Stochastic Gradient Hamiltonian Monte Carlo with Variance Reduction for Bayesian Inference

Zhize Li† Tianyi Zhang† Shuyu Cheng Jun Zhu Jian Li
Tsinghua University
{zz-li14, tianyi-z16, chengsy18}@mails.tsinghua.edu.cn
{dcszj, lijian83}@mail.tsinghua.edu.cn

Abstract

Gradient-based Monte Carlo sampling algorithms, like Langevin dynamics and Hamiltonian Monte Carlo, are important methods for Bayesian inference. In large-scale settings, full-gradients are not affordable and thus stochastic gradients evaluated on mini-batches are used as a replacement. In order to reduce the high variance of noisy stochastic gradients, Dubey et al. [2016] applied the standard variance reduction technique on stochastic gradient Langevin dynamics and obtained both theoretical and experimental improvements. In this paper, we apply the variance reduction tricks on Hamiltonian Monte Carlo and achieve better theoretical convergence results compared with the variance-reduced Langevin dynamics. Moreover, we apply the symmetric splitting scheme in our variance-reduced Hamiltonian Monte Carlo algorithms to further improve the theoretical results. The experimental results are also consistent with the theoretical results. As our experiment shows, variance-reduced Hamiltonian Monte Carlo demonstrates better performance than variance-reduced Langevin dynamics in Bayesian regression and classification tasks on real-world datasets.

1 Introduction

Gradient-based Monte Carlo algorithms are useful tools for sampling posterior distributions. Similar to gradient descent algorithms, gradient-based Monte Carlo generates posterior samples iteratively using the gradient of log-likelihood.

Langevin dynamics (LD) and Hamiltonian Monte Carlo (HMC) [Duane et al., 1987, Neal et al., 2011] are two important examples of gradient-based Monte Carlo sampling algorithms that are widely used in Bayesian inference. Since calculating likelihood on large datasets is expensive, people use stochastic gradients [Robbins and Monro, 1951] in place of full gradient, and have, for both Langevin dynamics and Hamiltonian Monte Carlo, developed their stochastic gradient counterparts [Welling and Teh, 2011, Chen et al., 2014]. Stochastic gradient Hamiltonian Monte Carlo (SGHMC) usually converges faster than stochastic gradient Langevin dynamics (SGLD) in practical machine learning tasks like covariance estimation of bivariate Gaussian and Bayesian neural networks for classification on MNIST dataset, as demonstrated in [Welling and Teh, 2011]. Similar phenomenon was also observed in [Chen et al., 2015] where SGHMC and SGLD were compared on both synthetic and real-world datasets. Intuitively speaking, comparing against SGLD, SGHMC has a momentum term that may enable it to explore the parameter space of posterior distribution much faster when the gradient of log-likelihood becomes smaller.

Very recently, Dubey et al. [2016] borrowed the standard variance reduction techniques from the stochastic optimization literature [Johnson and Zhang, 2013, Defazio et al., 2014] and applied them on SGLD to obtain two variance-reduced SGLD algorithms (called SAGA-LD and SVRG-LD) with improved theoretical results and practical performance. Because of the superiority of SGHMC over SGLD in terms of convergence rate in a wide range of machine learning tasks, it would be a natural question whether such variance reduction techniques can be applied on SGHMC to achieve better results than variance-reduced SGLD.

†denotes equal contribution

The challenge is that SGHMC is more complicated than SGLD, i.e., the extra momentum term (try to explore faster) and friction term (control the noise caused by SGHMC from HMC) in SGHMC. Note that the friction term in SGHMC is inherently different than SGLD since LD itself already has noise so it can be directly extended to SGLD, while HMC itself is deterministic. To the best of our knowledge, there is even no existing work to prove that SGHMC is better than SGLD. So in this paper we need to give some new approaches and insights in our analysis to prove that variance-reduced SGHMC is better than variance-reduced SGLD due to the existence of momentum term and friction term. Note that in stochastic optimization literature, the variance-reduced methods with momentum term (e.g., [Allen-Zhu, 2017, Lan et al., 2019]) indeed are better than variance-reduced methods without momentum term (e.g., [Johnson and Zhang, 2013]) especially for convex optimization.

Actually, it seems that the variance reduction in this stochastic Bayesian inference is more effective compared with stochastic optimization settings. Intuitively, the full gradient case (no variance) may converge to a saddle point or a local minimum (not a global minimum) in nonconvex optimization, and the variance of the stochastic gradient estimator may be useful for escaping saddle points or bad local minima. Thus, we may not want to reduce the variance. However, the full gradient case (no variance) will converge to the stationary posterior distribution for Bayesian inference. Thus, it is useful to reduce the variance of the stochastic gradient estimator for obtaining more approximate posterior distribution. Note that in large-scale settings, full-gradients (no variance) are not affordable and thus stochastic gradients evaluated on mini-batches are used as a replacement.

1.1 Our contribution

1. We propose two variance-reduced versions of Hamiltonian Monte Carlo algorithms (called SVRG-HMC and SAGA-HMC) using the standard approaches from [Johnson and Zhang, 2013, Defazio et al., 2014]. Compared with SVRG/SAGA-LD [Dubey et al., 2016], our algorithms guarantee improved theoretical convergence results due to the extra momentum term in HMC (see Corollary 3).

2. Moreover, we combine the proposed SVRG/SAGA-HMC algorithms with the symmetric splitting scheme [Chen et al., 2015, Leimkuhler and Shang, 2016] to extend them to 2nd-order integrators, which further improve the dependency on step size (see the difference between Theorem 2 and 5). We denote these two algorithms as SVRG2nd-HMC and SAGA2nd-HMC.

3. Finally, we evaluate our algorithms on real-world datasets and compare them with SVRG/SAGA-LD [Dubey et al., 2016]; as it turns out, our algorithms converge markedly faster than the benchmarks (vanilla SGHMC and SVRG/SAGA-LD).

1.2 Related work

Langevin dynamics and Hamiltonian Monte Carlo are two important sampling algorithms that are widely used in Bayesian inference. Many literatures studied how to develop the variants of them to achieve improved performance, especially for scalability for large datasets. Welling and Teh [2011] started this direction with the notable work stochastic gradient Langevin dynamics (SGLD). Ahn et al. [2012] proposed a modification to SGLD reminiscent of Fisher scoring to better estimate the gradient noise variance, with lower classification error rates on HHP dataset and MNIST dataset. Chen et al. [2014] developed the stochastic gradient version of HMC (SGHMC), with a quite nontrivial approach different from SGLD. Ding et al. [2014] further improved SGHMC by a new dynamics to better control the gradient noise, and the proposed stochastic gradient Nosé-Hoover thermostats (SGNHT) outperforms SGHMC on MNIST dataset.

Various settings of Markov Chain Monte Carlo (MCMC) are also considered. Girolami and Calderhead [2011] enhanced LD and HMC by exploring the Riemannian structure of the target distribution, with Riemannian manifold LD and HMC (RMLD and RMHMC, respectively). Byrne and Girolami [2013] developed geodesic Monte Carlo (GMC) that is applicable to Riemannian manifolds with no global coordinate systems. Large-scale variants of RMLD, RMHMC and GMC with stochastic gradient were developed by [Patterson and Teh, 2013], [Ma et al., 2015] and [Liu et al., 2016], respectively. Ahn et al. [2014] studied the behaviour of stochastic gradient MCMC algorithms for distributed posterior inference. Very recently, Zou et al. [2018] used a stochastic variance-reduced HMC for sampling

from smooth and strongly log-concave distributions which requires f is smooth and strongly convex. In this paper, we do not assume f is strongly convex or convex and we also use an efficient discretization scheme to further imporve the convergence results. Their results were measured with 2-Wasserstein distance, while ours are measured with mean square error. Note that the variance reduction techniques have already been used in nonconvex optimization literature (see e.g., [Allen-Zhu and Hazan, 2016, Reddi et al., 2016, Li and Li, 2018, Ge et al., 2019, Li, 2019]), and they achieved improved convergence results.

2 Preliminary

Let $X = \{x_i\}_{i=1}^n$ be a d-dimensional dataset that follows the distribution $\Pr(X|\theta) = \prod_{i=1}^n \Pr(x_i|\theta)$. Then, we are interested in sampling the posterior distribution $\Pr(\theta|X) \propto \Pr(\theta)\prod_{i=1}^n \Pr(x_i|\theta)$ based on Hamiltonian Monte Carlo algorithms. Let $[n]$ denote the set $\{1, 2, \ldots, n\}$. Define $f(\theta) = \sum_{i=1}^n f_i(\theta) - \log \Pr(\theta)$, where $f_i(\theta) = -\log \Pr(x_i|\theta)$ and $i \in [n]$. Similar to [Dubey et al., 2016], we assume that each f_i is L-smooth and G-Lipschitz, for all $i \in [n]$.

The general algorithmic framework maintains two sequences for $t = 0, 1, \ldots, T-1$ by the following discrete time procedure:

$$p_{t+1} = (1 - Dh)p_t - h\tilde{\nabla}_t + \sqrt{2Dh} \cdot \xi_t \tag{1}$$

$$\theta_{t+1} = \theta_t + hp_{t+1} \tag{2}$$

and then returns the samples $\{\theta_1, \theta_2, \ldots, \theta_T\}$ as an approximation to the stationary distribution $\Pr(\theta|X)$. θ_t is the parameter we wish to sample and p_t is an auxiliary variable conventionally called the "momentum". Here h is step size, D is a constant independent of θ and p, $\xi_t \sim \mathrm{N}(0, I_d)$ and $\tilde{\nabla}_t$ is a mini-batch approximation of the full gradient $\nabla f(\theta_t)$. If we set $\tilde{\nabla}_t = \frac{n}{b}\sum_{i\in I} \nabla f_i(\theta_t)$, I being a b-element index set uniformly randomly drawn (with replacement) from $\{1, 2, \ldots, n\}$ as introduced in [Robbins and Monro, 1951], then the algorithm becomes SGHMC.

The above discrete time procedure provides an approximation to the continuous Hamiltonian Monte Carlo diffusion process (θ, p):

$$\mathrm{d}\theta = p\mathrm{d}t \tag{3}$$

$$\mathrm{d}p = -\nabla_\theta f(\theta)\mathrm{d}t - Dp\mathrm{d}t + \sqrt{2D}\mathrm{d}W \tag{4}$$

Here W is a Wiener process. According to [Chen et al., 2015], the stationary joint distribution of (θ, p) is $\pi(\theta, p) \propto e^{-f(\theta) - \frac{p^\intercal p}{2}}$.

How do we evaluate the quality of the samples $\{\theta_1, \theta_2, \ldots, \theta_T\}$? Assuming $\phi : \mathbb{R}^d \to \mathbb{R}$ is a smooth test function, we wish to upper bound the Mean-Squared Error (MSE) $\mathbb{E}(\hat{\phi} - \bar{\phi})^2$, where $\hat{\phi} = \frac{1}{T}\sum_{t=1}^T \phi(\theta_t)$ is the empirical average, and $\bar{\phi} = \mathbb{E}_{\theta\sim\Pr(\theta|X)}\phi(\theta)$ is the population average. So, the objective of our algorithm is to carefully design $\tilde{\nabla}_t$ to minimize $\mathbb{E}(\hat{\phi} - \bar{\phi})^2$ in a faster way, where $\tilde{\nabla}_t$ is a stochastic approximation of $\nabla f(\theta_t)$.

To study how the choice of $\tilde{\nabla}_t$ influences the value of $\mathbb{E}(\hat{\phi} - \bar{\phi})^2$, define $\psi(\theta, p)$ to be the solution to the Poisson equation $\mathcal{L}\psi = \phi(\theta) - \bar{\phi}$, $\mathcal{L}$ being the generator of Hamiltonian Monte Carlo diffusion process. In order to analyze the theoretical convergence results related to the MSE $\mathbb{E}(\hat{\phi} - \bar{\phi})^2$, we inherit the following assumption from [Chen et al., 2015].

Assumption 1 ([Chen et al., 2015]) *Function ψ is bounded up to 3rd-order derivatives by some real-valued function $\Gamma(\theta, p)$, i.e. $\|\mathcal{D}^k\psi\| \le C_k\Gamma^{q_k}$ where $\mathcal{D}^k$ is the kth order derivative for $k = 0, 1, 2, 3$, and $C_k, q_k > 0$. Furthermore, the expectation of Γ on $\{(\theta_t, p_t)\}$ is bounded, i.e. $\sup_t \mathbb{E}[\Gamma^q(\theta_t, p_t)] < \infty$ and that Γ is smooth such that $\sup_{s\in(0,1)} \Gamma^q(s\theta + (1-s)\theta', sp + (1-s)p') \le C(\Gamma^q(\theta, p) + \Gamma^q(\theta', p'))$, $\forall\theta, p, \theta', p', q \le \max 2q_k$ for some constant $C > 0$.*

Define operator $\Delta V_t = (\tilde{\nabla}_t - \nabla f(\theta_t)) \cdot \nabla$ for all $t = 0, 1, 2, \ldots, T-1$. When the above assumption holds, we have the following theorem by [Chen et al., 2015].

For the rest of this paper, for any two values $A, B > 0$, we say $A \lesssim B$ if $A = O(B)$, where the notation $O(\cdot)$ only hides a constant factor independent of algorithm parameters T, n, D, h, G, b.

Theorem 1 ([Chen et al., 2015]) *Let $\tilde{\nabla}_t$ be an unbiased estimate of $\nabla f(\theta_t)$ for all t. Then under assumption 1, for a smooth test function ϕ, the MSE of SGHMC is bounded in the following way:*

$$\mathbb{E}(\hat{\phi} - \bar{\phi})^2 \lesssim \frac{\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}(\Delta V_t \psi(\theta_t, p_t))^2}{T} + \frac{1}{Th} + h^2 \tag{5}$$

Similar to the [A2] assumption in [Dubey et al., 2016], we also need to make the following assumption which relates $\Delta V_t \phi(\theta, p)$ to the difference $\|\tilde{\nabla}_t - \nabla f(\theta)\|^2$.

Assumption 2 $(\Delta V_t \psi(\theta_t, p_t))^2 \lesssim \|\tilde{\nabla}_t - \nabla f(\theta_t)\|^2$ *for all* $0 \leq t < T$.

Combined with Theorem 1, Assumption 2 immediately yields the following corollary.

Corollary 1 *Under Assumptions 1 and 2, we have:*

$$\mathbb{E}(\hat{\phi} - \bar{\phi})^2 \lesssim \frac{\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\tilde{\nabla}_t - \nabla f(\theta_t)\|^2}{T} + \frac{1}{Th} + h^2$$

As we mentioned before, if we take $\tilde{\nabla}_t$ to be the Robbins & Monro approximation of $\nabla f(\theta_t)$ [Robbins and Monro, 1951], then it becomes SGHMC and the following corollary holds since all f_i's are G-Lipschitz and $\mathbb{E}(X - \mathbb{E}X)^2 \leq \mathbb{E}X^2$ for any random variable X.

Corollary 2 *Under Assumptions 1 and 2, the MSE of SGHMC is bounded as:*

$$\begin{aligned}\mathbb{E}(\hat{\phi} - \bar{\phi})^2 &\lesssim \frac{\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\tilde{\nabla}_t - \nabla f(\theta_t)\|^2}{T} + \frac{1}{Th} + h^2 \\ &\leq \frac{n^2 G^2}{bT} + \frac{1}{Th} + h^2\end{aligned} \tag{6}$$

3 Variance Reduction for Hamiltonian Monte Carlo

In this section, we introduce two versions of variance-reduced Hamiltonian Monte Carlo based on SVRG [Johnson and Zhang, 2013] and SAGA [Defazio et al., 2014] respectively.

3.1 SVRG-HMC

In this subsection, we propose the SVRG-HMC algorithm (see Algorithm 1) which is based on the SVRG algorithm. As can be seen from Line 8 of Algorithm 1, we use $\tilde{\nabla}_{tK+k} = -\nabla \log \Pr(\theta_{tK+k}) + \frac{n}{b}\sum_{i \in I}\big(\nabla f_i(\theta_{tK+k}) - \nabla f_i(w)\big) + g$, where g is $\sum_{i=1}^{n} \nabla f_i(\theta_{tK})$, as the stochastic estimation for the full gradient $\nabla f(\theta_{tK+k})$.

Note that we initialize θ_0, p_0 to be zero vectors in the algorithm only to simplify the theoretical analysis. It would still work with an arbitrary initialization.

The following theorem shows the convergence result for MSE of SVRG-HMC (Algorithm 1). We defer all the proofs to Appendix B.

Theorem 2 *Under Assumptions 1 and 2, the MSE of SVRG-HMC is bounded as:*

$$\mathbb{E}[(\hat{\phi} - \bar{\phi})^2] \lesssim \min\left\{\frac{n^2 G^2}{bT}, \frac{L^2 n^2 K^2 h^2}{bT}\Big(\frac{\sqrt{n^2 G^2 + D^2 d}}{D - L^2 n^2 K^2 h^3 b^{-1}}\Big)^2\right\} + \frac{1}{Th} + h^2 \tag{7}$$

To see how the SVRG-HMC (Algorithm 1) is compared with SVRG-LD [Dubey et al., 2016], we restate their results as following.

Theorem 3 ([Dubey et al., 2016]) *Under Assumptions 1 and 2, the MSE of SVRG-LD is bounded as:*

$$\mathbb{E}[(\hat{\phi} - \bar{\phi})^2] \lesssim \frac{\min\{n^2 G^2, n^2 K^2 (n^2 L^2 h^2 G^2 + hd)\}}{bT} + \frac{1}{Th} + h^2 \tag{8}$$

Algorithm 1: SVRG-HMC

parameters $T, K, b, h > 0$, $Dh < 1$, $D \geq 1$;
initialize $\theta_0 = p_0 = 0$;
for $t = 0, 1, \ldots, T/K - 1$ **do**
  compute $g = \sum_{i=1}^{n} \nabla f_i(\theta_{tK})$;
  $w = \theta_{tK}$;
  for $k = 0, 1, \ldots, K - 1$ **do**
    uniformly sample an index subset $I \subseteq [n]$, $|I| = b$;
    $\tilde{\nabla}_{tK+k} = -\nabla \log \Pr(\theta_{tK+k}) + \frac{n}{b} \sum_{i \in I} \Big(\nabla f_i(\theta_{tK+k}) - \nabla f_i(w) \Big) + g$;
    $p_{tK+k+1} = (1 - Dh) p_{tK+k} - h \tilde{\nabla}_{tK+k} + \sqrt{2Dh} \xi_{tK+k}$;
    $\theta_{tK+k+1} = \theta_{tK+k} + h p_{tK+k+1}$;
return $\{\theta_t\}_{1 \leq t \leq T}$;

We assume $n^2 G^2 > n^2 K^2 (n^2 L^2 h^2 G^2 + hd)$. Otherwise the MSE upper bound of SVRG-LD would be equal to SGHMC (see (6) and (8)). We then omit the same terms (i.e., second and third terms) in the RHS of (7) and (8). Then we have the following lemma.

Lemma 1 *Let* $\mathsf{R}^{\mathsf{HMC}} = \frac{L^2 n^2 K^2 h^2}{bT} \big(\frac{\sqrt{n^2 G^2 + D^2 d}}{D - L^2 n^2 K^2 h^3 b^{-1}} \big)^2$ *(i.e., the first term in the RHS of (7)) and* $\mathsf{R}^{\mathsf{LD}} = \frac{n^2 K^2 (n^2 L^2 h^2 G^2 + hd)}{bT}$ *(i.e., the first term in the RHS of (8)). Then the following inequality holds.*

$$\mathsf{R}^{\mathsf{HMC}} \leq \max \Big\{ \frac{1}{D^2}, \frac{L}{nK} \Big\} \mathsf{R}^{\mathsf{LD}} \tag{9}$$

In particular, if $D \geq 1/L\sqrt{h}$*, then (9) becomes:*

$$\mathsf{R}^{\mathsf{HMC}} \leq \frac{L}{nK} \mathsf{R}^{\mathsf{LD}} \tag{10}$$

Note that K is suggested to be $2n$ by [Johnson and Zhang, 2013] or n/b by [Dubey et al., 2016]. we obtain the following corollary from Lemma 1.

Corollary 3 *If* K *is* n/b *as suggested by [Dubey et al., 2016], then (10) becomes:*

$$\mathsf{R}^{\mathsf{HMC}} \leq \frac{bL}{n^2} \mathsf{R}^{\mathsf{LD}} \tag{11}$$

In other words, the SVRG-HMC is $O(\frac{n^2}{bL})$ *times faster than SVRG-LD, in terms of the convergence bound related to the variance reduction (i.e., the first terms in RHS of (7) and (8)). Note that* n *is the size of dataset which can be very large,* b *is the mini-batch size which is usually a small constant and* L *is the Lipschitz smooth parameter for* $f_i(\theta)$*.*

We also want to mention that the convergence proof for SVRG-HMC (i.e., Theorem 2) is a bit more difficult than that for SVRG-LD Dubey et al. [2016] due to the momentum variable p (see Line 9 of Algorithm 1). Concretely, the main part of the proof in both SVRG-LD and SVRG-HMC is to bound the variance. Moreover, the variance can be bounded by the adjacent distance $\{\|\theta_t - \theta_{t-1}\|^2\}$ in both SVRG-LD and SVRG-HMC. In SVRG-LD [Dubey et al., 2016], they can directly bound each of $\|\theta_t - \theta_{t-1}\|^2$ for $t \in [T]$. However, due to the momentum variable p in our SVRG-HMC (see Line 9 of Algorithm 1), the distances $\{\|\theta_t - \theta_{t-1}\|^2\}_{t \in [T]}$ are more correlated. Thus, we cannot directly bound each of $\|\theta_t - \theta_{t-1}\|^2$ independently like SVRG-LD. We bound the variance as a whole, i.e., we bound the summation of the variance which is equivalent to bound the summation $\sum_{t \in [T]} \|\theta_t - \theta_{t-1}\|^2$. Then we get a quadratic inequality due to the correlation among $\{\|\theta_t - \theta_{t-1}\|^2\}_{t \in [T]}$. Finally, we solve this quadratic inequality to bound the variance.

3.2 SAGA-HMC

In this subsection, we propose the SAGA-HMC algorithm by applying the SAGA framework [Defazio et al., 2014] to the Hamiltonian Monte Carlo. The details are described in Algorithm 2. Similar to the SVRG-HMC, we initialize θ_0, p_0 to be zero vectors in the algorithm only to simplify the analysis; it would still work with an arbitrary initialization.

Algorithm 2: SAGA-HMC

parameters $T, b, h > 0$, $Dh < 1$, $D \geq 1$;
initialize $\theta_0 = 0, p_0 = 0$;
initialize an array $\alpha_0^i = \theta_0, \forall i \in [n]$;
compute $g = \sum_{i=1}^n \nabla f_i(\alpha_0^i)$;
for $t = 0, 1, \ldots, T-1$ **do**
  uniformly randomly pick a set $I \subseteq [n]$ such that $|I| = b$;
  $\tilde{\nabla}_t = -\nabla \log \Pr(\theta_t) + \frac{n}{b} \sum_{i \in I} (\nabla f_i(\theta_t) - \nabla f_i(\alpha_t^i)) + g$;
  $p_{t+1} = (1 - Dh) p_t - h\tilde{\nabla}_t + \sqrt{2Dh}\xi_t$;
  $\theta_{t+1} = \theta_t + h p_{t+1}$;
  update $\alpha_{t+1}^i = \theta_t, \forall i \in I$;
  $g \leftarrow g + \sum_{i \in I} (\nabla f_i(\alpha_{t+1}^i) - \nabla f_i(\alpha_t^i))$;
return $\{\theta_t\}_{t=1}^T$;

The following theorem shows the convergence result for MSE of SAGA-HMC. The proof is deferred to Appendix B.

Theorem 4 *Under Assumptions 1 and 2, the MSE of SAGA-HMC is bounded as:*

$$\mathbb{E}[(\hat{\phi} - \bar{\phi})^2] \lesssim \min\left\{\frac{n^2G^2}{bT}, \frac{L^2n^4h^2}{Tb^3}\left(\frac{\sqrt{n^2G^2 + D^2d}}{D - L^2n^4h^3b^{-3}}\right)^2\right\} + \frac{1}{Th} + h^2$$

Note that SAGA-HMC can be compared with SAGA-LD [Dubey et al., 2016] in a very similar manner to SVRG-HMC (i.e., Lemma 1 and Corollary 3). Thus we omit such a repetition.

4 Variance-reduced SGHMC with Symmetric Splitting

Symmetric splitting is a numerically efficient method introduced in [Leimkuhler and Shang, 2016] to accelerate the gradient-based algorithms. We note that one additional advantage of SGHMC over SGLD is that SGHMC can be combined with symmetric splitting while SGLD cannot [Chen et al., 2015]. So it is quite natural to combine symmetric splitting with the proposed SVRG-HMC and SAGA-HMC respectively to see if any further improvements can be obtained.

The symmetric splitting scheme breaks the original recursion into 5 steps:

$$\theta_t^{(1)} = \theta_t + \frac{h}{2} p_t \tag{12}$$

$$p_t^{(1)} = e^{-Dh/2} p_t \tag{13}$$

$$p_t^{(2)} = p_t^{(1)} - h\tilde{\nabla}_t + \sqrt{2Dh}\xi_t \tag{14}$$

$$p_{t+1} = e^{-Dh/2} p_t^{(2)} \tag{15}$$

$$\theta_{t+1} = \theta_t^{(1)} + \frac{h}{2} p_{t+1} \tag{16}$$

If we eliminate the intermediate variables, then

$$p_{t+1} = e^{-Dh/2}\big(e^{-Dh/2}p_t - h\tilde{\nabla}_t + \sqrt{2Dh}\xi_t\big) \tag{17}$$

$$\theta_{t+1} = \theta_t + \frac{h}{2}p_{t+1} + \frac{h}{2}p_t \tag{18}$$

Same as before $\xi_t \sim \mathrm{N}(0, I_d)$. Note that the stochastic gradient $\tilde{\nabla}_t$ is computed at $\theta_t^{(1)}$ (which is $\theta_t + \frac{h}{2}p_t$) instead of θ_t (see (1), (12) and (14)). As shown in [Chen et al., 2015], this symmetric splitting scheme is a 2nd-order local integrator. Then it improves the dependency of MSE on step size h, i.e., the third term in the RHS of (5) changes to be h^4, which is a higher order term than the original h^2. It means that we can allow larger step size h by using this symmetric splitting scheme (note that $h < 1$).

Similarly, we can further improve the convergence results for SVRG/SAGA-HMC by combining the symmetric splitting scheme. We give the details of the algorithms and theoretical results for SVRG2nd-HMC and SAGA2nd-HMC in the following subsections.

4.1 SVRG2nd-HMC

In this subsection, we propose the SVRG2nd-HMC algorithm (see Algorithm 3) by combining our SVRG-HMC (Algorithm 1) with the symmetric splitting scheme.

Algorithm 3: SVRG2nd-HMC

parameters $T, K, b, h > 0$, $Dh < 1$, $D \geq 1$;
initialize $\theta_0 = p_0 = 0$;
for $t = 0, 1, \ldots, T/K - 1$ **do**
  compute $g = \sum_{i=1}^n \nabla f_i(\theta_{tK} + \frac{h}{2}p_{tK})$;
  $w = \theta_{tK} + \frac{h}{2}p_{tK}$;
  for $k = 0, 1, \ldots, K - 1$ **do**
    uniformly sample an index subset $I \subseteq [n]$, $|I| = b$;
    $\tilde{\nabla}_{tK+k} = -\nabla \log \Pr(\theta_{tK+k} + \frac{h}{2}p_{tK+k}) + \frac{n}{b}\sum_{i\in I}(\nabla f_i(\theta_{tK+k} + \frac{h}{2}p_{tK+k}) - \nabla f_i(w)) + g$;
    $p_{tK+k+1} = e^{-Dh/2}(e^{-Dh/2}p_t - h\tilde{\nabla}_{tK+k} + \sqrt{2Dh}\xi_{tK+k})$;
    $\theta_{tK+k+1} = \theta_{tK+k} + \frac{h}{2}p_{tK+k+1} + \frac{h}{2}p_{tK+k}$;
return $\{\theta_t\}_{t=1}^T$;

The convergence result for SVRG2nd-HMC is provided in Theorem 5. It shows that the dependency of MSE on step size h can be improved from h^2 to h^4 (see (7) and (19)).

Theorem 5 *Under Assumptions 1 and 2, the MSE of SVRG2nd-HMC is bounded as:*

$$\mathbb{E}(\hat{\phi} - \bar{\phi})^2 \lesssim \min\Big\{\frac{n^2G^2}{bT}, \frac{L^2n^2K^2h^2}{bT}\Big(\frac{\sqrt{n^2G^2 + D^2d}}{D - L^2n^2K^2h^3b^{-1}}\Big)^2\Big\} + \frac{1}{Th} + h^4 \tag{19}$$

4.2 SAGA2nd-HMC

In this subsection, we propose the SAGA2nd-HMC algorithm (see Algorithm 4) by combining our SAGA-HMC (Algorithm 2) with the symmetric splitting scheme. The convergence result and algorithm details are described below.

Theorem 6 *Under Assumptions 1 and 2, the MSE of SAGA2nd-HMC is bounded as:*

$$\mathbb{E}(\hat{\phi} - \bar{\phi})^2 \lesssim \min\Big\{\frac{n^2G^2}{bT}, \frac{L^2n^4h^2}{Tb^3}\Big(\frac{\sqrt{n^2G^2 + D^2d}}{D - L^2n^4h^3b^{-3}}\Big)^2\Big\} + \frac{1}{Th} + h^4$$

Algorithm 4: SAGA2nd-HMC

parameters $T, b, h > 0$, $Dh < 1$, $D \geq 1$;
initialize $\theta_0 = 0, p_0 = 0$;
initialize an array $\alpha_0^i = 0, \forall i \in [n]$;
compute $g = \sum_{i=1}^n \nabla f_i(\alpha_0^i)$;
for $t = 0, 1, \ldots, T-1$ **do**
  uniformly randomly pick a set $I \subseteq [n]$ such that $|I| = b$;
  $\tilde{\nabla}_t = -\nabla \log \Pr(\theta_t + \frac{h}{2} p_t) + \frac{n}{b} \sum_{i \in I} (\nabla f_i(\theta_t + \frac{h}{2} p_t) - \nabla f_i(\alpha_t^i)) + g$;
  $p_{t+1} = e^{-Dh/2}(e^{-Dh/2} p_t - h\tilde{\nabla}_t + \sqrt{2Dh}\xi_t)$;
  $\theta_{t+1} = \theta_t + \frac{h}{2} p_{t+1} + \frac{h}{2} p_t$;
  update $\alpha_{t+1}^i = \theta_t + \frac{h}{2} p_t, \forall i \in I$;
  $g \leftarrow g + \sum_{i \in I} (\nabla f_i(\alpha_{t+1}^i) - \nabla f_i(\alpha_t^i))$;
return $\{\theta_t\}_{t=1}^T$;

5 Experiment

We present experimental results in this section. We compare the proposed SVRG-HMC (Algorithm 1), as well as its symmetric splitting variant SVRG2nd-HMC (Algorithm 3), against SVRG-LD [Dubey et al., 2016] on Bayesian regression, Bayesian classification and Bayesian Neural Networks. The experimental results of SAGA variants (Algorithm 2 and 4) are almost same as the SVRG variants. We report the corresponding SAGA experiments in Appendix A. In accordance with the theoretical analysis, all algorithms have fixed step size h, and all HMC-based algorithms have fixed friction parameter D; a grid search is performed to select the best step size and friction parameter for each algorithm. The minibatch size b is chosen to be 10 (same as SVRG/SAGA-LD [Dubey et al., 2016]) for all algorithms, and K is set to be n/b.

The experiments are tested on the real-world UCI datasets[1]. The information of the standard datasets used in our experiments are described in the following Table 1 and Table 2 (Section 5.3). For each dataset (regression or classification), we partition the dataset into training (70%), validation (10%) and test (20%) sets. The validation set is used to select step size as well as friction for HMC-based algorithms in an 8-fold manner.

Table 1: Summary of standard UCI datasets for Bayesian regression and classification

datasets	concrete	noise	parkinson	bike	pima	diabetic	eeg
size	1030	1503	5875	17379	768	1151	14980
features	8	5	21	12	8	20	15

The Bayesian regression experiments were conducted on the first four UCI regression datasets, the Bayesian classification experiments were conducted on the last three UCI classification datasets, and the more complicated Bayesian Neural Networks experiments were conducted on larger UCI datasets in Table 2.

5.1 Bayesian Regression

In this subsection we study the performance of those aforementioned algorithms on Bayesian linear regression. Say we are provided with inputs $Z = \{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. The distribution of y_i given x_i is modelled as $\Pr(y_i|x_i) = N(\beta^\intercal x_i, \sigma^2)$, where the unknown parameter β follows a prior distribution of $\mathrm{N}(0, I_d)$. The gradients of log-likelihood can thus be calculated as $\nabla_\beta \log \Pr(y_i|x_i, \beta) = (y_i - \beta^\intercal x_i)x_i$ and $\nabla_\beta \log \Pr(\beta) = -\beta$. The average test Mean-Squared Error (MSE) is reported in Figure 1.

[1]The UCI datasets can be downloaded from https://archive.ics.uci.edu/ml/datasets.html

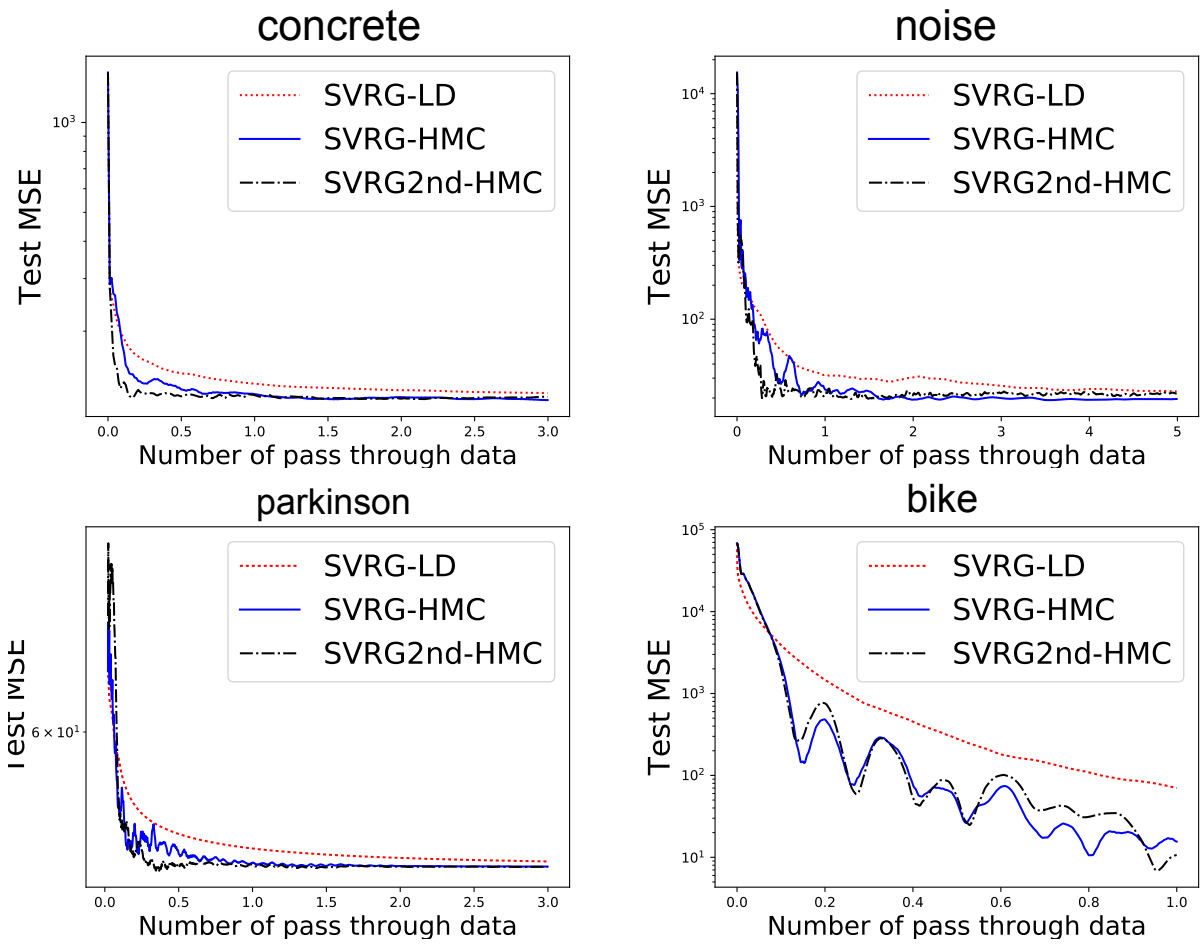

Figure 1: Performance comparison of SVRG variants on Bayesian regression tasks. The x-axis and y-axis represent number of passes through the entire training dataset and average test MSE respectively. For the bike dataset, we have omitted the first 10 MSE values from the diagram because otherwise the diagram would scale badly as MSE values are very large in the first several iterations.

As can be observed from Figure 1, SVRG-HMC as well its symmetric splitting counterpart SVRG2nd-HMC, converge markedly faster than SVRG-LD in the first pass through the whole dataset. The performance SVRG2nd-HMC is usually similar (no worse) to SVRG-HMC, and it turns out that a slightly larger step size can be chosen for SVRG2nd-HMC, which is also consistent with our theoretical results (i.e., allow larger step size).

5.2 Bayesian Classification

In this subsection we study classification tasks using Bayesian logistic classification. Suppose input data $Z = \{(x_i, y_i)\}$ where $x_i \in \mathbb{R}^d$, $y_i \in \{0, 1\}$. The distribution of the output y_i is modelled as $\Pr(y_i = 1) = 1/(1 + \exp(-\beta^\intercal x_i))$, where the model parameter follows a prior distribution of $\mathrm{N}(0, I_d)$. Then the gradient of log-likelihood and log-prior can be written as $\nabla_\beta \log \Pr(y_i|x_i, \beta) = \big(y_i - 1/(1 + \exp(-\beta^\intercal x_i))\big)x_i$ and $\nabla_\beta \log \Pr(\beta) = -\beta$. The average test log-likelihood is reported in Figure 2.

Similar to the Bayesian regression, SVRG-HMC as well its symmetric splitting counterpart SVRG2nd-HMC, converge markedly faster than SVRG-LD for the Bayesian classification tasks. Also, the experimental results suggest

that SVRG2nd-HMC converges more quickly than SVRG-HMC, which is consistent with Theorem 2 and 5.

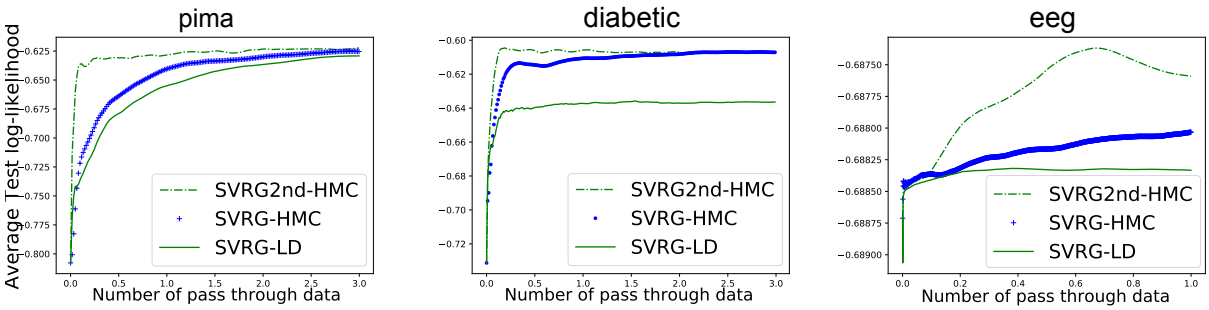

Figure 2: Performance comparison of SVRG variants on Bayesian classification tasks. The x-axis and y-axis represent number of passes through the entire training dataset and average test log-likelihood respectively.

In sum, for our four algorithms, we recommend SVRG2nd/SAGA2nd-HMC due to the better theoretical results (Theorem 2, 4, 5 and 6) and practical experimental results (Figure 1–6) compared with SVRG/SAGA-HMC. Further, we recommend SVRG2nd-HMC since SAGA2nd-HMC needs high memory cost and its implementation is a little bit complicated than SVRG2nd-HMC.

5.3 Bayesian Neural Networks

To show the scalability of variance reduced HMC to larger datasets and its application to nonconvex problems and more complicated models, we study Bayesian neural networks tasks. In our experiments, the model is a neural network with one hidden layer which has 50 hidden units (100 hidden units for 'susy' dataset) with ReLU activation, which is denoted by f_{NN}. Its unknown parameter β follows a prior distribution of $\mathrm{N}(0, \sigma_p^2 I_d)$. Let $t_i = f_{NN}(x_i, \beta)$ denotes output of the neural network with parameter value β and input x_i. The experiments are tested on larger UCI regression and classification datasets described in Table 2. Suppose we are provided with inputs $Z = \{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$. In regression tasks, $y_i \in \mathbb{R}$, $t_i \in \mathbb{R}$, and the distribution of y_i given x_i is modelled as $\Pr(y_i|x_i) = N(t_i, \sigma_l^2)$. In binary classification tasks, $y_i \in \{0, 1\}$, $t_i \in \mathbb{R}$, and $\Pr(y_i = 1) = 1/(1 + \exp(-t_i))$. In K-class classification tasks ($K \geq 3$), $y_i \in \{1, 2, ..., K\}$, $t_i \in \mathbb{R}^K$, and $\Pr(y_i = n) = \exp(t_{in}) / \sum_{m=1}^K \exp(t_{im})$. The code for experiments is implemented in TensorFlow. We conduct experiments for vanilla SGHMC and SVRG variants of LD and HMC algorithms. The test Root-Mean-Square Error (RMSE) for regression tasks is reported in Figure 3, and the average test log-likelihood for classification tasks is reported in Figure 4.

Table 2: Summary of larger UCI datasets for Bayesian neural networks experiments

datasets	protein	music	letter	susy
size	45730	515345	20000	5000000
features	9	90	16	18

The Bayesian neural network regression experiments were conducted on the first two UCI regression datasets and the classification experiments were conducted on the last two UCI classification datasets. The 'letter' dataset is 26-class and the 'susy' dataset is binary class.

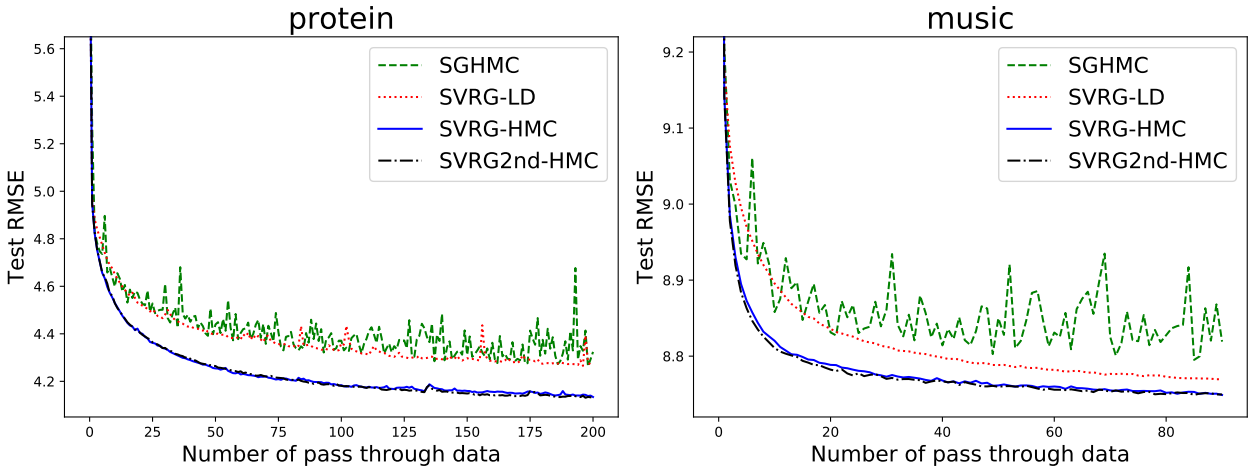

Figure 3: Performance comparison of vanilla SGHMC, SVRG-LD, SVRG-HMC, SVRG2nd-HMC on regression tasks using Bayesian neural networks. The x-axis and y-axis represent number of passes through the entire training dataset and average test RMSE respectively.

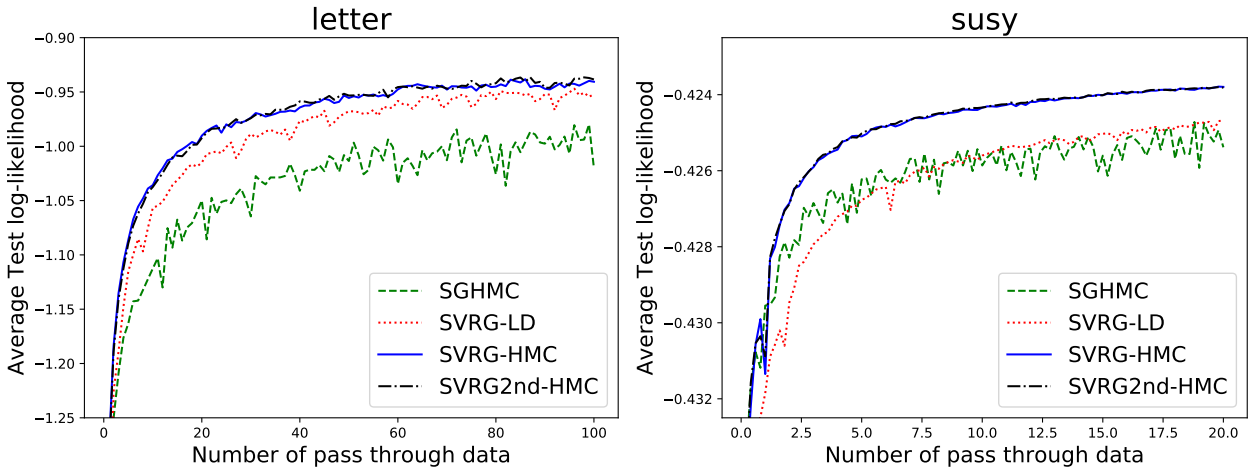

Figure 4: Performance comparison of vanilla SGHMC, SVRG-LD, SVRG-HMC, SVRG2nd-HMC on classification tasks using Bayesian neural networks. The x-axis and y-axis represent number of passes through the entire training dataset and average test log-likelihood respectively.

Experimental results show that SVRG/SVRG2nd-HMC outperforms vanilla SGHMC and SVRG-LD, often by a significant gap. In particularly, this means that variance reduction technique indeed helps the convergence of SGHMC, i.e., the performance gap between SVRG/SVRG2nd-HMC and SVRG-LD in Figure 1–4 is not only coming from the superiority of HMC compared with LD. Similar to previous Section 5.1 and 5.2, the performance SVRG2nd-HMC is usually similar (no worse) to SVRG-HMC, and our experiments found sometimes a slightly larger step size can be chosen for SVRG2nd-HMC (while the same step size brings SVRG-HMC to NaN), which is also consistent with our theoretical results Theorem 5.

6 Conclusion

In this paper, we propose four variance-reduced Hamiltonian Monte Carlo algorithms, i.e., SVRG-HMC, SAGA-HMC, SVRG2nd-HMC and SAGA2nd-HMC for Bayesian Inference. These proposed algorithms guarantee improved theoretical convergence results and converge markedly faster than the benchmarks (vanilla SGHMC and SVRG/SAGA-LD) in practice. In conclusion, the SVRG2nd/SAGA2nd-HMC are more preferable than SVRG/SAGA-HMC according to our theoretical and experimental results. We would like to note that, our variance-reduced Hamiltonian Monte Carlo samplers are not Markovian procedures, but fortunately our theoretical analysis does not rely on any properties of Markov processes, and so it does not affect the correctness of Theorem 2, 4, 5 and 6.

For future work, it would be interesting to study whether our analysis can be apply to vanilla SGHMC without variance reduction. To the best of our knowledge, there is no existing work to prove that SGHMC is better than SGLD. On the other hand, we note that stochastic thermostat [Ding et al., 2014] could outperform both SGLD and SGHMC. It might be interesting to study if a variance-reduced variant of stochastic thermostat could also beat SVRG-LD and SVRG/SVRG2nd-HMC both theoretically and experimentally.

Acknowledgments

We would like to thank Chang Liu for useful discussions.

References

Sungjin Ahn, Anoop Korattikara, and Max Welling. Bayesian posterior sampling via stochastic gradient fisher scoring. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1771–1778, 2012.

Sungjin Ahn, Babak Shahbaba, and Max Welling. Distributed stochastic gradient mcmc. In *International Conference on Machine Learning*, pages 1044–1052, 2014.

Zeyuan Allen-Zhu. Katyusha: the first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1200–1205. ACM, 2017.

Zeyuan Allen-Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. In *International Conference on Machine Learning*, pages 699–707, 2016.

Simon Byrne and Mark Girolami. Geodesic monte carlo on embedded manifolds. *Scandinavian Journal of Statistics*, 40(4):825–845, 2013.

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

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

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

Nan Ding, Youhan Fang, Ryan Babbush, Changyou Chen, Robert D Skeel, and Hartmut Neven. Bayesian sampling using stochastic gradient thermostats. In *Advances in Neural Information Processing Systems*, pages 3203–3211, 2014.

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

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

Rong Ge, Zhize Li, Weiyao Wang, and Xiang Wang. Stabilized svrg: Simple variance reduction for nonconvex optimization. In *Conference on Learning Theory*, 2019.

Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

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

Guanghui Lan, Zhize Li, and Yi Zhou. A unified variance-reduced accelerated gradient method for convex optimization. *arXiv preprint arXiv:1905.12412*, 2019.

Benedict Leimkuhler and Xiaocheng Shang. Adaptive thermostats for noisy gradient systems. *SIAM Journal on Scientific Computing*, 38(2):A712–A736, 2016.

Zhize Li. Ssrgd: Simple stochastic recursive gradient descent for escaping saddle points. *arXiv preprint arXiv:1904.09265*, 2019.

Zhize Li and Jian Li. A simple proximal stochastic gradient method for nonsmooth nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 5569–5579, 2018.

Chang Liu, Jun Zhu, and Yang Song. Stochastic gradient geodesic mcmc methods. In *Advances in Neural Information Processing Systems*, pages 3009–3017, 2016.

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

Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11), 2011.

Sam Patterson and Yee Whye Teh. Stochastic gradient riemannian langevin dynamics on the probability simplex. In *Advances in Neural Information Processing Systems*, pages 3102–3110, 2013.

Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabás Póczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pages 314–323, 2016.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

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

Difan Zou, Pan Xu, and Quanquan Gu. Stochastic variance-reduced hamilton monte carlo methods. *arXiv preprint arXiv:1802.04791*, 2018.

A SAGA Experiments

In this appendix, we report the corresponding experimental results of SAGA variants (i.e., SAGA-LD, SAGA-HMC and SAGA2nd-HMC) for Bayesian regression and Bayesian classification tasks. The settings are the same as those in Section 5.1 and 5.2.

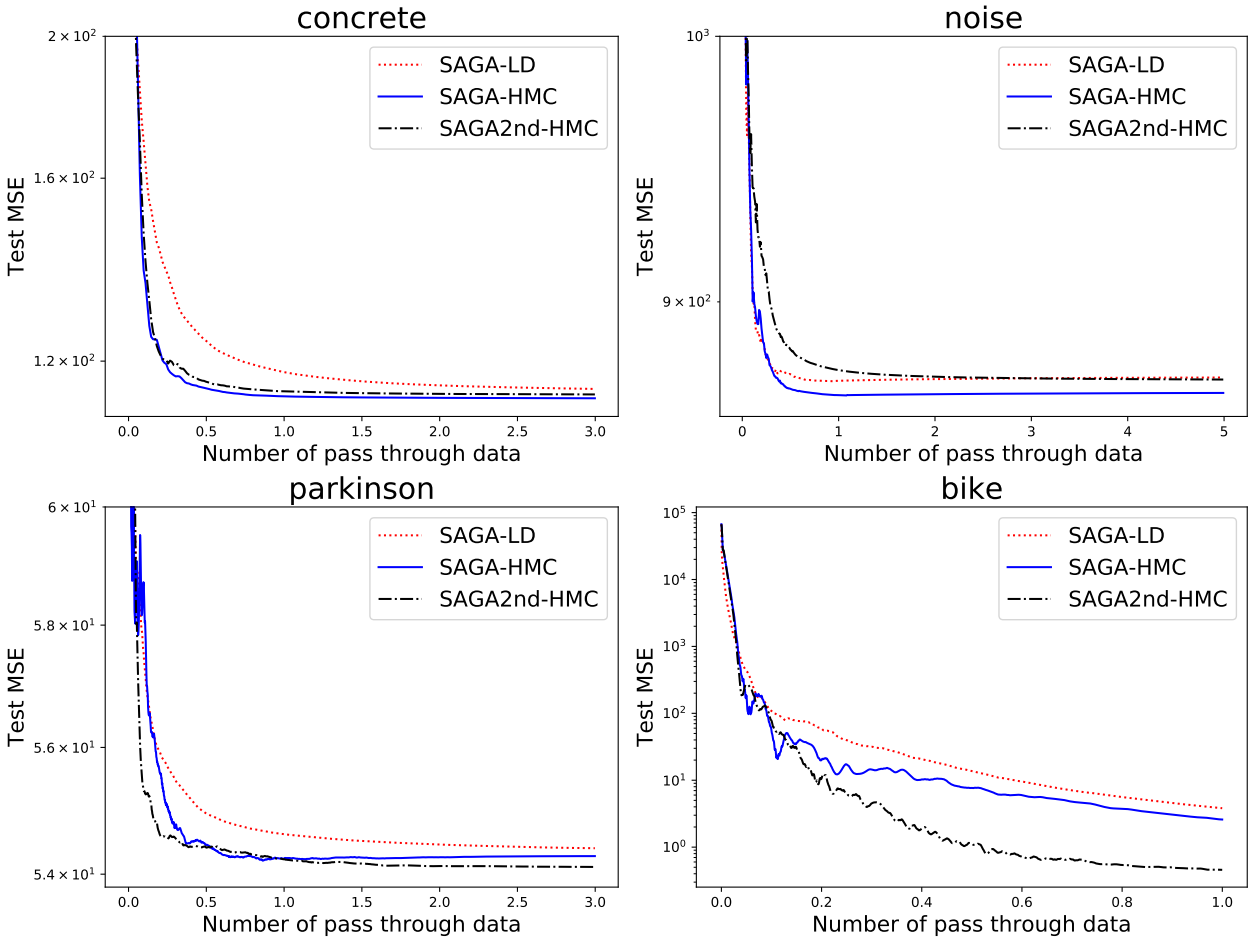

Figure 5: Performance comparison of SAGA variants on Bayesian regression tasks.

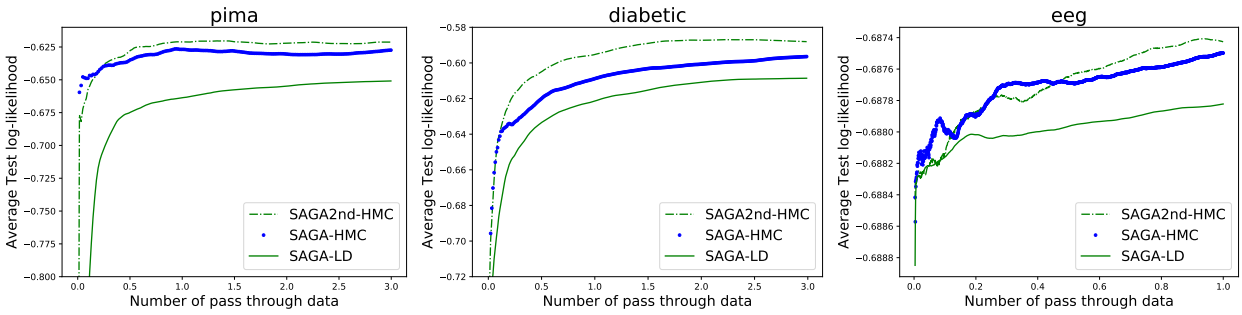

Figure 6: Performance comparison of SAGA variants on Bayesian classification tasks.

B Missing Proofs

In this appendix, we provide the detailed proofs for Corollary 2, Theorem 2, Lemma 1, and Theorem 4, 5 and 6.

B.1 Proof of Corollary 2

To prove this corollary, it is sufficient to show $\mathbb{E}\|\tilde{\nabla}_t - \nabla f(\theta_t)\|^2 \leq n^2G^2/b$. Recall that $\tilde{\nabla}_t = \frac{n}{b}\sum_{i\in I}\nabla f_i(\theta_t)$, I being a b-element index set uniformly randomly drawn (with replacement) from $\{1, 2, \ldots, n\}$ and $\nabla f(\theta_t) = \sum_{j=1}^n \nabla f_j(\theta_t)$. Now, we prove this inequality as follows:

$$
\begin{aligned}
\mathbb{E}_I\|\tilde{\nabla}_t - \nabla f(\theta_t)\|^2 &= \mathbb{E}_I\big\|\frac{n}{b}\sum_{i\in I}\nabla f_i(\theta_t) - \sum_{j=1}^n \nabla f_j(\theta_t)\big\|^2 \\
&= n^2\mathbb{E}_I\big\|\frac{1}{b}\sum_{i\in I}\nabla f_i(\theta_t) - \frac{1}{n}\sum_{j=1}^n \nabla f_j(\theta_t)\big\|^2 \\
&= n^2\mathbb{E}_I\big\|\frac{1}{b}\sum_{i\in I}\Big(\nabla f_i(\theta_t) - \frac{1}{n}\sum_{j=1}^n \nabla f_j(\theta_t)\Big)\big\|^2 \\
&= \frac{n^2}{b^2}\mathbb{E}_I\big\|\sum_{i\in I}\Big(\nabla f_i(\theta_t) - \frac{1}{n}\sum_{j=1}^n \nabla f_j(\theta_t)\Big)\big\|^2 \\
&= \frac{n^2}{b^2}\mathbb{E}_I\big\|\sum_{i\in I}\Big(\nabla f_i(\theta_t) - \frac{1}{n}\sum_{j=1}^n \nabla f_j(\theta_t)\Big)\big\|^2 \\
&= \frac{n^2}{b}\mathbb{E}_i\big\|\nabla f_i(\theta_t) - \frac{1}{n}\sum_{j=1}^n \nabla f_j(\theta_t)\big\|^2 \\
&\leq \frac{n^2}{b}\mathbb{E}_i\big\|\nabla f_i(\theta_t)\big\|^2 \\
&\leq \frac{n^2G^2}{b}
\end{aligned}
$$

where the last two inequalities hold since $\mathbb{E}(X - \mathbb{E}X)^2 \leq \mathbb{E}X^2$ for any random variable X and f_i is G-Lipschitz. □

B.2 Proof of Theorem 2

According to Corollary 1, we have:

$$
\mathbb{E}[(\hat{\phi} - \bar{\phi})^2] \lesssim \frac{1}{T^2}\sum_{t=0}^{T-1}\mathbb{E}[\|\Delta_t\|^2] + \frac{1}{Th} + h^2 \tag{20}
$$

where $\Delta_t = \tilde{\nabla}_t - \nabla f(\theta_t)$ is the additive error in estimating the full gradient $\nabla f(\theta_t)$. By applying the variance reduction technique, we need to upper bound the summation $\sum_{t=1}^T \mathbb{E}[\|\Delta_t\|^2]$.

Unpacking the definition of Δ_t and $\tilde{\nabla}_t$, we have:

$$\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}[\|\Delta_t\|^2]\\
=&\sum_{t=0}^{T-1}\mathbb{E}[\|-\nabla\log\Pr(\theta_t)+\frac{n}{b}\sum_{i\in I}(\nabla f_i(\theta_t)-\nabla f_i(\theta_{\lfloor\frac{t}{K}\rfloor K}))+g-\nabla f(\theta_t)\|^2]\\
=&\sum_{t=0}^{T-1}n^2\mathbb{E}[\|\frac{1}{b}\sum_{i\in I}(\nabla f_i(\theta_t)-\nabla f_i(\theta_{\lfloor\frac{t}{K}\rfloor K}))-\frac{1}{n}\sum_{j=1}^{n}(\nabla f_j(\theta_t)-\nabla f_j(\theta_{\lfloor\frac{t}{K}\rfloor K}))\|^2]\\
\leq&\frac{n^2}{b}\sum_{t=0}^{T-1}\mathbb{E}_{i\in[n]}\|\nabla f_i(\theta_t)-\nabla f_i(\theta_{\lfloor\frac{t}{K}\rfloor K})\|^2 \qquad (21)
\end{aligned}$$

The Inequality (21) is due to $\mathbb{E}(X-\mathbb{E}X)^2\leq\mathbb{E}X^2$ for any random variable X. In the rightmost summation, index i is picked uniformly random from $[n]=\{1,2,\ldots,n\}$.

Then, we bound the RHS of (21) as follows:

$$\begin{aligned}
\frac{n^2}{b}\sum_{t=0}^{T-1}\mathbb{E}_{i\in[n]}\|\nabla f_i(\theta_t)-\nabla f_i(\theta_{\lfloor\frac{t}{K}\rfloor K})\|^2 &\leq \frac{n^2}{b}\sum_{t=0}^{T-1}L^2\mathbb{E}\|\theta_t-\theta_{\lfloor\frac{t}{K}\rfloor K}\|^2\\
&\leq \frac{L^2n^2}{b}\sum_{t=0}^{T-1}K\sum_{j=\lfloor\frac{t}{K}\rfloor K}^{t-1}\mathbb{E}\|\theta_{j+1}-\theta_j\|^2\\
&\leq \frac{L^2n^2K^2}{b}\sum_{t=0}^{T-1}\mathbb{E}\|\theta_{t+1}-\theta_t\|^2 \qquad (22)
\end{aligned}$$

The first inequality is by L-smoothness of all f_i's, and the second one is by Cauchy's inequality.

By our algorithm, $\|\theta_{t+1}-\theta_t\|^2=h^2\|p_{t+1}\|^2$, so we need to upper bound $\mathbb{E}\|p_{t+1}\|^2$ for each $0\leq t<T$.

By the recursion of p_{t+1},

$$\begin{aligned}
&\mathbb{E}\|p_{t+1}\|^2\\
=&\mathbb{E}\|(1-Dh)p_t-h\tilde{\nabla}_t+\sqrt{2Dh}\xi_t\|^2\\
=&\mathbb{E}\|(1-Dh)p_t-h\nabla f(\theta_t)-h\Delta_t+\sqrt{2Dh}\xi_t\|^2\\
=&\mathbb{E}\|(1-Dh)p_t-h\nabla f(\theta_t)\|^2+h^2\mathbb{E}[\|\Delta_t\|^2]+2Dhd\\
\leq&(1-Dh)^2\mathbb{E}\|p_t\|^2+2Gnh(1-Dh)\sqrt{\mathbb{E}\|p_t\|^2}+h^2n^2G^2\\
&+h^2\mathbb{E}[\|\Delta_t\|^2]+2Dhd
\end{aligned}$$

The third equality holds because $\mathbb{E}[\Delta_t]=\mathbb{E}[\xi_t]=0$ and Δ_t and ξ_t are independent. The first inequality takes advantage of $\|\nabla f\|\leq nG$ and $\mathbb{E}\|p_t\|\leq\sqrt{\mathbb{E}\|p_t\|^2}$.

Define $S=\sum_{t=1}^{T}\mathbb{E}[\|p_t\|^2]$. Then, taking a grand summation over $t=0,1,\ldots,T-1$,

$$\begin{aligned}
S&\leq(1-Dh)^2S+2nGh(1-Dh)\sum_{t=0}^{T-1}\sqrt{\mathbb{E}\|p_t\|^2}\\
&\quad+T(h^2n^2G^2+2Dhd)+h^2\sum_{t=0}^{T-1}\mathbb{E}[\|\Delta_t\|^2]\\
&\leq(1-Dh)^2S+2nGh(1-Dh)\sqrt{T}\sqrt{S}\\
&\quad+T(h^2n^2G^2+2Dhd)+\frac{L^2n^2K^2h^4}{b}S
\end{aligned}$$

The second inequality again contains an implicit Cauchy's inequality.

Rearranging the terms we have:

$$(1-(1-Dh)^2-\frac{L^2n^2K^2h^4}{b})\frac{S}{T}-2nGh(1-Dh)\sqrt{\frac{S}{T}}-(h^2n^2G^2+2Dhd)\leq 0$$

Solving a quadratic equation with respect to $\sqrt{S/T}$ and ignoring constant factors, we have:

$$\sqrt{\frac{S}{T}}\lesssim\frac{nG+\sqrt{n^2G^2+D^2d}}{D-L^2n^2K^2h^3b^{-1}}\lesssim\frac{\sqrt{n^2G^2+D^2d}}{D-L^2n^2K^2h^3b^{-1}} \tag{23}$$

From (21) and (22), we have:

$$\sum_{t=0}^{T-1}\mathbb{E}[\|\Delta_t\|^2]\leq\frac{L^2n^2K^2}{b}\sum_{t=0}^{T-1}\mathbb{E}\|\theta_{t+1}-\theta_t\|^2$$

Recall that $\|\theta_{t+1}-\theta_t\|^2=h^2\|p_{t+1}\|^2$ and $S=\sum_{t=1}^{T}\mathbb{E}[\|p_t\|^2]$, we have:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\Delta_t\|^2]\lesssim\frac{L^2n^2K^2h^2}{b}\left(\frac{\sqrt{n^2G^2+D^2d}}{D-L^2n^2K^2h^3b^{-1}}\right)^2 \tag{24}$$

On the other hand, we can bound (21) as follows:

$$\begin{aligned}\sum_{t=0}^{T-1}\mathbb{E}[\|\Delta_t\|^2]&\leq\frac{n^2}{b}\sum_{t=0}^{T-1}\mathbb{E}_{i\in[n]}\|\nabla f_i(\theta_t)-\nabla f_i(\theta_{\lfloor\frac{t}{K}\rfloor K})\|^2\\&\leq\frac{4Tn^2G^2}{b}\end{aligned} \tag{25}$$

where (25) holds due to all f_i's are G-Lipschitz and Cauchy's inequality $\|a+b\|^2\leq 2(\|a\|^2+\|b\|^2)$.

Now, the proof of Theorem 2 is finished by combining (20), (24) and (25). □

B.3 Proof of Lemma 1

Since $G^2>K^2(n^2L^2h^2G^2+hd)>K^2n^2L^2h^2G^2$, we have $h<\frac{1}{nKL}$. Therefore, we have

$$\begin{aligned}D-\frac{h^3L^2n^2K^2}{b}&>D-\frac{1}{n^3K^3L^3}\frac{L^2n^2K^2}{b}\\&=D-\frac{1}{nKbL}\\&\gg D-0.1\geq 0.9D\end{aligned}$$

where the last line holds since $nKbL\gg 10$ and $D>1$ (see Line 1 of Algorithm 1). Thus, the proof is reduced to comparing $h^2L^2n^2G^2+hd$ and $\frac{h^2L^2n^2G^2}{D^2}+h^2L^2d$ asymptotically.

Clearly,

$$\begin{aligned}&\frac{h^2L^2n^2G^2}{D^2}+h^2L^2d\\&\leq\frac{h^2L^2n^2G^2}{D^2}+\frac{1}{nKL}hL^2d\\&=\frac{h^2L^2n^2G^2}{D^2}+hd\frac{L}{nK}\\&\leq\max\{\frac{1}{D^2},\frac{L}{nK}\}(h^2L^2n^2G^2+hd)\end{aligned}$$

Note that if $D \geq 1/L\sqrt{h}$, then $\max\{\frac{1}{D^2}, \frac{L}{nK}\}$ turns to be $\frac{L}{nK}$ which is very small. The reason is that:

$$D^2 \geq \frac{1}{L^2 h} \geq \frac{1}{L^2 \frac{1}{nKL}} = \frac{nK}{L}$$

where the second inequality holds due to $h < \frac{1}{nKL}$ which is mentioned above. □

B.4 Proof of Theorem 4

Defining $\Delta_t = \tilde{\nabla}_t - \nabla f(\theta_t)$, it suffices to upper bound $\sum_{t=0}^{T-1} \mathbb{E}[\|\Delta_t\|^2]$ according to (20). Unpacking the definition of Δ_t and $\tilde{\nabla}_t$, we have:

$$\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{E}[\|\Delta_t\|^2] &= \sum_{t=0}^{T-1} \mathbb{E}[\|\tilde{\nabla}_t - \nabla f(\theta_t)\|^2] \\
&= \sum_{t=0}^{T-1} \mathbb{E}[\|\frac{n}{b} \sum_{i \in I} (\nabla f_i(\theta_t) - \nabla f_i(\alpha_t^i)) - \sum_{j=1}^{n} (\nabla f_j(\theta_t) - \nabla f_j(\alpha_t^j))\|^2] \\
&= \sum_{t=0}^{T-1} n^2 \mathbb{E}[\|\frac{1}{b} \sum_{i \in I} (\nabla f_i(\theta_t) - \nabla f_i(\alpha_t^i)) - \frac{1}{n} \sum_{j=1}^{n} (\nabla f_j(\theta_t) - \nabla f_j(\alpha_t^j))\|^2] \\
&\leq \frac{n^2}{b} \sum_{t=0}^{T-1} \mathbb{E}_{i \in [n]}[\|\nabla f_i(\theta_t) - \nabla f_i(\alpha_t^i)\|^2] \qquad (26) \\
&\leq \frac{L^2 n^2}{b} \sum_{t=0}^{T-1} \mathbb{E}_{i \in [n]}[\|\theta_t - \alpha_t^i\|^2]
\end{aligned}$$

The first inequality is because $\mathbb{E}(X - \mathbb{E}X)^2 \leq \mathbb{E}X^2$ for any random variable X; the second inequality holds due to L-smoothness of f_i's.

Let $\gamma = 1 - (1 - 1/n)^b$. Next we upper bound each $\mathbb{E}[\|\theta_t - \alpha_t^i\|^2]$ in the following manner.

$$\begin{aligned}
\mathbb{E}[\|\theta_t - \alpha_t^i\|^2] &= \sum_{j=0}^{t-1} \mathbb{E}[\|\theta_t - \theta_j\|^2] \Pr(\alpha_t^i = \theta_j) \\
&= \sum_{j=0}^{t-1} \mathbb{E}[\|\theta_t - \theta_j\|^2] (1-\gamma)^{t-j-1} \gamma \\
&= h^2 \sum_{j=0}^{t-1} \mathbb{E}[\|p_t + p_{t-1} + \cdots + p_{j+1}\|^2] (1-\gamma)^{t-j-1} \gamma \\
&\leq h^2 \sum_{j=0}^{t-1} (t-j)(1-\gamma)^{t-j-1} \gamma (\mathbb{E}[\|p_t\|^2] + \mathbb{E}[\|p_{t-1}\|^2] + \cdots + \mathbb{E}[\|p_{j+1}\|^2]) \\
&= h^2 \sum_{j=1}^{t} \mathbb{E}[\|p_j\|^2] \sum_{k=0}^{j-1} (t-k)(1-\gamma)^{t-k-1} \gamma \\
&\leq h^2 \sum_{j=1}^{t} \mathbb{E}[\|p_j\|^2] \sum_{k=0}^{\infty} (t-j+k+1)(1-\gamma)^{t-j+k} \gamma \\
&< h^2 \sum_{j=1}^{t} \mathbb{E}[\|p_j\|^2] (\frac{1}{\gamma} + t - j)(1-\gamma)^{t-j}
\end{aligned}$$

The second equality is by direct calculation $\Pr(\alpha_t^i = \theta_j) = (1-\gamma)^{t-j-1}\gamma$; the first inequality is a direct application of Cauchy's inequality; the last inequality is a weighted summation of geometric series $(1-\gamma)^{t-j+k}, k \geq 0$.

Summing over all t and i, we then have:

$$\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{E}_{i\in[n]}[\|\theta_t - \alpha_t^i\|^2] &\leq h^2 \sum_{t=0}^{T-1}\sum_{j=1}^{t} \mathbb{E}[\|p_j\|^2](\frac{1}{\gamma} + t - j)(1-\gamma)^{t-j} \\
&\leq h^2 \sum_{t=1}^{T-1} \mathbb{E}[\|p_t\|^2] \sum_{j=t}^{T-1}(\frac{1}{\gamma} + j - t)(1-\gamma)^{j-t} \\
&\leq h^2 \sum_{t=1}^{T-1} \mathbb{E}[\|p_t\|^2] \sum_{j=t}^{\infty}(\frac{1}{\gamma} + j - t)(1-\gamma)^{j-t} \\
&\leq h^2 \sum_{t=1}^{T-1} \mathbb{E}[\|p_t\|^2] \frac{2}{\gamma^2} \\
&= \frac{2}{\gamma^2} h^2 \sum_{t=1}^{T-1} \mathbb{E}[\|p_t\|^2] \\
&\leq \frac{8h^2n^2}{b^2} \sum_{t=1}^{T} \mathbb{E}[\|p_t\|^2]
\end{aligned}$$

The last inequality is because $(1-1/n)^b \leq \frac{1}{1+\frac{b}{n-1}}$, and thus $\gamma = 1-(1-1/n)^b \geq \frac{\frac{b}{n-1}}{1+\frac{b}{n-1}} > \frac{b}{2n}$; the last inequality holds as mini-batch size b is smaller than dataset size n.

Similar to the previous subsection, we derive upper an upper bound on $\sum_{t=1}^{T} \mathbb{E}[\|p_t\|^2]$. By recursion of p_{t+1}'s, we have:

$$\begin{aligned}
\mathbb{E}\|p_{t+1}\|^2 &= \mathbb{E}\|(1-Dh)p_t - h\tilde{\nabla}_t + \sqrt{2Dh}\xi_t\|^2 \\
&= \mathbb{E}\|(1-Dh)p_t - h\nabla f(\theta_t) - h\Delta_t + \sqrt{2Dh}\xi_t\|^2 \\
&= \mathbb{E}\|(1-Dh)p_t - h\nabla f(\theta_t)\|^2 + h^2\mathbb{E}[\|\Delta_t\|^2] + 2Dhd \\
&\leq (1-Dh)^2\mathbb{E}\|p_t\|^2 + 2Gnh(1-Dh)\sqrt{\mathbb{E}\|p_t\|^2} + h^2n^2G^2 \\
&\quad + h^2\mathbb{E}[\|\Delta_t\|^2] + 2Dhd
\end{aligned}$$

The third equality holds because $\mathbb{E}[\Delta_t] = \mathbb{E}[\xi_t] = 0$ and Δ_t and ξ_t are independent. The first inequality takes advantage of $\|\nabla f\| \leq nG$ and $\mathbb{E}\|p_t\| \leq \sqrt{\mathbb{E}\|p_t\|^2}$.

Define $S = \sum_{t=1}^{T} \mathbb{E}[\|p_t\|^2]$. Then, taking a grand summation over $t = 0, 1, \ldots, T-1$,

$$\begin{aligned}
S &\leq (1-Dh)^2 S + 2nGh(1-Dh)\sum_{t=0}^{T-1}\sqrt{\mathbb{E}\|p_t\|^2} \\
&\quad + T(h^2n^2G^2 + 2Dhd) + h^2\sum_{t=0}^{T-1}\mathbb{E}[\|\Delta_t\|^2] \\
&\leq (1-Dh)^2 S + 2nGh(1-Dh)\sqrt{T}\sqrt{S} \\
&\quad + T(h^2n^2G^2 + 2Dhd) + \frac{8h^4L^2n^4}{b^3}S
\end{aligned}$$

Rearranging the terms we have:

$$(1-(1-Dh)^2 - \frac{8h^4L^2n^4}{b^3})\frac{S}{T} + 2nGh(1-Dh)\sqrt{\frac{S}{T}} + (h^2n^2G^2 + 2Dhd) \leq 0$$

Solving a quadratic equation with respect to $\sqrt{S/T}$ and ignoring constant factors, we have:

$$\sqrt{\frac{S}{T}} \lesssim \frac{nG}{D - h^3L^2n^4b^{-3}} + \frac{\sqrt{n^2G^2 + D^2d}}{D - h^3L^2n^4b^{-3}} \lesssim \frac{\sqrt{n^2G^2 + D^2d}}{D - h^3L^2n^4b^{-3}}$$

Plugging it in

$$\sum_{t=0}^{T-1} \mathbb{E}[\|\Delta_t\|^2] \leq \frac{8h^2L^2n^4}{b^3} \sum_{t=0}^{T-1} \mathbb{E}\|p_t\|^2$$

we have:

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\|\Delta_t\|^2] \lesssim \frac{h^2L^2n^4}{b^3} \left(\frac{\sqrt{n^2G^2 + D^2d}}{D - h^3L^2n^4b^{-3}} \right)^2 \tag{27}$$

Similar to (25), we can bound (26) as follows:

$$\begin{aligned} &\sum_{t=0}^{T-1} \mathbb{E}[\|\Delta_t\|^2] \\ &\leq \frac{n^2}{b} \sum_{t=0}^{T-1} \mathbb{E}_{i\in[n]}[\|\nabla f_i(\theta_t) - \nabla f_i(\alpha_t^i)\|^2] \\ &\leq \frac{4Tn^2G^2}{b} \end{aligned} \tag{28}$$

where (28) holds due to all f_i's are G-Lipschitz and Cauchy's inequality $\|a+b\|^2 \leq 2(\|a\|^2 + \|b\|^2)$.

Now, the proof of Theorem 4 is finished by combining (20), (27) and (28).

□

B.5 Proof of Theorem 5

To prove this theorem, we need the following theorem from [Chen et al., 2015]. The theorem shows that SGHMC with symmetric splitting can improve the dependency of MSE on step size h, thus allowing larger step size and faster MSE convergence.

Theorem 7 ([Chen et al., 2015]) *Let $\tilde{\nabla}_t$ be an unbiased estimate of $\nabla f(\theta_t)$ for all t. Then under Assumption 1, for a smooth test function ϕ, the MSE of SGHMC with symmetric splitting is bounded in the following way:*

$$\mathbb{E}(\hat{\phi} - \bar{\phi})^2 \lesssim \frac{\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}(\Delta V_t \psi(\theta_t, p_t))^2}{T} + \frac{1}{Th} + h^4 \tag{29}$$

Now, we define $\Delta_t = \tilde{\nabla}_t - \nabla f(\theta_t + \frac{h}{2}p_t)$. According to Assumption 2, we have:

$$\mathbb{E}(\hat{\phi} - \bar{\phi})^2 \leq \frac{1}{T^2}\sum_{t=0}^{T-1} \mathbb{E}\|\Delta_t\|^2 + \frac{1}{Th} + h^4 \tag{30}$$

According to (30), we mainly need to bound the term $\sum_{t=0}^{T-1} \mathbb{E}\|\Delta_t\|^2$ for our SVRG2nd-HMC algorithm. First, we unfold the definition of $\tilde{\nabla}_t$,

$$\begin{aligned} \tilde{\nabla}_t = &-\nabla \log \Pr(\theta_t + \frac{h}{2}p_t) + \frac{n}{b}\sum_{i\in I}(\nabla f_i(\theta_t + \frac{h}{2}p_t) - \nabla f_i(\theta_{\lfloor \frac{t}{K} \rfloor K} + \frac{h}{2}p_{\lfloor \frac{t}{K} \rfloor K})) \\ &+ \sum_{i=1}^{n} \nabla f_i(\theta_{\lfloor \frac{t}{K} \rfloor K} + \frac{h}{2}p_{\lfloor \frac{t}{K} \rfloor K}) \end{aligned}$$

Then,

$$\begin{aligned}
\mathbb{E}\|\Delta_t\|^2 &= \mathbb{E}\|\frac{n}{b}\sum_{i\in I}(\nabla f_i(\theta_t+\frac{h}{2}p_t)-\nabla f_i(\theta_{\lfloor\frac{t}{K}\rfloor K}+\frac{h}{2}p_{\lfloor\frac{t}{K}\rfloor K})) \\
&\qquad -\sum_{i=1}^{n}(\nabla f_i(\theta_t+\frac{h}{2}p_t)-\nabla f_i(\theta_{\lfloor\frac{t}{K}\rfloor K}+\frac{h}{2}p_{\lfloor\frac{t}{K}\rfloor K}))\|^2 \\
&\le \mathbb{E}\|\frac{n}{b}\sum_{i\in I}(\nabla f_i(\theta_t+\frac{h}{2}p_t)-\nabla f_i(\theta_{\lfloor\frac{t}{K}\rfloor K}+\frac{h}{2}p_{\lfloor\frac{t}{K}\rfloor K}))\|^2 \\
&\le \frac{n^2}{b}\mathbb{E}_{i\in[n]}\|\nabla f_i(\theta_t+\frac{h}{2}p_t)-\nabla f_i(\theta_{\lfloor\frac{t}{K}\rfloor K}+\frac{h}{2}p_{\lfloor\frac{t}{K}\rfloor K})\|^2 \\
&\le \frac{n^2L^2}{b}\mathbb{E}\|\theta_t+\frac{h}{2}p_t-\theta_{\lfloor\frac{t}{K}\rfloor K}-\frac{h}{2}p_{\lfloor\frac{t}{K}\rfloor K}\|^2 \\
&\le \frac{n^2L^2K}{b}\sum_{j=\lfloor\frac{t}{K}\rfloor K}^{t-1}\mathbb{E}\|\theta_{j+1}+\frac{h}{2}p_{j+1}-\theta_j-\frac{h}{2}p_j\|^2
\end{aligned} \tag{31}$$

Taking a summation we have:

$$\begin{aligned}
\sum_{t=0}^{T-1}\mathbb{E}\|\Delta_t\|^2 &\le \frac{n^2L^2K}{b}\sum_{t=0}^{T-1}\sum_{j=\lfloor\frac{t}{K}\rfloor K}^{t-1}\mathbb{E}\|\theta_{j+1}+\frac{h}{2}p_{j+1}-\theta_j-\frac{h}{2}p_j\|^2 \\
&\le \frac{n^2L^2K^2}{b}\sum_{t=0}^{T-1}\mathbb{E}\|\theta_{t+1}+\frac{h}{2}p_{t+1}-\theta_t-\frac{h}{2}p_t\|^2 \\
&= \frac{n^2L^2K^2h^2}{b}\sum_{t=0}^{T-1}\mathbb{E}\|p_{t+1}\|^2
\end{aligned}$$

The last equality follows by the recursion $\theta_{t+1}=\theta_t+\frac{h}{2}p_{t+1}+\frac{h}{2}p_t$.
By definition of p_{t+1} we have:

$$\begin{aligned}
&\mathbb{E}\|p_{t+1}\|^2 \\
=&\mathbb{E}\|e^{-Dh/2}\big(e^{-Dh/2}p_t-h\tilde{\nabla}_t+\sqrt{2Dh}\xi_t\big)\|^2 \\
=&e^{-Dh}\mathbb{E}\|e^{-Dh/2}p_t-h\Delta_t-h\nabla f(\theta_t+\frac{h}{2}p_t)+\sqrt{2Dh}\xi_t\|^2 \\
=&e^{-Dh}\Big(\mathbb{E}\|e^{-Dh/2}p_t-h\nabla f(\theta_t+\frac{h}{2}p_t)\|^2+\mathbb{E}\|\sqrt{2Dh}\xi_t\|^2+\mathbb{E}\|h\Delta_t\|^2\Big) \\
=&e^{-Dh}\Big(e^{-Dh}\mathbb{E}\|p_t\|^2+2e^{-Dh/2}nG\mathbb{E}\|p_t\|+n^2G^2h^2+2Dhd+h^2\mathbb{E}\|\Delta_t\|^2\Big) \\
\le&(1-\frac{Dh}{4})^2\Big((1-\frac{Dh}{4})^2\mathbb{E}\|p_t\|^2+2(1-\frac{Dh}{4})nG\sqrt{\mathbb{E}\|p_t\|^2}+n^2G^2h^2+2Dhd+h^2\mathbb{E}\|\Delta_t\|^2\Big)
\end{aligned}$$

Define $S=\sum_{t=1}^{T}\mathbb{E}\|p_t\|^2$ and $M=n^2G^2h^2+2Dhd$. Taking a grand summation of the above inequality for

$t = 0, 1, 2, \ldots, T-1$, we have:

$$
\begin{aligned}
S &\leq (1-\frac{Dh}{4})^2\Big((1-\frac{Dh}{4})^2 S + 2(1-\frac{Dh}{4})nG\sum_{t=0}^{T-1}\sqrt{\mathbb{E}\|p_t\|^2} + TM + h^2\sum_{t=0}^{T-1}\mathbb{E}\|\Delta_t\|^2\Big) \\
&\leq (1-\frac{Dh}{4})^2\Big((1-\frac{Dh}{4})^2 S + 2(1-\frac{Dh}{4})nG\sqrt{TS} + TM + h^2\sum_{t=0}^{T-1}\mathbb{E}\|\Delta_t\|^2\Big) \\
&\leq (1-\frac{Dh}{4})^2\Big((1-\frac{Dh}{4})^2 S + 2(1-\frac{Dh}{4})nG\sqrt{TS} + TM + h^2\sum_{t=0}^{T-1}\mathbb{E}\|\Delta_t\|^2\Big) \\
&\leq (1-\frac{Dh}{4})^2\Big((1-\frac{Dh}{4})^2 S + 2(1-\frac{Dh}{4})nG\sqrt{TS} + TM + \frac{n^2L^2K^2h^4}{b}S\Big)
\end{aligned}
$$

Rewriting it as a quadratic inequality with respect to $\sqrt{\frac{S}{T}}$, we have:

$$
\Big(1-(1-\frac{Dh}{4})^4 - (1-\frac{Dh}{4})^2\frac{n^2L^2K^2h^4}{b}\Big)\frac{S}{T} - 2(1-\frac{Dh}{4})^3 nG\sqrt{\frac{S}{T}} - (1-\frac{Dh}{4})^2 M \leq 0
$$

Solve the inequality and ignore constant factors:

$$
\begin{aligned}
\sqrt{\frac{S}{T}} &\lesssim \frac{nG + \sqrt{n^2G^2 + D^2 d}}{D - L^2n^2K^2h^3b^{-1}} \\
&\lesssim \frac{\sqrt{n^2G^2 + D^2 d}}{D - L^2n^2K^2h^3b^{-1}}
\end{aligned}
$$

Similar to the proof of Theorem 2, it easily follows that:

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\Delta_t\|^2 \lesssim \frac{L^2n^2K^2h^2}{b}\left(\frac{\sqrt{n^2G^2 + D^2 d}}{D - L^2n^2K^2h^3b^{-1}}\right)^2 \tag{32}
$$

Similar to (25), we can bound (31) as follows:

$$
\begin{aligned}
\mathbb{E}[\|\Delta_t\|^2] &\leq \frac{n^2}{b}\mathbb{E}_{i\in[n]}\|\nabla f_i(\theta_t + \frac{h}{2}p_t) - \nabla f_i(\theta_{\lfloor\frac{t}{K}\rfloor K} + \frac{h}{2}p_{\lfloor\frac{t}{K}\rfloor K})\|^2 \\
&\leq \frac{4n^2G^2}{b} \qquad (33)
\end{aligned}
$$

where (33) holds due to all f_i's are G-Lipschitz and Cauchy's inequality $\|a+b\|^2 \leq 2(\|a\|^2 + \|b\|^2)$.

Now, the proof of Theorem 5 is finished by combining (30), (32) and (33). □

B.6 Proof of Theorem 6

Similar to the proof of Theorem 5, we define $\Delta_t = \tilde{\nabla}_t - \nabla f(\theta_t + \frac{h}{2}p_t)$. By Assumption 2 and Inequality (29), we have:

$$
\begin{aligned}
\mathbb{E}(\hat{\phi} - \bar{\phi})^2 &\lesssim \frac{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}(\Delta V_t\psi(\theta_t, p_t))^2}{T} + \frac{1}{Th} + h^4 \\
&\leq \frac{1}{T^2}\sum_{t=0}^{T-1}\mathbb{E}\|\Delta_t\|^2 + \frac{1}{Th} + h^4
\end{aligned}
$$

Unpacking the definition of Δ_t and $\tilde{\nabla}_t$, we have:

$$\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{E}[\|\Delta_t\|^2] &= \sum_{t=0}^{T-1} \mathbb{E}[\|\tilde{\nabla}_t - \nabla f(\theta_t)\|^2] \\
&= \sum_{t=0}^{T-1} \mathbb{E}[\| \frac{n}{b} \sum_{i \in I} \Big(\nabla f_i(\theta_t + \frac{h}{2} p_t) - \nabla f_i(\alpha_t^i) \Big) - \sum_{j=1}^{n} \Big(\nabla f_j(\theta_t + \frac{h}{2} p_t) - \nabla f_j(\alpha_t^j) \Big) \|^2] \\
&= \sum_{t=0}^{T-1} n^2 \mathbb{E}[\| \frac{1}{b} \sum_{i \in I} \Big(\nabla f_i(\theta_t + \frac{h}{2} p_t) - \nabla f_i(\alpha_t^i) \Big) - \frac{1}{n} \sum_{j=1}^{n} \Big(\nabla f_j(\theta_t + \frac{h}{2} p_t) - \nabla f_j(\alpha_t^j) \Big) \|^2] \\
&\leq \frac{n^2}{b} \sum_{t=0}^{T-1} \mathbb{E}_{i \in [n]}[\|\nabla f_i(\theta_t + \frac{h}{2} p_t) - \nabla f_i(\alpha_t^i)\|^2] \qquad (34) \\
&\leq \frac{L^2 n^2}{b} \sum_{t=0}^{T-1} \mathbb{E}_{i \in [n]}[\|\theta_t + \frac{h}{2} p_t - \alpha_t^i\|^2] \qquad (35)
\end{aligned}$$

Let $\gamma = 1 - (1 - 1/n)^b$. Then,

$$\begin{aligned}
&\mathbb{E}[\|\theta_t + \frac{h}{2} p_t - \alpha_t^i\|^2] \\
&= \sum_{j=0}^{t-1} \mathbb{E}[\|\theta_t + \frac{h}{2} p_t - \theta_j - \frac{h}{2} p_j\|^2] \Pr(\alpha_t^i = \theta_j + \frac{h}{2} p_j) \\
&= \sum_{j=0}^{t-1} \mathbb{E}[\|\theta_t + \frac{h}{2} p_t - \theta_j - \frac{h}{2} p_j\|^2] (1-\gamma)^{t-j-1} \gamma \\
&= h^2 \sum_{j=0}^{t-1} \mathbb{E}[\|p_t + p_{t-1} + \cdots + p_{j+1}\|^2] (1-\gamma)^{t-j-1} \gamma \\
&\leq h^2 \sum_{j=0}^{t-1} (t-j)(1-\gamma)^{t-j-1} \gamma (\mathbb{E}[\|p_t\|^2] + \mathbb{E}[\|p_{t-1}\|^2] + \cdots + \mathbb{E}[\|p_{j+1}\|^2]) \\
&= h^2 \sum_{j=1}^{t} \mathbb{E}[\|p_j\|^2] \sum_{k=0}^{j-1} (t-k)(1-\gamma)^{t-k-1} \gamma \\
&\leq h^2 \sum_{j=1}^{t} \mathbb{E}[\|p_j\|^2] \sum_{k=0}^{\infty} (t-j+k+1)(1-\gamma)^{t-j+k} \gamma \\
&< h^2 \sum_{j=1}^{t} \mathbb{E}[\|p_j\|^2] (\frac{1}{\gamma} + t - j)(1-\gamma)^{t-j}
\end{aligned}$$

Summing over all t and i, we then have:

$$\begin{aligned}
\sum_{t=0}^{T-1}\mathbb{E}_{i\in[n]}[\|\theta_t+\frac{h}{2}p_t-\alpha_t^i\|^2] &\le h^2\sum_{t=0}^{T-1}\sum_{j=1}^{t}\mathbb{E}[\|p_j\|^2](\frac{1}{\gamma}+t-j)(1-\gamma)^{t-j}\\
&= h^2\sum_{t=0}^{T-1}\sum_{j=1}^{t}\mathbb{E}[\|p_j\|^2](\frac{1}{\gamma}+t-j)(1-\gamma)^{t-j}\\
&\le h^2\sum_{t=1}^{T-1}\mathbb{E}[\|p_t\|^2]\sum_{j=t}^{T-1}(\frac{1}{\gamma}+j-t)(1-\gamma)^{j-t}\\
&\le h^2\sum_{t=1}^{T-1}\mathbb{E}[\|p_t\|^2]\sum_{j=t}^{\infty}(\frac{1}{\gamma}+j-t)(1-\gamma)^{j-t}\\
&\le h^2\sum_{t=1}^{T-1}\mathbb{E}[\|p_t\|^2]\frac{2}{\gamma^2} = \frac{2}{\gamma^2}h^2\sum_{t=1}^{T-1}\mathbb{E}[\|p_t\|^2]\\
&\le \frac{8h^2n^2}{b^2}\sum_{t=1}^{T}\mathbb{E}[\|p_t\|^2] \qquad (36)
\end{aligned}$$

The last inequality is because $(1-1/n)^b \le \frac{1}{1+\frac{b}{n-1}}$, and thus $\gamma = 1-(1-1/n)^n \ge \frac{\frac{b}{n-1}}{1+\frac{b}{n-1}} > \frac{b}{2n}$; the last inequality holds as mini-batch size b is smaller than dataset size n.

Now, we derive an upper bound on $\sum_{t=1}^{T}\mathbb{E}[\|p_t\|^2]$. By recursion of p_{t+1}'s, we have:

$$\begin{aligned}
&\mathbb{E}\|p_{t+1}\|^2\\
&= \mathbb{E}\|e^{-Dh/2}(e^{-Dh/2}p_t - h\tilde{\nabla}_t + \sqrt{2Dh}\xi_t)\|^2\\
&= e^{-Dh}\mathbb{E}\|e^{-Dh/2}p_t - h\Delta_t - h\nabla f(\theta_t+\frac{h}{2}p_t) + \sqrt{2Dh}\xi_t\|^2\\
&= e^{-Dh}\Big(\mathbb{E}\|e^{-Dh/2}p_t - h\nabla f(\theta_t+\frac{h}{2}p_t)\|^2 + \mathbb{E}\|\sqrt{2Dh}\xi_t\|^2 + \mathbb{E}\|h\Delta_t\|^2\Big)\\
&= e^{-Dh}\Big(e^{-Dh}\mathbb{E}\|p_t\|^2 + 2e^{-Dh/2}nG\mathbb{E}\|p_t\| + n^2G^2h^2 + 2Dhd + h^2\mathbb{E}\|\Delta_t\|^2\Big)\\
&\le (1-\frac{Dh}{4})^2\Big((1-\frac{Dh}{4})^2\mathbb{E}\|p_t\|^2 + 2(1-\frac{Dh}{4})nG\sqrt{\mathbb{E}\|p_t\|^2} + n^2G^2h^2 + 2Dhd + h^2\mathbb{E}\|\Delta_t\|^2\Big)
\end{aligned}$$

Define $S = \sum_{t=1}^{T}\mathbb{E}\|p_t\|^2$ and $M = n^2G^2h^2 + 2Dhd$. Taking a grand summation of the above inequality for $t = 0, 1, 2, \ldots, T-1$, we have:

$$\begin{aligned}
S &\le (1-\frac{Dh}{4})^2\Big((1-\frac{Dh}{4})^2S + 2(1-\frac{Dh}{4})nG\sum_{t=0}^{T-1}\sqrt{\mathbb{E}\|p_t\|^2} + TM + h^2\sum_{t=0}^{T-1}\mathbb{E}\|\Delta_t\|^2\Big)\\
&\le (1-\frac{Dh}{4})^2\left((1-\frac{Dh}{4})^2S + 2(1-\frac{Dh}{4})nG\sqrt{TS} + TM + h^2\sum_{t=0}^{T-1}\mathbb{E}\|\Delta_t\|^2\right)\\
&\le (1-\frac{Dh}{4})^2\left((1-\frac{Dh}{4})^2S + 2(1-\frac{Dh}{4})nG\sqrt{TS} + TM + h^2\sum_{t=0}^{T-1}\mathbb{E}\|\Delta_t\|^2\right)\\
&\le (1-\frac{Dh}{4})^2\left((1-\frac{Dh}{4})^2S + 2(1-\frac{Dh}{4})nG\sqrt{TS} + TM + \frac{8n^4L^2h^4}{b^3}S\right)
\end{aligned}$$

Similar to the proof of theorem 5, we solve a quadratic inequality with respect to $\sqrt{\frac{S}{T}}$, and then,

$$\sqrt{\frac{S}{T}} \lesssim \frac{\sqrt{n^2G^2 + D^2d}}{D - h^3L^2n^4b^{-3}} \tag{37}$$

From (35), (36), (37) and the definition of S, we have:

$$\begin{aligned}
&\sum_{t=0}^{T-1} \mathbb{E}[\|\Delta_t\|^2] \\
&\leq \frac{8h^2L^2n^4}{b^3} \sum_{t=0}^{T-1} \mathbb{E}\|p_t\|^2 \\
&\leq \frac{8h^2L^2n^4T}{b^3} \left(\frac{\sqrt{n^2G^2 + D^2d}}{D - h^3L^2n^4b^{-3}} \right)^2
\end{aligned} \tag{38}$$

Similar to (25), we can bound (34) as follows:

$$\begin{aligned}
&\sum_{t=0}^{T-1} \mathbb{E}[\|\Delta_t\|^2] \\
&\leq \frac{n^2}{b} \sum_{t=0}^{T-1} \mathbb{E}_{i\in[n]}[\|\nabla f_i(\theta_t + \frac{h}{2}p_t) - \nabla f_i(\alpha_t^i)\|^2] \\
&\leq \frac{4Tn^2G^2}{b}
\end{aligned} \tag{39}$$

where (39) holds due to all f_i's are G-Lipschitz and Cauchy's inequality $\|a+b\|^2 \leq 2(\|a\|^2 + \|b\|^2)$.

Now, the proof of Theorem 6 is finished by combining (30), (38) and (39). □