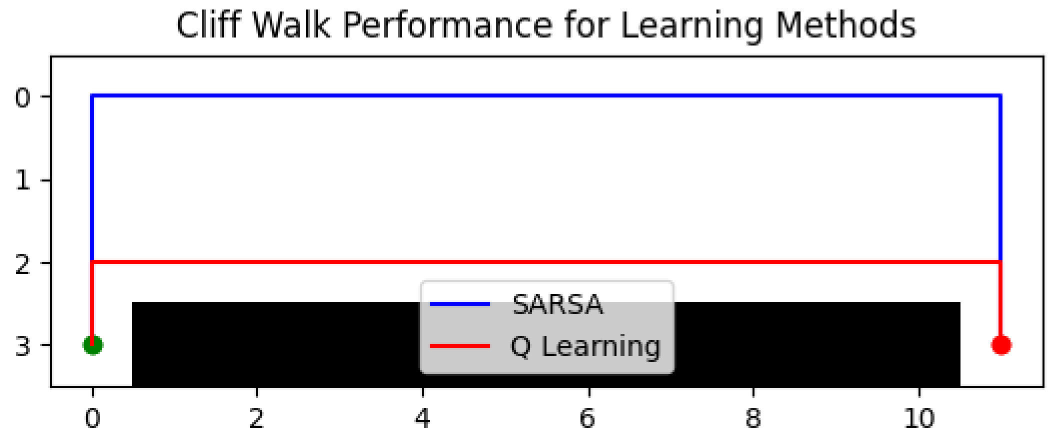
Shane Toma

RBE595

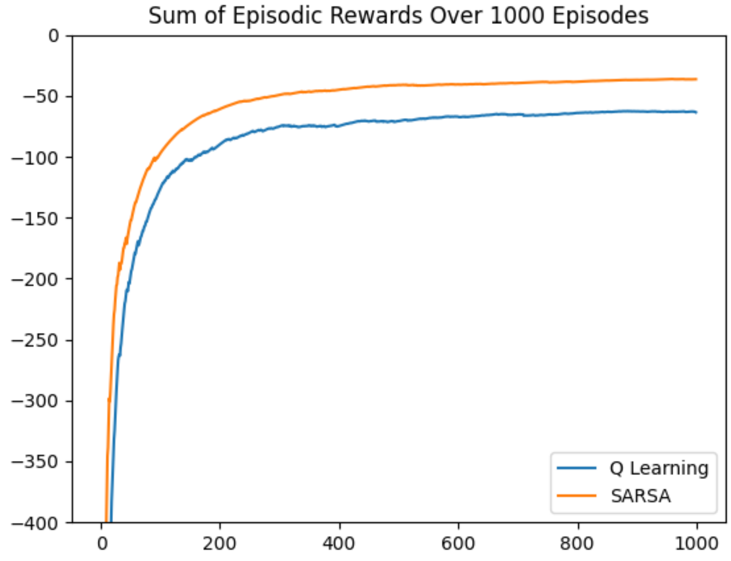
2/25/2024

Programming Assignment 4

1. As shown in the figure below, the SARSA model converges to the safer path away from the cliff, while the Q-Learning method converges to the optimal path to the goal, even though it is right next to the cliff. This is due to the Q-learning method always approximates the optimal action value function, which will lead it to the optimal path, but due to the ϵ greedy action selection, this means it will occasionally choose to step off of the cliff in an exploratory action. SARSA, on the other hand predicts action-values and takes a longer but safer path around the cliff.



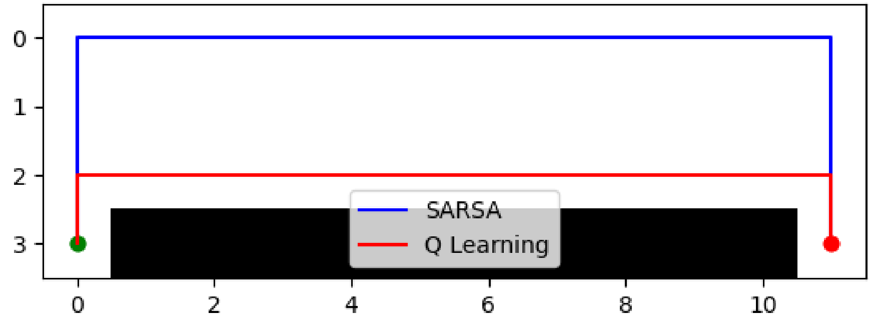
2.

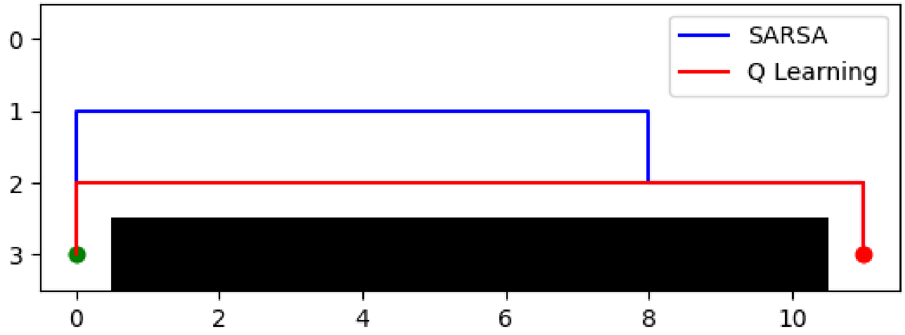


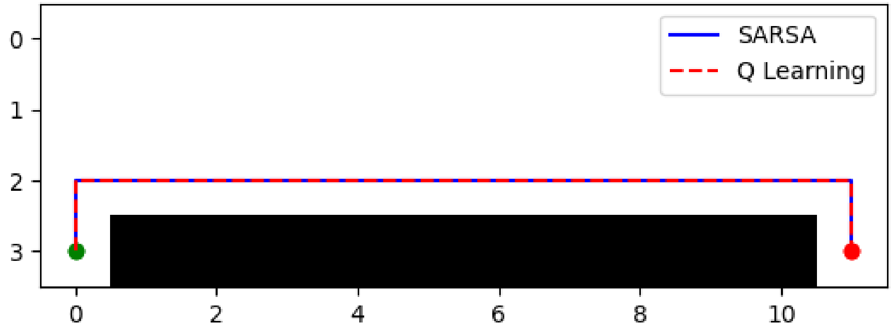
The figure above shows the running average of rewards received by each learning method over 1000 episodes. As expected the Q-learning method converges to a slightly lower average reward than SARSA. This is because although Q-learning learns the shortest, most optimal path to the goal, it will also occasionally fall off and receive a reward of -100. Although the SARSA method takes more steps to reach the goal and receives the reward of -1 more often, it will fall off the cliff and receive the -100 penalty far less, which keeps its average reward higher than the riskier Q-learning path.

3.

The following 3 figures show that both methods will converge to the optimal path by lowering the value of epsilon. This is because as the value of epsilon is lowered, less exploratory actions will be taken, and the SARSA method will not learn the additional value in taking the longer, safer path. Instead both functions will always choose the action with the maximum action value.


Cliff Walk Performance for Learning Methods, ε=0.1


Cliff Walk Performance for Learning Methods, ε=0.01


Cliff Walk Performance for Learning Methods, ε=0.0001