RBE 595 — Reinforcement Learning
Week #3 Assignment

Arjan Gupta

Problem 1

Suppose $\gamma = 0.8$ and we get the following sequence of rewards

$$R_1 = -2,\ R_2 = 1,\ R_3 = 3,\ R_4 = 4,\ R_5 = 1.0$$

Calculate the value of G_0 by using the equation 3.8 (work forward) and 3.9 (work backward) and show they yield the same results.

Answer

Work Forward

From the the book, the *discounted return* (equation 3.8), G_t, is defined as,

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

Plugging in the values from this problem, we get,

$$\begin{aligned} G_0 &= R_1 + \gamma R_2 + \gamma^2 R_3 + \gamma^3 R_4 + \gamma^4 R_5 \\ &= -2 + 0.8 \cdot 1 + 0.8^2 \cdot 3 + 0.8^3 \cdot 4 + 0.8^4 \cdot 1 \\ &= -2 + 0.8 + 0.64 \cdot 3 + 0.512 \cdot 4 + 0.4096 \\ &= 3.1776 \end{aligned}$$

Work Backward

From the book, the “recursive” representation of *discounted return* (equation 3.9), G_t, is defined as,

$$G_t \doteq R_{t+1} + \gamma G_{t+1} \tag{3.9}$$

Plugging in the values from this problem, we get,

$$\begin{aligned} G_0 &= R_1 + \gamma G_1 \\ &= -2 + 0.8 \cdot G_1 \end{aligned}$$

Where we apply 3.8 to G_1,

$$\begin{aligned} G_1 &= R_2 + \gamma R_3 + \gamma^2 R_4 + \gamma^3 R_5 \\ &= 1 + 0.8 \cdot 3 + 0.8^2 \cdot 4 + 0.8^3 \cdot 1 \\ &= 6.472 \end{aligned}$$

Therefore,

$$\begin{aligned} G_0 &= -2 + 0.8 \cdot G_1 \\ &= -2 + 0.8 \cdot 6.472 \\ &= 3.1776 \end{aligned}$$

Conclusion

We see that both methods yield the same result, $G_0 = 3.1776$.

Problem 2

Explain how a room temperature control system can be modeled as an MDP? What are the states, actions, rewards, and transitions.

Answer

A room temperature control system can be modeled as an MDP as follows.

Scope

Let us make some assumptions to define the scope of the solution.

- The temperatures are being measured in Fahrenheit.
- The temperature resolution of the temperature sensor in the room is 1°F.
- Given the climate of the area, the room naturally stays between the range of 40°F and 90°F.
- The humans in the room are comfortable with temperatures between 68°F and 72°F.

States

Therefore, the states of the system are the temperatures in the room, $S = \{s \in \mathbb{Z} \mid 40 \leq s \leq 90\}$.

Actions

The actions of the system are the temperature changes in the room. Assume that the control system can change the temperature by up to 5°F in either direction. Therefore, in general, the set of all actions are $A = \{a \in \mathbb{Z} \mid -5 \leq a \leq 5\}$. However, the action at each state is limited by the state itself. For example, if the current temperature is below 68°F, then the action cannot be to decrease the temperature further. Therefore, the set of actions can take on three possible sub-sets of A depending on the state, as follows,

- $A_{\text{low}} = \{a \in A \mid a \geq 0\}$, if $s \leq 68$
- $A_{\text{mid}} = \{a \in A \mid -1 \leq a \leq 1\}$, if $68 < s < 72$
- $A_{\text{high}} = \{a \in A \mid a \leq 0\}$, if $s \geq 72$

Rewards

The reward for the system is defined as the difference between the current temperature and the desired temperature. Therefore, the reward function is defined as,

$$r(s, a, s') = \begin{cases} |70 - s|, & \text{if } 68 \leq s \leq 72 \\ 68 - s, & \text{if } s < 68 \\ s - 72, & \text{if } s > 72 \end{cases}$$

Notice that the reward is always non-negative. If the temperature does not change, then the reward is zero. If the temperature changes (the direction of which is enforced by the action set), then the reward is positive.

Problem 2 continued on next page. . .

Transitions

The transitions are defined as follows,

$$p(s' \mid s, a) = \begin{cases} \alpha_{\text{low}}, & \text{if } s \leq 68 \text{ and } s' = s + a \\ \alpha_{\text{mid}}, & \text{if } 68 < s < 72 \text{ and } s' = s + a \\ \alpha_{\text{high}}, & \text{if } s \geq 72 \text{ and } s' = s + a \\ 1 - \alpha_{\text{low}}, & \text{if } s \leq 68 \text{ and } s' = s \\ 1 - \alpha_{\text{mid}}, & \text{if } 68 < s < 72 \text{ and } s' = s \\ 1 - \alpha_{\text{high}}, & \text{if } s \geq 72 \text{ and } s' = s \\ 0, & \text{otherwise} \end{cases}$$

where α_{low}, α_{mid}, and α_{high} are the probabilities of the actions being taken when the state is low, mid, and high respectively. The value of these α's would vary depending on how effective the cooling and heating systems are. For example, if the cooling system is very effective, then α_{low} would be high. Similarly, if the heating system is very effective, then α_{high} would be high.

Tabular Summary

The tabular summary of the MDP is as follows,

s	a	s'	$p(s' \mid s, a)$	$r(s, a, s')$
$40 \leq s \leq 68$	$a \geq 0$	$s + a$	α_{low}	$68 - s$
$40 \leq s \leq 68$	$a \geq 0$	s	$1 - \alpha_{\text{low}}$	$68 - s = 0$
$68 < s < 72$	$-1 \leq a \leq 1$	$s + a$	α_{mid}	$\lvert 70 - s \rvert$
$68 < s < 72$	$-1 \leq a \leq 1$	s	$1 - \alpha_{\text{mid}}$	$\lvert 70 - s \rvert = 0$
$72 \leq s \leq 90$	$a \leq 0$	$s + a$	α_{high}	$s - 72$
$72 \leq s \leq 90$	$a \leq 0$	s	$1 - \alpha_{\text{high}}$	$s - 72 = 0$

Problem 3

What is the reward hypothesis in RL?

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.)

Figure 3.2: Gridworld example: exceptional reward dynamics (left) and state-value function for the equiprobable random policy (right).

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 $p = 1$.

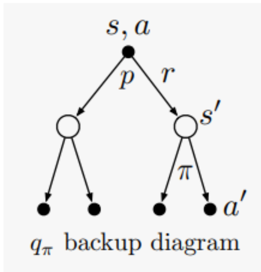
Hence, the state-value function for the center state is,

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

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,

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

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,

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

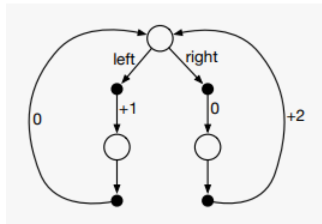
However we still need to eliminate the $v_\pi(s')$ term. To do this, we go through our policy, π, to get the action a' that we would take in the state s'. Now the equation becomes,

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

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,

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

Problem 8 continued on next page. . .

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.

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.