wang, 2016]. SG-MCMC is a family of scalable Bayesian sampling algorithms built on Itó diffusions, stochastic differential equations (SDEs) with appropriately designed coefficients whose stationary distributions match the target distributions. One potential issue of SG-MCMC is that samples may be highly correlated partially due to the nature of Markov chains, leading to undesired low sample-efficiency. SVGD, on the other hand, belongs to the family of particle-optimization-based sampling methods that optimize a set of interacting particles to minimize some distance metric (*e.g.*, KL-divergence) between the target distribution and the particle-induced approximate distribution. By optimization, one seeks to maintain an optimal set of particles. Recent development of SVGD has shown that the underlying mathematical principle is based on a family of *nonlinear* partial differential equations (PDEs) [Liu, 2017]. Although achieving significant practical successes [Liu and Wang, 2016, Feng et al., 2017, Liu et al., 2017, Haarnoja et al., 2017, Zhang et al., 2018a, Zhang et al., 2019, Liu and Zhu, 2018], little theory is available to fully understand its *non-asymptotic* convergence properties under numerical errors. A recent theoretical development has interpreted SVGD as a special type of gradient flows, and developed theory to disclose its *asymptotic* convergence behavior [Liu, 2017]. The asymptotic theory is also studied in [Lu et al., 2018]. A more recent work [Liu and Wang, 2018] investigated non-asymptotic properties of SVGD, limited to the region of finite particles and infinite time with restricted conditions. [Şimşekli et al., 2018] considers convergence property of the sliced-Wasserstein flow only under an infinite-particle setting.

Recently, [Chen et al., 2018] unified SG-MCMC and SVGD by proposing a particle-optimization-sampling (POS) framework to interpret both as Wasserstein gradient flows (WGFs). Generally, a WGF is a PDE defined on the space of probability measures, describing the evolution of a density over time. [Chen et al., 2018] defined a WGF by combining the corresponding PDEs for both SG-MCMC and SVGD, and solved it with deterministic particle approximations. However,

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

due to its diffusion nature, deterministic-particle approximation leads to a hard-to-control error, making it challenging for theoretical analysis.

Our contributions In this paper, we generalize POS to a stochastic setting, and develop a novel analytical framework based on granular media equations [Malrieu, 2003, Cattiaux et al., 2008] to analyze its non-asymptotic convergence properties. Our contributions are summarized as follows: *i*) We first identify a pitfall of standard SVGD, where particles tend to collapse under certain conditions and measurement, indicating challenges in developing non-asymptotic theory for SVGD (if possible at all). *ii*) Based on the unified framework in [Chen et al., 2018], we propose *stochastic* particle-optimization sampling (SPOS) by injecting Gaussian noise in particle updates to overcome the pitfall. *iii*) For the first time, we develop nonasymptotic convergence theory for the family of SPOS algorithms, considering both convex- and nonconvex-energy targets. Different from existing theory for SG-MCMC-based algorithms [Teh 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 PDEs*, 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 non-asymptotic error bounds in terms of 1-Wasserstein distance. More detailed distinctions between our work and existing work are discussed in Section N of the Supplementary Material (SM). Somewhat surprisingly, our theory indicates adopting more particles does not necessarily lead to better approximations, due to the numerical errors in the algorithms. This phenomenon is also observed for SVGD empirically. *iv*) Our theory and advantages of the algorithm are verified via various experiments, including synthetic experiments, Bayesian deep learning and Bayesian exploration for reinforcement learning.

2 Preliminaries

Notation We use bold letters to denote variables in *continuous-time diffusions and model definitions* (no numerical methods included), *e.g.*, $\boldsymbol{\theta}_\tau$ in (1) below (indexed by "time" τ). By contrast, *unbold letters* are used to denote parameters in *algorithms* (numerical solutions of continuous-time diffusions), *e.g.*, $\theta_k^{(i)}$ in (3) below (indexed by "iteration" k). For conciseness, all proofs, extra experimental results and a discussion on algorithmic complexity are presented in the SM.

2.1 Stochastic gradient MCMC

In Bayesian sampling, one aims to generate random samples from a posterior distribution $p(\boldsymbol{\theta}|\mathcal{X}) \propto p(\mathcal{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \mathbb{R}^d$ represents the model parameter with a prior distribution $p(\boldsymbol{\theta})$, and $\mathcal{X} \triangleq \{\mathbf{x}_q\}_{q=1}^N$ represents the observed data with likelihood $p(\mathcal{X}|\boldsymbol{\theta}) = \prod_q p(\mathbf{x}_q|\boldsymbol{\theta})$. Define the potential energy as: $U(\boldsymbol{\theta}) \triangleq -\log p(\mathcal{X}|\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}) = -\sum_{q=1}^N \left(\log p(\mathbf{x}_q|\boldsymbol{\theta}) + \frac{1}{N}\log p(\boldsymbol{\theta})\right) \triangleq \sum_{q=1}^N U_q(\boldsymbol{\theta})$. 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 SDE. For example, in stochastic gradient Langevin dynamic (SGLD) [Welling and Teh, 2011], the SDE endows the following form:

$$\mathrm{d}\boldsymbol{\theta}_\tau = -\beta^{-1}F(\boldsymbol{\theta}_\tau)\mathrm{d}\tau + \sqrt{2\beta^{-1}}\mathrm{d}\mathcal{W}_\tau \ , \qquad (1)$$

where $F(\boldsymbol{\theta}) \triangleq \nabla_{\boldsymbol{\theta}}U(\boldsymbol{\theta}) = \sum_{q=1}^N \nabla_{\boldsymbol{\theta}}U_q(\boldsymbol{\theta}) \triangleq \sum_{q=1}^N F_q(\boldsymbol{\theta})$; τ is the time index; $\beta > 0$ is the temperature parameter; and $\mathcal{W}_\tau \in \mathbb{R}^d$ 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 a more detailed description of general SG-MCMC algorithms. Denote the probability density function of $\boldsymbol{\theta}_\tau$ in (1) as ν_τ, and let $\mathbf{a}\cdot\mathbf{b} \triangleq \mathbf{a}^\top\mathbf{b}$ for two vectors $\mathbf{a}$ and $\mathbf{b}$. It is known that ν_t is characterized by the following Fokker-Planck (FP) equation [Risken, 1989]:

$$\partial_\tau\nu_\tau = \nabla_{\boldsymbol{\theta}}\cdot\left(\beta^{-1}\nu_\tau F(\boldsymbol{\theta}) + \beta^{-1}\nabla_{\boldsymbol{\theta}}\nu_\tau\right) . \qquad (2)$$

According to [Chiang and Hwang, 1987], the stationary distribution ν_∞ equals to our target distribution $p(\boldsymbol{\theta}|\mathcal{X})$. As a result, SGLD is designed to generates samples from $p(\boldsymbol{\theta}|\mathcal{X})$ by numerically solving the SDE (1). For scalability, it replaces $F(\theta_k)$ in each iteration with an unbiased evaluation by randomly sampling a subset of $\mathcal{X}$, *i.e.*, $F(\theta_k)$ is approximated by: $G_k \triangleq \frac{N}{B_k}\sum_{q\in\mathcal{I}_k}F_q(\theta_k)$, where $\mathcal{I}_k$ is a random subset of $[1, 2, \cdots, N]$ with size B_k in each iteration. As a result, SGLD uses the Euler method with stepsize h_k to numerically solve (1), resulting in the update equation: $\theta_{k+1} = \theta_k - \beta^{-1}G_kh_k + \sqrt{2\beta^{-1}h_k}\xi_k$, with $\xi_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

2.2 Stein variational gradient descent

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

$$= \theta_k^{(i)} + \frac{h_k}{M}\sum_{j=1}^{M}\left[\kappa(\theta_k^{(j)},\theta_k^{(i)})G_k^{(i)} + \nabla_{\theta_k^{(j)}}\kappa(\theta_k^{(j)},\theta_k^{(i)})\right] \quad (3)$$

where the first term in the bracket encourages moving particles to the density modes, and the second term serves as repulsive force that pushes away different particles. Particularly, the particle evolution (3) are numerical solutions of the ODEs: $\mathrm{d}\boldsymbol{\theta}_\tau^{(i)} = \frac{1}{M}\sum_{j=1}^M\left[\kappa(\boldsymbol{\theta}_\tau^{(j)},\boldsymbol{\theta}_\tau^{(i)})F(\boldsymbol{\theta}_\tau^{(i)}) + \nabla_{\boldsymbol{\theta}_\tau^{(j)}}\kappa(\boldsymbol{\theta}_\tau^{(j)},\boldsymbol{\theta}_\tau^{(i)})\right]\mathrm{d}\tau$. Different from SG-MCMC, typically 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

SG-MCMC and SVGD, though they may look closely related, behave very differently as algorithms, *e.g.*, stochastic and noninteractive versus deterministic and interactive particle updates. Recently, [Chen et al., 2018] proposed a deterministic particle-optimization framework that unified SG-MCMC and SVGD. Specifically, the authors viewed both SG-MCMC and SVGD as solutions of Wasserstein gradient flows (WGFs) on the space of probabilistic measures, and derived several deterministic particle-optimization techniques for particle evolution, like SVGD. 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 the nonlinear PDE literature, [Chen et al., 2018] proposed a general form of PDE to characterize the evolution of the density for the model parameter $\boldsymbol{\theta}$, denoted as ν_τ at time τ with ν_∞ matching our target (posterior) distribution, *i.e.*,

$$\partial_\tau\nu_\tau = \nabla_{\boldsymbol{\theta}}\cdot\left(\nu_\tau\beta^{-1}F(\boldsymbol{\theta}) + \nu_\tau\left(\mathcal{K}*\nu_\tau(\boldsymbol{\theta})\right) + \beta^{-1}\nabla_{\boldsymbol{\theta}}\nu_\tau\right), \quad (4)$$

where $\mathcal{K}$ is a function controlling the interaction of particles in the PDE system. For example, in SVGD, [Chen et al., 2018] showed that $\mathcal{K}$ and $\mathcal{K}*\nu_\tau(\boldsymbol{\theta})$ endow the following forms:

$$\mathcal{K}*\nu_\tau(\boldsymbol{\theta}) \triangleq \int\mathcal{K}(\boldsymbol{\theta},\boldsymbol{\theta}')\nu_\tau(\boldsymbol{\theta}')d\boldsymbol{\theta}', \quad (5)$$

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

$$\begin{aligned}\partial_\tau\nu_\tau =& \nabla_{\boldsymbol{\theta}}\cdot\left(\nu_\tau\beta^{-1}F(\boldsymbol{\theta}) + \nu_\tau\left(E_{Y\sim\nu_\tau}K(\boldsymbol{\theta}-Y)F(Y)\right.\right.\\ &\left.\left.-\nabla K*\nu_\tau(\boldsymbol{\theta})\right) + \beta^{-1}\nabla_{\boldsymbol{\theta}}\nu_\tau\right), \quad (6)\end{aligned}$$

where Y is a random sample from ν_τ independent of $\boldsymbol{\theta}$. Note our formula here is significantly different from standard granular media equations in the literature. Please refer to Section N of the SM for more details.

Proposition 1 **([Chen et al., 2018])** *The stationary distribution of (6) equals to our target distribution, which means $\nu_\infty(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathcal{X})$.*

[Chen et al., 2018] proposed to solve (4) numerically with deterministic particle-optimization algorithms, such as what is called the blob method. Specifically, the continuous density ν_τ is approximated by a set of M particles $\{\boldsymbol{\theta}_\tau^{(i)}\}_{i=1}^M$ that evolve over time τ, *i.e.* $\nu_\tau \approx \frac{1}{M}\sum_{i=1}^M\delta_{\boldsymbol{\theta}_\tau^{(i)}}(\boldsymbol{\theta})$, where $\delta_{\boldsymbol{\theta}_\tau^{(i)}}(\boldsymbol{\theta}) = 1$ if $\boldsymbol{\theta} = \boldsymbol{\theta}_\tau^{(i)}$ and 0 otherwise. Note $\nabla_{\boldsymbol{\theta}}\nu_\tau$ in (4) is no longer a valid definition when adopting particle approximation for ν_τ. Consequently, $\nabla_{\boldsymbol{\theta}}\nu_\tau$ 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 (SPOS)

We first introduce a pitfall of SVGD, which is overcame by SPOS. In the analysis for both SVGD and SPOS, we impose the following basic assumptions.

Assumption 1 *Assume F and K satisfy the following assumptions:*

1.1 F is L_F-Lipschitz continuous i.e., $\|F(\boldsymbol{\theta}) - F(\boldsymbol{\theta}')\| \le L_F\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|$.

1.2 K is L_K-Lipschitz continuous; ∇K is $L_{\nabla K}$-Lipschitz continuous.

1.3 $F(\mathbf{0}) = \mathbf{0}$ and K is an even function, i.e., $K(-\boldsymbol{\theta}) = K(\boldsymbol{\theta})$.

A few remarks: *i)* Assumptions 1.1 is widely adopted in the other theoretical works such as [Dalalyan and Karagulyan, 2017, Chatterji et al., 2018] *ii)* $F(\mathbf{0}) = \mathbf{0}$ in Assumption 1.3 is reasonable, as F in our setting corresponds to an unnormalized log-posterior, which can be shifted such that $F(\mathbf{0}) = \mathbf{0}$ for a specific problem. The assumptions of K are mild, and satisfied when adopting the RBF Kernel.

3.1 A pitfall of SVGD

First, we motivate SPOS by discovering a pitfall of standard SVGD, *i.e.*, particles in SVGD tend to collapse to a local mode under some particular conditions. Inspired by the work on analyzing granular media equations by [Malrieu, 2003, Cattiaux et al., 2008], we measure the collapse by calculating the expected distance between *exact particles* (without numerical errors), called expected particle distance (EPD) defined below.

Assumption 2 *F and K satisfy the following assumptions:*

2.1 There exists positive m_K such that $\langle \nabla K(\boldsymbol{\theta}) - \nabla K(\boldsymbol{\theta}'), \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle \leq -m_K \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2$.

2.2 F is bounded by H_F i.e., $\|F(\boldsymbol{\theta})\| \leq H_F$

For an RBF kernel, this assumption could be satisfied by setting the bandwidth large enough and only considering the concave region. This seems a little restrictive. However, this assumption is imposed only for the analysis of the pitfall property. It is not needed in the non-asymptotic convergence analysis below. Besides, we point out what might happen without this assumption in Remark 1.

Theorem 2 *Under Assumptions 1 and 2, for the particles $\boldsymbol{\theta}_\tau^{(i)}$ defined in Section 2.2, the EPD for SVGD is bounded as: EPD $\triangleq \sqrt{\sum_{i,j}^M \mathbb{E}\|\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}\|^2} \leq C_0 e^{-2\lambda\tau}$, where $C_0 = \sqrt{\sum_{i,j}^M \|\boldsymbol{\theta}_0^{(i)} - \boldsymbol{\theta}_0^{(j)}\|^2}$, $\lambda = m_K - H_F L_K$.*

Remark 1 *1) In the case of $\lambda \geq 0$, Theorem 2 indicates that particles in SVGD would collapse to a point when $\tau \to \infty$. In practice, we usually find that particles are trapped in a local mode instead of collapsing in practice. This might be due to two reasons: i) Particles in SVGD are numerical solutions instead of exact solutions as used in EPD, which induce extra numerical errors; ii) Some particles might be out of the concave region of K stated in Assumption 2 in SVGD, which is required for the result to hold. These make the empirical EPD behave not exactly the same as the true particle distance. 2) Theorem 2 and its proof in the SM also apply to the case of non-convex energy functions. 3) Even if the kernel is not concave, the result would still indicate that particles in the* concave regions *would collapse. 4) The pitfall indicates a challenge in developing non-asymptotic theory for SVGD (if possible at all), motivating the development of SPOS. 5) This is a complement to the result of [Liu et al., 2019], which proves SVGD is ill-pose under some conditions.*

3.2 Stochastic particle-optimization sampling to mitigate the pitfall

We argue the WGF framework proposed in [Chen et al., 2018], if solved appropriately, is able to overcome the pitfall of SVGD. Specifically, the original solution in [Chen et al., 2018] is based on a deterministic particle-approximation method for (4), which introduces hard-to-control approximation errors. Instead, we propose to solve (4) *stochastically* to replace the $\nabla_{\boldsymbol{\theta}} \nu_\tau$ term in (4) with a Brownian motion. Specifically, first note that the term $\beta^{-1}\nabla_{\boldsymbol{\theta}} \cdot \nabla_{\boldsymbol{\theta}} \nu_\tau$ is contributed from Brownian motion, *i.e.*, solving the SDE, $\mathrm{d}\boldsymbol{\theta}_\tau = \sqrt{2\beta^{-1}}\mathrm{d}\mathcal{W}_\tau$, is equivalent to solving the corresponding FP equation: $\partial \nu_\tau = \beta^{-1}\nabla_{\boldsymbol{\theta}} \cdot \nabla_{\boldsymbol{\theta}} \nu_\tau$. Consequently, we decompose the RHS of (4) into two parts: $F_1 \triangleq \nabla_{\boldsymbol{\theta}} \cdot \left(\nu_\tau \beta^{-1} F(\boldsymbol{\theta}_\tau) + (\mathcal{K} * \nu_\tau)\nu_\tau\right)$ and $F_2 \triangleq \beta^{-1}\nabla_{\boldsymbol{\theta}} \cdot \nabla_{\boldsymbol{\theta}} \nu_\tau$. 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 ν_τ, both solutions of F_1 and F_2 are represented in terms of particles $\{\boldsymbol{\theta}_\tau^{(i)}\}$. 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 [Liu, 2017], we first formally show in Theorem 3 that when approximating ν_τ with particles, *i.e.*, $\nu_\tau \approx \frac{1}{M}\sum_{i=1}^M \delta_{\boldsymbol{\theta}_\tau^{(i)}}(\boldsymbol{\theta})$, the PDE can be transformed into a system of deterministic differential equations with interacting particles.

Theorem 3 *When approximating ν_τ in (4) with particles $\{\boldsymbol{\theta}_\tau^{(i)}\}$, the PDE $\partial_\tau \nu_\tau = F_1$ reduces to the following system of differential equations describing evolutions of the particles over time: $\forall i$*

$$\mathrm{d}\boldsymbol{\theta}_\tau^{(i)} = -\beta^{-1}F(\boldsymbol{\theta}_\tau^{(i)})\mathrm{d}\tau - \frac{1}{M}\sum_{j=1}^M K(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)})F(\boldsymbol{\theta}_\tau^{(j)})\mathrm{d}\tau + \frac{1}{M}\sum_{j=1}^M \nabla K(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)})\mathrm{d}\tau \quad (7)$$

Consequently, by solving $\partial_\tau \nu_\tau = F_2$ stochastically from an SDE perspective, we arrive at the following differential equation system, describing evolution of the particles $\{\boldsymbol{\theta}_\tau^{(i)}\}$ over time τ: $\forall i$

$$\mathrm{d}\boldsymbol{\theta}_\tau^{(i)} = -\beta^{-1}F(\boldsymbol{\theta}_\tau^{(i)} - \frac{1}{M}\sum_{j=1}^M K(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)})F(\boldsymbol{\theta}_\tau^{(j)}) + \frac{1}{M}\sum_{j=1}^M \nabla K(\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}))\mathrm{d}\tau + \sqrt{2\beta^{-1}}\mathrm{d}\mathcal{W}_\tau^{(i)} \quad (8)$$

Algorithm 1 Stochastic Particle-Optimization Sampling

```
Input: Initial particles $\{\theta_0^{(i)}\}_{i=1}^M$ with $\theta_0^{(i)} \in \mathbb{R}^d$, step size $h_k$, batch size $B_k$
for iteration $k$= 0,1,...,T do
    Update $\theta_{k+1}^{(i)}$ with (9) for $\forall i$.
end for
Output:$\{\theta_T^{(i)}\}_{i=1}^M$
```

Our intuition is that if the particle evolution (8) can be solved exactly, the solution of (6) ν_τ will be well-approximated by the particles $\{\boldsymbol{\theta}_\tau^{(i)}\}_{i=1}^M$. In our theory, we show this intuition is true. In practice, however, solving (8) is typically infeasible, and thus numerical methods are adopted. Furthermore, in the case of big

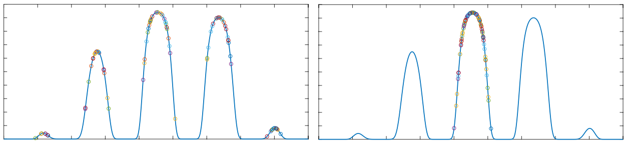

Figure 1: Comparison of SPOS (left) and SVGD (right) on a multi-mode distribution. The circles with different colors are the final 100 particles, which are able to spread over all modes for SPOS.

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: let $\xi_k^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ for $\forall i$,

$$\begin{aligned}\theta_{k+1}^{(i)} =& \theta_k^{(i)} - h_k\beta^{-1}G_k^{(i)} - \frac{h_k}{M}\sum_{j=1}^{M} K(\theta_k^{(i)} - \theta_k^{(j)})G_k^{(j)} \\ &+ \frac{h_k}{M}\sum_{j=1}^{M} \nabla K(\theta_k^{(i)} - \theta_k^{(j)}) + \sqrt{2\beta^{-1}h_k}\xi_k^{(i)} \quad (9)\end{aligned}$$

We call the algorithm with particle update equations (9) stochastic particle-optimization sampling (Algorithm 1), in the sense that particles are optimized stochastically with extra random Gaussian noise. Intuitively, the added noise enhances the ability of the algorithm to jump out of local modes, leading to better exploration 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. Note there are always positive probabilities to jump between modes in this example. 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. Theorem 4 below bounds the EPD of SPOS, in contrast with that for SVGD in Theorem 2, which is intuitively obtained by taking the $\beta \to \infty$ limit.

Theorem 4 *Under Assumption 1, further assuming every $\{\boldsymbol{\theta}_\tau^{(i)}\}$ of (8) for approximating ν_τ in (4) has the same initial probability law ν_0 and $\Gamma \triangleq \mathbb{E}_{\boldsymbol{\theta}\sim\nu_0,\boldsymbol{\theta}'\sim\nu_0}[\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2] < \infty$. Choose a β such that $\lambda = \frac{m_F}{\beta} + m_K - H_F L_K > 0$. Then the EPD of SPOS is bounded as: EPD $\triangleq \sqrt{\sum_{i,j}^{M}\mathbb{E}\|\boldsymbol{\theta}_\tau^{(i)} - \boldsymbol{\theta}_\tau^{(j)}\|^2} \leq C_1 e^{-2\lambda\tau} + 4\sqrt{\frac{d}{\beta}}\frac{M}{\lambda}$, where $C_1 = M(M-1)\Gamma - 4\sqrt{d\beta^{-1}}\frac{M}{\lambda}$.*

Remark 2 *There are two interesting cases: i) When $C_1 > 0$, the EPD would decrease to the bound $4\sqrt{d\beta^{-1}}M/\lambda$ 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.*

4 Non-Asymptotic Convergence Analysis

In this section, we prove non-asymptotic convergence rates for the proposed SPOS algorithm under the 1-Wasserstein metric W_1, a special case of p-Wasserstein metric defined as

$$W_p(\mu,\nu) = \left(\inf_{\zeta\in\Gamma(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|X_\mu - X_\nu\|^p d\zeta(X_\mu, X_\nu)\right)^{1/p}$$

where $\Gamma(\mu,\nu)$ is the set of joint distributions on $\mathbb{R}^d\times\mathbb{R}^d$ with marginal distribution μ and ν. 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 mostly missing, despite the asymptotic theory developed recently [Liu, 2017, Lu et al., 2018]. For analysis, we further impose the following assumptions.

Assumption 3 *Assume F and ν_0 satisfy the following assumptions:*

3.1 There exists positive m_F such that $\langle F(\boldsymbol{\theta}) - F(\boldsymbol{\theta}'), \boldsymbol{\theta} - \boldsymbol{\theta}'\rangle \geq m_F\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2$.

3.2 The initial probability law of each particle has a bounded and strictly positive density ν_0 with respect to the Lebesgue measure on $\mathbb{R}^d$, and $\gamma_0 \triangleq \log\int_{\mathbb{R}^d} e^{\|\boldsymbol{\theta}\|^2}\nu_0(\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$

A few remarks: *i)* Assumption 3.1 indicates U to be convex. Theory of non-convex U is presented in Section J of the SM with some extra assumptions. *ii)* Assumptions 3.1 is widely adopted in the other theoretical works such as [Dalalyan and Karagulyan, 2017, Chatterji et al., 2018] *iii)* Assumptions 3.2 has also been adopted in [Raginsky et al., 2017]

4.1 Basic setup and extra notation

Due to the exchangeability of the particle system $\{\boldsymbol{\theta}_\tau^{(i)}\}_{i=1}^M$ in (8), if we initialize all the particles $\boldsymbol{\theta}_\tau^{(i)}$ with the same distribution ρ_0, they would endow the same distribution for each time τ. *We denote the distribution of each $\boldsymbol{\theta}_\tau^{(i)}$ as ρ_τ.* Similar arguments hold for the particle system $\{\theta_k^{(i)}\}_{i=1}^M$ in (9), and thus we denote the distribution of each $\theta_k^{(i)}$ as μ_k $(k = 1, 2, \cdots, T)$. To this end, our analysis aims at bounding $W_1(\mu_T, \nu_\infty)$ since ν_∞ is our target distribution $p(\boldsymbol{\theta}|\mathcal{X})$ according to Proposition 1.

In the following, for conciseness, we use a summation of stepsizes to represent the "time index" of some density, *e.g.*, $\rho_{\sum_{k=0}^{T-1} h_k}$. The high-level idea of bounding $W_1(\mu_T, \nu_\infty)$ in this section is to decompose it as follows:

$$W_1(\mu_T, \nu_\infty) \le W_1\left(\mu_T, \rho_{\sum\limits_{k=0}^{T-1} h_k}\right) \tag{10}$$
$$+ W_1\left(\rho_{\sum\limits_{k=0}^{T-1} h_k}, \nu_{\sum\limits_{k=0}^{T-1} h_k}\right) + W_1\left(\nu_{\sum\limits_{k=0}^{T-1} h_k}, \nu_\infty\right) .$$

4.2 Bounds with stochastic particle approximation

In this section, we bound $W_1(\rho_{\sum_{k=0}^{T-1} h_k}, \nu_{\sum_{k=0}^{T-1} h_k})$ and $W_1(\nu_{\sum_{k=0}^{T-1} h_k}, \nu_\infty)$ in (10). The first term corresponds to a variant of granular media equation, but is much more challenging to bound.

Theorem 5 *Under Assumption 1&3 and letting $\rho_0 = \nu_0$, there exist positive constants c_1 and c_2 independent of (M, τ) and satisfying $c_2 < \beta^{-1}$ such that*

$$W_1(\rho_\tau, \nu_\tau) \le c_1(\beta^{-1} - c_2)^{-1} M^{-1/2}, \qquad \forall \tau. \tag{11}$$

Remark 3 *According to Theorem 5, we can bound the $W_1(\rho_{\sum_{k=0}^{T-1} h_k}, \nu_{\sum_{k=0}^{T-1} h_k})$ term as $W_1(\rho_{\sum_{k=0}^{T-1} h_k}, \nu_{\sum_{k=0}^{T-1} h_k}) \le \frac{c_1}{\sqrt{M}(\beta^{-1} - c_2)}$. Furthermore, by letting $\tau \to \infty$, we have $W_1(\rho_\infty, \nu_\infty) \le \frac{c_1}{\sqrt{M}(\beta^{-1} - c_2)}$, an important result to prove the following theorem.*

Theorem 6 *Under Assumption 1&3, the following holds: $W_1(\nu_\tau, \nu_\infty) \le c_3 e^{-2\lambda_1 \tau}$, where $\lambda_1 = \beta^{-1} m_F - L_F - 2L_K$ and c_3 is some positive constant independent of (M, τ). Furthermore, the $W_1(\nu_{\sum_{k=0}^{T} h_k}, \nu_\infty)$ term in (10) can be bounded as:*

$$W_1(\nu_{\sum_{k=0}^{T-1} h_k}, \nu_\infty) \le c_3 \exp\left(-2\lambda_1 (\sum_{k=0}^{T-1} h_k)\right) . \tag{12}$$

To ensure $W_1(\nu_{\sum_{k=0}^{T-1} h_k}, \nu_\infty)$ decreases over time, one needs to choose β small enough such that $\lambda_1 > 0$. This also sheds light on a failure case of SVGD (where $\beta \to \infty$) discussed in Section 3.1.

4.3 Bounds with a numerical solution

To bound the $W_1(\mu_T, \rho_{\sum_{k=0}^{T-1} h_k})$ term in (10), we adopt techniques from [Raginsky et al., 2017, Xu et al., 2018] on analyzing the behavior of SGLD, and derive the following results for our SPOS algorithm:

Theorem 7 *Under Assumptions 1&3, for a fixed step size $h_k = h$ $(\forall k)$ that is small enough, the corresponding $W_1(\mu_T, \rho_{Th})$ is bounded as:*

$$W_1(\mu_T, \rho_{Th}) \le c_4 M d^{\frac{3}{2}} \beta^{-3} (c_5 \beta^2 B^{-1} + c_6 h)^{\frac{1}{2}} T^{\frac{1}{2}} h^{\frac{1}{2}} \tag{13}$$

where B is the minibatch size and (c_4, c_5, c_6) are some positive constants independent of (M, T, h).

Combining bounds from Theorems 5 and (7), given T, the optimal bound over h can be seen to decrease at a rate of $O(M^{-1/2})$. Furthermore, the dependence of T in the bound of Theorem 7 makes the bound relatively loose. Fortunately, the bound can be made independent of T by considering a decreasing-stepsize SPOS algorithm, stated in Theorem 8.

Theorem 8 *Under Assumptions 1&3, for a decreasing step size $h_k = h_0/(k+1)$, and letting the minibatch size in each iteration k be $B_k = B_0 + [\log(k+1)]^{100/99}$ with B_0 the initial batch size, the corresponding $W_1(\mu_T, \rho_{\sum_{k=0}^{T-1} h_k})$ term is bounded, for some β small enough, as:*

$$W_1\left(\mu_T, \rho_{\sum\limits_{k=0}^{T-1} h_k}\right) \le c_4 \beta^{-3} M d^{\frac{3}{2}} \left(c_7 h_0^3 + c_8 \beta^3 h_0 / B_0 \right.$$
$$\left. + c_9 h_0^2 \beta^2\right)^{1/2} , \tag{14}$$

where (c_4, c_7, c_8, c_9) are positive constants independent of (M, T, h_0).

Note B_k increases at a very low speed, *e.g.*, only by 15 after 10^5 iterations, thus it does not affect algorithm efficiency. Consequently, $W_1(\mu_T, \rho_{\sum_{k=0}^{T-1} h_k})$ would approach zero when $h_0^{1/2} M \to 0$.

The Overall Non-Asymptotic Bounds By directly combining results from Theorems 5–8, one can easily bound the target $W_1(\mu_T, \nu_\infty)$, stated in Theorem 9 and Theorem 10.

Theorem 9 (Fixed Stepsize) *Under Assumption 1&3 and setting $h_k = h_0$, $B_k = B_0$, $W_1(\mu_T, \nu_\infty)$ is bounded as: $W_1(\mu_T, \nu_\infty) \le$*

$$\frac{c_1}{\sqrt{M}(\beta^{-1} - c_2)} + c_6 M d^{\frac{3}{2}} \beta^{-3} (c_4 \beta^2 B^{-1} + c_5 h)^{\frac{1}{2}} T^{\frac{1}{2}} h^{\frac{1}{2}}$$
$$+ c_3 \exp\left\{-2\left(\beta^{-1} m_F - L_F - 2L_K\right) Th\right\} , \tag{15}$$

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

Theorem 10 (Decreasing Stepsize) *Denote $\tilde{h}_T \triangleq \sum_{k=0}^{T-1} h_k$. Under Assumption 1&3, if we set $h_k =$*

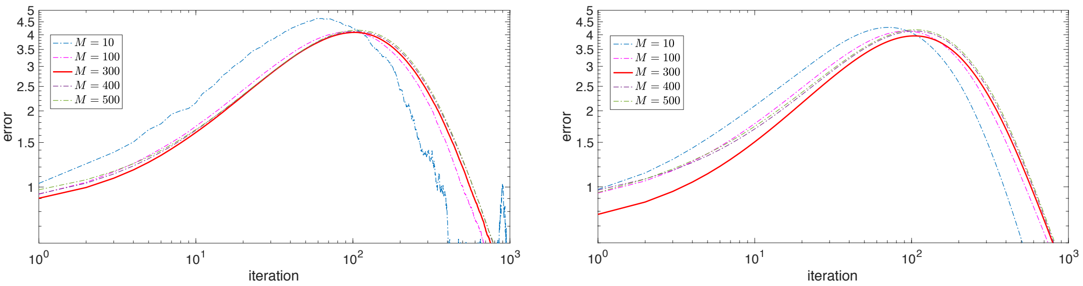

Figure 2: Estimation errors versus number of iterations for SPOS (left) and SVGD (right).

$h_0/(k{+}1)$ and $B_k = B_0 + [\log(k+1)]^{100/99}$, $W_1(\mu_T, \nu_\infty)$ is bounded as:

$$W_1(\mu_T, \nu_\infty) \leq \frac{c_1}{\sqrt{M}(\beta^{-1} - c_2)} \qquad (16)$$
$$+ c_3 \exp\{-2\left(\beta^{-1} m_F - L_F - 2L_K\right) \tilde{h}_T\}$$
$$+ c_{10} \beta^{-3} M d^{\frac{3}{2}} (c_9 h_0^3 + c_7 \beta^3 h_0 / B_0 + c_8 h_0^2 \beta^2)^{\frac{1}{2}} \ .$$

where $(c_1, c_2, c_3, c_7, c_7, c_8, c_9, c_10, \beta)$ are positive constants such that $\frac{1}{\beta} > c_2$ and $\frac{m_F}{\beta} > L_F + 2L_K$.

Remark 4 *Four implications are highlighted from the theorems: i) M and T play a similar role when bounding the numerical errors (the third term in the RHS of (15)). The bound increases with increasing M and T, which seems unavoidable and is consistent with the latest result for SGLD, whose bound is proved to increase w.r.t. T [Raginsky et al., 2017]. ii) The increasing bound w.r.t. T can be compromised by using decreasing stepsizes shown in Theorem 10. Unfortunately, this does not seem to eliminate the effect of M. To accommodate this, one should either use a smaller h or a larger β. We believe future work is needed to improve the bound w.r.t. M. However, this is nontrivial as recent theory shows coordinate-wise SGLD scales linearly w.r.t. parameter dimension [Shen et al., 2019] (corresponding to scaling linearly w.r.t. M in our case, consistent with our theory). iii) When $T \times M$ (proportional to computation cost) is not too large, the error is bounded above by $O(M^{-1/2} + M)$, indicating the existence of an optimal M, i.e., one should not choose arbitrary many particles as it would induce larger numerical-error bounds. This is somewhat surprising and counter-intuitive compared with the asymptotic theory [Liu, 2017, Lu et al., 2018]. However, we will demonstrate this is true with synthetic experiments below, where the phenomenon is also observed in SVGD. iv) When $T \times M$ is large enough, the $O(M)$ term dominates, indicating an increasing error w.r.t. M. This is verified by the experiments in Section 5.1 (Figure 3), although the bound might not be strictly tight.*

5 Experiments

We use a simple synthetic experiments to demonstrate the non-asymptotic convergence behaviors of SPOS indicated by our theory. For more experiments and real applications and comparisons of SPOS with SVGD and SGLD on Bayesian learning of deep neural network and Bayesian exploration in deep reinforcement learning (RL), please refer to Section O of the SM.

5.1 Sampling a Gaussian distribution

We apply the algorithms to sample from a simple 1-D Gaussian distribution with mean 2 and variance 1. Since the 1-Wasserstein distance is infeasible to calculate, we follow [Vollmer et al., 2016, Chen et al., 2015] and measure the convergence using err $\triangleq |\mathbb{E}_{\theta \sim \mu_T}[f(\theta)] - \mathbb{E}_{\theta \sim \mathcal{N}(2,1)}[f(\theta)]|$ with a test function $f(\theta) \triangleq \theta^2$. We fix $T = 1000$ and $h = 0.03$. Particles are initialized as being drawn from $\mathcal{N}(0, 1)$. Figure 2 plots the estimation errors versus the number of iterations for different particles M. For both SPOS and SVGD, it is observed that when T is not too large (≈ 100), the errors increase w.r.t. T, and the optimal M is around 300, consistent with our theory. When T is large enough, the errors decrease w.r.t. T, and larger M induces larger errors. This is also consistent with our theory because the last term in Theorem 9 dominates when T is large, leading to increasing errors with larger M. The only concern seems to be the tightness of the bound, which might be due to technical difficulty as current techniques for SGLD also indicate an increasing bound w.r.t. T [Raginsky et al., 2017]. The large optimal M also suggests using a relative large M should not be a problem in real applications.

Impact of particle number M In addition to the above result to demonstrate the existence of an optimal M, we further verify that when $T \times M$ is large enough, for a fixed T, we observe the errors increase with increasing M's. We use the same setting as above. Figure 3 plots the curves of errors versus number of particles. We see that errors indeed increase w.r.t. particle numbers, consistent with our theory. Although the rate of the bound from our theory might not match exactly with the experimental results, we believe this is still significant as the problem has never been discovered before, which is somewhat counter-intuitive. On the other hand, the results are also reasonable, as

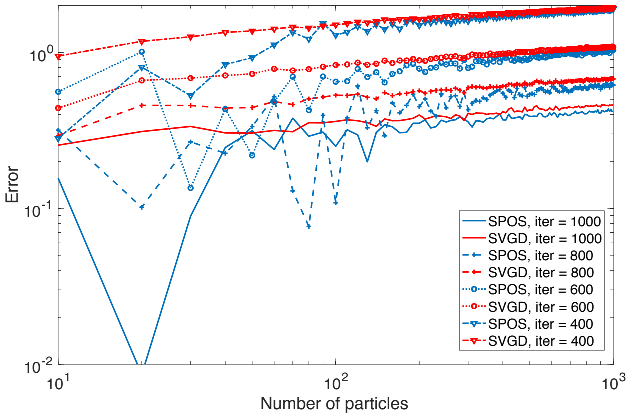

Figure 3: Errors versus Number of particles. Errors increase with increasing particle numbers.

more particles would need much more updates to fit a distribution well. The results indicate that to get a smaller error, one should increase number of iterations faster than increasing the number of particles.

Table 1: Averaged RMSE with standard deviations.

	Test RMSE		
Dataset	SGLD	SVGD	SPOS
Boston	3.114 ± 0.144	2.961 ± 0.109	**2.829 ± 0.126**
Concrete	5.508 ± 0.275	5.157 ± 0.082	**5.071 ± 0.150**
Energy	0.842 ± 0.060	1.291 ± 0.029	**0.752 ± 0.029**
Kin8nm	0.080 ± 0.001	0.090 ± 0.001	**0.079 ± 0.001**
Naval	0.004 ± 0.000	0.004 ± 0.000	**0.004 ± 0.000**
CCPP	4.059 ± 0.080	4.127 ± 0.027	**3.939 ± 0.049**
Wine	0.632 ± 0.022	0.604 ± 0.007	**0.598 ± 0.014**
Yacht	1.183 ± 0.263	1.597 ± 0.099	**0.840 ± 0.087**
Protein	4.281 ± 0.011	4.392 ± 0.015	**4.254 ± 0.005**
YearPredict	8.707 ± NA	8.684 ± NA	**8.681 ± NA**

5.2 BNNs for regression

We next conduct experiments for Bayesian learning of deep neural networks (DNNs) to empirically compare SGLD, SVGD and SPOS for posterior sampling of BNN weights with standard Gaussian priors. We use a RBF kernel with the bandwidth set to the medium of particles. 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. We use the same setting as [Zhang et al., 2018b]. The datasets are randomly split into 90% training and 10% testing. For a fair comparison, we use the same split of data (train, val and test) for all methods. We report the root mean squared error (RMSE) in Table 1. The proposed SPOS outperforms both SVGD and SGLD. More detailed settings and results are given in Section O of the SM.

5.3 Bayesian exploration in deep RL

It is well-accepted that RL performance directly measures how well the uncertainty is learned, due to the need for exploration. We apply SPOS for RL, and compare it with SVPG, a SVGD version of the policy gradient method [Liu et al., 2017]. Following [Liu et al., 2017, Zhang et al., 2018a], we define policies with Bayesian DNNs. This naturally introduces uncertainty into action selections, rendering Bayesian explorations to make policy learning more effective.

We follow the same setting as in [Liu et al., 2017] except using simpler policy-network architectures as in [Houthooft et al., 2016]. We conduct experiments on three classical continuous control tasks are considered: Cartpole Swing-Up, Double Pendulum, and Cartpole. Detailed experimental settings are given in the SM. Figure 4 plots the cumulative rewards over time on the Cartpole environment, which clearly shows the advantage of our method over SVPG. More results are provided in the SM.

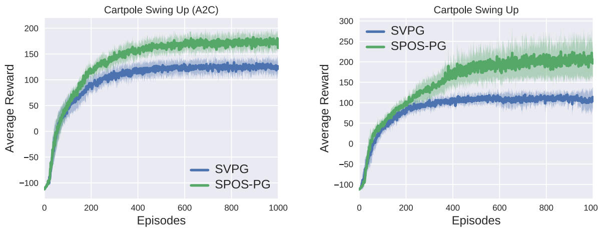

Figure 4: Policy learning with Bayesian exploration in policy-gradient methods with SVPG and SPOS-PG.

6 Conclusion

We propose an approach for particle-optimization-based sampling that overcomes a potential pitfall of standard SVGD. Notably, for the first time, 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 [Wang et al., 2017, Zhuo et al., 2018], is formally analyzed. Our theory is practically significant as it provides nonasymptotic theoretical guarantees for the recently proposed particle-optimization-based algorithms such as the SVGD, whose advantages have also been extensively examined in real applications. Surprisingly, our theory indicates the existence of an optimal particle size, *i.e.*, increasing particle size does not necessarily guarantee performance improvement. This is also observed for SVGD in a synthetic experiment. There are a number of interesting future works. For example, one might explore more recently developed techniques such as [Cheng et al., 2018, Liu and Wang, 2018] to improve the convergence bound; one can also adopt the SPOS framework for non-convex optimization like where SG-MCMC is used, and develop corresponding theory to study the convergence properties of the algorithm to the global optimum.

References

[Blundell et al., 2015] Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural networks. In *ICML*.

[Bolley and Villani, 2005] Bolley, F. and Villani, C. (2005). Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. *Annales de la Facultédes sciences de Toulouse : Mathématiques*, 14(6):331–352.

[Carrillo et al., 2017] Carrillo, J. A., Craig, K., and Patacchini, F. S. (2017). A blob method for diffusion. (arXiv:1709.09195).

[Cattiaux et al., 2008] Cattiaux, P., Guillin, A., and Malrieu, F. (2008). Probabilistic approach for granular media equations in the non-uniformly convex case. *Probability Theory and Related Fields*, 140(1–2):19–40.

[Chatterji et al., 2018] Chatterji, N. S., Flammarion, N., Ma, Y.-A., Bartlett, P. L., and Jordan, M. I. (2018). On the theory of variance reduction for stochastic gradient monte carlo.

[Chen et al., 2015] Chen, C., Ding, N., and Carin, L. (2015). On the convergence of stochastic gradient MCMC algorithms with high-order integrators. In *Neural Information Processing Systems (NIPS)*.

[Chen et al., 2018] Chen, C., Zhang, R., Wang, W., Li, B., and Chen, L. (2018). A unified particle-optimization framework for scalable Bayesian sampling. In *UAI*.

[Chen et al., 2014] Chen, T., Fox, E. B., and Guestrin, C. (2014). Stochastic gradient Hamiltonian Monte Carlo. In *International Conference on Machine Learning (ICML)*.

[Cheng et al., 2018] Cheng, X., Chatterji, N. S., Abbasi-Yadkori, Y., Bartlett, P. L., and Jordan, M. I. (2018). Sharp convergence rates for Langevin dynamics in the nonconvex setting. In *arXiv:1805.01648*.

[Chiang and Hwang, 1987] Chiang, T.-S. and Hwang, C.-R. (1987). Diffusion for global optimization in rn. *SIAM J. Control Optim.*, 25(3):737–753.

[Şimşekli et al., 2018] Şimşekli, U., Liutkus, A., Majewski, S., and Durmus, A. (2018). Sliced-Wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. Technical Report arXiv:1806.08141.

[Dalalyan and Karagulyan, 2017] Dalalyan, A. and Karagulyan, A. (2017). User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *Stochastic Processes and their Applications*.

[Ding et al., 2014] Ding, N., Fang, Y., Babbush, R., Chen, C., Skeel, R. D., and Neven, H. (2014). Bayesian sampling using stochastic gradient thermostats. In *Neural Information Processing Systems (NIPS)*.

[Durmus et al., 2018] Durmus, A., Eberle, A., Guillin, A., and Zimmer, R. (2018). An Elementary Approach To Uniform In Time Propagation Of Chaos. *ArXiv e-prints*.

[Durmus et al., 2018] Durmus, A., Eberle, A., Guillin, A., and Zimmer, R. (2018). An elementary approach to uniform in time propagation of chaos. In *arXiv:1805.11387*.

[Feng et al., 2017] Feng, Y., Wang, D., and Liu, Q. (2017). Learning to draw samples with amortized stein variational gradient descent. In *UAI*.

[Givens and Shortt, 1984] Givens, C. R. and Shortt, R. M. (1984). A class of wasserstein metrics for probability distributions. *Michigan Math. J.*, 31.

[Haarnoja et al., 2017] Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. (2017). Reinforcement learning with deep energy-based policies. In *ICML*.

[Hernández-Lobato and Adams, 2015] Hernández-Lobato, J. M. and Adams, R. P. (2015). Probabilistic backpropagation for scalable learning of Bayesian neural networks. In *ICML*.

[Houthooft et al., 2016] Houthooft, R., Chen, X., Duan, Y., Schulman, J., De Turck, F., and Abbeel, P. (2016). VIME: Variational information maximizing exploration. In *NIPS*.

[Li et al., 2016] Li, C., Chen, C., Carlson, D., and Carin, L. (2016). Preconditioned stochastic gradient Langevin dynamics for deep neural networks. In *AAAI*.

[Li et al., 2015] Li, Y., Hernández-Lobato, J., and Turner, R. E. (2015). Stochastic expectation propagation. In *NIPS*.

[Liu and Zhu, 2018] Liu, C. and Zhu, J. (2018). Riemannian Stein variational gradient descent for Bayesian inference. In *AAAI*.

[Liu et al., 2019] Liu, C., Zhuo, J., Cheng, P., Zhang, R., Zhu, J., and Carin, L. (2019). Understanding and accelerating particle-based variational inference. In *ICML*.

[Liu, 2017] Liu, Q. (2017). Stein variational gradient descent as gradient flow. In *NIPS*.

[Liu and Wang, 2016] Liu, Q. and Wang, D. (2016). Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Neural Information Processing Systems (NIPS)*.

[Liu and Wang, 2018] Liu, Q. and Wang, D. (2018). Stein variational gradient descent as moment matching. In *NIPS*.

[Liu et al., 2017] Liu, Y., Ramachandran, P., Liu, Q., and Peng, J. (2017). Stein variational policy gradient. In *UAI*.

[Louizos and Welling, 2016] Louizos, C. and Welling, M. (2016). Structured and efficient variational deep learning with matrix Gaussian posteriors. In *ICML*.

[Lu et al., 2018] Lu, J., Lu, Y., and Nolen, J. (2018). Scaling limit of the Stein variational gradient descent part I: the mean field regime. In *arXiv:1805.04035*.

[Ma et al., 2015] Ma, Y. A., Chen, T., and Fox, E. (2015). A complete recipe for stochastic gradient MCMC. In *NIPS*.

[Malrieu, 2003] Malrieu, F. (2003). Convergence to equilibrium granular media equations and their euler schemes. *The Annnals of Applied Probability*, 13(2):540–560.

[Mattingly et al., 2002] Mattingly, J. C., Stuartb, A. M., and Higham, D. J. (2002). Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic Processes and their Applications*, 101(2):185–232.

[Raginsky et al., 2017] Raginsky, M., Rakhlin, A., and Telgarsky, M. (2017). Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *COLT*.

[Rezende and Mohamed, 2015] Rezende, D. J. and Mohamed, S. (2015). Variational inference with normalizing flows. In *ICML*.

[Risken, 1989] Risken, H. (1989). *The Fokker-Planck equation*. Springer-Verlag, New York.

[Schulman et al., 2015] Schulman, J., Moritz, P., Levine, S., Jordan, M., and Abbeel, P. (2015). High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*.

[Shen et al., 2019] Shen, L., Balasubramanian, K., and Ghadimi, S. (2019). Non-asymptotic results for Langevin monte carlo: Coordinate-wise and black-box sampling. In *arXiv:1902.01373*.

[Teh et al., 2016] Teh, Y. W., Thiery, A. H., and Vollmer, S. J. (2016). Consistency and fluctuations for stochastic gradient Langevin dynamics. *JMLR*, 17(1):193–225.

[Villani, 2008] Villani, C. (2008). *Optimal transport: old and new*. Springer Science & Business Media.

[Vollmer et al., 2016] Vollmer, S. J., Zygalakis, K. C., and Teh, Y. W. (2016). (exploration of the (Non-)asymptotic bias and variance of stochastic gradient Langevin dynamics. *JMLR*, 1:1–48.

[Wang et al., 2017] Wang, D., Zeng, Z., and Liu, Q. (2017). Stein variational message passing for continuous graphical models. *arXiv preprint arXiv:1711.07168*.

[Welling and Teh, 2011] Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *ICML*.

[Williams, 1992] Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*.

[Xu et al., 2018] Xu, P., Chen, J., Zou, D., and Gu, Q. (2018). Global convergence of Langevin dynamics based algorithms for nonconvex optimization. In *NIPS*.

[Zhang et al., 2018a] Zhang, R., Chen, C., Li, C., and Carin, L. (2018a). Policy optimization as wasserstein gradient flows. In *ICML*.

[Zhang et al., 2018b] Zhang, R., Li, C., Chen, C., and Carin, L. (2018b). Learning structural weight uncertainty for sequential decision-making. In *AISTATS*.

[Zhang et al., 2019] Zhang, R., Wen, Z., Chen, C., and Carin, L. (2019). Scalable thompson sampling via optimal transport. In *AISTATS*.

[Zhang et al., 2017] Zhang, Y., Liang, P., and Charikar, M. (2017). A hitting time analysis of stochastic gradient Langevin dynamics. In *COLT*.

[Zhuo et al., 2018] Zhuo, J., Liu, C., Shi, J., Zhu, J., Chen, N., and Zhang, B. (2018). Message passing stein variational gradient descent. In *ICML*.