DS.GA 3001 DS-GA 3001 Special Topics in Data Science: MATHEMATICS OF DATA SCIENCE: Graphs and Networks (Spring 2018)

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

Shuyang Ling
sling@cims.nyu.edu
http://www.cims.nyu.edu/~sling

January 25, 2018

Lectures (Afonso): Thursdays 4.55pm-6.35pm at CDS110 (60 5th Av)

Lab (Shuyang): Wednesdays 7.45pm-8.35pm at CDS110 (60 5th Av)

Office Hours Afonso: Thursdays 3.45pm-4.45pm at CDS603 (60 5th Av). You are also welcome to email me and we'll schedule a time to meet at CIWW1123.

Office Hours Shuyang: Wednesday 4.30pm-5.30pm at CIWW1103.

Course Website: Lecture notes, optional homework, and announcements will be posted at: `https://cims.nyu.edu/~bandeira/Spring2018.DS.GA.3001.MathDataScienceGraphs.html`

Piazza: Course announcements will be made via Piazza, be sure to sign up: `https://piazza.com/class/jcs9ru6yxd366d?cid=4#`

Prerequisites: Working knowledge of linear algebra and basic probability is required. Some familiarity with the basics of optimization and algorithms is also recommended, but not required.

0.1 Syllabus

This is part of a two course series on Mathematics of Data Science. Each part can be taken independently, and they can be taken in any order, the other part can be found here. This part focus on algorithms on graphs and networks while the other in high dimensional data. This is a mostly self-contained research-oriented and fast-paced course designed for 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 (Applied) Mathematics with: Computer Science, Electrical Engineering, Statistics, Operations Research and/or Statistical Physics. 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 working knowledge of 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 Random Matrices, Approximation Algorithms, Convex Relaxations, Community detection in graphs, and several others.

The Syllabus includes: Matrix Concentration, Approximation Algorithms and Max-Cut, Community Detection and the Stochastic Block Model, Synchronization Problems and Alignment, Cheeger's Inequality, Semidefinite Programming relaxations, Approximate Message Passing algorithms, and (if time permits) statistical physics heuristics for computational limits of problem on networks.

Please email us if you have any question.

0.2 Grading and important dates:

Grading:

- The grade is based on homework sets (40%) and a project (60%). 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 on the week before Spring Breakand each student is expected to make a 5 minute presentation on class about their project in the last couple of weeks of the class.

  The homework will be roughly bi-weekly and due on Thursday before class. More information and instructions will be included in each problem set.

Important dates (subject to change (in particular depending on the number of different projects) – please check course website for announcements):

- March 8: A preliminary abstract of the project is due before class this day (by email to both instructors).
- April 26 and May 3: Each student will make a short presentation (5 minutes) about their project. Depending on the number of presents, some might have to present on April 19. The slides for the project are due a day before the presentation (by email to both instructors). I will merge all of the slides on the same pdf file to minimize the time spent in transitions between students. For groups of two students, which student should present a different part of the project.
- April 26: The project report is due before class this day (by email to both instructors). If presentations start on April 19, due date will be changed to that, we should know before Spring Break, be sure to check for announcements and sign up for Piazza announcements.

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. There is also a feedback form on the website, be sure to explore the course website.

Lecture notes will be posted or referred to, be sure to check the course website

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 and some will be new. 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 that were also present in the other part of this class, on the Fall of 2016 (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 $u_1, \dots, u_n \in \mathbb{R}^n$ 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.