1st Symposium on Advances in Approximate Bayesian Inference, 2018 1–6

Parallel-tempered Stochastic Gradient Hamiltonian Monte Carlo for Approximate Multimodal Posterior Sampling

Rui Luo* RUI.LUO@AIG.COM
Qiang Zhang* QIANG.ZHANG@AIG.COM
Yaodong Yang YAODONG.YANG@AIG.COM
Yuanyuan Liu YUANYUAN.LIU@AIG.COM
American International Group, Inc.

Abstract

We propose a new sampler that integrates the protocol of parallel tempering with the Nosé-Hoover (NH) dynamics. The proposed method can efficiently draw representative samples from complex posterior distributions with multiple isolated modes in the presence of noise arising from stochastic gradient. It potentially facilitates deep Bayesian learning on large datasets where complex multi-modal posteriors and mini-batch gradient are encountered.

1. Introduction

In Bayesian inference, one of the fundamental problems is to efficiently draw *i.i.d.* samples from the posterior distribution $\pi(\theta|\mathcal{D})$ given the dataset $\mathcal{D} = \{x\}$, where $\theta \in \mathbb{R}^D$ denotes the variable of interest. Provided the prior distribution $\pi(\theta)$ and the likelihood per datum $\ell(\theta; x)$, the posterior to be sampled can be formulated as

$$\pi(\theta|\mathcal{D}) = \pi(\theta) \prod_{x\in\mathcal{D}} \ell(\theta; x). \tag{1}$$

To facilitate posterior sampling, the framework of Markov chain Monte Carlo (MCMC) has been established, which has initiated a broad family of methods that generate Markov chains to propose new sample candidates and then apply tests of acceptance in order to guarantee the condition of detailed balance. Methods like the Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953; Hastings, 1970), the Gibbs sampler (Geman and Geman, 1984), and the hybrid/Hamiltonian Monte Carlo (HMC) (Duane et al., 1987; Neal, 2011) are famous representatives for the MCMC family where different generating procedures of Markov chains are adopted; each of those methods has achieved great success on various tasks in statistics and related fields.

Among MCMC methods, HMC, in particular, has attracted attention due to its exploitation of gradient information. In a typical HMC setting (Neal, 2011), the target posterior distribution $\pi(\theta|\mathcal{D})$ is embedded into a virtual physical system fixed at the standard temperature $T = 1$ with the potential energy defined in the form of

$$U(\theta) = -\log \pi(\theta|\mathcal{D}) = -\log \pi(\theta) - \sum_{x\in\mathcal{D}} \log \ell(\theta; x) - \text{const}\,. \tag{2}$$

The variable of interest θ is interpreted as the position of the system in the phase space; an auxiliary variable $p \in \mathbb{R}^D$ is then introduced as the conjugate momentum corresponding to the kinetic energy $p^\top M^{-1} p/2$. By defining the total energy, i.e. the Hamiltonian, as the sum of the potential and kinetic

* Equal

energy, the Hamiltonian dynamics that governs the physical system can therefore be derived from the Hamilton's formalism. From the perspective of sampling, new sample candidates are proposed via simulating the Hamiltonian dynamics, where the gradient of potential $\nabla U(\theta)$ is utilized.

Despite possessing numerous advantages against its alternatives within the MCMC family, HMC still suffers, however, from two major issues: 1. gradient noise arising from mini-batches may lead to a severe deviation of the dynamics from the desired orbit; 2. isolated modes may not be correctly sampled or even left undiscovered. Unfortunately, as one deals with deep neural networks training on large datasets, those two problems arise simultaneously: deep neural networks leads to complex posterior distributions for the parameters, which may contain numbers of isolated modes; efficient training on large datasets requires mini-batching, the gradient hence would be quite noisy as is evaluated on a small fraction of dataset.

It has long been known that the tempering mechanism is capable of helping the system to get across high energy barriers and hence improve the ergodicity (Marinari and Parisi, 1992; Earl and Deem, 2005). Recently, the research of incorporating tempering into MCMC methods has provided a practical approach towards efficient multimodal posterior sampling (Graham and Storkey, 2017; Luo et al., 2018). In the meantime, the advances in thermostatting techniques for molecular dynamics (Jones and Leimkuhler, 2011) have shed some light on adaptive control for noisy dynamics. In this paper, we propose a novel method that addresses the two issues previously mentioned for HMC; it combines the protocol of parallel tempering (Swendsen and Wang, 1986; Sugita and Okamoto, 1999) with the dynamics of Nosé-Hoover (NH) thermostat (Nosé, 1984; Hoover, 1985). The simulation shows the advantages w.r.t. the accuracy as well as efficiency of our method against the classic HMC (Neal, 2011) and one of its stochastic variants, Stochastic Gradient Nosé-Hoover Thermostat (SGNHT) (Ding et al., 2014).

2. Parallel-tempered Stochastic Gradient Hamiltonian Monte Carlo

The proposed method consists of two alternating subroutines: 1. the parallel dynamics simulation of system replicas, and 2. the configuration exchange between replicas. The first subroutine utilizes the Nosé-Hoover thermostat to adaptively detect and neutralize the noise within mini-batch gradient; the second incorporates a mini-batch acceptance test to ensure the detailed balance during exchanges.

2.1. Parallel Dynamics Simulation of System Replicas

We define an increasing ladder $\{T_j\}_{j=1}^{R}$ of temperature with R rungs; the temperature ranges from the standard $T_1 = 1$ to some higher temperature. On each rung j, a replica (θ_j, p_j) of the physical system is initialized and the actual potential energy for that replica is rescaled to $U(\theta_j)/T_j$.

As the datum x within each mini-batch $\mathcal{S}$ is independently selected at random, the mini-batch gradient can be approximated by a Gaussian variable due to the Central Limit Theorem (CLT):

$$\nabla \tilde{U}(\theta) = -\nabla \log \pi(\theta) - \frac{|\mathcal{D}|}{|\mathcal{S}|} \sum_{x \in \mathcal{S} \subset \mathcal{D}} \nabla \log \ell(\theta; x). \tag{3}$$

To retain the correct trajectory in simulating the system dynamics, we leverage the NH thermostat because of its capability of adaptive control of the gradient noise (Jones and Leimkuhler, 2011; Ding et al., 2014). According to the formulation of Hoover (1985), for each replica (θ_j, p_j), we augment the system with NH thermostat $\xi_j \in \mathbb{R}$ and then modify the dynamics as:

$$\frac{\mathrm{d}\theta_j}{\mathrm{d}t} = M^{-1} p_j, \quad \frac{\mathrm{d}p_j}{\mathrm{d}t} = -\nabla \tilde{U}(\theta_j)/T_j - \xi p_j, \quad \frac{\mathrm{d}\xi_j}{\mathrm{d}t} = \left[p_j^\top M^{-1} p_j - D \right] \Big/ Q, \tag{4}$$

where M denotes the mass, and Q the thermal inertia. It can be proved that the dynamics in Eq. (4) leads to a stationary distribution w.r.t. θ_j by the Fokker-Planck equation (Risken and Haken, 1989)

$$\pi_j(\theta_j) \propto e^{-U(\theta_j)/T_j}. \tag{5}$$

This guarantees that, during the simulation, one can readily recover the desired distribution at a certain temperature T_j by simply retaining the position θ_j and discarding the momentum p_j as well as the thermostat ξ_j. Note that for the replica on rung 1, the temperature is fixed at standard $T_1 = 1$ and the position $\theta_1 = \theta$ is distributed as the target posterior $\pi_1(\theta_1) = e^{-U(\theta_1)/T_1} = e^{-U(\theta)} = \pi(\theta|\mathbb{D})$.

2.2. Configuration Exchange between Replicas

The principles of statistical physics suggest that high temperature facilitates the physical systems to get across energy barriers, which means replicas at higher temperatures are more likely to traverse among different modes of the distributions. As a consequence, however, the distribution sampled at high temperature has a spread spectrum and is hence biased. To recover an unbiased distribution, we perform configuration exchange between replicas at higher temperatures and the one at the standard.

Consider the configuration exchange between the replicas on rung i and j; as is a non-physical process, the exchange has to satisfy the condition of detailed balance:

$$\pi_j(\theta_j)\pi_k(\theta_k)\alpha[(j,k) \to (k,j)] = \pi_j(\theta_k)\pi_k(\theta_j)\alpha[(k,j) \to (j,k)], \tag{6}$$

where the transition probability reads

$$\alpha[(i,j) \to (j,i)] = \frac{\pi_j(\theta_k)\pi_k(\theta_j)}{\pi_j(\theta_j)\pi_k(\theta_k) + \pi_j(\theta_k)\pi_k(\theta_j)} = \frac{1}{1+e^{-\delta E}}, \tag{7}$$

and $\delta E = \big[U(\theta_k) - U(\theta_j)\big]\big[(T_k - T_j)/T_jT_k\big]$. It is straightforward to verify that Eq. (6) holds. Note that the transition probability $\alpha[(j,k) \to (k,j)]$ resembles the logistic distribution; such logistic test of acceptance is developed by Barker (1965).

With mini-batching, the potential energy $\tilde{U}(\theta_j)$ becomes a r.v., and so is the difference $\tilde{U}(\theta_k) - \tilde{U}(\theta_j)$. By CLT, δE is asymptotically Gaussian with some certain variance σ^2. Seita et al. (2017) proposed a mini-batch version of Baker's logistic test of acceptance such that $\delta E + \mathscr{C} > 0$ must hold for the exchange to carry out, where $\mathscr{L}$ denotes an auxiliary correction r.v. that aims to bridge the gap between the logistic distribution and Gaussian. The probability density $p_{\mathscr{C}}$ of this correction variable $\mathscr{C}$ satisfies the convolution equation $p_{\mathscr{C}} * p_{\mathcal{N}_{\sigma^2}} = p_{\mathscr{L}}$; it is equivalent to solve the Gaussian deconvolution problem w.r.t. the standard logistic distribution.

With the convolution theorem for distributions, it is helpful to convert the Gaussian deconvolution into solving for the inverse Fourier transform w.r.t. quotient of characteristic functions

$$p_{\mathscr{C}} = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{\phi_{\mathscr{L}}(t)}{\phi_{\mathcal{N}_{\sigma^2}}(t)} e^{-ixt}\,\mathrm{d}t, \quad \text{since } \phi_{\mathscr{C}} = \phi_{\mathscr{L}}/\phi_{\mathcal{N}_{\sigma^2}}, \tag{8}$$

where $\phi_{\mathcal{N}_{\sigma^2}}$ and $\phi_{\mathscr{L}}$ denote the characteristic functions of $\mathcal{N}(0,\sigma^2)$ and the standard logistic r.v., respectively. As the logistic distribution has much heavier tails than the Gaussian, the exact solution of $p_{\mathscr{C}}$ does not exist: the "integrand" on the RHS of Eq. (8) is in fact not integrable. We can only approximate $p_{\mathscr{C}}$ by introducing the kernel $\psi = e^{-\gamma^2 t^4}$ of bandwidth $1/\gamma$ (see Fan, 1991) in Eq. (8):

$$\hat{p}_{\mathscr{C}} = \frac{1}{2\pi}\int_{-\infty}^{\infty} \psi \cdot \frac{\phi_{\mathscr{L}}}{\phi_{\mathcal{N}_{\sigma^2}}} e^{-itx}\,\mathrm{d}t = \frac{1}{2\pi}\int_{-\infty}^{\infty} \left[\frac{\psi}{\phi_{\mathcal{N}_{\sigma^2}}}\right] \phi_{\mathscr{L}}\, e^{-ixt}\,\mathrm{d}t. \tag{9}$$

Using the Hermite polynomials H_k (Abramowitz and Stegun, 1965), we now expand the quotient within the brackets of Eq. (9) as

$$\frac{\psi}{\phi_{\mathcal{N}_{\sigma^2}}} = e^{-\gamma^2 t^4 + \sigma^2 t^2/2} = \sum_{k=0}^{\infty} \frac{\gamma^k}{k!} H_k(\sigma^2/4\gamma) t^{2k}. \tag{10}$$

The correction distribution can be approximated via Fourier's differential theorem:

$$\hat{p}_{\mathscr{C}} = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} H_k(\sigma^2/4\gamma)\gamma^k \left[\frac{1}{2\pi}\int_{-\infty}^{\infty} (-it)^{2k} \phi_{\mathscr{L}} e^{-itx}\, \mathrm{d}t\right] = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} H_k(\sigma^2/4\gamma)\gamma^k p_{\mathscr{L}}^{(2k)}, \tag{11}$$

where $p_{\mathscr{L}}^{(j)}$ represents the $(j+1)$-th derivative of logistic function, which can be efficiently calculated in a recursive fashion (Minai and Williams, 1993).

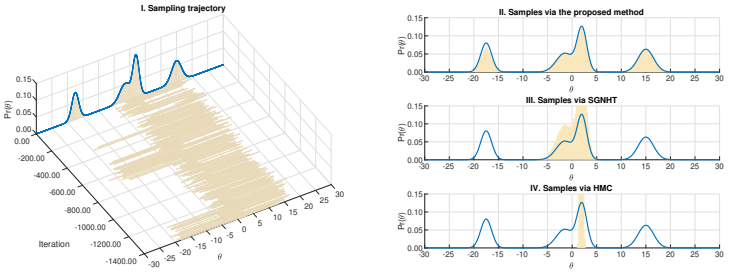

Figure 1: Experiment on sampling a $1d$ mixture of 4 Gaussians.

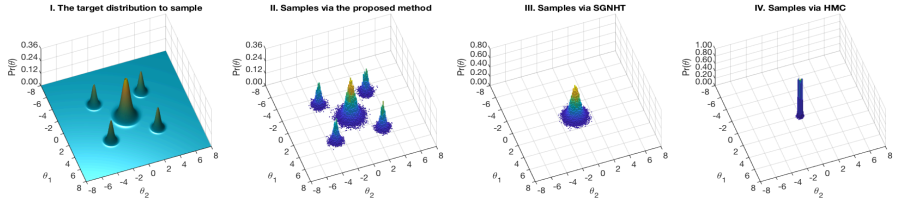

Figure 2: Experiment on sampling a $2d$ mixture of 5 Gaussians.

3. Experiment

We conduct two sets of experiments on synthetic distributions: the first is a mixture of 4 Gaussians in $1d$, and the second is a $2d$ Gaussian mixture with 5 isolated modes. The potential energy as well as its gradient is perturbed by zero-mean Gaussian noise with variance $\sigma^2 = 0.25$, and all samplers in test have no access to the actual parameters of that noise. We establish a ladder of temperature with $R = 10$ rungs ranging from $T_1 = 1$ to $T_R = 10$, i.e. totally 10 replicas are simulated in parallel. The baselines are the classic HMC (Neal, 2011) the adaptive variant SGNHT (Ding et al., 2014). It is demonstrated in Fig. 1 and 2 that, in both synthetic testing cases, our method has accurately sampled the target distributions with multiple isolated modes in the presence of noise within mini-batch gradient, where all baselines failed: SGNHT managed to control the gradient noise but did not discover the isolated modes while the classic HMC appears to be unable to correctly draw samples due to the deviated dynamics. Moreover, the subplot on the *left* of Fig. 1 illustrates the sampling trajectory of our method, indicating a good mixing property.

References

Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation, 1965.

Av A Barker. Monte carlo calculations of the radial distribution functions for a proton-electron plasma. *Australian Journal of Physics*, 18(2):119–134, 1965.

Nan Ding, Youhan Fang, Ryan Babbush, Changyou Chen, Robert D Skeel, and Hartmut Neven. Bayesian sampling using stochastic gradient thermostats. In *Advances in neural information processing systems*, pages 3203–3211, 2014.

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

David J Earl and Michael W Deem. Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7(23):3910–3916, 2005.

Jianqing Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, pages 1257–1272, 1991.

Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6): 721–741, 1984.

Matthew M. Graham and Amos J. Storkey. Continuously tempered hamiltonian monte carlo. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*, 2017.

W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

William G Hoover. Canonical dynamics: equilibrium phase-space distributions. *Physical review A*, 31(3):1695, 1985.

Andrew Jones and Ben Leimkuhler. Adaptive stochastic methods for sampling driven molecular systems. *The Journal of chemical physics*, 135(8):084125, 2011.

Rui Luo, Yaodong Yang, Jun Wang, and Yuanyuan Liu. Thermostat-assisted continuously-tempered hamiltonian monte carlo for multimodal posterior sampling on large datasets. In *Advances in Neural Information Processing Systems*, 2018.

Enzo Marinari and Giorgio Parisi. Simulated tempering: a new monte carlo scheme. *EPL (Europhysics Letters)*, 19(6):451, 1992.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

Ali A Minai and Ronald D Williams. On the derivatives of the sigmoid. *Neural Networks*, 6(6): 845–853, 1993.

Radford M Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2:113–162, 2011.

Shuichi Nosé. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of chemical physics*, 81(1):511–519, 1984.

H. Risken and H. Haken. *The Fokker-Planck Equation: Methods of Solution and Applications Second Edition*. Springer, 1989.

Daniel Seita, Xinlei Pan, Haoyu Chen, and John F. Canny. An efficient minibatch acceptance test for metropolis-hastings. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*, 2017.

Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical physics letters*, 314(1-2):141–151, 1999.

Robert H Swendsen and Jian-Sheng Wang. Replica monte carlo simulation of spin-glasses. *Physical Review Letters*, 57(21):2607, 1986.