Stats 230 Final Project Report

David Mwakima and Jizhi Zhang

03/10/2023

1 Introduction

What is MCMC for Bayesian inference?

[Add details here]

The Metropolis-Hastings algorithm developed by Metropolis et al. (1953) is used for MCMC.

The use of MCMC for Large datasets presents a new research frontier.

[Add details here]

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.

2 Main ideas of how it works

This methodology is "informed" because it makes use of a measure of similarity with respect to the full dataset through summary statistics. It is "sub-sampling" 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.

3 Comparison with other approaches

Other 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.

Another approach which the authors compare their approach with is an approach based on continuous time Markov processes (Zig-Zag process, Langevin diffusion) in Fearnhead et al. (2018). 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.

4 Example with logistic regression

References

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).

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.

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.*