Stats 230 Final Project Report

David Mwakima and Jizhi Zhang

03/10/2023

1 Introduction

Markov Chain Monte Carlo (MCMC) is a numerical integration technique. The main motivation for this technique is that it is often not feasible both mathematically and practically to evaluate certain integrals in high dimensions. In Bayesian statistics, in particular, all inference proceeds via the posterior distribution $p(\theta|\mathbf{X})$ where θ could be a vector of multiple parameters. This distribution is given by

$$p(\theta|\mathbf{X}) = \frac{1}{C} f(\mathbf{X}|\theta)p(\theta)$$

where $f(\mathbf{X}|\theta)$ is our sampling model for $\mathbf{X}$, $p(\theta)$ is our prior model for θ and $C = \int_\Theta f(\mathbf{X}|\theta)p(\theta)$ is the normalizing constant. But computing C requires integration if we are going to evaluate this posterior and use it for inference. This is difficult, especially in cases where it is not available in closed analytic form.

Since Gelfand and Smith (1990) rediscovered the Metropolis-Rosenbluth-Hastings (MRH) algorithm developed by Metropolis et al. (1953), this difficulty has been addressed. This algorithm relies on the theory of Markov stochastic processes. Assuming certain conditions (irreducibility, positive recurrence, detailed balance) are satisfied by a homogeneous Markov chain $\{X_n\}$ on state space E, then one can show that the chain possesses a stationary distribution π. One then appeals to the Ergodic Theorem (a dependent samples analogue of the Strong Law of Large Numbers for i.i.d samples) that guarantees that for any initial distribution:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} f(X_k) = \sum_{i\in E} f(i)\pi_i = E_\pi[f(X)]$$

Let $p(\theta|\mathbf{X})$ be the target distribution, where θ is a vector of parameters. Then with proposal densities $q(\theta^{(j)}|\theta^{(i)}, \mathbf{X})$, the MRH acceptance ratio is:

$$a(\theta^{(j)}, \theta^{(i)}, \mathbf{X}) = \min\left\{ \frac{p(\theta^{(j)}|\mathbf{X})}{p(\theta^{(i)}|\mathbf{X})} \frac{q(\theta^{(i)}|\theta^{(j)}, \mathbf{X})}{q(\theta^{(j)}|\theta^{(i)}, \mathbf{X})}, 1 \right\}$$

And the Metropolis-Rosenbluth-Hastings Algorithm (MRH) for approximating $E_{p(\theta|\mathbf{X})}[h(\mathbf{X})]$ is:

	Metropolis-Rosenbluth-Hastings Algorithm
1:	Start with some initial value $\theta = \theta^{(0)}$
2:	**for** i = 0 to N **do**
3:	Simulate $\theta^{(i+1)}$ with $q(\theta^{(i+1)}|\theta^{(i)}, \mathbf{X})$
4:	Compute $a(\theta^{(i+1)}, \theta^{(i)}, \mathbf{X})$
5:	Generate $U \sim \text{Unif}(0, 1)$
6:	Accept $\theta^{(i+1)}$ if $U \leq a(\theta^{(i+1)}, \theta^{(i)}, \mathbf{X})$
	Otherwise $\theta^{(i+1)} = \theta^{(i)}$
7:	**end for**
8:	**return** $\frac{1}{N}\sum_{i=1}^{N} h(X_i)$

However, the use of MCMC for large datasets presents a new research frontier, see Bardenet, Doucet, and Holmes (2014) and Bardenet, Doucet, and Holmes (2017). This is because when $\mathbf{X}$ is large $n \gg 1$ as is typically the case in genomics, spatial statistics and cosmology, evaluating the likelihood ratio appearing in the MRH acceptance ratio is computationally intensive $O(n^3)$.

$$\frac{p(\theta^{(j)}|\mathbf{X})}{p(\theta^{(i)}|\mathbf{X})} = \frac{p(\theta^{(j)})}{p(\theta^{(i)})} \frac{\prod_{k=1}^{n} f(X_k|\theta^{(j)})}{\prod_{k=1}^{n} f(X_k|\theta^{(i)})}$$

As a consequence MCMC with the MRH algorithm cannot be considered for reasonable runtime.

In our report we consider a paper by Maire, Friel, and Alquier (2019) that addresses the problem of

using MCMC for large datasets. This paper proposes a new methodology, which the authors call *Informed Sub-Sampling MCMC* (ISS-MCMC), for doing Bayesian MCMC approximation of the posterior distribution. This is a scalable version of the Metropolis-Hastings algorithm designed for situations when N is so big that to approximate the posterior distribution takes a very long time.

2 Main ideas of how it works

ISS-MCMC is "informed" because it makes use of a measure of similarity with respect to the full dataset through summary statistics. It is "subsampling" because it uses this measure to select a subset of the dataset that will be used by the Markov transition kernel at the k-th iteration of the algorithm. In this way, the Markov chain transition kernel uses only a fraction n/N of the entire dataset. They show using examples that choosing $n \ll N$ can lead to significant reductions in computational run-times while still retaining the simplicity of the standard Metropolis-Hastings algorithm. In the following subsections we consider in more detail the main ideas of how this algorithm works. See section 4.

2.1 Similarity through summary statistics

Summary statistics are a set of numerical measures that are used to describe the essential features of a data set. These measures provide a quick and easy way to understand the key characteristics of a dataset, such as its center, spread, and shape. The most commonly used summary statistics include measures of central tendency, such as the mean, median, and mode, which give an indication of where the data cluster around. Measures of variability, such as the standard deviation and range, provide information about how spread out the data are. Other summary statistics, such as skewness and kurtosis, give an indication of the shape of the distribution. By using summary statistics, researchers and data analysts can quickly get a sense of what the data looks like, which can help guide further analysis and interpretation.

The choice of the summary statistics in the algorithm is problem specific and is meant to be the counterpart of the sufficient statistic mapping for general models (hence sharing, slightly abusively, the same notation). Since the question of specifying summary statistics also arises in Approximate Bayesian Computation (ABC), one can take advantage of the abundant ABC literature on this topic to find some examples of summary statistics for usual likelihood models.

2.2 Transition kernel

[Jizhi add details here]

2.3 The Algorithm

[Copy-paste]

3 Comparison with other approaches

Other similar approaches to solve the same statistical problems are Quiroz et al. (2018) and the *Confidence Sampler* in Bardenet, Doucet, and Holmes (2017). Both of these approaches use sophisticated "control variates" to get positive unbiased estimators (based on a subset of data) for the likelihoods in the Metropolis-Hastings acceptance ratio. The authors note that these control variates are computationally intensive.

The noisy approaches due to Korattikara, Chen, and Welling (2014) and Alquier et al. (2016)

Maclaurin and Adams (2015) also addresses the computational issue are independent.

Another approach which the authors compare their approach with is an approach based on continuous time Markov processes (Langevin diffusion, Zig-Zag process) in Fearnhead et al. (2018) and Bierkens, Fearnhead, and Roberts (2019). Here, the authors note that the computational hurdle involves calculation of the gradient of the log-likelihood, which may not always be unbiased. Moreover, these approaches depart significantly from the simplicity of the original discrete M-H algorithm.

[David add details]

4 Example with logistic regression

Here we reproduce one example in their paper for the case of logistic regression. See section 6.3

[Meet to code example next week]

4.1 Compare convergence of ISS-MCMC with M-H.

[Add after coding example]

References

Alquier, Pierre, Nial Friel, Richard Everitt, and Aidan Boland. 2016. "Noisy Monte Carlo: Convergence of Markov Chains with Approximate Transition Kernels." *Statistics and Computing* 26 (1-2): 29–47.

Bardenet, Rémi, Arnaud Doucet, and Chris Holmes. 2014. "Towards Scaling up Markov Chain Monte Carlo: An Adaptive Subsampling Approach." In *International Conference on Machine Learning*, 405–13.

Bardenet, Rémi, Arnaud Doucet, and Christopher C Holmes. 2017. "On Markov Chain Monte Carlo Methods for Tall Data." *Journal of Machine Learning Research* 18 (47).

Bierkens, Joris, Paul Fearnhead, and Gareth Roberts. 2019. "The Zig-Zag Process and Super-Efficient Sampling for Bayesian Analysis of Big Data." *The Annals of Statistics* 47 (3): 1288–1320.

Fearnhead, Paul, Joris Bierkens, Murray Pollock, and Gareth O Roberts. 2018. "Piecewise Deterministic Markov Processes for Continuous-Time Monte Carlo." *Statistical Science* 33 (3): 386–412.

Gelfand, Alan E, and Adrian FM Smith. 1990. "Sampling-Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association* 85 (410): 398–409.

Korattikara, Anoop, Yutian Chen, and Max Welling. 2014. "Austerity in MCMC Land: Cutting the Metropolis-Hastings Budget." In *International Conference on Machine Learning*, 181–89.

Maclaurin, Dougal, and Ryan P. Adams. 2015. "Firefly Monte Carlo: Exact MCMC with Subsets of Data." In *Proceedings of the 24th International Conference on Artificial Intelligence*, 4289–95. AAAI Press.

Maire, Florian, Nial Friel, and Pierre Alquier. 2019. "Informed Sub-Sampling MCMC: Approximate Bayesian Inference for Large Datasets." *Statistics and Computing* 29: 449–82.

Metropolis, Nicholas, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. 1953. "Equation of State Calculations by Fast Computing Machines." *The Journal of Chemical Physics* 21 (6): 1087–92.

Quiroz, Matias, Robert Kohn, Mattias Villani, and Minh-Ngoc Tran. 2018. "Speeding up MCMC by Efficient Data Subsampling." *Journal of the American Statistical Association*.