Date: __/__/____

RL - Reinforcement Learning - CSE564
HW-3
Theory

NAME: Ansh Arora
ROLL NO: 2019022

(1) We have to make the code more efficient. We can do this by maintaining a mean and a count for each state-action pair and just update them incrementally.

Initialize:
$\pi(s) \in A(s)$ arbitrarily $\forall s \in S$
$Q(s,a) \leftarrow 0 \quad \forall s \in S, a \in A(s)$
$count(s,a) \leftarrow 0 \quad \forall s \in S, a \in A(s)$
Loop forever:
Choose $S_0 \in S, A_0 \in A$ (A_0 per $\in S$)
Generate an episode $S_0, A_0, R_1, S_1, A_1, R_2 \ldots, R_T$ by π.
$G \leftarrow 0$
Loop for $t = T-1, T-2, \ldots, 0$
$G \leftarrow \gamma G + R_{t+1}$
If S_t, A_t not in $S_0, A_0 \ldots S_{t-1}, A_{t-1}$:
$Q(S_t, A_t) \leftarrow (count(S_t, A_t) * Q(S_t, A_t) + G) / (count(S_t, A_t) + 1)$
$count(S_t, A_t) \leftarrow count(S_t, A_t) + 1$
$\pi(S_t) \leftarrow \underset{a}{\text{argmax}}\, Q(S_t, a)$

Update step of Q is -

$$Q \leftarrow \frac{nQ + G}{n+1}$$

$$n \leftarrow n+1$$

where n is number of times (S_t, A_t) has been encountered before.

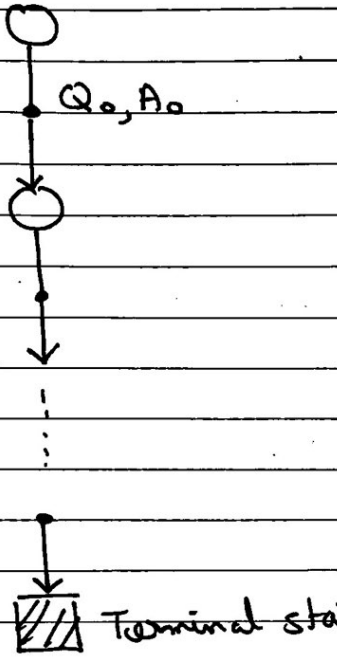
$$Q_{new} = \frac{G_1 + \ldots + G_n}{n}$$

$\Rightarrow G_1 + \ldots + G_n = nQ \quad \therefore \frac{nQ + Q_{new}}{n+1}$ is new mean.

#learnthesmarterway

Date: __/__/____

② Backup diagram for Monte Carlo estimation of q_π.



③ Weighted Importance Sampling—

$$V(s) \doteq \frac{\sum_{t \in \mathcal{J}(s)} \rho_{t:T(t)-1} G_t}{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1}}$$

We have to find equation analogous to this for action value pairs $Q(s,a)$

$$Pr\{S_t, A_t, S_{t+1}, A_{t+1}, \ldots, S_T \mid S_t = s_0, A_t = a\}$$

$$= P(S_t \mid A_t) \times P(S_{t+1} \mid S_t, A_t)\, \pi(A_{t+1} \mid S_{t+1}) \ldots P(S_T \mid S_{T-1}, A_{T-1})$$

$$= \prod_{t+1}^{T-1} \left(\pi(a_i \mid s_i)\, p(s_{i+1} \mid s_i, a_i)\right) \times p(s_{t+1} \mid s_t, a_t)$$

Date: __/__/____

Relative probability-

$$\rho_{t:T-1} \doteq \frac{\prod_{t+1}^{T-1} \pi(a_i|s_i)}{\prod_{t+1}^{T-1} b(a_i|s_i)}$$

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

(5) We have to imagine a scenario in which TD update would be better.

We take the example where our path to home after the highway point remains the same, but the building where car is parked has changed. Thus,

Initial → (A) → (B) → (C) → (D) → (E)

Now → (X) → (B) → (C) → (D) → (E)

where A is old office state
X is new parking state

We take the assumption that we have a lot of experience.

For MC Methods-

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

We generate complete episode to calculate G_t. Since G_t can have high variance given that

Date: __/__/____

it takes accounts for all the episodes in the state, it will take a long time for V to converge to v_π.

For TD Methods-

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

Since we have a large amount of experience, the value of $V(S_{t+1})$ would have low variance from the actual value. So even though the path taken is changed, not much variance is observed and V converges to v_π.

(6) a) Only the value of V(A) changes and becomes lesser than its initial value, indicating that the episode terminated on the left end.

The change to A can be explained using TD δ.

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

Now, $\gamma = 1$, $V(S_{t+1}) = 0$ (Left Terminal), $V(S_t) = 0.5$

$$\therefore \delta_t = 0 + 0 - 0.5 = -0.5$$

Taking $\alpha = 0.1$,

$$V(A) \leftarrow V(A) + \alpha \times \delta_t$$

$$\Rightarrow V(A) \leftarrow 0.5 + 0.1 \times (-0.5)$$

$$\Rightarrow V(A) = \underline{\underline{0.45}}$$

For other states, $\delta_t = 0 + \gamma \times 0.5 - 0.5$

$$= 0$$

$\therefore$ No change is observed.

Date: __/__/____

b) On increasing α rates too much, larger step updates take place that will result into non-converging noisy curves (as can be seen for $\alpha = 0.03$ and $\alpha = 0.04$). Thus further changing values of α would give a better result, and TD does perform better than MC.

c) This happens because of increase in error in $V(C)$. We initialize $V(C)$ to the correct value of $v_\pi(C)$. However, as we experiment more, this value diverges more and more, thus increasing the RMSE. For higher α, this effect is even higher since the change in $V(C)$ is higher for each step of TD learning.