Solutions to Reinforcement Learning by Sutton Chapter 4

Yifan Wang

May 2019

Exercise 4.1

if π is the equiprobable random policy:

$$q_\pi(11, down) = -1 + v_\pi(T) = -1 + 0 = -1$$
$$q_\pi(7, down) = -1 + v_\pi(11) = -1 + (-14) = -15$$

■

Exercise 4.2

Adding state 15 will result:

$$v_\pi(15) = -1 + 0.25(-20 - 22 - 14 + v_\pi(15)) = -15 + 0.25 v_\pi(15)$$
$$v_\pi(15) = -15/0.75 = -20$$

Changing the dynamics will not result the recalculation of the whole game: the Set S' of $S = 15$ is exactly as the one of $S = 13$. Thus, they must share the same state value as -20.

Here is my script implementation of this game. Feel free to add a $S = 15$ in it.

■

Exercise 4.3

$$
\begin{aligned}
q_\pi(s,a) &\doteq \mathbb{E}_\pi\big[G_t \,|\, S_t = s,\, A_t = a\big] \\
&= \mathbb{E}_\pi\big[R_{t+1} + \gamma G_{t+1} \,|\, S_t = s,\, A_t = a\big] \\
&= \mathbb{E}_\pi\big[R_{t+1} + \gamma \sum_{s',a'} q_\pi(s',a') \,|\, S_t = s,\, A_t = a\big] \\
&= \sum_{s',r} p(s',r|s,a)\Big[r + \gamma \sum_{a'} \pi(a'|s') q_\pi(s',a')\Big]
\end{aligned}
$$

$$
\begin{aligned}
q_{k+1}(s,a) &\doteq \mathbb{E}_\pi\big[R_{t+1} + \gamma G_{t+1} \,|\, S_t = s,\, A_t = a\big] \\
&= \sum_{s',r} p(s',r|s,a)\Big[r + \gamma \sum_{a'} \pi(a'|s') q_k(s',a')\Big]
\end{aligned}
$$

■

Exercise 4.4

In the step 3. Policy Improvement, it said:

If old-action$\neq \pi(s)$, then .....

It is a bug and one way to fix it is to say the following instead:

If old-action $\notin \{a_i\}$, which is the all equi-best solutions from $\pi(s)$, ...

■

Exercise 4.5

1. Initialization

$Q(s,a) \in \mathbb{R}$ and $\pi(s) \in A(s)$ arbitrarily for all $s \in S, a \in A$

2. Policy Evaluation

Loop:

$\Delta \leftarrow 0$

Loop for each $s \in S$ and $a \in A$:

$q = Q(s,a)$

$Q(s,a) \leftarrow \sum_{s',r} p(s',r|s,a)\Big[r + \gamma \sum_{a'} \pi(a'|s')Q(s',a')\Big]$

$\Delta \leftarrow \max(\Delta,\, |q - Q(s,a)|)$

until $\Delta < \theta$ (a small positive number determining the accuracy of estimation)

3. Policy Improvement

policy-stable $\leftarrow$ *true*

For each $s \in S$ and $a \in A$:

old-action $\leftarrow \pi(s)$

$\pi(s) \leftarrow \arg\max_a Q(s,a)$

If *old-action* $\notin \{a_i\}$, which is the set of equi-best solutions from $\pi(s)$

Then *policy-stable* $\leftarrow false$

If *policy-stable*, then stop and return $Q \approx q_*$ and $\pi \approx \pi_*$; else go to 2

■

Exercise 4.6

Step 3 Changes: We will only decide *policy-stable* is false under the condition that the policy does not explore.

Step 2 Changes: θ should not be set above the limit of any $soft$-ϵ method.

Step 1 Changes: π should be well defined as $soft$-ϵ method. ϵ should be given.

■

Exercise 4.7

Partial answer here. The programming implementation of DP is extremely time consuming. I did not get the exactly same answer from the book due to some unknown reason (algorithm difference/float precision etc.). Still, feel free to check it out and solve the whole picture! (Warning: It takes 30min to train.)

■

Exercise 4.8

The gambler's problem has such curious form of optimal policy because at the number 50, you can suddenly win with probability 0.5. Thus, the best policy will bet all when Capital=50 and the possible dividends of it, like 25.

Thinking capital of 51 as 50 plus 1. Of course we can bet all when we have 51, but the best policy is to see if we can earn much from the extra 1 dollar. If this return g is positive, we can say we have extra g money and bet it again until 75 when the sudden win chance is coming. That means, we have bonus opportunity of winning from 75. On the contrary, if we bet 50 out of 51 first, our chance of win is only ph and we lose the chance to reach 75. Instead, we will have to try our best to reach 25 with 1 dollar if we lose the bet, which is a much worse condition.

Conclusion: The indicated optimal policy creates more chances to win and guarantees the gambler be better off when he loses.

■

Exercise 4.9

Program here.

Plot A

Plot B

Plot C

Plot D

With proper thinking, you could easily recognize which plot is which. (Think of human playing technique!)

■

Exercise 4.10

$$\begin{aligned} q_{k+1}(s,a) &\doteq \mathbb{E}\big[R_{t+1} + \max_{a'} \gamma q_k(s',a')\big] \\ &= \sum_{s',r} p(s',r|s,a)\Big[r + \max_{a'} \gamma q_k(s',a')\Big] \end{aligned}$$

■