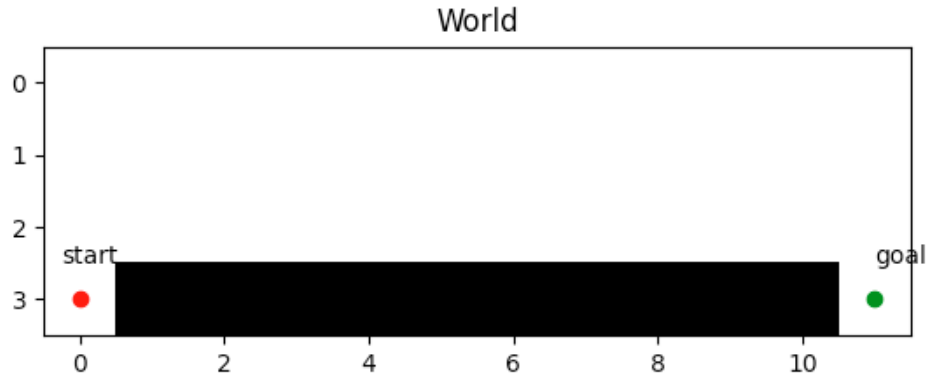
README.md

# Temporal-Difference

Cliff Walking Problem using SARSA and Q-Learning!

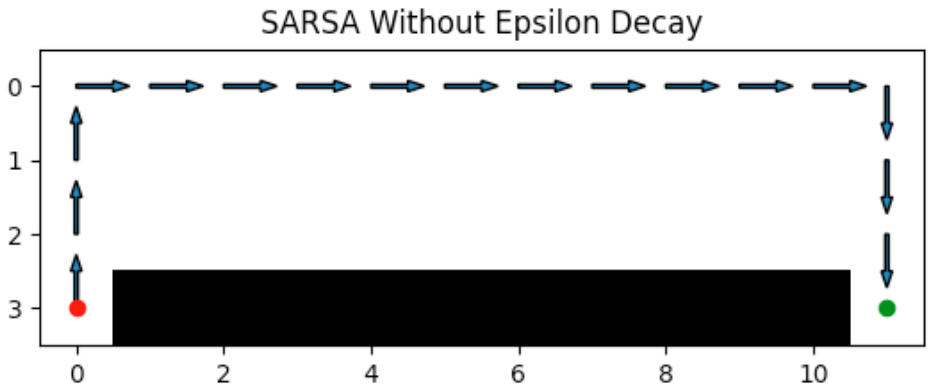
This grid world example compares SARSA and Q-learning, highlighting the difference between on-policy (SARSA) and off-policy (Q-learning) methods. Consider the grid world below. This is a standard undiscounted, episodic task, with start and goal states, and the usual actions causing movement up, down, right, and left. Reward is -1 on all transitions except those into the region marked “The Cliff.” Stepping into this region incurs a reward of -100 and sends the agent instantly back to the start.

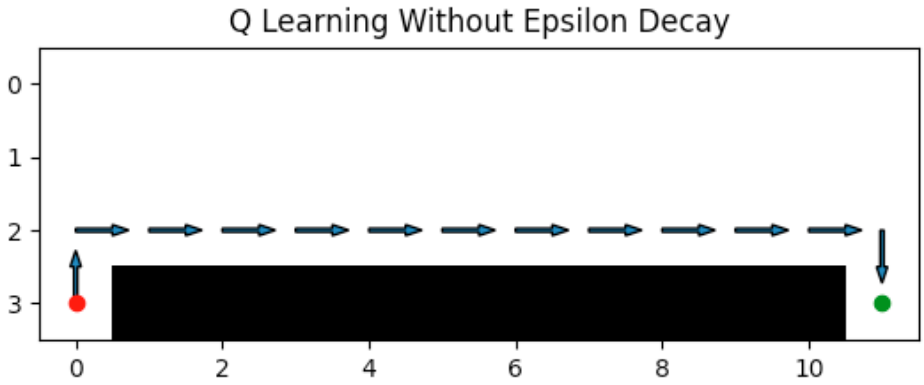



The following hyper prameters are used:

```
start = (3,0)
goal = (3,11)
alpha (step size) = 0.2
gamma (discount factor) = 0.9
epsilon (for epsilon-greedy)= 0.1
epsilon decay factor = 0.01
n_episodes = 1000
```

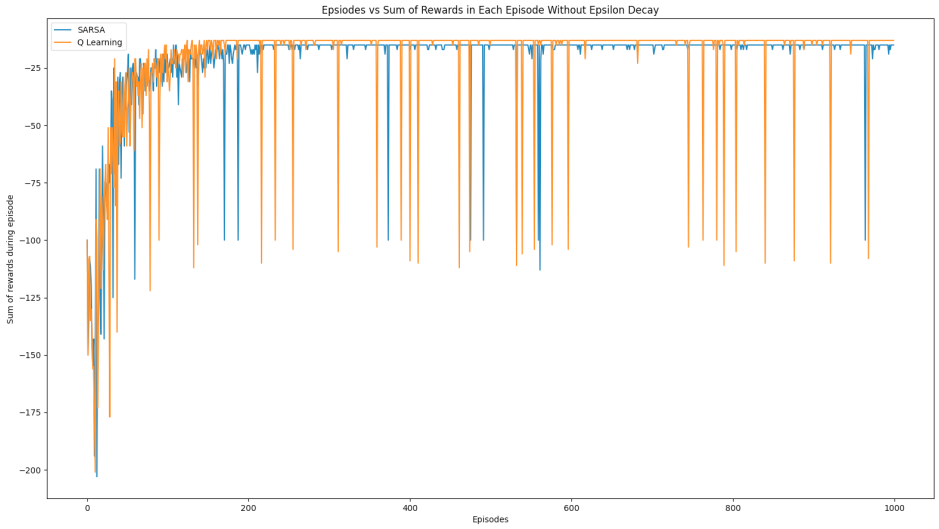
When there is no decay of epsilon in epsilon-greedy , SARSA converges to a longer safer path , where as the Q-learning converges to an optimal path. This is given by the following images.



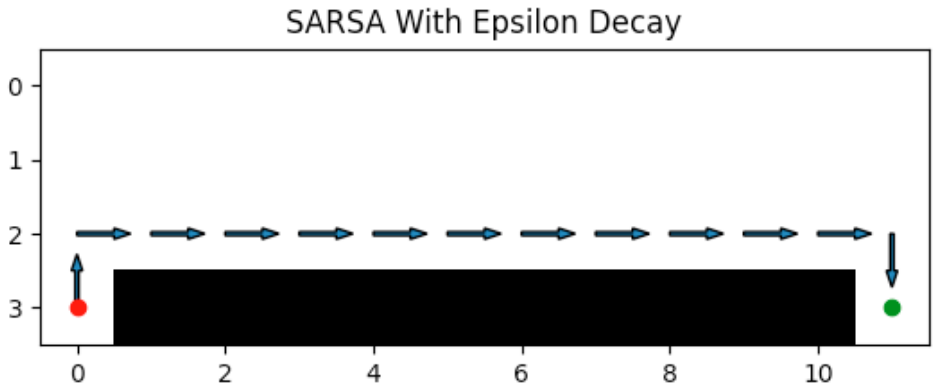



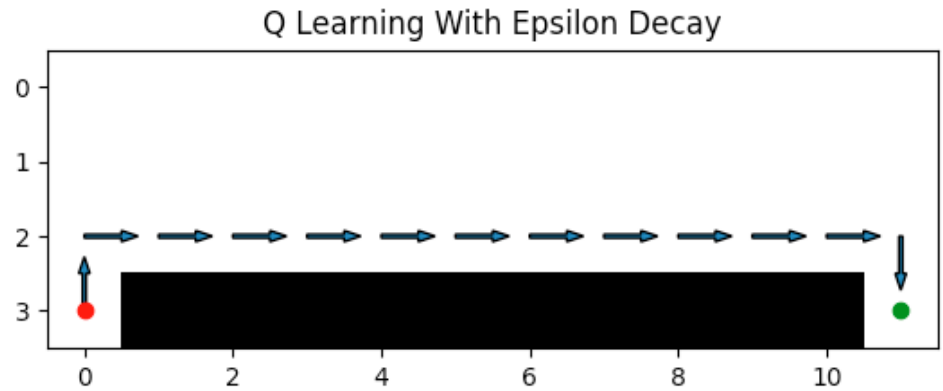

In Q-learning the exploring policy is epsilon-greedy , but the learning policy is greedy. But in SARSA both exploring and learning policies are epsilon-greedy! Because of this SARSA learns a safer path but Q-learning tries to learn optimal path,although occasionally it falls into the cliff , because of epsilon-greedy exploration!

The following is the plot of sum of rewards as function of episodes. In this , we can observe that the Q_learning receives more -100 rewards because it is occasionally trying the more riskier path thus obtaining the optimal path! That is why Q-learning has lower average reward compared to the SARSA!




But if the epsilon is gradually reduced after each episode, even the SARSA converges to the optimal path. This is happening because over the episodes since the epsilon is reduced , SARSA is acting as Q-learning ! That means when epsilon is reduced, the learning becomes greedy, because of the low epsilon value!



This can be clearly observed from the episodes vs cumulative rewards graph. In this the sum of rewards from both methods are same because they are acting as same at lower epsilon values!