Indian Institute of Technology, Kanpur
Department of Electrical Engineering
Introduction to Reinforcement Learning (EE932)
Quiz 2

Date: 26th May 2024 (4 to 5:30 PM) 2023-24 Quarter 4 Max points: 10

All the Questions are Objective type

1. What is the update equation for SARSA? **(1 pt)**

   (a) $V(s_t) = V(s_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(s_t)]$
   (b) $Q(s_t, a_t) = Q(s_t, a_t) + \alpha[R_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]$
   (c) $Q(s_t, a_t) = Q(s_t, a_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - Q(s_t, a_t)]$
   (d) $V(s_t) = V(s_t) + \alpha[R_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - V(s_t)]$

   Answer: (b)

2. Suppose we are interested in evaluating the value function of a policy π but interacting according to it is costly. Assume that someone offers you some episodes of data collected using another policy. Which of the following statements is correct about the Off-Policy MC method to estimate V_π? **(1 pt)**

   (a) We can use that data to estimate V_π only if we know exactly the policy used to collect it.
   (b) The data itself is sufficient since it does not matter which policy is used to collect it.

   Answer: (a)

   We need the details of the policy used to collect that data. It is required to compute the weights of the importance sampling.

3. Given that $q^\pi(s, a) > v^\pi(s)$, we can conclude **(1 pt)**

   (a) action a is the best action that can be taken in state s
   (b) π may be an optimal policy
   (c) π is not an optimal policy
   (d) none of the above

Answer: (c)

The inequality indicates that there exists an action that, if taken in state s, the expected return would be higher than the expected return of taking actions in state s as per policy π. While this indicates that π is not an optimal policy, it does not indicate that a is the best action that can be taken in state s, since there may exist another action a' such that $q^\pi(s,a') > q^\pi(s,a)$.

4. While following a policy π, the following two episodes of data is observed for an undiscounted MDP with two states P and Q and a terminal state T: **(1 pt)**

$$P, +3, P, +2, Q, -4, P, +4, Q, -3, T$$
$$Q, -2, P, +3, Q, -3, T$$

Estimate the state-value function V^π using first-visit Monte-Carlo evaluation.

(a) $V^\pi(P) = 2, V^\pi(Q) = -\frac{5}{2}$

(b) $V^\pi(P) = 2, V^\pi(Q) = 0$

(c) $V^\pi(P) = 1, V^\pi(Q) = -\frac{5}{2}$

(d) $V^\pi(P) = 1, V^\pi(Q) = 0$

Answer: (c)

For first-visit MC, we consider only the first occurrence of each state in each transition. Thus, we have

$$V^\pi(P) = \frac{2+0}{2} = 1$$

$$V^\pi(Q) = \frac{-3-2}{2} = -\frac{5}{2}$$

5. Considering the same transition data as above, estimate the state value function using the every-visit Monte-Carlo evaluation. **(1 pt)**

(a) $v(P) = 2, v(Q) = -\frac{5}{2}$

(b) $v(P) = 2, v(Q) = -\frac{11}{4}$

(c) $v(P) = \frac{1}{2}, v(Q) = -\frac{11}{4}$

(d) $v(P) = \frac{1}{4}, v(Q) = -\frac{5}{2}$

Answer: (c)

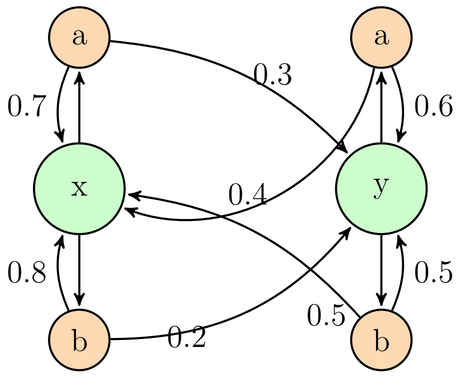
In the every-visit case, we consider each occurrence of each state in the transitions. Thus, we have

$$v(P) = \frac{2 + (-1) + 1 + 0}{4} = \frac{2}{4} = \frac{1}{2}$$

$$v(Q) = \frac{-3 - 3 - 2 - 3}{4} = \frac{-11}{4}$$

6. Consider the following Markov Decision Process (MDP) with two states x and y, and two actions a and b. The transition probabilities and rewards are given as follows:

(1+1+1+1+1 pts)



$$\begin{aligned} P(x|x,a) &= 0.7, \quad P(y|x,a) = 0.3 \\ P(x|x,b) &= 0.8, \quad P(y|x,b) = 0.2 \\ P(x|y,a) &= 0.4, \quad P(y|y,a) = 0.6 \\ P(x|y,b) &= 0.5, \quad P(y|y,b) = 0.5 \end{aligned}$$

Rewards:

$$\begin{aligned} R(x,a) &= 5, \quad R(x,b) = 0 \\ R(y,a) &= 10, \quad R(y,b) = 2 \end{aligned}$$

- From state x:
  - Taking action a yields a reward of 5 ($R(x,a) = 5$).
  - Taking action b yields a reward of 0 ($R(x,b) = 0$).

- From state y:
  - Taking action a yields a reward of 10 ($R(y, a) = 10$).
  - Taking action b yields a reward of 2 ($R(y, b) = 2$).

(i) Assume an initial uniform random policy π_0 where $\pi_0(a|x) = 0.5$ and $\pi_0(b|x) = 0.5$, $\pi_0(a|y) = 0.5$, and $\pi_0(b|y) = 0.5$. The discount factor γ is 0.9. Calculate the value function $V_1(s)$ after one iteration of policy evaluation with initial value $V_0(s) = [2.5, 6]$.

(a) $V_1(x) = 2.5, V_1(y) = 6$

(b) $V_1(x) = 5.53, V_1(y) = 9.98$

(c) $V_1(x) = 35.78, V_1(y) = 40.58$

(d) $V_1(x) = 2.5, V_1(y) = 0$

Answer: (b)

First, calculate the expected reward $R_\pi(s)$ for each state under the policy π_0:

$$R_\pi(x) = \pi_0(a|x) \cdot R(x, a) + \pi_0(b|x) \cdot R(x, b)$$

$$R_\pi(x) = 0.5 \cdot 5 + 0.5 \cdot 0 = 2.5$$

$$R_\pi(y) = \pi_0(a|y) \cdot R(y, a) + \pi_0(b|y) \cdot R(y, b)$$

$$R_\pi(y) = 0.5 \cdot 10 + 0.5 \cdot 2 = 5 + 1 = 6$$

Next, calculate the expected transition probabilities $P_\pi(s, s')$ for each state under the policy π_0:

$$P_\pi(x, x) = \pi_0(a|x) \cdot P(x|x, a) + \pi_0(b|x) \cdot P(x|x, b)$$

$$P_\pi(x, x) = 0.5 \cdot 0.7 + 0.5 \cdot 0.8 = 0.35 + 0.4 = 0.75$$

$$P_\pi(x, y) = \pi_0(a|x) \cdot P(y|x, a) + \pi_0(b|x) \cdot P(y|x, b)$$

$$P_\pi(x, y) = 0.5 \cdot 0.3 + 0.5 \cdot 0.2 = 0.15 + 0.1 = 0.25$$

$$P_\pi(y, x) = \pi_0(a|y) \cdot P(x|y, a) + \pi_0(b|y) \cdot P(x|y, b)$$

$$P_\pi(y, x) = 0.5 \cdot 0.4 + 0.5 \cdot 0.5 = 0.2 + 0.25 = 0.45$$

$$P_\pi(y, y) = \pi_0(a|y) \cdot P(y|y, a) + \pi_0(b|y) \cdot P(y|y, b)$$

$$P_\pi(y,y) = 0.5 \cdot 0.6 + 0.5 \cdot 0.5 = 0.3 + 0.25 = 0.55$$

Now, calculate the value function $V_1(s)$ for one iteration of policy evaluation:

$$V_1(x) = R_\pi(x) + \gamma \left[P_\pi(x,x)V_0(x) + P_\pi(x,y)V_0(y)\right]$$

$$V_1(x) = 2.5 + 0.9 \left[0.75 \cdot 2.5 + 0.25 \cdot 6\right] = 5.53$$

$$V_1(y) = R_\pi(y) + \gamma \left[P_\pi(y,x)V_0(x) + P_\pi(y,y)V_0(y)\right]$$

$$V_1(y) = 6 + 0.9 \left[0.45 \cdot 2.5 + 0.55 \cdot 6\right] = 9.9825$$

$$V_1(s) = \begin{bmatrix} 5.53 & 9.9825 \end{bmatrix}$$

(ii) Assume that the iterative policy evaluation for the uniform policy π_0 converges to $V_{\pi_0}(x) = 35.78, V_{\pi_0}(y) = 40.58$. Upon performing policy improvement, what will be the next π_1?

(a) $\pi_1(x) = a, \pi_1(y) = a$
(b) $\pi_1(x) = a, \pi_1(y) = b$
(c) $\pi_1(x) = b, \pi_1(y) = a$
(d) $\pi_1(x) = b, \pi_1(y) = b$

Answer: (a)

The value function from iterative policy evaluation is:

$$V(x) \approx 35.7877$$

$$V(y) \approx 40.5822$$

For each state s, we need to calculate the value of each action a and choose the action that maximizes this value. The value of taking action a in state s is given by:

$$Q(s,a) = R(s,a) + \gamma \sum_{s'} P(s'|s,a)V(s')$$

State x

Action a

$$Q(x,a) = R(x,a) + \gamma[P(x|x,a)V(x) + P(y|x,a)V(y)]$$

$$Q(x,a) = 5 + 0.9[0.7 \cdot 35.7877 + 0.3 \cdot 40.5822]$$
$$Q(x,a) = 5 + 0.9[25.05139 + 12.17466]$$
$$Q(x,a) = 5 + 0.9 \cdot 37.22605$$
$$Q(x,a) = 5 + 33.50344$$
$$Q(x,a) \approx 38.5034$$

Action b

$$Q(x,b) = R(x,b) + \gamma[P(x|x,b)V(x) + P(y|x,b)V(y)]$$

$$Q(x,b) = 0 + 0.9[0.8 \cdot 35.7877 + 0.2 \cdot 40.5822]$$
$$Q(x,b) = 0.9[28.63016 + 8.11644]$$
$$Q(x,b) = 0.9 \cdot 36.7466$$
$$Q(x,b) \approx 33.0719$$

State y

Action a

$$Q(y,a) = R(y,a) + \gamma[P(x|y,a)V(x) + P(y|y,a)V(y)]$$

$$Q(y,a) = 10 + 0.9[0.4 \cdot 35.7877 + 0.6 \cdot 40.5822]$$
$$Q(y,a) = 10 + 0.9[14.31508 + 24.34932]$$
$$Q(y,a) = 10 + 0.9 \cdot 38.6644$$
$$Q(y,a) = 10 + 34.7980$$
$$Q(y,a) \approx 44.7980$$

Action b

$$Q(y,b) = R(y,b) + \gamma[P(x|y,b)V(x) + P(y|y,b)V(y)]$$

$$Q(y,b) = 2 + 0.9[0.5 \cdot 35.7877 + 0.5 \cdot 40.5822]$$

$$Q(y,b) = 2 + 0.9[17.89385 + 20.2911]$$

$$Q(y,b) = 2 + 0.9 \cdot 38.18495$$

$$Q(y,b) = 2 + 34.3665$$

$$Q(y,b) \approx 36.3665$$

Improved Policy

We compare the action values $Q(s,a)$ and choose the action that maximizes the value for each state.

For state x:

$$Q(x,a) \approx 38.5034$$

$$Q(x,b) \approx 33.0719$$

Since $Q(x,a) > Q(x,b)$, the improved policy $\pi'(x)$ is:

$$\pi'(x) = a$$

For state y:

$$Q(y,a) \approx 44.7980$$

$$Q(y,b) \approx 36.3665$$

Since $Q(y,a) > Q(y,b)$, the improved policy $\pi'(y)$ is:

$$\pi'(y) = a$$

New Policy

The new improved policy is:

$$\pi'(x) = a$$

$$\pi'(y) = a$$

(iii) Consider the previous Markov Decision Process (MDP) with two states x and y, and two actions a and b. Using **value iteration**, compute the value function for the states x and y after the second iteration. Assume that the initial value function is $V_0(x) = 0$ and $V_0(y) = 0$, and the value function after the first iteration is $V_1(x) = 5$ and $V_1(y) = 10$. What are the values of $V_2(x)$ and $V_2(y)$ after the second iteration of value iteration?

(a) $V_2(x) = 8.05, \quad V_2(y) = 15.00$

(b) $V_2(x) = 9.50, \quad V_2(y) = 16.00$

(c) $V_2(x) = 10.85, \quad V_2(y) = 17.20$

(d) $V_2(x) = 11.50, \quad V_2(y) = 18.00$

Answer: (c)

- **For state x:**

$$\begin{aligned}
V_2(x) &= \max\left(R(x,a) + \gamma\left[P(x|x,a)V_1(x) + P(y|x,a)V_1(y)\right],\right.\\
&\qquad \left.R(x,b) + \gamma\left[P(x|x,b)V_1(x) + P(y|x,b)V_1(y)\right]\right)\\
&= \max\left(5 + 0.9\left[0.7 \cdot 5 + 0.3 \cdot 10\right],\right.\\
&\qquad \left.0 + 0.9\left[0.8 \cdot 5 + 0.2 \cdot 10\right]\right)\\
&= \max\left(5 + 0.9\left[3.5 + 3\right], 0.9\left[4 + 2\right]\right)\\
&= \max\left(5 + 0.9 \cdot 6.5, 0.9 \cdot 6\right)\\
&= \max\left(5 + 5.85, 5.4\right)\\
&= \max\left(10.85, 5.4\right)\\
&= 10.85
\end{aligned}$$

- **For state y:**

$$\begin{aligned}
V_2(y) &= \max\left(R(y,a) + \gamma\left[P(x|y,a)V_1(x) + P(y|y,a)V_1(y)\right],\right.\\
&\qquad \left.R(y,b) + \gamma\left[P(x|y,b)V_1(x) + P(y|y,b)V_1(y)\right]\right)\\
&= \max\left(10 + 0.9\left[0.4 \cdot 5 + 0.6 \cdot 10\right],\right.\\
&\qquad \left.2 + 0.9\left[0.5 \cdot 5 + 0.5 \cdot 10\right]\right)\\
&= \max\left(10 + 0.9\left[2 + 6\right], 2 + 0.9\left[2.5 + 5\right]\right)\\
&= \max\left(10 + 0.9 \cdot 8, 2 + 0.9 \cdot 7.5\right)\\
&= \max\left(10 + 7.2, 2 + 6.75\right)\\
&= \max\left(17.2, 8.75\right)\\
&= 17.2
\end{aligned}$$

Thus, the value function after the second iteration of value iteration is:

$$V_2(x) = 10.85, \quad V_2(y) = 17.2$$

(iv) Consider the same Markov Decision Process (MDP) as before. Assume the discount factor γ is 0.9, and the learning rate α is 0.1. Assume the initial Q-values are $Q(x, a) = 1$, $Q(x, b) = 2$, $Q(y, a) = 3$, and $Q(y, b) = 4$.

The agent follows the trajectory:

- Starts in state x, takes action a, transitions to state y, and receives a reward of $R(x, a) = 5$. In state y, it takes action b.
- In state y, takes action b, transitions back to state x, and receives a reward of $R(y, b) = 2$. In state x, it takes action b.
- In state x, takes action b, transitions to state y, and receives a reward of $R(x, b) = 0$. In state y, it takes action a.

What are the updated Q-values for $Q(x, a)$, $Q(y, b)$, and $Q(x, b)$ after these steps?

(a) $Q(x, a) = 1.76$, $Q(y, b) = 3.98$, $Q(x, b) = 2.07$

(b) $Q(x, a) = 1.76$, $Q(y, b) = 4.00$, $Q(x, b) = 2.07$

(c) $Q(x, a) = 1.50$, $Q(y, b) = 3.98$, $Q(x, b) = 2.00$

(d) $Q(x, a) = 1.76$, $Q(y, b) = 3.50$, $Q(x, b) = 2.20$

Answer: (a)

- **Initial Q-values**:

$$Q(x, a) = 1, \quad Q(x, b) = 2, \quad Q(y, a) = 3, \quad Q(y, b) = 4$$

- **First step**:
  - State x, Action a, Next State y, Next Action b
  - Reward $R(x, a) = 5$

$$Q(x, a) \leftarrow Q(x, a) + \alpha \left[R(x, a) + \gamma Q(y, b) - Q(x, a)\right]$$
$$Q(x, a) \leftarrow 1 + 0.1 \left[5 + 0.9 \cdot 4 - 1\right]$$
$$Q(x, a) \leftarrow 1 + 0.1 \left[5 + 3.6 - 1\right]$$
$$Q(x, a) \leftarrow 1 + 0.1 \cdot 7.6$$
$$Q(x, a) \leftarrow 1 + 0.76$$
$$Q(x, a) \leftarrow 1.76$$

- **Second step**:
  - State y, Action b, Next State x, Next Action b
  - Reward $R(y, b) = 2$

$$Q(y, b) \leftarrow Q(y, b) + \alpha \left[R(y, b) + \gamma Q(x, b) - Q(y, b)\right]$$
$$Q(y, b) \leftarrow 4 + 0.1 \left[2 + 0.9 \cdot 2 - 4\right]$$
$$Q(y, b) \leftarrow 4 + 0.1 \left[2 + 1.8 - 4\right]$$
$$Q(y, b) \leftarrow 4 + 0.1 \cdot -0.2$$
$$Q(y, b) \leftarrow 4 - 0.02$$
$$Q(y, b) \leftarrow 3.98$$

- **Third step**:
  - State x, Action b, Next State y, Next Action a
  - Reward $R(x, b) = 0$

$$Q(x, b) \leftarrow Q(x, b) + \alpha \left[R(x, b) + \gamma Q(y, a) - Q(x, b)\right]$$
$$Q(x, b) \leftarrow 2 + 0.1 \left[0 + 0.9 \cdot 3 - 2\right]$$
$$Q(x, b) \leftarrow 2 + 0.1 \left[0 + 2.7 - 2\right]$$
$$Q(x, b) \leftarrow 2 + 0.1 \cdot 0.7$$
$$Q(x, b) \leftarrow 2 + 0.07$$
$$Q(x, b) \leftarrow 2.07$$

Thus, the updated Q-values after these steps are:

$$Q(x, a) = 1.76, \quad Q(x, b) = 2.07, \quad Q(y, a) = 3, \quad Q(y, b) = 3.98$$

(v) Assume $\epsilon = 0.1$. Consider the trajectory data given in the previous question. In the third step of the trajectory, it was mentioned that action a was chosen in state y. What would have been the policy used by the agent to decide that action?

(a) $P(a|x) = 0.05$, $P(b|x) = 0.95$; $P(a|y) = 0.05$, $P(b|y) = 0.95$

(b) $P(a|x) = 0.1$, $P(b|x) = 0.9$; $P(a|y) = 0.1$, $P(b|y) = 0.9$

(c) $P(a|x) = 0.95$, $P(b|x) = 0.05$; $P(a|y) = 0.95$, $P(b|y) = 0.05$

(d) $P(a|x) = 0.9$, $P(b|x) = 0.1$; $P(a|y) = 0.9$, $P(b|y) = 0.1$

Answer: (a)

The action taken in state y during the third step would be based on the Q-values updated after the second step. Let's review the Q-values after the second step:

After the first step:

$$Q(x,a) = 1.76, \quad Q(x,b) = 2, \quad Q(y,a) = 3, \quad Q(y,b) = 4$$

After the second step:

$$Q(x,a) = 1.76, \quad Q(x,b) = 2, \quad Q(y,a) = 3, \quad Q(y,b) = 3.98$$

Using the Q-values updated after the second step, we apply the epsilon-greedy policy:

- The action b has the highest Q-value: $Q(y,b) = 3.98$. - The probability of taking the best action b is $1 - \epsilon + \frac{\epsilon}{2} = 1 - 0.1 + 0.05 = 0.95$. - The probability of taking the suboptimal action a is $\frac{\epsilon}{2} = 0.05$.

Thus, the probability of taking action a in state y is:

$$P(a|y) = 0.05, P(b|y) = 0.95$$

Similarly, we can compute for state x.