Assignment 3

Electrical and Electronics Engineering Department, Bilkent University

Instructor: Muhammed O. Sayin

Posted on Nov. 3 **Due Date: Nov. 20**

Content: Multi-armed Bandits and Temporal Difference Learning
Recommended Reading: Sutton and Barto: Chapters 2, 5, and 6

Problem 1. (20pt) Consider arm 1 and arm 2 generating i.i.d. samples from Bernoulli distributions, resp., $\mathrm{Ber}(0.4)$ and $\mathrm{Ber}(0.8)$.

a) Which arm is the best one with respect to the expected amount of rewards that can be collected over T-length horizon?

b) If we set $\delta = 0.05$, compute N (the exploration length for each arm) in the Explore & Commit Algorithm to guarantee $O(T^{2/3})$ regret with high probability over T-length horizon.

Problem 2. (50pt) For the multi-armed bandit problem described in Problem 1, we will **implement** the Explore & Commit Algorithm (in any computational tool or software you are confident with) and challenge the performance of the exploration length computed in Part b) of Problem 1 (based on the regret analysis). To this end, set $T = 500$ and consider the scenarios where

$$N = 1, 6, 11, 16, 21, 26, 31, 41, \text{ and } 46.$$

a) Implement and simulate the Explore & Commit Algorithm for each N listed above across at least 1000 independent trials.

- **Plot** N vs the total average rewards obtained with the Explore & Commit Algorithm with the associated exploration length.
- Which N is the best one according to your plot? Compare it with the one computed in Part b) of Problem 1. If they are different, can you explain why this was the case?

b) The guarantee of $O(T^{2/3})$ regret holds with at least probability $1 - K\delta$, where K is the number of arms, due to the conditioning the differences between the sample and true means based on the bound attained according to the Hoeffding inequality.

- **Re-run** and plot the simulations of Part a) by only focusing on the cases where the differences between the sample and true means satisfy the bound?
- **Report** the percentage of trials where this condition holds compared to the lower bound $1 - K\delta$ for each N.
- Which N is the best one according to your new plot? Compare it with the ones computed in Part b) of Problem 1 and Part a) of Problem 2.

Problem 3. (30pt) For the multi-armed bandit problem described in Problem 1, we will **implement** the ϵ-greedy algorithm (in any computational tool or software you are confident with) and compare its performance with the Explore & Commit Algorithm. In the Explore & Commit Algorithm, we explore for the initial KN stages out of T stages. On the other hand, in the ϵ-greedy algorithm, we explore in $\epsilon \times T$ stages out of T stages on average. Therefore, we will examine whether the performance of the ϵ-greedy algorithm is similar with the Explore & Commit Algorithm where $N \approx \epsilon T/K$. To this end, set $T = 500$ and consider the scenarios where $\epsilon \in \{KN/T \in (0,1) : N = 1, 6, 11, 16, 21, 26, 31, 41, 46\}$ as in Problem 2.

Implement and simulate the ϵ-greedy algorithm for each ϵ listed above across at least 1000 independent trials.

- **Plot** ϵ vs the total average rewards obtained with the ϵ-greedy algorithm with the associated exploration probability.
- Which ϵ is the best one according to your plot? Compare it with the plots drawn in Problem 2.