REINFORCEMENT LEARNING
(CS3316)

# Assignment Report

## Assignment 3

## CLIFF WALKING
## SARSA AND Q-LEARNING

Name: Kezhi Li ID: 520021911013
Date: 31 March 2023

# Contents

# 1 Introduction

This assignment requires us to use the on-policy learning method Sarsa and the off-policy learning method Q-learning for finding an optimal route from the start point to the goal point shown in Figure 1. The grid world specializes a cliff region where the player will return to the beginning and get a reward -100. Other movement will cause a reward -1 as usual until the player reaches the goal point.

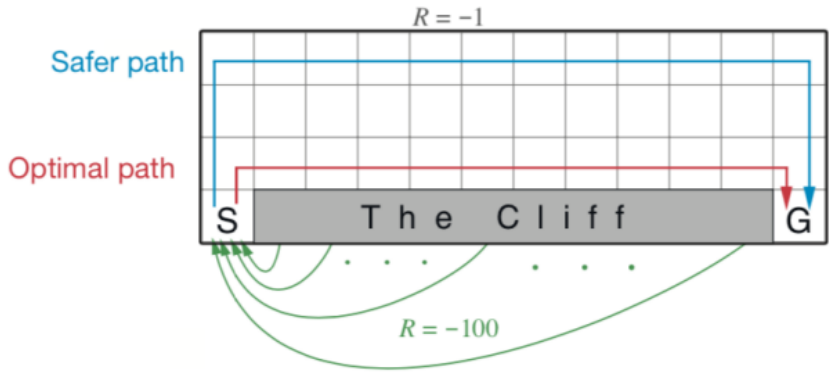


Figure 1: Cliff Walking

# 2 Experiment

To explore the abilities of Sarsa and Q-learning at finding the optimal solution of these model-free questions, the optimal actions gotten from these two methods will be shown. The setting $\epsilon$ also has a big influence on the results, so different choices of $\epsilon$ will also be used for comparison.

# 3 Results

## 3.1 Parameters Chosen

$\gamma$ of the whole assignment is chosen to be 1. $\alpha$ of Sarsa and Q-learning is chosen to be 0.05.

## 3.2 Sarsa

After running 1000 iterations on the grid world, Sarsa showed the optimal actions of each grid as Figure 2 with $\epsilon = 0.5$. Then, the optimal route will be going upward to the ceiling firstly, and then go left, and finally go downward.

Interestingly, when $\epsilon = 0.1$, the final optimal route altered with a shorter route and closer to the cliff, which can be shown in Figure 3.

With $\epsilon = 0.001$, the final optimal route became even closer to the cliff, and the route length became the shortest, shown in Figure 4. The final results shown in Figure 5 with $\epsilon = 0$ is similar.

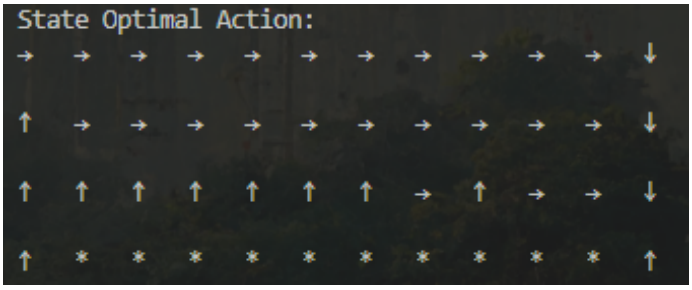


Figure 2: Optimal actions by Sarsa with $\epsilon = 0.5$

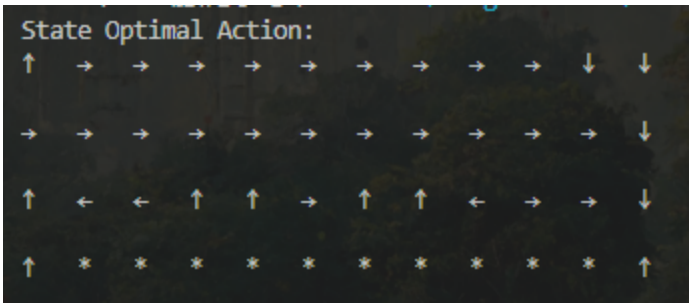


Figure 3: Optimal actions by Sarsa with $\epsilon = 0.1$

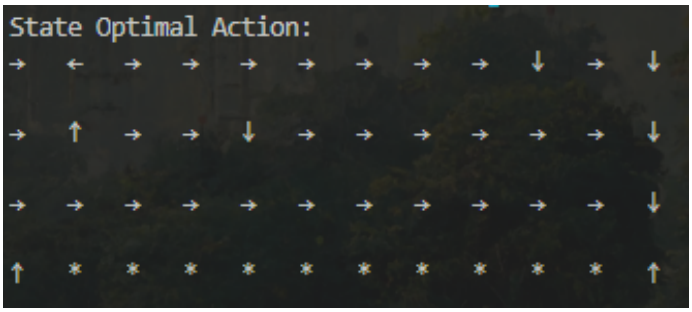


Figure 4: Optimal actions by Sarsa with $\epsilon = 0.01$

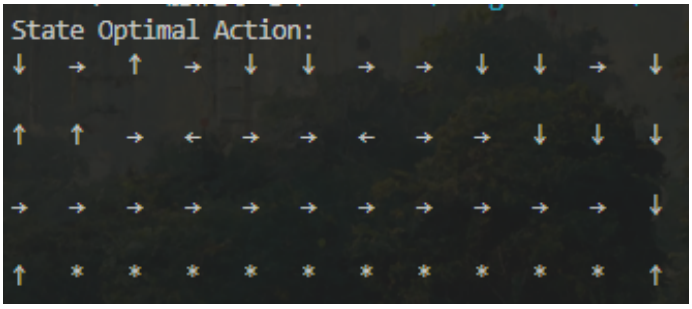


Figure 5: Optimal actions by Sarsa with $\epsilon = 0$

### 3.3 Q-learning

Also with 1000 iterations, the optimal travel route for $\epsilon = 0.1$ and $\epsilon = 0$ are shown in Figure 6 and Figure 7.

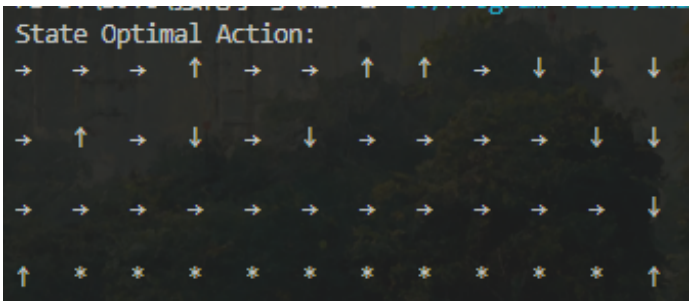


Figure 6: Optimal actions by Q-learning with $\epsilon = 0.1$

## 4 Conclusion

The difference of the optimal routes of Sarsa method with different epsilons is interesting. It seems that with higher $\epsilon$, the player becomes more prudent, and less likely to be close to the cliff, which is dangerous for him. When $\epsilon$ is smaller, the player becomes venturesome and finally find the true optimal route. In my opinion, since the interception of the $\epsilon$ is the likelihood of "explore" the unexplored action, when the $\epsilon$ is bigger, the player becomes more dangerous if he is close to the cliff. As a result, he is willing to take the longer but safer way.

In the Q-learning, the player always find the optimal route no matter the value of the $\epsilon$. This is probably because Q-learning is a off-policy, and the Q value of the next state will not "explore" when updating the Q value of the current state, but find the greedy one, even though the Q-learning still need to "explore".

By the way, there seems no difference when $\epsilon$ is chosen to be 0. It is probably the consequence of that this question is simple. If the question becomes more complex, the result of $\epsilon = 0$ may fall into a local optimum.

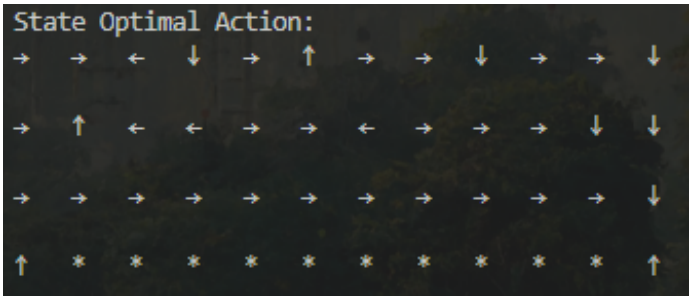


Figure 7: Optimal actions by Q-learning with $\epsilon = 0$