AIL722: Assignment 2 Report

Soumyodeep Dey
aiy237526

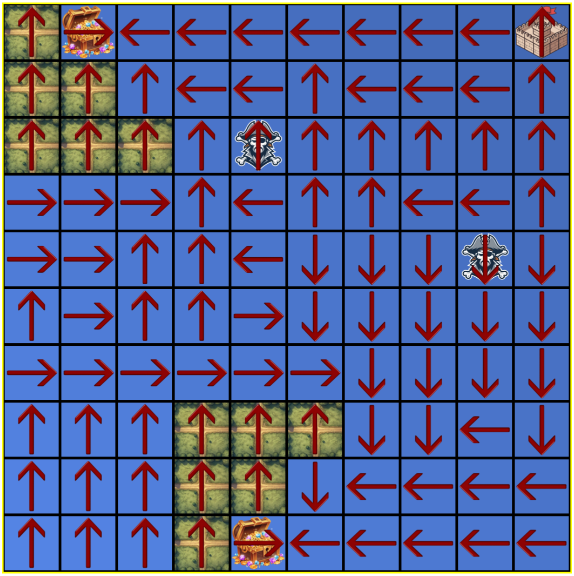
October 22, 2024

Contents

1 Model-Based Methods

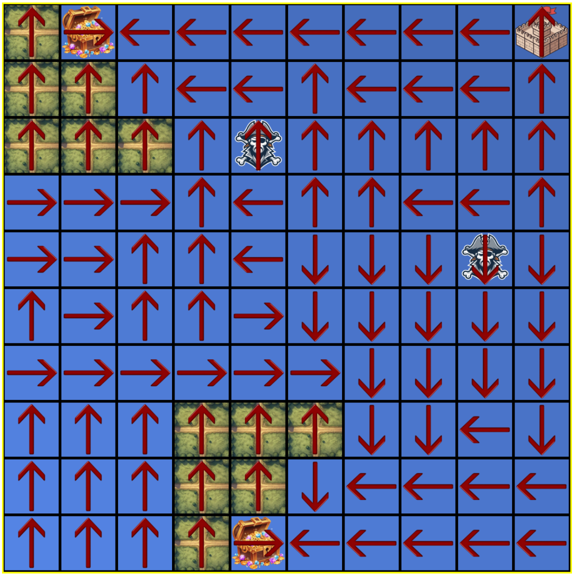
1.1 Policy Iteration

- **Time taken to converge**: 149 seconds.
- Plot of the Policy Learned:



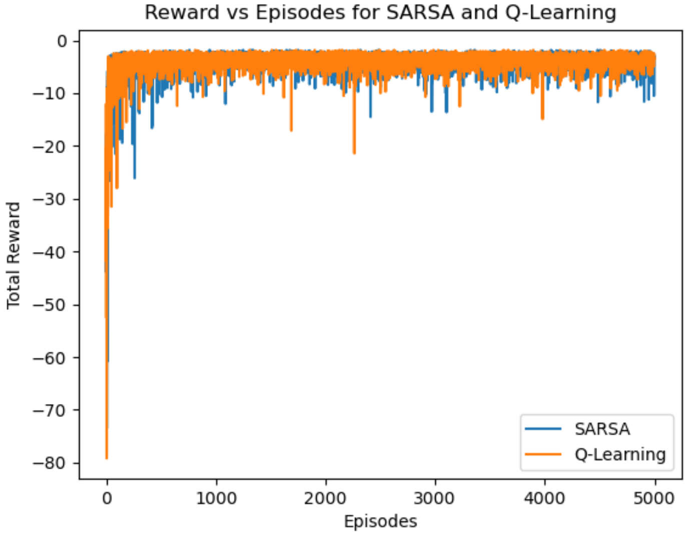
1.2 Value Iteration

- **Time taken to converge**: 149 seconds.
- Plot of the Policy Learned:

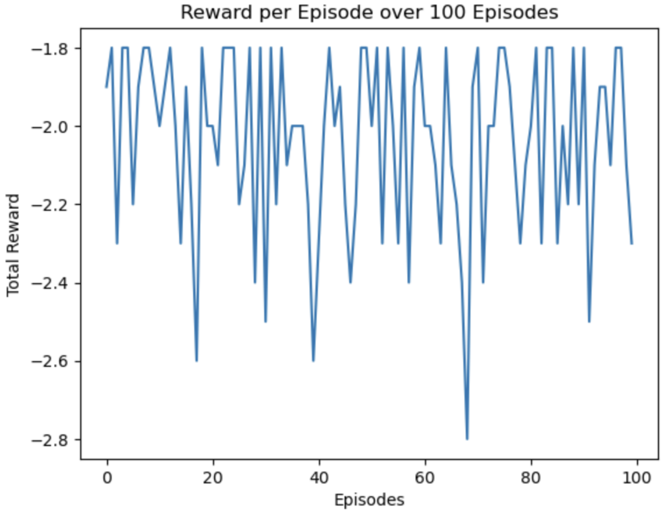


2 Model-Free Methods

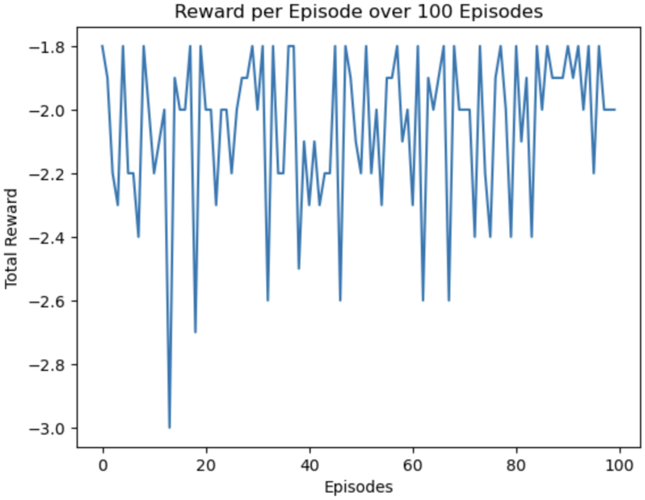
2.1 SARSA and Q-Learning on TreasureHunt-v1



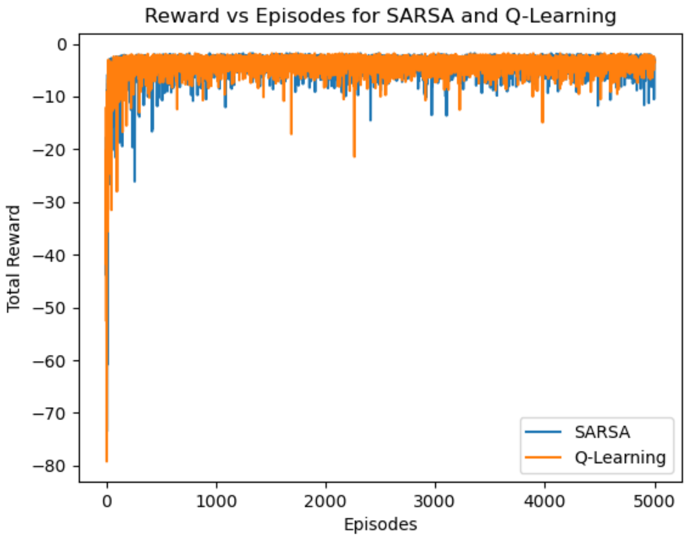
```
For Q−Learning
Results after 100 episodes:
Average reward per episode: −2.0350000000000015
```

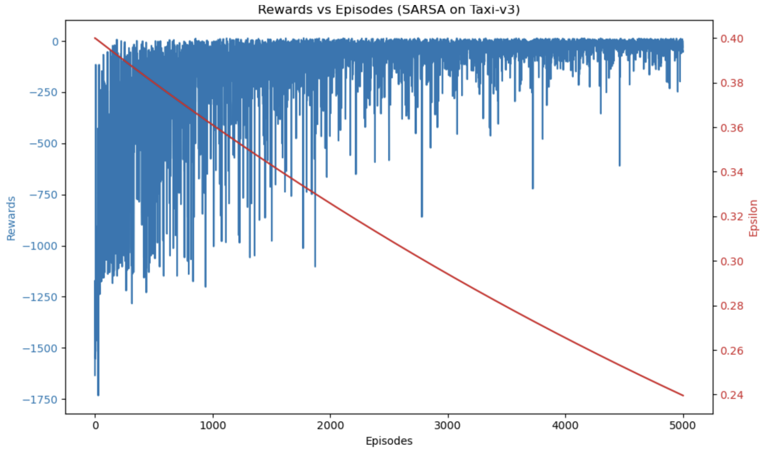


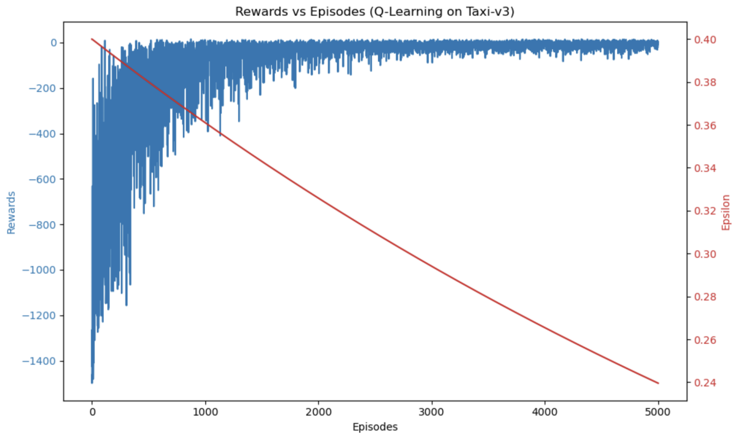
```
For SARSA:
Results after 100 episodes:
Average reward per episode: −2.0490000000000013
```



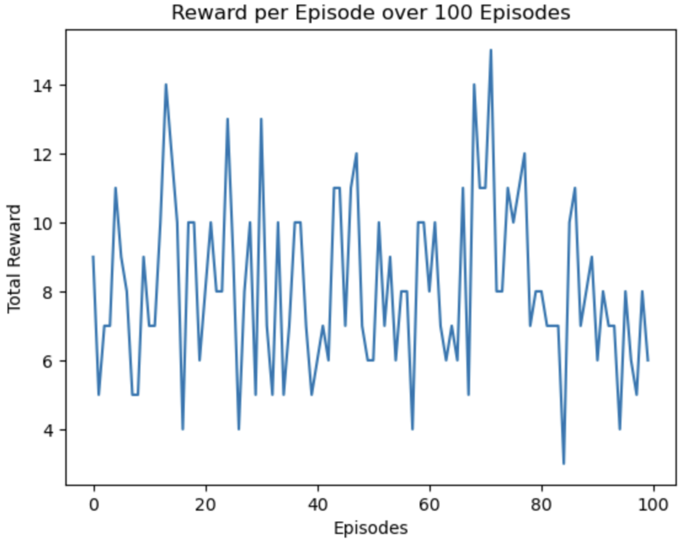
2.2 SARSA and Q-Learning on Taxi-v3



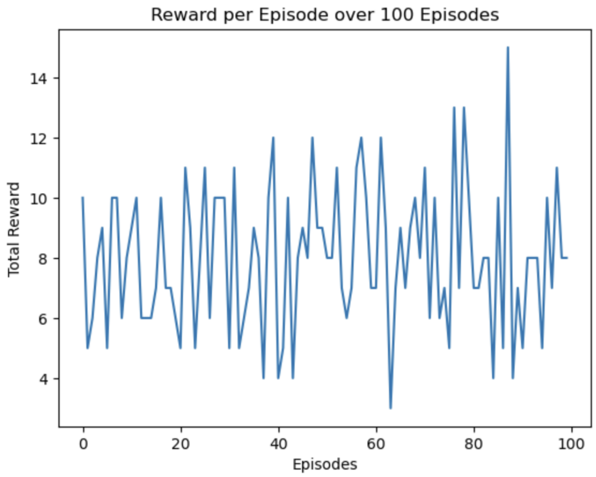




```
Results after 100 episodes:
Average reward per episode: 8.17
```



```
Results after 100 episodes:
Average reward per episode: 8.0
```



Q-Learning always converged quicker than SARSA. The convergence criteria is set on the Q value function that is being learnt after every episode. If the change in Q value does not change more than delta than the algorithm is assumed to converge. Delta set here is $\Delta = 1e - 3$.

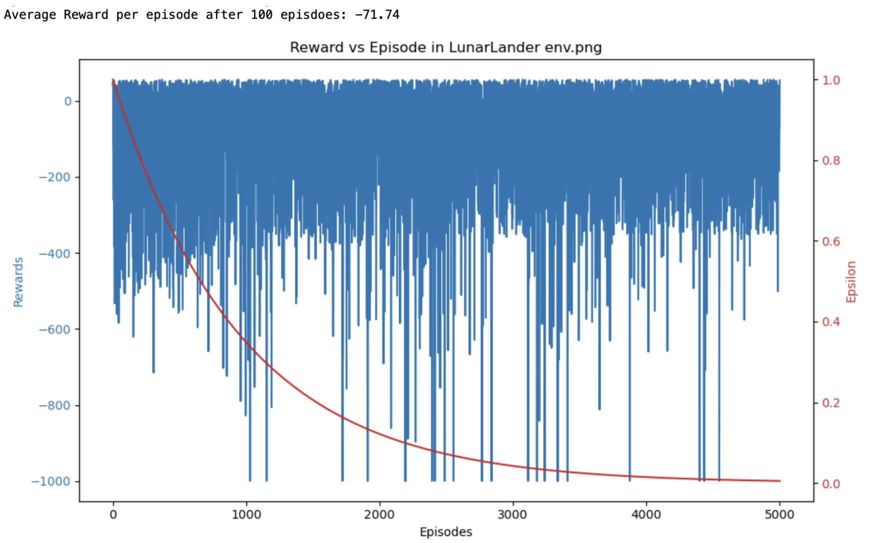
Algorithm	**CPU Time (s)**	**Number of Episodes**
Q-Learning (TreasureHunt)	7.50	2833
SARSA (TreasureHunt)	10.3	5655
Q-Learning (Taxi-v3)	58	7732
SARSA (Taxi-v3)	52	9465

3 Part 3

3.1 LunarLander-v2

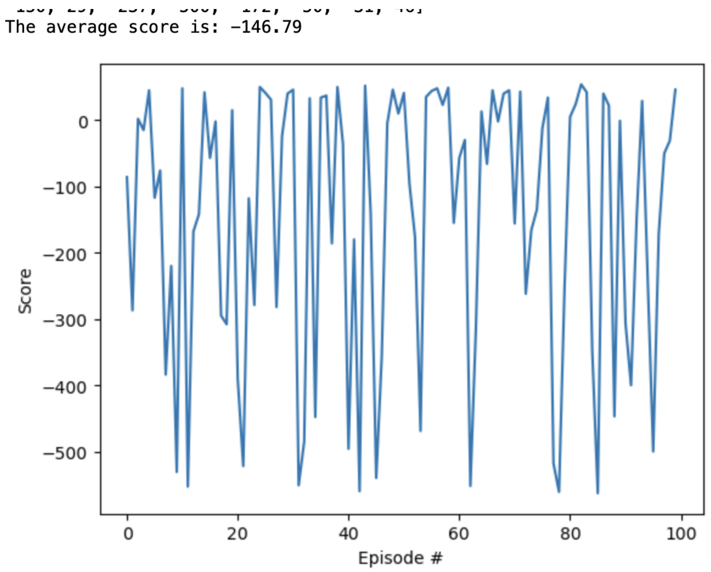
Time taken to run 50000 episodes: 28 minutes

Following is the best result of average reward per episode for $decayfactor = 0.85$ and all other hyperparameters as instructed.



```
Average Reward per episode after 100 episdoes: −71.74
```

The average score for a random untrained agent is as follows:



```
-150, 29, -237, -300, -172, -50, -51, 40]
The average score is: −146.79
```