Reinforcement learning - Frozen lake

Introduction

The frozen lake is an environment that is a standard in reinforcement learning. The objective is for an agent to grid from a start point to the goal point while trying to avoid holes. The environment can either be deterministic or stochastic, which can affect the performance

I implemented Q learning where the agent learns q values to achieve the most reward. For exploration I used:

- ε-greedy: This chooses a random action with a certain probability of the best action, respectively explore or exploit.
- Boltzmann: This chooses actions based on q values, in which temperature, where a high temperature means more exploration and low temperature is more exploitation.

I also experimented with the effect of step and hole penalty on the deterministic environment on learning efficiency and safety.

Methods

Softmax Function

The softmax function converts the q value of a state into a probability of the actions. The higher q values show a higher probability. The temperature controls the randomness of choices, high temperatures promote more exploration while lower temperature favors choosing the action with high probability.

Action Selection

The two chosen policies include:

- ε-greedy: This chooses a random action with a certain probability of the best action, respectively explore or exploit.
- Boltzmann: This chooses actions based on q values, in which temperature, where a high temperature means more exploration and low temperature is more exploitation.

Q-Learning Training

The training function initializes the environment and the q table. For each episode the agent will interact with the environment, selecting actions according to the chosen policy. Then it receives rewards and updates the q value. At the end of each episode, the total reward and success are recorded. Epsilon decays gradually to reduce exploration over time, allowing the agent to exploit the learned policy as training progresses.

Visualization

I used different visualization techniques:

- Learning curves: Plots of average reward and success rate per episode to observe the progress and convergence.
- Q-table heatmaps: Show which actions are favored in each state.
- Policy arrows: Overlay arrows on the grid to indicate the learned optimal action for each state.
- Comparison plots: Compare different strategies, temperature settings, and penalty values.

Experiments & Results

Deterministic Environment (slippery=False)

Using ε-greedy exploration with ε=0.2, the agent quickly learned almost perfectly: By episode 500, the average reward was 0.737 and by episode 2000, it reached 0.916.

In deterministic conditions and ε-greedy exploration the agent was able to navigate well to determine the goal.

While exploring with the boltzmann policy with different temperatures, it showed that a higher temperature slowed convergence due to excessive exploration and had a lower reward while lower temperatures was better but was not as good at the ε-greedy.

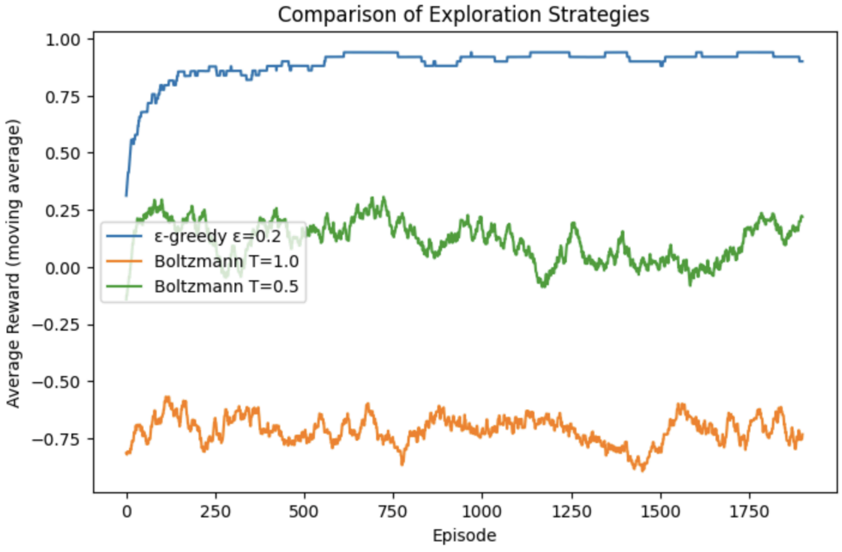

Fig 3

Deterministic vs Slippery with Boltzman

Next I explored the slippery = true vs false. Slippery = true means that even if the agent chooses the best action, it might not move as intended. Because of this the average reward was overall lower than teh deterministic conditions which showed higher rewards and more stable success rates.

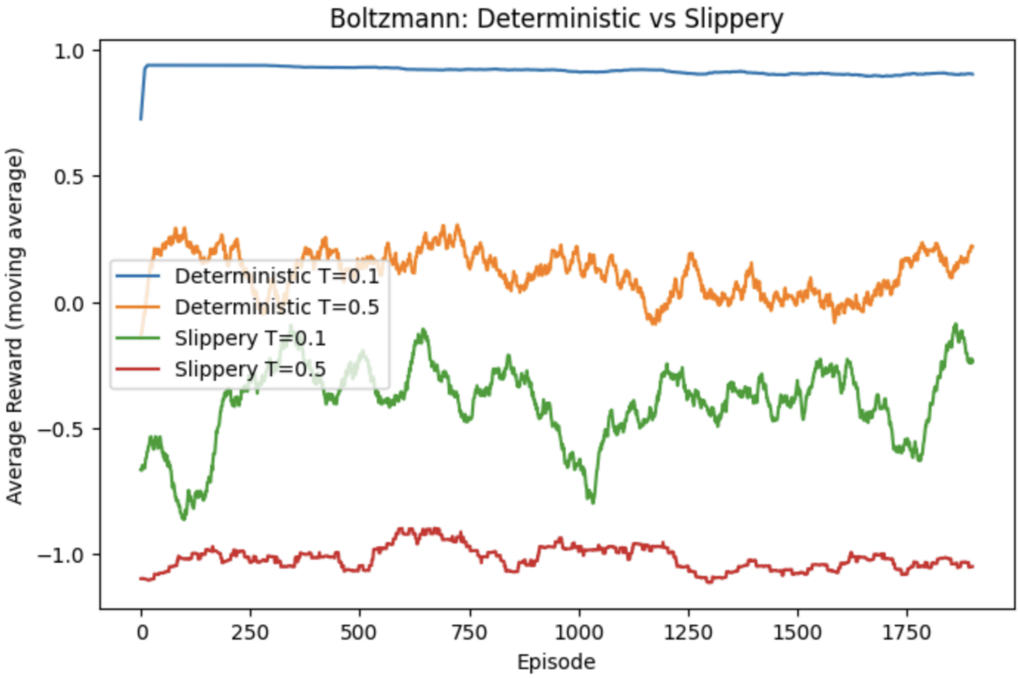

Fig 4

Step Penalty Experiments on

Next I explored step penalties of 0.0, -0.01, and -0.05 on ε-greedy. It was seen that the -0.01 penalty allowed the agent to be the best and get to the goal while the higher penalty slowed learning.

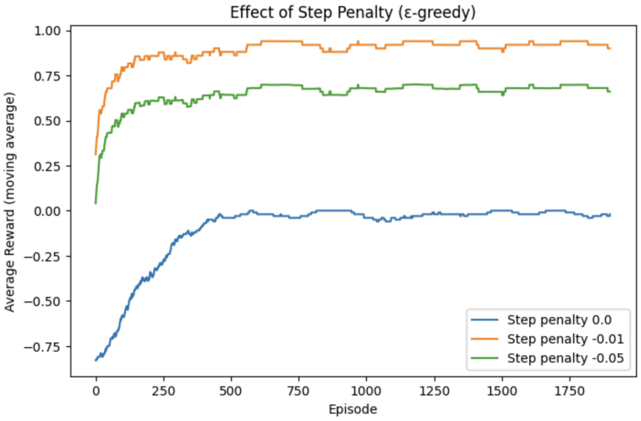

Fig 7

Hole Penalty Experiments with ε-greedy

The hole penalties of -1.0 and -5.0 influenced the agent's to be more cautious and achieved higher results but sometimes at the cost that the agent learns slower.

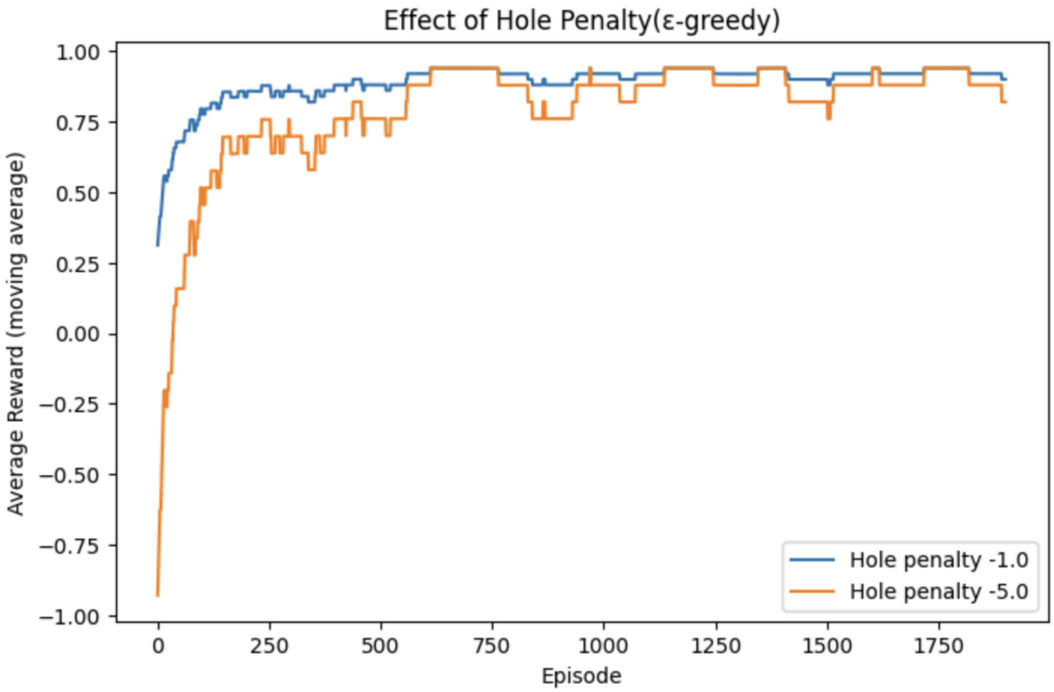

Fig 8

3.5 Learned Policy

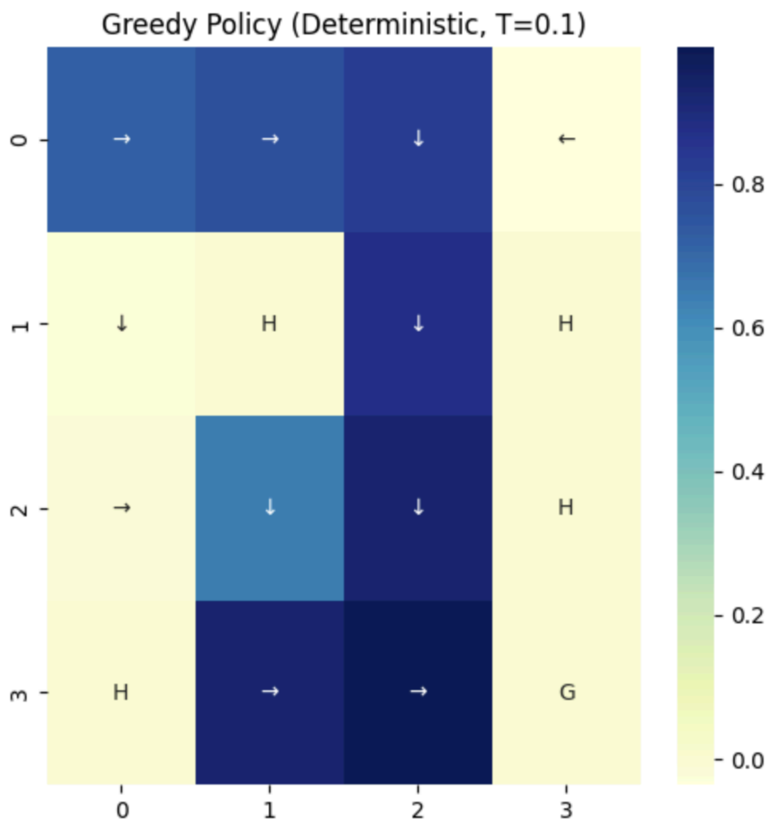

Fig 6

The arrows show the best for each state in the deterministic policy with temperature at 1. This shows that the agent learned a safe and efficient path to the goal.

Discussion

Key insights

Exploration vs Exploitation:

- ε-greedy has a simple way that manages exploration and exploitation. Too much exploration slows convergence; too little risks the best actions.
- Boltzmann exploration offers a probabilistic alternative, where temperature controls the balance continuously.

Effect of Environment:

- Deterministic environments allow consistent learning of the optimal policy.
- Slippery environments introduce uncertainty, reducing average rewards

Impact of Penalties:

- Step penalties make the agent take shorter paths.
- Hole penalties makes the agent avoids high-risk states.

Conclusion

Through this exploration I learned with the hyperparameters I tested:

- ε-greedy is highly effective in deterministic settings.
- Boltzmann exploration is more flexible for stochastic environments.
- Step and hole penalties significantly affect learning behavior, efficiency, and safety.

Finding the best hyperparameters such as ε, temperature, step penalties, and hole penalties is essential to determine the performance in a learning agent.