samsi
NSF·Duke·NCSU·UNC

Optimization Program WISO Workshop
February 8-10, 2017

Lecture: ***Sparse Learning and Distributed PCA with Control of Statistical Errors and Computing Resources***

Speaker: Jianqing Fan

Abstract:

High-dimensional sparse learning and analysis of Big Data data pose significant challenges on computation and communication. Scalable statistical procedures need to take into account both statistical errors and computing resource constraints. This talk illustrate this idea by using two important examples in statistical machine learning. The first one is to solve sparse learning via a computational framework named iterative local adaptive majorize-minimization (I-LAMM) to simultaneously control algorithmic complexity and statistical error when fitting high dimensional sparse models via a family of folded concave penalized quasi-likelihood. The algorithmic complexity and statistical errors are explicitly given and we show that the algorithm achieves the optimal statistical error rate under the weakest signal strength assumptions. The second problem is to study distributed PCA with communication constraints: each node machine computes the top eigenvectors and communicates to the central server; the central server then aggregates the information transmitted from the node machines and conduct another PCA based on the aggregated information. We investigate the bias and variance for such a distributed PCA. We derive the rate of convergence for distributed PCA, which depends explicitly on effective rank, eigen-gap, and the number of machines, and show that the distributed PCA performs as well as the whole sample PCA, even without full access of whole data.