**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 them applied everyday, 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 run. They are 4x4, 8x8 and 25x25. The results are plotted below. The number of episodes is kept constant at 1000 and max iterations are set 1000, 3000 and 10000 respectively.

**Value Iteration (4x4):**

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.

Chart

Description automatically generatedChart, line chart

Description automatically generated

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Chart

Description automatically generatedChart, line chart

Description automatically generated

Chart

Description automatically generatedChart

Description automatically generated

**Policy Iteration (4x4):**

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.

Chart

Description automatically generatedChart, line chart

Description automatically generated

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Chart

Description automatically generatedChart, line chart

Description automatically generated

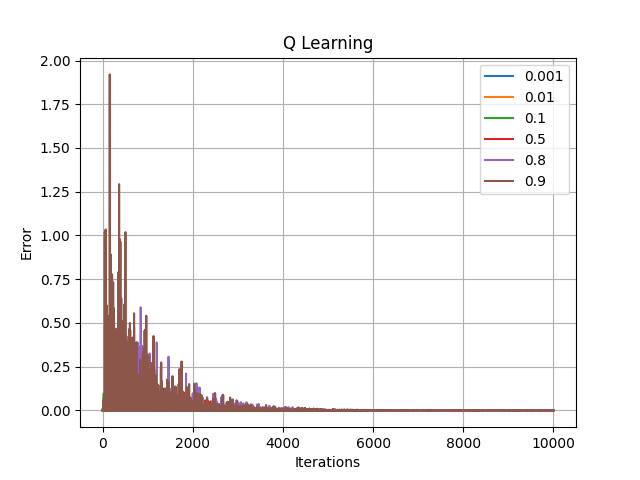
Chart

Description automatically generatedChart

Description automatically generated

**Q Learning (4x4):**

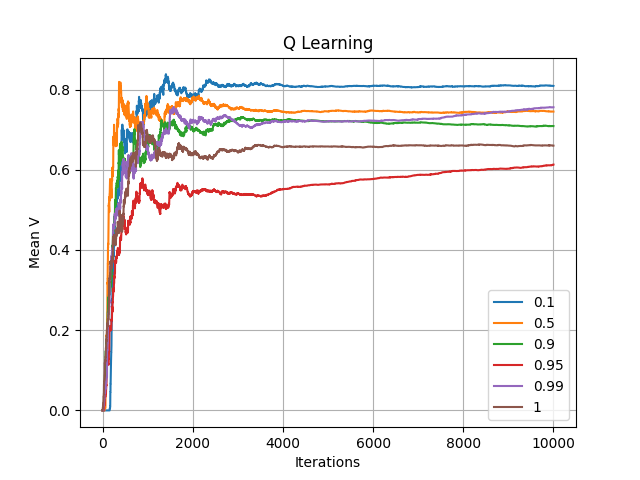
The following plots are based on varying values of alpha. Alpha = 0.1 is chosen for the experiments.

Chart

Description automatically generated

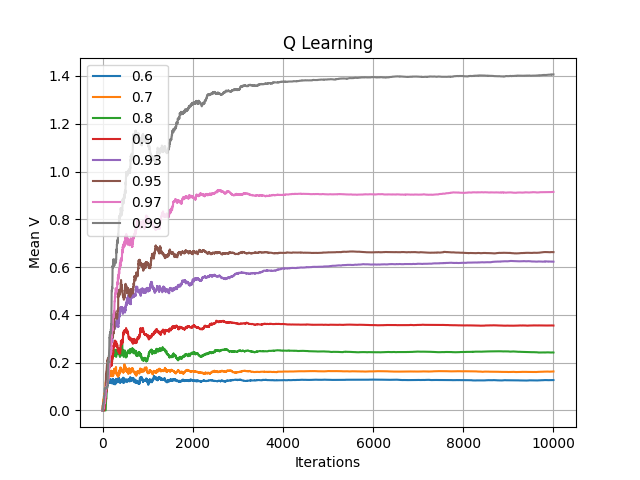
The following plots are based on varying values of epsilon. Epsilon = 0.95 is chosen for the experiment.

Chart, histogram

Description automatically generated

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

Chart, histogram

Description automatically generated

**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.

Chart

Description automatically generatedChart, line chart

Description automatically generatedChart, line chart

Description automatically generatedChart, line chart

Description automatically generatedChart

Description automatically generatedChart, line chart

Description automatically generated

Chart

Description automatically generatedChart, line chart

Description automatically generated

**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.

Chart

Description automatically generatedChart, line chart

Description automatically generatedChart, line chart

Description automatically generatedChart, line chart

Description automatically generatedChart, line chart

Description automatically generatedChart

Description automatically generatedChart

Description automatically generatedChart, line chart

Description automatically generated

**Q Learning (8x8):**

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

Chart, histogram

Description automatically generatedChart, line chart

Description automatically generated

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

Chart, histogram

Description automatically generatedChart

Description automatically generated

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

Chart, histogram

Description automatically generatedChart

Description automatically generated

**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)

**Forest**

Three states of sizes 3, 6 and 2000 are run. The results are plotted below.