SAHAS MARWAH

202237

HW-3

1. To make the pseudocode more efficient, to calculate the mean it would be better to maintain the mean and the count for each state-action pair and update them incrementally.

Acc. to section 2.4 of the course book; given Q_n and R_n, the new avg. for all n rewards can be found by : $Q_{n+1} = Q_n + \left(\frac{R_n - Q_n}{n}\right)$

Generally,

New Est. = Old est + step size (Target − Old Est)

update MC ES :

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

$Q(s,a) \in \mathbb{R} \quad \forall\ s \in S,\ a \in A \quad = 0$

$\text{counter}(s,a) \quad \forall\ s \in S,\ a \in A \quad = 0$

Loop forever :

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

Generate $S_0, A_0, R_1, S_1, A_1, R_2 \ldots S_{T-1}, A_{T-1}, R_T$.

$G = 0$

loop $t = T-1 : -1 : 0$

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

until S_t, A_t appears in $S_0, A_0, S_1, A_1 \ldots$:

$$Q(S_t, A_t) = \frac{Q(S_t, A_t) \times \text{Counter}(S_t, A_t) + G}{\text{Counter}(S_t, A_t) + 1}$$

$\text{Counter}(S_t, A_t) = \text{Counter}(S_t, A_t) + 1$

$n Q = \Sigma G_n$

$(n+1) Q' = \Sigma G_{n+1}$

$n Q' + Q' = \Sigma G_n + G_{n+1}$

$n (Q' - Q) \quad nQ + G_{n+1}$

Pseudocode is equivalent as we are calculating mean at every iteration.

$$Q(S_t, A_t) = \frac{G_1 + G_2 - - G_n}{n}$$ (similar to Section 2.4 of the book)

$$Q'(S_t, A_t) = \frac{G_1 + G_2 \_\_ G_{n+1}}{n+1}$$

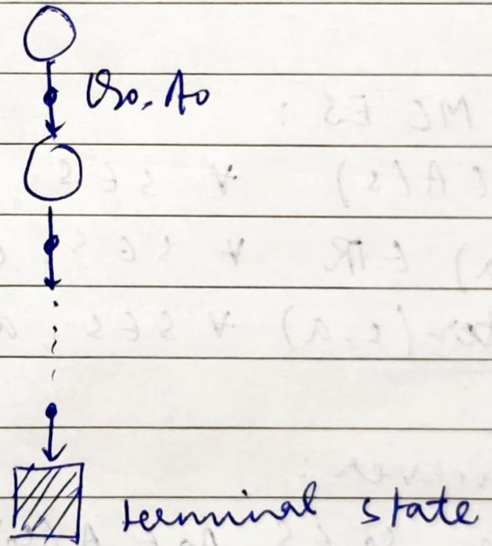
So, $$Q(S_t, A_t) = \frac{n\, Q(S_t, A_t) + G_{n+1})}{n+1}$$

$$= \frac{\text{Counter}(S_t, A_t) \cdot Q(S_t, A_t) + G_{n+1}}{\text{Counter}(S_t, A_t) + 1}$$

Checked as pseudocode.

2. MC est of q_π:

Backup diagram:



π

3– Eq. (5.6) $\Rightarrow$ $V(s) = \dfrac{\sum_t \rho_{t:T(t)-1} G_t}{\sum_t \rho_{t:T(t)-1}}$

In terms of $Q(s,a) = ?$

Consider.

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

$= p(S_t \mid A_t) \cdot P(S_{t+1} \mid S_t, A_t)\, \pi(A_{t+2} \mid S_{t+2}) \ldots P(S_T \mid S_T, A_{T-1})$

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

Relative Ratio :

$\rho_{t:T-1} = \dfrac{\prod \pi(a_i \mid s_i)}{\prod b(a_i \mid s_i)}$

$\boxed{\cancel{\dfrac{\prod \pi(a_i \mid s_i)}{\prod b(a_i \mid s_i)}} \quad \dfrac{\pi}{\pi b}}$

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

5. A scenario in which TD update would be better:

From the hint let us build an example.

Let X be the old parking space, Y be the new parking and H be the highway entry point.

~~[illegible]~~

OLD Scenario : $H \rightarrow \rightarrow \rightarrow \rightarrow X$

NEW " : $H \rightarrow \rightarrow \rightarrow \rightarrow Y$

As we have a lot of experience,

For MC Method : $V(S_t) = V(S_t) + \alpha [G_t - V(S_t)]$

~~[illegible]~~ Here we calculate, G_t by going on the entire episode but if we have a break in b/w we have to calculate G_t all once again.

This will take a long time to converge $v \rightarrow v_\pi$.

On the other hand,

TD Method: $v(s_t) = v(s_t) + \alpha\,[R_{t+1} + \gamma v(s_{t+1}) - v(s_t)]$

Here, we just make prediction of our new parking space, then we can use the previous experience for rest of the spaces.

Hence, $v \to v_\pi$ convergence is faster.

6. (6.3) In the first episode, only the value of V(A) decreased. So, the episode must have ended on the left most state (terminal).

~~TD δ: $\delta_t = R_{t+1} + \gamma v(S_{t+1}) - v(S_t)$ $= 0 + 1$ (~~

V(A) = 0.5 (graph)

$$V(A) = V(A) + \alpha\, [R_{t+1} + V(S_T) - V(A)]$$
$$= 0.5 + 0.1\, [0 + 0 - 0.5]$$
$$= 0.45$$

So, $\Delta V(A) = 0.5 - 0.45 = \underline{0.05}$

All other estimates : $\delta_t = 0 + \gamma(0.5) - 0.5)$
$= 0$

So, $v(k) = v(k) + 0.1\, (0 + \underbrace{v_\alpha(S_t) - v_\alpha(S_{t-1})}_{\text{same}})$
$= v(k) + 0$

$\therefore$ No change.

(6.4) From the graph we infer that on increasing α, we get noisy plots. Noisy plots are also non-converging.

~~At 0.02, 0.03 we see the~~ When $\alpha = 0.05$ in TD, the value is converging enough. In case of MC, $\alpha = 0.04$ is quite noisy.

Hence, I do not think there is a fixed value of α, at which either algo would perform better.

(6.5) RMS error of TD goes down and then up at high α as, $v_\pi(C)$ diverges from its initial estimate.

But, as we go on with the experiment the $v_\pi(C)$ diverges more, in turn, increasing the RMS error.

This ~~does~~ may not always occur and only be a function of how the values were initialized.

-

8. Given : Action selection is greedy.
Compare SARSA and Q-learning :

let us look at Q-learning:
Here we just find $\max\{Q(s', a) \ \forall a \in A(s')\}$
insted of $Q(S', A')$.

SARSA:
Here, we find $Q(S', A')$ by R and updating $Q(S, A)$
by finding A' through S'.

So, as the order of updating Q and finding A' greedly
is different, it is not guranteed the solutions
of the two algos will converge to the same
new action.

Here, the algos are different.