A Survey on Stochastic Gradient Markov Chain Monte Carlo

EE381K-6 Estimation Theory

Wireless Neworking Communication Group

Kiyeon Jeon(eid: kj22295)

May 5, 2016

Abstract

Markov Chain Monte Carlo algorithms are sampling techniques, which employ proposal distribution instead of target distribution. When MCMC is used in large scale dataset, intractable time compexity problem arises. Recently, plugging stochastic gradient into MCMC has been emerging as a solution to intractable problems. Stochastic gradient MCMC(SG-MCMC) are proposed as diverse varients of stochastic gradient Langevin dynamics, stochastic gradient Hamiltonian Monte Carlo, and stochastic gradient Nose-Hoover thermostat. Three versions of SG-MCMC are representatives of SG-MCMC. Also, a theorectical framework on SG-MCMC is established with theoretical proof of convergence analysis. This survey paper provides a brief overview of MCMC, three versions of SG-MCMC, and ongoing problems to discuss.

Keyword: SG-MCMC, Langevin Dynamics, Hamiltonian Dynamics, Nose-Hoover thermostats, Stochastic Optimization

1 Introduction

A. Motivation

In recent years, there has been many new attepts to apply subsampling to Markov Chain Monte Carlo(MCMC) for large scale data. As the data size increases, subsampling technique should be introduced into regular MCMC for efficient exploration. The most commonly used regular MCMC algorithms are Metropolis-Hastings(MH) and Gibbs sampler. Those sampling algorithms have played a key role in solving high-dimensional sampling problems in polynomial time [3]. The MH algorithm made it possible to sample complicated target distribution with simple proposal distribution like Gaussian ditribution. The special case of MH is Gibbs sampling where full conditional proposal distribution is used, and the modified algorithm induced by Hamiltonian dynamics is Hamiltonian(Hybrid) Monte Carlo(HMC). [4]

To realize how to apply subsampling to MCMC, we have to understand basic physical dynamics including Langevin dynamics and Hamiltonian dynamics and molecular dynamics like Nose-Hoover thermostat. Instead of standard random walk in Gibbs sampler, the recent Stochastic Gradient MCMC(SG-MCMC) algorithms bring those dynamic systems into MCMC for efficient epxloration. Through those SG-MCMC, we can handle complicated posterior distribution. Also recent research constructs a strong theoretical framework on convergence analysis.

MCMC algorithms were introduced in [3] overall and received much attention in the recent literature [1] [7] [15]. Also, the theoretical framework was established well in [6] [10]. Somehow surprisingly these new SG-MCMC can achieve a significant reduction in sampling time as well as good quality in sampling performance, compared with existing MCMC algorithms. In this paper, we provide an overview of recent related work and ongoing problems with SG-MCMC.

B. Background: Various MCMC Algorithms

1)Metropolis-Hastings

As we mentioned earlier, MH is a representative algorithm among MCMC algorithms. Instead of sampling from target distribution directly, we can sample from proposal distribution, typically Gaussian distribution, designed by user. Sampling quality depends on the variance of this designed proposal distribution. The reason why we can sample target distribution with MH is that we can get a stationary distribution of target distribution by setting accept/reject stage after sampling from proposal distribution. This stage made MH satisfy detailed balance condition on Markov Chain. Aperiodicity condition holds because rejection state means self-loop, and irreducibility also holds since we assume that the support of proposal distribution includes that of target distribution. By these three satisfied conditions, Discrete Time Markov Chain(DTMC) have a unique stationary distribution.

2)Gibbs sampler and Hamiltonian(Hybrid) Monte Carlo

The special case of MH is the Gibbs sampler which uses full conditionals as proposal distribution on MH. The accept/reject stage is ignored by using full conditionals as proposal because the accept/reject stage is always satisfied. Although Gibbs sampler is a simple MCMC algorithm, standard random walk of Gibbs sampler having full conditionals except each coordinate sequentially is inefficient compared to physical dynamics of Hamitonian Monte Carlo(HMC) having auxilary momentum variable. Although it seems like complicated to understand HMC algorithm that is a solution of stochastic differential equation(SDE) from hamiltonian dynamics, it is more efficient than Gibbs sampler.

Three versions of SG-MCMC have been considered in the literature: Stochastic Gradient Langevin Dynamics(SGLD), Stochastic Gradient Hamiltonian Monte Carlo(SGHMC), Stochastic Gradient Nose-Hoover thermostat(SGNHT). Before we cover these algorithms, we need to clarify how posterior sampling is helpful for Bayesian learning.

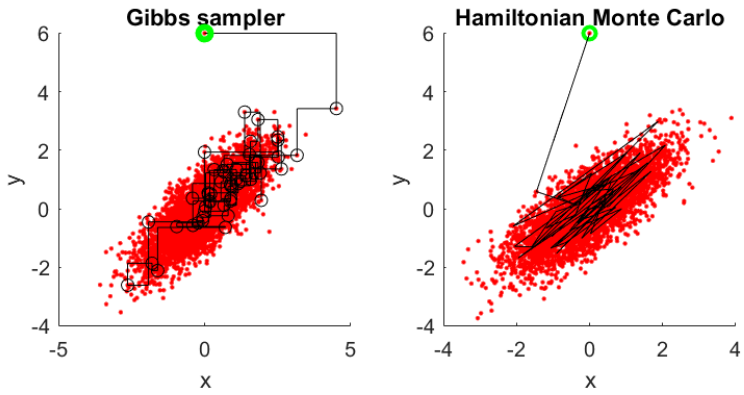

Figure 1: Gibbs sampler and Hamiltonian Monte Carlo sampling with 5000 points from two dimensional normal distribution

2 Bayesian Learning via Posterior distribution

If we know posterior distribution like $p(\theta|\mathcal{D})$ where $\mathcal{D}$ is dataset $\{x_1, x_2, \ldots, x_N\}$ and $x_i \in \mathbb{R}^n$, we can get $\theta^* = \arg\max_\theta p(\theta|\mathcal{D})$ through an optimization method like gradient ascent. Let's think about classification problem with logistic regression. We have feature dataset $X = \{x_1, x_2, \ldots, x_N\}$ where $x_i \in \mathbb{R}^n$ and labels $Y = \{y_1, y_2, \ldots, y_N\}$ where $y_i \in \{-1, +1\}$ Here, since we can assume θ, X are independent, $p(\theta|X, Y) \propto p(Y|X, \theta)p(\theta)$ (Bayes Theorem) Since each data instance is independent and log function is monotone increasing,

$$\begin{aligned} \theta^* &= \arg\max_\theta \left(\prod_{i=1}^{N} p(y_i|x_i, \theta)\right) \cdot p(\theta) \\ &= \arg\max_\theta \left(\sum_{i=1}^{N} \log p(y_i|x_i, \theta) + \log p(\theta)\right) \end{aligned} \tag{1}$$

By using gradient asent to this objective function, we can get the update equation:

$$\theta_{t+1} \leftarrow \theta_t + \epsilon(\nabla \log p(\theta) + \sum_{i=1}^{N} \nabla \log p(x_i|\theta)) \tag{2}$$

Based on this update equation, we can get the posterior distribution samples by adding Gaussian noise to the update. This technique is called Langevin Dynamics and we deal with it in the next chpater.

Posterior distribution sampling does not mean that we can get the global optimum for maxium a posteriori. In the case of Gibbs sampler, we know efficient algorithm getting global optimum through simulated annealing algorithm where the algorithm utilizes varaint prior distribution, $p^{1/T}(x)$, instead of original prior distribution, $p(x)$ [3]. The posterior distribution sampling that we dealt with in this paper can be enhanced by such an simulated annealing idea [5]

3 Model I : SGLD

As shown in [15], the SGLD is the first order Langevin dynamics solution with subsampling and Gaussian noise as followings:

$$\theta_{t+1} \leftarrow \theta_t + \epsilon_t(\nabla \log p(\theta_t) + \frac{N}{n} \sum_{i=1}^{n} \nabla \log p(x_{ti}|\theta_t)) + \eta_t \text{ where } \eta_t \sim \mathcal{N}(0, \epsilon_t) \tag{3}$$

where the stepsize should satisfy the following property to guarantee convergence to a local maximum [13]:

$$\sum_{t=1}^{\infty} \epsilon_t = \infty, \ \sum_{t=1}^{\infty} \epsilon_t^2 < \infty \tag{4}$$

[15] gives an intuitive argument that the above update equation becomes samplings from posterior distribution. The solid proof is shown in [14], where SGLD algorithm has a weak convergence not strong convergence so that invaraint distribution from SGLD converges to the posterior distribution.

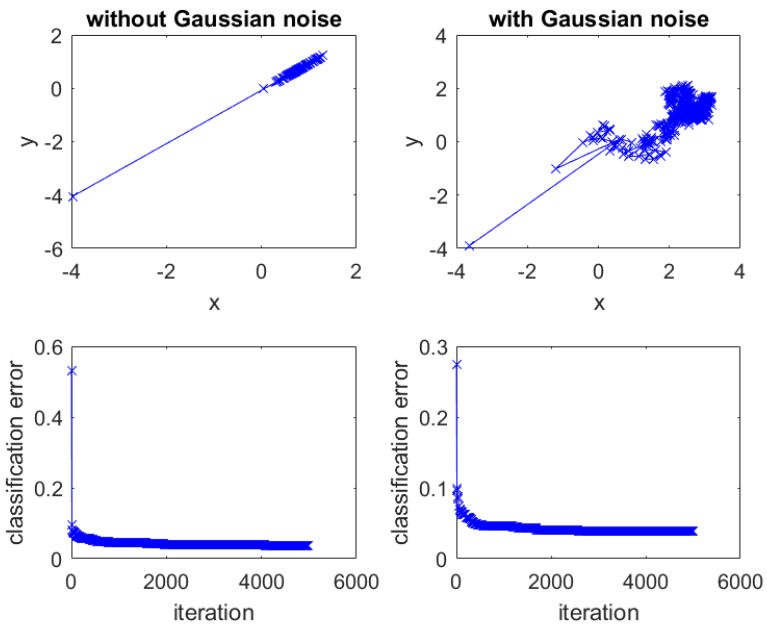

Figure 2: Illustration of stocahstic gradient Langevin dynamics and classification test error

In the Figure 2, I experiment update equations with and without Gaussian noise term about 7 and 9 digits of MNIST data. The above two plots show the movement of two selected coordinates among a lot of features of trained parameter. Only the right plot with Gaussian noise becomes posterior distribution sampling.

Contribution

SGLD is the first posterior sampling algorithm by introducing Langevin dynamics into stochastic optimization so that we can get the posterior distribution efficiently. The SGLD algorithm's drawback is that as stepsize decreases the mixing rate slows down. We can supplement this drawback by setting stepsize constant after the proper number of iteration.

The advanced algorithm of SGLD is covered in Section 2-A [1], where they use normal approximation of posterior distrubtion with preconditioned matrix when stepsize is large. In Section 2-B, we discuss varaints of SGLD corresponding to distributed SGLD.

A. Stochastic Gradient Fisher Scoring(SGFS)

SGFS uses different approach related to EQ (3) where SGLD introduces Gaussian noise to stochastic optimization [1]. The Bayesian asymptotic theory(Bayesian Central Limit Theorem) is as followings:

Theorem 3.1 (Bayesian Central Limit Theorem [11]). *Suppose $\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$ are independent observations, and prior on θ is $p(\theta)$. We assume that the posterior is proper, its mode exists, and the regularity condition,$\mathbb{E}[\nabla \log p(\theta|\mathcal{D})] = 0$, holds. Then,*

$$p(\theta|\mathcal{D}) \xrightarrow[N\to\infty]{} \mathcal{N}(\theta_0, \mathcal{I}_N^{-1}) \tag{5}$$

where θ_0 is posterior mode i.e. a solution of $\frac{\partial \log p(\theta|\mathcal{D})}{\partial \theta_i} = 0$ and $\mathcal{I}_N$ is Fisher information i.e. $\mathcal{I}_N = N\mathcal{I}(\theta_0) = N\mathbb{E}[\nabla \log p(\theta|\mathcal{D}) \cdot \nabla \log p(\theta|\mathcal{D})^T] = -N\mathbb{E}[\nabla^2 \log p(\theta|\mathcal{D})]$

SGFS uses this Bayesian Central Limit Theorem(BCLT) and preconditioned matrix on stepsize. BCLT implies $\nabla \log p(\theta|\mathcal{D}) \approx -\mathcal{I}_N(\theta - \theta_0)$ Therefore, we can make the following update equation:

$$\theta_{t+1} \leftarrow \frac{\epsilon C}{2}\{-\mathcal{I}_N(\theta_t - \theta_0)\} + w \quad \text{where } w \sim \mathcal{N}(0, \epsilon C - \frac{\epsilon^2}{4} C\mathcal{I}_N C) \tag{6}$$

If θ_t has a distribution $\mathcal{N}(\theta_0, I_N^{-1})$, we can prove that $\theta_i (i \geq t)$, has a stationary distribution $\mathcal{N}(\theta_0, I_N^{-1})$. Utilizing this approximate posterior concept, we can make the following update rules [1]:

$$\begin{aligned} \theta_{t+1} \quad &\leftarrow \theta_t + 2[\gamma\mathcal{I}_N + \frac{4B}{\epsilon}]^{-1} \cdot \{\nabla \log p(\theta_t) + N\bar{g}_n(\theta_t; X_n^t) + \eta\} \text{ where } \eta \sim \mathcal{N}(0, \frac{4B}{C}) \\ \hat{\mathcal{I}}_{1,t} \quad &= (1 - \kappa_t)\hat{\mathcal{I}}_{1,t-1} + \kappa V(\theta_t) \end{aligned} \tag{7}$$

where $V(\theta_t)$ is an empirical Fisher information and $\kappa_t = 1/t$ and B is any symmetric positive-definite matrix

Contribution

SGFS is an enhanced posterior sampling method which runs in high mixing rate with large stepsize converging to normal approximate posterior and operates like SGLD with small stepsize by preconditioned matrix.

B. Distributed Stochastic Gradient MCMC(D-SGLD)

There is a research introducing parallerization into SGLD. For the distributed environment, they propose trajectory sampling for short-communication-cycle problem, adaptive load balancing for balancing the workloads, and variance reduction gradient estimator for enhancing estimator's quality. They verified that D-SGLD outwighs SGLD on convergence time by some experiments. [2]

4 Model II : SGHMC

SGHMC is fundamentally based on a HMC algorithm. HMC algorithm makes use of Hamiltonian system with a newly introduced 'momentum′ variable to achieve stationary distribution. The Hamiltonian systems is:

$$\begin{aligned} \frac{d\theta_i}{dt} &= \quad \frac{\partial H}{\partial r_i} = \quad \frac{\partial K(r)}{\partial r_i} \Rightarrow \quad \frac{\partial}{\partial t}\theta = \quad \nabla_r K(r) = r \ (K(r) = \frac{1}{2} r^T M^{-1} r \text{ where } M = \mathcal{I}) \\ \frac{dr_i}{dt} &= - \quad \frac{\partial H}{\partial \theta_i} = \quad -\frac{\partial U(\theta)}{\partial \theta_i} \Rightarrow \quad \frac{\partial}{\partial t} r = \quad -\nabla_\theta U(\theta) \end{aligned} \tag{8}$$

Here, K is a kinetic energy, U is a potential energy and $H = K + U$ is a total energy. We use posterior distribution by potential energy: $p(\theta|\mathcal{D}) \propto \exp(-U(\theta))$ From Bayes' theory, $p(\theta|\mathcal{D}) \propto p(\theta) \prod_{x \in \mathcal{D}} p(x|\theta) = \exp(\log p(\theta) + \sum_{x \in \mathcal{D}} \log p(x|\theta))$ Thus, we can set our potential energy, $U(\theta) = \log p(\theta) + \sum_{x \in \mathcal{D}} \log p(x|\theta)$ where we just ignore the proportional constant. For reduce the computation cost of data-driven potential energy, we apply subsampling dataset to potential energy, $\tilde{U}(\theta) = \log p(\theta) + \frac{|\mathcal{D}|}{|\tilde{\mathcal{D}}|} \sum_{x \in \tilde{\mathcal{D}}} \log p(x|\theta)$ where $\tilde{\mathcal{D}} \subset \mathcal{D}$. We assume that

all observations are independent and Bayesian Central Limit Theorem(BCLT) holds. BCLT implies $\theta \sim \mathcal{N}(\theta_0, \mathcal{I}_n)$ Thus, $\nabla\tilde{U}(\theta) \xrightarrow[|\mathcal{D}|\to\infty]{d} \mathcal{N}(\nabla U(\theta), V(\theta))$ where $V(\theta) = \frac{1}{n}\mathcal{I}_n^{-1}$. So we just use $\nabla\tilde{U}(\theta) \approx \nabla U(\theta) + \mathcal{N}(0, V(\theta))$

EQ (8) can be translated as an discretized version of the above stochastic differential equation(SDE):

$$\begin{cases} \theta_t \leftarrow \theta_{t-1} + \epsilon M^{-1} r_{t-1} \\ r_t \leftarrow r_{t-1} - \epsilon\nabla\tilde{U}(\theta_t) \approx r_{t-1} - \epsilon\nabla U(\theta_t) + \mathcal{N}(0, \epsilon^2 V(\theta_t)) \end{cases}$$

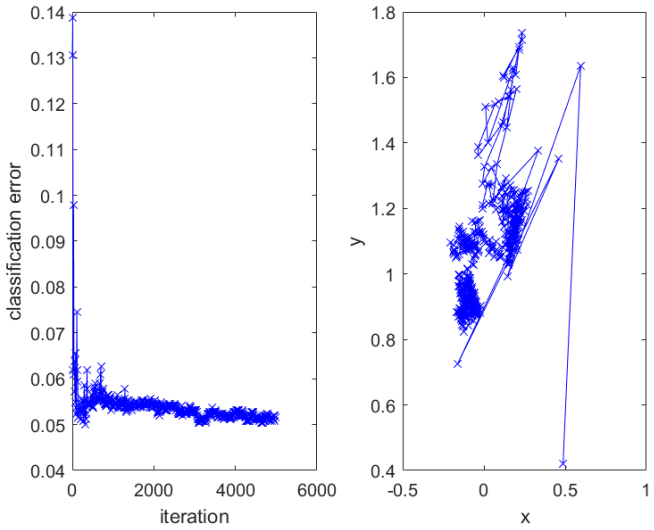

Figure 3: Illustration of stochastic gradient Hamiotonian Monte Carlo and classification error with MNIST dataset

However, according to [9], if we just apply subsampling on potential energy, the HMC process does not have stationary distribution any more. This subsampling potential energy have to be operated with another friction term in order to have invariant distribution. The new update equation with friction term is:

$$\begin{cases} \theta_t \leftarrow \theta_{t-1} + \epsilon r_{t-1} \\ r_t \leftarrow r_{t-1} - \epsilon\nabla U(\theta_t) - \frac{1}{2}\epsilon^2 V(\theta_t) M^{-1} r_{t-1} + \mathcal{N}(0, \epsilon^2 V(\theta_t)) \end{cases} \tag{9}$$

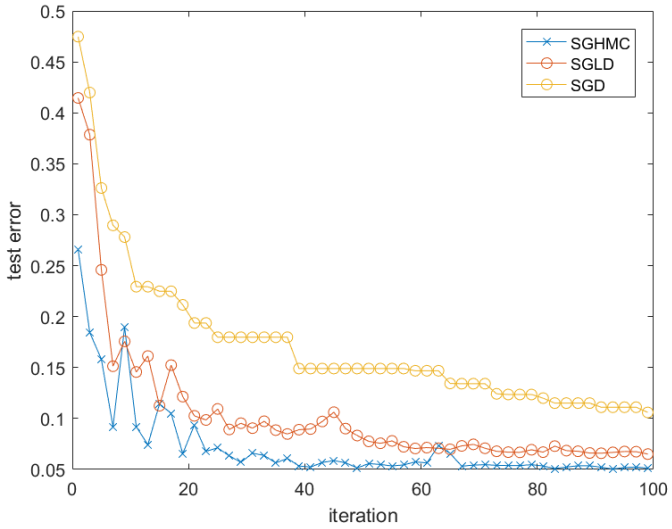

Figure 4: logistic regression's classification error with 4 and 9 digits of MNIST using SGD, SGLD, and SGHMC (Although posterior distribution sampling does not mean a global optimum satisfying maximum a posteriori exactly, we roughly estimate $\hat{\theta}$ which maximize posterior distribution through posterior distribution sampling)

Contribution

SGHMC is the extended version of SGLD with physical dynamics' behavior instead random walk-like behavior. In the Figure 4, we can know that both SGLD and SGHMC outweigh stochastic optimization method, and SGHMC has better performance than SGLD.

5 Model III : SGNHT

The previous two models, SGLD and SGHMC, introduce unknown noise to their stochastic algorithm so that we have to estimate the noise with Fisher information. To alleviate this estimating problem, we introduce another additional variable to SGHMC algorithm for stabilizing momentum fluctuation came from subsampling.

The state probability of canonical ensemble has canonical distribution $p(\theta, r) \propto \exp(-H(\theta, r)/(k_B T))$ where K_B is Boltzmann constant. The system temperature is expressed by the mean kinetic

energy satisfying the following thermal equilibrium condition:

$$\frac{k_B T}{2} = \frac{1}{n}\mathbb{E}[K(r)] \Leftrightarrow k_B T = \frac{1}{n}\mathbb{E}[r^T r] \tag{10}$$

SGNHT introduce an additional variable ξ from the above thermostat property with $k_B T = 1$ to SGHMC as followings [8]:

$$\begin{cases} r_t \leftarrow r_{t-1} - \epsilon \nabla U(\theta_{t-1}) - \xi_{t-1}\epsilon r_{t-1} + \sqrt{2A}\mathcal{N}(0, \epsilon) \\ \theta_t \leftarrow \theta_{t-1} + \epsilon r_t \\ \xi_t \leftarrow \xi_{t-1} + \frac{1}{n}(r_t^T r_t - 1)\epsilon \end{cases} \tag{11}$$

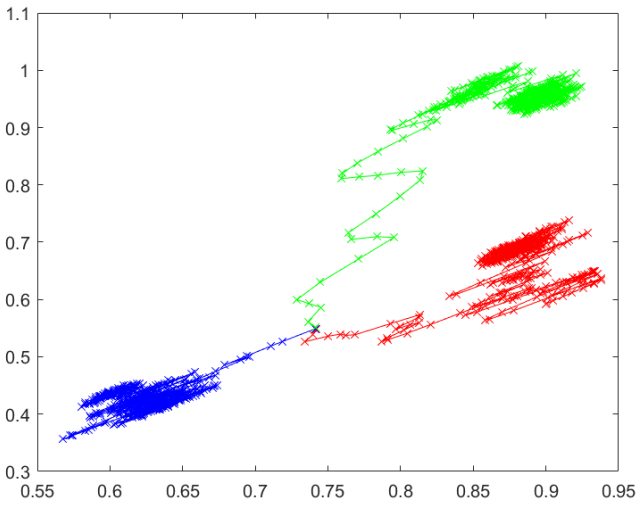

Figure 5: Illustration of stochastic gradient Nose Hoover thermostat with the same initial point and different path to stationary distribution with MNIST dataset

Constribution

SGNHT is the method to deal with noise term derived by stochastic gardient through termostat dynamics. In the SGHMC algorithm, we use Bayesian CLT and empirical Fisher

information to process the noise term for stationary distribution.

6 Discussion and Conclusions

We provided an overview of recent results about MCMC with posterior distribution sampling. Three versions of SG-MCMC are considered: SGLD, SGHMC, SGHNT. The SGLD is in essence an algorithm connecting MCMC to stochastic optimization. As we see before, there is a computational benefit by applying stochastic gradient to MCMC. The slow mixing rate with small stepsize was enhanced by SGFS, and through the continued research like D-SGLD the SGLD have been developed.

The SGHMC is related to the first order Langevin dynamics of SGLD. The SGHMC can be considered as second order Langevin dynamics having a friction term. In practice, the time coplexity of SGHMC is the same as that of SGLD [7]. As the stepsize decreases, SGLD achieves low estimation error with high autocorrelation time, which means a bad sampler. In contrast, SGHMC has low estimation error as weel as low autocorrelation time, thus concluding that distribution is efficiently explored by the sampler. In other words, SGLD makes it hard to explore distribution, but SGHMC enables the sampler to operate well following distrubtion.

The most advanced and complicated algorithm among three models is SGHNT. This model introduced another term related to thermostat dynmaics into Hamiltonian dynamics in order to alleviate difficulty with estimating the noise term from subsampling in likelihood function. From the experiment in [8], SGNHT is more stable than SGHMC when discretization step is large.

In addition to these three models on SG-MCMC, there are well-established theoretical framework in [6] [12]. A general recipe regarding to SG-MCMC samplers based on continuous Markov processes was given in [12]. The recipe proposes a new state-adaptive sampler: stochastic gradient Riemann Hamiltonian Monte Carlo(SGRHMC) and the benefits of

SGRHMC is verified with simulated data. [6] proposes theory to analyze finite-time ergodic errors and aymptotic invariant meaures of SG-MCMCs with high-order integrators. This theory indicates that the convergence rate of an SG-MCMC is able to theoretically approach the standard MCMC convergence rate.

Also, there is an attempt to investigate into relations between SG-MCMC and stochastic optimization [6] The gap is explored by introducing simulated annealing into SG-MCMC. The annealed SG-MCMC, called Santa in [6], can analyze parameter space in an efficient way and find the global optima closely and give a theory framework related to convergence analyis of the annealed SG-MCMC algorithm about an objective function.

References

[1] 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 (ICML-12)*, pages 1591–1598, 2012.

[2] Sungjin Ahn, Babak Shahbaba, Max Welling, et al. Distributed stochastic gradient mcmc. In *JMLR Workshop and Conference Proceedings*, number 32, pages 1044–1052. JMLR, 2014.

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

[4] George Casella and Edward I George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

[5] Changyou Chen, David Carlson, Zhe Gan, Chunyuan Li, and Lawrence Carin. Bridging the gap between stochastic gradient mcmc and stochastic optimization. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics(AISTATS 2016)*, pages 1051–1060.

[6] 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 2269–2277, 2015.

[7] Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1683–1691, 2014.

[8] 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.

[9] 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.

[10] G. Grimmett and D. Stirzaker. *Probability and random processes*. Oxford science publications. Clarendon Press, 1985.

[11] Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.

[12] Yi-An Ma, Tianqi Chen, and Emily Fox. A complete recipe for stochastic gradient mcmc. In *Advances in Neural Information Processing Systems*, pages 2899–2907, 2015.

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

[14] Issei Sato and Hiroshi Nakagawa. Approximation analysis of stochastic gradient langevin dynamics by using fokker-planck equation and ito process. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 982–990, 2014.

[15] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688, 2011.