Reinforcement Learning Course Project

Partha Sarathi Mohapatra
EE18D703

13th March 2020

Q-learning and SARSA

1 Q-learning

All three variants of the problem are solved using Q-learning for 1000 episodes and the average was taken over 50 independent runs. We get the following observations:

1.1 (a): Learning curves

For plotting the average reward and average steps to reach the terminal state per episodes, we used three learning rates ($\alpha = 0.1, 0.5, 0.9$).

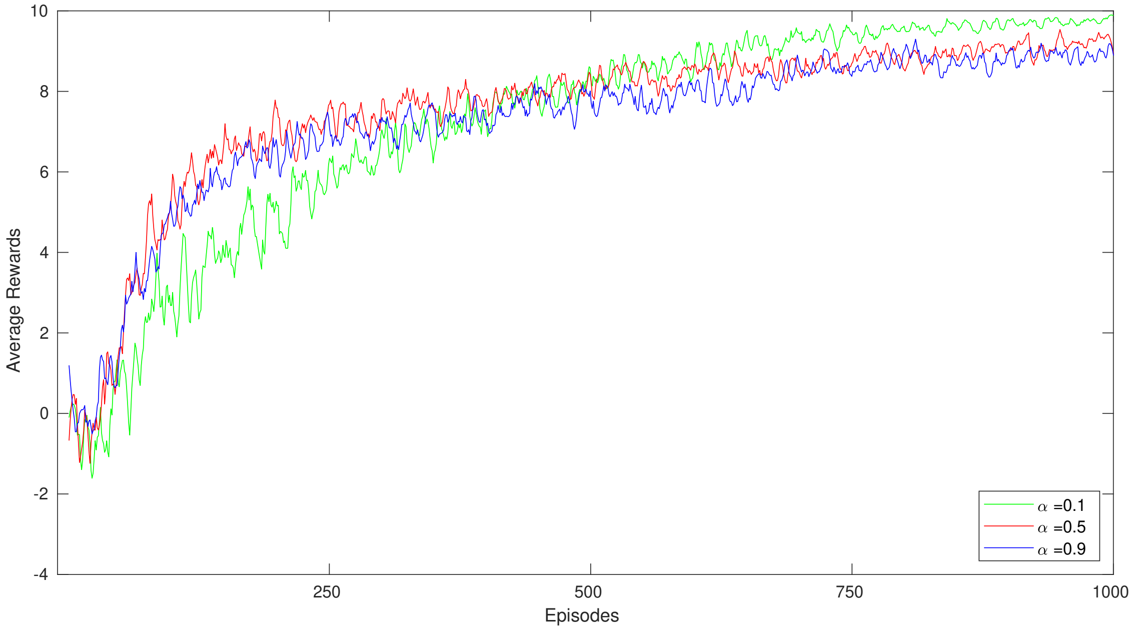

Figure 1: Average rewards per episodes in Q-learning for different α for terminal state A

(i) For $\alpha = 0.9$ the increase in the reward in initial episodes are high as compared to $\alpha = 0.1$, but after many episodes it get stuck in sub-optimal solution and the reward is around 9. On the other hand for $\alpha = 0.1$ during initial episodes though the rewards are less, the agent learns the optimal policy gradually and the rewards are very close to the optimal value of 10.

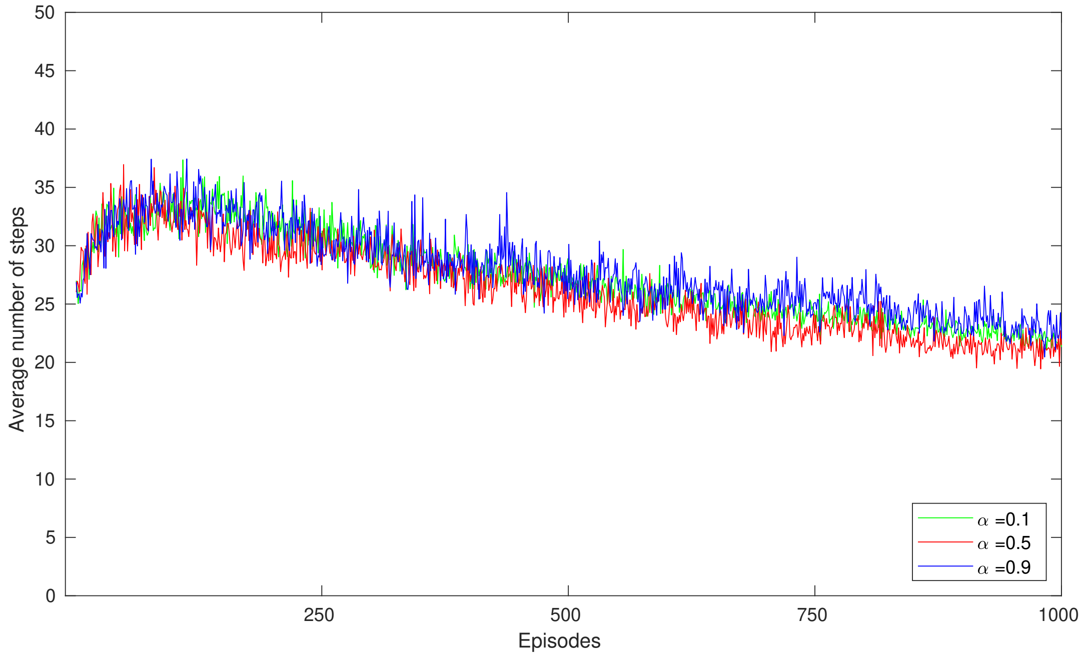

Figure 2: Average number of steps per episodes in Q-learning for different α for terminal A

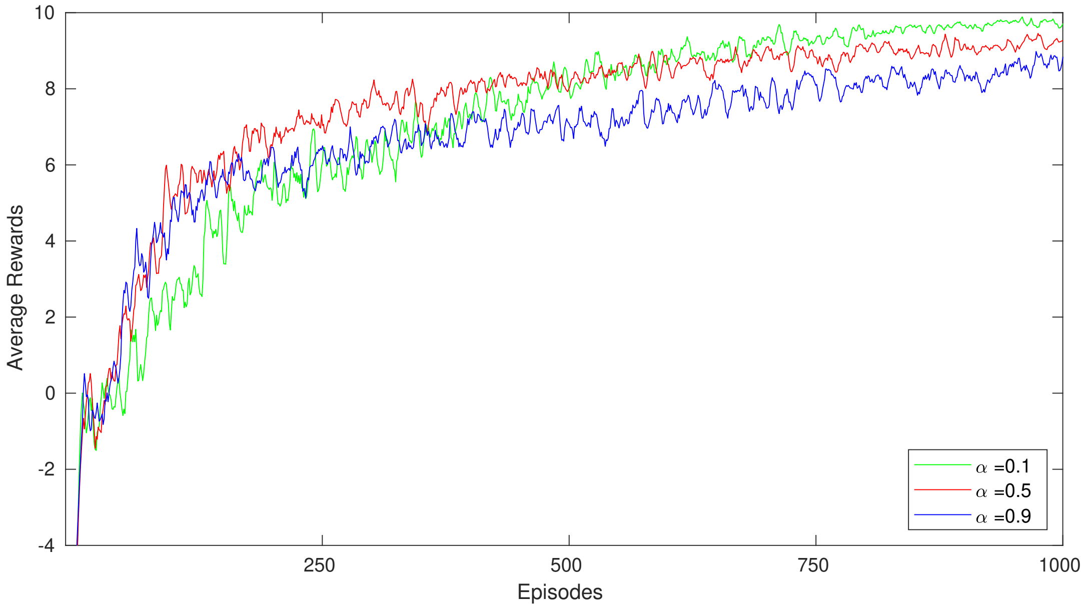

Figure 3: Average rewards per episodes in Q-learning for different α for terminal state B

(ii) The above observations in (i) are same for all the terminal states A, B and C. In case of C the optimal reward is less than 10 as the agent has to step in the puddle having negative reward of -1 atleast once to get to the terminal state.

(iii) The average number of steps to reach the terminal position decreases with increase in episodes.

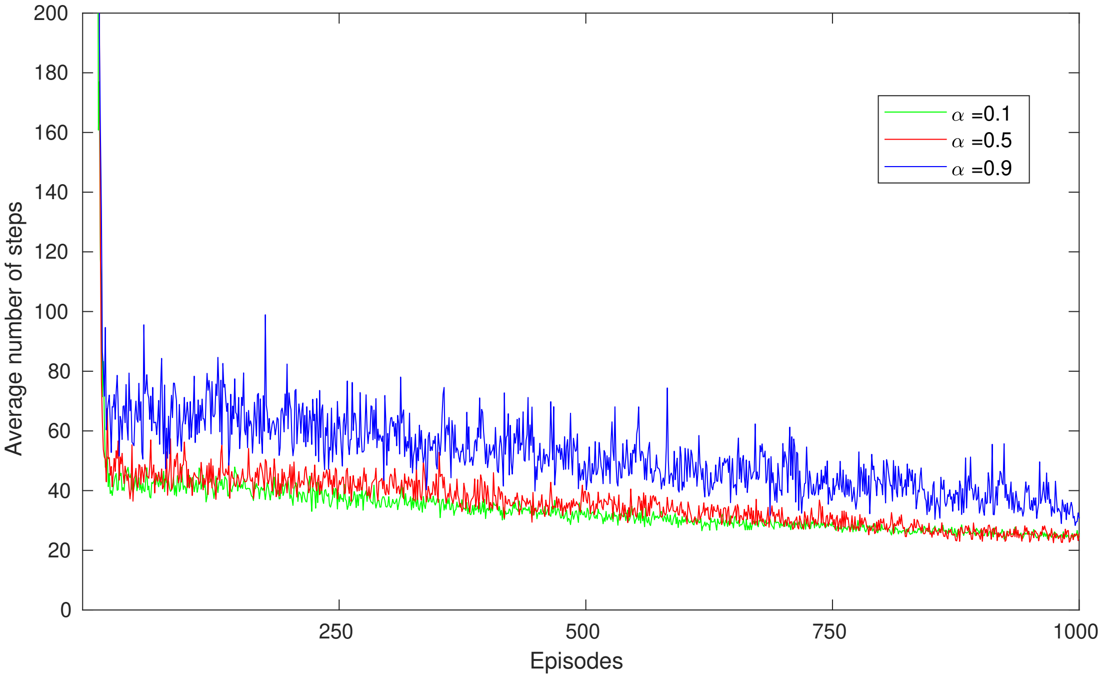

Figure 4: Average number of steps per episodes in Q-learning for different α for terminal B

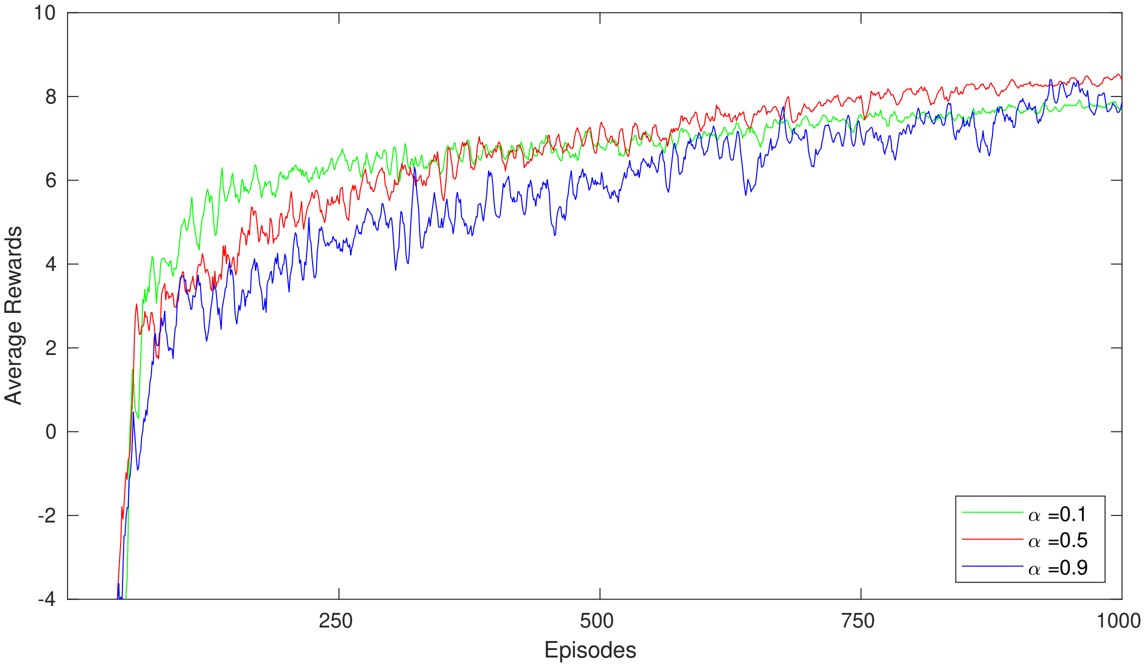

Figure 5: Average rewards per episodes in Q-learning for different α for terminal state C

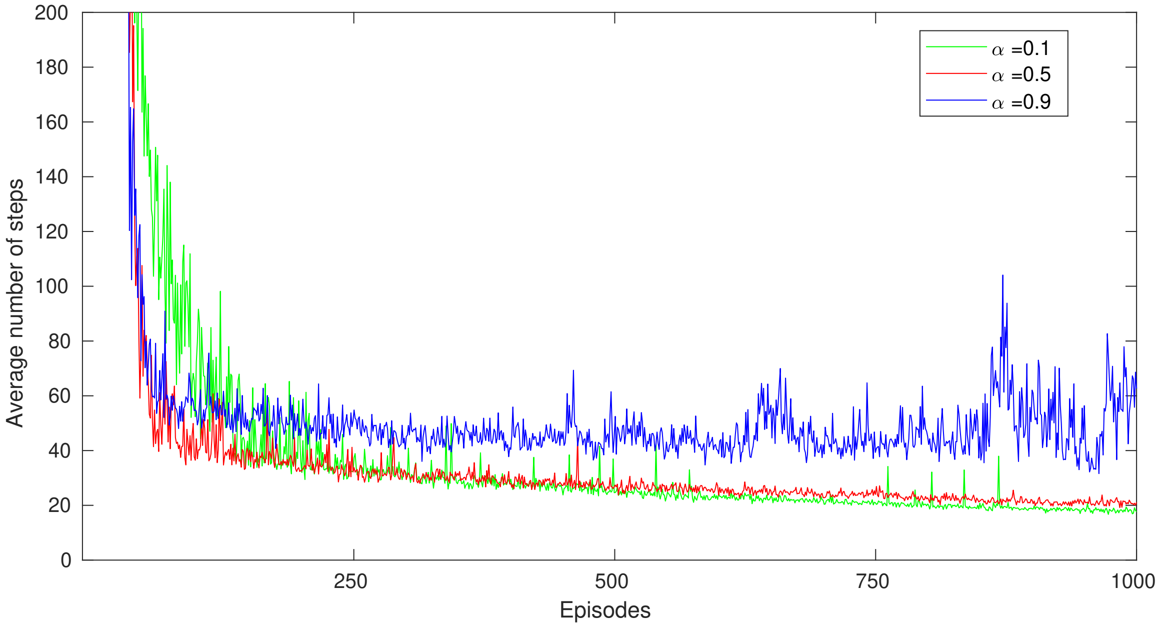

Figure 6: Average number of steps per episodes in Q-learning for different α for terminal C

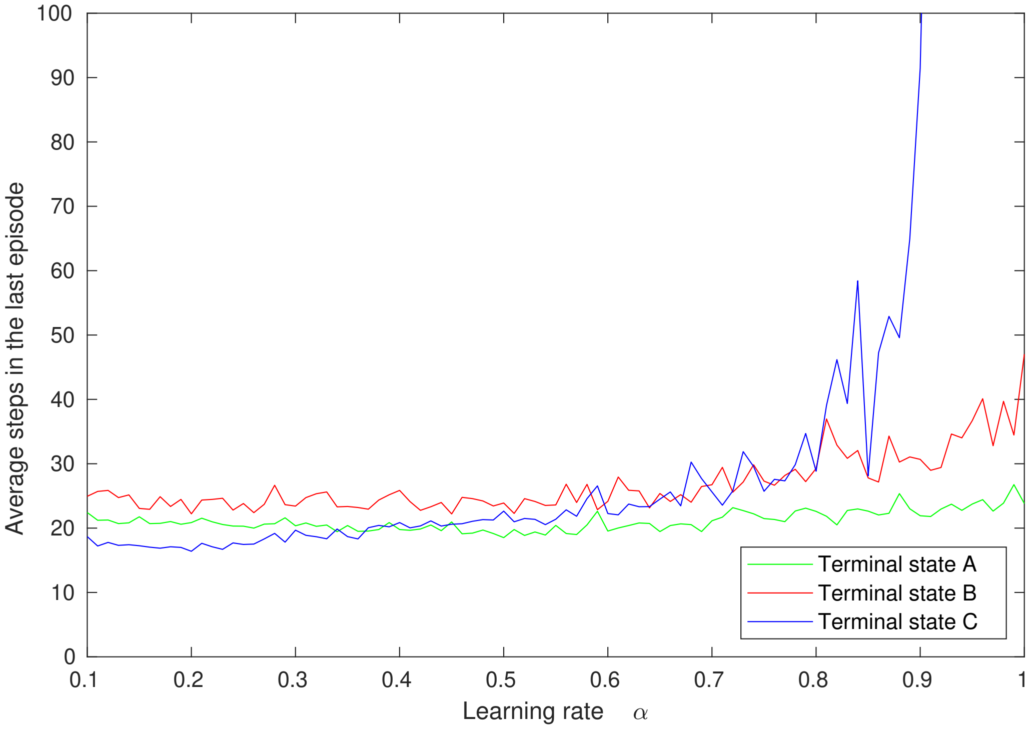

Figure 7: Average number of steps at the last episodes in Q-learning for different α

(iv) From figure (7), we observe that the number of steps to reach the terminal state in the last episode increases (asymptotic performance degrades) with increase in the learning rate (α).

1.2 (b): Optimal policy (trajectory)

The following figures shows the optimal trajectories for all the three variant problems for Q-learning. The green lines shows the trajectories from the lower two starting positions and the red lines are the trajectories from the middle starting positions. We used the dotted lines to show occasional deviations due to randomness of the world (gentle Westerly blowing).

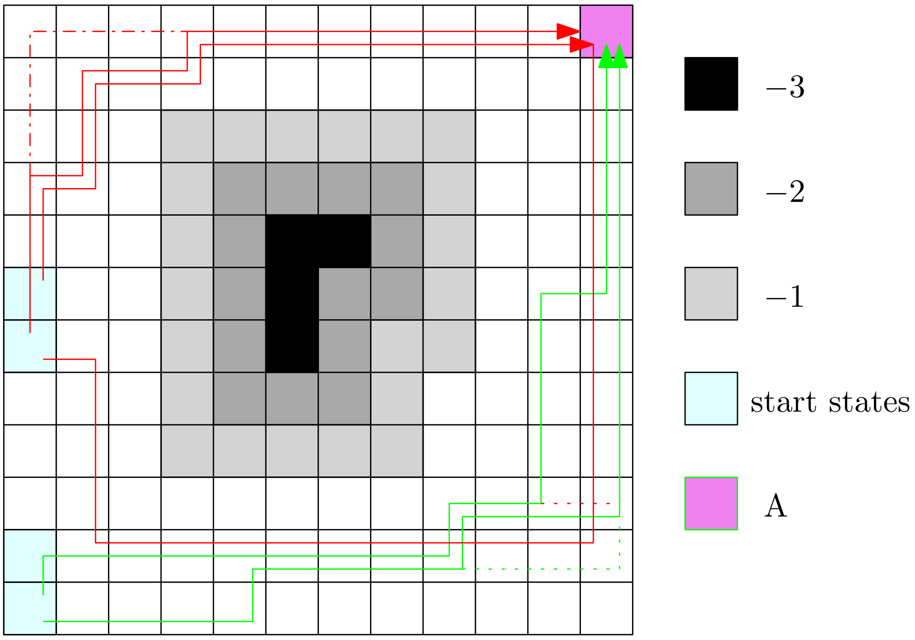

Figure 8: Optimal policy in Q-learning for terminal state A

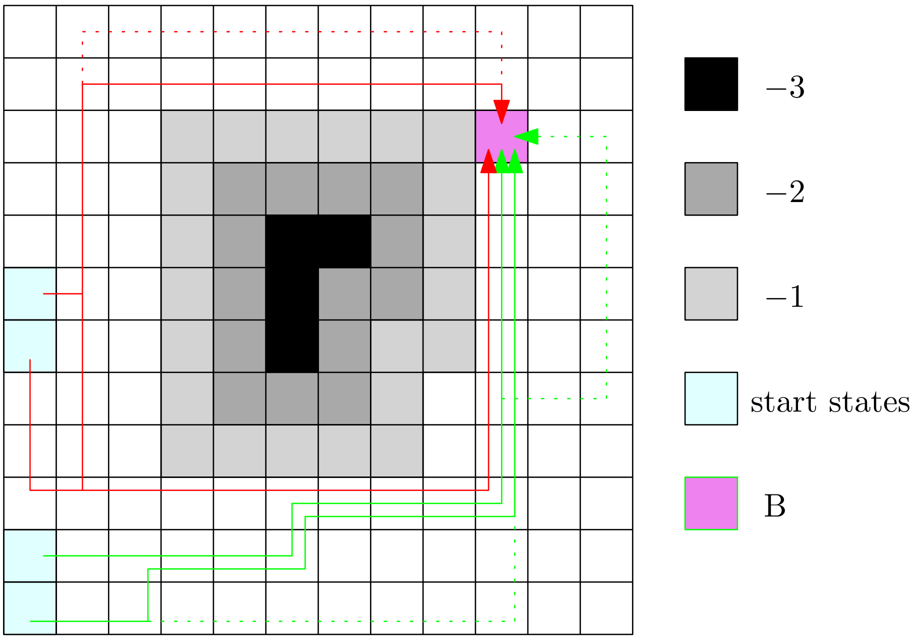

Figure 9: Optimal policy in Q-learning for terminal state B

(i) Optimal trajectories for Q-learning goes nearer to the puddle area.

(ii) For terminal state A and B optimal policies from the lower two starting positions goes below the puddle area, whereas from the two middle starting states most of times it goes above the puddle area (small fraction of time goes below the puddle area).

(iii) For terminal state C almost all time, from all stating position the agent follows one optimal policy (as the wind is off for this case).

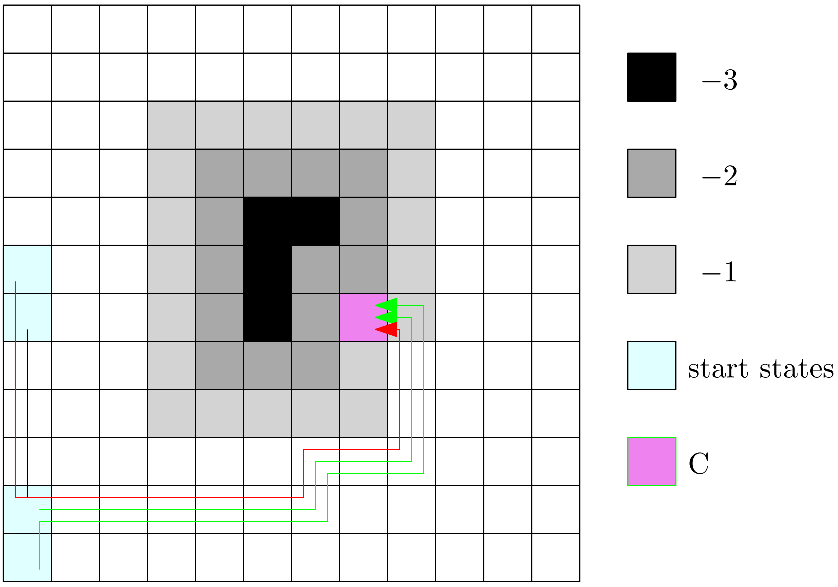

Figure 10: Optimal policy in Q-learning for terminal state C

2 SARSA

2.1 (a): Learning curves

For plotting the average reward and average steps to reach the terminal state per episodes, we used three learning rates ($\alpha = 0.1, 0.5, 0.9$).

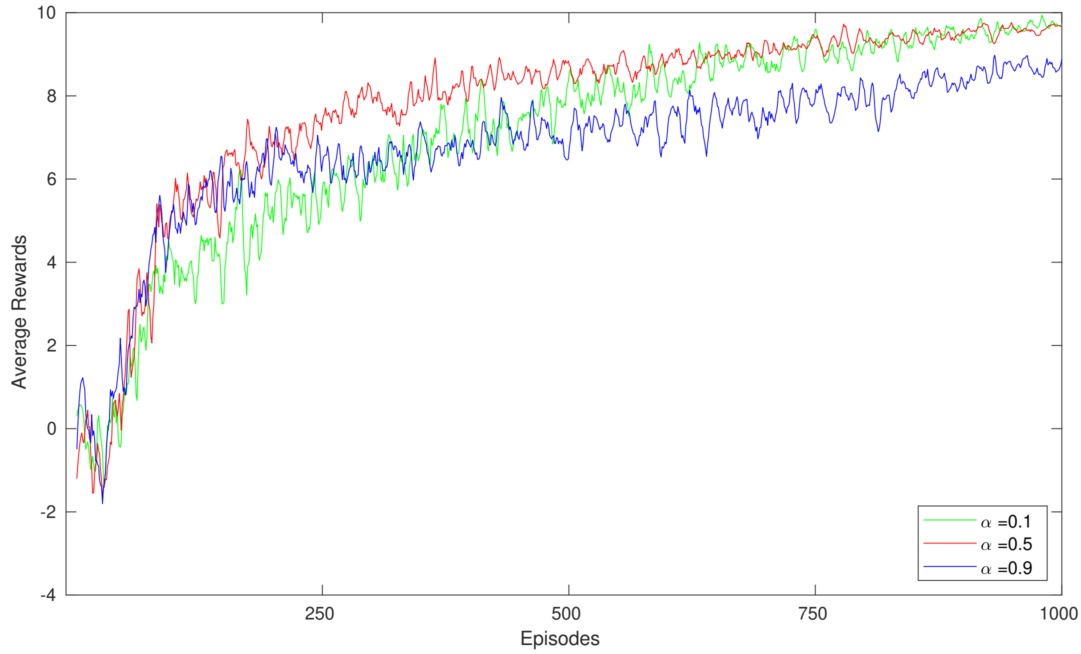

Figure 11: Average rewards per episodes in $SARSA$ for different α for terminal state A

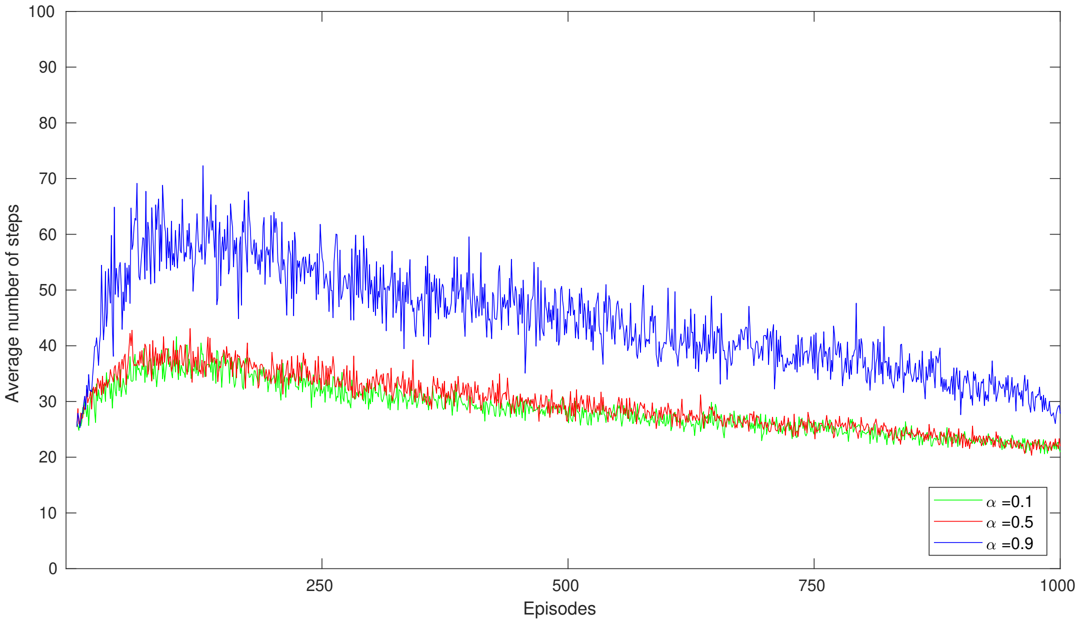

Figure 12: Average number of steps per episodes in $SARSA$ for different α for terminal A

(i) As for Q-learning we obtained similar results for SARSA. For $\alpha = 0.9$ the increase in the reward in initial episodes are high as compared to $\alpha = 0.1$, but after many episodes it get stuck in sub-optimal solution and the reward is around 9.5. On the other hand for $\alpha = 0.1$ during initial episodes though the rewards are less, the agent learns the optimal policy gradually and the rewards are very close to the optimal value of 10.

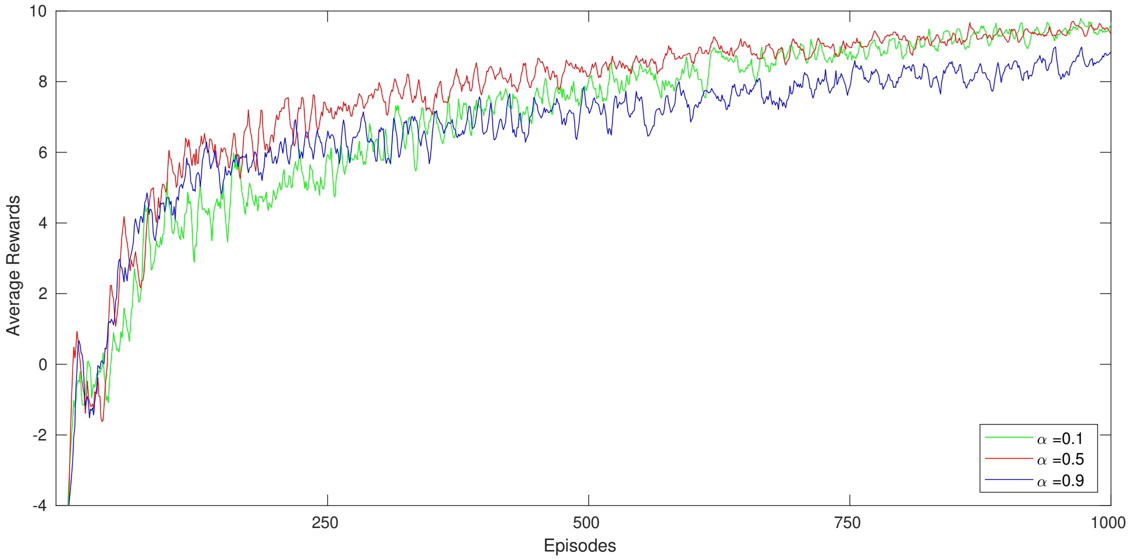

Figure 13: Average rewards per episodes in $SARSA$ for different α for terminal state B

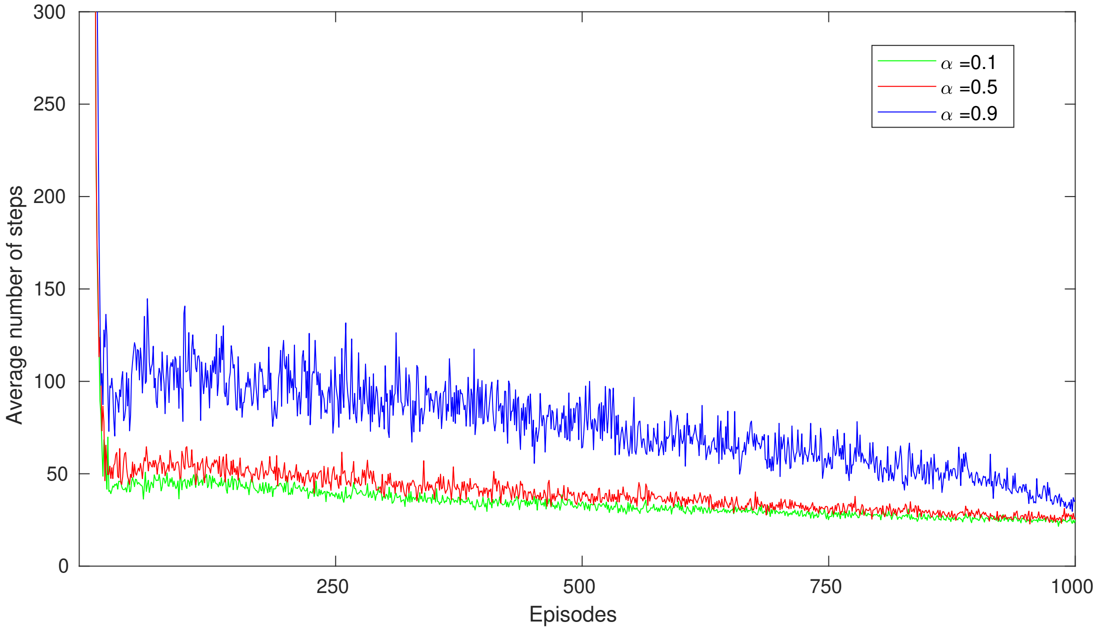

Figure 14: Average number of steps per episodes in $SARSA$ for different α for terminal B

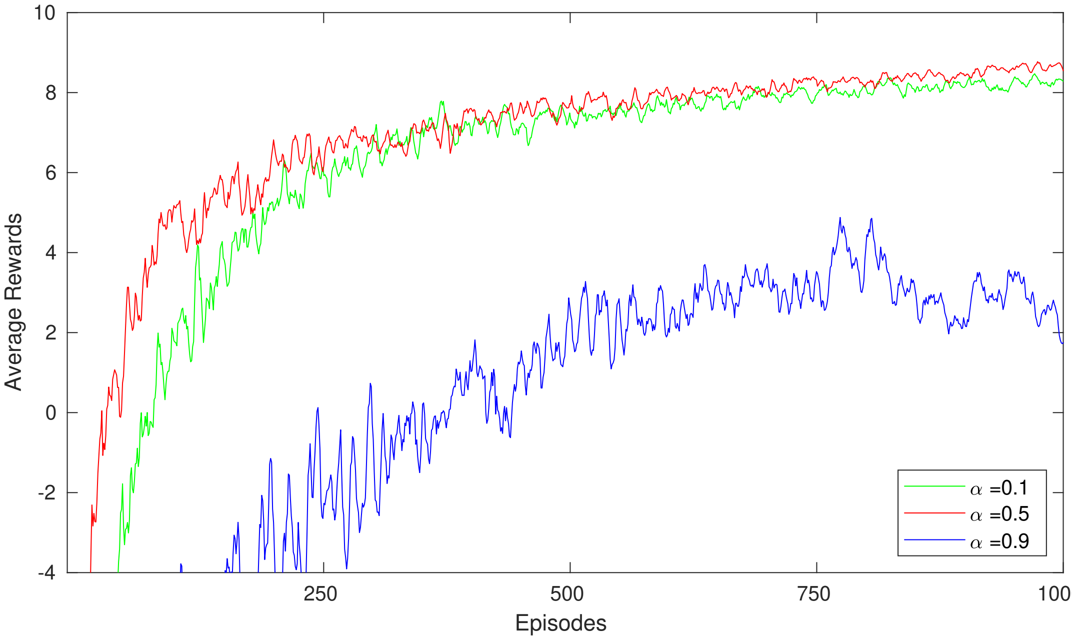

Figure 15: Average rewards per episodes in $SARSA$ for different α for terminal state C

(ii) In case of C the optimal reward is less than 10 as the agent has to step in the puddle having negative reward of -1 atleast once to get to the terminal state.

(iii) The average number of steps to reach the terminal position decreases with increase in episodes. With the increase in α value number of steps increases (the increase in steps for increase in α is more as compared Q-learning).

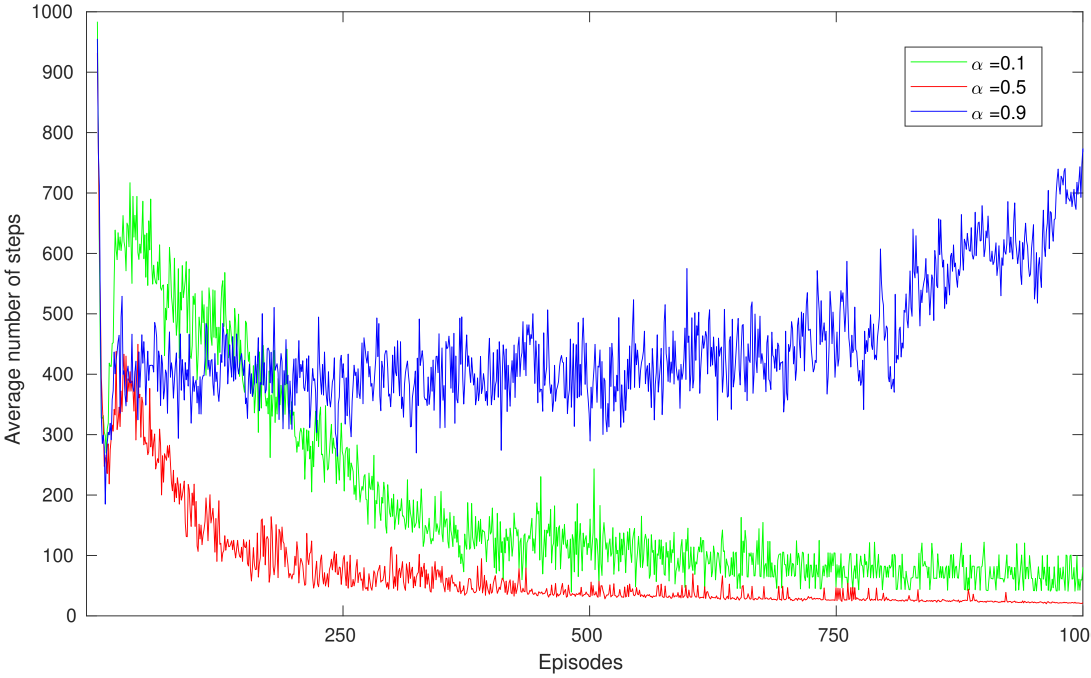

Figure 16: Average number of steps per episodes in $SARSA$ for different α for terminal C

(iv) For C the asymptotic performance degrades significantly with increase in learning rate (α) which is clear from the figure (17).

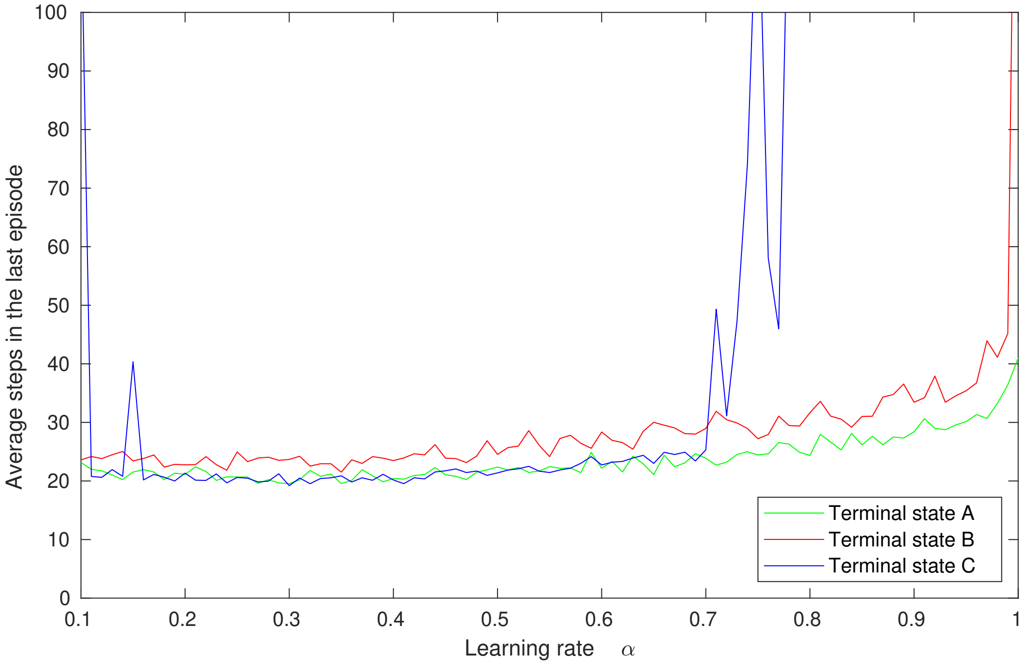

Figure 17: Average number of steps at the last episodes in $SARSA$ for different α

2.2 (b): Optimal policy (trajectory)

Similar to Q-learning the green lines shows the trajectories from the lower two starting positions and the red lines are the trajectories from the middle starting positions. We used the dotted lines to show occasional deviations due to randomness of the world (gentle Westerly blowing).

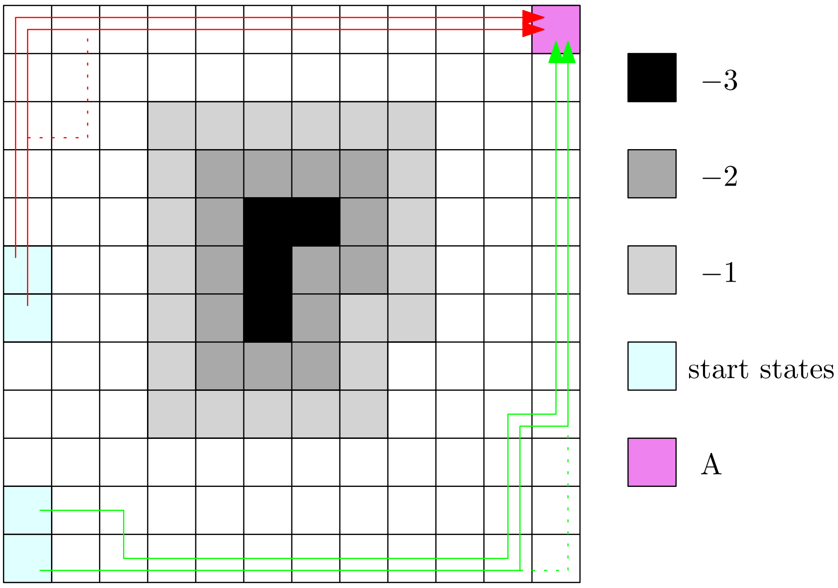

Figure 18: Optimal policy in $SARSA$-learning for terminal state A

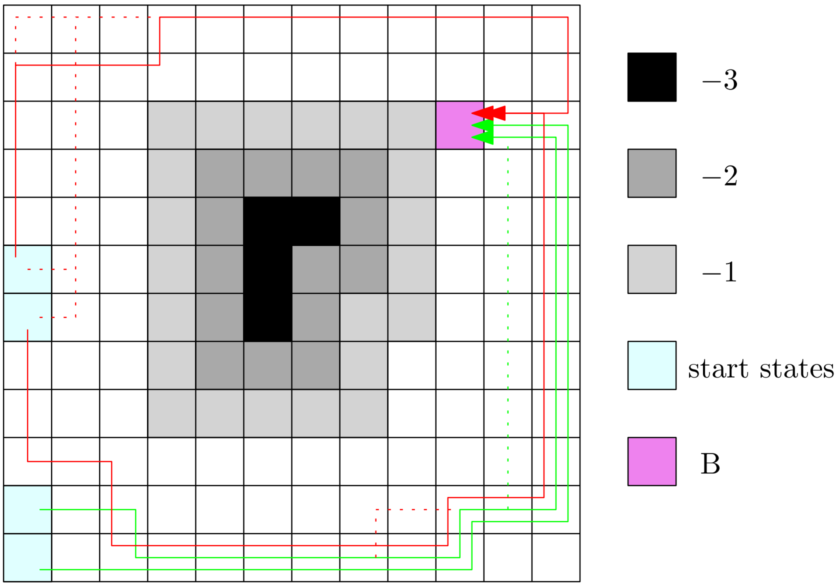

Figure 19: Optimal policy in $SARSA$-learning for terminal state B

(i) As contrast to Q-learning, optimal trajectories for SARSA goes near the edges (away from the puddle area).

(ii) For terminal state A and B optimal policies from the lower two starting positions goes below the puddle area, whereas from the two middle starting states most of the times goes above the puddle area (only a small fraction of time it goes below the puddle area).

(iii) For terminal state C almost all time, from all stating position the agent follows one optimal policy (as the wind is off for this case) which is again nearer to the edge.

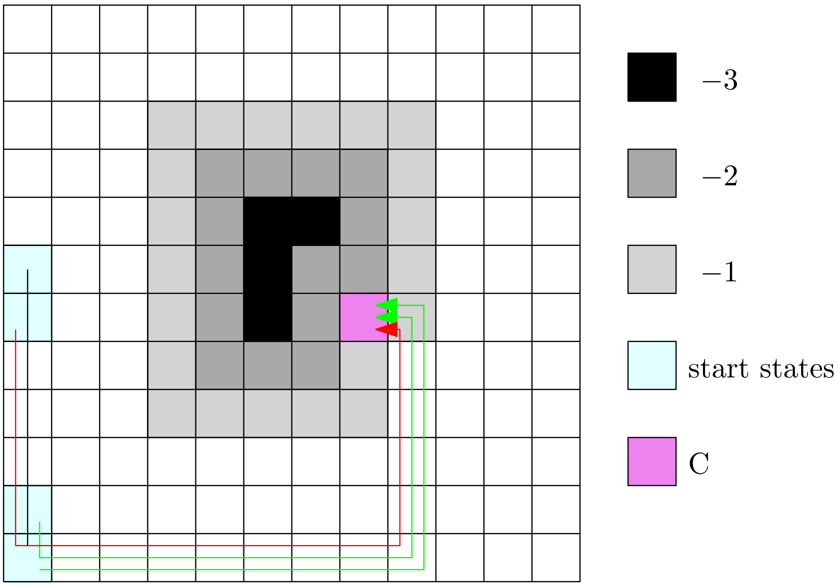

Figure 20: Optimal policy in $SARSA$-learning for terminal state C

3 SARSA (λ)

3.1 (a): Learning curves

The following figures show the average reward and average steps to reach the terminal state in each of the three variant problems for $\lambda = 0.0, 0.3, 0.5, 0.9, 0.99$ and 1.0.

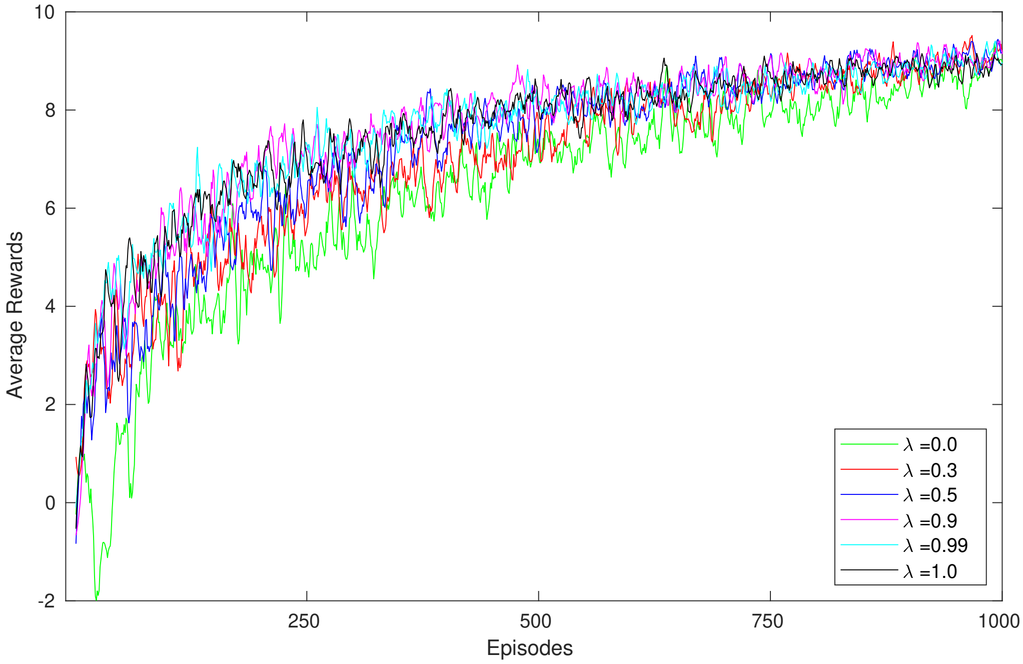

Figure 21: Average rewards per episodes for different λ for terminal state A

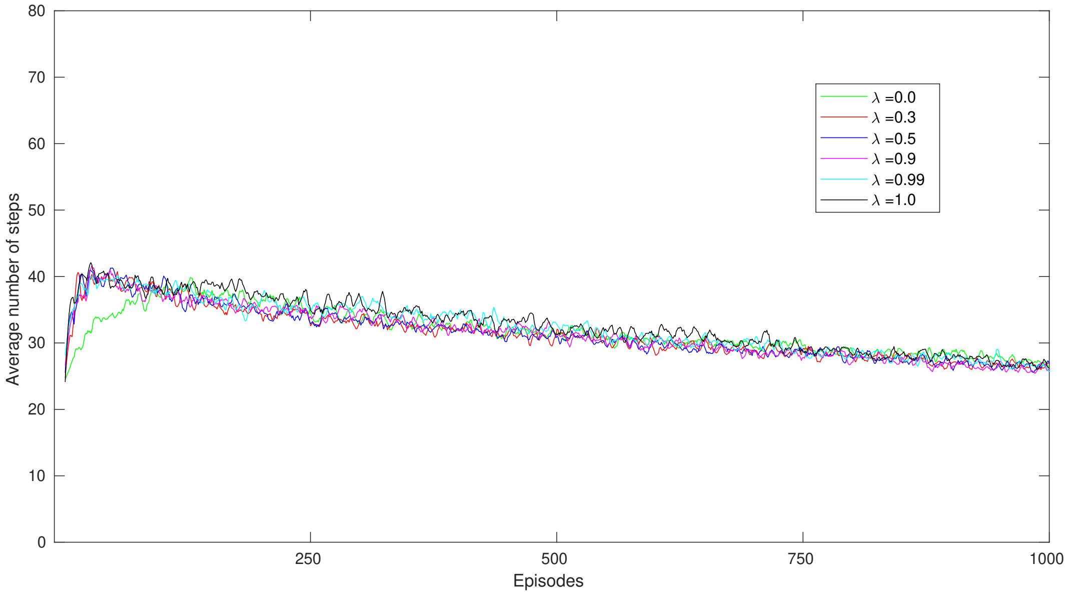

Figure 22: Average number of steps per episodes for different λ for terminal state A

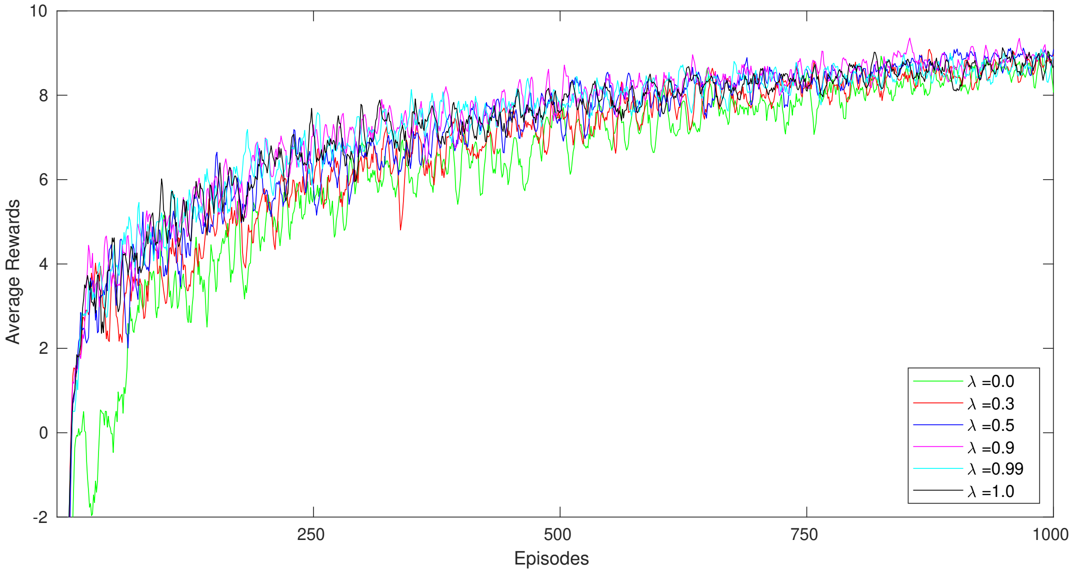

Figure 23: Average rewards per episodes for different λ for terminal state B

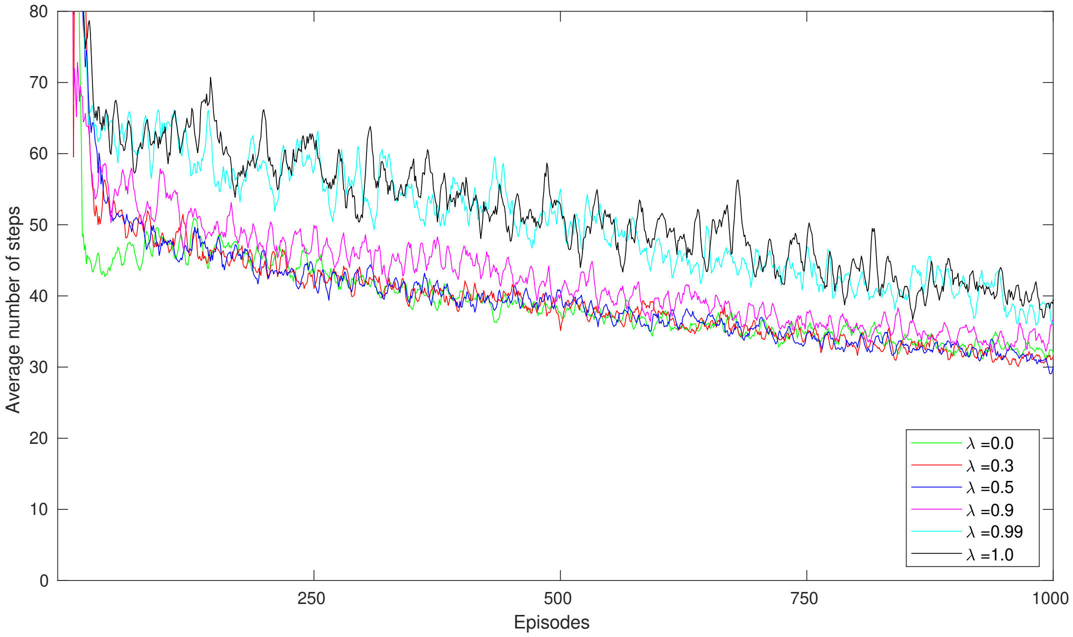

Figure 24: Average number of steps per episodes for different λ for terminal state B

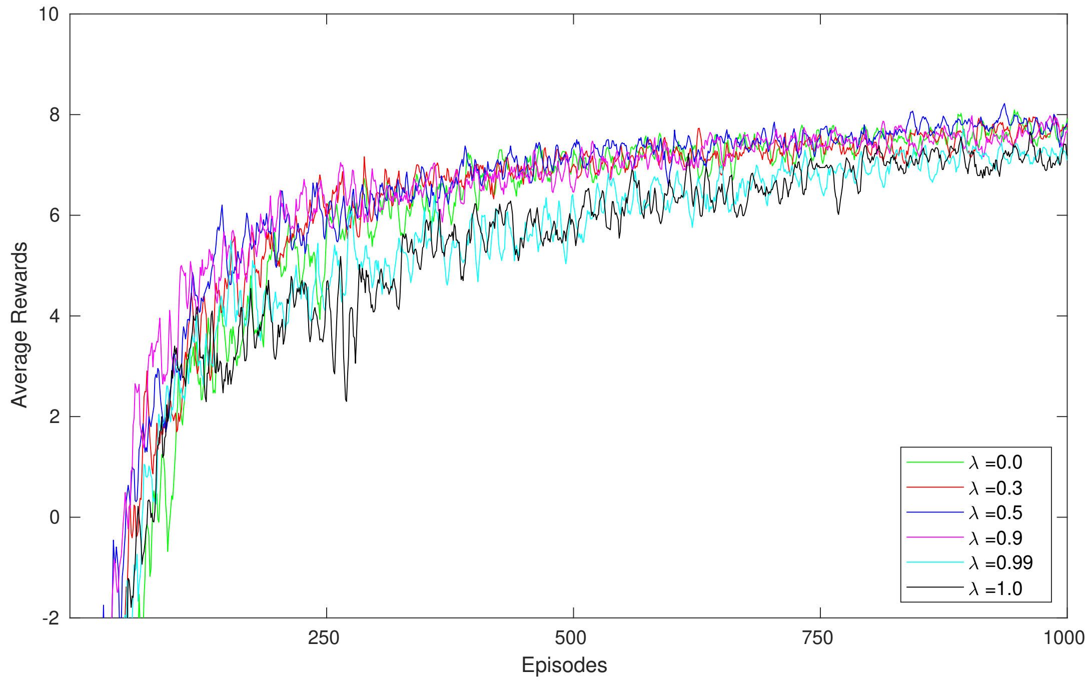

Figure 25: Average rewards per episodes for different λ for terminal state C

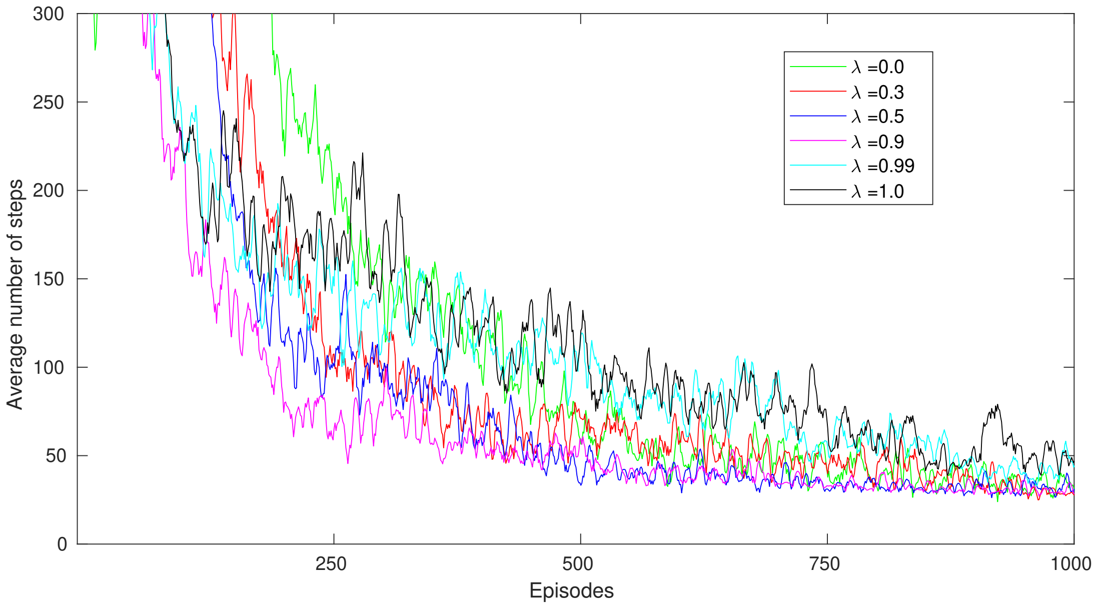

Figure 26: Average number of steps per episodes for different λ for terminal state C

Following figures shows the average number of steps and average reward in the last episode for different λ values.

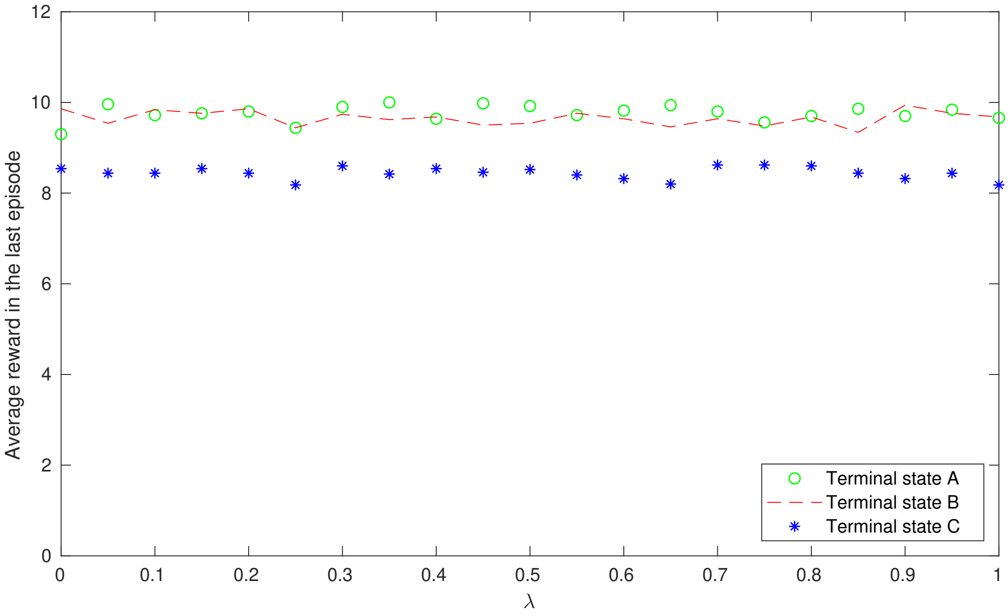

Figure 27: Average reward in last episode in $SARSA(\lambda)$ for different values of λ

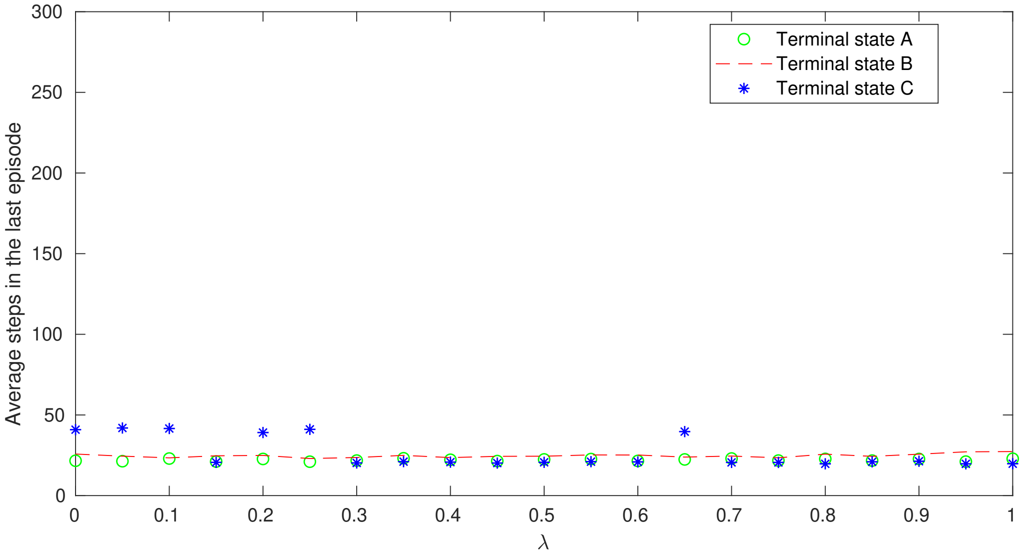

Figure 28: Average number of steps in last episode in $SARSA(\lambda)$ for different values of λ

3.2 (b): Optimal policy (trajectory)

The optimal trajectories for $SARSA(\lambda)$ are almost similar to the $SARSA$ case. The following trajectories are for the case of $\lambda = 0.5$. The observations from these trajectories are similar to the case of $SARSA$, so we did not repeat these here.

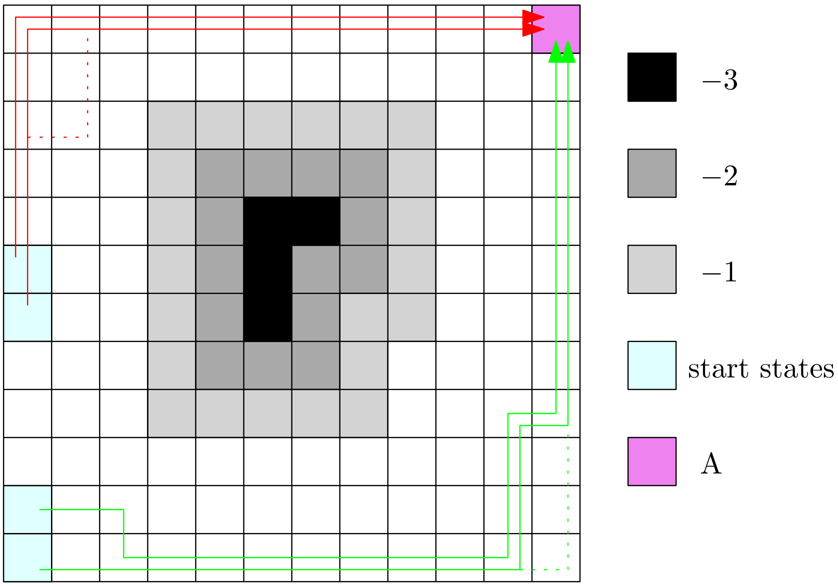

Figure 29: Optimal policy in $SARSA(\lambda)$ for terminal state A

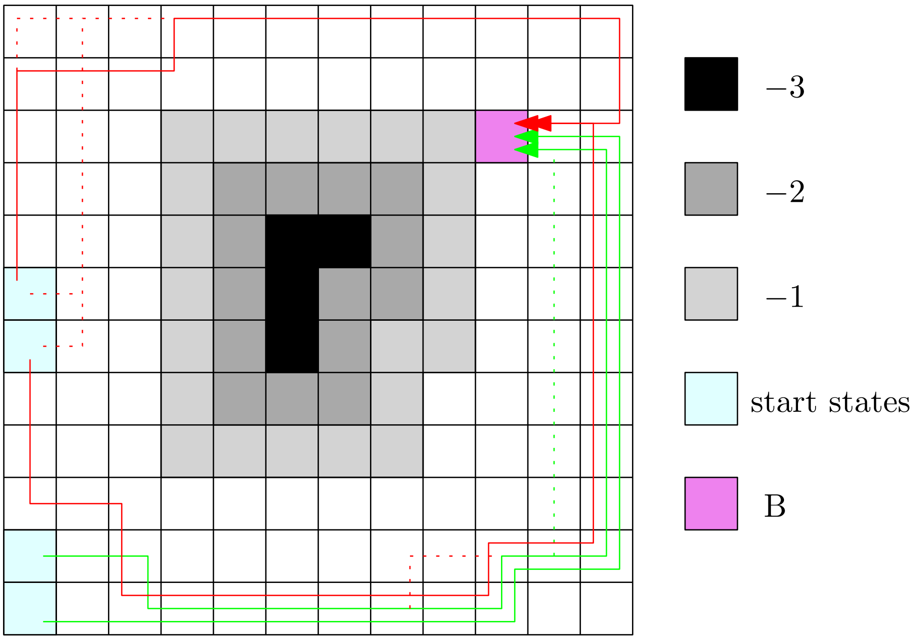

Figure 30: Optimal policy in $SARSA(\lambda)$ for terminal state B

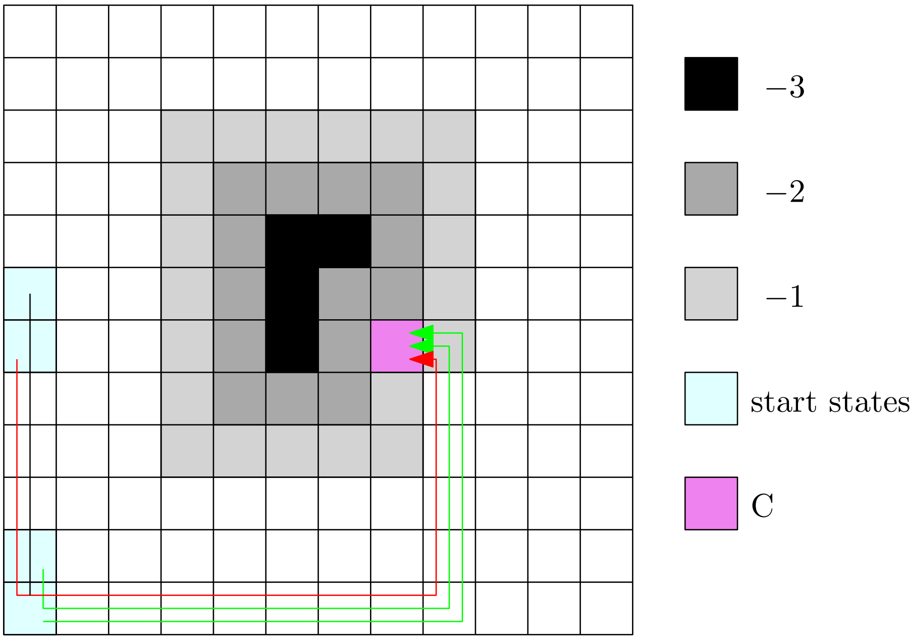

Figure 31: Optimal policy in $SARSA(\lambda)$ for terminal state C