Markov Decision Process and RL

Pburra6

# Introduction

The report explores the value and policy iteration algo for different MDP problems.

# MDPs

|  |  |  |
| --- | --- | --- |
| MDP Name | Type | State space size |
| Frozen Lake | Grid world - *Finite/discrete* | 20 \* 20 = 400 |
| Black Jack | *Non Grid World - Finite/ discrete* | 290 – Small |

Frozen lake is a simple problem to visualise and understand the dynamics of the decisions and shaping the reward structure. The transition prob can be adjusted to be stochastic or deterministic that should help us explore the change in policy as we adjust the map size, reward and gamma. Convergence for the frozen lake has bene measure as the smallest change observed in consecutive iterations.

Black Jack was considered as an option for a non-Grid MDP

# Frozen Lake – VI/PI

There are two ways to build the env, one with no slipping - deterministic and one with slipping - stochastic. Once the deterministic policy is set then the future paths reap the same reward. Stochastic will show the distribution.

### Compare slip and no slip with no holes:

The policy behaviour for no slip is to find the shortest way to goal following the Manhattan path along the edges. With the slip, the policy focus on moving to the goal through the centre of the grid to increase prob for reaching goal.

|  |  |  |
| --- | --- | --- |
|  | VI | PI |
| No Slip | 0.03s, 31 iterations | 0.02s, 27 iterations |
| Slip | 0.26s, 152 iterations | 0.23s, 153 iterations |

A graph of blue and white arrows

Description automatically generatedA graph of a graph

Description automatically generated

### Reward shaping:

Default reward structure was simple where reaching the goal had a reward of 1. We compare the difference in adding the reward shaping to include the penalty (-0.01) of step should see the changes in the path.

**Iterations required for convergence on slippery surface, (p\_frozen=0.8):**

1. Policy iterations are always faster in reaching convergence in the number of iterations.
2. Adding cost reduced the number of iterations required to reach convergence

A graph of different colored lines

Description automatically generated with medium confidence

**Runtimes for convergence on slippery surface (p\_frozen=0.8)**

A graph of different colored lines

Description automatically generated

1. Runtimes get exponentially higher with map size.
2. Runtimes are also made worse as gamma tends to 1 as remembering all past steps in a stochastic env leads to including the errors as well in the value matrix update.
3. More unpredictable the env the lower the gamma should be .

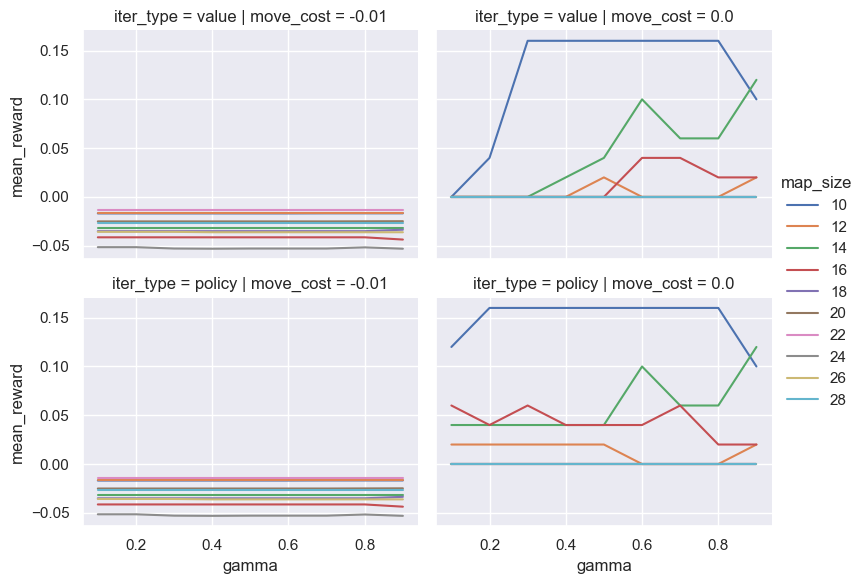
**Runtimes for convergence on slippery surface (p\_frozen=0.8)**

A graph of a number of numbers

Description automatically generated with medium confidence

1. Runtimes improve with adding move cost under policy iterations.

**Iterations required for convergence on slippery surface (p\_frozen=0.8)**



# References

<https://avandekleut.github.io/gym/>

<https://gymnasium.farama.org/tutorials/training_agents/FrozenLake_tuto/>

<https://gymnasium.farama.org/tutorials/training_agents/blackjack_tutorial/>