Assignment 1, Q4

Majid Ghasemi

September 2024

1 Multi-Armed Bandits – Q4

1.1 Part (a)

As stated in the question, we are following an $\epsilon-$greedy policy, and we initialize Q(a) prior to start with 0. Also, as we are using sample-average action-value estimates, we are going to update the values with the following equation:

$$Q(a) = Q(a) + \frac{1}{n}(R_n - Q(a)) \tag{1}$$

Where n shows the number of times a specific action (bandit) is taken.
We are to fill out the Q-estimates for each action before each action is taken from $t = 1$ to $t = 6$.
Here is the step-by-step completion:
1. **At** $t = 1$**:** Before any actions are taken, all Q-values are initialized to 0.

t	$Q(1)$	$Q(2)$	$Q(3)$	$Q(4)$
1	0	0	0	0

2. **After action** $A_1 = 3$ **with reward** $R_1 = 1$**:**
Update $Q(3)$ according to Eq. 1:

$$Q(3) = 0 + \frac{1}{1}(1 - 0)$$

Before $t = 2$**:**

t	$Q(1)$	$Q(2)$	$Q(3)$	$Q(4)$
2	0	0	1	0

3. **After action** $A_2 = 1$ **with reward** $R_2 = 2$**:**
Update $Q(1)$ according to Eq. 1:

$$Q(1) = 0 + \frac{1}{1}(2 - 0)$$

Before $t = 3$**:**

t	$Q(1)$	$Q(2)$	$Q(3)$	$Q(4)$
3	2	0	1	0

4. **After action** $A_3 = 3$ **with reward** $R_3 = 2$**:**
Update $Q(3)$ according to Eq. 1:

$$Q(3) = Q_{\text{old}}(3) + \frac{1}{N(3)}\left(R_3 - Q_{\text{old}}(3)\right) = 1 + \frac{1}{2}(2 - 1) = 1 + 0.5 = 1.5$$

Before $t = 4$**:**

t	$Q(1)$	$Q(2)$	$Q(3)$	$Q(4)$
4	2	0	1.5	0

5. **After action $A_4 = 2$ with reward $R_4 = 1$:**
Update $Q(2)$ according to Eq. 1:

$$Q(2) = 0 + \frac{1}{1}(1 - 0)$$

Before $t = 5$:

t	$Q(1)$	$Q(2)$	$Q(3)$	$Q(4)$
5	2	1	1.5	0

6. **After action $A_5 = 3$ with reward $R_5 = 0$:**
Update $Q(3)$ incrementally w.r.t Eq. 1:

$$Q(3) = Q_{\text{old}}(3) + \frac{1}{N(3)}\left(R_5 - Q_{\text{old}}(3)\right) = 1.5 + \frac{1}{3}(0 - 1.5) = 1.5 - 0.5 = 1$$

Before $t = 6$:

t	$Q(1)$	$Q(2)$	$Q(3)$	$Q(4)$
6	2	1	1	0

Now, we can create the complete table from $t = 1$ to $t = 6$:

t	$Q(1)$	$Q(2)$	$Q(3)$	$Q(4)$
1	0	0	0	0
2	0	0	1	0
3	2	0	1	0
4	2	0	1.5	0
5	2	1	1.5	0
6	2	1	1	0

1.2 Part (b)

We need to determine at each time step whether the agent **chose to explore** (the ε case) or **might have chosen to exploit** (the greedy case).

1. **At $t = 1$:**
   (a) **Q-values before action:** All are 0.
   (b) **Action taken:** $A_1 = 3$.
   (c) **Analysis:** All actions have the same Q-value (0). The agent could have either explored or exploited (since any action is as good as any other). **Cannot be told** whether it was exploration or exploitation.

2. **At $t = 2$:**
   (a) **Q-values before action:** $Q(1) = 0$, $Q(2) = 0$, $Q(3) = 1$, $Q(4) = 0$.
   (b) **Action taken:** $A_2 = 1$.
   (c) **Greedy action:** Action 3 (highest Q-value of 1).
   (d) **Analysis:** The agent chose action 1, which is not the greedy action. **The agent explored**.

3. **At $t = 3$:**
   (a) **Q-values before action:** $Q(1) = 2$, $Q(2) = 0$, $Q(3) = 1$, $Q(4) = 0$.

(b) **Action taken:** $A_3 = 3$.

(c) **Greedy action:** Action 1 (highest Q-value of 2).

(d) **Analysis:** The agent chose action 3, which is not the greedy action. **The agent decided to explore**.

4. **At** $t = 4$**:**

   (a) **Q-values before action:** $Q(1) = 2$, $Q(2) = 0$, $Q(3) = 1.5$, $Q(4) = 0$.

   (b) **Action taken:** $A_4 = 2$.

   (c) **Greedy action:** Action 1 (highest Q-value of 2).

   (d) **Analysis:** The agent chose action 2, which is not the greedy action. **The agent has explored**.

5. **At** $t = 5$**:**

   (a) **Q-values before action:** $Q(1) = 2$, $Q(2) = 1$, $Q(3) = 1.5$, $Q(4) = 0$.

   (b) **Action taken:** $A_5 = 3$.

   (c) **Greedy action:** Action 1 (highest Q-value of 2).

   (d) **Analysis:** The agent chose action 3, which is not the greedy action. **Exploration happened**.

In gist, to the best of my knowledge, the agent has decided to explore in all the timesteps except the first one that I cannot confidently say what happen because all the estimates had the same values. And also, even if the agent chooses the highest estimate, in $\epsilon-$greedy, again we cannot be confident to say whether it has exploited or explored as with probability ϵ, we are choosing actions randomly, and the best action has the same probability to be chosen like the other actions.

SIGNED AS: MAJID GHASEMI , `September 26`