# HW-3

1) Initialize

$\pi(s) \in A(s)$ (arbitrarily) for all $s \in S$

$Q(s,a) \in \mathbb{R}$ (arbitrarily) for all $s \in S, a \in A(s)$

$Count(s,a) \in$ Integers for all $s \in S, a \in A(s)$

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

Loop Forever (for each episode):

Choose $S_0 \in S$, $A_0 \in A(S_0)$ randomly s.t. all pairs have probability $> 0$

Generate an episode from $S_0, A_0$, following $\pi$:

$S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

For $t = T-1, T-2, \ldots, 0$

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

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

$$Q(S_t, A_t) \leftarrow \frac{Q(S_t, A_t) \cdot Count(S_t, A_t) + G}{Count(S_t, A_t) + 1}$$

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

This pseudocode is equivalent as we have assumed that $Q(S_t, A_t)$ already has the average returns for state $S_t$ and $A_t$ for $t$, say $n$ points

$$\Rightarrow Q(S_t, A_t) = \frac{G_1 + G_2 + \cdots + G_n}{n}$$

Now, to update $Q(S_t, A_t)$ for a new point $G_{n+1}$, we get $Q(S_t, A_t)^{\vee}$

$$Q(S_t, A_t)^{\vee} = \frac{G_1 + G_2 + \cdots + G_n + G_{n+1}}{n+1}$$

Here, we maintained a count in the count $(S_t, A_t)$ which was n.

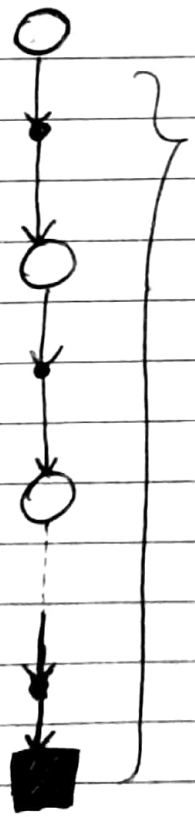
$$\Rightarrow Q(S_t, A_t) = \frac{n \cdot Q(S_t, A_t) + G_{n+1}}{n+1}$$

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

⇒ We have maintained the average by deriving a recursive relation

∴ It is the same as mentioned pseudocode

2)



3) $P[S_t, A_t, S_{t+1}, A_{t+1}, S_{t+2}, A_{t+2}, \dots, S_T \mid S_t = s_t, A_t = a_t, A_{t+1:T-1} \sim \pi]$

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

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

$= \left( \prod_{k=t+1}^{T-1} \pi(a_k \mid s_k) \, p(s_{k+1} \mid s_k, a_k) \right) p(s_{t+1} \mid s_t, a_t)$

$\Rightarrow Reward = \left( \prod_{k=t+1}^{T-1} \pi(a_k \mid s_k) \, p(s_{k+1} \mid s_k, a_k) \right) p(s_{t+1} \mid s_t, a_t)$

$$\left(\prod_{k=t+1}^{T-1} b(a_k|s_k)\, p(s_{k+1}|s_k,a_k)\right) p(\text{s}_{t+1}|s_t,a_t)$$

$$= \frac{\prod_{k=t+1}^{T-1} \pi(a_k|s_k)}{\prod_{k=t+1}^{T-1} b(a_k|s_k)} = \rho_{t+1:T-1}$$

$$\Rightarrow \quad q(s,a) = \sum_{t \in \mathcal{T}(s,a)} \rho_{t+1:T-1}\, G_t \Big/ \sum_{t \in \mathcal{T}(s,a)} \rho_{t+1:T-1}$$

5) In a nutshell, MC estimates need a lot of episodes in order to make accurate estimates about a new state introduced in a system with previous experience on a different set of states while TD uses previous experience to give initial set of estimates with low variance.

MC update: $V_\pi(S_t) \leftarrow V_\pi(S_t) + \alpha[G_t - V_\pi(S_t)]$

$S_{t+1} \to S_t$

For the new states using the new reward and previous state estimates

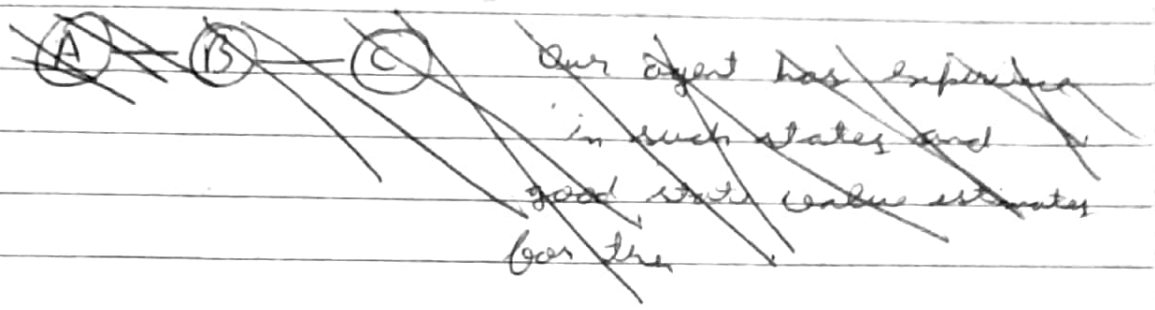
Now, $G_t$ may have high variance

TD Update: $V_\pi(S_t) = V_\pi(S_t) + \alpha[R_{t+1} + \gamma V_\pi(S_{t+1}) - V_\pi(S_t)]$

Here, $V_\pi(S_{t+1})$ has low variance since it has a good estimate of it → our assumption

Another case in which TD helps is when it is hard to reach the terminal state. For such cases MC becomes infeasible but TD can still learn significantly.

~~Example:~~




Our agent has experience in such states and good state value estimates for the

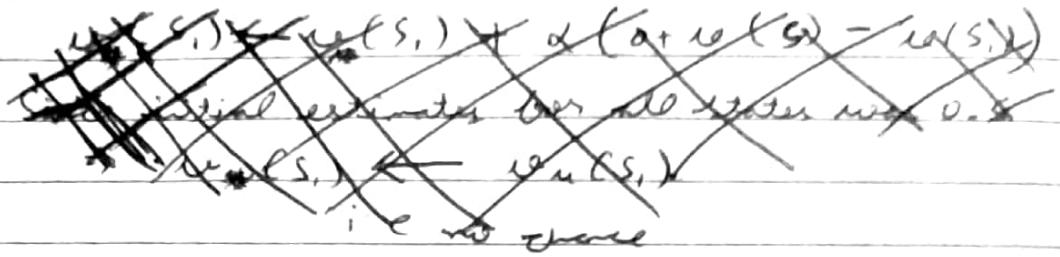
Example:

For the highway instance, these facts hold true as with the introduction of the new building, the initial estimates of MC would have extremely high variance due to lack of experience in that state while TD, on the other hand, would use previous experience and current reward to give a low variance estimate for the new building which would serve as great initial estimates.

87 The would not be the same in some cases. Reason being is the order of updates and action selection done in both the algorithms. In SARSA, we pick our next action $A^{\sim}$ and then update our current state-action pair. On the other hand, in Q-learning, we first make updates to the Q-value of current state action pair and then greedily choose from the updated set of Q-values, say that action is $A^{n'}$. So, it is definitely not guaranteed that $A^{n} = A^{n'}$, hence the algorithms are different.

6) (6.1) We can conclude that the episode ended at the left-most terminal. Say we have the following MRP,

C, 0, $S_1$, 0, $S_2$, 0, ---, $S_n$, 0, A, 0



$v_\pi(C) \leftarrow v_\pi(C) + \alpha(v_\pi(S_1) + 0 - v_\pi(C))$

~~$v_\pi(A) = 0$~~ ∀ state, the initial value is equal

$\therefore v_\pi(C) \leftarrow v_\pi(C)$ i.e no change

For, $S_1$ to $S_{n-1}$

$v_\pi(S_i) \leftarrow v_\pi(S_i) + 0.1(0 + v_\pi(S_{i+1}) - v_\pi(S_i))$

$\Rightarrow v_\pi(S_i) \leftarrow v_\pi(S_i)$ i.e no change

For $S_n$

$v_\pi(S_n) \leftarrow v_\pi(S_n) + 0.1(0 + v_\pi(A) - v_\pi(S_n))$

$\Rightarrow v_\pi(S_n) \leftarrow v_\pi(S_n)$, i.e no change

A → A

$v_\pi(A) \leftarrow v_\pi(A) + 0.1(0 + 0 - v_\pi(A))$

$\Rightarrow v_\pi(A) \leftarrow 0.9\, v_\pi(A)$ i.e change

=> $v_\pi(A)$ value changes by $\frac{1}{10} \times 0.5 = 0.05$

(6.4) The $\alpha$ ranges are sufficient as for both TD and MC, we have obtained results which converge to the best solution possible for the methods. For MC, the curves get noisy with increase in $\alpha$ and hence 0.04 is a sufficient place to stop. For $\alpha = 0.05$ in case of TD, we see that the system converges to a sufficiently low error, over a period of episodes, hence $\alpha$ ranges are sufficient to characterize the analysis and there is no such alpha that can give a significantly different results than these

(6.5) The error increases because the estimate of $v_\pi(C)$ changes from its initial estimate in between, which was the ground truth value of $v_\pi(C)$. So, initially, when there were less episodes, then $v_\pi(C)$ might have been stable near 0.5, thus causing a decrease in error but as episodes moved on, the value of C may have started to change significantly, thus causing an increase in error from that part. This phenomenon happens, clearly due to the initialization of state values, especially $v_\pi(C)$.