Exploring Q-Learning in Taxi-v3 Environment: Hyperparameters and Exploration Strategies

Karol Korszun

June 2024

1 Introduction

Q-learning is a model-free reinforcement learning algorithm that aims to find the optimal action-selection policy for a given finite Markov decision process (MDP). This report explores the application of Q-learning to the Taxi-v3 problem in the Gym environment, focusing on the effects of various hyperparameters and exploration strategies.

2 Q-Learning Algorithm

Q-learning is an off-policy reinforcement learning algorithm that seeks to learn the value of the optimal policy, denoted as Q^*. The Q-value $Q(s, a)$ represents the expected future rewards for an agent in state s taking action a and following the optimal policy thereafter. The Q-learning update rule is given by:

$$Q(s_t, a_t) \leftarrow (1 - \alpha)Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) \right]$$

where: - α is the learning rate, - γ is the discount factor, - r_{t+1} is the reward received after taking action a_t from state s_t, - s_{t+1} is the new state after taking action a_t.

3 Effect of hyperparameters

In this section, all of the simulations were conducted using the Epsilon-Greedy strategy with epsilon decay. To minimize the effect of randomness, each simulation was performed multiple times. For better visibility, the plots show the moving averages of the performance metrics.

3.1 Learning Rate and Epsilon Decay

The effect of different learning rates (α) and epsilon decays on the performance of the Q-learning algorithm were analyzed. The learning rates tested were 0.1, 0.5 and 0.9 and the epsilon decays were 0.9, 0.95 and 0.99.

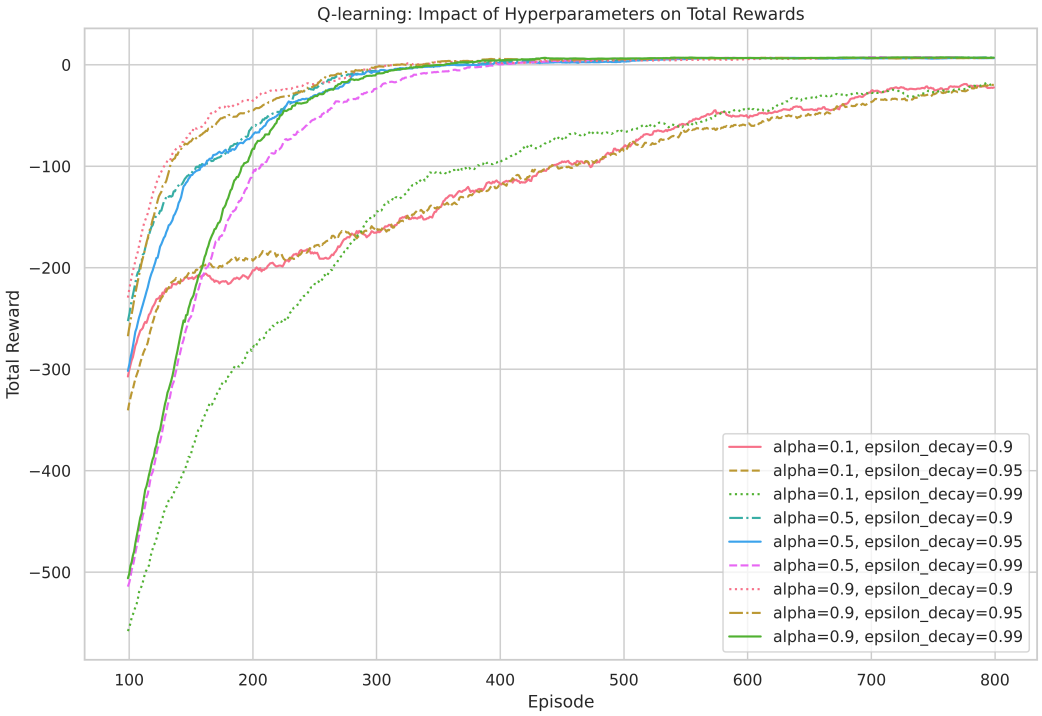

Figure 1: Performance of Q-learning with varying learning rates (α) using Epsilon-Greedy strategy with decay.

As shown in Figure 1, the Q-learning algorithm with a learning rate of 0.1 were significant outliers, failing to reach optimal solution after 800 episodes. For other learning rates, the models converged at roughly the same time; however, the initial phase of learning varied significantly. Higher learning rate $\alpha = 0.9$) showed generally quicker initial learning.

3.2 Epsilon parameter

Experiments were conducted to explore the impact of the epsilon parameter on the learning process. Different values of epsilon, including 0.1, 0.3, 0.5, 0.7, 0.9, and 0.99, were tested. However, as illustrated in Figure 2, the epsilon parameter did not exhibit a significant effect on the training process. All learning plots displayed similar patterns and converged at approximately the same time.

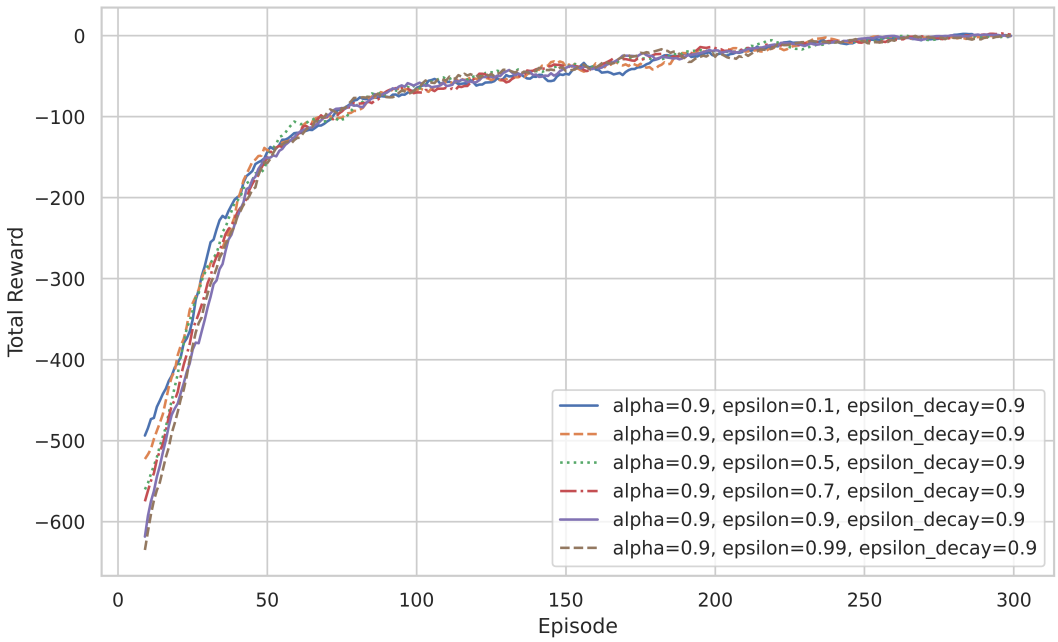

Figure 2: Effect of different values of epsilon on the learning process.

3.3 Gamma Parameter

In the context of Q-learning, the gamma parameter (γ) represents the discount factor, which determines the importance of future rewards in the agent's decision-making process. A higher gamma value indicates that the agent considers future rewards more heavily, while a lower gamma value prioritizes immediate rewards.

Experiments were conducted to evaluate the impact of different gamma parameters on the learning process. Generally speaking, higher gamma parameters tended to yield better results.

In the context of this task, it's logical for our agent to prioritize obtaining future rewards, as the ultimate goal is to complete the entire taxi sequence successfully.

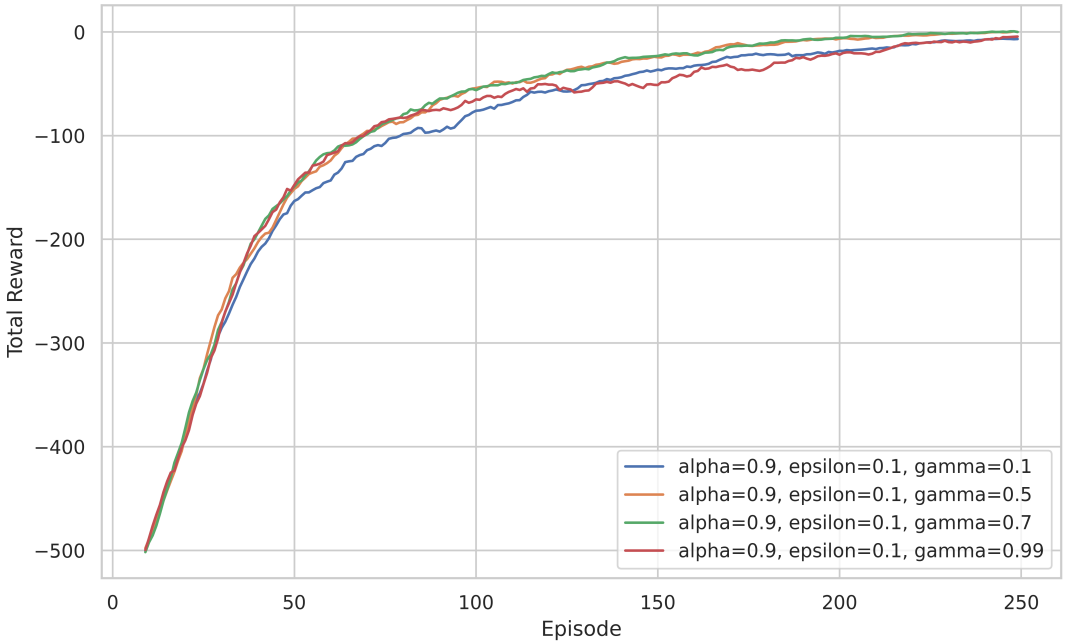

Figure 3: Effect of different values of gamma on the learning process.

4 Strategy Comparison

This section compares the performance of two strategies: Epsilon-Greedy and Epsilon First. In the Epsilon First strategy, exploration is emphasized for a specified number of initial episodes, such as 15 in this example, before transitioning to a more greedy approach. As observed in the results, the Epsilon-Greedy strategy demonstrates faster initial learning. However, over time, the Epsilon First strategy catches up, and both strategies eventually converge to similar outcomes.

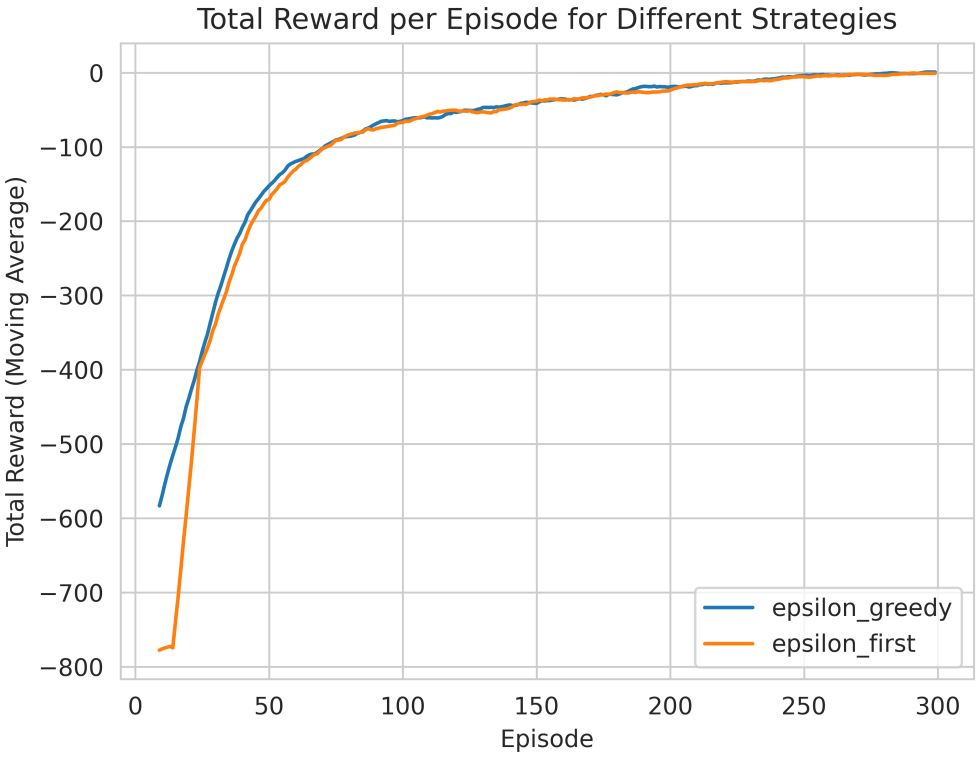

Figure 4: Strategy comparison

5 Conclusions

In conclusion, the performance of different strategies and hyperparameters in Q-learning applied to the Taxi-v3 problem was investigated. While variations in performance were observed, the simplicity of the problem may have limited the extent to which variaton in effectiveness between different hyperparameter combinations was visible.

Nevertheless, some trends were clearly visible, such as the influence of the gamma parameter on learning efficiency. Higher gamma values generally led to better performance, indicating the importance of considering future rewards in the agent's decision-making process.

Additionally, the comparison between different exploration strategies, such as Epsilon-Greedy and Epsilon First, revealed no significant differences in their effectiveness. Both strategies exhibited similar convergence behavior and achieved comparable outcomes over time.