On the Convergence of Stochastic Gradient MCMC with High-Order Integrators

Changyou Chen[†], Nan Ding[‡] and Lawrence Carin[†]

[†]Duke University, Durham NC 27708, USA
[‡]Google Inc., Venice, CA, USA

Introduction

Large-scale Bayesian learning becomes increasingly popular due to the necessity of processing big data.

Contributions:

- Develop theory to analyze convergence properties of general stochastic gradient MCMC (SG-MCMC) algorithms.
- Propose a more accurate 2nd-order integrator for SG-MCMC, with faster convergence rates.
- Experiments on both synthetic data and large-scale applications demonstrate the proposed theory.

Example SG-MCMC Algorithm

Setting: Given data $\mathbf{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$, a generative model $p(\mathbf{X}|\boldsymbol{\theta}) = \prod_{i=1}^N p(\boldsymbol{x}_i|\boldsymbol{\theta})$ with model parameter $\boldsymbol{\theta}$, and prior $p(\boldsymbol{\theta})$, we want to compute the posterior:

$$\pi(\boldsymbol{\theta}) \triangleq p(\boldsymbol{\theta}|\mathbf{X}) \propto p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \triangleq e^{-U(\boldsymbol{\theta})} ,$$

where $U(\boldsymbol{\theta})$ is called the potential energy.

Stochastic gradient Hamiltonian Monte Carlo (SGHMC):

- Conventional MCMC algorithms require processing the whole data in each iteration, which is computationally prohibited in big data setting.
- SG-MCMC algorithms overcome this problem by using a minibatch of the data in each iteration.

The SGHMC is based on the 2nd-order Langevin dynamic defined as:

$$\begin{cases} \mathrm{d}\boldsymbol{\theta} = \boldsymbol{p}\mathrm{d}t \\ \mathrm{d}\boldsymbol{p} = -\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})\mathrm{d}t - D\boldsymbol{p}\mathrm{d}t + \sqrt{2D}\mathrm{d}\mathcal{W} , \end{cases} \quad (1)$$

where $\boldsymbol{p}$ is the augmented momentum variable, $\mathcal{W}$ is the standard Brownian motion, t is the time, and D is a constant.

According to the Fokker-Planck equation, the equilibrium distribution of (1) is:

$$P(\boldsymbol{\theta}, \boldsymbol{p}) \propto e^{-U(\boldsymbol{\theta}) + \frac{\boldsymbol{p}^T\boldsymbol{p}}{2}} .$$

To generate approximate samples from (1), we use Algorithm 1 by discretizing (1) and using stochastic gradients.

Algorithm 1 Stochastic Gradient Hamiltonian Monte Carlo

Input: Parameters h, D.
Initialize $\boldsymbol{\theta}_0 \in \mathbb{R}^n$, $\boldsymbol{p}_0 \sim \mathcal{N}(0, \mathbf{I})$.
for $l = 1, 2, \ldots$ **do**
  Evaluate stochastic gradient $\nabla \tilde{U}_l(\boldsymbol{\theta}_{(l-1)h})$ from the l-th minibatch.
  $\boldsymbol{p}_{lh} = \boldsymbol{p}_{(l-1)h} - D\boldsymbol{p}_{(l-1)h}h - \nabla \tilde{U}_l(\boldsymbol{\theta}_{(l-1)h})h + \sqrt{2Dh}\mathcal{N}(0, \mathbf{I})$.
  $\boldsymbol{\theta}_{lh} = \boldsymbol{\theta}_{(l-1)h} + \boldsymbol{p}_{lh}h$.
end for

Convergence of SG-MCMC

Priliminary: Given an ergodic stochastic differential equation such as (1), with an invariant measure $\rho(\boldsymbol{x})$. In Bayesian learning, we are interested in the posterior average for some test function $\phi(\boldsymbol{x})$:

$$\bar{\phi} \triangleq \int_{\mathcal{X}} \phi(\boldsymbol{x})\rho(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

For a given SG-MCMC algorithm with generated samples $(\boldsymbol{x}_{lh})_{l=1}^L$, we use the *sample average* $\hat{\phi}$ to approximate $\bar{\phi}$, defined as

$$\hat{\phi} = \frac{1}{L}\sum_{l=1}^L \phi(\boldsymbol{x}_{lh}) \approx \bar{\phi} .$$

Order of integrators: When solving the discretized SDE such as in Algorithm 1, the samples are generated from numerical integrators, *e.g.*, the Euler integrator in Algorithm 1.

An integrator is said to be a Kth-order local integrator if for any smooth and bounded function f, the following holds:

$$\mathbb{E}f(\boldsymbol{x}) = e^{h\mathcal{L}}f(\boldsymbol{x}) + O(h^{K+1}) , \quad (2)$$

where $\mathcal{L}$ is the generator of the corresponding SDE, and the expectation is taken over the distribution of $\boldsymbol{x}$.

Theorem (SG-MCMC with fixed step sizes)

Let $\|\cdot\|$ be the operator norm. Under certain assumptions, the bias and MSE of an SG-MCMC with a Kth-order integrator at time $T = hL$ can be bounded, for some constants C_1 and C_2, as:

$$\text{Bias: } \left|\mathbb{E}\hat{\phi} - \bar{\phi}\right| \leq C_1\left(\frac{1}{Lh} + \frac{\sum_l \|\mathbb{E}\Delta V_l\|}{L} + h^K\right)$$

$$\text{MSE: } \mathbb{E}\left(\hat{\phi} - \bar{\phi}\right)^2 \leq C_2\left(\frac{\frac{1}{L}\sum_l \mathbb{E}\|\Delta V_l\|^2}{L} + \frac{1}{Lh} + h^{2K}\right) ,$$

where ΔV_l characterizes the error introduced by stochastic gradients in the l-th minibatch, e.g., in SGHMC, $\Delta V_l = (\nabla_{\theta}\tilde{U}_l - \nabla_{\theta}U) \cdot \nabla_p$.

Theorem (Decreasing step sizes)

Under certain assumptions, the bias and MSE of an SG-MCMC with a Kth-order integrator at time $S_L = \sum_{l=1}^L h_l$ can be bounded, for some constants C_1 and C_2, as:

$$\text{Bias: } \left|\mathbb{E}\tilde{\phi} - \bar{\phi}\right| \leq C_1\left(\frac{1}{S_L} + \frac{\sum_{l=1}^L h_l^{K+1}}{S_L}\right)$$

$$\text{MSE: } \mathbb{E}\left(\tilde{\phi} - \bar{\phi}\right)^2 \leq C_2\left(\sum_l \frac{h_l^2}{S_L^2}\mathbb{E}\|\Delta V_l\|^2 + \frac{1}{S_L} + \frac{(\sum_{l=1}^L h_l^{K+1})^2}{S_L^2}\right) .$$

Optimal convergence rates: $L^{-K/(K+1)}$ for the bias, $L^{-2K/(2K+1)}$ for the MSE.

Acknowledgements

This research was supported by ARO, DARPA, DOE, NGA and ONR.

Symmetric Splitting Integrators

The idea is to split the unfeasible SDE into several sub-SDEs, such that all the sub-SDEs are analytically solvable. Samples are then generated by sequentially evolving through these sub-SDEs.
For the SGHMC, (1) is split into

$$A: \begin{cases} \mathrm{d}\boldsymbol{\theta} = \boldsymbol{p}\mathrm{d}t \\ \mathrm{d}\boldsymbol{p} = 0 \end{cases}, B: \begin{cases} \mathrm{d}\boldsymbol{\theta} = 0 \\ \mathrm{d}\boldsymbol{p} = -D\boldsymbol{p}\mathrm{d}t \end{cases}, O: \begin{cases} \mathrm{d}\boldsymbol{\theta} = 0 \\ \mathrm{d}\boldsymbol{p} = -\nabla_{\boldsymbol{\theta}}U(\boldsymbol{\theta})\mathrm{d}t + \sqrt{2D}\mathrm{d}W \end{cases}$$

The corresponding updates for $\boldsymbol{x}_{lh} = (\boldsymbol{\theta}_{lh}, \boldsymbol{p}_{lh})$ consist of the following 5 steps:

$$\boldsymbol{\theta}_{lh}^{(1)} \overset{A}{=} \boldsymbol{\theta}_{(l-1)h} + \boldsymbol{p}_{(l-1)h}h/2 \Rightarrow \boldsymbol{p}_{lh}^{(1)} \overset{B}{=} e^{-Dh/2}\boldsymbol{p}_{(l-1)h}$$
$$\Rightarrow \boldsymbol{p}_{lh}^{(2)} \overset{O}{=} \boldsymbol{p}_{lh}^{(1)} - \nabla_{\boldsymbol{\theta}}\tilde{U}_l(\boldsymbol{\theta}_{lh}^{(1)})h + \sqrt{2Dh}\zeta_l$$
$$\Rightarrow \boldsymbol{p}_{lh} \overset{B}{=} e^{-Dh/2}\boldsymbol{p}_{lh}^{(2)} \Rightarrow \boldsymbol{\theta}_{lh} \overset{A}{=} \boldsymbol{\theta}_{lh}^{(1)} + \boldsymbol{p}_{lh}h/2 ,$$

where $(\boldsymbol{\theta}_{lh}^{(1)}, \boldsymbol{p}_{lh}^{(1)}, \boldsymbol{p}_{lh}^{(2)})$ are intermediate variables.

Experiments

I. Synthetic data: We consider a standard Gaussian model where $x_i \sim \mathcal{N}(\theta, 1), \theta \sim \mathcal{N}(0, 1)$, with 1000 data samples, minibatch size 10, and test function $\phi(\theta) \triangleq \theta^2$.

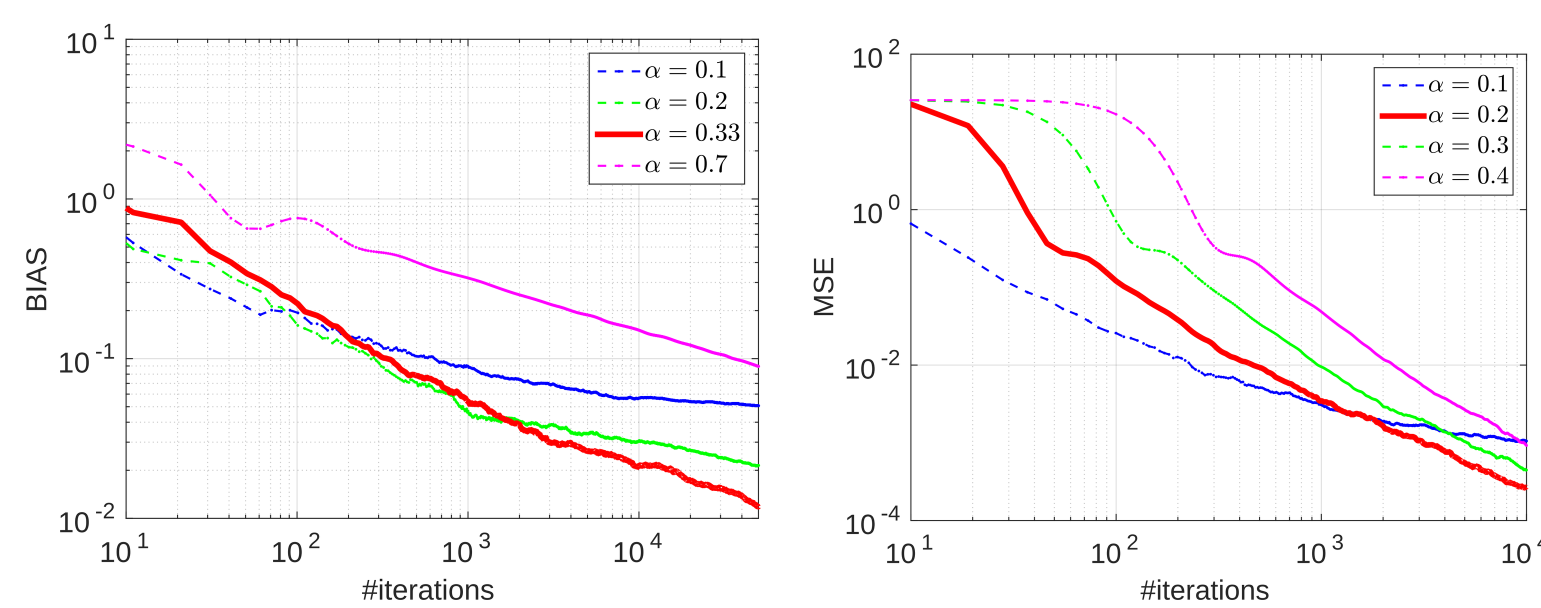

Figure : Bias of SGHMC-D (left) and MSE of SGHMC-F (right). Solid red curves correspond to theoretical optimal rates.

II. Large-scale applications: 1) Latent Dirichlet allocation model (LDA) on 10M Wikipedia data; standard test perplexity is calculated; 2) Sigmoid belief network (SBN) on the MNIST dataset; test likelihood is calculated.

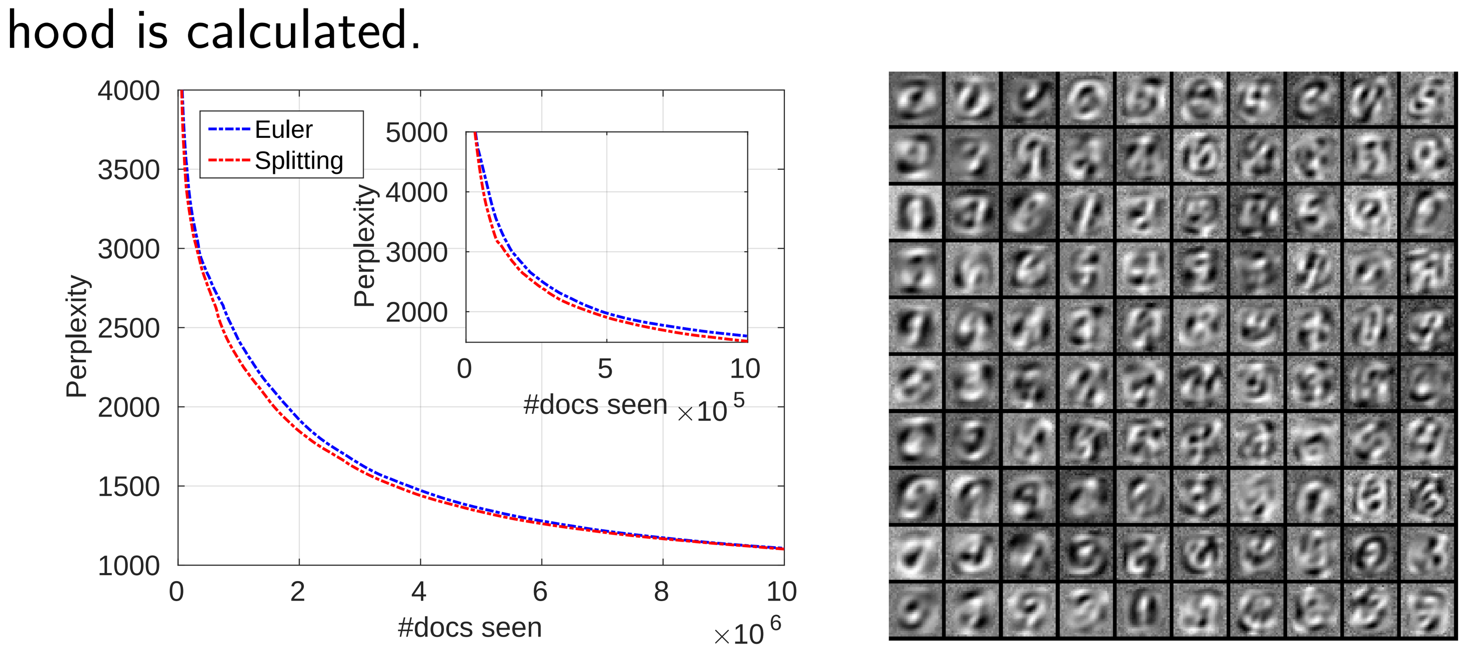

Figure : Test perplexity on LDA (left) and test likelihood on SBN (right).