

1.

$$Q(s, a)_n = \frac{1}{n} \sum_{i=1}^n G_{i, a, s}$$

$$Q(s, a)_{n+1} = \frac{1}{n+1} \sum_{i=1}^{n+1} G_{i, a, s}$$

$$= \frac{1}{n+1} \left[ \sum_{i=1}^n G_{i, a, s} + G_{n+1, a, s} \right]$$

$$= \frac{1}{n+1} \left[ \cancel{n} \times \frac{\sum_{i=1}^n G_{i, a, s}}{n} + G_{n+1, a, s} \right]$$

$$= \frac{1}{n+1} \left[ n Q(s, a)_n + G_{n+1, a, s} \right]$$

$$= \frac{1}{n+1} \left[ (n+1) Q(s, a)_n - Q(s, a)_n + G_{n+1, a, s} \right]$$

$$= Q(s, a)_n + \frac{1}{n+1} \left[ G_{n+1, a, s} - Q(s, a)_n \right]$$

# Pseudocode

Initialize:

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

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

Return  $(s, a) \leftarrow$  empty list, for all  $s \in S, a \in A(s)$

~~Count Times~~  $N(s, a) \leftarrow 0$  for all  $s \in S, a \in A(s)$

Loop forever (for each episode)

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

~~Generate~~

Generate an episode from  $s_0, A_0$ , following  $\pi$ :  $s_0, A_0, R_1, \dots, s_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

Loop for each step of episode,  $t = T-1, T-2, \dots, 0$ :

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

Unless the pair  $s_t, A_t$  appears in  $s_0, A_0, s_1, A_1, \dots, s_{t-1}, A_{t-1}$

Append  $G$  to  $\text{Return}(s_t, A_t)$   $N(s, a) \leftarrow N(s, a) + 1$

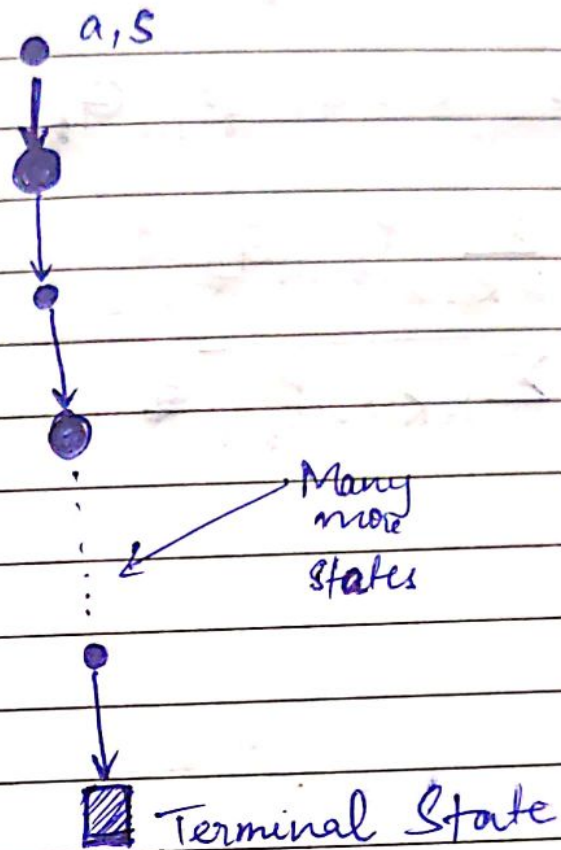
$Q(s_t, A_t) \leftarrow Q(s_t, A_t) + \frac{1}{n(s, a)} (G - Q(s_t, A_t))$

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

\*  
ranged



## 2: Backup Diagram for Monte Carlo estimation of $q_{\pi}$



$$3. P_{t:T-1} = \prod_{k=t}^{T-1} \frac{\pi(A_k | S_k)}{b(A_k | S_k)}$$

$$q_b(s, a) = E[G_t | S_t = s, A_t = a]$$

$$q_{\pi}(s, a) = E[P_{t:T-1} G_t | S_t = s, A_t = a]$$

$k(s, a) \rightarrow$  set of all time steps with current state  $s$ , followed by action  $a$  on it. for an every visit monte carlo

$G_t$  is the return upto the time of episode termination.

~~$$Q(s, a) = \sum_{t \in k(s, a)} P_{t:T-1} G_t$$~~

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

$$\sum_{t \in k(s, a)} P_{t:T-1}$$



## 5. Exercise 6.2

TD updates are better since, I will be able to use <sup>not</sup> prior knowledge of the highway and these will be incorporated with MC, ~~as~~ as it will note my observations after the completion of the entire journey

~~6. The road~~

## 6. 6.3

The walk must have ended in the terminal state with reward 0.

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

for all states with  $s'$  other than terminal states  
 $V(s')$  must be 0.5 for episode 1. & Reward was 0  
So

$$V(s) = 0.5 + \alpha [0 + 0.5 - 0.5] \\ = 0.5$$

for state A

$$V(s_A) = 0.5 + \alpha [0 + 0 - 0.5] < 0.5 \\ = 0.5 + 0.1(-0.5) \\ = 0.45$$

## 6.4

~~As seen before (prev. chapter)~~

$\alpha = \frac{1}{N(s)}$  can perform better where  $\alpha$  will

decrease with time.

Also the stationary nature of the problem is suited <sup>has</sup> the choice of  $\alpha$ .

6.5

Higher values of  $\alpha$ , mean a higher jump in the values of the state ~~value~~ at any timestep, which causes an increase in RMSE.

8 Yes, ~~both~~ the algorithms, make the same update in weights. If both follow a greedy policy, in a general scenario, both algorithms are same.