Shaurya goel
2016194

PAGE NO.:

RL Theory Assignment-1

Ans 2.6 Initial action-value estimates are equal & very large. So, there is equal probability to pick ~~the first~~ any bandit. The reward it produces would most likely be lower than the initial estimate. Hence action-value estimate would reduce. This will happen for 10 times for each bandit. on 11^{th} step we would chose the highest value estimate bandit which would likely be the optimal action. Hence, we would get a high reward & observe a spike.

Ans 2.7

$$\beta_n = \frac{\alpha}{\bar{o}_n} \quad , \quad \bar{o}_n = \bar{o}_{n-1} + \alpha(1 - \bar{o}_{n-1}), \quad \bar{o}_0 = 0$$

(β_n: step size)

We know that -

$$Q_{n+1} = Q_n + \beta_n(R_n - Q_n)$$
$$= Q_n(1-\beta_n) + \beta_n R_n$$
$$Q_{n+1} = \left[Q_{n-1} + \beta_{n-1}(R_{n-1} - Q_{n-1})\right](1-\beta_n) + \beta_n R_n$$
$$\Rightarrow Q_{n+1} = Q_{n-1}(1-\beta_n)(1-\beta_{n-1}) + (1-\beta_n)\beta_{n-1}R_{n-1} + \beta_n R_n$$

As we want to prove that Q_{n+1} is independent of Q_1, we will look only at coefficient of Q_i

$\Rightarrow$ Q_{n+1} depends on $(1-\beta_1)(1-\beta_2)\cdots(1-\beta_n)Q_1$ — (1)

now, $\beta_1 = \frac{\alpha}{\bar{o}_1} = \frac{\alpha}{\bar{o}_0 + \alpha(1-\bar{o}_0)} = 1$

Put in eq. (1) we see that coefficient of Q_1 is 0
$\Rightarrow$ Q_{n+1} is independent of Q_1 & only depends on $R_1, R_2 \cdots R_n$

$$Q_{n+1} = \beta_1(1-\beta_2)(1-\beta_3)\cdots(1-\beta_n)R_1 + \beta_2(1-\beta_3)\cdots(1-\beta_n)R_2 \cdots + \beta_n R_n$$

$$= \sum_{i=1}^{n} \beta_i R_i \prod_{j=i+1}^{n}(1-\beta_j)$$