Adaptive Bayesian Sampling with Monte Carlo EM

Anirban Roychowdhury Srinivasan Parthasarathy
Ohio State University

Abstract

We present a novel technique for learning the mass matrices in samplers obtained from discretized dynamics that preserve some energy function. Our approach provides a simpler alternative to Riemannian preconditioning techniques, by using existing dynamics in the sampling step of a Monte Carlo EM framework, and learning the mass matrices in the M step with a novel online technique. Along with a novel stochastic sampler based on Nosé-Poincaré dynamics, we use this framework with standard Hamiltonian Monte Carlo (HMC) as well as newer stochastic algorithms such as SGHMC and SGNHT, and show strong performance on synthetic and real high-dimensional sampling scenarios; we achieve sampling accuracies comparable to Riemannian samplers while being notably faster.

Outline

- Hamiltonian Monte Carlo (HMC) for Bayesian sampling treats the target density as an "energy function" augmented with auxiliary "momentum" parameters.
- A primary (hyper-)parameter of interest is the "mass" matrix of the kinetic energy term.
- Riemannian samplers are sensitive to the underlying geometry by reformulating it in terms of the target parameters to be sampled [1, 2, 3].
- We propose an alternative way to learn the mass using Monte Carlo EM (MCEM).
- MCEM is used to locally optimize maximum likelihood problems where the E step posterior probabilities in EM are not closed form.
- We perform existing dynamics derived from energy functions in the Monte Carlo E step while holding the mass fixed, and use the stored samples of the momentum term to learn the mass in the M step.

Experiments

- Synthetic : Parameter estimation for 1D standard normal distribution and 2D Bayesian logistic regression.
- Real world: Topic modeling with hierarchical Gamma processes (GPs).

Problem formulation for HMC

$$\max_{M \succ 0} \quad \mathcal{L}(\boldsymbol{\theta}) - \frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p} - \frac{1}{2}\log|M|$$

SGNHT [4] augmentation: $\max_{M \succ 0} \quad \mathcal{L}(\boldsymbol{\theta}) - \frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p} - \frac{1}{2}\log|M| + \mu(\xi - \bar{\xi})^2/2$

HMC-EM

Input: $\boldsymbol{\theta}^{(0)}, \epsilon, LP\_S, S\_\text{count}$
· Initialize M;
repeat
 · Sample $\mathbf{p}^{(t)} \sim N(0, M)$;
 for $i = 1$ **to** LP_S **do**
  · $\mathbf{p}^{(i)} \leftarrow \mathbf{p}^{(i+\epsilon-1)}$, $\boldsymbol{\theta}^{(i)} \leftarrow \boldsymbol{\theta}^{(i+\epsilon-1)}$;
  · $\mathbf{p}^{(i+\frac{\epsilon}{2})} \leftarrow \mathbf{p}^{(i)} - \frac{\epsilon}{2}\nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}^{(i)}, \mathbf{p}^{(i)})$;
  · $\boldsymbol{\theta}^{(i+\epsilon)} \leftarrow \boldsymbol{\theta}^{(i)} + \frac{\epsilon}{2}\nabla_{\mathbf{p}} H(\boldsymbol{\theta}^{(i)}, \mathbf{p}^{(i+\frac{\epsilon}{2})})$;
  · $\mathbf{p}^{(i+\epsilon)} \leftarrow \mathbf{p}^{(i+\frac{\epsilon}{2})} - \frac{\epsilon}{2}\nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}^{(i+\epsilon)}, \mathbf{p}^{(i+\frac{\epsilon}{2})})$;
 end for
 · Set $\left(\boldsymbol{\theta}^{(t+1)}, \mathbf{p}^{(t+1)}\right)$ from $\left(\boldsymbol{\theta}^{LP\_S+\epsilon}, \mathbf{p}^{LP\_S+\epsilon}\right)$ using Metropolis-Hastings
 · Store MC-EM sample $\mathbf{p}^{(t+1)}$;
 if $(t+1)$ mod $S\_\text{count} = 0$ **then**
  · Update M using MC-EM samples;
 end if
 · Update S_count as described in the paper;
until forever

SGNHT-EM

Input: $\boldsymbol{\theta}^{(0)}, \epsilon, A, LP\_S, S\_\text{count}$
· Initialize $\xi^{(0)}$, $\mathbf{p}^{(0)}$ and M;
repeat
 for $i = 1$ **to** LP_S **do**
  · $\mathbf{p}^{(i+1)} \leftarrow \mathbf{p}^{(i)} - \epsilon\xi^{(i)} M^{-1}\mathbf{p}^{(i)} - \epsilon\tilde{\nabla}\mathcal{L}(\boldsymbol{\theta}^{(i)}) + \sqrt{2A}\mathcal{N}(0, \epsilon)$;
  · $\boldsymbol{\theta}^{(i+1)} \leftarrow \boldsymbol{\theta}^{(i)} + \epsilon M^{-1}\mathbf{p}^{(i+1)}$;
  · $\xi^{(i+1)} \leftarrow \xi^{(i)} + \epsilon\left[\frac{1}{D}\mathbf{p}^{(i+1)T} M^{-1}\mathbf{p}^{(i+1)} - 1\right]$;
 end for
 · Set $\left(\boldsymbol{\theta}^{(t+1)}, \mathbf{p}^{(t+1)}, \xi^{(t+1)}\right) = \left(\boldsymbol{\theta}^{(LP\_S+1)}, \mathbf{p}^{(LP\_S+1)}, \xi^{(LP\_S+1)}\right)$;
 · Store MC-EM sample $\mathbf{p}^{(t+1)}$;
 if $(t+1)$ mod $S\_\text{count} = 0$ **then**
  · Update M using MC-EM samples;
 end if
 · Update S_count as described in the paper;
until forever

Results

Setup

- Synthetic Gaussian experiments
  - $5,000$ datapoints from standard normal distribution
  - Normal-Wishart priors
  - Metric tensor : Observed Fisher information plus negative Hessian of prior
- Synthetic Bayesian LR
  - $5,000$ datapoints from two Gaussians with means at $[-1, 1], [1, -1]$, and unit covariance
  - Linear classifier with weights $(\omega_1, \omega_2) = [1, -1]$
- Topic modeling using GP construction from [5]
  - Poisson factor analysis on term-document count matrix.
  - $\mathbf{D}_{V\times N} = \text{Poi}(\boldsymbol{\Phi}\boldsymbol{\Theta})$, $\theta_{n,k} \sim \Gamma(r_k, \frac{p_j}{1-p_j})$, r_ks being the GP weights.
  - Perplexities measured for 20-Newsgroups and Reuters corpora, with stochastic samplers and MCEM augmentations.

RMSE and runtimes

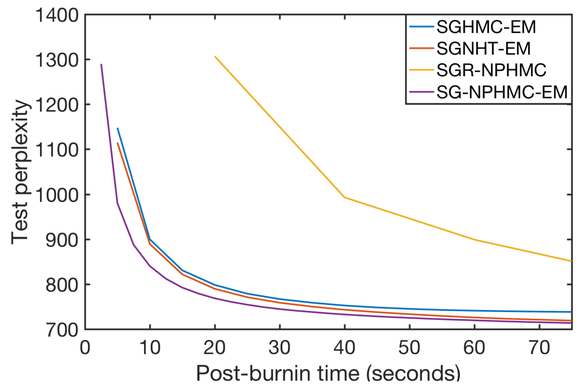
- Per-iteration runtimes and RMSE for HMC, HMC-EM and RHMC on synthetic data.
- Synthetic Gaussian dataset.

Method	RMSE(μ)	RMSE(τ)	Time
HMC	0.0196	0.0197	0.417ms
HMC-EM	0.0115	0.0104	0.423ms
RHMC	0.0111	0.0089	5.748ms

- Synthetic regression dataset.

Method	RMSE(W_0)	RMSE(W_1)	Time
HMC	0.0456	0.1290	1.435ms
HMC-EM	0.0145	0.0851	1.428ms
RHMC	0.0091	0.0574	1550ms

Results, contd.



- Test perplexities plotted against wall-clock time for the 20-Newsgroups dataset.
- The MCEM algorithms converge to perplexities within 3% of SGR-NPHMC [2], but are an order of magnitude faster.

References

[1] M. Girolami and B. Calderhead.
Riemann manifold Langevin and Hamiltonian Monte Carlo methods.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(2):123–214, 2011.

[2] A. Roychowdhury, B. Kulis, and S. Parthasarathy.
Robust Monte Carlo Sampling using Riemannian Nosé-Poincaré Hamiltonian Dynamics.
In *Proceedings of The 33rd International Conference on Machine Learning (ICML)*, pages 2673–2681, 2016.

[3] S. Patterson and Y. W. Teh.
Stochastic Gradient Riemannian Langevin Dynamics on the Probability Simplex.
In *Advances in Neural Information Processing Systems (NIPS) 26*, pages 3102–3110, 2013.

[4] N. Ding, Y. Fang, R. Babbush, C. Chen, R. D. Skeel, and H. Neven.
Bayesian Sampling using Stochastic Gradient Thermostats.
In *Advances in Neural Information Processing Systems (NIPS) 27*, pages 3203–3211, 2014.

[5] A. Roychowdhury and B. Kulis.
Gamma Processes, Stick-Breaking, and Variational Inference.
In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 800–808, 2015.

Contact Information

- AR: roychowdhury.7@osu.edu
- SP: srini@cse.ohio-state.edu