## Overview

This will be a mostly self-contained research-oriented course designed for undergraduate students (but also extremely welcoming to graduate students) with an interest in doing research in theoretical aspects of algorithms that aim to extract information from data. These often lie in overlaps of two or more of the following: Mathematics, Applied Mathematics, Computer Science, Electrical Engineering, Statistics, and/or Operations Research.

The topics covered include:

1. Principal Component Analysis (PCA) and some random matrix theory that will be used to understand the performance of PCA in high dimensions, through spike models.
2. Manifold Learning and Diffusion Maps: a nonlinear dimension reduction tool, alternative to PCA. Semisupervised Learning and its relations to Sobolev Embedding Theorem.
3. Spectral Clustering and a guarantee for its performance: Cheeger's inequality.
4. Concentration of Measure and tail bounds in probability, both for scalar variables and matrix variables.
5. Dimension reduction through Johnson-Lindenstrauss Lemma and Gordon's Escape Through a Mesh Theorem.
6. Compressed Sensing/Sparse Recovery, Matrix Completion, etc. If time permits, I will present Number Theory inspired constructions of measurement matrices.
7. Group Testing. Here we will use combinatorial tools to establish lower bounds on testing procedures and, if there is time, I might give a crash course on Error-correcting codes and show a use of them in group testing.
8. Approximation algorithms in Theoretical Computer Science and the Max-Cut problem.
9. Clustering on random graphs: Stochastic Block Model. Basics of duality in optimization.
10. Synchronization, inverse problems on graphs, and estimation of unknown variables from pairwise ratios on compact groups.
11. Some extra material may be added, depending on time available.

## Open Problems

A couple of open problems will be presented at the end of most lectures. They won't necessarily be the most important problems in the field (although some will be rather important), I have tried to select a mix of important, approachable, and fun problems. In fact, I take the opportunity to present two problems below.

### 0.2.1 Komlós Conjecture

We start with a fascinating problem in Discrepancy Theory.

**Open Problem 0.1 (Komlós Conjecture)** *Given $n$, let $K(n)$ denote the infimum over all real numbers such that: for all set of $n$ vectors $u_1, \ldots, u_n \in \mathbb{R}^n$ satisfying $\|u_i\|_2 \leq 1$, there exist signs $\epsilon_i = \pm 1$ such that*

$$\|\epsilon_1 u_1 + \epsilon_2 u_2 + \cdots + \epsilon_n u_n\|_\infty \leq K(n).$$

*There exists a universal constant $K$ such that $K(n) \leq K$ for all $n$.*

An early reference for this conjecture is a book by Joel Spencer [Spe94]. This conjecture is tightly connected to Spencer's famous *Six Standard Deviations Suffice* Theorem [Spe85]. Later in the course we will study semidefinite programming relaxations, recently it was shown that a certain semidefinite relaxation of this conjecture holds [Nik13], the same paper also has a good accounting of partial progress on the conjecture.

- It is not so difficult to show that $K(n) \leq \sqrt{n}$, **try it!**

### 0.4.2 Matrix AM-GM inequality

We move now to an interesting generalization of arithmetic-geometric means inequality, which has applications on understanding the difference in performance of with- versus without-replacement sampling in certain randomized algorithms (see [RR12]).

**Open Problem 0.2** *For any collection of $d \times d$ positive semidefinite matrices $A_1, \cdots, A_n$, the following is true:*

***(a)***

$$\left\| \frac{1}{n!} \sum_{\sigma \in \mathrm{Sym}(n)} \prod_{j=1}^{n} A_{\sigma(j)} \right\| \leq \left\| \frac{1}{n^n} \sum_{k_1,\ldots,k_n=1}^{n} \prod_{j=1}^{n} A_{k_j} \right\|,$$

*and*

***(b)***

$$\frac{1}{n!} \sum_{\sigma \in \mathrm{Sym}(n)} \left\| \prod_{j=1}^{n} A_{\sigma(j)} \right\| \leq \frac{1}{n^n} \sum_{k_1,\ldots,k_n=1}^{n} \left\| \prod_{j=1}^{n} A_{k_j} \right\|,$$

*where* $\mathrm{Sym}(n)$ *denotes the group of permutations of $n$ elements, and $\|\cdot\|$ the spectral norm.*

Morally, these conjectures state that products of matrices with repetitions are larger than without. For more details on the motivations of these conjecture (and their formulations) see [RR12] for conjecture **(a)** and [Duc12] for conjecture **(b)**.

Recently these conjectures have been solved for the particular case of $n = 3$, in [Zha14] for **(a)** and in [IKW14] for **(b)**.

## References


[Duc12] J. C. Duchi. Commentary on "towards a noncommutative arithmetic-geometric mean inequality" by b. recht and c. re. 2012.

[IKW14] A. Israel, F. Krahmer, and R. Ward. An arithmetic-geometric mean inequality for products of three matrices. *Available online at arXiv:1411.0333 [math.SP]*, 2014.

[Nik13] A. Nikolov. The komlos conjecture holds for vector colorings. *Available online at arXiv:1301.4039 [math.CO]*, 2013.

[RR12] B. Recht and C. Re. Beneath the valley of the noncommutative arithmetic-geometric mean inequality: conjectures, case-studies, and consequences. *Conference on Learning Theory (COLT)*, 2012.

[Spe85] J. Spencer. Six standard deviations suffice. *Trans. Amer. Math. Soc.*, (289), 1985.

[Spe94] J. Spencer. *Ten Lectures on the Probabilistic Method: Second Edition*. SIAM, 1994.

[Zha14] T. Zhang. A note on the non-commutative arithmetic-geometric mean inequality. *Available online at arXiv:1411.5058 [math.SP]*, 2014.

18.S096 Topics in Mathematics of Data Science
Fall 2015