Unbiased Posterior Expectations for Big Data

Heiko Strathmann[1], *Dino Sejdinovic*[2], *Mark Girolami*[3]

[1]Gatsby Unit for Computational Neuroscience, University College London
[2]Department of Statistics, University of Oxford
[3]Department of Statistics, Warwick University

Markov Chain Monte Carlo is a fundamental tool in Bayesian data analysis. However, the advent of Big Data resources to be analysed poses serious challenges for this methodology as it does not scale to today's, and future projected, dataset sizes. Recent research on practical variants of MCMC for large datasets has focused on devising a transition kernel that will lead to samples being drawn from an approximately correct full posterior measure, thus reducing the amount of computation at the expense of introducing bias. Since the goal of MCMC for Bayesian inference is almost always to estimate the expectations of certain functions of interest over the posterior, we challenge this paradigm. According to the dictum that one should never solve a harder problem than the one at hand, we construct a scheme that allows for the unbiased estimation of posterior expectations with MCMC without exact simulation from the full posterior. The average complexity of our scheme is sublinear in the number of observations and the variance of estimators is straightforward to control, leading to algorithms that are provably unbiased and naturally arrive at the desired error tolerance. We demonstrate the utility of the proposed methodology on a variety of test cases by way of illustration as well as examples from ongoing massive scale statistical data analysis.