Quiz# 1 : E1 277: Reinforcement Learning

Name: **Department:**
S.R.No.: **Programme:**

1. Consider a multi-armed bandit with three arms A, B, and C. The decision maker pulls these arms in the order A, B, C, B, A, C, B and gets the rewards $5, 4, 1, 6, 3, 7, 2$, respectively. Assuming initial Q-values for each arm as zero, find the final values of $Q(A)$, $Q(B)$ and $Q(C)$? (2 marks)

- $N(A) = 1$

  $Q(A) \leftarrow Q(A) + \frac{1}{N(A)}(R - Q(A))$

  $= 0 + 1(5-0) = 5$

- $N(B) = 1$

  $Q(B) \leftarrow Q(B) + \frac{1}{N(B)}(R - Q(B)) = 4$

- $N(C) = 1$

  $Q(C) \leftarrow Q(C) + \frac{1}{N(C)}(R - Q(C)) = 1$

- $N(B) = 2$

  $Q(B) \leftarrow Q(B) + \frac{1}{N(B)}(R - Q(B))$

  $= 4 + \frac{1}{2}(6 - 4) = 5$

- $N(A) = 2$

  $Q(A) \leftarrow Q(A) + \frac{1}{N(A)}(R - Q(A))$

  $= 5 + \frac{1}{2}(3 - 5) = 4$

- $N(C) = 2$

  $Q(C) \leftarrow Q(C) + \frac{1}{N(C)}(R - Q(C))$

  $= 1 + \frac{1}{2}(7 - 1) = 4$

- $N(B) = 3$

  $Q(B) \leftarrow Q(B) + \frac{1}{N(B)}(R - Q(B))$

  $= 5 + \frac{1}{3}(2 - 5) = 4$

Final values:

$Q(A) = Q(B) = Q(C) = 4.$

2. Consider a finite horizon MDP for which the single stage costs at each of the times $0, 1, \ldots, N-1$ are the same, i.e., the functions $g_0 = g_1 = \cdots = g_{N-1} \triangleq g$. Let the terminal cost be g_N and it only depends on the terminal state as before. Assume now that $J_{N-1}(x) \geq g_N(x)$, $\forall x \in S$ (where S denotes the state space). Show that this implies $J_k(x) \geq J_{k+1}(x)$, for all $k = 0, 1, \ldots, N-1$ and all $x \in S$. (3 marks)

As per DP algorithm,

$$J_N(x) = g_N(x) \quad \forall x.$$

$$J_{N-2}(x) = \min_{a \in A(x)} E_Y \left[g(x,a,Y) + J_{N-1}(Y) \right]$$

$$\geq \min_{a \in A(x)} E_Y \left[g(x,a,Y) + J_N(Y) \right]$$

$$= J_{N-1}(x), \quad \forall x.$$

Proceeding similarly, we obtain

$$J_k(x) \geq J_{k+1}(x) \quad \forall k = 0, 1, \ldots, N-1$$

and all x.