6.882 Project Proposal: An Overview of Techniques for Accelerated MCMC-MH

Edward Park, Michael Sun

March 2018

1 Problem Statement

In this project, we propose to replicate the 2-stage Metropolis-Hastings (MH) algorithm formulated by Payne and Mallick (2014) [3]. We will attempt replicate its results on the three datasets proposed in the paper: Portugese banking data[2], loan data from Freddie Mac dataset[1], and a simulated data set. Time permitting, we will compare these results to other methods of speeding up MH like consensus Monte Carlo and data subsampling.

2 Background

For the past several decades, Markov Chain Monte Carlo (MCMC) methods have been used to sample from nonstandard posterior distributions in Bayesian statistics. However, when dealing with a large data set, MCMC methods tend to scale rather poorly. Much of the computational complexity of the Metropolis-Hastings (MH) algorithm comes from the need to make a full pass over the observed data at each iteration. As such, there have been a multitude of papers that propose various techniques to improve this performance.

The paper by Payne and Mallick [3] does an excellent job researching past results and compiling a list of the various techniques used. The authors describe simple likelihood parallelization, consensus Monte Carlo [5], subsampling based methods [4], and their own proposed idea, two-stage MH. Importantly, they compare the posterior distributions and performance numbers for each of the methods, concluding that their own two-stage method is the most efficient.

3 Timeline/Division of Labor

We tentatively plan to have the following timeline:

1. Discuss and figure out the mathematical / algorithmic details that are left out of the paper. (This will likely involve reading through the references cited by the Payne paper) [Edward + Michael, by 4/6]
2. Create a framework + skeleton code in which we can implement Metropolis-Hastings [Edward + Michael, by 4/13]
3. Implement the standard MH technique. [Michael, by 4/20]
4. Implement the consensus Monte Carlo method. [Edward, by 4/27]
5. Implement the subsampling method. [Edward, by 4/27]
6. Implement the two-stage MH method. [Michael, by 4/27]
7. Perform the necessary data processing on the three datasets mentioned above [Edward, by 5/4]
8. Run experiments and generate plots [Michael, by 5/11]

4 Risks/Contingency Plan

Below, we outline some of the potential risks we discussed regarding our project and our proposed contingency plans:

- Unsuitable data set: Our proposed datasets (Freddie Mac, Portuguese, BMARS simulated data) are inaccessible, inefficient, or otherwise infeasible for us to use.
  Backup Plan: Our workflow is structured so that we will (hopefully) have a generic implementation of the MH algorithm and a framework that is dataset-agnostic. If this is the case, we will have tested our framework on some toy data before jumping into using one of the three proposed datasets. Thus, if we find that all the datasets are unsuited towards our needs, we can try to replicate the paper's results (with comparisons to other methods) on our toy data. In addition, if we find the datasets are unsuitable early in our workflow, we can switch to a dataset (ex. MNIST) that we know is more suitable for our needs.
- Incomplete algorithm: 2-stage MH incomplete by personal deadline
  Backup Plan: If we do not meet our personal deadlines for implementing the MH algorithm (of which the 2-stage MH is likely to give us the most trouble), we will expand the breadth of our project, researching more datasets to compare our methods across.

References

[1] Freddie Mac. *Single Family Loan-Level Dataset*. URL: http://www.freddiemac.com/research/datasets/sf_loanlevel_dataset.html.

[2] Sergio Moro, Paulo Cortez, and Paulo Rita. "A Data-Driven Approach to Predict the Success of Bank Telemarketing". In: (2014). URL: https://pdfs.semanticscholar.org/4a27/709545cfa225d8983fb4df8061fb205b911pdf.

[3] R. D. Payne and B. K. Mallick. "Two-Stage Metropolis-Hastings for Tall Data". In: *ArXiv e-prints* (Nov. 2014). arXiv: 1411.5653 [stat.ME].

[4] M. Quiroz et al. "Speeding Up MCMC by Efficient Data Subsampling". In: *ArXiv e-prints* (Apr. 2014). arXiv: 1404.4178 [stat.ME].

[5] Steven L. Scott et al. "Bayes and Big Data: The Consensus Monte Carlo Algorithm". In: *International Journal of Management Science and Engineering Management* 11 (2016), pp. 78–88. URL: http://www.tandfonline.com/doi/full/10.1080/17509653.2016.1142191.