RBE 595 — Reinforcement Learning
Week #5 Assignment

Arjan Gupta

Problem 1

When is it suited to apply Monte-Carlo to a problem?

Answer

Monte-Carlo methods are best suited to be applied to problems where we do not have a model of the environment (i.e., the dynamics of the environment are unknown). For example, sometimes it is simply not practical to model the complexity of the environment. In such cases, the agent must learn about the environment by interacting with it and using the obtained rewards to update its policy via the action-value function.

Problem 2

When does the Monte-carlo prediction performs the first update?

Answer

The Monte-Carlo prediction performs the first update after an episode terminates. This is because the Monte-Carlo method is an episodic method, i.e., it learns from a series of state, action, and reward tuples that occur in an episode.

Problem 3

What is off-policy learning and why it is useful?

Answer

Off-policy learning is a method of reinforcement learning where the agent learns about the environment by observing the behavior of another agent, called the *behavior policy*, which is the policy responsible for exploration and interaction. However, the agent performs evaluation and optimization using a different policy, called the *target policy*.

Off-policy learning is useful for the following reasons,

- It avoids the unlikely assumption of exploring starts. In some Monte Carlo algorithms, the exploratory behavior comes from random starting states, however this is not always possible. Off-policy learning allows the agent to learn about the environment without this assumption.
- Existing knowledge can be leveraged by learning from the behavior of other agents. The behavior policy can be a simple random policy, or it can be a policy that has been learned from past experience.
- It avoids the situation where the agent is stuck in a suboptimal policy because it is not exploring enough. This would happen if the agent is using a greedy and deterministic policy to learn about the environment.
- It avoids unexpected actions that may occur during exploration. Instead, the behavior policy can continuously explore while the target policy learns. This is particularly useful in cases where the environment is dangerous or expensive to explore, or if humans are involved.

Problem 4

(Exercise 5.5, page 105) Consider an MDP with a single nonterminal state and a single action that transitions back to the nonterminal state with probability p and transitions to the terminal state with probability $1 - p$. Let the reward be +1 on all transitions, and let $\gamma = 1$. Suppose you observe one episode that lasts 10 steps, with a return of 10. What are the first-visit and every-visit estimators of the value of the nonterminal state?

Answer

Since this problem does not involve a behavior and target policy, we will not use the importance sampling ratio. Instead, we can manually calculate the first-visit and every-visit estimators of the value of the nonterminal state.

Given episode

Let the nonterminal state be s (and let s_i denote the i^{th} time s was visited) and the terminal state be s'. Let the action be a (and let a_i denote the i^{th} time s was visited). The reward is $r = 1$ for all transitions. Also, $\gamma = 1$.

The given episode is as follows,

$$s_0 \xrightarrow{a_1} s_1 \xrightarrow{a_2} s_2 \xrightarrow{a_3} s_3 \xrightarrow{a_4} s_4 \xrightarrow{a_5} s_5 \xrightarrow{a_6} s_6 \xrightarrow{a_7} s_7 \xrightarrow{a_8} s_8 \xrightarrow{a_9} s_9 \xrightarrow{a_{10}} s'$$

As we can see by the subscript of a, there are 10 rewards of 1 each. Therefore, the total return is 10.

First-Visit Estimator

The first-visit estimator of the value of the nonterminal state is calculated as follows,

$$\begin{aligned} V(s) &= 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 \\ &= 1(10) = 10 \end{aligned}$$

Every-Visit Estimator

The every-visit estimator of the value of the nonterminal state is calculated as follows,

$$\begin{aligned} V(s) &= \frac{1}{10}(1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10) \\ &= \frac{1}{10}(55) = 5.5 \end{aligned}$$

Problem 5

(Exercise 5.7, page 108) In learning curves such as those shown in Figure 5.3 error generally decreases with training, as indeed happened for the ordinary importance-sampling method. But for the weighted importance-sampling method error first increased and then decreased. Why do you think this happened?

Answer

As shown in the lectures by Dr. Navid Dadkhah Tehrani, here is a table showing the bias and variance comparison between the ordinary importance-sampling method and the weighted importance-sampling method.

	Ordinary Importance-Sampling	**Weighted Importance-Sampling**
Bias	Un-biased	Biased (eventually unbiased)
Variance	Large	Low

For weighted importance-sampling, the bias is initially high, but it eventually becomes unbiased. That initial bias is what causes the error to increase initially. However, after a large number of episodes, the bias decreases and the error decreases as well.

Problem 6

(Exercise 3.14) The Bellman equation must hold for each state for the value function v_π shown in Figure 3.2 (right-side) of Example 3.5. Show numerically that this equation holds for the center state, valued at +0.7, with respect to its four neighboring states, valued at +2.3, +0.4, -0.4, and +0.7. (These numbers are accurate only to one decimal place.)

Answer

From the textbook, the state-value function for a policy π is defined as,

$$\begin{aligned} v_\pi(s) &\doteq \mathbb{E}_\pi\left[G_t \mid S_t = s\right] \\ &= \sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a)\left[r + \gamma v_\pi(s')\right] \end{aligned}$$

From Example 3.5, we also know the following given information:

- The action set $A = \{\text{up}, \text{down}, \text{left}, \text{right}\}$ in each state.
- An equiprobable random policy is used. Therefore, $\pi(a \mid s) = 0.25$ for all $a \in A$ and $s \in S$.
- The reward is always 0 for all transitions.
- $\gamma = 0.9$.
- Any action taken deterministically leads to the expected state, so $p = 1$.

Hence, the state-value function for the center state is,

$$\begin{aligned} v_\pi(s_{\text{center}}) &= \sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a)\left[r + \gamma v_\pi(s')\right] \\ &= \pi(\text{up} \mid s) p(s_{\text{up}}, r \mid s, \text{up})\left[r + \gamma v_\pi(s_{\text{up}})\right] + \pi(\text{down} \mid s) p(s_{\text{down}}, r \mid s, \text{down})\left[r + \gamma v_\pi(s_{\text{down}})\right] \\ &\quad + \pi(\text{left} \mid s) p(s_{\text{left}}, r \mid s, \text{left})\left[r + \gamma v_\pi(s_{\text{left}})\right] + \pi(\text{right} \mid s) p(s_{\text{right}}, r \mid s, \text{right})\left[r + \gamma v_\pi(s_{\text{right}})\right] \\ &= 0.25 \cdot 1 \cdot [0 + 0.9 \cdot 2.3] + 0.25 \cdot 1 \cdot [0 + 0.9 \cdot 0.4] + 0.25 \cdot 1 \cdot [0 + 0.9 \cdot (-0.4)] + 0.25 \cdot 1 \cdot [0 + 0.9 \cdot 0.7] \\ &= 0.25 \cdot 0.9 \cdot [2.3 + 0.4 - 0.4 + 0.7] \\ &= 0.25 \cdot 0.9 \cdot 3.0 \\ &= 0.675 \approx 0.7 \text{ (rounded to one decimal place, as mentioned in prompt)} \end{aligned}$$

Therefore, we see that the Bellman equation holds for the center state, valued at +0.7.

Problem 7

(Exercise 3.17) What is the Bellman equation for action values, that is, for q_π? It must give the action value $q_\pi(s, a)$ in terms of the action values, $q_\pi(s', a')$, of possible successors to the state-action pair (s, a). Hint: the backup diagram below corresponds to this equation. Show the sequence of equations analogous to (3.14), but for action values.

Answer

From the textbook, the action-value function for a policy π is defined as,

$$\begin{aligned} q_\pi(s, a) &\doteq \mathbb{E}_\pi \left[G_t \mid S_t = s, A_t = a \right] \\ &= \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = a \right] \\ &= \mathbb{E}_\pi \left[R_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} \mid S_t = s, A_t = a \right] \\ &= \mathbb{E}_\pi \left[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a \right] \\ &= \mathbb{E}_\pi \left[R_{t+1} \mid S_t = s, A_t = a \right] + \gamma \mathbb{E}_\pi \left[G_{t+1} \mid S_t = s, A_t = a \right] \end{aligned}$$

Now, let us consider the first and second terms of the above equation separately.

First Term

$$\mathbb{E}_\pi \left[R_{t+1} \mid S_t = s, A_t = a \right] = \sum_{r \in \mathcal{R}} r \cdot p(r \mid s, a) = \sum_{r \in \mathcal{R}} \sum_{s' \in \mathcal{S}} r \cdot p(s', r \mid s, a)$$

Second Term

$$\begin{aligned} \gamma \mathbb{E}_\pi \left[G_{t+1} \mid S_t = s, A_t = a \right] &= \gamma \sum_{g \in \mathcal{G}} g \cdot p(g \mid s, a) \\ &= \gamma \sum_{g \in \mathcal{G}} \sum_{r \in \mathcal{R}} \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} g \cdot p(g \mid s', a') \cdot p(s', r \mid s, a) \cdot \pi(a' \mid s') \end{aligned}$$

Where, $\sum_{g \in \mathcal{G}} g \cdot p(g \mid s', a') = \mathbb{E}_\pi \left[G_{t+1} \mid S_{t+1} = s', A_{t+1} = a' \right] = q_\pi(s', a')$

Therefore the second term is,

$$\gamma \mathbb{E}_\pi \left[G_{t+1} \mid S_t = s, A_t = a \right] = \gamma \sum_{r \in \mathcal{R}} \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} q_\pi(s', a') \cdot p(s', r \mid s, a) \cdot \pi(a' \mid s')$$

Now, combining the first and second terms, we get,

Problem 7 continued on next page. . .

$$
\begin{aligned}
q_\pi(s,a) &= \sum_{r\in\mathcal{R}}\sum_{s'\in\mathcal{S}} r\cdot p(s',r\mid s,a) + \gamma\sum_{r\in\mathcal{R}}\sum_{s'\in\mathcal{S}}\sum_{a'\in\mathcal{A}} q_\pi(s',a')\cdot p(s',r\mid s,a)\cdot\pi(a'\mid s') \\
q_\pi(s,a) &= \sum_{s',r} p(s',r\mid s,a)\left[r+\gamma\sum_{a'}\pi(a'\mid s')q_\pi(s',a')\right]
\end{aligned}
$$

Which is the Bellman equation for action values, i.e., for q_π.

Backup Diagram Confirmation

This equation can be verified by looking at the backup diagram given in the prompt. The backup diagram shows that we start with the state-action pair (s,a). To get to the next state, we are subjected to the environment $p(s',r\mid s,a)$. The reward r is added to the discounted return G_{t+1}. This brings us to our new state, s'. At this point, the equation would look as follows,

$$
q_\pi(s,a) = \sum_{s',r} p(s',r\mid s,a)\left[r+\gamma v_\pi(s')\right]
$$

However we still need to eliminate the $v_\pi(s')$ term. To do this, we go through our policy, π, to get the action a' that we would take in the state s'. Now the equation becomes,

$$
q_\pi(s,a) = \sum_{s',r} p(s',r\mid s,a)\left[r+\gamma\sum_{a'}\pi(a'\mid s')q_\pi(s',a')\right]
$$

So, the Bellman equation for action values, i.e., for q_π, is confirmed by the backup diagram.

Problem 8

(Exercise 3.22) Consider the continuing MDP shown below. The only decision to be made is that in the top state, where two actions are available, left and right. The numbers show the rewards that are received deterministically after each action. There are exactly two deterministic policies, π_{left} and π_{right}. What policy is optimal if $\gamma = 0$? If $\gamma = 0.9$? If $\gamma = 0.5$?

Answer

The discounted return is defined as,

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{3.8}$$

Case 1: $\gamma = 0$

When $\gamma = 0$, the left policy rewards are calculated as follows,

$$G_{\text{left}} = 1 + 0 + 0 + \cdots = 1$$

Similarly, the right policy rewards are calculated as follows,

$$G_{\text{right}} = 0 + 0 + \cdots = 0$$

In this case, the **left** policy is optimal.

Case 2: $\gamma = 0.9$

When $\gamma = 0.9$, the left policy rewards are calculated as follows,

$$\begin{aligned} G_{\text{left}} &= 1 + 0.9 \cdot 0 + 0.9^2 \cdot 1 + \cdots \\ &= 1 + 0.9^2 + 0.9^4 + \cdots \\ &= \sum_{k=0}^{\infty} 0.9^{2k} \\ &= \sum_{k=0}^{\infty} 0.81^k \\ &= \frac{1}{1 - 0.81} = \frac{1}{0.19} \\ &= 5.263 \end{aligned}$$

Similarly, the right policy rewards are calculated as follows,

$$\begin{aligned} G_{\text{right}} &= 0 + 0.9 \cdot 2 + 0 + 0.9^3 \cdot 2 + \cdots \\ &= 0.9 \cdot 2 + 0.9^3 \cdot 2 + \cdots \\ &= 2 \cdot \sum_{k=0}^{\infty} 0.9^{2k+1} = 2 \cdot \sum_{k=0}^{\infty} (0.9)(0.81)^k = 2 \cdot \frac{0.9}{1 - 0.81} \\ &= \frac{1.8}{0.19} = 9.474 \end{aligned}$$

In this case, the **right** policy is optimal.

Problem 8 continued on next page. . .

Case 3: $\gamma = 0.5$

When $\gamma = 0.5$, the left policy rewards are calculated as follows,

$$\begin{aligned} G_{\text{left}} &= 1 + 0.5 \cdot 0 + 0.5^2 \cdot 1 + \cdots \\ &= 1 + 0.5^2 + 0.5^4 + \cdots \\ &= \sum_{k=0}^{\infty} 0.5^{2k} = \sum_{k=0}^{\infty} 0.25^k \\ &= \frac{1}{1 - 0.25} = \frac{1}{0.75} \\ &= 1.333 \end{aligned}$$

Similarly, the right policy rewards are calculated as follows,

$$\begin{aligned} G_{\text{right}} &= 0 + 0.5 \cdot 2 + 0 + 0.5^3 \cdot 2 + \cdots \\ &= 0.5 \cdot 2 + 0.5^3 \cdot 2 + \cdots \\ &= 2 \cdot \sum_{k=0}^{\infty} 0.5^{2k+1} = 2 \cdot \sum_{k=0}^{\infty} (0.5)(0.25)^k = 2 \cdot \frac{0.5}{1 - 0.25} \\ &= \frac{1}{0.75} = 1.333 \end{aligned}$$

In this case, both the **left** and **right** policies are optimal.