Partner 1
Partner 2
Partner 3

CSCI 1550 / 2540
March 6th, 2025

Homework 3

Due: March 31, 2025
Remember to show your work for each problem to receive full credit.

Problem 1 [40 points]

k-Mean Clustering: Let $S = \{x_1, \dots, x_n\}$ be a set of points in a high dimension Euclidian space. A k-mean clustering of S is a partition of S into k disjoint sets (clusters), $C_1, \dots, C_k$, with centers $c_1, \dots, c_k$,

$$c_j = \frac{1}{|C_j|} \sum_{x \in C_j} x.$$

$||x||^2$ is the standard Euclidean norm.

The cost of k-mean a clustering $C_1, \dots, C_k$ solution, with centers $c_1, \dots, c_k$, is

$$Cost(C_1, \dots, C_k) = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - c_i||^2$$

where $||x||$ is the standard Euclidean norm. (If $x \in R^m$ then $||x||^2 = \sum_{i=1}^{m} x_i^2$.)

The goal is to find a clustering that minimizes the cost function. (In general the problem is NP-complete, but there are efficient approximations.) Here we show that we can speed-up the computation by solving the problem in a lower dimension space.

1. Prove that for any set of points $C \subset R^m$:

$$\sum_{x \in C} \sum_{y \in C} ||x - y||^2 = 2|C| \sum_{x \in C} ||x - c||^2, \quad \text{where} \quad c = \frac{1}{|C|} \sum_{x \in C} x.$$

   [Hint: prove first for dimension 1.]

2. Apply dimension reduction (JL-Lemma) to show that a cost of a solution for k-mean clustering in a low dimension vector space is within $1 \pm \epsilon$ times the cost of the corresponding solution in the original space. How small can the dimension of the projected space be?

Problem 2 [40 points]

1. Let $X(m, \epsilon)$ be a random variable with a Binomial distribution $B(m, \frac{1}{2} + \epsilon)$, for some $0 < \epsilon \leq \frac{1}{2}$. Prove that for any even $m \geq 2$ there is an $0 \leq \epsilon \leq \frac{1}{2}$, such that $Pr(X(m, \epsilon) \leq m/2) \geq 1/4$. [Hint: It's hard to compute an explicit value of ϵ that satisfies the requirement. Instead, write $Pr(X(m, \epsilon) \leq m/2)$ as a function of ϵ and use simple calculus to show that there is a value of ϵ that satisfies the requirement.]

2. Assume that we have k arms. $k - 1$ of the arms give 0 with probability 1/2 and 1 with probability 1/2. The k-th arm gives 0 with probability $1/2 - \epsilon$ and 1 with probability $1/2 + \epsilon$. Show that the Explore and Commit has linear expected regret for any m and some choice of ϵ.

Problem 3 [40 points]

Consider a doubling version of the UCB algorithm:

Algorithm UCB(α):

```
1. For all $i \in [k]$ activate arm $i$ once, update $M_i$ and set $N_i = 1$.
2. $s = k$
3. For $t = 1, \ldots, \log T$ do:
    (a) Activate $2^t$ times arm $j = \arg\max_{i \in [k]} \left( M_i + \sqrt{\frac{\alpha \log s}{2N_i}} \right)$.
    (b) Update $M_j$, and set $N_j = N_j + 2^t$, and $s = s + 2^t$
```

Prove that the expected regret of this algorithm is no more than:

$$\sum_{i \in [k],\ \Delta_i > 0} \left(\frac{2\Delta_i}{1 - 2^{-(\alpha-1)}} + \frac{4\alpha \log T}{\Delta_i} \right)$$

Problem 4 [25 points]

A parking-lot attendant has mixed up n keys for n cars. The n car owners arrive together. The attendant gives each owner a key according to a permutation chosen uniformly at random from all permutations. If an owner receives the key to their own car, they take it and leave; otherwise, they return the key to the attendant. The attendant now repeats the process with the remaining keys and car owners. This continues until all owners receive the keys to their cars.

Let R be the number of rounds until all car owners receive the keys to their cars. We want to compute $\mathbb{E}[R]$. Let X_i be the number of owners who receive their car keys in the ith round.

1. Prove that

$$Y_i = \sum_{j=1}^{i} (X_j - \mathbb{E}(X_j \mid X_1, \ldots, X_{j-1}))$$

is a martingale with respect to the X_i's.

2. Use the martingale stopping theorem to compute $\mathbb{E}[R]$.