MDPs and RL

The reinforcement learning algorithm chosen is Q Learning. The two MDPs chosen for this assignment are Frozen Lake and Forest Management. Both are available in mdptoolbox and were chosen primarily because of the ease of use. I can also see examples related to them applied every day, such as the Roomba which runs in the house and the need to manage forests to combat global warming, maintain wildlife and to prevent wildfires.

Frozen Lake:

It is a grid world problem which has four possible actions – up, down, left and right. The agent moves around the grid until it reaches the goal or falls in the hole. If it falls in the hole, it has to start over and is awarded 0 reward. Following links are used as a reference:

- https://towardsdatascience.com/value-iteration-to-solve-openai-gyms-frozenlake-6c5e7bf0a64d
- https://github.com/adodd202/GT-ML-Assignment4/blob/main/Frozen%20Lake%20Analysis.ipynb

Forest Management:

It is a discrete non grid world problem which has two possible actions – wait and cut. An action is decided each year with the first objective to maintain an old forest for wildlife and second to make money selling wood. Following link is used as a reference:

- https://medium.com/sequential-learning/optimistic-q-learning-b9304d079e11

Q Learning:

In this algorithm, the goal is to iteratively learn the optimal Q value function using the Bellman Optimality Equation. To prevent Q value function to converge on a local optimum, exponential decay is added. Learning rate alpha is used to help with convergence.

Frozen Lake

Three grids of varying sizes are generated via the random map generator and run. They are 4x4, 8x8 and 25x25. The results for 8x8 and 25x25 are plotted below. The plots for 4x4 are captured under plots/FrozenLake folder. The number of episodes is kept constant at 1000 and max iterations are set to 1000, 3000 and 10000 respectively.

- Both value iteration and policy iteration converge to the same optimum policy.
- As the MDP size increases, the time taken by value iteration is longer. Policy iteration converges faster than value iteration for larger MDPs. This is due to the value iteration needing to find the maximum value in each iteration.
- Q Learning Q value does not converge to the same as value and policy iterations. Balancing between exploration and exploitation has significant impact on the agent's learning performance. Epsilon was set with a decay to handle explore-exploit impact. The error is high in the initial phase as the agent is exploring and the error gradually reduces as it learns more about the environment and as the decay is applied. Alphas is set to 0.1 and works fine for smaller

MDPs. However, for larger MDPs such as the 25x25 grid, a higher alpha seems to be needed. The 25x25 grid episodes and max iteration values need to be revisited. The current setup is not optimal.

Optimal Policy Results:

4x4 Value Iteration Policy: (0, 0, 0, 0, 1, 0, 3, 1, 1, 0, 0, 0, 1, 1, 1, 0)
4x4 Policy Iteration Policy: (0, 0, 0, 0, 1, 0, 3, 1, 1, 0, 0, 0, 1, 1, 1, 0)
4x4 Q Learning Policy Gamma=0.95: (1, 0, 1, 0, 2, 0, 3, 0, 2, 2, 0, 0, 1, 1, 1, 0)
4x4 Q Learning Policy Gamma=0.99: (1, 0, 0, 0, 0, 0, 3, 1, 2, 1, 0, 0, 1, 1, 1, 0)

8x8 Value Iteration Policy: (3, 2, 2, 2, 2, 2, 2, 2, 3, 3, 2, 2, 2, 2, 1, 1, 0, 0, 2, 2, 2, 1, 1, 1, 3, 1, 3, 2, 2, 1, 1, 1, 3, 0, 0, 2, 2, 2, 1, 1, 3, 0, 0, 2, 2, 2, 1, 1, 0, 0, 1, 2, 3, 2, 1, 0, 0, 0, 0, 0, 0, 2, 1, 0)
8x8 Policy Iteration Policy: (3, 2, 2, 2, 2, 2, 2, 2, 3, 3, 2, 2, 2, 2, 1, 1, 0, 0, 2, 2, 2, 1, 1, 1, 3, 1, 3, 2, 2, 1, 1, 1, 3, 0, 0, 2, 2, 2, 1, 1, 3, 0, 0, 2, 2, 2, 1, 1, 0, 0, 1, 2, 3, 2, 1, 0, 0, 0, 0, 0, 0, 2, 1, 0)
8x8 Q Learning Policy Gamma=0.95: (1, 1, 3, 1, 1, 3, 0, 1, 3, 1, 1, 3, 2, 2, 1, 2, 0, 0, 1, 1, 1, 0, 3, 0, 1, 3, 3, 0, 3, 1, 0, 0, 0, 0, 0, 1, 2, 3, 1, 0, 0, 0, 0, 0, 2, 2, 2, 0, 0, 0, 2, 3, 2, 1, 2, 0, 0, 0, 0, 0, 0, 2, 1, 0)
8x8 Q Learning Policy Gamma=0.99: (2, 0, 2, 3, 1, 0, 0, 0, 3, 1, 3, 3, 3, 0, 2, 0, 0, 0, 0, 2, 2, 2, 0, 0, 0, 3, 3, 3, 2, 0, 1, 0, 3, 0, 0, 0, 0, 3, 1, 0, 0, 0, 0, 1, 3, 3, 2, 0, 0, 0, 3, 1, 3, 3, 1, 0, 0, 0, 0, 2, 0, 2, 0, 0)

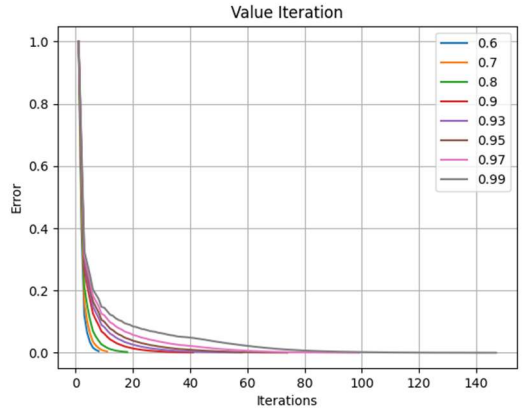
25x25 Value Iteration Policy: (3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 0, 0, 0, 0, 0, 2, 2, 2, 2, 2, 2, 2, 3, 2, 2, 3, 2, 2, 2, 3, 3, 3, 3, 3, 3, 2, 1, 1, 0, 0, 0, 0, 2, 2, 3, 3, 0, 0, 2, 1, 3, 2, 2, 3, 3, 3, 3, 0, 0, 0, 2, 1, 1, 1, 1, 1, 1, 2, 0, 0, 2, 2, 1, 1, 1, 3, 2, 0, 0, 0, 2, 0, 0, 1, 1, 2, 3, 2, 1, 3, 3, 3, 2, 0, 0, 0, 2, 3, 2, 1, 3, 3, 3, 1, 1, 0, 0, 0, 2, 1, 0, 0, 2, 0, 0, 0, 0, 2, 2, 1, 1, 0, 0, 2, 1, 0, 0, 0, 2, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 2, 2, 3, 2, 0, 0, 2, 3, 0, 0, 1, 2, 1, 1, 1, 1, 1, 3, 3, 2, 1, 1, 1, 0, 0, 2, 0, 0, 2, 1, 1, 0, 0, 0, 1, 3, 2, 1, 1, 1, 0, 0, 0, 0, 2, 2, 1, 3, 1, 0, 0, 2, 1, 1, 1, 1, 0, 1, 1, 0, 0, 2, 1, 1, 1, 0, 3, 1, 1, 3, 3, 0, 0, 2, 1, 1, 1, 1, 1, 1, 1, 0, 0, 3, 1, 1, 2, 1, 1, 1, 0, 0, 0, 2, 0, 0, 2, 1, 1, 3, 3, 2, 2, 2, 2, 2, 1, 1, 1, 0, 2, 2, 2, 1, 1, 1, 0, 1, 1, 0, 0, 1, 2, 1, 0, 0, 0, 2, 2, 3, 3, 2, 2, 1, 1, 2, 2, 2, 2, 1, 1, 0, 0, 0, 2, 1, 3, 2, 1, 0, 0, 1, 2, 0, 0, 0, 2, 2, 1, 0, 2, 2, 2, 2, 2, 1, 0, 1, 0, 0, 0, 0, 2, 1, 0, 0, 2, 3, 0, 1, 1, 2, 3, 2, 1, 3, 3, 3, 2, 2, 1, 1, 0, 0, 1, 0, 0, 2, 1, 1, 1, 0, 0, 2, 0, 0, 0, 0, 2, 0, 0, 2, 0, 0, 1, 2, 0, 0, 2, 1, 1, 3, 1, 1, 0, 0, 2, 3, 3, 3, 3, 2, 2, 2, 2, 2, 2, 1, 1, 1, 3, 1, 1, 1, 1, 3, 0, 2, 0, 0, 1, 2, 0, 0, 0, 0, 2, 3, 3, 3, 2, 2, 2, 1, 0, 0, 2, 1, 1, 0, 0, 0, 0, 0, 0, 2, 3, 3, 1, 3, 1, 3, 3, 0, 0, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 2, 3, 3, 0, 0, 0, 2, 1, 2, 3, 3, 2, 2, 2, 1, 1, 1, 1, 3, 3, 3, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 2, 0, 0, 2, 2, 2, 1, 1, 1, 1, 1, 3, 0, 0, 2, 1, 0, 0, 0, 1, 0, 0, 2, 0, 0, 0, 0, 1, 2, 3, 2, 1, 1, 1, 3, 1, 3, 3, 1, 0, 3, 0, 0, 0, 0, 2, 0, 0, 2, 0, 0, 1, 2, 0, 0, 2, 1,

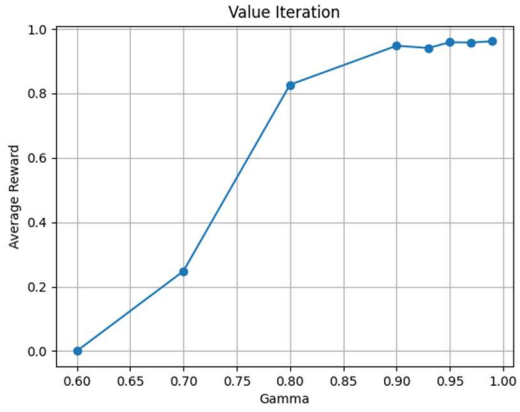
25x25 Policy Iteration Policy: (3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 0, 0, 0, 0, 0, 2, 2, 2, 2, 2, 2, 2, 3, 2, 2, 3, 2, 2, 2, 3, 3, 3, 3, 3, 3, 2, 1, 1, 0, 0, 0, 0, 2, 2, 3, 3, 0, 0, 2, 1, 3, 2, 2, 3, 3, 3, 3, 0, 0, 0, 2, 1, 1, 1, 1, 1, 1, 2, 0, 0, 2, 2, 1, 1, 1, 3, 2, 0, 0, 0, 2, 0, 0, 1, 1, 2, 3, 2, 1, 3, 3, 3, 2, 0, 0, 0, 2, 3, 2, 1, 3, 3, 3, 1, 1, 0, 0, 0, 2, 1, 0, 0, 2, 0, 0, 0, 0, 2, 2, 1, 1, 0, 0, 2, 1, 0, 0, 0, 2, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 2, 2, 3, 2, 0, 0, 2, 3, 0, 0, 1, 2, 1, 1, 1, 1, 1, 3, 3, 2, 1, 1, 1, 0, 0, 2, 0, 0, 2, 1, 1, 0, 0, 0, 1, 3, 2, 1, 1, 1, 0, 0, 0, 0, 2, 2, 1, 3, 1, 0, 0, 2, 1, 1, 1, 1, 0, 1, 1, 0, 0, 2, 1, 1, 1, 0, 3, 1, 1, 3, 3, 0, 0, 2, 1, 1, 1, 1, 1, 1, 1, 0, 0, 3, 1, 1, 2, 1, 1, 1, 0, 0, 0, 2, 0, 0, 2, 1, 1, 3, 3, 2, 2, 2, 2, 2, 1, 1, 1, 0, 2, 2, 2, 1, 1, 1, 0, 1, 1, 0, 0, 1, 2, 1, 0, 0, 0, 2, 2, 3, 3, 2, 2, 1, 1, 2, 2, 2, 2, 1, 1, 0, 0, 0, 2, 1, 3, 2, 1, 0, 0, 1, 2, 0, 0, 0, 2, 2, 1, 0, 2, 2, 2, 2, 2, 1, 0, 1, 0, 0, 0, 0, 2, 1, 0, 0, 2, 3, 0, 1, 1, 2, 3, 2, 1, 3, 3, 3, 2, 2, 1, 1, 0, 0, 1, 0, 0, 2, 1, 1, 1, 0, 0, 2, 0, 0, 0, 0, 2, 0, 0, 2, 0, 0, 1, 2, 0, 0, 2, 1, 1, 3, 1, 1, 0, 0, 2, 3, 3, 3, 3, 2, 2, 2, 2, 2, 2, 1, 1, 1, 3, 1, 1, 1, 1, 3, 0, 2, 0, 0, 1, 2, 0, 0, 0, 0, 2, 3, 3, 3, 2, 2, 2, 1, 0, 0, 2, 1, 1, 0, 0, 0, 0, 0, 0, 2, 3, 3, 1, 3, 1, 3, 3, 0, 0, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 2, 3, 3, 0, 0, 0, 2, 1, 2, 3, 3, 2, 2, 2, 1, 1, 1, 1, 3, 3, 3, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 2, 0, 0, 2, 2, 2, 1, 1, 1, 1, 1, 3, 0, 0, 2, 1, 0, 0, 0, 1, 0, 0, 2, 0, 0, 0, 0, 1, 2, 3, 2, 1, 1, 1, 3, 1, 3, 3, 1, 0, 3, 0, 0, 0, 0, 2, 0, 0, 2, 0, 0, 1, 2, 0, 0, 2, 1,

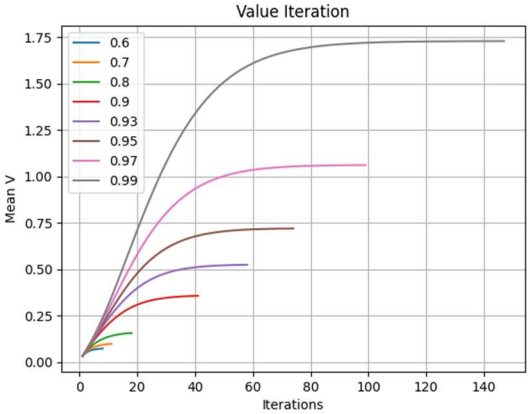
1, 0, 0, 2, 0, 0, 2, 0, 0, 2, 0, 0, 1, 2, 3, 0, 0, 0, 0, 0, 2, 0, 0, 2, 1, 1, 0, 0, 2, 0, 1, 2, 0, 1, 0, 0, 1, 3, 0, 0, 2, 1, 1,
1, 1, 3, 0, 0, 2, 2, 1, 1, 1, 0, 1, 1, 2, 0, 0, 0, 1, 0, 0, 2, 0, 0, 2, 2, 2, 0, 0, 0, 0, 2, 2, 2, 2, 1, 0)
25x25 Q Learning Policy Gamma=0.95: (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 2,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 2, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 2, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 1, 0,
0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 2, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 2, 1, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 3, 1, 1, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 1, 1, 1, 1, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 2, 1, 3,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
3, 0, 3, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 3, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 2, 3, 1, 0)
Q Learning Policy Gamma=0.99: (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 3, 1, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 2, 2, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 2, 3, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 1, 1, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 2, 1, 2, 0, 1, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 3, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 1, 1, 0, 1,
2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 2, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 2, 0, 1, 0, 0, 0, 0, 0,
0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 3, 1, 1, 3, 1, 1, 1, 3, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2,
1, 0, 0, 0, 2, 2, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 3, 3, 2, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 1, 1, 3, 1, 2, 3, 1, 3, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 2, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 3,
0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 3, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 2, 2, 1, 1, 0)

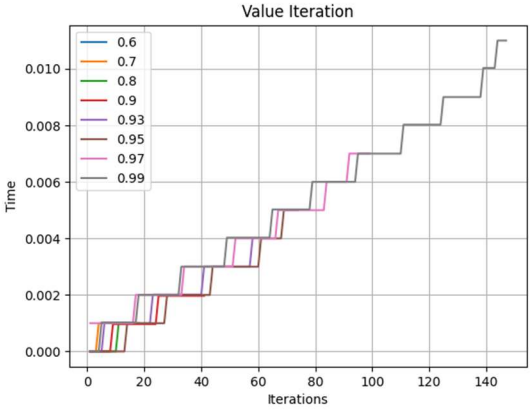
Value Iteration (8x8):

The following plots are based on varying values of gamma. Gamma = 0.99 was chosen based on the maximum rewards plotted after running the experiments.



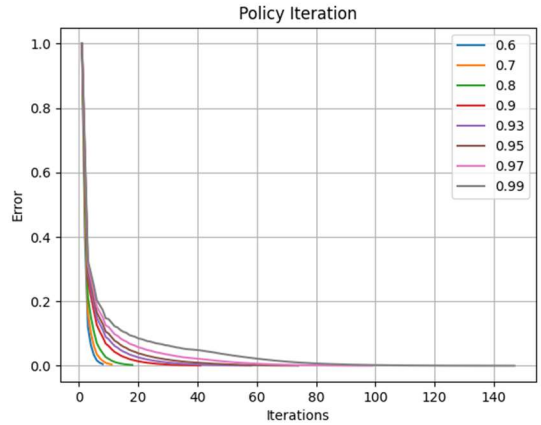


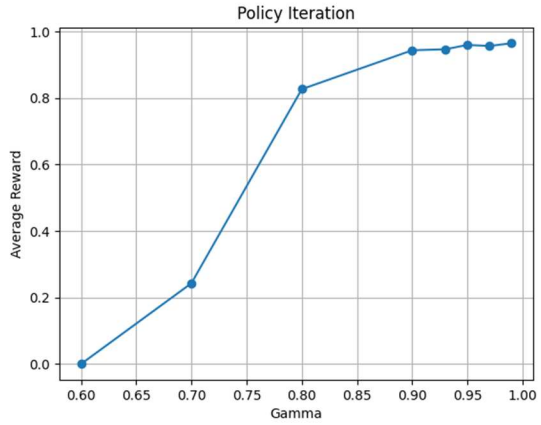


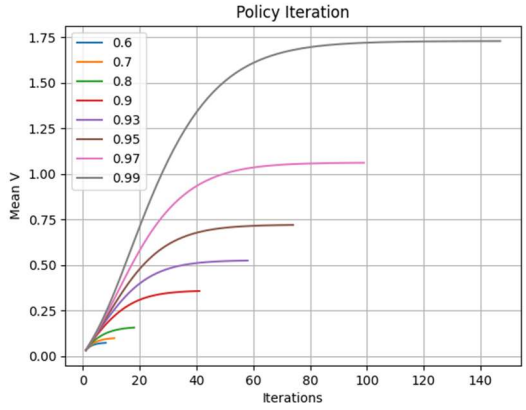


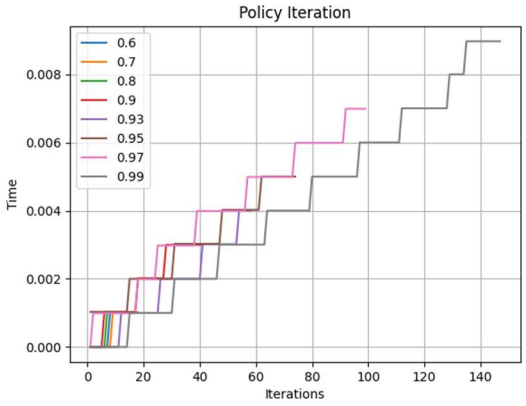
Policy Iteration (8x8):

The following plots are based on varying values of gamma. Gamma = 0.99 was chosen based on the maximum rewards plotted after running the experiments.



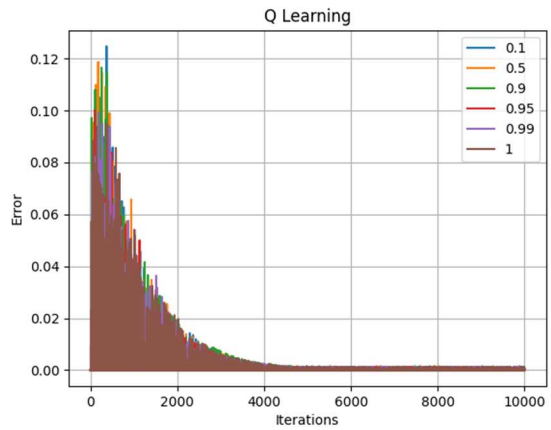


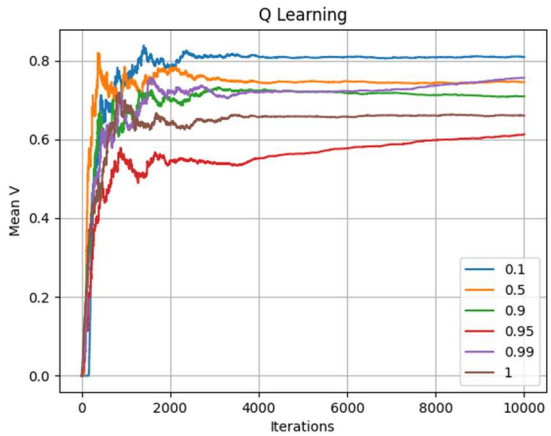




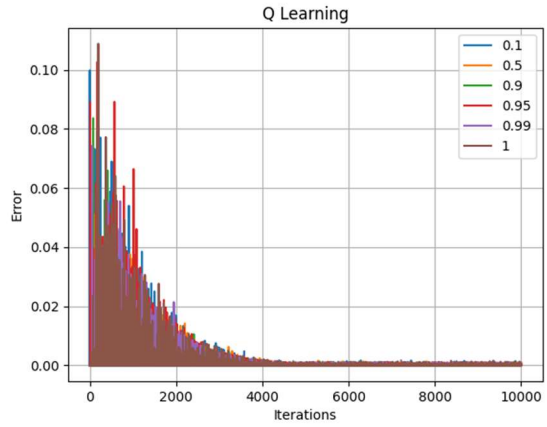
Q Learning (8x8):

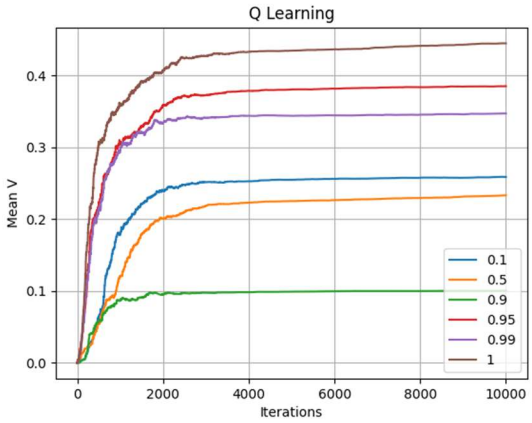
The following plots are based on varying values of alpha. Alpha = 0.1 is chosen for the experiments.



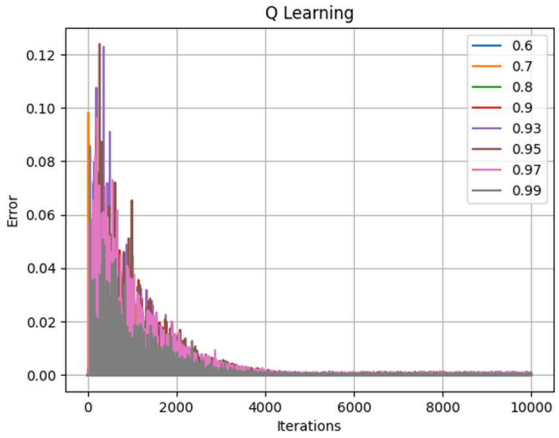


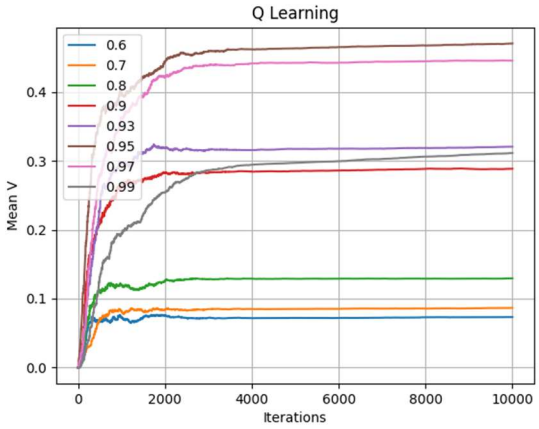
The following plots are based on varying values of epsilon. Epsilon = 0.95 is chosen for the experiment.





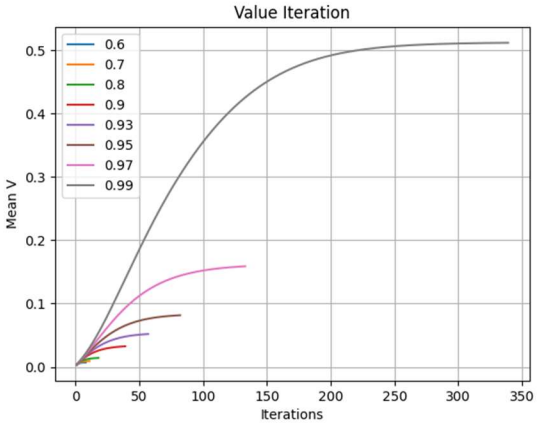
The following plots are based on varying values of gamma. Gamma = 0.99 is chosen for the experiment.

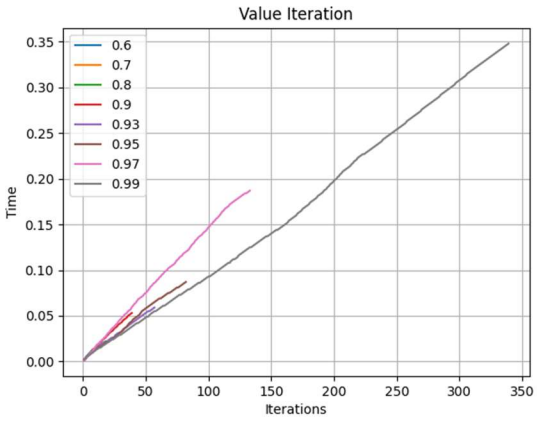




Value Iteration (25x25):

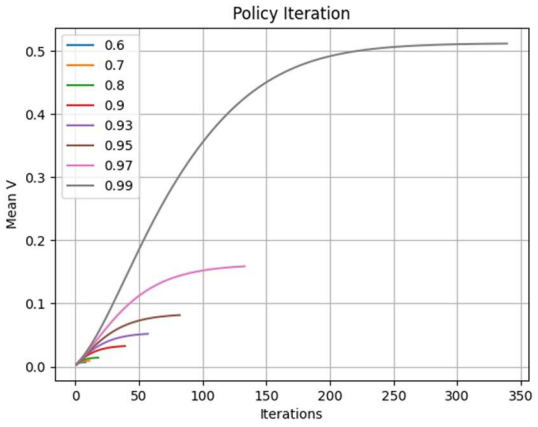
The following plots are based on varying values of gamma. Gamma = 0.99 was chosen based on the maximum rewards plotted after running the experiments.

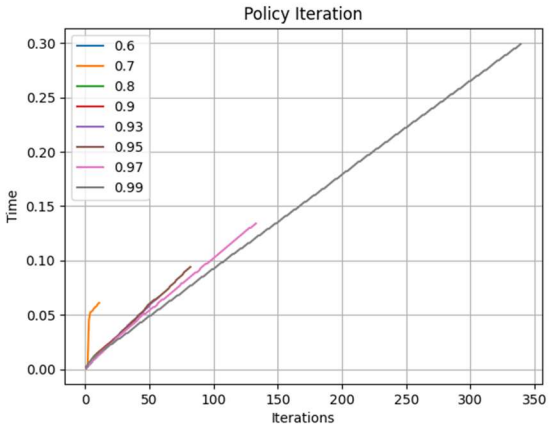




Policy Iteration (25x25):

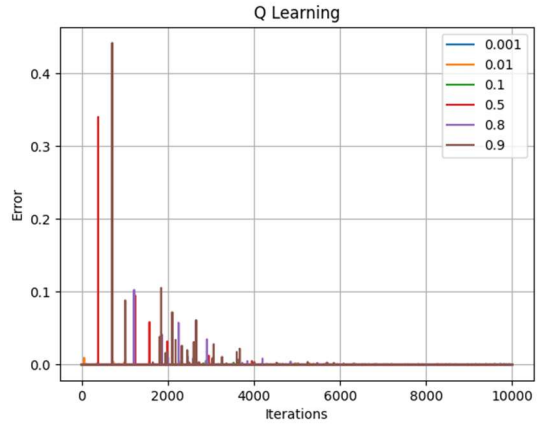
The following plots are based on varying values of gamma. Gamma = 0.99 was chosen based on the maximum rewards plotted after running the experiments.

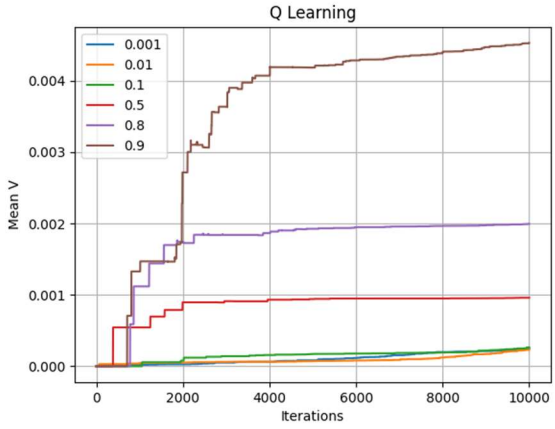




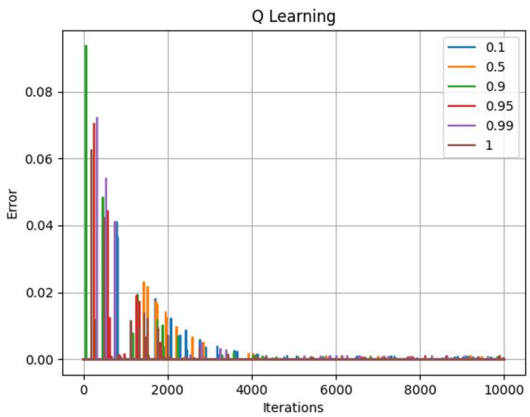
Q Learning (25x25):

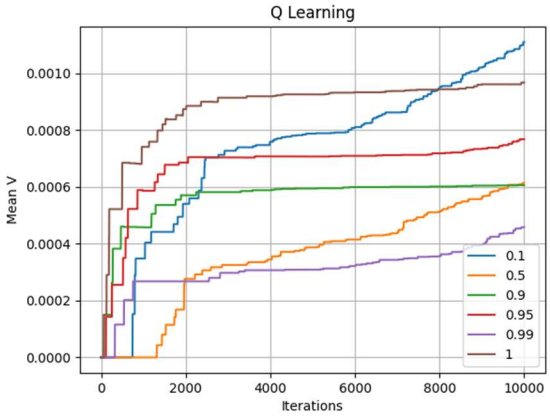
The following plots are based on varying values of alpha. Alpha = 0.1 is chosen for the experiments.



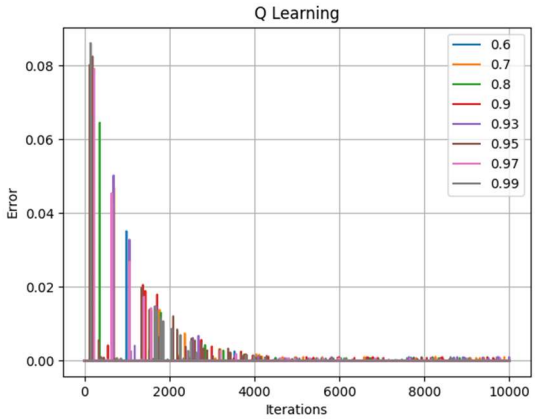


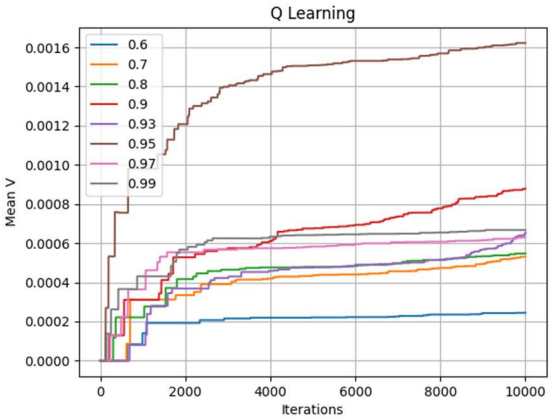
The following plots are based on varying values of epsilon. Epsilon = 0.95 is chosen for the experiment.





The following plots are based on varying values of gamma. Gamma = 0.99 is chosen for the experiment.





Forest

Three states of sizes 3, 6 and 2000 are run. The results for S = 2000 are plotted below. The plots for S = 3 and S = 6 are captured in the plots/Forest folder. The number of episodes is kept constant at 1000 and max iterations is set to 5000 for all three experiments. The reward for waiting is set to 10 and for cutting when past the older age barrier is set to 50.

- All three, value iteration, policy iteration and Q learning converge to the same optimum policy for S = 3 and 6. Q values do not match for S = 2000 however.
- As the MDP size increases, the time taken by value iteration is longer. Policy iteration converges faster than value iteration for larger MDPs. This is due to the value iteration needing to find the maximum value in each iteration.
- Balancing between exploration and exploitation has significant impact on the agent's learning performance. Epsilon was set with a decay to handle explore-exploit impact. The error is high in the initial phase as the agent is exploring and the error gradually reduces as it learns more about the environment and as the decay is applied.

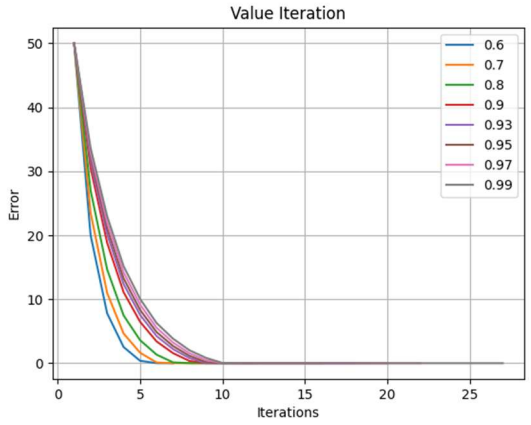
Optimal Policy Results:

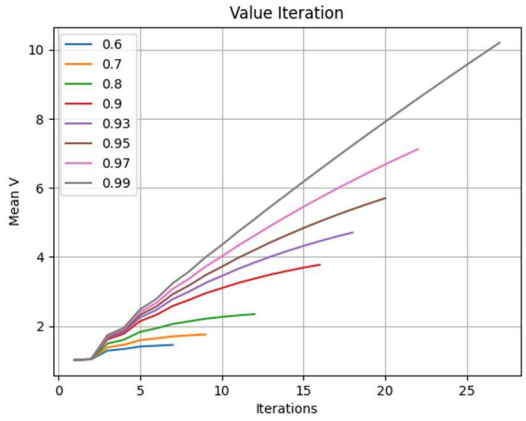
S = 3 Value Iteration Policy: (0, 0, 1)
S = 3 Policy Iteration Policy: (0, 0, 1)
S = 3 Q Learning Policy Gamma=0.95: (0, 0, 1)
S = 3 Q Learning Policy Gamma=0.99: (0, 0, 1)

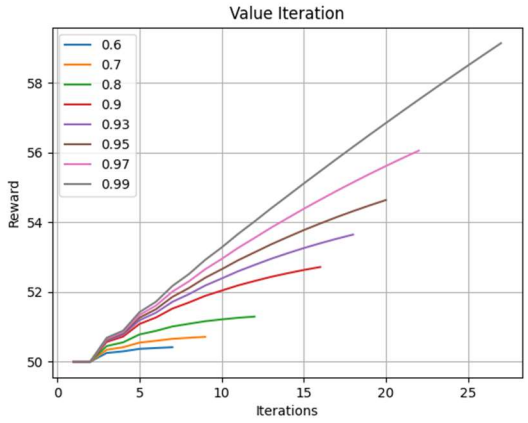
S = 6 Value Iteration Policy: (0, 0, 0, 0, 0, 1)
S = 6 Policy Iteration Policy: (0, 0, 0, 0, 0, 1)
S = 6 Q Learning Policy Gamma=0.95: (0, 0, 0, 0, 0, 1)
S = 6 Q Learning Policy Gamma=0.99: (0, 0, 0, 0, 0, 1)

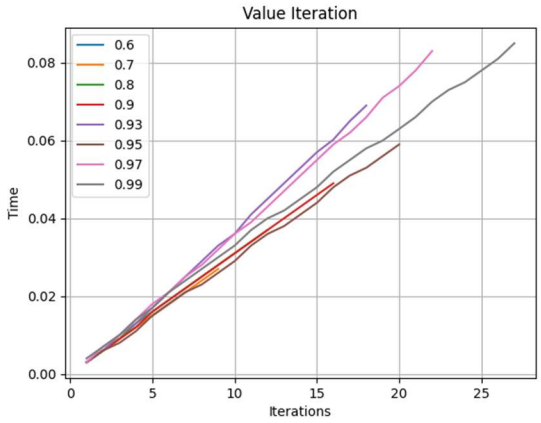
S = 2000, the policy values are too big to fit in the report.

Value Iteration (S = 2000):

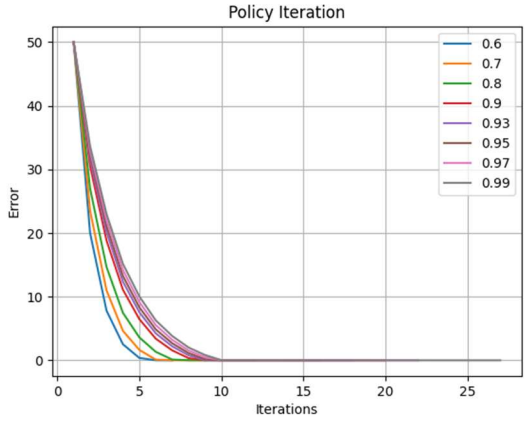


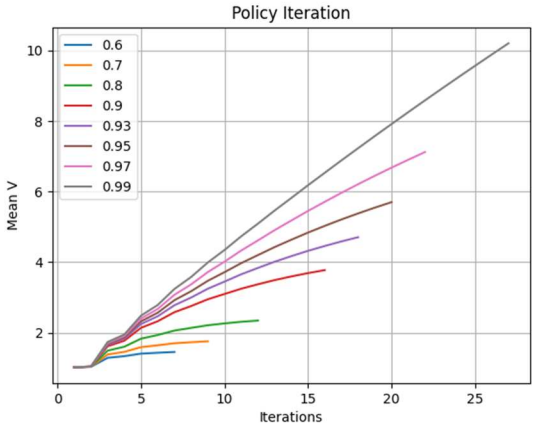


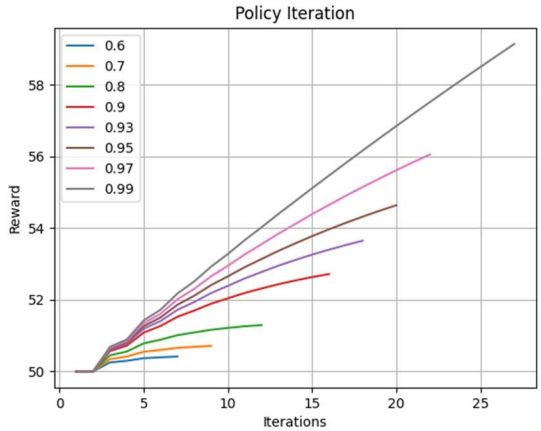


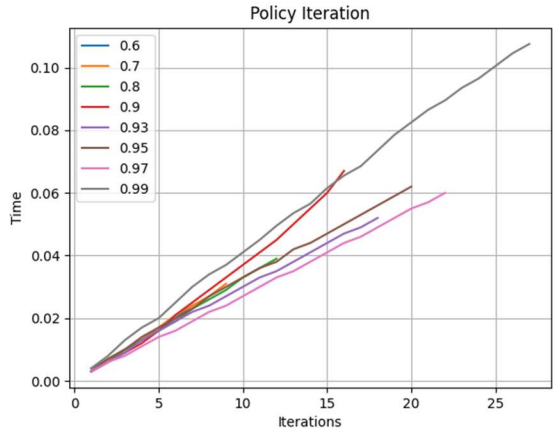


Policy Iteration (S = 2000):



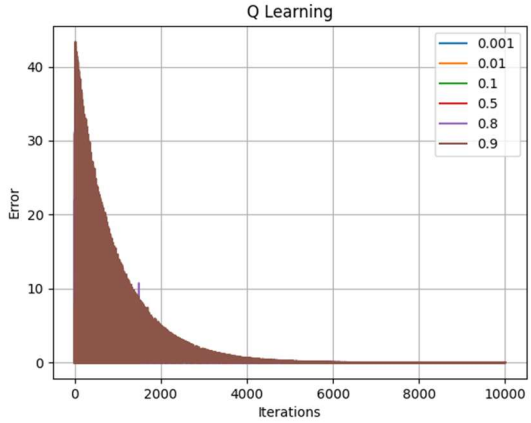


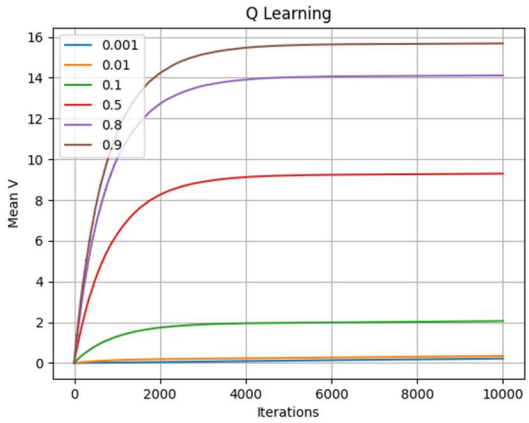




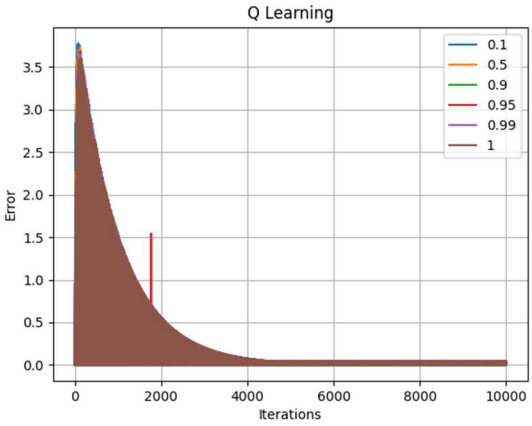
Q Learning (S = 2000):

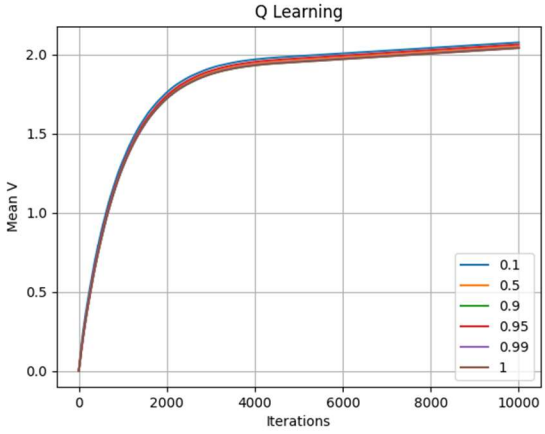
The following plots are based on varying values of alpha. Alpha = 0.1 is chosen for the experiments.





The following plots are based on varying values of epsilon. Epsilon = 0.95 is chosen for the experiment.





The following plots are based on varying values of gamma. Gamma = 0.99 is chosen for the experiment.