Accelerating the diffusion-based ensemble sampling by non-reversible dynamics

Futoshi Futami [1 2 *] **Issei Sato** [1 2] **Masashi Sugiyama** [2 1]

Abstract

Posterior distribution approximation is a central task in Bayesian inference. Stochastic gradient Langevin dynamics (SGLD) and its extensions have been practically used and theoretically studied. While SGLD updates a single particle at a time, ensemble methods that update multiple particles simultaneously have been recently gathering attention. Compared with the naive parallel-chain SGLD that updates multiple particles independently, ensemble methods update particles with their interactions. Thus, these methods are expected to be more particle-efficient than the naive parallel-chain SGLD because particles can be aware of other particles' behavior through their interactions. Although ensemble methods numerically demonstrated their superior performance, no theoretical guarantee exists to assure such particle-efficiency and it is unclear whether those ensemble methods are really superior to the naive parallel-chain SGLD in the non-asymptotic settings. To cope with this problem, we propose a novel ensemble method that uses a non-reversible Markov chain for the interaction, and we present a non-asymptotic theoretical analysis for our method. Our analysis shows that, for the first time, the interaction causes a faster convergence rate than the naive parallel-chain SGLD in the non-asymptotic setting if the discretization error is appropriately controlled. Numerical experiments show that we can control the discretization error by tuning the interaction appropriately.

1. Introduction

In Bayesian inference, a central task is to accurately and efficiently evaluate the posterior distribution (Bishop, 2006; Murphy, 2012). For many practical models, we cannot obtain an analytical expression of the normalizing constant; thus, we need to approximate the posterior. One of the most successfully used methods to approximate the posterior is stochastic gradient Langevin dynamics (SGLD)(Welling & Teh, 2011) and its variants (Ma et al., 2015; Chen et al., 2016; 2014). These are diffusion-based sampling methods and suitable for large-scale data by using not the full gradient but a stochastic version obtained through a randomly chosen subset of data. Each sample in SGLD moves toward the gradient direction with added Gaussian noise (hereinafter, we refer to a sample as a *particle*). Extensions of SGLD have been extensively developed (Ma et al., 2015; Chen et al., 2014) to focus on improving the sampling scheme, which updates one particle at a time, by extending its associated phase space.

On the other hand, ensemble methods that update multiple particles simultaneously have recently been gathering attention (Nusken & Pavliotis, 2019). Compared with naive parallel-chain SGLD, which also updates multiple particles independently at each step, recent ensemble methods introduced some interaction between particles. The advantage of these methods is that the multiple particles interact with each other while moving simultaneously; thus, they have correlations with each other. Because of these correlations, these particles can be aware of each other's behavior and can be more *particle-efficient* than naive parallel-chain SGLD, in which the particles are independent of each other (Liu et al., 2019a). Also, recent development of parallel-processing computation schemes has further encouraged the ensemble methods (Nusken & Pavliotis, 2019). Representative examples of diffusion-based ensemble methods include Stein variational gradient descent (SVGD) (Liu & Wang, 2016) and stochastic particle-optimization sampling (SPOS)(Zhang et al., 2018).

Although the ensemble methods showed superior performance numerically, no theoretical analysis has been conducted to clarify the theoretical advantage of introducing such "interactions" into diffusion-based sampling in a non-asymptotic setting and no work has clarified such improved "particle efficiency". To be more precise, the theoretical advantage of updating multiple particles simultaneously through their interactions compared to naive parallel-chain SGLD, which updates multiple particles independently at

[*]The author is now with NTT. [1]The Univiersity of Tokyo, Tokyo, Japan [2]RIKEN, Tokyo, Japan. Correspondence to: Futoshi Futami <futami@ms.k.u-tokyo.ac.jp>, Issei Sato <sato@g.ecc.u-tokyo.ac.jp>, Masashi Sugiyama <sugi@k.u-tokyo.ac.jp>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

each step, has not been clarified yet.

It is difficult to theoretically compare SVGD and SPOS with naive parallel-chain SGLD because SVGD and SPOS are Vlasov processes (Veretennikov, 2006; Bolley et al., 2010), which are nonlinear Markov processes. Thus, we raise a different, related question: Is it possible to construct an ensemble sampling that is theoretically superior to naive parallel-chain SGLD in a non-asymptotic setting ? We answer this question affirmatively by using the technique of a non-reversible Markov chain (Hwang et al., 2005; Kaiser et al., 2017; Hwang et al., 2015; Duncan et al., 2016; 2017). Although non-reversible methods introduce an additional drift function into the stochastic differential equation (SDE), the introduced drift never changes the stationary distribution of the original SDE and accelerates the convergence. Thus, we propose constructing the interaction between particles with the technique of such non-reversible methods. Then, we theoretically analyze the 2-Wasserstein (W_2) distance and the bias of the given target function in the non-asymptotic setting and compare it with the case of naive parallel-chain SGLD.

Our contributions: The major contributions of this work are as follows.

1. We propose a new ensemble sampling method based on the non-reversible Markov chain technique. Then, we theoretically analyze the proposed sampling scheme in terms of the W_2 distance. To obtain an upper bound on the W_2 distance for our proposed method, we first improve the existing upper bound for standard SGLD, given in Raginsky et al. (2017). Our new bound for standard SGLD shows a tighter upper bound on the constant of the logarithmic Sobolev inequality.

2. To clarify the advantage of using particle interaction, we compare theoretical properties of the proposed sampling method with those of naive parallel-chain SGLD (Chen et al., 2016; Ahn et al., 2014). We find that the interaction causes a trade-off between a larger discretization error and faster convergence to the stationary distribution.

3. We conduct numerical experiments to confirm that we can control the trade-off by tuning the interaction appropriately. Experiments on standard Bayesian models support our theoretical findings and show the superior performance of our method compared to SGLD and other ensemble methods.

Notations: The last page of Appendix gives a summary of the notations used in this paper. Note that $\cdot$ and $\|\cdot\|$ denote the Euclidean inner product and distance, respectively, and $|\cdot|$ is the absolute value. Capital letters such as X represent random variables, and lowercase letters such as x represent usual real values.

2. Preliminary

In this section, we briefly introduce the basic settings of SGLD and its theoretical behavior.

2.1. SGLD and its non-asymptotic behavior

First, we introduce the notations and basic settings of SGLD. Appendix B gives detailed explanations. Our aim is to approximate the target distribution with density $d\pi(x) \propto e^{-\beta U(x)}dx$, where the potential function $U(x)$ is the summation of $u : \mathbb{R}^d \times \mathbb{Z} \to \mathbb{R}$, thus $U(x) = \frac{1}{|\mathcal{Z}|}\sum_{i=1}^{|\mathcal{Z}|} u(x, z_i)$. Here, z_i denotes the data point in some space $\mathbb{Z}$, $|\mathcal{Z}|$ denotes the total number of data points and we express the tuple of data points as $Z = (z_1, \ldots, z_{|\mathcal{Z}|})$. $x \in \mathcal{X} \subset \mathbb{R}^d$ denotes a parameter of the given model.

The SGLD algorithm (Welling & Teh, 2011; Raginsky et al., 2017) is given as the recursion

$$X_{k+1} = X_k - hg(X_k, Q_{z,k}) + \sqrt{2h\beta^{-1}}\epsilon_k, \qquad (1)$$

where $h \in \mathbb{R}^+$ is a step size, $\epsilon_k \in \mathbb{R}^d$ is a standard Gaussian random vector, $g(X_k, Q_{z,k})$ is an conditionally unbiased estimator of the true gradient $\nabla U(X_k)$, and $Q_{z,k}$ is a random variable following the probability $P_z(Q_{z,k})$ that expresses the stochastic access to the subset of data points $\{z_i\}$ and satisfies $\mathbb{E}_{P_z(Q_{z,k})}[g(X_k, Q_{z,k})] = \nabla U(X_k)$ (see Appendix B for the detail). We assume that $X_0, \epsilon_k, Q_{z,k}$ are independent of each other.

The discrete time Markov process Eq.(1) can be regareded as the discretization of the continuous-time Langevin dynamics (Raginsky et al., 2017)

$$dX_t = -\nabla U(X_t) + \sqrt{2\beta^{-1}}dw(t), \qquad (2)$$

where $w(t)$ denotes standard Brownian motion in $\mathbb{R}^d$. The stationary measure of Eq.(2) is $d\pi(x) \propto e^{-\beta U(x)}dx$.

We denote the law of X_k induced by Eq.(1) as μ_{kh} and the law of X_t induced by Eq.(2) as ν_t. Our goal is to sample from the true target measure π. This goal can be naively achieved by taking samples from Eq.(2) according to the ergodic theory. However, Eq.(2) represents a continuous dynamics and we cannot simulate it exactly. Instead, we take samples from the discretized dynamics of Eq.(1). Thus, our interests are in how much μ_{kh} differs from π and in how much μ_{kh} differs from ν_{kh}. In this work, we measure this by the W_2 distance and the bias given a target function. The W_2 distance is expressed by $W_2(\mu_{kh}, \pi)$, where the cost function is Euclidean distance (see Appendix A for the definition). The bias of a given test function f is expressed by $\left|\mathbb{E}f(X_k) - \int_{\mathbb{R}^d} f d\pi\right|$.

We review the result of Raginsky et al. (2017), which established the convergence of the SGLD algorithm in terms of the W_2 distance. Although there are already sharper results

e.g., Xu et al. (2018), in terms of the dimension, our analysis relies on the result of convergence via the logarithmic Sobolev inequality (LSI) (see Appendix C). Thus, we follow the approach in Raginsky et al. (2017), which also used the LSI.

Assumptions: Before proceeding to the result, we introduce the assumptions used in this work, which are the same as those in Raginsky et al. (2017).

Assumption 1. *(Upper bound of the potential function at the origin) The function u takes nonnegative real values and is continuously differentiable on $\mathbb{R}^d$, and there exist constants A, B such that, for all $z \in \mathbb{Z}$,*

$$|u(0,z)| \leq A, \quad \|\nabla u(0,z)\| \leq B. \tag{3}$$

Assumption 2. *(Smoothness) The function u has Lipschitz continuous gradients; that is, for all $z \in \mathbb{Z}$, there exists a positive constant M for all $x, y \in \mathbb{R}^d$,*

$$\|\nabla u(x,z) - \nabla u(y,z)\| \leq M\|x-y\|. \tag{4}$$

Assumption 3. *(Dissipative condition) The function u satisfies the (m,b)-dissipative condition for all $z \in \mathbb{Z}$; that is, for all $x, \in \mathbb{R}^d$, there exist $m > 0, b \geq 0$, such that*

$$-x \cdot \nabla u(x,z) \leq -m\|x\|^2 + b. \tag{5}$$

Assumption 4. *(Initial condition) The initial probability distribution μ_0 of X_0 has a bounded and strictly positive density p_0 and for all $x \in \mathbb{R}^d$,*

$$\kappa_0 := \log \int_{\mathbb{R}^d} e^{\|x\|^2} p_0(x)dx < \infty. \tag{6}$$

Assumption 5. *(Stochastic gradient) There exists a constant $\delta \in [0,1)$ such that*

$$\mathbb{E}_{P(Q_{z,k})}[\|g(x,Q_{z,k}) - \nabla U(x)\|^2] \leq 2\delta\left(M^2\|x\|^2 + B^2\right). \tag{7}$$

The motivation to use the same assumptions as in Raginsky et al. (2017) is that we want to clarify the advantage of introducing interactions in terms of the W_2 distance compared to standard SGLD. Under the above assumptions, the error is bounded in the following way.

Theorem 1. *(Proposition 10 in Raginsky et al. (2017)) Under Assumptions 1 to 5, for any $k \in \mathbb{N}$ and any $h \in (0, 1 \wedge \frac{m}{4M^2})$ obeying $kh \geq 1$ and $\beta m \geq 2$, we have*

$$W_2(\mu_{kh}, \pi) \leq \tilde{C}kh + \sqrt{2\lambda_0 C'}e^{-\frac{kh}{\beta\lambda_0}}, \tag{8}$$

$$C_0 = \left(M^2\left(\kappa_0 + 2\left(1 \vee \tfrac{1}{m}\right)\left(b + 2B^2 + \tfrac{d}{\beta}\right)\right) + B^2\right),$$

$$C_1 = 6M^2(\beta C_0 + d),$$

$$\tilde{C}_0^2 = \left(12 + 8\left(\kappa_0 + 2b + \tfrac{2d}{\beta}\right)\right)(\beta C_0 + \sqrt{\beta C_0}),$$

$$\tilde{C}_1^2 = \left(12 + 8\left(\kappa_0 + 2b + \tfrac{2d}{\beta}\right)\right)(C_1 + \sqrt{C_1}),$$

$$\tilde{C} = \sqrt{\tilde{C}_0^2\sqrt{\delta} + \tilde{C}_1^2\sqrt{h}},$$

$$C' = \log\|p_0\|_\infty + \tfrac{d}{2}\log\tfrac{3\pi}{m\beta} + \beta\left(\tfrac{M\kappa_0}{3} + B\kappa_0^{1/2} + A + \tfrac{b\log 3}{2}\right),$$

and λ_0 is the constant of LSI shown in Eq.(11); see Section 2.2 for details.

In Eq.(8), the first term corresponds to the error due to the discretization and stochastic gradient, i.e., $W_2(\mu_{kh}, \nu_{kh})$ (hereinafter, we refer to this term as the discretization error for simplicity), and the second term corresponds to the convergence to the stationary measure, i.e., $W_2(\nu_{kh}, \pi)$.

2.2. Logarithmic Sobolev inequality

The constant of LSI, λ_0 plays an important role to analyze the SDEs including our non-reversible SGLD. Here, we introduce basic concepts (see Appendix C and Bakry et al. (2013) for more details). First, we introduce *the generator* associated to SDE of Eq.(2) as

$$\begin{aligned}\mathcal{L}f(X_t) :&= \lim_{s\to 0^+} \tfrac{\mathbb{E}(f(X_{t+s})|X_t) - f(X_t)}{s} \\ &= \left(-\nabla U(X_t)\cdot\nabla + \beta^{-1}\Delta\right)f(X_t),\end{aligned} \tag{9}$$

where Δ denotes a standard Laplacian on $\mathbb{R}^d$, $f \in \mathcal{D}(\mathcal{L})$ and $\mathcal{D}(\mathcal{L}) \subset \mathrm{L}^2(\pi)$ denotes the domain of $\mathcal{L}$. This $-\mathcal{L}$ is a self-adjoint operator, which has only discrete spectrums (eigenvalues). We say that π with $\mathcal{L}$ has a *spectral gap* if the smallest eigenvalue of $-\mathcal{L}$ other than 0 is positive. We refer to it as $\rho_0 (> 0)$ (see Appendix C). We say that π with $\mathcal{L}$ satisfies the *(tight) logarithmic Sobolev inequality* (LSI) with constant λ_0 (we call this LSI(λ_0)) if for any f that is integrable ($\int_{\mathbb{R}^d} f|\log f|d\pi < \infty$), π with $\mathcal{L}$ satisfies,

$$\mathrm{Ent}_\pi(f^2) \leq -2\lambda_0 \int_{\mathbb{R}^d} f\mathcal{L}f d\pi, \tag{10}$$

$$\mathrm{Ent}_\pi(f) := \textstyle\int_{\mathbb{R}^d} f\log f d\pi - \int_{\mathbb{R}^d} f d\pi \log\left(\int_{\mathbb{R}^d} f d\pi\right).$$

Then, Raginsky et al. (2017) clarified that under the conditions of Theorem 1, π with $\mathcal{L}$ of Eq.(9) satisfies LSI(λ_0) and an upper bound of λ_0 is given as

$$\lambda_0 \leq \lambda_l := D_1 + \rho_0^{-1}(D_2 + 2), \tag{11}$$

$$D_1 = \tfrac{2m^2 + 8M^2}{\beta m^2 M}, \quad D_2 \leq \tfrac{6M(d+\beta)}{m}, \tag{12}$$

$$\rho_0^{-1} \leq \tfrac{2C(d+b\beta)}{m\beta}\exp\left(\tfrac{2}{m}(M+B)(b\beta+d) + \beta(A+B)\right) + \tfrac{1}{m\beta(d+b\beta)}. \tag{13}$$

This constant λ_0 controls the convergence speed in Theorem 1. The smaller λ_0 means faster convergence. From D_2 and ρ, larger d means larger λ_0 (see Propositions 13 and 15, Appendix B in Raginsky et al. (2017) or Theorem 1.2 (2) in Cattiaux et al. (2010) for details).

3. Proposed ensemble sampling

As we mentioned in the introduction, we update N particles simultaneously. First, we introduce the notations to treat the multiple particles. We express the n-th particle at time t as $X_t^{(n)} \in \mathbb{R}^d$. We express the joint state of all the N particles at time t as $X_t^{\otimes N} := (X_t^{(1)}, \ldots, X_t^{(N)})^\top \in \mathbb{R}^{dN}$.

We express the joint stationary measure as $\pi^{\otimes N} := \pi \otimes \cdots \otimes \pi \propto e^{-U(X^{(1)})-U(X^{(2)})-\cdots-U(X^{(N)})}$.

3.1. Naive parallel-chain SGLD

First, we introduce naive parallel-chain SGLD. The N-parallel and independent chain is written as

$$dX_t^{\otimes N} = -\nabla U^{\otimes N}(X_t^{\otimes N})dt + \sqrt{2\beta^{-1}}dw_t, \quad (14)$$

$$\nabla U^{\otimes N}(X_t^{\otimes N}) := \left(\nabla U(X_t^{(1)}), \ldots, \nabla U(X_t^{(N)})\right)^\top, \quad (15)$$

and w_t is the dN-dimensional Wiener process. The discretized dynamics with the stochastic gradient is given as

$$X_{k+1}^{\otimes N} = X_k^{\otimes N} - g_k^{\otimes N}h + \sqrt{2\beta^{-1}}\epsilon_k, \quad (16)$$

$$g_k^{\otimes N} := (g(X_k^{(1)}, Q_{z,k}), \ldots, g(X_k^{(N)}, Q_{z,k}))^\top, \quad (17)$$

where each $g(X_k^{(n)}, Q_{z,k})$ is an unbiased estimator of the gradient $\nabla U(X_t^{(n)})$ and for simplicity, we assume the same random access to the data points for all n. Intuitively, this means we use the same subset of data for all n. $\epsilon_k \in \mathbb{R}^{dN}$ is a standard Gaussian random vector. Eq.(16) is the baseline method of the ensemble sampling since there is no interaction among particles. This dynamics is just the concatenation of the d-dimensional single chain introduced in Eq.(2). We assume that all the initial measures $\{X_0^{(n)}\}_{n=1}^N$ are the same. Then, all the marginal probability at any time $t \geq 0$ will be the same. We study theoretical properties of the dynamics in Eq.(16) in Section 4.2.

3.2. Proposed algorithm

Building on naive parallel-chain SGLD, we propose our sampling scheme. Motivated by existing ensemble methods, including SVGD and SPOS, we introduce an interaction term into naive parallel-chain SGLD. Specifically, we introduce the additional drift term γ in the following way:

$$dX_t^{\otimes N} = -\nabla U^{\otimes N}(X_t^{\otimes N})dt + \alpha\gamma(X_t^{\otimes N})dt + \sqrt{2\beta^{-1}}dw_t, \quad (18)$$

where $\alpha \in \mathbb{R}$ expresses the strength of the interaction term. Since the stationary measure should not be changed by the interaction, we assume that the interaction γ satisfies the divergence-free condition: $\nabla \cdot (\gamma\pi^{\otimes N}) = 0$. Then, we can easily confirm that the interaction never changes the stationary measure (see Appendix H.1). This type of drift term has been studied in Hwang et al. (2005), Kaiser et al. (2017), Hwang et al. (2015), Duncan et al. (2016), Duncan et al. (2017), Hu et al. (2020). There are multiple ways to construct such γ. Our strategy is using a skew-symmetric matrix J as

$$\gamma(X_t^{\otimes N}) = -J\nabla U^{\otimes N}(X_t^{\otimes N}), \quad J = -J^\top. \quad (19)$$

This surely satisfies the divergence-free condition. This is motivated by SVGD and SPOS, which use the derivative of a kernel function as the interaction. Note that the derivative of the kernel Gram matrix is a skew-symmetric matrix. Then, we introduce a discretized dynamics as

$$X_{k+1}^{\otimes N} = X_k^{\otimes N} - g_k^{\otimes N}h + \alpha\gamma_{g^{\otimes N}}h + \sqrt{2\beta^{-1}}\epsilon_k, \quad (20)$$

$$\gamma_{g^{\otimes N}} := -Jg_k^{\otimes N}. \quad (21)$$

We denote the law of $X_k^{\otimes N}$ induced by Eq.(20) as $\mu_{kh}^{\otimes N}$ and the law of $X_t^{\otimes N}$ induced by Eq.(18) as $\nu_{kh}^{\otimes N}$. We discuss theoretical properties of this dynamics in Section 4.3.

4. Theoretical properties

In this section, we first improve the bound of standard SGLD, Eq.(1) and then, analyze our proposed method.

4.1. Standard SGLD

First, we present our bound for standard SGLD, then discuss its difference from the Theorem 1 of Raginsky et al. (2017).

Theorem 2. *Under Assumptions 1 to 5, for any $k \in \mathbb{N}$ and any $h \in (0, 1 \wedge \frac{m}{4M^2})$ obeying $kh \geq 1$ and $\beta m \geq 2$, μ_{kh}, which is induced by Eq.(1), satisfies*

$$W_2(\mu_{kh}, \pi) < \sqrt{C_3\,(kh + C_4)\,kh} + \sqrt{2\lambda C'}e^{-\frac{kh}{\beta\lambda}}, \quad (22)$$

$$C_3 := \tfrac{6}{\beta}\,(C_1 h + \beta C_0 \delta),$$

$$C_4 := (6M^2)^{-1}\left(\sqrt{2\pi(3M^2)^{-1}}\exp\left(\tfrac{3M^2}{2}(kh)^2\right) - kh\right),$$

where C_0, C_1, and C' are given in Eq.(8). λ is the LSI constant.

We can obtain a tighter bound for the LSI constant than that of Raginsky et al. (2017).

Theorem 3. *Under the same conditions as Theorem 2 and the additional condition $(4d+9)\pi e^2 > \beta m \geq 16\pi e^2/3$, the LSI constant is upper-bounded by λ_e:*

$$\lambda \leq \lambda_e := ((1 + \rho_0^{-1}|C|)2\pi e^2)^{-1} + 3(2\rho_0)^{-1}, \quad (23)$$

$$-C := \inf_x \left\{\tfrac{\beta}{4}\|\nabla U(x)\|^2 - \tfrac{1}{2}\nabla^2 U(x) - \pi e^2 U(x)\right\}, \quad (24)$$

where ρ_0 is given in Eq.(13) and C is bounded by

$$0 < C \leq \frac{\beta B^2}{4} + \frac{b\pi e^2}{2}\log 3 + \frac{Md}{2}. \quad (25)$$

Moreover, λ_e is always smaller than λ_l of Eq.(11) estimated by Raginsky et al. (2017).

The proof of Theorem 2 is shown in Appendix E.1 and the proof of Theorem 3 is shown in Appendix K.1. We may further eliminate the additional assumption of Theorem 2. See Theorem 14 in Appendix L for details.

Outline of the proof: Our proof is similar to that of Raginsky et al. (2017). First, we decompose the W_2 distance in the following way:

$$W_2(\mu_{kh}, \pi) \leq W_2(\mu_{kh}, \nu_{kh}) + W_2(\nu_{kh}, \pi). \quad (26)$$

Then, we bound the convergence to the stationary, $W_2(\nu_{kh}, \pi)$, in the same way as Raginsky et al. (2017) using the property of LSI (see Appendix E.1). The difference is the discretization error, $W_2(\mu_{kh}, \nu_{kh})$. Similarly to Chen et al. (2019), we consider the continuous-time interpolation of Eq.(1) and denote by V_k, of which measure is the same as μ_{kh}. Then, we use the relation $W_2^2(\nu_{kh}, \mu_{kh}) \leq \mathbb{E}\|X_k - V_k\|^2$, and upper-bound the right-hand side of this inequality and applied Gronwall's inequality to it.

As for the estimation of the LSI constant, we use the method of Carlen & Loss (2004), which relies on a restricted LSI and a spectral gap. If $-C$, which is defined in Eq.(24), is lower-bounded, then, π admits an LSI and its constant is upper-bounded by $\lambda \leq ((1+\rho^{-1}|C|)2\pi e^2)^{-1} + 3(2\rho)^{-1}$ where ρ is a spectral gap. See Appendix K.1 for the proof.

Comparison with Theorem 1: The W_2 distance of our Theorem 2 shows better dependency on N compared to Theorem 1, especially for the discretization error and the LSI constant. First, we discuss the discretization error. In Theorem 1, the discretization error is $\tilde{C}kh$, which depends on d linearly due to the weighted CKP inequality in the derivation. On the other hand, our discretization error shows $d^{1/2}$-dependency. This gap is important when we consider ensemble sampling. Let us consider the bias of an ensemble sampling; that is, with N-particles, we approximate the integral of a test function f by $\frac{1}{N}\sum_{n=1}^N f(X_k^{(n)})$. If a test function f is L_f-lipschitz in $\mathbb{R}^d$, the bias $\left|\mathbb{E}\frac{1}{N}\sum_{n=1}^N f(X_k^{(n)}) - \int_{\mathbb{R}^d} f d\pi\right|$ can be upper-bounded by the W_2 distance multiplied by $L_f/\sqrt{N}$. Additionally, when we assume the W_2 distance of this ensemble sampling can be upper-bounded by the same approach as Theorem 1, then since the N-particle system is dN-dimensonal, its discretization error linearly depends on dN. Thus, the bias is $\mathcal{O}(\sqrt{N})$. This means that the more particles we use, the larger bias we suffer. This is an undesirable property as the ensemble sampling. Our approach in Theorem 2 does not suffer from this problem, since the discretization error depends on $\sqrt{dN}$; thus, the bias of ours has the constant order with respect to N. However, our discretization error is crude which entails $\sqrt{kh}e^{k^2h^2}$. See Appendix E.2 for more details. We may further improve the discretization error based on Vempala & Wibisono (2019). See Theorem 9 in Appendix F for details.

Next, we discuss the upper-bound of the LSI constant. In Raginsky et al. (2017), the LSI constant is estimated via the Lyapunov condition-based approach (Cattiaux et al., 2010). Its estimate is given by $\lambda \leq a + \rho_0^{-1}(a' + a''\int_{\mathbb{R}^d}\|x\|^2 d\pi)$, where a, a', a'' are some positive constants and independent of d. Thus, if we consider the dN-dimensional particle system, the estimated LSI constant becomes significantly larger than the single-particle system due to the second-moment term. Since the larger LSI constant means the slower convergence to the stationary measure, the convergence speed of the N-particle system is much slower than that of standard SGLD. Thus, this results in a larger bias. On the other hand, our estimation in Theorem 3 does not show such behavior. Moreover, as we will see in Section 4.3, we can show that our estimate of the LSI constant for the proposed ensemble sampling is smaller than that of standard SGLD. However, we need the stronger condition of $\beta m \geq 16\pi e^2$ than that of Raginsky et al. (2017), which is $\beta m \geq 2$. See Appendix K.1 for more details.

4.2. Naive parallel-chain SGLD

We analyze naive parallel-chain SGLD with N particles. Since naive parallel-chain SGLD is just the N concatenation of standard SGLD, its W_2 distance is $N^{1/2}$ times larger than Eq.(8). In addition to the W_2 distance, we consider bias here additionally. Our goal is to approximate the integral of the test function f with L_f-lipschitzness by the ensemble average $\frac{1}{N}\sum_{n=1}^N f(X_k^{(n)})$. Then we obtain,

Corollary 1. *Under the same conditions as Theorem 2, $X_k^{\otimes N}$ of Eq.(16) satisfies*

$$\left|\mathbb{E}\frac{1}{N}\sum_{n=1}^N f(X_k^{(n)}) - \int_{\mathbb{R}^d} f d\pi\right| < L_f\left(\sqrt{C_3\,(kh + C_4)\,kh} + \sqrt{2\lambda C'}e^{-\frac{kh}{\beta\lambda}}\right), \tag{27}$$

where the constants C_0, C_1, C' and λ are given in Theorem 2.

The proof is shown in Appendix G. Note that this bias does not depend on N, which means that using multiple chains will not contribute to reducing the bias.

4.3. Proposed method

Here, we analyze our proposed method. Since we control the magnitude of the interaction by α, we impose the additional condition about the norm of J:

Assumption 6. *A skew-symmetric matrix J is bounded as*

$$\|J\|_{\mathrm{F}} \leq 1, \tag{28}$$

where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm.

Then, we have our main theorem,

Theorem 4. *Under the same conditions as Theorem 3 and Assumption 6, $\mu_{kh}^{\otimes N}$, which is induced by Eq.(20), satisfies*

$$W_2(\mu_{kh}^{\otimes N}, \pi^{\otimes N}) < N^{1/2}\left(\sqrt{C_3'(\alpha)\,(kh + C_4'(\alpha))\,kh} + \sqrt{2\lambda C'}e^{-\frac{kh}{\beta\lambda}}\right), \tag{29}$$

where $C_3'(\alpha), C_4'(\alpha)$ are the positive constants, which are obtained by replacing $M \to (1+\alpha)M$, $B^2 \to (1+\alpha)^2 B^2$

in C_3 and C_4 of Eq.(22) (see Appendix H for details) and λ is the LSI constant bounded by $\lambda(\alpha, N)$:

$$\lambda \leq \lambda(\alpha, N) \begin{cases} \leq \lambda_e & \text{if } \alpha \neq 0 \\ = \lambda_e & \text{if } \alpha = 0 \end{cases}. \tag{30}$$

The proof is shown in Appendix H. It is clear that when we substitute $N = 1$ and $\alpha = 0$ in Theorem 4, the bound will be equal to standard SGLD, Eq.(22). This is natural since these conditions means that there is no interaction.

From the above theorem, we can easily find that the bias of our proposed method is

$$\left|\mathbb{E}\frac{1}{N}\sum_{n=1}^{N} f(X_k^{(n)}) - \int_{\mathbb{R}^d} f d\pi\right| < L_f\left(\sqrt{C_3'(\alpha)\left(kh + C_4'(\alpha)\right)kh} + \sqrt{2\lambda C'}e^{-\frac{kh}{\beta\lambda}}\right), \tag{31}$$

which never increases as N increases. N only appears through the upper-bound of the LSI constant. See Appendix H.7 for this proof. We may further improve the discretization error based on Vempala & Wibisono (2019). See Theorem 11 in Appendix I for details.

Outline of the proof of Theorem 4

Modified dissipative condition: First, we study how the parameters in the dissipative and smoothness assumptions are modified by the interaction. We define the drift function $\nabla u_\alpha(x^{\otimes N}, z) := \nabla u^{\otimes N}(x^{\otimes N}, z) + \alpha J \nabla u^{\otimes N}(x^{\otimes N}, z)$ and $\nabla U_\alpha^{\otimes N} = \sum_z \nabla u_\alpha(x^{\otimes N}, z)/|\mathcal{Z}|$. Then, we have

Lemma 1. *Let $x^{\otimes N}, y^{\otimes N} \in \mathbb{R}^{dN}$, for all z,*

$$-x^{\otimes N} \cdot \nabla u_\alpha(x^{\otimes N}, z) \leq -m\|x^{\otimes N}\|^2 + bN, \tag{32}$$

$$\|\nabla u_\alpha(x^{\otimes N}, z) - \nabla u_\alpha(y^{\otimes N}, z)\| \leq M(1+\alpha)\|x^{\otimes N} - y^{\otimes N}\|. \tag{33}$$

Here the dot $\cdot$ and $\|\cdot\|$ are the inner product and norm in $\mathbb{R}^{dN}$. Because of the skew-symmetric property of J, the dissipative constant m does not change. This is a crucial property in our analysis. The proofs and other conditions are discussed in Appendix H.

Based on these modified conditions, we bound the W_2 distance in a similar way to standard SGLD in Section 4.1. We just change the constants in the assumptions.

Smaller upper-bound of the LSI constant: Next, we discuss the estimation of the LSI constant. Note that the generator of Eq.(18) is

$$\mathcal{L}_\alpha := \left(-\nabla U_\alpha^{\otimes N}(X_t^{\otimes N}) \cdot \nabla + \beta^{-1}\Delta\right). \tag{34}$$

Then, under the same conditions as Theorem 2, $\pi^{\otimes N}$ with $\mathcal{L}_\alpha$ satisfies the LSI and there exists a spectral gap $\rho(\alpha, N)$ (see Appendix C). Then, our interest is how the upper-bound of the LSI constant $\lambda(\alpha, N)$ and $\rho(\alpha, N)$ depend on N, α. We answer this in the following lemma:

Lemma 2. *Under the same conditions as Theorem 4, we have*

$$\lambda(\alpha, N) \leq \lambda(\alpha = 0, N) = \lambda_e < \lambda_l. \tag{35}$$

This means that the upper-bound of the LSI constant of the proposed method can be smaller than that of naive parallel-chain SGLD. Moreover, it is bounded by that of standard SGLD. The proof is shown in Appendix K.2. Here, we briefly describe the outline of the proof. First, note that Eq.(260) is monotonically increasing function about ρ^{-1} if C is fixed. This means that the larger the spectral gap ρ is, the smaller the upper-bound of the LSI constant is. Thus, we need to evaluate the spectral gaps. We can prove $\rho(\alpha, N) \geq \rho(0, N)$ by the spectral decomposition of $\mathcal{L}_\alpha$. Then, since $\mathcal{L}_{\alpha=0}$ is the generator of naive parallel-chain SGLD, we can apply this tensorization property of a spectral gap. This results in $\rho(0, N) = \rho_0$. Next, we prove the constant C of Eq.(260) for $\mathcal{L}_\alpha$, $\mathcal{L}_{\alpha=0}$ and $\mathcal{L}$ are the same. Finally, combined with the inequality of spectral gaps and the equality of C, we get the lemma. See Appendix K.2 for more details.

We cannot obtain this lemma in the approach of Raginsky et al. (2017) because we cannot conclude that the larger the spectral gap ρ_0 is, the smaller the LSI constant is. This is because, when we use the Lyapunov condition-based approach, its estimation includes the term: $\rho_0^{-1}\mathbb{E}_\pi\|X\|^2$ and this second moment of the N-particle system can be N times larger than that of the single-particle system.

4.4. Comparison with naive parallel-chain SGLD

In Eq.(31), the first term is dominated by the discretization error and the second term is the convergence to the stationary. Compared to the naive parallel-chain bound in Eq.(27), the discretization error becomes larger due to the additional interaction term. On the other hand, since the upper-bound of the LSI constant becomes small, the convergence speed is improved. In conclusion, when we use the non-reversible interaction term, there is a trade-off between the larger discretization error and faster convergence speed.

From Theorem 4, we should set α to be small enough so that the discretization error will not become so large. Under the assumption that α is sufficiently small, we can evaluate how much the spectral gap is improved in the following way:

Theorem 5. *Let us denote the pairs of the eigenvalues and eigenvectors of $-\mathcal{L}_{\alpha=0}$ as $\{(\rho_k, e_k)\}_{k=0}^{\infty}$, which satisfies $0 < \rho_0 < \rho_1 < \cdots$. Then, the spectral gap is $\rho(0, N) = \rho_0$. Let $V := \mathcal{L}_{\alpha\neq 0} - \mathcal{L}_{\alpha=0}$. Under the same conditions as Theorem 4, we have*

$$\rho(\alpha, N) = \rho(0, N) + \alpha^2 \sum_{k=1}^{\infty} \frac{|\int e_k V e_0 d\pi^{\otimes N}|^2}{\rho_k - \rho_0} + \mathcal{O}(\alpha^3).$$

The proof is shown in Appendix J.3. This is the perturbation of the operator $\mathcal{L}_\alpha$. Note that the first-order of α is zero due

to the skew-symmetric property of J. The second term of the above equation is always positive since for all $k \geq 1$, $\rho_k > \rho_0$. Thus, up to the second-order of α, the spectral gap becomes large. In practice, since it is difficult to calculate the eigenvectors and eigenvalues of $\mathcal{L}_\alpha$, evaluation of the second term is difficult numerically.

5. Related work

In this section, we discuss the relation of our proposed method to other ensemble sampling methods and non-reversible Markov chain methods.

5.1. Comparison with other ensemble samplings

Although Stochastic particle-optimization sampling (SPOS) (Zhang et al., 2018) is the most closely related method to ours, it is a Vlasov process, of which drift function depends on the probability law at each time steps. Since we do not know the explicit expression of this law in practice, we need the empirical approximation for it by particles. This introduces an additional bias. To reduce this bias, we need a large number of particles, which causes high computational costs.

Another difference between SPOS and the proposed method is that we upper-bound the W_2 distance, while the bound of SPOS is upper-bounded in terms of the W_1 distance. Then, for the discretization error, they obtained the bound following the approach in Raginsky et al. (2017). Thus, the bias is $\mathcal{O}(\sqrt{N})$. On the other hand, as shown in Theorem 4, our bound does not depends on N explicitly and N only affects the LSI constant. As for the convergence rate, they showed the exponential convergence and its exponent depends on ρ^{dN}, where ρ is the positive constant, $\rho \in [0, 1)$. This means that as we increase the number of particles, the convergence speed drops significantly. Thus, it is hard to recognize the advantage of the ensemble method.

Another famous ensemble method is Nusken & Pavliotis (2019). While our method correlates particles by using the divergence-free drift, Nusken & Pavliotis (2019) correlates particles by the coupling technique, such as synchronous coupling, mirror coupling. Another difference is that we focused on non-asymptotic behavior, on the other hand, they focused on asymptotic behavior.

The existing parallel-chain SGLD methods, e.g. Chen et al. (2016); Ahn et al. (2014), focus on reducing the computational cost of calculating the gradient by the distributed framework. On the other hand, our method focuses on accelerating the sampling.

Other than sampling, Stein variational gradient descent (SVGD) (Liu & Wang, 2016) is the most widely used ensemble method. However, SVGD is not a valid sampling, which is pointed out in Zhang et al. (2018). Moreover, because it is a Vlasov process, it is hard to assure the theoretical guarantee under the non-asymptotic settings. Thus, the theoretical advantage as the ensemble method is unclear.

5.2. Comparison with the non-reversible drift work

Compared to existing non-reversible Markov chain work (Hwang et al., 2005; 2015; Duncan et al., 2016; 2017; Kaiser et al., 2017), our work has both theoretical and numerical contributions in this field. We believe that this work is the first step to clarify the non-asymptotic behavior of the non-reversible Markov chain with the non-convex potential function, which is widely used in the field of SGLD, while the existing work of non-reversible Markov chain has focused on the asymptotic settings. Although some work also focused on the convergence speed, they only took into account the Ornstein-Uhlenbeck (OU) processes, which have the convex potential functions and are limited. As for the convergence, we focused on the LSI under the non-reversible drift settings and derived the explicit formula (Theorem 5) about the improvement of a spectral gap.

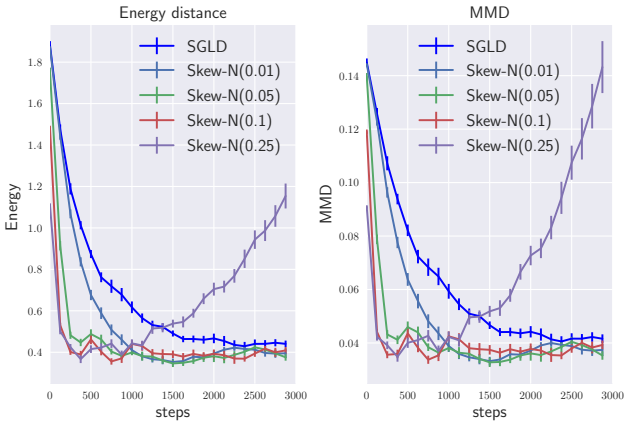
As for the numerical contributions, we believe that this work is the first attempt to apply the divergence-free drift method to the standard Bayesian models. Most existing work only took into account OU processes. In the next section, we numerically clarify that the divergence-free drift methods are promising for sampling in Bayesian inference.

6. Numerical experiments

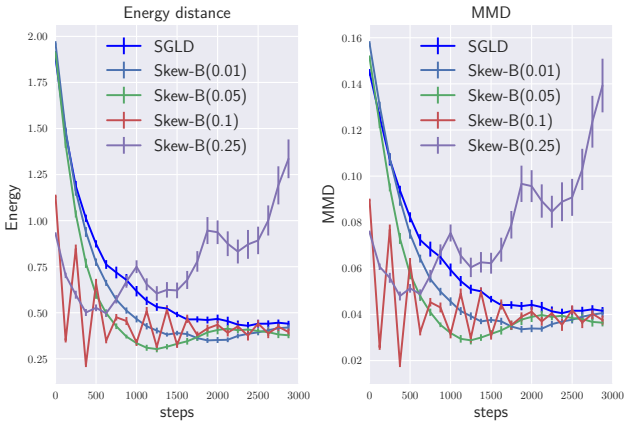
Detailed experimental settings are shown in Appendix M. From the theoretical analysis, we confirmed that there is a trade-off between discretization error and the convergence speed. Thus, it is natural to consider that if we tune the interaction α and J appropriately, we can improve the convergence speed while regulating the discretization error. We confirm this numerically since theoretical analysis does not tell us what is the optimal α and J.

Thus, the primal purpose of the numerical experiments is to confirm that our proposed ensemble methods enjoy better and faster performance compared to naive parallel-chain SGLD. Additionary, we compared the proposed method with other ensemble methods; SPOS, SVGD. We also changed the value of α so that how α affects the discretization error and the convergence rate. The models we used are simple and widely used Bayesian models including the Ornstein-Uhlenbeck process (OU), Bayesian logistic regression (BLR), Latent Dirichlet Allocation (LDA) and Bayesian neural net (BNN).

Another purpose of the experiments is to study the effect of the choice of J since it is unknown how to construct the skew-symmetric matrix J for the smaller bias theoretically. Thus, we prepared three types of J. We generated J in the following way: First, generated an upper triangular matrix

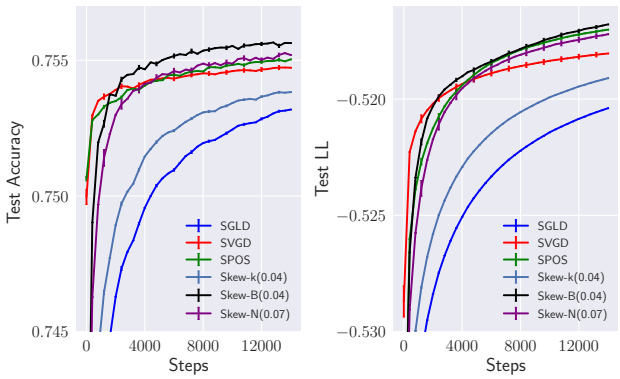


(a) Effect of different α with *Skew-N*

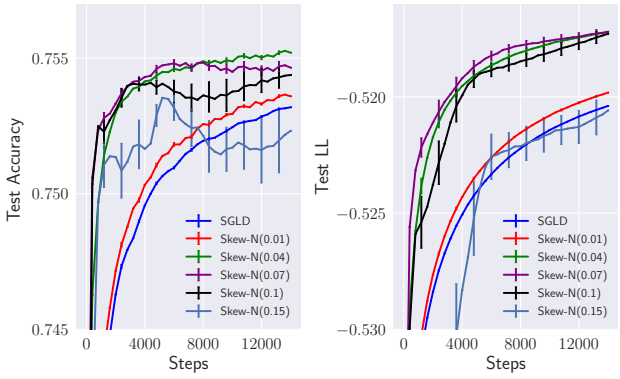


(b) Effect of different α with *Skew-B*

Figure 1. OU experiments (Averaged over 10 trials)



(a) Comparison with different methods



(b) Effect of different αs

Figure 2. BLR experiments (Averaged over 10 trials)

J' randomly and then calculated $J' - J'^{\top}$. We generated two types of J', of which each entry follows the Bernoulli distribution and the Gaussian distribution. Then, we normalized them to satisfy Assumption 6. We refer to this matrix multiplied α as *skew-B(α)* that is generated from the Bernoulli distribution and *skew-N(α)* that is generated from the Gaussian distribution in the followings. Another skew-symmetric matrix is that before taking the normalization in *skew-N*, we multiplied the kernel Gram matrix of RBF kernel, of which elements are X_0 from both left and right-hand side. This is expressed as *skew-k(α)*.

We used 20 particles for all the experiments except for OU. We repeated 10 trials for OU, BLR and LDA experiments, and 20 trials for BNN experiments. The following values and error bars are the mean and the standard deviation of these trials.

Ornstein-Ohlenbeck process: This process is given by

$$dX_t = \Sigma^{-1}(X_t - \mu)dt + \sqrt{2}dw(t) \tag{36}$$

for the standard SGLD. Its stationary distribution is $\pi = N(\mu, \Sigma)$. Theoretical properties of this dynamics and its discretized version have been widely studied (Wibisono, 2018). An important property is that there is a formula for $W_2(\nu_t, \pi)$ if the initial distribution is Gaussian (see Appendix M for the details). Thus, by studying the convergence behavior of OU, we can understand our proposed method more clearly.

In our experiments, we used 100 particles. Since calculating the W_2 distance is computationally demanding, we used the energy distance (Székely et al., 2004) and the maximum mean discrepancy (MMD) (Gretton et al., 2007) between $\mu_k^{\otimes N}$ and the stationary distribution as indicators to observe the convergence. The results are shown in Figure 1.

We can see that if α is set to be very small, its performance is close to naive parallel-chain SGLD, while if α is set to too large, it suffers from the large discretization error. This shows that there is a trade-off between the larger discretization error and faster convergence by the interaction, as our analysis clarified.

Bayesian logistic regression experiment: Following Liu & Wang (2016), we test on BLR using Covertype dataset

Table 1. Holdout perplexity (Averaged over 10 trials)

Method	Test perplexity
SGLD	1034.86 ± 1.46
SVGD	1029.97 ± 1.02
SPOS	1031.42 ± 1.15
Skew-k(0.01)	1029.12 ± 1.35
Skew-N(0.02)	1026.47 ± 1.72
Skew-B(0.01)	$\mathbf{1024.33 \pm 1.85}$

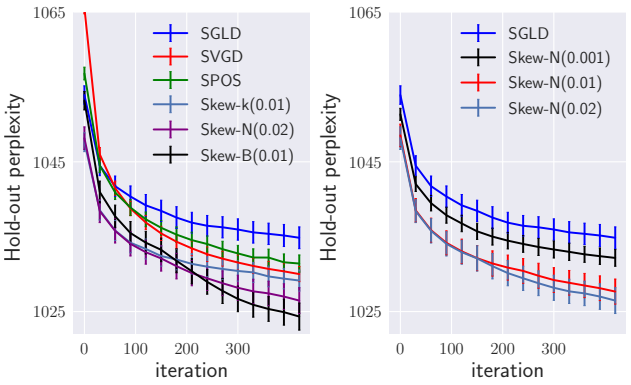


Figure 3. LDA experiments (Averaged over 10 trials)

Table 2. Results of BNN experiments (Averaged over 20 trials)

Method	Test RMSE ($\times 10^{-2}$)	Test LL
SGLD	6.92 ± 0.08	1.20 ± 0.01
SVGD	7.24 ± 0.07	1.16 ± 0.01
SPOS	6.88 ± 0.07	1.21 ± 0.01
Skew-N(0.05)	6.86 ± 0.08	1.21 ± 0.01
Skew-B(0.05)	6.90 ± 0.07	1.21 ± 0.01

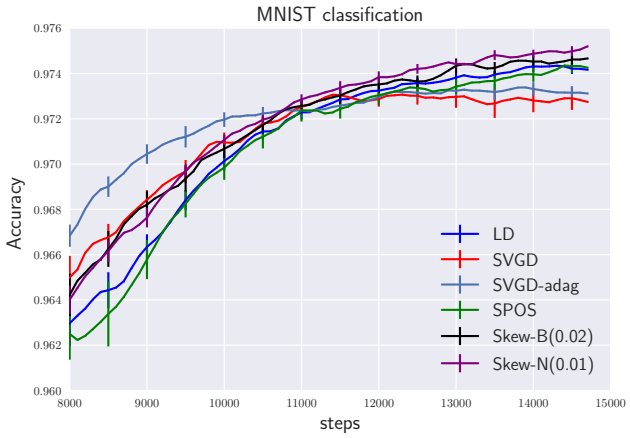


Figure 4. Mnist classification (Averaged over 10 trials)

(Dua & Graff, 2017) and the result is shown in Fig.2. In Fig.2(a), we can see that by the interaction, the convergence speed and performance is improved compared to naive parallel-chain SGLD. In Fig.2(b), we changed α in *skew-N*. We can see a trade-off between the larger discretization error and faster convergence, which is similar to the results of OU. The results of *skew-B* is shown in Appendix M.

Latent Dirichlet allocation experiment: We test on LDA model using the ICML dataset (Ding et al., 2014) following the same setting as Patterson & Teh (2013). The result is shown in Table.1 and Fig.3. From the left-figure of Fig.3 and Table.1, the proposed method shows faster and superiror performance compared to naive parallel-chain SGLD, and competitive performance with SVGD and SPOS. In the left-figure of Fig.3, we did the experiments with different α, and found that the result is robust to the choice of α.

Bayesian neural net regression: We test on the BNN regression task using Kin8nm dataset of UCI (Dua & Graff, 2017), following the same setting as Liu & Wang (2016). The results are shown in Table 2. We found that the proposed methods shows competitive performance with other ensemble methods. We show an additional Figure in Appendix M.

Bayesian neural net classification: We test on the BNN classification task using MNIST (LeCun & Cortes, 2010) dataset. We used a fully connected two-layer neural network with 100 and 50 hidden units. The detailed settings are shown in Appendix M. The result is shown in Figure 4. We found that proposed methods show competitive performance with other ensemble methods.

7. Conclusion

In this work, we proposed the new diffusion-based ensemble sampling, which updates many particles simultaneously with interaction by using the non-reversible drift term. We also derive the non-asymptotic bound and compare it with that of the naive parallel-chain SGLD. Introducing the interactions have resulted in the larger discretization error and faster convergence, which is a trade-off. Numerical experiments on standard Bayesian models clarified that by choosing the interaction carefully, we can enjoy faster convergence compared to naive parallel-chain SGLD.

Our work can be extended in various ways. Theoretically, it is still unclear how much the convergence speed is improved when α is not small and the discretization error is crude, and we leave them to the future work. It is still unclear how to choose an appropriate skew-symmetric matrix and α theoretically, although it is important in practice. This also should be clarified in future work.

Acknowledgements

FF was supported by JST ACT-X Grant Number JPMJAX190R, IS was supported by KAKENHI 17H04693, and MS was supported by KAKENHI 17H00757.

References

Ahn, S., Shahbaba, B., and Welling, M. Distributed stochastic gradient mcmc. In *International conference on machine learning*, pp. 1044–1052, 2014.

Bakry, D., Barthe, F., Cattiaux, P., and Guillin, A. A simple proof of the poincaré inequality for a large class of probability measures including the log-concave case. *Electron. Commun. Probab*, 13(60-66):21, 2008.

Bakry, D., Gentil, I., and Ledoux, M. *Analysis and geometry of Markov diffusion operators*, volume 348. Springer Science & Business Media, 2013.

Baksalary, J. K. and Puntanen, S. An inequality for the trace of matrix product. *IEEE Transactions on Automatic Control*, 37(2):239–240, Feb 1992. ISSN 2334-3303. doi: 10.1109/9.121626.

Bishop, C. M. *Pattern recognition and machine learning*. springer, 2006.

Bolley, F. and Villani, C. Weighted csiszár-kullback-pinsker inequalities and applications to transportation inequalities. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 14, pp. 331–352, 2005.

Bolley, F., Guillin, A., and Malrieu, F. Trend to equilibrium and particle approximation for a weakly selfconsistent vlasov-fokker-planck equation. *ESAIM: Mathematical Modelling and Numerical Analysis*, 44(5):867–884, 2010.

Carlen, E. and Loss, M. Logarithmic sobolev inequalities and spectral gaps. In *Recent Advances in the Theory and Applications of Mass Transport. Contemp. Math., vol. 353*, pp. 53–60. Am. Math. Soc. Providence, 2004.

Cattiaux, P., Guillin, A., and Wu, L.-M. A note on talagrand's transportation inequality and logarithmic sobolev inequality. *Probability theory and related fields*, 148(1-2): 285–304, 2010.

Chen, C., Ding, N., and Carin, L. On the convergence of stochastic gradient mcmc algorithms with high-order integrators. In *Advances in Neural Information Processing Systems*, pp. 2278–2286, 2015.

Chen, C., Ding, N., Li, C., Zhang, Y., and Carin, L. Stochastic gradient mcmc with stale gradients. In *Advances in Neural Information Processing Systems*, pp. 2937–2945, 2016.

Chen, T., Fox, E., and Guestrin, C. Stochastic gradient hamiltonian monte carlo. In *International conference on machine learning*, pp. 1683–1691, 2014.

Chen, Y., Chen, J., Dong, J., Peng, J., and Wang, Z. Accelerating nonconvex learning via replica exchange langevin diffusion. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

Ding, N., Fang, Y., Babbush, R., Chen, C., Skeel, R. D., and Neven, H. Bayesian sampling using stochastic gradient thermostats. In *Advances in neural information processing systems*, pp. 3203–3211, 2014.

Dragomir, S. S. Some gronwall type inequalities and applications. 2002.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Duncan, A. B., Lelièvre, T., and Pavliotis, G. A. Variance reduction using nonreversible langevin samplers. *Journal of Statistical Physics*, 163(3):457–491, May 2016. ISSN 1572-9613. doi: 10.1007/s10955-016-1491-2. URL https://doi.org/10.1007/s10955-016-1491-2.

Duncan, A. B., Nüsken, N., and Pavliotis, G. A. Using perturbed underdamped langevin dynamics to efficiently sample from probability distributions. *Journal of Statistical Physics*, 169(6):1098–1131, Dec 2017. ISSN 1572-9613. doi: 10.1007/s10955-017-1906-8. URL https://doi.org/10.1007/s10955-017-1906-8.

Franke, B., Hwang, C.-R., Pai, H.-M., and Sheu, S.-J. The behavior of the spectral gap under growing drift. *Transactions of the American Mathematical Society*, 362(3): 1325–1350, 2010.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. J. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pp. 513–520, 2007.

Hu, Y., Wang, X., Gao, X., Gurbuzbalaban, M., and Zhu, L. Non-convex stochastic optimization via non-reversible stochastic gradient langevin dynamics. *arXiv preprint arXiv:2004.02823*, 2020.

Hwang, C.-R., Hwang-Ma, S.-Y., and Sheu, S.-J. Accelerating gaussian diffusions. *The Annals of Applied Probability*, pp. 897–913, 1993.

Hwang, C.-R., Hwang-Ma, S.-Y., Sheu, S.-J., et al. Accelerating diffusions. *The Annals of Applied Probability*, 15 (2):1433–1444, 2005.

Hwang, C.-R., Normand, R., and Wu, S.-J. Variance reduction for diffusions. *Stochastic Processes and their Applications*, 125(9):3522–3540, 2015.

Kaiser, M., Jack, R. L., and Zimmer, J. Acceleration of convergence to equilibrium in markov chains by breaking detailed balance. *Journal of Statistical Physics*, 168(2): 259–287, Jul 2017.

LeCun, Y. and Cortes, C. MNIST handwritten digit database. 2010.

Liu, C., Zhuo, J., Cheng, P., Zhang, R., and Zhu, J. Understanding and accelerating particle-based variational inference. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4082–4092, Long Beach, California, USA, 09–15 Jun 2019a. PMLR.

Liu, C., Zhuo, J., and Zhu, J. Understanding mcmc dynamics as flows on the wasserstein space. *arXiv preprint arXiv:1902.00282*, 2019b.

Liu, Q. and Wang, D. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances In Neural Information Processing Systems*, pp. 2378–2386, 2016.

Ma, Y.-A., Chen, T., and Fox, E. A complete recipe for stochastic gradient mcmc. In *Advances in Neural Information Processing Systems*, pp. 2917–2925, 2015.

Murphy, K. P. *Machine learning: a probabilistic perspective*. 2012.

Nusken, N. and Pavliotis, G. Constructing sampling schemes via coupling: Markov semigroups and optimal transport. *SIAM/ASA Journal on Uncertainty Quantification*, 7(1):324–382, 2019.

Patterson, S. and Teh, Y. W. Stochastic gradient riemannian langevin dynamics on the probability simplex. In *Advances in neural information processing systems*, pp. 3102–3110, 2013.

Raginsky, M., Rakhlin, A., and Telgarsky, M. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pp. 1674–1703, 2017.

Székely, G. J., Rizzo, M. L., et al. Testing for equal distributions in high dimension. 2004.

Vempala, S. and Wibisono, A. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems*, pp. 8094–8106, 2019.

Veretennikov, A. Y. On ergodic measures for mckean-vlasov stochastic equations. In Niederreiter, H. and Talay, D. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2004*, pp. 471–486, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-31186-7.

Villani, C. Optimal transportation, dissipative pde's and functional inequalities. In *Optimal transportation and applications*, pp. 53–89. Springer, 2003.

Villani, C. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 681–688, 2011.

Wibisono, A. Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem. In *Conference On Learning Theory*, pp. 2093–3027, 2018.

Xu, P., Chen, J., Zou, D., and Gu, Q. Global convergence of langevin dynamics based algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pp. 3122–3133, 2018.

Zhang, J., Zhang, R., and Chen, C. Stochastic particle-optimization sampling and the non-asymptotic convergence theory. *arXiv preprint arXiv:1809.01293*, 2018.