Reinforcement Learning Assignment Report

Student Name: Suyash Kulkarni **Student Id:** 24239960

Topic: Q-Learning on Deterministic FrozenLake Environment

Introduction

This paper explores the application of the Q-learning algorithm over a deterministic 5x5 FrozenLake grid. The agent begins at top-left and must find its way to the bottom-right goal while avoiding holes that are placed immovably. Rewards are in place to encourage optimal path-following: +10 for target achievement, -5 for falling in the hole, and -1 per step.

Multiple Q-learning configurations were attempted through varying the learning rate (alpha), discount factor (gamma), and exploration rate (epsilon). All experiments were executed for 10,000 episodes, and performance was evaluated in terms of end Q-value estimates and learning curves. The report contains results from three example strategies and contains a comparative overview of all experiments.

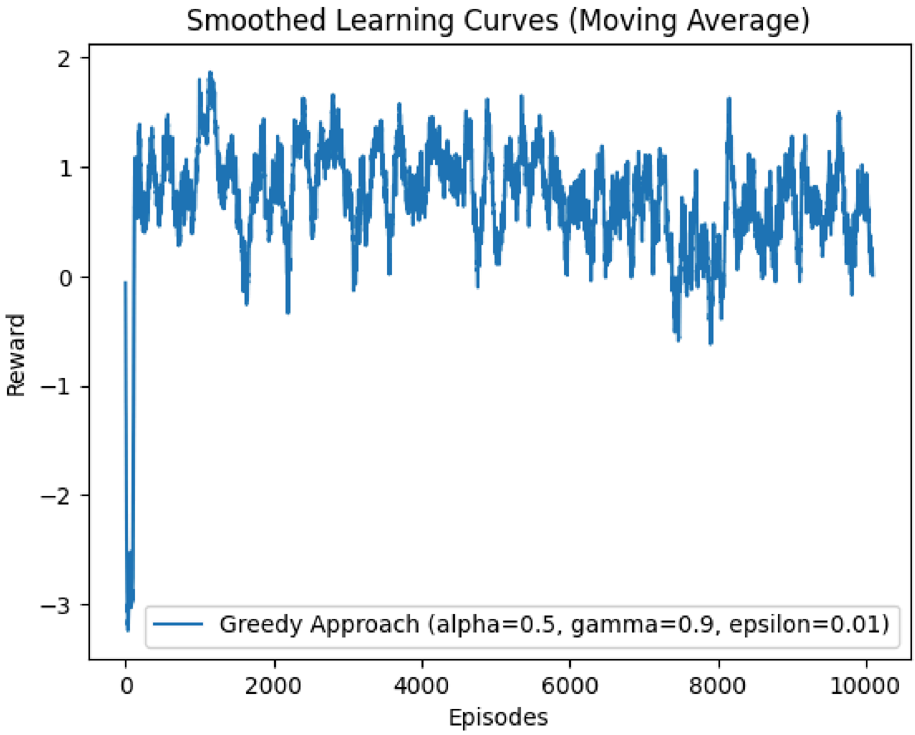
Q-Value Summary Table

Strategy	Max Q-value	Avg Q-value
Baseline (α=0.5, γ=0.9, ε=0.1)	10.00	2.83
Epsilon Decay (α=0.5, γ=0.9, ε=0.1 to 0)	10.00	2.89
Fast Learning (α=0.9, γ=0.9, ε=0.1)	10.00	2.26

1. Baseline (alpha = 0.5, gamma = 0.9, epsilon = 0.1)

The baseline strategy offers a well-balanced setting with moderate learning and discounting of delayed rewards. It has a consistent performance with a well-balanced average Q-value across the grid. The learning curve shows steady growth with oscillations, reflecting an optimal balance between exploration and exploitation.

This setup effectively solves the FrozenLake without overcommitting greedy or risky actions. The terminal Q-values reflect an obvious and best policy path towards the target, evading holes and optimizing strides. It is usable as a solid baseline of comparison for alternative strategies.

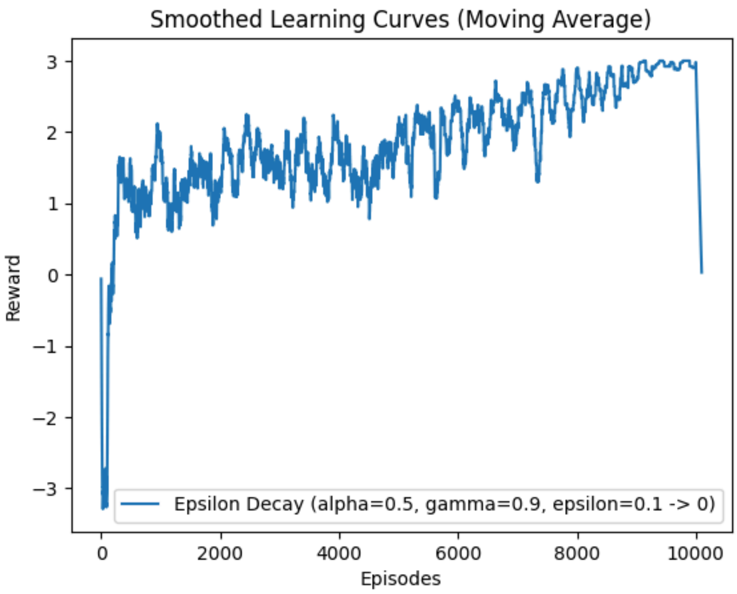


.

2. Epsilon Decay (alpha = 0.5, gamma = 0.9, epsilon = 0.1 to 0)

This approach begins from an exploration strategy and subsequently evolves step-by-step to be greedy when epsilon tends to zero. Though here the idea is to explore more first and then exploit, experiment proves that too early decay harms complete perception of the environment. The agent tends to remain in poor-quality trajectories if exploration gets shut too prematurely.

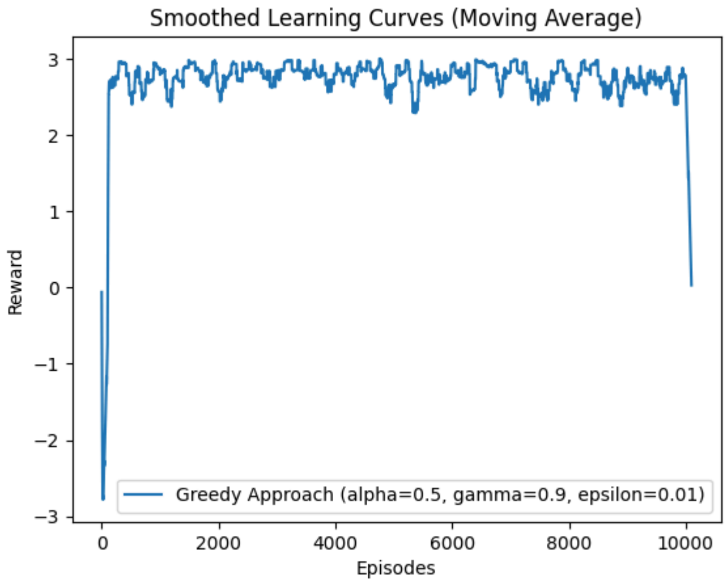
Average Q-values are below the baseline, and the learning curve is not as smooth with slow progress. This shows that while epsilon decay is intuitively useful, its use needs to be carefully planned in such a way so as to avoid slowing down the learning before max behavior is maximally revealed.



3. Greedy Approach (alpha = 0.5, gamma= 0.9, epsilon = 0.01)

The Greedy Approach, with a low exploration rate (ε = 0.01), showed the best overall performance among all tested configurations. Its learning curve showed consistent improvement with minimal noise, and the final Q-values reflected a strong, stable policy.

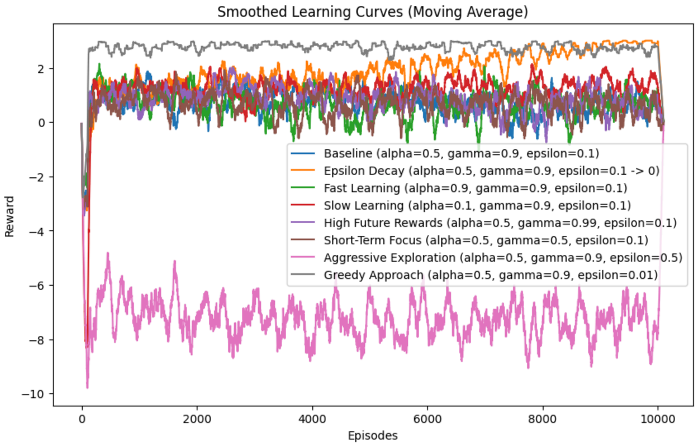
The agent quickly identified optimal paths in the deterministic FrozenLake environment, reinforcing efficient behavior early on, leading to faster convergence and higher average reward across episodes. Despite concerns about low exploration trapping the agent in suboptimal loops, the simplicity of the environment allowed the agent to quickly discover optimal actions and stick with them, highlighting the effectiveness of a near-greedy strategy in low-risk, fully observable environments.

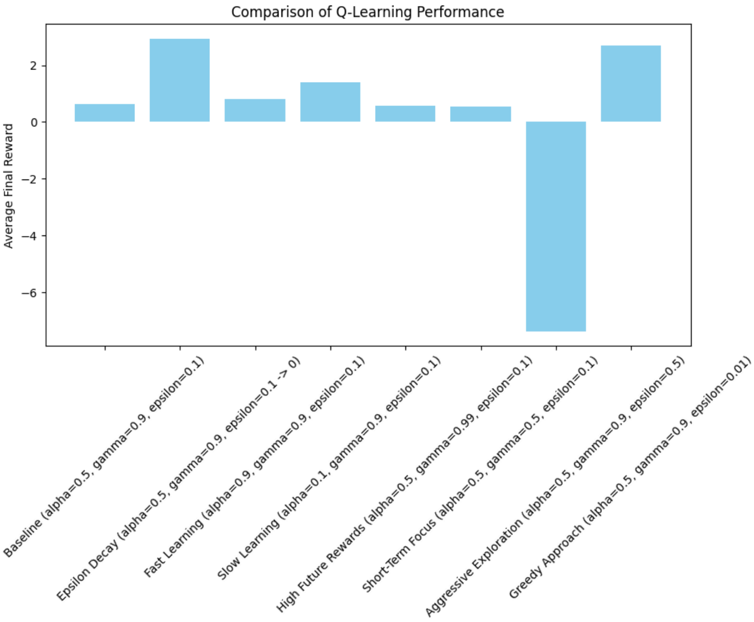


4. All other custom strategies

Fast Learning, Slow Learning, High Future Rewards, Short-Term Focus, and Aggressive Exploration—each exhibited different levels of performance depending on their learning rate, discount factor, and exploration setting. Fast Learning (α = 0.9) enabled the agent to quickly update its Q-values, with good early convergence, but this aggressive adaptation sometimes results in instability or overfitting in more complex environments. High Future Rewards (γ = 0.99) was successful by placing emphasis on delayed rewards so that the agent could appropriately assign value to the distant goal, even at the risk of diluting the impact of close future penalties like holes.

Slow Learning (α = 0.1), on the other hand, caused progress to be very slow due to its conservative updates, requiring more episodes to learn a good policy. Short-Term Focus (γ = 0.5) always performed worst, as the agent did not value much reaching the distant goal and hence had low mean Q-values and bad paths. Aggressive Exploration (ε = 0.5) ensured perfect state coverage, but the high randomness led to slow stable policy construction. In most cases, while some of these setups were worth exploring in principle, they were either too unstable or too conservative to beat out the simpler, more specialized strategies in a deterministic setting like FrozenLake.


Smoothed Learning Curves (Moving Average)
Reward
Episodes
Baseline (alpha=0.5, gamma=0.9, epsilon=0.1)
Epsilon Decay (alpha=0.5, gamma=0.9, epsilon=0.1 -> 0)
Fast Learning (alpha=0.9, gamma=0.9, epsilon=0.1)
Slow Learning (alpha=0.1, gamma=0.9, epsilon=0.1)
High Future Rewards (alpha=0.5, gamma=0.99, epsilon=0.1)
Short-Term Focus (alpha=0.5, gamma=0.5, epsilon=0.1)
Aggressive Exploration (alpha=0.5, gamma=0.9, epsilon=0.5)
Greedy Approach (alpha=0.5, gamma=0.9, epsilon=0.01)


Comparison of Q-Learning Performance
Average Final Reward
Baseline (alpha=0.5, gamma=0.9, epsilon=0.1)
Epsilon Decay (alpha=0.5, gamma=0.9, epsilon=0.1 -> 0)
Fast Learning (alpha=0.9, gamma=0.9, epsilon=0.1)
Slow Learning (alpha=0.1, gamma=0.9, epsilon=0.1)
High Future Rewards (alpha=0.5, gamma=0.99, epsilon=0.1)
Short-Term Focus (alpha=0.5, gamma=0.5, epsilon=0.1)
Aggressive Exploration (alpha=0.5, gamma=0.9, epsilon=0.5)
Greedy Approach (alpha=0.5, gamma=0.9, epsilon=0.01)