1 Biography

My name is Jeffrey Chan. I am a first year EECS PhD student working in Yun Song’s group. My research interests are centered around statistical and computational problems in population genetics, Bayesian inference methods, and statistical machine learning. I’d like to gain the requisite knowledge to design algorithms to fully utilize computational resources in my research.

2 Problem: Hogwild

Stochastic gradient descent (SGD) is widely used in data-heavy machine learning problems due to its robustness, rapid convergence rates, and small memory requirements. A difficult problem is understanding how to parallelize SGD to properly leverage our multicore systems and large datasets. SGD is inherently sequential with the each iteration determined by the most recent value making parallelizing SGD difficult. Many MapReduce-based algorithms are tailored towards batch algorithms and do not perform well on online data. Hogwild! proposes a parallel lock-free approach that simply allows each processor to overwrite the memory and works well with sparse data.

The strength of Hogwild! is that it removes the overhead of locking mechanism and converges linearly. However, this only holds if the optimized variable gets minimally changed during each overwrite. This result is particularly interesting since it clearly scales well since there is no overhead. Hogwild! also scales well when the gradients are particularly hard to compute and seems to do well against other methods such as Round Robin. However, it is essentially allowing the cores to act independently and overwrite the memory which seems naive initially. It is unclear in practice how often problems have ‘sparse’ decision variables to allow for Hogwild! to achieve its theoretical guarantees. In their experiments, they have done a few different optimization problems on various data sets and seemed to achieve good results. The experiments were run in C++ on a dual Xeon X650 CPU with Raid-0 software over 7 2TB disks.