Stochastic Gradient Hamiltonian Monte Carlo Methods with Recursive Variance Reduction

Difan Zou
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
knowzou@cs.ucla.edu

Pan Xu
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
panxu@cs.ucla.edu

Quanquan Gu
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
qgu@cs.ucla.edu

Abstract

Stochastic Gradient Hamiltonian Monte Carlo (SGHMC) algorithms have received increasing attention in both theory and practice. In this paper, we propose a Stochastic Recursive Variance-Reduced gradient HMC (SRVR-HMC) algorithm. It makes use of a semi-stochastic gradient estimator that recursively accumulates the gradient information to reduce the variance of the stochastic gradient. We provide a convergence analysis of SRVR-HMC for sampling from a class of non-log-concave distributions and show that SRVR-HMC converges faster than all existing HMC-type algorithms based on underdamped Langevin dynamics. Thorough experiments on synthetic and real-world datasets validate our theory and demonstrate the superiority of SRVR-HMC.

1 Introduction

Monte Carlo Markov Chain (MCMC) has been widely used in Bayesian learning [1] as a powerful tool for posterior sampling, inference and decision making. More recently, Hamiltonian MCMC approaches based on the Hamiltonian Langevin dynamics [24, 43] have received extensive attention in both theory and practice [16, 5, 40, 14, 6, 18, 55, 28] due to their widespread empirical successes. Hamiltonian Langevin dynamics (a.k.a., underdamped Langevin dynamics) [19] is described by the following stochastic differential equation:

$$\begin{aligned} \mathrm{d}\boldsymbol{V}_t &= -\gamma \boldsymbol{V}_t \mathrm{d}t - u\nabla f(\boldsymbol{X}_t)\mathrm{d}t + \sqrt{2\gamma u}\mathrm{d}\boldsymbol{B}_t, \\ \mathrm{d}\boldsymbol{X}_t &= \boldsymbol{V}_t \mathrm{d}t, \end{aligned} \tag{1.1}$$

where $\gamma > 0$ is called the friction parameter, $u > 0$ is the inverse mass, $\boldsymbol{X}_t, \boldsymbol{V}_t \in \mathbb{R}^d$ are the position and velocity variables of the continuous-time dynamics respectively, and $\boldsymbol{B}_t \in \mathbb{R}^d$ is the standard Brownian motion. Under mild assumptions on the function $f(\mathbf{x})$, the Markov process $(\boldsymbol{X}_t, \boldsymbol{V}_t)$ has a unique stationary distribution which is proportional to $\exp\{-f(\mathbf{x}) - \|\mathbf{v}\|_2^2/(2u)\}$ and the marginal distribution of $\boldsymbol{X}_t$ converges to a stationary distribution $\pi \propto \exp\{-f(\mathbf{x})\}$. Hence, we can apply numerical integrators to discretize the continuous-time dynamics (1.1) in order to sample from the target distribution π. Direct Euler-Maruyama discretization [34] of (1.1) gives rise to

$$\begin{aligned} \mathbf{v}_{k+1} &= \mathbf{v}_k - \gamma\eta\mathbf{v}_k - \eta u\nabla f(\mathbf{x}_k) + \sqrt{2\gamma u\eta}\epsilon_k, \\ \mathbf{x}_{k+1} &= \mathbf{x}_k + \eta\mathbf{v}_k, \end{aligned} \tag{1.2}$$

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

which is known as underdamped Langevin MCMC (UL-MCMC) and can also be viewed as a type of Hamiltonian Monte Carlo (HMC) methods [43, 6]. Cheng et al. [18] studied a modified version of UL-MCMC in (1.2) and proved its convergence rate to the stationary distribution in 2-Wasserstein distance for sampling from strongly log-concave densities. When the target distribution is non-log-concave but admits certain good properties, the convergence guarantees of UL-MCMC in Wasserstein metric have also been established in [27, 17, 8, 30].

In practice, $f(\mathbf{x})$ in (1.2) can be chosen as the negative log-likelihood function on the training data:

$$f(\mathbf{x}) = n^{-1} \sum_{i=1}^{n} f_i(\mathbf{x}), \tag{1.3}$$

where n is the size of training data and $f_i(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$ is the negative log-likelihood function on the i-th data point. For a large dataset, it can be extremely inefficient to compute the full gradient $\nabla f(\mathbf{x})$ which consists of gradients $\nabla f_i(\mathbf{x})$'s for all data points. To alleviate this computational burden, stochastic gradient Hamiltonian Monte Carlo (SGHMC) methods [16, 40] and stochastic gradient UL-MCMC (SG-UL-MCMC) [18] were proposed, which replace the full gradient in (1.2) with a mini-batch stochastic gradient. While SGHMC is much more efficient than HMC methods, it comes at the cost of a slower mixing rate due to the large variance caused by stochastic gradients [5, 6, 23]. To resolve this dilemma, Zou et al. [55], Li et al. [37] proposed stochastic variance-reduced gradient HMC methods using variance reduction techniques [33, 36] and proved that variance reduction can accelerate the convergence of both HMC and SGHMC for sampling and Bayesian inference. For sampling from a class of non-log-concave densities, Gao et al. [30] showed that SGHMC converges to the stationary distribution of (1.1) up to an ϵ-error in 2-Wasserstein distance with $\widetilde{O}(\epsilon^{-8}\mu_*^{-5})$[1] gradient complexity[2], where μ_* is a lower bound of the spectral gap of the Markov process generated by (1.1) and is in the order of $\exp(-\widetilde{O}(d))$ in the worst case [27]. This gradient complexity of SGHMC is very high even for a moderate sampling error ϵ.

In this paper, we aim to reduce the gradient complexity of SGHMC for sampling from non-log-concave densities. The fundamental challenge in speeding up HMC-type methods lies in the control of the discretization error between the Hamiltonian Langevin dynamics (1.1) and discrete algorithms. We propose a novel algorithm, namely stochastic recursive variance-reduced gradient HMC (SRVR-HMC), which employs a recursively updated semi-stochastic gradient estimator to reduce the variance of stochastic gradient and improve the discretization error. Note that such a recursively updated semi-stochastic gradient estimator was originally proposed in [44, 29] for finding stationary points in stochastic nonconvex optimization. Nevertheless, our analysis is fundamentally different from that in [44, 29] since their goal is just to find a stationary point of $f(\mathbf{x})$, while we aim to sample from the target distribution $\pi \propto \exp(-f(\mathbf{x}))$ that concentrates on the global minimizer of $f(\mathbf{x})$, which is substantially more challenging.

1.1 Our contributions

We summarize our major contributions as follows.

- We propose a new HMC algorithm called SRVR-HMC for approximate sampling, which is built on a recursively updated semi-stochastic gradient estimator that significantly decreases the discretization error and speeds up the sampling process.
- We establish the convergence guarantee of SRVR-HMC for sampling from non-log-concave densities satisfying certain dissipativeness condition. Specifically, we show that its gradient complexity for achieving ϵ-error in 2-Wasserstein distance is $\widetilde{O}((n + \epsilon^{-2}n^{1/2}\mu_*^{-3/2}) \wedge \epsilon^{-4}\mu_*^{-2})$. Remarkably, the convergence guarantee of SRVR-HMC is better than the $\widetilde{O}(\epsilon^{-4}\mu_*^{-3}n)$ gradient complexity of HMC [30] by a factor of at least $\widetilde{O}(\epsilon^{-2}\mu_*^{-3/2}n^{1/2})$, and better than the $\widetilde{O}(\epsilon^{-8}\mu_*^{-5})$ gradient complexity of SGHMC [30] by a factor of at least $\widetilde{O}(\epsilon^{-4}\mu_*^{-3})$.
- With a proper choice of parameters, our algorithm can reduce to UL-MCMC [18] and SG-UL-MCMC [18], which are originally proposed for sampling from strongly-log-concave distributions.

[1] $\widetilde{O}(\cdot)$ hides constant and logarithm factors.

[2] Gradient complexity is the total number of stochastic gradients $\nabla f_i(\mathbf{x})$ an algorithm needs to compute in order to achieve ϵ-error in terms of certain measurement.

Our theoretical analysis shows that these two algorithms can be used for sampling from non-log-concave distributions as well, and they enjoy lower gradient complexities than HMC and SGHMC [30], which is of independent interest.

- We compare our algorithm with many state-of-the-art baselines through experiments on sampling from Gaussian mixture distributions, independent component analysis (ICA) and Bayesian logistic regression, which further validates the superiority of our algorithm.

1.2 Additional related work

There is also a vast literature of MCMC methods based on the overdamped Langevin dynamics [35]:

$$\mathrm{d}\boldsymbol{X}_t = -\nabla f(\boldsymbol{X}_t)\mathrm{d}t + \sqrt{2\beta}\mathrm{d}\boldsymbol{B}_t, \tag{1.4}$$

where $\beta > 0$ is the temperature parameter and $\boldsymbol{B}_t$ is Brownian motion. The convergence analysis of Langevin based algorithms dates back to [46]. Mattingly et al. [41] established convergence rates for a class of discrete approximation of Langevin dynamics. When the target distribution is smooth and strongly log-concave, the convergence of Langevin Monte Carlo (LMC) based on the discretization of (1.4) has been widely studied in terms of both total variation (TV) distance [21, 26] and 2-Wasserstein distance [22, 20]. Welling and Teh [50] proposed the stochastic gradient Langevin dynamics (SGLD) algorithm to avoid full gradient computation. Teh et al. [47] proposed to apply decreasing step size with SGLD and proved its convergence in terms of mean square error (MSE). Vollmer et al. [48] characterized the bias of SGLD and further proposed a modified SGLD algorithm that removes the bias. [10] establish a link between LMC, SGLD, SGLDFP (a variant of SGLD) and SGD, which shows that the stationary distribution of LMC and SGLDFP can be closer to the target density π as the sample size increases, while the dynamics of SGLD is more similar to that of SGD. Barkhagen et al. [4], Chau et al. [13] studied the convergence of SGLD when the training data in (1.3) are dependent. In order to reduce the variance of SGLD, SVRG-LD and SAGA-LD have been proposed by Dubey et al. [25] and their convergence have been studied in terms of MSE [25, 15] and 2-Wasserstein distance [56, 12]. Baker et al. [2] proposed to use control variate in SGLD which can also reduce the variance and improve the convergence rate. Mou et al. [42] studied the generalization performance of SGLD from both stability and PAC-Bayesian perspectives. For nonconvex optimization, Raginsky et al. [45] proved the non-asymptotic convergence rate of SGLD and Zhang et al. [52] analyzed the hitting time of SGLD to local minima. Xu et al. [51] further studied the global convergence of a class of Langevin dynamics based algorithms.

Table 1: Gradient complexity of different methods to achieve ϵ-error in 2-Wasserstein distance for sampling from non-log-concave densities.

Methods	Gradient Complexity	
LMC	$\widetilde{O}\big(\epsilon^{-4}\lambda_*^{-5}n\big)$	[45]
SGLD	$\widetilde{O}\big(\epsilon^{-8}\lambda_*^{-9}\big)$	[45]
SVRG-LD	$\widetilde{O}\big(n+\epsilon^{-2}\lambda_*^{-4}n^{3/4}+\epsilon^{-4}\lambda_*^{-4}n^{1/2}\big)$	[57]
HMC	$\widetilde{O}\big(\epsilon^{-4}\mu_*^{-3}n\big)$	[30]
UL-MCMC	$\widetilde{O}\big(\epsilon^{-2}\mu_*^{-3/2}n\big)$	▷ Corollary 3.9
SGHMC	$\widetilde{O}\big(\epsilon^{-8}\mu_*^{-5}\big)$	[30]
SG-UL-MCMC	$\widetilde{O}\big(\epsilon^{-6}\mu_*^{-5/2}\big)$	▷ Corollary 3.9
SRVR-HMC	$\widetilde{O}\big((n+\epsilon^{-2}n^{1/2}\mu_*^{-3/2})\wedge\epsilon^{-4}\mu_*^{-2}\big)$	▷ Corollary 3.5

In Table 1, we compare the gradient complexity of different methods to achieve ϵ-error in 2-Wasserstein distance for sampling from non-log-concave densities[3]. LMC, SGLD and SVRG-LD are based on overdamped Langevin dynamics (1.4) and HMC, UL-MCMC, SGHMC, SG-UL-MCMC and SRVR-HMC are based on underdamped Langevin dynamics (1.1). The HMC/SGHMC algorithm studied in [30] and the UL-MCMC/SG-UL-MCMC algorithm [18] analyzed in this paper are

[3]The original results for LMC/SGLD in [45] and for HMC/SG-HMC in [30] are about the global convergence in nonconvex optimization. Yet their results can be adapted to sampling from non-log-concave distributions, and the corresponding gradient complexities can be spelled out from their convergence rates.

slightly different since they rely on different discretization methods to the Hamiltonian Langevin dynamics (1.1). In addition, note that λ_* denotes the spectral gap of the Markov process generated by overdamped Langevin dynamics (1.4), which is also in the order of $\exp(-\widetilde{O}(d))$ [9, 45] in the worst case.

From Table 1, we can see that the proposed SRVR-HMC algorithm strictly outperforms HMC, UL-MCMC, SGHMC and SG-UL-MCMC, and also outperforms LMC, SGLD and SVRG-LD in terms of the dependency on target accuracy ϵ and training sample size n. We remark that for a general non-log-concave target density, λ_* and μ_* are not directly comparable, though both of them are exponential in dimension d. However, it is shown that for a class of target densities, μ_* can be in the order of $O(\lambda_*^{1/2})$ [27, 30], which suggests that SRVR-HMC is also strictly better than LMC, SGLD and SVRG-LD for sampling from such densities.

Notation. We denote discrete update by lower case bold symbol $\mathbf{x}_k$ and continuous-time dynamics by upper case italicized bold symbol $\boldsymbol{X}_t$. For a vector $\mathbf{x} \in \mathbb{R}^d$, we denote by $\|\mathbf{x}\|_2$ the Euclidean norm. For random vectors $\mathbf{x}_k, \boldsymbol{X}_t \in \mathbb{R}^d$, we denote their probability distribution functions by $\mathbb{P}(\mathbf{x}_k)$ and $\mathbb{P}(\boldsymbol{X}_t)$ respectively. For a probability measure μ, we denote by $\mathbb{E}_\mu[\boldsymbol{X}]$ the expectation of $\boldsymbol{X}$ under probability measure u. The 2-Wasserstein distance between two probability measures u and v is

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

where the infimum is taken over all joint distributions ζ with u and v being its marginal distributions. $\mathbb{1}(\cdot)$ denotes the indicator function. We denote index set $[n] = \{1, 2, \ldots, n\}$ for an integer n. We use $a_n = O(b_n)$ to denote that $a_n \leq C b_n$ for some constant $C > 0$ independent of n, and use $a_n = \widetilde{O}(b_n)$ to hide the logarithmic factors in b_n. The Vinogradov notation $a_n \lesssim b_n$ is also used synonymously with $a_n = O(b_n)$. We denote $\min\{a, b\}$ and $\max\{a, b\}$ by $a \wedge b$ and $a \vee b$ respectively. The ceiling function $\lceil x \rceil$ outputs the least integer greater than or equal to x.

2 The proposed algorithm

In this section, we present our algorithm, SRVR-HMC, for sampling from a target distribution in the form of $\pi \propto \exp\{-f(\mathbf{x})\}$. Our algorithm is shown in Algorithm 1, which has a multi-epoch structure. In detail, there are $\lceil K/L \rceil$ epochs, where K is the number of total iterations and L denotes the epoch length, i.e., the number of iterations within each inner loop.

Recall that the update rule of HMC in (1.2) requires the computation of full gradient $\nabla f(\mathbf{x}_k)$ at each iteration, which is the average of n stochastic gradients. This causes a high per-iteration complexity when n is large. Therefore, we propose to leverage the stochastic gradient to offset the computational burden. At the beginning of the j-th epoch, we compute a stochastic gradient $\widetilde{\mathbf{g}}_j$ based on a batch of training data (uniformly sampled from $[n]$ without replacement) as shown in Line 4 of Algorithm 1, where the batch is denoted by $\widetilde{\mathcal{B}}_j$ with batch size $|\widetilde{\mathcal{B}}_j| = B_0$. In each epoch, we make use of the stochastic path-integrated differential estimator [29] to compute the following semi-stochastic gradient

$$\mathbf{g}_k = 1/B \textstyle\sum_{i \in \mathcal{B}_k} \big[\nabla f_i(\mathbf{x}_k) - \nabla f_i(\mathbf{x}_{k-1})\big] + \mathbf{g}_{k-1}, \tag{2.1}$$

where $\mathcal{B}_k$ is another uniformly sampled (without replacement) mini-batch from $[n]$ with mini-batch size $|\mathcal{B}_k| = B$. Unlike the unbiased stochastic gradient estimators in SGHMC [16] and SVR-HMC [55], $\mathbf{g}_k$ is a biased estimator of the full gradient $\nabla f(\mathbf{x}_k)$ conditioned on $\mathbf{x}_k$. However, we can show that while being biased, the variance of $\mathbf{g}_k$ is substantially smaller than that of unbiased ones. This is the key reason why our algorithm can achieve a faster convergence rate than existing HMC-type algorithms. Based on the semi-stochastic gradient in (2.1), we update the position and velocity variables as follows

$$\begin{aligned} \mathbf{v}_{k+1} &= \mathbf{v}_k e^{-\gamma\eta} - u\gamma^{-1}(1 - e^{-\gamma\eta})\mathbf{g}_k + \boldsymbol{\epsilon}_k^v, \\ \mathbf{x}_{k+1} &= \mathbf{x}_k + \gamma^{-1}(1 - e^{-\gamma\eta})\mathbf{v}_k + u\gamma^{-2}(\gamma\eta + e^{-\gamma\eta} - 1)\mathbf{g}_k + \boldsymbol{\epsilon}_k^x, \end{aligned} \tag{2.2}$$

where η is the step size and u, γ are the inverse mass and friction parameter defined in (1.1), which are usually treated as tunable hyper parameters in practice. Moreover, $\boldsymbol{\epsilon}_k^v, \boldsymbol{\epsilon}_k^x \in \mathbb{R}^d$ are zero mean

Algorithm 1 Stochastic Recursive Variance-Reduced gradient HMC (SRVR-HMC)

input: Initial points $\widetilde{\mathbf{x}}_0 = \mathbf{x}_0 = \mathbf{x}_0, \mathbf{v}_0$; step size η; batch sizes B_0 and B; total number of iterations K; epoch length L
for $j = 0, \ldots, \lceil K/L \rceil$ **do**
  Uniformly sample a subset of index $\widetilde{\mathcal{B}}_j \subset [n]$ with $|\widetilde{\mathcal{B}}_j| = B_0$
  Compute $\widetilde{\mathbf{g}}_j = 1/B_0 \sum_{i \in \widetilde{\mathcal{B}}_j} \nabla f_i(\widetilde{\mathbf{x}}_j)$
  for $l = 0, \ldots, L-1$ **do**
    $k = jL + l$
    if $l = 0$ **then**
      $\mathbf{g}_k = \widetilde{\mathbf{g}}_j$
    else
      Uniformly sample a subset of index $\mathcal{B}_k \subset [n]$ with $|\mathcal{B}_k| = B$
      Compute $\mathbf{g}_k = 1/B \sum_{i \in \mathcal{B}_k} (\nabla f_i(\mathbf{x}_k) - \nabla f_i(\mathbf{x}_{k-1})) + \mathbf{g}_{k-1}$
    end if
    $\mathbf{x}_{k+1} = \mathbf{x}_k + \gamma(1 - e^{-\gamma\eta})\mathbf{v}_k + u\gamma^{-2}(\gamma\eta + e^{-\gamma\eta} - 1)\mathbf{g}_k + \boldsymbol{\epsilon}_k^x$
    $\mathbf{v}_{k+1} = \mathbf{v}_k e^{-\gamma\eta} - u\gamma^{-1}(1 - e^{-\gamma\eta})\mathbf{g}_k + \boldsymbol{\epsilon}_k^v$
  end for
  $\widetilde{\mathbf{x}}_{j+1} = \mathbf{x}_{(j+1)L}$
end for
output: $\mathbf{x}_K$

Gaussian random vectors with covariance matrices satisfying

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\epsilon}_k^v(\boldsymbol{\epsilon}_k^v)^\top] &= u(1 - e^{-2\gamma\eta}) \cdot \mathbf{I}, \\
\mathbb{E}[\boldsymbol{\epsilon}_k^x(\boldsymbol{\epsilon}_k^x)^\top] &= u\gamma^{-2}(2\gamma\eta + 4e^{-\gamma\eta} - e^{-2\gamma\eta} - 3) \cdot \mathbf{I}, \\
\mathbb{E}[\boldsymbol{\epsilon}_k^v(\boldsymbol{\epsilon}_k^x)^\top] &= u\gamma^{-1}(1 - 2e^{-\gamma\eta} + e^{-2\gamma\eta}) \cdot \mathbf{I},
\end{aligned} \tag{2.3}$$

where $\mathbf{I} \in \mathbb{R}^{d \times d}$ is the identity matrix. The covariance of the Gaussian noises in (2.3) is obtained by integrating the Hamiltonian Langevin dynamics (1.1) over a time period of length η. It is worth noting our update rule in (2.2) and the construction of the Gaussian noises in (2.3) follow Cheng et al. [18], Zou et al. [55], Cheng et al. [17], except that we use a different semi-stochastic gradient estimator as shown in (2.1). In contrast, Cheng et al. [18] uses full gradient and noisy gradient, and Zou et al. [55] uses an unbiased semi-stochastic gradient based on SVRG [33].

We remark here that the semi-stochastic gradient estimator in (2.1) was originally proposed in finding stationary points in finite-sum optimization [44, 29] and further extended in [49, 32]. In addition, another semi-stochastic gradient estimator called SNVRG [54, 53] has also been demonstrated to achieve similar convergence rate in finite-sum optimization. Despite using the same semi-stochastic gradient estimator, our work differs from [44, 29] in at least two aspects: (1) the sampling problem studied in this paper is different from the optimization problem studied in [44, 29], where our goal is to sample from a target distribution concentrating on the global minimizer of $f(\mathbf{x})$ such that the sample distribution is close to the target distribution in 2-Wasserstein distance. In contrast, Nguyen et al. [44], Fang et al. [29] aim at finding a stationary point of $f(\mathbf{x})$ with small gradient; and (2) the algorithms in [44, 29] only have one update variable, while our SRVR-HMC algorithm has an additional Hamiltonian momentum term and therefore has two update variables (i.e., velocity and position variables). The Hamiltonian momentum is essential for underdamped Langevin Monte Carlo methods to achieve a smaller discretization error than overdamped methods such as SGLD [50] and SVRG-LD [25]. At the same time, this also introduces a great technical challenge in our theoretical analysis and requires nontrivial efforts.

3 Main theory

In this section, we provide the convergence guarantee for Algorithm 1. In particular, we characterize the 2-Wasserstein distance between the distribution of the output of Algorithm 1 and the target distribution $\pi \propto e^{-f(\mathbf{x})}$. We focus on sampling from non-log-concave densities that satisfy the smoothness and dissipativeness conditions, which are formally defined as follows.

Assumption 3.1 (Smoothness)**.** Each f_i in (1.3) is M-smooth, i.e., there exists a positive constant $M > 0$, such that the following holds

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|_2 \leq M\|\mathbf{x} - \mathbf{y}\|_2, \quad \text{for any } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

Note that Assumption 3.1 directly implies that function $f(\mathbf{x})$ is also M-smooth.

Assumption 3.2 (Dissipativeness)**.** There exist constants $m, b > 0$, such that the following holds

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

Different from the smoothness assumption, Assumption 3.2 is only required for $f(\mathbf{x})$ rather than $f_i(\mathbf{x})$. The dissipativeness assumption is standard in the analysis for sampling from non-log-concave densities and is essential to guarantee the convergence of underdamped Langevin dynamics [46, 41].

3.1 Convergence analysis of the proposed algorithm

Now we state our main theorem that establishes the convergence rate of Algorithm 1.

Theorem 3.3. Suppose Assumptions 3.1 and 3.2 hold and the initial points are $\mathbf{x}_0 = \mathbf{v}_0 = \mathbf{0}$. If set $\gamma \leq 2\sqrt{Mu}$ and the step size $\eta \leq O(mM^{-3} \wedge m^{1/2}M^{-3/2}L^{-1/2})$, the output $\mathbf{x}_K$ of Algorithm 1 satisfies

$$\mathcal{W}_2\big(\mathbb{P}(\mathbf{x}_K), \pi\big) \leq \Gamma_1 \bigg(\bigg(1 + \frac{L}{B}\bigg) K\eta^3 + \frac{K\eta}{\gamma^2 B_0} \cdot \mathbb{1}(B_0 < n) \bigg)^{1/4} + \Gamma_0 e^{-\mu_* K\eta},$$

where B_0, B are the batch and minibatch sizes, L is the epoch length and $\mu_* = \exp(-\widetilde{O}(d))$ is a lower bound of the spectral gap of the Markov process generated by (1.1). $\Gamma_0 = \widetilde{O}(\mu_*^{-1})$ and $\Gamma_1 = 2D_1(M^2\gamma^3 u D_2)^{1/4}$ are problem-dependent parameters with constants D_1, D_2 defined as

$$D_1 = \frac{8}{\gamma}\sqrt{\frac{um(f(\mathbf{0}) - f(\mathbf{x}^*)) + 2Mu(4d + 2b + m\|\mathbf{x}^*\|_2^2\gamma^2) + (12um + 3\gamma^2)}{m}},$$
$$D_2 = \frac{8um(f(\mathbf{0}) - f(\mathbf{x}^*)) + 8Mu\big(20(d+b) + m\|\mathbf{x}^*\|_2^2\big)}{\gamma^2 m} + \max_{i\in[n]} \frac{\|\nabla f_i(\mathbf{0})\|_2^2}{M^2},$$

and $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$ is the global minimizer of f.

Theorem 3.3 states that the 2-Wasserstein distance between the output of SRVR-HMC and the target distribution is upper bounded by two terms: the first term is the discretization error between the discrete-time Algorithm 1 and the continuous-time dynamics (1.1), which goes to zero when the step size η goes to zero; the second term represents the ergodicity of the Markov process generated by (1.1) which converges to zero exponentially fast.

Remark 3.4. The result in Theorem 3.3 encloses a term μ_* with an exponential dependence on the dimension d, which is a lower bound of the spectral of the Markov process generated by (1.1). When f is nonconvex, the exponential dependence of μ_* on dimension is unavoidable under the dissipativeness assumption [9]. However, this exponential dependency on d can be weakened by imposing stronger assumptions on $f(\mathbf{x})$. For instance, Eberle et al. [27], Gao et al. [30] showed that for a symmetric double-well potential $f(\mathbf{x})$, μ_* is in the order of $\Omega(1/a)$, where a is the distance between these two wells, and is typically polynomial in the dimension d. Another example is shown by Cheng et al. [17]: when $f(\mathbf{x})$ is strongly convex outside a ℓ_2 ball centered at the origin with radius R, μ_* is in the order of $\exp(-O(MR^2))$ where M is the smoothness parameter.

From Theorem 3.3, we can obtain the gradient complexity of SRVR-HMC by optimizing the choice of minibatch size B and batch size B_0 in the following corollary.

Corollary 3.5. Under the same assumptions in Theorem 3.3, if set $B_0 = \widetilde{O}(\epsilon^{-4}\mu_*^{-1} \wedge n)$, $B \lesssim B_0^{1/2}$, $L = O(B_0/B)$, and $\eta = \widetilde{O}(\epsilon^2 B_0^{-1/2}\mu_*^{1/2}B)$, then Algorithm 1 requires $\widetilde{O}((n + \epsilon^{-2}n^{1/2}\mu_*^{-3/2}) \wedge \epsilon^{-4}\mu_*^{-2})$ stochastic gradient evaluations to achieve ϵ-error in 2-Wasserstein distance.

Remark 3.6. Recall the gradient complexities of HMC and SGHMC in Table 1, it is evident that the gradient complexity of Algorithm 1 is lower than that of HMC [30] by a factor of $\widetilde{O}(\epsilon^{-2}n^{1/2}\mu_*^{3/2} \vee n\mu_*)$ and is lower than that of SGHMC [30] by a factor of $\widetilde{O}(\epsilon^{-6}n^{-1/2}\mu_*^{-7/2} \vee \epsilon^{-4}\mu_*^{-3})$.

Remark 3.7. As shown in Table 1, the gradient complexities of overdamped Langevin dynamics based algorithms, including LMC, SGLD and SVRG-LD, depend on the spectral gap λ_* of the Markov chain generated by (1.4). Although the magnitudes of μ_* and λ_* are not directly comparable, they are generally in the same order in the worst case [9, 45, 27]. Thus we treat them the same in the following comparison. In specific, the gradient complexity of SRVR-HMC is better than those of LMC [45] SGLD [45] and SVRG-LD [57] by factors of $\widetilde{O}(\epsilon^{-2}n^{1/2} \vee n)$, $\widetilde{O}(\epsilon^{-6}n^{-1/2} \vee \epsilon^{-4})$ and $\widetilde{O}(\epsilon^{-2} \vee n^{1/2})$ respectively.

3.2 Implication for UL-MCMC and SG-UL-MCMC

Recall the proposed SRVR-HMC algorithm in Algorithm 1, if we set the epoch length to be $L = 1$, Algorithm 1 degenerates to SG-UL-MCMC [18], with the following update formulation:

$$\begin{aligned} \mathbf{v}_{k+1} &= \mathbf{v}_k e^{-\gamma\eta} - u\gamma^{-1}(1 - e^{-\gamma\eta})\widetilde{\mathbf{g}}_k + \boldsymbol{\epsilon}_k^v, \\ \mathbf{x}_{k+1} &= \mathbf{x}_k + \gamma^{-1}(1 - e^{-\gamma\eta}\mathbf{v}_k) + u\gamma^{-2}(\gamma\eta + e^{-\gamma\eta} - 1)\widetilde{\mathbf{g}}_k + \boldsymbol{\epsilon}_k^x, \end{aligned} \tag{3.1}$$

where $\widetilde{\mathbf{g}}_k = |\widetilde{\mathcal{B}}_k|^{-1}\sum_{i=1}^n \nabla f_i(\mathbf{x}_k)$ denotes the stochastic gradient computed in the k-th iteration. In addition, if we replace $\widetilde{\mathbf{g}}_k$ with the full gradient $\nabla f(\mathbf{x}_k)$, SG-UL-MCMC in (3.1) further reduces to UL-MCMC [18]. Although these two algorithms were originally proposed for sampling from strongly-log-concave densities [18], in this subsection, we show that our analysis of SRVR-HMC can be easily adapted to derive the gradient complexity of UL-MCMC/SG-UL-MCMC for sampling from non-log-concave densities. We first state the convergence of SG-UL-MCMC in the following theorem.

Theorem 3.8. Under the same assumptions in Theorem 3.3, the output $\mathbf{x}_K$ of the SG-UL-MCMC algorithm in (3.1) satisfies

$$\mathcal{W}_2\big(\mathbb{P}(\mathbf{x}_K), \pi\big) \leq \Gamma_1\big[2K\eta^3 + K\eta/(\gamma^2 B_0)\cdot \mathbb{1}(B_0 < n)\big]^{1/4} + \Gamma_0 e^{-\mu_* K\eta},$$

where B_0 denotes the mini-batch size, μ_*, Γ_0 and Γ_1 are defined in Theorem 3.3.

Similar to the results in Theorem 3.3, the sampling error of SG-UL-MCMC in 2-Wasserstein distance is also controlled by the discretization error of the discrete algorithm (3.1) and the ergodicity rate of Hamiltonian Langevin dynamics (1.1). In particular, the main difference in the convergence results of SG-UL-MCMC and SRVR-HMC lies in the discretization error term, which leads to a different gradient complexity for SG-UL-MCMC.

Corollary 3.9. Under the same assumptions in Theorem 3.3, if we set $\eta = \widetilde{O}(\epsilon^2\mu_*^{1/2})$ and $B_0 = \widetilde{O}(\epsilon^{-4}\mu_*^{-1})$, SG-UL-MCMC in (3.1) requires $\widetilde{O}(\epsilon^{-6}\mu_*^{-5/2})$ stochastic gradient evaluations to achieve ϵ-error in 2-Wasserstein distance. Moreover, UL-MCMC requires $\widetilde{O}(\epsilon^{-2}\mu_*^{-3/2}n)$ stochastic gradient evaluations to achieve ϵ-error in 2-Wasserstein distance.

Remark 3.10. Our theoretical analysis suggests that the gradient complexity of UL-MCMC is better than that of HMC [30] by a factor of $O(\epsilon^{-2}\mu_*^{-3/2})$ and the gradient complexity of SG-UL-MCMC is better than that of SGHMC [30] by a factor of $O(\epsilon^{-2}\mu_*^{-5/2})$. We note that Cheng et al. [17] proved $O(1/\epsilon)$ convergence rate of UL-MCMC for sampling from a smaller class of non-log-concave densities in 1-Wasserstein distance. Their result is not directly comparable to our result since 1-Wasserstein distance is strictly smaller than 2-Wasserstein distance and more importantly, their results rely on a stronger assumption than the dissipativeness assumption used in our paper as we commented in Remark 3.4.

4 Experiments

In this section, we evaluate the empirical performance of SRVR-HMC on both synthetic and real datasets. We compare our proposed algorithm with existing overdamped and underdamped Langevin based stochastic gradient algorithms including SGLD [50], SVRG-LD [25], SGHMC [16], SG-UL-MCMC [18] and SVR-HMC [55].

4.1 Sampling from Gaussian mixture distributions

We first demonstrate the performance of SRVR-HMC for fitting a Gaussian mixture model on synthetic data . In this case, the density on each data point is defined as

$$e^{-f_i(\mathbf{x})} = 2e^{\|\mathbf{x}-\mathbf{a}_i\|_2^2/2} + e^{\|\mathbf{x}+\mathbf{a}_i\|_2^2/2},$$

which is proportional to the probability density function (PDF) of two-component Gaussian mixture density with weights $1/3$ and $2/3$. By simple calculation, it can be verified that when $\|\mathbf{a}_i\|_2 \geq 1$, $f_i(\mathbf{x})$ is nonconvex but satisfies Assumption 3.2, and so does $f(\mathbf{x}) = 1/n \sum_{i=1}^n f_i(\mathbf{x})$.

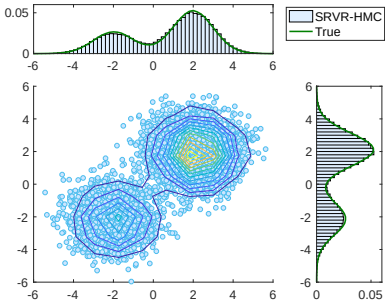
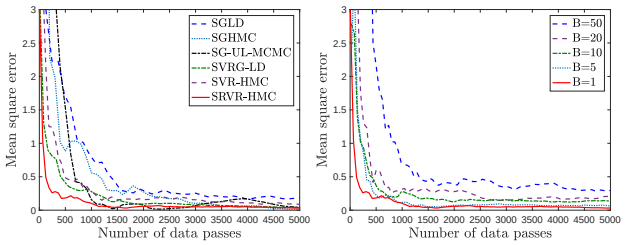


Figure 1: Kernel density estimation for Gaussian mixture distribution.

We generated $n = 500$ vectors $\{\mathbf{a}_i\}_{i=1,\ldots,n} \in \mathbb{R}^2$ to construct the target density functions. We first show that the proposed algorithm can well approximate the target distribution. Specifically, we run SRVR-HMC for 10^4 data passes, and use the last 10^5 iterates to visualize the estimated distribution, where the batch size, minibatch size and epoch length are set to be $B_0 = n$, $B = 1$ and $L = n$ respectively. As a reference, we run MCMC with Metropolis-Hasting (MH) correction to represent the underlying distribution. Following [3], we display the kernel densities of random samples generated by SRVR-HMC in Figures 4.1, which shows that the random samples generated by SRVR-HMC well approximate Gaussian mixture distribution.

In Figure 2(a), we compare the performance of SRVR-HMC with baseline algorithms for sampling from Gaussian mixture distribution. Since directly computing the 2-Wasserstein distance is expensive, we resort to the mean square error (MSE) $\mathbb{E}[\|\widehat{\mathbf{x}}-\bar{\mathbf{x}}\|_2^2]$, where $\bar{\mathbf{x}} = \mathbb{E}_\pi[\mathbf{x}]$ is obtained via running MCMC with MH correction and $\widehat{\mathbf{x}} = \sum_{s=1001}^k \mathbf{x}_s/(k-1000)$ is the sample path average, where $\mathbf{x}_s$ denotes the s-th position iterate of the algorithms and we discard the first 1000 iterates as burn-in. We report the MSE results of all algorithms in Figure 2(a) by repeating each algorithms for 20 times. It can be seen that SRVR-HMC converges faster than all baseline algorithms, which is well aligned with our theory. In addition, it can be seen SG-UL-MCMC outperforms SGHMC, which is consistent with our results in Table 1. We also compare the convergence performance of SRVR-HMC with different batch sizes in Figure 2(b). It can be observed that SRVR-HMC works well for all small batch sizes ($B < 20$) but becomes significantly worse when B is large ($B = 50$). This observation is consistent with Corollary 3.5 where we prove that when $B \lesssim B_0^{1/2}$ the gradient complexity maintains the same.



(a) (b)

Figure 2: Experiment results for sampling from Gaussian mixture distribution, where X-axis represents the number of data passes and Y-axis represents MSE: (a) Comparison with baseline algorithms. (b) Convergence of SRVR-HMC with varying batch size B.

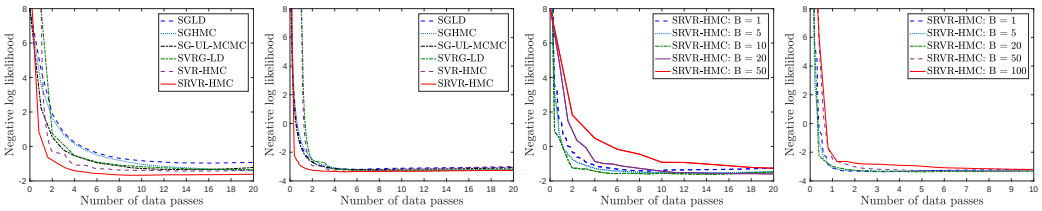
4.2 Independent components analysis

We further run the sampling algorithms for independent components analysis (ICA) tasks. In the ICA model, the input are examples $\{\mathbf{x}_i\}_{i=1}^n$, and the likelihood function can be written as $p(\mathbf{x}|\mathbf{W}) = |\det(\mathbf{W})| \prod_{j=1}^l p(\mathbf{w}_j^\top \mathbf{x})$, where $\mathbf{W} \in \mathbb{R}^{d\times l}$ is the model matrix, d is the problem dimension, l denotes the number of independent components and $\mathbf{w}_j$ denotes the j-th column of $\mathbf{W}$. Following [50, 25] we set $p(\mathbf{w}_j^\top \mathbf{x}) = 1/(4\cosh^2(\mathbf{w}_j^\top \mathbf{x}/2))$ with a Gaussian prior $p(\mathbf{W}) \sim \mathcal{N}(0, \lambda^{-1}\mathbf{I})$. Then the negative log-posterior can be written as $f(\mathbf{W}) = 1/n \sum_{i=1}^n f_i(\mathbf{W})$, where

$$f_i(\mathbf{W}) = -n\log(|\det(\mathbf{W})|) - 2n\sum_{j=1}^l \log\left(\cosh(\mathbf{w}_j^\top \mathbf{x}_i/2)\right) + \lambda\|\mathbf{W}\|_F^2/2.$$

We compare the performance of SRVR-HMC with all the baseline algorithms on MEG dataset[4], which consists of 17730 time-points in 122 channels. In order to explore the performance of our

[4] http://research.ics.aalto.fi/ica/eegmeg/MEG_data.html



(a) n=500, B_0=100 (b) n=5000, B_0=1000 (c) n=500, B_0=100 (d) n=5000, B_0=1000

Figure 3: Experiment results for ICA, where X-axis represents the number of data passes, and Y-axis represents the negative log likelihood on the test dataset: (a)-(b) Comparison with different baselines (c)-(d) Convergence of SRVR-HMC with varying batch size B.

algorithm for different sample size, we extract two subset with sizes $n = 500$ and $n = 5000$ from the original dataset for training, and regard the rest 12730 examples as test dataset. For inference, we compute the sample path average while discarding the first 100 iterates as burn-in. We first compare the convergence performance of SRVR-HMC with baseline algorithms and report the negative log likelihood on test dataset in Figures 3(a)-3(b), where the batch size, minibatch size and epoch length are set to be $B_0 = n/5$, $B = 10$ and $L = B_0/B$, and the rest hyper parameters are tuned to achieve the best performance. It is worth noting that we do not perform the normalization when evaluating the test likelihood, thus the negative log likelihood results may be smaller than 0. From Figures 3(a)-3(b) it can be clearly seen that SRVR-HMC outperforms all baseline algorithms, which validates its superior theoretical properties. Again, we can see that SG-UL-MCMC can decrease the negative log likelihood much faster than SGHMC, which is well aligned with our theory. Furthermore, we evaluate the convergence for different minibatch size, which are displayed in Figures 3(c)-3(d), where the batch size B_0 is fixed as $n/5$ for both scenarios. It can be seen that SRVR-HMC attains similar convergence performance for all small minibatch sizes ($B \leq 10$ when $B_0 = 100$ and $B \leq 20$ when $B_0 = 1000$), which again corroborates our theory that when $B \lesssim B_0^{1/2}$ the gradient complexity maintains the same.

We also evaluate our proposed algorithm SRVR-HMC on Bayesian logistic regression. We defer the additional experimental results to Appendix E due to space limit.

5 Conclusions

We propose a novel algorithm SRVR-HMC based on Hamiltonian Langevin dynamics for sampling from a class of non-log-concave target densities. We show that SRVR-HMC achieves a lower gradient complexity in 2-Wasserstein distance than all existing HMC-type algorithms. In addition, we show that our algorithm reduces to UL-MCMC and SG-UL-MCMC with properly chosen parameters. Our analysis of SRVR-HMC directly applies to these two algorithms and suggests that UL-MCMC/SG-UL-MCMC are faster than HMC/SGHMC for sampling from non-log-concave densities.

Acknowledgement

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

References

[1] Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to mcmc for machine learning. *Machine learning*, 50(1-2):5–43, 2003.

[2] Jack Baker, Paul Fearnhead, Emily B Fox, and Christopher Nemeth. Control variates for stochastic gradient MCMC. *Statistics and Computing*, 2018. ISSN 1573-1375. doi: 10.1007/s11222-018-9826-2.

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

[4] M Barkhagen, NH Chau, É Moulines, M Rásonyi, S Sabanis, and Y Zhang. On stochastic gradient langevin dynamics with dependent data streams in the logconcave case. *arXiv preprint arXiv:1812.02709*, 2018.

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

[6] Michael Betancourt, Simon Byrne, Sam Livingstone, Mark Girolami, et al. The geometric foundations of Hamiltonian monte carlo. *Bernoulli*, 23(4A):2257–2298, 2017.

[7] Francois Bolley and Cedric Villani. Weighted csiszár-kullback-pinsker inequalities and applications to transportation inequalities. *Annales de la Faculté des Sciences de Toulouse. Série VI. Mathématiques*, 14, 01 2005. doi: 10.5802/afst.1095.

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

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

[10] Nicolas Brosse, Alain Durmus, and Eric Moulines. The promises and pitfalls of stochastic gradient langevin dynamics. In *Advances in Neural Information Processing Systems*, pages 8268–8278, 2018.

[11] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

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

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

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

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

[16] Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient Hamiltonian monte carlo. In *International Conference on Machine Learning*, pages 1683–1691, 2014.

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

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

[19] William Coffey and Yu P Kalmykov. *The Langevin equation: with applications to stochastic problems in physics, chemistry and electrical engineering*, volume 27. World Scientific, 2012.

[20] Arnak Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In *Conference on Learning Theory*, pages 678–689, 2017.

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

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

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

[24] Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

[25] Kumar Avinava Dubey, Sashank J Reddi, Sinead A Williamson, Barnabas Poczos, Alexander J Smola, and Eric P Xing. Variance reduction in stochastic gradient Langevin dynamics. In *Advances in Neural Information Processing Systems*, pages 1154–1162, 2016.

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

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

[28] Murat A Erdogdu, Lester Mackey, and Ohad Shamir. Global non-convex optimization with discretized diffusions. In *Advances in Neural Information Processing Systems*, pages 9671–9680, 2018.

[29] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pages 686–696, 2018.

[30] Xuefeng Gao, Mert Gürbüzbalaban, and Lingjiong Zhu. Global convergence of stochastic gradient Hamiltonian monte carlo for non-convex stochastic optimization: Non-asymptotic performance bounds and momentum-based acceleration. *arXiv preprint arXiv:1809.04618*, 2018.

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

[32] Kaiyi Ji, Zhe Wang, Yi Zhou, and Yingbin Liang. Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization. In *International Conference on Machine Learning*, pages 3100–3109, 2019.

[33] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

[34] Peter E Kloeden and Eckhard Platen. Higher-order implicit strong numerical schemes for stochastic differential equations. *Journal of statistical physics*, 66(1):283–314, 1992.

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

[36] Lihua Lei, Cheng Ju, Jianbo Chen, and Michael I Jordan. Non-convex finite-sum optimization via scsg methods. In *Advances in Neural Information Processing Systems*, pages 2345–2355, 2017.

[37] Zhize Li, Tianyi Zhang, and Jian Li. Stochastic gradient Hamiltonian monte carlo with variance reduction for bayesian inference. *arXiv preprint arXiv:1803.11159*, 2018.

[38] M. Lichman. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

[39] Robert S Liptser and Albert N Shiryaev. *Statistics of random processes: I. General theory*, volume 5. Springer Science & Business Media, 2013.

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

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

[42] Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. In *Conference on Learning Theory*, pages 605–638, 2018.

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

[44] Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. *arXiv preprint arXiv:1703.00102*, 2017.

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

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

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

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

[49] Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh. Spiderboost: A class of faster variance-reduced algorithms for nonconvex optimization. *arXiv preprint arXiv:1810.10690*, 2018.

[50] Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pages 681–688, 2011.

[51] Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 3126–3137, 2018.

[52] Yuchen Zhang, Percy Liang, and Moses Charikar. A hitting time analysis of stochastic gradient Langevin dynamics. In *Conference on Learning Theory*, pages 1980–2022, 2017.

[53] Dongruo Zhou, Pan Xu, and Quanquan Gu. Finding local minima via stochastic nested variance reduction. *arXiv preprint arXiv:1806.08782*, 2018.

[54] Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduced gradient descent for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 3925–3936, 2018.

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

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

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

A Proof of main theory

In this section, we prove our main theorems and corollaries.

A.1 Proof of Theorem 3.3

We first prove the convergence of Algorithm 1 in terms of 2-Wasserstein distance. To simplify the proof, we define the following concatenated vectors

$$\mathbf{z}_k = \begin{pmatrix} \mathbf{x}_k \\ \mathbf{v}_k \end{pmatrix}, \qquad \boldsymbol{Z}_t = \begin{pmatrix} \boldsymbol{X}_t \\ \boldsymbol{V}_t \end{pmatrix} \tag{A.1}$$

where $\mathbf{x}_k, \mathbf{v}_k$ are the iterates in Algorithm 1 and $\boldsymbol{X}_t, \boldsymbol{V}_t$ are the variables in the continuous-time dynamics (1.1). Instead of directly bounding $\mathcal{W}_2(\mathbb{P}(\mathbf{x}_k), \pi)$, we aim to prove that its upper bound $\mathcal{W}_2(\mathbb{P}(\mathbf{z}_k), \pi_{\mathbf{z}})$ converges to ϵ-precision, where

$$\pi_{\mathbf{z}} \propto \exp(-\|\mathbf{v}\|_2^2/(2u) + f(\mathbf{x})) \tag{A.2}$$

denotes the stationary distribution of Hamiltonian dynamics (1.1) with respect to both $\mathbf{x}$ and $\mathbf{v}$. By triangle inequality, it holds that

$$\mathcal{W}_2\big(\mathbb{P}(\mathbf{z}_k), \pi_{\mathbf{z}}\big) \leq \mathcal{W}_2\big(\mathbb{P}(\mathbf{z}_k), \mathbb{P}(\boldsymbol{Z}_{k\eta})\big) + \mathcal{W}_2\big(\mathbb{P}(\boldsymbol{Z}_{k\eta}), \pi_{\mathbf{z}}\big). \tag{A.3}$$

The first term on the R.H.S. of (A.3) represents the discretization error of Algorithm 1, and the second term is typically referred to the ergodicity of the continuous-time dynamics (1.1), which characterizes the mixing time of the Markov process $(\boldsymbol{X}_t, \boldsymbol{V}_t)$. These two terms can be upper bounded by the following lemmas respectively.

Lemma A.1. Suppose the initial point of Algorithm 1 is $\mathbf{x} = \mathbf{v} = \mathbf{0}$. $\mathbf{z}_k$ and $\boldsymbol{Z}_t$ are defined as in (A.1). Under Assumptions 3.1 and 3.2, if we set the step size $\eta = O(mM^{-3} \wedge m^{1/2}M^{-3/2}L^{-1/2})$, the 2-Wasserstein distance between the iterate $\mathbf{z}_k$ generated by Algorithm 1 and the point $\boldsymbol{Z}_{k\eta}$ generated by Hamiltonian dynamics (1.1) is upper bounded as follows,

$$\mathcal{W}_2\big(\mathbb{P}(\mathbf{z}_k), \mathbb{P}(\boldsymbol{Z}_{k\eta})\big) \leq 2\bar{\Lambda}\bigg(M^2\gamma^3 u\bar{\mathcal{E}}\bigg(1 + \frac{L}{B}\bigg)K\eta^3 + \frac{M^2\gamma u\bar{\mathcal{E}}K\eta}{B_0} \cdot \mathbb{1}(B_0 < n)\bigg)^{1/4},$$

where $\bar{\Lambda}$ and $\bar{\mathcal{E}}$ are defined as

$$\bar{\Lambda} = \frac{8}{\gamma}\sqrt{\frac{um(f(\mathbf{x}_0) - f(\mathbf{x}^*)) + 2Mu(4d + 2b + m\|\mathbf{x}^*\|_2^2\gamma^2) + (12um + 3\gamma^2)}{m}},$$
$$\bar{\mathcal{E}} = \frac{8um(f(\mathbf{x}_0) - f(\mathbf{x}^*)) + 8Mu\big(20(d+b) + m\|\mathbf{x}^*\|_2^2\big)}{\gamma^2 m} + \frac{G^2}{M^2},$$

$G = \max_{i\in n}\|\nabla f_i(0)\|_2$ and $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}\mathbb{R}^d} f(\mathbf{x})$ is the global minimizer of f.

Lemma A.2. $\boldsymbol{Z}_t$ and $\pi_{\mathbf{z}}$ are defined as in (A.1) and (A.2) respectively. Under Assumptions 3.1 and 3.2, then we have

$$\mathcal{W}_2\big(\mathbb{P}(\boldsymbol{Z}_t), \pi_{\mathbf{z}}\big) \leq \Gamma_0 e^{-\mu_* t},$$

where μ_* denotes the contraction rate of Hamiltonian Langevin dynamics (1.1), which is in the order of $e^{-\widetilde{O}(d)}$ under Assumption 3.2, and Γ_0 is a constant of order $O(1/\mu_*)$.

Here μ^* serves as a lower bound of the spectral gap of the spectral gap of the Markov process generated by (1.1), and in the worst case the exponential dependency on d is unavoidable [27].

Based on the above two lemmas, the proof of Theorem 3.3 is straightforward.

Proof of Theorem 3.3. By Lemmas A.1 and A.2, it holds that

$$\begin{aligned}\mathcal{W}_2\big(\mathbb{P}(\mathbf{z}_K), \pi_{\mathbf{z}}\big) &\leq \mathcal{W}_2\big(\mathbb{P}(\mathbf{z}_K), \mathbb{P}(\boldsymbol{Z}_{K\eta})\big) + \mathcal{W}_2\big(\mathbb{P}(\boldsymbol{Z}_{K\eta}), \pi_{\mathbf{z}}\big) \\ &\leq \Gamma_1\bigg[\bigg(1 + \frac{L}{B}\bigg)K\eta^3 + \frac{K\eta}{\gamma^2 B_0} \cdot \mathbb{1}(B_0 < n)\bigg]^{1/4} + \Gamma_0 e^{-\mu_* K\eta},\end{aligned}$$

where $\Gamma_1 = 2D_1(M^2\gamma^3 u D_2)^{1/4}$, Γ_0 is defined in A.2, and D_1, D_2 correspond to $\bar{\Lambda}$ and $\bar{\mathcal{E}}$ in Lemmas A.1 respectively. By plugging in the definition of 2-Wasserstein distance, we obtain the fact that $\mathcal{W}_2\big(\mathbb{P}(\mathbf{x}_K), \pi\big) \leq \mathcal{W}_2\big(\mathbb{P}(\mathbf{z}_K), \pi_{\mathbf{z}}\big)$, which completes the proof. □

A.2 Proof of Corollary 3.5

Proof. In order to ensure the 2-Wasserstein distance $\mathcal{W}_2\big(\mathbb{P}(\mathbf{x}_k), \pi\big) \leq \epsilon$, we can set

$$\Gamma_1 \bigg[\bigg(1 + \frac{L}{B}\bigg) K\eta^3 + \frac{K\eta}{\gamma^2 B_0} \cdot \mathbb{1}(B_0 < n)\bigg]^{1/4} \leq \frac{\epsilon}{2} \quad \text{and} \quad \Gamma_0 e^{-\mu_* K\eta} = \frac{\epsilon}{2}. \tag{A.4}$$

For the first equation in (A.4), we further set η sufficiently small and B_0 sufficiently large such that

$$\bigg(1 + \frac{L}{B}\bigg) K\eta^3 = \frac{1}{2}\bigg(\frac{\epsilon}{2\Gamma_1}\bigg)^4 \quad \text{and} \quad \frac{K\eta}{\gamma^2 B_0} \cdot \mathbb{1}(B_0 < n) \leq \frac{1}{2}\bigg(\frac{\epsilon}{2\Gamma_1}\bigg)^4.$$

Solving the second equation in (A.4), we obtain $K\eta = \mu_*^{-1}\log(2\Gamma_0/\epsilon)$. Plugging this into the above equations, we have

$$\eta = \frac{\epsilon^2 \mu_*^{1/2}}{4\sqrt{2}\Gamma_1^2\sqrt{(1 + L/B)\log(2\Gamma_0/\epsilon)}}, \qquad B_0 = \frac{32\Gamma_1^4 \log(2\Gamma_0/\epsilon)}{\epsilon^4 \gamma^2 \mu_*} \wedge n.$$

Combining the choice of η and the fact that $K\eta = \mu_*^{-1}\log(2\Gamma_0/\epsilon)$, we get

$$K = \frac{4\sqrt{2}\Gamma_1^2 (1 + L/B)^{1/2} \log^{3/2}(2\Gamma_0/\epsilon)}{\epsilon^2 \mu_*^{3/2}}.$$

Now we can calculate the total gradient complexity of Algorithm 1 as follows:

$$T_g = KB + KB_0/L + B_0.$$

In order to minimize the gradient complexity, it requires to set $BL = O(B_0)$ which implies that $\eta = O(\epsilon^2 \mu_*^{1/2} B_0^{-1/2} B)$. Then we have

$$T_g = \widetilde{O}\big(B_0 + (B^2 + LB)^{1/2} \mu_*^{-3/2} \epsilon^{-2}\big).$$

Note that $B_0 = \widetilde{O}(\epsilon^{-4} \wedge n)$. Thus we can chose the set size B such that $B^2 \lesssim B_0$ and get

$$T_g = \widetilde{O}(\epsilon^{-2} \mu_*^{-3/2} B_0^{1/2}) + O(B_0) = \widetilde{O}\big((\epsilon^{-2} \mu_*^{-3/2} n^{1/2} + n) \wedge \epsilon^{-4} \mu_*^{-2}\big).$$

This completes the proof. □

A.3 Proof of Theorem 3.8

Proof of Theorem 3.8. As we mentioned before, the proposed algorithm can reduce to a variant of SGHMC algorithm by setting $L = 1$. Therefore, the convergence guarantee of SGHMC can be directly generalized from Theorem 3.3, i.e.,

$$\mathcal{W}_2\big(\mathbb{P}(\mathbf{x}_K), \pi\big) \leq \Gamma_1 \bigg[\bigg(1 + \frac{1}{B}\bigg) K\eta^3 + \frac{K\eta}{\gamma^2 B_0} \cdot \mathbb{1}(B_0 < n)\bigg]^{1/4} + \Gamma_0 e^{-\mu_* K\eta}.$$

Since $B \geq 1$, we have $1/B \leq 1$. Plugging this into the above inequality, we can complete the proof. □

A.4 Proof of Corollary 3.9

Proof of Corollary 3.9. In order to guarantee the distance $\mathcal{W}_2\big(\mathbb{P}(\mathbf{x}_K), \pi\big)$ be smaller than ϵ, we can set

$$2K\eta^3 = \frac{\epsilon^4}{32\Gamma_1^4}, \quad \frac{K\eta}{\gamma^2 B_0}\mathbb{1}(B_0 < n) = \frac{\epsilon^4}{32\Gamma_1^4} \quad \text{and} \quad \Gamma_0 e^{-\mu_* K\eta} = \frac{\epsilon}{2}.$$

Solving the above equations, we get

$$K\eta = \mu_*^{-1}\log(2\Gamma_0/\epsilon), \quad \eta = \frac{\epsilon^2 \mu_*^{1/2}}{8\Gamma_1^2\sqrt{\log(2\Gamma_0/\epsilon)}}, \quad B_0 = \frac{32\Gamma_1^4 \log(2\Gamma_0/\epsilon)}{\epsilon^4 \gamma^2 \mu_*} \wedge n.$$

Solving the above we further obtain

$$K = \frac{8\Gamma_1^2 \log^{3/2}(2\Gamma_0/\epsilon)}{\epsilon^2 \mu_*^{3/2}},$$

which implies that the gradient complexity of SG-UL-MCMC is

$$T = KB_0 = \widetilde{O}\big(\epsilon^{-6} \mu^{-5/2} \wedge \epsilon^{-2} \mu^{-3/2} n\big).$$

This completes the proof. □

B Proof of technical lemmas

In this section, we provide the proofs of the two key lemmas presented in the analysis in Appendix A.

B.1 Proof of Lemma A.1

We first lay down the supporting lemmas that would be useful in our proof.

Lemma B.1 (Lemma 10 in [18])**.** The Hamiltonian Langevin dynamics (1.1) has the following solution

$$\boldsymbol{V}_t = \boldsymbol{V}_0 e^{-\gamma t} - u\int_0^t e^{-\gamma(t-s)}\nabla f(\boldsymbol{X}_t)\mathrm{d}s + \widetilde{\boldsymbol{\epsilon}}_t^v, \tag{B.1}$$

$$\boldsymbol{X}_t = \boldsymbol{X}_0 + \frac{1-e^{-\gamma t}}{\gamma}\boldsymbol{V}_0 + u\int_0^t\int_0^s e^{-\gamma(s-r)}\nabla f(\boldsymbol{X}_r)\mathrm{d}r\mathrm{d}s + \widetilde{\boldsymbol{\epsilon}}_t^x, \tag{B.2}$$

where $\widetilde{\boldsymbol{\epsilon}}_t^v = \sqrt{2\gamma u}\int_0^t e^{-\gamma(t-s)}\mathrm{d}\boldsymbol{B}_s$ and $\widetilde{\boldsymbol{\epsilon}}_t^x = \sqrt{2\gamma u}\int_0^t\int_0^s e^{-\gamma(s-r)}\mathrm{d}\boldsymbol{B}_r\mathrm{d}s$ are Gaussian random variables with mean $\mathbf{0}$ and their covariance matrices are as follows:

$$\begin{aligned}
\mathbb{E}[\widetilde{\boldsymbol{\epsilon}}_t^v(\widetilde{\boldsymbol{\epsilon}}_t^v)^\top] &= u(1-e^{-2\gamma t})\cdot\mathbf{I}_{d\times d}\\
\mathbb{E}[\widetilde{\boldsymbol{\epsilon}}_t^x(\widetilde{\boldsymbol{\epsilon}}_t^x)^\top] &= \frac{u}{\gamma^2}(2\gamma t + 4e^{-\gamma t} - e^{-2\gamma t} - 3)\cdot\mathbf{I}_{d\times d}\\
\mathbb{E}[\widetilde{\boldsymbol{\epsilon}}_t^v(\widetilde{\boldsymbol{\epsilon}}_t^x)^\top] &= \frac{u}{\gamma}(1-2e^{-\gamma t}+e^{-2\gamma t})\cdot\mathbf{I}_{d\times d}.
\end{aligned}$$

To prove the convergence of Algorithm 1, we define a Lyapunov function for all $(\mathbf{x},\mathbf{v})\in\mathbb{R}^d\times\mathbb{R}^d$ as follows

$$\mathcal{E}(\mathbf{x},\mathbf{y}) = \|\mathbf{x}\|_2^2 + \|\mathbf{x}+2\mathbf{v}/\gamma\|_2^2 + 8u\big(f(\mathbf{x})-f(\mathbf{x}^*)\big)/\gamma^2. \tag{B.3}$$

Note that $\|\mathbf{a}\|_2^2+\|\mathbf{b}\|_2^2 \geq \|\mathbf{a}-\mathbf{b}\|_2^2/2$. By the definition of $\mathcal{E}$ and the fact that $f(\mathbf{x})\geq f(\mathbf{x}^*)$, we have

$$\mathcal{E}(\mathbf{x},\mathbf{v}) \geq \|\mathbf{x}\|_2^2 + \|\mathbf{x}+2\mathbf{v}/\gamma\|_2^2 \geq \max\{\|\mathbf{x}\|_2^2, 2\|\mathbf{v}/\gamma\|_2^2\}. \tag{B.4}$$

Lemma B.2. Under Assumptions 3.1 and 3.2, if we set the step size of Algorithm 1 according to the following condition:

$$\begin{aligned}
\eta \leq \min\Bigg(&\frac{\gamma}{4(8Mu+u\gamma+22\gamma^2)}, \sqrt{\frac{4u^2}{4Mu+3\gamma^2}}, \frac{6\gamma bu}{(4Mu+3\gamma^2)d},\\
&\frac{\gamma^4 m}{48(46\gamma^2+288u\gamma+32u)M^3u}, \frac{\gamma m^{1/2}}{48M^{3/2}(\gamma^2+u)^{1/2}L^{1/2}}, \frac{\gamma}{\sqrt{6Mu}}, \frac{\gamma\bar{\mathcal{E}}^{1/2}}{2Gu}, \frac{1}{2\sqrt{L}\gamma}\Bigg),
\end{aligned}$$

and $B_0\geq\min\{1/\eta, n\}$, then for all $k\geq 0$, $\mathbb{E}[\|\mathbf{x}_k\|_2^2]$, $\mathbb{E}[\|\mathbf{v}_k\|_2^2]$ and $\mathbb{E}[\|\mathbf{g}_k\|_2^2]$ can be bounded as follows,

$$\mathbb{E}[\|\mathbf{x}_k\|_2^2]\leq\bar{\mathcal{E}},\quad \mathbb{E}[\|\mathbf{v}_k\|_2^2]\leq\gamma^2\bar{\mathcal{E}}/2,\quad \text{and}\quad \mathbb{E}[\|\mathbf{g}_k\|_2^2]\leq 14M^2\bar{\mathcal{E}},$$

where $\bar{\mathcal{E}}$ is defined as

$$\bar{\mathcal{E}} = \mathcal{E}(\mathbf{x}_0,\mathbf{v}_0) + \frac{8Mu\big[16(d+b)+m\|\mathbf{x}^*\|_2^2\big]}{\gamma^2 m} + \frac{G^2}{M^2},\qquad G=\max_{i\in n}\|\nabla f_i(\mathbf{0})\|_2,$$

and $\mathcal{E}(\mathbf{x},\mathbf{y})$ is the Lyapunov function defined in (B.3).

The following lemma characterizes the expected distance between the semi-stochastic gradient $\mathbf{g}_k$ and the full gradient $\nabla f(\mathbf{x}_k)$.

Lemma B.3. Suppose Assumptions 3.1 and 3.2 hold. For Algorithm 1, if we choose the same step size η used in Lemma B.2, then it holds that

$$\mathbb{E}[\|\mathbf{g}_k-\nabla f(\mathbf{x}_k)\|_2^2] \leq \frac{4LM^2\gamma^2\eta^2\bar{\mathcal{E}}}{B} + \frac{4M^2\bar{\mathcal{E}}}{B_0}\cdot\mathbb{1}(B_0<n),$$

where $\bar{\mathcal{E}}$ is defined in Lemma B.2.

The next lemma is referred to as the exponential integrability.

Lemma B.4. Suppose Assumptions 3.1 and 3.2 hold. Let $\theta > 0$ be any constant such that $\theta \leq \min\{\gamma^2/(128u), m/32\}$. Then, it holds that

$$\log\Big(\mathbb{E}\big[e^{\theta(\|\boldsymbol{X}_t\|_2^2+\|\boldsymbol{V}_t\|_2^2)}\big]\Big) \leq 2\theta\mathcal{E}(\boldsymbol{X}_0,\boldsymbol{V}_0) + \frac{32M\theta u\big(4d+2b+m\|\mathbf{x}^*\|_2^2\big)}{\gamma^2 m},$$

where $\mathcal{E}(\mathbf{x},\mathbf{y})$ is the Lyapunov function defined in (B.3).

The following weighted CKP inequality gives a tight connection between 2-Wasserstein distance and KL divergence.

Lemma B.5 (Weighted CKP Inequality [7])**.** For any two probability measures P and Q, if they have finite second moments, the following holds,

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

where $\Lambda = 2\inf_{\theta>0}\sqrt{1/\theta(3/2+\log\mathbb{E}_{\mathbf{x}\sim P}[e^{\theta\|\mathbf{x}\|_2^2}])}$.

Now we are ready to prove our first key lemma on the discretization error of Algorithm 1.

Proof of Lemma A.1. By the weighted CKP inequality in Lemma B.5, we have

$$\mathcal{W}_2\big(\mathbb{P}(\mathbf{z}_K),\mathbb{P}(\boldsymbol{Z}_{K\eta})\big) \leq \Lambda\Big(\sqrt{D_{KL}(\mathbb{P}(\mathbf{z}_K)||\mathbb{P}(\boldsymbol{Z}_{K\eta}))} + \sqrt[4]{D_{KL}(\mathbb{P}(\mathbf{z}_K)||\mathbb{P}(\boldsymbol{Z}_{K\eta}))}\Big), \tag{B.5}$$

where $\Lambda = 2\inf_{\theta>0}\sqrt{1/\theta(3/2+\log\mathbb{E}_{\mathbb{P}(\boldsymbol{Z}_T)}[e^{\theta\|\boldsymbol{Z}_T\|_2^2}])}$ and $T = K\eta$. By (A.1) it holds that $\|\boldsymbol{Z}_T\|_2^2 = \|\boldsymbol{X}_T\|_2^2 + \|\boldsymbol{V}_T\|_2^2$. Applying Lemma B.4, we obtain

$$\begin{aligned}
\Lambda &= 2\inf_{\theta>0}\sqrt{1/\theta(3/2+\log\mathbb{E}_{\mathbb{P}_T}[e^{\theta(\|\boldsymbol{X}_T\|_2^2+\|\boldsymbol{V}_T\|_2^2)}])} \\
&\leq 2\inf_{0<\theta\leq\min\{\frac{\gamma^2}{128u},\frac{m}{32}\}}\sqrt{\frac{1}{\theta}\bigg(\frac{3}{2}+2\theta\mathcal{E}(\boldsymbol{X}_0,\boldsymbol{V}_0)+\frac{32M\theta u\big(4d+2b+m\|\mathbf{x}^*\|_2^2\big)}{\gamma^2 m}\bigg)} \\
&\leq 2\sqrt{2\mathcal{E}(\boldsymbol{X}_0,\boldsymbol{V}_0)+\frac{32Mu\big(4d+2b+m\|\mathbf{x}^*\|_2^2\big)+16(12um+3\gamma^2)}{\gamma^2 m}} := \bar{\Lambda},
\end{aligned} \tag{B.6}$$

where in the last inequality we used the fact that the infimum value is attained at $\theta = \min\{\gamma^2/(128u), m/32\}$ and the fact that $1/\theta \leq 128u/\gamma^2 + 32/m$. Therefore, it remains to prove the upper bound of the KL divergence between distributions $\mathbb{P}(\mathbf{z}_K)$ and $\mathbb{P}(\boldsymbol{Z}_{K\eta})$, which can be done by following the standard techniques in [21, 45, 51] to construct a continuous-time Markov process. In particular, based on the update rule in Algorithm 1, we define the following continuous-time interpolation of $(\mathbf{v}_k,\mathbf{x}_k)$

$$\begin{aligned}
\mathrm{d}\widetilde{\boldsymbol{V}}_t &= -\gamma\widetilde{\boldsymbol{V}}_t\mathrm{d}t - u\widetilde{\mathbf{G}}_t\mathrm{d}t + \sqrt{2\gamma u}\cdot\mathrm{d}\boldsymbol{B}_t \\
\mathrm{d}\widetilde{\boldsymbol{X}}_t &= \widetilde{\boldsymbol{V}}_t\mathrm{d}t,
\end{aligned} \tag{B.7}$$

where $\widetilde{\mathbf{G}}_t = \sum_{k=0}^{\infty}\mathbf{g}_k\,\mathbb{1}\{t\in[k\eta,(k+1)\eta)\}$ remains invariant in each interval $[k\eta,(k+1)\eta)$ and $\mathbf{g}_k$ is the semi-stochastic gradient at the k-th iteration of Algorithm 1. It can be verified that the distribution of $(\mathbf{v}_k,\mathbf{x}_k)$ is identical to that of $(\widetilde{\boldsymbol{V}}_{k\eta},\widetilde{\boldsymbol{X}}_{k\eta})$. Integrating (B.7) from 0 to t gives

$$\begin{aligned}
\widetilde{\boldsymbol{V}}_t &= \widetilde{\boldsymbol{V}}_0 - \int_0^t\gamma\widetilde{\boldsymbol{V}}_s\mathrm{d}s - \int_0^t u\widetilde{\mathbf{G}}_s\mathrm{d}s + \int_0^t\sqrt{2\gamma u}\cdot\mathrm{d}\boldsymbol{B}_s, \\
\widetilde{\boldsymbol{X}}_t &= \widetilde{\boldsymbol{X}}_0 + \int_0^t\widetilde{\mathbf{V}}_s\mathrm{d}s.
\end{aligned}$$

Due to the semi-stochastic gradient $\mathbf{g}_k$, (B.7) does not form a Markov chain since $\widetilde{\mathbf{G}}_s$ contains additional randomness introduced by the stochastic gradient. Nevertheless, Gyöngy [31] showed

that we can use the following Markov chain whose one-time marginal distribution mimics that of $(\widetilde{\boldsymbol{V}}_t, \widetilde{\boldsymbol{X}}_t)$,

$$\widehat{\boldsymbol{V}}_t = \widehat{\boldsymbol{V}}_0 - \int_0^t \gamma \widehat{\boldsymbol{V}}_s \mathrm{d}s - \int_0^t u \widehat{\mathbf{G}}_s \mathrm{d}s + \int_0^t \sqrt{2\gamma u} \cdot \mathrm{d}\boldsymbol{B}_s,$$
$$\widehat{\boldsymbol{X}}_t = \widehat{\boldsymbol{X}}_0 + \int_0^t \widehat{\boldsymbol{V}}_s \mathrm{d}s,$$

where $\widehat{\mathbf{G}}_s = \mathbb{E}[\widetilde{\mathbf{G}}_s | \widetilde{\mathbf{V}}_s = \widehat{\mathbf{V}}_s]$. Next, we let $\mathbb{P}_t$ denote the probability measure of the point $(\boldsymbol{V}_t, \boldsymbol{X}_t)$ in Hamiltonian Langevin dynamics and $\mathbb{Q}_t$ denote the probability measure of $(\widehat{\boldsymbol{V}}_t, \widehat{\boldsymbol{X}}_t)$. By Girsanov formula [39] we can derive the Radon-Nikodym derivative of $\mathbb{P}_t$ with respect to $\mathbb{Q}_t$ as follows:

$$\frac{\mathrm{d}\mathbb{P}_t}{\mathrm{d}\mathbb{Q}_t} = \exp\left\{ \sqrt{\frac{\gamma u}{2}} \int_0^t (\nabla f(\widehat{\boldsymbol{X}}_s) - \widehat{\mathbf{G}}_s) \cdot \mathrm{d}\boldsymbol{B}_s - \frac{\gamma u}{4} \int_0^t \|\nabla f(\widehat{\boldsymbol{X}}_s) - \widehat{\mathbf{G}}_s\|_2^2 \mathrm{d}s \right\}.$$

When we choose $T = K\eta$, it follows that

$$\begin{aligned} D_{KL}(\mathbb{Q}_T || \mathbb{P}_T) &= \mathbb{E}_{\mathbb{Q}_T}\left[\log\left(\frac{\mathrm{d}\mathbb{P}_T}{\mathrm{d}\mathbb{Q}_T} \right) \right] \\ &= \frac{\gamma u}{4} \int_0^T \mathbb{E}\big[\|\nabla f(\widehat{\boldsymbol{X}}_s) - \widehat{\mathbf{G}}_s\|_2^2 \big] \mathrm{d}s \\ &= \frac{\gamma u}{4} \int_0^T \mathbb{E}\big[\|\nabla f(\widetilde{\boldsymbol{X}}_s) - \widetilde{\mathbf{G}}_s\|_2^2 \big] \mathrm{d}s \\ &= \frac{\gamma u}{4} \sum_{k=0}^{K-1} \int_{k\eta}^{(k+1)\eta} \mathbb{E}\big[\|\nabla f(\widetilde{\boldsymbol{X}}_s) - \widetilde{\mathbf{G}}_s\|_2^2 \big] \mathrm{d}s, \end{aligned} \tag{B.8}$$

where the third equality holds since $\widehat{\boldsymbol{X}}_s$ has the same distribution as $\widetilde{\boldsymbol{X}}_s$. Moreover, note that $\widetilde{\mathbf{G}}_s$ is a step function based on semi-stochastic gradients $\{\mathbf{g}_k\}_{k=1,\ldots,K}$, and equals $\mathbf{g}_k$ when $s \in [k\eta, (k+1)\eta)$ for all $k < K$. Therefore, in the k-th interval, i.e., $s \in [k\eta, (k+1)\eta)$, we have

$$\mathbb{E}[\|\nabla f(\widetilde{\boldsymbol{X}}_s) - \widetilde{\mathbf{G}}_s\|_2^2] \le 2\mathbb{E}[\|\nabla f(\widetilde{\boldsymbol{X}}_s) - \nabla f(\mathbf{x}_k)\|_2^2] + 2\mathbb{E}[\|\nabla f(\mathbf{x}_k) - \mathbf{g}_k\|_2^2], \tag{B.9}$$

where $\mathbf{x}_k = \widetilde{\boldsymbol{X}}_{k\eta}$ is the k-th iterate in Algorithm 1. We then upper bound two terms on the R.H.S. of (B.9) separately. Regarding the first term $\mathbb{E}[\|\nabla f(\widetilde{\boldsymbol{X}}_s) - \nabla f(\mathbf{x}_k)\|_2^2]$, Assumption 3.1 implies

$$\mathbb{E}[\|\nabla f(\widetilde{\boldsymbol{X}}_s) - \nabla f(\mathbf{x}_k)\|_2^2] \le M^2 \mathbb{E}[\|\widetilde{\boldsymbol{X}}_s - \widetilde{\boldsymbol{X}}_{k\eta}\|_2^2], \tag{B.10}$$

where we replaced $\mathbf{x}_k$ with $\widetilde{\boldsymbol{X}}_{k\eta}$. Multiplying $e^{\gamma t}$ to both sides of the first equation in (B.7) yields

$$(\mathrm{d}\widetilde{\boldsymbol{V}}_t + \gamma \widetilde{\boldsymbol{V}}_t \mathrm{d}t) e^{\gamma t} = -u \widetilde{\mathbf{G}}_t e^{\gamma t} \mathrm{d}t + \sqrt{2\gamma u} \cdot e^{\gamma t} \cdot \mathrm{d}\boldsymbol{B}_t.$$

Note that $(\mathrm{d}\widetilde{\boldsymbol{V}}_t + \gamma \widetilde{\boldsymbol{V}}_t dt) e^{\gamma t} = \mathrm{d}\big(\widetilde{\boldsymbol{V}}_t e^{\gamma t}\big)$, integrating both sides over t from $k\eta$ to r gives

$$\widetilde{\boldsymbol{V}}_r e^{\gamma r} - \widetilde{\boldsymbol{V}}_{k\eta} e^{\gamma k\eta} = \int_{k\eta}^r -u \widetilde{\mathbf{G}}_z e^{\gamma z} \mathrm{d}z + \int_{k\eta}^r \sqrt{2\gamma u} \cdot e^{\gamma z} \cdot \mathrm{d}\boldsymbol{B}_z,$$

which can be further simplified as

$$\widetilde{\boldsymbol{V}}_r = \widetilde{\boldsymbol{V}}_{k\eta} \cdot e^{-\gamma(r-k\eta)} - \int_{k\eta}^r u \widetilde{\mathbf{G}}_z e^{-\gamma(r-z)} \mathrm{d}z + \int_{k\eta}^r \sqrt{2\gamma u} \cdot e^{-\gamma(r-z)} \cdot \mathrm{d}\boldsymbol{B}_z.$$

Thus by the second equation in (B.7) we have

$$\begin{aligned} \widetilde{\boldsymbol{X}}_s &= \widetilde{\boldsymbol{X}}_{k\eta} + \int_{k\eta}^s \widetilde{\boldsymbol{V}}_r dr \\ &= \widetilde{\boldsymbol{X}}_{k\eta} + \int_{k\eta}^s \left(\widetilde{\boldsymbol{V}}_{k\eta} e^{-\gamma(r-k\eta)} - u\left(\int_{k\eta}^r e^{-\gamma(r-z)} \widetilde{\mathbf{G}}_{k\eta} \mathrm{d}z \right) + \sqrt{2\gamma u} \int_{k\eta}^r e^{-\gamma(r-z)} \mathrm{d}\boldsymbol{B}_z \right) \mathrm{d}r, \end{aligned}$$

where $\widetilde{\mathbf{G}}_z = \widetilde{\mathbf{G}}_{k\eta}$ for $z \in [k\eta, (k+1)\eta)$ by definition. This further implies that

$$\begin{aligned}
\|\widetilde{\boldsymbol{X}}_s - \widetilde{\boldsymbol{X}}_{k\eta}\|_2^2 &= \bigg\| \int_{k\eta}^{s} \Big(\widetilde{\boldsymbol{V}}_{k\eta} e^{-\gamma(r-k\eta)} - u \int_{k\eta}^{r} e^{-\gamma(r-z)} \widetilde{\mathbf{G}}_{k\eta} \mathrm{d}z + \sqrt{2\gamma u} \int_0^r e^{-\gamma(r-z)} \mathrm{d}\boldsymbol{B}_z \Big) \mathrm{d}r \bigg\|_2^2 \\
&\leq 3 \bigg\| \int_{k\eta}^{s} \widetilde{\boldsymbol{V}}_{k\eta} e^{-\gamma(r-k\eta)} \mathrm{d}r \bigg\|_2^2 + 3u^2 \bigg\| \int_{k\eta}^{s} \int_{k\eta}^{r} e^{-\gamma(r-z)} \widetilde{\mathbf{G}}_{k\eta} \mathrm{d}z \mathrm{d}r \bigg\|_2^2 \\
&\quad + 6\gamma u \bigg\| \int_{k\eta}^{s} \int_0^r e^{-\gamma(r-z)} \mathrm{d}\boldsymbol{B}_z \mathrm{d}r \bigg\|_2^2 \\
&\leq 3\eta^2 \|\mathbf{v}_k\|_2^2 + 3u^2\eta^4 \|\mathbf{g}_k\|_2^2 + 6\gamma u \bigg\| \int_{k\eta}^{s} \int_0^r e^{-\gamma(r-z)} \mathrm{d}\boldsymbol{B}_z \mathrm{d}r \bigg\|_2^2,
\end{aligned}$$

where the second inequality follows from the fact that $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$ and the last inequality follows from facts that $s \in [k\eta, (k+1)\eta))$, $\widetilde{\boldsymbol{V}}_{k\eta} = \mathbf{v}_k$, $\widetilde{\mathbf{G}}_{k\eta} = \mathbf{g}_k$ and $e^{-\gamma(r-z)} \leq 1$. Moreover, by Lemma B.1, we have

$$\mathbb{E}\bigg[\bigg\| \int_{k\eta}^{s} \int_0^r e^{-\gamma(r-z)} \mathrm{d}\boldsymbol{B}_z \mathrm{d}r \bigg\|_2^2 \bigg] = \frac{d}{\gamma^2} \big(2\gamma(s-k\eta) + 4e^{-\gamma(s-k\eta)} - e^{-2\gamma(s-k\eta)} - 3 \big) \leq 2d\eta^2,$$

where we use inequality $1 - x \leq e^{-x} \leq 1 - x + x^2/2$ for positive x and $0 \leq s - k\eta \leq \eta$ to get the last inequality. Combining the above analysis and (B.10), we have

$$\mathbb{E}[\|\nabla f(\widetilde{\boldsymbol{X}}_s) - \nabla f(\mathbf{x}_k)\|_2^2] \leq 3M^2\eta^2 \big(\mathbb{E}[\|\mathbf{v}_k\|_2^2] + u^2\eta^2 \mathbb{E}[\|\mathbf{g}_k\|_2^2]/4 + 4\gamma u d \big).$$

Applying Lemma B.2 and setting $\eta^2 \leq \min\{\gamma^2/(4M^2u^2), \gamma^2\bar{\mathcal{E}}/(2G^2u^2)\}$, we have

$$\mathbb{E}[\|\nabla f(\widetilde{\boldsymbol{X}}_s) - \nabla f(\mathbf{x}_k)\|_2^2] \leq 4M^2\gamma^2\eta^2\bar{\mathcal{E}}.$$

Then in terms of the second term on the R.H.S. of (B.9), we have the following by Lemma B.3,

$$\mathbb{E}[\|\nabla f(\mathbf{x}_k) - \mathbf{g}_k\|_2^2] \leq \frac{4LM^2\gamma^2\eta^2\bar{\mathcal{E}}}{B} + \frac{4M^2\bar{\mathcal{E}}}{B_0} \cdot \mathbb{1}(B_0 < n).$$

Plugging the above inequalities into (B.9) and further (B.8), we have

$$\begin{aligned}
D_{KL}(\mathbb{Q}_T || \mathbb{P}_T) &= \frac{\gamma u}{4} \sum_{k=0}^{K-1} \int_{k\eta}^{(k+1)\eta} \mathbb{E}[\|\nabla f(\widetilde{\boldsymbol{X}}_s) - \widetilde{\mathbf{G}}_s\|_2^2] \\
&\leq M^2\gamma^3 u\bar{\mathcal{E}} \bigg(1 + \frac{L}{B} \bigg) K\eta^3 + \frac{M^2\gamma u\bar{\mathcal{E}} K\eta}{B_0} \cdot \mathbb{1}(B_0 < n). \qquad \text{(B.11)}
\end{aligned}$$

Combining (B.5), (B.6) and (B.11) and assuming that $D_{KL}(\mathbb{Q}_T || \mathbb{P}_T) \leq 1$, we get

$$\mathcal{W}_2(\mathbb{Q}_{K\eta}, \mathbb{P}_{K\eta}) \leq 2\bar{\Lambda} \bigg(M^2\gamma^3 u\bar{\mathcal{E}} \bigg(1 + \frac{L}{B} \bigg) K\eta^3 + \frac{M^2\gamma u\bar{\mathcal{E}} K\eta}{B_0} \cdot \mathbb{1}(B_0 < n) \bigg)^{1/4},$$

which completes the proof. □

B.2 Proof of Lemma A.2

Now we prove Lemma A.2 which characterizes the exponential mixing rate of the Hamiltonian dynamics (1.1). Our analysis will be built based on the contraction results of Langevin dynamics in [27]. We first lay down some useful lemmas that will be used in our analysis.

The following lemma is a direct implication of Assumption 3.2.

Lemma B.6. If $f(\mathbf{x})$ satisfies Assumption 3.2, then for all $\mathbf{x} \in \mathbb{R}^d$ it holds that

$$\langle \nabla f(\mathbf{x}), \mathbf{x} \rangle / 2 \geq \lambda \big(f(\mathbf{x}) + u^{-1}\gamma^2 \|\mathbf{x}\|_2^2/4 \big) - A, \qquad \text{(B.12)}$$

where λ and A are parameters defined as

$$\lambda = \frac{2m}{4M + u^{-1}\gamma^2} \quad \text{and} \quad A = \frac{2m(f(\mathbf{x}^*) + M\|\mathbf{x}^*\|_2^2)}{4M + u^{-1}\gamma^2} + \frac{b}{2}. \qquad \text{(B.13)}$$

Before we present the contraction results provided in [27], we first define a semi-metric $\mathcal{W}_\rho(\cdot,\cdot)$. For any concatenated vectors $(\mathbf{x},\mathbf{v}),(\mathbf{x}',\mathbf{v}')\in\mathbb{R}^{2d}$ (equivalently, $\mathbf{x},\mathbf{v},\mathbf{x}',\mathbf{v}'\in\mathbb{R}^d$), we define

$$\begin{aligned} r((\mathbf{x},\mathbf{v}),(\mathbf{x}',\mathbf{v}')) &= \alpha\|\mathbf{x}-\mathbf{x}'\|_2 + \|\mathbf{x}-\mathbf{x}'+\gamma^{-1}(\mathbf{v}-\mathbf{v}')\|_2, \\ \rho((\mathbf{x},\mathbf{v}),(\mathbf{x}',\mathbf{v}')) &= h(r((\mathbf{x},\mathbf{v}),(\mathbf{x}',\mathbf{v}')))(1+\theta\mathcal{V}(\mathbf{x},\mathbf{v})+\theta\mathcal{V}(\mathbf{x}',\mathbf{v}')), \end{aligned} \tag{B.14}$$

where $\alpha,\theta\in(0,\infty)$ are constants. $h:[0,\infty)\to[0,\infty)$ (1) is a continuous, non-decreasing concave function which is C^2 continuous on $(0,R_1)$ for some constant $R_1>0$; (2) is a constant function on $[0,\infty)$; (3) and satisfies $h(0)=0$, $h'_+(0)=1$ and $h'_-(R_1)>0$. $\mathcal{V}:\mathbb{R}^{2d}\to\mathbb{R}$ is defined as follows

$$\mathcal{V}(\mathbf{x},\mathbf{v}) = f(\mathbf{x}) + \frac{\gamma^2}{4u}\big(\|\mathbf{x}+\gamma^{-1}\mathbf{v}\|_2^2 + \|\gamma^{-1}\mathbf{v}\|_2^2 - \lambda\|\mathbf{x}\|_2^2\big),$$

where γ,u are the parameter of dynamics in (1.1), λ is defined in (B.13). For any two probability measures μ and ν, we define

$$\mathcal{W}_\rho(\mu,\nu) = \inf_{\zeta\in\Gamma(\mu,\nu)}\int\rho((\mathbf{x},\mathbf{v}),(\mathbf{x}',\mathbf{v}'))\mathrm{d}\zeta((\mathbf{x},\mathbf{v}),(\mathbf{x}',\mathbf{v}')), \tag{B.15}$$

where the infimum is over all couplings of μ and ν. As is pointed out by Eberle et al. [27], $\mathcal{W}_\rho$ may not necessarily be a metric and thus triangle inequality does not hold. Therefore, we call $\mathcal{W}_\rho$ a semi-metric.

Recall the solution of Hamiltonian dynamics in Lemma B.1. We use $\mathcal{L}_t$ to denote the operator of integration on the dynamics from time 0 to t. That is, $\mathcal{L}_t\boldsymbol{V}_0=\boldsymbol{V}_t$ and $\mathcal{L}_t\boldsymbol{X}_0=\boldsymbol{X}_t$ denote the velocity and the position of the random process. Suppose the initial point $\boldsymbol{Z}_0=(\boldsymbol{X}_0^\top,\boldsymbol{V}_0^\top)^\top\in\mathbb{R}^{2d}$ follows a distribution μ. Then with a slight abuse of notation, we also use $\mathcal{L}_t\mu$ to denote the distribution of $\boldsymbol{Z}_t=(\boldsymbol{X}_t^\top,\boldsymbol{V}_t^\top)^\top$. Built on the above preliminaries and notations, the following lemma is about the contraction of Hamiltonian dynamics in terms of semi-metric $\mathcal{W}_\rho$.

Lemma B.7 (Theorem 2.3 and Corollary 2.6 in [27])**.** Suppose Assumptions 3.1 and 3.2 hold and thus (B.12) is true. There exist constants $\alpha,\theta>0$ and a continuous non-decreasing concave function $h:[0,\infty)\to[0,\infty)$ as required in (B.14) such that for all probability measures μ,ν, it holds that

$$\mathcal{W}_2(\mathcal{L}_t\mu,\mathcal{L}_t\nu)\le C_0\sqrt{\mathcal{W}_\rho(\mu,\nu)}e^{-\mu_* t}$$

for all $t\ge0$, where μ_* is a lower bound of the spectral gap of Markov chain (1.1) and satisfies

$$\begin{aligned} \mu_* &= \frac{1}{768\gamma e^\Lambda}\min\{\lambda Mue^\Lambda,\Lambda^{1/2}Mu,\gamma\Lambda^{1/2}\}, \\ \Lambda &= \frac{12(1+2\alpha+2\alpha^2)(d+A)Mu}{5\gamma^2\lambda(1-2\lambda)}, \\ C_0 &= \frac{\sqrt{2}e^{1+\Lambda/2}}{\min\{1,\alpha\}}\max\left\{1,\frac{2\sqrt{2+4\alpha+4\alpha^2}(d+A)^{1/2}u^{1/2}\gamma^{-1/2}\mu_*^{-1/2}}{\min\{1,(8\Lambda/M)^{1/4}\}}\right\}, \end{aligned} \tag{B.16}$$

γ,u are the parameters in dynamics (1.1) and λ,A are defined in (B.13).

In particular, for Lemma B.7, the function h in the definition of semi-metric $\mathcal{W}_\rho$ in (B.15) is chosen as follows:

$$h(r) = \int_0^{r\wedge R_1}\phi(s)g(s)\mathrm{d}s,$$

where $R_1=\sqrt{8\Lambda/M}$ and the auxiliary functions are defined as

$$\begin{aligned} \phi(s) &= e^{-\frac{(1+\eta)Ms^2}{8}-\frac{\gamma^2s^2\max\{1,1/(2\alpha)\}}{2u}}, \quad \Phi(s)=\int_0^s\phi(x)\mathrm{d}x, \\ g(s) &= 1-\frac{9\lambda^*\gamma}{4u}\int_0^s\Phi(x)\phi(x)^{-1}\mathrm{d}x. \end{aligned}$$

Now we are ready to complete the proof of Lemma A.2.

Proof of Lemma A.2. By (2.11) in [27], we know that the function $h(r)$ is upper bounded by $R_1 = \sqrt{8\Lambda/M}$ defined in Lemma B.7. Thus, the distance function $\rho((\mathbf{x},\mathbf{v}),(\mathbf{x}',\mathbf{v}'))$ can be bounded as

$$\rho((\mathbf{x},\mathbf{v}),(\mathbf{x}',\mathbf{v}')) \le R_1(1+\theta\mathcal{V}(\mathbf{x},\mathbf{v})+\theta\mathcal{V}(\mathbf{x}',\mathbf{v}')).$$

Let μ_0 denote the distribution of $(\mathbf{x}_0,\mathbf{v}_0)$. It follows that

$$\begin{aligned}\mathcal{W}_\rho(\mu_0,\pi_{\mathbf{z}}) &= \inf_{\zeta\in\Gamma(\mu_0,\pi_{\mathbf{z}})}\int \rho((\mathbf{x}_0,\mathbf{v}_0),(\mathbf{x}_\pi,\mathbf{v}_\pi))d\zeta((\mathbf{x}_0,\mathbf{v}_0),(\mathbf{x}^\pi,\mathbf{v}^\pi))\\ &\le R_1\big(1+\theta\mathbb{E}[\mathcal{V}(\mathbf{x}_0,\mathbf{v}_0)]+\theta\mathbb{E}[\mathcal{V}(\mathbf{x}^\pi,\mathbf{v}^\pi)]\big). \end{aligned} \tag{B.17}$$

Moreover, recall the definition of function $\mathcal{V}(\mathbf{x},\mathbf{v})$ we have

$$\begin{aligned}\mathcal{V}(\mathbf{x},\mathbf{v}) &= f(\mathbf{x})+\frac{\gamma^2}{4u}\big(\|\mathbf{x}+\gamma^{-1}\mathbf{v}\|_2^2+\|\gamma^{-1}\mathbf{v}\|_2^2-\lambda\|\mathbf{x}\|_2^2\big)\\ &\le f(\mathbf{x}^*)+\frac{M}{2}\|\mathbf{x}-\mathbf{x}^*\|_2^2+\frac{\gamma^2}{4u}\big(\|\mathbf{x}+\gamma^{-1}\mathbf{v}\|_2^2+\|\gamma^{-1}\mathbf{v}\|_2^2-\lambda\|\mathbf{x}\|_2^2\big)\\ &\le f(\mathbf{x}^*)+\bigg(M+\frac{\gamma^2(2-\lambda)}{4u}\bigg)\|\mathbf{x}\|_2^2+M\|\mathbf{x}^*\|_2^2+\frac{3\|\mathbf{v}\|_2^2}{4u},\end{aligned}$$

where the first inequality comes from the smoothness of f (Assumption 3.1) and the second inequality is due to the fact that $(a+b)^2\le 2a^2+2b^2$ for any a,b. Note that the stationary distribution π is proportional to the function $e^{-f(\mathbf{x})-\|\mathbf{v}\|_2^2/(2u)}$. As is shown in [45] (Section 3.5, equation (3.19)), we know that

$$\mathbb{E}[\|\mathbf{x}^\pi\|_2^2]\le\frac{b+d}{m}.$$

In addition, since the marginal distribution of $\mathbf{v}^\pi$ is a d dimensional Gaussian distribution, we have $\mathbb{E}[\|\mathbf{v}^\pi\|_2^2]=du$. Plugging these into (B.17), we have

$$\begin{aligned}\mathcal{W}_\rho(\mu_0,\pi_{\mathbf{z}}) &\le R_1\big(1+\theta\mathbb{E}[\mathcal{V}(\mathbf{x}_0,\mathbf{v}_0)]+\theta\mathbb{E}[\mathcal{V}(\mathbf{x}^\pi,\mathbf{v}^\pi)]\big)\\ &\le R_1\bigg(1+\theta\bigg(\mathcal{V}(\mathbf{x}_0,\mathbf{v}_0)+f(\mathbf{x}^*)+\frac{\big(4Mu+\gamma^2(2-\lambda)\big)(b+d)}{4um}+d+M\|\mathbf{x}^*\|_2^2\bigg)\bigg).\end{aligned}$$

Moreover, note the fact that $\mathbf{x}_0=\mathbf{v}_0=\mathbf{0}$. We further obtain

$$\mathcal{W}_\rho(\mu_0,\pi_{\mathbf{z}})\le R_1\bigg(1+\theta\bigg(2f(\mathbf{x}^*)+\frac{\big(4Mu+\gamma^2(2-\lambda)\big)(b+d)}{4um}+d+2M\|\mathbf{x}^*\|_2^2\bigg)\bigg):=\Theta.$$

For the stationary distribution $\pi_{\mathbf{z}}$, it is invariant under the Hamiltonian dynamics, i.e., $\mathcal{L}_t\pi_{\mathbf{z}}=\pi_{\mathbf{z}}$ for any $t\ge 0$. Note that $\mathcal{L}_t\mu_0=\mathbb{P}(\boldsymbol{Z}_t)$ by definition. By Lemma B.7, we have

$$\begin{aligned}\mathcal{W}_2\big(\mathbb{P}(\boldsymbol{Z}_t),\pi_{\mathbf{z}}\big) &\le C_0\sqrt{\mathcal{W}_\rho\big(\mathbb{P}(\boldsymbol{Z}_0),\pi_{\mathbf{z}}\big)}e^{-\mu_* t}\\ &\le C_0\Theta^{1/2}e^{-\mu_* t},\end{aligned}$$

where μ_* and C_0 are defined in (B.16) in Lemma B.7. Let $\Gamma_0=C_0\Theta^{1/2}$, it can be seen that both Γ_0 and $1/\mu_*$ are in the order of $\exp(\widetilde{O}(d))$. This completes the proof. □

C Proof of technical lemmas in Appendix B

In this section, we prove the technical lemmas used in the proof of our key lemmas.

C.1 Proof of Lemma B.2

We first present the following lemma on upper bound of the gradient norm, which is a straightforward implication of the smoothness of f.

Lemma C.1. Under Assumption 3.1, for all $\mathbf{x}\in\mathbb{R}^d$ and $i\in[n]$, it holds that

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

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

Lemma C.2. Under Assumption 3.1, let $k = jm + l$, it holds that

$$\mathbb{E}[\|\mathbf{g}_k - \nabla f(\mathbf{x}_k)\|_2^2] \leq \frac{M^2}{B} \sum_{s=jm}^{jm+l} \mathbb{E}[\|\mathbf{x}_{s+1} - \mathbf{x}_s\|_2^2] + \frac{2}{B_0} \mathbb{E}[\|\mathbf{x}_{jm}\|_2^2 + G^2] \cdot \mathbb{1}(B_0 < n),$$

where G follows the same definition in Lemma C.1.

Proof of Lemma B.2. Recall the Lyapunov function defined in (B.3), we have

$$\mathcal{E}(\mathbf{x}_{k+1}, \mathbf{v}_{k+1}) = \|\mathbf{x}_{k+1}\|_2^2 + \|\mathbf{x}_{k+1} + 2\mathbf{v}_{k+1}/\gamma\|_2^2 + 8u\big(f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)\big)/\gamma^2. \tag{C.1}$$

By Assumption 3.1, it holds that

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + M\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2/2 - f(\mathbf{x}^*). \tag{C.2}$$

For the first two terms in (C.1), we have

$$\begin{aligned}
\|\mathbf{x}_{k+1}\|_2^2 &= \|\mathbf{x}_k\|_2^2 + 2\langle \mathbf{x}_k, \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2, \\
\|\mathbf{x}_{k+1} + 2\mathbf{v}_{k+1}/\gamma\|_2^2 &= \|\mathbf{x}_k + 2\mathbf{v}_k/\gamma\|_2^2 + 2\langle \mathbf{x}_k + 2\mathbf{v}_k/\gamma, \mathbf{x}_{k+1} - \mathbf{x}_k + 2(\mathbf{v}_{k+1} - \mathbf{v}_k)/\gamma \rangle \\
&\quad + \|\mathbf{x}_{k+1} - \mathbf{x}_k + 2(\mathbf{v}_{k+1} - \mathbf{v}_k)/\gamma\|_2^2,
\end{aligned}$$

Substituting the above two equations and (C.2) into (C.1) yields

$$\begin{aligned}
&\mathbb{E}[\mathcal{E}(\mathbf{x}_{k+1}, \mathbf{v}_{k+1})] \\
&\leq \mathbb{E}[\mathcal{E}(\mathbf{x}_k, \mathbf{v}_k)] + 4\mathbb{E}[\langle \mathbf{x}_k, \mathbf{x}_{k+1} - \mathbf{x}_k \rangle] + \frac{4}{\gamma}\mathbb{E}[\langle \mathbf{x}_k, \mathbf{v}_{k+1} - \mathbf{v}_k \rangle] + \frac{4}{\gamma}\mathbb{E}[\langle \mathbf{v}_k, \mathbf{x}_{k+1} - \mathbf{x}_k \rangle] \\
&\quad + \frac{8}{\gamma^2}\mathbb{E}[\langle \mathbf{v}_k, \mathbf{v}_{k+1} - \mathbf{v}_k \rangle] + \frac{8u}{\gamma^2}\mathbb{E}\big[\langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + M/2\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2\big] \\
&\quad + \mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2] + \mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}_k + 2(\mathbf{v}_{k+1} - \mathbf{v}_k)/\gamma\|_2^2].
\end{aligned} \tag{C.3}$$

Next, we need to upper bound inner products terms $\langle \mathbf{x}_k, \mathbf{x}_{k+1} - \mathbf{x}_k \rangle$, $\langle \mathbf{x}_k, \mathbf{v}_{k+1} - \mathbf{v}_k \rangle$, $\langle \mathbf{v}_k, \mathbf{x}_{k+1} - \mathbf{x}_k \rangle$, and $\langle \mathbf{v}_k, \mathbf{v}_{k+1} - \mathbf{v}_k \rangle$ respectively. Recall the update formula of Algorithm 1 as follows,

$$\begin{aligned}
\mathbf{v}_{k+1} &= \mathbf{v}_k e^{-\gamma\eta} - \frac{u(1 - e^{-\gamma\eta})}{\gamma}\mathbf{g}_k + \boldsymbol{\epsilon}_k^v, \\
\mathbf{x}_{k+1} &= \mathbf{x}_k + \frac{1 - e^{-\gamma\eta}}{\gamma}\mathbf{v}_k + \frac{u(\gamma\eta + e^{-\gamma\eta} - 1)}{\gamma^2}\mathbf{g}_k + \boldsymbol{\epsilon}_k^x.
\end{aligned} \tag{C.4}$$

Note that $\boldsymbol{\epsilon}_k^v$ and $\boldsymbol{\epsilon}_k^x$ are zero mean and independent of $\mathbf{v}_k, \mathbf{x}_k$ and $\mathbf{g}_k$. Then we have

$$\begin{aligned}
\mathbb{E}[\langle \mathbf{x}_k, \mathbf{x}_{k+1} - \mathbf{x}_k \rangle] &= \frac{1 - e^{-\gamma\eta}}{\gamma}\mathbb{E}[\langle \mathbf{x}_k, \mathbf{v}_k \rangle] + \frac{u(\gamma\eta + e^{-\gamma\eta} - 1)}{\gamma^2}\mathbb{E}[\langle \mathbf{x}_k, \mathbf{g}_k \rangle], \\
\mathbb{E}[\langle \mathbf{x}_k, \mathbf{v}_{k+1} - \mathbf{v}_k \rangle] &= -\big(1 - e^{-\gamma\eta}\big)\mathbb{E}[\langle \mathbf{x}_k, \mathbf{v}_k \rangle] - \frac{u(1 - e^{-\gamma\eta})}{\gamma}\mathbb{E}[\langle \mathbf{x}_k, \mathbf{g}_k \rangle], \\
\mathbb{E}[\langle \mathbf{v}_k, \mathbf{x}_{k+1} - \mathbf{x}_k \rangle] &= \frac{1 - e^{-\gamma\eta}}{\gamma}\mathbb{E}[\|\mathbf{v}_k\|_2^2] + \frac{u(\gamma\eta + e^{-\gamma\eta} - 1)}{\gamma^2}\mathbb{E}[\langle \mathbf{v}_k, \mathbf{g}_k \rangle], \\
\mathbb{E}[\langle \mathbf{v}_k, \mathbf{v}_{k+1} - \mathbf{v}_k \rangle] &= -\big(1 - e^{-\gamma\eta}\big)\mathbb{E}[\|\mathbf{v}_k\|_2^2] - \frac{u(1 - e^{-\gamma\eta})}{\gamma}\mathbb{E}[\langle \mathbf{v}_k, \mathbf{g}_k \rangle].
\end{aligned}$$

Plugging the above bounds for inner products and (C.4) into (C.3) yields

$$\begin{aligned}
&\mathbb{E}[\mathcal{E}(\mathbf{x}_{k+1}, \mathbf{v}_{k+1})] \\
&\leq \mathbb{E}[\mathcal{E}(\mathbf{x}_k, \mathbf{v}_k)] - \frac{4u(2 - \gamma\eta - 2e^{-\gamma\eta})}{\gamma^2}\mathbb{E}[\langle \mathbf{x}_k, \mathbf{g}_k \rangle] - \frac{4(1 - e^{-\gamma\eta})}{\gamma^2}\mathbb{E}[\|\mathbf{v}_k\|_2^2] \\
&\quad + \frac{4u\big(\gamma\eta + e^{-\gamma\eta} - 1\big)}{\gamma^3}\mathbb{E}[\langle \mathbf{v}_k, \mathbf{g}_k \rangle] + \frac{8u(1 - e^{-\gamma\eta})}{\gamma^3}\mathbb{E}[\langle \mathbf{v}_k, \nabla f(\mathbf{x}_k) - \mathbf{g}_k \rangle] \\
&\quad + \frac{8u^2\big(\gamma\eta + e^{-\gamma\eta} - 1\big)}{\gamma^4}\mathbb{E}[\langle \nabla f(\mathbf{x}_k), \mathbf{g}_k \rangle] + \left(\frac{4Mu}{\gamma^2} + 3\right)\mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2]
\end{aligned}$$

$$+\frac{8}{\gamma^2}\mathbb{E}[\|\mathbf{v}_{k+1}-\mathbf{v}_k\|_2^2]. \tag{C.5}$$

By Assumption 3.2, we know that $\langle \mathbf{x}_k, \nabla f(\mathbf{x}_k)\rangle \geq m\|\mathbf{x}_k\|_2^2 - b$. We then assume $\eta \leq 1/(8\gamma)$ and use the inequality $-x \leq e^{-x} - 1 \leq x^2/2 - x$ for any $x \geq 0$, it follows that

$$\begin{aligned}
&\frac{4u(2-\gamma\eta-2e^{-\gamma\eta})}{\gamma^2}\mathbb{E}[\langle \mathbf{x}_k, \mathbf{g}_k\rangle] \\
&= \frac{4u(2-\gamma\eta-2e^{-\gamma\eta})}{\gamma^2}\big[\mathbb{E}[\langle \mathbf{x}_k, \nabla f(\mathbf{x}_k)\rangle] + \mathbb{E}[\langle \mathbf{x}_k, \mathbf{g}_k - \nabla f(\mathbf{x})\rangle]\big] \\
&\geq \frac{4u(2-\gamma\eta-2e^{-\gamma\eta})}{\gamma^2}\big(m\|\mathbf{x}_k\|_2^2 - b\big) - \frac{4u(2-\gamma\eta-2e^{-\gamma\eta})}{\gamma^2}\bigg(\frac{1}{8}\mathbb{E}[\|\mathbf{x}_k\|_2^2] + 2\mathbb{E}[\|\mathbf{g}_k - \nabla f(\mathbf{x}_k)\|_2^2]\bigg) \\
&\geq \frac{3mu\eta}{\gamma}\|\mathbf{x}_k\|_2^2 - \frac{4u\eta b}{\gamma} - \frac{8u\eta}{\gamma}\mathbb{E}[\|\mathbf{g}_k - \nabla f(\mathbf{x}_k)\|_2^2],
\end{aligned}$$

where the first inequality is by Young's inequality and the last one is based on the inequality $\gamma\eta - (\gamma\eta)^2 \leq 2 - \gamma\eta - 2e^{-\gamma\eta} \leq \gamma\eta$. Similarly, by Young's inequality, we also have

$$\begin{aligned}
\frac{8u(1-e^{-\gamma})}{\gamma^3}\mathbb{E}[\langle \mathbf{v}_k, \nabla f(\mathbf{x}_k) - \mathbf{g}_k\rangle] &\leq \frac{8u(1-e^{-\gamma\eta})}{\gamma^3}\bigg[\frac{\gamma}{8u}\mathbb{E}[\|\mathbf{v}_k\|_2^2] + \frac{2u}{\gamma}\mathbb{E}[\|\nabla f(\mathbf{x}_k) - \mathbf{g}_k\|_2^2]\bigg] \\
&\leq \frac{1-e^{-\gamma\eta}}{\gamma^2}\mathbb{E}[\|\mathbf{v}_k\|_2^2] + \frac{16u^2\eta}{\gamma^3}\mathbb{E}[\|\nabla f(\mathbf{x}_k) - \mathbf{g}_k\|_2^2].
\end{aligned}$$

Then again by Young's inequalities $\mathbb{E}[\langle \mathbf{v}_k, \mathbf{g}_k\rangle] \leq 1/2\mathbb{E}[\|\mathbf{v}_k\|_2^2] + 1/2\mathbb{E}[\|\mathbf{g}_k\|_2^2]$ and $\mathbb{E}[\langle \nabla f(\mathbf{x}_k), \mathbf{g}_k\rangle] \leq 1/2\mathbb{E}[\|\nabla f(\mathbf{x}_k)\|_2^2] + 1/2\mathbb{E}[\|\mathbf{g}_k\|_2^2]$, (C.5) can be further simplified as

$$\begin{aligned}
&\mathbb{E}[\mathcal{E}(\mathbf{x}_{k+1}, \mathbf{v}_{k+1})] \\
&\leq \mathbb{E}[\mathcal{E}(\mathbf{x}_k, \mathbf{v}_k)] - \frac{3um\eta}{\gamma}\mathbb{E}[\|\mathbf{x}_k\|_2^2] + \frac{4u\eta b}{\gamma} - \frac{3(1-e^{-\gamma\eta}) - u\gamma\eta^2}{\gamma^2}\mathbb{E}[\|\mathbf{v}_k\|_2^2] \\
&\quad + \frac{8u\gamma^2\eta + 16u^2\eta}{\gamma^3}\mathbb{E}[\|\nabla f(\mathbf{x}_k) - \mathbf{g}_k\|_2^2] + \frac{(2u+\gamma)u\eta^2}{\gamma^2}\mathbb{E}[\|\mathbf{g}_k\|_2^2] + \frac{2u^2\eta^2}{\gamma^2}\mathbb{E}[\|\nabla f(\mathbf{x}_k)\|_2^2] \\
&\quad + \bigg(\frac{4Mu}{\gamma^2} + 3\bigg)\mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2] + \frac{8}{\gamma^2}\mathbb{E}[\|\mathbf{v}_{k+1} - \mathbf{v}_k\|_2^2],
\end{aligned} \tag{C.6}$$

where we use the inequality $-x \leq e^{-x} - 1 \leq x^2/2 - x$ again. We then focus on bounding terms $\mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2]$ and $\mathbb{E}[\|\mathbf{v}_{k+1} - \mathbf{v}_k\|_2^2]$. According to (C.4), we have

$$\begin{aligned}
\mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2] &= \mathbb{E}\bigg[\bigg\|\frac{1-e^{-\gamma\eta}}{\gamma}\mathbf{v}_k + \frac{u(\gamma\eta + e^{-\gamma\eta} - 1)}{\gamma^2}\mathbf{g}_k\bigg\|_2^2\bigg] + \mathbb{E}[\|\boldsymbol{\epsilon}_k^x\|_2^2] \\
&\leq 2\eta^2\mathbb{E}[\|\mathbf{v}_k\|_2^2] + \frac{u^2\eta^4}{2}\mathbb{E}[\|\mathbf{g}_k\|_2^2] + \mathbb{E}[\|\boldsymbol{\epsilon}_k^x\|_2^2],
\end{aligned} \tag{C.7}$$

where the first equation is due to the independence between $\boldsymbol{\epsilon}_k^x$ and $\mathbf{v}_k, \mathbf{g}_k$, and the inequality come from the fact that $-x \leq e^{-x} - 1 \leq x^2/2 - x$ and Young's inequality to (C.4). Similarly, we also have

$$\mathbb{E}[\|\mathbf{v}_{k+1} - \mathbf{v}_k\|_2^2] \leq 2\gamma^2\eta^2\mathbb{E}[\|\mathbf{v}_k\|_2^2] + 2u^2\eta^2\mathbb{E}[\|\mathbf{g}_k\|_2^2] + \mathbb{E}[\|\boldsymbol{\epsilon}_k^v\|_2^2]. \tag{C.8}$$

Furthermore, by (2.3) it can be easily verified that $\mathbb{E}[\|\boldsymbol{\epsilon}_k^v\|_2^2] \leq 2\gamma ud\eta$ and $\mathbb{E}[\|\boldsymbol{\epsilon}_k^x\|_2^2] \leq 2ud\eta^2$. Plugging (C.7) and (C.8) into (C.6) gives

$$\begin{aligned}
&\mathbb{E}[\mathcal{E}(\mathbf{x}_{k+1}, \mathbf{v}_{k+1})] \\
&\leq \mathbb{E}[\mathcal{E}(\mathbf{x}_k, \mathbf{v}_k)] - \frac{3um\eta^2}{\gamma}\mathbb{E}[\|\mathbf{x}_k\|_2^2] - \frac{3(1-e^{-\gamma\eta}) - \eta^2\big(8Mu + u\gamma + 22\gamma^2\big)}{\gamma^2}\mathbb{E}[\|\mathbf{v}_k\|_2^2] \\
&\quad + \frac{36u^2\eta^2 + 2\gamma u\eta^2 + (4Mu + 3\gamma^2)\eta^4}{2\gamma^2}\mathbb{E}[\|\mathbf{g}_k\|_2^2] + \frac{2u^2\eta^2}{\gamma^2}\mathbb{E}[\|\nabla f(\mathbf{x}_k)\|_2^2] \\
&\quad + \frac{8u\eta(\gamma^2 + 2u)}{\gamma^3}\mathbb{E}[\|\nabla f(\mathbf{x}_k) - \mathbf{g}_k\|_2^2] + \frac{(8Mu + 6\gamma^2)ud\eta^2 + 4(4d+b)u\gamma\eta}{\gamma^2},
\end{aligned} \tag{C.9}$$

where we the fact that $-x \leq e^{-x} - 1 \leq x^2/2 - x$. Note that $1 - \exp(x) \geq 3x/4$ when $x \leq 1/2$. Thus, we set

$$\eta \leq \min\left\{\frac{\gamma}{4(8Mu + u\gamma + 22\gamma^2)}, \sqrt{\frac{4u^2}{4Mu + 3\gamma^2}}, \frac{6\gamma bu}{(4Mu + 3\gamma^2)d}\right\},$$

and obtain the following according to (C.9),

$$\begin{aligned}\mathbb{E}[\mathcal{E}(\mathbf{x}_{k+1}, \mathbf{v}_{k+1})] \leq \mathbb{E}[\mathcal{E}(\mathbf{x}_k, \mathbf{v}_k)] - \frac{3um\eta}{\gamma}\mathbb{E}[\|\mathbf{x}_k\|_2^2] - \frac{2\eta}{\gamma}\mathbb{E}[\|\mathbf{v}_k\|_2^2] + \frac{(20u + \gamma)u\eta^2}{\gamma^2}\mathbb{E}[\|\mathbf{g}_k\|_2^2] \\ + \frac{2u^2\eta^2}{\gamma^2}\mathbb{E}[\|\nabla f(\mathbf{x}_k)\|_2^2] + \frac{8u\eta(\gamma^2 + 2u)}{\gamma^3}\mathbb{E}[\|\nabla f(\mathbf{x}_k) - \mathbf{g}_k\|_2^2] + \frac{16(d+b)u\eta}{\gamma}.\end{aligned} \tag{C.10}$$

We complete the proof via induction. In particular, we aim to prove the following

$$\begin{aligned}\mathbb{E}[\mathcal{E}(\mathbf{x}_k, \mathbf{v}_k)] \leq \bar{\mathcal{E}} = \mathcal{E}(\mathbf{x}_0, \mathbf{v}_0) + \frac{8Mu\big[20(d+b) + m\|\mathbf{x}^*\|_2^2\big]}{\gamma^2 m} + \frac{G^2}{M^2}, \\ \mathbb{E}[\|\mathbf{g}_k\|_2^2] \leq 2M^2\bar{\mathcal{E}} + 2\mathbb{E}[\|\nabla f(\mathbf{x}_k)\|_2^2].\end{aligned} \tag{C.11}$$

First, it is easy to verify that (C.11) holds for $(\mathbf{x}_0, \mathbf{v}_0)$. Then we assume it holds for all $(\mathbf{x}_s, \mathbf{v}_s)$ with $s \leq k$, and prove it remains true for $(\mathbf{x}_{k+1}, \mathbf{v}_{k+1})$. It is worthy noting that by (B.4), we have $\mathbb{E}[\|\mathbf{x}_s\|_2^2] \leq \bar{\mathcal{E}}$ and $\mathbb{E}[\|\mathbf{v}_s\|_2^2] \leq \gamma^2\bar{\mathcal{E}}/2$ for all $s \leq k$.

Induction for $\mathbb{E}[\|\mathbf{g}_{k+1}\|_2^2]$: Regarding $\mathbb{E}[\|\mathbf{g}_{k+1}\|_2^2]$, we aim to show that $\mathbb{E}[\|\mathbf{g}_{k+1}\|_2^2] \leq 2M^2\bar{\mathcal{E}} + 2\mathbb{E}[\|\nabla f(\mathbf{x}_{k+1})\|_2^2]$. By Lemma C.2, we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{g}_{k+1} - \nabla f(\mathbf{x}_{k+1})\|_2^2] &\leq \frac{M^2}{B}\sum_{s=jm}^{jm+l}\mathbb{E}[\|\mathbf{x}_{s+1} - \mathbf{x}_s\|_2^2] + \frac{2}{B_0}\big(\mathbb{E}[M^2\|\mathbf{x}_{jm}\|_2^2] + G^2\big)\cdot\mathbb{1}(B_0 < n) \\ &\leq \frac{M^2}{B_0}\sum_{s=jm}^{k}\left[2\eta^2\mathbb{E}[\|\mathbf{v}_s\|_2^2] + \frac{u^2\eta^4}{2}\mathbb{E}[\|\mathbf{g}_s\|_2^2] + \mathbb{E}[\|\boldsymbol{\epsilon}_s^x\|_2^2]\right] \\ &\quad + \frac{2}{B}\big(\mathbb{E}[M^2\|\mathbf{x}_{jm}\|_2^2] + G^2\big).\end{aligned} \tag{C.12}$$

Note that by the induction assumption, Lemma C.1 and Young's inequality, we have

$$\mathbb{E}[\|\nabla f(\mathbf{x}_s)\|_2^2] \leq 2M^2\mathbb{E}[\|\mathbf{x}_s\|_2^2] + 2G^2 \leq 2M^2\bar{\mathcal{E}} + 2G^2,$$

for all $s \leq k$, which implies $\mathbb{E}[\|\mathbf{g}_s\|_2^2] \leq 6M^2\bar{\mathcal{E}} + 4G^2$ for all $s \leq k$. In addition we have $\mathbb{E}[\|\mathbf{x}_{jm}\|_2^2] \leq \bar{\mathcal{E}}$ since $jm \leq k$. Therefore, we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{g}_{k+1} - \nabla f(\mathbf{x}_{k+1})\|_2^2] &\leq \frac{M^2}{B}\sum_{s=jm}^{k}\left[2\eta^2\mathbb{E}[\|\mathbf{v}_s\|_2^2] + \frac{u^2\eta^4}{2}\mathbb{E}[\|\mathbf{g}_s\|_2^2] + \mathbb{E}[\|\boldsymbol{\epsilon}_s^x\|_2^2]\right] + \frac{2}{B_0}(M^2\bar{\mathcal{E}} + G^2) \\ &\leq \frac{LM^2}{B}\big[\gamma^2\eta^2\bar{\mathcal{E}} + u^2(3M\bar{\mathcal{E}} + 2G^2)\eta^4 + 2ud\eta^2\big] + \frac{4M^2\bar{\mathcal{E}}}{B_0}\cdot\mathbb{1}(B_0 < n),\end{aligned}$$

where we use the fact that $\mathbf{G}^2 \leq M^2\bar{\mathcal{E}}$. Let $\eta^2 \leq \min\{\gamma^2/(6M^2u^2), \gamma^2\bar{\mathcal{E}}/(4G^2u^2)\}$, and use the fact that $u \leq \gamma^2\bar{\mathcal{E}}/d$, we have

$$\mathbb{E}[\|\mathbf{g}_{k+1} - \nabla f(\mathbf{x}_{k+1})\|_2^2] \leq \frac{4LM^2\gamma^2\bar{\mathcal{E}}\eta^2}{B} + \frac{4M^2\bar{\mathcal{E}}}{B_0}\cdot\mathbb{1}(B_0 < n). \tag{C.13}$$

Moreover, by Young's inequality, we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{g}_{k+1}\|_2^2] &\leq 2\mathbb{E}[\|\nabla f(\mathbf{x}_{k+1}) - \mathbf{g}_{k+1}\|_2^2] + 2\mathbb{E}[\|\nabla f(\mathbf{x}_k)\|_2^2] \\ &\leq 2\mathbb{E}[\|\nabla f(\mathbf{x}_k)\|_2^2] + \frac{8LM^2\gamma^2\bar{\mathcal{E}}\eta^2}{B} + 4M^2\bar{\mathcal{E}}\cdot\mathbb{1}(B_0 < n).\end{aligned}$$

Let $\eta^2 \leq 4^{-1}L^{-1}\gamma^{-2}$, it is evident that

$$\mathbb{E}[\|\mathbf{g}_{k+1}\|_2^2] \leq \frac{2M^2}{B}\bar{\mathcal{E}} + 2\mathbb{E}[\|\nabla f(\mathbf{x}_{k+1})\|_2^2] \leq 6M^2\bar{\mathcal{E}} + 2\mathbb{E}[\|\nabla f(\mathbf{x}_{k+1})\|_2^2],$$

which completes the induction for $\mathbf{g}_k$.

Induction for $\mathcal{E}(\mathbf{x}_{k+1}, \mathbf{v}_{k+1})$: By assumption (C.11), we have $\mathbb{E}[\|\mathbf{g}_k\|_2^2] \leq 6M^2\bar{\mathcal{E}} + 2\mathbb{E}[\|\nabla f(\mathbf{x}_k)\|_2^2] \leq 14M^2\bar{\mathcal{E}}$. Moreover, by (C.13) we have

$$\mathbb{E}[\|\mathbf{g}_{k+1} - \nabla f(\mathbf{x}_{k+1})\|_2^2] \leq \frac{4LM^2\gamma^2\bar{\mathcal{E}}\eta^2}{B} + \frac{4M^2\bar{\mathcal{E}}}{B_0} \cdot \mathbb{1}(B_0 < n).$$

Note that by Young's inequality $\|\mathbf{a}+\mathbf{b}\|_2^2 \leq 3\|\mathbf{a}\|_2^2/2 + 3\|\mathbf{b}\|_2^2$ we have

$$\mathcal{E}(\mathbf{x}, \mathbf{v}) \leq 5/2\|\mathbf{x}\|_2^2 + \frac{12}{\gamma^2}\|\mathbf{v}\|_2^2 + \frac{2uM}{\gamma^2}\big(3\|\mathbf{x}\|_2^2 + 6\|\mathbf{x}^*\|_2^2\big),$$

where we used the inequality

$$f(\mathbf{x}) - f(\mathbf{x}^*) \leq \frac{M}{2}\|\mathbf{x}-\mathbf{x}^*\|_2^2 \leq \frac{M}{4}\big(3\|\mathbf{x}\|_2^2 + 6\|\mathbf{x}^*\|_2^2\big).$$

Then if $\gamma^2 \leq 4Mu$, we have

$$\mathcal{E}(\mathbf{x}, \mathbf{v}) \leq \frac{12}{\gamma^2}\|\mathbf{v}\|_2^2 + \frac{16uM}{\gamma^2}\|\mathbf{x}\|_2^2 + \frac{12uM}{\gamma^2}\|\mathbf{x}^*\|_2^2. \tag{C.14}$$

Therefore, by (C.10) and the fact that $B_0 \geq \min\{1/\eta, n\}$, we have

$$\begin{aligned}
\mathbb{E}[\mathcal{E}(\mathbf{x}_{k+1}, \mathbf{v}_{k+1})] &\leq \mathbb{E}[\mathcal{E}(\mathbf{x}_k, \mathbf{v}_k)] - \frac{3um\eta}{\gamma}\mathbb{E}[\|\mathbf{x}_k\|_2^2] - \frac{2\eta}{\gamma}\mathbb{E}[\|\mathbf{v}_k\|_2^2] + \frac{32(\gamma^2+u)uLM^2\bar{\mathcal{E}}\eta^3}{\gamma B} \\
&\quad + \frac{32(\gamma^2+u)uM^2\eta^2}{\gamma^3}\bar{\mathcal{E}} + \frac{(288u+14\gamma)M^2u\eta^2}{\gamma^2}\bar{\mathcal{E}} + \frac{16(d+b)u\eta}{\gamma} \\
&\leq \bigg(1 - \frac{\gamma m\eta}{6M}\bigg)\mathbb{E}[\mathcal{E}(\mathbf{x}_k, \mathbf{v}_k)] + \frac{(46\gamma^2 + 288u\gamma + 32u)M^2u\eta^2}{\gamma^3}\bar{\mathcal{E}} \\
&\quad + \frac{32(\gamma^2+u)uLM^2\eta^3}{\gamma B}\bar{\mathcal{E}} + \frac{16(d+b)u\eta + 2um\|\mathbf{x}^*\|_2^2\eta}{\gamma},
\end{aligned}$$

where the last inequality follows from (C.14). We then set the step size as

$$\eta \leq \min\bigg\{\frac{\gamma^4 m}{48(46\gamma^2 + 288u\gamma + 32u)M^3u}, \frac{\gamma m^{1/2}}{48M^{3/2}(\gamma^2+u)^{1/2}L^{1/2}}\bigg\},$$

and use the fact that $B \geq 1$, the following holds,

$$\begin{aligned}
\mathbb{E}[\mathcal{E}(\mathbf{x}_{k+1}, \mathbf{v}_{k+1})] &\leq \bigg(1 - \frac{\gamma m\eta}{6M}\bigg)\mathbb{E}[\mathcal{E}(\mathbf{x}_k, \mathbf{v}_k)] + \frac{\gamma m\eta}{24M}\bar{\mathcal{E}} + \frac{20(d+b)u\eta + 2um\|\mathbf{x}^*\|_2^2\eta}{\gamma} \\
&\leq \bigg(1 - \frac{\gamma m\eta}{8M}\bigg)\bar{\mathcal{E}} + \frac{16(d+b)u\eta + 2um\|\mathbf{x}^*\|_2^2\eta}{\gamma},
\end{aligned}$$

where the last inequality follows from the assumption that $\mathbb{E}[\mathcal{E}(\mathbf{x}_k, \mathbf{v}_k)] \leq \bar{\mathcal{E}}$. Since we have set

$$\bar{\mathcal{E}} = \mathcal{E}(\mathbf{x}_0, \mathbf{v}_0) + \frac{8M\big[16(d+b)u + um\|\mathbf{x}^*\|_2^2\big]}{\gamma^2 m} + \frac{G^2}{M^2},$$

it is evident that $\mathbb{E}[\mathcal{E}(\mathbf{x}_{k+1}, \mathbf{v}_{k+1})] \leq \bar{\mathcal{E}}$ holds as well. This completes the induction for $\mathcal{E}(\mathbf{x}_{k+1}, \mathbf{v}_{k+1})$.

□

C.2 Proof of Lemma B.3

Proof. Note that the semi-stochastic gradient $\mathbf{g}_k$ takes form

$$\mathbf{g}_k = \frac{1}{B}\sum_{i\in\mathcal{B}_k}\big(\nabla f_i(\mathbf{x}_k) - \nabla f_i(\mathbf{x}_{k-1})\big) + \mathbf{g}_{k-1}.$$

By Lemma C.2 and (C.12), we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{g}_k - \nabla f(\mathbf{x}_k)\|_2^2] &\le \frac{M^2}{B}\sum_{s=jm}^{jm+l}\mathbb{E}[\|\mathbf{x}_{s+1}-\mathbf{x}_s\|_2^2] + \frac{2}{B_0}\big(M^2\mathbb{E}[\|\mathbf{x}_{jm}\|_2^2]+G^2\big)\cdot\mathbb{1}(B_0<n)\\ &\le \frac{LM^2}{B}\bigg(2\eta^2\mathbb{E}[\|\mathbf{v}_s\|_2^2] + \frac{u^2\eta^4}{2}\mathbb{E}[\|\mathbf{g}_s\|_2^2] + \mathbb{E}[\|\boldsymbol{\epsilon}_s^x\|_2^2]\bigg)\\ &\quad + \frac{2}{B_0}\big(M^2\mathbb{E}[\|\mathbf{x}_{jm}\|_2^2]+G^2\big)\cdot\mathbb{1}(B_0<n).\end{aligned}$$

Then by Lemma B.2 and (C.13), set $\eta^2 \le \min\{\gamma^2/(6M^2u^2), \gamma^2\bar{\mathcal{E}}/(4G^2u^2)\}$ and use the fact that $G^2 \le M^2\bar{\mathcal{E}}$, we obtain

$$\mathbb{E}[\|\mathbf{g}_k - \nabla f(\mathbf{x}_k)\|_2^2] \le \frac{4LM^2\gamma^2\eta^2\bar{\mathcal{E}}}{B} + \frac{4M^2\bar{\mathcal{E}}}{B_0}\cdot\mathbb{1}(B_0<n).$$

This completes the proof. □

C.3 Proof of Lemma B.4

Proof. Similar to the proof of Lemma B.2, we define

$$\mathcal{E}(\mathbf{x},\mathbf{v}) = \|\mathbf{x}\|_2^2 + \|\mathbf{x}+2\mathbf{v}/\gamma\|_2^2 + 8u\big(f(\mathbf{x})-f(\mathbf{x}^*)\big)/\gamma^2.$$

Performing operator $\mathcal{L}$ on $\mathcal{A}(\mathbf{x},\mathbf{v}) = e^{\lambda\mathcal{E}(\mathbf{x},\mathbf{v})}$ gives

$$\begin{aligned}\mathcal{L}\mathcal{A} &= \langle\nabla_{\mathbf{x}}\mathcal{A},\mathbf{v}\rangle - \langle\nabla_{\mathbf{v}}\mathcal{A},\gamma\mathbf{v}+u\nabla f(\mathbf{x})\rangle + \langle\nabla_{\mathbf{v}}^2\mathcal{A},\gamma u\mathbf{I}\rangle\\ &= \lambda\mathcal{A}\big(\langle\nabla_{\mathbf{x}}\mathcal{E},\mathbf{v}\rangle - \langle\nabla_{\mathbf{v}}\mathcal{E},\gamma\mathbf{v}+u\nabla f(\mathbf{x})\rangle + \langle\nabla_{\mathbf{v}}^2\mathcal{E},\gamma u\mathbf{I}\rangle\big) + \lambda^2\mathcal{A}\gamma u\|\nabla_{\mathbf{v}}\mathcal{E}\|_2^2\\ &= \lambda\mathcal{A}\big(-4\|\mathbf{v}\|_2^2/\gamma - 4um\|\mathbf{x}\|_2^2/\gamma + 4u(2d+b)/\gamma\big) + \lambda^2\mathcal{A}\gamma u\big\|4\mathbf{x}/\gamma + 8\mathbf{v}/\gamma^2\big\|_2^2\\ &\le \lambda\mathcal{A}\big(-4\|\mathbf{v}\|_2^2/\gamma - 4um\|\mathbf{x}\|_2^2/\gamma + 4u(2d+b)/\gamma\big) + \lambda^2\mathcal{A}\gamma u\big(32\|\mathbf{x}\|^2/\gamma^2 + 128\|\mathbf{v}\|^2/\gamma^4\big\|_2^2,\end{aligned}$$

where the last inequality is by Young's inequality. Let

$$\lambda \le \min\bigg\{\frac{\gamma^2}{64u}, \frac{m}{16}\bigg\},$$

we get

$$\mathcal{L}\mathcal{A} \le \lambda\mathcal{A}\big(-2\|\mathbf{v}\|_2^2/\gamma - 2um\|\mathbf{x}\|_2^2/\gamma + 4u(2d+b)/\gamma\big). \tag{C.15}$$

Moreover, by Young's inequality we have

$$\mathcal{E}(\mathbf{x},\mathbf{v}) \le 5/2\|\mathbf{x}\|_2^2 + \frac{12}{\gamma^2}\|\mathbf{v}\|_2^2 + \frac{2uM}{\gamma^2}\big(3\|\mathbf{x}\|_2^2 + 6\|\mathbf{x}^*\|_2^2\big).$$

where we use the inequality

$$f(\mathbf{x}) - f(\mathbf{x}^*) \le \frac{M}{2}\|\mathbf{x}-\mathbf{x}^*\|_2^2 \le \frac{M}{4}\big(3\|\mathbf{x}\|_2^2 + 6\|\mathbf{x}^*\|_2^2\big).$$

Assume $\gamma^2 \le 4\mu M$, we have

$$\mathcal{E}(\mathbf{x},\mathbf{v}) \le \frac{12}{\gamma^2}\|\mathbf{v}\|_2^2 + \frac{16uM}{\gamma^2}\|\mathbf{x}\|_2^2 + \frac{12uM}{\gamma^2}\|\mathbf{x}^*\|_2^2. \tag{C.16}$$

Plugging the above into (C.15) gives

$$\mathcal{L}\mathcal{A} \le \lambda\mathcal{A}\bigg(-\frac{\gamma m}{8M}\mathcal{E} + 4u(2d+b)/\gamma + \frac{2um}{\gamma}\|\mathbf{x}^*\|_2^2\bigg).$$

Therefore, we have the following for the Hamiltonian Langevin dynamics 1.1,

$$\begin{aligned}\frac{\mathrm{d}\mathbb{E}[\mathcal{A}(\boldsymbol{X}_t,\boldsymbol{V}_t)]}{\mathrm{d}t} &= \mathbb{E}[\mathcal{L}\mathcal{A}(\boldsymbol{X}_t,\boldsymbol{V}_t)] \\ &\leq \mathbb{E}\bigg[\mathcal{A}(\boldsymbol{X}_t,\boldsymbol{V}_t)\bigg(-\frac{\gamma m}{8M}\log\big(\mathcal{A}(\boldsymbol{X}_t,\boldsymbol{V}_t)\big)+\frac{4\lambda u(2d+b)}{\gamma}+\frac{2\lambda u m}{\gamma}\|\mathbf{x}^*\|_2^2\bigg)\bigg].\end{aligned} \tag{C.17}$$

Note that $g(x) = x\log(x)$ is convex with respect to x, thus we have $\mathbb{E}[-A\log(A)] \leq -\log(\mathbb{E}[\mathcal{A}])\mathbb{E}[\mathcal{A}]$. Plugging this into (C.17) yields

$$\frac{\mathrm{d}\mathbb{E}[\mathcal{A}]}{\mathrm{d}t} \leq \mathbb{E}[\mathcal{A}]\bigg(-\frac{\gamma m}{8M}\log\big(\mathbb{E}[\mathcal{A}]\big)+\frac{4\lambda u(2d+b)}{\gamma}+\frac{2\lambda u m}{\gamma}\|\mathbf{x}^*\|_2^2\bigg), \tag{C.18}$$

where we abuse the notation $\mathcal{A}$ for simplification. Dividing $\mathbb{E}[\mathcal{A}]$ on both sides of (C.18) and rearranging terms give

$$\frac{\mathrm{d}\log(\mathbb{E}[\mathcal{A}])}{\mathrm{d}t} \leq -\frac{\gamma m}{8M}\log\big(\mathbb{E}[\mathcal{A}]\big)+\frac{4\lambda u(2d+b)}{\gamma}+\frac{2\lambda u m}{\gamma}\|\mathbf{x}^*\|_2^2.$$

This further lead to

$$\begin{aligned}\log(\mathbb{E}[\mathcal{A}(\boldsymbol{X}_t,\boldsymbol{V}_t)]) &\leq \log(\mathbb{E}[\mathcal{A}(\boldsymbol{X}_0,\boldsymbol{V}_0)])+\frac{16M\lambda u\big[4d+2b+m\|\mathbf{x}^*\|_2^2)\big]}{\gamma^2 m} \\ &= \lambda\mathcal{E}(\boldsymbol{X}_0,\boldsymbol{V}_0)+\frac{16M\lambda u\big[4d+2b+m\|\mathbf{x}^*\|_2^2)\big]}{\gamma^2 m}.\end{aligned} \tag{C.19}$$

Moreover, note that we have $\|\mathbf{a}\|_2^2+\|\mathbf{b}\|_2^2 \geq \|\mathbf{a}-\mathbf{b}\|_2^2/2$, therefore,

$$\mathcal{E}(\mathbf{x},\mathbf{v}) \geq \|\mathbf{x}\|_2^2/2+\|\mathbf{v}/\gamma\|_2^2. \tag{C.20}$$

Let $\gamma < \sqrt{2}$, we have $\mathcal{E}(\mathbf{x},\mathbf{v}) \geq (\|\mathbf{x}\|_2^2+\|\mathbf{v}\|_2^2)/2$. Thus

$$\begin{aligned}\log\Big(\mathbb{E}\big[e^{\lambda(\|\boldsymbol{X}_t\|_2^2+\|\boldsymbol{V}_t\|_2^2)}\big]\Big) &\leq \log\Big(\mathbb{E}\big[e^{2\lambda\mathcal{E}(\boldsymbol{X}_t,\boldsymbol{V}_t)}\big]\Big) \\ &\leq 2\lambda\mathcal{E}(\boldsymbol{X}_0,\boldsymbol{V}_0)+\frac{32M\lambda u\big[4d+2b+m\|\mathbf{x}^*\|_2^2)\big]}{\gamma^2 m},\end{aligned}$$

where the last inequality is obtained by replacing λ with 2λ in (C.19), and thus we require $\lambda \leq \min\{\gamma^2/(128u), m/32\}$. This completes the proof. □

C.4 Proof of Lemma B.6

Proof. By Assumption 3.1, we have

$$f(\mathbf{x}) \leq f(\mathbf{x}^*)+\frac{M}{2}\|\mathbf{x}-\mathbf{x}^*\|_2^2 \leq f(\mathbf{x}^*)+M\|\mathbf{x}\|_2^2+M\|\mathbf{x}^*\|_2^2,$$

where the second inequality is by Yong's inequality, which implies

$$f(\mathbf{x})+u^{-1}\gamma^2\|\mathbf{x}\|_2^2/4-f(\mathbf{x}^*)+M\|\mathbf{x}^*\|_2^2 \leq (M+u^{-1}\gamma^2/4)\|\mathbf{x}\|_2^2.$$

Divide both side by $(M+u^{-1}\gamma^2/4)/m$, and we have

$$\frac{m(f(\mathbf{x})+u^{-1}\gamma^2\|\mathbf{x}\|_2^2/4-f(\mathbf{x}^*)+-M\|\mathbf{x}^*\|_2^2)}{M+u^{-1}\gamma^2/4} \leq m\|\mathbf{x}\|_2^2.$$

According to Assumption 3.2, we have

$$\langle\nabla f(\mathbf{x}),\mathbf{x}\rangle \geq m\|\mathbf{x}\|_2^2-b \geq \frac{m(f(\mathbf{x})+u^{-1}\gamma^2\|\mathbf{x}\|_2^2/4}{M+u^{-1}\gamma^2/4}-\frac{f(\mathbf{x}^*)+M\|\mathbf{x}^*\|_2^2)}{M+u^{-1}\gamma^2/4}-b,$$

which directly completes the proof by dividing both side by 2. □

D Proof of additional lemmas

In this section we prove the additional supporting lemmas.

D.1 Proof of Lemma C.1

Proof. Applying Assumption 3.1 and noting that $\mathbf{x}_0 = \mathbf{0}$, we have

$$\|\nabla f_i(\mathbf{x})\|_2 \leq \|\nabla f_i(\mathbf{x}_0)\|_2 + M\|\mathbf{x} - \mathbf{x}_0\|_2 = \|\nabla f_i(\mathbf{0})\|_2 + M\|\mathbf{x}\|_2.$$

By setting $G = \max_{i\in[n]} \|\nabla f_i(\mathbf{0})\|_2$, we complete the proof. □

D.2 Proof of Lemma C.2

Proof. By the formula of $\mathbf{g}_k$, and let $k = jm + l$ denote the l-th iterate in the j-th epoch of Algorithm 1, we have

$$\begin{aligned}
\mathbb{E}[\|\mathbf{g}_{k+1} - \nabla f(\mathbf{x}_{k+1})\|_2^2] &= \mathbb{E}\bigg[\bigg\|\frac{1}{B}\bigg(\sum_{i\in\mathcal{B}_{k+1}}\big[\nabla f_i(\mathbf{x}_{k+1}) - \nabla f_i(\mathbf{x}_k)\big]\bigg) + \mathbf{g}_k - \nabla f(\mathbf{x}_{k+1})\bigg\|_2^2\bigg] \\
&= \mathbb{E}\bigg[\bigg\|\frac{1}{B}\bigg(\sum_{i\in\mathcal{B}_{k+1}}\big[\nabla f_i(\mathbf{x}_{k+1}) - \nabla f_i(\mathbf{x}_k)\big]\bigg) - \big(\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\big)\bigg\|_2^2\bigg] \\
&\quad + \mathbb{E}[\|\mathbf{g}_k - \nabla f(\mathbf{x}_k)\|_2^2].
\end{aligned}$$

By Lemma A.1 in [36], we know that

$$\begin{aligned}
&\mathbb{E}\bigg[\bigg\|\frac{1}{B}\bigg(\sum_{i\in\mathcal{B}_{k+1}}\big[\nabla f_i(\mathbf{x}_{k+1}) - \nabla f_i(\mathbf{x}_k)\big]\bigg) - \big(\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\big)\bigg\|_2^2\bigg] \\
&\leq \frac{1}{B}\mathbb{E}[\|\nabla f_i(\mathbf{x}_{k+1}) - \nabla f_i(\mathbf{x}_k)\|_2^2].
\end{aligned}$$

Thus, it follows that

$$\begin{aligned}
\mathbb{E}[\|\mathbf{g}_{k+1} - \nabla f(\mathbf{x}_{k+1})\|_2^2] &\leq \frac{1}{B}\mathbb{E}[\|\nabla f_i(\mathbf{x}_{k+1}) - \nabla f_i(\mathbf{x}_k)\|_2^2] + \mathbb{E}[\|\mathbf{g}_k - \nabla f(\mathbf{x}_k)\|_2^2] \\
&\leq \frac{1}{B}\sum_{s=jm}^{jm+l}\mathbb{E}[\|\nabla f_i(\mathbf{x}_{s+1}) - \nabla f_i(\mathbf{x}_s)\|_2^2] + \mathbb{E}[\|\mathbf{g}_{jm} - \nabla f(\mathbf{x}_{jm})\|_2^2] \\
&\leq \frac{M^2}{B}\sum_{s=jm}^{jm+l}\mathbb{E}[\|\mathbf{x}_{s+1} - \mathbf{x}_s\|_2^2] + \frac{1}{B_0}\mathbb{E}[\|\nabla f_i(\mathbf{x}_{jm})\|_2^2]\cdot\mathbb{1}(B_0 < n) \\
&\leq \frac{M^2}{B}\sum_{s=jm}^{jm+l}\mathbb{E}[\|\mathbf{x}_{s+1} - \mathbf{x}_s\|_2^2] + \frac{2}{B_0}\mathbb{E}[\|\mathbf{x}_{jm}\|_2^2 + G^2]\cdot\mathbb{1}(B_0 < n),
\end{aligned}$$

where the first inequality is by Young's inequality, the second inequality is by Assumption 3.1, the third inequality follows Lemma A.1 in [36], the last inequality is by Lemma C.1. This completes the proof. □

E Additional experimental results

In this section, we provide additional experimental results.

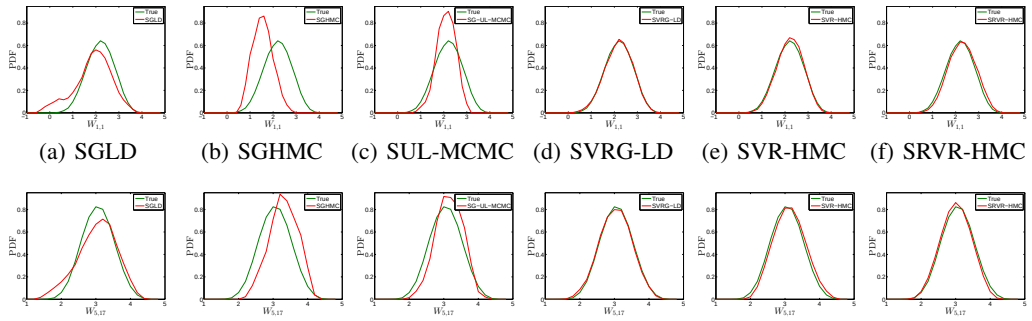
E.1 Comparison of posterior distributions

Here we conduct additional comparison in terms of sampled posterior distributions for ICA. In detail, we use HMC with metropolis hasting correction to generate the ground truth. Similar to [23], we randomly choose two variables ($W_{1,1}$ and $W_{5,17}$) from the parameter matrix $\mathbf{W}$ and display their

Table 2: Summary of datasets for Bayesian logistic classification.

Dataset	*pima*	*a3a*	*mushroom*	*a9a*
n (training)	600	3185	6000	32,561
n (test)	168	29376	2124	16281
d	8	123	122	123

marginal distributions after 1000 data passes in Figures 4(a)-4(f) (row 1) and Figures 4(g)-4(l) (row 2) respectively. It can be observed that the proposed SRVR-HMC (as well as SVRG-LD and SVR-HMC) can well approximate the ground truth, while SGLD and SGHMC cannot provide accurate approximation. This further validates the superior performance of SRVR-HMC and other variance reduced algorithms (SVRG-LD, SVR-HMC).



(a) SGLD (b) SGHMC (c) SUL-MCMC (d) SVRG-LD (e) SVR-HMC (f) SRVR-HMC

(g) SGLD (h) SGHMC (i) SUL-MCMC (j) SVRG-LD (k) SVR-HMC (l) SRVR-HMC

Figure 4: Marginal distributions of the posterior samples generated by Langevin dynamics based algorithms (red line) including SGLD, SGHMC, SG-UL-MCMC, SVRG-LD, SVR-HMC and SRVR-HMC, as well as the ground truth (green line). (Here we use SUL-MCMC to denote SG-UL-MCMC due to the space limit.)

E.2 ICA with larger dataset

We also ran additional experiments for ICA on a larger dataset (extract a larger subset from the original dataset, i.e., $n = 10000$), which is displayed in Figure 5. It can be seen that the proposed SRVR-HMC algorithm achieves the best performance among all methods.

E.3 Bayesian Logistic Regression

Assume we are given data $\{\mathbf{x}_i, y_i\}_{i=1,\ldots,n}$ where $\mathbf{x}_i$ denotes the feature vector and $y_i \in \{-1, 1\}$ denotes the corresponding label. Then the probability density function of the label y given the feature $\mathbf{x}_i$ and model vector $\boldsymbol{\beta}$ is modeled as $p(y|\mathbf{x}_i, \boldsymbol{\beta}) = 1/(1 + e^{-y_i \boldsymbol{\beta}^\top \mathbf{x}_i})$. We further assume the model vector $\boldsymbol{\beta}$ follows a Gamma prior $p(\boldsymbol{\beta}) \propto \|\boldsymbol{\beta}\|_2^{-\lambda} \exp(-\theta\|\boldsymbol{\beta}\|_2)$, where λ and $\boldsymbol{\beta}$ are fixed parameter. In the Bayesian logistic classification task, we aim to sample the posterior distribution

$$p(\boldsymbol{\beta}|\{\mathbf{x}_i, y_i\}_{i=1,\ldots,n}) = p(\boldsymbol{\beta}) \prod_{i=1}^{n} p(y_i|\mathbf{x}_i, \boldsymbol{\beta}).$$

Let $f(\boldsymbol{\beta}) = -\log p(\boldsymbol{\beta}|\{\mathbf{x}_i, y_i\}_{i=1,\ldots,n})$. Each function $f_i(\boldsymbol{\beta})$ in (1.3) takes the form of $f_i(\boldsymbol{\beta}) = -n \log(p(y_i|\mathbf{x}_i, \boldsymbol{\beta})) + \lambda \log(\|\boldsymbol{\beta}\|_2) + \theta\|\boldsymbol{\beta}\|_2$.

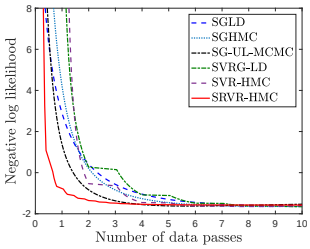
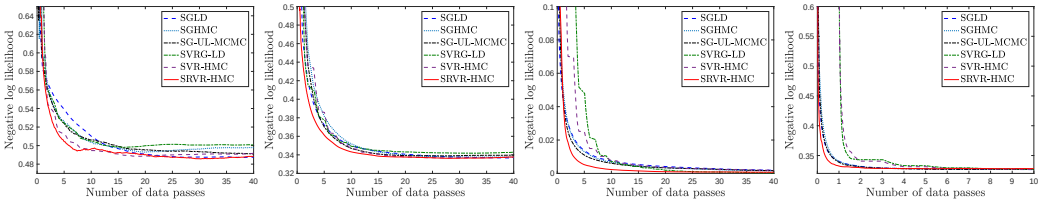


Figure 5: Results for ICA on a larger dataset (training sample size: $n = 10000$). X-axis represents the number of data passes and Y-axis represents the negative log likelihood on the test dataset.



(a) *pima* (b) *a3a* (c) *mushroom* (d) *a9a*

Figure 6: Comparison of different algorithms for Bayesian logistic regression, where Y-axis denotes the negative log likelihood on test datasets and X-axis denotes number of data passes.

We compare the performance of the proposed algorithm with SGLD [50], SGHMC [16], SG-UL-MCMC [18], SVRG-LD [25], and SVR-HMC [55] on *pima*, *a3a*, *mushroom*, and *a9a* dataset, which are available in UCI[5] [11] and LibSVM[6] [38] libraries. We summarize the detail of these datasets in Table 2. We run all algorithms on the training dataset, where the hyper parameters are tuned under the guidance of their theory. Moreover, we compute the sample path average of the position variable as the output. Then, such output is applied to conduct classification tasks on the test datasets, and we plot the negative log likelihood in Figures 6(a) - 6(d). It can be seen that the proposed algorithm SRVR-HMC outperforms all baseline algorithms on these four dataset, which is consistent with our theory.

[5] https://archive.ics.uci.edu/ml/

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