EE 290: Mathematics of Data Science

Jiantao Jiao

Fall, 2019

TuTh 2:00PM - 3:29PM, Cory 540AB
Web: `https://people.eecs.berkeley.edu/~jiantao/ee290/`
Office Hours: Tu 4:00PM - 5:00PM, Cory 257M   Class Hours: Tu/Th 2-3:29pm
GSI: Banghua Zhu
Course staff e-mail: `ee290-mds [at] lists.eecs.berkeley.edu`

Course Description

This course covers selected topics in the mathematical, statistical, and computational aspects of data science. We characterize the information-theoretic (statistical) limit for inference problems, investigate whether the statistical limits can be attained computationally efficiently, and analyze algorithmic techniques such as spectral methods, semidefinite programming relaxations, kernel methods, wavelet shrinkage. Specific topics will include spectral clustering, planted clique and partition problem, adaptive estimation, sparse PCA, community detection on stochastic block models, nonparametric function estimation and Lepski's method.

Prerequisites

Solid background of probability theory and mathematical statistics (at the level of Stat 135 or EECS 126), convex optimization (at the level of EE 127 or CS 189), and linear algebra. Knowledge about information theory would help but is not a prerequisite.

Tentative Outline

1. *Introduction:* statistical decision theory, comparison of statistical procedures, definitions of optimality, common loss functions
2. *Spectral methods:* preliminaries from linear algebra, perturbation bound
3. *Information theoretic tools:* entropy and mutual information, Kullback–Leibler divergence, total variation, Hellinger distance, hypothesis testing, data processing inequality, rate-distortion function, Shannon lower bound
4. *Planted clique:* degree test, spectral methods

5. *Community detection:* stochastic block models, correlated recovery and mutual information, almost exact and exact recovery, degree corrected block models, first and second moment methods

6. *Semidefinite programming (SDP) relaxation:* KKT conditions and exact recovery threshold, Grothendieck inequality and consequences on clustering, robustness in semi-random models

7. *Computational limits:* Polynomial-time randomized reduction, Planted dense subgraph problem, Sparse PCA

8. *Adaptive estimation:* Lepski's method, Stein's estimator, mean estimation with heavy tails, constrained risk inequality

9. *Nonparametric estimation:* Bias-variance trade-off, kernel-based estimator, limitations of Holder ball assumption, estimation over Sobolev balls, adaptive bandwidth, wavelet shrinkage

Administrivia

1. Grading: 30% participation, 30% homework, 40% final project, 5% lecture scribing (extra points)

2. Participation: class attendance is required. Each enrolled student is expected to scribe the notes for at least one lecture, which is due in one week from the lecture. LaTeX template is available online

3. Homework: three to four problem sets

4. Final project: (group) presentation and report

5. Lecture notes and reading materials will be posted online

6. **Save the date:** in class final project presentations on Dec.3 and Dec. 5