Faster Convergence of Stochastic Gradient Langevin Dynamics for Non-Log-Concave Sampling

Difan Zou[1] **Pan Xu**[1] **Quanquan Gu**[1]

[1]Department of Computer Science, University of California, Los Angeles, CA 90095, USA

Abstract

We provide a new convergence analysis of stochastic gradient Langevin dynamics (SGLD) for sampling from a class of distributions that can be non-log-concave. At the core of our approach is a novel conductance analysis of SGLD using an auxiliary time-reversible Markov Chain. Under certain conditions on the target distribution, we prove that $\widetilde{O}(d^4\epsilon^{-2})$ stochastic gradient evaluations suffice to guarantee ϵ-sampling error in terms of the total variation distance, where d is the problem dimension. This improves existing results on the convergence rate of SGLD [Raginsky et al., 2017, Xu et al., 2018]. We further show that provided an additional Hessian Lipschitz condition on the log-density function, SGLD is guaranteed to achieve ϵ-sampling error within $\widetilde{O}(d^{15/4}\epsilon^{-3/2})$ stochastic gradient evaluations. Our proof technique provides a new way to study the convergence of Langevin based algorithms, and sheds some light on the design of fast stochastic gradient based sampling algorithms.

1 INTRODUCTION

We study the problem of sampling from a target distribution using Langevin dynamics [Langevin, 1908] based algorithms. Mathematically, Langevin dynamics (a.k.a., overdamped Langevin dynamics) is defined by the following stochastic differential equation (SDE)

$$\mathrm{d}\boldsymbol{X}(t) = -\nabla f\big(\boldsymbol{X}(t)\big)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}\boldsymbol{B}(t), \qquad (1.1)$$

where $\beta > 0$ is called the inverse temperature parameter and $\boldsymbol{B}(t) \in \mathbb{R}^d$ is the Brownian motion at time t. It has been proved in Chiang et al. [1987], Roberts and Tweedie [1996] that under certain conditions on the drift term $-\nabla f(\boldsymbol{X}(t))$, the Langevin dynamics will converge to a unique stationary distribution $\pi(\mathrm{d}\mathbf{x}) \propto e^{-\beta f(\mathbf{x})}\mathrm{d}\mathbf{x}$. To approximately sample from such a target distribution π, we can apply the Euler-Maruyama discretization onto (1.1), leading to the Langevin Monte Carlo algorithm (LMC), which iteratively updates the parameter $\mathbf{x}_k$ as follows

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta\nabla f(\mathbf{x}_k) + \sqrt{2\eta\beta^{-1}}\cdot\boldsymbol{\epsilon}_k, \qquad (1.2)$$

where $k = 0, 1, \ldots$ denotes the time step, $\{\boldsymbol{\epsilon}_k\}_{k=0,1,\ldots}$ are i.i.d. standard Gaussian random vectors in $\mathbb{R}^d$, and $\eta > 0$ is the step size of the discretization.

In large scale machine learning problems that involve a large amount of training data, the log-density function $f(\mathbf{x})$ can be typically formulated as the average of the log-density functions over all the training data points, i.e., $f(\mathbf{x}) = n^{-1}\sum_{i=1}^n f_i(\mathbf{x})$[1], where n is the size of training dataset and $f_i(\mathbf{x})$ denotes the log-density function for the i-th training data point. In these problems, the computation of the full gradient over the entire dataset can be very time-consuming. In order to save the cost of gradient computation, one can replace the full gradient $\nabla f(\mathbf{x})$ with a stochastic gradient computed only over a small subset of the dataset, which gives rise to stochastic gradient Langevin dynamics (SGLD) [Welling and Teh, 2011].

When the target distribution π is log-concave, SGLD provably converges to π at a sublinear rate in 2-Wasserstein distance [Dalalyan and Karagulyan, 2019, Dalalyan, 2017a, Wang et al., 2019]. However, it becomes much more challenging to establish the convergence of SGLD when the target distribution is not log-concave. When the negative log-density function $f(\mathbf{x})$ is smooth and dissipative, the global convergence guarantee of SGLD has been firstly established in Raginsky et al. [2017][2] via the optimal control

[1]In some cases, the log-density function $f(\mathbf{x})$ is formulated as the sum of the log-density functions for training data points instead of the average. To cover these cases, we can simply transform the temperature parameter $\beta \to n\beta$ and thus the target distribution remains the same.

[2]Although this paper mainly focuses on the convergence anal-

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).

theory and further improved in Xu et al. [2018] by a direct analysis of the ergodicity of LMC. Nonetheless, these two works require extremely large mini-batch size (e.g., $B = \Omega(\epsilon^{-4})$) to ensure sufficiently small sampling error, which is prohibitively large or even unrealistic compared with the practical setting. Zhang et al. [2017] studied the hitting time of SGLD for nonconvex optimization, but can only provide the convergence guarantee for finding a local minimum rather than converging to the target distribution. Recently, Chau et al. [2019], Zhang et al. [2019] studied the global convergence of SGLD for nonconvex stochastic optimization problems and proved faster convergence rates than those in Raginsky et al. [2017], Xu et al. [2018]. However, their convergence results require an additional Lipschitz condition in terms of the input data (rather than the model parameter) on the stochastic gradients, which restricts their applications to a small class of SGLD-based sampling problems.

In this paper, we consider the same setting in Raginsky et al. [2017], Xu et al. [2018] and aim to establish faster convergence rates for SGLD with an arbitrary mini-batch size. In particular, we provide a new convergence analysis for SGLD based on an auxiliary time-reversible Markov chain called Metropolized SGLD [Zhang et al., 2017], which is constructed by adding a Metropolis-Hasting step to SGLD[3]. The key idea is that as long as the transition kernel of the constructed Metropolized SGLD chain is sufficiently close to that of SGLD, we can prove the convergence of SGLD to the target distribution. Compared with existing proof techniques that typically take LMC or Langevin dynamics as an auxiliary sequence, the advantage of using Metropolized SGLD as the auxiliary sequence is that it is closer to SGLD in distribution as its transition distribution also covers the randomness of stochastic gradients, thus can better characterize the convergence behavior of SGLD and lead to sharper convergence guarantees. To sum up, we highlight our main contributions as follows:

- We provide a new convergence analysis of SGLD for sampling a large class of distributions that can be non-log-concave. In contrast to Raginsky et al. [2017], Xu et al. [2018] that require a very large mini-batch size, our convergence guarantee holds for an arbitrary choice of mini-batch size.
- We prove that SGLD can achieve ϵ-sampling error in total variation distance within $\widetilde{O}(d^4\beta^2\rho^{-4}\epsilon^{-2})$ stochastic gradient evaluations, where d is the problem dimension, β is the inverse temperature parameter, and ρ is the Cheeger constant (See Definition 4.2) of a truncated version of the target distribution. We also prove the convergence of SGLD under the measure of polynomial growth functions, which suggests that the number of required stochastic gradient evaluations is $\widetilde{O}(\epsilon^{-2})$. This improves the state-of-the-art result proved in Xu et al. [2018] by a factor of $\widetilde{O}(\epsilon^{-3})$.
- We further establish sharper convergence guarantees for SGLD under an additional Hessian Lipschitz condition on the negative log density function $f(\mathbf{x})$. We show that $\widetilde{O}(d^{15/4}\beta^{7/4}\rho^{-7/2}\epsilon^{-3/2})$ stochastic gradient evaluations suffice to achieve ϵ-sampling error in total variation distance. Our proof technique is much simpler and more intuitive than existing analysis for proving the convergence of Langevin algorithms under the Hessian Lipschitz condition [Dalalyan and Karagulyan, 2019, Mou et al., 2019, Vempala and Wibisono, 2019], which can be of independent interest.

Notation. We use the notation $x \wedge y$ and $x \vee y$ to denote $\min\{x, y\}$ and $\max\{x, y\}$ respectively. We denote by $\mathcal{B}(\mathbf{u}, r)$ the Euclidean of radius $r > 0$ centered at $\mathbf{u} \in \mathbb{R}^d$. For any distribution μ and set $\mathcal{A}$, we use $\mu(\mathcal{A})$ to denote the probability measure of $\mathcal{A}$ under the distribution μ. For any two distributions μ and ν, we use $\|\mu-\nu\|_{TV}$ and $D_{KL}(\mu, \nu)$ to denote the total variation distance and Kullback–Leibler divergence between μ and ν respectively. For $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$, we use $\mathcal{T}_{\mathbf{u}}(\mathbf{v})$ to denote the probability of transiting to $\mathbf{v}$ after one step SGLD update from $\mathbf{u}$. Similarly, $\mathcal{T}_{\mathbf{u}}(\mathcal{A})$ and $\mathcal{T}_{\mathcal{A}'}(\mathcal{A})$ are the probabilities of transiting to a set $\mathcal{A} \subseteq \mathbb{R}^d$ after one step SGLD update starting from $\mathbf{u}$ and the set $\mathcal{A}'$ respectively. For any two sequences $\{a_n\}$ and $\{b_n\}$, we denote $a_n = O(b_n)$ and $a_n = \Omega(b_n)$ if $a_n \leq C_1 b_n$ or $a_n \geq C_2 b_n$ for some absolute constants C_1 and C_2. We use notations $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ to hide polylogarithmic factors in $O(\cdot)$ and $\Omega(\cdot)$ respectively.

2 RELATED WORK

Markov Chain Monte Carlo (MCMC) methods, such as random walk Metropolis [Mengersen et al., 1996], ball walk [Lovász and Simonovits, 1990], hit-and-run [Smith, 1984] and Langevin algorithms [Parisi, 1981], have been extensively studied for sampling from a target distribution, and widely used in many machine learning applications. There are a large number of works focusing on developing fast MCMC algorithms and establishing sharp theoretical guarantees. We will review the most related works among them due to the space limit.

Langevin dynamics (1.1) based algorithms have recently aroused as a promising method for accurate and efficient Bayesian sampling in both theory and practice [Welling and Teh, 2011, Dalalyan, 2017b]. The non-asymptotic convergence rate of LMC has been extensively investigated in the literature when the target distribution is strongly log-concave [Durmus and Moulines, 2016, Dalalyan, 2017b,

ysis of SGLD for nonconvex optimization, part of its theoretical results also reveal the convergence rate for sampling from a target distribution.

[3]This Markov chain is practically intractable and is only used for the sake of theoretical analysis.

Durmus et al., 2017b], weakly log-concave [Dalalyan, 2017a, Mangoubi and Vishnoi, 2019], and non-log-concave but admits certain good isoperimetric properties [Raginsky et al., 2017, Ma et al., 2018, Lee et al., 2018, Xu et al., 2018, Vempala and Wibisono, 2019], to mention a few. The stochastic variant of LMC, i.e., SGLD, is often studied together in the above literature and the convex/nonconvex optimization field [Raginsky et al., 2017, Zhang et al., 2017, Xu et al., 2018, Gao et al., 2018, Chen et al., 2019a, Deng et al., 2020]. Another important Langevin based algorithm is the Metropolis Adjusted Langevin Algorithms (MALA) [Roberts and Tweedie, 1996], which is developed by introducing a Metropolis-Hasting step into LMC. Theoretically, it has been proved that MALA converges to the target distribution at a linear rate for sampling from both strongly log-concave [Dwivedi et al., 2018] and non-log-concave [Bou-Rabee and Hairer, 2013] distributions.

Beyond first-order MCMC methods, there has also emerged extensive work on high-order MCMC methods. One popular algorithm among them is Hamiltonian Monte Carlo (HMC) [Neal et al., 2011], which introduces a Hamiltonian momentum and leapfrog integrator to accelerate the mixing rate. From the theoretical perspective, Durmus et al. [2017a] established general conditions under which HMC can be guaranteed to be geometrically ergodic. Mangoubi and Vishnoi [2018, 2019] proved the convergence rate of HMC for sampling both log-concave and non-log-concave distributions. Bou-Rabee et al. [2018], Chen et al. [2019b] studied the convergence of Metropolized HMC (MHMC) for sampling strongly log-concave distributions. Another important high-order MCMC method are built upon the underdamped Langevin dynamics, which incorporates the velocity into the Langevin dynamics (1.1). For continuous-time underdamped Langevin dynamics, its mixing rate has been studied in Eberle [2016], Eberle et al. [2017]. The convergence of its discrete version has also been widely studied for sampling from both log-concave [Chen et al., 2017, Zou et al., 2018] and non-log-concave distributions [Chen et al., 2015, Cheng et al., 2018, Gao et al., 2018, Zou et al., 2019b].

3 REVIEW OF THE SGLD ALGORITHM

For the completeness, we present the SGLD algorithm [Welling and Teh, 2011] in Algorithm 1, which is built upon the Euler-Maruyama discretization of the continuous-time Langevin dynamics (1.1) while using mini-batch stochastic gradient in each iteration.

In the k-th iteration, SGLD samples a mini-batch of data points without replacement, denoted by $\mathcal{I}$, and computes the stochastic gradient at the current iterate $\mathbf{x}_k$, i.e., $\mathbf{g}(\mathbf{x}_k, \mathcal{I}) = 1/B \sum_{i\in\mathcal{I}} \nabla f_i(\mathbf{x}_k)$, where $B = |\mathcal{I}|$ is the mini-batch size. Based on the stochastic gradient, the model parameter is

Algorithm 1 Stochastic Gradient Langevin Dynamics (SGLD)

input: step size η; mini-batch size B; inverse temperature parameter β;
Randomly draw $\mathbf{x}_0$ from initial distribution μ_0.
for $k = 0, 1, \ldots, K$ **do**
  Randomly pick a subset $\mathcal{I}$ from $\{1, \ldots, n\}$ of size $|\mathcal{I}| = B$; randomly draw $\boldsymbol{\epsilon}_k \sim N(\mathbf{0}, \mathbf{I})$
  Compute the stochastic gradient $\mathbf{g}(\mathbf{x}_k, \mathcal{I}) = 1/B \sum_{i\in\mathcal{I}} \nabla f_i(\mathbf{x}_k)$
  Update: $\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \mathbf{g}(\mathbf{x}_k, \mathcal{I}) + \sqrt{2\eta/\beta}\boldsymbol{\epsilon}_k$
end for
output: $\mathbf{x}_K$

updated using the following rule,

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \mathbf{g}(\mathbf{x}_k, \mathcal{I}) + \sqrt{2\eta/\beta} \cdot \boldsymbol{\epsilon}_k,$$

where $\boldsymbol{\epsilon}_k$ is randomly drawn from a standard normal distribution $N(\mathbf{0}, \mathbf{I})$ and $\eta > 0$ is the step size.

4 MAIN RESULTS

In this section, we present our main theoretical results. We start with the following two definitions. The first one quantifies the goodness of the initial distribution compared with the target distribution, and the second one characterizes the isoperimetric profile of a given distribution. Both definitions are widely used in the convergence analysis of MCMC methods [Lovász and Simonovits, 1993, Vempala, 2007, Dwivedi et al., 2018, Mangoubi and Vishnoi, 2019].

Definition 4.1 (λ-warm start)**.** Let ν be a distribution on Ω. We say the initial distribution μ_0 is a λ-warm start with respect to ν if

$$\sup_{\mathcal{A}:\mathcal{A}\subseteq\Omega} \frac{\mu_0(\mathcal{A})}{\nu(\mathcal{A})} \leq \lambda.$$

Definition 4.2 (Cheeger constant)**.** Let μ be a probability measure on Ω. We say μ satisfies the isoperimetric inequality with Cheeger constant ρ if for any $\mathcal{A} \in \Omega$, it holds that

$$\liminf_{h\to 0^+} \frac{\mu(\mathcal{A}_h) - \mu(\mathcal{A})}{h} \geq \rho \min\big\{\mu(\mathcal{A}), 1 - \mu(\mathcal{A})\big\},$$

where $\mathcal{A}_h = \{\mathbf{x} \in \Omega : \exists \mathbf{y} \in \mathcal{A}, \|\mathbf{x} - \mathbf{y}\|_2 \leq h\}$.

Next, we introduce some common assumptions on the negative log density function $f(\mathbf{x})$ and stochastic gradients $\mathbf{g}(\mathbf{x}, \mathcal{I})$.

Assumption 4.3 (Dissipativeness)**.** There are absolute constants $m > 0$ and $b \geq 0$ such that

$$\langle \nabla f(\mathbf{x}), \mathbf{x} \rangle \geq m\|\mathbf{x}\|_2^2 - b, \quad \text{for all } \mathbf{x} \in \mathbb{R}^d.$$

This assumption has been conventionally made in the convergence analysis for sampling form non-log-concave distributions [Raginsky et al., 2017, Xu et al., 2018, Zou et al., 2019a]. Basically, this assumption implies that the log density function $f(\mathbf{x})$ grows like a quadratic function when $\mathbf{x}$ is outside a ball centered at the origin. Note that a strongly convex function $f(\mathbf{x})$ simply satisfies Assumption 4.3, but not vice versa.

Assumption 4.4 (Smoothness)**.** There exists a positive constant L such that for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and all functions $f_i(\mathbf{x})$,

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2.$$

This assumption has also been made in many prior works [Raginsky et al., 2017, Zhang et al., 2017, Xu et al., 2018].

We now define the following function that will be repeatedly used in the subsequent theoretical results:

$$\bar{R}(z) = \bigg[\max\bigg\{\frac{625d\log(4/z)}{m\beta}, \frac{4d\log(4L/m) + 4\beta b}{m\beta}, \frac{4d + 8\sqrt{d\log(1/z)} + 8\log(1/z)}{m\beta}\bigg\}\bigg]^{1/2}. \tag{4.1}$$

Based on all aforementioned assumptions, we present the convergence result of SGLD in the following theorem.

Theorem 4.5. For any $\epsilon \in (0,1)$, let $\pi^* \propto e^{-\beta f(\mathbf{x})}\mathbb{1}\big(\mathbf{x} \in \mathcal{B}(0,R)\big)$ be the truncated target distribution in $\Omega = \mathcal{B}(0,R)$ with $R = \bar{R}(\epsilon K^{-1}/12)$, and ρ be the Cheeger constant of π^*. Under Assumptions 4.3 and 4.4, we suppose $\mathbb{P}(\|\mathbf{x}_0\|_2 \leq R/2) \leq \epsilon/16$, and set the step size as $\eta = \widetilde{O}(\rho^2 d^{-2}\beta^{-1} \wedge B^2\rho^2 d^{-4}\beta^{-1})$, then for any λ-warm start with respect to π, the output of Algorithm 1 satisfies

$$\|\mu_K^{\mathrm{SGLD}} - \pi\|_{TV} \leq \lambda(1 - C_0\eta)^K + \frac{C_1\eta^{1/2}}{B} + C_2\eta^{1/2} + \frac{\epsilon}{2},$$

where $C_0 = \widetilde{O}\big(\rho^2\beta^{-1}\big)$, $C_1 = \widetilde{O}\big(Rd\rho^{-1}\beta^{3/2}\big)$ and $C_2 = \widetilde{O}\big(d\rho^{-1}\beta^{1/2}\big)$ are problem-dependent constants.

Theorem 4.5 shows that the total variation distance between the distributions μ_K^{SGLD} and π can be upper bounded by the sum of four terms. Specifically, the first term corresponds to the sampling error of Metropolized SGLD, which converges to zero at a linear rate. The second and third terms correspond to the approximation error between SGLD and Metropolized SGLD, which is in the order of $O(\eta^{1/2})$. Moreover, the third the third term corresponds to the variance of stochastic gradients, which decreases when increasing the mini-batch size B. The last term can be understood as an approximation error that comes from the technical proof.

Remark 4.6. For a general non-log-concave distribution, it is difficult to prove a tight bound on the Cheeger constant ρ. One possible lower bound of ρ can be obtained via Buser's inequality [Buser, 1982, Ledoux, 1994], which shows that the Cheeger constant ρ can be lower bounded by $\Omega(d^{-1/2}c_p)$ under Assumption 4.4, where c_p is the Poincaré constant of the distribution $\pi^\star$. Moreover, Bakry et al. [2008] gave a simple lower bound of c_p, showing that $c_p \geq e^{-\beta \mathrm{Osc}_R f}/(2R^2)$, where $\mathrm{Osc}_R f = \sup_{\mathbf{x}\in\mathcal{B}(\mathbf{0},R)} f(\mathbf{x}) - \inf_{\mathbf{x}\in\mathcal{B}(\mathbf{0},R)} f(\mathbf{x}) \leq LR^2/2$. Assuming $R = \widetilde{O}(d^{1/2})$, this further implies that $\rho = \Omega(d^{-1}) \cdot e^{-O(R^2)} = e^{-\widetilde{O}(d)}$. In addition, better lower bounds of ρ can be proved when the target distribution enjoys better properties. When the target distribution is a mixture of strongly log-concave distributions, the lower bound of ρ can be improved to $1/\mathrm{poly}(d)$ [Lee et al., 2018]. Strengthening Assumption 4.3 to a local nonconvexity condition yields $\rho = e^{-O(L)}$ [Ma et al., 2018]. The lower bound of Cheeger constant has been extensively studied for log-concave distributions [Kannan et al., 1995, Lee and Vempala, 2017, Chen, 2021], among them Lee and Vempala [2017] proved that the Cheeger constant ρ can be lower bounded by $\rho = \Omega\big(1/(\mathrm{Tr}(\mathbf{\Sigma}^2))^{1/4}\big)$, where $\mathbf{\Sigma}$ is the covariance matrix of the distribution $\pi^\star$. When the target distribution is m-strongly log-concave, based on Cousins and Vempala [2014], Dwivedi et al. [2018], it can be shown that $\rho = \Omega(\sqrt{m})$.

Note that the upper bound of the sampling error proved in Theorem 4.5 relies on the step size, mini-batch size, and the goodness of the initialization (i.e., λ). In order to guarantee ϵ-sampling error of SGLD, we need to specify the choices of these hyper-parameters. In particular, we present the iteration complexity of SGLD in the following corollary.

Corollary 4.7. Under the same assumptions made in Theorem 4.5, consider Gaussian initialization $\mu_0 = N\big(\mathbf{0}, \mathbf{I}/(2\beta L)\big)$, then for any mini-batch size $B \leq n$ and $\epsilon \in (0,1)$, if set the step size and maximum iteration number as

$$\eta = \widetilde{O}\bigg(\frac{\rho^2\epsilon^2}{d^2\beta} \wedge \frac{B^2\rho^2\epsilon^2}{d^4\beta}\bigg),$$
$$K = \widetilde{O}\bigg(\frac{d^3\beta^2}{\rho^4\epsilon^2} \vee \frac{d^5\beta^2}{B^2\rho^4\epsilon^2}\bigg),$$

SGLD can achieve an ϵ sampling error in total variation distance.

It is worth noting that the iteration complexity in Corollary 4.7 holds for any mini-batch size $1 \leq B \leq n$, as opposed to Raginsky et al. [2017], Xu et al. [2018] that require the mini-batch size to be $\mathrm{poly}(\epsilon^{-1})$ in order to guarantee vanishing sampling error. Moreover, if we set the mini-batch size to be $B = O(d)$, the number of stochastic gradient evaluations needed to achieve ϵ-sampling error is $K \cdot B = \widetilde{O}(d^4\beta^2\rho^{-4}\epsilon^{-2})$.

Based on Corollary 4.7, we further show prove the convergence of SGLD under the measure of any polynomial

growth function.

Corollary 4.8. Under the same assumptions and hyperparameter configurations as in Corollary 4.7, let $h(\mathbf{x})$ be a polynomial growth function with degree D, i.e., $h(\mathbf{x}) \leq C(1 + \|\mathbf{x}\|_2^D)$ for some constant C, and K be defined in Corollary 4.7, then the output of SGLD satisfies

$$\mathbb{E}[h(\mathbf{x}_K)] - \mathbb{E}[h(\mathbf{x}^\pi)] \leq C'\epsilon,$$

where $\mathbf{x}^\pi \sim \pi$ denotes the random vector sampled from π and $C' = \widetilde{O}\big(d^{D/2}\big)$ is a problem-dependent constant.

Remark 4.9. Similar results have been presented in Sato and Nakagawa [2014], Chen et al. [2015], Vollmer et al. [2016], Erdogdu et al. [2018]. However, Sato and Nakagawa [2014] only analyzed the finite-time approximation error between SGLD and the SDE (1.1) rather than the convergence to the target distribution. The convergence results in Chen et al. [2015], Vollmer et al. [2016], Erdogdu et al. [2018] also differ from ours as their guarantees are made on the sample path average rather than the last iterate. In addition, these works assume that the Poisson equation solution of the SDE (1.1) has polynomially bounded i-th order derivative ($i \in \{2, 3, 4\}$), which is not required in our result.

Let us consider a special case that $h(\cdot) = f(\cdot)$, which was studied in Raginsky et al. [2017], Xu et al. [2018]. Assumption 4.4 implies that $h(\mathbf{x})$ is a quadratic growth function. Then Corollary 4.8 shows that in order to guarantee $\mathbb{E}[f(\mathbf{x}_k)] - \mathbb{E}[f(\mathbf{x}^\pi)] \leq \epsilon$, SGLD requires $\widetilde{O}(\epsilon^{-2})$ stochastic gradient evaluations. In contrast, in order to achieve the same error, Raginsky et al. [2017], Xu et al. [2018] require $\widetilde{O}(\epsilon^{-8})$ and $\widetilde{O}(\epsilon^{-5})$ stochastic gradient evaluations respectively, both of which are worse than ours.

5 IMPROVED CONVERGENCE RATES UNDER HESSIAN LIPSCHITZ CONDITION

In this section, we will show that the convergence rate of SGLD can be improved if the log density function additionally satisfies the Hessian Lipschitz condition, which is defined as follows.

Assumption 5.1 (Hessian Lipschitz)**.** There exists a positive constant H such that for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, it holds that

$$\big\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\big\|_{\text{op}} \leq H\|\mathbf{x} - \mathbf{y}\|_2.$$

This assumption has been made in many recent papers to prove faster convergence rate of LMC [Dalalyan and Karagulyan, 2019, Vempala and Wibisono, 2019, Mou et al., 2019] for sampling from both log-concave and non-log-concave distributions.

With this additional assumption, we state the convergence result of SGLD in the following theorem.

Theorem 5.2. For any $\epsilon \in (0, 1)$, let $\pi^* \propto e^{-\beta f(\mathbf{x})}\mathbb{1}\big(\mathbf{x} \in \mathcal{B}(0, R)\big)$ be the truncated target distribution in $\Omega = \mathcal{B}(0, R)$ with $R = \bar{R}(\epsilon K^{-1}/12)$, and ρ be the Cheeger constant of π^*. Under Assumptions 4.3, 4.4, and 5.1, suppose $\mathbb{P}(\|\mathbf{x}_0\|_2 \leq R/2) \leq \epsilon/16$. Setting the step size $\eta = \widetilde{O}\big(\rho^2 d^{-2}\beta^{-1}B^2 \wedge \rho/(d^{3/2} + d\beta^{1/2})\big)$, then for any λ-warm start with respect to π, the output of Algorithm 1 satisfies

$$\|\mu_K^{\text{SGLD}} - \pi\|_{TV} \leq \lambda(1 - C_0\eta)^K + \frac{C_1\eta^{1/2}}{B} + C_2\eta + \frac{\epsilon}{2},$$

where $C_0 = O(\beta^{-1}\rho^2)$, $C_1 = \widetilde{O}(R^2 d\rho^{-1}\beta^{3/2})$ and $C_2 = \widetilde{O}(d^{3/2}\rho^{-1} + Rd^{1/2}\beta\rho^{-1})$ are problem-dependent constants.

The four terms in Theorems 5.2 have the same meaning as those in Theorem 4.5. Compared with the convergence result in Theorem 4.5, the improvement brought by Hessian Lipschitz condition lies in the approximation error between the transition distributions of SGLD and Metropolized SGLD, which is improved from $O(\eta^{1/2})$ to $O\big(B^{-1}\eta^{1/2} + \eta\big)$.

Dalalyan and Karagulyan [2019], Mou et al. [2019], Vempala and Wibisono [2019] also improved the convergence rate of LMC using the Hessian Lipschitz condition. However, Dalalyan and Karagulyan [2019] only focused on strongly log-concave distributions and the theoretical results in Mou et al. [2019], Vempala and Wibisono [2019] cannot be easily extended to SGLD.

Corollary 5.3. Under the same assumptions made in Theorem 5.2, consider Gaussian initialization $\mu_0 = N\big(\mathbf{0}, \mathbf{I}/(2\beta L)\big)$, then for any mini-batch size $B \leq n$, if set the step size and maximum iteration number as

$$\eta = \widetilde{O}\bigg(\frac{\rho^2 B^2\epsilon^2}{d^2\beta} \wedge \frac{\rho\epsilon}{d^{3/2} + d\beta^{1/2}}\bigg),$$
$$K = \widetilde{O}\bigg(\frac{d^5\beta^2}{\rho^4 B^2\epsilon^2} + \frac{d^{5/2}\beta + d^2\beta^{3/2}}{\rho^3\epsilon}\bigg),$$

SGLD can achieve an ϵ sampling error in terms of total variation distance.

Note that the required number of stochastic gradient evaluations is $K \cdot B = \widetilde{O}\big(d^5\beta^2/(B\rho^4\epsilon^2) + Bd^{5/2}\beta^{3/2}/(\rho^3\epsilon)\big)$. Therefore, if setting the mini-batch size as $B = \widetilde{O}\big([d^{5/2}\beta^{1/2}\rho\epsilon]^{1/2}\big)$, it can be derived that the gradient complexity of SGLD is $\widetilde{O}(d^{15/4}\beta^{7/4}\rho^{-7/2}\epsilon^{-3/2})$. This strictly improves the stochastic gradient complexity (i.e., number of stochastic gradient evaluations to achieve ϵ-sampling error) of SGLD without Assumption 5.1 by a factor of $\widetilde{O}(d^{1/4}\beta^{1/4}\rho^{-1/2}\epsilon^{-1/2})$.

6 PROOF OUTLINE

In this section, we will sketch the proof of the main results (Theorem 4.5). The missing proofs for the other theorems,

corollaries and lemmas are deferred to the appendix. We first highlight the key proof technique and its novelty and difference compared with prior works. Then we will go over each of the key steps in detail.

6.1 PROOF TECHNIQUE AND NOVELTY

Proof Technique. Our proof relies on two sequences (green arrows in Figure 1): **Projected SGLD** ($\mathbf{x}_k^{\text{Proj-SGLD}}$) and **Metropolized SGLD** ($\mathbf{x}_k^{\text{MH}}$). Projected SGLD is constructed by adding an accept/reject step to the standard SGLD algorithm, which was first studied in Zhang et al. [2017]. Metropolized SGLD is a "virtual" sequence constructed by further adding a Metropolis Hasting step into Projected SGLD (the Metropolis Hasting step is computationally intractable so that Metropolized SGLD is not a practical algorithm and we only use it for theoretical analysis). Due to such Metropolis Hasting step, Metropolized SGLD is a time-reversible Markov chain and thus enjoys good conductance properties. Based on these two auxiliary sequences, we will prove the convergence of SGLD following three steps: (1) show that the output of Projected SGLD is close to that of SGLD in distribution (see Lemma 6.1); (2) show that the transition distribution of Projected SGLD is close to that of Metropolized SGLD (see Lemma 6.2); and (3) prove the convergence of Projected SGLD based on the conductance of Metropolized SGLD (see Lemma 6.4).

Technical Novelty. In order to prove the convergence rate of SGLD, prior works [Raginsky et al., 2017, Xu et al., 2018] typically make use of the LMC iterates $\mathbf{x}_k^{\text{LMC}}$ and decompose the sampling error of SGLD (the error between $\mathbf{x}_k$ and $\mathbf{x}^\pi$) into two parts: (1) the error between SGLD iterates and LMC iterates; and (2) the sampling error of LMC (though Raginsky et al. [2017], Xu et al. [2018] bound the sampling error of $\mathbf{x}_k^{\text{LMC}}$ in different ways). We illustrate the roadmap of different proof techniques in Figure 1. Note that their results on the error between $\mathbf{x}_k$ and $\mathbf{x}_k^{\text{LMC}}$ diverge as k increases, due to the uncertainty of stochastic gradients. This suggests that LMC may not be a good enough auxiliary chain for studying SGLD. In contrast, our constructed auxiliary sequences (i.e., Projected SGLD and Metropolized SGLD) are closer to SGLD since they also cover the randomness of stochastic gradients (this randomness can be included as part of the transition distribution, see Section 6.3 for more details). Therefore, our proof technique can lead to a sharper convergence analysis than those in Raginsky et al. [2017], Xu et al. [2018], which consequently gives a faster convergence rate of SGLD for sampling from non-log-concave distributions.

We would also like to point out that while the construction of Metropolized SGLD follows the same spirit of Zhang et al. [2017], it has a different goal and thus the corresponding analysis is not the same. Specifically, Zhang et al. [2017] only characterizes the hitting time of SGLD to a certain set by lower bounding the restricted conductance of SGLD, but does not prove its convergence to π. In contrast, we focus on the ability of SGLD for sampling from a certain target distribution. Thus we not only need to analyze the conductance of SGLD, but also need to bound the approximation error between the distribution of $\mathbf{x}_k$ and the target one (see Lemma 6.4 and B.3 and their proofs for more details), which is more challenging. As a consequence, we prove that the sampling error of SGLD to the target distribution can be upper bounded by $O(\sqrt{\eta})$, while the analysis in Zhang et al. [2017] can only give $O(1)$ sampling error.

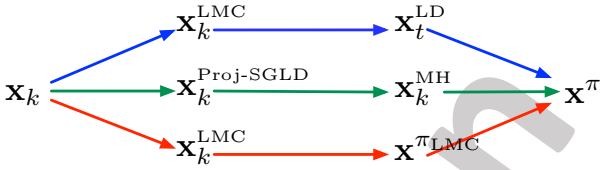

Figure 1: Illustration of the analysis framework of SGLD in different works: **Raginsky et al. (2017)**, **Xu et al. (2018)**, **this work**. The goal is to prove the convergence of SGLD iterates $\mathbf{x}_k$ to the point following the target distribution $\mathbf{x}^\pi$. Note that, $\mathbf{x}_k^{\text{LMC}}$, $\mathbf{x}_k^{\text{Proj-SGLD}}$ and $\mathbf{x}_k^{\text{MH}}$ denote the k-th iterates of LMC, Proj-SGLD, and Metropolized SGLD respectively; $\mathbf{x}_t^{\text{LD}}$ denotes the solution of (1.1) at time t; $\mathbf{x}^{\pi_{\text{LMC}}}$ denotes the point following the stationary distribution of LMC.

6.2 PROJECTED SGLD AND ITS EQUIVALENCE TO SGLD

Projected SGLD is constructed by adding an extra step in Algorithm 2 with the following accept/reject rule:

$$\mathbf{x}_{k+1} = \begin{cases} \mathbf{x}_{k+1} & \mathbf{x}_{k+1} \in \mathcal{B}(\mathbf{x}_k, r) \cap \mathcal{B}(\mathbf{0}, R); \\ \mathbf{x}_k & \text{otherwise.} \end{cases} \tag{6.1}$$

This step ensures each new iterate $\mathbf{x}_{k+1}$ does not go too far away from the current iterate and all iterates are restricted in a (relatively) large region $\mathcal{B}(\mathbf{0}, R)$. The entire algorithm is summarized in Algorithm 2. Due to the above accept/reject rule, Projected SGLD is slightly different from the standard SGLD algorithm (see Algorithm 1). However, we can show that Projected SGLD is nearly the same as SGLD given proper choices of R and r. In particular, in the following lemma, we will show that the total variance distance between the distributions of the outputs of both algorithms can be arbitrarily small.

Lemma 6.1. Let μ_K^{SGLD} and $\mu_K^{\text{Proj-SGLD}}$ be the distributions of the outputs of the standard SGLD algorithm and the projected SGLD algorithm. For any $\epsilon \in (0, 1)$, set

$$R = \bar{R}(\epsilon K^{-1}/4), \quad r = \sqrt{2\eta d/\beta}\big(2 + \sqrt{2\log(8K/\epsilon)/d}\big).$$

Algorithm 2 Projected SGLD

input: step size η; mini-batch size B; inverse temperature parameter β; radius R, r;
Randomly draw $\mathbf{x}_0$ from initial distribution μ_0.
for $k = 0, 1, \ldots, K$ **do**
  Randomly pick a subset $\mathcal{I}$ from $\{1, \ldots, n\}$ of size $|\mathcal{I}| = B$; randomly draw $\boldsymbol{\epsilon}_k \sim N(\mathbf{0}, \mathbf{I})$
  Compute the stochastic gradient $\mathbf{g}(\mathbf{x}_k, \mathcal{I}) = 1/B \sum_{i \in \mathcal{I}} \nabla f_i(\mathbf{x}_k)$
  Update: $\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \mathbf{g}(\mathbf{x}_k, \mathcal{I}) + \sqrt{2\eta/\beta}\boldsymbol{\epsilon}_k$
  if $\mathbf{x}_{k+1} \notin \mathcal{B}(\mathbf{x}_k, r) \cap \mathcal{B}(\mathbf{0}, R)$ **then**
    $\mathbf{x}_{k+1} = \mathbf{x}_k$
  end if
end for
output: $\mathbf{x}_K$

Suppose $\mathbb{P}(\|\mathbf{x}_0\|_2 \leq R/2) \leq \epsilon/16$ and setting $\eta \leq (LR + G)^{-2}\beta^{-1}d$, then we have

$$\big\|\mu_K^{\text{SGLD}} - \mu_K^{\text{Proj-SGLD}}\big\|_{TV} \leq \frac{\epsilon}{4}.$$

6.3 CONSTRUCTION OF METROPOLIZED SGLD

Projected SGLD will approximately generate samples from the following truncated target distribution since it restricts all iterates to the region $\Omega := \mathcal{B}(0, R)$,

$$\pi^\star(\mathrm{d}\mathbf{x}) = \begin{cases} \frac{e^{-\beta f(\mathbf{x})}}{\int_\Omega e^{-\beta f(\mathbf{y})}\mathrm{d}\mathbf{y}}\mathrm{d}\mathbf{x} & \mathbf{x} \in \Omega; \\ 0 & \text{otherwise.} \end{cases} \tag{6.2}$$

Then we will characterize the convergence of Projected SGLD to π^*. In particular, we will introduce an useful auxiliary Markov chain called Metropolized SGLD, i.e., SGLD with a Metropolis-Hasting step. We will first give the transition distribution of the Markov chain corresponding to Projected SGLD.

Transition distribution of Projected SGLD. Let $\mathbf{g}\big(\mathbf{x}, \mathcal{I}\big)$ be the stochastic gradient computed at the point $\mathbf{x}$, where $\mathcal{I}$ denotes the mini-batch of data points queried in the stochastic gradient computation. Then it is clear that Algorithm 2 can be described as a Markov process. More specifically, let $\mathbf{u}$ and $\mathbf{w}$ be the starting point and the point obtained after one-step iteration of Algorithm 2, the Markov chain in this iteration can be formed as $\mathbf{u} \to \mathbf{v} \to \mathbf{w}$, where $\mathbf{v}$ is generated based on the following conditional probability density function,

$$\begin{aligned} &P(\mathbf{v}|\mathbf{u}) = \mathbb{E}_{\mathcal{I}}[P(\mathbf{v}|\mathbf{u}, \mathcal{I})] \\ &= \mathbb{E}_{\mathcal{I}}\bigg[\frac{1}{(4\pi\eta/\beta)^{d/2}} \exp\bigg(-\frac{\|\mathbf{v} - \mathbf{u} + \eta\mathbf{g}(\mathbf{u}, \mathcal{I})\|_2^2}{4\eta/\beta}\bigg)\bigg|\mathbf{u}\bigg], \end{aligned} \tag{6.3}$$

which is exactly the transition probability of standard SGLD (i.e., without any accept/reject step). Let $R > 0$ be a tunable radius and recall that $\Omega = \mathcal{B}(\mathbf{0}, R)$, the process $\mathbf{v} \to \mathbf{w}$ can be formulated as

$$\mathbf{w} = \begin{cases} \mathbf{v} & \mathbf{v} \in \mathcal{B}(\mathbf{u}, r) \cap \Omega; \\ \mathbf{u} & \text{otherwise.} \end{cases} \tag{6.4}$$

Let $p(\mathbf{u}) = \mathbb{P}_{\mathbf{v} \sim P(\cdot|\mathbf{u})}[\mathbf{v} \in \mathcal{B}(\mathbf{u}, r) \cap \Omega]$ be the acceptance probability in (6.4), and $Q(\mathbf{w}|\mathbf{u})$ be the conditional PDF that describes $\mathbf{u} \to \mathbf{w}$, we have

$$\begin{aligned} Q(\mathbf{w}|\mathbf{u}) = (1 - p(\mathbf{u}))\delta_{\mathbf{u}}(\mathbf{w}) \\ + P(\mathbf{w}|\mathbf{u}) \cdot \mathbb{1}\big[\mathbf{w} \in \mathcal{B}(\mathbf{u}, r) \cap \Omega\big], \end{aligned}$$

where $P(\mathbf{w}|\mathbf{u})$ is computed by replacing $\mathbf{v}$ with $\mathbf{w}$ in (6.3). Similar to Zhang et al. [2017], Dwivedi et al. [2018], we consider the $1/2$-lazy version of the above Markov process, i.e., a Markov process with the following transition distribution

$$\mathcal{T}_{\mathbf{u}}(\mathbf{w}) = \frac{1}{2}\delta_{\mathbf{u}}(\mathbf{w}) + \frac{1}{2}Q(\mathbf{w}|\mathbf{u}), \tag{6.5}$$

where $\delta_{\mathbf{u}}(\cdot)$ is the Dirac-delta distribution at $\mathbf{u}$. However, it is difficult to directly prove the ergodicity of the Markov process with transition distribution $\mathcal{T}_{\mathbf{u}}(\mathbf{w})$, and it is also hard to tell whether its stationary distribution exists or not. Besides, SGLD is known to be asymptotically biased [Teh et al., 2016, Vollmer et al., 2016], which does not converge to the target distribution π even when it runs for infinite steps. It remains unclear whether Projected SGLD can converge to the target distribution given the formula of its transition distribution.

Metropolized SGLD. In order to quantify the sampling error for the output of Projected SGLD in Algorithm 2 and prove its convergence, we follow the idea of Zhang et al. [2017], which constructs an auxiliary Markov process by adding an extra Metropolis-Hasting correction step into Algorithm 2. We call it Metropolized SGLD. Given the starting point $\mathbf{u}$, let $\mathbf{w}$ be the candidate state generated from the distribution $\mathcal{T}_{\mathbf{u}}(\cdot)$. Metropolized SGLD will accept the candidate $\mathbf{w}$ with the following probability,

$$\alpha_{\mathbf{u}}(\mathbf{w}) = \min\bigg\{1, \frac{\mathcal{T}_{\mathbf{w}}(\mathbf{u})}{\mathcal{T}_{\mathbf{u}}(\mathbf{w})} \cdot \exp\big[-\beta\big(f(\mathbf{w}) - f(\mathbf{u})\big)\big]\bigg\}.$$

Let $\mathcal{T}_{\mathbf{u}}^\star(\cdot)$ denote the transition distribution of such auxiliary Markov process, i.e.,

$$\mathcal{T}_{\mathbf{u}}^\star(\mathbf{w}) = (1 - \alpha_{\mathbf{u}}(\mathbf{w}))\delta(\mathbf{u}) + \alpha_{\mathbf{u}}(\mathbf{w})\mathcal{T}_{\mathbf{u}}(\mathbf{w}),$$

which is time-reversible and easy to verify. Due to this Metropolis-Hastings correction step, the Markov chain can converge to a unique stationary distribution $\pi^\star \propto e^{-\beta f(\mathbf{x})} \cdot \mathbb{1}(\mathbf{x} \in \Omega)$ [Zhang et al., 2017]. It is worth pointing out that Metropolized SGLD cannot be implemented in practice since we are only allowed to query a subset of the training data in each iteration of SGLD, thus we are not be able to exactly calculate the accept probability $\alpha_{\mathbf{u}}(\mathbf{w})$, which

involves the expectation computation over the stochastic mini-batch of data points. Nevertheless, we will only use this auxiliary Markov chain in our theoretical analysis to show the convergence of Algorithm 2.

We will further show that the transition distribution of Projected SGLD ($\mathcal{T}_{\mathbf{u}}(\cdot)$) can be δ-close to that of Metropolized SGLD ($\mathcal{T}^*_{\mathbf{u}}(\cdot)$) for some small quantity δ governed by η, which is provided in the following lemma.

Lemma 6.2. Under Assumption 4.4, let $G = \max_{i\in[n]} \|\nabla f_i(\mathbf{0})\|_2$, and set $r = \sqrt{10\eta d/\beta}\big(1 + \sqrt{\log(8K/\epsilon)/d}\big)$, where K is the total number of iterations of Projected SGLD. Then there exists a constant

$$\delta = \Big[10Ld\eta + 10L(LR+G)d^{1/2}\beta^{1/2}\eta^{3/2} + 12\beta(LR+G)^2 d\eta/B + 2\beta^2(LR+G)^4\eta^2/B\Big] \cdot \big(1+\sqrt{\log(8K/\epsilon)/d}\big)^2$$

such that for any set $\mathcal{A} \subseteq \Omega$ and any point $\mathbf{u} \in \Omega$,

$$(1-\delta)\mathcal{T}^\star_{\mathbf{u}}(\mathcal{A}) \le \mathcal{T}_{\mathbf{u}}(\mathcal{A}) \le (1+\delta)\mathcal{T}^\star_{\mathbf{u}}(\mathcal{A}). \tag{6.6}$$

6.4 CONVERGENCE OF PROJECTED SGLD

In this part, we will characterize the convergence of Projected SGLD, which consists of two steps: (1) given the δ-closeness result in Lemma 6.2, we prove that Projected SGLD can converge to the truncated target distribution $\pi^\star$ up to some approximation error determined by δ; and (2) we prove that with a proper choice of the truncation radius R, the total variation distance between $\pi^\star$ and the target distribution π can be sufficiently small.

Convergence of Projected SGLD to $\pi^\star$. We first provide the definition of the conductance for a time-reversible Markov chain as follows.

Definition 6.3 (Conductance)**.** The conductance of a time-reversible Markov chain with transition distribution $\mathcal{T}^\star_{\mathbf{u}}(\cdot)$ and stationary distribution $\pi^\star$ is defined by,

$$\phi := \inf_{\mathcal{A}:\mathcal{A}\subseteq\Omega,\pi^\star(\mathcal{A})\in(0,1)} \frac{\int_{\mathcal{A}} \mathcal{T}^\star_{\mathbf{u}}(\Omega\backslash\mathcal{A})\pi^\star(\mathrm{d}\mathbf{u})}{\min\{\pi^\star(\mathcal{A}), \pi^\star(\Omega\backslash\mathcal{A})\}},$$

where Ω is the support of the state of the Markov chain.

In Lemma 6.2, we have already shown that the transition distribution of Algorithm 2, i.e., $\mathcal{T}_{\mathbf{u}}(\cdot)$ is δ-close to that of Metropolized SGLD, i.e., $\mathcal{T}^\star_{\mathbf{u}}(\cdot)$, for some small quantity δ. Besides, from Lovász and Simonovits [1993], Vempala [2007], we know that a time-reversible Markov chain can converge to its stationary distribution at a linear rate depending on its conductance. Therefore, we aim to characterize the convergence rate of $\mathcal{T}_{\mathbf{u}}(\cdot)$ based on the ergodicity of $\mathcal{T}^\star_{\mathbf{u}}(\cdot)$. We utilize the conductance parameter of $\mathcal{T}^\star_{\mathbf{u}}(\cdot)$, denoted by ϕ, and establish the convergence of $\mathcal{T}_{\mathbf{u}}(\cdot)$ in total variation distance in the following lemma.

Lemma 6.4. Let $\mu_K^{\text{Proj-SGLD}}$ be the distribution of the output of Algorithm 2. Under Assumption 4.4, if $\mathcal{T}_{\mathbf{u}}(\cdot)$ is δ-close to $\mathcal{T}^\star_{\mathbf{u}}(\cdot)$ with $\delta \le \min\{1-\sqrt{2}/2, \phi/16\}$, then for any λ-warm start initial distribution with respect to $\pi^\star$, it holds that

$$\|\mu_K^{\text{Proj-SGLD}} - \pi^\star\|_{TV} \le \lambda\big(1-\phi^2/8\big)^K + 16\delta/\phi.$$

Lemma 6.4 shows that Projected SGLD converges to $\pi^\star$ in total variance distance with approximation error up to $16\delta/\phi$. The next step is to characterize the conductance parameter ϕ and reveal its dependency on the problem dependent parameters, which we state in the following lemma.

Lemma 6.5. Under Assumptions 4.3 and 4.4, if the step size satisfies $\eta \le \big[35(Ld + (LR+G)^2\beta d/B)\big]^{-1} \wedge [25\beta(LR+G)^2]^{-1}$, there exists an absolute constant c_0 such that

$$\phi \ge c_0\rho\sqrt{\eta/\beta},$$

where ρ is the Cheeger constant of the distribution $\pi^\star$.

Bounding the difference between π and $\pi^\star$. Lemmas 6.4 and 6.5 together guarantee that Algorithm 2 converges to the truncated target distribution $\pi^\star$. Thus the last thing remaining to be done is ensuring that $\pi^\star$ is sufficiently close to π. The following lemma characterizes the total variation distance between the target distribution π and its truncated version π^* in $\mathcal{B}(\mathbf{0}, R)$.

Lemma 6.6. For any $\epsilon \in (0,1)$, set $R = \bar{R}(\epsilon/12)$ and let $\Omega = \mathcal{B}(\mathbf{0}, R)$ and $\pi^\star$ be the truncated target distribution in Ω. Then the total variation distance between $\pi^\star$ and π can be upper bounded by $\|\pi^\star - \pi\|_{TV} \le \epsilon/4$.

Proof of Theorem 4.5. The rest proof of Theorem 4.5 is straightforward by combining Lemmas 6.1, 6.4, and 6.6 using the triangle inequality. We defer the detailed proof to Appendix A. □

7 CONCLUSION

In this paper, we proved a faster convergence rate of SGLD for sampling from a broad class of distributions that can be non-log-concave. In particular, we developed a new proof technique for characterizing the convergence of SGLD. Different from the existing works that mainly study the convergence of SGLD based on full-gradient based Markov chain suchs as LMC or continuous Langevin dynamics, the key of our proof technique relies on two auxiliary Markov chains: Projected SGLD and Metropolized SGLD, which can better capture the behavior of SGLD since they also cover the randomness of the stochastic gradients. Our proof technique is of independent technical interest and can be potentially adapted to study the convergence of other stochastic gradient-based sampling algorithms.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments. DZ is supported by the Bloomberg Data Science Ph.D. Fellowship. QG is partially supported by the National Science Foundation CAREER Award 1906169. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Dominique Bakry, Franck Barthe, Patrick Cattiaux, Arnaud Guillin, et al. A simple proof of the poincaré inequality for a large class of probability measures. *Electronic Communications in Probability*, 13:60–66, 2008.

Nawaf Bou-Rabee and Martin Hairer. Nonasymptotic mixing of the mala algorithm. *IMA Journal of Numerical Analysis*, 33(1):80–110, 2013.

Nawaf Bou-Rabee, Andreas Eberle, and Raphael Zimmer. Coupling and convergence for Hamiltonian monte carlo. *arXiv preprint arXiv:1805.00452*, 2018.

Peter Buser. A note on the isoperimetric constant. *Annales scientifiques de l'École Normale Supérieure*, Ser. 4, 15 (2):213–230, 1982.

Ngoc Huy Chau, Éric Moulines, Miklos Rásonyi, Sotirios Sabanis, and Ying Zhang. On stochastic gradient Langevin dynamics with dependent data streams: the fully non-convex case. *arXiv preprint arXiv:1905.13142*, 2019.

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

Changyou Chen, Wenlin Wang, Yizhe Zhang, Qinliang Su, and Lawrence Carin. A convergence analysis for a class of practical variance-reduction stochastic gradient mcmc. *arXiv preprint arXiv:1709.01180*, 2017.

Yi Chen, Jinglin Chen, Jing Dong, Jian Peng, and Zhaoran Wang. Accelerating nonconvex learning via replica exchange Langevin diffusion. In *ICLR*, 2019a.

Yuansi Chen. An almost constant lower bound of the isoperimetric coefficient in the kls conjecture. *Geometric and Functional Analysis*, 31(1):34–61, 2021.

Yuansi Chen, Raaz Dwivedi, Martin J Wainwright, and Bin Yu. Fast mixing of metropolized hamiltonian monte carlo: Benefits of multi-step gradients. *arXiv preprint arXiv:1905.12247*, 2019b.

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

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

Ben Cousins and Santosh Vempala. A cubic algorithm for computing gaussian volume. In *SODA*, pages 1215–1228. SIAM, 2014.

Arnak S Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In *COLT*, pages 678–689, 2017a.

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

Arnak S Dalalyan and Avetik Karagulyan. User-friendly guarantees for the Langevin monte carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12):5278–5311, 2019.

Wei Deng, Qi Feng, Liyao Gao, Faming Liang, and Guang Lin. Non-convex learning via replica exchange stochastic gradient mcmc. In *ICML*, 2020.

Alain Durmus and Eric Moulines. Sampling from strongly log-concave distributions with the unadjusted Langevin algorithm. 2016.

Alain Durmus, Eric Moulines, and Eero Saksman. On the convergence of hamiltonian monte carlo. *arXiv preprint arXiv:1705.00166*, 2017a.

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

Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-hastings algorithms are fast! In *COLT*, pages 793–797, 2018.

Andreas Eberle. Reflection couplings and contraction rates for diffusions. *Probability theory and related fields*, 166 (3-4):851–886, 2016.

Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. *arXiv preprint arXiv:1703.01617*, 2017.

Murat A Erdogdu, Lester Mackey, and Ohad Shamir. Global non-convex optimization with discretized diffusions. In *NeurIPS*, pages 9671–9680, 2018.

Xuefeng Gao, Mert Gürbüzbalaban, and Lingjiong Zhu. Global convergence of stochastic gradient Hamiltonian monte carlo for non-convex stochastic optimization: Non-asymptotic performance bounds and momentum-based

acceleration. *arXiv preprint arXiv:1809.04618*, 2018.

Ravi Kannan, László Lovász, and Miklós Simonovits. Isoperimetric problems for convex bodies and a localization lemma. *Discrete & Computational Geometry*, 13 (3-4):541–559, 1995.

Paul Langevin. On the theory of brownian motion. *CR Acad. Sci. Paris*, 146:530–533, 1908.

Michel Ledoux. A simple analytic proof of an inequality by p. buser. *Proceedings of the American mathematical society*, 121(3):951–959, 1994.

Holden Lee, Andrej Risteski, and Rong Ge. Beyond log-concavity: Provable guarantees for sampling multi-modal distributions using simulated tempering Langevin monte carlo. In *NeurIPS*, pages 7857–7866, 2018.

Yin Tat Lee and Santosh Srinivas Vempala. Eldan's stochastic localization and the kls hyperplane conjecture: An improved lower bound for expansion. In *FOCS*, pages 998–1007. IEEE, 2017.

László Lovász and Miklós Simonovits. The mixing rate of markov chains, an isoperimetric inequality, and computing the volume. In *FOCS*, pages 346–354. IEEE, 1990.

László Lovász and Miklós Simonovits. Random walks in a convex body and an improved volume algorithm. *Random structures & algorithms*, 4:359–412, 1993.

Yi-An Ma, Yuansi Chen, Chi Jin, Nicolas Flammarion, and Michael I. Jordan. Sampling can be faster than optimization. *arXiv preprint arXiv:1811.08413*, 2018.

Oren Mangoubi and Nisheeth Vishnoi. Dimensionally tight bounds for second-order hamiltonian monte carlo. In *NeurIPS*, pages 6027–6037, 2018.

Oren Mangoubi and Nisheeth K Vishnoi. Nonconvex sampling with the metropolis-adjusted Langevin algorithm. In *COLT*, pages 2259–2293, 2019.

Kerrie L Mengersen, Richard L Tweedie, et al. Rates of convergence of the hastings and metropolis algorithms. *The annals of Statistics*, 24(1):101–121, 1996.

Wenlong Mou, Nicolas Flammarion, Martin J Wainwright, and Peter L Bartlett. Improved bounds for discretization of Langevin diffusions: Near-optimal rates without convexity. *arXiv preprint arXiv:1907.11331*, 2019.

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

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

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

Gareth O Roberts and Richard L Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.

Issei Sato and Hiroshi Nakagawa. Approximation analysis of stochastic gradient Langevin dynamics by using fokker-planck equation and ito process. In *ICML*, pages 982–990, 2014.

Robert L Smith. Efficient monte carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research*, 32(6):1296–1308, 1984.

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

Santosh Vempala. Geometric random walks: a survey. In *Combinatorial and Computational Geometry*, pages 577–616. Cambridge University Press, 2007.

Santosh S Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Log-sobolev suffices. *arXiv preprint arXiv:1903.08568*, 2019.

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

Bao Wang, Difan Zou, Quanquan Gu, and Stanley Osher. Laplacian smoothing stochastic gradient markov chain monte carlo. *arXiv preprint arXiv:1911.00782*, 2019.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *ICML*, pages 681–688, 2011.

Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. In *NeurIPS*, pages 3126–3137, 2018.

Ying Zhang, Ömer Deniz Akyildiz, Theo Damoulas, and Sotirios Sabanis. Nonasymptotic estimates for stochastic gradient Langevin dynamics under local conditions in nonconvex optimization. *arXiv preprint arXiv:1910.02008*, 2019.

Yuchen Zhang, Percy Liang, and Moses Charikar. A hitting time analysis of stochastic gradient Langevin dynamics. In *COLT*, pages 1980–2022, 2017.

Difan Zou, Pan Xu, and Quanquan Gu. Stochastic variance-reduced Hamilton Monte Carlo methods. In *ICML*, pages 6028–6037, 2018.

Difan Zou, Pan Xu, and Quanquan Gu. Sampling from

non-log-concave distributions via variance-reduced gradient Langevin dynamics. In *Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2936–2945. PMLR, 2019a.

Difan Zou, Pan Xu, and Quanquan Gu. Stochastic gradient hamiltonian monte carlo methods with recursive variance reduction. In *NeurIPS*, pages 3830–3841, 2019b.