RBE 595 — Reinforcement Learning
Week #7 Assignment
Temporal Difference Learning

Arjan Gupta

Problem 1

Between DP (Dynamic Programming), MC (Monte-Carlo) and TD (Temporal Difference), which one of these algorithms use bootstrapping? Explain.

Answer

Bootstrapping is the process of updating the value of a state based on the value of a future state.

- **Dynamic Programming** (DP) uses bootstrapping. This is because DP uses the Bellman equation to update the value of a state based on the value of a future state.
- **Monte-Carlo** (MC) does not use bootstrapping. This is because MC does not use the Bellman equation to update the value of a state based on the value of a future state. Instead, MC uses the actual return value to update the value of a state.
- **Temporal Difference** (TD) uses bootstrapping. This is because TD uses the Bellman equation to update the value of a state based on the value of a future state.

Problem 2

We mentioned that the target value for TD is $[R_{t+1} + \gamma V(s_{t+1})]$. What is the target value for Monte-carlo, Q-learning, SARSA and Expected-SARSA?

Answer

- **Monte-Carlo** (MC) does not use bootstrapping. Therefore, the target value is the actual return value, G_t.
- **Q-Learning** is an off-policy TD control algorithm. Therefore, the target value is $R_{t+1}+\gamma \max_a Q(S_{t+1}, a)$.
- **SARSA** is an on-policy TD control algorithm. Therefore, the target value is $R_{t+1} + \gamma Q(S_{t+1}, A_{t+1})$.
- **Expected-SARSA** is an on-policy TD control algorithm. Therefore, the target value is $R_{t+1} + \gamma \mathbb{E}_\pi \left[Q(S_{t+1}, A_{t+1}) \mid S_{t+1}\right]$.

Problem 3

What are the similarities of TD and MC?

Answer

The textbook states the *reward hypothesis* as follows,

> "That all of what we mean by goals and purposes can be well thought of as the maximization of the expected value of the cumulative sum of a received scalar signal (called reward)."

Here is a simplified break-down of what the reward hypothesis means:

- In RL, we talk about goals and purposes, which is to find best way to solve a problem.
- Any solution to a complex problem can be broken down into a series of steps, and each step can have a value associated with it.
- We design this 'value' associated with each step as a scalar signal which is received from the environment. This scalar signal is called the *reward.*
- Therefore, **we hypothesize that** our all goals can be achieved by the maximization of the expected cumulative reward.
- A paper from 2021 titled "Reward is enough" by David Silver, Satinder Singh, Doina Precup, and Richard S. Sutton discusses this hypothesis in detail.

Problem 4

We have an agent in maze-like world. We want the agent to find the goal as soon as possible. We set the reward for reaching the goal equal to +1 with $\gamma = 1$. But we notice that the agent does not always reach the goal as soon as possible. How can we fix this?

Answer

As stated in the textbook, the *discounted return* (equation 3.8), G_t, is defined as,

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

Here, as γ approaches 1, the discounted return takes far-sighted rewards into account. Therefore, if the agent is not reaching the goal as soon as possible, then the agent is likely too far-sighted. Therefore, we can reduce the value of γ to make the agent more near-sighted and reach the goal sooner.

Problem 5

What is the difference between policy and action?

Answer

An *action* is a choice made by the agent at a given state. It is an attempted modification of the environment which leads to a new state or the same state. We give an agent an associated reward for each action.

In contrast, a policy determines how good it is for the agent to perform an action in a given state. Formally, a *policy* is a mapping from states to probabilities of selecting each possible action. It defines a probability distribution over actions for each state.

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.