Reinforcement Learning

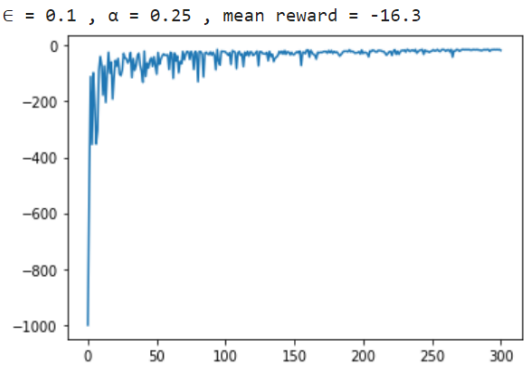
Assignment 2 Report

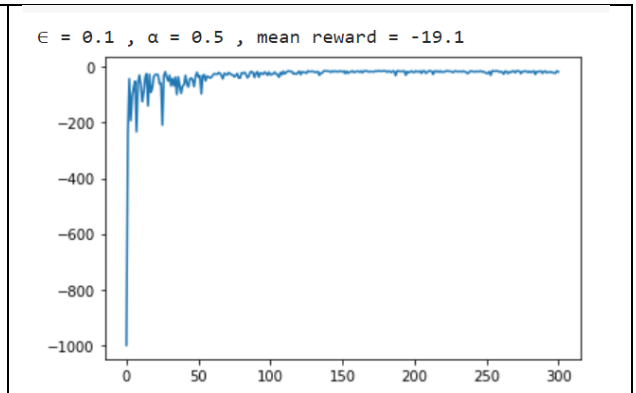
- **Introduction:**

SARSA'S policy performs better because its value function considers its Epsilon-greedy behavior. Q-Learning will learn faster because it directly learns the optimal policies value function.

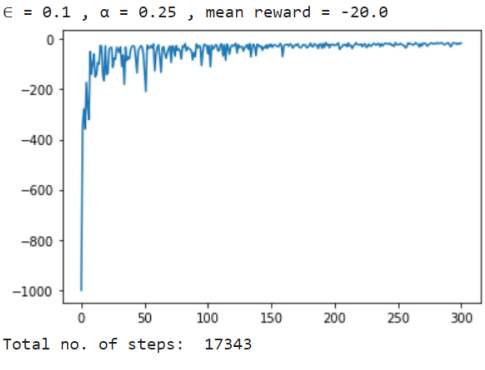
- **Windy Grid world**

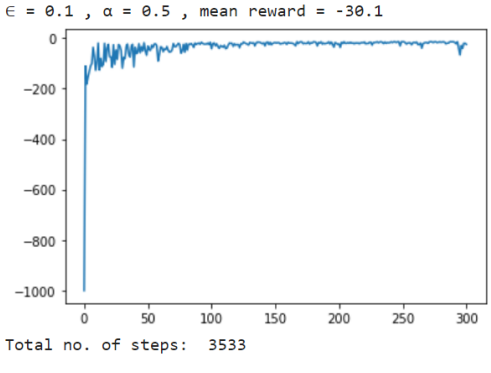
1- Q-Learning





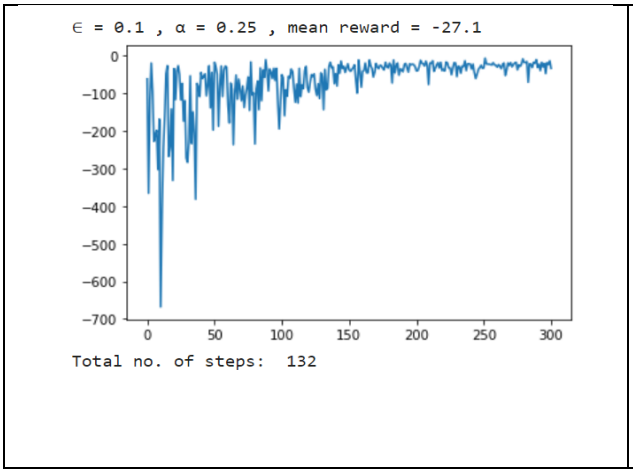
2- SARSA

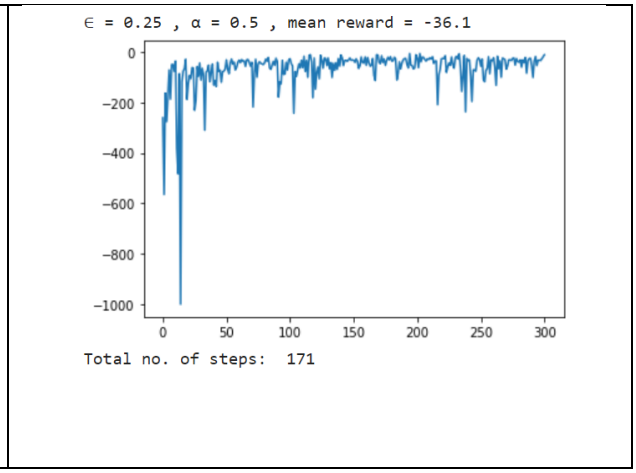




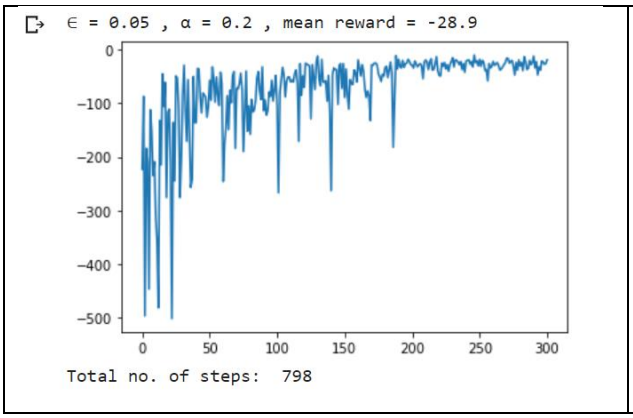
- **Stochastic Windy Grid world**

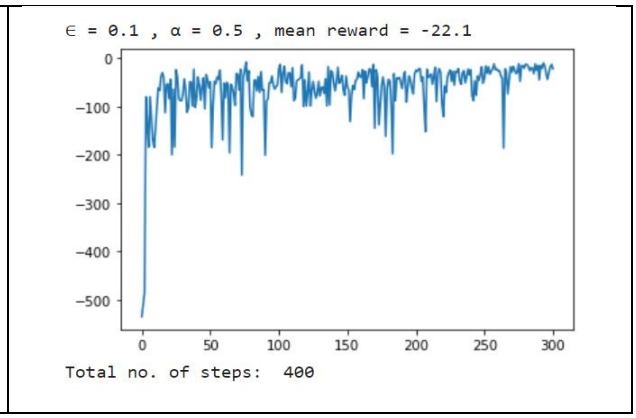
1- Q-Learning





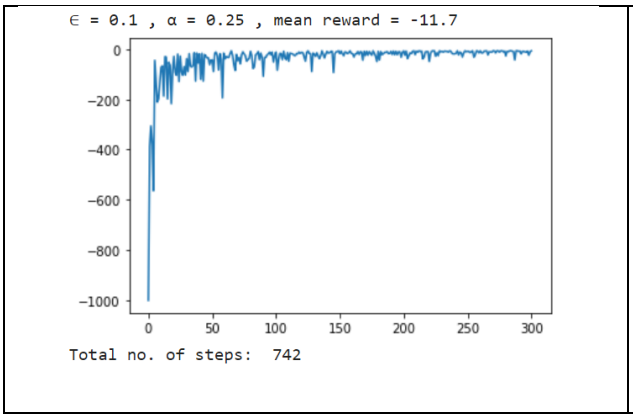
2- SARSA

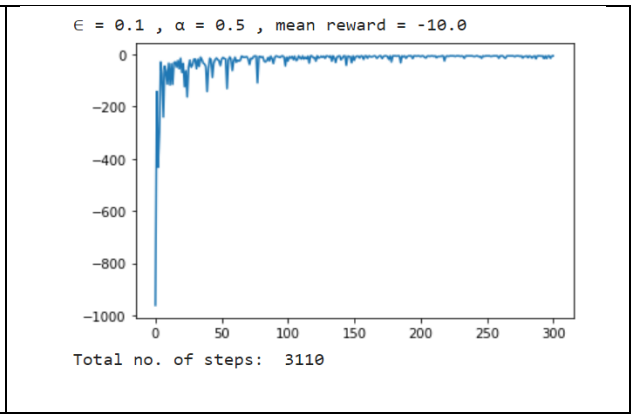




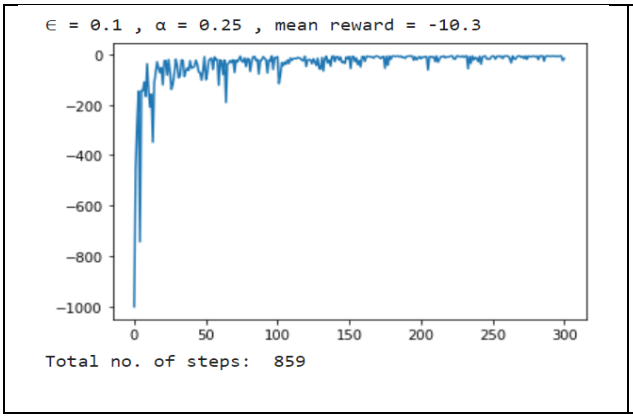
- **King Windy Grid world**

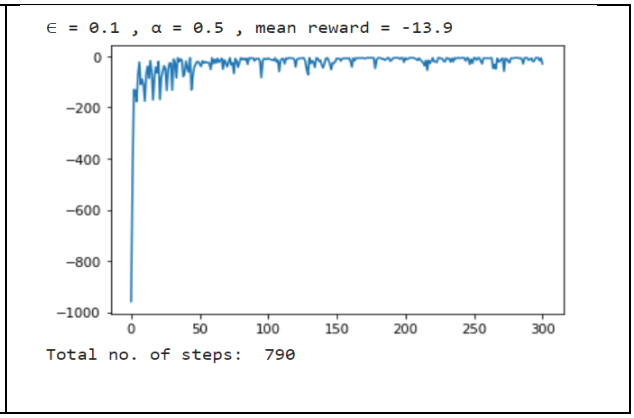
1- Q-Learning





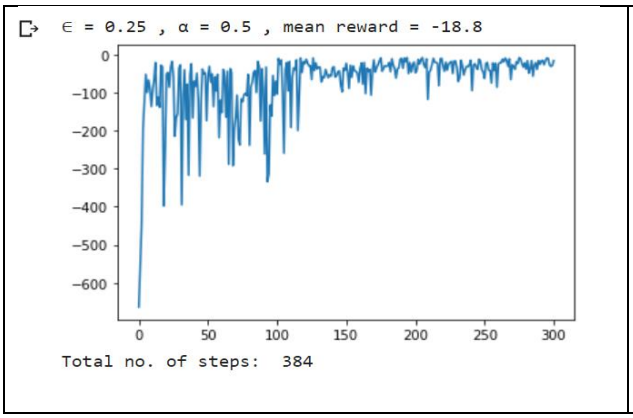
2- SARSA

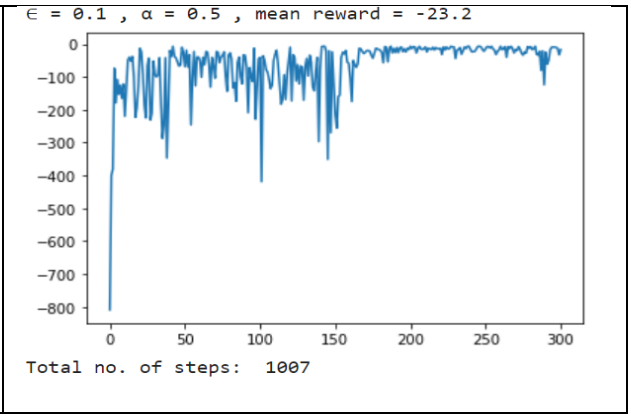




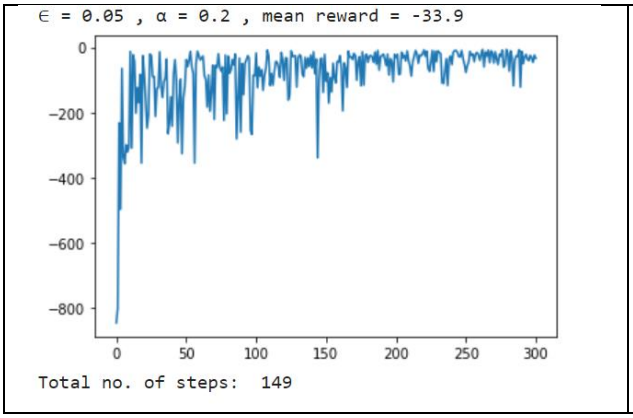
- **Stochastic King Windy Grid world**

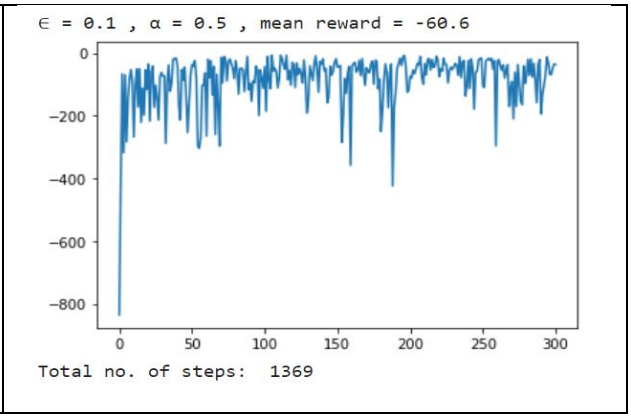
1- Q-Learning





2- SARSA

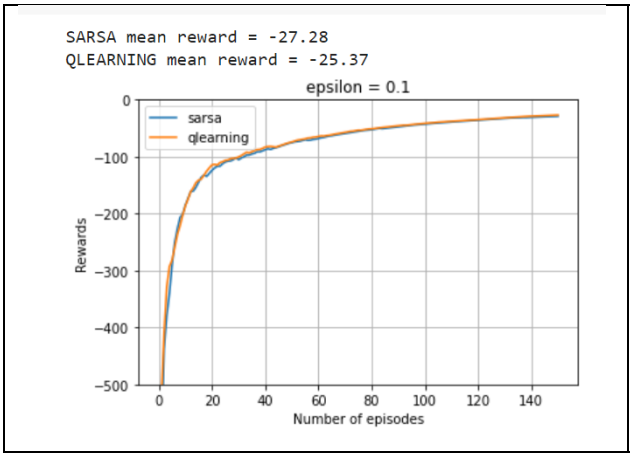


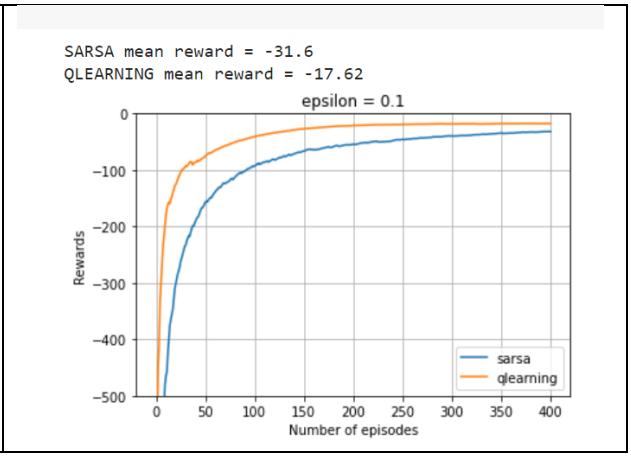


- **Compare Q-Learning & SARSA at same alpha and epsilon**

1- alpha=0.5, epsilon=0.1 at both.

2- alpha=0.1, epsilon=0.1 on SARSA, alpha=0.5, epsilon=0.1 on Q-Learning





- **Optimal Vaues:**

Windy Type	Epsilon	Alpha	Time Steps	Type
Windy Grid World	0.1	0.5	1160	Q-Learning
Windy Grid World	0.1	0.5	2402	SARSA
Stochastic Windy Grid World	0.1	0.25	18	Q-Learning
Stochastic Windy Grid World	0.1	0.1	457	SARSA
King Windy Grid World	0.1	0.25	63	Q-Learning

King Windy Grid World	0.05	0.2	867	SARSA
Stochastic King Windy Grid World	0.1	0.5	598	Q-Learning
Stochastic King Windy Grid World	0.1	0.5	232	SARSA

- **Conclusion**

At King Windy Grid World, when I selected learning rate 0.5, Q-Learning performed well but SARSA did not because it is too large for SARSA So I decreased it to 0.2 to see what would happen. In this way SARSA is learning slowly, but we give agent chance to find better policy.

Both are different because Q-Learning takes the max over the next action values. So, it only changes when the agent learns that one action is better than another. In contrast, SARSA uses the estimate of the next action value in its target. This changes every time the agent takes an exploratory action. SARSA learns the same final policy as Q-Learning, but more slowly. We know both algorithms have to converge to the same policy because the slopes of the lines are equal. Equal slopes mean that both agents are completing episodes at the same rate.