arXiv:1504.07107v1 [stat.ML] 27 Apr 2015

Stochastic Subgradient MCMC Methods

Wenbo Hu HWB13@MAILS.TSINGHUA.EDU.CN
Jun Zhu DCSZJ@MAIL.TSINGHUA.EDU.CN
Bo Zhang DCSZB@MAIL.TSINGHUA.EDU.CN
Dept. of Comp. Sci. & Tech., LITS Lab, TNList Lab, Tsinghua University, Beijing 100084, China

Abstract

Many Bayesian models involve continuous but non-differentiable log-posteriors, including the sparse Bayesian methods with a Laplace prior and the regularized Bayesian methods with max-margin posterior regularization that acts like a likelihood term. In analogy to the popular stochastic subgradient methods for deterministic optimization, we present the stochastic subgradient MCMC for efficient posterior inference in such Bayesian models in order to deal with large-scale applications. We investigate the variants that use adaptive stepsizes and thermostats to improve mixing speeds. Experimental results on a wide range of problems demonstrate the effectiveness of our approach.

1. Introduction

Bayesian methods are becoming increasingly relevant in big data applications to protect rich models from overfitting, to account for uncertainty, and to adaptively infer the model complexity via nonparametric techniques. However, with the fast growth of data volume and model size, developing scalable inference algorithms has become a key step in order to apply Bayesian models. Fortunately, much progress has been done on scalable inference with both variational and Markov chain Monte Carlo (MCMC) methods. We refer the readers to (Zhu et al., 2014a) for a review. In particular, stochastic gradient-based MCMC methods have proven effective in exploring high-dimensional sampling spaces, by conjoining the stochastic gradient-based optimization techniques (Robbins & Monro, 1951) and the Markov chain theory. Representative examples include the stochastic gradient Langevin dynamics (SGLD) (Welling & Teh, 2011) and its successors on dealing with manifolds (Patterson & Teh, 2013) and exploring higher-order Hamiltonian dynamics (Chen et al., 2014).

However, one implicit assumption typically made in such gradient-based methods is that the energy function (i.e., log-posteriors) is differentiable, while little work has been done to systematically analyze the continuous but non-differentiable cases, which are not uncommon. For example, sparse Bayesian methods (Park & Casella, 2008) with a doubly exponential prior is non-smooth in the log-space. Another example arises from Bayesian support vector machines (SVMs) (Polson et al., 2011) and the recent work on max-margin Bayesian latent variable models (Zhu et al., 2013; 2014b), whose posterior distributions are regularized by the non-smooth hinge loss. For such non-differentiable models, the posterior inference has been largely hindered by the slow batch algorithms. First, the straightforward application of Metropolis-Hastings MCMC with a random walk proposal (Metropolis et al., 1953; Chib & Greenberg, 1995) is likely to have very low sample efficiency in high dimensional spaces. Second, the Gibbs samplers with data augmentation techniques (Park & Casella, 2008; Polson et al., 2011) are not efficient either in high-dimensional spaces as they often involve inverting large matrices. Moreover, the benefit of introducing extra variables would be counteracted in the view of the extra computation on dealing with the extra sampling variables (Roberts & Stramer, 2002).

In contrast, stochastic subgradient descent (SSGD) methods (Boyd & Mutapcic, 2008) have been extensively analyzed in optimizing non-differentiable objectives with many examples, including the stochastic subgradient solvers (Pegasos) for SVMs (Shalev-Shwartz et al., 2011), the SSGD method for structured SVMs (Ratliff et al., 2007), and the SSGD methods for sparse Lasso regressor (Bertsekas, 1999) and its various extensions (Duchi et al., 2008; Usai et al., 2009). However, none of them has been systematically investigated for efficient Bayesian inference.

In this paper, we conjoin the ideas of stochastic subgradient optimization and Markov chain theory and systematically investigate such stochastic subgradient-based MCMC

methods for Bayesian models with non-differentiable log-posteriors. By generalizing the Hamiltonian dynamics to the subgradient case, we are able to analyze its properties, including volume preservation and the detailed balance. We further explore the stochastic techniques to reduce the computational cost on calculating the full subgradient by replacing it with an unbiased stochastic estimate. By annealing the stepsizes, our stochastic subgradient MCMC methods can converge efficiently to the target posteriors. We empirically demonstrate the effectiveness on the tasks of learning Bayesian SVMs and sparse Bayesian methods with a diverse range of datasets.

The rest of the paper is organized as follows. Section 2 reviews the stochastic gradient-based MCMC methods. Section 3 presents the stochastic subgradient sampling methods. Section 4 presents the experimental results on both Bayesian SVMs and sparse Bayesian logistic regression. Finally, Section 5 concludes.

2. Preliminaries

2.1. Hamiltonian Dynamics

Hamiltonian dynamics is a physical dynamics that generally describes the mechanics (Arnold et al., 2007). The dynamics is described by two variables, namely, the position variable q and the momentum variable p. The time-evolution of the Hamiltonian is determined by the Hamilton's equation:

$$\frac{\mathrm{d}q}{\mathrm{d}t} = \frac{\partial H}{\partial p} \qquad \frac{\mathrm{d}p}{\mathrm{d}t} = -\frac{\partial H}{\partial q}, \tag{1}$$

where the Hamiltonian H is a function of p and q. If we define $U(q)$ as the potential energy, M as the symmetric positive-definite mass matrix, $K(p) = p^\top Mp/2$ as the the kinetic energy, we can usually write the Hamiltonian function as $H(q,p) = U(q) + K(p)$. Then, the Hamiltonian equation can be written as follows:

$$\frac{\mathrm{d}q}{\mathrm{d}t} = M^{-1}p \qquad \frac{\mathrm{d}p}{\mathrm{d}t} = -\frac{\partial U}{\partial q}. \tag{2}$$

2.2. Hamiltonian Monte Carlo

Hamiltonian Monte Carlo, also known as the Hybrid Monte Carlo (HMC) (Duane et al., 1987; Neal, 2012), is one of the classic MCMC methods which combine the physical dynamic simulations with the statistical sampling.

Formally, we consider the most general sense of the statistical learning, where a posterior distribution is any well-defined distribution that takes into account of the prior belief and data (Ghosh & Ramamoorthi, 2003). Besides the conventional Bayes' rule, a posterior distribution may be derived from other rules, such as regularized Bayesian inference (RegBayes) (Zhu et al., 2014b) which solves an optimization problem with nontrivial constraints. A posterior distribution can be represented as $P(\theta|\mathcal{D}) \propto \exp(-U(\theta;\mathcal{D}))$, where $\mathcal{D}$ is the observation dataset and U is a potential energy function. For Bayesian models, U is typically written as:

$$U(\theta|\mathcal{D}) = -\log P_0(\theta) - \sum_{i=1}^{N} \log P(x_i|\theta), \tag{3}$$

a combination of prior $P_0(\theta)$ and some likelihood function $P(\mathcal{D}|\theta) = \prod_{i=1}^{N} P(x_i|\theta)$ with the common i.i.d assumption. Note that by convention, we use θ as the variable of interest in HMC sampling methods which is viewed as the position variable q in Hamiltonian dynamics. Then, an HMC sampler simulates the joint distribution over (θ, p):

$$P(\theta, p) \propto \exp\left(-U(\theta;\mathcal{D}) - p^\top M^{-1}p/2\right). \tag{4}$$

Since it is possible to simulate the dynamics using some discretization methods such as the Euler or leapfrog method, we can derive an HMC sampler to infer the posterior distribution with the assumption that the potential energy $U(q)$ is differentiable. Specifically, using the conventional leapfrog integrator, the HMC method updates the iterations with the stepsize ϵ through:

$$\begin{cases} p_{t+1/2} = p_t - \frac{\epsilon}{2}\nabla U(\theta_t|\mathcal{D}) \\ \theta_{t+1} = \theta_t - \epsilon M^{-1} p_{t+1/2} \\ p_{t+1} = p_{t+1/2} - \frac{\epsilon}{2}\nabla U(\theta_{t+1}|\mathcal{D}), \end{cases} \tag{5}$$

where p_0 is initialized as $p_0 \sim \mathcal{N}(0, M)$. After the samples of (θ, p) are drawn, by simply discarding the augmented momentum variable p, we get the samples from the marginal distribution, which is our target posterior.

In particular, if only one leapfrog step is used and M is set as the identity matrix I, the Langevin Monte Carlo (LMC) can be derived by omitting the momentum variable not used

$$\begin{cases} \theta_{t+1} = \theta_t - \frac{\epsilon^2}{2}\nabla U(\theta_{t+1}|\mathcal{D}) + \epsilon p_t \\ p_{t+1} = p_t - \frac{\epsilon}{2}\nabla U(\theta_t|\mathcal{D}) - \frac{\epsilon}{2}\nabla U(\theta_{t+1}|\mathcal{D}). \end{cases} \tag{6}$$

To compensate for the inaccuracy due to discretization error, a Metropolis-Hastings (MH) correction step is needed to retain the invariance of the target distribution.

2.3. Stochastic Gradient HMC and LMC

One challenge of these gradient-based MCMC methods on dealing with massive data is that the evaluation of the full gradient $\nabla_\theta U(\theta;\mathcal{D})$ is often too expensive. To address this problem, a noisy gradient estimate $\widetilde{\nabla}_\theta U(\theta;\mathcal{D})$ can be constructed by drawing some samples from the whole dataset, as used in the stochastic optimization setting (Robbins & Monro, 1951; Zhang, 2004):

$$\widetilde{\nabla} U(\theta|\mathcal{D}) = -\nabla \log P(\theta) - \tfrac{N}{\widehat{N}} \nabla \log P(\widehat{\mathcal{D}}|\theta), \tag{7}$$

where $\widehat{\mathcal{D}}$ is the randomly drawn subset with size $\widetilde{N}$. Since $\widetilde{N} \ll N$, computing this noisy gradient estimate is much cheaper and the overall algorithm is scalable. This idea has been implemented in (Welling & Teh, 2011) to develop the stochastic gradient Langevin dynamics (SGLD), in (Chen et al., 2014) to develop the stochastic gradient HMC with friction and in (Ding et al., 2014) to develop the stochastic gradient HMC with thermostats.

We briefly review the stochastic gradient HMC with thermostats, or stochastic gradient Nose-Hoover thermostat (SGNHT). SGNHT uses the simple Euler integrator and adds the augmented thermostat variable to control the momentum fluctuations and the injected noise. It simulates via the following equations:

$$\begin{cases} p_{t+1} = p_t - \epsilon\xi_t p_t - \epsilon\widetilde{\nabla}U(\theta_t|\mathcal{D}) + \sqrt{2A}\mathcal{N}(0,\epsilon) \\ \theta_{t+1} = \theta_t + \epsilon p_{t+1} \\ \xi_{t+1} = \xi_t + \epsilon(\frac{1}{n}p_t^\top p_t - 1), \end{cases} \tag{8}$$

where ϵ is the stepsize parameter and A is the diffusion factor parameter. p_0 is initialized as $\mathcal{N}(0, \mathbf{I})$ and ξ_0 is initialized as A.

Similarly, the stochastic gradient LMC (Welling & Teh, 2011) gives samples through:

$$\begin{cases} \theta_{t+1} = \theta_t - \frac{\epsilon_t^2}{2}\widetilde{\nabla}U(\theta_t|\mathcal{D}) + \epsilon_t\eta_t \\ \eta_t \sim \mathcal{N}(0, I), \end{cases} \tag{9}$$

where ϵ_t^2 is the stepsize at step t. In (Welling & Teh, 2011), the polynomially decaying stepsize $\epsilon_t^2 = a(b+t)^{-\gamma}$ is used to omit the MH correction. SGLD can also be considered as a special case of the stochastic gradient Hamiltonian dynamics by setting the leapfrog step as 1.

3. Stochastic Subgradient Sampling Methods

If the log-posterior is non-differentiable, the gradient-based LMC and HMC are not directly applicable. Using the more general subgradients could potentially address this problem, as proven in the deterministic subgradient descent methods (Boyd & Mutapcic, 2008). In this section, we investigate the properties of subgradient-based HMC methods and further propose two stochastic subgradient MCMC sampling methods.

3.1. Generalized Hamiltonian Dynamics

We first show that the non-stochastic subgradient HMC gives the true posterior samples. Similar results hold for the subgradient LMC as it can be seen as a simplified version of HMC.

Formally, we consider the generalized Hamiltonian dynamics with a non-differentiable but continuous potential energy $U(q)$. In this case, we generalize the Hamiltonian equation (2) as

$$\frac{\mathrm{d}q}{\mathrm{d}t} = M^{-1}p \qquad \frac{\mathrm{d}p}{\mathrm{d}t} = -\mathcal{G}_q U(q), \tag{10}$$

where $\frac{\mathrm{d}p}{\mathrm{d}t}$ is the gradient of p over t and $\mathcal{G}_q U(q)$ is the subgradient of U over q.

To analyze the properties of the generalized Hamiltonian dynamics, we restrict ourself to the case where the energy function (i.e., log-posterior) is differentiable, except a finite number of points. We call the functions that satisfy this condition the *piece-wise differentiable* functions. This is of wide interest and covers the max-margin Bayesian models and the sparse Bayesian methods with a Laplace prior. We show that the generalized Hamiltonian dynamics shares the similar properties as the ordinary Hamiltonian dynamics (Neal, 2012) with a differentiable energy function. These properties are significant in constructing the stochastic subgradient sampling methods from the generalized Hamiltonian dynamics.

3.1.1. Reversibility

Reversibility means that any mapping from one state to another state corresponds to an inverse mapping which turns back to the primary state. For the ordinary Hamiltonian dynamics, the reversibility is proved by simply negating the Hamiltonian equation 2 (Neal, 2012). For the generalized Hamiltonian dynamics, we still can show the reversibility by negating the Hamiltonian equation 10. This property indicates the reversibility of the Markov chain transitions of the subgradient HMC.

3.1.2. Conservation of the Hamiltonian

Even though the potential energy field is non-differentiable, we can still show that the Hamiltonian is kept invariant:

$$\mathcal{G}_t H = \sum_i \left\{ \frac{\partial K}{\partial p_i}\mathcal{G}_{q_i}U - \mathcal{G}_{q_i}U\frac{\partial K}{\partial p_i} \right\} = 0. \tag{11}$$

3.1.3. Volume Preservation

Volume preservation, known as Liouville theorem (Arnold et al., 2007), is an important property of the dynamics we considered. Here, volume means the integral of the phase-space volume $\rho = \mathrm{d}^n p \mathrm{d}^n q$, where n is the feature dimension.

In this part, we prove the volume preservation of the generalized Hamiltonian dynamics whose potential energy function has only one non-differentiable point. The extension to the finite case (i.e., piece-wise differentiable) can be done similarly using induction. The basic idea of the analysis is to construct a sequence of ordinary Hamiltonian systems and use them to approximate the generalized Hamiltonian dynamics

Let $\mathcal{H}_0$ be a generalized Hamiltonian dynamics with the volume ρ_0 and the bounded potential energy U_0, which has only one non-differentiable point q_0. We can construct a sequence of ordinary Hamiltonian dynamics H_ϵ with volume ρ_ϵ and a differentiable potential energy $U_\epsilon(q)$ which satisfies $U_\epsilon(q) = U_0(q)$ when $|q - q_0| > \epsilon$. We defer the construction details to Appendix A. Then, we have the following result:

Proposition 1. *The volume of the sequence of the ordinary Hamiltonian dynamics, ρ_ϵ, converges to ρ_0 when ϵ goes to zero which is exactly 0,*

$$0 = \lim_{\epsilon \to 0} \frac{\mathrm{d}\rho_\epsilon}{\mathrm{d}t} = \frac{\mathrm{d}\rho_0}{\mathrm{d}t}. \tag{12}$$

Proof: See Appendix B. □

This result implies that the presence of a finite number of non-differentiable points of the potential energy function does not influence the volume preservation property of the generalized Hamiltonian system.

3.1.4. Subgradient HMC and Its Detailed Balance

By applying the subgradient information to the existing HMC, we can develop the subgradient HMC with leapfrog method as:

$$\begin{cases} p_{t+1/2} = p_t - \frac{\epsilon}{2}\mathcal{G}U(\theta_t|\mathcal{D}) \\ \theta_{t+1} = \theta_t - \epsilon M^{-1} p_{t+1/2} \\ p_{t+1} = p_{t+1/2} - \frac{\epsilon}{2}\mathcal{G}U(\theta_{t+1}|\mathcal{D}), \end{cases} \tag{13}$$

where p_0 is initialized as $p_0 \sim \mathcal{N}(0, M)$ and ϵ is the discretization stepsize.

With the properties of the generalized Hamiltonian dynamics, we can show the detailed balance of the subgradient HMC as follows:

Theorem 2. *The subgradient HMC with an MH test and a discretization integrator (either Euler or leapfrog) leaves the canonical distribution invariant.*
Proof: See Appendix C. □

The detailed balance of the subgradient HMC means that the Metropolis update of the subgradient HMC proposal leaves the canonical distribution invariant.

3.2. Stochastic Subgradient MCMC

We can derive the straightforward version of the stochastic subgradient Langevin dynamics (SSGLD) by simply replacing the posterior gradient with the posterior subgradient. Formally, as outlined in Algorithm 1, SSGLD simulates the samples by running the following dynamics:

$$\begin{cases} \theta_{t+1} = \theta_t - \frac{\epsilon^2}{2}\widetilde{\mathcal{G}}U(\theta_{t+1}|\mathcal{D}) + \epsilon\eta_t \\ \eta_t \sim \mathcal{N}(0, I), \end{cases} \tag{14}$$

Algorithm 1 Stochastic subgradient Langevin Monte Carlo

Input: data X, batchsize $\widetilde{N}$, stepsize ϵ
Initialize θ_0.
for $i = 1, 2, ...$ **do**
  $\theta_{t+1} = \theta_t - \frac{\epsilon^2}{2}\widetilde{\mathcal{G}}U(\theta_{t+1}|X) + \epsilon\mathcal{N}(0, I)$
end for

where $\widetilde{\mathcal{G}}U(\theta|\mathcal{D})$ is the stochastic estimate of the subgradient information $\mathcal{G}U(\theta|\mathcal{D})$

$$\widetilde{\mathcal{G}}U(\theta|\mathcal{D}) = -\mathcal{G}\log P(\theta) - \frac{N}{\widetilde{N}}\mathcal{G}\log P(\widehat{\mathcal{D}}|\theta). \tag{15}$$

In the existing SGLD methods (Welling & Teh, 2011), it is recommended to use the polynomial decaying stepsize to ignore the MH correction step of the Langevin proposals. When the stepsize properly decays, the Markov chain of the SSGLD would converge to the target posterior.

One subtle part of the SGLD method on tuning the discretization stepsizes. A prefixed annealing scheme (if not chosen properly) would make the chain miss the target or bounce around the target. The work (Teh et al., 2014) recommends some relatively optimal scheme for SGLD. Inspired by the an adaptive stepsize for (sub)gradient descent methods (Duchi et al., 2011), we adopt the same adaptive stepsize setting for the SSGLD method. As we shall see in experiments, the adaptive annealing of the stepsizes is often beneficial to give better mixing speeds.

We can similarly derive stochastic subgradient Hamiltonian dynamics using the posterior subgradient. Here, we adopt the optimized version of stochastic HMC (i.e., stochastic gradient Nose-Hoover Thermostat) to derive our stochastic subgradient Nose-Hoover Thermostat (SSNHT), which is outlined in Algorithm 2 and simulates samples via m steps of the following dynamics for each iteration:

$$\begin{cases} p_{t+1} = p_t - \xi_t p_t \epsilon - \epsilon\widetilde{\mathcal{G}}U(\theta_t|X) + \sqrt{2A}\mathcal{N}(0, \epsilon) \\ \theta_{t+1} = \theta_t + \epsilon p_{t+1} \\ \xi_{t+1} = \xi_t + \epsilon(\frac{1}{n}p_t^\top p_t - 1). \end{cases} \tag{16}$$

We also omit the MH correction by adopting the decaying stepsizes. With the properly chosen stepsizes and the thermostats initialization, the SSGNHT simulations would give the posterior samples correctly.

4. Experiments

We now empirically show how our stochastic subgradient sampling methods work on the two representative models, i.e., sparse Bayesian learning with a Laplace prior (Williams, 1995) and Bayesian linear SVMs (Vapnik, 2000), with both low dimensional and high dimensional datasets. Our results demonstrate that the stochastic sub-

Algorithm 2 Stochastic subgradient Nose-Hoover thermostat

Input: data X, batchsize $\widetilde{N}$, stepsize ϵ, step number m
repeat
  Initialize θ_0, A, p_0, $\xi_0 = A$
  for $t = 1$ **to** m **do**
    $p_{t+1} = p_t - \epsilon\xi_t p_t - \epsilon\widetilde{\mathcal{G}}U(\theta_t|X) + \sqrt{2A}\mathcal{N}(0,\epsilon)$
    $\theta_{t+1} = \theta_t + \epsilon p_{t+1}$
    $\xi_{t+1} = \xi_t + \epsilon(\frac{1}{n}p_t^\top p_t - 1)$
  end for
until Converge

gradient MCMC can achieve dramatic improvement on time efficiency while getting accurate posterior samples.

4.1. Bayesian Linear SVMs

Our first example is the Bayesian SVMs for binary classification. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ be the given training dataset, where $x_i \in \mathbb{R}^n$ denotes the input data i and $y_i \in \{+1, -1\}$ is the binary label. We consider the linear classifier, where the decision boundary is determined by a weight vector w and the prediction is done by the sign rule: $\hat{y} = \text{sign}(w^\top x)$. We consider the Bayesian setting, where we are interested in learning a posterior distribution $P(w|\mathcal{D})$, with a prior which is commonly chosen as the standard normal distribution $P_0(w) = \mathcal{N}(0, I)$. As formulated in (Polson et al., 2011), the posterior is

$$P(w|x, y) \propto P_0(w)P(y|x, w) = P_0(w)\prod_i P(y_i|x_i, w),$$

where the per-data unnormalized likelihood is

$$P(y_i|x_i, w) = \exp(-c \cdot \max(0, 1 - y_i w^T x_i)),$$

and c is a nonnegative regularization parameter. Due to the non-conjugacy between the normal prior and the non-Gaussian likelihood, the posterior inference is challenging. The work (Polson et al., 2011) presents a Gibbs sampler by using data augmentation techniques, but it involves the inversion of matrices that are of dimension $n \times n$, which is not scalable for high-dimensional data. We treat this method as a strong baseline. We also compare with the random walk Metropolis (Chib & Greenberg, 1995) on one synthetic dataset and two real datasets. The stochastic MH test (Korattikara et al., 2014) is considered for the random walk Metropolis and we call it the *stochastic random walk Metropolis* (SRWM).

For our stochastic subgradient-based methods, we take the subgradient of the non-differentiable hinge loss as

$$\mathcal{G}\log P(y_i|x_i, w) = \begin{cases} -cy_i x_i & 1 - y_i w^\top x_i > 0 \\ 0 & 1 - y_i w^\top x_i \leq 0. \end{cases} \quad (17)$$

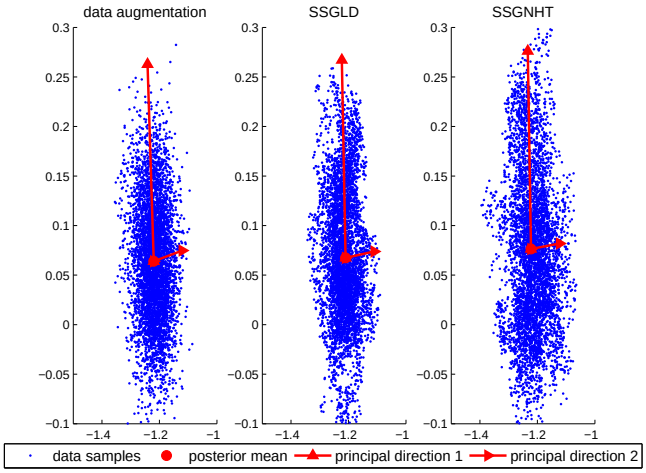


Figure 1. Posterior samples drawn by SSGLD, SSGNHT and the data augmentation sampler.

4.1.1. Results on Synthetic Data

We first test on a two dimensional synthetic dataset to show that our methods give the correct samples from the posterior distribution. Specifically, we generate 1,000 input observations from a uniform distribution $x_i \overset{i.i.d}{\sim} U(0, 1)$ and the two-dimensional coefficient vector as $w \sim \mathcal{N}(0, \lambda^{-1}I)$, where $\lambda = 3$. Given the input observations and the coefficients, the labels are generated from a Bernoulli distribution $B(\alpha, 1 - \alpha)$, where the parameter is

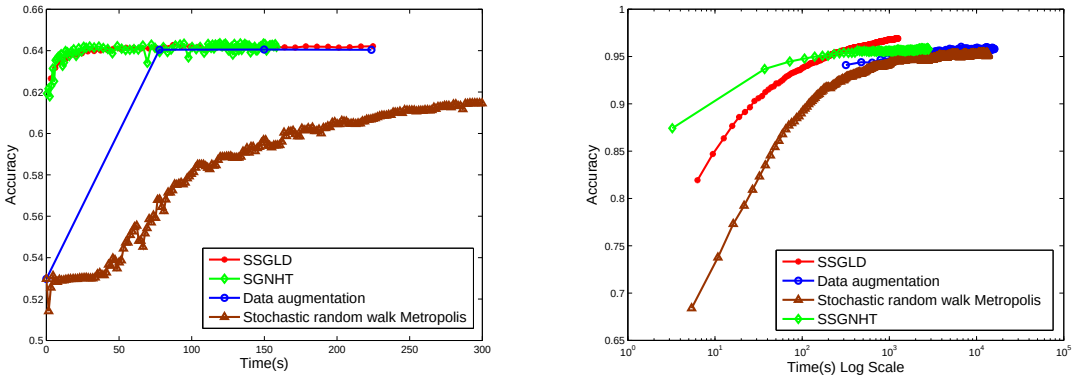
$$\alpha = \frac{P(y = 1|x_i, w)}{P(y = 1|x_i, w) + P(y = -1|x_i, w)}.$$

We compare the samples obtained from SSGLD and SSGNHT with that from the data augmentation method, which is known as an accurate sampler for Bayesian SVMs. We set the batchsize $\widetilde{N} = 10$ for stochastic subgradient MCMC methods. As the data is rather simple, we just set the stepsize as small as 0.0001 for SSGLD and 0.0005 for SSGNHT and omit the MH correction. Besides, for SSGNHT we set the diffusion parameter $A = 1$ and the step number $m = 20$. We take 5,000 samples for each method after a sufficiently long burn-in stage and give the comparison in Figure 1, where the posterior sample mean and the principal directions of the samples are also marked. We can see that our stochastic subgradient MCMC methods are very accurate, almost indistinguishable from the data augmentation sampler.

4.1.2. Results on Real Data

We then test our methods on two real world datasets, namely, the low dimensional higgs dataset (Wang et al., 2014) and the high dimensional realsim dataset [1].

[1] http://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/binary.html



(a) higgs dataset (b) realsim dataset

Figure 2. Accuracy convergence of the binary linear SVM using SSGNHT, SSGLD, data augmentation and SRWM.

The higgs dataset contains 1.1×10^7 samples in a 28-dimensional feature space. We randomly choose 10^7 samples as the training set and use the rest as the testing set. We choose the stochastic batchsize $\widetilde{N}$ as $1,000$ for both SSGLD and the stochastic random walk Metropolis. For SSGHNT, we choose the batchsize as 100, which is a good setting as analyzed in Section 4.1.3. We use the adaptive stepsize for SSGLD, which has been successfully applied in the stochastic (sub)gradient descent (Duchi et al., 2011), and use the polynomial decaying stepsize $0.0001 \times t^{-0.2}$ for SSGNHT. We choose the diffusion factor $A = 1$ and step number $m = 20$ for SSGNHT. For the stochastic random walk Metropolis, the variance parameter is 0.01. The tuning of the parameters is done through the cross validation step of the experiment and we also explain how the batchsize parameter is chosen in Section 4.1.3.

Figure 2(a) shows the convergence curves of various methods on the higgs dataset with respect to the running time. We can see that our stochastic subgradient MCMC methods spend less time to reach the best accuracy obtained by linear SVMs. More specifically, although stochastic subgradient MCMC methods often take more iterations, in order to get a high accuracy, than the data augmentation method, our methods are overall more efficient because at each iteration they draw posterior samples using only a very small mini-batch of the large dataset. Furthermore, compared with the stochastic random walk Metropolis, the gradient information used in our subgradient MCMC methods provides the right direction to the true posterior. Finally, we observe that the two stochastic subgradient MCMC methods have comparable convergence speeds, although in theory SSGNHT would mix faster than SSGLD by using the momenta information (Neal, 2012; Chen et al., 2014).

Figure 3 shows the convergence curves of the SSGLD with adaptive stepsizes and the SSGLD with tuned polynomial decaying stepsizes. We can see that using adaptive stepsizes leads to a better convergence. We also show the good mixing of adopting the adaptive stepsizes in Appendix D.

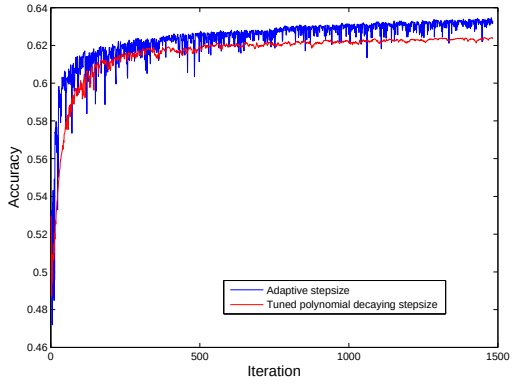


Figure 3. Performance comparison between adaptive stepsize and tuned decaying polynomial

We then test on the high-dimensional realsim dataset, which contains $92,309$ observations in a $20,958$ dimension feature space. We randomly draw $50,000$ observations as the training data and use the rest as the testing data. We also take the stochastic random walk Metropolis and the data augmentation sampler as the baseline methods. We take 10 as the stochastic batchsize for the three stochastic sampling methods. For stochastic random walk Metropolis, we set the variance parameter at 0.05. For SSGNHT, we choose diffusion factor $A = 1$, decaying stepsize $0.001 \times t^{-0.2}$ and the step number $m = 10$. We take the adaptive stepsize for SSGLD. Figure 2(b) shows the convergence of the above methods with respect to the running time in the log scale. We can see that our methods are about about ten times faster than the two baseline methods (i.e., data augmentation and random walk Metropolis), which would miss their sampling direction due to the curse of dimensionality. In contrast, our stochastic subgradient MCMC methods give the more practical solution to the challenging inference task in a high dimensional space.

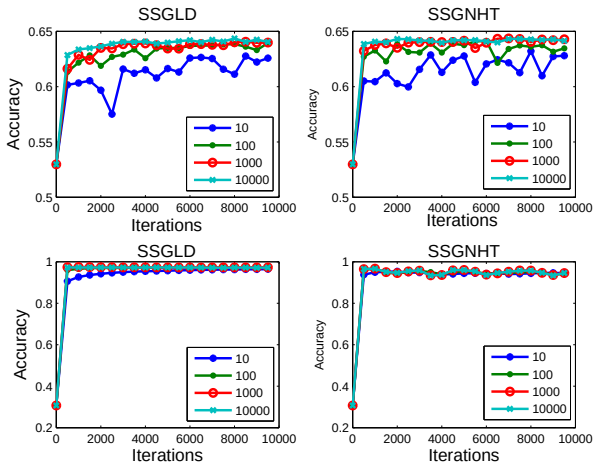


Figure 4. Sensitivity analysis of the batchsize parameter for both SSGLD and SSGNHT on the higgs dataset (**first row**); and the realsim dataset (**second row**).

4.1.3. Sensitivity Analysis

Figure 4 presents the sensitivity analysis of the batchsize $\widetilde{N}$ for the two stochastic subgradient MCMC methods on both the higgs and realsim datasets. We can see that tuning the batchsize $\widetilde{N}$ represents an accuracy-efficiency tradeoff, analogous to the bias-variance tradeoff in stochastic Monte Carlo sampling (Korattikara et al., 2014). In general, using a smaller batchsize often leads to a larger injected noise, but the computation cost at each iteration is reduced, which is linear to the batchsize (i.e., $O(\widetilde{N})$. When doing the cross validation to choose parameters, both the accuracy and time efficiency are key factors that should be taken into consideration. Take the results of SSGNHT on the higgs data as a concrete example (i.e., the upper right subgraph in Figure 4). We can see that using the batchsize of 1,000 (i.e., the circled curve) leads to a slightly more stable curve than the curve with the batchsize of 100 (i.e., the dotted curve). When we take both the accuracy and efficiency into consideration, the mild fluctuation of the green line is not that serious, especially when we consider the computation benefits resulted from the smaller batchsize. Therefore, we choose the batchsize as 100 for the results shown in Figure 2(a).

4.2. Sparse Bayesian Learning

The second example is the sparse Bayesian logistic regression problem. We still consider the binary classification setting with inputs $x_i \in \mathbb{R}^n$ and class labels $y_i \in \{+1, -1\}$. Using a sigmoid function $\sigma(z) = \frac{1}{1+\exp(-z)}$, the posterior distribution of the model is:

$$P(w|\mathcal{D}) = P_0(w)\prod_i P(y_i|x_i, w) = P_0(w)\prod_i \sigma(y_i w^\top x_i).$$

Here the prior of the vector w is the Laplace distribution $P_0(w) = \exp\left(-\lambda^{-1}|w|_1\right)$ with which we can do the sparse feature selection. The posterior computing is also difficult due to the non-conjugacy between the Laplace prior and the sigmoid likelihood. In (Korattikara et al., 2014), the reverse-jump MCMC method is used to do the stochastic sampling for this model with the Metropolis Hastings test. In RJ-MCMC, the MH iterations are finished by adding a binary vector to the corresponding feature vector w and the RJ-MCMC proposal gives samples through a mixture of 3 types of moves (i.e., birth, death and update). This method has already been successfully applied in the sampling of Bayesian lasso (Chen et al., 2011).

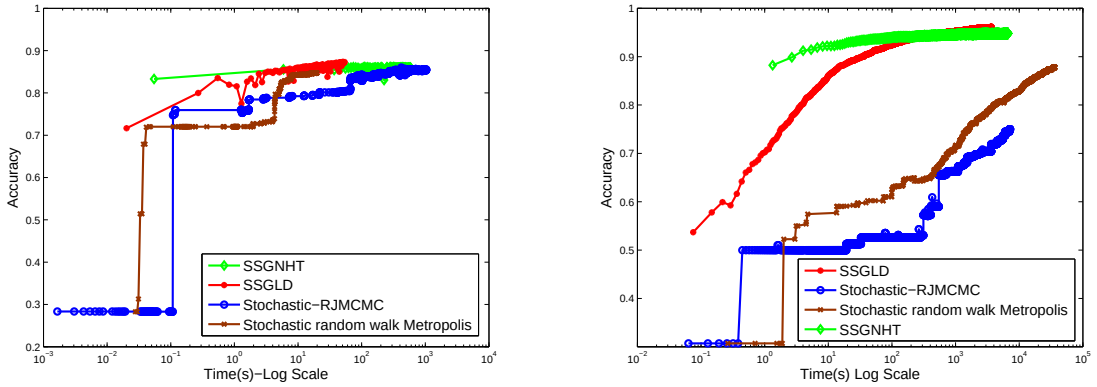
For the stochastic subgradient MCMC methods, the subgradient of the log likelihood is

$$\mathcal{G}_w \log P(y_i|x_i, w) = \sigma(y_i w^\top x_i) y_i x_i,$$

and the subgradient of the log Laplace prior is just taken as $\mathcal{G}_w \log P_0(w) = -\lambda^{-1}\,\text{sign}(w)$. As for the stochastic subgradient MCMC, (Welling & Teh, 2011) ever give a description of SGLD to solve this problem using the subgradient information, but without careful justification. Here, we give a systematic study on both stochastic subgradient LMC and HMC.

We first test our methods on the MiniBooNE dataset from the UCI machine learning repository (Korattikara et al., 2014). The MiniBooNE dataset contains $130,064$ data samples in a 50-dimensional feature space. We randomly choose $100,000$ samples as the training data and use the rest as the testing data. We compared our stochastic subgradient MCMC methods with the stochastic RJ-MCMC (Korattikara et al., 2014) and the stochastic random walk Metropolis (Metropolis et al., 1953). We choose 0.01 as the variance parameter for the stochastic random walk Metropolis and the same variance setting with (Korattikara et al., 2014) for RJ-MCMC. For SSGNHT, the diffusion factor A is set as 1, the stepsize is $0.0001 \times t^{-0.6}$ and the step number is 30. We take 1,000 as the batchsize for the four stochastic sampling methods. We also use the adaptive stepsizes in the SSGLD method for better mixing. We start all the sampling methods with only the first dimension of the coefficient being 1 and all the others being 0.

Figure 5(a) presents the results of various methods with respect to the running time (in log-scale). For the stochastic random walk Metropolis and the stochastic RJ-MCMC, it takes a long time for them in the high dimension space to get a meaningful move, and the zigzag pattern of their convergence curves reveals the low sampling efficiency. For the stochastic subgradient MCMC methods, the accuracy of the sparse Bayesian logistic regression converges within a few steps. The accuracy of the RJ-MCMC method



(a) MiniBooNE dataset (b) Realsim dataset

Figure 5. Convergence of the SSGLD, SSGNHT, SRMW and stochastic RJ-MCMC on the sparse Bayesian logistic regression model

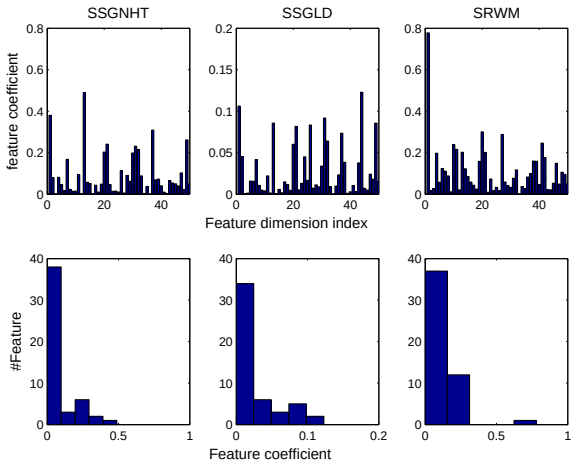


Figure 6. Feature analysis of the sparse Bayesian logistic regression on the MiniBooNE dataset. **First row**: feature coefficients; and **Second row**: feature frequency histogram.

converges to about 0.85 with 9 selected features. Our stochastic subgradient MCMC methods get the ten times faster convergence compared with the stochastic random-walk Metropolis and the stochastic augmented RJ-MCMC method.

We also analyze the features selected by the stochastic subgradient MCMC methods and the stochastic random-walk Metropolis. Figure 6 shows the feature coefficient and the standardized feature frequency histogram. We also show the top 10 features with the largest absolute weights in Table 1. We can see that the two stochastic subgradient MCMC methods produce very similar feature ranks (e.g., 8 of the top-10 features are shared), while the similarity to the random-walk method is smaller (e.g., only 4 of the top-10 features are shared between SRWM and SSGLD).

Again, we test the methods on the high dimensional realsim dataset. We set the batchsize to 10 for all the stochastic

Table 1. Selected feature indexes of sparse Bayesian logistic regression by various samplers on the MiniBooNE dataset

METHOD	SELECTED FEATURE INDEXES
SRWM	1, 4, 10, 11, 13, 20, 21, 27, 41, 42
SSGLD	1, 7, 13, 20, 21, 30, 31, 32, 37, 49
SSGNHT	1, 13, 20, 21, 26, 31, 32, 37, 44, 49

sampling methods. For the RJ-MCMC method, the two sigma variances are set as the same value of 1, and for the stochastic random walk Metropolis, the variance parameter is also set as 1. For SSGNHT, the stepsize is taken as $0.01 \times t^{-0.4}$ and the step number is set as 30. For SSGLD, the adaptive stepsize is used. Figure 5(b) presents the results of various methods with respect to the running time (in log-scale). We can see that the stochastic random walk Metropolis method has a low sampling efficiency, again due to the high dimensionality of the sample space. The stochastic RJ-MCMC suffers from the local minima, which are caused by introducing auxiliary binary variables (Korattikara et al., 2014), and we select a relatively good convergence among the local minima to report. Overall, in the high dimensional sparse learning case, our stochastic subgradient MCMC methods are faster than the chosen baseline methods by several orders of magnitude in terms of running time.

5. Conclusions

We systematically investigate the subgradient MCMC methods for a large class of Bayesian models whose posteriors involve some non-differentiable terms in the log space, such as the sparse Bayesian models with a Laplace prior and the Bayesian SVMs with a piecewise linear hinge loss. We show that by using subgradients the common properties (e.g., volume preservation and detail balance) of an ordinary Hamiltonian Monte Carlo sampler still hold. We further propose to do stochastic subgradient MCMC for

dealing with large-scale applications, with extensive empirical results on both low-dimensional and high-dimensional data. Our results demonstrate the effectiveness of the stochastic subgradient MCMC methods on improving the time efficiency and the high accuracy on drawing posterior samples.

References

Arnold, V. I, Kozlov, V. V., and Neishtadt, A. I. *Mathematical aspects of classical and celestial mechanics*, volume 3. Springer Science & Business Media, 2007.

Atkinson, K. E. *An introduction to numerical analysis*. John Wiley & Sons, 2008.

Bertsekas, D. P. *Nonlinear Programming*. Athena scientific Belmont, 1999.

Boyd, S. and Mutapcic, A. Stochastic subgradient methods. *Lecture Notes for EE364b, Stanford University*, 2008.

Chen, T., Fox, E., and Guestrin, C. Stochastic gradient Hamiltonian Monte Carlo. In *ICML*, 2014.

Chen, X., Wang, Z. J., and McKeown, M. J. A Bayesian lasso via reversible-jump MCMC. *Signal Processing*, 91 (8):1920–1932, 2011.

Chib, S. and Greenberg, E. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4): 327–335, 1995.

Ding, N., Fang, Y., Babbush, R., Chen, C., Skeel, R. D, and Neven, H. Bayesian sampling using stochastic gradient thermostats. In *NIPS*, 2014.

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

Duchi, J., Gould, S., and Koller, D. Projected subgradient methods for learning sparse gaussians. In *UAI*, 2008.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *JMLR*, 12:2121–2159, 2011.

Ghosh, J. K and Ramamoorthi, R. V. *Bayesian nonparametrics*, volume 1. Springer, 2003.

Korattikara, A., Chen, Y., and Welling, M. Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In *ICML*, 2014.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H, and Teller, E. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

Neal, R. M. MCMC using Hamiltonian dynamics. *arXiv preprint arXiv:1206.1901*, 2012.

Park, T. and Casella, G. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

Patterson, S. and Teh, Y. W. Stochastic gradient Riemannian Langevin dynamics on the probability simplex. In *NIPS*, 2013.

Polson, N. G., Scott, S. L., et al. Data augmentation for support vector machines. *Bayesian Analysis*, 6(1):1–23, 2011.

Ratliff, N. D., Bagnell, J. A., and A., Zinkevich M. (Online) subgradient methods for structured prediction. In *AISTATS*, 2007.

Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, pp. 400–407, 1951.

Roberts, G. O. and Stramer, O. Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and Computing in Applied Probability*, 4(4):337–357, 2002.

Shalev-Shwartz, S., Singer, Y., Srebro, N., and Cotter, A. Pegasos: Primal estimated subgradient solver for SVM. *Mathematical programming*, 127(1):3–30, 2011.

Teh, Y. W., Thiéry, A., and Vollmer, S. Consistency and fluctuations for stochastic gradient langevin dynamics. *arXiv preprint arXiv:1409.0578*, 2014.

Usai, M. G., Goddard, M. E., and Hayes, B. J. Lasso with cross-validation for genomic selection. *Genetics Research*, 91(06):427–436, 2009.

Vapnik, V. *The nature of statistical learning theory*. Springer Science & Business Media, 2000.

Wang, X., Peng, P., and Dunson, D. B. Median selection subset aggregation for parallel inference. In *NIPS*, 2014.

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

Williams, P. M. Bayesian regularization and pruning using a Laplace prior. *Neural Computation*, 7(1):117–143, 1995.

Zhang, T. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *ICML*, 2004.

Zhu, J., Chen, N., Perkins, H., and Zhang, B. Gibbs max-margin topic models with fast sampling algorithms. In *ICML*, 2013.

Zhu, J., Chen, J., and Hu, W. Big learning with Bayesian methods. *arXiv preprint arXiv:1411.6370*, 2014a.

Zhu, J., Chen, N., and Xing, E. P. Bayesian inference with posterior regularization and applications to infinite latent svms. *JMLR*, 15:1799–1847, 2014b.

Supplementary Material

A. Polynomial Construction of $U_\epsilon(q)$

Here we give a construction of $U_\epsilon(q)$ using the polynomial smooth:

$$U_\epsilon(q) = \begin{cases} U_0(q), \|q - q_0\| > \epsilon \\ \mathcal{P}(q), \ \|q - q_0\| \leq \epsilon \end{cases} \tag{18}$$

where the polynomial function $\mathcal{P}$ satisfies the following conditions:

$$\begin{cases} U_\epsilon(q) = U_0(q), q = q_0 \pm \epsilon, \\ \frac{\mathrm{d}U_\epsilon(q)}{\mathrm{d}q} = \mathcal{G}_q U_0(q), q = q_0 \pm \epsilon. \end{cases} \tag{19}$$

For a given $U_0(q)$, the multi-dimensional polynomial interpolation (Atkinson, 2008) ensures the existence of the $U_\epsilon(q)$ as defined in Eqn. (18).

B. Proof of the Proposition 1

From the construction of the polynomial smooth in Eqn. (18), it can simply be derived that

$$\lim_{\epsilon \to 0} U_\epsilon(q) = U_0(q). \tag{20}$$

Then, by the definitions of the ordinary Hamiltonian dynamics and the generalized Hamiltonian dynamics, we know that the series of the space-phase distribution converges to that of the generalized dynamics when ϵ goes to zero.

These show that the volume also has the same convergence:

$$\lim_{\epsilon \to 0} \frac{\mathrm{d}\rho_\epsilon}{\mathrm{d}t} = \frac{\mathrm{d}\rho_0}{\mathrm{d}t}. \tag{21}$$

As we know, the series of the Hamiltonian dynamics H_ϵ satisfies the volume preservation, i.e.,

$$\frac{\mathrm{d}\rho_\epsilon}{\mathrm{d}t} = 0, \epsilon \in (0, \infty]. \tag{22}$$

Putting Eqn. (21) and Eqn. (22) together finishes the proof of Proposition 1.

C. Proof of Theorem 2

With the proof of the properties of the generalized Hamiltonian dynamics, reversibility, conservation of the Hamiltonian and the volume preservation, the detailed balance of the HMC method can be similarly proved as in the proof for the Hamiltonian dynamics (Neal, 2012).

Another significant issue for applying the subgradient HMC is the volume preservation for the commonly used discretization integrators, e.g., the Euler method and the leapfrog method. Similarity, the volume preservation and the reversibility holds in the subgradient HMC (Neal, 2012).

D. The Mixing of the Adaptive stepsizes for SSGLD

In this part, we show that using adaptive stepsizes for SSGLD leads to a good mixing through the running mean plot illustration. A running mean plot is a plot of the iterations against the mean of the draws up to each iteration. If the running mean get converged, the Markov chain probably mixes.

Figure 7 shows the running mean plot of SSGLD using the adaptive stepsizes. We can see that the SSGLD with the adaptive stepsizes has a good mixing, and the accuracy gets converged when the chain gets mixed.

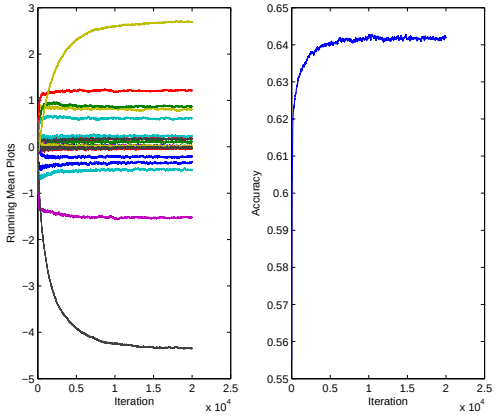
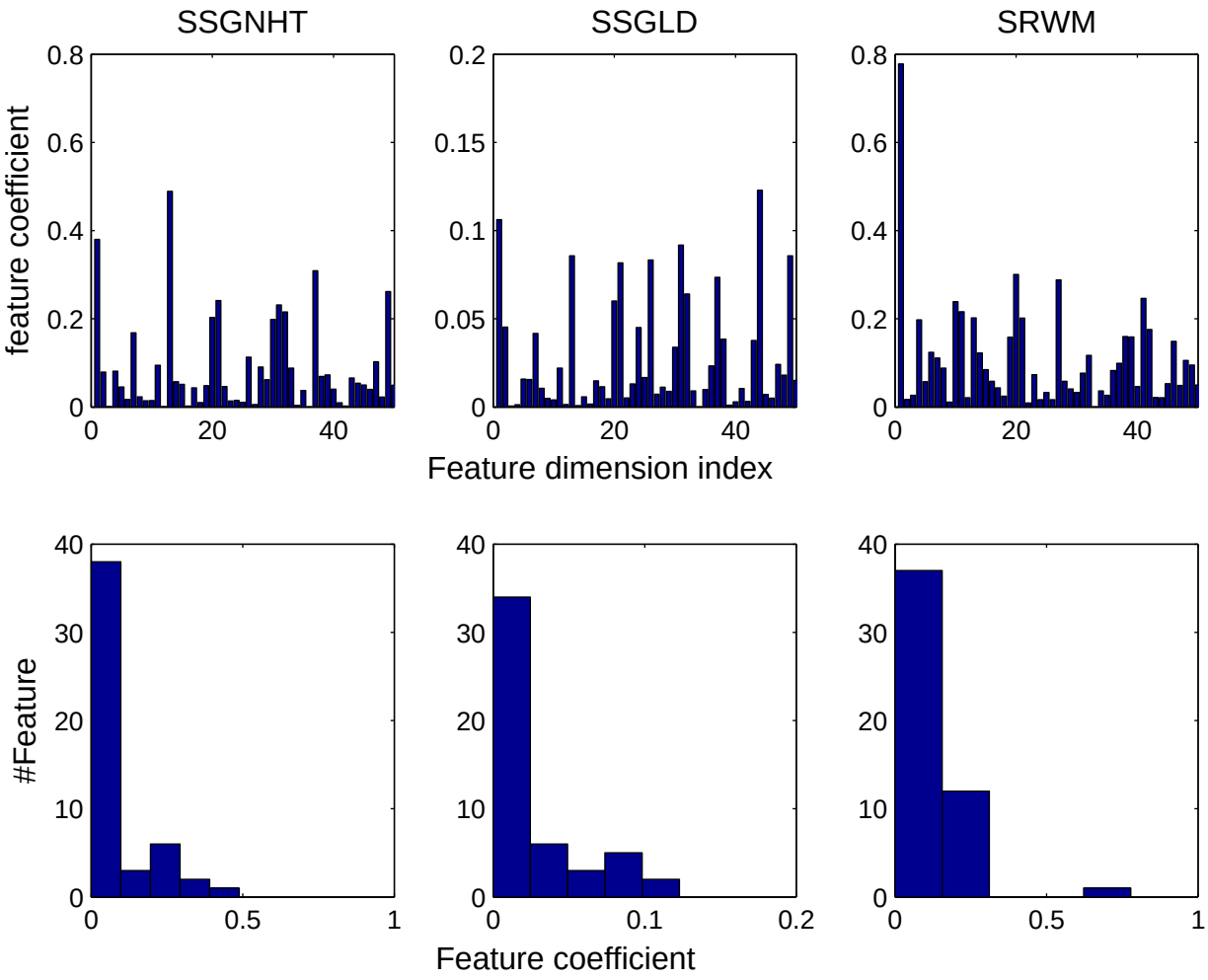


Figure 7. Performance of the SSGLD method with adaptive stepsizes on the higgs dataset, including (**Left**) running mean plot; and (**Right**) the convergence curve of testing accuracy.


SSGNHT
SSGLD
SRWM
feature coefficient
Feature dimension index
#Feature
Feature coefficient