entropy

MDPI

Article

Accelerated Diffusion-Based Sampling by the Non-Reversible Dynamics with Skew-Symmetric Matrices

Futoshi Futami [1,*], Tomoharu Iwata [1], Naonori Ueda [1] and Issei Sato [2]

[1] Communication Science Laboratories, NTT, Hikaridai, Seika-cho, "Keihanna Science City", Kyoto 619-0237, Japan; tomoharu.iwata.gy@hco.ntt.co.jp (T.I.); naonori.ueda.fr@hco.ntt.co.jp (N.U.)
[2] Department of Computer Science, Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan; sato@g.ecc.u-tokyo.ac.jp
* Correspondence: futoshi.futami.uk@hco.ntt.co.jp

Abstract: Langevin dynamics (LD) has been extensively studied theoretically and practically as a basic sampling technique. Recently, the incorporation of non-reversible dynamics into LD is attracting attention because it accelerates the mixing speed of LD. Popular choices for non-reversible dynamics include underdamped Langevin dynamics (ULD), which uses second-order dynamics and perturbations with skew-symmetric matrices. Although ULD has been widely used in practice, the application of skew acceleration is limited although it is expected to show superior performance theoretically. Current work lacks a theoretical understanding of issues that are important to practitioners, including the selection criteria for skew-symmetric matrices, quantitative evaluations of acceleration, and the large memory cost of storing skew matrices. In this study, we theoretically and numerically clarify these problems by analyzing acceleration focusing on how the skew-symmetric matrix perturbs the Hessian matrix of potential functions. We also present a practical algorithm that accelerates the standard LD and ULD, which uses novel memory-efficient skew-symmetric matrices under parallel-chain Monte Carlo settings.

Keywords: Markov Chain Monte Carlo; Langevin dynamics; Hamilton Monte Carlo; non-reversible dynamics

Citation: Futami, F.; Iwata, T.; Ueda, N.;Sato, I. Accelerated Diffusion-Based Sampling by the Non-Reversible Dynamics with Skew-Symmetric Matrices. *Entropy* **2021**, *23*, 993. https://doi.org/10.3390/e23080993

Academic Editor: Pierre Alquier

Received: 21 June 2021
Accepted: 27 July 2021
Published: 30 July 2021

1. Introduction

Sampling is one of the most widely used techniques for the approximation of posterior distribution in Bayesian inference [1]. Markov Chain Monte Carlo (MCMC) is widely used to obtain samples. In MCMC, Langevin dynamics (LD) is a popular choice for sampling from high-dimensional distributions. Each sample in LD moves toward a gradient direction with added Gaussian noise. LD efficiently explore around a mode of a target distribution using the gradient information without being trapped by local minima thanks to added Gaussian noise. Many previous studies theoretically and numerically proved LD's superior performance [2–5]. Since non-reversible dynamics generally improves mixing performance [6,7], research on introducing non-reversible dynamics to LD for better sampling performance is attracting attention [8].

There are two widely known non-reversible dynamics for LD. One is underdamped Langevin dynamics (ULD) [9], which uses second-order dynamics. The other introduces perturbation, which consists of multiplying the skew-symmetric matrix by a gradient [8]. Here, we refer to the matrix as skew matrices for simplicity and this perturbation technique as skew acceleration. Much research has been done on ULD theoretically [9–11] and ULD is widely used in practice, which is also known as stochastic gradient Hamilton Monte Carlo [12]. In contrast, the application of the skew acceleration for standard Bayesian models is quite limited even though it is expected to show superior performance theoretically [8].

For example, skew acceleration has been analyzed focusing on sampling from Gaussian distributions [13–17], although assuming Gaussian distributions in Bayesian models is restrictive in practice. A recent study [8] theoretically showed that skew acceleration accelerates the dynamics around the local minima and saddle points for non-convex functions. Another work [18] clarified that the skew acceleration theoretically and numerically improves mixing speed when used as interactions between chains in parallel sampling schemes for non-convex Bayesian models.

Compared to ULD, what seems to be lacking for skew acceleration is a theoretical understanding of issues that are important to practitioners. The most significant problem is that no theory exists for selecting skew matrices. In existing studies, introducing a skew matrix into LD results in equal or faster convergence, denoting that a bad choice of skew matrix results in no acceleration. Thus, choosing appropriate skew matrices is critical. Furthermore, although ULD's acceleration has been analyzed quantitatively, existing studies have only analyzed skew acceleration qualitatively. Thus, it is difficult to justify the usefulness of skew acceleration in practice compared to ULD. Another issue is that introducing skew matrices requires a vast memory cost in many practical Bayesian models.

The purpose of this study is to solve these problems from theoretical and numerical viewpoints and establish a practical algorithm for skew acceleration. The following are the two major contributions of this work.

Our contribution 1: We present a convergence analysis of skew acceleration for standard Bayesian model settings, including non-convex potential functions using Poincaré constants [19]. The major advantage of Poincaré constants is that we can analyze skew acceleration through a Hessian matrix and its eigenvalues and develop a practical theory about the selection of J and the quantitative assessment of skew acceleration.

Furthermore, we propose skew acceleration for ULD and present convergence analysis for the first time. Since ULD shows faster convergence than LD, combining skew acceleration with ULD is promising.

Our contribution 2: We develop a practical skew accelerated sampling algorithm for a parallel sampling setting with novel memory-efficient skew matrices. Since a naive implementation of skew acceleration requires a large memory cost to store skew matrices, memory-efficiency is critical in practice. We also present a non-asymptotic theoretical analysis for our algorithm in both LD and ULD settings under a stochastic gradient and Euler discretization. We clarify that introducing skew matrices accelerates the convergence of continuous dynamics, although it increases the discretization and stochastic gradient error. Then to the best of our knowledge, we propose the first algorithm that adaptively controls this trade-off using the empirical distribution of the parallel sampling scheme.

Finally, we verify our algorithm and theory in practical Bayesian problems and compare it with other sampling methods.

Notations: I_d denotes a $d \times d$ identity matrix. Capital letters such as X represent random variables, and lowercase letters such as x represent non-random real values. $\cdot$, $\|\cdot\|$ and $|\cdot|$ denote Euclidean inner products, distances and absolute values.

2. Preliminaries

In this section, we briefly introduce the basic settings of LD and non-reversible dynamics for the posterior distribution sampling in Bayesian inference.

2.1. LD and Stochastic Gradient LD

First, we introduce the notations and the basic settings of LD and stochastic gradient LD (SGLD), which is a practical extension of LD. Here z_i denotes a data point in space $\mathbb{Z}$, $|\mathcal{Z}|$ denotes the total number of data points, and $x \in \mathbb{R}^d$ corresponds to the parameters of a given model, which we want to sample. Our goal is to sample from the target distribution with density $d\pi(x) \propto e^{-\beta U(x)}dx$, where potential function $U(x)$ is the summation of $u : \mathbb{R}^d \times \mathbb{Z} \to \mathbb{R}$, i.e., $U(x) = \frac{1}{|\mathcal{Z}|}\sum_{i=1}^{|\mathcal{Z}|} u(x, z_i)$. Function $u(\cdot, \cdot)$ is continuous and non-

convex. The explicit assumptions made for it are discussed in Section 3.1. The SGLD algorithm [2,3] is given as a recursion:

$$X_{k+1} = X_k - h\nabla\hat{U}(X_k) + \sqrt{2h\beta^{-1}}\epsilon_k, \tag{1}$$

where $h \in \mathbb{R}^+$ is a step size, $\epsilon_k \in \mathbb{R}^d$ is a standard Gaussian random vector, β is a temperature parameter of π, and $\nabla\hat{U}(X_k)$ is a conditionally unbiased estimator of true gradient $\nabla U(X_k)$. This unbiased estimate of the true gradient is suitable for large-scale data set since we can use not the full gradient, but a stochastic version obtained through a randomly chosen subset of data at each time step. This means that we can reduce the computational cost to calculate the gradient at each time step.

The discrete time Markov process in Equation (1) is the discretization of the continuous-time LD [2]:

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

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

2.2. Poincaré Inequality and Convergence Speed

In sampling, we are interested in the convergence speed to the stationary measure. The speed is often characterized by the *the generator* associated with Equation (2) and defined as:

$$\begin{aligned}\mathcal{L}f(X_t) :&= \lim_{s\to 0^+} \frac{\mathbb{E}(f(X_{t+s})|X_t) - f(X_t)}{s} \\ &= \left(-\nabla U(X_t)\cdot\nabla + \beta^{-1}\Delta\right)f(X_t),\end{aligned} \tag{3}$$

where Δ denotes a standard Laplacian on $\mathbb{R}^d$ and $f \in \mathcal{D}(\mathcal{L})$ and $\mathcal{D}(\mathcal{L}) \subset \mathrm{L}^2(\pi)$ denote the $\mathcal{L}$ domain. This $-\mathcal{L}$ is a self-adjoint operator, which has only discrete spectrums (eigenvalues). 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 $\rho_0(>0)$. This spectral gap is closely related to Poincaré inequality. Internal energy is defined:

$$\mathcal{E}(f) := -\int_{\mathbb{R}^d} f\mathcal{L}f d\pi. \tag{4}$$

Please note that $\mathcal{E}(f) > 0$ is satisfied. Then π with $\mathcal{L}$ satisfies the Poincaré inequality with constant c, if for any $f \in \mathcal{D}(\mathcal{L})$, π with $\mathcal{L}$ satisfies:

$$\int f^2 d\pi - \left(\int f d\pi\right)^2 \le c\mathcal{E}(f). \tag{5}$$

The spectral gap characterizes this constant $c \le \frac{1}{\rho_0}$, which holds (see Appendix A.2 for details). We refer to best constant c as the Poincaré constant [19]. For notational simplicity, we define $m_0 := \frac{1}{c}$ and refer to this m_0 as the Poincaré constant.

In sampling, crucially, Poincaré inequality dominates the convergence speed in χ^2 divergence:

$$\int\left(\frac{d\mu_t}{d\pi} - 1\right)^2 d\pi := \chi^2(\mu_t\|\pi) \le e^{-\frac{2m_0}{\beta}t}\chi^2(\mu_0\|\pi), \tag{6}$$

where μ_t denotes the measure at time t induced by Equation (2) and μ_0 is the initial measure (see Appendix A.3 for details). Thus, the larger Poincaré constant m_0 is, the faster convergence we have.

2.3. Non-Reversible Dynamics

In this section, we introduce the non-reversible dynamics. π with $\mathcal{L}$ is reversible if for any test function $f, g \in \mathcal{D}(\mathcal{L})$, π with $\mathcal{L}$ satisfies

$$\int_{\mathbb{R}^d} f\mathcal{L}g d\pi = \int_{\mathbb{R}^d} g\mathcal{L}f d\pi. \tag{7}$$

If this is not satisfied, π with $\mathcal{L}$ is non-reversible [19].

We introduce two non-reversible dynamics for LD. The first is ULD, which is given as

$$\begin{aligned} dX_t &= \Sigma^{-1} V_t dt, \\ dV_t &= -\nabla U(X_t)dt - \gamma\Sigma^{-1}V_t dt + \sqrt{2\gamma\beta^{-1}}dw_t, \end{aligned} \tag{8}$$

where $V \in \mathbb{R}^d$ is an auxiliary random variable, $\gamma \in \mathbb{R}$ is a positive constant, and Σ is the variance of the stationary distribution of auxiliary random variable V. The stationary distribution is $\tilde{\pi} := \pi \otimes \mathcal{N}(0, \Sigma) \propto e^{-\beta U(x) - \frac{1}{2}\Sigma^{-1}\|v\|^2}$, where $\mathcal{N}$ denotes a Gaussian distribution. The superior performance of ULD compared with LD has been studied rigorously [9–11]. ULD's convergence speed is also characterized by the Poincaré constant [20]. In practice, we use discretization and the stochastic gradient for ULD, which is called the stochastic gradient Hamilton Monte Carlo (SGHMC) [10]. The second non-reversible dynamics is the skew acceleration given as

$$dX_t = -(I + \alpha J)\nabla U(X_t)dt + \sqrt{2\beta^{-1}}dw_t, \tag{9}$$

where J is a real value skew matrix and $\alpha \in \mathbb{R}^+$ is a positive constant. We call this dynamics S-LD. The stationary distribution of S-LD is still π, and S-LD shows faster convergence and smaller asymptotic variance [13–15,18].

3. Theoretical Analysis of Skew Acceleration

In this section, we present a theoretical analysis of skew acceleration in LD and ULD in standard Bayesian settings. We analyze acceleration through the Poincaré constant and connect it with the eigenvalues of the Hessian matrix, which allows us to obtain a practical criterion to choose skew matrices and quantitatively evaluate acceleration. We focus on a setting where a continuous SDE and a full gradient of the potential function is used in this section. The discretized SDE and stochastic gradient are discussed in Section 4.

3.1. Acceleration Characterization by the Poincaré Constant

First, we introduce the same four assumptions as a previous work [2], which showed the existence of the Poincaré constant about m_0 for LD (see Appendix C for details).

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

$$|u(0,z)| \leq A, \quad \|\nabla u(0,z)\| \leq B. \tag{10}$$

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

$$\|\nabla u(x,z) - \nabla u(y,z)\| \leq M\|x - y\|. \tag{11}$$

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

$$-x \cdot \nabla u(x,z) \leq -m\|x\|^2 + b. \tag{12}$$

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

$$\kappa_0 := \log \int_{\mathbb{R}^d} e^{\|x\|^2} p_0(x) dx < \infty. \tag{13}$$

Please note that these assumptions allow us to consider the non-convex potential functions, which are common in practical Bayesian models. Furthermore, we make the following assumption about J.

Assumption 5. *The operator norm of J is bounded:*

$$\|J\|_2 \leq 1. \tag{14}$$

This means that the largest singular value of J is below 1.

Under these assumptions, we present the convergence behavior of skew acceleration using the Poincaré constant. First, we present the following S-LD result.

Theorem 1. *Under Assumptions 1–5, the S-LD of Equation (9 has exponential convergence,*

$$\chi^2(\mu_t^\alpha \| \pi) \leq e^{-\frac{2m(\alpha)}{\beta}t} \chi^2(\mu_0 \| \pi), \tag{15}$$

where μ_t^α is the measure at time t induced by S-LD and $m(\alpha)$ is the Poincaré constant of S-LD defined by its generator

$$\mathcal{L}_\alpha f(x) := \Big(-(I + \alpha J)\nabla U(x) \cdot \nabla + \beta^{-1}\Delta\Big) f(x). \tag{16}$$

Furthermore, $m(\alpha)$ satisfies $m(\alpha) \geq m_0$.

The proof is shown in Appendix C. This theorem states that introducing the skew matrices accelerates the convergence of LD by improving the convergence rate from m_0 to $m(\alpha)$. Although [18] obtained a similar result, we used the Poincaré constant and derived an explicit criterion when $m(\alpha) = m_0$ holds, as we discuss below.

Next, we also introduce skew acceleration in ULD. Since ULD shows faster convergence than LD in standard Bayesian settings [10,11], it is promising to combine skew acceleration with ULD to obtain a more efficient sampling algorithm. For that purpose, we propose the following SDE:

$$dX_t = \Sigma^{-1} V_t dt + \alpha_1 J_1 \nabla U(X_t) dt, \tag{17}$$

$$dV_t = -\nabla U(X_t) dt - \gamma(\Sigma^{-1} + \alpha_2 J_2) V_t dt + \sqrt{2\gamma\beta^{-1}} dw_t, \tag{18}$$

where J_1 and J_2 are real value skew matrices and α_1 and α_2 are positive constants. We assume that J_1 and J_2 satisfy Assumption 5. We refer to this method as skew underdamped Langevin dynamics (S-ULD) whose stationary distribution is $\tilde{\pi} = \pi \otimes \mathcal{N}(0, \Sigma) \propto e^{-\beta U(x) - \frac{1}{2}\Sigma^{-1}\|v\|^2}$. See Appendix B for details, which include discussions on other combinations of skew matrices. As for S-ULD, we need an additional assumption about the initial condition of V_0:

Assumption 6. *(Initial condition) Initial probability distribution $\mu_0(x,v)$ of (X_0,V_0) has a bounded and strictly positive density p_0 that satisfies,*

$$\kappa_0 := \log \int_{\mathbb{R}^{2d}} e^{\|x\|^2+\|v\|^2} p_0((x,v))dxdv < \infty. \tag{19}$$

We then provide the following convergence theorem that resembles S-LD.

Theorem 2. *Under Assumptions 1–3, 5, 6, S-ULD has exponential convergence in χ^2 divergence and its convergence rate is also characterized by $m(\alpha)$ as defined in Theorem 1. S-ULD's convergence equals or exceeds ULD, of which convergence rate is characterized by m_0.*

See Appendix C.2 for details. From these theorems, we confirmed that skew acceleration is effective in both S-LD and S-ULD, and the convergence speed is characterized by Poincaré constant $m(\alpha)$ defined by Equation (16).

3.2. Skew Acceleration from the Hessian Matrix

Our goal is to clarify what choices of J induce $m(\alpha) > m_0$, which leads to acceleration. Therefore, we discuss how Poincaré constant $m(\alpha)$ is connected to the eigenvalues and eigenvectors of the perturbed Hessian matrix $(I+\alpha J)\nabla^2 U(x)$. Next, we introduce the notations. We express the Hessian of $U(x)$ as $H(x)$ and the perturbed Hessian matrix as $H'(x) := (I+\alpha J)H(x)$. Please note that H is a real symmetric matrix, which has real eigenvalues and diagonalizable. On the other hand, since H' is not symmetric, it has complex eigenvalues, although diagonalization is not assured (see Appendix E). We express pairs of eigenvectors and eigenvalues of $H'(x)$ as $\{(v_i^\alpha(x), \lambda_i^\alpha(x))\}_{i=1}^d$, which are ordered as $\mathrm{Re}(\lambda_1^\alpha(x))) \leq \cdots \leq \mathrm{Re}(\lambda_d^\alpha(x))$. Here, $\mathrm{Re}(\lambda_1^\alpha(x))$ expresses the real part of complex value λ_1^α and Im denotes the imaginary part. We express those of $H(x)$ as $\{(v_i^0(x), \lambda_i^0(x))\}_{i=1}^d$ and order them as $\lambda_1^0(x) \leq \cdots \leq \lambda_d^0(x)$.

3.2.1. Strongly Convex Potential Function

Assume that U is an m-strongly convex function, where for all $x \in \mathbb{R}^d$, $m \leq \lambda_1^0(x)$ holds. Poincaré constant m_0 of LD satisfies $m_0 = m$ [19]. For the skew acceleration, since Poincaré constant satisfies $m(\alpha) = m'(\alpha)$, where $m'(\alpha)$ is the best constant that satisfies, for all x, $m'(\alpha) \leq \mathrm{Re}\big(\lambda_1^\alpha(x)\big)$ (see Appendix D.1). Therefore, studying the Poincaré constant is equivalent to studying the smallest (real part of the) eigenvalue of the Hessian matrix. Thus, the relation between $\lambda_1^0(x)$ and $\mathrm{Re}\big(\lambda_1^\alpha(x)\big)$ must be studied. The following theorem describes how the skew matrices change the smallest eigenvalue.

Theorem 3. *For all $x \in \mathbb{R}^d$, the real parts of the eigenvalues of H' satisfy*

$$m \leq \lambda_1^0(x) \leq \mathrm{Re}(\lambda_1^\alpha(x)) \leq \cdots \leq \mathrm{Re}(\lambda_d^\alpha(x)) \leq \lambda_d^0(x). \tag{20}$$

The condition of $\lambda_1^0(x) = \mathrm{Re}\big(\lambda_1^\alpha(x)\big)$ is shown in Remark 1.

Remark 1. *Denote the set of the eigenvectors of eigenvalue $\lambda_1^0(x)$ as V_1^0. If $V_1^0 = \{v\}$ and $Jv = 0$, then $\lambda_1^0(x) = \mathrm{Re}\big(\lambda_1^\alpha(x)\big)$ holds. If the cardinality of set V_1^0 is larger than 1, and vectors $v, v' \in V_1^0$ exist, such that $\lambda_1^0 \alpha Jv = (\mathrm{Im}\big(\lambda_1^\alpha\big))v'$ and $\lambda_1^0 \alpha Jv' = -(\mathrm{Im}\big(\lambda_1^\alpha\big))v$, then $\lambda_1^0(x) = \mathrm{Re}\big(\lambda_1^\alpha(x)\big)$ holds.*

Refer to Appendix F for the proof. This is an extension of previous work [8,13]. If $\lambda_1^0(x) < \mathrm{Re}\big(\lambda_1^\alpha(x)\big)$ is satisfied for all x, we have $m_0 < m(\alpha)$, i.e., acceleration occurs. We discuss how to construct J such that $\lambda_1^0(x) < \mathrm{Re}\big(\lambda_1^\alpha(x)\big)$ holds in Section 3.3.

3.2.2. Non-Convex Potential Function

The previous work [21] clarified that the Poincaré constant of the non-convex function is characterized by the negative eigenvalue of the saddle point. As shown in Figure 1, denote x_1 as the global minima, and x_2 is the local minima which has the second smallest value in $U(x)$. We express the saddle point with index one, i.e., there is only one negative eigenvalue at the point, between x_1 and x_2 as x^*. This means that the eigenvalues of $H(x^*)$ satisfies $\lambda_1^0(x^*) < 0 < \lambda_2^0(x^*) < \cdots < \lambda_d^0(x^*)$. [21] clarified that the saddle point x^* characterizes the Poincaré constant as

$$m_0^{-1} \propto \frac{1}{|\lambda_1(x^*)|} e^{\beta(U(x^*)-U(x_1)-U(x_2))}. \tag{21}$$

When skew matrices are introduced, [8] clarified the following relation:

Theorem 4. *([8]) $\lambda_1^\alpha(x^*) \leq \lambda_1^0(x^*) < 0$ and equality holds only if $Jv_1^\alpha(x^*) = 0$.*

Note $\lambda_1^\alpha(x^*)$ is not a complex number. Thus, the skew acceleration reduces the negative eigenvalue and leads to a larger Poincaré constant (see Appendix D.2) and results in faster convergence.

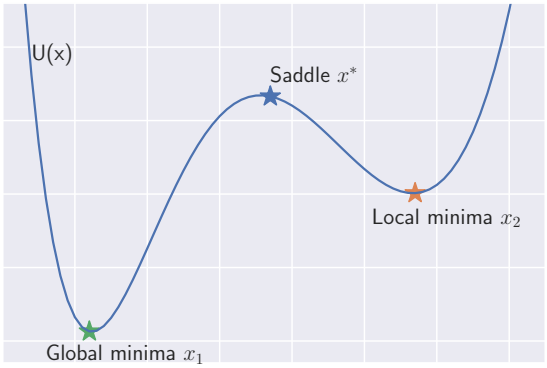

Figure 1. Double-potential example: Poincaré constant is related to the eigenvalue at x^*.

In conclusion, introducing the skew matrix changes the Hessian's eigenvalues and increase the Poincaré constant. If $\lambda_1^0(x) \neq \mathrm{Re}(\lambda_1^\alpha(x))$ is satisfied, this leads to faster convergence for both convex and non-convex potential functions.

3.3. Choosing J

In this section, we present a method for choosing J that leads to $\lambda_1^0(x) \neq \mathrm{Re}(\lambda_1^\alpha(x))$ to ensure the acceleration based on the equality conditions in Theorems 3 and 4. Combining these theorems, we obtain the following criterion:

Remark 2. *Given a point x, $\lambda_1^0(x) \neq \mathrm{Re}(\lambda_1^\alpha(x))$ holds if either the following conditions are satisfied: (i) when $V_1^0 = \{v\}$, $Jv \neq 0$ is satisfied. (ii) when $|V_1^0| > 1$, $Jv \neq 0$ holds for any $v \in V_1^0$, and for any $v, v' \in V_1^0$, $\lambda_1^0 \alpha Jv = (\mathrm{Im}(\lambda_1^\alpha))v'$ and $\lambda_1^0 \alpha Jv' = -(\mathrm{Im}(\lambda_1^\alpha))v$ are not satisfied.*

The first condition (i) is easily satisfied if we choose J such that $\mathrm{Ker}J = \{0\}$. On the other hand, the second condition (ii) is difficult to verify since H and its eigenvalues and eigenvectors generally depend on the current position of X_t. Instead of evaluating eigenvalues and eigenvectors of H and H' directly, we use the random matrix property shown in the next theorem.

Theorem 5. *Suppose the upper triangular entries of J follow a probability distribution that is absolutely continuous with respect to the Lebesgue measure. If* $\mathrm{Ker}J = \{0\}$ *is satisfied, then given a point* $x \in \mathbb{R}^d$, $\lambda_1^0(x) \neq \mathrm{Re}(\lambda_1^\alpha(x))$ *holds with probability 1.*

The proof is given in Appendix G.1. From this theorem, we simply generate J from some probability distribution, such as the Gaussian distribution. Then, we check whether $\mathrm{Ker}J = \{0\}$ holds. If $\mathrm{Ker}J = \{0\}$ does not hold, we generate a random matrix J again.

The above theorem is valid only at a given evaluation point x. We can extend the above theorem to all the points over the path of the discretized dynamics (see Appendix G.3). With this procedure, we can theoretically ensure that acceleration occurs with probability one for discretized dynamics.

3.4. Qualitative Evaluation of The Acceleration

So far, we have discussed skew acceleration qualitatively but not quantitatively. Although acceleration's quantitative evaluation is critical for practical purposes, to the best of our knowledge, no existing work has addressed it. In this section, we present a formula that quantitatively assesses skew acceleration by analyzing the eigenvalues of the Hessian matrix.

Theorem 6. *With the identical notation as in Theorem 3, for all x, we have*

$$\mathrm{Re}(\lambda_1^\alpha(x)) = \lambda_1^0(x) + \alpha^2 \sum_{k=2}^{d} \frac{\lambda_1^0(x)\lambda_k^0(x)|v_k^0(x)Jv_1^0(x)|^2}{\lambda_k^0(x) - \lambda_1^0(x)} + \mathcal{O}(\alpha^3). \tag{22}$$

In particular, at saddle point x^*, we have

$$\lambda_1^\alpha(x^*) = \lambda_1^0(x^*) + \alpha^2 \sum_{k=2}^{d} \frac{\lambda_1^0(x^*)\lambda_k^0(x^*)|v_k^0(x^*)Jv_1^0(x^*)|^2}{\lambda_k^0(x^*) - \lambda_1^0(x^*)} + \mathcal{O}(\alpha^3). \tag{23}$$

The proofs are shown in Appendix H. When focusing on Equation (22, if $U(x)$ is a strongly convex function, since for all $k > 1$, $\lambda_k(x) > \lambda_1(x) > 0$ holds and the second term in Equation (22 is positive. From this, $\mathrm{Re}(\lambda_1^\alpha(x)) > \lambda_1^0(x)$ holds. A similar relation holds for $\mathrm{Re}(\lambda_d^\alpha(x))$. In Equation (23, $\lambda_1^\alpha(x^*) < \lambda_1^0(x^*) < 0$ holds. Thus, the changes of the Poincaré constants are proportional to α^2. With these formulas, we can quantitatively evaluate the acceleration. We present numerical experiments to confirm our theoretical findings in Section 6.1.

4. Practical Algorithm for Skew Acceleration

In this section, we discuss skew acceleration in more practical settings compared to Section 3. First, we discuss the memory issue for storing J and the discretization of SDE and the stochastic gradient, which are widely used techniques in Bayesian inference. Finally, we present a practical algorithm for skew acceleration.

4.1. Memory Issue of Skew Acceleration and Ensemble Sampling

For d-dimensional Bayesian models, we need $\mathcal{O}(d^2)$ memory space to store skew matrices Js, and this is difficult for high-dimensional models. Instead of storing J, we can randomly generate Js at each time step following Theorem 5. However, we experimentally confirmed that using different Js at each step does not accelerate the convergence (see Section 6). Thus, we need to use a fixed J during the iterations.

As discussed below, we found that the previously proposed accelerated parallel sampling [18] can be a practical algorithm to resolve this memory issue. In that method, we simultaneously updated N samples of the model's parameters with correlation. In such a parallel sampling scheme, a correlation exists among multiple Markov chains, it is more efficient than a naive parallel-chain MCMC, where the samples are independent.

We express the n-th sample at time t as $X_t^{(n)} \in \mathbb{R}^d$ and the joint state of all samples 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 $\pi^{\otimes N} := \pi \otimes \cdots \otimes \pi(x^{\otimes N}) \propto e^{-\beta \sum_{i=1}^N U(x^{(i)})}$. We express the sum of the potential function as $U^{\otimes N} := \sum_{i=1}^N U(x^{(i)})$. We then consider the following dynamics:

$$dX_t^{\otimes N} = -(I_{dN} + \alpha J)\nabla U^{\otimes N}(X_t^{\otimes N})dt + \sqrt{2\beta^{-1}}dw_t, \tag{24}$$

$$\nabla U^{\otimes N}(X_t^{\otimes N}) := \left(\nabla U(X_t^{(1)}), \ldots, \nabla U(X_t^{(N)})\right)^\top. \tag{25}$$

We call this dynamics skew parallel LD (S-PLD). N-independent parallel LD (PLD) is coupled with the skew matrix. Since each chain in PLD is independent of the other, the Poincaré constant of PLD is also m_0. [18] argued that the Poincaré constant of S-PLD, $m(\alpha, N)$, satisfies $m(\alpha, N) \geq m_0$. This means S-PLD shows faster convergence than PLD. As discussed in Section 3.2, these Poincaré constants are characterized by the smallest eigenvalue of the Hessian matrix $\nabla^2 U^{\otimes N}(x^{\otimes N})$ and $(I_{dN} + \alpha J)\nabla^2 U^{\otimes N}(x^{\otimes N})$ where $x^{\otimes N} \in \mathbb{R}^{dN}$. We denote these smallest eigenvalues as $\lambda_1^0(x^{\otimes N})$ and $\mathrm{Re}\lambda_1^\alpha(x^{\otimes N})$. As discussed in Section 3.2, acceleration occurs if $\lambda_1^0(x^{\otimes N}) \neq \mathrm{Re}\lambda_1^\alpha(x^{\otimes N})$ is satisfied.

In [18], they failed to specify the choice of J whose naive construction of J requires $\mathcal{O}(d^2N^2)$ memory cost. To reduce the memory cost, we propose the following skew matrix:

$$J := J_0 \otimes I_d, \tag{26}$$

where J_0 is a $N \times N$ skew matrix and $\otimes$ is a Kronecker product. We then have the following lemma:

Lemma 1. *If J_0 is generated based on Theorem 5 and $\mathrm{Ker} J_0 = \{0\}$ is satisfied, then given a point $x^{\otimes N}$, J does not satisfy the equality condition in Theorems 3,4, which means $\lambda_1^0(x^{\otimes N}) \neq \mathrm{Re}\lambda_1^\alpha(x^{\otimes N})$ with probability* 1.

See Appendix G.2 for the proof. Thus, from this lemma, we only need to prepare and store J_0, which requires $\mathcal{O}(N^2)$ memory, which does not depend on d. In practical settings, this is a significant reduction of the memory size since the number of parallel chains is smaller than the dimension of models. Please note that we can ensure the acceleration with this J,

Lemma 2. *Under Assumptions 1–5, assume J satisfies the condition of Lemma 1. Then S-PLD shows*

$$\chi^2(\mu_t^{\alpha,\otimes N} \| \pi^{\otimes N}) \leq e^{-\frac{2m(\alpha,N)}{\beta}t} \chi^2(\mu_0^{\otimes N} \| \pi^{\otimes N}), \tag{27}$$

where $\mu_t^{\alpha,\otimes N}$ is the measure at time t induced by S-PLD, and $\mu_0^{\otimes N}$ is the initial measure defined as the product measure of μ_0.

See Appendix I.1 for the proofs. Thus, combined with Lemma 2, S-PLD converges faster than PLD. We also considered the ensemble version of ULD (parallel ULD (PULD)) and its skew accelerated version:

$$\begin{aligned} dX_t^{\otimes N} &= \Sigma^{-1} V_t^{\otimes N} dt + \alpha_1 J_1 \nabla U^{\otimes N}(X_t^{\otimes N})dt, \\ dV_t^{\otimes N} &= -\nabla U^{\otimes N}(X_t^{\otimes N})dt - \gamma(\Sigma^{-1} + \alpha_2 J_2)V_t^{\otimes N} dt + \sqrt{2\gamma\beta^{-1}}dw_t, \end{aligned} \tag{28}$$

where J_1 and $J_2 \in \mathbb{R}^{dN \times dN}$ are real-valued skew-symmetric matrices, and α_1 and $\alpha_2 \in \mathbb{R}_+$ are positive constants and $V_t^{\otimes N} = \left(V_t^{(1)}, \ldots, V_t^{(N)}\right)^\top \in \mathbb{R}^{dN}$. We refer to this dynamics as skew PULD (S-PULD) whose faster convergence can be assured similar to Lemma 2 as shown in Appendix I.2.

4.2. Discussion of the Discretization of SDE and Stochastic Gradient and Practical Algorithm

In this section, we further consider practical settings for S-PLD and S-PULD. We discretize these continuous dynamics, e.g., by the Euler-Maruyama method, and approximate the gradient by the stochastic gradient. Although introducing skew matrices accelerates the convergence of continuous dynamics, it simultaneously increases the discretization and stochastic gradient error, resulting in a trade-off. We present a practical algorithm that controls this trade-off.

4.2.1. Trade-Off Caused by Discretization and Stochastic Gradient

We consider the following discretization and stochastic gradient for S-PLD and S-PULD:

$$X_{k+1}^{\otimes N} = X_k^{\otimes N} - h(I_{dN} + \alpha J)\nabla \hat{U}^{\otimes N}(X_k^{\otimes N}) + \sqrt{2h\beta^{-1}}\epsilon_k, \tag{29}$$

and

$$\begin{aligned} X_{k+1}^{\otimes N} &= X_k^{\otimes N} + \Sigma^{-1}V_k^{\otimes N}h + \alpha J\nabla \hat{U}^{\otimes N}(X_k^{\otimes N})h \\ V_{k+1}^{\otimes N} &= V_k^{\otimes N} - \nabla \hat{U}^{\otimes N}(X_k^{\otimes N})h - \gamma\Sigma^{-1}V_k^{\otimes N}h + \sqrt{2\gamma\beta^{-1}h}\epsilon_k, \end{aligned} \tag{30}$$

where $\epsilon_k \in \mathbb{R}^{dN}$ is a standard Gaussian random vector. $\nabla \hat{U}^{\otimes N}(X^{\otimes N})$ is an unbiased estimator of the gradient $\nabla U^{\otimes N}(X^{\otimes N})$. We refer to Equation (29 as skew-SGLD and Equation (30) as skew-SGHMC. For skew-SGHMC, we dropped J_2 of S-PULD to decrease the parameters, shown in Appendix B. Please note that skew-SGLD is the identical as the previous dynamics [18]. We introduce an assumption about the stochastic gradient:

Assumption 7. *(Stochastic gradient) There exists a constant $\delta \in [0, 1)$ such that*

$$\mathbb{E}[\|\nabla \hat{U}(x) - \nabla U(x)\|^2] \leq 2\delta\Big(M^2\|x\|^2 + B^2\Big). \tag{31}$$

Given a test function f with L_f lipschitzness, we approximate $\int f d\pi$ by skew-SGLD or skew-SGHMC, with estimator $\frac{1}{N}\sum_{n=1}^{N} f(X_k^{(n)})$. The bias of skew-SGLD is upper-bounded as

Theorem 7. *Under Assumptions 1–7, 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*

$$\left|\mathbb{E}\frac{1}{N}\sum_{n=1}^{N} f(X_k^{(n)}) - \int_{\mathbb{R}^d} f d\pi\right| \leq L_f(\underbrace{C_1(\alpha)kh}_{(i)} + \underbrace{C_2 e^{-\beta^{-1}m(\alpha,N)kh}}_{(ii)}) \tag{32}$$

and C_1 and C_2 depends on the constants of Assumptions 1–7, for the details see Appendix J.

We present a tighter bias bound in Section 4.3 under a stronger assumption. We can show a similar upper bound for the skew-SGHMC using the same proof strategy. This bound resembles of a previous one [18]; ours shows improved dependency on kh. The previous results of [18] are also limited to LD, not including skew-SGHMC.

Please note that (i) corresponds to the discretization and stochastic gradient error and (ii) corresponds to the convergence behavior of S-PLD, which is continuous dynamics. Since $C_1(\alpha) \geq C_1(\alpha = 0)$, skew acceleration increases the discretization and stochastic gradient error. On the other hand, since $m(\alpha, N) \geq m_0$, the convergence of the continuous dynamics is accelerated. Thus, skew acceleration causes a trade-off. When α is suffi-

ciently small, we derive the explicit dependency of α for this trade-off from an asymptotic expansion. Using the quantitative evaluation of skew acceleration in Theorem 6, we obtain

$$\left|\mathbb{E}\frac{1}{N}\sum_{n=1}^{N} f(X_k^{(n)}) - \int_{\mathbb{R}^d} f d\pi\right| \leq \underbrace{(d_1\alpha + d_2\alpha^2)kh}_{(i)} - \underbrace{\alpha^2 d_0 e^{-\beta^{-1}m_0 kh}}_{(ii)} + \mathcal{O}(\alpha^3) + \text{const}, \quad (33)$$

where d_0 to d_2 are positive constants obtained by the asymptotic expansion. See Appendix K for the details. In the above expression, (i) and (ii) correspond to (i) and (ii) of Equation (32). Thus, by choosing appropriate α, we can control the trade-off.

4.2.2. Practical Algorithm Controlling the Trade-Off

Since calculating the optimal α that minimizes Equation (33) at each step is computationally demanding, we adaptively tune the value of α by measuring the acceleration with kernelized Stein discrepancy (KSD) [22]. Our idea is to update samples under different α and $\alpha + \eta$, and compare KSD between the stationary and empirical distributions of these different interaction strengths. Here, $\eta \in \mathbb{R}^+$ is a small increment of α. We denote the samples at the $(k+1)$th step, which is obtained by Equation (29) as $X_{k+1,\alpha}^{\otimes N} := X_{k,\alpha}^{\otimes N} - h(I_{dN} + \alpha J)\nabla\hat{U}^{\otimes N}(X_{k,\alpha}^{\otimes N}) + \sqrt{2h\beta^{-1}}\epsilon_k$, (or (30) as $X_{k+1,\alpha}^{\otimes N} := X_k^{\otimes N} + \Sigma^{-1}V_k^{\otimes N}h + \alpha J\nabla\hat{U}^{\otimes N}(X_k^{\otimes N})h$). We denote the samples, which are obtained by replacing the above α by $\alpha + \eta$, as $X_{k+1,\alpha+\eta}^{\otimes N}$. We denote the KSD between the measure of $X_{k+1,\alpha}^{\otimes N}$ and stationary measure π as $KSD(k+1,\alpha)$ and estimate the differences of empirical KSD:

$$\Delta := K\hat{S}D(k+1,\alpha) - K\hat{S}D(k+1,\alpha+\eta), \quad (34)$$

where KSD is estimated by

$$K\hat{S}D(k,\alpha) = \frac{1}{N(N-1)}\sum_{i=1}^{N} u_q(X_{k,\alpha}^{(i)}, X_{k,\alpha}^{(j)}), \quad (35)$$

$$u_q(x,x') := \nabla_x \log\pi(x)^\top l(x,x')\nabla_x \log\pi(x') + \nabla_x \log\pi(x)^\top \nabla_{x'} l(x,x') + \nabla_x l(x,x')^\top \nabla_x \log\pi + \mathrm{Tr}\nabla_{x,x'} l(x,x'), \quad (36)$$

where l denotes a kernel and we use an RBF kernel. If $\Delta > 0$, which indicates that the empirical distribution of $X_{k+1,\alpha+\eta}^{\otimes N}$ is closer to the stationary distribution than that of $X_{k+1,\alpha}^{\otimes N}$. Thus, we should increase the interaction strength from α to $\alpha + \eta$. If $\Delta < 0$, we decrease it to $\alpha - \eta$. We also update η to $c\eta$ where $c \in (0,1]$. The overall process is shown in Algorithm 1. Detailed discussions of the algorithm including how to select α_0, η_0, and c are shown in Appendix L.

Algorithm 1 Tuning α

Input: $X_k^{\otimes N}, \eta_k, \alpha_k, c$
Output: α_{k+1}, η_{k+1}
1: Calculate $X_{k+1,\alpha_k}^{\otimes N}$ and $X_{k+1,\alpha_k+\eta_k}^{\otimes N}$.
2: Calculate $\Delta := K\hat{S}D(k+1,\alpha_k) - K\hat{S}D(k+1,\alpha_k+\eta_k)$
3: **if** $\Delta > 0$ **then**
4: Update $\alpha_{k+1} = \alpha_k + \eta_k$
5: Update $\eta_{k+1} = \eta_k$
6: **else**
7: Update $\alpha_{k+1} = |\alpha_k - \eta_k|$
8: Update $\eta_{k+1} = c\eta_k$
9: **end if**

Finally, we present Algorithm 2, which describes the whole process. We update the value of α once every k' step. Please note that its computational cost is not much larger than that of Equation (30. We only calculate the eigenvalues of J once, which requires $\mathcal{O}(N^3)$. The calculation of different KSDs is computationally inexpensive since we can re-use the gradient, which is the most computationally demanding part.

Algorithm 2 Proposed algorithm

Input: $X_0^{\otimes N}, h, \alpha_0, \eta, k', K, c, (V_0^{\otimes N}, \gamma, \Sigma^{-1})$
Output: $X_K^{\otimes N}$
1: Make a $N \times N$ random matrix J_0 and check $\ker J_0 = \{0\}$
2: Set $J = J_0 \otimes I_d$
3: **for** $k = 0$ to K **do**
4: **if** $\lfloor \frac{k}{k'} \rfloor = 0$ **then**
5: Update α by Algorithm 1
6: **end if**
7: Update $X_k^{\otimes N}$ by Equation (29) (for skew-SGLD)
8: (Update $(X_k^{\otimes N}, V_k^{\otimes N})$ by Equation (30) for skew-SGHMC)
9: **end for**

4.3. Refined Analysis for the Bias of Skew-SGLD

When using a constant step size for skew-SGLD, the bound in Theorem 7 is meaningless since the first term of Equation (32 will diverge. Here, following [23], we present a tighter bound for the bias of skew-SGLD under a stronger assumption.

Theorem 8. *Under Assumptions 1–7, for any $k \in \mathbb{N}$ and any $h \in (0, 1 \wedge \frac{\lambda(\alpha,N)}{4\sqrt{2}M^2} \wedge \frac{m}{4M^2})$ obeying $kh \geq 1$ and $\beta m \geq 2$, we have*

$$\left| \mathbb{E} \frac{1}{N} \sum_{n=1}^{N} f(X_k^{(n)}) - \int_{\mathbb{R}^d} f d\pi \right| \leq L_f \sqrt{\frac{2}{\lambda(\alpha, N)}} \sqrt{e^{-\lambda(\alpha,N)kh} \mathrm{KL}(\mu_0|\pi) + \frac{C_3(\alpha)}{\lambda(\alpha, N)}}, \tag{37}$$

where

$$\lambda(\alpha, N) := \left(\frac{1}{(1 + m(\alpha, N)^{-1} \beta C(m_0)) 2\pi e^2} + \frac{3}{2} m(\alpha, N)^{-1} \right)^{-1} \tag{38}$$

and constants $C_3(\alpha)$ and $C(m_0)$ depend on the constants of Assumptions 1–7. Moreover, $\lambda(\alpha, N)$ satisfies $\lambda(\alpha, N) \geq \lambda(\alpha = 0, N)$. For the details, see Appendix M.

Proof is shown in Appendix M. Please note that even if we use a constant step size for skew-SGLD, the bound in Theorem 8 will not diverge. Here we need the stronger assumption about a step size compared to Theorem 7. From Equation (37), the convergence behavior is characterized by $\lambda(\alpha, N)$ and the bias bound become smaller when $\lambda(\alpha, N)$ become larger. From the definition of $\lambda(\alpha, N)$, the larger $m(\alpha, N)$ is, the larger $\lambda(\alpha, N)$ we obtain. Thus, as we had seen so far, introducing the skew matrices leads to the larger Poincaré constant, and thus, this leads to larger $\lambda(\alpha, N)$.

Previous work [18] clarified that if α is sufficiently small, introducing skew matrices improves the Poincaré constant by a constant factor, which means that we have $m(\alpha, N) - m_0 \approx \mathcal{O}(\alpha^2)$, where $\mathcal{O}(\alpha^2)$ depends on the eigenvector and eigenvalues of the generator $\mathcal{L}$. On the other hand, from Theorem 8, for any $\xi > 0$, to achieve the bias smaller than ξ, it suffice to run skew-SGLD at least for $k \geq \frac{2}{\lambda(\alpha,N)h} \ln \frac{L_f}{\xi} \sqrt{\frac{\mathrm{KL}(\mu_0|\pi)}{2\lambda(\alpha,N)}}$ iterations using the appropriate step size h and under the assumption that δ and α are small enough (see Appendix M.2 for details). Combined with these observations, introducing skew matrices into SGLD improves the computational complexity for a constant order. Our numerical ex-

periments show that even constant improvement results in faster convergence in practical Bayesian models.

5. Related Work

In this section, we discuss the relationship between our method and other sampling methods.

5.1. Relation to Non-Reversible Methods

As we discussed in Section 1, our work extends the existing analysis of non-reversible dynamics [8,18] and presents a practical algorithm. Compared to those previous works, we focus on the practical setting of Bayesian sampling and derive the explicit condition about J for acceleration. We also derived a formula to quantitatively evaluate skew acceleration based on the asymptotic expansion of the eigenvalues of the perturbed Hessian matrix. A previous work [24], which derived the optimal skew matrices when the target distribution is Gaussian, requires $\mathcal{O}(d^3)$ computational cost to derive optimal skew matrices, and it is unclear whether it works for non-convex potential functions. On the other hand, our construction method for skew matrices is simple, computationally cheap, and can be applied to general Bayesian models.

Our work analyzes skew acceleration for ULD, which is more effective than LD in practical problems. Another work [8,18] only analyzed skew acceleration for LD. A previous work [17] combined a non-reversible drift term with ULD. Unlike our method, this work's purpose was to reduce the asymptotic variance of the expectation of a test function and is mainly focusing on sampling from Gaussian distribution.

To the best of our knowledge, our work is the first to focus on the memory issue of skew acceleration and develop a memory-efficient skew matrix for ensemble sampling. Our work also presents an algorithm that controls the trade-off for the first time. Another work [18] identified the trade-off and handled it by cross-validation, which is computationally inefficient, unfortunately.

Finally, we point out an interesting connection between our skew-SGHMC and the magnetic HMC (M-HMC) [25]. M-HMC accelerates HMC's mixing time by introducing a "magnetic" term into the Hamiltonian. That magnetic term is expressed by special skew matrices. Although a previous work [25] argued that M-HMC is numerically superior to a standard HMC, its theoretical property remains unclear. Thus, our work can analyze the theoretical behavior of magnetic HMC.

5.2. Relation to Ensemble Methods

Our proposed algorithm is based on ensemble sampling [26]. Ensemble sampling, in which multiple samples are simultaneously updated with interaction, has been attracting attention numerically and theoretically because of improvements in memory size, computational power, and parallel processing computation schemes [26]. There are successful, widely used ensemble methods, including SVGD [27] and SPOS [28], with which we compare our proposed method numerically in Section 6. Although both show numerically good performance, it is unclear how the interaction term theoretically accelerates the convergence since they are formulated as a McKean–Vlasov process, which is non-linear dynamics, complicating establishing a finite sample convergence rate. Our algorithm is an extension of another work [18], where the interaction was composed of a skew-acceleration term and can be rigorously analyzed. Compared to that previous work [18], we analyzed skew acceleration, focused on the Hessian matrix, and developed practical algorithms, as discussed in Section 4.2, and derived the explicit condition when acceleration occurs, which was unclear [18].

Another difference among SPOS, SVGD, and [18] is that they use first-order methods; our approach uses the second-order method. Little work has been done on ensemble sampling for second-order dynamics. Recently a second-order ensemble method was proposed [29], based on gradient flow analysis. Although its method showed good numerical

performance, its theoretical property for finite samples remains unclear since it proposed a scheme as a finite sample approximation of the gradient flow. In contrast, our proposed method is a valid sampling scheme with a non-asymptotic guarantee.

6. Numerical Experiments

The purpose of our numerical experiments is to confirm the acceleration of our algorithm proposed in Section 4 in various commonly used Bayesian models including Gaussian distribution (toy data), latent Dirichlet allocation (LDA), and Bayesian neural net regression and classification (BNN). We compared our algorithm's performance with other ensemble sampling methods: SVGD, SPOS, standard SGLD, and SGHMC. In all the experiments, the values and the error bars are the mean and the standard deviation of repeated trials. For all the experiments we set $\gamma = 1$ and $\Sigma^{-1} = 300$ for SGHMC and Skew-SGHMC. As for the hyperparameters of our proposed algorithm, the selection criterion is discussed in Appendix L.

6.1. Toy Data Experiment

The target distribution is the multivariate Gaussian distribution, $\pi = N(\mu, \Omega)$ where we generated $\Omega^{-1} = A^\top A$ and each element of $A \in \mathbb{R}^{2d \times d}$ is drawn from the standard Gaussian distribution. The dimension of the target distribution is $d = 50$, we approximate by 20 samples using the proposed ensemble methods. We tested these toy data because the LD for this target distribution is known as the Ornstein–Uhlenbeck process, and its theoretical properties have been studied extensively e.g., [30]. Thus, by studying the convergence behavior of these toy data, we can understand our proposed method more clearly.

First, we confirmed how the skew-symmetric matrix affects the eigenvalues of the Hessian matrix, as discussed in Section 3, where we only showed the asymptotic expansion for the smallest real part of the eigenvalues and saddle point. Here we can show a similar expansion for the largest real part:

$$\mathrm{Re}(\lambda^\alpha_{dN}) = \lambda^0_{dN} + \alpha^2 \sum_{k=1}^{dN-1} \frac{\lambda_d N^0 \lambda^0_k |v^{0\top}_k J v^0_{dN}|^2}{\lambda^0_k - \lambda^0_{dN}} + \mathcal{O}(\alpha^3). \tag{39}$$

$\mathrm{Re}\left(\lambda^\alpha_{dN}\right) \leq \lambda^\alpha_{dN}$ holds.

Then we observed how the largest and smallest real parts of the eigenvalues of $(I + \alpha J)\Omega^{-1}$ depend on α. The results are shown in Figure 2, where we averaged 10 trials over a randomly made J with fixed A. The upper-left, upper-right, and lower figures show $\mathrm{Re}(\lambda_1(\alpha))$, $\mathrm{Re}(\lambda_{dN}(\alpha))$, and $\mathrm{Re}(\lambda_1(\alpha))/\mathrm{Re}(\lambda_{dN}(\alpha))$. These behaviors are consistent with Theorem 3. When α is small, its behavior is close to the quadratic function proved in Theorem 3.

Next, we observed the convergence behavior of skew-SGLD and skew-SGHMC. We measured the convergence by maximum mean discrepancy (MMD) [31] between the empirical and stationary distributions. For MMD, we used 2000 samples for the target distribution, and we used the Gaussian kernel whose bandwidth is set to the median distance of these 2000 samples. We used gradient descent (GD), with step size $h = 1e-4$. The results are shown in Figure 3. The proposed method shows faster convergence than naive parallel sampling, which is consistent with Table 2.

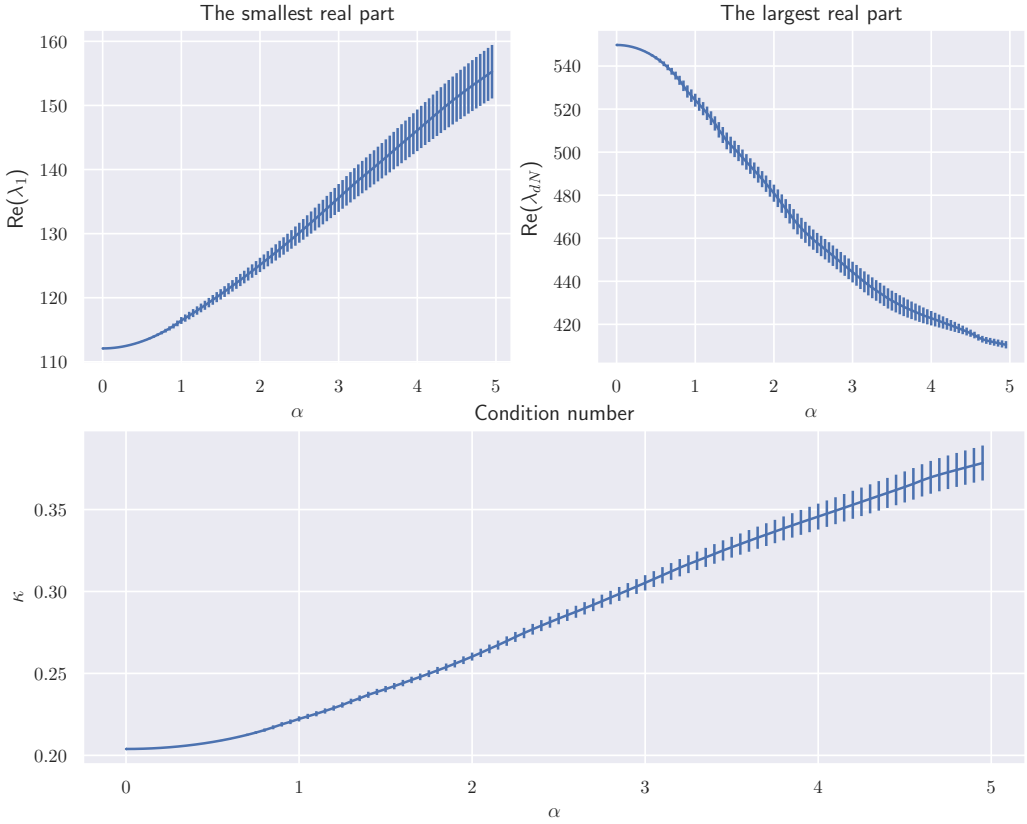

Figure 2. Eigenvalue changes (averaged over ten trials).

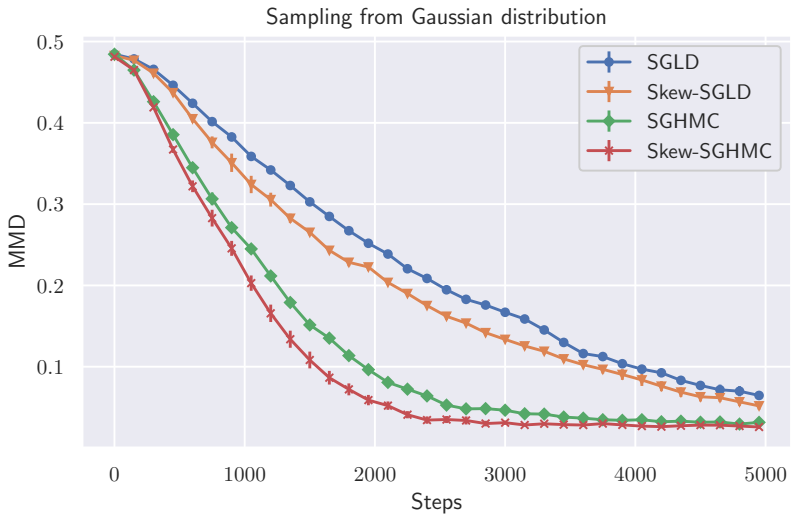

Figure 3. Convergence behavior of toy data in MMD (averaged over ten trials)

6.2. LDA Experiment

We tested with an LDA model using the ICML dataset [32] following the same setting as [33]. We used 20 samples for all the methods. Minibatch size is 100. We used step size $h = 5e-4$. First, we confirmed the effectiveness of our proposed Algorithm 1, which adaptively tunes α values. For that purpose, we compared the final performance obtained by our methods with a previous method [18], in which α is selected by cross-validation (CV). Here instead of CV, we just fixed α during the sampling and refer to it as fixed α. We also tested the case when J is generated randomly at each step with fixed α, as discussed in Section 4.1. We refer to it as random J. The results are shown in Figure 4 where skew-SGLD was used. We found that our method showed competitive performance with

the best performance of fixed α. For the computational cost, we used $k' = 2$ in Algorithm 2, and our method needed twice the wall clock time than each fixed α. This means that our algorithm greatly reduces the total computational time since we tried more than two αs in the fixed α for CV. We also found that since using different Js at each step did not accelerate the performance, we need to store and fix J during the sampling for acceleration. Next, we compared our method with other ensemble sampling schemes and observed the convergence speed. The result is shown in Figure 5. Skew-SGLD and skew-SGHMC outperformed SGLD and SGHMC, which is consistent with our theory.

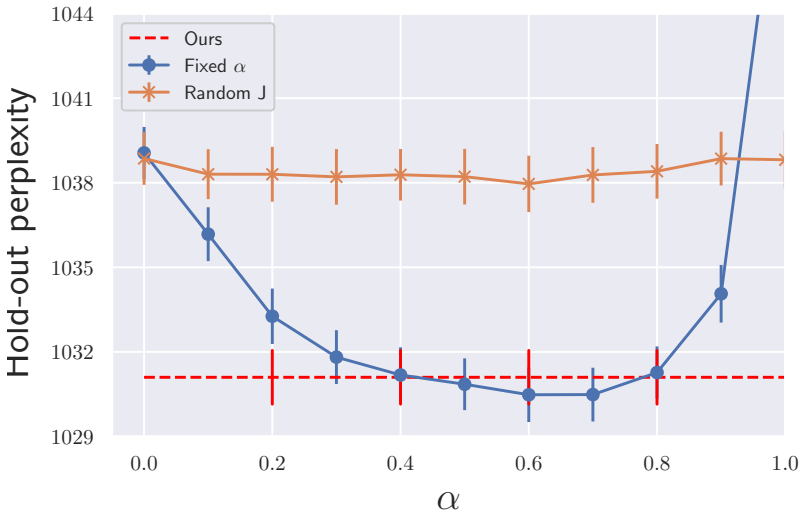

Figure 4. Final performances of LDA under different values of α (averaged over ten trials).

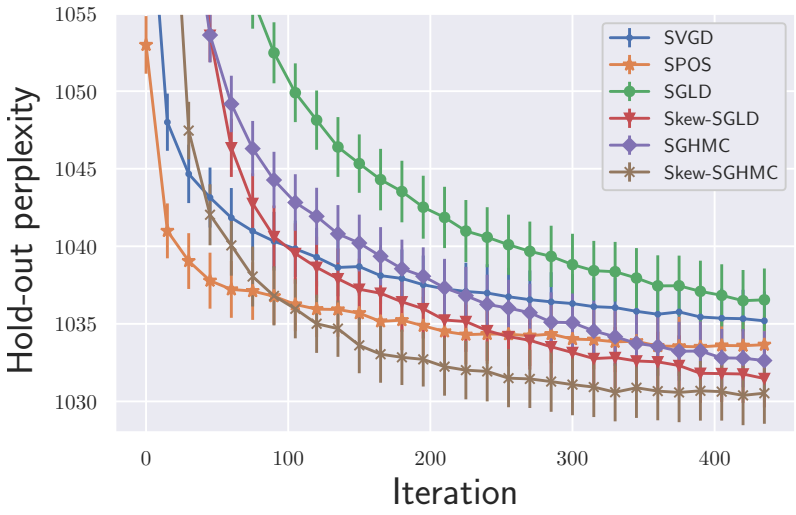

Figure 5. LDA experiments (Averaged over 10 trials).

6.3. BNN Regression and Classification

We tested with the BNN regression task using the UCI dataset [34], following a previous setting Liu and Wang [27]. We used one hidden layer neural network model with ReLU activation and 100 hidden units. We used 10 samples for all the methods. We used the minibatch size 100. We used step size $h = 5e - 5$. The results are shown in Tables 1 and 2. We also tested on BNN classification task using the MNIST dataset. The result is shown in Figure 6. We used one hidden layer neural network model with ReLU activation and 100 hidden units. Batchsize is 500 and we set step size $h = 5e - 5$. Our proposed methods outperformed other ensemble methods. Please note that skew-SGHMC and skew-SGLD consistently outperformed SGHMC and SGLD.

Table 1. Benchmark results on test RMSE for regression task

Dataset	Avg. Test RMSE					
	SVGD	**SPOS**	**SGLD**	**Skew-SGLD**	**SGHMC**	**Skew-SGHMC**
Concrete	5.709 ± 0.040	5.239 ± 0.199	5.009 ± 0.091	4.973 ± 0.057	4.949 ± 0.144	4.790 ± 0.081
Kin8nm	0.0731 ± 0.0006	0.0688 ± 0.0003	0.0693 ± 0.0006	0.0689 ± 0.0005	0.0687 ± 0.0001	0.0683 ± 0.0003
Energy	0.520 ± 0.060	0.456 ± 0.030	0.428 ± 0.045	0.412 ± 0.045	0.406 ± 0.019	0.403 ± 0.008
Bostonhousing	3.306 ± 0.005	3.107 ± 0.173	2.948 ± 0.084	2.930 ± 0.095	3.053 ± 0.093	2.986 ± 0.143
Winequality	0.619 ± 0.001	0.618 ± 0.007	0.641 ± 0.003	0.634 ± 0.004	0.614 ± 0.004	0.613 ± 0.004
PowerPlant	4.219 ± 0.012	4.160 ± 0.009	4.129 ± 0.002	4.118 ± 0.006	4.112 ± 0.009	4.105 ± 0.008
Yacht	0.475 ± 0.049	0.467 ± 0.110	0.464 ± 0.058	0.442 ± 0.046	0.464 ± 0.078	0.432 ± 0.051

Table 2. Benchmark results on test negative log likelihood for regression task

Dataset	Avg. Test Negative Log Likelihood					
	SVGD	**SPOS**	**SGLD**	**Skew-SGLD**	**SGHMC**	**Skew-SGHMC**
Concrete	-3.157 ± 0.008	-3.124 ± 0.025	-3.052 ± 0.009	-3.049 ± 0.012	-3.046 ± 0.025	-3.033 ± 0.021
Kin8nm	1.153 ± 0.0084	1.212 ± 0.008	1.223 ± 0.002	1.223 ± 0.005	1.230 ± 0.0015	1.235 ± 0.0025
Energy	-0.816 ± 0.102	-0.976 ± 0.079	-0.867 ± 0.056	-0.845 ± 0.021	-0.843 ± 0.045	-0.844 ± 0.041
Bostonhousing	-2.98 ± 0.000	-2.644 ± 0.027	-2.548 ± 0.016	-2.539 ± 0.002	-2.574 ± 0.019	-2.561 ± 0.017
Winequality	-1.012 ± 0.000	-0.959 ± 0.007	-0.976 ± 0.006	-0.968 ± 0.005	-0.941 ± 0.007	-0.938 ± 0.005
PowerPlant	-2.871 ± 0.004	-2.850 ± 0.004	-2.844 ± 0.002	-2.842 ± 0.001	-2.838 ± 0.004	-2.835 ± 0.003
Yacht	-1.184 ± 0.06	-1.372 ± 0.07	-1.077 ± 0.066	-1.078 ± 0.030	-1.083 ± 0.030	-1.079 ± 0.051

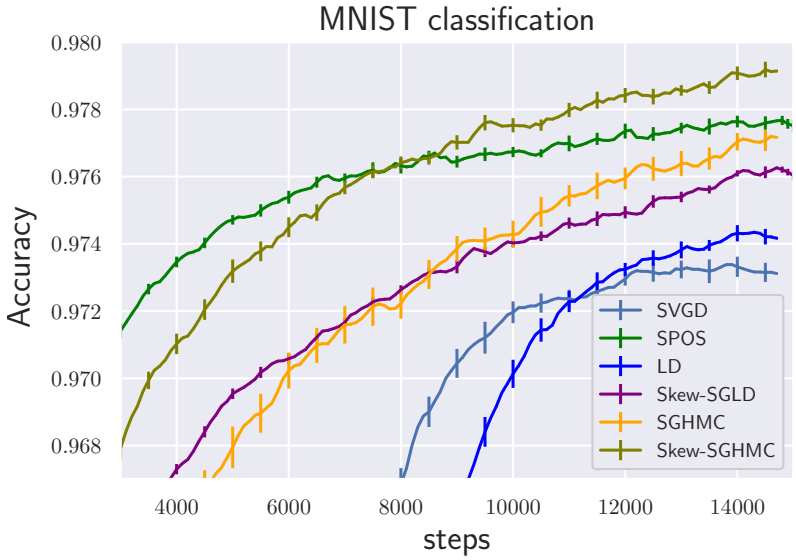

Figure 6. MNIST classification (Averaged over ten trials)

7. Conclusions

We studied skew acceleration for LD and ULD from practical viewpoints and concluded that the improved eigenvalues of the perturbed Hessian matrix caused acceleration and derived the explicit condition for acceleration. We described a novel ensemble sampling method, which couples multiple SGLD or SGHMC with memory-efficient skew matrices. We also proposed a practical algorithm that controls the trade-off of faster convergence and larger discretization and stochastic gradient error and numerically confirmed the effectiveness of our proposed algorithm.

Author Contributions: Conceptualization, F.F. and T.I.; methodology, F.F. and T.I.; software, F.F.; validation, F.F., T.I., N.U. and I.S.; formal analysis, F.F. and I.S.; writing—original draft preparation, F.F.; project administration, F.F.; funding acquisition, F.F. All authors have read and agreed to the published version of the manuscript.

Funding: JST ACT-X: Grant Number JPMJAX190R.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: http://archive.ics.uci.edu/ml

Acknowledgments: FF was supported by JST ACT-X Grant Number JPMJAX190R.

Conflicts of Interest: The authors declare no conflict of interest.The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

LD	Langevin Dynamics
MCMC	Markov Chain Monte Carlo
ULD	Underdamped Langevin Dynamics
SGLD	Stochastic Gradient Langevin Dynamics
SGHMC	Stochastic Gradient Hamilton Monte Carlo
PLD	Parallel Langevin Dynamics
PULD	Parallel Underdamped Langevin Dynamics
SLD	Skew Langevin Dynamics
S-ULD	Skew Underdamped Langevin Dynamics
S-PLD	Skew Parallel Langevin Dynamics
S-PULD	Skew Parallel Underdamped Langevin Dynamics
KSD	Kernelized Stein Discrepancy

Appendix A. Additional Backgrounds

We introduce additional backgrounds which are used in our Proof.

Appendix A.1. Wasserstein Distance and Kullback–Leibler Divergence

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

$$W_p^d(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \left(\int \int d(x,y)^p d\pi(x,y) \right)^{1/p}, \tag{A1}$$

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. The various properties of Wasserstein distance are summarized in [35]. We define the Kullback–Leibler (KL) divergence as

$$\mathrm{KL}(\nu \| \mu) = \begin{cases} \int \log \frac{d\nu}{d\mu} d\nu, & \nu \ll \mu, \\ +\infty, & \text{otherwise.} \end{cases} \tag{A2}$$

Appendix A.2. Markov Diffusion and Generator

Here we introduce the additional explanation about the generator of the Markov diffusion process. Given an SDE,

$$dX_t = -\nabla U(X_t)dt + \sqrt{2\beta^{-1}}dw(t), \tag{A3}$$

and we denote the corresponding Markov semigroup as $P = \{P_t\}_{t>0}$ and define the Kolmogorov operator as P_s which is defined as $P_s f(X_t) = \mathbb{E}[f(X_{t+s})|X(t)]$, 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, $\int_{\mathbb{R}^d} P_t f d\pi = \int_{\mathbb{R}^d} f d\pi$.

We denote the infinitesimal generator of the associated Markov group as $\mathcal{L}$ and we call it a 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

$$\frac{1}{h}(P_{t+h} - P_t) = P_t \frac{1}{h}(P_h - Id) = \frac{1}{h}(P_h - Id)P_t, \tag{A4}$$

where Id is the identity map. In addition, taking $h \to 0$, we have $\partial P_t = \mathcal{L}P_t = P_t\mathcal{L}$. From the Hille–Yoshida theory [19], 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 Equation (A3), the generator can be written as

$$\mathcal{L}f(X_t) := \lim_{h\to 0^+} \frac{\mathbb{E}(f(X_{t+h})|X_t) - f(X_t)}{h} = \Big(-\nabla U(X_t)\cdot\nabla + \beta^{-1}\Delta\Big)f(X_t), \tag{A5}$$

where Δ is the Laplacian in the standard Euclidean space. The generator satisfies $\mathcal{L}1 = 0, \quad \int_{\mathbb{R}^d} \mathcal{L}f d\pi = 0$.

Appendix A.3. Poincaré Inequality

We use the Poincaré inequality to measure the speed of convergence to the stationary distribution. In this section, we summarize definitions and useful properties of them and see [19] for more details. We define the Dirichlet form $\mathcal{E}(f)$ for all bounded functions $f \in \mathcal{D}(\mathcal{L})$ where $\mathcal{D}(\mathcal{L})$ denotes the domain of $\mathcal{L}$ as

$$\mathcal{E}(f) := -\int_{\mathbb{R}^d} f\mathcal{L}f d\pi. \tag{A6}$$

$\mathcal{E}(f) > 0$ is satisfied. By the partial integration, we have $\mathcal{E}(f) = -\int_{\mathbb{R}^d} f\mathcal{L}f d\pi = \frac{1}{\beta}\int_{\mathbb{R}^d} \|\nabla f\|^2 d\pi$. 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 Poincaré inequality* with a positive constant c if for any $f \in \mathcal{D}(\mathcal{E})$, π with $\mathcal{L}$ satisfies,

$$\int f^2 d\pi - \left(\int f d\pi\right)^2 \le c\mathcal{E}(f). \tag{A7}$$

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

$$\lambda := \inf_{f\in\mathcal{D}(\mathcal{E})}\left\{\frac{\mathcal{E}(f)}{\int f^2 d\pi} : f \neq 0, \int f d\pi = 0\right\}. \tag{A8}$$

From this, a constant c which satisfies $c \geq 1/\lambda$, can also satisfy the Poincaré inequality. To check the existence of the spectral gap, one approach is to use the Lyapunov function, which is developed by Bakry et al. [36].

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

$$\chi^2(\mu\|\pi) := \left\| \frac{d\mu}{d\pi} - 1 \right\|^2_{L^2_\pi} = \int_{\mathbb{R}^d} \left| \frac{d\mu}{d\pi} - 1 \right|^2 d\pi. \tag{A9}$$

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

$$\chi^2(\mu\|\pi) \leq c\, \mathcal{E}\left(\sqrt{\frac{d\mu}{d\pi}} \right). \tag{A10}$$

We obtain the following exponential convergence results from the above functional inequalities for measures.

Theorem A1. *(Exponential convergence in the variance, Theorem 4.2.5 in [19]) When π satisfies the Poincaré inequality with a constant c, it implies the exponential convergence in the variance with a rate $2/c$, i.e., for every bounded function $f : \mathbb{R}^d \to \mathbb{R}$,*

$$\mathrm{Var}_\pi(P_t f) \leq e^{-2t/c} \mathrm{Var}_\pi(f), \tag{A11}$$

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

We also introduce the important property of Poincaré inequality as for the product measures. These relations play important roles in our analysis.

Theorem A2. *(Stability under the product, Proposition 4.3.1 in [19]) If μ_1 and μ_2 on $\mathbb{R}^d$ satisfy the Poincaré 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 Poincaré inequality with the constant* $\max(c_1, c_2)$.

Appendix B. Generator of the Underdamped Langevin Dynamics (ULD)

Following [10], we define the infinitesimal generator of the ULD as

$$\mathcal{L}f(x,v) := -(\gamma v + \nabla U(x))\nabla_v f(x,v) + \gamma\beta^{-1}\Delta f(x,v) + v\nabla_x f(x,v). \tag{A12}$$

Then, we define the generator of S-ULD as

$$\begin{aligned} \mathcal{L}f(x,v) := &-(\gamma v + \nabla U^(x))\nabla_v f(x,v) + \gamma\beta^{-1}\Delta f(x,v) \\ &+ v\nabla_x f(x,v) + \alpha_1 J_1 \nabla U(x)\nabla_x f(x,v) + \alpha_1 J_2 \Sigma^{-1} v \nabla_v f(x,v), \end{aligned} \tag{A13}$$

where the second line corresponds to the interaction terms. Then it is easily to confirm $\int_{\mathbb{R}^{2d}} \mathcal{L}f(x,v)d\tilde{\pi} = 0$, where $\tilde{\pi} := \pi \otimes \mathcal{N}(0,\Sigma) \propto e^{-\beta U(x) - \frac{1}{2}\Sigma^{-1}\|v\|^2}$. Thus, the stationary distribution of S-ULD is $\tilde{\pi}$. We can prove this by simply using the partial integral and using the property of the skew-symmetric matrix. Thus, the stationary distribution of S-ULD is $\tilde{\pi}$.

We consider other combinations the skew matrices with ULD. For example, we can consider the following more general combination;

$$\begin{aligned} dX_t &= \Sigma^{-1} V_t dt + \alpha_1 J_1 \nabla U(X_t) dt + \alpha_2 \Sigma^{-1} J_2 V_t dt \\ dV_t &= -\nabla U(X_t) dt - \gamma \Sigma^{-1} V_t dt + \alpha_3 J_3 V_t dt + \alpha_4 J_4 \nabla U(X_t) dt + \sqrt{2\gamma\beta^{-1}} dw_t, \end{aligned} \tag{A14}$$

compared to S-ULD, there are new two terms are included. We can also derive the infinitesimal generator of this Markov process. We express it as $\tilde{\mathcal{L}}$. Then we calculate the infinitesimal change of the expectation of f

$$\int_{\mathbb{R}^{2d}} \tilde{\mathcal{L}} f(x,v) d\tilde{\pi} \neq 0, \tag{A15}$$

which suggests that the stationary distribution of Equation (A14) is different form $\tilde{\pi}$.

It is widely known that underdamped Langevin dynamics converges to (overdamped) Langevin dynamics. Here we observe that S-ULD converges to Skew-LD in [18]. The limiting procedure is widely known, for example, see [17,37,38]. We cite Proposition 1 in [17]; given a stochastic process

$$\begin{aligned} dX_t &= \Sigma^{-1} V_t dt + \alpha_1 J_1 \nabla U(X_t) dt, \\ dV_t &= -\nabla U(X_t) dt - \gamma \Sigma^{-1} V_t dt - \alpha_2 \Sigma^{-1} J_2 V_t dt + \sqrt{2\gamma} dw_t, \end{aligned} \tag{A16}$$

and we rescale it by introducing ϵ which expresses the small mass limit as

$$\begin{aligned} dX_t &= \frac{1}{\epsilon} \Sigma^{-1} V_t dt + \alpha_1 J_1 \nabla U(X_t) dt, \\ dV_t &= -\frac{1}{\epsilon} \nabla U(X_t) dt - \frac{1}{\epsilon^2} \gamma \Sigma^{-1} V_t dt - \frac{1}{\epsilon}^2 \alpha_2 \Sigma^{-1} J_2 V_t dt + \frac{1}{\epsilon} \sqrt{2\gamma} dw_t, \end{aligned} \tag{A17}$$

and by taking the limit $\epsilon \to 0$, the dynamics converges to

$$dX_t = -(\alpha_2 J_2 + \gamma)^{-1} \nabla U(X_t) dt - \alpha_1 J_1 \nabla U(X_t) + (\alpha_2 J_2 + \gamma)^{-1} \sqrt{2\gamma} dw_t. \tag{A18}$$

See Proposition 1 in [17], for the precise statements. Please note that the term related J_2 works as preconditioning. Thus, if we set $\alpha_2 J_2 = 0$, the obtained dynamics are equivalent to the continuous dynamics of skew-SGLD. Thus, our skew-SGHMC is the natural extension of skew-SGLD.

Appendix C. Proof of Theorem 1

Appendix C.1. Proof for S-LD

First, under Asuumptions 1–5, LD has a spectral gap, and its Poincaré constant is upper bounded as

$$\frac{1}{m_0} \leq \frac{2C(d+b\beta)}{m\beta} \exp\left(\frac{2}{m}(M+B)(b\beta+d) + \beta(A+B) \right) + \frac{1}{m\beta(d+b\beta)}. \tag{A19}$$

and this is derived in [2].

Next, we introduce the generator of S-LD

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

where $\nabla U_\alpha(x) := \nabla U(x) + \alpha J \nabla U(x)$.

The proof is almost similar to [18] of Theorem 12.

Proof of Theorem 1. Since the generator $\mathcal{L}_{\alpha=0}$ is self-adjoint, and the suitable growth condition, the spectral of $\mathcal{L}_{\alpha=0}$ is discrete [19]. 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 $0 > \lambda_0 > \lambda_1 > \dots$. Thus, $m_0 = -\lambda_0$.

As for $\mathcal{L}_\alpha$, although it is not a self-adjoint operator, from Proposition 1 in Franke et al. [39], 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

$$\mathcal{L}_\alpha(u + iv) = (\lambda + i\mu)(u + iv). \tag{A20}$$

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

$$\mathcal{L}_\alpha u = \lambda u - \mu v, \tag{A21}$$
$$\mathcal{L}_\alpha v = \lambda v + \mu u. \tag{A22}$$

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

$$\int g(\mathcal{L}_{\alpha=0} - \mathcal{L}_\alpha) g d\pi = \int \alpha g \gamma \cdot \nabla g d\pi = -\int \alpha g \gamma \cdot \nabla g d\pi, \tag{A23}$$

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

$$\int g \mathcal{L}_{\alpha=0} g d\pi = \int g \mathcal{L}_\alpha g d\pi. \tag{A24}$$

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

$$\int u \mathcal{L}_{\alpha=0} u d\pi + \int v \mathcal{L}_{\alpha=0} v d\pi = \int u \mathcal{L}_\alpha u d\pi + \int v \mathcal{L}_\alpha v d\pi = \lambda \left(\int u^2 d\pi + \int v^2 d\pi \right) = \lambda. \tag{A25}$$

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

$$\begin{aligned} \lambda = \int u \mathcal{L}_{\alpha=0} u d\pi + \int v \mathcal{L}_{\alpha=0} v d\pi &= \sum_k \lambda_k \left(\left(\int u e_k d\pi \right)^2 + \left(\int v e_k d\pi \right)^2 \right) \\ &\leq \lambda_0 \sum_k \left(\left(\int u e_k d\pi \right)^2 + \left(\int v e_k d\pi \right)^2 \right) \leq \lambda_0. \end{aligned} \tag{A26}$$

Thus, the real part of the eigenvalue of $\mathcal{L}_\alpha$ is smaller than the smallest eigenvalue of $\mathcal{L}_\alpha$. This means that the spectral gap of $\mathcal{L}_\alpha$ is larger than that of $\mathcal{L}_{\alpha=0}$, i.e., $m(\alpha) \geq m_0$ holds. □

Appendix C.2. Proof of Theorem 2 (S-ULD)

Proof of Theorem 2 . To prove the S-ULD, we use the result of [20], which characterize the convergence of ULD via the Poincaré constant. Let us denote $\tilde{\mu}_t$ as the measure induced by ULD. Then from Theorem 1 of [20], if π with $\mathcal{L}$ has the Poincaré constant m_0, we have

$$\chi^2(\tilde{\mu}_t \| \tilde{\pi}) \leq \frac{1 + \bar{\epsilon}}{1 - \bar{\epsilon}} e^{-\lambda_\gamma t} \chi^2(\tilde{\mu}_t \| \tilde{\pi}). \tag{A27}$$

where $\bar{\epsilon}$ and λ_γ is given as follows.

$$\lambda_\gamma = \frac{\Lambda(\gamma, \bar{\epsilon} \min(\gamma, \gamma^{-1}))}{1 + \bar{\epsilon} \min(\gamma, \gamma^{-1})}, \tag{A28}$$

where

$$\Lambda(\gamma,\epsilon) = \frac{\gamma\Sigma^{-1} - \frac{1}{1+\frac{m_0\Sigma^{-1}}{\beta}}}{2} - \frac{1}{2}\sqrt{(S_{--}-S_{++})^2+(S_{-+})^2}, \tag{A29}$$

$$S_{--} = \epsilon\lambda_{ham}, \tag{A30}$$

$$S_{-+} = -\epsilon(R_{ham} + \gamma\Sigma^{-1}/2), \tag{A31}$$

$$S_{++} = \gamma\Sigma^{-1} - \epsilon, \tag{A32}$$

$$\lambda_{ham} = 1 - \left(1 + \frac{m_0\Sigma^{-1}}{\beta}\right)^{-1}, \tag{A33}$$

$$\epsilon = \bar{\epsilon}\min(\gamma,\gamma^{-1}), \tag{A34}$$

where $\bar{\epsilon}$ is arbitrary sufficiently small positive value such that $\Lambda(\gamma,\bar{\epsilon}\min(\gamma,\gamma^{-1})) > 0$ is satisfies. As for R_{ham}, if there exists a positive constant K, such that $\nabla^2 U \geq -KI$, then $R_{ham} \leq \sqrt{\max\{K,2\}}$. In our assumption, this corresponds to βM, thus $R_{ham} \leq \sqrt{\max\{\beta M,2\}}$. From the above definitions, we can see that the larger m_0 is, i.e., the larger the Poincaré constant is the faster convergence ULD shows.

This can also be confirmed numerically, see Figure A1, which shows how the Λ changes under different m_0. We set $\Sigma^{-1} = 100$. From the figure, the larger the Poincaré constant is, the larger Λ becomes.

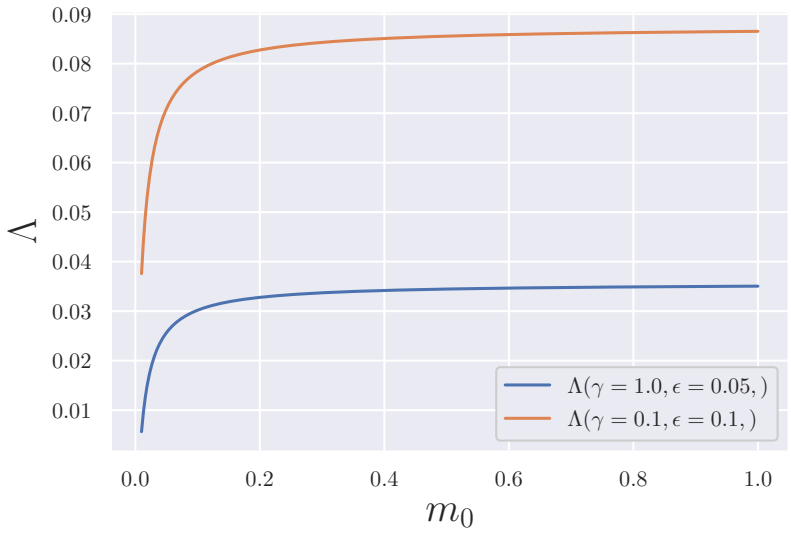

Figure A1. The convergence rate of ULD under the different Poincaré constants.

So far, we confirmed that the convergence speed of S-ULD is characterized by the Poincaré constant of $\mathcal{L}$. When we consider S-ULD, we simply add the skew matrices term to the generator of the ULD in the proof of Proposition 1 in [20]. This means that we simply replace the Poincaré constant from m_0 to $m(\alpha)$ in the proof of Proposition 1 in [20]. Then, m_0 will be replaced with $m(\alpha)$ that indicates the faster convergence. □

Appendix D. Eigenvalue and Poincaré Constant

In this section, we discuss the relation between eigenvalues of the Hessian matrix and Poincaré constant.

Appendix D.1. Strongly Convex Potential Function

When we consider LD with m-strongly convex potential function, then the Poincaré constant is m, this means exponential convergence with rate m (See [19] for the detail).

We then consider the S-LD with m-strongly convex function. In this setting, by considering the synchronous coupling technique [11], we can show that the variance decays exponentially with the rate of the smallest real part of the eigenvalue. This is because that by preparing two S-LD (X_t, Y_t) given as

$$dX_t = -(I + \alpha J)\nabla U(X_t)dt + \sqrt{2\beta^{-1}}dw_t, \qquad dY_t = -(I + \alpha J)\nabla U(Y_t)dt + \sqrt{2\beta^{-1}}dw'_t. \tag{A35}$$

Then we evaluate the behavior of $\|X_t - Y_t\|^2$. From Ito lemma and considering the synchronous coupling, we obtain

$$\frac{d}{dt}\|X_t - Y_t\|^2 = -(X_t - Y_t) \cdot \frac{(I + \alpha J)}{\beta}(\nabla U(X_t) - \nabla U(Y_t)) \leq -\frac{2m(\alpha)}{\beta}\|X_t - Y_t\|^2, \tag{A36}$$

where $m(\alpha)$ is the constant that satisfies $m(\alpha) \leq \mathrm{Re}\lambda_1^\alpha(x)$ for all x, see Appendix E for details. This means that variance decays exponentially with the rate $\frac{2m(\alpha)}{\beta}$. From the fundamental property of the Poincaré constant (Theorem 4.2.5 in [19]), $m(\alpha)$ is the Poincaré constant. Thus the imaginary part has no effect on the continuous dynamics. Thus, the Poincaré inequality is the smallest real part of the perturbed Hessian matrix.

Appendix D.2. Non-Convex Potential Function

As we discussed in Section 3.1, [21] derived the sharper estimation for the Poincaré constant for the non-convex potential function. It is easy to verify that their assumptions are satisfied under our assumption 1–5. Following the main paper, we denote x_1 global minima, and x_2 is the local minima which have the second smallest value in $U(x)$. We express the saddle point between x_1 and x_2 as x^*. To be more precise, the saddle point that characterizes the Poincaré constant is known as the critical point with index one defined as

$$U(x^*) = \inf\left\{\max_{s\in[0,1]} U(\gamma(s)) : \gamma \in C([0,1], \mathbb{R}^d), \gamma(0) = x_1, \gamma(1) = x_2\right\}, \tag{A37}$$

and the eigenvalue of $\nabla^2 U(x^*)$ has one negative eigenvalue and $d-1$ positive eigenvalues. We express them as $\lambda_1(x^*) < 0 < \lambda_2(x^*) < \ldots, \lambda_d(x^*)$.

[21] studied the Poincaré constant by decomposing the non-convex potential focusing on attractors. By focusing on attractors, they showed that the non-convex potential can be decomposed into the sum of *approximately* Gaussian distributions. They proved that the Poincaré constant is characterized by the local Poincaré constants, these are derived by the approximate Gaussian distribution on the attractors and their surrounding regions. In addition, they proved that the dominant term of the Poincaré constant is specified by the saddle points between the global minima and the point which takes the second smallest value for $U(x)$. From Theorem 2.12 and Corollary 2.15 in [21], the Poincaré constant is characterized by

$$m_0^{-1} \approx \frac{\sqrt{\det H(x^*)}}{\sqrt{Z|\lambda_1(x^*)|\det H(x_1)}\sqrt{\det H(x_2)}} e^{\beta(U(x^*)-U(x_1)-U(x_2))} \propto \frac{1}{|\lambda_1(x^*)|} e^{\beta(U(x^*)-U(x_1)-U(x_2))}, \tag{38}$$

where Z is the normalizing constant of $e^{-\beta U(x)}$.

Next, we discuss how this estimate changes when skew matrices are applied. When the skew matrices are introduced, from lemma A.1 in [40], at the saddle point, there exists a unique negative real eigenvalue $\lambda_1^\alpha(x^*) < 0$ for the perturbed Hessian matrix even if $(I + \alpha J)H$ is not a symmetric matrix.

Then from Proposition 5 in [8], that negative eigenvalue of the perturbed Hessian is smaller than that of the un-perturbed Hessian matrix at the saddle point. This means that $\lambda_1^\alpha(x^*) \leq \lambda_1(x^*) < 0$ holds.

Finally, from Theorem 5.1 in [41] and Theorem 2.12 in [21], this improvement of the negative eigenvalue of the saddle point directly leads to the larger Poincaré constant.

Appendix E. Properties of a Skew-Symmetric Matrix

Here, we introduce the basic properties of the skew-symmetric matrices. Let us consider assume that $d \times d$ matrix $H' = (I + \alpha J)H$ is diagonalizable. Then assume that matrix H' has l real eigenvalues $\lambda_1, \ldots, \lambda_l$ and $2m$ complex eigenvalues, $\mu_1 = \alpha_1 \pm i\beta_1, \ldots, \mu_m = \alpha_m \pm i\beta_m$. Thus, $d = l + 2m$. We denote the corresponding eigenvectors as $\{v_j\}_{j=1}^{l}$ for real eigenvalues and $\{w_j = a_j + ib_j\}_{j=1}^{m}$ for complex eigenvalues $\{\mu_j\}_{j=1}^{m}$ and $\{\bar{w}_j\}$ for corresponding conjugate eigenvalues. Then, let us define a $d \times d$ matrix V as

$$V = [v_1, \ldots, v_l, a_1, b_1, \ldots, a_m, b_m]. \tag{A39}$$

Then, we can decompose H' into a block diagonal matrix [42];

$$H'V = VD \tag{A40}$$

$$D = \underbrace{\begin{pmatrix} \lambda_1 & & & & & & & \\ & \ddots & & & & & & \\ & & \lambda_l & & & & & \\ & & & \alpha_1 & 0 & & & \\ & & & 0 & \alpha_1 & & & \\ & & & & & \ddots & & \\ & & & & & & \alpha_m & 0 \\ & & & & & & 0 & \alpha_m \end{pmatrix}}_{:=A} + \underbrace{\begin{pmatrix} 0 & & & & & & & \\ & \ddots & & & & & & \\ & & 0 & & & & & \\ & & & 0 & \beta_1 & & & \\ & & & -\beta_1 & 0 & & & \\ & & & & & \ddots & & \\ & & & & & & 0 & \beta_m \\ & & & & & & -\beta_m & 0 \end{pmatrix}}_{:=B}. \tag{41}$$

Thus, $D := A + B$. Then, from the Taylor expansion and expressing its residual by integral, by defining $H(x) := \nabla^2 U(x)$ we have

$$(x-y)^\top (I + \alpha J)(\nabla U(x) - \nabla U(y)) = (x-y)^\top \left(\int_0^1 (I + \alpha J) H(y + \tau(x-y))(x-y) d\tau \right). \tag{42}$$

Then, let us apply the Jordan canonical form here. If $(I + \alpha J)H$ is diagonalizable, and it is decomposable by the Jordan canonical form shown in Equation (A40. Then, we can decompose $(I + \alpha J)H$ as

$$(I + \alpha J)H(x^* + \tau(x(t) - x^*)) = VDV^{-1}. \tag{43}$$

Then, we obtain

$$\begin{aligned} (x-y)^\top (I + \alpha J)(\nabla U(x) - \nabla U(y)) &= (x-y)^\top \left(\int_0^1 (I + \alpha J) H(y + \tau(x-y))(x-y) d\tau \right) \\ &= \left(\int_0^1 (x-y)^\top V(A+B)V^{-1}(x-y) dt \right) \\ &= \left(\int_0^1 (x-y)^\top VAV^{-1}(x(t) - x^*) dt \right) \\ &\leq m(\alpha) \|x(t) - x^*\|^2. \end{aligned} \tag{44}$$

where $m(\alpha)$ is the constant that satisfies $m(\alpha) \leq \min\{\lambda_1, \ldots, \lambda_l, \alpha_1, \ldots, \alpha_m\}$ for all x. Thus, the imaginary part never appears to the upper bound and we only need to focus on the largest real part of the eigenvalues, if the matrix is diagonalizable. Next subsection describes when the non-symmetric matrix H' is diagonalizable by focusing on the random matrix.

Appendix F. Proof of Theorem 3

Proof. Since the potential function is m-strongly convex, the smallest eigenvalue of the Hessian matrix H is m, which is larger than 0. Thus, H and $H^{1/2}$ are regular matrices. With

this in mind, we consider $H + H^{1/2}JH^{1/2}$ as a similar matrix of $H' := (I + J)H$. This is easily confirmed by

$$H^{-1/2}(H + H^{1/2}JH^{1/2})H^{1/2} = H'. \tag{A45}$$

This means that to study the eigenvalues of H', we only need to study the similar matrix $A := H + H^{1/2}JH^{1/2}$. By doing this, A is composed of symmetric and skew-symmetric matrices, which are easy to treat compared to H', where the term JH is difficult to analyze. For simplicity, we omit the dependency of H and H' on x in this section.

Remark A1. *Please note that we can eliminate the strong convexity of U, if H is a regular matrix. This means that H does not have 0 as an eigenvalue.*

For simplicity, we assume that the dimension d is an even number. We assume that the eigenvalues and eigenvectors of A are expressed as

$$Aw_j = \mu_j w_j \Leftrightarrow A(a_j + ib_j) = (\alpha_j + i\beta_j)(a_j + ib_j). \tag{A46}$$

and α_j is ordered as $\alpha_1 \leq \alpha_2, \ldots$. In this section, we only consider the setting where all the eigenvalue and eigenvector are imaginary for notational simplicity. The extension to the general settings similar to Appendix E and the setting when is d is odd is straightforward.

We denote the eigenvalues and eigenvectors of H as $\{\lambda_j, v_j\}_{j=1}^d$ and v_js are linearly independent. In addition, we assume that $\lambda_1 \leq, \ldots, \lambda_d$. From this definition, by checking the real parts and complex parts, the following relations are derived

$$Aa_j = \alpha_j a_j - \beta b_j, \tag{A47}$$
$$Ab_j = \alpha_j b_j + \beta a_j. \tag{A48}$$

thus, by the skew-symmetric property

$$a_j^\top Aa_j + b_j^\top Ab_j = \alpha_j(\|a_j\|^2 + \|b_j\|^2) = \alpha_j \tag{A49}$$
$$= a_j^\top Ha_j + b_j^\top Hb_j, \tag{A50}$$

and in the third equality, we used the property

$$a_j^\top H^{1/2}JH^{1/2}a_j = b_j^\top H^{1/2}JH^{1/2}b_j = 0, \tag{A51}$$

since $H^{1/2}JH^{1/2}$ is a skew-symmetric matrix. Then, we expand a_j and b_j by v_j as

$$a_k = \sum_{j=1}^d a_k^\top v_j \tag{A52}$$

$$b_k = \sum_{j=1}^d b_k^\top v_j v_j, \tag{A53}$$

since v_js are eigenvalues of H, which can be used as the basis for $\mathbb{R}^d$. Then we substitute this into Equation (A50) and we have

$$\alpha_k = \sum_{j=1}^d \lambda_j (a_k^\top v_j)^2 + \sum_{j=1}^d \lambda_j (b_k^\top v_j)^2 \geq \lambda_1 \sum_{j=1}^d (a_k^\top v_j)^2 + (b_k^\top v_j)^2) = \lambda_1. \tag{A54}$$

This means that any real part of the eigenvalue of A is larger than λ_1 which is the smallest eigenvalue of H. Thus, if the α_1 is the smallest real part of the eigenvalue of A, that is larger than the smallest eigenvalue of H. This concludes the proof.

In the same way,

$$\alpha_k = \sum_{j=1}^{d} \lambda_j (a_k^\top v_j)^2 + \sum_{j=1}^{d} \lambda_j (b_k^\top v_j)^2 \leq \lambda_d \sum_{j=1}^{d} (a_k^\top v_j)^2 + (b_k^\top v_j)^2) = \lambda_d, \tag{A55}$$

which means any real part of the eigenvalues of A is smaller than the largest eigenvalue of H. Thus, if α is the largest real part of the eigenvalues of A, it is smaller than the largest eigenvalue of H.

Equality condition:

Next, we discuss when the equality holds for $\alpha_1 = \lambda_1$. First, we assume that eigenvalues of H are distinct, thus, there is only one eigenvector for λ_1. Later, we discuss if eigenvalues are not distinct. From Equation (A54), we have

$$\alpha_1 = \sum_{j=1}^{d} \lambda_j (a_1^\top v_j)^2 + \sum_{j=1}^{d} \lambda_j (b_1^\top v_j)^2 \geq \lambda_1 \sum_{j=1}^{d} (a_1^\top v_j)^2 + (b_1^\top v_j)^2) = \lambda_1, \tag{A56}$$

in general. Please note that if a_1 and b_1 does not correspond to v_1, then $\lambda_{j\neq 1} > \lambda_1$ must appear in the summation and equality never holds. So, the condition is

$$a_1, b_1 \propto v_1, \tag{A57}$$

must hold for the equality.

Based on this, let us assume that $w_1 = ca_1 + ic'b_1$ where $c^2 + c'^2 = 1$. We consider the case $a_1 = b_1 = v_1$. Then we need to solve the simultaneous equations

$$A(ca_1 + ic'b_1) = (\lambda_1 + i\beta_1)(ca_1 + ic'b_1) = (\lambda_1 c - c'\beta_1)v_1 + i(c\beta_1 + \lambda_1 c')v_1, \tag{A58}$$

this is obtained by the definition of the eigenvalue of A and

$$A(ca_1 + ic'b_1) = \lambda_1^{1/2} c(I\lambda_1^{1/2} + \alpha H^{1/2} J)v_1 + i\lambda_1^{1/2} c'(I\lambda_1^{1/2} + \alpha H^{1/2} J)v_1, \tag{A59}$$

this is obtained from the definition of eigenvalues of H. Then multiplying v_1 from the left, we obtain $c\beta_1 = 0$ and $c'\beta_1 = 0$. Thus, $\beta_1 = 0$. $\beta_1 = 0$ means$b_1 = 0$ from the property of the complex eigenvectors. Thus, we obtain $w_1 = a_1 = v_1$ for $\lambda_1 = \alpha_1$. Then, the following relation holds,

$$\lambda_1 v_1 = Av_1 = Hv_1 + \alpha H^{1/2} J H^{1/2} v_1 = \lambda_1 v_1 + \alpha \lambda_1^{1/2} H^{1/2} J v_1. \tag{A60}$$

Since $\lambda_1 \neq 0$ and $H^{1/2}$ has the inverse matrix, this condition indicates that

$$\alpha J v_1 = 0. \tag{A61}$$

This is the condition that $\lambda_1 = \alpha_1$ holds. The same relation can be derived for $\lambda_d = \alpha_d$.

Next, we assume that eigenvalues of H are not distinct. Let us denote the set of eigenvectors of the eigenvalue λ_1^0 as $\{v_1^0\}$. Please note that if a_1 and b_1 does not included in V_1^0, then $\lambda_{j\neq 1} > \lambda_1$ must appear and equality never holds. Thus

$$a_1, b_1 \in V_1^0 \tag{A62}$$

must hold for equality. Based on this, let us assume that $w_1 = ca_1 + ic'b_1$ where $c^2 + c'^2 = 1$. We consider the case $a_1 \neq b_1$. Then

$$\begin{aligned} &H^{-1/2} A(ca_1 + ic'b_1) = \lambda_1^{-1/2}(\lambda_1 + i\beta_1)(ca_1 + ic'b_1) \\ &H^{-1/2}(H + \alpha H^{1/2} J H^{1/2})(ca_1 + ic'b_1) = \lambda_1^{1/2} c(I + \alpha J)a_1 + i\lambda_1^{1/2} c'(I + \alpha J)b_1, \end{aligned} \tag{A63}$$

then we obtain the condition

$$\lambda_1 c\alpha J a_1 = -\beta_1 c' b_1 \tag{A64}$$
$$\lambda_1 c'\alpha J b_1 = \beta_1 c a_1. \tag{A65}$$

□

Appendix G. Proofs of Random Matrices

Appendix G.1. Proof of Theorem 5

Proof. The proof is the straightforward consequence of lemma in [43], that is

Lemma in ([43]) *If $f(x_1, \ldots, x_m)$ is a polynomial in real variables $x_1, \ldots, x_m$, which is not identically zero, then the subset $N_m = \{(x_1, \ldots, x_m) | f(x_1, \ldots, x_m) = 0\}$ of the Euclidean m-space $\mathbb{R}^m$ has the Lebesgue measure zero.*

We use this lemma to prove that the probability of $\lambda_1 = \alpha_1$ is 0 by showing that the probability mass of $\lambda_1 = \alpha_1$ has Lebesgue measure zero.

We use the same notation as in Appendix F. Recall Equation (A64), which is the condition of equality about $\lambda_1 = \alpha_1$. We express the elements of a_1 and b_1 as $a_1 = (a_1^1, \ldots, a_1^d)^\top$ and $b_1 = (b_1^1, \ldots, b_1^d)^\top$. Then the equality condition can be written as

$$\sum_{i=1}^{d}(\sum_{j=1}^{d} \lambda_1 c\alpha J_{i,j} a_1^j + \beta_1 c' b_1^i))^2 + \sum_{i=1}^{d}(\sum_{j=1}^{d} \lambda_1 c'\alpha J_{i,j} b_1^j - \beta_1 c a_1^i))^2 = 0. \tag{A66}$$

Then we define the polynomial about $\{J_{i,j}\}$

$$f(J_{1,2}, \ldots, J_{d-1,d}) = \sum_{i=1}^{d}(\sum_{j=1}^{d} \lambda_1 c\alpha J_{ij} a_1^j + \beta_1 c' b_1^i))^2 + \sum_{i=1}^{d}(\sum_{j=1}^{d} \lambda_1 c'\alpha J_{ij} b_1^j - \beta_1 c a_1^i))^2. \tag{A67}$$

To apply lemma of [43], we must confirm that $f(J_{1,2}, \ldots, J_{d-1,d})$ is not always 0. This is clear from the definition of f since we generate $J_{1,2}, \ldots, J_{d-1,d}$ randomly from the distribution that is absolutely continuous with respect to Lebesgue measure and $\lambda_1 \neq 0$ and $c^2 + c'^2 = 1$ and either $a_1, b_1 \neq 0$.

Then, given an evaluation point x, from lemma of [43], the subset of $\{J_{i,j}\} \in \mathbb{R}^{d(d-1)/2}$ that satisfies $f(J_{1,2}, \ldots, J_{d-1,d}) = 0$ has Lebesque measure zero. Thus, if we generate $\{J_{i,j}\}$ from the probability measure which is absolutely continuous with respect to Lebesque measure, (such as Gaussian distribution), $f(J_{1,2}, \ldots, J_{d-1,d}) = 0$ holds probability 0. This concludes the proof. □

Appendix G.2. Proof of Lemma 1

Proof. We first discuss the condition about $\mathrm{Ker} J_0 = \{0\}$. Since $J = J_0 \otimes I_d$, and we denote the set of eigenvalues of J_0 as $\{\omega_i\}$. In general, the eigenvalues of the matrix that is composed of the Kronecker product with two matrices, e.g., A and B, are given as the product of each eigenvalue of A and B [44]. Thus, since J is the Kronecker product of J_0 and I_d, if J_0 does not have 0 as an eigenvalue, J does not have 0 as an eigenvalue.

Next, we discuss another equality condition. We use the similar notation as in Appendix F, but now the dimension of the matrix J is dN. We express the eigenvalue which has the smallest real part as λ_1^α and its eigenvector as $\omega_1^\alpha = a_1 + ib_1$. The elements of a_1 and b_1 as $a_1 = (a_1^1, \ldots, a_1^d, a_1^{d+1}, \ldots, a_1^{dN})^\top \in \mathbb{R}^{dN}$ and $b_1 = (b_1^1, \ldots, b_1^d, b_1^{d+1}, \ldots, b_1^{dN})^\top$. We also express these as $a_1 = (a_1^{(1)}, \ldots, a_1^{(N)})^\top \in \mathbb{R}^{dN}$ where $a_1^{(i)} = (a_1^{(i-1)d+1}, \ldots, a_1^{id})^\top \in \mathbb{R}^d$.

We use the Kronecker product property:

$$Ja_1 = (J_0 \otimes I_d)a_1 = \left(\sum_{i=1}^{N} J_{0|i,1} a_1^{(i)}, \ldots, \sum_{i=1}^{N} J_{0|i,N} a_1^{(i)}\right)^\top, \tag{A68}$$

where $J_{0|i,j}$ indicates the element of i-th row and j-th column of J_0 where we use the property of the Kronecker product and the Vec operator in the second equality [44].

The proof is almost similar to Appendix G.1. Then the equality condition can be written as

$$\sum_{n=1}^{N}\left\|\lambda_1 c\alpha\sum_i J_{0|i,n}a_1^{(i)}+\beta_1 c' b_1^{(n)}\right\|^2+\sum_{n=1}^{N}\left\|\lambda_1 c'\alpha\sum_i J_{0|i,n}b_1^{(i)}+\beta_1 c a_1^{(n)}\right\|^2=0, \tag{A69}$$

where $\|\cdot\|$ is the d-dimensional Euclidean norm since $a_1^{(n)}, b_1^{(n)} \in \mathbb{R}^d$. Then we define the polynomial about $\{J_{i.j}\}$

$$f(J_{1,2},\ldots,J_{N-1,N})=\sum_{n=1}^{N}\left\|\lambda_1 c\alpha\sum_i J_{0|i,n}a_1^{(i)}+\beta_1 c' b_1^{(n)}\right\|^2+\sum_{n=1}^{N}\left\|\lambda_1 c'\alpha\sum_i J_{0|i,n}b_1^{(i)}+\beta_1 c a_1^{(n)}\right\|^2. \tag{A70}$$

In a similar discussion with Appendix G.1, it is clear that f is not always 0. Thus, given an evaluation point x, from lemma of [43], the subset of $\{J_{i,j}\} \in \mathbb{R}^{N(N-1)/2}$ that satisfies $f(J_{1,2},\ldots,J_{N-1,d})=0$ has Lebesque measure zero. Thus, if we generate $\{J_{i,j}\}$ from the probability measure which is absolutely continuous with respect to Lebesque measure, (such as Gaussian distribution), $f(J_{1,2},\ldots,J_{N-1,N})=0$ holds probability 0. This concludes the proof. □

Appendix G.3. Extending the Theorem to the Path

About Theorem 5 and Lemma 1, the statement holds true when we fix an evaluation point x. To ensure the acceleration, we need to extend Theorem 5 and Lemma 1 from a single evaluation point to the path of the stochastic process for S-LD, S-PLD, S-ULD, and S-PULD.

First, the condition of $\mathrm{Ker}J_0=\{0\}$ is not related to the evaluation point. Thus, we need to consider the equality condition for $\mathrm{Re}\lambda_1^{\alpha}=\lambda_1^0$. As for this condition, as we had seen in Theorem 5 and Lemma 1, if we generate the random matrix J which is absolutely continuous with respect to Lebesgue measure, then the equality condition is not satisfied with probability 1 at the given evaluation point. The important point in those proof is to prove that the event when the equality holds has Lebesgue measure 0 at the given evaluation point using the lemma of [43].

Let us consider when two evaluation points are given (e.g., x_1, x_2), and we check whether the random matrix J satisfies the above equality condition or not. We can easily prove that at each evaluation point, such an event (we express them as S_1 and S_2) has Lebesgue measure 0 using the lemma of [43] (We refer to this as $P(S_1)=0$ and $P(S_2)=0$ where P is the law induced by generating the random matrix that has independent $d(d-1)/2$ elements). So, the volume of the event of sum of S_1 and S_2 are also 0 ($P(S_1\bigcup S_2=0)$. By repeating this procedure, when given a finite number of evaluation points, $(x_1,\ldots,x_k)$, the sum of such probability is 0 (this indicates $P(S_1\bigcup S_2,\ldots,\bigcup S_k)=0$).

When we consider the discretized dynamics of S-LD, S-PLD, and so on, and update samples up to k-iterations, then there exist k evaluation points. So, by applying the above discussion, we can ensure that along the path of the discretized dynamics, the equality condition does not hold with probability 1. On the other hand, as for the continuous dynamics, the evaluation point is infinite, thus when we cannot conclude that the probability that the equality does not hold is 1.

Appendix H. Proof of Theorem 6

We use the same notation as in Appendix F. We consider the expansion concerning α and we consider the following setting,

$$w_j := v_j + \delta v_j \tag{A71}$$

$$\mu_j := \lambda_j + \delta\lambda_j, \tag{A72}$$

which indicates that by introducing the skew-acceleration terms, the pairs of eigenvalues and eigenvectors of H' are expressed by the small perturbation for the eigenvalues and eigenvectors of H. Since $\{v_j\}_{j=1}^d$ are the eigenvalues of H and they can be used as an orthogonal basis, thus we expand δv by this basis. We obtain

$$\delta v_j = \sum_{k \neq j}^{d} c_{jk} v_k, \tag{A73}$$

where $c_{jk} = \delta v_j^\top v_k$.

Appendix H.1. Asymptotic Expansion When the Smallest Eigenvalue of $H(x)$ Is Positive

We work on the similar matrix of H', that is $H + \alpha V$ where $V := H^{1/2} J H^{1/2}$. See Appendix G.1 for the detail. Please note that this similar matrix only exists when the smallest eigenvalue of $H(x)$ is positive. Thus, the following discussion cannot apply to the case at the saddle point, where negative eigenvalues appear. We discuss the saddle point expansion later.

From the definition, we have

$$H' w_j = H w_j + \alpha V w_j = \mu_j w_j = (\lambda_j + \delta\lambda_j)(v_j + \delta v_j), \tag{A74}$$

We rearrange this equation as

$$H v_j + H \delta v_j + \alpha V v_j + \alpha V \delta v_j = \lambda_j v_j + \delta\lambda_j v_j + \lambda_j \delta v_j + \delta\lambda_j \delta v_j. \tag{A75}$$

First, we focus on the first-order expansion. This means we neglect high-order terms. Then, we have

$$H v_j + H \delta v_j + \alpha V v_j = \lambda_j v_j + \delta\lambda_j v_j + \lambda_j \delta v_j. \tag{A76}$$

By multiplying v_j to Equation (A76 from the left-hand side, we have

$$\lambda_j + \lambda_j v_j^\top \delta v_j + \alpha v_j^\top V v_j = \lambda_j + \delta\lambda_j + \lambda_j v_j^\top \delta v_j, \tag{A77}$$

Since $v_j^\top V v_j = 0$ due to the skew-symmetric property of V. Thus, we have

$$\delta\lambda_j = 0, \tag{A78}$$

up to the first-order expansion. Then we substitute this into Equation (A76 and multiplying v_i where $i \neq j$, we have

$$\lambda_i c_{ji} + \alpha v_i^\top V v_j = \lambda_j c_{ji}. \tag{A79}$$

Then we have

$$c_{ji} = \frac{\alpha v_i^\top V v_j}{\lambda_j - \lambda_i}. \tag{A80}$$

Then we obtain

$$\delta v_j = \alpha \sum_{i \neq j}^{d} \frac{v_i^\top V v_j}{\lambda_j - \lambda_i} v_i. \tag{A81}$$

We substitute this into Equation (A75), and multiplying $v_j^\top$, we have

$$v_j^\top H\alpha \sum_{i\neq j}^{d} \frac{v_i^\top V v_j}{\lambda_j - \lambda_i} v_i + \alpha v_j^\top V v_j + \alpha v_j^\top V \alpha \sum_{i\neq j}^{d} \frac{v_i^\top V v_j}{\lambda_j - \lambda_i} v_i$$
$$= \delta\lambda_j v_j^\top v_j + \lambda_j v_j^\top \alpha \sum_{i\neq j}^{d} \frac{v_i^\top V v_j}{\lambda_j - \lambda_i} v_i + \delta\lambda_j v_j^\top \alpha \sum_{i\neq j}^{d} \frac{v_i^\top V v_j}{\lambda_j - \lambda_i} v_i. \tag{A82}$$

Since $v_j^\top V v_j = 0$ and $v_j^\top v_i = 0$ and $v_j^\top v_j = 1$, we have

$$\alpha^2 \sum_{i\neq j}^{d} \frac{v_i^\top V v_j}{\lambda_j - \lambda_i} v_j^\top V v_i = \delta\lambda_j. \tag{A83}$$

Thus, we have

$$\mu_j - \lambda_j = \alpha_j + i\beta_j - \lambda_j = -\alpha^2 \sum_{i\neq j}^{d} \frac{(v_i^\top V v_j)^2}{\lambda_j - \lambda_i}. \tag{A84}$$

Thus, by taking the real part, and note that $\mathrm{Re}\lambda_j(\alpha) = \alpha_j$, we have

$$\mathrm{Re}\lambda_j(\alpha) - \lambda_j = \alpha^2 \mathrm{Re} \sum_{i\neq j}^{d} \frac{(v_i^\top V v_j)^2}{\lambda_i - \lambda_j} + \mathcal{O}(\alpha^3) = \alpha^2 \sum_{i\neq j}^{d} \frac{\lambda_i \lambda_j (v_i^\top J v_j)^2}{\lambda_i - \lambda_j} + \mathcal{O}(\alpha^3). \tag{A85}$$

This concludes the proof.

Appendix H.2. Expansion of the Eigenvalue at the Saddle Point

Here we derive the formula of the expansion of the eigenvalue at the saddle point. Since the smallest eigenvalue is negative, we cannot use the similar matrix as shown above. Instead, we use the relation,

$$\mu_j H w_j = H \mu_j w_j = H(I + \alpha J) H w_j \tag{A86}$$

where we used the definition of the eigenvalues and eigenvectors. Here, we express $H' := (I + \alpha J)H$ and its pairs of eigenvalues and eigenvectors as $\{(\mu_i, w_i)\}_{i=1}^d$. As introduced in the above, we substitute the expansion to Equation (A86), then we obtain

$$(\lambda_j + \delta\lambda_j) H(v_j + \delta v_j) = H(I + \alpha J) H(v_j + \delta v_j) \tag{A87}$$

Then, in the same way as above, since $\{v_j\}_{j=1}^d$ are the eigenvalues of H and they can be used as an orthogonal basis, we expand δv by this basis. This means

$$\delta v_j = \sum_{k=1}^{d} c_{jk} v_k, \tag{A88}$$

where $c_{jk} = \delta v_j^\top v_k$. By multiplying v_i to Equation (A87) where $i \neq j$ from left-hand side and neglecting high-order terms, we have

$$c_{ji} = \frac{\lambda_j}{\lambda_j - \lambda_i} (v_i^\top \alpha J v_i). \tag{A89}$$

Next, Then by multiplying v_j to Equation (A87) from left-hand side, we have

$$v_j H(\alpha J) H \delta v_j = (\delta\lambda_j)(\lambda_j + \lambda_j v_j^\top \delta v_j) \tag{A90}$$

Then by substituting δv_j with coefficient Equation (A89), we have

$$\delta\lambda_j = \alpha^2 \sum_{i\neq j}^{d} \frac{\lambda_i\lambda_j (v_i^\top J v_j)^2}{\lambda_i - \lambda_j} + \mathcal{O}(\alpha^3) \tag{A91}$$

This concludes the proof.

Appendix I. Convergence Rate of Parallel Sampling Schemes

Appendix I.1. Proof of Lemma 2

First, we introduce the notations. we express the random variables of S-PLD as $Y_t^{\otimes N}$. We express the measure induced by S-PLD as $\mu_t^{\otimes N}(\alpha)$, which uses the αJ as an interaction term. Thus, we express the measure of PLD as $\mu_{kh}^{\otimes N}(0)$, we can decompose the measure as marginals. We also denote the marginal measure of S-PLD for $Y_t^{(n)}$ $\nu_t^{(n)}(\alpha)$. Please note that initial distribution is $\mu_0^{\otimes N}$ and its marginals are μ_0 as defined in Assumption 4.

Please note that the marginal measure of PLD is the same as those of LD if the initial measures are all the same, thus each marginal satisfy the Poincaré constant m_0. This is also the result of the tensorization property of the spectral gap (Proposition 4.3.1 in Bakry et al. [19]).

As for the initial condition, from the fact that χ^2 divergence is the special case of Renyi divergence ($\alpha = 4$), and from the tensorization property of the Renyi divergence (See Theorem 28 in [45]), we have

$$\chi^2(\mu_t^{\otimes N}(0), \pi^{\otimes N}) \leq e^{-2\beta^{-1}m_0 t}\chi^2(\mu_0^{\otimes N}, \pi^{\otimes N}) = \sum_{n=1}^{N} e^{-2\beta^{-1}m_0 t}\chi^2(\mu_0, \pi). \tag{A92}$$

Then we have

$$\chi^2(\mu_t^{\otimes N}(0), \pi^{\otimes N}) \leq e^{-2\beta^{-1}m_0 t}\chi^2(\mu_0^{\otimes N}, \pi^{\otimes N}) = Ne^{-2\beta^{-1}m_0 t}\chi^2(\mu_0, \pi). \tag{A93}$$

If the skew acceleration is applied, from the same discussion as S-LD (see Appendix C.1), S-PLD has the Poincaré constant which is larger than m_0. We express it as $m(\alpha, N)(\geq m_0)$. Then we have

$$\chi^2(\mu_t^{\otimes N}(\alpha), \pi^{\otimes N}) \leq Ne^{-2\beta^{-1}m(\alpha,N)t}\chi^2(\mu_0, \pi). \tag{A94}$$

At first, since there exists a constant N in the convergence bound, this bound seems not useful. However, as we discussed below, when we bound the bias or variance, these bound is meaningful. For example, let us consider approximating the true expectation $\int f(x)d\pi(x)$ by the ensemble samples $\frac{1}{N}\sum_{n=1}^{N} f(X_t^{(n)})$. Then we are interested in bounding the error

$$\left| \mathbb{E}\frac{1}{N}\sum_{n=1}^{N} f(X_k^{(n)}) - \int_{\mathbb{R}^d} f d\pi \right|. \tag{A95}$$

For this purpose, we can bound this by 2-Wasserstein distance as

$$\left| \mathbb{E}\frac{1}{N}\sum_{n=1}^{N} f(X_k^{(n)}) - \int_{\mathbb{R}^d} f d\pi \right|, \leq \frac{L_f}{\sqrt{N}} W_2(\mu_{kh}^{\otimes N}(\alpha), \pi^{\otimes N}) \tag{A96}$$

where we assumed that f shows L_f lipschitzness and used the fact that $\frac{1}{N}\sum_{n=1}^{N} f(x^{(n)})$ shows $L_f/\sqrt{N}$ lipschitzness.

To bound the distance, we use the basic relation

$$W_2^2(\nu_{kh}(\alpha), \pi^{\otimes N}) \leq 2\frac{1}{m(\alpha, N)}\chi^2(\mu_{kh}^{\otimes N}(\alpha), \pi^{\otimes N}), \tag{A97}$$

where $m(\alpha, N)$ is the Poincaré constant. This is established by the definition of Wasserstein distance and χ^2-divergence, see [46] for the detail. Then combined with above relations, we obtain the bias bound of S-PLD as

$$\left|\mathbb{E}\frac{1}{N}\sum_{n=1}^{N} f(X_k^{(n)}) - \int_{\mathbb{R}^d} f d\pi\right| \leq L_f\sqrt{\frac{2}{m(\alpha, N)}}e^{-\beta^{-1}m(\alpha,N)kh}\chi^2(\mu_0, \pi)^{1/2}. \tag{A98}$$

In the same way, we obtain the bias bound of PLD as

$$\left|\mathbb{E}\frac{1}{N}\sum_{n=1}^{N} f(X_k^{(n)}) - \int_{\mathbb{R}^d} f d\pi\right| \leq L_f\sqrt{\frac{2}{m_0}}e^{-\beta^{-1}m_0 kh}\chi^2(\nu_0, \pi)^{1/2}. \tag{A99}$$

Thus, while the explicit dependency on N disappeared, but S-PLD shows faster convergence through the relation of $m(\alpha, N) \geq m_0$. Moreover, if we use the skew matrices, which does not satisfy the equality condition, we have $m(\alpha, N) > m_0$.

Appendix I.2. Proof for S-ULD

We can characterize the convergence rate almost in the same way as Appendix C.2. The derivation is the same above, thus we only show the result

$$\left|\mathbb{E}\frac{1}{N}\sum_{n=1}^{N} f(X_k^{(n)}) - \int_{\mathbb{R}^d} f d\pi\right| \leq L_f\sqrt{\frac{2}{m(\alpha, N)}}\sqrt{\frac{1+\bar{\epsilon}}{1-\bar{\epsilon}}}e^{-\lambda_\gamma/2kh}\chi^2(\nu_0^0, \pi)^{1/2}. \tag{A100}$$

where $\bar{\epsilon}$ and λ_γ is given as follows.

$$\lambda_\gamma = \frac{\Lambda(\gamma, \bar{\epsilon}\min(\gamma, \gamma^{-1}))}{1 + \bar{\epsilon}\min(\gamma, \gamma^{-1})}, \tag{A101}$$

and

$$\Lambda(\gamma, \epsilon) = \frac{\gamma\Sigma^{-1} - \frac{1}{1+\frac{m_0\Sigma^{-1}}{\beta}}}{2} - \frac{1}{2}\sqrt{(S_{--} - S_{++})^2 + (S_{-+})^2}, \tag{A102}$$

$$S_{--} = \epsilon\lambda_{ham}, \tag{A103}$$

$$S_{-+} = -\epsilon(R_{ham} + \gamma\Sigma^{-1}/2), \tag{A104}$$

$$S_{++} = \gamma\Sigma^{-1} - \epsilon, \tag{A105}$$

$$\lambda_{ham} = 1 - \left(1 + \frac{m(\alpha, N)\Sigma^{-1}}{\beta}\right)^{-1}, \tag{A106}$$

$$\epsilon = \bar{\epsilon}\min(\gamma, \gamma^{-1}), \tag{A107}$$

where $\bar{\epsilon}$ is arbitrary sufficiently small positive value such that $\Lambda(\gamma, \bar{\epsilon}\min(\gamma, \gamma^{-1})) > 0$ is satisfies. and

$$R_{ham} \leq \sqrt{\max\{M, 2\}}. \tag{A108}$$

Appendix J. Proof of Theorem 7

We show our theorem again with explicit constants

Theorem A3. *Under Assumptions 1–7, 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*

$$\begin{aligned}&\left|\mathbb{E}\frac{1}{N}\sum_{n=1}^{N} f(X_k^{(n)}) - \int_{\mathbb{R}^d} f d\pi\right| \\ &\leq L_f\sqrt{\tilde{C}_0^2\sqrt{\delta} + \tilde{C}_1^2\sqrt{h}}k\eta + L_f\sqrt{\frac{2}{m(\alpha,N)}}\chi^2(\mu_0,\pi)^{1/2}e^{-\beta^{-1}m(\alpha,N)kh}.\end{aligned} \tag{A109}$$

where

$$\tilde{C}_0^2 = \left(12 + 8\left(\kappa_0 + 2b + \frac{2d}{\beta}\right)\right)\left(\beta C_0 + \sqrt{\beta C_0}\right), \tag{A110}$$

$$\tilde{C}_1^2 = \left(12 + 8\left(\kappa_0 + 2b + \frac{2d}{\beta}\right)\right)\left(C_1 + \sqrt{C_1}\right) \tag{A111}$$

$$C_0 = (1+\alpha)^2\left(M^2\left(\kappa_0 + 2\left(1 \vee \frac{1}{m}\right)\left(b + 2(1+\alpha)^2B^2 + \frac{d}{\beta}\right)\right) + B^2\right), \tag{A112}$$

$$C_1 = 6(1+\alpha^2)M^2(\beta C_0 + d), \tag{A113}$$

Then obtained bound is $\mathcal{O}(kh \cdot h^{1/4})$, which is independent of N. Thus, this result is much better than those in [18]. Additionally, note that we can derive the similar bias bound for skew-SGHMC in the same way as skew-SGLD.

Proof. For notational simplicity, we express the random variables of skew-SGLD which uses the αJ as an interaction term as $X_k^{\otimes N}$ and those of S-PLD as $Y_k^{\otimes N}$. In this section, for simplicity, we express them as X_k and Y_k. We denote the measure of X_k and Y_k as $\nu_{kh}^{\otimes N}$ and $\mu_{kh}^{\otimes N}$. We also denote the marginal measure of $X_k^{(n)}$ and $Y_k^{(n)}$ as $\mu_{kh}^{(n)}$ and $\nu_{kh}^{(n)}$.

Then, we first decompose the bias as

$$\begin{aligned}&\left|\mathbb{E}\frac{1}{N}\sum_{n=1}^{N} f(X_k^{(n)}) - \int_{\mathbb{R}^d} f d\pi\right| \\ &= \left|\mathbb{E}\frac{\sum_{n=1}^N f(X_k^{(n)})}{N} - \mathbb{E}\frac{\sum_{n=1}^N f(Y_k^{(n)})}{N} + \mathbb{E}\frac{\sum_{n=1}^N f(Y_k^{(n)})}{N} - \int_{\mathbb{R}^d} f d\pi\right| \\ &\leq \left|\mathbb{E}\frac{1}{N}\sum_{n=1}^{N} f(X_k^{(n)}) - \mathbb{E}\frac{1}{N}\sum_{n=1}^{N} f(Y_k^{(n)})\right| + \left|\mathbb{E}\frac{1}{N}\sum_{n=1}^{N} f(Y_k^{(n)}) - \int_{\mathbb{R}^d} f d\pi\right| \\ &\leq \frac{L_f}{N}\sum_{i=1}^{N} W_2(\nu_{kh}^{(n)}(\alpha), \mu_{kh}^{(n)}(\alpha)) + \frac{L_f}{\sqrt{N}}\underbrace{W_2(\mu_{kh}^{\otimes N}(\alpha), \pi^{\otimes N})}_{(i)},\end{aligned} \tag{A114}$$

where we used the Jensen inequality for the first term in the last inequality and we move $\frac{1}{N}\sum_{i=1}^N$ outside the $|\cdot|$. In addition, each expectation only depends on the marginal measures $\mu^{(i)}$ in the first term and we use the property of the 2-Wasserstein (2-W) distance. Furthermore, we decompose the first term as

$$\frac{L_f}{N}\sum_{n=1}^{N} W_2(\mu_{kh}^{(n)}(\alpha), \nu_{kh}^{(n)}(\alpha)) \leq \frac{L_f}{N}\left(\sum_{n=1}^{N}\underbrace{W_2(\nu_{kh}^{(n)}(\alpha), \mu_{kh}^{(n)}(0))}_{(ii)} + \underbrace{W_2(\mu_{kh}^{(n)}(\alpha), \mu_{kh}^{(n)}(0))}_{iii)}\right), \tag{A115}$$

where $\mu_{kh}^{(n)}(0)$ denotes the measure induced by *PLD*, which is the naive parallel sampling without a skew-symmetric interaction.

In conclusion, our task is to bound each (i), (ii), (iii) terms in the above. Bounding (i) is already discussed in Appendix I.1.

Next, we work on (ii) and(iii). Following [10], we use weighted CKP inequality to bound the 2-W distance. From Bolley and Villani [47], using the weighted CKP inequality, we can bound each 2-W distance by the relative entropy (KL divergence). This weighted CKP inequality indicates that

$$W_2(\nu_{kh}^{(n)}(\alpha), \mu_{kh}^{(n)}(0)) \leq C_{\mu_{kh}^{(n)}(0)} \left(\mathrm{KL}(\nu_{kh}^{(n)}(\alpha)|\mu_{kh}^{(n)}(0))^{1/2} + \left(\frac{\mathrm{KL}(\nu_{kh}^{(n)}(\alpha)|\mu_{kh}^{(n)}(0))}{2} \right)^{1/4} \right), \quad \text{(A116)}$$

with

$$C_{\mu_{kh}^{(n)}(0)} = 2 \inf_{\lambda>0} \left(\frac{1}{\lambda} \left(\frac{3}{2} + \log \int_{\mathbb{R}^d} e^{\lambda \|x^{(n)}\|^2} d\mu_{kh}^{(n)}(0) \right) \right)^{1/2}. \quad \text{(A117)}$$

and

$$W_2(\mu_{kh}^{(n)}(\alpha), \mu_{kh}^{(n)}(0)) \leq C_{\mu_{kh}^{(n)}(0)} \left(\mathrm{KL}(\mu_{kh}^{(n)}(\alpha)|\mu_{kh}^{(n)}(0))^{1/2} + \left(\frac{\mathrm{KL}(\mu_{kh}^{(n)}(\alpha)|\mu_{kh}^{(n)}(0))}{2} \right)^{1/4} \right), \quad \text{(A118)}$$

with

$$C_{\mu_{kh}^{(n)}(0)} = 2 \inf_{\lambda>0} \left(\frac{1}{\lambda} \left(\frac{3}{2} + \log \int_{\mathbb{R}^d} e^{\lambda \|x^{(n)}\|^2} d\mu_{kh}^{(n)}(0) \right) \right)^{1/2}. \quad \text{(A119)}$$

We point out that using $C_{\mu_{kh}^{(i)}(0)}$ not $C_{\nu_{kh}^{(i)}(\alpha)}$ and $C_{\mu_{kh}^{(i)}(\alpha)}$ in weighted CKP inequality is important. This is because since $\mu_{kh}^{(i)}(0)$ is the constant based on the parallel-chain Monte Carlo without skew-symmetric term, thus the parallel chain can be decomposed each independent chains. Thus, $C_{\mu_{kh}^{(i)}}$ actually does not depend on i and it does not depend on N and shows $\mathcal{O}(d)$ dependency. However, $C_{\nu_{kh}^{(i)}(\alpha)}$ and $C_{\mu_{kh}^{(i)}(\alpha)}$ show $\mathcal{O}(dN)$ which shows linear dependency on N since there is an interaction term between parallel chains and we cannot decompose the parallel chain easily. Thus, this results in unsatisfactory dependency on N. This is the reason we introduced $\mu_{kh}^{(i)}(0)$ in our theoretical analysis.

Please note that since $\mu_{kh}^{(n)}(0)$ is induced by the naive parallel chain, each marginal is independent with each other and takes the same measure if the initial measure is the same. Thus, $\mu_{kh}^{(1)}(0) = \cdots = \mu_{kh}^{(N)}(0)$. From now on, we express the marginal as $\mu_{kh}(0)$ for simplicity. Thus, $C_{\mu_{kh}^{(1)}(0)} = \cdots = C_{\mu_{kh}^{(N)}(0)} = C_{\mu_{kh}(0)}$.

Then substituting the above WKP inequalities and using the Jensen inequality, we obtain

$$\begin{aligned}
&\left| \mathbb{E} \frac{1}{N} \sum_{n=1}^{N} f(X_k^{(n)}) - \mathbb{E} \frac{1}{N} \sum_{n=1}^{N} f(Y_k^{(n)}) \right| \\
&\leq L_f C_{\mu_{kh}(0)} \frac{1}{N} \sum_{n=1}^{N} \left(\mathrm{KL}(\nu_{kh}^{(n)}(\alpha)|\mu_{kh}(0))^{1/2} + \left(\frac{\mathrm{KL}(\nu_{kh}^{(n)}(\alpha)|\mu_{kh}(0))}{2} \right)^{1/4} \right. \\
&\qquad \left. + \mathrm{KL}(\mu_{kh}^{(n)}(\alpha)|\mu_{kh}(0))^{1/2} + \left(\frac{\mathrm{KL}(\mu_{kh}^{(n)}(\alpha)|\mu_{kh}(0))}{2} \right)^{1/4} \right)
\end{aligned}$$

$$\leq L_f C_{\mu_{kh}(0)} \left(\left(\sum_{n=1}^{N} \frac{\mathrm{KL}(\nu_{kh}^{(n)}(\alpha)|\mu_{kh}(0))}{N} \right)^{\frac{1}{2}} + \left(\sum_{n=1}^{N} \frac{\mathrm{KL}(\nu_{kh}^{(n)}(\alpha)|\mu_{kh}(0))}{2N} \right)^{\frac{1}{4}} + \left(\sum_{n=1}^{N} \frac{\mathrm{KL}(\mu_{kh}^{(n)}(\alpha)|\mu_{kh}(0))}{N} \right)^{\frac{1}{2}} + \left(\sum_{n=1}^{N} \frac{\mathrm{KL}(\mu_{kh}^{(n)}(\alpha)|\mu_{kh}(0))}{2N} \right)^{\frac{1}{4}} \right). \tag{A120}$$

To analyze the discretization error, we use the following key lemma:

Lemma A1. *Assume that there exist random variables* $\{X_i \in \Omega_i\}_{i=1}^N$ *and* $\{Y_i \in \Omega_i\}_{i=1}^N$. *We denote the product space as* $\Omega^{\otimes N} := \Omega_1 \times \ldots \Omega_N$. *Let us introduce* $X = (X_1, \ldots, X_N) \in \Omega^{\otimes N}$ *and* $Y = (Y_1, \ldots, Y_N) \in \Omega^{\otimes N}$. *Let us express their joint probability measures as expressed as* $P(X) := P(X_1, \ldots, X_N)$, $Q(Y) := Q(Y_1, \ldots, Y_N)$, *let us denote the marginal measures of each Xs and Ys as* $\{P_i(X_i)\}_{i=1}^N$ *and* $\{Q_i(Y_i)\}_{i=1}^N$. *If* $P_i << Q_i$ *holds, we have*

$$\sum_{i=1}^{N} \mathrm{KL}(P_i(X_i) \| Q_i(Y_i)) \leq \mathrm{KL}(P(X) \| Q(Y)), \tag{A121}$$

A proof is given in Appendix J.1. We apply this lemma as

$$\sum_{n=1}^{N} \mathrm{KL}(\mu_{kh}^{(n)}|\mu_{kh}(0)) \leq \mathrm{KL}(\nu_{kh}^{\otimes N}|\mu_{kh}^{\otimes N}(0)), \tag{A122}$$

$$\sum_{n=1}^{N} \mathrm{KL}(\mu_{kh}^{(n)}(\alpha)|\mu_{kh}(0)) \leq \mathrm{KL}(\mu_{kh}^{\otimes N}(\alpha))|\mu_{kh}^{\otimes N}(0))). \tag{A123}$$

Combining these results with the above bias bound, we obtain

$$\left| \mathbb{E} \frac{1}{N} \sum_{n=1}^{N} f(X_k^{(n)}) - \mathbb{E} \frac{1}{N} \sum_{n=1}^{N} f(Y_k^{(n)}) \right| \leq L_f C_{\mu_{kh}(0)} \left(\left(\frac{\mathrm{KL}(\nu_{kh}^{\otimes N}(\alpha)|\mu_{kh}^{\otimes N}(0))}{N} \right)^{\frac{1}{2}} + \left(\frac{\mathrm{KL}(\nu_{kh}^{\otimes N}(\alpha)|\mu_{kh}^{\otimes N}(0))}{2N} \right)^{\frac{1}{4}} + \left(\frac{\mathrm{KL}(\mu_{kh}^{\otimes N}(\alpha)(\alpha)|\mu_{kh}^{\otimes N}(0))}{N} \right)^{\frac{1}{2}} + \left(\frac{\mathrm{KL}(\mu_{kh}^{\otimes N}(\alpha)|\mu_{kh}^{\otimes N}(0))}{2N} \right)^{\frac{1}{4}} \right). \tag{A124}$$

Thus, we need to bound $\mathrm{KL}(\mu_{kh}^{(i)}(\alpha)|\mu_{kh}^{\otimes N}(0))$ and $\mathrm{KL}(\nu_{kh}^{\otimes N}(\alpha)|\mu_{kh}^{\otimes N}(0))$ and $C_{\mu_{kh}(0)}$. We can upper-bound them using the results of [2]. For that purpose, we need to replace the constants in [2] as we show in the below. Here, we discuss how the constants in the assumption are changed in the ensemble scheme. We define

$$\nabla u^{\otimes N}(x^{\otimes N}) := (\nabla u(x^{(1)}), \ldots, \nabla u(x^{(N)})) \tag{A125}$$

First, we focus on the smoothness condition. From Assumption 2 and lemma 8 in [18], we have

$$\|(I + \alpha J)\nabla u^{\otimes N}(x^{\otimes N}, z) - (I + \alpha J)\nabla u^{\otimes N}(y^{\otimes N}, z))\| \leq M(1 + \alpha)\|x^{\otimes N} - y^{\otimes N}\|. \tag{A126}$$

where the norm in the right-hand side is the Euclidean norm in $\mathbb{R}^{dN}$.

Next, we discuss the smoothness condition. Define $\nabla U_\alpha(x^{\otimes N}) := \nabla U^{\otimes N}(x^{\otimes N}) + \alpha J \nabla U^{\otimes N}(x^{\otimes N})$. Then, Let $x^{\otimes N} \in \mathbb{R}^{dN}$ and under the assumptions 1 to 6, we have

$$x^{\otimes N} \cdot \nabla U_\alpha(x^{\otimes N}) \geq m\|x^{\otimes N}\|^2 - bN. \tag{A127}$$

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

$$\|(I+\alpha J)\nabla U^{\otimes N}(0^{\otimes N})\| \leq B\sqrt{N}(1+\alpha). \tag{A128}$$

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

$$\begin{aligned}
&\mathbb{E}[\|(I+\alpha J)\nabla\hat{U}^{\otimes N}(x^{\otimes N})-(I+\alpha J)\nabla U^{\otimes N}(x^{\otimes N})\|^2]\\
&\leq (1+\alpha)^2\mathbb{E}[\nabla\hat{U}^{\otimes N}(x^{\otimes N})-\nabla U^{\otimes N}(x^{\otimes N})\|^2]\\
&\leq (1+\alpha)^2\sum_{i=1}^{N}\mathbb{E}[\nabla\hat{U}(x^{(i)})-\nabla U(x^{(i)})\|^2]\\
&\leq (1+\alpha)^2\sum_{i=1}^{N}2\delta\left(M^2\|x^{(i)}\|^2+B^2\right)\\
&\leq 2\delta(1+\alpha)^2\left(M^2\|x^{\otimes N}\|^2+NB^2\right).
\end{aligned} \tag{A129}$$

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

$$\kappa_0^{\otimes N} := \log\int_{\mathbb{R}^{dN}} e^{\|x^{\otimes N}\|^2}\mu_0^{\otimes N}(x^{\otimes N})dx^{\otimes N} = \log\prod_{n=1}^{N}\left(\int_{\mathbb{R}^d} e^{\|x^{(n)}\|^2}\mu_0(x^{(n)})dx\right) = N\kappa_0. \tag{A130}$$

In this way, the constants in the assumptions are modified and expressed with N and α. Then combined with the results of [2], we can derive the following relations

$$C_{\nu_{kh}(0)} \leq 12+8\left(\kappa_0+2b+\frac{2d}{\beta}\right), \tag{A131}$$

$$\mathrm{KL}(\nu_{kh}^{\otimes N}|\mu_{kh}^{\otimes N}(0)) \leq N(C_0\beta\delta+C_1\eta)k\eta, \tag{A132}$$

$$\mathrm{KL}(\mu_{kh}^{\otimes N}(\alpha)|\mu_{kh}^{\otimes N}(0)) \leq N\frac{\beta}{2}\alpha^2M^2(\kappa_0+\frac{b+d/\beta}{m})k\eta, \tag{A133}$$

where

$$C_0 = (1+\alpha)^2\left(M^2\left(\kappa_0+2\left(1\vee\frac{1}{m}\right)\left(b+2(1+\alpha)^2B^2+\frac{d}{\beta}\right)\right)+B^2\right), \tag{A134}$$

$$C_1 = 6(1+\alpha^2)M^2(\beta C_0+d). \tag{A135}$$

This concludes the proof. □

Appendix J.1. Proof of Lemma A1

Proof. We prove this lemma using the Donsker–Varadhan representation of the relative entropy [48]. The relative entropy admits the dual representation as:

$$\mathrm{KL}(P(X)\|Q(Y)) = \sup_{T:\Omega^{\otimes N}\to\mathbb{R}}\mathbb{E}_{P(X)}[T]-\log\mathbb{E}_{Q(Y)}[e^T], \tag{A136}$$

where supremum is taken over all function T of which the expectation of e^T and T are finite. We then restrict the function class into a class $\mathcal{F}(T) = \{T(X)|\exists T_i:\Omega_i\to\mathbb{R}, s.t. T(X) = \sum_{i=1}^N T_i(X_i)\}$ where each expectation of e^{T_i} and T_i are finite. Then by definition,

$$\mathrm{KL}(P(X)\|Q(Y)) = \sup_{T:\Omega\to\mathbb{R}}\mathbb{E}_{P(X)}[T]-\log\mathbb{E}_{Q(Y)}[e^T] \geq \sup_{T\in\mathcal{F}}\mathbb{E}_{P(X)}\left[\sum_i T_i\right]-\log\mathbb{E}_{Q(Y)}\left[e^{\sum_i T_i}\right]. \tag{A137}$$

Then we have

$$\begin{aligned}\mathrm{KL}(P(X)\|Q(Y)) &\geq \sup_{T\in\mathcal{F}} \sum_i \mathbb{E}_{P_i(X_i)}[T_i] - \log \prod_i \mathbb{E}_{Q_i(Y_i)}\left[e^{T_i}\right] \\ &= \sup_{T\in\mathcal{F}} \sum_i \left(\mathbb{E}_{P_i(X_i)}[T_i] - \log \mathbb{E}_{Q_i(Y_i)}\left[e^{T_i}\right]\right) \\ &= \sum_i \sup_{T_i:\Omega_i\to\mathbb{R}} \mathbb{E}_{P_i(X_i)}[T_i] - \log \mathbb{E}_{Q_i(Y_i)}\left[e^{T_i}\right] \\ &= \sum_{i=1}^N \mathrm{KL}(P_i(X_i)\|Q_i(Y_i)). \end{aligned} \tag{A138}$$

□

Appendix K. Order Expansion

Appendix K.1. Bias Expansion for S-PLD

Recall that the bias of S-PLD is

$$\begin{aligned}&\left|\mathbb{E}\frac{1}{N}\sum_{n=1}^N f(X_k^{(n)}) - \int_{\mathbb{R}^d} f d\pi\right| \\ &\leq L_f\sqrt{\tilde{C}_0^2\sqrt{\delta} + \tilde{C}_1^2\sqrt{h}}k\eta + L_f\sqrt{\frac{2}{m(\alpha,N)}}\chi^2(\mu_0),\pi)^{1/2}e^{-\beta^{-1}m(\alpha,N)kh}. \end{aligned} \tag{A139}$$

where

$$\tilde{C}_0^2 = \left(12 + 8\left(\kappa_0 + 2b + \frac{2d}{\beta}\right)\right)\left(\beta C_0 + \sqrt{\beta C_0}\right), \tag{A140}$$

$$\tilde{C}_1^2 = \left(12 + 8\left(\kappa_0 + 2b + \frac{2d}{\beta}\right)\right)\left(C_1 + \sqrt{C_1}\right) \tag{A141}$$

$$C_0 = (1+\alpha)^2\left(M^2\left(\kappa_0 + 2\left(1\vee\frac{1}{m}\right)\left(b + 2(1+\alpha)^2B^2 + \frac{d}{\beta}\right)\right) + B^2\right), \tag{A142}$$

$$C_1 = 6(1+\alpha^2)M^2(\beta C_0 + d), \tag{A143}$$

First, we discuss the convergence of the continuous dynamics. Using the eigenvalue expansion in Theorem 6 , with some positive constant d_0, we have

$$m(\alpha, N) \approx m_0 + \alpha^2 d_0 + \mathcal{O}(\alpha^3). \tag{A144}$$

Then by assuming α^2 is small enough and considering the Tayler expansion, we have

$$L_f\sqrt{\frac{2}{m(\alpha,N)}}\chi^2(\mu_0,\pi)^{1/2}e^{-\beta^{-1}m(\alpha,N)t} \approx L_f\chi^2(\mu_0,\pi)^{1/2}\sqrt{2}\left(\frac{1}{\sqrt{m_0}} - \frac{d_0}{2m_0^{3/2}}\alpha^2\right)e^{-\beta^{-1}m_0 t}. \tag{A145}$$

As for the discretization and stochastic gradient error, using the Taylor expansion, there exists a positive constant d_1 and d_2, such that

$$L_f\sqrt{\tilde{C}_0^2\sqrt{\delta} + \tilde{C}_1^2\sqrt{h}}k\eta \approx (d_1\alpha + d_2\alpha^2 + \mathrm{Const})kh. \tag{A146}$$

Combining these terms, we have

$$\left|\mathbb{E}\frac{1}{N}\sum_{n=1}^N f(X_k^{(n)}) - \int_{\mathbb{R}^d} f d\pi\right| \leq (d_1\alpha + d_2\alpha^2)kh - \alpha^2 L_f\chi^2(\mu_0,\pi)^{1/2}\frac{1}{\sqrt{2}m_0^{3/2}}e^{-\beta^{-1}m_0 t} + \mathrm{Const}. \tag{A147}$$

Thus, there exists an optimal α^*, which minimizes the bias. Please note that at $k = 0$, acceleration always occurs. As k goes to infinity, the second third terms 0, thus the first

term will be dominant, which means we have larger discretization and stochastic gradient error.

Appendix L. Hyperparameters of the Proposed Algorithm

Here we discuss how to set hyperparameters in the algorithm. There are three hyperparameters, α_0, η, and c. We numerically found that setting $c = 0.95$ work well for real dataset including LDA experiment, and Bayesian neural network regression and classification. For toy dataset, we set $c = 0.9$.

As for α_0 and η, we empirically found that using the following scaling trick works well for real dataset including LDA experiment, and Bayesian neural network regression and classification,

$$\alpha_0 \approx \frac{1}{\sqrt{\frac{1}{N^2}\sum_n \nabla U(x_0^{(n)})^2}} Nh. \tag{A148}$$

and using $\eta \approx 0.1\alpha_0$. The intuition is that the magnitude of the gradient can be very different in each dimension, so we introduce the scaling by the gradient. We also multiply h so that the stochastic gradient and discretization error of the skew term will not be dominant compared to usual gradient term. Finally, we multiply some constant so that α_0 will not be too small.

Appendix M. Proof of Theorem 8

In this section, we derive the upper-bound of the bias of skew-SGLD based on [23]. This approach requires us to use the logarithmic Sobolev inequality [19], which is stronger than the Poincaré inequality. First, we present the definition of the logarithmic Sobolev inequality. We say that π on $\mathbb{R}^d$ with $\mathcal{L}$ satisfies the logarithmic Sobolev inequality with constant λ in case for all function f on $\mathbb{R}^d$ with $\int_{\mathbb{R}^d} u^2 d\pi = 1$,

$$\int_{\mathbb{R}^d} f^2 \ln f^2 d\pi \leq \frac{2}{\lambda}\int_{\mathbb{R}^d} -f\mathcal{L}f d\pi. \tag{A149}$$

This logarithmic Sobolev inequality is stronger than the Poincaré inequality and induces the convergence in KL divergence. See [19] for details. It was proved in [2,18] that our dynamics, LD, SLD, PLD, S-PLD, and skew-SGLD satisfy the logarithmic Sobolev inequalities under our assumptions. We express the constant of the logarithmic Sobolev inequality for skew-SGLD as $\lambda(\alpha, N)$. This constant depends on the skew matrices and the Poincaré constant. We estimate this constant in Appendix M.1.

To upper-bound the bias, here we control the KL divergence. We denote the law of skew-SGLD at iteration k with interaction strength α as $\mu_{kh}^{\otimes N}(\alpha)$. We upper-bound the bias by 2-Wasserstein distance

$$\left|\mathbb{E}\frac{1}{N}\sum_{n=1}^{N} f(X_k^{(n)}) - \int_{\mathbb{R}^d} f d\pi\right|, \leq \frac{L_f}{\sqrt{N}} W_2(\mu_{kh}^{\otimes N}(\alpha), \pi^{\otimes N}). \tag{A150}$$

Then, from the transportation inequality [19],

$$W_2(\mu_k, \pi) \leq \sqrt{\frac{2}{\lambda(\alpha, N)}\mathrm{KL}(\mu_{kh}^{\otimes N}(\alpha)|\pi^{\otimes N})}. \tag{A151}$$

Thus, we will upper bound the KL divergence using the technique in [23]. However, in the original proof, a full gradient ∇U is used so we replace it with the stochastic gradient. Moreover, we introduce the skew interaction term.

First, Lemma 11 in [23] is modified to

$$\mathbb{E}_{\pi^{\otimes N}}\|\nabla U^{\otimes N}\|^2 \leq \frac{dNM}{\beta}. \tag{A152}$$

Then Lemma 12 in [23] is modified to

$$\mathbb{E}_{\mu}\|\nabla U^{\otimes N}\|^2 \leq 4M^2\lambda \mathrm{KL}(\mu|\pi^{\otimes N}) + \frac{2dNM}{\beta}, \tag{A153}$$

for any integrable μ.

Herein after, we drop $\otimes N$ from $X^{\otimes N}$, $\nabla U^{\otimes N}$, and $\nabla \hat{U}^{\otimes N}$ for notational simplicity. We focus on skew-SGLD at iteration k, we consider the following SDE for $t \in (kh, (k+1)h]$

$$dX_t = -(I + \alpha J)\nabla \tilde{U}(X_k)dt + \sqrt{2\beta^{-1}}dw_t, \tag{A154}$$

where $\nabla \tilde{U}(X_k)$ is the stochastic gradient conditioned on X_k. The solution of this SDE is

$$X_{(k+1)} = X_k - (I + \alpha J)\nabla \tilde{U}(X_k)h + \sqrt{2\beta^{-1}}\epsilon. \tag{A155}$$

We would like to derive the continuity equation correspond to Equation (A154). Following [23], we express X_t as x_t and X_k as x_0 for simplicity. Let $\rho_{0t}(x_0, x_t)$ denote the joint distribution of (x_0, x_t). Then, the conditional and marginal relations are written as

$$\rho_{0t}(x_0, x_t) = \rho_0(x_0)\rho_{t|0}(x_t|x_0) = \rho_t(x_t)\rho_{0|t}(x_0|x_t). \tag{A156}$$

The conditional density $\rho_{t|0}(x_t|x_0)$ follows the FP equation

$$\frac{\partial \rho_{t|0}(x_t|x_0)}{\partial t} = \nabla \cdot (\rho_{t|0}(x_t|x_0)(I + \alpha J)\nabla \tilde{U}(x_0)) + \beta^{-1}\Delta \rho_{t|0}(x_t|x_0), \tag{A157}$$

Then following [23], to derive the evolution of ρ_t, we take the expectation over $\rho_0(x_0)$

$$\begin{aligned}\frac{\partial \rho_t(x)}{\partial t} &= \int_{\mathbb{R}^d} \frac{\partial \rho_{t|0}(x_t|x_0)}{\partial t}\rho_0(x_0)dx_0 \\ &= \nabla \cdot (\rho_t(x_t)\mathbb{E}_{\rho_{0|t}}[(I + \alpha J)\nabla \tilde{U}(x_0)|x_t = x]) + \beta^{-1}\Delta \rho_t(x).\end{aligned} \tag{A158}$$

Then, we take the expectation regarding for the stochastic gradient in the above equation and include it into $\mathbb{E}_{\rho_{0|t}}$ for notational simplicity. Then following the discussion of Lemma 3 in [23], we obtain

$$\begin{aligned}\frac{\partial \mathrm{KL}(\mu_t|\pi)}{\partial t} \leq &-\frac{3}{4}I(\mu_t^{\otimes N}|\pi^{\otimes N}) + 2\mathbb{E}_{\rho_{0t}}[\|\nabla U(X_t) - \nabla U(X_0)\|^2] \\ &+ 2(1+\alpha)^2\mathbb{E}_{\rho_{0t}}[\|\nabla U(X_0) - \nabla \tilde{U}(X_0)\|^2] + 2\alpha^2\mathbb{E}_{\rho_{0t}}[\nabla U(x_0)\|^2],\end{aligned} \tag{A159}$$

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

$$X_t = X_k - t(I + \alpha J)\nabla U(X_k) + \sqrt{2t\beta^{-1}}\epsilon. \tag{A160}$$

Then, from [18], we can upper-bound the second term by

$$\mathbb{E}_{\rho_{0t}}[\|\nabla U(X_0) - \nabla \tilde{U}(X_0)\|^2] \leq NC_0'\delta, \tag{A161}$$

$$C_0' := 2\left(M^2\left(\kappa_0 + 2\left(1 \vee \frac{1}{m}\right)\left(b + 2(1+\alpha)^2B^2 + \frac{d}{\beta}\right)\right) + B^2\right) \tag{A162}$$

and the third term is upper-bounded by

$$\begin{aligned}\mathbb{E}_{\rho_{0t}}[\|\nabla U(X_0) - \nabla \mathbb{E}_{\rho_{0t}}[\nabla U(x_0)\|^2] &\leq 2M^2\|x_0\|^2 + 2NB^2 \\ &\leq NC_0',\end{aligned} \tag{A163}$$

where we used lemma 2 and 7 in [2]. Finally, from the original proof of [23] we obtain

$$2\mathbb{E}_{\rho_{0t}}[\|\nabla U(X_t) - \nabla U(X_0)\|^2] \leq 8t^2M^4\lambda \mathrm{KL}(\mu_k^{\otimes N}|\pi^{\otimes N}) + \frac{4t^2dNM^3}{\beta} + \frac{4tdNM^2}{\beta}. \tag{A164}$$

Then, in conclusion, under $h \in (0, 1 \wedge \frac{m}{4M^2})$ obeying $kh \geq 1$ and $\beta m \geq 2$, we obtain

$$\begin{aligned}\frac{d}{dt}\mathrm{KL}(\mu_t^{\otimes N}|\pi^{\otimes N}) \leq -\frac{3}{4}I(\mu_t^{\otimes N}|\pi^{\otimes N}) + 8t^2M^4\lambda(\alpha, N)\mathrm{KL}(\mu_k^{\otimes N}|\pi^{\otimes N}) \\ + \frac{4t^2dNM^3}{\beta} + \frac{4tdNM^2}{\beta} + 2NC_0'(\delta(1+\alpha)^2 + \alpha^2).\end{aligned} \tag{A165}$$

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

$$\begin{aligned}\frac{d}{dt}\mathrm{KL}(\mu_t^{\otimes N}|\pi^{\otimes N}) \leq -\frac{3}{4}I(\mu_t^{\otimes N}|\pi^{\otimes N}) + 8t^2M^4\lambda(\alpha, N)\mathrm{KL}(\mu_k^{\otimes N}|\pi^{\otimes N}) \\ + \frac{t^2dNM}{\beta}(m+4M) + 2NC_0'(\delta(1+\alpha)^2 + \alpha^2).\end{aligned} \tag{A166}$$

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

$$\begin{aligned}\mathrm{KL}(\mu_{k+1}^{\otimes N}|\pi^{\otimes N}) \leq& e^{-\frac{3}{2}\lambda(\alpha,N)h}\Big(1 + 16h^3M^4\lambda\Big)\mathrm{KL}(\mu_k^{\otimes N}|\pi^{\otimes N}) \\ &+ e^{-\frac{3}{2}\lambda(\alpha,N)h}\left(\frac{2hdNM}{\beta}(m+4M) + 8hNC_0'(\delta(1+\alpha)^2 + \alpha^2)\right).\end{aligned} \tag{A167}$$

If $h \in (0, \frac{\lambda(\alpha,N)}{4\sqrt{2}M^2})$, we obtain

$$\mathrm{KL}(\mu_{k+1}^{\otimes N}|\pi^{\otimes N}) \leq e^{-\lambda(\alpha,N)h}\mathrm{KL}(\mu_k^{\otimes N}|\pi^{\otimes N}) + \frac{2h^2dNM}{\beta}(m+4M) + 8hNC_0'(\delta(1+\alpha)^2 + \alpha^2). \tag{A168}$$

From this one step inequality, we obtain

$$\begin{aligned}&\mathrm{KL}(\mu_k^{\otimes N}|\pi^{\otimes N}) \\ &\leq e^{-\lambda(\alpha,N)kh}\mathrm{KL}(\mu_0^{\otimes N}|\pi^{\otimes N}) + \frac{1}{1-e^{-\lambda(\alpha,N)h}}\left(\frac{2h^2dNM}{\beta}(m+4M) + 8hNC_0'(\delta(1+\alpha)^2 + \alpha^2)\right) \\ &\leq e^{-\lambda(\alpha,N)kh}\mathrm{KL}(\mu_0^{\otimes N}|\pi^{\otimes N}) + \frac{2N}{\lambda(\alpha,N)}\left(\frac{hdM}{\beta}(m+4M) + 4C_0'(\delta(1+\alpha)^2 + \alpha^2)\right).\end{aligned} \tag{A169}$$

Then, finally we obtain

$$\begin{aligned}&\left|\mathbb{E}\frac{1}{N}\sum_{n=1}^{N} f(X_k^{(n)}) - \int_{\mathbb{R}^d} f d\pi\right| \\ &\leq \frac{L_f}{\sqrt{N}}\sqrt{\frac{2}{\lambda(\alpha,N)}\mathrm{KL}(\mu_{kh}^{\otimes N}(\alpha)|\pi^{\otimes N})} \\ &\leq L_f\sqrt{\frac{2}{\lambda(\alpha,N)}}\sqrt{e^{-\lambda(\alpha,N)kh}\mathrm{KL}(\mu_0|\pi) + \frac{2}{\lambda(\alpha,N)}\left(\frac{hdM}{\beta}(m+4M) + 4C_0'(\delta(1+\alpha)^2 + \alpha^2)\right)} \\ &\leq L_f\sqrt{\frac{2}{\lambda(\alpha,N)}}\sqrt{e^{-\lambda(\alpha,N)kh}\mathrm{KL}(\mu_0|\pi) + \frac{C_3(\alpha)}{\lambda(\alpha,N)}},\end{aligned} \tag{A170}$$

where

$$C_3(\alpha) := 2\frac{hdM}{\beta}(m+4M) + 8C_0'(\delta(1+\alpha)^2+\alpha^2), \tag{A171}$$

$$C_0' := 2\left(M^2\left(\kappa_0 + 2\left(1\vee\frac{1}{m}\right)\left(b + 2(1+\alpha)^2B^2 + \frac{d}{\beta}\right)\right) + B^2\right). \tag{A172}$$

Moreover, from Appendix M.1, the logarithmic Sobolev constant is

$$\lambda(\alpha, N) := \left(\frac{1}{(1+\beta m(\alpha,N)^{-1}|C(m_0)|)2\pi e^2} + \frac{3}{2m(\alpha,N)}\right), \tag{A173}$$

where

$$-C(m_0) := \mathbb{E}_{\pi^{\otimes N}}[\|\nabla U^{\otimes N}(x)\|]^{1/2} + \sqrt{\frac{8}{m_0}\mathbb{E}_{\pi^{\otimes N}}[\|\nabla U^{\otimes N}(x)\|^2]^{1/2}}. \tag{A174}$$

Appendix M.1. Estimation of the Logarithmic Sobolev Constant

In this section, we estimate the logarithmic Sobolev constants using the technique of restricted logarithmic Sobolev inequality, which was introduced in [49].

The technique of [49] estimates the constant of the logarithmic Sobolev inequality as follows. Assume that π on $\mathbb{R}^d$ with $\mathcal{L}$ satisfies the Poincaré inequality with constant m. Then, for any function u on $\mathbb{R}^d$ that satisfies

$$\int_{\mathbb{R}^d} u d\pi = 0 \quad \text{and} \quad \int_{\mathbb{R}^d} u^2 d\pi = 1, \tag{A175}$$

we find a constant b that satisfies

$$\int_{\mathbb{R}^d} u^2 \ln u^2 d\pi \le b\int_{\mathbb{R}^d} -u\mathcal{L}u d\pi. \tag{A176}$$

Then the logarithmic constant is larger than $2(b+\frac{3}{m})^{-1}$. Thus, we only need to focus on the restricted function class to estimate a constant b. We slightly change the Lemma 3.2 of [49] that estimate the constant b in Equation (A176) to apply it in our setting. In Lemma 3.2 of [49], it was proved that if u on $\mathbb{R}^d$ satisfies the conditions in Equation (A175), then for any $t \in (0,1)$, we have

$$\int_{\mathbb{R}^d} -u\mathcal{L}u d\pi - t\pi e^2\int_{\mathbb{R}^d} u^2\ln u^2 d\pi \ge (1-t)m + t\beta\int_{\mathbb{R}^d}(-\frac{1}{2}\mathcal{L}U(x) - \pi e^2 U(x))u^2 d\pi, \tag{A177}$$

where we assume that $\pi \propto e^{-\beta U(x)}$ satisfies the Poincaré inequality with constant m. If there exists a constant C such that

$$-C \ge \beta\int_{\mathbb{R}^d}(-\frac{1}{2}\mathcal{L}U(x) - \pi e^2 U(x))u^2 d\pi > -\infty, \tag{A178}$$

then by setting $t = m/(m+|C|)$, we can show that

$$\int_{\mathbb{R}^d} -u\mathcal{L}u d\pi - m/(m+|C|)\pi e^2\int_{\mathbb{R}^d} u^2\ln u^2 d\pi > 0. \tag{A179}$$

Thus, the constant b in Equation (A176) is $b = t = m/(m+|C|)$ and the logarithmic constant is $2(m/(m+|C|)+\frac{3}{m})^{-1}$.

Thus, We analyze the constant C. The first term of the integral in Equation (A178) is lower-bounded bounded by

$$-\mathbb{E}_\pi[\mathcal{L}U(x)u^2] \ge -|\mathbb{E}_\pi[U(x)\mathcal{L}U(x)]|^{1/2}|\mathbb{E}_\pi[u^2\mathcal{L}u^2]|^{1/2} \ge -2\mathbb{E}_\pi[\|\nabla U(x)\|^2]^{1/2}, \tag{A180}$$

where we used the property of $\mathcal{L}$, see [19] for details. As for the second term, it is lower-bounded by

$$-|\mathbb{E}_\pi[U(x)u^2] \geq -\sqrt{|\mathbb{E}_\pi[U^2(x)u^2]} \geq -\sqrt{\frac{1}{m}|\mathbb{E}_\pi[(U(x)|u|)\mathcal{L}(U(x)|u|)]|} \geq -\sqrt{\frac{8}{m}\mathbb{E}_\pi[\|\nabla U(x)\|^2]^{1/2}}. \tag{A181}$$

Thus, by setting

$$-C := \mathbb{E}_\pi[\|\nabla U(x)\|]^{1/2} + \sqrt{\frac{8}{m_0}\mathbb{E}_\pi[\|\nabla U(x)\|^2]^{1/2}}, \tag{A182}$$

we can estimate the logarithmic constant as $2(m/(m+|C|)+\frac{3}{m})^{-1}$.

In our setting, this is modified to

$$\lambda(\alpha, N) = \left(\frac{1}{(1+\beta m(\alpha,N)^{-1}|C(m_0)|)2\pi e^2} + \frac{3}{2m(\alpha,N)}\right)^{-1}. \tag{A183}$$

where

$$-C(m_0) := \mathbb{E}_{\pi^{\otimes N}}[\|\nabla U^{\otimes N}(x)\|]^{1/2} + \sqrt{\frac{8}{m_0}\mathbb{E}_{\pi^{\otimes N}}[\|\nabla U^{\otimes N}(x)\|^2]^{1/2}}. \tag{A184}$$

Finally, if we increase $m(\alpha, N)$, $\lambda(\alpha, N)$ increases. Thus, since $m(\alpha, N) \geq m(\alpha = 0, N)$, we obtain $\lambda(\alpha, N) \geq \lambda(\alpha = 0, N)$.

Appendix M.2. Computational Complexity

To derive the computational complexity, for simplicity, we assume that $\delta \leq h$ and We also set $\alpha^2 \leq h$ for simplicity. This means that the variance of the stochastic gradient is small enough and we use small α. Then the bias is

$$\left|\mathbb{E}\frac{1}{N}\sum_{n=1}^N f(X_k^{(n)}) - \int_{\mathbb{R}^d} f d\pi\right| \leq L_f\sqrt{\frac{2}{\lambda(\alpha,N)}}\sqrt{e^{-\lambda(\alpha,N)kh}\mathrm{KL}(\mu_0|\pi) + \frac{C_3(\alpha)}{\lambda(\alpha,N)}} \leq L_f\sqrt{\frac{2}{\lambda(\alpha,N)}}\left(\sqrt{e^{-\lambda(\alpha,N)kh}\mathrm{KL}(\mu_0|\pi)} + \sqrt{\frac{C_3(\alpha)}{\lambda(\alpha,N)}}\right), \tag{A185}$$

where

$$C_3(\alpha) := h\left(2\frac{dM}{\beta}(m+4M) + 8C_0'((1+h^{1/2})^2+1)\right), \tag{A186}$$

$$C_0' := 2\left(M^2\left(\kappa_0 + 2\left(1\vee\frac{1}{m}\right)\left(b + 2(1+h^{1/2})^2B^2 + \frac{d}{\beta}\right)\right) + B^2\right). \tag{A187}$$

Then we define

$$C_3' := 2\frac{dM}{\beta}(m+4M) + 8C_0'((1+h^{1/2})^2+1), \tag{A188}$$

and use the step size that satisfies $h = \frac{\lambda(\alpha,N)\xi}{2\sqrt{2}C_3'L_f}$. Then when we use

$$k \geq \frac{2}{\lambda(\alpha,N)h}\ln\frac{L_f}{\xi}\sqrt{\frac{\mathrm{KL}(\mu_0|\pi)}{2\lambda(\alpha,N)}}, \tag{A189}$$

we have

$$\left| \mathbb{E} \frac{1}{N} \sum_{n=1}^{N} f(X_k^{(n)}) - \int_{\mathbb{R}^d} f d\pi \right| \leq \frac{\xi}{2} + \frac{\xi}{2} \leq \xi. \tag{A190}$$

References

1. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; MIT Press: Cambridge, MA, USA, 2012.
2. Raginsky, M.; Rakhlin, A.; Telgarsky, M. Non-convex learning via Stochastic Gradient Langevin Dynamics: A nonasymptotic analysis. In Proceedings of the Conference on Learning Theory, Amsterdam, The Netherlands, 7–10 July 2017; pp. 1674–1703.
3. Welling, M.; Teh, Y.W. Bayesian learning via stochastic gradient Langevin dynamics. In Proceedings of the International Conference on Machine Learning, Washington, DC, USA, 28 June–2 July 2011; pp. 681–688.
4. Livingstone, S.; Girolami, M. Information-Geometric Markov Chain Monte Carlo Methods Using Diffusions. *Entropy* **2014**, *16*, 3074–3102, doi:10.3390/e16063074.
5. Hartmann, C.; Richter, L.; Schütte, C.; Zhang, W. Variational Characterization of Free Energy: Theory and Algorithms. *Entropy* **2017**, *19*, 626, doi:10.3390/e19110626.
6. Neal, R.M. Improving asymptotic variance of MCMC estimators: Non-reversible chains are better. *arXiv* **2004**, arXiv:math/0407281.
7. Neklyudov, K.; Welling, M.; Egorov, E.; Vetrov, D. Involutive mcmc: a unifying framework. In Proceedings of the International Conference on Machine Learning, Vienna, Austria, 13–18 July 2020; pp. 7273–7282.
8. Gao, X.; Gurbuzbalaban, M.; Zhu, L. Breaking Reversibility Accelerates Langevin Dynamics for Non-Convex Optimization. In Proceedings of the Advances in Neural Information Processing Systems, Online, 6–12 December 2020; pp. 17850–17862.
9. Eberle, A.; Guillin, A.; Zimmer, R.; others. Couplings and quantitative contraction rates for Langevin dynamics. *Ann. Probab.* **2019**, *47*, 1982–2010.
10. Gao, X.; Gürbüzbalaban, M.; Zhu, L. Global convergence of stochastic gradient Hamiltonian Monte Carlo for non-convex stochastic optimization: Non-asymptotic performance bounds and momentum-based acceleration. *arXiv* **2018**, arXiv:1809.04618.
11. Cheng, X.; Chatterji, N.S.; Abbasi-Yadkori, Y.; Bartlett, P.L.; Jordan, M.I. Sharp convergence rates for Langevin dynamics in the nonconvex setting. *arXiv* **2018**, arXiv:1805.01648.
12. Chen, T.; Fox, E.; Guestrin, C. Stochastic gradient hamiltonian monte carlo. In Proceedings of the International conference on machine learning, Beijing, China, 21–26 June 2014; pp. 1683–1691.
13. Hwang, C.R.; Hwang-Ma, S.Y.; Sheu, S.J. Accelerating gaussian diffusions. *Ann. Appl. Probab.* **1993**, 3, 897–913.
14. Hwang, C.R.; Hwang-Ma, S.Y.; Sheu, S.J. Accelerating diffusions. *Ann. Appl. Probab.* **2005**, *15*, 1433–1444.
15. Hwang, C.R.; Normand, R.; Wu, S.J. Variance reduction for diffusions. *Stoch. Process. Their Appl.* **2015**, *125*, 3522–3540.
16. Duncan, A.B.; Lelièvre, T.; Pavliotis, G.A. Variance Reduction Using Nonreversible Langevin Samplers. *J. Stat. Phys.* **2016**, *163*, 457–491, doi:10.1007/s10955-016-1491-2.
17. Duncan, A.B.; Nüsken, N.; Pavliotis, G.A. Using Perturbed Underdamped Langevin Dynamics to Efficiently Sample from Probability Distributions. *J. Stat. Phys.* **2017**, *169*, 1098–1131, doi:10.1007/s10955-017-1906-8.
18. Futami, F.; Sato, I.; Sugiyama, M. Accelerating the diffusion-based ensemble sampling by non-reversible dynamics. In Proceedings of the International Conference on Machine Learning, Vienna, Austria, 13–18 July 2020; pp. 3337–3347.
19. Bakry, D.; Gentil, I.; Ledoux, M. *Analysis and Geometry of Markov Diffusion Operators*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013; Volume 348,.
20. Roussel, J.; Stoltz, G. Spectral methods for Langevin dynamics and associated error estimates. *ESAIM Math. Model. Numer. Anal.* **2018**, *52*, 1051–1083.
21. Menz, G.; Schlichting, A. Poincaré and logarithmic Sobolev inequalities by decomposition of the energy landscape. *Ann. Probab.* **2014**, *42*, 1809–1884.
22. Liu, Q.; Lee, J.; Jordan, M. A kernelized Stein discrepancy for goodness-of-fit tests. In Proceedings of the International Conference on Machine Learning, New York , NY, USA, 24–26 June 2016; pp. 276–284.
23. Vempala, S.; Wibisono, A. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 8094–8106.
24. Lelièvre, T.; Nier, F.; Pavliotis, G.A. Optimal non-reversible linear drift for the convergence to equilibrium of a diffusion. *J. Stat. Phys.* **2013**, *152*, 237–274.
25. Tripuraneni, N.; Rowland, M.; Ghahramani, Z.; Turner, R. Magnetic Hamiltonian Monte Carlo. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 3453–3461.
26. Nusken, N.; Pavliotis, G. Constructing sampling schemes via coupling: Markov semigroups and optimal transport. *SIAM/ASA J. Uncertain. Quantif.* **2019**, *7*, 324–382.
27. Liu, Q.; Wang, D. Stein variational gradient descent: A general purpose bayesian inference algorithm. In Proceedings of the Advances In Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 2378–2386.
28. Zhang, J.; Zhang, R.; Chen, C. Stochastic particle-optimization sampling and the non-asymptotic convergence theory. *arXiv* **2018**, arXiv:1809.01293.
29. Wang, Y.; Li, W. Information Newton's flow: second-order optimization method in probability space. *arXiv* **2020**, arXiv:2001.04341.

30. Wibisono, A. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In Proceedings of the Conference On Learning Theory, Stockholm, Sweden, 6–9 July 2018; pp. 2093–3027.
31. Gretton, A.; Borgwardt, K.M.; Rasch, M.J.; Schölkopf, B.; Smola, A. A kernel two-sample test. *J. Mach. Learn. Res.* **2012**, *13*, 723–773.
32. Ding, N.; Fang, Y.; Babbush, R.; Chen, C.; Skeel, R.D.; Neven, H. Bayesian sampling using stochastic gradient thermostats. In Proceedings of the Advances in neural information processing systems, Montreal, QC, Canada, 8–11 December 2014; pp. 3203–3211.
33. Patterson, S.; Teh, Y.W. Stochastic gradient Riemannian Langevin dynamics on the probability simplex. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–8 December 2013; pp. 3102–3110.
34. Dua, D.; Graff, C. *UCI Machine Learning Repository*. Available online: http://archive.ics.uci.edu/ml (accessed on 21 July 2021).
35. Villani, C. Optimal transportation, dissipative PDE's and functional inequalities. In *Optimal Transportation and Applications*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 53–89.
36. Bakry, D.; Barthe, F.; Cattiaux, P.; Guillin, A. A simple proof of the Poincaré inequality for a large class of probability measures including the log-concave case. *Electron. Commun. Probab* **2008**, *13*, 21.
37. Nelson, E. *Dynamical Theories of Brownian Motion*; Princeton University Press: 1967; Volume 3.
38. Pavliotis, G.A. *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*; Springer: Berlin/Heidelberg, Germany, 2014; Volume 60.
39. Franke, B.; Hwang, C.R.; Pai, H.M.; Sheu, S.J. The behavior of the spectral gap under growing drift. *Trans. Am. Math. Soc.* **2010**, *362*, 1325–1350.
40. Landim, C.; Seo, I. Metastability of Nonreversible Random Walks in a Potential Field and the Eyring-Kramers Transition Rate Formula. *Commun. Pure Appl. Math.* **2018**, *71*, 203–266.
41. Landim, C.; Mariani, M.; Seo, I. Dirichlet's and Thomson's principles for non-selfadjoint elliptic operators with application to non-reversible metastable diffusion processes. *Arch. Ration. Mech. Anal.* **2019**, *231*, 887–938.
42. Golub, G.H.; Van Loan, C.F. *Matrix Computations*; JHU Press: Baltimore, MD, USA, 2012; Volume 3, .
43. Okamoto, M. Distinctness of the Eigenvalues of a Quadratic form in a Multivariate Sample. *Ann. Statist.* **1973**, *1*, 763–765, doi:10.1214/aos/1176342472.
44. Petersen, K.B.; Pedersen, M.S. *The Matrix Cookbook*. Available online: http://www2.compute.dtu.dk/pubdb/pubs/3274-full.html (accessed on 21 July 2021).
45. Van Erven, T.; Harremos, P. Rényi divergence and Kullback-Leibler divergence. *IEEE Trans. Inf. Theory* **2014**, *60*, 3797–3820.
46. Chewi, S.; Le Gouic, T.; Lu, C.; Maunu, T.; Rigollet, P.; Stromme, A. Exponential ergodicity of mirror-Langevin diffusions. In Proceedings of the Advances in Neural Information Processing Systems, Online, 6–12 December 2020; 2020; pp. 19573–19585.
47. Bolley, F.; Villani, C. Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. In *Annales de la Faculté des Sciences de Toulouse: Mathématiques*; Université Paul Sabatier: Toulouse, France, 2005; Volume 14, pp. 331–352.
48. Donsker, M.D.; Varadhan, S.S. Asymptotic evaluation of certain Markov process expectations for large time. IV. *Commun. Pure Appl. Math.* **1983**, *36*, 183–212.
49. Carlen, E.; Loss, M. Logarithmic Sobolev inequalities and spectral gaps. *Contemporary Mathematics.* **2004**, *353*, 53–60.