Solutions to Reinforcement Learning by Sutton Chapter 5

Yifan Wang

May 2019

Exercise 5.1

1. The estimated value jump up for the last two rows because the dealer never hits after having 17 or greater. It makes the winning rate extremely higher if the player gets more than 17.

2. It drop off for the whole last row on the left because if Dealer showing an ACE, It has very high possibility of getting higher score than the player when it counts as 11. Thus, the value of dealer's A has contained the dealer's winning rate of making it usable or not. Other cards have no such condition thus A is a special which makes the gap.

3. Frontmost values are higher in the upper diagrams because A represent dual values of being used as 1 and 11 in the upper diagram. It makes the player better off and is similar with the condition of having drop in the leftmost rows.

■

Exercise 5.2

No. Every visit will make the same result because in Blackjack, it is not possible of having same conditions in the series, unless the player sticks. But after making sticks, game is over without any additional new states.

One may think of after sticking, the dealer could hit and change the game. But the state is defined with current sum of player, dealer's one showing card and usable ace or not. That is saying, the last legal state S_T is after player's stick:

$$... \to S_{T-1} \to A_{stick} \to R_T \to S_T$$

You can think of dealer's after actions as given instantly as S_T but we only record until R_T. Thus, we cannot find any state exactly same in one episode even considering dealer's move.

■

Exercise 5.3

Drawing problem. I will not draw it here. The diagram has max of q_π in the bottom.

■

Exercise 5.4

Recall from Chapter 2.4:

$$
\begin{aligned}
Q_{n+1} &= \frac{1}{n}\sum_{i=1}^{n} R_i \\
&= \frac{1}{n}\Big(R_n + \sum_{i=1}^{n-1} R_i\Big) \\
&= \frac{1}{n}\Big(R_n + (n-1)\frac{1}{n-1}\sum_{i=1}^{n-1} R_i\Big) \\
&= \frac{1}{n}\Big(R_n + (n-1)Q_n\Big) \\
&= \frac{1}{n}\Big(R_n + nQ_n - Q_n\Big) \\
&= Q_n + \frac{1}{n}\Big(R_n - Q_n\Big)
\end{aligned}
$$

Then, the pseudo-code for Monte Carlo ES has similar improvement:

$$
\begin{aligned}
Q_n(S_t, A_t) &= \frac{1}{n}\sum_{i=1}^{n} G_i(S_t, A_t) \\
&= \frac{1}{n}\Big(G_n(S_t, A_t) + \sum_{i=1}^{n-1} G_i(S_t, A_t)\Big) \\
&= \frac{1}{n}\Big(G_n(S_t, A_t) + (n-1)\frac{1}{n-1}\sum_{i=1}^{n-1} G_i(S_t, A_t)\Big) \\
&= \frac{1}{n}\Big(G_n(S_t, A_t) + (n-1)\, Q_{n-1}(S_t, A_t)\Big) \\
&= \frac{1}{n}\Big(G_n(S_t, A_t) + nQ_{n-1}(S_t, A_t) - Q_{n-1}(S_t, A_t)\Big) \\
&= Q_{n-1}(S_t, A_t) + \frac{1}{n}\Big(G_n(S_t, A_t) - Q_{n-1}(S_t, A_t)\Big)
\end{aligned}
$$

$\blacksquare$

Exercise 5.5

Because

$$|J(s)| = \sum_{t \in J(s)} \rho_{t:T(t)-1} = 10$$

$$\rho_{t:T(t)-1} = 1$$

We have off-policy = on-policy, thus:

first-visit:

$$v_s = 10$$

all-visit:

$$v_s = \frac{1}{10}(1+2+3+4+5+6+7+8+9+10) = 5.5$$

■

Exercise 5.6

eq (5.6):

$$V(s) \doteq \frac{\sum_{t \in J(s)} \rho_{t:T(t)-1} G_t}{\sum_{t \in J(s)} \rho_{t:T(t)-1}}$$

For Q:

$$Q(s,a) \doteq \frac{\sum_{t \in J(s,a)} \rho_{t:T(t)-1} G_t}{\sum_{t \in J(s,a)} \rho_{t:T(t)-1}}$$

■

Exercise 5.7

The weighted average algorithm will need few episodes to decrease its bias. Especially when the $\rho_{t:T(t)-1}$ is big, the weighted average algorithm would be shifted by those data. When we get enough episodes, the average begins to be stable and decreasing the bias.

$\blacksquare$

Exercise 5.8

For first visit, we have:

$$\mathbb{E}_b\left[\left(\prod_{t=0}^{T-1}\frac{\pi(A_t|S_t)}{b(A_t|S_t)}G_0\right)^2\right]$$

For every visit, we will some difference

$$\mathbb{E}_b\left[\left(\frac{1}{T-1}\sum_{k=1}^{T-1}\prod_{t=0}^{k}\frac{\pi(A_t|S_t)}{b(A_t|S_t)}G_0\right)^2\right]$$

$$\begin{aligned}
&= 0.5\,(\text{policy possibility})\cdot 0.1\,(\text{length of 1 possibility})\cdot 2^{2\cdot 1}\,(\text{square of }\rho)\\
&+\frac{1}{2}\big[0.5\cdot 0.9\cdot 0.5\cdot 0.1\cdot 2^{2\cdot 2}(\text{second visit})+0.5\cdot 0.1\cdot 2^{2\cdot 1}(\text{first visit})\big]\\
&+\frac{1}{3}\big[0.5\cdot 0.9\cdot 0.5\cdot 0.9\cdot 0.5\cdot 0.1\cdot 2^{2\cdot 3}(\text{third visit})\\
&+0.5\cdot 0.9\cdot 0.5\cdot 0.1\cdot 2^{2\cdot 2}(\text{second visit})+0.5\cdot 0.1\cdot 2^{2\cdot 1}(\text{first visit})\big]\\
&\cdots\\
&=0.1\sum_{k=1}^{\infty}\frac{1}{k}\sum_{l=0}^{k-1}0.9^l\cdot 2^l\cdot 2\\
&=0.2\sum_{k=1}^{\infty}\frac{1}{k}\sum_{l=0}^{k-1}1.8^l=\infty.
\end{aligned}$$

On the other hand, considering weighted average method:

$$\mathbb{E}_b\left[\left(\frac{1}{|J(s)|}\sum_{k=1}^{T-1}\prod_{t=0}^{k}\frac{\pi(A_t|S_t)}{b(A_t|S_t)}G_0\right)^2\right]$$

$$\begin{aligned}
&\dots \\
&\leq 0.1\sum_{k=1}^{\infty}\frac{1}{k}\frac{1}{(\sum_{i=0}^{k}2^i)^2}\sum_{l=0}^{k-1}0.9^l\cdot 2^l\cdot 2 \\
&\leq 0.2\sum_{k=1}^{\infty}\frac{1}{k}\frac{1}{2^{2k}}\sum_{l=0}^{k-1}1.8^l \\
&\leq 0.2\sum_{k=1}^{\infty}\frac{1}{k}2^{-2k}k\,1.8^k \\
&\leq \sum_{k=1}^{\infty}\frac{1}{k}2^{-2k}k\,2^k \\
&= \sum_{k=1}^{\infty}2^{-k} = 1
\end{aligned}$$

■

Exercise 5.9

Similar with *Exercise 5.4:*

$$\begin{aligned}
V_n(S_t) &= \frac{1}{n}\sum_{i=1}^{n}G_i(t) \\
&= \dots \\
&= V_{n-1}(S_t) + \frac{1}{n}\Big(G_n(t) - V_{n-1}(S_t)\Big)
\end{aligned}$$

■

Exercise 5.10

$$
\begin{aligned}
V_{n+1} &\doteq \frac{\sum_{k=1}^{n} W_k G_k}{\sum_{k=1}^{n} W_k} \\
&= \frac{W_n G_n + \sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n-1} W_k} \, \frac{\sum_{k=1}^{n-1} W_k}{\sum_{k=1}^{n} W_k} \\
&= \Big[\frac{W_n G_n}{C_{n-1}} + V_n\Big] \frac{C_{n-1}}{C_n} \\
&= \frac{W_n G_n}{C_n} + \frac{V_n C_{n-1}}{C_n} \\
&= V_n + \frac{W_n G_n}{C_n} + \frac{V_n C_{n-1}}{C_n} - V_n \\
&= V_n + \frac{W_n G_n}{C_n} + \frac{V_n C_{n-1} - V_n C_n}{C_n} \\
&= V_n + \frac{W_n G_n}{C_n} + \frac{-V_n W_n}{C_n} \\
&= V_n + \frac{W_n}{C_n} \big[G_n - V_n\big]
\end{aligned}
$$

■

Exercise 5.11

Because π, the target policy, is deterministic and is redefined as the $\arg\max_a Q(S_t, a)$ **just before the update of W happens. Thus we always have $\pi(A_t|S_t) = 1$.**

■

Exercise 5.12

Programming problem. See Github ■

Exercise 5.13

from 5.12

$$\rho_{t:T-1}R_{t+1} = \frac{\pi(A_t|S_t)}{b(A_t|S_t)} \frac{\pi(A_{t+1}|S_{t+1})}{b(A_{t+1}|S_{t+1})} \frac{\pi(A_{t+2}|S_{t+2})}{b(A_{t+2}|S_{t+2})} \cdots \frac{\pi(A_{T-1}|S_{T-1})}{b(A_{T-1}|S_{T-1})} R_{t+1}$$

However, we have 5.13

$$\mathtt{E}\left[\frac{\pi(A_t|S_t)}{b(A_t|S_t)}\right] \doteq \sum_a \pi(a|S_k) = 1$$

And, we know after t, any importance-sampling ratio becomes independent with R_{t+1}. This must follows:

$$\mathtt{E}(XY) = \mathtt{E}(X)\mathtt{E}(Y)$$

Go back to 5.12 using 5.13, we will have 5.14 indeed:

$$\mathtt{E}[\rho_{t:T-1}R_{t+1}] = \mathtt{E}[\rho_t R_{t+1}]$$

■

Exercise 5.14

Skip ■