MATH-GA 2830.002 Advanced Topics in Applied Math: MATHEMATICS OF DATA SCIENCE (Fall 2016)

Afonso S. Bandeira
bandeira@cims.nyu.edu
http://www.cims.nyu.edu/~bandeira

September 7, 2016

Lectures: Wednesdays 11am-12.50pm CIWW-517

Office Hours: Tue 10.30am-12.30pm at CIWW1123. You are also welcome to email me and we'll schedule a time to meet.

Course Website: Lecture notes, optional homework, and announcements will be posted at: http://www.cims.nyu.edu/~bandeira/Fall2016.MATH.GA.2830.002.MathDataScience.html

Prerequisites: Working knowledge of linear algebra and basic probability is required. Some familiarity with the basics of optimization and algorithms is also recommended.

0.1 Syllabus

This will be a mostly self-contained research-oriented and fast-paced course designed for PhD 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 (Applied) Mathematics with: Computer Science, Electrical Engineering, Statistics, and/or Operations Research. Each lecture will feature a couple of Mathematical Open Problem(s) with relevance in Data Science. The main mathematical tools used will be Probability and Linear Algebra, and a basic familiarity with these subjects is required. There will also be some (although knowledge of these tools is not assumed) Graph Theory, Representation Theory, Applied Harmonic Analysis, among others. The topics treated will include Dimension reduction, Manifold learning, Sparse recovery, Random Matrices, Approximation Algorithms, Community detection in graphs, and several others.

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. Error-correcting codes. Message Passing Algorithms. Belief Propagation. 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.

0.2 Grading and important dates:

Grading:

- The grade is based on a project. The project (which can be done individually or in groups of two) can be a literature review, but I recommended attempting to do original research, either by trying to make partial progress on (or completely solve!) one of the open problems posed in class (see below), or by pursuing another research direction. A preliminary abstract of the project will be due roughly a month before the end of classes and each group is expected to make a 5 minute presentation on class about their project before the due date.

Important dates (subject to change – please check course website for announcements):

- November 16: A preliminary abstract of the project is due before class this day.
- December 7: Each group will make a short presentation (5 minutes) about their project. If there are too many groups to present on just one class, some groups will present on December 1. The slides for the project are due a day before the presentation. I will merge all of the slides on the same pdf file to minimize the time spent in transitions between groups.
- December 7: The project report is due before class this day.

I am here to help: if you have any question or concern, want to discuss a problem, or brainstorm about any research idea, just stop by during office hours or email me and we'll schedule a time to meet. Also, please let me know of your goals for your project and keep me up to date on your progress on it.

0.3 Open Problems

A couple of open problems will be presented at the end of most lectures, some will be from the open problem list from a previous iteration of this course (they'll have an A after the number) and some will be new (will have a B after the number). 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 (a similar exposition of this problems is also available on my blog [**?**]).

0.3.1 Komlós Conjecture

We start with a fascinating problem in Discrepancy Theory.

Open Problem 0.1 ***A****. 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!**

0.3.2 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 ***A****. 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.