# Project 3 Report

## Problem 1

### a) Path Attained

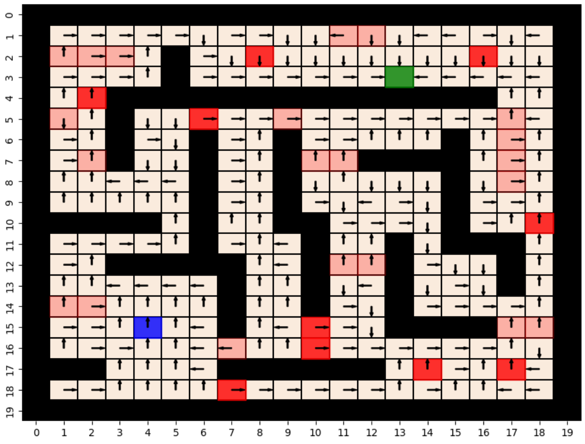
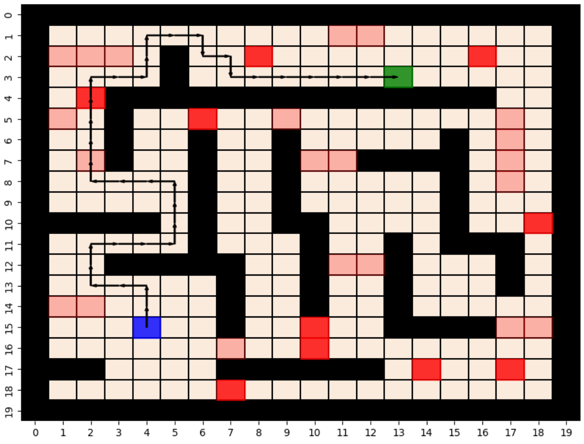
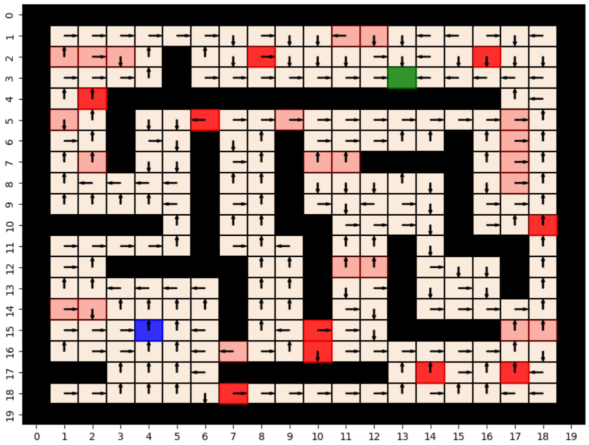
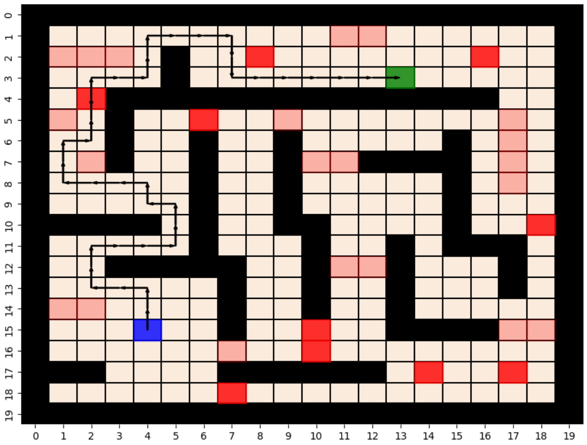
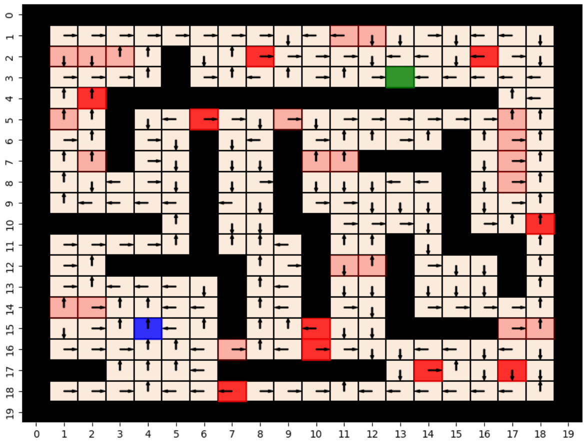
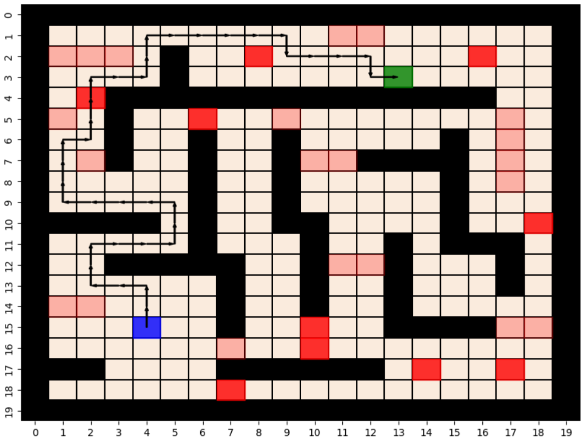
| | Paths Obtained |
|---|---|
| Q Learning | 10 |
| SARSA | 10 |
| Actor-Critic (β=0.3) | 7 |

For the Actor-Critic method, modification in parameters was needed to consistently create a policy that produces an optimal path to goal. A sweeping function was made to sweep β from 0 to 1 to find the β value that produces the most paths out of 10 runs.

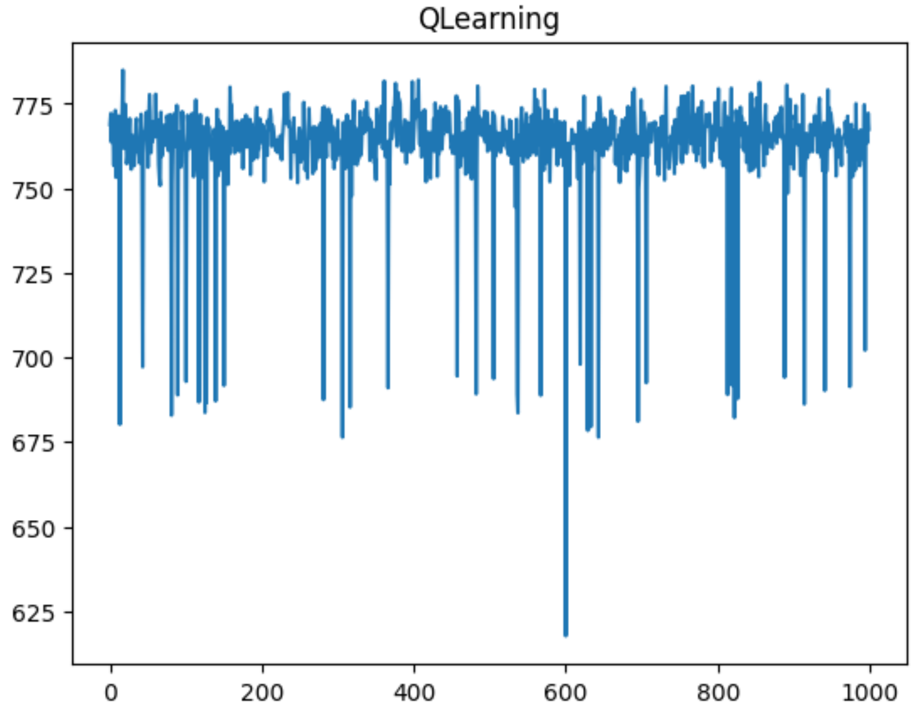
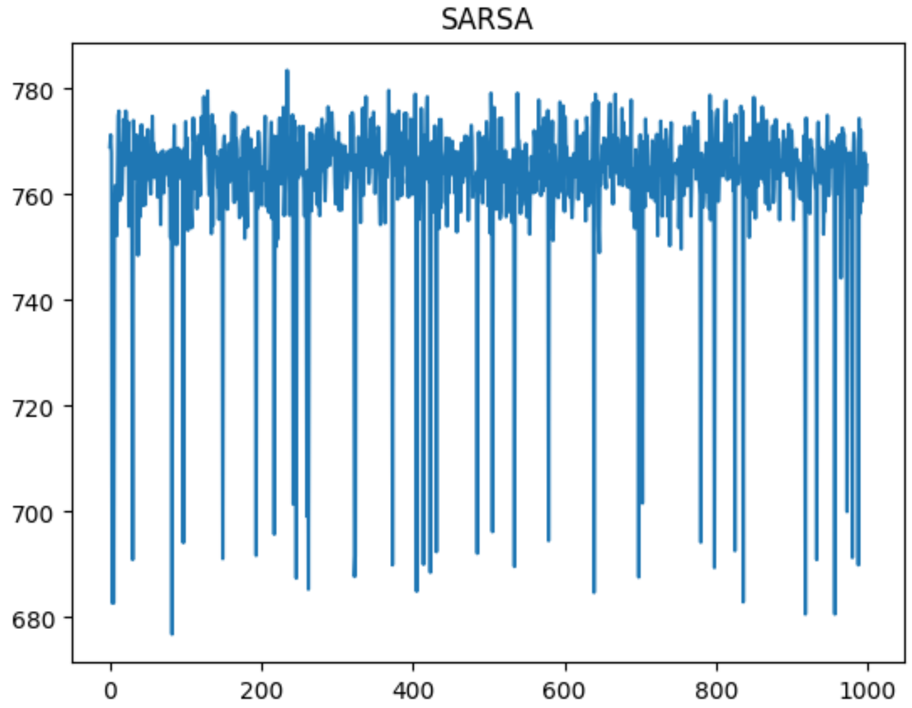
The final β value was set to 0.3, as this performed the best.

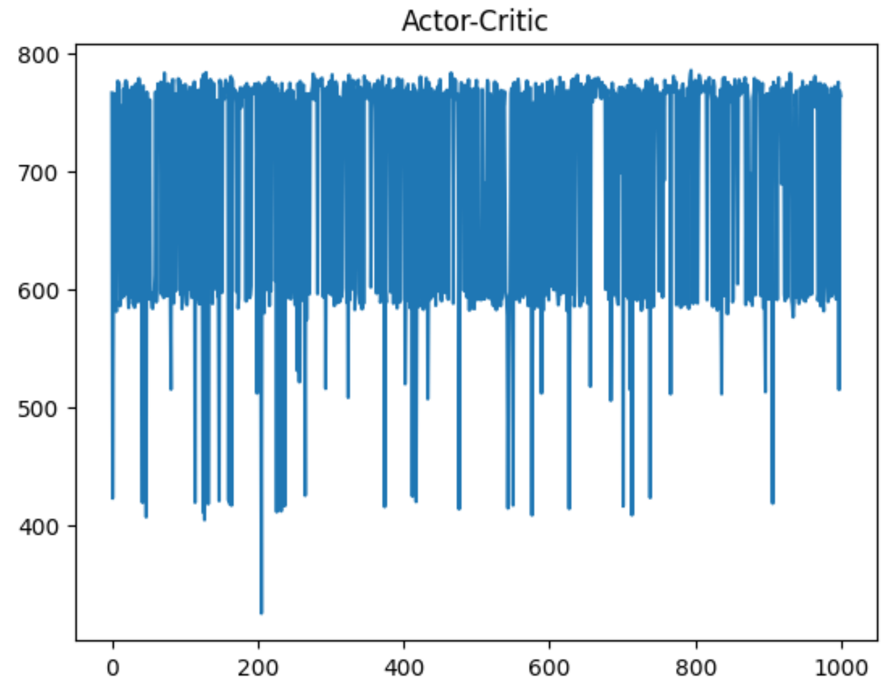
$$\gamma = 0.95, \qquad \alpha = 0.3, \qquad \epsilon = 0.1, \qquad \beta = 0.3$$

### b) Optimal Policy + Optimal Path

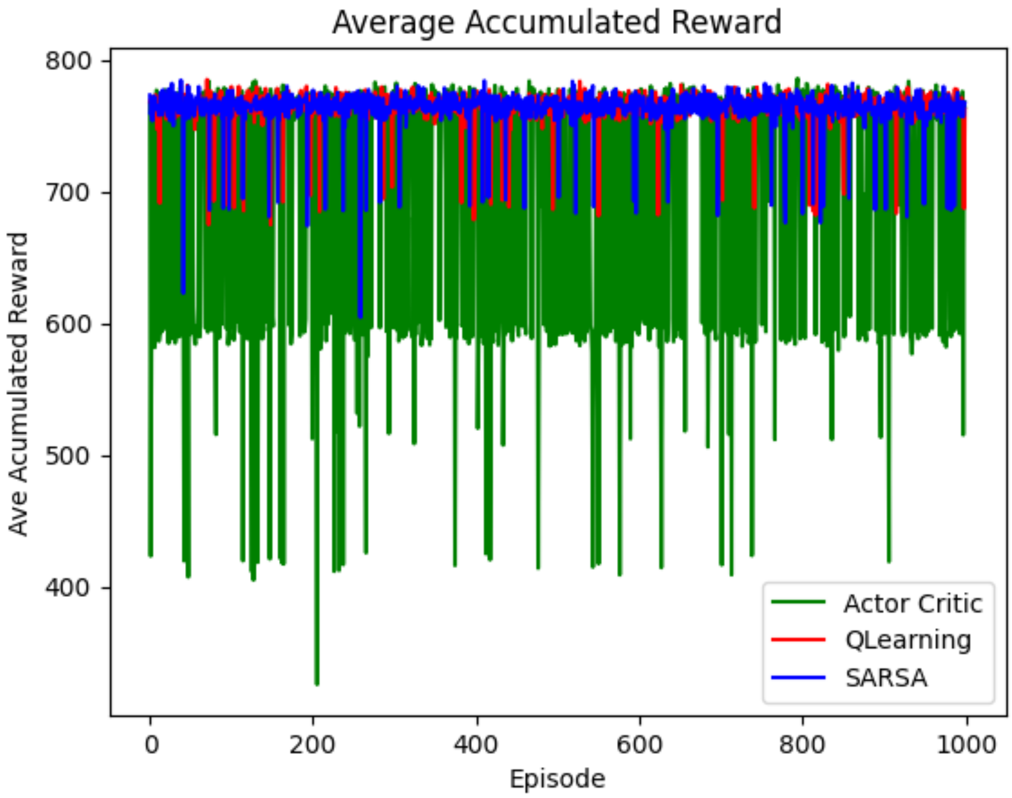
| | Optimal Policy | Optimal Path |
|---|---|---|
| Q Learning |  |  |
| SARSA |  |  |
| Actor Critic (β=0.3) |  |  |

### c) Average Accumulated Reward

| | Average Accumulated Reward |
|---|---|
| Q Learning |  |
| SARSA |  |

| | |
|---|---|
| Actor Critic (β=0.3) |  |

**d) All Accumulated Reward Comparison**



**Discussion:**

The results for Q-Learning and SARSA are very close, with the SARSA with a bit more variation. This is due to the SARSA target being less greedy than Q-Learning, giving it a bit more exploration.

The results for Actor-Critic varies greatly as it does not always find a path to goal and is much more inconsistent.

All algorithms converged quite quickly, it is difficult to see which one converged faster.

## Problem 2

### a) Optimal Policy

$$Let\ a^1 = 1, a^2 = 2, a^3 = 3, a^4 = 4$$

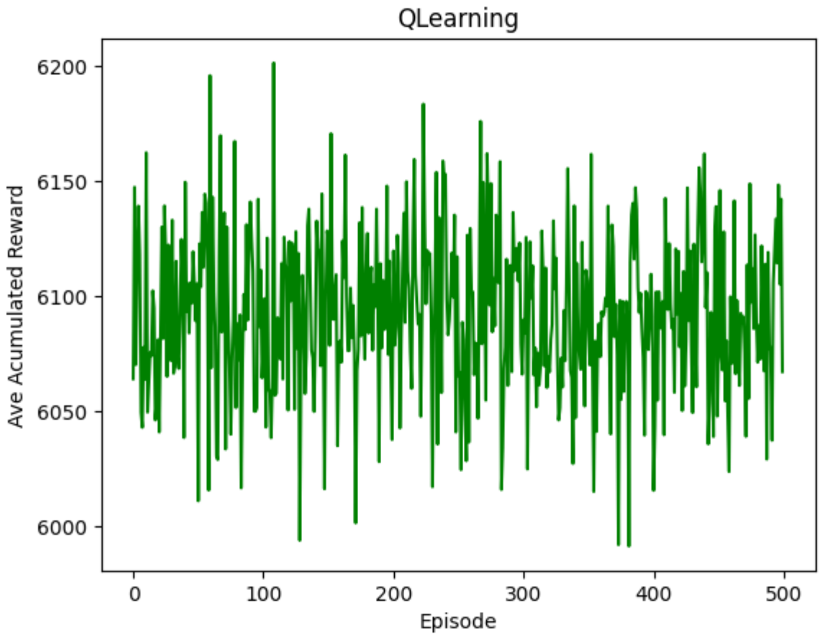
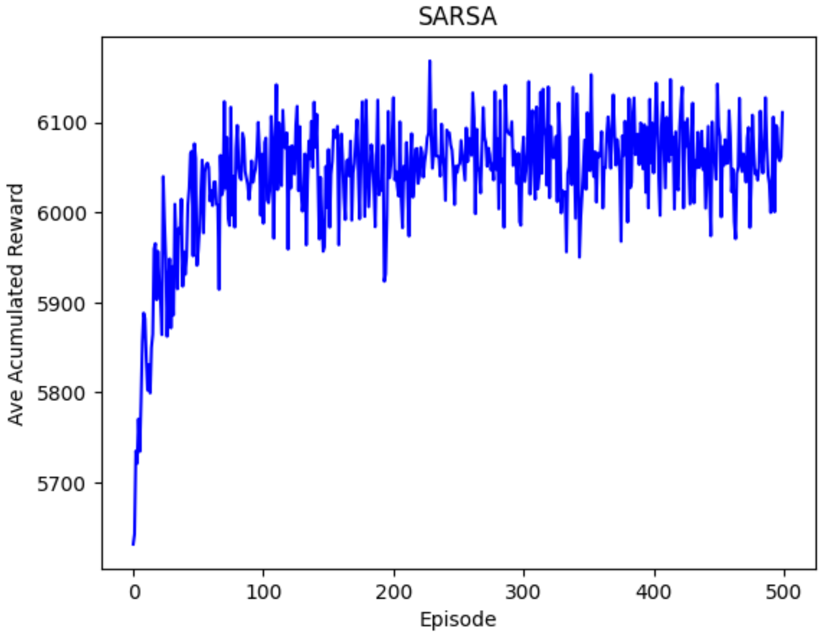
| Run | Q-Learning Action |
|---|---|
| 1 | [2 2 2 2 2 2 2 2 3 2 2 2 4 2 2 2] |
| 2 | [2 2 2 2 2 2 2 2 3 2 2 2 4 2 2 2] |
| 3 | [2 2 2 2 2 2 2 2 3 2 2 2 4 2 2 2] |
| 4 | [2 2 2 2 2 2 2 2 3 2 2 2 1 2 2 2] |
| 5 | [2 2 2 2 2 2 2 2 1 2 2 2 4 2 2 2] |
| 6 | [2 2 2 2 2 2 2 2 3 2 2 2 4 2 2 2] |
| 7 | [2 2 2 2 2 2 2 2 3 2 2 2 4 2 2 2] |
| 8 | [2 2 2 2 2 2 2 2 3 2 2 2 1 2 2 2] |
| 9 | [2 2 2 2 2 2 2 2 3 2 2 2 4 2 2 2] |
| 10 | [2 2 2 2 2 2 2 2 4 2 2 2 4 2 2 2] |

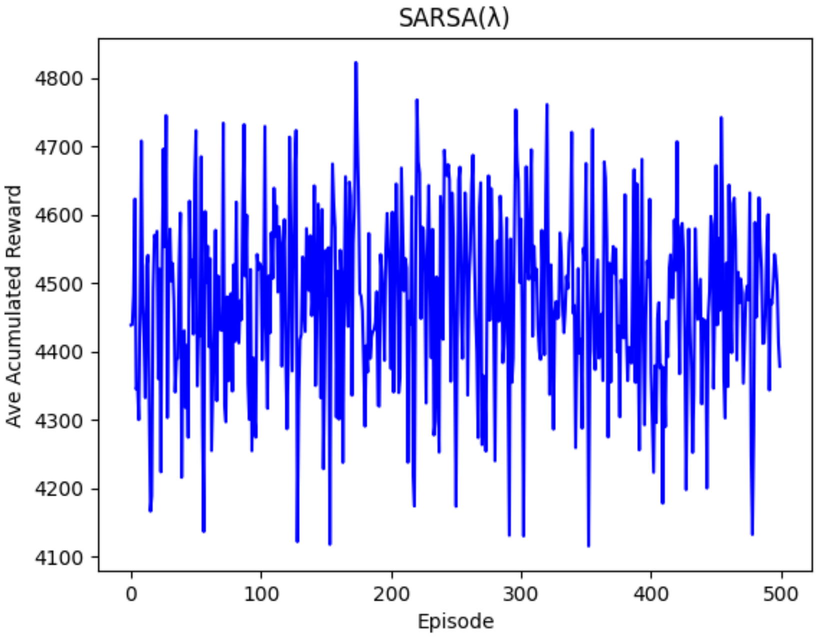
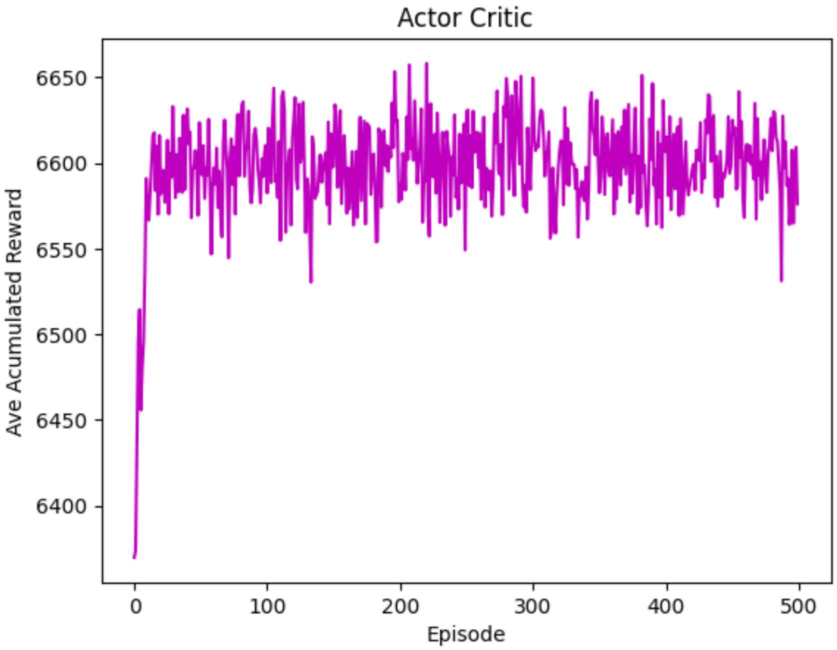
| Run | SARSA Action |
|---|---|
| 1 | [2 2 2 2 2 2 2 2 1 1 1 2 1 2 1 2] |
| 2 | [2 2 2 2 2 2 2 2 1 2 2 3 1 2 2 2] |
| 3 | [2 2 2 2 2 2 2 2 1 2 2 2 1 2 2 2] |
| 4 | [2 2 2 2 2 2 2 2 1 2 2 2 1 2 2 2] |
| 5 | [2 2 2 2 2 2 2 2 3 2 2 2 1 2 2 2] |
| 6 | [2 2 2 2 2 2 2 2 4 2 2 2 4 2 2 2] |
| 7 | [2 3 2 2 2 2 2 2 3 2 2 2 4 2 2 2] |
| 8 | [2 2 2 2 2 2 2 2 1 2 2 2 1 2 2 2] |
| 9 | [2 2 2 2 2 2 2 2 3 2 2 2 4 2 2 2] |
| 10 | [2 2 2 2 2 2 2 2 3 2 2 2 1 2 2 2] |

| Run | SARSA (λ=0.95) Action |
|---|---|
| 1 | [3 2 2 2 2 1 1 1 3 4 3 3 1 4 2 2] |
| 2 | [4 3 2 2 2 2 3 1 1 2 2 4 4 2 2 4] |
| 3 | [3 2 2 2 1 2 2 1 3 4 2 2 1 4 2 2] |
| 4 | [4 2 2 2 2 2 1 1 2 2 3 3 4 4 2 2] |
| 5 | [2 3 3 4 2 1 1 2 3 2 3 2 1 4 2 2] |
| 6 | [3 1 2 4 1 2 2 4 3 2 3 2 1 4 2 3] |
| 7 | [3 2 2 1 2 1 2 1 3 3 3 3 3 1 1 2] |
| 8 | [2 2 2 2 2 2 2 1 3 2 2 2 1 2 4 1] |
| 9 | [3 3 3 2 1 2 1 1 3 2 1 3 1 1 2 2] |
| 10 | [2 2 2 2 2 2 1 1 4 2 2 2 4 2 2 3] |

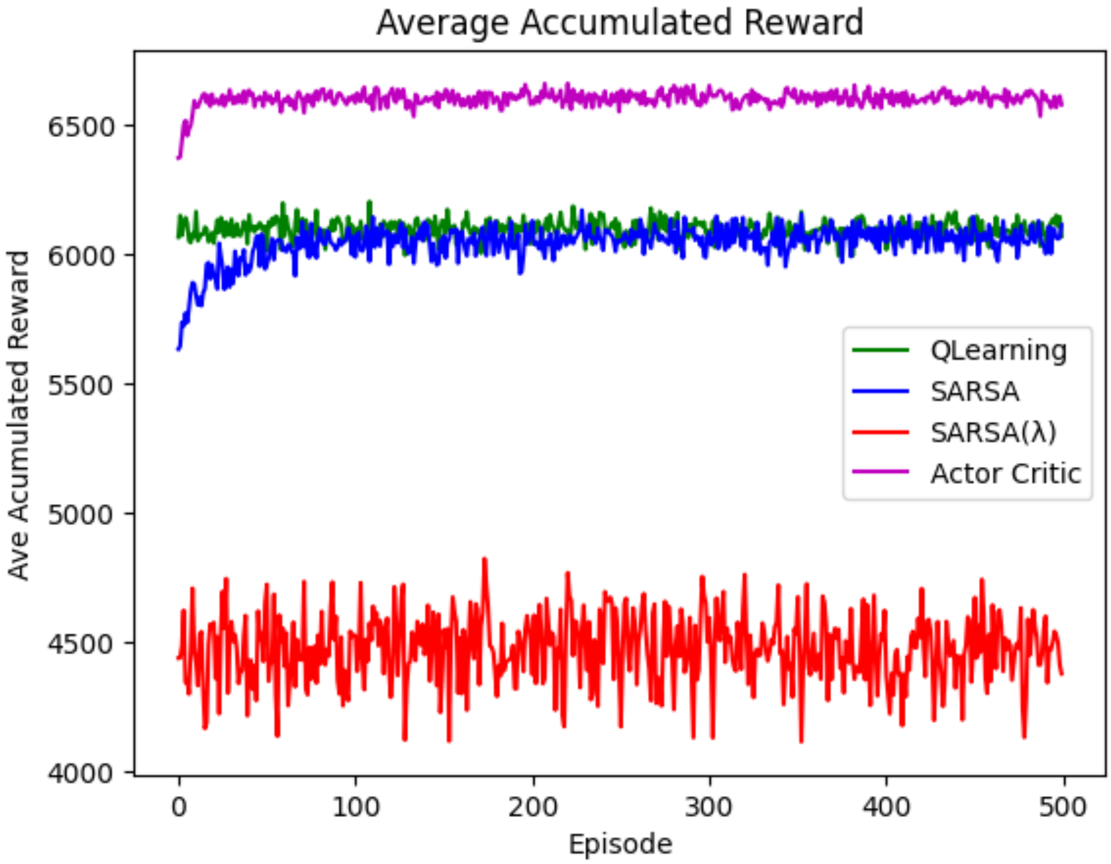
| **Run** | **Actor-Critic (β=0.05) (Action)** |
|---|---|
| 1 | [2 2 2 2 2 2 2 2 4 2 3 2 1 2 3 2] |
| 2 | [2 2 2 2 2 2 2 2 4 2 3 2 1 2 2 2] |
| 3 | [2 2 2 2 2 2 2 2 4 2 3 2 1 2 2 2] |
| 4 | [2 2 2 2 2 2 2 2 4 2 3 2 1 2 2 2] |
| 5 | [2 2 2 2 2 2 2 2 4 2 3 2 1 2 2 2] |
| 6 | [2 2 2 2 2 2 2 2 4 2 3 2 1 2 2 2] |
| 7 | [2 2 2 2 2 2 2 2 4 2 3 2 1 2 2 2] |
| 8 | [2 2 2 2 2 2 2 2 4 2 3 2 1 2 2 2] |
| 9 | [2 2 2 2 2 2 2 2 4 2 3 2 1 2 2 2] |
| 10 | [2 2 2 2 2 2 2 2 4 2 3 2 1 2 2 2] |

## b) Average Accumulated Reward

| | Average Accumulated Reward |
|---|---|
| Q-Learning |  |
| SARSA |  |

| | |
|---|---|
| SARSA (λ=0.95) |  |
| Actor-Critic<br>(β=0.05) |  |

### c) Comparison of Average Accumulated Reward



The Q-Learning and SARSA results are quite similar. The Q-Learning had less variation and converged much faster than SARSA. SARSA had more exploration and ended with a similar average.

The SARSA lambda algorithm had slower convergence, much more variation in reward, and a poorer performance compared to the others. This may be because all the Q values for actions that are not chosen are downsized and discouraged, which would cause a lower accumulated reward.

The Actor Critic algorithm had a convergence similar to that of SARSA, though slightly faster. The converged accumulated reward was slightly higher. This may be related to the high alpha values favoring the target delta calculated by the critic.