BIOMATH

Ph.D. Dissertation Defense Seminar

Projection Algorithms for Large Scale Optimization and Genomic Data Analysis

Kevin Keys
Doctoral Graduate Student
Department of Biomathematics
UCLA

Friday, June 10, 2016
10:00 AM
Gonda Conference Room 1357

ABSTRACT:

The advent of the Big Data era has spawned intense interest in scalable optimization methods. Traditional approaches such as Newton's method fall apart whenever the features outnumber the examples in a data set. Consequently, researchers have intensely developed first-order methods that rely only on gradients and subgradients of a cost function. In this dissertation we focus on projected gradient methods for large-scale constrained optimization. We develop a particular case of a proximal gradient method called the *proximal distance algorithm* that combines the classical penalty method of constrained minimization with distance majorization. To optimize the loss function $f(x)$ over a constraint set C, the proximal distance principle mandates minimizing the penalized loss $f(x) + t\ \mathrm{dist}(x,C)^2$ and following the solution x to its limit as t tends to infinity. At each iteration $\mathrm{dist}(x,C)^2$ is majorized by $\| x - \Pi_C(x_k) \|^2$, where $\Pi_C(x_k)$ denotes the projection of the current iterate x_k onto C. The minimum of $f(x) + t \| x - \Pi_C(x_k) \|^2$ is given by the proximal map $\mathrm{prox}_{(1/t)f}\,[\Pi_C(x_k)]$. Since many projections and proximal maps are known in analytic or computable form, the proximal distance algorithm provides a scalable computational framework for a variety of constraints. For the particular case of sparse linear regression, we implement a projected gradient algorithm known as *iterative hard thresholding* (IHT) for *genome-wide association studies*. A genome-wide association study (GWAS) correlates marker variation with trait variation in a sample of individuals genotyped at a multitude of SNPs (single nucleotide polymorphisms) spanning the genome. The massive amount of data produced in these studies present unique computational challenges. Penalized regression with LASSO or MCP penalties is capable of selecting a handful of associated SNPs from millions of potential SNPs. Unfortunately, type I and type II errors can complicate model selection and obscure the genetic underpinning of a trait. Our parallel implementation of IHT accommodates SNP genotype compression and exploits both multicore CPUs and massively parallel GPUs. This allows statistical geneticists to leverage desktop workstations and to eschew expensive supercomputing resources. We evaluate IHT performance on both simulated and real GWAS data and conclude that it reduces type I and type II errors while maintaining compute speeds competitive with LASSO and MCP.

Doctoral Committee: Ken Lange, Ph.D. (Chair), Van Savage, Ph.D.,
Marc Suchard, M.D., Ph.D., Lieven Vandenberghe, Ph.D.