Introduction to Reinforement learning using Bandit Problem

Saurav Singh Chandel[†]

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

E-mail: sschandel@mun.ca

Abstract: In our research, we explored how computers can learn to make smart decisions, much like a person might learn through trial and error. Our focus was on a problem called the multi-armed bandit, where a computer tries to figure out which of ten slot machines gives the best rewards, without any prior knowledge. We used a method known as the ε-greedy strategy, which mixes sticking with the best-known option so far and trying new ones. We were particularly interested in how the strategy performs when we start with two different types of guesses: all zeros or random numbers. Our tests, carried out across 2000 simulations for each setup, showed some surprising results. When we set ε (the measure of exploration) to 0.2 and used random initial guesses, the strategy did better than when we started with zeros. This was particularly true for $\varepsilon = 0.2$, suggesting that a bit of randomness in the beginning can be helpful, especially when we're more open to trying new options. However, this benefit wasn't as clear for lower ε values. This study highlights how the starting point and the willingness to explore can significantly affect how well and how quickly computers can learn in situations where they have to make decisions based on incomplete information.

Keywords: Reinforcement Learning, K-armed Bandit, Epsilon-Greedy

1 Introduction

Reinforcement learning is like a self-improvement journey for computers. Instead of having everything figured out from the start, the computer learns by trying different things and getting rewards. It's a bit like teaching a robot to navigate the world and make smart decisions.

Now, imagine a situation called the multi-armed bandit problem. It's like a game where you have several slot machines, each with a different prize. The catch is, you don't know which machine gives the best prize. The challenge is to figure that out by trying different machines and balancing between sticking to what seems good and trying new options.

In our study, we narrow our focus to ten of these metaphorical slot machines, each endowed with its unique potential for rewards. At the outset, our artificial intelligence agent is blissfully unaware of which machine holds the key to the highest payouts. To unravel this mystery, we've devised two strategies for our digital explorer.

The first strategy adopts a "*greedy*" approach, where the agent consistently chooses the option that appears most promising in the current context. The second strategy introduces an element of randomness, occasionally prompting the system to try new options, even when certainty is lacking. Additionally, we're exploring the impact of starting with an optimistic view, essentially injecting a burst of confidence into the system from the beginning.

Our main goal is to understand how well these strategies work. It's like watching a robot learn how to play a slot machine game and figuring out which approach helps it quickly find the best machine. These experiments are not just about games; they also teach us a lot about how computers learn and make choices in different and changing situations.

2 Methods

2.1 Objective and Setup

Our experiment aimed to investigate the effectiveness of different decision-making strategies in a scenario called the multi-armed bandit problem. In this setup, imagine a series of ten slot machines (referred to as bandits), each offering a different and unknown average reward. The challenge is to maximize the total reward over many trials. Our focus was on evaluating the ε-greedy strategy, which involves a balance between exploitation (choosing the machine that has so far seemed to give the best rewards) and exploration (occasionally trying other machines to discover if they offer better rewards).[3]

2.2 Strategies and Variations

We tested the ε-greedy strategy with varying levels of exploration, determined by the ε value. Specifically, we examined how the strategy performs with ε values of 0.01, 0.1, and 0.2. A smaller ε value indicates less frequent exploration, meaning the strategy predominantly sticks with the machine that has given the highest average reward to date. Furthermore, we explored the effects of different initial Q value estimates on the performance of the ε-greedy strategy. Two initial conditions were used: setting all initial Q estimates to 0, which provides a neutral, unbiased starting point, and random initialization, where initial Q values were set randomly from a normal distribution, mirroring the true value initialization. This allowed us to assess the impact of initial biases on the learning process. Additionally, to enhance the computational efficiency of our simulations, we employed a mathematical formula for the incremental update of Q values, denoted as

$$Q_{n+1} = Q_n + \frac{1}{n}(R_n - Q_n)$$

where Q_n is the estimated value after n plays, and R_n is the reward received in the n^{th} play. This formula enables faster recalculations of Q values, essential for handling a large number of iterations. [4]

2.3 Procedure

The experiment was conducted by running 2000 simulations (or runs) for each strategy variant, ensuring a comprehensive evaluation. In each run, the strategy interacted with the bandits over 1000 plays (or steps). We meticulously recorded the reward received at each step. This extensive data collection enabled us to analyze the average reward performance across all runs for every step, providing a robust assessment of each strategy's effectiveness over time. [1]

2.4 Expected Outcomes and Significance

Our primary goal was to determine the optimal balance between exploration and exploitation for maximizing long-term rewards. We hypothesized that a strategy incorporating some level of exploration is crucial, as relying solely on the initially perceived best option could lead to missed opportunities for greater rewards. This experiment holds significant implications for understanding decision-making processes in various real-world scenarios, where choices often have to be made with incomplete information, such as in finance, healthcare, and artificial intelligence.

3 Results

In this experiment, we employed the ε-greedy strategy to solve a multi-armed bandit problem with 10 arms. The experiment was conducted under two different initial conditions:

1. The Q values (q *estimated*) were initially set to zero.

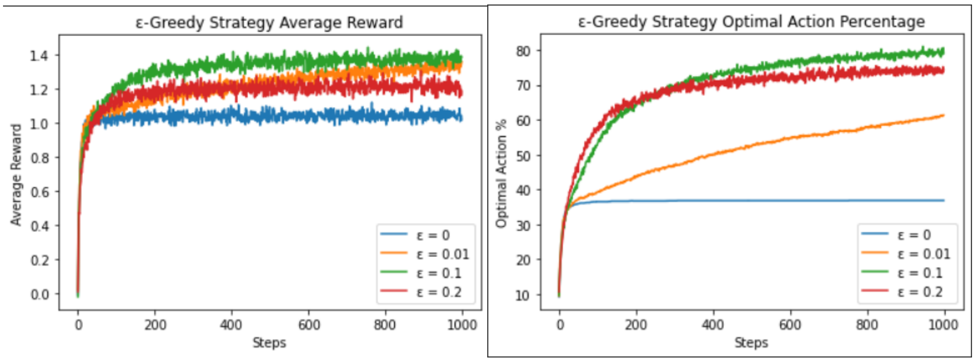
2. The Q values were initialized randomly from a normal distribution, the same as the true values (q *true*).

The ε-greedy strategy was tested with four different ε values: *0* (*pure exploitation*), *0.01*, *0.1*, and *0.2*, across 2000 simulations, each consisting of 1000 steps. The primary metrics for evaluation were the average reward and the percentage of optimal actions taken.

3.1 Zero Initial Q Estimates

3.1.1 Average Rewards

The average reward tended to increase over time for all ε values, indicating that the strategy was effective in learning the best bandits. A notable observation was that with ε=0 (pure exploitation), the average reward was generally lower than the other ε values, especially in the early stages. This suggests that some degree of exploration (non-zero ε) is beneficial in identifying more rewarding arms.



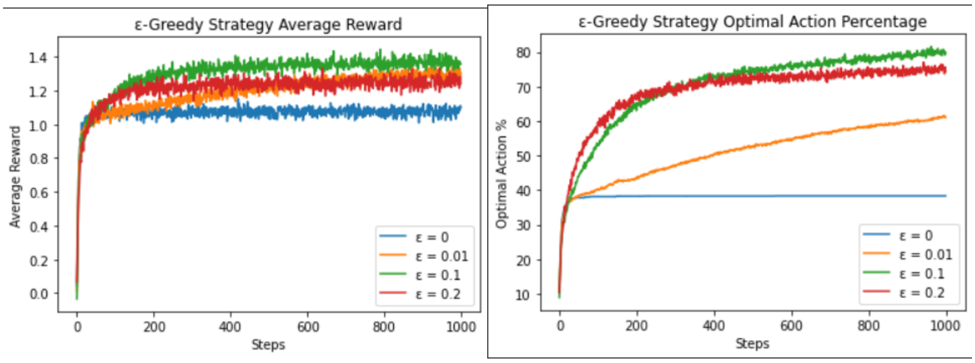
(a) Average Rewards (b) Optimal Action Percentage

Figure 1. Q Estimates as Zeros

3.1.2 Optimal Action Percentage

The percentage of optimal actions taken was noticeably higher for lower ε values. With ε=0, the strategy consistently exploited the bandit with the highest estimated value, leading to a high percentage of optimal actions once the best bandit was identified. As ε increased, the frequency of exploration (and thus the chance of selecting suboptimal bandits) increased, leading to a lower percentage of optimal actions.

3.2 Random Q Estimates



(a) Average Rewards (b) Optimal Action Percentage

Figure 2. Q Estimates as Normal Distribution

3.2.1 Average Rewards

Interestingly, the average rewards for $\varepsilon = 0.2$ were notably higher when starting with random initial Q values, as opposed to zero initialization. This suggests that higher randomness in initial guesses may benefit the strategy when exploration is more frequent (as is the case with a higher ε value). For $\varepsilon = 0.2$, the random starting points likely encouraged broader exploration early on, leading to the discovery of more rewarding options. However, for lower ε values (0.01 and 0.1), the results were largely similar to the zero-initialized scenario, indicating that the impact of initial estimates diminishes with less frequent exploration.

3.2.2 Optimal Action Percentage

Consistent with the average reward findings, the percentage of optimal actions taken showed a significant improvement for $\varepsilon = 0.2$ with random initial Q values, compared to zero initialization. This improvement reflects that the strategy with higher exploration ($\varepsilon = 0.2$) benefited from the diverse starting points, quickly aligning its choices with more rewarding options. For lower ε values, however, the optimal action percentages remained comparable to those observed in the zero-initialized setup. This pattern suggests that while random initial estimates can enhance the learning process under conditions of high exploration, their influence is less pronounced when exploration is less frequent.

3.3 Analysis

3.3.1 Effect of Exploration (ε Value)

Our results show that trying out new things (exploration) is really important in the ε-greedy strategy. When we just stick to what we know best (pure exploitation, ε=0), it can work well eventually, but it's usually better to mix in a little bit of trying new options (having a small ε value). This mix helps us learn better and faster. [2]

3.3.2 Impact of Initial Q Value Estimates

The way we first guess the value of each option (initial Q values) really affects how well the strategy works. When we start with random guesses, it takes longer to learn and doesn't work as well at first. This shows us that our first guesses can really influence how we learn in these kinds of problems.

3.3.3 Balancing Exploration and Exploitation

Getting the right balance between sticking with what we know (exploitation) and trying new things (exploration) depends on the situation. Some exploration is definitely important, but too much can lead us to make sub-optimal choices more often.

3.3.4 Learning Dynamics

The way we learn changes based on how we begin. Starting with all guesses at zero means we are not leaning any particular way at first, and we get to build up our understanding from the ground up. But starting with random guesses means we have some early ideas that might not be right, and we need to work through these to learn properly.

4 Conclusion

In our test of the ε-greedy strategy with different settings, we found something interesting, especially when ε was set to 0.2 and we started with random guesses for each option's value. With $\varepsilon = 0.2$, which means trying new things quite often, the strategy worked better when we began with random guesses instead of starting all guesses at zero. This shows us that when we're open to exploring more, tarting off with a variety of guesses can help us find good options faster. This was especially true for $\varepsilon = 0.2$, but not as much for smaller ε values. Our experiment shows how important it is to think about how much we want to explore and how we first guess the value of our options when we are learning to make good choices.

Acknowledgements

We extend our heartfelt gratitude to Professor Alexander Bihlo, whose invaluable guidance and expertise greatly enriched the content of this study. His insights and suggestions were crucial in shaping the research direction and interpreting the results.

References

[1] Collier M. and Llorens H.U., Deep contextual multi-armed bandits, *arXiv preprint arXiv:1807.09809* (2018).

[2] Kuleshov V. and Precup D., Algorithms for multi-armed bandit problems, *arXiv preprint arXiv:1402.6028* (2014).

[3] Maes F., Wehenkel L. and Ernst D., Learning to play k-armed bandit problems, in *4th International Conference on Agents and Artificial Intelligence (ICAART 2012)*, 2012 .

[4] Morimoto J., Foraging decisions as multi-armed bandit problems: Applying reinforcement learning algorithms to foraging data, *Journal of Theoretical Biology* **467** (2019), 48–56, doi:https://doi.org/10.1016/j.jtbi.2019.02.002.
URL https://www.sciencedirect.com/science/article/pii/S0022519319300566