# Principles of Online Decision-Making (CS-303)
## Problem Set 1

## Problem 1

(a) We had derived the optimal solution for the secretary problem in class, and seen that (asymptotically for large $n$), the right approach is to observe the maximum score over the first $n/e$ items, and then to stop as soon as this maximum is exceeded.

Suppose now that you know that the score distribution: $x_i \sim \text{unif}(0,1)$, i.i.d. How would you modify the algorithm to maximize the probability of hitting the winner? Will the success probability be higher/lower/the same than for the original version?

(b) Suppose again the same score distribution as above, and it is again known to the player. But assume the player tries to maximize $\mathbb{E}x_Z$ instead of the probability of finding the absolute best. Can you figure out a good policy? You may want to run some simulations to confirm.

## Problem 2

Prove Lemma 4.4 in L&S.

## Problem 3

In the given codes, there are two python files: `bandit_algorithm.py` and `bandit_environment.py`. The first file implements the Explore-then-commit algorithm and the $\epsilon$-greedy algorithm. The second file implements the bernoulli bandit environment, and gaussian bandit environment, that you have to complete. We provide `experiment_1.ipynb` to run the algorithms on the environments. Alternatively, you can also run the same code with `experiment_1.py`.

(a) Explore the explore-then-commit algorithm by setting different random seeds.

(b) Try the explore-then-commit algorithm with several seed to find a failing case *i.e.,* the case where the explore-then-commit algorithm chooses the sub-optimal arm at the end of the exploration phase (time step 200). Let $\mu_i$ be the expected reward of arm $i$. Let $\hat{\mu}_i$ be the empirical mean of arm $i$ at the end of the exploration phase. In the code arm 1 is the optimal arm. Under which condition on $\mu_i$ and $\hat{\mu}_i$, does the explore-then-commit algorithm fail?

(c) Let $\hat{p}_{sub}^{200}$ be the empirical probability that the explore-then-commit algorithm chooses the sub-optimal arm at the end of the exploration phase (time step 200). Compute $\hat{p}_{sub}^{200}$ by simulation.

(d) Change the sub-optimality gap $\Delta$ and make a plot to see the evolution of $\hat{p}_{sub}^{200}$ as a function of $\Delta$.

(e) Change the exploration period and make a plot to see the evolution of $\hat{p}_{sub}^{200}$ as a function of the exploration period.

(f) One can consider a variant of the explore-then-commit algorithm that switches arms during the commitment phase if the empirical mean of the arm being pulled becomes lower than that of the other arm. We call this variant the "explore-then-weak-commit" algorithm. Specifically, explore-then-weak-commit algorithm works as follows (with the same notations as in Algorithm 1 on pp. 92 of "Bandit Algorithms" by Lattimore and Szepesvári):

1. Input $m$ (exploration period per arm)

2. In round $t$ chooses action

$$A_t = \begin{cases} (t \mod k) & \text{if } t \leq mk; \\ \arg\max_{i \in [k]} \hat{\mu}_i(t-1) & \text{if } t > mk. \end{cases}$$

Implement this variant.

(g) Try the explore-then-weak-commit algorithm with seed 14766, explain the behavior you observe with this seed. Give an example of conditions on $\mu_i$ and $\hat{\mu}_i$, under which this kind of behavior can be observed.

(h) Can there any runs where the explore-then-weak-commit algorithm performs better than the original explore-then-commit algorithm. If there are some such cases, give the example of conditions of such cases in terms of $\mu_i$ and $\hat{\mu}_i$. (Recall that $\mu_i$ and $\hat{\mu}_i$ are defined in (b).) On contrary, is there any case where original the explore-then-commit algorithm performs better than the explore-then-weak-commit algorithm? If there is such a case, give an example condition on $\mu_i$ and $\hat{\mu}_i$.

(i) Explore the $\epsilon$-greedy algorithm algorithm by setting different random seeds.

(j) Try the $\epsilon$-greedy algorithm with seed 4195. Are the results similar to those with other seeds? If not, explain what happens here.

(k) Let $\hat{q}_{sub}^{200}$ be the empirical probability that the $\epsilon$-greedy algorithm estimates the sub-optimal arm (incorrectly) as the optimal arm at time step 200.

(l) Change the sub-optimality gap $\Delta$ and make a plot to see the evolution of $\hat{q}_{sub}^{200}$ as a function of $\Delta$.

(m) Change the exploration period and make a plot to see the evolution of $\hat{q}_{sub}^{200}$ as a function of the $\epsilon$.