CS253A QR: Reinforcement Learning: Assignment №4

Ayat Ospanov

September 27, 2018

Contents

1 Exercise 3.6

The return at each time would be $-\gamma^L$, where L is the number of steps until failure. It differs from continuing formulation by the return value of $-\gamma^L - \gamma^{L_1} - \cdots - \gamma^{L_n} - \cdots$, where L_i are the next failures.

2 Exercise 3.7

The reward in this problem is not discounted, thus at each time step we get the expected total reward of +1. It is not informative as we are not getting close to the exit from a maze. We can add discounting or punish by -1 for each step in the maze.

3 Exercise 3.8

$$
\begin{aligned}
&\gamma = 0.5 \\
&R_1 = -1, R_2 = 2, R_3 = 6, R_4 = 3, R_5 = 2 \\
&G_5 = 0 \\
&G_4 = R_5 + \gamma G_5 = 2 \\
&G_3 = R_4 + \gamma G_4 = 3 + 0.5 * 2 = 4 \\
&G_2 = R_3 + \gamma G_3 = 6 + 0.5 * 4 = 8 \\
&G_1 = R_2 + \gamma G_2 = 2 + 0.5 * 8 = 6 \\
&G_0 = R_1 + \gamma G_1 = -1 + 0.5 * 6 = 2
\end{aligned}
$$

4 Exercise 3.9

$$
\begin{aligned}
&\gamma = 0.9 \\
&R_i = 2, 7, 7, 7, ... \\
&G_1 = \sum_{i=0}^{\infty} \gamma^i R_{i+2} = R \sum_{i=0}^{\infty} \gamma^i = R \frac{1}{1-\gamma} = \frac{7}{0.1} = 70 \\
&G_0 = R_1 + \gamma G_1 = 2 + 0.9 * 70 = 65
\end{aligned}
$$

5 Exercise 3.11

$$r(s) = \mathbb{E}[R_{t+1} | S_t = s] = \sum_r r \sum_{a,s'} \pi(a|s) p(s', r|s, a)$$

6 Exercise 3.12

$$v_\pi(s) = \sum_a \pi(a|s) q_\pi(s, a)$$

7 Exercise 3.13

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

8 Exercise 3.14

$$\begin{aligned}
v_\pi(s) &= \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)[r + \gamma v_\pi(s')] = \\
&= \{\pi(a|s) = \frac{1}{4}, p(s',r|s,a) = 1, r = 0, \gamma = 0.9\} = \\
&= \frac{1}{4} 0.9[0.7 + 2.3 + 0.4 - 0.4] = \frac{27}{40} = \\
&= 0.675 \approx 0.7
\end{aligned}$$

9 Exercise 3.15

Signs are important as they tell which action is good and which is bad.
Let's show, that adding a constant to all rewards adds a constant to $v_\pi(s)$:

$$\begin{aligned}
v'_\pi(s) &= \mathbb{E}[G'_t|S_t = s] = \\
&= \mathbb{E}[\sum_{i=0}^{\infty} \gamma^i (R_{t+k+1} + c)|S_t = s] = \\
&= \mathbb{E}[G_t + c\sum_{i=0}^{\infty} \gamma^i |S_t = s] = \\
&= \mathbb{E}[G_t|S_t = s] + \mathbb{E}[\frac{c}{1-\gamma}|S_t = s] = \\
&= v_\pi(s) + \frac{c}{1-\gamma}
\end{aligned}$$

Thus, $v_c = \frac{c}{1-\gamma}$. From that we can conclude, that adding a constant doesn't affect relative values of states, as difference of the values is the same.

10 Exercise 3.16

If we add a constant to all rewards in an episodic task, we change value of a state by cT, where T is the steps until the end of the episode. As we have different length and it is dynamic as well, we can't guarantee that relative values have not been affected. Moreover, episodic tasks are sensitive to a reward value. For example, maze running. If the reward for every move -1, we force the algorithm to find the fastest way. But if we add $c = 2$ to all rewards, we lose the aim of the task and the algorithms will not exit the maze as we are maximizing overall reward.

11 Exercise 3.17

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

12 Exercise 3.18

$$v_\pi(s) = \mathbb{E}_\pi[q_\pi(S_t = s, A_t)|S_t = s] = \sum_a \pi(a|s) q_\pi(s, a)$$

13 Exercise 3.19

$$q_\pi(s, a) = \mathbb{E}[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s, A_t = a] = \sum_{s', r} p(s', r|s, a)[r + \gamma v_\pi(s')]$$

14 Exercise 3.22

Let s be the top state and s' one of the bottom ones. Using Bellman equation, we get the next equations:

π_{left}	π_{right}
$v_*(s) = 1 + \gamma v_*(s')$	$v_*(s) = \gamma v_*(s')$
$v_*(s') = \gamma v_*(s)$	$v_*(s') = 2 + \gamma v_*(s)$
$\downarrow$	$\downarrow$
$v_*(s) = \frac{1}{1-\gamma^2}$	$v*(s) = \frac{2\gamma}{1-\gamma^2}$

Using different γ, we get the next table (we can compare only numerators, optimal value if bold):

γ	π_{left}	π_{right}
0	**1**	0
0.5	**1**	**1**
0.9	1	**1.8**

15 Exercise 3.23

All states and action names are shortened to their first letter.

$$q_*(h, s) = \alpha[r_s + \gamma \max_{a \in \{s,w\}} q_*(h, a)] + (1 - \alpha)[r_s + \gamma \max_{a \in \{s,w,r\}} q_*(l, a)]$$
$$q_*(h, w) = r_w + \gamma \max_{a \in \{s,w\}} q_*(h, a)$$
$$q_*(l, s) = \beta[r_s + \gamma \max_{a \in \{s,w,r\}} q_*(l, a)] + (1 - \beta)[-3 + \gamma \max_{a \in \{s,w\}} q_*(h, a)]$$
$$q_*(l, r) = \gamma \max_{a \in \{s,w\}} q_*(h, a)$$
$$q_*(l, w) = r_w + \gamma \max_{a \in \{s,w,r\}} q_*(l, a)$$

After a few simplifications:

$$
\begin{aligned}
q_*(h,s) &= \alpha\gamma \max_{a\in\{s,w\}} q_*(h,a) + (1-\alpha)\gamma \max_{a\in\{s,w,r\}} q_*(l,a) \\
q_*(h,w) &= r_w + \gamma \max_{a\in\{s,w\}} q_*(h,a) \\
q_*(l,s) &= \beta[r_s + \gamma \max_{a\in\{s,w,r\}} q_*(l,a)] + (1-\beta)[-3 + \gamma \max_{a\in\{s,w\}} q_*(h,a)] \\
q_*(l,r) &= \gamma \max_{a\in\{s,w\}} q_*(h,a) \\
q_*(l,w) &= r_w + \gamma \max_{a\in\{s,w,r\}} q_*(l,a)
\end{aligned}
$$

16 Exercise 3.24

As from A we get to A' only, we get the next equation:

$$v_*(A) = r_A + \gamma v_*(A')$$

From A' we get to A by the direct route of going up. Let's mark them A', A'' and etc. As we have reward of 0 for those moves, we get the next equations:

$$
\begin{aligned}
v_*(A') &= \gamma v_*(A'') \\
v_*(A'') &= \gamma v_*(A''') \\
v_*(A''') &= \gamma v_*(A'''') \\
v_*(A'''') &= \gamma v_*(A)
\end{aligned}
$$

From these equations it is obvious we get $v_*(A') = \gamma^4 v_*(A)$
Solving the next system:

$$
\begin{aligned}
v_*(A) &= r_A + \gamma v_*(A') \\
v_*(A') &= \gamma^4 v_*(A)
\end{aligned}
$$

gives us $v_*(A) = r_A + \gamma^5 v_*(A)$. Given $r_A = 10$ and $\gamma = 0.9$, the value for $v_*(A)$ is the next:

$$v_*(A) = \frac{10}{1-\gamma^5} = \frac{10}{0.40951} = 24.419$$

17 Exercise 3.25

$$v_*(s) = \max_{a\in A(s)} q_*(s,a)$$

18 Exercise 3.26

$$q_*(s,a) = \sum_{s',r} p(s',r|s,a)[r + \gamma v_*(s')]$$

19 Exercise 3.27

$$\pi_*(a|s) = \begin{cases} 1, & \text{if } a \in Arg \max\limits_{a \in A(s)} q_*(s,a) \\ 0, & \text{otherwise} \end{cases}$$

20 Exercise 3.28

$$\pi_*(a|s) = \begin{cases} 1, & \text{if } a \in Arg \max\limits_{a \in A(s)} \sum\limits_{s',r} p(s',r|s,a)[r + \gamma v_*(s')] \\ 0, & \text{otherwise} \end{cases}$$