SIAM J. SCI. COMPUT.
Vol. 43, No. 1, pp. A26–A53

LAPLACIAN SMOOTHING STOCHASTIC GRADIENT MARKOV CHAIN MONTE CARLO*

BAO WANG†, DIFAN ZOU‡, QUANQUAN GU‡, AND STANLEY J. OSHER§

Abstract. As an important Markov chain Monte Carlo (MCMC) method, the stochastic gradient Langevin dynamics (SGLD) algorithm has achieved great success in Bayesian learning and posterior sampling. However, SGLD typically suffers from a 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 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.

Key words. Langevin dynamics, stochastic gradient, Laplacian smoothing

AMS subject classifications. 60H35, 60J22, 65C05, 65C60

DOI. 10.1137/19M1294356

1. Introduction.

Let $\boldsymbol{x} \in \mathbb{R}^d$ be a machine learning model's parameter with prior distribution $p(\boldsymbol{x})$ and let $p(\boldsymbol{d}|\boldsymbol{x})$ be the likelihood function of the observed data $\boldsymbol{d}$. Suppose the training data points are generated independently from some unknown distribution; then the posterior distribution of the model parameter $\boldsymbol{x}$ given the entire training dataset $\mathcal{D} = \{\boldsymbol{d}_i\}_{i=1}^n$ is computed as $p(\boldsymbol{x}|\mathcal{D}) \propto p(\boldsymbol{x})\Pi_{i=1}^n p(\boldsymbol{d}_i|\boldsymbol{x})$. While optimization algorithms can be used to find the maximum a posterior point estimator, i.e., $\boldsymbol{x}_{\text{MAP}} = \arg\max_{\boldsymbol{x}} \log p(\boldsymbol{x}|\mathcal{D})$, 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 Markov chain Monte Carlo (MCMC) algorithms to sample $e^{-f(\boldsymbol{x})}$, where

$$f(\boldsymbol{x}) := \frac{1}{n}\sum_{i=1}^n f_i(\boldsymbol{x}) = -\frac{1}{n}\sum_{i=1}^n \log p(\boldsymbol{d}_i|\boldsymbol{x}). \tag{1.1}$$

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

*Submitted to the journal's Methods and Algorithms for Scientific Computing section October 21, 2019; accepted for publication (in revised form) October 2, 2020; published electronically January 4, 2021. B. Wang and D. Zou contributed equally; Q. Gu and S. Osher are co-corresponding authors.
https://doi.org/10.1137/19M1294356

Funding: This work was supported by the National Science Foundation under grants DMS-1924935, DMS-1952339, DMS-1554564 (STROBE), and SaTC-1717950, by the Air Force Research Laboratory under grants FA9550-18-0167 and MURI FA9550-18-1-0502, by the Office of Naval Research under grant N00014-18-1-2527, and by the Department of Energy under grant DE-SC0021142.

†Scientific Computing and Imaging Institute and Department of Mathematics, University of Utah, Salt Lake City, UT 84112 USA (wangbaonj@gmail.com).

‡Department of Computer Science, University of California, Los Angeles, Los Angeles, CA 90095 USA (knowzou@cs.ucla.edu, qgu@cs.ucla.edu).

§Department of Mathematics, University of California, Los Angeles, Los Angeles, CA 90095 USA (sjo@math.ucla.edu).

The first-order LD reads

$$d\boldsymbol{X}_t = -\nabla f(\boldsymbol{X}_t)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}\boldsymbol{B}_t, \tag{1.2}$$

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 (1.2) converges to a unique invariant distribution $\pi \propto e^{-\beta f(\boldsymbol{x})}$ [14]. Therefore, one can apply an numerical integrator to approximate (1.2) in order to obtain samples that follow the posterior distribution. One simple integrator is to apply the Euler–Maruyama discretization [23] to (1.2), which gives

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta\nabla f(\boldsymbol{x}_k) + \sqrt{2\beta^{-1}\eta}\boldsymbol{\epsilon}_k, \tag{1.3}$$

and it is known as the Langevin Monte Carlo (LMC) (a.k.a., unadjusted Langevin algorithm [34]). When the target density, i.e., posterior distribution, is strongly log-concave and log-smooth, Dalalyan and others 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 [15, 19]. In the same setting, [20] studied the Metropolis-adjusted Langevin algorithm, realized by introducing a Metropolis–Hastings (MH) correction step into LMC, and proved its linear convergence rate in total variation distance. Furthermore, the convergence guarantee of LMC for sampling from non-log-concave distributions has also been established in [35, 40].

Note that the posterior distribution is defined on the whole dataset $\mathcal{D}$, which is typically extremely large in modern machine learning 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 (1.3) with a subsampled one, which gives rise to stochastic gradient Langevin dynamics (SGLD) [39]. From the theoretical perspective, the convergence guarantee of SGLD has been proved for both strongly log-concave distributions [16] and non-log-concave distributions [35, 40] in 2-Wasserstein distance. The generalization performance of SGLD for nonconvex optimization has been further investigated in [31]. 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 [36, 37]. In order to reduce the variance of stochastic gradient as well as to improve the convergence rate, Dubey et al. incorporated variance reduction techniques into SGLD [18], which gives rise to a family of variance-reduced LD-based algorithms such as SVRG-LD and SAGA-LD. Chatterji et al. 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 [8, 43, 44]. 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 *whether we can reduce the variance of stochastic gradients while maintaining similar computation and memory costs of SGLD.*

Recently, Osher et al. integrated Laplacian smoothing and related high-order smoothing techniques into stochastic gradient descent (SGD) to reduce the variance of stochastic gradient on-the-fly [33]. 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 training a large number of deep learning models with good generalization ability. Laplacian smoothing

also demonstrates some abilities to avoid saddle point in gradient descent [24]. Most recently, Wang et al. leveraged Laplacian smoothing to improve the utility of machine learning models trained with privacy guarantee [38].

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 (FFT) 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. Moreover, we prove the convergence rate of LS-SGLD for sampling from both log-concave and non-log-concave densities in 2-Wasserstein distance and show that there exists a trade-off between the discretization error and ergodicity rate. Experimental results show that compared with SGLD, LS-SGLD can achieve much smaller discretization error and similar ergodicity rate, and demonstrate the superior performance of LS-SGLD for a variety of machine learning applications.

1.1. Our contributions. We summarize our main contributions 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 counterpart 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 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 the 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 (BLR), 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-known high-order MCMC methods is Hamiltonian Monte Carlo (HMC) [32, 41], which incorporates a Hamiltonian momentum term into the first-order MCMC method in order to improve the mixing time. Other variants of HMC include the reduced-order HMC [1], which has broad applications in scientific computing such as background flow field estimation [6]. Similar to SGLD, a stochastic version of HMC (namely SGHMC) has been further established in [10] and was shown to be able to achieve a faster convergence rate than SGLD in experiments. Ma, Chen, and Fox investigated a family of SGHMC methods and proposed a new state-adaptive sampler on the Riemannian

manifold [28]. Chen, Ding, and Carin provided theoretical convergence guarantees of SGHMC in terms of mean square error (MSE) and proposed a second-order symmetric splitting integrator to further improve the discretization error [9]. When the target density is strongly log-concave and log-smooth, Cheng et al. proposed underdamped MCMC and stochastic gradient underdamped MCMC and obtained convergence rates in 2-Wasserstein distance [13]. The convergence rates of these two algorithms have been further established for sampling from non-log-concave densities [12]. Chen et al. proposed the stochastic gradient HMC [11]. However, due to the large variance of stochastic gradients and lack of the MH correction step, SGHMC has also been observed to have a highly biased sampling trajectory [4, 17]. 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 [42, 27, 45].

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 BLR, and convolutional neural nets (CNNs). We conclude this work in section 5.

1.4. Notation. Throughout this paper we use bold uppercase letters $\mathbf{A}$, $\mathbf{B}$ to denote matrices, bold lowercase letters $\boldsymbol{x}$, $\boldsymbol{y}$ to denote vectors, and lowercase letters x, y and α, β to denote scalars. We use $x \wedge y$ and $x \vee y$ to denote $\min\{x, y\}$ and $\max\{x, y\}$, respectively. For continuous-time random vectors, we denote them with italics bold uppercase letters $\boldsymbol{X}$, $\boldsymbol{Y}$ with sub/superscripts. 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 a semipositive 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 (1D) Laplacian matrix, respectively. Therefore,

$$\mathbf{A}_\sigma := \begin{bmatrix} 1+2\sigma & -\sigma & 0 & \ldots & 0 & -\sigma \\ -\sigma & 1+2\sigma & -\sigma & \ldots & 0 & 0 \\ 0 & -\sigma & 1+2\sigma & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ -\sigma & 0 & 0 & \ldots & -\sigma & 1+2\sigma \end{bmatrix}_{d \times d}. \tag{2.1}$$

To optimize the loss function $f(\boldsymbol{x}) = 1/n \sum_{i=1}^n f_i(\boldsymbol{x})$, LSSGD [33] takes the following iteration:

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \eta_k \mathbf{A}_\sigma^{-1} \nabla f_{i_k}(\boldsymbol{x}^k), \tag{2.2}$$

where $\eta_k > 0$ is the learning rate, and i_k is a random sample from $[n] := \{1, 2, \ldots, n\}$. 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 $\boldsymbol{d} = [-2, 1, 0, \ldots, 0, 1]^T$. 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 coordinatewise 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. [33] showed that for an 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 gradient descent'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):

$$d\boldsymbol{X}_t = -\mathbf{A}_\sigma^{-1}\nabla f(\boldsymbol{X}_t) + \sqrt{2\beta^{-1}}\mathbf{A}_\sigma^{-1/2}\mathrm{d}\boldsymbol{B}_t. \tag{2.3}$$

Note that we premultiply 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 formally state this property in the following proposition.

PROPOSITION 2.1. *The stationary distribution, π, of the LS-LD,* (2.3)*, satisfies* $\pi \propto e^{-\beta f(\boldsymbol{x})}$.

The proof of Proposition 2.1 can be found in Appendix A. If we apply the Euler–Maruyama scheme to discretize (2.3), we end up with the following discrete algorithm, namely Laplacian smoothing gradient Langevin dynamics (LS-GLD):

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta\mathbf{A}_\sigma^{-1}\nabla f(\boldsymbol{x}_k) + \sqrt{2\beta^{-1}\eta}\mathbf{A}_\sigma^{-1/2}\boldsymbol{\epsilon}_k, \tag{2.4}$$

where $\boldsymbol{\epsilon}_k \sim N(\mathbf{0}, \mathbf{I}_{d\times d})$. In practice, we use the minibatch 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 (2.4), and we arrive at the following LS-SGLD:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta\mathbf{A}_\sigma^{-1}\mathbf{g}_k + \sqrt{2\beta^{-1}\eta}\mathbf{A}_\sigma^{-1/2}\boldsymbol{\epsilon}_k. \tag{2.5}$$

We summarize LS-SGLD in Algorithm 2.1. In the remaining part of this section, we will present FFT-based fast algorithms for computing $\mathbf{A}_\sigma^{-1}\mathbf{g}_k$ and $\mathbf{A}_\sigma^{-1/2}\boldsymbol{\epsilon}_k$.

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 an FFT-based algorithm for solving (2.4).

Algorithm 2.1. 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 $k = 0, 1, \ldots, K-1$ **do**

Uniformly sample $\mathcal{I}_k \subset [n]$ with $|\mathcal{I}_k| = B$.

Compute the minibatch 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: $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_K$.

LEMMA 2.2 (see [21]). *The normalized eigenvectors of the $d \times d$ circulant matrix,*

$$\mathbf{C} = \begin{bmatrix} c_0 & c_{d-1} & \ldots & c_2 & c_1 \\ c_1 & c_0 & c_{d-1} & \ldots & c_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ c_{d-2} & \ldots & \ldots & \ldots & c_{d-1} \\ c_{d-1} & c_{d-2} & \ldots & c_1 & c_0 \end{bmatrix}, \tag{2.6}$$

are given by

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

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

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

LEMMA 2.3 (see [21]). *The inverse of a circulant matrix is circulant.*

LEMMA 2.4 (see [21]). *The square root of a circulant matrix is circulant.*

LEMMA 2.5. *For any circulant matrix $\mathbf{C}$ of the form in* (2.6)*, and for any given vector $\boldsymbol{v}$, let $\boldsymbol{u} = \mathbf{C}^{-1}\boldsymbol{v}$; then $\boldsymbol{u}$ can be computed by the FFT with sublinear scaling in the following way:*

$$\boldsymbol{u} = \text{ifft}\left(\frac{\text{fft}(\boldsymbol{v})}{\text{fft}(\boldsymbol{c})}\right), \tag{2.7}$$

where $\boldsymbol{c}$ is the first row of the matrix $\mathbf{C}$, and the division in (2.7) *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 $\text{fft}(\boldsymbol{v}) = \text{fft}(\boldsymbol{c}) \cdot \text{fft}(\boldsymbol{u})$. ∎

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 2.3 and 2.4, $\mathbf{A}_\sigma^{-1/2}$ 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 $\mathbf{Q} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_d]^T$ with $\boldsymbol{v}_i$ being the eigenvector associated with the eigenvalue $\lambda_i > 0$, and $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$. Therefore, we have

$$\mathbf{A}_\sigma^{-1/2} = \mathbf{Q}\sqrt{\Lambda}\mathbf{Q}^{-1}, \tag{2.8}$$

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

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 2.6. In computing (2.8), there is no need to store the matrix $\mathbf{Q}$; according to Lemma 2.2, each row of $\mathbf{Q}$ and $\sqrt{\Lambda}$ can be written 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 theoretical 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 [30, 35, 40, 44], which is essential to guarantee the convergence of the continuous-time Langevin dynamics (1.2).

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 functions $f_i(\boldsymbol{x})$.

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 3.1 (logarithmic Sobolev inequality). *Let μ be a probability measure; then we say μ satisfies the 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 the logarithmic Sobolev inequality.

Proposition 3.2 (see [35]). *Under Assumptions 1 and 2, the target density $\pi \propto e^{-f(\boldsymbol{x})}$ satisfies the logarithmic Sobolev inequality with some constant $\lambda > 0$.*

It has been shown in [19, 35] that if the function $f(\boldsymbol{x})$ is smooth and strongly convex (which is stronger than Assumption 1), the logarithmic Sobolev constant λ is a universal constant. However, if the function $f(\boldsymbol{x})$ is nonconvex, in the worst case the logarithmic Sobolev constant λ has exponential dependency on the problem dimension d and inverse temperature β [7, 35].

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 3.3. *Under Assumptions* 1, 2, 3, *and* 4, *if we 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*

$$\begin{aligned}\mathcal{W}_2(\mathbb{P}(\boldsymbol{x}_K), \pi) \leq &\left(\frac{2\gamma_1 K\eta^2\beta d\omega^2}{B}\right)^{1/2} + \left[8\gamma_2 K\eta^2(K+1)\beta d\eta\right]^{1/2} \\ &+ \left[2\lambda\big(\beta f(0) + \log(\Lambda)\big)\right]^{1/2} e^{-c_0 K\eta/(2\beta\lambda)},\end{aligned} \tag{3.1}$$

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 3.4. We emphasize that the three terms on the right-hand side (R.H.S.) of (3.1) have their respective meanings. In particular, the first and second terms represent the discretization errors introduced by the stochastic gradient estimator and the numerical integrator of (2.3), respectively. The third term represents the ergodicity of the continuous-time Markov process (2.3), which characterizes the mixing time of LS-LD (2.3). Moreover, we remark here that the convergence rate of LS-GLD (LS-SGLD with full gradient) can be directly implied from Theorem 3.3 by removing the first term on the R.H.S. of (3.1).

Based on Theorem 3.3, 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 3.3 are all 1's. We formally state the convergence result of SGLD in the following corollary.

COROLLARY 3.5. *Under the same assumptions in Theorem* 3.3, *the output of standard SGLD, denoted by* $\boldsymbol{y}_K$, *satisfies*

$$\begin{aligned}\mathcal{W}_2(\mathbb{P}(\boldsymbol{y}_K), \pi) \leq &\left(\frac{2K\eta^2 d\omega^2}{B}\right)^{1/2} + \left[8K\eta^2(K+1)\beta^{-1}d\eta\right]^{1/2} \\ &+ \left[2\lambda\big(\beta f(0) + \log(\Lambda)\big)\right]^{1/2} e^{-K\eta/(2\beta\lambda)}.\end{aligned} \tag{3.2}$$

Remark 3.6. 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 (3.1) and (3.2)), 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.

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

3.2. 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 3.7. *Under Assumptions* 1, 2, *and* 3, *if set the step size* $\eta \le C m \beta^{-1}/M^2$ *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* $\bar{\Gamma} = \big(3/2+2(b+\beta^{-1}d)\big)^{1/2}$ *such that the output of LS-SGLD satisfies*

$$\begin{aligned}\mathcal{W}_2\big(\mathbb{P}(\boldsymbol{x}_K), \pi\big) \le \bar{\Gamma}(K\eta)^{1/2}\Bigg[\bigg(\frac{\gamma_1\beta d\omega^2}{B}K\eta + 2\gamma_2 M^2 dK\eta^2\bigg)^{1/2} \\ + \bigg(\frac{\gamma_1\beta d\omega^2}{B}K\eta + 2\gamma_2 M^2 dK\eta^2\bigg)^{1/4}\Bigg] \\ + \big[2\lambda\big(\beta f(0) + \log(\Lambda)\big)\big]^{1/2} e^{-c_0 K\eta/(2\beta\lambda)},\end{aligned} \tag{3.3}$$

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 3.8. The convergence rate of SGLD in 2-Wasserstein distance can also be obtained from Theorem 3.7 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 [35]. As a clear comparison, the discretization error induced by both the stochastic gradient and the numerical integrator of LS-SGLD (the first bracket term of (3.3)) is smaller than that of SGLD, while the ergodicity term of LS-SGLD (the last term of (3.3)) 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 BLR, and training CNNs. Throughout all the experiments, we regard SGLD [39] and preconditioned SGLD (pSGLD) [26], which considers local curvature of $f(\boldsymbol{x})$ with RMSProp type adaptive step size, as benchmarks. In addition, we also incorporated the precondition technique, proposed in [26], into LS-SGLD, which leads to a variant of LS-SGLD, namely Laplacian smoothing preconditioned SGLD (LS-pSGLD).

4.1. Numerical simulations on synthetic datasets.

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

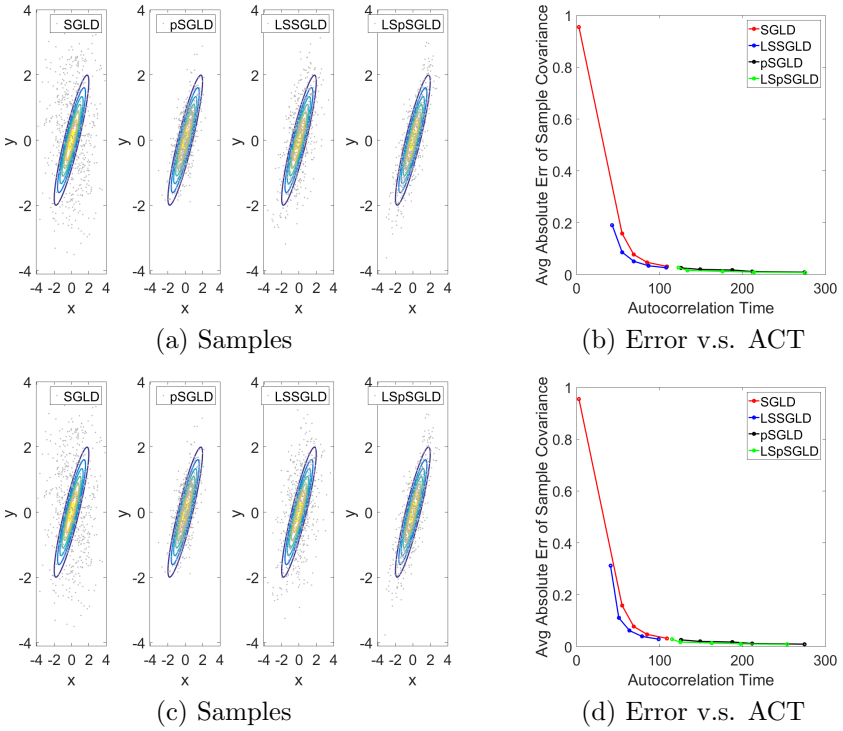
$$\mathbf{A}_\sigma = \begin{bmatrix} 1+\sigma & -\sigma \\ -\sigma & 1+\sigma \end{bmatrix}. \tag{4.1}$$

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

$$\tau = \frac{1}{2} + \sum_{t=1}^{\infty} \frac{A(t)}{A(0)}, \tag{4.2}$$

where $A(t) = \mathbb{E}[(\bar{\phi}_\eta - \phi(\boldsymbol{x}_0))(\bar{\phi}_\eta - \phi(\boldsymbol{x}_t))]$ for any given bounded function $\phi(\boldsymbol{x})$, $\bar{\phi} = \int_{\mathcal{X}} \phi(\boldsymbol{x}) p(\boldsymbol{x}|\mathcal{D}) d\boldsymbol{x}$ 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 tth step and $S_T = \sum_{t=1}^{T} \eta_t$.

Figures 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 four samplers in the experiments shown in Figure 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 Figure 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



(a) Samples (b) Error v.s. ACT

(c) Samples (d) Error v.s. ACT

FIG. 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* $(1+4\sigma)^{1/4}$.

$\{0.19, 0.19 \times 0.8, 0.19 \times 0.8^2, 0.19 \times 0.8^3, 0.19 \times 0.8^4\}$. 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) $\{0.19, 0.19 \times 0.8, 0.19 \times 0.8^2, 0.19 \times 0.8^3, 0.19 \times 0.8^4\} \times (1 + 4\sigma)^{1/4}$. Figures 1(b) and (d) plot the autocorrelation time versus reconstruction error of the covariance matrix. In (b) we use the same set of step sizes for all 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.

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

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

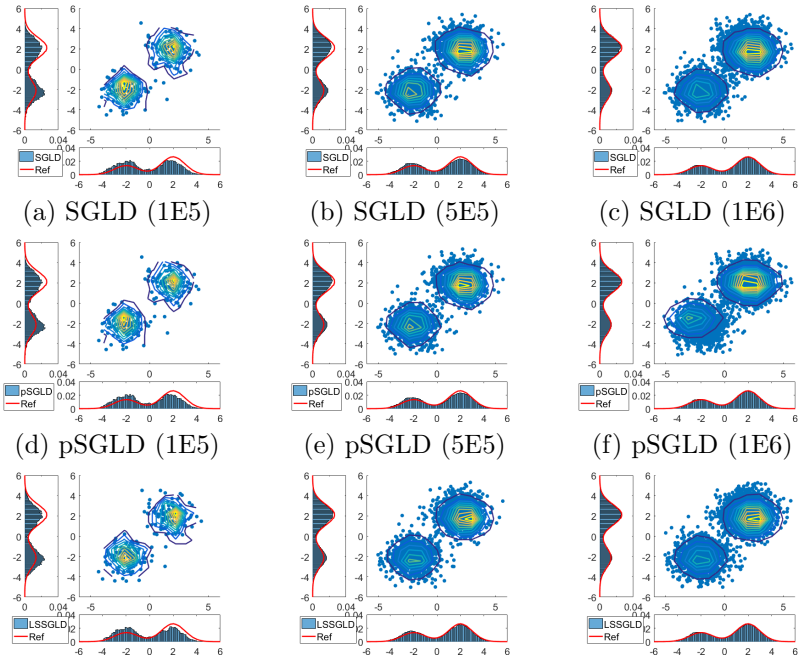
$$\nabla f_i(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{a}_i + \frac{2\boldsymbol{a}_i}{2 + \exp\left(2\langle \boldsymbol{x}, \boldsymbol{a}_i\rangle\right)}.$$

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 [3, 44, 45] 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 iterations where we set the batch size to be 10. We plot the distributions generated by different samplers with different numbers of iterations in Figure 2. As shown in Figures 2(c), (f), and (i), when the number of iterations is large enough, e.g., 10^6, the sample distributions of all 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 Figures 2(a), (d), and (g). With a moderate



(a) SGLD (1E5) (b) SGLD (5E5) (c) SGLD (1E6)

(d) pSGLD (1E5) (e) pSGLD (5E5) (f) pSGLD (1E6)

(g) LS-SGLD (1E5) (h) LS-SGLD (5E5) (i) LS-SGLD (1E6)

FIG. 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 MH correction and different SGLD.*

# 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

number of iterations, say, 5×10^5, the sample distribution from LS-SGLD is better than the other two (Figures 2(b), (e), and (h)).

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.

4.1.3. Comparison of the mixing time between LD and LS-LD. To verify that Laplacian smoothing in practice does not slow down the mixing rate of the continuous-time Markov process, we conduct the following experiments. First, we apply the MCMC with MH correction step to sample 10K points, respectively, from the following two distributions:

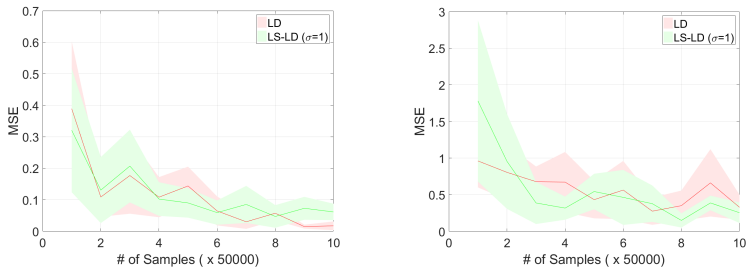


FIG. 3. *MSE between the true and reconstructed means from different numbers of samples generated by LD and LS-LD (*10 *independent runs).*

- Gaussian distribution with the probability density function

$$p(x,y) = \frac{1}{9\pi}\exp\left(-\left(\frac{(x-1)^2}{3^2}+\frac{(y-2)^2}{3^2}\right)\right), \tag{4.3}$$

- the Gaussian mixture distribution described in subsection 4.1.2.

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.

4.2. Bayesian logistic regression. Suppose we observe n independent and identically distributed samples $\{\boldsymbol{d}_i, y_i\}_{i=1,2,\ldots,n}$ where $\boldsymbol{d}_i \in \mathbb{R}^d$ and $y_i \in \{-1,1\}$ denote the feature and the corresponding label of the ith sample instance. The likelihood of the BLR model is given by

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

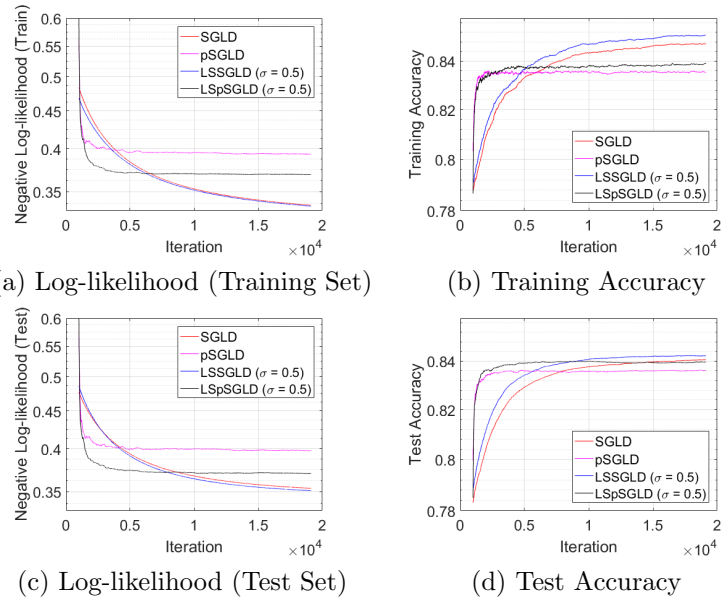
where $\boldsymbol{x}$ is the parameter to be learned. We use a Gamma prior $p(\boldsymbol{x}) \propto \|\boldsymbol{x}\|_2^{-\lambda}$ $\exp\left(-\theta\|\boldsymbol{x}\|_2\right)$ with $\lambda = 1$ and $\theta = 10^{-2}$. Then we formulate the logarithmic posterior distribution as follows:

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

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 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 four samplers. After burn-in, we compute the moving average of the sample parameters to estimate the regression parameters $\boldsymbol{x}$. We plot iteration versus negative

[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

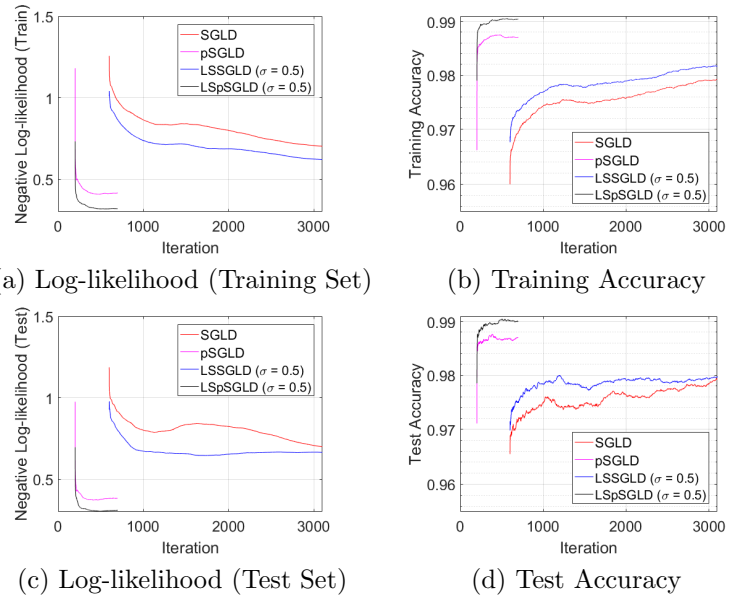
FIG. 4. *Convergence comparison for BLR, where the X-axis represents the number of iterations and the Y-axis represents the negative log-likelihood/accuracy.* (a) *Negative log-likelihood on training dataset;* (b) *accuracy on training dataset;* (c) *negative log-likelihood on test dataset;* (d) *accuracy on test dataset.*

log-likelihood and accuracy in Figure 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 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



(a) Log-likelihood (Training Set) (b) Training Accuracy

(c) Log-likelihood (Test Set) (d) Test Accuracy

FIG. 5. *Convergence comparison for Bayesian convolutional neural network, where the X-axis represents the number of iterations and the 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.*

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

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

The notation $\text{conv}_{c,k,m}$ denotes a 2D convolutional layer with c output channels, each of which is the sum of a channelwise 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 multiclass logistic function.

Similar to BLR, we use a Gamma prior $p(\boldsymbol{x}) \propto \|\boldsymbol{x}\|_2^{-\lambda} \exp(-\theta\|\boldsymbol{x}\|_2)$ 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 Figure 5, where 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.

5. Conclusions. In this paper, we integrate Laplacian smoothing with 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 FFT-based algorithms are employed to resolve the dynamics of the resulting 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 MCMC algorithms, e.g., the stochastic gradient HMC [10].

Appendix A. Missing proof in section 2. In this section, we provide the proof of Proposition 2.1.

Proof of Proposition 2.1. 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:

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

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. ∎

Appendix 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 (2.3), denoted by $\{\boldsymbol{X}_t\}_{t\geq 0}$. Letting $\boldsymbol{X}_0 = \boldsymbol{x}_0$, by the 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 (2.3), 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 B.1. *Under Assumptions* 1 *and* 2*, there exists a constant* $c_0 \in [\|\mathbf{A}_\sigma\|_2^{-1}, 1]$

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

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 B.1 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 3.3 and 3.7.

B.1. Proof of Theorem 3.3. 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 B.2. *Under Assumptions* 1, 2, 3, *and* 4, *if we 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:*

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

Then we can complete the proof of Theorem 3.3 as follows.

Proof of Theorem 3.3. By the triangle inequality and Lemmas B.1 and B.2, it is evident that

$$\begin{aligned}\mathcal{W}_2(\mathbb{P}(\boldsymbol{x}_K), \pi) &\le \mathcal{W}_2(\mathbb{P}(\boldsymbol{x}_K), \mathbb{P}(\boldsymbol{X}_{K\eta})) + \mathcal{W}_2(\mathbb{P}(\boldsymbol{X}_{K\eta}), \pi)\\ &\le \left(\frac{2\gamma_1 d\omega^2 K\eta^2}{B}\right)^{1/2} + \big[8\gamma_2 K(K+1)\beta^{-1}d\eta^3\big]^{1/2}\\ &\quad + \big[2\lambda\big(\beta f(0) + \log(\Lambda)\big)\big]^{1/2} e^{-c_0 K\eta/(2\beta\lambda)},\end{aligned}$$

which completes the proof. ∎

B.2. Proof of Theorem 3.7. Similar to the proof of Theorem 3.3, we provide the following lemma that characterizes the discretization error $\mathcal{W}_2\big(\mathbb{P}(\boldsymbol{x}_k), \mathbb{P}(\boldsymbol{X}_{k\eta})\big)$ for sampling from non-log-concave densities.

LEMMA B.3. *Under Assumptions* 1 *and* 2, *if we set the step size* $\eta \le Cm\beta^{-1}/M^2$ *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* $\bar{\Gamma} = \big(3/2 + 2(b + \beta^{-1}d)\big)^{1/2}$ *such that the following holds:*

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

Proof of Theorem 3.7. By the triangle inequality and Lemmas B.1 and B.3, it is evident that

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

which completes the proof. ∎

Appendix C. Proof of lemmas in Appendix B.

C.1. Proof of Lemma B.1. In order to prove Lemma B.1, we require the following lemma.

LEMMA C.1 (Theorem 9.6.1 in [2]). *Suppose the target density* π *satisfies logarithmic Sobolev inequality with a positive constant* λ; *for any density* μ *it holds that*

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

Proof of Lemma B.1. Recall that for LS-LD (2.3), the distribution of $\boldsymbol{X}_t$, denoted by $p(\boldsymbol{x}, t)$, can be described by the following Fokker–Planck equation:

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

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 shorthand 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

$$\begin{aligned}\frac{\mathrm{d}D_{KL}(\mathbb{P}_t||\pi)}{\mathrm{d}t} &= \int_{\mathbb{R}^d} \frac{\partial}{\partial t}\left[\mathbb{P}_t \log\left(\frac{\mathbb{P}_t}{\pi}\right)\right]\mathrm{d}\boldsymbol{x} \\ &= \int_{\mathbb{R}^d} \frac{\partial \mathbb{P}_t}{\partial t}\big[\log(\mathbb{P}_t) + 1 - \log(\pi)\big]\mathrm{d}\boldsymbol{x}.\end{aligned}$$

Similar to the proof of Proposition 2 in [31], by (C.1) we further have

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

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

$$\begin{aligned}\frac{\mathrm{d}D_{KL}(\mathbb{P}_t||\pi)}{\mathrm{d}t} &= -\frac{1}{\beta}\int_{\mathbb{R}_d}\big\langle \mathbf{A}_\sigma^{-1}\big(\nabla\log(\mathbb{P}_t) - \nabla\log(\pi)\big), \nabla\log(\mathbb{P}_t) - \nabla\log(\pi)\big\rangle \mathbb{P}_t\mathrm{d}\boldsymbol{x} \\ &= -\frac{1}{\beta}\int_{\mathbb{R}_d}\|\nabla\log(\mathbb{P}_t) - \nabla\log(\pi)\|^2_{\mathbf{A}_\sigma^{-1}}\mathbb{P}_t\mathrm{d}\boldsymbol{x}.\end{aligned}$$

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

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

where $\mathbf{I}(\mathbb{P}_t||\pi)$ denotes the Fisher information between $\mathbb{P}_t$ and π. By Proposition 3.2, we know that the target density π satisfies the logarithmic Sobolev inequality with constant $\lambda > 0$. Then, from [29], we have

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

Plugging the above inequality into (C.2), we obtain

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

which implies that

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

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 C.1, we have the following regarding the 2-Wasserstein distance $\mathcal{W}_2\big(\mathbb{P}(\boldsymbol{X}_{k\eta}), \pi\big)$:

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

which completes the proof. ∎

C.2. Proof of Lemma B.2. We first deliver the following useful lemmas.

LEMMA C.2. *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*

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

LEMMA C.3. *Under Assumptions* 1 *and* 2*, if we 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 B.2.

Proof of Lemma B.2. 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

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{x}_K - \boldsymbol{X}_{K\eta}\|_{\mathbf{A}_\sigma}^2] &= \mathbb{E}[\|\boldsymbol{x}_K - \mathcal{G}_\eta\boldsymbol{x}_{K-1} + \mathcal{G}_\eta\boldsymbol{x}_{K-1} - \boldsymbol{X}_{K\eta}\|_{\mathbf{A}_\sigma}^2] \\ &= \mathbb{E}[\|\boldsymbol{x}_K - \mathcal{G}_\eta\boldsymbol{x}_{K-1}\|_{\mathbf{A}_\sigma}^2] + \mathbb{E}[\|\mathcal{G}_\eta\boldsymbol{x}_{K-1} - \boldsymbol{X}_{K\eta}\|_{\mathbf{A}_\sigma}^2], \qquad \text{(C.3)}\end{aligned}$$

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 the stochastic gradient is independent of the iterate. Regarding the first term on the R.H.S. of (C.3), we have

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{x}_K - \mathcal{G}_\eta\boldsymbol{x}_{K-1}\|_{\mathbf{A}_\sigma}^2] &= \eta^2\mathbb{E}[\|\mathcal{S}_\eta\boldsymbol{x}_{K-1} - \mathcal{G}_\eta\boldsymbol{x}_{K-1}\|_{\mathbf{A}_\sigma}^2] \\ &\leq \eta^2\mathbb{E}[\|\mathbf{A}_\sigma^{-1}\mathbf{g}_{K-1} - \mathbf{A}_\sigma^{-1}\nabla f(\boldsymbol{x}_{K-1})\|_{\mathbf{A}_\sigma}^2] \\ &\leq \frac{\eta^2}{B}\mathbb{E}[\|\mathbf{A}_\sigma^{-1}\nabla f_i(\boldsymbol{x}_{K-1}) - \mathbf{A}_\sigma^{-1}\nabla f(\boldsymbol{x}_{K-1})\|_{\mathbf{A}_\sigma}^2] \\ &\leq \frac{\gamma_1\eta^2 d\omega^2}{B}, \qquad \text{(C.4)}\end{aligned}$$

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 [25], and the last inequality is by Assumption 3. In terms of the second term on the R.H.S. of (C.3), we have

$$\begin{aligned}\mathbb{E}[\|\mathcal{G}_\eta \boldsymbol{x}_{K-1} - \boldsymbol{X}_{K\eta}\|_{\mathbf{A}_\sigma}^2] &= \mathbb{E}[\|\mathcal{G}_\eta \boldsymbol{x}_{K-1} - \mathcal{L}_\eta \boldsymbol{x}_{K-1} + \mathcal{L}_\eta \boldsymbol{x}_{K-1} - \boldsymbol{X}_{K\eta}\|_{\mathbf{A}_\sigma}^2] \\ &\le (1+\alpha)\mathbb{E}[\|\mathcal{G}_\eta \boldsymbol{x}_{K-1} - \mathcal{L}_\eta \boldsymbol{x}_{K-1}\|_{\mathbf{A}_\sigma}^2] \\ &\quad + (1+1/\alpha)\mathbb{E}[\|\mathcal{L}_\eta \boldsymbol{x}_{K-1} - \mathcal{L}_\eta \boldsymbol{X}_{(K-1)\eta}\|_{\mathbf{A}_\sigma}^2] \\ &\le 4(1+\alpha)\gamma_2\beta^{-1}d\eta^3 + (1+1/\alpha)\mathbb{E}[\|\boldsymbol{x}_{K-1} - \boldsymbol{X}_{(K-1)\eta}\|_{\mathbf{A}_\sigma}^2], \end{aligned} \tag{C.5}$$

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 C.2 and C.3. Plugging (C.4) and (C.5) into (C.3) gives

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

Then, by recursively applying the above inequality, we obtain

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

Letting $\alpha = K$ and applying the inequality $(1+1/K)^K - 1 \le e - 1 \le 2$, the above inequality implies

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

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

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

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

C.3. Proof of Lemma B.3. In order to prove Lemma B.3, we require the following lemmas.

LEMMA C.4. *Under Assumptions* 1 *and* 2, *for all* $k \ge 0$, *there exists a constant* $c_1 \in [\|\mathbf{A}_\sigma\|_2^{-1}, 1)$ *such that*

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

LEMMA C.5 (Theorem 2.3 in [5]). *Letting* μ, ν *be two probability measures with finite exponential second moments, it holds that*

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

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 C.6 (Lemma 4 in [38]). *Letting $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ be the standard Gaussian random vector with dimension d, it holds that*

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

LEMMA C.7. *Under Assumptions 1 and 2, let $\boldsymbol{X}_t$ denote the solution of LS-LD (2.3) 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 B.3.

Proof of Lemma B.3. By Lemma C.5, 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 [15, 35, 40], we leverage the following continuous-time interpolation of LS-SGLD:

$$\tilde{\boldsymbol{X}}_t = \int_0^t -\mathbf{A}_\sigma^{-1}\mathbf{G}_s \mathrm{d}s + \int_0^t \sqrt{2\beta^{-1}}\mathbf{A}_\sigma^{-1/2}\mathrm{d}\boldsymbol{B}_s, \tag{C.6}$$

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 (C.6) does not form a Markov chain since it contains randomness of the stochastic gradient. To tackle this, we leverage the results in [22] and construct the following Markov chain to mimic (C.6):

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

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 [22]. Then letting $\mathbb{P}_t$ and $\mathbb{Q}_t$ denote the distribution of $\boldsymbol{X}_t$ and $\hat{\boldsymbol{X}}_t$ respectively, by the Girsanov formula, the Radon–Nikodym derivative of $\mathbb{P}_t$ with respect to $\mathbb{Q}_t$ can be derived as follows:

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

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

$$\begin{aligned} D_{KL}(\mathbb{Q}_T||\mathbb{P}_T) &= -\int_{\mathbb{R}^d} \log\bigg(\frac{\mathrm{d}\mathbb{P}_T}{\mathrm{d}\mathbb{Q}_T}\bigg)\mathrm{d}\mathbb{Q}_T \\ &= \frac{\beta}{4}\int_0^T \mathbb{E}\big[\|\mathbf{A}_\sigma^{-1}\nabla f(\hat{\boldsymbol{X}}_s) - \mathbf{A}_\sigma^{-1}\hat{\mathbf{G}}_s\|_2^2\big]\mathrm{d}s \\ &= \frac{\beta}{4}\sum_{k=0}^{K-1}\int_{k\eta}^{(k+1)\eta} \mathbb{E}\big[\|\mathbf{A}_\sigma^{-1}\nabla f(\tilde{\boldsymbol{X}}_s) - \mathbf{A}_\sigma^{-1}\mathbf{g}_k\|_2^2\big]\mathrm{d}s, \end{aligned}$$

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

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

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

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

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 $s-k\eta\le\eta$, it follows that

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

Regarding I_2, based on Lemma A.1 in [25] and Assumption 3, we have

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

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

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

By Lemma D.1 and Young's inequality, we know that

$$\mathbb{E}\big[\|\mathbf{g}_k\|_{\mathbf{A}_\sigma^{-2}}^2\big]\le\mathbb{E}[\|\mathbf{g}_k\|_2^2]\le 2M^2\mathbb{E}[\|\boldsymbol{x}_k\|_2^2]+2G^2\le\frac{4M^2(2b+\beta^{-1}d)}{c_1m}+2G^2,$$

where $G=\max_{i\in[n]}\|\nabla f_i(0)\|_2$. Then by Lemma C.6, we have $\mathbb{E}[\|\boldsymbol{\epsilon}_k\|_{\mathbf{A}_\sigma^{-1}}^2]\le\gamma_2 d$ 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,

$$D_{KL}(\mathbb{Q}_T||\mathbb{P}_T)\le\frac{2\beta M^4(2b+\beta^{-1}d)+\beta c_1mM^2G^2}{c_1m}K\eta^3+\gamma_2M^2dK\eta^2+\frac{\gamma_1\beta d\omega^2}{2B}K\eta.$$

For sufficiently small step size such that

$$\eta\le\frac{c_1\beta^{-1}\gamma_2md}{2M^2(2b+\beta^{-1}d)+c_1mG^2},$$

we have

$$D_{KL}(\mathbb{P}(\boldsymbol{x}_K)||\mathbb{P}(\boldsymbol{X}_{K\eta}))\le\frac{\gamma_1\beta d\omega^2}{2B}K\eta+2\gamma_2M^2dK\eta^2.$$

Then, by Lemma C.5, we have

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

where Γ can be further bounded as

$$\begin{aligned}\Gamma &\le \left(\frac{3}{2} + \log \mathbb{E}[e^{\|\boldsymbol{X}_{K\eta}\|_2^2}]\right)^{1/2} \\ &\le \big(3/2 + 2(b+\beta d)K\eta\big)^{1/2} \\ &\le \big(3/2 + 2(b+\beta d)\big)^{1/2} \cdot (K\eta)^{1/2},\end{aligned}$$

where the first inequality is by the choice $\alpha = 1$, the second inequality is by Lemma C.7, and the last inequality is by the assumption that $K\eta > 1$. Therefore, defining $\bar{\Gamma} = \big(3/2 + 2(b+\beta^{-1}d)\big)^{1/2}$, the 2-Wasserstein distance $\mathcal{W}_2\big(\mathbb{P}(\boldsymbol{x}_K), \mathbb{P}(\boldsymbol{X}_{K\eta})\big)$ can be bounded by

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

which completes the proof. ∎

Appendix D. Proof of lemmas in Appendix C.

D.1. Proof of Lemma C.2.

Proof of Lemma C.2. Assuming shared Brownian motions in $\boldsymbol{W}_t$ and $\boldsymbol{V}_t$, we have

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

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

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

which completes the proof. ∎

D.2. Proof of Lemma C.3.

Proof of Lemma C.3. 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

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

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

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

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

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

where the inequality is by Jensen's inequality and Lemma C.6 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 D.1, 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 an iterate of LS-SGLD; by Lemma C.4 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

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

Noting that $\tau, t \leq \eta$, plugging the above inequality into (D.1), 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

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

where $\wedge$ stands for maximum of two numbers,

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 C.4. In order to prove Lemma C.4, we need the following lemma.

LEMMA D.1 (Lemma 3.1 in [35]). *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 of Lemma C.4. 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

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

where the second equality follows from the fact that $\mathbb{E}[\|\boldsymbol{\epsilon}_k\|_2^2] = d$. Noting 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 D.1, 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

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

Since $\eta \leq 1/(c_1 m) \wedge b/G$, (D.2) 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. ∎

D.4. Proof of Lemma C.7.

Proof of Lemma C.7. 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

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

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

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

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

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

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

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

which completes the proof. ∎

REFERENCES

[1] A. Attia, R. Ştefănescu, and A. Sandu, *The reduced-order hybrid Monte Carlo sampling smoother*, Internat. J. Numer. Methods Fluids, 83 (2017), pp. 28–51.

[2] D. Bakry, I. Gentil, and M. Ledoux, *Analysis and Geometry of Markov Diffusion Operators*, Grundlehren Math. Wiss. 348, Springer, Berlin, 2013.

[3] R. Bardenet, A. Doucet, and C. Holmes, *On Markov chain Monte Carlo methods for tall data*, J. Mach. Learn. Res., 18 (2017), pp. 1515–1557.

[4] M. Betancourt, *The fundamental incompatibility of scalable hamiltonian Monte Carlo and naive data subsampling*, in Proceedings of the International Conference on Machine Learning, 2015, pp. 533–540.

[5] F. Bolley and C. Villani, *Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities*, Ann. Fac. Sci. Toulouse Math. (6), 14 (2005), https://doi.org/10.5802/afst.1095.

[6] J. Borggaard, N. Glatt-Holtz, and J. Krometis, *A Bayesian approach to estimating background flows from a passive scalar*, SIAM/ASA J. Uncertain. Quantif., 8 (2020), pp. 1036–1060.

[7] A. Bovier, M. Eckhoff, V. Gayrard, and M. Klein, *Metastability in reversible diffusion processes* I: *Sharp asymptotics for capacities and exit times*, J. Eur. Math. Soc. (JEMS), 6 (2004), pp. 399–424.

[8] N. S. Chatterji, N. Flammarion, Y.-A. Ma, P. L. Bartlett, and M. I. Jordan, *On the Theory of Variance Reduction for Stochastic Gradient Monte Carlo*, preprint, arXiv:1802.05431, 2018.

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

[10] T. Chen, E. Fox, and C. Guestrin, *Stochastic gradient Hamiltonian Monte Carlo*, in Proceedings of the International Conference on Machine Learning, 2014.

[11] Y. Chen, R. Dwivedi, M. J. Wainwright, and B. Yu, *Fast Mixing of Metropolized Hamiltonian Monte Carlo: Benefits of Multi-Step Gradients*, preprint, arXiv:1905.12247, 2019.

[12] X. Cheng, N. S. Chatterji, Y. Abbasi-Yadkori, P. L. Bartlett, and M. I. Jordan, *Sharp Convergence Rates for Langevin Dynamics in the Nonconvex Setting*, preprint, arXiv:1805.01648, 2018.

[13] X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan, *Underdamped Langevin mcmc: A non-asymptotic analysis*, in Proceedings of the 31st Conference on Learning Theory, Vol. 75, 2018, pp. 300–323.

[14] T.-S. Chiang, C.-R. Hwang, and S. J. Sheu, *Diffusion for global optimization in r^n*, SIAM J. Control Optim., 25 (1987), pp. 737–753.

[15] A. S. Dalalyan, *Theoretical guarantees for approximate sampling from smooth and log-concave densities*, J. R. Stat. Soc. Ser. B Stat. Methodol., 79 (2017), pp. 651–676.

[16] A. S. Dalalyan and A. G. Karagulyan, *User-Friendly Guarantees for the Langevin Monte Carlo with Inaccurate Gradient*, preprint, arXiv:1710.00095, 2017.
[17] K.-D. Dang, M. Quiroz, R. Kohn, M.-N. Tran, and M. Villani, *Hamiltonian Monte Carlo with energy conserving subsampling*, J. Mach. Learn. Res., 20 (2019), pp. 1–31.
[18] K. A. Dubey, S. J. Reddi, S. A. Williamson, B. Poczos, A. J. Smola, and E. P. Xing, *Variance reduction in stochastic gradient Langevin dynamics*, in Advances in Neural Information Processing Systems, 2016, pp. 1154–1162.
[19] A. Durmus and E. Moulines, *Nonasymptotic convergence analysis for the unadjusted Langevin algorithm*, Ann. Appl. Probab., 27 (2017), pp. 1551–1587.
[20] R. Dwivedi, Y. Chen, M. J. Wainwright, and B. Yu, *Log-concave sampling: Metropolis-Hastings algorithms are fast!*, in Proceedings of the 31st Conference on Learning Theory, 2018, pp. 793–797.
[21] G. Golub and C. Van Loan, *Matrix Computation*, 3rd ed., Johns Hopkins University Press, Baltimore, 1996.
[22] I. Gyöngy, *Mimicking the one-dimensional marginal distributions of processes having an Itô differential*, Probab. Theory Related fields, 71 (1986), pp. 501–516.
[23] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Springer, Berlin, 1992.
[24] L. Kreusser, S. Osher, and B. Wang, *A Deterministic Approach to Avoid Saddle Points*, preprint, arXiv:1901.06827, 2019.
[25] L. Lei, C. Ju, J. Chen, and M. I. Jordan, *Non-convex finite-sum optimization via SCSG methods*, in Advances in Neural Information Processing Systems 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., 2017, pp. 2348–2358.
[26] C. Li, C. Chen, D. Carlson, and C. Lawrence, *Preconditioned stochastic gradient Langevin dynamics for deep neural networks*, in Proceedings of the Association for the Advancement of Artificial Intelligence, 2016.
[27] Z. Li, T. Zhang, and J. Li, *Stochastic Gradient Hamiltonian Monte Carlo with Variance Reduction for Bayesian Inference*, preprint, arXiv:1803.11159, 2018.
[28] Y.-A. Ma, T. Chen, and E. Fox, *A complete recipe for stochastic gradient MCMC*, in Advances in Neural Information Processing Systems, 2015, pp. 2917–2925.
[29] P. A. Markowich and C. 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 19, Brazilian Mathematical Society, 1999.
[30] J. C. Mattingly, A. M. Stuart, and D. J. Higham, *Ergodicity for SDEs and approximations: Locally Lipschitz vector fields and degenerate noise*, Stochastic Process. Appl., 101 (2002), pp. 185–232.
[31] W. Mou, L. Wang, X. Zhai, and K. Zheng, *Generalization Bounds of SGLD for Non-Convex Learning: Two Theoretical Viewpoints*, preprint, arXiv:1707.05947, 2017.
[32] R. M. Neal, *MCMC using hamiltonian dynamics*, in Handbook of Markov Chain Monte Carlo, Handbooks Modern Statist. Methods 2, Routledge, 2011, pp. 113–162.
[33] S. Osher, B. Wang, P. Yin, X. Luo, M. Pham, and A. Lin, *Laplacian Smoothing Gradient Descent*, preprint, arXiv:1806.06317, 2018.
[34] G. Parisi, *Correlation functions and computer simulations*, Nuclear Phys. B, 180 (1981), pp. 378–384.
[35] M. Raginsky, A. Rakhlin, and M. Telgarsky, *Non-convex learning via stochastic gradient Langevin dynamics: A nonasymptotic analysis*, in Proceedings of the Conference on Learning Theory, 2017, pp. 1674–1703.
[36] Y. W. Teh, A. H. Thiery, and S. J. Vollmer, *Consistency and fluctuations for stochastic gradient Langevin dynamics*, J. Mach. Learn. Res., 17 (2016), pp. 193–225.
[37] S. J. Vollmer, K. C. Zygalakis, and Y. W. Teh, *Exploration of the (non-) asymptotic bias and variance of stochastic gradient Langevin dynamics*, J. Mach. Learn. Res., 17 (2016), pp. 5504–5548.
[38] B. Wang, Q. Gu, M. Boedihardjo, F. Barekat, and S. Osher, *DP-LSSGD: A Stochastic Optimization Method to Lift the Utility in Privacy-Preserving ERM*, preprint, arXiv:1906.12056, 2019.
[39] M. Welling and T. Y. Whye, *Bayesian learning via stochastic gradient Langevin dynamics*, in Proceedings of the International Conference on Machine Learning, 2011.
[40] P. Xu, J. Chen, D. Zou, and Q. Gu, *Global convergence of Langevin dynamics based algorithms for nonconvex optimization*, in Advances in Neural Information Processing Systems, 2018, pp. 3122–3133.

[41] C. Zhang, B. Shahbaba, and H. Zhao, *Hamiltonian Monte Carlo acceleration using surrogate functions with random bases*, Statist. Comput., 27 (2017), pp. 1473–1490.
[42] D. Zou, P. Xu, and Q. Gu, *Stochastic variance-reduced Hamilton Monte Carlo methods*, in Proceedings of the 35th International Conference on Machine Learning, 2018, pp. 6028–6037.
[43] D. Zou, P. Xu, and Q. Gu, *Subsampled stochastic variance-reduced gradient Langevin dynamics*, in Proceedings of the International Conference on Uncertainty in Artificial Intelligence, 2018.
[44] D. Zou, P. Xu, and Q. Gu, *Sampling from non-log-concave distributions via variance-reduced gradient Langevin dynamics*, in Artificial Intelligence and Statistics, Proc. Machine Learning Res. 89, MLR Press, 2019, pp. 2936–2945.
[45] D. Zou, P. Xu, and Q. Gu, *Stochastic gradient Hamiltonian Monte Carlo methods with recursive variance reduction*, in Advances in Neural Information Processing Systems, 2019.