Homework 3

Ishaan Bassi, 2016238

Q1: Storing the past returns is unnecessary.

The state-action value for pair (s, a) can be done as -

$$Q(s,a) \leftarrow Q(s,a) + \frac{1}{\text{count}(s,a)} [\text{Return} - Q(s,a)]$$

Hence the pseudocode will be -

Initialize:

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

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

$C(s,a) \in \mathbb{Z}$, count of pair (s,a) $\forall s \in S, a \in A(s)$

Loop over each episode:

choose S_0, A_0 randomly with equal probability

Generate episode from S_0, A_0 following π

$G \leftarrow 0$

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

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

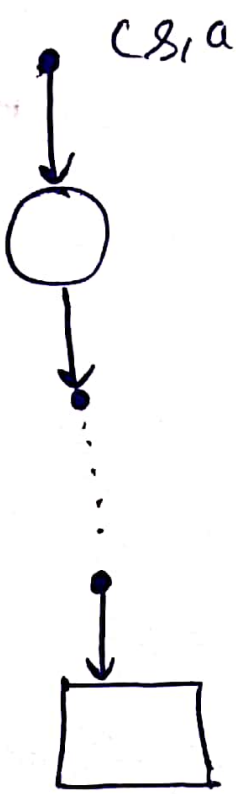
If pair S_t, A_t is not in $S_0, A_0 \cdots S_{t-1}, A_{t-1}$:

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

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

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

Q2. With monte carlo exploring starts, we start by choosing a state action pair randomly. Hence our root node will be this pair.



Q3. The formula for state - action pair (S, a) is given by -

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

where $Q(s,a)$ is the value of state-action pair (s,a) and $T(s,a)$ is the set of all time steps when state is 's' and action a is taken.

Q5. TD updates would work better because we have prior knowledge of the return (our learnt estimate) for the state of exiting the highway. We can update the state values for other ~~states~~ or previous state in the episode ~~for~~ before it ends, i.e we don't have to wait for the agent to reach his home. Yes, the same thing would happen if the state value is close to true value.

Q8. In Q-learning, the value of $Q(s,a)$ is updated using the greedy approach i.e find the best action (a') for which target $R + Q(s', a')$ is maximum (argmax), whereas in case of sarsa the update is done using ~~the~~ a' found in ϵ-greedy way. Hence even if action a is chosen using greedy policy the update method differs from sarsa and hence the sequence of actions and updates.