M22AI608- ADV_AI-ASSIGN1 ---- GRIDWORLD – Q LEARNING

Objective

Implement a Q-learning algorithm in a 5x5 grid world where an agent can move around to reach a goal state. The environment should have specific rewards and penalties for different states and actions.

Specifications

1. **Grid Size**: 5x5 grid world.
2. **Agent**: An agent that can move around the grid.
3. **Actions**: Four possible actions (Up, Down, Left, Right).
4. **Goal State**: A specific goal state within the grid.
5. **Rewards**:
    - Goal state: +5
    - Another terminal state (e.g., a trap): -5
    - Anywhere else: 0
    - Any action that takes the agent outside the boundary: -1
6. **Episodes**: Run 100,000 episodes.
7. **Random Seed**: Use a random number of seed for reproducibility.

Introduction:

This report examines how different values of the discount factor gamma affect the strategies developed by an agent using Q-learning in a GridWorld environment.

Grid Setup

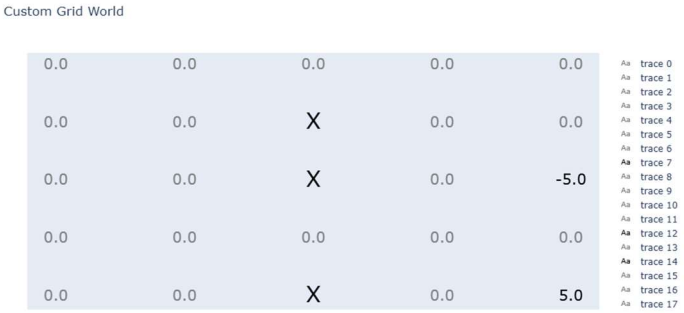
We conducted Q-learning experiments in a GridWorld setup with the following parameters:

- **Grid Size**: 5x5
- **Actions**: NORTH, SOUTH, WEST, EAST
- **Rewards**:
    - Goal Reward: +5
    - Terminal Reward: -5
    - Boundary Reward: -1
- **Learning Rate (Alpha)**: 0.1
- **Episodes**: 100,000

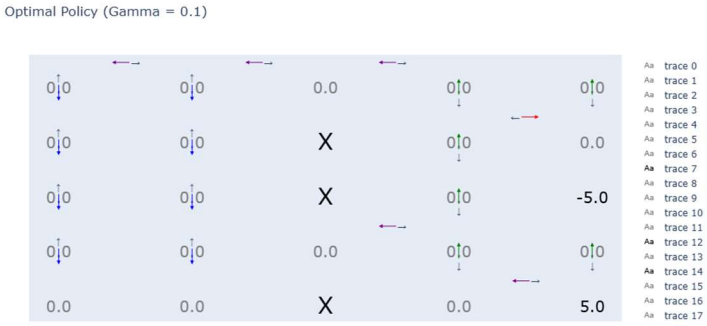
Q-learning Implementation:

- I used the Q-table to store the expected rewards for each action in every state.
- The three values of gamma (0.1, 0.5, 0.9) to observe their impact on the agent's decision-making process.
- Gamma determines how much weight the agent gives to future rewards compared to immediate rewards.
- The arrows on the grid visually represented the optimal action from each state, showing the path the agent would take to maximize rewards.
- Plots illustrated the expected cumulative reward from each state, highlighting the paths leading towards the goal state.

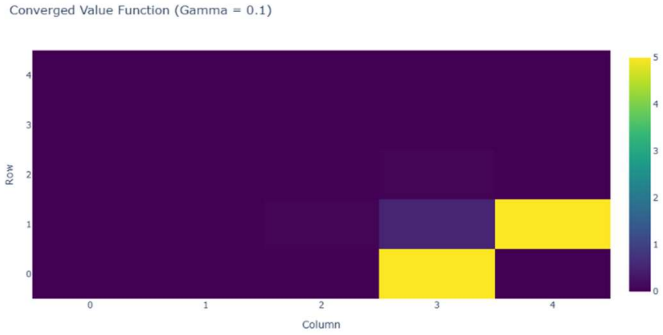
Custom grid Environment:



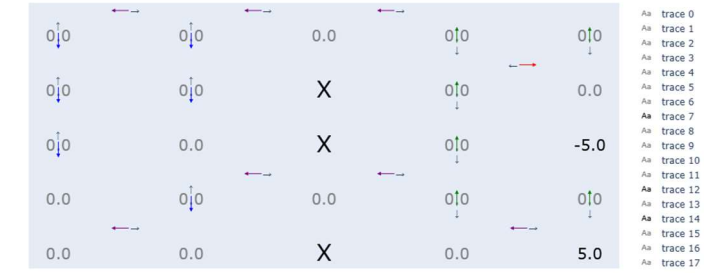
Optimal policy (Gamma = 0.1)



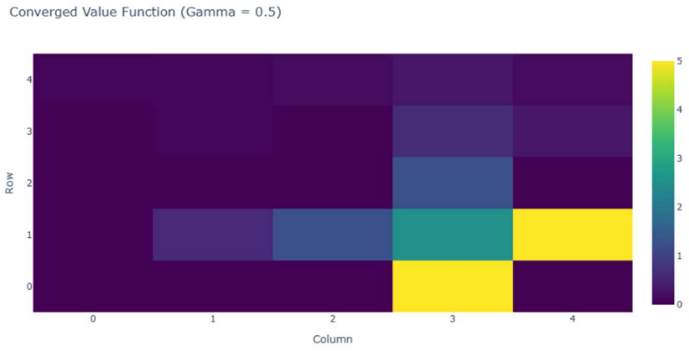
Converged value gamma=0.1



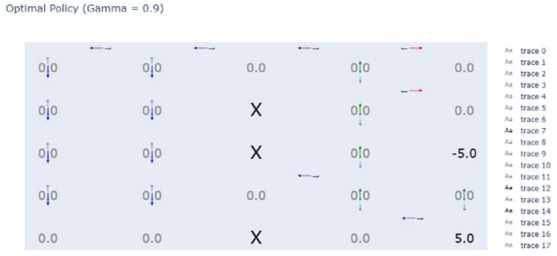
Optimal policy gamma =0.5



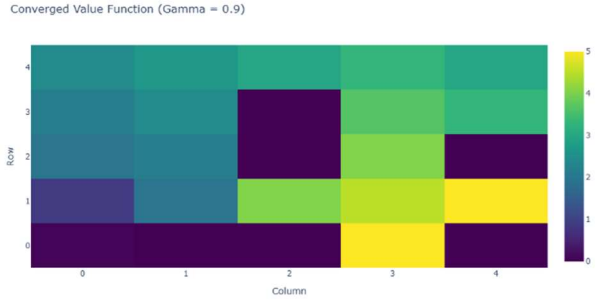
Converged value gamma=0.5



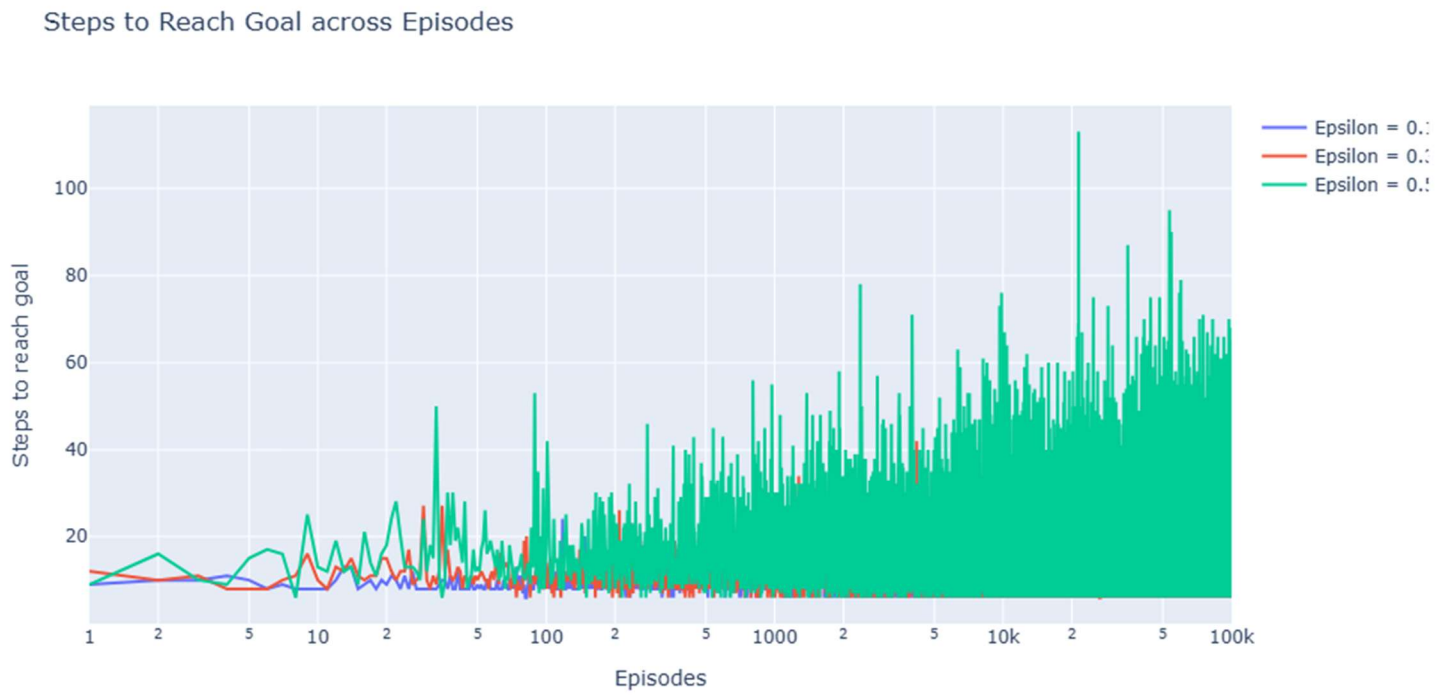
Converged value gamma=0.9



Optimal policy gamma =0.9



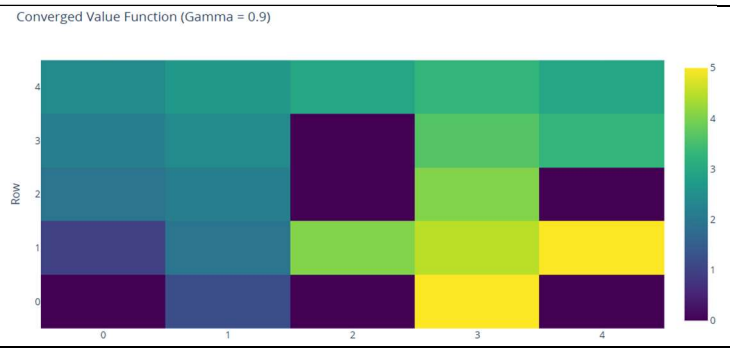
Steps to reach goal across Episodes:



Task 1: Plot the Converged Policy and Value Function for Epsilon = 0.1, alpha=0.1, Gamma = 0.9

Episode: 0, Goal reached: False
Episode: 10000, Goal reached: True
Episode: 20000, Goal reached: True
Episode: 30000, Goal reached: True
Episode: 40000, Goal reached: True
Episode: 50000, Goal reached: True
Episode: 60000, Goal reached: True
Episode: 70000, Goal reached: True
Episode: 80000, Goal reached: True
Episode: 90000, Goal reached: True
Total episodes where goal was reached: 94496/100000

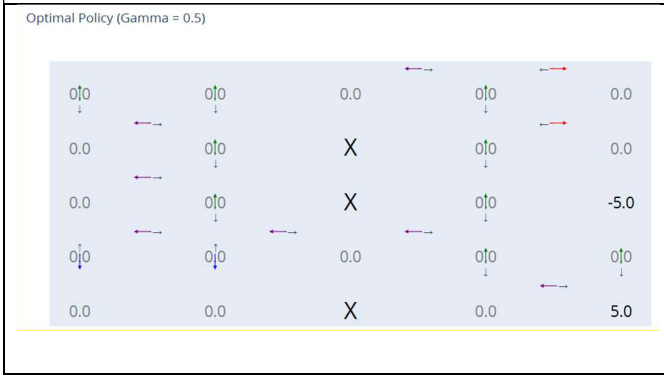
```
Q-values for Task 1 (Gamma = 0.9, Epsilon = 0.1):
[[[ 1.15233605  1.93710244  1.15233605  2.3914845 ]
  [ 1.3914845   2.15233605  2.15233605  2.657205  ]
  [ 1.657205    1.657205    2.3914845   2.95245   ]
  [ 1.95245     3.2805      2.657205    2.657205  ]
  [ 1.65686083  2.95063561  2.95245     1.6569469 ]]

 [[ 2.15233605  1.73487889  0.93667923  2.14551709]
  [ 2.3914845   1.936819    1.9369928   1.15098744]
  [ 0.          0.          0.          0.        ]
  [ 2.95245     3.645       2.2805      2.95245   ]
  [ 2.657205   -5.          3.2805      1.95126544]]

 [[ 1.93612342  0.1844705  -0.37944651  0.19371024]
  [ 2.15233605  1.27431617  0.92401301  0.06786362]
  [ 0.          0.          0.          0.        ]
  [ 3.2805      4.05        2.645      -5.        ]
  [ 0.          0.          0.          0.        ]]

 [[ 0.13627309  0.         -0.94607879  0.99254555]
  [ 1.93710244  0.47403148  0.16772766  1.66554146]
  [ 2.63934663  2.64418698  1.7433922   4.05      ]
  [ 3.645       4.5         3.645       4.5       ]
  [-5.          5.          4.05        3.5       ]]

 [[ 0.         -0.1        -0.40951     0.        ]
  [ 1.1954656  -0.6861894   0.         -0.5217031 ]
  [ 0.          0.          0.          0.        ]
  [ 4.05        3.5         3.5         5.        ]
  [ 0.          0.          0.          0.        ]]]
```

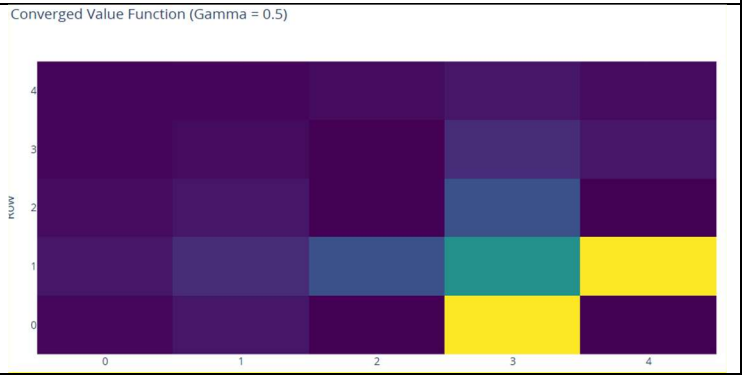


Task 2: Plot the Converged Policy and Value Function for Epsilon = 0.1, alpha=0.1,Gamma = 0.5

```
Episode: 0, Goal reached: False
Episode: 10000, Goal reached: True
Episode: 20000, Goal reached: True
Episode: 30000, Goal reached: True
Episode: 40000, Goal reached: True
Episode: 50000, Goal reached: True
Episode: 60000, Goal reached: True
Episode: 70000, Goal reached: True
Episode: 80000, Goal reached: True
Episode: 90000, Goal reached: True
Total episodes where goal was reached: 95263/100000
```

```
Q-values for Task 2 (Gamma = 0.5, Epsilon = 0.1):
[[[-0.98046875  0.0390625  -0.98046875  0.0390625 ]
  [-0.9609375   0.078125    0.01953125  0.078125  ]
  [-0.921875   -0.921875    0.0390625   0.15625   ]
  [-0.84375     0.3125      0.078125    0.078125  ]
  [-0.92188162  0.156133    0.15625    -0.92203108]]

 [[ 0.01953125  0.078125   -0.9609375   0.078125  ]
  [ 0.0390625   0.15625     0.0390625  -0.921875  ]
  [ 0.          0.          0.          0.        ]
  [ 0.15625     0.625      -0.6875      0.15625   ]
  [ 0.07812495 -5.          0.3125     -0.84465085]]

 [[ 0.02488866  0.08938321 -0.94975915  0.15625   ]
  [ 0.078125    0.3125      0.078125   -0.84375   ]
  [ 0.          0.          0.          0.        ]
  [ 0.3125      1.25       -0.375      -5.        ]
  [ 0.          0.          0.          0.        ]]

 [[ 0.06543126  0.006932   -0.91390064  0.3125    ]
  [ 0.15625     0.15625     0.15625     0.625     ]
  [-0.375      -0.375       0.3125      1.25      ]
  [ 0.625       2.5         0.625       2.5       ]
  [-5.          5.          1.25        1.5       ]]

 [[ 0.09620491 -0.271      -0.26730582  0.015625  ]
  [ 0.3125     -0.47935953  0.01636394 -0.68507797]
  [ 0.          0.          0.          0.        ]
  [ 1.25        1.5         1.5         5.        ]
  [ 0.          0.          0.          0.        ]]]
```

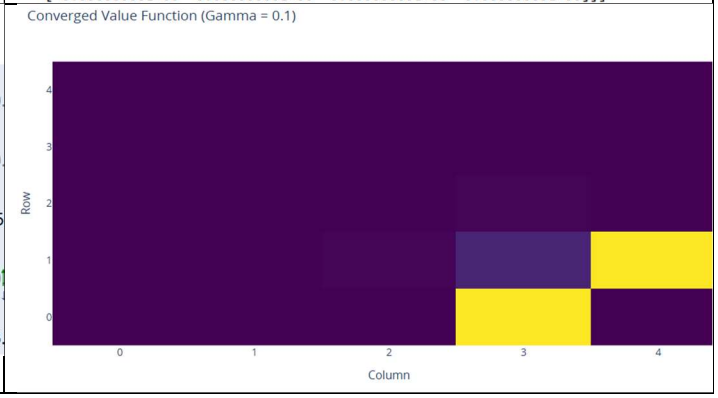




Task 2: Plot the Converged Policy and Value Function for Epsilon = 0.1, alpha=0.1,Gamma = 0.1

```
Episode: 0, Goal reached: False
Episode: 10000, Goal reached: True
Episode: 20000, Goal reached: True
Episode: 30000, Goal reached: True
Episode: 40000, Goal reached: True
Episode: 50000, Goal reached: True
Episode: 60000, Goal reached: True
Episode: 70000, Goal reached: True
Episode: 80000, Goal reached: True
Episode: 90000, Goal reached: True
Total episodes where goal was reached: 93789/100000
```

```
Q-values for Task 2 (Gamma = 0.1, Epsilon = 0.1):
[[[-9.99999950e-01  5.00000000e-07 -9.99999950e-01  5.00000000e-07]
  [-9.99999500e-01  5.00000000e-06  5.00000000e-08  5.00000000e-06]
  [-9.99995000e-01 -9.99995000e-01  5.00000000e-07  5.00000000e-05]
  [-9.99950000e-01  5.00000000e-04  5.00000000e-06  5.00000000e-06]
  [-9.99995001e-01  4.98406066e-05  5.00000000e-05 -9.99995005e-01]]

 [[ 5.00000000e-08  5.00000000e-06 -9.99999500e-01  5.00000000e-06]
  [ 5.00000000e-07  5.00000000e-05  5.00000000e-07 -9.99995000e-01]
  [ 0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00]
  [ 5.00000000e-05  5.00000000e-03 -9.99500000e-01  5.00000000e-05]
  [ 5.00000000e-06 -5.00000000e+00  5.00000000e-04 -9.99950210e-01]]

 [[ 4.93508309e-07  3.67430296e-05 -9.99996846e-01  5.00000000e-05]
  [ 5.00000000e-06  5.00000000e-04  5.00000000e-06 -9.99950000e-01]
  [ 0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00]
  [ 5.00000000e-04  5.00000000e-02 -9.95000000e-01 -5.00000000e+00]
  [ 0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00]]

 [[ 3.55405124e-06  1.29360417e-06 -9.99976909e-01  5.00000000e-04]
  [ 5.00000000e-05  5.00000000e-05  5.00000000e-05  5.00000000e-03]
  [-9.95000000e-01 -9.95000000e-01  5.00000000e-04  5.00000000e-02]
  [ 5.00000000e-03  5.00000000e-01  5.00000000e-03  5.00000000e-01]
  [-5.00000000e+00  5.00000000e+00  5.00000000e-02 -5.00000000e-01]]

 [[ 3.63675216e-05 -2.71000000e-01 -4.09510000e-01  0.00000000e+00]
  [ 5.00000000e-04 -7.71218105e-01  8.29572055e-07 -7.94080395e-01]
  [ 0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00]
  [ 5.00000000e-02 -5.00000000e-01 -5.00000000e-01  5.00000000e+00]
  [ 0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00]]]
```



Optimal Policies Analysis

Gamma = 0.9:

Characteristics: Gamma value of 0.9 prioritizes long-term rewards, emphasizing future gains over immediate rewards. The agent is willing to take longer paths if they promise a higher cumulative reward, aiming directly towards the goal even if it means taking more steps.

Movement Pattern: The movement pattern involves moving down (South) and then right (East) towards the goal, as these actions incrementally approach the goal state while potentially avoiding obstacles or terminal states.

Gamma = 0.5:

Characteristics: A Gamma value of 0.5 balances immediate gains with future rewards, reflecting a more moderate approach between short-term and long-term objectives. The agent considers both the immediate rewards and the potential future rewards but with less emphasis on distant outcomes compared to Gamma = 0.9.

Movement Pattern:

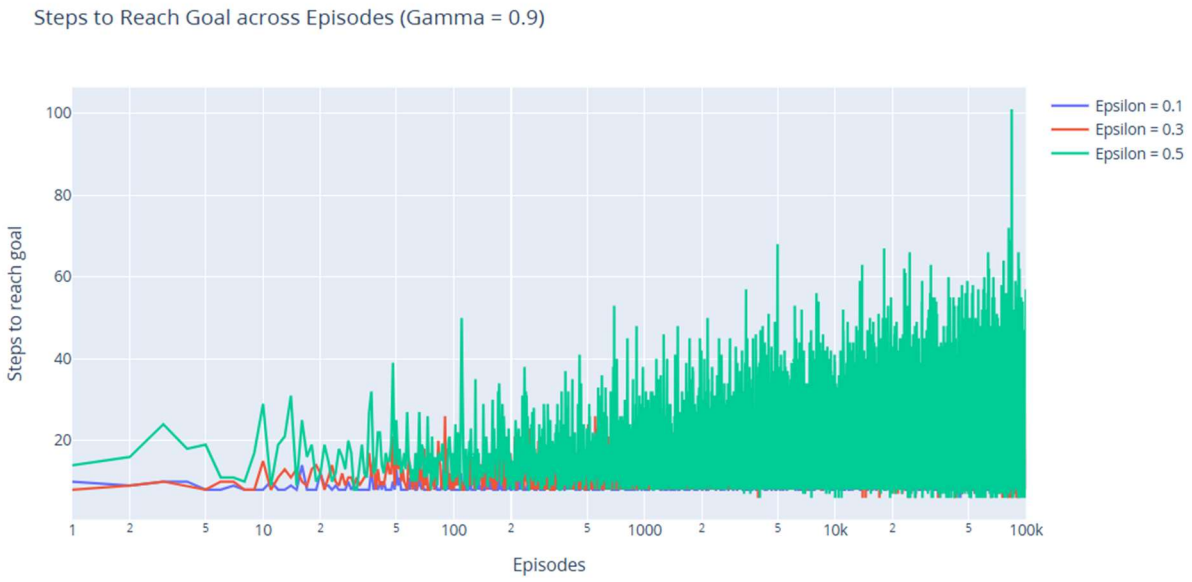
The movement pattern tends to prioritize safer routes towards the goal, avoiding unnecessary risks while still aiming to reach the goal state.

Gamma = 0.1:

Characteristics: With a Gamma value of 0.1, the agent focuses primarily on short-term rewards, prioritizing immediate gains and penalties avoidance over long-term planning. The agent makes decisions based predominantly on the immediate rewards associated with each action, with little consideration for future rewards or penalties.

Movement Pattern: The movement pattern for Gamma = 0.1 is localized and reactive, focusing on immediate gains and penalties. This behavior results in a more reactive and short-sighted approach compared to higher gamma values, where the agent might sacrifice immediate rewards for higher long-term gains.

Task3: For gamma = 0.9, plot the no. of steps to reach the goal across episodes for epsilon = 0.1, 0.3 and 0.5.



Epsilon = 0.5 : The number of steps to reach the goal increases significantly as episodes progress. This is because an epsilon value of 0.5 encourages extensive exploration, causing the agent to frequently deviate from known paths in search of potentially better strategies.

The variance is high plot shows frequent spikes, indicating substantial variability in the number of steps taken in each episode.

Epsilon = 0.3 : The increase in the number of steps is less steep compared to epsilon = 0.5. A moderate epsilon value of 0.3 still encourages exploration but to a lesser extent, allowing the agent to balance between exploration and exploitation more effectively.

The variance is Moderate and plot shows fewer spikes compared to epsilon = 0.5, indicating relatively stable performance.

Epsilon = 0.1 : The number of steps remains relatively low and increases at a slower rate compared to higher epsilon values

The variance is Low and plot shows minimal spikes, signifying consistent behavior in reaching the goal. The agent tends to exploit known optimal paths rather than exploring new ones.

----------Thank You----------