Sanchit Malhotra
2016264
RL - Assignment 3.

Ex 5.4) Given numbers $a_1, a_2 \cdots, a_n$

we the mean $\theta_n = \frac{a_1 + a_2 + \cdots + a_n}{n}$

$$= \frac{n-1}{n}\left(\frac{a_1 + a_2 \cdots + a_{n-1}}{n-1}\right) + \frac{a_n}{n}$$

$$\theta_n = \theta_{n-1} + \frac{1}{n}\left(a_n - \theta_{n-1}\right).$$

So if we know the count & previous mean we ~~[illegible]~~ can simply use the above formula.

Initialize:

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

$Q(s,a) \in \mathbb{R}$ (" "), $\forall$ $s \in S, a \in A(s)$

$Returns(s,a) \leftarrow$ empty list, $\forall$ $s \in S$, $a \in A(s)$

Loop forever (for each episode):

Choose $S_0 \in S$, $A_0 \in A(S_0)$ randomly s.t all pairs have probability > 0.

Generate an episode from S_0, A_0, following π: $S_0, A_0, R_1, \cdots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

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

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

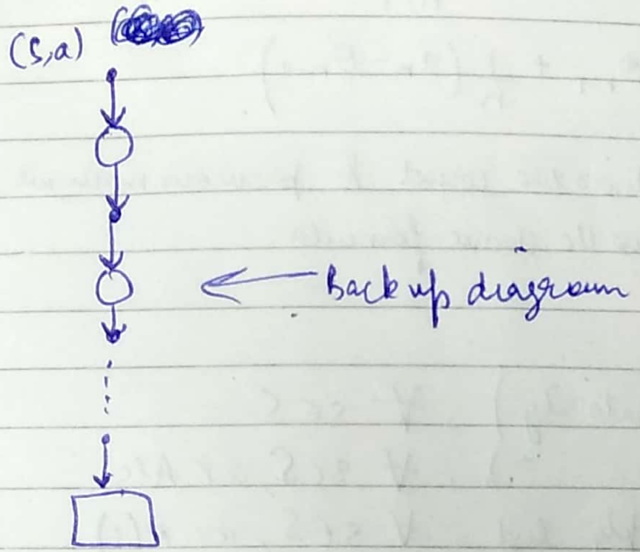
Unless pair S_t, A_t appears in $S_0, A_0, S_1, A_1, \cdots, S_{t-1}, A_{t-1}$:

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

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{1}{N(A_t, S_t)}\left[G - Q(A_t, S_t)\right]$$

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

Ex 5.3) In monte Carlo Exploring Starts,
All state action pair have non-zero probability of being chosen. As this ensures that all state-action pair will be visited an infinite number of times in the limit of an infinite number of episodes.



Ex 5.6) We know $V(s) = \dfrac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1} G_t}{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1}}$

Then

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

$Q(s,a) \rightarrow$ State-action Value function

$\mathcal{T}(s,a) \rightarrow$ set of all time steps when we are at state 's' & action 'a' is taken

~~Sa-Ch~~

Ex-6.12)

No,

In Q-learning, ~~the~~ Q(S,a) is updated using greedy approach
whereas in Sarsa, Q(S,a) is " " " ϵ-greedy approach.

There is non-zero probability of not choosing the current max.

Ex 6.2).