SOLUTION OF A MODIFIED MULTI-ARMED BANDIT PROBLEM BASED ON REINFORCEMENT LEARNING

Xiang Zheng
21307110169
mango789

Jiaao Wu
21307130203
Julius Woo

Zihao Cheng
21307130080
football prince

December 10, 2023

ABSTRACT

In this work, we present a novel approach to solving a modified multi-armed bandit problem using reinforcement learning. We extend traditional policy strategies by introducing a "misguide" mechanism, where the agent strategically introduces randomness in its decision-making to interfere with the opponent's learning process. Additionally, we explore the combination of off-line recorded information and on-line learning to strike a balance between specialization and generalization of our agent. Our empirical results demonstrate the effectiveness of these innovations in improving the performance of our agent in competitive scenarios.

1 Introduction

The multi-armed bandit problem has been extensively studied in the context of reinforcement learning, where an agent must make a sequence of decisions to maximize its cumulative reward. In this work, we consider a modified version of the problem where two agents compete against each other. Traditionally, agents focus on maximizing their own rewards based on historical information, overlooking the possibility of minimizing their opponent's rewards.

We propose an extension to existing policies by introducing a "misguide" module, allowing our agent to strategically choose sub-optimal actions to interfere with the opponent's learning process. This introduces a game-theoretic element to the problem, where the agent not only seeks to maximize its own rewards but also aims to disrupt the opponent's decision-making.

Furthermore, we investigate the combination of off-line recorded information and on-line learning. Our analysis reveals a trade-off between the early-term performance of pure off-line agents and the long-term adaptability of on-line learners. To address this, we introduce a weighted sum approach, dynamically adjusting the influence of off-line and on-line information over time.

Innovative common-ideas, such as the use of global variables to store historical information and starting the search from the opponent's last action, are discussed. These ideas contribute to the robustness and efficiency of our agent, leading to improved performance in both the *Connectx* and *Santa* scenarios.

The remainder of this paper is organized as follows: Section 2 explores the extension of policies, Section 3 discusses the combination of off-line and on-line information, and Section 4 introduces innovative common-ideas. We conclude with a discussion of our findings in Section 5.

2 Extension of policy

The competition of two different agents in this *Santa* problem can be formulated as

$$\min_x \max_y f(x, y) \tag{1}$$

which is a minmax problem often discussed in optimization. The y stands for the opponent's agent and x stands for our agent. If y is a very "optimize" agent, then we always want to find the Nash equilibrium in the outer layer of the problem. This is exactly what we did in Final Project 2 before.

However, the minmax problem can be relaxed as

$$\max_y \min_x f(x, y) \leq \min_x \max_y f(x, y) \tag{2}$$

where the turns of two agents are exchanged. The solution of these two problems are quite closed under some mild smooth or Lipschitz conditions. This offers us a brand new perspective of this problem: Previously, we are always trying our best to maximize our own reward based on the historical information, neglecting the fact that we can also achieve it by minimizing our opponent's reward. Notice that our choice in each step is visible to our opponent's agent, and their agent must learn some information from our agent, we decide to extend our policies by adding the **misguide** module. The misguide mechanism is defined as

```
if accumulate_step % 35 == 0 and 100 <= step <= 1000:
    return np.random.choice(my_sub_best)
```

where we choose a random bandit from our sub-optimal choices in the medium-term to interfere opponent's approximation of the threshold. Figure 1 and 2 are the experiment results which showcase the effectiveness of our method.

Figure 1: The total reward over steps

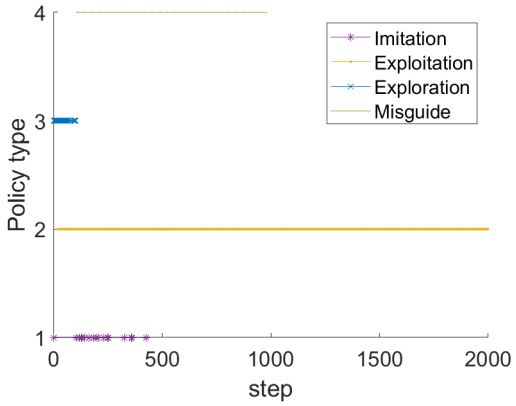

Figure 2: The type of policy chosen over steps

3 Combination of off-line record and on-line information

The agents we developed before in Final Project 2 are all pure on-line learners, which indicates that these agents only have access to the information in a specific competition that they are competing for. This assures the speciality of our agents while ignoring the generalization of the combat between our agent and various kinds of agents such as ϵ-greedy, UCB and softmax.

Based on our empirical results, the pure off-line agent will perform much better in the early-term of the game, but it'll lag behind other on-line agent as the game progresses. Conversely, the on-line agent will behave poorly at the early stage of the game and grows wiser and wiser as the game progresses. Hence there is a trade-off between off-line and on-line(generalization and speciality) in the development of our agent.

The idea of weighted sum is implemented here to determine the different weight of off-line and on-line over steps. At first, the weight of the off-line record is 1 and the weight of the on-line information is 0. As the time progresses, the weight of the off-line record decreases to 0 and the weight of the on-line information increases to 1. This insures that our agent is always at its optimality in each step.

4 Innovative common-ideas

In this section, we'll introduce two innovative common-ideas used through *Connectx* and *Santa* which greatly improve the performance of our agent. One is the global variable, the other one is starting from opponent's last action.

4.1 Global variable

In Final Project 1, we use the global variable `current_state` to maintain the Monte Carlo Tree already constructed in previous rounds. This reduces the time for repeated search greatly and our agent is able to search deeper to achieve better result.

In Final Project 2, we use several global variables to record the historical information of our agent and opponent's agent for the approximation of `threshold` and choices of policies. It simplifies the code structure and guarantees the robustness of our agent.

4.2 Opponent's last action

In Final Project 1, the search in the current round is started from opponent's last action. This implies that our agent will search from opponent's perspective to retaliate our opponent. The strategy is proved to slightly improve our agent in the medium-term of the game by the analysis of *Kaggle* competition results.

In Final Project 2, the last bandit taken by our opponent is taken full consideration in the choices of policies since our goal is to achieve a win rate higher than 50%. In the early-term of the game, our agent will imitate our opponent if the bandit taken by our opponent is not chosen by our agent to avoid the losses of information. In the medium-term, the imitation policy dominates other policies to learn information from our opponent for a better approximation of `threshold`. As our agent gains more information about the `threshold`, it will almost follow its own approximation, but if the "last bandit" is taken more frequently by opponent than us, our agent will imitate the last action to compensate for the gap. This is a correction mechanism for a higher win rate.

5 Discussion

Our experiments and innovations have shed light on several key aspects of the modified multi-armed bandit problem. The introduction of the "misguide" mechanism, which strategically introduces randomness into decision-making, has proven to be a powerful tool in disrupting the opponent's learning process. The empirical results, as depicted in Figures 1 and 2, showcase the effectiveness of this approach in achieving a competitive edge over various opponent strategies.

The combination of off-line recorded information and on-line learning introduces a dynamic balance between specialization and generalization. Our weighted sum approach allows our agent to leverage the strengths of both paradigms, resulting in a versatile strategy that performs well across different stages of the game. This trade-off, as demonstrated in our experiments, highlights the importance of adaptability in competitive scenarios.

The incorporation of global variables and the consideration of the opponent's last action have proven to be innovative common-ideas that significantly enhance the efficiency and

robustness of our agent. The use of global variables facilitates information storage and retrieval, reducing computational overhead. Starting the search from the opponent's last action provides a strategic advantage, allowing our agent to retaliate effectively.

One interesting observation is the correction mechanism introduced in the imitation policy based on the opponent's last action. This correction mechanism serves as a dynamic adjustment, compensating for performance gaps that may arise over the course of the game. As our agent gains more information about the optimal threshold, the correction mechanism ensures a higher win rate by adapting to the opponent's choices.

In summary, our approach to the modified multi-armed bandit problem, incorporating the "misguide" mechanism, balancing off-line and on-line learning, and utilizing innovative common-ideas, has demonstrated promising results. The flexibility and adaptability of our agent make it well-suited for competitive scenarios where opponents may employ diverse strategies. Further research could explore additional refinements and extensions to enhance the agent's performance in a broader range of environments.