Problem 1

Q-Learning:

Set P=0.02,γ=0..95，α=0.3，ε=0.1 implementing 10 times independent Q-Learning, 10 path from start to goal has been obtained.The plot of optimal policy, optimal path and average accumulated reward with respect to episode is as follow:

```
path has been obtained
path has been obtained
path has been obtained
path has been obtained
path has been obtained
path has been obtained
path has been obtained
path has been obtained
path has been obtained
path has been obtained
number of successful runs = 10
```

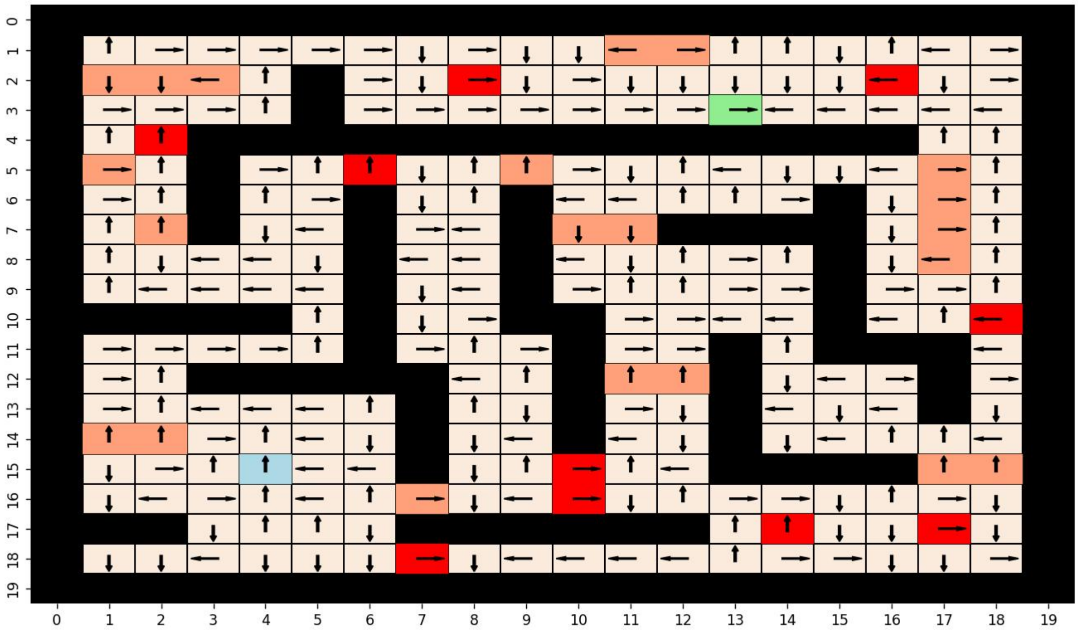

Figure 1.1 The optimal policy obtained by Q-Learning

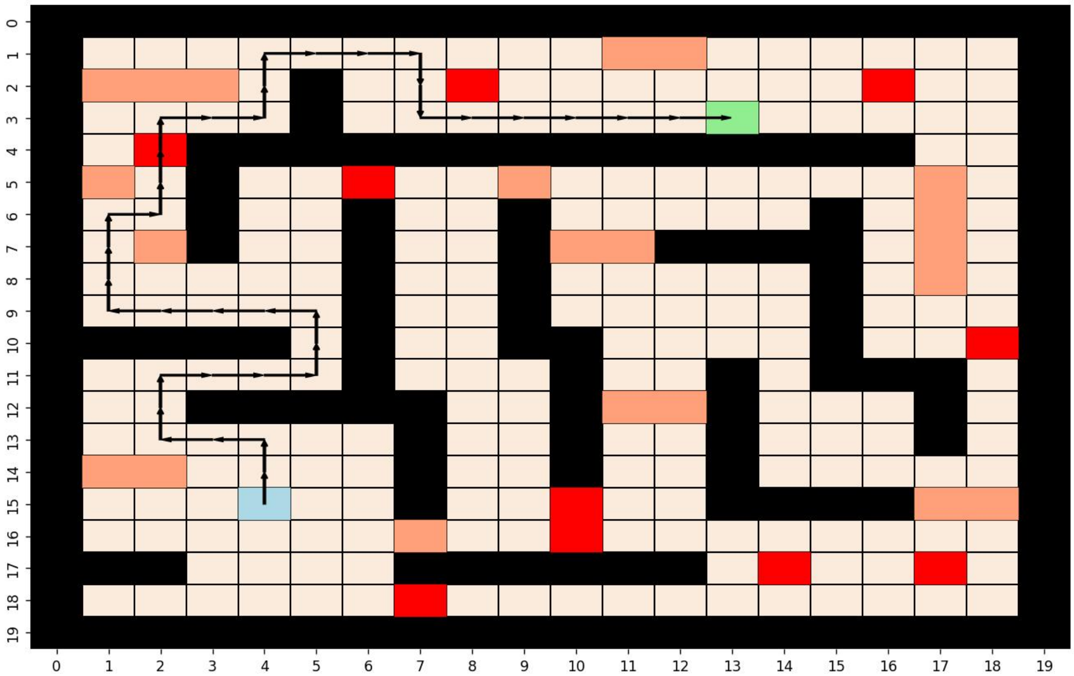

Figure 1.2 The optimal path obtained by Q-Learning

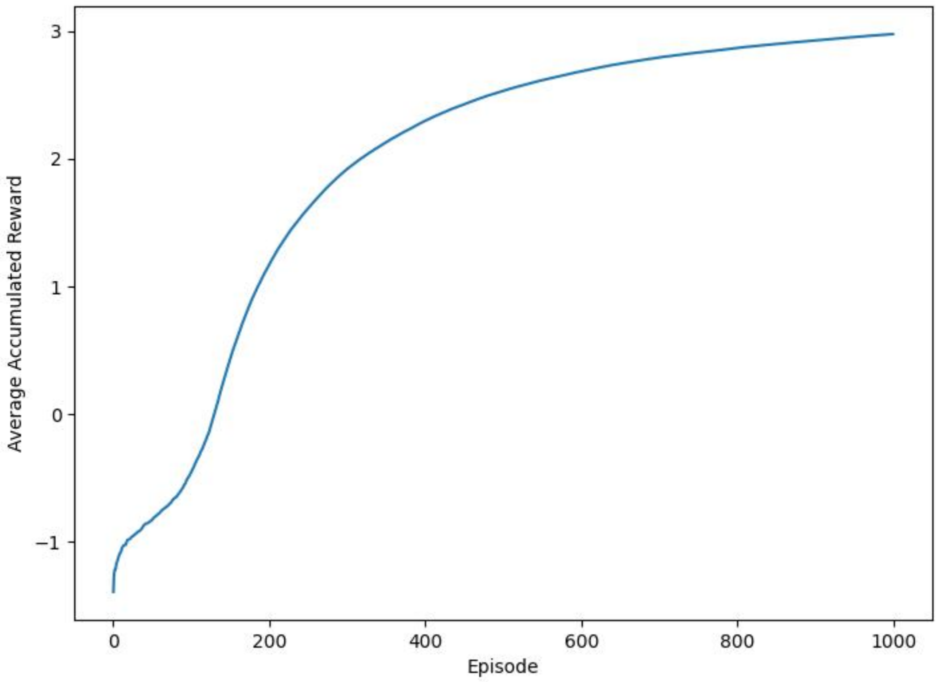

Figure 1.3 The average accumulated reward with respect to the episode obtained by Q-Learning

SARSA:

Set P=0.02,γ=0..95，α=0.3，ε=0.1 implementing 10 times independent SARSA, 7 path from start to goal has been obtained. The plot of optimal policy, optimal path and average accumulated reward with respect to episode is as follow:

```
path has been obtained
path has been obtained
path has been obtained
path has been obtained
path has been obtained
path has been obtained
path has been obtained
number of successful runs = 7
```

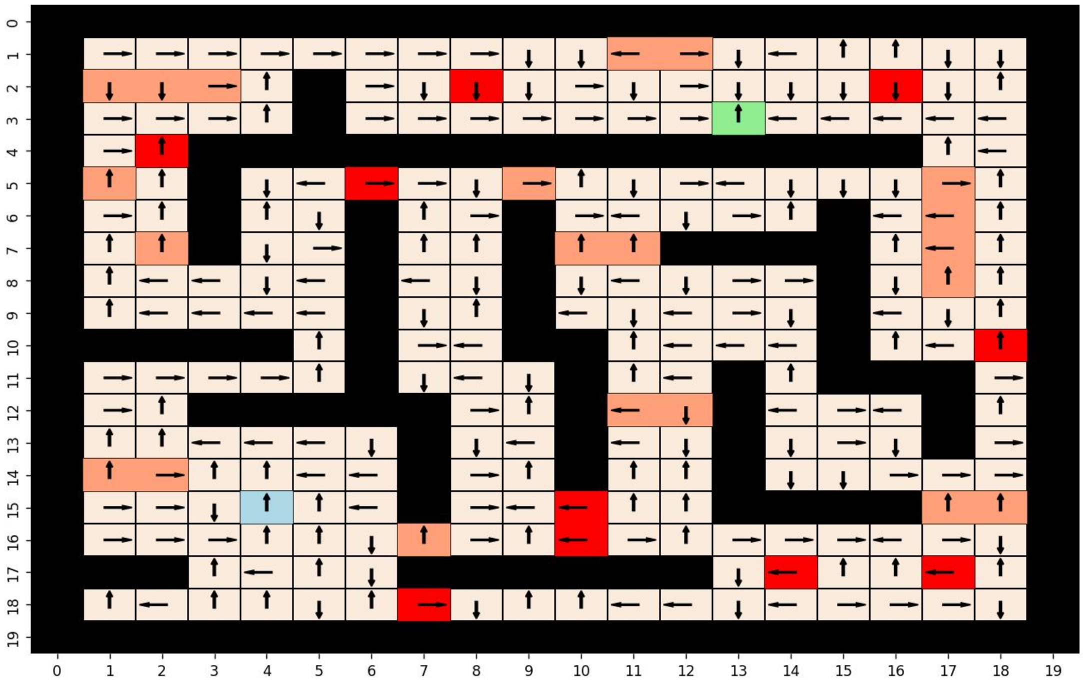

Figure 1.4 The optimal policy obtained SARSA

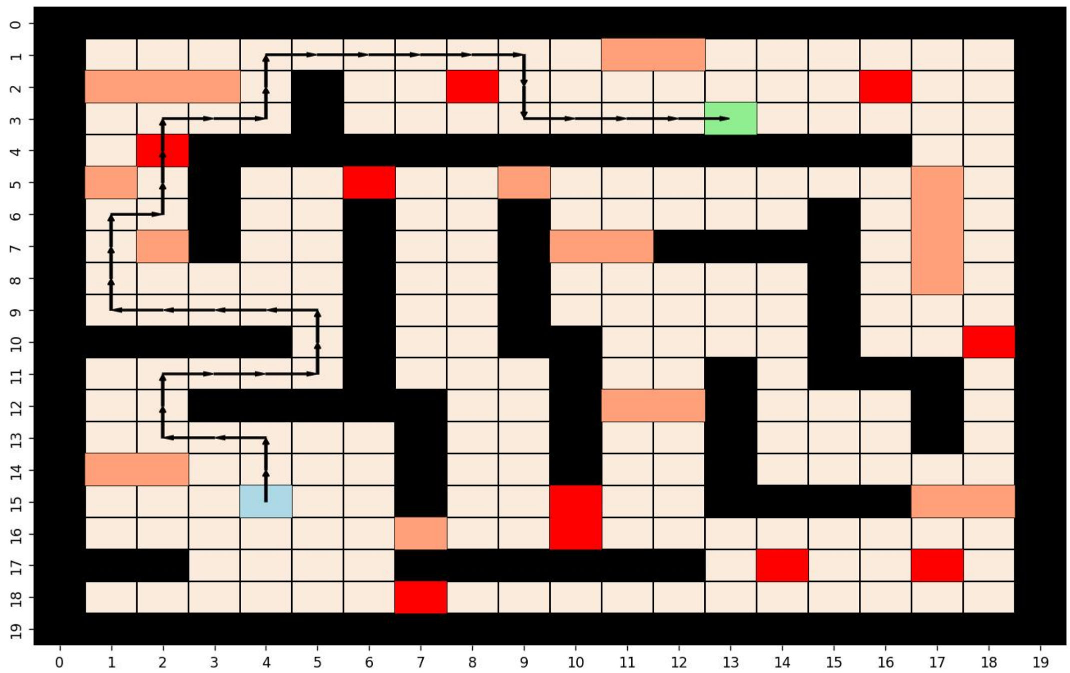

Figure 1.5 The optimal path obtained by SARSA

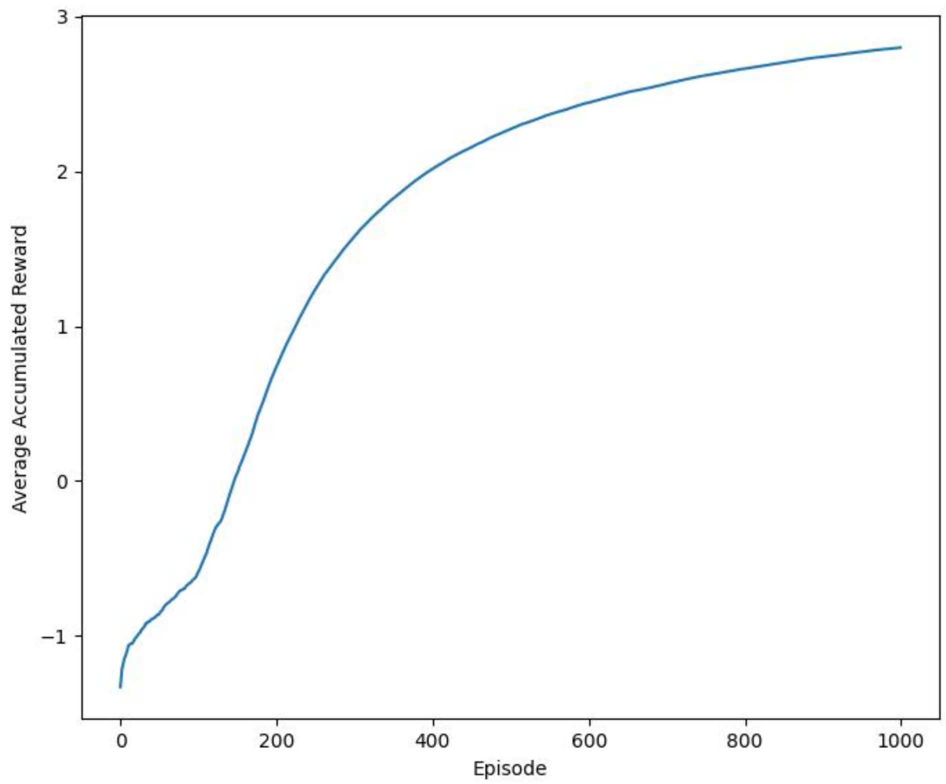

Figure 1.6 The average accumulated reward with respect to the episode obtained by SARSA

Actor-Critic:

Set P=0.02,γ=0..95，α=0.3，β=0.05 implementing 10 times independent Actor - Critic:, 1 path from start to goal has been obtained.The plot of optimal policy, optimal path and average accumulated reward with respect to episode is as follow:

```
path has been obtained
number of successful runs = 1
```

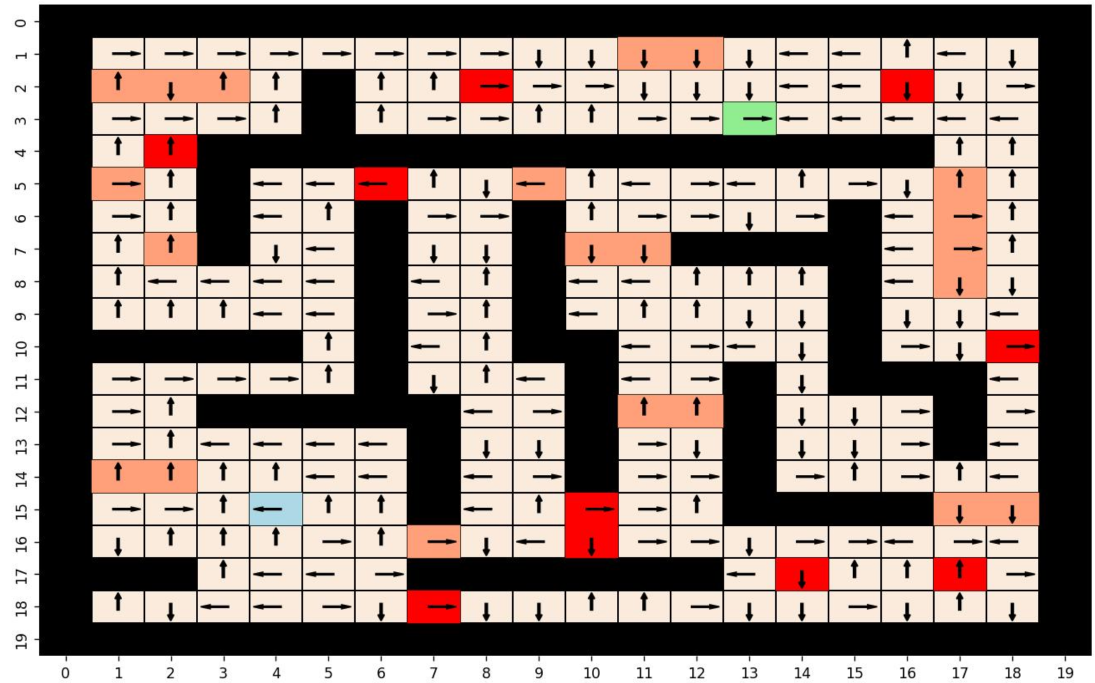

Figure 1.7 The optimal policy obtained Actor-Critic

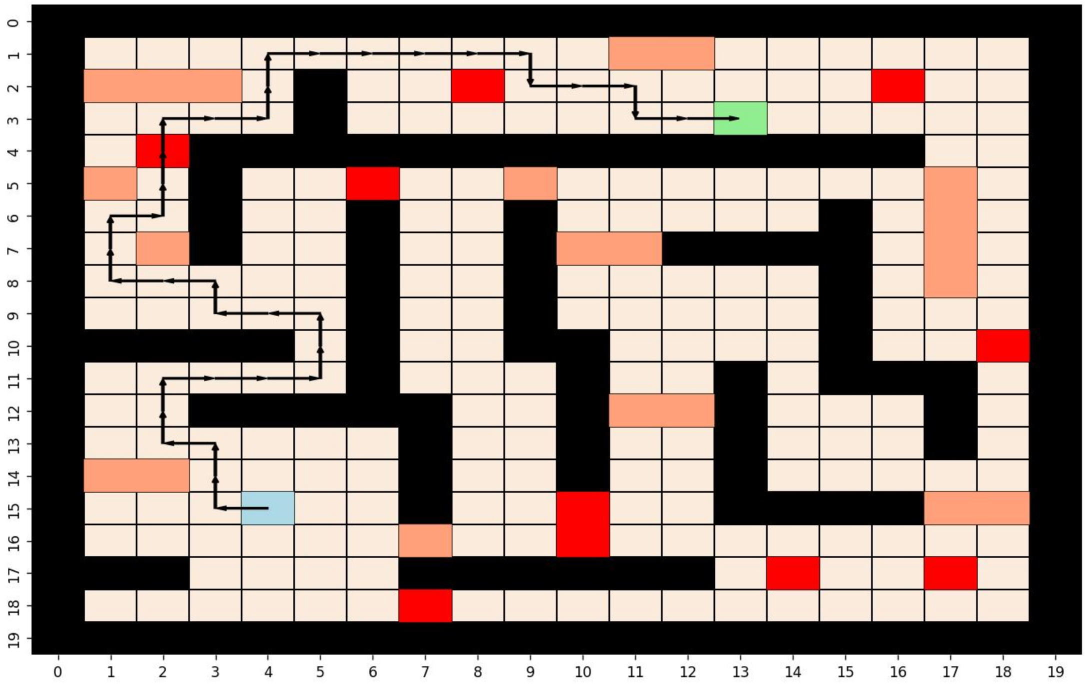

Figure 1.8 The optimal path obtained by Actor-Critic

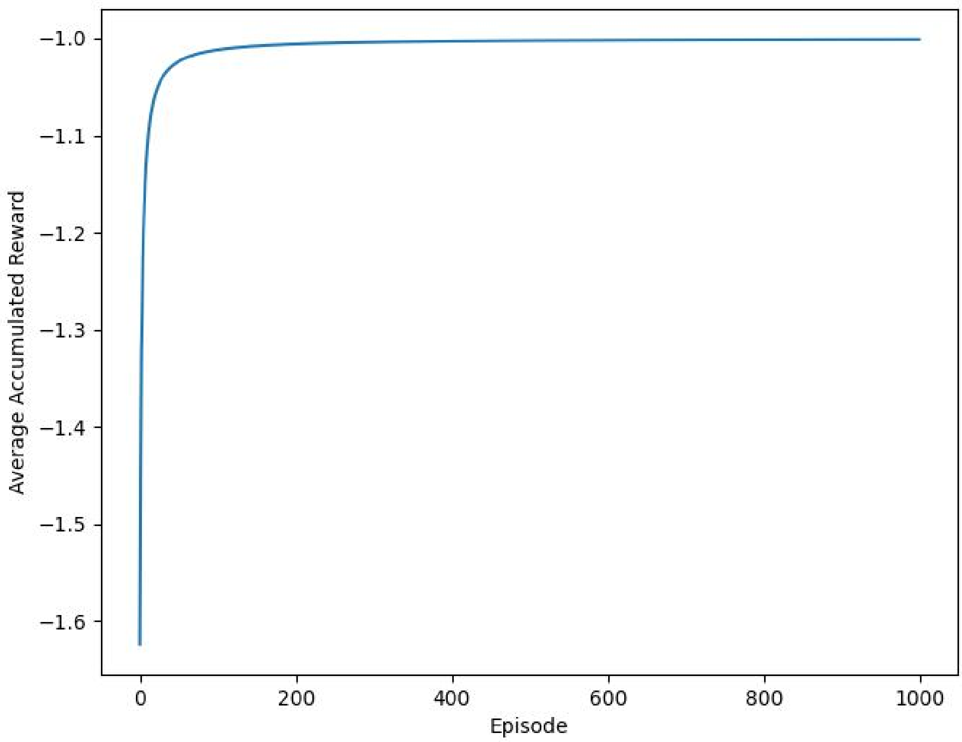

Figure 1.9 The average accumulated reward with respect to the episode obtained by Actor-Critic

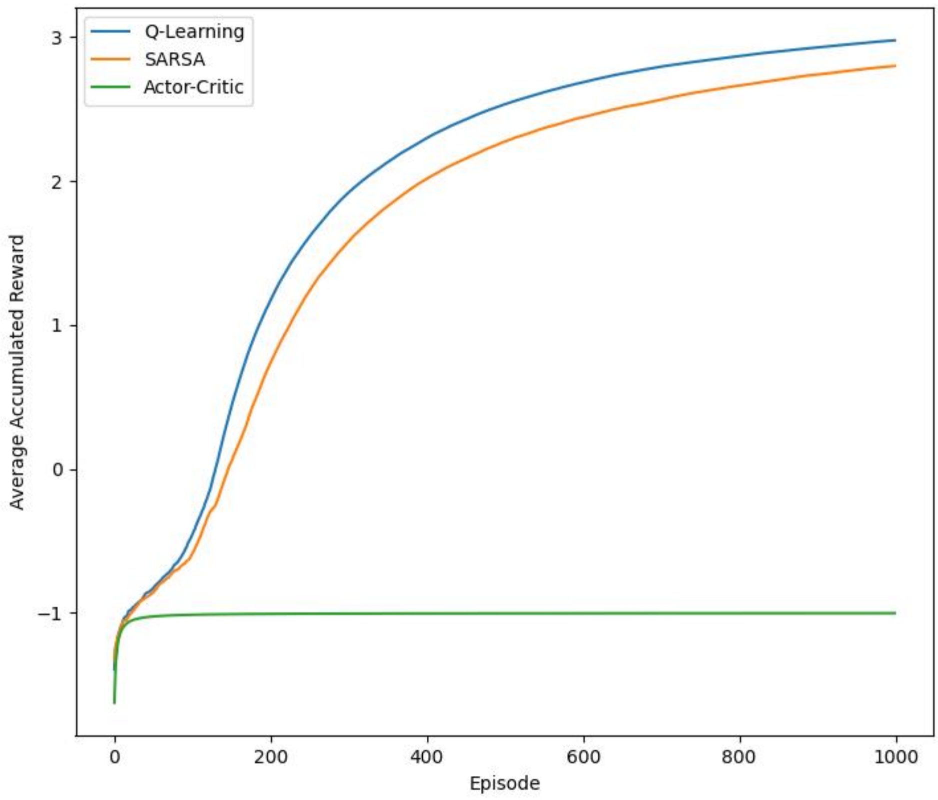

Figure 1.10 The average accumulated reward with respect to the episode obtained by Q-Learning , SARSA, Actor - Critic

According to the figure 1.10 we can conclude that because Q-learning is off-policy learning mechanism, which allows it to update its Q-values based on a policy that may be different from the one currently being followed, and Q-learning will always maximum the next state's q value which allows the Q-learning is expected to achieve a higher average accumulated reward than SARSA and Actor-Critic. Due to SARSA follows on-policy learning meaning that it updates its Q-values based on the policy it is currently following, so it will gain low accumulated reward and converge slowly than Q-Learning. Actor-Critic improve the policy and the value function simultaneously, so it converge faster than Q-Learning.

Problem 2

Q-Learning:

Set P=0.02,γ=0..95， α=0.2， ε=0.1 implementing 10 times independent Q-Learning. The optimal policy and average accumulated reward with respect to episode is as follow:

```
optimal policy for all 10 independent runs when implementing Q-Learning
['a2', 'a4', 'a2', 'a1', 'a2', 'a3', 'a2', 'a1', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2']
['a2', 'a1', 'a2', 'a2', 'a2', 'a3', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2']
['a2', 'a3', 'a2', 'a1', 'a2', 'a1', 'a2', 'a1', 'a2', 'a2', 'a2', 'a2', 'a2', 'a3', 'a2', 'a2']
['a2', 'a4', 'a2', 'a1', 'a2', 'a1', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2']
['a2', 'a3', 'a2', 'a3', 'a2', 'a3', 'a2', 'a4', 'a2', 'a1', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2']
['a2', 'a4', 'a2', 'a1', 'a2', 'a2', 'a2', 'a4', 'a2', 'a3', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2']
['a2', 'a4', 'a2', 'a2', 'a2', 'a3', 'a2', 'a1', 'a2', 'a1', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2']
['a2', 'a4', 'a2', 'a4', 'a2', 'a4', 'a2', 'a1', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2']
['a2', 'a4', 'a2', 'a1', 'a2', 'a4', 'a2', 'a2', 'a2', 'a3', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2']
['a2', 'a3', 'a2', 'a4', 'a2', 'a3', 'a2', 'a2', 'a2', 'a1', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2']
```

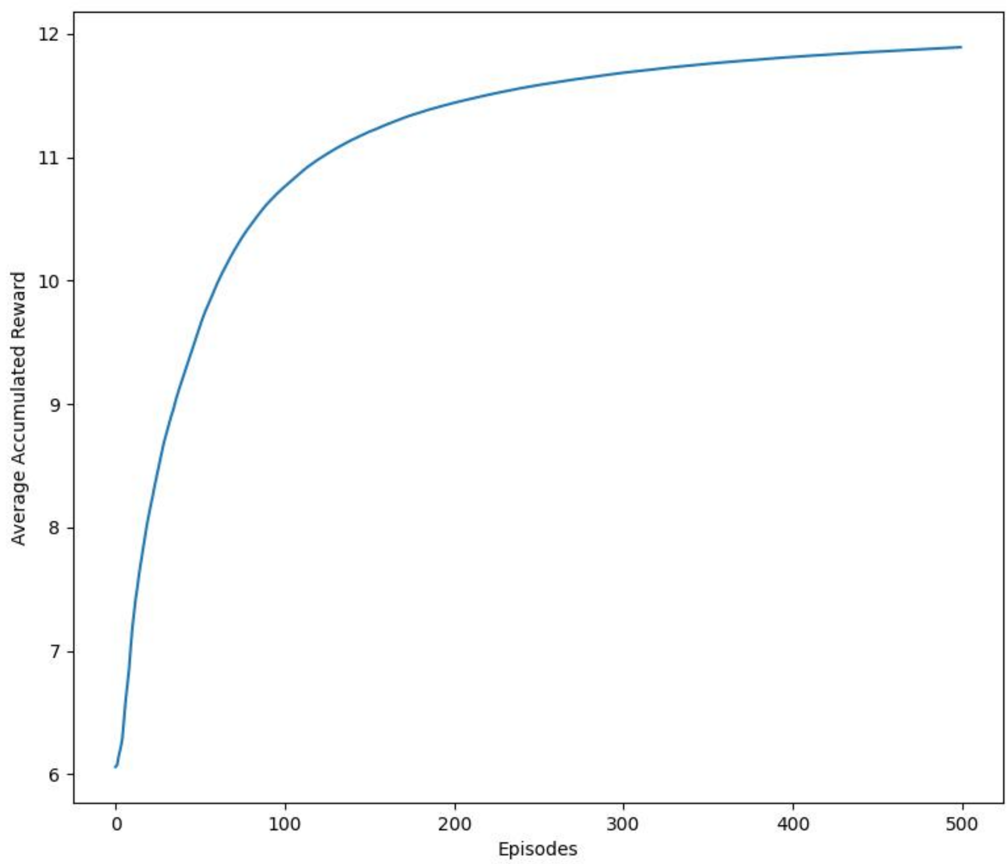

Figure 2.1 The average accumulated reward with respect to the episodes obtained by Q-Learning ,

SARSA:

Set P=0.05,γ=0..95, α=0.2, ε=0.1 implementing 10 times independent SARSA. The optimal policy and average accumulated reward with respect to episode is as follow:

```
optimal policy for all 10 independent runs when implementing SARSA
['a3', 'a4', 'a2', 'a1', 'a2', 'a2', 'a2', 'a2', 'a2', 'a3', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2']
['a2', 'a3', 'a2', 'a4', 'a2', 'a2', 'a2', 'a3', 'a2', 'a3', 'a2', 'a4', 'a2', 'a2', 'a2', 'a2']
['a2', 'a2', 'a2', 'a4', 'a2', 'a4', 'a2', 'a1', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2']
['a3', 'a4', 'a2', 'a1', 'a2', 'a2', 'a2', 'a2', 'a2', 'a1', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2']
['a2', 'a2', 'a2', 'a1', 'a2', 'a2', 'a2', 'a1', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2']
['a2', 'a1', 'a2', 'a1', 'a2', 'a3', 'a2', 'a2', 'a2', 'a3', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2']
['a2', 'a4', 'a2', 'a3', 'a2', 'a3', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2']
['a2', 'a3', 'a2', 'a1', 'a2', 'a1', 'a2', 'a2', 'a3', 'a4', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2']
['a2', 'a3', 'a2', 'a2', 'a2', 'a4', 'a2', 'a1', 'a2', 'a3', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2']
['a2', 'a3', 'a2', 'a1', 'a2', 'a2', 'a2', 'a2', 'a2', 'a4', 'a2', 'a2', 'a2', 'a3', 'a2', 'a2']
```

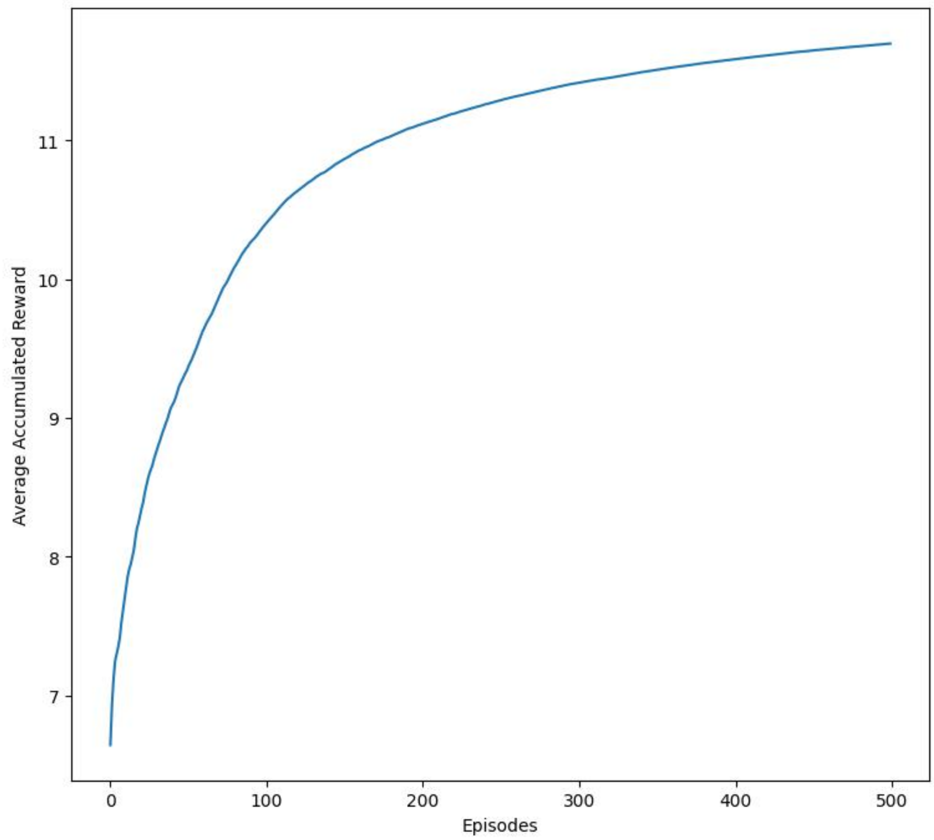

Figure 2.2 The average accumulated reward with respect to the episodes obtained by SARSA

SARSA-λ:

Set P=0.05,γ=0..95, α=0.2, ε=0.1 implementing 10 times independent SARSA-λ. The optimal policy and average accumulated reward with respect to episode is as follow:

```
optimal policy for all 10 independent runs when implementing SARSA-Lambda
['a2', 'a2', 'a2', 'a4', 'a2', 'a2', 'a2', 'a1', 'a3', 'a4', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2']
['a2', 'a3', 'a2', 'a1', 'a2', 'a2', 'a2', 'a1', 'a2', 'a4', 'a2', 'a4', 'a2', 'a2', 'a2', 'a2']
['a2', 'a1', 'a2', 'a4', 'a2', 'a4', 'a2', 'a1', 'a2', 'a4', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2']
['a3', 'a2', 'a2', 'a1', 'a2', 'a3', 'a2', 'a2', 'a2', 'a3', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2']
['a2', 'a1', 'a2', 'a4', 'a2', 'a2', 'a2', 'a2', 'a3', 'a3', 'a2', 'a4', 'a2', 'a2', 'a2', 'a2']
['a2', 'a2', 'a2', 'a1', 'a2', 'a2', 'a2', 'a1', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2']
['a2', 'a3', 'a2', 'a3', 'a2', 'a3', 'a2', 'a1', 'a2', 'a4', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2']
['a2', 'a1', 'a2', 'a1', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2']
['a3', 'a4', 'a2', 'a1', 'a2', 'a4', 'a2', 'a2', 'a2', 'a2', 'a2', 'a4', 'a2', 'a2', 'a2', 'a2']
['a2', 'a2', 'a2', 'a4', 'a2', 'a3', 'a2', 'a4', 'a2', 'a3', 'a2', 'a2', 'a2', 'a2', 'a2', 'a2']
```

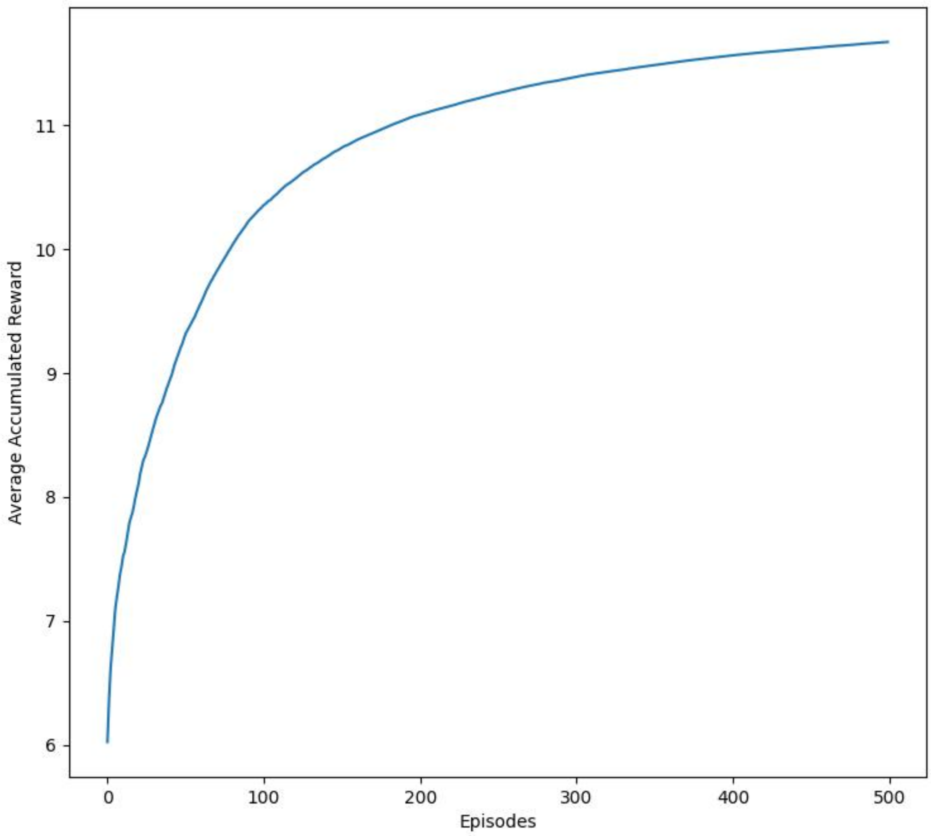

Figure 2.3 The average accumulated reward with respect to the episodes obtained by SARSA-λ

Actor-Critic:

Set P=0.05,γ=0..95, α=0.2, β=0.05 implementing 10 times independent Actor-Critic:, The optimal policy and average accumulated reward with respect to episode is as follow:

```
optimal policy for all 10 independent runs when implementing Actor_Critic
['a3', 'a4', 'a3', 'a4', 'a4', 'a1', 'a2', 'a2', 'a2', 'a3', 'a2', 'a1', 'a2', 'a2', 'a2', 'a2']
['a3', 'a2', 'a3', 'a3', 'a4', 'a2', 'a2', 'a2', 'a2', 'a4', 'a2', 'a2', 'a2', 'a1', 'a2', 'a2']
['a3', 'a2', 'a4', 'a2', 'a2', 'a4', 'a4', 'a1', 'a1', 'a2', 'a1', 'a3', 'a2', 'a2', 'a1', 'a2']
['a2', 'a3', 'a3', 'a4', 'a4', 'a2', 'a1', 'a4', 'a2', 'a3', 'a2', 'a3', 'a2', 'a1', 'a2', 'a1']
['a2', 'a1', 'a3', 'a4', 'a2', 'a2', 'a4', 'a2', 'a2', 'a3', 'a3', 'a3', 'a2', 'a1', 'a1', 'a1']
['a3', 'a3', 'a3', 'a3', 'a4', 'a2', 'a2', 'a2', 'a2', 'a4', 'a2', 'a3', 'a2', 'a1', 'a2', 'a2']
['a3', 'a1', 'a4', 'a2', 'a2', 'a3', 'a4', 'a2', 'a2', 'a2', 'a3', 'a2', 'a2', 'a2', 'a2', 'a2']
['a4', 'a2', 'a3', 'a4', 'a2', 'a4', 'a2', 'a2', 'a2', 'a4', 'a4', 'a4', 'a2', 'a2', 'a2', 'a2']
['a4', 'a1', 'a3', 'a3', 'a2', 'a4', 'a1', 'a1', 'a2', 'a1', 'a3', 'a3', 'a2', 'a1', 'a2', 'a1']
['a3', 'a3', 'a4', 'a3', 'a2', 'a4', 'a4', 'a1', 'a2', 'a2', 'a3', 'a3', 'a1', 'a2', 'a1', 'a1']
```

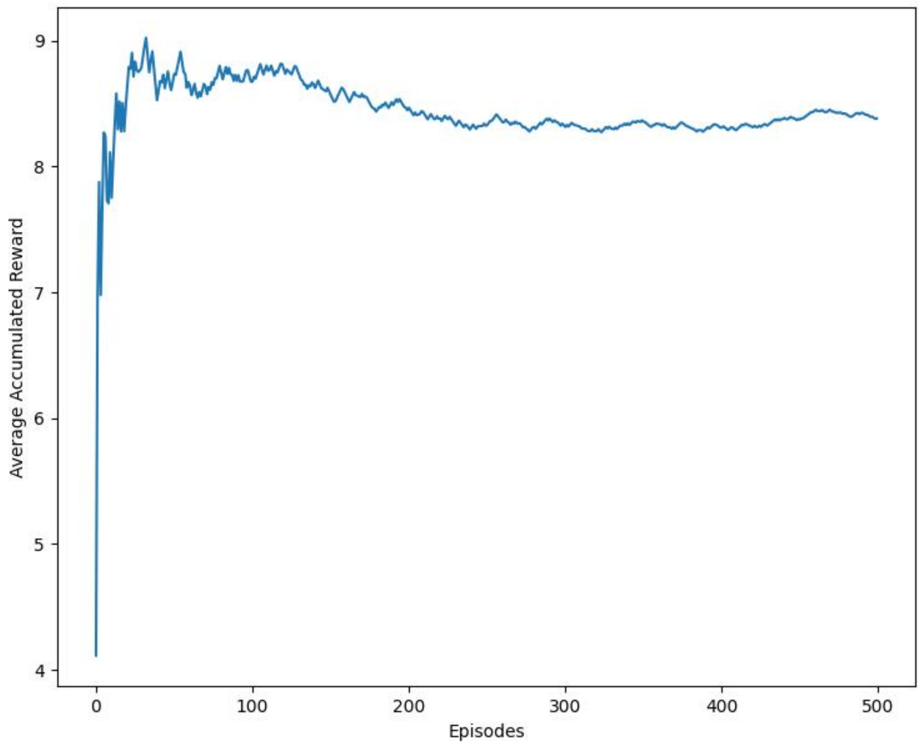

Figure 2.4 The average accumulated reward with respect to the episode obtained by Actor - Critic

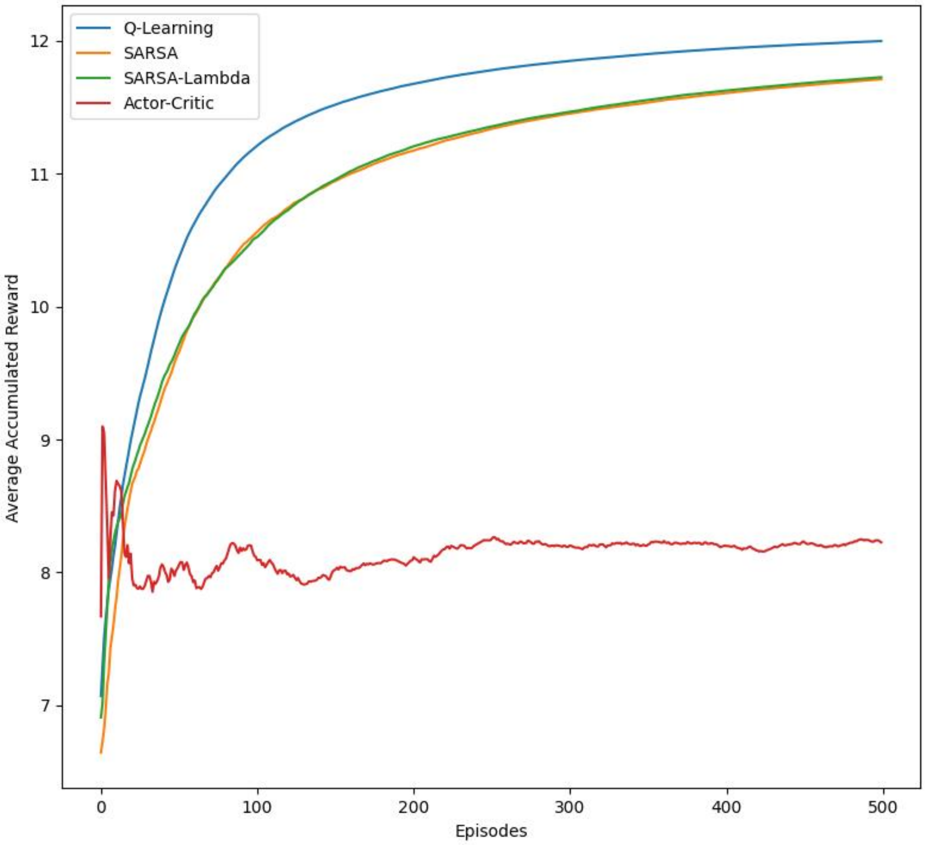

Figure 2.5 The average accumulated reward with respect to the episode obtained by Q-Learning , SARSA, SARSA-λ, Actor - Critic

Based on figure 2.5, Actor-Critic converge faster than other learning algorithms because it can balance exploration and exploitation more effectively but we can also see it also brings lower accumulated reward. Q-learning can explore suboptimal policies during training, leading to better exploration and a higher probability of finding the optimal policy. So Moreover, Q-learning is less prone to being stuck in a suboptimal policy compared to SARSA, SARSA-λ, Actor-Critic. SARSA-λ is a more sophisticated algorithm than SARSA that can provide faster convergence and a better bias-variance trade-off. However, it also has additional hyperparameters to tune so it may be more sensitive to their values.