Reinforcement Learning
Homework 3

__/__/__

Ques 8 Exercise 6.12

Q learning is not same as SARSA if the action selection is greedy. This is because Q learning is off policy while SARSA is on policy. In Q learning the update equation is:

$Q(s,a) = Q(s,a) + \alpha [R + \max_{a'} Q(s',a') - Q(s,a)]$.

Here, (a) is found using ε greedy (or any other behaviour policy) s' is then observed from s,a. s is updated to s' the but a is not updated to greedy a' selection ∴ although we take the greedy action in each update, our underlying policy is still the same. On the other hand in SARSA, update eqn is

$Q(s,a) = Q(s,a) + \alpha [R + Q(s',a') - Q(s,a)]$

where a' → ε greedy (next greedy) policy. Hence, both s and a are updated to s' and a'. ∴ the Q values used for updation are different ∴ the α are different.

Ques 1 . For calculating the average value, we donot need to store all the returns for a given s,a and recompute mean every time. The mean can be easily calculated given the current mean and number of times the (s,a) has been visited.

Say current mean $\bar{x} = \frac{v_1 + v_2 + \cdots v_n}{n}$ ①

new mean $\bar{x}' = \frac{(v_1 + v_2 + \cdots + v_n) + v_{n+1}}{n+1}$ ②

From ①, $v_1 + v_2 + \cdots v_n = \boxed{\bar{x} \times n}$, put into ②,

$$\bar{x}' = \boxed{\frac{\bar{x}\,n + v_{n+1}}{n+1}}$$

2. the line $Q(S_t, A_t) \leftarrow$ Average (Returns (S_t, A_t))

can be modified as $Q(S_t, A_t) \leftarrow \frac{Q(S_t, A_t) \times (\text{count}(S,A) - 1) + G}{\text{count}(S,A)}$

Pseudo code :

Initialize

$\pi(s) \in A(s) \quad \forall s \in S$

$Q(s,a) \in R \quad \forall s \in S, a \in A(s)$.

$count(s,a) = 0 \quad \forall s \in S, a \in A(s)$.

loop forever:

Choose $S_0 \in S, A_0 \in A(S_0)$ randomly

Generate episode S_0, A_0 following π ?

$G \leftarrow 0$.

loop for each step of episode, $t = T-1, T-2, \ldots 0$:

#get S_t, A_t of episode

$count(S_t, A_t) += 1$

$G \leftarrow \gamma G + R_{t+1}$

unless S_t, A_t appears in $S_0, A_0, S_1, A_1, \ldots S_{t-1}, A_{t-1}$:

$Q(S_t, A_t) = \left(Q(S_t, A_t) \times (count(S_t, A_t) - 1) + G\right) / count(S_t, A_t)$

$\pi(S_t) \leftarrow \underset{a}{\text{argmax}}\, Q(S_t, a)$

ques 3 Exercise 5-6

$$V(s) = \frac{\sum_{t \in \tau(s)} \rho_{t:T(t)-1} G_t}{\sum_{t \in \tau(s)} \rho_{t:T(t)-1}}$$ in terms of $Q(s,a)$.

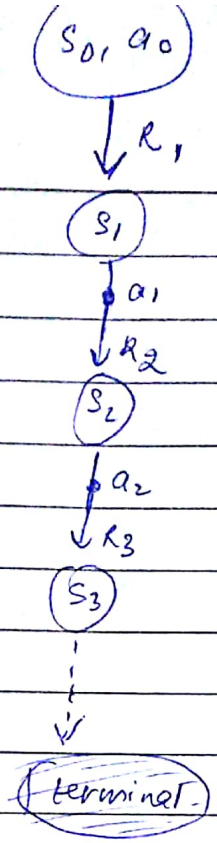
$\tau(s)$ determines the starting time / first time a ~~state~~ state s was visited in the sequence of states.

Now we need $\tau(s,a)$, for first visit of tuple s, a

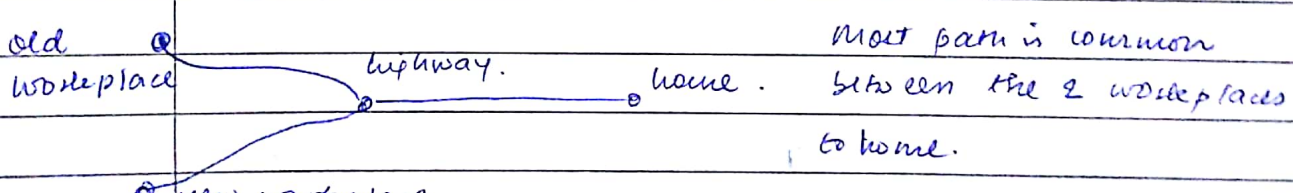
$$Q(s,a) = \frac{\sum_{t \in \tau(s,a)} \rho_{t+1:T(t)} G_t}{\sum_{t \in \tau(s,a)} \rho_{t+1:T(t)}}$$ → modified expression.

ques 2 Backup diagram for MC estimation of q_π

$Q(s,a)$ → start with s, a tuple



Ques Yes a same thing happens for the original scenario. Here, if we move to a new work place, the situation is somewhat like:



Since TD updates values at each step and does not need to wait for the episode to end, there will not be a large change in values of states common to both parts. As a result, optimal values can be found faster in case of TD.