Sampling from Non-Log-Concave Distributions via Stochastic Variance-Reduced Gradient Langevin Dynamics

Difan Zou
Department of Computer Science
University of California,
Los Angeles

Pan Xu
Department of Computer Science
University of California,
Los Angeles

Quanquan Gu
Department of Computer Science
University of California,
Los Angeles

Abstract

We study stochastic variance reduction-based Langevin dynamic algorithms, SVRG-LD and SAGA-LD (Dubey et al., 2016), for sampling from non-log-concave distributions. Under certain assumptions on the log density function, we establish the convergence guarantees of SVRG-LD and SAGA-LD in 2-Wasserstein distance. More specifically, we show that both SVRG-LD and SAGA-LD require $\widetilde{O}\big(n+n^{3/4}/\epsilon^2+n^{1/2}/\epsilon^4\big)\cdot\exp\big(\widetilde{O}(d+\gamma)\big)$ stochastic gradient evaluations to achieve ϵ-accuracy in 2-Wasserstein distance, which outperforms the $\widetilde{O}\big(n/\epsilon^4\big)\cdot\exp\big(\widetilde{O}(d+\gamma)\big)$ gradient complexity achieved by Langevin Monte Carlo Method (Raginsky et al., 2017). Experiments on synthetic data and real data back up our theory.

1 INTRODUCTION

In the past decade, there has been an increasing interest in applying gradient based Markov Chain Monte Carlo (MCMC) methods for sampling from posterior distributions in Bayesian machine learning (Neal et al., 2011; Welling and Teh, 2011; Ahn et al., 2012; Chen et al., 2014; Ma et al., 2015; Cheng et al., 2018). In detail, this class of MCMC methods is based on the Langevin dynamics, which is described by the following stochastic differential equation (SDE)

$$\mathrm{d}\boldsymbol{X}(t) = -\nabla F(\boldsymbol{X}(t))\mathrm{d}t + \sqrt{2/\gamma}\mathrm{d}\boldsymbol{B}(t), \tag{1.1}$$

where $\gamma > 0$ is the inverse temperature parameter and $\{\boldsymbol{B}(t)\}_{t\geq 0}$ is the standard Brownian motion in $\mathbb{R}^d$. Under certain assumptions on the drift term $\nabla F(\mathbf{x})$, the distribution of $\boldsymbol{X}(t)$ can be described by Fokker-Planck equation, and is able to converge to an invariant stationary distribution $\pi \propto \exp\big(-\gamma F(\mathbf{x})\big)$ (Chiang et al., 1987). In Bayesian inference, one aims to sample the target distribution with the form $\pi \propto \exp\big(-\gamma F(\mathbf{x})\big)$, and a typical way is to apply Euler-Maruyama discretization (Kloeden and Platen, 1992) to (1.1), which gives rise to the celebrated Langevin Monte Carlo (LMC) method (Roberts and Tweedie, 1996a),

$$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k - \nabla F(\boldsymbol{X}_k)\eta + \sqrt{2\eta/\gamma}\boldsymbol{\epsilon}_k, \tag{1.2}$$

where $\boldsymbol{\epsilon}_k$ follows a standard multivariate normal distribution, and $\eta > 0$ denotes the step size. When the target distribution is strongly log-concave, i.e., function $F(\mathbf{x})$ is strongly convex, the convergence property of LMC has been widely studied based on total variation (Durmus and Moulines, 2015, 2016; Dalalyan, 2017b) and 2-Wasserstein (Dalalyan, 2017a; Dalalyan and Karagulyan, 2017) distances. On the other hand, for many machine learning problems involving extremely large amount of data, the function $F(\mathbf{x})$ on the drift term of (1.1) can be written as an average of n component functions

$$F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}),$$

where $f_i(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$ is the negative log likelihood function on the i-th example. When the data sample size n is enormous, the computation of the full gradient $\nabla F(\boldsymbol{X})$ in LMC is often very expensive. To overcome this computational burden, one resorts to using stochastic gradient to approximate the drift term in (1.1), which gives rise to the celebrated stochastic gradient Langevin dynamics (SGLD) method (Welling and Teh, 2011). In practice, the SGLD algorithm has achieved great success in Bayesian learning (Welling and Teh, 2011; Ahn et al., 2012) and Bayesian deep learning (Chaudhari et al., 2016; Ye et al., 2017). However, the SGLD algorithm requires more iteration steps

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

to achieve a high sampling precision compared with LMC due to the large variance of stochastic gradients. In order to alleviate this issue as well as save the gradient computation, Dubey et al. (2016) incorporated the idea of variance reduction (Johnson and Zhang, 2013; Reddi et al., 2016) into SGLD, and proposed two types of stochastic variance reduced algorithms based on gradient Langevin dynamics, namely SVRG-LD and SAGA-LD. Recently, Chatterji et al. (2018) proved the convergence rate of SVRG-LD and SAGA-LD in 2-Wasserstein distance when the target distribution is strongly log-concave, which characterizes the feasible regime where SVRG-LD and SAGA-LD outperform LMC and SGLD. The convergence rate of SVRG-LD was further improved by Zou et al. (2018b) recently. However, the current convergence analyses (Chatterji et al., 2018; Zou et al., 2018b) of stochastic variance-reduced gradient Langevin dynamics are mostly restricted to the strongly log-concave distributions, except Dubey et al. (2016); Chen et al. (2017). Nevertheless, Dubey et al. (2016); Chen et al. (2017) only investigated the mean square error of the sample path average. It is of more interest to establish the nonasymptotic convergence guarantee in terms of certain distance between the target distribution and that of the current iterate, which provides a fine-grained characterization of the sampling algorithms.

In this paper, we provide convergence analyses of SVRG-LD and SAGA-LD in 2-Wasserstein distance for non-log-concave target distributions. Different from the analysis of sampling from strongly log-concave distributions, the contraction property of 2-Wasserstein distance along the Langevin diffusion (1.1) no longer holds, which poses a great challenge for our analysis and makes existing proof techniques (Chatterji et al., 2018) for strongly log-concave distribution not applicable to our case. To address this challenge, we provide a new proof technique by extending the idea of Raginsky et al. (2017) for analyzing SGLD in nonconvex optimization. More specifically, our proof technique is based on a coupled Brownian motion between the discrete-time Markov chain and a continuous-time Markov chain generated by (1.1) and decomposes the 2-Wasserstein distance between the target distribution and that of the current iterate into two parts: the 2-Wasserstein distance between distributions of the current iterate and the corresponding continuous-time Markov Chain, and the distance between the distribution of the position in the coupled Markov chain and its stationary distribution, i.e., the target distribution π.

Our Contributions The major contributions of this paper are highlighted as follows.

- We study the SVRG-LD and SAGA-LD methods for sampling from non-log-concave distributions and prove their nonasymptotic convergence to the target distribution in terms of 2-Wasserstein distance. Specifically, we show that both SVRG-LD and SAGA-LD require $\widetilde{O}\big(n + n^{3/4}/\epsilon^2 + n^{1/2}/\epsilon^4\big) \cdot \exp\big(\widetilde{O}(d+\gamma)\big)$ stochastic gradient evaluations to achieve ϵ-accuracy, where n is the number of samples, γ is the inverse temperature and d is the problem dimension, which outperforms the gradient complexities of LMC and SGLD.
- We conduct experiments on both synthetic and real-world data to compare different first-order Langevin methods (SVRG-LD, SAGA-LD, SGLD, LMC) for sampling from non-log-concave distributions. The comparison suggests that the SVRG-LD and SAGA-LD have similar performance, and attain faster mixing time and perform better than their counterparts even when the target distribution is non-log-concave.

Notation We denote a deterministic vector by lower case bold symbol $\mathbf{x}$ and a random vector by upper case italicized bold symbol $\boldsymbol{X}$. We also use $\boldsymbol{X}_k$ (with subscript k) to denote the iterate of a discrete-time algorithm and $\boldsymbol{X}(t)$ (with index t in a parenthesis) to denote the continuous-time random process. For a vector $\mathbf{x} \in \mathbb{R}^d$, we denote by $\|\mathbf{x}\|_2$ the Euclidean norm. For a matrix $\mathbf{X}$, we denote $\|\mathbf{X}\|_F$ as the Frobenius norm.For a random vector $\boldsymbol{X} \in \mathbb{R}^d$, we denote its probability distribution function by $P(\boldsymbol{X})$. We denote by $\mathbb{E}_u(\boldsymbol{X})$ the expectation of $\boldsymbol{X}$ under probability measure u. We denote the 2-Wasserstein distance between two probability measures u and v as

$$\mathcal{W}_2^2(u,v) = \inf_{\zeta \in \Gamma(u,v)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\boldsymbol{X}_u - \boldsymbol{X}_v\|_2^2 \mathrm{d}\zeta(\boldsymbol{X}_u, \boldsymbol{X}_v),$$

where the infimum is over all joint distributions ζ with u and v being its marginal distributions. We denote by $\mathrm{KL}(p_1||p_2)$ the KL-divergence between probability measures p_1 and p_2. We use $a_n = O(b_n)$ to denote that $a_n \leq C b_n$ for some universal constant $C > 0$, and use $a_n = \widetilde{O}(b_n)$ to hide some logarithmic terms of b_n. We also use $a \wedge b$ to denote $\min\{a, b\}$.

2 RELATED WORK

In this section, we review the literature on generic Langevin dynamics based algorithms.

Langevin Monte Carlo (LMC) (1.2) have been widely used for approximate sampling. Dalalyan (2017b) proved that the distribution of the last iterate in LMC converges to the stationary distribution within $O(d/\epsilon^2)$ iterations in variation distance. Durmus and

Moulines (2015) improved the results by showing the same result holds for any starting point and in Wasserstein distance. Recently Dalalyan (2017a) improved the existing results in terms of the 2-Wasserstein distance and provided further insights on the close relation between approximate sampling and optimization. Bubeck et al. (2015) analyzed sampling from log-concave distributions with compact support via projected LMC. Brosse et al. (2017) proposed a proximal LMC algorithm. The Euler discretization on SDEs introduces a bias, and might fail to converge to the target distribution (Roberts and Tweedie, 1996a,b). An effective way to address this issue is incorporating the metropolis hasting correction step (Hastings, 1970) into LMC, which gives rise to metropolis adjusted Langevin algorithm (MALA) (Roberts and Rosenthal, 1998). Following this line of research, Bou-Rabee and Hairer (2012) provided nonasymptotic bounds on the mixing time of MALA, but the explicit dependence on the dimension d and target accuracy remains implicit. Eberle et al. (2014) established a clearer mixing time bound of MALA in terms of a modified Wasserstein distance for log-concave densities. Dwivedi et al. (2018) investigated MALA for strongly log-concave densities, and proved a linear rate of convergence in total variation distance.

Due to the increasing amount of data in modern machine learning problems, stochastic gradient Langevin dynamics (SGLD) (Welling and Teh, 2011; Ahn et al., 2012; Ma et al., 2015) has received extensive attentions. Vollmer et al. (2016) analyzed the nonasymptotic bias and variance of SGLD using Poisson equations. Dalalyan and Karagulyan (2017) proved $\widetilde{O}(d\sigma^2/\epsilon^2)$ convergence rate for SGLD in 2-Wasserstein distance when the target distribution is strongly log-concave. Moreover, Neal et al. (2011) introduced fictitious momentum term in Hamilton dynamics, which gives rise to Hamiltonian Monte Carlo (HMC). Similar to SGLD, stochastic gradient Hamiltonian Monte Carlo (SGHMC) (Chen et al., 2014) was proposed to overcome the limitation of gradient evaluation on large datasets, and demonstrated better performance in learning Bayesian neural networks and online Bayesian matrix factorization (Chen et al., 2014). Under a similar framework, Chen et al. (2014) studied the stochastic MCMC method with higher-order integrator in terms of the MSE of the average sample path. Cheng et al. (2018) proposed the underdamped MCMC method and proved its convergence guarantee in 2-Wasserstein distance for strongly log-concave distribution. Despite the great success of SGLD and SGHMC, the large variance of stochastic gradient may lead to unavoidable bias due to the lack of metropolis hasting (MH) correction. To overcome this, Teh et al. (2016) proposed to decrease the step size to alleviate the bias and proved the asymptotic rate of SGLD in terms of MSE. Betancourt (2015) pointed out that SGHMC may also lead to poor sampling performance, and there exists a tradeoff between the step size and acceptance probability in MH correction. This issue has been addressed by Dang et al. (2017) where they proposed a modified HMC algorithm that uses a subset of data to estimate both the dynamics and the subsequent MH acceptance probability.

Another way to alleviate the variance of stochastic gradient and save gradient computation is applying variance-reduction technique. Dubey et al. (2016) proposed a variance-reduced stochastic gradient Langevin dynamics for Bayesian posterior inference, and proved that it improves the mean square error upon SGLD. Baker et al. (2017) applied zero variance control variates to stochastic MCMC method, and showed that it is able to reduce the computational cost of SGMCMC to $O(1)$. Chatterji et al. (2018) studied two variants of variance-reduced stochastic Langevin dynamics proposed in Dubey et al. (2016), and proved their convergence guarantees for strongly log-concave distributions. Moreover, by replacing the full gradient in the outer loop of SVRG-LD with a subsampled one, Chen et al. (2017) and Zou et al. (2018b) studied the convergence rate of subsampled SVRG-LD method in MSE and 2-Wasserstein distance respectively. The variance-reduced HMC has also been investigated recently in Zou et al. (2018a); Li et al. (2018).

There is also a line of work that utilize Langevin dynamics to design better algorithms for nonconvex optimization. In particular, Raginsky et al. (2017); Zhang et al. (2017) studied the nonasymptotic convergence of SGLD to global and local minimum of nonconvex functions. Xu et al. (2018) studied the global convergence of a family of Langevin dynamics based algorithm. Chen et al. (2019) studied the algorithm that swaps between two Langevin diffusions with different temperatures.

In Table 1, we summarize the gradient complexity[1] of LMC, SGLD, SVRG-LD and SAGA-LD in 2-Wasserstein distance for sampling from strongly log-concave and non-log-concave densities. To the best of our knowledge, there is no convergence result in 2-Wasserstein distance for sampling from general log-concave densities using Langevin dynamics based algorithms. It should be noted that for sampling from a non-log-concave the dependence on dimension d is inevitably exponential. In fact, it is proved in Bovier et al. (2004) that the lower bound of metastable exit time of SDE is exponential in d when the nonconvex function F in (1.2) has multiple local minima.

[1]Gradient complexity is defined as the number of

Table 1: Gradient complexities to converge to the stationary distribution in 2-Wasserstein distance. Note that Raginsky et al. (2017) shows that SGLD dose not converge in 2-Wasserstein distance for non-log-concave densities.

	Strongly log-concave[2]	Non-log-concave
LMC	$\widetilde{O}\Big(\frac{nd^{1/2}}{\epsilon}\Big)$(Dalalyan, 2017a)	$\widetilde{O}\big(\frac{n}{\epsilon^4}\big)\cdot e^{\widetilde{O}(d)}$(Raginsky et al., 2017)
SGLD	$\widetilde{O}\Big(\frac{d}{\epsilon^2}\Big)$(Dalalyan, 2017a)	–
SVRG-LD	$\widetilde{O}\Big(n+\frac{n^{1/2}d^{1/2}}{\epsilon}\Big)$(Zou et al., 2018b)	$\widetilde{O}\Big(n+\frac{n^{3/4}}{\epsilon^2}+\frac{n^{1/2}}{\epsilon^4}\Big)\cdot e^{\widetilde{O}(d)}$**(This paper)**
SAGA-LD	$\widetilde{O}\Big(n+\frac{n^{1/2}d^{1/2}}{\epsilon}\Big)$ (Chatterji et al., 2018)	$\widetilde{O}\Big(n+\frac{n^{3/4}}{\epsilon^2}+\frac{n^{1/2}}{\epsilon^4}\Big)\cdot e^{\widetilde{O}(d)}$**(This paper)**

Algorithm 1 Stochastic Variance-Reduced Gradient Langevin Dynamics (SVRG-LD)

1: **input:** step size $\eta>0$; batch size B; epoch length m; inverse temperature parameter $\gamma>0$
2: **initialization:** $\boldsymbol{X}_0=\mathbf{0}$, $\widetilde{\boldsymbol{X}}^{(0)}=\boldsymbol{X}_0$
3: **for** $s=0,1,\ldots,(K/m)$ **do**
4: $\quad\widetilde{\boldsymbol{G}}=\nabla F(\widetilde{\boldsymbol{X}}^{(s)})$
5: $\quad$**for** $\ell=0,\ldots,m-1$ **do**
6: $\quad\quad k=sm+\ell$
7: $\quad\quad$randomly pick a subset I_k from $\{1,\ldots,n\}$ of size $|I_k|=B$; randomly draw $\boldsymbol{\epsilon}_k\sim N(\mathbf{0},\mathbf{I}_{d\times d})$
8: $\quad\quad\widetilde{\nabla}_k=\frac{1}{B}\sum_{i_k\in I_k}\big(\nabla f_{i_k}(\boldsymbol{X}_k)-\nabla f_{i_k}(\widetilde{\boldsymbol{X}}^{(s)})+\widetilde{\boldsymbol{G}}\big)$
9: $\quad\quad\boldsymbol{X}_{k+1}=\boldsymbol{X}_k-\eta\widetilde{\nabla}_k+\sqrt{2\eta/\gamma}\boldsymbol{\epsilon}_k$
10: $\quad$**end for**
11: $\quad\widetilde{\boldsymbol{X}}^{(s+1)}=\boldsymbol{X}_{(s+1)m}$
12: **end for**

Algorithm 2 Stochastic Average Gradient Langevin Dynamics (SAGA-LD)

1: **input:** step size $\eta>0$; batch size B; epoch length m; inverse temperature parameter $\gamma>0$
2: **initialization:** $\boldsymbol{X}_0=\mathbf{0}$, $\widetilde{\boldsymbol{G}}=[\nabla f_1(\boldsymbol{X}_0),\ldots,\nabla f_n(\boldsymbol{X}_0)]$
3: **for** $k=0,1,\ldots,K$ **do**
4: $\quad\widetilde{\boldsymbol{g}}_k=n^{-1}\sum_{i=1}^n\widetilde{\boldsymbol{G}}_i$, where $\widetilde{\boldsymbol{G}}_i$ denotes the i-th column of Matrix $\widetilde{\boldsymbol{G}}$
5: $\quad$randomly pick a subset I_k from $\{1,\ldots,n\}$ of size $|I_k|=B$; randomly draw $\boldsymbol{\epsilon}_k\sim N(\mathbf{0},\mathbf{I}_{d\times d})$
6: $\quad\widetilde{\nabla}_k=\frac{1}{B}\sum_{i_k\in I_k}\big(\nabla f_{i_k}(\boldsymbol{X}_k)-\widetilde{\boldsymbol{G}}_{i_k}+\widetilde{\boldsymbol{g}}_k\big)$
7: $\quad\boldsymbol{X}_{k+1}=\boldsymbol{X}_k-\eta\widetilde{\nabla}_k+\sqrt{2\eta/\gamma}\boldsymbol{\epsilon}_k$
8: $\quad\widetilde{\boldsymbol{G}}_{i_k}=\nabla f_{i_k}(\boldsymbol{X}_k)$ for $i_k\in I_k$
9: **end for**

3 REVIEW OF SVRG-LD AND SAGA-LD

In this section, we review the SVRG-LD and SAGA-LD algorithms, which incorporates the variance reduction technique into the Langevin based algorithm.

Algorithm 1 displays the detail of SVRG-LD, which consists of multiple epochs. In the beginning of the s-th epoch, we compute the full gradient of $F(\widetilde{\boldsymbol{X}}^{(s)})$ by scanning all samples

$$\widetilde{G}=\nabla F(\widetilde{\boldsymbol{X}}^{(s)})=\frac{1}{n}\sum_{i=1}^n\nabla f_i(\widetilde{\boldsymbol{X}}^{(s)}).$$

Regarding the l-th inner iteration in the s-th epoch (the k-th update in the total iteration sequence), the semi-stochastic gradient $\widetilde{\nabla}_k$ is computed based on the snapshot gradient $\widetilde{G}$ and a new minibatch of samples I_k, which yields

$$\widetilde{\nabla}_k=\frac{1}{B}\sum_{i_k\in I_k}\big(\nabla f_{i_k}(\boldsymbol{X}_k)-\nabla f_{i_k}(\widetilde{\boldsymbol{X}}^{(s)})+\widetilde{\boldsymbol{G}}\big),$$

where i_k is uniformly sampled from $[n]=\{1,2,\ldots,n\}$, and $|I_k|=B$ denotes the minibatch size. Then we perform the following update based on the semi-stochastic gradient with an injected Gaussian noise $\boldsymbol{\epsilon}_k$,

$$\boldsymbol{X}_{k+1}=\boldsymbol{X}_k-\eta\widetilde{\nabla}_k+\sqrt{2\eta/\gamma}\boldsymbol{\epsilon}_k.$$

At the end of the epoch, we use the last iterate as the starting point of the next epoch, i.e., $\widetilde{\boldsymbol{X}}^{(s+1)}=\boldsymbol{X}_{(s+1)m}$.

Now we present SAGA-LD in Algorithm 2. Compared with SVRG-LD, SAGA-LD requires higher memory cost, since it explicitly stores n stochastic gradients in memory, which formulates n columns of a matrix $\widetilde{\boldsymbol{G}}$. $\widetilde{\boldsymbol{G}}$ is initialized as $[\nabla f_1(\boldsymbol{X}_0),\ldots,\nabla f_n(\boldsymbol{X}_0)]$. In the k-th update, we first compute the average of the column vectors in $\widetilde{\boldsymbol{G}}$, i.e., $\widetilde{\boldsymbol{g}}_k=n^{-1}\sum_{i=1}^n\widetilde{\boldsymbol{G}}_i$ as a snapshot gradient, where $\widetilde{\boldsymbol{G}}_i$ is the i-th column of $\widetilde{\boldsymbol{G}}$. Then an index set I_k is uniformly generated from $[n]$ to compute the following approximated gradient

$$\widetilde{\nabla}_k=\frac{1}{B}\sum_{i_k\in I_k}\big(\nabla f_{i_k}(\boldsymbol{X}_k)-\widetilde{\boldsymbol{G}}_{i_k}+\widetilde{\boldsymbol{g}}_k\big),$$

where $B=|I_k|$ is the size of index set I_k. Then we apply such approximated gradient to perform one-step update on the iterate $\boldsymbol{X}_k$, as shown in the line 7 of

stochastic gradient evaluations.

[2]LMC, SVRG-LD and SAGA-LD require hessian Lipschitz assumption in the strongly log-concave regime.

Algorithm 2. At the end of each iteration, we update the columns in $\widetilde{\boldsymbol{G}}$ whose indexes belong to I_k with the stochastic gradients computed in the current iteration, i.e., we set $\widetilde{\boldsymbol{G}}_{i_k} = \nabla f_{i_k}(\boldsymbol{X}_k)$ for all $i_k \in I_k$.

Algorithms 1 and 2 stem from Dubey et al. (2016). However, they only analyzed the mean square error of averaged the sample path based on all iterates $\{\boldsymbol{X}_k\}_{k=0}^K$, while we aim at developing a nonasymptotic analyses of SVRG-LD and SAGA-LD in terms of 2-Wasserstein distance and Algorithms 1 and 2 only require the last iterate $\boldsymbol{X}_K$.

4 MAIN THEORY

In this section, we present our main theoretical results, which characterize the convergence rates of SVRG-LD and SAGA-LD for sampling from non-log-concave distributions. We first lay out the assumptions that are necessary for our theory.

Assumption 4.1 (Smoothness)**.** The function $f_i(\mathbf{x})$ is M-smooth with $M > 0$, i.e., for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, $i = 1, \ldots, n$, we have

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

The smoothness assumption is also known as gradient Lipschitzness in the literature.

Assumption 4.2 (Dissipative)**.** There exist constants $a, b > 0$, such that for all $\mathbf{x} \in \mathbb{R}^d$ we have

$$\langle \nabla F(\mathbf{x}), \mathbf{x} \rangle \geq b\|\mathbf{x}\|_2^2 - a.$$

It is worthy noting that the smoothness assumption is made on all component function $f_i(\mathbf{x})$, while the dissipative assumption is only made on the average of the component functions. Assumption 4.2 is a typical assumption for the ergodicity analysis of stochastic differential equations (SDE) and diffusion approximation (Mattingly et al., 2002; Vollmer et al., 2016; Raginsky et al., 2017; Zhang et al., 2017). It means that, starting from a position that is sufficiently far away from the origin, the Markov process defined by (1.1) moves towards the origin on average. Note that the class of distribution satisfying dissipative assumption covers many densities of interest such as Gaussian mixture model (Lee et al., 2018).

4.1 Convergence Guarantee for SVRG-LD

Now we present our main theoretical results on the nonasymptotic convergence of SVRG-LD.

Theorem 4.3. Under Assumptions 4.1 and 4.2, consider $\{\boldsymbol{X}_k\}_{k=0,1,\ldots,K}$ generated by Algorithm 1 with initial point $\boldsymbol{X}_0 = \mathbf{0}$. The 2-Wasserstein distance between the distribution of $\boldsymbol{X}_k$ and the target distribution π is bounded by

$$\begin{aligned}
&\mathcal{W}_2\big(P(\boldsymbol{X}_k), \pi\big) \\
&\leq D_1\bigg[D_2\Big(\frac{m^2}{B} + 1\Big)k\eta^3 + D_3\Big(\frac{m}{B} + 1\Big)k\eta^2\bigg]^{1/4} \\
&\quad + D_4 e^{-\frac{k\eta}{\gamma D_5}},
\end{aligned} \tag{4.1}$$

where the parameters are defined as

$$\begin{aligned}
D_1 &= 4\sqrt{3/2 + (2b + d/\gamma)k\eta}, \\
D_2 &= 3\gamma M^2\big(2M^2(1 + 1/b)(a + G^2 + d/\gamma) + G^2\big), \\
D_3 &= M^2 d,
\end{aligned}$$

and $G = \max_{i\in[n]} \|f_i(0)\|_2$. Moreover, B is the batch size, m is inner loop length of Algorithm 1, and parameters D_4, D_5 are both in the order of $\exp(O(d + \gamma))$.

Based on Theorem 4.3, we are able to characterize the gradient complexity of Algorithm 1 as well as the choices of hyper parameters including η, m and B. We state these results in the following corollary.

Corollary 4.4. Under the identical assumptions in Theorem 4.3, in order to guarantee that the target accuracy satisfies $\mathcal{W}_2\big(P(\mathbf{x}_k), \pi\big) \leq \epsilon$, we set $mB = O(n)$, $\eta = \widetilde{O}(\epsilon^2 B^{3/2}/n^2 \wedge \epsilon^4 B^2/n) \cdot \exp\big(-\widetilde{O}(\gamma + d)\big)$. Then the gradient complexity of Algorithm 1 is

$$T_g = \widetilde{O}\Big(\frac{nB^{-1/2}}{\epsilon^2} + \frac{n/B + B}{\epsilon^4} + n\Big) \cdot \exp\big(\widetilde{O}(\gamma + d)\big).$$

Moreover, if we set $B = O\big(n^{1/2}\big)$ and $\eta = \widetilde{O}\big(\epsilon^2/n^{1/4} \wedge \epsilon^4\big) \cdot \exp\big(-\widetilde{O}(\gamma + d)\big)$, the gradient complexity is

$$T_g = \widetilde{O}\Big(n + \frac{n^{3/4}}{\epsilon^2} + \frac{n^{1/2}}{\epsilon^4}\Big) \cdot \exp\big(\widetilde{O}(\gamma + d)\big).$$

Remark 4.5. Under identical assumptions in Theorem 4.3, LMC achieves ϵ-accuracy in 2-Wasserstein distance after $T_g = \widetilde{O}\big(n/\epsilon^4\big) \cdot \exp\big(\widetilde{O}(d + \gamma)\big)$ stochastic gradient evaluations (Raginsky et al., 2017). It is obvious that SVRG-LD requires less stochastic gradient evaluations to achieve ϵ-accuracy than LMC.

4.2 Convergence Guarantee for SAGA-LD

Next, we present the following theorem that spells out the convergence rate of SAGA-LD.

Theorem 4.6. Under Assumptions 4.1 and 4.2, consider $\{\boldsymbol{X}_k\}_{k=0,1,\ldots,K}$ generated by Algorithm 2 with initial point $\boldsymbol{X}_0 = \mathbf{0}$. The 2-Wasserstein distance between the distribution of $\boldsymbol{X}_k$ and the target distribu-

tion π is bounded by

$$\begin{aligned}&\mathcal{W}_2\big(P(\boldsymbol{X}_k),\pi\big)\\&\leq D_1\bigg[D_2\bigg(\frac{48n^2}{B^3}+1\bigg)k\eta^3+D_3\bigg(\frac{4n}{B^2}+1\bigg)k\eta^2\bigg]^{1/4}\\&\quad+D_4e^{-\frac{k\eta}{\gamma D_5}},\end{aligned}\tag{4.2}$$

where $G=\max_{i\in[n]}\|f_i(0)\|_2$, B is the batch size, parameters D_1, D_2, D_3, D_4 and D_5 are identical to those in Theorem 4.3.

Based on Theorem 4.6, we present the gradient complexity of SAGA-LD in the following corollary.

Corollary 4.7. Under the same assumptions as in Theorem 4.6, in order to guarantee that the target accuracy satisfies $\mathcal{W}_2\big(P(\mathbf{x}_k),\pi\big)\leq\epsilon$, we set $\eta=\widetilde{O}(\epsilon^2B^{3/2}/n^2\wedge\epsilon^4B^2/n)\cdot\exp\big(-\widetilde{O}(\gamma+d)\big)$, and the gradient complexity of Algorithm 2 is

$$T_g=\widetilde{O}\Big(\frac{nB^{-1/2}}{\epsilon^2}+\frac{n/B+B}{\epsilon^4}+n\Big)\cdot\exp\big(\widetilde{O}(\gamma+d)\big).$$

Moreover, if we set $B=O\big(n^{1/2}\big)$ and $\eta=\widetilde{O}\big(\epsilon^2/n^{1/4}\wedge\epsilon^4\big)\cdot\exp\big(-\widetilde{O}(\gamma+d)\big)$, the gradient complexity becomes

$$T_g=\widetilde{O}\Big(n+\frac{n^{3/4}}{\epsilon^2}+\frac{n^{1/2}}{\epsilon^4}\Big)\cdot\exp\big(\widetilde{O}(d+\gamma)\big).$$

Remark 4.8. It can be clearly observed that the gradient complexities of SVRG-LD and SAGA-LD are essentially identical when we set $mB=O(n)$ in SVRG-LD. This observation also matches the result in Dubey et al. (2016) and Zou et al. (2018b), where the former focuses on the mean squared error of sample path average and the latter only establishes the convergence guarantees for strongly log-concave densities.

Remark 4.9. It is worth noting that our analyses on SVRG-LD and SAGA-LD do not imply the convergence rate of SGLD. However, the convergence rate of SGLD in 2-Wasserstein distance is similar to Equation (3.2) in Raginsky et al. (2017). Based on the argument in Raginsky et al. (2017), the SGLD algorithm cannot be guaranteed to converge to the target distribution if the batch size is not carefully specified. However, empirical study shows that SGLD converges in most cases, which indicates a gap between the theory and the experiment. In particular, we found that SGLD actually converge to the target distribution in our experiment, even when the batch size is set to be 1, and enjoys faster rate than LMC.

5 EXPERIMENTS

In order to explore the behavior of SVRG-LD and SAGA-LD for sampling from non-log-concave densities, we carry out numerical experiments on both synthetic and real dataset in this section. Specifically, we compare the SVRG-LD and SAGA-LD algorithms with LMC and SGLD for sampling from non-log-concave density, independent component analysis (ICA) and Bayesian logistic regression.

5.1 Sampling for Gaussian Mixture Distribution

We first compare the performances of SVRG-LD, SAGA-LD, LMC and SGLD on synthetic data. In particular, we consider the target distribution with form $\pi\propto\exp(-F(\mathbf{x}))=\exp\big(-\sum_{i=1}^n f_i(\mathbf{x})/n\big)$, where each component $\exp(-f_i(\mathbf{x}))$ is defined as

$$\exp(-f_i(\mathbf{x}))=e^{-\|\mathbf{x}-\mathbf{a}_i\|_2^2/2}+e^{-\|\mathbf{x}+\mathbf{a}_i\|_2^2/2},\quad\mathbf{a}_i\in\mathbb{R}^d.$$

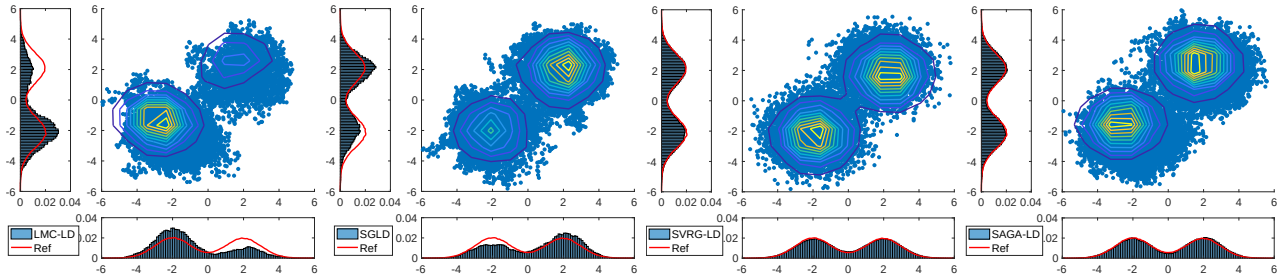
It is easy to verify that $\exp(-f_i(\mathbf{x}))$ is proportion to the PDF of a Gaussian mixture distribution. The function $f_i(\mathbf{x})$ and its gradient can be further simplified as

$$\begin{aligned}f_i(\mathbf{x})&=\frac{1}{2}\|\mathbf{x}-\mathbf{a}_i\|_2^2-\log\big(1+\exp(-2\mathbf{x}^\top\mathbf{a}_i)\big),\\\nabla f_i(\mathbf{x})&=\mathbf{x}-\mathbf{a}_i+\frac{2\mathbf{a}_i}{1+\exp(2\mathbf{x}^\top\mathbf{a}_i)}.\end{aligned}$$

According to Dalalyan (2017b); Dwivedi et al. (2018), when the parameter $\mathbf{a}_i$ is chosen such that $\|\mathbf{a}_i\|_2^2>1$, function $f_i(\mathbf{x})$ defined as above is nonconvex. Moreover, it can be seen that

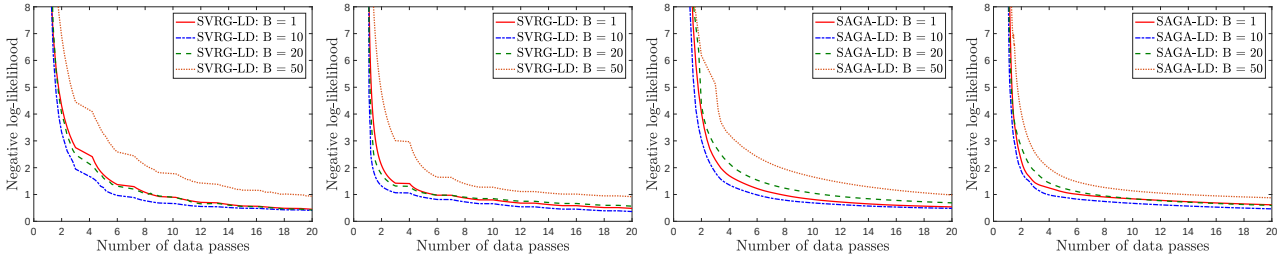
$$\begin{aligned}\langle\nabla f_i(\mathbf{x}),\mathbf{x}\rangle&=\|\mathbf{x}\|_2^2+\frac{1-\exp(2\mathbf{x}^\top\mathbf{a}_i)}{1+\exp(2\mathbf{x}^\top\mathbf{a}_i)}\langle\mathbf{a}_i,\mathbf{x}\rangle\\&\geq\frac{1}{2}\|\mathbf{x}\|_2^2-\frac{1}{2}\|\mathbf{a}_i\|_2^2,\end{aligned}$$

which suggests that function $f_i(\mathbf{x})$ satisfies Dissipative Assumption 4.2 with $b=1/2$ and $a=\|\mathbf{a}_i\|_2^2/2$ and further implies that $F(\mathbf{x})$ is also dissipative. Then we set sample size $n=500$ and dimension $d=10$, and randomly generate parameters $\mathbf{a}_i\sim N(\mu,\boldsymbol{\Sigma})$ with $\mu=(2,\ldots,2)^\top$ and $\boldsymbol{\Sigma}=\mathbf{I}_{d\times d}$. Since it takes a large number of samples to characterize the distribution, which makes repeated experiments computationally expensive, we instead follow Bardenet et al. (2017) to use iterates along one Markov chain to visualize the distribution of iterates obtained by MCMC algorithms. Specifically, we run all four algorithms for 2×10^4 data passes, and make use of the iterates in the last 10^4 data passes to visualize distributions, where the batch sizes for SGLD, SVRG-LD and SAGA-LD are all set to be 10. In Figures 1(a) - 1(d), We compare the distributions generated by LMC, SGLD, SVRG-LD and SAGA-LD while using MCMC with Metropolis-Hasting correction as a reference. It can be observed that both SVRG-LD and SAGA-LD can well approximate the target distribution within



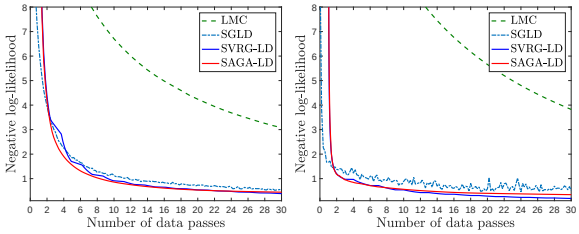
(a) LMC (b) SGLD (c) SVRG-LD (d) SAGA-LD

Figure 1: 2D projection of the kernel densities of random samples generated after 10^4 data passes. (a) - (d) represent 4 different algorithms.



(a) $n = 500$ (b) $n = 5000$ (c) $n = 500$ (d) $n = 5000$

Figure 2: Experiment results for independent components analysis, where x axis indicates the number of data pass and y axis shows the negative log-likelihood on the test data. (a)-(b) Experiment results for SVRG-LD with different batch size. (c)-(d) Experiment results for SAGA-LD with different batch size.



(a) $n = 500$ (b) $n = 5000$

Figure 3: Experiment results of ICA for different algorithms.

2×10^4 datapass, while the distributions generated by LMC and SGLD have obvious deviation from the true one. This suggests that SVRG-LD and SAGA-LD enjoy faster convergence rate than LMC and SGLD, which verifies our theory. However, if we run SGLD and LMC for more iterations, SGLD and LMC can both well approximate the target distribution. More interestingly, we find that SGLD actually requires less gradient evaluations than LMC to well approximate the target distribution, which does not well align with the existing theory.

5.2 Independent Components Analysis

We further apply the SVRG-LD and SAGA-LD algorithms to a Bayesian Independent Component Analysis (ICA) model, and compare their performance with LMC and SGLD. In the ICA model, we are given a dataset with n examples $\mathbf{X} = \{\mathbf{x}_i\}_{i=1,\ldots,n}$. The probability of samples $\mathbf{x}_i$ given the model matrix $\mathbf{W}$ can be written as follows (Welling and Teh, 2011; Dubey et al., 2016)

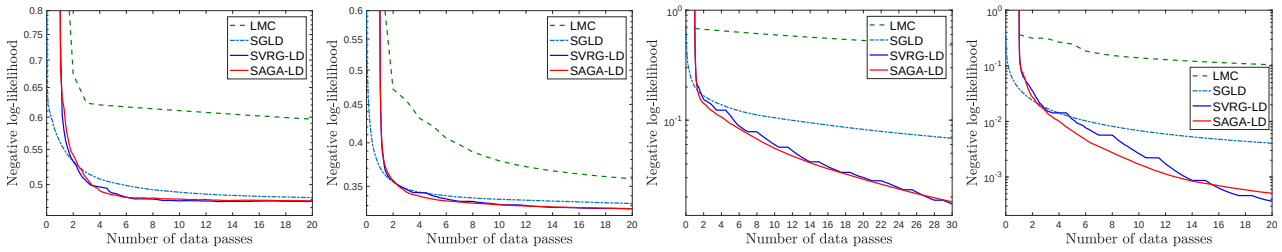
$$p(\mathbf{x}_i|\boldsymbol{W}) = |\det(\mathbf{W})| \prod_i p(\mathbf{w}_i^\top \mathbf{x}_i),$$

where $p(\mathbf{w}_i^\top \mathbf{x}_i) = 1/(4\cosh^2(\mathbf{w}_i^\top \mathbf{x}_i/2))$. We consider Gaussian prior over $\mathbf{W}$, i.e., $p(\mathbf{W}) \sim \mathcal{N}(0, \lambda^{-1}\mathbf{I})$. Then we formulate the log-posterior as the average of n component functions, i.e., $\sum_{i=1}^n f_i(\mathbf{W})/n$, where

$$f_i(\mathbf{W}) = -n\big[\log(|\det(\mathbf{W})|) + 2\sum_{i=1}^{d} \log\big(\cosh(\mathbf{w}_i^\top \mathbf{x}_i/2)\big)\big] + \lambda\|\mathbf{W}\|_F^2.$$

We perform the ICA algorithm on EEG dataset[3], which contains 125337 samples with 34 channels. In this experiment, we consider two regimes with different sample size n. To achieve this, we extract two subsets with size 500 and 5000 from the original dataset, and extract 5000 samples from the rest dataset for test. Follow the same procedures in Welling and Teh (2011); Chen et al. (2014); Zou et al. (2018a), we discard the first 50 iterates as burnin and compute the sample path average to estimate the model matrix parameter $\mathbf{W}$. We first run SVRG-LD and SAGA-LD with different batch sizes $B = 1$, $B = 10$, $B = 20$ and $B = 50$ (the epoch length is set to be $m = 2n/B$ for SVRG-LD),

[3] https://mmspg.epfl.ch/cms/page-58322.html



(a) *pima* (b) *a3a* (c) *gisette* (d) *mushroom*

Figure 4: Experiment results for Bayesian logistic regression, where x axis indicates the number of data pass and y axis shows the negative log-likelihood on the test dataset. (a) - (d) represent 4 different datasets.

and plot the negative log-likelihood on test dataset with respect to the number of effective data pass in Figures 2(a)-2(d). It can be seen that both SVRG-LD and SAGA-LD algorithms have the best performance when the batch size is $B = 10$. Next, we set batch size to be $B = 10$ for both SVRG-LD and SAGA-LD, and compare their convergence performances with those of LMC and SGLD, which are displayed in Figures 3(a)-3(b). It should be noted that in the first epoch, SVRG-LD and SAGA-LD compute the full gradient using all n samples, thus the curves of SVRG-LD and SAGA-LD should start from the first data pass. Moreover, we observe that SVRG-LD and SAGA-LD have comparable performance and both converge faster than SGLD and LMC, this supports our theory.

5.3 Bayesian Logistic Regression

We also apply LMC, SGLD, SVRG-LD and SAGA-LD to a Bayesian logistic regression problem. In this problem, n i.i.d samples $\{\mathbf{x}_i, y_i\}_{i=1,\dots,n}$ are observed, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$ denote the feature and the corresponding label of the i-th sample. In Bayesian logistic model, the likelihood function takes the form $p(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = 1/\big(1+\exp(-y_i\mathbf{x}_i^\top\boldsymbol{\beta})\big)$ where $\boldsymbol{\beta}$ is the regression parameter that requires to be trained. In order to evaluate the performance of SVRG-LD and SAGA-LD when dealing with non-log-concave densities, we consider Gamma prior $p(\boldsymbol{\beta}) \propto \|\boldsymbol{\beta}\|_2^{-\lambda}\exp(-\theta\|\boldsymbol{\beta}\|_2)$. Then we formulate the logarithmic posterior distribution as follows,

$$\log\big[p(\boldsymbol{\beta}|\mathbf{x}_1,\dots,\mathbf{x}_n;y_1,\dots,y_n)\big] \propto -\frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{\beta}),$$

where $f_i(\beta) = n\log\big(1+e^{-y_i\mathbf{x}_i^\top\boldsymbol{\beta}}\big) + \lambda\log(\|\boldsymbol{\beta}\|_2) + \theta\|\boldsymbol{\beta}\|_2$. We compare SVRG-LD and SAGA-LD with the baseline algorithms on four datasets from UCI[4] and Libsvm[5] libraries, which are *pima*, *a3a*, *gisette*, and *mushroom*. Since *pima* and *mushroom* do not have test data, we manually split the whole dataset into training and test parts. Again, we compute the sample path average to estimate the regression parameter $\boldsymbol{\beta}$. The comparison between different algorithms for different datasets are displayed in Figure 4(a) - 4(d). Similarly, SVRG-LD and SAGA-LD start from the first data pass. It can be observed that the performances of SVRG-LD and SAGA-LD are quite similar, and both converge faster than another two baseline algorithms, which suggests that the SVRG-LD and SAGA-LD methods serve as better choices for Bayesian logistic regression with non-log-concave prior compared with LMC and SGLD.

[4] `https://archive.ics.uci.edu/ml/`

[5] `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`

6 CONCLUSIONS AND FUTURE WORK

We studied the SVRG-LD and SAGA-LD methods for sampling from non-log-concave densities, and proved the corresponding convergence rate as well as the gradient complexity when the sampling error is measured as 2-Wasserstein distance. Experimental results showed that SVRG-LD and SAGA-LD achieve similar performance, and converge faster than LMC and SGLD when the target distribution is non-log-concave, which is consistent with our theory.

There are many possible future directions that demand to be explored, such as the convergence rate of SGLD in Wasserstein distance when the target distribution is non-log-concave. In addition, it is also of interest to investigate whether the metropolis hasting step can be applied to further improve the current results.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments. This research was sponsored in part by the National Science Foundation IIS-1906169 and BIGDATA IIS-1855099. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Ahn, S., Korattikara, A. and Welling, M. (2012). Bayesian posterior sampling via stochastic gradient fisher scoring. In *Proceedings of the 29th International Conference on Machine Learning*.

Baker, J., Fearnhead, P., Fox, E. B. and Nemeth, C. (2017). Control variates for stochastic gradient mcmc. *arXiv preprint arXiv:1706.05439* .

Bakry, D., Barthe, F., Cattiaux, P., Guillin, A. et al. (2008). A simple proof of the poincaré inequality for a large class of probability measures. *Electronic Communications in Probability* **13** 60–66.

Bakry, D., Gentil, I. and Ledoux, M. (2013). *Analysis and geometry of Markov diffusion operators*, vol. 348. Springer Science & Business Media.

Bardenet, R., Doucet, A. and Holmes, C. (2017). On markov chain monte carlo methods for tall data. *The Journal of Machine Learning Research* **18** 1515–1557.

Betancourt, M. (2015). The fundamental incompatibility of scalable hamiltonian monte carlo and naive data subsampling. In *International Conference on Machine Learning*.

Bolley, F. and Villani, C. (2005). Weighted csiszár-kullback-pinsker inequalities and applications to transportation inequalities. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, vol. 14.

Bou-Rabee, N. and Hairer, M. (2012). Nonasymptotic mixing of the mala algorithm. *IMA Journal of Numerical Analysis* **33** 80–110.

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

Brosse, N., Durmus, A., Moulines, É. and Pereyra, M. (2017). Sampling from a log-concave distribution with compact support with proximal langevin monte carlo. *arXiv preprint arXiv:1705.08964* .

Bubeck, S., Eldan, R. and Lehec, J. (2015). Sampling from a log-concave distribution with projected langevin monte carlo. *arXiv preprint arXiv:1507.02564* .

Cattiaux, P., Guillin, A. and Wu, L.-M. (2010). A note on talagrands transportation inequality and logarithmic sobolev inequality. *Probability theory and related fields* **148** 285–304.

Chatterji, N. S., Flammarion, N., Ma, Y.-A., Bartlett, P. L. and Jordan, M. I. (2018). On the theory of variance reduction for stochastic gradient monte carlo. *arXiv preprint arXiv:1802.05431* .

Chaudhari, P., Choromanska, A., Soatto, S. and LeCun, Y. (2016). Entropy-sgd: Biasing gradient descent into wide valleys. *arXiv preprint arXiv:1611.01838* .

Chen, C., Wang, W., Zhang, Y., Su, Q. and Carin, L. (2017). A convergence analysis for a class of practical variance-reduction stochastic gradient mcmc. *arXiv preprint arXiv:1709.01180* .

Chen, T., Fox, E. and Guestrin, C. (2014). Stochastic gradient hamiltonian monte carlo. In *International Conference on Machine Learning*.

Chen, Y., Chen, J., Dong, J., Peng, J. and Wang, Z. (2019). Accelerating nonconvex learning via replica exchange Langevin diffusion. In *International Conference on Learning Representations*.

Cheng, X., Chatterji, N. S., Bartlett, P. L. and Jordan, M. I. (2018). Underdamped langevin mcmc: A non-asymptotic analysis. In *Conference On Learning Theory*.

Chiang, T.-S., Hwang, C.-R. and Sheu, S. J. (1987). Diffusion for global optimization in rˆn. *SIAM Journal on Control and Optimization* **25** 737–753.

Dalalyan, A. (2017a). Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent. In *Conference on Learning Theory*.

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

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

Dang, K.-D., Quiroz, M., Kohn, R., Tran, M.-N. and Villani, M. (2017). Hamiltonian Monte Carlo with energy conserving subsampling. *arXiv preprint arXiv:1708.00955* .

Dubey, K. A., Reddi, S. J., Williamson, S. A., Poczos, B., Smola, A. J. and Xing, E. P. (2016). Variance reduction in stochastic gradient langevin dynamics. In *Advances in Neural Information Processing Systems*.

Durmus, A. and Moulines, E. (2015). Nonasymptotic convergence analysis for the unadjusted langevin algorithm. *arXiv preprint arXiv:1507.05021* .

Durmus, A. and Moulines, E. (2016). Sampling from strongly log-concave distributions with the unadjusted langevin algorithm. *arXiv preprint arXiv:1605.01559* .

Dwivedi, R., Chen, Y., Wainwright, M. J. and Yu, B. (2018). Log-concave sampling: Metropolis-hastings algorithms are fast! In *Conference On Learning Theory*.

Eberle, A. et al. (2014). Error bounds for metropolis–hastings algorithms applied to perturbations of gaussian measures in high dimensions. *The Annals of Applied Probability* **24** 337–377.

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

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109.

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*.

Kloeden, P. E. and Platen, E. (1992). Higher-order implicit strong numerical schemes for stochastic differential equations. *Journal of statistical physics* **66** 283–314.

Lee, H., Risteski, A. and Ge, R. (2018). Beyond log-concavity: Provable guarantees for sampling multi-modal distributions using simulated tempering langevin monte carlo. In *Advances in Neural Information Processing Systems*.

Li, Z., Zhang, T. and Li, J. (2018). Stochastic gradient hamiltonian monte carlo with variance reduction for bayesian inference. *arXiv preprint arXiv:1803.11159* .

Ma, Y.-A., Chen, T. and Fox, E. (2015). A complete recipe for stochastic gradient mcmc. In *Advances in Neural Information Processing Systems*.

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

Neal, R. M. et al. (2011). Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo* **2** 113–162.

Raginsky, M., Rakhlin, A. and Telgarsky, M. (2017). Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*.

Reddi, S. J., Hefny, A., Sra, S., Poczos, B. and Smola, A. (2016). Stochastic variance reduction for nonconvex optimization. In *International Conference on Machine Learning*.

Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60** 255–268.

Roberts, G. O. and Tweedie, R. L. (1996a). Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli* 341–363.

Roberts, G. O. and Tweedie, R. L. (1996b). Geometric convergence and central limit theorems for multidimensional hastings and metropolis algorithms. *Biometrika* **83** 95–110.

Teh, Y. W., Thiery, A. H. and Vollmer, S. J. (2016). Consistency and fluctuations for stochastic gradient langevin dynamics. *The Journal of Machine Learning Research* **17** 193–225.

Vollmer, S. J., Zygalakis, K. C. and Teh, Y. W. (2016). Exploration of the (non-) asymptotic bias and variance of stochastic gradient langevin dynamics. *Journal of Machine Learning Research* **17** 1–48.

Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*.

Xu, P., Chen, J., Zou, D. and Gu, Q. (2018). Global convergence of langevin dynamics based algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems*.

Ye, N., Zhu, Z. and Mantiuk, R. K. (2017). Langevin dynamics with continuous tempering for training deep neural networks. In *Advances in Neural Information Processing Systems*.

Zhang, Y., Liang, P. and Charikar, M. (2017). A hitting time analysis of stochastic gradient langevin dynamics. In *Conference on Learning Theory*.

Zou, D., Xu, P. and Gu, Q. (2018a). Stochastic variance-reduced Hamilton Monte Carlo methods. In *Proceedings of the 35th International Conference on Machine Learning*.

Zou, D., Xu, P. and Gu, Q. (2018b). Subsampled stochastic variance-reduced gradient langevin dynamics. In *UAI*.

A Proof of the Main Results

Let $\{\boldsymbol{X}_k\}_{k=0,1,\ldots,K}$ be the discrete-time time-inhomogeneous Markov chain generated by Algorithm 1, and let $\boldsymbol{X}(k\eta)$ be the Langevin dynamics (1.1) at time $k\eta$, which satisfies $\boldsymbol{X}(0) = \boldsymbol{X}_0$. Consider the target distribution $\pi = \exp(-\gamma F(\mathbf{x}))/\int \exp(-\gamma F(\mathbf{x}))d\mathbf{x}$, we decompose the 2-Wasserstein distance $\mathcal{W}_2\big(P(\boldsymbol{X}_k), \pi\big)$ into the following two terms based on triangle inequality.

$$\mathcal{W}_2\big(P(\boldsymbol{X}_k), \pi\big) \leq \mathcal{W}_2\big(P(\boldsymbol{X}_k), P(\boldsymbol{X}(k\eta))\big) + \mathcal{W}_2\big(P(\boldsymbol{X}(k\eta), \pi)\big). \tag{A.1}$$

The first term in (A.1) stands for the discretization error between the continuous-time Langevin dynamics at time $k\eta$ and the k-th iteration of SVRG-LD in 2-Wasserstein distance. The second term describes the convergence of the probability density of Markov process $\{\boldsymbol{X}(k\eta)\}_{t\geq 0}$ to its stationary distribution, and is referred to as the ergordicity of a Markov process. In what follows, we aim at establishing upper bounds for these two terms, respectively.

A.1 Proof of Theorem 4.3

We first study the discretization error between the distribution of continuous Markov process at time $k\eta$ and that of the discrete iterate at the k-th update in Algorithm 1.

Lemma A.1. Under Assumptions 4.1 and 4.2, consider $\{\boldsymbol{X}_k\}_{k=0,1,\ldots,K}$ generated by Algorithm 1 with initial point $\boldsymbol{X}_0 = \mathbf{0}$. The 2-Wasserstein distance between distributions of the iterate $\boldsymbol{X}_k$ in Algorithm 1 and the point $\boldsymbol{X}(k\eta)$ in the Langevin dynamic sequence (1.1) is upper bounded by

$$\mathcal{W}_2\big(P(\boldsymbol{X}_k), P(\boldsymbol{X}(k\eta))\big) \leq D_A\bigg[\bigg(\frac{6\gamma m^2 M^2(n-B)}{B(n-1)} + 3\gamma M^2\bigg)(M^2 D_B^2 + G^2)k\eta^3 + \bigg(\frac{2dmM^2(n-B)}{B(n-1)} + M^2 d\bigg)k\eta^2\bigg]^{1/4},$$

where $D_A = 4\sqrt{3/2 + (2b + d/\gamma)k\eta}$ and $D_B = \sqrt{2(1 + 1/b)(a + G^2 + d/\gamma)}$.

In what follows, we show that the continuous-time process $\{\boldsymbol{X}(t)\}_{t\geq 0}$ converges to its stationary distribution with linear rate.

Lemma A.2. Under Assumptions 4.1 and 4.2, the continuous-time Markov chain $\boldsymbol{X}(t)$ generated by Langevin dynamics (1.1) converges exponentially to the stationary distribution π, i.e.,

$$\mathcal{W}_2\big(P(\boldsymbol{X}(t), \pi)\big) \leq D_4 e^{-\frac{t}{\gamma D_5}},$$

where both D_4 and D_5 are in the order of $\exp\big(\widetilde{O}(\gamma + d)\big)$.

It can be seen that the 2-Wasserstein distance diminishes exponentially fast, and the crucial factor that determines the rate is the parameter in the exponential term, i.e., D_5. It is worth noting that D_5 has an exponential dependence on γ and d.

Proof of Theorem 4.3. In previous parts, we have shown the upper bounds on terms $\mathcal{W}_2\big(P(\boldsymbol{X}_k), P(\boldsymbol{X}(k\eta))\big)$ and $\mathcal{W}_2\big(P(\boldsymbol{X}(k\eta), \pi)\big)$ in (A.1), thus we are ready to prove the main theorem 4.3. It can be seen that by combining Lemmas A.1 and A.2, together with the triangle inequality and the fact that $(n-B)/(n-1) \leq 1$, we have

$$\begin{aligned}
\mathcal{W}_2\big(P(\boldsymbol{X}_k), \pi\big) &\leq \mathcal{W}_2\big(P(\mathbf{x}_k), P(\boldsymbol{X}(k\eta))\big) + \mathcal{W}_2\big(P(\boldsymbol{X}(k\eta), \pi)\big) \\
&\leq D_1\bigg[D_2\bigg(\frac{2m^2}{B} + 1\bigg)k\eta^3 + D_3\bigg(\frac{2m}{B} + 1\bigg)k\eta^2\bigg]^{1/4} + D_4 e^{-\frac{k\eta}{\gamma D_5}},
\end{aligned}$$

where

$$\begin{aligned}
&D_1 = D_A = 4\sqrt{3/2 + (2b + d/\gamma)k\eta}, \\
&D_2 = 3\gamma M^2(M^2 D_B^2 + G^2) = 3\gamma M^2\big(2M^2(1 + 1/b)(a + G^2 + d/\gamma) + G^2\big), \\
&D_3 = M^2 d.
\end{aligned}$$

This completes the proof. □

A.2 Proof of Theorem 4.6

Similar to the proof of SVRG-LD, we first present the following lemma that characterizes the discretization error between the continuous Markov process at time $k\eta$ and the discrete iterate at the k-th update in Algorithm 2.

Lemma A.3. Under Assumptions 4.1 and 4.2, consider $\{\boldsymbol{X}_k\}_{k=0,1,\ldots,K}$ generated by Algorithm 2 with initial point $\boldsymbol{X}_0 = \mathbf{0}$. The 2-Wasserstein distance between distributions of the iterate $\boldsymbol{X}_k$ in Algorithm 1 and the point $\boldsymbol{X}(k\eta)$ in the Langevin dynamic sequence (1.1) is upper bounded by

$$\mathcal{W}_2\big(P(\boldsymbol{X}_k), P(\boldsymbol{X}(k\eta))\big) \leq D_A\bigg[\bigg(\frac{144n^2(n-B)M^2}{B^3(n-1)} + 3M^2\bigg)(M^2D_B^2 + G^2)\gamma k\eta^3 + \bigg(\frac{4n(n-B)M^2d}{B^2(n-1)} + M^2d\bigg)k\eta^2\bigg]^{1/4},$$

where $D_A = 4\sqrt{3/2 + (2b + d/\gamma)k\eta}$ and $D_B = \sqrt{2(1 + 1/b)(a + G^2 + d/\gamma)}$.

In terms of the sequence of continuous-time Langevin dynamics $\{\boldsymbol{X}(t)\}_{t\geq 0}$, Lemma A.2 is also applicable. Thus we are able to complete the proof by combining Lemmas A.2 and A.3.

Proof of Theorem 4.6. Straightforwardly, combining Lemmas A.3 and A.2 together with triangle inequality, and use the fact that $(n-B)/(n-1) \leq 1$, we obtain

$$\begin{aligned}\mathcal{W}_2\big(P(\boldsymbol{X}_k), \pi\big) &\leq \mathcal{W}_2\big(P(\boldsymbol{X}_k), P(\boldsymbol{X}(k\eta))\big) + \mathcal{W}_2\big(P(\boldsymbol{X}(k\eta), \pi)\big)\\ &\leq D_1\bigg[D_2\bigg(\frac{48n^2}{B^3} + 1\bigg)k\eta^3 + D_3\bigg(\frac{4n}{B^2} + 1\bigg)k\eta^2\bigg]^{1/4} + D_4e^{-\frac{k\eta}{\gamma D_5}},\end{aligned}$$

where D_1, D_2, D_3, D_4 and D_5 are identical to those in Theorem 4.3. This completes the proof. □

B Proof of Corollaries

In this section, we provide the proofs of our corollaries in Section 4.

Proof of Corollary 4.4. In order to ensure the ϵ-accuracy in 2-Wasserstein distance, we set

$$\begin{aligned}&D_1\bigg[D_2\Big(\frac{2m^2}{B} + 1\Big)k\eta^3 + D_3\Big(\frac{2m}{B} + 1\Big)k\eta^2\bigg]^{1/4} = \frac{\epsilon}{2},\\ &D_4e^{-\frac{k\eta}{\gamma D_5}} = \frac{\epsilon}{2}.\end{aligned} \tag{B.1}$$

Based on the second equation in (B.1), it can be derived that

$$T \triangleq k\eta = \gamma D_5 \log\Big(\frac{2D_4}{\epsilon}\Big).$$

Then, note that if we have $a + b = c$ for positive constants a, b and c, it either follows that $c \leq 2a$ or $c \leq 2b$. Then we have the following according to the first equation in (B.1),

$$\eta \geq \min\bigg\{\sqrt{\frac{\epsilon^4}{32D_1^4D_2(2m^2/B + 1)T}}, \frac{\epsilon^4}{32D_1^4D_3(2m/B + 1)T}\bigg\}.$$

Combine the above two results, we have

$$k = \frac{T}{\eta} \leq \gamma D_5 \log\Big(\frac{2D_4}{\epsilon}\Big)\bigg(\sqrt{\frac{32D_1^4D_2(2m^2/B + 1)T}{\epsilon^4}} + \frac{32D_1^4D_3(2m/B + 1)T}{\epsilon^4}\bigg).$$

From Lemma A.2, we know that $D_5 = \exp\big(\widetilde{O}(\gamma + d)\big)$, thus the required iteration number k exponentially depends on dimension d and inverse temperature γ. Then, we focus on figuring out dependence on ϵ. Ignoring constants that have no dependence in ϵ and only polynomially depends on γ and d, we have

$$k = \widetilde{O}\Big(\frac{m/B^{1/2} + 1}{\epsilon^2} + \frac{m/B + 1}{\epsilon^4}\Big) \cdot \exp\big(\widetilde{O}(\gamma + d)\big).$$

Note that we have to compute full gradient for k/m times, thus the total gradient complexity is

$$T_g \leq kB + n(k/m \vee 1) \leq kB + \frac{kn}{m} + n.$$

Obviously, minimizing T_g requires $mB = n$. Then, the gradient complexity becomes

$$T_g \leq 2kB + n = \widetilde{O}\Big(\frac{nB^{-1/2}}{\epsilon^2} + \frac{n/B + B}{\epsilon^4} + n\Big) \cdot \exp\big(\widetilde{O}(\gamma + d)\big).$$

Let $B = O(n^{1/2})$, we straightforwardly obtain

$$T_g = \widetilde{O}\Big(n + \frac{n^{3/4}}{\epsilon^2} + \frac{n^{1/2}}{\epsilon^4}\Big) \cdot \exp\big(\widetilde{O}(\gamma + d)\big),$$

which completes the proof. □

Proof of Corollary 4.7. Analogous to the proof of Corollary 4.4, we set

$$D_1\bigg[D_2\bigg(\frac{48n^2}{B^3} + 1\bigg)k\eta^3 + D_3\bigg(\frac{4n}{B^2} + 1\bigg)k\eta^2\bigg]^{1/4} = \frac{\epsilon}{2}$$
$$D_4 e^{-\frac{k\eta}{\gamma D_5}} = \frac{\epsilon}{2}.$$

From the second equation, we obtain

$$k\eta = \gamma D_5 \log\bigg(\frac{2D_4}{\epsilon}\bigg).$$

Let $T = k\eta$, the first equation yields that

$$\eta \geq \min\bigg\{\sqrt{\frac{\epsilon^4}{32D_1^4 D_2(48n^2/B^3 + 1)T}}, \frac{\epsilon^4}{32D_1^4 D_3(4n/B^2 + 1)T}\bigg\}.$$

Then the required number of iterations satisfies

$$k = \frac{T}{\eta} \leq \gamma D_5 \log\bigg(\frac{2D_4}{\epsilon}\bigg)\bigg(\sqrt{\frac{32D_1^4 D_2(48n^2/B^3 + 1)T}{\epsilon^4}} + \frac{32D_1^4 D_3(4n/B^2 + 1)T}{\epsilon^4}\bigg).$$

From Lemma A.2, we know that $D_5 = \exp\big(\widetilde{O}(\gamma + d)\big)$, the complexity k must exponentially depends on dimension d and inverse temperature γ. Then, we focus on figuring out the dependence on ϵ. Ignoring constants that have no dependence in ϵ, we have

$$k = \widetilde{O}\bigg(\frac{n/B^{3/2} + 1}{\epsilon^2} + \frac{n/B^2 + 1}{\epsilon^4}\bigg).$$

Then the corresponding gradient complexity is

$$T_g = n + kB = \widetilde{O}\bigg(n + \frac{n/B^{1/2}}{\epsilon^2} + \frac{n/B + B}{\epsilon^4}\bigg).$$

Plugging the dependence on d and γ, we complete the proof.

□

C Proof of Technical Lemmas

In this section, we prove the technical lemmas in Appendix A.

C.1 Proof of Lemma A.1

We first lay out the following 5 lemmas which is useful for proving Lemma A.1.

Lemma C.1. For all $\mathbf{x} \in \mathbb{R}^d$ and $i = 1, \ldots, n$, we have

$$\|\nabla f_i(\mathbf{x})\|_2 \leq M\|\mathbf{x}\|_2 + G.$$

Moreover, it follows that

$$\|\nabla f_i(\mathbf{x})\|_2^2 \leq 2M\|\mathbf{x}\|_2^2 + 2G.$$

Lemma C.2. Under Assumptions 4.1 and 4.2, for sufficiently small step size η, suppose the initial point is chosen at $\boldsymbol{X}_0 = \mathbf{0}$, the expectation of the ℓ^2 norm of the iterates generated by Algorithm 1 is bounded by

$$\mathbb{E}[\|\boldsymbol{X}_k\|_2^2] \leq 2\bigg(1 + \frac{1}{b}\bigg)\bigg(a + G^2 + \frac{d}{\gamma}\bigg) \triangleq D_B.$$

Lemma C.3. (Bolley and Villani, 2005) For any two probability measures P and Q, if they have finite second moments, the following holds,

$$\mathcal{W}_2(Q, P) \leq \Lambda(\sqrt{D_{KL}(Q||P)} + \sqrt[4]{D_{KL}(Q||P)}),$$

where $\Lambda = 2\inf_{\lambda>0}\sqrt{1/\lambda(3/2 + \log \mathbb{E}_P[e^{\lambda\|\mathbf{x}\|_2^2}])}$, where $\mathbf{x}$ satisfies probability measure P.

Lemma C.4. Under Assumptions 4.1 and 4.2, for sufficiently small step size η and $\beta \geq 2/m$, we have

$$\log \mathbb{E}[\exp(\|\boldsymbol{X}(t)\|_2^2)] \leq \|\mathbf{X}_0\|_2^2 + (2b + d/\gamma)k\eta,$$

where we consider the fact that $\eta \leq 1$, and require that $\gamma > 4$.

Lemma C.5. Under Assumption 4.1, we have the following upper bound on the variance of semi-stochastic gradient $\widetilde{\nabla}_k$ in the SVRG-LD update,

$$\mathbb{E}[\|\widetilde{\nabla}_k - \nabla F(\boldsymbol{X}_k)\|_2^2] \leq \frac{M^2(n-B)}{B(n-1)}\mathbb{E}\|\boldsymbol{X}_k - \widetilde{\boldsymbol{X}}\|_2^2.$$

In order to analyze the long-time behaviour of the error between the discrete-time algorithm and continuous-time Langevin dynamics, we follow the similar technique used in Dalalyan (2017b); Raginsky et al. (2017); Xu et al. (2018), in which a continuous-time Markov process $\{\boldsymbol{D}(t)\}_{t\geq 0}$ is introduced to describe the numerical approximation sequence $\{\boldsymbol{X}_k\}_{k=0,1,\ldots,K}$. Define

$$d\boldsymbol{D}(t) = -b(\boldsymbol{D}(t))dt + \sqrt{2\gamma^{-1}}d\boldsymbol{B}(t), \tag{C.1}$$

where $b(\boldsymbol{D}(t)) = \sum_{k=0}^{\infty}\widetilde{\nabla}_k \mathbb{1}\{t \in \big[\eta k, \eta(k+1)\big)\}$. Integrating (C.1) on interval $\big[\eta k, \eta(k+1)\big)$ yields

$$\boldsymbol{D}(\eta(k+1)) = \boldsymbol{D}(\eta k) - \eta\nabla F(\boldsymbol{D}(\eta k)) + \sqrt{2\eta\gamma^{-1}}\cdot\boldsymbol{\epsilon}_k,$$

where $\boldsymbol{\epsilon}_k \sim N(\mathbf{0}, \mathbf{I}_{d\times d})$ and $\widetilde{\nabla}_k$ is the semi-stochastic gradient at k-th iteration of VR-SGLD. This implies that the distribution of random vector $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k, \ldots)$ is equivalent to that of $(\boldsymbol{D}(\eta), \ldots, \boldsymbol{D}(\eta k), \ldots)$. Note that (C.1) is not a time-homogeneous Markov chain since the semi-stochastic gradient $b(\boldsymbol{D}(t))$ also depends on some historical iterates. However, Gyöngy (1986) showed that one can construct an alternative Markov chain which enjoys the same one-time marginal distribution as that of $\boldsymbol{D}(t)$, which is formulated as follows,

$$d\widetilde{\boldsymbol{D}}(t) = -\widetilde{b}(\widetilde{\boldsymbol{D}}(t))dt + \sqrt{2\gamma^{-1}}d\boldsymbol{B}(t),$$

where $\widetilde{b}(\widetilde{\boldsymbol{D}}(t)) = \mathbb{E}[b(\boldsymbol{D}(t))|\widetilde{D}(t) = D(t)]$. Then we let $\mathbb{P}_t$ denote the distribution of $\widetilde{\boldsymbol{D}}(t)$, which is identical to that of $\boldsymbol{D}(t)$. Recall the SDE of Langevin dynamics, i.e.,

$$d\boldsymbol{X}(t) = -q(\boldsymbol{X}(t))dt + \sqrt{2\gamma^{-1}}d\boldsymbol{B}(t),$$

where $q(\boldsymbol{X}(t)) = \nabla F(\boldsymbol{X}(t))$ and define by $\mathbb{Q}_t$ the distribution of $\boldsymbol{X}(t)$. Now, we have constructed two continuous continuous process. Thus, the Radon-Nykodim derivative of $\mathbb{P}_t$ with respective to $\mathbb{Q}_t$ can be obtained by the Girsanov formula

$$\frac{d\mathbb{P}_t}{d\mathbb{Q}_t}(\widetilde{\boldsymbol{D}}(s)) = \exp\bigg\{\int_0^t (q(\widetilde{\boldsymbol{D}}(s)) - b(\widetilde{\boldsymbol{D}}(s)))^\top d\boldsymbol{B}(s) - \frac{\gamma}{4}\int_0^t \|q(\widetilde{\boldsymbol{D}}(s)) - b(\widetilde{\boldsymbol{D}}(s))\|_2^2 ds\bigg\}.$$

This suggests that the KL divergence between $\mathbb{P}_t$ and $\mathbb{Q}_t$ has the following form

$$D_{KL}(\mathbb{Q}_t||\mathbb{P}_t) = -\mathbb{E}\bigg[\log\bigg(\frac{d\mathbb{P}_t}{d\mathbb{Q}_t}(\widetilde{\boldsymbol{D}}(s))\bigg)\bigg] = \frac{\gamma}{4}\int_0^t \mathbb{E}\big[\|q(\widetilde{\boldsymbol{D}}(s)) - b(\widetilde{\boldsymbol{D}}(s))\|_2^2\big] ds. \tag{C.2}$$

This result gives us an opportunity to estimate the 2-Wasserstein distance $\mathcal{W}_2\big(P(\boldsymbol{X}_k, P(\boldsymbol{X}(k\eta))\big)$, since we are able to apply KL divergence $D_{KL}(\mathbb{Q}_{k\eta}||\mathbb{P}_{k\eta})$ to generate an upper bound based on Lemma C.3. Now, we are going to complete the proof for Lemma A.1 in the following.

Proof of Lemma A.1. Denote P_k, Q_k as the probability density functions of $\boldsymbol{X}_k$ and $\boldsymbol{X}(k\eta)$ respectively. By Lemma C.3, we know that the 2-Wasserstein distance is upper bounded as follows,

$$\mathcal{W}_2(Q_k, P_k) \le \Lambda(\sqrt{D_{KL}(Q_k||P_k)} + \sqrt[4]{D_{KL}(Q_k||P_k)}).$$

Moreover, by data-processing theorem in terms of KL divergence, we have

$$\begin{aligned} D_{KL}(Q_k||P_k) \le D_{KL}(\mathbb{Q}_{k\eta}||\mathbb{P}_{k\eta}) &= \frac{\gamma}{4}\int_0^{k\eta} \mathbb{E}\big[\|q(\widetilde{\boldsymbol{D}}(s)) - b(\widetilde{\boldsymbol{D}}(s))\|_2^2\big] ds \\ &= \frac{\gamma}{4}\int_0^{k\eta} \mathbb{E}\big[\|q(\boldsymbol{D}(s)) - b(\boldsymbol{D}(s))\|_2^2\big] ds, \end{aligned}$$

where the second equality holds due to the fact that $\widetilde{\boldsymbol{D}}(s)$ and $\boldsymbol{D}(s)$ have same one-time distribution. Note that $\boldsymbol{D}(k\eta)$ is generated based on $\boldsymbol{X}_k$. By definition, we know that $b(\boldsymbol{D}(s))$ is a step function and remains constant when $s \in [\eta k, \eta(k+1))$ for any k, and $q(\boldsymbol{D}(s))$ is a continuous function for any s. Based on this observation, it follows that

$$\begin{aligned} &\int_0^{\eta k} \mathbb{E}\big[\|q(\boldsymbol{D}(s)) - b(\boldsymbol{D}(s))\|_2^2\big] ds \\ &\quad= \sum_{v=0}^{k-1}\int_{v\eta}^{\eta(v+1)} \mathbb{E}\big[\|\widetilde{\nabla}_v - \nabla F(\boldsymbol{D}(s))\|_2^2\big] ds \\ &\quad\le 2\eta\sum_{v=0}^{k-1}\mathbb{E}\big[\|\widetilde{\nabla}_v - \nabla F(\boldsymbol{X}_v)\|_2^2\big] + 2\sum_{v=0}^{k-1}\int_{v\eta}^{\eta(v+1)} \mathbb{E}\big[\|\nabla F(\boldsymbol{D}(v\eta)) - \nabla F(\boldsymbol{D}(s))\|_2^2\big] ds, \end{aligned}$$

where the second inequality is due to Jensen's inequality and the convexity of function $\|\cdot\|_2^2$, and $\nabla F(\boldsymbol{X}_v) = \nabla F(\boldsymbol{D}(v\eta))$ denotes the gradient of $F(\cdot)$ at $\boldsymbol{X}_v$. Combine the above results we obtain

$$\begin{aligned} D_{KL}(Q_k||P_k) \le{}& \frac{\gamma\eta}{2}\sum_{v=0}^{k-1}\mathbb{E}[\|\widetilde{\nabla}_v - \nabla F(\boldsymbol{X}_v)\|_2^2] \\ &+ \frac{\gamma}{2}\sum_{v=0}^{k-1}\int_{v\eta}^{\eta(v+1)} \mathbb{E}\big[\|\nabla F(\boldsymbol{D}(v\eta)) - \nabla F(\boldsymbol{D}(s))\|_2^2\big] ds, \end{aligned} \tag{C.3}$$

where the first term on the R.H.S. can be further bounded by

$$\frac{\gamma\eta}{2}\sum_{v=0}^{k-1}\mathbb{E}[\|\widetilde{\nabla}_v - \nabla F(\boldsymbol{X}_v)\|_2^2] \le \frac{\gamma\eta}{2}\sum_{i=0}^{s}\sum_{j=0}^{m-1}\mathbb{E}[\|\widetilde{\nabla}_{im+j} - \nabla F(\boldsymbol{X}_{im+j})\|_2^2],$$

where we use the fact that $k = sm + \ell \leq (s+1)m$ for some $\ell = 0, 1, \ldots, m-1$. Applying Lemma C.5, the inner summation satisfies

$$\sum_{j=0}^{m-1} \mathbb{E}[\|\widetilde{\nabla}_{im+j} - \nabla F(\boldsymbol{X}_{im+j})\|_2^2] \leq \sum_{j=0}^{m-1} \frac{M^2(n-B)}{B(n-1)} \mathbb{E}\big\|\boldsymbol{X}_{im+j} - \widetilde{\boldsymbol{X}}^{(i)}\big\|_2^2. \tag{C.4}$$

Note that we have

$$\begin{aligned}
&\mathbb{E}\big\|\boldsymbol{X}_{im+j} - \widetilde{\boldsymbol{X}}^{(i)}\big\|_2^2 \\
&\quad= \mathbb{E}\bigg\|\sum_{u=0}^{j-1} \eta\big(\nabla f_{i_{im+u}}(\boldsymbol{X}_{im+u}) - \nabla f_{i_{im+u}}(\widetilde{\boldsymbol{X}}^{(i)}) + \nabla F(\widetilde{\boldsymbol{X}}^{(i)})\big) - \sum_{u=0}^{j-1} \sqrt{\frac{2\eta}{\gamma}}\epsilon_j\bigg\|_2^2 \\
&\quad\leq j\sum_{u=0}^{j-1} \mathbb{E}\big[2\eta^2\big\|\nabla f_{i_{im+u}}(\boldsymbol{X}_{im+u}) - \nabla f_{i_{im+u}}(\widetilde{\boldsymbol{X}}^{(i)}) + \nabla F(\widetilde{\boldsymbol{X}}^{(i)})\big\|_2^2\big] + \sum_{u=0}^{j-1} \frac{4\eta d}{\gamma} \\
&\quad\leq j\sum_{u=0}^{j-1} \mathbb{E}\big[6\eta^2\big(\|\nabla f_{i_{im+u}}(\boldsymbol{X}_{im+u})\|_2^2 + \|\nabla f_{i_{im+u}}(\widetilde{\boldsymbol{X}}^{(i)})\|_2^2 + \|\nabla F(\widetilde{\boldsymbol{X}}^{(i)})\big\|_2^2)\big] + \sum_{u=0}^{j-1} \frac{4\eta d}{\gamma} \\
&\quad\leq 36j^2\eta^2(M^2D_B^2 + G^2) + \frac{4j\eta d}{\gamma},
\end{aligned} \tag{C.5}$$

where the first and second inequalities follow from Young's inequality and the last one follows from Lemma C.1 and Lemma C.2, and $D_B = \sqrt{2(1+1/b)(a + G^2 + d/\gamma)}$. Submit (C.5) back into (C.4) we have

$$\begin{aligned}
\sum_{j=0}^{m-1} \mathbb{E}[\|\widetilde{\nabla}_{im+j} - \nabla F(\boldsymbol{X}_{im+j})\|_2^2] &\leq \sum_{j=0}^{m-1} \frac{4M^2(n-B)}{B(n-1)}\bigg(9j^2\eta^2(M^2D_B^2 + G^2) + \frac{j\eta d}{\gamma}\bigg) \\
&\leq \frac{4M^2(n-B)}{B(n-1)}\bigg(3m^3\eta^2(M^2D_B^2 + G^2) + \frac{m^2\eta d}{\gamma}\bigg).
\end{aligned} \tag{C.6}$$

Submitting (C.6) into (C.3) yields

$$\sum_{v=0}^{k-1} \mathbb{E}[\|\widetilde{\nabla}_v - \nabla F(\boldsymbol{X}_v)\|_2^2] \leq \frac{4kM^2(n-B)}{B(n-1)}\bigg(3m^2\eta^2(M^2D_B^2 + G^2) + \frac{m\eta d}{\gamma}\bigg). \tag{C.7}$$

Next, we are going to upper bound the second term on the R.H.S of (C.3). According to the smoothness assumption on $F(\mathbf{x})$, we have

$$\mathbb{E}[\|\nabla F(\boldsymbol{D}(v\eta)) - \nabla F(\boldsymbol{D}(s))\|_2^2] \leq M^2\mathbb{E}[\|\boldsymbol{D}(s) - \boldsymbol{D}(v\eta)\|_2^2],$$

which yields that

$$\begin{aligned}
&\sum_{v=0}^{k-1} \int_{v\eta}^{\eta(v+1)} \mathbb{E}\big[\|\nabla F(\boldsymbol{X}_v) - \nabla F(\boldsymbol{N}(s))\|_2^2\big] ds \\
&\qquad\leq \sum_{v=0}^{k-1} \int_{v\eta}^{\eta(v+1)} M^2\mathbb{E}\big[\|\boldsymbol{D}(s) - \boldsymbol{D}(v\eta)\|_2^2\big] ds \\
&\qquad= \sum_{v=0}^{k-1} \int_{v\eta}^{\eta(v+1)} M^2\bigg((s-v\eta)^2\mathbb{E}[\|\widetilde{\nabla}_v\|_2^2] + \frac{2(s-v\eta)d}{\gamma}\bigg) ds \\
&\qquad\leq \frac{M^2\eta^3}{3}\sum_{v=0}^{k-1} \mathbb{E}[\|\widetilde{\nabla}_v\|_2^2] + \frac{2kM^2\eta^2 d}{\gamma}.
\end{aligned} \tag{C.8}$$

By Lemma C.1, we know that

$$\begin{aligned}\mathbb{E}[\|\widetilde{\nabla}_v\|_2^2] &= \mathbb{E}\bigg[\bigg\|1/B\sum_{i_k\in I_k}\Big(\nabla f_{i_k}(\boldsymbol{X}_v) - \nabla f_{i_k}(\widetilde{\boldsymbol{X}}^{(s)}) + \nabla F(\widetilde{\boldsymbol{X}}^{(s)})\Big)\bigg\|_2^2\bigg] \\ &\le 3\mathbb{E}\big[(M\|\boldsymbol{X}_v\|_2+G)^2 + 2(M\|\widetilde{\boldsymbol{X}}^{(s)}\|_2+G)^2\big] \\ &\le 6M^2\mathbb{E}[\|\boldsymbol{X}_v\|_2^2] + 12N^2\mathbb{E}[\|\widetilde{\boldsymbol{X}}^{(s)}\|_2^2] + 18G^2 \\ &\le 18M^2D_B^2 + 18G^2,\end{aligned}$$

where $D_B = \sqrt{2(1+1/b)(a+G^2+d/\gamma)}$ is defined in Lemma C.2, and the last second inequality follows from the fact that $(M\|\boldsymbol{X}_v\|_2+G)^2 \le 2M^2\|\boldsymbol{X}_v\|_2^2 + 2G^2$. Thus, combining (C.3), (C.8), (C.7), we arrive at

$$\begin{aligned}D_{KL}(Q_k||P_k) &\le \frac{2k\eta\gamma M^2(n-B)}{B(n-1)}\bigg(3m^2\eta^2(M^2D_B^2+G^2) + \frac{m\eta d}{\gamma}\bigg) \\ &\quad + \frac{\gamma}{2}\bigg(6M^2k\eta^3(M^2D_B^2+G^2) + \frac{2kM^2\eta^2 d}{\gamma}\bigg) \\ &= \bigg(\frac{6m^2M^2(n-B)}{B(n-1)} + 3\gamma M^2\bigg)(M^2D_B^2+G^2)k\eta^3 \\ &\quad + \bigg(\frac{2dmM^2(n-B)}{B(n-1)} + M^2d\bigg)k\eta^2 \end{aligned} \tag{C.9}$$

Combining (C.9) and Lemma C.3, assume that $D_{KL}(Q_k||P_k) \le 1$, and choose $\lambda = 1$ in Lemma C.3 we obtain

$$\begin{aligned}&\mathcal{W}_2\big(P(\boldsymbol{X}_k), P(\boldsymbol{X}(k\eta))\big) \\ &\le D_A\bigg[\bigg(\frac{6\gamma m^2M^2(n-B)}{B(n-1)} + 3\gamma M^2\bigg)(M^2D_B^2+G^2)k\eta^3 + \bigg(\frac{2dmM^2(n-B)}{B(n-1)} + M^2d\bigg)k\eta^2\bigg]^{1/4},\end{aligned}$$

where $D_A = 2\Lambda = 4\sqrt{3/2 + (2b+d/\gamma)k\eta}$ since $\|\boldsymbol{X}_0\|_2 = 0$. □

C.2 Proof of Lemma A.2

In the following, we adopt the method in Bakry et al. (2013) to show the exponential ergodicity of Langevin diffusion (1.1). In detail, following Bakry et al. (2013), we show the exponential decay in terms of KullbackLeibler divergence (KL divergence) between the probability measure $P_L^t(\cdot)$ and the stationary distribution π, characterize the convergence rate of Langevin dynamics, and link the 2-Wasserstein distance and KL divergence using Otto-Villani theorem (Bakry et al., 2013). We first present the following lemma, which is necessary for the estimation of constant D_3 in Lemma A.2.

Lemma C.6. (Raginsky et al. (2017)) Consider Langevin diffusion (1.1), under Assumptions 4.1 and 4.2, its stationary distribution π satisfies the logarithmic Sobolev inequality with constant C, i.e., for any function h such that $\int_{\mathbb{R}^d} h|\log h|d\pi < \infty$ and $\int_{\mathbb{R}^d} h^2 d\pi = 1$, we have

$$\int_{\mathbb{R}^d} 2h^2\log h d\pi \le 2\Gamma\cdot\int_{\mathbb{R}}^d \|\nabla h\|_2^2 d\pi, \tag{C.10}$$

where $\Gamma = \exp\big(\widetilde{O}(\gamma+d)\big)$.

Proof of Lemma A.2. By Lemma C.6, the stationary distribution π satisfies logarithmic Sobolev inequality with constant Γ. According to Bakry et al. (2013) (Theorem 5.2.1), we know that for Langevin diffusion (1.1), the KL divergence between probability measure of $\boldsymbol{X}(t)$ and the stationary distribution π satisfies the following inequality for any $t \ge 0$,

$$D\big(P_L^t(\cdot)\|\pi\big) \le D\big(P_L^0(\cdot)\|\pi\big)e^{-\frac{2t}{\gamma\Gamma}}. \tag{C.11}$$

where Γ is the constant in logarithmic Sobolev inequality. It can be seen that the above result gives the form of exponential decay, and the corresponding rate relies on the constant Γ, which is specified in Lemma C.6.

Moreover, according to Bakry et al. (2013) (Theorem 9.6.1), it can be seen that if (C.10) holds for stationary distribution π with constant Γ, we have the following hold for probability measure $P_L^t(\cdot)$,

$$\mathcal{W}_2(P_L^t(\cdot),\pi) \leq \sqrt{2\Gamma \cdot D\big(P_L^t(\cdot)\|\pi\big)}, \tag{C.12}$$

where $\mathcal{W}_2(u,v)$ is the 2-Wasserstein distance between probability measures u and v. Submit (C.12) into (C.11), we have the following

$$\mathcal{W}_2(P_L^t(\cdot),\pi) \leq \sqrt{2\Gamma \cdot D\big(P_L^0(\cdot)\|\pi\big)}e^{-\frac{t}{\gamma\Gamma}}.$$

Let $D_4 = \sqrt{2\Gamma \cdot D\big(P(\boldsymbol{X}(0))\|\pi\big)}$ and $D_5 = \Gamma$, we have

$$\mathcal{W}_2(P_L^t(\cdot),\pi) \leq D_4 e^{-\frac{t}{\gamma D_5}},$$

which completes the proof.

□

C.3 Proof of Lemma A.3

We first lay out the following Lemmas which will be used to prove Lemma A.3

Lemma C.7. Under Assumptions 4.1 and 4.2, for sufficiently small step size η, suppose the initial point is $\boldsymbol{X}_0 = \mathbf{0}$, the expectation of the squared ℓ^2 norm of the iterates generated by Algorithm 2 is bounded by

$$\mathbb{E}[\|\boldsymbol{X}_k\|_2^2] \leq 2\bigg(1+\frac{1}{b}\bigg)\bigg(a + G^2 + \frac{d}{\gamma}\bigg) = D_B.$$

Lemma C.8. Under Assumption 4.1, we have the following upper bound on the variance of semi-stochastic gradient $\widetilde{\nabla}_k$ in the SAGA-LD update,

$$\mathbb{E}[\|\widetilde{\nabla}_k - \nabla F(\boldsymbol{X}_k)\|_2^2] \leq \frac{n-B}{B(n-1)}\mathbb{E}\|\nabla f_{i_k}(\boldsymbol{X}_k) - \widetilde{\mathbf{G}}_{i_k}\|_2^2.$$

Similar to the proof of Lemma A.1, we have two continuous Markov chains, one of them is generated by the Langevin dynamics, i.e.,

$$d\boldsymbol{X}(t) = -q(\boldsymbol{X}(t))dt + \sqrt{2\gamma^{-1}}d\boldsymbol{B}(t),$$

where $q(\boldsymbol{X}(t)) = \nabla f(\boldsymbol{X}(t))$, and the other one, denoted as $\{\boldsymbol{H}(t)\}_{t\geq 0}$, follows from the iterate sequence $\{\boldsymbol{X}_k\}_{k=0,1,\ldots,K}$ generated by Algorithm 2, and takes the following form

$$d\boldsymbol{H}(t) = -h(\boldsymbol{H}(t))dt + \sqrt{2\gamma^{-1}}d\boldsymbol{B}(t),$$

where the drift term $h(\boldsymbol{H}(t)) = \widetilde{\nabla}_k$ is defined in Algorithm 2. Similar to the proof of SVRG-LD, $\{\boldsymbol{H}(t)\}_{t\geq 0}$ does not form a Markov Chain since the drift term $h(\boldsymbol{H}(t))$ depends on some history iterates $\{\boldsymbol{H}(\tau), \tau \leq t\}$. However, we can again construct a Markov chain $\{\widetilde{\boldsymbol{H}}(t)\}_{t\geq 0}$ which possesses the identical one-time distribution of $\{\boldsymbol{H}(t)\}_{t\geq 0}$. $\{\widetilde{\boldsymbol{H}}(t)\}_{t\geq 0}$ is defined by the following SDE

$$d\widetilde{\boldsymbol{H}}(t) = -\widetilde{h}(\widetilde{\boldsymbol{H}}(t))dt + \sqrt{2\gamma^{-1}}d\boldsymbol{B}(t),$$

where $\widetilde{h}(\widetilde{\boldsymbol{H}}(t)) = \mathbb{E}[h(\boldsymbol{H}(t))|\widetilde{\boldsymbol{H}}(t) = \boldsymbol{H}(t)]$. Let $\mathbb{P}_t$ and $\mathbb{Q}_t$ denote the distributions of $\widetilde{\boldsymbol{H}}(t)$ and $\boldsymbol{X}(t)$ respectively. Using the Radon-Nykodim derivative of $\mathbb{P}_t$ with respect to $\mathbb{Q}_t$, we obtain the following formula in terms of the KL divergence between $\mathbb{P}_t$ and $\mathbb{Q}_t$,

$$D_{KL}(\mathbb{Q}_t||\mathbb{P}_t) = -\mathbb{E}\bigg[\log\bigg(\frac{d\mathbb{P}_t}{d\mathbb{Q}_t}(\widetilde{\boldsymbol{H}}(t))\bigg)\bigg] = \frac{\gamma}{4}\int_0^t \mathbb{E}\big[\|q(\widetilde{\boldsymbol{H}}(s)) - h(\widetilde{\boldsymbol{H}}(s)\|\big].$$

Then we are going to complete the proof.

Proof of Lemma A.3. Note that $h(\boldsymbol{H}(s))$ is a step function and remains constant when $s \in [v\eta, (v+1)\eta]$ for any v, then we have

$$\begin{aligned}
\int_0^{k\eta} \mathbb{E}\big[\|q(\widetilde{\boldsymbol{H}}(s)) - h(\widetilde{\boldsymbol{H}}(s))\|_2^2\big] &= \int_0^{k\eta} \mathbb{E}\big[\|q(\boldsymbol{H}(s)) - h(\boldsymbol{H}(s))\|_2^2\big] \\
&= \sum_{v=0}^{k-1} \int_{v\eta}^{(v+1)\eta} \mathbb{E}\big[\|\widetilde{\nabla}_v - \nabla F(\boldsymbol{H}(s))\|_2^2\big] \\
&\le 2\sum_{v=0}^{k-1} \int_{v\eta}^{(v+1)\eta} \mathbb{E}\big[\|\widetilde{\nabla}_v - \nabla F(\boldsymbol{X}_v)\|_2^2\big] \\
&\quad + 2\sum_{v=0}^{k-1} \int_{v\eta}^{(v+1)\eta} \mathbb{E}\big[\|\nabla F(\boldsymbol{H}(v\eta)) - \nabla F(\boldsymbol{H}(s))\|_2^2\big], \qquad \text{(C.13)}
\end{aligned}$$

where the first equality holds since $\widetilde{\boldsymbol{H}}(s)$ and $\boldsymbol{H}(s)$ has identical distribution, and the inequality is by Young's inequality and the fact that $\boldsymbol{X}_v = \boldsymbol{H}(v\eta)$. In terms of the first term on the R.H.S of the above inequality, the following holds according to Lemma C.8,

$$\mathbb{E}\big[\|\widetilde{\nabla}_v - \nabla F(\boldsymbol{X}_v)\|_2^2\big] \le \frac{n-B}{B(n-1)} \mathbb{E}\big[\|\nabla f_{i_v}(\boldsymbol{X}_v) - \widetilde{\boldsymbol{G}}_{i_v}\|_2^2\big].$$

Note that $\widetilde{\boldsymbol{G}}_{i_v} = \nabla f_{i_v}(\boldsymbol{X}_u)$ for some u satisfying $0 \le u < v$. Then we have

$$\begin{aligned}
\mathbb{E}\big[\|\widetilde{\nabla}_v - \nabla F(\boldsymbol{X}_v)\|_2^2\big] &= \frac{(n-B)}{B(n-1)} \mathbb{E}\big[\|\nabla f_{i_v}(\boldsymbol{X}_v) - \nabla f_{i_v}(\boldsymbol{X}_u)\|_2^2\big] \\
&\le \frac{(n-B)M^2}{B(n-1)} \mathbb{E}[\|\boldsymbol{X}_v - \boldsymbol{X}_u\|_2^2]. \qquad \text{(C.14)}
\end{aligned}$$

Note that

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{X}_u - \boldsymbol{X}_v\|_2^2 | u] &= \mathbb{E}\bigg[\bigg\|\sum_{j=u}^{v-1} \eta\widetilde{\nabla}_j + \sum_{j=u}^{v-1} \sqrt{\frac{2\eta}{\gamma}}\boldsymbol{\epsilon}_j\bigg\|_2^2\bigg] \\
&\le 2(u-v)\sum_{j=u}^{v-1} \mathbb{E}[\|\widetilde{\nabla}_j\|_2^2] + \frac{4(u-v)\eta d}{\gamma} \\
&\le 36(u-v)^2\eta^2(M^2D_B^2 + G^2) + \frac{4(u-v)\eta d}{\gamma},
\end{aligned}$$

where the first inequality follows from Jensen's inequality, and the second inequality is by Young's inequality and Lemma C.7, where $D_B = \sqrt{2(1+1/b)(a + G^2 + d/\gamma)}$. Then we have

$$\mathbb{E}[\|\boldsymbol{X}_v - \boldsymbol{X}_u\|_2^2] = \mathbb{E}\mathbb{E}[\|\boldsymbol{X}_u - \boldsymbol{X}_v\|_2^2 | u, v] \le \mathbb{E}\big[36(u-v)^2\eta^2(M^2D_B^2 + G^2) + \frac{4(u-v)\eta d}{\gamma}\big].$$

Let $q = 1 - (1-1/n)^B$ be the probability of choosing a particular index, then

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{X}_v - \boldsymbol{X}_u\|_2^2] &\le 36\eta^2(M^2D_B^2 + G^2)\mathbb{E}[(u-v)^2] + \frac{4\eta d}{\gamma}\mathbb{E}[(u-v)] \\
&= 36\eta^2(M^2D_B^2 + G^2)\sum_{t=0}^{v-1}(v-t)^2(1-q)^{v-t-1}q + \frac{4\eta d}{\gamma}\sum_{t=0}^{v-1}(v-t)(1-q)^{v-t-1}q \\
&\le 36\eta^2(M^2D_B^2 + G^2)\sum_{t=0}^{\infty} t^2(1-q)^{t-1}q + \frac{4\eta d}{\gamma}\sum_{t=0}^{\infty} t(1-q)^{t-1}q \\
&\le \frac{72\eta^2(M^2D_B^2 + G^2)}{q^2} + \frac{4\eta d}{q\gamma}.
\end{aligned}$$

From Dubey et al. (2016) we know that $q = 1-(1-1/n)^B \geq B/(2n)$, thus

$$\mathbb{E}[\|\boldsymbol{X}_v - \boldsymbol{X}_u\|_2^2] \leq \frac{288n^2\eta^2(M^2D_B^2+G^2)}{B^2} + \frac{8n\eta d}{B\gamma}. \tag{C.15}$$

In the following, we are going to bound the second term on the R.H.S of (C.13). Based on the definition of $\boldsymbol{H}(s)$, the following holds,

$$\begin{aligned}
\int_{v\eta}^{(v+1)\eta} \mathbb{E}\big[\|\nabla F(N(v\eta)) - \nabla F(\boldsymbol{H}(s))\|_2^2\big]ds &\leq \int_{v\eta}^{(v+1)\eta} M^2\mathbb{E}[\|\boldsymbol{H}(s) - \boldsymbol{H}(v\eta)\|_2^2]ds \\
&= \int_{v\eta}^{(v+1)\eta} M^2\bigg((s-v\eta)^2\mathbb{E}[\|\widetilde{\nabla}_v\|_2^2] + \frac{2(s-v\eta)d}{\gamma}\bigg)ds \\
&\leq \frac{M^2\eta^3}{3}\mathbb{E}[\|\widetilde{\nabla}_v\|_2^2] + \frac{2M^2\eta^2 d}{\gamma} \\
&\leq 6M^2\eta^3(M^2D_B^2+G^2) + \frac{2M^2\eta^2 d}{\gamma}.
\end{aligned} \tag{C.16}$$

where the last inequality follows from Lemma C.7. Then, plugging (C.16), (C.15), (C.14) into (C.13), we arrive at

$$\begin{aligned}
D_{KL}(Q_k||P_k) &\leq \frac{\gamma}{4}\int_0^{k\eta}\mathbb{E}\big[\|q(\boldsymbol{H}(s) - h(\boldsymbol{H}(s))\|_2^2\big] \\
&\leq \frac{\gamma}{2}\bigg[\frac{k\eta^2 n(n-B)M^2}{B^2(n-1)}\bigg(\frac{288n\eta(M^2D_B^2+G^2)}{B} + \frac{8d}{\gamma}\bigg) + k\eta^2M^2\bigg(6\eta(M^2D_B^2+G^2) + \frac{2d}{\gamma}\bigg)\bigg] \\
&= \bigg(\frac{144n^2(n-B)M^2}{B^3(n-1)} + 3M^2\bigg)(M^2D_B^2+G^2)\gamma k\eta^3 + \bigg(\frac{4n(n-B)M^2d}{B^2(n-1)} + M^2d\bigg)k\eta^2.
\end{aligned}$$

Apply Lemma C.3, and choose $\lambda = 1$ in Lemma C.3 we obtain

$$\begin{aligned}
&\mathcal{W}_2\big(P(\boldsymbol{X}_k), P(\boldsymbol{X}(k\eta))\big) \\
&\leq D_A\bigg[\bigg(\frac{144n^2(n-B)M^2}{B^3(n-1)} + 3M^2\bigg)(M^2D_B^2+G^2)\gamma k\eta^3 + \bigg(\frac{4n(n-B)M^2d}{B^2(n-1)} + M^2d\bigg)k\eta^2\bigg]^{1/4},
\end{aligned}$$

where $D_A = 4\sqrt{3/2 + (2b+d/\gamma)k\eta}$.

□

D Proof of Auxiliary Lemmas in Appendix C

In this section, we prove the technical lemmas in Appendix C.

D.1 Proof of Lemma C.1

Proof. Let $G = \max_{i=1,\ldots,n}\|f_i(\mathbf{0})\|$, then we have

$$\|\nabla f_i(\mathbf{x})\|_2 \leq \|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{0})\|_2 + \|\nabla f_i(\mathbf{0})\|_2 \leq M\|\mathbf{x}\|_2 + G,$$

where the first inequality follows from triangle inequality and the second inequality follows from Assumption 4.1. This completes the proof. □

D.2 Proof of Lemma C.2

Proof. We prove the bound for $\mathbb{E}[\|\boldsymbol{X}_k\|_2^2]$ by mathematical induction. Since $\widetilde{\nabla}_k = 1/B\sum_{i_k\in I_k}\big(\nabla f_{i_k}(\boldsymbol{X}_k) - \nabla f_{i_k}(\widetilde{\boldsymbol{X}}^{(s)}) + \nabla F(\widetilde{\boldsymbol{X}}^{(s)})\big)$, we have

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{X}_{k+1}\|_2^2] &= \mathbb{E}[\|\boldsymbol{X}_k - \eta\widetilde{\nabla}_k\|_2^2] + \sqrt{\frac{8\eta}{\gamma}}\mathbb{E}[\langle\boldsymbol{X}_k - \eta\widetilde{\nabla}_k, \boldsymbol{\epsilon}_k\rangle] + \frac{2\eta}{\gamma}\mathbb{E}[\|\boldsymbol{\epsilon}_k\|_2^2] \\
&= \mathbb{E}[\|\boldsymbol{X}_k - \eta\widetilde{\nabla}_k\|_2^2] + \frac{2\eta d}{\gamma},
\end{aligned} \tag{D.1}$$

where the second equality follows from the fact that $\boldsymbol{\epsilon}_k$ is independent of $\boldsymbol{X}_k$ and standard Gaussian.

We prove it by induction. First, consider the case when $k = 1$. Since we choose the initial point at $\boldsymbol{X}_0 = \mathbf{0}$, we immediately have

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{X}_1\|_2^2] &= \mathbb{E}[\|\boldsymbol{X}_0 - \eta\widetilde{\nabla}_0\|_2^2] + \sqrt{\frac{8\eta}{\gamma}}\mathbb{E}[\langle \boldsymbol{X}_0 - \eta\widetilde{\nabla}_0, \boldsymbol{\epsilon}_0\rangle] + \frac{2\eta}{\gamma}\mathbb{E}[\|\boldsymbol{\epsilon}_0\|_2^2] \\ &= \eta^2\mathbb{E}[\|\nabla F(\boldsymbol{X}_0)\|_2^2] + \frac{2\eta d}{\gamma} \\ &\leq \eta^2 G^2 + \frac{2\eta d}{\gamma},\end{aligned}$$

where the second equality holds due to the fact that $\widetilde{\nabla}_0 = \nabla F(\boldsymbol{X}_0)$ and the inequality follows from Lemma C.1. For sufficiently small η we can easily make the conclusion holds for $\mathbb{E}[\|\boldsymbol{X}_1\|_2^2]$.

Now assume that the conclusion holds for all iteration from 1 to k, then for the $(k+1)$-th iteration, by (D.1) we have,

$$\mathbb{E}[\|\boldsymbol{X}_{k+1}\|_2^2] = \mathbb{E}[\|\boldsymbol{X}_k - \eta\widetilde{\nabla}_k\|_2^2] + \frac{2\eta d}{\gamma}, \tag{D.2}$$

For the first term on the R.H.S of (D.2) we have

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{X}_k - \eta\widetilde{\nabla}_k\|_2^2] &= \mathbb{E}[\|\boldsymbol{X}_k - \eta\nabla F(\boldsymbol{X}_k)\|_2^2] + 2\eta\mathbb{E}\langle \boldsymbol{X}_k - \eta\nabla F(\boldsymbol{X}_k), \nabla F(\boldsymbol{X}_k) - \widetilde{\nabla}_k\rangle \\ &\quad + \eta^2\mathbb{E}[\|\nabla F(\boldsymbol{X}_k) - \widetilde{\nabla}_k\|_2^2] \\ &= \underbrace{\mathbb{E}[\|\boldsymbol{X}_k - \eta\nabla F(\boldsymbol{X}_k)\|_2^2]}_{T_1} + \underbrace{\eta^2\mathbb{E}[\|\nabla F(\boldsymbol{X}_k) - \widetilde{\nabla}_k\|_2^2]}_{T_2},\end{aligned} \tag{D.3}$$

where the second equality holds due to the fact that $\mathbb{E}[\widetilde{\nabla}_k] = \nabla F(\boldsymbol{X}_k)$. For term T_1, we can further bound it by

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{X}_k - \eta\nabla F(\boldsymbol{X}_k)\|_2^2] &= \mathbb{E}[\|\boldsymbol{X}_k\|_2^2] - 2\eta\mathbb{E}[\langle \boldsymbol{X}_k, \nabla F(\boldsymbol{X}_k)\rangle] + \eta^2\mathbb{E}[\|\nabla F(\boldsymbol{X}_k)\|_2^2] \\ &\leq \mathbb{E}[\|\boldsymbol{X}_k\|_2^2] + 2\eta(a - b\mathbb{E}[\|\boldsymbol{X}_k\|_2^2]) + 2\eta^2(M^2\mathbb{E}[\|\boldsymbol{X}_k\|_2^2] + G^2) \\ &= (1 - 2\eta b + 2\eta^2 M^2)\mathbb{E}[\|\boldsymbol{X}_k\|_2^2] + 2\eta a + 2\eta^2 G^2,\end{aligned} \tag{D.4}$$

where the inequality follows from Lemma C.1 and triangle inequality. For term T_2, by Lemma C.5 we have

$$\mathbb{E}\|\nabla F(\boldsymbol{X}_k) - \widetilde{\nabla}_k\|_2^2 \leq \frac{M^2(n-B)}{B(n-1)}\mathbb{E}\big\|\boldsymbol{X}_k - \widetilde{\boldsymbol{X}}^{(s)}\big\|_2^2 \leq \frac{2M^2(n-B)}{B(n-1)}\Big(\mathbb{E}\big\|\boldsymbol{X}_k\big\|_2^2 + \mathbb{E}\big\|\widetilde{\boldsymbol{X}}^{(s)}\big\|_2^2\Big).$$

Submit the above bound back into (D.1) we have

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{X}_{k+1}\|_2^2] \leq &\bigg(1 - 2\eta b + 2\eta^2 M^2\Big(1 + \frac{n-B}{B(n-1)}\Big)\bigg)\mathbb{E}[\|\boldsymbol{X}_k\|_2^2] \\ &+ \frac{2\eta^2 M^2(n-B)}{B(n-1)}\mathbb{E}\big\|\widetilde{\boldsymbol{X}}^{(s)}\big\|_2^2 + 2\eta a + 2\eta^2 G^2 + \frac{2\eta d}{\gamma}.\end{aligned} \tag{D.5}$$

Note that by assumption we have $\mathbb{E}\big\|\boldsymbol{X}_j\big\|_2^2 \leq C_\psi$ for all $j = 1, \ldots, k$ where $C_\psi = 2\big(1 + 1/b\big)\big(a + G^2 + d/\gamma\big)$, thus (D.5) can be further bounded as:

$$\mathbb{E}[\|\boldsymbol{X}_{k+1}\|_2^2] \leq \underbrace{\bigg(1 - 2\eta b + 2\eta^2 M^2\Big(1 + \frac{2(n-B)}{B(n-1)}\Big)\bigg)}_{C_\lambda} C_\psi + 2\eta a + 2\eta^2 G^2 + \frac{2\eta d}{\gamma}. \tag{D.6}$$

For sufficient small η that satisfies

$$\eta \leq \min\left(1, \frac{b}{2M^2\big(1 + 2(n-B)/(B(n-1))\big)}\right),$$

there are only two cases we need to take into account:
If $C_\lambda \leq 0$, then from (D.6) we have

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{X}_{k+1}\|_2^2] &\leq 2\eta a + 2\eta^2 G^2 + \frac{2\eta d}{\gamma} \\ &\leq 2\bigg(a + G^2 + \frac{d}{\gamma}\bigg).\end{aligned} \tag{D.7}$$

If $0 < C_\lambda \leq 1$, then iterate (D.6) and we have

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{X}_{k+1}\|_2^2] &\leq C_\lambda^{k+1}\|\boldsymbol{X}_0\|_2^2 + \frac{\eta a + \eta^2 G^2 + \frac{\eta d}{\gamma}}{\eta b - \eta^2 M^2\Big(1 + \frac{2(n-B)}{B(n-1)}\Big)} \\ &\leq \frac{2}{b}\bigg(a + G^2 + \frac{d}{\gamma}\bigg).\end{aligned} \tag{D.8}$$

Combining (D.7) and (D.8), we have

$$\mathbb{E}[\|\boldsymbol{X}_{k+1}\|_2^2] \leq 2\bigg(1 + \frac{1}{b}\bigg)\bigg(a + G^2 + \frac{d}{\gamma}\bigg).$$

Thus we show that when $\mathbb{E}[\|\boldsymbol{X}_j\|_2^2], j = 1, \ldots, k$ are bounded, $\mathbb{E}[\|\boldsymbol{X}_{k+1}\|_2^2]$ is also bounded. By mathematical induction we complete the proof. □

D.3 Proof of Lemma C.5

Proof. Since by Algorithm 1 we have $\widetilde{\nabla}_k = (1/B)\sum_{i_k \in I_k}\big(\nabla f_{i_k}(\boldsymbol{X}_k) - \nabla f_{i_k}(\widetilde{\boldsymbol{X}}^{(s)}) + \nabla F(\widetilde{\boldsymbol{X}}^{(s)})\big)$, therefore,

$$\mathbb{E}[\|\widetilde{\nabla}_k - \nabla F(\boldsymbol{X}_k)\|_2^2] = \mathbb{E}\bigg\|\frac{1}{B}\sum_{i_k \in I_k}\big(\nabla f_{i_k}(\boldsymbol{X}_k) - \nabla f_{i_k}(\widetilde{\boldsymbol{X}}^{(s)}) + \nabla F(\widetilde{\boldsymbol{X}}^{(s)}) - \nabla F(\boldsymbol{X}_k)\big)\bigg\|_2^2.$$

Let $\mathbf{v}_i = \nabla F(\mathbf{x}_k) - \nabla F(\widetilde{\mathbf{x}}^{(s)}) - \big(\nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\widetilde{\mathbf{x}}^{(s)})\big)$.

$$\begin{aligned}\mathbb{E}\bigg\|\frac{1}{B}\sum_{i \in I_k}\mathbf{v}_i(\mathbf{x})\bigg\|_2^2 &= \frac{1}{B^2}\mathbb{E}\bigg[\sum_{i \neq i', \{i,i'\} \in I_k}\mathbf{v}_i(\mathbf{x})^\top \mathbf{v}_{i'}(\mathbf{x})\bigg] + \frac{1}{B}\mathbb{E}\|\mathbf{v}_i(\mathbf{x})\|_2^2 \\ &= \frac{B-1}{Bn(n-1)}\mathbb{E}\bigg[\sum_{i \neq i'}\mathbf{v}_i(\mathbf{x})^\top \mathbf{v}_{i'}(\mathbf{x})\bigg] + \frac{1}{B}\mathbb{E}\|\mathbf{v}_i(\mathbf{x})\|_2^2 \\ &= \frac{B-1}{Bn(n-1)}\mathbb{E}\bigg[\sum_{i,i'}\mathbf{v}_i(\mathbf{x})^\top \mathbf{v}_{i'}(\mathbf{x})\bigg] - \frac{B-1}{B(n-1)}\mathbb{E}\|\mathbf{v}_i(\mathbf{x})\|_2^2 + \frac{1}{B}\mathbb{E}\|\mathbf{v}_i(\mathbf{x})\|_2^2 \\ &= \frac{n-B}{B(n-1)}\mathbb{E}\|\mathbf{v}_i(\mathbf{x})\|_2^2,\end{aligned} \tag{D.9}$$

where the last equality is due to the fact that $\frac{1}{n}\sum_{i=1}^n \mathbf{v}_i(\mathbf{x}) = 0$.

Therefore, we have

$$\begin{aligned}\mathbb{E}[\|\widetilde{\nabla}_k - \nabla F(\mathbf{x}_k)\|_2^2] &\leq \frac{n-B}{B(n-1)}\mathbb{E}\|\mathbf{v}_i\|_2^2 \\ &= \frac{n-B}{B(n-1)}\mathbb{E}\|\nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\widetilde{\mathbf{x}}) - \mathbb{E}[\nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\widetilde{\mathbf{x}})]\|_2^2 \\ &\leq \frac{n-B}{B(n-1)}\mathbb{E}\|\nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\widetilde{\mathbf{x}})\|_2^2 \\ &\leq \frac{M^2(n-B)}{B(n-1)}\mathbb{E}\|\mathbf{x}_k - \widetilde{\mathbf{x}}\|_2^2,\end{aligned} \tag{D.10}$$

where the second inequality holds due to the fact that $\mathbb{E}[\|\mathbf{x} - \mathbb{E}[\mathbf{x}]\|_2^2] \leq \mathbb{E}[\|\mathbf{x}\|_2^2]$ and the last inequality follows from Assumption 4.1. This completes the proof. □

D.4 Proof of Lemma C.6

Although the similar proof has been shown in Raginsky et al. (2017), we provide a refined version to make this paper self-contained.

In order to prove Lemma C.6, we need the following three lemmas.

Lemma D.1. (Raginsky et al. (2017)) In terms of the Langevin dynamics (1.1), under Assumption 4.2, we have the following upper bound on the expectation $\mathbb{E}[\|\boldsymbol{X}(t)\|_2^2]$

$$\mathbb{E}\big[\|\boldsymbol{X}(t)\|_2^2\big] \leq e^{-2bt}\|\boldsymbol{X}(0)\|_2^2 + \frac{a+d/\gamma}{b}(1-e^{-2bt}).$$

Lemma D.2. (Bakry et al. (2008)). Suppose that there exists constants $k_0, \lambda_0 > 0$, $R \geq 0$ and a C^2 function $V : \mathbb{R}^d \to [1, \infty)$ such that

$$\mathcal{L}V(\mathbf{w}) \leq -\lambda_0 V(\mathbf{w}) + k_0 \mathbf{1}\{\|\mathbf{w}\|_2 \leq R\},$$

where the operator $\mathcal{L}$ is Itô differential operator. Then the stationary distribution, i.e., π, satisfies a Poincaré inequality with constant

$$c_p \leq \frac{1}{\lambda_0}\bigg(1 + C_p k_0 R^2 e^{Osc_R(g)}\bigg),$$

where $C_p > 0$ is a universal constant and $Osc_R(f) := \max_{\|\mathbf{w}\|_2 \leq R} f(\mathbf{w}) - \min_{\|\mathbf{w}\|_2 \leq R} f(\mathbf{w})$.

Lemma D.3. (Cattiaux et al. (2010)) Suppose the following conditions hold:

1. There exist constants $k, \lambda > 0$ and a C^2 function $V : \mathbb{R}^d \to [1, \infty)$ such that
$$\frac{\mathcal{L}V(\mathbf{w})}{V(\mathbf{w})} \leq k - \lambda\|\mathbf{w}\|_2^2$$
for all $\mathbf{w} \in \mathbb{R}^d$.
2. π satisfies a Poincaré inequality with constant c_p.
3. There exists some constant $K \geq 0$, such that $\nabla^2 f \geq -K\mathbf{I}$.

Let $\widetilde{C}_1$ and $\widetilde{C}_2$ be defined, for some $\epsilon > 0$, by

$$\widetilde{C}_1 = \frac{2}{\lambda}\bigg(\frac{1}{\epsilon} + \frac{K}{2}\bigg) + \epsilon \quad \text{and} \quad \widetilde{C}_2 = \frac{2}{\lambda}\bigg(\frac{1}{\epsilon} + \frac{K}{2}\bigg)\bigg(K + \lambda\int_{\mathbb{R}^d}\|\mathbf{w}\|_2^2 d\pi\bigg).$$

Then π satisfies a logarithmic Sobolev inequality with constant $\Gamma = \widetilde{C}_1 + (\widetilde{C}_2 + 2)c_p$.

Based on the above two lemmas, we are able to complete the proof.

Proof of Lemma C.6. We first give the upper bound of the constant c_p in Poincaré inequality. Following from Lemma D.2, we can establish a Lyapunov function $V(\mathbf{w})$ and then derive a upper bound of c_p. In this proof, we apply the same Lyapunov as Raginsky et al. (2017). Let $V(\mathbf{w}) = e^{-b\gamma\|\mathbf{w}\|_2^2/4}$, and we have

$$\begin{aligned}
\mathcal{L}V(\mathbf{w}) &= -\gamma\langle\nabla V, \nabla F\rangle + \nabla^2 V : \mathbf{I} \\
&= \bigg(-\frac{b\gamma^2}{2}\langle\mathbf{w}, \nabla F\rangle + \frac{b\gamma d}{2} + \frac{(b\gamma)^2}{4}\|\mathbf{w}\|_2^2\bigg)V \\
&\leq \bigg(\frac{b\gamma(d+a\gamma)}{2} - \frac{(b\gamma)^2}{4}\|\mathbf{w}\|_2^2\bigg)V, \qquad \text{(D.11)}
\end{aligned}$$

where the last inequality follows from Assumption 4.2. Thus, let $R^2 = 4(d+a\gamma)/(b\gamma)$, we have

$$\mathcal{L}V(\mathbf{w}) \leq -\frac{b\gamma(d+a\gamma)}{2}V(\mathbf{w}) + \max_{\|\mathbf{w}\|_2 \leq R}\bigg(\frac{b\gamma(d+a\gamma)}{2} - \frac{(b\gamma)^2}{4}\|\mathbf{w}\|_2^2\bigg)V(\mathbf{w})\mathbf{1}\{\|\mathbf{w}\|_2 \leq R\}.$$

Let

$$\lambda_0 = \frac{b\gamma(d+a\gamma)}{2}, \quad \text{and} \quad k_0 = \frac{b\gamma(d+a\gamma)e^{b\gamma R^2/4}}{2},$$

we immediately have

$$\mathcal{L}V(\mathbf{w}) \leq -\lambda_0 V(\mathbf{w}) + k_0 \mathbf{1}\{\|\mathbf{w}\|_2 \leq R\}.$$

Under Assumption 4.1, it follows that

$$F(\mathbf{x}) - F(\mathbf{y}) \leq \langle \nabla F(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle + \frac{M}{2}\|\mathbf{x}-\mathbf{y}\|_2^2,$$

for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. By taking $\mathbf{y} = \mathbf{0}$, we obtain that there exists a constant $K_0 > 0$, such that

$$F(\mathbf{x}) \leq F(0) + \langle \nabla F(0), \mathbf{x}\rangle + \frac{M}{2}\|\mathbf{x}\|_2^2 \leq K_0(1+\|\mathbf{x}\|_2^2), \tag{D.12}$$

where

$$K_0 = \max\left\{F(0) + \frac{1}{2}\|\nabla F(0)\|_2^2, \frac{M+1}{2}\right\}.$$

By (D.12), we have

$$Osc_R(\gamma F) \leq 2\gamma K_0(1+R^2).$$

Thus, based on Lemma D.2, the stationary distribution π satisfies a Poincaré inequality with constant

$$c_p \leq \frac{1}{b\gamma(d+a\gamma)} + \frac{4C_p(d+a\gamma)}{b\gamma}\exp\left(2\gamma K_0 + \frac{(8K_0+b)(d+a\gamma)}{b}\right).$$

Next, we are going to prove the upper bound of constant Γ in logarithmic Sobolev inequality. According to (D.11), we know that

$$\frac{\mathcal{L}V(\mathbf{w})}{V(\mathbf{w})} \leq k - \lambda\|\mathbf{w}\|_2^2$$

holds with

$$k = \frac{b\gamma(d+a\gamma)}{2}, \quad \text{and} \quad \lambda = \frac{(b\gamma)^2}{4}.$$

In addition, for function $f(\mathbf{x}) = \gamma F(\mathbf{x})$, we have $\nabla^2 f \geq -M\gamma\mathbf{I}$ according to Assumption 4.1. Then substitute the above parameters into Lemma D.3, choose $\epsilon = \frac{2}{M}$, we obtain

$$\widetilde{C}_1 = \frac{2b^2+8M^2}{M\gamma b^2}.$$

Moreover, from Lemma D.1, constant $\widetilde{C}_2$ is bounded by

$$\widetilde{C}_2 \leq \frac{6M(d+a\gamma)}{b},$$

Submitting $\widetilde{C}_1$ and $\widetilde{C}_2$ back to Lemma D.3, we have

$$C \leq \frac{2b^2+8M^2}{b^2M\gamma} + c_p\left(\frac{6M(d+a\gamma)}{b}+2\right),$$

note that $c_p = e^{\widetilde{O}(\gamma+d)}$, we also have $\Gamma = e^{\widetilde{O}(\gamma+d)}$, which completes the proof. □

D.5 Proof of Lemma C.7

Proof of Lemma C.7. The proof for Lemma C.7 is quite similar to that for Lemma C.2. Based on the update form of $\boldsymbol{X}_k$ in Algorithm 2, we have

$$\mathbb{E}[\|\boldsymbol{X}_{k+1}\|_2^2] = \mathbb{E}[\|\boldsymbol{X}_k - \eta\widetilde{\nabla}_k + \sqrt{2\eta/\gamma}\boldsymbol{\epsilon}_k\|_2^2] = \mathbb{E}[\|\boldsymbol{X}_k - \eta\widetilde{\nabla}_k\|_2^2] + \frac{2\eta d}{\gamma}.$$

Similar to (D.3), we further have

$$\mathbb{E}[\|\boldsymbol{X}_k - \eta\widetilde{\nabla}_k\|_2^2] = \mathbb{E}[\|\boldsymbol{X}_k - \eta\nabla F(\boldsymbol{X}_k)\|_2^2] + \eta^2\mathbb{E}[\|\nabla F(\boldsymbol{X}_k) - \widetilde{\nabla}_k\|_2^2]. \tag{D.13}$$

Compared with the argument in (D.3), the first term on the R.H.S of the above inequality can be upper bounded in the same way as we did in (D.4), which is stated as follows,

$$\mathbb{E}[\|\boldsymbol{X}_k - \eta\nabla F(\boldsymbol{X}_k)\|_2^2] \leq (1 - 2\eta b + 2\eta^2 M^2)\mathbb{E}[\|\boldsymbol{X}_k\|_2^2] + 2\eta a + 2\eta^2 G^2.$$

Regarding the second term on the R.H.S of (D.13), we have the following based on Lemma C.5

$$\mathbb{E}[\|\nabla F(\boldsymbol{X}_k) - \widetilde{\nabla}_k\|_2^2] \leq \frac{n-B}{B(n-1)}\mathbb{E}[\|\nabla f_{i_k}(\boldsymbol{X}_k) - \widetilde{\boldsymbol{G}}_{i_k}\|_2^2] = \frac{n-B}{B(n-1)}\mathbb{E}[\|\nabla f_{i_k}(\boldsymbol{X}_k) - \nabla f_{i_k}(\boldsymbol{X}_u)\|_2^2],$$

where u is an index satisfying $u < k$. Applying smoothness assumption we have

$$\begin{aligned}\mathbb{E}[\|\nabla F(\boldsymbol{X}_k) - \widetilde{\nabla}_k\|_2^2] &\leq \frac{(n-B)M^2}{B(n-1)}\mathbb{E}[\|\boldsymbol{X}_k - \boldsymbol{X}_u\|_2^2] \\ &\leq 2\big(\mathbb{E}[\|\boldsymbol{X}_k\|_2^2] + \mathbb{E}[\|\boldsymbol{X}_u\|_2^2]\big),\end{aligned}$$

where the second inequality follows from Young's inequality and the fact that $B \geq 1$. Now, we are able to upper bound $\mathbb{E}[\|\boldsymbol{X}_{k+1}\|_2^2]$ as follows

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{X}_{k+1}\|_2^2] &\leq (1 - 2\eta b + 2\eta^2 M^2)\mathbb{E}[\|\boldsymbol{X}_k\|_2^2] + 2\eta a + 2\eta^2 G^2 + 2\eta^2(\mathbb{E}[\|\boldsymbol{X}_k\|_2^2] + \mathbb{E}[\|\boldsymbol{X}_u\|_2^2]) + \frac{2\eta d}{\gamma} \\ &\leq \big(1 - 2\eta b + 2\eta^2(M^2+4)\big)\max\{\mathbb{E}[\|\boldsymbol{X}_k\|_2^2], \mathbb{E}[\|\boldsymbol{X}_u\|_2^2]\} + 2\eta(a + d/\gamma) + 2\eta^2 G^2\end{aligned}$$

Then we apply induction to prove that $\mathbb{E}[\|\boldsymbol{X}_k\|_2^2] \leq 2(1+1/b)(a+G^2+d/\gamma)$. It is easy to verify that $\mathbb{E}[\|\boldsymbol{X}_0\|_2^2] = 0$ satisfies the argument. Then we assume that the argument holds for all iterates from 0 to k. Note that $u < k$, which implies that

$$\max\{\mathbb{E}[\|\boldsymbol{X}_k\|_2^2], \mathbb{E}[\|\boldsymbol{X}_u\|_2^2]\} \leq 2\bigg(1 + \frac{1}{b}\bigg)\bigg(a + G^2 + \frac{d}{\gamma}\bigg).$$

Then, for sufficiently small η such that

$$\eta \leq \min\bigg\{1, \frac{b}{2(M^2+4)}\bigg\},$$

it follows that

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{X}_{k+1}\|_2^2] &\leq 2\bigg[(1 - \eta b)\bigg(1 + \frac{1}{b}\bigg) + \eta\bigg]\bigg(a + G^2 + \frac{d}{\gamma}\bigg) \\ &\leq 2\bigg(1 + \frac{1}{b} - \eta b\bigg)\bigg(a + G^2 + \frac{d}{\gamma}\bigg) \\ &\leq 2\bigg(1 + \frac{1}{b}\bigg)\bigg(a + G^2 + \frac{d}{\gamma}\bigg),\end{aligned}$$

which indicates that $\mathbb{E}[\|\boldsymbol{X}_{k+1}\|_2^2]$ also satisfies the argument. Thus we are able to complete the proof. □

D.6 Proof of Lemma C.8

Proof. Since by Algorithm 2 we have $\widetilde{\nabla}_k = \frac{1}{B}\sum_{i_k \in I_k}\big(\nabla f_{i_k}(\boldsymbol{X}_k) - \widetilde{\boldsymbol{G}}_{i_k} + \widetilde{\boldsymbol{g}}_k\big)$, therefore,

$$\mathbb{E}[\|\widetilde{\nabla}_k - \nabla F(\boldsymbol{X}_k)\|_2^2] = \mathbb{E}\bigg\|\frac{1}{B}\sum_{i_k \in I_k}\big(\nabla f_{i_k}(\boldsymbol{X}_k) - \widetilde{\mathbf{G}}_{i_k} + \widetilde{\mathbf{g}}_k - \nabla F(\boldsymbol{X}_k)\big)\bigg\|_2^2.$$

Let $\mathbf{v}_i = \nabla f_{i_k}(\boldsymbol{X}_k) - \widetilde{\mathbf{G}}_{i_k} + \widetilde{\mathbf{g}}_k - \nabla F(\boldsymbol{X}_k)$, following the same procedure in (D.9) we have

$$\mathbb{E}\bigg\|\frac{1}{B}\sum_{i \in I_k}\mathbf{v}_i(\mathbf{x})\bigg\|_2^2 = \frac{n-B}{B(n-1)}\mathbb{E}\|\mathbf{v}_i(\mathbf{x})\|_2^2.$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}[\|\widetilde{\nabla}_k - \nabla F(\mathbf{x}_k)\|_2^2] &= \frac{n-B}{B(n-1)}\mathbb{E}\|\mathbf{v}_i\|_2^2 \\
&= \frac{n-B}{B(n-1)}\mathbb{E}\|\nabla f_{i_k}(\boldsymbol{X}_k) - \widetilde{\mathbf{G}}_{i_k} - (\nabla F(\boldsymbol{X}_k) - \widetilde{\mathbf{g}}_k)\|_2^2 \\
&\leq \frac{n-B}{B(n-1)}\mathbb{E}\|\nabla f_{i_k}(\boldsymbol{X}_k) - \widetilde{\mathbf{G}}_{i_k}\|_2^2,
\end{aligned}$$

where the inequality holds due to the fact that $\mathbb{E}[\|\mathbf{x} - \mathbb{E}[\mathbf{x}]\|_2^2] \leq \mathbb{E}[\|\mathbf{x}\|_2^2]$, which completes the proof. □