Attempts to exercise in Reinforcement Learning book Chapter 5

Mengliao Wang

July 16, 2017

Exercise 5.1:

The last two rows are state with hand 20 and 21, which according to the policy will stop. Thus these states will never take the risk of fail comparing to other states, and resulting in much higher values.
For the most left column which means opponent shows a card ace, it indicates that opponent might have a usable ace. As we can tell from the state values, that states with a usable ace is normally higher than the states without a usable ace, because it gives more flexibility and reduces the risk of going bust. That is also why the most font row in the upper diagram has higher values than the lower diagram.

Exercise 5.2:

Figure 1: The backup diagram for q_π estimation

The backup diagram for q_π is shown as in Figure 1

Exercise 5.3:

According to the definition of $q_\pi(s,a)$, we can easily have as the following anaglos formula. Here we updated the timing for ratio ρ to from t to $t+1$, because the action for time t is already determined.

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

Exercise 5.4:

The MSE increased at the beginning because the weighted estimate is bias towards V_b with few training samples (with only one sample the expectation is exactly V_b). Here the behaviour policy is randomly stick or hit, which is very different from the target policy (stick only on a sum of 20 or 21). Thus initially the weighted estimate shows a bigger difference from the real values for target policy V_π. However as the episode number increases, estimate will converge to the expectation of V_π instead of V_b, thus the MSE decreased later.

Exercise 5.5:

Yes the variance of the estimate will still be infinite. For both method, $\mathbb{E}[X^2]$ can be written as:

$$\mathbb{E}\left[\left(\sum_{k\in\mathcal{T}(s)}\prod_{t=k}^{T-1}\frac{\pi(A_t|S_t)}{\mu(A_t|S_t)}G_k\right)^2\right] \tag{2}$$

For first-visit method we will define $\mathcal{T}(s)$ to be the first time we observe state s, which is obviously time step 0, then we have the same expression as in the book. For every-visit method, we will define $\mathcal{T}(s)$ to be the all the times we could observe state s, which obviously is 0, 1, 2, ... Thus for every-visit method it is:

$$\mathbb{E}\left[\left(\sum_{k=0}^{T-1}\prod_{t=k}^{T-1}\frac{\pi(A_t|S_t)}{\mu(A_t|S_t)}G_k\right)^2\right] \tag{3}$$

Follow the same calculation as shown in the book can find that $\mathbb{E}\left[\left(\prod_{t=k}^{T-1}\frac{\pi(A_t|S_t)}{\mu(A_t|S_t)}G_k\right)^2\right], \forall k <= T-1$ is the same as $\mathbb{E}\left[\left(\prod_{t=0}^{T-1}\frac{\pi(A_t|S_t)}{\mu(A_t|S_t)}G_0\right)^2\right]$. Thus obviously the whole expectation Expression (3) is still be infinite.

Exercise 5.6:

To make the algorithm for first-visit method, we would need to hold the updated $Q(S_t, A_t)$ in a placeholder $H(S_t, A_t)$, and keep overwriting $H(S_t, A_t)$ during the for loop, so eventually $H(S_t, A_t)$ will only hold the values for the first time visit of (S_t, A_t). Then we assign it to $Q(S_t, A_t)$ outside of the for loop. See the modified algorithm below:

Initialize, for all $s \sin \mathcal{S}, a \in \mathcal{A}(s)$:
- $Q(s,a) \leftarrow$ arbitrary
- $H(s,a) \leftarrow Q(s,a)$
- $C(s,a) \leftarrow 0$
- $\mu(a \mid s) \leftarrow$ an arbitrary soft behavior policy
- $\pi(a \mid s) \leftarrow$ an arbitrary target policy

repeat
- Generate an episode using μ
  - $S_0, A_0, R_1, ..., S_{T-1}, A_{T-1}, R_T, S_T$
- $G \leftarrow 0$
- $W \leftarrow 1$
- **for** $t = T-1, T-2, ...$ down to 0: **do**
  - $G \leftarrow \gamma G + R_{t+1}$
  - $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$

$H(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)}[G - Q(S_t, A_t)]$

$W \leftarrow W \frac{\pi(A_t \mid S_t)}{\mu(A_t \mid S_t)}$

if $W = 0$: **then**

ExistForLoop

for Each unique pair S_t, A_t appeared in the episode: **do**

$Q(S_t, A_t) \leftarrow H(S_t, A_t)$

until $1 = 1$

Exercise 5.7:

According to 5.6 we have:

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

Exercise 5.8:

Program attached. Here are the results after learing 50, 200, 1000, 10000 episodes in Figure 2. The green area is the track, and the yellow line is the destination.

Exercise 5.9:

The algorithm would be:

Initialize, for all $s \sin \mathcal{S}, a \in \mathcal{A}(s)$:

$Q(s, a) \leftarrow$ arbitrary

$C(s, a) \leftarrow 0$

$\pi(a \mid s) \leftarrow$ a deterministic policy that is greedy with respect to Q

repeat

Generate an episode using any soft policy μ

$S_0, A_0, R_1, ..., S_{T-1}, A_{T-1}, R_T, S_T$

$G \leftarrow 0$

for $t = T - 1, T - 2, ...$ down to 0: **do**

$\rho \leftarrow 1$

$\bar{G} \leftarrow 0$

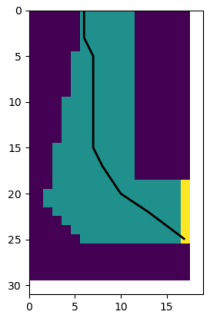
$W \leftarrow 0$

for $k = t + 1, t + 2, ...$ until $T - 1$: **do**

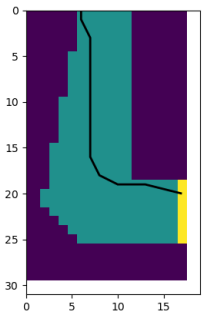
$\bar{G} \leftarrow \bar{G} + R_k$

$\rho \leftarrow \rho \frac{\pi(A_k \mid S_k)}{\mu(A_k \mid S_k)}$

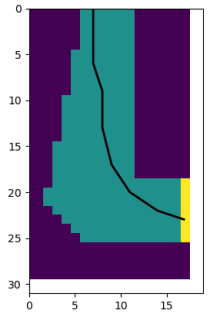
$W \leftarrow W + (1 - \gamma)\gamma^{k-t-1}\rho$



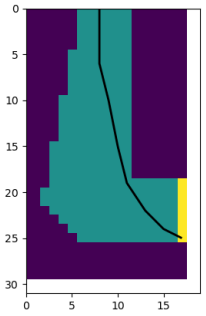
(a) Results after training with 50 episodes



(b) Results after training with 200 episodes



(c) Results after training with 1000 episodes



(d) Results after training with 10000 episodes

Figure 2: RaceTrack results after training with 50, 200, 1000, 10000 episodes

$G \leftarrow G + W\bar{G}$

$W \leftarrow W + \gamma^{T-t-1}\rho$

$G \leftarrow G + W\bar{G}$

$C(S_t, A_t) = C(S_t, A_t) + W$

$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)}[G - Q(S_t, A_t)]$

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

if $A_t \neq \pi(S_t)$: **then**

ExistForLoop

until $1 = 1$