Reinforcement Learning Unveiled: Q-Learning in Practice

Name: Jasser Abdelfattah, UH ID: 21033101

December 21, 2024

1 Introduction

This report investigates the performance and behavior of agents employing Q-learning within reinforcement learning environments. Q-learning, a widely used model-free reinforcement learning algorithm, enables agents to derive optimal policies by interacting with their environment. The central principle of reinforcement learning is to guide agents through a reward-punishment system, reinforcing desired behaviors while discouraging undesired ones.

2 Q-Learning Parameters

2.1 Learning Rate (α)

- **Definition:** Specifies the weight assigned to new information during Q-value updates.
- **Impact:**
  - α close to 1 (e.g., 0.9): Promotes fast learning but may lead to instability.
  - α close to 0 (e.g., 0.01): Results in slow but steady learning.
- **Default Value:** $\alpha = 0.1$, providing a balance between learning speed and stability.

2.2 Discount Factor (γ)

- **Definition:** Governs the importance of future rewards relative to immediate rewards.
- **Impact:**
  - High γ (close to 1): Encourages long-term planning.
  - Low γ (close to 0): Focuses on short-term gains.
- **Default Value:** $\gamma = 0.9$, emphasizing future rewards significantly.

2.3 Exploration Rate (ϵ)

- **Definition:** Balances exploration (random actions) and exploitation (choosing the best-known action).
- **Impact:**
  - High ϵ (e.g., 0.9): Promotes extensive exploration, useful for discovering new strategies.
  - Low ϵ (e.g., 0.01): Prioritizes exploitation, leveraging learned strategies but risking suboptimal solutions.
- **Default Value:** $\epsilon = 0.1$, ensuring a predominance of exploitation with occasional exploration.

2.4 Parameter Interactions

- α and γ : A high γ requires a carefully balanced α to prevent unstable updates.
- ϵ and α : Lower ϵ increases exploitation, necessitating a stable α for robust learning.
- Dynamic Adjustments: Techniques like *epsilon decay*, where ϵ reduces over time, can improve learning efficiency.

3 Single-Agent Learning Behavior

3.1 Initial Phase (Episodes 0–200)

- Initial rewards are low (-800), reflecting poor performance.
- Rapid improvement occurs as the agent explores and learns.

3.2 Middle Phase (Episodes 200–600)

- Rewards show a consistent upward trend and increased stability.
- The agent transitions from exploration to exploiting its learned policy.

3.3 Convergence Phase (Episodes 600–1000)

- Rewards stabilize near 0, indicating near-optimal policy learning.
- Variations in rewards are minor, driven by occasional exploration.

3.4 Key Observations

- **Learning Speed:** Rapid improvement occurs in the first 200 episodes.
- **Stability:** Convergence is achieved by episode 600.
- **Efficiency:** Task completion becomes faster over episodes.

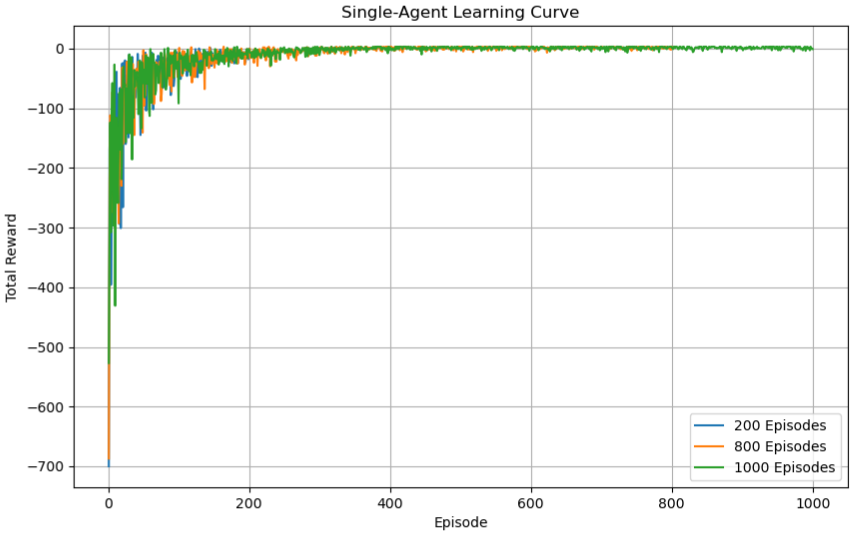

Figure 1: Single-Agent Learning Curve

4 Multi-Agent Learning Behavior

4.1 Learning Performance

- All agents converge to a stable reward close to 0 after approximately 300 episodes.
- The initial learning phase exhibits rapid improvement, followed by gradual stabilization.
- Post-convergence, stable rewards indicate consistent performance.

4.2 Collaboration or Competition

- **Collaboration:**
  - Smooth convergence may suggest cooperative behavior among agents.
  - Minimal reward fluctuations post-convergence reflect synchronized learning.
- **Competition:**

- Overlapping curves indicate stable competitive strategies with no major interference.
- Absence of reward drops suggests agents avoid detrimental competition.

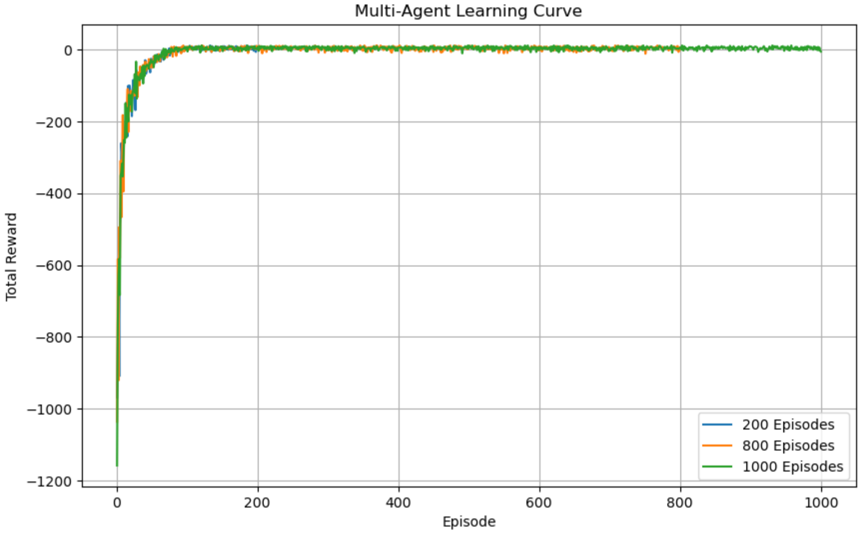

Figure 2: Multi-Agent Learning Curve

4.3 Additional Observations

- Results from 200, 800, and 1000 episodes are nearly identical post-convergence, suggesting fewer episodes may suffice.
- Analysis of individual agent strategies could provide deeper insights.

5 Conclusion

The analysis demonstrates that agents employing Q-learning improve performance efficiently, achieving stable and consistent policies over time. The results emphasize the importance of parameter tuning and the potential for enhanced strategies in multi-agent systems.