Large-Scale Stochastic Sampling from the Probability Simplex

Jack Baker[1], Paul Fearnhead[1], Emily Fox[2] and Christopher Nemeth[1]

[1]STOR-i CDT, Lancaster University [2]University of Washington

EPSRC Engineering and Physical Sciences Research Council

1. Big Data MCMC & SGLD

Problem: MCMC requires calculations over full dataset each iteration. High computational cost at large datasets.

Popular Solution: SGLD [3], which uses a subset of data at each iteration. First, construct unbiased estimate of log posterior gradient using subsample of the data:

$$\nabla \log \hat{\pi}(\theta) = \nabla \log \pi_0(\theta) + \frac{N}{|S|} \sum_{i \in S} \nabla \log f(x_i \,|\, \theta), \quad S \subset \{1, \ldots, N\}.$$

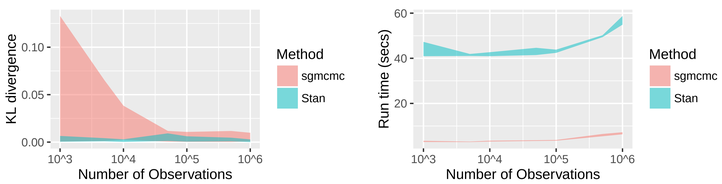
SGLD Algorithm

- Set starting value θ_0 and stepsize h.
- Iterate and store the following:

$$\theta_{m+1} = \theta_m + h \nabla \log \hat{\pi}(\theta) + \sqrt{2h}\zeta_m,$$

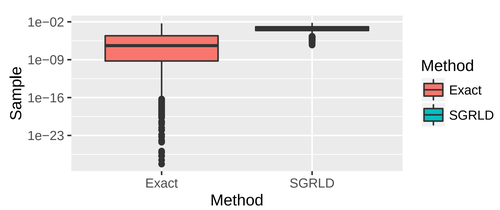
where $\zeta_m \sim N(0, 1)$.

When $h \to 0$ & $m \to \infty$, SGLD samples from the true posterior. In practice use small h; approximation quality depends on dataset size N. As N increases, approximation of SGLD gets better.



2. Problem

Problem: SGLD designed for unbounded spaces $\mathbb{R}^d$. Biased on constrained spaces like $[0, \infty)$; especially at boundary. SGRLD [2] aimed to fix problem for simplex space but still biased due to discretisation by h.



3. CIR Process

Aim: Construct a scalable, approximate MCMC algorithm for the simplex (i.e., Dirichlet distribution) with no discretisation error. This removes biases at the boundary.

Useful Transformation: If $\theta_j \sim \text{Gamma}(a_j, 1)$, for $j = 1, \ldots, d$; then $\omega := (\theta_1, \ldots, \theta_d)/\sum_j \theta_j \sim \text{Dirichlet}(\mathbf{a})$.

Useful Process: Cox-Ingersoll-Ross (CIR) process [1]. Fix $h > 0$. Aim to target Gamma$(a, 1)$ distribution. Update is:

$$\theta_{m+1} = \frac{1 - e^{-h}}{2} W, \qquad W \,|\, \theta_m \sim \chi^2\left(2a, 2\theta_m \frac{e^{-h}}{1 - e^{-h}}\right),$$

$\chi^2(\nu, \mu)$ is a noncentral Chi sq distribution. As $m \to \infty$ CIR samples Gamma$(a, 1)$ for any $h > 0$. So h no longer discretising.

4. Making Scalable: SCIR Process

Idea: Use CIR process to target Gamma$(a_j, 1)$, $j = 1, \ldots, d$; then use transformation to get simplex sample $\omega \sim \text{Dirichlet}(\mathbf{a})$.

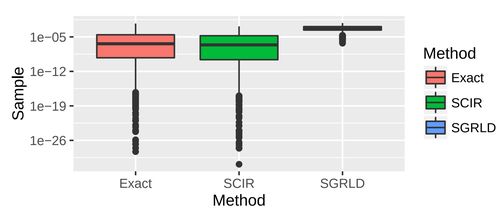
Problem: Typically Dirichlet posterior of form

$$\text{Dirichlet}\left(a_{01} + \sum_{i=1}^{N} z_{i1}, \ldots, a_{0d} + \sum_{i=1}^{N} z_{id}\right);$$

i.e., CIR process calculates $O(N)$ sum each iteration. Expensive!

Solution – SCIR Process: Similar to SGLD, run CIR process with unbiased estimate of a using subsample of data: $\hat{a} = a_0 + \frac{N}{|S|} \sum_{i \in S} z_i$.

SCIR process still discretisation free! h only determines how often $\hat{a}$ resampled. Similar to SGLD: as $m \to \infty$, $h \to 0$ SCIR samples from target Gamma$(a, 1)$. As discretisation free, provably unbiased.



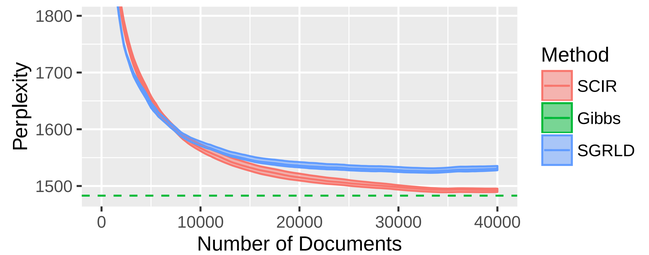
5. Extensions and Theory

Extensions: Method not limited to simplex. SCIR proceess can sample from Gamma distributed $[0, \infty)$ spaces. Common for inference of variances. Also, processes similar to CIR can be exploited; e.g., geometric Brownian process can be used for scalable inference on lognormal distributed $[0, \infty)$ spaces.

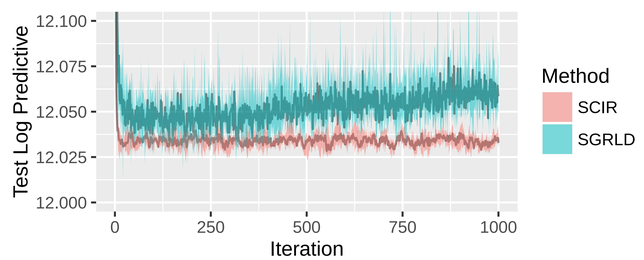
Theory: The CIR process is tractable. Allows us to derive theoretical results for SCIR. Can prove asymptotic unbiasedness and derive the non-asymptotic moment generating function.

6. Experiments

LDA: Comparison between SCIR, SGRLD and (non-scalable) Gibbs on a Latent Dirichlet Allocation model on Wikipedia corpus.



Dirichlet Process Mixture: Comparison between SCIR, SGRLD on a Dirichlet Process Mixture model on the Microsoft user dataset. Stochastic DP sampler not been considered before.



[1] John C. Cox, Jonathan E. Ingersoll, and Stephen A. Ross. A theory of the term structure of interest rates. *Econometrica*, 53(2):385–407, 1985.

[2] Sam Patterson and Yee Whye Teh. Stochastic gradient Riemannian Langevin dynamics on the probability simplex. In *Advances in Neural Information Processing Systems 26*, pages 3102–3110. Curran Associates, Inc., 2013.

[3] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pages 681–688. PMLR, 2011.