# SHPOS: A Theoretical Guaranteed Accelerated Particle Optimization Sampling Method

**Zhijian Li**[1] , **Chao Zhang**[2,3*] , **Hui Qian**[2,3] , **Xin Du**[1†] and **Lingwei Peng**[2]
[1]Information Science and Electronic Engineering, Zhejiang University
[2]College of Computer Science and Technology, Zhejiang University
[3]Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies
{lizhijian, zczju, qianhui, duxin, penglingwei}@zju.edu.cn

## Abstract

Recently, the Stochastic Particle Optimization Sampling (SPOS) method is proposed to solve the particle-collapsing pitfall of deterministic Particle Variational Inference methods by ultilizing the stochastic Overdamped Langevin dynamics to enhance exploration. In this paper, we propose an accelerated particle optimization sampling method called Stochastic Hamiltonian Particle Optimization Sampling (SHPOS). Compared to the first-order dynamics used in SPOS, SHPOS adopts an augmented second-order dynamics, which involves an extra momentum term to achieve acceleration. We establish a non-asymptotic convergence analysis for SHPOS, and show that it enjoys a faster convergence rate than SPOS. Besides, we also propose a variance-reduced stochastic gradient variant of SHPOS for tasks with large-scale datasets and complex models. Experiments on both synthetic and real data validate our theory and demonstrate the superiority of SHPOS over the state-of-the-art.

## 1 Introduction

Sampling from a Bayesian posterior distribution lies at the core of many modern machine learning tasks, such as topic modelling [Gan *et al.*, 2015], reinforcement learning [Liu *et al.*, 2017], and Bayesian neural networks [Hernández-Lobato and Adams, 2015]. Particle based Variational Inference (ParVI) methods have recently drawn great attention due to their empirical success in approximating the target posterior distribution [Liu and Wang, 2016; Liu *et al.*, 2017; Feng *et al.*, 2017; Chen *et al.*, 2018; Liu and Zhu, 2018]. Typically, these methods update a finite set of interacting particles *deterministically* to approximately simulate infinite-particle gradient flows on the Wasserstein space $P_2(\mathcal{X})$.

One representative method of this type is the Stein Variational Gradient Descent (SVGD) method [Liu and Wang, 2016], which updates the particles according to a gradient flow described by the Vlasov equation [Liu, 2017; Braun and Hepp, 1977]. Subsequently, by exploiting the Riemannian structure of the Wasserstein space $P_2(\mathcal{X})$, [Liu *et al.*, 2019] proposed a Nesterov's-acceleration variant of SVGD called SVGD Wasserstein Nesterov's method (SVGD-WNes).

However, ParVI methods have an intractable pitfall that particles tend to collapse under certain condition due to the deterministic-update fashion with a limited number of particles [Zhang *et al.*, 2020a; Zhuo *et al.*, 2018], which indicates a large deviation of the particles' empirical distribution to target distribution. To understand the influence of finite-particle approximation to the infinite-particle gradient flows, Liu *et al.* [2019] provided a unified theory on the approximation property of different ParVI methods. They show that existing finite-particle ParVIs can be regarded as essentially smoothing operations on gradient flows, in the form of either smoothing the density or smoothing functions. Though the underlying gradient flows usually evolve towards the target distribution, the smoothing operations deteriorate the convergence of the particles' empirical distributions in ParVIs to the target distribution.

Inspired by the random exploration in the dynamics-based Markov Chain Monte Carlo methods, recent researches relieve the particle-collapsing phenomenon by introducing additional stochasticity into ParVI methods [Zhang *et al.*, 2020a; Zhang *et al.*, 2020b]. Zhang *et al.* [2020a] integrated the gradient flow of the first-order Overdamped Langevin Dynamics (OLD) into the Vlasov equation used in SVGD, and proposed a new method called Stochastic Particle Optimization Sampling (SPOS). Specifically, SPOS updates a set of particles following the same mechanism as in SVGD with an additional drift term and an extra Gaussian random noise induced by OLD. Following the analysis framework of OLD based MCMC methods, they conduct the non-asymptotic convergence analysis of SPOS, and establish a bound on the 2-Wasserstein distance between the empirical distribution of particles and the target distribution. However, observations in the dynamics-based MCMC indicate that OLD is not the most efficient dynamics although it is simple and easy-to-analyze. Actually, the error bound in SPOS increases with the iterations, therefore the step size should be samll in order to restrict the error in a desirable level.

In this work, we propose an accelerated particle optimization sampling method called Stochastic Hamiltonian Particle Optimization Sampling (SHPOS) by introducing the second-order Underdamped Langevin Dynamics (ULD) [Jaakkola

[*] Zhijian Li and Chao Zhang contribute equally.
[†] Contact Author.

and Jordan, 1997] into the flow of SVGD. Compared with the first-order OLD, ULD possesses an auxiliary momentum term, and can be regarded as an accelerated second-order dynamics of OLD. Typically, ULD based methods have both better theoretical guarantees and more competitive practical performance than their OLD based counterparts [Chen *et al.*, 2014; Cheng *et al.*, 2018; Zou *et al.*, 2018; Zou *et al.*, 2019]. We utilize a variant of the first order exponential integrator scheme to discretize the corresponding stochastic differential equation of the augmented flow, which is obtained by integrating the gradient flow of ULD into the Vlasov flow. The contributions of our paper are listed as follows:

- As far as we know, SHPOS firstly introduce second-order ULD into the Vlasov flow to accelerate particle optimization sampling. The particles in SHPOS can be regarded as drifted by the (stochastic-)gradient of the log-posterior of the target distribution, the repulsive force between particles, and an extra random standard Gaussian force. By balancing these forces, particles explore the high probability space of the target distribution rapidly and avoid particle-collapsing (refer to the experiment for more details). We also propose a variance-reduced stochastic gradient variant of SHPOS by using the historical gradient information to build a recursively updated gradient estimator. This variant is shown to be more suitable for Bayesian learning tasks with large-scale datasets and complex models.
- We establish the non-asymptotic convergence guarantee of SHPOS. Specifically, we show that the 2-Wasserstein distance between the particles' distribution in the $k$-th iteration of SHPOS and the target distribution is in the order of $\tilde{\mathcal{O}}(M^{-1/2} + \exp(-k\eta) + \eta d^{1/2})$, where $M$ is the number of particles, $d$ denotes the dimension of variable, and $\eta$ is the step size. The result shows that the approximation error decreases as the particle number $M$ and the iteration $k$ increase, while the error bound $\tilde{\mathcal{O}}(M^{-1/2} + \exp(-k\eta) + M\eta d^{3/2}k^{1/2})$ of SPOS increases with particle numbers and iterations. Thus, SHPOS could use a much larger step size compared with SPOS to achieve the same sampling accuracy.

We evaluate our method on a list of tasks, including both synthetic and real datasets. The empirical results demonstrate the superiority of our method over the state-of-the-art.

## 2 Preliminaries

In this paper, we focus on Bayesian learning tasks, i.e., sampling from the target posterior distribution

$$p^* = p(\mathbf{x}|\mathcal{D}) \propto \exp(-U(\mathbf{x})),$$

where $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$ is the model parameter, $\mathcal{D} = \{D_i\}_{i=1}^N$ denotes the dataset and $U(\mathbf{x}) = -\log p(\mathbf{x}) - \log p(D|\mathbf{x})$ denotes the potential energy function. SVGD and ULD based MCMC methods are two recently proposed inference methods for solving Bayesian learning tasks.

### 2.1 Stein Variational Gradient Descent

SVGD is firstly proposed by [Liu and Wang, 2016] as an efficient approximate inference method for sampling from the target distribution. Basically, it simulates the following density gradient flow

$$\frac{\partial \varphi_t}{\partial t} = \nabla_{\mathbf{x}} \cdot (\varphi_t \boldsymbol{\phi}_{\varphi_t}), \tag{1}$$

where $\varphi_t$ denotes the evolutionary density at time $t$ and $\boldsymbol{\phi}_{\varphi_t}$ is the evolution function of the following form

$$\boldsymbol{\phi}_{\varphi_t}(\mathbf{x}) = \mathbb{E}_{x' \sim \varphi_t}\left[-\kappa(\mathbf{x}, \mathbf{x}')F(\mathbf{x}') + \nabla_{\mathbf{x}'}\kappa(\mathbf{x}, \mathbf{x}')\right], \tag{2}$$

with $F(\mathbf{x}) = \nabla U(\mathbf{x})$ denoting the gradient of the potential function and $\kappa(\mathbf{x}, \mathbf{x}')$ denoting a kernel function, such as RBF kernel. Actually, $\boldsymbol{\phi}_{\varphi_t}$ is a direction of steepest descent that minimize the KL divergence between the evolutionary density $\varphi_t$ and the target distribution $p^*$ over the unit ball in Reproducing Kernel Hilbert Space induced by the kernel function $\kappa$ [Liu and Wang, 2016]. The density gradient flow (1) is a special type of Vlasov equation and its stationary distribution is $p^*$ under certain regularity conditions [Liu, 2017].

SVGD draws a set of particles $\{\mathbf{x}_{i,0}\}_{i=1}^M$ from an initial distribution $q_0$, and updates the particles iteratively according to the following rule:

$$\mathbf{x}_{i,k+1} = \mathbf{x}_{i,k} - \frac{\eta}{M}\sum_{j=1}^{M} F(\mathbf{x}_{j,k})\kappa(\mathbf{x}_{i,k}, \mathbf{x}_{j,k}) - \nabla_{\mathbf{x}_{j,k}}\kappa(\mathbf{x}_{i,k}, \mathbf{x}_{j,k}),$$

where $\eta$ denotes the step size. On the RHS of the update rule, the $-F(\mathbf{x}_{j,k})\kappa(\mathbf{x}_{i,k}, \mathbf{x}_{j,k})$ term pushes particles towards high probability regions, while the second term $\nabla_{\mathbf{x}_{j,k}}\kappa(\mathbf{x}_{i,k}, \mathbf{x}_{j,k})$ forces the particles to move away from each other. The stochastic gradient technique, i.e., constructing a stochastic approximation from a mini-batch of data to replace the full gradient, has been adopted by SVGD to reduce the per-iteration computational cost in large-scale and complex Bayesian learning tasks. Moreover, SVGD uses empirical distribution of the particles to approximate target distribution. However, Zhang *et al.*[2020a] pointed out that SVGD has an unintended particle-collapsing pitfall, i.e., particles tend to collapse to a local mode under particular conditions.

### 2.2 Dynamics-Based MCMC

Recently, different dynamics-based MCMC methods have been proposed by discretizing certain dynamics whose stationary distribution (or its marginal) is the target distribution. Two mostly adopted dynamics are the Overdamped Langevin Dynamics (OLD) and the Underdamped Langevin Dynamics (ULD). OLD is a first-order dynamics with only one variable, while ULD involves an extra momentum term, thus can be viewed as a second-order dynamics. Basically, ULD based methods usually outperform their OLD based counterparts in both practice and theory, due to the extra momentum term. However, their analysis are much more complicated and elaborate as their structures are more complex than their OLD based counterparts.

The Underdamped Langevin Dynamics is described by the following stochastic differential equation (SDE):

$$\begin{aligned} \mathrm{d}\mathbf{x}_t &= \mathbf{v}_t \mathrm{d}t, \\ \mathrm{d}\mathbf{v}_t &= -\gamma\mathbf{v}_t\mathrm{d}t - uF(\mathbf{x}_t)\mathrm{d}t + \sqrt{2u\gamma}\,\mathrm{d}\mathbf{B}_t, \end{aligned} \tag{3}$$

| Method | SVGD | SPOS | UL-MCMC | SVGD-WNes | SHPOS |
|---|---|---|---|---|---|
| momentum acceleration | × | × | ✓ | ✓ | ✓ |
| repulsive force | ✓ | ✓ | × | ✓ | ✓ |
| random exploration | × | ✓ | ✓ | × | ✓ |

Table 1: Feature-by-feature comparison of different methods.

where $\gamma > 0$ is the friction parameter, $u$ is the inverse mass, $\mathbf{x}_t, \mathbf{v}_t \in \mathbb{R}^d$ denote the position and velocity variables of the dynamics respectively, and $\mathbf{B}_t$ represents standard Brownian motion. Let $\mathbf{z}_t^T = \left[\mathbf{x}_t^T, \mathbf{v}_t^T\right]$, according to the Fokker-Planck (FP) equation [Risken, 1996], the density gradient flow of $\mathbf{z}$ is as follow:

$$\frac{\partial \Psi_t}{\partial t} = -\nabla_{\mathbf{z}} \cdot \left( \Psi_t \begin{pmatrix} \mathbf{v_t} \\ -uF(\mathbf{x}_t) - \gamma \mathbf{v}_t \end{pmatrix} \right) + u\gamma \nabla_{\mathbf{v}} \cdot (\nabla_{\mathbf{v}} \Psi_t), \quad (4)$$

where $\Psi_t$ denotes the density of $\mathbf{z}_t$. The stationary distribution of variable $\mathbf{z}_t$ is $p(\mathbf{z}) \propto \exp\left(-U(\mathbf{x}) - \|\mathbf{v}\|_2^2/(2u)\right)$, and the marginal distribution of $\mathbf{x}_t$ converges to the target distribution $p^*$.

Directly using the Euler-Maruyama discretization scheme of ULD (3) gives rise to the Underdamped Langevin MCMC (UL-MCMC) method [Kloeden and Platen, 1992]:

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{x}_k + \eta \mathbf{v}_k, \\ \mathbf{v}_{k+1} &= \mathbf{v}_k - \gamma\eta\mathbf{v}_k - u\eta F(\mathbf{x}_k) + \sqrt{2u\gamma\eta}\boldsymbol{\epsilon}_k, \end{aligned} \quad (5)$$

where $\eta > 0$ is the step size, $\boldsymbol{\epsilon}_k \sim N(\mathbf{0}, \mathbf{I}_{d\times d})$ is a standard Gaussian random vector, and $\mathbf{I}_{d\times d}$ is a $d \times d$ identity matrix. To make the algorithm scalable in large-scale tasks, SGHMC method [Chen *et al.*, 2014] is proposed by replacing the full gradient with a mini-batch estimator

$$G_k(\mathbf{x}_k) = \frac{1}{|B|} \sum_{\xi_k \in B_k} F(\mathbf{x}_k, \xi_k), \quad (6)$$

where $B_k$ denotes a mini-batch of size $|B|$, and $F(\mathbf{x}_k, \xi_k)$ is an unbiased estimate of $F(\mathbf{x}_k)$. Ever since, different variance-reduced stochastic gradient ULD MCMC methods has been proposed to reduce the influence of stochastic variance, e.g. SVR-HMC [Zou *et al.*, 2018], SAGA-HMC [Li *et al.*, 2019] and HSG-HMC [Zhang *et al.*, 2021].

## 3 Methodology

In this section, we introduce our Stochastic Hamiltonian Particle Optimization Sampling method as in Algorithm 1. We first construct the following mixed density gradient flow by integrating the gradient flow of ULD into the Vlasov flow:

$$\frac{\partial \Phi_t}{\partial t} = -\nabla_{\mathbf{z}} \cdot \begin{pmatrix} \Phi_t \mathbf{v}_t \\ \Phi_t \left(-uF(\mathbf{x}_t) - \gamma\mathbf{v}_t + \beta \boldsymbol{\phi}_{\varphi_t}\right) \end{pmatrix} + u\gamma\nabla_{\mathbf{v}} \cdot (\nabla_{\mathbf{v}} \Phi_t), \quad (7)$$

where $\mathbf{z}_t^T = \left[\mathbf{x}_t^T, \mathbf{v}_t^T\right]$, $\Phi_t$ denotes the density of $\mathbf{z}_t$, $\varphi_t$ denotes the marginal distribution of variable $\mathbf{x}_t$, and $\boldsymbol{\phi}_{\varphi_t}$ is defined as in (2).

In the following lemma, we give the corresponding SDE of flow (7) and show that the marginal distribution $\varphi_t$ of variable $\mathbf{x}_t$ converges to the target distribution $p^*$.

**Lemma 1.** *The density flow of the following SDE is exact* (7)*.*

$$\begin{aligned} \mathrm{d}\mathbf{x} =& \mathbf{v}\mathrm{d}t, \\ \mathrm{d}\mathbf{v} =& -\gamma\mathbf{v}\mathrm{d}t - uF(\mathbf{x})\mathrm{d}t - \beta\mathbb{E}_{Y\sim\varphi_t}\left[F(Y)\kappa(\mathbf{x}, Y)\right]\mathrm{d}t \\ &+ \beta\mathbb{E}_{Y\sim\varphi_t}\left[\nabla_Y \kappa(\mathbf{x}, Y)\right]\mathrm{d}t + \sqrt{2u\gamma}\mathrm{d}\mathbf{B}_t, \end{aligned} \quad (8)$$

*where $Y \in \mathbb{R}^d$ is a random sample from evolutionary density $\varphi_t$ but independent of $\mathbf{x}$ and $\mathbf{v}$, and $\mathbf{B}_t$ represents standard Brownian motion. The stationary distribution of this SDE is $p(\mathbf{z}) \propto \exp\left(-U(\mathbf{x}) - \|\mathbf{v}\|_2^2/(2u)\right)$, and its marginal distribution $\varphi_t$ converges to the target distribution $p^*$.*

**Algorithm 1** Stochastic Hamiltonian Particle Optimization Sampling

```
Input: Initial particles $\{\mathbf{x}_{i,0}\}_{i=1}^M$ and the corresponding momentum variable $\{\mathbf{v}_{i,0}\}_{i=1}^M$. Weight parameter $\beta$, friction parameter $\gamma$, inverse mass $u$, and step size $\eta$.
for $k = 0, 1, ..., T-1$ do
    Uniformly sample a subset of index $B_k \subset [n]$
    for $i = 1, 2, ..., M$ do
        Update $\mathbf{x}_{i,k+1}$ and $\mathbf{v}_{i,k+1}$ as
            $\mathbf{x}_{i,k+1} = \mathbf{x}_{i,k} + \eta\mathbf{v}_{i,k} + \boldsymbol{\epsilon}_{i,k}^{\mathbf{x}},$   (9)
            $\mathbf{v}_{i,k+1} = (1-\gamma\eta)\mathbf{v}_{i,k} - u\eta g_{i,k} + \boldsymbol{\epsilon}_{i,k}^{\mathbf{v}} + \frac{\eta\beta}{M}$
                $\cdot \sum_j [\nabla K(\mathbf{x}_{i,k} - \mathbf{x}_{j,k}) - g_{j,k} K(\mathbf{x}_{i,k} - \mathbf{x}_{j,k})]$,   (10)
        with $g_{i,k}$ as the full gradient $F(\mathbf{x}_{i,k})$ (or its stochastic estimate $G_{i,k}^{\mathrm{mini}}$ (14) or $G_{i,k}^{\mathrm{vr}}$ (15)).
    end for
end for
Output: $\{\mathbf{x}_{i,T}\}_{i=1}^M$.
```

According to the system (8), $\mathbf{x}$ and $\mathbf{v}$ can be regarded as the position of a particle and its associated velocity. Besides, it can be verified that the Vlasov flow endows an repulsive force on the particle and $\beta$ controls the magnitude of particle interaction. As this system contains an extra Hamiltonian velocity part compared to the system used in SPOS, we call our method a Hamiltonian method. We list a feature-by-feature comparison of different sampling methods in Table 1 to give a more intuitive understanding of SHPOS method.

We construct our SHPOS method by simulating system (8). Note that (8) contains the marginal distribution $\varphi_t$ of $\mathbf{x}_t$, which is intractable to calculate. Following the idea in ParVI, we maintains a set of $M$ particles and use the empirical distribution to approximate $\varphi_t$ in SDE (8). Consequently, we obtain an multi-particle approximate dynamics of (8), where the dynamics of each particle is as follow:

$$\begin{aligned} \mathrm{d}\tilde{\mathbf{x}}_{i,t} =& \tilde{\mathbf{v}}_{i,t}\mathrm{d}t, \\ \mathrm{d}\tilde{\mathbf{v}}_{i,t} =& -\gamma\tilde{\mathbf{v}}_{i,t}\mathrm{d}t - uF(\tilde{\mathbf{x}}_{i,t})\mathrm{d}t + \frac{\beta}{M}\sum_{j=1}^{M}\left[-F(\tilde{\mathbf{x}}_{j,t})\right. \\ & \left. K(\tilde{\mathbf{x}}_{i,t} - \tilde{\mathbf{x}}_{j,t}) + \nabla K(\tilde{\mathbf{x}}_{i,t} - \tilde{\mathbf{x}}_{j,t})\right]\mathrm{d}t + \sqrt{2u\gamma}\mathrm{d}\tilde{\mathbf{B}}_{i,t}. \end{aligned} \quad (11)$$

Here, we focus on the RBF kernel $\kappa(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|_2^2/h)$, and replace $\kappa(\mathbf{x}, \mathbf{x}')$ with $K(\mathbf{x} - \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|_2^2/h)$. Note that this kernel is also adopted in SPOS and many ParVIs [Liu and Wang, 2016; Liu *et al.*, 2019].

In order to sample from the target distribution, we can apply numerical integrators to discretize the multi-particle dynamics (11). Specifically, we utilize a variant of the first order exponential integrator scheme [Zou *et al.*, 2018], which leads to the following updates for variables $\mathbf{x}_{i,k}$ and $\mathbf{v}_{i,k}$:

$$\begin{aligned}\mathbf{x}_{i,k+1} =& \mathbf{x}_{i,k} + \eta \mathbf{v}_{i,k} + \boldsymbol{\epsilon}^{\mathbf{x}}_{i,k},\\ \mathbf{v}_{i,k+1} =& (1-\gamma\eta)\mathbf{v}_{i,k} - u\eta F(\mathbf{x}_{i,k}) + \frac{\eta\beta}{M}\sum_{j=1}^{M}\\ & [-F(\mathbf{x}_{j,k})K(\mathbf{x}_{i,k}-\mathbf{x}_{j,k}) + \nabla K(\mathbf{x}_{i,k}-\mathbf{x}_{j,k})] + \boldsymbol{\epsilon}^{\mathbf{v}}_{i,k},\end{aligned} \tag{12}$$

where $\gamma$, $u$, $\eta$, $\beta$ are tunable parameters, $\boldsymbol{\epsilon}^{\mathbf{x}}_{i,k}$ and $\boldsymbol{\epsilon}^{\mathbf{v}}_{i,k}$ are Gaussian random vectors with zero mean and covariance matrices satisfy

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{\epsilon}^{\mathbf{v}}_{k}(\boldsymbol{\epsilon}^{\mathbf{v}}_{k})^{T}\right] =& u\left(1-e^{-2\gamma\eta}\right)\cdot \mathbf{I}_{d\times d},\\ \mathbb{E}\left[\boldsymbol{\epsilon}^{\mathbf{x}}_{k}(\boldsymbol{\epsilon}^{\mathbf{x}}_{k})^{T}\right] =& \frac{u}{\gamma^2}\left(2\gamma\eta + 4e^{-\gamma\eta} - e^{-2\gamma\eta} - 3\right)\cdot \mathbf{I}_{d\times d},\\ \mathbb{E}\left[\boldsymbol{\epsilon}^{\mathbf{v}}_{k}(\boldsymbol{\epsilon}^{\mathbf{x}}_{k})^{T}\right] =& \frac{u}{\gamma}\left(1-2e^{-\gamma\eta} + e^{-2\gamma\eta}\right)\cdot \mathbf{I}_{d\times d}.\end{aligned} \tag{13}$$

In many real-world Bayesian learning tasks with large-scale datasets and complex models, obtaining the exact gradient $F(\mathbf{x})$ is computationally expensive or even prohibitive, and only unbiased estimates $F(\mathbf{x},\xi)$ of $F(\mathbf{x})$ are available. To deal with this situation, one typical approach is to replace the full gradient with a mini-batch estimator

$$G^{\text{mini}}_{i,k} = \frac{1}{|B|}\sum_{\xi_k \in B_k} F(\mathbf{x}_{i,k},\xi_k). \tag{14}$$

However, this estimator usually suffers from high stochastic variance, which will degrade the performance of algorithms. To relieve the influence caused by stochastic variance, we propose to use the following estimator

$$\begin{aligned}G^{\text{vr}}_{i,k} =& (1-\rho_k)\underbrace{(G^{\text{vr}}_{i,k-1} + \tilde{F}_k(\mathbf{x}_{i,k}) - \tilde{F}_k(\mathbf{x}_{i,k-1}))}_{g_{b,k}} +\\ & \rho_k \underbrace{\tilde{F}_k(\mathbf{x}_{i,k})}_{g_{u,k}},\end{aligned} \tag{15}$$

where $\tilde{F}_k(\mathbf{x}) = \frac{1}{|B|}\sum_{\xi_k\in B_k} F(\mathbf{x},\xi_k)$. This estimator has been widely adopted in the optimization literature [Cutkosky and Orabona, 2019] and stochastic gradient MCMC methods [Zhang *et al.*, 2021] to relieve the influence of variance. Specifically, $G^{\text{vr}}_{i,k}$ is a combination of two parts, $g_{u,k}$ is an unbiased stochastic gradient estimator with high variance, while $g_{b,k}$ is a biased estimator with low variance. This merit lies in using a proper weight parameter $\rho_k$ to strike a balance between bias and variance.

With different gradient estimators, we obtain different variants of SHPOS. We denote the variant with $G^{\text{mini}}$ and $G^{\text{vr}}$ as SHPOS-MINI and SHPOS-VR, respectively.

## 4 Theoretical Analysis

In this section, we provide the non-asymptotic convergence analysis for the proposed method under 2-Wasserstein distance $W_2$. Given two probability measures $\rho_1$ and $\rho_2$, the $W_2$ distance is defined as

$$W_2(\rho_1,\rho_2) = \left(\inf_{\pi\in\Gamma(\rho_1,\rho_2)}\int \|x-y\|_2^2 \mathrm{d}\pi(x,y)\right)^{1/2}, \tag{16}$$

where $\Gamma(\rho_1,\rho_2)$ denotes a set of joint distributions on $\mathbb{R}^d\times\mathbb{R}^d$ with marginal distribution $\rho_1$ and $\rho_2$. In our analysis, we first provide the approximate error between the stationary distribution $\nu_\infty$ of $\tilde{\mathbf{x}}_{i,t}$ in the continuous-time dynamics (11) and target distribution in terms of $W_2$ distance, then bound the $W_2$ distance between $\nu_\infty$ and the density $\mu_{k\eta}$ of the discrete-time iterative $\mathbf{x}_{i,k}$ (9), and draw our conclusion by combining these results together as

$$W_2(\mu_{k\eta},p^*) \le W_2(\mu_{k\eta},\nu_\infty) + W_2(\nu_\infty,p^*). \tag{17}$$

If we initialize all the particles with the same distribution, they would endow the same distribution during the evolution due to the exchangeability of the particles according to [Zhang *et al.*, 2020a], and we denote the density of all particles in the $k$-th iteration as $\mu_{k\eta}$. We only list the main theorems here due to the limit of space. We refer readers to Appendix for more detailed proofs.

### 4.1 Assumptions

Following [Zhang *et al.*, 2020a; Zhang *et al.*, 2020b], we make the following standard assumptions on $F$ and $K$.

**Assumption 1.** *$F$ satisfies the following conditions:*

- *$F$ is $L_F$-Lipschitz, i.e., $\|F(\mathbf{x})-F(\mathbf{y})\|_2 \le L_F\|\mathbf{x}-\mathbf{y}\|_2$.*
- *There exists positive $m_F$ such that $\langle F(\mathbf{x}) - F(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle \ge m_F\|\mathbf{x}-\mathbf{y}\|_2$.*

Assumption 1 means that the potential function is smooth and strongly convex. This assumption is quite standard in the analysis of existing dynamics-based sampling methods, including SPOS and many underdamped Langevin MCMC methods [Cheng *et al.*, 2018; Zou *et al.*, 2018]. It is satisfied in many Bayesian learning tasks, such as Bayesian Logistic Regression and Bayesian Ridge Regression. It is techinically possible to extend our theoretical analysis to the non-convex setting, and we leave this as future work.

**Assumption 2.** *$F(\mathbf{x})$ is bounded, i.e. there exists an positive constant $H_F$ such that $\|F(\mathbf{x})\|_2 \le H_F$, and $F(\mathbf{0}) = \mathbf{0}$.*

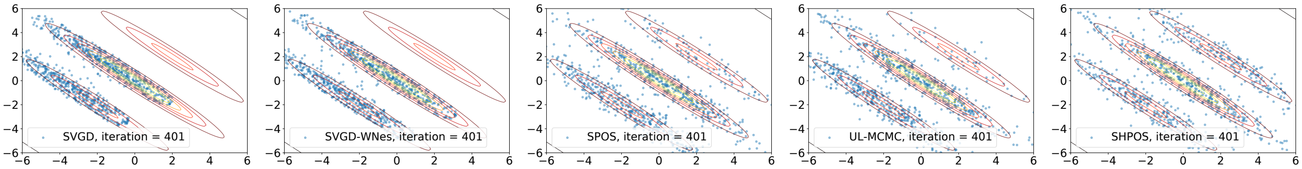
It is easy to be varified that the first part of this assumption holds if $F$ is Lipschitz and the domain $\mathcal{X}$ of $\mathbf{x}$ is a compact set. Besides, we can make $F(\mathbf{0}) = \mathbf{0}$ by transforming coordinate system.

**Assumption 3.** *The kernel function $K$ is $L_K$-Lipschitz and $L_{\nabla K}$ smooth, i.e. $\|\nabla K(\mathbf{x}_1-\mathbf{x}_2) - \nabla K(\mathbf{y}_1-\mathbf{y}_2)\|_2 \le L_{\nabla K}\|\mathbf{x}_1-\mathbf{x}_2-\mathbf{y}_1+\mathbf{y}_2\|_2$.*

Since we focus on RBF kernel in this paper, this assumption can be satisfied by setting the bandwidth $h$ large enough.

Besides, we also assume that the stochastic estimator $F(\mathbf{x},\xi)$ is unbiased and of bounded variance.

**Assumption 4** (Unbiasedness and Bounded Variance)**.** *For all $\mathbf{x}\in\mathbb{R}^d$, the unbiased estimator $F(\mathbf{x},\xi)$ of $F(\mathbf{x})$ has a bounded variance $\mathbb{E}\|F(\mathbf{x},\xi)-F(\mathbf{x})\|_2^2 \le \sigma^2$, where the sample variable $\xi$ is drawn from certain fixed distribution (e.g. the uniform distribution on the indices set or the hyperparameter prior).*



(a) SVGD (b) SVGD-WNes (c) SPOS (d) UL-MCMC (e) SHPOS

Figure 1: Results on sampling from a Gaussian mixture distribution.

## 4.2 Main Theorems

In the following theorem, we establish an upper bound for the 2-Wasserstein distance $W_2(\nu_\infty, p^*)$ between the stationary distribution $\nu_\infty$ of $\tilde{\mathbf{x}}_{i,t}$ in the continuous-time dynamics (11) and the target distribution $p^*$.

**Theorem 1** (Finite Particle Error)**.** *Under assumption 1, 2, and 3, let $\nu_0 = \varphi_0$, there exists a positive constant $c_1$ such that*

$$W_2(\nu_\infty, p^*) \le \frac{c_1}{\lambda_1\sqrt{M}}, \tag{18}$$

*where $\lambda_1 = \frac{m_F}{2\beta L_F} - L_{\nabla K} - \frac{1}{2}L_F - H_F L_K$.*

Recall that $p^*$ is the stationary distribution of the continuous dynamics (8), and (11) is an finite-particle approximation of (8). This theorem shows that the approximation accuracy increases with the growth of particle number .

The following theorem spells out the $W_2$ distance between $\nu_\infty$ and the density $\mu_{k\eta}$ of the discrete-time iterative $\mathbf{x}_{i,k}$ (9).

**Theorem 2** (Time Discretization and Stochastic Gradient Error)**.** *Under assumption 1, 2, 3, and 4, if we run Algorithm 1 with stochastic gradient estimator* (14) *and small step size $\eta$ such that* $\exp(-\lambda\eta) + c_2\eta^2 < 1$*, the 2-Wasserstein distance $W_2(\mu_{k\eta}, \nu_\infty)$ is bounded as follow:*

$$W_2(\mu_{k\eta}, \nu_\infty) \le c_3 e^{-\lambda k\eta} + \frac{\eta^2\Omega}{1-e^{-\lambda\eta}} + \sqrt{\frac{2\eta^2\sigma^2(1+\beta^2 L_F^2)}{|B| L_F^2(1-e^{-\lambda\eta})}}, \tag{19}$$

*where $\lambda = \frac{m_F}{2L_F} - \beta L_{\nabla K} - \beta\frac{1}{2}L_F - \beta H_F L_K$, $(c_2, c_3)$ are positive constants independent of $(M, \eta, k)$, and $\Omega$ is in the order of $\tilde{\mathcal{O}}(\sqrt{d})$.*

Note that the third term in the RHS of (19) belongs to the stochastic gradient error, and this term diminishes if we use the full gradient $F$ in SHPOS.

Based on Theorem 1 and 2, we present our main theorem about the 2-Wasserstein distance between the density $\mu_{k\eta}$ of particles generated in SHPOS and the target distribution $p^*$.

**Theorem 3** (The Overall Bound)**.** *Under the assumptions of theorem 1 and 2, the 2-Wasserstein distance between $\mu_{k\eta}$ and target distribution $p^*$ is in the order of*

$$\tilde{\mathcal{O}}\left(M^{-1/2} + \exp(-\lambda k\eta) + \eta d^{1/2} + \eta^{1/2}|B|^{-1/2}\sigma\right). \tag{20}$$

*If we use the full gradient $F$, the 2-Wasserstein distance between $\mu_{k\eta}$ and target distribution $p^*$ is in the order of*

$$\tilde{\mathcal{O}}\left(M^{-1/2} + \exp(-\lambda k\eta) + \eta d^{1/2}\right). \tag{21}$$

The error bound of SHPOS in Theroem 3 is better than that of SPOS under the same condition. In SPOS [Zhang *et al.*, 2020a], the authors give a bound in terms of 1-Wasserstein distance between the empirical distribution of particles and the target distribution $p^*$, which is in the order of $\tilde{\mathcal{O}}(M^{-1/2} + \exp(-\lambda' k\eta) + M\eta d^{3/2}k^{1/2})$, where $\lambda'$ is a constant in the same order as $\lambda$. This result demonstrates that the sample accuracy of SPOS decreases with particle numbers and iterarions, and a small step size $\eta$ is needed to restrict the 1-Wasserstein distance in a desirable level. Note that 1-Wasserstein distance is a weaker metric than 2-Wasserstein distance ($W_1(\rho_1, \rho_2) \le W_2(\rho_1, \rho_2)$), thus our results in Theorem 3 also hold in terms of the 1-Wasserstein distance. Consequently, it can be verified that the bound (21) of SHPOS is strictly better than the result of SPOS. Futhermore, the sampling accuracy of SHPOS increases as the particle number $M$ and the iteraion $k$ increase, and a much larger step size can be used compared with SPOS to achieve the same accuracy.

# 5 Experiments

We follow the conventions in dynamics-based sampling literature [Liu and Wang, 2016; Zhang *et al.*, 2020b] and conduct empirical studies on one synthetic experiment and two real-world applications. Three types of methods are selected as baselines: ParVI methods (SVGD, SVGD-WNes), dynamics-based MCMC method (UL-MCMC), and Stochastic Particle Optimization Sampling method (SPOS). For a fair comparison, we use a parallel version of the sequential UL-MCMC method, i.e., maintain $M$ independent chains simultaneously. We focus on the stochastic gradient methods in the two real-world applications since it is difficult to calculate the full gradient. We compare our SHPOS-MINI and SHPOS-VR methods with the mini-batch estimator based methods, i.e. SVGD-MINI, SPOS-MINI and SVGD-WNes-MINI, and the variance-reduced stochastic gradient methods SVRG-SPOS [Zhang *et al.*, 2020b] and UL-MCMC-VR [Zhang *et al.*, 2021]. We list the information of datasets and parameter tuning of different methods in our Appendix.

## 5.1 Synthetic Experiment

We consider sampling from a 2-D Gaussian mixture distribution with three modes, whose density is defined as follow:

$$\pi(\mathbf{x}) \propto \exp(-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}) + \frac{1}{2} * \exp(-\frac{1}{2}(\mathbf{x}-\mathbf{a})^T\Sigma^{-1}(\mathbf{x}-\mathbf{a})) + \frac{1}{2} * \exp(-\frac{1}{2}(\mathbf{x}+\mathbf{a})^T\Sigma^{-1}(\mathbf{x}+\mathbf{a})),$$

where $\mathbf{a} = [2, 2]^T$, and $\Sigma_{11} = \Sigma_{22} = 6$ with correlation $\Sigma_{21} = \Sigma_{12} = -0.98 * 6 = -5.88$. We use 1000 particles

| Method | SVGD | SVGD-WNes | SPOS | UL-MCMC | SHPOS |
|---|---|---|---|---|---|
| $W_2$ | 0.9336 | 0.8759 | 0.2174 | 0.2365 | 0.2108 |

Table 2: The $W_2$ distance between the sample distribution and the target distribution of different methods in the synthetic experiment.

that initialized by drawing from a Gaussian distribution with mean $[-4, 2]^T$ and variance $0.25^2$.

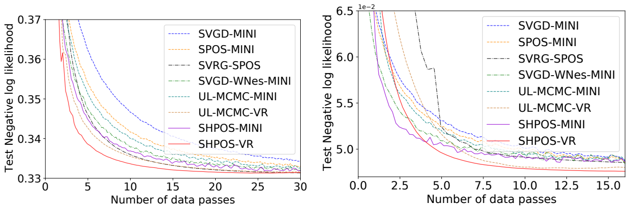
In Figure 1, we report the sampling results generated by each algorithm. From the results, the particle-collapsing phenomenon can be observed in the deterministic ParVI methods SVGD and the accelerated SVGD-WNes method. The particles in these two methods only aggregate near two of the three modes and fail to explore region near the third mode. For the methods with stochastic exploration, i.e. SPOS, UL-MCMC and SHPOS, though SPOS and UL-MCMC could explore all the three high probability space, our SHPOS method has the best performance as it contains both an extra momentum term and a repulsive force to explore the whole space more efficiently. In Table 2, we report the $W_2$ distance between density of random samples generated by each algorithm and the true distribution. We use Sinkhorn method with penalty $10^{-2}$ to calculate the $W_2$ distance. The results show that our SHPOS method achieves the best sampling accuracy.

### 5.2 Bayesian Logistic Regression

Bayesian Logistic Regression (BLR) is a robust binary classification task. Given a dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$, the likelihood of the BLR task is Sigmoid$(y_i \mathbf{w}^T \mathbf{x}_i)$, where $\mathbf{x}_i \in \mathbb{R}^d$ denotes the sample covariate vector, $y_i \in \{-1, 1\}$ denotes the label, and $\mathbf{w}$ is the model parameter with a standard multivariate Gaussian prior $\mathcal{N}(\mathbf{0}, \mathrm{I})$. Four publicly available benchmark datasets from LIBSVM[1], a3a, w8a, a8a, and ijcnn1 are used for evaluation. We report negative log-likelihood versus the number of data passes with 50 particles on datasets a3a and w8a in Figure 2, and we refer readers to Appendix for more results. The experimental results demonstrate that our variance-reduced SHPOS method SHPOS-VR achieves the best performance among all the comparisons. Moreover, it can be verified that the variance-reduced methods outperform their mini-batch counterparts, and both the particle interaction and the momentum contribute to a better performance.

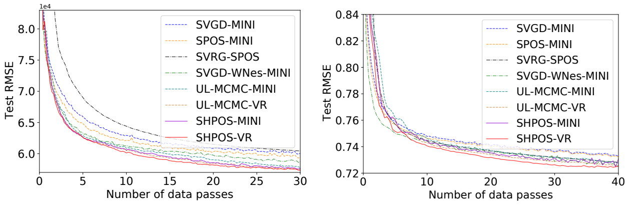
### 5.3 Bayesian Neural Network

In this experiment, we study a regression Bayesian posterior learning task with Bayesian Neural Network (BNN) on 6 datasets from UCI[2] and LIBSVM. Following the settings from [Liu and Wang, 2016], we use a Gamma$(1, 0.1)$ prior for the inverse covariance and adopt a one-hidden-layer neural network with 50 hidden units. In Figure 3, we report the root mean squared error (RMSE) versus the number of data passes with 20 particles, and we refer readers to Appendix for more results. The proposed SHPOS-VR method achieves the best performance, which shows its superiority over other methods. Besides, the results demonstrate that all the stochastic noise, the repulsive force, the extra momentum term, and the variance reduction technique improve the performance. Note that the SVRG estimator based method SVRG-SPOS performs badly in this neural network task. Actually, similar phenomenon has been widely observed in the deep learning literature, and [Defazio and Bottou, 2019] shows that the SVRG estimator is not suitable for nonconvex models. Conversely, the recursive estimator $G^{vr}$ used in SHPOS-VR and UL-MCMC-VR has been shown to be a better estimator for nonconvex problem [Cutkosky and Orabona, 2019; Zhang *et al.*, 2021], which agrees with our results.



(a) a3a (b) w8a

Figure 2: Test negative log-likelihood versus number of data passes



(a) cadata (b) WineWhite

Figure 3: Test RMSE versus the number of data passes.

## 6 Conclusion

In this paper, we propose a second-order particle optimization sampling method called Stochastic Hamiltonian Particle Optimization Sampling (SHPOS), which involves an extra momentum term to achieve acceleration. In SHPOS, the particles can be regarded as drifted by the (stochastic-)gradient of the target distribution, the repulsive force between particles, and an extra random standard Gaussian force. By balancing these forces, particles in SHPOS explore the high probability space of the target distribution rapidly and avoid particle-collapsing. Besides, we also propose a variance-reduced stochastic gradient variant of SHPOS for large-scale tasks. We establish the non-asymptotic convergence analysis for SHPOS, and show that it enjoys a faster convergence rate than the first-order SPOS. The empirical results also demonstrate the superiority of our methods over the state-of-the-art.

## Acknowledgements

This work is supported by National Key Research and Development Program of China under Grant 2020AAA0107400, Zhejiang Provincial Natural Science Foundation of China under Grant No. LZ18F020002, Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies, and National Natural Science Foundation of China (Grant No: 61672376, 61751209, 61472347).

[1] https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/

[2] http://archive.ics.uci.edu/ml/datasets.php

## References

[Braun and Hepp, 1977] Werner Braun and K Hepp. The vlasov dynamics and its fluctuations in the 1/n limit of interacting classical particles. *Communications in mathematical physics*, 56(2):101–113, 1977.

[Chen *et al.*, 2014] Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In *International conference on machine learning*, pages 1683–1691, 2014.

[Chen *et al.*, 2018] Wilson Ye Chen, Lester Mackey, Jackson Gorham, François-Xavier Briol, and Chris J Oates. Stein points. *arXiv preprint arXiv:1803.10161*, 2018.

[Cheng *et al.*, 2018] Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped langevin mcmc: A non-asymptotic analysis. In *Conference on Learning Theory*, pages 300–323. PMLR, 2018.

[Cutkosky and Orabona, 2019] Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. In *Advances in Neural Information Processing Systems*, pages 15236–15245, 2019.

[Defazio and Bottou, 2019] Aaron Defazio and Léon Bottou. On the ineffectiveness of variance reduced optimization for deep learning. In *Advances in Neural Information Processing Systems*, pages 1755–1765, 2019.

[Feng *et al.*, 2017] Yihao Feng, Dilin Wang, and Qiang Liu. Learning to draw samples with amortized stein variational gradient descent. *arXiv preprint arXiv:1707.06626*, 2017.

[Gan *et al.*, 2015] Zhe Gan, Changyou Chen, Ricardo Henao, David Carlson, and Lawrence Carin. Scalable deep poisson factor analysis for topic modeling. In *International Conference on Machine Learning*, pages 1823–1832, 2015.

[Hernández-Lobato and Adams, 2015] José Miguel Hernández-Lobato and Ryan Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International Conference on Machine Learning*, pages 1861–1869, 2015.

[Jaakkola and Jordan, 1997] Tommi Jaakkola and Michael Jordan. A variational approach to bayesian logistic regression models and their extensions. In *Sixth International Workshop on Artificial Intelligence and Statistics*, volume 82, 1997.

[Kloeden and Platen, 1992] Peter E Kloeden and Eckhard Platen. Higher-order implicit strong numerical schemes for stochastic differential equations. *Journal of statistical physics*, 66(1-2):283–314, 1992.

[Li *et al.*, 2019] Zhize Li, Tianyi Zhang, Shuyu Cheng, Jun Zhu, and Jian Li. Stochastic gradient hamiltonian monte carlo with variance reduction for bayesian inference. *Machine Learning*, 108(8-9):1701–1727, 2019.

[Liu and Wang, 2016] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in neural information processing systems*, 29:2378–2386, 2016.

[Liu and Zhu, 2018] Chang Liu and Jun Zhu. Riemannian stein variational gradient descent for bayesian inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[Liu *et al.*, 2017] Yang Liu, Prajit Ramachandran, Qiang Liu, and Jian Peng. Stein variational policy gradient. *arXiv preprint arXiv:1704.02399*, 2017.

[Liu *et al.*, 2019] Chang Liu, Jingwei Zhuo, Pengyu Cheng, Ruiyi Zhang, and Jun Zhu. Understanding and accelerating particle-based variational inference. In *International Conference on Machine Learning*, pages 4082–4092. PMLR, 2019.

[Liu, 2017] Qiang Liu. Stein variational gradient descent as gradient flow. In *Advances in neural information processing systems*, pages 3115–3123, 2017.

[Risken, 1996] Hannes Risken. Fokker-planck equation. In *The Fokker-Planck Equation*, pages 63–95. Springer, 1996.

[Zhang *et al.*, 2020a] Jianyi Zhang, Ruiyi Zhang, Lawrence Carin, and Changyou Chen. Stochastic particle-optimization sampling and the non-asymptotic convergence theory. In *International Conference on Artificial Intelligence and Statistics*, pages 1877–1887. PMLR, 2020.

[Zhang *et al.*, 2020b] Jianyi Zhang, Yang Zhao, and Changyou Chen. Variance reduction in stochastic particle-optimization sampling. In *International Conference on Machine Learning*, pages 11307–11316. PMLR, 2020.

[Zhang *et al.*, 2021] Chao Zhang, Zhijian Li, Zebang Shen, Jiahao Xie, and Hui Qian. A hybrid stochastic gradient hamiltonian monte carlo method. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[Zhuo *et al.*, 2018] Jingwei Zhuo, Chang Liu, Jiaxin Shi, Jun Zhu, Ning Chen, and Bo Zhang. Message passing stein variational gradient descent. In *International Conference on Machine Learning*, pages 6018–6027. PMLR, 2018.

[Zou *et al.*, 2018] Difan Zou, Pan Xu, and Quanquan Gu. Stochastic variance-reduced hamilton monte carlo methods. *arXiv preprint arXiv:1802.04791*, 2018.

[Zou *et al.*, 2019] Difan Zou, Pan Xu, and Quanquan Gu. Stochastic gradient hamiltonian monte carlo methods with recursive variance reduction. In *Advances in Neural Information Processing Systems*, pages 3835–3846, 2019.