Assignment-8

2023-02-27

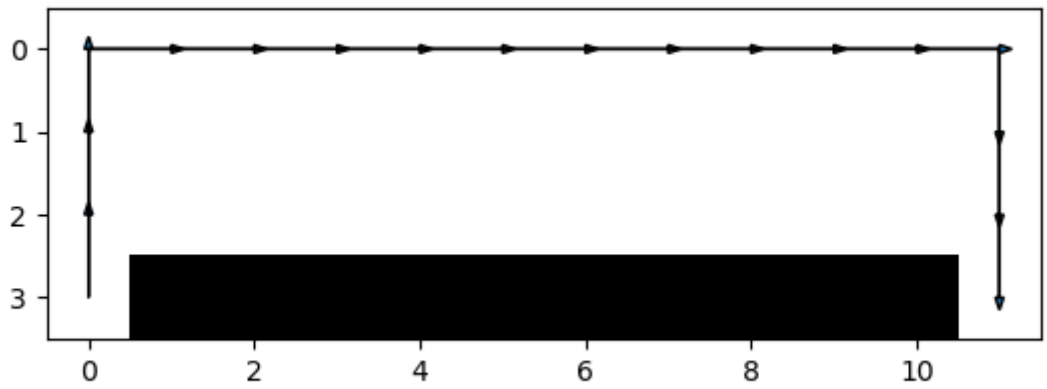
Name: Varun Gampa - 773686296, Rohin Siddhartha - 808068806

ID: vgampa, rvenkateswaran

Q1.

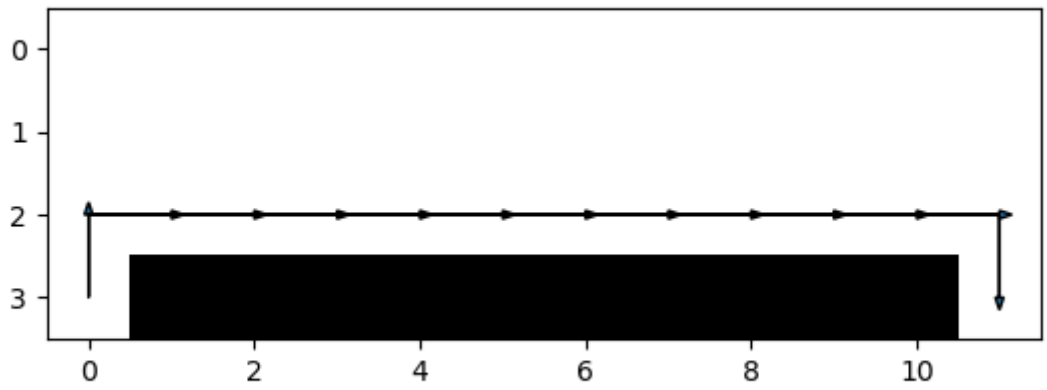
The hyper-parameters chosen in this exercise are:
num_episodes = 5000
alpha = 0.05
e = 0.2
gamma = 0.90

Sarsa-method- policy:



Optimal Path

Q-learning-Policy:



Optimal Path

SARSA is an ON policy learning method. In this case epsilon (the probability of exploration) is a constant and will be there in the final policy as well. As a result as per this policy, should the path chosen be close to the cliff, there will always be an epsilon/4 probability that the agent falls of the cliff and reaches the initial position.

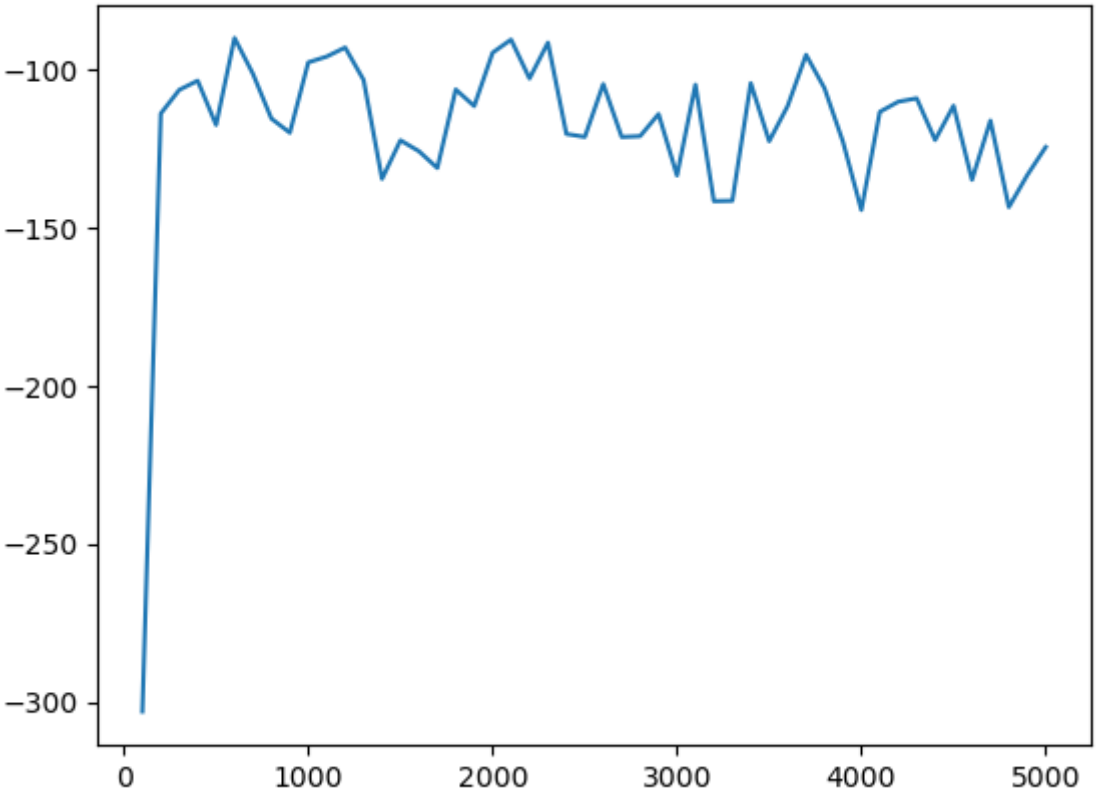
Q-learning is an OFF policy learning method as a result, it policy at any state is deterministic and will not choose an action which can make it fall of the cliff.

Hence SARSA converges to the top (blue in question) path and Q-learning converges to the bottom path (red in question).

Q2.

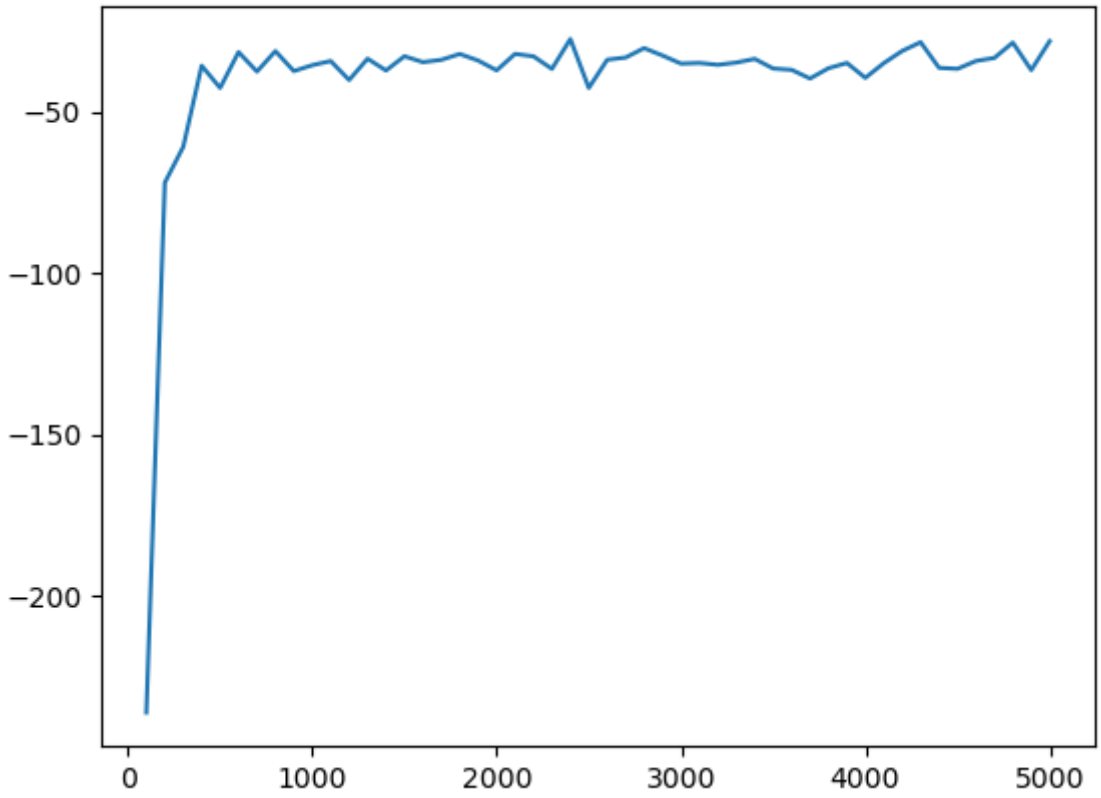
The plot of sum of rewards versus number of episodes is to be shown.
Since the graph is too noisy, the sum of rewards was averaged over hundred episodes.

Q-learning plot of rewards obtained during trajectory to goal:



Number of rewards(averaged over hundred episodes) vs number of episodes

SARSA plot of rewards obtained during trajectory to goal:



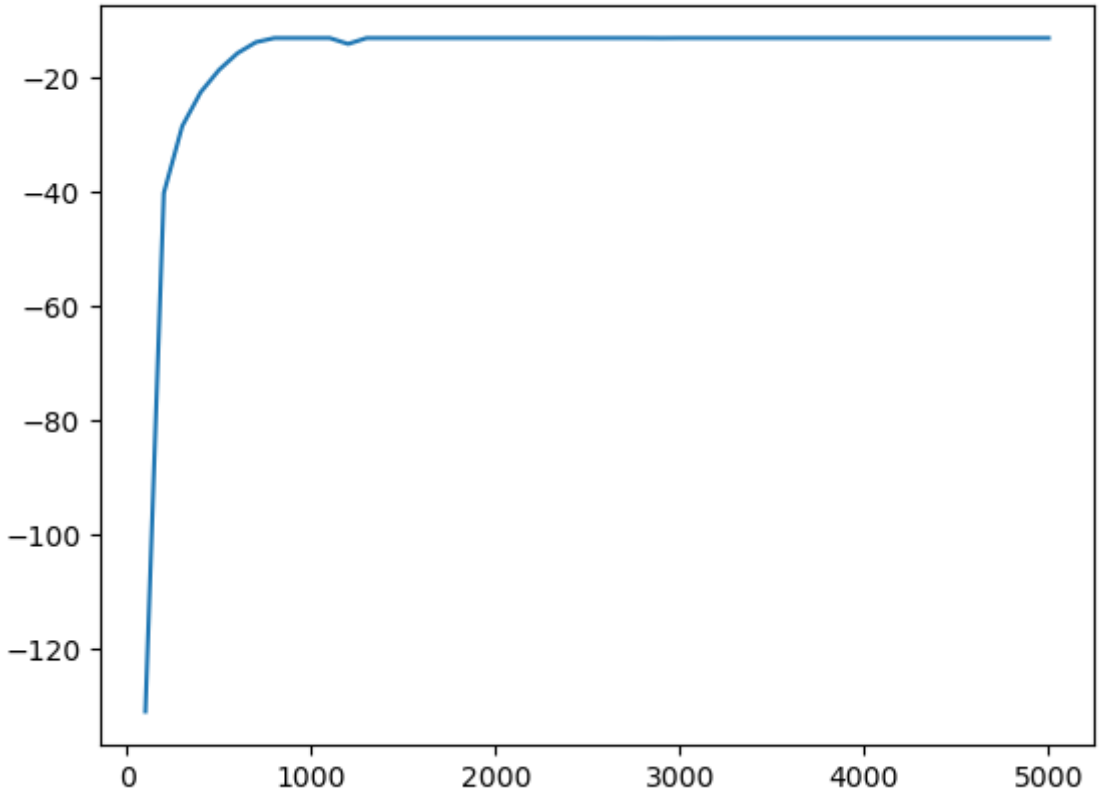
Number of rewards(averaged over hundred episodes) vs number of episodes

In both the methods the trajectory is generated by the epsilon-greedy algorithm. In the SARSA, since it is an ON policy method it's state-action value matrix suggests a trajectory far away from the cliff, so even if with epsilon/4 probability the agent moves towards the cliff it will not immediately fall off. Whereas in Q-learning the state-action value matrix suggests a trajectory close to the edge, so with epsilon/4 probability the agent falls incurring an expensive reward of -100. As a result the reward obtained by the agent in Q - learning training even after converging on the state-action value matrix, is more negative.

Q3.

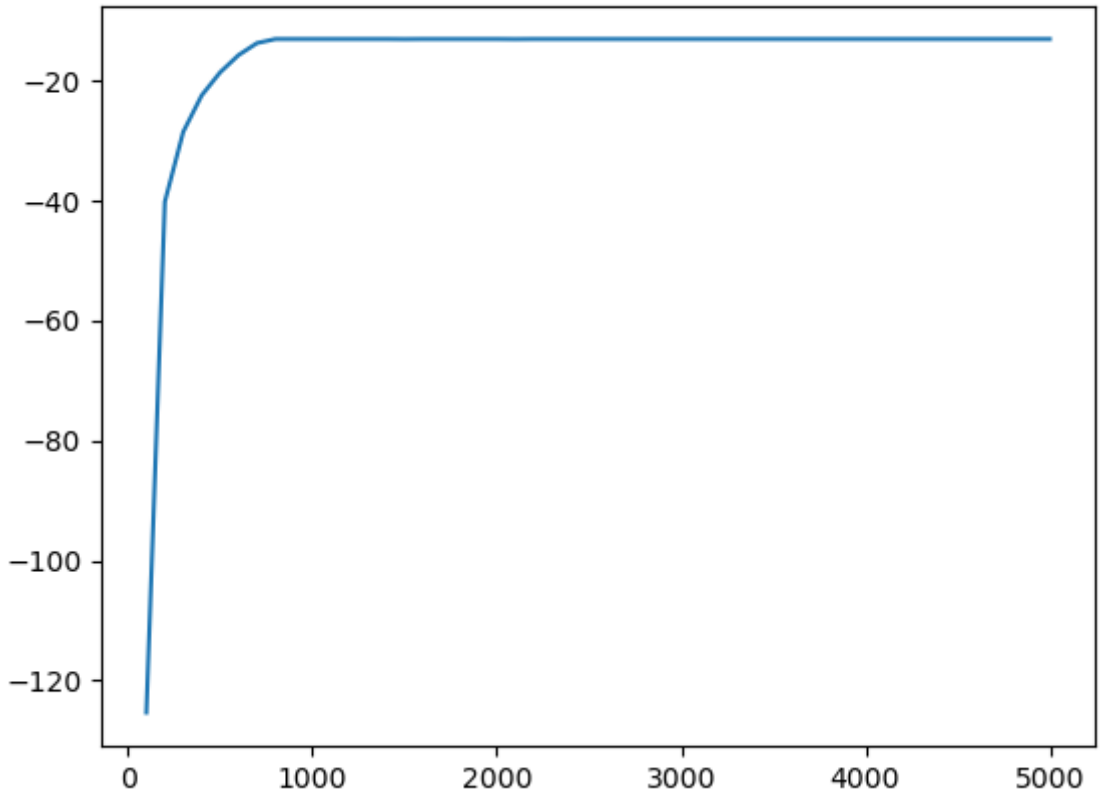
The decay method is chosen as epsilon = $epsilon\_start/episode^{decay\_rate}$
Decay rate = 1.25

Q learning reward in each episode graph with decreasing epsilon:



Number of rewards(averaged over hundred episodes) vs number of episodes

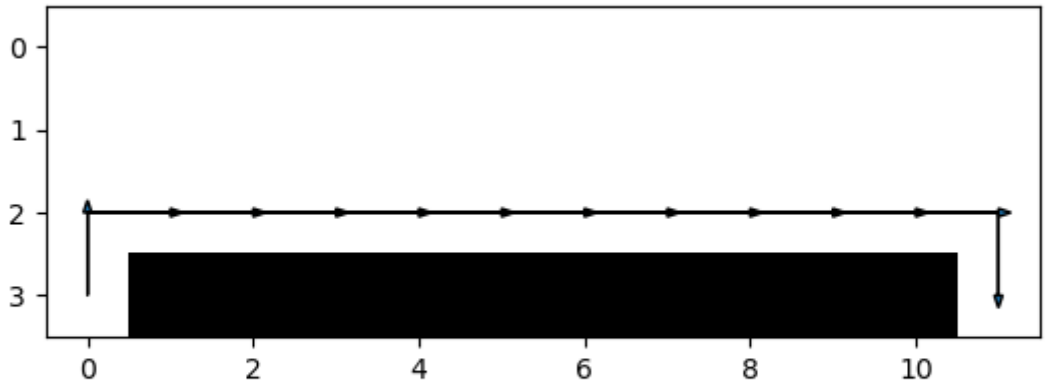
SARSA reward in each episode graph with decreasing epsilon:



Number of rewards(averaged over hundred episodes) vs number of episodes

It can be seen that both policy converge to same rewards, indeed the optimal path generated by both is same as well. This is because the exploring policy is converging to the optimal path.

Q learning optimal path:



SARSA optimal path: