Reinforcing Learning. Q-learning for Blackjack

Egor Sementul (i6290310), Maja Gójska (i6274446)

March 22, 2024

Abstract

1 Introduction

In this report, we will explore a Q-learning method, a famous reinforcement learning technique, to design an intelligent agent for the game of blackjack. Reinforcement learning enables agents to learn optimal behaviors through interactions with their environment. Q-learning, a model-free reinforcement learning algorithm, offers a framework for agents to learn the optimal value of actions in various states without requiring a model of the environment (Watkins and Dayan 1992). This makes Q-learning particularly suited for complex problems like blackjack, where the agent must make decisions based on incomplete information and dynamic conditions.

2 Research Questions

Our research aims to explore the effects of Q-learning parameters: namely the Learning Rate (α), Discount Factor (γ), and Exploration Rate (ϵ) on the performance and learning efficiency of Q agent.

(1) Optimizing Parameter Combinations for Peak Performance
Research Question: What is the optimal combination of the α, γ and ϵ parameters to achieve the highest performance in a Q-learning agent?
Hypothesis: We hypothesize that there exists a unique combination of parameters that maximizes the Q-learning agent's efficiency and effectiveness in learning the target policy. This optimal combination minimizes the time to convergence while ensuring the stability of the learning process.

(2) Impact of Individual Parameters on Agent Learning
Research Question: How do variations in the α , γ and ϵ individually affect the learning and performance of a Q-learning agent?
Hypothesis: We hypothesize that each parameter uniquely influences the Q-learning agent's learning trajectory and efficiency.

(3) Dynamic Exploration Strategies in Learning
Research Question: How does a dynamically decreasing exploration rate throughout the learning process affect the Q-learning agent's performance compared to a static exploration rate, and what are the implications for convergence and learning efficiency?
Hypothesis: We hypothesize that a dynamic exploration rate, which decreases as the agent learns, will lead to more efficient learning process by initially encouraging exploration and gradually shifting towards exploitation of the learned policy. For this approach we anticipate to achieve faster convergence and a more robust final policy compared to a static exploration rate.

3 Experiments and Results

3.1 Optimizing Parameter Combinations

To identify the optimal configuration of the parameters α, γ, and ϵ, each within the range of 0 to 1 in increments of 0.1, we used a grid search strategy. This involved running a distinct Q-learning agent for every possible parameter combination across 1000 games. The performance of each agent was assessed by calculating the average reward throughout the games played. The combination of $\alpha = 0.3$, $\gamma = 0.5$, and $\epsilon = 0.3$ was the most effective, achieving an average reward of -0.03. This is notably better when contrasted with the overall average reward of -0.1306572 for all agents and is a substantial improvement over the lowest observed average reward of -0.257.

This configuration demonstrates the highest efficiency in balancing the exploration-exploitation trade-off and optimizing long-term rewards. Learning process of the Q-Agent with this optimal parameter combination is depicted in the Figure 1.

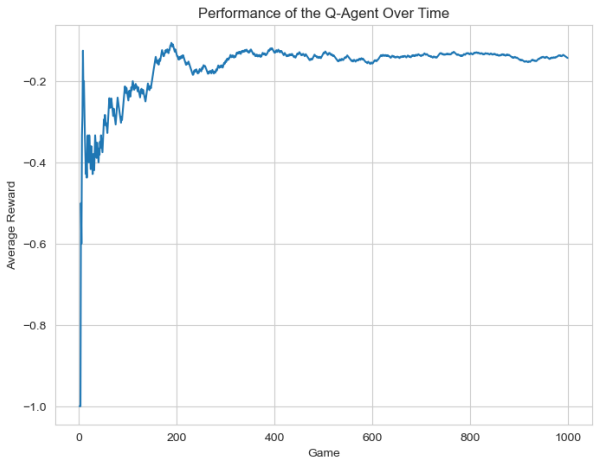

Figure 1: Performance of the Q-Agent with Optimal Parameters Over Time

3.2 Impact of Individual Parameters on Agent Learning

In this set of experiments, we used the previously identified optimal parameter set.By varying one parameter while holding the others fixed, we assessed the individual impact on the Q-learning agent's performance.

Analysis began with the learning rate, α. As Figure 2 illustrates, the peak performance occurs at $\alpha = 0.3$.

Furthermore the graph reveals a non-linear relationship between α and the average reward. Notably, poor rewards are recorded at the lower α values ($\alpha \in \{0.1, 0.2\}$). This might be due to underfitting, when the agent does not learn adequately from its experiences. On the other hand values of α above 0.3 are also suboptimal most likely due to overfitting when agent does not perform well across a broader range of scenarios due to a lack of generalization.

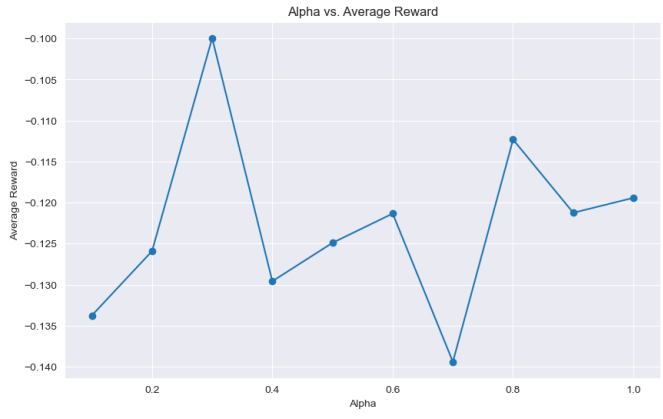

Figure 2: Comparison of Performance of the Q-Agent with Different Values of α Parameter

Next, we examined the discount factor, γ, with findings illustrated in Figure 3. From the plot we can notice a sharp decline in average reward as γ increases beyond 0.4, with the lowest point around γ of 0.6, indicating that overemphasizing future rewards may negatively impact the agent's performance. Surprisingly, the average reward increases again as γ approaches 0.8 before falling at 1.0, showing a non-monotonic relationship between γ and performance. This suggest variability in the average reward with changes in γ.

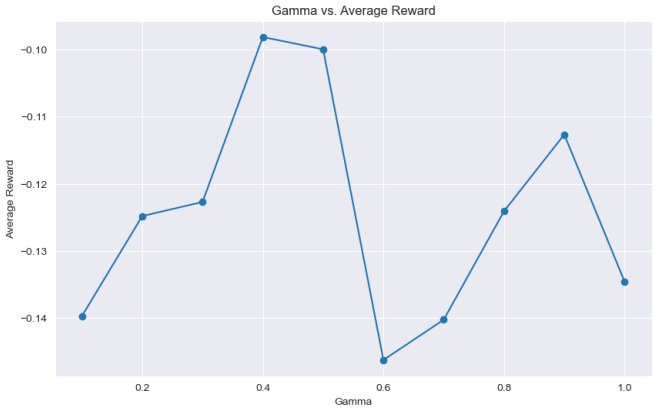

Figure 3: Comparison of Performance of the Q-Agent with Different Values of γ Parameter

Finally, we analyzed the exploration rate, ϵ, and its effects, as detailed in Figure 4. A peak performance at $\epsilon = 0.3$ suggests optimal balance between exploration and exploitation. As ϵ increases towards 1, the average reward generally decreases, indicating that excessive exploration may detract from the agent's performance. This variation in rewards across different ϵ values points to the exploration's complex and non-linear influence on learning.

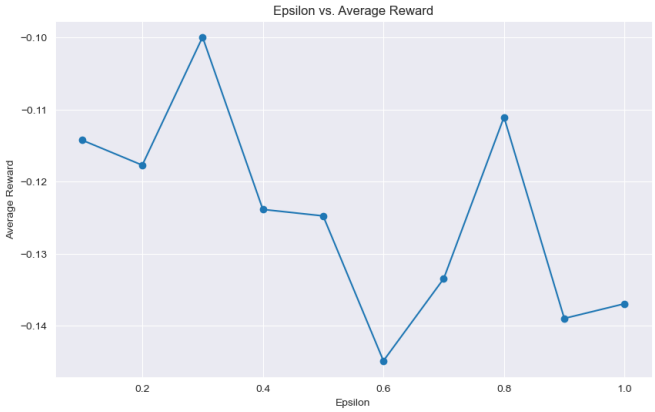

Figure 4: Comparison of Performance of the Q-Agent with Different Values of ϵ Parameter

3.3 Dynamic Exploration Strategies in Learning

In order to assess the impact of dynamically decreasing exploration rates on the performance of a Q-learning agent, a comparative analysis was conducted between agents utilizing a static exploration rate and those with rates that decrease as learning progresses. The performance was evaluated based on the average reward over a series of games, with the expectation that dynamic rates would promote a balance between exploration and exploitation, potentially resulting in more efficient learning and faster convergence to robust policies.

A dataset encompassing 5000 games across 11 different exploration rates, including a static rate, was collected. The average reward for each game was recorded, resulting in a dataset for analysis. To test the hypothesis that dynamic exploration rates lead to improved learning efficiency, statistical tests were performed comparing the final average rewards of agents with dynamic rates against those with the static rate.

The hypothesis testing yielded significant results, with p-values such as 6.58×10^{-132} and many others effectively zero, indicating a robust statistical difference between static and dynamic exploration rates in terms of their impact on the Q-learning agent's performance.

An illustrative depiction of the learning process for the various decay rates can be observed in Figure 5. This plot encapsulates the average reward trends over the initial 1000 games, providing visual substantiation of the impact that the exploration rate has on the learning dynamics of the agents.

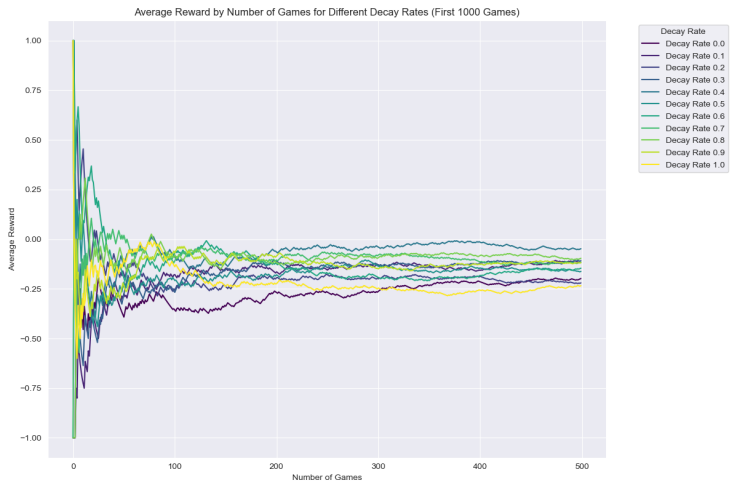

Figure 5: Average reward trends under various exploration decay rates.

4 Conclusion

In conclusion we have supported our hypothesis that dynamic exploration rate strategy significantly enhances the agent's performance in a blackjack environment. Statistical tests confirmed with high confidence that dynamic rates improve average rewards compared to a static approach. This suggest that an adaptive exploration strategy can improve the learning process, leading to faster convergence and more robust policy formation.

Furthermore, we have found that there indeed exists an optimal combination of the parameters yielding the greatest performance of the Q-learning agent: $\alpha = 0.3$, $\gamma = 0.4$, and $\epsilon = 0.3$. This specific values effectively balance the exploration-exploitation trade-off and mitigates the risks of overfitting and underfitting.

Lastly, our investigation also clarified the individual impact of each parameter on the agent's learning and performance. In the complex environment and task presented by the game of blackjack, the influence of these parameters proved to be non-linear, revealing the nuanced challenges of tuning the Q-learning model for optimal results.

References

Watkins, Christopher J. C. H. and Peter Dayan (May 1992). "Q-learning". In: *Machine Learning* 8.3, pp. 279–292. DOI: 10.1007/BF00992698. URL: https://doi.org/10.1007/BF00992698.