Laplacian Smoothing Stochastic Gradient Markov Chain Monte Carlo

Bao Wang[*]
Department of Mathematics
University of California, Los Angeles
wangbaonj@gmail.com

Difan Zou[*]
Department of Computer Science
University of California, Los Angeles
knowzou@cs.ucla.edu

Quanquan Gu[†]
Department of Computer Science
University of California, Los Angeles
qgu@cs.ucla.edu

Stanley J. Osher[†]
Department of Mathematics
University of California, Los Angeles
sjo@math.ucla.edu

Abstract

As an important Markov Chain Monte Carlo (MCMC) method, stochastic gradient Langevin dynamics (SGLD) algorithm has achieved great success in Bayesian learning and posterior sampling. However, SGLD typically suffers from slow convergence rate due to its large variance caused by the stochastic gradient. In order to alleviate these drawbacks, we leverage the recently developed Laplacian Smoothing (LS) technique and propose a Laplacian smoothing stochastic gradient Langevin dynamics (LS-SGLD) algorithm. We prove that for sampling from both log-concave and non-log-concave densities, LS-SGLD achieves strictly smaller discretization error in 2-Wasserstein distance, although its mixing rate can be slightly slower. Experiments on both synthetic and real datasets verify our theoretical results, and demonstrate the superior performance of LS-SGLD on different machine learning tasks including posterior sampling, Bayesian logistic regression and training Bayesian convolutional neural networks. The code is available at https://github.com/BaoWangMath/LS-MCMC.

1 Introduction

Given a dataset $\mathcal{D} = \{\boldsymbol{d}_i\}_{i=1}^n$, the posterior of a machine learning (ML) model's parameters $\boldsymbol{x} \in \mathbb{R}^d$ with prior $p(\boldsymbol{x})$ and likelihood $\Pi_{i=1}^n p(\boldsymbol{d}_i|\boldsymbol{x})$ is computed as $p(\boldsymbol{x}|\mathcal{D}) \propto p(\boldsymbol{x})\Pi_{i=1}^n p(\boldsymbol{d}_i|\boldsymbol{x})$. Optimization algorithms are used to find the maximum a posterior (MAP) estimate, $\boldsymbol{x}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{x}} \log p(\boldsymbol{x}|\mathcal{D})$. Sampling algorithms such as Langevin Dynamics (LD) are used to sample the posterior or the log posterior. In this paper, we consider applying LD-based MCMC algorithms to sample $e^{-f(\boldsymbol{x})}$, where

$$f(\boldsymbol{x}) := \frac{1}{n}\sum_{i=1}^n f_i(\boldsymbol{x}) = -\frac{1}{n}\sum_{i=1}^n \log p(\boldsymbol{d}_i|\boldsymbol{x}). \tag{1}$$

Here, we normalized the log-likelihood by a factor n for the ease of presentation in the remaining part of this paper.

The first-order LD reads:

$$d\boldsymbol{X}_t = -\nabla f(\boldsymbol{X}_t)\mathrm{d}t + \sqrt{2\beta^{-1}} \cdot \mathrm{d}\boldsymbol{B}_t, \tag{2}$$

---

[*]Equal Contribution
[†]Co-Corresponding Authors

where $\boldsymbol{X}_t \in \mathbb{R}^d$ denotes the point at time t, β denotes the inverse temperature and $\boldsymbol{B}_t \in \mathbb{R}^d$ is the standard Brownian term. Under certain assumptions on the negative log posterior (i.e., $f(\boldsymbol{x})$), the LD (2) converges to an unique invariant distribution $\pi \propto e^{-\beta f(\boldsymbol{x})}$ (Chiang et al., 1987). Therefore, one can apply numerical integrator to approximate (2) in order to obtain samples that follow the posterior distribution. One simple integrator is to apply the Euler-Maruyama discretization (abd E. Platen, 1992) to (2), which gives:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta \nabla f(\boldsymbol{x}_k) + \sqrt{2\beta^{-1}\eta} \cdot \boldsymbol{\epsilon}_k, \tag{3}$$

and it is known as the Langevin Monte Carlo (LMC), (a.k.a., unadjusted Langevin algorithm (Parisi, 1981)). When the target density, i.e., posterior distribution, is strongly log-concave and log-smooth, Dalalyan (2017); Durmus et al. (2017) proved that LMC is able to converge to the target density up to an arbitrarily small sampling error in both total variation and 2-Wasserstein distances. Furthermore, the convergence guarantee of LMC for sampling from non-log-concave distributions has also been established in Raginsky et al. (2017); Xu et al. (2018).

Note that the posterior distribution is defined on the whole dataset $\mathcal{D}$, which is typically extremely large in modern ML tasks. Therefore, computing the full gradient $\nabla f(\boldsymbol{x})$ is inefficient and may dramatically slow down the convergence of sampling algorithms. One solution is to replace the full gradient in (3) with a subsampled one, which gives rise to Stochastic Gradient Langevin Dynamics (SGLD) (Welling and Whye, 2011). From the theoretical perspective, the convergence guarantee of SGLD has been proved for both strongly log-concave distributions (Dalalyan and Karagulyan, 2017) and non-log-concave distributions (Raginsky et al., 2017; Xu et al., 2018) in 2-Wasserstein distance. Mou et al. (2017) further studied the generalization performance of SGLD for nonconvex optimization. Although SGLD can drastically reduce the computational cost, it is also observed to have a slow convergence rate due to the large variance caused by the stochastic gradient (Teh et al., 2016; Vollmer et al., 2016). In order to reduce the variance of stochastic gradient as well as to improve the convergence rate, Dubey et al. (2016) incorporated variance reduction techniques into SGLD, which gives rise to a family of variance-reduced LD-based algorithms such as SVRG-LD and SAGA-LD. Chatterji et al. (2018); Zou et al. (2018b, 2019a) further proved that SVRG-LD and SAGA-LD are able to converge to the target density with fewer stochastic gradient evaluations than SGLD and LMC in certain regimes. However, both SVRG-LD and SAGA-LD require a large amount of extra computation and memory costs and can only be shown to achieve faster convergence on small to moderate datasets. Therefore, it is natural to ask *if we can reduce the variance of stochastic gradients while maintaining similar computation and memory costs of SGLD?*

Recently, Osher et al. (2018) integrated Laplacian smoothing and related high-order smoothing techniques into Stochastic Gradient Descent (SGD) to reduce the variance of stochastic gradient on-the-fly. Laplacian Smoothing SGD (LSSGD) allows us to take a significantly larger step size than vanilla SGD and reduces the optimality gap in convex optimization when constant step size is used. Empirically, LSSGD preconditions the gradient when the objective function has a large condition number and can avoid local minima. Because of this, LSSGD is applicable to train a large number of deep learning models with good generalizability. Laplacian smoothing also demonstrates some ability to avoid saddle point in gradient descent (Kreusser et al., 2019). Most recently, Wang et al. (2019) leveraged Laplacian smoothing to improve the utility of machine learning models trained with privacy guarantee.

In this paper, we integrate Laplacian smoothing with SGLD, and we call the resulting algorithm Laplacian Smoothing SGLD (LS-SGLD). The extra computation of LS-SGLD compared with SGLD is that we need to compute the products of the inverse of two circulant matrices with vectors. We leverage the Fast Fourier Transform to develop fast algorithms to compute these matrix-vector products efficiently, and the resulting algorithms can compute the matrix-vector products with a negligible overhead in both time and memory.

1.1 Our Contributions

We summarize the main contributions of this work as follows:

- We propose a simple modification on the SGLD, which applies the Laplacian smoothing matrix and its squared root to the stochastic gradient and Gaussian noise vectors, respectively. The continuous and full-gradient counter-part of the modified LS-SGLD has the same stationary distribution as the LD.

- We proposed FFT-based fast algorithms to compute the product of the inverse of circulant matrices with any given vector. By leveraging the structure of eigenvalues and eigenvectors of the circulant matrices, we can compute these products very efficiently with a negligible overhead in both time and memory.

- We prove the convergence rate of LS-SGLD for sampling from both log-concave and non-log-concave densities in 2-Wasserstein distance. Specifically, we decompose the sampling error into the discretization error and the ergodicity rate. Moreover, we show that there exists a trade-off between the discretization error and ergodicity rate of LS-SGLD, as adding Laplacian smoothing can reduce the discretization error but slow down the mixing time.

- We conduct extensive experiments to evaluate the performance of LS-SGLD. First, we show that compared with SGLD, LS-SGLD can achieve a significantly smaller discretization error but similar ergodicity rate, which implies that the overall sampling error of LS-SGLD can be much smaller. Second, we conduct experiments on both synthetic and real data for posterior sampling, Bayesian logistic regression and training Bayesian convolutional networks, all of which demonstrate the superior performance of LS-SGLD.

1.2 Additional Related Work

In addition to the first-order Langevin based algorithms we discussed in the introduction, there also emerges a vast body of work focusing on higher-order Langevin based algorithms. One of the well-know high-order MCMC method is Hamiltonian Monte Carlo (HMC) (Neal et al., 2011), which incorporates an Hamiltonian momentum term into the first-order MCMC method in order to improve the mixing time. Similar to SGLD, a stochastic version of HMC (namely SGHMC) has been further established in Chen et al. (2014), which was shown to be able to achieve a faster convergence rate than SGLD in experiments. Ma et al. (2015) investigated a family of SGHMC methods and proposed a new state-adaptive sampler on the Riemannian manifold. Chen et al. (2015) provided theoretical convergence guarantees of SGHMC in terms of mean square error (MSE) and proposed a 2nd-order symmetric splitting integrator to further improve the discretization error. When the target density is strongly log-concave and log-smooth, Cheng et al. (2018b) proposed underdamped MCMC (U-MCMC) and stochastic underdamped MCMC (SG-U-MCMC), and obtained convergence rates in 2-Wasserstein distance. The convergence rates of these two algorithms have been further established for sampling from non-log-concave densities (Cheng et al., 2018a). However, due to the large variance of stochastic gradients and lacking of the Metropolis Hasting (MH) correction step, SGHMC has also been observed to have highly biased sampling trajectory (Betancourt, 2015; Dang et al., 2019). One way to address this issue is to make use of a variance-reduction technique to alleviate the variance of stochastic gradients in SGHMC, which gave rise to stochastic variance-reduced HMC methods (Zou et al., 2018a; Li et al., 2018; Zou et al., 2019b).

1.3 Organization

We organize this paper as follows: We present LS-SGLD and derive FFT-based fast algorithms for LS-SGLD in Section 2. In Section 3, we give theoretical guarantees for the performance of LS-SGLD in both log-concave and non-log-concave settings. In Section 4, we numerically verify the performance of LS-SGLD on sampling different distributions, training Bayesian logistic regression, and convolutional neural nets. We conclude this work in Section 5.

1.4 Notations

Throughout this paper we use bold upper-case letters $\mathbf{A}$, $\mathbf{B}$ to denote matrices, bold lower-case letters $\boldsymbol{x}$, $\boldsymbol{y}$ to denote vectors, and lower cases letters x, y and α, β to denote scalars. For continuous-time random vectors, we denote them with the tilt bold upper-case letters $\boldsymbol{X}$, $\boldsymbol{Y}$ with sub/super-scripts. For vector $\boldsymbol{x} = (x_1, \ldots, x_d)^\top$, we use $\|\boldsymbol{x}\|_2 = \sqrt{x_1^2 + \cdots + x_d^2}$ to represent its ℓ_2-norm and use $\|\boldsymbol{x}\|_{\mathbf{A}} = \sqrt{\boldsymbol{x}^\top \mathbf{A} \boldsymbol{x}}$ to represent its $\mathbf{A}$-norm, where $\mathbf{A}$ is an semi-positive definite matrix. We use $\mathbb{P}(\boldsymbol{x})$ to denote the distribution of $\boldsymbol{x}$, and $\mathcal{W}_2(\cdot, \cdot)$ and

$D_{KL}(\cdot||\cdot)$ denote the 2-Wasserstein distance and Kullback–Leibler (KL) divergence between two distributions, respectively. For a function $f : \mathbb{R}^d \to \mathbb{R}$, we use $\nabla f(\cdot)$ and $\nabla^2 f(\cdot)$ to denote its gradient and Hessian.

2 Algorithms

2.1 Laplacian Smoothing (Stochastic) Gradient Descent

For $\sigma \geq 0$, let $\mathbf{A}_\sigma := \mathbf{I} - \sigma \mathbf{L}$ where $\mathbf{I} \in \mathbb{R}^{d \times d}$ and $\mathbf{L} \in \mathbb{R}^{d \times d}$ are the identity and the discrete one-dimensional Laplacian matrix, respectively. Therefore,

$$\mathbf{A}_\sigma := \begin{bmatrix} 1+2\sigma & -\sigma & 0 & \ldots & 0 & -\sigma \\ -\sigma & 1+2\sigma & -\sigma & \ldots & 0 & 0 \\ 0 & -\sigma & 1+2\sigma & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ -\sigma & 0 & 0 & \ldots & -\sigma & 1+2\sigma \end{bmatrix}_{d \times d} \tag{4}$$

To find $\boldsymbol{x}_{\mathrm{MAP}}$ of (1), LSSGD (Osher et al., 2018) takes the following iteration

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \eta_k \mathbf{A}_\sigma^{-1} \nabla f_{i_k}(\boldsymbol{x}^k), \tag{5}$$

where $\eta_k > 0$ is the learning rate, i_k is a random sample from $[n] := \{1, 2, \cdots, n\}$. When $\sigma = 0$, LSSGD reduces to SGD. Since $\mathbf{A}_\sigma$ is a circulant matrix, for any vector $\boldsymbol{v}$, $\mathbf{A}_\sigma^{-1}\boldsymbol{v} := \boldsymbol{u}$ can be computed via the FFT in the following way

$$\mathbf{A}_\sigma^{-1} \boldsymbol{v} = \boldsymbol{u} \Longrightarrow \boldsymbol{v} = \mathbf{A}_\sigma \boldsymbol{u} = \boldsymbol{u} - \sigma \boldsymbol{d} * \boldsymbol{u},$$

where $*$ is the convolution operator, and $\boldsymbol{d} = [-2, 1, 0, \cdots, 0, 1]^T$. By the convolution theorem, we have

$$\mathrm{fft}(\boldsymbol{v}) = \mathrm{fft}(\boldsymbol{u}) - \sigma \mathrm{fft}(\boldsymbol{d})\mathrm{fft}(\boldsymbol{u}).$$

Finally, we arrive at the following FFT-based algorithm for computing $\mathbf{A}_\sigma^{-1}\boldsymbol{v}$

$$\mathbf{A}_\sigma^{-1} \boldsymbol{v} = \mathrm{ifft}\left(\frac{\mathrm{fft}(\boldsymbol{v})}{\mathbf{1} - \sigma \cdot \mathrm{fft}(\boldsymbol{d})} \right),$$

where $\mathbf{1}$ is an all-one vector with the same dimension as $\boldsymbol{v}$, and the division of two vectors is defined in the coordinate-wise way. fft and ifft denote FFT and inverse FFT operators, respectively.

The Laplacian matrix $\mathbf{A}_\sigma^{-1}$ can reduce the variance of stochastic gradient and guarantee at least the same convergence rate as SGD. Osher et al. (2018) showed that for a L-gradient Lipschitz function $f(\boldsymbol{x})$, i.e., $\|\nabla f(\boldsymbol{x})\|_2 \leq L$, the largest step size for LSSGD is $(1 + 4\sigma)^{1/4}/L$ (with high probability) which is larger than GD's by a factor $(1 + 4\sigma)^{1/4}$.

2.2 Laplacian Smoothing Langevin Dynamics

We integrate Laplacian smoothing with LD and obtain the following Lapacian Smoothing LD (LS-LD)

$$d\boldsymbol{X}_t = -\mathbf{A}_\sigma^{-1} \nabla f(\boldsymbol{X}_t) + \sqrt{2\beta^{-1}} \mathbf{A}_\sigma^{-1/2} d\boldsymbol{B}_t. \tag{6}$$

Note that we pre-multiply the Brownian motion term by $\mathbf{A}_\sigma^{-1/2}$ instead of $\mathbf{A}_\sigma^{-1}$ to guarantee that the stationary distribution of the LS-LD remains to be $\exp(-\beta f(\boldsymbol{x}))$. We can easily verify that LD and LS-LD have the same stationary distribution by looking at the associated Fokker-Planck equation. We formally state this property in the following proposition.

Proposition 1. *The stationary distribution, π, of the LS-LD, (6), satisfies $\pi \propto e^{-\beta f(\boldsymbol{x})}$.*

If we apply the Euler-Maruyama scheme to discretize (6), we end up with the following discrete algorithm, namely Laplacian smoothing gradient Langevin dynamics (LS-GLD)

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta \mathbf{A}_\sigma^{-1} \nabla f(\boldsymbol{x}_k) + \sqrt{2\beta^{-1}\eta} \mathbf{A}_\sigma^{-1/2} \boldsymbol{\epsilon}_k, \tag{7}$$

where $\boldsymbol{\epsilon}_k \sim N(\mathbf{0}, \mathbf{I}_{d \times d})$. In practice, we use the mini-batch gradient $\mathbf{g}_k = \sum_{i \in \mathcal{I}_k} \nabla f_i(\boldsymbol{x}_k)/|\mathcal{I}_k|$ with $\mathcal{I}_k \subset [n]$ to replace the gradient in (7), and we arrive at the following LS-SGLD

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta \mathbf{A}_\sigma^{-1} \mathbf{g}_k + \sqrt{2\beta^{-1}\eta} \mathbf{A}_\sigma^{-1/2} \boldsymbol{\epsilon}_k. \tag{8}$$

We summarize LS-SGLD in Algorithm 1. It is worth noting that computing the inverse of both $\mathbf{A}_\sigma$ and $\mathbf{A}_\sigma^{1/2}$ can be expensive. Moreover, multiplying the vectors by the inverse of these two matrices is also expensive. So in the remaining part of this section, we will present FFT-based fast algorithms for implementing (7).

Algorithm 1 LS-SGLD

Input: Training data, learning rate η, minibatch size B, inverse temperature β, Laplacian smoothing constant σ.
Initialization: Set $\boldsymbol{x}_0 = 0$.
for $k = 0, 1, \cdots, K-1$ **do**
    Uniformly sample $\mathcal{I}_k \subset [n]$ with $|\mathcal{I}_k| = B$.
    Compute the mini-batch stochastic gradient $\sum_{i \in \mathcal{I}_k} \nabla f_i(\boldsymbol{x}_k)/B$.
    $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta \mathbf{A}_\sigma^{-1} \mathbf{g}_k + \sqrt{2\beta^{-1}\eta} \mathbf{A}_\sigma^{-1/2} \boldsymbol{\epsilon}_k$, where $\boldsymbol{\epsilon}_k \sim N(\mathbf{0}, \mathbf{I}_{d \times d})$.
Output: $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_K$.

2.3 FFT-based Implementation of LS-SGLD

2.3.1 Circulant Matrix and Convolutional Operation

In this subsection, we list a few results on the circulant matrix which will be the basic recipes for designing FFT-based algorithm to solving (7).

Lemma 1 (Golub and Van Loan (1996)). *The normalized eigenvectors of the following $d \times d$ circulant matrix,*

$$\mathbf{C} = \begin{bmatrix} c_0 & c_{d-1} & \ldots & c_2 & c_1 \\ c_1 & c_0 & c_{d-1} & \ldots & c_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ c_{d-2} & \ldots & \ldots & \ldots & c_{d-1} \\ c_{d-1} & c_{d-2} & \ldots & c_1 & c_0 \end{bmatrix}, \tag{9}$$

are given by

$$\boldsymbol{v}_j = \frac{1}{\sqrt{d}} \left(1, w_j, w_j^2, \cdots, w_j^{n-1}\right), \quad j = 0, 1, \cdots, d-1,$$

where $w_j = \exp\left(i\frac{2\pi j}{d}\right)$ are the j-th roots of unity and i is the imaginary unit. The corresponding eigenvalues are then given by

$$\lambda_j = c_0 + c_{d-1}w_j + c_{d-2}w_j^2 + \cdots + c_1 w_j^{d-1}, \quad j = 0, 1, \cdots, d-1.$$

Lemma 2 (Golub and Van Loan (1996)). *The inverse of a circulant matrix is circulant.*

Lemma 3 (Golub and Van Loan (1996)). *The square root of a circulant matrix is circulant.*

Lemma 4. *For any circulant matrix $\mathbf{C}$ of the form in (9), and for any given vector $\boldsymbol{v}$. Let $\boldsymbol{u} = \mathbf{C}^{-1}\boldsymbol{v}$, then $\boldsymbol{u}$ can be computed by the fast Fourier transform with sublinear scaling in the following way*

$$\boldsymbol{u} = \text{ifft}\left(\frac{\text{fft}(\boldsymbol{v})}{\text{fft}(\boldsymbol{c})}\right), \tag{10}$$

where $\boldsymbol{c}$ is the first row of the matrix $\mathbf{C}$, and the division in (10) is defined coordinate-wise.

Proof. Since $\mathbf{C}$ is a circulant matrix we have $\boldsymbol{v} = \mathbf{C}\boldsymbol{u} = \boldsymbol{c} * \boldsymbol{u}$, therefore $\text{fft}(\boldsymbol{v}) = \text{fft}(\boldsymbol{c}) \cdot \text{fft}(\boldsymbol{u})$. $\qquad\square$

2.3.2 Fast Algorithm for Computing the Square Root of Laplacian Smoothing

We will derive an FFT-based algorithm for computing $\mathbf{A}_\sigma^{-1/2}\boldsymbol{\epsilon}_k$ in this subsection. According to Lemmas 2 and 3, $\mathbf{A}_\sigma^{-1/2}$ is circulant. Note $\mathbf{A}_\sigma^{-1}$ is positive definite, we denote its eigen-decomposition as

$$\mathbf{A}_\sigma^{-1} = \mathbf{Q}\Lambda\mathbf{Q}^{-1},$$

where $\mathbf{Q} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_d]^T$ with $\boldsymbol{v}_i$ being the eigenvector associated with the eigenvalue $\lambda_i > 0$, and $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_d)$. Therefore, we have

$$\mathbf{A}_\sigma^{-1/2} = \mathbf{Q}\sqrt{\Lambda}\mathbf{Q}^{-1}, \tag{11}$$

where $\sqrt{\Lambda} = \mathrm{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \cdots, \sqrt{\lambda_d})$.

Furthermore, note that $\mathbf{A}_\sigma$ is symmetric, therefore $\mathbf{Q}^{-1} = \mathbf{Q}^T$. It follows that we can compute $\mathbf{A}_\sigma^{-1/2}$ without inverting the matrix $\mathbf{Q}$. By the fact that $\mathbf{A}_\sigma^{-1/2}$ is circulant, we have $\mathbf{A}_\sigma^{-1/2}\boldsymbol{\epsilon}_k = \mathrm{ifft}(\mathrm{fft}(\boldsymbol{b}) \cdot \mathrm{fft}(\boldsymbol{\epsilon}_k))$, where $\boldsymbol{b}$ is the first row of $\mathbf{A}_\sigma^{-1/2}$.

Remark 1. *In computing* (11), *there is no need to store the matrix* $\mathbf{Q}$, *according to Lemma 1, each row of* $\mathbf{Q}$ *and* $\sqrt{\Lambda}$ *can be written down explicitly which enables us to compute* $\mathbf{A}_\sigma^{-1/2}$ *quickly with negligible memory overhead and scalable to very high dimensional problems.*

3 Main Results

We first make the following three assumptions regarding the function $f(\boldsymbol{x})$.

Assumption 1 (Dissipativeness). *For any $\boldsymbol{x} \in \mathbb{R}^d$, there exist constants m and b such that*

$$\langle \nabla f(\boldsymbol{x}), \boldsymbol{x} \rangle \geq m\|\boldsymbol{x}\|_2^2 - b.$$

This assumption has been widely made to study the convergence of Langevin based sampling algorithms (Mattingly et al., 2002; Raginsky et al., 2017; Xu et al., 2018; Zou et al., 2019a), which is essential to guarantee the convergence of the continuous-time Langenvin dynamics (2).

Assumption 2 (smoothness). *For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, there exists a positive constant M such that for all $i = 1, \ldots, n$, it holds that*

$$\|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{y})\|_2 \leq M\|\boldsymbol{x} - \boldsymbol{y}\|_2.$$

Unlike Assumption 1, Assumption 2 is made for all component function $f_i(\boldsymbol{x})$.

Assumption 3 (Bounded Variance). *For any $\boldsymbol{x} \in \mathbb{R}^d$, there exists a constant ω such that the variance of the stochastic gradient is bounded as follows,*

$$\mathbb{E}[\|\nabla f_i(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|_2^2] \leq d\omega^2.$$

Definition 1 (Logarithmic Sobolev inequality). *Let μ be a probability measure, then we say μ satisfies logarithmic Sobolev inequality with constant λ if for any smooth function g, the following holds:*

$$\int g^2 \log g^2 \, d\mu - \int g^2 \, d\mu \log \int g^2 \, d\mu \leq \lambda \int \|\nabla g\|_2^2 \, d\mu.$$

Then the following proposition states that if the function $f(\cdot)$ satisfies Assumptions 1 and 2, the target density $\pi \propto e^{-f(\boldsymbol{x})}$ satisfies Logarithmic Sobolev inequality.

Proposition 2 (Raginsky et al. (2017)). *Under Assumptions 1 and 2, the target density $\pi \propto e^{-f(\boldsymbol{x})}$ satisfies Logarithmic Sobolev inequality with some constant $\lambda > 0$.*

It has been shown in Durmus et al. (2017); Raginsky et al. (2017) that if the function $f(\boldsymbol{x})$ is smooth and strongly convex (which is stronger than Assumption 1), the logarithmic Sobolev constant λ is an universal constant. However, if the function $f(\boldsymbol{x})$ is nonconvex, in the worst case the logarithm Sobolev constant λ can have exponential dependency on the problem dimension d and inverse temperature β (Bovier et al., 2004; Raginsky et al., 2017).

Table 1: The values of γ_2 corresponding to some σ and d.

σ	1	2	3	4	5
$d = 1000$	0.268	0.185	0.149	0.128	0.114
$d = 10000$	0.268	0.185	0.149	0.128	0.114
$d = 100000$	0.268	0.185	0.149	0.128	0.114

3.1 Convergence Analysis of Sampling from Log-concave Densities

In this subsection, we assume that the target density is log-concave, which is equivalent to the following assumption on the function $f(\boldsymbol{x})$.

Assumption 4 (Convexity). *For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, it holds that*

$$f(\boldsymbol{x}) - f(\boldsymbol{y}) \geq \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle.$$

Then we are ready to establish the convergence rate of LS-SGLD for sampling from log-concave densities, which is stated in the following theorem.

Theorem 1. *Under Assumptions 1, 2, 3 and 4, if set the step size $\eta \leq Cm\beta^{-1}/M^2$ for some sufficiently small constant C, there exist constants $c_0 \in [\|\mathbf{A}_\sigma\|_2^{-1}, 1]$, $\gamma_1 \in [\|\mathbf{A}_\sigma\|_2^{-2}, 1]$ and $\gamma_2 = d^{-1} \sum_{i=1}^d (1 + 2\sigma - 2\sigma \cos(2\pi i/d))^{-1}$ such that the output of LS-SGLD satisfies,*

$$\mathcal{W}_2(\mathbb{P}(\boldsymbol{x}_K), \pi) \leq \left(\frac{2\gamma_1 K \eta^2 \beta d \omega^2}{B} \right)^{1/2} + \left[8\gamma_2 K \eta^2 \cdot (K+1)\beta d\eta \right]^{1/2} + \left[2\lambda \big(\beta f(0) + \log(\Lambda) \big) \right]^{1/2} \cdot e^{-c_0 K \eta/(2\beta\lambda)},$$

$$(12)$$

where $\Lambda = \int_{\mathbb{R}^d} e^{-\beta f(\boldsymbol{x})} d\boldsymbol{x}$ and λ denotes the logarithmic Sobolev constant of the target distribution $\pi \propto e^{-\beta f(\boldsymbol{x})}$.

Remark 2. *We emphasize that the the three terms on the R.H.S. of (12) have their respective meanings. In particular, the first and second terms represent the discretization errors introduced by stochastic gradient estimator and numerical integrator of (6), respectively. The third term represents the ergodicity of the continuous-time Markov process (6), which characterizes the mixing time of LS-LD (6). Moreover, we remark here that the convergence rate of LS-GLD (LS-SGLD with full gradient) can be directly implied from Theorem 1 by removing the first term on the R.H.S. of (12).*

Based on Theorem 1, we can also derive the convergence rate of SGLD in the same setting by setting $\mathbf{A}_\sigma = \mathbf{I}$ (i.e., $\sigma = 0$), which implies that the constants γ_1, γ_2 and c_0 in Theorem 1 are all 1's. We formally state the convergence result of SGLD in the following corollary.

Corollary 1. *Under the same assumptions in Theorem 1, the output of standard SGLD, denoted by $\boldsymbol{y}_K$, satisfies*

$$\mathcal{W}_2(\mathbb{P}(\boldsymbol{y}_K), \pi) \leq \left(\frac{2K \eta^2 d \omega^2}{B} \right)^{1/2} + \left[8K \eta^2 \cdot (K+1)\beta^{-1} d\eta \right]^{1/2} + \left[2\lambda \big(\beta f(0) + \log(\Lambda) \big) \right]^{1/2} \cdot e^{-c_0 K \eta/(2\beta\lambda)},$$

$$(13)$$

Remark 3. *We can now compare the convergence rates of LS-SGLD and SGLD. In terms of the discretization error, it is clear that LS-SGLD is strictly better since the constants γ_1 and γ_2 are strictly less than 1 (some values of γ_2 corresponding to different choices of σ and d can be found in Table 1). In terms of the ergodicity of the continuous-time Markov process (the third terms in (12) and (13)), LS-SGLD is worse than SGLD due to the fact that $c_0 \leq 1$. Therefore, there exists a trade-off between the discretization error and the ergodicity rate of LS-SGLD. In our experiments we will conduct numerical evaluations of these error terms and demonstrate that LS-LD and LD achieve similar ergodicity performance (i.e., mixing time), but LS-SGLD can achieve a significantly smaller discretization error.*

3.2 Convergence Analysis of Sampling from Non-log-concave Densities

Here we consider the setting where the target density is no longer log-concave. The following theorem states the convergence rate of LS-SGLD in 2-Wasserstein distance.

Theorem 2. *Under Assumptions 1, 2 and 3, if set the step size $\eta \le Cm\beta^{-1}/M^2$ for some sufficiently small constant C, there exist constants $c_0 \in [\|\mathbf{A}_\sigma\|_2^{-1}, 1]$, $\gamma_1 \in [\|\mathbf{A}_\sigma\|_2^{-2}, 1]$, $\gamma_2 = d^{-1}\sum_{i=1}^{d}(1+2\sigma-2\sigma\cos(2\pi i/d))^{-1}$ and $\bar{\Gamma} = \left(3/2 + 2(b+\beta^{-1}d)\right)^{1/2}$ such that the output of LS-SGLD satisfies,*

$$\mathcal{W}_2\big(\mathbb{P}(\boldsymbol{x}_K), \pi\big) \le \bar{\Gamma}(K\eta)^{1/2}\left[\left(\frac{\gamma_1\beta d\omega^2}{B}K\eta + 2\gamma_2 M^2 dK\eta^2\right)^{1/2} + \left(\frac{\gamma_1\beta d\omega^2}{B}K\eta + 2\gamma_2 M^2 dK\eta^2\right)^{1/4}\right]$$
$$+ \left[2\lambda\big(\beta f(0) + \log(\Lambda)\big)\right]^{1/2} \cdot e^{-c_0 K\eta/(2\beta\lambda)}, \tag{14}$$

where $\Lambda = \int_{\mathbb{R}^d} e^{-\beta f(\boldsymbol{x})} d\boldsymbol{x}$ and λ denotes the logarithmic Sobolev constant of the target distribution $\pi \propto e^{-\beta f(\boldsymbol{x})}$.

Remark 4. *The convergence rate of SGLD in 2-Wasserstein distance can also be obtained from Theorem 2 by setting $\mathbf{A}_\sigma = \mathbf{I}$, which implies that the constants c_0, γ_1, γ_2 become all 1's. It can be verified that the resulting convergence rate matches that proved in Raginsky et al. (2017). As a clear comparison, the discretization error induced by both stochastic gradient and numerical integrator of LS-SGLD (the first bracket term of (14)) is smaller than that of SGLD, while the ergodicity term of LS-SGLD (the last term of (14)) is worse than that of SGLD. Again, we will experimentally demonstrate that the mixing time of LS-LD is not much slower compared with LD, but LS-SGLD can achieve significantly smaller discretization error than SGLD.*

4 Numerical Results

In this section, we will perform numerical experiments on sampling 2D distributions, training Bayesian Logistic Regression (BLR), and training Convolutional Neural Nets (CNNs). Throughout all the experiments, we regard SGLD (Welling and Whye, 2011) and preconditioned SGLD (pSGLD) (Li et al., 2016), which considers local curvature of $f(\boldsymbol{x})$ with RMSProp type of adaptive step size, as benchmarks. In addition, we also incorporated the precondition technique, proposed in Li et al. (2016), into LS-SGLD, which leads to a variant of LS-SGLD, namely Laplacian smoothing preconditioned SGLD (LS-pSGLD).

4.1 Numerical Simulations on Synthetic Dataset

4.1.1 2D Gaussian Distribution

As a simple illustration, we apply the proposed LS-SGLD and LS-pSGLD to sample a 2D Gaussian distribution, studied in Chen et al. (2014), with the probability density function $e^{f(\boldsymbol{x})} = \exp\left(\frac{1}{2}\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}\right)$ with $\boldsymbol{x} \in \mathbb{R}^2$ where $\Sigma = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$. We let the prior to be $p(\boldsymbol{x}) = \mathcal{N}(0, \nu^2\mathbf{I})$ with $\nu = 1$. For both SGLD and pSGLD, we let the step size to be 0.19 which is obtained based on the grid search. For LS-SGLD and LS-pSGLD, we let the Laplacian smoothing parameter σ to be 0.1 with step size to be either 0.19 or $0.19(1+4\sigma)^{1/4}$. It is worth noting that in 2D $\mathbf{A}_\sigma$ becomes

$$\mathbf{A}_\sigma = \begin{bmatrix} 1+\sigma & -\sigma \\ -\sigma & 1+\sigma \end{bmatrix}. \tag{15}$$

To measure the quality of samples, we consider the MSE between the true and reconstructed covariance matrices; and we calculate the autocorrelation time of the samples to verify the efficacy of different samplers in sampling the correlated distribution above. The autocorrelation time is defined as following

$$\tau = \frac{1}{2} + \sum_{t=1}^{\infty} \frac{A(t)}{A(0)}, \tag{16}$$

where $A(t) = \mathbb{E}\left[(\bar{\phi}_\eta - \phi(\boldsymbol{x}_0))(\bar{\phi}_\eta - \phi(\boldsymbol{x}_t))\right]$ for any given bounded function $\phi(\boldsymbol{x})$, $\bar{\phi} = \int_\chi \phi(\boldsymbol{x})p(\boldsymbol{x}|\mathcal{D})d\boldsymbol{x}$ is the population mean of ϕ, and the empirical mean $\hat{\phi} = \frac{1}{S_T}\sum_{t=1}^{T}\eta_t\phi(\boldsymbol{x}_t)$ with η_t being the step size at the t-th step and $S_T = \sum_{t=1}^{T}\eta_t$.

Figure 1 (a) and (c) plot the first 600 samples from the target distribution by different samplers. We use the same step size 0.19 for all the four samplers in the experiments shown in Fig. 1 (a), and use a larger step size $0.19(1 + 4\sigma)^{1/4}$ for LS-SGLD and LS-pSGLD in experiments shown in Fig. 1 (c). Qualitatively, Laplacian smoothing can enhance the quality of samples, and the improvement becomes more remarkable when we use a larger step size. Next, we draw 2×10^5 samples from the target distribution by different samplers and we use these samples to reconstruct the covariance matrix. For SGLD and pSGLD, we use a set of step size $\{0.19, 0.19 \times 0.8, 0.19 \times 0.8^2, 0.19 \times 0.8^3, 0.19 \times 0.8^4\}$. For LS-SGLD and LS-pSGLD we test two sets of step sizes: (i) $\{0.19, 0.19 \times 0.8, 0.19 \times 0.8^2, 0.19 \times 0.8^3, 0.19 \times 0.8^4\}$, (ii) $\{0.19, 0.19 \times 0.8, 0.19 \times 0.8^2, 0.19 \times 0.8^3, 0.19 \times 0.8^4\} \times (1 + 4\sigma)^{1/4}$. Figure 1 (b) and (d) plot the autocorrelation time v.s. reconstruction error of the covariance matrix. In (b) we use the same set of step sizes for all the four samplers, and in (d) we use a larger step size for LS-SGLD and LS-pSGLD. We see that reconstruction errors can be reduced significantly when Laplacian smoothing is used. Moreover, Laplacian smoothing can also reduce the autocorrelation time in pSGLD.
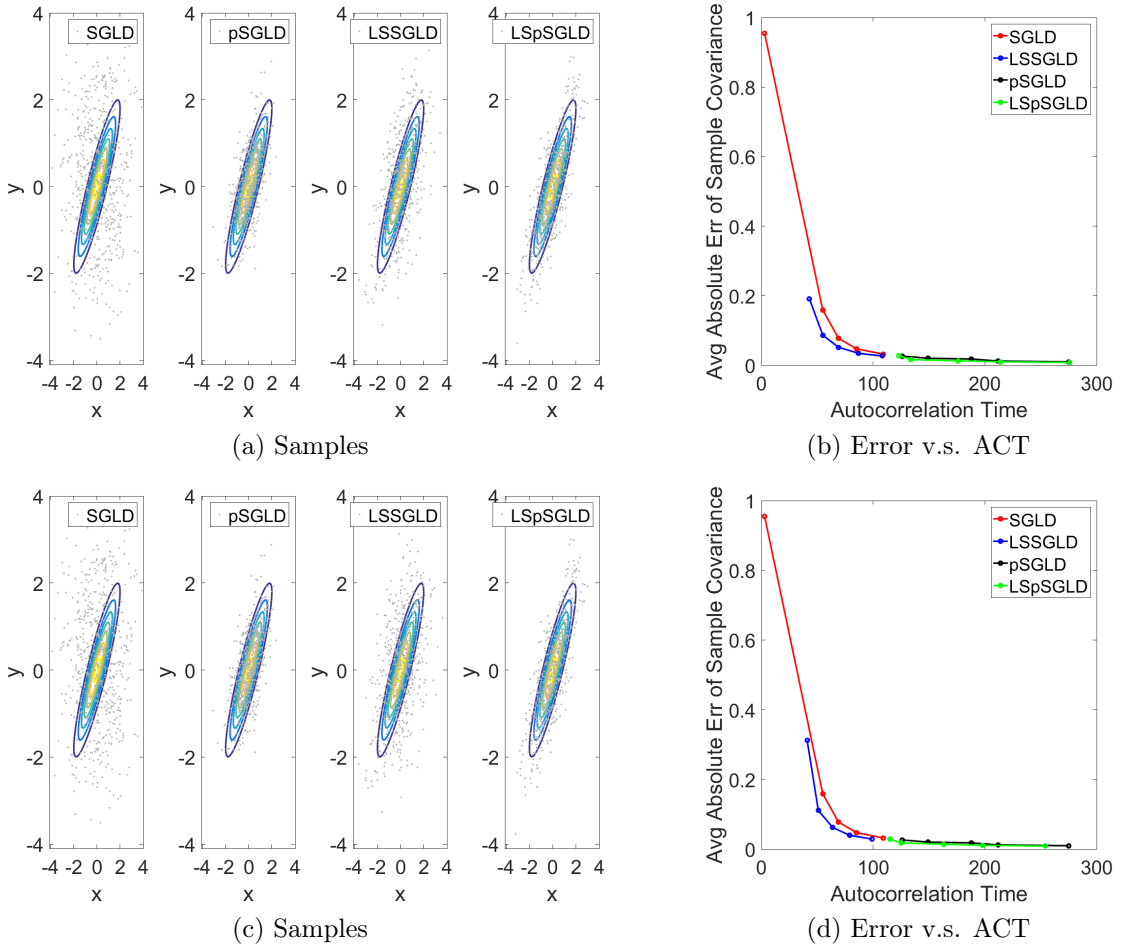

(a) Samples

(b) Error v.s. ACT

(c) Samples

(d) Error v.s. ACT

Figure 1: Contrasting sampling of a 2D Gaussian distribution with covariance matrix Σ, with $\Sigma_{11} = \Sigma_{22} = 1$ and $\Sigma_{12} = \Sigma_{21} = 0.9$, using different samplers. (a) and (c): the first 600 samples draw by SGLD, pSGLD. (b) and (d): LS-SGLD, and LS-pSGLD. In (c) and (d), we multiply the step size for LS-SGLD and LS-pSGLD by a factor $(1 + 4\sigma)^{1/4}$.

4.1.2 2D Gaussian Mixture Distribution

In this subsection, we compare the performance of SGLD, pSGLD, LS-SGLD and LS-pSGLD on a Gaussian mixture distribution. In particular, we consider the target distribution

$$\pi \propto \exp\left(-f(\boldsymbol{x})\right) = \exp\left(-\frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x})\right), \quad n = 500,$$

where each component $\exp\left(-f_i(\boldsymbol{x})\right)$ is defined as

$$\exp\left(-f_i(\boldsymbol{x})\right) = \frac{2}{3}e^{-\frac{\|\boldsymbol{x}-\boldsymbol{a}_i\|_2^2}{2}} + \frac{1}{3}e^{-\frac{\|\boldsymbol{x}+\boldsymbol{a}_i\|_2^2}{2}},$$

where we sample $\boldsymbol{a}_i$ by the MCMC sampler with MH correction from the following 2D Gaussian distribution

$$\mathcal{N}\left(\begin{bmatrix} 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}\right).$$

The function $f_i(\boldsymbol{x})$ and its gradient can be simplified as

$$f_i(\boldsymbol{x}) = \frac{\|\boldsymbol{x}-\boldsymbol{a}_i\|_2^2}{2} - \log\left(\frac{2}{3} + \frac{1}{3}\exp\left(-2\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle\right)\right),$$

$$\nabla f_i(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{a}_i + \frac{2\boldsymbol{a}_i}{2 + \exp\left(2\langle \boldsymbol{x}, \boldsymbol{a}_i\rangle\right)}.$$

It can be easily verified that if $\boldsymbol{a}_i$ satisfies $\|\boldsymbol{a}_i\|_2 > 3/2$, the function $f_i(\boldsymbol{x})$ defined above is nonconvex. Moreover, it can be seen that

$$\langle \nabla f_i(\boldsymbol{x}), \boldsymbol{x}\rangle = \|\boldsymbol{x}\|_2^2 - \frac{\exp\left(2\langle \boldsymbol{x}, \boldsymbol{a}_i\rangle\right)}{2 + \exp\left(2\langle \boldsymbol{x}, \boldsymbol{a}_i\rangle\right)}\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle \geq \frac{1}{2}\|\boldsymbol{x}\|_2^2 - \frac{1}{2}\|\boldsymbol{a}_i\|_2^2,$$

which suggests that the function $f_i(\boldsymbol{x})$ satisfies the Dissipative Assumption 1 with $m = \frac{1}{2}$ and $b = \frac{\|\boldsymbol{a}_i\|_2^2}{2}$, and it further implies that $f(\boldsymbol{x})$ is also dissipative.

Since it takes a large number of samples to characterize the distribution, which makes repeated experiments computationally expensive, we instead follow Bardenet et al. (2017); Zou et al. (2019a,b) to use iterates along one Markov chain to visualize the distribution of iterates obtained by MCMC algorithms. We run the four samplers with different numbers of iteration where we set the batch size to be 10. We plot the distributions generated by different samplers with different numbers of iterations in Fig. 2. As shown in Fig. 2 (c), (f), and (u), when the number of iterations is large enough, e.g. 10^6, the sample distributions of all the three samplers matches well with the reference distribution (sampled by ground-truth sampler, e.g., MCMC with MH step). However, when the number of iterations is not enough, there is a large discrepancy between the sample and target distributions, as shown in Fig. 2 (a), (d), and (g). With a moderate number of iterations, say 5×10^5, the sample distribution from LS-SGLD is better than the other two (Fig. 2 (b), (e), and (d)).

Let us further evaluate the sample quality in a quantitative approach. We first apply the MCMC with MH step to sample $10K$ samples from the above target distribution. Then we apply SGLD, pSGLD, and LS-SGLD to sample different numbers of samples, respectively, from the target distribution. We measure the 2-Wasserstein distance between the last $10K$ samples of the different number of samples by the above three stochastic gradient samplers with the MH samples. We list the Wasserstein distance between the last $10K$ samples of different numbers of samples from different samplers with the MH samples in Table 2. These results show that the samples generated by LS-SGLD are consistently closer to the samples from MCMC with MH correction.
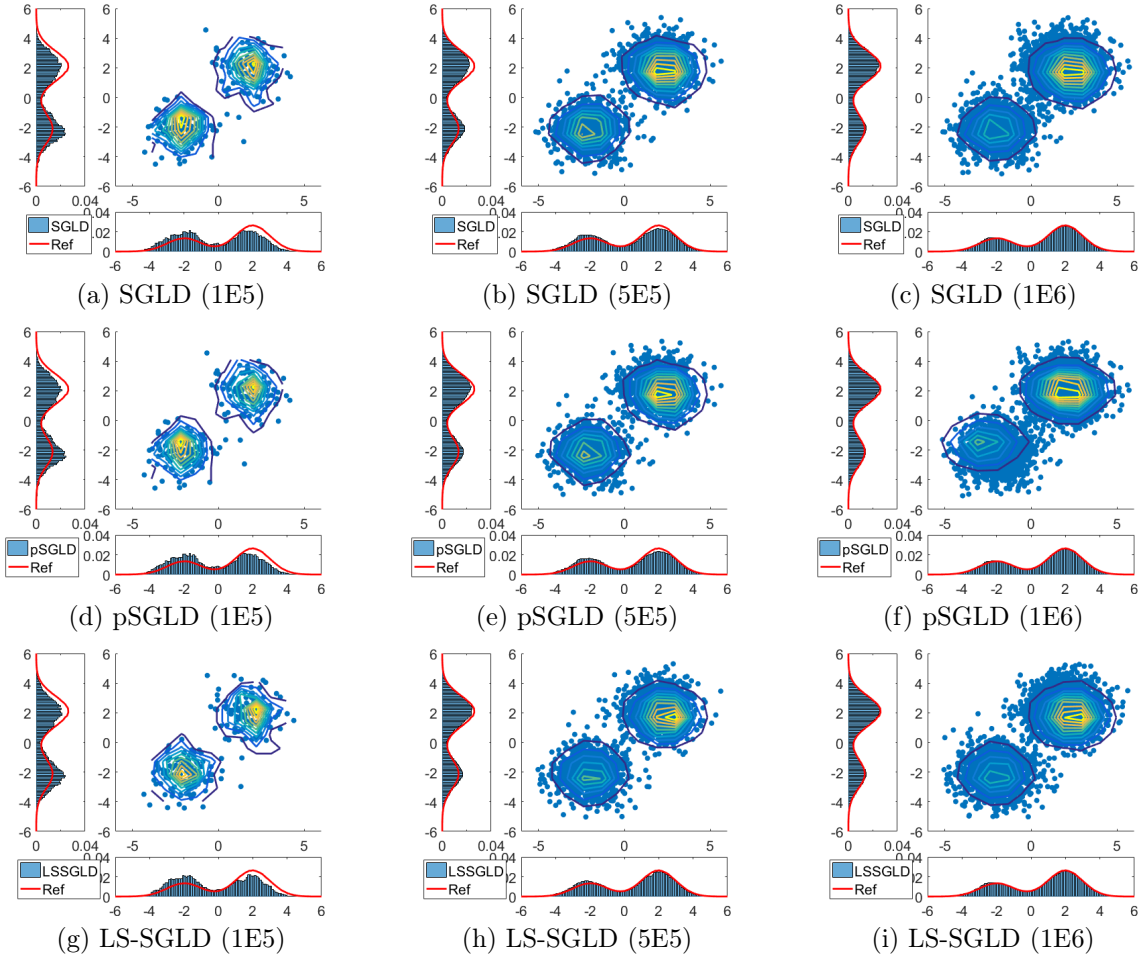
Figure 2: Kernel density plots of samples generated from Gaussian mixture distribution using SGLD, LS-SGLD, pSGLD and LS-pSGLD. We set $\sigma = 1.0$ for LS-SGLD and LS-pSGLD.

Table 2: 2-Wasserstein distance between samples sampled by MCMC with Metropolis–Hastings correction and different stochastic gradient Langevin dynamics.

# of Samples	1E5	5E5	9E5
SGLD	0.695	6.726	0.285
pSGLD	5.364	0.286	6.728
LS-SGLD	0.421	0.414	0.418

4.1.3 Comparison of the mixing time between LD and LS-LD

To verify that Laplacian smoothing does not slow down the mixing rate of the continuous-time Markov process, we conduct the following experiments. First, we apply the MCMC with Metropolis-Hasting correction step to sample 10K points, respectively, from the following two distributions.

- Gaussian distribution with the following probability density function

$$p(x, y) = \frac{1}{9\pi} \exp\left(-\left(\frac{(x-1)^2}{3^2} + \frac{(y-2)^2}{3^2} \right) \right). \tag{17}$$

- The Gaussian mixture distribution described in subsection 4.1.2.

11

Second, we use either LD or LS-LD (which can be approximated by Euler-Maruyama discretization with very small step size), to draw samples from the above two distributions and use these samples to estimate the mean of the target densities. Figure 3 plots the MSE between the true and reconstructed (from a different number of samples) means, and they show that LD and LS-LD perform similarly in reconstructing the mean of the target densities.
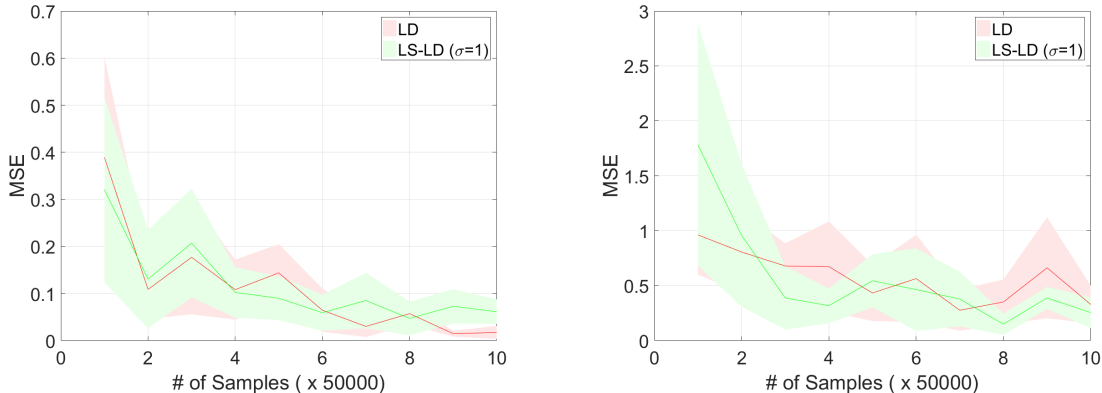

Figure 3: MSE between the true and reconstructed means from different numbers of samples generated by LD and LS-LD (10 independent runs).

4.2 Bayesian Logistic Regression

Suppose we observe n i.i.d samples $\{\boldsymbol{d}_i, y_i\}_{i=1,2,\cdots,n}$ where $\boldsymbol{d}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$ denote the feature and the corresponding label of the i-th sample instance. The likelihood of BLR model is given by

$$p(y_i|\boldsymbol{d}_i, \boldsymbol{x}) = \frac{1}{1 + \exp\left(-y_i\langle\boldsymbol{d}_i, \boldsymbol{x}\rangle\right)}$$

where $\boldsymbol{x}$ is the parameter to be learned. We use a Gamma prior $p(\boldsymbol{x}) \propto \|\boldsymbol{x}\|_2^{-\lambda} \exp\left(-\theta\|\boldsymbol{x}\|_2\right)$ with $\lambda = 1$ and $\theta = 10^{-2}$. Then we formulate the logarithmic posterior distribution as follows

$$\log\left[p(\boldsymbol{x}|\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_n; y_1, y_2, \cdots, y_n)\right] \propto -\frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}),$$

where $f_i(\boldsymbol{x}) = n\log\left(1 + e^{-y_i\langle\boldsymbol{d}_i, \boldsymbol{x}\rangle}\right) + \lambda\log\left(\|\boldsymbol{x}\|_2\right) + \theta\|\boldsymbol{x}\|_2$.

We use SGLD, pSGLD, LS-SGLD, and LS-pSGLD with batch size 5 to train a Bayesian logistic regression (BLR) model on the benchmark *a3a* dataset from the UCI machine learning repository [1]. The *a3a* dataset contains 3185 training data and 29376 test instances, each data instance is of dimension 122. We use the grid search to determine the optimal learning rate for SGLD (0.001) and pSGLD (0.002), and then we multiply them by $(1 + 4\sigma)^{1/4}$ to get the learning rate for LS-SGLD and LS-pSGLD. We set the burn-in to be 1000 for all these four samplers. After burn-in, we compute the moving average of the sample parameters to estimate the regression parameters $\boldsymbol{x}$. We plot iteration v.s. negative log-likelihood and accuracy in Fig. 4, and we see that Laplacian smoothing reduces the negative log-likelihood and increases the accuracy. The preconditioning accelerates mixing initially, however, the gap between sampling and target distribution is remarkably larger than the case without preconditioning.

4.2.1 Variance reduction in stochastic gradient

We numerically verify the efficiency of variance reduction on BLR for *a3a* dataset classification. We first compute a path by full batch SGLD with the same learning rate as before, and meanwhile, we record the

---
[1] https://archive.ics.uci.edu/ml/index.php

(a) Log-likelihood (Training Set)

(b) Training Accuracy

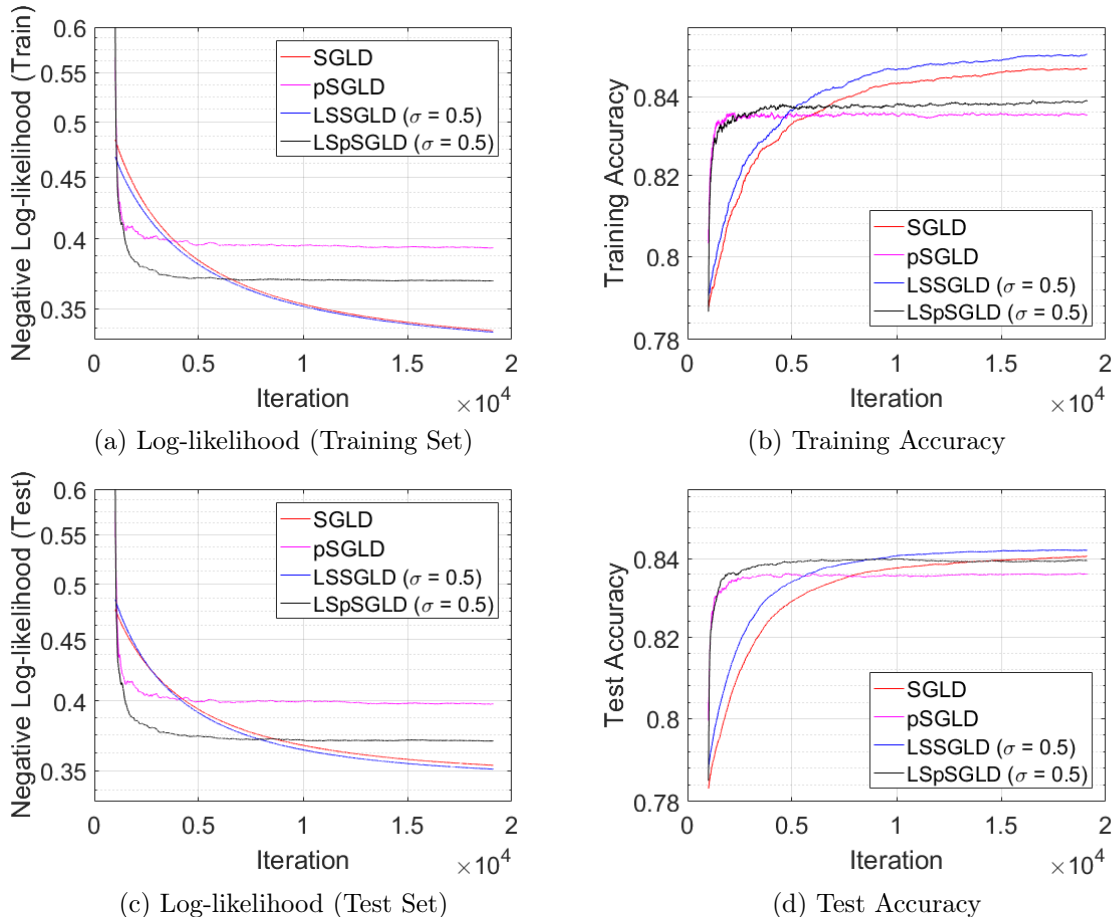(c) Log-likelihood (Test Set)

(d) Test Accuracy

Figure 4: Convergence comparison for Bayesian logistic regression, where the X-axis represents the number of iterations and Y-axis represents the negative log-likelihood/accuracy. (a) Negative log-likelihood on training dataset; (b) Accuracy on training dataset; (c) Negative log-likelihood on test dataset; (d) Accuracy on test dataset.

Laplacian smoothing gradient on each point along the path. Then we compute the Laplacian smoothing stochastic gradients on each point along the path by using different batch size and σ. We run 100 independent experiments to acquire the Laplacian smoothing stochastic gradients, and then we compute the variance of these stochastic gradients by regarding the full batch Laplacian smoothing gradient as the mean. In Table 3, we report the maximum variance, among all coordinates of the gradient and all points on the descent path, for each pair of batch size and σ.

Table 3: The maximum variance of the stochastic gradients generated by LS-SGLD on training BLR on the *a3a* data. $\sigma = 0$ reduces to SGLD.

Batch Size	10	15	50
$\sigma = 0$	7.69E-1	3.17E-1	5.69E-2
$\sigma = 0.5$	2.56E-1	1.06E-2	1.96E-2
$\sigma = 1.0$	1.54E-1	6.37E-2	1.21E-2
$\sigma = 2.0$	8.52E-2	3.54E-2	7.04E-3

4.3 Bayesian Convolutional Neural Network

We consider training a CNN by SGLD, pSGLD, LS-SGLD, and LS-pSGLD on the MNIST benchmark with batch size 100, the architecture of the CNN is

$$\text{CNN: input}_{28\times28} \to \text{conv}_{20,5,2} \to \text{conv}_{20,20,5} \to \text{fc}_{128} \to \text{softmax}.$$

The notation $\text{conv}_{c,k,m}$ denotes a 2D convolutional layer with c output channels, each of which is the sum of a channel-wise convolution operation on the input using a learnable kernel of size $k \times k$, it further adds ReLU nonlinearity and max-pooling with stride size m. fc_{128} is an affine transformation that transforms the input to a vector of dimension 128. Finally, the tensors are activated by a multi-class logistic function.

Similar to BLR, we use a Gamma prior $p(\boldsymbol{x}) \propto \|\boldsymbol{x}\|_2^{-\lambda} \exp\left(-\theta\|\boldsymbol{x}\|_2\right)$ with $\lambda = 1$ and $\theta = 5e^{-4}$. Again, we use the grid search to find the optimal step size for SGLD and pSGLD which is 0.02 and $2e-4$, respectively. We multiply the optimal step size for SGLD and pSGLD by a factor $(1+4\sigma)^{1/4}$ to get the step size for LS-SGLD and LS-pSGLD, and we let $\sigma = 0.5$ for Laplacian smoothing. The comparisons between different sampling algorithms are plotted in Fig. 5, we see that Laplacian smoothing reduces the negative log-likelihood and increases the accuracy of both training and test datasets. The preconditioning accelerates mixing and reduces the gap between sampling and target distribution. Here, we applied early stopping in training CNN by pSGLD and LS-pSGLD.

(a) Log-likelihood (Training Set)

(b) Training Accuracy
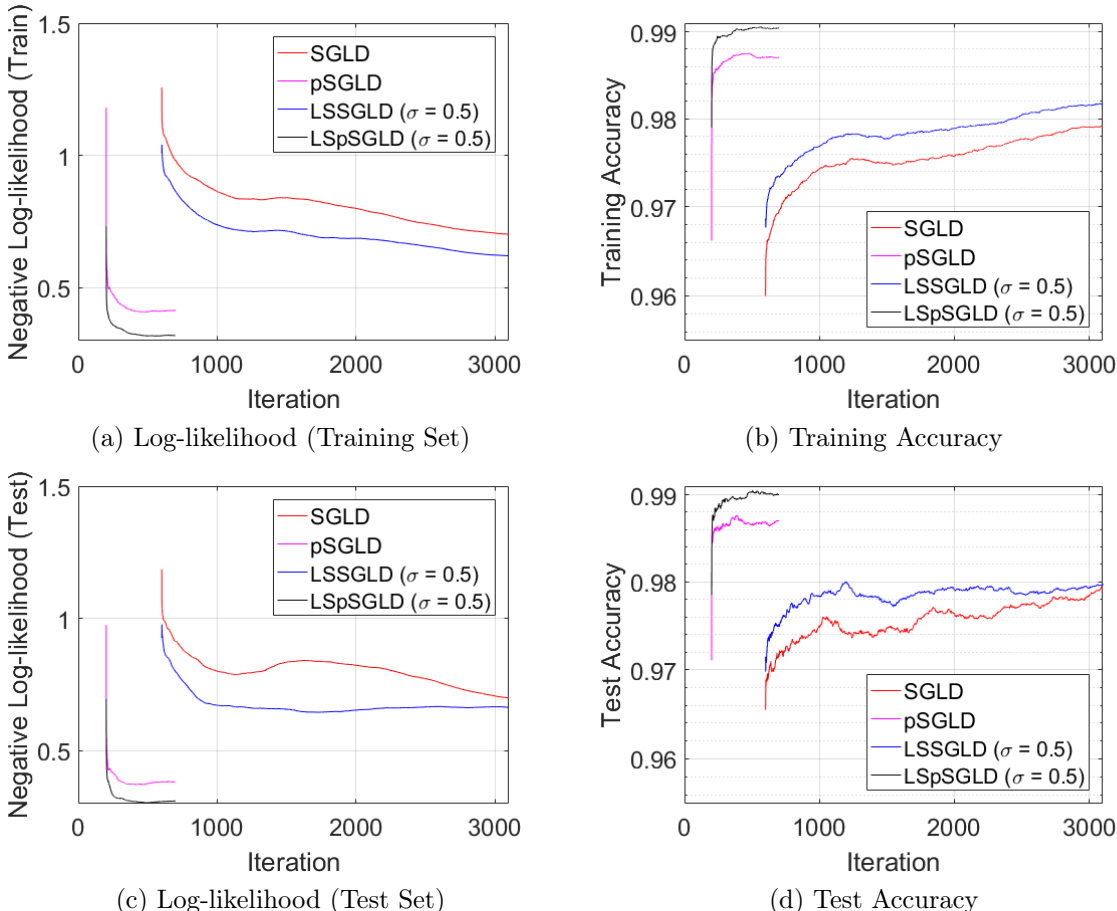
(c) Log-likelihood (Test Set)

(d) Test Accuracy

Figure 5: Convergence comparison for Bayesian convolutional neural network, where the X-axis represents the number of iterations and Y-axis represents the negative log-likelihood/accuracy. (a) Negative log-likelihood on training dataset; (b) Accuracy on training dataset; (c) Negative log-likelihood on test dataset; (d) Accuracy on test dataset.

14

5 Conclusions

In this paper, we integrate Laplacian smoothing with Stochastic Gradient Langevin Dynamics (SGLD) to reduce the gap between the sample and target distributions. The resulting algorithm also allows us to take a larger step size. The proposed algorithm is simple to implement and the extra computation and memory costs compared with the SGLD are negligible when the Fast Fourier Transform (FFT)-based algorithms is employed to resolve the dynamics of the resulting Laplacian Smoothing SGLD (LS-SGLD). We show, both theoretically and empirically, that LS-SGLD can improve the sample quality. It is straightforward to extend Laplacian smoothing to the other Markov Chain Monte Carlo (MCMC) algorithms, e.g., the stochastic gradient Hamiltonian Monte Carlo (Chen et al., 2014).

Acknowledgments

A   Missing proof in Section 2

In this section, we provide the proof of Proposition 1.

Proof of Proposition 1 . Let $p(\boldsymbol{x}, t)$ be the distribution of $\boldsymbol{X}_t$. Then we know that $p(\boldsymbol{x}, t)$ satisfies the following Fokker-Planck equation

$$\frac{\partial p(\boldsymbol{x}, t)}{\partial t} = \frac{1}{\beta}\langle \mathbf{A}_\sigma^{-1}, \nabla^2 p(\boldsymbol{x}, t)\rangle + \langle \nabla, p(\boldsymbol{x}, t)\mathbf{A}_\sigma^{-1}\nabla f(\boldsymbol{x})\rangle$$
$$= \frac{1}{\beta}\langle \nabla, \mathbf{A}_\sigma^{-1}\nabla p(\boldsymbol{x}, t)\rangle + \langle \nabla, p(\boldsymbol{x}, t)\mathbf{A}_\sigma^{-1}\nabla f(\boldsymbol{x})\rangle, \tag{18}$$

where $\langle \nabla, \mathbf{h}(\boldsymbol{x})\rangle$ denotes the divergence of the vector field $\mathbf{h}(\boldsymbol{x})$. Since the stationary distribution π satisfies $\partial \pi / \partial t = 0$, we have

$$\frac{1}{\beta}\langle \nabla, \mathbf{A}_\sigma^{-1}\nabla p(\boldsymbol{x}, t)\rangle + \langle \nabla, p(\boldsymbol{x}, t)\mathbf{A}_\sigma^{-1}\nabla f(\boldsymbol{x})\rangle = 0,$$

which further implies that $\beta^{-1}\nabla \pi + \pi \nabla f(\boldsymbol{x}) = 0$. Solving this equation directly gives $\pi \propto e^{-\beta f(\boldsymbol{x})}$, which completes the proof. $\qquad\square$

B   Proof of Main Theory

In order to bound sampling error between the distribution of the output of LS-SGLD and the target distribution $\pi \propto e^{-\beta f(\boldsymbol{x})}$, we consider a reference sequence generated by LS-LD (6), denoted by $\{\boldsymbol{X}_t\}_{t \geq 0}$. Let $\boldsymbol{X}_0 = \boldsymbol{x}_0$, by triangle inequality, we can decompose the 2-Wasserstein distance $\mathcal{W}_2(\mathbb{P}(\boldsymbol{x}_K), \pi)$ as follows

$$\mathcal{W}_2(\mathbb{P}(\boldsymbol{x}_K), \pi) \leq \mathcal{W}_2(\mathbb{P}(\boldsymbol{x}_K), \mathbb{P}(\boldsymbol{X}_{K\eta})) + \mathcal{W}_2(\mathbb{P}(\boldsymbol{X}_{K\eta}), \pi).$$

The first term on the R.H.S. stands for the discretization error of the numerical integrator, and the second term denotes the ergodicity of LS-LD (6), which characterizes the mixing time of LS-LD. In what follows, we first deliver the following lemma that characterizes the error term $\mathcal{W}_2(\mathbb{P}(\boldsymbol{X}_{K\eta}), \pi)$.

Lemma 5. *Under Assumptions 1 and 2, there exists a constant $c_0 \in [\|\mathbf{A}_\sigma\|_2^{-1}, 1]$*

$$\mathcal{W}_2\big(\mathbb{P}(\boldsymbol{X}_{K\eta}), \pi\big) \leq \big[2\lambda\big(\beta f(0) + \log(\Lambda)\big)\big]^{1/2} \cdot e^{-c_0 K\eta/(2\beta\lambda)},$$

where $\Lambda = \int_{\mathbb{R}^d} e^{-\beta f(\boldsymbol{x})}\, d\boldsymbol{x}$ and λ denotes the logarithmic Sobolev constant of the target distribution $\pi \propto e^{-\beta f(\boldsymbol{x})}$.

Note that Lemma 5 does not require that the target density is log-concave, which can be utilized to prove the convergence rate of LS-SGLD for sampling both log-concave and non-log-concave densities. In the following, we are going to complete the proofs of Theorems 1 and 2.

B.1   Proof of Theorem 1

We first provide the following lemma which proves an upper bound of the discretization error $\mathcal{W}_2\big(\mathbb{P}(\boldsymbol{x}_K), P(\boldsymbol{X}_{K\eta})\big)$ for sampling log-concave densities.

Lemma 6. *Under Assumptions 1, 2, 3 and 4, if set the step size $\eta \leq Cm\beta^{-1}/M^2$ for some sufficiently small constant C, there exist constants $\gamma_1 \in [\|\mathbf{A}_\sigma\|_2^{-2}, 1]$ and $\gamma_2 = d^{-1}\sum_{i=1}^d (1 + 2\sigma - 2\sigma\cos(2\pi i/d))^{-1}$ such that the following holds*

$$\mathcal{W}_2\big(\mathbb{P}(\boldsymbol{x}_K), P(\boldsymbol{X}_{K\eta})\big) \leq \left(\frac{2\gamma_1 d\omega^2 K\eta^2}{B}\right)^{1/2} + \big[8\gamma_2 K(K+1)\beta^{-1}d\eta^3\big]^{1/2}$$

Then we can complete the proof of Theorem 1 as follows.

Proof of Theorem 1. By triangle inequality and Lemmas 5 and 6, it is evident that

$$\mathcal{W}_2(\mathbb{P}(\boldsymbol{x}_K), \pi) \leq \mathcal{W}_2(\mathbb{P}(\boldsymbol{x}_K), \mathbb{P}(\boldsymbol{X}_{K\eta})) + \mathcal{W}_2(\mathbb{P}(\boldsymbol{X}_{K\eta}), \pi)$$

$$\leq \left(\frac{2\gamma_1 d\omega^2 K\eta^2}{B}\right)^{1/2} + \left[8\gamma_2 K(K+1)\beta^{-1} d\eta^3\right]^{1/2} + \left[2\lambda(\beta f(0) + \log(\Lambda))\right]^{1/2} \cdot e^{-c_0 K\eta/(2\beta\lambda)},$$

which completes the proof.

$\square$

B.2   Proof of Theorem 2

Similar to the proof of Theorem 1, we provide the following lemma that characterizes the discretization error $\mathcal{W}_2(\mathbb{P}(\boldsymbol{x}_k), \mathbb{P}(\boldsymbol{X}_{k\eta}))$ for sampling from non-log-concave densities.

Lemma 7. *Under Assumptions 1 and 2, if set the step size $\eta \leq Cm\beta^{-1}/M^2$ for some sufficiently small constant C, there exist constants $\gamma_1 \in [\|\mathbf{A}_\sigma\|_2^{-2}, 1]$, $\gamma_2 = d^{-1}\sum_{i=1}^d (1 + 2\sigma - 2\sigma\cos(2\pi i/d))^{-1}$ and $\bar{\Gamma} = (3/2 + 2(b + \beta^{-1}d))^{1/2}$ such that the following holds,*

$$\mathcal{W}_2(\mathbb{P}(\boldsymbol{x}_K), \mathbb{P}(\boldsymbol{X}_{K\eta})) \leq \bar{\Gamma}(K\eta)^{1/2}\left[\left(\frac{\gamma_1\beta d\omega^2}{2B}K\eta + 2\gamma_2\beta M^2 dK\eta^2\right)^{1/2} + \left(\frac{\gamma_1\beta d\omega^2}{2B}K\eta + 2\gamma_2\beta M^2 dK\eta^2\right)^{1/4}\right].$$

Proof of Theorem 2. By triangle inequality and Lemmas 5 and 7, it is evident that

$$\mathcal{W}_2(\mathbb{P}(\boldsymbol{x}_K), \pi) \leq \mathcal{W}_2(\mathbb{P}(\boldsymbol{x}_K), \mathbb{P}(\boldsymbol{X}_{K\eta})) + \mathcal{W}_2(\mathbb{P}(\boldsymbol{X}_{K\eta}), \pi)$$

$$\leq \bar{\Gamma}(K\eta)^{1/2}\left[\left(\frac{\gamma_1\beta d\omega^2}{2B}K\eta + 2\gamma_2 M^2 dK\eta^2\right)^{1/2} + \left(\frac{\gamma_1\beta d\omega^2}{2B}K\eta + 2\gamma_2 M^2 dK\eta^2\right)^{1/4}\right]$$

$$+ \left[2\lambda(\beta f(0) + \log(\Lambda))\right]^{1/2} \cdot e^{-c_0 K\eta/(2\beta\lambda)},$$

which completes the proof.

$\square$

C   Proof of lemmas in Appendix B

C.1   Proof of Lemma 5

In order to prove Lemma 5, we require the following lemma.

Lemma 8 (Theorem 9.6.1 in Bakry et al. (2013)). *Suppose the target density π satisfies logarithmic Sobolev inequality with a positive constant λ, for any density μ it holds that*

$$\mathcal{W}_2(\mu, \pi) \leq \sqrt{2\lambda D_{KL}(\mu\|\pi)}.$$

Proof of Lemma 5. Recall that for LS-LD (6), the distribution of $\boldsymbol{X}_t$, denoted by $p(\boldsymbol{x}, t)$, can be described by the following Fokker-Planck equation

$$\frac{\partial p(\boldsymbol{x}, t)}{\partial t} = \frac{1}{\beta}\langle \mathbf{A}_\sigma^{-1}, \nabla^2 p(\boldsymbol{x}, t)\rangle + \langle \nabla, p(\boldsymbol{x}, t)\mathbf{A}_\sigma^{-1}\nabla f(\boldsymbol{x})\rangle$$

$$= \frac{1}{\beta}\langle \nabla, \mathbf{A}_\sigma^{-1}\nabla p(\boldsymbol{x}, t)\rangle + \langle \nabla, p(\boldsymbol{x}, t)\mathbf{A}_\sigma^{-1}\nabla f(\boldsymbol{x})\rangle, \tag{19}$$

where $\langle\nabla, \mathbf{h}(\boldsymbol{x})\rangle$ denotes the divergence of the vector field $\mathbf{h}(\boldsymbol{x})$. Let $\mathbb{P}_t$ be the short-hand notation of $p(\boldsymbol{x}, t)$, and denote by $D_{KL}(\mathbb{P}_t\|\pi)$ the KL-divergence between the distribution $\mathbb{P}_t$ and the target distribution π. Then, we have

$$\frac{\mathrm{d}D_{KL}(\mathbb{P}_t\|\pi)}{\mathrm{d}t} = \int_{\mathbb{R}^d} \frac{\partial}{\partial t}\left[\mathbb{P}_t \log\left(\frac{\mathbb{P}_t}{\pi}\right)\right]\mathrm{d}\boldsymbol{x}$$

$$= \int_{\mathbb{R}^d} \frac{\partial \mathbb{P}_t}{\partial t}\left[\log(\mathbb{P}_t) + 1 - \log(\pi)\right]\mathrm{d}\boldsymbol{x}.$$

Similar to the proof of Proposition 2 in Mou et al. (2017), by (19) we further have

$$\frac{\mathrm{d}D_{KL}(\mathbb{P}_t||\pi)}{\mathrm{d}t} = -\int_{\mathbb{R}^d} \left\langle \frac{1}{\beta}\mathbf{A}_\sigma^{-1}\nabla\mathbb{P}_t + \mathbb{P}_t\mathbf{A}_\sigma^{-1}\nabla f(\boldsymbol{x}), \nabla\log(\mathbb{P}_t) - \nabla\log(\pi) \right\rangle \mathrm{d}\boldsymbol{x}$$

$$= -\int_{\mathbb{R}^d} \left\langle \mathbf{A}_\sigma^{-1}\left(\frac{1}{\beta}\mathbb{P}_t\nabla\log(\mathbb{P}_t) + \mathbb{P}_t\nabla f(\boldsymbol{x})\right), \nabla\log(\mathbb{P}_t) - \nabla\log(\pi) \right\rangle \mathrm{d}\boldsymbol{x},$$

where the second equality holds due to $\nabla\mathbb{P}_t = \mathbb{P}_t\nabla\log(\mathbb{P}_t)$. In addition, note that $\pi \propto e^{-\beta f(\boldsymbol{x})}$, we have $\nabla\log(\pi) = -\beta\nabla f(\boldsymbol{x})$. Then we have

$$\frac{\mathrm{d}D_{KL}(\mathbb{P}_t||\pi)}{\mathrm{d}t} = -\frac{1}{\beta}\int_{\mathbb{R}_d} \left\langle \mathbf{A}_\sigma^{-1}\big(\nabla\log(\mathbb{P}_t) - \nabla\log(\pi)\big), \nabla\log(\mathbb{P}_t) - \nabla\log(\pi)\right\rangle\mathbb{P}_t\mathrm{d}\boldsymbol{x}$$

$$= -\frac{1}{\beta}\int_{\mathbb{R}_d} \|\nabla\log(\mathbb{P}_t) - \nabla\log(\pi)\|_{\mathbf{A}_\sigma^{-1}}^2 \mathbb{P}_t\mathrm{d}\boldsymbol{x}.$$

Since $\mathbf{A}_\sigma$ is a positive definite matrix, there exists a constant $c_0 \in [\|\mathbf{A}_\sigma\|_2^{-1}, 1]$ such that

$$\frac{\mathrm{d}D_{KL}(\mathbb{P}_t||\pi)}{\mathrm{d}t} \leq -\frac{c_0}{\beta}\int_{\mathbb{R}_d} \|\nabla\log(\mathbb{P}_t) - \nabla\log(\pi)\|_2^2 \mathbb{P}_t\mathrm{d}\boldsymbol{x} = -\frac{c_0}{\beta}\mathbf{I}(\mathbb{P}_t||\pi), \tag{20}$$

where $\mathbf{I}(\mathbb{P}_t||\pi)$ denotes the fisher information between $\mathbb{P}_t$ and π. By Proposition 2, we know that the target density π satisfies logarithmic Sobolev inequality with constant $\lambda > 0$. Then, from Markowich and Villani (1999), we have

$$D_{KL}(\mathbb{P}_t||\pi) \leq \frac{1}{\lambda}\mathbf{I}(\mathbb{P}_t||\pi).$$

Plugging the above inequality into (20), we obtain

$$\frac{\mathrm{d}D_{KL}(\mathbb{P}_t||\pi)}{\mathrm{d}t} \leq -\frac{c_0}{\lambda\beta}D_{KL}(\mathbb{P}_t||\pi),$$

which implies that

$$D_{KL}(\mathbb{P}_t||\pi) \leq D_{KL}(\mathbb{P}_0||\pi)e^{-c_0 t/(\beta\lambda)}.$$

Note that we have $\mathbb{P}_0 = \delta(0)$, where $\delta(\cdot)$ is the Dirac delta function, thus,

$$D_{KL}(\mathbb{P}_0||\pi) = \int_{\mathbb{R}^d} \mathbb{P}_0\big[\log(\mathbb{P}_0) - \log(\pi)\big]\mathrm{d}\boldsymbol{x} = -\log(\pi)|_{\boldsymbol{x}=0} = \beta f(0) + \log(\Lambda),$$

where $\Lambda = \int_{\mathbb{R}^d} e^{-\beta f(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}$. Then by Lemma 8, we have the following regarding the 2-Wasserstein distance $\mathcal{W}_2\big(\mathbb{P}(\boldsymbol{X}_{k\eta}), \pi\big)$,

$$\mathcal{W}_2\big(\mathbb{P}(\boldsymbol{X}_{k\eta}), \pi\big) \leq \sqrt{2\lambda D_{KL}(\mathbb{P}(\boldsymbol{X}_0)||\pi)} \cdot e^{-c_0 t/(2\beta\lambda)} = \big[2\lambda\big(\beta f(0) + \log(\Lambda)\big)\big]^{1/2} \cdot e^{-c_0 t/(2\beta\lambda)},$$

which completes the proof. □

C.2 Proof of Lemma 6

We first deliver the following useful lemmas.

Lemma 9. *Consider any two LS-LD sequences $\{\boldsymbol{W}_t\}_{t\geq 0}$ and $\{\boldsymbol{V}_t\}_{t\geq 0}$, and assume that $\boldsymbol{W}_t$ and $\boldsymbol{V}_t$ have shared Brownian motion terms. Under Assumption 4, for any $t > 0$ it holds that,*

$$\mathbb{E}[\|\boldsymbol{W}_t - \boldsymbol{V}_t\|_{\mathbf{A}_\sigma}^2] \leq \mathbb{E}[\|\boldsymbol{W}_0 - \boldsymbol{V}_0\|_{\mathbf{A}_\sigma}^2].$$

18

Lemma 10. *Under Assumptions 1 and 2, if set the step size $\eta \leq Cm\beta^{-1}/M^2$ for some sufficiently small constant C, there exists a constant $\gamma_2 = d^{-1}\sum_{i=1}^{d}(1 + 2\sigma - 2\sigma\cos(2\pi i/d))^{-1}$ such that for any $\boldsymbol{x}_k$ with $k \geq 0$,*

$$\mathbb{E}[\|\mathcal{L}_\eta \boldsymbol{x}_k - \mathcal{G}_\eta \boldsymbol{x}_k\|_{\mathbf{A}_\sigma}^2] \leq 4\gamma_2\beta^{-1}d\eta^3.$$

Now we are ready to complete the proof of Lemma 6.

Proof of Lemma 6. For the sake of simplicity, we first define three operators $\mathcal{L}_t$, $\mathcal{G}_t$ and $\mathcal{S}_t$ as follows: for any $\boldsymbol{x} \in \mathbb{R}^d$ we denote by $\mathcal{L}_t\boldsymbol{x}$ the random point generated by LS-LD at time t starting from $\boldsymbol{x}$, $\mathcal{G}_t\boldsymbol{x}$ the point after performing one-step LS-SGLD with full gradient at $\boldsymbol{x}$ with step size t, and $\mathcal{S}_t\boldsymbol{x}$ the point after performing one-step LS-SGLD at $\boldsymbol{x}$ with step size t. Then we have

$$
\begin{aligned}
\mathbb{E}[\|\boldsymbol{x}_K - \boldsymbol{X}_{K\eta}\|_{\mathbf{A}_\sigma}^2] &= \mathbb{E}[\|\boldsymbol{x}_K - \mathcal{G}_\eta\boldsymbol{x}_{K-1} + \mathcal{G}_\eta\boldsymbol{x}_{K-1} - \boldsymbol{X}_{K\eta}\|_{\mathbf{A}_\sigma}^2] \\
&= \mathbb{E}[\|\boldsymbol{x}_K - \mathcal{G}_\eta\boldsymbol{x}_{K-1}\|_{\mathbf{A}_\sigma}^2] + \mathbb{E}[\|\mathcal{G}_\eta\boldsymbol{x}_{K-1} - \boldsymbol{X}_{K\eta}\|_{\mathbf{A}_\sigma}^2],
\end{aligned}
\tag{21}
$$

where the second equality follows from the fact that $\mathbb{E}[\langle \boldsymbol{x}_K - \mathcal{G}_\eta\boldsymbol{x}_{K-1}, \mathbf{A}_\sigma(\mathcal{G}_\eta\boldsymbol{x}_{K-1} - \boldsymbol{X}_{K\eta})\rangle] = 0$ since at any iteration the randomness of stochastic gradient is independent of the iterate. Regarding the first term on the R.H.S. of (21), we have

$$
\begin{aligned}
\mathbb{E}[\|\boldsymbol{x}_K - \mathcal{G}_\eta\boldsymbol{x}_{K-1}\|_{\mathbf{A}_\sigma}^2] &= \eta^2 \mathbb{E}[\|\mathcal{S}_\eta\boldsymbol{x}_{K-1} - \mathcal{G}_\eta\boldsymbol{x}_{K-1}\|_{\mathbf{A}_\sigma}^2] \\
&\leq \eta^2 \mathbb{E}[\|\mathbf{A}_\sigma^{-1}\mathbf{g}_{K-1} - \mathbf{A}_\sigma^{-1}\nabla f(\boldsymbol{x}_{K-1})\|_{\mathbf{A}_\sigma}^2] \\
&\leq \frac{\eta^2}{B}\mathbb{E}[\|\mathbf{A}_\sigma^{-1}\nabla f_i(\boldsymbol{x}_{K-1}) - \mathbf{A}_\sigma^{-1}\nabla f(\boldsymbol{x}_{K-1})\|_{\mathbf{A}_\sigma}^2] \\
&\leq \frac{\gamma_1\eta^2 d\omega^2}{B},
\end{aligned}
\tag{22}
$$

where $\gamma_1 \in [\|\mathbf{A}_\sigma\|_2^{-1}, 1)$ is a problem-dependent parameter, the first inequality follows the definitions of operators $\mathcal{S}_\eta$ and $\mathcal{G}_\eta$, the second inequality follows from Lemma A.1 in Lei et al. (2017) and the last inequality is by Assumption 3. In terms of the second term on the R.H.S. of (21), we have

$$
\begin{aligned}
\mathbb{E}[\|\mathcal{G}_\eta\boldsymbol{x}_{K-1} - \boldsymbol{X}_{K\eta}\|_{\mathbf{A}_\sigma}^2] &= \mathbb{E}[\|\mathcal{G}_\eta\boldsymbol{x}_{K-1} - \mathcal{L}_\eta\boldsymbol{x}_{K-1} + \mathcal{L}_\eta\boldsymbol{x}_{K-1} - \boldsymbol{X}_{K\eta}\|_{\mathbf{A}_\sigma}^2] \\
&\leq (1+\alpha)\mathbb{E}[\|\mathcal{G}_\eta\boldsymbol{x}_{K-1} - \mathcal{L}_\eta\boldsymbol{x}_{K-1}\|_{\mathbf{A}_\sigma}^2] + (1+1/\alpha)\mathbb{E}[\|\mathcal{L}_\eta\boldsymbol{x}_{K-1} - \mathcal{L}_\eta\boldsymbol{X}_{(K-1)\eta}\|_{\mathbf{A}_\sigma}^2] \\
&\leq 4(1+\alpha)\gamma_2\beta^{-1}d\eta^3 + (1+1/\alpha)\mathbb{E}[\|\boldsymbol{x}_{K-1} - \boldsymbol{X}_{(K-1)\eta}\|_{\mathbf{A}_\sigma}^2],
\end{aligned}
\tag{23}
$$

where α is a positive constant that will be specified later, the first inequality is by Young's inequality, and the second inequality follows from Lemmas 9 and 10. Plugging (22) and (23) into (21) gives

$$\mathbb{E}[\|\boldsymbol{x}_K - \boldsymbol{X}_{K\eta}\|_{\mathbf{A}_\sigma}^2] \leq 4(1+\alpha)\gamma_2\beta^{-1}d\eta^3 + \frac{\gamma_1\eta^2 d\omega^2}{B} + (1+1/\alpha)\mathbb{E}[\|\boldsymbol{x}_{K-1} - \boldsymbol{X}_{(K-1)\eta}\|_{\mathbf{A}_\sigma}^2].$$

Then, by recursively applying the above inequality, we obtain

$$
\begin{aligned}
\mathbb{E}[\|\boldsymbol{x}_K - \boldsymbol{X}_{K\eta}\|_{\mathbf{A}_\sigma}^2] &\leq (1+1/\alpha)^K \mathbb{E}[\|\boldsymbol{x}_0 - \boldsymbol{X}_0\|_{\mathbf{A}_\sigma}^2] + \sum_{k=0}^{K-1}(1+1/\alpha)^k\left[4(1+\alpha)\gamma_2\beta^{-1}d\eta^3 + \frac{\gamma_1\eta^2 d\omega^2}{B}\right] \\
&= \alpha\left[(1+1/\alpha)^K - 1\right]\cdot\left[4(1+\alpha)\gamma_2\beta^{-1}d\eta^3 + \frac{\gamma_1\eta^2 d\omega^2}{B}\right].
\end{aligned}
$$

Let $\alpha = K$ and apply the inequality $(1+1/K)^K - 1 \leq e - 1 \leq 2$, the above inequality implies

$$\mathbb{E}[\|\boldsymbol{x}_K - \boldsymbol{X}_{K\eta}\|_{\mathbf{A}_\sigma}^2] \leq 2K\eta^2\cdot\left[\frac{\gamma_1 d\omega^2}{B} + 4(K+1)\gamma_2\beta^{-1}d\eta\right].$$

Based on the definition of 2-Wasserstein distance, we have

$$\mathcal{W}_2^2\big(\mathbb{P}(\boldsymbol{x}_K), \mathbb{P}(\boldsymbol{X}_{K\eta})\big) \leq \sqrt{\mathbb{E}[\|\boldsymbol{x}_K - \boldsymbol{X}_{K\eta}\|_2^2]} \leq \sqrt{\mathbb{E}[\|\boldsymbol{x}_K - \boldsymbol{X}_{K\eta}\|_{\mathbf{A}_\sigma}^2]} \leq \left(\frac{2\gamma_1 d\omega^2 K\eta^2}{B}\right)^{1/2} + \left[8\gamma_2 K(K+1)\beta^{-1}d\eta^3\right]^{1/2},$$

where the last inequality is by the fact that $\sqrt{x^2 + y^2} \leq |x| + |y|$. This completes the proof.

$\square$

C.3 Proof of Lemma 7

In order to prove Lemma 7, we require the following lemmas.

Lemma 11. *Under Assumptions 1 and 2, for all $k \geq 0$, there exists a constant $c_1 \in [\|\mathbf{A}_\sigma\|_2^{-1}, 1)$ such that*

$$\mathbb{E}[\|\boldsymbol{x}_k\|_2^2] \leq \mathbb{E}[\|\boldsymbol{x}_k\|_{\mathbf{A}_\sigma}^2] \leq \frac{2(2b + \beta^{-1}d)}{c_1 m}.$$

Lemma 12 (Theorem 2.3 in Bolley and Villani (2005))**.** *Let μ, ν be two probability measures with finite exponential second moments, it holds that*

$$\mathcal{W}_2(\mu, \nu) \leq \Gamma\left[\sqrt{D_{KL}(\mu||\nu)} + \left[D_{KL}(\mu||\nu)\right]^{1/4}\right],$$

where

$$\Gamma = \inf_{\alpha > 0}\left(\frac{1}{\alpha}\left(\frac{3}{2} + \log\mathbb{E}_\nu[e^{\alpha\|\boldsymbol{x}\|_2^2}]\right)\right)^{1/2}.$$

Lemma 13 (Lemma 4 in Wang et al. (2019))**.** *Let $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ be the standard Gaussian random vector with dimension d, it holds that*

$$\mathbb{E}\left[\|\mathbf{A}_\sigma^{-1}\boldsymbol{\epsilon}\|_2^2\right] = \sum_{i=1}^d \frac{1}{\left(1 + 2\sigma - 2\sigma\cos(2\pi i/d)\right)^2}.$$

Lemma 14. *Under Assumptions 1 and 2, let $\boldsymbol{X}_t$ denote the solution of LS-LD (6) at time t with initial point $\boldsymbol{X}_0 = 0$. Then if the inverse temperature satisfies $\beta \geq 2\|\mathbf{A}_\sigma\|_2/m$, it holds that*

$$\mathbb{E}[e^{\|\boldsymbol{X}_t\|_2^2}] \leq e^{2(b + \beta^{-1}d)t}.$$

Based on the above lemmas, we are able to complete the proof of Lemma 7.

Proof of Lemma 7. By Lemma 12, we know that the 2-Wasserstein distance between any two probability measures can be bounded by their KL divergence. Therefore, the remaining part will focus on deriving the upper bound of the KL divergence $D_{KL}\big(\mathbb{P}(\boldsymbol{x}_k)||\mathbb{P}(\boldsymbol{X}_{k\eta})\big)$. Similar to the proof technique used in Dalalyan (2017); Raginsky et al. (2017); Xu et al. (2018), we leverage the following continuous-time interpolation of LS-SGLD

$$\tilde{\boldsymbol{X}}_t = \int_0^t -\mathbf{A}_\sigma^{-1}\mathbf{G}_s \mathrm{d}s + \int_0^t \sqrt{2\beta^{-1}}\mathbf{A}_\sigma^{-1/2}\mathrm{d}\boldsymbol{B}_s, \tag{24}$$

where $\mathbf{G}_t = \sum_{k=0}^\infty \mathbf{g}_k \mathbb{1}\{t \in [k\eta, (k+1)\eta)\}$. It can be easily verified that $\tilde{\boldsymbol{X}}_{k\eta}$ follows the same distribution as $\boldsymbol{x}_k$. However, it is worth noting that (24) does not form a Markov chain since it contains randomness of the stochastic gradient. To tackle this, we leverage the results in Gyöngy (1986) and construct the following Markov chain to mimic (24),

$$\hat{\boldsymbol{X}}_t = \int_0^t -\mathbf{A}_\sigma^{-1}\hat{\mathbf{G}}_s \mathrm{d}s + \int_0^t \sqrt{2\beta^{-1}}\mathbf{A}_\sigma^{-1/2}\mathrm{d}\boldsymbol{B}_s,$$

where $\hat{\mathbf{G}}_s = \mathbb{E}[\mathbf{G}_s|\hat{\boldsymbol{X}}_s = \tilde{\boldsymbol{X}}_s]$. It was shown that $\hat{\boldsymbol{X}}_t$ and $\tilde{\boldsymbol{X}}_t$ has the same one-time marginal distribution (Gyöngy, 1986). Then let $\mathbb{P}_t$ and $\mathbb{Q}_t$ denote the distribution of $\boldsymbol{X}_t$ and $\hat{\boldsymbol{X}}_t$ respectively, by Girsanov formula, the Radon-Nikodym derivative of $\mathbb{P}_t$ with respect to $\mathbb{Q}_t$ can be derived as follows,

$$\frac{\mathrm{d}\mathbb{P}_t}{\mathrm{d}\mathbb{Q}_t} = \exp\left\{\frac{\beta}{2}\int_0^t \langle\nabla f(\hat{\boldsymbol{X}}_s) - \hat{\mathbf{G}}_s, \mathbf{A}_\sigma^{-1/2}\mathrm{d}\boldsymbol{B}_s\rangle - \frac{\beta}{4}\int_0^t \|\mathbf{A}_\sigma^{-1}\nabla f(\hat{\boldsymbol{X}}_s) - \mathbf{A}_\sigma^{-1}\hat{\mathbf{G}}_s\|_2^2\mathrm{d}s\right\}.$$

Therefore, let $T = K\eta$, the KL divergence $D_{KL}(\mathbb{P}_T||\mathbb{Q}_T)$ satisfies

$$D_{KL}(\mathbb{Q}_T||\mathbb{P}_T) = -\int_{\mathbb{R}^d} \log\left(\frac{\mathrm{d}\mathbb{P}_T}{\mathrm{d}\mathbb{Q}_T}\right)\mathrm{d}\mathbb{Q}_T$$

$$= \frac{\beta}{4}\int_0^T \mathbb{E}\big[\|\mathbf{A}_\sigma^{-1}\nabla f(\hat{\boldsymbol{X}}_s) - \mathbf{A}_\sigma^{-1}\hat{\mathbf{G}}_s\|_2^2\big]\mathrm{d}s$$

$$= \frac{\beta}{4}\sum_{k=0}^{K-1}\int_{k\eta}^{(k+1)\eta} \mathbb{E}\big[\|\mathbf{A}_\sigma^{-1}\nabla f(\tilde{\boldsymbol{X}}_s) - \mathbf{A}_\sigma^{-1}\mathbf{g}_k\|_2^2\big]\mathrm{d}s,$$

where the second equality holds due to $\mathbb{E}[\langle \nabla f(\hat{\boldsymbol{X}}_s) - \hat{\mathbf{G}}_s, \mathbf{A}_\sigma^{-1/2}\mathrm{d}\boldsymbol{B}_s\rangle] = 0$ and the second equality follows from the fact that $\hat{\boldsymbol{X}}_s$ and $\tilde{\boldsymbol{X}}_s$ follow the same distribution. Using Young's inequality, we have

$$\mathbb{E}[\|\mathbf{A}_\sigma^{-1}\nabla f(\tilde{\boldsymbol{X}}_s) - \mathbf{A}_\sigma^{-1}\mathbf{g}_k\|_2^2] \leq 2\underbrace{\mathbb{E}[\|\mathbf{A}_\sigma^{-1}\nabla f(\tilde{\boldsymbol{X}}_s) - \mathbf{A}_\sigma^{-1}\nabla f(\boldsymbol{x}_k)\|_2^2]}_{I_1} + 2\underbrace{\mathbb{E}[\|\mathbf{A}_\sigma^{-1}\nabla f(\boldsymbol{x}_k) - \mathbf{A}_\sigma^{-1}\mathbf{g}_k\|_2^2]}_{I_2}.$$

Then we are going to tackle I_1 and I_2 separately. Note that $\|\mathbf{A}_\sigma^{-1}\|_2 \leq 1$, thus by Assumption 2, we have the following for I_1,

$$I_1 \leq \mathbb{E}[\|\nabla f(\tilde{\boldsymbol{X}}_s) - \nabla f(\boldsymbol{x}_k)\|_2^2] \leq M^2\mathbb{E}[\|\tilde{\boldsymbol{X}}_s - \boldsymbol{x}_k\|_2^2].$$

Based on the definition of $\tilde{\boldsymbol{X}}_s$, we have $\tilde{\boldsymbol{X}}_s - \boldsymbol{x}_k = (s-k\eta)\mathbf{A}_\sigma^{-1}\mathbf{g}_k + \sqrt{2\beta^{-1}(s-k\eta)}\mathbf{A}_\sigma^{-1/2}\boldsymbol{\epsilon}_k$. Since $s-k\eta \leq \eta$, it follows that

$$I_1 \leq M^2\mathbb{E}[\|\tilde{\boldsymbol{X}}_s - \boldsymbol{x}_k\|_2^2] \leq \eta^2 M^2\mathbb{E}[\|\mathbf{A}_\sigma^{-1}\mathbf{g}_k\|_2^2] + 2\eta M^2\beta^{-1}\mathbb{E}[\|\mathbf{A}_\sigma^{-1/2}\boldsymbol{\epsilon}_k\|_2^2].$$

Regarding I_2, based on Lemma A.1 in Lei et al. (2017) and Assumption 3, we have

$$I_2 \leq \frac{1}{B}\mathbb{E}[\|\mathbf{A}_\sigma^{-1}\nabla f(\boldsymbol{x}_k) - \mathbf{A}_\sigma^{-1}\nabla f_i(\boldsymbol{x}_k)\|_2^2] \leq \frac{\gamma_1 d\omega^2}{B},$$

where $\gamma_1 \in [\|\mathbf{A}_\sigma\|_2^{-2}, 1)$ is a problem-dependent parameter. Putting everything together, we have

$$D_{KL}(\mathbb{Q}_T||\mathbb{P}_T) \leq \sum_{k=0}^{K-1}\eta\bigg\{\frac{\beta}{2}\eta^2 M^2\mathbb{E}[\|\mathbf{g}_k\|_{\mathbf{A}_\sigma^{-2}}^2] + \eta M^2\mathbb{E}[\|\boldsymbol{\epsilon}_k\|_{\mathbf{A}_\sigma^{-1}}^2] + \frac{\gamma_1\beta d\omega^2}{2B}\bigg\}.$$

By Lemma 15 and Young's inequality, we know that

$$\mathbb{E}[\|\mathbf{g}_k\|_{\mathbf{A}_\sigma^{-2}}^2] \leq \mathbb{E}[\|\mathbf{g}_k\|_2^2] \leq 2M^2\mathbb{E}[\|\boldsymbol{x}_k\|_2^2] + 2G^2 \leq \frac{4M^2(2b+\beta^{-1}d)}{c_1 m} + 2G^2,$$

where $G = \max_{i\in[n]}\|\nabla f_i(0)\|_2$. Then by Lemma 13, we have $\mathbb{E}[\|\boldsymbol{\epsilon}_k\|_{\mathbf{A}_\sigma^{-1}}^2] \leq \gamma_2 d$ with $\gamma_2 = d^{-1}\sum_{i=1}^d(1 + 2\sigma - 2\sigma\cos(2\pi i/d))^{-1}$, which is strictly smaller than 1. Therefore,

$$D_{KL}(\mathbb{Q}_T||\mathbb{P}_T) \leq \frac{2\beta M^4(2b+\beta^{-1}d) + \beta c_1 mM^2 G^2}{c_1 m}K\eta^3 + \gamma_2 M^2 dK\eta^2 + \frac{\gamma_1\beta d\omega^2}{2B}K\eta.$$

For sufficiently small step size such that

$$\eta \leq \frac{c_1\beta^{-1}\gamma_2 md}{2M^2(2b+\beta^{-1}d) + c_1 mG^2},$$

we have

$$D_{KL}(\mathbb{P}(\boldsymbol{x}_K)||\mathbb{P}(\boldsymbol{X}_{K\eta})) \leq \frac{\gamma_1\beta d\omega^2}{2B}K\eta + 2\gamma_2 M^2 dK\eta^2.$$

Then, by Lemma 12, we have

$$\mathcal{W}_2\big(\mathbb{P}(\boldsymbol{x}_K), \mathbb{P}(\boldsymbol{X}_{K\eta})\big) \leq \Gamma\Big[\sqrt{D_{KL}\big(\mathbb{P}(\boldsymbol{x}_K)||\mathbb{P}(\boldsymbol{X}_{K\eta})\big)} + \big[D_{KL}\big(\mathbb{P}(\boldsymbol{x}_K)||\mathbb{P}(\boldsymbol{X}_{K\eta})\big)\big]^{1/4}\Big],$$

where Γ can be further bounded as

$$\Gamma \leq \Big(\frac{3}{2} + \log \mathbb{E}[e^{\|\boldsymbol{X}_{K\eta}\|_2^2}]\Big)^{1/2} \leq \big(3/2 + 2(b+\beta d)K\eta\big)^{1/2} \leq \big(3/2 + 2(b+\beta d)\big)^{1/2} \cdot (K\eta)^{1/2},$$

where the first inequality is by the choice $\alpha = 1$, the second inequality is by Lemma 14 and the last inequality is by the assumption that $K\eta > 1$. Therefore, define by $\bar{\Gamma} = \big(3/2 + 2(b+\beta^{-1}d)\big)^{1/2}$, the 2-Wasserstein distance $\mathcal{W}_2\big(\mathbb{P}(\boldsymbol{x}_K), \mathbb{P}(\boldsymbol{X}_{K\eta})\big)$ can be bounded by

$$\mathcal{W}_2\big(\mathbb{P}(\boldsymbol{x}_K), \mathbb{P}(\boldsymbol{X}_{K\eta})\big) \leq \bar{\Gamma}(K\eta)^{1/2}\Big[\Big(\frac{\gamma_1 \beta d\omega^2}{2B}K\eta + 2\gamma_2 M^2 dK\eta^2\Big)^{1/2} + \Big(\frac{\gamma_1 \beta d\omega^2}{2B}K\eta + 2\gamma_2 M^2 dK\eta^2\Big)^{1/4}\Big],$$

which completes the proof. $\qquad\square$

D Proof of Lemmas in Appendix C

D.1 Proof of Lemma 9

Proof of Lemma 9. Assuming shared Brownian motions in $\boldsymbol{W}_t$ and $\boldsymbol{V}_t$, we have

$$\begin{aligned}
\mathrm{d}\mathbb{E}[\|\boldsymbol{W}_t - \boldsymbol{V}_t\|_{\mathbf{A}_\sigma}^2] &= -\mathbb{E}\big[\big\langle \mathbf{A}_\sigma^{-1}\big(\nabla f(\boldsymbol{W}_t) - \nabla f(\boldsymbol{V}_t)\big), \mathbf{A}_\sigma(\boldsymbol{W}_t - \boldsymbol{V}_t)\big\rangle\big]\mathrm{d}t \\
&= -\mathbb{E}\big[\big\langle \big(\nabla f(\boldsymbol{W}_t) - \nabla f(\boldsymbol{V}_t)\big), \boldsymbol{W}_t - \boldsymbol{V}_t\big\rangle\big]\mathrm{d}t \\
&\geq 0,
\end{aligned}$$

where the first equality follows from the fact that we assume shared Brownian motion terms on both dynamics $\{\boldsymbol{W}_t\}_{t\geq 0}$ and $\{\boldsymbol{V}_t\}_{t\geq 0}$ and the inequality is due to the convexity of $f(\boldsymbol{x})$. Therefore, it can be evidently concluded that

$$\mathbb{E}[\|\boldsymbol{W}_t - \boldsymbol{V}_t\|_{\mathbf{A}_\sigma}^2] \leq \mathbb{E}[\|\boldsymbol{W}_0 - \boldsymbol{V}_0\|_{\mathbf{A}_\sigma}^2],$$

which completes the proof. $\qquad\square$

D.2 Proof of Lemma 10

Proof of Lemma 10. To simplify the analysis, let $\boldsymbol{x}$ be any iterate of LS-SGLD and define $\boldsymbol{x} = \boldsymbol{X}_0$. Then the operators $\mathcal{G}_\eta$ and $\mathcal{L}_\eta$ satisfy

$$\begin{aligned}
\mathcal{G}_\eta \boldsymbol{x} &= \boldsymbol{X}_0 - \eta \mathbf{A}_\sigma^{-1}\nabla f(\boldsymbol{X}_0) + \sqrt{2\beta^{-1}\eta}\mathbf{A}_\sigma^{-1/2}\boldsymbol{\epsilon} \\
&= \boldsymbol{X}_0 - \int_0^\eta \mathbf{A}_\sigma^{-1}\nabla f(\boldsymbol{X}_0)\mathrm{d}t + \int_0^\eta \sqrt{2\beta^{-1}}\mathbf{A}_\sigma^{-1/2}\mathrm{d}\boldsymbol{B}_t; \\
\mathcal{L}_\eta \boldsymbol{x} &= \boldsymbol{X}_0 - \int_0^\eta \mathbf{A}_\sigma^{-1}\nabla f(\boldsymbol{X}_t)\mathrm{d}t + \int_0^\eta \sqrt{2\beta^{-1}}\mathbf{A}_\sigma^{-1/2}\mathrm{d}\boldsymbol{B}_t.
\end{aligned}$$

Consider synchronous Brownian terms in $\mathcal{G}_\eta$ and $\mathcal{L}_\eta$, we have

$$\begin{aligned}
\mathbb{E}[\|\mathcal{L}_\eta \boldsymbol{x} - \mathcal{G}_\eta \boldsymbol{x}\|_{\mathbf{A}_\sigma}^2] &= \mathbb{E}\Big[\Big\|\int_0^\eta \big[\mathbf{A}_\sigma^{-1}\nabla f(\boldsymbol{X}_0) - \mathbf{A}_\sigma^{-1}\nabla f(\boldsymbol{X}_t)\big]\mathrm{d}t\Big\|_{\mathbf{A}_\sigma}^2\Big] \\
&\leq \mathbb{E}\Big[\eta\int_0^\eta \big\|\mathbf{A}_\sigma^{-1}\big[\nabla f(\boldsymbol{X}_0) - \nabla f(\boldsymbol{X}_t)\big]\big\|_{\mathbf{A}_\sigma}^2 \mathrm{d}t\Big] \\
&\leq M^2\Big[\eta\int_0^\eta \mathbb{E}[\|\boldsymbol{X}_t - \boldsymbol{X}_0\|_2^2]\mathrm{d}t\Big],
\end{aligned} \qquad (25)$$

where the second inequality follows from Jensen's inequality and the last inequality follows from Assumption 2 and the fact that $\|\mathbf{A}_\sigma\|_2 \geq 1$. We further have

$$\mathbb{E}[\|\boldsymbol{X}_t - \boldsymbol{X}_0\|_2^2] = \mathbb{E}\left[\left\|\int_0^t \mathbf{A}_\sigma^{-1} \nabla f(\boldsymbol{X}_\tau) \mathrm{d}\tau\right\|_2^2\right] + 2\beta^{-1} t \mathbb{E}[\|\mathbf{A}_\sigma^{-1/2}\boldsymbol{\epsilon}\|_2^2]$$
$$\leq \mathbb{E}\left[t \int_0^t \|\nabla f(\boldsymbol{X}_\tau)\|_2^2 \mathrm{d}\tau\right] + 2\gamma_2 \beta^{-1} dt$$

where the inequality is by Jensen's inequality and Lemma 13 and $\gamma_2 = d^{-1} \sum_{i=1}^d (1 + 2\sigma - 2\sigma\cos(2\pi i/d))^{-1}$ is strictly smaller than 1. By Lemma 15, we have

$$\mathbb{E}[\|\nabla f(\boldsymbol{X}_\tau)\|_2^2] \leq 2M^2 \mathbb{E}[\|\boldsymbol{X}_\tau\|_2^2] + 2G^2.$$

Note that by Ito's lemma we have for any $0 \leq s \leq \tau$,

$$\frac{\mathrm{d}\mathbb{E}[\|\boldsymbol{X}_s\|_{\mathbf{A}_\sigma}^2]}{\mathrm{d}s} = -2\mathbb{E}[\langle \boldsymbol{X}_s, \nabla f(\boldsymbol{X}_s)\rangle] + \beta^{-1} d \leq -2m\mathbb{E}[\|\boldsymbol{X}_s\|_2^2] + 2b + \beta^{-1} d \leq 2b + \beta^{-1} d,$$

where the second inequality follows from Assumption 1. Therefore,

$$\mathbb{E}[\|\boldsymbol{X}_\tau\|_2^2] \leq \mathbb{E}[\|\boldsymbol{X}_\tau\|_{\mathbf{A}_\sigma}^2] = \mathbb{E}[\|\boldsymbol{X}_0\|_{\mathbf{A}_\sigma}^2] + \int_0^\tau \frac{\mathrm{d}\mathbb{E}[\|\boldsymbol{X}_s\|_{\mathbf{A}_\sigma}^2]}{\mathrm{d}s} \mathrm{d}s \leq \mathbb{E}[\|\boldsymbol{X}_0\|_{\mathbf{A}_\sigma}^2] + \tau(2b + \beta^{-1} d).$$

Note that $\boldsymbol{X}_0 = \boldsymbol{x}$ is a iterate of LS-SGLD, by Lemma 11 we have $\mathbb{E}[\|\boldsymbol{X}_0\|_{\mathbf{A}_\sigma}^2] \leq (2b + \beta^{-1} d)/(c_1 m)$ for some constant $c_1 \in [\|\mathbf{A}_\sigma\|_2^{-1}, 1]$. Therefore,

$$\mathbb{E}[\|\nabla f(\boldsymbol{X}_\tau)\|_2^2] \leq 2M^2 \mathbb{E}[\|\boldsymbol{X}_\tau\|_2^2] + 2G^2 \leq \frac{4M^2(2b + \beta^{-1} d)}{c_1 m} + 2G^2 + 2M^2\tau(2b + \beta^{-1} d).$$

Thus, it follows that

$$\mathbb{E}[\|\boldsymbol{X}_t - \boldsymbol{X}_0\|_2^2] \leq \left(\frac{4M^2(2b + \beta^{-1} d)}{c_1 m} + 2G^2 + 2M^2\tau(2b + \beta^{-1} d)\right)t^2 + + 2\gamma_2 \beta^{-1} dt.$$

Note that $\tau, t \leq \eta$, plugging the above inequality into (25), we have

$$\mathbb{E}[\|\mathcal{L}_\eta \boldsymbol{x} - \mathcal{G}_\eta \boldsymbol{x}\|_2^2] \leq M^2\left[\left(\frac{4M^2(2b + \beta^{-1} d)}{c_1 m} + 2G^2 + 2M^2(2b + \beta^{-1} d)\eta\right)\eta^4 + 2\gamma_2 \beta^{-1} d\eta^3\right].$$

For sufficiently small step size satisfying

$$\eta \leq \frac{c_1 \beta^{-1} \gamma_2 md}{4M^2(2b + \beta^{-1} d) + 2c_1 mG^2} \wedge \sqrt{\frac{\gamma_2 \beta^{-1} d}{M^2(2b + \beta^{-1} d)}},$$

we have

$$\mathbb{E}[\|\mathcal{L}_\eta \boldsymbol{x} - \mathcal{G}_\eta \boldsymbol{x}\|_2^2] \leq 4\gamma_2 \beta^{-1} d\eta^3.$$

This completes the proof. □

D.3 Proof of Lemma 11

Lemma 15 (Lemma 3.1 in Raginsky et al. (2017)). *For any $\boldsymbol{x} \in \mathbb{R}^d$ and $i \in [n]$, it holds that*

$$\|\nabla f_i(\boldsymbol{x})\|_2 \leq M\|\boldsymbol{x}\|_2 + G,$$

where $G = \max_{i \in [n]} \|\nabla f_i(0)\|_2$.

Proof. Recall the update formula of $\boldsymbol{x}_k$,

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta \mathbf{A}_\sigma^{-1} \mathbf{g}_k + \sqrt{2\beta^{-1}\eta} \mathbf{A}_\sigma^{-1/2} \boldsymbol{\epsilon}_k.$$

Therefore, it holds that

$$
\begin{aligned}
\mathbb{E}[\|\boldsymbol{x}_{k+1}\|_{\mathbf{A}_\sigma}^2] &= \mathbb{E}[\|\boldsymbol{x}_k - \eta \mathbf{A}_\sigma^{-1} \mathbf{g}_k\|_{\mathbf{A}_\sigma}^2] + 2\eta\beta^{-1}\mathbb{E}[\|\mathbf{A}_\sigma^{-1/2}\boldsymbol{\epsilon}_k\|_{\mathbf{A}_\sigma}^2] \\
&= \mathbb{E}[\|\boldsymbol{x}_k\|_{\mathbf{A}_\sigma}^2] - 2\eta\mathbb{E}[\langle \boldsymbol{x}_k, \mathbf{g}_k \rangle] + \eta^2 \mathbb{E}[\|\mathbf{g}_k\|_{\mathbf{A}_\sigma^{-1}}^2] + 2\eta\beta^{-1}d,
\end{aligned}
$$

where the second equality follows from the fact that $\mathbb{E}[\|\boldsymbol{\epsilon}_k\|_2^2] = d$. Note that all eigenvalues of $\mathbf{A}_\sigma$ are greater than 1, it follows that

$$
\begin{aligned}
\mathbb{E}[\|\boldsymbol{x}_{k+1}\|_{\mathbf{A}_\sigma}^2] &= \mathbb{E}[\|\boldsymbol{x}_k\|_{\mathbf{A}_\sigma}^2] - 2\eta\mathbb{E}[\langle \boldsymbol{x}_k, \nabla f(\boldsymbol{x}) \rangle] + \eta^2 \mathbb{E}[\|\mathbf{g}_k\|_2^2] + 2\eta\beta^{-1}d \\
&\leq \mathbb{E}[\|\boldsymbol{x}_k\|_{\mathbf{A}_\sigma}^2] - 2\eta m \mathbb{E}[\|\boldsymbol{x}_k\|_2^2] + 2\eta b + 2\eta^2(M^2\mathbb{E}[\|\boldsymbol{x}_k\|_2^2] + G^2) + 2\eta\beta^{-1}d,
\end{aligned}
$$

where the inequality follows from Assumption 1, Lemma 15 and Young's inequality. Since the step size η satisfies $\eta \leq m/(2M^2)$, we further have

$$\mathbb{E}[\|\boldsymbol{x}_{k+1}\|_{\mathbf{A}_\sigma}^2] \leq \mathbb{E}[\|\boldsymbol{x}_k\|_{\mathbf{A}_\sigma}^2] - \eta m \mathbb{E}[\|\boldsymbol{x}_k\|_2^2] + 2\eta(b + \beta^{-1}d + \eta G^2).$$

Recall that all eigenvalues of $\mathbf{A}_\sigma$ are greater than 1, there exists a constant $\|\mathbf{A}_\sigma\|_2^{-1} \leq c_1 \leq 1$ such that

$$\mathbb{E}[\|\boldsymbol{x}_{k+1}\|_{\mathbf{A}_\sigma}^2] \leq (1 - c_1\eta m)\mathbb{E}[\|\boldsymbol{x}_k\|_{\mathbf{A}_\sigma}^2] + 2\eta(b + \beta^{-1}d + \eta G^2). \tag{26}$$

Since $\eta \leq 1/(c_1 m) \wedge b/G$, (26) implies that the following holds for all $k \geq 0$,

$$\mathbb{E}[\|\boldsymbol{x}_k\|_{\mathbf{A}_\sigma}^2] \leq (1 - c_1\eta m)^k \|\boldsymbol{x}_0\|_{\mathbf{A}_\sigma}^2 + \frac{2(2b + \beta^{-1}d)}{c_1 m}.$$

Since at the initialization $\boldsymbol{x}_0 = 0$, we have

$$\mathbb{E}[\|\boldsymbol{x}_k\|_2^2] \leq \mathbb{E}[\|\boldsymbol{x}_k\|_{\mathbf{A}_\sigma}^2] \leq \frac{2(2b + \beta^{-1}d)}{c_1 m}.$$

This completes the proof. $\qquad\square$

D.4 Proof of Lemma 14

Proof. We first define the function $L(t) = e^{\|\boldsymbol{X}_t\|_{\mathbf{A}_\sigma}^2}$, then by Ito's formula, we have

$$
\begin{aligned}
\mathrm{d}\mathbb{E}[L(t)] &= -2\mathbb{E}[\langle \mathbf{A}_\sigma \boldsymbol{X}_t, \mathbf{A}_\sigma^{-1}\nabla f(\boldsymbol{X}_t) \rangle L(t)]\mathrm{d}t + \mathbb{E}[\langle 4\mathbf{A}_\sigma \boldsymbol{X}_t \boldsymbol{X}_t^\top \mathbf{A}_\sigma + 2\mathbf{A}_\sigma, \beta^{-1}\mathbf{A}_\sigma^{-1}\mathbf{I} \rangle L(t)]\mathrm{d}t \\
&= -2\mathbb{E}\big[\big(\langle \boldsymbol{X}_t, \nabla f(\boldsymbol{X}_t) \rangle - \beta^{-1}d - 2\beta^{-1}\|\boldsymbol{X}_t\|_{\mathbf{A}_\sigma}^2\big)L(t)\big]\mathrm{d}t.
\end{aligned}
$$

By Assumption (1), we further have

$$\mathrm{d}\mathbb{E}[L(t)] \leq 2\mathbb{E}\big[\big((-m\|\boldsymbol{X}_t\|_2^2 + 2\beta^{-1}\|\boldsymbol{X}_t\|_{\mathbf{A}_\sigma}^2) + b + \beta^{-1}d\big)L(t)\big]\mathrm{d}t.$$

Therefore, assume $\beta \geq 2\|\mathbf{A}_\sigma\|_2/m$, we have

$$\mathrm{d}\mathbb{E}[L(t)] \leq 2(b + \beta^{-1}d)\mathbb{E}[L(t)]\mathrm{d}t.$$

Since $L(t)$ is always positive, it holds that

$$\mathbb{E}[L(t)] \leq L(0)e^{2(b + \beta^{-1}d)t}.$$

Note that $\|\boldsymbol{X}_t\|_{\mathbf{A}_\sigma}^2 \geq \|\boldsymbol{X}_t\|_2^2$, we immediately have

$$\mathbb{E}[e^{\|\boldsymbol{X}_t\|_2^2}] \leq \mathbb{E}[e^{\|\boldsymbol{X}_t\|_{\mathbf{A}_\sigma}^2}] \leq L(0)e^{2(b + \beta^{-1}d)t},$$

which completes the proof. $\qquad\square$

References

P. Kloeden abd E. Platen. *Numerical Solution of Stochastic Differential Equations.* Springer, 1992.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348. Springer Science & Business Media, 2013.

Rémi Bardenet, Arnaud Doucet, and Chris Holmes. On markov chain monte carlo methods for tall data. *The Journal of Machine Learning Research*, 18(1):1515–1557, 2017.

Michael Betancourt. The fundamental incompatibility of scalable hamiltonian monte carlo and naive data subsampling. In *International Conference on Machine Learning*, pages 533–540, 2015.

Francois Bolley and Cedric Villani. Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. *Annales de la Faculté des Sciences de Toulouse. Série VI. Mathématiques*, 14, 01 2005. doi: 10.5802/afst.1095.

Anton Bovier, Michael Eckhoff, Véronique Gayrard, and Markus Klein. Metastability in reversible diffusion processes i: Sharp asymptotics for capacities and exit times. *Journal of the European Mathematical Society*, 6(4):399–424, 2004.

Niladri S Chatterji, Nicolas Flammarion, Yi-An Ma, Peter L Bartlett, and Michael I Jordan. On the theory of variance reduction for stochastic gradient monte carlo. *arXiv preprint arXiv:1802.05431*, 2018.

Changyou Chen, Nan Ding, and Lawrence Carin. On the convergence of stochastic gradient mcmc algorithms with high-order integrators. In *Advances in Neural Information Processing Systems*, pages 2278–2286, 2015.

Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic Gradient Hamiltonian Monte Carlo. In *International Conference on Machine Learning*, 2014.

Xiang Cheng, Niladri S Chatterji, Yasin Abbasi-Yadkori, Peter L Bartlett, and Michael I Jordan. Sharp convergence rates for langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*, 2018a.

Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. Underdamped Langevin mcmc: A non-asymptotic analysis. In *Proceedings of the 31st Conference On Learning Theory*, volume 75, pages 300–323, 2018b.

Tzuu-Shuh Chiang, Chii-Ruey Hwang, and Shuenn Jyi Sheu. Diffusion for global optimization in r^n. *SIAM Journal on Control and Optimization*, 25(3):737–753, 1987.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017.

Arnak S Dalalyan and Avetik G Karagulyan. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *arXiv preprint arXiv:1710.00095*, 2017.

Khue-Dung Dang, Matias Quiroz, Robert Kohn, Minh-Ngoc Tran, and Mattias Villani. Hamiltonian monte carlo with energy conserving subsampling. *Journal of machine learning research*, 20(100):1–31, 2019.

Kumar Avinava Dubey, Sashank J Reddi, Sinead A Williamson, Barnabas Poczos, Alexander J Smola, and Eric P Xing. Variance Reduction in Stochastic Gradient Langevin Dynamics. In *Advances in Neural Information Processing Systems*, pages 1154–1162, 2016.

Alain Durmus, Eric Moulines, et al. Nonasymptotic convergence analysis for the unadjusted langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.

Gene Golub and Charles Van Loan. *Matrix Computation.* Johns Hopkin, 3rd edition, 1996.

István Gyöngy. Mimicking the one-dimensional marginal distributions of processes having an itô differential. *Probability theory and related fields*, 71(4):501–516, 1986.

Lisa Kreusser, Stanley Osher, and Bao Wang. A deterministic approach to avoid saddle points. *arXiv preprint arXiv:1901.06827*, 2019.

Lihua Lei, Cheng Ju, Jianbo Chen, and Michael I Jordan. Non-convex finite-sum optimization via scsg methods. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2348–2358. 2017.

Chunyuan Li, Changyou Chen, David Carlson, and Carin Lawrence. Preconditioned stochastic gradient langevin dynamics for deep neural networks. In *Association for the Advancement of Artificial Intelligence*, 2016.

Zhize Li, Tianyi Zhang, and Jian Li. Stochastic Gradient Hamiltonian Monte Carlo with Variance Reduction for Bayesian Inference. *arXiv preprint arXiv:1803.11159*, 2018.

Yi-An Ma, Tianqi Chen, and Emily Fox. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems*, pages 2917–2925, 2015.

Peter A Markowich and Cédric Villani. On the trend to equilibrium for the fokker-planck equation: an interplay between physics and functional analysis. In *Physics and Functional Analysis, Matematica Contemporanea (SBM) 19*. Citeseer, 1999.

Jonathan C Mattingly, Andrew M Stuart, and Desmond J Higham. Ergodicity for sdes and approximations: locally lipschitz vector fields and degenerate noise. *Stochastic processes and their applications*, 101(2): 185–232, 2002.

Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. *arXiv preprint arXiv:1707.05947*, 2017.

Radford M Neal et al. MCMC using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2: 113–162, 2011.

Stanley Osher, Bao Wang, Penghang Yin, Xiyang Luo, Minh Pham, and Alex Lin. Laplacian smoothing gradient descent. *arXiv preprint arXiv:1806.06317*, 2018.

G Parisi. Correlation functions and computer simulations. *Nuclear Physics B*, 180(3):378–384, 1981.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703, 2017.

Yee Whye Teh, Alexandre H Thiery, and Sebastian J Vollmer. Consistency and fluctuations for stochastic gradient langevin dynamics. *The Journal of Machine Learning Research*, 17(1):193–225, 2016.

Sebastian J Vollmer, Konstantinos C Zygalakis, and Yee Whye Teh. Exploration of the (non-) asymptotic bias and variance of stochastic gradient langevin dynamics. *The Journal of Machine Learning Research*, 17 (1):5504–5548, 2016.

Bao Wang, Quanquan Gu, March Boedihardjo, Farzin Barekat, and Stanley Osher. Dp-lssgd: A stochastic optimization method to lift the utility in privacy-preserving erm. *arXiv preprint arXiv:1906.12056*, 2019.

Max Welling and Teh Yee Whye. Bayesian learning via stochastic gradient langevin dynamics. In *International Conference on Machine Learning*, 2011.

Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of langevin dynamics based algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 3122–3133, 2018.

Difan Zou, Pan Xu, and Quanquan Gu. Stochastic variance-reduced Hamilton Monte Carlo methods. In *Proceedings of the 35th International Conference on Machine Learning*, pages 6028–6037, 2018a.

Difan Zou, Pan Xu, and Quanquan Gu. Subsampled stochastic variance-reduced gradient Langevin dynamics. In *Proceedings of International Conference on Uncertainty in Artificial Intelligence*, 2018b.

Difan Zou, Pan Xu, and Quanquan Gu. Sampling from non-log-concave distributions via variance-reduced gradient langevin dynamics. In *Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2936–2945. PMLR, 2019a.

Difan Zou, Pan Xu, and Quanquan Gu. Stochastic gradient Hamiltonian monte carlo methods with recursive variance reduction. In *Advances in Neural Information Processing Systems 2019*, 2019b.