Indian Institute of Technology, Kanpur
Department of Electrical Engineering
Introduction to Reinforcement Learning (EE932)
Theory Assignment 1

Deadline: 20th April 2024 | 2023-24 Quarter 4 | Max mark: 15

- **9 objective questions followed by 3 subjective questions.**

1 Objective Questions [9 Marks]

1. We have labelled training data in
   (a) Unsupervised Learning
   (b) Supervised learning **(1)**

2. In the reinforcement learning framework, who generates the rewards for the actions taken?
   (a) Agent
   (b) Environment **(1)**

3. In the multi-arm bandits, each arm has an underlying probability distribution from which the rewards are generated.
   (a) True
   (b) False **(1)**

4. Consider a 3-arm bandit problem with arms a, b, c. In Explore-Then-Commit algorithm, which arm will be played at round t=10, if each arm is explored thrice in the order $a, a, a, b, b, b, c, c, c$ and the observed rewards are $1, 2, 3, 4, -4, 3, 9, -10, 2$.
   (a) Arm a
   (b) Arm b
   (c) Arm c
   (d) A random arm is chosen **(1)**

5. Repeat the above problem for an epsilon greedy algorithm, assuming the first 9 rounds of data are the same as above. Consider $\epsilon = 0.3$. At the round $t = 10$, what is the probability of playing arm a?
   (a) 0.7
   (b) 0.8
   (c) 0.1
   (d) 0.3 **(1)**

6. The first 4 rounds of a UCB algorithm are: (a, +1), (b, +2), (c, +3), (c, +1.5). Which arm will be picked in the 5th round? Assume $T = 1000$. Use natural logarithm (with base e) for calculations.
   (a) Arm a
   (b) Arm b
   (c) Arm c
   (d) Any arm at random. **(1)**

7. Let there be two arms a, b. Assume total round $T = 5$. Assume that the true means (which are unknown to the agent) of the arms be $\mu(a) = 4$, $\mu(b) = 3$. If I follow an algorithm that picks arms uniformly at random in all the rounds, what is the expected regret performance of that algorithm? $T\mu^* - \sum_{t=1}^{T} \mathbb{E}[R_t]$
   (a) 3.5
   (b) 2.5
   (c) 15
   (d) 5 **(1)**

8. In the epsilon-greedy algorithm, which of the following two approaches make sense?
   (a) Gradually increase the value of ϵ
   (b) Gradually decrease the value of ϵ **(1)**

9. When we use clustering technique to solve the conextual bandits, we assume all the user in a cluster to have same expected rewards.
   (a) True
   (b) False **(1)**

2 Subjective Questions [6 Marks]

1. Consider a contextual bandits scenario in which the true mean $\mu_a(x) = {\theta_a}^T x$ of an arm a is a linear function of the context vector x. Here θ_a and x are $n \times 1$ vectors if n is the number of features in the context vector. Assume that we have two arms a_1 and a_2, and the samples *(context, action, reward)* observed by the agent in the first 6 rounds are as follows:

$$
\begin{aligned}
&(\begin{bmatrix}1\\3\end{bmatrix}, a_1, r = 17),\\
&(\begin{bmatrix}7\\13\end{bmatrix}, a_2, r = 2),\\
&(\begin{bmatrix}5\\7\end{bmatrix}, a_1, r = 2),\\
&(\begin{bmatrix}5\\3\end{bmatrix}, a_2, r = 1),\\
&(\begin{bmatrix}11\\13\end{bmatrix}, a_1, r = 23),\\
&(\begin{bmatrix}5\\7\end{bmatrix}, a_2, r = 9)
\end{aligned}
$$

   If the context seen in the 7^{th} round is $\begin{bmatrix}2\\1\end{bmatrix}$, what arm is played by the agent in that round if it uses an ETC policy? Upload an image showing your work. (2)

2. If LinUCB algorithm is used, what are the UCB scores of arm a_1 and arm a_2 for the above problem w.r.t to the context seen in the 7th round? Upload an image showing your work. (2)

3. In the contextual bandits, we have used a feature vector to represent users since there are too many users to handle individually. Now consider a case where the number of arms is also too large. Do you have any suggestions for handling this situation? Please type your answer and upload it. (2)