The University of Vermont

CS253A QR: Reinforcement Learning
Assignment №6

Ayat Ospanov

October 10, 2018

Contents

1 Exercise 5.3

$(S_t, A_t) \rightarrow (S_{t+1}) \rightarrow (A_{t+1}) \rightarrow \cdots \rightarrow \blacksquare$

2 Exercise 5.4

The update formula will be altered as follows:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{1}{n}(G - Q(S_t, A_t))$$

Thus, we need to keep the number of times (n) we saw (S_t, A_t). So, instead of keeping n numbers, we keep 1, and instead of n operations we do 3 (+, -, /)

3 Exercise 5.5

As when we go to the terminal state an episode ends, and as we have the episode of length 10, we have the next episode: $S \rightarrow S \rightarrow \cdots \rightarrow \blacksquare$

As for each transition we get +1, G_t will be the amount of steps until the end of the episode. For first-visit estimation we have only one $G_t, t = 0$, and get

$$V(S) = \frac{G_0}{1} = 10$$

. For every-visit estimation we should average all of them and in this case

$$V(S) = \frac{G_0 + G_1 + \cdots + G_9}{10} = \frac{10 + 9 + \cdots + 1}{10} = 5.5$$

4 Exercise 5.6

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

The only change is $\rho_{t+1:T(t)-1}$ while for V(s) we had $\rho_{t:T(t)-1}$. It happens because we don't have the probability of getting action a as we already took it.

5 Exercise 5.7

For the weighted importance-sampling method error first increased and then decreased, because at the first steps we have low variance as we estimate behavior policy and we know the policy. But error starts increasing later as we start converging to the target policy. On the other hand, ordinary importance-sampling estimates the target policy from the beginning, thus we have high variance and high error. But starting at some point, both methods start decreasing as we start estimating target policy more accurate.

6 Exercise 5.8

When we use every-visit method, at every n-length episode we add all previous length episode (n-1, n-2, ..., 1) besides what we added in first-visit method. Schematically it would be as follows:

$$\begin{aligned}
\mathbb{E}[...] = &\\
&\text{length 1 episode}\\
&+\text{length 2 episode}+\text{length 1 episode}\\
&+\text{length 3 episode}+\text{length 2 episode}+\text{length 1 episode}\\
&...
\end{aligned}$$

If we look at it column-wise, we can see that we are adding ∞ to ∞ ∞ times and thus get ∞.

7 Exercise 5.9

The modification will be as in Exercise 5.4:

$$V(S_t) \leftarrow V(S_t) + \frac{1}{n}(G - V(S_t))$$

where n is the number of times we met S_t in all episodes.

8 Exercise 5.10

$$\begin{aligned}
V_{n+1} = \frac{\sum\limits_{k=1}^{n} W_k G_k}{\sum\limits_{k=1}^{n} W_k} = \frac{\sum\limits_{k=1}^{n-1} W_k G_k + W_n G_n}{\sum\limits_{k=1}^{n} W_k} &=\\
= \frac{\sum\limits_{k=1}^{n-1} W_k G_k}{\sum\limits_{k=1}^{n} W_k} + \frac{W_n G_n}{\sum\limits_{k=1}^{n} W_k} = \frac{\sum\limits_{k=1}^{n-1} W_k \sum\limits_{k=1}^{n-1} W_k G_k}{\sum\limits_{k=1}^{n-1} W_k \sum\limits_{k=1}^{n} W_k} + \frac{W_n G_n}{\sum\limits_{k=1}^{n} W_k} &=\\
= \frac{\sum\limits_{k=1}^{n-1} W_k}{\sum\limits_{k=1}^{n} W_k} V_n + \frac{W_n G_n}{\sum\limits_{k=1}^{n} W_k} == \frac{\sum\limits_{k=1}^{n} W_k - W_n}{\sum\limits_{k=1}^{n} W_k} V_n + \frac{W_n G_n}{\sum\limits_{k=1}^{n} W_k} &=\\
= (1 - \frac{W_n}{C_n})V_n + \frac{W_n}{C_n} G_n = V_n + \frac{W_n}{C_n}\Big[G_n - V_n\Big]&
\end{aligned}$$

9 Exercise 5.11

As we converge over time, target policy will do action A_t on state S_t when we get the optimal policy. Thus, probability of this is 1, i.e. $\pi(A_t|S_t) = 1$. Thus instead of $\pi(A_t|S_t)$ they put its limit value, i.e. 1. I think it will increase convergence speed.