

Q-1) Initialize:

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

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

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

Loop forever:

Pick $s_0 \in S \geq A_0 \in A(s_0)$ randomly such that all (s, a) have prob. > 0 Generate episode $s_0, A_0, R_1, s_1, A_1, \dots, s_{T-1}, A_{T-1}, R_T$ from π

$$G \leftarrow 0$$

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

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

if s_t, A_t not in $s_0, A_0, s_1, A_1, \dots, s_{t-1}, A_{t-1}$:

$$Q(s_t, A_t) \leftarrow \frac{(\text{RetCount} * Q(s_t, A_t) + G)}{(\text{RetCount}(s_t, A_t) + 1)}$$

$$\text{RetCount}(s_t, A_t) \leftarrow \text{RetCount}(s_t, A_t) + 1$$

$$\pi(s_t) \leftarrow \arg\max_a Q(s_t, a)$$

The update step of Q is

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

$$n \leftarrow n+1 \quad \text{--- (2)}$$

where n is the number of time (s_t, A_t) actions have been encountered.Imagine a list of returns for s_t, A_t

$$[G_1, G_2, \dots, G_n]$$

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

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

The old estimate of Q is

$$Q_{old} = \underbrace{G_1 + G_2 + G_3 + \dots + G_n}_n$$

$$\Rightarrow G_1 + G_2 + \dots + G_n = n Q_{old} \quad \text{--- (4)}$$

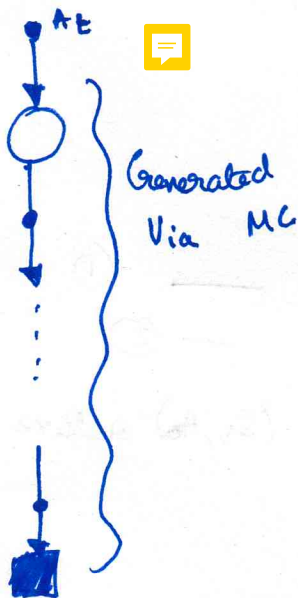
By (3) & (4)

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

This is same as eqⁿ (1)

Q^n is updated n for next updates.

Q-2)



Q-3)

$$\begin{aligned}
 & P_r \{ S_t, A_t, S_{t+1}, A_{t+1}, S_{t+2}, A_{t+2}, \dots, S_T \mid S_t = s_t, A_t = a_t, A_{t+1}, \dots, A_{T-1} \} \\
 &= P(S_t | A_t) P(S_{t+1} | S_t, A_t) \pi(A_{t+2} | S_{t+2}) \dots P(S_T | S_{T-1}, A_{T-1}) \\
 &= 1 \cdot p(s_{t+1} | s_t, a_t) \pi(a_{t+1} | s_{t+1}) \dots p(s_T | s_{T-1}, a_{T-1}) \\
 &= \left(\prod_{k=t+1}^{T-1} \pi(a_k | s_k) p(s_{k+1} | s_k, a_k) \right) p(s_T | s_{T-1}, a_{T-1})
 \end{aligned}$$

Relative probability

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

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



Q-5) Consider the past experience 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 F 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 takes 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 experience). 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

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

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

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

|||

$$\forall \quad 1 \leq i \leq T-1$$

$$\begin{aligned} V_{\pi}(s_i) &\leftarrow V_{\pi}(s_i) + 0.1 (0 + V_{\pi}(s_{i+1}) - V_{\pi}(s_i)) \\ &\leftarrow V_{\pi}(s_i) + 0.1 (0 + 0.5 - 0.5) \\ &\leftarrow V_{\pi}(s_i) \end{aligned}$$

for s_T

$$\begin{aligned} V_{\pi}(s_T) &\leftarrow V_{\pi}(s_T) + 0.1 (0 + V_{\pi}(A) - V_{\pi}(s_T)) \\ &\leftarrow V_{\pi}(s_T) \end{aligned}$$

for A

$$\begin{aligned} V_{\pi}(A) &\leftarrow V_{\pi}(A) + (0.1)(0 - V_{\pi}(A)) \\ &\leftarrow V_{\pi}(A) - \frac{V_{\pi}(A)}{10} \\ &\leftarrow \frac{9}{10} V_{\pi}(A) \end{aligned}$$

(As next state for A is terminal)

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

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

6.4.) The alpha ranges shown are sufficient.
Increasing α will result in larger 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 useful.

Similar argument holds for TD as well.

\therefore 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 propagate from ends to middle.

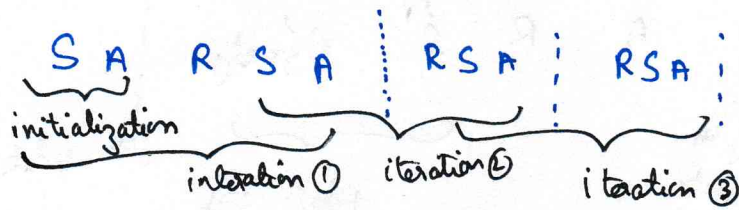
Since these estimates can be noisy, after some episodes, these noisy estimates propagate till c and start changing $v_{\pi}(c)$. This increases error.

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

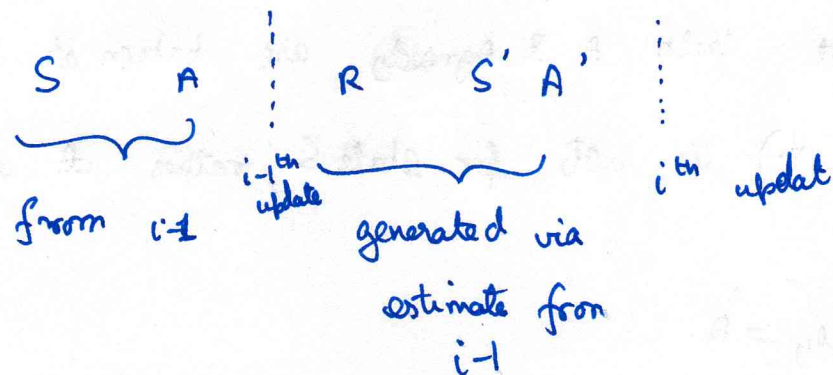
The seq. of state, action, rewards for SARSA is as follows



where \vdots represents Q updates

S and A of previous iteration are used in current iteration

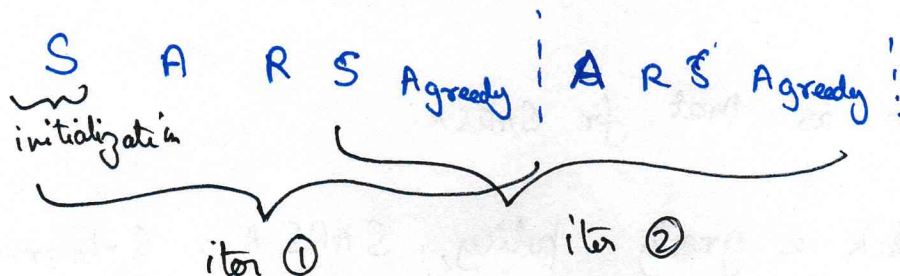
For an iteration i :



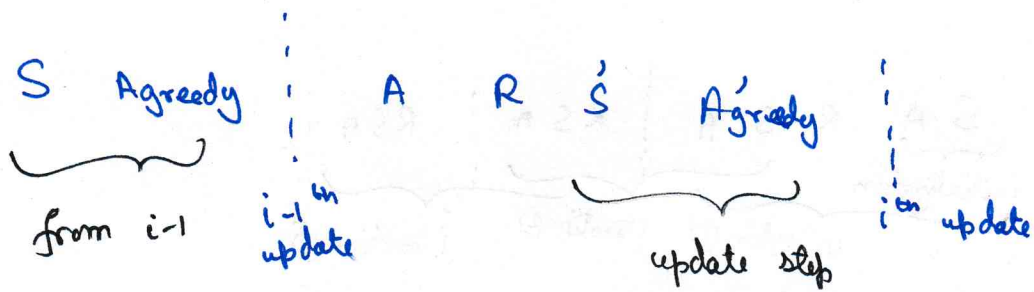
Suppose we pick A' greedily based on $i-1^{\text{th}}$ estimates

Q-learning

Its seq. for Q learning is



For i^{th} iteration

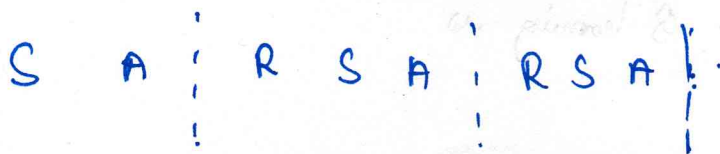


If $A_{\text{greedy}} = A$, then $\text{SARSA} = Q\text{-learning}$

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

$\therefore A_{\text{greedy}} = A$.

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



which is same as that for SARSA.

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

They will have same action selection & updates.