Exercise Set 2

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.

Explainable Clustering

1 *(half a *)* Consider the special case of the explainable k-median problem in which all centers of the reference clustering have the same pairwise distance Δ. So the cost increase by separating a single point from its closest center is at most Δ in this case.

Show that there is an explainable clustering whose cost is at most $O(\log k)$ times the cost of the reference clustering.

Hint: How many random cuts do you need to separate all centers from each other?

Explainable k-means. The explainable k-means problem is equivalent to the explainable k-median problem except that the distance between two points $x, y \in \mathbb{R}^d$ is now $\|x - y\|_2^2$ instead of $\|x - y\|_1$. As we will see, this "small" change greatly impacts the price of explainability.

2 **Lower bound for k-means** Consider the following reference clustering for k-means:

- There are $d = k!$ dimensions, one for each permutation $\pi_\ell : [k] \to [k]$ of the centers.
- The position of center i in dimension ℓ is $\pi_\ell(i)$.
- For each center i and dimension ℓ there are two points. Both points are colocated with i along all dimensions except for dimension ℓ. In dimension ℓ, one of the points is located at $\pi_\ell(i) - 1$ and the other at $\pi_\ell(i) + 1$.

2a Observe that this reference clustering costs $2k \cdot d$.

2b Argue that the pairwise distance between any two centers is $\Omega(dk^2)$.

2c Note that any axis-aligned cut will separate one point from its closest center and will thus cause a cost increase of dk^2 (which one can see implies a price of explainability of k).

3 **Upper bound for 2-means.** Show that k-means with $k = 2$ has constant price of explainability.

How is the price of explainability impacted for 2-means if we allow for general hyperplane cuts instead of only axis-aligned cuts?

Exponential Clocks

4 **Max of identically distributed exponential variables.** Suppose $X_1, X_2, \ldots, X_k \sim \exp(1)$. Show that $\mathbb{E}[\max(X_1, \ldots, X_k)] = H_k = 1/k + 1/(k-1) + \ldots + 1$.

Solution is here: *https://mikespivey.wordpress.com/2013/05/13/expectedmaxexponential/*

5 *(* *)* **Beautiful rounding of set cover LP.** The input for a set cover instance consists of a universe of $E = \{e_1, \ldots, e_n\}$ n elements and m subsets $\mathcal{S} = \{S_1, \ldots, S_m\} \subseteq 2^E$. Each set $S \in \mathcal{S}$ is associated with a cost c_S. The standard linear programming relaxation has a variable x_S for each set $S \in \mathcal{S}$ and the goal is to minimize $\sum_{S \in \mathcal{S}} c_S x_S$, subject to $\sum_{S: e \in S} x_S \geq 1$ for all $e \in E$. Show that the following algorithm returns an integral set cover of cost at most $1 + \ln(n)$ the cost of the linear programming solution. Specifically, prove that each set is taken with probability at most $(1 + \ln|S|) \cdot x_S$.

Algorithm 5: Set-Cover($\{x_S\}_{S \in \mathcal{S}}$).

1 choose i.i.d random variables $Z_S \sim \exp(1)$, $\forall S \in \mathcal{S}$.
2 output $\cup_{e \in E} \operatorname{argmin} \left\{ \frac{Z_S}{x_S} \ : \ e \in S \right\}$.

For the answer, see Appendix A in one of my favorite papers, "Simplex Partitioning via Exponential Clocks and the Multiway-Cut Problem" by Niv Buchbinder, Seffi Naor, and Roy Schwartz.

Preparation for tomorrow

6 **Ski rental with predictions.** In the classic ski rental problem, you will ski for an unknown number of days. Renting costs 1 EUR a day and buying costs B EUR. The optimal deterministic strategy is to rent for B days and then buy. In that way, you will never pay more than twice the cost of the optimal strategy that knew the number of days you would ski in advance.

Now suppose that you have access to an untrusty prediction P on the number of days you will ski. Devise a strategy, parameterized by $\lambda \in (0, 1)$ with the following guarantee:

- If the prediction P is correct, you pay at most $1 + \lambda$ times the optimal cost.
- If the prediction P is incorrect (even adversarial), you pay at most $O(1/\lambda)$ times the optimal cost.