401-4944-20L Mathematics of Data Science: Problem Set 3

(Spring 2020)

Afonso S. Bandeira
bandeira@math.ethz.ch
https://people.math.ethz.ch/~abandeira/

May 5, 2020

This homework is optional and it won't be graded. If you want to discuss a solution (to make sure it is correct) or want to ask questions about a problem stop by office hours or write a TA or myself an email and we can schedule a time to talk. Date is of last update (e.g. correction of typos)

Try not to look up the answers, you'll learn much more if you try to think about the problems without looking up the solutions. If you need hints, feel free to email me.

Problem 3.1 (Multidimensional Scaling) *Suppose you want to represent n data points in $\mathbb{R}^d$ and all you are given is estimates for their Euclidean distances $\delta_{ij} \approx \|x_i - x_j\|_2^2$. Multiimensional scaling attempts to find an d dimensions that agrees, as much as possible, with these estimates. Organizing $X = [x_1, \ldots, x_n]$ and consider the matrix Δ whose entries are δ_{ij}.*

1. *Show that, if $\delta_{ij} = \|x_i - x_j\|_2^2$ then there is a choice of x_i (note that the solution is not unique, as a translation of the points will preserve the pairwise distances, e.g.) for which*

$$X^T X = -\frac{1}{2} H \Delta H,$$

where $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$.

2. *If the goal is to find points in $\mathbb{R}^d$, how would you do it (keep part 1 of the question in mind)?*

(The procedure you have just derived is known as Multidimensional Scaling)

This motivates a way to embed a graph in d dimensions. Given two nodes we take δ_{ij} to be the square of some natural distance on a graph such as, for example, the geodesic distance (the distance of the shortest path between the nodes) and then use the ideas above to find an embedding in $\mathbb{R}^d$ for which Euclidean distances most resemble geodesic distances on the graph. This is the motivation behind a dimension reduction technique called ISOMAP (J. B. Tenenbaum, V. de Silva, and J. C. Langford, Science 2000).

Problem 3.2 *Given n i.i.d. non-negative random variables $x_1, \dots, x_n$, show that*

$$\mathbb{E} \max_i x_i \lesssim \left(\mathbb{E} \left[x_1^{\log n} \right] \right)^{\frac{1}{\log n}}.$$

Problem 3.3 (Little Grothendieck problem) *Let $C \succeq 0$ (C is positive semidefinite). In this homework you'll show an approximation ratio of $\frac{2}{\pi}$ to the problem*

$$\max_{x_i = \pm 1} \sum_{i,j=1}^{n} C_{ij} x_i x_j.$$

Similarly to `Max-Cut`*, we consider*

$$\max_{\substack{v_i \in \mathbb{R}^n \\ \|v_i\|^2 = 1}} \sum_{i,j=1}^{n} C_{ij} v_i^T v_j.$$

The goal is to show that, for $r \sim \mathcal{N}(0, I_{n\times n})$, taking $x_i^\natural = \operatorname{sign}(v_i^T r)$ a randomized rounding,

$$\mathbb{E} \left[\sum_{i,j=1}^{n} C_{ij} x_i^\natural x_j^\natural \right] \geq \frac{2}{\pi} \sum_{i,j=1}^{n} C_{ij} v_i^T v_j$$

Hints:

1. *The main difficulty is that $\mathbb{E} \left[\operatorname{sign}(v_i^T r) \operatorname{sign}(v_j^T r) \right]$ is not linear in $v_i^T v_j$ and C_{ij} might be negative for some (i,j)'s.*

2. *Show that that $\mathbb{E} \left[\operatorname{sign}(v_i^T r) v_j^T r \right]$ is linear in $v_i^T v_j$. What is it equal to?*

3. *Construct* S *with entries* $S_{ij} = \left(v_i^T r - \sqrt{\frac{\pi}{2}}\,\mathrm{sign}(v_i^T r)\right)\left(v_j^T r - \sqrt{\frac{\pi}{2}}\,\mathrm{sign}(v_j^T r)\right)$

4. *Show that* $\mathrm{Tr}(CS) \geq 0$.