Assignment 3

Anushka Bhandari
2016134

Q1 Initialise

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

$Q(s,a) \in R$ $\forall$ $a \in A$ and $s \in S$

count $(s,a) = 0$ $\forall$ $a \in A$ and $s \in S$

~~Loop for ever that is for~~ .
~~for each episode~~

for each episode : loop for ever

Choose $s_0 \in S$, $A_0 \in A(s_0)$ randomly such that all pairs have prob > 0

Generate episode from S_0 A_0 following π

$\pi : S_0\ A_0\ R_1 \ldots S_T\ A_T\ R_{T+1}$

$G \leftarrow 0$

loop for each step $t = T-1, T-2 \ldots 0$:

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

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

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

Count $(S_t, A_t) += 1$

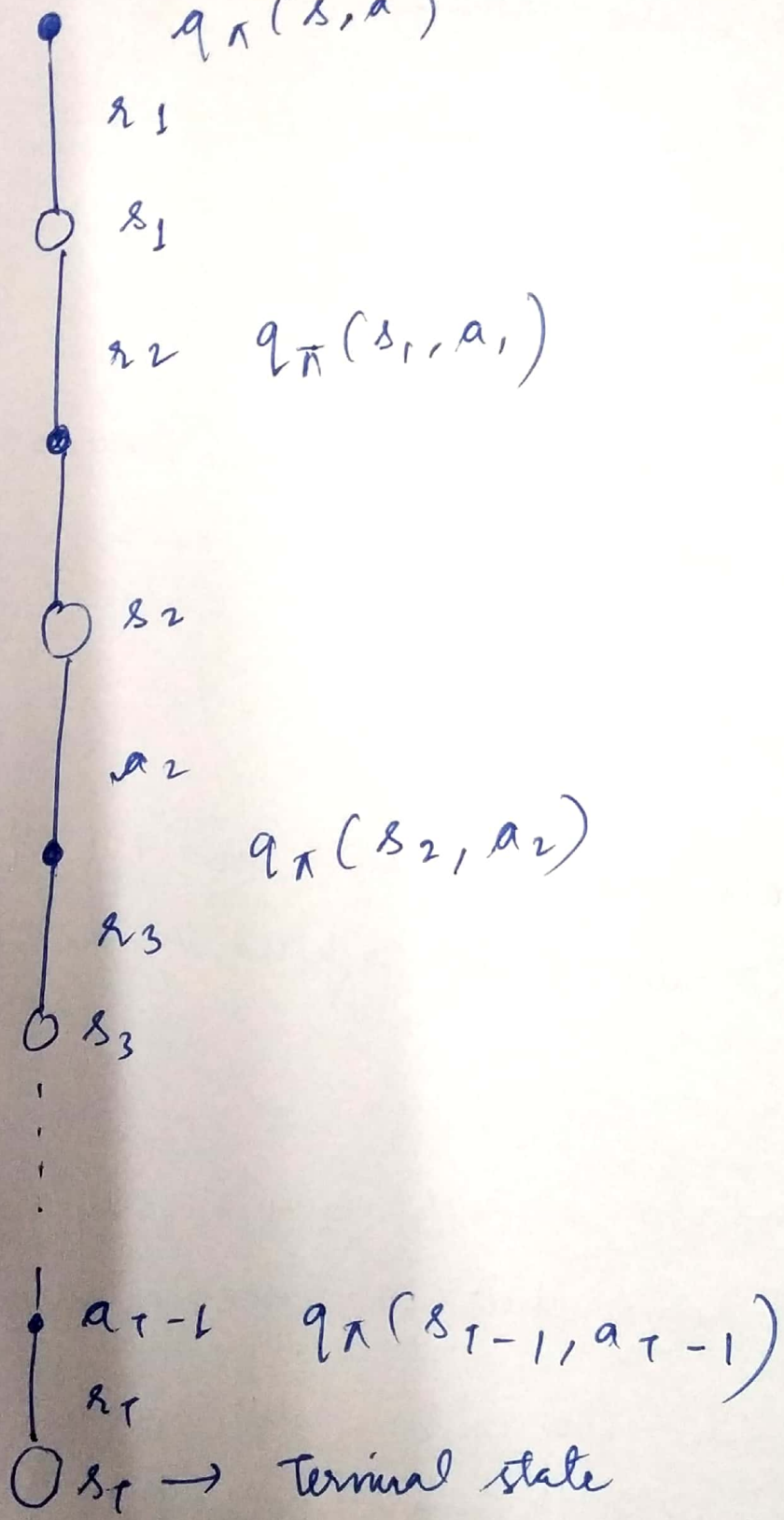
$\pi(S_t) \leftarrow \underset{a}{\text{argmax}}\, Q(S_t, A_t)$

Q2

Back up diagram for MC estimation of q_π

• ⟶ representation action

o ⟶ representation state



Q3

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

$\Upsilon(s,a)$ are set of all ~~timestamp~~ the time steps s and a are visited.
This is for every visit method.
so for
$Q(s,a)$, $T(s)$ is replaced

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

Q5

(1) TD learns directly from episodes of experience
(2) TD is model free. No knowledge of MDP transitions.

Lets take a driver has a lot of experience driving from office to home.

OFFICE ⟶ lane 1 ⟶ lane 2 ⟶ HOME

But on the next day suppose lane 2 is closed due to construction, the route then changes to

OFFICE ⟶ lane 1 ⟶ lane 3 ⟶ lane 4 ⟶ HOME

Had we used MC to update we would have to wait till we reached home.
TD updates at every step.
We can use the learnt values from office and lane 1 from the start

Q6.3

Only the V(A) was changed
This is because the updates that follow the below equation.

$$V(s) \leftarrow V(s) + \alpha[R + \gamma V(s')] - V(s)$$

$\because V(s') = 0$ as it is the terminal state

$R = 0$ and $V(s) = 0.5$ and $\alpha = 0.1$

$$V(A) = V(A) - \alpha V(A)$$
$$= 0.5 - 0.1(0.5) = 0.45$$

Only V(A) was updated because the trajectory ended at A and other V(s) balanced out.

Q6.4

Since TD updates using rewards and MC updates using returns, TD will be more affected by α.

Q6.5

Since the V(s) for each S is steadily updated to optimal, hence there are fluctuations in TD RMS.
After it reaches optimal the updating still continues.
Hence it causes V(s) to move away from $V^*(s)$.
The fluctuation is greater for larger α because the step size derives the value of updation.

Q 8

There is only one term difference between SARSA and Q learning i.e. $Q(S', A')$ and $\max_a Q(S', a)$

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t))$$

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma Q(S', A') - Q(S_t, A_t))$$

In case every selection is greedy with every other condition being the same.

then $Q(S' A') = \max_a Q(S', a)$

This will make the same action selection and hence the same weight update.