Matias Quiroz and Mattias Villani
Division of Statistics and Machine Learning
Dept. of Computer and Information Science
Linköping University

November 10, 2016

Master's thesis proposal

Stochastic Gradient Langevin Dynamics for large data sets

Stochastic Gradient Langevin Dynamics (SGLD) is a prominent posterior sampling algorithm for learning from large scale data sets in Machine Learning (Welling and Teh, 2011). SGLD use stochastic optimization where each step is perturbed by random noise. Stochastic optimization is, in principle, a randomized version of Newton's method for optimization. The randomization comes from estimating the gradient of the likelihood using only a small random subset of observations, hence its prominence for large data sets. The noise perturbation prevents the stochastic optimization to collapse to a mode and, more importantly, is such that the iterates of the algorithm form samples from the posterior distribution. Posterior sampling is extremely useful as it allows to easily quantify essentially any type of uncertainty that might be of interest.

Bardenet et al. (2015) demonstrate that the posterior samples generated by SGLD can sometimes be a poor approximation. One possible explanation is that the random subset of observations includes every observation with the same probability, causing highly variable estimates of the gradient of the likelihood. This project aims to develop methods to estimate this gradient efficiently and study the potential improvement of the posterior approximation. Efficient variance reduction techniques, although in the context of Markov Chain Monte Carlo (MCMC) simulation where the likelihood is estimated using a random subset, have been developed in Quiroz et al. (2016a) (see also Quiroz et al. 2016b). This project investigates the potential of adapting these techniques for estimating the gradient of the likelihood in SGLD. It is also of interest to compare this approach to the so called subsampling MCMC in Quiroz et al. (2016a, 2016b).

This project will give you a good insight into the research frontier in large scale data challenges in Machine Learning. Except learning a great deal about SGLD you will also learn about subsampling MCMC.

References

Bardenet, R., Doucet, A., and Holmes, C. (2015). On Markov chain Monte Carlo methods for tall data. *arXiv preprint arXiv:1505.02827*.

Quiroz, M., Villani, M., and Kohn, R. (2016a). Speeding up MCMC by efficient data subsampling. *arXiv preprint arXiv:1404.4178v3*.

Quiroz, M., Villani, M., and Kohn, R. (2016b). Exact subsampling MCMC. *arXiv preprint arXiv:1603.08232*.

Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688.