CS – 541 Artificial Intelligence
Assignment 3

Conclusion

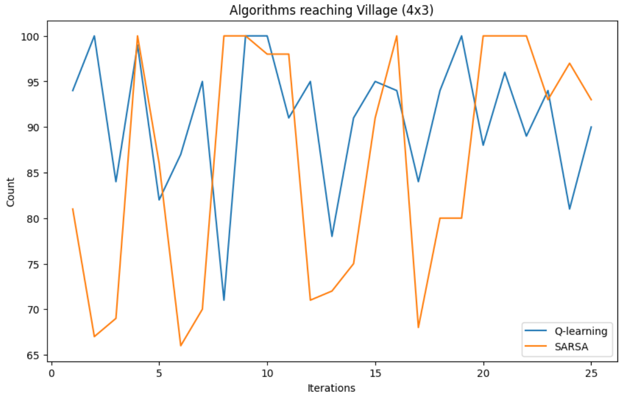
Below are outputs of running SARSA and Q-learning 25 times and recording how many times each algorithm reaches either Village, Volcano, or Beautiful View. To the left are outputs on the 4x3 grid and on the right are the outputs on 8x6 grid.

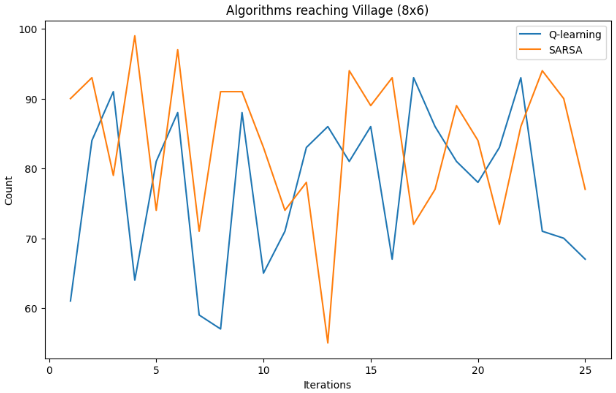
Regarding the exploration probabilities, both algorithms make use of this metric. Both algorithms use the epsilon-greedy policy which means that the agent will take a random action instead of following the action that is learned. The difference between both algorithms is the way states and actions are changed.

It was observed that increasing the exploration probability makes the algorithms reach optimal policy faster thus, faster convergence. However, this comes at the cost of high variance in Q-values. The reverse is true for low exploration probabilities, the algorithm will converge slowly with reduced variance in Q-values.

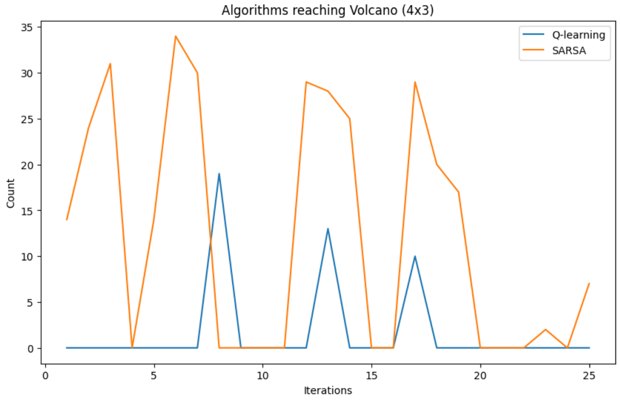
Furthermore, it was observed that the average rewards in iterations were greater for Q-learning than SARSA. This could be because Q-learning directly learns the optimal policy, whilst SARSA learns a near-optimal policy whilst exploring.

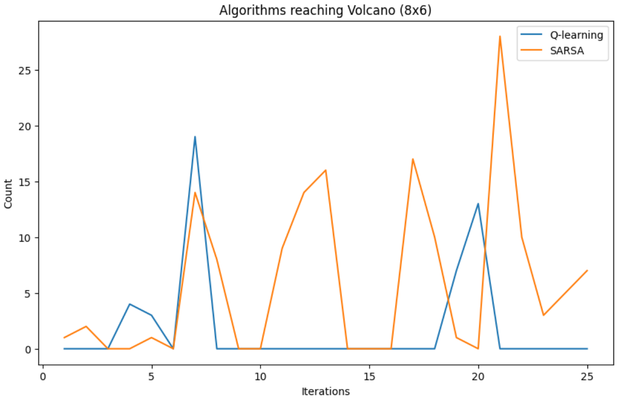
End state: Village



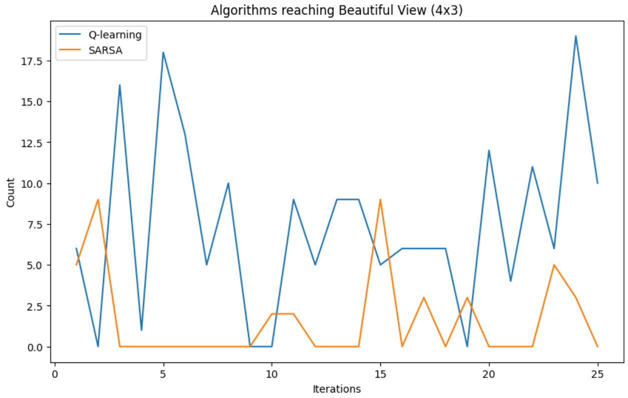


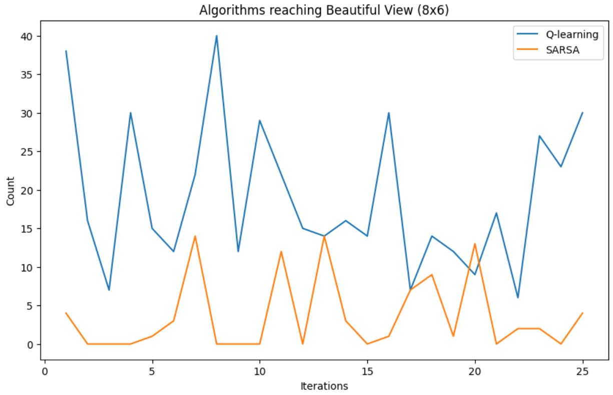
End state: Volcano



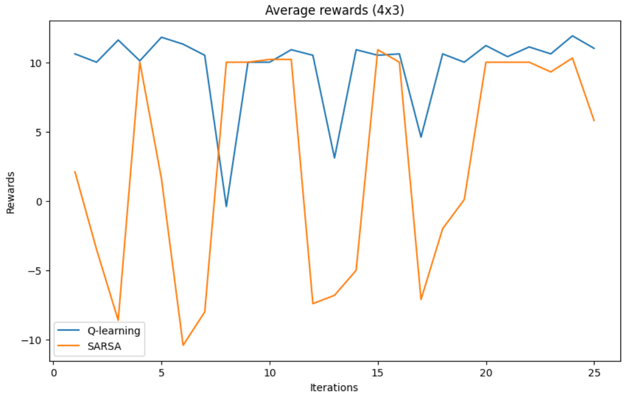


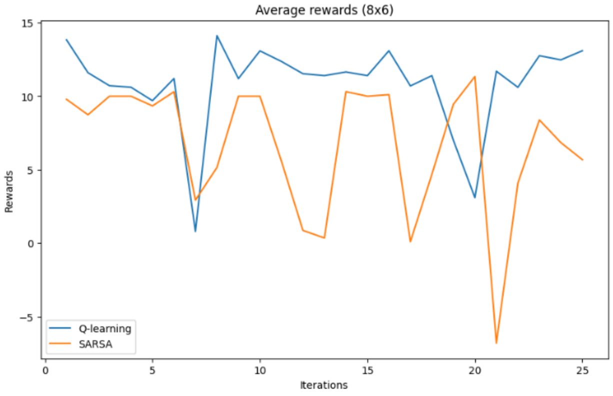
End state: Beautiful View





Average reward


Average rewards (4x3)
Q-learning
SARSA
Rewards
Iterations


Average rewards (8x6)
Q-learning
SARSA
Rewards
Iterations