CS369M: Algorithms for Modern Massive Data Set Analysis

Instructor: Michael Mahoney (mmahoney@cs.stanford.edu)

Teaching Assistant: Alex Shkolnik (ads2@stanford.edu)

Office Hours: TBA

Time and Location: MW 11:00-12:15, Building 540 Room 103. (First meeting is Mon, Sept 21, 2009.)

Course Web Page: TBA

Course Requirements: Most likely, three homeworks (ca. 15–20% each), scribe a lecture (ca. 5%), and a research project (ca. 50%, which includes initial proposal, then literature review report, then final report).

Representative topics:

- Introduction and overview—algorithmic and statistical perspectives.
- Randomized algorithms for matrix problems—motivating applications; The Johnson-Lindenstrauss lemma; Random projections; Matrix multiplication and norm estimation; Random sampling of columns and elements from a matrix; Sampling algorithms for L2 regression and relative error low-rank matrix approximation.
- Data analysis and machine learning uses of matrix computations—Algorithmic basics of kernels and machine learning; Basics of manifold-based machine learning; Connections with kernels and eigenfunction computation.
- Algorithmic approches to graph partitioning problems—motivating applications; Flow-based partitioning methods; Spectral-based partitioning methods: Combining spectral and flow-based methods; Local graph partitioning methods. Embeddings and geometric structure related to graphs; Expanders for algorithms and real networks; Connections to spectral clustering in machine learning.
- Novel data-motivated matrix factorizations—Sparse PCA; Maximum margin methods; Matrix rank minimization; Bregmann divergences; CUR and related decompositions; Nyström-based methods.
- Relationship to statistics and large-scale computation—Boosting, ensembles, and relationships with multiplicative updates; Some numerical issues; Some statistical issues; Some large-scale computational and implementation issues.