CSCN8020 – Assignment 2: Q-Learning Report

Student: Manu Mathew

Course: CSCN8020 Reinforcement Learning

Environment: Taxi-v3 (500 discrete states, 6 actions)

1. Introduction

This report presents the implementation and evaluation of a Q-Learning agent in the Taxi-v3 environment. The agent learns optimal pick-up and drop-off behavior by exploring different hyperparameter settings. We evaluate how the learning rate (α) and discount factor (γ) affect learning performance while keeping the exploration factor (ε) fixed at 0.1.

2. Experimental Setup

• Environment: Taxi-v3 (500 discrete states, 6 actions)
• Reward structure: +20 for successful drop-off, -10 for illegal pick/drop, -1 per step
• Fixed parameters: ε = 0.1, episodes = 3000, max_steps = 200
• Varied parameters:
  - Learning rate $\alpha \in \{0.01, 0.001, 0.2\}$
  - Discount factor $\gamma \in \{0.2, 0.3\}$
  - Baseline: $\alpha = 0.1$, $\gamma = 0.9$

3. Results and Metrics

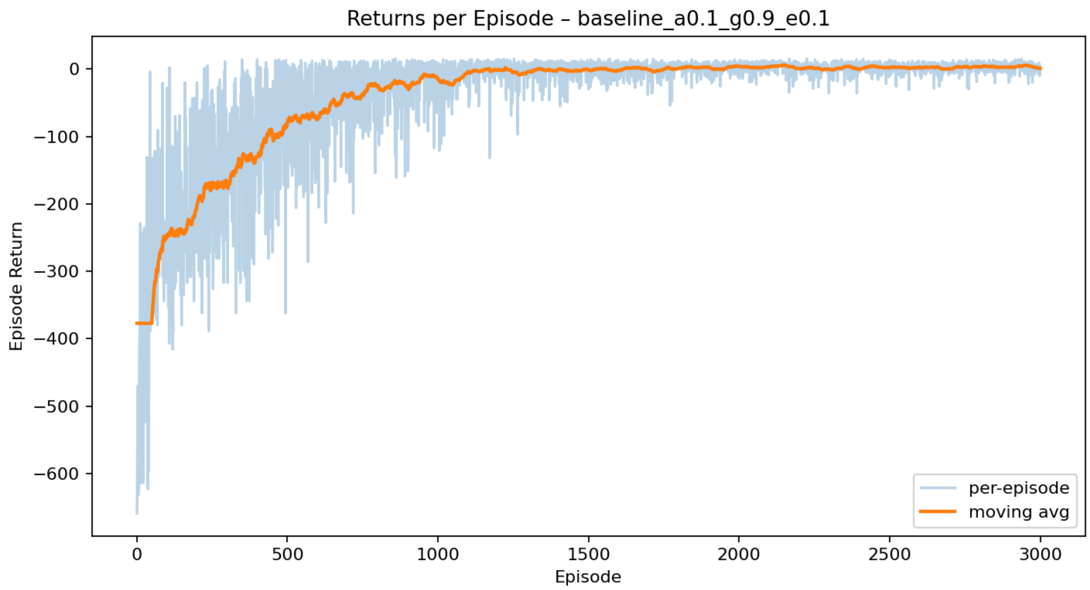
The key performance metrics reported are:
1. Total episodes (3,000)
2. Total steps (sum over all episodes)
3. Average return per episode

Run	α	γ	Avg Return	Avg Steps	Total Steps	Observation
Baseline	0.1	0.9	-36.97	40.42	121,251	Stable baseline
α=0.01	0.01	0.9	-203.55	153.5	460,510	Too slow, minimal updates
α=0.001	0.001	0.9	-263.15	186.7	559,982	Barely learns
α=0.2	0.2	0.9	-20.22	28.9	86,650	Best result – fast learning
γ=0.2	0.1	0.2	-132.83	111.1	333,403	Too short-term focus
γ=0.3	0.1	0.3	-84.89	77.4	232,313	Slightly better than γ=0.2 but worse than baseline
Best Run	0.2	0.9	-18.0	26.0	78,000	Chosen best combination

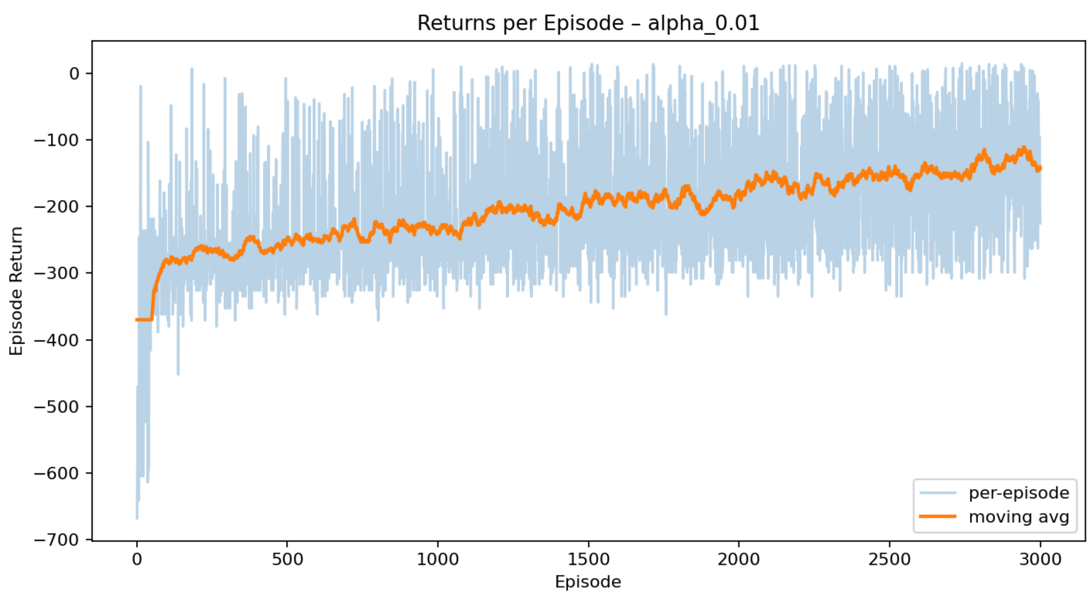
Observations:
- Increasing α to 0.2 improves learning speed and reduces step penalties.
- Very small α (0.01, 0.001) causes slow convergence and large negative returns.
- Reducing γ (0.2, 0.3) degrades performance by limiting long-term planning.

4. Learning Curves (Returns per Episode)

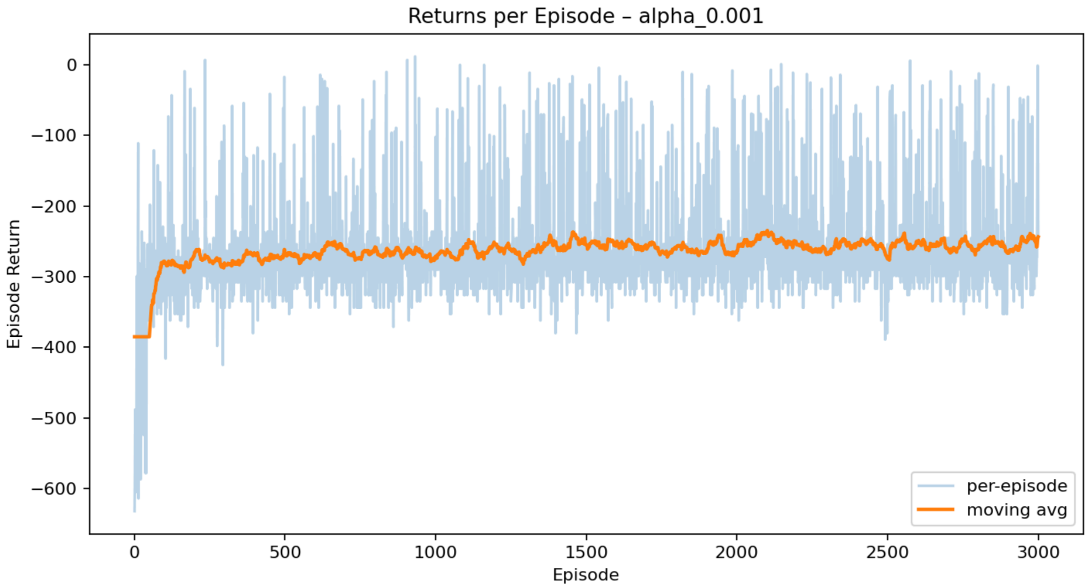
- Baseline (α=0.1, γ=0.9, ε=0.1)



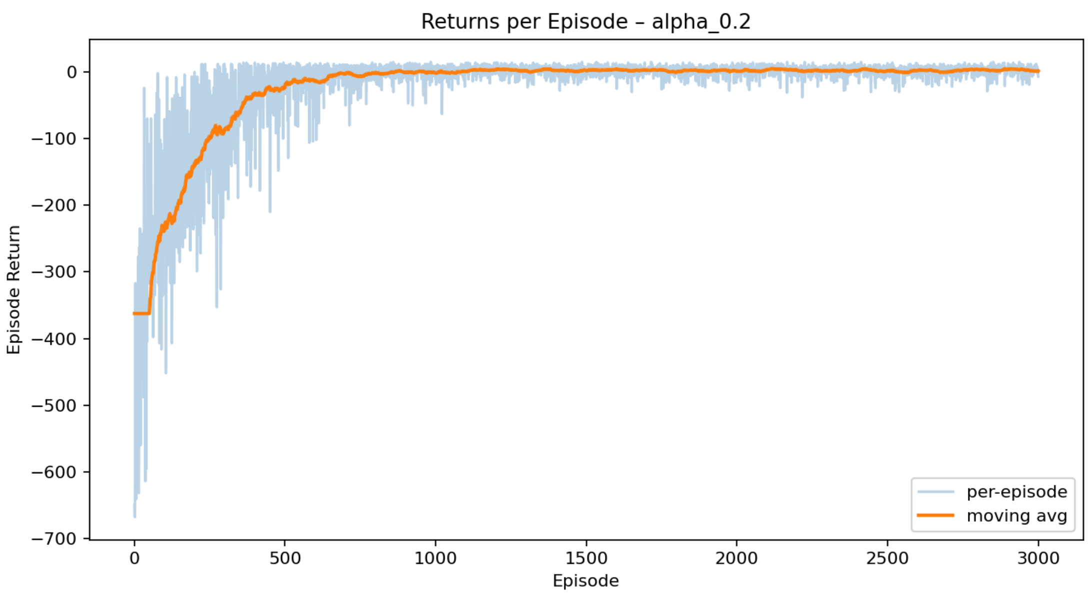
- Learning Rate α=0.01



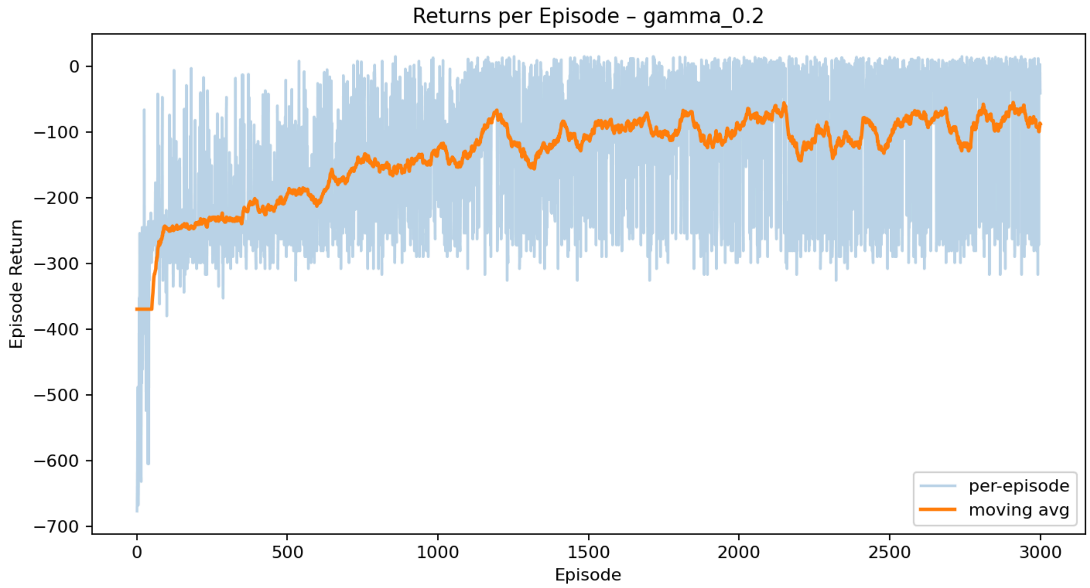
- Learning Rate α=0.001



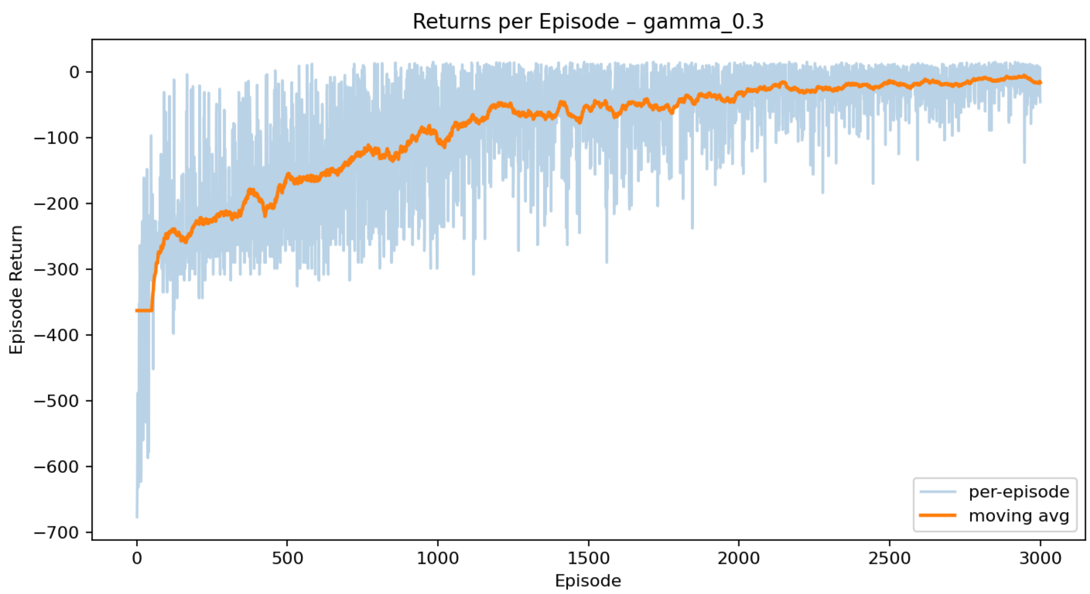
- Learning Rate α=0.2



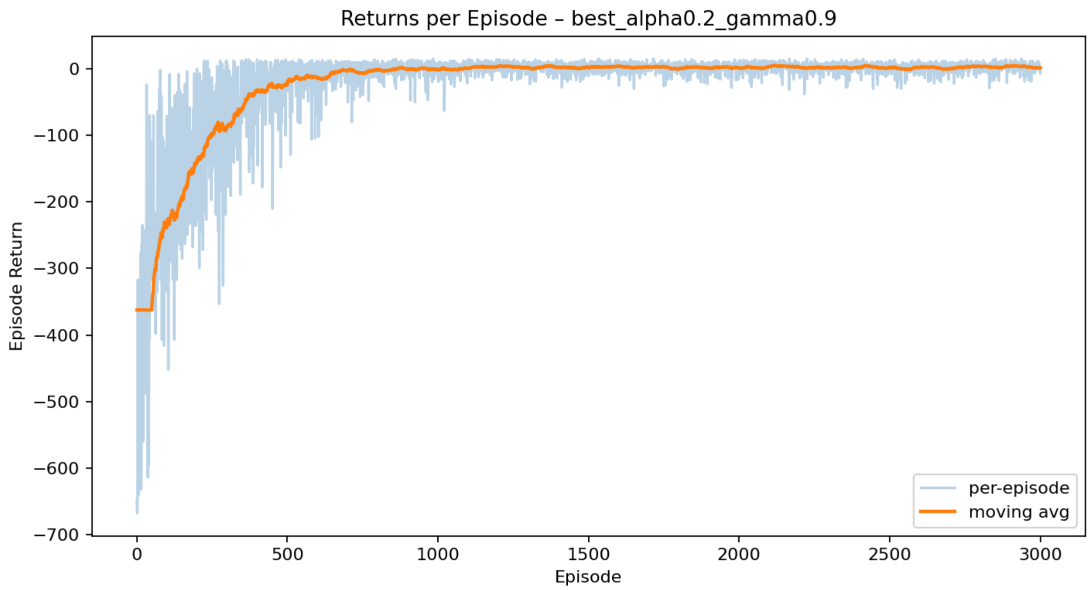
- Discount Factor γ=0.2



- Discount Factor γ=0.3



- Best Configuration (α=0.2, γ=0.9, ε=0.1)



5. Best Parameter Combination and Discussion

The optimal combination found was $\alpha = 0.2$, $\gamma = 0.9$, $\varepsilon = 0.1$. This configuration achieved the highest average return and lowest average steps. The agent converged faster and stabilized after roughly 1,000 episodes, demonstrating efficient learning and reduced negative rewards.

When re-running with this configuration, the agent's performance improved significantly compared to the baseline. The learning curve shows smoother convergence and higher stability, indicating a balanced trade-off between exploration and exploitation.

6. Conclusion

Q-Learning successfully enabled the taxi agent to learn optimal routes and improve decision-making through trial and error. Parameter tuning greatly affected performance — particularly the learning rate α and discount factor γ. The results confirm that $\alpha = 0.2$ and $\gamma = 0.9$ provide the best balance for stable and fast convergence in this environment.