CS-747 Assignment 3 Report
Arun Verma, 154190002

PART 2: Average Cumulative Reward v/s **Lambda value** plot for Sarsa(lambda) with **Accumulating**-trace method and **Replacing**-trace method for Instance **0** (Figure 1) and Instance **1** (Figure 2). From both figures, it is experimentally verified that accumulating-trace method is better than replacing-trace method for given problem instances.

Experiment Setup Details: Side length: 32, slipping probability: 0.02, epsilon: 0.01, alpha: 0.1, #Episodes: 500, #timesteps: 1000, gamma: 1.

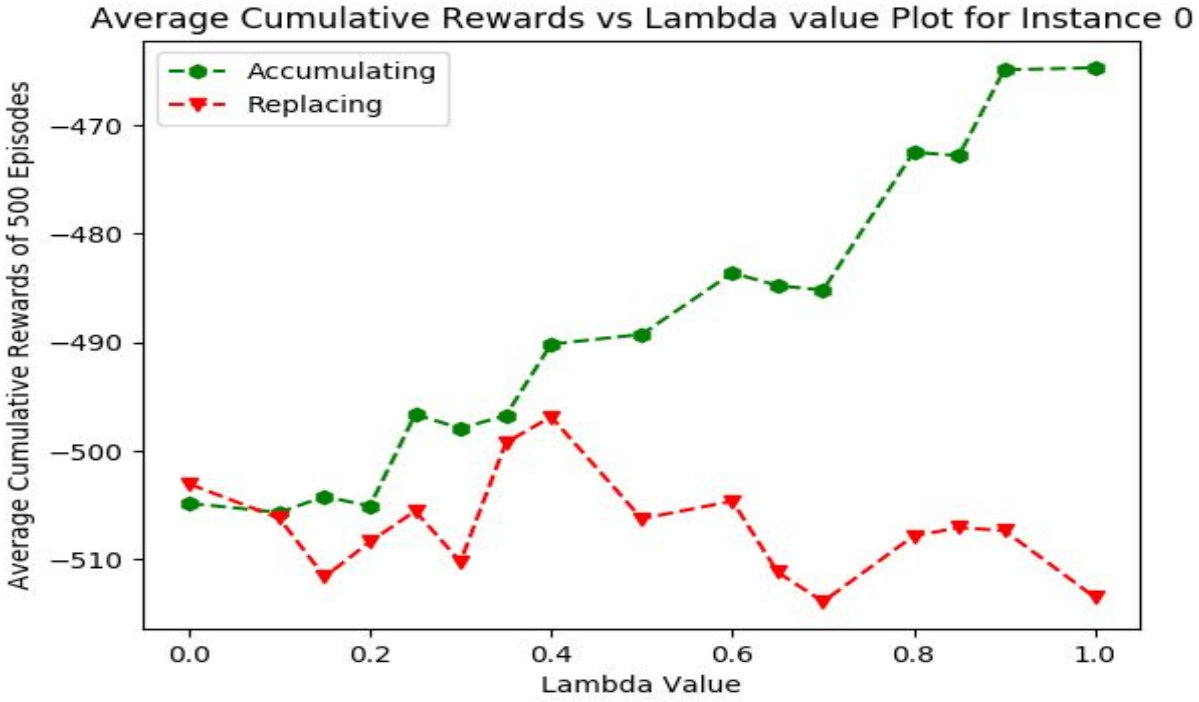

Figure 1

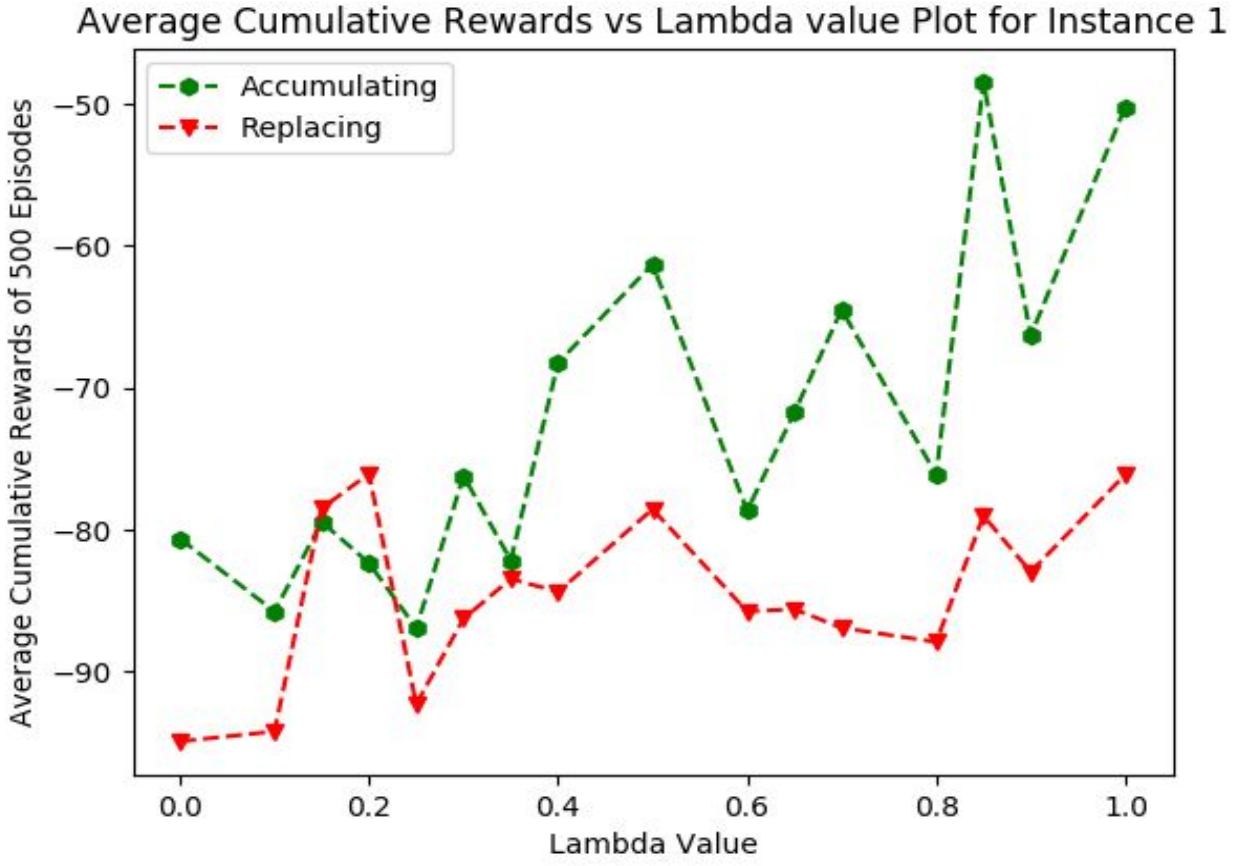

Figure 2

PART 1: Average Reward of each episode v/s **episode number** plot for **Q-learning** and **Sarsa(lambda)** for Instance **0** (Figure 3) and Instance **1** (Figure 4).
Experiment Setup Details: Side length: 32, slipping probability: 0.02, epsilon: 0.01, alpha: 0.1, #Episodes: 1500, #timesteps: 1000, gamma: 1, lambda: **0.9** for Instance 0 and **0.85** for Instance 1.

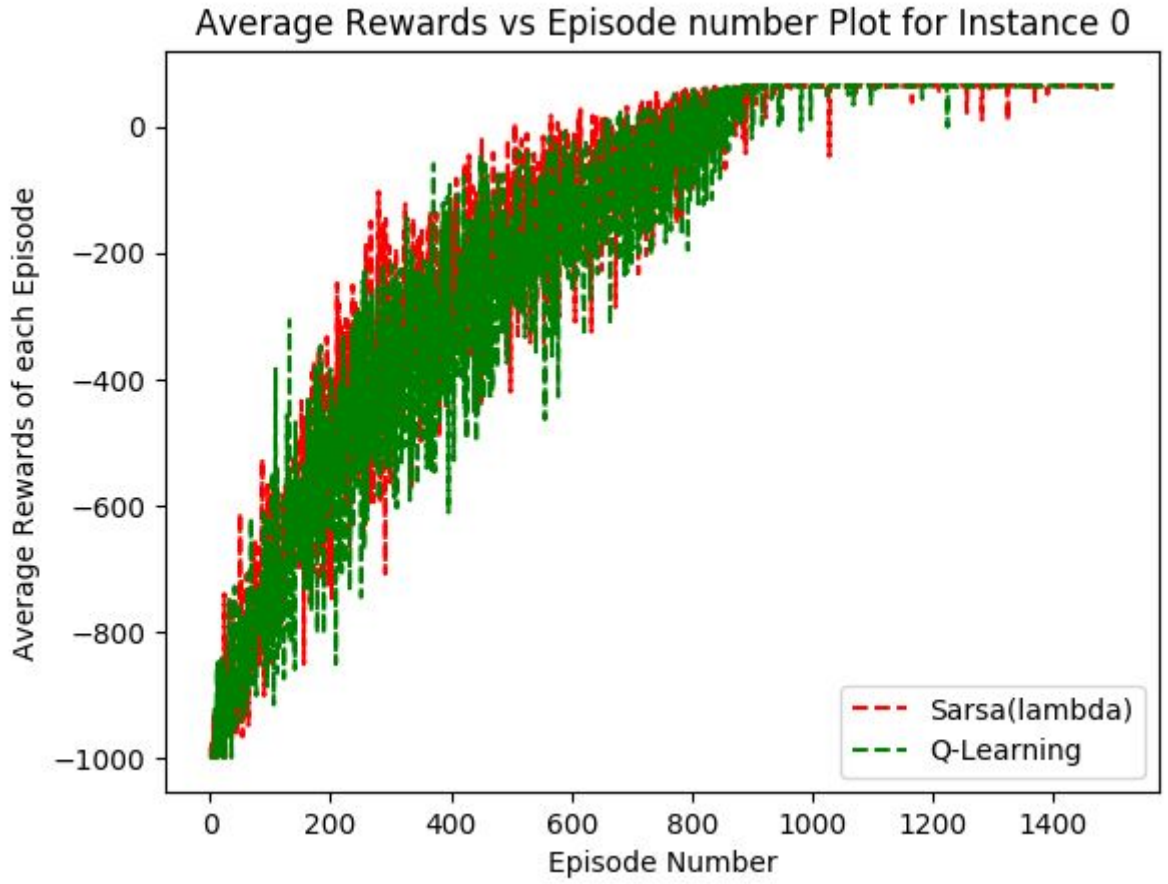

Figure 3

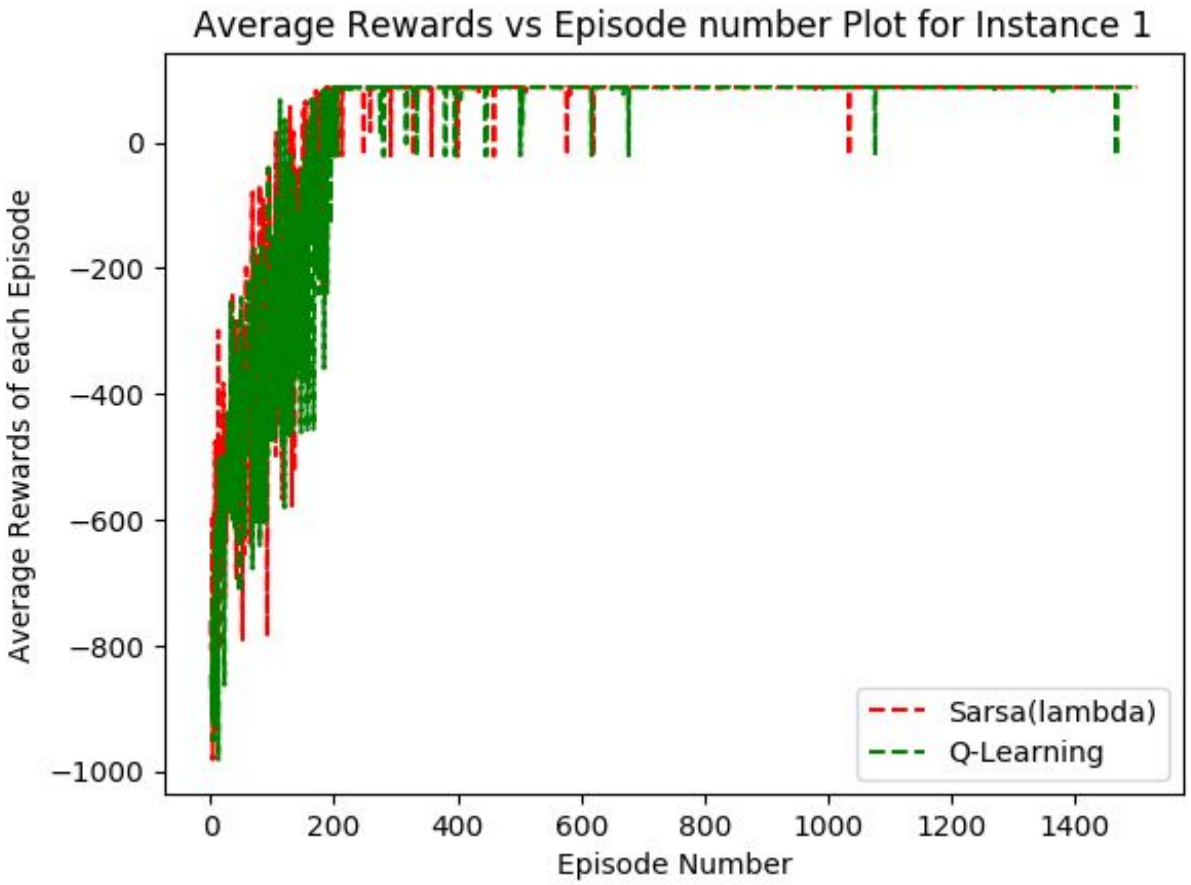

Figure 4

Explanation: After an initial transient, Q-learning learns values for the optimal policy, that which travels in the direction of destination. Unfortunately, this results in its occasionally getting into obstacles because of the ε-greedy action selection. Sarsa, on the other hand, takes the action selection into account and learns the longer but safer path. Although Q-learning actually learns the values of the optimal policy, its online performance is worse than that of Sarsa (as seen in **Figure 3 and 4**), which learns the roundabout policy. Of course, if ε were gradually reduced, then both methods would asymptotically converge to the optimal policy. [Explanation is taken from Reinforcement Learning: An Introduction, Sutton and Barto @2017 Book Edition, Page no. 141]

I have further tuned the **lambda value** for sarsa(lambda) algorithm using optimal interval given by **PART 2** results. For same experiment configuration, **Average Cumulative Reward** v/s **Lambda value** plot for Sarsa(lambda) with **accumulating** trace method for Instance **0** (Figure 5) and Instance **1** (Figure 6). **Average Reward of each episode** v/s **episode number** plot for **Q-learning** and **Sarsa(lambda)** for Instance **0** with l**ambda: 1.0** (Figure 7) and Instance **1** with **lambda: 0.98** (Figure 8).

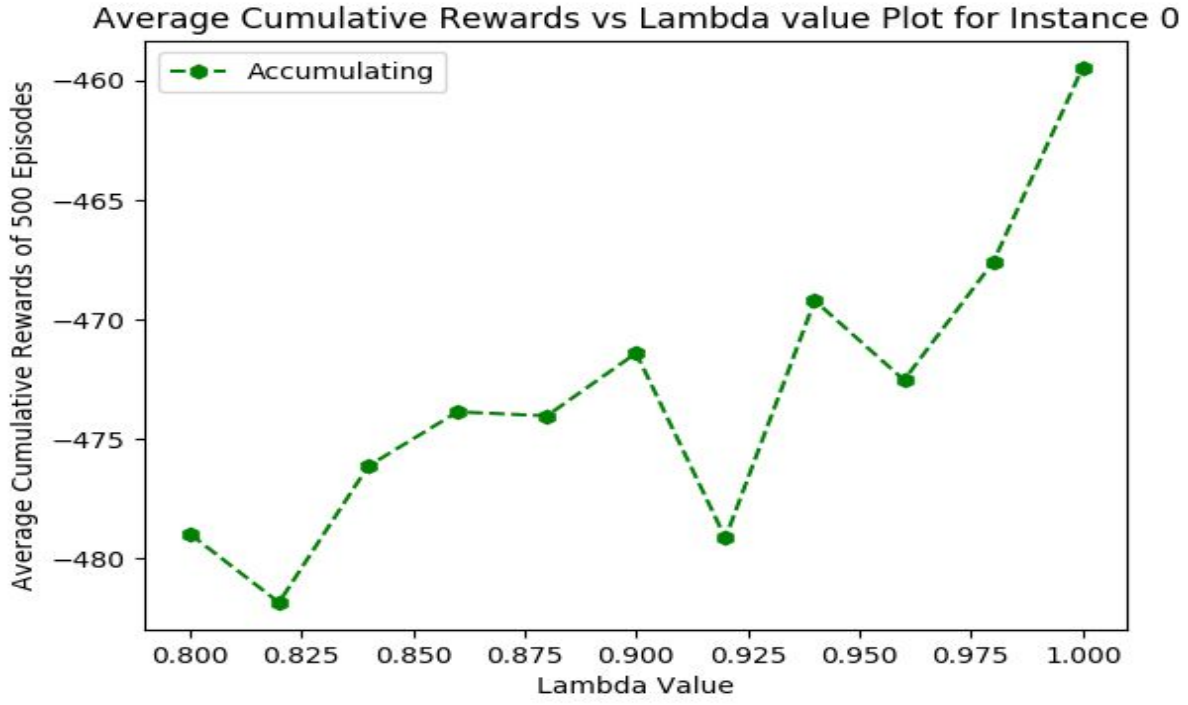

Figure 5

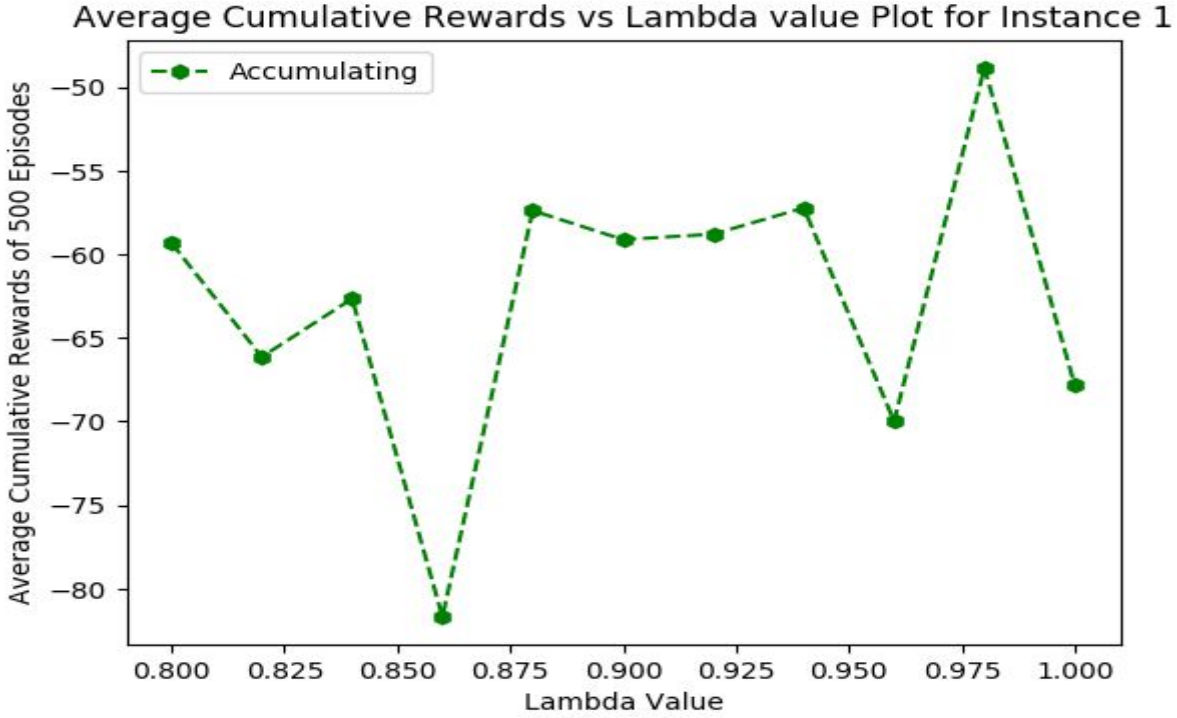

Figure 6

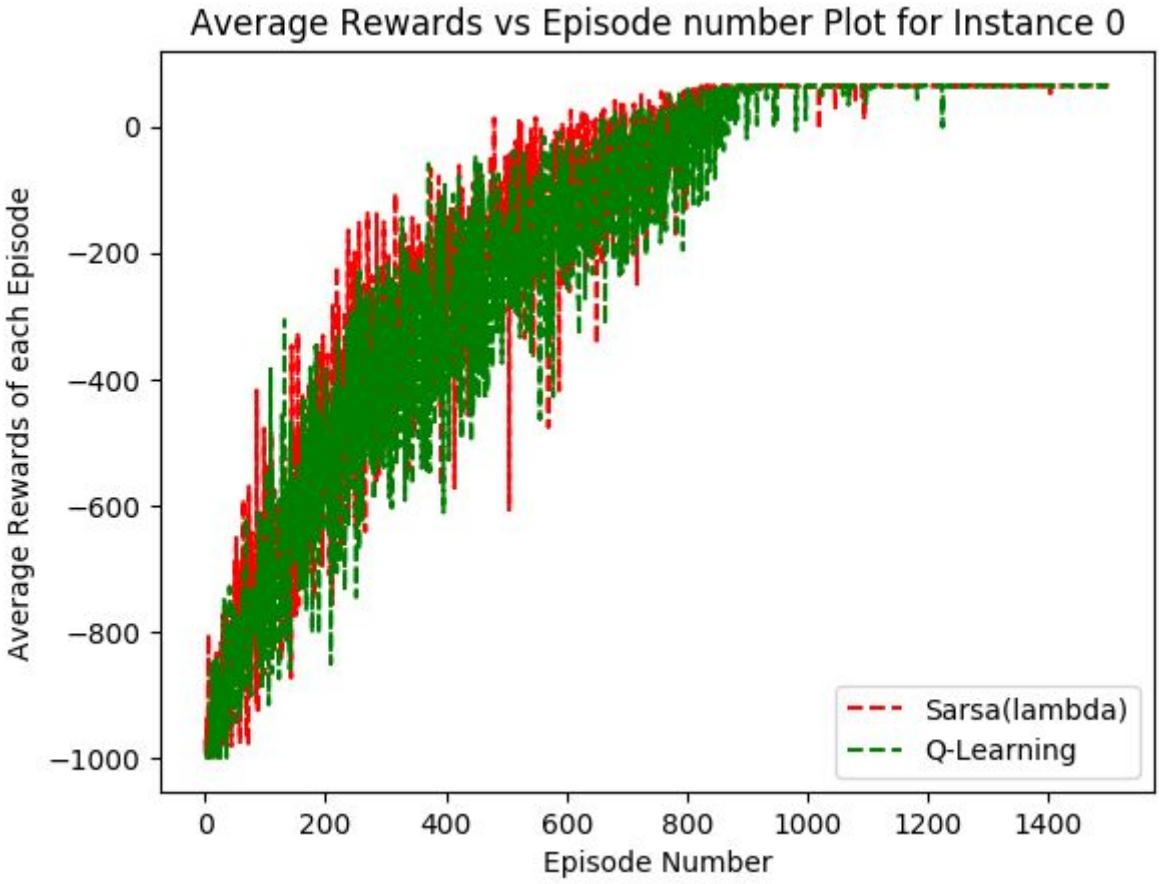

Figure 7

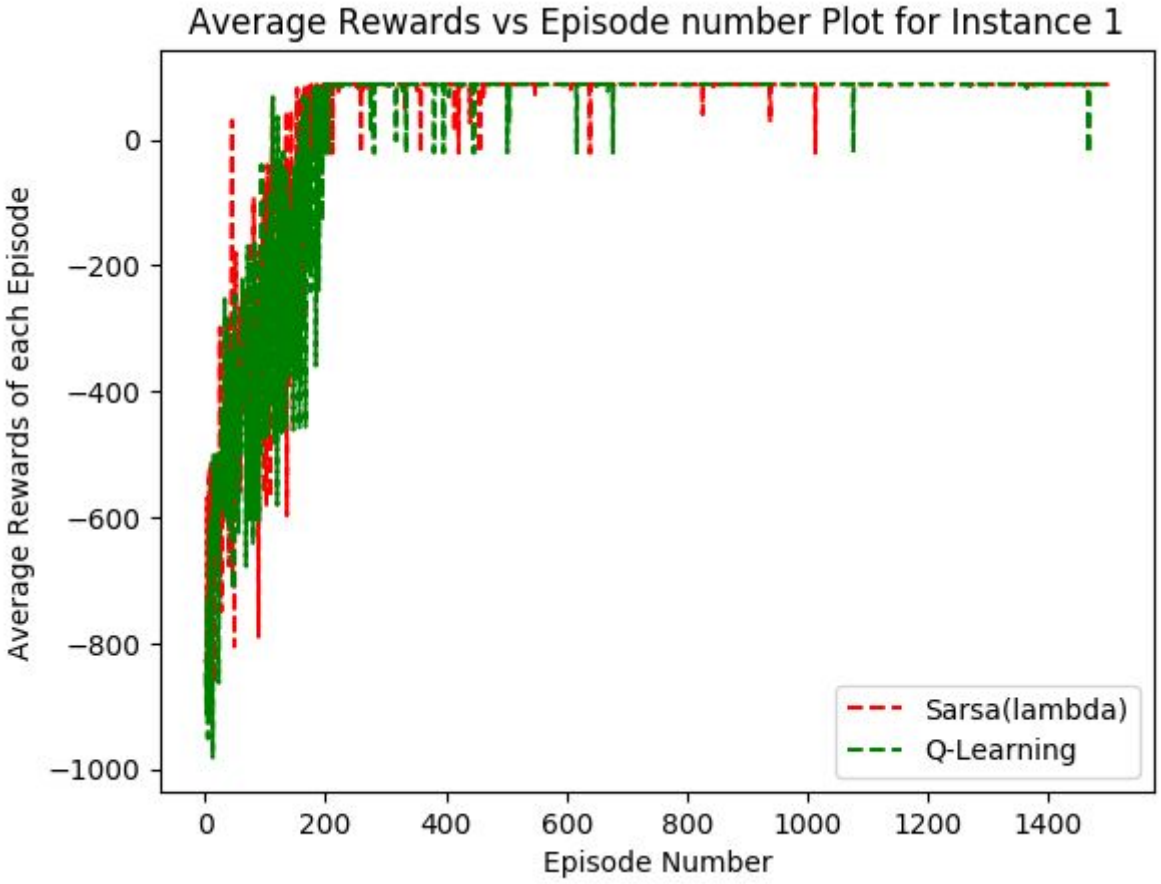

Figure 8