Exercise Set 1

These problems are taken from various sources. Problems marked * are more difficult but also more fun :). If you are done early, find me and discuss; there are always more problems.

k-Center

1 **Another 2-approximation for k-center.** Let r be the optimum radius and consider the following algorithm for the k center problem.

> While there is a remaining (input) point, select it as a center and remove all points within distance $2r$.

Argue that you must select at most k balls (and thus at most k centers).

2 Prove that it is NP-hard to approximate the k-center problem within a factor less than 2.

Hint: Use a reduction from the NP-complete dominating set problem. In the dominating set problem, we are given a graph $G = (V, E)$ and an integer k, and we must decide if there exists a set $S \subseteq V$ of size k such that each vertex is either in S or adjacent to a vertex in S.

k-Means

3 In this exercise, you prove the missing "part" of the analysis of k-means++ seeding. That is, prove the lemma

> If some centers T have already been chosen by k-means++ and $Z \in C_i$ is added next, then
>
> $$\mathbb{E}[cost(C_i, T \cup \{Z\}) \mid T, \{Z \in C_i\}] \leq 8 \cdot cost(C_i, z_i)\,,$$
>
> where $z_i = mean(C_i)$.

Hint: see Lemma 6 in the excellent notes by Dasgupta: *https://cseweb.ucsd.edu/ dasgupta/291-geom/kmeans.pdf*

4 Consider the "bad" instance for k-means++:

- We have k balls each of radius 1 and n/k points.
- The balls have pairwise distance $\Delta \gg 1$.

In contrast to k-means++, show that both Greedy k-Means++ and and D^α sampling with say $\alpha = 4$ give a constant-factor approximation algorithm for these instances when Δ tends to infinity.

(*) Prove the above statement for smaller Δ.

5 (*) Give $\omega(\log k)$ lower bounds for D^α sampling with say $\alpha = 4$ and Greedy k-Means++ with a large parameter t.

Preparation for tomorrow

6 Consider an undirected graph $G = (V, E)$ and let $s \neq t \in V$. Show that there is an s,t-cut of value at most the optimal value of the following program

$$\begin{aligned} \textbf{minimize} \quad & \sum_{\{u,v\}\in E} |x_u - x_v| \\ \textbf{subject to} \quad & x_s = 0, x_t = 1, \text{ and } x_v \in [0,1] \text{ for every } v \in V \end{aligned}$$

The above program can be made linear by introducing a variable y_e for each edge and minimizing $\sum_{e\in E} y_e$ and adding the constraints that $y_e \geq x_u - x_v$ and $y_e \geq x_v - x_u$ for every $e = \{u, v\} \in E$.

Hint: Show that the expected value of the following randomized rounding equals the value of the linear program. Select θ *uniformly at random from* $[0, 1]$ *and output the cut* $S = \{v \in V : x_v \leq \theta\}$.