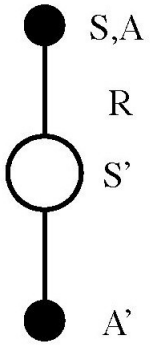
**Assignment 2**
**Advances in Robotics**
Harsh Shukla
2019701012

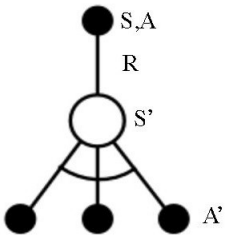
## Backup Diagrams

Backup diagrams for SARSA:




$$Q(S,A) \leftarrow Q(S,A) + \alpha\left(R + \gamma Q(S',A') - Q(S,A)\right)$$

Backup diagrams for Q-Learning:




$$Q(S,A) \leftarrow Q(S,A) + \alpha\left(R + \gamma \max_{a'} Q(S',a') - Q(S,A)\right)$$

## Differences between MC and TD methods:

1. MC must wait until the end of the episode before the return is known while TD can learn online after every step and does not need to wait until the end of episode.
2. MC has high variance and low bias while TD has low variance and some decent bias.
3. MC does not exploit the Markov property while TD exploits the Markov property.

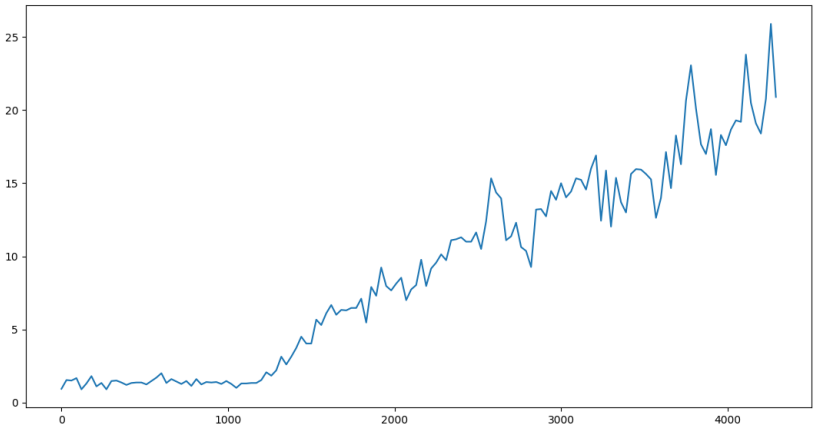



Figure 1: Performance of DQN

Figure 2: Caption

## Greedy Action Selection

No,they can't be same. SARSA is an on policy learning algorithm while Q-learning is off policy learning algorithm. If action selection is greedy, there will be no exploration so there is no guarantee of convergence to optimal policy. Weights updates and action selection will be same if behavioral policy is equal to the greedy policy.

## DQN

**1. Click here** for the video of the final performance of the network. Figure 1 shows the performance of DQN. x and y axis indicate number of time-steps and mean reward for past 30 episodes respectively.

**2.** Screenshots:

- Figure-2 Q values [2.5831807 2.5948505 2.6011992 **2.8010435**] The bold value is the highest value which represents 'right' action
- Figure-3 Q-values [3.490177 **3.540487** 3.53936 3.4871817] The bold value is the highest value which represents 'fire' action.
- Figure-4 Q-values [2.4859753 2.4880767 **2.5151093** 2.4915476] The bold value is the highest value which represents 'left' action

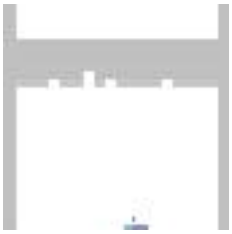


Figure 3: Caption

Figure 4: Caption

**3.** DQN performance was studied with variation of exploration policy parameter(epsilon). Figure 5 shows the rewards with epsilon 0.8, Figure 6 with epsilon 0.6 and Figure 1 with epsilon value equal to 1. Let, DQN with epsilon 0.6 be DQN-A, DQN with epsilon 0.8 be DQN-B.
It was observed that with same number of episodes(3000) trained, DQN-A converges to a higher average reward than epsilon DQN-B. This is partly because the number of decay steps were same in both the cases. In experiment it was obesrved, after 3000 episodes, value of the hyperparameter reduced to 0.1 in DQN-A and to around 0.5 in DQN-B i.e DQN-B was still in it's exploratory phase when the training ended.
I think, since DQN-B has higher value of exploration policy parameter(epsilon), it will visit more states and will give us better estimate of Q-values but will require a larger amount of training episodes. So, if trained for sufficiently longer time, DQN-B will finally converge to a higher reward value than DQN-A in the end.

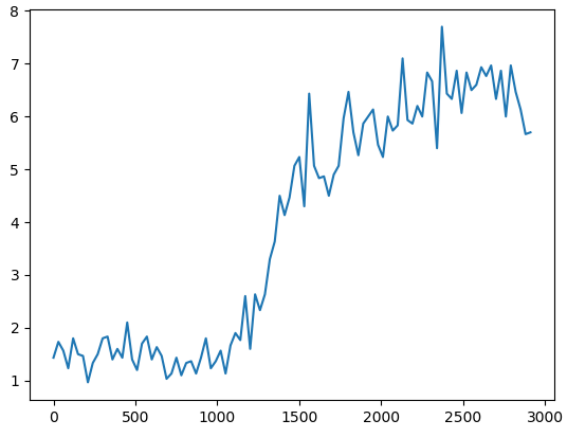



Figure 5: Performance of DQN with epsilon 0.8

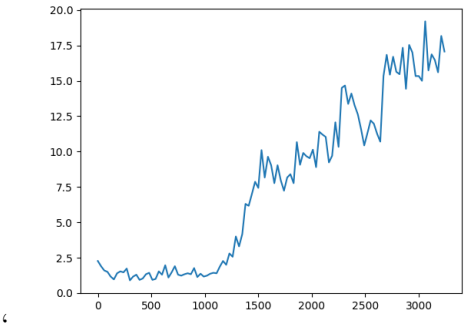



Figure 6: Performance of DQN with epsilon 0.6