Exploring Reinforcement Learning: Multi-Armed Bandits with Varied Strategies

Rusmia Sattar[†]

[†] *Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's (NL) A1C 5S7, Canada*

E-mail: rsattar@mun.ca

Abstract

The project investigates the field of reinforcement learning using ten-armed bandits problem with various strategies. It begins by analyzing the performance of greedy and epsilon-greedy strategies on a single set of ten-armed bandits with randomly generated action values. The experiment runs for 1000 steps, with epsilon values of 0.01, 0.1, and 0.2. The project then expands its scope by repeating the experiment across 2000 different ten-armed bandit scenarios, each featuring unique action values. This broader analysis provides insights into the adaptability and effectiveness of different strategies when facing diverse bandit configurations. Overall, this project advances our understanding of decision-making processes in artificial intelligence and machine learning.

Keywords: Reinforcement learning, Bandits, epsilon, epsilon-greedy, strategies

1 Introduction

Reinforcement learning is a dynamic subfield of artificial intelligence which has gained widespread recognition for its capacity to solve complex decision-making problems. Within this domain, multi-armed bandit problems stand as foundational models representing the trade-off between exploration and exploitation. They serve as an ideal platform for understanding and experimenting with reinforcement learning strategies.[1]

The development of computing power and growing interest in machine learning in the late 20th century led to a major surge in interest in reinforcement learning. The advancement of reinforcement learning in numerous fields has been expedited by the creation of algorithms like policy gradients, Q-learning, temporal difference learning, and Deep Q Networks (DQN). The convergence of these many domains, together with ongoing developments in computer power, technology, and algorithmic design, played a role in the current surge in the development and usefulness of reinforcement learning. It is still developing and making great progress toward resolving challenging issues in a variety of industries, including robotics, gaming, finance, healthcare, and more.[1]

This project embarks on an insightful journey to investigate the effectiveness of different strategies when faced with diverse bandit scenarios. The study encompasses both the fundamental greedy strategy and the versatile epsilon-greedy approach.[2] Our exploration unfolds in two significant phases. Initially, we set the stage with a solitary set of ten-armed bandits, where each action's true value is drawn from a normal distribution with a mean of 0 and a variance of 1. We execute a 1000-step experiment, varying the epsilon parameter across values of 0.01, 0.1, and 0.2. This phase offers a closer look at how these strategies perform in a controlled environment. Furthermore, we extend our viewpoint by repeating the experiment over 2000 different ten-armed bandit problems. In these cases, action values come from normal Gaussian

distributions with a variance of 1 and a mean of 0. We obtain a thorough grasp of the techniques' flexibility and efficacy when faced with a variety of bandit configurations by combining and evaluating the data from several trials.

This research aims to find useful insights applicable to real-world decision-making processes in a variety of domains, from recommendation systems to autonomous agents, along with making a theoretical contribution to the study of reinforcement learning.[3] We hope to shed light on the strategies' capacity to negotiate the fine line between exploration and exploitation in a heterogeneous and dynamic environment by exploring the dynamics of multi-armed bandit issues.[4]

2 Methods

The method employed in this project involved the generation of diverse multi-armed bandit scenarios, the implementation of epsilon-greedy strategies, and the analysis of resulting rewards to investigate the exploration-exploitation trade-off. This project operates within the realm of a fundamental problem in reinforcement learning known as the multi-armed bandit problem. Initially, we set the stage by establishing parameters as 10 bandit arms, the duration of the problem in 1000 time steps, and 2000 iterations or runs to collect data. The essential goal is to make decisions that maximize cumulative reward when confronted with several actions or bandits.

Here we start with writing a Python code that provides two different action selection strategies: a greedy strategy, which primarily exploits the action with the highest expected value, and an epsilon-greedy strategy, which balances exploration and exploitation.[5] These strategies can help the agent to decide which bandit to choose at each time step. To do that, we generate the greedy and epsilon greedy action.

```
def greedy(Q, epsilon):
    if np.random.rand() < epsilon:
        return np.random.choice(num_bandits)
    else:
        return np.argmax(Q)

def epsilon_greedy(Q, epsilon):
    if np.random.rand() < epsilon:
        return np.random.choice(num_bandits)
    else:
        return np.argmax(Q)
```

Then we generate the bandit problem for various scenarios to maximize the rewards.

```
    def run_bandit(epsilon):
    total_rewards = np.zeros(num_steps)
    for _ in range(num_runs):
        Q = np.zeros(num_bandits)
        action_counts = np.zeros(num_bandits)
        rewards = []

        for step in range(num_steps):
            action = epsilon_greedy(Q, epsilon)
```

```
            reward = np.random.normal(true_action_values[action], 1)
            rewards.append(reward)
            action_counts[action] += 1
            Q[action] += (reward - Q[action]) / action_counts[action]

        total_rewards += np.array(rewards)

    average_rewards = total_rewards / num_runs
    return average_rewards
```

It simulates the bandit play by generating rewards based on the action chosen at each time step. The rewards are normally distributed around a mean, simulating the real-world variance in reward outcomes. It then accumulates the rewards over multiple runs to compute average rewards, specifically focusing on the last time step of 1000 steps. The process is conducted for different epsilon values, which control the degree of exploration, allowing an analysis of their influence on the overall rewards gained.

Moreover, the code extends the analysis by altering the initial value estimates (Q1(a)). By changing these initial values to very optimistic (Q1(a) = 5) or extremely pessimistic (Q1(a) = -5) settings, it aims to explore how these initial assumptions can guide the agent's behavior. This comparison seeks to uncover how initial estimations influence the learning process, impacting the choices made by the agent and, consequently, the rewards garnered. This method served as the foundation for the project's findings and insights into reinforcement learning strategies in bandit problems.

3 Results

The results indicate the average reward obtained for 1000 steps using different epsilon-greedy strategies in the context of multi-armed bandit problems. The exploration of the results illustrates the influence of the epsilon-greedy strategies on reward outcomes in multi-armed bandit problems. Each strategy operates with a distinct epsilon value. Firstly, the project explores the impact of different epsilon values on average rewards over 1000 steps in epsilon-greedy strategies for multi-armed bandit problems interpreted below:

Epsilon = 0.01 reveals a conservative approach, favoring exploitation over exploration, resulting in an average reward of 0.8136. This lower exploration rate leads to a more consistent selection of actions deemed as the best based on initial estimations. At Epsilon = 0.1, the strategy strikes a balance between exploration and exploitation. This moderate epsilon value allows the occasional exploration of other actions while still favoring those with higher estimated values. This balance yields a slightly higher average reward of 0.8995, suggesting a more effective learning process and improved decision-making. Meanwhile, Epsilon = 0.2 significantly increases the exploration aspect of the strategy, leading to a decreased average reward of 0.7851. The higher exploration rate diverts the strategy from actions with higher estimated values, emphasizing the trade-off between exploration and exploitation. A lower epsilon leans towards exploitation, while a higher value leans more towards exploration. The moderate epsilon value of 0.1 appears to strike the best balance, resulting in higher average rewards over 1000 steps, implying that an optimal mix of exploration and exploitation often proves most effective in multi-armed bandit scenarios.

In the second part of the project, we attempt to find the average reward over 2000 Bandit

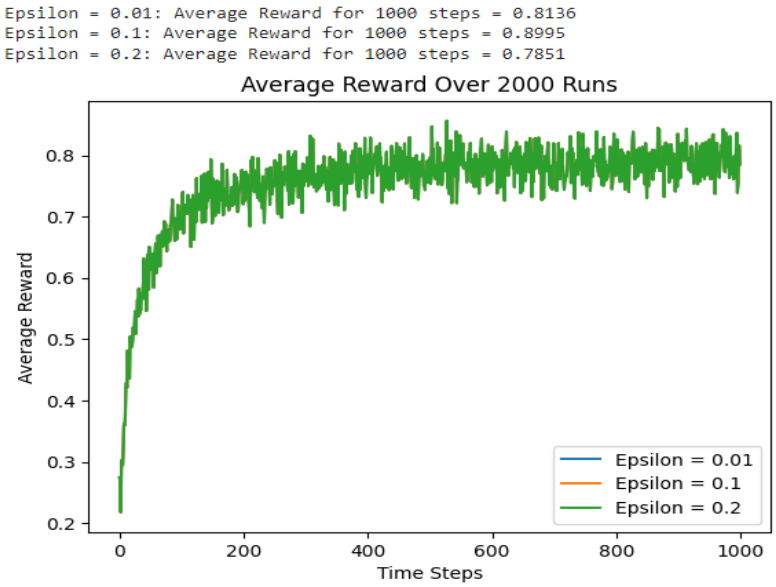

Figure 1. Average Reward over 2000 runs

Instances. Here, the results demonstrate the impact of different epsilon values on the average rewards obtained over 1000 steps in epsilon-greedy strategies, offering insights into exploration and exploitation trade-offs in multi-armed bandit problems. With Epsilon = 0.01, the strategy tends to be highly conservative, leaning towards exploitation over exploration. This cautious approach results in an average reward of 1.3306. The strategy predominantly chooses actions it deems best based on initial estimations, potentially missing out on higher-reward actions. At Epsilon = 0.1, the strategy finds a middle ground, enabling both exploration and exploitation. This balanced epsilon value allows occasional exploration of other actions while favoring those with higher estimated values, leading to a slightly improved average reward of 1.4257. This suggests a more effective learning process and enhanced decision-making.

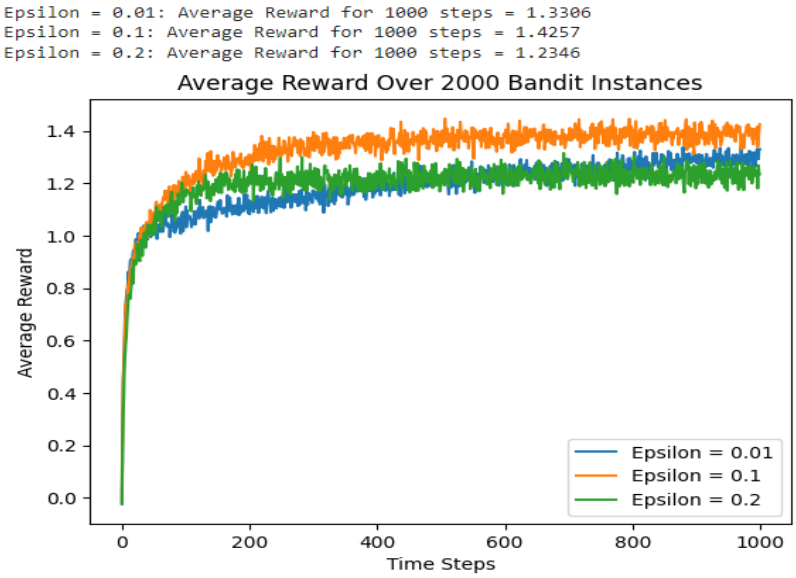

Figure 2. Average Reward over 2000 Bandit Instances

However, Epsilon = 0.2 significantly emphasizes exploration over exploitation. While exploration diversifies the strategy's choices, it leads to a reduced average reward of 1.2346. The strategy's increased tendency to deviate from actions with higher estimated values highlights the trade-off between exploration and exploitation. In essence, these findings underscore the critical influence of the epsilon value in epsilon-greedy strategies. A lower epsilon prioritizes exploitation, while a higher value prioritizes exploration. The moderate epsilon value of 0.1 demonstrates a superior balance, yielding higher average rewards over 1000 steps, indicating that an optimal blend of exploration and exploitation is more effective in multi-armed bandit scenarios.

Lastly, on the final part of our project we try to interpret the influence of initial value, q1 = 5 and -5 and estimate the average reward over 2000 Bandit Instances. Here, the outcomes present the influence of both initial values (Q1) and the epsilon parameter on the performance of epsilon-greedy strategies in multi-armed bandit problems over 1000 steps. The initial value estimates (Q1) set the starting point for the agent's confidence in its action values before any actions are taken. These values influence the agent's inclination to explore or exploit its environment. When Initial Q1 is set to 5, the agent starts with an optimistic estimate, assuming actions have high values over real values. However, this could lead to reduced exploration. The agent might be less motivated to explore further as it believes it has already found a high-reward action, potentially overlooking better options. Conversely, setting Initial Q1 to -5 introduces pessimism, assuming actions have lower values. In this scenario, the agent might initially explore more in search of higher rewards. However, an extremely pessimistic estimate could overly emphasize exploration, and this may hinder the agent's ability to exploit high-reward actions it may come across early in the process.

```
Epsilon = 0.01, Initial Q1 = 5: Average Reward for 1000 steps = 0.2634
Epsilon = 0.01, Initial Q1 = 0: Average Reward for 1000 steps = 0.2510
Epsilon = 0.01, Initial Q1 = -5: Average Reward for 1000 steps = 0.0330
Epsilon = 0.1, Initial Q1 = 5: Average Reward for 1000 steps = -0.0100
Epsilon = 0.1, Initial Q1 = 0: Average Reward for 1000 steps = 0.0853
Epsilon = 0.1, Initial Q1 = -5: Average Reward for 1000 steps = 0.0284
Epsilon = 0.2, Initial Q1 = 5: Average Reward for 1000 steps = 0.0105
Epsilon = 0.2, Initial Q1 = 0: Average Reward for 1000 steps = -0.0343
Epsilon = 0.2, Initial Q1 = -5: Average Reward for 1000 steps = 0.0302
Epsilon = 0.01, Initial Q1 = 5: Average Reward for 1000 steps = 1.4335
Epsilon = 0.01, Initial Q1 = 0: Average Reward for 1000 steps = 1.2649
Epsilon = 0.01, Initial Q1 = -5: Average Reward for 1000 steps = 1.2120
Epsilon = 0.1, Initial Q1 = 5: Average Reward for 1000 steps = 1.3446
Epsilon = 0.1, Initial Q1 = 0: Average Reward for 1000 steps = 1.3950
Epsilon = 0.1, Initial Q1 = -5: Average Reward for 1000 steps = 1.3905
Epsilon = 0.2, Initial Q1 = 5: Average Reward for 1000 steps = 1.1944
Epsilon = 0.2, Initial Q1 = 0: Average Reward for 1000 steps = 1.2261
Epsilon = 0.2, Initial Q1 = -5: Average Reward for 1000 steps = 1.2764
```

Figure 3. Average Rewards for 1000 steps for initial values, Q1 = 0, 5, -5

At Epsilon = 0.01 (low exploration), Initial Q1 of 5 presents a modest average reward, indicating the early overestimation of action values. On the other hand, setting Initial Q1 to -5 results in an even lower reward, reflecting the conservative exploration due to the initial pessimism. Under Epsilon = 0.1 (a more balanced exploration-exploitation strategy), both Initial Q1 values, 5 and -5, yield similarly low rewards. This implies that the balance between exploration and exploitation primarily dictates the agent's decisions, overshadowing the initial

estimates. With Epsilon = 0.2 (a higher focus on exploration), Initial Q1 of 5 generates slightly higher rewards. This indicates that while the initial optimistic estimate influences the initial actions, the increased exploration is beneficial, helping the agent learn from potential higher-reward actions. Conversely, at Initial Q1 of -5, the rewards remain low, highlighting the negative impact of excessive initial pessimism on the agent's ability to learn effectively. In summary, the initial value estimates influence the agent's actions by dictating its confidence levels and, to some extent, the balance between exploration and exploitation. While optimism or pessimism can guide the agent's early decisions, a balanced approach, influenced more by the epsilon parameter, often yields better long-term rewards.

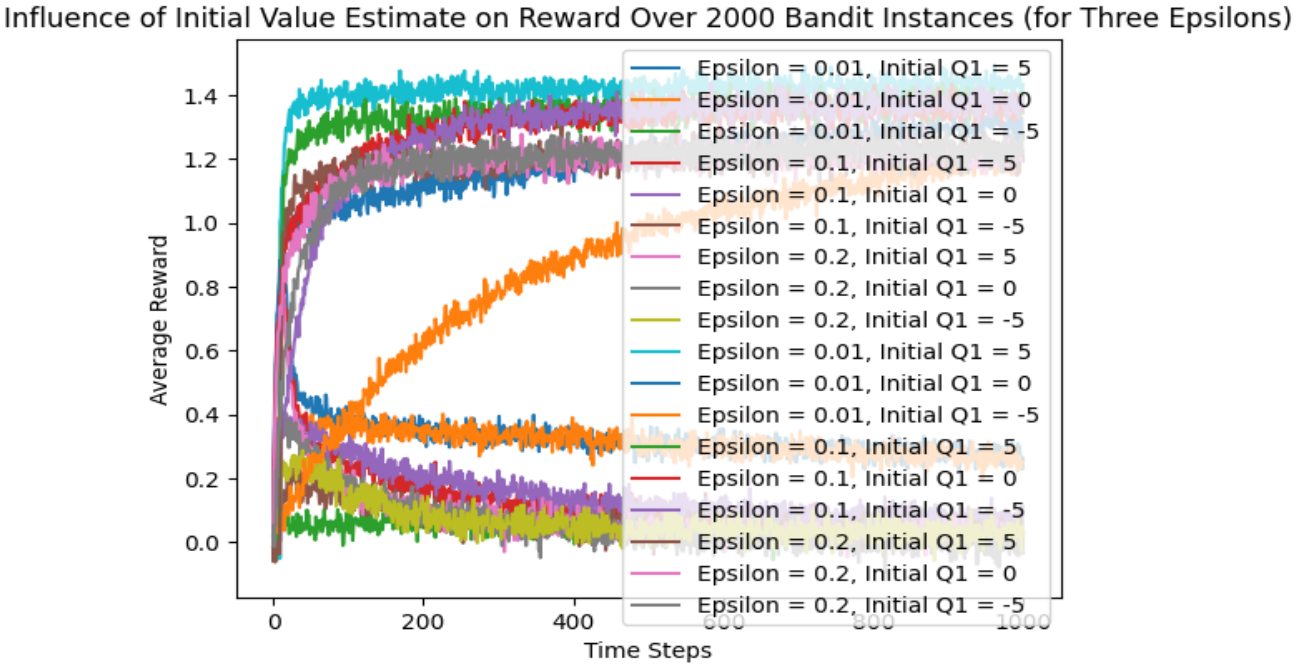

Figure 4. Influence of Initial Value Estimate on Average Rewards

The study's three parts analyze how different parameters affect the strategies of the epsilon-greedy approach in multi-armed bandit problems. The initial experiment reveals the impact of varying epsilon values (0.01, 0.1, and 0.2), demonstrating that a moderate epsilon (0.1) offers the most balanced and effective strategy over 1000 steps. For the random nature of algorithm, each time we run the code, we might get slightly different rewards but not significantly much difference are seen each time. The subsequent exploration of epsilon with varying initial Q1 values (5 and -5) reinforces the critical balance between exploration and exploitation. A balanced approach, especially when Epsilon = 0.1, shows consistent efficacy in gaining higher long-term rewards.

These results have practical applications in a variety of situations. Epsilon-greedy techniques, for example, assist achieve a balance between taking use of established effective recommendations and investigating novel, perhaps better suggestions in recommendation systems, such as content recommendation systems for users. While avoiding stagnation or drastic changes that can annoy consumers, a moderate approach permits incremental modification based on user preferences.[1]This balance is essential in many real-world situations because determining the best long-term outcomes requires balancing the investigation of new approaches versus the application of successful strategies for optimal long-term outcomes.

4 Conclusion

In this project, we used multi-armed bandit situations as a testing ground to conduct a thorough investigation of reinforcement learning. The balance between exploitation such as selecting known, high-reward activities to maximize profits, and exploration is the core conundrum in these challenges. The popular strategy known as "epsilon-greedy" techniques, which differ in their exploration rate (epsilon), was the subject of our investigation.

After conducting numerous tests, we came to a number of significant findings. First and foremost, a key factor in the effectiveness of these tactics is the selection of epsilon. Over time, we found that the highest average payouts were frequently obtained with a moderate epsilon value of 0.1. This suggests that optimizing cumulative benefits can be achieved more successfully with a balanced strategy that incorporates both exploration and exploitation.

Additionally, the techniques were quite sensitive to epsilon. It is important to precisely adjust this parameter in practical applications because little changes in it can have a big impact on the results. Even though our experiment focused on a simplified scenario, real-world domains are affected by its consequences. Decisions in internet advertising, autonomous agents, and recommendation systems are continuously weighing the trade-off between exploration and exploitation.

Our investigation emphasizes the fine art of achieving this precarious equilibrium in the quest for optimal decision-making, adding to the growing volume of knowledge in reinforcement learning. Essentially, the research offers a basic grasp of reinforcement learning and its application in real-world scenarios where decisions frequently center around the exploration-exploitation paradox.

Acknowledgements

I gratefully acknowledge the cordial help of my respected Professor and the valuable assistance of my peers for this project.

References

[1] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[2] N Hariharan and Anand G Paavai. A brief study of deep reinforcement learning with epsilon-greedy exploration. *International Journal of Computing and Digital Systems*, 11(1):541–552, 2022.

[3] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

[4] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory: 20th International Conference, ALT 2009, Porto, Portugal, October 3-5, 2009. Proceedings 20*, pages 23–37. Springer, 2009.

[5] Epsilon-Greedy Algorithm in Reinforcement Learning - GeeksforGeeks — geeksforgeeks.org. https://www.geeksforgeeks.org/epsilon-greedy-algorithm-in-reinforcement-learning/. [Accessed 12-11-2023].