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.

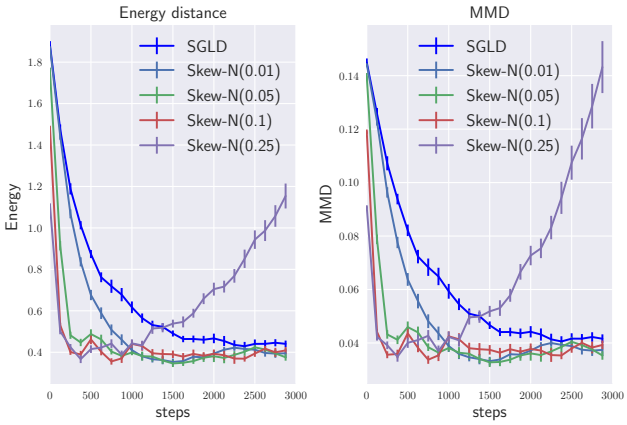
As for the numerical contributions, we believe that this work is the first attempt to apply the divergence-free drift method to the standard Bayesian models. Most existing work only took into account OU processes. In the next section, we numerically clarify that the divergence-free drift methods are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

6. Numerical experiments

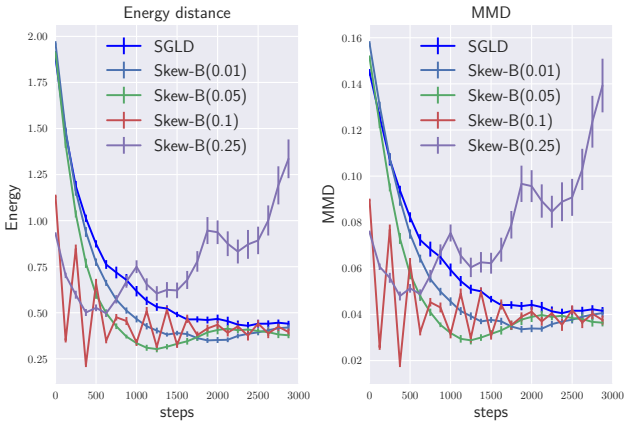
Detailed experimental settings are shown in Appendix M. From the theoretical analysis, we confirmed that there is a trade-off between discretization error and the convergence speed. Thus, it is natural to consider that if we tune the interaction α and J appropriately, we can improve the convergence speed 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

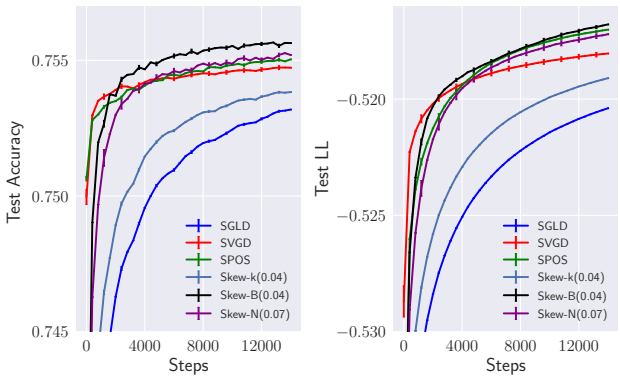


(a) Effect of different α with *Skew-N*

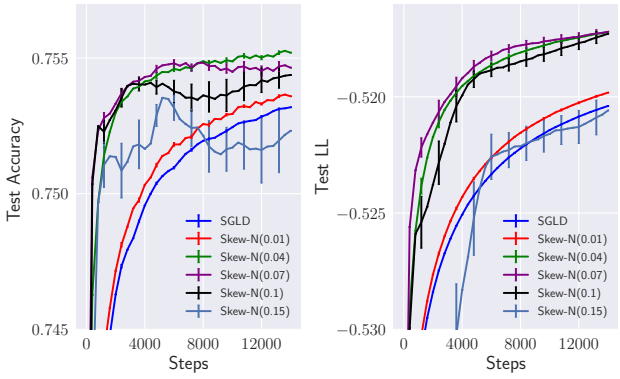


(b) Effect of different α with *Skew-B*

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)	$\mathbf{1024.33 \pm 1.85}$

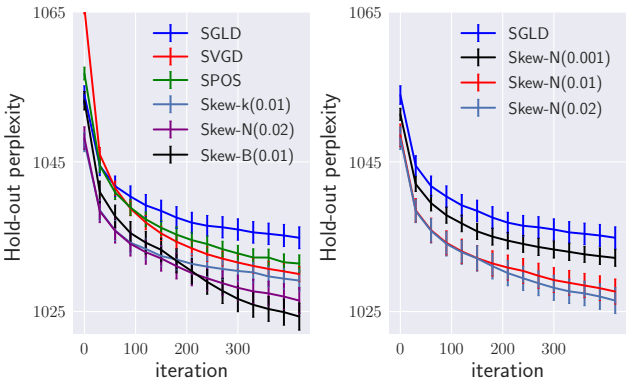


Figure 3. LDA experiments (Averaged over 10 trials)

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

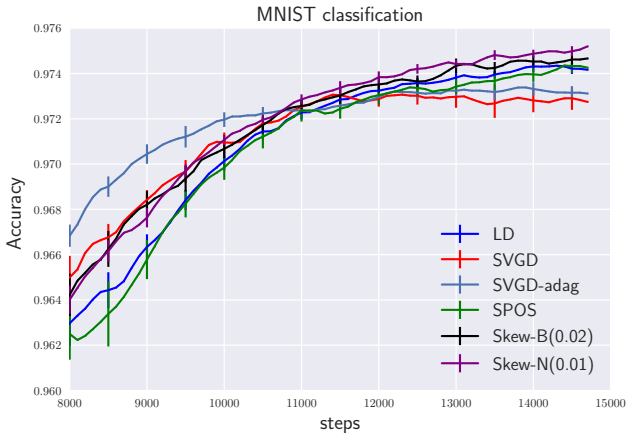


Figure 4. Mnist classification (Averaged over 10 trials)

(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.

A. Wasserstein distance

In this paper, we use the Wasserstein distance. Let us define the Wasserstein distance. Let (E, d) be a metric space (appropriate space such as Polish space) with σ field $\mathcal{A}$, where $d(\cdot,\cdot)$ is $\mathcal{A}\times\mathcal{A}$-measurable. Let μ, ν are probability measures on E, and $p \geq 1$. The Wasserstein distance of order p with cost function d between μ and ν is defined as

$$W_p^d(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \left(\int\int d(x,y)^p d\pi(x,y)\right)^{1/p}, \tag{37}$$

where $\Pi(\mu,\nu)$ is the set of all joint probability measures on $E\times E$ with marginals μ and ν. In this paper, we work on the space $\mathbb{R}^d$. As for the distance, we use the Euclidean distance, $\|\cdot\|$. For simplicity, we express the p-Wasserstein distance with the Euclidean distance as W_p. When we use the Wasserstein distance with which the cost function other than the Euclidean distance, we express it as W_p^d explicitly. The various properties of Wasserstein distance are summarized in (Villani, 2003).

We also define the Kullback leibler (KL) divergence as

$$\mathrm{KL}(\nu\|\mu) = \begin{cases} \int \log\frac{d\nu}{d\mu} d\nu, & \nu \ll \mu, \\ +\infty, & \text{otherwise.} \end{cases} \tag{38}$$

B. Preliminaries of SDE and Markov diffusion

B.1. Notations

Before going to the detailed analysis, we formally define the notations and the stochastic gradient Langevin dynamics algorithm. Our aim is to approximate the target distribution with the density $d\pi(x) \propto e^{-\beta U(x)}dx$ where

$$U(x) = \frac{1}{|\mathcal{Z}|}\sum_{i=1}^{|\mathcal{Z}|} u(x, z_i), \tag{39}$$

where z_i denotes the each data in some space $\mathbb{Z}$ and $|\mathcal{Z}|$ denotes the total number of the data points, $x \in \mathcal{X} \subset \mathbb{R}^d$ denotes the parameter of the model. For simplicity, we express the tuple of data points as $z = (z_1, \ldots, z_{|\mathcal{Z}|}) \in \mathbb{Z}^{\otimes|\mathcal{Z}|}$. The potential function $U(x)$ is the summation of $u : \mathbb{R}^d \times \mathbb{Z} \to \mathbb{R}$. The stochastic gradient Langevin dynamics algorithm is given as the recursion

$$X_{k+1} = X_k - hg(X_k, Q_{z,k}) + \sqrt{2h\beta^{-1}}\epsilon_k, \tag{40}$$

where $h \in \mathbb{R}^+$ is the 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 in some space Q following the probability $P_z(Q_{z,k})$. Following Raginsky et al. (2017), we consider g and $Q_{z,k}$ as a stochastic gradient oracle, which access the gradient of $U(X_k)$ at each iteration. Thus, a mapping, $g : \mathbb{R}^d \times \mathrm{Q} \to \mathbb{R}^d$ is the unbiased estimator;

$$\mathbb{E}_{P_z(Q_{z,k})}[g(X_k, Q_{z,k})] = \nabla U(X_k). \tag{41}$$

For example, this g expresses the stochastic access to the subset of data points $\{z_i\}$. Then, $\{Q_{z,k}\}_{k=0}^{\infty}$ is a sequence of i.i.d random variable of Q with law $P_z(Q_{z,k})$. We assume that $X_0, \epsilon_k, Q_{z,k}$ are independent of each other.

On the other hand, the continuous-time Langevin dynamics is written as

$$dX(t) = -\nabla U(X(t)) + \sqrt{2\beta^{-1}}dw(t), \tag{42}$$

where $w(t)$ denotes the standard Brownian motion in $\mathbb{R}^d$. The stationary measure of Eq.(42) is $d\pi(x) \propto e^{-\beta U(x)}dx$. Since Eq.(40) can be regarded as the discretization of Eq.(42), we will study the relation between them.

We denote the law of X_k induced by Eq.(40) as μ_{kh} and the law of X_t induced by Eq.(42) as ν_t.

B.2. Markov diffusion and generator

In this section, we introduce basic Markov diffusion operators. Given SDE

$$dX_t = -\nabla U(X_t)dt + \sqrt{2\beta^{-1}}dw(t), \tag{43}$$

then we denote the corresponding Markov semigroup as $P = \{P_t\}_{t>0}$ and the Kolmogorov operator as P_s which is defined as

$$P_s f(X_t) = \mathbb{E}[f(X_{t+s})|X(t)], \tag{44}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is some bounded test function in $\mathrm{L}^2(\mu)$. A property $P_{s+t} = P_s \circ P_t$ is called Markov property. A probability measure π is the stationary distribution when it satisfies for all measurable bounded function f and t

$$\int_{\mathbb{R}^d} P_t f d\pi = \int_{\mathbb{R}^d} f d\pi. \tag{45}$$

We denote the infinitesimal generator of the associated Markov group as $\mathcal{L}$ and we call it generator for simplicity. The linearity of the operators of P_t with the semigroup property indicates that $\mathcal{L}$ is the derivative of P_t as

$$\frac{1}{h}(P_{t+h} - P_t) = P_t\frac{1}{h}(P_h - Id) = \frac{1}{h}(P_h - Id)P_t \tag{46}$$

where Id is the identity map. And taking $h \to 0$, we have

$$\partial P_t = \mathcal{L}P_t = P_t\mathcal{L}. \tag{47}$$

From the Hille-Yoshida theory, there exists a dense linear subspace of $\mathrm{L}^2(\pi)$ on which $\mathcal{L}$ exists. We refer it as $\mathcal{D}(\mathcal{L})$.

If the Markov semigroup is associated with the SDE of Eq.(43), the generator can be written as

$$\begin{aligned}\mathcal{L}f(X_t) :&= \lim_{h\to 0^+} \frac{\mathbb{E}(f(X_{t+h})|X_t) - f(X_t)}{h} \\ &= \left(-\nabla U(X_t)\cdot\nabla + \beta^{-1}\Delta\right) f(X_t), \end{aligned} \tag{48}$$

where Δ is the Laplacian in the standard Euclidean space. The generator satisfies

$$\mathcal{L}1 = 0, \quad \int_{\mathbb{R}^d} \mathcal{L}f d\pi = 0. \tag{49}$$

B.3. Backward equation

The generator $\mathcal{L}$ is associated with the Kolmogorov backward equation,

$$\mathbb{E}(f(X_t)|X_0) = e^{t\mathcal{L}} f(X_0). \tag{50}$$

We can also express this identity as

$$\partial_t \mathbb{E}(f(X_t)|X_0) = \mathcal{L}\mathbb{E}(f(X_t)|X_0). \tag{51}$$

If we define $\phi(x,t) := \mathbb{E}(f(X_t)|X_0 = x)$ then we can rewrite above as

$$\partial_t \phi(x,t) = \mathcal{L}\phi(x,t) \quad \text{with} \quad \phi(x,0) = f(x) \tag{52}$$

and this is called the Kolmogorov backward equation. By the Taylor expansion, we obtain

$$\begin{aligned}\phi(x,t) &= \phi(x,0) + \partial_t\phi(x,t)|_{t=0}(t-0) + \mathcal{O}(t^2) \\ &= f(x) + t\mathcal{L}f(x) + \mathcal{O}(t^2). \end{aligned} \tag{53}$$

For more details, see Chen et al. (2015); Xu et al. (2018).

B.4. Fokker-Planck equation

We can write the evolution of the probabilty density $p(x)$. Given the initial density as $p_0(x)$ and express the density at time t as $p(x,t)$, then

$$\partial_t p(x,t) = \mathcal{L}^* p(x,t), \quad p(x,0) = p_0(x), \tag{54}$$

where $\mathcal{L}^*$ is the adjoint of $\mathcal{L}$.

C. Poincare and logarithmic Sobolev inequalities

Following Raginsky et al. (2017), we use the Poincare and logarithmic Sobolev inequalities to measure the speed of convergence to the stationary distribution. In this section, we review definitions and useful properties of them.

C.1. Poincare inequality

First, we define the Dirichlet form $\mathcal{E}(f)$ for all bounded function $f \in \mathcal{D}(\mathcal{L})$ where $\mathcal{D}(\mathcal{L})$ denotes the domain of $\mathcal{L}$ as

$$\mathcal{E}(f) := -\int_{\mathbb{R}^d} f\mathcal{L}f d\pi. \tag{55}$$

$\mathcal{E}(f) > 0$ is satisfied. If $\mathcal{L}$ is given by Eq.(48), by the partial integration, we have

$$\mathcal{E}(f) = -\int_{\mathbb{R}^d} f\mathcal{L}f d\pi = \frac{1}{\beta}\int_{\mathbb{R}^d} \|\nabla f\|^2 d\pi. \tag{56}$$

Also, we define a Dirichlet domain, $\mathcal{D}(\mathcal{E})$, which is the set of functions $f \in L^2(\pi)$ and satisfies $\mathcal{E}(f) < \infty$.

We say that π with $\mathcal{L}$ satisfies *a Poincare inequality* with a positive constant c if for any $f \in \mathcal{D}(\mathcal{E})$, π with $\mathcal{L}$ satisfies,

$$\begin{aligned}&\mathrm{Var}_\pi(f) \le c\mathcal{E}(f), \\ &\mathrm{Var}_\pi(f) := \int f^2 d\pi - \left(\int f d\pi\right)^2. \end{aligned} \tag{57}$$

This constant c is closely related to a spectral gap. If the smallest eigenvalue of $\mathcal{L}$, λ, is greater than 0, then it is called the spectral gap. If the spectral gap $\lambda > 0$ exists, then it is written as

$$\lambda := \inf_{f\in\mathcal{D}(\mathcal{E})} \left\{ \frac{\mathcal{E}(f)}{\int f^2 d\pi} : f \neq 0, \int f d\pi = 0 \right\}. \tag{58}$$

From this, a constant c which satisfies $c \ge 1/\lambda$, can also satisfy the Poincare inequality. To check the existence of the spectral gap, one approach is to use the Lyapunov function, which is developed by Bakry et al. (2008).

We can also express the Poincare inequality via chi divergence. Let us define the χ^2 divergence for $\mu \ll \pi$ as

$$\chi^2(\mu\|\pi) := \left\| \frac{d\mu}{d\pi} - 1 \right\|^2_{L^2_\pi} = \int_{\mathbb{R}^d} \left| \frac{d\mu}{d\pi} - 1 \right|^2 d\pi. \tag{59}$$

Then, we express the Poincare inequality with a constant c for all $\mu \ll \pi$ as

$$\chi^2(\mu\|\pi) \le c\,\mathcal{E}\left(\sqrt{\frac{d\mu}{d\pi}}\right). \tag{60}$$

C.2. Logarithmic Sobolev inequality

We say that π with $\mathcal{L}$ satisfies *the defective logarithmic Sobolev inequality* (LSI) with constant D_1, D_2 (we call this as LSI(D_1, D_2)) if for any f which is a integrable ($\int_{\mathbb{R}^d} f|\log f| d\pi < \infty$), π with $\mathcal{L}$ satisfies,

$$\begin{aligned}&\mathrm{Ent}_\pi(f^2) \le 2D_1\mathcal{E}(f) + D_2 \int f^2 d\pi, \\ &\mathrm{Ent}_\pi(f) := \int_{\mathbb{R}^d} f\log f d\pi - \int_{\mathbb{R}^d} f d\pi \log\left(\int_{\mathbb{R}^d} f d\pi\right). \end{aligned} \tag{61}$$

If $D_1 = 0$, the LSI is called tight. It is known that if there exists the poincare constant, $\rho \in \mathbb{R}^+$, $\mathrm{LSI}(D_1, D_2)$ becomes tight LSI, that is, $\mathrm{LSI}(\lambda_0, 0)$ with $\lambda_0 = D_1 + \rho(1 + D_2/2)$ (See Theorem 5.1.3 in (Bakry et al., 2013)). Hereinafter, we only consider the tight LSI and call it LSI simply.

We can express the LSI of π with $\mathcal{L}$ and a constant λ_0 by using the KL divergence. For all $\mu \ll \pi$,

$$\mathrm{KL}(\mu\|\pi) \leq 2\lambda\, \mathcal{E}\left(\sqrt{\frac{d\mu}{d\pi}}\right). \tag{62}$$

We can express the LSI by KL divergence and Fisher information. We define Carre du champ operator as

$$\Gamma(f, g) := \frac{1}{2}\left[\mathcal{L}(fg) - f\mathcal{L}g - g\mathcal{L}f\right], \tag{63}$$

and we also denote $\Gamma(f) := \Gamma(f, f)$. This satisfies

$$\int \Gamma(f, g) d\pi = -\int f\mathcal{L}g d\pi. \tag{64}$$

Then, for a function $\sqrt{f}$, the above definition can be rewritten as

$$\mathrm{Ent}_\pi(f) \leq \frac{\lambda}{2}\int \frac{\Gamma(f)}{f} d\pi. \tag{65}$$

Then we rewrite this by using the Fisher information. Before that we write the $\mathrm{Ent}_\nu(f)$ by using the KL divergence when f is the density function. Under the condition that $\int_{\mathbb{R}^d} f d\nu = 1$, $d\mu = f d\nu$,

$$\mathrm{Ent}_\nu(f) = \int_{\mathbb{R}^d} f \log f d\nu = \mathrm{KL}(\mu|\nu). \tag{66}$$

Then we can write the Fisher information as

$$I(\mu|\nu) = \int_{\mathbb{R}^d} \frac{\Gamma(f)}{f} d\nu = 4\int_{\mathbb{R}^d} \Gamma(\sqrt{f})\nu. \tag{67}$$

Then we can rewrite the LSI as

$$\mathrm{KL}(\mu|\nu) \leq \frac{\lambda}{2} I(\mu|\nu). \tag{68}$$

The important consequence of the LSI is that it implies the transportation cost inequality and the Poincare inequality. We say that π satisfies 2-transportation cost inequality (the T_2 inequality) if it satisfies for all $\mu \ll \pi$ with a constant λ,

$$W_2(\mu, \pi) \leq \sqrt{2\lambda\, \mathrm{KL}(\mu\|\pi)}. \tag{69}$$

For the details, see Villani (2008)

To check whether the logarithmic Sobolev inequality for the given Markov process, a famous approach is a Lyapunov function-based approach developed by Cattiaux et al. (2010).

C.3. Consequence of the inequality

From the above functional inequalities for measures, we obtain the following exponential convergence results. First, we state the consequence of Poincare inequality.

Theorem 6. *(Exponential convergence in the variance, Theorem 4.2.5 in (Bakry et al., 2013)) When π satisfies the Poincare inequality with a constant c, it implies the exponential convergence in the variance with a rate $2/c$, that is, for every bounded function $f : \mathbb{R}^d \to \mathbb{R}$,*

$$\mathrm{Var}_\pi(P_t f) \leq e^{-2t/c}\mathrm{Var}_\pi(f), \tag{70}$$

where $\mathrm{Var}_\pi(f) := \int_{\mathbb{R}^d} f^2 d\pi - \left(\int_{\mathbb{R}^d} f d\pi\right)^2$.

Next, we state the consequence of the logarithmic Sobolev inequality.

Theorem 7. *(Exponential convergence in the entropy, Theorem 5.2.1 in (Bakry et al., 2013)) When π satisfies the logarithmic Sobolev inequality with a constant c, it implies the exponential convergence in the entropy with a rate $2/c$, that is, for every bounded function $f \in L^1 : \mathbb{R}^d \to \mathbb{R}$,*

$$\mathrm{Ent}_\pi(P_t f) \leq e^{-2t/c}\mathrm{Ent}_\pi(f), \tag{71}$$

where $\mathrm{Ent}_\pi(f) := \int_{\mathbb{R}^d} f \log f d\pi - \int_{\mathbb{R}^d} f d\pi \log\left(\int_{\mathbb{R}^d} f d\pi\right)$. *By setting $f = \frac{d\mu}{d\pi}$, we get the exponential convergence of the KL divergence,*

$$\mathrm{KL}(\mu_t\|\pi) \leq e^{-2t/c}\mathrm{KL}(\mu_0\|\pi). \tag{72}$$

These exponential convergence play an central role to prove the convergence of the diffusion algorithm.

C.4. Functional inequalities and product measures

Here we review the important properties of the functional inequalities which are related to the product measures. These relations play important roles in our analysis.

Proposition 1. *(Stability under the product, proposition 4.3.1 in (Bakry et al., 2013)) If μ_1 and μ_2 on $\mathbb{R}^d$ satisfy the Poincare inequalities with a constant c_1 and c_2, then the product $\mu_1 \otimes \mu_2$ on $\mathbb{R}^d \otimes \mathbb{R}^d$ satisfies the Poincare inequality with the constant* $\max(c_1, c_2)$.

Proposition 2. *(Stability under the product, proposition 5.2.7 in (Bakry et al., 2013)) If μ_1 and μ_2 on $\mathbb{R}^d$ satisfy the logarithmic Sobolev inequalities with a constant c_1 and c_2, then the product $\mu_1 \otimes \mu_2$ on $\mathbb{R}^d \otimes \mathbb{R}^d$ satisfies the logarithmic Sobolev inequality with the constant* $\max(c_1, c_2)$.

D. Review of Raginsky et al. (2017)

D.1. Review of the bound of Raginsky et al. (2017)

In this section, we review the result of Raginsky et al. (2017), which establish the convergence of the SGLD algorithm in 2-Wasserstein sense.

Theorem 8. *(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*

$$W_2(\mu_{kh}, \pi) \leq Ckh + \sqrt{2\lambda_0 C'} e^{-\frac{kh}{\beta\lambda_0}}, \tag{73}$$
$$C_0 = \left(M^2\left(\kappa_0 + 2\left(1 \vee \tfrac{1}{m}\right)\left(b + 2B^2 + \tfrac{d}{\beta}\right)\right) + B^2\right),$$
$$C_1 = 6M^2(\beta C_0 + d),$$
$$\tilde{C}_0^2 = \left(12 + 8\left(\kappa_0 + 2b + \tfrac{2d}{\beta}\right)\right)\left(\beta C_0 + \sqrt{\beta C_0}\right),$$
$$\tilde{C}_1^2 = \left(12 + 8\left(\kappa_0 + 2b + \tfrac{2d}{\beta}\right)\right)\left(C_1 + \sqrt{C_1}\right),$$
$$C = \sqrt{\tilde{C}_0^2\sqrt{\delta} + \tilde{C}_1^2\sqrt{h}},$$
$$C' = \log\|p_0\|_\infty + \tfrac{d}{2}\log\tfrac{3\pi}{m\beta} + \beta\left(\tfrac{M\kappa_0}{3} + B\kappa_0^{1/2} + A + \tfrac{b\log 3}{2}\right),$$

We review how this bound was derived in Raginsky et al. (2017).In Raginsky et al. (2017), the error is decomposed to the convergence to the stationary by the exact continuous Langevin dynamics and the discretization error:

$$W_2(\mu_{kh}, \pi) \leq W_2(\mu_{kh}, \nu_{kh}) + W_2(\nu_{kh}, \pi). \tag{74}$$

Here, an important point is the discretization error $W_2(\mu_{kh}, \nu_{kh})$. First, the error in a sense of the relative entropy is bounded as

Lemma 3. *(Lemma 7 in (Raginsky et al., 2017)) For any k and h, we have*

$$\mathrm{KL}(\mu_{kh}|\nu_{kh}) \leq (C_0\beta\delta + C_1 h)kh, \tag{75}$$

where

$$C_0 = \left(M^2\left(\kappa_0 + 2\left(1 \vee \tfrac{1}{m}\right)\left(b + 2B^2 + \tfrac{d}{\beta}\right)\right) + B^2\right), \tag{76}$$

$$C_1 = 6M^2(\beta C_0 + d). \tag{77}$$

Since the relative entropy is not enough to bound the bias since it does not satisfy the triangle inequality, then from Bolley & Villani (2005), by using the weighted CKP inequality, we can bound W_2 distance by the relative entropy.

$$W_2(\mu, \nu) \leq C_\nu\left(\mathrm{KL}(\mu|\nu)^{1/2} + \left(\frac{\mathrm{KL}(\mu|\nu)}{2}\right)^{1/4}\right) \tag{78}$$

with

$$C_\nu = 2\inf_{\lambda>0}\left(\frac{1}{\lambda}\left(\frac{3}{2} + \log\int_{\mathbb{R}^d} e^{\lambda\|w\|^2} d\nu\right)\right)^{1/2}. \tag{79}$$

Next, we briefly introduce the LSI constant. As we mentioned in the main paper, the generator of the continuous dynamics is that

$$\mathcal{L}f(X_t) = \left(-\nabla U(X_t)\cdot\nabla + \beta^{-1}\Delta\right) f(X_t). \tag{80}$$

Instead of this original generator, in Raginsky et al. (2017), they used the generator

$$\mathcal{L}'f(X_t) = \left(-\beta\nabla U(X_t)\cdot\nabla + \Delta\right) f(X_t), \tag{81}$$

which satisfies $\beta\mathcal{L} = \mathcal{L}'$ to use the Lyapunov function-based approach of Cattiaux et al. (2010) to estimate LSI constant. Then $\mathcal{L}'$ corresponds to the SDE

$$dX_t = -\beta\nabla U(X_t) + \sqrt{2}dw(t), \tag{82}$$

which has the same stationary measure as $\mathcal{L}$. The difference is that the time scale is changed from $t \to \beta t$. In conclusion, we estimate LSI constant for $\mathcal{L}'$ then multiply $beta$ in the convergence, we recover the result of $\mathcal{L}$. We use this rescaling in the following.

D.2. The problem of the Theorem 1

Dimensional dependency: The important point is that C_ν depends on d as $d^{1/2}$ and C_0 and C_1 depends on d linealy. Thus, $\mathrm{KL}(\mu_{z,k}|\nu_{kh})$ depends on d linealy. In total, due to the weighted CKP inequality, the bound of Eq.(78) depends on d linearly. On the other hand, since we use the usual Euclidean distance for the cost fucntion in W_2 distance, we expect that W_2 depends on d as $d^{1/2}$ optimally. Thus, there is a gap.

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 an L_f-lipschitz function 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 8, which is based on the LSI inequality, then since the N-particle system is a dN-dimensonal, its discretization error depends on dN linearly. 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.

As for the problem of the LSI constant, please see the main paper.

E. Results of standard SGLD

E.1. Proposed Error control

Our proof strategy is almost as same as the Raginsky et al. (2017). That is, we decompose the distance into the convergence to the stationary by the exact continuous Langevin dynamics and the discretization error.

$$W_2(\mu_{kh}, \pi) \leq W_2(\mu_{kh}, \nu_{kh}) + W_2(\nu_{kh}, \pi). \tag{83}$$

We bound the convergence to the stationary; $W_2(\nu_h, \pi)$ in the same way as Raginsky et al. (2017). The difference is the approach for the discretization error.

For that purpose, we need the interpolated and one-time marginal dynamics for the analysis. See Appendix C.4 of Raginsky et al. (2017) or proof of theorem 6 in Zhang et al. (2018). First, we introduce the interpolated dynamics as

$$\bar{X}_t = X_0 - \int_0^t g(\bar{X}_{\lfloor s/h \rfloor h}, Q_{z,s})ds + \sqrt{2\beta^{-1}} \int_0^t dw(s), \tag{84}$$

where $Q_{z,s} := Q_{z,k}$ for $s \in [kh, (k+1)h)$. Then this $\bar{X}_t$ has the same probability law μ_t with the original discretized dynamics.

However, this is not a Markov process due to the stochasticity of the stochastic gradient oracle. Thus, we define the following one-time marginal dynamics which has the same marginal distribution μ_{kh},

$$V_t = X_0 - \int_0^t g_{z,s}(V(s))ds + \sqrt{2\beta^{-1}} \int_0^t dw(s) \tag{85}$$

with

$$g_{z,s}(v) := \mathbb{E}\left[g(\bar{X}_{\lfloor s/h \rfloor h}, Q_{z,s}) \mid \bar{X}_s = v \right]. \tag{86}$$

Note that this new SDE for V_t has a weak solution, thus the marginals of V_t is the same as that of $\bar{X}_t$ for all t. Under this setting, to bound $W_2(\mu_{kh}, \nu_{kh})$, we use the basic relation, which follows from the definition of the Wasserstein distance,

$$W_2^2(\nu_{kh}, \mu_{kh}) \leq \mathbb{E}\|X_{kh} - V_{kh}\|^2. \tag{87}$$

We consider the synchronous coupling for X_t and V_t then by using the Ito lemma,

$$X_t - V_t = -\int_0^t \left(\nabla U(X_s) - g_{z,s}(V(s))\right) ds. \tag{88}$$

By using this, we first apply Jensen inequality

$$\begin{aligned}
&\mathbb{E}\|X_t - V_t\|^2 = \mathbb{E}\left\| \int_0^t \nabla U(X_s) - g_{z,s}(V(s))ds \right\|^2 \\
&\leq t\mathbb{E} \int_0^t \|\nabla U(X_s) - g_{z,s}(V(s))\|^2 ds \\
&\leq t \sum_{k=0}^{\lfloor t/h \rfloor} \mathbb{E} \int_{kh}^{(k+1)h} \|\nabla U(X_s) - g_{z,s}(V(s))\|^2 ds.
\end{aligned} \tag{89}$$

Then, we consider the following decomposition

$$\begin{aligned}
&\|\nabla U(X_s) - g_{z,s}(V(s))\|^2 \\
&\leq \|\nabla U(X_s) - \nabla U(V_s) + \nabla U(V_s) - g_{z,s}(V(s))\|^2 \\
&\leq 3M^2\|X_s - V_s\|^2 + 3\|\nabla U(V_s) - \nabla U(V_{\lfloor s/h \rfloor h})\|^2 \\
&\qquad + 3\|\nabla U(V_{\lfloor s/h \rfloor h}) - g_{z,s}(V(s))\|^2
\end{aligned} \tag{90}$$

We will use the Gronwall inequality later, thus for that purpose we need to bound the second and third term in the above. These two terms are bounded in the proof of Lemma 7 in Appendix D of Raginsky et al. (2017), so we use it. They are bounded as

Lemma 4. *From Raginsky et al. (2017),*

$$\begin{aligned}
&\int_{kh}^{(k+1)h} \mathbb{E}\|\nabla U(V_s) - \nabla U(V_{\lfloor s/h \rfloor h})\|^2 ds \\
&\qquad \leq 6M^2 \left(2C_0 + \frac{d}{\beta}\right) h^2,
\end{aligned} \tag{91}$$

$$\int_{kh}^{(k+1)h} \mathbb{E}\|\nabla U(V_{\lfloor s/h \rfloor h}) - g_{z,s}(V(s))\|^2 ds \leq 2C_0 h\delta, \tag{92}$$

where the constant C_0 is shown in Theorem 8

Thus, we have

$$\begin{aligned}
&\mathbb{E}\|X_t - V_t\|^2 \\
&\leq 3M^2 t \int_0^t \|X_s - V_s\|^2 ds \\
&+ 6C_0 h(\lfloor t/h \rfloor)\delta t + 18M^2 \left(2C_0 + \frac{d}{\beta}\right) h^2(\lfloor t/h \rfloor)t.
\end{aligned} \tag{93}$$

For simplicity, we express the second and third term as

$$6C_0 h(\lfloor t/h \rfloor)\delta t + 18M^2 \left(2C_0 + \frac{d}{\beta}\right) h^2(\lfloor t/h \rfloor)t = Dt^2. \tag{94}$$

Then, we use the Gronwall inequality, from Dragomir (2002). We use the following type of Gronwall inequality. Let x, k are continuous and a, b are integrable functions on some interval, $J = [\alpha, \beta]$ and let b, k are non-negative on J. If

$$x(t) \leq a(t) + b(t) \int_\alpha^t k(s)x(s)ds, \quad t \in J, \tag{95}$$

then, we have

$$x(t) \leq a(t) + b(t) \int_\alpha^t a(s)k(s)\mathrm{e}^{\int_s^t b(r)k(r)dr} ds, \quad t \in J. \tag{96}$$

In our case, $\alpha = 0$, $a(t) = Dt^2$, $b(t) = 3M^2 t$ and $k(s) = 1$. Since we need to integrate about s, the second term of the Gronwall inequality is

$$\begin{aligned}
t \int_0^t Ds^2 \mathrm{e}^{\int_s^t r dr} ds &= t \int_0^t Ds^2 \mathrm{e}^{\frac{3M^2}{2}(t^2 - s^2)} ds \\
&= Dt\mathrm{e}^{\frac{3M^2}{2}t^2} \int_0^t s^2 \mathrm{e}^{-\frac{3M^2}{2}s^2} ds.
\end{aligned} \tag{97}$$

The integral can be evaluated by

$$\int_0^t s^2 \mathrm{e}^{-as^2} ds = -\frac{t\mathrm{e}^{-at^2}}{2a} + \int_0^t \frac{\mathrm{e}^{-as^2}}{2a} ds. \tag{98}$$

Thus, the second term is the error function. We can evaluate the asymptotic behavior of this function by using the property of the error function, but here for simplicity, we use the following relation,

$$\int_0^t \frac{\mathrm{e}^{-as^2}}{2a} ds < \int_0^\infty \frac{\mathrm{e}^{-as^2}}{2a} ds = \frac{1}{4a}\sqrt{\frac{\pi}{a}}. \tag{99}$$

Then, we have

$$\int_0^t s^2 \mathrm{e}^{-as^2} ds < -\frac{t\mathrm{e}^{-at^2}}{2a} + \frac{1}{4a}\sqrt{\frac{\pi}{a}}. \tag{100}$$

Thus, we get

$$\begin{aligned} &t\int_0^t Ds^2 \mathrm{e}^{\int_s^t r dr} ds \\ &< Dt\mathrm{e}^{\frac{3M^2}{2}t^2}\left(-\frac{t\mathrm{e}^{-\frac{3M^2}{2}t^2}}{3M^2} + \frac{1}{6M^2}\sqrt{\frac{2\pi}{3M^2}}\right). \end{aligned} \tag{101}$$

In conclusion, we have

$$\begin{aligned} &\mathbb{E}\|X_t - V_t\|^2 \\ &< Dt^2 + Dt\mathrm{e}^{\frac{3M^2}{2}t^2}\left(-\frac{t\mathrm{e}^{-\frac{3M^2}{2}t^2}}{3M^2} + \frac{1}{6M^2}\sqrt{\frac{2\pi}{3M^2}}\right) \end{aligned} \tag{102}$$

$$:= Dt\left(t + C_6\right). \tag{103}$$

Then we substitute $t = kh$, we get

$$\begin{aligned} Dt^2 &= 6C_0 hk\delta(kh) + 18M^2\left(2C_0 + \frac{d}{\beta}\right)h^2 k(kh) \\ &= \frac{6}{\beta}\left(C_1 kh^2 + \beta C_0 kh\delta\right)kh \qquad (104) \\ &:= C_7 (kh)^2, \qquad (105) \end{aligned}$$

where C_1 is shown in Theorem 8. We transformed in this way so that we can easily compare our result with the Theorem 8.

Then we get

$$W_2^2(\nu_{kh}, \mu_{kh}) \le \mathbb{E}\|X_{kh} - V_{kh}\|^2 < C_7\left(kh + C_6\right)kh \tag{106}$$

and

$$C_6 := \mathrm{e}^{\frac{3M^2}{2}(kh)^2}\left(-\frac{(kh)\mathrm{e}^{-\frac{3M^2}{2}(kh)^2}}{3M^2} + \frac{1}{6M^2}\sqrt{\frac{2\pi}{3M^2}}\right), \tag{107}$$

$$C_7 := \frac{6}{\beta}\left(C_1 h + \beta C_0 \delta\right). \tag{108}$$

Note that C_7 depends on d linearly, and C_6 does not depend on d.

Thus our bound of $W_2^2(\nu_{kh}, \mu_{kh})$ depends on d by $d^{1/2}$, which is crucial for the analysis of the ensemble sampling.

Finally, we briefly mension about the bound of the convergence to the stationary: $W_2(\nu_{kh}, \pi)$. This approach is same as Raginsky et al. (2017).

From Appendix Raginsky et al. (2017), π satisfies the LSI with the constant λ_0 (We discuss this constant in the Appendix K.1). Since LSI implies the exponential convergence of the relative entropy (Bakry et al., 2013),

$$\mathrm{KL}(\nu_{kh}|\pi) \le \mathrm{KL}(\nu_0|\pi)e^{-2\frac{kh}{\beta\lambda_0}}, \tag{109}$$

where we used the rescaled generator and the LSI constant, see Eq.(81). Also, the LSI implies T_2 inequality (Bakry et al., 2013),

$$W_2(\nu_{kh}|\pi) \le \sqrt{2\lambda_0 \mathrm{KL}(\nu_{kh}|\pi)}. \tag{110}$$

Combine with these relations, we get

$$W_2(\nu_{kh}|\pi) \le \sqrt{2\lambda_0 \mathrm{KL}(\nu_0|\pi)}e^{-\frac{kh}{\beta\lambda_0}}. \tag{111}$$

Thus, we conclude that

$$\begin{aligned} &W_2(\mu_{kh}, \pi) \\ &< \sqrt{C_7\left(kh + C_6\right)kh} + \sqrt{2\lambda_0 \mathrm{KL}(\nu_{kh}|\pi)}e^{-\frac{kh}{\beta\lambda_0}} \end{aligned} \tag{112}$$

with

$$C_6 := \mathrm{e}^{\frac{3M^2}{2}(kh)^2}\left(-\frac{(kh)\mathrm{e}^{-\frac{3M^2}{2}(kh)^2}}{3M^2} + \frac{1}{6M^2}\sqrt{\frac{2\pi}{3M^2}}\right), \tag{113}$$

$$C_7 := \frac{6}{\beta}\left(C_1 h + \beta C_0 \delta\right). \tag{114}$$

E.2. Comparison of discretization error

As we discussed in the main paper, our bound of $W_2(\nu_{kh}, \mu_{kh})$ is $\mathcal{O}(d^{1/2})$ and this is optimal since our cost function for the W_2 distance is Euclidean distance. On the other hand, the bound of Raginsky et al. (2017) is $\mathcal{O}(d)$.

On the other hand, with respect to the kh, ours is much worse than that of Raginsky et al. (2017). Due to the Cauchy-Schwartz inequality and the Gronwall inequality, ours has $\mathcal{O}(khe^{(kh)^2})$, which is far from satisfactory. On the other hand, that of Raginsky et al. (2017) is $\mathcal{O}(kh)$. This gap should be resolved in future work.

F. Improved discretizatoin error

In this section, we derive the discretization error based on Vempala & Wibisono (2019).

Theorem 9. *Under Assumptions 1 to 5 and an additional assumption* $\frac{m}{4M^2} \leq 1$, *for any* $k \in \mathbb{N}$ *and any* $h \in (0, \frac{1}{4\sqrt{2}M^2\lambda} \wedge \frac{m}{4M^2})$ *obeying* $kh \geq 1$ *and* $\beta m \geq 2$, μ_{kh}, *which is induced by Eq.(1), satisfies*

$$W_2(\mu_{kh}, \pi) \leq \sqrt{2\lambda e^{-\frac{kh}{\beta\lambda}}\mathrm{KL}(\mu_0|\pi) + \frac{8\lambda^2}{3}\left(\frac{2hdM}{\beta}(m+4M) + 8C_0\delta\right)}. \tag{115}$$

where λ *is the LSI constant for* π *with* $\mathcal{L}'$.

Note that the smaller the LSI constant is, the smaller 2-W distance we obtain. The third term corresponds to the error induced by the stochastic gradient and the second term corresponds to the discretization error. This shows much better behavior with respect to the step size compared to Theorem 2 in the main paper. However, we need to control the step size so that $h \in (0, \frac{1}{4\sqrt{2}M^2\lambda} \wedge \frac{m}{4M^2})$ holds which contains the LSI constant. So we need the information about the LSI constant to tune the step size. Additional remark is that due to the discretization error, the convergence rate of the exponential function is $\frac{1}{2\lambda}$, while the rate of that function in Raginsky et al. (2017) is $\frac{1}{\lambda}$

Proof. To prove this theorem, we use the lemma 3 in Vempala & Wibisono (2019),

Lemma 5. *(lemma 3 in Vempala & Wibisono (2019)) Suppose* π *satisfies LSI with constant* $1/\lambda > 0$ *and is* M*-smooth. If* $0 < h \leq \frac{1}{4\lambda M^2}$ *then the Langevin algorithm converges*

$$\mathrm{KL}(\mu_{(k+1)h}|\pi) \leq e^{-h/\lambda}\mathrm{KL}(\mu_{kh}|\pi) + 6h^2dM^2. \tag{116}$$

Since this lemma considers full gradient, we need to extend it to stochastic gradient. This is straight forward. We get the following lemma.

Lemma 6. *Suppose* π *satisfies LSI with constant* $\lambda > 0$ *and is* M*-smooth. If* $h \in (0, \frac{1}{4\sqrt{2}M^2\lambda} \wedge \frac{m}{4M^2})$ *obeying* $kh \geq 1$ *and* $\beta m \geq 2$, *then the Langevin algorithm converges*

$$\mathrm{KL}(\mu_{(k+1)h}|\pi) \leq e^{-\frac{1}{\lambda}h}\mathrm{KL}(\mu_{kh}|\pi) + \left(\frac{2h^2dM}{\beta}(m+4M) + 8hC_0\delta\right). \tag{117}$$

The proof of this lemma is given in Appendix F.1. Then we apply for k-steps, we get

$$\mathrm{KL}(\mu_k|\pi) \leq e^{-\frac{1}{\lambda}kh}\mathrm{KL}(\mu_0|\pi) + \frac{1}{1-e^{-\frac{1}{\lambda}h}}\left(\frac{2h^2dM}{\beta}(m+4M) + 8hC_0\delta\right). \tag{118}$$

Then $1 - e^{-c} \geq \frac{3}{4}c$ for $0 < c = \frac{h}{\lambda} \leq \frac{1}{4}\sqrt{2}$, which holds since $h \leq \frac{1}{4\sqrt{2}M^2\lambda}$ and $\frac{1}{\lambda} \leq M$. Then we get

$$\mathrm{KL}(\mu_k|\pi) \leq e^{-\frac{1}{\lambda}kh}\mathrm{KL}(\mu_0|\pi) + \frac{4\lambda}{3}\left(\frac{2hdM}{\beta}(m+4M) + 8C_0\delta\right). \tag{119}$$

Finally, from the transportation inequality, we get

$$W_2(\mu_k, \pi) \leq \sqrt{2\lambda\,\mathrm{KL}(\mu_k|\pi)} \leq \sqrt{2\lambda e^{-\frac{1}{\lambda}kh}\mathrm{KL}(\mu_0|\pi) + \frac{8\lambda^2}{3}\left(\frac{2hdM}{\beta}(m+4M) + 8C_0\delta\right)}. \tag{120}$$

□

F.1. Proof of lemma 6

Proof. The proof is almost similar to the original proof of lemma 3 in Vempala & Wibisono (2019). However, in the original proof, a full gradient ∇U is used so we replace it to the stochastic gradient. We use the notations in Appendix E.1.

First, lemma 11 in Vempala & Wibisono (2019) is modified to

$$\mathbb{E}_\pi\|\nabla U\|^2 \leq \frac{dM}{\beta}. \tag{121}$$

This is eqsily confirmed by the definition of π and using the integration by parts.

Then lemma 12 in Vempala & Wibisono (2019) is modified to

$$\mathbb{E}_\rho\|\nabla U\|^2 \leq 4M^2\lambda\mathrm{KL}(\rho|\pi) + \frac{2dM}{\beta}, \tag{122}$$

for any integrable μ. Recall the interpolated dynamics as

$$\bar{X}_t = X_0 - \int_0^t g(\bar{X}_{\lfloor s/h \rfloor h}, Q_{z,s})ds + \sqrt{2\beta^{-1}}\int_0^t dw(s), \tag{123}$$

where $Q_{z,s} := Q_{z,k}$ for $s \in [kh, (k+1)h)$. Then this $\bar{X}_t$ has the same probability law μ_t with the original discretized dynamics. When we focus on the step k, we consider the following SDE for $t \in (kh, (k+1)h]$

$$d\bar{X}_t = -g(\bar{X}_k, Q_{z,k})dt + \sqrt{2\beta^{-1}}dw(t), \tag{124}$$

and the solution to this SDE at time $t = h$ is given by

$$\bar{X}_{(k+1)h} = \bar{X}_k - g(\bar{X}_k, Q_{z,k})h + \sqrt{2\beta^{-1}}\epsilon, \tag{125}$$

We would like to derive the continuity equation correspond to Ew.(124). Following Vempala & Wibisono (2019), we express $\bar{X}_t$ as x_t and $\bar{X}_k$ as x_0 for simplicity. Let $\rho_{0t}(x_0, x_t)$ denote the joint distribution of (x_0, x_t) conditioned on the stochastic gradient oracle at step k. The conditional and marginal relations are written as

$$\rho_{0t}(x_0, x_t) = \rho_0(x_0)\rho_{t|0}(x_t|x_0) = \rho_t(x_t)\rho_{0|t}(x_0|x_t). \tag{126}$$

Then, conditioned on x_0, the conditional density $\rho_{t|0}(x_t|x_0)$ follows the FP equation

$$\begin{aligned}&\frac{\partial \rho_{t|0}(x_t|x_0)}{\partial t}\\ &= \nabla \cdot (\rho_{t|0}(x_t|x_0) g(\bar{X}_k, Q_{z,k})) + \beta^{-1}\Delta \rho_{t|0}(x_t|x_0),\end{aligned} \tag{127}$$

Then following Vempala & Wibisono (2019), to derive the evolution of ρ_t, we take the expection over $\rho_0(x_0)$

$$\begin{aligned}&\frac{\partial \rho_t(x)}{\partial t}\\ &= \int_{\mathbb{R}^d} \frac{\partial \rho_{t|0}(x_t|x_0)}{\partial t}\rho_0(x_0)dx_0\\ &= \nabla \cdot (\rho_t(x_t)\mathbb{E}_{\rho_{0|t}}[g(x_0, Q_{z,k})|x_t = x]) + \beta^{-1}\Delta\rho_t(x).\end{aligned} \tag{128}$$

and take the expectation over $Q_{z,k}$, we have

$$\begin{aligned}&\frac{\partial \mu_t(x)}{\partial t}\\ &= \nabla \cdot \mathbb{E}_{P(Q_{z,k})}[(\rho_t(x_t)\mathbb{E}_{\rho_{0|t}}[g(x_0, Q_{z,k})|x_t = x]]) + \beta^{-1}\Delta\mu_t(x).\end{aligned} \tag{129}$$

Recall that

$$\frac{\partial \mathrm{KL}(\mu_t|\pi)}{\partial t} = \int \frac{\partial \mu_t(x)}{\partial t} \ln \frac{\mu_t}{\pi} dx. \tag{130}$$

Then following Vempala & Wibisono (2019), we get

$$\begin{aligned}&\frac{\partial \mathrm{KL}(\mu_t|\pi)}{\partial t}\\ &\le -\frac{3}{4}I(\mu_t|\pi)\\ &\quad + \mathbb{E}_{\rho_{0t}}\mathbb{E}_{P(Q_{z,k})}[\|\nabla U(\bar{X}_t) - g(\bar{X}_k, Q_{z,k})\|^2],\end{aligned} \tag{131}$$

where $t \in (kh, (k+1)h]$ and

$$X_t = \bar{X}_k - t\nabla U(\bar{X}_k) + \sqrt{2t\beta^{-1}}\epsilon. \tag{132}$$

Then, following the proof of Vempala & Wibisono (2019), we need to upper bound

$$\mathbb{E}_{P(Q_{z,k})}\|\nabla U(X_t') - g(\bar{X}_k, Q_{z,k})\|^2. \tag{133}$$

To bound this, we consider the decomposition

$$\begin{aligned}&\mathbb{E}_{P(Q_{z,k})}\|\nabla U(X_t) - g(\bar{X}_k, Q_{z,k})\|^2\\ &\le \mathbb{E}_{P(Q_{z,k})}\|\nabla U(X_t) - \nabla U(\bar{X}_k) + \nabla U(\bar{X}_k) - g(\bar{X}_k, Q_{z,k})\|^2\\ &\le 2\|\nabla U(X_t) - \nabla U(\bar{X}_k)\|^2 + 2\mathbb{E}_{P(Q_{z,k})}\|\nabla U(\bar{X}_k) - g(\bar{X}_k, Q_{z,k})\|^2\\ &\le 2M^2\|X_t - \bar{X}_k\|^2 + 2\mathbb{E}_{P(Q_{z,k})}\|\nabla U(\bar{X}_k) - g(\bar{X}_k, Q_{z,k})\|^2\\ &\le 2M^2\|X_t - \bar{X}_k\|^2 + 4C_0\delta,\end{aligned} \tag{134}$$

in the last line, we used Eq.(92). Then, we can substitute above inequality into the original proof of Vempala & Wibisono (2019) and $h \in (0, 1 \wedge \frac{m}{4M^2})$ obeying $kh \ge 1$ and $\beta m \ge 2$, we get

$$\begin{aligned}&\frac{d}{dt}\mathrm{KL}(\mu_t|\pi)\\ &\le -\tfrac{3}{4}I(\mu_t|\pi) + 8t^2M^4\lambda\mathrm{KL}(\mu_t|\pi) + \tfrac{4t^2dM^3}{\beta} + \tfrac{4tdM^2}{\beta} + 4C_0\delta.\end{aligned} \tag{135}$$

For simplicity, we assume that $h \in (0, \frac{m}{4M^2})$ and $\frac{m}{4M^2} < 1$, then we get

$$\begin{aligned}&\frac{d}{dt}\mathrm{KL}(\mu_t|\pi)\\ &\le -\tfrac{3}{4}I(\mu_t|\pi) + 8t^2M^4\lambda\mathrm{KL}(\mu_t|\pi) + \tfrac{tdM}{\beta}(m+4M) + 4C_0\delta.\end{aligned} \tag{136}$$

Then by using $t \in (kh, (k+1)h]$, we get

$$\begin{aligned}&\mathrm{KL}(\mu_{k+1}|\pi)\\ &\le e^{-\frac{3}{2\lambda}h}\left(1 + 16h^3M^4\lambda\right)\mathrm{KL}(\mu_k|\pi)\\ &\quad + e^{-\frac{3}{2\lambda}h}\left(\frac{2h^2dM}{\beta}(m+4M) + 8hC_0\delta\right).\end{aligned} \tag{137}$$

If $h \in (0, \frac{1}{4\sqrt{2}M^2\lambda})$, we get

$$\begin{aligned}&\mathrm{KL}(\mu_{k+1}|\pi)\\ &\le e^{-\frac{1}{\lambda}h}\mathrm{KL}(\mu_k|\pi) + \left(\frac{2h^2dM}{\beta}(m+4M) + 8hC_0\delta\right).\end{aligned} \tag{138}$$

□

F.2. Bias of the improved discretization error

We can derive the bias for the test function f with Lipschitz constant L_f as

$$\begin{aligned}&\left|\mathbb{E}f(X_k) - \int_{\mathbb{R}^d} f d\pi\right|\\ &\le L_f W_2(\mu_k, \pi)\\ &\le L_f\sqrt{2\lambda e^{-\frac{kh}{\beta\lambda}}\mathrm{KL}(\mu_0|\pi) + \lambda C_8},\end{aligned} \tag{139}$$

where we set

$$C_8 := \frac{8}{3}\left(\frac{2hdM}{\beta}(m+4M) + 8\bar{C}_0\bar{\delta}\right). \tag{140}$$

for simplicity

G. N-parallel and interacting dynamics and its theoretical results

In this section, we review the theoretical results of the N-parallel chain dynamics. Since there is no interaction between 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. This means that

$$\left| \mathbb{E} \frac{1}{N} \sum_{n=1}^{N} f(X_k^{(n)}) - \int_{\mathbb{R}^d} f d\pi \right| = \left| \mathbb{E} f(X_k) - \int_{\mathbb{R}^d} f d\pi \right|. \tag{141}$$

Thus, the bias of the naive parallel chain SGLD is the same as the standard SGLD.

H. Divergence-free accelerated dynamics and its theoretical results

H.1. The stationary distribution of the proposed dynamics

Here, we briefly check how the interaction term affects the stationary distribution. When π satisfies

$$\int \mathcal{L} f d\pi = 0, \tag{142}$$

then we say π is the stationary (invariant) measure of the Markov semigroup P_t with the generator $\mathcal{L}$ (Bakry et al., 2013).

Then, when we add the divergence-free term as the interaction,

$$\mathcal{L}_\alpha = \mathcal{L} + \alpha \langle \gamma, \nabla \rangle \tag{143}$$

with $\nabla \cdot (\gamma \pi) = 0$. Thus, from partial integration,

$$\int \mathcal{L}_\alpha f d\pi = \int \mathcal{L} f d\pi = 0, \tag{144}$$

holds. Thus, the divergence-free additional term never changes the stationary measure.

H.2. The effect of the divergence-free drift

First, we discuss the effect of the divergence-free drift on the dissipative condition and the smoothness condition since these conditions directly affect the convergence and discretization error.

H.3. Dissipative condition

In this section, we analyze how the dissipative condition $x \cdot \nabla u(x,z) \geq m\|x\|^2 - b$ will be modified by the divergence-free drift. This analysis is crucial in our setting since the constants m, b directly influence the convergence speed. Since the original dissipative condition is the lower bound of the inner product

$$x \cdot \nabla u(x,z) \geq m\|x\|^2 - b, \tag{145}$$

what we will analyze is to estimate the lower bound of

$$X_t^{\otimes N} \cdot (I + \alpha J) \nabla U^{\otimes N}(X_t^{\otimes N}), \tag{146}$$

where $\cdot$ is the inner product of dN-dimensional euclidean space and $\nabla U_\alpha(x^{\otimes N}) := \nabla U^{\otimes N}(x^{\otimes N}) + \alpha \nabla U^{\otimes N}(x^{\otimes N})$. Then,

Lemma 7. *Let $x^{\otimes N} \in \mathbb{R}^{dN}$ and under the assumprions 1 to 6, we have*

$$x^{\otimes N} \cdot \nabla U_\alpha(x^{\otimes N}) \geq m\|x^{\otimes N}\|^2 - bN. \tag{147}$$

Proof. First of all, we study how the dissipative condition will change under the naive parallel chain SGLD. Recall that $X_t^{\otimes N} = (X_t^{(1)}, \dots, X_t^{(N)})^\top \in \mathbb{R}^{dN}$ and $\nabla U^{\otimes N}(X_t^{\otimes N}) := (\nabla U(X_t^{(1)}), \dots, \nabla U(X_t^{(N)}))^\top$. Based on this definition,

$$\begin{aligned} X_t^{\otimes N} \cdot \nabla U^{\otimes N}(X_t^{\otimes N}) &= \sum_{n=1}^{N} X_t^{(n)} \cdot U(X_t^{(n)}) \\ &\overset{(i)}{\geq} \sum_{n=1}^{N} (m\|X_t^{(n)}\|^2 - b) \\ &\overset{(ii)}{=} m\|X_t^{\otimes N}\|^2 - Nb, \end{aligned} \tag{148}$$

and in (i), we used the dissipative condition for each $X_t^{(n)}$ and the norm is the euclidean norm in $\mathbb{R}^d$ and in (ii), the norm is the euclidean norm in $\mathbb{R}^{dN}$. Thus, we can observe that the constants of m, b changed to m, Nb.

Before going to the analysis of the proposed method, we reformulate the dissipative condition as trace inequality. To do that, we use the trace form of the inner product, that is,

$$x \cdot \nabla u(x,z) = \mathrm{Tr}(\nabla u(x,z) x^\top), \tag{149}$$

thus, the condition of the naive parallel chain can be written as

$$\mathrm{Tr}(\nabla U^{\otimes N}(X_t^{\otimes N}) X_t^{\otimes N \top}) \geq m\|X_t^{\otimes N}\|^2 - Nb. \tag{150}$$

To bound the trace, we use the following theorem in Baksalary & Puntanen (1992)

Theorem 10. *(Upper and lower bound of the trace Baksalary & Puntanen (1992)) Let A be any real value $N \times N$ matrix, $\bar{A} = \frac{A + A^\top}{2}$, and B is any non-negative definite $N \times N$ matrix. Furthur, let $\rho_*(\bar{A})$ and $\rho^*(\bar{A})$ be the smallest and the largest negative eigenvalue of $\bar{A}$ if they exist,*

otherwise $\rho_*(\bar{A}) = \rho^*(\bar{A}) = 0$*, and* $\omega_*(\bar{A})$ *and* $\omega^*(\bar{A})$ *be the smallest and the largest positive eigenvalue of* $\bar{A}$ *if they exist, otherwise* $\omega_*(\bar{A}) = \omega^*(\bar{A}) = 0$*. Then*

$$\begin{aligned}\big(\rho_*(\bar{A}) + \omega_*(\bar{A})\big)\,\mathrm{Tr}(B) \leq &\mathrm{Tr}(AB) \\ &\leq \big(\rho^*(\bar{A}) + \omega^*(\bar{A})\big)\,\mathrm{Tr}(B).\end{aligned} \tag{151}$$

To apply Theorem 10, we introduce the following matrix

$$\Theta := \nabla U^{\otimes N}(X_t^{\otimes N}) X_t^{\otimes N,\top} \tag{152}$$

which satisfies

$$\mathrm{Tr}\Theta := X_t^{\otimes N} \cdot \nabla U^{\otimes N}(X_t^{\otimes N}) \geq m\|X_t^{\otimes N}\|^2 - Nb. \tag{153}$$

Then, what we will analyze is to estimate the lower bound of

$$\begin{aligned}&X_t^{\otimes N} \cdot (I + \alpha J)\,\nabla U^{\otimes N}(X_t^{\otimes N}) \\ &= \mathrm{Tr}((I + \alpha J)\,\nabla U^{\otimes N} X_t^{\otimes N\top}) \\ &= \mathrm{Tr}((I + \alpha J)\,\Theta).\end{aligned} \tag{154}$$

From the elementary calculus of the matrix, we can say that Θ is the rank 1 matrix of which eigenvalues are $X_t^{\otimes N} \cdot \nabla U^{\otimes N}(X_t^{\otimes N})$ and 0. To apply Theorem 10, the matrix B must be non-negative definite. Thus, we consider the several settings, which depend on the sign of $X_t^{\otimes N} \cdot \nabla U^{\otimes N}(X_t^{\otimes N})$.

- When $m\|X_t^{\otimes N}\|^2 - Nb \geq 0$, that is, outside the ball of which radius $R^2 = \frac{Nb}{m}$. Then we set $A := I + \alpha J$ and $B := \Theta$ which satisfies the assumption of Theorem 10. From the left handside of Eq.(151), we get

$$\mathrm{Tr}((I + \alpha J)\Theta) \geq m\|X_t^{\otimes N}\|^2 - Nb. \tag{155}$$

- When $0 \geq m\|X_t^{\otimes N}\|^2 - Nb$ and $X_t^{\otimes N} \cdot \nabla U^{\otimes N}(X_t^{\otimes N}) > 0$. Then we assume that $A := I + \alpha J$ and $B := \Theta$ which satisfies the assumption of Theorem 10. From the left handside of Eq.(151), we get

$$\mathrm{Tr}((I + \alpha J)\Theta) \geq m\|X_t^{\otimes N}\|^2 - Nb. \tag{156}$$

- When $0 \geq m\|X_t^{\otimes N}\|^2 - Nb$ and $0 \geq X_t^{\otimes N} \cdot \nabla U^{\otimes N}(X_t^{\otimes N})$. Then we assume that $A := I + \alpha J$ and $B := -\Theta$ which satisfies the assumption of Theorem 10. From the right handside of Eq.(151), we get

$$-\mathrm{Tr}((I + \alpha J)\Theta) \leq -(m\|X_t^{\otimes N}\|^2 - Nb). \tag{157}$$

Combining above three cases, we have

$$\mathrm{Tr}((I + \alpha J)\Theta) \geq m\|X_t^{\otimes N}\|^2 - Nb. \tag{158}$$

□

H.4. Smoothness condition

In this section, we study how the smoothness condition

$$\|\nabla u(x, z) - \nabla u(y, z)\| \leq M\|x - y\|, \tag{159}$$

changes by using the non-reversible drift. Recall that the non-reversible drift is

$$(I + \alpha J)\,\nabla U^{\otimes N}(X_t^{\otimes N}), \tag{160}$$

where $\cdot$ is the inner product of dN-dimensional euclidean space and $\nabla U_\alpha(x^{\otimes N}) := \nabla U^{\otimes N}(x^{\otimes N}) + \alpha J \nabla U^{\otimes N}(x^{\otimes N})$. Then,

Lemma 8. *Let* $x^{\otimes N}, y^{\otimes N} \in \mathbb{R}^{dN}$ *and under the assumprions 1 to 6, we have*

$$\|\nabla U_\alpha(x^{\otimes N}) - \nabla U_\alpha(y^{\otimes N})\| \leq M(1 + \alpha)\|x^{\otimes N} - y^{\otimes N}\|. \tag{161}$$

Proof. First of all, for the drift of the naive parallel chain SGLD, following smoothness condition holds,

$$\begin{aligned}&\|\nabla U^{\otimes N}(x_t^{\otimes N}) - \nabla U^{\otimes N}(y_t^{\otimes N})\|^2 \\ &\overset{(i)}{=} \sum_{n=1}^N \|\nabla U(x_t^{(n)}) - \nabla U(y_t^{(n)})\|^2 \\ &\overset{(ii)}{\leq} \sum_{n=1}^N M^2\|x_t^{(n)} - y_t^{(n)}\|^2 \\ &\overset{(iii)}{=} M^2\|x_t^{\otimes N} - y_t^{\otimes N}\|^2,\end{aligned} \tag{162}$$

and in (i), we changed the norm from $\mathbb{R}^{Nd}$ to $\mathbb{R}^d$ and in (ii), we used the smoothness condition in $\mathbb{R}^d$, and in (iii), we changed the norm from $\mathbb{R}^d$ to $\mathbb{R}^{Nd}$. Based on this basic relation, we upper bound the drift function of the non-reversible chain by using the matrix norm,

$$\begin{aligned}&\|\nabla U_\alpha(x^{\otimes N}) - \nabla U_\alpha(y^{\otimes N})\| \\ &= \|\,(I + \alpha J)\,\big(\nabla U^{\otimes N}(x_t^{\otimes N}) - \nabla U^{\otimes N}(y_t^{\otimes N})\big)\,\| \\ &\overset{(i)}{\leq} \|I + \alpha J\|_2 \|\nabla U^{\otimes N}(x_t^{\otimes N}) - \nabla U^{\otimes N}(y_t^{\otimes N})\| \\ &\overset{(ii)}{\leq} M\|I + \alpha J\|_2 \|x_t^{\otimes N} - y_t^{\otimes N}\| \\ &\overset{(iii)}{\leq} M(\|I\|_2 + \alpha\|J\|_F)\|x_t^{\otimes N} - y_t^{\otimes N}\| \\ &\overset{(iv)}{\leq} M(1 + \alpha)\|x_t^{\otimes N} - y_t^{\otimes N}\|\end{aligned} \tag{163}$$

and in (i), we used the product inequality about the norm of the matrix and vector product, and in (ii), we used the fact that 2-norm is smaller than the Frobenius norm and used the smoothness condition and in (iii), we used the decomposition inequality of the matrix norm and finally in (iv), we used the assumption 6. This ends the proof. □

H.5. Other conditions

In this section, we study how the conditions other than dissipativeness and smoothness change by the non-reversible drift function.

First, we check about the condition of the drift function at the origin: $\|\nabla u(0,z)\| \le B$. We can calculate in the same way as the smoothness condition. Then we have

$$\| (I+\alpha J)\nabla U^{\otimes N}(0^{\otimes N})\| \le B(1+\alpha). \tag{164}$$

Next, we study the condition about thte stochastic gradient: $\mathbb{E}[\|g(x,V_z)-\nabla U(x)\|^2] \le 2\delta\left(M^2\|x\|^2+B^2\right)$. This can be easily modified to

$$\begin{aligned}
&\mathbb{E}[\| (I+\alpha J)\, g^{\otimes N}(x^{\otimes N},V_z) - (I+\alpha J)\nabla U^{\otimes N}(x^{\otimes N})\|^2] \\
&\le (1+\alpha)^2\mathbb{E}[g^{\otimes N}(x^{\otimes N},V_z)-\nabla U^{\otimes N}(x^{\otimes N})\|^2] \\
&\le (1+\alpha)^2\sum_{i=1}^N \mathbb{E}[g(x^{(i)},V_z)-\nabla U(x^{(i)})\|^2] \\
&\le (1+\alpha)^2\sum_{i=1}^N 2\delta\left(M^2\|x^{(i)}\|^2+B^2\right) \\
&\le 2\delta(1+\alpha)^2\left(M^2\|x^{\otimes N}\|^2+NB^2\right).
\end{aligned} \tag{165}$$

Finally, we discuss about the initial condition: $\kappa_0 := \log\int_{\mathbb{R}^d} e^{\|x\|^2}p_0(x)dx < \infty$. We assume that the initial probability distribution is $\mu_0^{\otimes N}(X_0^{\otimes N}) = \mu_0(X_0^{(1)})\times\cdots\times\mu_0(X_0^{(N)})$, which means that all the marginal probability is the same. Then

$$\begin{aligned}
\kappa_0^{\otimes N} &:= \log\int_{\mathbb{R}^{dN}} e^{\|x^{\otimes N}\|^2}\mu_0^{\otimes N}(x^{\otimes N})dx^{\otimes N} \\
&= \log\prod_{n=1}^N\left(\int_{\mathbb{R}^d} e^{\|x^{(n)}\|^2}\mu_0(x^{(n)})dx\right) \\
&= N\kappa_0.
\end{aligned} \tag{166}$$

H.6. Error control

We analyze our proposed dynamics in the same way as the standard SGLD. For that purpose, we decompose the error in the following way.

$$W_2(\mu_{kh}^{\otimes N},\pi^{\otimes N}) \le W_2(\mu_{kh}^{\otimes N},\nu_{kh}^{\otimes N}) + W_2(\nu_{kh}^{\otimes N},\pi^{\otimes N}). \tag{167}$$

About the convergence to the stationary, the strategy is the same. As shown in in Appendi K.2, our proposed dynamics satisfies LSI($\lambda(\alpha,N)$). Thus, the logarithmic Sobolev inequality implies T_2 inequality and use the exponential convergence to the stationary in the meaning of the relative entropy, we get

$$\begin{aligned}
W_2(\nu_{kh}^{\otimes N}|\pi^{\otimes N}) &\le \sqrt{2\lambda(\alpha,N)\mathrm{KL}(\nu_{kh}^{\otimes N}|\pi^{\otimes N})} \\
&\le \sqrt{2\lambda(\alpha,N)\mathrm{KL}(\nu_0^{\otimes N}|\pi^{\otimes N})}e^{-\frac{kh}{\beta\lambda(\alpha,N)}}.
\end{aligned} \tag{168}$$

where we substitute β in the LSI constant following the time rescaling introduced in Appendix D.1. Then, we assume that the initial measures are all the same,

$$\mathrm{KL}(\nu_0^{\otimes N}|\pi^{\otimes N}) = N\mathrm{KL}(\nu_0|\pi), \tag{169}$$

holds. Thus, our interest is how the logarithmic constant λ_α depends on N and α. Our analysis clarified that the logarithmic Sobolev constant of our scheme can be smaller than that of the naive parallel chain SGLD, which is shown in Appendix K.2.

Thus, next, we consider the discretization error. Since the discretization error of the standard SGLD is written as

$$W_2^2(\nu_{kh},\mu_{kh}) \le \mathbb{E}\|X_{kh}-V_{kh}\|^2 < C_7\,(kh+C_6)\,kh, \tag{170}$$

our interest is how the constant C_6 and C_7 is modified due to the interaction.

For that analysis, we use the discussion of the previous section, Appendix H.2. Then, due to the interaction and ensembling, constants are modified in the following way: $M \to (1+\alpha)M$, $B^2 \to (1+\alpha)^2NB^2$, $\kappa_0 \to N\kappa_0$, $\kappa \to N\kappa$, $d \to Nd$. The definition of C_6 and C_7 is

$$C_6 := \mathrm{e}^{\frac{3M^2}{2}(kh)^2}\left(-\frac{(kh)\mathrm{e}^{-\frac{3M^2}{2}(kh)^2}}{3M^2}+\frac{1}{6M^2}\sqrt{\frac{2\pi}{3M^2}},\right) \tag{171}$$

$$C_7 := \frac{6}{\beta}\left(C_1h+\beta C_0\delta\right). \tag{172}$$

Thus, we easily find that C_6 is modified by replacing $M \to (1+\alpha)M$. Thus, we find that C_6 never depends on N. We express this modified C_6 as C_6'. Since this C_6' depends on α, we write it as $C_6'(\alpha)$ and $C_6'(\alpha=0) = C_6$. About C_7, recall that

$$\begin{aligned}
C_0 &= \left(M^2\left(\kappa_0+2\left(1\vee\tfrac{1}{m}\right)\left(b+2B^2+\tfrac{d}{\beta}\right)\right)+B^2\right), \\
C_1 &= 6M^2(\beta C_0+d).
\end{aligned}$$

Then, this is modified to

$$\begin{aligned}
\bar{C}_0 &= N(1+\alpha)^2M^2\left(\kappa_0+2\left(1\vee\tfrac{1}{m}\right)\left(b+2(1+\alpha)^2B^2+\tfrac{d}{\beta}\right)\right) \\
&\quad +(1+\alpha)^2NB^2, \\
\bar{C}_1 &= 6(1+\alpha)^2M^2(N\beta\tilde{C}_0+Nd).
\end{aligned}$$

We define $\bar{C}_0 = NC_0'', \bar{C}_1 = NC_1''$, then we get the modified C_7

$$\bar{C}_7 = N\frac{6}{\beta}\left(C_1''h + \beta C_0''\delta\right)kh := NC_7'. \tag{173}$$

where C_7' never depends on N. Since this C_7' depends on α, we write it as $C_7'(\alpha)$ and $C_7'(\alpha = 0) = C_7$. Thus, we get

$$W_2^2(\nu_{kh}^{\otimes N}, \mu_{kh}^{\otimes N}) < NC_7'(\alpha)\left(kh + C_6'(\alpha)\right)kh. \tag{174}$$

and

$$C_6' = (6((1+\alpha)M)^2)^{-1}\times \left(\sqrt{2\pi(3((1+\alpha)M)^2)^{-1}}e^{\left(\frac{3((1+\alpha)M)^2}{2}(kh)^2\right)} - 2kh\right), \tag{175}$$

$$C_7'(\alpha) = \frac{6}{\beta}\left(C_1''h + \beta C_0''\delta\right)kh, \tag{176}$$

$$C_0'' = (1+\alpha)^2B^2 + (1+\alpha)^2M^2, \times \left(\kappa_0 + 2\left(1 \vee \frac{1}{m}\right)\left(b + 2(1+\alpha)^2B^2 + \frac{d}{\beta}\right)\right) \tag{177}$$

$$C_1'' = 6(1+\alpha)^2M^2(\beta C_0'' + d). \tag{178}$$

Thus, in conclusion, the W_2 distance of our proposed method is

$$W_2(\mu_{kh}^{\otimes N}, \pi^{\otimes N}) < \sqrt{NC_7'(\alpha)\left(kh + C_6'(\alpha)\right)kh} + \sqrt{2N\lambda(\alpha, N)\mathrm{KL}(\nu_0|\pi)}e^{-\frac{kh}{\beta\lambda(\alpha,N)}}. \tag{179}$$

H.7. Derivation of the bias Eq (31)

If a test function f is an L_f-lipschitz function 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 $L_f/\sqrt{N}$. Thus, we get

$$\left|\mathbb{E}\frac{1}{N}\sum_{n=1}^N f(X_k^{(n)}) - \int_{\mathbb{R}^d} f d\pi\right| < L_f\left(\sqrt{C_3'(\alpha)\left(kh + C_4'(\alpha)\right)kh} + \sqrt{2\lambda C'}e^{-\frac{kh}{\beta\lambda(\alpha,N)}}\right), \tag{180}$$

H.8. Comparison with the bound of Theorem 1

Note that, if we set $N = 1$ and $\alpha = 0$, which is the setting of the standard SGLD, then this bound reduce to $W_2^2(\nu_{kh}, \mu_{kh}) < C_7\left(kh + C_6\right)kh$, which is the bound of the standard SGLD. Thus, this is a natural result.

On the other hand, let us consider to bound the proposed dynamics with the bound of Theorem 1. The problem is already discussed in Appendix D.2, that is, the bias becomes larger as we increase the number of particles.

On the other hand, our bias is

$$\left|\mathbb{E}\frac{1}{N}\sum_{n=1}^N f(X_{kh}^{(n)}) - \int_{\mathbb{R}^d} f d\pi\right| < L_f\left(\sqrt{C_3'(\alpha)\left(kh + C_4'(\alpha)\right)kh} + \sqrt{2\lambda C'}e^{-\frac{kh}{\beta\lambda(\alpha,N)}}\right), \tag{181}$$

which never increases as N increases.

I. Improved discretizatoin error for the ensemble system

In this section, we derive the discretization error based on Vempala & Wibisono (2019). For that puperse, we just replace the constants in Theorem 9 in the following way: $M \to (1+\alpha)M$, $B^2 \to (1+\alpha)^2NB^2$, $\kappa_0 \to N\kappa_0$, $\kappa \to N\kappa$, $d \to Nd$, $\lambda \to \lambda(\alpha, N)$, and $C_0 \to \bar{C}_0$ Then we get

$$\begin{aligned} &W_2(\mu_k^{\otimes N}, \pi^{\otimes N})^2 \\ &\le 2\lambda(\alpha, N)\,\mathrm{KL}(\mu_k^{\otimes N}|\pi^{\otimes N}) \\ &\le 2\lambda(\alpha, N)e^{-\frac{kh}{\beta\lambda(\alpha,N)}}N\mathrm{KL}(\mu_0|\pi) \\ &\quad + \frac{8\lambda(\alpha,N)^2}{3}\left(\frac{2hdN(1+\alpha)M}{\beta}(m + 4(1+\alpha)M) + 8\bar{C}_0\bar{\delta}\right) \\ &:= 2N\lambda(\alpha, N)e^{-\frac{kh}{\beta\lambda(\alpha,N)}}\kappa_0 + N\lambda(\alpha, N)^2C_8(\alpha), \end{aligned} \tag{182}$$

where we set

$$C_8(\alpha) := \frac{8}{3}\left(\frac{2hd(1+\alpha)M}{\beta}(m + 4(1+\alpha)M) + 8\bar{C}_0\bar{\delta}\right), \tag{183}$$

for simplicity.

We summarize this result as

Theorem 11. *Under Assumptions 1 to 5 and an additional assumption $\frac{m}{4M^2} \le 1$, for any $k \in \mathbb{N}$ and any $h \in (0, \frac{1}{4\sqrt{2}M^2\lambda} \wedge \frac{m}{4M^2})$ obeying $kh \ge 1$ and $\beta m \ge 2$, $\mu_{kh}^{\otimes N}$, which is induced by Eq.(20), satisfies*

$$\begin{aligned} &W_2(\mu_k^{\otimes N}, \pi^{\otimes N})^2 \\ &:= 2N\lambda(\alpha, N)e^{-\frac{kh}{\beta\lambda(\alpha,N)}}\kappa_0 + N\lambda(\alpha, N)^2C_8(\alpha), \end{aligned} \tag{184}$$

where

$$C_8(\alpha) := \frac{8}{3}\left(\frac{2hd(1+\alpha)M}{\beta}(m + 4(1+\alpha)M) + 8\bar{C}_0\bar{\delta}\right). \tag{185}$$

If a test function f is an L_f-lipschitz function 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 $L_f/\sqrt{N}$. Thus, we get

$$\left|\mathbb{E}\frac{1}{N}\sum_{n=1}^{N} f(X_k^{(n)}) - \int_{\mathbb{R}^d} f d\pi\right| \leq L_f\sqrt{2\lambda(\alpha,N)e^{-\frac{kh}{\beta\lambda(\alpha,N)}}\kappa_0 + \lambda(\alpha,N)^2 C_8(\alpha)}. \tag{186}$$

Compared to the bias Eq.(139),

$$\left|\mathbb{E}f(X_{kh}) - \int_{\mathbb{R}^d} f d\pi\right| \leq L_f\sqrt{2\lambda e^{-\frac{1}{\beta\lambda}kh}\mathrm{KL}(\mu_0|\pi) + \lambda C_8},$$

where $C_8 = C_8(\alpha = 0)$, we can see that the first term becomes small due to the interaction, while the second term corresponds to the stochastic gradient and discretization error can be larger due to the interaction. Thus, we can observe the trade-off between faster convergence and larger discretization error.

J. Discussion about a spectral gap

In this section, we discuss the spectral gap, which is closely related to the Poincaré constant. We focus on the situation when we introduce the divergence-free drift.

J.1. Definition of a spectral gap

First, recall that the definition of the generator for standard Langevin dynamics is

$$\mathcal{L}f(X_t) = \left(-\nabla U(X_t)\cdot\nabla + \beta^{-1}\Delta\right) f(X_t),$$

and that of the N-naive parallel chain as

$$\mathcal{L}_{\alpha=0}f(X_t^{\otimes N}) = \left(-\nabla U^{\otimes N}(X_t^{\otimes N})\cdot\nabla + \beta^{-1}\Delta\right) f(X_t^{\otimes N}).$$

On the other hand, the generator of the proposed dynamics is

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

where $\nabla U_\alpha(x^{\otimes N}) := \nabla U^{\otimes N}(x^{\otimes N}) + \alpha J\nabla U^{\otimes N}(x^{\otimes N})$.

Then, if $\rho_0 > 0$ which is defined as

$$\rho_0 := \inf_{f\in\mathcal{D}(\mathcal{E})}\left\{\frac{\mathcal{E}(f)}{\int f^2 d\pi} : f\neq 0, \int f d\pi = 0\right\},$$

π with $\mathcal{L}$ has the spectral gap ρ_0. From this definition, ρ_0 is the smallest eigenvalue other than 0. From the definition of the spectral gap, a constant c which satisfies $\rho^{-1} \geq c$ can satisfy a Poincare inequality. We can also define the spectral gap of $\pi^{\otimes N}$ with $\mathcal{L}_{\alpha=0}$ and denote it $\rho(\alpha = 0, N)$.

On the other hand, the spectrum of $\mathcal{L}_{\alpha\neq 0}$ are not the real values but takes complex values. Although $-\mathcal{L}_\alpha$ is not self adjoint operator, from proposition 1 in Franke et al. (2010), it has discrete complex spectrums. Following Franke et al. (2010), we define the spectral gap of $\mathcal{L}_\alpha$ as the smallest real part of the eigenvalue which is larger than 0 and denote it $\rho(\alpha, N)$.

We review the result of Hwang et al. (2005), which shows that a real part of the spectral gap of the nonreversible dynamics is larger than those of the reversible one. We denote the spectral gap of the proposed dynamics as $\rho(\alpha, N)$. Note that Then ρ_0 denotes the case of standard Langevin dynamics.

J.2. Relation of spectral gaps

Theorem 12. *Assume that the stationary distribution π with $\mathcal{L}$ has the spectral gap ρ_0 and $\pi^{\otimes N}$ with $\mathcal{L}_\alpha$ has the spectral gap $\rho(\alpha, N)$. Then we have*

$$\rho(\alpha, N) \geq \rho_0. \tag{187}$$

Proof. Our proof follow a similar line with Hwang et al. (1993; 2005).

Since the generator $\mathcal{L}_{\alpha=0}$ is self-adjoint, and the suitable growth condition, the spectral of $\mathcal{L}_{\alpha=0}$ is discrete Bakry et al. (2013). We denote the spectrum of $\mathcal{L}_{\alpha=0}$ as $\{\lambda_k\}_{k=0}^\infty \in \mathbb{R}$ and corresponding normalized eigenvectors as $\{e_k\}_{k=0}^\infty$, which are the real functions. We order the spectrum as $0 > \lambda_0 > \lambda_1 > \ldots$. Thus, $\rho(\alpha = 0, N) = -\lambda_0$.

As for $\mathcal{L}_\alpha$, although it is not self adjoint operator, from proposition 1 in Franke et al. (2010), it has discrete complex spectrums. We denote the spectrum of $\mathcal{L}_\alpha$ as $\lambda + i\mu \in \mathbb{C}$ where $\lambda, \mu \in \mathbb{R}$ and corresponding normalized eigenvector as $u + iv$ where u, v are the real functions and then we have

$$\mathcal{L}_\alpha(u + iv) = (\lambda + i\mu)(u + iv). \tag{188}$$

From this definition, by checking the real parts and complex parts, following relations are derived

$$\mathcal{L}_\alpha u = \lambda u - \mu v, \tag{189}$$
$$\mathcal{L}_\alpha v = \lambda v + \mu u. \tag{190}$$

Due to the divergence-free drift property, for any bounded real value test function $g(x)$,

$$\int g(\mathcal{L}_{\alpha=0} - \mathcal{L}_\alpha) g d\pi = \int \alpha g\gamma\cdot\nabla g d\pi = -\int \alpha g\gamma\cdot\nabla g d\pi \tag{191}$$

where we used the partial integral. This means that for any bounded real function $g(x)$,

$$\int g\mathcal{L}_{\alpha=0} g d\pi = \int g\mathcal{L}_\alpha g d\pi. \tag{192}$$

(This only holds for real functions.) Then, we can evaluate the real part of the eigenvalue λ as follows,

$$\begin{aligned}\int u\mathcal{L}_{\alpha=0}ud\pi + \int v\mathcal{L}_{\alpha=0}vd\pi &= \int u\mathcal{L}_{\alpha}ud\pi + \int v\mathcal{L}_{\alpha}vd\pi \\ &= \lambda\left(\int u^2d\pi + \int v^2d\pi\right) \\ &= \lambda. \end{aligned} \tag{193}$$

Then, by expanding the eigenfunction u, v by the eigenfunction $\{e_k\}$,

$$\begin{aligned}\lambda &= \int u\mathcal{L}_{\alpha=0}ud\pi + \int v\mathcal{L}_{\alpha=0}vd\pi \\ &= \sum_k \lambda_k\left(\left(\int ue_kd\pi\right)^2 + \left(\int ve_kd\pi\right)^2\right) \\ &\leq \lambda_0\sum_k\left(\left(\int ue_kd\pi\right)^2 + \left(\int ve_kd\pi\right)^2\right) \\ &\leq \lambda_0. \end{aligned} \tag{194}$$

Thus the real part of the eigenvalue of $\mathcal{L}_\alpha$ is smaller than the smallest eigenvalue of $\mathcal{L}_\alpha$. This means that $\rho(\alpha, N) \geq \rho(\alpha = 0, N) = -\lambda_0$.

Finally, we can conclude $\rho(\alpha = 0, N) = \rho_0$ from the tensorization property of the spectral gap (Proposition 4.3.1 in Bakry et al. (2013)).

Then we conclude $\rho(\alpha, N) \geq \rho(\alpha = 0, N) = \rho_0$. □

J.3. Evaluation of the spectral gap

As our analysis clarified, we should set α sufficiently small so that the discretization error will not become large. In such a situation, we can regard the non-reversible drift term as the perturbation to the original process. This means that when we write the perturbation term as V, the generator is written as

$$\mathcal{L}_\alpha f = \mathcal{L}_{\alpha=0}f + \alpha Vf. \tag{195}$$

(Hereinafter in this section, we denote $\mathcal{L}_{\alpha=0}$ as $\mathcal{L}$ for simplicity.) Then, our question is that how much the largest eigenvalue will be perturbed by the αV. Following the notations in the previous section, we write the eigenvalues and vectors of $\mathcal{L}$ as $\{\lambda_k\}$ and $\{e_k\}$. We write the eigenvalues and eigenvectors of $\mathcal{L}_\alpha$ as $\gamma = \lambda + \mathrm{i}\mu$, $w = u + \mathrm{i}v$. Our goal is to derive the relation between the λ_k and γ, especially it's real part $\mathrm{Re}\gamma$.

When α is sufficiently small, we can use the perturbation theory for that purpose. Let us start from the definition of the basic relation,

$$\mathcal{L}_\alpha w = (\mathcal{L} + \alpha V)w = \gamma w = (\lambda + \mathrm{i}\mu)(u + \mathrm{i}v). \tag{196}$$

Then by multiplying e_k from the left-hand side and take the average with respect to μ, we get

$$\int e_k(\mathcal{L} + \alpha V)wd\pi = (\lambda + \mathrm{i}\mu)\int e_kwd\pi. \tag{197}$$

As for the left-hand side in the above equation, by using

$$\mathcal{L}e_k = \lambda_k e_k, \tag{198}$$

we get

$$\int e_k(\mathcal{L} + \alpha V)wd\pi = \lambda_k\int e_kwd\pi + \alpha\int e_kVwd\pi. \tag{199}$$

Then we get

$$((\lambda + \mathrm{i}\mu) - \lambda_k)\int e_kwd\pi = \alpha\int e_kVwd\pi, \tag{200}$$

if w and e_k are not orthogonal,

$$(\lambda + \mathrm{i}\mu) - \lambda_k = \frac{\alpha\int e_kVwd\pi}{\int e_kwd\pi}. \tag{201}$$

Thus,

$$\lambda - \lambda_k = \mathrm{Re}\frac{\alpha\int e_kVwd\pi}{\int e_kwd\pi}. \tag{202}$$

Then, we will study the right-hand side by the perturbation theory. For that purpose, we introduce two projection operators, $\{Q_k\}$ and $\{P_k\}$. Since $\{e_k\}$ is the complete orthogonal system, we define Q_k as the projection into the space spanned by e_k. Thus we define the projection Q_k as

$$Q_kw := \left(\int e_kwd\pi\right)w = \langle e_k, w\rangle_\pi e_k. \tag{203}$$

Then, we define the projection P_k as the orthogonal projection to Q_k;

$$P_kw := (I - Q_k)w = w - \langle e_k, w\rangle_\pi e_k. \tag{204}$$

Since Q_k and P_k are projections, we can easily comfirm that $Q_k^2 = Q_k$, $P_k^2 = P_k$, $Q_kP_k = P_kQ_k = 0$, $P_k + Q_k = I$.

Based on this projections, let us start the perturbation estimation of the eigenvalues. Back to the definition,

$$(\mathcal{L} + \alpha V)w = \gamma w, \tag{205}$$

first, we add the baseline $\epsilon \in \mathbb{R}$ from both-sides, then we get

$$(\mathcal{L} + \alpha V + \epsilon)w = (\gamma + \epsilon)w, \tag{206}$$

then rearrange the both-sides as

$$(\alpha V + \epsilon - \gamma)w = (\epsilon - \mathcal{L})w. \tag{207}$$

Then, we apply the projection P_k on both-hands,

$$\begin{aligned} P_k(\alpha V + \epsilon - \gamma)w &= P_k(\epsilon - \mathcal{L})w \\ &= (\epsilon - \mathcal{L})P_k w \\ &= (\epsilon - \mathcal{L})(w - \langle e_k, w\rangle_\pi e_k), \end{aligned} \tag{208}$$

then by applying the inverse; $(\epsilon - \mathcal{L})^{-1}$ (thus, we need to choose ϵ so that this inverse exists. Later we set this $\epsilon = \lambda_k$). Then, by rearranging the both-hand sides, we get

$$w = \langle e_k, w\rangle_\pi e_k + (\epsilon - \mathcal{L})^{-1} P_k(\alpha V + \epsilon - \gamma)w. \tag{209}$$

From this expression, we recursively substituting w,

$$\begin{aligned} w &= \langle e_k, w\rangle_\pi e_k + (\epsilon - \mathcal{L})^{-1} P_k(\alpha V + \epsilon - \gamma)w \\ &= \langle e_k, w\rangle_\pi e_k + (\epsilon - \mathcal{L})^{-1} P_k(\alpha V + \epsilon - \gamma)\langle e_k, w\rangle_\pi e_k \\ &\qquad + \left((\epsilon - \mathcal{L})^{-1} P_k(\alpha V + \epsilon - \gamma)\right)^2 w \\ &\vdots \\ &= \sum_{n=0}^{\infty} D_k(\alpha, \epsilon)^n \langle e_k, w\rangle_\pi e_k. \end{aligned} \tag{210}$$

where

$$D_k(\alpha, \epsilon) := (\epsilon - \mathcal{L})^{-1} P_k(\alpha V + \epsilon - \gamma). \tag{211}$$

Then, we substitute this expression to Eq.(202), then we get

$$\lambda - \lambda_k = \alpha \mathrm{Re} \int e_k V \sum_{n=0}^{\infty} D_k(\alpha, \epsilon)^n e_k d\pi. \tag{212}$$

Let us set $\epsilon = \lambda_l$ where $l \neq k$. First, the when we consider only up to $n = 0$, we get the first order expansion with respect to α as

$$\lambda - \lambda_k = \alpha \mathrm{Re} \int e_k V e_k d\pi + \mathcal{O}(\alpha^2). \tag{213}$$

Since $Ve_k = J\nabla U \cdot \nabla e_k$ where J is the skew symmetric matrix, $\int e_k V e_k d\pi = 0$. Thus,

$$\lambda = \lambda_k + \mathcal{O}(\alpha^2). \tag{214}$$

This means that the real part of the eigenvalue of $\mathcal{L}_\alpha$ never changes up to the first order concerning α. Thus, we need to check the higher-order to observe the α dependency. Let us consider the term up to $n = 1$,

$$\begin{aligned} \lambda - \lambda_k &= \alpha \mathrm{Re} \int e_k V D_k(\alpha, \epsilon) e_k d\pi \\ &= \alpha \mathrm{Re} \int e_k V (\epsilon - \mathcal{L})^{-1} P_k(\alpha V + \epsilon - \gamma) e_k d\pi, \end{aligned} \tag{215}$$

where we omit $\mathcal{O}(\alpha^3)$. From now on, we set $k = 0$ and $\epsilon = \lambda_0$ for simplicity. The following discussion holds in the case of $k \neq 0$. Note that $\{e_k\}$ is the complete orthonormal system (CONS), thus for any $f, g \in \mathcal{D}(f)$, $\langle f, g\rangle_\pi = \sum_{k=0}^{\infty} \langle f, e_k\rangle_\pi \langle e_k, g\rangle_\pi$ (This is the property of CONS), thus we use this relation in Eq.(215), we get

$$\begin{aligned} &\lambda - \lambda_0 \\ &= \alpha \mathrm{Re} \sum_{k=0}^{\infty} \left(\int e_0 V (\lambda_0 - \mathcal{L})^{-1} P_0 e_k d\pi \right) \\ &\qquad \times \left(\int e_k (\alpha V + \epsilon - \gamma) e_0 d\pi \right) \\ &= \alpha \sum_{k=1}^{\infty} \left(\frac{\int e_0 V e_k d\pi}{\lambda_0 - \lambda_k} \right) \left(\int e_k \alpha V e_0 d\pi \right) \\ &= \alpha^2 \sum_{k=1}^{\infty} \frac{|\int e_k V e_0 d\pi|^2}{\lambda_k - \lambda_0}. \end{aligned} \tag{216}$$

Thus, in conclusion, we get the expression that up to the second order of α,

$$\lambda = \lambda_0 + \alpha^2 \sum_{k=1}^{\infty} \frac{|\langle e_k, V e_0\rangle_\pi|^2}{\lambda_k - \lambda_0} + \mathcal{O}(\alpha^3). \tag{217}$$

This concludes the proof.

J.4. Lower bound of the Poincare constant

In Raginsky et al. (2017), the upper bound of the Poincare constant is derived for the standard Langevin dynamics under the assumptions 1 to 5. On the other hand, in this section, we discuss the lower bound of the Poincare constant under the same assumptions, which will be used in the comparison of the LSI constants. Since we are interested in the lower bound of the Poincare constant, what we will investigate is the tightest Poincare inequality, which satisfies the assumptions 1 to 5. For that purpose, we focus on the assumption 3, which is a dissipative assumption. This assumption is closely related to a convex property. In usual, we say that a function $U(x) : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$ is a convex function , if there exists a positive constant m for any $x, y \in \mathcal{X} \subset \mathbb{R}^d$ such that

$$(\nabla U(x) - \nabla U(y)) \cdot (x - y) \geq m\|x - y\|^2. \tag{218}$$

If we substitute $y = 0$, then we get the assumption 3 with $b = 0$. Thus, the assumption 3 is weaker than the convex assumption. Thus, we focus on the Poincare constant when the potential is a convex function.

Convexity is closely related to the curvature dimension conditions (Bakry et al., 2013), and that condition directly results in various functional inequalities. Recall that Γ is the Carre du champ operator and then let us define the following operator

$$\Gamma_2(f, g) := \frac{1}{2} \left(\mathcal{L}\Gamma(f, g) - \Gamma(f, \mathcal{L}g) - \Gamma(\mathcal{L}f, g) \right). \tag{219}$$

We express $\Gamma_2(f) := \Gamma_2(f,f)$. Then, we say that $\mathcal{L}$ satisfies the curvature dimension condition $CD(\rho,\infty)$ if it satisfies

$$\Gamma_2(f) \geq \rho\Gamma(f), \tag{220}$$

for every function f in a sufficiently rich families of functions in $\mathcal{D}(\mathcal{L})$ (See Bakry et al. (2013) for the details). From the discussion in Collorary 4.8.2 in Bakry et al. (2013), if $U - \frac{m}{2}\|x\|^2$ is a convex function, then it satisfies $CD(m,\infty)$. Then from Collorary 4.8.2 in Bakry et al. (2013), such a Langevin dynamics satisfies the Poincare inequality with constant $\frac{1}{m}$. Furthurmore, if $\pi \propto e^{-U}$ is a compactly supported Riemannian measure, then from 4.8.1 in Bakry et al. (2013), it satisfies the Poincare inequality $\frac{d-1}{dm}$. We cannot improve this constant by the definition. See Bakry et al. (2013) for details.

Back to our setting, our potential function can satisfy $CD(2\beta m, d)$, thus it can satisfy the Poincare inequality $\frac{d-1}{d2\beta m}$, which cannot be improved. This is achieved if U is a convex function.

K. Estimation of the Logarithmic Sobolev constant

In this section, we estimate the upper bound of the Logarithmic Sobolev constants in standard SGLD, naive parallel chain SGLD, and particle interacting system.

K.1. LSI constant of standard SGLD

Proof of Theorem 3: To estimate the logarithmic Sobolev constant, we rely on the technique of restricted logarithmic Sobolev inequality, which was introduced in Carlen & Loss (2004).

Theorem 13. *(Theorem 3.3 in Carlen & Loss (2004)) Suppose that $\pi \propto \exp(-\beta U(x))$ and U is C^2 and π admits a spectral gap $\rho > 0$, and*

$$-C = \inf_x \left\{ \frac{\beta}{4}\|\nabla U(x)\|^2 - \frac{1}{2}\Delta U(x) - \pi e^2 U(x) \right\} > -\infty, \tag{221}$$

then π admits a logarithmic Sobolev inequality with constant λ no larger than

$$\lambda \leq \frac{1}{(1+\rho^{-1}\beta|C|)2\pi e^2} + \frac{3}{2}\rho^{-1}, \tag{222}$$

that means for all functions v on $\mathbb{R}^d$ with $\int_{\mathbb{R}^d} f^2 d\mu = 1$,

$$\int_{\mathbb{R}^d} v^2 \ln v^2 d\mu \leq \lambda \int_{\mathbb{R}^d} \|\nabla v\|^2 d\mu. \tag{223}$$

This means that if the right-hand side of Eq.(285) is lower bounded, then π satisfies the logarithmic Sobolev inequality.

Also we can transform Eq.(223) by partial integration,

$$\begin{aligned}\int_{\mathbb{R}^d} v^2 \ln v^2 d\mu &\leq \lambda \int_{\mathbb{R}^d} \|\nabla v\|^2 d\mu \\ &= -\lambda \int_{\mathbb{R}^d} v\mathcal{L}' v d\mu \\ &= -\lambda\beta \int_{\mathbb{R}^d} v\mathcal{L} v d\mu \end{aligned} \tag{224}$$

where $\mathcal{L}$ and $\mathcal{L}'$ are defined by

$$\mathcal{L}f(X_t) = \left(-\nabla U(X_t)\cdot\nabla + \beta^{-1}\Delta\right) f(X_t). \tag{225}$$

and

$$\mathcal{L}'f(X_t) = \left(-\beta\nabla U(X_t)\cdot\nabla + \Delta\right) f(X_t), \tag{226}$$

which satisfies $\beta\mathcal{L} = \mathcal{L}'$. These are defined in Appendix D.1. Thus, from Appendix D.1, $\mathcal{L}'$ induces

$$dX_t = -\beta\nabla U(X_t) + \sqrt{2}dw(t), \tag{227}$$

and its time scale is changed from $t \to \beta t$. So by using Theorem 3.3 in Carlen & Loss (2004), we can understand the LSI constant of $\mathcal{L}'$. After that we just multiply β to get the LSI constant of $\mathcal{L}'$.

We study the right-hand side of Eq.(285) of standard SGLD under the assumptions 1 to 5.

From lemma 2 in (Raginsky et al., 2017), for all $x \in \mathbb{R}^d$,

$$\|\nabla U\| \leq M\|x\| + B, \tag{228}$$

and

$$\frac{m}{3}\|x\|^2 - \frac{b}{2}\log 3 \leq U(x) \leq \frac{M}{2}\|x\|^2 + B\|x\| + A, \tag{229}$$

and since the U is M-smooth, thus

$$\Delta U \leq Md. \tag{230}$$

From the dissipative condition, we have

$$\|x\|\|\nabla U(x)\| \geq x\cdot\nabla U(x) \geq m\|x\|^2 - b. \tag{231}$$

Thus

$$\|\nabla U(x)\| \geq m\|x\| - b/\|x\|. \tag{232}$$

If $\|x\| \geq \sqrt{2b/m}$, then

$$\|\nabla U\| \geq \frac{1}{2}m\|x\|. \tag{233}$$

Then,

$$\begin{aligned}&\frac{\beta}{4}\|\nabla U(x)\|^2 - \frac{1}{2}\Delta U(x) - \pi e^2 U(x) \\ &\geq \frac{\beta m^2}{16}\|x\|^2 - Md/2 - \pi e^2\left(\frac{M}{2}\|x\|^2 + B\|x\| + A\right) \\ &= \left(\frac{\beta m^2}{16} - \frac{M\pi e^2}{2}\right)\|x\|^2 - \pi e^2 B\|x\| - (Md/2 + A\pi e^2).\end{aligned} \tag{234}$$

Thus,

$$-C = \inf_x \frac{\beta}{4}\|\nabla U(x)\|^2 - \frac{1}{2}\Delta U(x) - \pi e^2 U(x) \geq \inf_x \left(\frac{\beta m^2}{16} - \frac{M\pi e^2}{2}\right)\|x\|^2 - \pi e^2 B\|x\| - (Md/2 + A\pi e^2). \tag{235}$$

Then, the coefficient of $\|x\|^2$ should be larger than 0 so that the right-hand side of the above inequality is bounded below. This means

$$C' := \frac{\beta m^2}{16} - \frac{M\pi e^2}{2} > 0. \tag{236}$$

This is equivalent to

$$\frac{\beta m^2}{M} > 8\pi e^2. \tag{237}$$

On the other hand, from Eq.(229), following relation holds,

$$M/2 \geq m/3 \Leftrightarrow 3/2 \geq m/M. \tag{238}$$

Combined with Eq.(237,238), we get

$$\beta m > \frac{16\pi e^2}{3}. \tag{239}$$

Under this assumption, π satisfies the logarithmic Sobolev inequality. Compared with (Raginsky et al., 2017), where $\beta m > 2$ holds, this is stronger condition. However, as we can see later this shows better dependency about the dimension d in the LSI constant.

Back to the estimate of the upper bound of the LSI, under the above assumption, the larger the absolute value of C means the smaller LSI constant, which means stronger inequality. We can easily find the infimum of Eq.(235) under the assumption of $\|x\| \geq \sqrt{2b/m}$. However, for simplicity, we simply upper bound the LSI constant as

$$\lambda \leq \frac{1}{(1+\rho^{-1}\beta|C|)2\pi e^2} + \frac{3}{2}\rho^{-1} \leq \frac{1}{2\pi e^2} + \frac{3}{2}\rho^{-1}. \tag{240}$$

Thus, the estimate of the LSI constant is

$$\lambda_e := \frac{1}{(1+\rho^{-1}\beta|C|)2\pi e^2} + \frac{3}{2}\rho^{-1}. \tag{241}$$

We can estimate the lower bound of the LSI constant. Note the right-hand side of Eq.(285) as

$$\frac{\beta}{4}\|\nabla U(x)\|^2 - \frac{1}{2}\Delta U(x) - \pi e^2 U(x) \leq \frac{\beta(M\|x\|+B)^2}{4} - \pi e^2\left(\frac{m}{3}\|x\|^2 - \frac{b}{2}\log 3\right) + Md/2. \tag{242}$$

Thus,

$$-C = \inf_x \frac{\beta}{4}\|\nabla U(x)\|^2 - \frac{1}{2}\Delta U(x) - \pi e^2 U(x) \leq \inf_x \frac{\beta(M\|x\|+B)^2}{4} - \pi e^2\left(\frac{m}{3}\|x\|^2 - \frac{b}{2}\log 3\right) + Md/2 \leq \frac{\beta B^2}{4} + \frac{b\pi e^2}{2}\log 3 + \frac{Md}{2}. \tag{243}$$

in the last inequality on the above, we substitute 0 for $\|x\|$ for simplicity. Thus, we get the upper bound

$$\lambda \leq \lambda_e := \frac{1}{(1+\rho^{-1}\beta C)2\pi e^2} + \frac{3}{2}\rho^{-1}, \tag{244}$$

$$0 < C \leq \frac{\beta B^2}{4} + \frac{b\pi e^2}{2}\log 3 + \frac{Md}{2}. \tag{245}$$

Next let us compare the estimated upper bound of the logarithmic constant with that of (Raginsky et al., 2017), which shows that

$$\lambda \leq \lambda_l := D_1 + \rho^{-1}D_2, \quad D_1 = \frac{2m^2+8M^2}{\beta m^2 M}, \quad D_2 = \left(\frac{6M(d+\beta)}{m} + 2\right). \tag{246}$$

Thus, we compare λ_e and λ_l. From Cauchy-Shwartz inequality,

$$D_1 = \frac{2}{M\beta} + \frac{8M}{\beta m^2} \geq 2\sqrt{\frac{16}{\beta^2 m^2}} = \frac{8}{\beta m} \tag{247}$$

Then

$$\begin{aligned}
&\lambda_l - \lambda_e \\
&= D_1 + \rho^{-1}D_2 - \left(\frac{1}{(1+\rho^{-1}\beta|C|)2\pi e^2} + \frac{3}{2}\rho^{-1}\right) \\
&\geq \tfrac{8}{\beta m} + \rho^{-1}\left(\tfrac{6M(d+\beta)}{m} + 2\right) - \left(\tfrac{1}{(1+\rho^{-1}\beta|C|)2\pi e^2} + \tfrac{3}{2}\rho^{-1}\right) \\
&\geq \frac{8}{\beta m} + \rho^{-1}\left(4(d+\beta)+1/2\right) - \frac{1}{(1+\rho^{-1}\beta|C|)2\pi e^2} \\
&\geq \frac{8}{\beta m} + \frac{1}{2\beta m}\left(4(d+\beta)+1/2\right) - \frac{1}{2\pi e^2}.
\end{aligned} \tag{248}$$

In the last line, we used the tightest Poincare constant from Appendix J.4. Note that this ρ is associated with $\mathcal{L}'$ due to the time rescaling which is introduced in Appendix D.1. So, if the last line is larger than 0, $\lambda_l > \lambda_e$ holds. This is the condition.

K.2. Smaller logarithmic Sobolev constants for the non-reversible ensemble system

In this section, we show that a logarithmic Sobolev constant of the non-reversible chain is smaller than the reversible

one. The important question is that how the constant $-C$ in Eq.(285) changes by the divergence-free drift function. We express that constant as $C_{\alpha,N}$ to make it clear.

First let us consider the case of the naive parallel-chain SGLD and its corresponding constant which is related to Eq.(285) as $C_{0,N}$. By following the proof of Theorem 3.3 in Carlen & Loss (2004), we can easily find that the constant $C_{0,N}$ is derived by the following,

$$-C_{0,N} = \inf_{x^{\otimes N}\in\mathbb{R}^{dN}} \sum_{n=1}^{N} \frac{\beta}{4}\|\nabla U^{(n)}(x^{(n)})\|^2 - \frac{1}{2}\Delta U^{(n)}(x^{(n)}) - \pi e^2 U^{(n)}(x^{(n)}) > -\infty, \tag{249}$$

This is intuitive since the naive parallel-chain SGLD is the concatenation of standard SGLD.

Then, we calculate the case of the non-reversible chain, $C_{\alpha,N}$. We can easily find that the term which depends on J does not appear it is the skew-symmetric matrix. Thus, the objective function of the non-reversible chain is also Eq.(249). Thus, we can find that $C_{\alpha,N} = C_{0,N}$.

From this, the difference of the LSI constant between the naive-parallel chain SGLD and our proposed method comes from the difference of the spectral gap ρ. Recall that the estimate of LSI constant is given as

$$\lambda \le \lambda_e := \frac{1}{(1+\rho^{-1}\beta C)2\pi e^2} + \frac{3}{2}\rho^{-1}, \tag{250}$$

$$0 < C \le \frac{\beta B^2}{4} + \frac{b\pi e^2}{2}\log 3 + \frac{Md}{2}. \tag{251}$$

The important point is that the larger spectral gap ρ means the smaller LSI constant. This is easily confirmed by the fact that the right-hand side of Eq.(250) is a monotonically increasing function of about ρ^{-1}.

We express the upper bound of the LSI constant of standard SGLD as λ_e and that of the naive parallel chain as $\lambda(\alpha = 0, N)$, and our proposed method as $\lambda(\alpha, N)$.

First of all, we need to evaluate the spectral gaps. This is already finished in Appendix J.2,

$$\rho(\alpha, N) \ge \rho(0, N) = \rho_0. \tag{252}$$

Thus, we conclude that

$$\lambda(\alpha, N) \le \lambda(0, N). \tag{253}$$

Next, we discuss the relation between the LSI constant of naive N-parallel chain SGLD and standard SGLD. From the tensorization property of the LSI constant, we can see that the LSI inequality for naive N-parallel chain has the constant of λ_e, which is also the LSI constant of the standard SGLD, thus

$$\tilde{\lambda}(0, N) = \lambda_e = \frac{1}{(1+\rho_0^{-1}\beta C)2\pi e^2} + \frac{3}{2}\rho_0^{-1}. \tag{254}$$

Finally, let us compare $\tilde{\lambda}(0, N)$ and $\lambda(0, N)$. Then from the definition of C and $C_{0,N}$, the relation $C_{0,N} \ge C$ and $\rho(0, N) = \rho_0$ holds, thus, we can conclude that $\tilde{\lambda}(0, N) \ge \lambda(0, N)$. Thus, we get

$$\lambda(\alpha, N) \le \tilde{\lambda}(0, N) = \lambda_e. \tag{255}$$

Lyapunov function-based approach: On the other hand, when we estimate the LSI constat by the Lyapunov function-based approach, LSI constant of the standard SGLD is estimated as

$$\lambda_l \le a + \rho_0^{-1}(a' + a''\int_{\mathbb{R}^d}\|x\|^2 d\pi), \tag{256}$$

where a, a', a'' are some positive constants which are independent from the dimension d. This is also monotonically decreasing function about ρ, thus we can conclude that the upper bounds of the proposed and naive parallel chain SGLD are estimated as

$$\lambda(\alpha, N) \le \lambda(0, N) = b + \rho_0^{-1}(b' + b''\int_{\mathbb{R}^{dN}}\|x\|^2 d\pi^{\otimes N}), \tag{257}$$

where b, b', b'' are some positive constants which are independent of the dimension d. On the other hand, from the tensorization property of the LSI constant, the LSI constant of the naive parallel chain SGLD is estimated differently,

$$\tilde{\lambda}(0, N) = \lambda_l = a + \rho_0^{-1}(a' + a''\int_{\mathbb{R}^d}\|x\|^2 d\pi). \tag{258}$$

Thus, there is two different estimate about the upper bound of the LSI constant for the naive parallel chain SGLD. Different from our evaluation method, since Lyapunov function based-estimation has the second moment term $\int_{\mathbb{R}^{dN}}\|x\|^2 d\pi^{\otimes N}$, under the additional mild conditions,

$$\tilde{\lambda}(0, N) \le \lambda(0, N) \tag{259}$$

can hold since $\int_{\mathbb{R}^{dN}}\|x\|^2 d\pi^{\otimes N}$ can be N times larger than $\int_{\mathbb{R}^d}\|x\|^2 d\pi$. Thus, we cannot conclude that $\lambda(\alpha, N)$ is smaller than λ_l. This is the undesirable result. Thus, Lyapunov function-based approach is not appropriate in our analysis.

L. Removing the additional assumption of Theorem 3

In this section, we remove the additional condition of Theorem 3 in the main paper. Our goal is to prove

Theorem 14. *Under the same conditions as Theorem 2, the LSI constant is upper-bounded by λ_e:*

$$\lambda \le \lambda_e := \frac{\sqrt{12}}{(1+\rho^{-1}|C|)2\sqrt{\beta^2 m^2+8}} + 3(2\rho_0)^{-1}, \tag{260}$$

where

$$C := \inf_x \frac{\beta^2\|\nabla U\|^2}{4} - \frac{\beta\Delta U}{2} - \frac{\beta}{\gamma}V - \frac{\gamma^2-2}{4}\|x\|^2 + \frac{d}{2} > -\infty, \tag{261}$$

, $\gamma = \sqrt{\frac{\beta^2m^2+8}{12}}$, and ρ_0 is given in Eq.(13). Moreover, λ_e is always smaller than λ_l of Eq.(11) estimated by Raginsky et al. (2017).

Note: In Theorem 2 in the main paper, there is an additional assumption about βm. This assumption is strong. For example, if we consider the Gaussian potential function, which has $\beta = 1$ and $m = 2$, then it cannot satisfy the assumption of Theorem 2. On the other hand, it satisfy the assumption of Theorem 14, which requires $\beta m \ge 2$.

Proof. The proof is based on the improved result of Theorem 3.3 Carlen & Loss (2004) with respect to the constants.

Theorem 15. *Let us define $\pi \propto \exp(-U(x))$ and U is C^2 and π admits a spectral gap $\rho > 0$. Given a positive constant γ, if*

$$C := \inf_x \frac{\beta^2\|\nabla U\|^2}{4} - \frac{\beta\Delta U}{2} - \frac{\beta}{\gamma}V - \frac{\gamma^2-2}{4}\|x\|^2 + \frac{d}{2} > -\infty, \tag{262}$$

is satisfied, then, π admits a logarithmic Sobolev inequality with constant λ no larger than

$$\lambda \le \frac{1}{(1+\rho^{-1}|C|)2\gamma} + \frac{3}{2}\rho^{-1}. \tag{263}$$

The proof of this theorem is given in Appendix L.1

Discussion: Before the proof, we discuss why this is the improved version. Remember that Theorem 3.3 Carlen & Loss (2004) is given by

$$-C = \inf_x \frac{\beta}{4}\|\nabla U(x)\|^2 - \frac{1}{2}\Delta U(x) - \pi e^2 U(x) > -\infty, \tag{264}$$

then π admits a logarithmic Sobolev inequality with constant λ no larger than

$$\lambda \le \frac{1}{(1+\rho^{-1}\beta|C|)2\pi e^2} + \frac{3}{2}\rho^{-1}. \tag{265}$$

Let us compare this with Theorem 15. For example let us consider $U(x) = \|x\|^2$, which is the Gaussian. Then

$$\frac{\beta}{4}\|\nabla U(x)\|^2 - \frac{1}{2}\Delta U(x) - \pi e^2 U(x) = \|x\|^2 - d - \pi e^2\|x\|^2) = (1-\pi e^2)\|x\|^2 - d. \tag{266}$$

Since $(1-\pi e^2) < 0$, $\inf_x(1-\pi e^2)\|x\|^2 - d = -\infty$. On the other hand

$$\begin{aligned}&\frac{\beta^2\|\nabla U\|^2}{4} - \frac{\beta\Delta U}{2} - \frac{\beta}{\gamma}V - \frac{\gamma^2-2}{4}\|x\|^2 + \frac{d}{2}\\ &= \|x\|^2 - \frac{d}{2} - \frac{1}{\gamma}\|x\|^2 - \frac{\gamma^2-2}{4}\|x\|^2\\ &= (1 - \frac{1}{\gamma} - \frac{\gamma^2-2}{4})\|x\|^2 - \frac{d}{2}.\end{aligned} \tag{267}$$

So for example $\gamma = 1$, then $(1 - \frac{1}{\gamma} - \frac{\gamma^2-2}{4}) > 0$ thus $\inf_x(1 - \frac{1}{\gamma} - \frac{\gamma^2+2}{4})\|x\|^2 - \frac{d}{2} > -\infty$. Thus, by tuning γ appropriately, we can apply this improved Theorem to the standard Gaussian measure to estimate the LSI constant.

Back to the proof, following Appendix K.1, we will study when C of Theorem 15 is lower bounded. Following Appendix K.1, we substitute the relations about the upper bound of $U(x)$ and $\Delta U(x)$, and the lower bound of the $\nabla U(x)$, we get the condition that C of Theorem 15 is lower bounded if

$$\frac{\beta^2m^2\gamma}{16} - \frac{\beta M}{2} - \frac{\gamma^3-2\gamma}{4} \ge 0, \tag{268}$$

is satisfied. Then, we optimize above left handside with respect to γ which is a positive. We can easily find that if we define $f(\gamma) := \frac{\beta^2m^2\gamma}{16} - \frac{\beta M}{2} - \frac{\gamma^3-2\gamma}{4}$, then $f(\gamma)$ take the minimum when

$$\gamma = \sqrt{\frac{\beta^2m^2+8}{12}}, \tag{269}$$

and moreover $f(\sqrt{\frac{\beta^2m^2+8}{12}}) > 0$. Thus, if we set $\gamma = \sqrt{\frac{\beta^2m^2+8}{12}}$, then C of Theorem 15 is always lower bounded. Thus from Theorem 15, the LSI constant is upper bounded by

$$\lambda \le \frac{\sqrt{12}}{(1+\rho^{-1}|C|)2\sqrt{\beta^2m^2+8}} + \frac{3}{2}\rho^{-1} := \lambda_e. \tag{270}$$

Then we compare it in the following way

$$\begin{aligned}&\lambda_l - \lambda_e\\ &\ge \tfrac{8}{\beta m} + \rho^{-1}\left(\tfrac{6M(d+\beta)}{m} + 2\right) - \left(\tfrac{\sqrt{12}}{(1+\rho^{-1}|C|)2\sqrt{\beta^2m^2+8}} + \tfrac{3}{2}\rho^{-1}\right)\\ &\ge \frac{8}{\beta m} + \rho^{-1}\left(4(d+\beta) + 1/2\right) - \frac{\sqrt{3}}{(1+\rho^{-1}\beta|C|)\beta m} > 0.\end{aligned} \tag{271}$$

Thus, we confirmed that we do not need additional assumptions like Theorem 3. □

L.1. Proof of Theorem 15

Proof. Theorem 3.3 in Carlen & Loss (2004) is the direct consequence of lemma 3.1 and 3.2 in Carlen & Loss (2004).

The lemma 3.1 is that if π admits a spectral gap $\rho > 0$, and for some finite b,

$$\int_{\mathbb{R}^d} v^2 \ln v^2 d\mu \leq b \int_{\mathbb{R}^d} \|\nabla v\|^2 d\mu, \tag{272}$$

whenever

$$\int_{\mathbb{R}^d} v d\mu = 0, \tag{273}$$

and

$$\int_{\mathbb{R}^d} v^2 d\mu = 1. \tag{274}$$

Then μ admits a LSI with constant no larger than $\frac{b}{2} + \frac{3}{2}\rho$.

On the other hand, lemma 3.2 is that U is C^2 and π admits a spectral gap $\rho > 0$, and

$$-C = \inf_x \frac{\beta}{4}\|\nabla U(x)\|^2 - \frac{1}{2}\Delta U(x) - \pi e^2 U(x) > -\infty. \tag{275}$$

Then for all v satisfying

$$\int_{\mathbb{R}^d} v d\mu = 0, \tag{276}$$

and

$$\int_{\mathbb{R}^d} v^2 d\mu = 1, \tag{277}$$

we have

$$\int_{\mathbb{R}^d} v^2 \ln v^2 d\mu \leq \frac{1}{1 + \rho^{-1}\pi e^2} \int_{\mathbb{R}^d} \|\nabla v\|^2 d\mu. \tag{278}$$

We improve this lemma 3.2. In the proof of lemma 3.2, the key point is that the Lebesgue measure is used as a reference measure. There is an LSI for Lebesgue measure in the following way. For any function v on $\mathbb{R}^n$, which satisfies $\int_{\mathbb{R}^d} v^2 d^n x = 1$, we have

$$\int_{\mathbb{R}^d} v^2 \ln v^2 d^n x \leq \frac{1}{\pi e^2} \int_{\mathbb{R}^d} \|\nabla v\|^2 d^n x. \tag{279}$$

This relation plays an important role. We replace this relation to the Gaussian measure. Given a Gaussian measure $dN \propto e^{-\gamma \frac{\|x\|^2}{2}} d^n x$, which satisfies $CD(\gamma, \infty)$ from Proposition 5.7.1 in Bakry et al. (2013), we have

$$\int_{\mathbb{R}^d} v^2 \ln v^2 dN \leq \frac{1}{\gamma} \int_{\mathbb{R}^d} \|\nabla v\|^2 dN. \tag{280}$$

We use this Gaussian measure as a reference measure and improve the lemma 3.2.

Lemma 9. *(Improved result for lemma 3.2. in Carlen & Loss (2004)) U is C^2 and π admits a spectral gap $\rho > 0$, and γ is a positive constant and*

$$C := \inf_x \frac{\beta^2 \|\nabla U\|^2}{4} - \frac{\beta \Delta U}{2} - \frac{\beta}{\gamma} U - \frac{\gamma^2 + 2}{4}\|x\|^2 + \frac{d}{2} > -\infty, \tag{281}$$

Then for all v satisfying

$$\int_{\mathbb{R}^d} v d\mu = 0, \tag{282}$$

and

$$\int_{\mathbb{R}^d} v^2 d\mu = 1, \tag{283}$$

we have

$$\int_{\mathbb{R}^d} v^2 \ln v^2 d\mu \leq \frac{1}{(1 + \rho^{-1})\gamma} \int_{\mathbb{R}^d} \|\nabla v\|^2 d\mu. \tag{284}$$

This proof is given in Appendix L.1.1. We combined this result with lemma 3.1 in Carlen & Loss (2004), we have, if

$$C := \inf_x \frac{\beta^2 \|\nabla U\|^2}{4} - \frac{\beta \Delta U}{2} - \frac{\beta}{\gamma} V - \frac{\gamma^2 - 2}{4}\|x\|^2 + \frac{d}{2} > -\infty, \tag{285}$$

is satisfied, then π admits a logarithmic Sobolev inequality with constant λ no larger than

$$\lambda \leq \frac{1}{(1 + \rho^{-1}|C|)2\gamma} + \frac{3}{2}\rho^{-1}. \tag{286}$$

□

L.1.1. Proof of Lemma 9

Proof. For simplicity, we assume $d\mu = e^{-\beta U}$. Let us define

$$g := u e^{-\beta \frac{U}{2}} e^{\gamma \frac{\|x\|^2}{4}}. \tag{287}$$

Then for any $t \in (0, 1]$

$$\begin{aligned} &\int_{\mathbb{R}^d} \|\nabla v\|^2 d\mu - t\gamma^{-1} \int_{\mathbb{R}^d} v^2 \ln v^2 d\mu \\ &= (1-t) \textstyle\int_{\mathbb{R}^d} \|\nabla v\|^2 d\mu + t \left(\int_{\mathbb{R}^d} \|\nabla v\|^2 d\mu - \gamma^{-1} \int_{\mathbb{R}^d} v^2 \ln v^2 d\mu \right) \\ &\geq (1-t)\rho \textstyle\int_{\mathbb{R}^d} v^2 d\mu + t \left(\int_{\mathbb{R}^d} \|\nabla v\|^2 d\mu - \gamma^{-1} \int_{\mathbb{R}^d} v^2 \ln v^2 d\mu \right). \end{aligned} \tag{288}$$

Then, we will evaluate the second term. From the definition, we have

$$u = g e^{\beta \frac{U}{2}} e^{-\gamma \frac{\|x\|^2}{4}}. \tag{289}$$

Thus, we get

$$\nabla u = \left(\nabla g + \frac{\beta \nabla U}{2} g - \frac{\gamma}{2} x g\right) e^{\beta \frac{U}{2}} e^{-\gamma \frac{\|x\|^2}{4}}. \quad (290)$$

Thus, we have

$$\begin{aligned}
&\int_{\mathbb{R}^d} \|\nabla v\|^2 d\mu \\
&= \int_{\mathbb{R}^d} \left(\nabla g + \frac{\beta \nabla U}{2} g - \frac{\gamma}{2} x g\right)^2 dN \\
&= \int_{\mathbb{R}^d} \|\nabla g\|^2 dN \\
&+ \int_{\mathbb{R}^d} \left(\frac{\beta^2 \|\nabla U\|^2}{4} - \frac{\beta \Delta U}{2} - \frac{\gamma^2}{4}\|x\|^2 + \frac{d}{2}\right) g^2 dN.
\end{aligned} \quad (291)$$

Also, we have

$$\begin{aligned}
&\int_{\mathbb{R}^d} v^2 \ln v^2 d\mu \\
&= \int_{\mathbb{R}^d} g^2 \ln g^2 dN + \int_{\mathbb{R}^d} g^2 (\beta V - \frac{\gamma \|x\|^2}{2}) dN.
\end{aligned} \quad (292)$$

Thus, we have

$$\begin{aligned}
&\int_{\mathbb{R}^d} \|\nabla v\|^2 d\mu - t\gamma^{-1} \int_{\mathbb{R}^d} v^2 \ln v^2 d\mu \\
&\geq (1-t)\rho \int_{\mathbb{R}^d} g^2 dN \\
&+ t \int_{\mathbb{R}^d} \|\nabla g\|^2 dN \\
&+ t \int_{\mathbb{R}^d} \left(\frac{\beta^2 \|\nabla U\|^2}{4} - \frac{\beta \Delta U}{2} - \frac{\gamma^2}{4}\|x\|^2 + \frac{d}{2}\right) g^2 dN \\
&- t \int_{\mathbb{R}^d} g^2 \ln g^2 dN - t \int_{\mathbb{R}^d} g^2 (\beta V - \frac{\gamma \|x\|^2}{2}) dN \\
&\geq (1-t)\rho \int_{\mathbb{R}^d} g^2 dN \\
&+ t \int_{\mathbb{R}^d} \left(\frac{\beta^2 \|\nabla U\|^2}{4} - \frac{\beta \Delta U}{2} - \frac{\beta}{\gamma} V - \frac{\gamma^2 - 2}{4}\|x\|^2 + \frac{d}{2}\right) g^2 dN.
\end{aligned} \quad (293)$$

So if for any x,

$$\begin{aligned}
&(1-t)t^{-1}\rho \\
&+ \frac{\beta^2 \|\nabla U\|^2}{4} - \frac{\beta \Delta U}{2} - \frac{\beta}{\gamma} V - \frac{\gamma^2 - 2}{4}\|x\|^2 + \frac{d}{2} \geq 0,
\end{aligned} \quad (294)$$

is satisfied, we have

$$\int_{\mathbb{R}^d} \|\nabla v\|^2 d\mu - t\gamma^{-1} \int_{\mathbb{R}^d} v^2 \ln v^2 d\mu \geq 0. \quad (295)$$

For that purpose if we define

$$C' := \frac{\beta^2 \|\nabla U\|^2}{4} - \frac{\beta \Delta U}{2} - \frac{\beta}{\gamma} V - \frac{\gamma^2 - 2}{4}\|x\|^2 + \frac{d}{2} \quad (296)$$

and if this $|C'|$ is finite, and we set

$$t = \frac{1}{1 + \rho^{-1}|C'|}, \quad (297)$$

Eq.(295) wiil be satisfied. Thus, if

$$\begin{aligned}
C := &\inf_x \frac{\beta^2 \|\nabla U\|^2}{4} - \frac{\beta \Delta U}{2} \\
&- \frac{\beta}{\gamma} V - \frac{\gamma^2 - 2}{4}\|x\|^2 + \frac{d}{2} > -\infty,
\end{aligned} \quad (298)$$

we set $t = \frac{1}{1+\rho^{-1}|C|}$, we have

$$\int_{\mathbb{R}^d} v^2 \ln v^2 d\mu \leq \frac{1}{(1+\rho^{-1})\gamma} \int_{\mathbb{R}^d} \|\nabla v\|^2 d\mu. \quad (299)$$

□

M. Experimental settings

M.1. Construction of J

Here we explain how to build J in our experiments. First of all, we restrict the structure of J as

$$J = \overbrace{\begin{pmatrix}
\overbrace{\mathbf{0}}^{d} & \overbrace{J'_{12} I}^{d} & \dots & \overbrace{J'_{1N} I}^{d} \\
-J'_{12} I & \overbrace{\mathbf{0}}^{d} & \dots & \\
\vdots & & \ddots & \vdots \\
-J'_{1N} I & \dots & & \overbrace{\mathbf{0}}^{d}
\end{pmatrix}}^{dN}. \quad (300)$$

where $\overbrace{\mathbf{0}}^{d}$ means the $d \times d$ matrix whose entries are all zero and $\overbrace{J'_{12} I}^{d}$ means the $d \times d$ matrix whose entries are $J'_{12} \in \mathbb{R}$ times an identity matrix. Thus, this J is surely skew-symmetric.

Since the drift function is defined as $\nabla u_\alpha(x^{\otimes N}, z) := \nabla u^{\otimes N}(x^{\otimes N}, z) + \alpha J \nabla u^{\otimes N}(x^{\otimes N}, z)$ and $\nabla U_\alpha^{\otimes N} = \sum_z \nabla u_\alpha(x^{\otimes N}, z)/|\mathrm{Z}|$, for example, the first particle $X_t^{(1)}$ moves with the dynamics;

$$dX_t^{(1)} = -\left(\nabla U(X_t^{(1)}) + \sum_{n=2}^N J'_{1n} \nabla U(X_t^{(n)})\right) dt + \sqrt{2\beta^{-1}} dw_t. \quad (301)$$

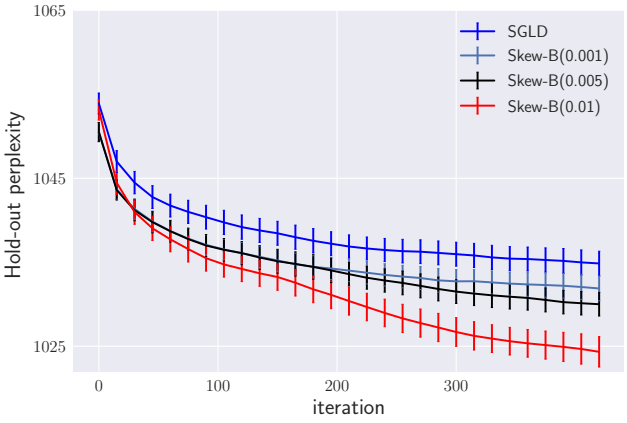


Figure 5. LDA experiments (Averaged over 10 trials)

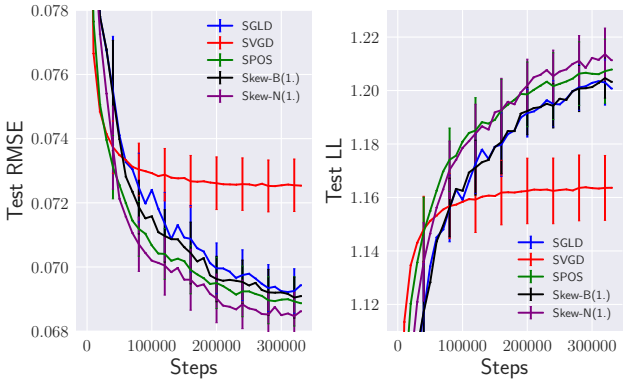


Figure 6. BNN experiments for kin8 (Averaged over 20 trials)

This means that the interaction term, $J'_{1n}\nabla U(X_t^{(n)})$, has the equal weights J'_{1n} to each dimension so that the gradient information $\nabla U(X_t^{(n)})$ is preserved. Also by setting the block diagonal element of J as 0 ($\{J_{nn}\}_{n=1}^N = 0$), the self interaction term becomes 0 ($J'_{11}\nabla U(X_t^{(1)}) = 0$).

In this work, as we described in the main paper, we prepared three types of J, *skew-N*, *skew-B*, and *skew-k*. As for *skew-N*, each entry follows standard Gaussian distribution. As for *skew-B*, each entries follow Bernoulli distribution, so each $J'_{i\neq j}$ takes $+1$ or 0 randomly with equal probability. After that, since we assumed that the Frobenius norm of J is below 1, we divided each element by N^2 where N is the number of particles. Next, as for *skew-k*, we first prepared the Gram matrix K, of which element is the Gaussian kernel $K_{ij} = k(X_0^{(i)}, X_0^{(j)}) = \exp(-\|X_0^{(i)} - X_0^{(j)}\|^2/(2h))$ and bandwidth h is calculated by the median trick, which is the median of all the combination of the distances between particles. Then, we calculate $K^{1/2}J(\textit{skew-B})K^{1/2}$. Then, we divide $K^{1/2}J(\textit{skew-N})K^{1/2}$ by N^2. The motivation of introducing *skew-k* is to smooth *skew-B* matrix by the initial position of the particles via the kernel Gram matrix.

M.2. Model settings and additional results

The settings of BLR and BNN experiments are exactly as same as Liu & Wang (2016). We used adagrad optimizer for SVGD.

The settings of LDA is the same as Patterson & Teh (2013) and Liu et al. (2019b).

In all the experiments, we found that setting $\alpha = 0$ in the early stage of the sampling is useful to make the algorithm stable. Thus, we set $\alpha = 0$ for the first several steps. After that, we set $\alpha \neq 0$.

Ornstein-Uhlenbeck process

There is a formula of the W_2 distance between two Gaussian distributions (Wibisono, 2018). Given two Gaussian, $N_1(\mu_1, \Sigma_1)$ and $N_2(\mu_2, \Sigma_2)$,

$$W_2(N_1, N_2)^2 = \|\mu_1 - \mu_2\|^2 + \mathrm{Tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}). \tag{302}$$

Then, given a OU

$$dX_t = \Sigma^{-1}(X_t - \mu)dt + \sqrt{2}dw(t), \tag{303}$$

its solution is given by

$$x(t) = x_0 e^{-\Sigma^{-1}t} + \mu(1 - e^{-\Sigma^{-1}t}) + \sqrt{2}\int_0^t e^{-\Sigma^{-1}(t-s)}dw(s), \tag{304}$$

thus, its expectation and variance is

$$\begin{aligned}\mathbb{E}x(t) &= x_0 e^{-\Sigma^{-1}t} + \mu(1 - e^{-\Sigma^{-1}at}) \\ &= \mu + (x_0 - \mu)e^{-\Sigma^{-1}t}\end{aligned} \tag{305}$$

and

$$\mathrm{Var}(x(t)) = \Sigma(1 - e^{-2\Sigma^{-1}t}) \tag{306}$$

Thus, if we write the distribution at time t of above OU process as $N_t(\mu(t), \Sigma(t))$, the wasserstein distance between N_t and its stationary $N(\mu, \Sigma)$ is given by

$$\begin{aligned}&W_2(N_t, N)^2 \\ &= \|x_0 - \mu\|^2 e^{-\Sigma^{-1}t} + \|\Sigma^{1/2} - \Sigma^{1/2}(1 - e^{-2\Sigma^{-1}t})^{1/2}\|_F^2.\end{aligned} \tag{307}$$

Thus, we can see that W_2 distance decreases exponentially.

In the numerical experiments, to calculate the MMD and the Energy distance, we draw 2000 samples from the target distribution. Then we calculate the MMD and the Energy distance between those samples and the particles of the ensemble method at each time steps. We used RBF kernel for MMD.

BNN classification

We performed the experiments of MNIST classification on a neural network model with 2 hidden layers, of which unit numbers are 100 and 50. We used Relu activation functions.

Symbolslist

Sign	Description
$U(x)$	A potential function of the gibbs measure $\pi \sim e^{-\beta U(x)}$
W_2	2-Wasserstein distance of which cost function is the Euclidean distance
$g(x, Q)$	An unbiased estimator of the gradient of $U(x)$
π	Stationary (target) measure $\pi \sim e^{-\beta U}$
$X_t^{\otimes N}$	A random variable following the ensemble method at time t
$X_t^{(i)}$	The i-th random variable of the $X_t^{\otimes N}$
X_t	A random variable following standard SGLD at time t
$\nu_{kh}^{\otimes N}$	The measure at time kh induced by the continuous proposed ensemble method
ν_{kh}	The measure at time kh induced by the continuous SGLD dynamics
$\mu_{kh}^{\otimes N}$	The measure at time kh induced by the discretized proposed ensemble method
μ_{kh}	The measure at time kh induced by the discretized SGLD dynamics
α	A magnitude of a interaction
J	A skew-symmetric matrix of an interaction
k	An iteration of the algorithm
h	A step size
t	A (continusous) time
f	A test function with lipschitzness L_f
$\lambda(\alpha, N)$	An upper bound of the LSI constant of the proposed ensemble dynamics
λ_e	An upper bound of the LSI constant of the standard Langevin dynamics
$\rho(\alpha, N)$	A spectral gap of the proposed ensemble method
ρ_0	A spectral gap of the standard Langevin dynamics