Q-learning Hyperparameter Study in a 10×10 GridWorld

Md Raihan Subhan
SID: 20585071
Department: Computer Science with Interdisciplinary Applications

09/24/2025

Abstract

This report provides a detailed analysis of the effect of key Q-learning hyperparameters—learning rate (α), discount factor (γ), and exploration rate (ε)—on the performance of Q-learning in a deterministic 10×10 GridWorld. The environment, which includes a fixed start state $(0, 0)$, goal state $(9, 9)$, and exactly 20 obstacles, is used to assess how different configurations of these hyperparameters affect convergence speed, stability, and performance. The results show clear dependencies of performance on α, γ, and ε, demonstrating expected trade-offs between exploration and exploitation, planning horizons, and learning stability. Practical recommendations for optimal hyperparameter settings are provided based on the findings.

1 Environment and Methods

GridWorld Setup. The 10×10 grid environment consists of a start state at $(0, 0)$, a goal state at $(9, 9)$, and exactly 20 obstacles placed randomly. The agent has four possible actions: up, down, left, and right. If an invalid move (e.g., hitting a wall or obstacle) is attempted, the agent remains in the same position. The agent receives a reward of -1 for each step and a reward of 0 for reaching the goal. Thus, the optimal return for any successful episode is the negative of the shortest path length from the start to the goal.

Q-learning Algorithm. The study utilizes tabular Q-learning with an ε-greedy policy, where the Q-values are updated using the following rule:

$$Q(s, a) \leftarrow Q(s, a) + \alpha \Big[r + \gamma \max_{a'} Q(s', a') - Q(s, a) \Big]. \tag{1}$$

Where: - s and a represent the current state and action taken, - α is the learning rate, - γ is the discount factor, - r is the reward received, - s' is the next state, and - $\max_{a'} Q(s', a')$ is the maximum future reward.

The agent explores the environment based on an ε-greedy policy, with ε controlling the exploration-exploitation trade-off. The algorithm is run for 500 episodes with a fixed random seed for reproducibility.

Obstacle Layout. The obstacle configuration remains fixed across all experiments, ensuring that observed changes in learning curves are due to hyperparameter variations and not changes in the environment layout. This fixed layout is visualized in Figure 1.

2 Experimental Design

Three experimental conditions are tested, each varying a single hyperparameter while keeping the others fixed at baseline values:

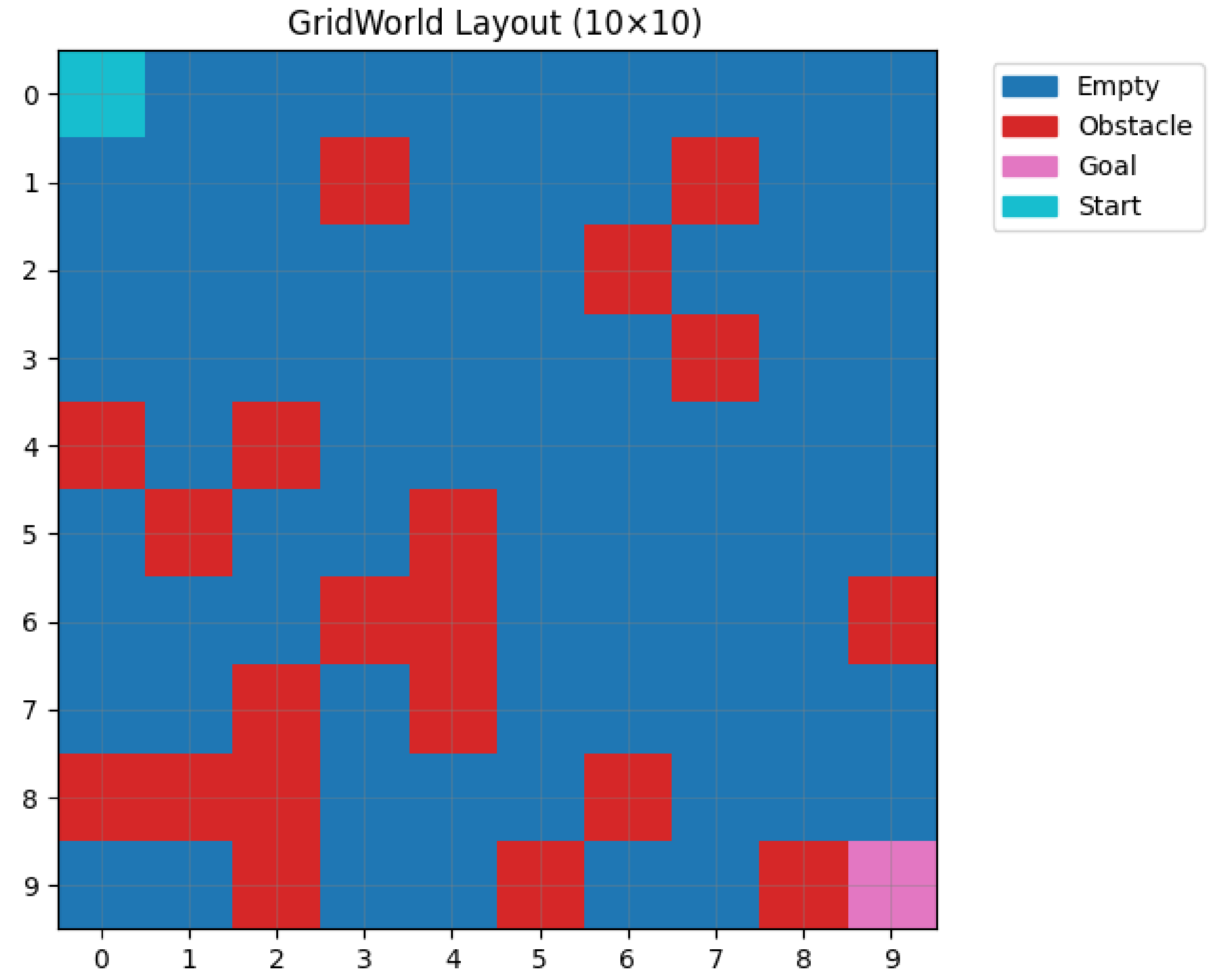

Figure 1: GridWorld obstacle layout used in all experiments. Start $(0, 0)$ and goal $(9, 9)$ are fixed; exactly 20 obstacles are shown.

- **Learning rate** (α): Varying α values from $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ with $\gamma = 0.9$, $\varepsilon = 0.1$ fixed.
- **Discount factor** (γ): Varying γ values from $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ with $\alpha = 0.1$, $\varepsilon = 0.1$ fixed.
- **Exploration rate** (ε): Varying ε values from $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ with $\alpha = 0.1$, $\gamma = 0.9$ fixed.

For each experiment, the returns (cumulative rewards) over episodes are plotted for all five variations of the hyperparameter on a single graph for comparison, ensuring consistency in layout and random seed across all runs.

3 Results and Discussion

3.1 Learning Rate (α): Speed vs. Stability

Smaller α values, such as $\alpha = 0.1$, yield smoother learning curves with slower convergence. In contrast, larger values, like $\alpha = 0.7$ or $\alpha = 0.9$, speed up initial learning but can introduce oscillations

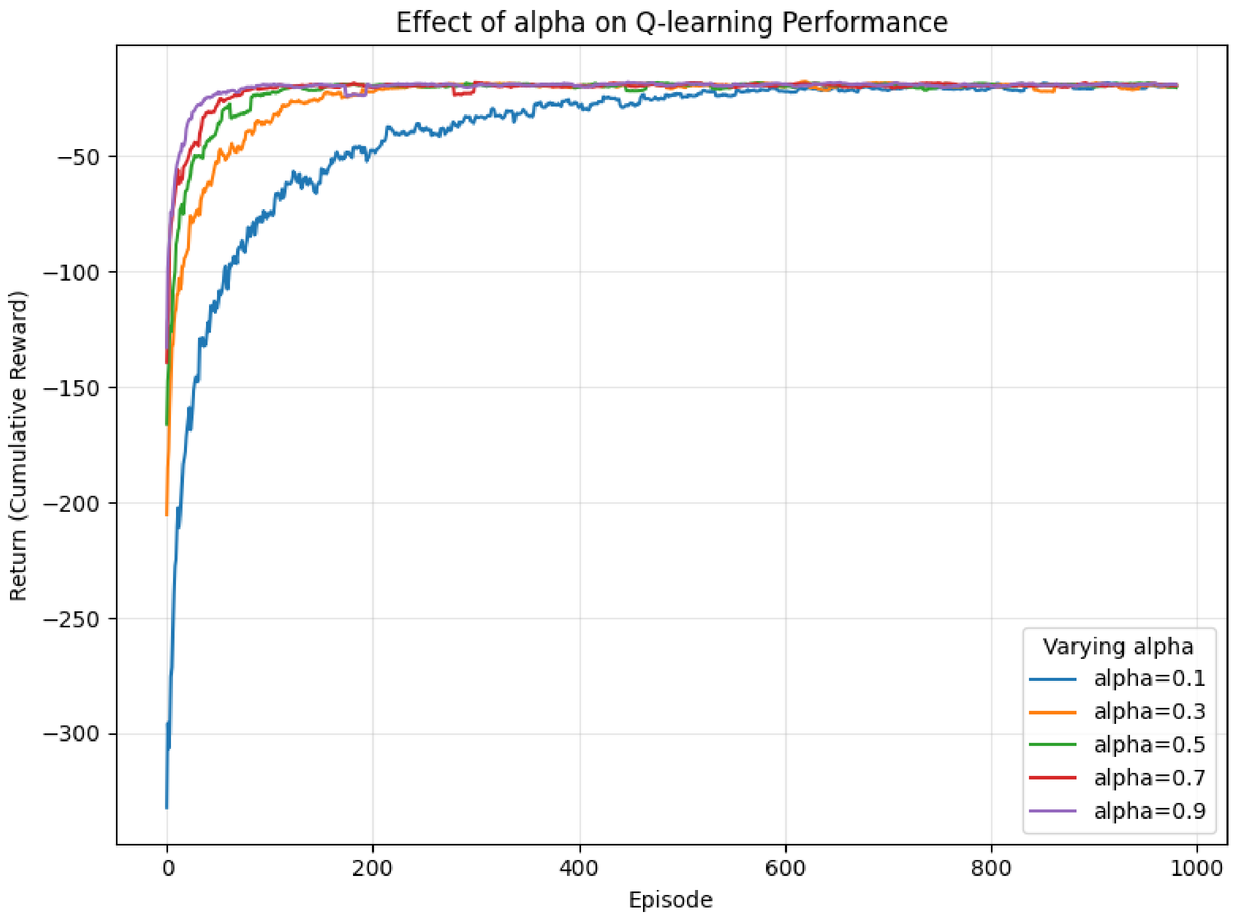

Figure 2: Effect of learning rate α on Q-learning performance (higher is better; returns are negative due to step penalties). All five α values are plotted together for comparison; moving-average smoothing applied.

and instability as they attempt to "chase" recent rewards. The ideal α strikes a balance between fast learning and stable convergence, with $\alpha \in [0.3, 0.5]$ producing the most consistent performance (Figure 2).

3.2 Discount Factor (γ): Planning Horizon

Low values of γ (e.g., 0.1) lead to short-term focus, where immediate rewards dominate over distant outcomes. This results in inefficient paths and longer learning times. As γ increases, the agent increasingly values long-term rewards, which allows it to identify and follow shorter paths. The results indicate that $\gamma = 0.9$ is optimal for this setup, enabling the agent to plan effectively over multiple steps (Figure 3).

3.3 Exploration Rate (ε): Exploration–Exploitation Trade-off

A very low ε (e.g., 0.1) restricts exploration, causing slower discovery of optimal paths but stronger exploitation of known paths. A moderate ε (e.g., 0.3–0.5) enhances early learning by encouraging exploration, though it may lead to noisier learning curves. Very high values of ε (e.g., 0.7 or

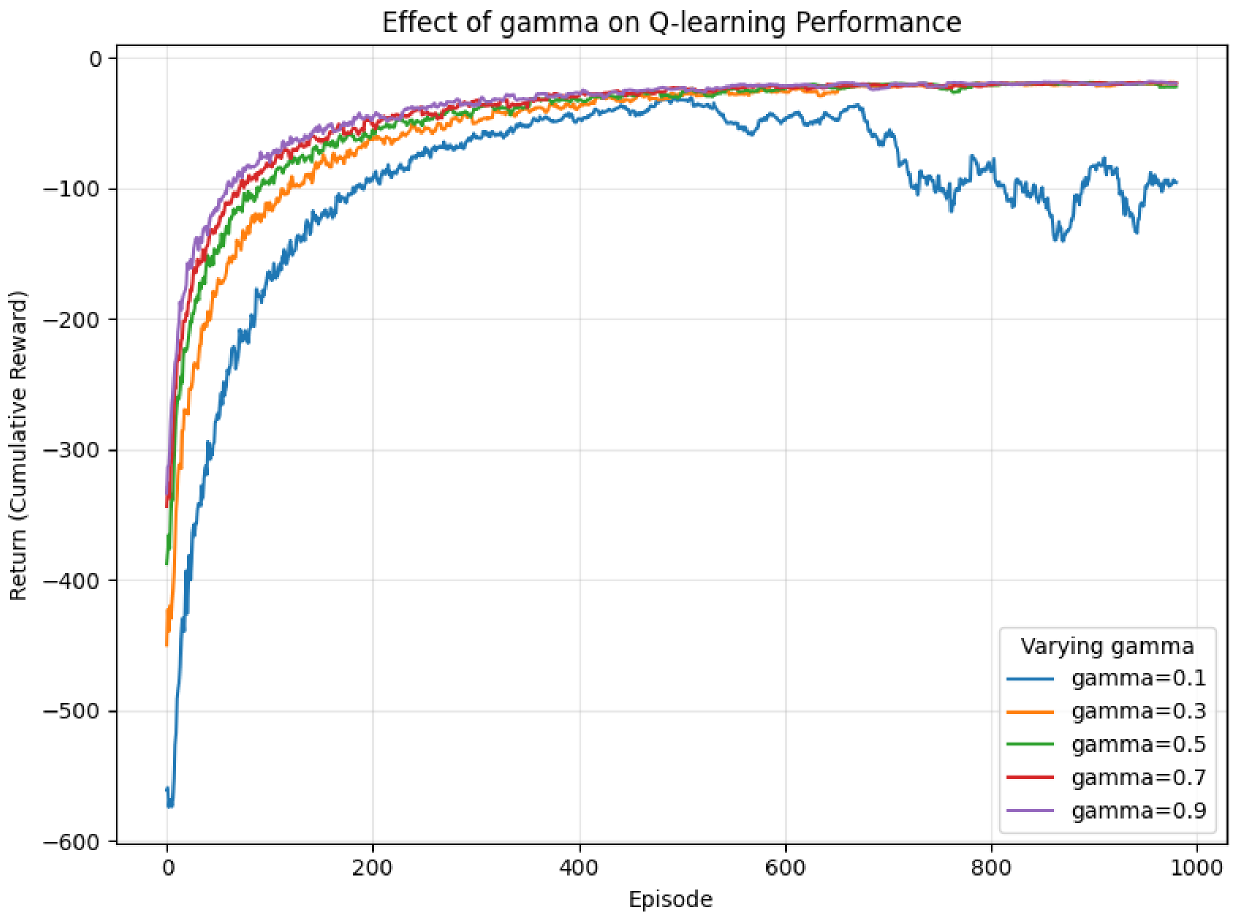

Figure 3: Effect of discount factor γ on Q-learning performance. Curves with higher γ typically approach the best asymptotic return more closely by valuing long-term outcomes.

higher) result in excessive randomness, delaying convergence. The ideal ε balances exploration and exploitation, with $\varepsilon \in [0.1, 0.3]$ typically producing the best performance (Figure 4).

3.4 Influence of the Obstacle Layout

Using a fixed obstacle layout ensures that differences in the learning curves across experiments are due to hyperparameter variations rather than changes in the environment. The narrow corridors and dead-ends present in the layout increase the need for exploration, particularly in early episodes, which amplifies the effect of ε on performance. If the obstacle layout were randomized, results might vary based on the layout's complexity, introducing additional variance in the learning process.

4 Practical Recommendations

Based on the experiments, the following hyperparameter settings are recommended for optimal Q-learning performance in deterministic grid environments:

- **Learning rate (α):** Use $\alpha \in [0.3, 0.5]$ for fast yet stable convergence.
- **Discount factor (γ):** Set γ to 0.9 to value long-term outcomes without slowing convergence.

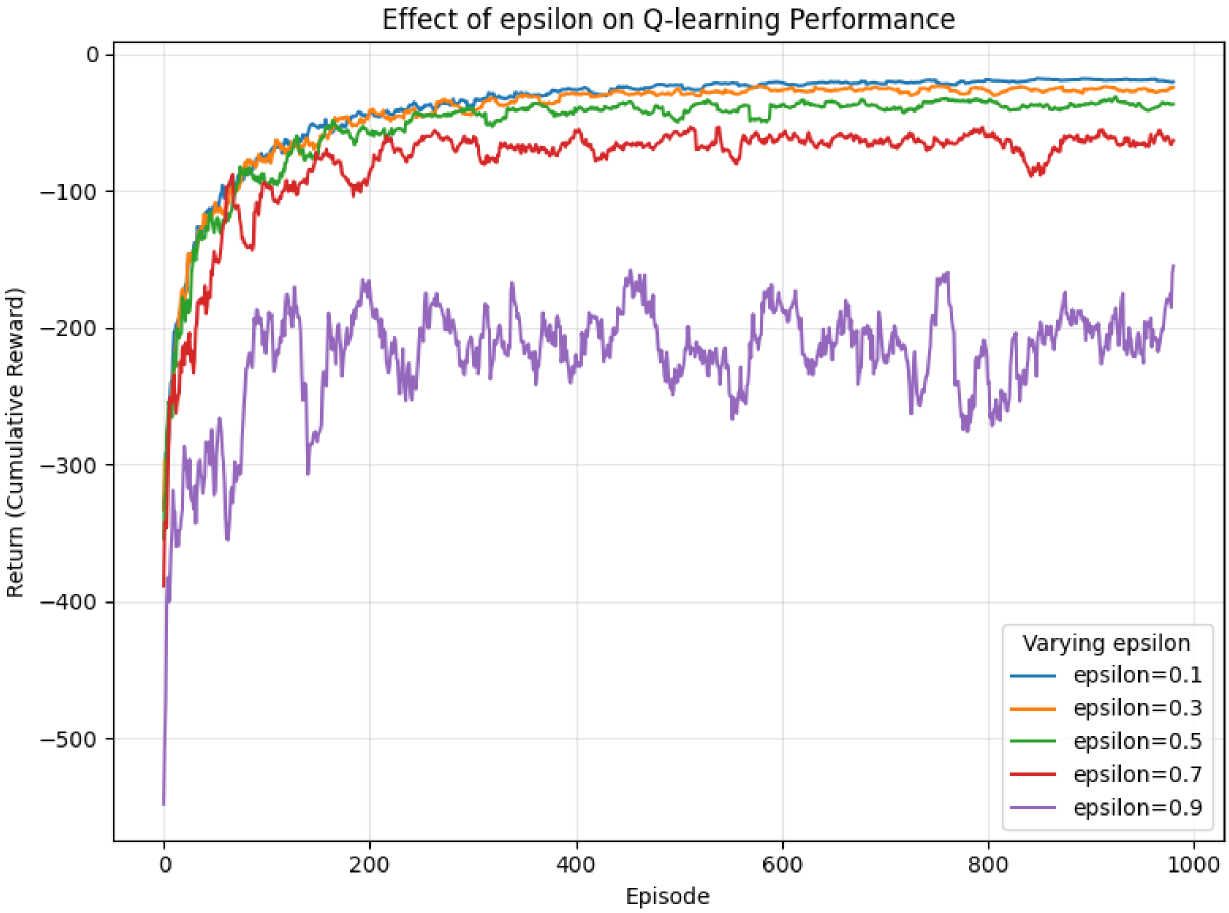

Figure 4: Effect of exploration rate ε on Q-learning performance. Moderate exploration accelerates early discovery; excessive exploration slows late-stage exploitation.

- **Exploration rate (ε):** Set $\varepsilon \in [0.1, 0.3]$ to balance exploration and exploitation.

It is important to document the smoothing window and ensure that axes and legends are consistent across figures. Plots should be saved as image files for clarity, and the obstacle map should be included in the report.

5 Conclusion

The experiments confirm the expected behavior of Q-learning with respect to the chosen hyperparameters. A moderate learning rate provides the best balance between speed and stability, the discount factor significantly influences planning behavior with $\gamma = 0.9$ yielding optimal performance, and the exploration rate plays a critical role in balancing exploration and exploitation. These findings offer important guidelines for tuning Q-learning in similar environments. The use of fixed random seeds and a consistent layout ensures reproducibility and fairness in the comparisons across hyperparameter settings.

Reproducibility. Random seeds were fixed for map generation and training, ensuring that differences across the curves are attributable to hyperparameter settings rather than layout variations. Code, plots, and the layout image are provided in the accompanying notebook/script.