RL Assignment 3 Theory

Apoorv Khattar
2016016

Ques 1. Exercise 5.4

Similar to the bandit problem where we maintained avg. return and count for each bandit, here we keep avg. return and count for each state action pair.

Now when we get a new return G_n for state S_t and action A_t,

$$Q_n(S_t, A_t) = \frac{\left(\overset{Q_{n-1}(S_t,A_t)}{\text{Avg.Return}(S_t, A_t)} * \overset{(n-1)}{\text{Count}(S_t, A_t)} + G_n \right)}{\text{Count}(S_t, A_t) + 1}$$

$$= \frac{Q_{n-1}(S_t, A_t)\{n-1\} + G_n}{n}$$

$$= Q_{n-1}(S_t, A_t) + \frac{1}{n}\left(G_n - Q_{n-1}(S_t, A_t)\right)$$

[* The code for figure 5.1 uses this idea as maintaing the list in my code for all returns.]

Pseudocode: Same as ES with slight change

1. Initialize $\pi(s)$, $Q(s,a)$, $\text{Count}(S_t, A_t) = 0$
2. Generate a random episode with exploring starts, $\pi(s)$
3. $G = 0$
4. Loop for $t = T-1, T-2 \dots 0$:

   (a) $G = \gamma G + R_{t+1}$

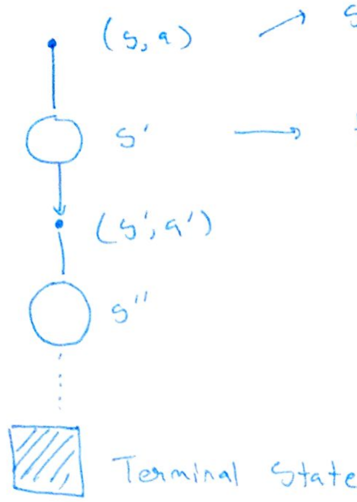
   (b) $\text{Count}(S_t, A_t) \mathrel{+}= 1$

   (c) $Q(S_t, A_t) \mathrel{+}= \frac{1}{\text{Count}(S_t, A_t)}\left(G - Q(S_t, A_t)\right)$

   (d) $\pi(S_t) = \underset{a}{\text{argmax}}\left(Q(S_t, a)\right)$

This is same as code provided in Sutton but with a slight change

Ques 2. Back up Diagram for 5.3 :-



(s, a) → starting state for one episode

s' → transition to new state s' based on (a)

Ques 3. Exercise 5.6

Equation 5.6 ⇒ $V(s) = \dfrac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1} G_t}{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1}}$

Analogous Equation for $Q(s,a)$ is given by

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

Ques 6.

(a) Exercise 6.3:-

In the first episode, based on the graph we know that A is visited and then new state is terminal state 1 with reward 0 as a result,

$$V(A) = V(A) + (0.1)\{0 + V(\text{Terminal}) - V(S)\}$$
$$= 0.5 + (0.1)\{-0.5\}$$
$$= 0.45$$

⇒ $V(A)_{new} \neq V(A)_{initial}$

Consider any other state, let it be B

$V(B) = V(B) + (0.1)\{0 + V(A) - V(B)\} = 0$ $\{V(A) = V(B)$ initially$\}$

(b) Exercise 6.4

~~Considering α ... upon running simulations~~

Alphas for the two algorithms are very small and
α provides weight to the reward earned at time t.
Higher values of α would lead to greater fluctuations
and a lower value would lead to smooth RMSE plots.

⇒ Since α are small enough thus there will not be
any other α that would significantly improve either
algorithm.

(c) Exercise 6.5

After large number of episodes, $V(S') - V(S)$ becomes
constant as they have converged so $V(S)$ is updated by
α (r) and $V(S') - V(S) \to 0$, this affects the estimates
of $V(S)$ thus error goes up.

Ques 8. Exercise 6.12

Even if action picking is greedy in Q-Learning then
also it won't be the same.
In SARSA, we select A' for new state S' then
update $Q(S,A)$ but in Q-Learning we update
$Q(S,A)$ using $\max_a (Q(S',a))$

⇒ Thus they won't be same

Ques 5 Exercise 6.2

In TD bootstrapping occurs, thus on change in building
we have some new states and some old states. For old
states, we can use previously determined value as
it is close to true value. This will lead to faster
convergence due to bootstrapping.