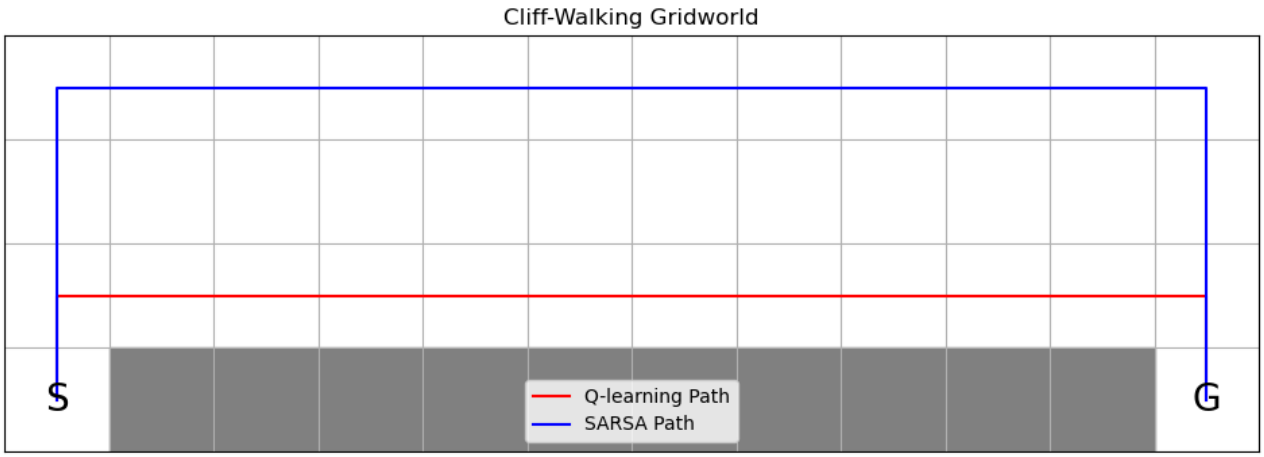
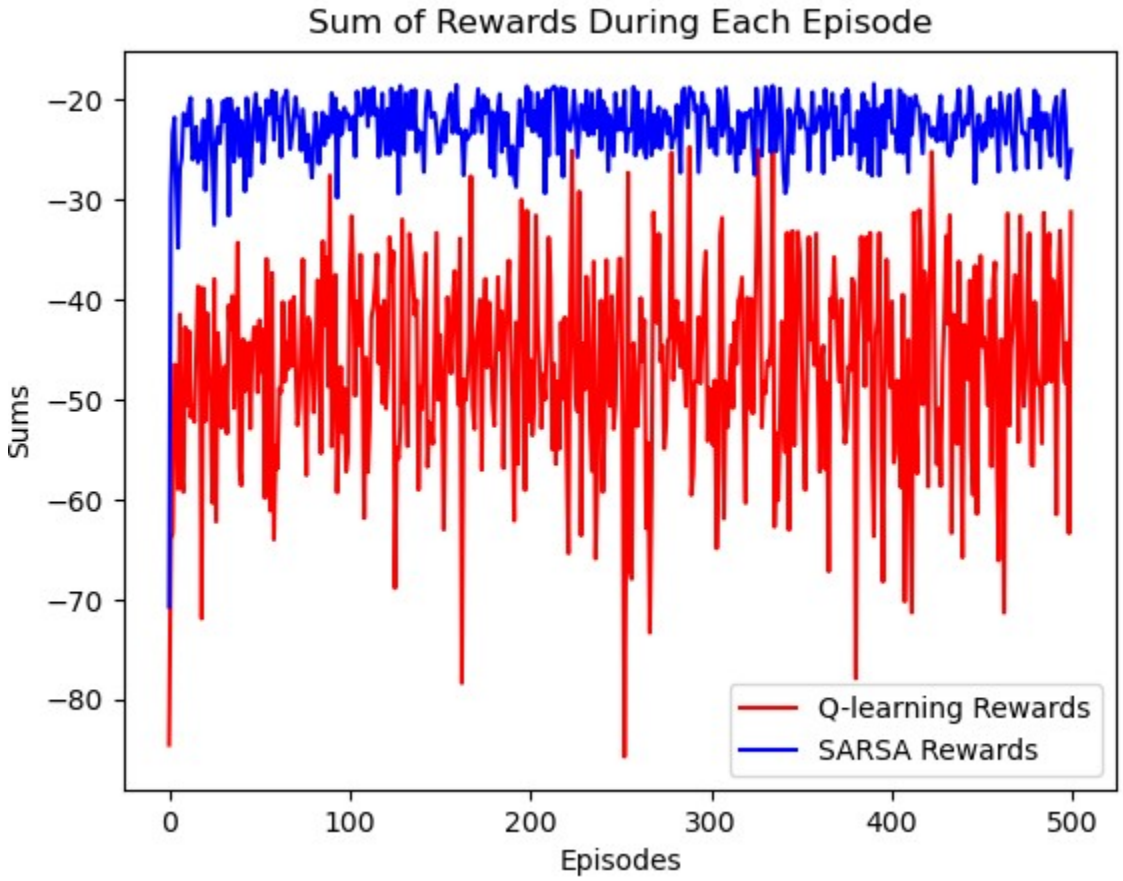
Taylor Bergeron and Arjan Gupta

Week 7 Temporal Difference Exercise

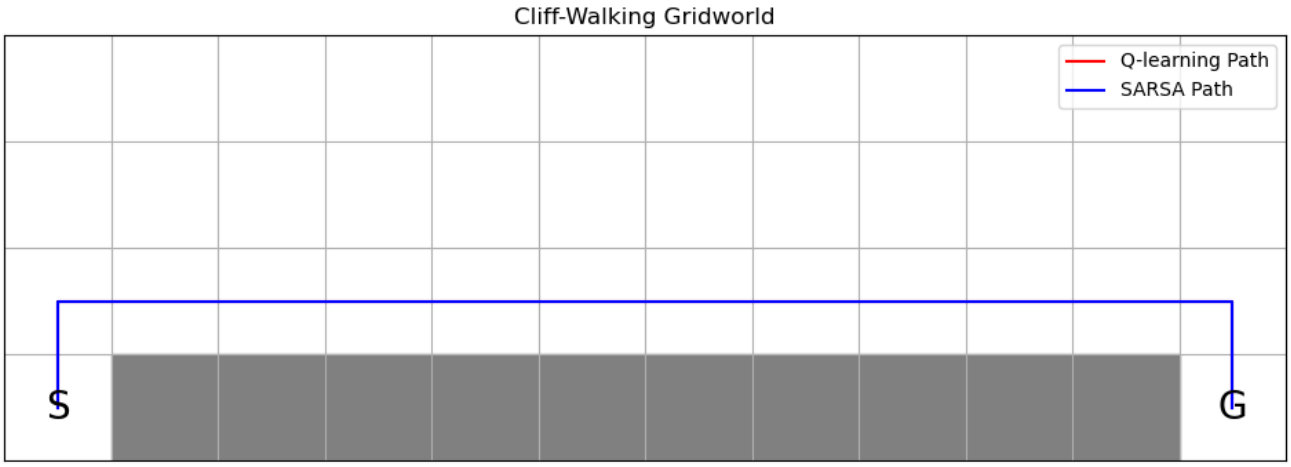
Plots

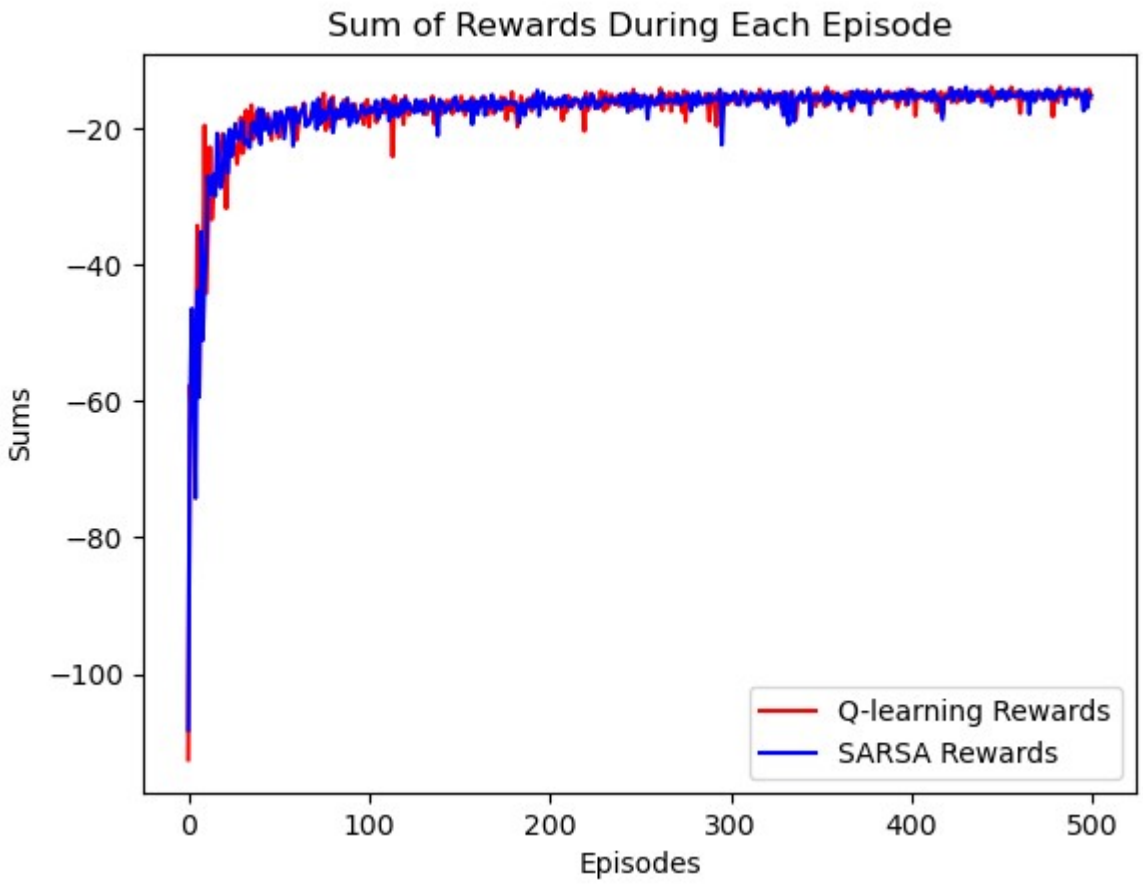




Questions

1. SARSA has randomness, due to epsilon, in both the exploring and learning policy. Q-learning only has randomness in the exploring policy, not the learning. Because SARSA has randomness in the learning it has more randomness in the actions chosen. Q-learning attempts to learn the optimal path but occasionally falls into the cliff because of the randomness from the epsilon exploring. However, it ends with the optimal policy, which has the least amount of -1 rewards returned, as the path is the shortest possible without falling off the cliff. This means that the performance of Q-learning during the episodes (i.e. its online performance) is worse than SARSA
2. Q-learning converges to a lower average rewards since it falls into the cliff occasionally due to the randomness in exploring due to epsilon-greedy. Therefore, it encounters more -100 rewards than SARSA.
3. After implementing epsilon decay over each episode, SARSA converges to the optimal path (the Q-learning path is rendered under the SARSA path). In order to get SARSA to fully converge to the optimal path, we used a low alpha of 0.01 and a decay rate of 0.9. The convergence to the optimal path is also reflected in the rewards graph in case of epsilon decay.




Sum of Rewards During Each Episode
−20
−40
−60
−80
−100
Sums
0
100
200
300
400
500
Episodes
Q-learning Rewards
SARSA Rewards