On the Convergence of Hamiltonian Monte Carlo with Stochastic Gradients

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

Abstract

Hamiltonian Monte Carlo (HMC), built based on the Hamilton's equation, has been witnessed great success in sampling from high-dimensional posterior distributions. However, it also suffers from computational inefficiency, especially for large training datasets. One common idea to overcome this computational bottleneck is using stochastic gradients, which only queries a mini-batch of training data in each iteration. However, unlike the extensive studies on the convergence analysis of HMC using full gradients, few works focus on establishing the convergence guarantees of stochastic gradient HMC algorithms. In this paper, we propose a general framework for proving the convergence rate of HMC with stochastic gradient estimators, for sampling from strongly log-concave and log-smooth target distributions. We show that the convergence to the target distribution in 2-Wasserstein distance can be guaranteed as long as the stochastic gradient estimator is unbiased and its variance is upper bounded along the algorithm trajectory. We further apply the proposed framework to analyze the convergence rates of HMC with four standard stochastic gradient estimators: mini-batch stochastic gradient (SG), stochastic variance reduced gradient (SVRG), stochastic average gradient (SAGA), and control variate gradient (CVG). Theoretical results explain the inefficiency of mini-batch SG, and suggest that SVRG and SAGA perform better in the tasks with high-precision requirements, while CVG performs better for large dataset. Experiment results verify our theoretical findings.

1. Introduction

Monte Carlo Markov Chain (MCMC) methods have been witnessed great success in many machine learning applications such as Bayesian inference, reinforcement learning, and computer vision. In the past decades, many MCMC algorithms, such as random walk Metropolis (Mengersen et al., 1996), ball walk (Lovász & Simonovits, 1990), hit and run (Smith, 1984), Langevin dynamics (LD) based algorithms (Langevin, 1908; Parisi, 1981), and Hamiltonian Monte Carlo (HMC) (Duane et al., 1987), have been invented and studied. Among them HMC has been recognized as the most effective MCMC algorithm due to its rapid mixing rate and small discretization error. In practice, HMC has been deployed as the default sampler in many open packages such as Stan (Carpenter et al., 2017) and Tensorflow (Abadi et al., 2016). In specific, HMC simulates the trajectory of a particle in the Hamiltonian system, which is described by the following Hamilton's equation

$$\begin{aligned} \frac{\mathrm{d}\mathbf{q}(t)}{\mathrm{d}t} &= \frac{\partial H(\mathbf{q}(t), \mathbf{p}(t))}{\partial \mathbf{p}} \\ \frac{\mathrm{d}\mathbf{p}(t)}{\mathrm{d}t} &= -\frac{\partial H(\mathbf{q}(t), \mathbf{p}(t))}{\partial \mathbf{q}}, \end{aligned} \tag{1.1}$$

where $\mathbf{q}$ and $\mathbf{p}$ are position and momentum variables, and $H(\mathbf{q}, \mathbf{p})$ is the so-called Hamiltonian function, which is typically defined as the sum of the potential energy $f(\mathbf{q})$ and the kinetic energy $\|\mathbf{p}\|_2^2/2$. In each step, HMC solves (1.1) using the sample generated in the last step as the initial position $\mathbf{q}(0)$ and an independently generated Gaussian random vector as the initial momentum $\mathbf{p}(0)$, then outputs the solution at a certain time τ, i.e., $\mathbf{q}(\tau)$, as the next sample. It is well known that if the Hamilton's equation can be exactly solved and the potential energy function $f(\mathbf{x})$ admits certain good properties, the sample sequence generated by HMC asymptotically converges to the target distribution $\pi \propto \exp(-f(\mathbf{x}))$ (Lee et al., 2018; Chen & Vempala, 2019). However, it is generally intractable to exactly solve (1.1), and numerical integrators are needed to solve it approximately. One of the most popular HMC algorithms adopts the leapfrog integrator for solving (1.1) following by a Metropolis-Hasting (MH) correction step (Neal et al., 2011). Aside from the algorithmic development, the convergence rate of HMC has also been extensively studied in recent literature (Bou-Rabee et al., 2018; Lee et al., 2018; Mangoubi & Smith, 2017; Durmus et al., 2017; Mangoubi & Vishnoi, 2018; Chen & Vempala, 2019; Chen et al., 2019), which demonstrate its superior performance compared with other MCMC methods.

[1]Department of Computer Science, UCLA. Correspondence to: Quanquan Gu <qgu@cs.ucla.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

However, as the data size grows rapidly nowadays, the standard HMC algorithm suffers from huge computational cost. For a Bayesian inference problem, the energy function $f(\mathbf{x})$ (a.k.a., negative log-posterior in Bayesian learning problem) is formulated as the sum of the negative log-likelihood functions over all observations (i.e., $f(\mathbf{x}) = \sum_{i=1}^{n} f_i(\mathbf{x})$). When the number of observations (i.e., n) becomes extremely large, the standard HMC algorithm may fail as it requires to query the entire dataset to compute the full gradient $\nabla f(\mathbf{x})$. To overcome the computational burden, a common idea is to leverage stochastic gradient in each update, i.e., we only compute the gradient approximately using a mini-batch of training data, which gives rise to stochastic gradient HMC (SG-HMC) (Chen et al., 2014)[1]. This idea has also triggered a bunch of work focusing on improving the scalability of other gradient-based MCMC methods (Welling & Teh, 2011; Chen et al., 2015; Ma et al., 2015; Baker et al., 2018). Despite the efficiency improvements for large-scale Bayesian inference problems, SG-HMC has many drawbacks. The variance of stochastic gradients may lead to inaccurate solutions to the Hamilton's equation (1.1). Additionally, it is no longer tractable to perform MH correction step since (1) the proposal distribution of HMC does not have an explicit formula and is not time-reversible, and (2) one cannot exactly query the entire training dataset to compute the MH acceptance probability. These two shortcomings prevent SG-HMC from achieving as accurate sampling as the standard HMC (Betancourt, 2015; Bardenet et al., 2017; Dang et al., 2019), and hurdle the application of SG-HMC in many sampling tasks with a high-precision requirement.

Despite the pros and cons of SG-HMC discussed in the aforementioned works, most of them are empirical studies. Unlike stochastic gradient Langevin dynamics (SGLD) and stochastic gradient underdamped Langevin dynamics (SGULD)[2] that have been extensively studied in theory, little work has been done to provide a theoretical understanding of SG-HMC. It remains illusive whether SG-HMC can be guaranteed to converge and how it performs in different regimes. Moreover, there has also emerged many other stochastic gradient estimators that exhibit smaller variance than the standard mini-batch stochastic gradient estimator, such as stochastic variance reduced gradient (SVRG) (Johnson & Zhang, 2013), stochastic averaged gradient (SAG) (Defazio et al., 2014), and control variates gradient (CVG) (Baker et al., 2018). These stochastic gradient estimators have been successfully incorporated into Langevin based algorithms for faster sampling (Dubey et al., 2016; Chatterji et al., 2018; Baker et al., 2018; Brosse et al., 2018; Zou et al., 2018a; Li et al., 2018). It is unclear whether these estimators can be adapted in the HMC algorithm to overcome the drawbacks of SG-HMC.

In this paper, we propose a general framework for proving the convergence rate of HMC with stochastic gradients for sampling from strongly log-concave and log-smooth distributions. At the core of our analysis is a sharp characterization of the solution to the Hamilton's equation (1.1) obtained using stochastic gradients, which is in the order of $O(\sqrt{\eta})$, where η is the step size of the numerical integrator. Under the proposed proof framework, the convergence rate of HMC with a variety of stochastic gradient estimators can be derived. We summarize the main contributions of our paper as follows:

- We develop a general framework for characterizing the convergence rate of HMC algorithm when using stochastic gradients. In particular, we prove that as long as the stochastic gradient is unbiased, and its variance along the algorithm trajectory is upper bounded (not require a uniform upper bound), the stochastic gradient HMC algorithm provably converges to the target distribution $\pi \propto \exp(-f(\mathbf{x}))$ in 2-Wasserstein distance with a sampling error up to $O(\sqrt{\eta})$.
- We apply four commonly used stochastic gradient estimators to the HMC algorithm for sampling from the target distribution of form $\pi \propto \exp(-\sum_{i=1}^{n} f_i(\mathbf{x}))$, which gives rise to four variants of stochastic gradient HMC algorithms, including SG-HMC, SVRG-HCM, SAGA-HMC, and CVG-HMC. We establish their convergence guarantees under the proposed framework. Our analysis suggests that in order to achieve $\epsilon/\sqrt{n}$-sampling error in 2-Wasserstein distance, the gradient complexity[3] of SG-HMC, CVG-HMC, SVRG-HMC and SAGA-HMC are $\widetilde{O}(n/\epsilon^2)$, $\widetilde{O}(1/\epsilon^2 + 1/\epsilon)$, $\widetilde{O}(n^{2/3}/\epsilon^{2/3} + 1/\epsilon)$, and $\widetilde{O}(n^{2/3}/\epsilon^{2/3} + 1/\epsilon)$ respectively. This explains the inefficiency of SG-HMC observed in prior work, and reveals the prospects of CVG-HMC, SVRG-HMC and SAGA-HMC for large-scale sampling problems.
- We carry out numerical experiments on both synthetic and real-world dataset. The results show all stochastic gradient HMC algorithms converge but SG-HMC has a significantly larger bias compared with other algorithms. Additionally, SVRG-HMC performs the best when the sample size is small while CVG-HMC becomes more efficient and effective when the sample

[1] In fact, in addition to making use of stochastic gradients, Chen et al. (2014) also introduces a friction term and an additional Brownian term to mitigate the bias and variance brought by stochastic gradients.

[2] In some existing works this algorithm is also referred to as SGHMC (Zou et al., 2018a; Gao et al., 2018b;a). We highlight that their SGHMC algorithm is different from the SG-HMC algorithm studied in this paper, which we will clearly discuss in Section 2.

[3] The gradient complexity is defined by the number of stochastic gradient evaluations to achieve the target accuracy.

size increases. This well corroborates our theoretical findings.

Notation. Given two scalars a and b, we use $a \wedge b$ to denote $\min\{a, b\}$ and use $a \vee b$ to denote $\max\{a, b\}$. Given a vector $\mathbf{x} \in \mathbb{R}^d$, we define by $\|\mathbf{x}\|_2 = \sqrt{x_1^2 + \cdots + x_d^2}$ its Euclidean norm. We use $O(\cdot)$ and $\Omega(\cdot)$ notations to hide constant factors and use $\widetilde{O}(\cdot)$ to hide the poly-logarithmic factors in $O(\cdot)$ notation. Given two sequences $\{x_k\}$ and $\{y_k\}$, we further define $x_k = \Theta(y_k)$ if $x_k = \Omega(y_k)$ and $x_k = O(y_k)$. We Given two distributions μ and ν, the 2-Wasserstein distance is defined by $\mathcal{W}_2^2(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \int \|\mathbf{x} - \mathbf{y}\|_2^2 \mathrm{d}\gamma(\mathbf{x}, \mathbf{y})$.

2. Additional Related Work: Langevin Dynamics Based Algorithms

2.1. Mathematical Description of Langevin Dynamics.

Aside from the HMC algorithm, another important family of MCMC methods is built upon the Langevin dynamics, including both overdamped Langevin dynamics (LD) (Roberts & Tweedie, 1996) and underdamped Langevin dynamics (ULD) (Chen et al., 2017). Formally, the overdamped Langevin dynamics can be described by the following stochastic differential equation (SDE):

$$\mathrm{d}\mathbf{x}_t = -\nabla f(\mathbf{x}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t, \tag{2.1}$$

where $\mathbf{B}_t$ is the Brownian term. The underdamped Langevin dynamics takes the form of the following SDE,

$$\begin{aligned} \mathrm{d}\mathbf{v}_t &= -\gamma \mathbf{v}_t \mathrm{d}t - u\nabla f(\mathbf{x}_t) + \sqrt{2\gamma u}\mathrm{d}\mathbf{B}_t \\ \mathrm{d}\mathbf{x}_t &= \mathbf{v}_t \mathrm{d}t, \end{aligned} \tag{2.2}$$

where $\mathbf{x}_t$ and $\mathbf{v}_t$ are the position and velocity variables at time t respectively, $\gamma > 0$ is the "friction" parameter, and $u > 0$ is referred to the "inverse mass" parameter. Notably, both overdamped and underdamped Langevin dynamics converges to the (marginal) stationary distribution $\pi \propto e^{-f(\mathbf{x})}$. The goal of Langevin dynamics based algorithms is to approximately solve the SDEs in (2.1) and (2.2). As a comparison, the focus of HMC based algorithms is to solve the ODE (1.1), which could be done more accurately or even exactly.

2.2. Existing Convergence Results of Langevin Dynamics Based Algorithms

The convergence rate of Langevin dynamics based algorithms have been widely studied for various machine learning problems such as sampling (Chen et al., 2015; Li et al., 2016; Dubey et al., 2016; Chen et al., 2017; Dalalyan & Karagulyan, 2019; Li et al., 2018; Cheng et al., 2018; Zou et al., 2018b;a; Shen & Lee, 2019; Zou et al., 2019; Dalalyan et al., 2020; Simsekli et al., 2020) and nonconvex optimization (Raginsky et al., 2017; Zhang et al., 2017; Xu et al., 2018; Ma et al., 2018; Gao et al., 2018a;b; Chau & Rasonyi, 2019; Deng et al., 2020; Zou et al., 2020). Among them, the most relevant works to this paper are focusing on establishing the convergence rate of Langevin dynamics based algorithm for sampling from strongly log-concave and log-smooth distributions (Dalalyan & Karagulyan, 2019; Dalalyan, 2017; Chen et al., 2017; Baker et al., 2018; Zou et al., 2018a; Chatterji et al., 2018). In particular, based on the overdamped Langevin dynamics, Dalalyan & Karagulyan (2019); Dalalyan (2017) established the convergence guarantee of Langevin Monte Carlo (LMC, Euler discretization of (2.1) using full gradient) and stochastic gradient Langevin dynamics (SGLD, Euler discretization of (2.1) using stochastic gradients). Zou et al. (2018b) further showed that using SVRG or subsampled SVRG gradient estimator can help improve the convergence rate for both LMC and SGLD. Baker et al. (2018) proposed to use a control-variate gradient estimator in SGLD and also demonstrated its efficiency in terms of the sample size. Based on the underdamped Langevin dynamics, Chen et al. (2015) showed that using naive Euler discretization on (2.2) cannot give faster convergence rate than LMC/SGLD and instead one may need to use a high-order discretization mechanism. However, Chen et al. (2017) showed that part of (2.2) can be solved analytically and proposed an accurate first-order discretization method for solving (2.2), which provably achieves faster convergence rate than LMC. Following this line of research, Zou et al. (2018a); Chatterji et al. (2018) considered using the SVRG and CVG estimators in the algorithm developed by Chen et al. (2017), which can also help reduce the discretization error and thus lead to faster convergence rates.

2.3. Comparison between Langevin Dynamics and HMC based algorithms

We would like to highlight that these two types of algorithms (especially ULD based algorithms vs. HMC based algorithms) are different in terms of both algorithm designs and their underlying SDE/ODE (see (1.1) and (2.2) for their formulas). In particular, HMC-based algorithms focus on solving the Hamilton's equation (which is an ODE) in each proposal while ULD based algorithms are derived from the discretization of an SDE. From the algorithmic perspective, HMC based algorithms have a double loop structure: the inner loop solves the ODE and makes a proposal, the outer loop updates the proposals until convergence. ULD-based algorithms exhibit a single loop structure and are designed as a discretization of the underlying SDE.

To better position our algorithms and results, we also summarize the gradient complexities of the HMC based algorithms and Langevin dynamics based algorithms in Table

Table 1. Comparison of different stochastic sampling algorithms, where the target distribution is $\pi \propto e^{-\sum_{i=1}^n f_i(\mathbf{x})}$ and the target sampling error is $\epsilon/\sqrt{n}$ in 2-Wasserstein distance. Besides, the gradient complexities for SG-ULD, CV-ULD, SGLD, SVRG-LD, and SAGA-LD are derived in Chatterji et al. (2018), the gradient complexity of SVRG-ULD is derived in Zou et al. (2018a).

Algorithm	Complexity	Type
SGLD (Welling & Teh, 2011)	$\widetilde{O}\big(\frac{n}{\epsilon^2}\big)$	LD
SVRG-LD (Dubey et al., 2016)	$\widetilde{O}\big(\frac{n}{\epsilon}\big)$	LD
SAGA-LD (Dubey et al., 2016)	$\widetilde{O}\big(\frac{n}{\epsilon}\big)$	LD
SG-ULD (Chen et al., 2017)	$\widetilde{O}\big(\frac{n}{\epsilon^2}\big)$	ULD
SVRG-ULD (Zou et al., 2018a)	$\widetilde{O}\big(\frac{n^{2/3}}{\epsilon^{2/3}} + \frac{1}{\epsilon}\big)$	ULD
CV-ULD (Chatterji et al., 2018)	$\widetilde{O}\big(\frac{1}{\epsilon^3}\big)$	ULD
SG-HMC	$\widetilde{O}\big(\frac{n}{\epsilon^2}\big)$	HMC
SVRG-HMC	$\widetilde{O}\big(\frac{n^{2/3}}{\epsilon^{2/3}} + \frac{1}{\epsilon}\big)$	HMC
SAGA-HMC	$\widetilde{O}\big(\frac{n^{2/3}}{\epsilon^{2/3}} + \frac{1}{\epsilon}\big)$	HMC
CVG-HMC	$\widetilde{O}\big(\frac{1}{\epsilon^2}\big)$	HMC

1, where SG-ULD, SVRG-ULD, CVG-ULD are referred to as the algorithms that apply the discretization approach in (Chen et al., 2017) on (2.2) using SG, SVRG, and CVG estimators respectively; SAGA-LD and SVRG-LD are referred to as the algorithms that apply Euler discretization on (2.2) using SVRG and SAGA estimators respectively.

Here we calibrate the bounds proved in the original papers to fit into our setting (i.e., the target distribution is $\pi \propto e^{-\sum_{i=1}^n f_i(\mathbf{x})}$ and the target sampling error is $\epsilon/\sqrt{n}$)[4]. First, when comparing between different HMC based algorithms, it can be seen that SG-HMC has the worst dependency on both dataset size n and accuracy parameter ϵ. Moreover, if the dataset size satisfies $n = \Omega(\epsilon^{-2})$, CVG-HMC enjoys better gradient complexity than both SVRG-HMC and SAGA-HMC, which suggests that CVG-HMC performs better for very large datasets. On the other hand, if $\epsilon = O(n^{-1/2})$, SVRG-HMC and SAGA-HMC will outperform CVG-HMC, implying that SAGA-HMC and SVRG-HMC are better for sampling with high-precision requirements. Then we will compare the HMC based algorithms with Langevin dynamics based algorithms. Clearly, it can be seen that when using the same stochastic gradient estimator, the complexity bound of SG-HMC is the same as those of SG-ULD and SGLD, the complexity bounds of SVRG-HMC and SAGA-HMC are the same as that of SVR-ULD and better than those of SAGA-LD and SVRG-LD, and the complexity bound of CVG-HMC is better than that of CV-ULD.

[4]This is to make the Wasserstein metric invariant under different scalings (Lee et al., 2018). Some of related work consider averaged log-likelihood functions (i.e., $\pi \propto \exp(-n^{-1}\sum_{i=1}^n f_i(\mathbf{x}))$) and set the target sampling error as ϵ.

3. HMC with Stochastic Gradients

Let $H(\mathbf{q},\mathbf{p}) = f(\mathbf{q}) + \|\mathbf{p}\|_2^2/2$ be the Hamiltonian function, then the Hamilton's equation (3.1) can be formulated as.

$$\frac{\mathrm{d}\mathbf{q}(t)}{\mathrm{d}t} = \mathbf{p}(t), \quad \frac{\mathrm{d}\mathbf{p}(t)}{\mathrm{d}t} = -\nabla f(\mathbf{q}(t)). \tag{3.1}$$

Let $\{\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(t)}, \ldots\}$ be the sequence of generated samples by HMC. Given $\mathbf{x}^{(t)}$, an idealized HMC generates the next sample $\mathbf{x}^{(t+1)}$ by solving the differential equation (3.1) at a certain time τ (i.e., $\mathbf{q}(\tau)$) with initial position $\mathbf{q}(0) = \mathbf{x}^{(t)}$ and initial momentum $\mathbf{p}(0) \sim N(\mathbf{0}, \mathbf{I})$ being independently drawn from the standard Gaussian distribution. In practice, one typically applies the leapfrog numerical integrator to solve (3.1). In particular, the numerical integrator first divides the time interval $[0,\tau]$ into K subintervals with length equaling to $\eta = \tau/K$. Let $\mathbf{p}_0 = \mathbf{p}(0)$ and $\mathbf{q}_0 = \mathbf{q}(0)$, the one-step second-order leapfrog update for $(\mathbf{q}_k, \mathbf{p}_k)$ is defined as follows

$$\begin{aligned} \mathbf{p}_{k+1/2} &= \mathbf{p}_k - \frac{\eta}{2}\nabla f(\mathbf{q}_k) \\ \mathbf{q}_{k+1} &= \mathbf{q}_k + \eta \mathbf{p}_{k+1/2} \\ \mathbf{p}_{k+1} &= \mathbf{p}_{k+1/2} - \frac{\eta}{2}\nabla f(\mathbf{q}_{k+1}). \end{aligned} \tag{3.2}$$

Similarly, stochastic gradient HMC can be designed by replacing the full gradient $\nabla f(\mathbf{q}_k)$ and $\nabla f(\mathbf{q}_{k+1})$ with stochastic gradient estimators. Let $\mathbf{g}(\mathbf{q},\boldsymbol{\xi})$ be an unbiased stochastic gradient estimator of $\nabla f(\mathbf{q})$, where $\boldsymbol{\xi}$ represents randomness. Then reformulating (3.2) and replacing $\nabla f(\mathbf{q}_k)$ and $\nabla f(\mathbf{q}_{k+1})$ with stochastic gradients yields

$$\begin{aligned} \mathbf{q}_{k+1} &= \mathbf{q}_k + \eta \mathbf{p}_k - \frac{\eta^2}{2}\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) \\ \mathbf{p}_{k+1} &= \mathbf{p}_k - \frac{\eta}{2}\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) - \frac{\eta}{2}\mathbf{g}(\mathbf{q}_{k+1}, \boldsymbol{\xi}_{k+1/2}), \end{aligned} \tag{3.3}$$

where η is the step size of leapfrog integrator and the randomness $\boldsymbol{\xi}_k$ and $\boldsymbol{\xi}_{k+1/2}$ are independent. Here we use $\boldsymbol{\xi}_{k+1/2}$ rather than $\boldsymbol{\xi}_{k+1}$ to guarantee that the randomness at two subsequent leapfrog updates are independent. We summarize the entire algorithm in Algorithm 1. $\mathbf{q}_K$ can be seen as an approximate solution to (3.1), which will be passed to the next proposal of HMC.

4. General Convergence Results for Stochastic Gradient HMC Algorithms

In this section, we will present the general theoretical results on the convergence rate of HMC with stochastic gradients. Before presenting the main theory, we first make the following assumptions on the potential energy function $f(\cdot)$ and the stochastic gradient estimator $\mathbf{g}(\cdot,\cdot)$.

Assumption 4.1 (Strongly Convex)**.** There exists a positive

Algorithm 1 Noisy Gradient Hamiltonian Monte Carlo

input: Step size η, number of leapfrog steps K, number of HMC proposals T, initial point $\mathbf{x}^{(0)}$
for $t = 0, \ldots, T$ **do**
  Set $\mathbf{q}_0 = \mathbf{x}^{(t)}$
  Sample $\mathbf{p}_0$ from $N(\mathbf{0}, \mathbf{I})$
  for $k = 0, \ldots, K-1$ **do**
    $\mathbf{q}_{k+1} = \mathbf{q}_k + \eta \mathbf{p}_k - \frac{\eta^2}{2}\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k)$
    $\mathbf{p}_{k+1} = \mathbf{p}_k - \frac{\eta}{2}\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) - \frac{\eta}{2}\mathbf{g}(\mathbf{q}_{k+1}, \boldsymbol{\xi}_{k+1/2})$
  end for
  Set $\mathbf{x}^{(t+1)} = \mathbf{q}_K$
end for
output: $\mathbf{x}^{(T)}$

constant μ such that for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|_2^2.$$

Assumption 4.2 (Smoothness)**.** There exists a positive constant L such that for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

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

The above two assumptions on the target distribution are commonly made in the convergence analysis of sampling algorithms (Chen et al., 2017; Erdogdu et al., 2018; Chen et al., 2019; Dalalyan & Karagulyan, 2019).

Assumption 4.3 (Bounded Variance)**.** For any $\mathbf{q}_k$ in the leapfrog update, the variance of the unbiased stochastic gradient estimator $\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k)$ is upper bounded by

$$\mathbb{E}\big[\|\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) - \nabla f(\mathbf{q}_k)\|_2^2\big] \leq \sigma^2,$$

where the expectation is taken over both $\mathbf{q}_k$ and $\boldsymbol{\xi}_k$.

We remark that while this assumption is made on the algorithm path which we will verify it for several widely used stochastic gradient estimators in Section 5. In other words, this assumption is only needed for the general convergence analysis. When specializing to a specific algorithm with certain stochastic gradient estimator, it can be proved.

Now we are ready to present our main result, which characterizes the convergence rate of Algorithm 1 for sampling from strongly log-concave and log-smooth distribution π.

Theorem 4.4. Under Assumptions 4.1, 4.2, and 4.3, let $\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x})$, $D = \|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2$, and μ_t be the distribution of the iterate $\mathbf{x}^{(t)}$, if the step size satisfies $\eta = O(L^{1/2}\sigma^{-2}\kappa^{-1} \wedge L^{-1/2})$ and $K = 1/(4\sqrt{L}\eta)$, the output of Algorithm 1 satisfies

$$\mathcal{W}_2(\mu_T, \pi) \leq \big(1 - (128\kappa)^{-1}\big)^{T/2}(2D + 2d/\mu)^{1/2} + \Gamma_1\eta^{1/2} + \Gamma_2\eta,$$

where $\kappa = L/\mu$ is the condition number of $f(\mathbf{x})$ and the constants Γ_1 and Γ_2 satisfy,

$$\Gamma_1^2 = O\big(L^{-3/2}\sigma^2\kappa^2\big)$$
$$\Gamma_2^2 = O\big(\kappa^2\big(LD + \kappa d + L^{-1/2}\sigma^2\eta\big)\big).$$

Remark 4.5. Theorem 4.4 provides a general convergence result for HMC with stochastic gradients. The first term in the upper bound represents the mixing of HMC, the second term represents the error brought by the variance of stochastic gradients, and the last term is mainly from the discretization error of the numerical integrator. Clearly, stochastic gradient HMC provably converges with a sampling error governed by the step size and the variance of stochastic gradients along the Markov chain (in a rate of $O(\sigma\eta^{1/2})$). In order achieve the target sampling error ϵ, one may have to choose a sufficiently small step size η, and run $T = O(\log(1/\epsilon))$ HMC proposals for mixing. Then the total number of iterations in Algorithm 1 is $TK = \widetilde{O}(L^{-1/2}\eta^{-1})$.

Note that the sampling error proved in Theorem 4.4 depends on the magnitude of the variance ($\Gamma_1 \sim \sqrt{\sigma^2}$), and different stochastic gradient estimators may lead to different convergence rates. Intuitively speaking, smaller variance leads to smaller parameter Γ_1, implying that we can use larger step size to achieve the same accuracy. This in turn speeds up the convergence of HMC since the iteration complexity is proportion to η^{-1}. In the next section, we will apply the general convergence result in Theorem 4.4 to some specific stochastic gradient estimators and establish the corresponding convergence guarantees.

5. Application to Commonly Used Stochastic Gradient Estimators

Note that given n observations, the target distribution can be described as $\pi \propto \exp(-f(\mathbf{x}))$ with $f(\mathbf{x}) = \sum_{i=1}^n f_i(\mathbf{x})$, where $f_i(\cdot)$ corresponds to the i-th observation. In each step, we will only query a subset of observations to estimate the gradient and update the variables accordingly. In this section, we will prove the convergence rates of HMC with four commonly used stochastic gradient estimators, including mini-batch stochastic gradient (SG), stochastic variance reduced gradient (SVRG) (Johnson & Zhang, 2013), stochastic average gradient (SAGA) (Defazio et al., 2014), and control variates gradient (CVG) (Baker et al., 2018). We list these four stochastic gradient estimators in Algorithm 2.

5.1. Review of Stochastic Gradient Estimators

The mini-batch stochastic gradient samples a mini-batch of training examples $\mathcal{I}_k$ of size $|\mathcal{I}_k| = B$ to compute the stochastic gradient and is identical to that used in SGLD (Welling & Teh, 2011). The SVRG and SAGA estimators

follow from (Dubey et al., 2016). SVRG estimator adopts a reference gradient $\nabla f(\widetilde{\mathbf{q}})$ associate with a reference point $\widetilde{\mathbf{q}}$, both of which are updated in a low frequency (updated every N leapfrog steps). In each update, we will sample a fresh mini-batch of training examples and leverage $\nabla f(\widetilde{\mathbf{q}})$ and $\widetilde{\mathbf{q}}$ as control variate to help reduce the variance. SAGA estimator maintains a table $\mathbf{G}$ that stores all stochastic gradients $\{f_i(\mathbf{x})\}_{k=1,\ldots,n}$. In each iteration, it queries a mini-batch of training examples $\mathcal{I}_k$ and computes the stochastic gradient by combining the mini-batch stochastic gradients on the new examples and the most recent history gradients in the table, including the stochastic gradients for new examples $\{\mathbf{G}_i\}_{i\in\mathcal{I}_k}$ and the sum of all stochastic gradients in the table (i.e., $\widetilde{\mathbf{g}}_k = \sum_{i=1}^n \mathbf{G}_i$). Afterward, the newly computed mini-batch stochastic gradient will be used to update the table. Similar to the SVRG estimator, CVG estimator also maintains a reference point $\widehat{\mathbf{q}}$, which is typically set to be an approximate minimizer of the function $f(\mathbf{x})$, and queries a new mini-batch of training examples to compute the stochastic gradient jointly. Different from the SVRG estimator that slowly updates the reference point, the reference point $\widehat{\mathbf{q}}$ adopted in CVG is fixed during the entire algorithm.

Algorithm 2 Stochastic Gradient Estimators

input: Current point $\mathbf{q}_k$, index of the HMC proposal t, random sampled mini-batch $\mathcal{I}_k$
Mini-batch Stochastic gradient
$\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) = \frac{n}{B}\sum_{i\in\mathcal{I}_k}\nabla f_i(\mathbf{q}_k)$
Stochastic variance reduced gradient
if $k + Kt \mod N = 0$ **then**
  $\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) = f(\mathbf{q}_k),\ \widetilde{\mathbf{q}} = \mathbf{q}_k$
else
  $\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) = \frac{n}{B}\sum_{i\in\mathcal{I}_k}\big[\nabla f_i(\mathbf{q}_k) - \nabla f_i(\widetilde{\mathbf{q}})\big] + f(\widetilde{\mathbf{q}})$
end if
Stochastic averaged gradient
if $k + Kt = 0$ **then**
  $\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) = \nabla f(\mathbf{q}_k),\ \mathbf{G} = \{\nabla f_i(\mathbf{q}_k)\}_{i=1,\ldots,n}$
else
  $\widetilde{\mathbf{g}}_k = \sum_{i=1}^n \mathbf{G}_i,\ \mathbf{G}_i \leftarrow \nabla f_i(\mathbf{q}_k)$ for all $i \in \mathcal{I}_k$,
  $\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) = \frac{n}{B}\sum_{i\in\mathcal{I}_k}\big[\nabla f_i(\mathbf{q}_k) - \mathbf{G}_i\big] + \widetilde{\mathbf{g}}_k,$
end if
Control variate gradient
$\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) = \nabla f(\widehat{\mathbf{q}}) + \frac{n}{B}\sum_{i\in\mathcal{I}_k}[\nabla f_i(\mathbf{q}_k) - \nabla f_i(\widehat{\mathbf{q}})]$
output: $\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k)$

5.2. Convergence Results of Specific Stochastic Gradient HMC Algorithms

Note that the convergence guarantee of stochastic gradient HMC in Theorem 4.4 is established based on Assumption 4.3. Therefore, in order to prove the convergence rates for HMC equipped with the aforementioned stochastic gradient estimators, it suffices the verify Assumption 4.3 and characterize the magnitude of the variance parameter σ. In the subsequent analysis, we will use a stronger version of Assumption 4.2 by requiring all component functions $\{f_i(\mathbf{x})\}_{i=1}^n$ are L/n-smooth.

Assumption 5.1. For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $i \in [n]$, there exists a positive constant L such that

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

This Assumption has also been made in many prior works (Baker et al., 2018; Chatterji et al., 2018; Brosse et al., 2018) for studying the convergence of stochastic gradient Langevin MCMC algorithms. Note that Assumption 5.1 immediately implies Assumption 4.2 and thus the result in Theorem 4.4 applies. We would also like to point out that we only need all component functions to be smooth but not necessarily to be strongly convex. Additionally, we follow the similar setting in Baker et al. (2018); Chatterji et al. (2018) that assumes L/n and μ/n are in the constant order, which implies that $L, \mu = O(n)$

By combining Algorithm 1 and the corresponding stochastic gradient estimator presented in Algorithm 2, we can obtain four specific stochastic gradient HMC algorithms, namely SG-HMC, SVRG-HMC, SAGA-HMC and CVG-HMC. We assume that the initial point $\mathbf{x}^{(0)}$ satisfies $\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2 \leq d/\mu$. Note that this can be achieved by running SGD for roughly $O(n)$ steps (Baker et al., 2018; Brosse et al., 2017). In the sequel, we will provide the convergence guarantees for these four algorithms.

Mini-batch stochastic gradient HMC (SG-HMC). The following theorem characterizes the convergence results of SG-HMC in 2-Wasserstein distance.

Theorem 5.2. Under Assumptions 4.1 and 5.1, assume $\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2 \leq d/\mu$ and let μ_t be the distribution of $\mathbf{x}^{(t)}$, then if the step size satisfies $\eta = O(L^{-1/2} \wedge d\mu^{-1}\Gamma_1^{-1})$ and set $K = 1/(4\sqrt{L}\eta)$, the output of SG-HMC satisfies

$$\mathcal{W}_2\big(\mu_T, \pi\big) \leq 2\sqrt{\frac{d}{\mu}}\big(1 - (128\kappa)^{-1}\big)^{T/2} + \Gamma_1\eta^{1/2} + \Gamma_2\eta,$$

where the constants Γ_1 and Γ_2 satisfy,

$$\Gamma_1^2 = O\big(L^{-1/2}B^{-1}\kappa^3 d + L^{-3/2}B^{-1}\kappa^2 n^2 d\big)$$
$$\Gamma_2^2 = O\big(\kappa^3 d + L^{-1/2}B^{-1}n^2\kappa^2 d\eta\big).$$

Gradient complexity of SG-HMC. Similar to (Chatterji et al., 2018; Baker et al., 2018; Brosse et al., 2018), we assume $L = O(n)$ for simplicity. This further implies that $K = O(n^{-1/2}\eta^{-1})$ and $\eta = O(L^{-1/2}) = O(n^{1/2})$. Then if ignoring the dependency on the condition number κ and dimension d but only pay attention to the dependency on ϵ,

B, and η, the sampling error—corresponding to the last two terms of the bound—is $O(n^{1/4}B^{-1/2}\eta^{1/2})$. Let the target sampling error be $\epsilon/\sqrt{n}$ for arbitrary $\epsilon \in (0,1)$, it suffices to set $\eta = \Theta(n^{-3/2}B\epsilon^2)$. Note that HMC requires to make $T = O(\log(1/\epsilon))$ proposals to ensure good mixing. As a result, the gradient complexity of SG-HMC is $KTB = \widetilde{O}(n^{-1/2}\eta^{-1}B) = \widetilde{O}(n\epsilon^{-2})$.

Stochastic variance reduced gradient HMC (SVRG-HMC). We deliver the convergence rate of SVGR-HMC in the following theorem.

Theorem 5.3. Under the same assumptions made in Theorem 5.2 and let μ_t be the distribution of $\mathbf{x}^{(t)}$. Then if $\eta = O(L^{-1} \wedge d\mu^{-1}\Gamma_1^{-1})$ and set $K = 1/(4\sqrt{L}\eta)$, the output of SVRG-HMC satisfies

$$\mathcal{W}_2(\mu_T, \pi) \leq 2\sqrt{\frac{d}{\mu}}\big(1-(128\kappa)^{-1}\big)^{T/2} + \Gamma_1\eta^{1/2} + \Gamma_2\eta,$$

where the constants Γ_1 and Γ_2 satisfy,

$$\Gamma_1^2 = O\big(L^{1/2}B^{-1}N^2\kappa^3 d\eta^2\big)$$
$$\Gamma_2^2 = O\big(\kappa^3 d + L^{3/2}B^{-1}N^2\kappa^3 d\eta^3\big).$$

Gradient complexity of SVRG-HMC. We first set $BN = \Theta(n)$, then Theorem 5.3 suggests that the sampling error of SVRG-HMC is $O(n^{5/4}B^{-3/2}\eta^{3/2} + \eta)$. Then it suffices to set the step size $\eta = \Theta\big(n^{-7/6}B\epsilon^{2/3} \wedge n^{-1/2}\epsilon\big)$ to guarantee $\epsilon/\sqrt{n}$-sampling error, which further implies that the gradient complexity of SVRG-HMC is $KTB = \widetilde{O}\big(n^{2/3}\epsilon^{-2/3} + B\epsilon^{-1}\big)$, where we use the fact that $T = O(\log(1/\epsilon))$ and $K = O(n^{-1/2}\eta^{-1})$. Then we can set the batch size $B = O(n^{2/3}\epsilon^{1/3} \vee 1)$ and get a $O(n^{2/3}\epsilon^{-2/3} + \epsilon^{-1})$ gradient complexity for SVRG-HMC.

Stochastic averaged gradient HMC (SAGA-HMC). We present the convergence rate of SAGA-HMC in the following theorem.

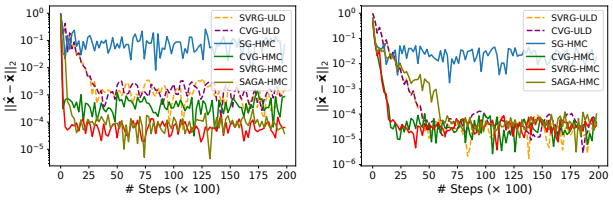
Theorem 5.4. Under the same assumptions made in Theorem 5.2. Let μ_t be the distribution of $\mathbf{x}^{(t)}$, then if $\eta = O(L^{-1} \wedge d\mu^{-1}\Gamma_1^{-1})$ and set $K = 1/(4\sqrt{L}\eta)$, the output of SAGA-HMC satisfies

$$\mathcal{W}_2(\mu_T, \pi) \leq 2\sqrt{\frac{d}{\mu}}\big(1-(128\kappa)^{-1}\big)^{T/2} + \Gamma_1\eta^{1/2} + \Gamma_2\eta,$$

where the constants Γ_1 and Γ_2 satisfy,

$$\Gamma_1^2 = O\big(L^{1/2}B^{-3}n^2\kappa^3 d\eta^2\big)$$
$$\Gamma_2^2 = O\big(\kappa^3 d + L^{3/2}B^{-3}n^2\kappa^3 d\eta^3\big).$$

Gradient complexity of SAGA-HMC. Theorem 5.4 suggests that the sampling error of SAGA-HMC is $O(n^{5/4}B^{-3/2}\eta^{3/2} + \eta)$, which is identical to that of SVRG-HMC. Then we can similarly set the step size $\eta = \Theta\big(n^{-7/6}B\epsilon^{2/3} \wedge \epsilon\big)$ to guarantee $\epsilon/\sqrt{n}$-sampling error, and further set $B = O(n^{2/3}\epsilon^{1/3} \vee 1)$ to get a $\widetilde{O}\big(n^{2/3}\epsilon^{-2/3} + \epsilon^{-1}\big)$ gradient complexity for SAGA-HMC.



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

Figure 1. Sampling error of SG-HMC, CVG-HMC, SVRG-HMC, SAGA-HMC, SVRG-ULD, and CVG-ULD on synthetic data. X-axis represents the error between the estimated mean $\widehat{\mathbf{x}}$ and the true one $\bar{\mathbf{x}}$, Y-axis represents the total number of steps.

Control variates gradient HMC (CVG-HMC). Note that control variates gradient adopts a fixed reference point $\widehat{\mathbf{q}}$ in the entire algorithm. In the following analysis, we will simply set $\widehat{\mathbf{q}} = \mathbf{x}^{(0)}$, which also satisfies that $\|\widehat{\mathbf{q}} - \mathbf{x}^*\|_2^2 \leq d/\mu$. The following theorem characterizes the convergence results of CVG-HMC in 2-Wasserstein distance.

Theorem 5.5. Under the same Assumptions made in Theorem 5.2. Let μ_t be the distribution of $\mathbf{x}^{(t)}$, then if $\eta = O(L^{-1} \wedge d\mu^{-1}\Gamma_1^{-1})$ and set $K = 1/(4\sqrt{L}\eta)$, the output of CVG-HMC satisfies

$$\mathcal{W}_2(\mu_T, \pi) \leq 2\sqrt{\frac{d}{\mu}}\big(1-(128\kappa)^{-1}\big)^{T/2} + \Gamma_1\eta^{1/2} + \Gamma_2\eta,$$

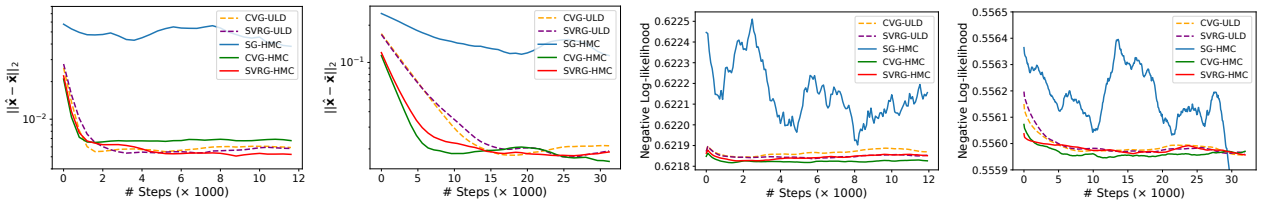
where the constants Γ_1 and Γ_2 satisfy,

$$\Gamma_1^2 = O\big(L^{-1/2}B^{-1}\kappa^3 d\big) \quad \text{and} \quad \Gamma_2^2 = O\big(\kappa^3 d\big).$$

Gradient complexity of CVG-HMC. Theorem 5.5 shows that the sampling error of CVG-HMC is $O(n^{-1/4}B^{-1/2}\eta^{1/2} + \eta)$, which implies that we can set the step size as $\eta = \Theta\big(n^{-1/2}B\epsilon^2 \wedge n^{-1/2}\epsilon\big)$ to achieve $\epsilon/\sqrt{n}$-sampling error in 2-Wasserstein distance. Then similarly, by setting $B = O(\epsilon^{-1})$ we can derive that the gradient complexity of CVG-HMC is $KTB = \widetilde{O}\big(\epsilon^{-2}\big)$.

6. Experiments

In this section, we will evaluate the empirical performance of the aforementioned four stochastic gradient HMC algorithms, including SG-HMC, SVRG-HMC, SAGA-HMC and CVG-HMC, on both synthetic and real-world datasets. Moreover, we will also include SVRG-ULD and CVG-ULD



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

Figure 2. Experimental results of Bayesian logistic regression on Covtype dataset. (a)-(b): Sampling error of SG-HMC, CVG-HMC, SVRG-HMC, SVG-ULD, and SVRG-ULD, function of the number of iterations. (c)-(d): Negative log-likelihood on the test dataset for SG-HMC, CVG-HMC, SVRG-HMC, SVG-ULD, and SVRG-ULD, function of the number of iterations.

for comparison since they have been demonstrated to perform well in both theory and experiment (Zou et al., 2018a; Chatterji et al., 2018).

6.1. Sampling from Multivariate Gaussian Distribution

We first evaluate the performances of these four stochastic gradient HMC algorithms for sampling from a multivariate Gaussian distribution. Specifically, given n mean vectors $\{\boldsymbol{\mu}_i\}_{i=1}^n$ and positive definite matrices $\{\boldsymbol{\Sigma}_i\}_{i=1}^n$, we set each component function as $f_i(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i(\mathbf{x} - \boldsymbol{\mu}_i)/2$. Then it can be seen that each component distribution, i.e., $\pi_i \propto \exp(-f_i(\mathbf{x}))$ is a Gaussian distribution with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i^{-1}$, and thus the target distribution $\pi \propto \exp(-f(\mathbf{x}))$ is also a Gaussian distribution. In our experiment, we generate two synthetic dataset with size $n = 500$ and $n = 5000$. For HMC based algorithms, We run all four algorithms using the same step size ($\eta = \{2\times10^{-3}, 3\times10^{-4}\}$ for $n = \{500, 5000\}$) and mini-batch size $B = 16$ with 2×10^4 steps (2000 proposals with 10 internal leapfrog steps each). For ULD based algorithms, we follow the same configuration in Chen et al. (2017); Chatterji et al. (2018) by setting the friction parameter as $\gamma = 1/n$ and the inverse mass parameter as $u = 2$. The mini-batch size and iteration number are identical to those of HMC based algorithms, and the step size are tuned such that the algorithm can converge fast.

In order to characterize the convergence performance in terms of the distance between distributions, we run all of these four algorithms for 10^5 times in parallel, which gives 10^5 independent samples at each iteration. Since it is not computation efficient to exactly compute the 2-Wasserstein distance, we instead evaluate the error between the estimated mean $\widehat{\mathbf{x}}$ and the true one $\bar{\mathbf{x}}$ (which can be exactly computed based on $\{\boldsymbol{\mu}_i\}_{i=1}^n$ and $\{\boldsymbol{\Sigma}_i\}_{i=1}^n$). We display the experimental results in Figure 1. Besides, we also characterize the estimation errors of the second moment $\bar{\mathbf{z}} = \mathbb{E}_{\mathbf{x}\sim\pi}[\mathbf{x} \odot \mathbf{x}]$, which are reported in Table 1. It can be observed that all these four stochastic gradient HMC converges, while the mini-batch stochastic gradient leads to significantly larger sampling error than other three stochastic gradient estimators. Additionally, when n increases, the performance of CVG become closer to those of SVRG-HMC and SAGA-HMC. These observations align well with our theoretical results on the HMC based algorithms stated in Table 1. Moreover, we also observe that SVRG-HMC and SAGA-HMC can outperform SVRG-ULD on small dataset, though in theory they have the same gradient complexity.

Table 2. Error of estimating the quantity $\bar{\mathbf{z}} = \mathbb{E}_{\mathbf{x}\sim\pi}[\mathbf{x} \odot \mathbf{x}]$

Algorithm (HMC)	SG	SVRG	SAGA	CVG
Error ($\|\widehat{\mathbf{z}} - \bar{\mathbf{z}}\|_2$)	0.070	0.0022	0.0018	0.0017

6.2. Bayesian Logistic Regression

We then perform Bayesian logistic regression to evaluate the empirical performances of all stochastic gradient HMC algorithms. In particular, let $\{\mathbf{z}_i, y_i\}_{i=1}^n$ be the observed training data, where $\mathbf{z}_i \in \mathbb{R}^d$ and y_i are the feature vector and label of the i-th observation respectively. The likelihood function given the observation $\{\mathbf{z}_i, y_i\}$ is modeled by $p(y_i|\mathbf{z}_i, \mathbf{x}) = 1/\big(1 + \exp(-y_i\mathbf{x}^\top\mathbf{z}_i)\big)$. Then assuming that the model parameter $\mathbf{x}$ follows from a Gaussian prior $p(\mathbf{x}) = N(\mathbf{0}, \lambda^{-1}\mathbf{I})$. We aim to sample the posterior $p\big(\mathbf{x}|\{\mathbf{z}_i, y_i\}_{i=1}^n\big) = p(\mathbf{x})\prod_{i=1}^n p(y_i|\mathbf{z}_i, \mathbf{x})$. Therefore, it can be derived that the negative log-posterior function is $f(\mathbf{x}) = \sum_{i=1}^n f_i(\mathbf{x})$ with the component function $f_i(\mathbf{x})$ defined by $f_i(\mathbf{x}) = \log\big(1 + \exp(-y_i\mathbf{x}^\top\mathbf{z}_i)\big) + \lambda\|\mathbf{x}\|_2^2/(2n)$.

We carry out the experiments on Covtype dateset [5], which has 581012 instances with 54 attributes. We further extract two training dataset with size $n = \{500, 5000\}$ from the original dataset, and take the rest for test. Similar to the experiments on the synthetic data, we use the same mini-batch size ($B = 16$) and step size ($\eta = \{2 \times 10^{-3}, 4 \times 10^{-4}\}$ for $n = \{500, 5000\}$) for SH-HMC, SVRG-HMC, and CVG-HMC [6]. For ULD based algorithms we use the same batch size and tune the step size such that they converge fast.

[5] Available at *https://archive.ics.uci.edu/ml/datasets/covertype*

[6] We point out that SAGA-HMC is extremely inefficient in generating independent samples in a parallel manner since it requires huge memory cost. So we do not include SAGA-HMC in this part.

Moreover, we run all algorithms for 2×10^4 times in parallel and obtain 2×10^4 independent samples at each iteration. Given the generated samples, we compute the mean (in order to increase the precision of the estimation, we further apply moving average with size 100, so in total we use 2×10^6 samples) and compared it to the ground truth, which is obtained by running standard HMC algorithms (using full gradient and MH correction), and display the errors in Figures 2(a) and 2(b). It can be clearly observed that SG-HMC performs significantly worse than other algorithms. Besides, we also observe SVRG-HMC slightly outperforms CVG-HMC, SVRG-ULD, and CVG-ULD on the small dataset, which is consistent with the observation on the synthetic dataset. Moreover, given the estimated mean at different iterations, we evaluate the negative log-likelihood of all algorithms on the test dataset. The results are displayed in Figures 2(c) and 2(d). The plots show that the output of SG-HMC has a significantly larger bias than those of other algorithms, while SVRG-HMC, CVG-HMC, CVG-ULD, and SVRG-ULD give similar results. This again explains the inefficiency of SG-HMC and verifies our theory.

7. Conclusion and Future Work

In this paper, we provided a general framework for proving the convergence rate of HMC with stochastic gradients. Our result shows that as long as the variance of stochastic gradient is upper bounded along the Markov chain, stochastic gradient HMC algorithms with properly chosen step size provably converge to the target distribution. We applied the general convergence result to four specific stochastic gradient HMC algorithms: SG-HMC, CVG-HMC, SVRG-HMC and SAGA-HMC, and established their convergence guarantees. The results explain the inefficiency of SG-HMC, and reveal the potential prospects of the applications of CVG-HMC, SVRG-HMC, and SAGA-HMC.

One interesting future direction is to explore whether adding Metropolis-Hasting (MH) correction in certain ways to the stochastic HMC algorithm can help mitigating the bias caused by stochastic gradients in theory, which is supported by some empirical evidence (Dang et al., 2019).

Acknowledgement

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

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

Baker, J., Fearnhead, P., Fox, E. B., and Nemeth, C. Control variates for stochastic gradient MCMC. *Statistics and Computing*, 2018. ISSN 1573-1375. doi: 10.1007/s11222-018-9826-2.

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

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

Bou-Rabee, N., Eberle, A., and Zimmer, R. Coupling and convergence for Hamiltonian monte carlo. *arXiv preprint arXiv:1805.00452*, 2018.

Brosse, N., Durmus, A., Moulines, É., and Pereyra, M. Sampling from a log-concave distribution with compact support with proximal Langevin Monte Carlo. In *Conference on Learning Theory*, pp. 319–342, 2017.

Brosse, N., Durmus, A., and Moulines, E. The promises and pitfalls of stochastic gradient langevin dynamics. In *Advances in Neural Information Processing Systems*, pp. 8268–8278, 2018.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

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

Chau, H. N. and Rasonyi, M. Stochastic gradient hamiltonian monte carlo for non-convex learning. *arXiv preprint arXiv:1903.10328*, 2019.

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

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

Chen, T., Fox, E., and Guestrin, C. Stochastic gradient Hamiltonian monte carlo. In *International Conference on Machine Learning*, pp. 1683–1691, 2014.

Chen, Y., Dwivedi, R., Wainwright, M. J., and Yu, B. Fast mixing of metropolized hamiltonian monte carlo: Benefits of multi-step gradients. *arXiv preprint arXiv:1905.12247*, 2019.

Chen, Z. and Vempala, S. S. Optimal convergence rate of hamiltonian monte carlo for strongly logconcave distributions. *arXiv preprint arXiv:1905.02313*, 2019.

Cheng, X., Chatterji, N. S., Abbasi-Yadkori, Y., Bartlett, P. L., and Jordan, M. I. Sharp convergence rates for Langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*, 2018.

Dalalyan, A. S. 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.

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

Dalalyan, A. S., Riou-Durand, L., et al. On sampling from a log-concave density using kinetic langevin diffusions. *Bernoulli*, 26(3):1956–1988, 2020.

Dang, K.-D., Quiroz, M., Kohn, R., Tran, M.-N., and Villani, M. Hamiltonian monte carlo with energy conserving subsampling. *Journal of machine learning research*, 20 (100):1–31, 2019.

Defazio, A., Bach, F., and Lacoste-Julien, S. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pp. 1646–1654, 2014.

Deng, W., Feng, Q., Gao, L., Liang, F., and Lin, G. Non-convex learning via replica exchange stochastic gradient mcmc. In *International Conference on Machine Learning*, pp. 2474–2483. PMLR, 2020.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

Dubey, K. A., Reddi, S. J., Williamson, S. A., Poczos, B., Smola, A. J., and Xing, E. P. Variance reduction in stochastic gradient Langevin dynamics. In *Advances in Neural Information Processing Systems*, pp. 1154–1162, 2016.

Durmus, A., Moulines, E., and Saksman, E. On the convergence of hamiltonian monte carlo. *arXiv preprint arXiv:1705.00166*, 2017.

Durmus, A., Moulines, E., et al. High-dimensional bayesian inference via the unadjusted langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

Erdogdu, M. A., Mackey, L., and Shamir, O. Global non-convex optimization with discretized diffusions. In *Advances in Neural Information Processing Systems*, pp. 9671–9680, 2018.

Gao, X., Gurbuzbalaban, M., and Zhu, L. Breaking reversibility accelerates langevin dynamics for global non-convex optimization. *arXiv preprint arXiv:1812.07725*, 2018a.

Gao, X., Gürbüzbalaban, M., and Zhu, L. 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*, 2018b.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pp. 315–323, 2013.

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

Lee, Y. T., Song, Z., and Vempala, S. S. Algorithmic theory of odes and sampling from well-conditioned logconcave densities. *arXiv preprint arXiv:1812.06243*, 2018.

Li, C., Chen, C., Carlson, D., and Carin, L. Preconditioned stochastic gradient langevin dynamics for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

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

Lovász, L. and Simonovits, M. The mixing rate of markov chains, an isoperimetric inequality, and computing the volume. In *Proceedings [1990] 31st annual symposium on foundations of computer science*, pp. 346–354. IEEE, 1990.

Ma, Y.-A., Chen, T., and Fox, E. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems*, pp. 2917–2925, 2015.

Ma, Y.-A., Chen, Y., Jin, C., Flammarion, N., and Jordan, M. I. Sampling can be faster than optimization. *arXiv preprint arXiv:1811.08413*, 2018.

Mangoubi, O. and Smith, A. Rapid mixing of hamiltonian monte carlo on strongly log-concave distributions. *arXiv preprint arXiv:1708.07114*, 2017.

Mangoubi, O. and Vishnoi, N. Dimensionally tight bounds for second-order hamiltonian monte carlo. In *Advances in neural information processing systems*, pp. 6027–6037, 2018.

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

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

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

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

Roberts, G. O. and Tweedie, R. L. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pp. 341–363, 1996.

Shen, R. and Lee, Y. T. The randomized midpoint method for log-concave sampling. *arXiv preprint arXiv:1909.05503*, 2019.

Simsekli, U., Zhu, L., Teh, Y. W., and Gurbuzbalaban, M. Fractional underdamped langevin dynamics: Retargeting sgd with momentum under heavy-tailed gradient noise. In *International Conference on Machine Learning*, pp. 8970–8980. PMLR, 2020.

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

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 681–688, 2011.

Xu, P., Chen, J., Zou, D., and Gu, Q. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pp. 3126–3137, 2018.

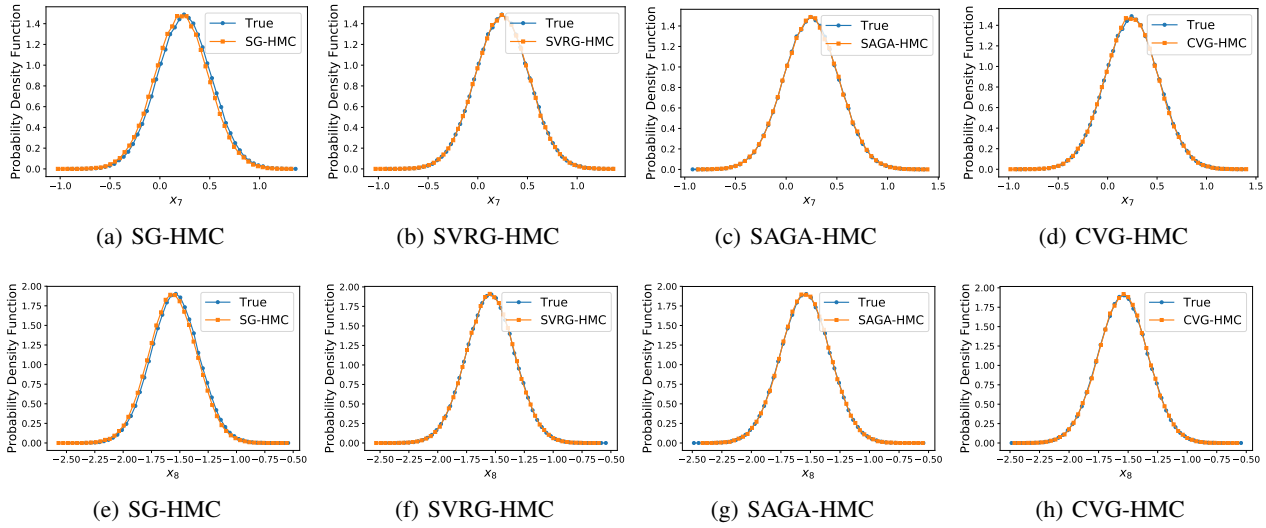
Zhang, Y., Liang, P., and Charikar, M. A hitting time analysis of stochastic gradient Langevin dynamics. In *Conference on Learning Theory*, pp. 1980–2022, 2017.

Zou, D., Xu, P., and Gu, Q. Stochastic variance-reduced Hamilton Monte Carlo methods. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 6028–6037, 2018a.

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

Zou, D., Xu, P., and Gu, Q. 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*, pp. 2936–2945. PMLR, 2019.

Zou, D., Xu, P., and Gu, Q. Faster convergence of stochastic gradient langevin dynamics for non-log-concave sampling. *arXiv preprint arXiv:2010.09597*, 2020.



(a) SG-HMC (b) SVRG-HMC (c) SAGA-HMC (d) CVG-HMC

(e) SG-HMC (f) SVRG-HMC (g) SAGA-HMC (h) CVG-HMC

Figure 3. Visualization of the marginal distributions of samples generated by SG-HMC, CVG-HMC, SVRG-HMC, and SAGA-HMC. (a)-(d): Marginal distribution of x_7 (e)-(f): Marginal distribution of x_8.

A. Additional Experiments

We further conduct the Bayesian logistic regression experiments on the Pima dataset, which consists of 768 instances with 9 attributes (including an additional bias coordinate). In order the demonstrate the distributional convergence of all stochastic gradient HMC algorithms, we plot the marginal distributions of samples obtained by the algorithms and compare them with the true one (obtained by running full-gradient HMC with MH correction) in Figure 3. Here we pick the 7-th and 8-th coordinates (x_7 and x_8) of the model parameter for plotting the distributions. It can be seen that all the four algorithms can well recover the target distribution, although there is some slight mismatch for SG-HMC. This is consistent with our general result in Theorem 4.4.

B. Proof of the Main Results

In order to simplify the proof, we define the following operators that will be frequently leveraged in the remaining part of this paper. Specifically, given the iterate $(\mathbf{q}_k, \mathbf{p}_k)$ and step size η, the operator $\mathcal{S}_\eta$ is defined as

$$\mathcal{S}_\eta \mathbf{q}_k = \mathbf{q}_{k+1} = \mathbf{q}_k + \eta \mathbf{p}_k - \frac{\eta^2}{2}\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k)$$
$$\mathcal{S}_\eta \mathbf{p}_k = \mathbf{p}_{k+1} = \mathbf{p}_k - \frac{\eta}{2}\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) - \frac{\eta}{2}\mathbf{g}\big(\mathbf{q}_k + \eta \mathbf{p}_k - \eta^2 \mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k)/2, \boldsymbol{\xi}_{k+1/2}\big),$$

where the randomness $\boldsymbol{\xi}_k$ and $\boldsymbol{\xi}_{k+1/2}$ are independently drawn. Accordingly, the operator $\mathcal{G}_\eta$ is defined in a way that $\mathcal{G}_\eta \mathbf{q}_K = \mathbb{E}_{\boldsymbol{\xi}_k}[\mathcal{S}_\eta \mathbf{q}_k]$ and $\mathcal{G}_\eta \mathbf{p}_k = \mathbb{E}_{\boldsymbol{\xi}_k, \boldsymbol{\xi}_{k+1/2}}[\mathcal{S}_\eta \mathbf{q}_k]$, which computes the mean of the trajectory induced by $\mathcal{S}_\eta$. It is worth noting that $\mathcal{G}_\eta \mathbf{q}_k$ is identical to one-step HMC using full gradient, while $\mathcal{G}_\eta \mathbf{p}_k$ is not since the $\mathcal{S}_\eta \mathbf{p}_k$ contains double randomness. The operator $\mathcal{H}_\eta$ is defined as the solution of ODE (3.1) after time η, i.e., let $(\mathbf{q}(0), \mathbf{p}(0)) = (\mathbf{p}_k, \mathbf{q}_k)$, we have

$$\mathcal{H}_\eta \mathbf{q}_k = \mathbf{q}_k + \int_0^\eta \mathbf{p}(t) \mathrm{d}t$$
$$\mathcal{H}_\eta \mathbf{p}_k = \mathbf{p}_k - \int_0^\eta \nabla f(\mathbf{q}(t)) \mathrm{d}t$$

Then the proof roadmap of Theorem 4.4 are divided into three steps. In particular, in the first two steps, we will assume that the second moments of $\mathbf{q}_k$ and $\mathbf{p}_k$ are upper bounded along the interpolation of the training trajectory. More specifically, the first two steps are 1) we will prove the error bound of the approximated solution of the Hamilton's equation (3.1) found using stochastic gradient; 2) We will combine the error bound obtained in Step 1) and the contraction results of the Hamilton's

equation (3.1) to prove the convergence of stochastic gradient descent. Then in the third step, we will explicitly prove the upper bounds on the second moments of $\mathbf{q}_k$ and $\mathbf{p}_k$ to validate the results we obtained in the first two steps.

For the step 1), the following Lemma provides the approximation error of the solution to (3.1) solved with stochastic gradients, which is the key to guarantee the convergence of stochastic gradient HMC algorithms.

Lemma B.1. For any $t \leq T$, let $\{(\mathbf{q}_k, \mathbf{p}_k)\}_{k=1,\ldots,K}$ be all iterates in the leapfrog update starting from $\mathbf{x}^{(t)}$, define by

$$E_p = \max_{k,\tau} \mathbb{E}[\| \max_{s \leq \eta} \mathcal{H}_s \mathbf{p}_k \|_2^2] \quad \text{and} \quad E_q = \max_{k,\tau} \mathbb{E}[\max_{s \leq \eta} \|\nabla f(\mathcal{H}_s \mathbf{p}_k)\|_2^2].$$

Then under Assumptions 4.2 and 4.3, if $K\eta \leq 1/(4L^{1/2})$, the discretization error $\mathbb{E}[\|\mathcal{S}_\eta^k \mathbf{z}_0 - \mathcal{H}_\eta^k \mathbf{z}_0\|_2^2]$ can be upper bounded as follows,

$$\mathbb{E}[\|\mathcal{S}_\eta^k \mathbf{z}_0 - \mathcal{H}_\eta^k \mathbf{z}_0\|_2^2] \leq \gamma_1 L^{-1/2} \eta + \gamma_2 \big(L^{-1/2}\eta + L^{-1}\big)\eta^2,$$

where γ_1 and γ_2 are defined as follows,

$$\gamma_1 = \frac{\sigma^2}{2L}, \gamma_2 = \frac{\eta^2 L^2 E_p}{36} + \frac{\eta^2 L \sigma^2}{8} + 8L(\eta^2 E_q + 2E_p).$$

In step 2), we first present the following lemma that gives an tight contraction bound on the Hamilton's equation (3.1). equation (3.1).

Lemma B.2 (Lemma 6 in Chen & Vempala (2019))**.** Under Assumptions 4.2 and 4.1, let $\mathbf{q}(0)$ and $\mathbf{q}'(0)$ be two different initial positions and $\mathbf{p}(0) = \mathbf{p}'(0)$ be the initial velocities. Then for any $t \leq 1/(2\sqrt{L})$, it holds that

$$\|\mathcal{H}_t \mathbf{q}(0) - \mathcal{H}_t \mathbf{q}'(0)\|_2^2 \leq \big(1 - \mu t^2/4\big)\|\mathbf{q}(0) - \mathbf{q}'(0)\|_2^2.$$

Based on the above two lemmas, we further provide the following lemma that gives the convergence rate of stochastic gradient HMC under the assumption that E_p and E_q defined in Lemma B.1 are upper bounded.

Lemma B.3. Under Assumptions 4.1, 4.2, and 4.3, if set the step size $\eta \leq 1/(4K\sqrt{L})$, it holds that

$$\mathbb{E}[\|\mathbf{x}^{(t)} - \mathbf{x}^\pi\|_2^2] \leq \big(1 - (128\kappa)^{-1}\big)^t \mathbb{E}[\|\mathbf{x}^{(0)} - \mathbf{x}^\pi\|_2^2] + 16512\kappa^2 \big[\gamma_1 L^{-1/2}\eta + 2\gamma_2 L^{-1}\eta^2\big],$$

where $\mathbf{x}^\pi$ denotes the random vector following the target distribution π, the parameters γ_1 and γ_2 are defined as follows,

$$\gamma_1 = \frac{\sigma^2}{2L}, \gamma_2 = \frac{\eta^2 L^2 E_p}{36} + \frac{\eta^2 L \sigma^2}{8} + 8L(\eta^2 E_q + 2E_p),$$

where E_q and E_p are defined in Lemma B.1.

Then we are ready to complete step 3) by showing that the quantities E_p and E_q are upper bounded. We clearly state this in the following lemma.

Lemma B.4. Under Assumptions 4.1, 4.2, and 4.3, and assume that $L \geq 4$, then if the step size satisfies $\eta = O(L^{1/2}\sigma^{-2}\kappa^{-1} \wedge L^{-1/2})$, the constants E_p and E_q defined in Lemma B.1 can be upper bounded by

$$\begin{aligned} E_p &\leq 12(2DL + 5\kappa d + d/2 + K\eta^2\sigma^2) \\ E_q &\leq 30L(2DL + 5\kappa d + d/2 + K\eta^2\sigma^2), \end{aligned}$$

where $D = \|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2$ and $\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x})$.

Proof of Theorem 4.4. Now we can simply combine Lemmas B.3 and B.4 to complete the proof of Theorem 4.4. Additionally, by Young's inequality we have

$$\begin{aligned} \mathbb{E}[\|\mathbf{x}^{(0)} - \mathbf{x}^\pi\|_2^2] &\leq 2\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2 + 2\mathbb{E}[\|\mathbf{x}^* - \mathbf{x}^\pi\|_2^2] \\ &\leq 2(D + d/\mu), \end{aligned}$$

where $D = \|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2$ and the second inequality follows from Proposition 1 in (Durmus et al., 2019) that shows that $\mathbb{E}[\|\mathbf{x}^* - \mathbf{x}^\pi\|_2^2] \leq d/\mu$. Then we are able to complete the proof. □

C. Proof of Theorems in Section 5

In order to prove the theorems in Section 5, it suffices to verify that the variance along the Markov chain is upper bounded. Then in the subsequent proof for each theorem, we will first present a lemma that characterizes the variance of the corresponding stochastic gradient estimator, and then complete the proof using the general result in Theorem 4.4.

C.1. Proof of Theorem 5.2

The following lemma provides an upper bound on the variance of the mini-batch stochastic gradient along the entire Markov Chain generated by SG-HMC.

Lemma C.1. Under Assumption 5.1, if the initial point $\mathbf{x}^{(0)}$ satisfies $\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2 \leq d/\mu$, it holds that

$$\mathbb{E}[\|\mathbf{g}_{SG}(\mathbf{q}_k, \boldsymbol{\xi}_k) - \nabla f(\mathbf{q}_k)\|_2^2] \leq \frac{1072L\kappa d + 4n^2\mathbb{E}[\|\nabla f_i(\mathbf{q}^*)\|_2^2]}{B},$$

where $D = \|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2$.

Proof of Theorem 5.2. Based on Lemma C.1, we are able to complete the proof of Theorem 1 using the general results in Theorem 4.4. Specifically, Theorem 4.4 states that for a stochastic gradient estimator with variance upper bounded by σ^2 along the iteration trajectory, the output of Algorithm 1 satisfies

$$\mathcal{W}_2\big(P(\mathbf{x}^{(t)}), \pi\big) \leq \big(1 - (128\kappa)^{-1}\big)^{T/2} \cdot \big(2d/\mu + 2D\big)^{1/2} + \Gamma_1\eta^{1/2} + \Gamma_2\eta,$$

where

$$\begin{aligned}\Gamma_1^2 &= O\bigg(\frac{\kappa^2\sigma^2}{L^{3/2}}\bigg)\\ \Gamma_2^2 &= O\big(\kappa^2(\kappa d + L^{-1/2}\sigma^2\eta)\big).\end{aligned} \tag{C.1}$$

Plugging the bound on σ^2 proved in Lemma C.1 and using the assumption in Theorem 5.2 that $D \leq L/\mu$, we have

$$\begin{aligned}\big(2d/\mu + 2D\big)^{1/2} &\leq 2\sqrt{d/\mu}\\ \Gamma_1^2 &= O\bigg(\frac{L^{-1/2}\kappa^3 d + L^{-3/2}\kappa^2 n^2 d}{B}\bigg)\\ \Gamma_2^2 &= O\bigg(\kappa^3 d + \frac{L^{-1/2}n^2\kappa^2 d\eta}{B}\bigg).\end{aligned}$$

This completes the proof.

□

C.2. Proof of Theorem 5.3

Lemma C.2 (Upper bound of the variance of SVRG estimator)**.** Under assumptions 5.1, it holds that

$$\mathbb{E}\big[\|\mathbf{g}_{\mathrm{SVRG}}(\mathbf{q}_k, \boldsymbol{\xi}_k) - \nabla f(\mathbf{q}_k)\|_2^2\big] \leq \frac{672N^2L^2\eta^2\kappa d}{B},$$

where E_p and E_q are defined in Lemma D.2.

Proof of Theorem 5.3. Similar to the proof of Theorem 5.2, we only need to combine the upper bound of the variance of stochastic gradients proved in Lemma C.2 and the general convergence results in Theorem 4.4. Specifically, by Theorem 4.4 we know that

$$\mathcal{W}_2\big(P(\mathbf{x}^{(t)}), \pi\big) \leq \big(1 - (128\kappa)^{-1}\big)^{T/2} \cdot \big(2d/\mu + 2D\big)^{1/2} + \Gamma_1\eta^{1/2} + \Gamma_2\eta,$$

with Γ_1 and Γ_2 defined in Theorem 4.4. Then applying Lemma C.2, we obtain

$$\Gamma_1^2 = O\bigg(\frac{L^{1/2}N^2\kappa^3 d\eta^2}{B}\bigg)$$
$$\Gamma_2^2 = O\bigg(\kappa^3 d + \frac{L^{3/2}N^2\kappa^3 d\eta^3}{B}\bigg).$$

This completes the proof.

□

C.3. Proof of Theorem 5.4

Lemma C.3 (Upper bound of the variance of SAG estimator)**.** Under assumptions 5.1, it holds that

$$\mathbb{E}\big[\|\mathbf{g}_{\mathrm{SAG}}(\mathbf{q}_k, \boldsymbol{\xi}_k) - \nabla f(\mathbf{q}_k)\|_2^2\big] \leq \frac{672\eta^2 n^2 L^2 \kappa d}{B^3},$$

where $D = \|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2$.

Proof of Theorem 5.3. Similar to the proof of Theorem 5.2, we only need to combine the upper bound of the variance of stochastic gradients proved in Lemma C.3 and the general convergence results in Theorem 4.4. Specifically, by Theorem 4.4, we know that

$$\mathcal{W}_2\big(P(\mathbf{x}^{(t)}), \pi\big) \leq \big(1 - (128\kappa)^{-1}\big)^{T/2} \cdot \big(2d/\mu + 2D\big)^{1/2} + \Gamma_1\eta^{1/2} + \Gamma_2\eta,$$

with Γ_1 and Γ_2 defined in Theorem 4.4. Then applying C.3 and using the fact that $K\eta \leq 1/(2\sqrt{L})$, we obtain

$$\Gamma_1^2 = O\bigg(\frac{L^{1/2}N^2\kappa^3 d\eta^2}{B}\bigg)$$
$$\Gamma_2^2 = O\bigg(\kappa^3 d + \frac{L^{3/2}n^2\kappa^3 d\eta^3}{B^3}\bigg).$$

This completes the proof. □

C.4. Proof of Theorem 5.5

Lemma C.4 (Upper bound of the variance of CV estimator)**.** Under assumptions 5.1, it holds that

$$\mathbb{E}\big[\|\mathbf{g}_{\mathrm{CV}}(\mathbf{q}_k, \boldsymbol{\xi}_k) - \nabla f(\mathbf{q}_k)\|_2^2\big] \leq \frac{688L\kappa d}{B},$$

Proof of Theorem 5.5. Similar to the proof of Theorem 5.2, we only need to combine the upper bound of the variance of stochastic gradients proved in Lemma C.4 and the general convergence results in Theorem 4.4. Specifically, by Theorem 4.4 we know that

$$\mathcal{W}_2\big(P(\mathbf{x}^{(t)}), \pi\big) \leq \big(1 - (128\kappa)^{-1}\big)^{T/2} \cdot \big(2d/\mu + 2D\big)^{1/2} + \Gamma_1\eta^{1/2} + \Gamma_2\eta,$$

with Γ_1 and Γ_2 defined in Theorem 4.4. Then applying Lemma and C.2 and using the fact that $K\eta \leq 1/(2\sqrt{L})$, we obtain

$$\Gamma_1^2 = O\bigg(\frac{L^{-1/2}\kappa^3 d}{B}\bigg)$$
$$\Gamma_2^2 = O\big(\kappa^3 d\big).$$

This completes the proof.

□

D. Proof of Lemmas in Appendix A

D.1. Proof of Lemma B.1

We first present the following lemma that states the difference between the operator $\mathcal{S}_\eta$ and $\mathbf{g}_\eta$.

Lemma D.1. Under Assumptions 4.2 and 4.3, for any $(\mathbf{q}_k, \mathbf{p}_k)$, it holds that

$$\mathbb{E}[\|\mathcal{S}_\eta \mathbf{q}_k - \mathcal{G}_\eta \mathbf{q}_k\|_2^2] \leq \frac{\eta^4 \sigma^2}{4}$$
$$\mathbb{E}[\|\mathcal{S}_\eta \mathbf{p}_k - \mathcal{G}_\eta \mathbf{p}_k\|_2^2] \leq \frac{\eta^2(4\sigma^2 + \eta^4 L^2 \sigma^2)}{4},$$

where the first expectation is taken over the randomness of $\mathbf{q}_k$ and $\boldsymbol{\xi}_k$ and the second expectation is taken over $\mathbf{p}_k$, $\boldsymbol{\xi}_k$ and $\boldsymbol{\xi}_{k+1/2}$

The following Lemma characterizes the upper bound of the one-step error between the operators $\mathcal{G}_\eta$ and $\mathcal{H}_\eta$.

Lemma D.2. Let E_p and E_q be defined in Lemma B.1. Then under Assumptions 4.2 and 4.3, it holds for all $(\mathbf{q}_k, \mathbf{p}_k)$ that

$$\mathbb{E}[\|\mathcal{G}_\eta \mathbf{q}_k - \mathcal{H}_\eta \mathbf{q}_k\|_2^2] \leq \frac{L^2 \eta^6}{36} E_p.$$
$$\mathbb{E}[\|\mathcal{G}_\eta \mathbf{p}_k - \mathcal{H}_\eta \mathbf{p}_k\|_2^2] \leq \frac{\eta^6 L^2 \sigma^2}{8} + 8\eta^4 L^2 (\eta^2 E_q + 2E_p),$$

where constants E_p and E_q are defined in Lemma B.1.

We further present the following lemma that gives an contraction bound of (3.1) for both position and momentum variables.

Lemma D.3. Under Assumptions 4.2, let $\mathbf{q}(0)$ and $\mathbf{q}'(0)$ be two different initial positions, $\mathbf{p}(0)$ and $\mathbf{p}'(0)$ be two different initial velocities, then for any $t \geq 0$ it holds that

$$\|\mathcal{H}_t \mathbf{q}(0) - \mathcal{H}_t \mathbf{q}'(0)\|_2^2 + L^{-1} \|\mathcal{H}_t \mathbf{p}(0) - \mathcal{H}_t \mathbf{p}'(0)\|_2^2 \leq e^{2\sqrt{L}t} \big[\|\mathbf{q}(0) - \mathbf{q}'(0)\|_2^2 + L^{-1} \|\mathbf{p}(0) - \mathbf{p}'(0)\|_2^2\big]$$

Now we are ready to complete the proof of Lemma B.1.

Proof of Lemma B.1. Let $\mathbf{z}_k = (\mathbf{q}_k, L^{-1/2}\mathbf{p}_k) = \mathcal{S}_\eta^k \mathbf{z}_0$, we will focus on proving the upper bound of $\mathbb{E}[\|\mathcal{S}_\eta^k \mathbf{z}_0 - \mathcal{H}_\eta^k \mathbf{z}_0\|_2^2]$. In particular, we have

$$\begin{aligned}
\mathcal{E}_k := \mathbb{E}[\|\mathcal{S}_\eta^k \mathbf{z}_0 - \mathcal{H}_\eta^k \mathbf{z}_0\|_2^2] &= \mathbb{E}\big[\|\mathcal{S}_\eta^k \mathbf{z}_0 - \mathcal{G}_\eta \mathcal{S}_\eta^{k-1} \mathbf{z}_0 + \mathcal{G}_\eta \mathcal{S}_\eta^{k-1} \mathbf{z}_0 - \mathcal{H}_\eta^k \mathbf{z}_0\|_2^2\big] \\
&= \underbrace{\mathbb{E}\big[\|\mathcal{S}_\eta^k \mathbf{z}_0 - \mathcal{G}_\eta \mathcal{S}_\eta^{k-1} \mathbf{z}_0\|_2^2\big]}_{I_1} + \underbrace{\mathbb{E}\big[\|\mathcal{G}_\eta \mathcal{S}_\eta^{k-1} \mathbf{z}_0 - \mathcal{H}_\eta^k \mathbf{z}_0\|_2^2\big]}_{I_2},
\end{aligned}$$

where the second equality is due to the fact that $\mathbb{E}[\mathcal{S}_\eta^k \mathbf{z}_0 | \mathcal{S}_\eta^{k-1} \mathbf{z}_0] = \mathcal{G}_\eta \mathcal{S}_\eta^{k-1} \mathbf{z}_0$. Then we will upper bound the two terms I_1 and I_2 separately. Regarding I_1, note that $\mathcal{S}_\eta^{k-1} \mathbf{z}_0 = (\mathbf{q}_{k-1}, L^{-1/2}\mathbf{p}_{k-1})$, applying Lemma D.1 gives

$$I_1 = \mathbb{E}[\|\mathcal{S}_\eta \mathbf{q}_{k-1} - \mathcal{G}_\eta \mathbf{q}_{k-1}\|_2^2] + L^{-1} \mathbb{E}[\|\mathcal{S}_\eta \mathbf{p}_{k-1} - \mathcal{G}_\eta \mathbf{p}_{k-1}\|_2^2] \leq \frac{\eta^4 \sigma^2}{4} + \frac{\eta^2(4\sigma^2 + \eta^4 L^2 \sigma^2)}{4L} := \gamma_1 \eta^2,$$

where

$$\gamma_1 := 2L^{-1}\sigma^2 \geq \frac{\eta^2 \sigma^2}{4} + \frac{4\sigma^2 + \eta^4 L^2 \sigma^2}{4L},$$

where the inequality follows from the assumption that $\eta \leq L^{-1/2}$. Regarding I_2, we can further expand it as follows,

$$\begin{aligned}
I_2 &= \mathbb{E}\big[\|\mathcal{G}_\eta \mathcal{S}_\eta^{k-1} \mathbf{z}_0 - \mathcal{H}_\eta \mathcal{S}_\eta^{k-1} \mathbf{z}_0 + \mathcal{H}_\eta \mathcal{S}_\eta^{k-1} \mathbf{z}_0 - \mathcal{H}_\eta^k \mathbf{z}_0\|_2^2\big] \\
&\leq (1 + 1/\alpha) \mathbb{E}\big[\|\mathcal{G}_\eta \mathcal{S}_\eta^{k-1} \mathbf{z}_0 - \mathcal{H}_\eta \mathcal{S}_\eta^{k-1} \mathbf{z}_0\|_2^2\big] + (1 + \alpha) \underbrace{\mathbb{E}\big[\|\mathcal{H}_\eta \mathcal{S}_\eta^{k-1} \mathbf{z}_0 - \mathcal{H}_\eta^k \mathbf{z}_0\|_2^2\big]}_{I_3},
\end{aligned}$$

where the last inequality follows from Young's inequality and $\alpha > 0$ is a constant that will be specified later. The first term on the R.H.S. of the above inequality can be further bounded using Lemma D.2. Note that $\mathcal{S}_\eta^{k-1}\mathbf{z}_0 = (\mathbf{q}_{k-1}, \mathbf{p}_{k-1}/L)$, we have

$$\begin{aligned}(1+1/\alpha)\mathbb{E}\big[\|\mathcal{G}_\eta\mathcal{S}_\eta^{k-1}\mathbf{z}_0 - \mathcal{H}_\eta\mathcal{S}_\eta^{k-1}\mathbf{z}_0\|_2^2\big] &\leq (1+1/\alpha)\mathbb{E}\big[\|\mathcal{G}_\eta\mathbf{q}_{k-1} - \mathcal{H}_\eta\mathbf{q}_{k-1}\|_2^2 + 1/L\|\mathcal{G}_\eta\mathbf{p}_{k-1} - \mathcal{H}_\eta\mathbf{p}_{k-1}\|_2^2\big] \\ &\leq (1+1/\alpha)\gamma_2\eta^4,\end{aligned}$$

where

$$\gamma_2 = \frac{\eta^2L^2E_p}{36} + \frac{\eta^2L\sigma^2}{8} + 8L(\eta^2E_q + 2E_p).$$

Moreover, in terms of I_3, we have the following based on Lemma D.3,

$$I_3 = \mathbb{E}\big[\|\mathcal{H}_\eta\mathcal{S}_\eta^{k-1}\mathbf{z}_0 - \mathcal{H}_\eta\mathcal{H}_\eta^{k-1}\mathbf{z}_0\|_2^2\big] \leq e^{2L^{1/2}\eta}\mathbb{E}\big[\|\mathcal{S}_\eta^{k-1}\mathbf{z}_0 - \mathcal{H}_\eta^{k-1}\mathbf{z}_0\|_2^2\big] \leq e^{2L^{1/2}\eta}\mathcal{E}_{k-1}.$$

Combining the above results, the discretization error $\mathcal{E}_k$ can be recursively upper bounded as follows,

$$\mathcal{E}_k \leq I_1 + I_2 \leq \gamma_1\eta^2 + (1+1/\alpha)\gamma_2\eta^4 + e^{2L^{1/2}\eta+\alpha}\mathcal{E}_{k-1}.$$

Therefore, based on the above recursive upper bound, it can be derived that

$$\begin{aligned}\mathcal{E}_k &\leq e^{2L^{1/2}\eta k+\alpha k}\mathcal{E}_0 + \sum_{t=0}^{k-1} e^{(2L^{1/2}\eta+\alpha)t}\big[\gamma_1\eta^2 + (1+1/\alpha)\gamma_2\eta^4\big] \\ &\leq \frac{e^{(2L^{1/2}\eta+\alpha)k}}{2L^{1/2}\eta+\alpha}\big[\gamma_1\eta^2 + (1+1/\alpha)\gamma_2\eta^4\big],\end{aligned}$$

where the second inequality follows from the fact that $\mathcal{E}_0 = 0$. Then we can set $\alpha = 2L^{1/2}\eta$. Assume that $k\eta \leq 1/(4L^{1/2})$, we have

$$\mathcal{E}_k \leq \frac{e^{4L^{1/2}\eta k}}{4L^{1/2}\eta}\big[\gamma_1\eta^2 + (1+L^{-1/2}\eta^{-1})\gamma_2\eta^4\big] \leq \gamma_1L^{-1/2}\eta + 2\gamma_2L^{-1}\eta^2.$$

This completes the proof.

□

D.2. Proof of Lemma B.3

Proof of Lemma B.3. By Lemmas B.1 and B.2, set $K\eta = 1/(4\sqrt{L})$, for any two different initial positions $\mathbf{q}_0$ and $\mathbf{q}_0'$ with the same initial velocities, we have

$$\begin{aligned}\mathbb{E}[\|\mathcal{S}_\eta^K\mathbf{q}_0 - \mathcal{H}_\eta^K\mathbf{q}_0'\|_2^2] &= \mathbb{E}[\|\mathcal{S}_\eta^K\mathbf{q}_0 - \mathcal{H}_\eta^K\mathbf{q}_0 + \mathcal{H}_\eta^K\mathbf{q}_0 - \mathcal{H}_\eta^K\mathbf{q}_0'\|_2^2] \\ &\leq (1+\beta)\|\mathcal{H}_\eta^K\mathbf{q}_0 - \mathcal{H}_\eta^K\mathbf{q}_0'\|_2^2 + (1+1/\beta)\mathbb{E}[\|\mathcal{S}_\eta^K\mathbf{q}_0 - \mathcal{H}_\eta^K\mathbf{q}_0'\|_2^2] \\ &\leq (1+\beta)\big(1-1/(64\kappa)\big)\mathbb{E}[\|\mathbf{q}_0 - \mathbf{q}_0'\|_2^2] \\ &\quad + (1+1/\beta)\big[\gamma_1L^{-1/2}\eta + 2\gamma_2L^{-1}\eta^2\big].\end{aligned}$$

Then we can set $\beta = 1/(128\kappa)$ and obtain

$$\begin{aligned}\mathbb{E}[\|\mathcal{S}_\eta^K\mathbf{q}_0 - \mathcal{H}_\eta^K\mathbf{q}_0'\|_2^2] \leq \big(1-(128\kappa)^{-1}\big)\mathbb{E}[\|\mathbf{q}_0 - \mathbf{q}_0'\|_2^2] \\ + 129\kappa\big[\gamma_1L^{-1/2}\eta + 2\gamma_2L^{-1}\big)\eta^2\big].\end{aligned} \tag{D.1}$$

Then we can consider a reference sequence $\{\widehat{\mathbf{x}}^{(t)}\}_{t=0,\ldots,T}$ with $\widehat{\mathbf{x}}^{(0)} \sim \pi$ and

$$\widehat{\mathbf{x}}^{(t+1)} = \mathcal{H}_\eta^K\widehat{\mathbf{x}}^{(t)},$$

where the initial momentum variable is the same as that used to compute $\mathbf{x}^{(t+1)}$. Then it can be readily verify that all iterates in such sequence follows the stationary distribution π. Therefore, applying (D.1) for T times gives

$$\mathbb{E}[\|\mathbf{x}^{(T)} - \widehat{\mathbf{x}}^{(T)}\|_2^2] \leq \big(1 - (128\kappa)^{-1}\big)^T \mathbb{E}[\|\mathbf{x}^{(0)} - \widehat{\mathbf{x}}^{(0)}\|_2^2] + 16512\kappa^2\big[\gamma_1 L^{-1/2}\eta + 2\gamma_2 L^{-1}\eta^2\big]. \tag{D.2}$$

Note that $\widehat{\mathbf{x}}^{(T)}$ and $\widehat{\mathbf{x}}^{(t)}$ are both following the target distribution π, we are able to complete the proof.

□

D.3. Proof of Lemma B.4

We first provide the following two useful lemmas that will be used during the proof.

Lemma D.4. Under Assumptions 4.1, 4.2 and 4.3, for any inner loop of Algorithm 1, given the initial quantities $\mathbf{q}_0$ and $\mathbf{p}_0$, it holds that for any $k \leq K$,

$$\begin{aligned}
\mathbb{E}[f(\mathbf{q}_k)] &\leq 3\bigg(f(\mathbf{q}_0) + \frac{\|\mathbf{p}_0\|_2^2}{2}\bigg) + 3K\eta^2(\sigma^2 - Lf(\mathbf{q}^*)) \\
\mathbb{E}[\|\nabla f(\mathbf{q}_k)\|_2] &\leq 3L\bigg(f(\mathbf{q}_0) + \frac{\|\mathbf{p}_0\|_2^2}{2}\bigg) + 3LK\eta^2(\sigma^2 - Lf(\mathbf{q}^*)) - Lf(\mathbf{q}^*) \\
\mathbb{E}[\|\mathbf{p}_k\|_2^2] &\leq 6\bigg(f(\mathbf{q}_0) + \frac{\|\mathbf{p}_0\|_2^2}{2}\bigg) + 6K\eta^2(\sigma^2 - Lf(\mathbf{q}^*)) - 2f(\mathbf{q}^*).
\end{aligned}$$

Lemma D.5. Under assumptions 4.2, if $\eta \leq 1/\sqrt{L}$, then it holds that for any $(\mathbf{q}_k, \mathbf{p}_k)$,

$$\begin{aligned}
\max_{\tau \leq \eta} \|\mathcal{H}_\tau \mathbf{p}_k\|_2^2 &\leq \|\mathbf{p}_k\|_2^2 + 2f(\mathbf{q}_k) - 2f(\mathbf{q}^*) \\
\max_{\tau \leq \eta} \|\nabla f(\mathcal{H}_\tau \mathbf{q}_k)\|_2^2 &\leq 2L\big(\|\mathbf{p}_k\|_2^2 + 2f(\mathbf{q}_k) - 2f(\mathbf{q}^*)\big) + 2\|\nabla f(\mathbf{q}_k)\|_2^2.
\end{aligned}$$

Proof of Lemma B.4. Instead of directly proving the upper bounds of E_p and E_q, we will prove the upper bound of $\mathbb{E}[f(\mathbf{x}^{(t)})]$ for all $t \geq 0$. Then applying Lemma D.4 and D.5, the upper bounds of E_p and E_q can be proved accordingly.

In particular, we will use mathematical induction prove the upper bounds of $\mathbb{E}[f(\mathbf{x}^{(t)})]$. In particular, we will prove the hypothesis that $\mathbb{E}[f(\mathbf{x}^{(t)})] - f(\mathbf{x}^*) \leq L\big[2D + 5d/\mu\big]$ based on the assumption that $\mathbb{E}[f(\mathbf{x}^{(\tau)})] \leq L\big[2D + 5d/\mu\big]$ holds for all $\tau \leq t$. Then let $E_f := L\big[2D + 5d/\mu\big]$, by Lemma D.4, applying $\mathbf{q}_0 = \mathbf{x}^{(t-1)}$, we get

$$\begin{aligned}
\mathbb{E}[f(\mathbf{q}_k)] - f(\mathbf{q}^*) &\leq 3\bigg(f(\mathbf{q}_0) - f(\mathbf{q}^*) + \frac{\|\mathbf{p}_0\|_2^2}{2}\bigg) + 3K\eta^2\sigma^2 \\
&\leq 3(E_f + d/2 + K\eta^2\sigma^2) \\
\mathbb{E}[\|\nabla f(\mathbf{q}_k)\|_2^2] &\leq 3L\bigg(f(\mathbf{q}_0) - f(\mathbf{q}^*) + \frac{\|\mathbf{p}_0\|_2^2}{2}\bigg) + 3LK\eta^2\sigma^2 \\
&\leq 3L(E_f + d/2 + K\eta^2\sigma^2) \\
\mathbb{E}[\|\mathbf{p}_k\|_2^2] &\leq 6\bigg(f(\mathbf{q}_0) - f(\mathbf{q}^*) + \frac{\|\mathbf{p}_0\|_2^2}{2}\bigg) + 6K\eta^2\sigma^2 \\
&\leq 6(E_f + d/2 + K\eta^2\sigma^2).
\end{aligned}$$

Then recall the definitions of E_p and E_q in Lemma B.1, by Lemma D.5, the quantities E_p and E_q can be upper bounded by

$$\begin{aligned}
E_p &\leq \mathbb{E}[\|\mathbf{p}_k\|_2^2] + 2\mathbb{E}[f(\mathbf{q}_k)] - 2f(\mathbf{q}^*) \leq 12(E_f + d/2 + K\eta^2\sigma^2) = O(\kappa d + K\eta^2\sigma^2) \\
E_q &\leq 2L\big(\mathbb{E}[\|\mathbf{p}_k\|_2^2] + 2\mathbb{E}[f(\mathbf{q}_k)] - 2f(\mathbf{q}^*)\big) + 2\mathbb{E}[\|\nabla f(\mathbf{q}_k)\|_2^2] \\
&\leq 30(E_f + d/2 + K\eta^2\sigma^2) = O(L\kappa d + LK\eta^2\sigma^2).
\end{aligned}$$

These are exactly the results we want to prove in this lemma. In what follows we will use these bounds to complete the induction of $\mathbb{E}[f(\mathbf{x}^{(t)})]$. In order to upper bound $\mathbb{E}[f(\mathbf{x}^{(t)})]$, we can leverage Assumption 4.2 and resort to bounding the

R.H.S. of the following inequality

$$\mathbb{E}[f(\mathbf{x}^{(t)})] \leq \frac{L\mathbb{E}[\|\mathbf{x}^{(t)} - \mathbf{x}^*\|_2^2]}{2} + f(\mathbf{x}^*). \tag{D.3}$$

Let $\mathbf{x}^\pi \sim \pi \propto e^{-f(\mathbf{x})}$, applying Young's inequality gives

$$\mathbb{E}[\|\mathbf{x}^{(t)} - \mathbf{x}^*\|_2^2] = \mathbb{E}[\|\mathbf{x}^{(t)} - \mathbf{x}^\pi + \mathbf{x}^\pi - \mathbf{x}^*\|_2^2] \leq 2\big[\mathbb{E}[\|\mathbf{x}^{(t)} - \mathbf{x}^\pi\|_2^2] + \mathbb{E}[\|\mathbf{x}^\pi - \mathbf{x}^*\|_2^2]\big]. \tag{D.4}$$

By Proposition 1 in (Durmus et al., 2019), we know that $\mathbb{E}[\|\mathbf{x}^\pi - \mathbf{x}^*\|_2^2] \leq d/\mu$. Then applying Lemma B.3, we know that

$$\begin{aligned}
\mathbb{E}[\|\mathbf{x}^{(t)} - \mathbf{x}^\pi\|_2^2] &\leq \mathbb{E}[\|\mathbf{x}^{(0)} - \mathbf{x}^\pi\|_2^2] + 16512\kappa^2\big[\gamma_1 L^{-1/2}\eta + 2\gamma_2 L^{-1}\eta^2\big] \\
&\leq 2\mathbb{E}[\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2] + 2\mathbb{E}[\|\mathbf{x}^{(\pi)} - \mathbf{x}^*\|_2^2] + 16512\kappa^2\big[\gamma_1 L^{-1/2}\eta + 2\gamma_2 L^{-1}\eta^2\big] \\
&\leq 2\mathbb{E}[\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2] + \frac{2d}{\mu} + 16512\kappa^2\big[\gamma_1 L^{-1/2}\eta + 2\gamma_2 L^{-1}\eta^2\big],
\end{aligned}$$

where $\gamma_1 = O(\sigma^2 L^{-1})$ and $\gamma_2 = O(L\kappa d + \sigma^2)$ are defined in Lemma B.1. Therefore, as long as the step size satisfies

$$\eta \leq \min\left\{\frac{d}{16512\gamma_1\kappa L^{1/2}}, \left(\frac{d}{33124\gamma_2\kappa}\right)^{1/2}\right\} = O\bigg(\frac{dL^{1/2}}{\sigma^2\kappa} \wedge \frac{d^{1/2}}{L^{1/2}\kappa d^{1/2} + \kappa^{1/2}\sigma}\bigg),$$

we have

$$\mathbb{E}[\|\mathbf{x}^{(t)} - \mathbf{x}^\pi\|_2^2] \leq 2\mathbb{E}[\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2] + \frac{4d}{\mu}. \tag{D.5}$$

Then plugging the above bound into (D.4) and (D.3), we get

$$\begin{aligned}
\mathbb{E}[f(\mathbf{x}^{(t)})] &\leq L\big[\mathbb{E}[\|\mathbf{x}^{(t)} - \mathbf{x}^\pi\|_2^2] + d/\mu\big] + f(\mathbf{x}^*) \\
&\leq L\big[2D + 5d/\mu\big] + f(\mathbf{x}^*),
\end{aligned}$$

where $D = \|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2$ is an absolute constant. This verifies the hypothesis for $\mathbf{x}^{(t)}$ and thus we are able to complete the proof. □

E. Proof of Lemmas in Appendix C

E.1. Proof of Lemma C.1

We first present the following lemma that characterizes the bound on the second moment of $\mathbf{q}_k - \mathbf{x}^*$.

Lemma E.1. Under Assumptions 4.2, 4.1, and 4.3, let $D = \|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2$. Then for any $\mathbf{q}_k$ with $k \leq K$, if $K\eta \leq 1/(2\sqrt{L})$, it holds that

$$\mathbb{E}[\|\mathbf{q}_k - \mathbf{x}^*\|_2^2] \leq L^{-1}\big[E_p + \eta^2(\sigma^2 + E_q)\big] + 8D + \frac{20d}{\mu}.$$

Proof of Lemma C.1. During the proof we will use $\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k)$ to denote the mini-batch stochastic gradient $\mathbf{g}_{SG}(\mathbf{q}_k, \boldsymbol{\xi}_k)$ for simplicity. Based on the definition, we have

$$\begin{aligned}
\mathbb{E}[\|\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) - \nabla f(\mathbf{q}_k)\|_2^2] &= \mathbb{E}\bigg[\bigg\|\frac{n}{B}\sum_{i\in\mathcal{I}_k}\nabla f_i(\mathbf{q}_k) - \nabla f(\mathbf{q}_k)\bigg\|_2^2\bigg] \\
&\leq \frac{1}{B}\mathbb{E}\big[\|n\nabla f_i(\mathbf{q}_k) - \nabla f(\mathbf{q}_k)\|_2^2\big].
\end{aligned}$$

By assumption 5.1, we know that $nf_i(\mathbf{x})$ and $f(\mathbf{x})$ are L-smooth, thus by Young's inequality, it holds that

$$\begin{aligned}
\|n\nabla f_i(\mathbf{q}_k) - \nabla f(\mathbf{q}_k)\|_2^2 &\leq 2\|n\nabla f_i(\mathbf{q}_k)\|_2^2 + 2\|\nabla f(\mathbf{q}_k)\|_2^2 \\
&\leq 2\|n\nabla f_i(\mathbf{q}_k) - nf_i(\mathbf{q}^*)\|_2^2 + 2n^2\|\nabla f_i(\mathbf{q}^*)\|_2^2 + 2\|\nabla f(\mathbf{q}_k)\|_2^2 \\
&\leq 2L^2\|\mathbf{q}_k - \mathbf{q}^*\|_2^2 + 2n^2\|\nabla f_i(\mathbf{q}^*)\|_2^2 + 2\|\nabla f(\mathbf{q}_k)\|_2^2.
\end{aligned}$$

By Lemma E.1 and the fact that $D \leq d/\mu$, we further have

$$\begin{aligned}\mathbb{E}[\|n\nabla f_i(\mathbf{q}_k) - \nabla f(\mathbf{q}_k)\|_2^2] &\leq 2L^2\mathbb{E}[\|\mathbf{q}_k - \mathbf{q}^*\|_2^2] + 2n^2\mathbb{E}[\|\nabla f_i(\mathbf{q}^*)\|_2^2] + 2E_q \\ &\leq 2L\big[E_p + \eta^2(\sigma^2 + E_q)\big] + 56L\kappa d + 2n^2\mathbb{E}[\|\nabla f_i(\mathbf{q}^*)\|_2^2] + 2E_q\end{aligned}$$

where the expectation is taken over both the randomness of $\mathbf{q}_k$ and i. Then it follows that

$$\mathbb{E}[\|\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) - \nabla f(\mathbf{q}_k)\|_2^2] \leq \frac{2L\big[E_p + \eta^2(\sigma^2 + E_q)\big] + 56L\kappa d + 2n^2\mathbb{E}[\|\nabla f_i(\mathbf{q}^*)\|_2^2] + 2E_q}{B}. \tag{E.1}$$

Let $\sigma_{\max}^2 = \arg\max_k \mathbb{E}[\|\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) - \nabla f(\mathbf{q}_k)\|_2^2]$. Then by Lemma B.4 and Assumption 4.3, assume that $L \geq 1$ and define by $\bar{E} = 12(2LD + 5\kappa d + d/2) \leq 96\kappa d$, where we use the fact that $D \leq d/\mu$, we know that

$$\begin{aligned}E_p &\leq \bar{E} + 12K\eta^2\sigma_{\max}^2 \\ E_q &\leq L\bar{E} + 12LK\eta^2\sigma_{\max}^2.\end{aligned} \tag{E.2}$$

Note that $K\eta \leq 1/(2\sqrt{L})$, plugging the above inequalities into (E.1) gives

$$\begin{aligned}\sigma_{\max}^2 &\leq \frac{2L\big[E_p + \eta^2(\sigma_{\max}^2 + E_q)\big] + 56L\kappa d + 2n^2\mathbb{E}[\|\nabla f_i(\mathbf{q}^*)\|_2^2] + 2E_q}{B} \\ &\leq \frac{28L^{1/2}\sigma_{\max}^2\eta + 5L\bar{E} + 56L\kappa d + 2n^2\mathbb{E}[\|\nabla f_i(\mathbf{q}^*)\|_2^2]}{B}.\end{aligned}$$

Then if $\eta \leq BL^{-1/2}/56$, the above inequality yields that

$$\begin{aligned}\sigma_{\max}^2 &\leq \frac{10L\bar{E} + 112L\kappa d + 4n^2\mathbb{E}[\|\nabla f_i(\mathbf{q}^*)\|_2^2]}{B} \\ &\leq \frac{1072L\kappa d + 4n^2\mathbb{E}[\|\nabla f_i(\mathbf{q}^*)\|_2^2]}{B},\end{aligned}$$

where the last inequality follows from the fact that $\bar{E} \leq 96\kappa d$. This completes the proof. □

E.2. Proof of Lemma C.4

Proof of Lemma C.4 . By the definition of the control variate gradient estimator, we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) - \nabla f(\mathbf{q}_k)\|_2^2] &= \mathbb{E}\bigg[\bigg\|\frac{n}{B}\sum_{i\in\mathcal{I}_k}[\nabla f_i(\mathbf{q}_k) - \nabla f_i(\widehat{\mathbf{q}})] + \nabla f(\widehat{\mathbf{q}}) - \nabla f(\mathbf{q})\bigg\|_2^2\bigg] \\ &\leq \frac{n^2}{B}\mathbb{E}\big[\|\nabla f_i(\mathbf{q}_k) - \nabla f_i(\widehat{\mathbf{q}}) + \nabla f(\widehat{\mathbf{q}}) - \nabla f(\mathbf{q})\|_2^2\big] \\ &\leq \frac{n^2}{B}\mathbb{E}\big[\|\nabla f_i(\mathbf{q}_k) - \nabla f_i(\widehat{\mathbf{q}})\|_2^2\big] \\ &\leq \frac{L^2}{B}\mathbb{E}[\|\mathbf{q}_k - \widehat{\mathbf{q}}\|_2^2],\end{aligned} \tag{E.3}$$

where the last inequality is by Assumption 5.1. By Lemma E.1 and Young's inequality, we know that

$$\begin{aligned}\mathbb{E}[\|\mathbf{q}_k - \widehat{\mathbf{q}}\|_2^2] &\leq 2\mathbb{E}[\|\mathbf{q}_k - \mathbf{q}^*\|_2^2] + 2\|\widehat{\mathbf{q}} - \mathbf{q}^*\|_2^2 \\ &\leq 2L^{-1}\big[E_p + \eta^2(\sigma^2 + E_q)\big] + 54d/\mu + 2\|\widehat{\mathbf{q}} - \mathbf{q}^*\|_2^2,\end{aligned}$$

where we use the fact that $D \leq d/\mu$. Note that we set $\mathbf{x}^{(0)} = \widehat{\mathbf{q}}$, which implies that $\|\widehat{\mathbf{q}} - \mathbf{q}^*\|_2^2 = D$. Then plugging the above inequality into (E.3) yields

$$\mathbb{E}[\|\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) - \nabla f(\mathbf{q}_k)\|_2^2] \leq \frac{2L\big[E_p + \eta^2(\sigma^2 + E_q)\big] + 56L\kappa d}{B}.$$

Note that the above inequality holds for any $\mathbf{q}_k$, then set $\sigma_{\max}^2 = \max_k \mathbb{E}[\|\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) - \nabla f(\mathbf{q}_k)\|_2^2]$. By (E.2) and using the fact that $K\eta \leq 1/(2\sqrt{L})$, we further have

$$\sigma_{\max}^2 \leq \frac{16L^{1/2}\sigma_{\max}^2\eta + 3L\bar{E} + 56L\kappa d}{B}.$$

Then if $\eta \leq BL^{-1/2}/32$, the above inequality yields

$$\sigma_{\max}^2 \leq \frac{6L\bar{E} + 112L\kappa d}{B} \leq \frac{688L\kappa d}{B},$$

where the second inequality is due to the fact that $\bar{E} \leq 96\kappa d$. This completes the proof.

□

E.3. Proof of Lemma C.2

Proof of Lemma C.2. In order to simplify the proof, we will use the short-hand notation $\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k)$ to denote the SVRG estimator $\mathbf{g}_{\mathrm{SVRG}}(\mathbf{q}_k, \boldsymbol{\xi}_k)$. Recall that

$$\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) = \frac{n}{B}\sum_{i\in\mathcal{I}_k}\big[\nabla f_i(\mathbf{q}_k) - \nabla f_i(\widetilde{\mathbf{q}})\big] + \nabla f(\widetilde{\mathbf{q}}).$$

Then based on the definition of the variance, we have

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) - \nabla f(\mathbf{q}_k)\|_2^2\big] &= \mathbb{E}\bigg[\bigg\|\frac{n}{B}\sum_{i\in\mathcal{I}_k}\big[\nabla f_i(\mathbf{q}_k) - \nabla f_i(\widetilde{\mathbf{q}})\big] + \nabla f(\widetilde{\mathbf{q}}) - \nabla f(\mathbf{q}_k)\bigg\|_2^2\bigg] \\ &\leq \frac{n^2}{B}\mathbb{E}\big[\|\nabla f_i(\mathbf{q}_k) - \nabla f_i(\widetilde{\mathbf{q}})\|_2^2\big],\end{aligned}$$

where the expectation is taken over both the random choice of $i \in [n]$ and the randomness of $\widetilde{\mathbf{q}}$ and $\mathbf{q}_k$. Then by Assumption 5.1, we have

$$\mathbb{E}\big[\|\nabla f_i(\mathbf{q}_k) - \nabla f_i(\widetilde{\mathbf{q}})\|_2^2\big] \leq \frac{L^2}{n^2}\mathbb{E}[\|\mathbf{q}_k - \widetilde{\mathbf{q}}\|_2^2]. \tag{E.4}$$

Note that $\widetilde{\mathbf{q}}$ is the iterate computed before $\mathbf{q}_k$. Assume its index is k', we have $k' \geq k - N + 1$ due to the definition of the SVRG estimator. Then, based on the update form (3.3), we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{q}_k - \widetilde{\mathbf{q}}\|_2^2] &= \mathbb{E}\bigg[\bigg\|\sum_{s=k'}^{k-1}\eta\mathbf{p}_s - \frac{\eta^2}{2}\mathbf{g}(\mathbf{q}_s, \boldsymbol{\xi}_s)\bigg\|_2^2\bigg] \\ &\leq N\sum_{s=k'}^{k-1}\mathbb{E}\big[\big\|\eta\mathbf{p}_s - \eta^2\mathbf{g}(\mathbf{q}_s, \boldsymbol{\xi}_s)/2\big\|_2^2\big] \\ &\leq 2N\sum_{s=k'}^{k-1}\mathbb{E}\big[\eta^2\|\mathbf{p}_s\|_2^2 + \eta^4\|\mathbf{g}(\mathbf{q}_s, \boldsymbol{\xi}_s)\|_2^2\big],\end{aligned} \tag{E.5}$$

where the first and second inequalities are due to Young's inequality. Note that we have $\mathbb{E}[\|\mathbf{p}_s\|_2^2] \leq E_p$ and

$$\mathbb{E}[\|\mathbf{g}(\mathbf{q}_s, \boldsymbol{\xi}_s)\|_2^2] = 2\mathbb{E}[\|\mathbf{g}(\mathbf{q}_s, \boldsymbol{\xi}_s) - \nabla f(\mathbf{q}_s)\|_2^2] + 2\mathbb{E}[\|\nabla f(\mathbf{q}_s)\|_2^2] \leq 2\mathbb{E}[\|\mathbf{g}(\mathbf{q}_s, \boldsymbol{\xi}_s) - \nabla f(\mathbf{q}_s)\|_2^2] + 2E_q, \tag{E.6}$$

where the first inequality is by Young's inequality and the second one is by the definition of E_q. Then plugging this back to (E.4) gives

$$\mathbb{E}\big[\|\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) - \nabla f(\mathbf{q}_k)\|_2^2\big] \leq \frac{2NL^2}{B}\sum_{s=k'}^{k-1}\big\{\eta^2 E_p + 2\eta^4\big[\|\mathbf{g}(\mathbf{q}_s, \boldsymbol{\xi}_s) - \nabla f(\mathbf{q}_s)\|_2^2 + E_q\big]\big\}.$$

Note that the above inequality holds for all k, then define $\sigma_{\max}^2 := \max_k \mathbb{E}\big[\|\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) - \nabla f(\mathbf{q}_k)\|_2^2\big]$, we have

$$\sigma_{\max}^2 \leq \frac{2N^2L^2}{B}\big[\eta^2 E_p + 2\eta^4\big(\sigma_{\max}^2 + E_q\big)\big].$$

By (E.2) and the fact that $K\eta \leq 1/(2\sqrt{L})$, the above inequality implies that

$$\begin{aligned}\sigma_{\max}^2 &\leq \frac{2N^2L^2}{B}\big[\eta^2(\bar{E} + 12K\eta^2\sigma_{\max}^2) + 2\eta^4\big(\sigma_{\max}^2 + L\bar{E} + 12LK\eta^2\sigma_{\max}^2\big)\big] \\ &\leq \frac{2N^2L^2}{B}\big[2\eta^2\bar{E} + 10\eta^3 L^{-1/2}\sigma_{\max}^2\big],\end{aligned}$$

where $\bar{E} = 12(2LD + 5\kappa d + d/2) \leq 96\kappa d$. Then if $\eta^3 \leq B/(40N^2L^{3/2})$, we have

$$\sigma_{\max}^2 \leq \frac{8N^2L^2\bar{E}\eta^2}{B} \leq \frac{768N^2L^2\eta^2\kappa d}{B},$$

where the last inequality follows from the fact that $\bar{E} \leq 96\kappa d$. This completes the proof. □

E.4. Proof of Lemma C.3

Proof of Lemma C.3. Similar to the proof of Lemma C.2, we denote by $\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) = \mathbf{g}_{\mathrm{SAG}}(\mathbf{q}_k, \boldsymbol{\xi}_k)$ in the subsequent proof. Recall that

$$\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) = \frac{1}{B}\sum_{i\in\mathcal{I}_k}\big[\nabla f_i(\mathbf{q}_k) - \mathbf{G}_i\big] + \widetilde{\mathbf{g}}_k.$$

Then the variance can be formulated as

$$\begin{aligned}\mathbb{E}[\|\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) - \nabla f(\mathbf{q}_k)\|_2^2] &= \mathbb{E}\bigg[\bigg\|\frac{n}{B}\sum_{i\in\mathcal{I}_k}\big[\nabla f_i(\mathbf{q}_k) - \mathbf{G}_i\big] + \widetilde{\mathbf{g}}_k - \nabla f(\mathbf{q}_k)\bigg\|\bigg] \\ &\leq \frac{n^2}{B}\mathbb{E}[\|\nabla f_i(\mathbf{q}_k) - \mathbf{G}_i + \widetilde{\mathbf{g}}_k - \nabla f(\mathbf{q}_k)\|_2^2] \\ &\leq \frac{n^2}{B}\mathbb{E}[\|\nabla f_i(\mathbf{q}_k) - \mathbf{G}_i\|_2^2],\end{aligned}$$

the first inequality is due to the fact that the estimation variance of sampling without replacement is always less than that of sampling with replacement, and the second inequality is due to the fact that $\mathbb{E}_i[\nabla f_i(\mathbf{q}) - \mathbf{G}_i] = \nabla f(\mathbf{q}_k) - \widetilde{\mathbf{g}}_k$. Based on the construction of the SAG estimator in Algorithm 2, we know that $\mathbf{G}_i = \nabla f_i(\mathbf{q}_u)$ for some $u \leq k$. Then by Assumption 5.1, it follows that

$$\mathbb{E}[\|\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) - \nabla f_i(\mathbf{q}_u)\|_2^2] \leq \frac{L^2\mathbb{E}[\|\mathbf{q}_k - \mathbf{q}_u\|_2^2]}{B}. \tag{E.7}$$

Then based on the update rule of $\mathbf{q}_k$, following (E.5) gives

$$\begin{aligned}\|\mathbf{q}_k - \mathbf{q}_u\|_2^2 &= \mathbb{E}\bigg[\bigg\|\sum_{s=u}^{k-1}\eta\mathbf{p}_s - \frac{\eta^2}{2}\mathbf{g}(\mathbf{q}_s, \boldsymbol{\xi}_s)\bigg\|_2^2\bigg] \\ &\leq 2(k-u)\sum_{s=u}^{k-1}\mathbb{E}\big[\eta^2\|\mathbf{p}_s\|_2^2 + \eta^4\|\mathbf{g}(\mathbf{q}_s, \boldsymbol{\xi}_s)\|_2^2\big].\end{aligned}$$

Define $\sigma_{\max}^2 := \max_k \mathbb{E}\big[\|\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) - \nabla f(\mathbf{q}_k)\|_2^2\big]$, we have the following based on (E.6),

$$\mathbb{E}[\|\mathbf{g}(\mathbf{q}_s, \boldsymbol{\xi}_s)\|_2^2] \leq 2\sigma_{\max}^2 + 2E_q.$$

Then using the fact that $\mathbb{E}[\|\mathbf{p}_s\|_2^2] \leq E_p$, it follows that

$$\|\mathbf{q}_k - \mathbf{q}_u\|_2^2 \leq 2(k-u)^2\eta^2(E_p + 2\eta^2\sigma_{\max}^2 + 2\eta^2 E_q).$$

Plugging the above inequality into (E.7) yields

$$\mathbb{E}[\|\mathbf{g}(\mathbf{q}_k,\boldsymbol{\xi}_k)-\nabla f(\mathbf{q}_k)\|_2^2] \leq \frac{L^2\mathbb{E}[\|\mathbf{q}_k-\mathbf{q}_u\|_2^2]}{B} \leq \frac{2\eta^2L^2(E_p+2\eta^2\sigma_{\max}^2+2\eta^2E_q)\mathbb{E}[(k-u)^2]}{B}.$$

Let $q = 1-(1-1/n)^B$ be the probability of choosing one particular index, by Dubey et al. (2016) we know that $\mathbb{E}[(k-u)^2] \leq 2q^{-2}$ (see proof of Theorem 2 in Dubey et al. (2016) for more detail). Moreover, it is easy to verify that $q \geq B/(2n)$, which implies that

$$\mathbb{E}[\|\mathbf{g}(\mathbf{q}_k,\boldsymbol{\xi}_k)-\nabla f(\mathbf{q}_k)\|_2^2] \leq \frac{2\eta^2L^2(E_p+2\eta^2\sigma_{\max}^2+2\eta^2E_q)n^2}{B^3}.$$

Note that the above holds for any $\mathbf{q}_k$, then applying (E.2) and the fact that $K\eta \leq 1/(2\sqrt{L})$ gives

$$\sigma_{\max}^2 \leq \frac{2\eta^2n^2L^2\big((1+2\eta^2L)\bar{E}+(12K\eta^2+12LK\eta^4+2\eta^2)\sigma_{\max}^2\big)}{B^3} \leq \frac{2\eta^2n^2L^2\big(2\bar{E}+10\eta L^{-1/2}\sigma_{\max}^2\big)}{B^3}.$$

Then if the step size satisfies $\eta^3 \leq B^3/(40n^2L^{3/2})$, we have

$$\sigma_{\max}^2 \leq \frac{8\eta^2n^2L^2\bar{E}}{B^3} \leq \frac{768\eta^2n^2L^2\kappa d}{B^3},$$

where the last inequality follows from the fact that $\bar{E} \leq 96\kappa d$. This completes the proof.

□

F. Proof of Lemmas in Appendices D and E

F.1. Proof of Lemma D.1

Proof of Lemma D.1. In terms of $\mathbf{q}_k$, based on Assumption 4.3, we have

$$\mathbb{E}[\|\mathcal{S}_\eta\mathbf{q}_k-\mathcal{G}_\eta\mathbf{q}_k\|_2^2] = \frac{\eta^4}{4}\mathbb{E}[\|\mathbf{g}(\mathbf{q}_k,\boldsymbol{\xi}_k)-\nabla f(\mathbf{q}_k)\|_2^2] \leq \frac{\eta^4\sigma^2}{4}.$$

In terms of $\mathbf{p}_k$, we have

$$\begin{aligned}\mathbb{E}[\|\mathcal{S}_\eta\mathbf{q}_k-\mathcal{G}_\eta\mathbf{q}_k\|_2^2] &= \frac{\eta^2}{4}\mathbb{E}[\|\mathbf{g}(\mathbf{q}_k,\boldsymbol{\xi}_k)+\mathbf{g}(\mathbf{q}_{k+1},\boldsymbol{\xi}_{k+1/2})-\mathbb{E}_{\boldsymbol{\xi}_k,\boldsymbol{\xi}_{k+1/2}}[\mathbf{g}(\mathbf{q}_k,\boldsymbol{\xi}_k)+\mathbf{g}(\mathbf{q}_{k+1},\boldsymbol{\xi}_{k+1/2})]\|_2^2]\\ &\leq \frac{\eta^2}{2}\big[\underbrace{\mathbb{E}[\|\mathbf{g}(\mathbf{q}_k,\boldsymbol{\xi}_k)-\nabla f(\mathbf{q}_k)\|_2^2]}_{I_1}\\ &\quad+\underbrace{\mathbb{E}[\mathbf{g}(\mathbf{q}_{k+1},\boldsymbol{\xi}_{k+1/2})-\mathbb{E}_{\boldsymbol{\xi}_k,\boldsymbol{\xi}_{k+1/2}}[\mathbf{g}(\mathbf{q}_{k+1},\boldsymbol{\xi}_{k+1/2})]\|_2^2]}_{I_2}\big],\end{aligned}$$

where the inequality is based on Young's inequality. Then we will bound the two terms on the R.H.S. separately. It is clearly that by Assumption 4.3, we have

$$I_1 \leq \sigma^2.$$

Additionally, in terms of I_2, since $\boldsymbol{\xi}_k$ and $\boldsymbol{\xi}_{k+1}$ are independent, we have the following by the law of total variance

$$\begin{aligned}I_2 &= \underbrace{\mathbb{E}_{\mathbf{q}_k,\boldsymbol{\xi}_k}\big[\mathbb{E}_{\boldsymbol{\xi}_{k+1/2}|\boldsymbol{\xi}_k}[\|\mathbf{g}(\mathbf{q}_{k+1},\boldsymbol{\xi}_{k+1/2})-\mathbb{E}_{\boldsymbol{\xi}_{k+1/2}|\boldsymbol{\xi}_k}[\mathbf{g}(\mathbf{q}_{k+1},\boldsymbol{\xi}_{k+1/2})]\|_2^2]\big]}_{I_3}\\ &\quad+\underbrace{\mathbb{E}_{\mathbf{q}_k,\boldsymbol{\xi}_k}\big[\|\mathbb{E}_{\boldsymbol{\xi}_{k+1/2}|\boldsymbol{\xi}_k}[\mathbf{g}(\mathbf{q}_{k+1},\boldsymbol{\xi}_{k+1/2})]-\mathbb{E}_{\boldsymbol{\xi}_k,\boldsymbol{\xi}_{k+1/2}}[\mathbf{g}(\mathbf{q}_{k+1},\boldsymbol{\xi}_{k+1/2})]\|_2^2\big]}_{I_4}.\end{aligned}$$

By Assumption 4.3, we have

$$I_3 \leq \sigma^2.$$

In terms of I_4, note that $\mathbb{E}_{\boldsymbol{\xi}_{k+1/2}|\boldsymbol{\xi}_k}[\mathbf{g}(\mathbf{q}_{k+1}, \boldsymbol{\xi}_{k+1/2})] = \nabla f(\mathbf{q}_{k+1})$, we have

$$\begin{aligned}
I_4 &\leq \mathbb{E}\big[\|\nabla f(\mathbf{q}_{k+1}) - \mathbb{E}_{\boldsymbol{\xi}_k}[\nabla f(\mathbf{q}_{k+1})]\|_2^2\big] \\
&\leq 2\mathbb{E}\big[\|\nabla f(\mathbf{q}_{k+1}) - \mathbb{E}_{\boldsymbol{\xi}_k}[\nabla f(\mathbf{q}_{k+1})]\|_2^2\big] \\
&\leq 2\mathbb{E}\big[\|\nabla f(\mathbf{q}_{k+1}) - \nabla f(\mathbb{E}_{\boldsymbol{\xi}_k}[\mathbf{q}_{k+1}])]\|_2^2\big] \\
&\leq 2L^2\mathbb{E}\big[\|\mathbf{q}_{k+1} - \mathbb{E}_{\boldsymbol{\xi}_k}[\mathbf{q}_{k+1}]]\|_2^2\big],
\end{aligned}$$

where the second inequality is by Young's inequality, the third one is due to the fact that $\nabla f(\mathbf{q}_{k+1})$ is an unbiased estimator of $\mathbb{E}_{\boldsymbol{\xi}_k}[\nabla f(\mathbf{q}_{k+1})]$ and the last inequality is by Assumption 4.2. Then by Assumption 4.3 and the definition of $\mathcal{S}_\eta$, we have

$$\mathbb{E}\big[\|\mathbf{q}_{k+1} - \mathbb{E}_{\boldsymbol{\xi}_k}[\mathbf{q}_{k+1}]]\|_2^2\big] \leq \frac{\eta^4\sigma^2}{4}.$$

Therefore, combining the above results gives

$$\mathbb{E}_{\boldsymbol{\xi}_k}[\|\mathcal{S}_\eta\mathbf{q}_k - \mathcal{G}_\eta\mathbf{q}_k\|_2^2] \leq \frac{\eta^2}{2}(I_1 + I_2) \leq \frac{\eta^2}{4}\big(4\sigma^2 + \eta^4L^2\sigma^2\big),$$

which completes the proof. □

F.2. Proof of Lemma D.2

Proof of Lemma D.2. We first prove the first result in this lemma. By the definitions of $\mathcal{H}_\eta$ and $\mathcal{G}_\eta$, note that $\mathbb{E}_{\boldsymbol{\xi}_k}[\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k)] = \nabla f(\mathbf{q}_k)$, defining $(\mathbf{q}(0), \mathbf{p}(0)) = (\mathbf{p}_k, \mathbf{q}_k)$, we have

$$\begin{aligned}
\mathcal{G}_\eta\mathbf{q}_k - \mathcal{H}_\eta\mathbf{q}_k &= \eta\mathbf{p}_k - \frac{\eta^2}{2}\nabla f(\mathbf{q}_k) - \int_0^\eta \mathbf{p}(t)\mathrm{d}t \\
&= \int_0^\eta\int_0^t [\nabla f(\mathbf{q}(s)) - \nabla f(\mathbf{q}(0))]\mathrm{d}s\mathrm{d}t.
\end{aligned}$$

Therefore, by Assumption 4.2, we have

$$\begin{aligned}
\|\mathcal{G}_\eta\mathbf{q}_k - \mathcal{H}_\eta\mathbf{q}_k\|_2^2 &\leq \bigg[\int_0^\eta\int_0^t \|\nabla f(\mathbf{q}(s)) - \nabla f(\mathbf{q}(0))\|_2\mathrm{d}s\mathrm{d}t\bigg]^2 \\
&\leq L^2\bigg[\int_0^\eta\int_0^t \|\mathbf{q}(s) - \mathbf{q}(0)\|_2\mathrm{d}s\mathrm{d}t\bigg]^2.
\end{aligned}$$

Note that $\mathrm{d}\mathbf{q}(s)/\mathrm{d}s = \mathbf{p}(s)$ and $s \leq \eta$, thus it holds that $\|\mathbf{q}(s) - \mathbf{q}(0)\|_2 \leq s\max_{\tau\in[0,\eta]}\|\mathbf{p}(\tau)\|_2$. Then, we have

$$\begin{aligned}
\mathbb{E}[\|\mathcal{G}_\eta\mathbf{q}_k - \mathcal{H}_\eta\mathbf{q}_k\|_2^2] &\leq L^2\mathbb{E}\bigg[\bigg[\int_0^\eta\int_0^t \|\mathbf{q}(s) - \mathbf{q}(0)\|_2\mathrm{d}s\mathrm{d}t\bigg]^2\bigg] \\
&\leq \frac{L^2\eta^6}{36}\mathbb{E}[\max_{\tau\in[0,\eta]}\|\mathbf{p}(\tau)\|_2^2] \\
&\leq \frac{L^2\eta^6}{36}E_p.
\end{aligned}$$

Then we will prove the upper bound of $\mathbb{E}\|\mathcal{G}_\eta\mathbf{p}_k - \mathcal{H}_\eta\mathbf{p}_k\|_2^2]$. We first introduce an intermediate momentum variable $\widehat{\mathbf{p}}_{k+1}$ defined by

$$\widehat{\mathbf{p}}_{k+1} = \mathbf{p}_k - \frac{\eta}{2}\nabla f(\mathbf{q}_k) - \frac{\eta}{2}\nabla f\big(\mathbf{q}_k + \eta\mathbf{p}_k - \eta^2\nabla f(\mathbf{q}_k)/2\big).$$

It can be observed that $\widehat{\mathbf{p}}_{k+1}$ is exactly the momentum variable obtained after one step full-gradient update of HMC. Then, by Young's inequality, we have

$$\mathbb{E}[\|\mathcal{G}_\eta \mathbf{p}_k - \mathcal{H}_\eta \mathbf{p}_k\|_2^2] \leq 2\underbrace{\mathbb{E}[\|\mathcal{G}_\eta \mathbf{p}_k - \widehat{\mathbf{p}}_{k+1}\|_2^2]}_{I_1} + 2\underbrace{\mathbb{E}[\|\mathcal{H}_\eta \mathbf{p}_k - \widehat{\mathbf{p}}_{k+1}\|_2^2]}_{I_2}.$$

Then regarding I_1, note that $\mathbb{E}_{\boldsymbol{\xi}_k}[\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k)] = \nabla f(\mathbf{q}_k)$ and $\mathbb{E}_{\boldsymbol{\xi}_{k+1/2}|\boldsymbol{\xi}_k}[\mathbf{g}(\mathbf{q}_{k+1}, \boldsymbol{\xi}_{k+1/2})] = \nabla f(\mathbf{q}_{k+1})$, we have

$$\begin{aligned} I_1 &= \frac{\eta^2}{4}\mathbb{E}\big[\big\|\nabla f(\mathbf{q}_k) + \mathbb{E}_{\boldsymbol{\xi}_k, \boldsymbol{\xi}_{k+1/2}}[\mathbf{g}(\mathbf{q}_{k+1}, \boldsymbol{\xi}_{k+1/2})] - \nabla f(\mathbf{q}_k) - \nabla f\big(\mathbf{q}_k + \eta \mathbf{p}_k - \eta^2 \nabla f(\mathbf{q}_k)/2\big)\big\|_2^2\big] \\ &\leq \frac{\eta^2}{4}\mathbb{E}\big[\big\|\mathbb{E}_{\boldsymbol{\xi}_k}[\nabla f(\mathbf{q}_{k+1})] - \nabla f\big(\mathbf{q}_k + \eta \mathbf{p}_k - \eta^2 \nabla f(\mathbf{q}_k)/2\big)\big\|_2^2\big]. \end{aligned}$$

Note that $\mathbf{q}_{k+1} = \mathbf{q}_k + \eta \mathbf{p}_k - \eta^2 \mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k)/2$, by Assumptions 4.2 and 4.3 we further have

$$\begin{aligned} \mathbb{E}\big[\|\mathbb{E}_{\boldsymbol{\xi}_k}[\nabla f(\mathbf{q}_{k+1})] - \nabla f\big(\mathbf{q}_k + \eta \mathbf{p}_k - \eta^2 \nabla f(\mathbf{q}_k)/2\big)\|_2^2\big] &\leq \mathbb{E}\big[\|\nabla f(\mathbf{q}_{k+1}) - \nabla f\big(\mathbf{q}_k + \eta \mathbf{p}_k - \eta^2 \nabla f(\mathbf{q}_k)/2\big)\|_2^2\big] \\ &\leq \frac{\eta^4 L^2}{4}\mathbb{E}\big[\|\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) - \nabla f(\mathbf{q}_k)\|_2^2\big] \\ &\leq \frac{\eta^4 L^2 \sigma^2}{4}. \end{aligned}$$

This immediately implies that

$$I_1 \leq \frac{\eta^6 L^2 \sigma^2}{16}.$$

In terms of I_2, we can follow the similar idea used in the proof of Lemma 9.1 in (Mangoubi & Vishnoi, 2018). In particular, we further define an intermediate point $\widetilde{\mathbf{p}}_{k+1}$ as follows

$$\widetilde{\mathbf{p}}_{k+1} = \mathbf{p}_k - \eta \nabla f(\mathbf{q}_k) - \frac{\eta^2}{2}\nabla^2 f(\mathbf{q}_k)\mathbf{p}_k.$$

Therefore, by triangle inequality and Young's inequality, I_2 can be upper bounded by

$$I_2 \leq 2\mathbb{E}[\|\widetilde{\mathbf{p}}_{k+1} - \widehat{\mathbf{p}}_{k+1}\|_2^2] + 2\mathbb{E}[\|\mathcal{H}_\eta \mathbf{p}_k - \widehat{\mathbf{p}}_{k+1}\|_2^2].$$

Based on the definition of $\widetilde{\mathbf{p}}_{k+1}$ and $\widehat{\mathbf{p}}_{k+1}$, we have

$$\begin{aligned} \|\widetilde{\mathbf{p}}_{k+1} - \widehat{\mathbf{p}}_{k+1}\|_2 &= \frac{\eta}{2}\big\|\nabla f(\mathbf{q}_k + \eta \mathbf{p}_k - \eta^2 \nabla f(\mathbf{q}_k)/2) - \nabla f(\mathbf{q}_k) - \eta \nabla^2 f(\mathbf{q}_k)\mathbf{p}_k\big\|_2 \\ &\leq \frac{\eta}{2}\big(\|\nabla f(\mathbf{q}_k + \eta \mathbf{p}_k - \eta^2 \nabla f(\mathbf{q}_k)/2) - \nabla f(\mathbf{q}_k + \eta \mathbf{p}_k)\|_2 \\ &\quad + \|\nabla f(\mathbf{q}_k + \eta \mathbf{p}_k) - \nabla f(\mathbf{q}_k) - \eta \nabla^2 f(\mathbf{q}_k)\mathbf{p}_k\|_2\big). \end{aligned}$$

Based on Assumption 4.2, we have

$$\|\nabla f(\mathbf{q}_k + \eta \mathbf{p}_k - \eta^2 \nabla f(\mathbf{q}_k)/2) - \nabla f(\mathbf{q}_k + \eta \mathbf{p}_k)\|_2 \leq \frac{\eta^2 L \|\nabla f(\mathbf{q}_k)\|_2}{2}$$

Additionally, we have

$$\begin{aligned} \|\nabla f(\mathbf{q}_k + \eta \mathbf{p}_k) - \nabla f(\mathbf{q}_k) - \eta \nabla^2 f(\mathbf{q}_k)\mathbf{p}_k\|_2 &= \bigg\|\int_0^\eta \big[\nabla^2 f(\mathbf{q}_k + t\mathbf{p}_k) - \nabla^2 f(\mathbf{q}_k)\big]\mathbf{p}_k \mathrm{d}t\bigg\|_2 \\ &\leq \int_0^\eta \big\|[\nabla^2 f(\mathbf{q}_k + t\mathbf{p}_k) - \nabla^2 f(\mathbf{q}_k)]\mathbf{p}_k\big\|_2 \mathrm{d}t \\ &\leq 2\eta L \|\mathbf{p}_k\|_2 \end{aligned}$$

where the last inequality follows from Assumption 4.2. Combining the above two inequalities gives

$$\mathbb{E}[\|\widetilde{\mathbf{p}}_{k+1}-\widehat{\mathbf{p}}_{k+1}\|_2^2]\leq\mathbb{E}\bigg[\frac{\eta^4(L\eta\|\nabla f(\mathbf{q}_k)\|_2+4L\|\mathbf{p}_k\|)^2}{16}\bigg]\leq\frac{\eta^6(L^2E_q+16L^2E_p)}{8}.$$

Then we will be proving the upper bound of $\|\mathcal{H}_\eta\mathbf{p}_k-\widehat{\mathbf{p}}_{k+1}\|_2$. In specific, note that $(\mathbf{q}(0),\mathbf{p}(0))=(\mathbf{q}_k,\mathbf{p}_k)$, we have

$$\begin{aligned}\|\mathcal{H}_\eta\mathbf{p}_k-\widehat{\mathbf{p}}_{k+1}\|_2&=\bigg\|\int_0^\eta\int_0^t\big[\nabla^2 f(\mathbf{q}(s))\mathbf{p}(s)-\nabla^2 f(\mathbf{q}(0))\mathbf{p}(0)\big]\mathrm{d}s\mathrm{d}t\bigg\|_2\\&\leq\int_0^\eta\int_0^t\big\|\nabla^2 f(\mathbf{q}(s))\mathbf{p}(s)-\nabla^2 f(\mathbf{q}(0))\mathbf{p}(0)\big\|_2\mathrm{d}s\mathrm{d}t.\end{aligned}$$

We further have

$$\begin{aligned}\big\|\nabla^2 f(\mathbf{q}(s))\mathbf{p}(s)-\nabla^2 f(\mathbf{q}(0))\mathbf{p}(0)\big\|_2&\leq\big\|\nabla^2 f(\mathbf{q}(s))\big(\mathbf{p}(s)-\mathbf{p}(0)\big)\big\|_2+\big\|\big[\nabla^2 f(\mathbf{q}(s))-\nabla^2 f(\mathbf{q}(s))\big]\mathbf{p}(0)\big\|_2\\&\leq L\|\mathbf{p}(s)-\mathbf{p}(0)\|_2+2L\|\mathbf{p}(0)\|_2.\end{aligned}$$

Note that $s\leq\eta$, we further have

$$\begin{aligned}\mathbb{E}[\big\|\nabla^2 f(\mathbf{q}(s))\mathbf{p}(s)-\nabla^2 f(\mathbf{q}(0))\mathbf{p}(0)\big\|_2^2]&\leq 2s^2L^2\mathbb{E}[\max_{\tau\leq\eta}\|\nabla f(\mathbf{q}(\tau))\|_2^2]+4L^2\mathbb{E}[\|\mathbf{p}(0)\|_2^2]\\&\leq 2\eta^2L^2E_q+4L^2E_p^2.\end{aligned}$$

Combining the above results we have

$$\mathbb{E}[\|\mathcal{H}_\eta\mathbf{p}_k-\widehat{\mathbf{p}}_{k+1}\|_2^2]\leq\eta^4\big(\eta^2L^2E_q+2L^2E_p\big).$$

Then, we can bound I_2 as follows,

$$\begin{aligned}I_2&\leq 2\mathbb{E}[\|\widetilde{\mathbf{p}}_{k+1}-\widehat{\mathbf{p}}_{k+1}\|_2^2]+2\mathbb{E}[\|\mathcal{H}_\eta\mathbf{p}_k-\widehat{\mathbf{p}}_{k+1}\|_2^2]\\&\frac{\eta^6(L^2E_q+16L^2E_p)}{4}+2\eta^4\big(\eta^2L^2E_q+2L^2E_p\big)\\&\leq 4\eta^4(\eta^2L^2E_q+2L^2E_p).\end{aligned}$$

Then, it follows that

$$\mathbb{E}[\|\mathcal{G}_\eta\mathbf{p}_k-\mathcal{H}_\eta\mathbf{p}_k\|_2^2]\leq 2(I_1+I_2)\leq\frac{\eta^6L^2\sigma^2}{8}+8\eta^4L^2(\eta^2E_q+2E_p),$$

which completes the proof.

□

F.3. Proof of Lemma D.3

Proof of Lemma D.3. By (3.1), let $\mathbf{q}(t)=\mathcal{H}_t\mathbf{q}(0)$ and $\mathbf{p}(t)=\mathcal{H}_t\mathbf{p}(0)$, we have

$$\begin{aligned}&\frac{\mathrm{d}\big[\|\mathbf{q}(t)-\mathbf{q}'(t)\|_2^2+L^{-1}\|\mathbf{p}(t)-\mathbf{p}'(t)\|_2^2\big]}{\mathrm{d}t}\\&=2\langle\mathbf{q}(t)-\mathbf{q}'(t),\mathbf{p}(t)-\mathbf{p}'(t)\rangle+2L^{-1}\langle\mathbf{p}(t)-\mathbf{p}'(t),\nabla f(\mathbf{q}'(t))-\nabla f(\mathbf{q}(t))\rangle\\&\leq 4\|\mathbf{q}(t)-\mathbf{q}'(t)\|_2\|\mathbf{q}(t)-\mathbf{q}'(t)\|\\&\leq 2\big[L^{1/2}\|\mathbf{q}(t)-\mathbf{q}'(t)\|_2^2+L^{-1/2}\|\mathbf{p}(t)-\mathbf{p}'(t)\|_2^2\big]\\&=2L^{1/2}\big[\|\mathbf{q}(t)-\mathbf{q}'(t)\|_2^2+L^{-1}\|\mathbf{p}(t)-\mathbf{p}'(t)\|_2^2\big],\end{aligned}$$

where the first inequality is based on Assumption 4.2 and Cauchy-Schwarz inequality, and the second inequality is due the fact that $2ab\leq\beta a^2+\beta^{-1}b^2$ holds for all a, b and $\beta>0$. Solving the above inequality directly implies the desired result. □

F.4. Proof of Lemma D.4

Proof of Lemma D.4. Recall the update rules of $\mathbf{q}_k$ and $\mathbf{p}_k$

$$\mathbf{q}_{k+1} = \mathbf{q}_k + \eta \mathbf{p}_k - \frac{\eta^2}{2}\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k)$$
$$\mathbf{p}_{k+1} = \mathbf{p}_k - \frac{\eta}{2}\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) - \frac{\eta}{2}\mathbf{g}(\mathbf{q}_{k+1}, \boldsymbol{\xi}_{k+1/2}).$$

Then motivated by the fact that the Hamiltonian energy maintains invariant along the continuous-time evolution (3.1), we will leverage the energy function $f(\mathbf{q}_{k+1}) + \|\mathbf{p}_{k+1}\|_2^2/2$ to prove the desired upper bounds. Based on the update rule of $\mathbf{p}_{k+1}$ and $\mathbf{q}_{k+1}$, by Assumption 4.2 we have

$$\begin{aligned}
f(\mathbf{q}_{k+1}) &\leq f(\mathbf{q}_k) + \langle \nabla f(\mathbf{q}_k), \mathbf{q}_{k+1} - \mathbf{q}_k \rangle + \frac{L\|\mathbf{q}_{k+1} - \mathbf{q}_k\|_2^2}{2} \\
&= f(\mathbf{q}_k) + \eta \langle \nabla f(\mathbf{q}_k), \mathbf{p}_k - \eta \mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k)/2 \rangle + \frac{\eta^2 L \|\mathbf{p}_k - \eta \mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k)/2\|_2^2}{2}; \\
\|\mathbf{p}_{k+1}\|_2^2 &= \Big\| \mathbf{p}_k - \frac{\eta}{2}\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) - \frac{\eta}{2}\mathbf{g}(\mathbf{q}_{k+1}, \boldsymbol{\xi}_{k+1/2}) \Big\|_2^2 \\
&= \|\mathbf{p}_k\|_2^2 - \eta \langle \mathbf{p}_k, \mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) + \mathbf{g}(\mathbf{q}_{k+1}, \boldsymbol{x}_{k+1/2}) \rangle + \frac{\eta^2}{4} \big\| \mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) + \mathbf{g}(\mathbf{q}_{k+1}, \boldsymbol{\xi}_{k+1/2}) \big\|_2^2.
\end{aligned}$$

Then taking expectation over the randomness $\boldsymbol{\xi}_k$ and $\boldsymbol{\xi}_{k+1/2}$ conditioned on $\mathbf{q}_k$ and $\mathbf{p}_k$, we have

$$\begin{aligned}
&\mathbb{E}[f(\mathbf{q}_{k+1})] + \frac{\mathbb{E}[\|\mathbf{p}_{k+1}\|_2^2]}{2} \\
&\leq f(\mathbf{q}_k) + \eta \langle \nabla f(\mathbf{q}_k), \mathbf{p}_k - \eta \nabla f(\mathbf{q}_k)/2 \rangle + \frac{\eta^2 L \mathbb{E}[\|\mathbf{p}_k - \eta \mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k)/2\|_2^2]}{2} \\
&\quad + \frac{\|\mathbf{p}_k\|_2^2}{2} - \frac{\eta}{2}\big\langle \mathbf{p}_k, \nabla f(\mathbf{q}_k) + \mathbb{E}[\nabla f(\mathbf{q}_{k+1})] \big\rangle + \frac{\eta^2}{8}\mathbb{E}\big[\big\| \mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) + \mathbf{g}(\mathbf{q}_{k+1}, \boldsymbol{\xi}_{k+1/2}) \big\|_2^2 \big] \\
&= f(\mathbf{q}_k) + \frac{\|\mathbf{p}_k\|_2^2}{2} - \frac{\eta^2 \|\nabla f(\mathbf{q}_k)\|_2^2}{2} + \frac{\eta}{2}\big\langle \mathbf{p}_k, \nabla f(\mathbf{q}_k) - \mathbb{E}[\nabla f(\mathbf{q}_{k+1})] \big\rangle \\
&\quad + \frac{\eta^2 L \mathbb{E}[\|\mathbf{p}_k - \eta \mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k)\|_2^2]}{2} + \frac{\eta^2}{8}\mathbb{E}\big[\big\| \mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) + \mathbf{g}(\mathbf{q}_{k+1}, \boldsymbol{\xi}_{k+1/2}) \big\|_2^2 \big].
\end{aligned} \tag{F.1}$$

Then we first focus on the term $\big\langle \mathbf{p}_k, \nabla f(\mathbf{q}_k) - \mathbb{E}[\nabla f(\mathbf{q}_{k+1})] \big\rangle$, which can be upper bounded as follows,

$$\begin{aligned}
\big\langle \mathbf{p}_k, \nabla f(\mathbf{q}_k) - \mathbb{E}[\nabla f(\mathbf{q}_{k+1})] \big\rangle &\leq \|\mathbf{p}_k\|_2 \|\nabla f(\mathbf{q}_k) - \mathbb{E}[\nabla f(\mathbf{q}_{k+1})]\|_2 \\
&\leq \|\mathbf{p}_k\|_2 \mathbb{E}[\|\nabla f(\mathbf{q}_k) - \nabla f(\mathbf{q}_{k+1})\|_2] \\
&\leq \eta L \|\mathbf{p}_k\|_2 \mathbb{E}\big[\|\mathbf{p}_k - \eta \mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k)/2\|_2\big] \\
&\leq \frac{\eta L}{2}\big(\|\mathbf{p}_k\|_2^2 + \mathbb{E}[\|\mathbf{p}_k - \eta \mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k)/2\|_2^2] \big),
\end{aligned}$$

where the first inequality is by Cauchy-Schwartz inequality, the second inequality is by triangle inequality, the third inequality is based on Assumption 4.2 and the last one follows from Young's inequality. We next focus on bounding $\mathbb{E}\big[\big\| \mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) + \mathbf{g}(\mathbf{q}_{k+1}, \boldsymbol{\xi}_{k+1/2}) \big\|_2^2\big]$, which relies on the following inequality.

$$\begin{aligned}
\mathbb{E}[\|\nabla f(\mathbf{q}_{k+1})\|_2^2] &\leq \mathbb{E}\big[\big(\|\nabla f(\mathbf{q}_k)\|_2 + \|\nabla f(\mathbf{q}_{k+1}) - \nabla f(\mathbf{q}_k)\|_2 \big)^2 \big] \\
&\leq \mathbb{E}\big[\big(\|\nabla f(\mathbf{q}_k)\|_2 + \eta L \|\mathbf{p}_k - \eta \mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k)/2\|_2 \big)^2 \big] \\
&= \|\nabla f(\mathbf{q}_k)\|_2^2 + 2\eta L \mathbb{E}\big[\|\nabla f(\mathbf{q}_k)\|_2 \cdot \|\mathbf{p}_k - \mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k)/2\|_2 \big] \\
&\quad + \eta^2 L^2 \mathbb{E}[\|\mathbf{p}_k - \mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k)/2\|_2^2] \\
&\leq (1 + \eta L)\|\nabla f(\mathbf{q}_k)\|_2^2 + (\eta L + \eta^2 L^2)\mathbb{E}[\|\mathbf{p}_k - \mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k)/2\|_2^2],
\end{aligned} \tag{F.2}$$

where the first inequality is by triangle inequality, the second one follows from Assumption 4.2, and the last one follows from Young's inequality. Based on the above inequality, we further have

$$\begin{aligned}\mathbb{E}\big[\big\|\mathbf{g}(\mathbf{q}_k,\boldsymbol{\xi}_k)+\mathbf{g}(\mathbf{q}_{k+1},\boldsymbol{\xi}_{k+1/2})\big\|_2^2\big] &\leq 2\mathbb{E}\big[\big\|\mathbf{g}(\mathbf{q}_k,\boldsymbol{\xi}_k)\|_2^2+\|\mathbf{g}(\mathbf{q}_{k+1},\boldsymbol{\xi}_{k+1/2})\big\|_2^2\big]\\ &\leq 2\big[\|\nabla f(\mathbf{q}_k)\|_2^2+2\sigma^2+\mathbb{E}[\|\nabla f(\mathbf{q}_{k+1})\|_2^2]\big]\\ &= 4\big[\|\nabla f(\mathbf{q}_k)\|_2^2+\sigma^2\big]+2\mathbb{E}[\|\nabla f(\mathbf{q}_{k+1})\|_2^2-\|\nabla f(\mathbf{q}_k)\|_2^2]\\ &\leq 4\big[\|\nabla f(\mathbf{q}_k)\|_2^2+\sigma^2\big]+2\eta L\|\nabla f(\mathbf{q}_k)\|_2^2\\ &\quad+2(\eta L+\eta^2L^2)\mathbb{E}[\|\mathbf{p}_k-\mathbf{g}(\mathbf{q}_k,\boldsymbol{\xi}_k)/2\|_2^2].\end{aligned}$$

where the first inequality follows from Young's inequality, the second follows from Assumption 4.3, and the last follows from (F.2). Plugging the above results into (F.1), we get

$$\begin{aligned}&\mathbb{E}[f(\mathbf{q}_{k+1})]+\frac{\mathbb{E}[\|\mathbf{p}_{k+1}\|_2^2]}{2}\\ &\leq f(\mathbf{q}_k)+\frac{\|\mathbf{p}_k\|_2^2}{2}-\frac{\eta^2\|\nabla f(\mathbf{q}_k)\|_2^2}{2}+\eta^2L\big(\|\mathbf{p}_k\|_2^2+\mathbb{E}[\|\mathbf{p}_k-\eta\mathbf{g}(\mathbf{q}_k,\boldsymbol{\xi}_k)/2\|_2^2]\big)\\ &\quad+\frac{\eta^2L\mathbb{E}[\|\mathbf{p}_k-\eta\mathbf{g}(\mathbf{q}_k,\boldsymbol{\xi}_k)\|_2^2]}{2}+\frac{\eta^2}{8}\Big(4\big[\|\nabla f(\mathbf{q}_k)\|_2^2+\sigma^2\big]+2\eta L\|\nabla f(\mathbf{q}_k)\|_2^2\\ &\quad+2(\eta L+\eta^2L^2)\mathbb{E}[\|\mathbf{p}_k-\mathbf{g}(\mathbf{q}_k,\boldsymbol{\xi}_k)/2\|_2^2]\Big)\\ &=f(\mathbf{q}_k)+\frac{\|\mathbf{p}_k\|_2^2}{2}+\eta^2L\|\mathbf{p}_k\|_2^2+\bigg(\frac{3\eta^2L}{2}+\frac{\eta^2(\eta L+\eta^2L^2)}{4}\bigg)\mathbb{E}[\|\mathbf{p}_k-\eta\mathbf{g}(\mathbf{q}_k,\boldsymbol{\xi}_k)\|_2^2]\\ &\quad+\frac{\eta^3L\|\nabla f(\mathbf{q}_k)\|_2^2}{4}+\frac{\eta^2\sigma^2}{2}.\end{aligned}$$

Moreover, by Young's inequality and Assumption 4.3, we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{p}_k-\eta\mathbf{g}(\mathbf{q}_k,\boldsymbol{\xi}_k)\|_2^2] &\leq 2\|\mathbf{p}_k\|_2^2+2\eta^2\mathbb{E}[\|\mathbf{g}(\mathbf{q}_k,\boldsymbol{\xi}_k)\|_2^2]\\ &\leq 2\|\mathbf{p}_k\|_2^2+2\eta^2\|\nabla f(\mathbf{q}_k)\|_2+2\eta^2\sigma^2.\end{aligned}$$

Therefore, assume $\eta\leq\min\{1/(2\sqrt{L}),1/2\}$, we have

$$\mathbb{E}[f(\mathbf{q}_{k+1})]+\frac{\mathbb{E}[\|\mathbf{p}_{k+1}\|_2^2]}{2}\leq f(\mathbf{q}_k)+\frac{\|\mathbf{p}_k\|_2^2}{2}+3\eta^2L\|\mathbf{p}_k\|_2^2+4\eta^3L\|\nabla f(\mathbf{q}_k)\|_2^2+\eta^2\sigma^2.$$

Note that for convex and smooth function $f(\cdot)$, we have $\|\nabla f(\mathbf{q}_k)\|_2^2\leq L(f(\mathbf{q}_k)-f(\mathbf{q}^*))$, therefore,

$$\mathbb{E}[f(\mathbf{q}_{k+1})]+\frac{\mathbb{E}[\|\mathbf{p}_{k+1}\|_2^2]}{2}\leq f(\mathbf{q}_k)+\frac{\|\mathbf{p}_k\|_2^2}{2}+3\eta^2L\|\mathbf{p}_k\|_2^2+4\eta^2L\big[f(\mathbf{q}_k)-f(\mathbf{q}^*)\big]+\eta^2\sigma^2,$$

which further implies that

$$\mathbb{E}[f(\mathbf{q}_{k+1})-f(\mathbf{q}^*)]+\frac{\mathbb{E}[\|\mathbf{p}_{k+1}\|_2^2]}{2}\leq\big(1+4\eta^2L\big)\bigg(f(\mathbf{q}_k)-f(\mathbf{q}^*)+\frac{\|\mathbf{p}_k\|_2^2}{2}\bigg)+\eta^2\sigma^2$$

Then, assume that $4K\eta^2L\leq 1$, we have $(1+4\eta^2L)^K\leq e\leq 3$, which implies that for all $k\leq K$,

$$\begin{aligned}\mathbb{E}[f(\mathbf{q}_k)-f(\mathbf{q}^*)]+\frac{\mathbb{E}[\|\mathbf{p}_k\|_2^2]}{2} &\leq\big(1+4\eta^2L\big)^k\bigg(f(\mathbf{q}_0)-f(\mathbf{q}^*)+\frac{\|\mathbf{p}_0\|_2^2}{2}\bigg)+\eta^2\sum_{s=0}^{k-1}\big(1+4\eta^2L\big)^s\sigma^2\\ &\leq 3\bigg(f(\mathbf{q}_0)-f(\mathbf{q}^*)+\frac{\|\mathbf{p}_0\|_2^2}{2}\bigg)+3K\eta^2\sigma^2.\end{aligned}$$

Using the fact that $\|\nabla f(\mathbf{q}_k)\|_2^2 \leq L(f(\mathbf{q}_k) - f(\mathbf{q}^*))$, the above inequality further implies that

$$\mathbb{E}[f(\mathbf{q}_k)] \leq 3\bigg(f(\mathbf{q}_0) - f(\mathbf{q}^*) + \frac{\|\mathbf{p}_0\|_2^2}{2}\bigg) + 3K\eta^2\sigma^2$$
$$\mathbb{E}[\|\nabla f(\mathbf{q}_k)\|_2^2] \leq 3L\bigg(f(\mathbf{q}_0) - f(\mathbf{q}^*) + \frac{\|\mathbf{p}_0\|_2^2}{2}\bigg) + 3LK\eta^2\sigma^2$$
$$\mathbb{E}[\|\mathbf{p}_k\|_2^2] \leq 6\bigg(f(\mathbf{q}_0) - f(\mathbf{q}^*) + \frac{\|\mathbf{p}_0\|_2^2}{2}\bigg) + 6K\eta^2\sigma^2.$$

This completes the proof. □

F.5. Proof of Lemma D.5

Proof of Lemma D.5. Regarding $\mathcal{H}_\tau \mathbf{p}_k$, we have

$$\|\mathcal{H}_\tau \mathbf{p}_k\|_2^2 = \|\mathbf{p}_k\|_2^2 + 2f(\mathbf{q}_k) - 2f(\mathcal{H}_\tau \mathbf{q}_k) \leq \|\mathbf{p}_k\|_2^2 + 2f(\mathbf{q}_k) - 2f(\mathbf{q}^*).$$

Then for any $k \leq K$, we have

$$\begin{aligned}
\max_{\tau \leq \eta} \|\nabla f(\mathcal{H}_\tau \mathbf{q}_k)\|_2 &\leq \max_{\tau \leq \eta} \|\nabla f(\mathcal{H}_\tau \mathbf{q}_k) - \nabla f(\mathbf{q}_k)\|_2 + \|\nabla f(\mathbf{q}_k)\|_2 \\
&\leq L \max_{\tau \leq \eta} \|\mathcal{H}_\tau \mathbf{q}_k - \mathbf{q}_k\|_2 + \|\nabla f(\mathbf{q}_k)\|_2 \\
&\leq L\eta \max_{\tau \leq \eta} \|\mathcal{H}_\tau \mathbf{p}_k\|_2 + \|\nabla f(\mathbf{q}_k)\|_2 \\
&\leq L\eta \sqrt{\|\mathbf{p}_k\|_2^2 + 2f(\mathbf{q}_k) - 2f(\mathbf{q}^*)} + \|\nabla f(\mathbf{q}_k)\|_2.
\end{aligned}$$

Thus, note that $\eta \leq 1/\sqrt{L}$, we have

$$\max_{\tau \leq \eta} \|\nabla f(\mathcal{H}_\tau \mathbf{q}_k)\|_2^2 \leq 2L\big(\|\mathbf{p}_k\|_2^2 + 2f(\mathbf{q}_k) - 2f(\mathbf{q}^*)\big) + 2\|\nabla f(\mathbf{q}_k)\|_2^2.$$

This completes the proof. □

F.6. Proof of Lemma E.1

Proof of Lemma E.1. Note that $\mathbf{q}_0 = \mathbf{x}^{(t)}$ for some $t \geq 0$. Then based on (D.5) (which holds for any $t \geq 0$), we obtain

$$\begin{aligned}
\mathbb{E}[\|\mathbf{q}_0 - \mathbf{x}^*\|_2^2] &\leq 2\mathbb{E}[\|\mathbf{x}^{(t)} - \mathbf{x}^\pi\|_2^2] + 2\mathbb{E}[\|\mathbf{x}^\pi - \mathbf{x}^*\|_2^2] \\
&\leq 8D + \frac{20d}{\mu},
\end{aligned}$$

where the last inequality follows from Proposition 1 in (Durmus et al., 2019) that $\mathbb{E}[\|\mathbf{x}^\pi - \mathbf{x}^*\|_2^2] \leq d/\mu$ and $D = \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2$. Then based on the update rules of $\mathbf{q}_k$, by Young's inequality we have

$$\begin{aligned}
\mathbb{E}[\|\mathbf{q}_k - \mathbf{x}^*\|_2^2] &\leq 2\mathbb{E}[\|\mathbf{q}_k - \mathbf{q}_0\|_2^2] + 2\mathbb{E}[\|\mathbf{q}_0 - \mathbf{x}^*\|_2^2] \\
&= 2\eta^2 \mathbb{E}\bigg[\bigg\|\sum_{s=0}^{k-1} \mathbf{p}_s - \frac{\eta}{2}\mathbf{g}(\mathbf{q}_s, \boldsymbol{\xi}_s)\bigg\|_2^2\bigg] + 4D + \frac{12d}{\mu}.
\end{aligned}$$

Then regarding the first term on the R.H.S. of the above inequality, we further have

$$\begin{aligned}
\mathbb{E}\bigg[\bigg\|\sum_{s=0}^{k-1} \mathbf{p}_s - \frac{\eta}{2}\mathbf{g}(\mathbf{q}_s, \boldsymbol{\xi}_s)\bigg\|_2^2\bigg] &\leq 2k\bigg[\sum_{s=0}^{k-1} \|\mathbf{p}_s - \eta\mathbf{g}(\mathbf{q}_s, \boldsymbol{\xi}_s)\|_2^2/2\bigg] \\
&\leq 4k\eta^2\bigg[\sum_{s=0}^{k-1} \mathbb{E}[\|\mathbf{p}_s\|_2^2] + \eta^2\mathbb{E}[\|\mathbf{g}(\mathbf{q}_s, \boldsymbol{\xi}_s)\|_2^2]/4\bigg] \\
&\leq 4k\eta^2\bigg[kE_p + \sum_{s=0}^{k-1} \eta^2\mathbb{E}[\|\mathbf{g}(\mathbf{q}_s, \boldsymbol{\xi}_s)\|_2^2]/4\bigg].
\end{aligned}$$

Note that

$$\mathbb{E}[\|\mathbf{g}(\mathbf{q}_s, \boldsymbol{\xi}_s)\|_2^2] = \mathbb{E}[\|\mathbf{g}(\mathbf{q}_s, \boldsymbol{\xi}_s) - \nabla f(\mathbf{q}_s)\|_2^2] + \mathbb{E}[\|\nabla f(\mathbf{q}_s)\|_2^2] \leq \sigma^2 + E_q.$$

Then it follows that

$$\mathbb{E}[\|\mathbf{q}_k - \mathbf{x}^*\|_2^2] \leq 4k^2\eta^2\big[E_p + \eta^2(\sigma^2 + E_q)\big] + 8D + \frac{20d}{\mu}.$$

Note that $k\eta \leq K\eta \leq 1/(2\sqrt{L})$, we further have

$$\mathbb{E}[\|\mathbf{q}_k - \mathbf{x}^*\|_2^2] \leq L^{-1}\big[E_p + \eta^2(\sigma^2 + E_q)\big] + 8D + \frac{20d}{\mu},$$

which completes the proof. □