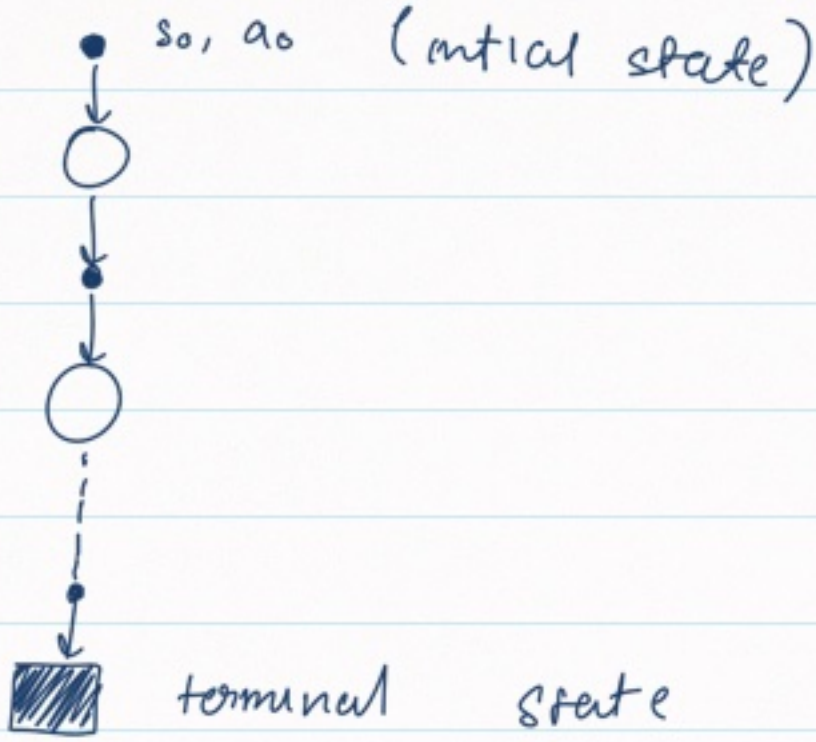
RL HW-3 Theory

Name :- Vishal Samal

Rollno :- 2017209.

2.



5. For the given example where the path is now different initially TD will perform better than MC. This because the update rule for MC is,

$$V(S_t) \leftarrow V(S_t) + \alpha\left[G_t - V(S_t)\right]$$

which would require as to have an entire sequence to calculate state values.

Whereas, from TD the update rule is,

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

which requires only the step reward to do this.

Due to this initially TD will be better.

8. Update rule for SARSA,

$$Q(S,a) \leftarrow Q(S,a) + \alpha\left[R + \gamma Q(S',a') - Q(S,a)\right]$$

where $a \in A(s)$, $a' \in A(s')$

and they selected using ϵ-greedy,

Update rule for Q-Learning,

$$Q(S,a) \leftarrow Q(S,a) + \alpha\left[R + \gamma \max_a (Q(S,'a)) - Q(S,a)\right]$$

where $a \in A(S)$

and selected using ϵ-greedy.

Now, if ϵ-greedy has $\epsilon = 0$ that is, we are following greedy policy. The all action values are optimal action values,

$$\Rightarrow \quad Q(s', a') = \max_{a} \left(Q(s', a) \right)$$

Thus, they both will be equal.

P.T.O

1. Pseudo code,

Initialise,

$\pi(s) \in A(s)$ [arbitrary] $\forall s \in S$

$Q(s,a) \in R$ [arbitrary] $\forall s \in S,\ a \in A(s)$

$N(s,a) \leftarrow 0$ $\forall s \in S,\ a \in A(s)$

Loop forever (for each episode):

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

Generate episode from S_0, A_0

$G \leftarrow 0$

Loop for each step, $t = T-1, T-2, \ldots, 0$

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

$N(S_t, A_t) \leftarrow N(S_t, A_t) + 1$

$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{1}{N(S_t, A_t)} \left[G_t - Q(S_t, A_t) \right]$

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

We would a variable $N(S_t, a)$ which stores how many times we visit, any state - action pair.

From section 2.4 we know the update rule and we can incorporate that for MC by replacing rewards with returns.

3. For given target policy $\pi(a|s)$ and behaviour $b(a|s)$,

$$V_b(s) = E_b[G_{t+1} \mid S_t = s]$$

If importance sampling ratio is,

$$P_{t:T-1} = \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{b(A_k|S_k)}$$

Then, $$V_\pi(s) = E_\pi[P_{t:T-1} G_{t+1} \mid S_t = s]$$

Similarly,

$$q_\pi(s,a) = E_\pi[P_{t:T-1} G_{t+1} \mid S_t = s, A_t = a]$$

We can define a set of all time steps where s, a action pair was visited, $\mathcal{T}(s,a)$

then, Ordinary importance sampling,

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

And, weighted importance sampling,

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