Setu Gupta

Reinforcement Learning

HW-3

2018190

Q-1) Initialize:

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

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

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

Loop forever:

Pick $S_0 \in S$ & $A_0 \in A(S_0)$ randomly such that all (s,a) have prob. > 0

Generate episode $S_0, A_0, R_1, S_1, A_1, \cdots, S_{T-1}, A_{T-1}, R_T$ from π

$G \leftarrow 0$

Loop for $t = T-1, T-2, \cdots 0$:

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

if S_t, A_t not in $S_0, A_0, S_1, A_1, \cdots, S_{t-1}, A_{t-1}$:

$Q(S_t, A_t) \leftarrow (RetCount * Q(S_t, A_t) + G) / (RetCount(S_t, A_t) + 1)$

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

$\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$

The update step of Q is

$$Q \leftarrow \frac{nQ + G}{n+1} \quad \text{—— (1)}$$

— (2)

$$n \leftarrow n+1$$

where n is the number of time (S_t, A_t) returns have been encountered.

Imagine a list of returns for S_t, A_t

$[G_1, G_2, \cdots G_n]$

We want to add G_{n+1} to it & word the new average as Q

$$\therefore Q = \frac{(G_1 + G_2 + \cdots G_n) + G_{n+1}}{n+1} \quad \text{— (3)}$$

The old estimate of Q is

$$Q_{old} = \frac{G_1 + G_2 + G_3 + \cdots + G_n}{n}$$

$$\Rightarrow G_1 + G_2 + \cdots + G_n = n\, Q_{old} \quad \text{— (4)}$$

By (3) & (4)

$$Q = \frac{n Q_{old} + G_{n+1}}{n+1}$$

This is same as eq^n (1)

Eq^n 2 updates n for next updates.

(8-2)

8-3)

$$\Pr\{ S_t, A_t, S_{t+1}, A_{t+1}, S_{t+2}, A_{t+2}, \cdots, S_T \mid S_t = s_t, A_t = a_t, A_{t+1,T-1} \sim \bar{\pi} \}$$

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

$$= 1 \cdot p(s_{t+1} \mid s_t, a_t)\, \pi(a_{t+1} \mid s_{t+1}) \cdots p(s_T \mid s_{T-1}, a_{T-1})$$

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

Relative probability

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

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

(Q-5.) Consider the last expirience of travelling from A to E

We have a pretty good estimate of time needed to reach E from state A, B, C, D.

Now suppose we follow another trajectory

a.) Now if we use MC, while updating value for F, we will have to generate the complete episode to get G_t. We use this G_t to update F as

$$V_\pi(F) = V_\pi(F) + \alpha [G_t - V_\pi(F)]$$

where $V_\pi(F)$ is the estimate of state S initialized to random value. and $0 \leq \alpha \leq 1$ is the step of update at every episode.

Now G_t can be high variance & hence it take time to converge to true value of $V_\pi(F)$

b.) If we use TD, we can update as

$$V_\pi(F) = V_\pi(F) + \alpha [R_{t+1} + \gamma V_\pi(B) - V_\pi(F)]$$

This step update is low variance as $V_\pi(B)$ is low variance (due to expirience). Hence we can learn $V_\pi(F)$ faster if we use TD.

Q-6.)

6.3) Since only the estimate of leftmost state A changed, the episode must have ended at left most terminal state

Suppose the episode was

$$B, 0, s_1, 0, s_2, 0, \cdots, s_T, 0, A, 0$$

The TD updates were

$$V_\pi(B) = V_\pi(B) + \alpha(V_\pi(B) + 0 - V_\pi(B))$$

Since initial estimates are $= 0.5$ $\forall$ states

$$V_\pi(B) \leftarrow V_\pi(B) + 0.1\,(0.5 + 0 - 0.5) = V_\pi(B)$$

Illy $\forall \quad 1 \le i \le T-1$

$$V_\pi(s_i) \leftarrow V_\pi(s_i) + 0.1\,(0 + V_\pi(s_{i+1}) - V_\pi(s_i))$$

$$\leftarrow V_\pi(s_i) + 0.1\,(0 + 0.5 - 0.5)$$

$$\leftarrow V_\pi(s_i)$$

for s_T

$$V_\pi(s_T) \leftarrow V_\pi(s_T) + 0.1\,(0 + V_\pi(\pi) - V_\pi(s_T))$$

$$\leftarrow V_\pi(s_T)$$

for A

$$V_\pi(A) \leftarrow V_\pi(A) + (0.1)(0 - V_\pi(A)) \qquad \text{(As next state for A is terminal)}$$

$$\leftarrow V_\pi(A) - \frac{V_\pi(A)}{10}$$

$$\leftarrow \frac{9}{10} V_\pi(A)$$

$$\therefore \text{ Change} = \frac{1}{10} \times V_\pi(A) = 0.05$$

Hence only estimate of A changed with $\delta = 0.05$

6.4) The alpha ranges shown are sufficient.
Increasing α will result in bigger step updates and we will get less smooth curves.

As we increase α, the MC curves get more and more noisy. $\alpha = 0.04$ is sufficiently noisy to claim that increasing α beyond 0.04 is not usefull.

Similar argument holds for TD as well.

∴ We can see from the given curves that TD learns faster.

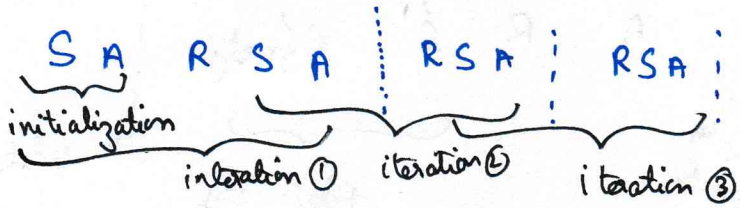
6.5) The reason why TD error increases for more episodes is because the estimate of $v_\pi(C)$ is getting contaminated.
We initialize $v_\pi(C)$ to its true value.
As we learn, we start to estimate value of all states and they propogate from ends to middle.
Since these estimates can be noisy, after some episodes, those noisy estimates propogate till C and start changing $v_\pi(C)$. This increases error.

This is more pronounced for higher α as higher α causes more change in every step update.

This is also an artifact of initialization of $v_\pi(C)$ to truth.

Q-7.) SARSA

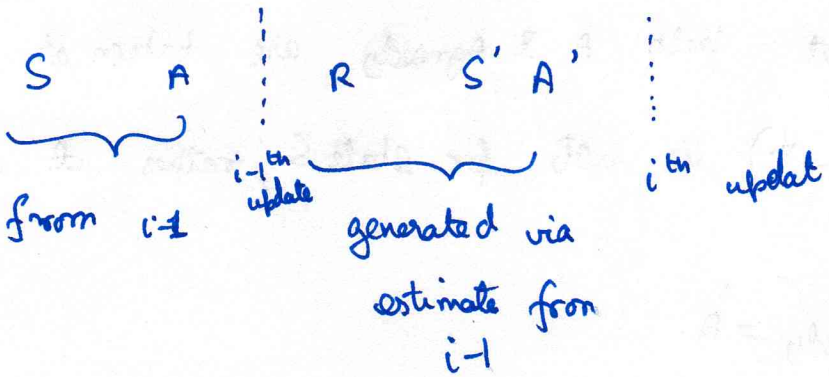
The seq. of state, action, rewards for SARSA is as follows



where ⋮ represents Q updates

S and A of previous iteration are use in current iteration

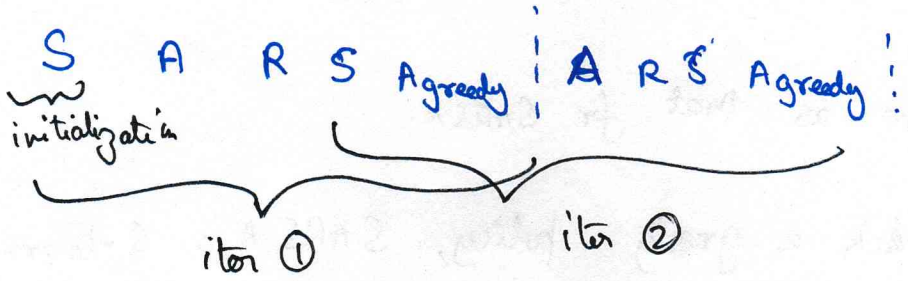
For an iteration i :



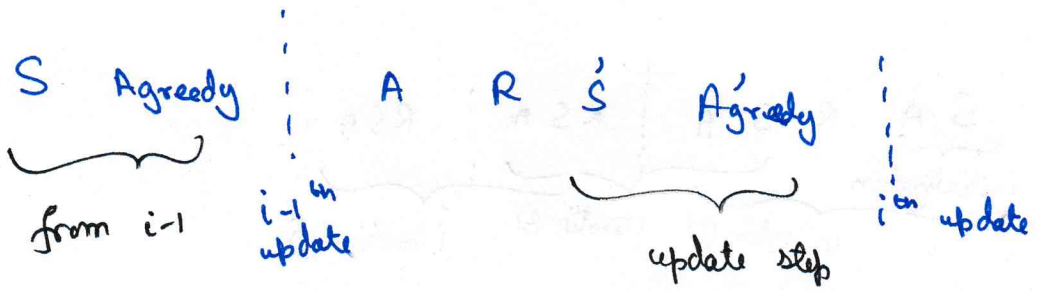
Suppose we pick A' greedily based on i-1 th estimates

Q-learning

Illy two seq. for Q learning is



For i^{th} iteration



If $A_{greedy} = A$, then SARSA = Q learning

Note that both A & A_{greedy} are taken at state S and the update ($i-1^{th}$) is not for state S, rather it is for previous state.

$\therefore A_{greedy} = A$.

$\therefore$ We can simplify the seq for Q-learning to

S A : R S A : R S A :.

which is same as that for SARSA.

$\therefore$ If we pick a greedy policy, SARSA = Q-learning.

They will have same action selection & updates.