Stochastic Gradient Monomial Gamma Sampler

Presenter: Yizhe Zhang

Joint with: Changyou Chen, Zhe Gan, Ricardo Henao, Lawrence Carin

Duke University

August 9, 2017

Background: Stochastic Gradient MCMC

- Sampling from $f(\theta) \propto \exp(-U(\theta, X))$
- Bayesian model averaging; uncertainty estimation;
- SG-MCMC replaces $U(\theta, X)$ with an unbiased *stochastic likelihood*, $\tilde{U}(\theta, x_\tau)$, evaluated from a subset of data, x_τ

$$\tilde{U}(\theta) = -\frac{N}{N'}\sum_{i=1}^{N'} \log p(x_{\tau_i}|\theta) - \log p(\theta)\,, \tag{1}$$

where $\{\tau_1, \cdots, \tau_{N'}\}$ are random subsets.

Background: Stochastic Gradient MCMC

- Driven by a continuous-time Markov stochastic process.

$$d\Gamma = V(\Gamma)dt + D(\Gamma)dW\,, \tag{2}$$

  where Γ denotes the parameters of the *augmented* system, *e.g.*, p and θ, $V(\cdot)$ and $D(\cdot)$ are referred as *drift* and *diffusion* vectors, respectively, and W denotes a standard Wiener process.

- To have a stationary distribution $p(\Gamma)$, *Fokker-Planck equation* needs to be hold.

$$\nabla_\Gamma \cdot p(\Gamma)V(\Gamma) = \nabla_\Gamma \nabla_\Gamma^T : [p(\Gamma)D(\Gamma)]$$

Background: Stochastic Gradient Hamiltonian Monte Carlo

- SGHMC (stochastic gradient Hamiltonian Monte Carlo) [Chen et. al., 2014] use stochastic gradient $\partial_\theta \tilde{U}(\theta)$, and introduce a friction term $B(\theta)$ to account for stochastic noise. The SDE is given as

$$d\theta = \partial_p K(p)dt \tag{3}$$

$$dp = -\partial_\theta \tilde{U}(\theta)dt - B(\theta)\partial_p K(p)dt + \mathcal{N}(0, 2B(\theta)dt). \tag{4}$$

where $K(p)$ is the kinetics, $K(p) = p^T p/m$

Background: Stochastic Gradient Nosé-Hoover thermostat

- SGNHT (stochastic gradient Nosé-Hoover thermostat) [Ding et. al, 2015] generalize the SGHMC to use thermostat for estimating the stochastic noise.

$$d\theta = \partial_p K(p) dt \tag{5}$$

$$dp = -\partial_\theta \tilde{U}(\theta) dt - \xi \partial_p K(p) dt + \mathcal{N}(0, 2B(\theta) dt) \tag{6}$$

$$d\xi = (p^T p - 1) dt. \tag{7}$$

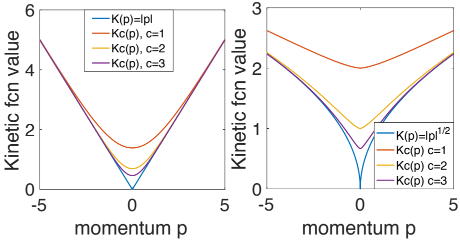
Improving over SGMCMC

We propose three techniques for improving efficiency of SGMCMC.

- Use *generalized kinetics* which delivers superior mixing rate.
- Use *additional dynamic* which helps convergence, and has better ergodic properties.
- Use *stochastic resampling* which helps convergence.

More efficient kinetics

- We consider *monomial Gamma* (MG) [Zhang et. al. 2016] kinetics $K(p) = |p|^{1/a}$, where $a \geq 1$.
- 1) Better stationary mixing 2) Better exploring multimodal distribution.
- However, directly applying such $K(p)$ will not satisfy FP equation.
- We use a softened version of MG kinetics.



Additional First Order Dynamics

- Hamiltonian system with a generalized form of kinetics and thermostat variable (stochastic noise).

$$H = K(p) + U(\theta) + F(\xi)\,, \tag{8}$$

- Consider SDE of SGNHT under this generalized form

$$\begin{aligned}
d\theta =& \nabla K(p)dt && (9)\\
dp =& -(\sigma_p + \gamma \nabla F(\xi)) \odot \nabla K(p)dt && (10)\\
& - \nabla U(\theta)dt + \sqrt{2\sigma_p}dW, && (11)\\
d\xi =& \gamma \left[\nabla K_c(p) \odot \nabla K(p) - \nabla^2 K(p)\right] dt. && (12)
\end{aligned}$$

- With **numerical integrator**, $\nabla U(\theta_t)$ is large $\rightarrow p_{t+1}$ is large.
- For $a > 1$, $\nabla K(p) \approx |p|^{1/a-1}$. p_{t+1} is large $\rightarrow \nabla K(p)$ is small $\rightarrow \theta$ won't change.

Additional First Order Dynamics (Cont'd)

- We consider adding first-order dynamics to θ and ξ

$$
\begin{aligned}
d\theta =& \nabla K_c(p)dt - \sigma_\theta \nabla U(\theta)dt + \sqrt{2\sigma_\theta} dW \\
dp =& -(\sigma_p + \gamma \nabla F(\xi)) \odot \nabla K_c(p)dt \\
& - \nabla U(\theta)dt + \sqrt{2\sigma_p} dW, \\
d\xi =& \gamma \left[\nabla K_c(p) \odot \nabla K_c(p) - \nabla^2 K_c(p) \right] dt \\
& - \sigma_\xi \nabla F(\xi)dt + \sqrt{2\sigma_\xi} dW \,.
\end{aligned} \tag{13}
$$

- Fortunately, the first order Langevin directly *compensate* this with large updating signal $\nabla U(\theta_{t+1})$
- On the other hand, when $\nabla U(\theta)$ is small, $\nabla K(p)$ would be large.
- The proposed SDE also has *better theoretic guarantee* on the existence and convergence of bounded solutions for a particular differential equation.

Stochastic resampling

- Resample p and ξ from their marginal distribution ($\propto \exp[-K(p)]; \exp[-F(\xi)]$) with a fixed frequency
- Move on a higher energy level is less efficient
- Make the sampler to **immediately** move to a lower energy level.
- Converge to stationary distribution

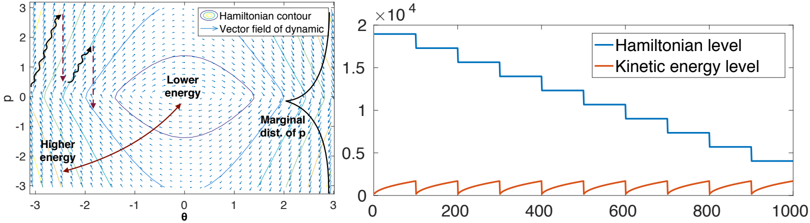


Figure: Stochastic resampling.

Theoretical properties

- Quantifying how fast the sample average, $\hat{\phi}_T$, converges to the true posterior average, $\bar{\phi} \triangleq \int \phi(\theta)\pi(\theta|X)\mathrm{d}\theta$, for $\hat{\phi}_T \triangleq \frac{1}{T}\sum_{t=1}^{T} \phi(\theta_t)$, where T is number of iterations.

Theorem

For the proposed SGMGT and SGMGT-D algorithms, if a fixed stepsize h is used, we have:

$$\textit{Bias: } \left|\mathbb{E}\hat{\phi}_T - \bar{\phi}\right| = O\left(1/(Th) + h\right) ,$$

$$\textit{MSE: } \mathbb{E}\left(\hat{\phi} - \bar{\phi}\right)^2 = O\left(1/(Th) + h^2\right) .$$

Experiments overview

- We evaluate our model on various tasks:
  1. Toy task: multiple-well synthetic potential
  2. Bayesian Logistic Regression
  3. Latent Dirichlet Allocation
  4. Discriminative RBM
  5. Bayesian Recurrent Neural Network

Multiple-well Synthetic Potential

- Generate samples from a complex multimodal distribution.
- SGMGT-D: w/ 1st dynamics and resampling

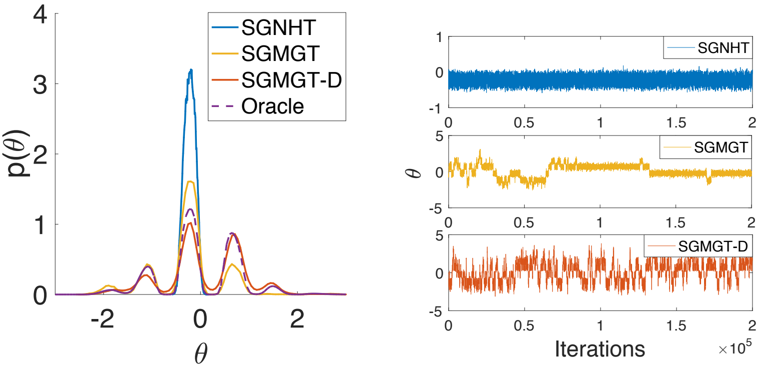


Figure: Synthetic multimodal distribution. Left: empirical distributions for different methods. Right: traceplot for each method.

Bayesian Logistic Regression

Table: Average AUROC and median ESS. Dataset dimensionality is indicated in parenthesis after the name of each dataset.

AUROC (D)	A (15)	G (25)	H (14)	P(8)	R (7)	C (87)
SGNHT	0.89	0.75	0.90	0.86	0.95	0.65
SGMGT(a=1)	0.92	0.78	0.91	0.86	0.87	0.70
SGMGT-D(a=1)	0.95	0.86	0.95	**0.93**	**0.98**	**0.73**
SGMGT(a=2)	0.93	0.79	0.93	0.88	0.86	0.62
SGMGT-D(a=2)	**0.95**	**0.90**	**0.95**	0.90	0.97	0.69
ESS (D)	A (15)	G (25)	H (14)	P(8)	R (7)	C (87)
SGNHT	869	941	1911	2077	1761	1873
SGMGT-D(a=1)	**3147**	**2131**	2448	**4244**	1494	**3605**
SGMGT-D(a=2)	2700	1989	**2768**	3430	**2265**	2969

Discriminative RBM for MNIST

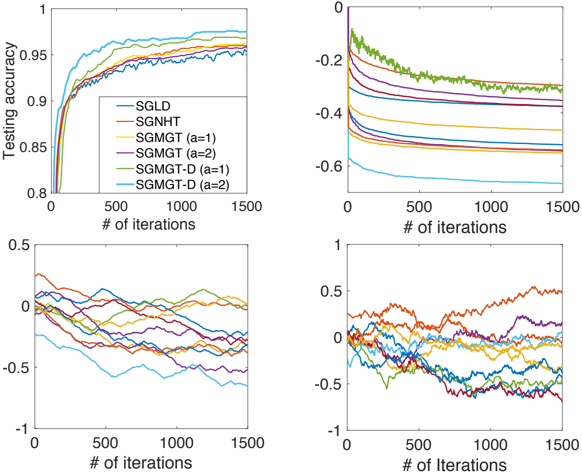


Figure: Experimental results for DRBM. Upper-left: testing accuracies for SGLD, SGNHT, SGMGT and SGMGT-D. Upper-right through lower-right: traceplots for SGLD, SGNHT and SGMGT-D with $a = 2$, respectively.

Bayesian Recurrent Neural Network

Table: Test negative log-likelihood results on polyphonic music datasets and test perplexities on PTB using RNN.

Algorithms	Piano	Nott	Muse	JSB	PTB
SGLD	11.37	6.07	10.83	11.25	127.47
SGNHT	9.00	4.24	7.85	9.27	131.3
SGMGT (a=1)	7.90	4.35	8.42	8.67	120.6
SGMGT (a=2)	10.17	4.64	8.51	8.84	250.5
SGMGT-D (a=1)	**7.51**	**3.33**	7.11	8.46	113.8
SGMGT-D (a=2)	7.53	3.35	**7.09**	**8.43**	**109.0**
SGD	11.13	5.26	10.08	10.81	120.44
RMSprop	7.70	3.48	7.22	8.52	120.45
ADAM	8.00	3.70	7.56	8.51	120.45

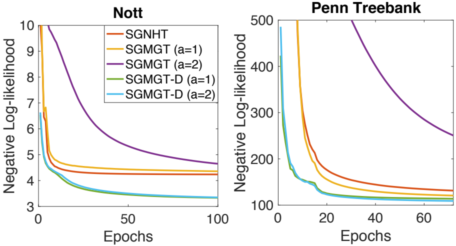


Figure: Learning curves of different SG-MCMC methods for RNN.

Conclusion and Future study

Conclusion:

- Scalable MCMC inference with improved stationary mixing efficiency.
- Remedies to alleviate practical issues with generalized HMC kinetics.
- Better theoretical guarantees.

Future research:

- Connection to optimization methods.

Q&A