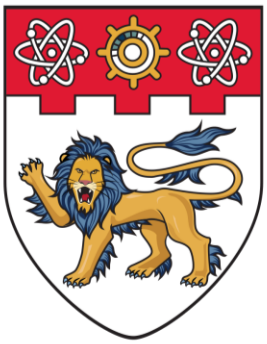

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

School of Computer Science and Engineering

Nanyang Technological University

AY2023/24

AI6101 Introduction to AI & AI Ethics

Reinforcement Learning Assignment

Submitted by: Kee Ming Yuan

Matriculation Number: G2304842E

Table of Contents

1. Introduction..... 1

2. Compare SARSA and Q-Learning Exploration Techniques…… 1

3. Final V-table..... 12

4. Policy Visualization..... 13

5. Bonus Component..... 19

6. Conclusion.....21

1. Introduction

This report seeks to implement a Reinforcement Learning algorithm for the CliffBoxPushing grid-world game with the goal of guiding the agent to reach the goal state in the fewest possible episodes. In a 6x14 grid space where the starting point of the agent, box and goal position are all fixed, the game ends under the following scenarios:

1. The agent enters the dangerous region where the agent will fall off a cliff.
2. The agent took 100 steps and still did not reach the goal state or fall off a cliff.
3. Agent reached the goal state.

The agent can only move up, down, left or right in the grid and the point system are as follows:

1. for each step taken:
    a. -1
    b. - (Manhattan distance between the box and goal)
    c. - (Manhattan distance between the box and agent)
2. If entered danger region: -1000
3. If entered goal: +1000

If the box or agent collide with the wall, either the box or the agent will stay in the same position.

2. Compare SARSA and Q-Learning Exploration Techniques

2.1. SARSA and Q-Learning techniques with the same hyperparameters

SARSA and Q-Learning differs by their Q-value computing:

SARSA Q-Value = $Q(S,A) + \alpha(R + \gamma Q(S',A') - Q(S,A)$ ---------- (1)

Q-Learning Q-Value = $Q(S,A) + \alpha(R + \gamma Q_{max}(S',A) - Q(S,A)$ ---------- (2)

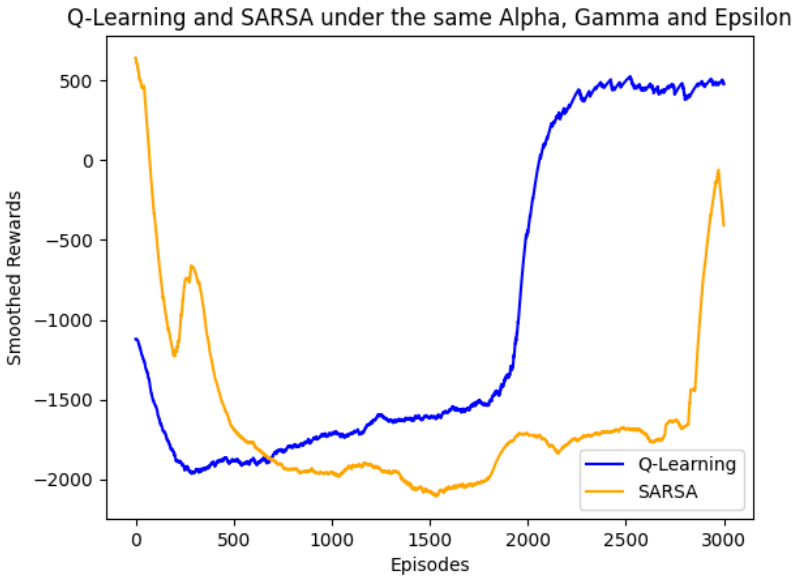

Figure 1: SARA and Q-Learning techniques with the same hyperparameters

Looking at figure 1, the performance of SARSA and Q-Learning techniques do not perform the same under the same hyperparameters of Learning Rate (α=0.5), Discount Factor (γ=0.99) and exploration (ε=0.01). This is due to their Q-value being computed differently as shown in equation 1 and 2 above.

2.4 Optimization of Hyperparameters

In this section, we seek to find the optimum hyperparameter values which give the maximum rewards. Optimization methods of reinforcement learning are namely grid search, random search, and Bayesian optimisation but only Bayesian Optimization is conducted in this report. All graphs plotted are with exploration strategy present.

2.4.1 Bayesian Optimization

A random range of values for alpha, gamma, epsilon, and number of episodes were used as per figure table 1. The graph displayed in Figure 2 represents the outcomes of executing the Q-Learning algorithm with hyperparameters derived from Bayesian Optimization, with the only exception being that the number of episodes was approximated to 2165. Interestingly, the smoothed reward achieved was 0, which is notably distant from the expected value of 642.

Table 1: Values input and output from Bayesian Optimization for Q-Learning

Range of values input into Bayesian Optimization		Values from the optimization	
Alpha	0.3-0.7	Alpha	0.6883185189590643
Gamma	0.7-0.99	Gamma	0.9882245846605997
Epsilon	0.005-0.15	Epsilon	0.012096065877324085
Number of episodes	1000-5000	Number of episodes	2495.6850642070167
		Maximum rewards	0

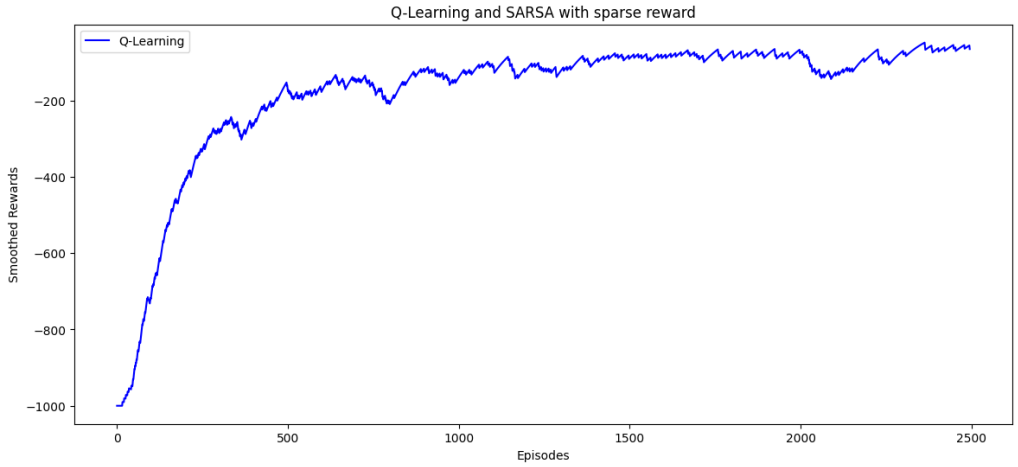

Figure 2: Smoothed rewards against episodes graph for Q-Learning from Bayesian Optimization

This shows that Bayesian optimization relies greatly on an initial exploration of the search space to build a model. The performance of this approach is notably influenced by the quality of this initial exploration. Inadequate sampling during the initial phase may result in suboptimal outcomes. It is not wise to input the full range of possible values for each hyperparameter as Bayesian Optimization can only identify the local maximum and not the global one. Consequently, the subsequent section of the report will focus on determining

optimal values for alpha, gamma, and epsilon in the context of Q-Learning and SARSA to maximize rewards. This investigation will help identify a suitable range for alpha, gamma, and epsilon that can be subsequently used in Bayesian Optimization.

2.2 Optimizing SARSA hyperparameters

Our optimization approach involves altering a single hyperparameter in isolation. The entire spectrum of values for each hyperparameter but this report will only narrow our focus to a smaller range of values of the hyperparameters.

2.2.1 Changing α

The below graphs are plotted with γ=0.99 and ε=0.01 while α changes.

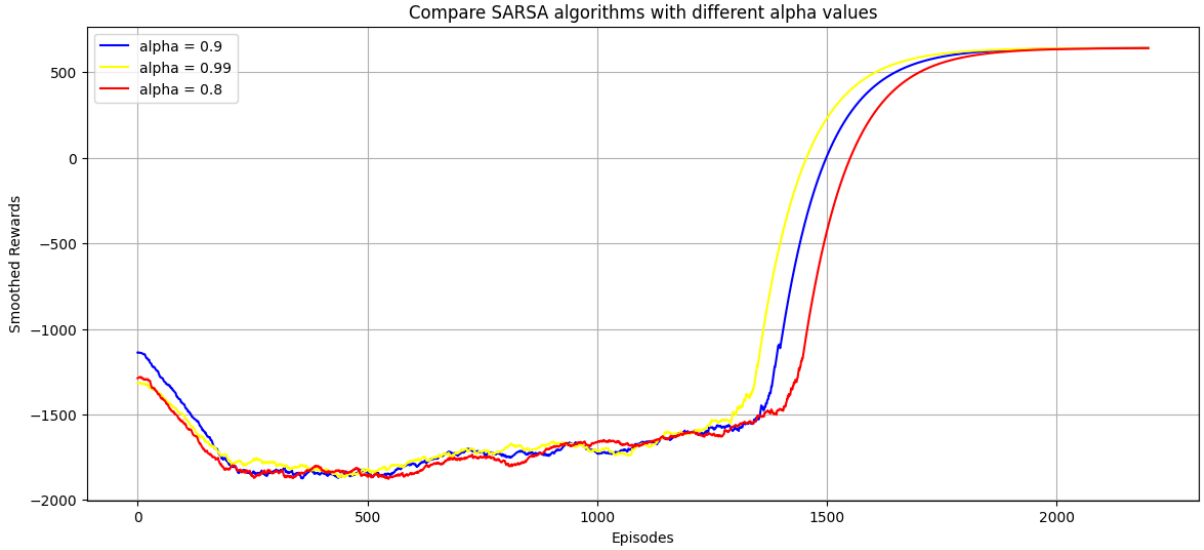

Figure 3: Change in smoothed rewards-episodes graph when changing α

The maximum absolute reward value found in this report is 642. Examining figure 3, it becomes evident that a considerably large range of α values can achieve the maximum return when γ is set at 0.99, and ε is at 0.01. Although the smoothed rewards exhibit a faster increase for α=0.99 compared to α=0.9 and α=0.8, all three curves appear to reach a stable smoothed reward level of around 600 simultaneously. This suggests that the maximum reward is not highly sensitive to variations in the α value when γ is held at 0.99 and ε at 0.01.

In principle, a larger α value enables more substantial steps in the gradient ascent algorithm, leading to a quicker rise in smoothed rewards. However, a higher learning rate can result in overshooting away from the point of maximum smoothed rewards. This might explain the smaller increase in smoothed rewards as the number of episodes progresses from 1500 to 2000. Fortunately, the large α values displayed in figure 1 does not result in divergence, where SARSA fails to converge to a solution.

Table 2 below reveals that α=0.99 gives the fewest episodes needed to achieve the 642 reward. Hence, we explore different α values close to 0.99 to determine if even fewer episodes can be found to reach the target.

Table 2: The minimum episodes for maximum rewards as alpha changes

α value	Minimum episodes for maximum reward of 642
0.8	1422
0.9	1346
0.99	1295

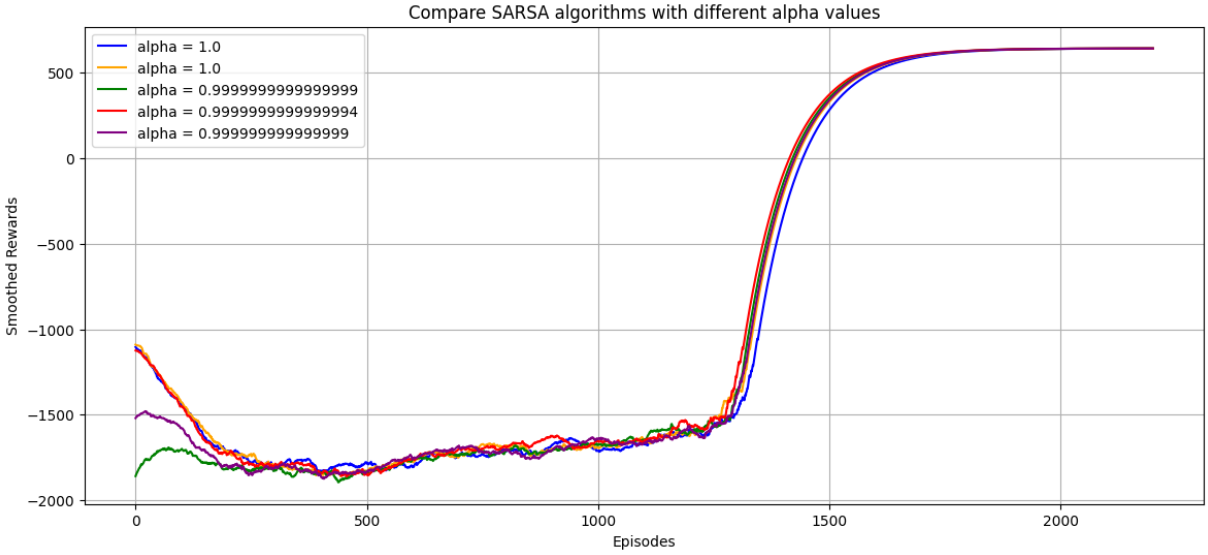

Figure 4: 1st run with the α values stated in the graph

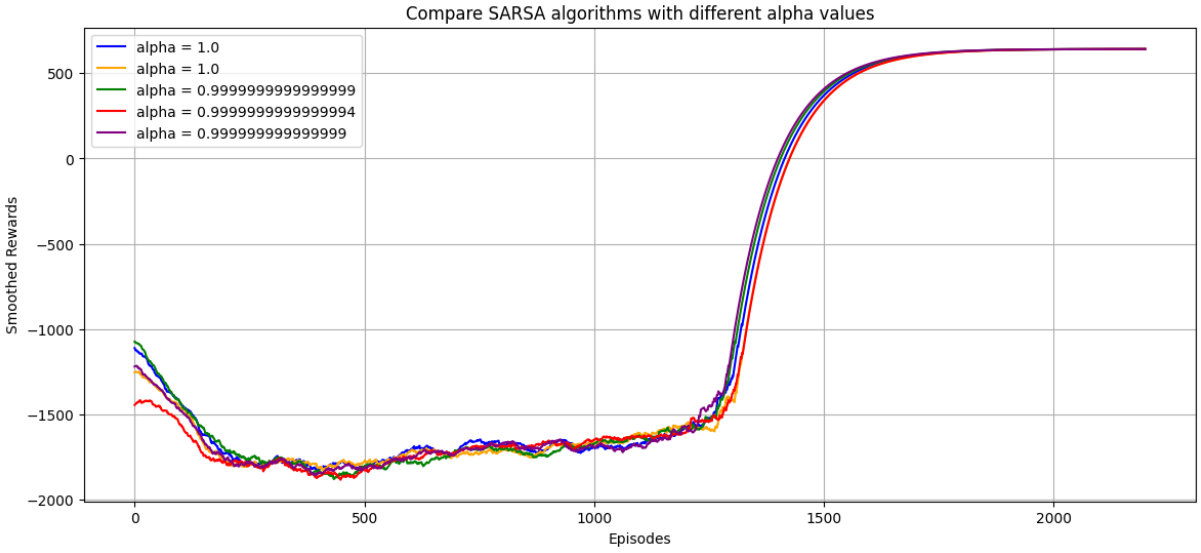

Figure 5: 2nd run with the α values stated in the graph

Table 3: Alpha values after 1st and 2nd run

1st run		2nd run	
α value	Minimum episodes for maximum rewards	α value	Minimum episodes for maximum rewards
0.999999999999999	1240	0.999999999999999	1263
0. 9999999999999994	1217	0. 9999999999999994	1259
0.9999999999999999	1256	0.9999999999999999	1189
0.99999999999999995	1263	0.99999999999999995	1260
0.9999999999999999	1279	0.99999999999999999	1288

As observed in figures 4 and 5, as well as table 3, the minimum number of episodes needed to get 642 reward exhibits slight variations from one run under the same alpha values. Therefore, it can be concluded that small changes in the alpha value do not significantly impact the minimum number of episodes required to achieve the 642 reward. Consequently, alpha=0.9999999999999994 is selected as the optimal alpha value for the subsequent hyperparameter optimization, which will be discussed later in this report.

Although it is not possible to entirely eliminate the inherent randomness in the reinforcement learning process, one can enhance the reproducibility of the results across different runs by utilizing a seed value at the beginning of the script.

2.2.2 Changing γ

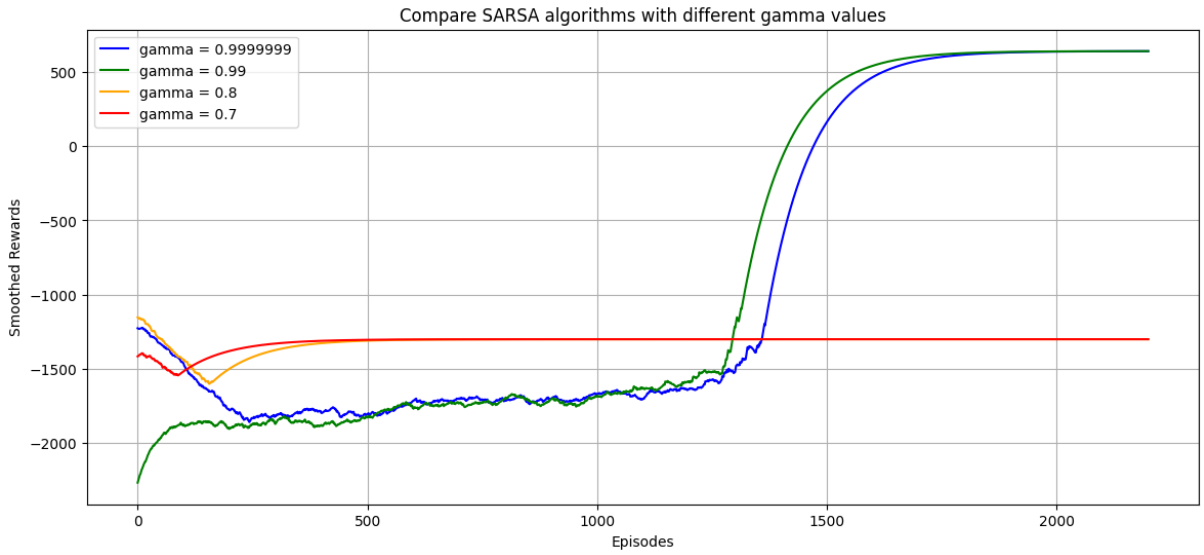

Figure 6: Change in smoothed value and episodes as gamma changes

Table 4: The minimum episodes for maximum rewards as alpha changes

γ value	Maximum rewards	Minimum episodes for maximum rewards
0.7	-1068	4
0.8	-1053	5
0.99	642	1274
0.9999999	642	1308

Table 4 and figure 6 provide insight into the impact of different gamma values. It is evident that gamma=0.8 and gamma=0.7 fail to achieve the 642 reward, while the range of gamma=0.9907 to gamma=0.9923 does give the maximum reward of 642.

In our specific game problem, employing larger gamma values aligns with our intuition. This is because our game gives the agent additional points only when it successfully reaches the goal state, and any other actions result in negative point rewards. Therefore, prioritizing future rewards, which occurs only when reaching the goal state, makes the agent place greater importance in long-term consequences rather than being fixated on immediate rewards.

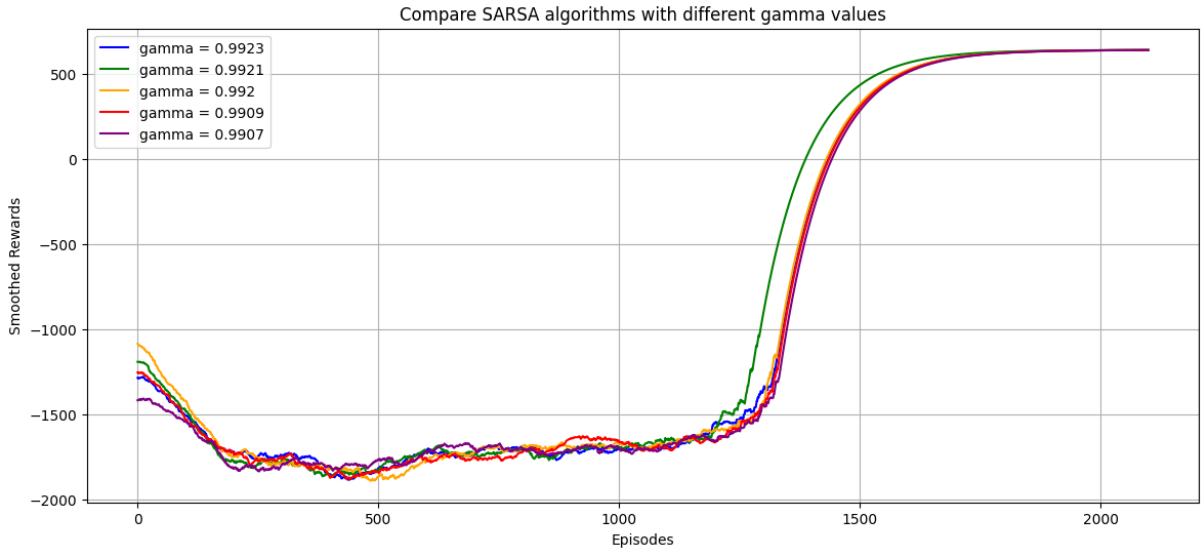

Figure 7: Optimizing gamma values for SARSA

Table 5: Minimum episodes needed to get maximum reward with different gamma values for SARSA

γ value	Minimum episodes needed to get maximum reward of 642
0.9923	1295
0.9921	1294
0.992	1278
0.9909	1259
0.9907	1306

Figure 7 and table 5 provide insights regarding gamma values. It's evident that gamma values within the range of 0.9907 to 0.9923 yield a similar minimum number of episodes required to attain the maximum reward of 642. As explained earlier, the minimum number of episodes varies slightly in different runs. Therefore, gamma=0.9923 is selected as the gamma value when optimizing other hyperparameters below.

2.2.3 Changing ε

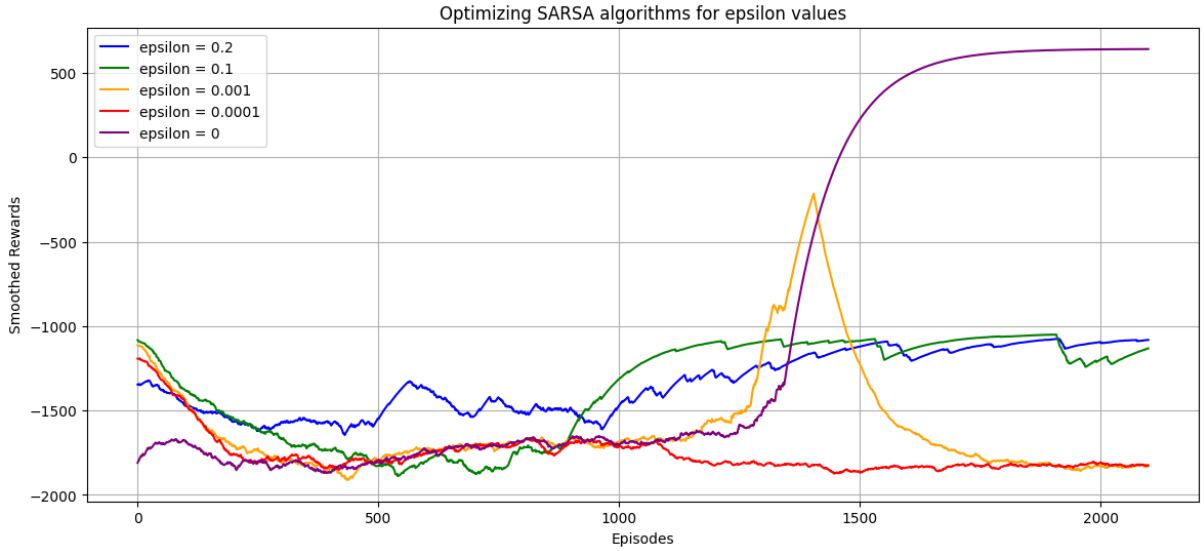

Figure 8: Optimizing epsilon value for SARSA

Table 6: The minimum episodes for maximum rewards as epsilon changes

ε value	Maximum rewards	Minimum episodes for maximum rewards
0.2	-1039	7
0.1	-1039	1388
0.001	642	1265
0.0001	-882	700
0	642	1325

Table 6 reveals that both epsilon=0.001 and epsilon=0 result in the maximum reward of 642, and the number of episodes required to achieve this maximum reward is nearly identical for both. However, examining figure 8, we observe that the smoothed reward for epsilon=0.001 displays instability as it starts decreasing around 1450 episodes and continues to decline at higher episode numbers. This decline can potentially lead to deviations from the goal state as the number of episodes increase.

The preference for epsilon=0 suggests that, in the context of our game environment, adopting a conservative and greedy strategy, which exploits the best action based on the learned policy, outperforms the riskier approach of exploring the environment. This is because the game's action space is limited to only four choices, and the game environment is relatively small.

Figure 8 further illustrates that when epsilon is set to 0, the agent experiences the most negative rewards at the beginning episodes compared to other epsilon values. This is attributed to the absence of exploration in the environment, causing the agent to make numerous incorrect moves at the beginning episodes due to its limited knowledge, until it refines the Q-values through exploitation and converge to an optimum policy and reach the goal state at later episodes.

2.3 Optimizing Q-Learning hyperparameters

2.3.1 Changing ε

The below graphs are plotted with γ=0.99 and α=0.5 while ε changes.

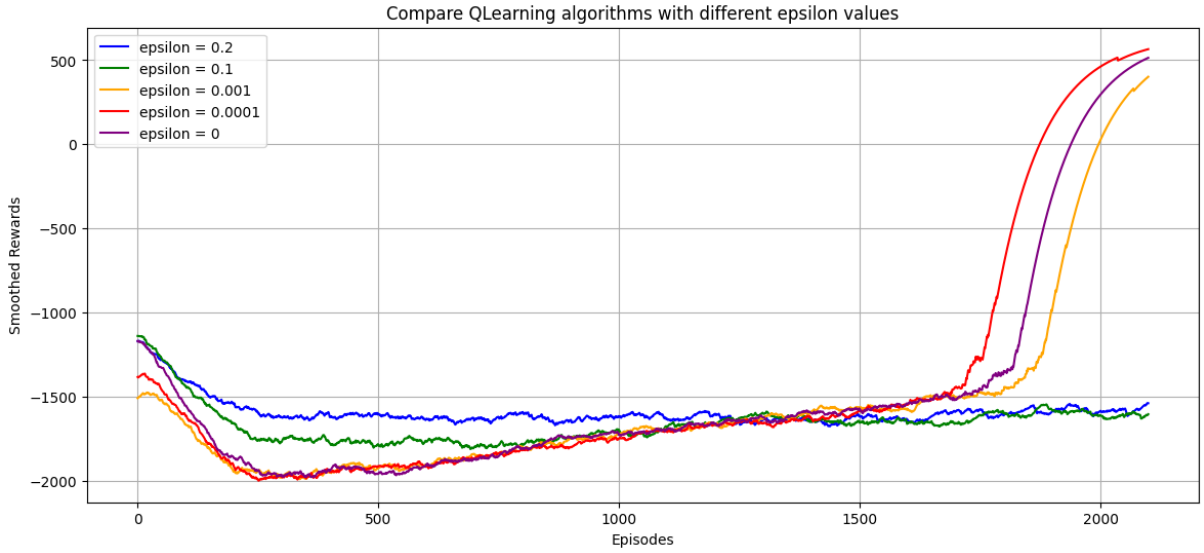

Figure 8: Changing epsilon for Q-Learning

Table 7: The minimum episodes for maximum rewards as epsilon changes

ε value	Maximum rewards	Minimum episodes for maximum rewards
0.2	-921	1926
0.1	-851	1540
0.001	642	1847
0.0001	642	1729
0	642	1776

We observe that epsilon values ranging from 0 to 0.001 result in the same maximum reward of 642 at approximately the same number of episodes, as indicated in Table 6. Hence, we narrow our optimization efforts within this epsilon range.

Table 8: The minimum episodes for maximum rewards as epsilon changes

ε value	Minimum episodes for maximum rewards
0.0001	1795
0.00009	1841
0.00008	1752

As previously mentioned, it's impossible to completely eradicate the randomness inherent in the reinforcement learning algorithm, resulting in varying minimum number of episodes needed to achieve a reward of 642 across different runs. Upon examining Table 8, we observe that these epsilon values from 0.00008 to 0.0001 yield approximately the same number of maximum rewards. Therefore, we opt for epsilon=0.00008 as our epsilon value while optimizing the other hyperparameters.

2.3.2 Changing α

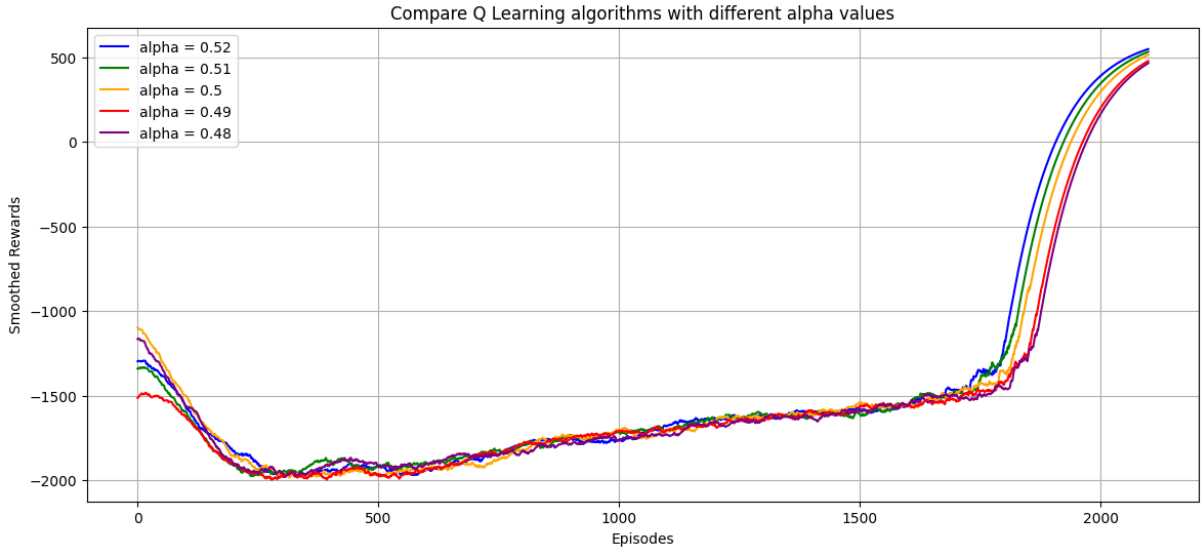

Figure 9: Observing the change in smoothed value over a range of alpha for Q-Learning

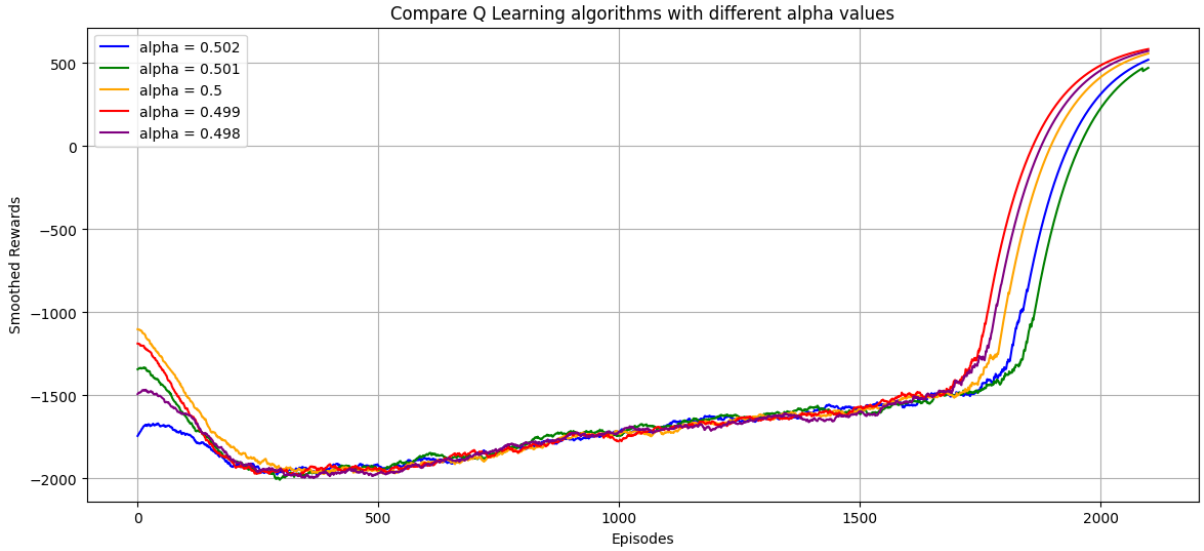

Figure 10: Observing the change in smoothed rewards as alpha value changes for Q-Learning

Table 9: Minimum episodes for 642 when changing alpha based on alpha values in figure 9 and 10

α value	Minimum episodes for 642 reward	α value	Minimum episodes for 642 reward
0.52	1791	0.502	1754
0.51	1722	0.501	1732
0.5	1768	0.5	1707
0.49	1817	0.499	1815
0.48	1787	0.498	1772

Alpha=0.5 was executed twice in Table 9, yielding different minimum episode counts required to achieve a reward of 642 due to the inherent randomness in the reinforcement learning process. The minimum number of episodes for reaching a reward of 642 appears to

be quite consistent across alpha values ranging from 0.48 to 0.52. Therefore, we select alpha=0.5 and will employ it as the alpha value while optimizing the other hyperparameters.

It's worth noting that the alpha value in Q-Learning is smaller compared to that in SARSA. This distinction may arise from the fact that Q-Learning considers the maximum Q-value from all available actions in the current state when transitioning to the next state to determine the new Q(S,A). This approach introduces an overestimation bias, making a smaller learning rate preferable for increased learning stability. In contrast, SARSA has a reduced risk of overestimation because it considers the action taken in the next state and relies on the Q-value of that specific action. Hence, SARSA is able to employs a larger learning rate while still converging.

2.3.1 Changing γ

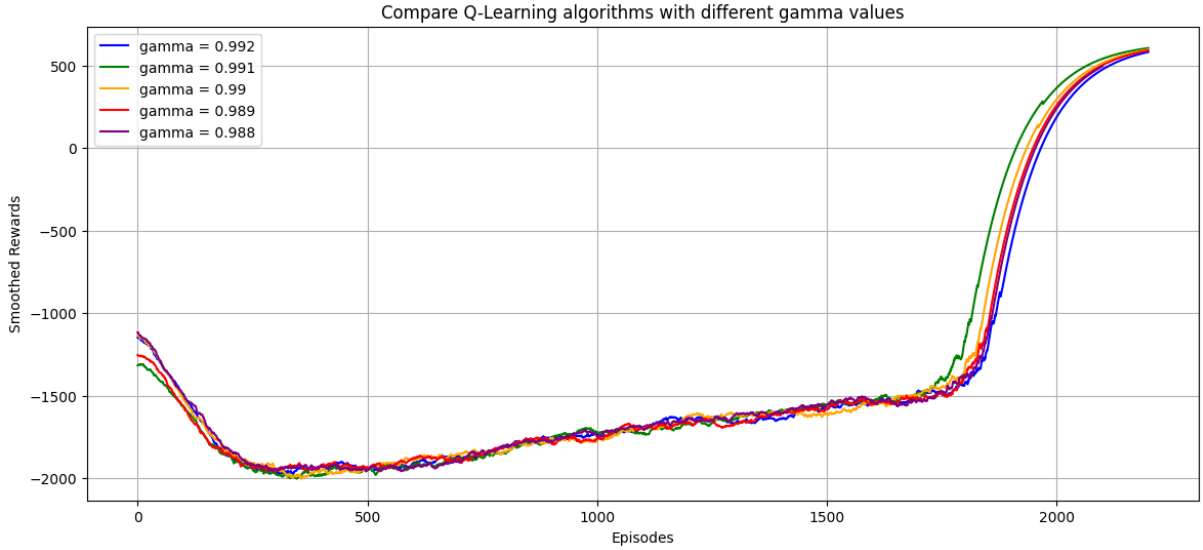

Figure 11: Observing the change in smoothed rewards when gamma changes

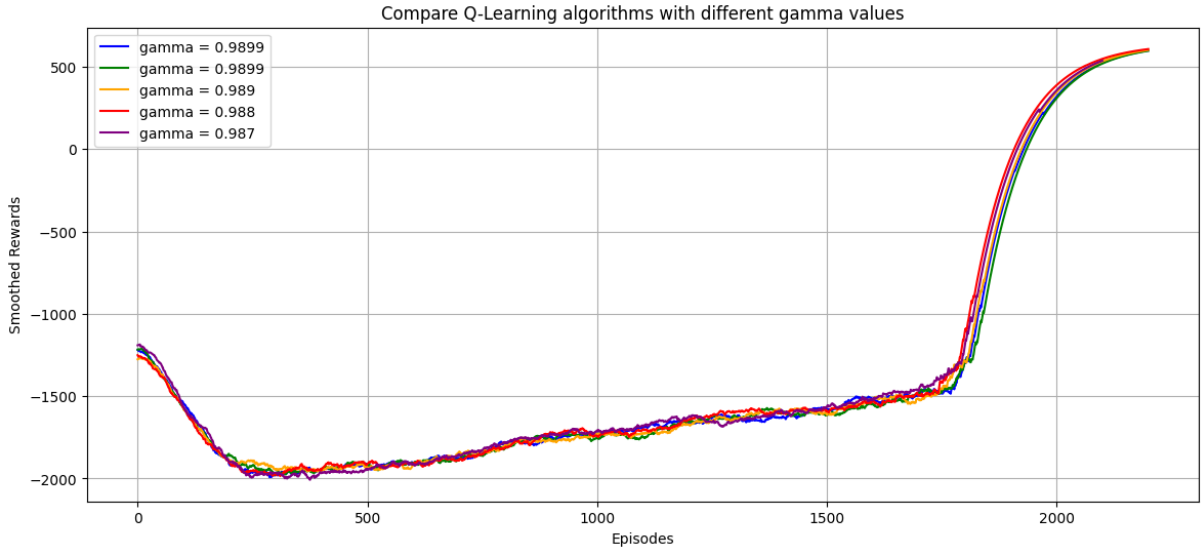

Figure 12: Observing the change in smoothed rewards when gamma changes

Table 10: Gamma values from figure 11 and 12 and their minimum episodes for 642 reward

γ value	Minimum episodes for 642 reward	γ value	Minimum episodes for 642 reward
0.992	1775	0.9899	1790
0.991	1819	0.9899	1754
0.99	1782	0.989	1797
0.989	1781	0.988	1803
0.988	1827	0.987	1794

As indicated in Table 10, there is minimal change to the minimum episode counts required to achieve the reward of 642 as gamma changes within a narrow range of values from 0.987 to 0.992. Consequently, we opt for gamma=0.987 as the optimal value for Q-Learning.

2.4 Bayesian Optimisation with better hyperparameters input range

The Bayesian Optimization is run for the Q-Learning and SARSA algorithm and the results are shown below.

Table 11: Values input output from Bayesian Optimization for Q-Learning

Values input to optimization		Values from the optimization	
Alpha	0.48-0.52	Alpha	0.4949816047538945
Gamma	0.98-0.99	Gamma	0.987319939418114
Epsilon	0-0.0001	Epsilon	0.00009507143064099161
Number of episodes	1000-5000	Number of episodes	3394.6339367881465
Maximum rewards			
642			
Minimum episodes to reach maximum rewards			
1750			

Table 12: Values input output from Bayesian Optimization for SARSA

Values input to optimization		Values from the optimization	
Alpha	0.99-0.99999999	Alpha	0.9915601848442379
Gamma	0.99-0.994	Gamma	0.9902323344486728
Epsilon	0-0.00001	Epsilon	0.000001559945203362026 6
Number of episodes	1000-5000	Number of episodes	4464.704583099741
Maximum rewards			
642			
Minimum episodes to reach maximum rewards			
1270			

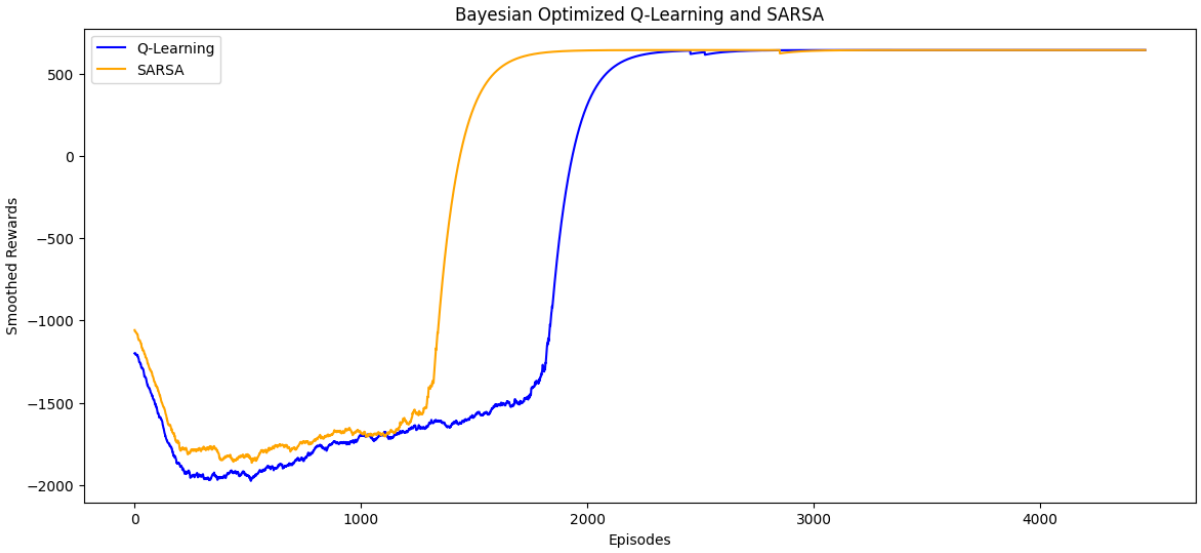

Figure 13: Smoothed rewards against episodes graph for Q-Learning and SARSA with Bayesian Optimization

Considering the data presented in Table 11 and 12, it is evident that within the defined range of hyperparameter inputs, the maximum attainable reward in this gaming environment is 642. Given that SARSA demonstrates a shorter minimum episode requirement to reach the maximum reward of 642, we will employ SARSA for further analysis in this report.

3. Final V-Table

The following analysis will be done for SARSA at epsilon, alpha and gamma values as per the Bayesian Optimization and number of episodes=2000.

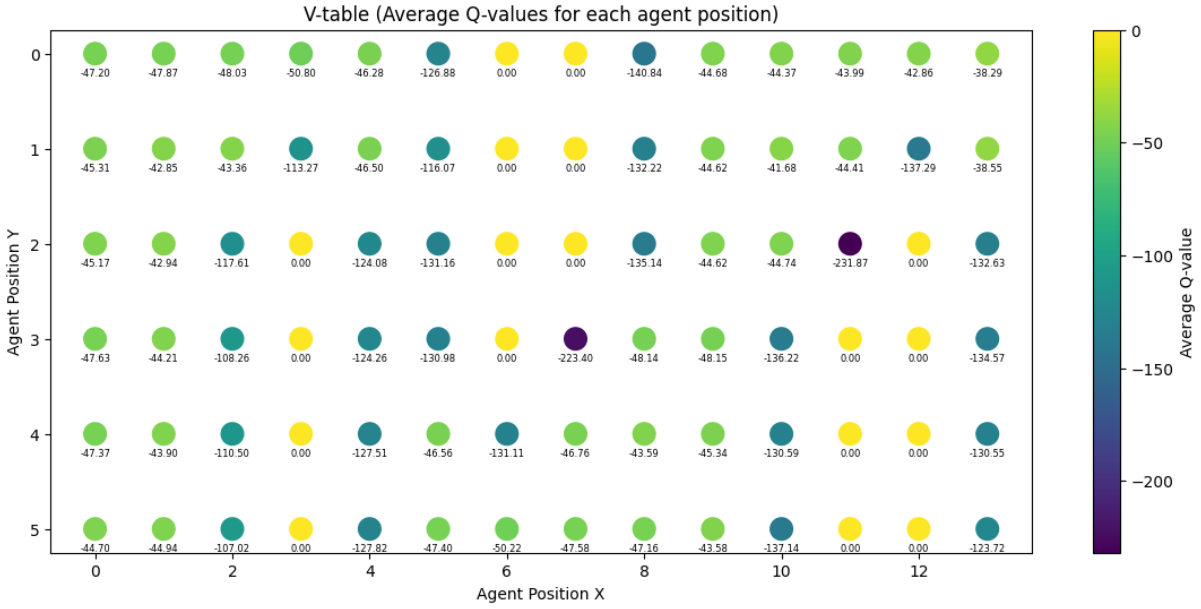

Figure 12: V-Table with respect to the agent position for SARSA with 2000 episodes

Figure 14 is the V-table showing the average reward when the agent is at any of the grid position based on the Q-Table. In each grid position, the average values tend to be predominantly negative. This is because every step taken by the agent will generate a -1 score

on top of the score of - (Manhattan distance between the box and goal) and - (Manhattan distance between the box and agent).

Blue areas, which are the more negative area of the graph appear beside the danger zones as the action can only take 1 action to enter the danger zone, which gives -1000 points. The dark blue part of the graph is where the agent can take 2 different actions to enter the danger zone. Hence, those areas have the most negative value. The area indicated in yellow has the most positive average values since the game will end immediate when the agent enters there.

There is no increase in average values at the goal state position as seen from figure 14 where the goal state grid is in blue colour. This is because the agent gets awarded +1000 point only when the **box** reaches the goal state, not when the **agent** reaches the goal state.

4. Policy Visualization

```
[[b'_' b'_' b'_' b'_' b'_' b'_' b'x' b'x' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'x' b'x' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'x' b'_' b'_' b'x' b'x' b'_' b'_' b'_' b'_' b'x' b'_']
 [b'_' b'_' b'_' b'x' b'_' b'_' b'x' b'_' b'_' b'_' b'_' b'x' b'x' b'_']
 [b'_' b'B' b'_' b'x' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'x' b'x' b'G']
 [b'A' b'_' b'_' b'x' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'x' b'x' b'_']]
step: 1, state: (5, 1, 4, 1), action: 4, reward: -14
Action: 4
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'B' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'A' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 2, state: (4, 1, 3, 1), action: 1, reward: -15
Action: 1
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'B' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'A' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 3, state: (3, 1, 2, 1), action: 1, reward: -16
Action: 1
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'B' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'A' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 4, state: (2, 1, 1, 1), action: 1, reward: -17
Action: 1

[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'B' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'A' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
```

```
step: 5, state: (2, 0, 1, 1), action: 3, reward: -18
Action: 3
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'B' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'A' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 6, state: (1, 0, 1, 1), action: 1, reward: -17
Action: 1
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'A' b'B' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 7, state: (1, 1, 1, 2), action: 4, reward: -16
Action: 4
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'A' b'B' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 8, state: (1, 2, 1, 3), action: 4, reward: -15
Action: 4

[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'A' b'B' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 9, state: (1, 3, 1, 4), action: 4, reward: -14
Action: 4
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'A' b'B' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 10, state: (1, 4, 1, 5), action: 4, reward: -13
Action: 4
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'A' b'B' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 11, state: (0, 4, 1, 5), action: 1, reward: -14
Action: 1
[[b'_' b'_' b'_' b'_' b'A' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'B' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
```

```
step: 12, state: (0, 5, 1, 5), action: 4, reward: -13
Action: 4
[[b'_' b'_' b'_' b'_' b'_' b'A' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'B' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 13, state: (1, 5, 2, 5), action: 2, reward: -12
Action: 2
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'A' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'B' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 14, state: (2, 5, 3, 5), action: 2, reward: -11
Action: 2
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'A' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'B' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 15, state: (3, 5, 4, 5), action: 2, reward: -10
Action: 2
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'A' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'B' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 16, state: (3, 4, 4, 5), action: 3, reward: -11
Action: 3
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'A' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'B' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 17, state: (4, 4, 4, 5), action: 2, reward: -10
Action: 2
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'A' b'B' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 18, state: (4, 5, 4, 6), action: 4, reward: -9
Action: 4
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'A' b'B' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
```

```
step: 19, state: (4, 6, 4, 7), action: 4, reward: -8
Action: 4
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'A' b'B' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 20, state: (4, 7, 4, 8), action: 4, reward: -7
Action: 4
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'A' b'B' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 21, state: (4, 8, 4, 9), action: 4, reward: -6
Action: 4
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'A' b'B' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 22, state: (4, 9, 4, 10), action: 4, reward: -5
Action: 4
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'A' b'B' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 23, state: (5, 9, 4, 10), action: 2, reward: -6
Action: 2
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'B' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'A' b'_' b'_' b'_']]
step: 24, state: (5, 10, 4, 10), action: 4, reward: -5
Action: 4
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'B' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'A' b'_' b'_' b'_']]
step: 25, state: (4, 10, 3, 10), action: 1, reward: -6
Action: 1
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'B' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'A' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
```

```
step: 26, state: (3, 10, 2, 10), action: 1, reward: -7
Action: 1
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'B' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'A' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 27, state: (2, 10, 1, 10), action: 1, reward: -8
Action: 1
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'B' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'A' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 28, state: (2, 9, 1, 10), action: 3, reward: -9
Action: 3
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'B' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'A' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 29, state: (1, 9, 1, 10), action: 1, reward: -8
Action: 1
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'A' b'B' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 30, state: (1, 10, 1, 11), action: 4, reward: -7
Action: 4
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'A' b'B' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 31, state: (1, 11, 1, 12), action: 4, reward: -6
Action: 4
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'A' b'B' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 32, state: (1, 12, 1, 13), action: 4, reward: -5
Action: 4
```

```
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'A' b'B']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 33, state: (0, 12, 1, 13), action: 1, reward: -6
Action: 1
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'A' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'B']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 34, state: (0, 13, 1, 13), action: 4, reward: -5
Action: 4
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'A']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'B']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 35, state: (1, 13, 2, 13), action: 2, reward: -4
Action: 2
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'A']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'B']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 36, state: (2, 13, 3, 13), action: 2, reward: -3
Action: 2
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'A']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'B']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
step: 37, state: (3, 13, 4, 13), action: 2, reward: 998
Action: 2
[[b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'A']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'B']
 [b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_' b'_']]
Total reward: 642
Action history: [4, 1, 1, 1, 3, 1, 4, 4, 4, 4, 1, 4, 2, 2, 2, 3, 2, 4, 4,

    4, 4, 4, 2, 4, 1, 1, 1, 3, 1, 4, 4, 4, 1, 4, 2, 2, 2]
```

Figure 13: Policy visualization of SARSA agent

The figure above shows the policy visualization of the SARSA agent at epsilon, alpha and gamma values as per the Bayesian Optimization and number of episodes=2000 from the start position and can push the box to the goal state with 37 actions and got the maximum reward of 642.

5. Bonus Component

5.1 Greedy and Non-Greedy strategy for Q-Learning and SARSA with exploration

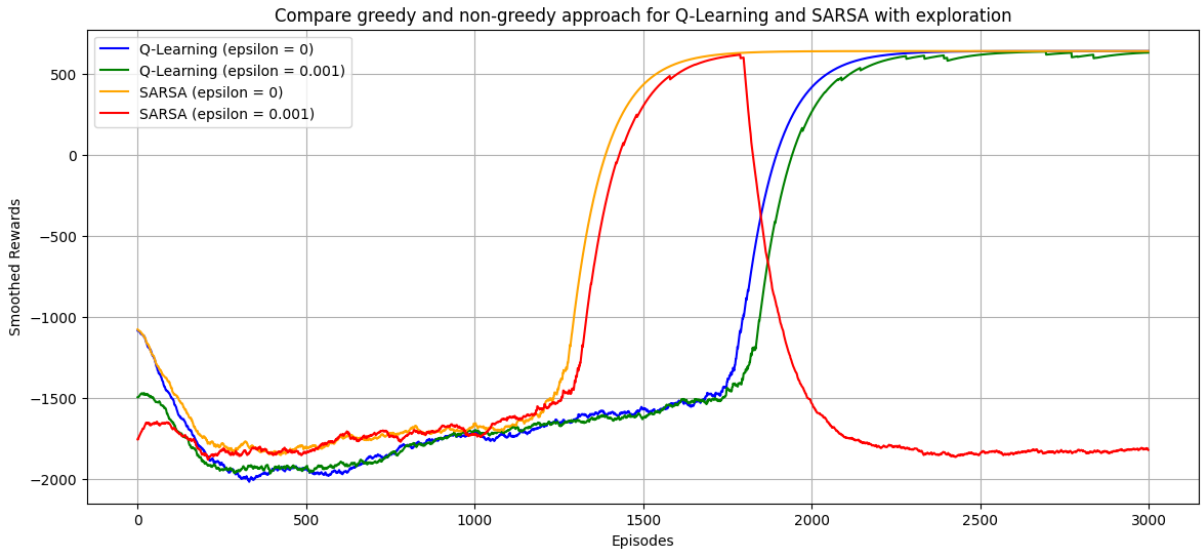

Figure 14: Compare Greedy and Non-Greedy strategy for Q-Learning and SARSA with exploration

Epsilon=0 signifies a complete absence of an exploration strategy. As observed in the above figure, the blue and yellow graphs behave as if there is no exploration strategy implemented, while only the red and green graphs incorporate an exploration strategy.

For the Q-Learning algorithm, both epsilon values of 0 and 0.001 result in the smoothed rewards reaching high values. However, Q-Learning when epsilon=0.001 is unstable as even after numerous episodes, we see the smoother rewards jittering up and down from episode 2000 to 3000. It can be concluded the epsilon-greedy strategy employed in Q-Learning leads to stabilization at a high smoothed reward value.

For SARSA, the greedy strategy allows the graph to stabilize at the maximum reward the fastest. When epsilon=0.001 for SARSA algorithm, the graph holds very similar graph shape as the greedy strategy of SARSA until around episode 1750, there is a sharp decrease in smoothed rewards and the smoothed reward value eventually reaches negative as the episode increases.

In most scenarios, relying on a purely greedy strategy may not yield the best results, as the agent heavily depends on the currently explored space. However, in this specific game environment, the greedy strategy leads to the stabilization of smoothed rewards at their maximum value, whereas exploration strategies exhibit signs of instability. This phenomenon can be attributed to the small size of the search space, where exploration does not yield significant benefits.

5.2 Greedy and Non-Greedy strategy for Q-Learning and SARSA without exploration

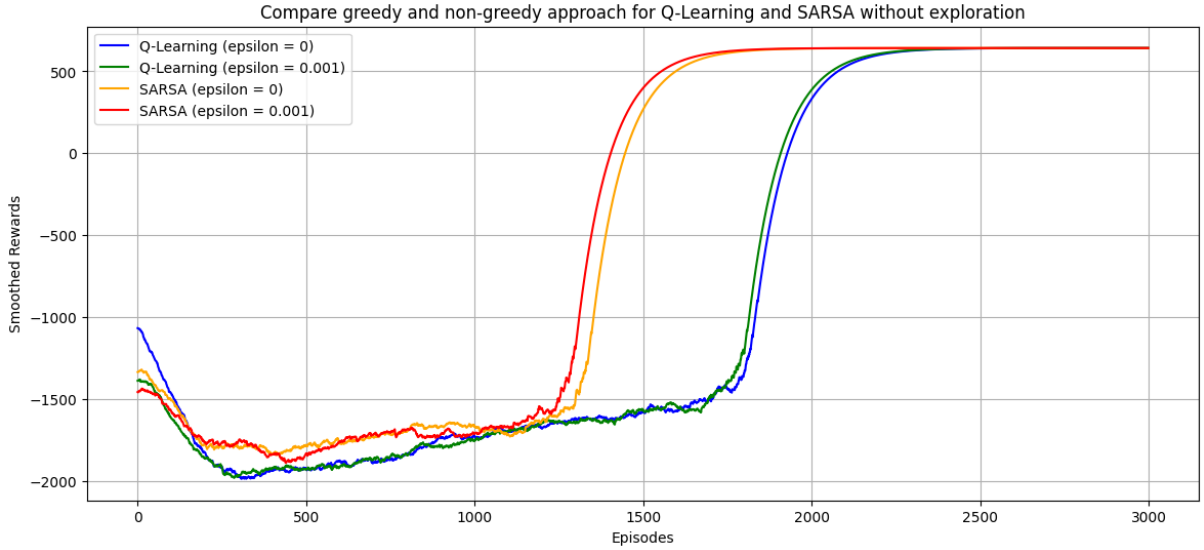

Figure 15: Compare Greedy and Non-Greedy strategy for Q-Learning and SARSA without exploration

When epsilon is set to 0, indicating the absence of an exploration strategy, all four graphs in the above figure operate as if epsilon is 0. This includes the green and red graphs, which had epsilon values of 0.001, as the exploration-related code has been completely removed.

Both Q-Learning graphs exhibit a similar pattern, which is conceptually correct since they both employ the same algorithm. Therefore, they produce similar smoothed rewards for each episode. The same principle applies to the SARSA graph.

Any slight distinctions observed in the two Q-Learning graphs and the two SARSA graphs can be attributed to the inherent randomness in reinforcement learning, resulting in variations in graph shapes between different runs.

5.3 Sparse Reward

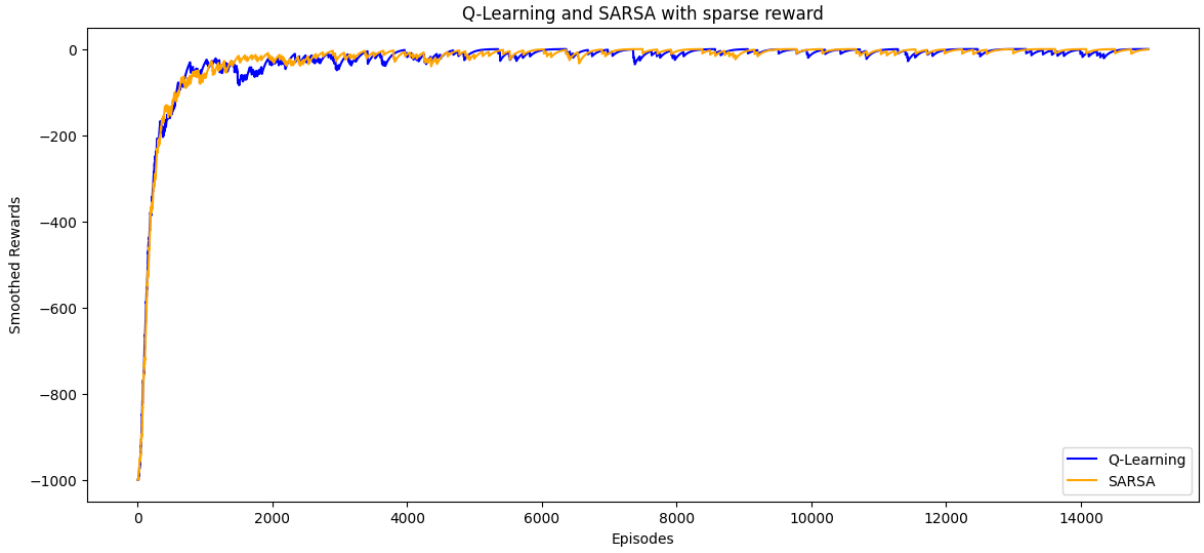

Figure 16: Q-Learning and SARSA with sparse reward using hyperparameters from Bayesian Optimization

Sparse Reward points conditions:

1. +1000 if agent reaches goal state.
2. -1000 if agent reaches danger zone.

Examining the figure depicted in figure 18, when sparse rewards are in place, both Q-Learning and SARSA algorithms converge to a reward value close to zero even after 15,000 episodes. This phenomenon occurs because sparse reward conditions significantly complicate reinforcement learning. The agent only receives reinforcement after accomplishing specific achievements, such as entering the goal state or the danger zone, rather than for each individual step. Consequently, the agent will face substantial challenges in discerning the correct sequence of actions required to reach the goal.

6. Conclusion

This report aims to identify the optimal hyperparameters for achieving the goal state using Bayesian Optimization for SARSA and Q-Learning. Given that both reinforcement learning algorithms yielded the same maximum reward of 642, it can be inferred that 642 is the highest achievable reward in this specific game problem. The V-Table based on the SARSA algorithm illustrates the average rewards the agent can expect when being at each grid position within the game world. Additionally, the Policy visualization showcases the path the SARSA agent followed to reach the goal state. Furthermore, the report explores both exploration and greedy strategies, as well as the implications of sparse rewards within this game context.