Assignment 2

Reinforcement Learning

Prof. B. Ravindran

1. In the UCB 1 algorithm, how is it ensured that the confidence bounds around the estimated values for each arm shrink?

   (a) the intervals do not shrink

   (b) the confidence intervals of each arm are made to shrink after every arm selection

   (c) in calculating the bounds we use the number of times an arm is selected in the denominator

   (d) on selecting an arm its interval is shrunk by half every time

2. After 12 iterations of the UCB 1 algorithm applied on a 4-arm bandit problem, we have $n_1 = 3$, $n_2 = 4$, $n_3 = 3$, $n_4 = 2$ and $Q_{12}(1) = 0.55$, $Q_{12}(2) = 0.63$, $Q_{12}(3) = 0.61$, $Q_{12}(4) = 0.40$. Which arm should be played next?

   (a) 1

   (b) 2

   (c) 3

   (d) 4

3. In the proof of the UCB 1 algorithm, we had the following expression:

   $$\{min_{0<s<m} \left(Q_s(a^*) + C_{m-1,T_{a^*}(s)}\right) \leq max_{l \leq s_i \leq m}(Q_{s_i}(i) + C_{m-1,T_i(s_i)})\}$$

   What does this expression stand for?

   (a) the number of instances where the minimum value among the upper confidence bound values of an optimal arm up till the m^{th} time step is less than or equal to the maximum value among the upper confidence bound values of the i^{th} arm between the times steps l and m

   (b) the condition that the minimum value among the upper confidence bound values of an optimal arm up till the m^{th} time step is less than or equal to the maximum value among the upper confidence bound values of the i^{th} arm between the times steps l and m

   (c) the condition that there exists an upper confidence bound value of an optimal arm that is smaller than the corresponding upper confidence value of the chosen arm i

   (d) the number of instances where the upper confidence bound value of an optimal arm is smaller than the corresponding upper confidence value of the chosen arm i

4. In the initial stage of the UCB 1 algorithm, each arm is selected one time. Thereafter, the arm that is selected depends on the arm with the maximum upper confidence bound. Suppose that one of the arms has not been selected after its initial selection in the first stage of the algorithm. What happens to the upper confidence bound of this arm?

   (a) it increases

   (b) it decreases

   (c) it remains constant until it is selected

5. Suppose that we apply the naive PAC algorithm on a 10-arm bandit problem where the required (ϵ, δ) values are: $\epsilon = 0.5$ and $\delta = 0.05$. How many iterations would the algorithm take to output an arm selection? (Note: each arm is sampled $\frac{2}{\epsilon^2} ln\left(\frac{2k}{\delta}\right)$ times).

   (a) 48

   (b) 21

   (c) 210

   (d) 480

6. In the proof of the Naive PAC algorithm, we have the expression

$$P(Q(a') > Q(a^*)) \leq P(Q(a') > q_*(a') + \epsilon/2 \text{ OR } Q(a^*) < q_*(a^*) - \epsilon/2)$$

   Why is the LHS "less than or equal to" the RHS here?

   (a) because the occurrence of at least one of the events on the RHS is a necessary but not sufficient condition for the event on the LHS to occur

   (b) because the event on the LHS does not require that the events on the RHS occur

   (c) because the event on the LHS requires that both the events on the RHS occur

7. Suppose we have a 64-arm bandit problem. We apply the median elimination algorithm where the (ϵ, δ) values are: $\epsilon = 0.5$ and $\delta = 0.01$. How many times do we sample arms in the first round? (Note: in each round, each arm is sampled $\frac{1}{\epsilon_l^2/2} ln\left(\frac{3}{\delta_l}\right)$ times).

   (a) 52404

   (b) 818

   (c) 52416

   (d) 819

8. Continuing with the previous example, what is the number of samples in the third round? Also, what is the total number of rounds required to identify the (ϵ, δ)-optimal arm?

   (a) 50384, 6

   (b) 50384, 7

   (c) 50378, 6

   (d) 201514, 7

9. Consider a bandit problem in which there is a single optimal arm, a^*. Applying the median elimination algorithm to this problem, is it possible that arm a^* is eliminated in the first round? In case it is possible, does this mean that the algorithm cannot output an arm that is ϵ-close to a^*?

   (a) no

   (b) yes, no

   (c) yes, yes

10. We know that there is an exploration/exploitation dilemma in reinforcement learning problems. Considering the Thompson sampling algorithm for solving bandit problems, which step of the algorithm ensures that we perform exploration?

    (a) initialisation of each arm's reward distribution

    (b) updating the reward distribution based on observing actual reward for an arm

    (c) sampling the expected payoff of each arm from the corresponding reward distribution

    (d) identifying the correct arm given the set of expected payoffs for each arm sampled from each arm's reward distribution