Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

A Study of Exploration-Exploitation Strategies in Unconventional Situations

Zhangqi Zheng *

College of Sciences, Nanjing Agricultural University, Nanjing, China

* Corresponding Author Email: 23123223@stu.njau.edu.cn

Abstract. The exploration-exploitation problem is a central challenge in Reinforcement Learning (RL), and the Multi-Armed Bandits (MAB) serve as its foundation, providing a classical paradigm for exploration and exploitation strategies. With the development of big data and deep learning, the application of RL models in online learning, recommender systems, and other fields has become increasingly complex, giving rise to variants of models such as multi-objective optimization and stochastic adversarial. This paper reviews the limitations of classical algorithms such as ε-greedy, Upper Confidence Bound (UCB), and Thompson sampling in multi-armed bandit systems. It explores potential improvements in unconventional environments as far as the problem of rewards is concerned, which includes the case where the reward signal is time-varying and comes with some delay. And the limitations of traditional MAB, i.e., the inability to utilize contextual information, are explored in a relevant way. Meanwhile, scenario-oriented application-oriented MAB that are differentiated for real-world situations are mainly investigated as multi-objective, adversarial two major application-driven MAB. The cross-disciplinary characteristics of its variant algorithms are also examined to provide relevant algorithmic references for future research.

Keywords: Exploration-exploitation problem; multi-armed bandits; reinforcement learning.

1. Introduction

With the gradual increase in computer arithmetic and the creation of several relevant databases in the last decade. Machine learning has become a focal point for development in many specialized fields. In reinforcement learning, deep learning is gradually becoming the direction in which many learning algorithms converge. Based on this, excellent algorithms such as Implicit Q-Learning (IQL) [1] and Conservative Q-Learning (CQL) [2] were born. Meanwhile, in the field of online reinforcement learning, i.e., real-time exploration and learning during the learning process, excellent algorithms such as Deep Reinforcement Learning (DQN) [3], Proximal Policy Optimization (PPO) [4], were gradually born [5]. However, at a time when the emphasis is on specialization and lightweight, the above algorithms require larger resources due to their higher thresholds. Therefore, some unconventional reinforcement learning content is either poorly adapted or trapped in local optimal solutions. They have difficulty accomplishing the trade-off between exploration and utilization with low complexity requirements.

Exploration-exploitation is the core problem of reinforcement learning and directly affects algorithm performance and convergence. MAB provides a classic paradigm for this problem as an important precursor to reinforcement learning. Intelligentsia must interactively trade-off between exploring the unknown and executing optimal decisions without full knowledge of the environment, thus gradually learning the optimal strategy.

Focusing on MAB systems, this paper will review and analyze recent algorithmic research in reinforcement learning targeting exploration and exploitation strategies, especially the challenges and solutions in unconventional environments (e.g., noise, dynamic changes, reward delays). Starting from the theoretical foundations, this paper will present existing algorithms and applications step by step and discuss the key issues in current research and the potential for future research.

2. Traditional strategies and their limitations

Starting from a multi-armed slot machine problem, classical reinforcement learning algorithms possess several of the most commonly used exploration-exploitation strategy algorithms. They include the ε-greedy method, the UCB algorithm, and the Thompson sampling algorithm.

2.1. ε-greedy methods

In most cases, ε-greedy is a simple and efficient exploration-exploitation strategy with low cumulative regret values. Its exploration is based on a fixed value of ε, e.g., one exploration is expected in every 10 actions for ε = 0.1. This allows ε-greedy to traverse all options at larger action counts while maintaining 9 times no increase in cumulative regret. However, its cumulative regret is linear and cannot be converged.

As shown in figure 1, in a 10-arm MAB experiment with 5000 runs, the ε-greedy method has not converged despite exhibiting low cumulative regret.

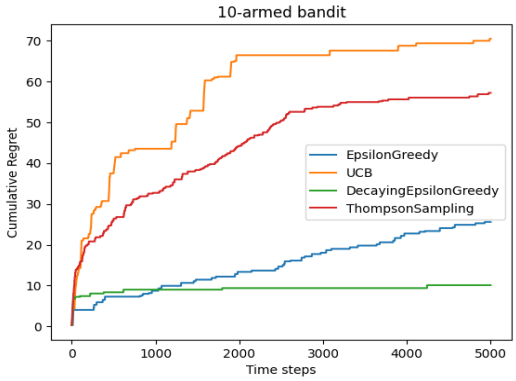

Fig 1. Comparison of Algorithm Cumulative Regret.

To solve this problem, the researcher proposed a variant of ε-greedy, ε-decaying. ε-decaying introduces a dynamic tuning approach to optimize learning by gradually decreasing the exploration probability ε. In the early stages of reinforcement learning, a larger ε promotes environment exploration. At the same time, ε gradually decays as the intelligence gains more experience, allowing the strategy to utilize the optimal action to improve long-term rewards. It is a precursor to value function-based methods such as Q-learning and DQN.

The decay parameter of ε-decaying is critical. If exploration is over-encouraged, especially when there are fewer options, it can lead to higher regret values; conversely, if the attenuation parameter is too low, full exploration is difficult when there are more options. Thus, the regret stability of ε-greedy is poor for fixed parameters.

As shown in figure 2 the cumulative regret of the ε-decaying algorithm with 200 arms and 1000 runs. It is clear that the parameter settings significantly impact the formulation of optimal solutions for atypical models (e.g., with a low number of runs).

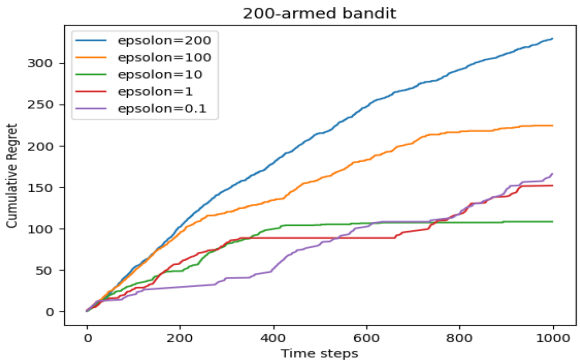

Fig 2. Comparison of Algorithm Cumulative Regret.

2.2. Upper Confidence Bound

The UCB algorithm is based on probability theory and mathematical statistics, and its core idea is to be optimistic about "not fully explored" options. Specifically, when uncertainty about an option is high, it may not have been fully explored. Therefore, when choosing an action, it is important to consider its uncertainty and the average return of the current action.

$$a_t = \arg\max\left[Q(a) + c\sqrt{\frac{\log t}{2(N(a_t)+1)}}\right] \quad (1)$$

a_t where the action is selected at the time step t. $Q(t)$ is the current average reward estimate for the action a. $N(t)$ is the number of times the action has been selected. $c$is a hyperparameter controlling the degree of exploration.

In the early stages of the UCB algorithm, due to the low number of action choices, the under-explored actions have a larger upper confidence bound, i.e., the optimistic estimates are weighted more heavily, and thus the algorithm tends to select these unexplored options, encouraging exploration. As the stage progresses, the increase in the number of action choices leads to a gradual decrease in the upper confidence bound, and the choices are gradually biased towards the option with the higher expected payoff, and convergence is achieved.

However, it has the problems of high cumulative regret and slow convergence. In addition, the UCB algorithm is less robust to delay perturbations and provides little adaptation to adversarial environments.

The following figure demonstrates the cumulative regret of the UCB algorithm in the face of an adversarial environment. By running the purple curve ten times and averaging the values, it can be seen that it exhibits a linear accumulation of regret. Following this, the failure of the UCB algorithm in complex scenarios is verified.

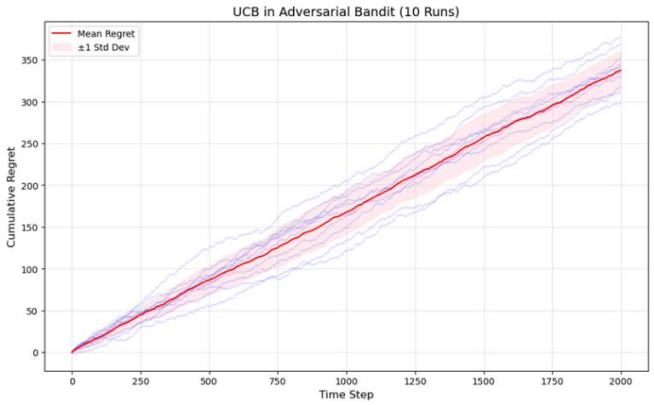

Fig 3. UCB in Adversarial Bandit (10 Runs).

2.3. Thompson Sampling

Thompson Sampling is very friendly to covariate distributions, such as Beta-Bernoulli and Normal-Normal, with an easy update process. However, when encountering non-conjugate distributions, Thompson Sampling requires the use of numerical methods or variational inference, which is computationally expensive.

$$a_t = \arg max\theta_i^{(t)}, \ \ \theta_i \sim Beta(\alpha_i, \beta_i) \quad (2)$$

a_t where the action is selected at the timestep t. $\theta_i^{(t)}$ is the probability distribution of each arm at the current time. Thompson Sampling is selecting the optimal reward arm from each option and

updating the distribution information to obtain information about the arm. However, Thompson Sampling mostly uses the beta distribution for iteration. While it is very friendly to many covariate distributions (e.g., normal-normal) and has an easy update process, it also lays the groundwork for its drawbacks.

Similar to the UCB algorithm, Thompson Sampling assigns the same initialization reward to each rocker at an early stage, thus overcoming the error of "extracting rockers" and traversing all options during the learning process. When the number of actions reaches a certain number, based on the law of large numbers, it can be shown that the probability distribution predicted by Thompson Sampling will converge to the actual probability distribution and tend to a normal distribution, proving its prediction's unbiasedness. Therefore, as the number of actions increases, the rocker arm with the largest selection reward will gradually dominate, thus ensuring the algorithm's convergence.

However, as mentioned earlier, Thompson Sampling is computationally expensive when encountering non-conjugate distributions that require numerical methods (e.g., Markov chain Monte Carlo MCMC) or variational inference. Furthermore, while Thompson Sampling performs well in independent multi-arm slot machine problems, the algorithm requires more complex modeling and inference when extended to associative arms or high-dimensional state-action spaces in reinforcement learning. More importantly, if the reward function changes over time, Thompson Sampling may have difficulty adapting in time [6].

3. Optimisation of conventional strategies

With the rapid development of big data and deep learning technologies, the MAB problem in application areas such as online learning, recommender systems, and advertisement optimization has evolved into various variants with special challenges. Especially in the big data environment, the dramatic growth of data volume and the complexity of information expose the limitations of the exploration-exploitation strategy based on the traditional MAB model.

These variants involve challenges regarding delayed feedback, contextual information, and non-smoothness. Traditional dobby slot machine models struggle to meet the demands of complex and dynamic real-world applications.

3.1. Situational feedback mechanisms

Conventional MAB environments ignore changes in their behavior to the environment and the feedback given by the environment. Their strategy of simply using rewards is not very efficient in real-world scenarios such as personalized recommender system scenarios. This is because intelligences are likely to fall into local optimal solutions and converge to suboptimal ones during decision-making [7].

The Situational feedback mechanisms problem introduces additional contextual information that allows the learner to make a decision and obtain relevant information about the current environment or state when choosing an action in each round. This contextual information typically includes features relevant to decision-making, such as user behavioral data, historical information, or the state of the environment. The central feature is that rewards depend not only on the chosen action but are also influenced by the current context. The learner aims to optimize decisions based on context to maximize long-term rewards.

Depending on the actual situation, the utilization of situational information can be divided into two categories.

One is that Bastani et al. encourage conservative strategies under the huge cost of exploration in areas such as healthcare. At its core, it switches to an exploratory algorithm for minimal exploration by analyzing the randomness of the context only when inaccurate estimates are detected [8]. That is, it protects against the necessity of exploration and the emergence of huge losses.

The other, typified by short-form video platforms, is a situation where exploration is less costly and encouraged. Therefore, Yang et al. designed a ULCB-like algorithm to set up two confidence intervals, optimistic and pessimistic, to reduce exploration when the information above exhibits signals of abandonment [7]. As shown in the formulas below, when the previous reward is 0, a lower confidence level should be chosen, i.e.,c_0, whereas when a higher reward is obtained, a larger confidence level can be chosen, i.e., c_1.

$$a_t = \arg\max\left[Q(a) + c_0\sqrt{\frac{\log t}{2(N(a_t)+1)}}\right], r_{t-1} = 0 \tag{3}$$

$$a_t = \arg\max\left[Q(a) + c_1\sqrt{\frac{\log t}{2(N(a_t)+1)}}\right], r_{t-1} = 1 \tag{4}$$

Strategies in this scenario are therefore commonly analyzed in terms of interaction information. Either by predicting future information or judging based on past information in order to choose whether to explore or not.

3.2. Delays in the granting of rewards

In most cases in real life, rewards are not instantly available, but there is a time delay. In the case of medical treatment, for example, the effects of treatment may take some time to be observed. Delayed Feedback The rewards in the dobby slot machine problem are not instantly available, but rather, there is a time delay, which prevents the intelligent body from knowing the effect of its decision immediately after making the choice.

It is mainly characterized by the presence of delayed feedback, where the intelligences need to simultaneously process the reward feedback of past choices during each round of decision making. This latency increases the uncertainty and complexity of decision making, and the key difficulty lies in how to effectively estimate and utilize the delayed reward information for real-time decision making.

Masoudian et al. proposed an improved algorithm based on the FTRL framework by introducing regularized equilibrium terms in cumulative observed losses as a balancing term and combining skipping techniques and self-definition methods [9].

Not coincidentally, Howson et al. targeting delayed feedback in generalized linear dobby slot machines, similarly introduced a regularization term as a balancing term in maximum likelihood estimation based on the optimism principle. This method effectively improves the learning efficiency by reducing the delay penalty from the multiplicative term in previous methods to the additive term through delay adaptive confidence sets [10]. The formula is shown below:

$$L_t(\theta, \alpha) = \sum_{s=1}^{t} C_s^t \log\big(f(Y_t|X_t)\big) - \frac{\alpha}{2}||\theta||_2^2 \tag{5}$$

Included among these, $L_t(\theta, \alpha)$ denotes the regularized log-likelihood value at round t with respect to the parameter θ, which combines the observed data (via the log-likelihood component) as well as the regularization penalty. θ is the vector of unknown parameters to be estimated, which determines the relationship between actions and rewards in the model. α is the regularization parameter, which is used to control the strength of the penalty term, the larger α is, the more obvious the regularization effect is, so that overfitting can be avoided or the uniqueness of the objective function can be guaranteed. $\log\big(f(Y_t|X_t)\big)$ denotes the corresponding log-likelihood value of the reward Y_s when it is observed in rounds where X_s is the eigenvector or action selected for that round, and f is the probability density function or probability mass function of the reward as defined according to the model (e.g., generalized linear model).

In practice, delayed feedback often produces a lagging effect on the updating of the reward prediction of an intelligent body, which in turn affects the timeliness and accuracy of decision-making. To address this challenge, the introduction of a balancing term becomes an effective compensation mechanism. The balancing term aims to optimize between the stability and detectability of decision-making, ensuring that the intelligent body can maintain a stable use of existing knowledge while sensitively capturing and responding to new information when necessary, thus facilitating the dynamic adjustment of strategies. This approach not only mitigates the negative impact of delayed feedback on reward prediction updating but also provides a reasonable trade-off framework for the intelligence between exploration and utilization, which ultimately enhances the robustness and adaptability of the overall decision-making system.

3.3. Non-constancy of rewards

Conventional MAB environments have constant rewards, which ignores the fact that the reward equation in real-world situations is a time-varying equation. This question seems to contradict the situational feedback mechanism mentioned above, as it seems that prior information is not as useful in the present. However, since reward distributions are mostly slow to change in real-world situations, this means that unbalanced MAB does not have different weights for prior information: the overvalue is extremely low, and the near-term value is not very different from traditional MAB.

Therefore, the strategy adopted for this problem should be to focus on near-term rewards while reducing the impact of the forward distribution. Cavenaghi et al. proposed the f-Discounted-Sliding-Window Thompson Sampling algorithm. The algorithm combines a discount factor, which reduces forward distributional effects, with a sliding window approach, which focuses on recent effects, resulting in a dynamic balance between historical rewards and the most recent observations [11].

The NCC-UCRL2 algorithm was proposed by Ghoorchian et al. This algorithm, based on UCRL2, realizes a dynamic balance between historical data and the latest changes by introducing sliding window estimation and confidence interval construction techniques. The sliding window approach focuses on capturing the most recent information in the environment, while the confidence intervals provide a robust guarantee for the estimation of rewards and costs, thus achieving excellent decision performance in a non-stationary environment where feature observation is costly [12].

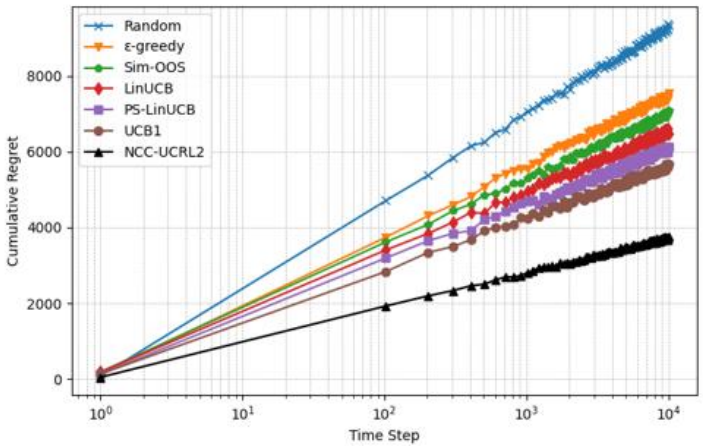

Fig 4. Cumulative regret of different policies in Non-constancy problem [12].

Although a linear relationship is presented in the above image, the NCC-UCRL2 algorithm presents a better convergence state since the horizontal x-axis is not linear. Since the state of the environment and data distribution often show dynamic changes over time, overly stale information may not accurately reflect the current real situation and may even introduce noise or bias, thus affecting the accuracy of prediction and the responsiveness of the model. In order to solve this problem, a commonly adopted strategy is to introduce a sliding window mechanism, which effectively weakens the interference of past information by setting a reasonable time horizon and using only data from the most recent period of time for distribution estimation. At the same time, since changes in reality are

often characterized by slow evolution, it is also necessary to adopt discount factors to gradually attenuate the previous information in time series data. The discount factor can retain historical information to a certain extent while making the model more sensitive to recent information, thus achieving a better regulation effect in balancing stability and flexibility. The combination of these two methods provides theoretical support and a practical basis for the accuracy of distribution estimation in dynamic environments.

4. Strategies geared toward real-world application differentiation

The conventional strategy optimization described above incorporates production practices, but still fails to meet the requirements of many reinforcement learning tasks. In order to enrich the application scope of traditional models and make them better able to cope with complex and changing practical problems, MAB models have moved towards the stage of intersection with other disciplines.

4.1. Multi-objective issues

One problem with MAB is that it can only optimize a single objective, and if the global optimal solution is considered, it needs to be combined with other optimization algorithms. In a multi-objective dobby slot machine problem, each action generates multiple associated rewards, each representing a different objective. The goal of an intelligent body is to optimize all objectives simultaneously rather than a single objective. These objectives may be independent or interdependent. Intelligentsia needs to make trade-offs between goals, and the challenge is to balance the conflicts and synergies between goals to achieve overall optimization. The central feature is the mutual trade-offs between goals, which brings up a key difficulty: how to effectively balance multiple goals, especially when there are conflicts between them. In addition, since objectives may be non-cooperative, the optimization strategy must ensure that the effectiveness of each objective is enhanced as much as possible without compromising other objectives.

Ahmadianshalchi et al. split Bayesian multi-objective optimization into two main problems, Improving the quality and diversity of Pareto allocations [13]. One is the use of MAB to dynamically select the acquisition function, and the other solves the batch selection problem through a determinant point process in a stochastic process.

Similarly, Crepon et al. extended the classical MAB model to a multi-objective setting by introducing the concept of vector-valued rewards and employing the Pareto frontier to transform the traditional single-objective optimization into a multi-objective optimization problem as a criterion for the optimal solution [14].

The intersection of multi-objective optimization and Pareto distribution has been studied in various fields such as operations research, statistics, machine learning and control theory. Multi-objective optimization focuses on optimization strategies that trade off different objectives, while Pareto distribution is used to characterize the distribution of inhomogeneous resources or benefits. In this cross-cutting context, stability analysis, decision theory, and probabilistic modeling are combined to help construct optimization methods that balance efficiency and fairness. Research in this direction not only involves mathematical optimization and statistical inference, but also relies on dynamic system analysis to ensure the reasonableness and applicability of solutions.

4.2. Battle network

With the development of a large number of intelligences, reinforcement learning considers not only the interactive learning process of one intelligence but also scenarios involving multiple intelligences against each other. The traditional MAB environment does not apply because its rewards involve opponent draws as well as uncertainty. Therefore, the cross-disciplinary characterization of MAB has the prospect of a superior exploration strategy.

Whereas adversarial environments are often associated with the automated control of bits of intelligence. Therefore, Huang et al. combined Lyapunov drift minimization with MAB learning with respect to system stability analysis. The proposed SoftMW and SSMW algorithms ensure the stability of queue length through time-varying learning rate and exploration rate, thus realizing efficient scheduling of complex queueing systems in adversarial environments [15].

The intersection of the adversarial dobby slot machine problem and cybernetics is mainly in dynamic decision-making and adaptive optimization. The uncertainty associated with the adversarial environment makes the decision-making process similar to a perturbation suppression problem in a control system, while the feedback mechanism and stability theory provided by cybernetics can be used to ensure the convergence and robustness of the strategy. Optimal control methods are used in this framework to adjust the trade-off between exploration and exploitation, while Lyapunov stability analysis ensures the feasibility of the decision strategy in the long-term game. By introducing a cybernetic perspective, the adversarial dobby slot machine problem can be modeled as an adaptive control system with disturbances, thus providing theoretical support for strategy optimization.

4.3. Integrated system

In integrated systems, MAB disciplines cross over more than ever, and scholars even combine multiple aspects of multidimensional thinking in order to determine better exploration-exploitation strategies. Take the Tsallis-INF algorithm proposed by Zimmert and Seldin as an example. Its based on Tsallis entropy regularization in information theory, which strikes a balance between exploration and exploitation. It is also combined with the analysis of the impact of adversarial perturbations in robust statistics, which incorporates the impact of adversarial perturbations into the analysis through self-constraints to improve the model fault tolerance. Optimal regret bounds in stochastic and adversarial multi-armed slot machine problems are achieved, successfully addressing the challenge of balancing algorithmic performance in both environments [11].

Similarly, the censored-UCB algorithm proposed by Tang et al. utilizes the reward truncation technique to deal with the strong dependency between reward and delay. By introducing a truncated-tail mechanism to reduce the bias caused by delayed feedback and combining it with an upper confidence bound strategy, an accurate estimation of the reward distribution under partial observation information is realized, which in turn achieves a near-optimal regret performance in a strictly delayed environment [12].

These examples fully prove that only through the deep integration with other disciplines, such as information theory and robust statistics, the traditional MAB model can realize a breakthrough at the theoretical and practical levels and better cope with the complex and changing problems in practical applications.

5. Conclusion

The current exploration-exploitation strategy in the direction of MAB shows a multipolar development trend, which optimizes the failure of the exploration-exploitation strategy to deal with feedback information in its own MAB as well as its heuristic algorithms and its inability to adapt to the lagging, transformative nature of rewards in a timely manner. Current research on MAB problem has shown a multipolar trend in exploration-exploitation strategies, which not only pushes the optimization of MAB itself, but also improves the limitations of its heuristic algorithms in coping with the insufficient processing of feedback information. At the same time, the MAB algorithm shows a trend of multidisciplinary cross-fertilization, which makes it able to adapt to complex and changing real situations.

In this paper, in a review of conventional MAB model optimization, although variants such as multi-objective, delayed feedback, and non-constant reward and their applications are systematically sorted out, there are still some shortcomings. First, there is a lack of unified theoretical framework between various types of extended models, which makes it difficult to achieve synergistic optimization in

complex scenarios. Second, some of the studies are only presented in a general way, lacking in-depth comparison and analysis of their methodology, theoretical basis, and application effects, which limits the systematicity and depth of the review to a certain extent.

Future research should focus on the unity and theoretical support of the model, explore an integrated framework that incorporates multi-objective optimization, situational feedback, and non-constant rewards, and incorporate methods such as Lyapunov stability and deep learning to improve the generality and scalability of the model. At the same time, the comparative analysis of different models in terms of convergence, stability, and performance should be strengthened to summarize their advantages, disadvantages, and applicable scenarios.

References

[1] Scott Fujimoto, Shixiang Gu. A Minimalist Approach to Offline Reinforcement Learning. Advances in neural information processing systems, 2021, 34: 20132-20145.

[2] Aviral Kumar, Aurick Zhou, George Tucker, Sergey Levine. Conservative Q-Learning for Offline Reinforcement Learning. Advances in neural information processing systems, 2020, 33: 1179-1191.

[3] Volodymyr Mnih, Volodymyr Mnih, Volodymyr Mnih, et al. Human-Level Control through Deep Reinforcement Learning. Nature, 2015, 518(7540): 529-533.

[4] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, Oleg Klimov. Proximal Policy Optimization Algorithms. arXiv preprint arXiv:1707.06347, 2017.

[5] Aleksandrs Slivkins. Introduction to Multi-Armed Bandits. Foundations and Trends® in Machine Learning, 2019, 12(1-2): 1-286.

[6] Julian Zimmert, Yevgeny Seldin. Tsallis-inf: An optimal algorithm for Stochastic and Adversarial Bandits. Journal of Machine Learning Research, 2021, 22(28): 1-49.

[7] Zixian Yang, Xin Liu, Lei Ying. Exploration, Exploitation, and Engagement in Multi-Armed Bandits with Abandonment. Journal of Machine Learning Research, 2024, 25(9): 1-55.

[8] Hamsa Bastani, Mohsen Bayati, Khashayar Khosravi. Mostly Exploration-Free Algorithms for Contextual Bandits. Management Science, 2021, 67(3): 1329-1349.

[9] Saeed Masoudian, Julian Zimmert, Yevgeny Seldin. A Best-of-Both-Worlds Algorithm for Bandits with Delayed Feedback. Advances in Neural Information Processing Systems, 2022, 35: 11752-11762.

[10] Benjamin Howson, Ciara Pike-Burke, Sarah Filippi. Delayed Feedback in Generalised Linear Bandits Revisited.International Conference on Artificial Intelligence and Statistics. PMLR, 2023: 6095-6119.

[11] Emanuele Cavenaghi, Gabriele Sottocornola, Fabio Stella, Markus Zanker. Non Stationary Multi-Armed Bandit: Empirical Evaluation of a New Concept Drift-Aware Algorithm. Entropy, 2021, 23(3): 380.

[12] Yifu Tang, Yingfei Wang, Zeyu Zheng. Stochastic Multi-Armed Bandits with Strongly Reward-Dependent Delays.International Conference on Artificial Intelligence and Statistics. PMLR, 2024: 3043-3051.

[13] Alaleh Ahmadianshalchi, Syrine Belakaria, Janardhan Rao Doppa. Pareto Front-Diverse Batch Multi-Objective Bayesian Optimization. Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(10): 10784-10794.

[14] Garivier, Aurélien, and Wouter M. Koolen. Sequential Learning of the Pareto Front for Multi-Objective Bandits. International Conference on Artificial Intelligence and Statistics. PMLR, 2024: 3583-3591.

[15] Jiatai Huang, Leana Golubchik, Longbo Huang. When Lyapunov Drift Based Queue Scheduling Meets Adversarial Bandit Learning. IEEE.ACM Transactions on Networking, 2024, 32(4): 3034-3044.