A Hybrid Stochastic Gradient Hamiltonian Monte Carlo Method

Chao Zhang,[1] **Zhijian Li,**[1] **Zebang Shen,**[2] **Jiahao Xie,**[1] **Hui Qian**[1*]
[1]Zhejiang University
[2]University of Pennsylvania
{zczju,lizhijian}@zju.edu.cn, zebang@seas.upenn.edu, {xiejh,qianhui}@zju.edu.cn

Abstract

Recent theoretical analyses reveal that existing Stochastic Gradient Markov Chain Monte Carlo (SG-MCMC) methods need large mini-batches of samples (exponentially dependent on the dimension) to reduce the mean square error of gradient estimates and ensure non-asymptotic convergence guarantees when the target distribution has a nonconvex potential function. In this paper, we propose a novel SG-MCMC algorithm, called Hybrid Stochastic Gradient Hamiltonian Monte Carlo (HSG-HMC) method, which needs merely one sample per iteration and possesses a simple structure with only one hyperparameter. Such improvement leverages a hybrid stochastic gradient estimator that exploits historical stochastic gradient information to control the mean square error. Theoretical analyses show that our method obtains the best-known overall sample complexity to achieve epsilon-accuracy in terms of the 2-Wasserstein distance for sampling from distributions with nonconvex potential functions. Empirical studies on both simulated and real-world datasets demonstrate the advantage of our method.

Introduction

In this paper, we consider the problem of sampling from a probability measure on $\mathbb{R}^d$, which admits a density $p^*_{\mathbf{x}}$ with respect to the Lebesgue measure on all $\mathbf{x} \in \mathbb{R}^d$ by

$$p^*_{\mathbf{x}} \propto \exp(-f(\mathbf{x})). \tag{1}$$

Here, $f(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$ is a smooth potential function. This problem lies at the core of many Bayesian learning tasks in the artificial intelligence literature (Gelman et al. 2014; Andrieu et al. 2003; Ahn et al. 2015; Blundell et al. 2015; Jaakkola and Jordan 1997). Dynamics based Markov Chain Monte Carlo (MCMC) methods, such as Langevin Monte Carlo (LMC) method (Roberts and Stramer 2002), Underdamped Langevin MCMC (UL-MCMC) method (Kloeden and Platen 1992), and modified UL-MCMC method (Cheng et al. 2017), have been widely adopted to solve this problem, due to their simplicity and effectiveness. Generally, these methods generate iterates by discretizing continuous dynamics whose stationary distribution (or its marginal) is the target distribution $p^*_{\mathbf{x}}$, and the expensive Metropolis Hastings correction step (Hastings 1970) is eschewed. It has been proved that by utilizing the full gradient of the potential function $f(\mathbf{x})$, the distributions of the iterates are driven towards the target distribution $p^*_{\mathbf{x}}$ efficiently (Roberts, Tweedie et al. 1996; Durmus and Moulines 2016a,b; Cheng et al. 2017). Nowadays, the stochastic gradient technique, i.e. constructing a stochastic approximation from a mini-batch of samples to replace the full gradient, has been widely adopted to reduce the per-iteration computational cost, especially in large-scale and complex Bayesian learning tasks (Welling and Teh 2011; Ma, Chen, and Fox 2015; Chen, Fox, and Guestrin 2014; Ahn, Shahbaba, and Welling 2014; Baker et al. 2017; Brosse, Durmus, and Moulines 2018). In this paper, we refer to MCMC methods with stochastic gradients as SG-MCMC methods.

However, the mean square error of the stochastic gradient approximation has to be controlled at a desirable level to ensure non-asymptotic convergence to the target distribution (Chen, Ding, and Carin 2015). To achieve this goal, a common approach is to enlarge the per-iteration mini-batch size. Recent analyses reveal that for tasks with *nonconvex* potential functions (e.g., Bayesian Neural Networks and Gaussian Mixture Models), the required mini-batch size is on the order of $\tilde{\mathcal{O}}(\epsilon^{-4} \exp(\mathcal{O}(d)))$, where ϵ denotes the target accuracy and d is the dimensionality (Raginsky, Rakhlin, and Telgarsky 2017; Gao, Gurbuzbalaban, and Zhu 2018; Zou, Xu, and Gu 2019b). This sample size may be gigantic even for a moderate target accuracy and dimensionality, which would result in a high overall sample complexity.

Inspired by the recent advances in the stochastic gradient descent methods, another class of SG-MCMC methods instead resort to the variance reduction techniques to control the mean square error (Zou, Xu, and Gu 2018a; Li et al. 2019; Zou, Xu, and Gu 2019b; Zhang et al. 2020). In these methods, historical gradient information is reused to reduce the per-iteration mini-batch size. Among them, the SRVR-HMC method (Zou, Xu, and Gu 2019b) employs a recursively updated biased semi-stochastic gradient estimator and achieves the best overall stochastic sample complexity for nonconvex potentials. Though the amortized mini-batch size can be small, SRVR-HMC still needs to periodically draw a large mini-batch of samples ($\tilde{\mathcal{O}}(\epsilon^{-4} \exp(\mathcal{O}(d)))$ to trade off between bias and variance in the mean square error. This renders a nested-loop algorithm with four hyperparameters

[*] Corresponding Author

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

METHOD	OVERALL SAMPLE COMPLEXITY	MINI-BATCH SIZE	HYPERPARAMETERS
SGLD	$\tilde{\mathcal{O}}(\epsilon^{-8}\lambda_*^{-9})$	$\tilde{\mathcal{O}}(\epsilon^{-4}\lambda_*^{-2})$	2
SGHMC	$\tilde{\mathcal{O}}(\epsilon^{-8}\mu_*^{-5})$	$\tilde{\mathcal{O}}(\epsilon^{-4}\mu_*^{-2})$	2
SG-UL-MCMC	$\tilde{\mathcal{O}}(\epsilon^{-6}\mu_*^{-5/2})$	$\tilde{\mathcal{O}}(\epsilon^{-4}\mu_*^{-1})$	2
SRVR-HMC	$\tilde{\mathcal{O}}(\boldsymbol{\epsilon^{-4}\mu_*^{-2}})$	$\tilde{\mathcal{O}}(\epsilon^{-4}\mu_*^{-1})$	4
HSG-HMC(THIS PAPER)	$\tilde{\mathcal{O}}(\boldsymbol{\epsilon^{-4}\mu_*^{-2}})$	**1**	**1**

Table 1: The overall sample complexity, required mini-batch size, and number of hyperparamters of different methods to achieve ϵ-accuracy in terms of 2-Wasserstein distance for sampling from probability measures with nonconvex potential functions $f(\mathbf{x})$. Here, μ_* and λ_* denote the spectral gaps of the Markov processes generated by the Underdamped Langevin Dynamics (3) and the Overdamped Langevin dynamics (Dalalyan 2017b), respectively. Both μ_* and λ_* are on the order of $\exp(-\mathcal{O}(d))$ in the worst case (Raginsky, Rakhlin, and Telgarsky 2017; Gao, Gurbuzbalaban, and Zhu 2018) and μ_* can be on the order of $\mathcal{O}(\sqrt{\lambda_*})$ for a class of target densities (Eberle et al. 2019; Gao, Gurbuzbalaban, and Zhu 2018). Though the amortized mini-batch size of SRVR-HMC can be reduced to $\mathcal{O}(1)$, it still needs to draw a large mini-batch of samples ($\tilde{\mathcal{O}}(\epsilon^{-4}\mu_*^{-1})$) periodically.

(i.e. outer-loop mini-batch size, inner-loop mini-batch size, inner-loop length and stepsize).

The large mini-batches (exponentially dependent on dimensionality) and a complicated structure with multiple hyperparameters herald a shift in the nature of stochastic gradient techniques, i.e. using a small mini-batch of samples to reduce per-iteration computational cost and derive easy-to-implement methods. In this paper, we propose a *one-hyperparameter* SG-MCMC method that requires only *one sample* per iteration, called Hybrid Stochastic Gradient HMC (HSG-HMC). Our method leverages a novel hybrid stochastic gradient estimator to control the mean square error. The contributions of our paper are listed as follows.

- We leverage a convex combination of a pair of stochastic gradients (one is recursively updated and biased, the other is unbiased) to construct a hybrid estimator that backbones our HSG-HMC method. We show that if the combination weight ρ_k in the hybrid estimator decays as $\rho_k = 1/k$, the mean square error could be upper bounded at a desired level, even using a single sample per iteration. With this estimator, our HSG-HMC possesses a simple structure with only one hyperparameter, i.e. the stepsize.
- We carry out theoretical analyses for HSG-HMC aiming at sampling from probability measures with nonconvex potential functions $f(\mathbf{x})$ that satisfy the mean-square-smoothness and dissipativeness conditions. We prove that our method achieves the best-known overall sample complexity to obtain ϵ-accuracy in terms of the $\mathcal{W}_2$ distance with only one sample per-iteration [1]. Actually, HSG-HMC is the *first* SG-MCMC method that dispense with the $\tilde{\mathcal{O}}(\epsilon^{-4}\exp(\mathcal{O}(d))$ mini-batch size.

We compare HSG-HMC with existing theoretically guaranteed stocahstic gradient MCMC methods through experiments on three tasks, including Gaussian Mixture Density Sampling, Bayesian Logistic Regression and Bayesian Neural Network. Empirical results demonstrate the superiority of HSG-HMC, i.e. it achieves the best performance while requiring least samples.

Notation and Preliminaries

Notation. We use $\mathbf{0}$ to denote the d-dimensional vector with all entries being 0 and $\mathbf{I}_d$ to denote an identity matrix of d dimension. For $a, b \in \mathbb{R}^+$, we use $a = \mathcal{O}(b)$ to denote $a \leq Cb$ for some $C > 0$, and use $a = \tilde{\mathcal{O}}(b)$ to hide some logarithmic terms of b. Given a random variable $\mathbf{x}$, $p_{\mathbf{x}}$ and $\mathbb{E}[\mathbf{x}]$ denote its probability density and expectation, respectively. Given two probability measures ω and ν, the $\mathcal{W}_2$ distance between them is defined as

$$\mathcal{W}_2(\omega, \nu) = \left(\inf_{\pi \in \Gamma(\omega,\nu)} \Big(\int \|\mathbf{x} - \mathbf{y}\|_2^2 \mathbf{d}\pi(\mathbf{x}, \mathbf{y}) \Big) \right)^{1/2},$$

where $\Gamma(\omega, \nu)$ denotes the collection of joint distributions with ω and ν being their marginal distributions.

Throughout this paper, we use $\nabla F(\mathbf{x}, \xi)$ to denote a simple unbiased stochastic approximation of $\nabla f(\mathbf{x})$, i.e.,

$$\mathbb{E}_{\xi}[\nabla F(\mathbf{x}, \xi)] = \nabla f(\mathbf{x}), \qquad (2)$$

where ξ is a random variable drawn from a fixed distribution.

Stochastic Gradient Hamiltonian Monte Carlo Methods

A large portion of the recent SG-MCMC methods are based on the Underdamped Langevin Dynamics (ULD) (Jaakkola and Jordan 1997), which is described by the following stochastic differential equation:

$$\begin{cases} \mathbf{dV}_t = -\gamma \mathbf{V}_t \mathbf{d}t - u \nabla f(\mathbf{X}_t)\mathbf{d}t + \sqrt{2\gamma u}\mathbf{dB}_t, \\ \mathbf{dX}_t = \mathbf{V}_t \mathbf{d}t, \end{cases} \qquad (3)$$

where $\gamma > 0$ is called the viscosity parameter, $u > 0$ is the inverse mass, and $\mathbf{B}_t \in \mathbb{R}^d$ is the standard Brownian motion. In dynamics (3), $\mathbf{V}_t$ and $\mathbf{X}_t$ are referred to as the velocity and position variables, respectively. According to the Fokker-Planck equation (Risken and Frank 1996), the stationary distribution of the position variable $\mathbf{X}_t$ is the target distribution $p^*_{\mathbf{x}}$. As (3) contains a Hamiltonian momentum component, its discretization can be viewed as a form of Hamiltonian MCMC(Cheng et al. 2017). Hence, we follow the convention in this literature and refer to SG-MCMC methods based on this dynamics as stochastic gradient Hamiltonian Monte Carlo methods.

[1] Generate a sample $\mathbf{x}$ whose distribution $p_{\mathbf{x}}$ satisfies $\mathcal{W}_2(p_{\mathbf{x}}, p^*_{\mathbf{x}}) \leq \epsilon$.

SG-UL-MCMC (Cheng et al. 2017) utilizes the first-order exponential integrator discretization scheme (Stetter 1973) of the Underdamped Langevin Dynamics (3) to generate iterates $\mathbf{x}^{(k)}$'s and $\mathbf{v}^{(k)}$'s in an iterative way. Moreover, it utilizes the following mini-batch stochastic gradient estimate to replace the full gradient in the k-th iteration

$$\mathbf{g}_u^{(k)} = \frac{1}{B} \sum_{i \in \mathcal{B}^{(k)}} \nabla F(\mathbf{x}^{(k)}, \xi_i^{(k)})$$

where $B = |\mathcal{B}^{(k)}|$ denotes the mini-batch size and $\nabla F(\mathbf{x}^{(k)}, \xi_i^{(k)})$'s are simple stochastic gradients whose expectation w.r.t $\xi_i^{(k)}$ is $\nabla f(\mathbf{x}^{(k)})$. When $f(\mathbf{x})$ is nonconvex, mini-batch size of order $\tilde{\mathcal{O}}(\epsilon^{-4} \exp(\mathcal{O}(d)))$ is needed to guarantee that the distribution of the position variable $\mathbf{x}^{(k)}$ converges to the target $p_{\mathbf{x}}^*$ (Zou, Xu, and Gu 2019b).

To alleviate the influence of variances of stochastic gradients, SRVR-HMC utilizes historical stochastic gradient information to construct the following recursively updated variance-reduced estimator

$$\mathbf{g}_b^{(k)} = \mathbf{g}_b^{(k-1)} + \frac{1}{B} \sum_{i \in \mathcal{B}^{(k)}} (\nabla F(\mathbf{x}^{(k)}, \xi_i^{(k)}) - \nabla F(\mathbf{x}^{(k-1)}, \xi_i^{(k)})),$$

where $\mathcal{B}^{(k)}$ denotes a mini-batch of size B. It can be verified that $\mathbf{g}_b^{(k)}$ is a biased estimator of $\nabla f(\mathbf{x}^k)$ w.r.t. $\xi_i^{(k)}$'s. To balance variance and bias in the mean square error, SRVR-HMC resets $\mathbf{g}_b^{(k)}$ to $\frac{1}{B_0} \sum_{\xi_i^k \in \mathcal{B}_0^{(k)}} \nabla F(\mathbf{x}^{(k)}, \xi_i^{(k)})$ with a mini-batch $\mathcal{B}_0^{(k)}$ of size B_0 every L iterations. While B can be small, B_0 is required to be on the order of $\tilde{\mathcal{O}}(\epsilon^{-4} \exp(\mathcal{O}(d)))$ to ensure the distributions of iterates $\mathbf{x}^{(k)}$'s converge to the target $p_{\mathbf{x}}^*$ in terms of the $\mathcal{W}_2$ distance. As a result, SRVR-HMC involves four parameters (i.e. L, B, B_0, and the stepsize η) to tune.

The required sample size in SG-UL-MCMC and SRVR-HMC is large for tasks with high dimensionality d due to its exponential dependent on d. Actually, all existing theoretical guaranteed SG-MCMC methods, including SGLD (Raginsky, Rakhlin, and Telgarsky 2017), SGHMC (Gao, Gurbuzbalaban, and Zhu 2018), SG-UL-MCMC and SRVR-HMC, need mini-batches of size $\tilde{\mathcal{O}}(\epsilon^{-4} \exp(\mathcal{O}(d)))$ when dealing with tasks with nonconvex potentials. In Table 1, we summarize the overall stochastic sample complexity, required mini-batch size, and hyperparameter numbers of these methods and HSG-HMC to achieve ϵ-accracy in terms of the $\mathcal{W}_2$ distance[2].

Methodology

Algorithm 1 Hybrid Stochastic Gradient Hamiltonian Monte Carlo (HSG-HMC) method

```
Require: initial iterate $\mathbf{x}^{(0)}$, $\mathbf{v}^{(0)}$, unbiased estimate $\mathbf{g}^{(0)}$ of $\nabla f(\mathbf{x}^{(0)})$, stepsize $\eta$, and total number of iterations $K$.
for $k = 0$ to $K - 1$ do
    Update $\mathbf{x}^{(k+1)}$ and $\mathbf{v}^{(k+1)}$ according to (4) and (5), respectively.
    Calculate the next exploration direction $\mathbf{g}^{(k+1)}$ according to (7).
end for
```

We present our Hybrid Stochastic Gradient Hamiltonian Monte Carlo method (HSG-HMC) in Algorithm 1. We utilize the first-order exponential integrator discretization scheme of the Underdamped Langevin Dynamics (3). It has been shown that this scheme would result in better overall gradient complexity than the Euler discretization based methods for distributions with nonconvex potentials $f(\mathbf{x})$ (Gao, Gurbuzbalaban, and Zhu 2018). In the k-th iteration, HSG-HMC updates $\mathbf{x}^{(k+1)}$ and $\mathbf{v}^{(k+1)}$ as follows,

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

$$\mathbf{v}^{(k+1)} = e^{-\gamma\eta}\mathbf{v}^{(k)} - u\gamma^{-1}(1 - e^{-\gamma\eta})\mathbf{g}^{(k)} + \boldsymbol{\epsilon}_v^{(k)}, \tag{5}$$

where $\boldsymbol{\epsilon}_x^{(k)}$ and $\boldsymbol{\epsilon}_v^{(k)} \in \mathbb{R}^d$ are zero-mean Gaussian random variables whose covariance matrices satisfy

$$\begin{cases} \mathbb{E}[\boldsymbol{\epsilon}_v^{(k)}(\boldsymbol{\epsilon}_v^{(k)})^T] = u(1 - e^{-2\gamma\eta})\mathbf{I}_d, \\ \mathbb{E}[\boldsymbol{\epsilon}_x^{(k)}(\boldsymbol{\epsilon}_x^{(k)})^T] = \frac{u}{\gamma^2}(2\eta\gamma + 4e^{-\eta\gamma} - e^{-2\gamma\eta} - 3)\mathbf{I}_d, \\ \mathbb{E}[\boldsymbol{\epsilon}_v^{(k)}(\boldsymbol{\epsilon}_x^{(k)})^T] = \frac{u}{\gamma}(1 - 2e^{-\gamma\eta} + e^{-2\gamma\eta})\mathbf{I}_d. \end{cases} \tag{6}$$

The gradient estimator $\mathbf{g}^{(k+1)}$ is constructed in the following hybrid way:

$$\begin{aligned} \mathbf{g}^{(k+1)} &= \rho_{k+1} \underbrace{\nabla F(\mathbf{x}^{(k+1)}, \xi^{(k+1)})}_{\hat{\mathbf{g}}_u^{(k+1)}} + \\ &(1 - \rho_{k+1}) \underbrace{(\mathbf{g}^{(k)} + \nabla F(\mathbf{x}^{(k+1)}, \xi^{(k+1)}) - \nabla F(\mathbf{x}^{(k)}, \xi^{(k+1)}))}_{\hat{\mathbf{g}}_b^{(k+1)}}, \end{aligned} \tag{7}$$

where ρ_k is the weight parameter, and $\nabla F(\mathbf{x}^{(k+1)}, \xi^{(k+1)})$ is an unbiased estimate of $\nabla f(\mathbf{x}^{(k+1)})$ (See the definition (2) in Notation). Specifically, $\mathbf{g}^{(k+1)}$ is a convex combination of two parts: i) $\hat{\mathbf{g}}_u^{(k+1)}$, an unbiased high-variance stochastic gradient estimator; ii) $\hat{\mathbf{g}}_b^{(k+1)}$, a biased variance-reduced stochastic gradient estimator. Note that if $\rho_k = 0$, we recover the biased estimator used in SRVR-HMC (with batch-size $B = 1$). The merit lies in that we compensate the biased estimator $\hat{\mathbf{g}}_b^{(k+1)}$ with an unbiased one $\hat{\mathbf{g}}_u^{(k+1)}$ in each iteration to strike a balance between bias and variance, instead of periodically reset it to an unbiased low-variance estimate. Theoretical analyses in Section show that if we choose ρ_k according to $\rho_k = 1/k$, the mean square error of this estimator would be controlled at a desired level. Thus, HSG-HMC can achieve ϵ-accuracy in terms of the $\mathcal{W}_2$ distance with merely one stochastic sample $\xi^{(k)}$ in each iteration.

Remark 1. The hybrid stochastic gradient technique is originally proposed in the optimization literature for solving nonconvex stochastic minimization problems (Cutkosky and

[2] We exclude SVRG-LD/SAGA-LD as they assume $f(\mathbf{x})$ to be decomposable and are not suitable for general $f(\mathbf{x})$

Orabona 2019; Tran-Dinh et al. 2019a). However, our work differs from theirs in at least three aspects: (1) Our method is designed to generate samples whose distribution is close to the target $p^*_{\mathbf{x}}$ in terms of the $\mathcal{W}_2$ distance. In our analyses, it is needed to show that the iterates sufficiently explore all the high probability area of $p^*_{\mathbf{x}}$. In contrast, their methods aim to find a stationary point of a optimization problem. (2) Their algorithms only update one variable per iteration, while our algorithm has an additional Hamiltonian momentum term and therefore updates two variables. The extra Hamiltonian term introduces a great technical challenge in our theoretical analyses. (3) Our weight strategy $\rho_k = 1/k$, which is crucial in deriving the theoretical bound of HSG-HMC, is different from those used in the optimization literature, i.e. the constant weight strategy (Tran-Dinh et al. 2019a) and the adaptive weight strategy (Cutkosky and Orabona 2019). Directly using existing weight strategies in the optimization *would not* result in any favourable theoretical guarantee. Moreover, our weight strategy does not introduce extra hyperparameter while their weight strategies involve additional hyperparameter tuning.

Theoretical Analysis

In this section, we give the theoretical analyses of HSG-HMC. All detailed proofs are deferred to the appendix.

We use the 2-Wasserstein ($\mathcal{W}_2$) distance as our criterion. $\mathcal{W}_2$ distance is widely used in the analysis of dynamics based MCMC methods since it is a more suitable measurement of the closeness between two distributions than metrics like the total variation and the Kullback-Leibler divergence as it can deal with distributions with different supports (Zou, Xu, and Gu 2018a; Dalalyan 2017a; Cheng et al. 2017).

We make the following assumptions on the potential function$f(\mathbf{x})$, which are commonly used in the analysis for sampling from density with nonconvex potentials (Zou, Xu, and Gu 2019a; Raginsky, Rakhlin, and Telgarsky 2017; Gao, Gurbuzbalaban, and Zhu 2018).

Assumption 1 (Unbiasedness and Bounded Variance)**.** *For all $\mathbf{x} \in \mathbb{R}^d$, the unbiased estimator $\nabla F(\mathbf{x}, \xi)$ of $\nabla f(\mathbf{x})$ has a bounded variance $\mathbb{E}\|\nabla F(\mathbf{x}, \xi) - \nabla f(\mathbf{x})\|^2 \leq G^2$, where the sample variable ξ is drawn from certain fixed distribution.*

Assumption 2 (Mean-squared smoothness)**.** *The stochastic gradient $\nabla F(\mathbf{x}, \xi)$ satisfies the following mean-squared smoothness property*

$$\mathbb{E}_\xi\|\nabla F(\mathbf{x}, \xi) - \nabla F(\mathbf{y}, \xi)\|^2 \leq M^2\|\mathbf{x} - \mathbf{y}\|^2, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

Acoordingly, it can be verified that f is M-smooth.

Assumption 3 (Dissipativeness)**.** *There exists constants $m, b > 0$ such that for all $x \in \mathbb{R}^d$, $\nabla f(\mathbf{x})$ satisfies*

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

The dissipativeness assumption is standard for the ergodicity analysis of stochastic differential equations and diffusion approximations, and is essential to guarantee the convergence of Underdamped Langevin dynamics (3) (Roberts, Tweedie et al. 1996; Mattingly, Stuart, and Higham 2002).

First, we show that $(\mathbf{x}^{(k)}, \mathbf{v}^{(k)})$ in Algorithm 1 is an approximate discretization of the ULD(3).

Lemma 1. $(\mathbf{x}^{(k+1)}, \mathbf{v}^{(k+1)})$ *is the solution of the following stochastic differential equation starting from $(\widetilde{\mathbf{X}}_0, \widetilde{\mathbf{V}}_0) = (\mathbf{x}^{(k)}, \mathbf{v}^{(k)})$ at time $t = \eta$,*

$$\begin{cases} \mathbf{d}\widetilde{\mathbf{V}}_t = -\gamma\widetilde{\mathbf{V}}_t\mathbf{d}t - u\mathbf{g}^{(k)}\mathbf{d}t + \sqrt{2\gamma u}\mathbf{d}\mathbf{B}_t, \\ \mathbf{d}\widetilde{\mathbf{X}}_t = \widetilde{\mathbf{V}}_t\mathbf{d}t, \end{cases} \tag{8}$$

where $\mathbf{g}^{(k)}$ is defined as (7).

Based on this, we establish an iterative relation on the mean square error of the gradient estimate $\mathbf{g}^{(k)}$ in two consecutive iterations.

Lemma 2. *Under Assumptions 1 and 2, for $\mathbf{g}^{(k+1)}$ defined in (7), we have*

$$\mathbb{E}\|\mathbf{g}^{(k+1)} - \nabla f(\mathbf{x}^{(k+1)})\|^2 \leq 2\rho_{k+1}^2 G^2 + (1-\rho_{k+1})^2$$
$$(\mathbb{E}\|\mathbf{g}^{(k)} - \nabla f(\mathbf{x}^{(k)})\|^2 + 2M^2\mathbb{E}\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|^2)$$

The analyses in SG-UL-MCMC and SRVR-HMC indicate that if we choose $\rho_k = 0$ or $\rho_k = 1$, the distributions of iterates $\mathbf{x}^{(k)}$'s would not converge to the target distribution as we only use one $\xi^{(k)}$ and require no periodically reconstruction of g^k as in SRVR-HMC.

Next, we show that the mean square error $\mathbb{E}\|\mathbf{g}^{(k)} - \nabla f(\mathbf{x}^k)\|^2$ can be upper bounded explicitly with the weight strategy $\rho_k = 1/k$.

Lemma 3. *Under Assumptions 1 and 2, if we start from $(\mathbf{x}^{(0)}, \mathbf{v}^{(0)}) = (\mathbf{0}, \mathbf{0})$ in Algorithm 1, set $\rho_{k+1} = 1/(k+1)$ and choose proper η which satisfies*

$$\eta \leq \min\left\{\frac{1}{2\gamma}, \frac{6M}{\gamma m}, \frac{\gamma}{4(8Mu + u\gamma + 22\gamma^2)}, \frac{2}{\sqrt{4Mu + 3\gamma^2}},\right.$$
$$\left.\frac{6\gamma bu}{(4Mu + 3\gamma^2)d}, \frac{\gamma^3 m}{48M^3u(288u + 14\gamma)}, \sqrt{\frac{24uM}{5}}\right\}, \tag{9}$$

then we have, for all $k \geq 0$, the mean square error of $\mathbf{g}^{(k)}$ is bounded as

$$\mathbb{E}\|\mathbf{g}^{(k)} - \nabla f(\mathbf{x}^k)\|^2 \leq \frac{2G^2}{k} + CM^2\bar{\mathcal{E}}\eta^2 k, \tag{10}$$

where

$$C = 6(\gamma^2 + 7u^2M^2 + 2ud), \tag{11}$$

and

$$\bar{\mathcal{E}} = \frac{8Mu[16(d+b) + 2m\|\mathbf{x}^*\|^2 + 16u(\gamma^2 + 2u)G^2/\gamma^2]}{\gamma^2 m}$$
$$+ \frac{8u(f(0) - f(\mathbf{x}^*))}{\gamma^2} + \frac{G^2}{M^2} + \gamma ud \tag{12}$$

with $G \geq \max\{\|\nabla F(0, \xi)\|, \sigma^2\} + 1$ a.s..

Proof Sketch. In the proof, we expand $\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|^2$ on the r.h.s. of Lemma 2 according to the update rule of $\mathbf{x}^{(k+1)}$ (4). Then we construct a novel Lyapunov function to bound certain terms in the expansion, and show that (10) holds for all $k \geq 0$ by induction. □

This lemma shows that with $\rho_{k+1} = 1/(k+1)$, the mean square error of $\mathbf{g}^{(k)}$ can be bounded as (10) even with only one sample $\xi^{(k)}$ per iteration. The first term $2G^2/k$ in (10) comes from the unbiased part of $\mathbf{g}^{k+1}$ and decreases with time k, while the second part $CM^2\bar{\mathcal{E}}\eta^2 k$ results from the biased part $\mathbf{g}_b^{k+1}$ and accumulates as k increases. With a proper stepsize η, the second term could be bounded from above. Since the first term becomes small as k increases, the mean square error of $\mathbf{g}^{(k)}$ will be bounded at a desirable level eventually. Note that it is hard to bound the mean square errors if we directly adopt those strategies used in the optimization literature (Cutkosky and Orabona 2019; Tran-Dinh et al. 2019a,b).

Now, we establish the main theorem to bound the Wasserstein distance between the distribution of the k-th iterate $p_{\mathbf{x}^{(k)}}$ and the target $p^*_{\mathbf{x}}$ as follows.

Theorem 1. *Under the Assumptions 1, 2, and 3, if we start from* $(\mathbf{x}^{(0)}, \mathbf{v}^{(0)}) = (\mathbf{0}, \mathbf{0})$*, set* $\rho_{k+1} = 1/(k+1)$ *and choose a step-size* η *satisfying the condition (9) in Algorithm 1, then for all* $0 \le k \le \frac{2}{C\eta^2}\min\{1, \frac{\gamma^2 m}{768uM^3}\}$ *where* C *is defined as (11), we have*

$$\mathcal{W}_2(p_{\mathbf{x}^{(k)}}, p^*_{\mathbf{x}}) \le \Gamma_1(\gamma^2 K\eta^3 + 4\eta\ln K + C\eta^3 K^2)^{1/4} + \Gamma_0\exp(-\mu_* K\eta), \tag{13}$$

where $\Gamma_0 = \mathcal{O}(1/\mu_*)$ *and* $\Gamma_1 = 2\bar{\Lambda}(u\gamma M^2\bar{\mathcal{E}})^{1/4}$ *are parameters with constants* $\bar{\mathcal{E}}$ *defined as (12),*

$$\bar{\Lambda} = \frac{8}{\gamma}\left(\frac{um(f(\mathbf{x}^{(0)}) - f(\mathbf{x}^*)) + 2Mu(4d + 2b + m\|\mathbf{x}^*\|^2\gamma^2) + 12um + 3\gamma^2}{m}\right)^{\frac{1}{2}},$$

and μ_* *denotes the spectral gap of the Markov process generated by the Underdamped Langevin Dynamics (3).*

The first part in the r.h.s. of (13) results from the discretization error between $(\mathbf{x}^{(k)}, \mathbf{v}^{(k)})$ and the continuous Underdamped Langevin Dynamics at time $k\eta$ (i.e. $(\mathbf{X}_{k\eta}, \mathbf{V}_{k\eta})$), and the second part $\mathcal{O}(\exp(-\mu_* K\eta))$ is an upper bound of the $\mathcal{W}_2$ distance between $p_{\mathbf{X}_{k\eta}}$ and the target $p^*_{\mathbf{x}}$. From Theorem 1, we can obtain the overall stochastic sample complexity of HSG-HMC to achieve ϵ-accuracy in terms of the $\mathcal{W}_2$ distance by specifying the stepsize η and iteration number K. We establish the following corollary by requiring both parts in r.h.s of (13) to be less than $\epsilon/2$.

Corollary 1. *Under the same assumptions of Theorem 1, if we set* $\eta = \tilde{\mathcal{O}}(\epsilon^4\mu_*^2)$ *and* $K = \tilde{\mathcal{O}}(\epsilon^{-4}\mu_*^{-3})$*, then HSG-HMC requires* $\tilde{\mathcal{O}}(\epsilon^{-4}\mu_*^{-3})$ *stochastic gradients to achieve* ϵ*-accuracy in terms of the* $\mathcal{W}_2$ *distance.*

Note that for large k, we need a very small η to bound the second part of the mean square error in (10). Actually, as k increases, $\rho_k = \frac{1}{k} \to 0$ and $\mathbf{g}^{(k)}$ would finally degenerate to the biased part $\hat{\mathbf{g}}_b^{(k)}$. Thus, we propose to refresh ρ_k periodically to prevent such degeneration. In the following corollary, we show that a larger η can be used if we reset $\rho_k = 1$ every η^{-1} times. Moreover, under this strategy of ρ_k, HSG-HMC achieves a better overall stochastic sample complexity $\tilde{\mathcal{O}}(\mu_*^{-2}\epsilon^{-4})$.

Corollary 2. *Under the same assumptions of Theorem 1, if we set* $\rho_{k+1} = \frac{1}{mod(k,\lceil\eta^{-1}\rceil)+1}$ *with* $\eta = \tilde{\mathcal{O}}(\epsilon^4\mu_*)$ *and* $K = \tilde{\mathcal{O}}(\epsilon^{-4}\mu_*^{-2})$*, HSG-HMC requires* $\tilde{\mathcal{O}}(\epsilon^{-4}\mu_*^{-2})$ *stochastic gradients to achieve* ϵ*-accuracy in terms of the* $\mathcal{W}_2$ *distance.*

The stochastic sample complexity of HSG-HMC with this weight parameter strategy is the best-known result in the SG-MCMC literature(the same as SRVR-HMC). However, SRVR-HMC needs to tune more hyperparameters than HSG-HMC. As observed in our experiments , all the four hyperparameter L, B_0, B and stepsize η effect the performance of SRVR-HMC greatly and should be tuned carefully. In HSG-HMC, we only need to tune the step-size η, which significantly alleviates the burden of parameter tuning.

Remark 2. Practically, we need not tune u and γ in Underdamped Langevin Dynamics based methods. For example, in SG-UL-MCMC/SRVR-HMC/HSG-MCMC, u is usually fixed to 1, and γ is chosen to make $e^{-\gamma\eta} = 0.9$. Thus, we do not include them as hyperparameters. We will discuss this more in the appendix.

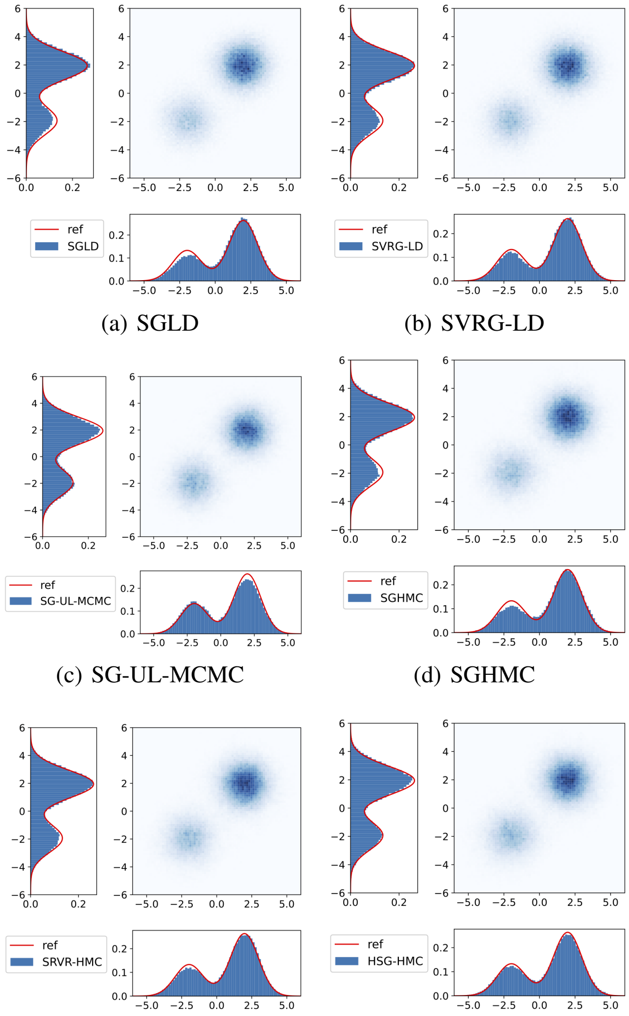
Related Work

In certain Bayesian learning tasks, obtaining the exact gradient $\nabla f(\mathbf{x})$ is computationally expensive or even prohibitive. We list two important examples as follows.

1. In *large-scale* Bayesian posterior inference tasks, $f(\mathbf{x})$ can be chosen as the negative log-posterior, i.e. $f(\mathbf{x}) = -(\sum_{i=1}^n \log p(d_i|\mathbf{x}) - \log p_\theta(\mathbf{x}))$, where $p_\theta(\mathbf{x})$ denotes the prior of $\mathbf{x}$ and $p(d_i|\mathbf{x})$ is the likelihood of data point d_i. In the massive data setting, i.e. n is on the magnitude of millions or billions, the computation of full gradients can be extremely computationally demanding. We can construct a stochastic approximation of $\nabla f(\mathbf{x})$ as $\nabla F(\mathbf{x}, \xi) = -n\nabla\log p(d_\xi|\mathbf{x}) + \nabla\log p_\theta(\mathbf{x})$ with ξ sampled uniformly from $\{1, \cdots, n\}$.

2. Another typical example comes from *hierarchical Bayesian models* (e.g. Latent Dirichlet Allocation model), where the target distribution is $p^*_{\mathbf{x}} = \mathbb{E}_{p_\theta(\xi)}[p(\mathbf{x}|\xi)]$ with $p_\theta(\xi)$ as the prior of the hyperparameter ξ. In this case, $f(\mathbf{x})$ can be set to the negative log-density $f(\mathbf{x}) = -\log p^*_{\mathbf{x}}$. While the exact gradient $\nabla f(\mathbf{x}) = -\mathbb{E}_{p_\theta(\xi)}[\nabla p(\mathbf{x}|\xi)]/\mathbb{E}_{p_\theta(\xi)}[p(\mathbf{x}|\xi)]$ is hard to calculate, its Monte Carlo approximations (Mooney 1997) are easy to obtain.

To handle these tasks, several stochastic gradient MCMC methods have been proposed such as SGLD (Welling and Teh 2011), SVRG-LD (Dubey et al. 2016), SAGA-LD (Dubey et al. 2016), SGHMC (Chen, Fox, and Guestrin 2014), SG-UL-MCMC (Cheng et al. 2017), SVR-HMC (Zou, Xu, and Gu 2018a), SAGA-HMC (Li et al. 2019), SVRG2nd-HMC\SAGA2nd-HMC (Li et al. 2019), and SRVR-HMC (Zou, Xu, and Gu 2019b). Among them, SGLD, SGHMC, SG-UL-MCMC and SRVR-HMC have theoretical guarantees for sampling from densities with general nonconvex $f(\mathbf{x})$.



(a) SGLD

(b) SVRG-LD

(c) SG-UL-MCMC

(d) SGHMC

(e) SRVR-HMC

(f) HSG-HMC

Figure 1: Results on sampling from a Gaussian Mixture Density. The red line denotes the projection of the target distribution $p^*_{\mathbf{x}}$.

Experiments

We follow the settings in the literature (Zou, Xu, and Gu 2018b; Dubey et al. 2016; Chatterji et al. 2018; Welling and Teh 2011; Zou, Xu, and Gu 2019a) and conduct empirical studies on one simulated experiment (sampling from a Gaussian Mixture Density) and two real-world applications (Bayesian Neural Network and Bayesian Logistic Regression). We include SGLD, SVRG-LD, SGHMC, SG-UL-MCMC, and SRVR-HMC as our baselines. In HSG-HMC, we use the weight parameter strategy $\rho_{k+1} = \frac{1}{mod(k,\lceil \eta^{-1} \rceil)+1}$ as indicated in Corollary 2. In all the experiments, we grid search the hyperparameters of each methods. Besides, we also report the results of comparisons between SRVR-HMC with different hyperparameter settings and HSG-HMC, which is deferred to the appendix.

Method	$\mathcal{W}_2$
SGLD	$1.46 * 10^{-2}$
SGHMC	$1.23 * 10^{-2}$
SG-UL-MCMC	$9.58 * 10^{-2}$
SVRG-LD	$7.82 * 10^{-3}$
SRVR-HMC	$5.33 * 10^{-3}$
HSG-HMC	$5.13 * 10^{-3}$

Table 2: The $\mathcal{W}_2$ distance result on sampling from a Gaussian Mixture Density.

Dataset	dim	datasize(training)	datasize(test)
a9a	123	32561	16281
mushroom	112	6000	2124
phishing	68	9000	2055
pima	8	600	168
a3a	123	3185	29376
ijcnn	22	49990	91701

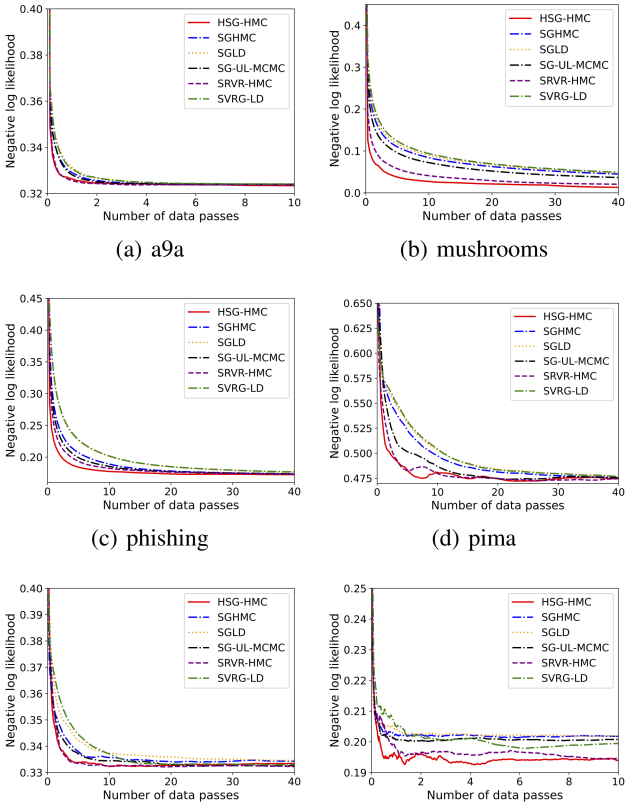
Table 3: Statistics of datasets used in BLR

Sampling from a Gaussian Mixture Density

We consider sampling from distribution $p^*_{\mathbf{x}} \propto \exp(-f(x)) = \exp(-\sum_{i=1}^{N} f_i(x)/N)$, where each component $\exp(-f_i(x))$ is defined as $\exp(-f_i(x)) = 2e^{-\|x-a_i\|_2^2/2} + e^{-\|x+a_i\|2/2}, a_i \in \mathbb{R}^d$. It can be verified that each $\exp(-f_i(x))$ is proportional to the probability density function of two-component Gaussian mixture density with weights $1/3$ and $2/3$. According to (Dalalyan 2017b), when the data point a_i is chosen such that $\|a_i\|^2 \geq 1$, $f_i(x)$ is nonconvex and satisfies Assumption 3. We set the sample size $N = 500$ and dimension $d = 2$, and randomly generate data $a_i \sim \mathbf{N}(\mu, \Sigma)$ with $\mu = (2, \cdots, 2)^T$ and $\Sigma = \mathbf{I}_{d \times d}$. We run each method for 2×10^5 iterations, and make use of the last 10^5 iterates to visualize distributions and calculate the $\mathcal{W}_2$ distance.

In Figure 1, we report the 2D projection of the densities of random samples generated by each algorithm. The results show that SRVR-HMC and HSG-HMC explore both components efficiently while the other methods (SGLD, SL-UL-MCMC, SGHMC and SVRG-LD) concentrate more on one particular component. These results suggest that both the variance reduction techniques and Hamiltonian component are crucial to generate high-quality iterates. Note that, the sample complexity of HSG-HMC is much lower than other methods when generating the same number of iterates.

In Table 2, we report the $\mathcal{W}_2$ distance between each density of random samples generated by each algorithm and the true posterior. To calculate the $\mathcal{W}_2$, we run HMC with MH correction for 2×10^5 iterations and discard the first 10^5 iterates as burn-in to generate 10^5 samples from the true posterior. We use Sinkhorn method with penalty 10^{-3} to calculate the $\mathcal{W}_2$ distance. We can also observe that HSG-HMC and SRVR-HMC perform better than other methods, which confirms our theoretical analysis.



(a) a9a (b) mushrooms

(c) phishing (d) pima

(e) a3a (f) ijcnn

Figure 2: Results for Bayesian Logistic Regression, where X-axis represents the number of data passes, and Y-axis represents the test negative log likelihood.

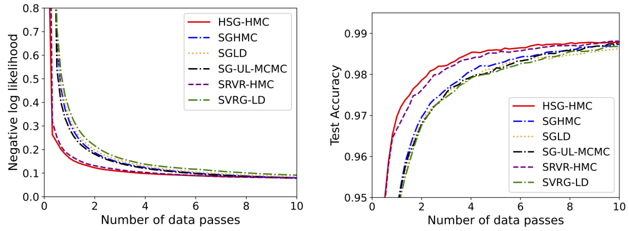
Bayesian Logistic Regression

Bayesian Logistic Regression(BLR) is a robust binary classification task. Let $\mathbf{Z} = \{x_i, y_i\}_{i=1}^N$ be a dataset with $y_i \in \{-1, 1\}$ denoting the sample label and $x_i \in \mathbb{R}^d$ denoting the sample covariate vector. The conditional distribution of label y is modeled by $p(y|x, \mathbf{w}) = \phi(y_i \mathbf{w}^T x_i)$, where $\phi(\cdot)$ is the sigmoid function and the prior of w is $p(w) = \mathbf{N}(0, \lambda \mathbf{I}_{d\times d})$. Six publicly available benchmark datasets, *a9a*, *mushrooms*, *phishing*, *pima*, *a3a* and *ijcnn* are used for evaluation [3]. The statistics of datasets are listed in Table 3.

Follow the convention in (Zou, Xu, and Gu 2019b,a; Welling and Teh 2011), we show the test negative log-likelihood of the test examples on these 6 datasets in Figure 2. We use the number of effective passes (epoch) of the dataset as the x-axis, which is proportional to the overall stochastic sample complexity. It can be observed that HSG-HMC has the best performance on all the datasets, which shows its superiority overall other methods.

Bayesian Neural Network

In this experiment, we study a multiclass Bayesian posterior learning task with Bayesian Neural Network(BNN). Given dataset $\mathbf{Z} = \{x_i, y_i\}_{i=1}^N$, $y_i \in \{0, 1\}^m$ denotes the sample label and $x_i \in \mathbb{R}^d$ denote the sample covariate vector. Note that each y_i is a m-dimensional vector with $y_i^{[k]} = 1$ and all other coordinates 0 if x_i belongs to the k-class. The negative log-likelihood of the conditional distribution of label y_i is proportional to $-\ln p(y_i|x_i, \mathbf{w}) \propto -y_i^T \ln f_{\mathbf{w}}(x_i)$, where $f_{\mathbf{w}}(\cdot) : \mathbb{R}^d \to [0, 1]^k$ denotes a Neural Network. Here, we choose the LeNet (LeCun et al. 1998). All the methods are tested on the standard MNIST dataset, consisting of 28×28 images (thus 784-dimensional input vectors) from 10 different classes (digits from 0 to 9), with 6×10^4 training samples and 10^4 test samples.



(a) Test NLL (b) TestAccuracy

Figure 3: Results for Feedforward Neural Network, where X-axis represents the number of data passes and Y-axes in the left and right subfigures represent the negative log likelihood and the accuracy on the test dataset, respectively

We run all the experiments for 10 times and report the average test log-likelihood and accuracy versus the data passes in Figure 3. The experimental results demonstrate that Underdamped Langevin Dynamics (3) based methods, i.e. SGHMC, SG-UL-MCMC, SRVR-HMC, HSG-HMC have better performance than the overdamped Langevin Dynamics based methods, i.e. SGLD and SVRG-LD. It can also be observed that the variance-reduced methods, i.e. SRVR-HMC and HSG-HMC, outperform ones without variance reduction(i.e. SGHMC and SG-UL-MCMC), and HSG-HMC achieves the best performance.

Conclusion

In this paper, we propose a novel one-sample stochastic gradient Hamiltonian Monte Carlo method, called Hybrid Stochastic Gradient HMC (HSG-HMC). HSG-HMC utilizes a hybrid stochastic gradient estimator, which is a convex combination of an unbiased high-variance estimator and a biased low-variance estimator. We prove that the overall stochastic sample complexity of HSG-HMC is $\tilde{O}(\epsilon^{-4}\mu_*^{-2})$, which achieves the best-known result (the same as SRVR-HMC) when applied in sampling tasks with nonconvex potential functions. While SRVR-HMC needs to alternate between large mini-batch and small mini-batch and possesses a nested-loop structure with 4 parameters, HSG-HMC merely needs one sample in each iteration and is of a simple one-loop structure, which renders much simpler parameter tuning (1 v.s. 4). Empirical results also demonstrate the advantages of our methods over existing works.

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

Acknowledgements

This work is supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LZ18F020002, and National Natural Science Foundation of China (Grant No: 61672376, 61751209, 61472347).

References

Ahn, S.; Korattikara, A.; Liu, N.; Rajan, S.; and Welling, M. 2015. Large-scale distributed Bayesian matrix factorization using stochastic gradient MCMC. In *SIGKDD*, 9–18. ACM.

Ahn, S.; Shahbaba, B.; and Welling, M. 2014. Distributed stochastic gradient MCMC. In *International conference on machine learning*, 1044–1052.

Andrieu, C.; De Freitas, N.; Doucet, A.; and Jordan, M. I. 2003. An introduction to MCMC for machine learning. *Machine learning* 50(1-2): 5–43.

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

Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; and Wierstra, D. 2015. Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424* .

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

Chatterji, N. S.; Flammarion, N.; Ma, Y.-A.; Bartlett, P. L.; and Jordan, M. I. 2018. On the Theory of Variance Reduction for Stochastic Gradient Monte Carlo. *arXiv preprint arXiv:1802.05431* .

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

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

Cheng, X.; Chatterji, N. S.; Bartlett, P. L.; and Jordan, M. I. 2017. Underdamped Langevin MCMC: A non-asymptotic analysis. *arXiv preprint arXiv:1707.03663* .

Cutkosky, A.; and Orabona, F. 2019. Momentum-Based Variance Reduction in Non-Convex SGD. *arXiv preprint arXiv:1905.10018* .

Dalalyan, A. S. 2017a. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. *arXiv preprint arXiv:1704.04752* .

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

Dubey, K. A.; Reddi, S. J.; Williamson, S. A.; Poczos, B.; Smola, A. J.; and Xing, E. P. 2016. Variance Reduction in Stochastic Gradient Langevin Dynamics. In *Advances in Neural Information Processing Systems*, 1154–1162.

Durmus, A.; and Moulines, E. 2016a. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *arXiv preprint arXiv:1605.01559* .

Durmus, A.; and Moulines, E. 2016b. Sampling from strongly log-concave distributions with the Unadjusted Langevin Algorithm. *arXiv preprint arXiv:1605.01559* 5: 3.

Eberle, A.; Guillin, A.; Zimmer, R.; et al. 2019. Couplings and quantitative contraction rates for Langevin dynamics. *The Annals of Probability* 47(4): 1982–2010.

Gao, X.; Gurbuzbalaban, M.; and Zhu, L. 2018. 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* .

Gelman, A.; Carlin, J. B.; Stern, H. S.; and Rubin, D. B. 2014. Bayesian data analysis (Vol. 2). *Boca Raton, FL: Chapman* .

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57(1): 97.

Jaakkola, T.; and Jordan, M. 1997. A variational approach to Bayesian logistic regression models and their extensions. In *AISTATS*, volume 82, 4.

Kloeden, P. E.; and Platen, E. 1992. Higher-order implicit strong numerical schemes for stochastic differential equations. *Journal of statistical physics* 66(1-2): 283–314.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11): 2278–2324.

Li, Z.; Zhang, T.; Cheng, S.; Zhu, J.; and Li, J. 2019. Stochastic gradient hamiltonian monte carlo with variance reduction for bayesian inference. *Machine Learning* 108(8-9): 1701–1727.

Ma, Y.-A.; Chen, T.; and Fox, E. 2015. A complete recipe for stochastic gradient MCMC. In *NIPS*, 2917–2925.

Mattingly, J. C.; Stuart, A. M.; and Higham, D. J. 2002. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic processes and their applications* 101(2): 185–232.

Mooney, C. Z. 1997. *Monte carlo simulation*, volume 116. Sage Publications.

Raginsky, M.; Rakhlin, A.; and Telgarsky, M. 2017. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. *arXiv preprint arXiv:1702.03849* .

Risken, H.; and Frank, T. 1996. *The Fokker-Planck Equation: Methods of Solution and Applications*. Springer Series in Synergetics. Springer Berlin Heidelberg. ISBN 9783540615309. URL https://books.google.com/books?id=MG2V9vTgSgEC.

Roberts, G. O.; and Stramer, O. 2002. Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and computing in applied probability* 4(4): 337–357.

Roberts, G. O.; Tweedie, R. L.; et al. 1996. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* 2(4): 341–363.

Stetter, H. J. 1973. *Analysis of discretization methods for ordinary differential equations*, volume 23. Springer.

Tran-Dinh, Q.; Pham, N. H.; Phan, D. T.; and Nguyen, L. M. 2019a. Hybrid stochastic gradient descent algorithms for stochastic nonconvex optimization. *arXiv preprint arXiv:1905.05920* .

Tran-Dinh, Q.; Pham, N. H.; Phan, D. T.; and Nguyen, L. M. 2019b. A Hybrid Stochastic Optimization Framework for Composite Nonconvex Optimization. *arXiv preprint arXiv:1907.03793* .

Welling, M.; and Teh, Y. W. 2011. Bayesian learning via stochastic gradient Langevin dynamics. In *ICML*, 681–688.

Zhang, C.; Xie, J.; Shen, Z.; Zhao, P.; Zhou, T.; and Qian, H. 2020. Aggregated Gradient Langevin Dynamics. *AAAI* .

Zou, D.; Xu, P.; and Gu, Q. 2018a. Stochastic Variance-Reduced Hamilton Monte Carlo Methods. *arXiv preprint arXiv:1802.04791* .

Zou, D.; Xu, P.; and Gu, Q. 2018b. Subsampled Stochastic Variance-Reduced Gradient Langevin Dynamics. *UAI* .

Zou, D.; Xu, P.; and Gu, Q. 2019a. Sampling from Non-Log-Concave Distributions via Variance-Reduced Gradient Langevin Dynamics. In *AISTATS*, 2936–2945.

Zou, D.; Xu, P.; and Gu, Q. 2019b. Stochastic Gradient Hamiltonian Monte Carlo Methods with Recursive Variance Reduction. In *NeurIPS*, 3830–3841.