Hom Work 3

Ans 1

Let count (S_t, A_t) denotes the number of returns of (S_t, A_t) we have got so for.

$$Q(S_t, A_t) = \frac{\sum_{r \in Returns(S_t, A_t)} r}{count(S_t, A_t)}.$$

~~Let r' be the last return of~~

Now suppose we get a new return r', then let the new Q be $Q'(S_t, A_t)$

$$Q'(S_t, A_t) = \frac{\sum_{r \in Return(S_t, A_t)} r + r'}{count(S_t, A_t) + 1}.$$

$$Q'(S_t, A_t) = \frac{Q(S_t, A_t)\, count(S_t, A_t) + r'}{count(S_t, A_t) + 1}$$

and we will update $count(S_t, A_t)$ by 1

∴ pseudo code would be →

Initialize:

$\pi(s) \in A(s)$ (arbitrarily), for $\forall s \in S$

$Q(S, a) = 0 \quad \ldots\ldots\ldots, a \in A(S)$

$count(S, a) = 0 \quad \ldots\ldots\ldots$

Loop forever (for each episode):

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

Generate episode from $S_0 A_0 \ldots$

$G \leftarrow 0$

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

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

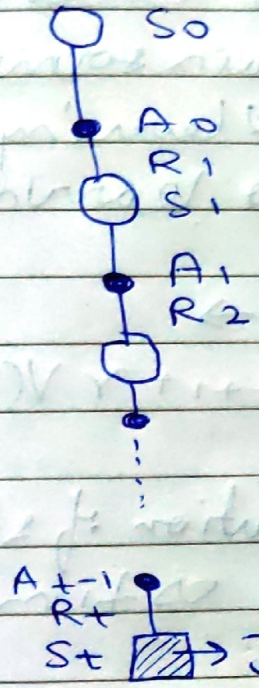
Unless the pair S_t, A_t appears in $S_0, A_0, S_1, A_1, \ldots S_{t-1}, A_{t-1}$:

$$Q(S_t, A_t) = \frac{Q(S_t, A_t)\, count(S_t, A_t) + G}{count(S_t, A_t) + 1}$$

$count(S_t, A_t) += 1$

$\pi(S_t) = \arg\max_a Q(S_t, a)$.

A.2



A.3

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

A.5

→ code

Ans5 TD learning will be better than MC learning if the predicted time to go is very close to the actual value.

Now suppose you have very good estimate (almost true) of predicted time to go for all the states.

Now you change your office, in this case only the first state & 2nd state would change but the estimation of predicted time from the moment he enters highway and thereafter will remain same.

Thus in this case TD learning would be better than Monte carlo learning.

The update in TD learning is →

$$V(S_t) = V(S_t) + \alpha [R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$$

Here better is the estimation of state $V(S_{t+1})$, better would be estimate of $V(S_t)$.

Thus if we know the true value of $V(S_{t+1})$ we can easily calculate the true value of $V(S_t)$. Thus, the ~~example~~ situation we considered is better for TD.

Ans8 No, Q-learning and Sarsa wouldn't be same in case of Greedy selection. This is because according to the pseudo code of both the methods, Sarsa first chooses the action and then updates Q whereas Q-learning first updates the value

of a and then chooses the action.