Reinforcement Learning - Monsoon 2022
IIITD
Assignment 1
Ojus Singhal | 2020094 | ojus20094@iiitd.ac.in

Answer 1

Using Sample Mean:

$$Q_{10}(arm1) = \frac{-4+1}{2} = -1.5$$
$$Q_{10}(arm2) = \frac{-5+9}{2} = 2$$
$$Q_{10}(arm3) = \frac{9+10+2}{3} = 7$$
$$Q_{10}(arm4) = \frac{5+2}{2} = 3.5$$

Using Exponential Weighted Averages:

$$Q_{10}(arm1) = 3.79$$
$$Q_{10}(arm2) = 6.93$$
$$Q_{10}(arm3) = -2.545$$
$$Q_{10}(arm4) = 0.65$$

Dependency of the sample mean on initial Q values:

$$Q_n(a) = \frac{n-1}{n} * Q_{n-1} + \frac{1}{n} * R_n$$

Plugging $n = 1$, we get
$Q_1(a) = R_1$

Hence, the sample mean isn't affected by the initial Q values.

Dependency of exponentially weighted averages on initial Q values:

$$Q_n = Q_{n-1} + \alpha * (R_n - Q_{n-1})$$

Plugging $n = 1$, we get
$Q_1 = Q_0 + \alpha * (R_1 - Q_0)$

This shows that exponentially weighted averages are dependent on initial estimates.

Answer 5

The UCB spikes are observed at the $(K+1)^{th}$ time-step where K is the total number of arms. The reason for this spike is that the UCB algorithm explores all the arms once in the first K time steps and takes a greedy action at the $(K+1)^{th}$ time step, causing the spike.

After the $(K+1)^{th}$ time-step, other non-greedy actions are again preferred due to the relatively high "variance" of these other actions, hence causing the drop after the peak.