1.

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

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

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

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

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

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

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

Pseudocode

Initialize:

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

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

~~Returns$(s,a) \leftarrow$ 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 $\pi: S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

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

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

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

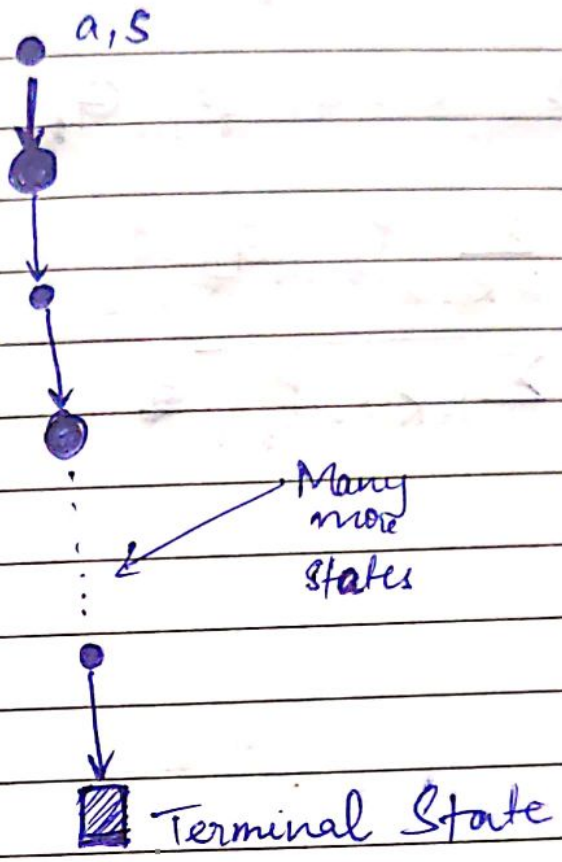
~~Append G to Returns(S_t, A_t)~~ $N(s,a) \leftarrow N(s,a) + 1$

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

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

* changed

2. Backup Diagram for Monte Carlo estimation of q_π



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

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

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

$k(s,a)$ → 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 uptil the time of episode termination.

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

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

5. Exercise 6.2

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

~~6 The end~~

6 6.3

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

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

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

So

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

for state A

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

6.4

~~Mc learn befor (prev. chapter.~~

$\alpha = \frac{1}{N(s)}$ can perform better where α will decrease with time.

Also the stationary nature of the problem is suited by the choice of α.

6.5

Higher values of α, mean a higher jump in the values of the state at any timestep which causes an increase in RMS.

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.