Sumeet Shanbhag

Student ID : 641020714

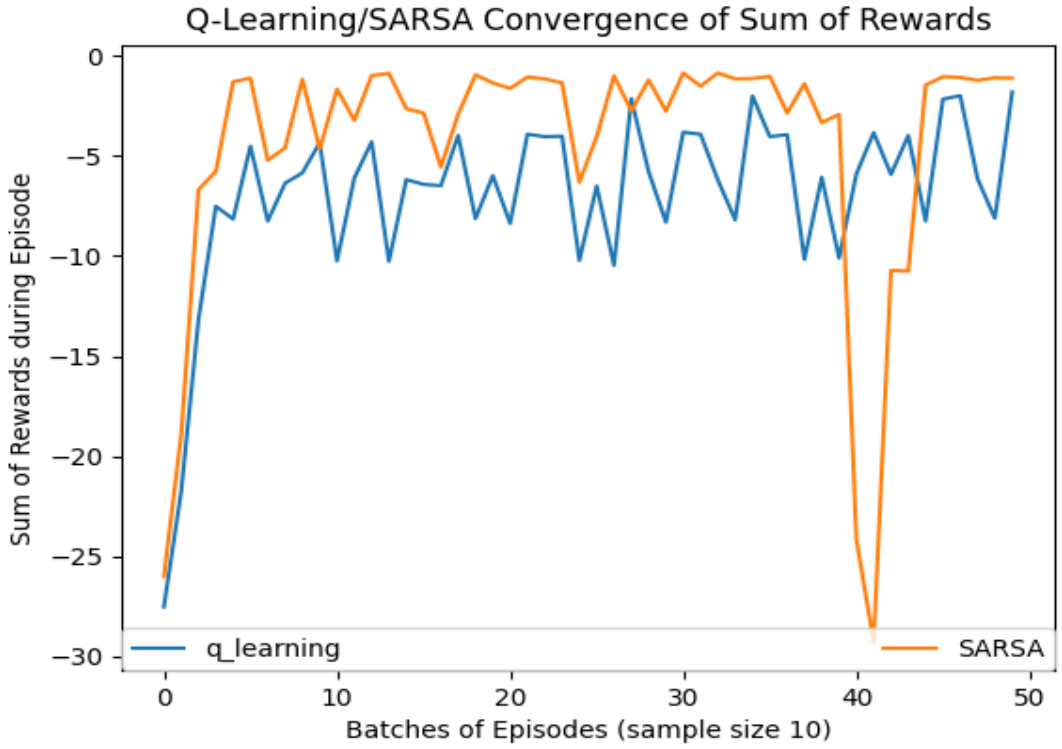
Temporal Different Programming Assignment

1. **SARSA converges to the blue path and Q-learning converges to the red path (shortest path). Explain why is that?**

   SARSA is an on-policy learning algorithm. When using an epsilon-greedy policy, SARSA tends to learn a more conservative policy, especially if there are risks of high negative rewards nearby. The cliff represents a high negative reward. When the agent is close to the cliff, even if there's a small chance of falling, SARSA considers the negative reward and tends to learn a safer path. This is why it converges to the blue path, which is a longer, but safer route.

   Q-Learning, on the other hand, is an off-policy learning algorithm. It essentially learns the optimal greedy policy while following a more exploratory policy. So, even if the agent falls off the cliff during training, over time, it learns that the shortest path (the red path) gives the highest reward, as it's not considering the future exploratory actions. This is why Q-Learning converges to the red path, which is riskier but shorter.

2. **Generate the plot of sum of the rewards as a function of episodes. Explain why Q-leaning converges to lower average rewards even though it can find the optimal path?**



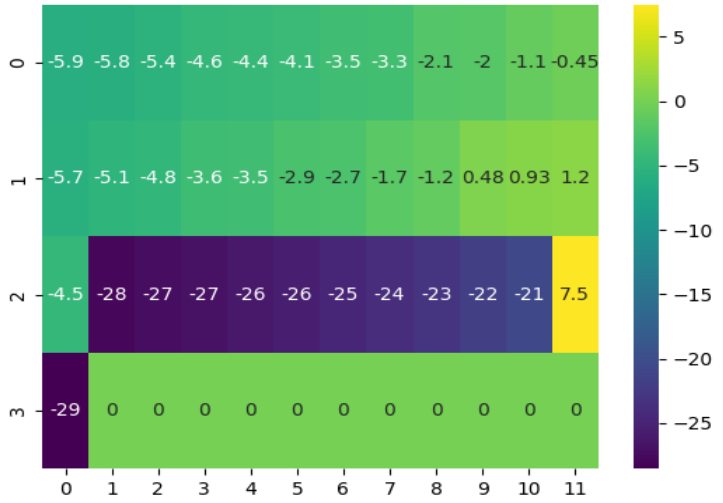
From the plotted graph of cumulative rewards, Q-Learning has lower rewards initially compared to SARSA. This is because Q-Learning is exploring the shortest path, which means it frequently falls off the cliff during the learning process. Each fall off the cliff results in a high negative

reward. So even though Q-Learning eventually learns the optimal path, the high negative rewards it accumulates during the exploration phase reduce its average reward.

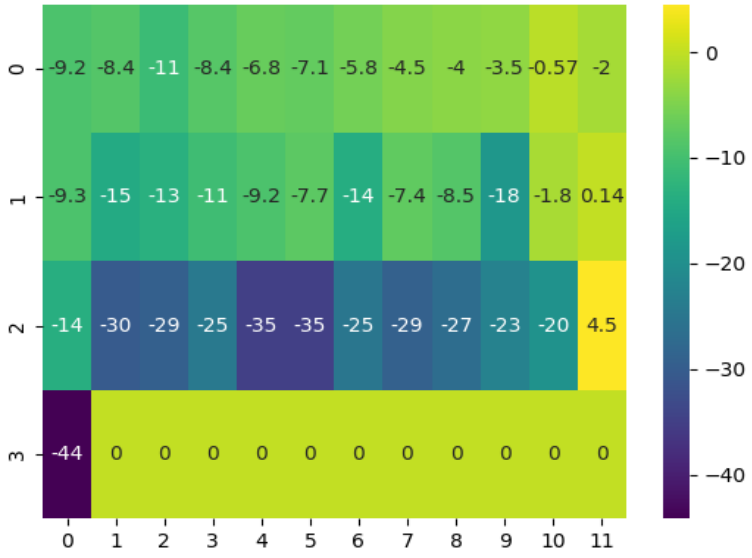
SARSA, taking the longer and safer route, doesn't fall off the cliff as frequently, resulting in it accumulating a higher average reward initially.

3. **For both methods gradually reduce the ϵ (in epsilon-greedy) and show that both algorithms converge to optimal path and explain why.**

Observing the heatmaps after gradually reducing ϵ:



Q-Learning with varying ϵ: As expected, the heatmap shows a strong preference for the red path (shortest path) when ϵ is reduced. This is because as exploration is reduced exploitation increases, Q-learning consistently chooses the most optimal path it has learned so far.



SARSA with varying ϵ: SARSA also starts to lean towards the optimal path. Initially, SARSA, being an on-policy algorithm, was learning a more conservative policy due to the chances of

exploration leading to the cliff. However, as exploration decreases with the reduction of ϵ, SARSA starts to trust its learned policy more and takes the shorter path.

In conclusion, as ϵ decreases, both Q-learning and SARSA algorithms tend to converge to the optimal path (red path). This is because the agent starts to rely more on its learned policy and less on exploration, leading it to consistently choose the most rewarding path.